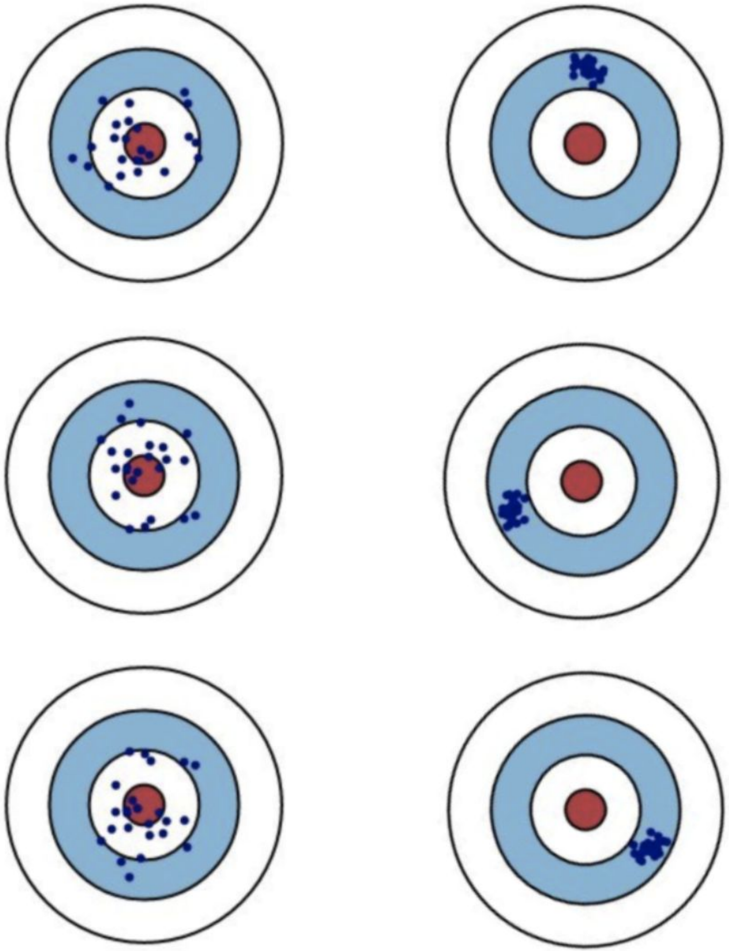
Chapter 15: Model Quality and the Bias-Variance Tradeoff

Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

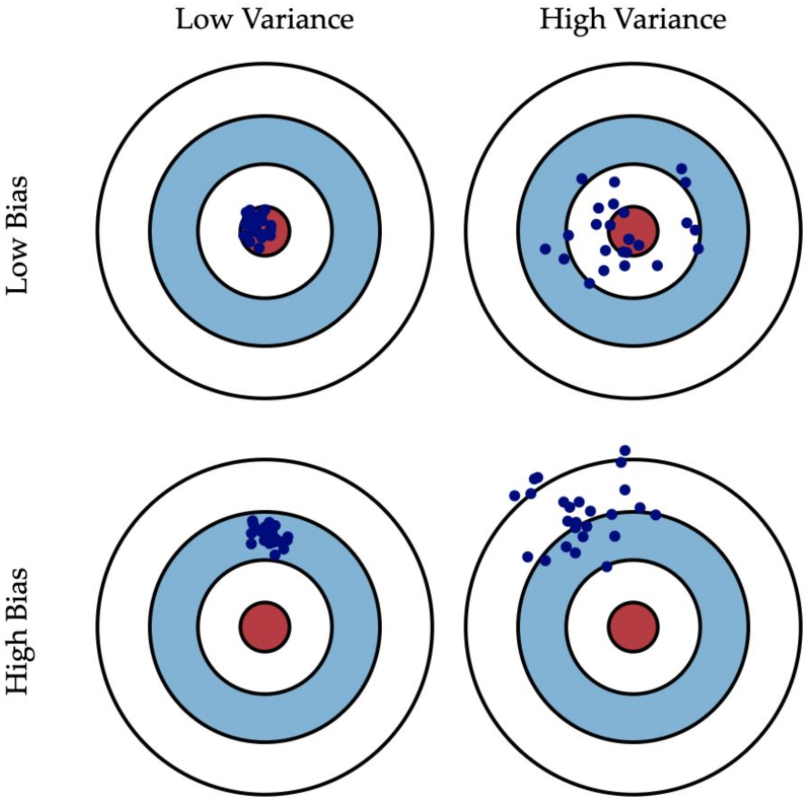
How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 30 min, max) while following along in the notes



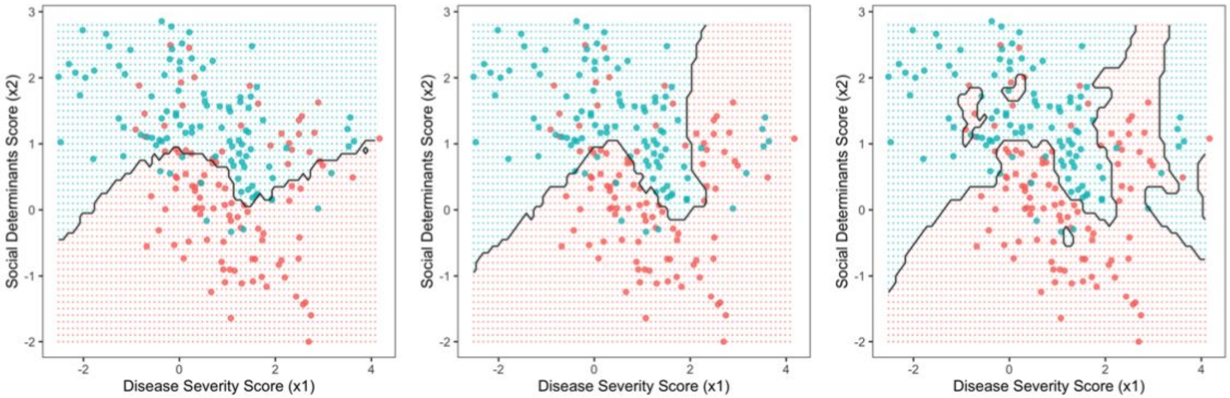
Each dot is one tree trained on a slightly different dataset, making a prediction on a single test example.

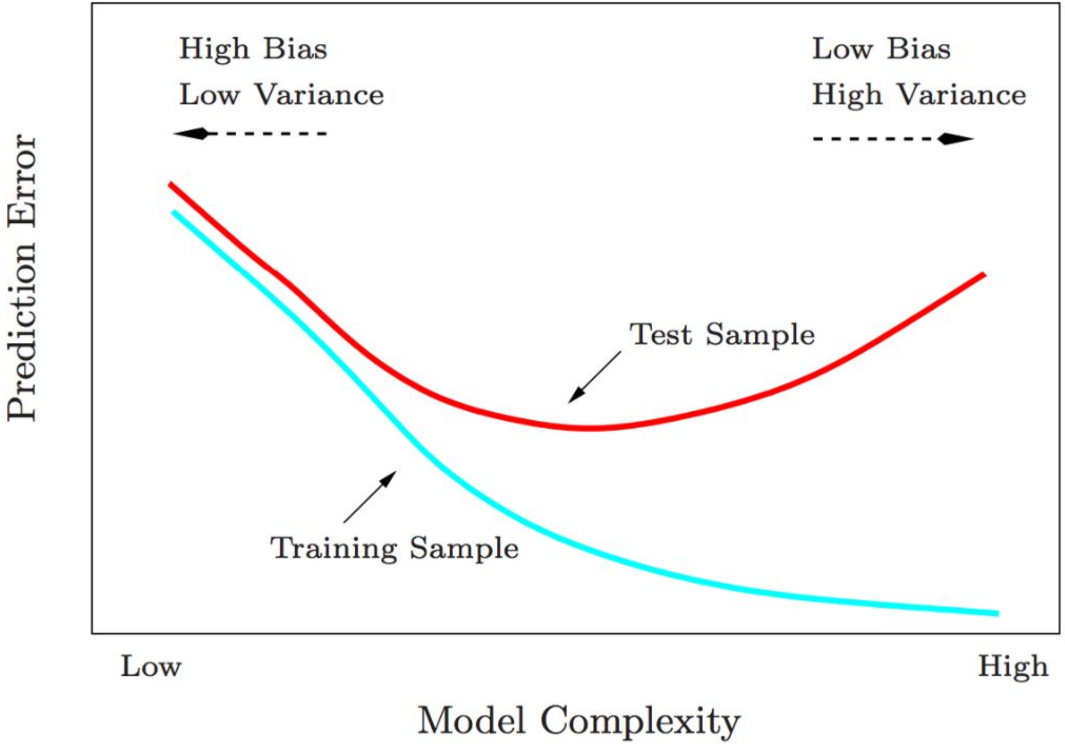
Which column represents boosting and which represents a random forest?


Low Variance
High Variance
Low Bias
High Bias

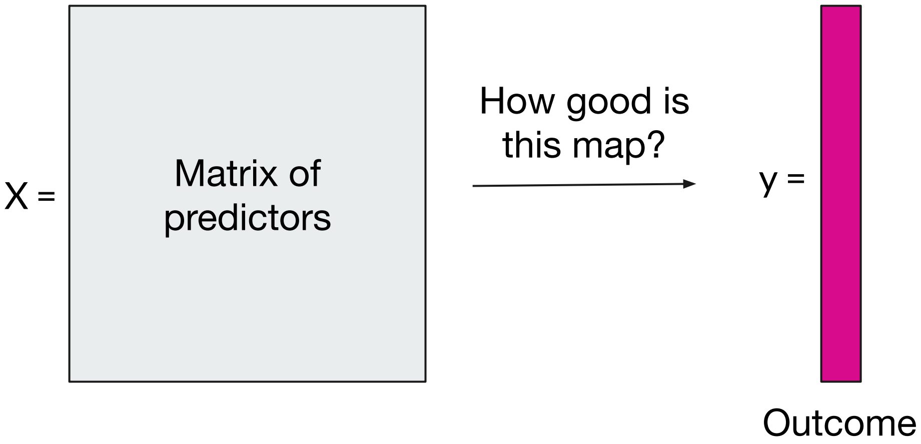
Question 15.3

Here we see three decision boundaries for KNN with different values of K (the number of neighbors considered in making a prediction). The data are for the two-class classification problem first discussed in Chapter 2. From left to right, $K = 50, 15$, and 3. What are the tradeoffs in moving from left to right in terms of (a) training error/goodness of fit and (b) test error/generalizability?



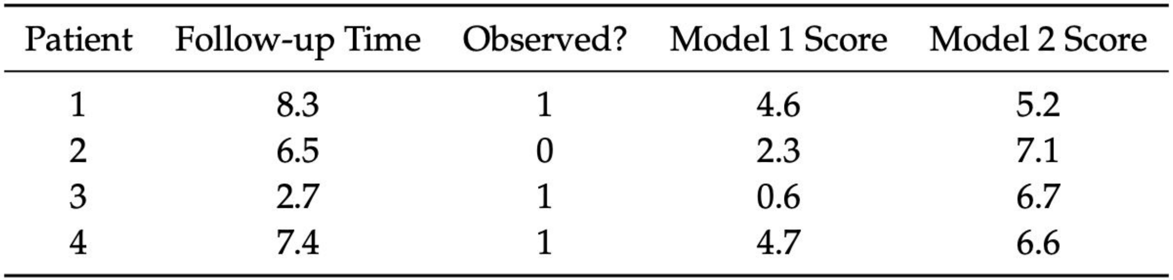

High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

Loss function:
A measurement of model error over an entire dataset (training or test).

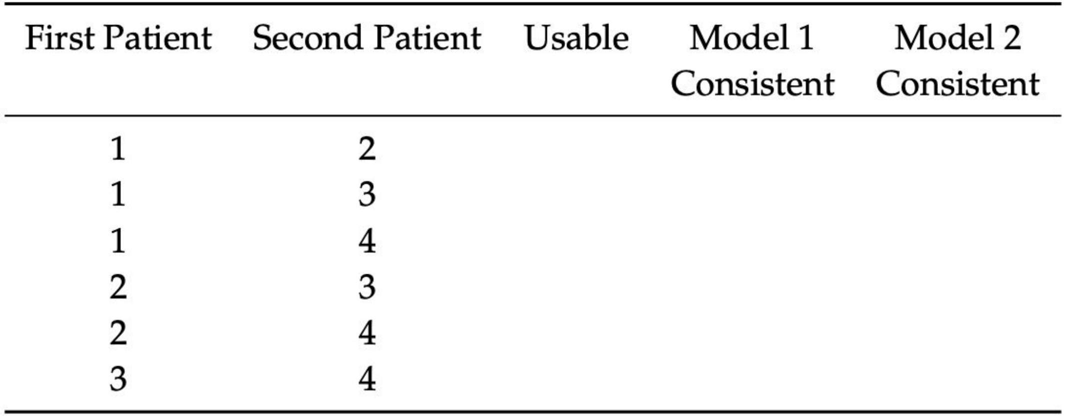


Error in survival analysis:

Harrell's Concordance Index



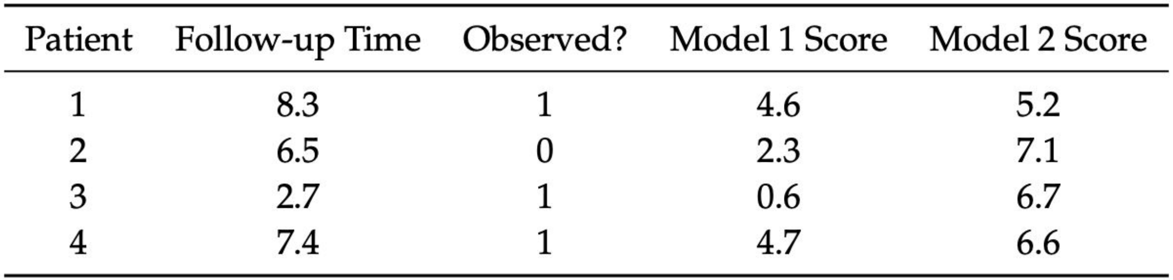
Patient	Follow-up Time	Observed?	Model 1 Score	Model 2 Score
1	8.3	1	4.6	5.2
2	6.5	0	2.3	7.1
3	2.7	1	0.6	6.7
4	7.4	1	4.7	6.6



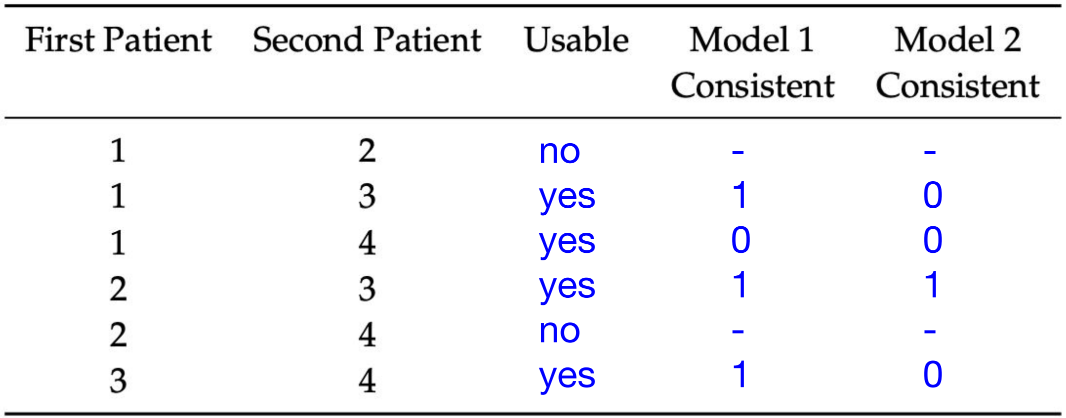
First Patient	Second Patient	Usable	Model 1 Consistent	Model 2 Consistent
1	2			
1	3			
1	4			
2	3			
2	4			
3	4			

Error in survival analysis:

Harrell's Concordance Index



Patient	Follow-up Time	Observed?	Model 1 Score	Model 2 Score
1	8.3	1	4.6	5.2
2	6.5	0	2.3	7.1
3	2.7	1	0.6	6.7
4	7.4	1	4.7	6.6



First Patient	Second Patient	Usable	Model 1 Consistent	Model 2 Consistent
1	2	no	-	-
1	3	yes	1	0
1	4	yes	0	0
2	3	yes	1	1
2	4	no	-	-
3	4	yes	1	0