Introduction

Cross-Validation

Will Doyle

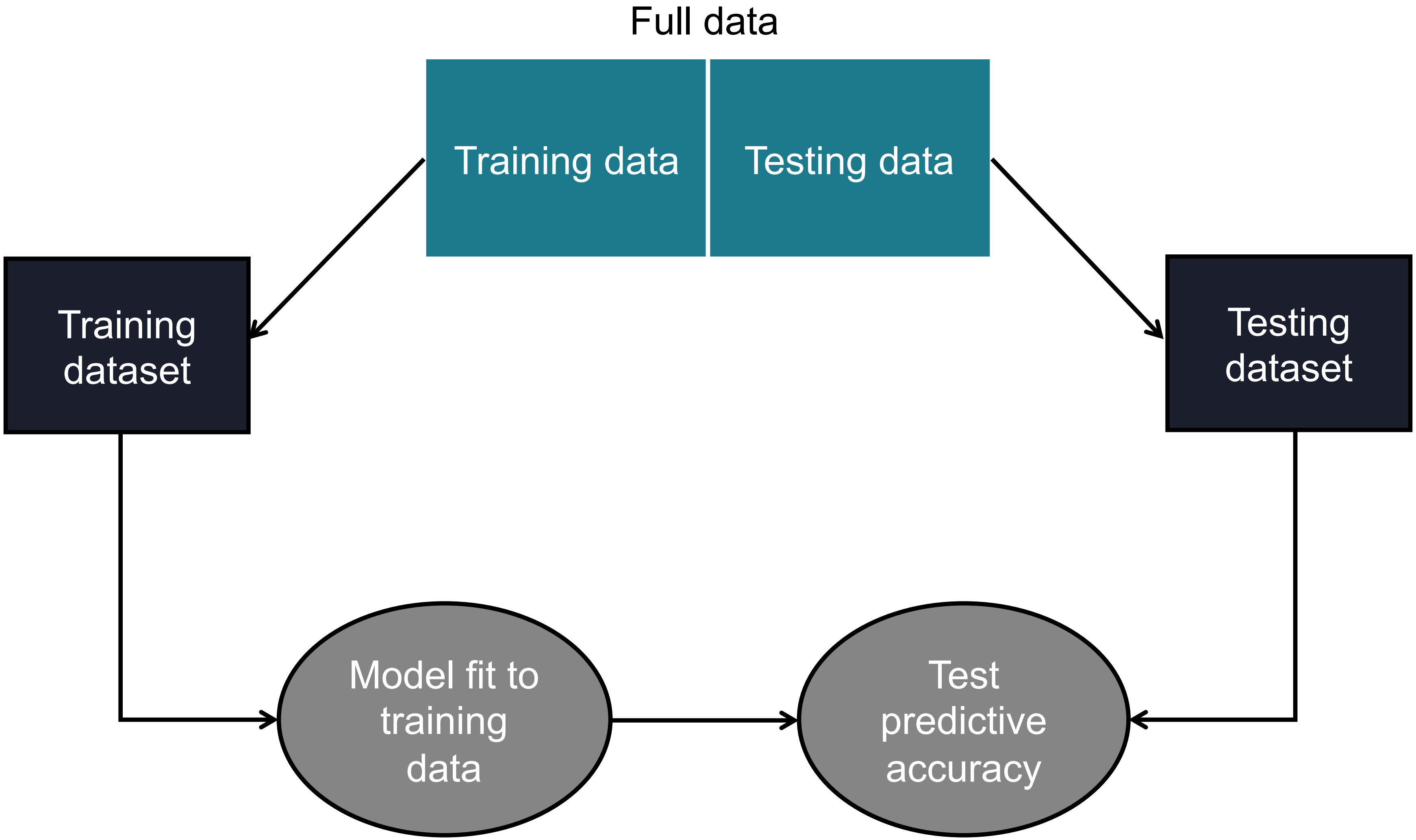
Review: The Problem of Prediction

- Our models can provide predictions given an observed set of characteristics.
- We can compare these to the actual data, but…
- We’ll likely be overconfident, as the predictions will be based on the data at hand.

Review: Training and Testing

- A “training” dataset is used to create a model.
- A testing dataset is used to (you guessed it) test the predictions made by the training dataset.
- The testing dataset must **not** be used to fit the model.
- Instead, the predictions from the model are compared with the actual values of the outcome from the testing dataset.

Training and Testing



Full data

Training data

Testing data

Training dataset

Testing dataset

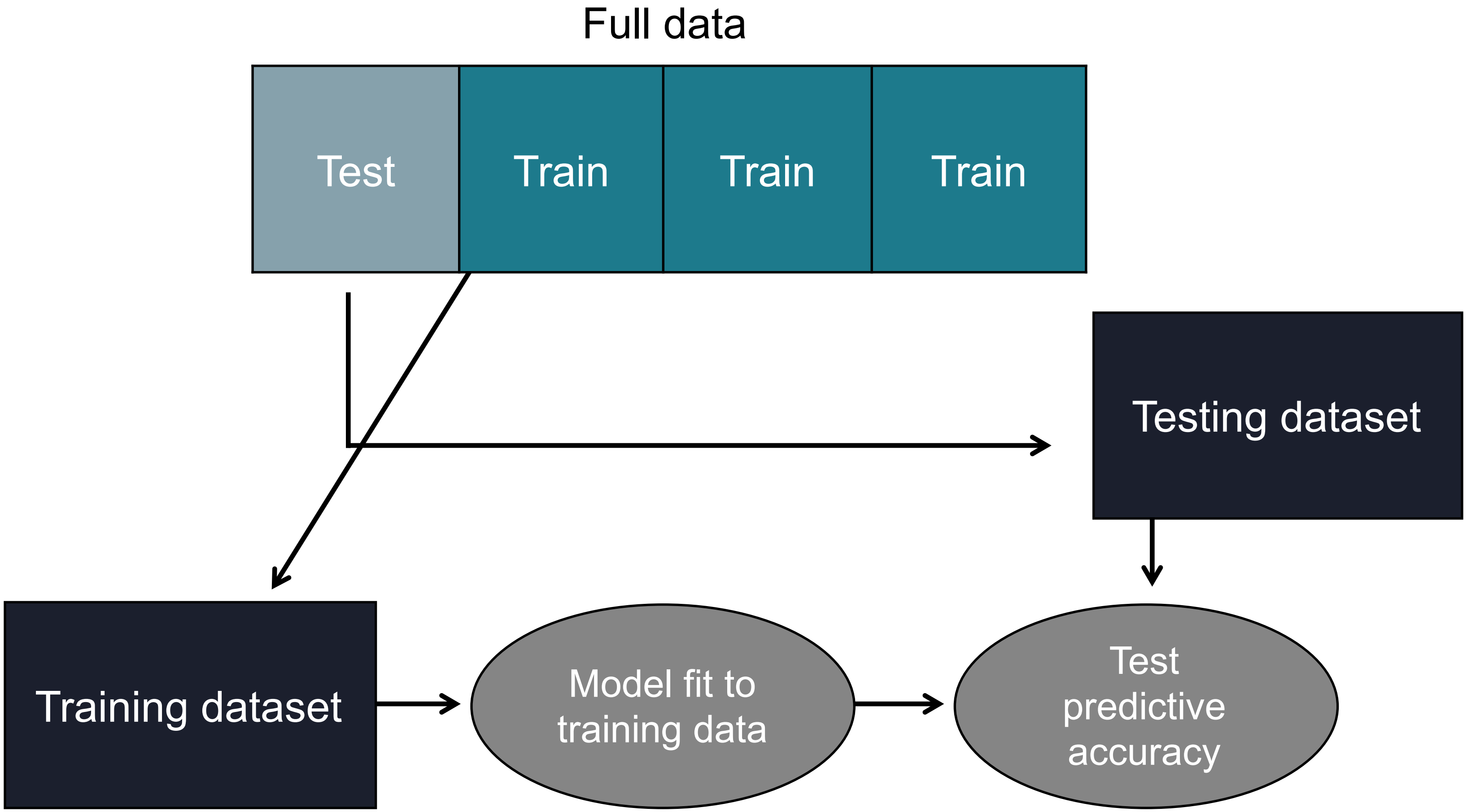
Model fit to training data

Test predictive accuracy

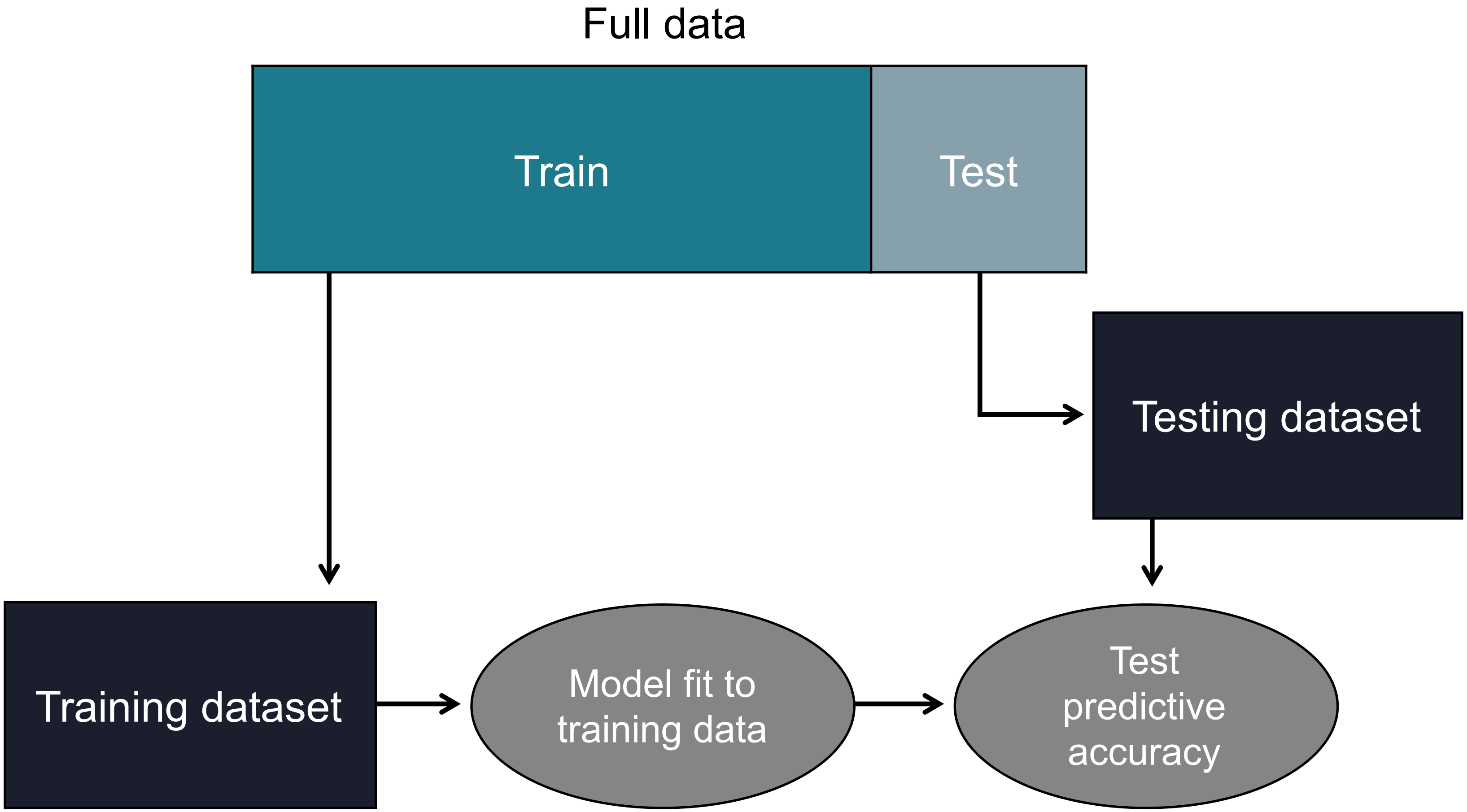
Cross-Validation

- Expands training and testing to involve multiple validation datasets
- Two ways to do this:
  1. K-fold cross-validation cuts the dataset into k non-overlapping, equal-sized sub-samples; one sample is retained for validation, the others are used for training the model
  2. Random partitioning cuts the dataset using a given proportion, then trains the model on the remaining data

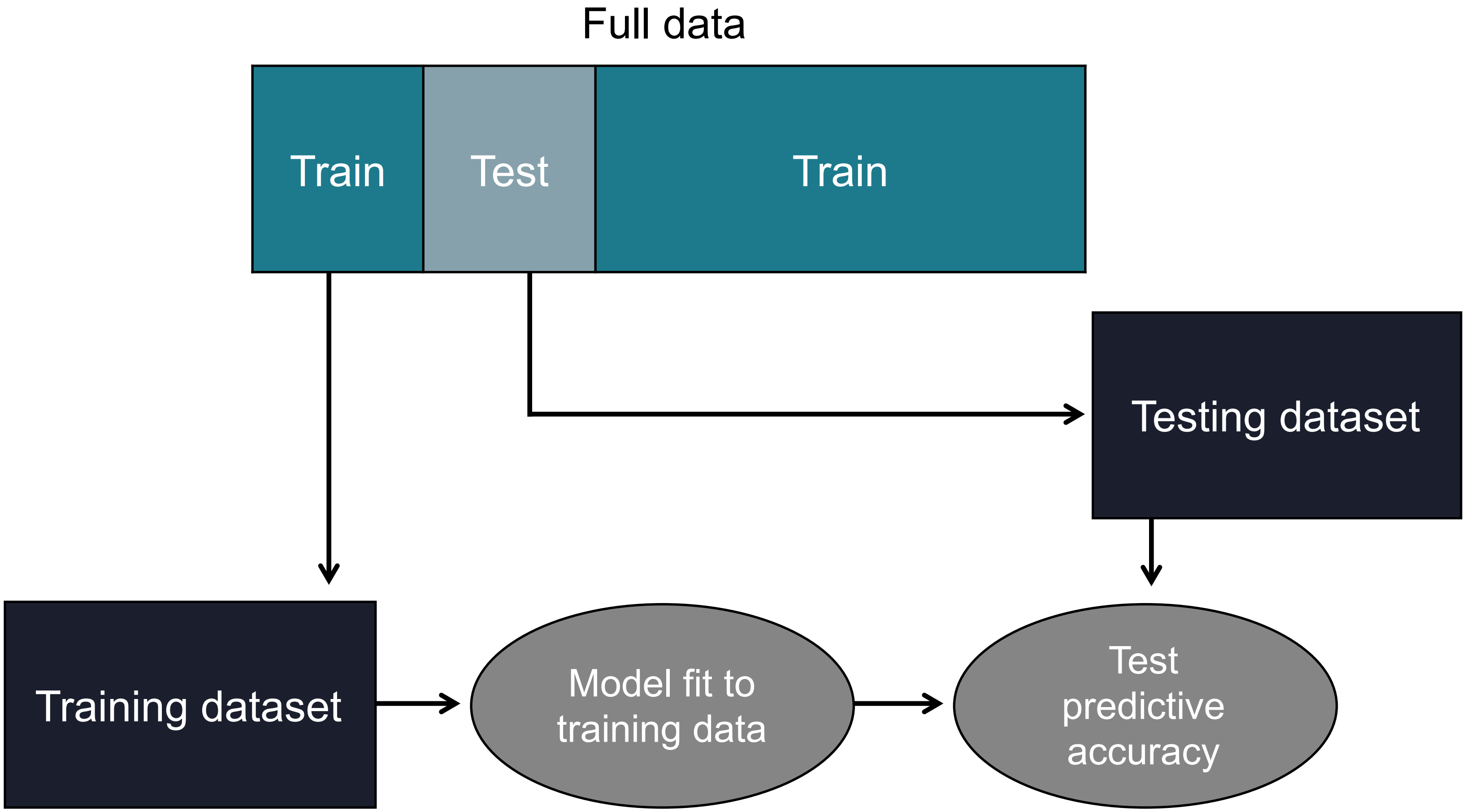
K-Fold Cross-Validation With K = 4



Random Partitioning



Random Partitioning



Repeat as many times as desired

RMSE and Cross-Validation

- We calculate RMSE from the testing dataset with each additional cross-validation
- We care about the distribution of the RMSE from these repeated cross-validations
- Can use mean and standard deviation OR
- Five-number summary
- Graphical summaries also can be helpful

VANDERBILT
PEABODY COLLEGE

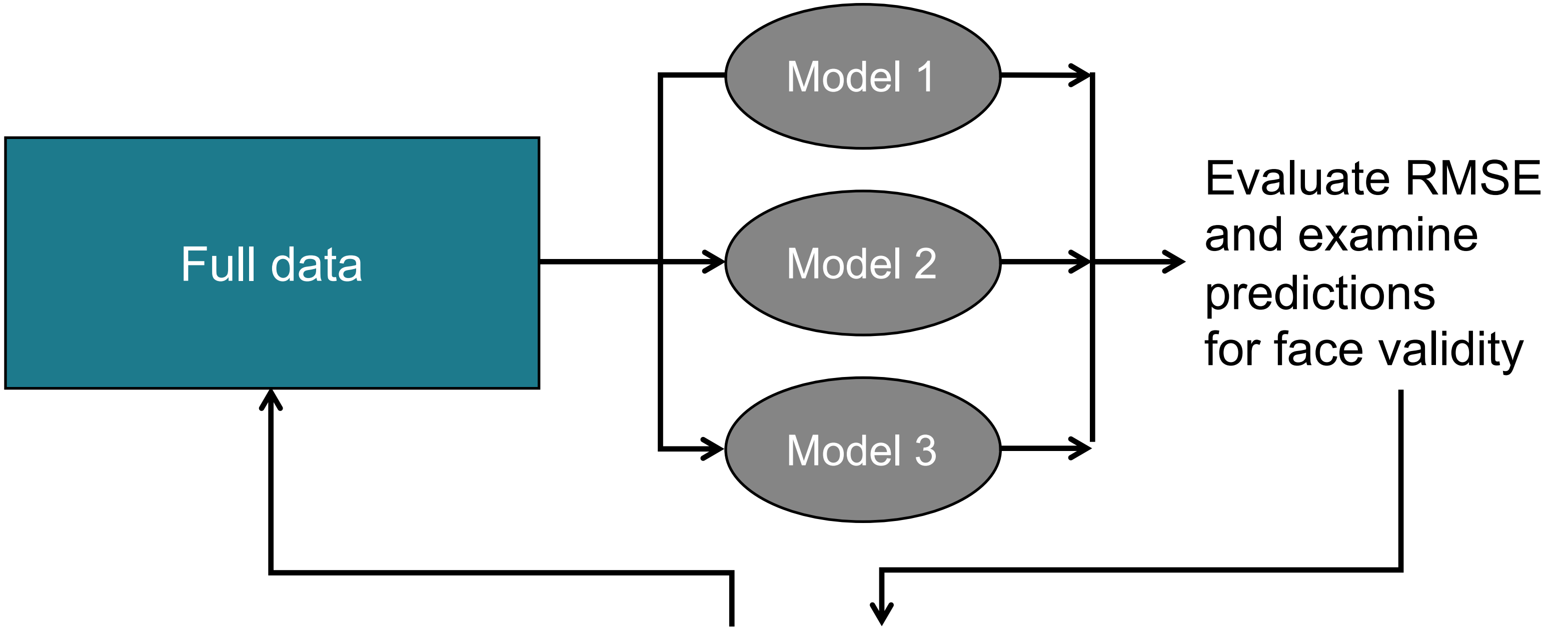
The Process of
Model Building and Evaluation

Will Doyle

Steps in Model Building and Validation

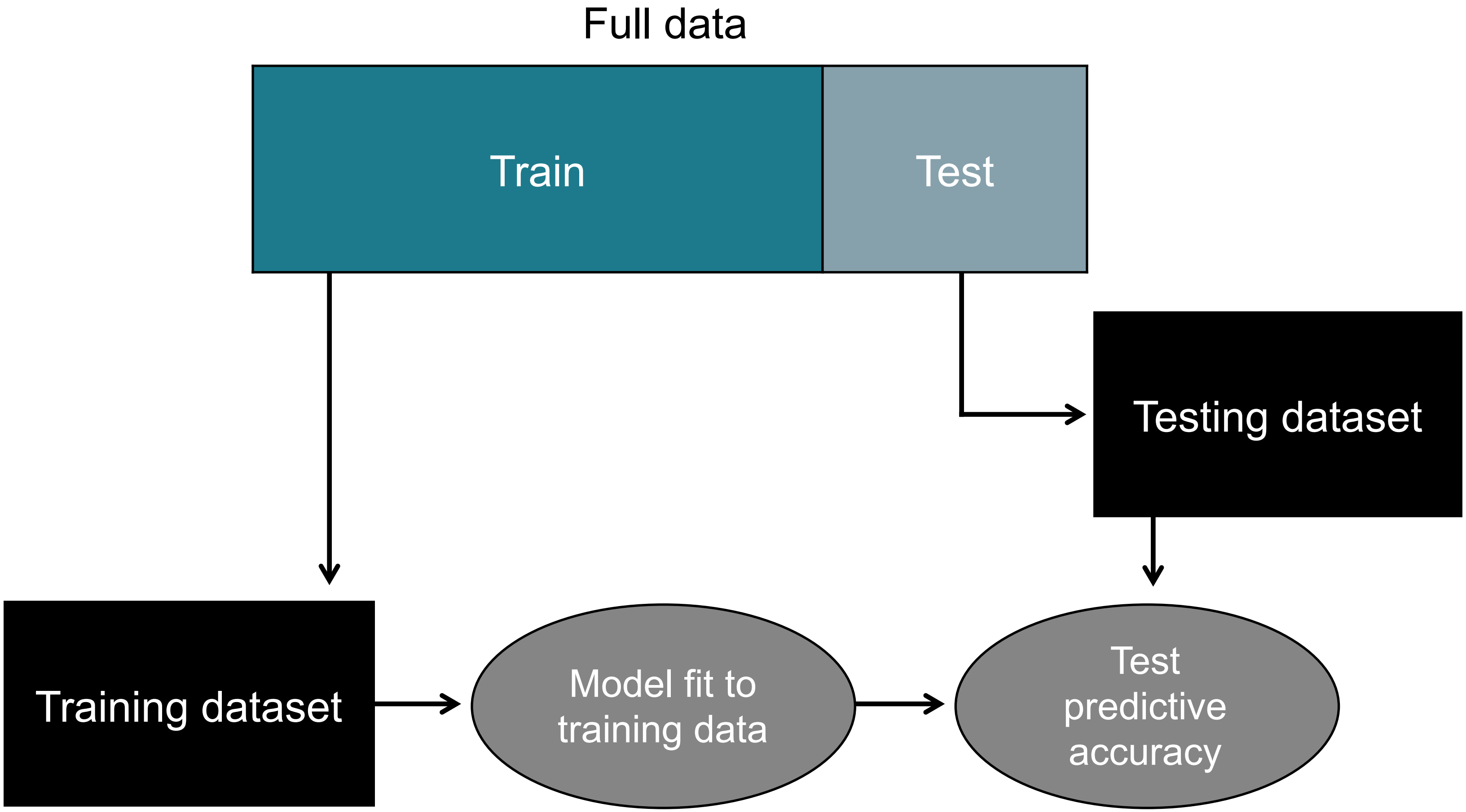
1. Exploratory data analysis
2. Build initial models
3. Check for face validity of predictions using full data (generate RMSE or similar for full dataset)
4. Repeat steps 1–3 as necessary
5. Cross-validate each model
6. Examine distribution of RMSE for each model
7. Choose model with best properties
8. Estimate parameters from full dataset, and use these on incoming real-time data
9. Evaluate and reassess

Process of Model Building and Validation

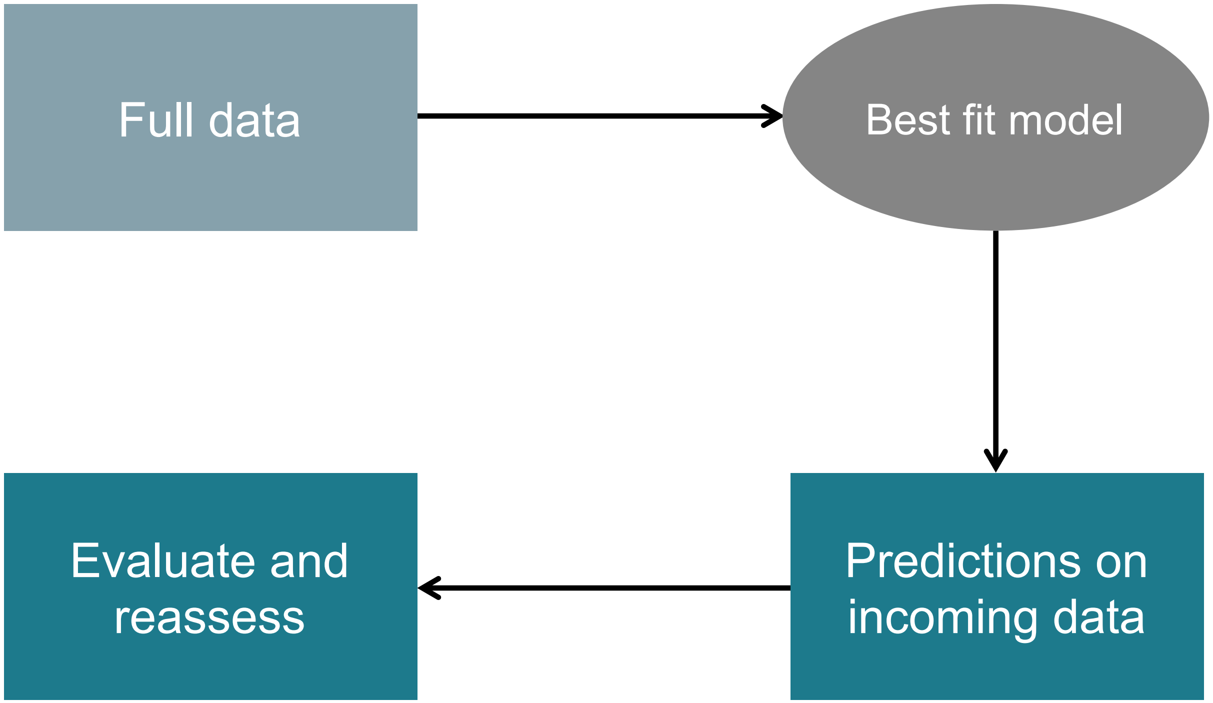


Repeat as necessary until you find appropriate models

Random Partitioning



Final Step



Full data

Best fit model

Predictions on incoming data

Evaluate and reassess

VANDERBILT
PEABODY COLLEGE