Tree based methods

- Decision trees
- Bagging
- Random forests
- Boosting
- BART

- What (not) to do

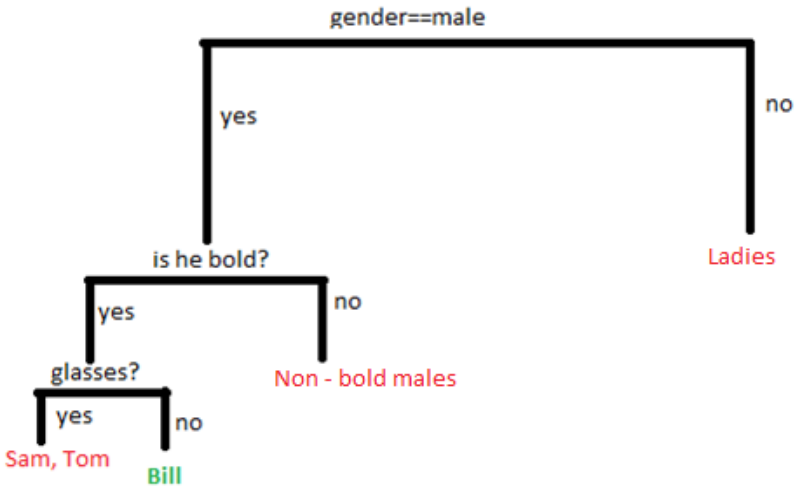
One slide about me

- Mathematics, Molecular biology, PhD in bioinformatics
- Running, crossfit
- Husband, daughter

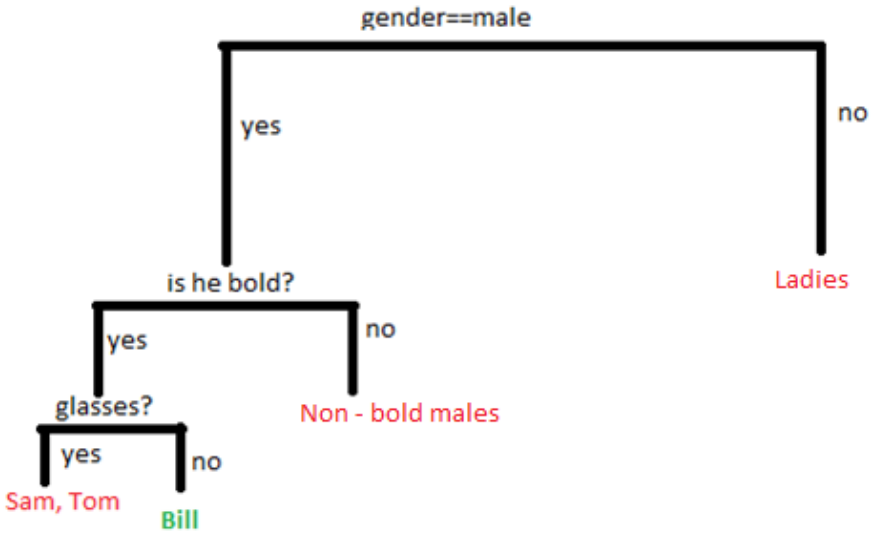
Decision trees

Decision trees

	gender	isBold	longHair	hat	galsses	blondHair
Sam	male	yes	no	no	yes	no
Tom	male	yes	no	no	yes	no
Maria	female	no	yes	yes	no	no
Bill	male	yes	no	no	**no**	no
Charles	male	no	no	no	no	yes
Paul	male	no	no	no	yes	no
Joe	male	no	no	no	yes	yes
David	male	no	no	no	no	yes
Anita	female	no	yes	no	no	yes
Claire	female	no	yes	yes	yes	no
	female	no	no	no	no	no
	male	no	no	yes	no	yes
	male	no	no	no	no	no
	male	no	no	no	no	no
	female	no	yes	no	no	no
	male	no	no	yes	no	no



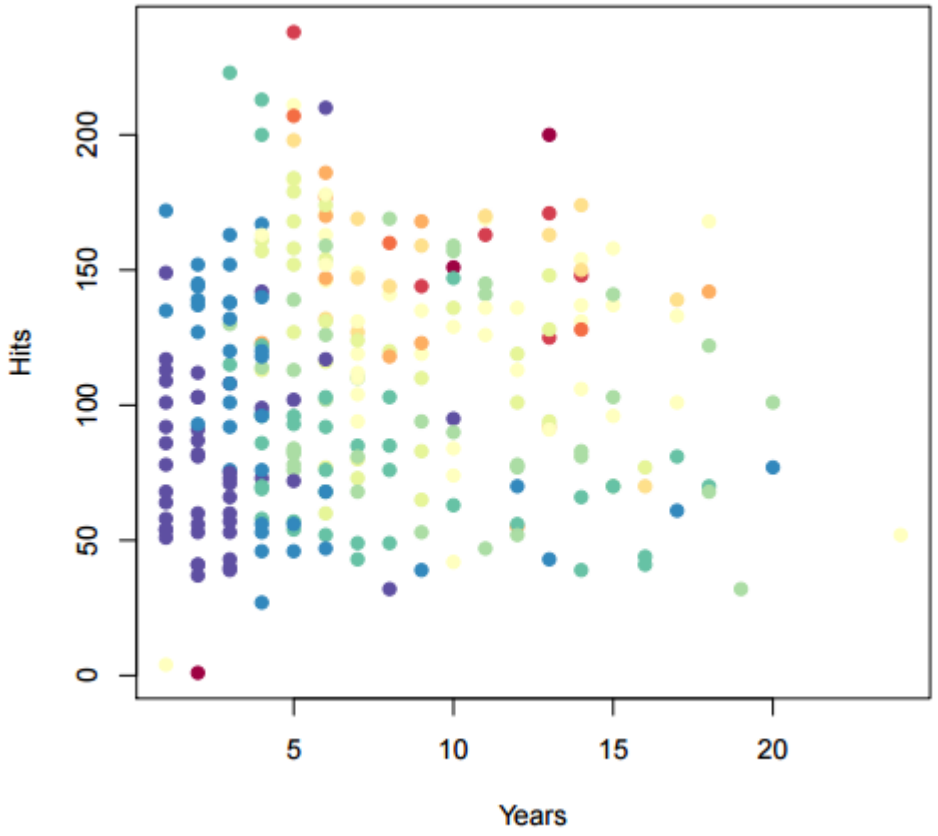
Decision trees



Terminology:

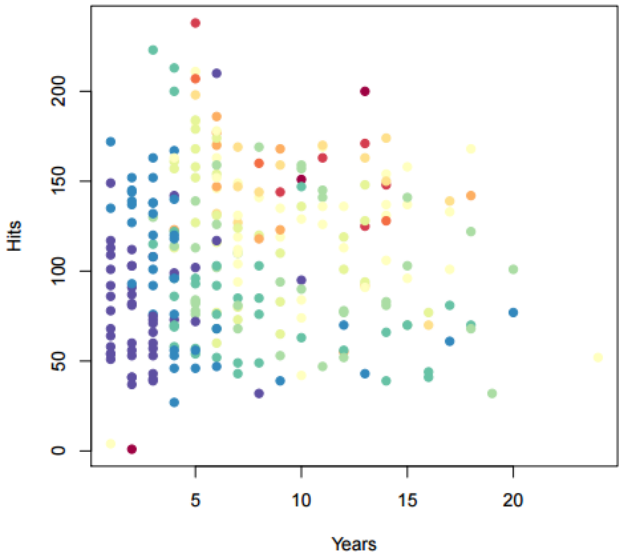
Classification tree
3 internal nodes
4 terminal nodes = leafs, they represent the mean of the response for the observations that fall there.

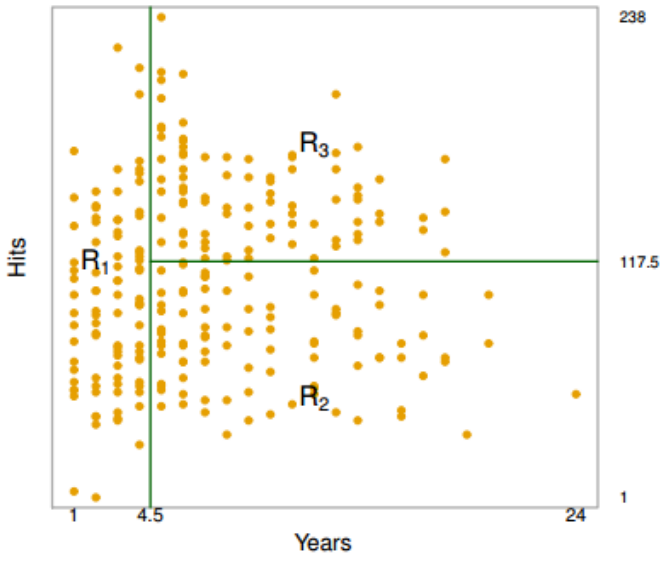
Decision trees

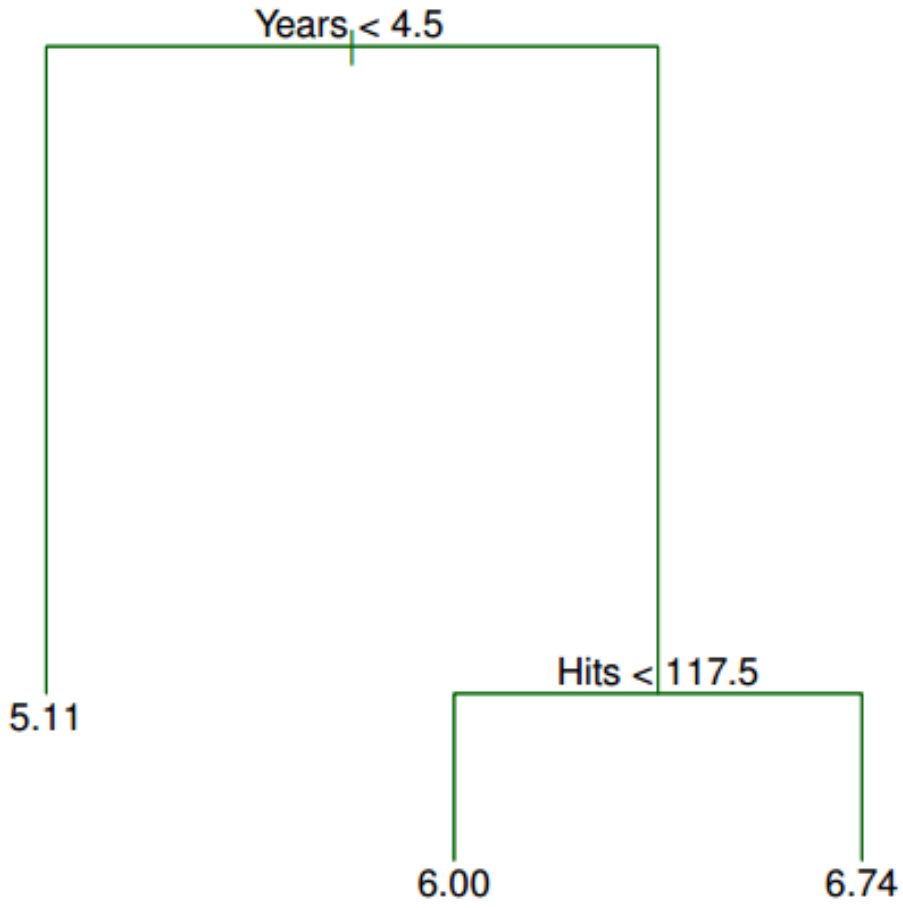


Hitters	Years	Hits	logSalary
-Alan Ashby	14	81	6.163315
-Alvin Davis	3	130	6.173786
-Andre Dawson	11	141	6.214608
-Andres Galarraga	2	87	4.516339
-Alfredo Griffin	11	169	6.620073
-Al Newman	2	37	4.248495
-Argenis Salazar	3	73	4.605170
-Andres Thomas	2	81	4.317488
-Andre Thornton	13	92	7.003065
-Alan Trammell	10	159	6.248319
-Alex Trevino	9	53	6.239301
-Andy VanSlyke	4	113	6.309918

Decision trees

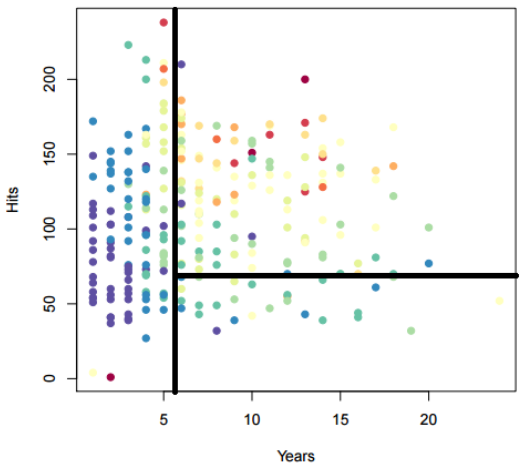




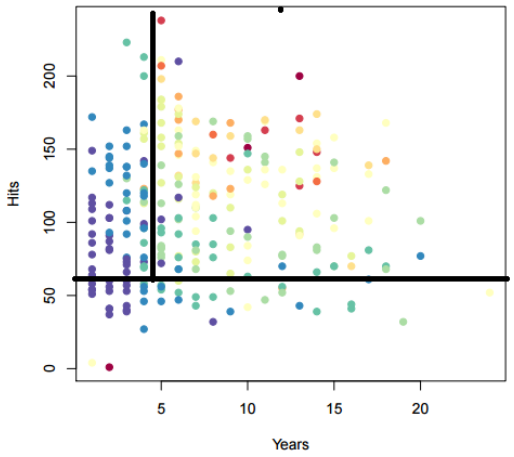


How to build a tree? – General idea

Divide predictor space into J distinct and nonoverlaping boxes.



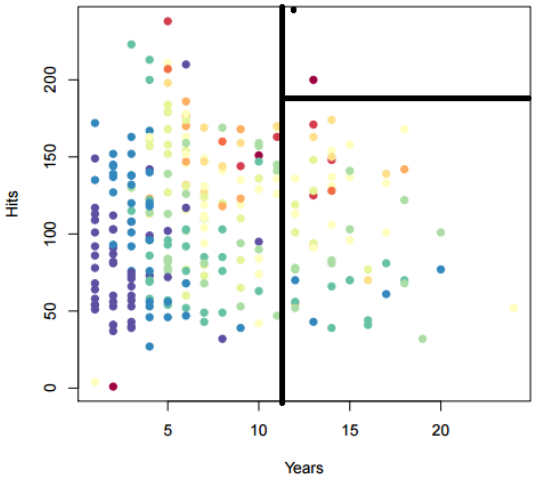
Possible division 1



Possible division 2

. . .

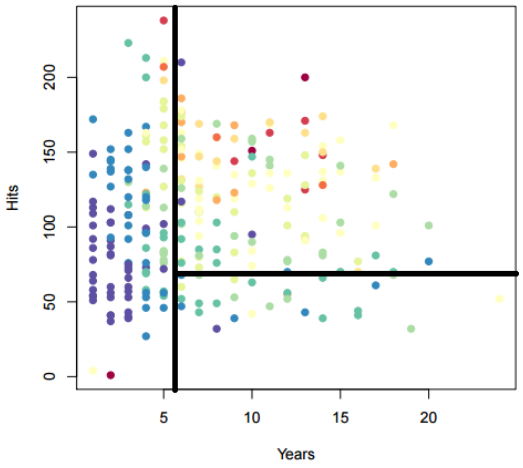
. . .



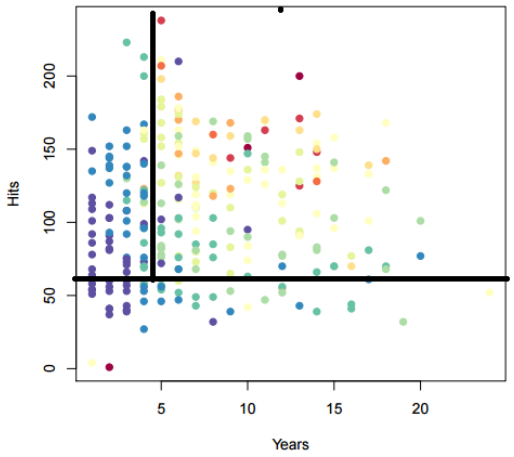
Possible division n

How to build a tree?

Divide predictor space into J distinct and nonoverlaping boxes.

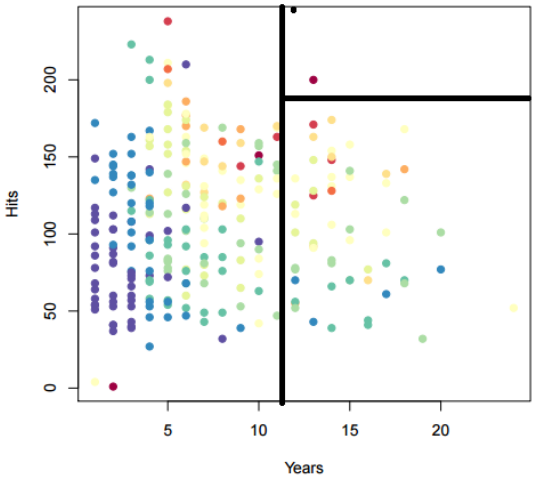


Possible division 1



Possible division 2

. . .

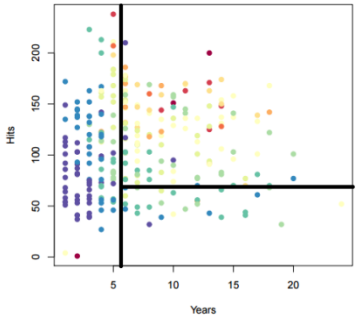


Possible division n

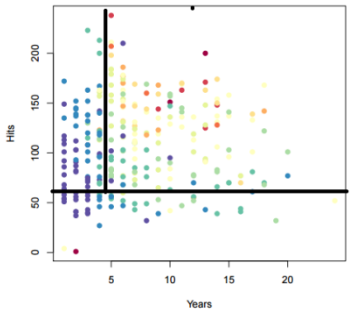
. . .

Which division is better?

How to build a tree? – General idea



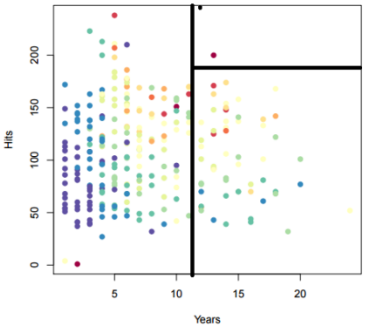
Possible division 1



Possible division 2

. . .

. . .



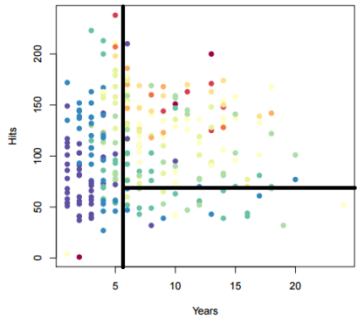
Possible division n

Which division is better? – use RSS

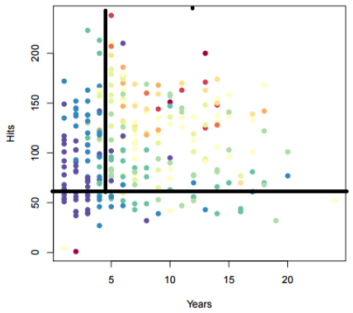
$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

The goal is to find boxes R1, . . . , RJ that minimize the RSS, where ˆyRj is the mean response for the training observations within the jth box

How to build a tree? – General idea

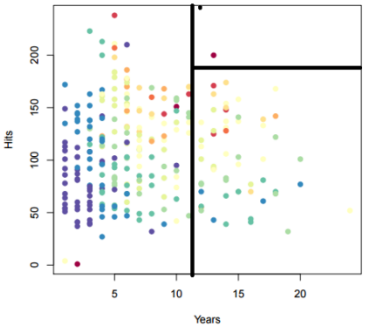


Possible division 1



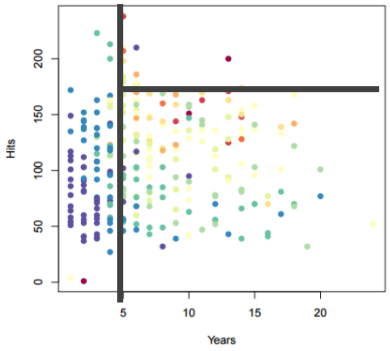
Possible division 2

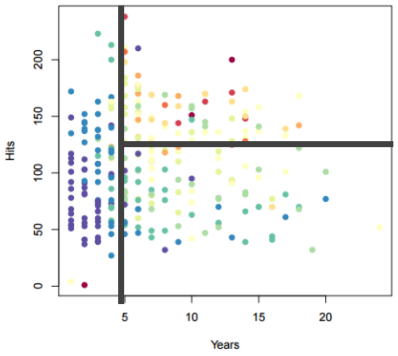
. . .

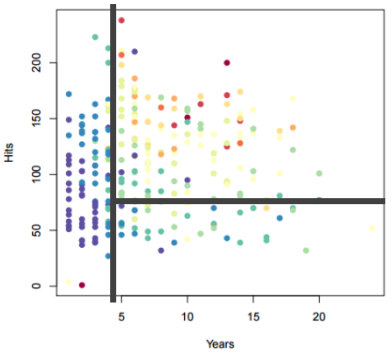


Possible division n

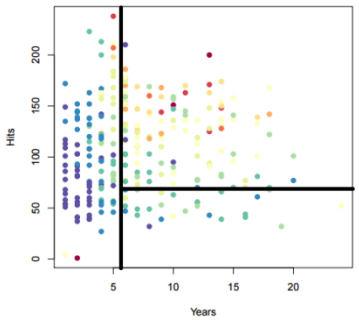
. . .



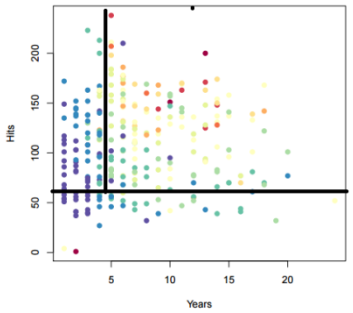




How to build a tree? – General idea

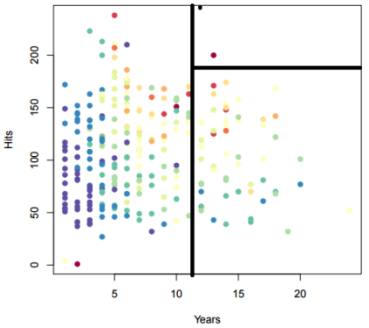


Possible division 1



Possible division 2

. . .



Possible division n

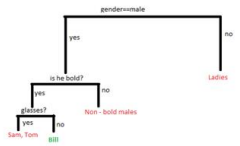
. . .

There are too many possible divisions…

Where to start?

Decision trees

	gender	isBold	longHair	hat	glasses	blondHair
Sam	male	yes	no	no	yes	no
Tom	male	yes	no	no	yes	no
Maria	female	no	yes	yes	no	no
Bill	male	yes	no	no	no	no
Charles	male	no	no	no	no	yes
Paul	male	no	no	no	yes	no
Joe	male	no	no	no	yes	yes
David	male	no	no	no	no	yes
Anita	female	no	yes	no	no	yes
Claire	female	no	yes	yes	yes	no
	female	no	no	no	no	no
	male	no	no	yes	no	yes
	male	no	no	no	no	no
	male	no	no	no	no	no
	female	no	yes	no	no	no
	male	no	no	yes	no	no

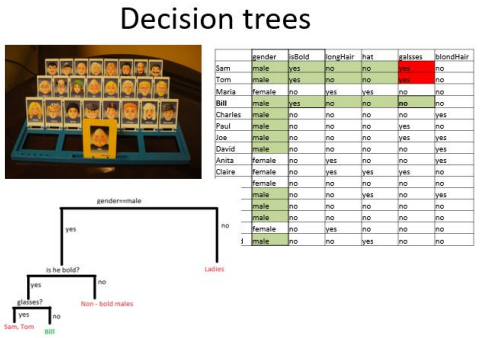


How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:

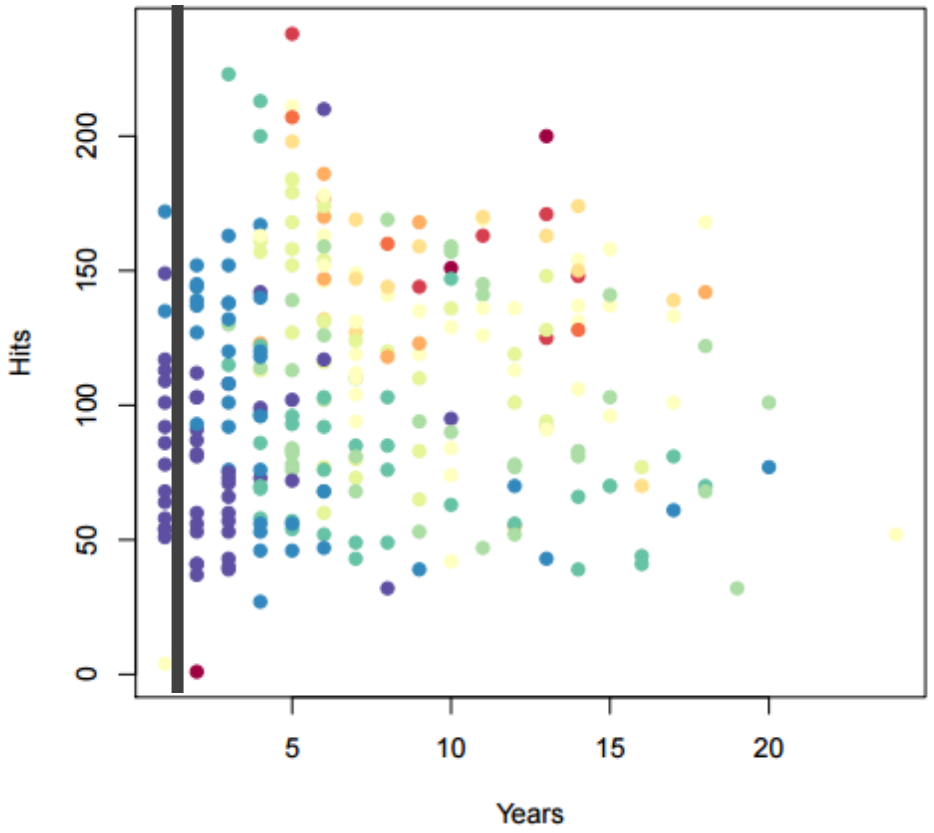
Top down : begin at the top and act successively
greedy : don't look ahead

First ask about the gender because it will filter out most subjects in most cases.



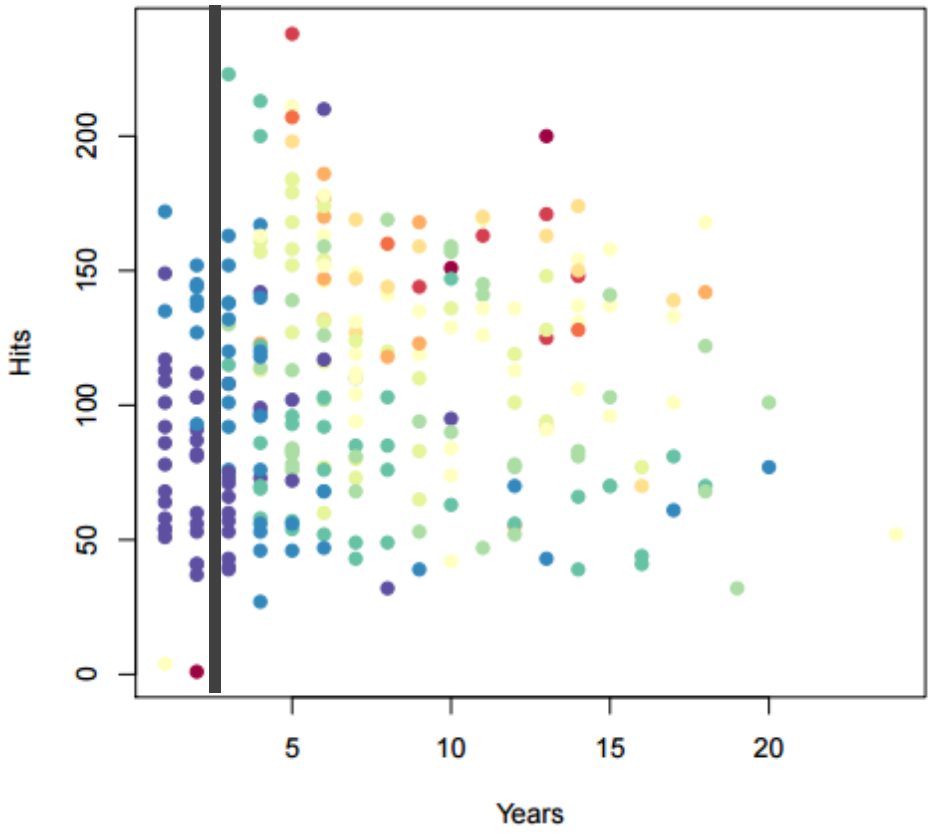
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



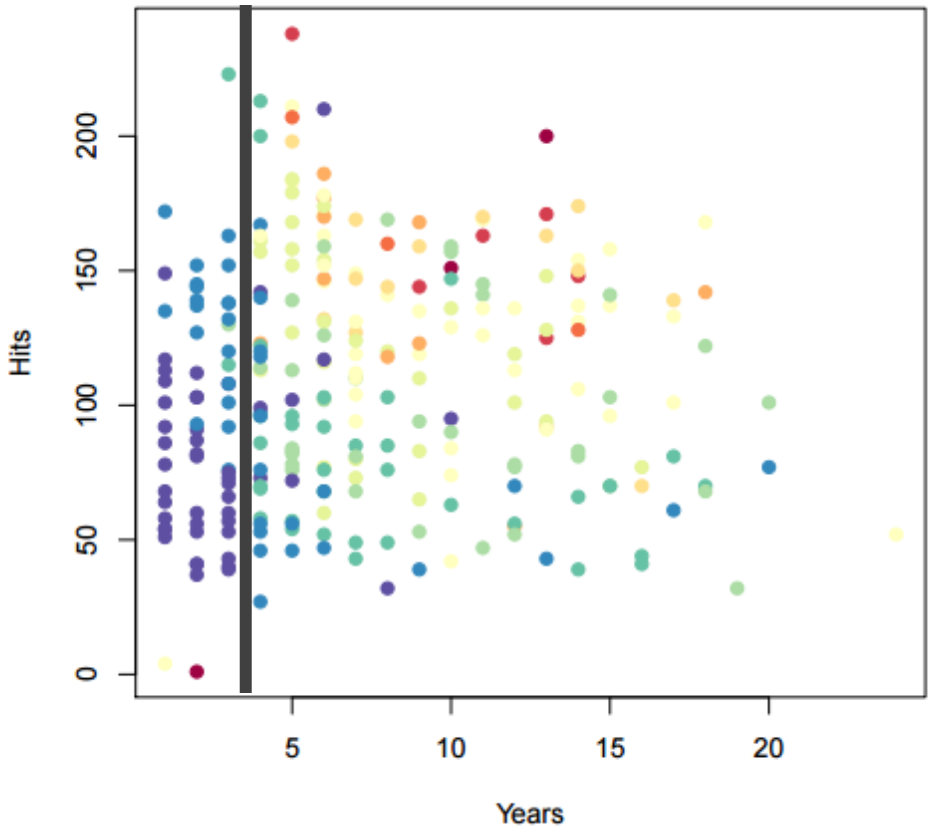
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



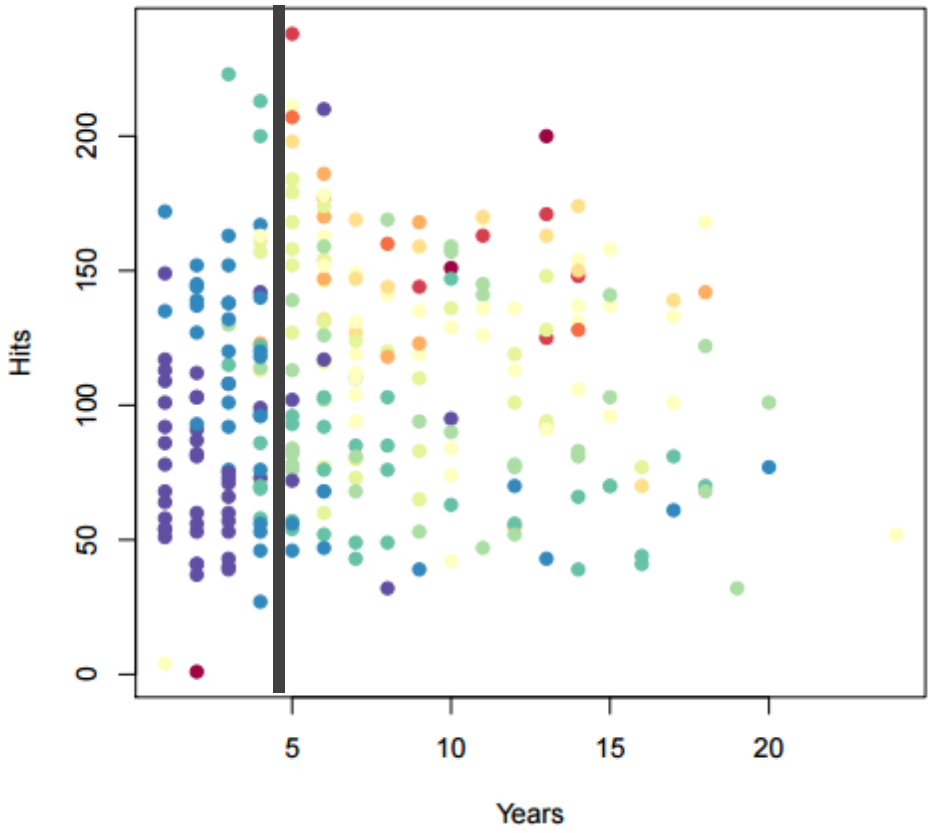
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



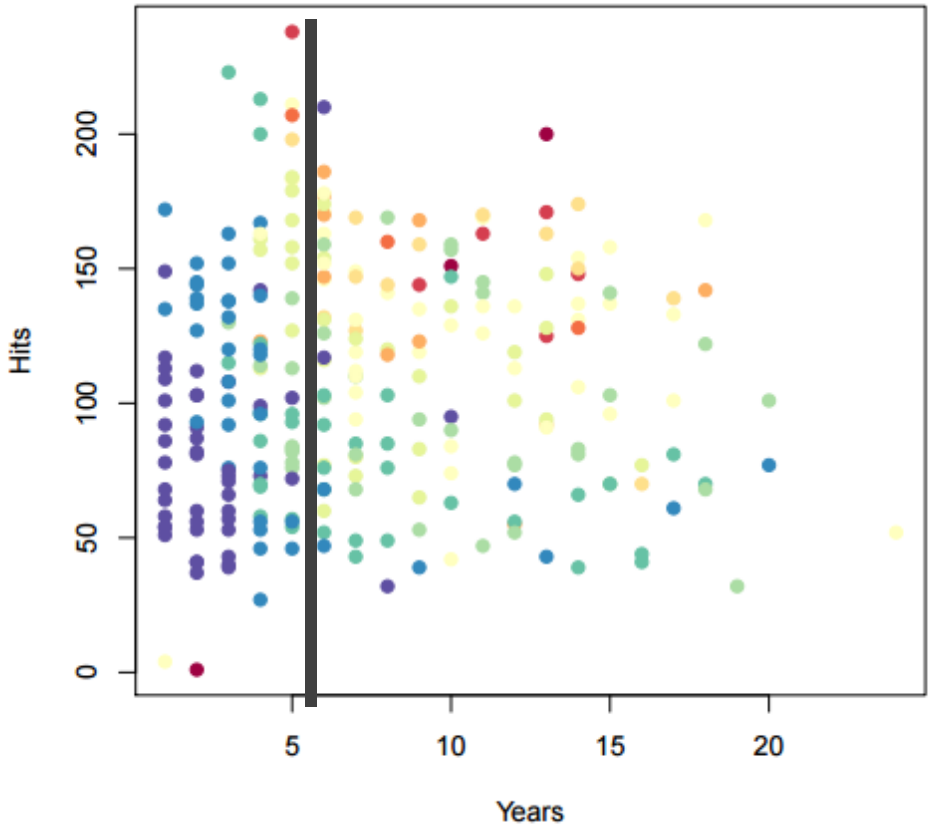
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



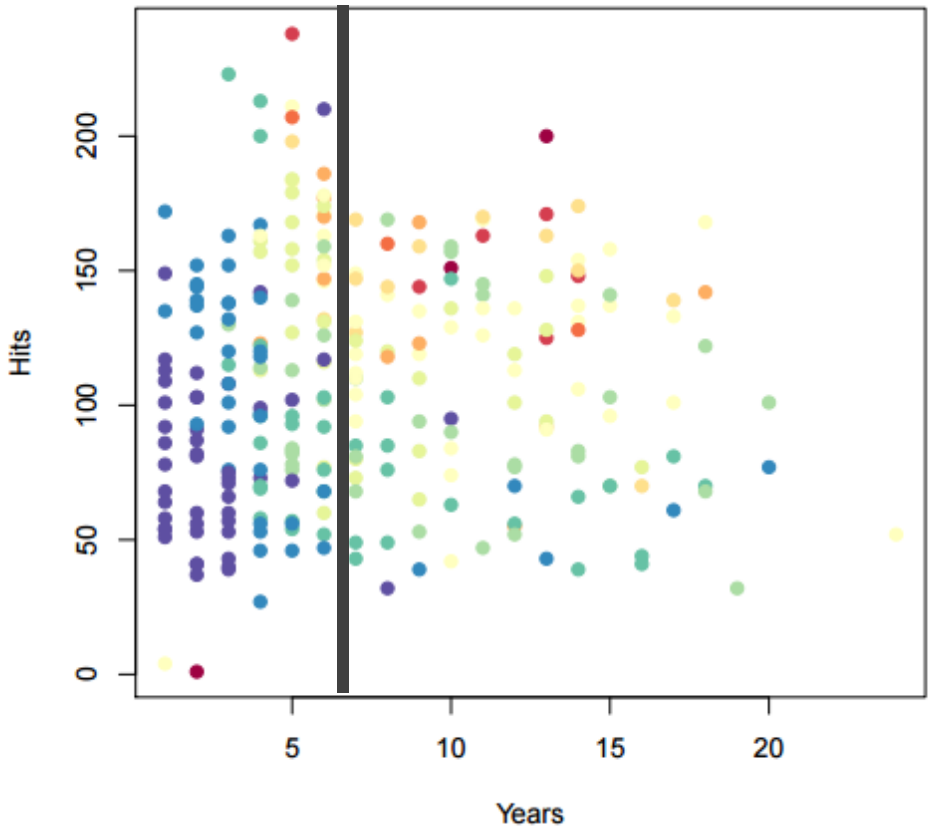
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



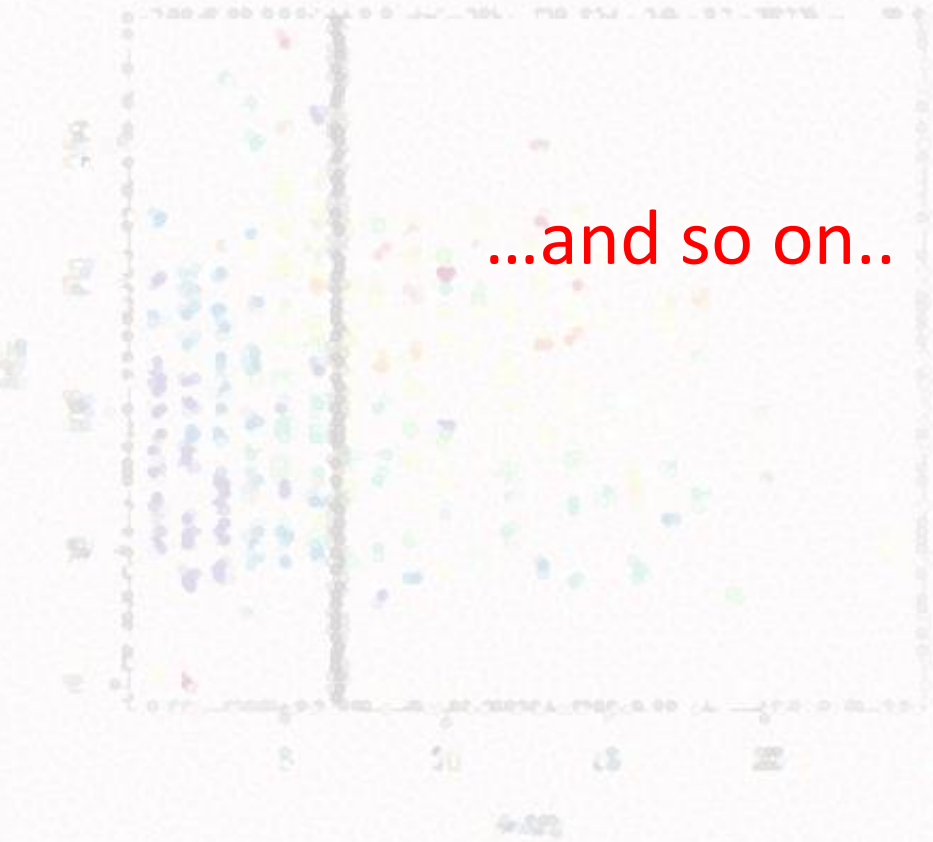
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



How to build a tree? – General idea

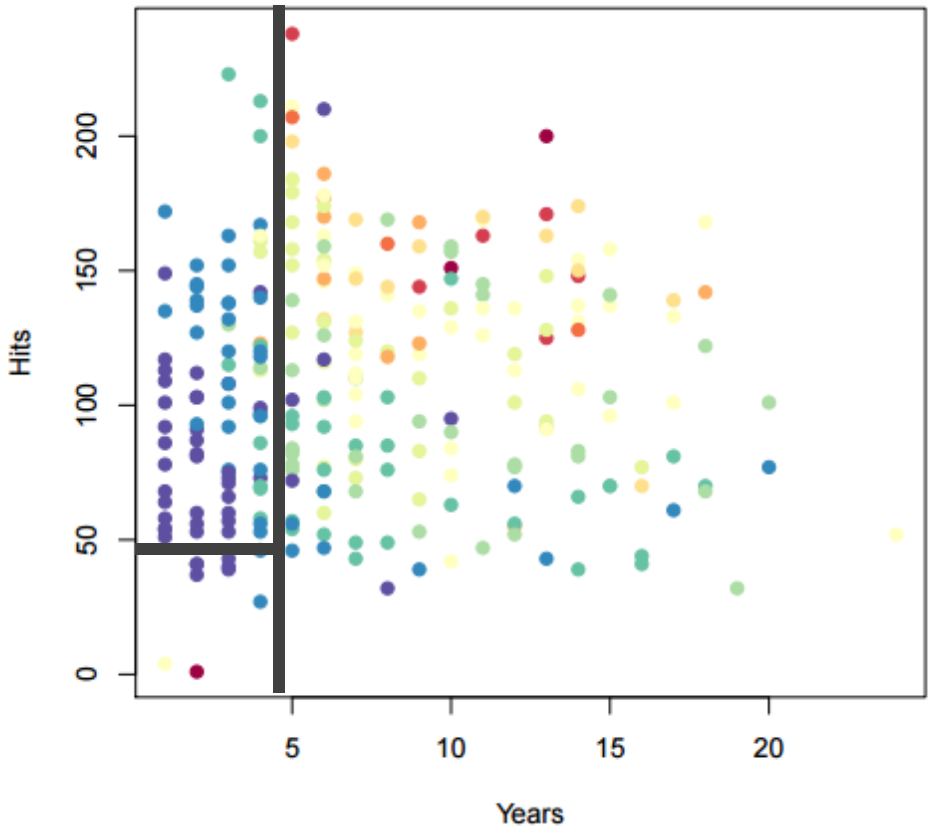
There are too many possible partitions? –
Use binary splitting approach:



…and so on..

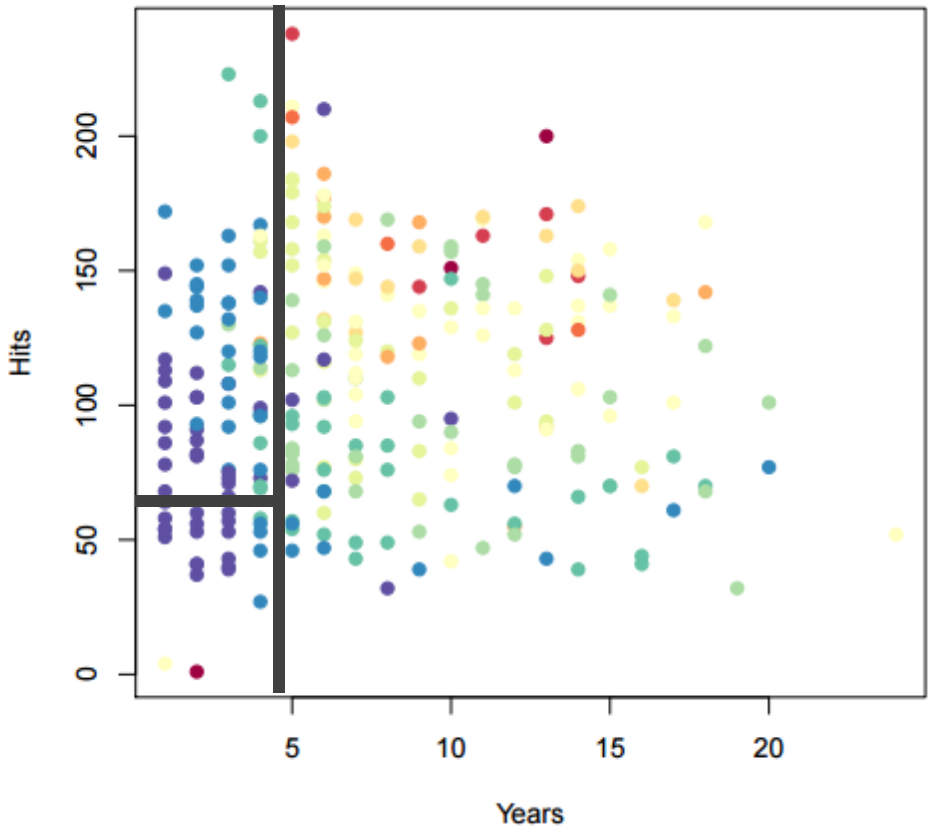
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



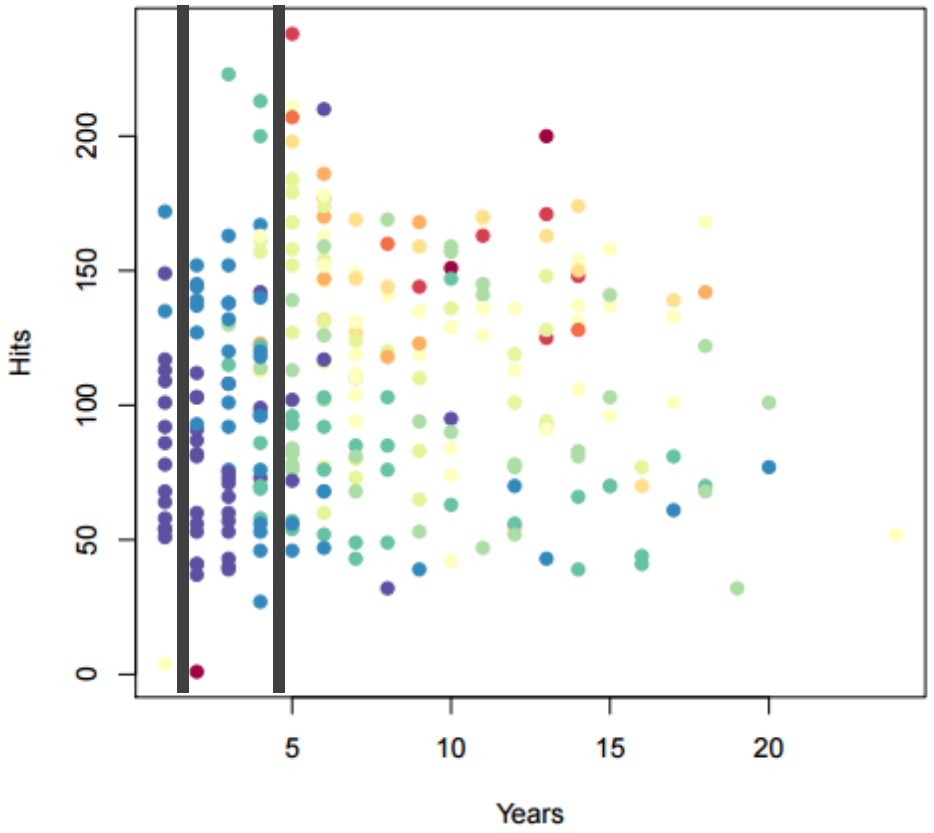
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



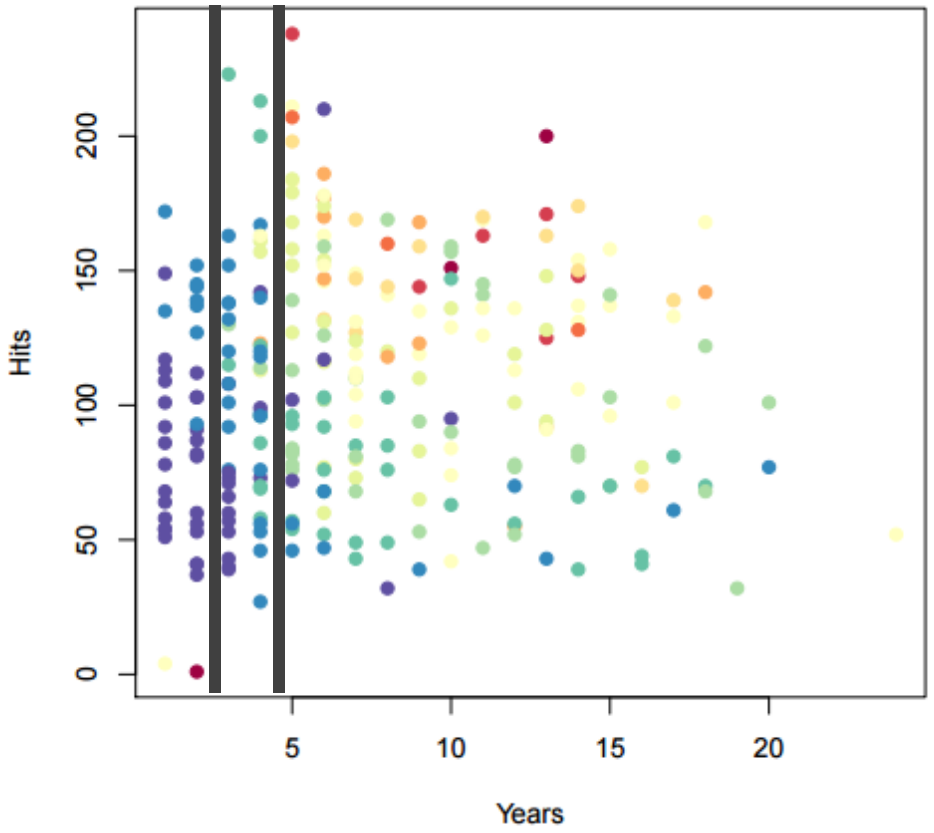
How to build a tree? – General idea

There are too many possible partitions? –
Use binary splitting approach:



How to build a tree? – General idea

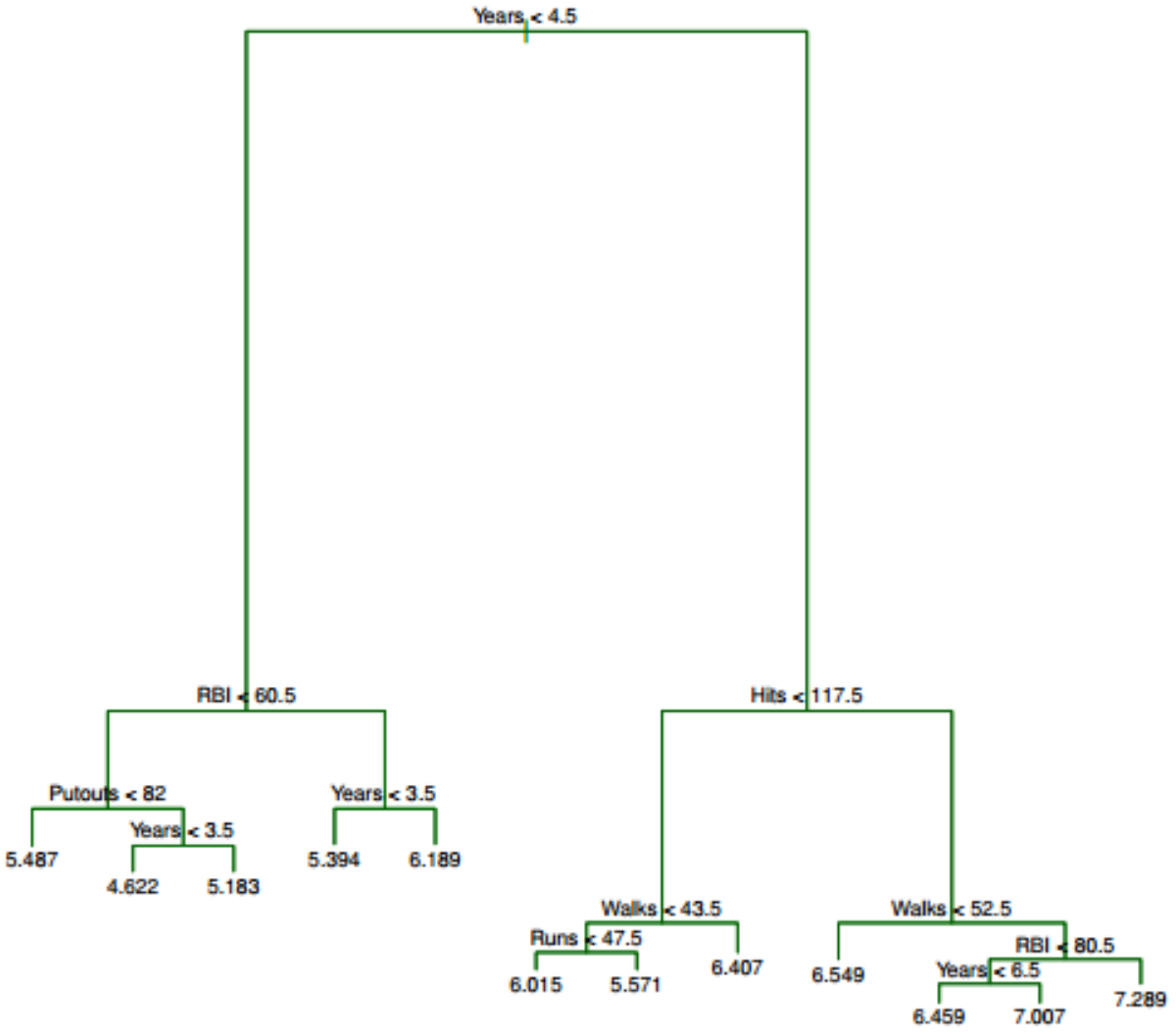
There are too many possible partitions? –
Use binary splitting approach:



Building a tree in details (1,2,3)

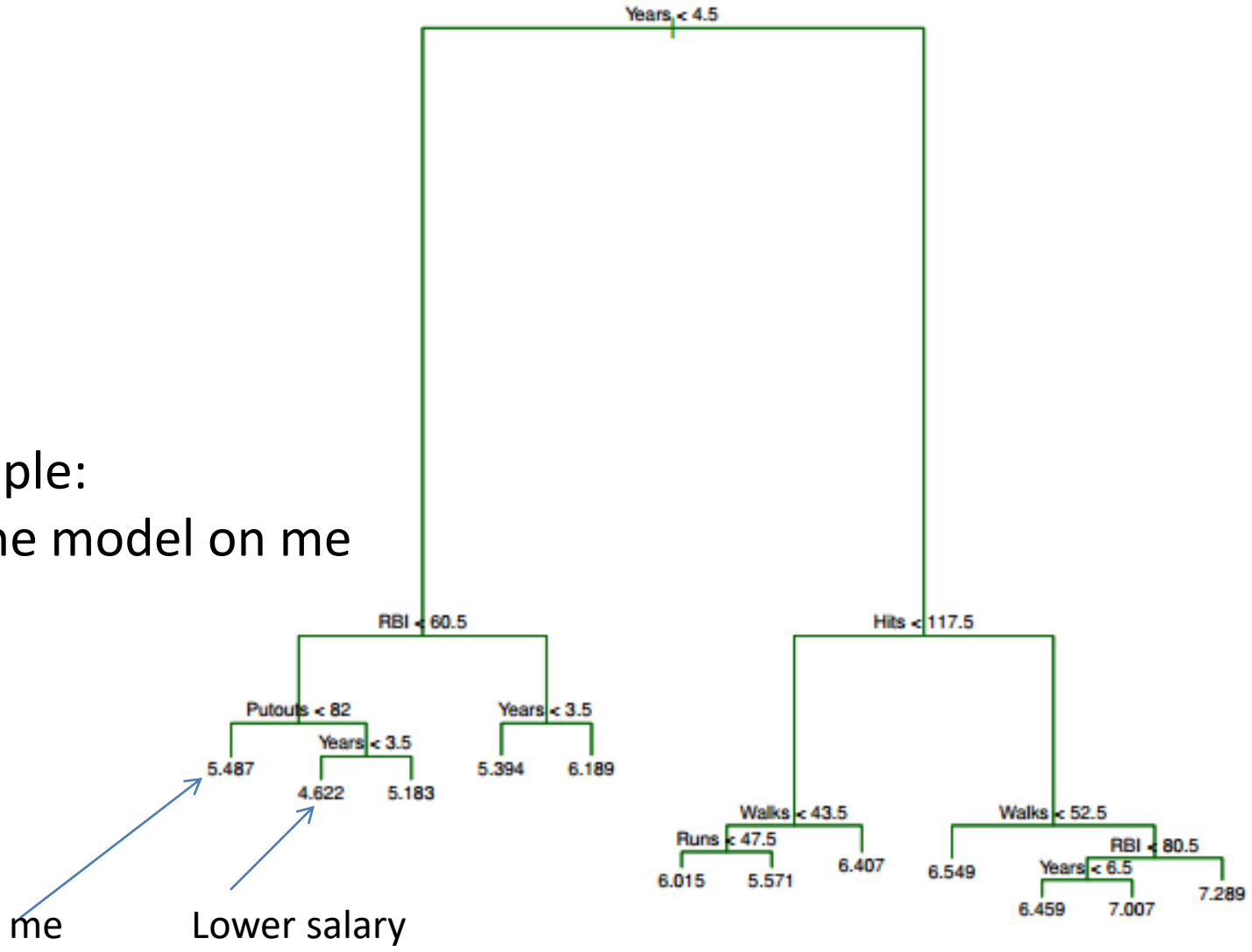
1. Select a predictor and cutpoint such that that split will lead to greatest reduction in RSS
2. Repeat within defined regions
3. Stop at some point. (for example when a region after the next splitting would contain less than 5 observations)
4. Predict response for new data using this tree as a model.

How to predict?



Problem: overfitting.

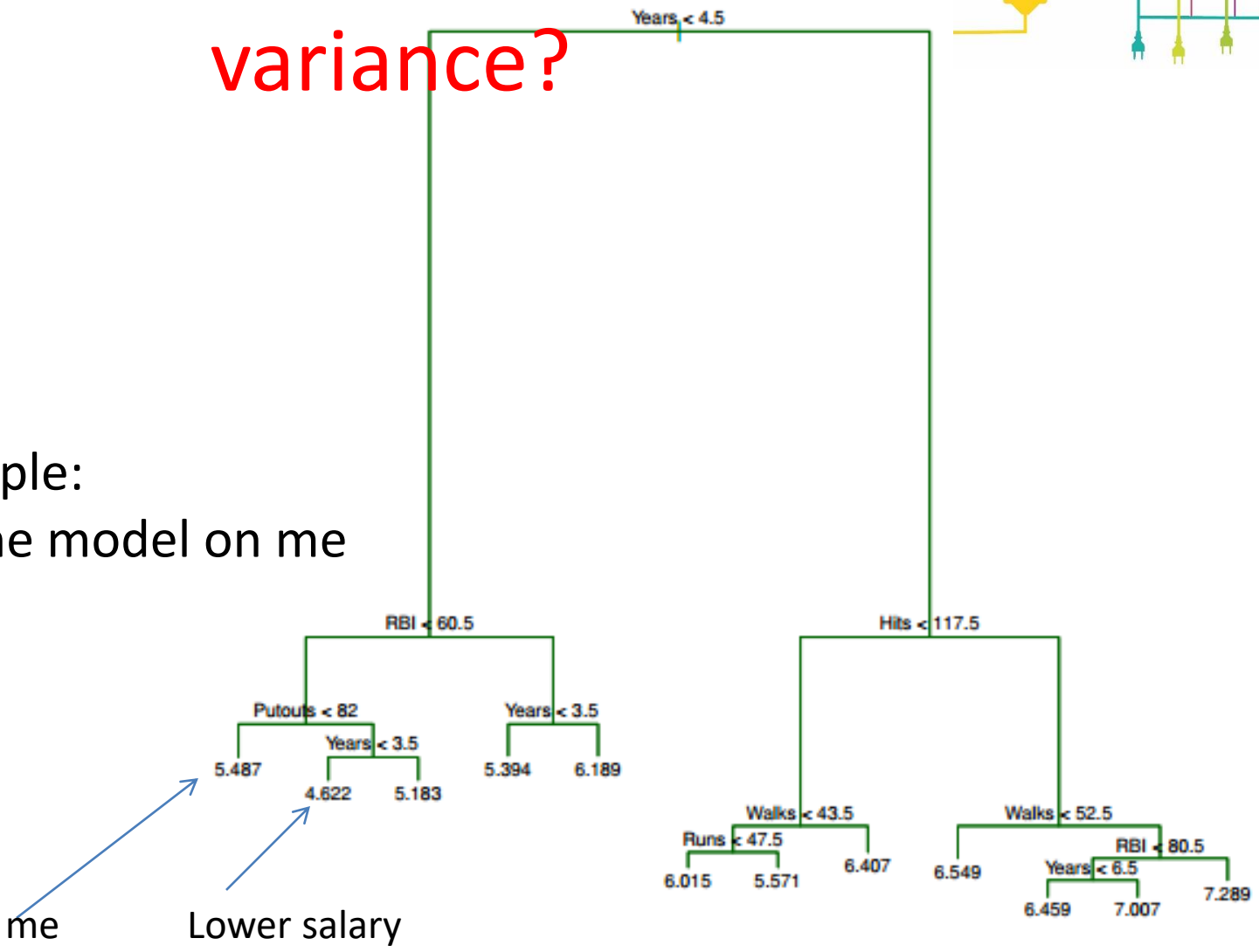
Example:
Try the model on me



What can we do to this tree to reduce the variance?

Example:

Try the model on me

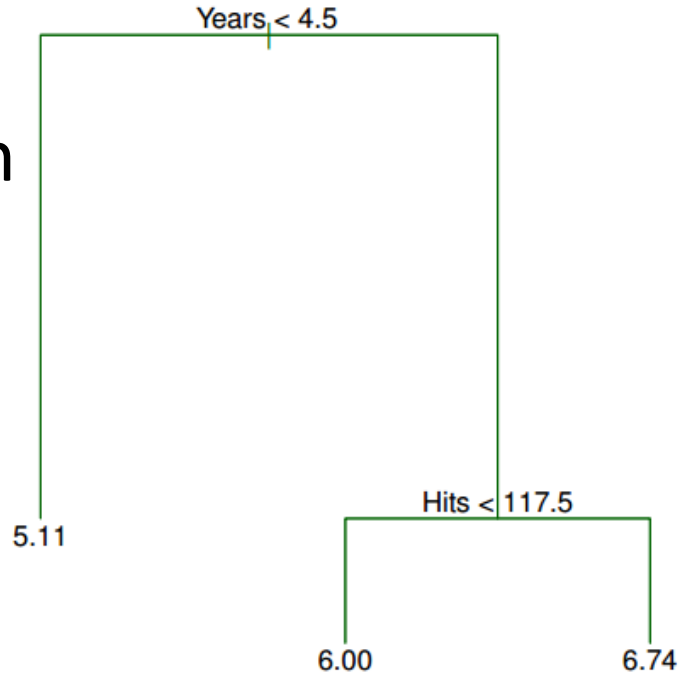


Solution: pruning

Pay a cost for complexity:

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

Select optimal α by cross validation



Classification trees – no RSS

$$E = 1 - \max_k(\hat{p}_{mk}).$$

Classification error rate - not sufficiently sensitive.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

Gini index – a measure of node „purity”

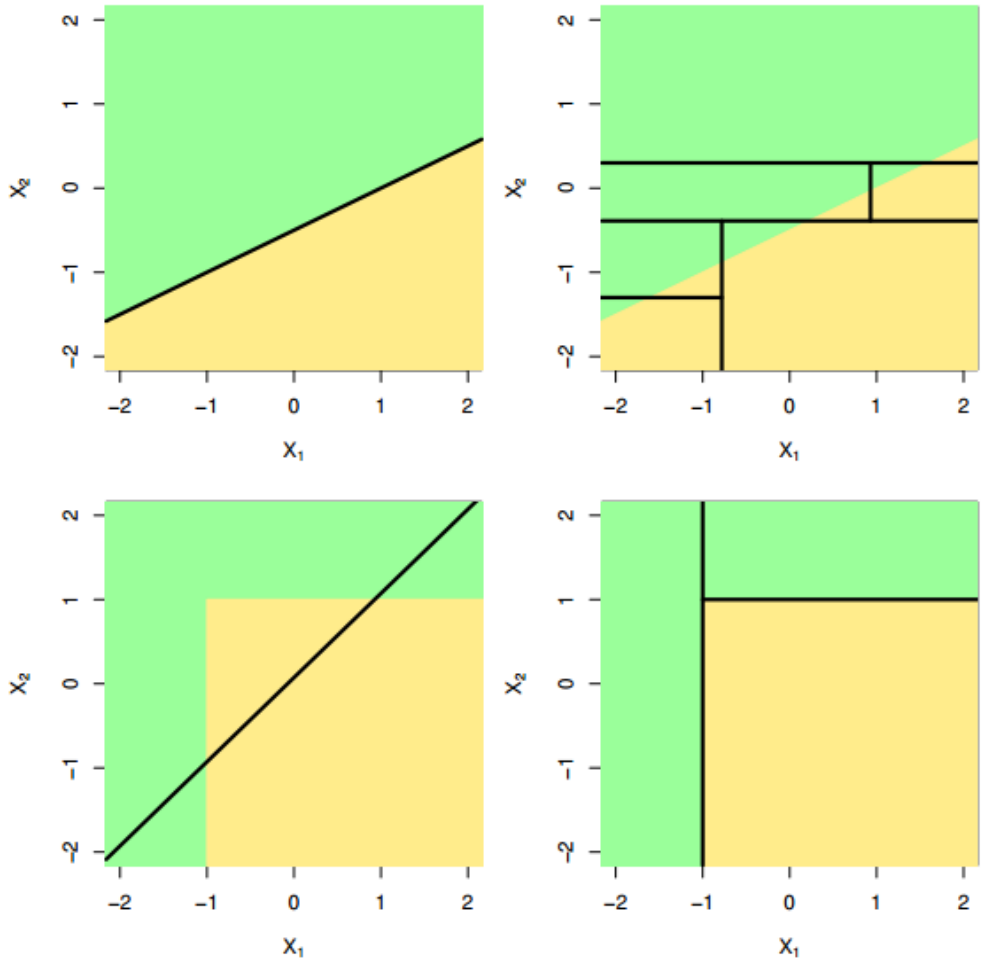
Conclude by formula

– which is better- larger or smaller G?

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}.$$

Cross entropy – similar to gini index.

Trees vs linear models

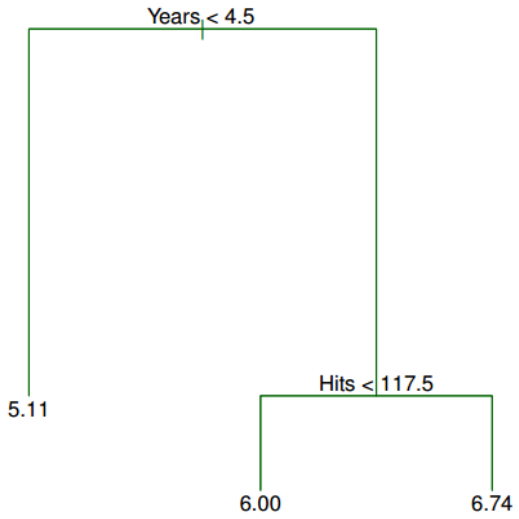


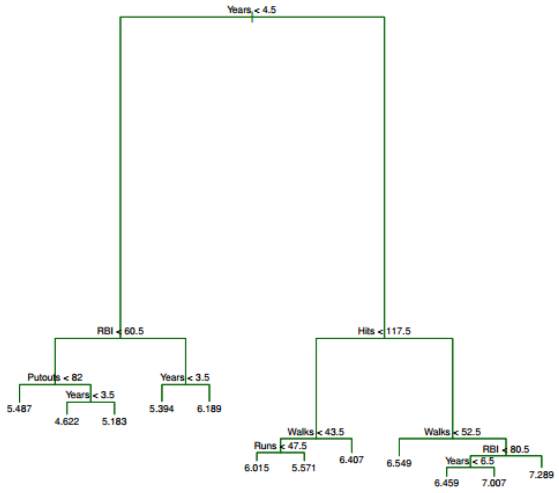
Advantages and Disadvantages

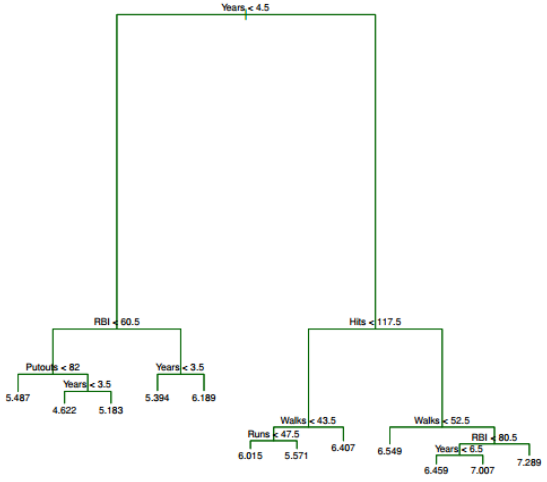
+ easy to explain
+ mirror human decision making
+ easy interpretation and display
+ easily handle qualitative predictors

Advantages and Disadvantages

-Not really as good as other regression and classification approaches







Bagging

- Averaging reduces the variance:

For a set of n independent sets of observations, Z1,…Zn, each with variance σ^2, the variance of the mean of the observations is σ^2/n

- If we had more than 1 training set, grow a tree on each of them and take average over trees (majority vote for classification)

Solution2: Bagging

bootstrap aggregation

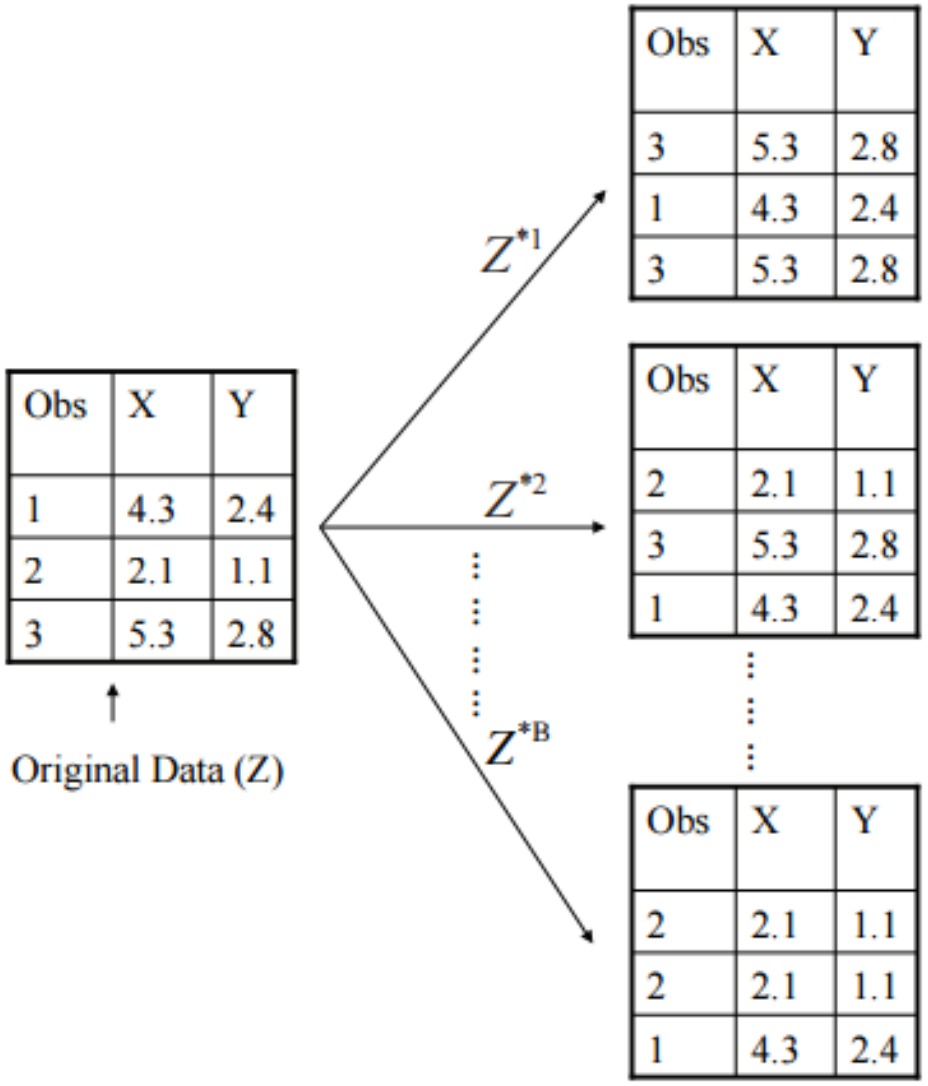
The bootstrap

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original
TRAINING set

Sample WITH REPLACEMENT from original training set to make "new" samples.

The bootstrap..

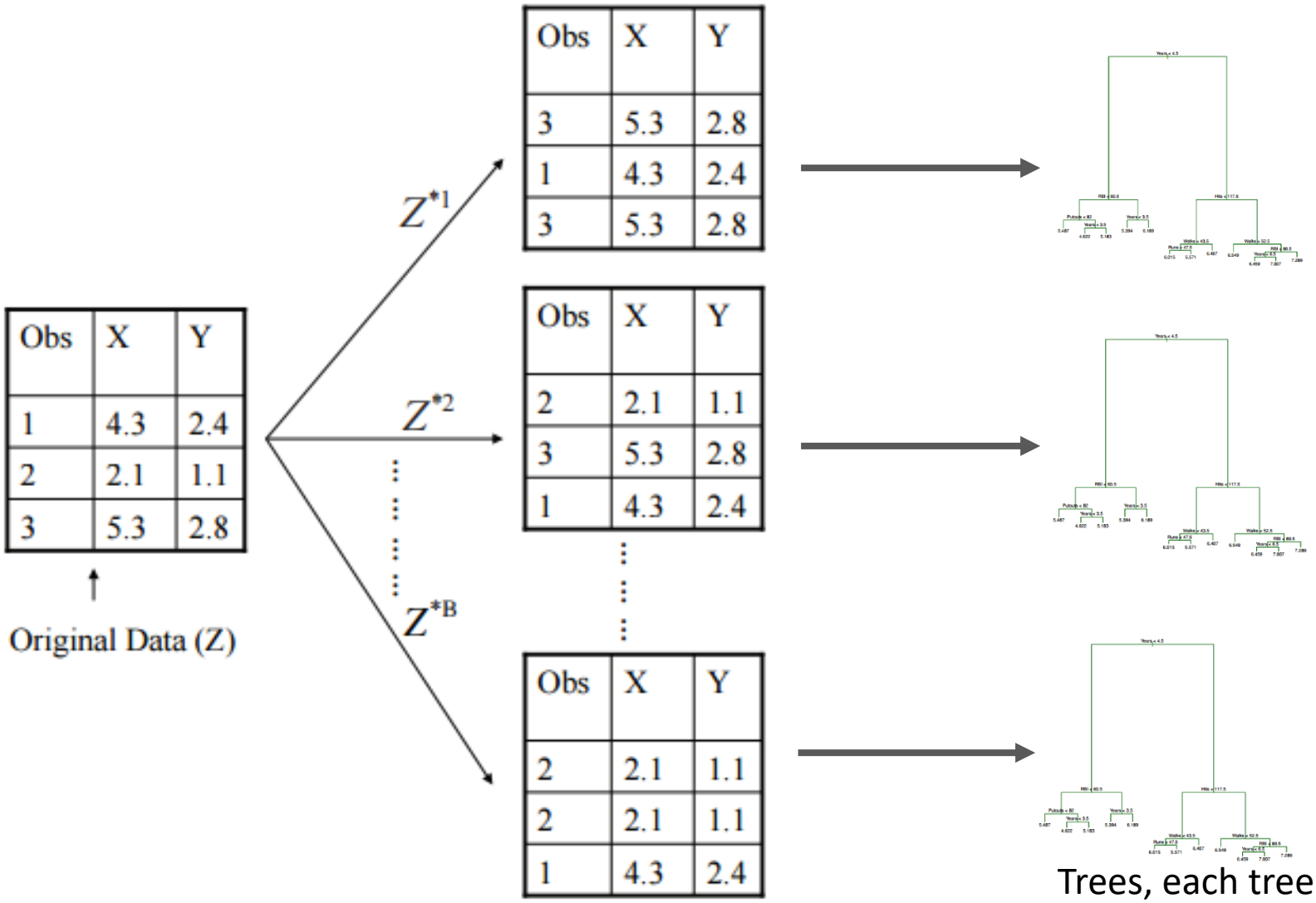


On average; between pairs 2/3 observations are the same

Original TRAINING set

Bootstrapped TRAINING sets

The bootstrap …

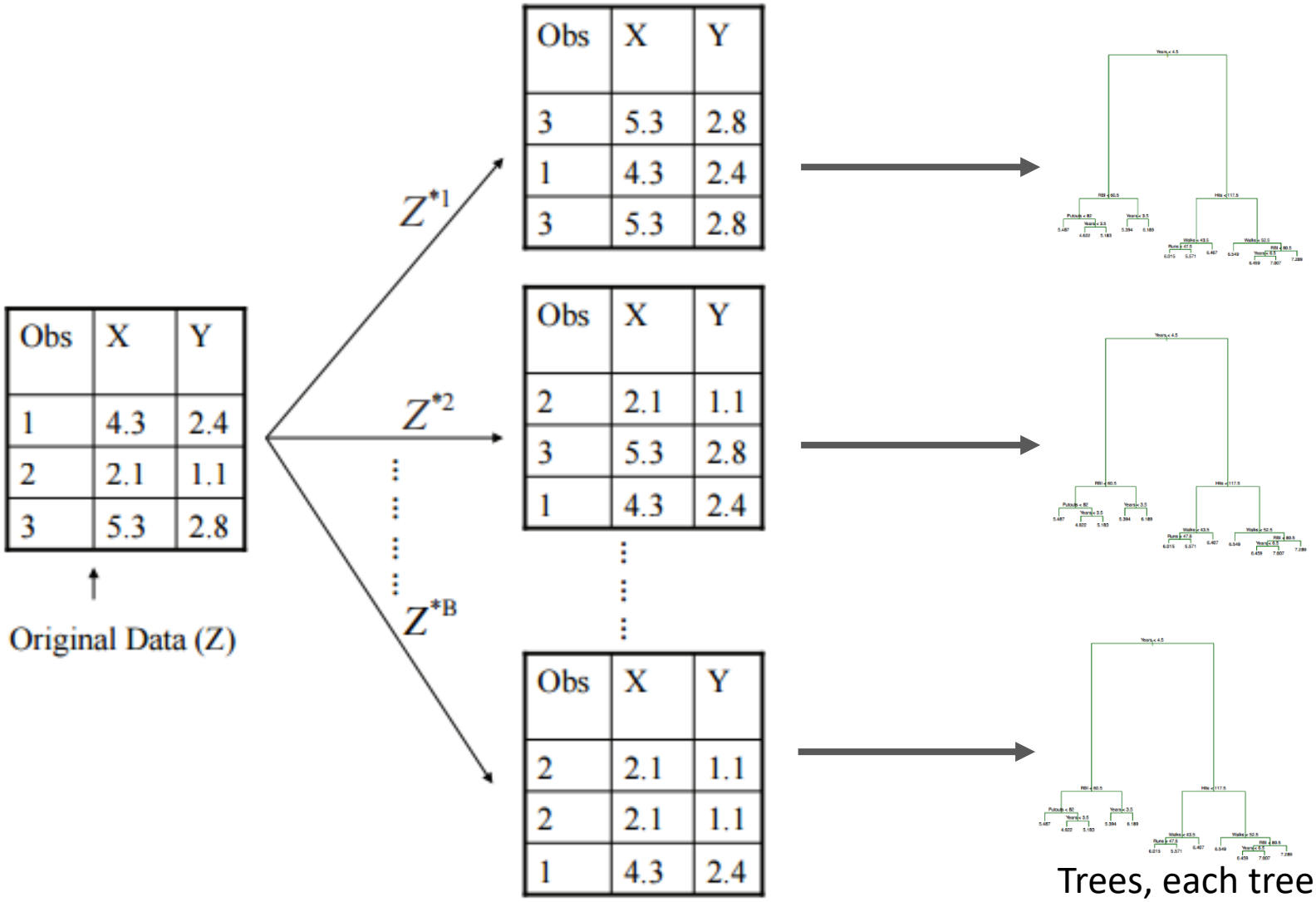


Original TRAINING set

Bootstrapped TRAINING sets

Trees, each tree built on one bootstrapped set

Great! (…right?)



Original TRAINING set

Bootstrapped TRAINING sets

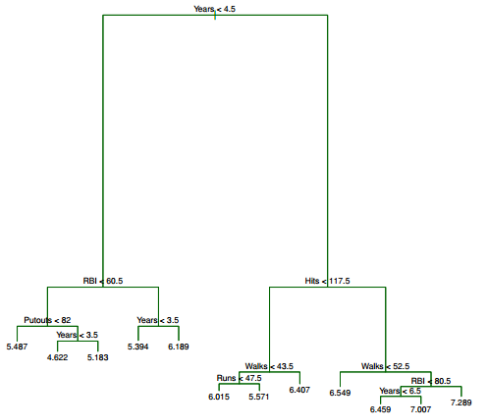
Trees, each tree built on one bootstrapped set

Predicting?

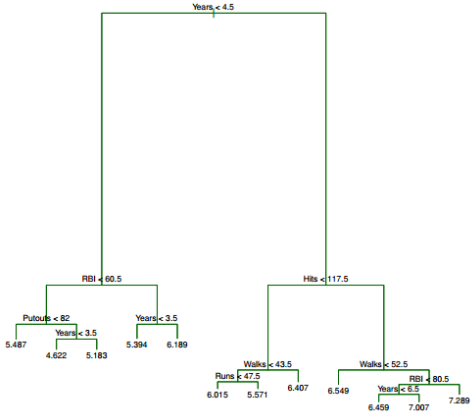
Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

new TEST set

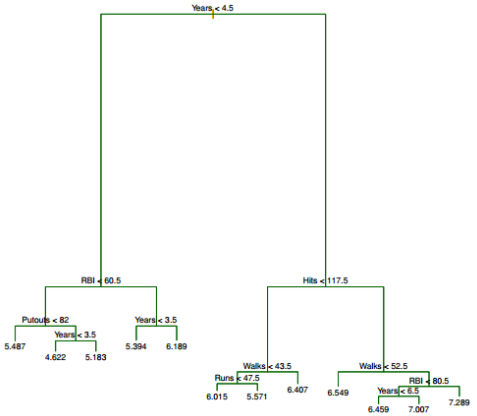
TEST row 1



5.2



6.1

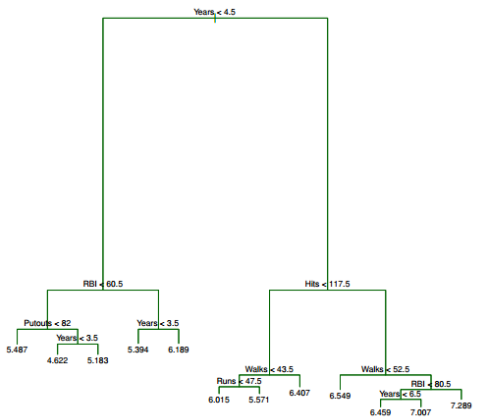


5.5

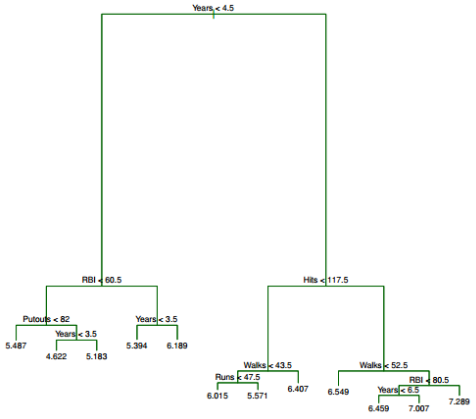
Predicting?

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

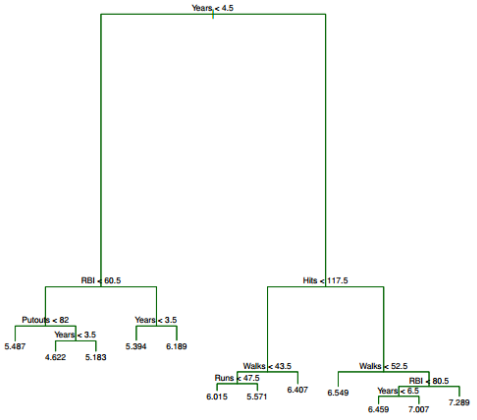
TEST row 1



5.2



6.1

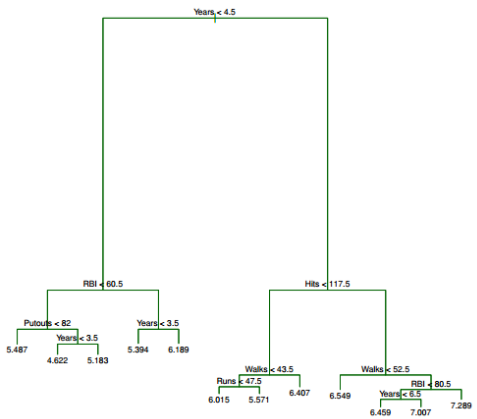


5.5

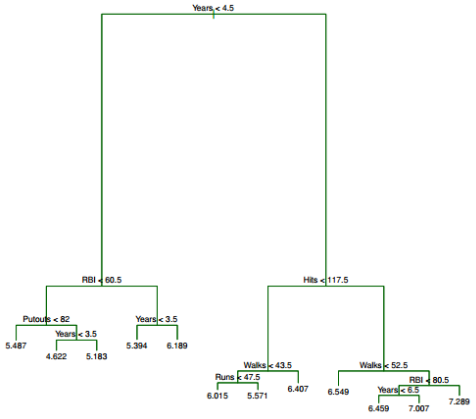
Predicting?

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

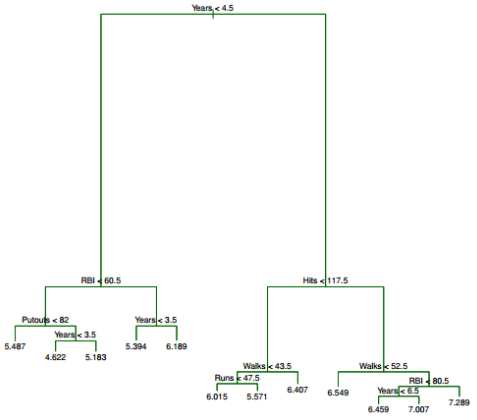
TEST row 1



5.2



6.1

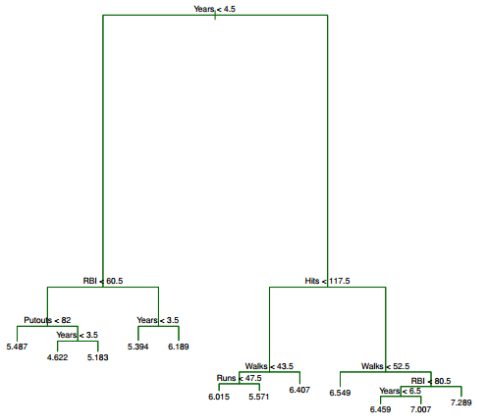


5.5

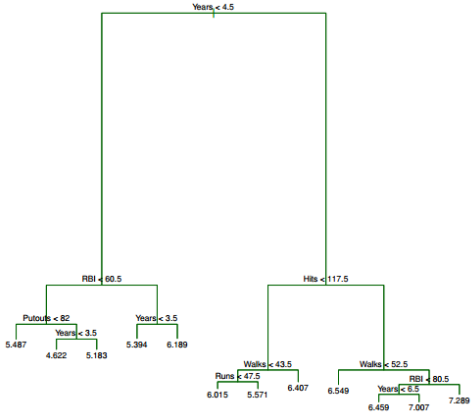
Predicting?

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

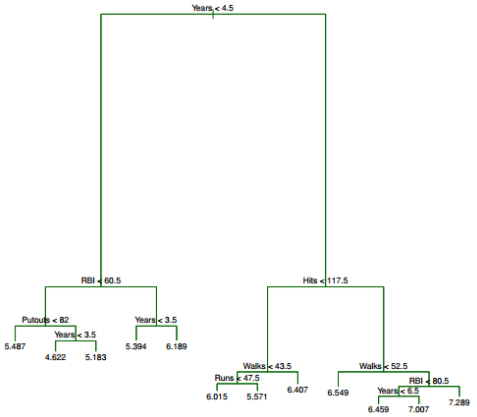
TEST row 1



5.2



6.1



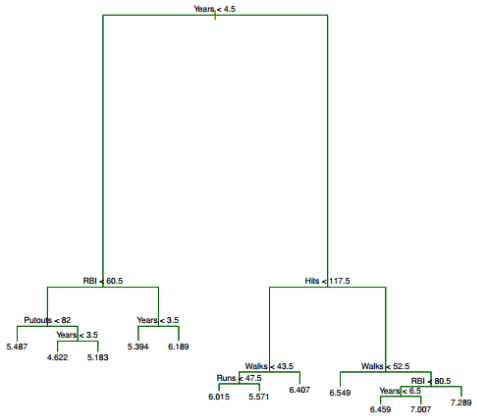
5.5

5.6

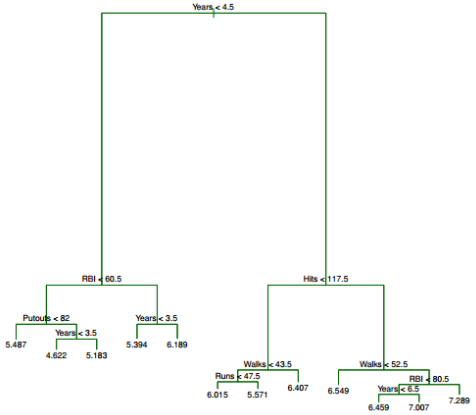
Predicting?

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

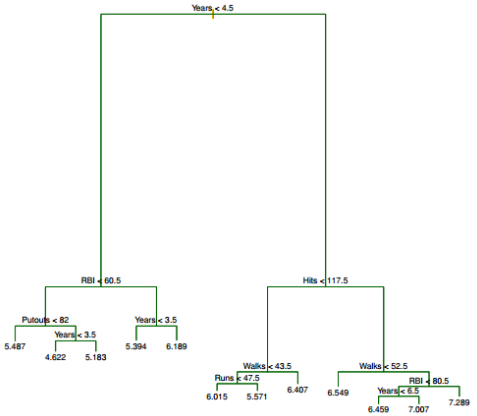
TEST row 2



6.0



4.7



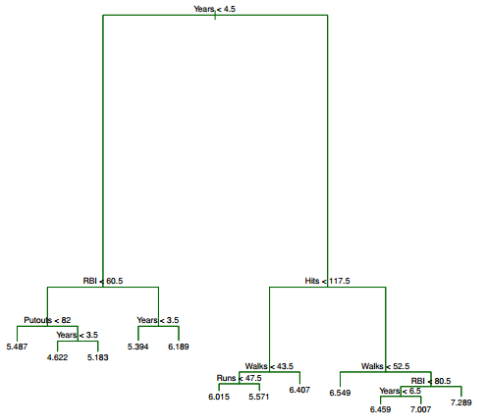
5.7

5.5

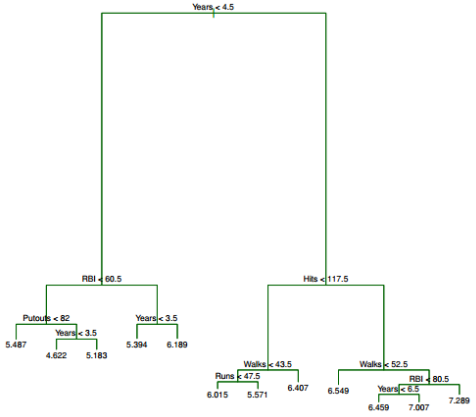
Predicting?

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

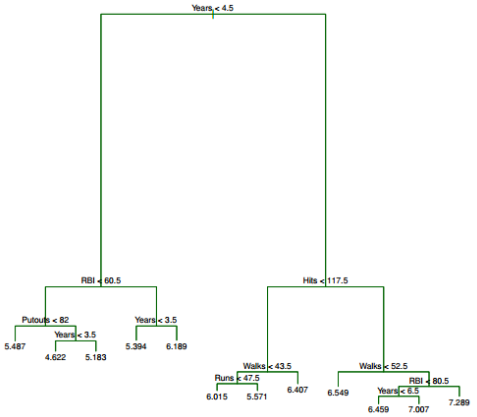
TEST row 3



7.0



4.7



5

5.6

Not independent observations…

2/3 of observations are shared between 2 resamples

Problem?

No! .. And yes.

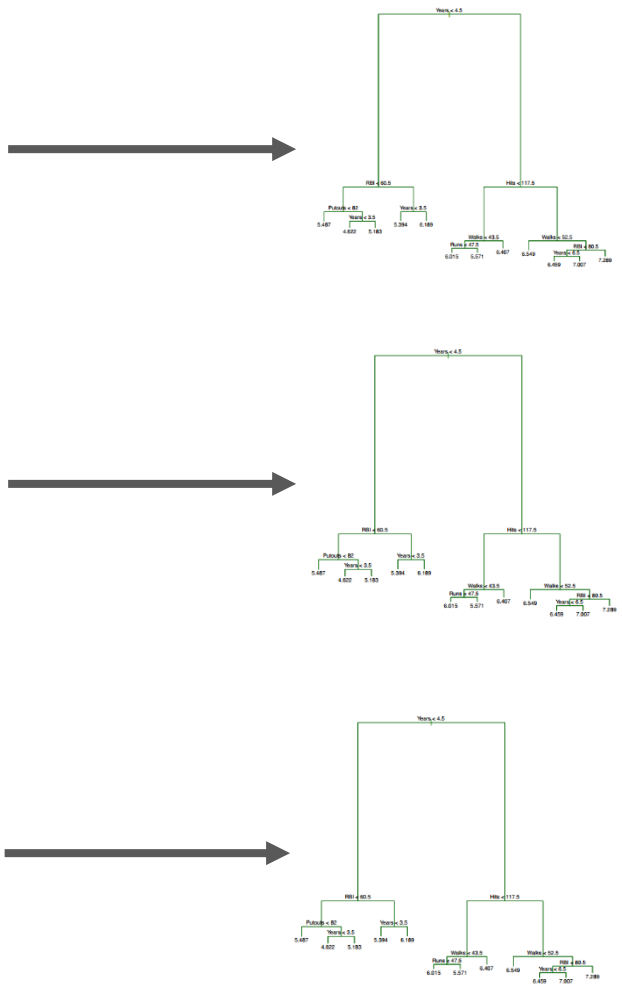
2/3 of observations are shared between each pair of samples

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

⋮

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4



Same variable is chosen as most important for splitting the data -> trees are very correlated

Bootstrapped TRAINING sets

Random Forests

Random Forests

- While taking averages, the error would be lower if correlation is lower (remember loocv vs cv)
- So do bagging but decorrelate the trees:

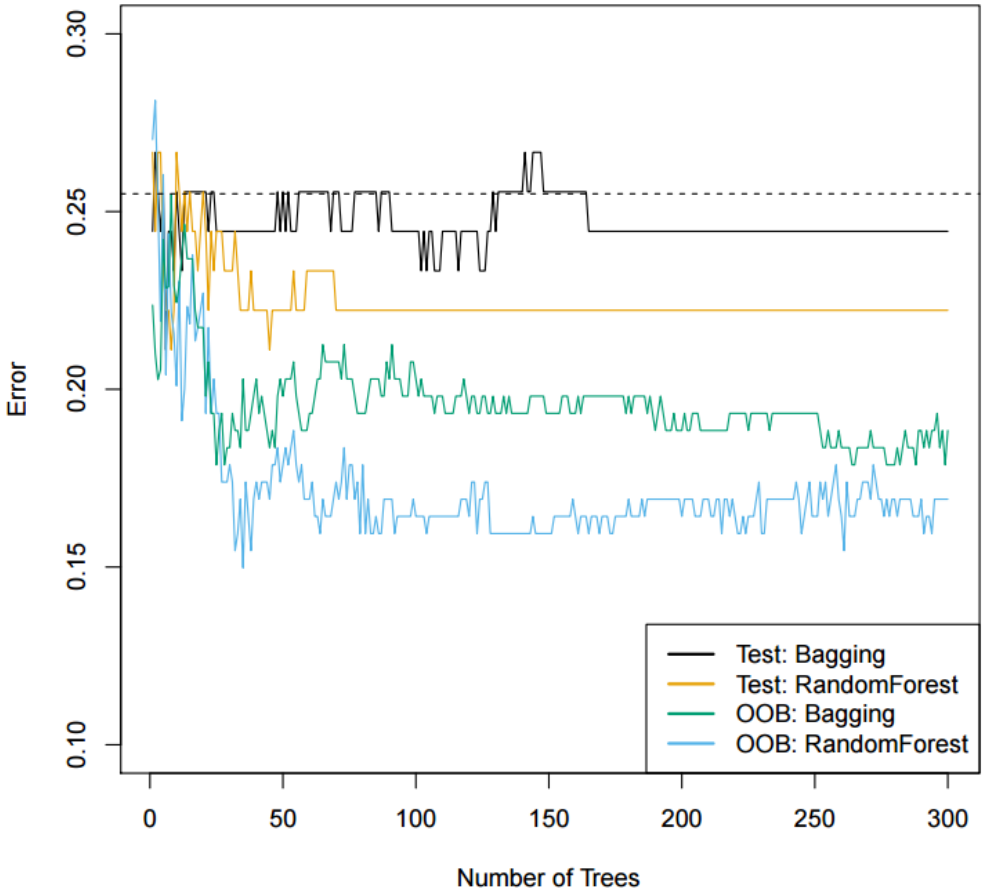
Randomly select m out of p possible predictors for each possible split!

$m \approx \sqrt{p}$ or $m \approx p/3$

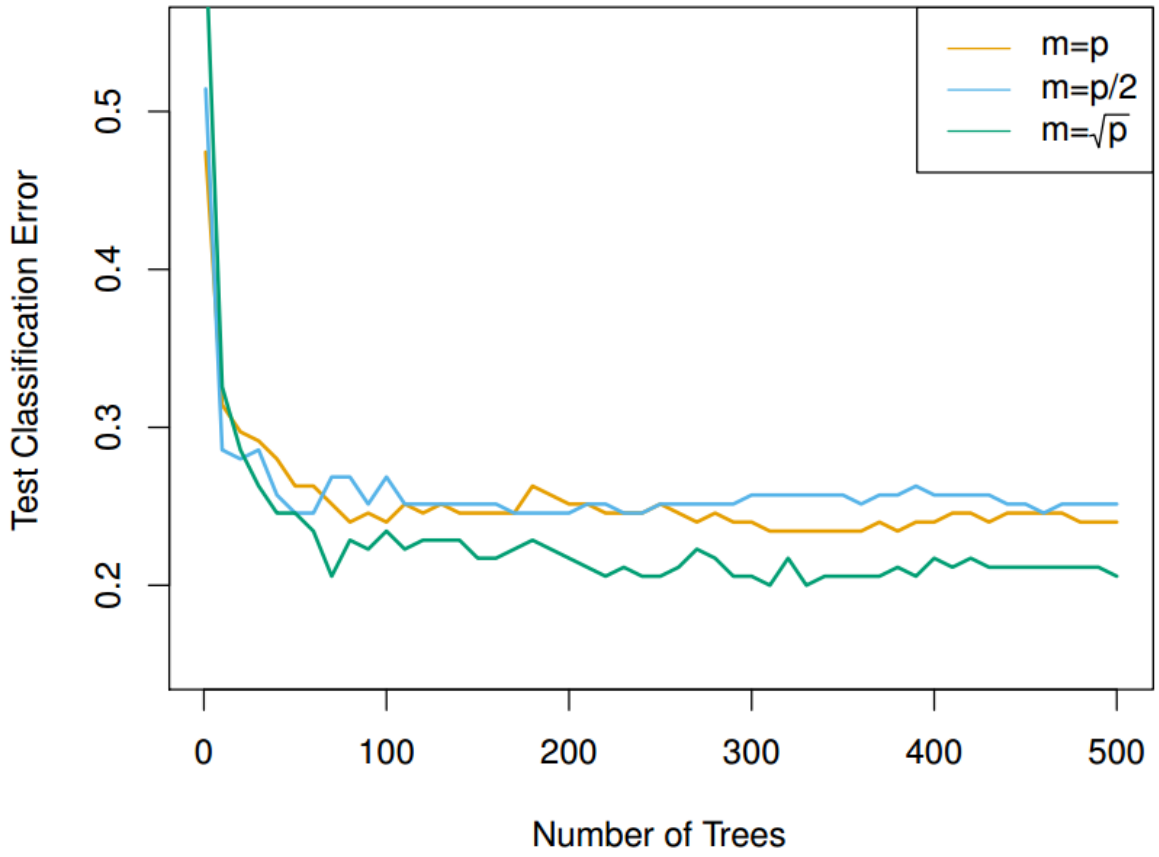
Out of Bag Error Estimation

- Straightforward way to estimate the error of a bagged model
- Make a tree using 2/3 observations, predict on the remaining 1/3= out of bag observations
- Average over trees

Prediction errors on Heart dataset with bagging and random forests

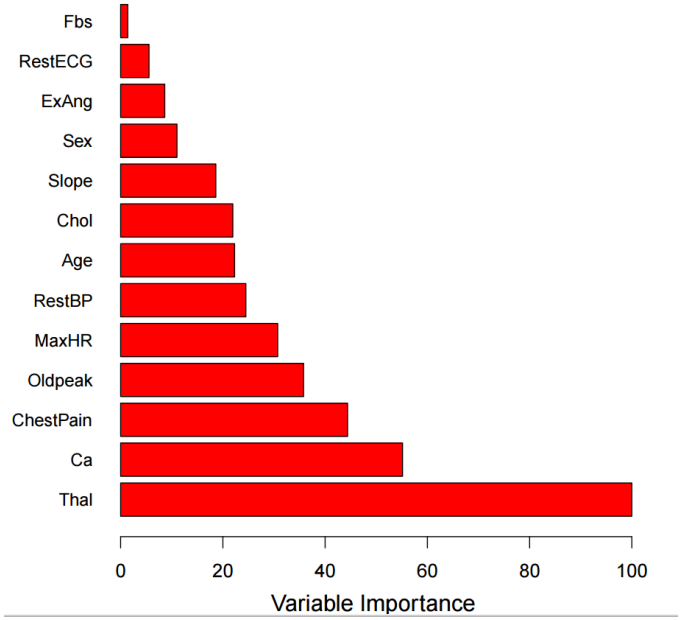


OOB errors for deciding on number of trees and predictors (m)



Variable importance measure

- For bagged / RF trees, we record the total amount that the RSS (Gini indeks for classification trees) is decreased due to splits over a given predictor, averaged over all B trees. (the higher the better)



Boosting

Boosting

- Trees are added sequentially to improve performance of the previous collection of trees

Boosting algorithm for regression trees

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
   - 2.1 Fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r).
   - 2.2 Update $\hat{f}$ by adding in a shrunken version of the new tree:

$$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$

   - 2.3 Update the residuals,

$$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$

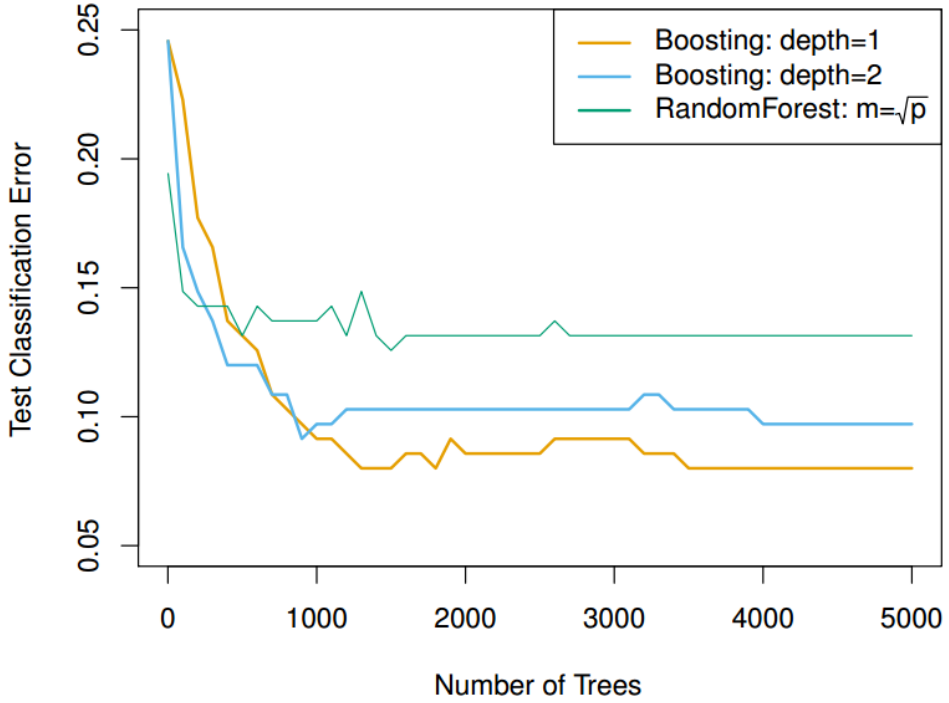
3. Output the boosted model,

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Boosting

- Fit very slowly so there is no overfitting
- Boosting can also be done for classification.

(details Elements of Statistical Learning, chapter 10)



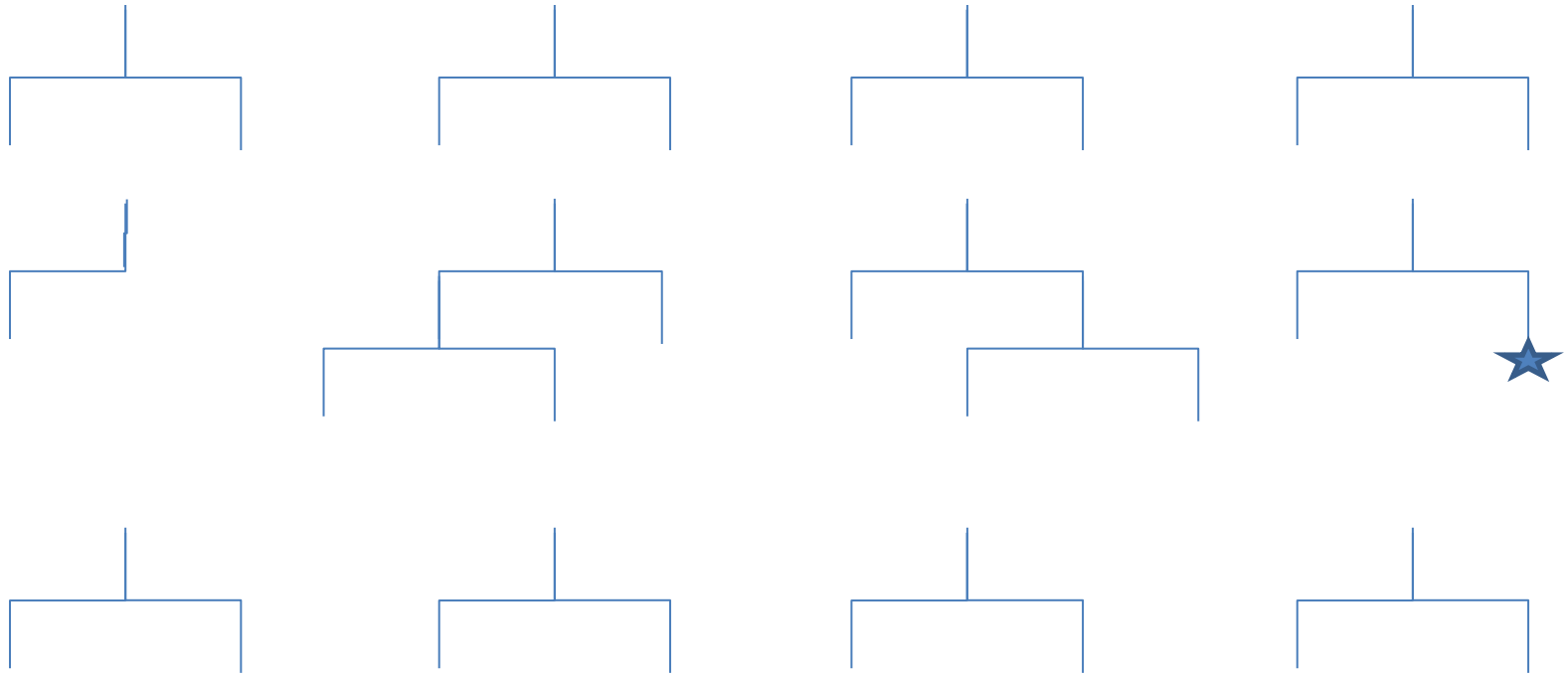
Tuning parameters for boosting

1. The number of trees B. Unlike bagging and random forests, boosting can overfit if B is too large, although this overfitting tends to occur slowly if at all. We use cross-validation to select B.
2. The shrinkage parameter λ, a small positive number. This controls the rate at which boosting learns. Typical values are 0.01 or 0.001, and the right choice can depend on the problem. Very small λ can require using a very large value of B in order to achieve good performance.
3. The number of splits d in each tree, which controls the complexity of the boosted ensemble. Often d = 1 works well, in which case each tree is a stump, consisting of a single split and resulting in an additive model. More generally d is the interaction depth, and controls the interaction order of the boosted model, since d splits can involve at most d variables.

BART
Bayesian additive regression trees

BART

- Prior instead of lambda (depth, y^, sd of the residual noise)



What to do

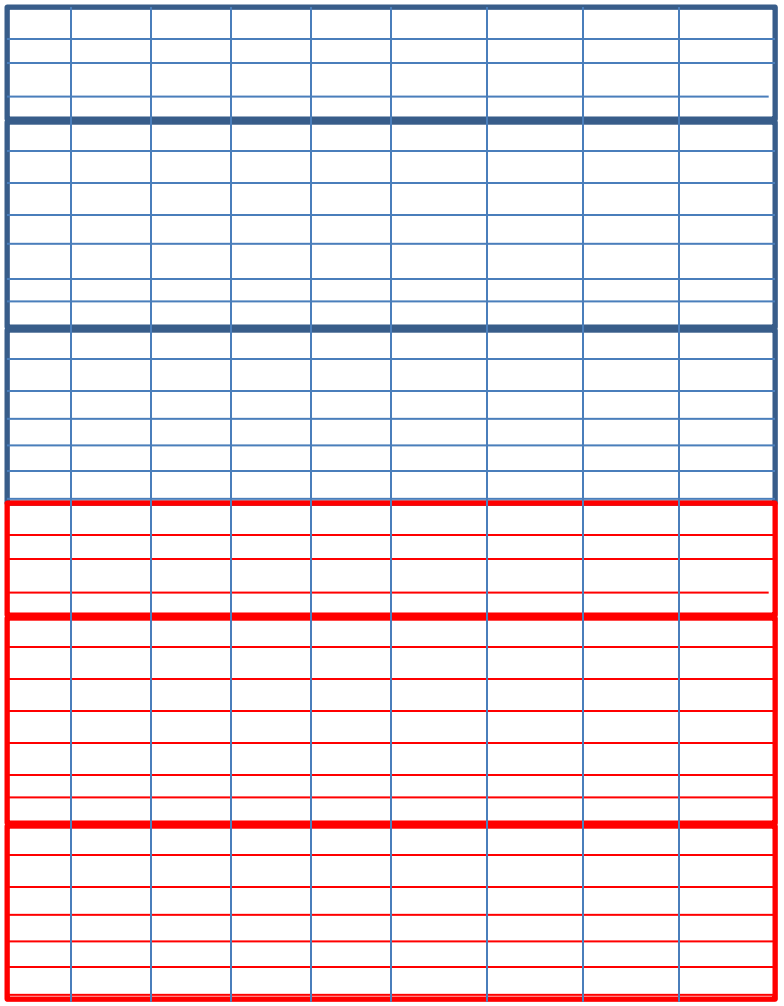
(and what not)

Example

- Large data set: 630730 genes in patients with liver cirrhosis and 578064 genes in healthy people.
- Around 20 variables extrapolated from our data set: M and A values for some statistic, on different levels of enrichment
- Cross validated OOB error: 10,5%
- What do we expect our test error to be?

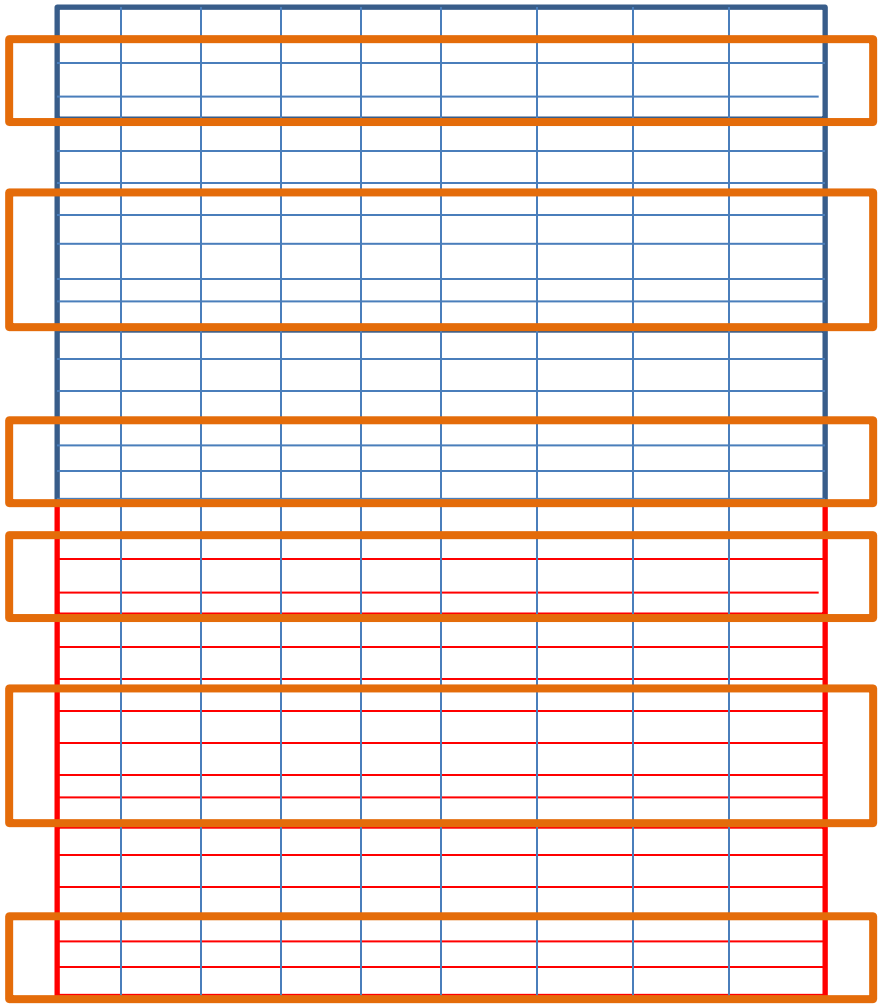
Example

- Large data set: 630730 genes in patients with liver cirrhosis and 578064 genes in healthy people.
- Some variables calculated based on each patient separately



Example

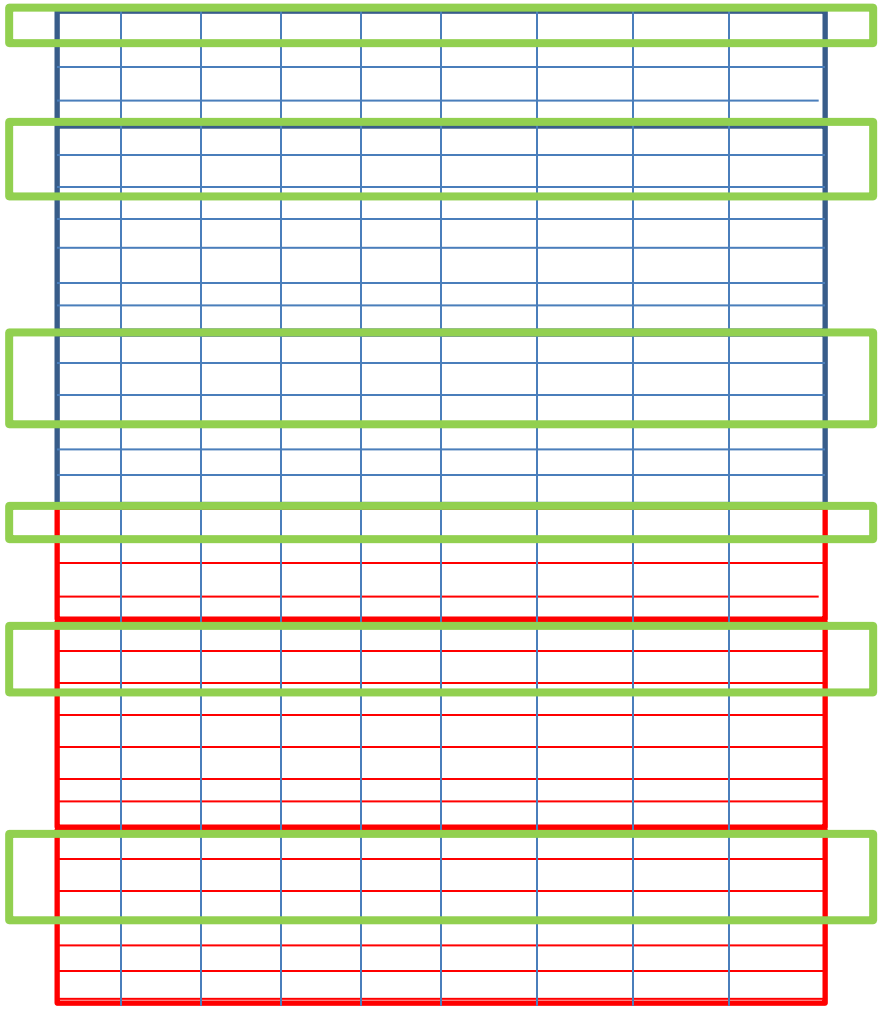
- Large data set: 630730 genes in patients with liver cirrhosis and 578064 genes in healthy people.
- Some variables calculated based on each patient separately



Example

test

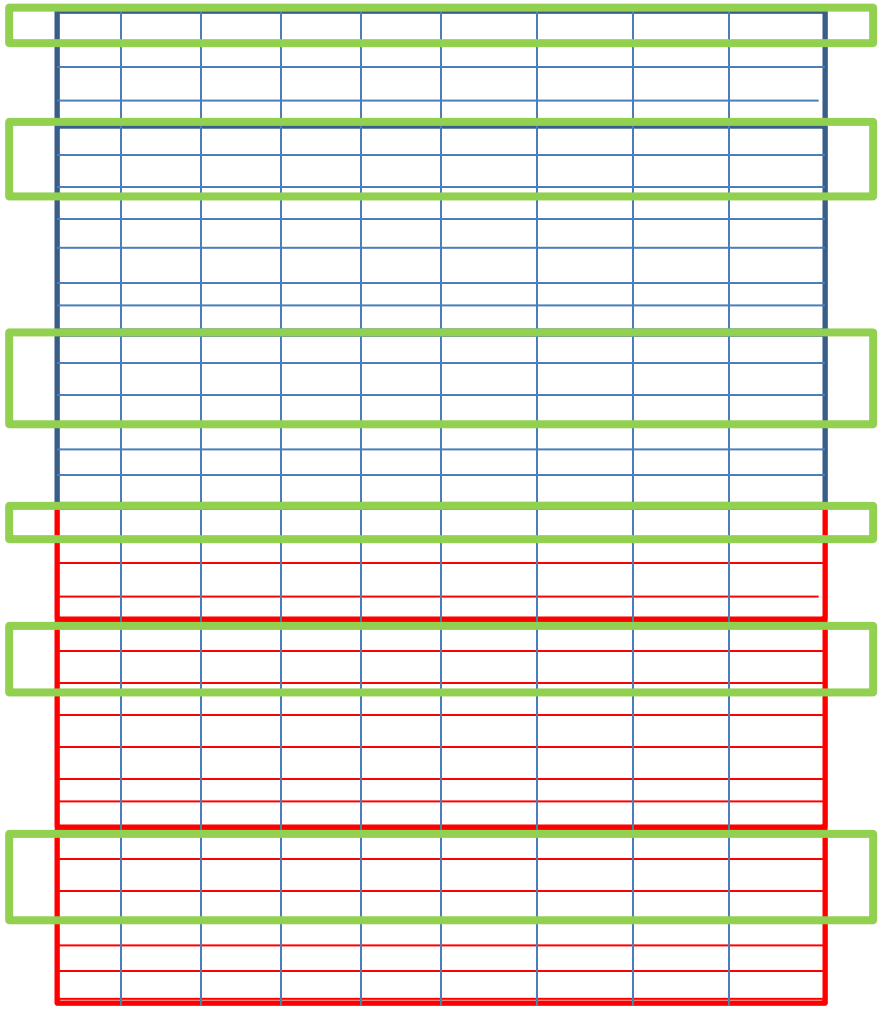
- Large data set: 630730 genes in patients with liver cirrhosis and 578064 genes in healthy people.
- Some variables calculated based on each patient separately



Example

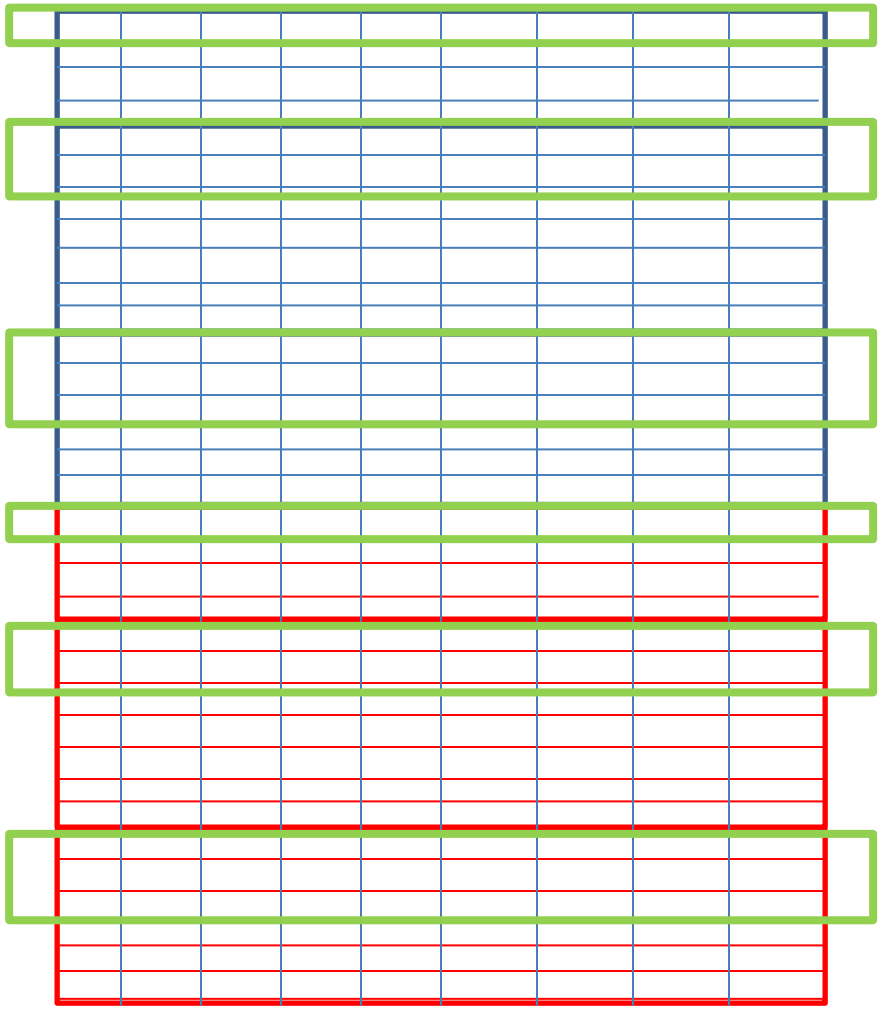
test

- Large data set: 630730 genes in patients with liver cirrhosis and 578064 genes in healthy people.
- Some variables calculated based on each patient separately

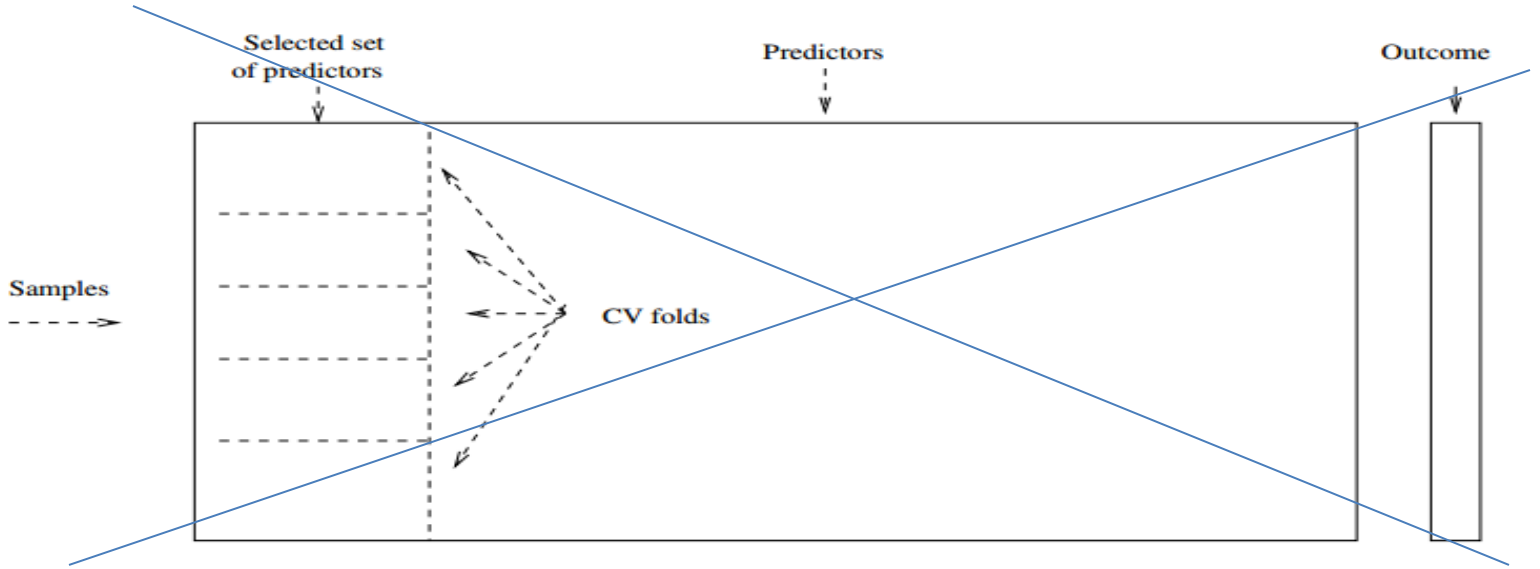


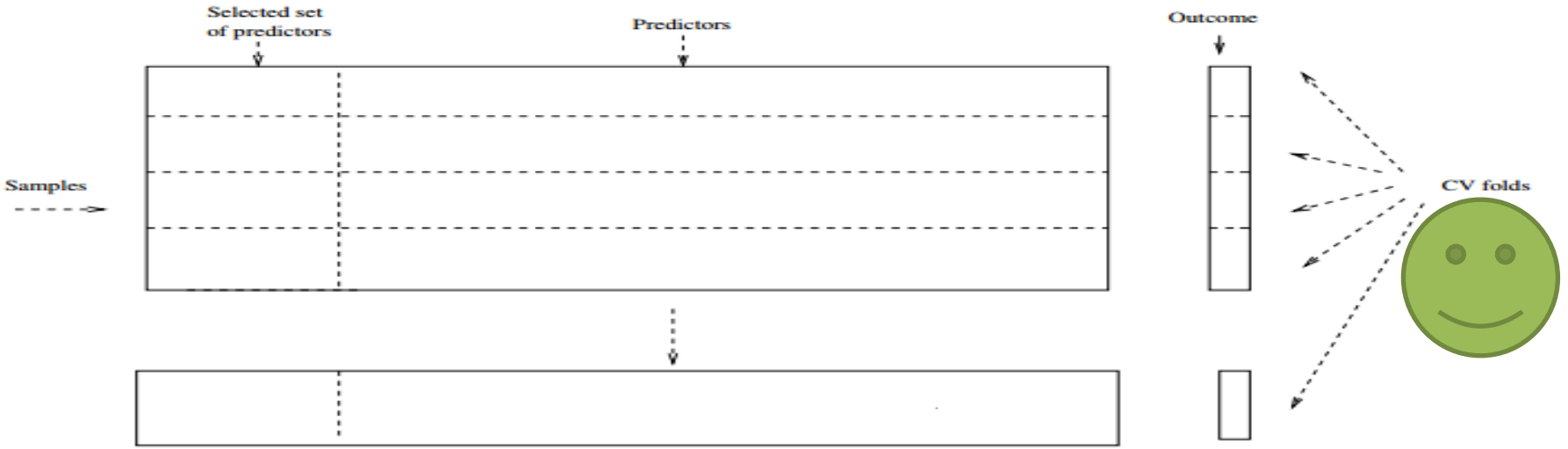
Example

- OOB error 10.5%
- What error do we expect on test set?



What NOT to do

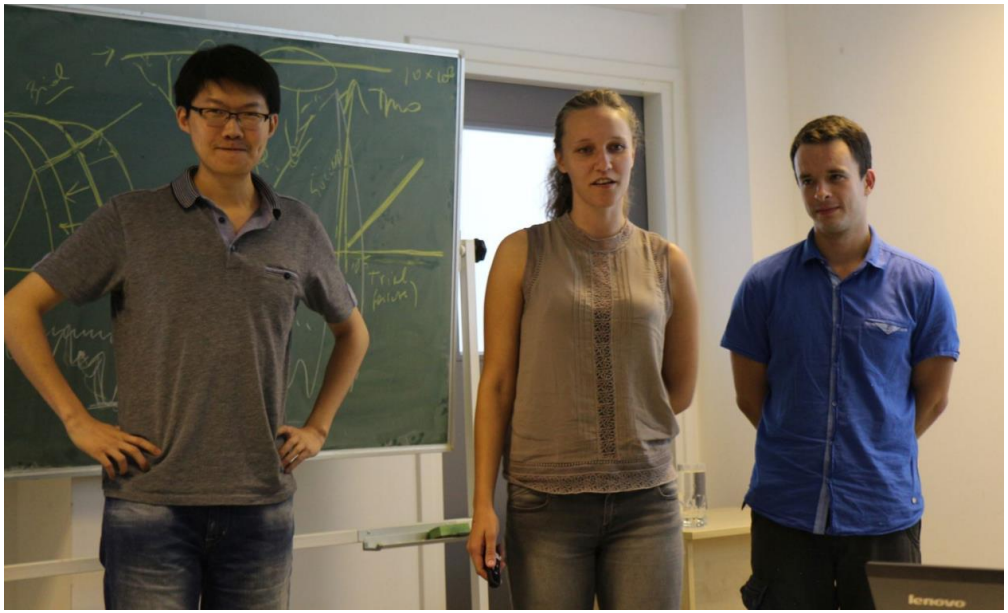




Example2: Choosing from predictor space

- We apply random forests to a high dimensional biological data set containing expression measurements of 4718 genes measured on tissue samples from 349 patients.
- Each sample is either normal or labeled as 1 of 14 types of cancer.
- We choose set of 500 predictors that have the highest variance in the training set and build random forest to predict cancer type.
- **Are we cheating?**

RSSSO 2015 Random Forest Team



Zheng Ning
German Demidov
Maja Fabijanić

Project leaders:
dr. Julia Dimitrieva
(+ husband dr. Dmitri Filippov)
mr. Alexander Kurilshikov
dr. Olga Zaitseva

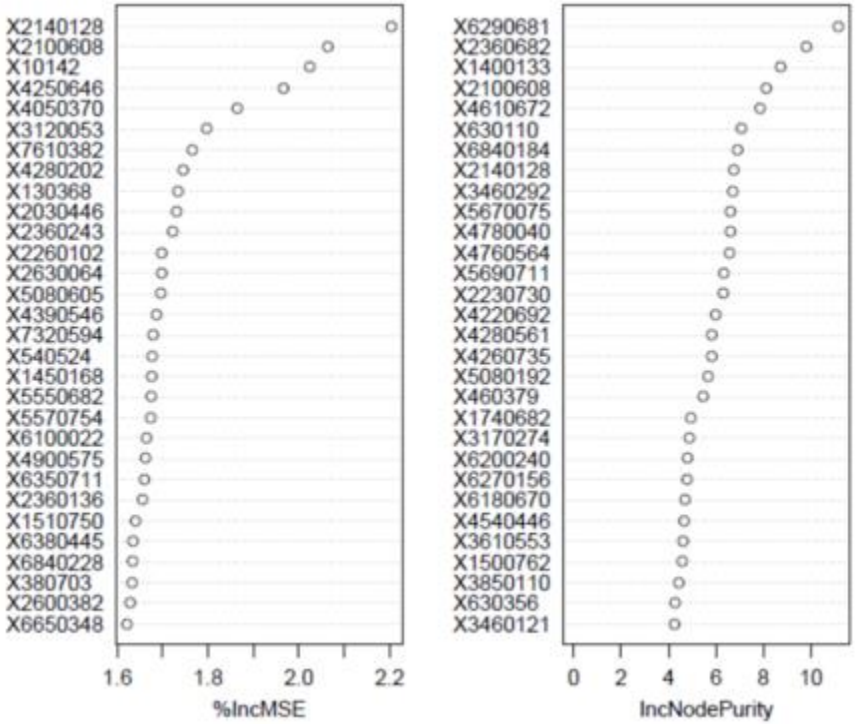
The problem

- Expression data for 20 tissues (~ 10 000 probes):
  - **Illeum**
  - CD4+
  - CD8+
  - …
- Select important probes to predict:
  - **BMI**
  - Risk Scores for Ulcerative colitis
  - Risk Scores for Chron's disease

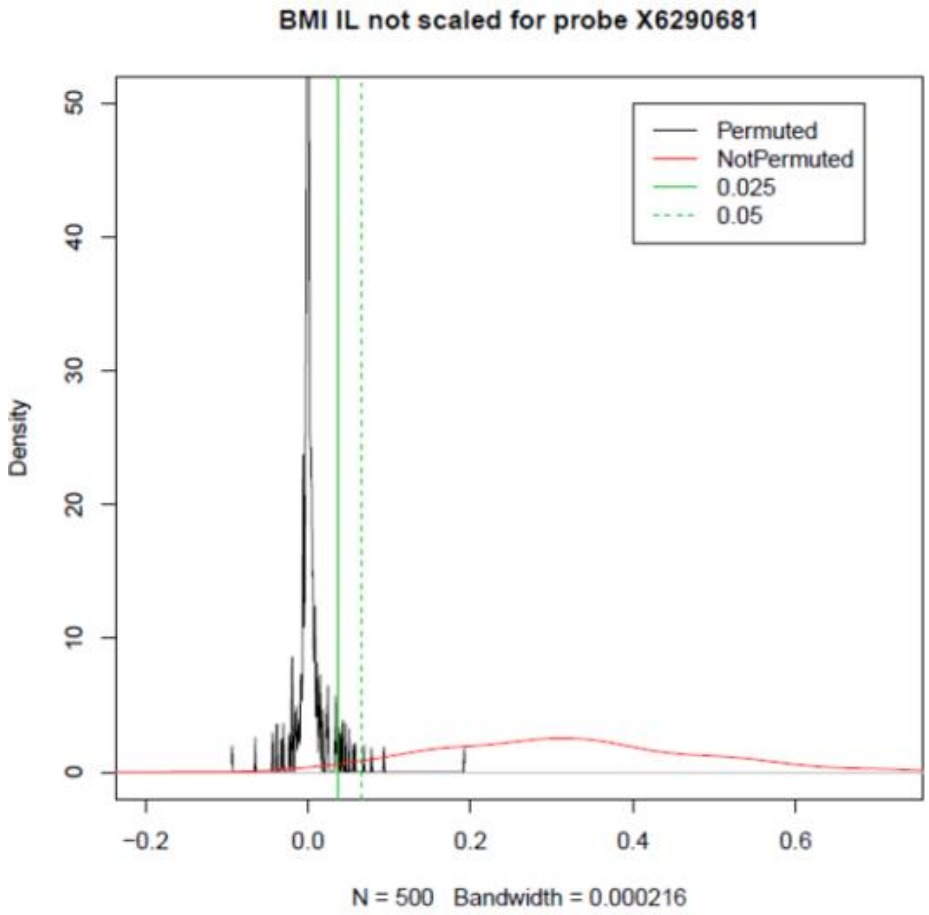
The approach: LR, RF, MIC

Permutation test for variable importance

For 1 forest:



Permutation test for variable importance



Oops

- Different variables each time!?
    - m for classification is 1/3*p, not sqrt(p)
- Too slow RF (~20 minutes for 1)
    - RF formula!
    - Packages!

Ileum - BMI: RF top20 + LR top20 intersection

	probes	pval.x	padj.x	genenames.x	slope	pval.LR	padj.LR
1	X110719	0.052	0.9855035	SLC16A3	3.257902	2.066555e-05	0.05778933
2	X4480341	0.006	0.9855035	DHCR24	4.066691	2.523919e-05	0.05778933
3	X6290681	0.002	0.9855035	TOMM20	-6.020913	9.889393e-06	0.05778933

	genenames.y	pval.y	padj.y	genenames
1	SLC16A3	0.306	0.9566392	SLC16A3
2	DHCR24	0.285	0.9566392	DHCR24
3	TOMM20	0.024	0.9566392	TOMM20

Conclusions

- Be careful when selecting variables not to introduce bias (Don’t leak data)
- Do a valid cross validation
- Check your results (if they are too good to be true, they probably aren’t true)
- Try a simpler approach first
- You WILL make mistakes.
- Use google. READ THE MANUAL.

Thank you for attention.