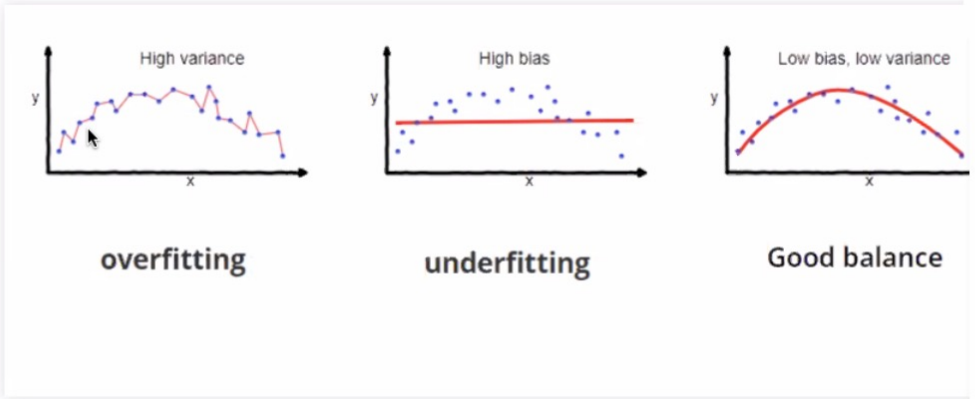
Supervisions/revision

A fundamental concept in machine learning

Bias-variance tradeoff



- Bias is residual error from fitting the Training data
- Variance is generalization error when applying the model fit to

A fundamental concept in machine learning

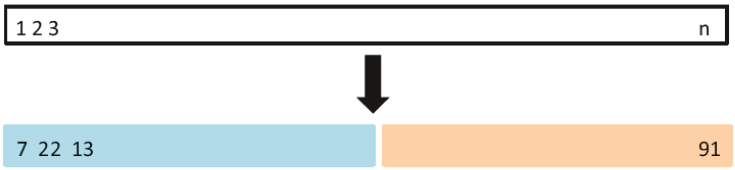
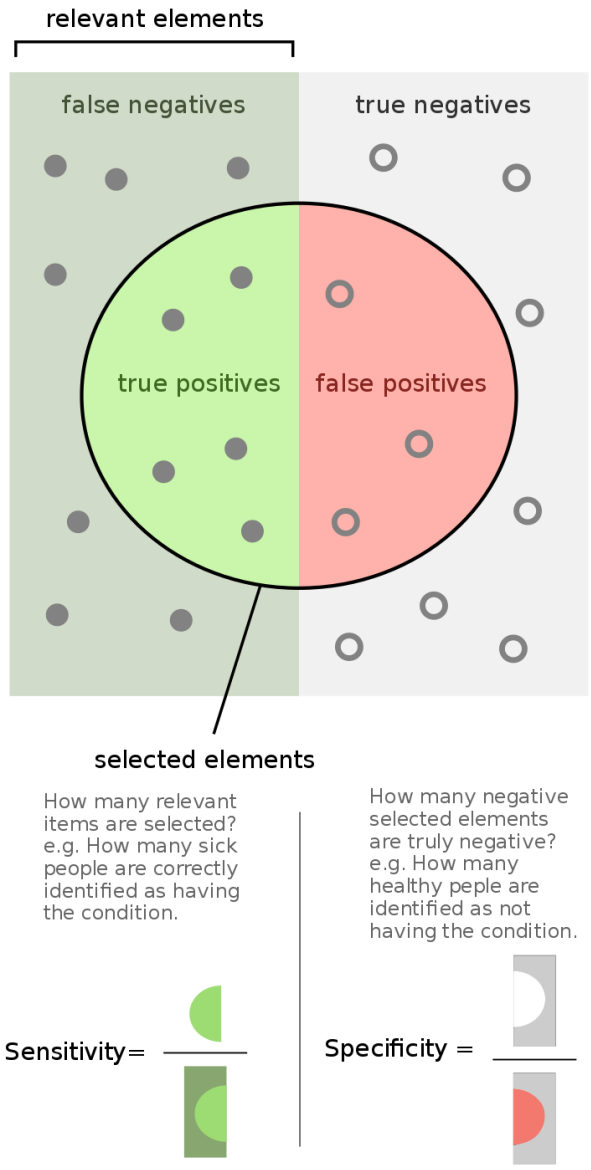
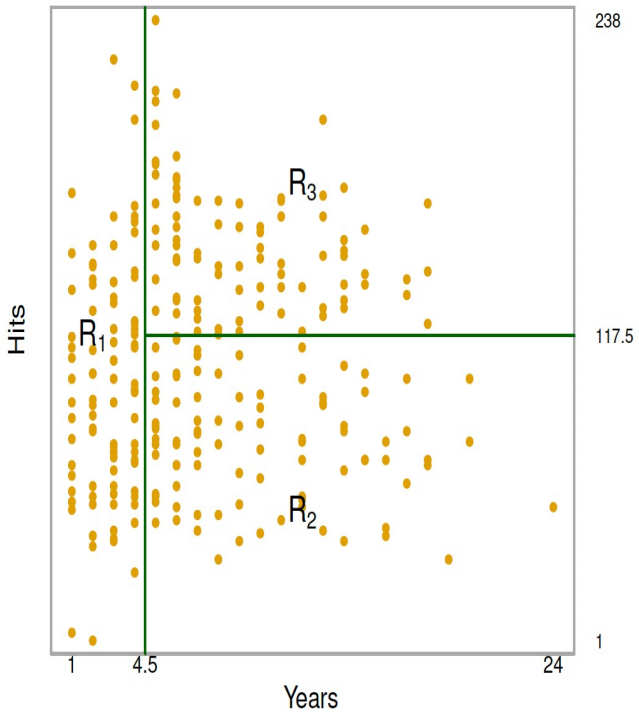


FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many relevant items are selected? e.g. How many sick people are correctly identified as having the condition.
How many negative selected elements are truly negative? e.g. How many healthy peple are identified as not having the condition.
Sensitivity=
Specificity =



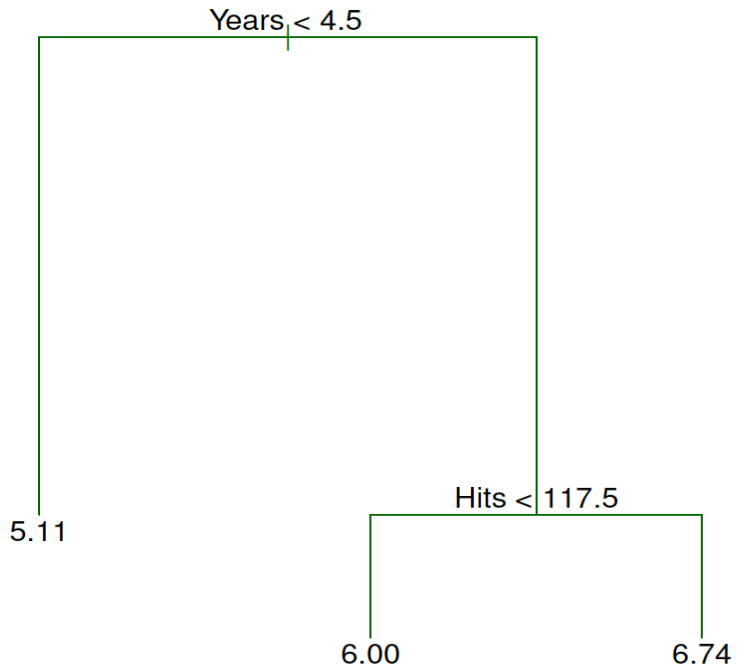
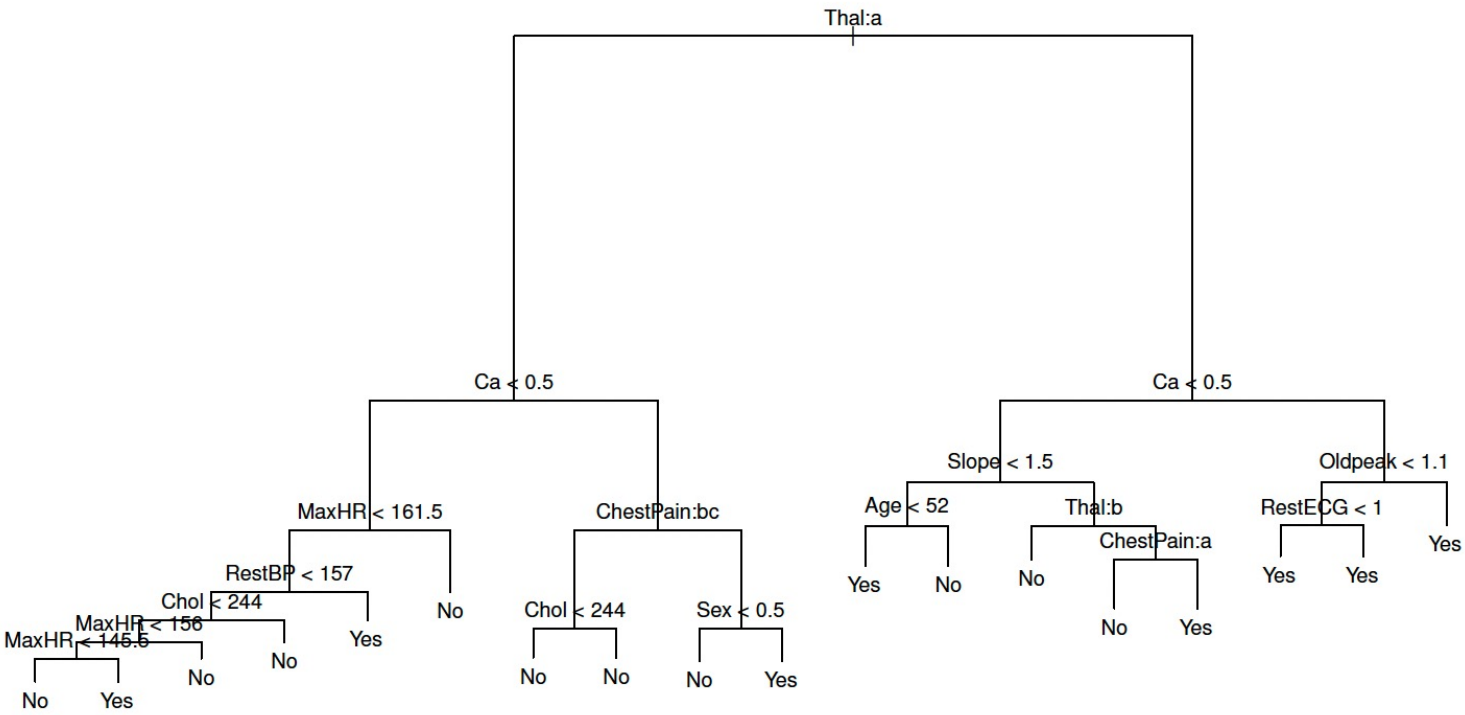
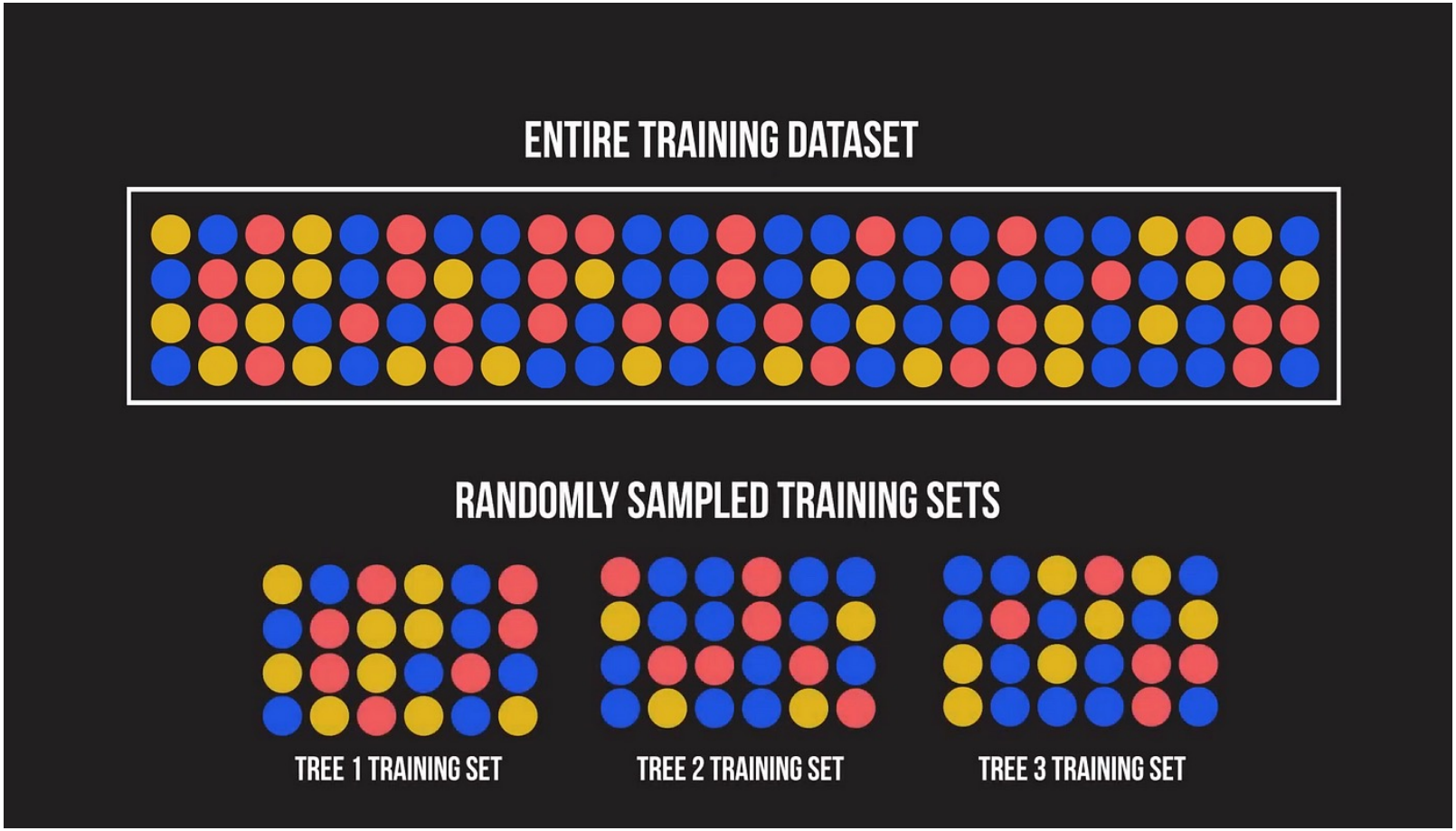


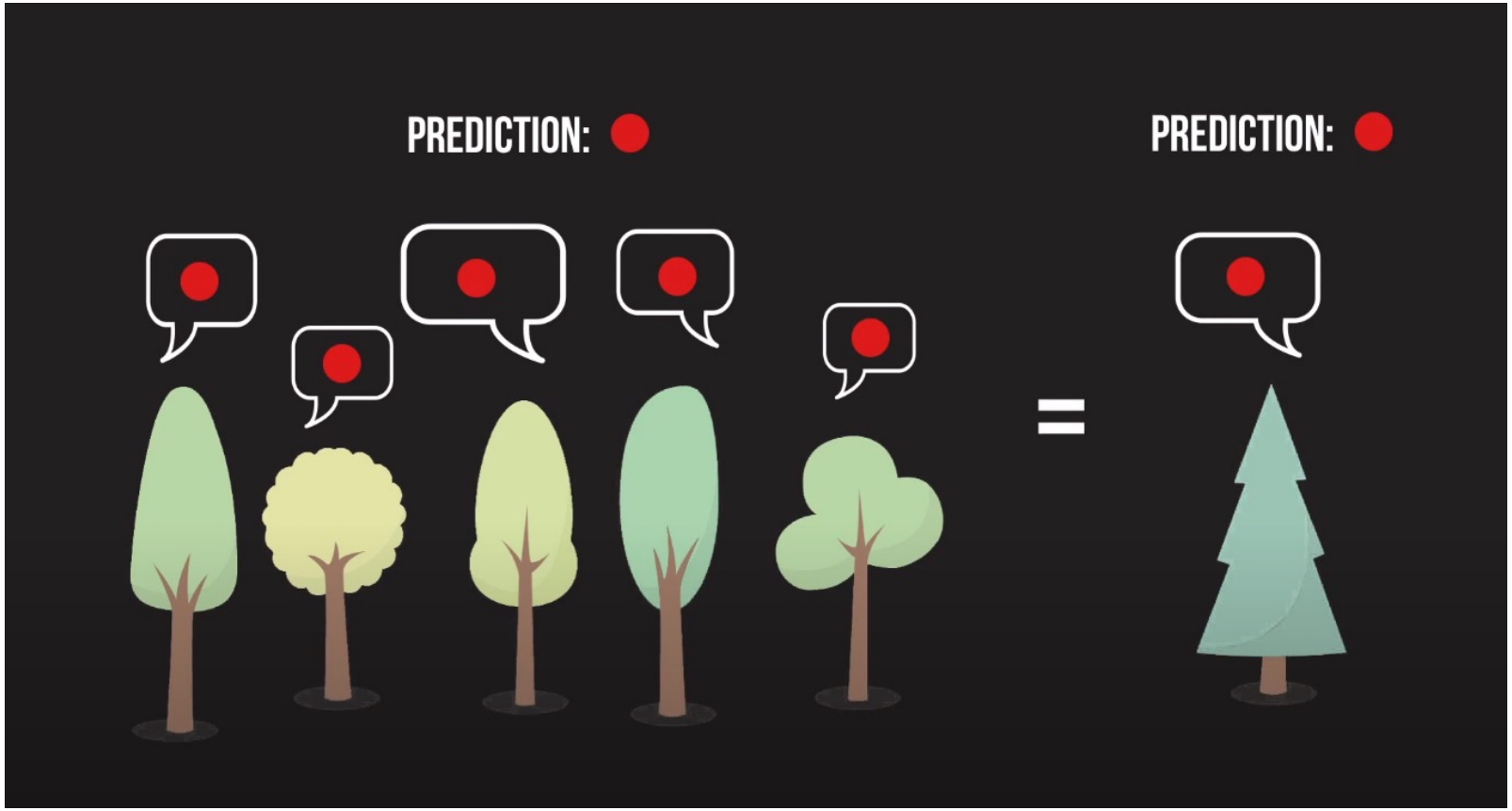
FIGURE 8.2. *The three-region partition for the* `Hitters` *data set from the regression tree illustrated in Figure 8.1.*


Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
Slope < 1.5
Oldpeak < 1.1
Age < 52
Thal:b
RestECG < 1
ChestPain:a
RestBP < 157
Chol < 244
MaxHR < 156
MaxHR < 145.5
Chol < 244
Sex < 0.5
No
Yes
No
No
Yes
No
No
No
No
Yes
Yes
No
No
Yes
Yes
Yes
Yes
No
Yes

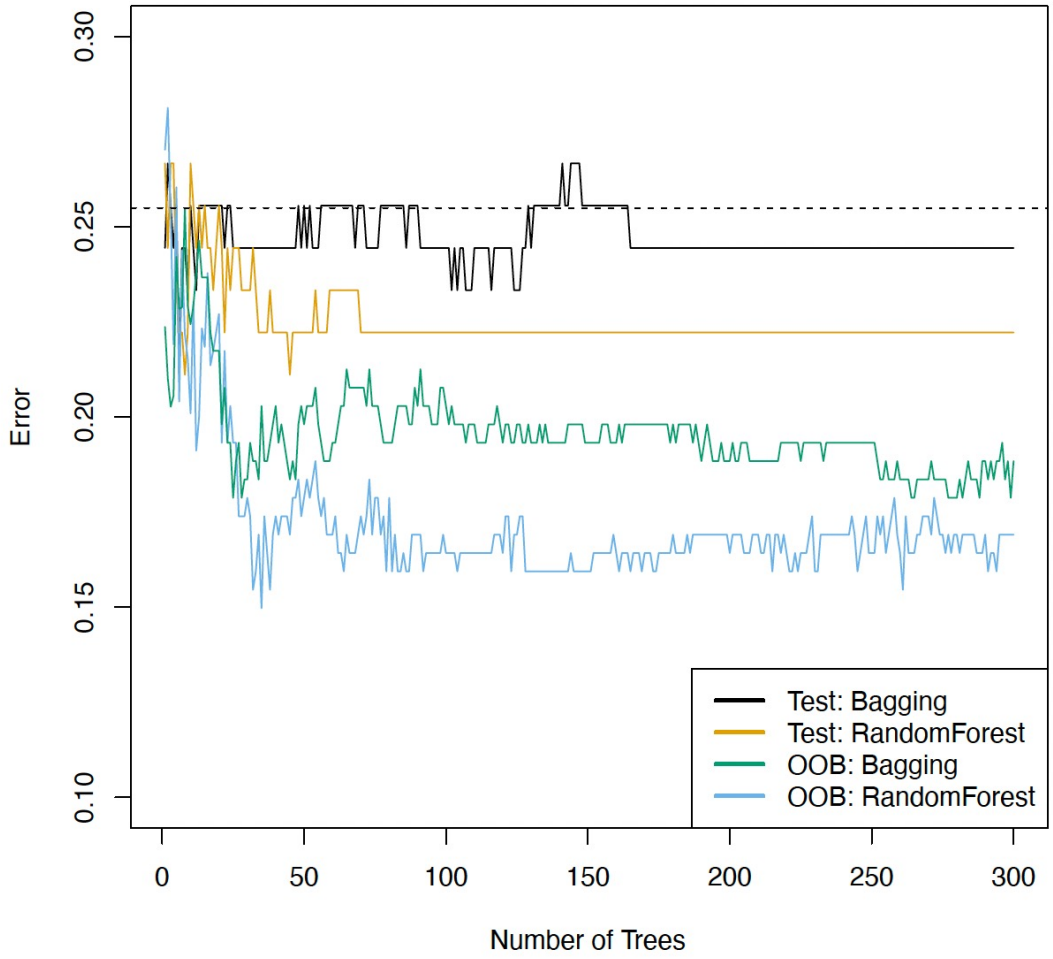


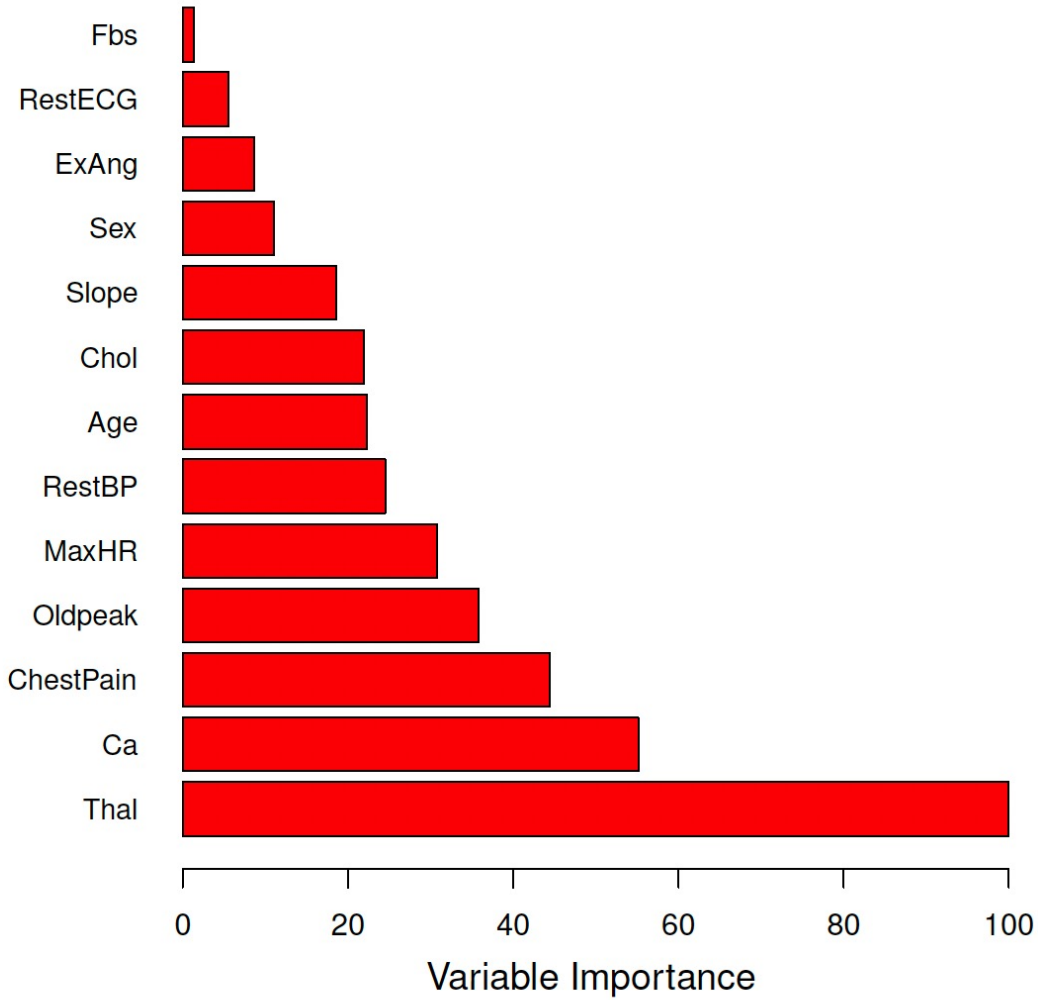
https://towardsdatascience.com/a-visual-guide-to-random-forests-b3965f453135

Random Forests

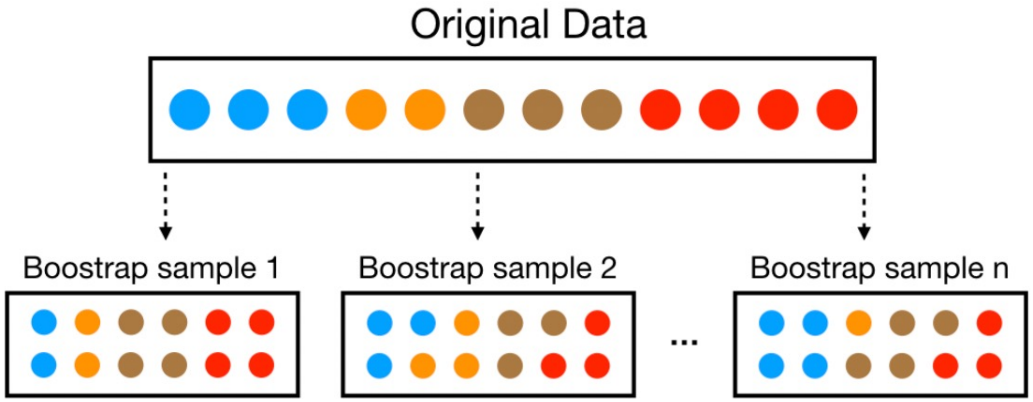


https://towardsdatascience.com/a-visual-guide-to-random-forests-b3965f453135


Error
0.10
0.15
0.20
0.25
0.30
0
50
100
150
200
250
300
Number of Trees
Test: Bagging
Test: RandomForest
OOB: Bagging
OOB: RandomForest

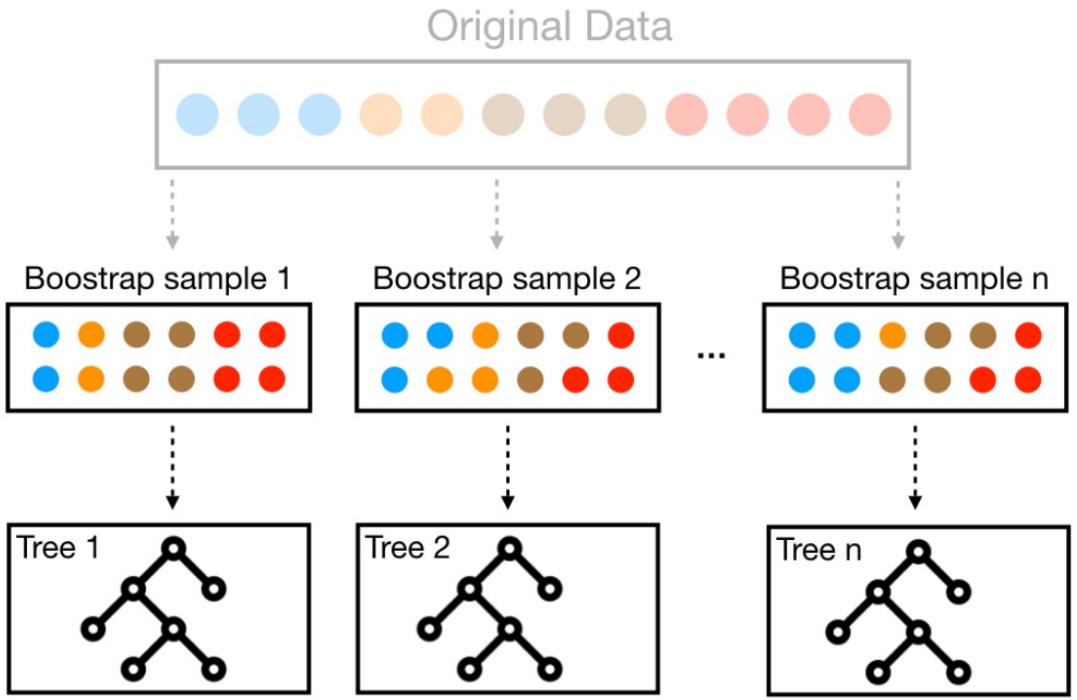

Fbs
RestECG
ExAng
Sex
Slope
Chol
Age
RestBP
MaxHR
Oldpeak
ChestPain
Ca
Thal
0
20
40
60
80
100
Variable Importance

Bagging



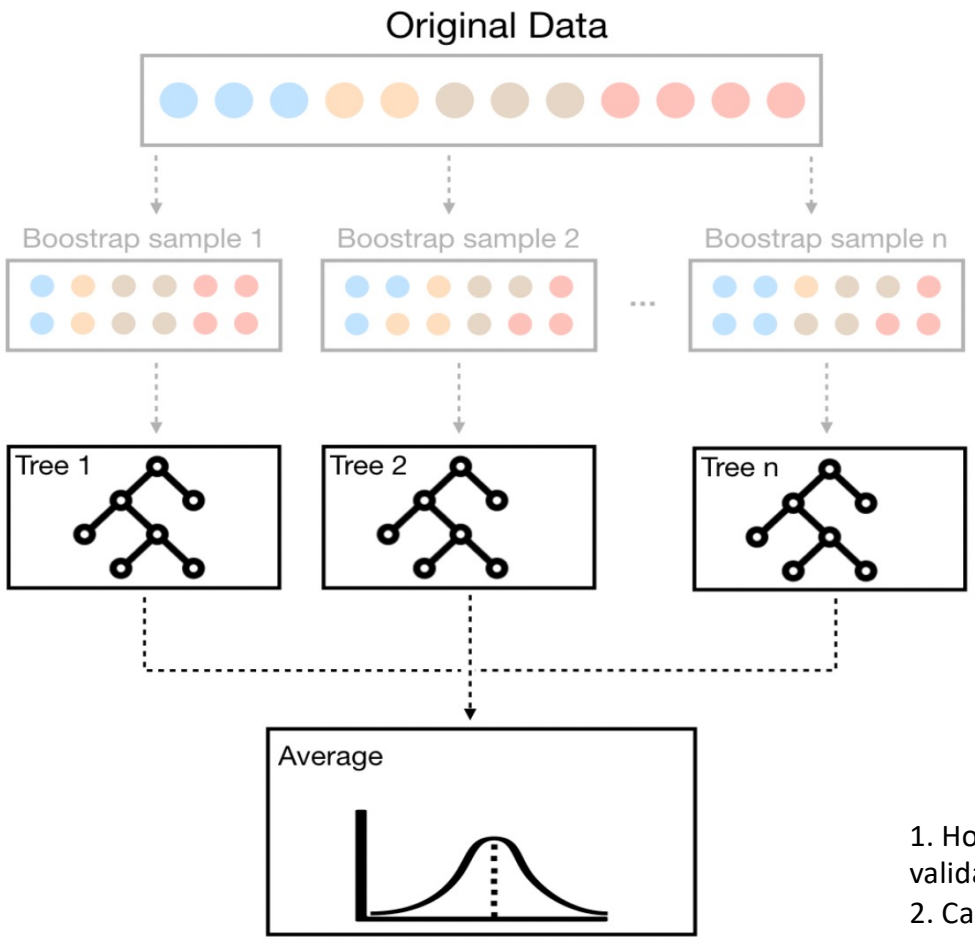
https://bradleyboehmke.github.io/random-forest-training/slides-source.html#33

Bagging



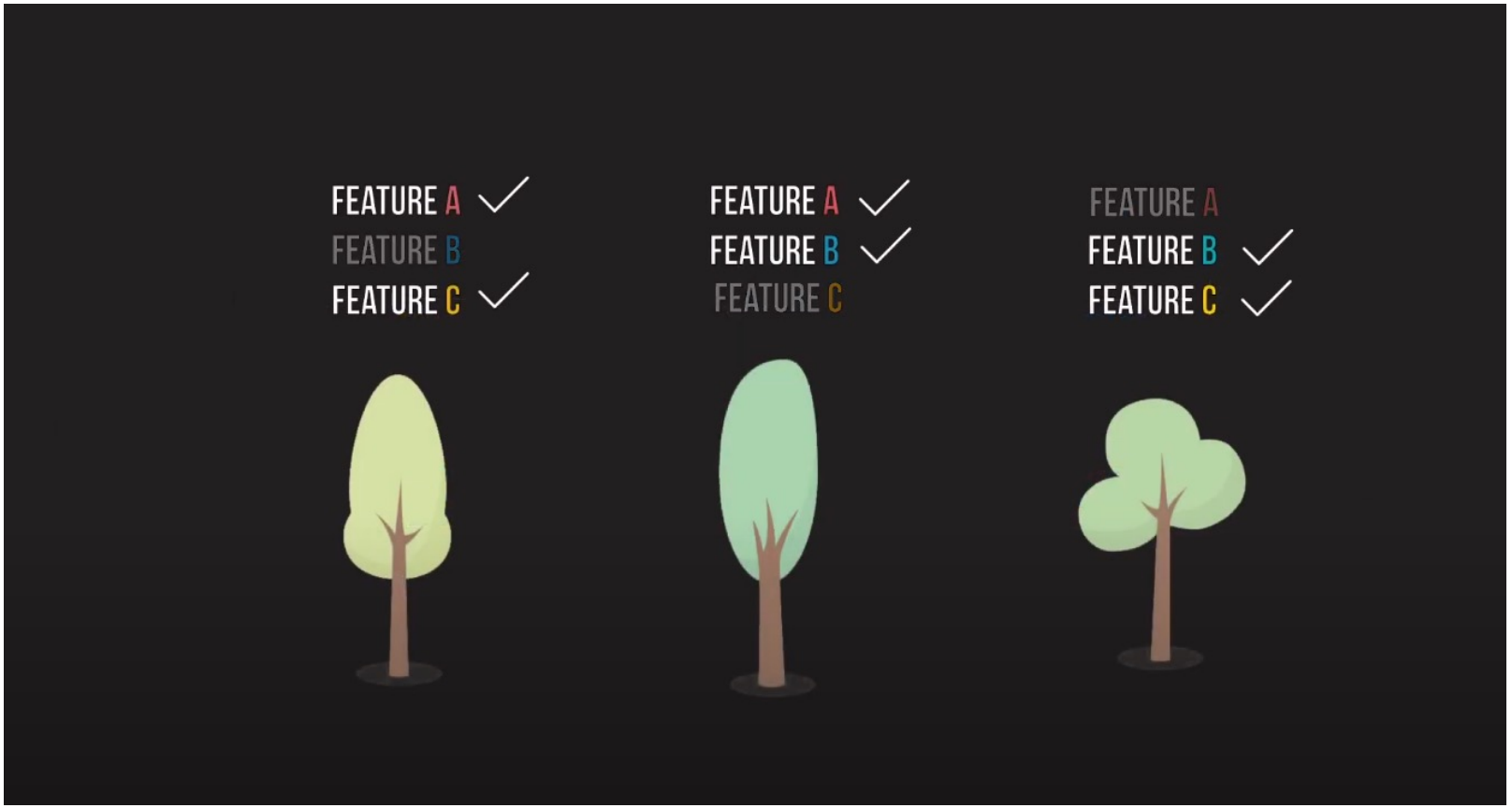
https://bradleyboehmke.github.io/random-forest-training/slides-source.html#33

Bagging



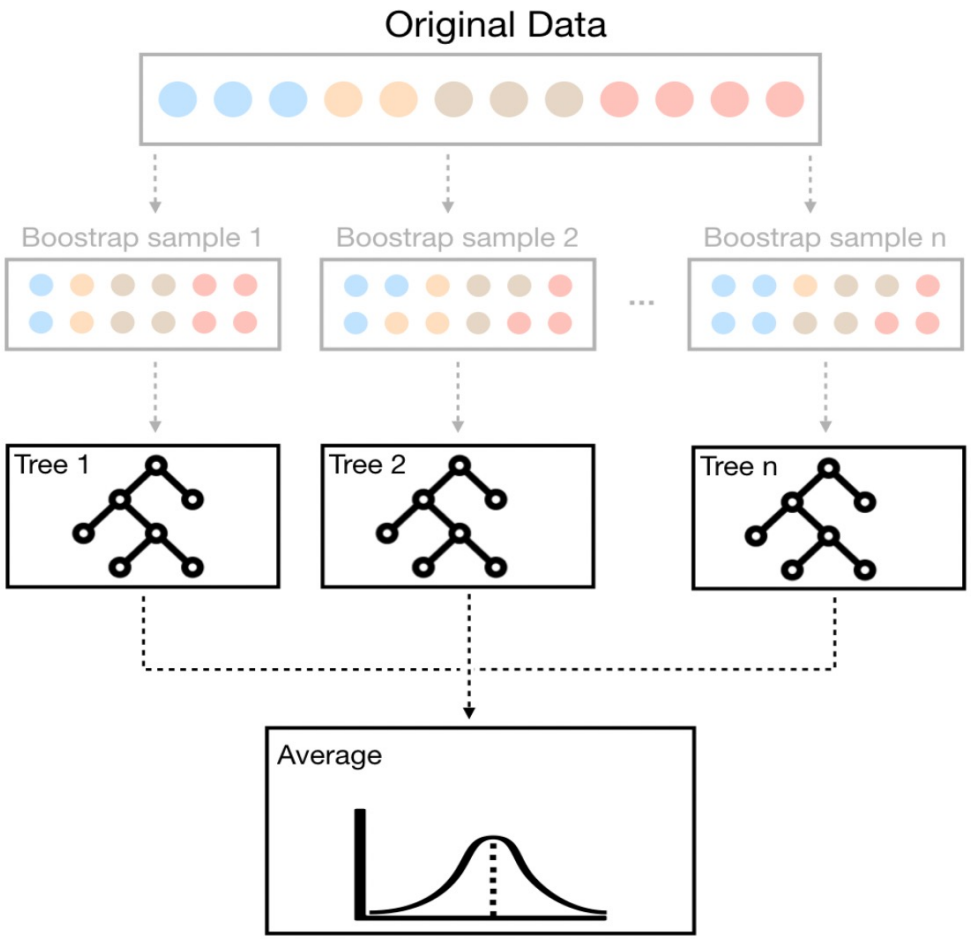
1. How would you cross-validate?
2. Can this lead to any problems?

Solution



https://towardsdatascience.com/a-visual-guide-to-random-forests-b3965f453135

Random forests + Bagging



1. Decorrelate trees
2. When deciding on a split, a random sample of features is taken
3. Each tree is built on a different random sample of features!

Question

What happens if we use a simple classifier that just predicts randomly on a disease?

For example, say you are trying to predict a particular disease.

1. Assume you have features like age, gender and medications.
2. You have to predict whether this patient will get a disease or not.
3. Assume the disease occurs in 90% of the population.

4. How should you build your algorithm? What metric should you look at?

Question

What happens if we use a simple classifier that just predicts randomly on a disease?

For example, say you are trying to predict a particular disease.

1. Assume you have features like age, gender and medications.
2. You have to predict whether this patient will get a disease or not.
3. Assume the disease occurs in 90% of the population.

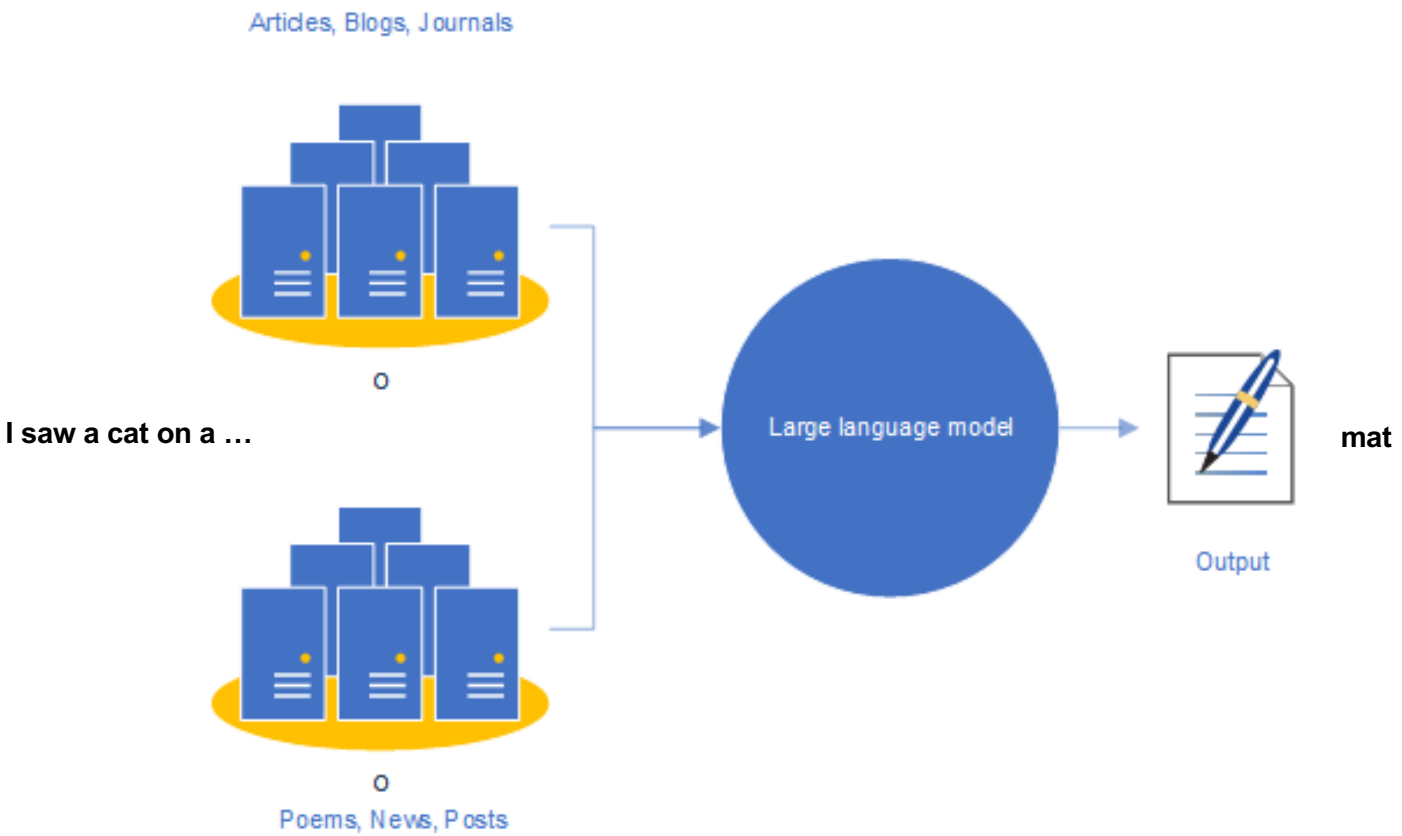
4. How should you build your algorithm? What metric should you look at?

5. **Answer**: You should look at class-specific error (sensitivity/specificity).
   a. This is unbalanced data.
   b. Even an algorithm that randomly outputs 1 [the person has disease], (regardless of the input features), would give 90% accuracy on the data.

Question

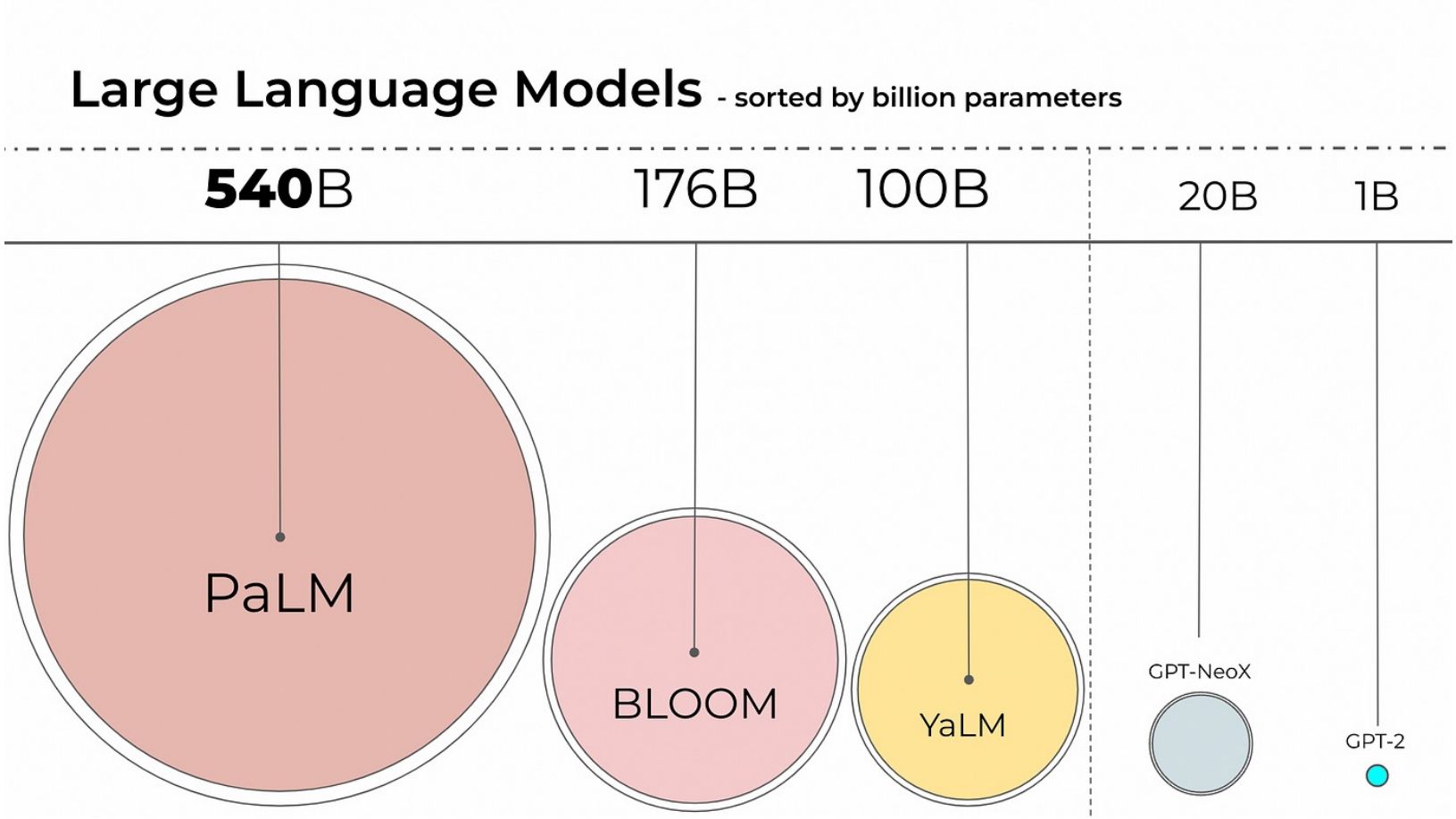
1. What is the sensitivity for the cases we talked about?
2. What is the specificity?
3. Fixes to balance the data (downsample, etc.)

Applications of this concept



https://becominghuman.ai/a-quick-introduction-to-the-large-language-model-chatgpt-a2f5f54b4d5e

Applications of this concept



https://ashukumar27.medium.com/decoding-large-language-models-quantization-ff58964c0f31