# Classification and Regression Trees

Sapna Sharma

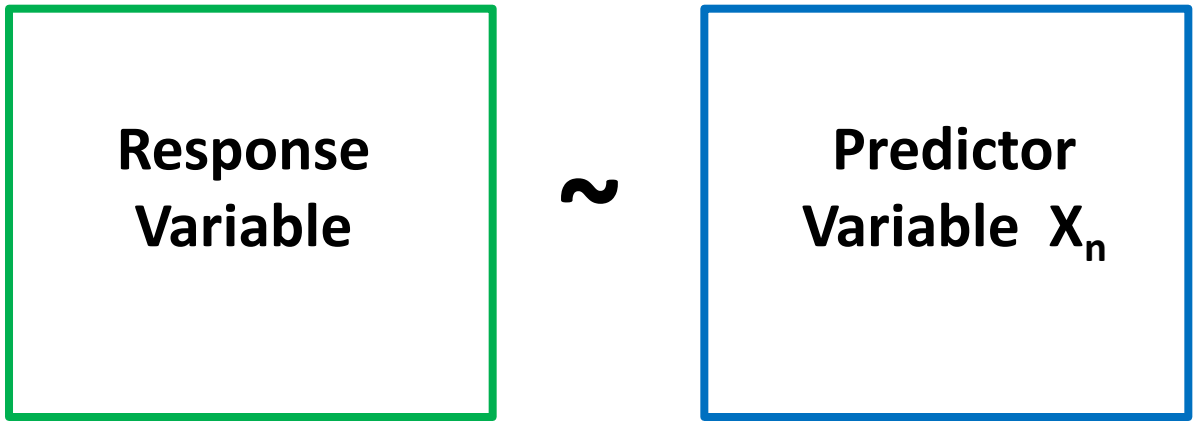
York University

YORK
UNIVERSITÉ
UNIVERSITY

# Road map

- Classification Trees
- Regression Trees
- Multivariate Regression Trees

# How do we develop a model?



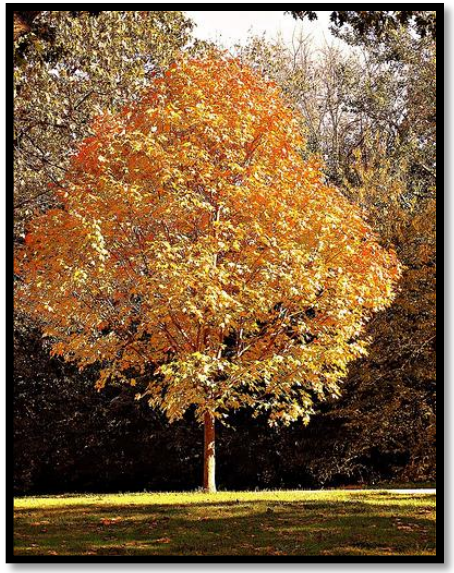
# Response variables

Species presence/ absence
Species abundance
Toxin concentrations
Temperatures

**Response Variable**

~

Predictor Variable $X_n$



# Predictor variables

Response Variable

~

Predictor Variable $X_n$

# Types of response data

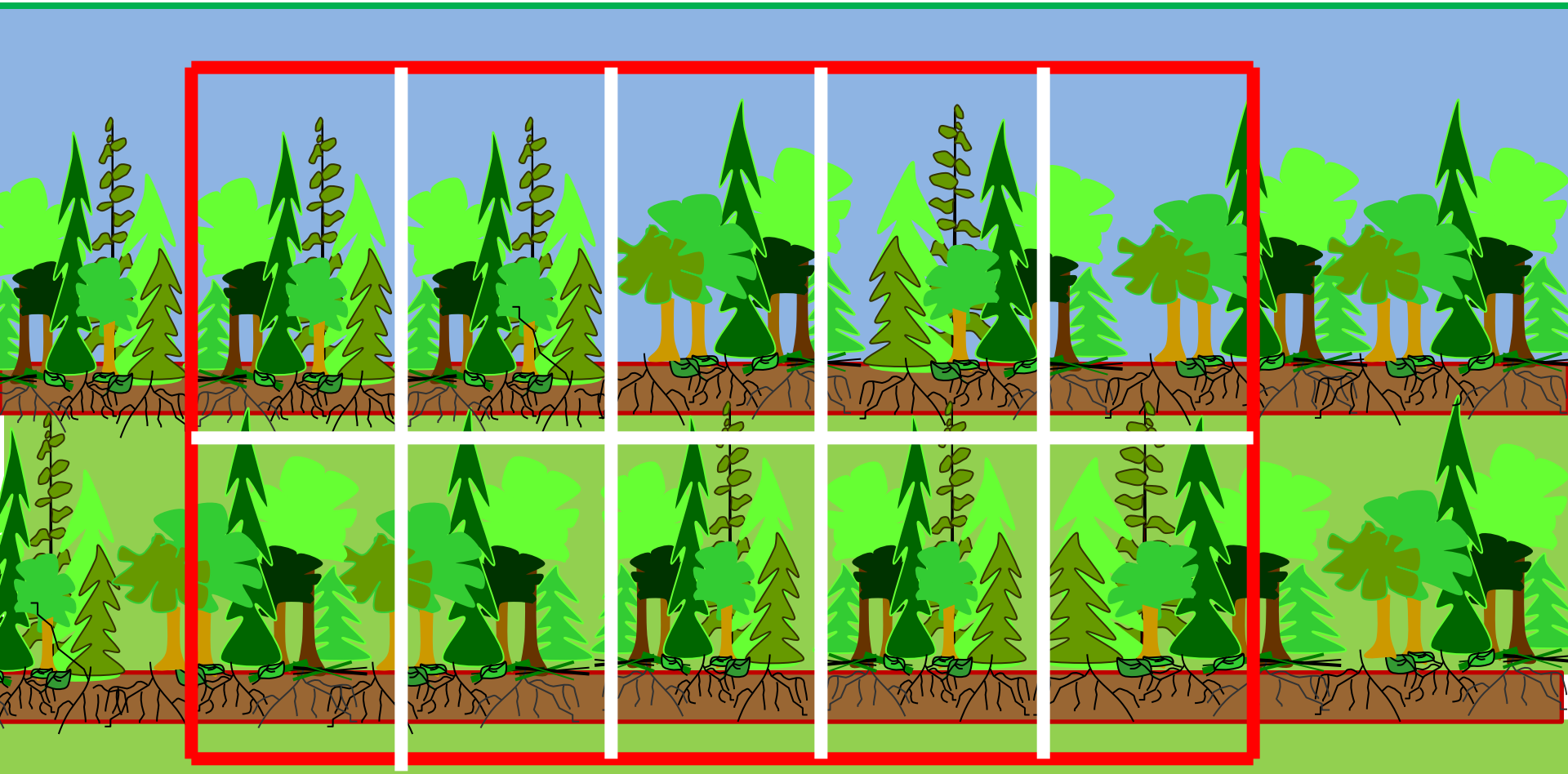
- Binary data
  - 0 or 1
- Continuous data
  - 0 to ∞
- Multivariate data
  - >1 Response variable (0/1 or 0- ∞)

***What do your response and predictor variables look like?***

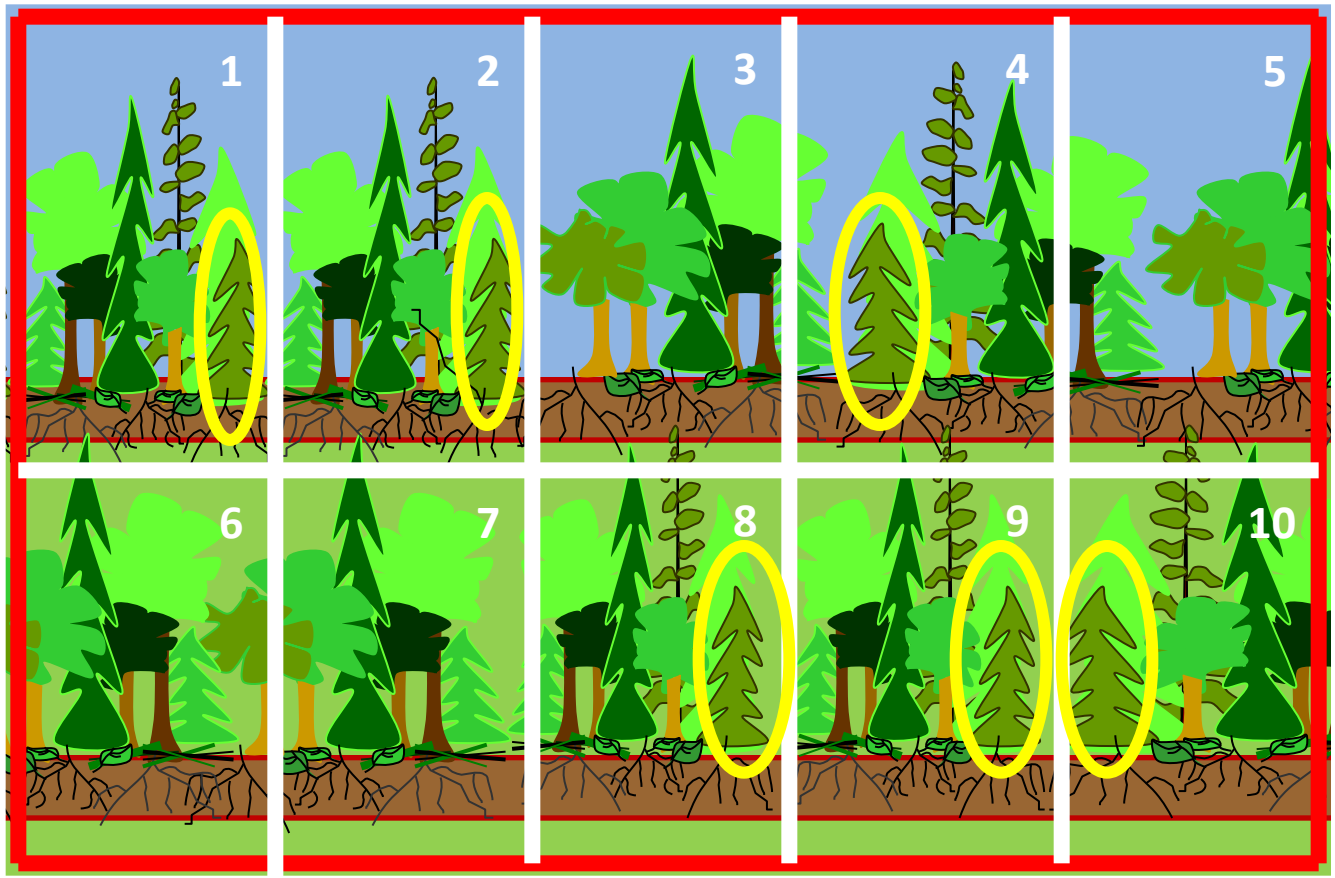
# Road map

- **Classification Trees:**
  - **Response variable is BINARY**
- Regression Trees
- Multivariate Regression Trees

# Sampling a forest



# Sampling the forest for Pine

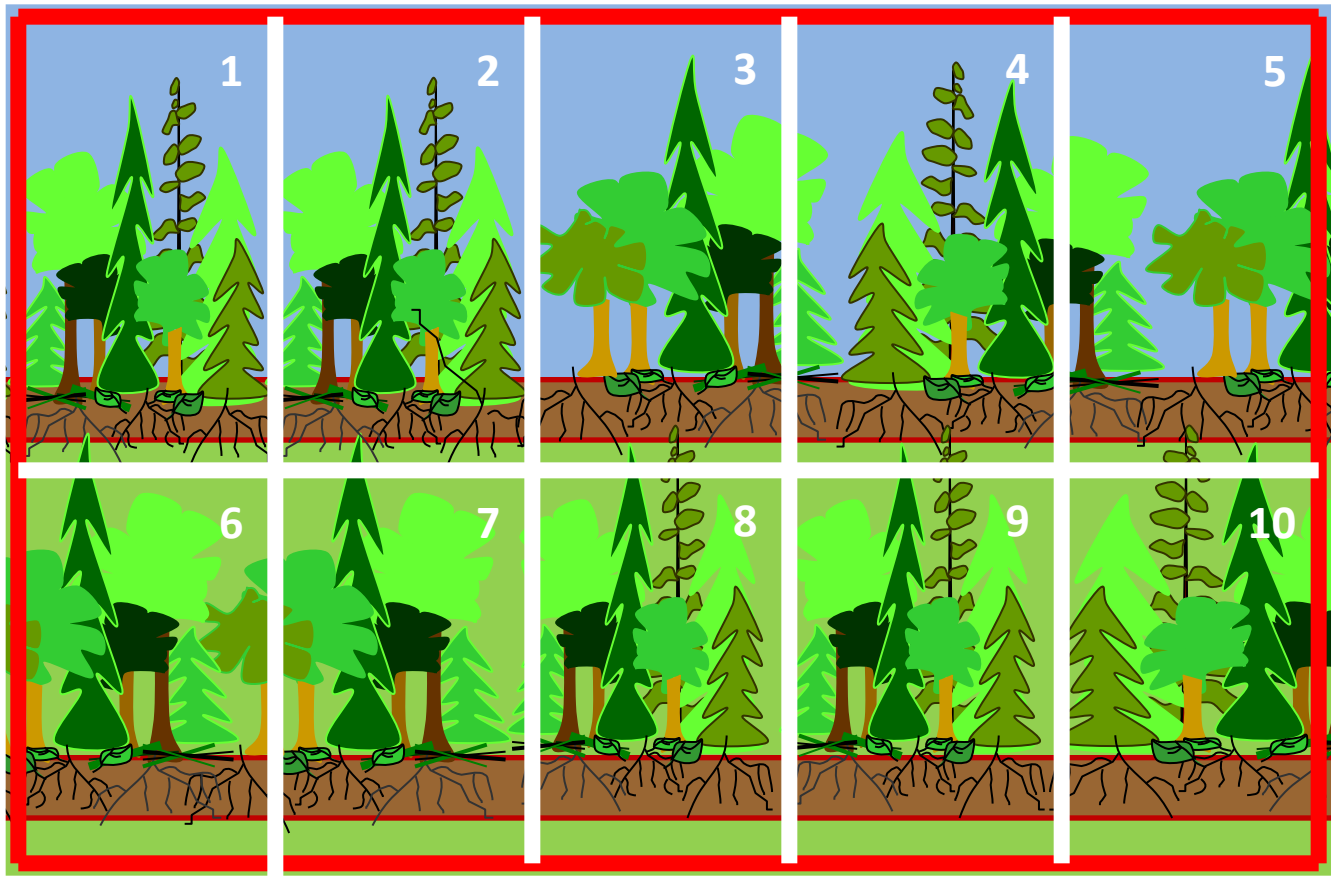



In which sites is this tree present?

# Data Table

| Site | Present/Absent |
|---|---|
| 1 | Present |
| 2 | Present |
| 3 | Absent |
| 4 | Present |
| 5 | Absent |
| 6 | Absent |
| 7 | Absent |
| 8 | Present |
| 9 | Present |
| 10 | Present |

# Sampling environmental conditions

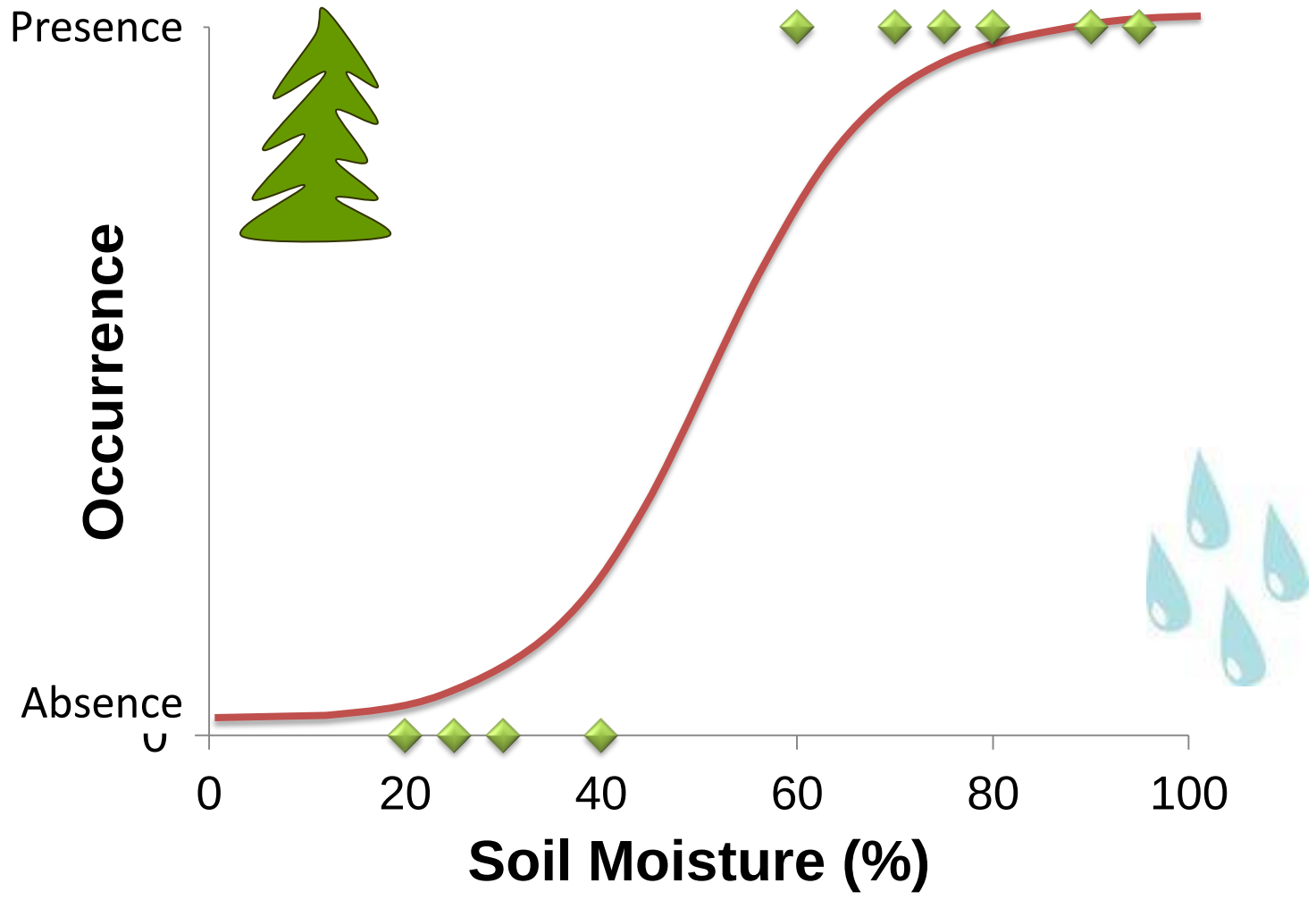



What are the environmental conditions in each site? (e.g., soil moisture)

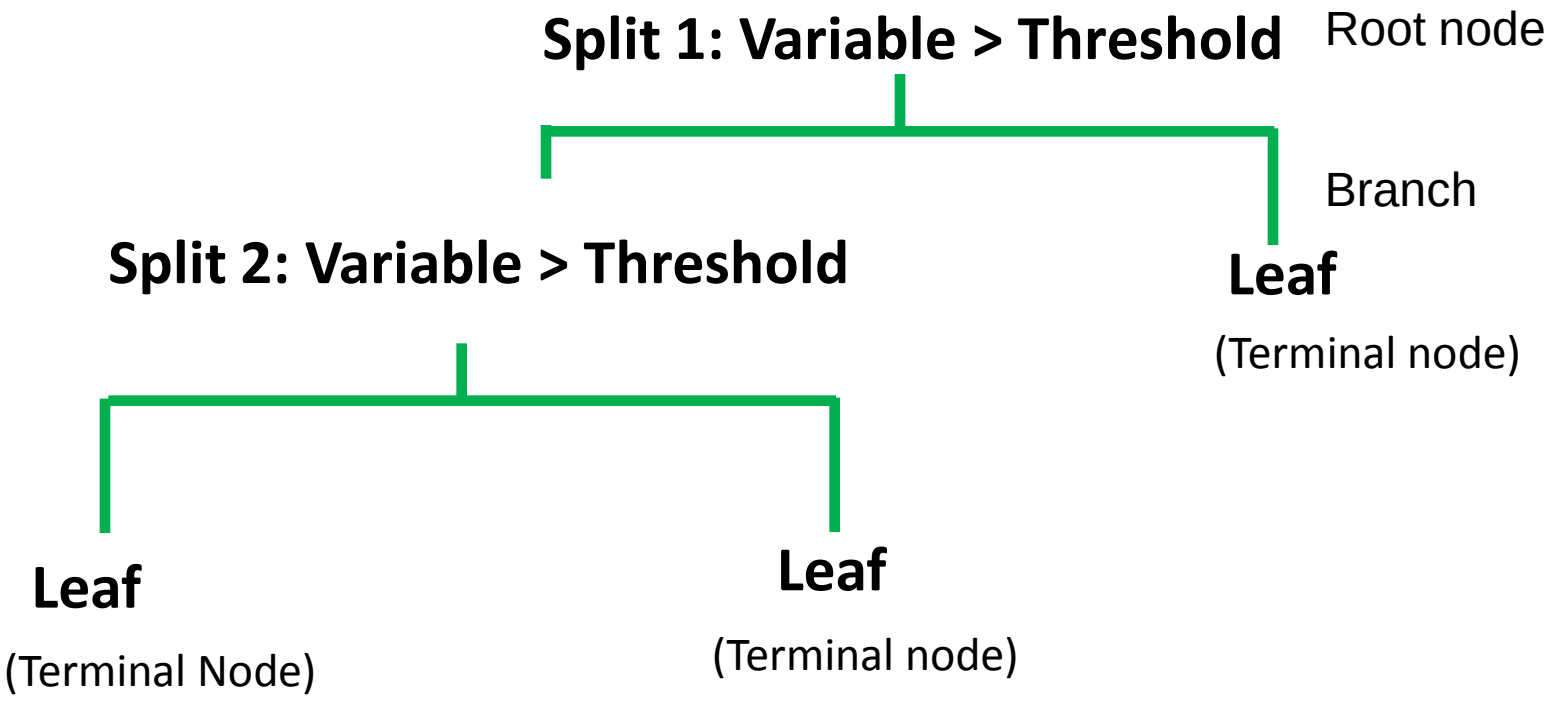
# Data Table

| Site | Present/Absent | Soil Moisture (%) |
|---|---|---|
| 1 | Present | 80 |
| 2 | Present | 75 |
| 3 | Absent | 40 |
| 4 | Present | 95 |
| 5 | Absent | 20 |
| 6 | Absent | 30 |
| 7 | Absent | 25 |
| 8 | Present | 60 |
| 9 | Present | 90 |
| 10 | Present | 70 |

# I. Traditional approach: Logistic Regression



# II. Classification Trees



# II. Classification Trees

**Pine**

**Soil moisture > 35%**

(Presence)

**Soil moisture < 35%**

(Absence)

# Data Table

| Site | Occurrence | Moisture (%) | Soil Type |
|---|---|---|---|
| 1 | Present | 80 | C |
| 2 | Present | 75 | C |
| 3 | Absent | 40 | A |
| 4 | Present | 95 | C |
| 5 | Absent | 20 | B |
| 6 | Absent | 30 | A |
| 7 | Absent | 25 | B |
| 8 | Present | 60 | C |
| 9 | Present | 90 | C |
| 10 | Present | 70 | C |

# II. Classification Trees

**Pine**

**Soil moisture > 35%**

**Soil C**

(Presence)

**Soils A&B**

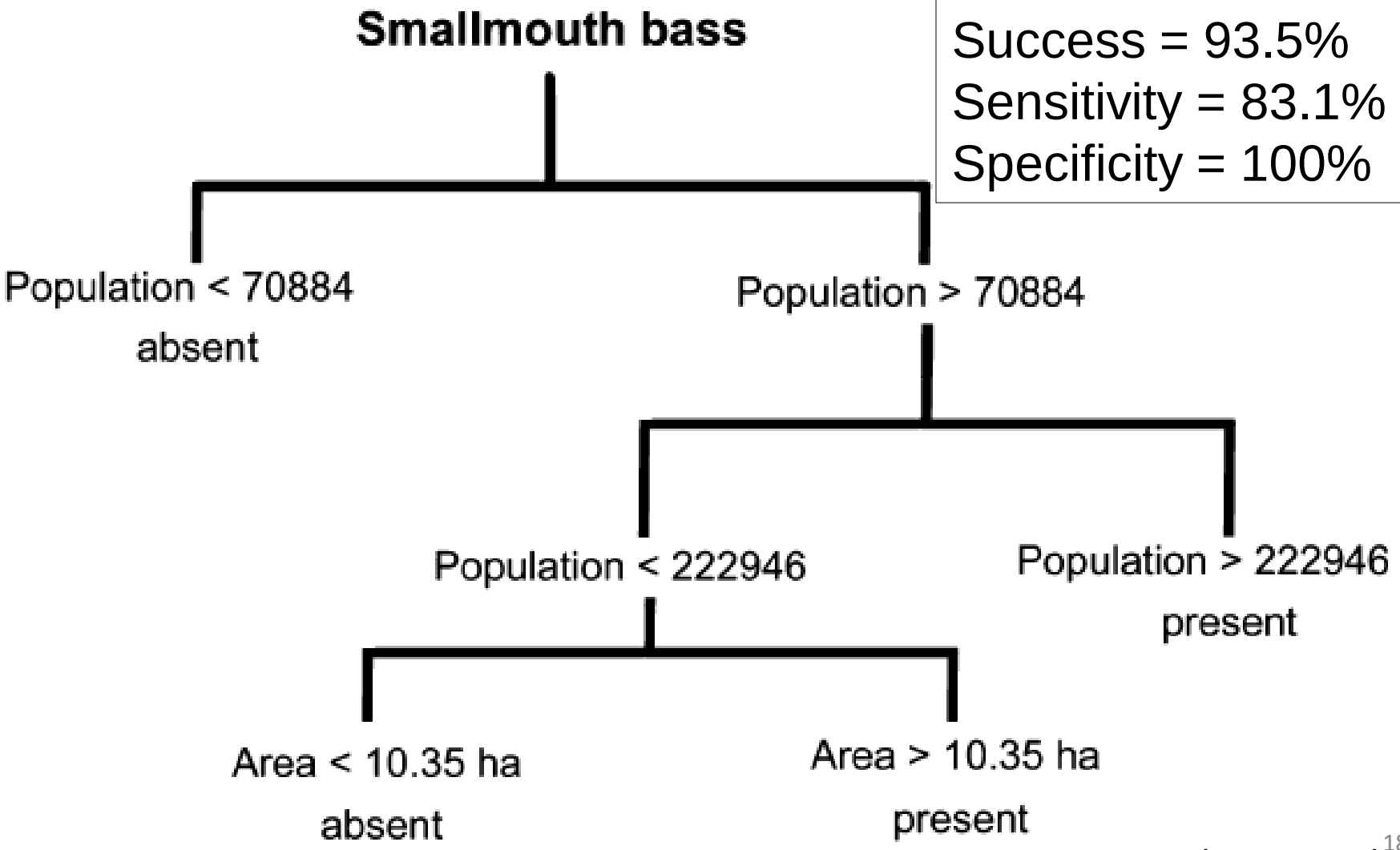
(Absence)

**Soil moisture < 35%**

(Absence)

# II. Characterizing Classification Trees

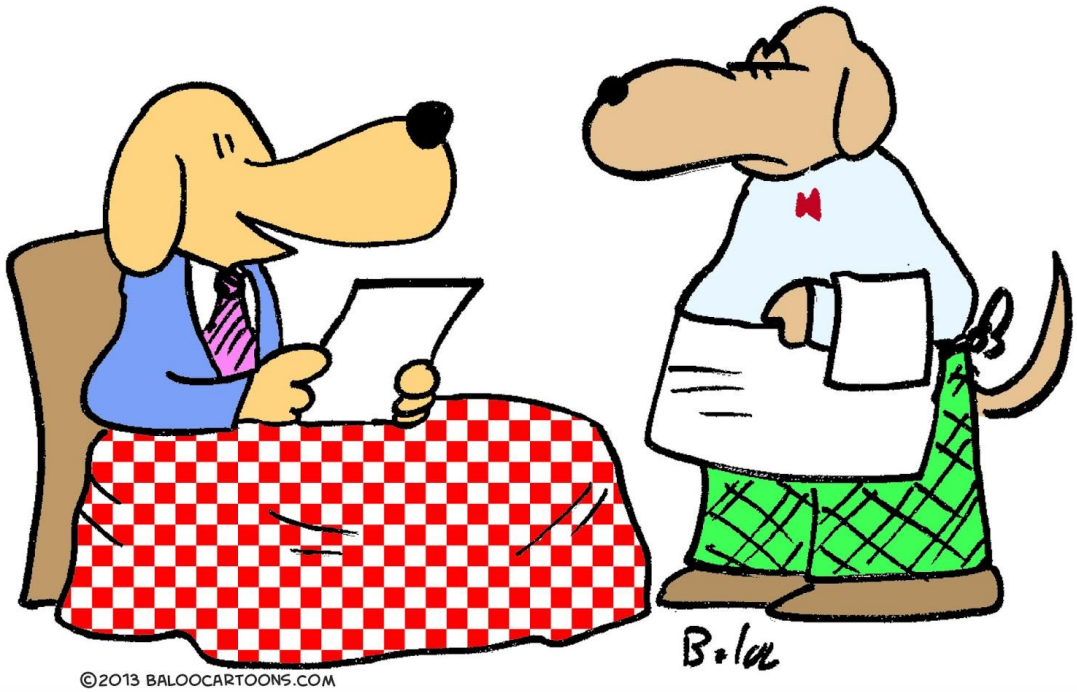
Overall Classification
Success = 93.5%
Sensitivity = 83.1%
Specificity = 100%




Sharma et al. 2009.

# II. Classification Tree Example

- Should you eat lunch at a restaurant or in a cafeteria?




"How's the homework today?"

# II. Summary: Classification Trees

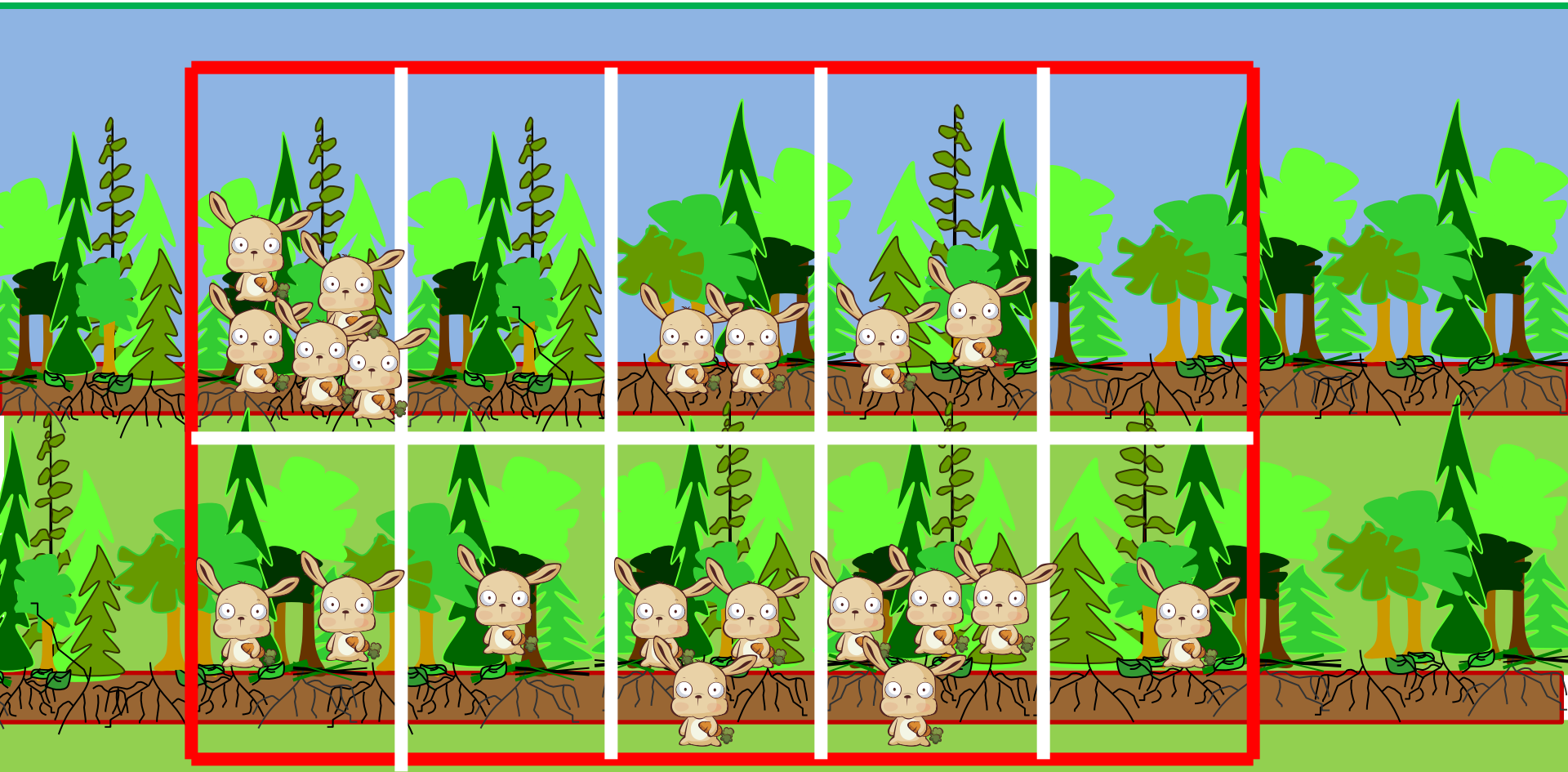
- Response variable is binary
- Can use continuous and/or categorical predictor variables
- Divide data into two groups
- Groups are as mutually exclusive and homogenous as possible
- Minimize misclassification rates at each split

De'ath and Fabricius 2000

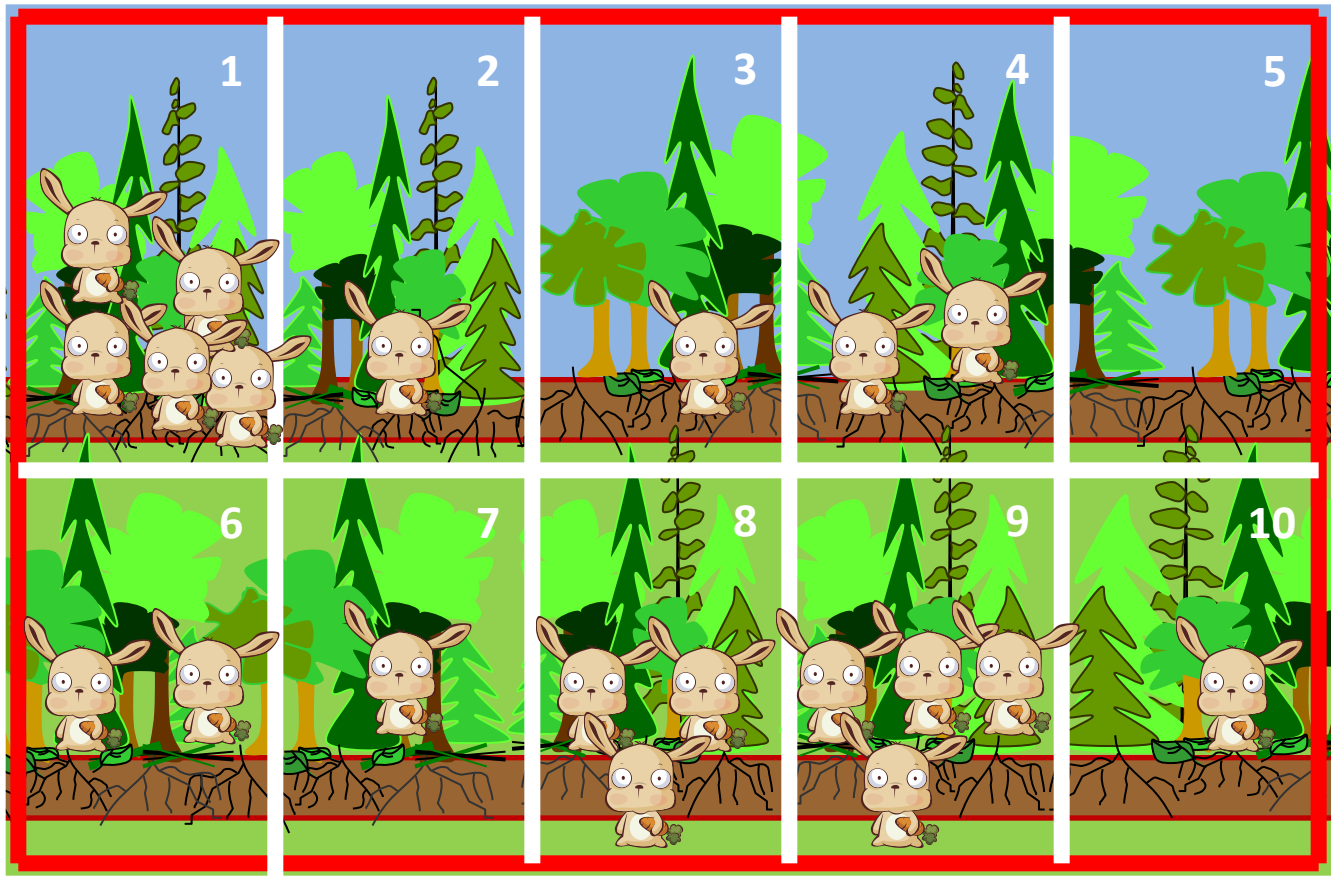
# Road map

- Classification Trees
- **Regression Trees:**
  - **Response variable is CONTINUOUS**
- Multivariate Regression Trees

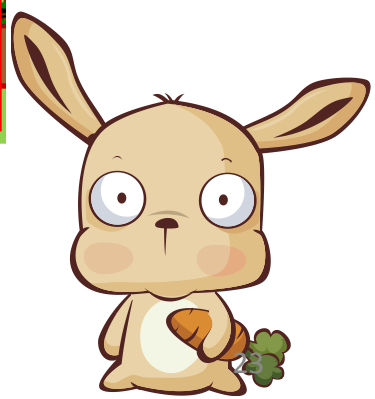
# Sampling the forest



# Sampling the forest for Bunnies



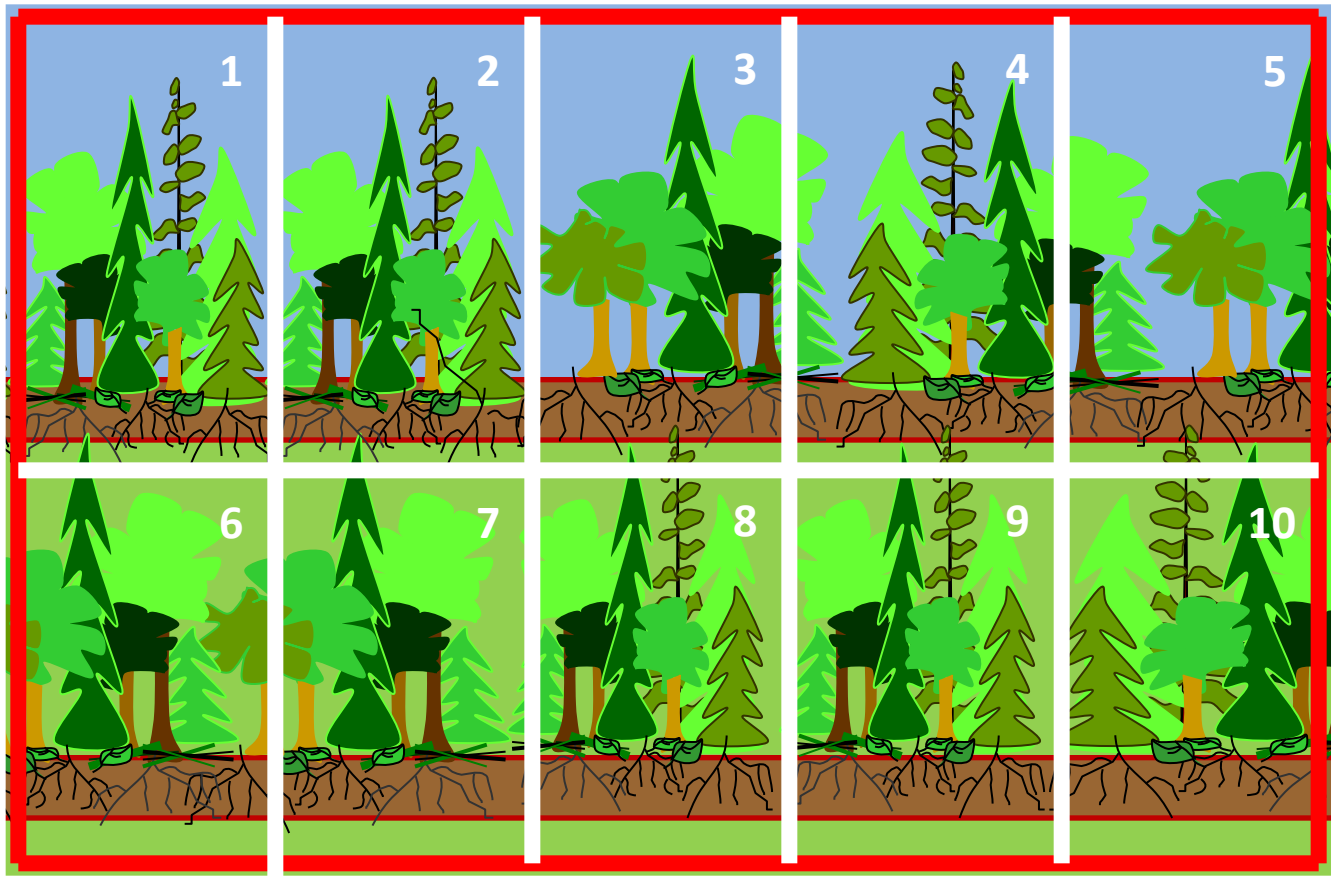

What is the abundance of bunnies in each site?



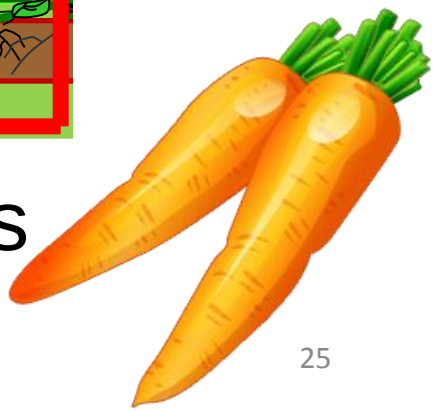
# Data Table

| Site | Bunny Abundance |
|---|---|
| 1 | 5 |
| 2 | 1 |
| 3 | 2 |
| 4 | 2 |
| 5 | 0 |
| 6 | 2 |
| 7 | 1 |
| 8 | 3 |
| 9 | 4 |
| 10 | 1 |

# Sampling environmental conditions



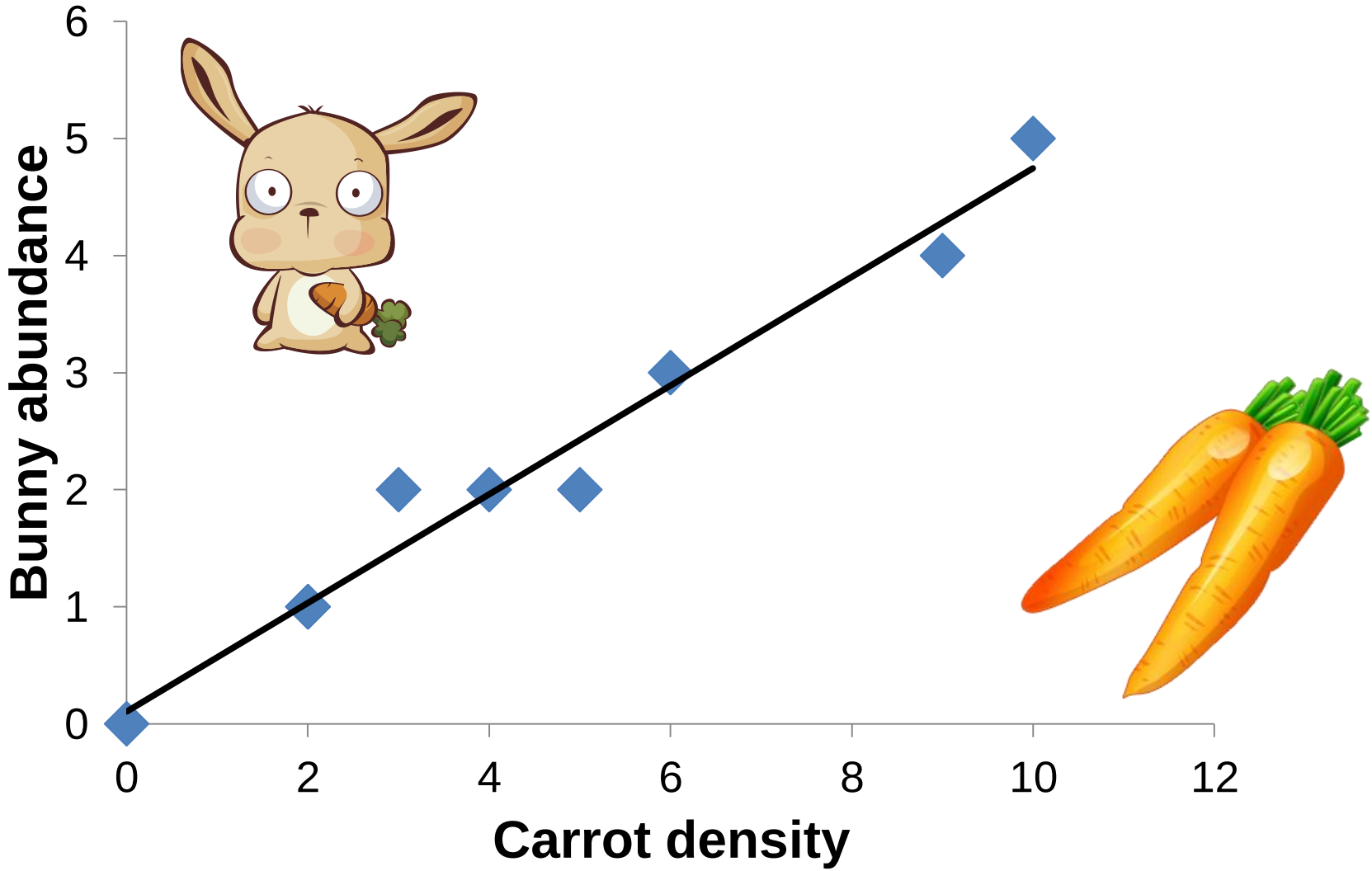

What are the environmental conditions in each site? (e.g., carrot density)



# Data Table

| Site | Bunny Abundance | Carrot density |
|---|---|---|
| 1 | 5 | 10 |
| 2 | 1 | 2 |
| 3 | 2 | 4 |
| 4 | 2 | 3 |
| 5 | 0 | 0 |
| 6 | 2 | 5 |
| 7 | 1 | 2 |
| 8 | 3 | 6 |
| 9 | 4 | 9 |
| 10 | 1 | 2 |

# I. Traditional Approach: Linear regression



# II. Regression Trees

**Bunny abundance**

**Carrot density > 2**

**Carrot density < 2**

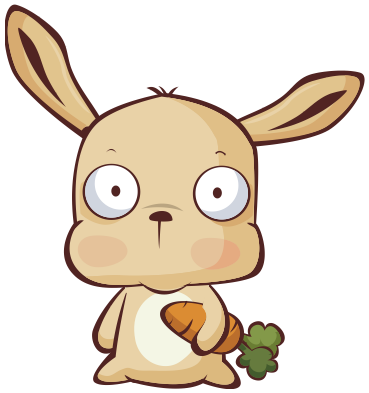
Mean = 1 bunny

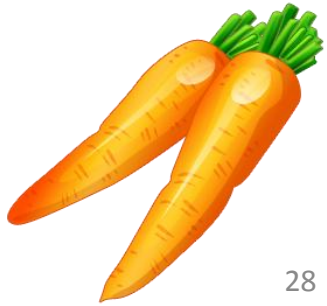
**2<Carrot density>8**

Mean = 2.5 bunnies

**Carrot density > 8**

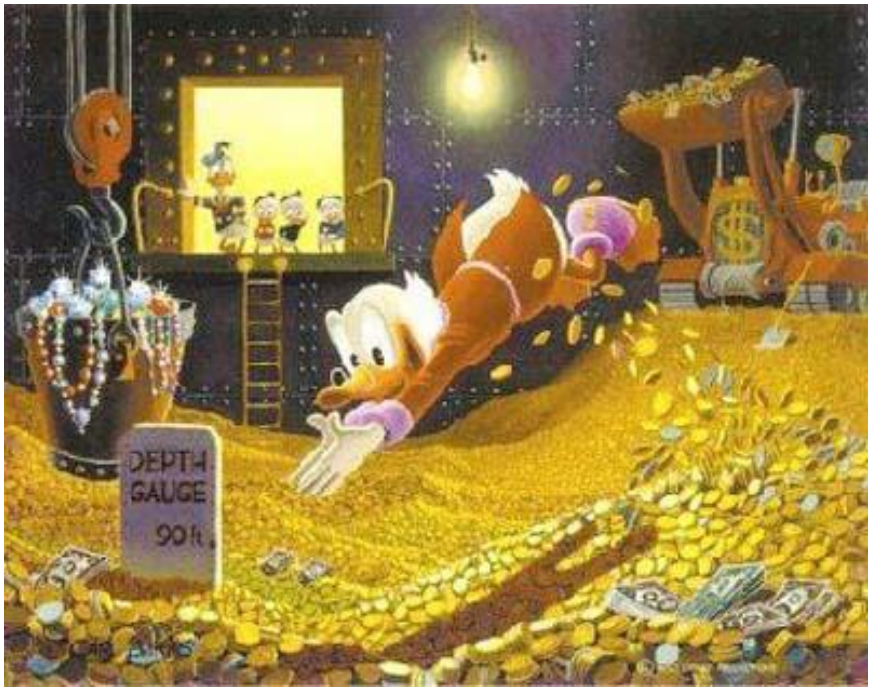
Mean = 4.5 bunnies



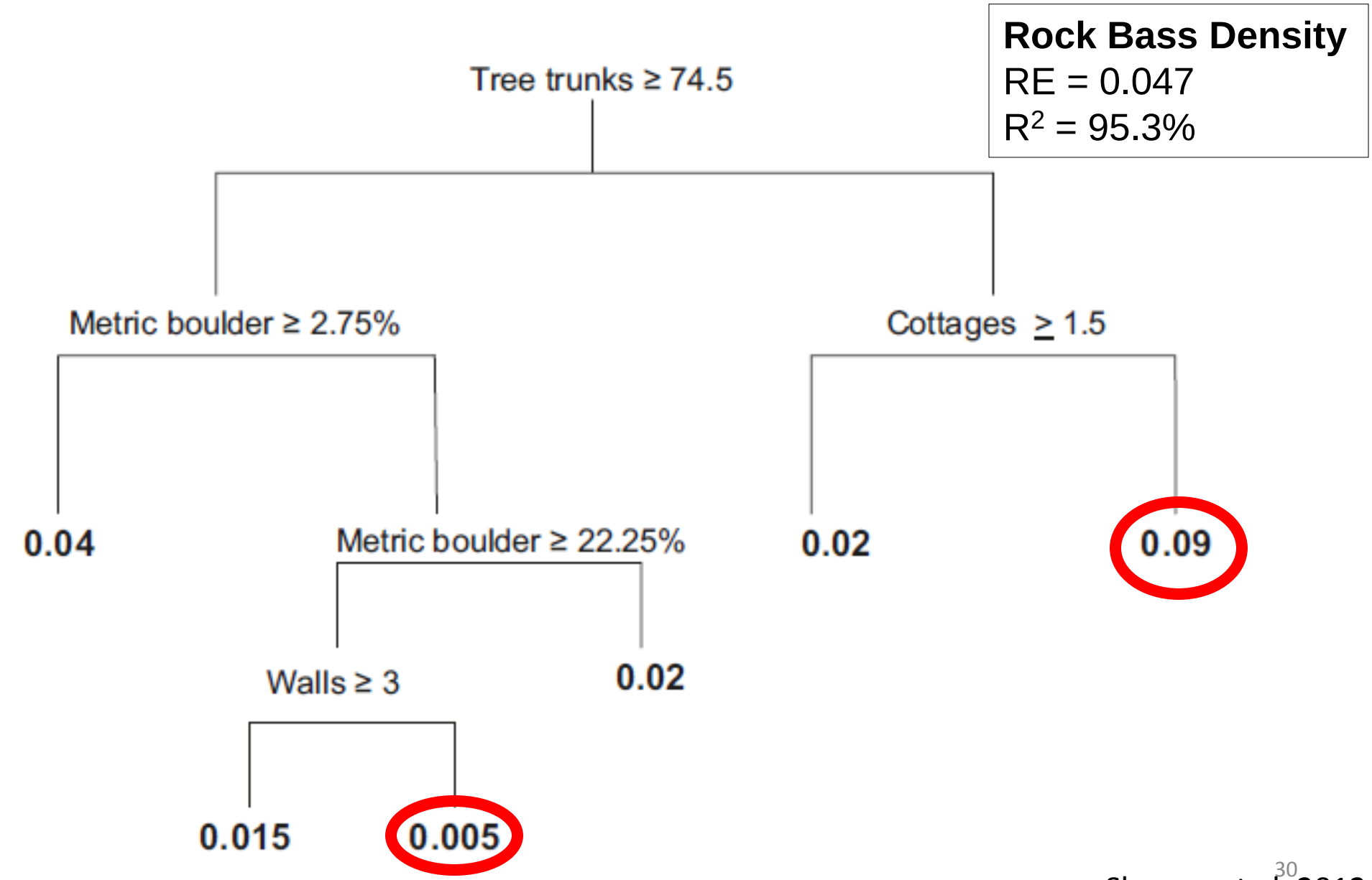


# II. Regression Tree Example

- How much money will you spend at lunch?



# II. Characterizing Regression Trees






Sharma et al. 2012

# II. Regression Trees

- Response variable is continuous
- Can use continuous and/or categorical predictor variables
- Divide data into two groups
- Groups are as mutually exclusive and homogenous as possible
- Minimize sum of square error at each split

De'ath and Fabricius 2000

# CART: Pros & Cons

- ⊞ Complex ecological data
- ⊞ Non-linear relationships
- ⊞ Missing values
- ⊞ Interactions among predictor variables
- ⊟ Over-fitting

De'ath and Fabricius 2000

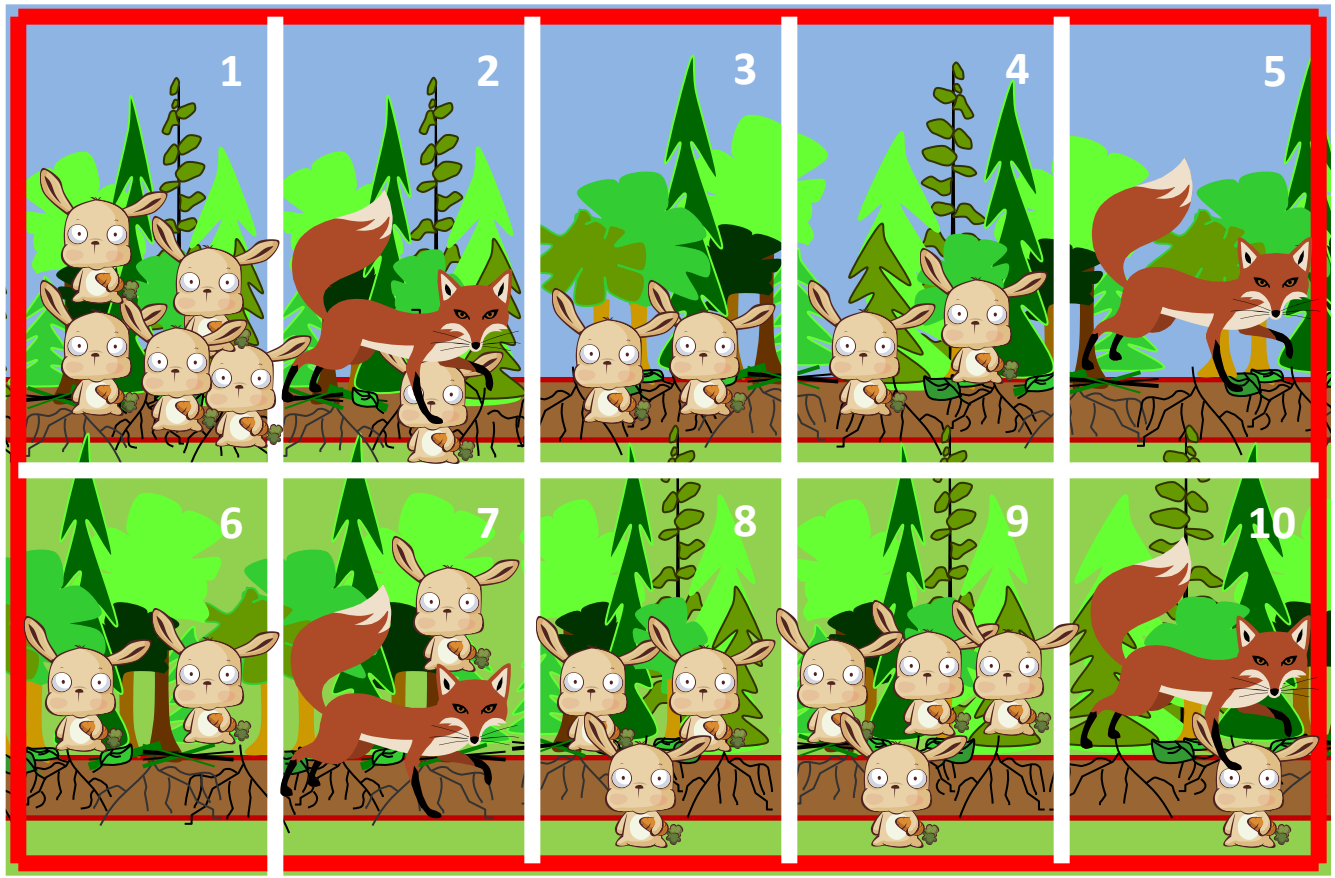
# Road map

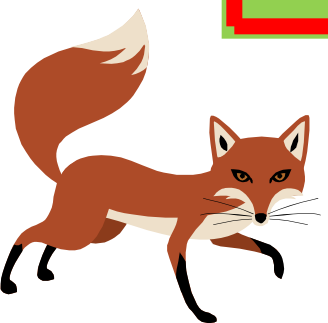
- Classification Trees
- Regression Trees
- **Multivariate Regression Trees:**
  - **>1 Response variables which are CONTINUOUS**

# Multivariate analyses

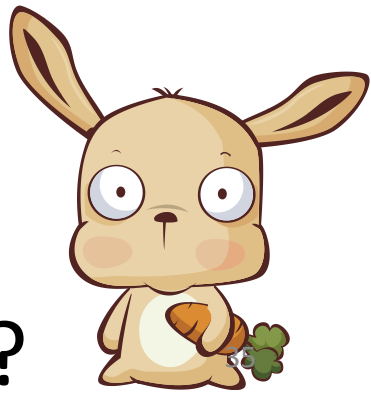
- > 1 Response variable
- Many approaches to use depending upon data structure and your research question
  - Ordinations, e.g., PCA, CA, CCA, RDA etc. (Response variables can be BINARY and/or CONTINUOUS)
  - Multivariate Regression Trees

# Sampling the forest for Bunnies and Foxes



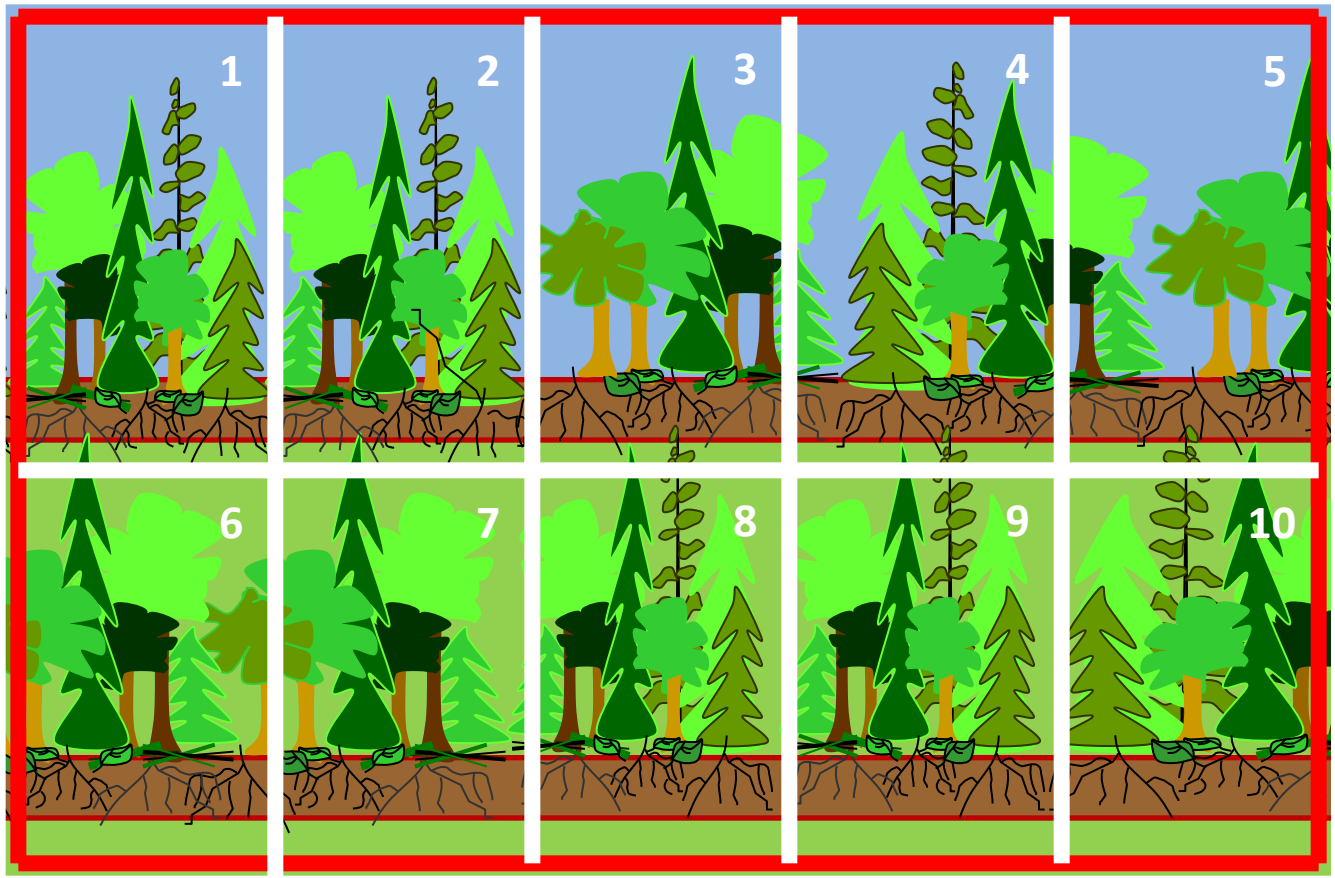
What is the abundance of bunnies and foxes in each site?



# Data Table

| Site | Bunny Abundance | Fox Abundance |
|---|---|---|
| 1 | 5 | 0 |
| 2 | 1 | 1 |
| 3 | 2 | 0 |
| 4 | 2 | 0 |
| 5 | 0 | 1 |
| 6 | 2 | 0 |
| 7 | 1 | 1 |
| 8 | 3 | 0 |
| 9 | 4 | 0 |
| 10 | 1 | 1 |

# Sampling environmental conditions

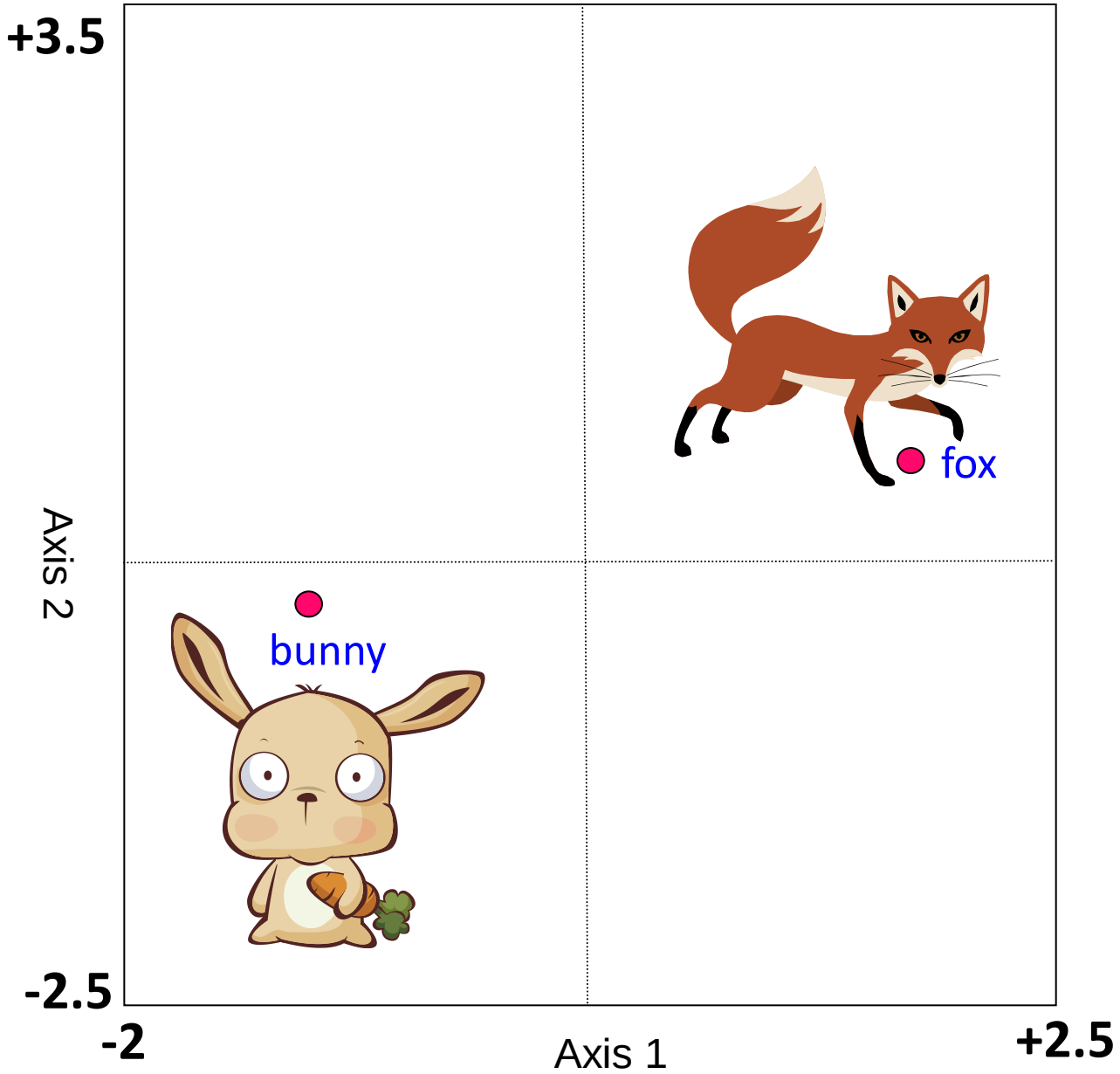



What are the environmental conditions in each site? (e.g., carrot density and soil moisture)

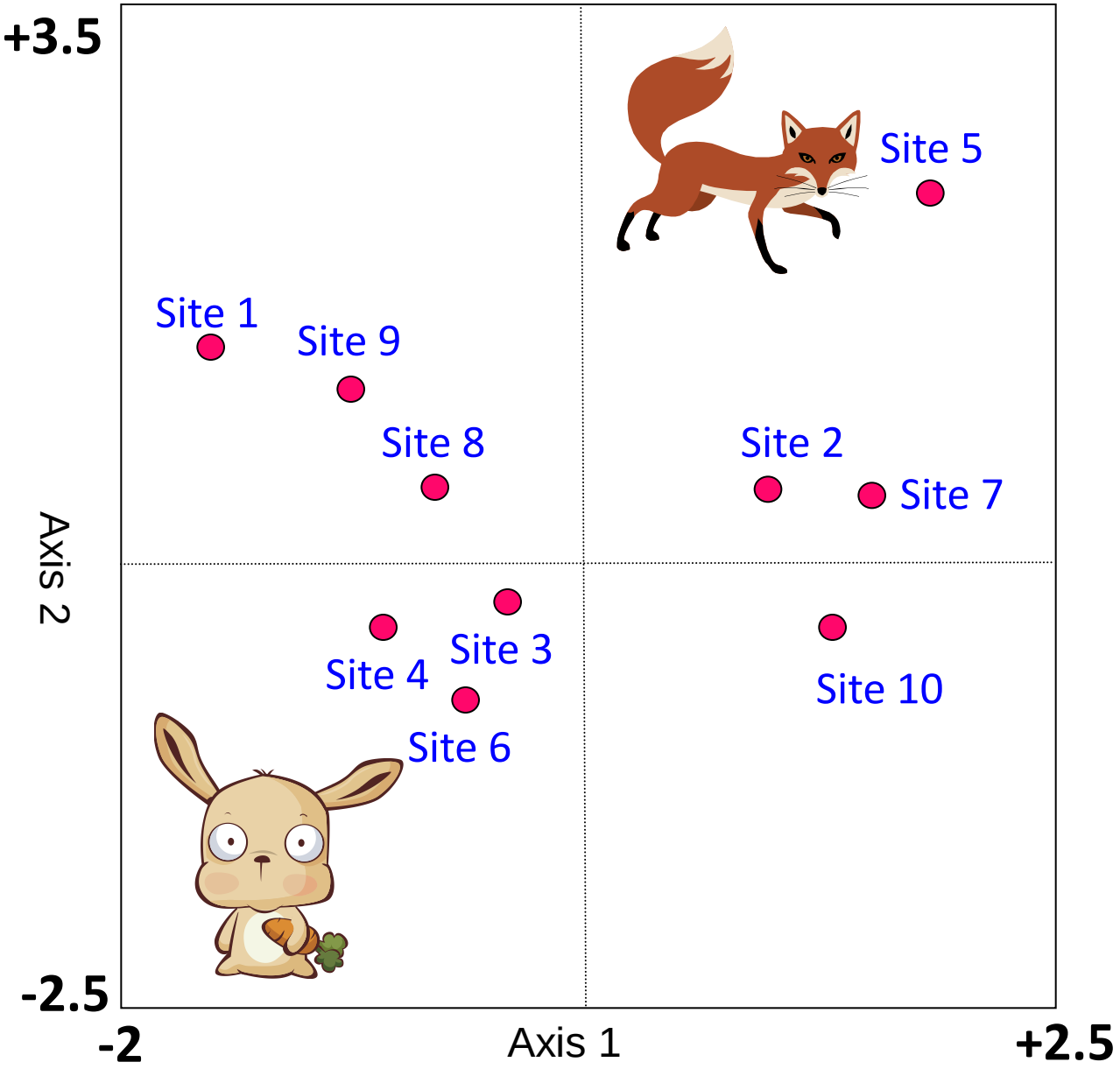
# Data Table

| Site | Bunny | Fox | Carrots | Soil Moisture |
|---|---|---|---|---|
| 1 | 5 | 0 | 10 | 80 |
| 2 | 1 | 1 | 2 | 75 |
| 3 | 2 | 0 | 4 | 40 |
| 4 | 2 | 0 | 3 | 95 |
| 5 | 0 | 1 | 0 | 20 |
| 6 | 2 | 0 | 5 | 30 |
| 7 | 1 | 1 | 2 | 25 |
| 8 | 3 | 0 | 6 | 60 |
| 9 | 4 | 0 | 9 | 90 |
| 10 | 1 | 1 | 2 | 70 |

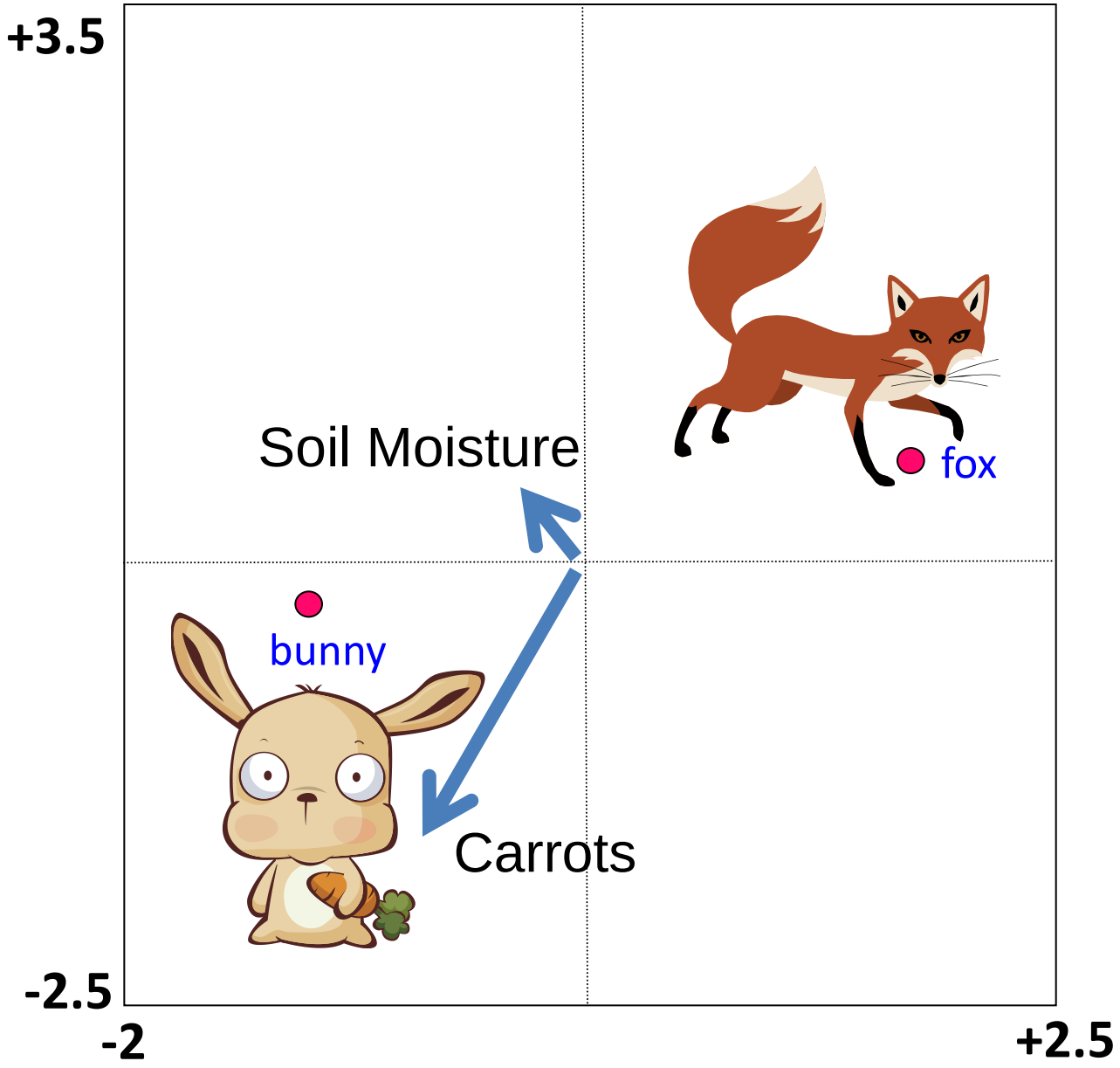
# I. Ordinations



# I. Indirect gradient analysis



# I. Direct Gradient Ordination



# II. Multivariate Regression Trees

- Multiple response variables
- Response variables are continuous
- Can use continuous and/or categorical predictor variables
- Divide data into two groups
- Groups are as mutually exclusive and homogenous as possible. Minimize dissimilarity of sites within clusters.
- Minimize sum of square error at each split



De'ath 2002

# II. Multivariate Regression Trees

**Animal abundance**
**(Bunny + Fox)**

**Carrot Density > 2**

**Carrot Density <2**

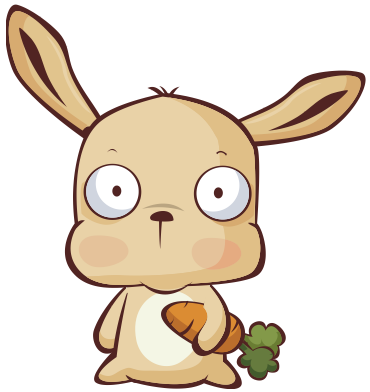
(Low abundance)

**Soil moisture > 35%**

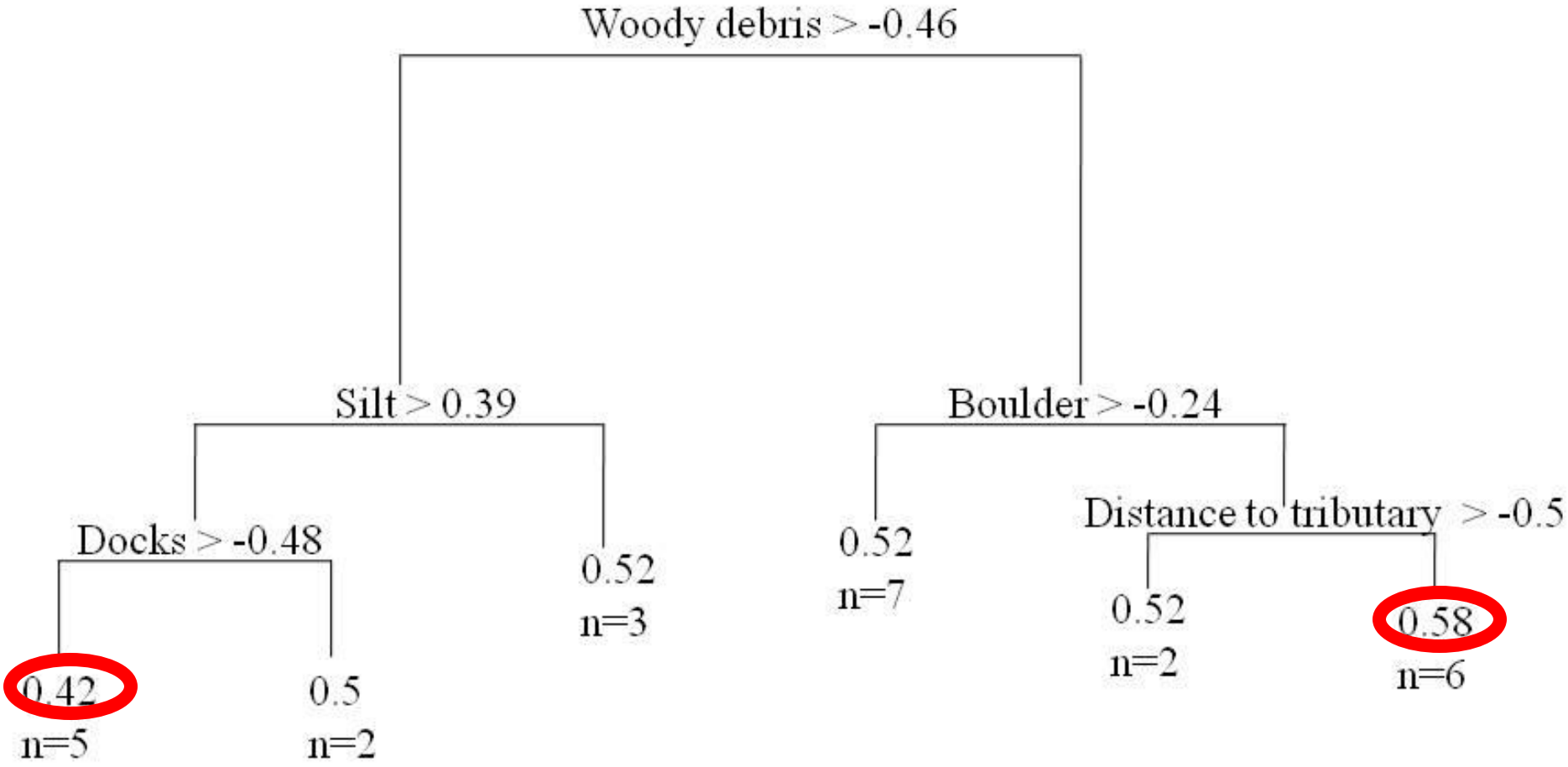
(High abundance)

**Soil Moisture <35%**

(Low abundance)



# II. MRT Example: Fish densities




Sharma et al. 2012

# Multivariate Regression Trees: Pros

- Complex ecological data
- Non-linear relationships
- Invariant to transformations of predictors
- Interactions among predictor variables
- Collinearity between predictors
- Local structure and interactions between environment, whereas ordinations determine global structure. Complementary techniques!

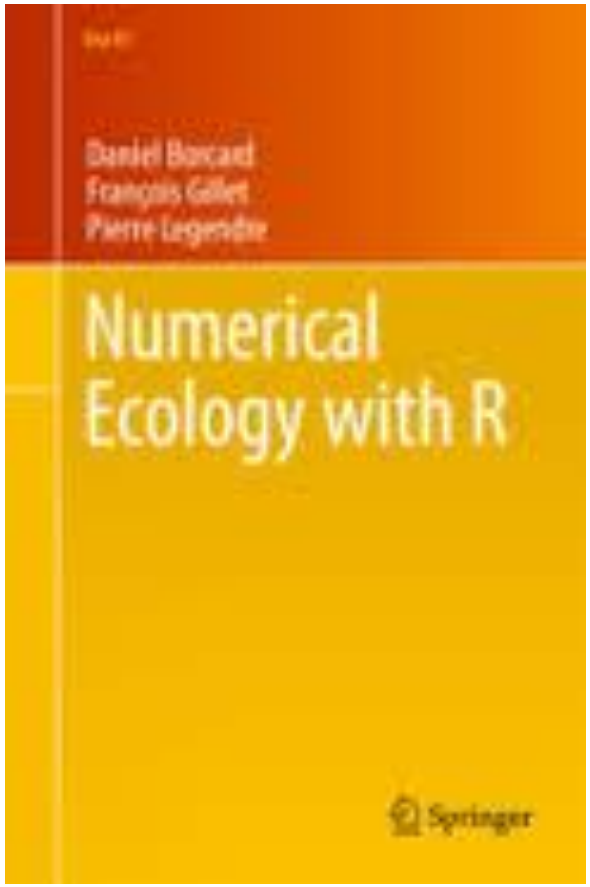


De'ath 2002

# Other Tree-Based Approaches

- Boosted regression trees (De'ath 2007; Buhlmann 2004; Elith et al. 2008)
- Bagging regression trees (Breiman 1996; Buhlmann 2004)
- Random forest (Breiman 2001; Evans et al. 2011).

# Lab Exercises




- Classification and Regression Trees: rpart
- Multivariate Regression Trees: mvpart
- Modified examples from “Numerical Ecology with R” by Daniel Borcard, Francois Gillet, and Pierre Legendre

Borcard et al. 2012

# Steps to generate a tree

1. For each predictor variable, produce all possible partitions of sites into 2 groups.
2. Compute sum of within-group sum of squared distances to the group mean (within group SS for response data.  Minimize this value.
3. Repeat until you have the best partition along with corresponding predictor variable and threshold value.
4. Continue until all objects are in a group

Borcard et al. 2012

# Steps to generate a tree

5. Calculate Relative Error (RE):

- RE = sum of within-group SS over all leaves / overall SS
- ~ fraction of variance NOT explained by tree
- Without pruning, $R^2 = 1 - RE$
- Calculate cross-validation RE to get a better estimate of $R^2$

Borcard et al. 2012

# Steps to generate a tree

6. Prune the tree

- Retain tree size for which CVRE is smallest (or where the minimal CVRE value plus 1 SE)
- Run the tree a large number of times (100, 500, 1000 times)
- Final tree has the smallest CVRE over all permutations or one that respects the 1 SE rule most often

Borcard et al. 2012

# CART Example: Car Mileage

- Car mileage predicted based on Price, Country, Reliability, Mileage, and Type

cu. summary

# CART Example: Grow Tree

```
library(rpart)
fit <- rpart(Mileage~Price + Country +
Reliability + Type, method="anova", xval
=100, data=cu.summary)
```

Note: If response variable was binary and you were interested in developing a classification tree, method = “class”

# CART Example: Interpret results

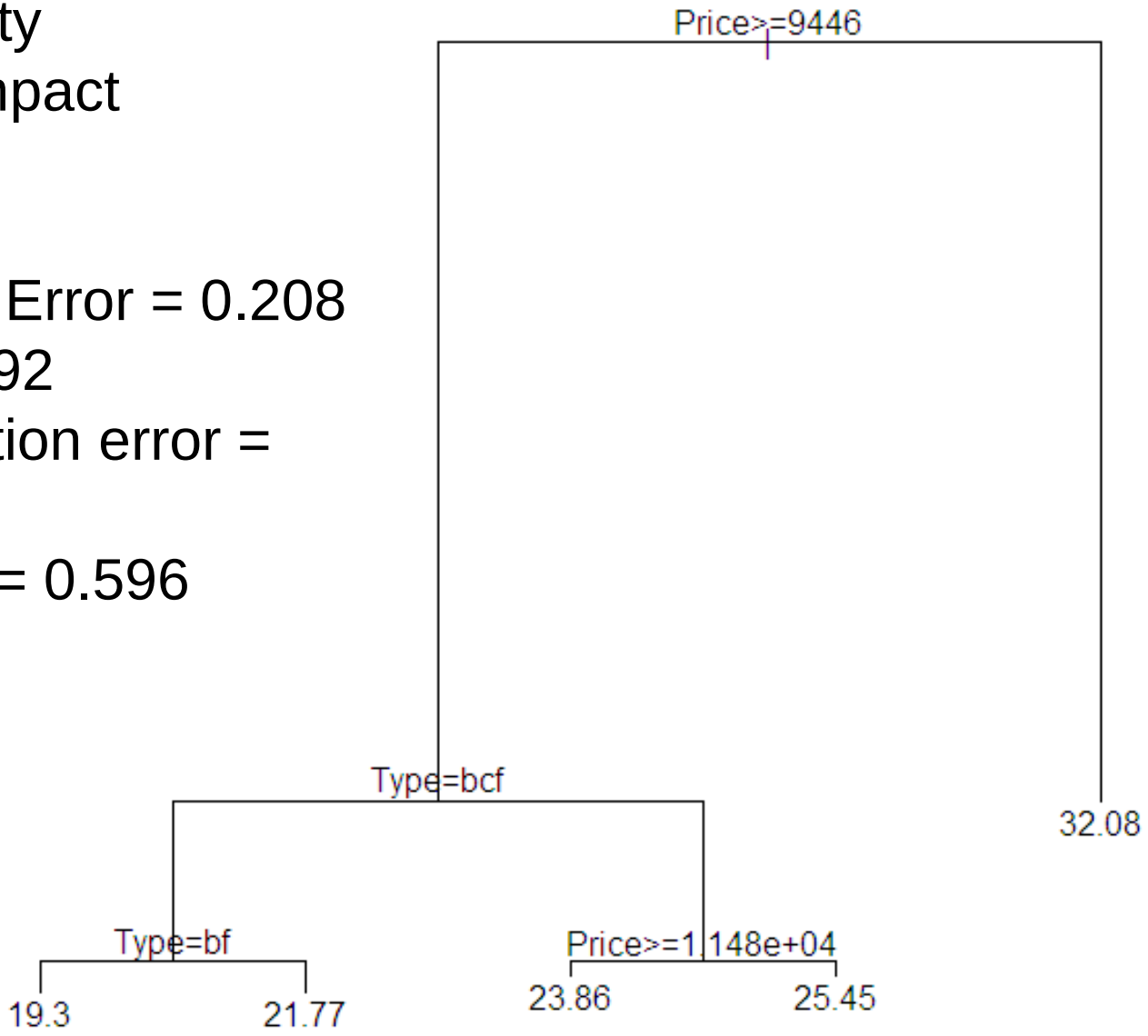
```
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit)
```

```
plot(fit)
text (fit)
```

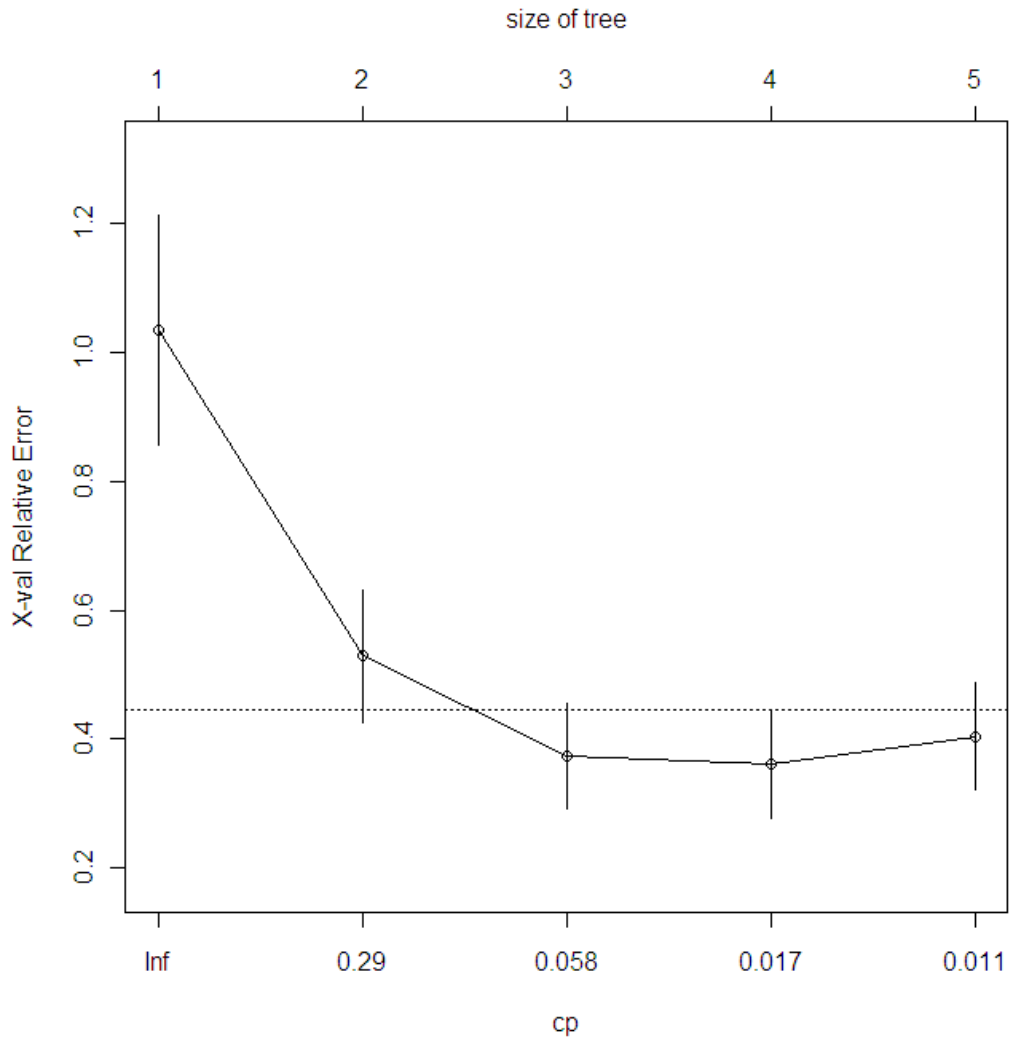
# CART Example: Interpret results

- B = sporty
- C = Compact
- F = Van

- Relative Error = 0.208
- $R^2$ = 0.792
- X-validation error = 0.404
- $R^2$ (CV) = 0.596



# CART Example: Prune Tree




- Pick tree with min. cross-validation error

```
pfit<- prune(fit, cp=
 fit$cptable[which.m
 in(fit$cptable[,"xerr
      or"]),"CP"])
      plot(pfit)
      text(pfit)
    summary(pfit)
```

# CART Example: Pick your own tree size

```
dfit <- rpart(Mileage~Price + Country + Reliability
+ Type, method="anova", maxdepth=2,
data=cu.summary)
plot(dfit)
text(dfit)
summary(dfit)
```

# MRT Example: Fish abundances

- Fish abundances from 30 sites along the Doubs River (Vernaux 1973)
- Data Info: http://www.inside-r.org/packages/cran/ade4/docs/doubs

```
library(ade4)
data(doubs)
env=doubs$mil
spe = doubs$poi
```

Borcard et al. 2012

# MRT Example: Transform data

- Transform species data using hellinger transformation
- Each value is expressed as a proportion of the sum of all values and the square root of the resulting value (Legendre and Gallagher 2001).
- Decreases the importance of the most abundant species.

```
library (vegan)
spe.norm=decostand(spe, "hellinger")
```

Borcard et al. 2012

# MRT Example: Tree Development

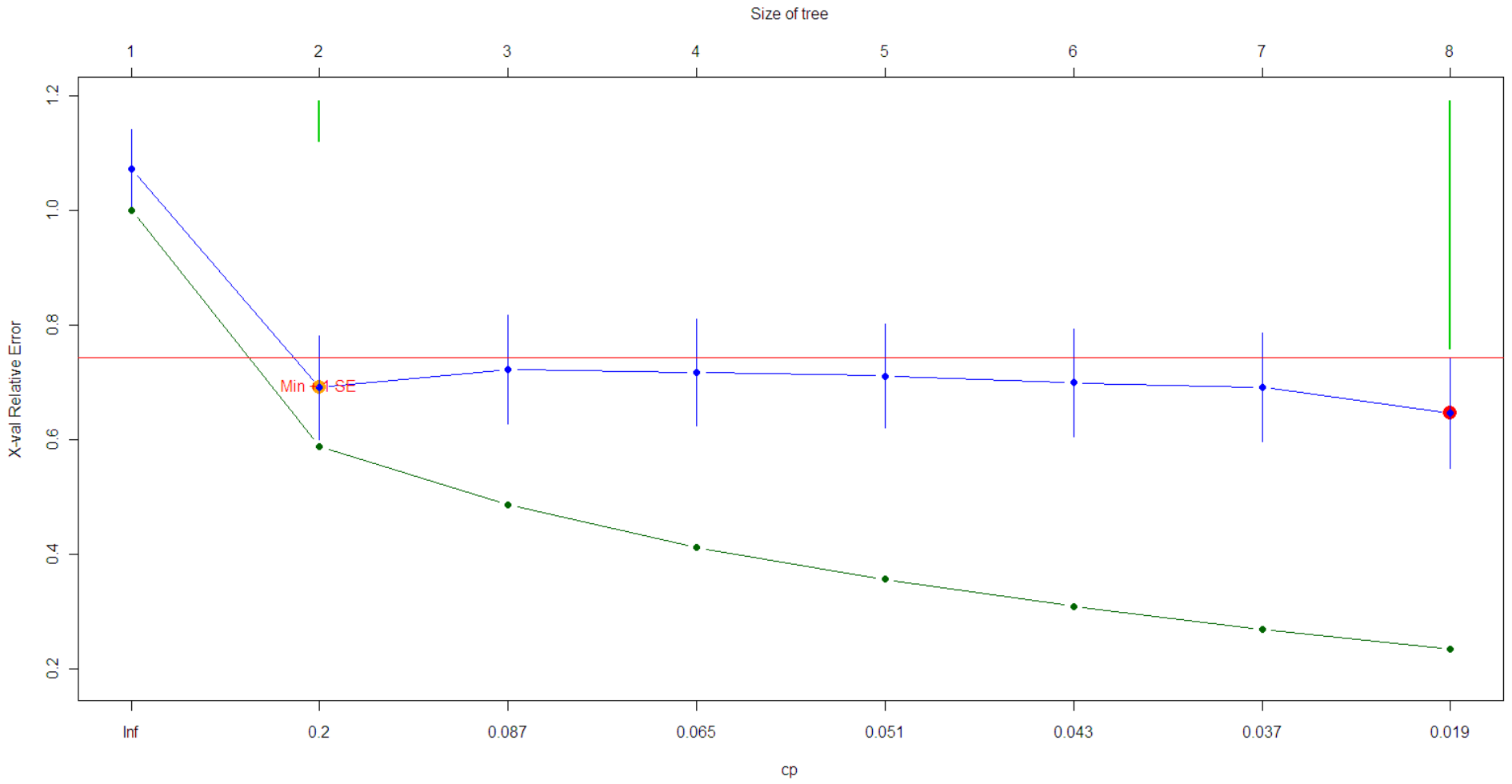
```
library(mvpart)
spe.ch.mvpart<-mvpart(data.matrix(spe.norm)
~., env, xv="pick", xval=nrow(spe),
xvmult=100, which=4)
```

- xv = selection of tree by cross-validation: "1se" - gives best tree within one SE of the overall best, "min" - the best tree, "pick" - pick the tree size interactively, "none" - no cross-validation
- xval = no. of cross-validation groups
- xvmult = No. of multiple cross-validations
- which = which split labels and where to plot them



Borcard et al. 2012

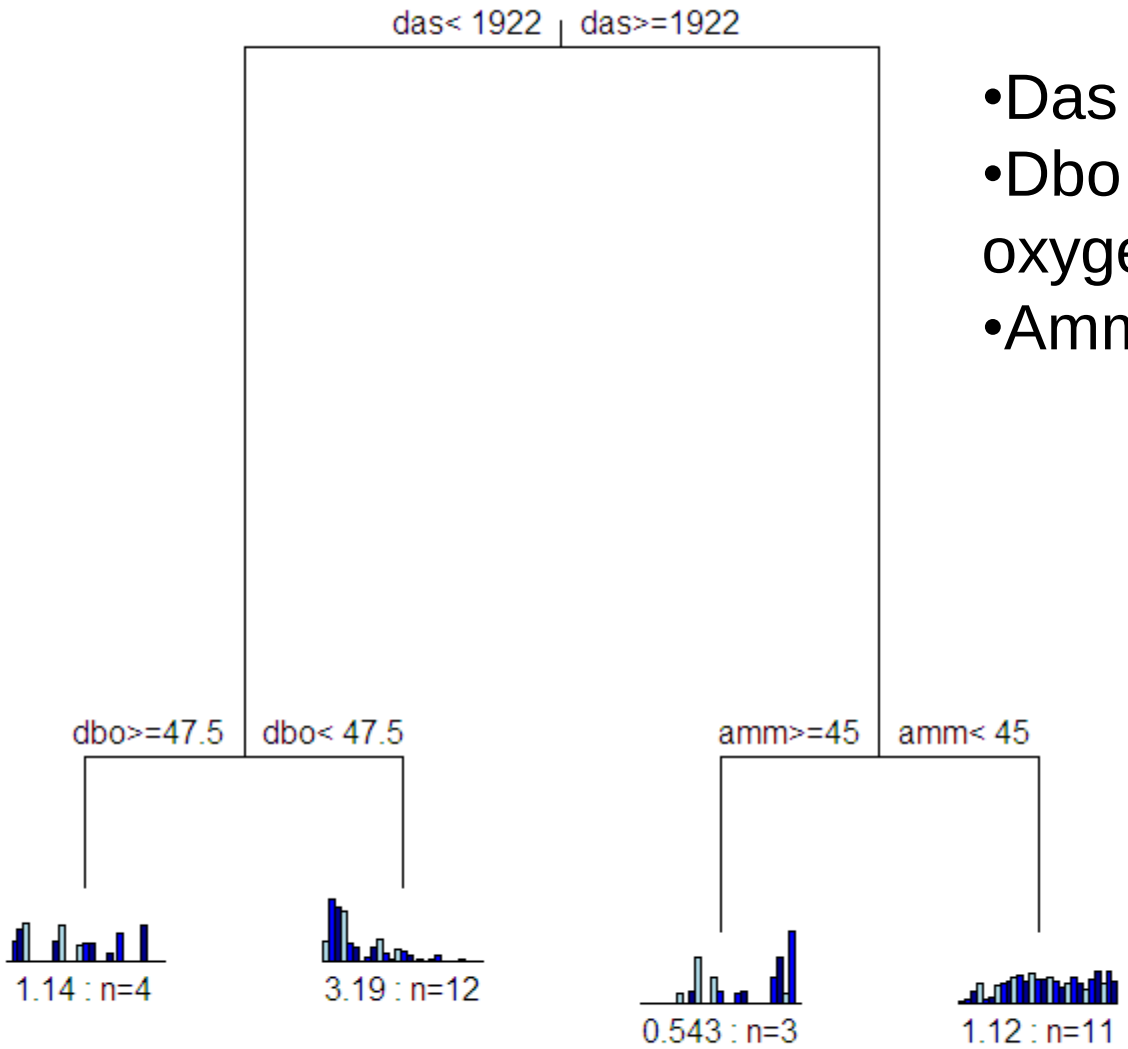
# MRT Example: Pick tree size






Borcard et al. 2012

# MRT Example: Interpret tree




Error : 0.411 CV Error : 0.719 SE : 0.0932

- Das = distance from source
- Dbo = biological demand for oxygen
- Amm = ammonia nitrogen



Borcard et al. 2012

# MRT Example: Interpret results

```
summary(spe.ch.mvpart)
printcp(spe.ch.mvpart)
```

- Summary of splits
- $R^2$ = 1 – Relative Error



Borcard et al. 2012

# CART and MRT: Pros & Cons

- ✚ Complex ecological data
- ✚ Non-linear relationships
- ✚ Invariant to transformations of predictors
- ✚ Interactions among predictor variables
- ✚ Collinearity between predictors
- ✚ Over-fitting
- ❑ Use the approach that is most appropriate to your data and question



De'ath and Fabricus 2000; De'ath 2002; Borcard et al. 2012; Sharma et al. 2012


Thank you!
Contact:
sharma11@yorku.ca