Model Evaluation & Selection

MCDC Virtual Workshop on
Teaching Introductory Machine Learning

2023-11-01

Quick overview of context

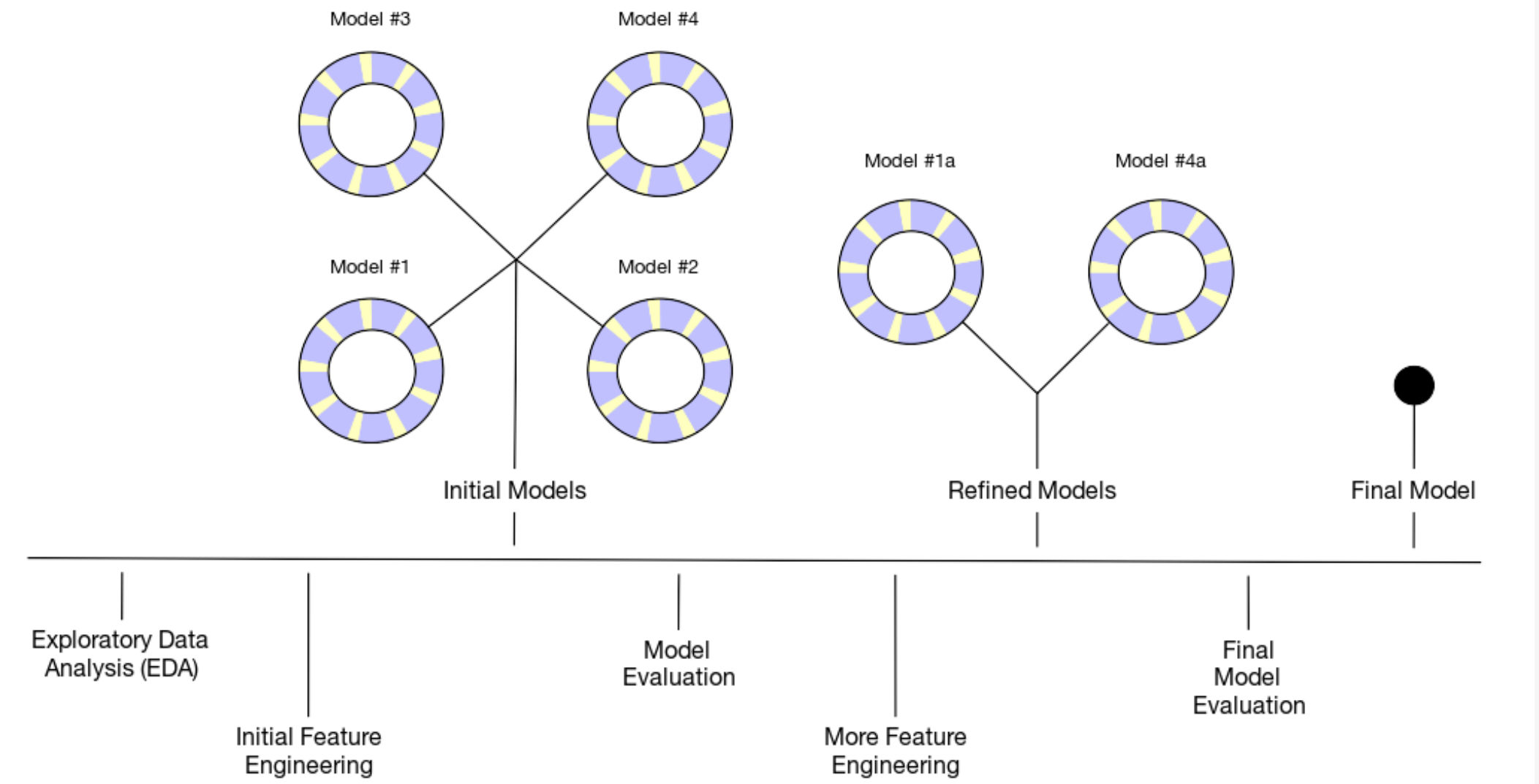
Where does model evaluation and selection come up during an introductory machine learning course?

1. Early on, maybe first day — How do we know if a model is any good?
2. During model building/training — How to pick the "best" model?
3. Discussion of final model — How will the final model preform?

Important related concepts/topics:

Data spending/allocation, performance vs comparisons, overfitting, bias-variance trade-off, metrics, & explaining models

Supervised learning schema



A schematic for the typical modeling process from Tidymodeling with R

How do we know if a model is any good?

Main objective is prediction!

Take a set of predictors/features X and use them to predict an outcome/target Y

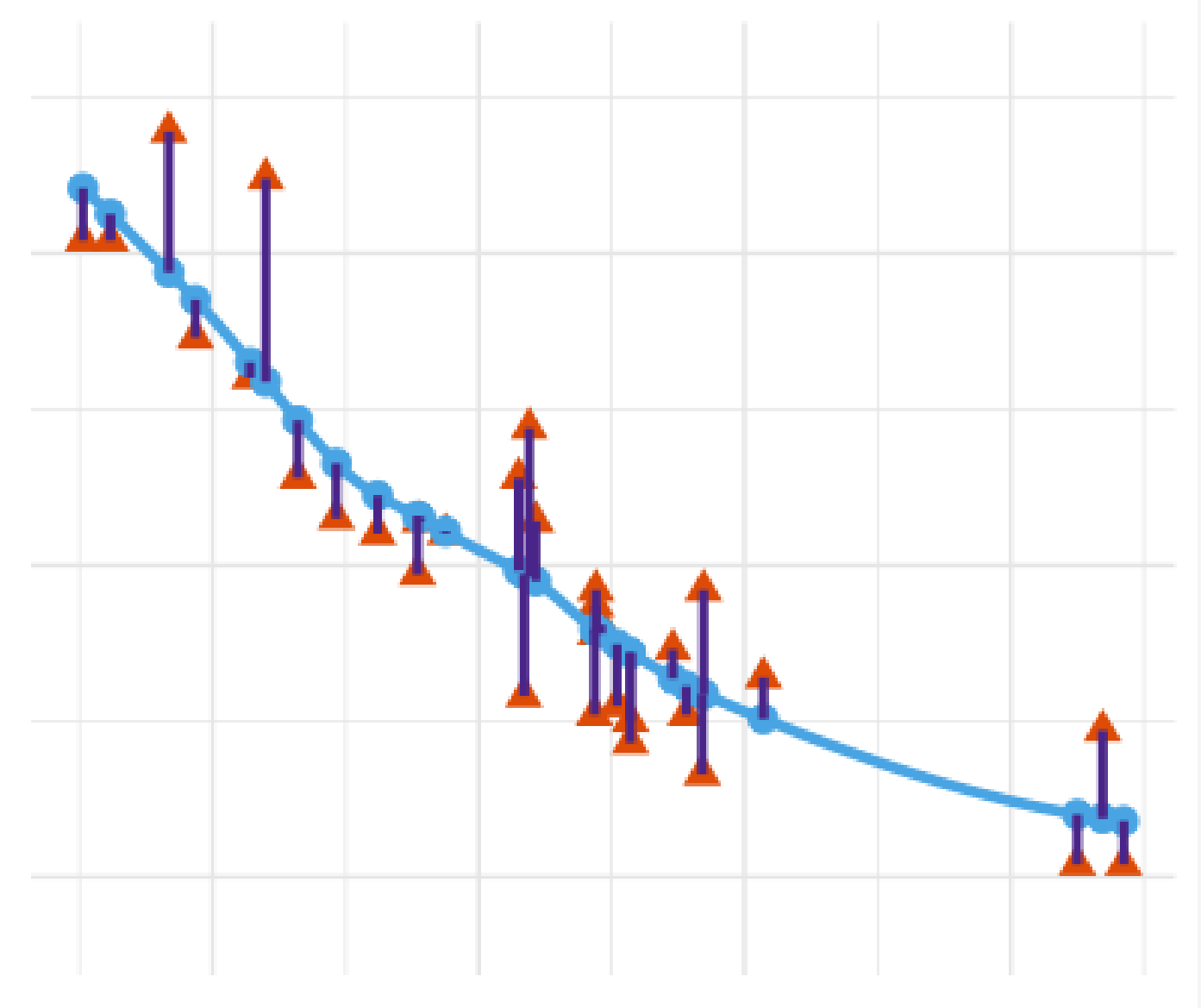
$$\widehat{Y} = \widehat{f}(X)$$

This assumes there exists an $f()$ such that $Y = f(X) + \epsilon$ … the truth!

Assess the residuals/errors!

$$Y - \widehat{Y} = f(X) - \widehat{f}(X) + \epsilon$$

Residuals



Mean Squared Error

$$E[(Y - \widehat{Y})^2] = E[(f(X) - \widehat{f}(X))^2] + var(\epsilon)$$

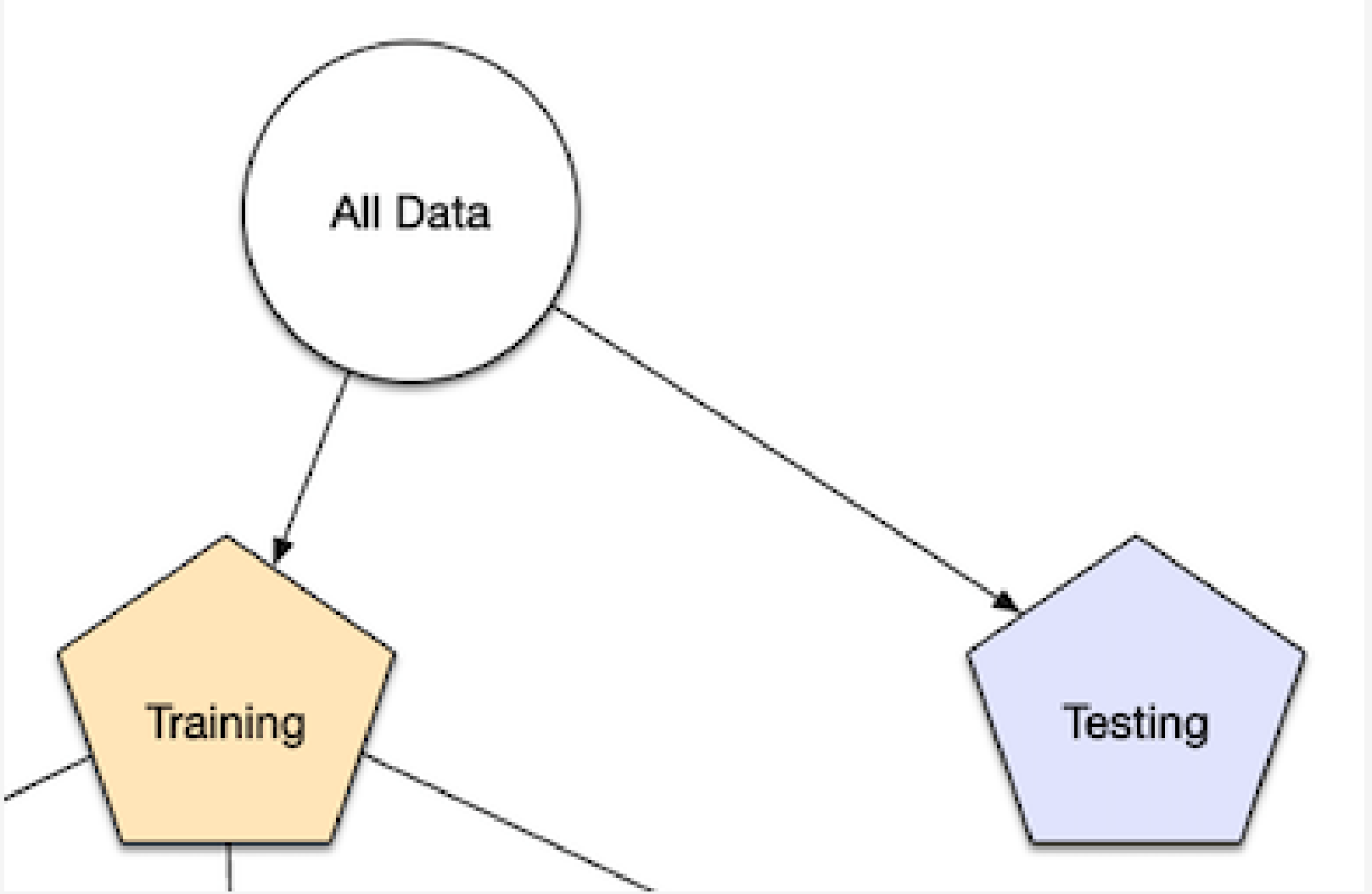
- Reducible error vs irreducible error
- Overfitting
- Training vs Testing (Data spending/ allocation)

Classification

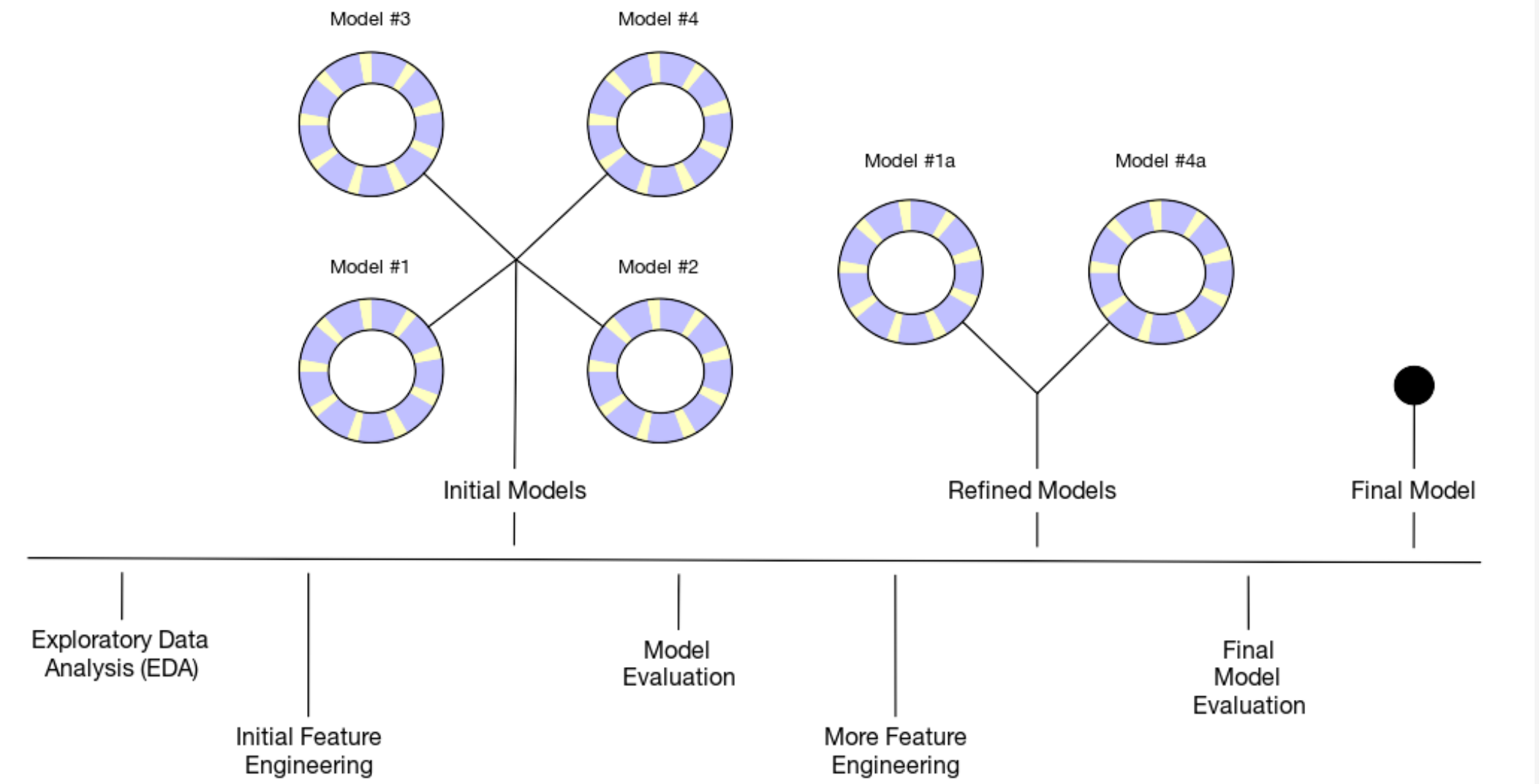
n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Initial data split

Avoiding overfitting & the goal is out of sample performance estimate



Supervised learning schema



A schematic for the typical modeling process from Tidymodeling with R

Splitting data for training

Pick an evaluation metric for model comparisons!

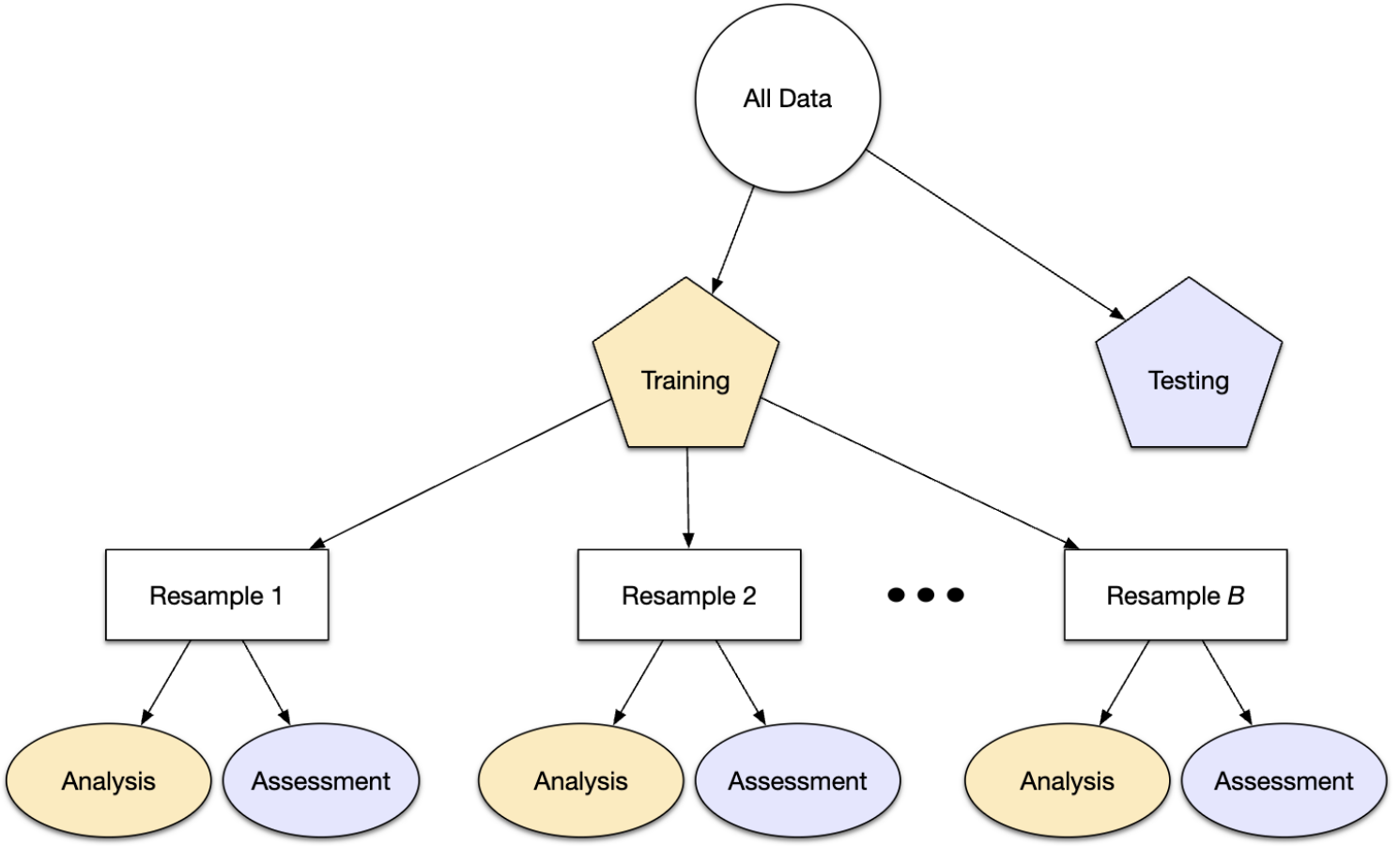


Figure 10.1: Data splitting scheme from the initial data split to resampling

Which metric matters!

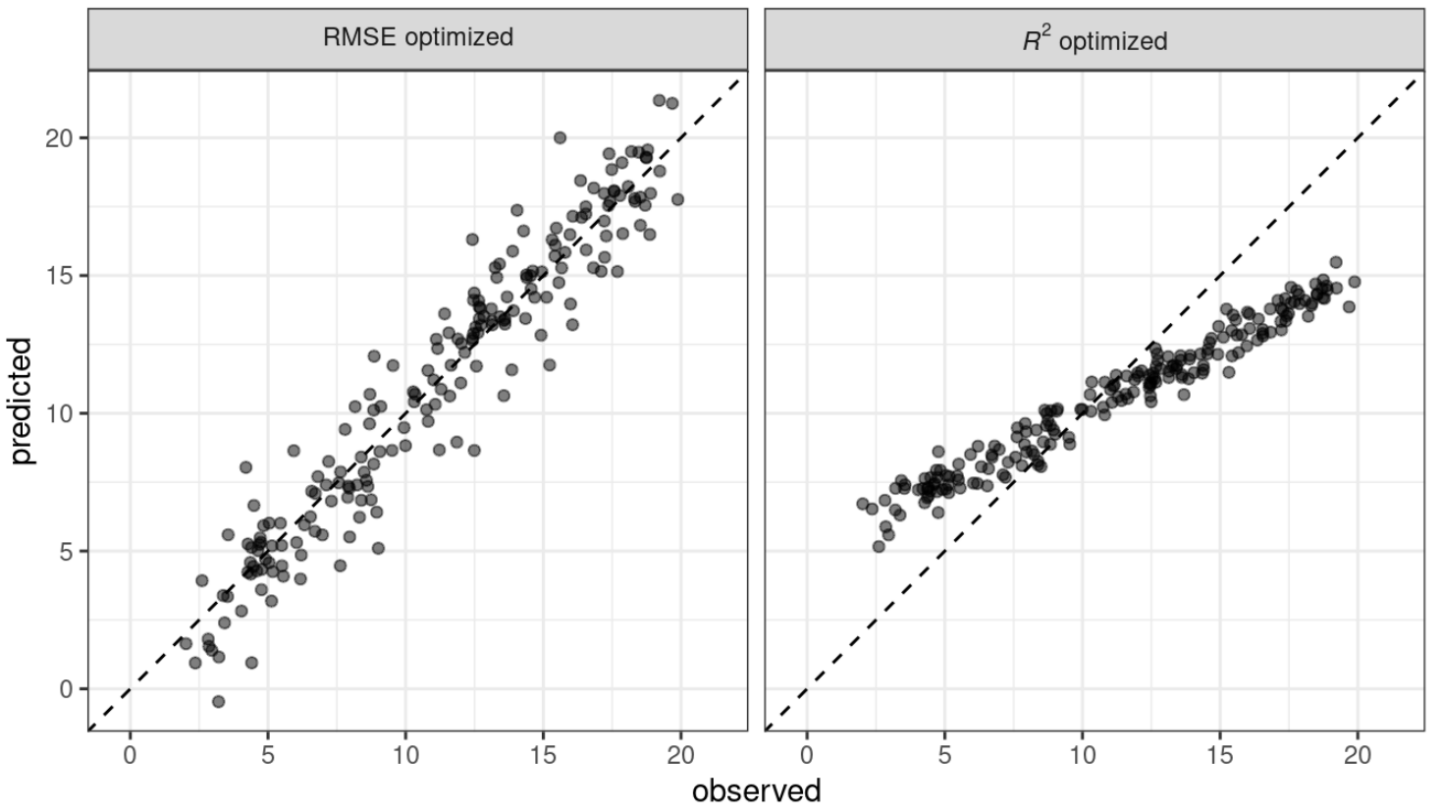


Figure 9.1: Observed versus predicted values for models that are optimized using the RMSE compared to the coefficient of determination

Common Metrics

Regression

- Mean squared error (MSE)
- Root Mean Squared error (RMSE)
- R-squred (RSQ or R^2)
- Mean absolute error/difference (MAE or MAD)
- many more ...

Classification

- Accuracy
- Area under the curve (AUC)
- Recall
- Sensitivity
- many more …

Selecting the “best” model

- Don’t always have to go with best performing model
- 1 standard error rule: Are the competing models really preforming differently?
- Occam’s Razor or KISS

Evaluating the final model

- Using the test data
- Not restricted to metric used for comparison!
- Might use metrics that are easier to explain
- Be careful of causal interpretations

Things to keep in mind

- Measurement error
- Data quality
- Missing data
- Baseline model or null model
- Selecting which models to fit