Workshop

Decision tree and random forests in R

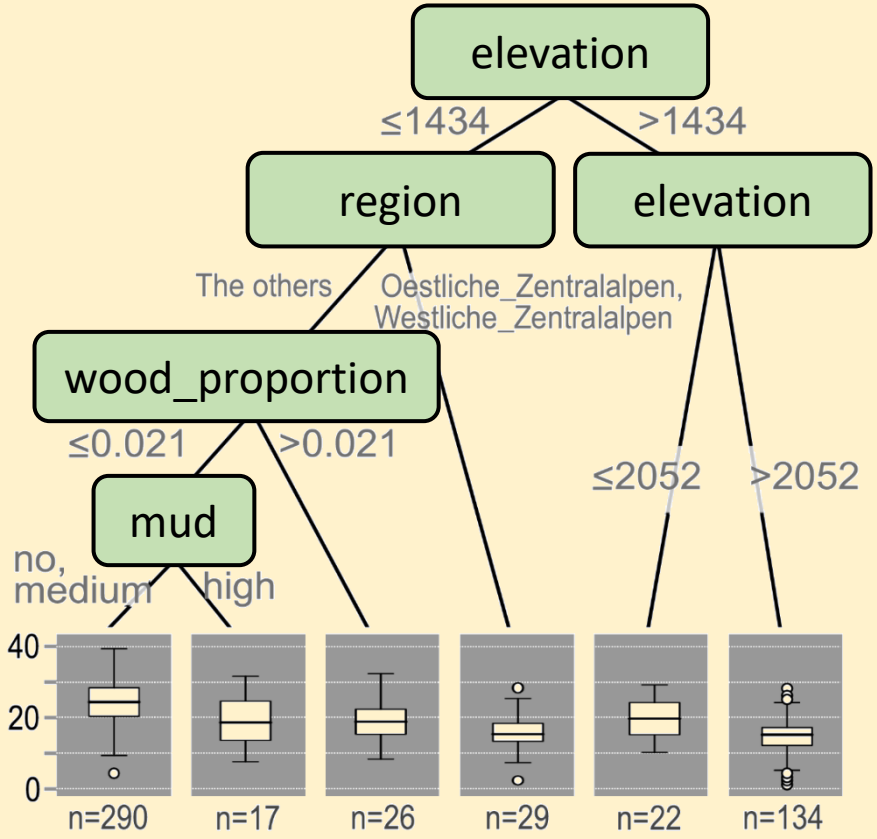
Masahiro Ryo

Free University of Berlin

Berlin-Brandenburg Institute of Advanced Biodiversity Research

Decision tree Classification and Regression Tree (CART)

(Breiman et al. 1984)

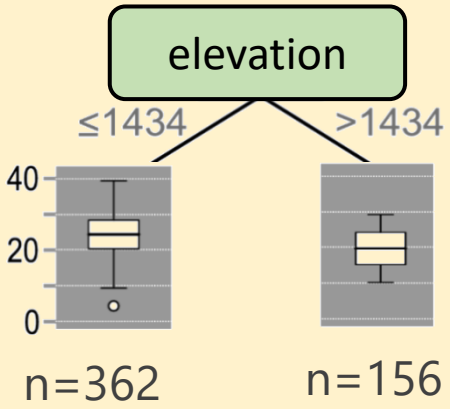


Points

1) No need for *a priori* selection of data & statistical assumptions (just run)
2) Missing values allowed
3) Nonlinearity
4) Indication for variable interactions

Decision tree Classification and Regression Tree (CART)

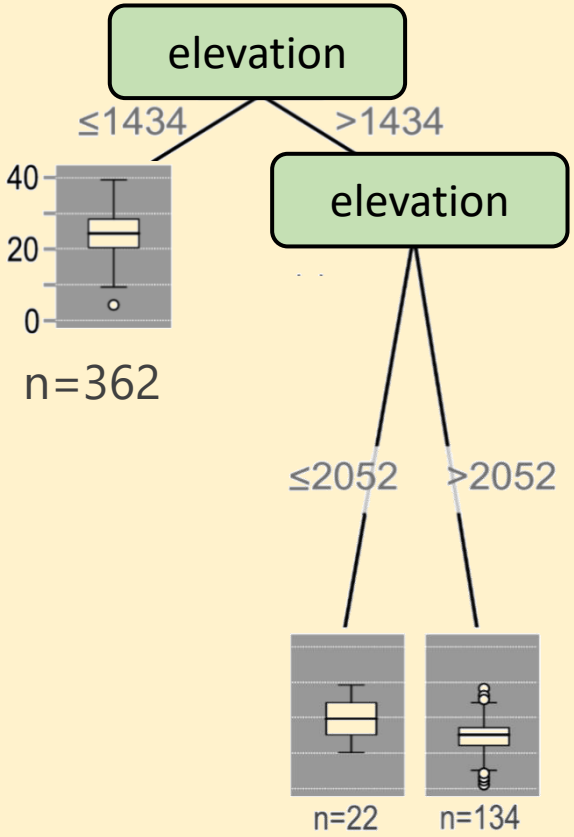
(Breiman et al. 1984)



1) Find a predictor & a threshold value which separate the data into two the most distinctively.
   ⇒ *If "elevation" is > 1434 m or not*

Decision tree Classification and Regression Tree (CART)

(Breiman et al. 1984)

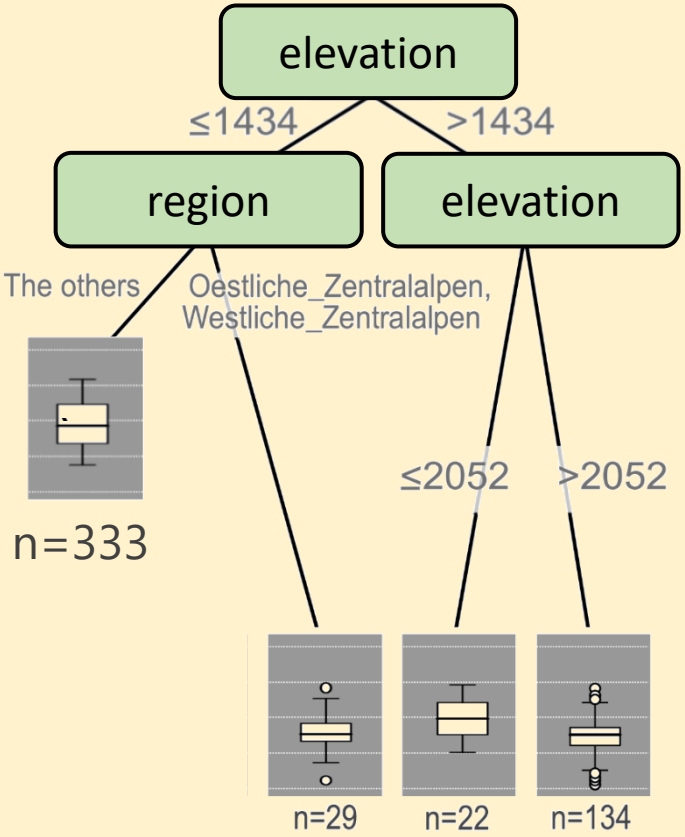


1) Find a predictor & a threshold value which separate the data into two the most distinctively.
   ⇒ *If "elevation" is > 1434 m or not*

2) For each of the separated data, repeat.
   ⇒ *If "elevation" is > 2052 m or not (for right-hand side)*

Decision tree Classification and Regression Tree (CART)

(Breiman et al. 1984)



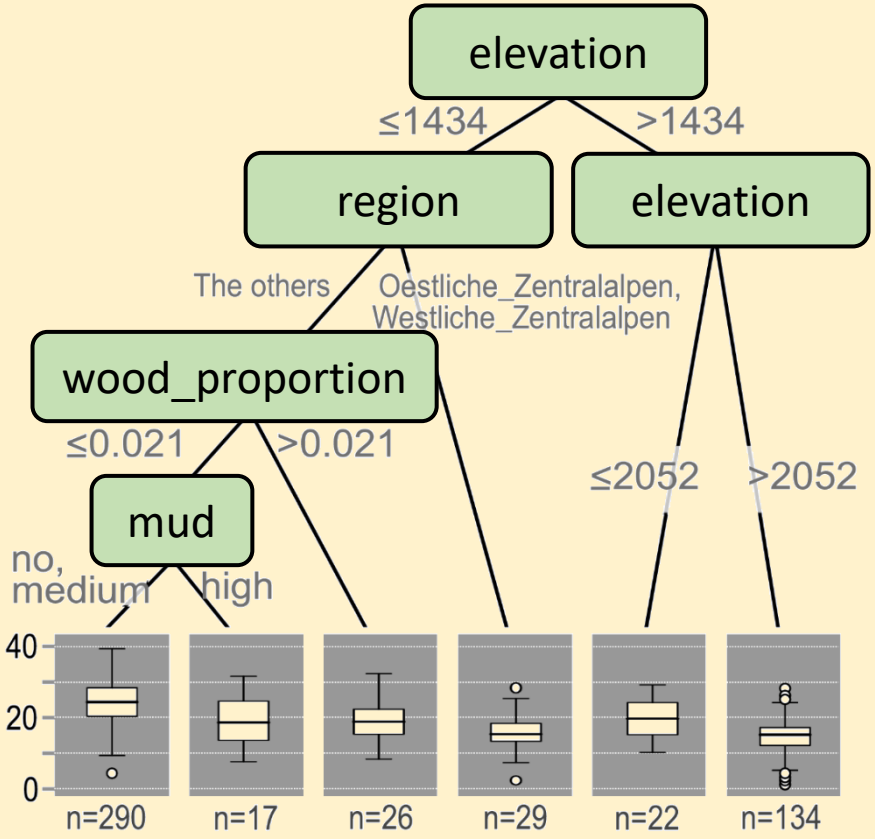
1) Find a predictor & a threshold value which separate the data into two the most distinctively.
   ⇒ *If "elevation" is > 1434 m or not*

2) For each of the separated data, repeat.
   ⇒ *If "elevation" is > 2052 m or not (for right-hand side)*

3) Stop separation when a set of rules are achieved (i.e. no more improvement).

Decision tree Classification and Regression Tree (CART)

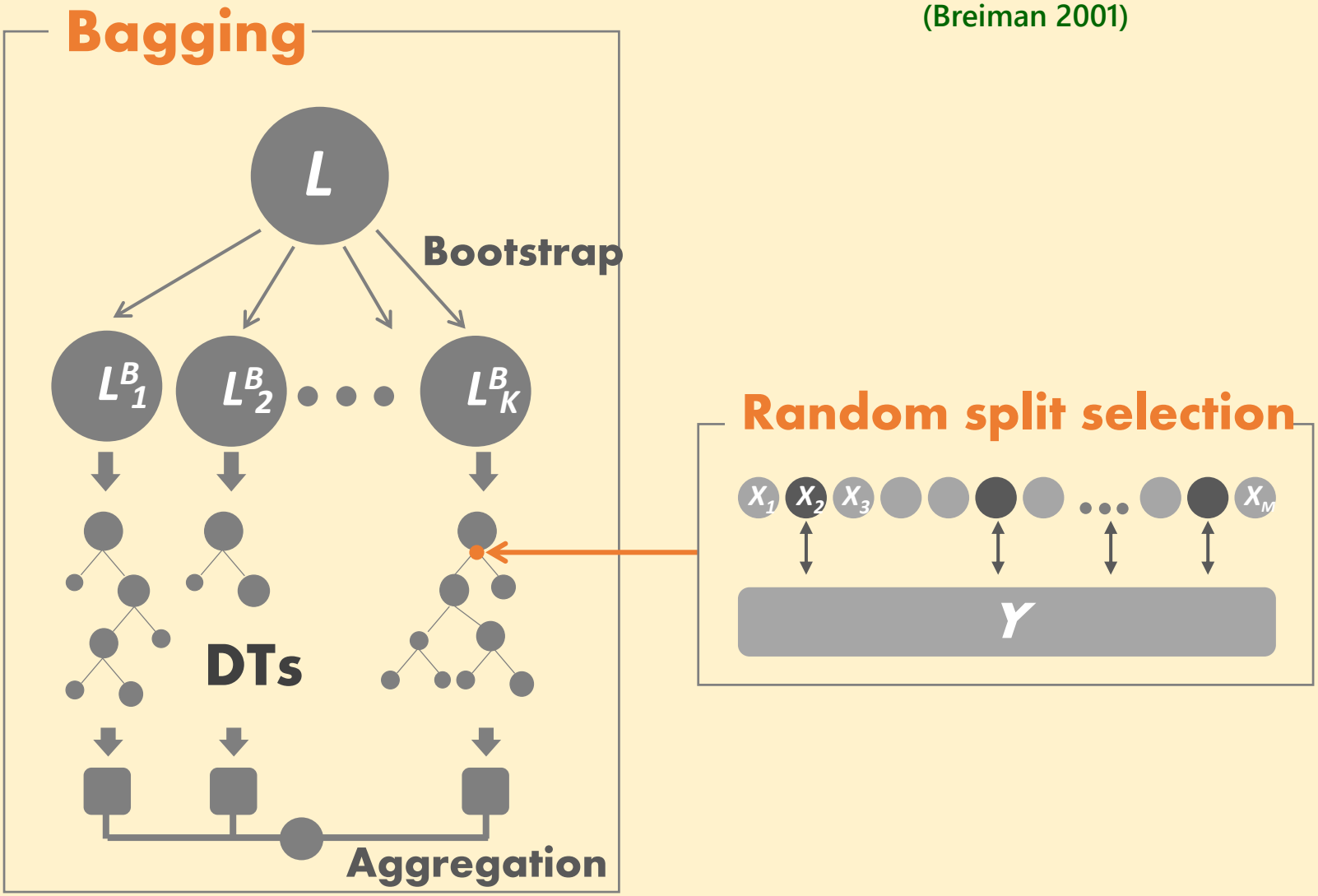
(Breiman et al. 1984)



From decision tree to random forests

Random forests model ensemble approach

(Breiman 2001)

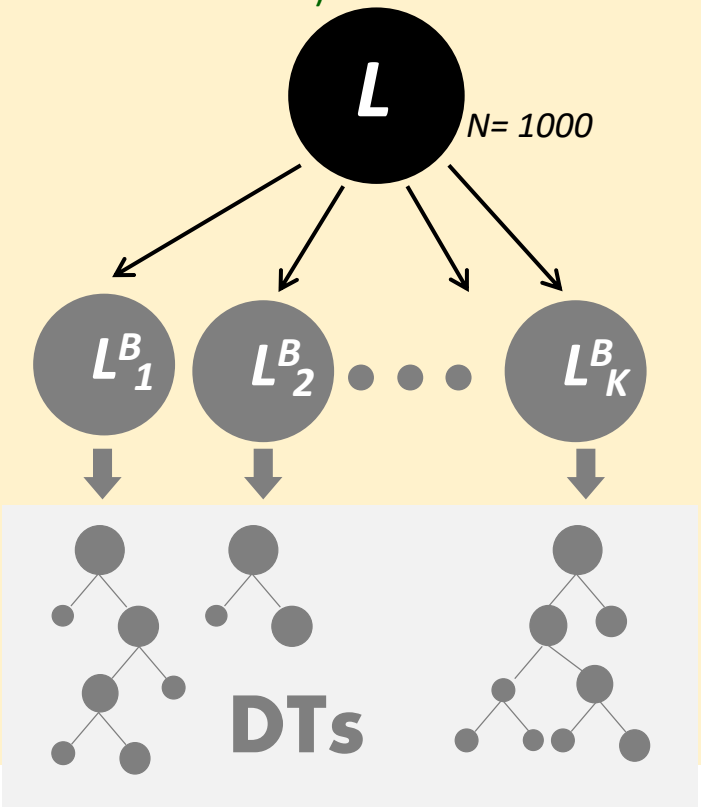


Random forests model ensemble approach

(Breiman 2001)

Bootstrap

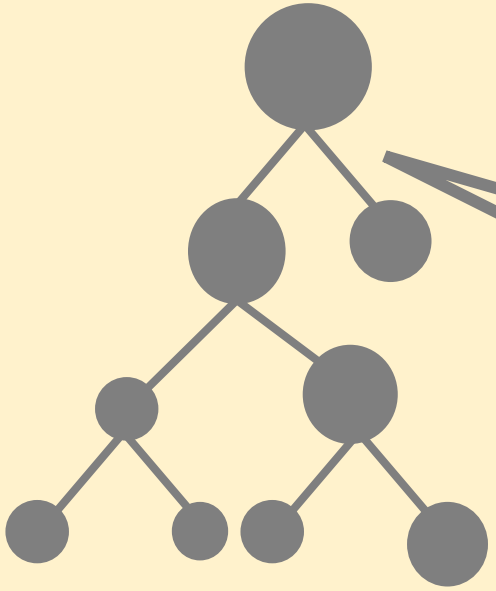
- To generate many replicates L^B from the original dataset L
- Each consisting of N cases, drawn at **RANDOM**, but with replacement (ca. **63.2%** of the original data is chosen)



Efron and Tibshirani, 1993

Random forests model ensemble approach

(Breiman 2001)



At each node…

- Do not compare all predictor variables
- But RANDOMLY pick up some and then compare

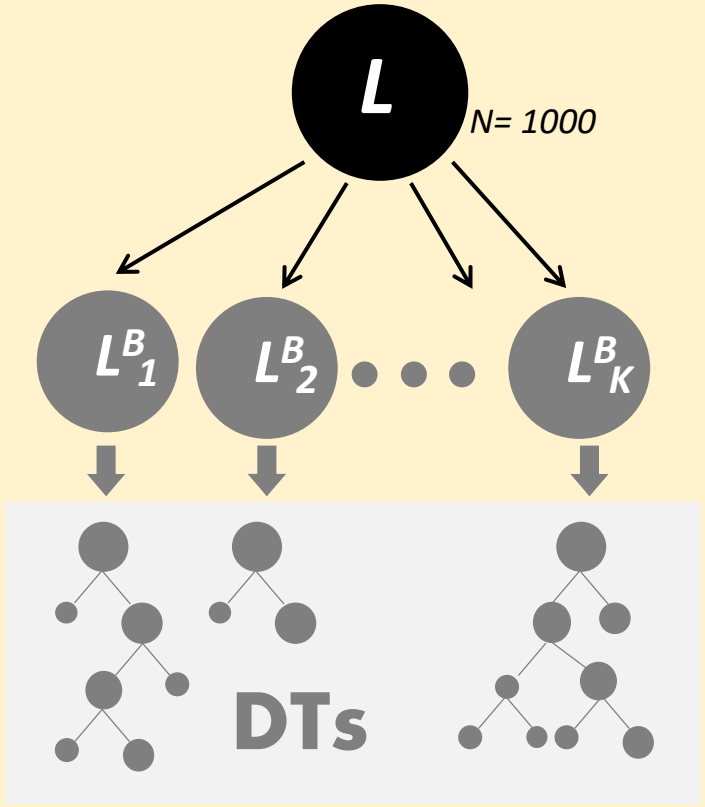
e.g.
Even though you prepare 80 predictor variables, only a handful of those are compared.

Dietterich, 2000

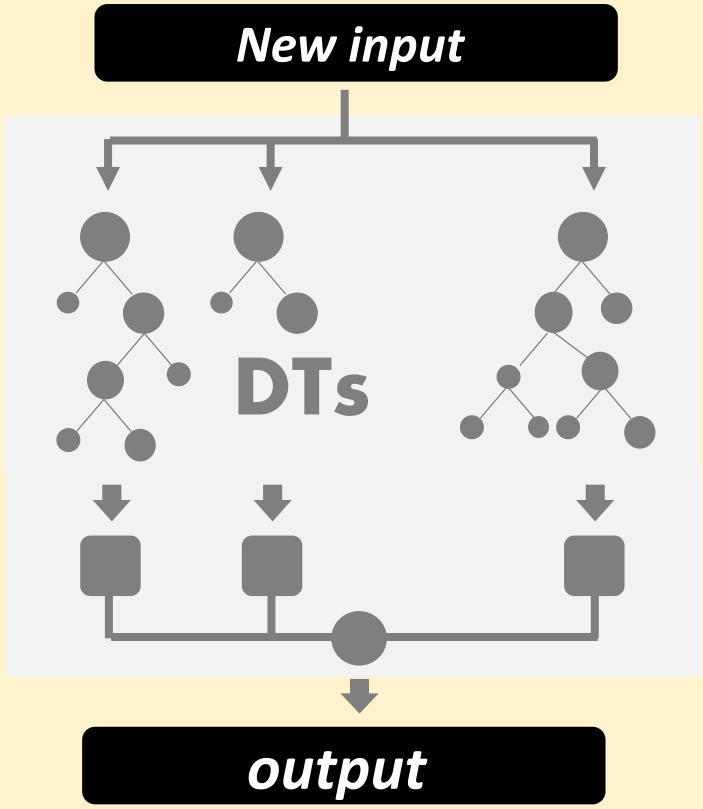
Random forests model ensemble approach

(Breiman 2001)

Step 1: building models



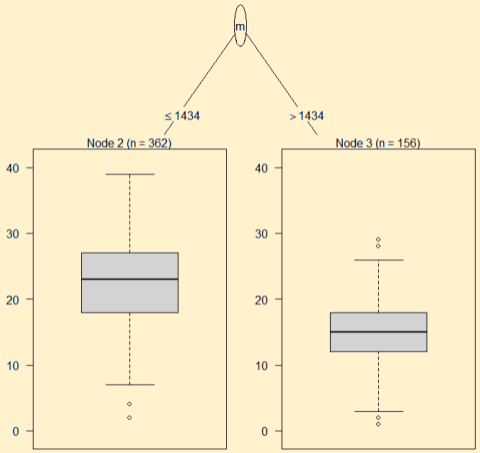
Step 2: running models

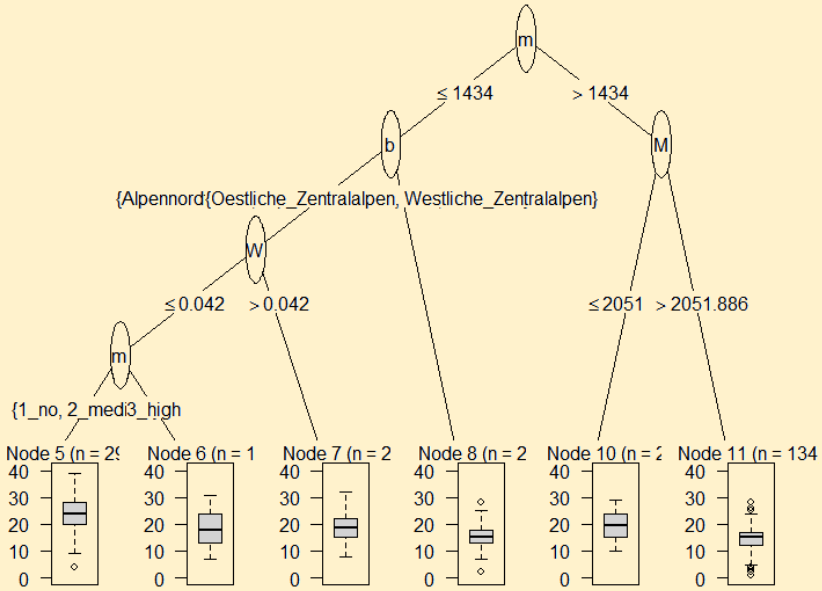


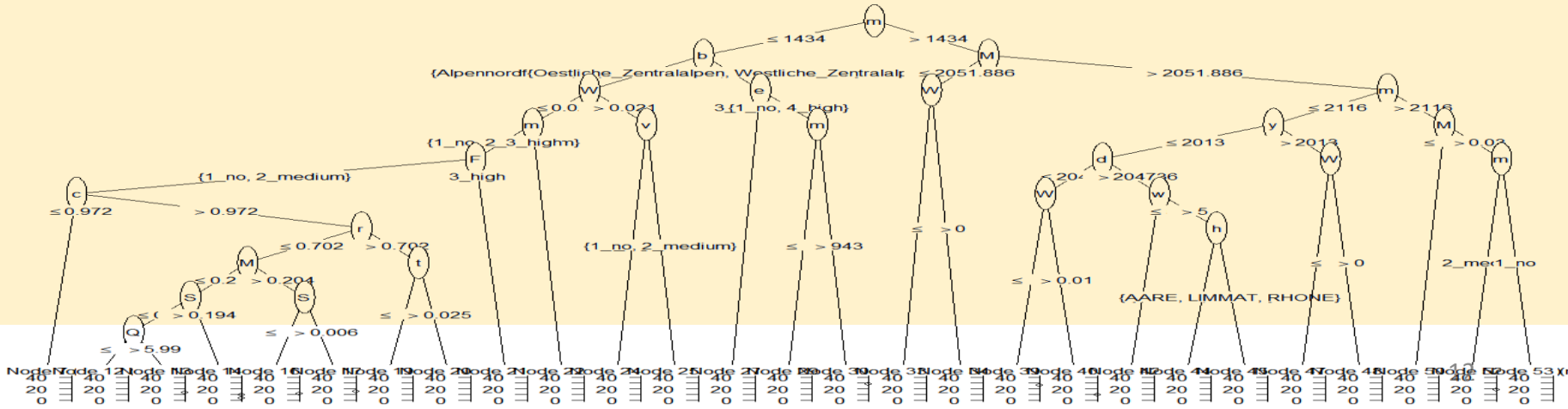
Breiman, *Machine Learning*, 1996

Two problems of decision tree algorithms

Over-fitting problem







Variable selection bias

Preferential order

Binary < categorical < continuous

Random forests?

Biased estimation on relative variable importance

History

1963: Morgan & Sonquist first developed the tree model protocol

1984: Breiman radically improved

1987: Mingers et al. reported **the two problems**

1994: White & Liu proposed statistical approach to solve

1999: Strasser & Weber proposed permutation test

(several attempts exist here)

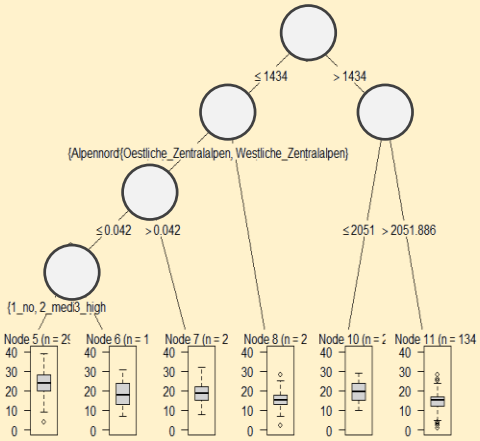
(De'ath et al. (2000) introduced it to ecology)

(2001: Breiman proposed Random forests)

2006: Hothorn et al. solved the problem

Statistically-reinforced decision trees

Conditional inference tree by Hothorn et al. (2006)



1. Estimate **p-values** for all covariates **x** based on permutation
   (p-value of test statistic: χ^2 & t)

Test type & test statistic		Covariate x	
		categorical	numeric
Response y	categorial	CMH (χ^2)	KW (χ^2)
	numeric	KW (χ^2)	Pearson (t)

CMH: Cochran-Mantel-Haenszel, KW: Kruskal-Wallis

2. Choose the covariate $\mathbf{x}_*$ with minimum p-value;
   stop if no covariates fall below significance level (α)
   (with Bonferroni correction)

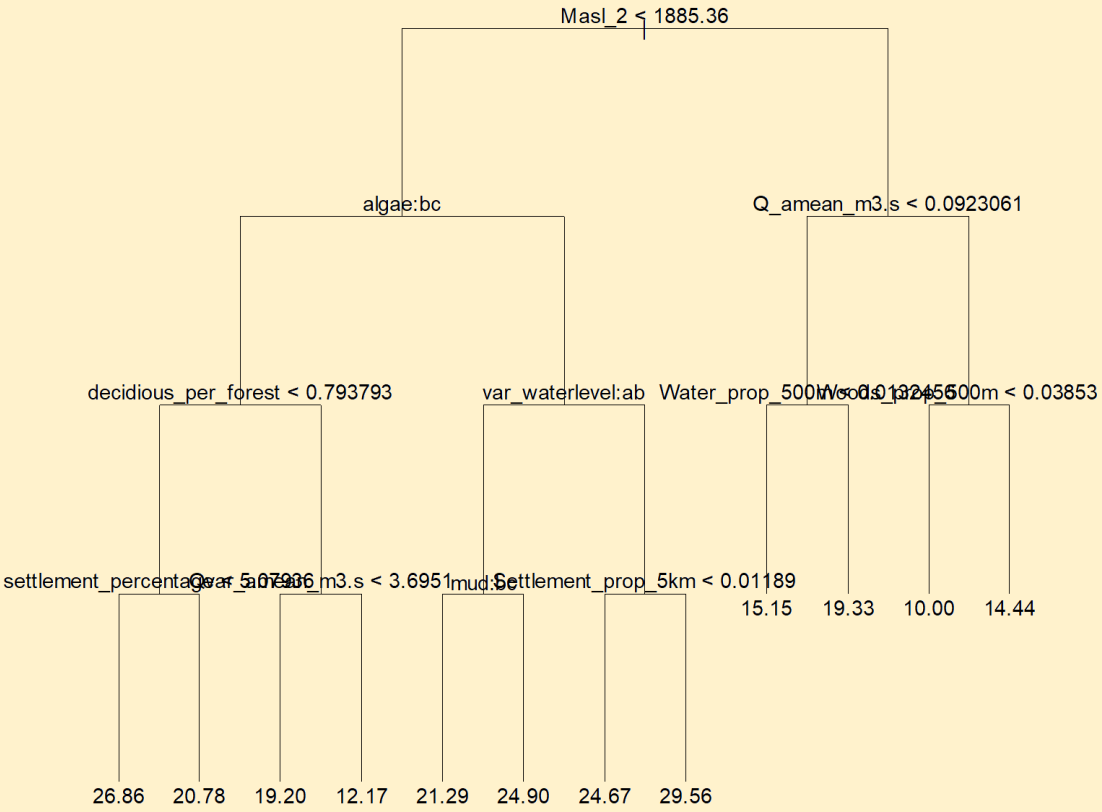
3. Find the value of the covariate x_* which best splits the sample into two subsamples and split (entropy or MSE)

4. Repeat steps 1-3 until being stopped

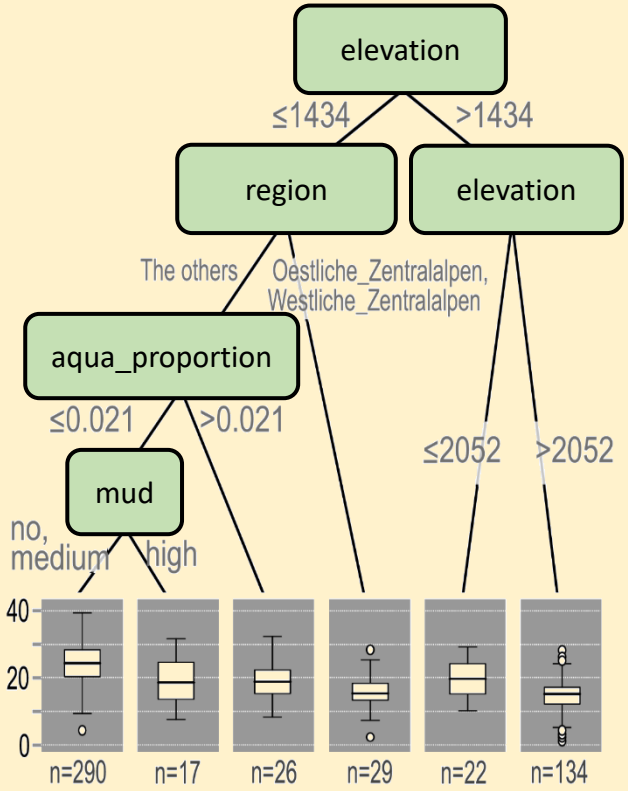
Statistically-reinforced decision trees

Conditional inference tree by Hothorn et al. (2006)

conventional tree

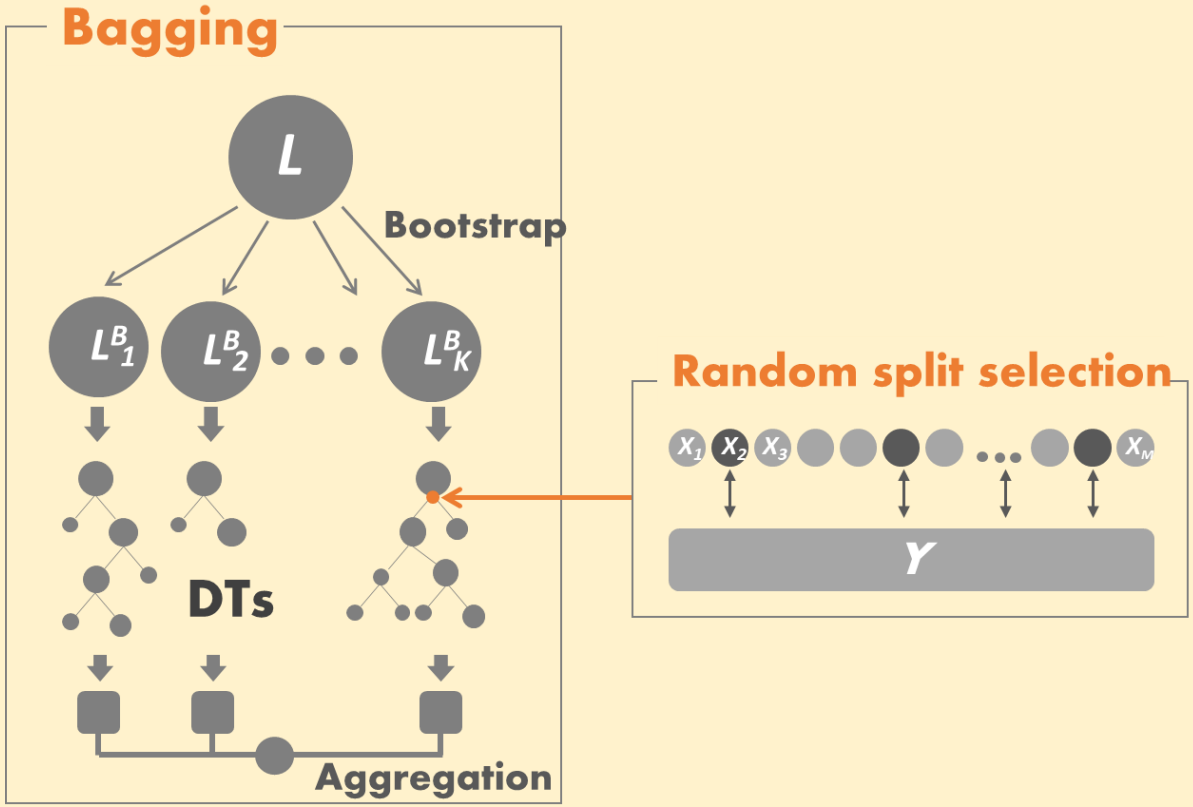


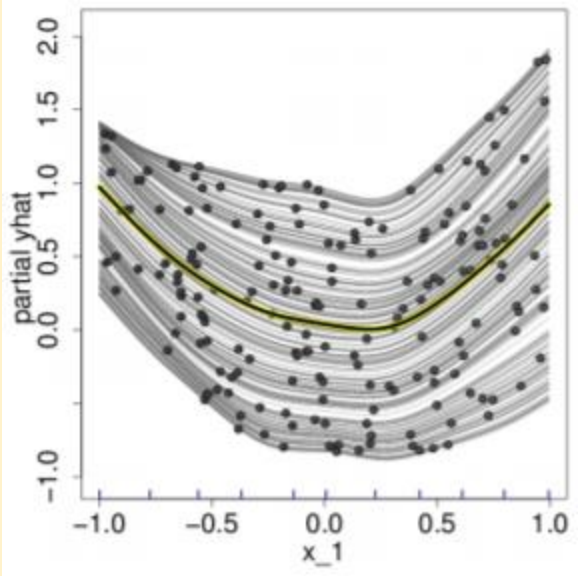
p < 0.01 tree

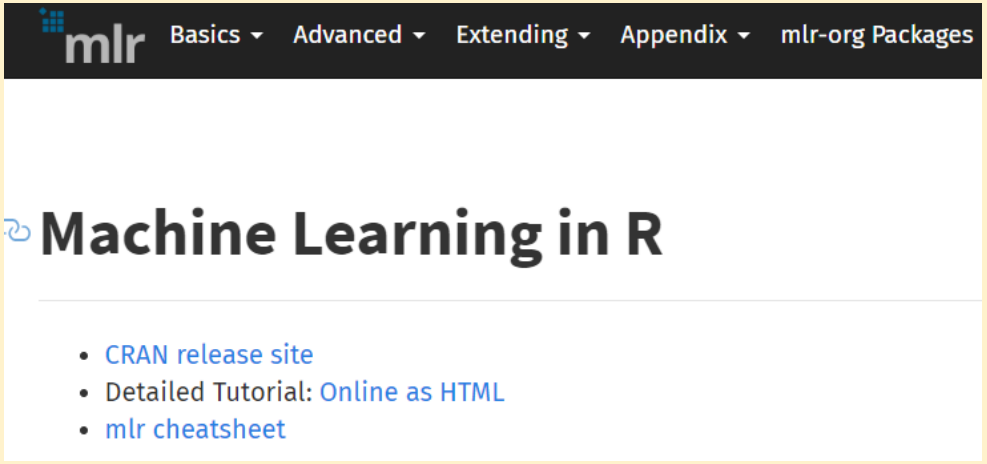


Statistically-reinforced random forests

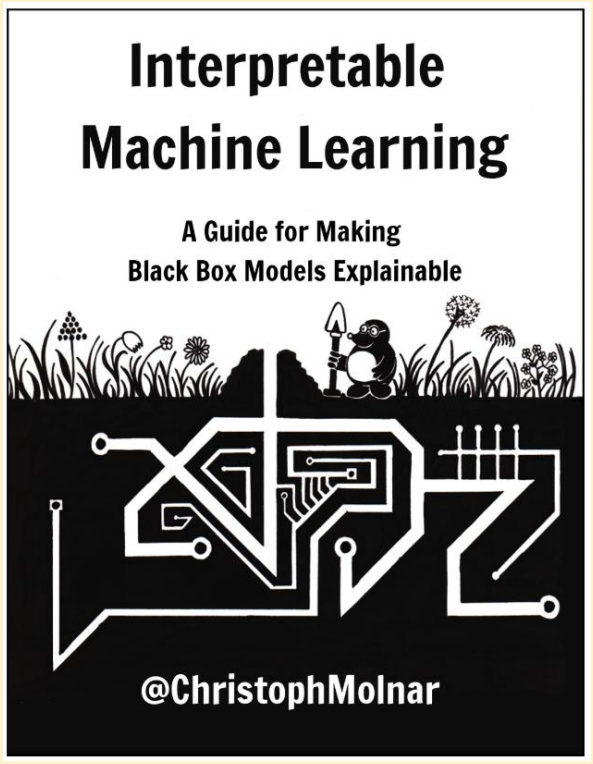
Conditional random forest by Strobl et al. (2008)







https://mlr.mlr-org.com/



https://christophm.github.io/interpretable-ml-book/