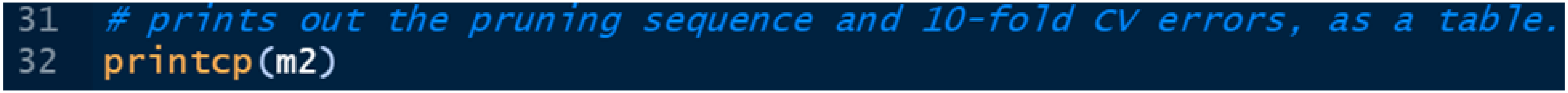
10 fold Cross Validation and 1 SE Rule

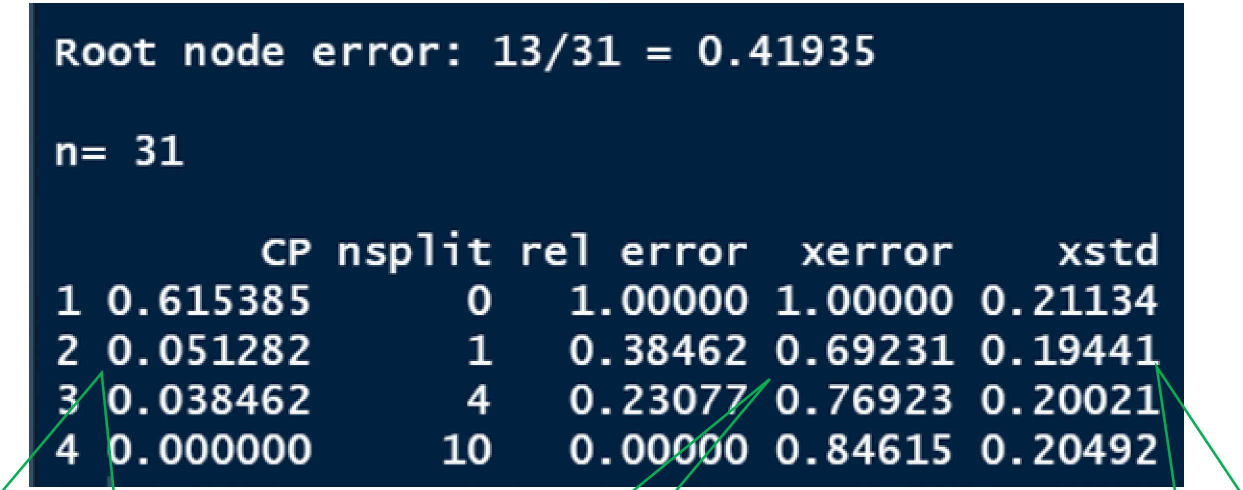
CART

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 8.

CP Table shows min CV error and its 1SE.



```
# prints out the pruning sequence and 10-fold CV errors, as a table.
printcp(m2)
```



```
Root node error: 13/31 = 0.41935

n= 31

        CP nsplit rel error  xerror    xstd
1 0.615385      0   1.00000 1.00000 0.21134
2 0.051282      1   0.38462 0.69231 0.19441
3 0.038462      4   0.23077 0.76923 0.20021
4 0.000000     10   0.00000 0.84615 0.20492
```

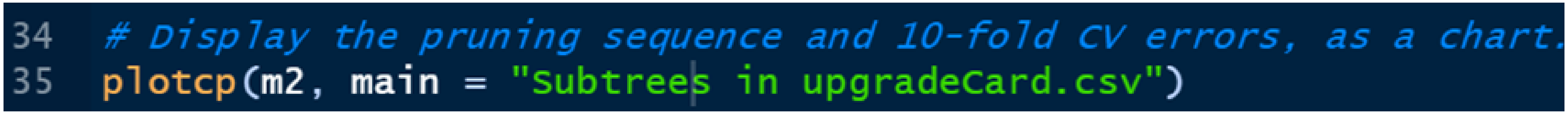
Geometric mean CP of the 2nd Tree = sqrt(0.0512 * 0.615) ≈ 0.18

Min CV Error = 0.69231

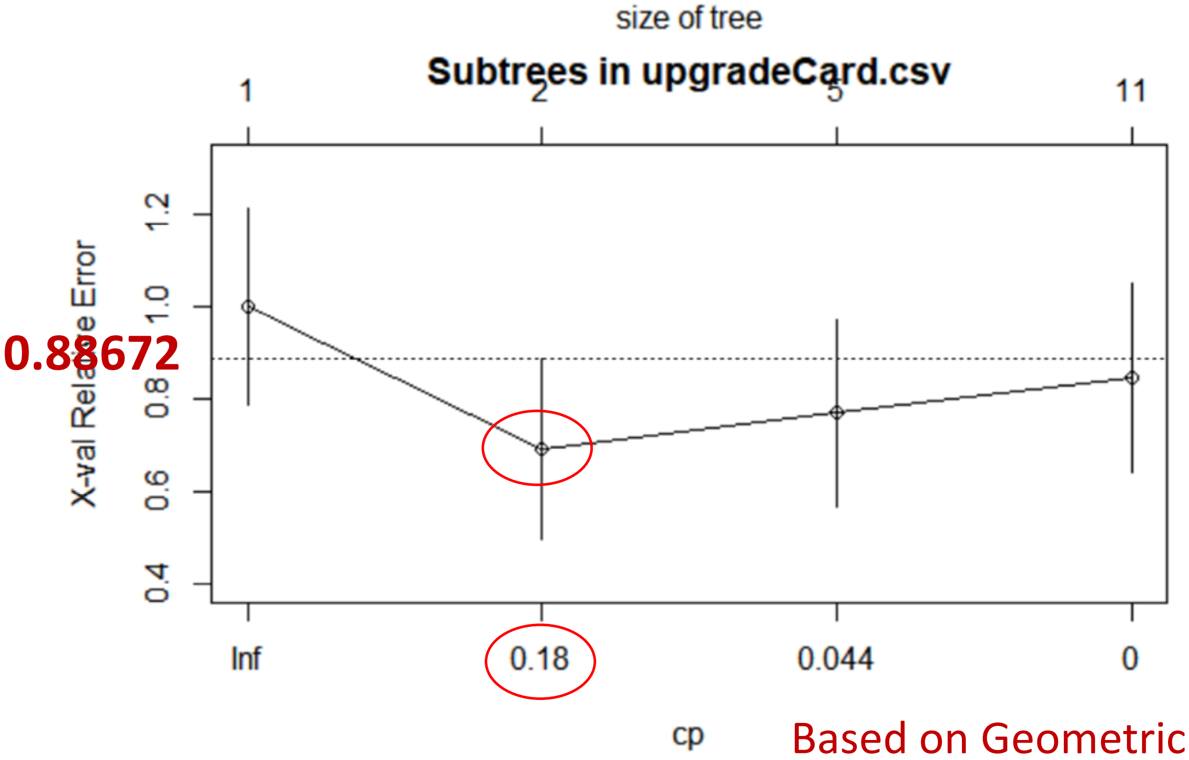
1SE = 0.19441

CV Error Cap = 0.69231 + 0.19441 = **0.88672**

CV error cap is displayed in plotcp()



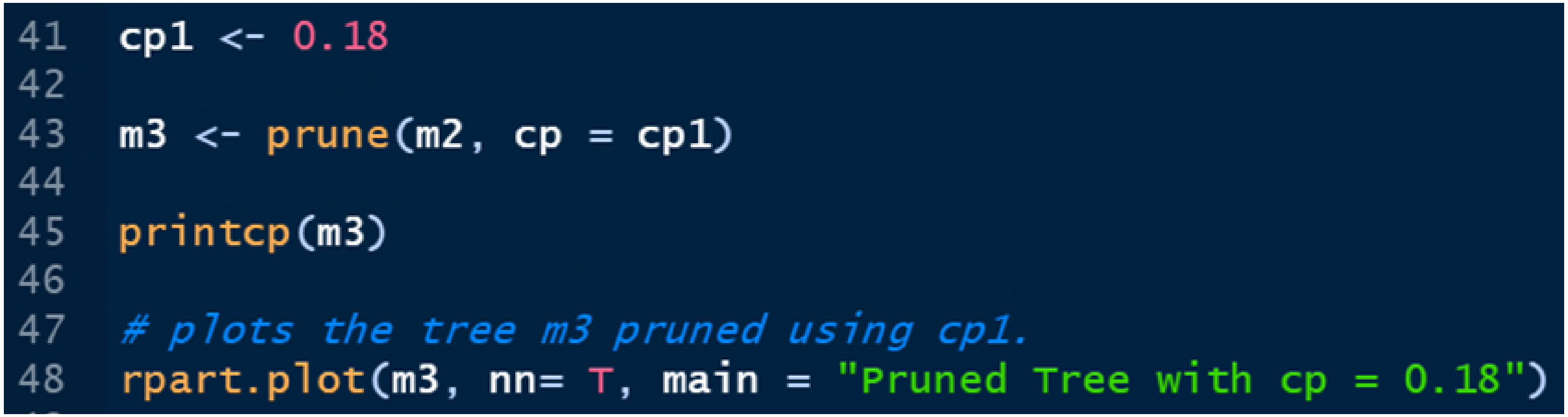
```
# Display the pruning sequence and 10-fold CV errors, as a chart.
plotcp(m2, main = "Subtrees in upgradeCard.csv")
```



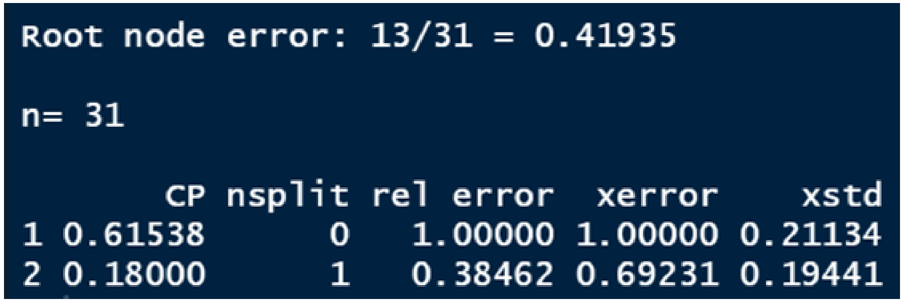
0.88672

Based on Geometric mean

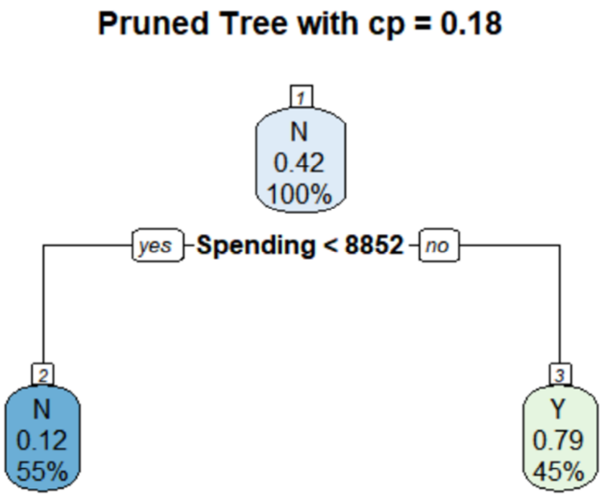
Get a specific subtree via prune() by pruning the maximal tree m2 with a specific value of cp



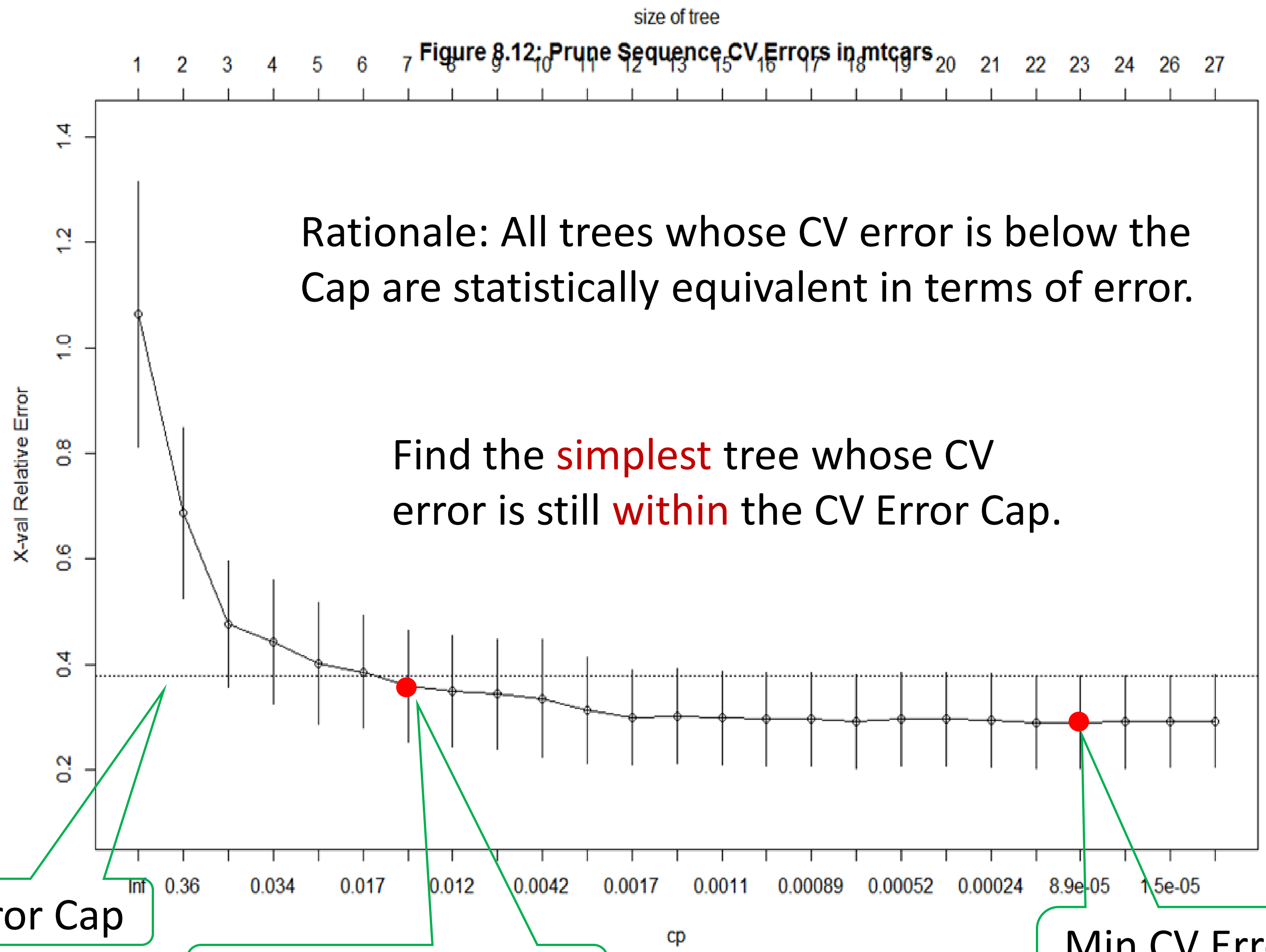
```r
cp1 <- 0.18

m3 <- prune(m2, cp = cp1)

printcp(m3)

# plots the tree m3 pruned using cp1.
rpart.plot(m3, nn= T, main = "Pruned Tree with cp = 0.18")
```



```
Root node error: 13/31 = 0.41935

n= 31

        CP nsplit rel error  xerror    xstd
1 0.61538      0   1.00000 1.00000 0.21134
2 0.18000      1   0.38462 0.69231 0.19441
```



CART on mtcars dataset shows 25 trees



Rationale: All trees whose CV error is below the Cap are statistically equivalent in terms of error.

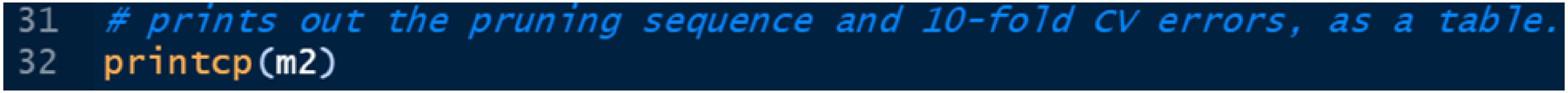
Find the simplest tree whose CV error is still within the CV Error Cap.

CV Error Cap

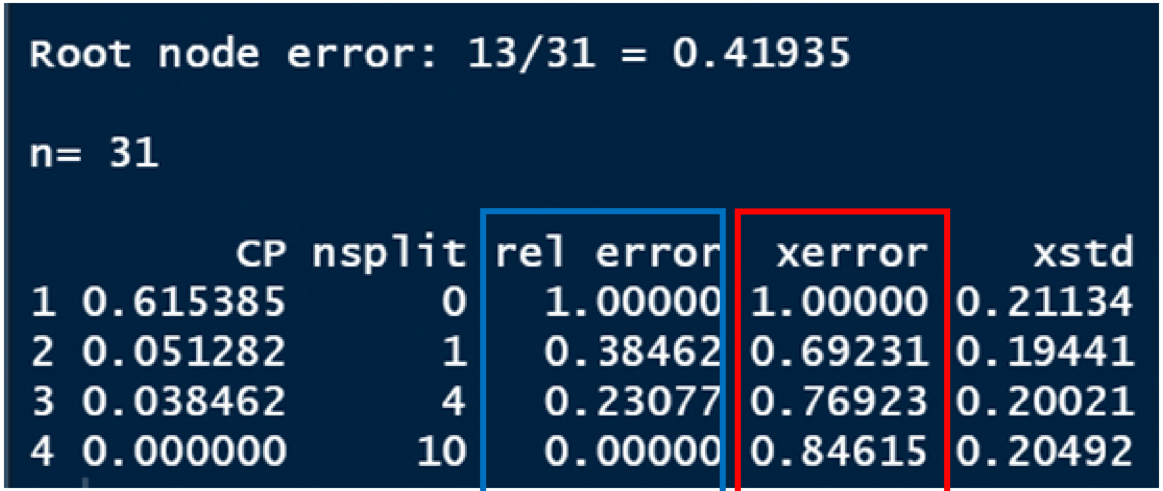
Optimal Tree #7

Min CV Error Tree #22

Trainset error vs 10 fold Cross Validation (CV) error



```
# prints out the pruning sequence and 10-fold CV errors, as a table.
printcp(m2)
```



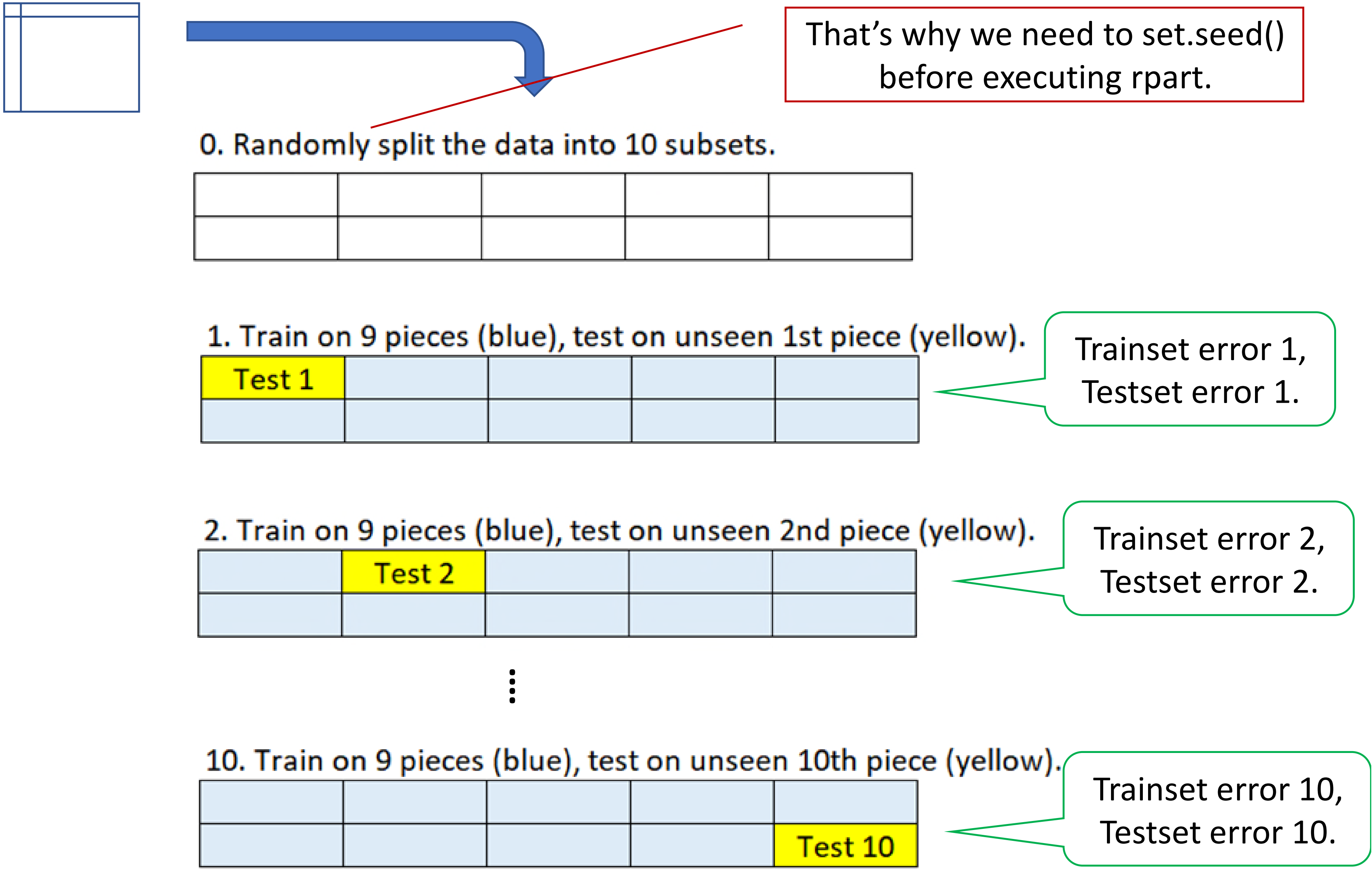
```
Root node error: 13/31 = 0.41935

n= 31

        CP nsplit rel error  xerror    xstd
1 0.615385      0   1.00000 1.00000 0.21134
2 0.051282      1   0.38462 0.69231 0.19441
3 0.038462      4   0.23077 0.76923 0.20021
4 0.000000     10   0.00000 0.84615 0.20492
```

Trainset error

10-fold CV error

Plotted as the y-coor in plotcp() chart.

What is 10-fold Cross Validation Error?



That’s why we need to set.seed() before executing rpart.

0. Randomly split the data into 10 subsets.

1. Train on 9 pieces (blue), test on unseen 1st piece (yellow).

Trainset error 1, Testset error 1.

2. Train on 9 pieces (blue), test on unseen 2nd piece (yellow).

Trainset error 2, Testset error 2.

⋮

10. Train on 9 pieces (blue), test on unseen 10th piece (yellow).

Trainset error 10, Testset error 10.

1 SE Rule is just a guideline to select the optimal tree.

- Min CV error tree is an unstable solution.
  - A small change in data could lead to a different solution.
  - Depends on the random subsets in 10 fold CV.
- 1 SE rule is more stable.
  - Many trees are statistically equivalent in terms of errors.
  - Choose the simplest tree that still perform well.

Next Video: CART for Continuous Y

- Continuous Y:
  - How to choose the best split?
  - How to evaluate node error and overall Tree error.
  - Variable Importance.
    - How important are each of those X variables?