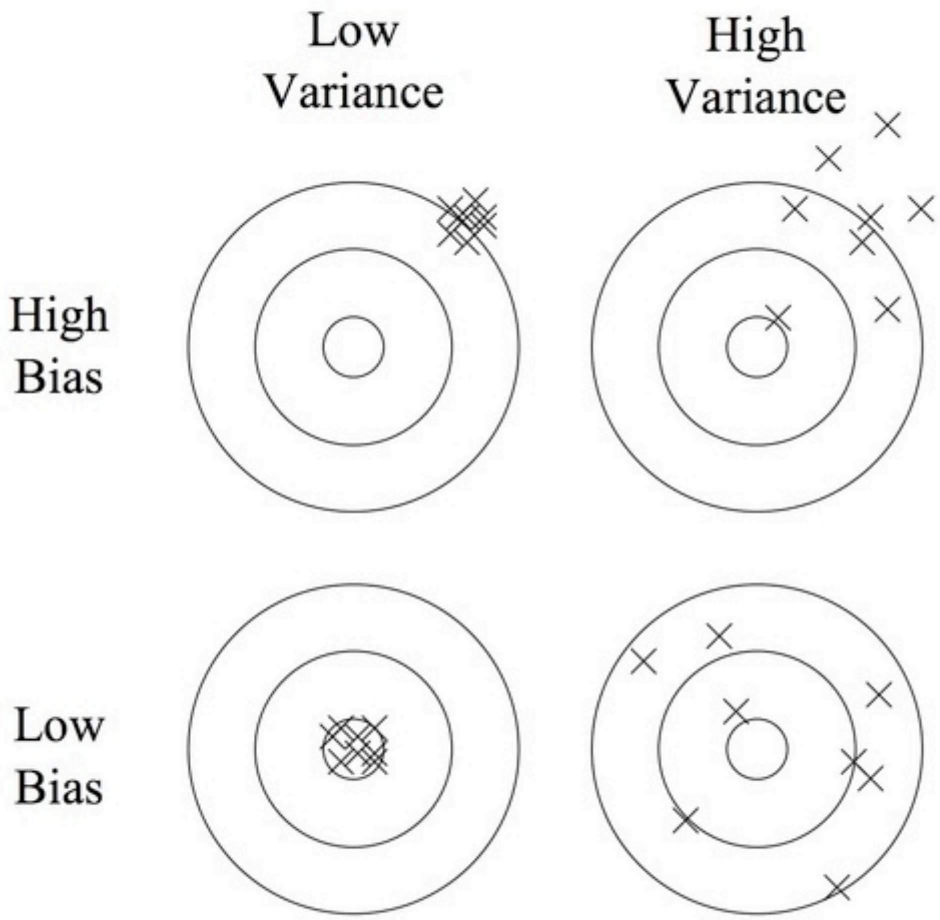
Machine learning:
Random forests

Over-fit, variance/bias dilemma

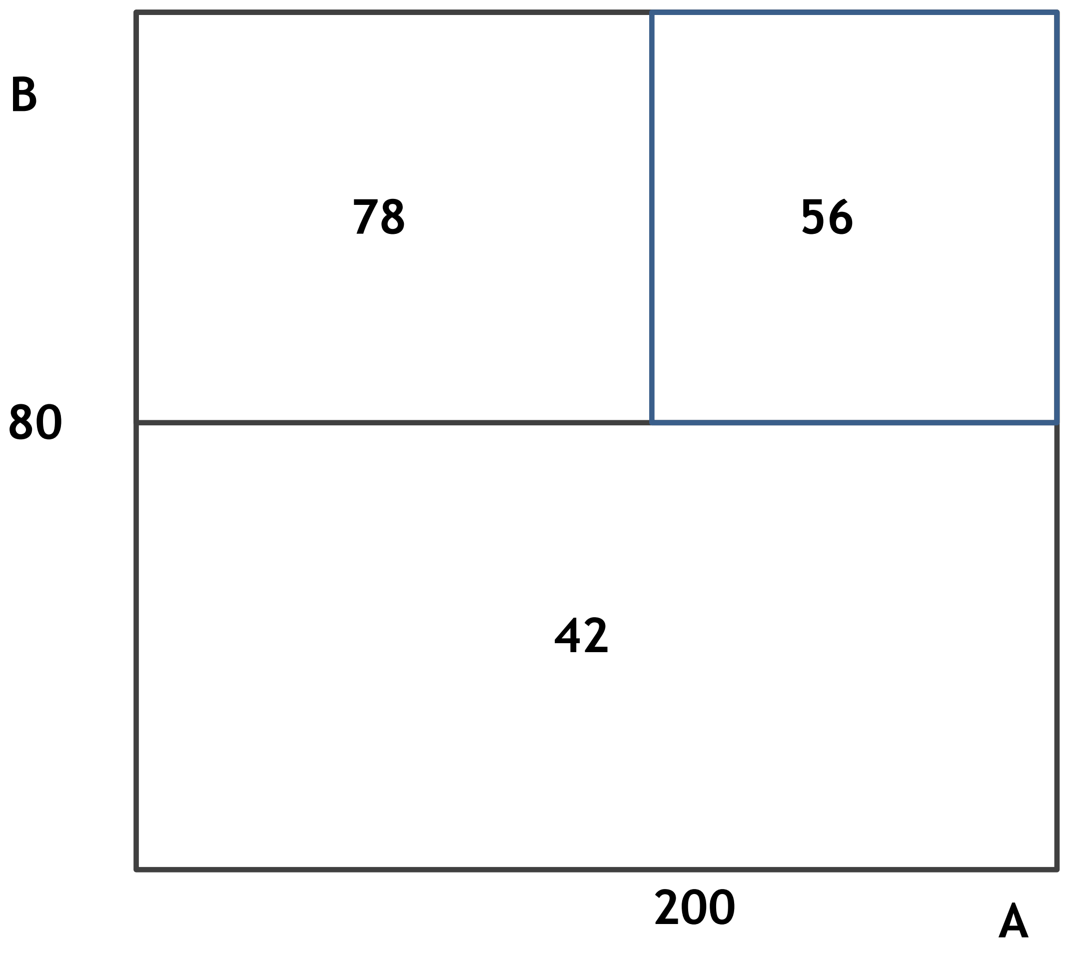


Regression trees

```
if Predictor B >= 80 then
  if Predictor A >= 200 then dependent measure
is 56
  else dependent measure is 78
else dependent measure is 42
```

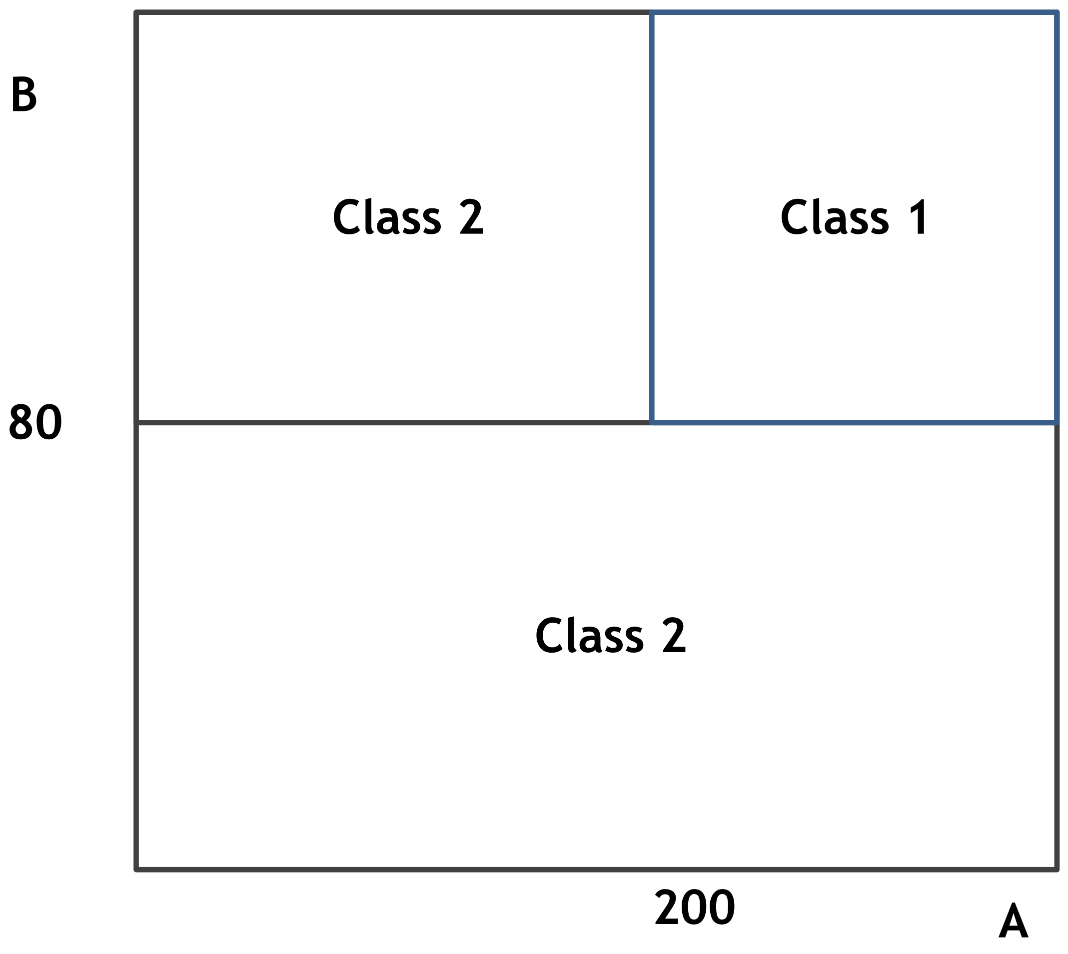
Regression trees

```
if Predictor B >= 80 then
  if Predictor A >= 200 then dependent measure is 56
  else dependent measure is 78
else dependent measure is 42
```

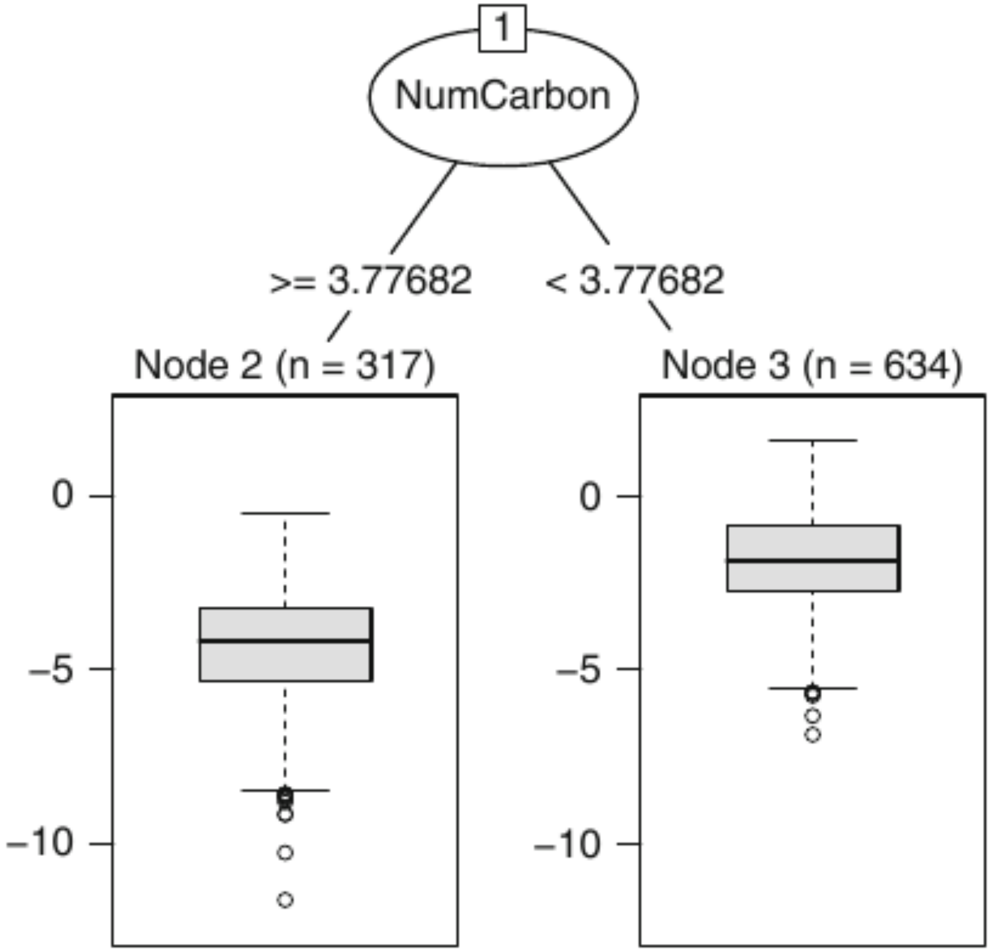


Categorization trees

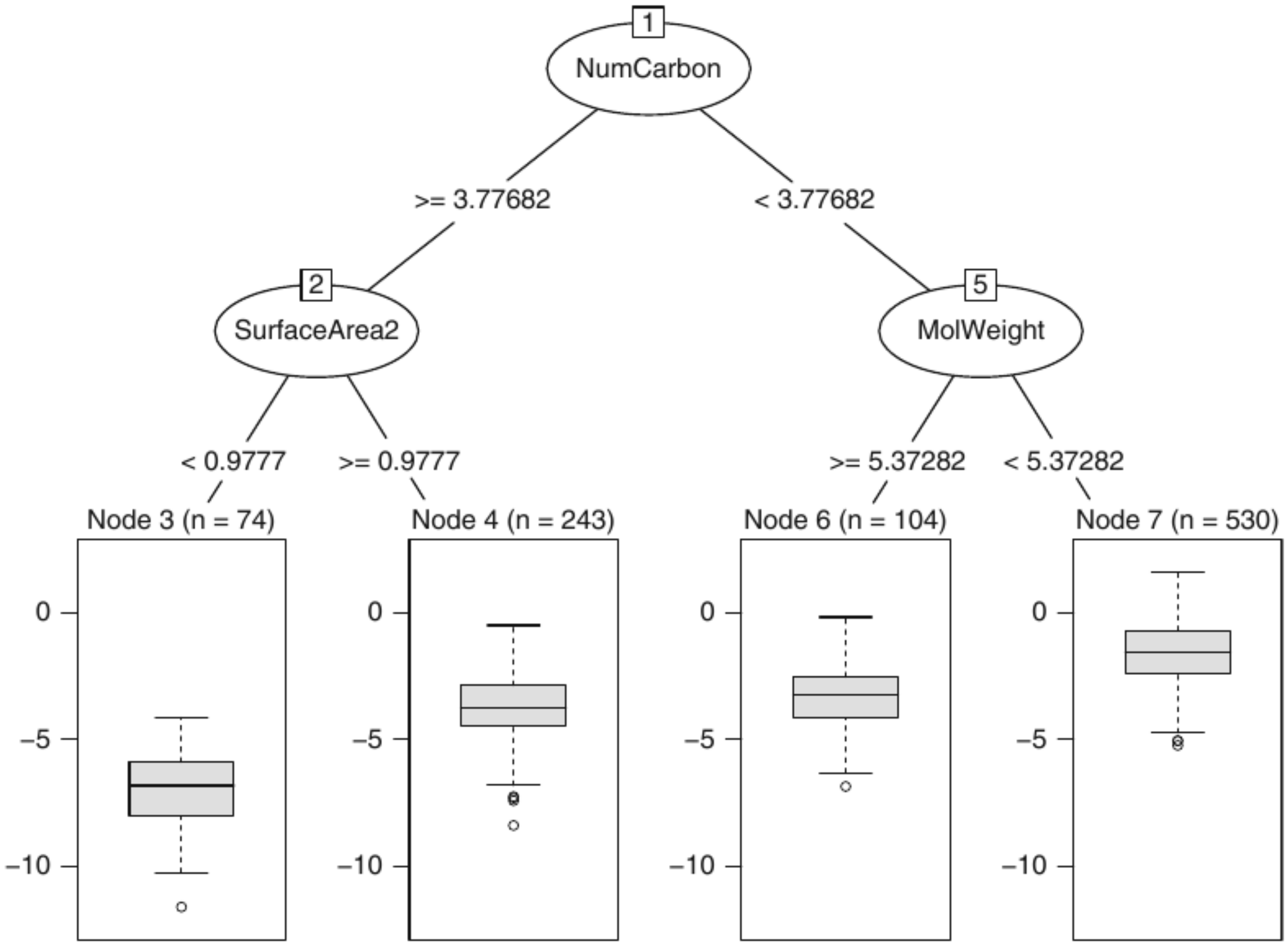
```
if Predictor B >= 80 then
  if Predictor A >= 200 then dependent
measure is class 1
  else dependent measure is class 2
else dependent measure is class 2
```



Trees



Trees



1
NumCarbon
>= 3.77682
< 3.77682
2
SurfaceArea2
5
MolWeight
< 0.9777
>= 0.9777
>= 5.37282
< 5.37282
Node 3 (n = 74)
Node 4 (n = 243)
Node 6 (n = 104)
Node 7 (n = 530)
0
−5
−10

Problems with decision trees

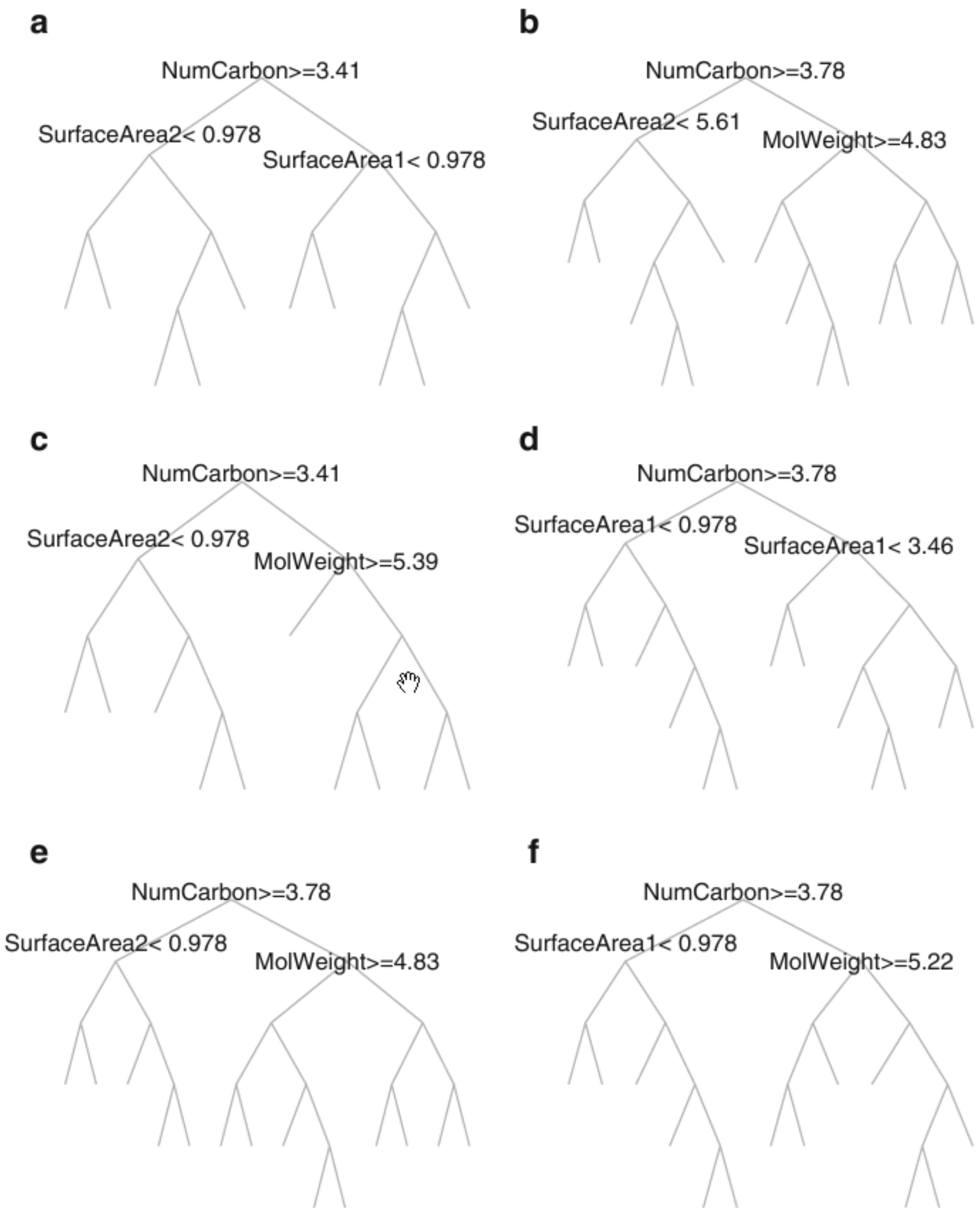
- Single regression trees are more likely to have sub-optimal predictive performance compared to other modeling approaches
- Decision boundaries are linear, trouble if your data is not linearly separable
- If the data are slightly altered, a completely different set of splits might be found
- Selection bias: predictors with a higher number of distinct values are favored

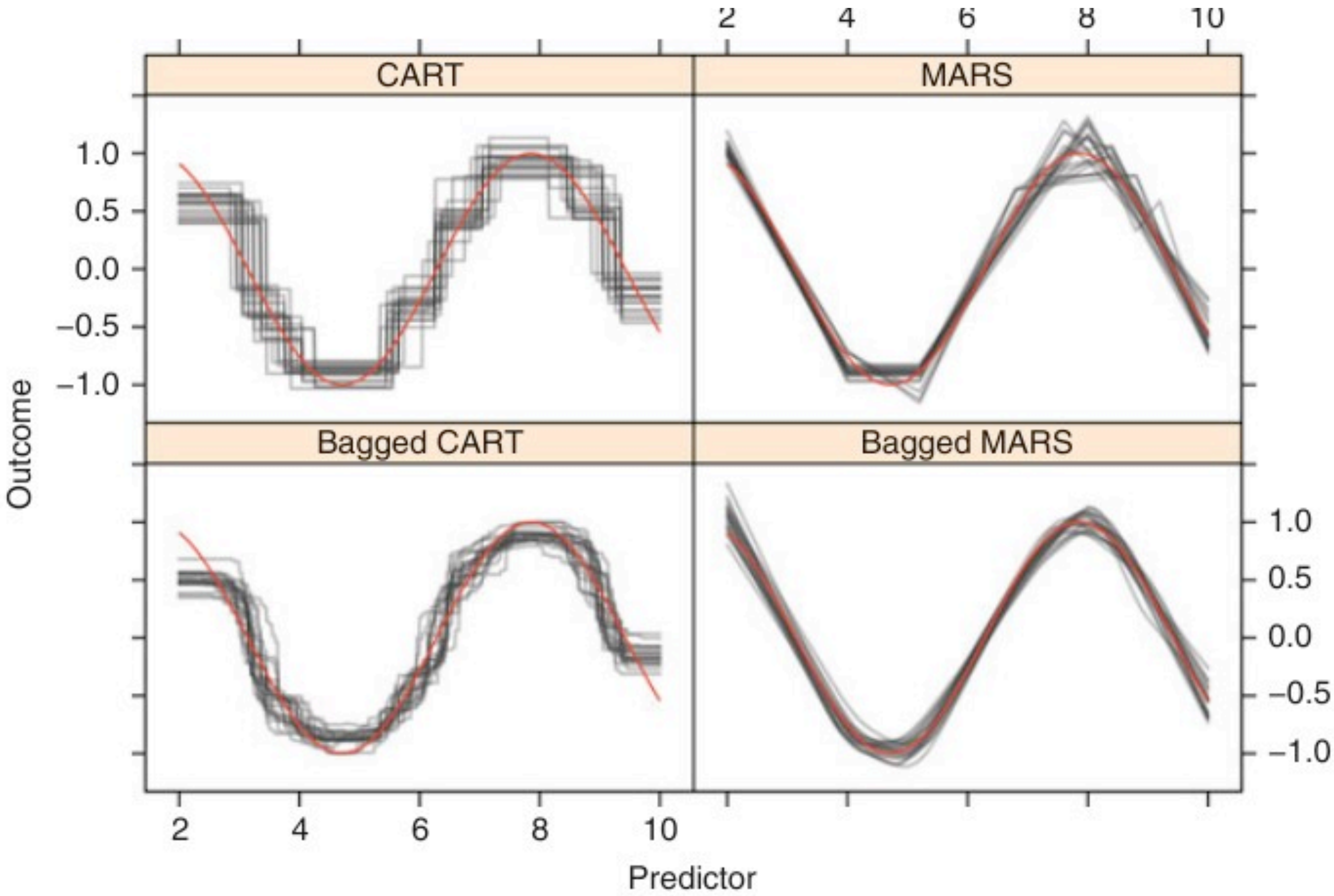
Solution

- Generating bootstrap samples introduces a random component into the tree building process

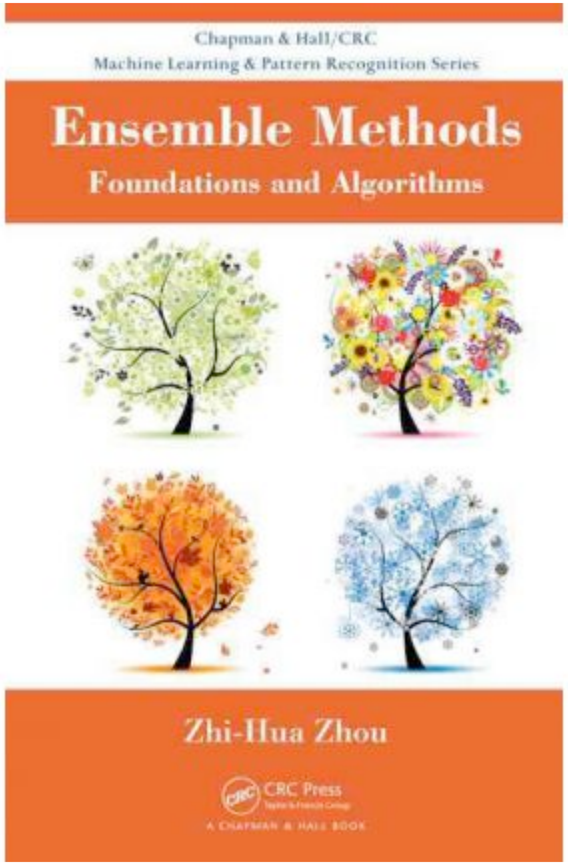
Bagging

- Boostrapping and Aggregating (B-Agg-ing)

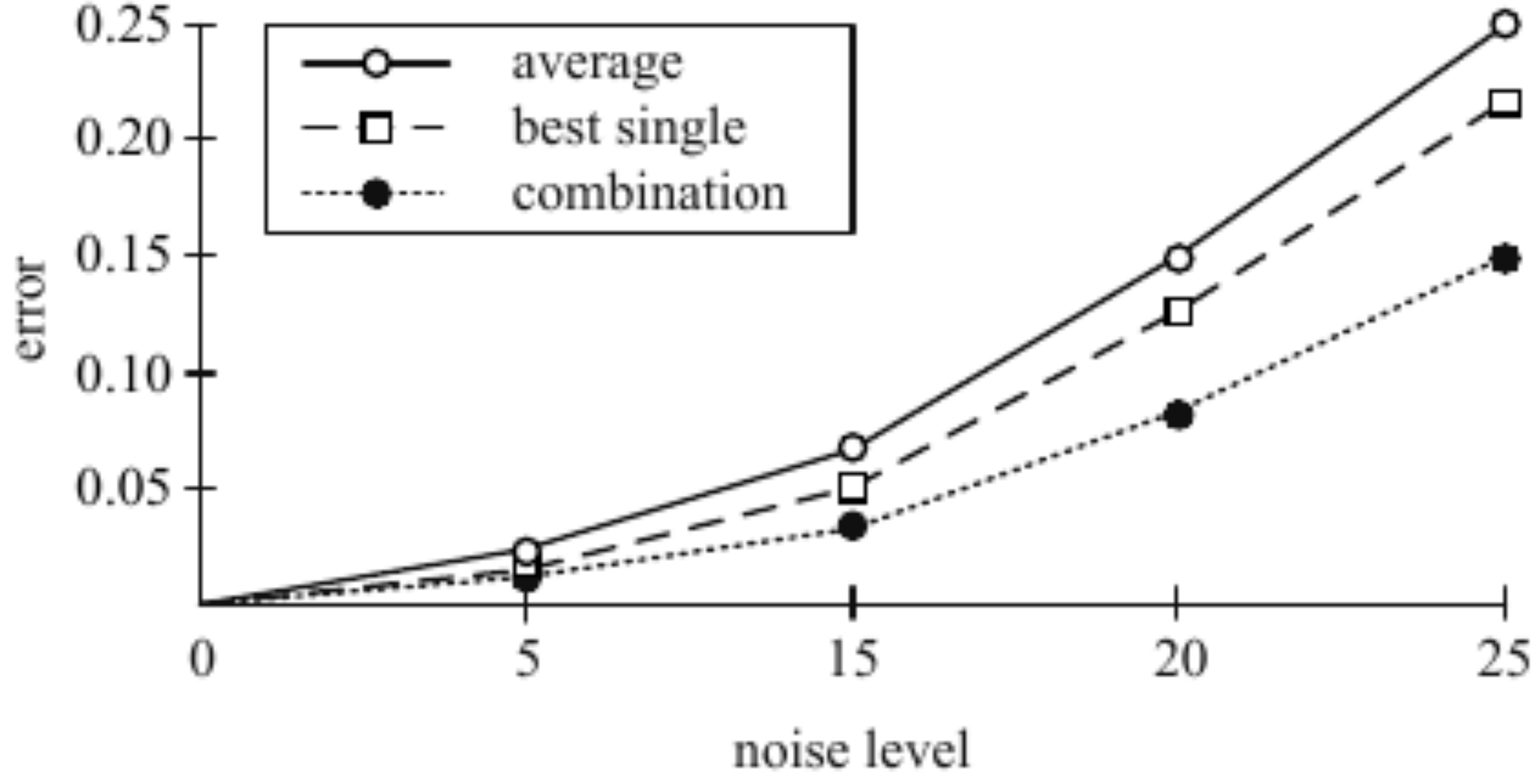



CART
MARS
Bagged CART
Bagged MARS
Outcome
Predictor
1.0
0.5
0.0
−0.5
−1.0
2
4
6
8
10

Ensemble Methods
Foundations and Algorithms



Hansen and Salamon (1990)'s observation: Ensemble is often better than the best single



- Switch to detailed slides