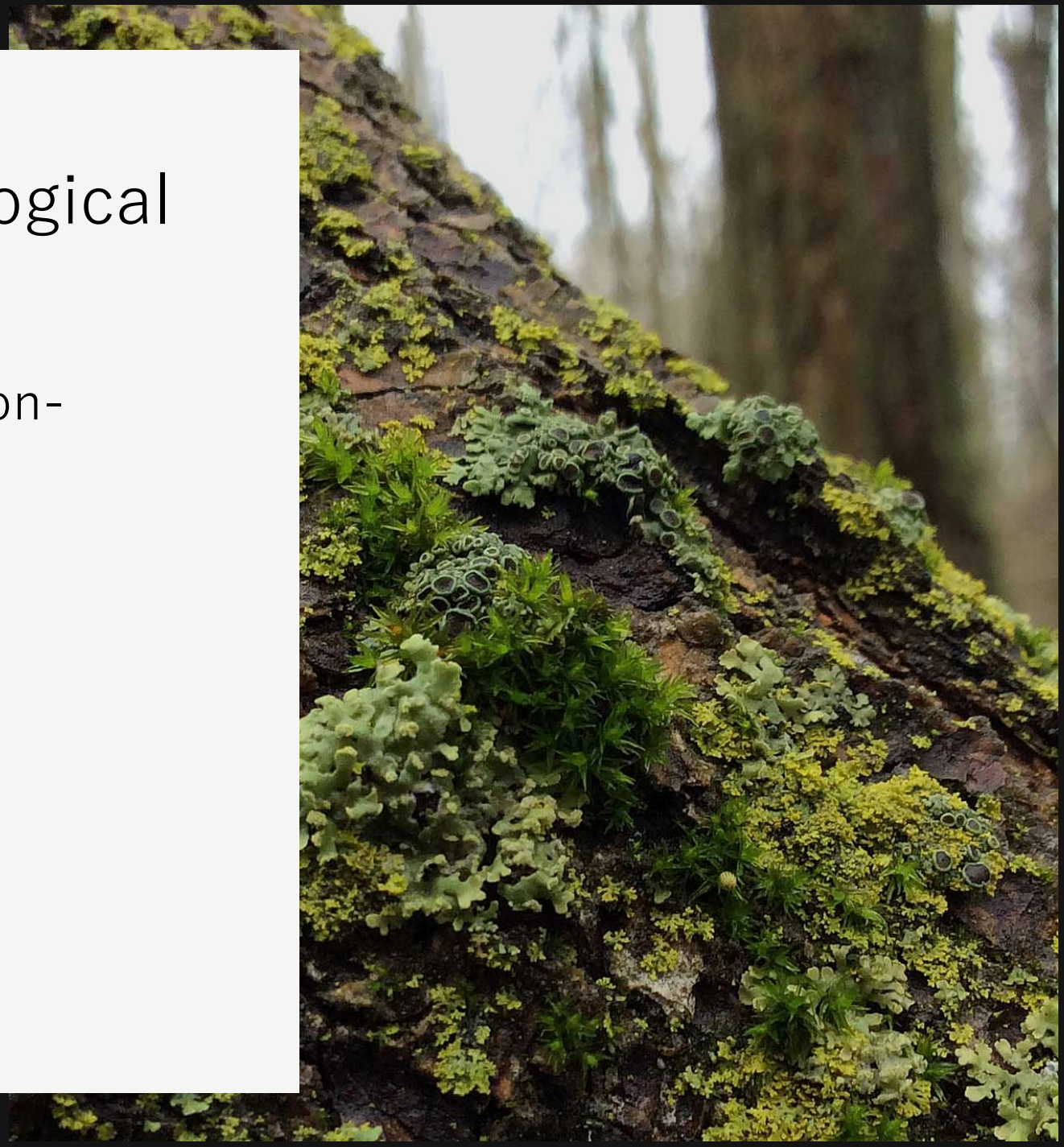
FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

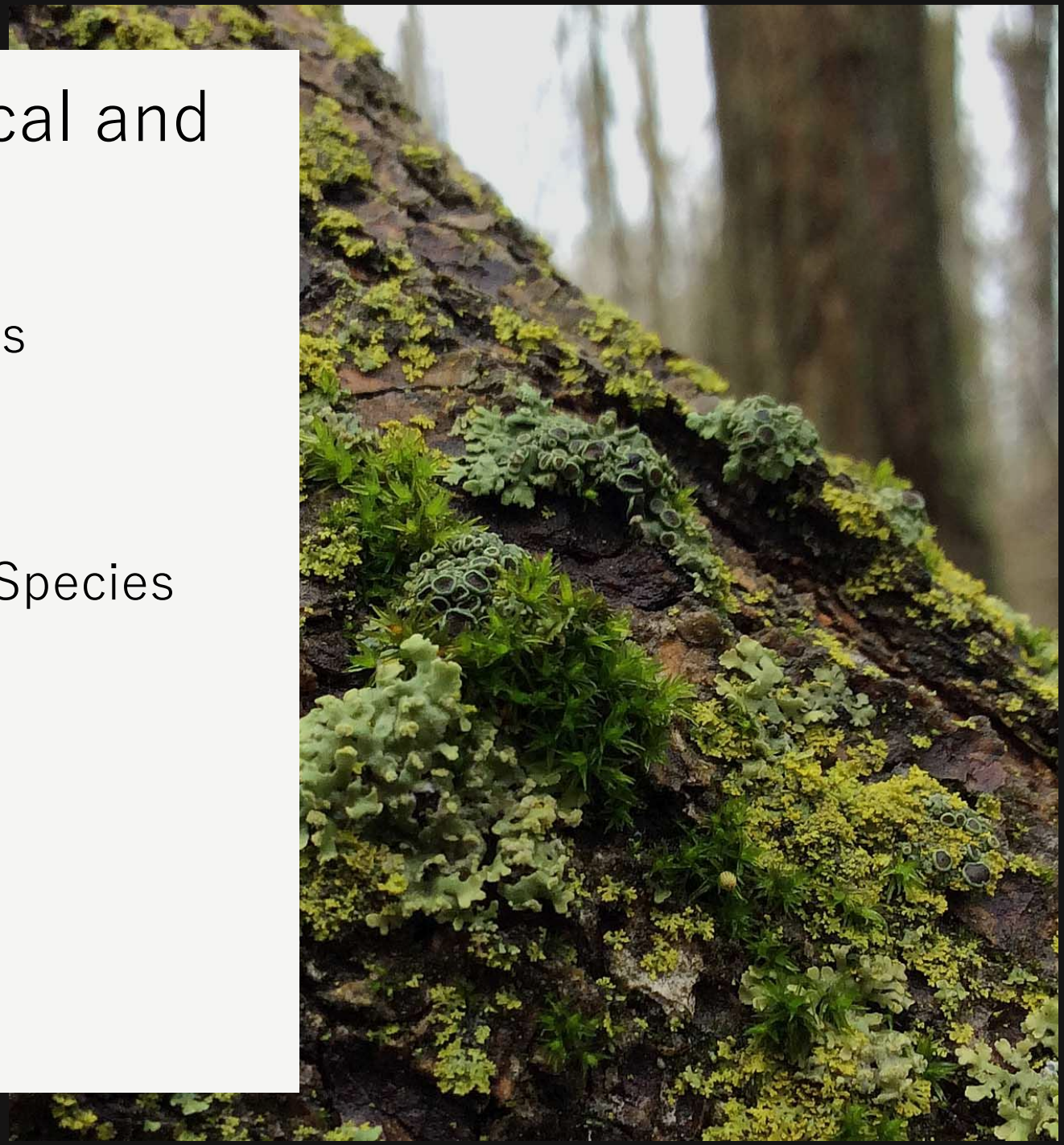
Lecture 5: Divisive Hierarchical and Non-Hierarchical Clustering

Tuesday, October 15, 2024

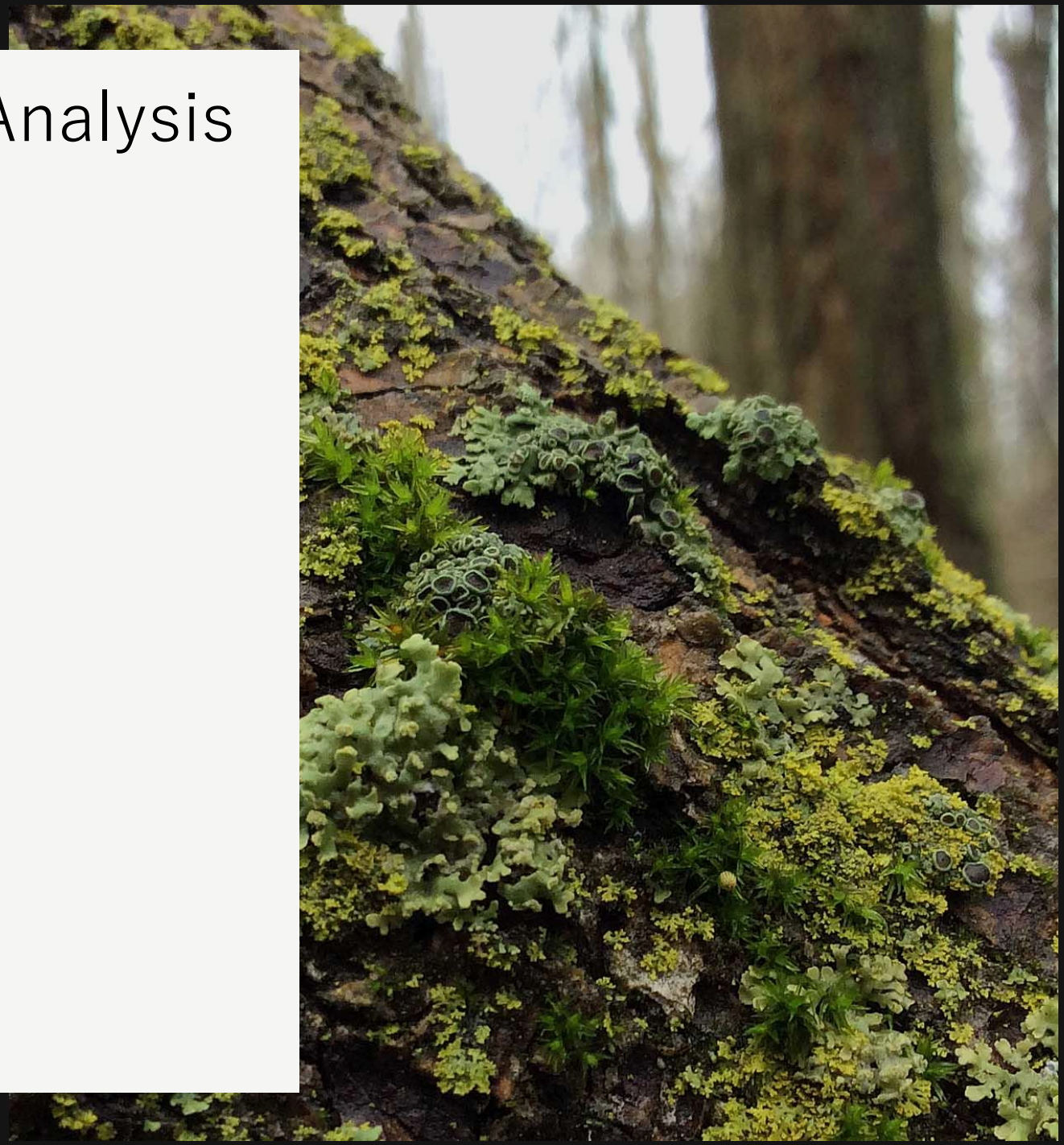


Lecture 5: Divisive Hierarchical and Non-Hierarchical Clustering

- Divisive Hierarchical Cluster Analysis
- K-means Partitioning
- Species Associations and Indicator Species



Recap: Hierarchical Cluster Analysis



Recap: Hierarchical Cluster Analysis

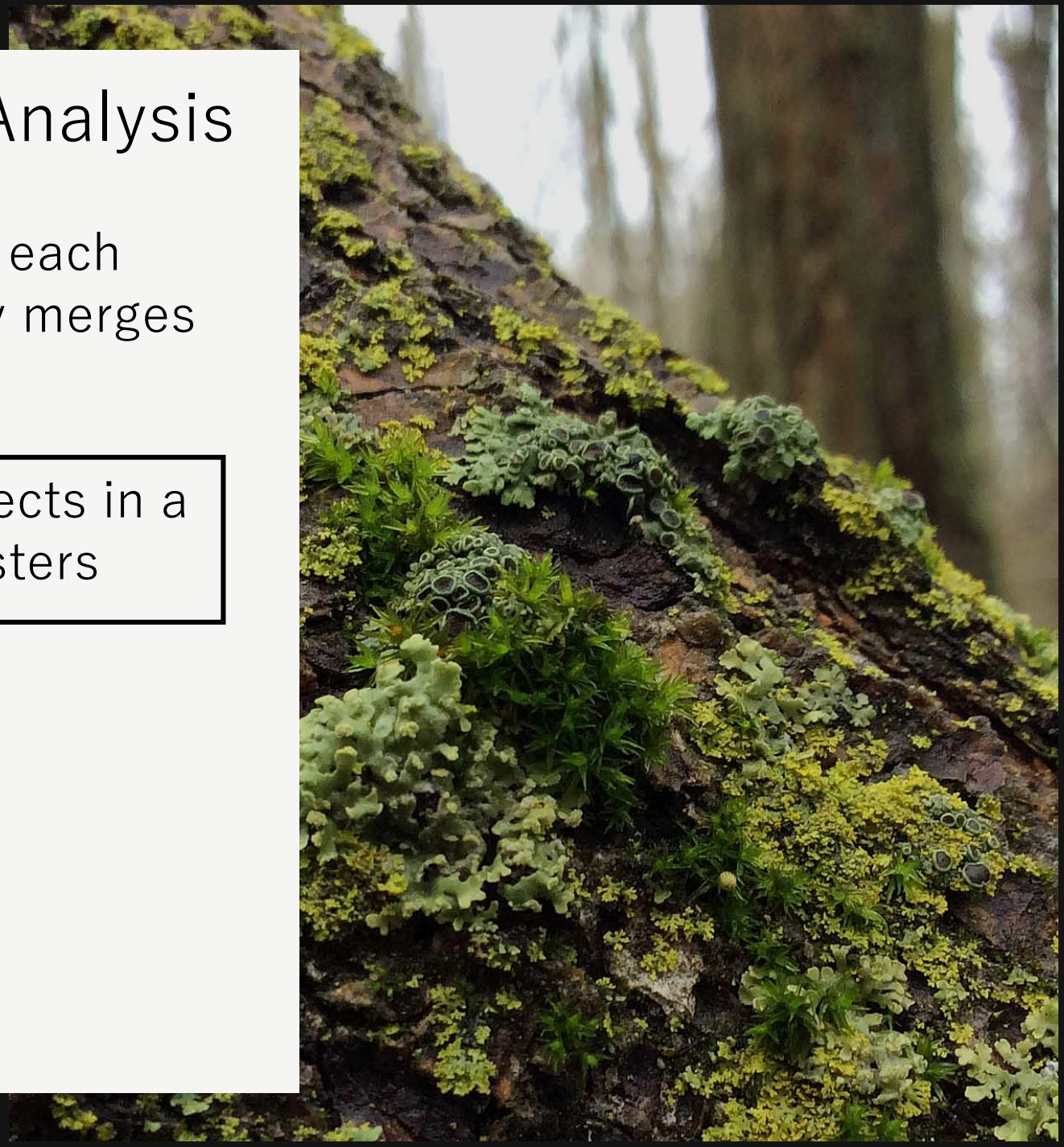
Hierarchical cluster analysis is used to classify objects, such as species, habitats, or environmental variables, into clusters based on their similarities or dissimilarities.

This technique helps ecologists to identify natural groupings and patterns within ecological data.

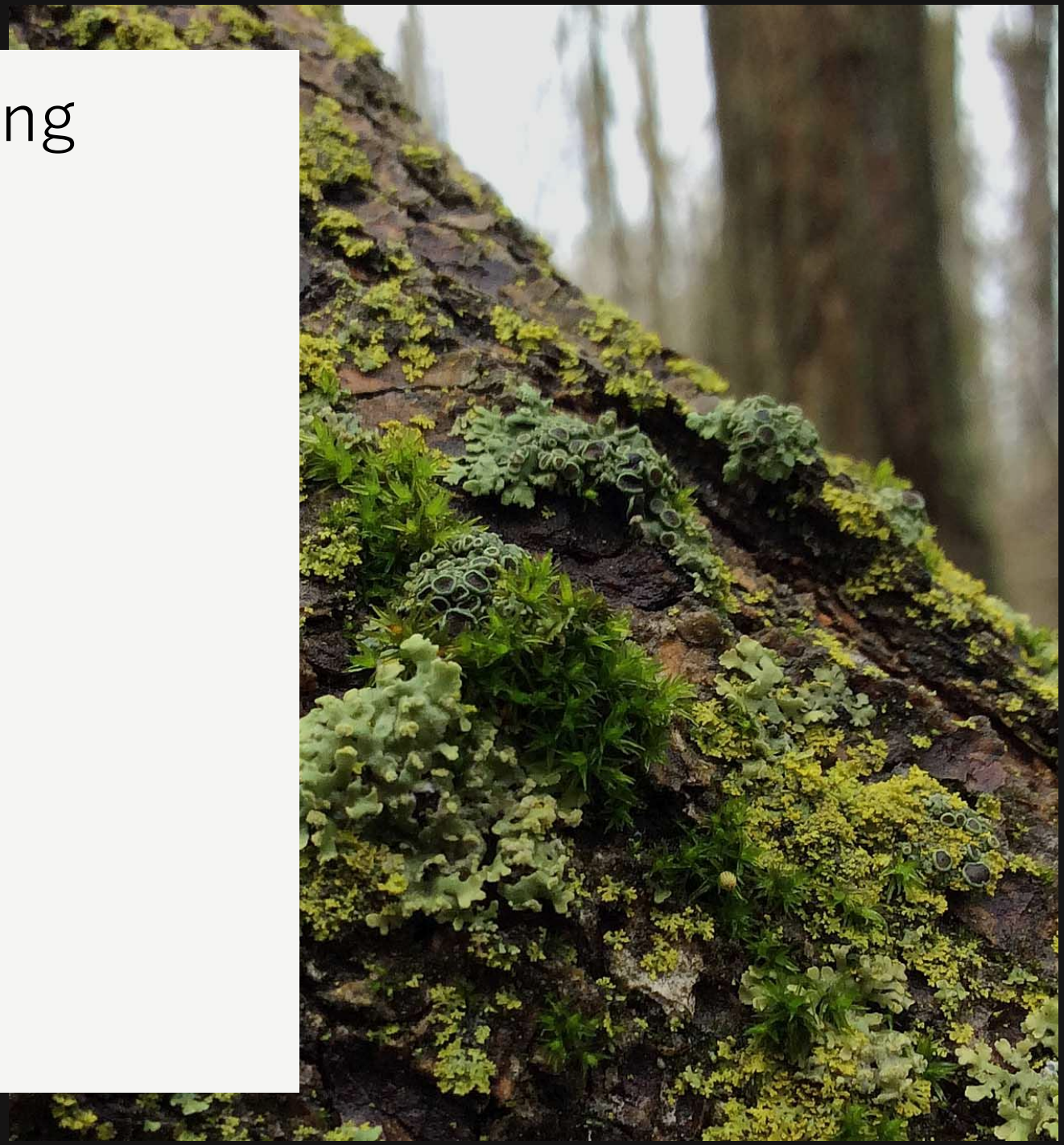
Recap: Hierarchical Cluster Analysis

Agglomerative Approach: Starts with each object in its own cluster and iteratively merges clusters

Divisive Approach: Starts with all objects in a single cluster and iteratively splits clusters

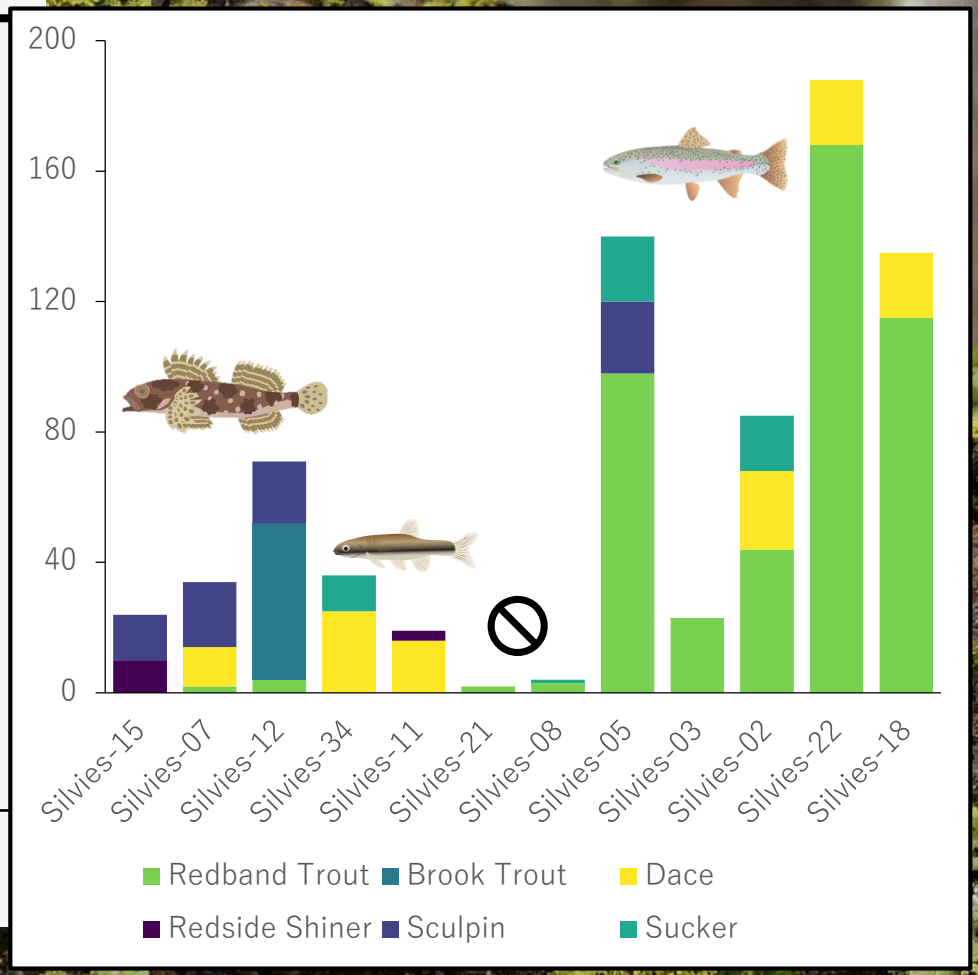


Divisive Hierarchical Clustering



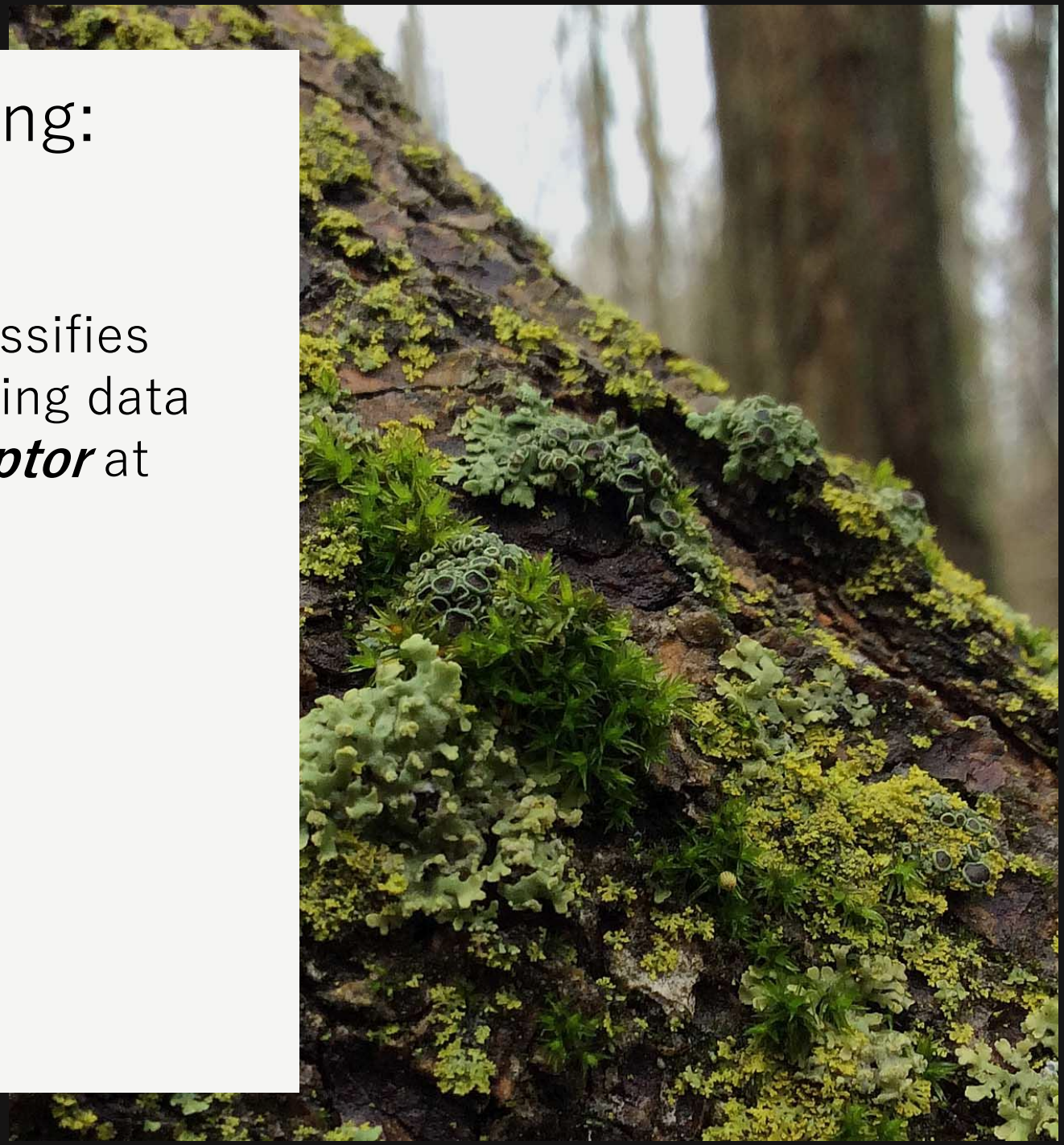
Divisive Hierarchical Clustering

Site ID	Redband Trout	Brook Trout	Dace	Redside Shiner	Sculpin	Sucker
Silvies-15	0	0	0	10	14	0
Silvies-07	2	0	12	0	20	0
Silvies-12	4	48	0	0	19	0
Silvies-34	0	0	25	0	0	11
Silvies-11	0	0	16	3	0	0
Silvies-21	2	0	0	0	0	0
Silvies-08	3	0	0	0	0	1
Silvies-05	98	0	0	0	22	20
Silvies-03	23	0	0	0	0	0
Silvies-02	44	0	24	0	0	17
Silvies-22	168	0	20	0	0	0
Silvies-18	115	0	20	0	0	0



Divisive Hierarchical Clustering: Monothetic Methods

Monothetic hierarchical clustering classifies ecological data by recursively partitioning data into subsets based on a ***single descriptor*** at each step.



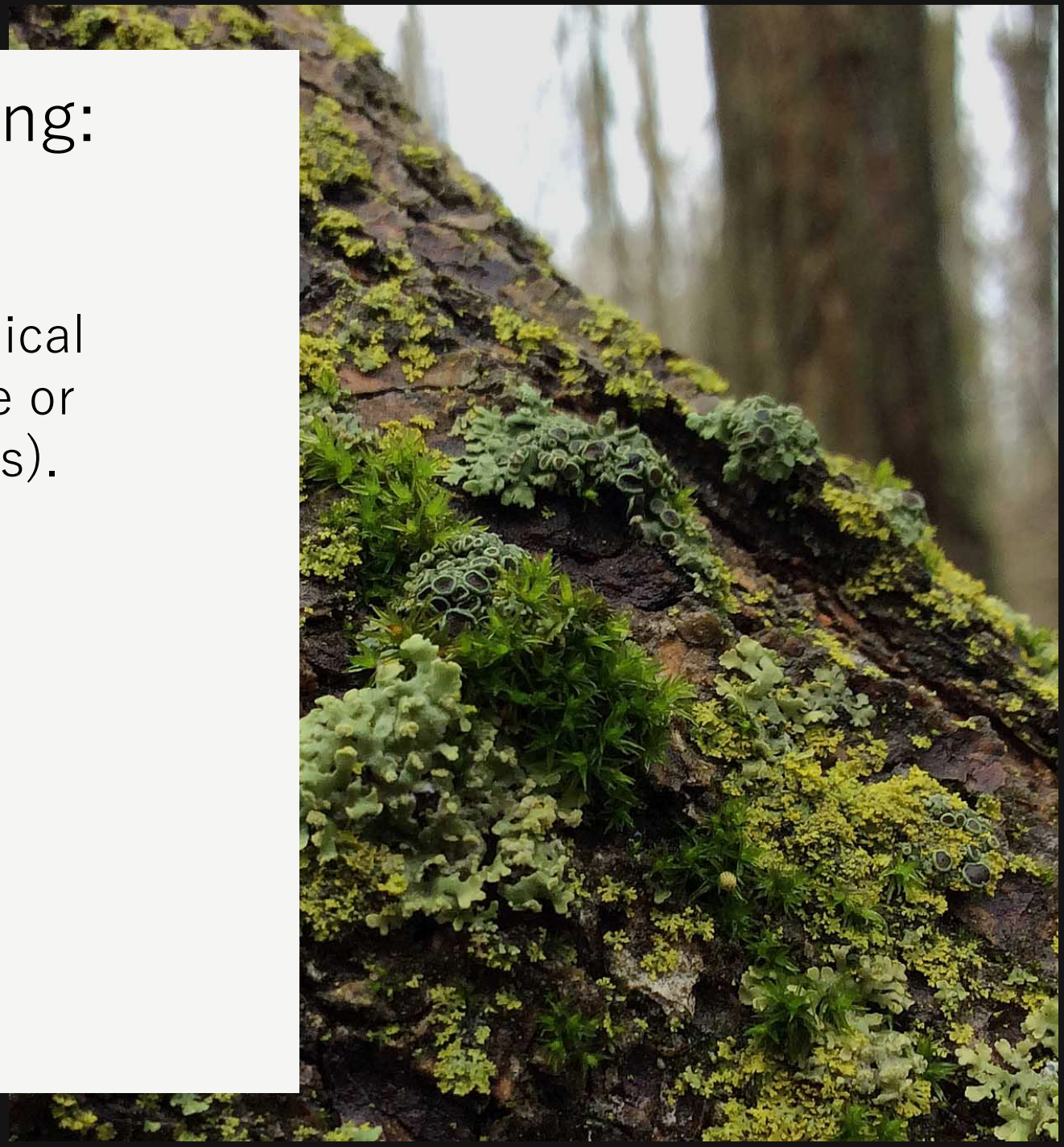
Divisive Hierarchical Clustering: Monothetic Methods

Monothetic hierarchical clustering classifies ecological data by recursively partitioning data into subsets based on a ***single descriptor*** at each step.

- Association Analysis using Binary Data
- Two-Way Indicator Species Analysis (TWINSPAN)
- Indicator Species (Indicator Value, IndVal)
- Classification and Regression Trees (CART)

Divisive Hierarchical Clustering: Monothetic Methods

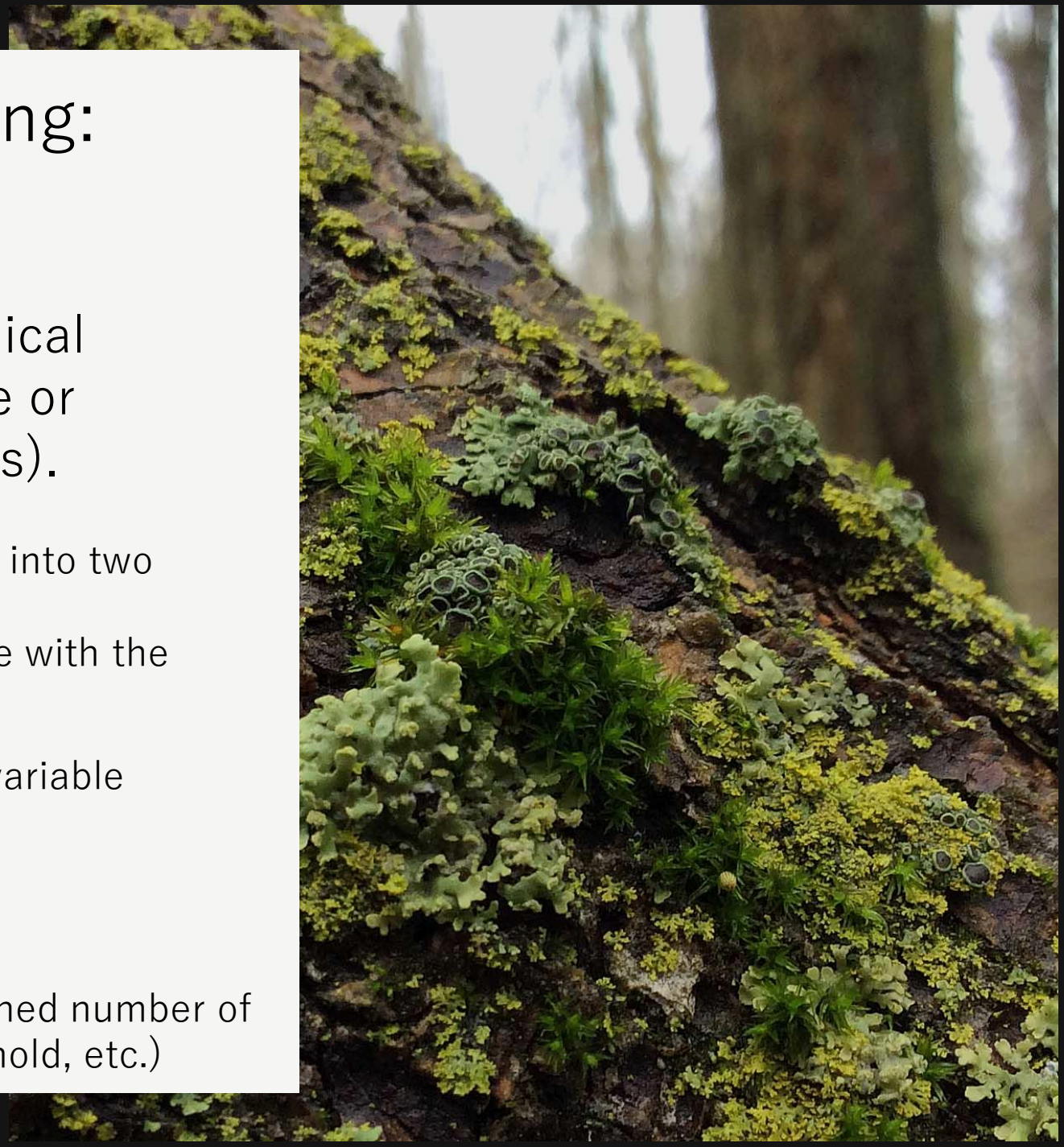
Association Analysis creates hierarchical clusters of data based on the presence or absence of descriptors (usually species).



Divisive Hierarchical Clustering: Monothetic Methods

Association Analysis creates hierarchical clusters of data based on the presence or absence of descriptors (usually species).

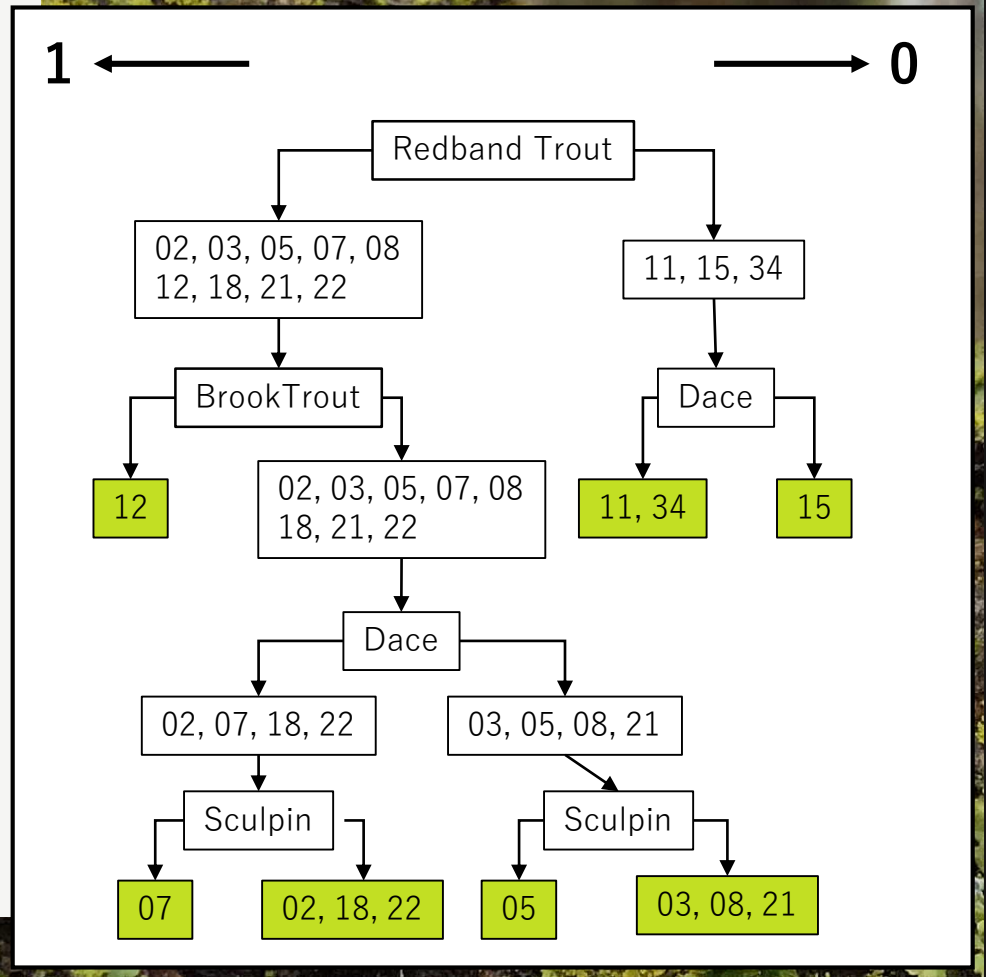
1. Identify binary variable that best separates data into two groups
   - Usually computed by identifying the variable with the highest χ^2 statistic or Gini index
2. Split dataset into two clusters based on binary variable (presence vs. absence)
3. Repeat steps
4. Continue until stopping criterion is met (predefined number of clusters, all variables used, minimum size threshold, etc.)



Divisive Hierarchical Clustering: Monothetic Methods

Association Analysis creates hierarchical clusters of data based on the presence or absence of descriptors (usually species).

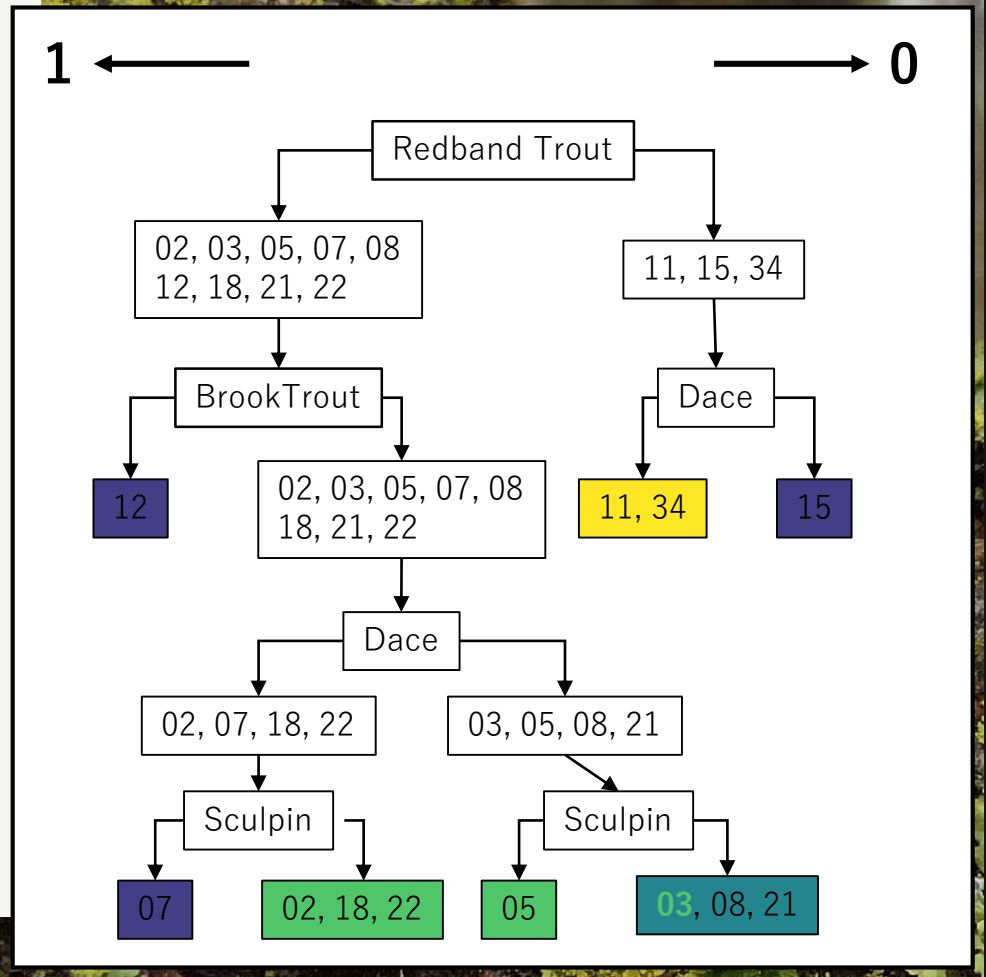
1. Identify binary variable that best separates data into two groups
   - Usually computed by identifying the variable with the highest χ^2 statistic or Gini index
2. Split dataset into two clusters based on binary variable (presence vs. absence)
3. Repeat steps
4. Continue until stopping criterion is met (predefined number of clusters, all variables used, minimum size threshold, etc.)



Divisive Hierarchical Clustering: Monothetic Methods

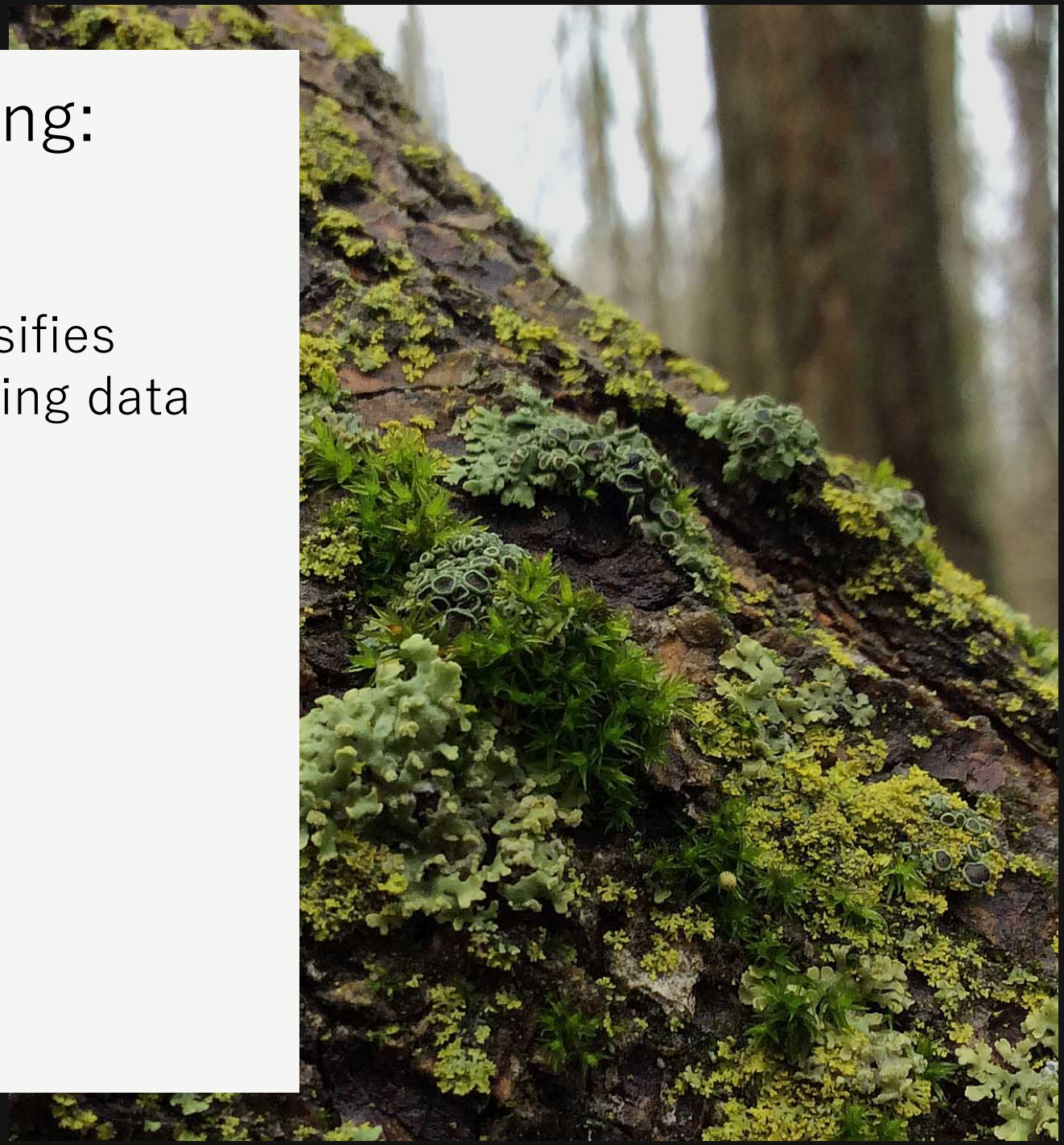
Association Analysis creates hierarchical clusters of data based on the presence or absence of descriptors (usually species).

1. Identify binary variable that best separates data into two groups
   - Usually computed by identifying the variable with the highest χ^2 statistic or Gini index
2. Split dataset into two clusters based on binary variable (presence vs. absence)
3. Repeat steps
4. Continue until stopping criterion is met (predefined number of clusters, all variables used, minimum size threshold, etc.)



Divisive Hierarchical Clustering: Polythetic Methods

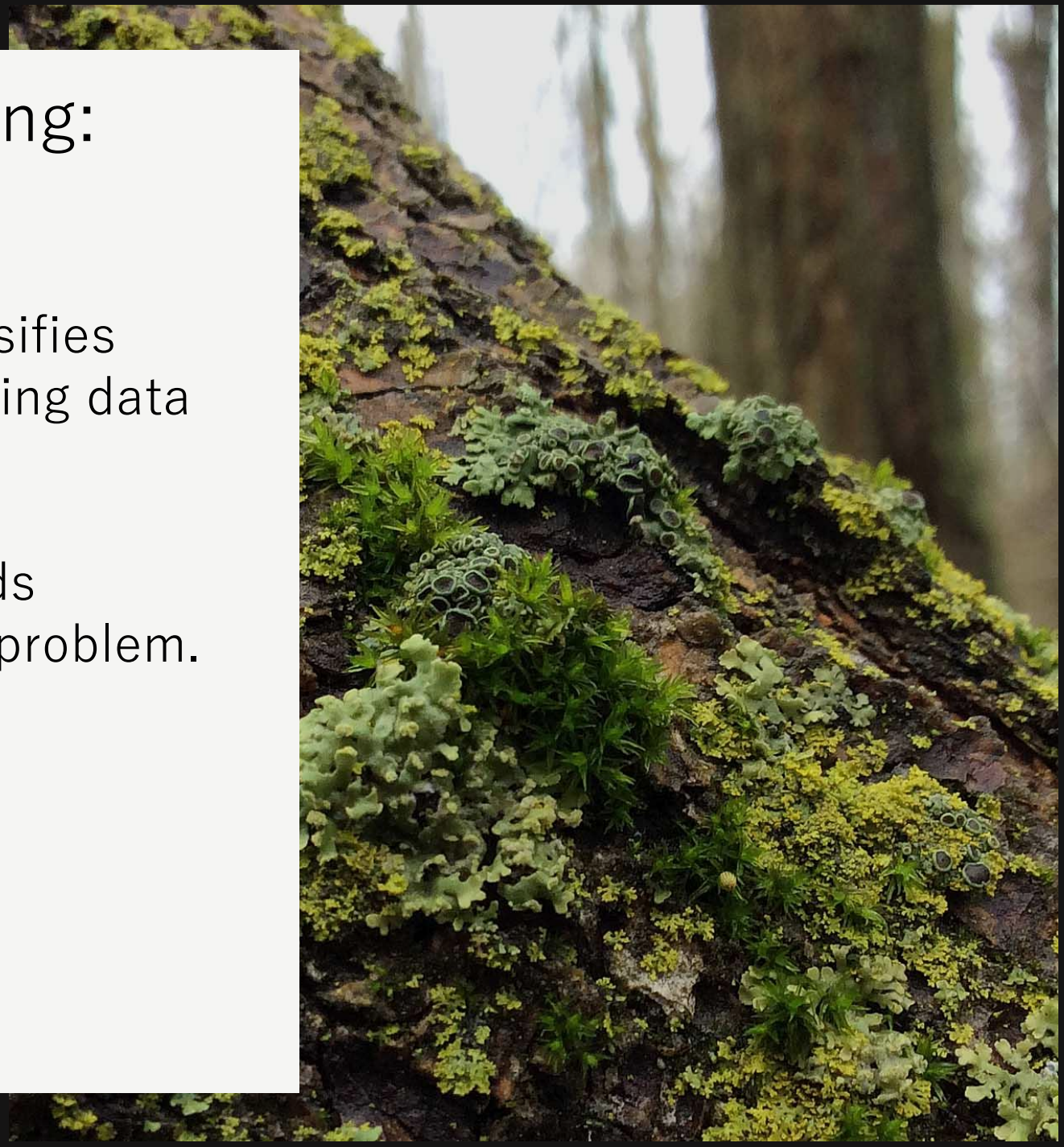
Polythetic hierarchical clustering classifies ecological data by recursively partitioning data into subsets based on ***all descriptors***.



Divisive Hierarchical Clustering: Polythetic Methods

Polythetic hierarchical clustering classifies ecological data by recursively partitioning data into subsets based on ***all descriptors***.

At this point in time, polythetic methods represent an **NP-Hard** computational problem.

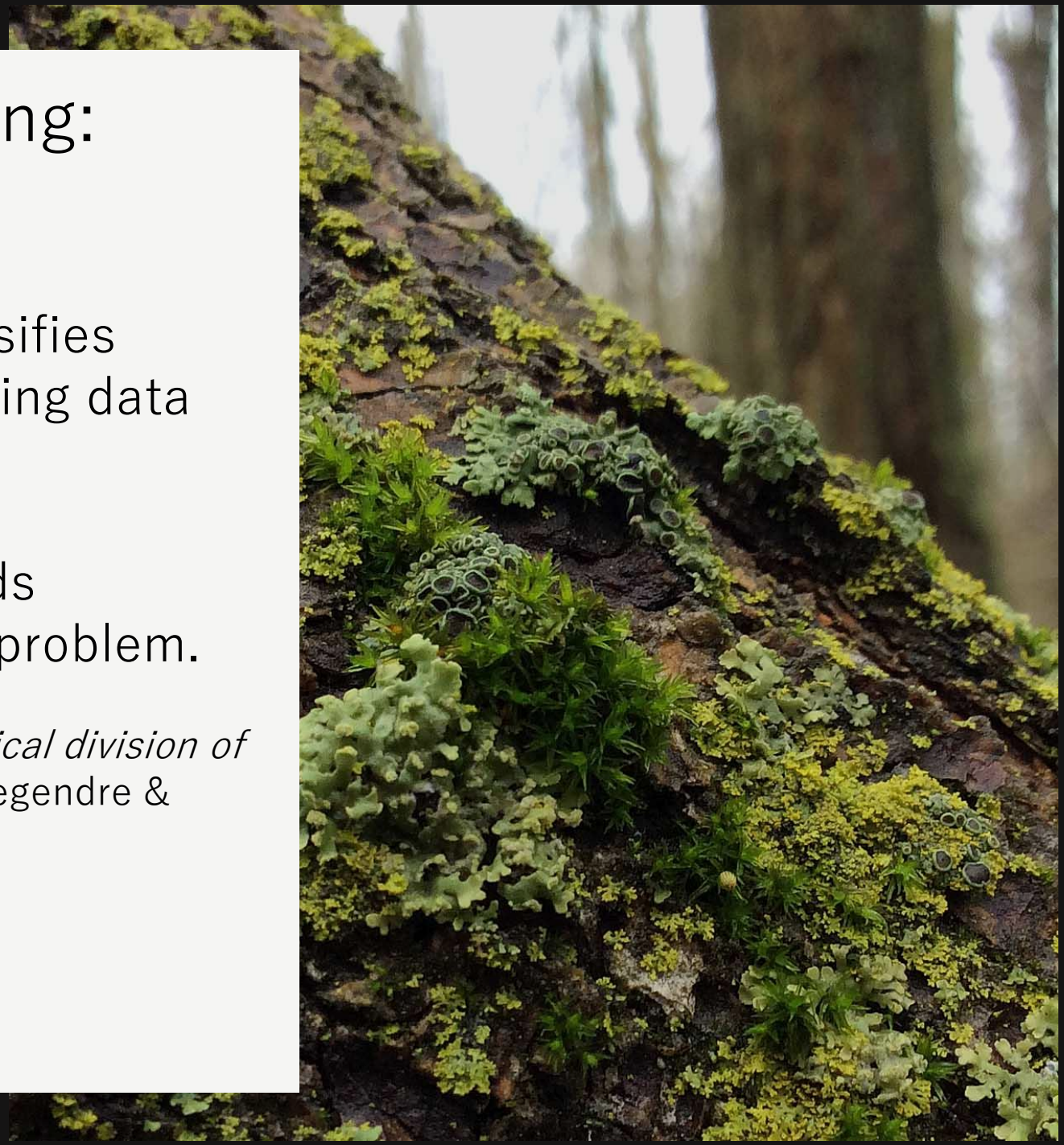


Divisive Hierarchical Clustering: Polythetic Methods

Polythetic hierarchical clustering classifies ecological data by recursively partitioning data into subsets based on ***all descriptors***.

At this point in time, polythetic methods represent an **NP-Hard** computational problem.

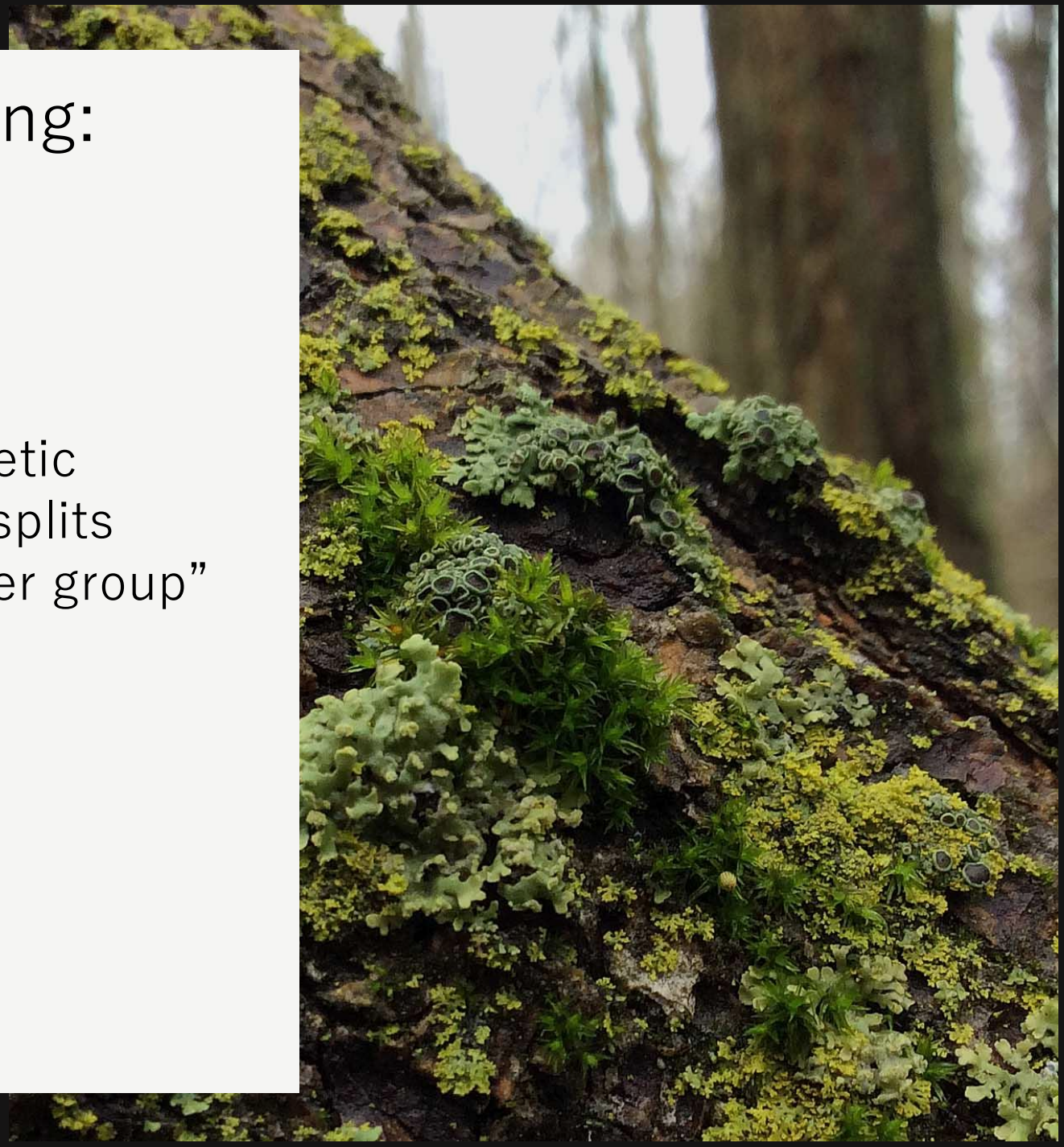
"*There is no satisfactory algorithm for the hierarchical division of objects based on the entire set of descriptors*." – Legendre & Legendre



Divisive Hierarchical Clustering: Polythetic Methods

Just kidding?

Divisive Analysis (**DIANA**) is a polythetic hierarchical clustering technique that splits clusters recursively based on a “splinter group” of dissimilar observations.



Divisive Hierarchical Clustering: Polythetic Methods

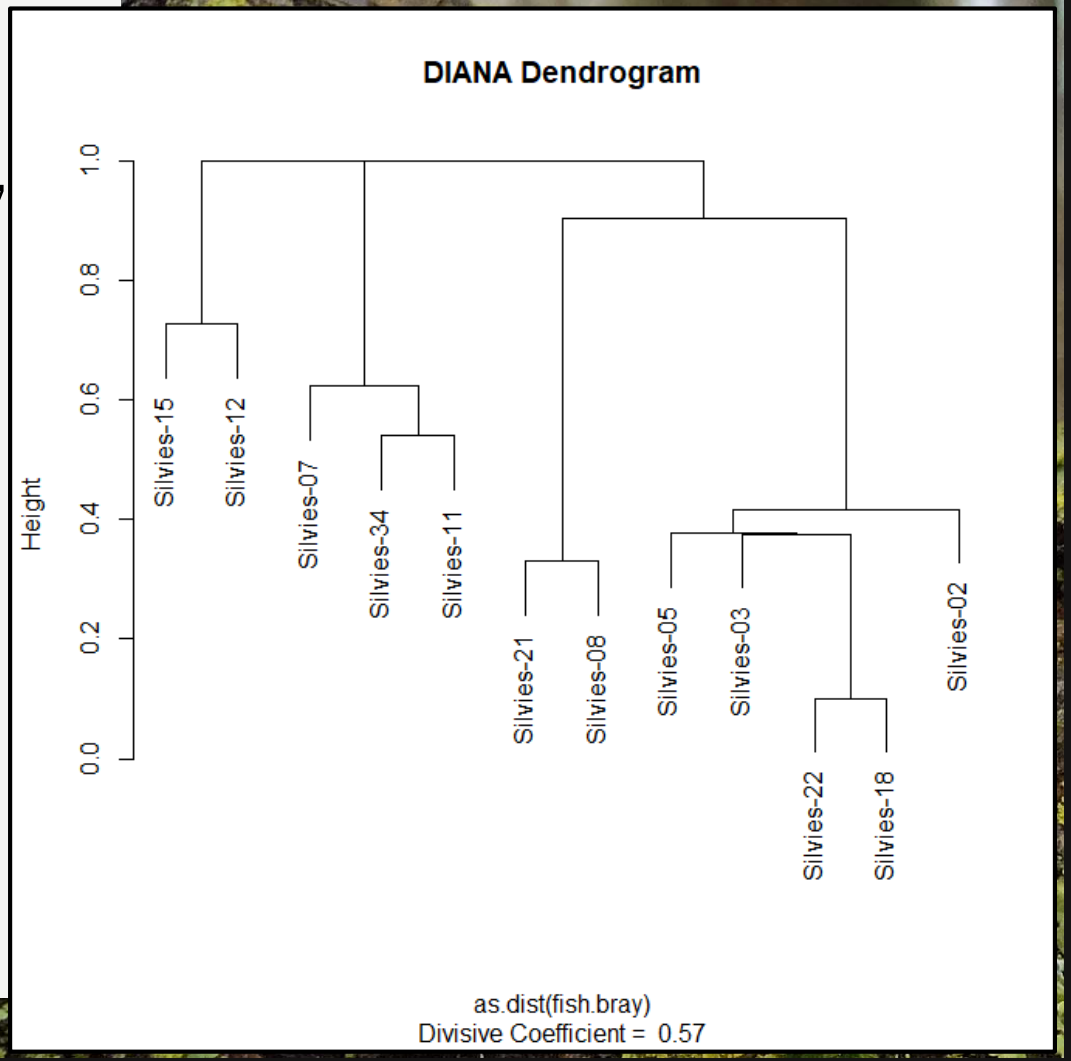
Divisive Analysis (DIANA) is a polythetic hierarchical clustering technique that splits clusters recursively based on a “splinter group” of dissimilar observations.

Advantages

- Easily interpretable
- Hierarchical structure

Disadvantages

- Computationally intensive
- Sensitive to noise and outliers



Divisive Hierarchical Clustering: Polythetic Methods

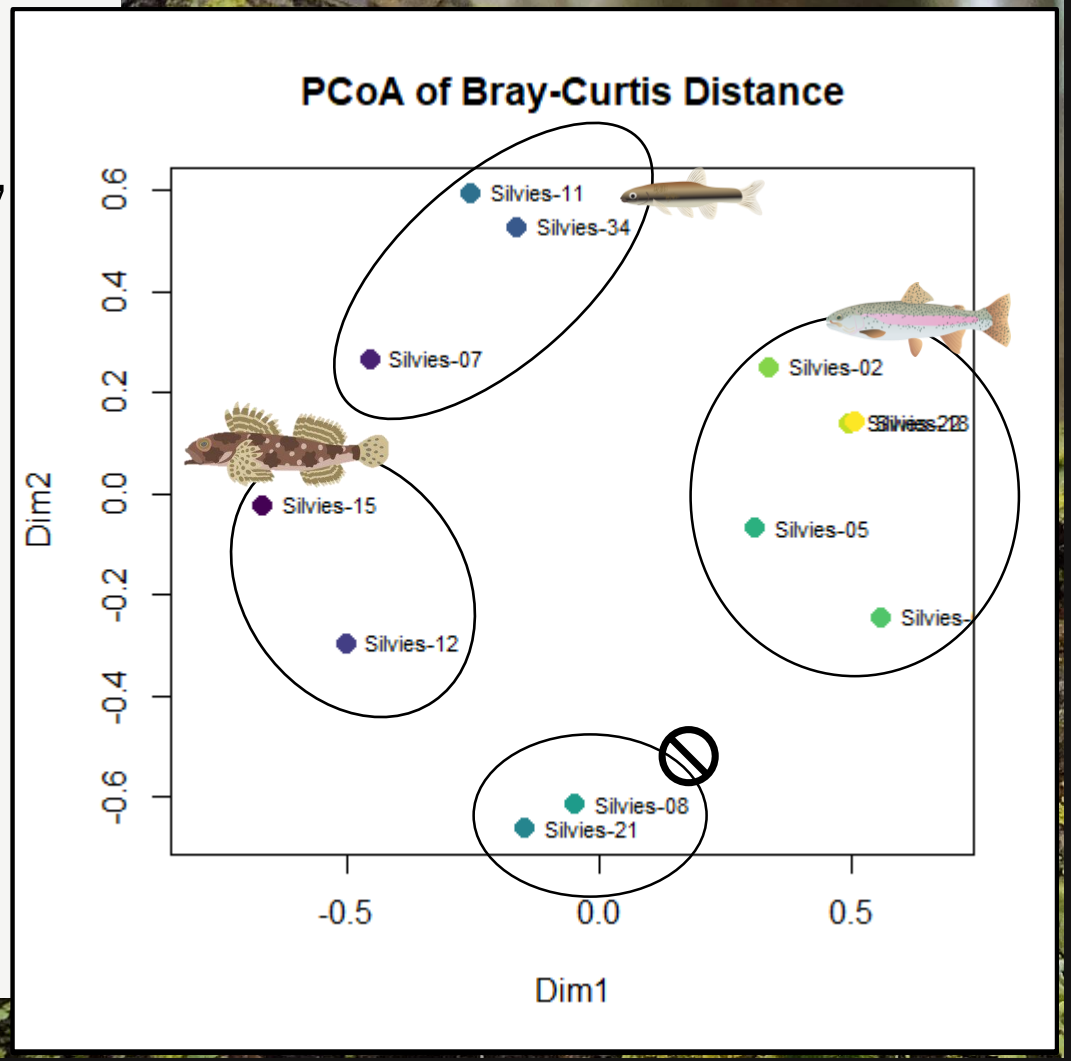
Divisive Analysis (DIANA) is a polythetic hierarchical clustering technique that splits clusters recursively based on a “splinter group” of dissimilar observations.

Advantages

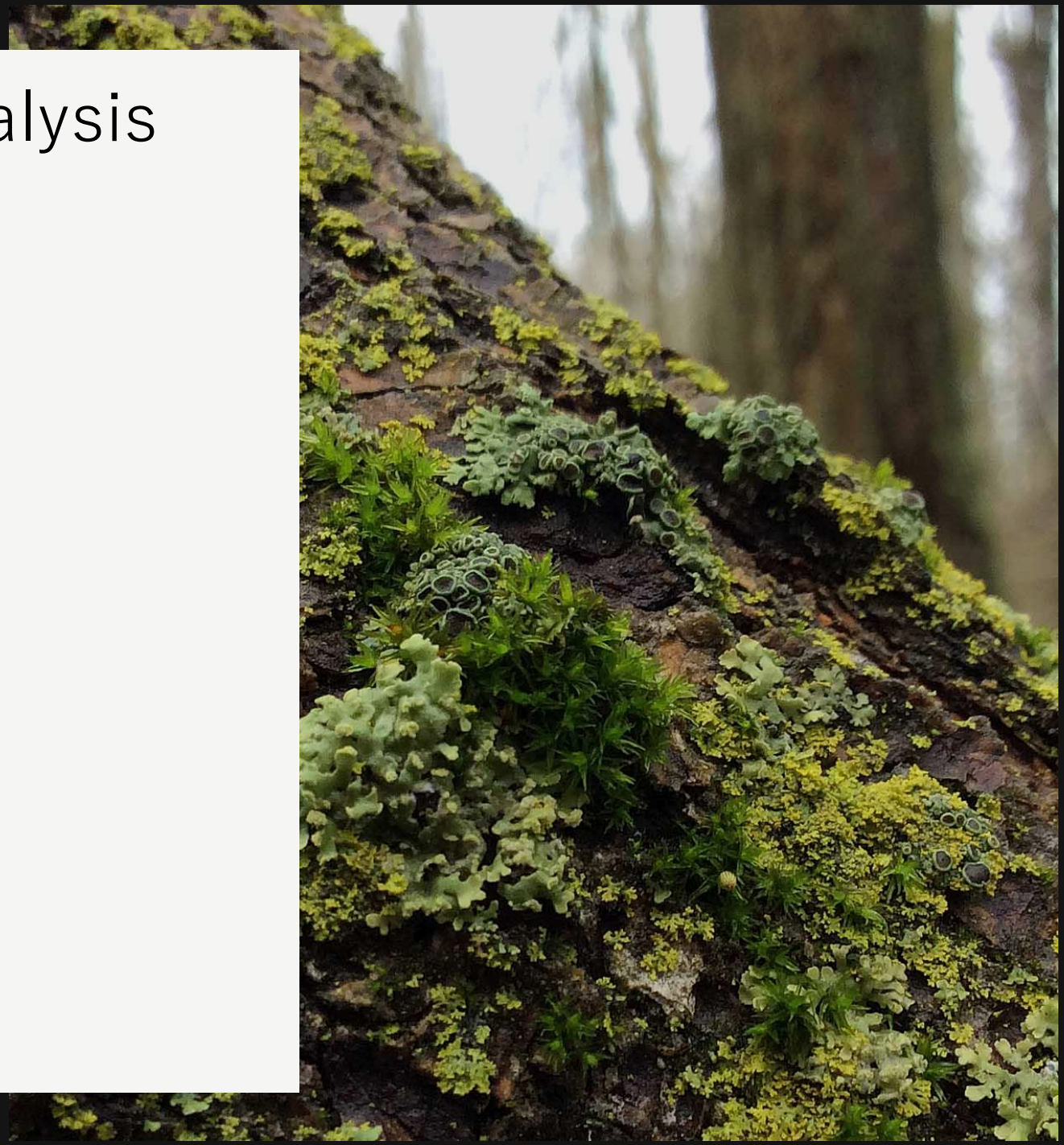
- Easily interpretable
- Hierarchical structure

Disadvantages

- Computationally intensive
- Sensitive to noise and outliers

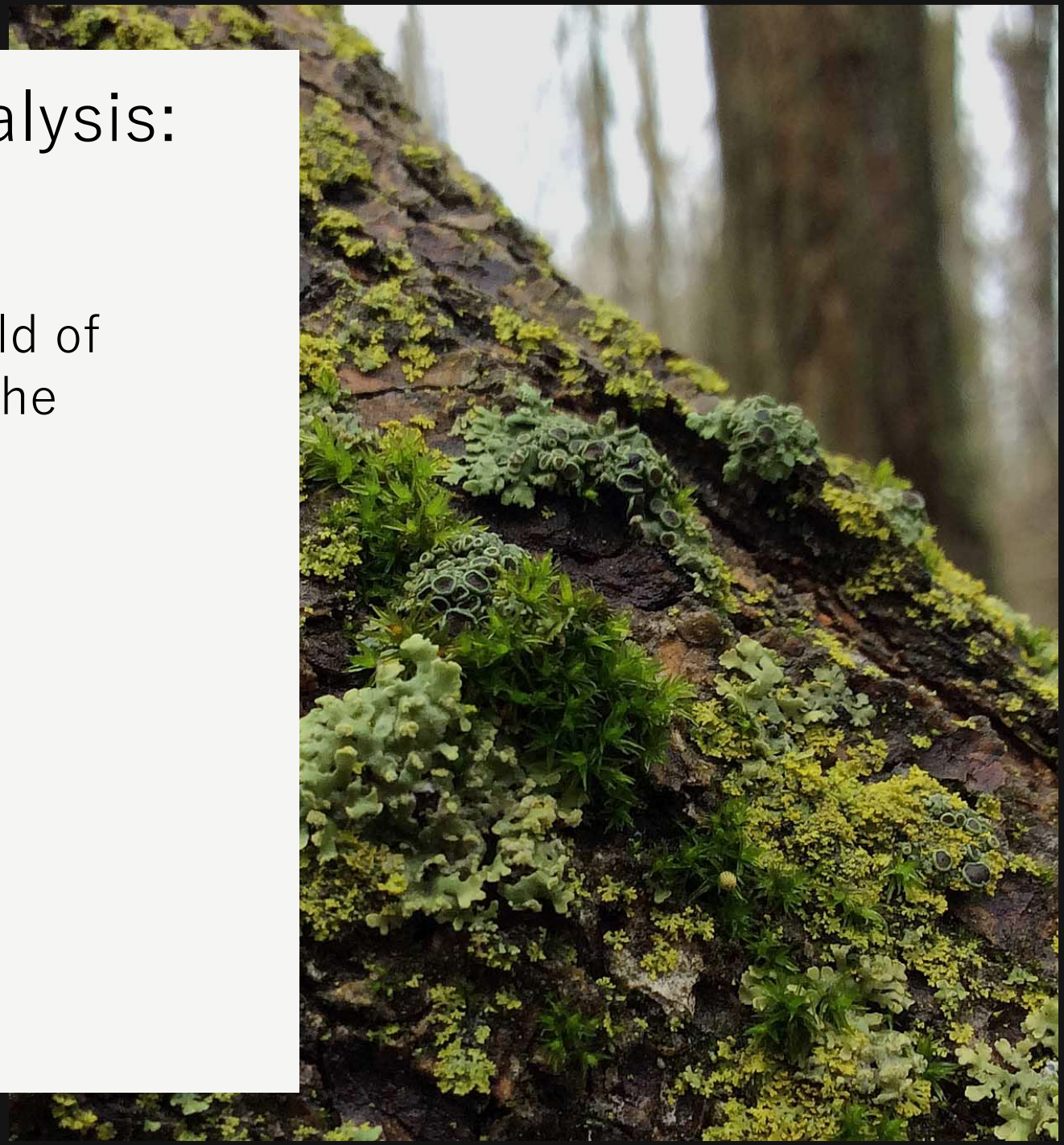


Non-Hierarchical Cluster Analysis



Non-Hierarchical Cluster Analysis: Complete Linkage

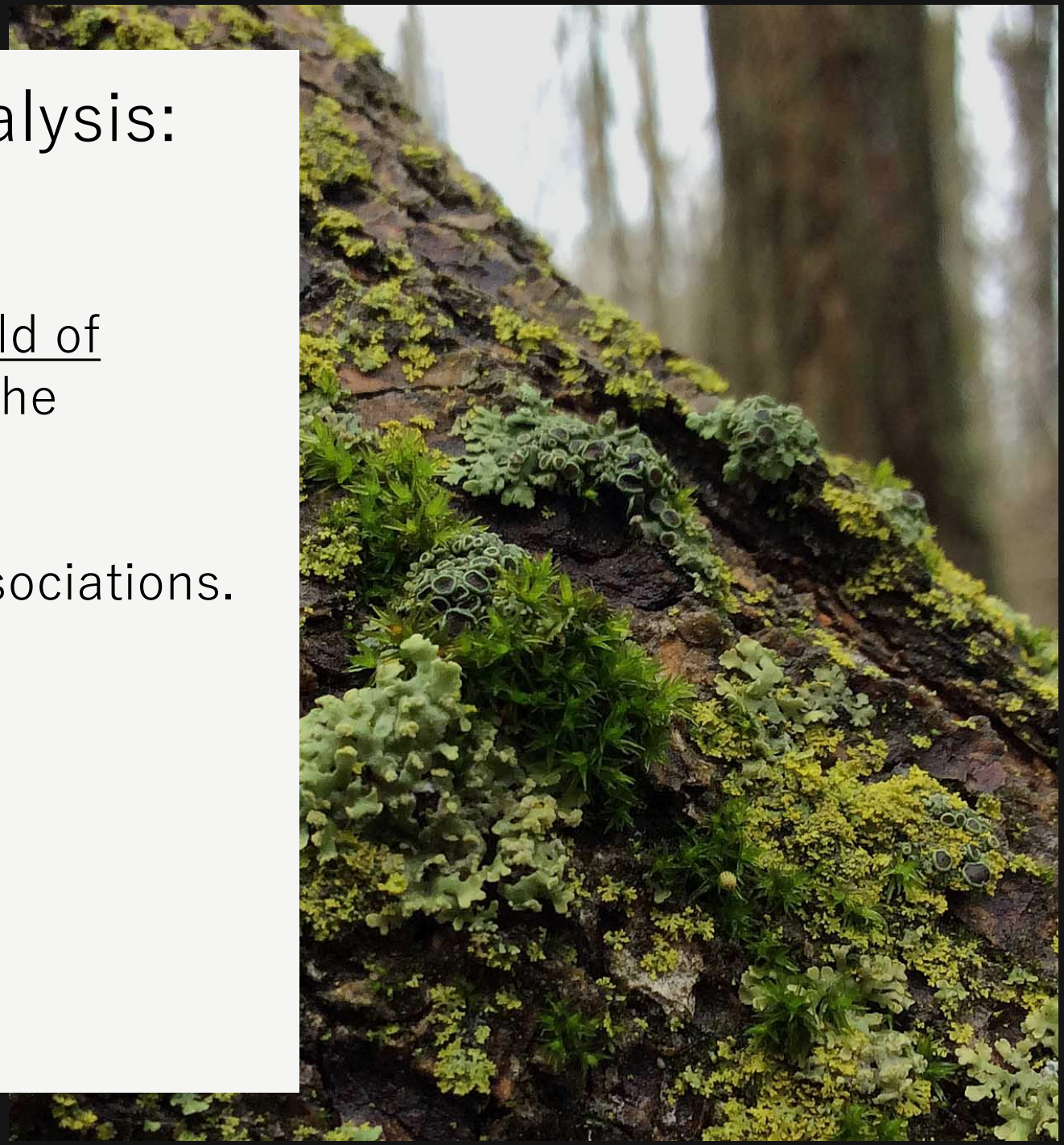
Identifies clusters formed at a threshold of similarity without taking into account the hierarchical cluster structure.



Non-Hierarchical Cluster Analysis: Complete Linkage

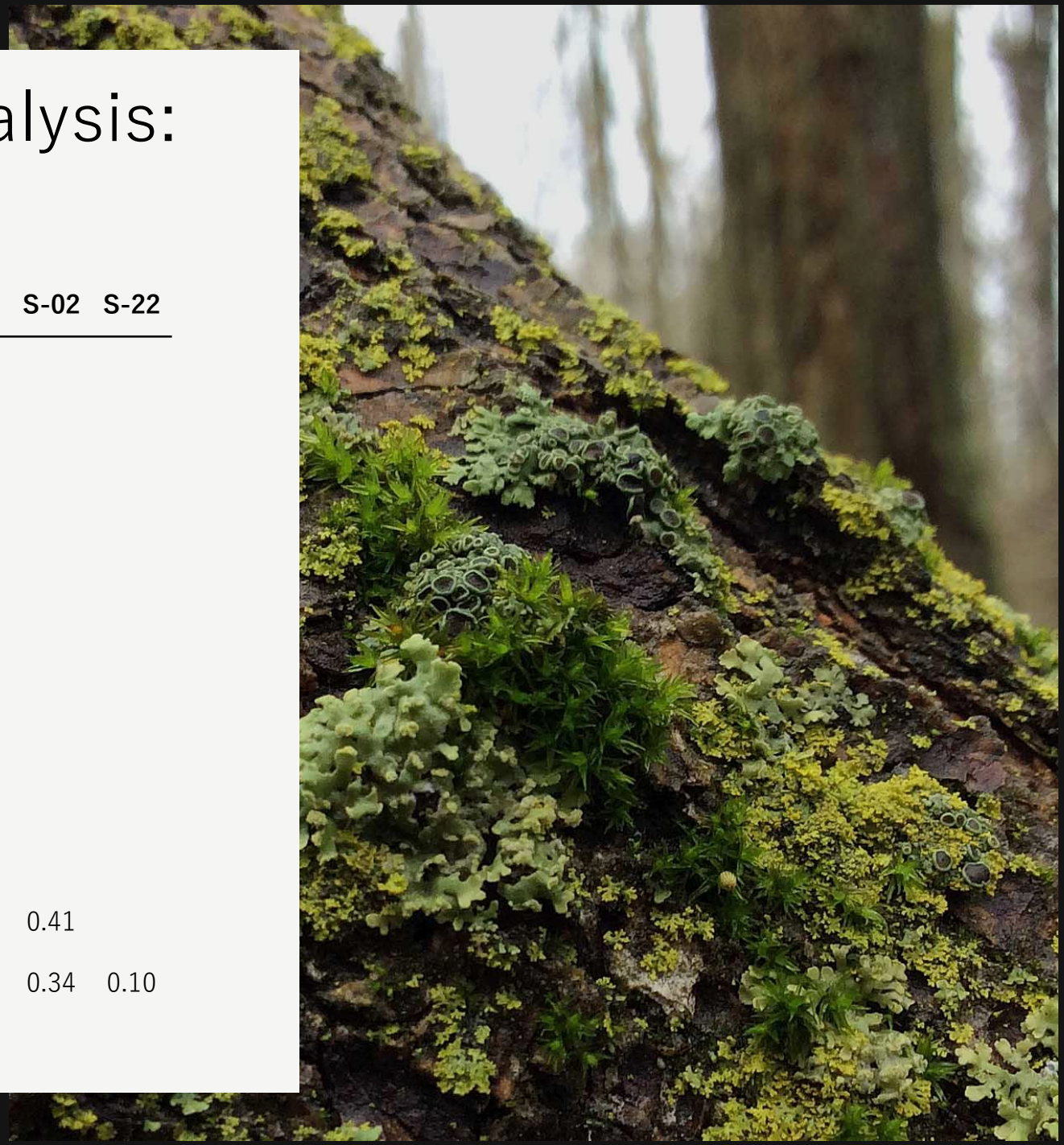
Identifies clusters formed at a threshold of similarity without taking into account the hierarchical cluster structure.

Generally used to discover species associations.



Non-Hierarchical Cluster Analysis: Complete Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10



Non-Hierarchical Cluster Analysis: Complete Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10

$S < 0.3$ **S-22 — S-18 — S-03**

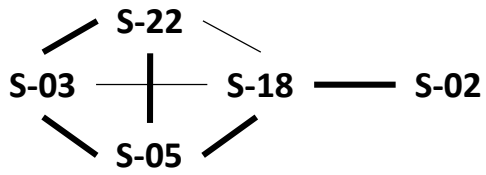
Non-Hierarchical Cluster Analysis: Complete Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10

$S < 0.3$

S-22 — S-18 — S-03

$S < 0.4$



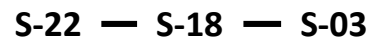
S-08 — S-21

Non-Hierarchical Cluster Analysis: Complete Linkage

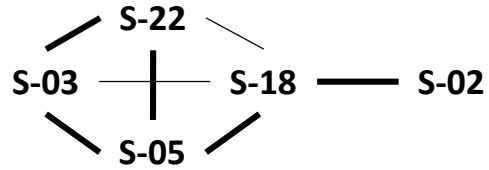
	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10

Either stop at a pre-determined threshold (0.5 is standard)…

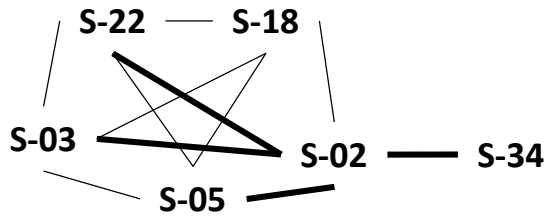
$S < 0.3$

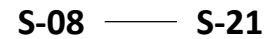


$S < 0.4$



$S < 0.5$



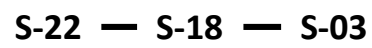


Non-Hierarchical Cluster Analysis: Complete Linkage

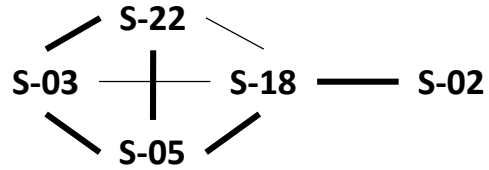
	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10

Continue until all sites are assigned to a cluster…

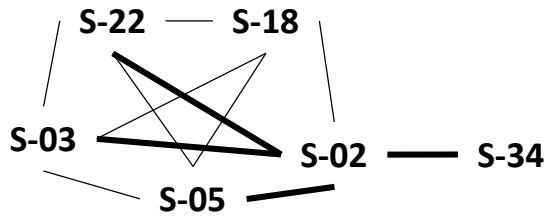
$S < 0.3$

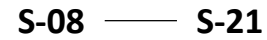


$S < 0.4$

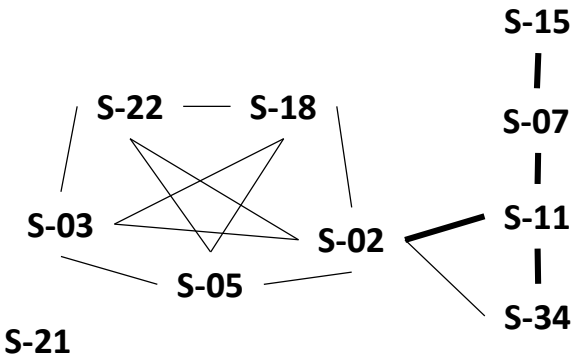


$S < 0.5$



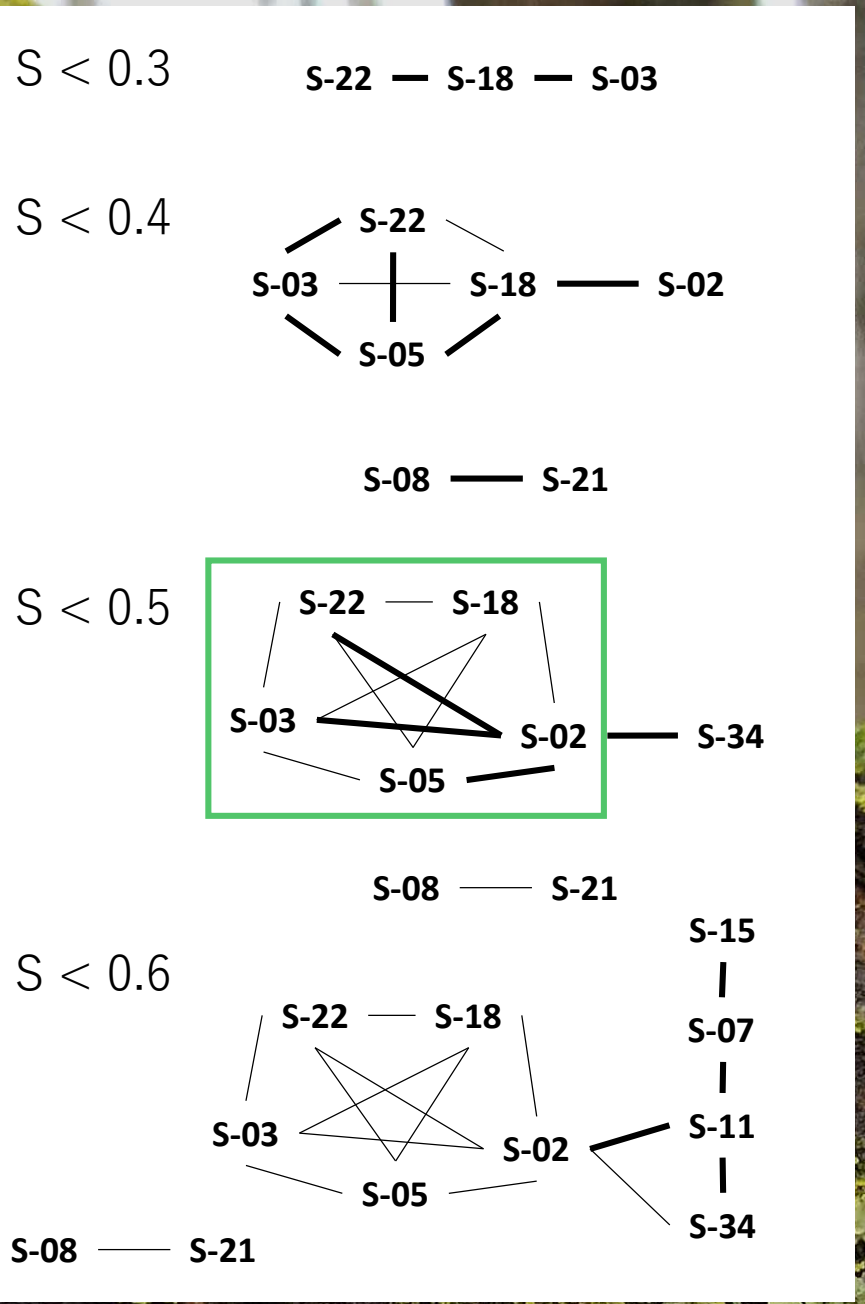


$S < 0.6$



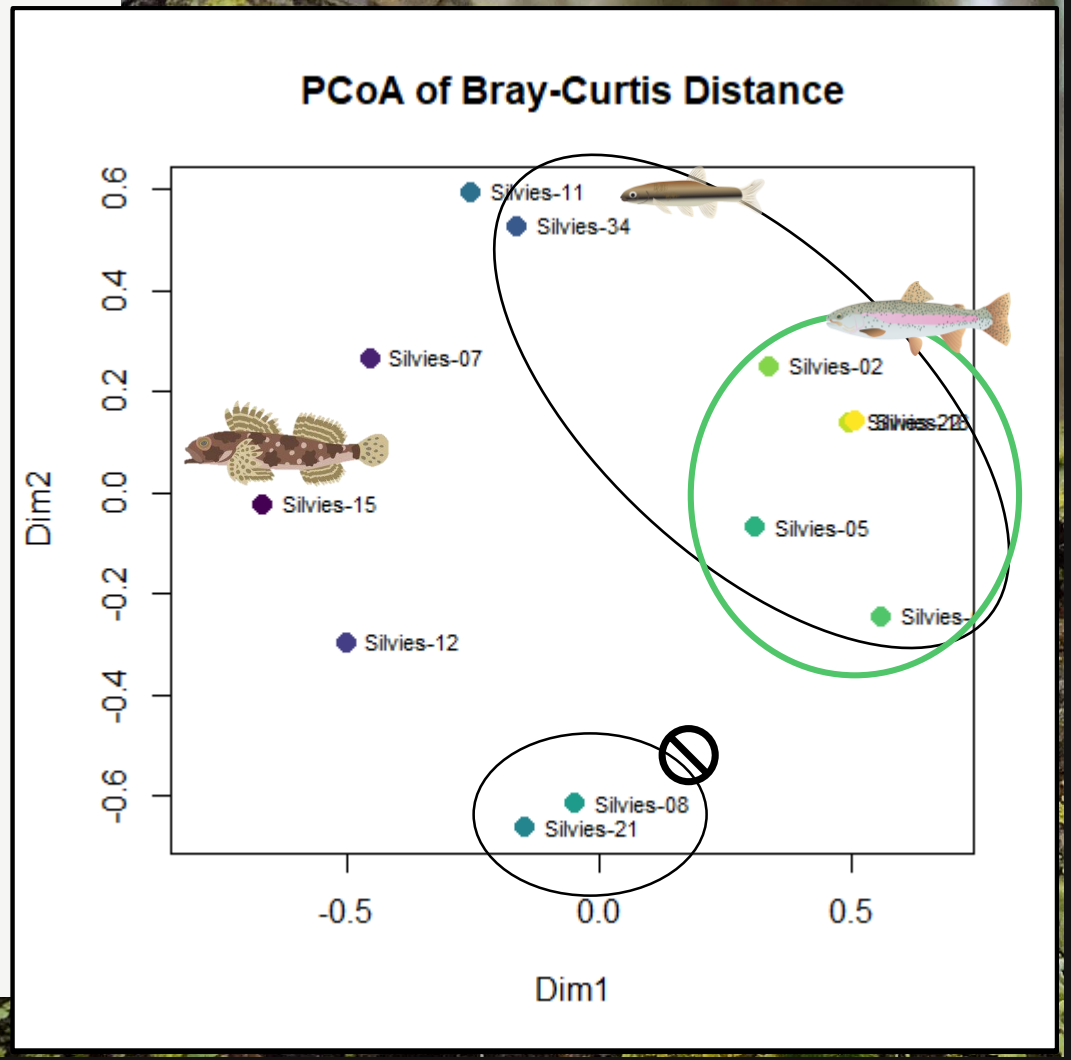
Non-Hierarchical Cluster Analysis: Complete Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02	S-22
Silvies-07	0.52										
Silvies-12	0.73	0.72									
Silvies-34	1.00	0.62	1.00								
Silvies-11	0.88	0.55	1.00	0.54							
Silvies-21	1.00	0.87	0.88	1.00	1.00						
Silvies-08	1.00	0.88	0.89	0.93	1.00	0.33					
Silvies-05	0.75	0.72	0.76	0.80	1.00	0.90	0.81				
Silvies-03	1.00	0.91	0.93	1.00	1.00	0.80	0.73	0.36			
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	0.41		
Silvies-22	1.00	0.76	0.96	0.81	0.75	0.90	0.86	0.38	0.38	0.41	
Silvies-18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34	0.10



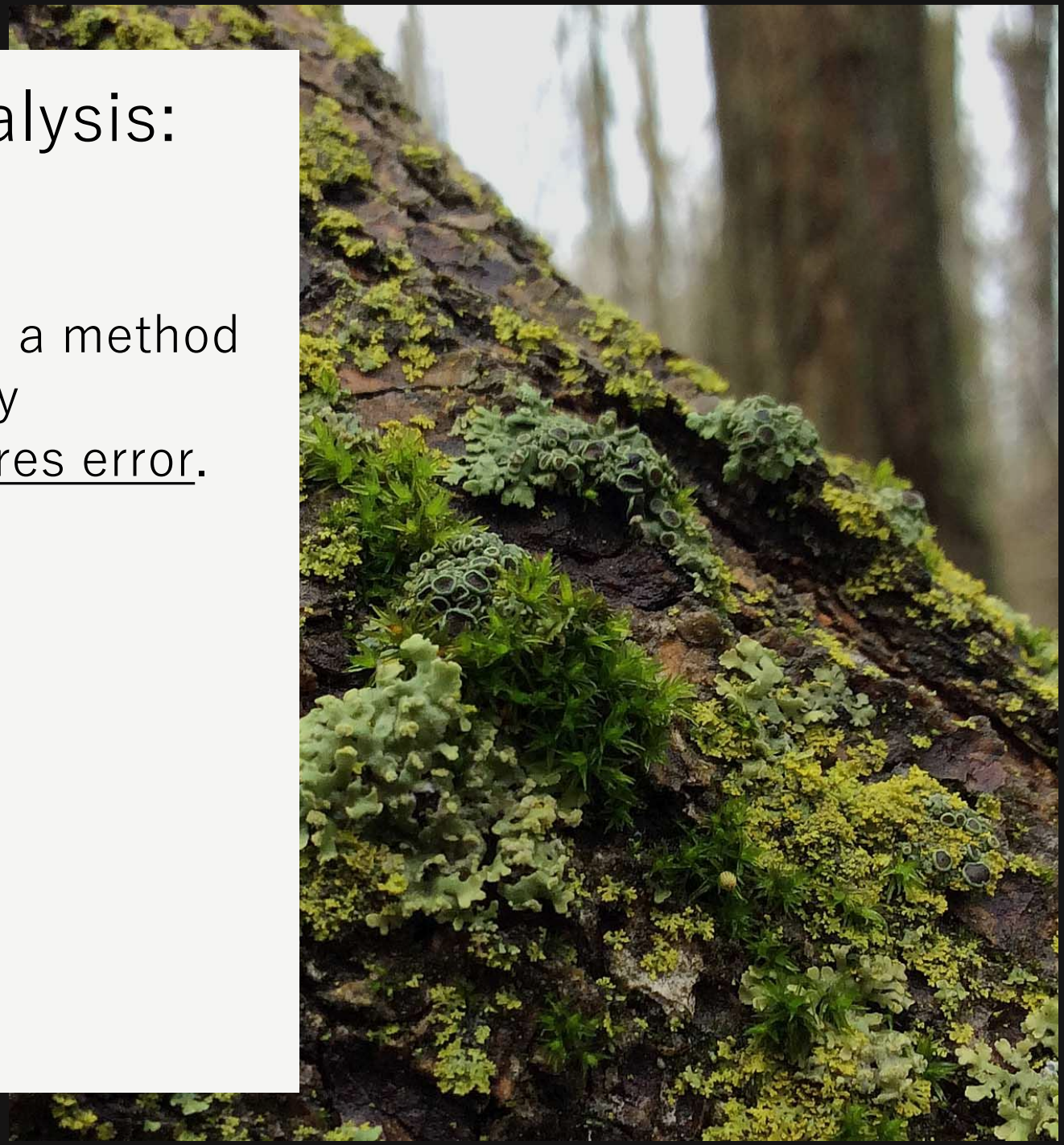
Non-Hierarchical Cluster Analysis: Complete Linkage

What do we think of this clustering outcome for the 0.5 level?



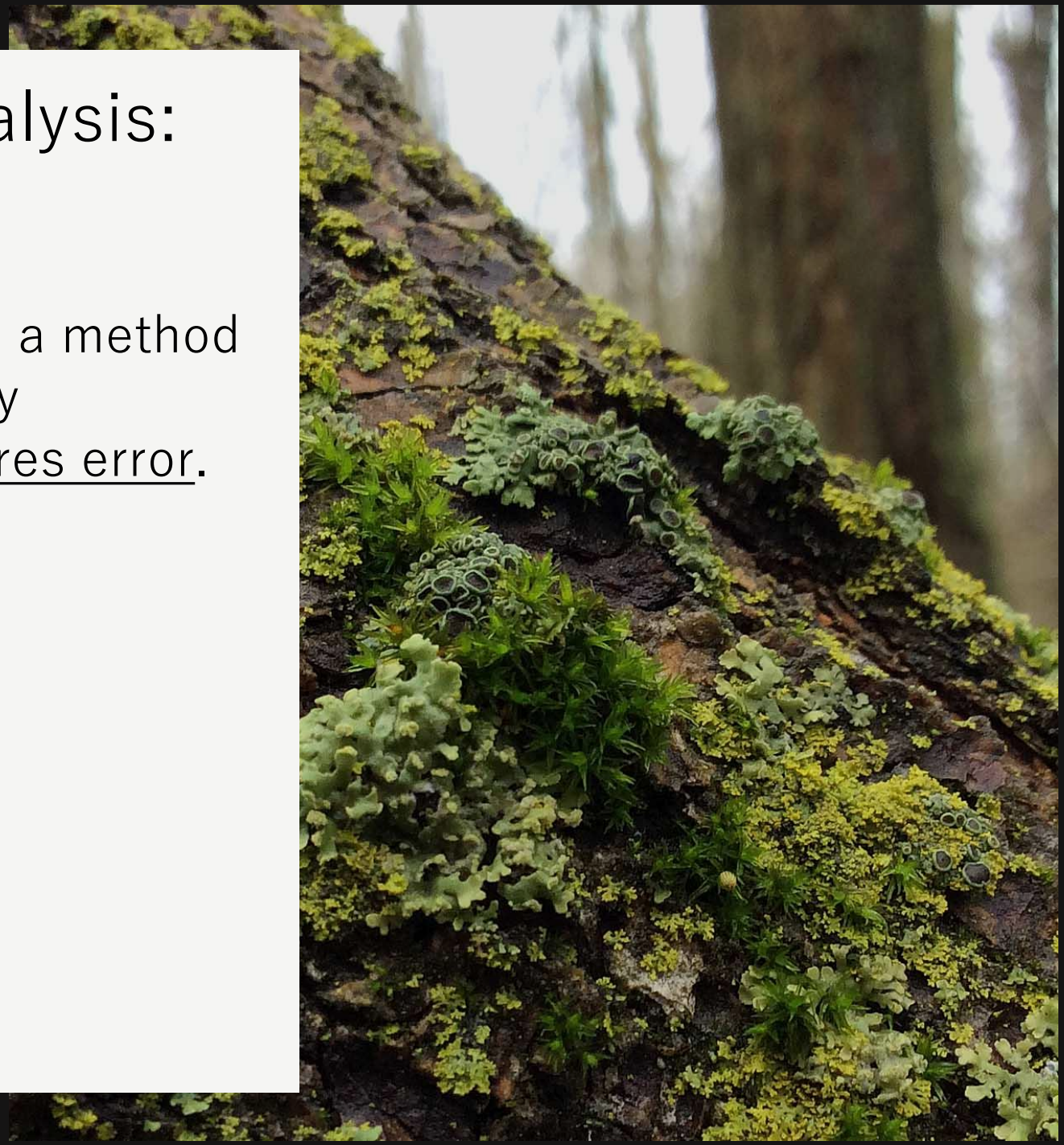
Non-Hierarchical Cluster Analysis: K-means Partitioning

K-means partitioning or clustering is a method used to partition data into *k* clusters by minimizing within-cluster sum of squares error.



Non-Hierarchical Cluster Analysis: K-means Partitioning

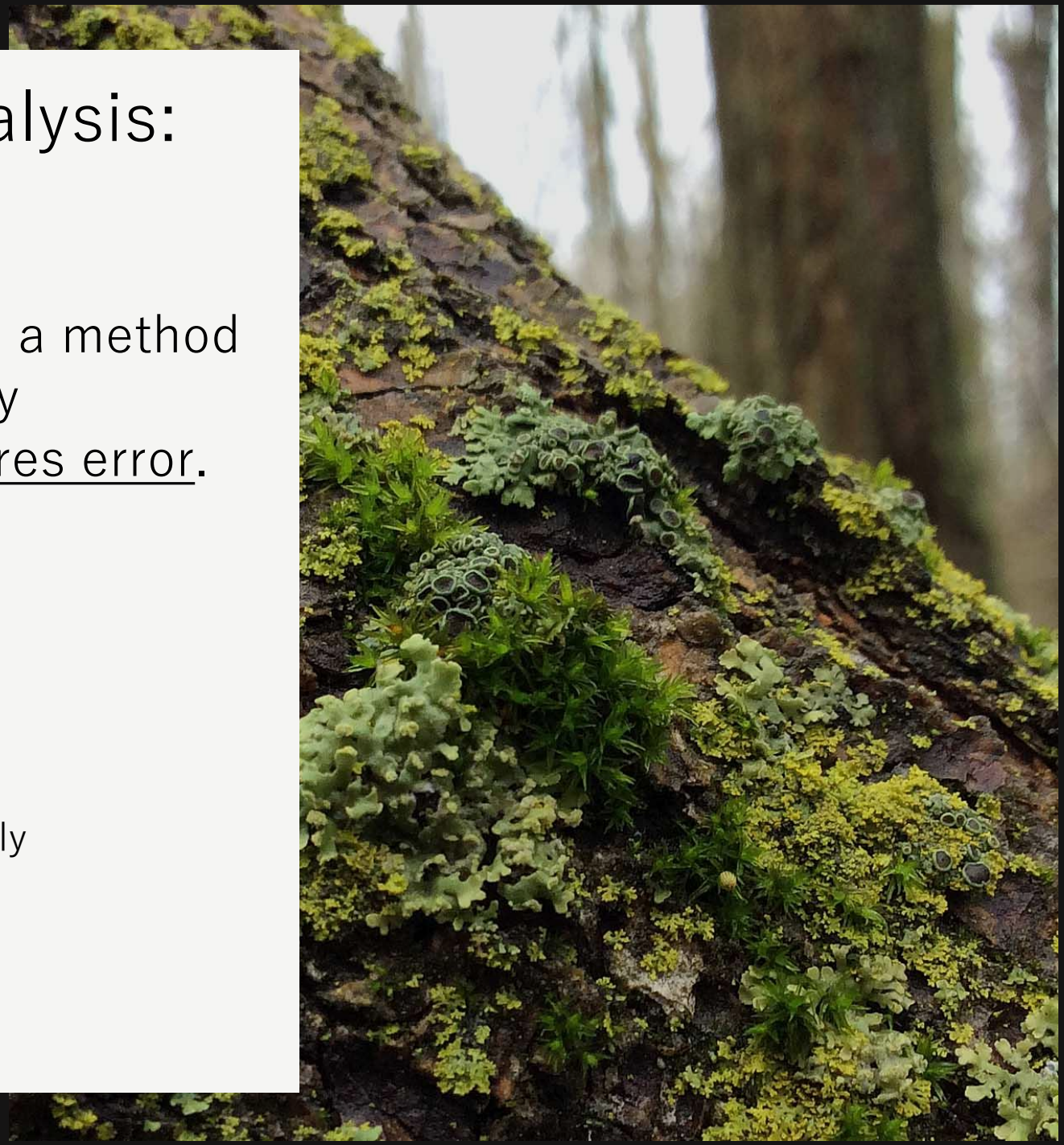
K-means partitioning or clustering is a method used to partition data into *k* clusters by minimizing within-cluster sum of squares error.



Non-Hierarchical Cluster Analysis: K-means Partitioning

K-means partitioning or clustering is a method used to partition data into *k* clusters by minimizing within-cluster sum of squares error.

1. Randomly select *k* initial centroids
2. Assign each data point to nearest centroid
3. Calculate new centroids
4. Repeat until centroids do not change significantly

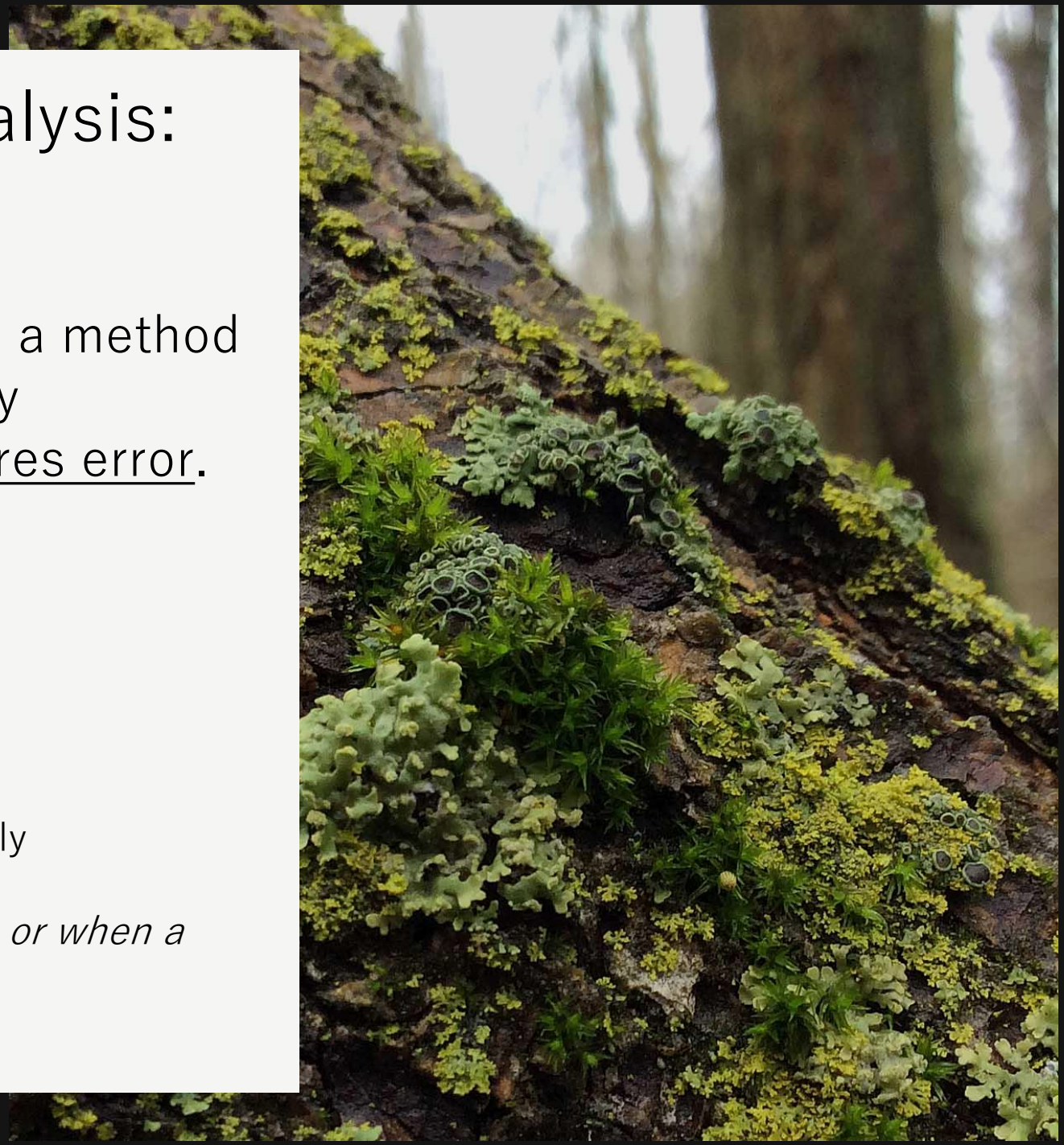


Non-Hierarchical Cluster Analysis: K-means Partitioning

K-means partitioning or clustering is a method used to partition data into *k* clusters by minimizing within-cluster sum of squares error.

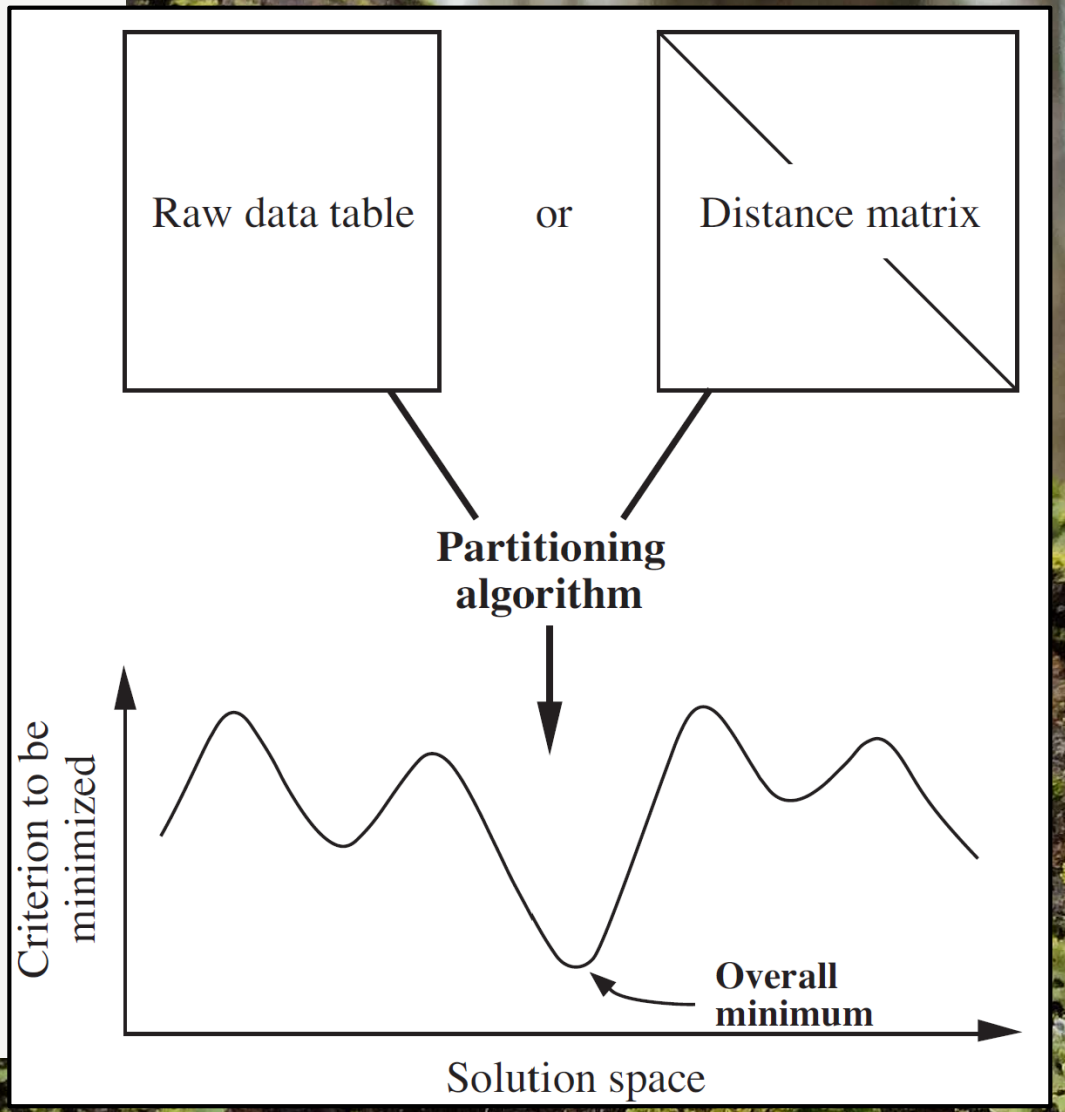
1. Randomly select *k* initial centroids
2. Assign each data point to nearest centroid
3. Calculate new centroids
4. Repeat until centroids do not change significantly

Convergence occurs when centroids do not change or when a maximum number of iterations is reached.



Non-Hierarchical Cluster Analysis: K-means Partitioning

The "local minimum" problem: the solution on which the computation eventually converges depends on the initial centroid positions.

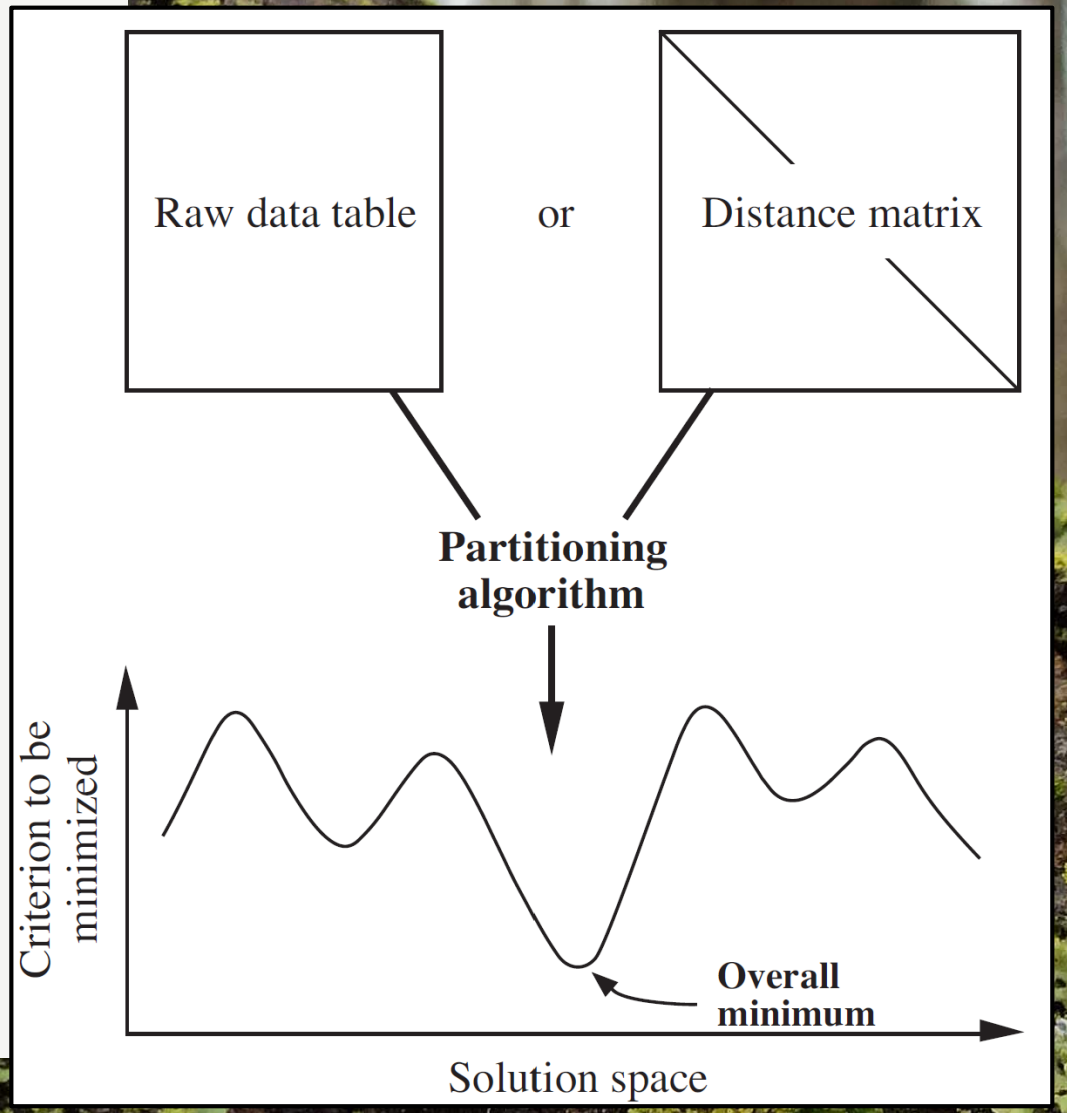


Legendre & Legendre Fig. 8.17

Non-Hierarchical Cluster Analysis: K-means Partitioning

The "local minimum" problem: the solution on which the computation eventually converges depends on the initial centroid positions.

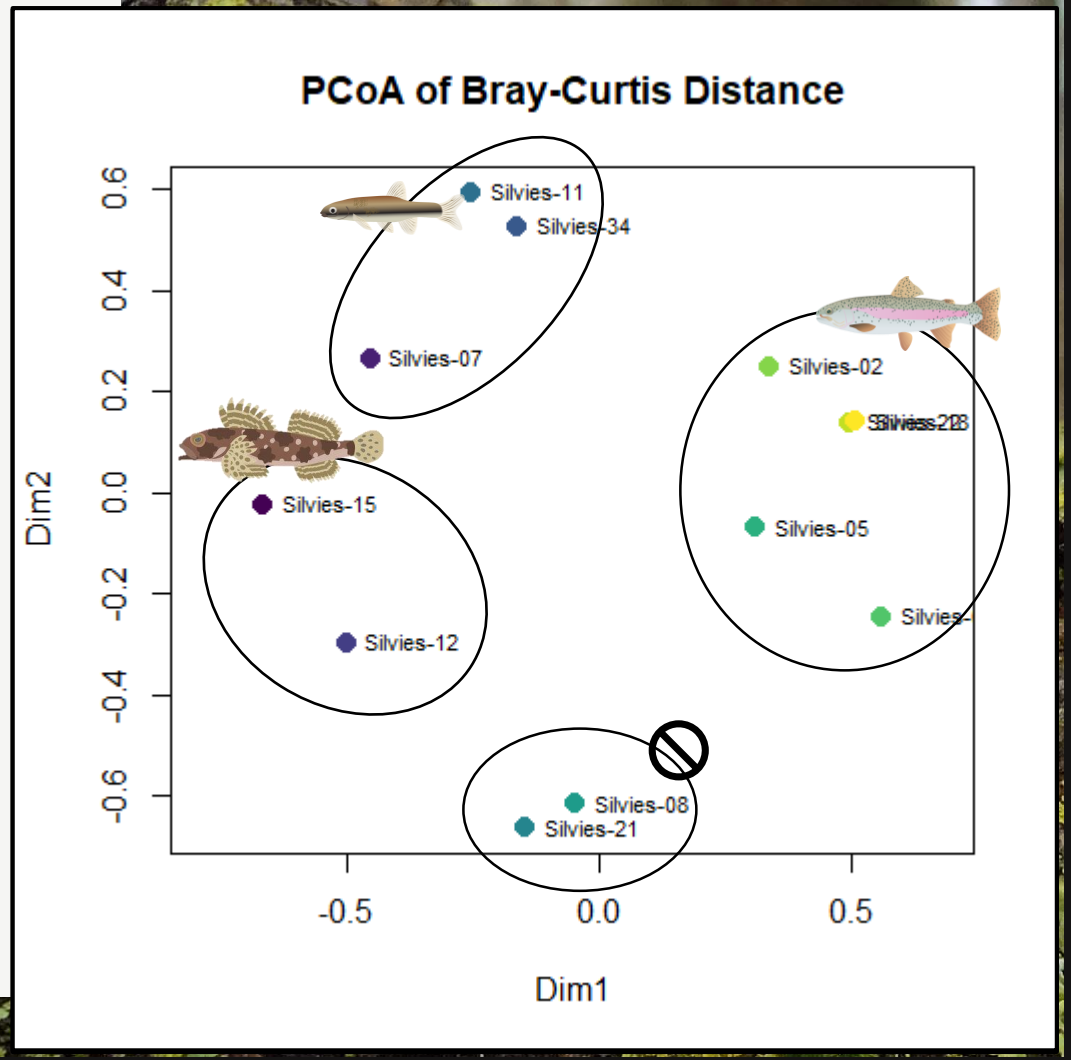
- Pre-set centroids
- Run procedure many times and retain solution that minimizes SSE



Legendre & Legendre Fig. 8.17

Non-Hierarchical Cluster Analysis: K-means Partitioning

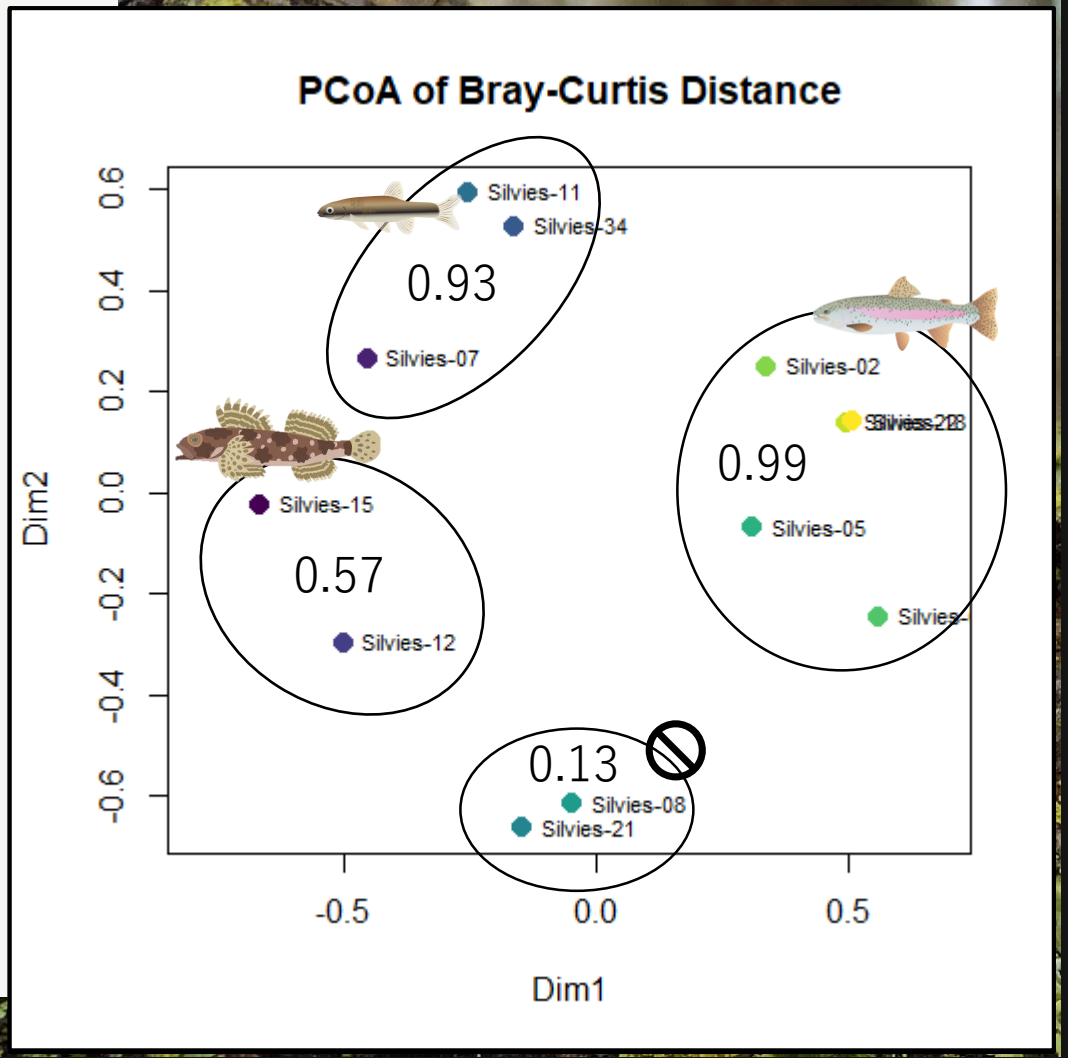
K-means partitioning or clustering is a method used to partition data into *k* clusters by minimizing within-cluster sum of squares error.



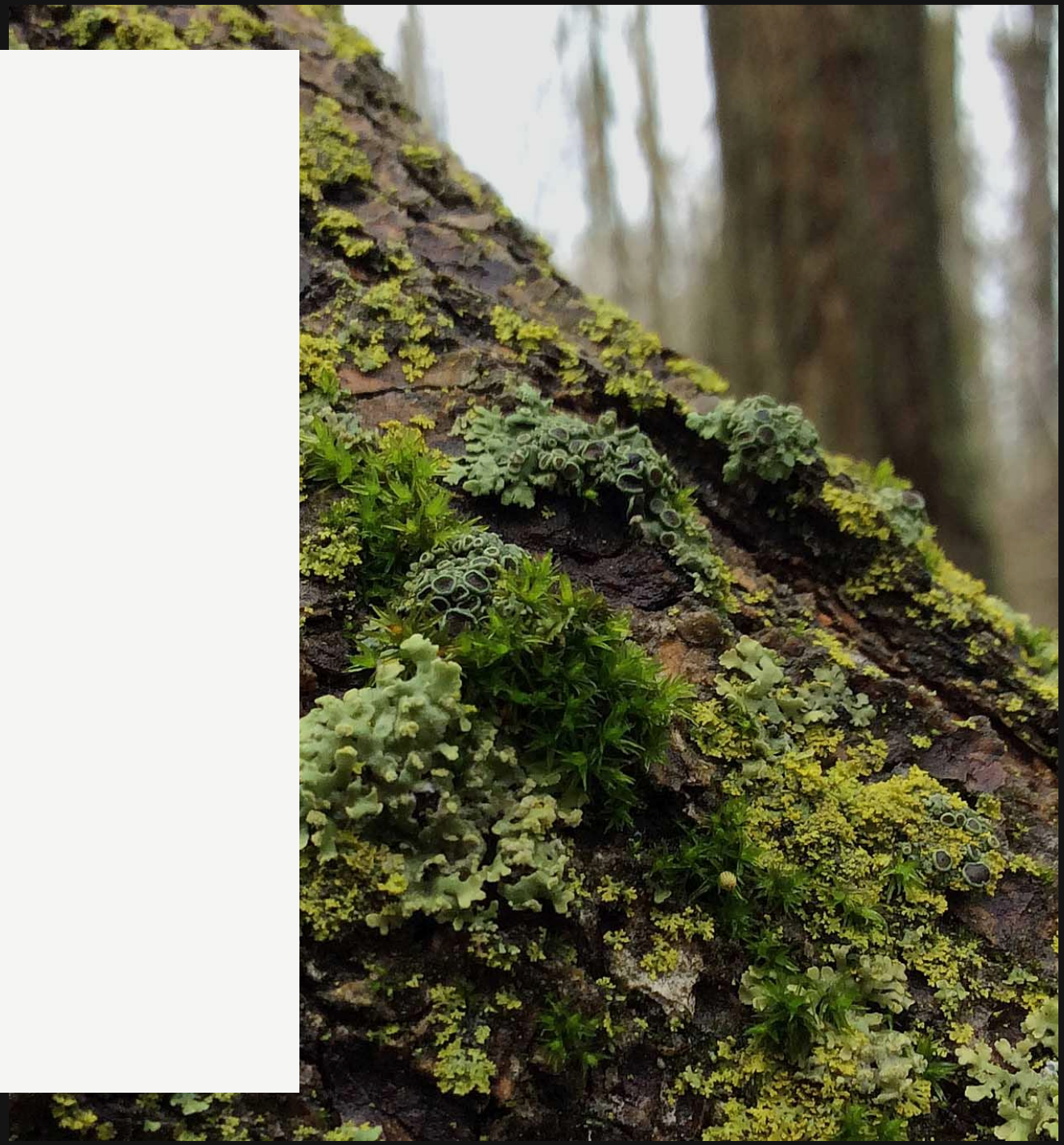
Non-Hierarchical Cluster Analysis: K-means Partitioning

K-means partitioning or clustering is a method used to partition data into *k* clusters by minimizing within-cluster sum of squares error.

Within Cluster SSE

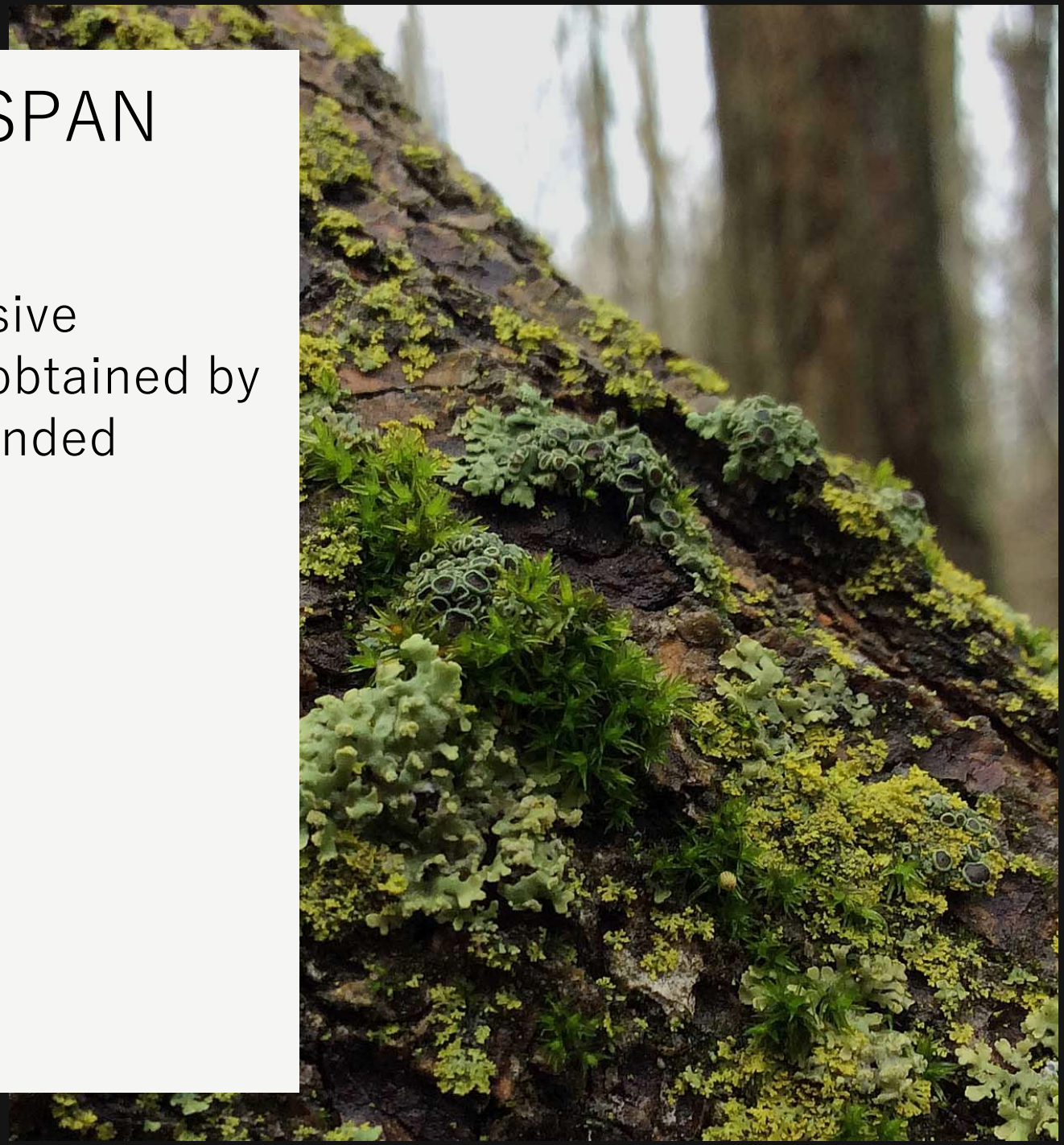


Species Associations



Species Associations: TWINSPAN

Two-Way Indicator Species Analysis (TWINSPAN) is based on the progressive refinement of a single ordination axis obtained by correspondence analysis (CA) or detrended correspondence analysis (DCA).



Species Associations: TWINSPAN

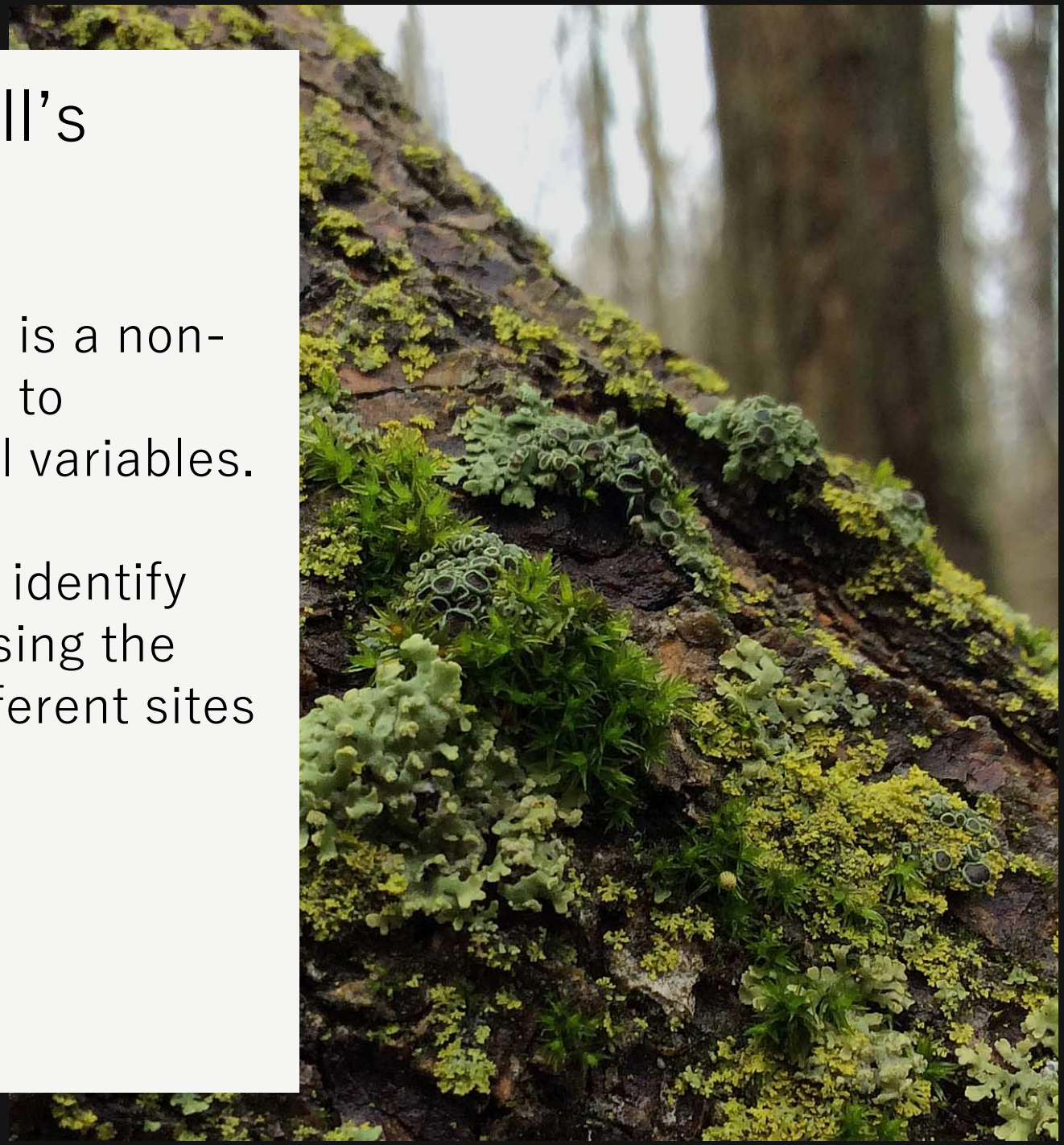
Two-Way Indicator Species Analysis (TWINSPAN) is based on the progressive refinement of a single ordination axis obtained by correspondence analysis (CA) or detrended correspondence analysis (DCA).

Simultaneously sorts sites/objects and indicator species/descriptors

Species Associations: Kendall’s Coefficient of Concordance

Kendall's coefficient of concordance is a non-parametric (rank-based) statistic used to measure the agreement among several variables.

In ecological studies, it can be used to identify significant groups of species by assessing the agreement in their rankings across different sites or conditions.

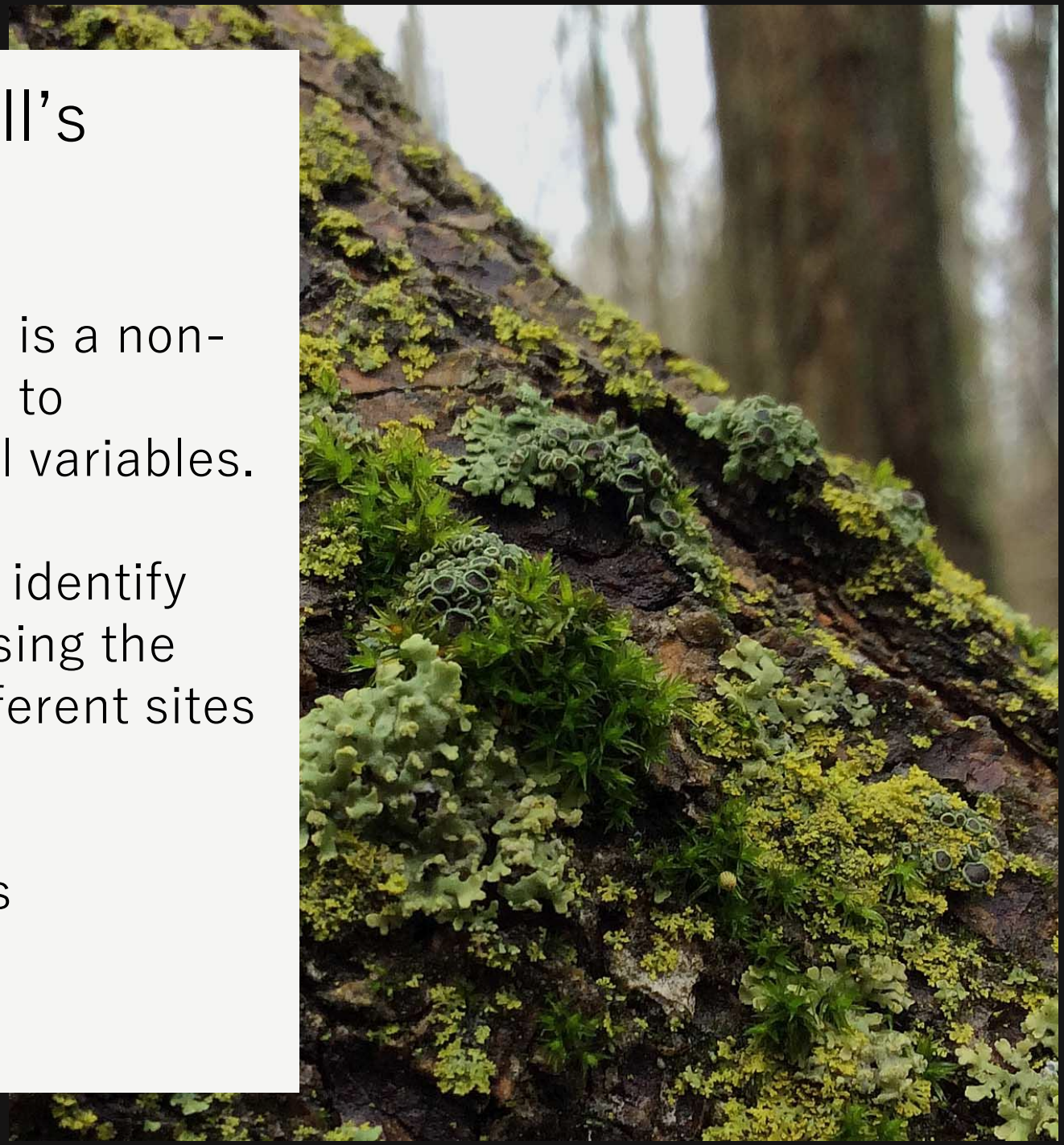


Species Associations: Kendall's Coefficient of Concordance

Kendall's coefficient of concordance is a non-parametric (rank-based) statistic used to measure the agreement among several variables.

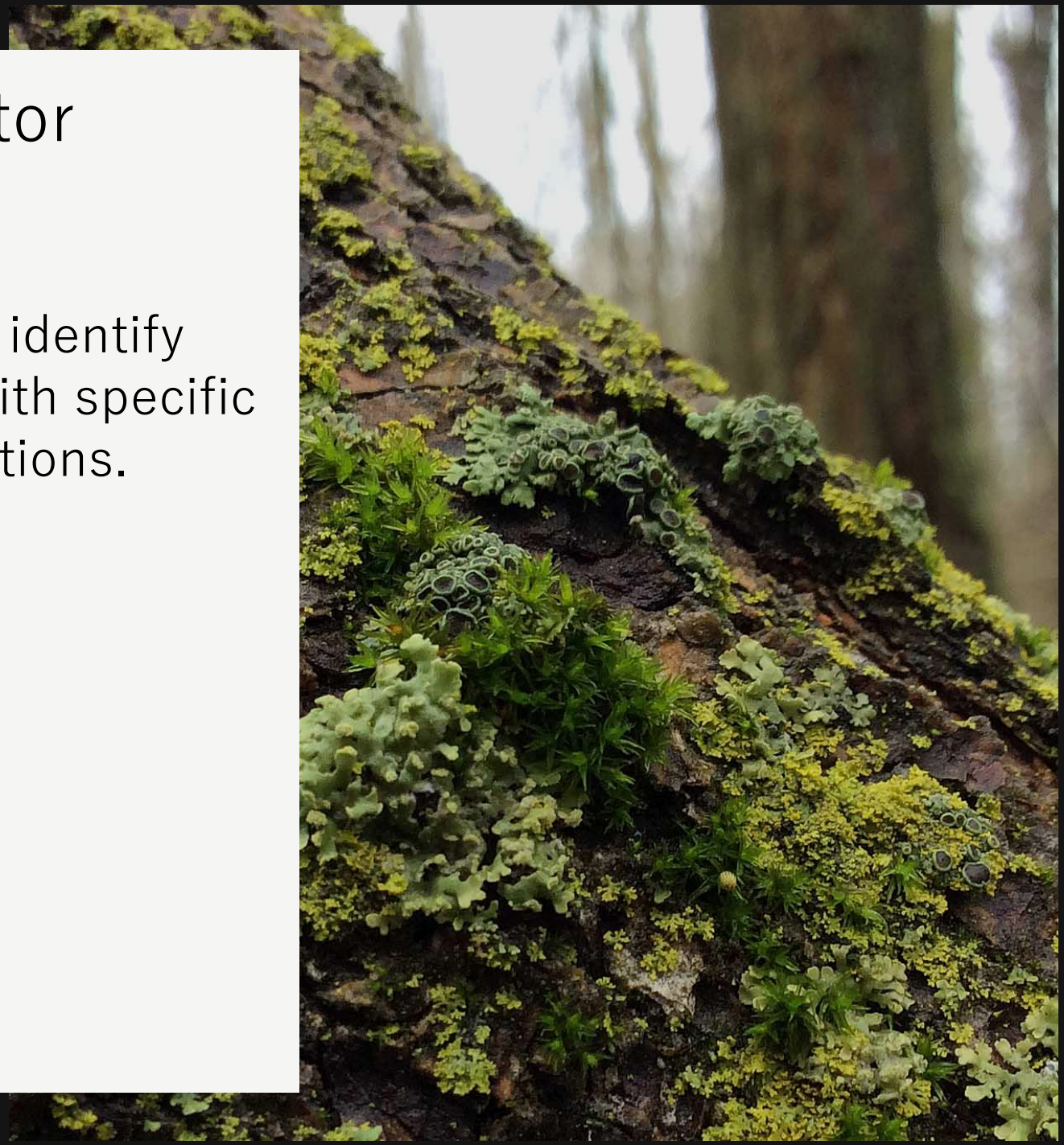
In ecological studies, it can be used to identify significant groups of species by assessing the agreement in their rankings across different sites or conditions.

This is an **R-mode** analysis for species abundance data!



Species Associations: Indicator Species

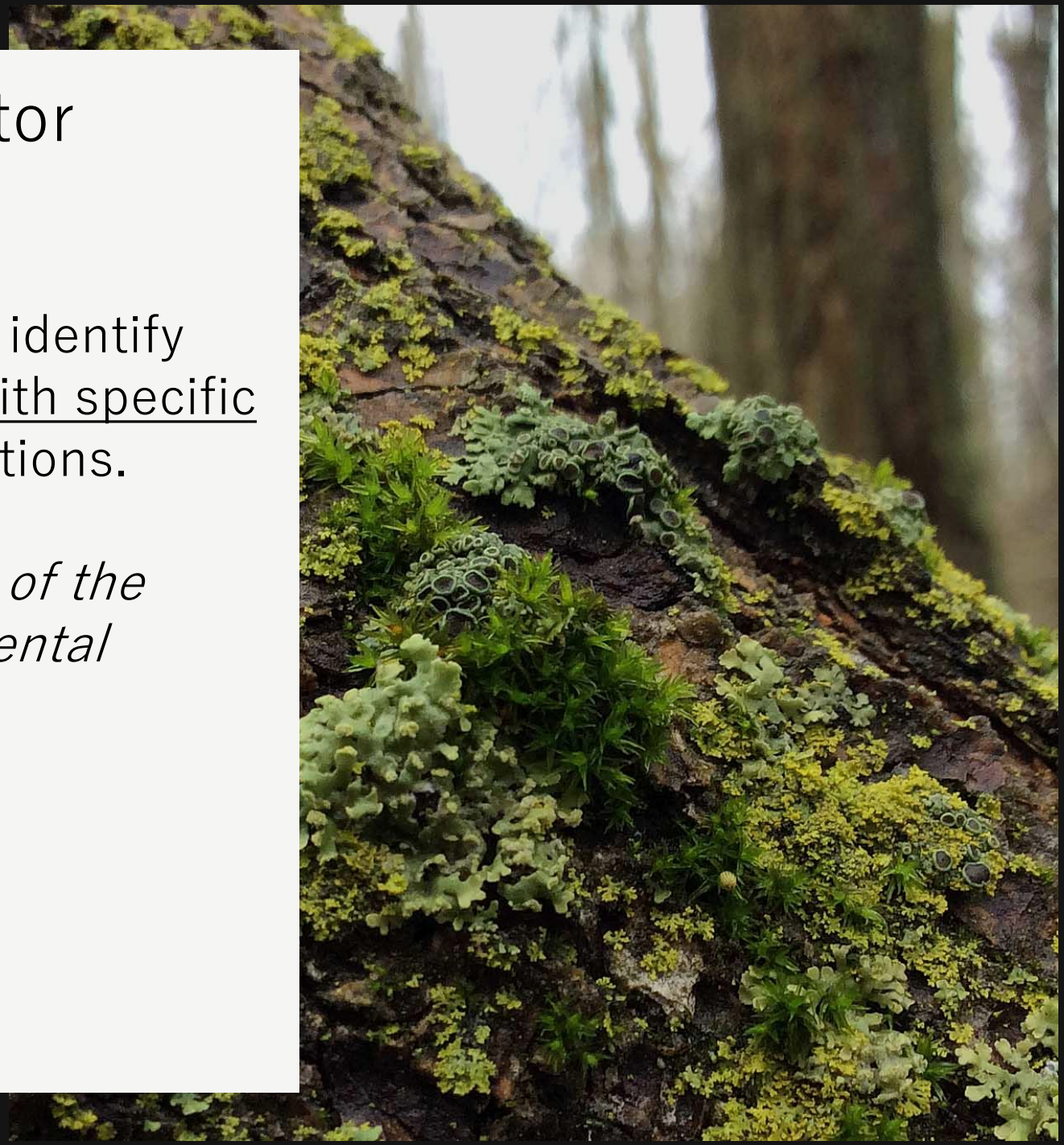
Indicator Species Analysis is used to identify species that are strongly associated with specific groups of sites or environmental conditions.



Species Associations: Indicator Species

Indicator Species Analysis is used to identify species that are strongly associated with specific groups of sites or environmental conditions.

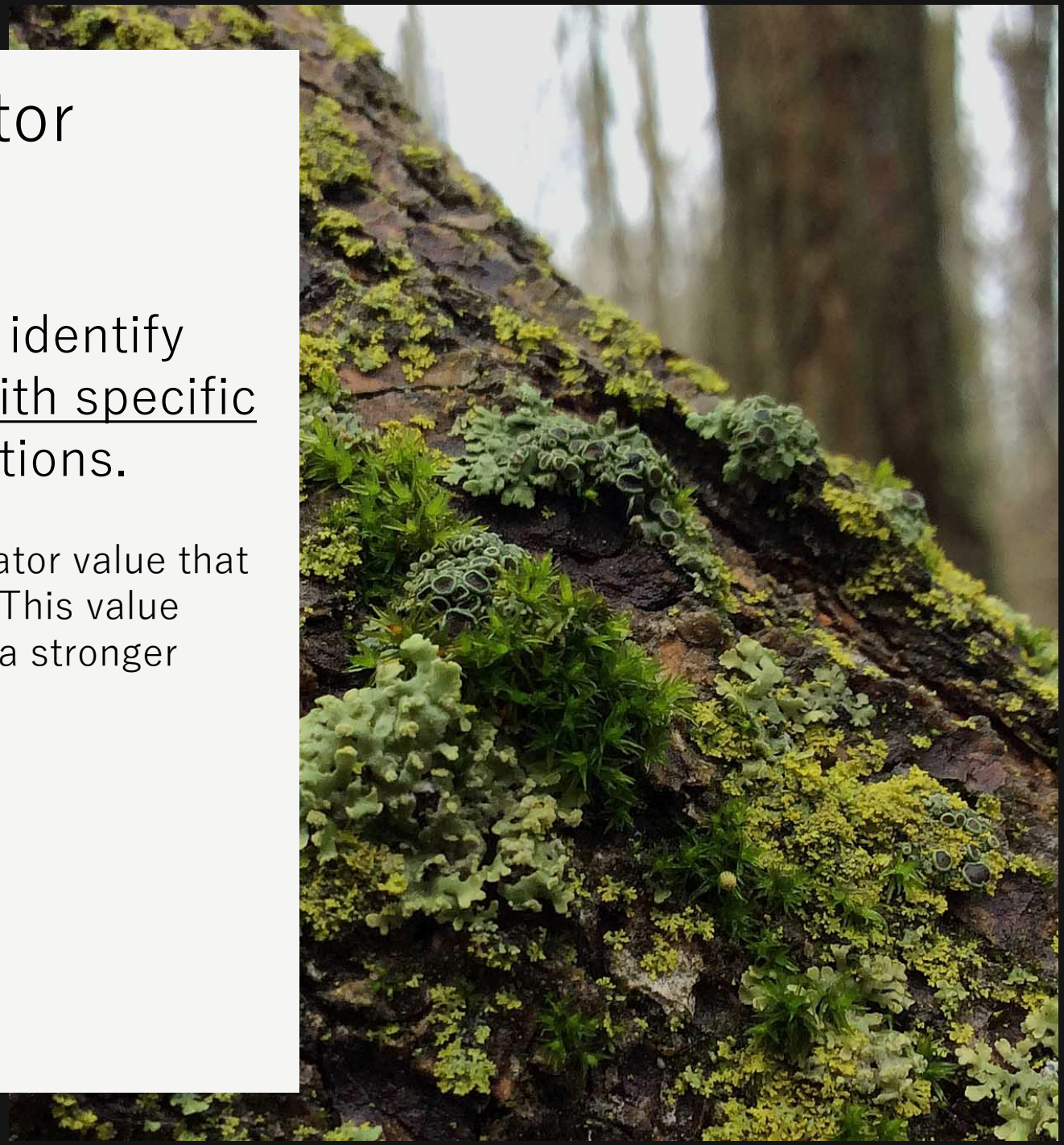
These species can serve as indicators of the ecological characteristics or environmental health of particular habitats.



Species Associations: Indicator Species

Indicator Species Analysis is used to identify species that are strongly associated with specific groups of sites or environmental conditions.

Indicator Value: Each species is assigned an indicator value that quantifies its association with specific site groups. This value ranges from 0 to 100, with higher values indicating a stronger association.

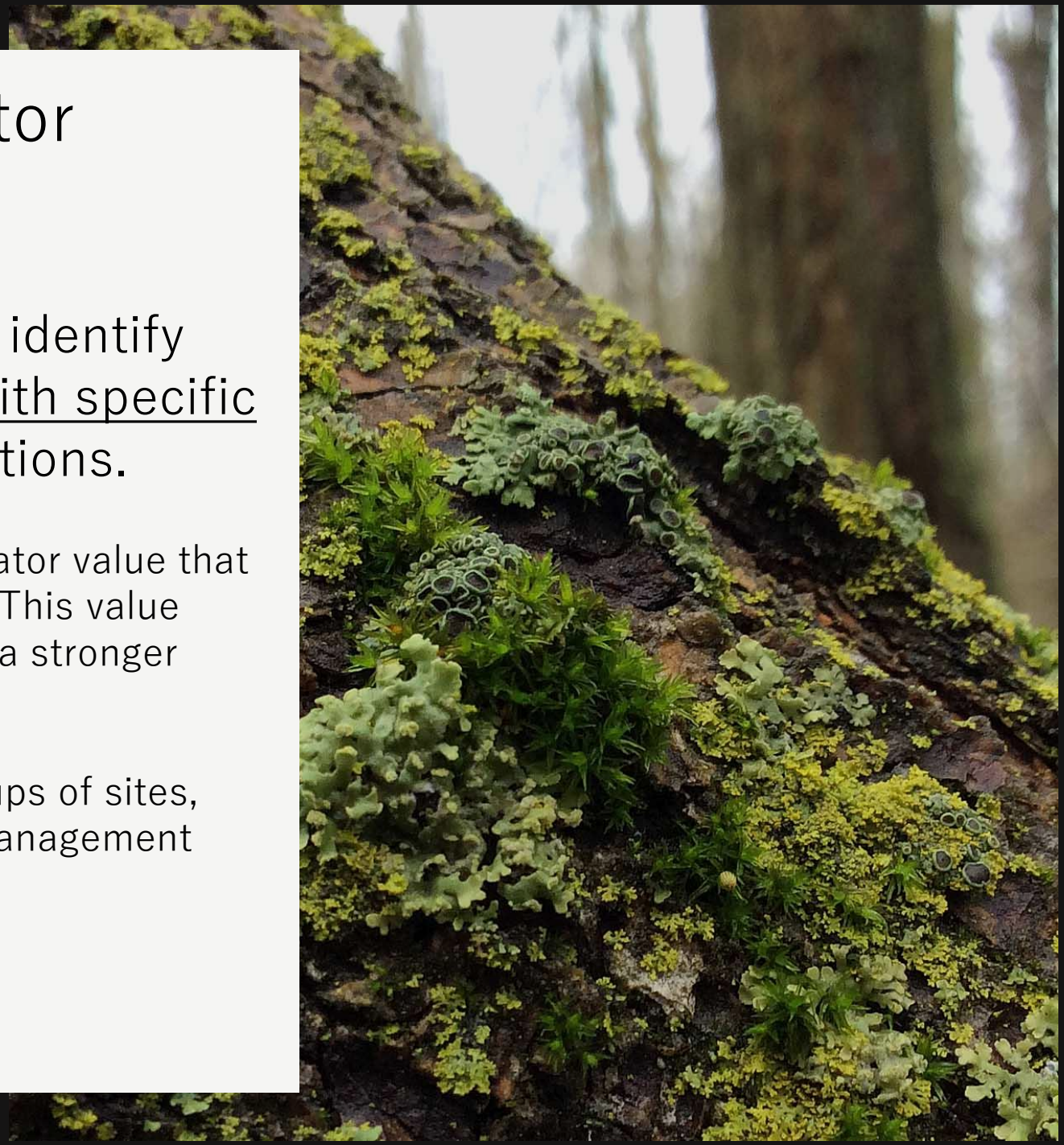


Species Associations: Indicator Species

Indicator Species Analysis is used to identify species that are strongly associated with specific groups of sites or environmental conditions.

Indicator Value: Each species is assigned an indicator value that quantifies its association with specific site groups. This value ranges from 0 to 100, with higher values indicating a stronger association.

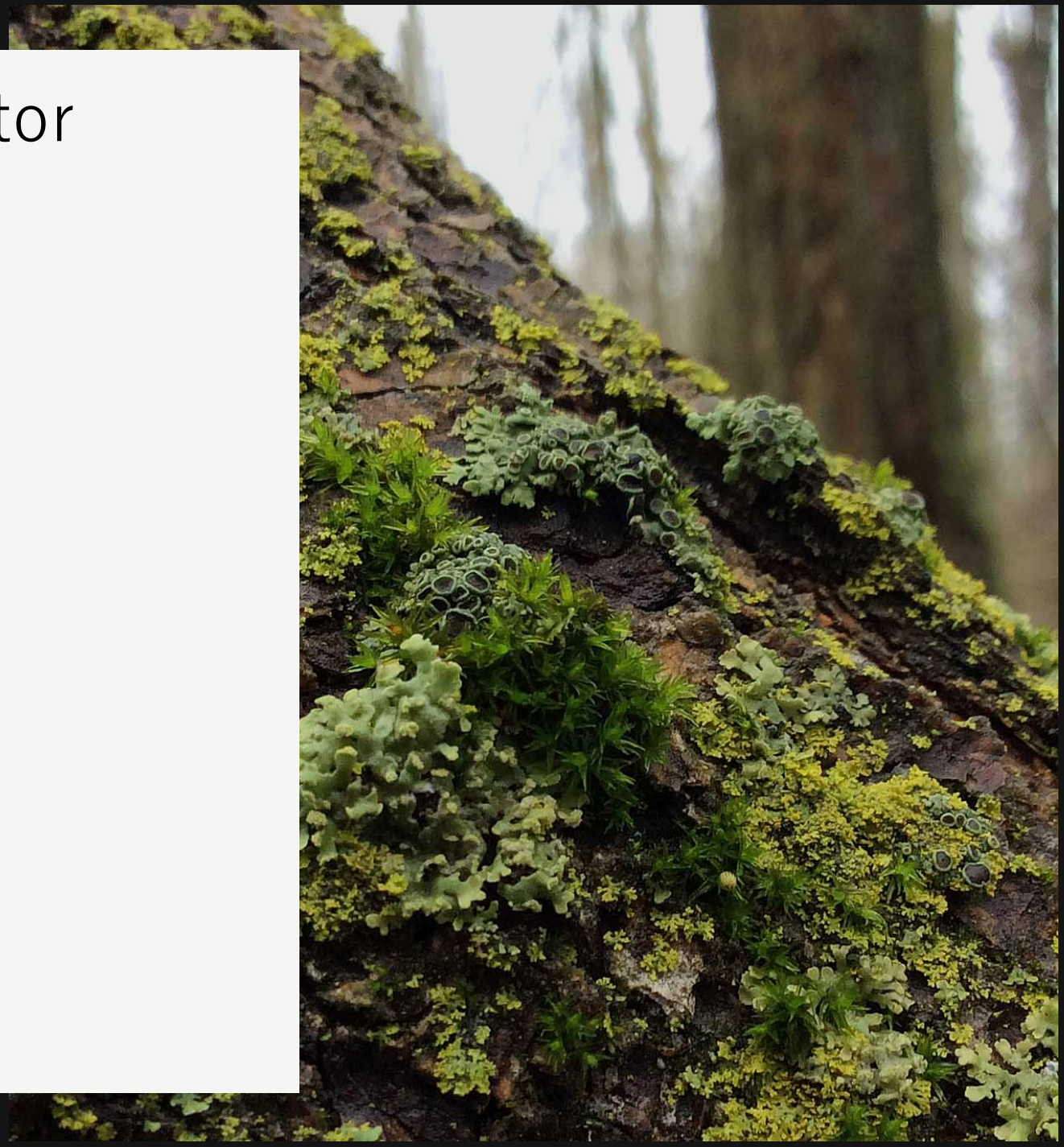
Site Groups: The analysis requires predefined groups of sites, which can be based on environmental gradients, management practices, or other criteria.



Species Associations: Indicator Species

Indicator Species Analysis

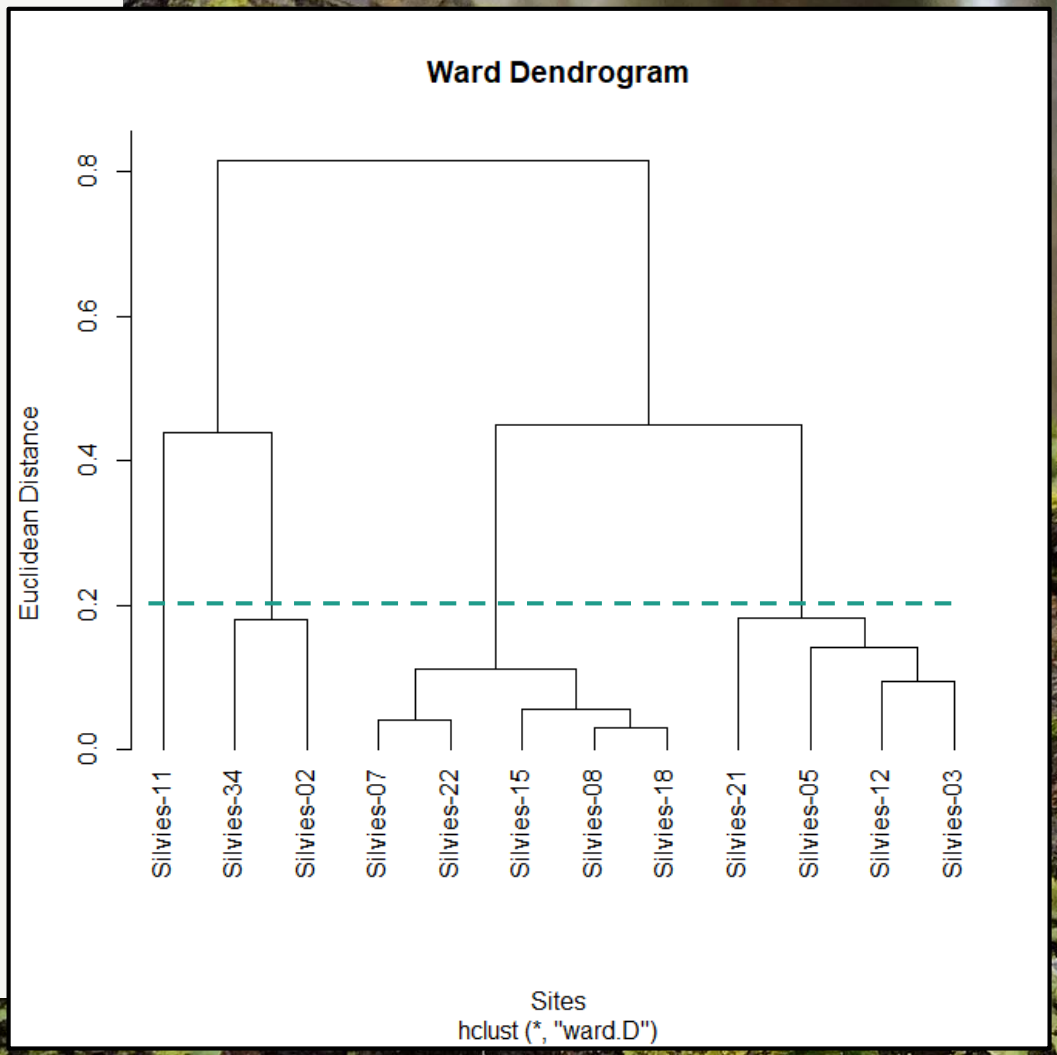
1. Define **site groups** based on relevant criteria
2. Calculate **indicator values**
3. Assess significance
4. Interpret results



Species Associations: Indicator Species

Indicator Species Analysis

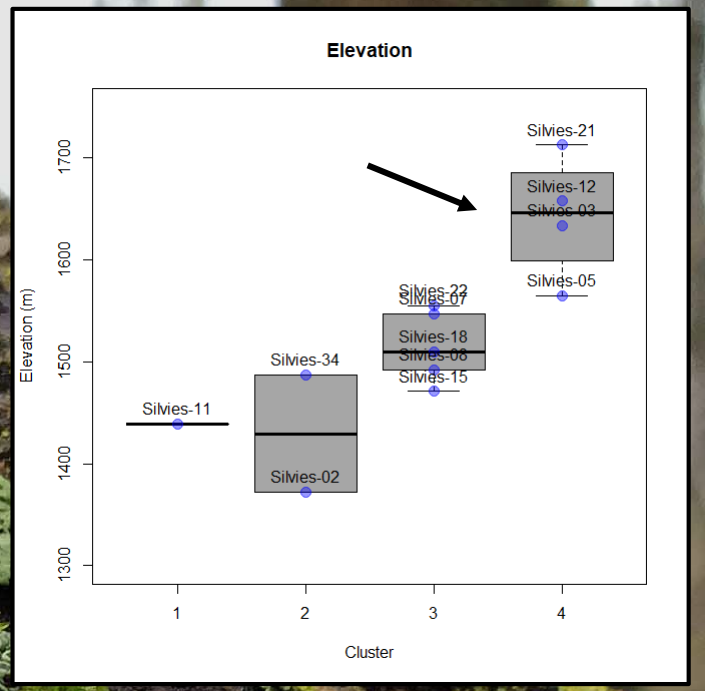
1. Define **site groups** based on relevant criteria
2. Calculate **indicator values**
3. Assess significance
4. Interpret results

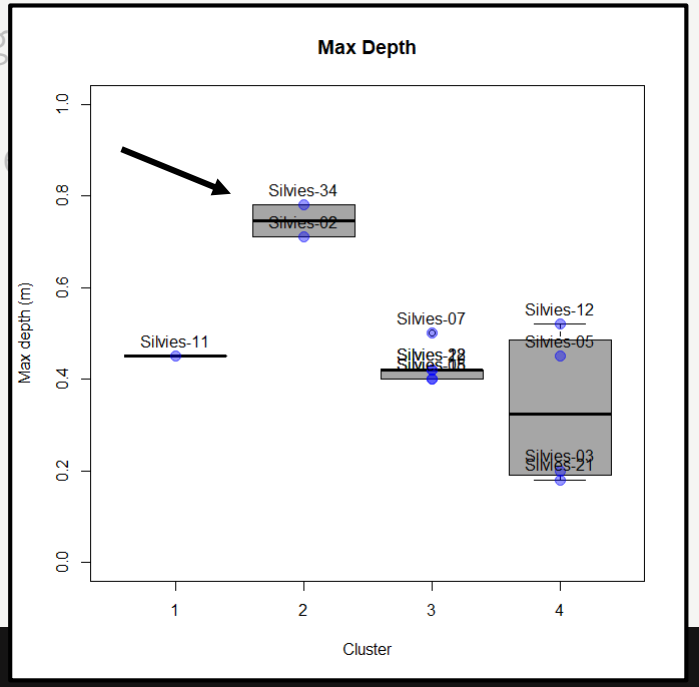


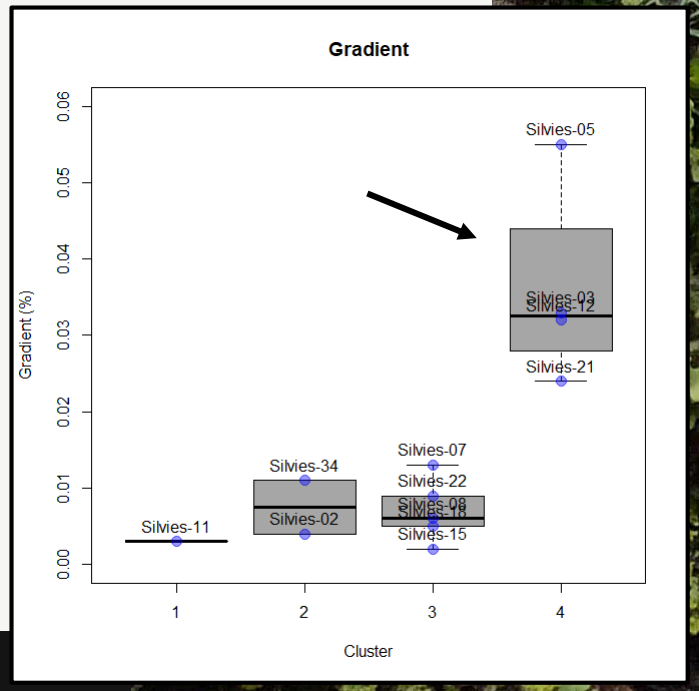
Species Associations: Indicator Species

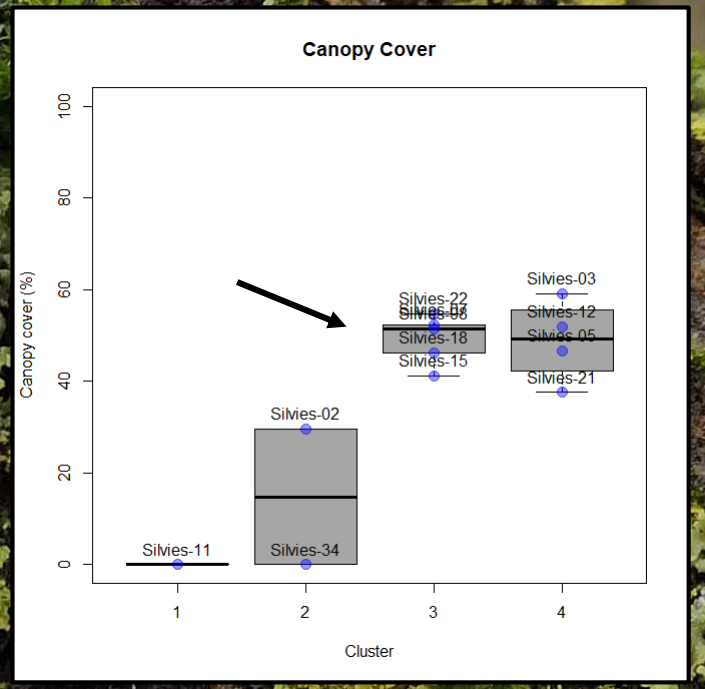
Indicator Species Analysis

1. Define **site groups** based on relevant criteria
2. Calculate **indicator values**
3. Assess sig
4. Interpret r







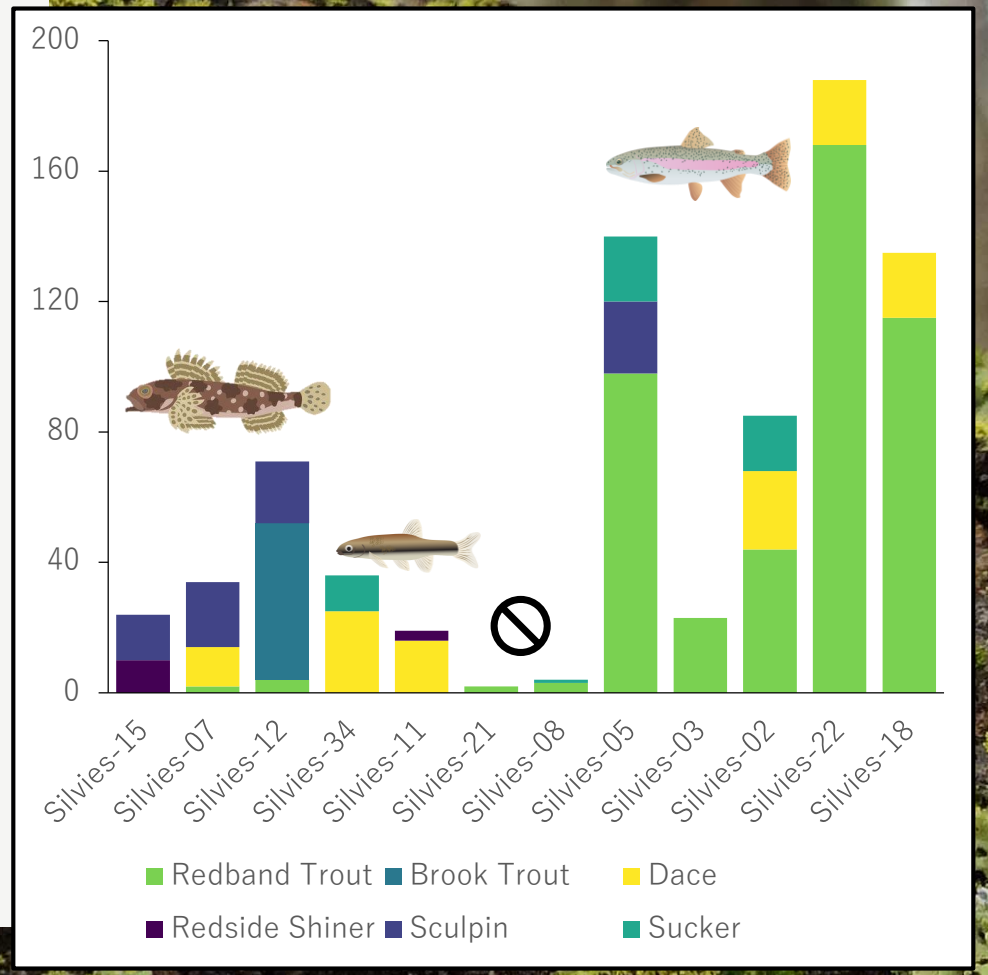


Species Associations: Indicator Species

Indicator Species Analysis

1. Define **site groups** based on relevant criteria
2. Calculate **indicator values**
3. Assess significance
4. Interpret results

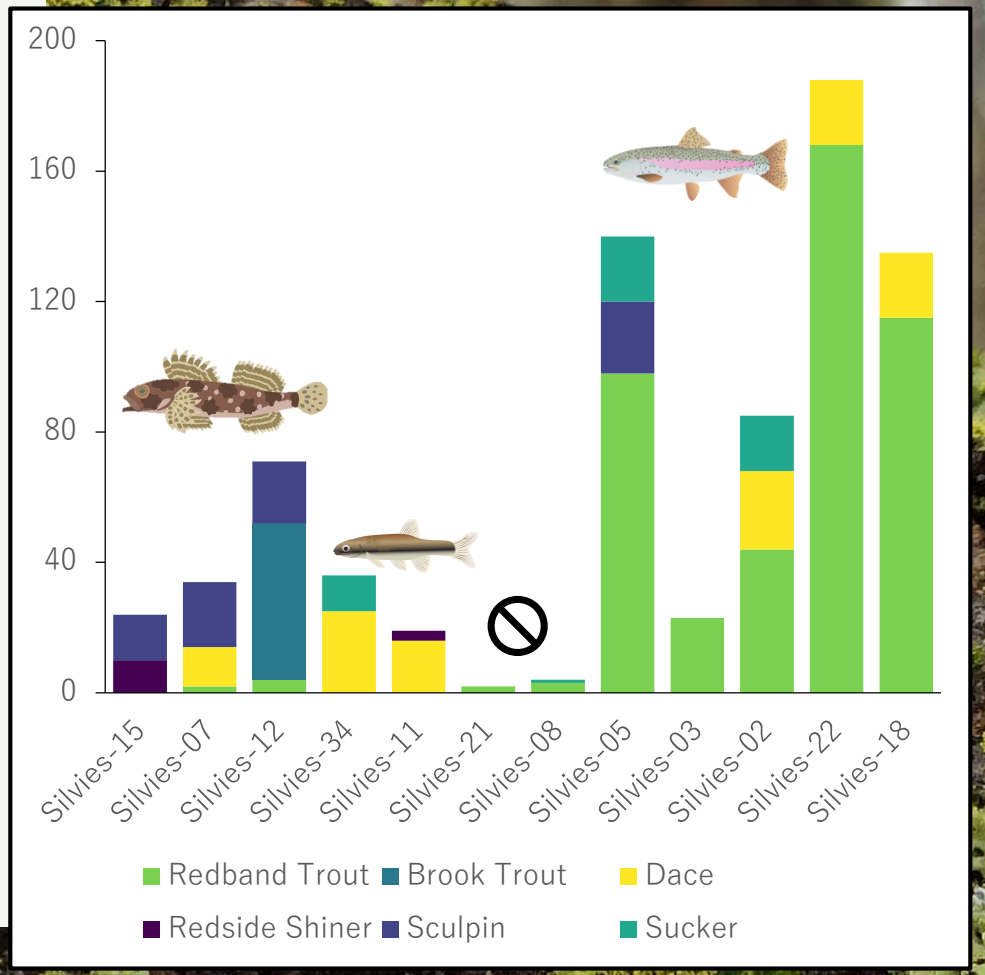
Shrub-Scrub and Low-Forest:
DACE IndVal = 0.866



Species Associations: Indicator Species

Indicator Species Analysis

1. Define **site groups** based on relevant criteria
2. Calculate **indicator values**
3. Assess significance
4. Interpret results

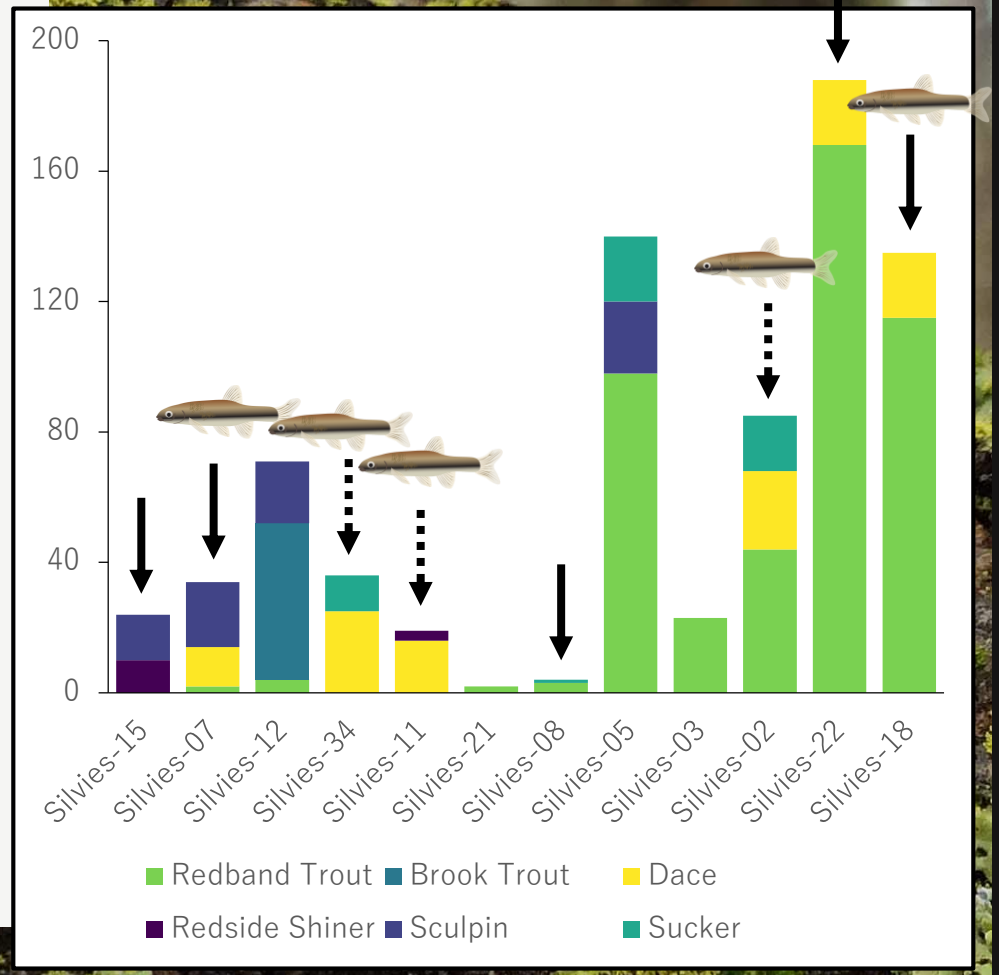


Shrub-Scrub and Low-Forest:
DACE IndVal = 0.866
$P = 0.048$

Species Associations: Indicator Species

Indicator Species Analysis

1. Define **site groups** based on relevant criteria
2. Calculate **indicator values**
3. Assess significance
4. Interpret results



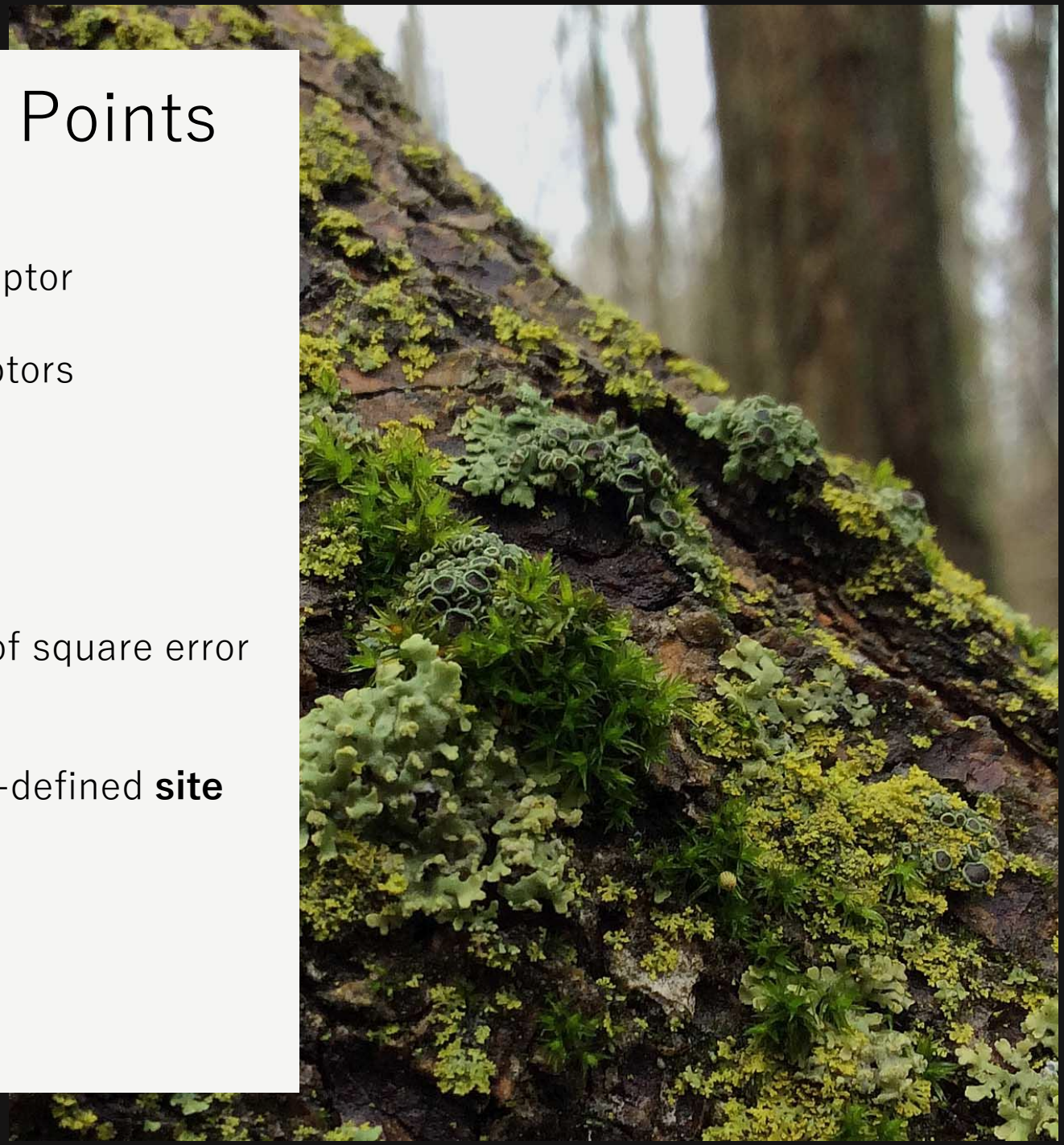
Shrub-Scrub and Low-Forest:
DACE IndVal = 0.866
$P = 0.048$

Conclusion: Summary of Key Points

- Divisive Hierarchical Cluster Analysis
  - **Monothetic** methods rely on a single descriptor
    - e.g., association analysis
  - **Polythetic** methods rely on multiple descriptors
    - e.g., divisive analysis a.k.a. DIANA

- Non-hierarchical Cluster Analysis
  - **Non-hierarchical Complete Linkage**
  - **K-means Partitioning/Clustering**
    - Goal is to minimize within-cluster sum of square error

- Indicator Species Analysis
  - Calculates **indicator values** (IndVal) for pre-defined **site groups**



Questions?