“Intelligence is 10 million rules” (Doug Lenat)….but Rules are meant to be generalizable

Reading

What are decision trees?

Carl Kingsford & Steven L Salzberg

Decision trees have been applied to problems such as assigning protein function and predicting splice sites. How do these classifiers work, what types of problems can they solve and what are their advantages over alternatives?

NATURE BIOTECHNOLOGY VOLUME 26 NUMBER 9 SEPTEMBER 2008

Learning from the data

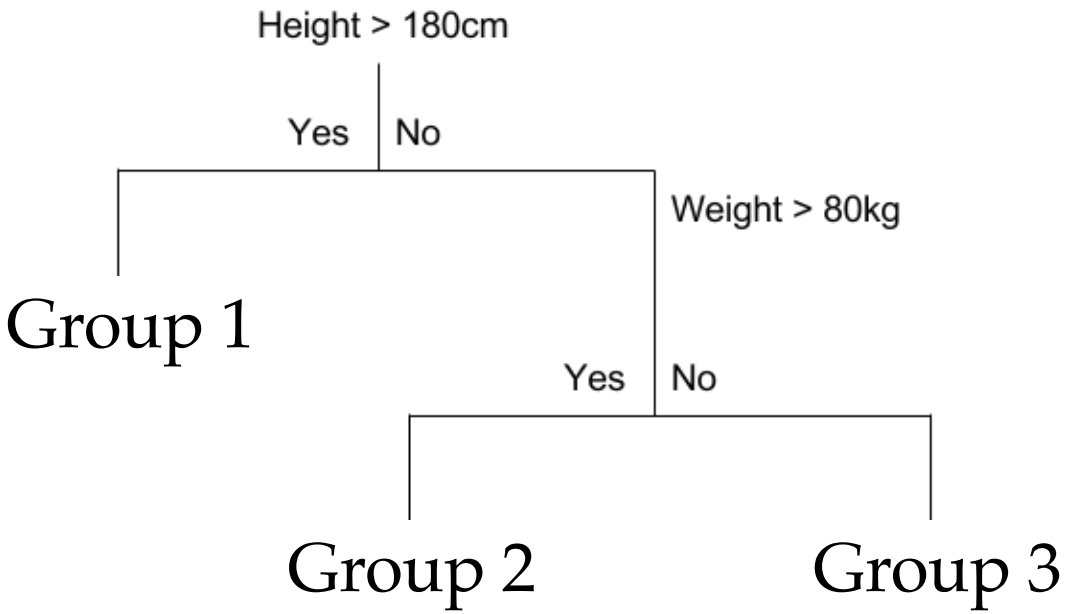
Pattern recognition

Learning from the data
Machine learning algorithms - Two main types

Unlabeled data

Unsupervised Learning Algorithm

Prediction based on finding patterns in the data



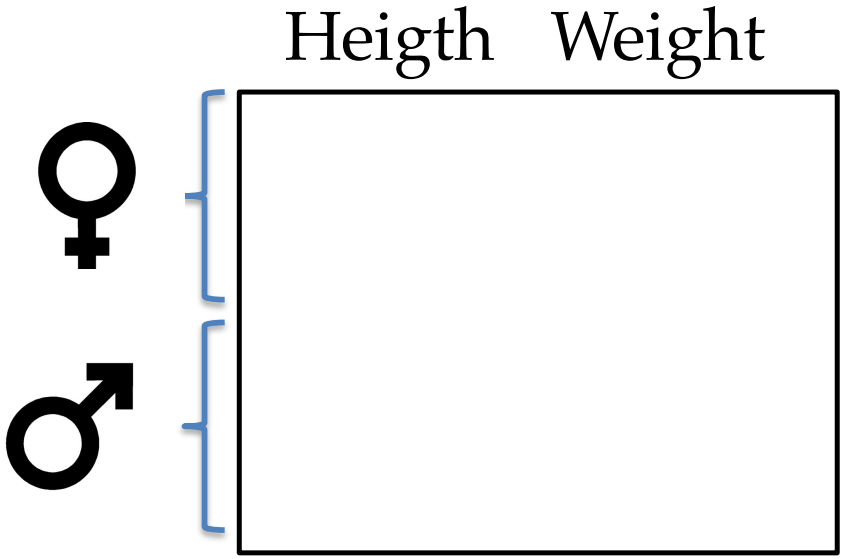
e.g., Finding number of groups in data and ways to classify (predict) observations based on their characteristics (height/weight)

Learning from the data
Machine learning algorithms - Two main types

Learning from the data
Machine learning algorithms - Two main types

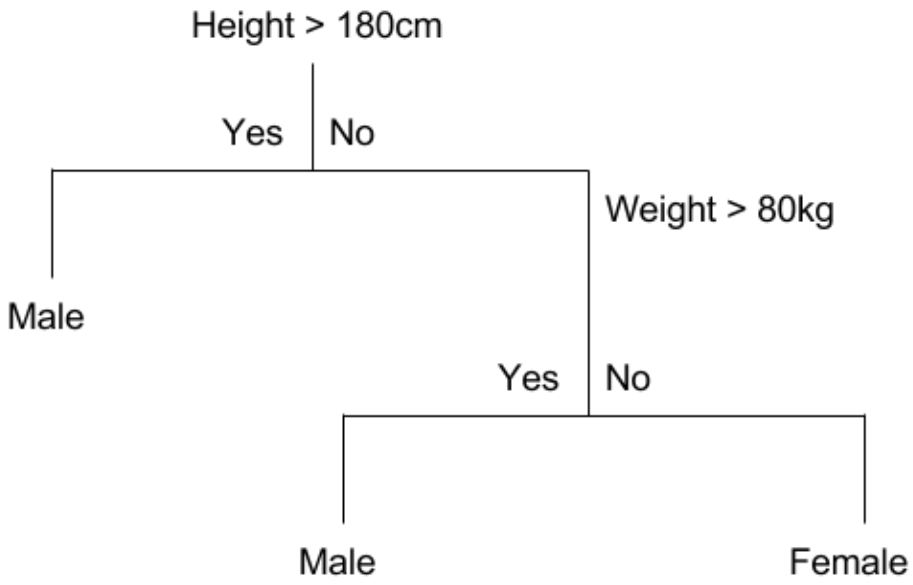
Labeled data → ***Supervised Learning Algorithm*** → Prediction based on knowing the label

Label = gender



Learning from the data
Machine learning algorithms - Two main types

Labeled data → ***Supervised Learning Algorithm*** → Prediction based on knowing the label



Predicting gender on the basis of Height and Weight

Label = gender

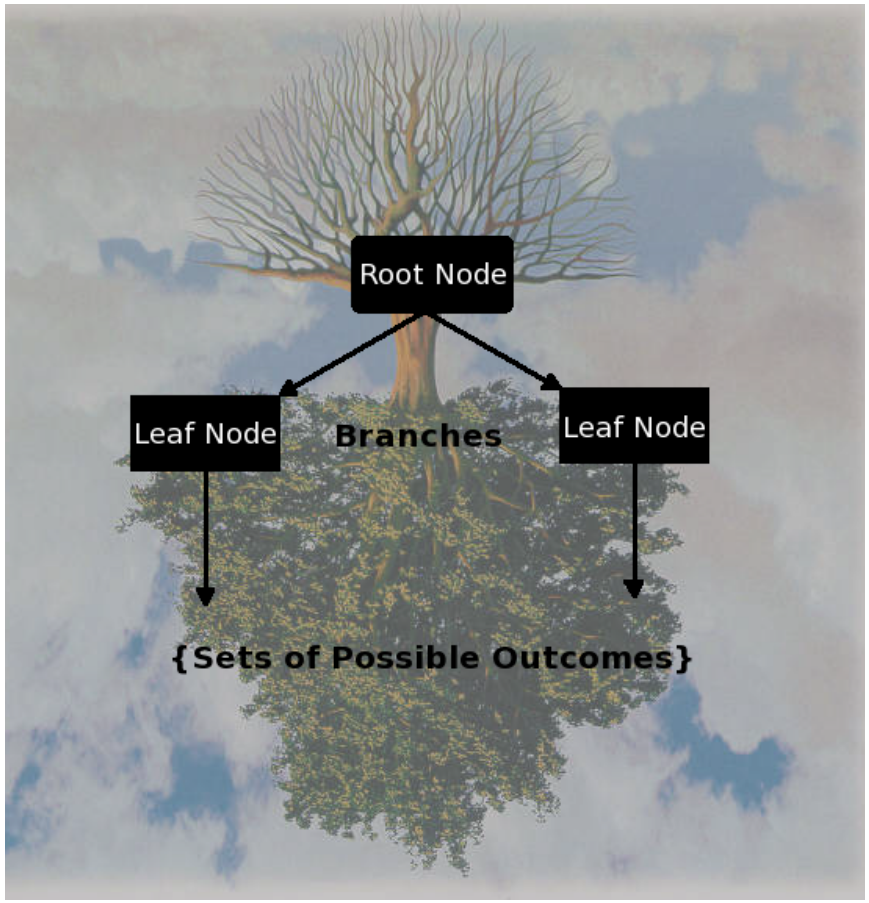
CART: Classification and Regression Trees – a powerful (machine learning) yet simple analytical tool for multivariate pattern description

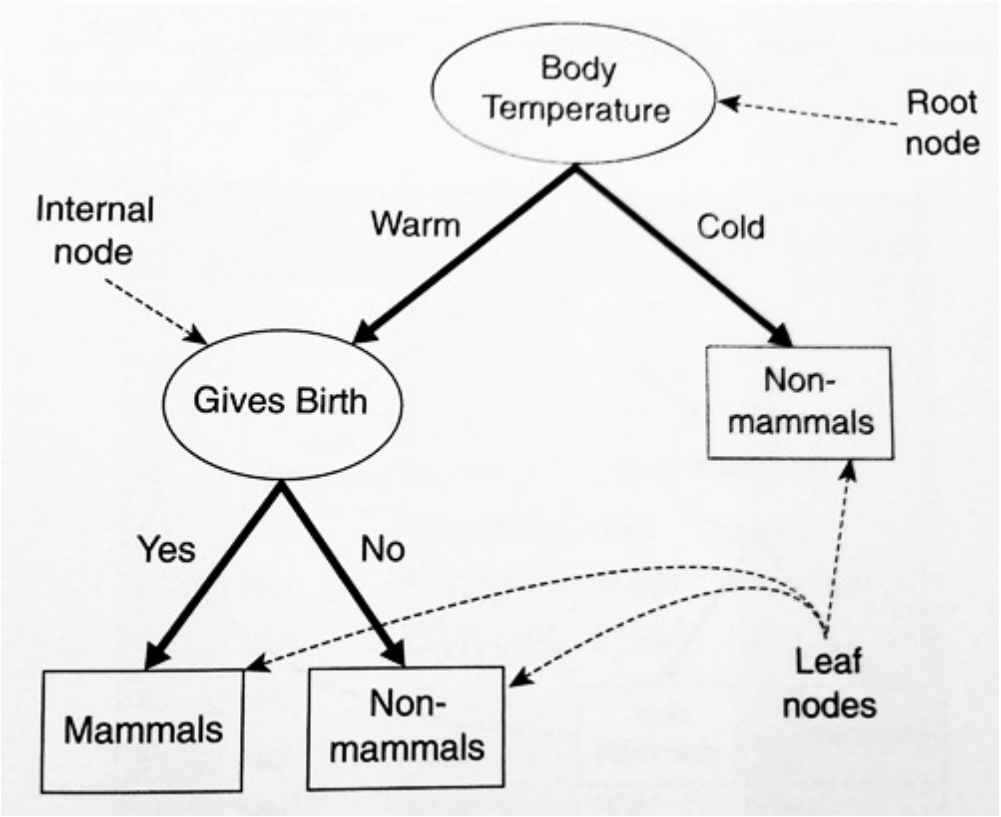
(Leo Breiman and colleagues 1984)

"Decision tree learning is among the most popular machine learning techniques used for ecological modelling. Decision trees can be used to predict the value of one or several (dependent) variables. " Jopp et al. (2011)

Tree anatomy

"Decision trees are hierarchical structures, where each internal node contains a test on an attribute, each branch corresponding to an outcome of the test, and each leaf node giving a prediction for the value of the class variable." (Jopp et al. 2011)

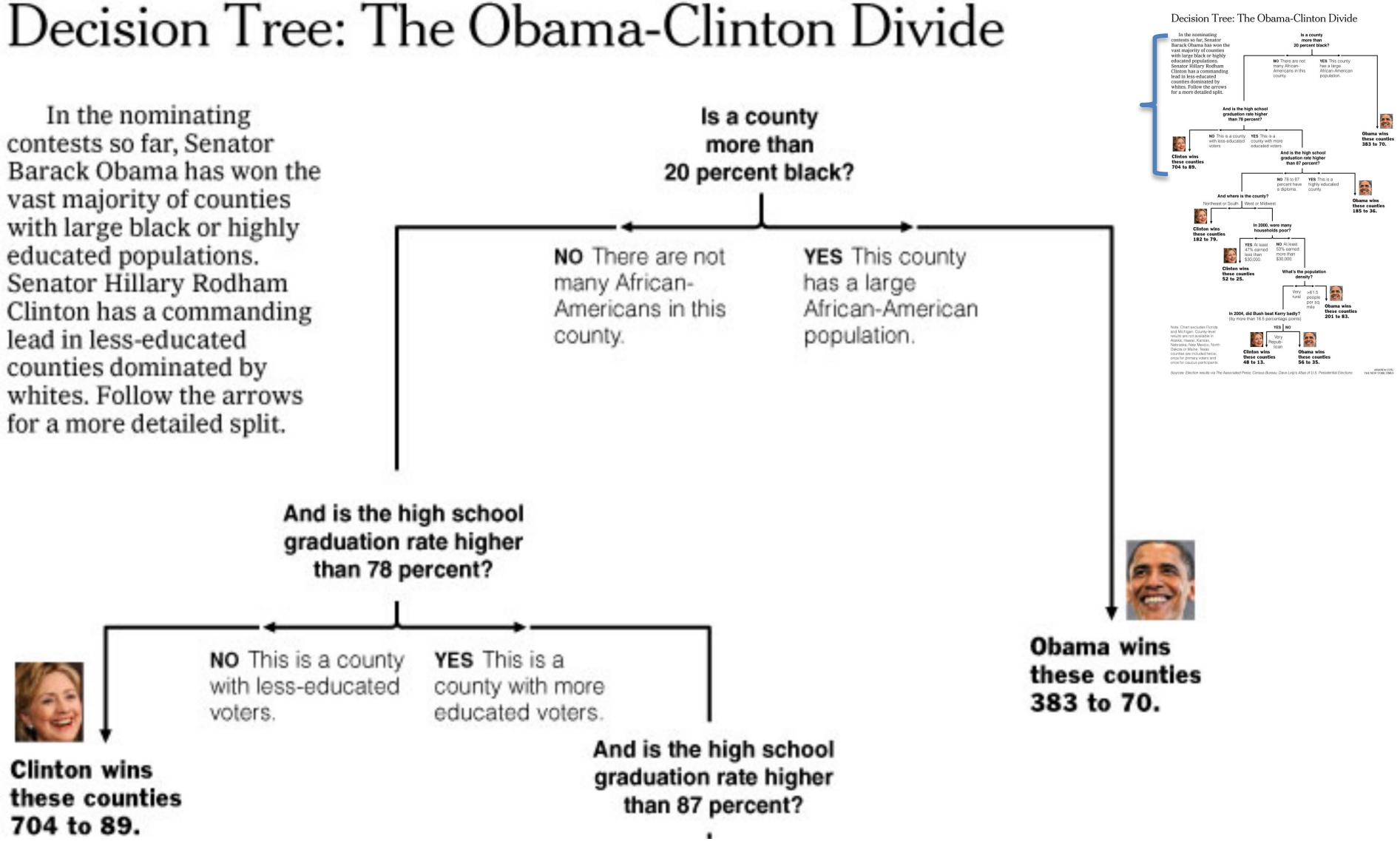




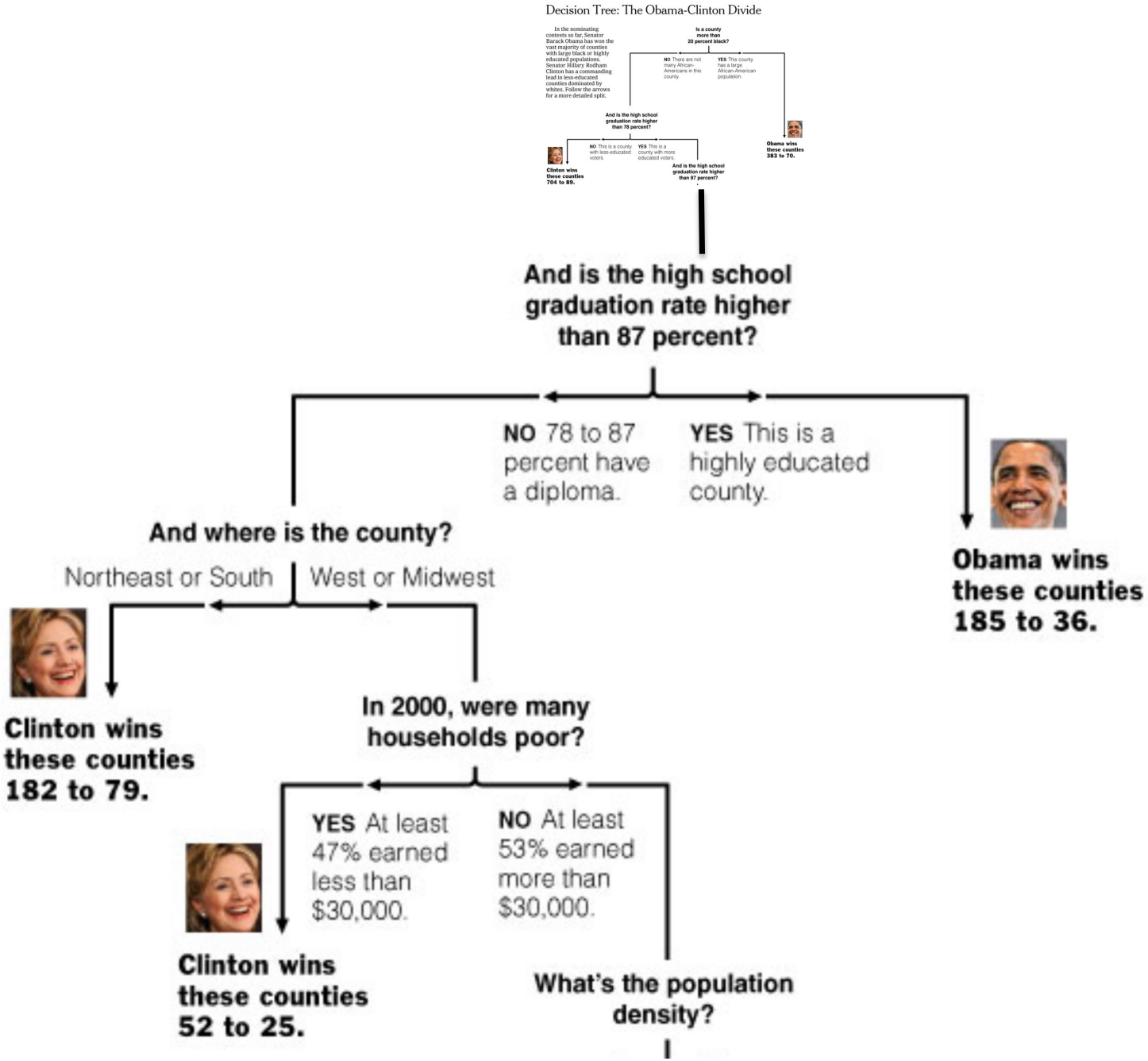
Source
http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/sdaugherty/decisiontree.html

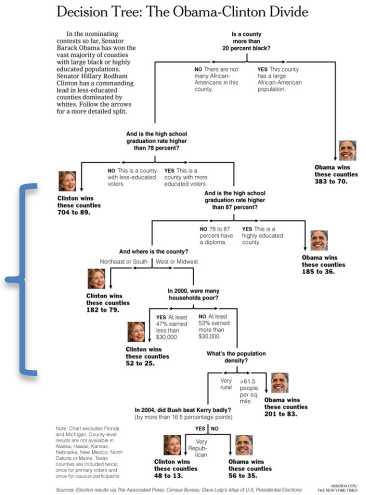
Learning from the data – Classification Trees

Deal with complex data but easy to convey results

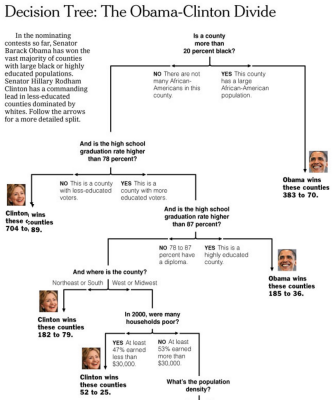


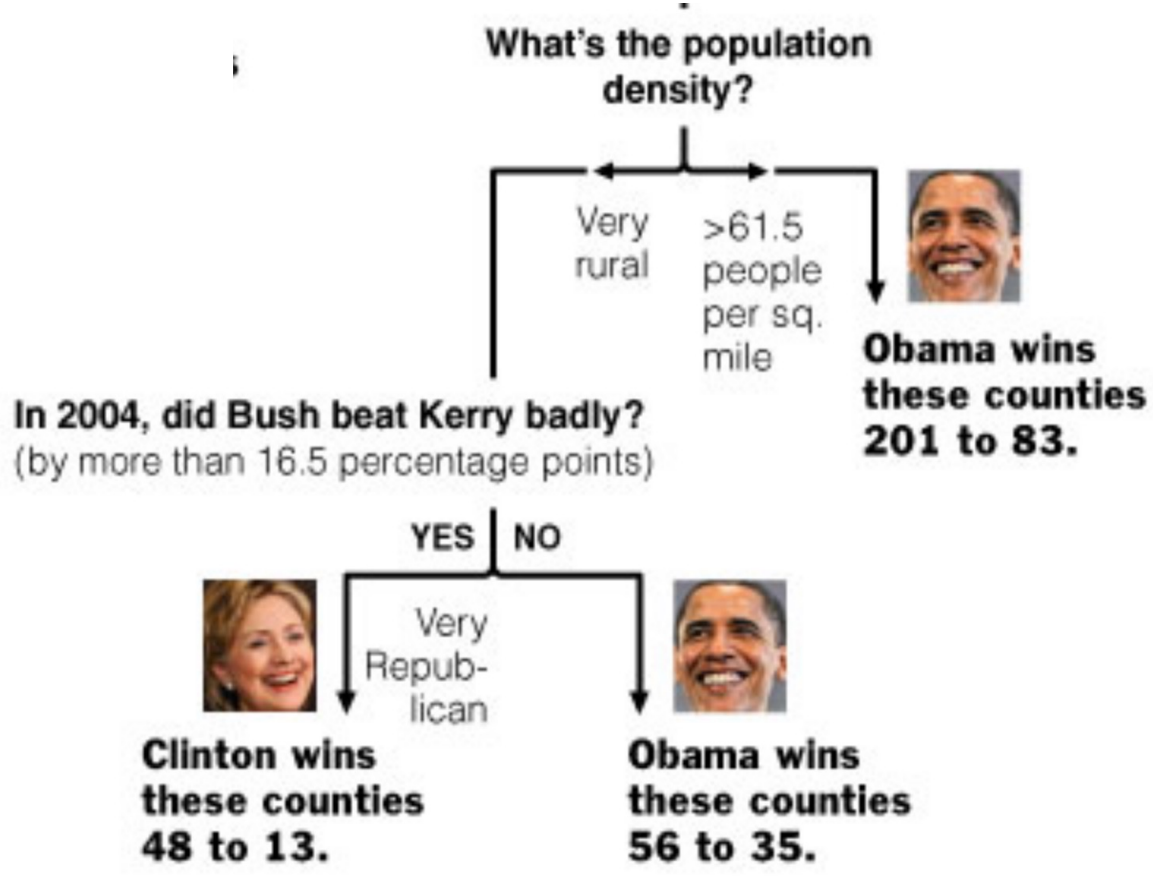
Learning from the data – ClassificationTrees

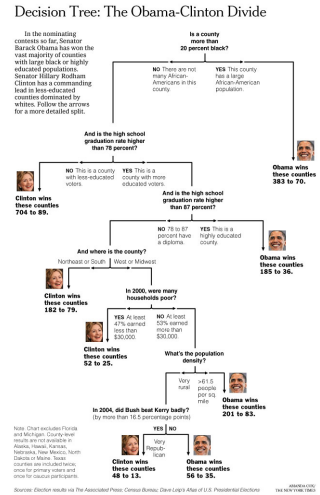




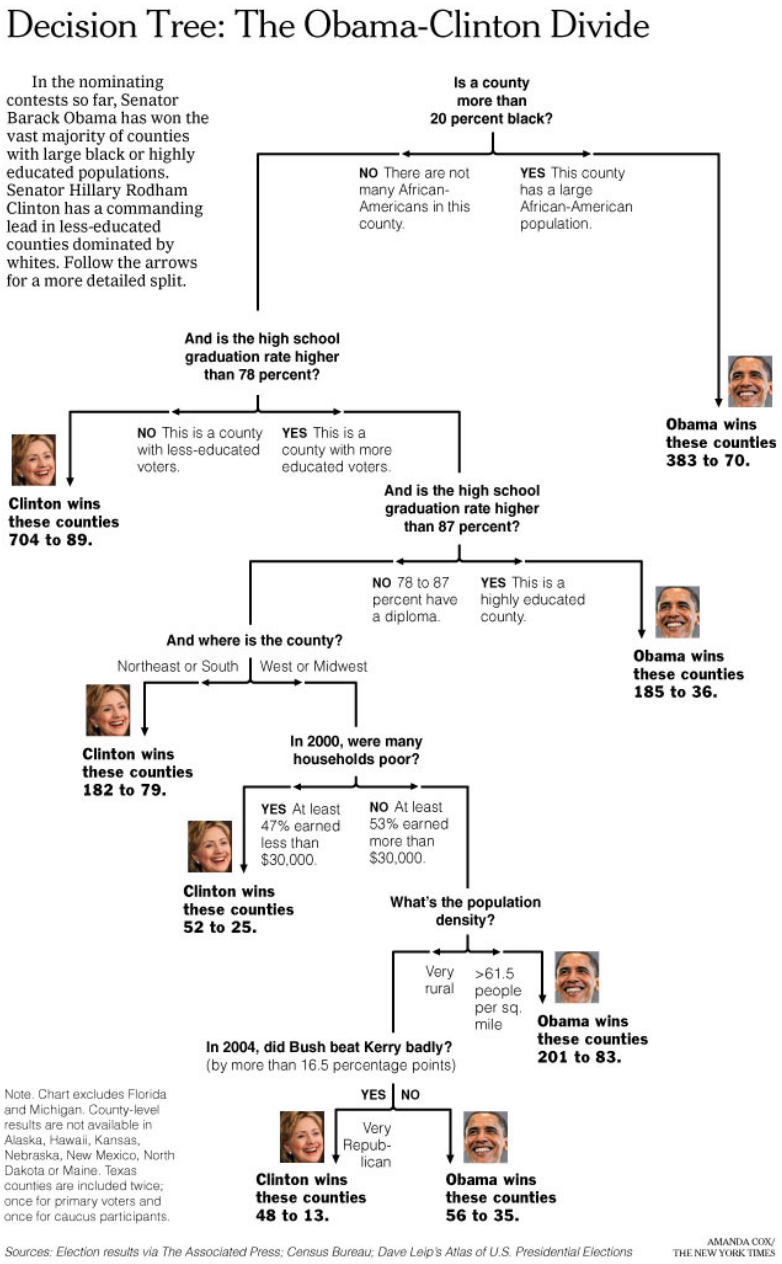
Learning from the data – ClassificationTrees







Learning from the data – ClassificationTrees



Classification trees model categories, including 1s and 0s (male / female, presence / absence, non-infected / recovery / infected)

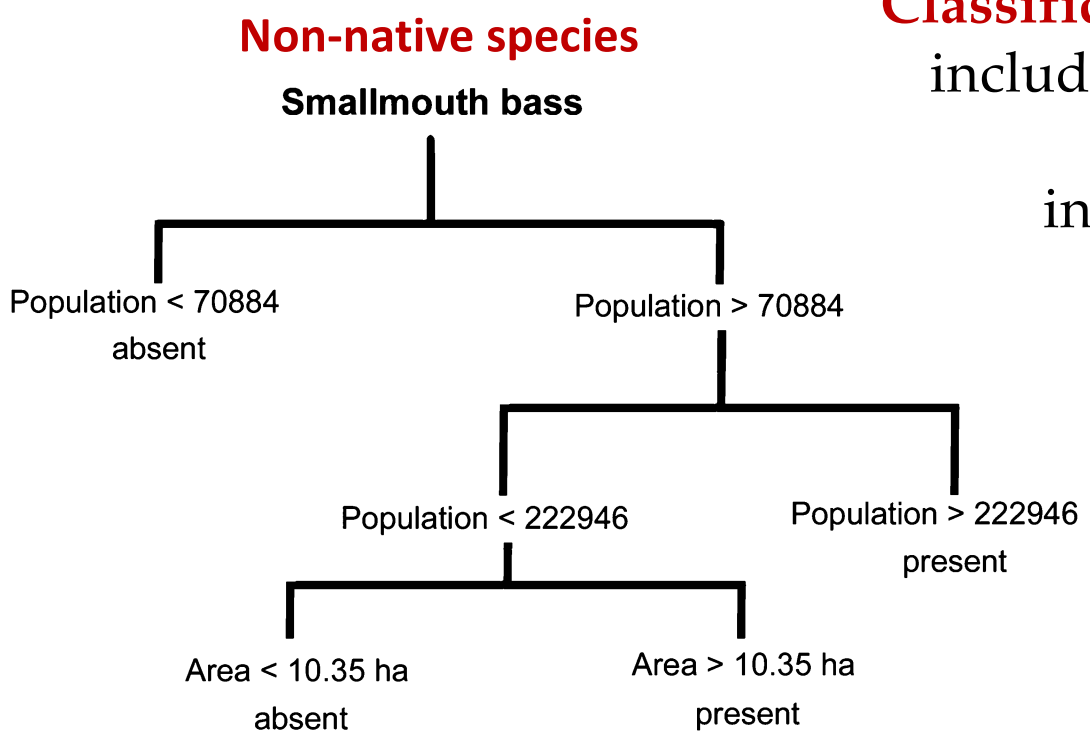


Figure 1 Summary of classification tree analysis predicting smallmouth bass occurrence in British Columbia based on lake morphology, distance to road and human population census data.

Probabilities of presence are calculated by the classification tree, but then transformed into 1 (e.g., >0.50) or absence (e.g., <=0.50)

Columbia based on the classification tree analysis (Fig. 1). Evaluation of the independent validation dataset showed that overall classification success was 93.5%, with 83.1% sensitivity and 100% specificity. Extrapolation of the classification tree

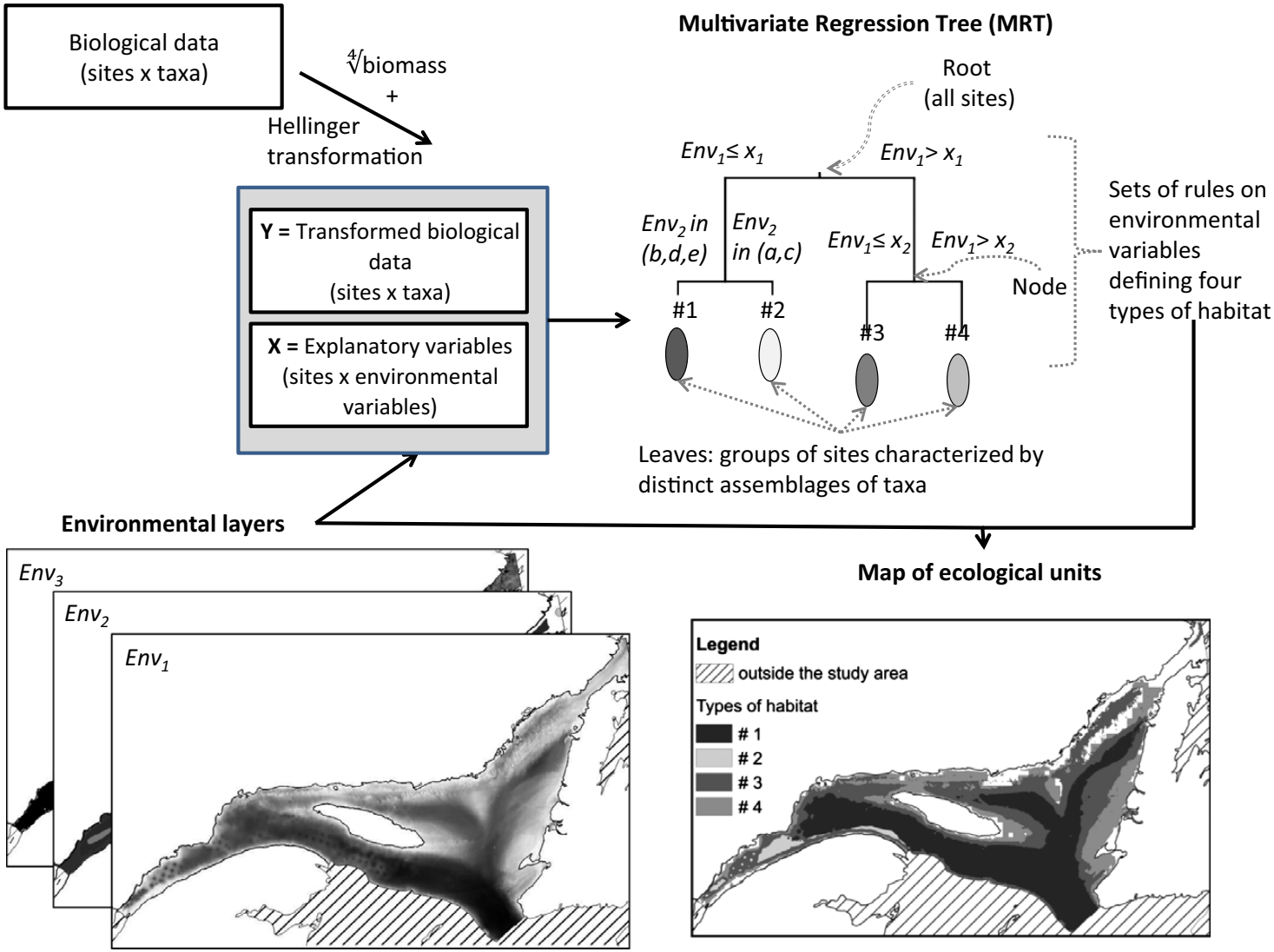
Diversity and Distributions, (Diversity Distrib.) (2009) **15**, 831–840

BIODIVERSITY RESEARCH

Predicting introduction, establishment and potential impacts of smallmouth bass

Sapna Sharma[1]*, Leif-Matthias Herborg[2,3] and Thomas W. Therriault[2]

Regression trees model quantitative variables (e.g., species abundances)



Diversity and Distributions, (Diversity Distrib.) (2015) **21**, 1167–1180

BIODIVERSITY RESEARCH

Delineating marine ecological units: a novel approach for deciding which taxonomic group to use and which taxonomic resolution to choose

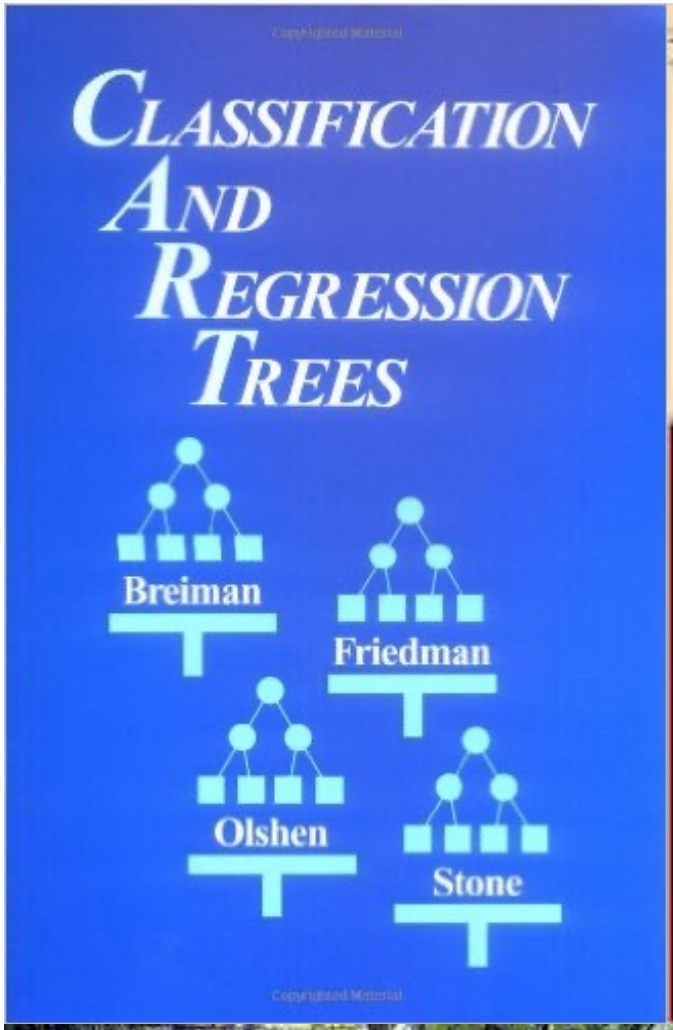
Anne Fontaine[1,2]*, Rodolphe Devillers[3], Pedro R. Peres-Neto[4] and Ladd E. Johnson[1]

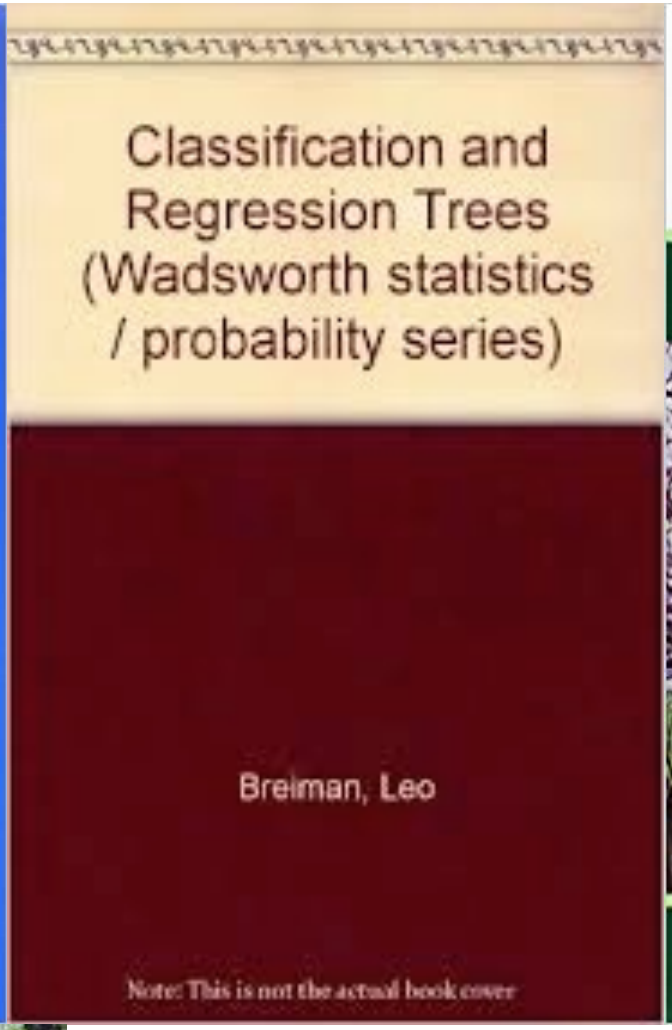
Classification versus Regression Trees (CART)

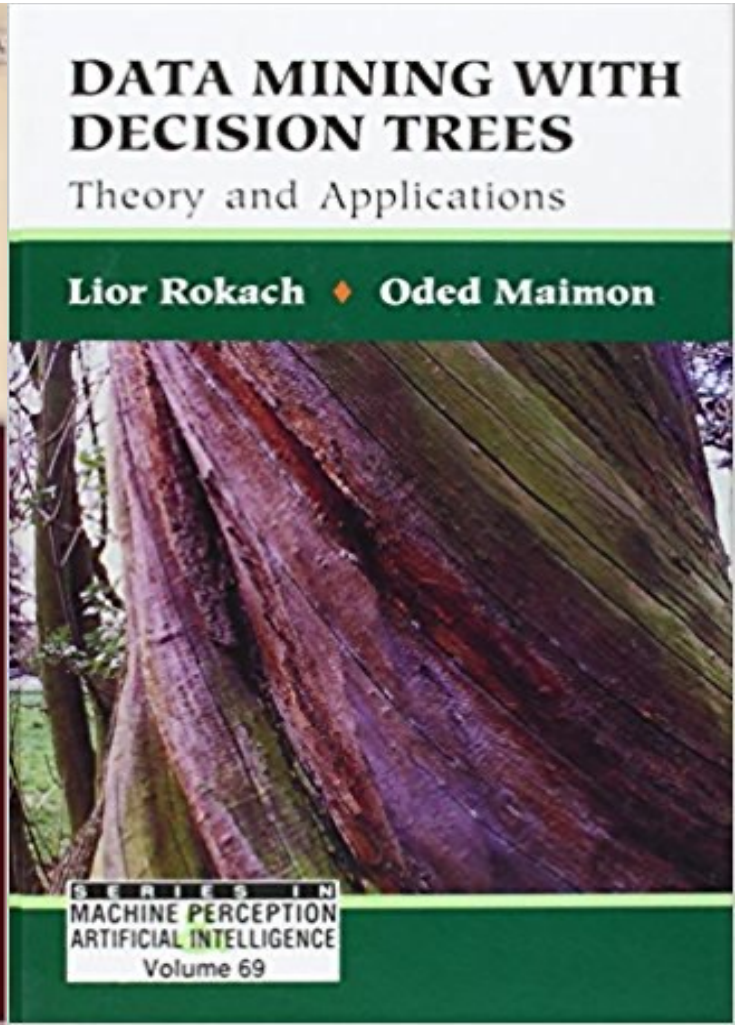
- Classification (sometimes referred as to decision trees) trees model dependent variables that have a finite number of categories (unordered values) - This lecture.

- Regression trees model dependent variables that are continuous.

The classification tree algorithm





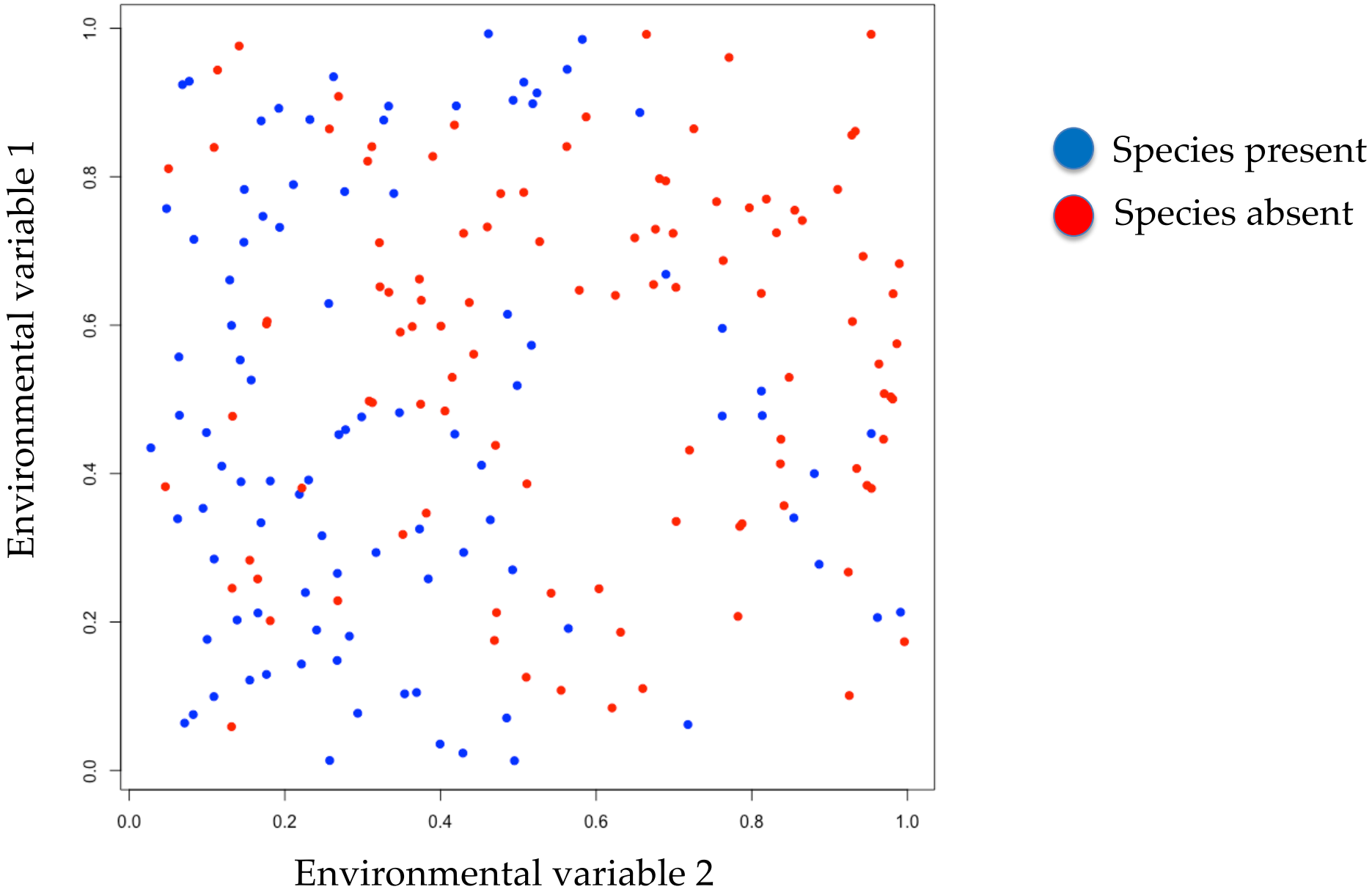


Ecology, 81(11), 2000, pp. 3178–3192

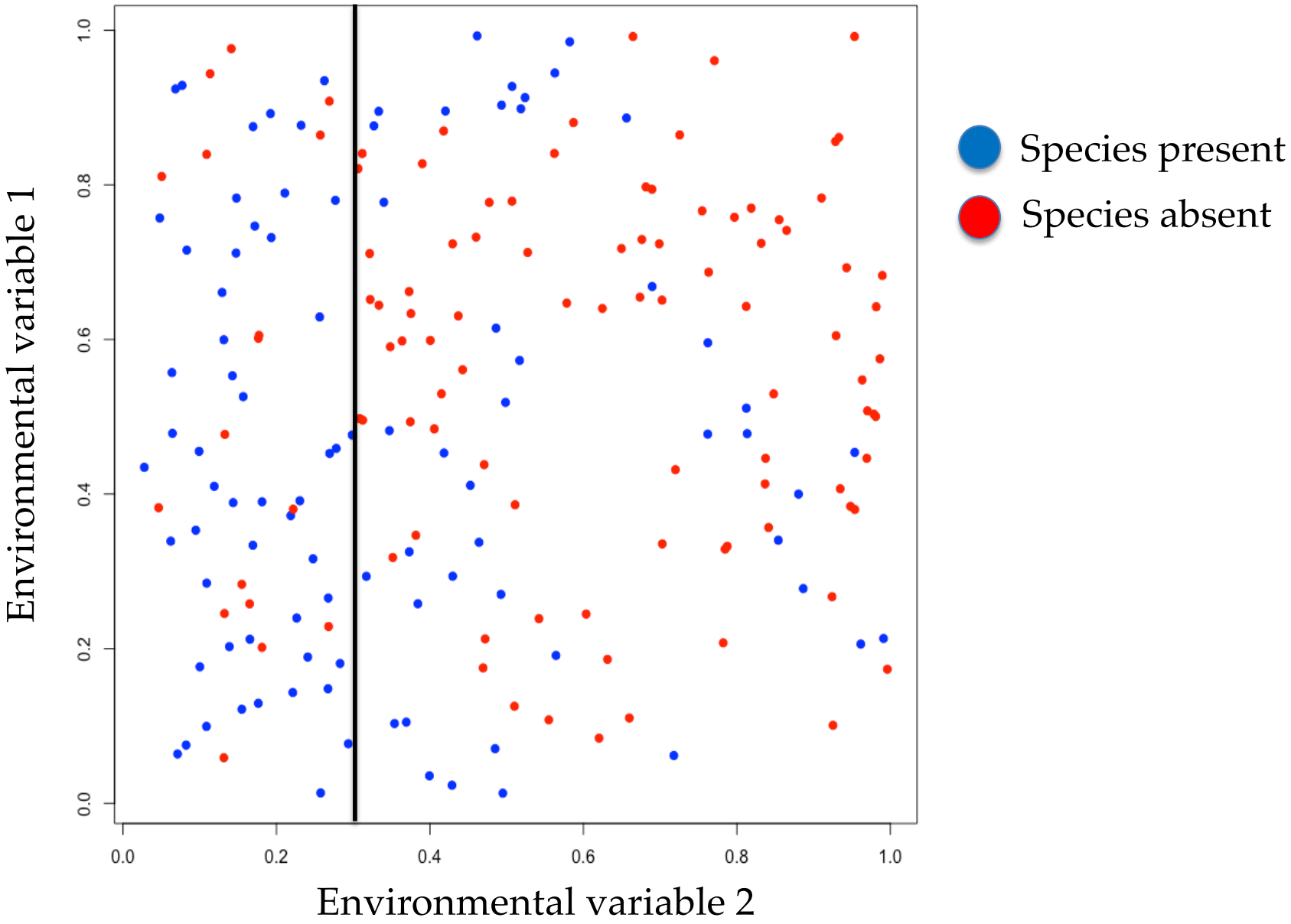
CLASSIFICATION AND REGRESSION TREES: A POWERFUL YET SIMPLE TECHNIQUE FOR ECOLOGICAL DATA ANALYSIS

GLENN DE'ATH[1] AND KATHARINA E. FABRICIUS[2]

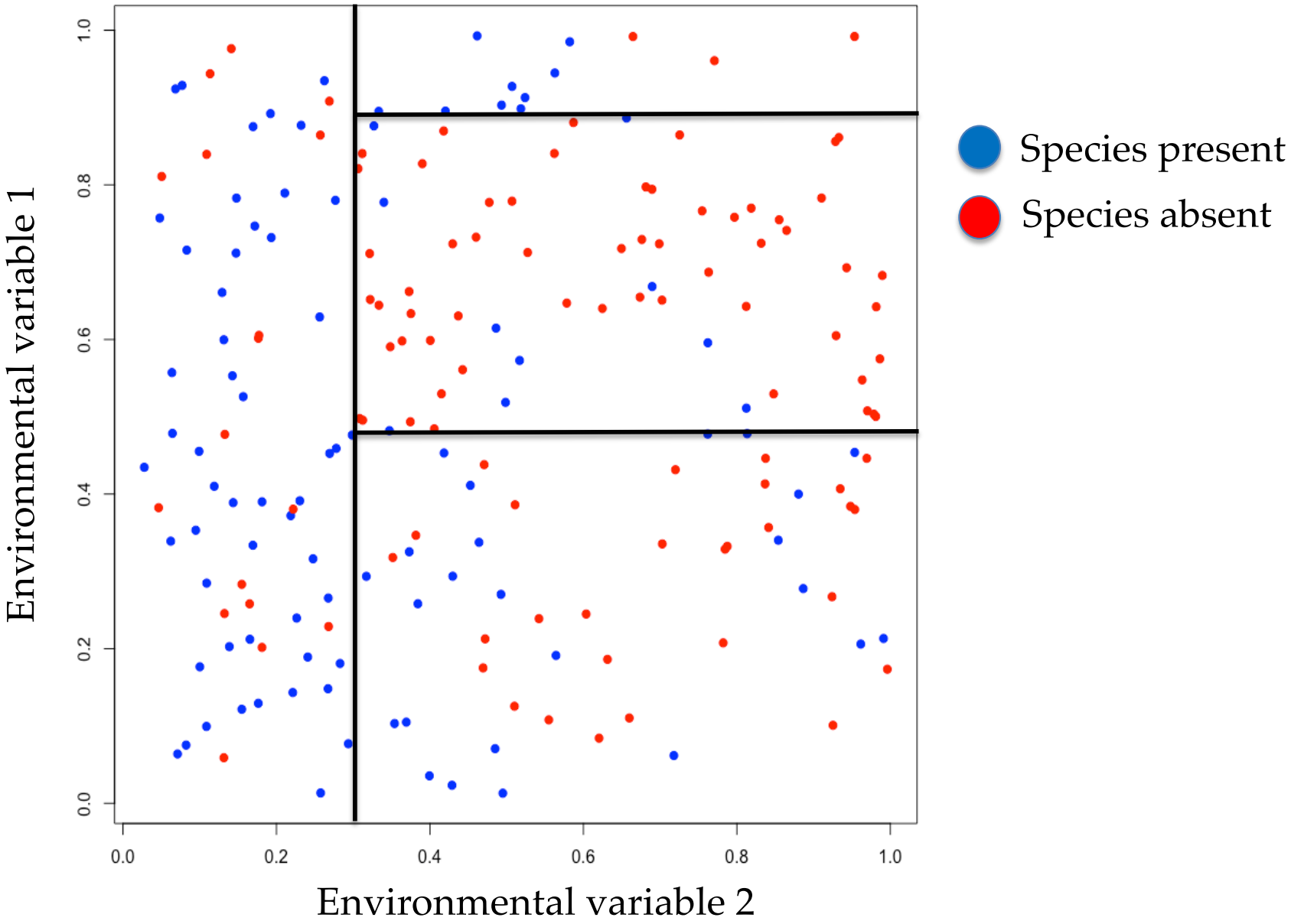
How to model these data?



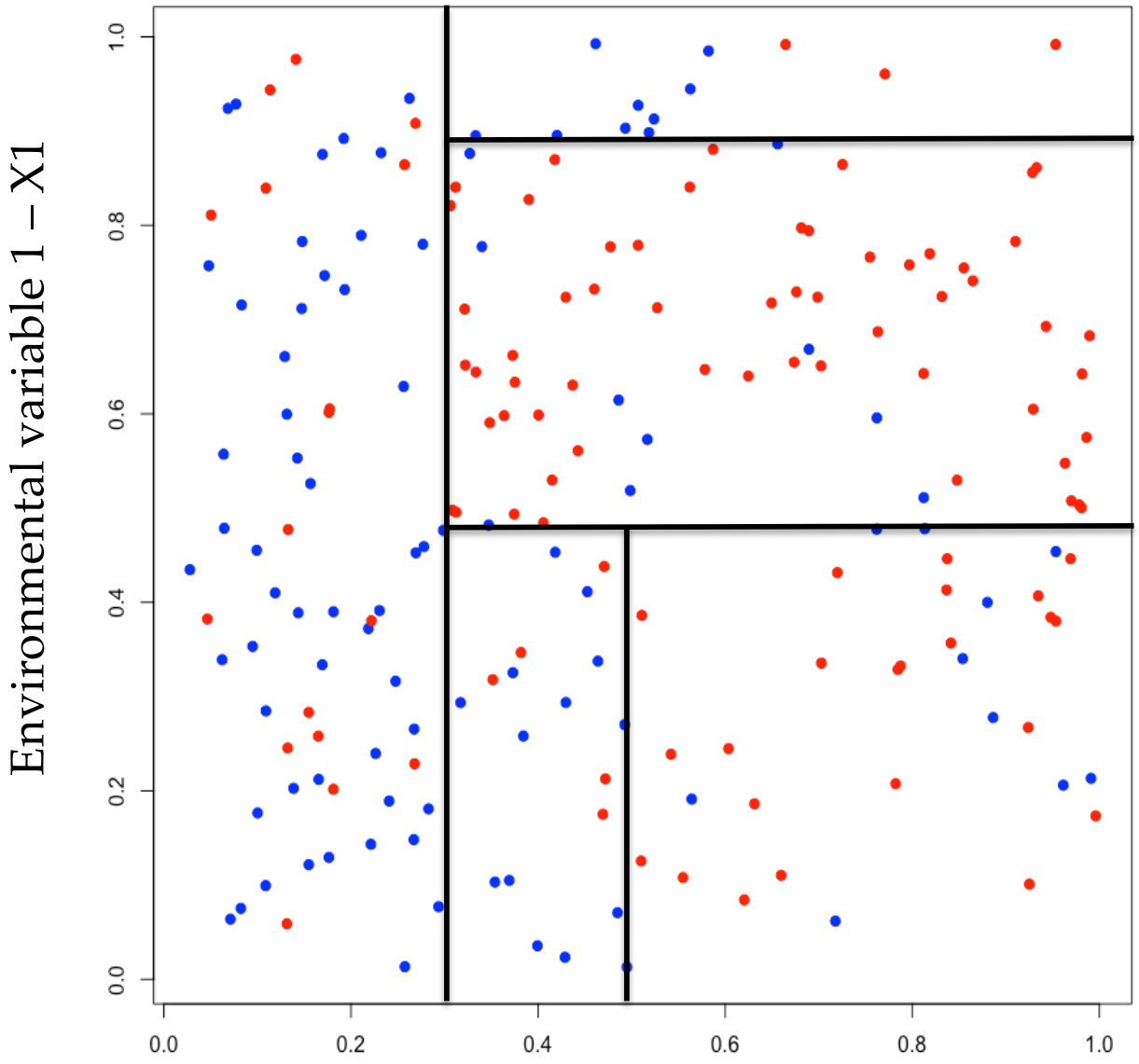
Mosaic or partition plot



Mosaic or partition plot

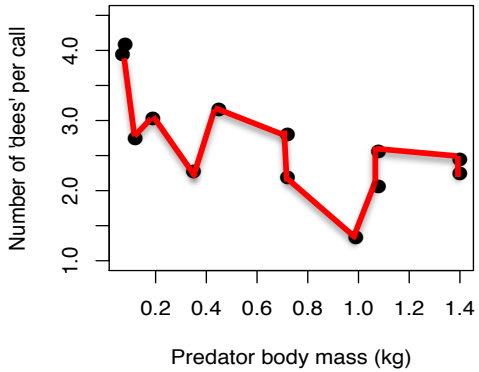


Mosaic or partition plot (best partitioning possible without too much fitting; many ways to determine final model)

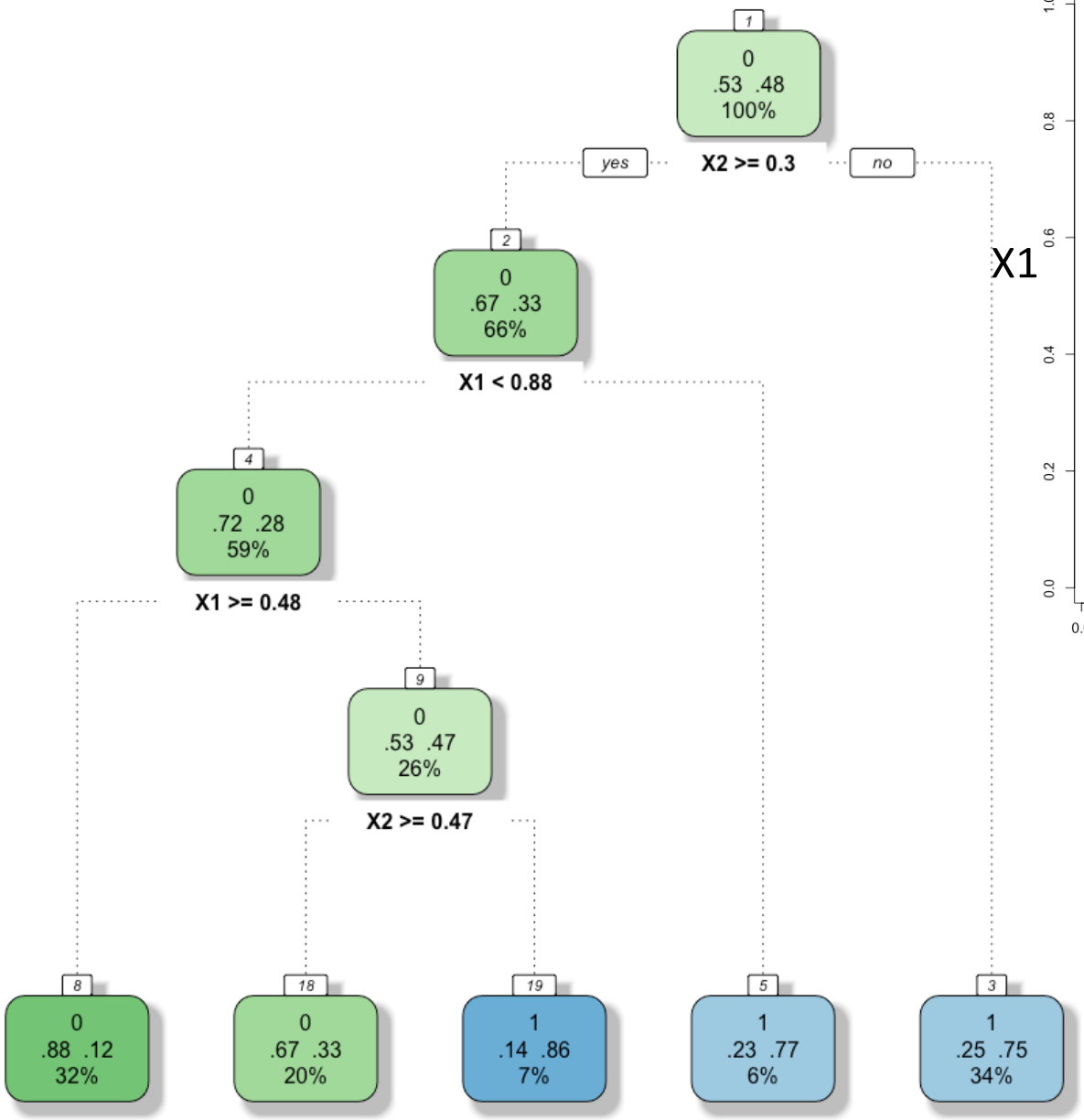


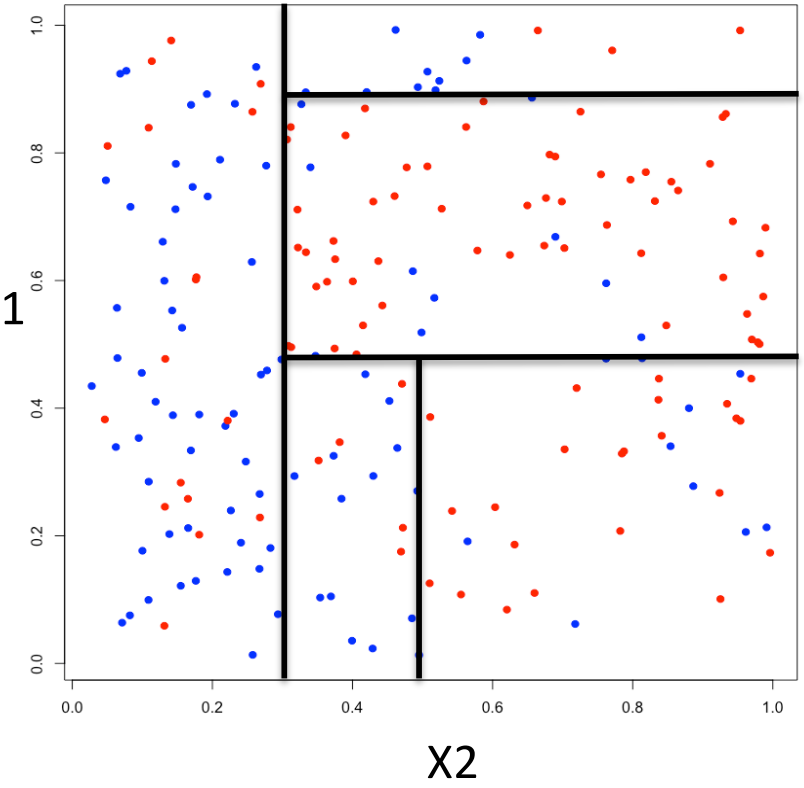
Species present

Species absent

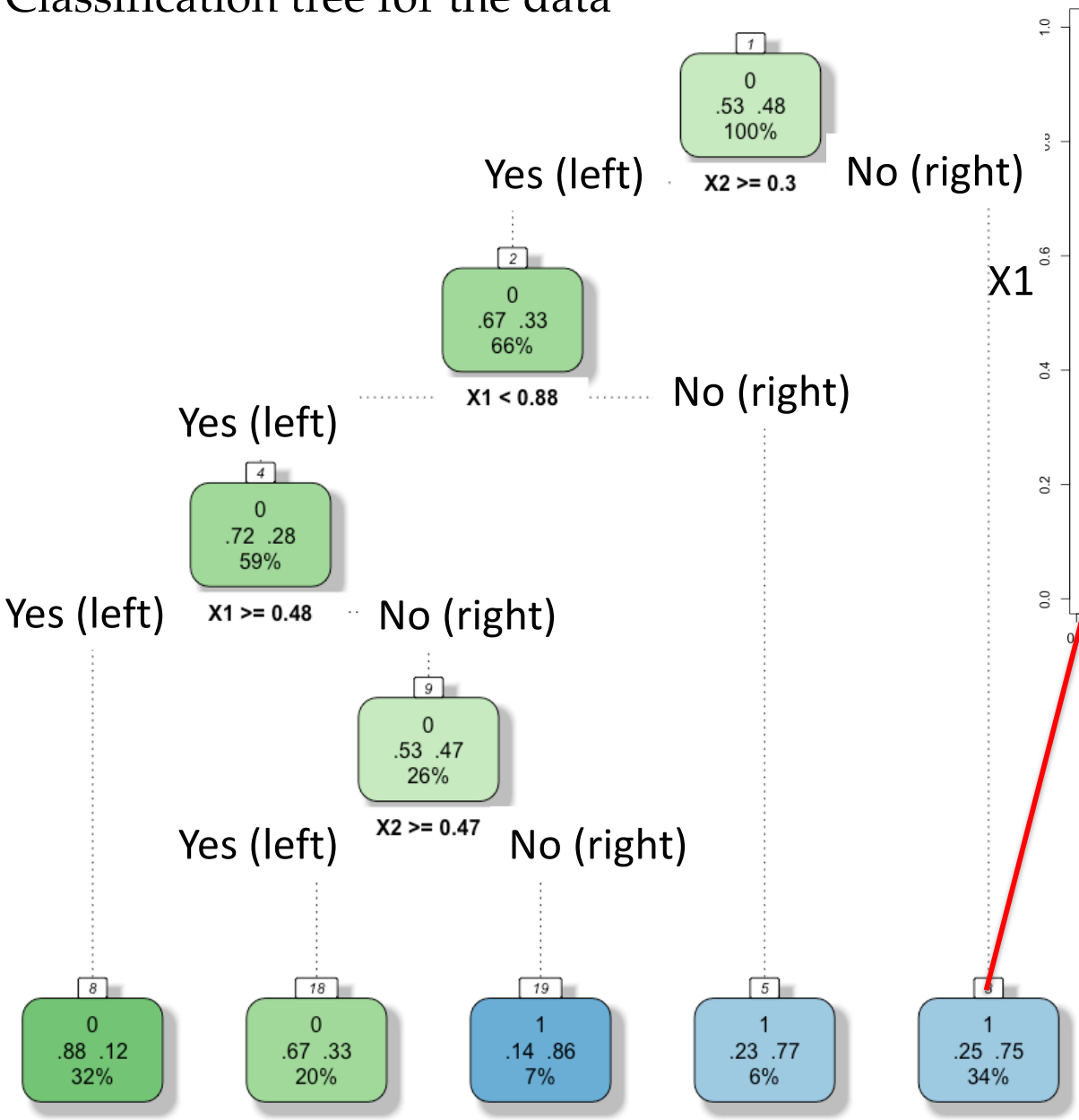


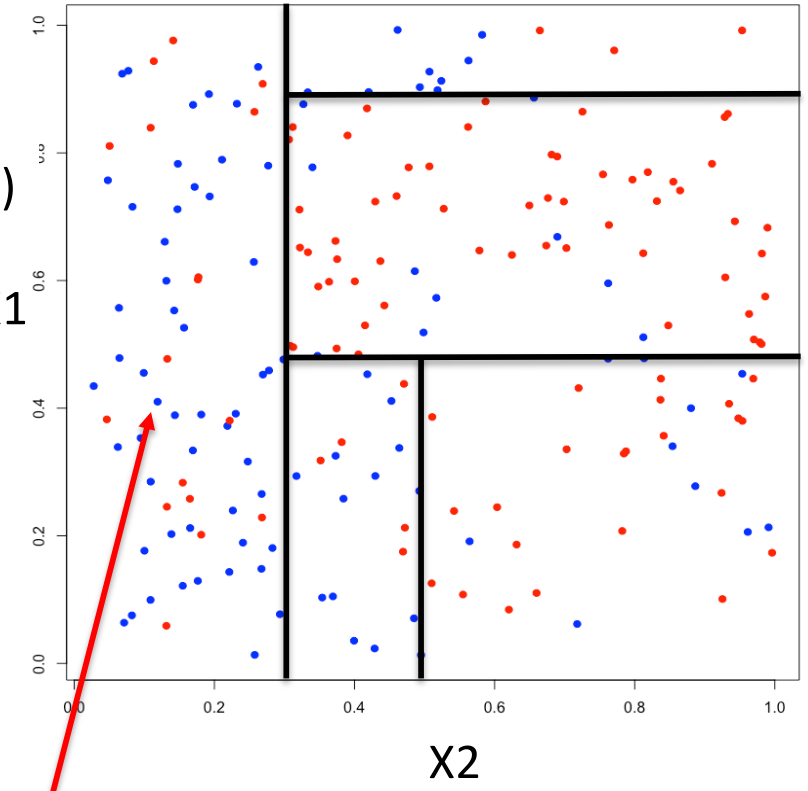
Classification tree for the data



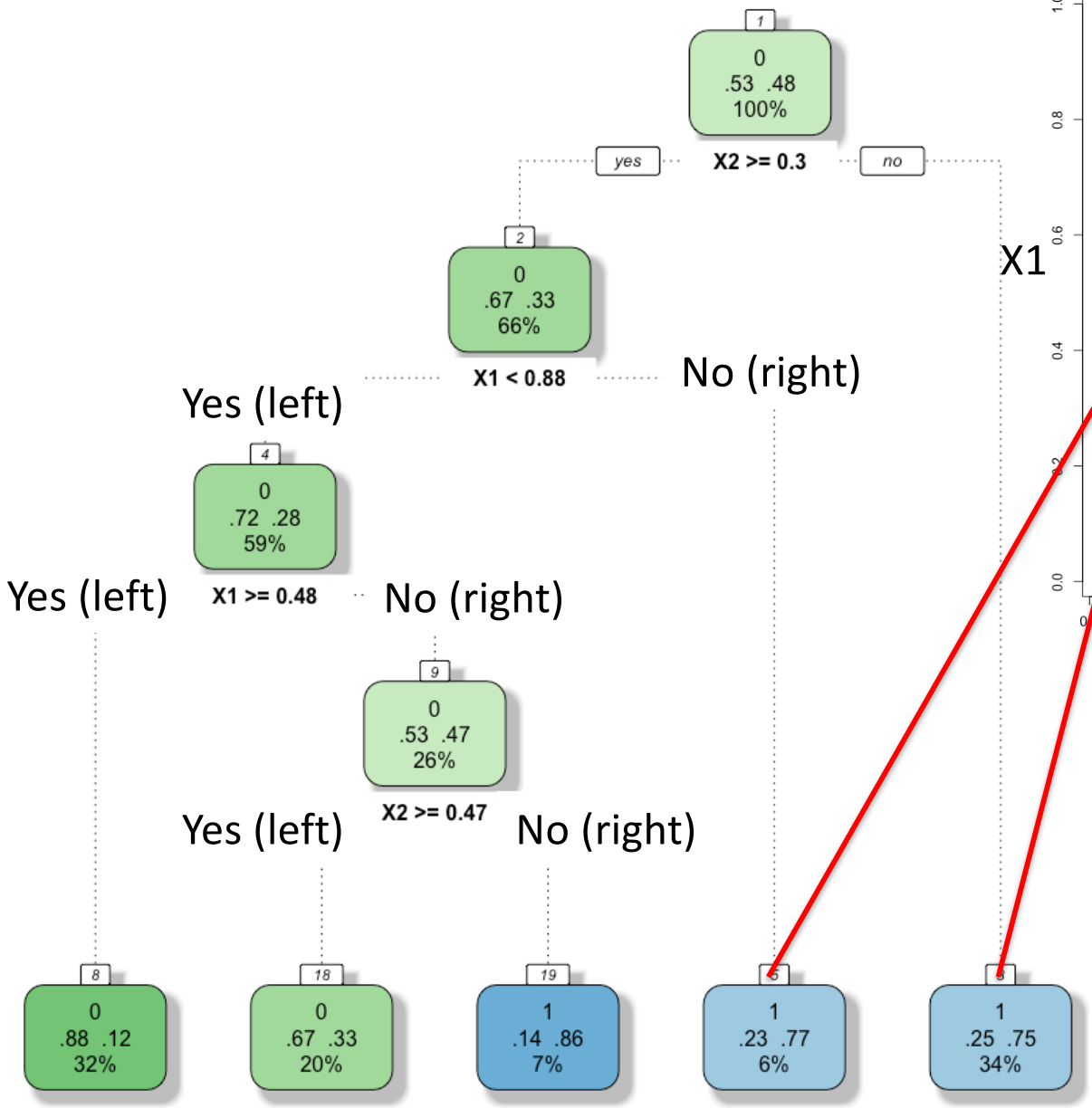


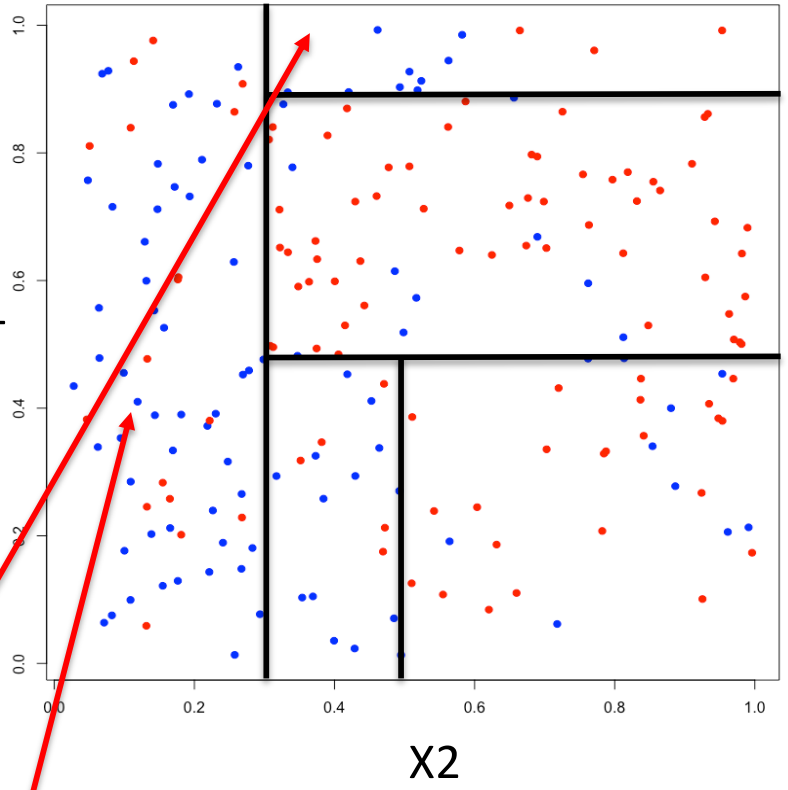
Classification tree for the data



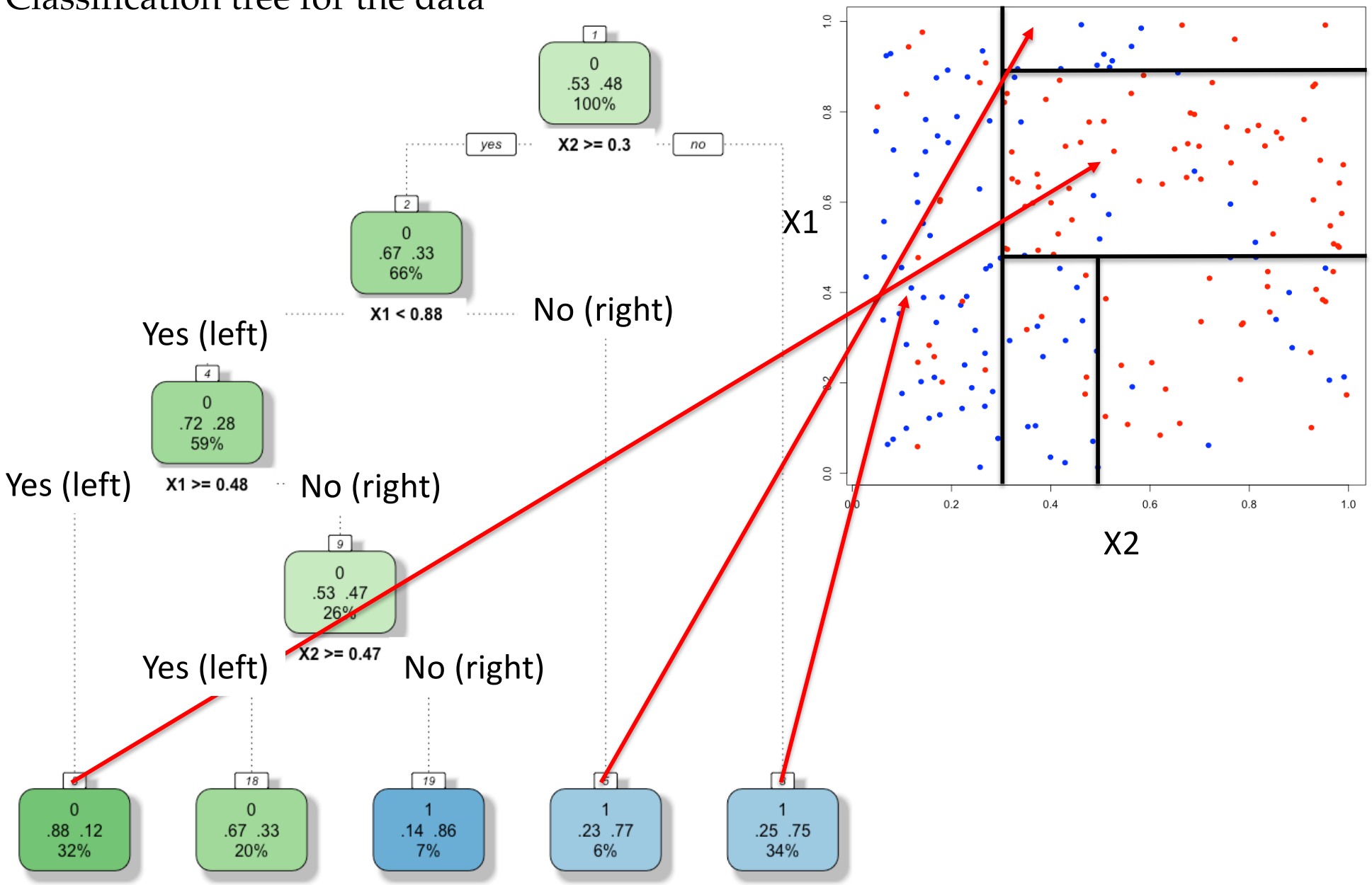


Classification tree for the data

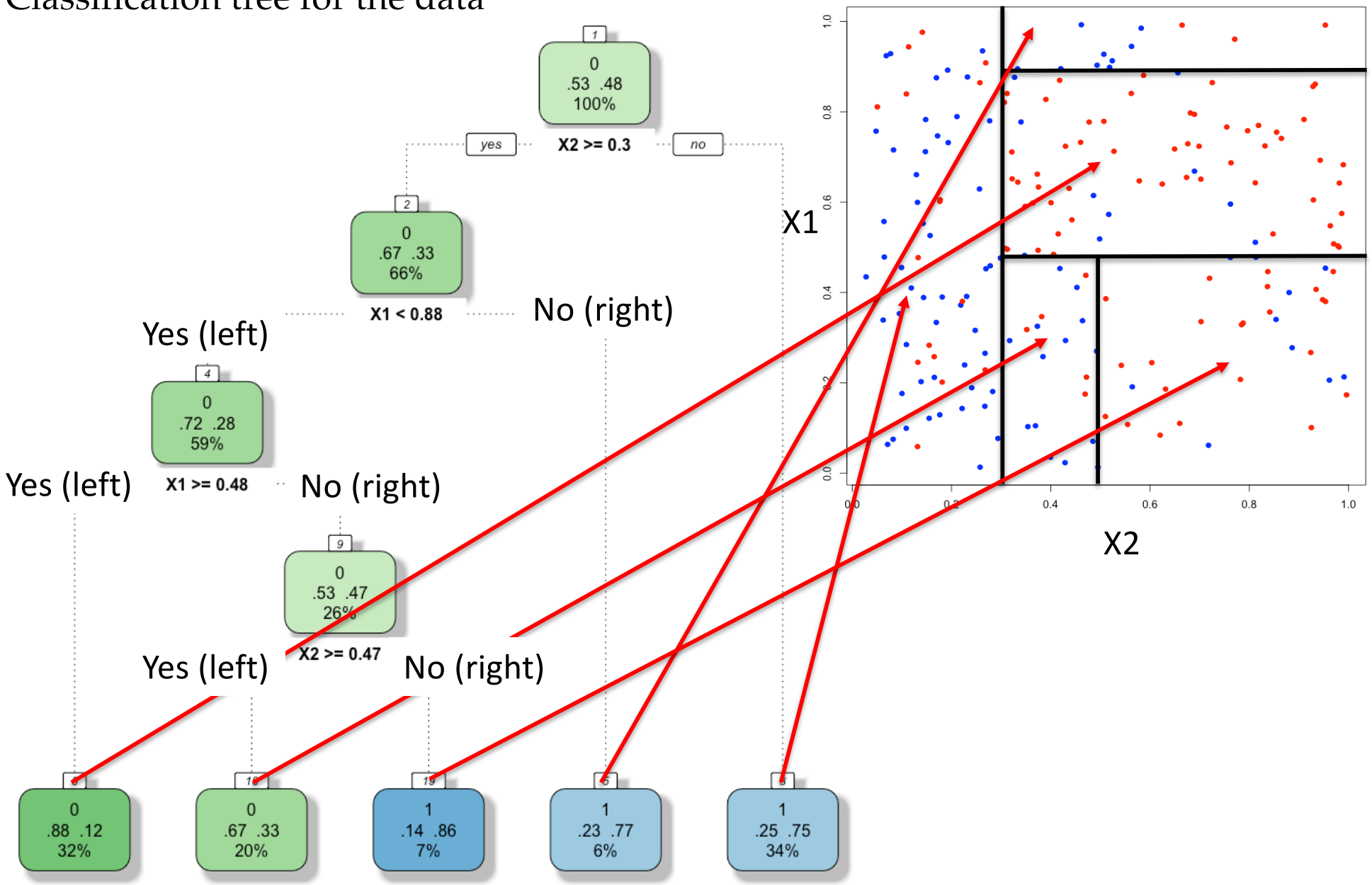




Classification tree for the data



Classification tree for the data



Den Boer et al. 2009

A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study

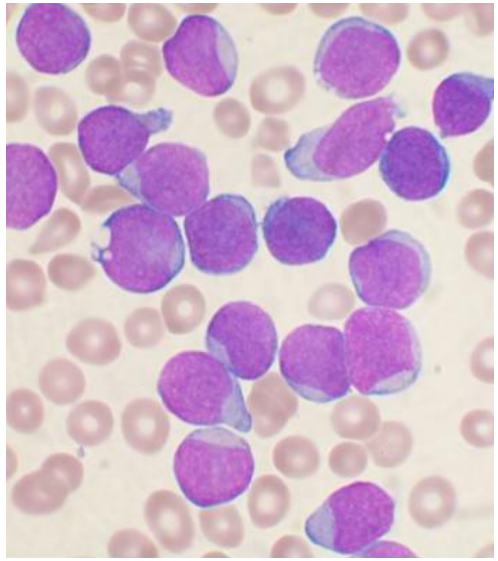
Monique L Den Boer, Marjon van Slegtenhorst*, Renée X De Menezes, Meyling H Cheok, Jessica G C A M Buijs-Gladdines, Susan T C J M Peters, Laura J C M Van Zutven, H Berna Beverloo, Peter J Van der Spek, Gaby Escherich†, Martin A Horstmann†, Gritta E Janka-Schaub†, Willem A Kamps‡, William E Evans, Rob Pieters‡*

Background - In childhood acute lymphoblastic leukemia (ALL) genetic subtypes are recognized that determine the risk-group for further treatment. However, 25% of precursor B-cell ALL (most common type of ALL) are currently genetically unclassified and have an intermediate prognosis. The present study used genome-wide strategies to reveal new biological insights and advance the prognostic classification of childhood ALL.

Lancet Oncol 2009; 10: 125–34

Published Online
January 9, 2009
DOI:10.1016/S1470-

The expression of 22283 genes across 190 patients were considered

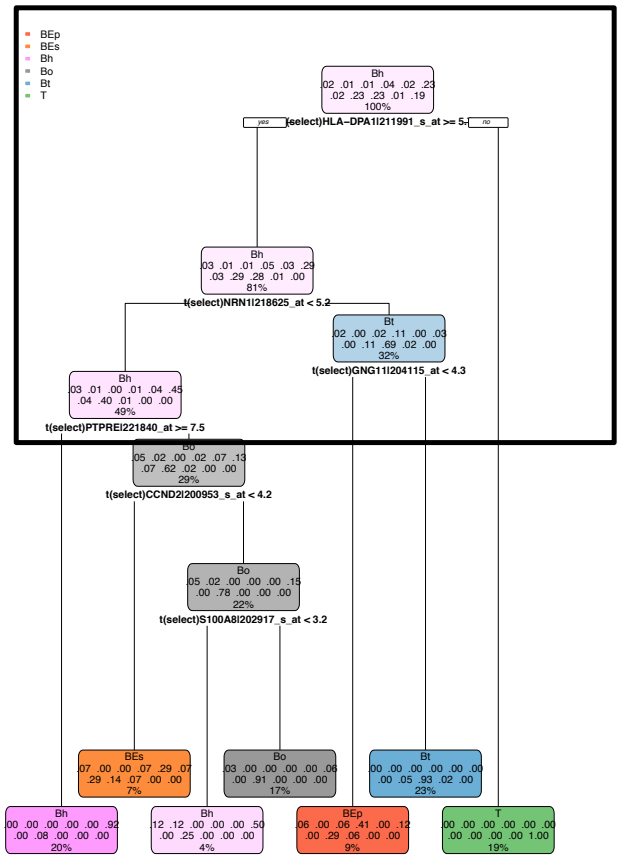


Several subgroups of childhood acute lymphoblastic leukaemia (ALL) have unfavourable prognosis

BEp
BEs
Bh
Bo
Bt
T

Bo (B-other are about 25% of patients and remains unclassified)

Can we improve prognosis based on gene expression?



The expression of 22283 genes across 190 patients were considered to build the model (calibration); 107 independent patients were predicted by the model (validation). The model was 87.7% accurate!

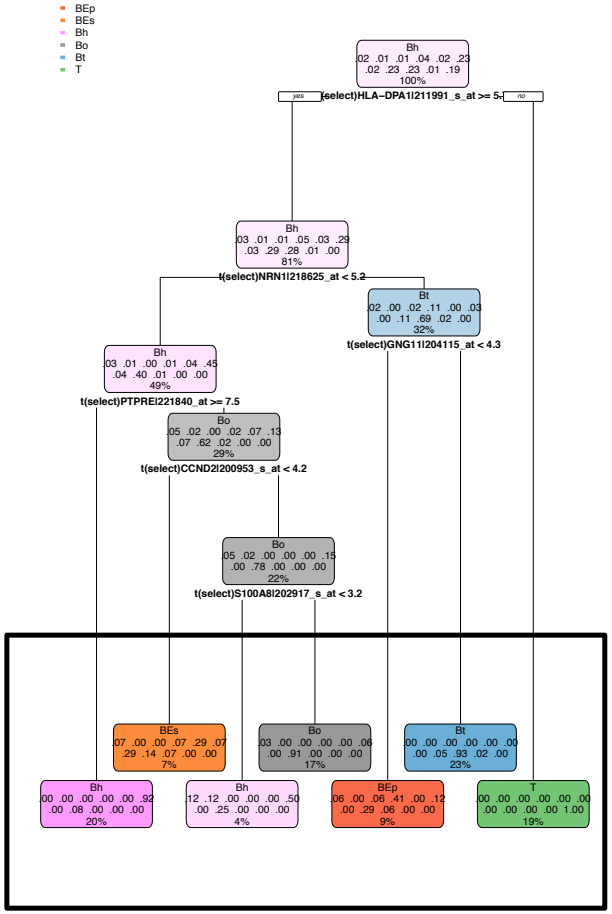
A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study

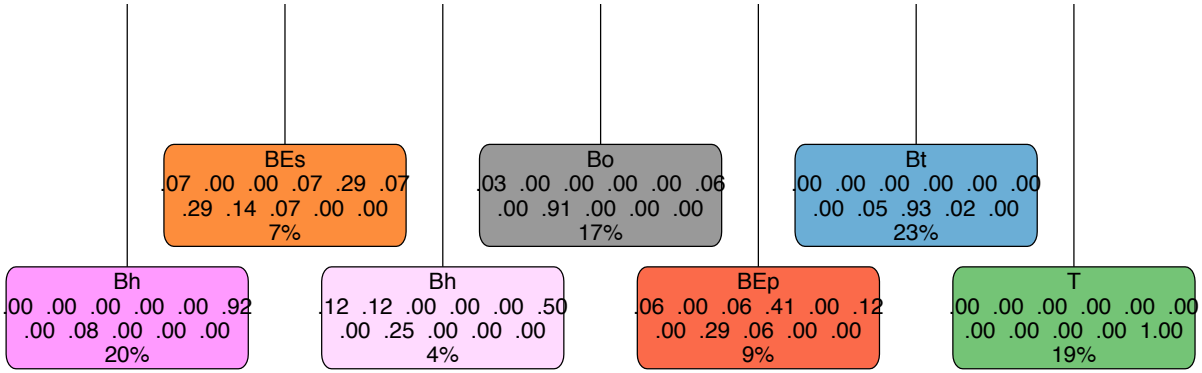
Monique L Den Boer*, Marjon van Slegtenhorst*, Renée X De Menezes, Meyling H Cheok, Jessica G C A M Buijs-Gladdines, Susan T C J M Peters, Laura J C M Van Zutven, H Berna Beverloo, Peter J Van der Spek, Gaby Escherich†, Martin A Horstmann†, Gritta E Janka-Schaub†, Willem A Kamps‡, William E Evans, Rob Pieters‡

Summary

Background Genetic subtypes of acute lymphoblastic leukaemia (ALL) are used to determine risk and treatment in children. 25% of precursor B-ALL cases are genetically unclassified and have intermediate prognosis. We aimed to use a genome-wide study to improve prognostic classification of ALL in children.

Lancet Oncol 2009; 10: 125–34
Published Online January 9, 2009
DOI:10.1016/S1470-





New patients can have their B-ALL types classified according to this model

Growing a tree

- There are many ways of building CARTs and many complex and advanced ways of doing it.

- Search and establishing hierarchy among variables Partition values of a variable: X<=c and X>c for ”all” possible c values. Compare fit using (for example) *pseudo* R^2 (correlation between predicted and observed).

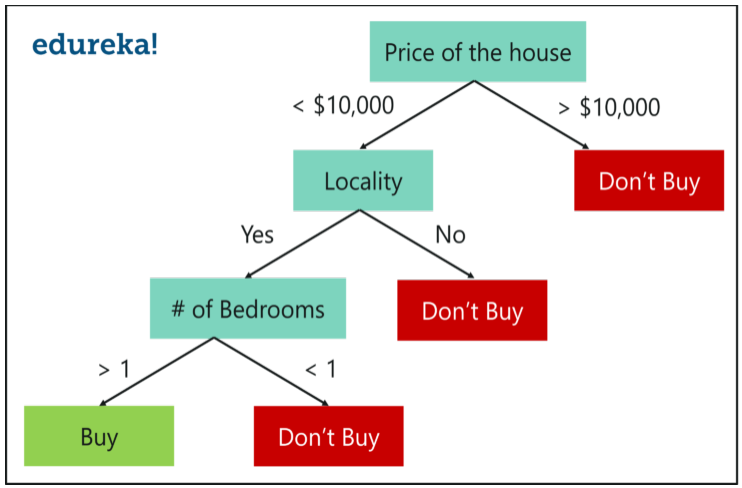
- Order of variables are important and may influence the tree – bagging & random forests deal with this issue via building multiple trees (bootstrap) and selecting trees that maximize R^2 or average trees.

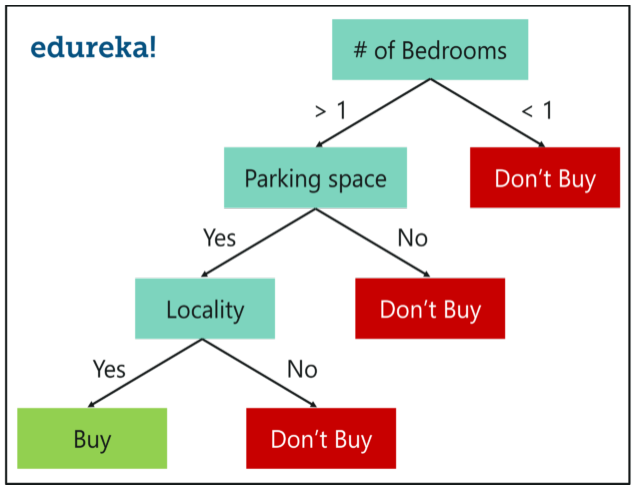
More complex models for building trees
bagging: bootstrapping objects but keeping all predictors

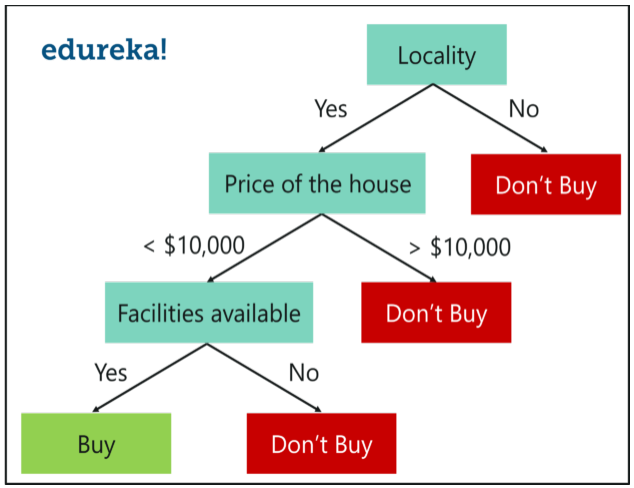
Model for determining factors that influenced house purchasing

Data subsample 1

Data subsample 2……. Data subsample 1000

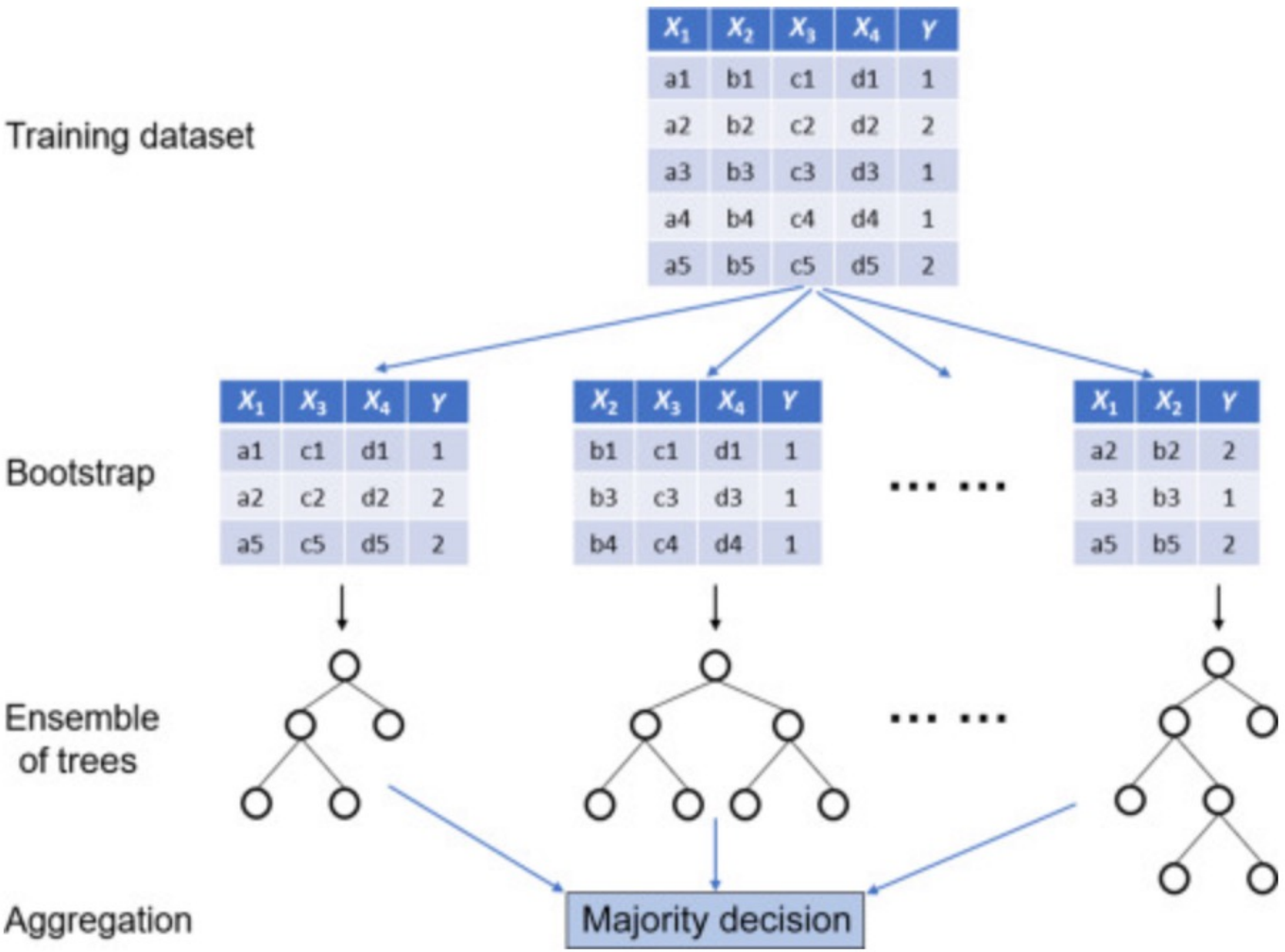






Build separate trees for each subsample (bootstrap) of houses. For each house, make a separate prediction for each tree (buy/not buy). Then make a decision for that house based on the majority rule (if the majority of trees let you the decision to buy that house), then buy it). This is called majority rule. In regression trees (continuous responses), we take the average of the predicted value for any observation of interest.

More complex models for building trees
Random forest: bootstrapping predictors



Classification and Regression trees

- Presenting a complex model as a tree that is easy to interpret is the key why CART became such a popular method.

- "There is no need to understand statistics to fit and interpret CARTs"…but one should understand the basis to feel comfortable with the method and outputs.

- It treats data without a mechanism (as in OLS regressions, GLMs, etc); the thinking is in the algorithm and not about the mechanism that generated the response variable.