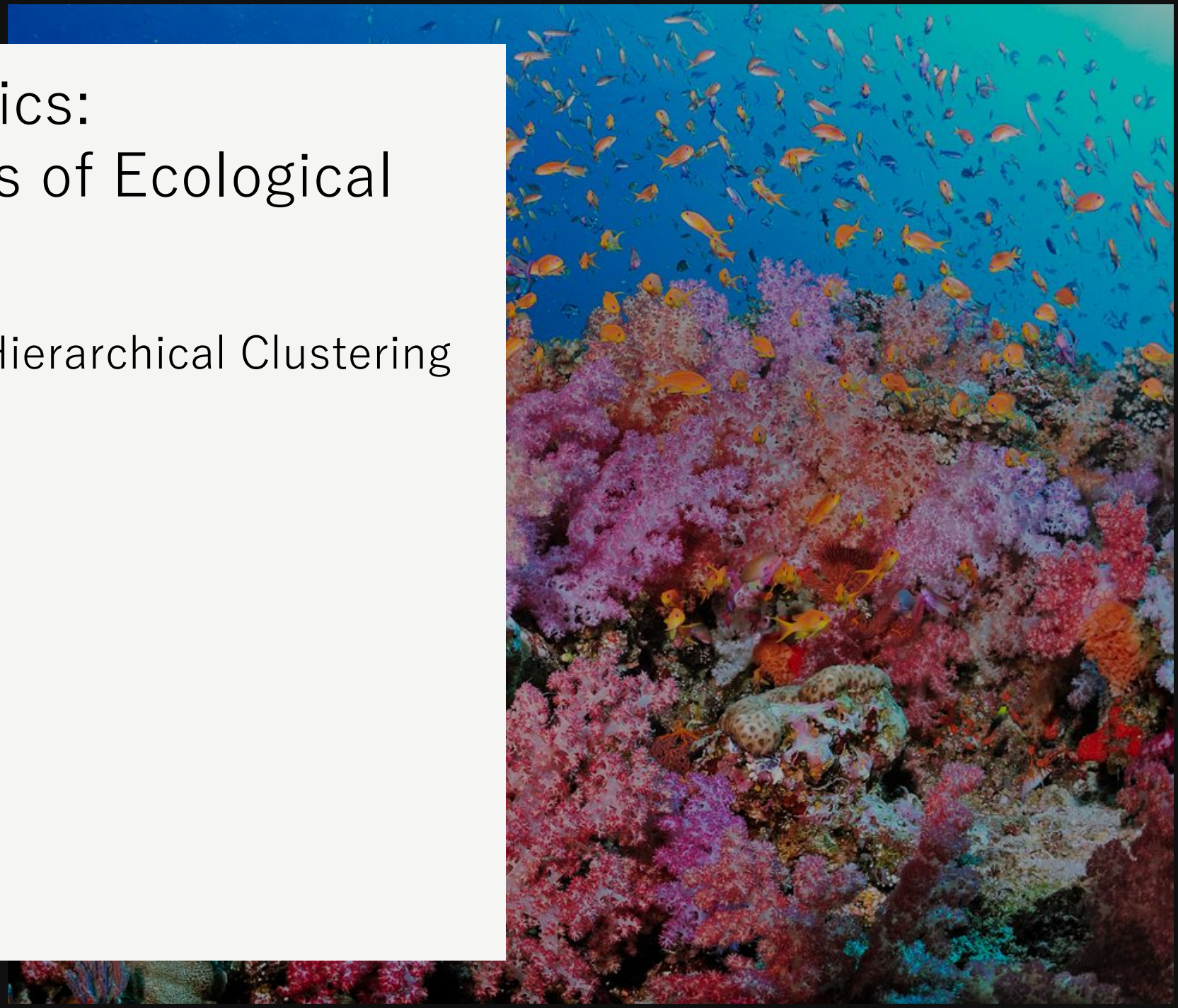
FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

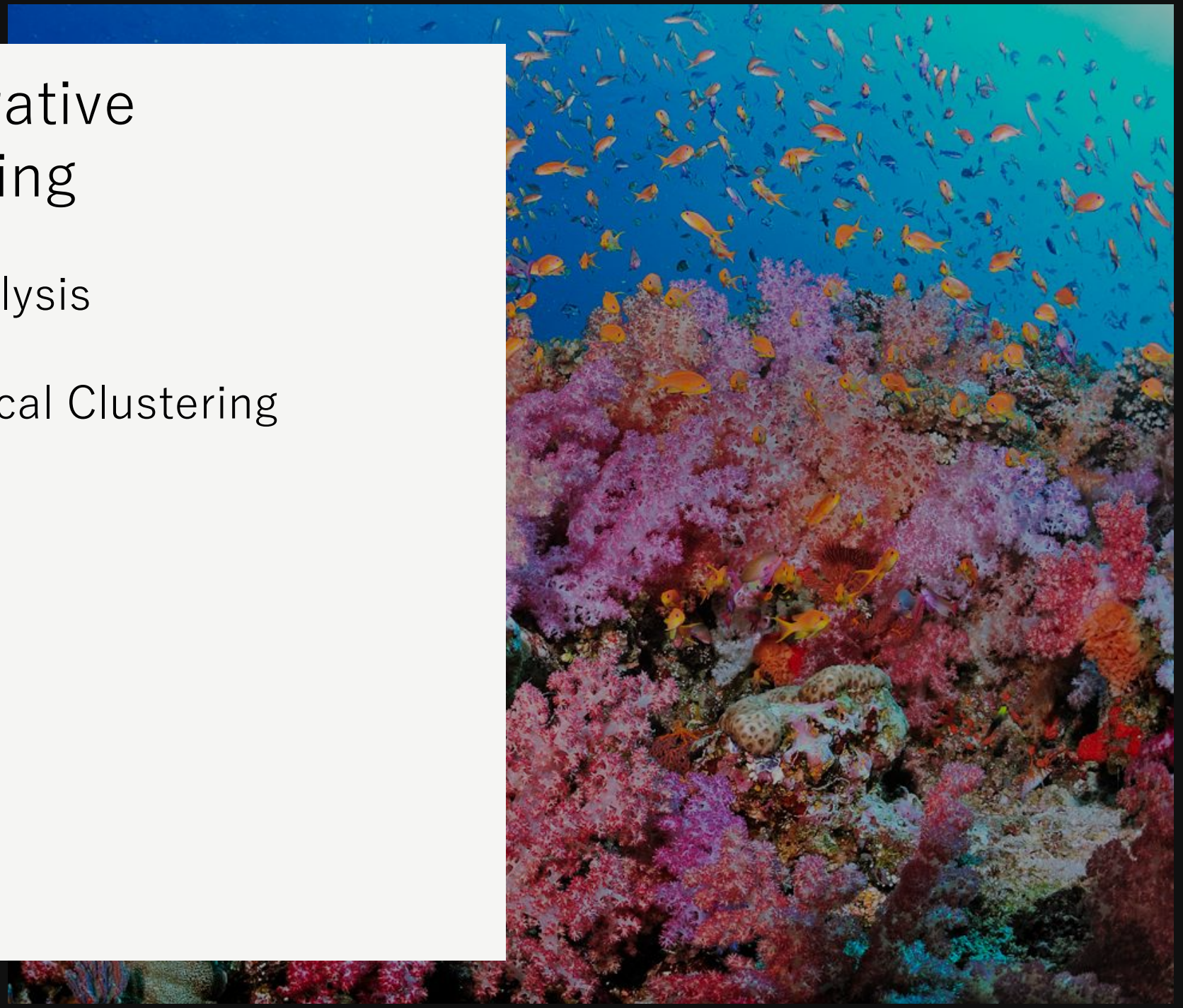
Lecture 4: Agglomerative Hierarchical Clustering

Thursday, October 10, 2024

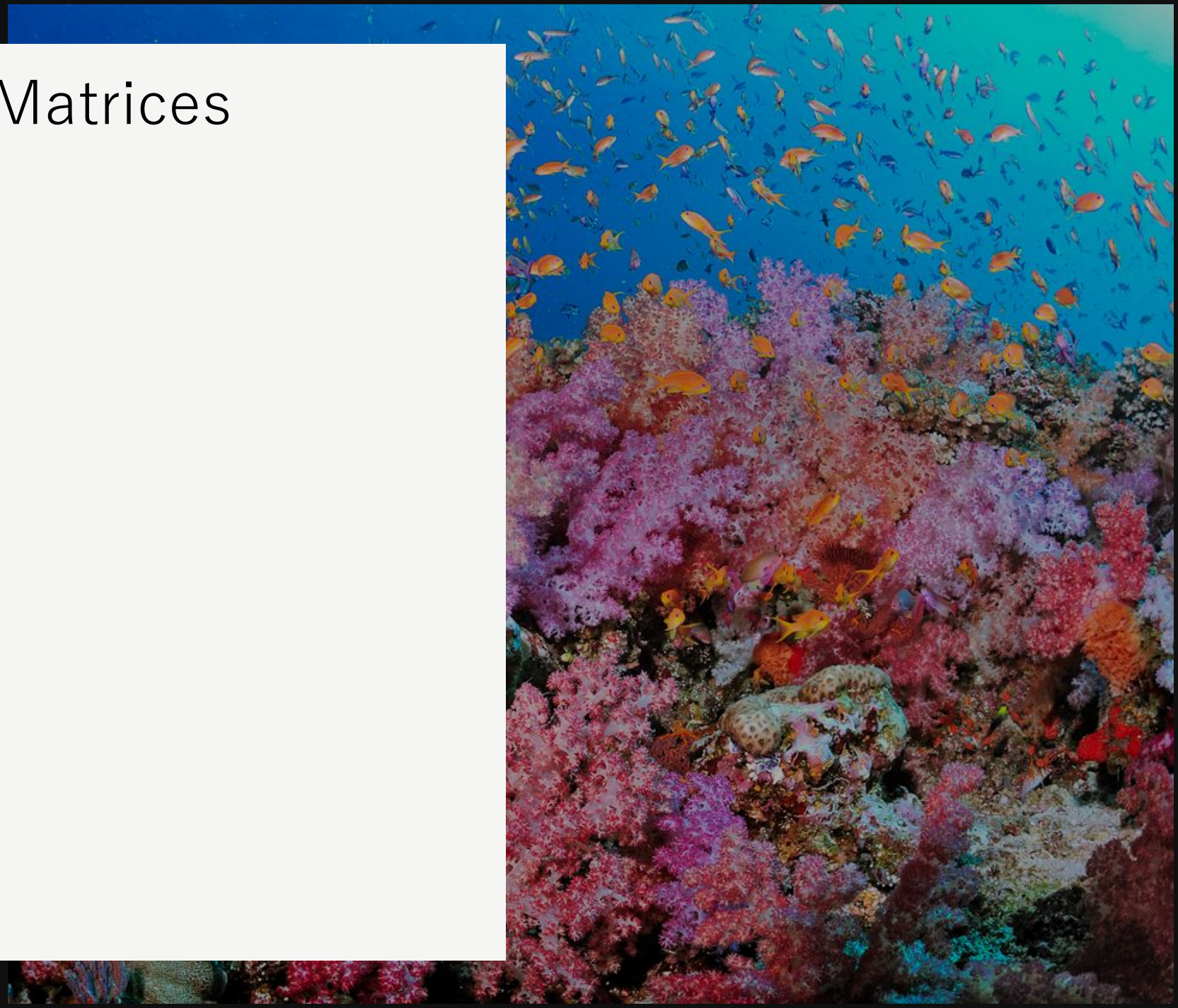


Lecture 4: Agglomerative Hierarchical Clustering

- Hierarchical Cluster Analysis
- Agglomerative Hierarchical Clustering Methods
- Clustering Statistics
- Cluster Validation



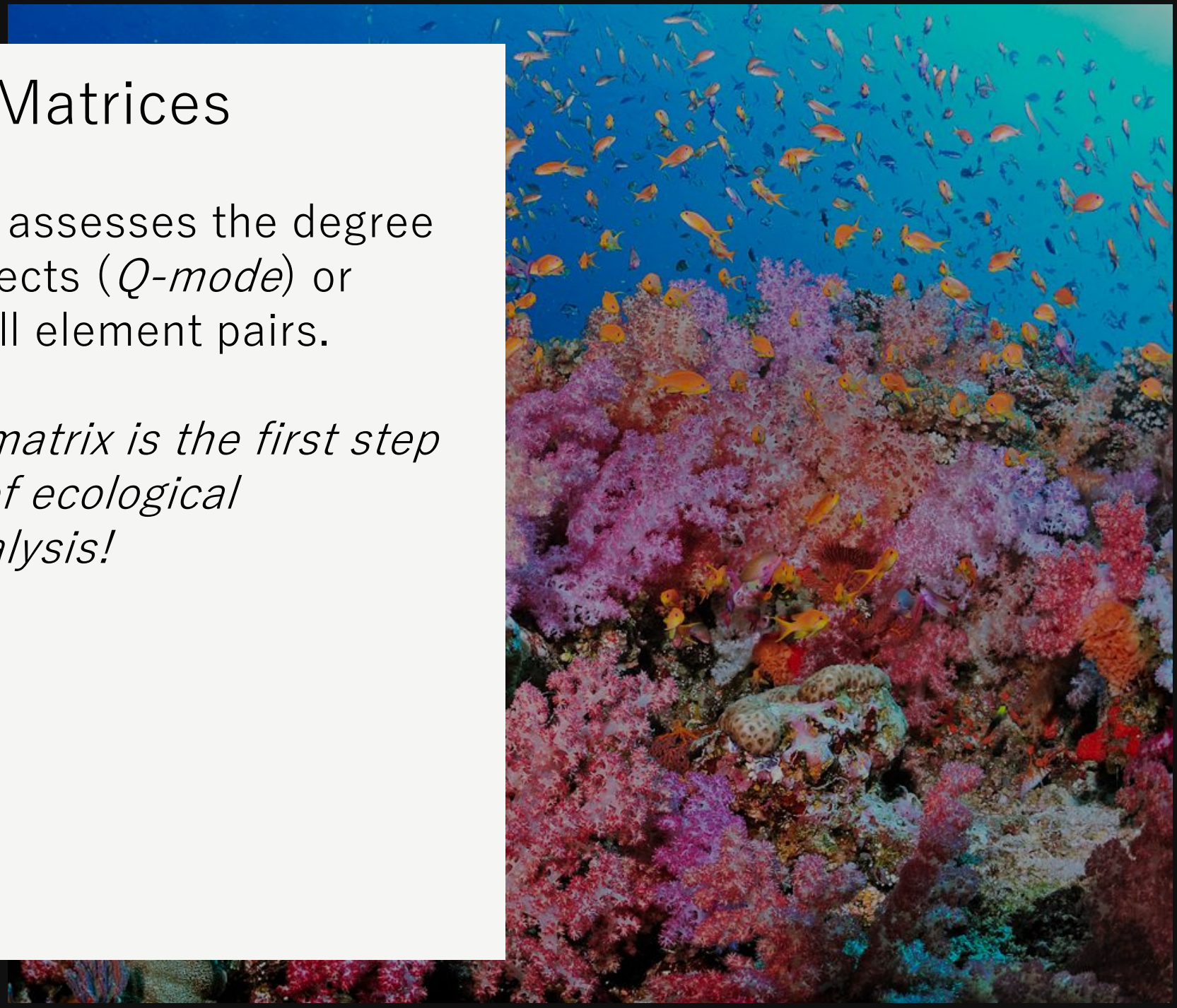
Recap: Association Matrices



Recap: Association Matrices

An **association matrix (A)** assesses the degree of resemblance among objects (*Q-mode*) or descriptors (*R-mode*) for all element pairs.

Producing an association matrix is the first step in the numerical analysis of ecological data…including cluster analysis!



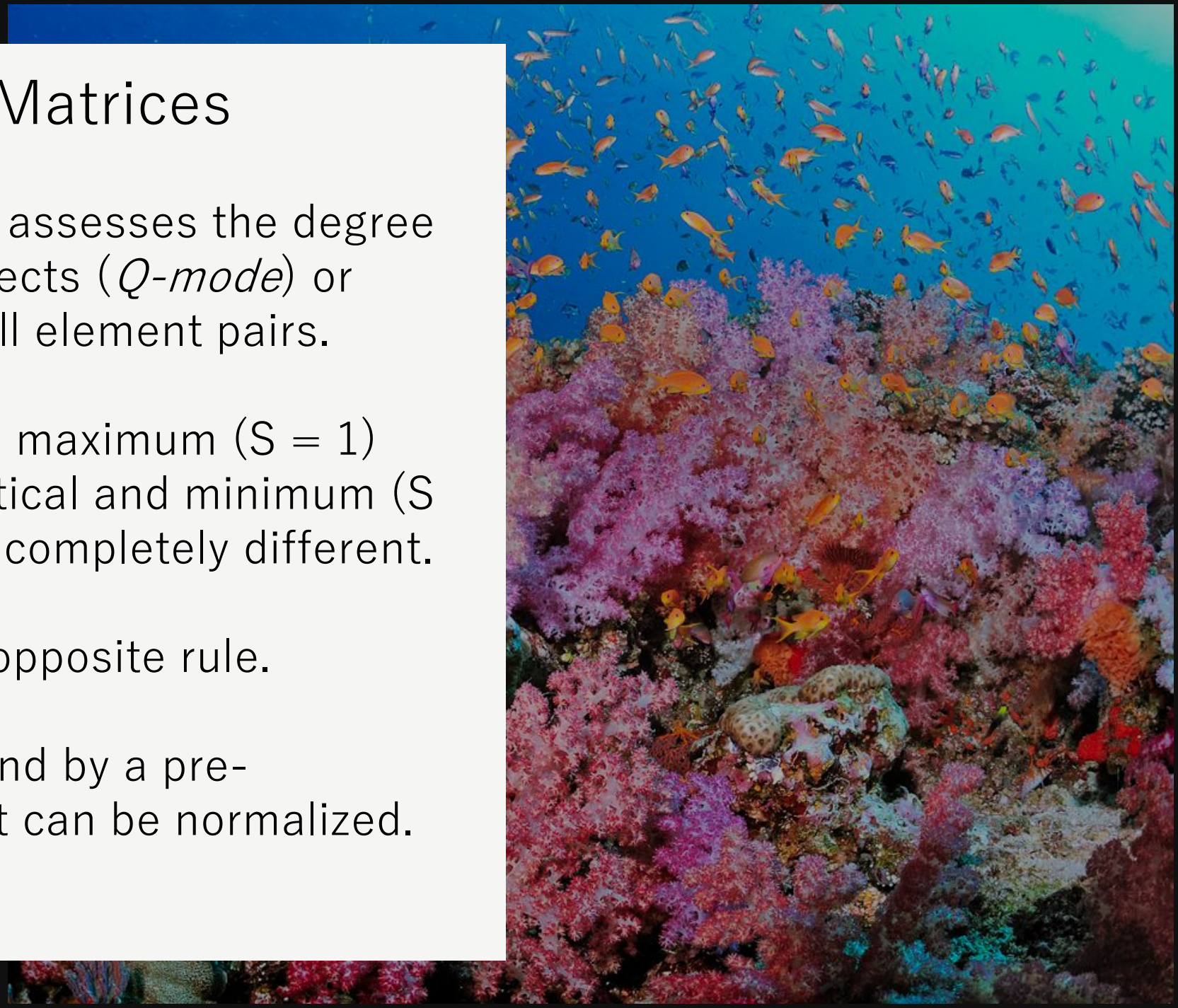
Recap: Association Matrices

An **association matrix (A)** assesses the degree of resemblance among objects (*Q-mode*) or descriptors (*R-mode*) for all element pairs.

Similarity coefficients are maximum (S = 1) when two objects are identical and minimum (S =0) when two objects are completely different.

Dissimilarities follow the opposite rule.

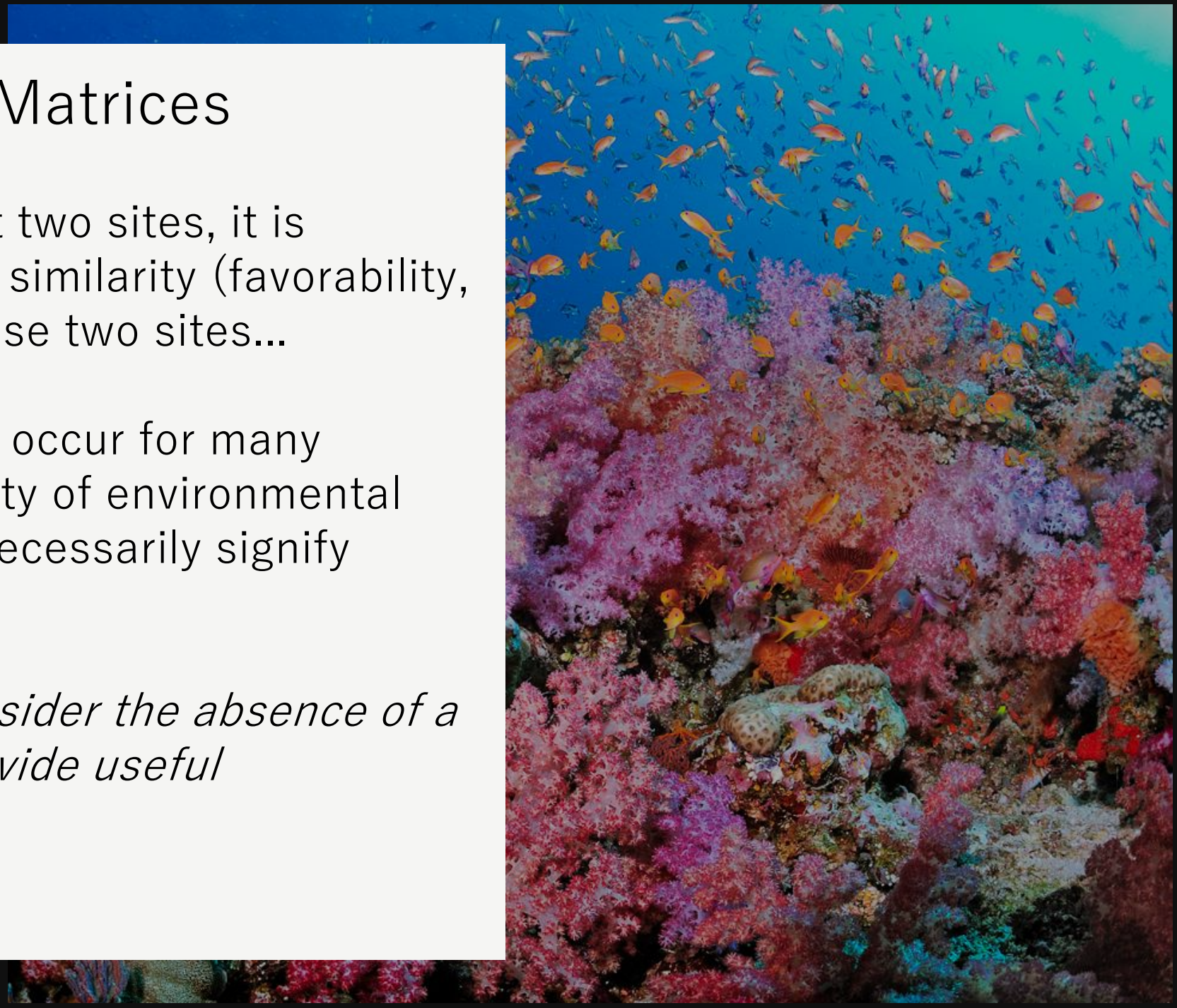
Distances may not be bound by a pre-determined upper limit, but can be normalized.



Recap: Association Matrices

- If a species is present at two sites, it is generally an indicator of similarity (favorability, tolerability) between these two sites…
- ***However***, absences can occur for many reasons, indicate a variety of environmental conditions, and do not necessarily signify environmental similarity

Most scientists do not consider the absence of a species at two sites to provide useful information.



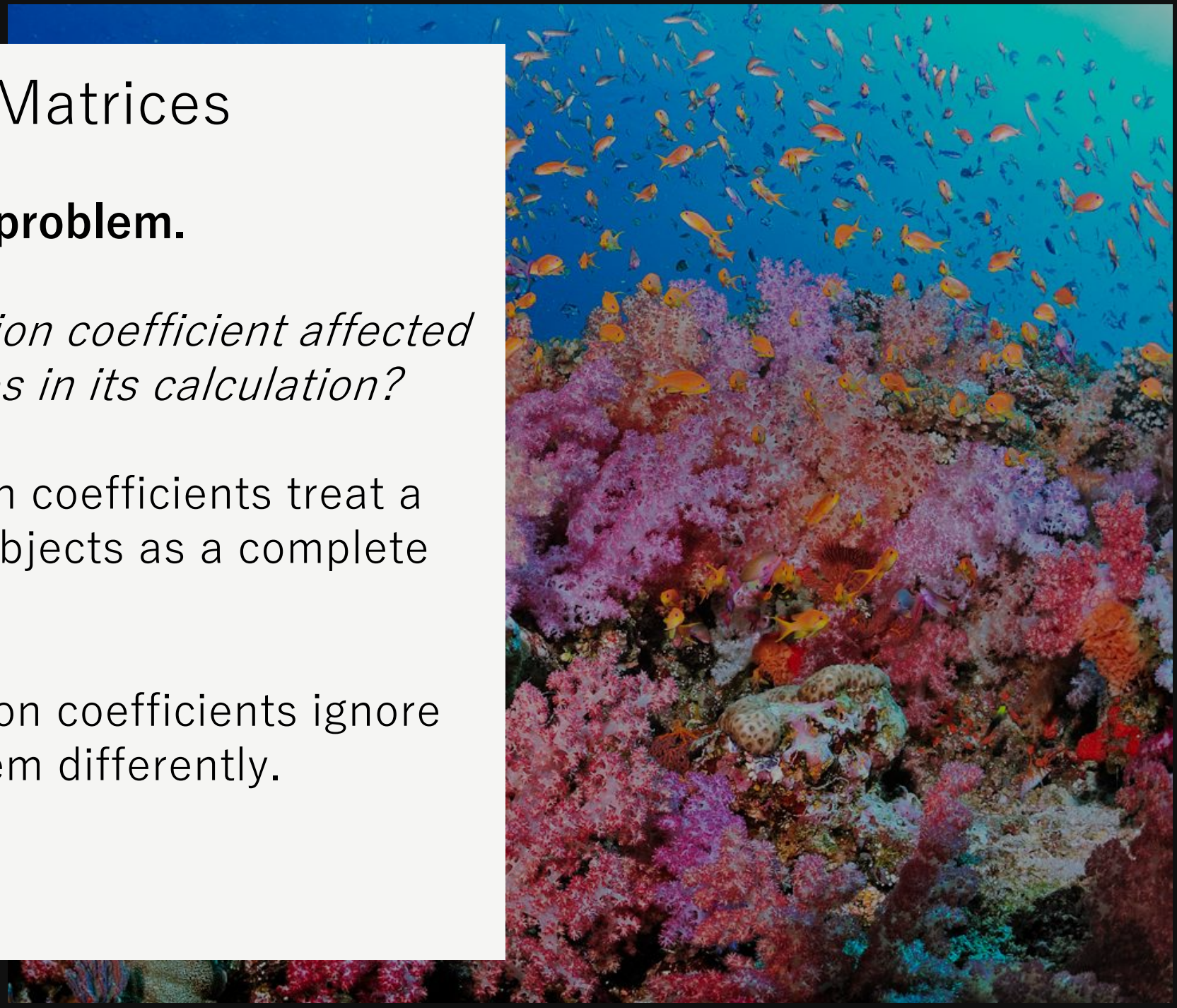
Recap: Association Matrices

- This is the **double zero problem.**

Is the value of an association coefficient affected by inclusion of double zeros in its calculation?

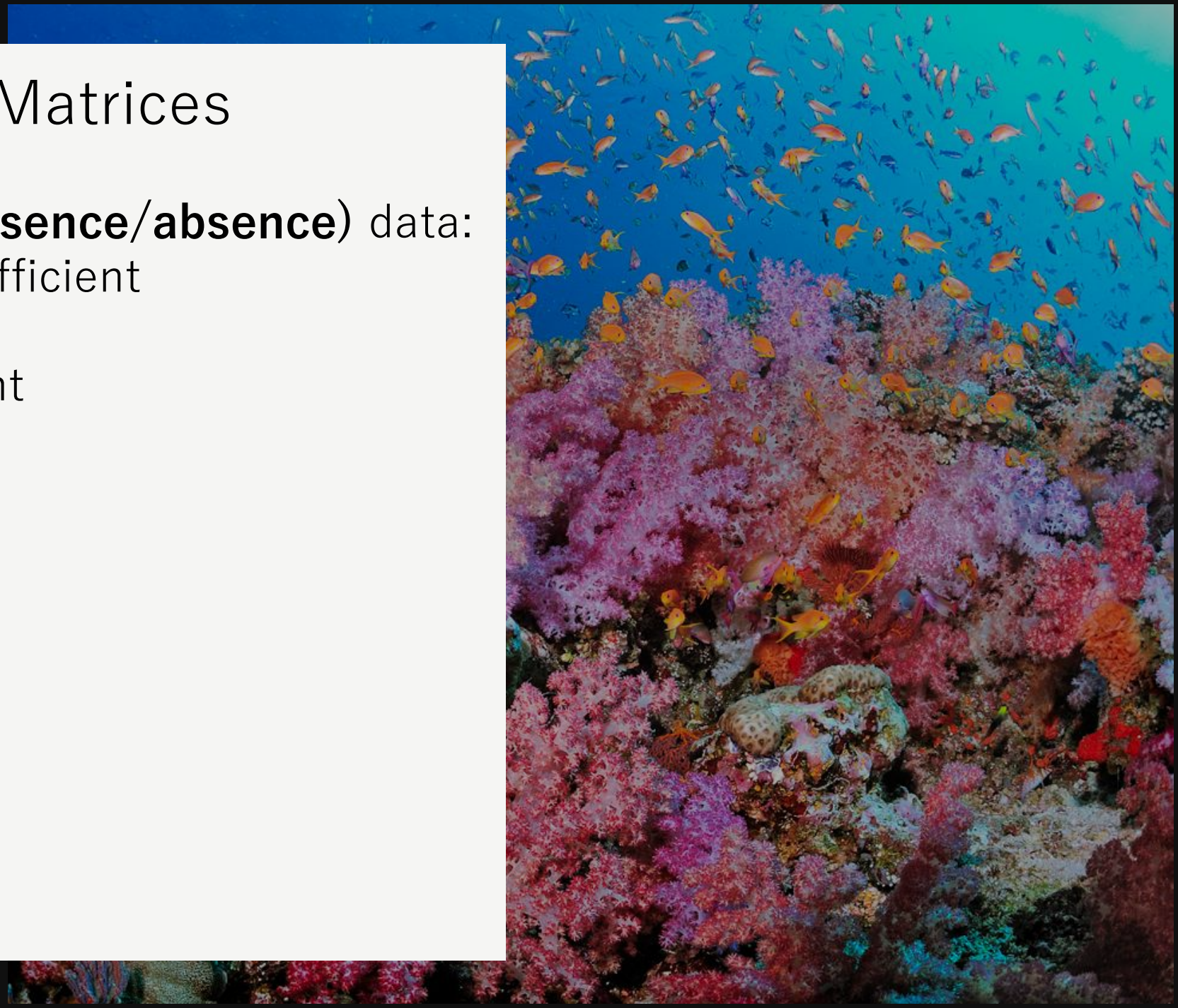
- **Symmetrical** association coefficients treat a zero value for a pair of objects as a complete similarity.

- **Asymmetrical** association coefficients ignore double zeros or treat them differently.



Recap: Association Matrices

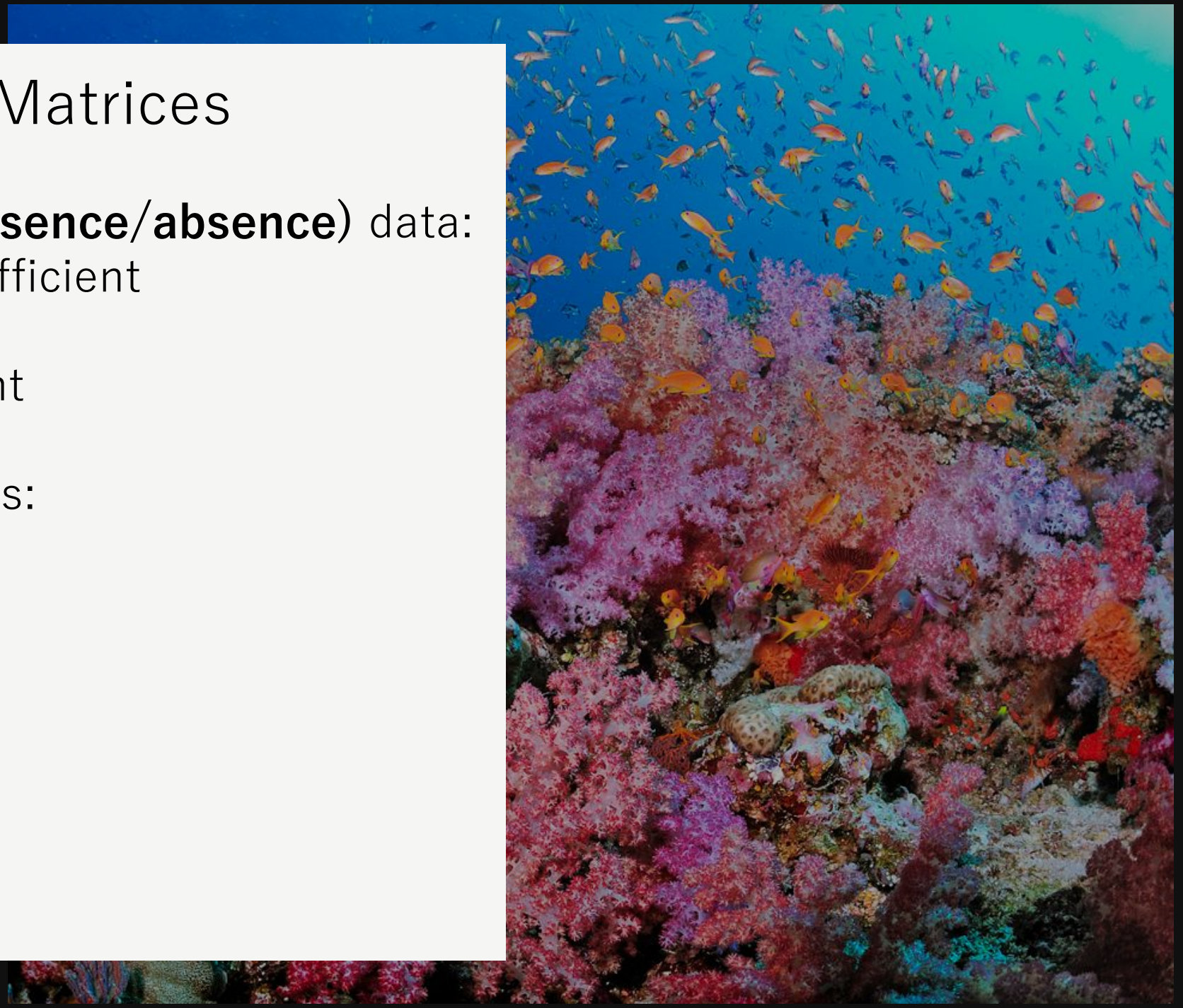
- Matrices for **binary** (**presence**/**absence**) data:
  - Simple matching coefficient
  - Jaccard's coefficient
  - Sørensen's coefficient



Recap: Association Matrices

- Matrices for **binary** (**presence**/**absence**) data:
  - Simple matching coefficient
  - Jaccard's coefficient
  - Sørensen's coefficient

- **Symmetrical** coefficients:
  - Euclidean distance
  - Manhattan distance
  - Gower's coefficient

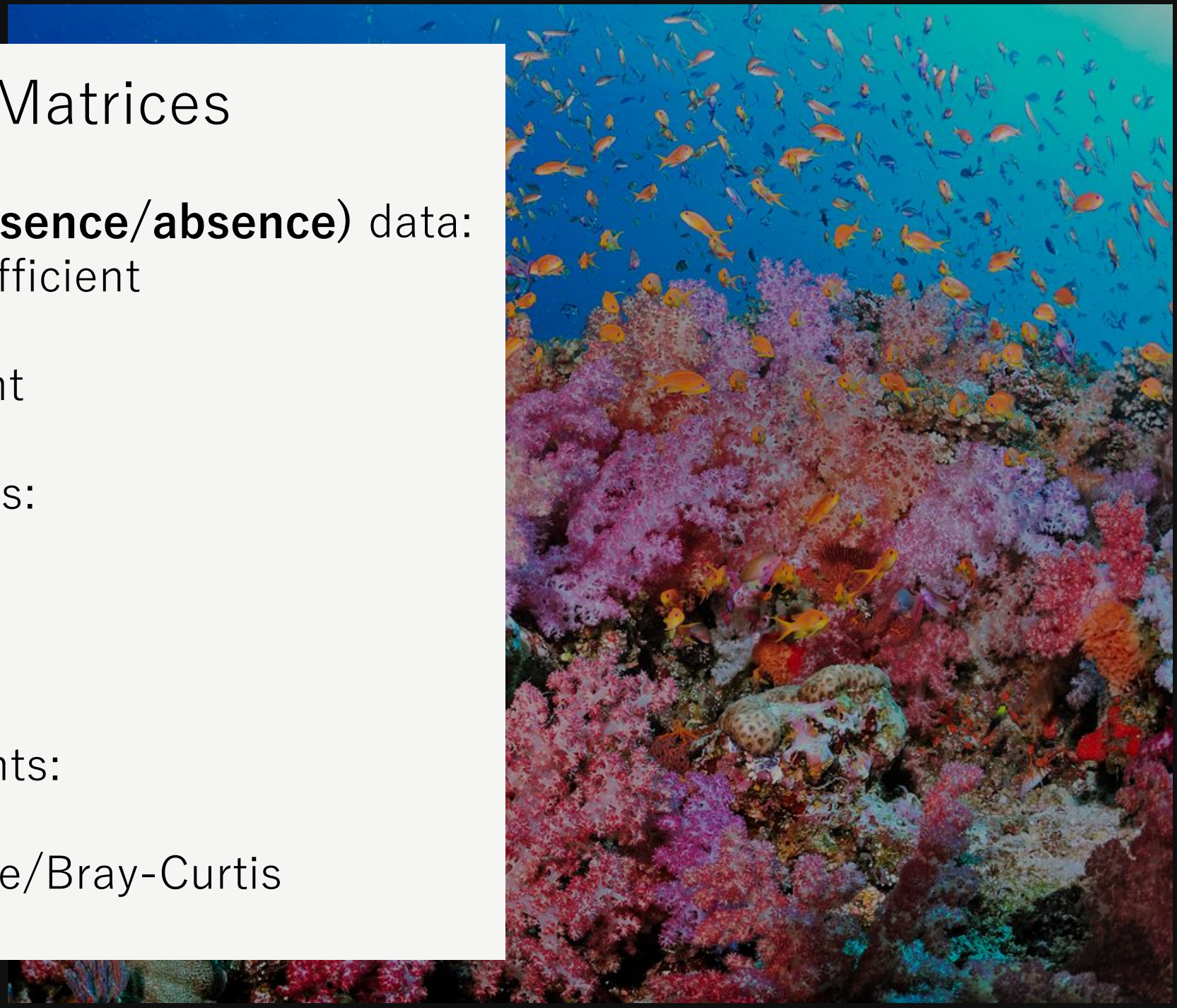


Recap: Association Matrices

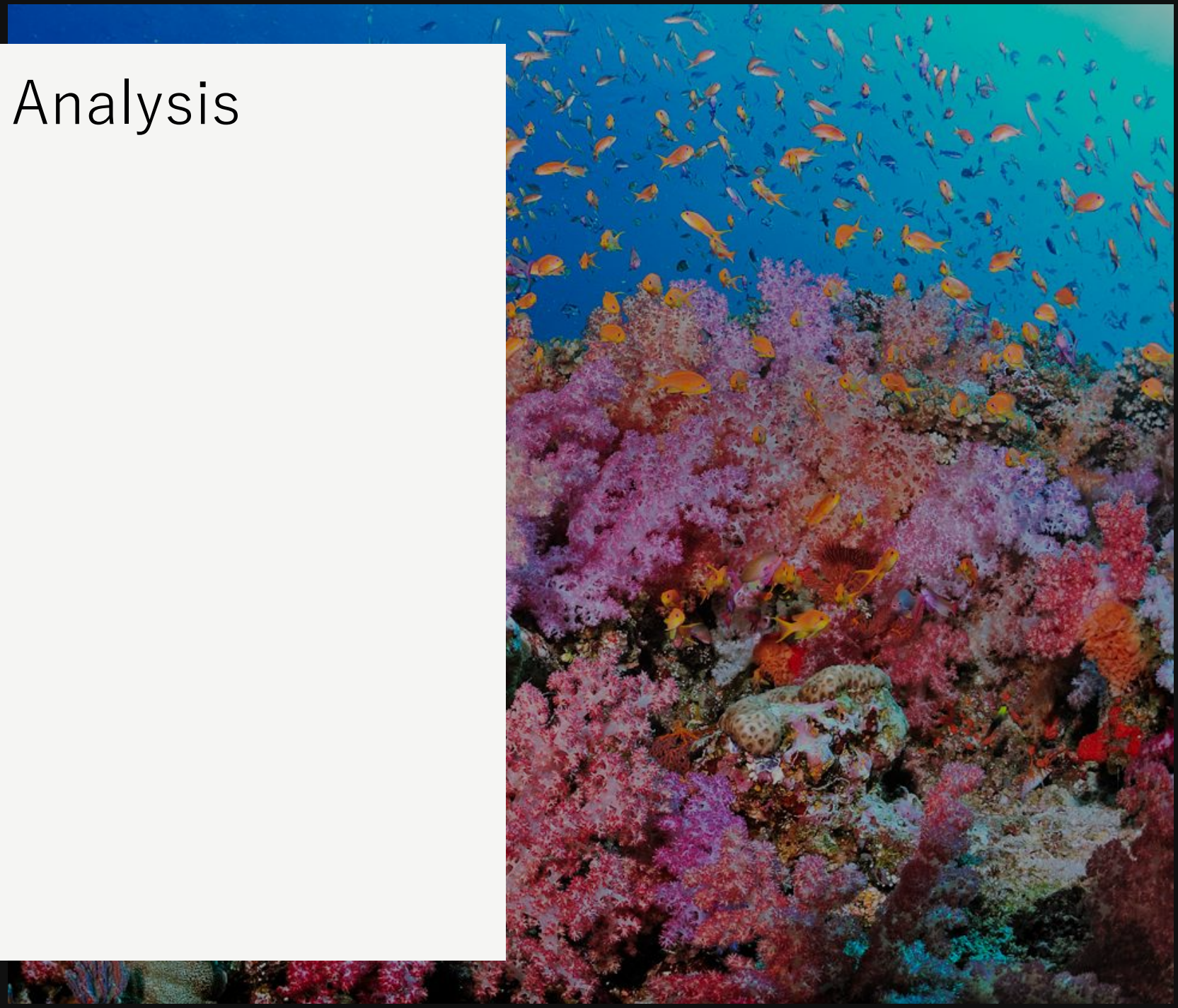
- Matrices for **binary** (**presence**/**absence**) data:
  - Simple matching coefficient
  - Jaccard's coefficient
  - Sørensen's coefficient

- **Symmetrical** coefficients:
  - Euclidean distance
  - Manhattan distance
  - Gower's coefficient

- **Asymmetrical** coefficients:
  - Chi-square distance
  - Percentage difference/Bray-Curtis

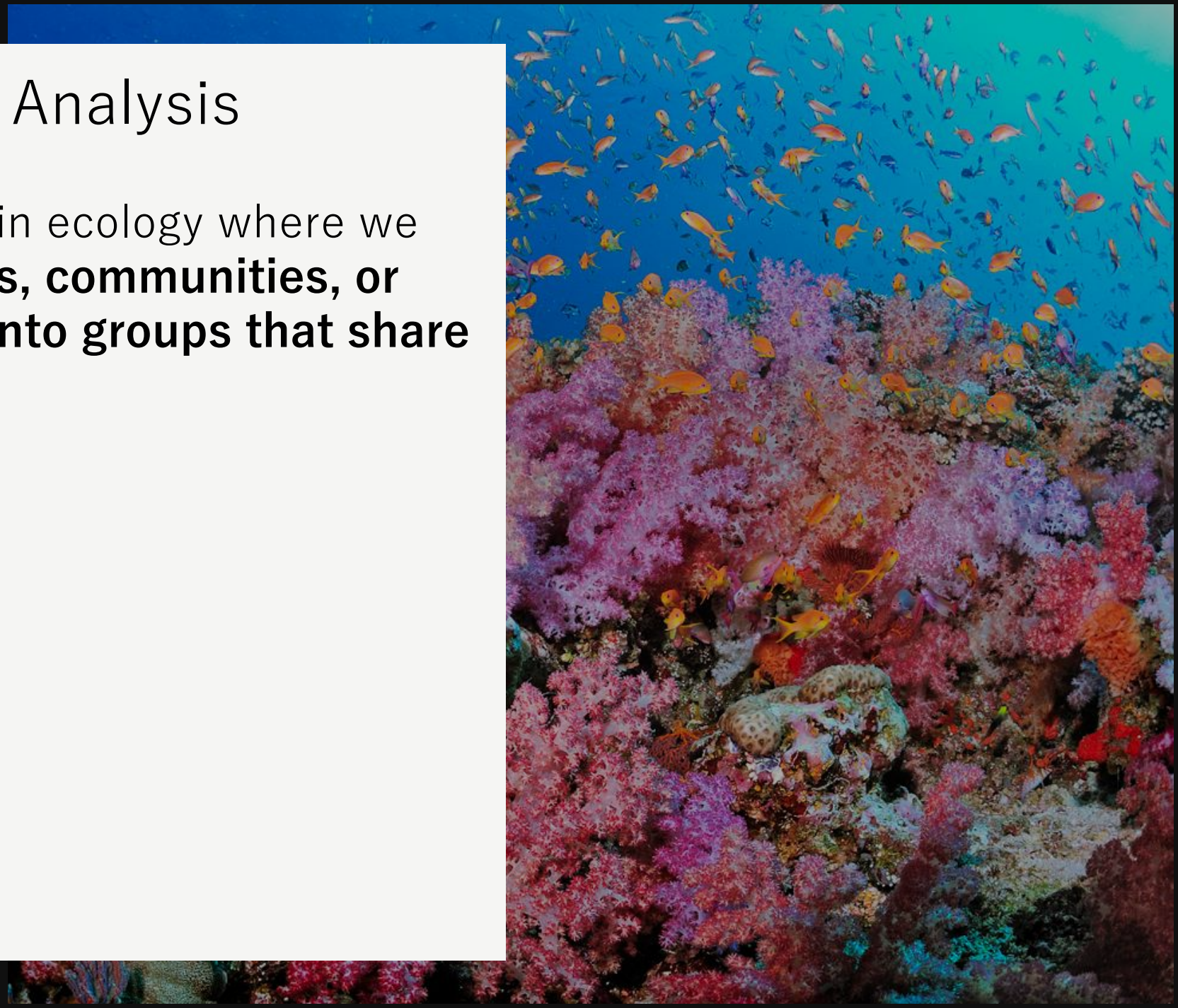


Hierarchical Cluster Analysis



Hierarchical Cluster Analysis

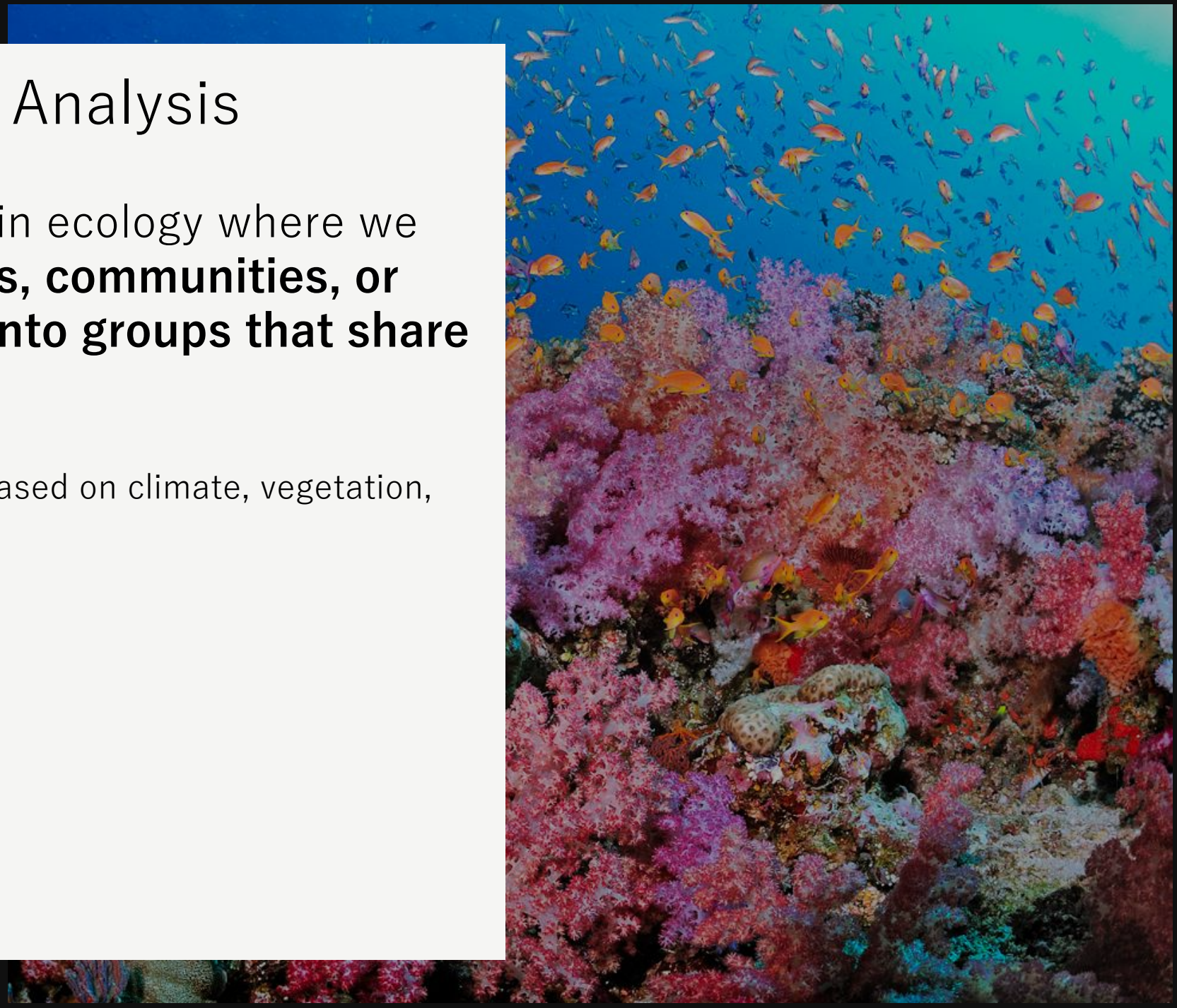
There are many instances in ecology where we need to **classify organisms, communities, or environmental variables into groups that share similar characteristics**:



Hierarchical Cluster Analysis

There are many instances in ecology where we need to **classify organisms, communities, or environmental variables into groups that share similar characteristics**:

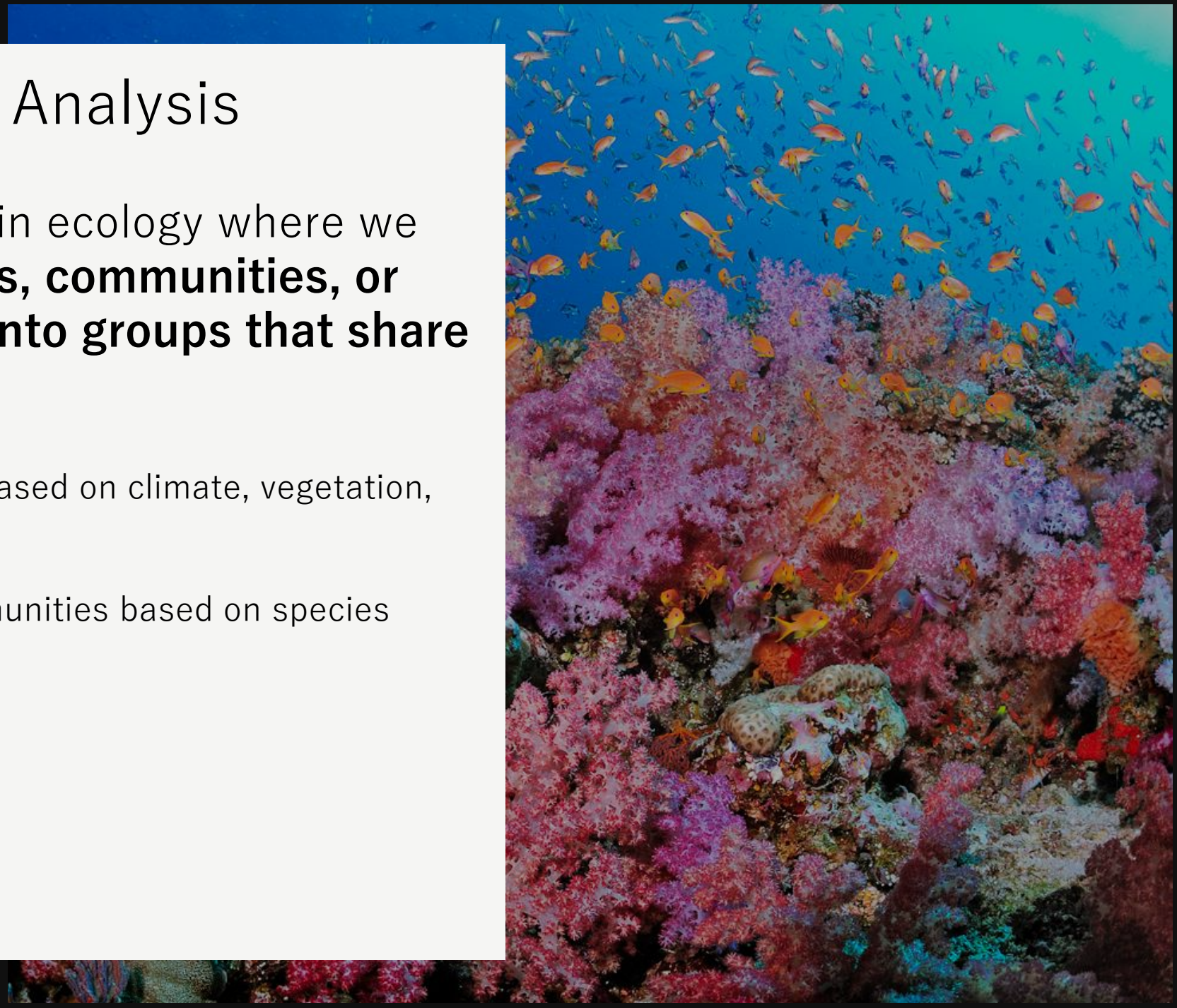
- Classifying regions or habitats based on climate, vegetation, and characteristic species



Hierarchical Cluster Analysis

There are many instances in ecology where we need to **classify organisms, communities, or environmental variables into groups that share similar characteristics**:

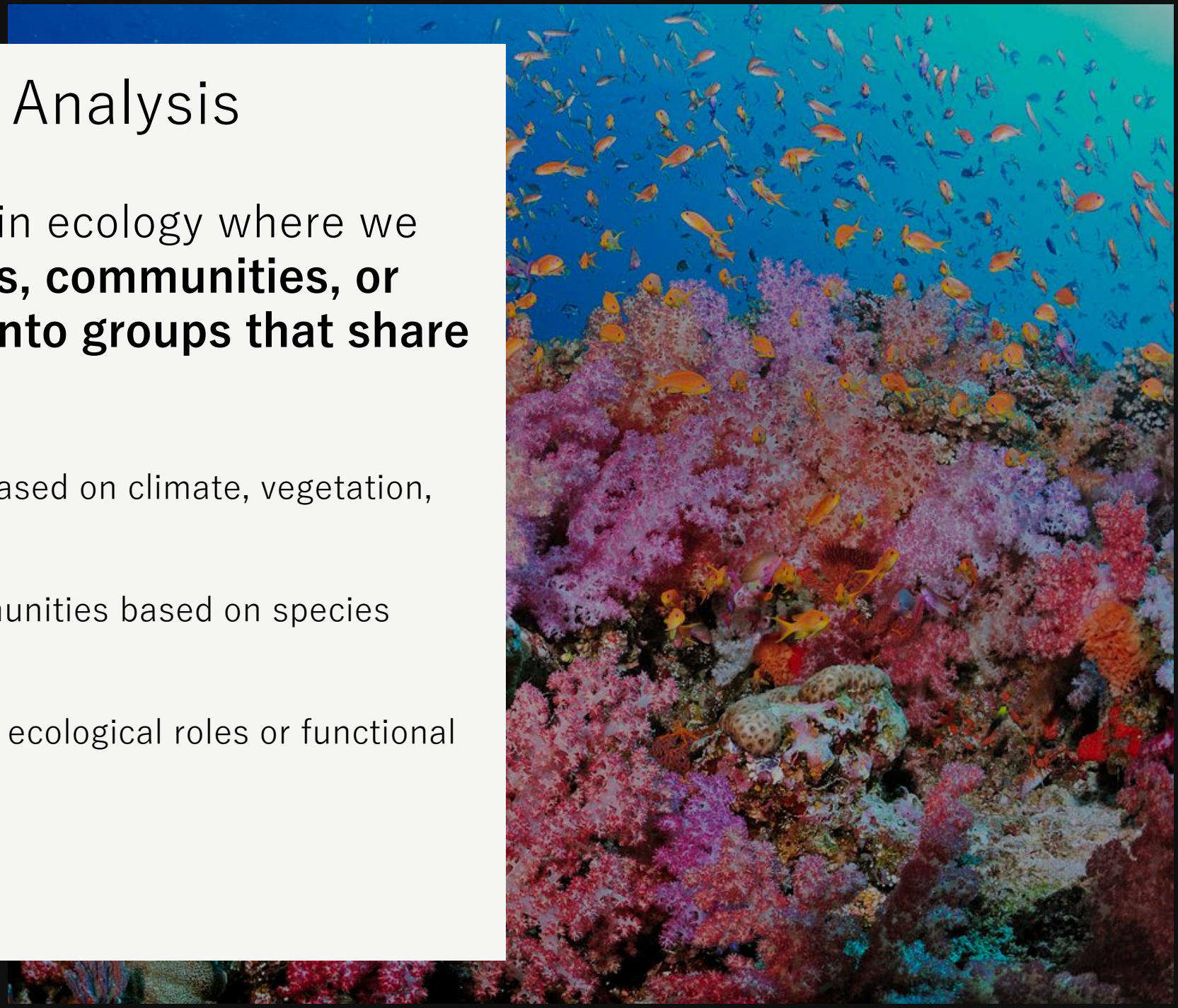
- Classifying regions or habitats based on climate, vegetation, and characteristic species
- Identifying and classifying communities based on species composition and abundance



Hierarchical Cluster Analysis

There are many instances in ecology where we need to **classify organisms, communities, or environmental variables into groups that share similar characteristics**:

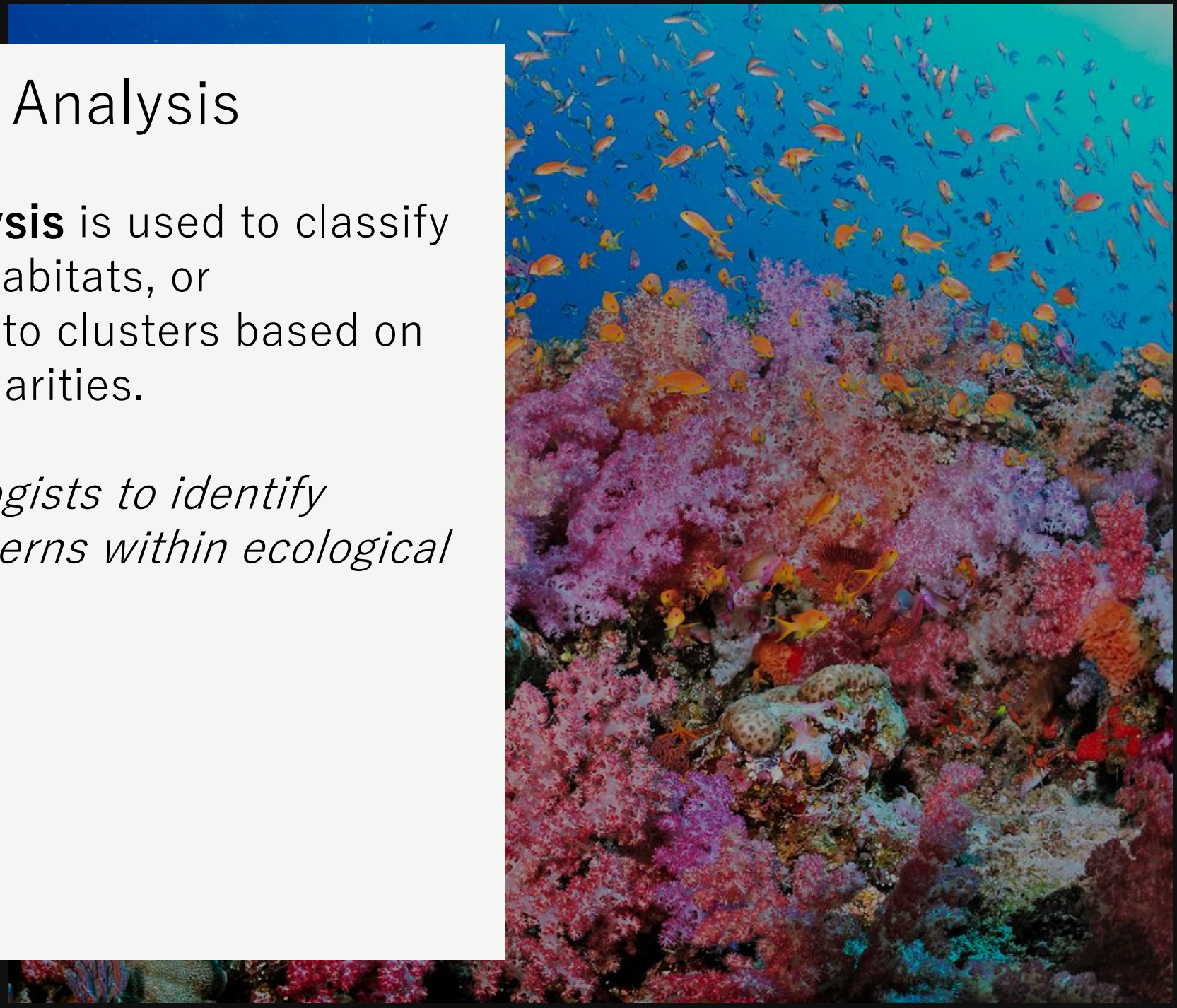
- Classifying regions or habitats based on climate, vegetation, and characteristic species
- Identifying and classifying communities based on species composition and abundance
- Grouping species based on their ecological roles or functional traits



Hierarchical Cluster Analysis

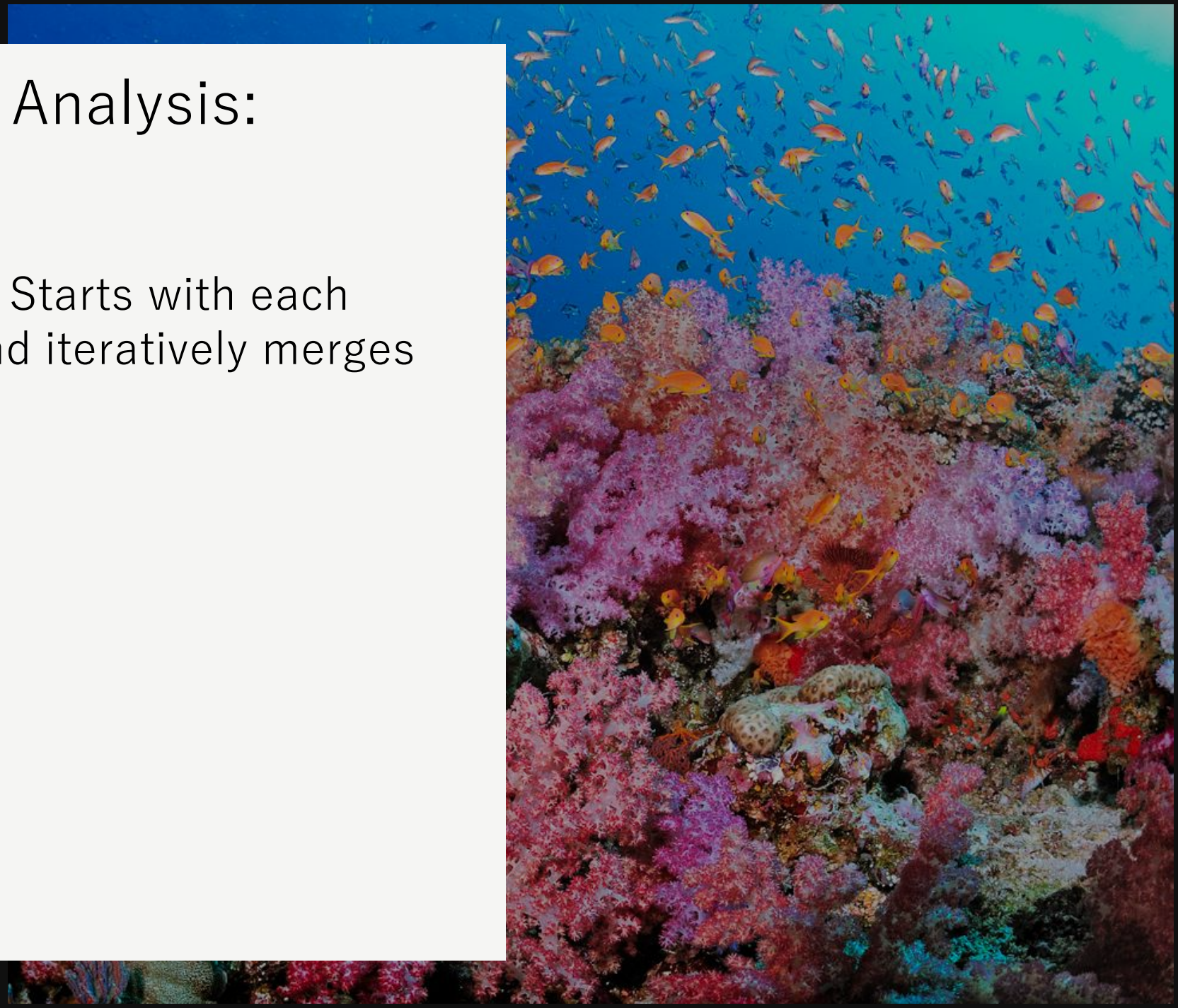
Hierarchical cluster analysis is used to classify objects, such as species, habitats, or environmental variables, into clusters based on their similarities or dissimilarities.

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



Hierarchical Cluster Analysis: Terminology

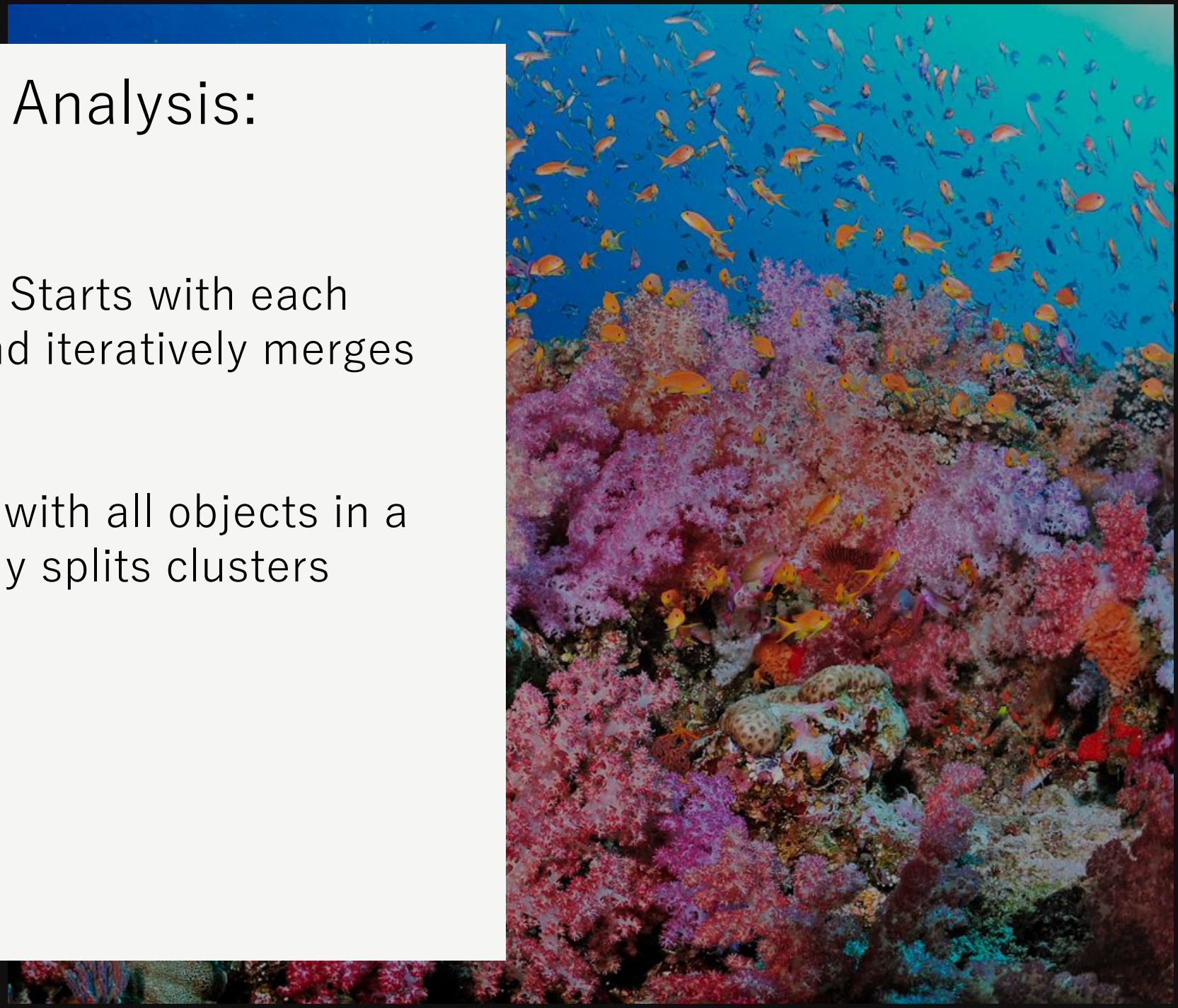
Agglomerative Approach: Starts with each object in its own cluster and iteratively merges clusters



Hierarchical Cluster Analysis: Terminology

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

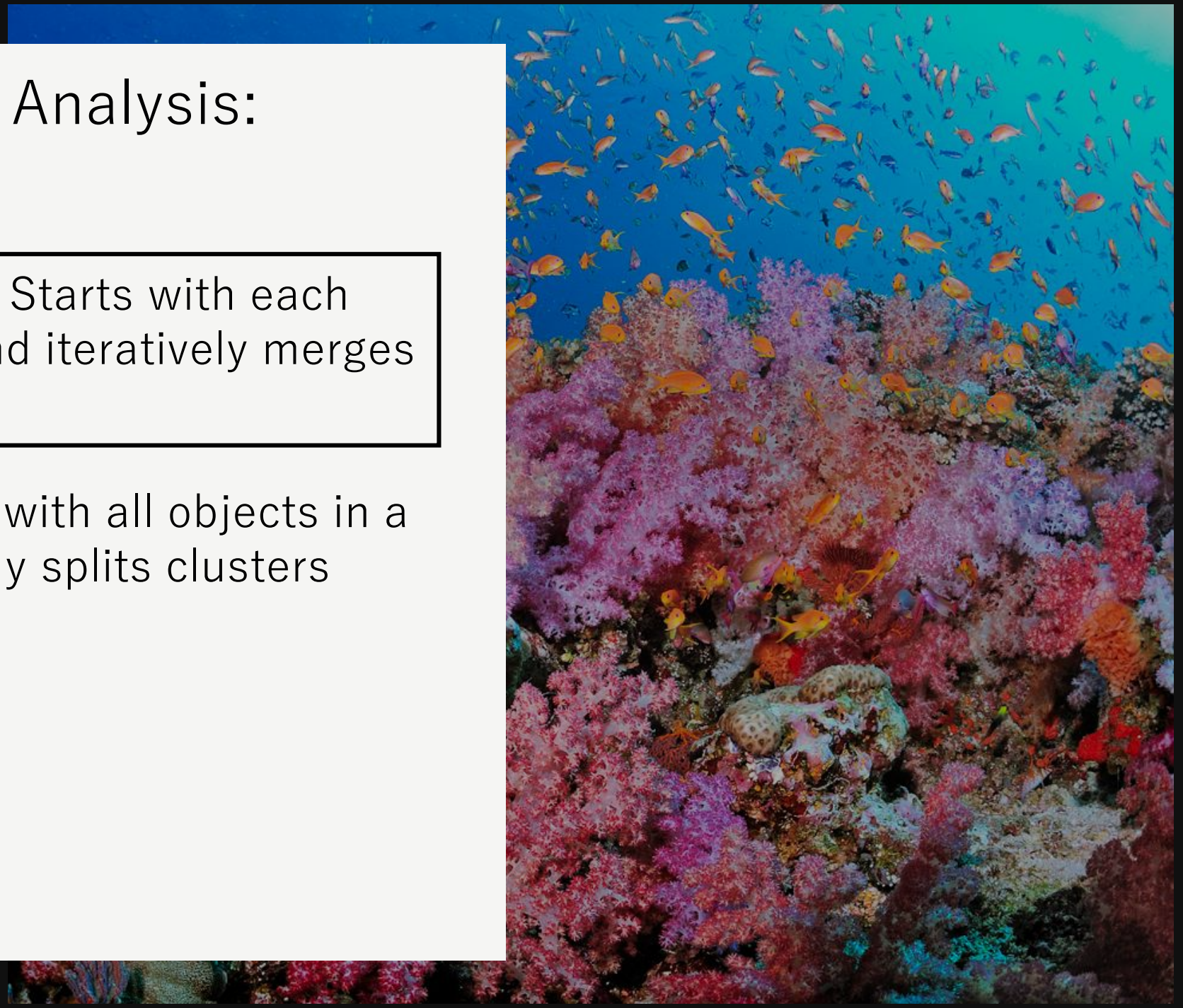
Divisive Approach: Starts with all objects in a single cluster and iteratively splits clusters



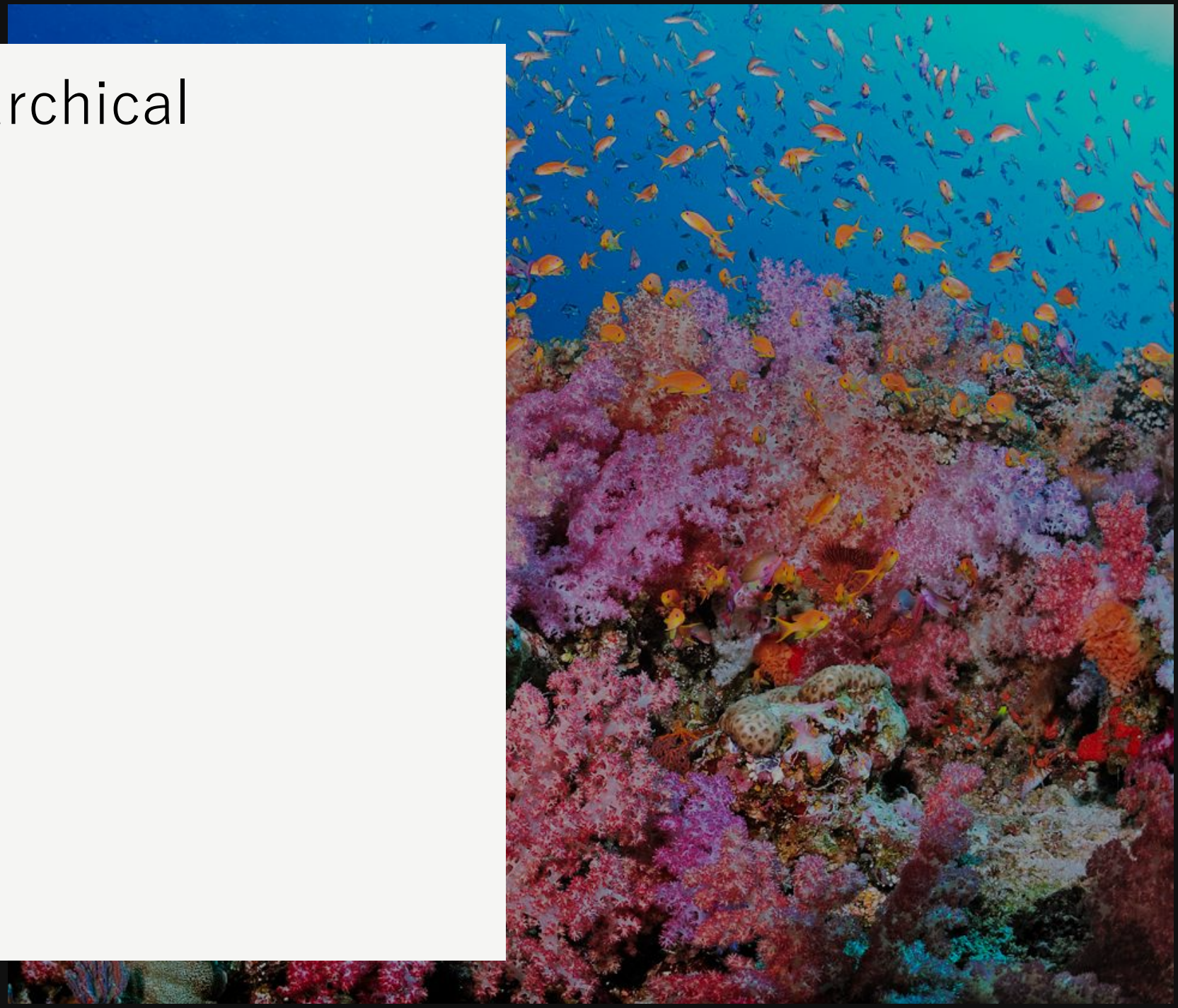
Hierarchical Cluster Analysis: Terminology

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

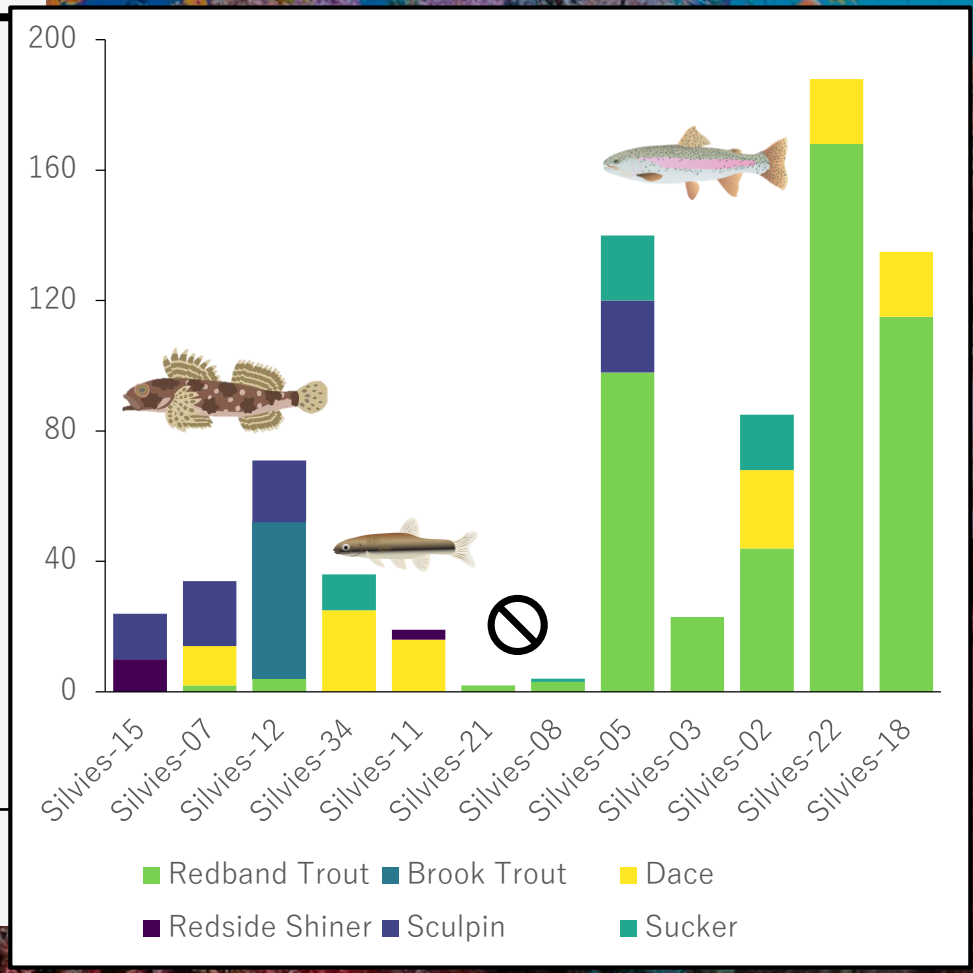


Agglomerative Hierarchical Clustering Methods



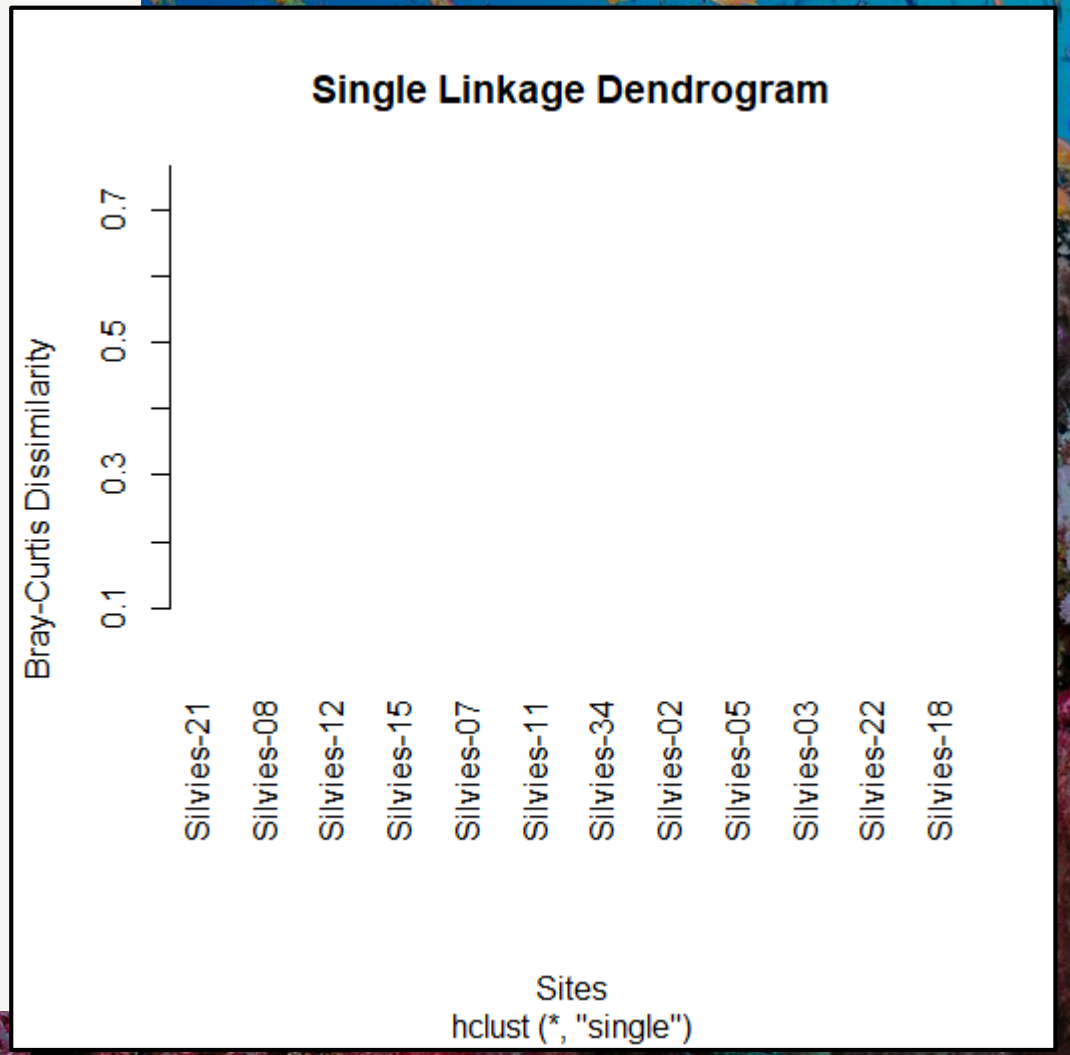
Agglomerative Hierarchical Clustering Methods

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



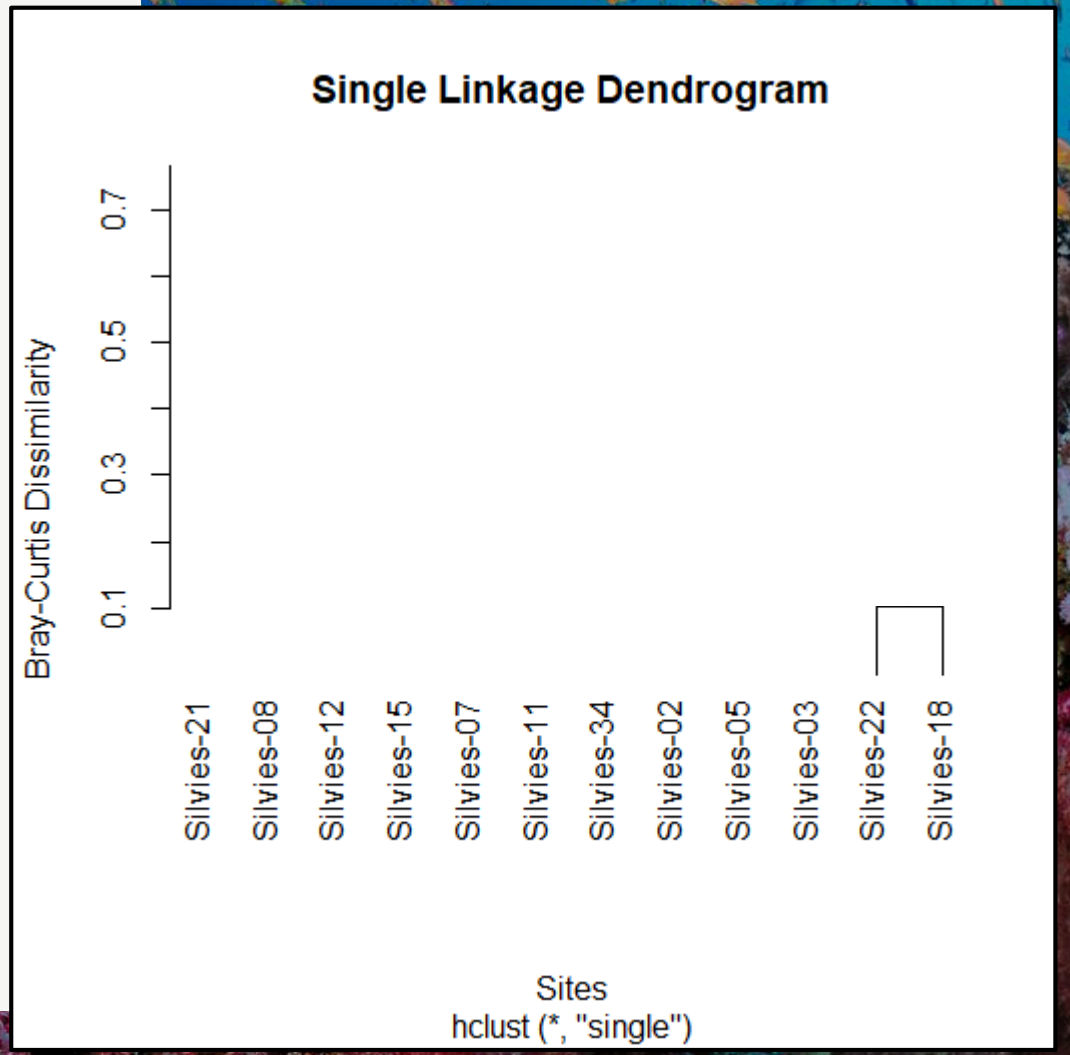
Agglomerative Hierarchical Clustering Methods: Single Linkage

	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



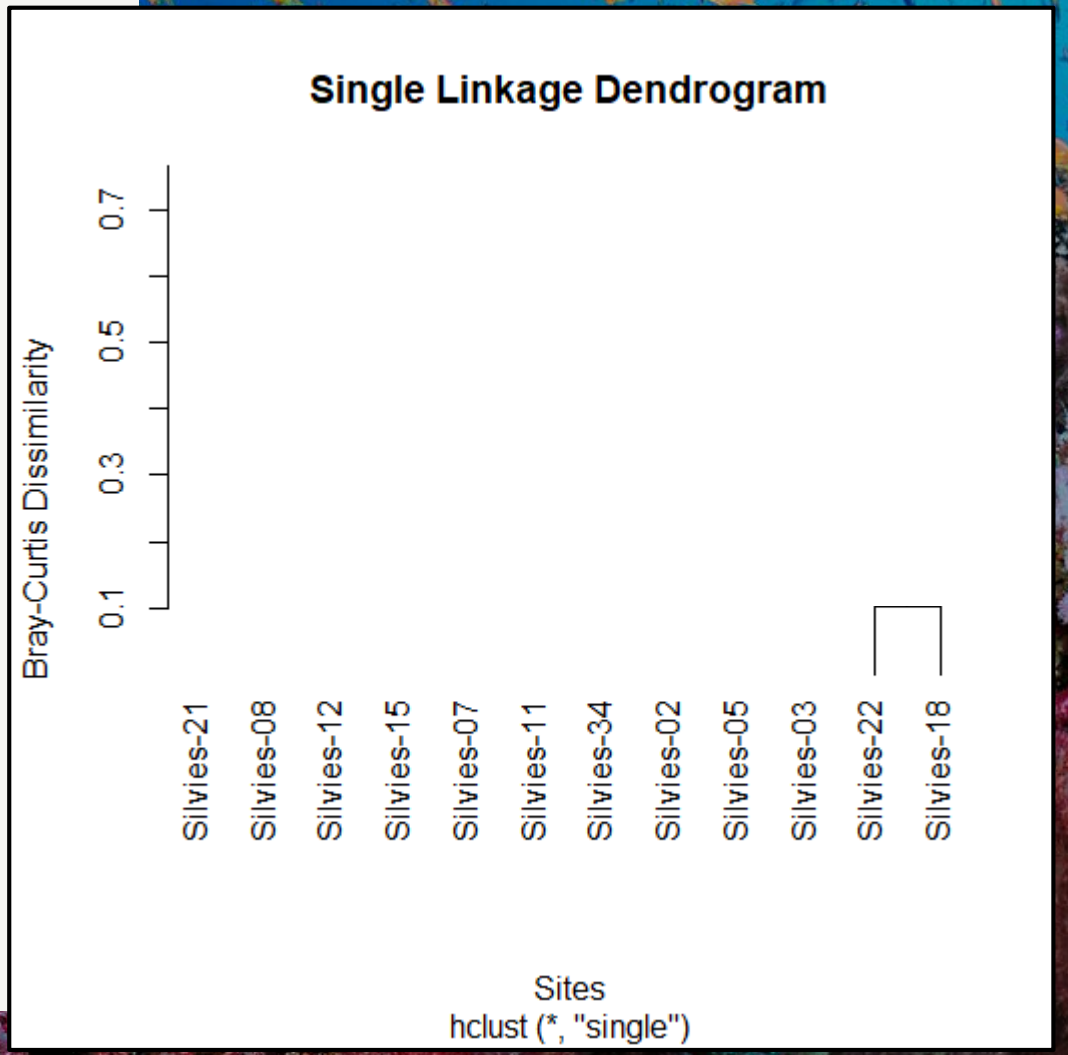
Agglomerative Hierarchical Clustering Methods: Single Linkage

	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



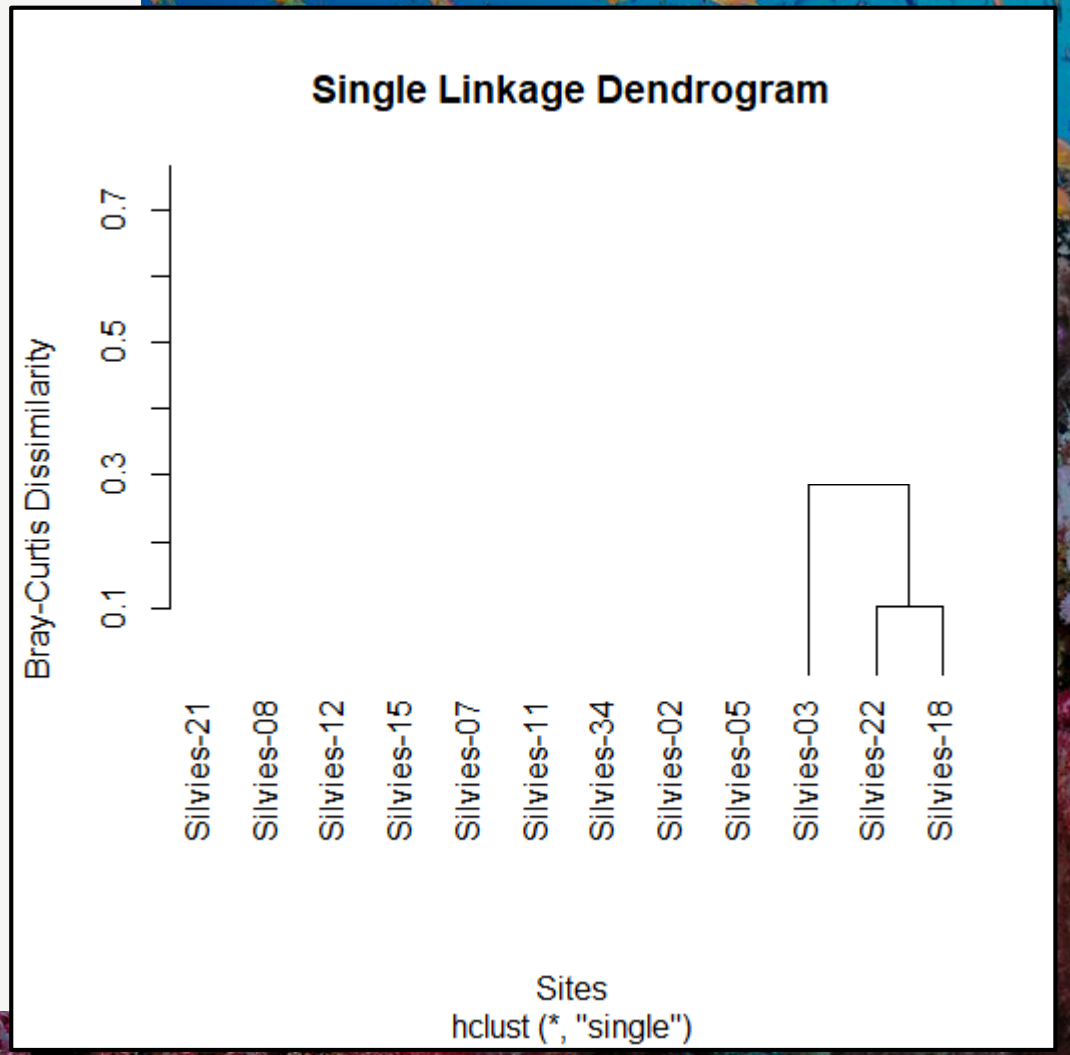
Agglomerative Hierarchical Clustering Methods: Single Linkage

	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



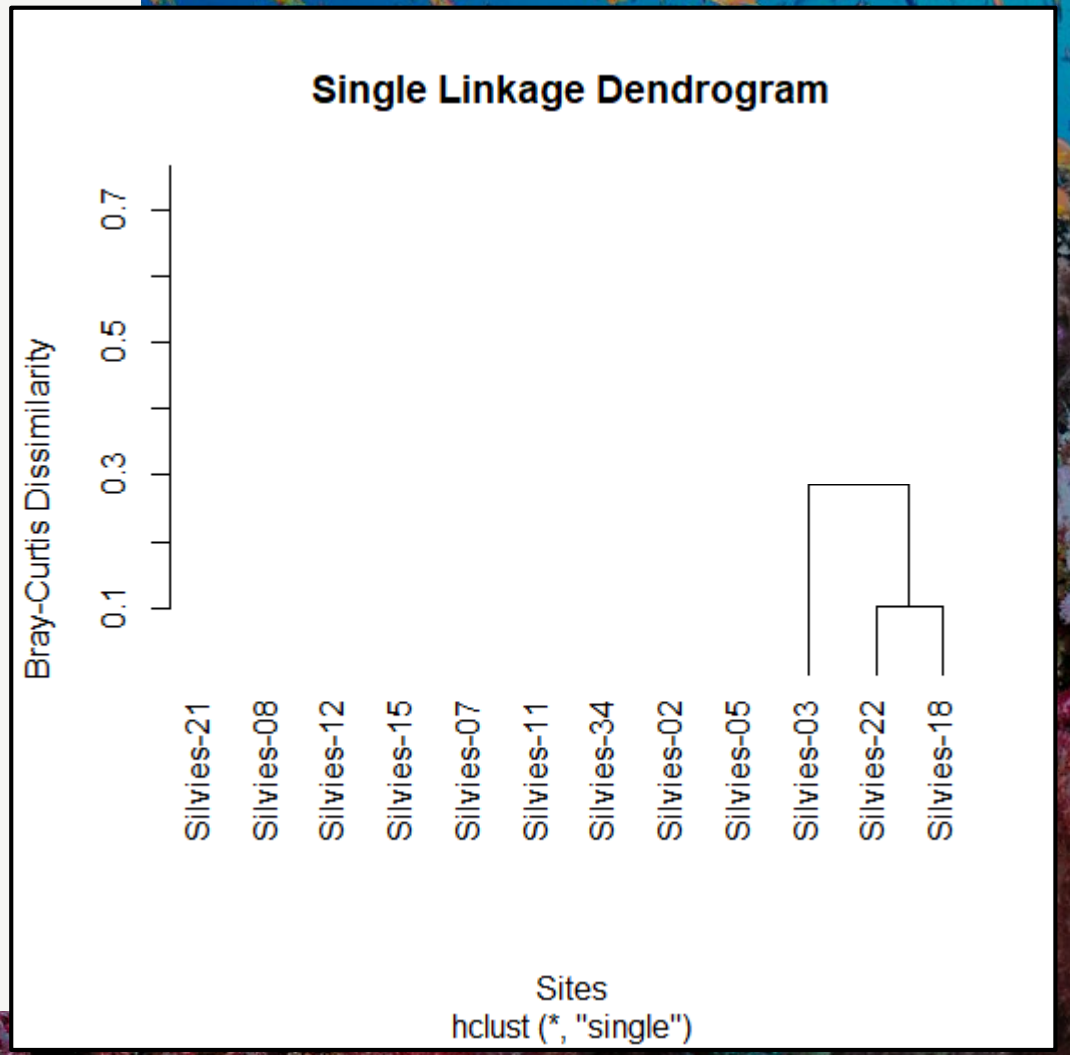
Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02
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-22/18	1.00	0.73	0.95	0.79	0.70	0.88	0.83	0.31	0.29	0.34



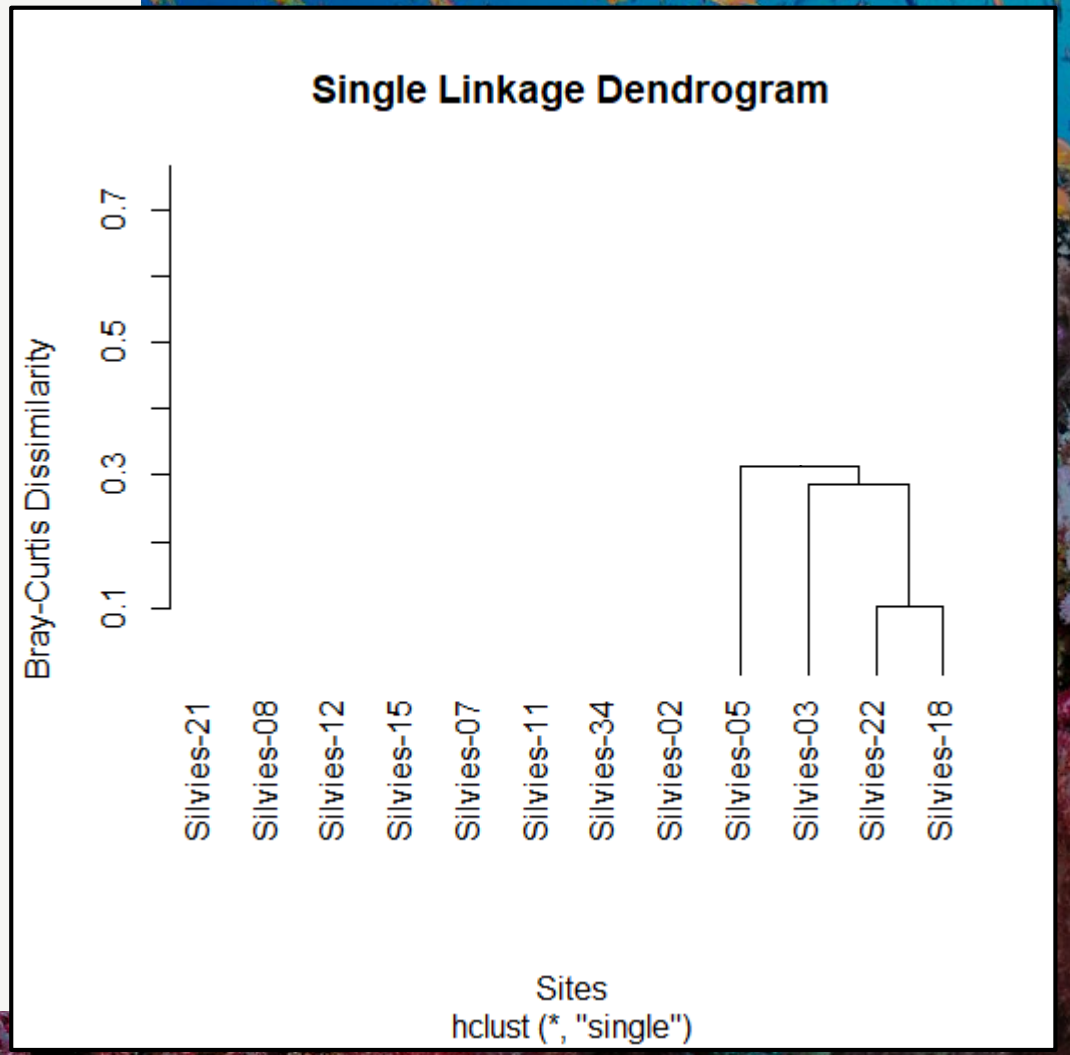
Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-03	S-02
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-22/18	**1.00**	**0.73**	0.95	**0.79**	**0.70**	0.88	0.83	**0.31**	0.29	**0.34**



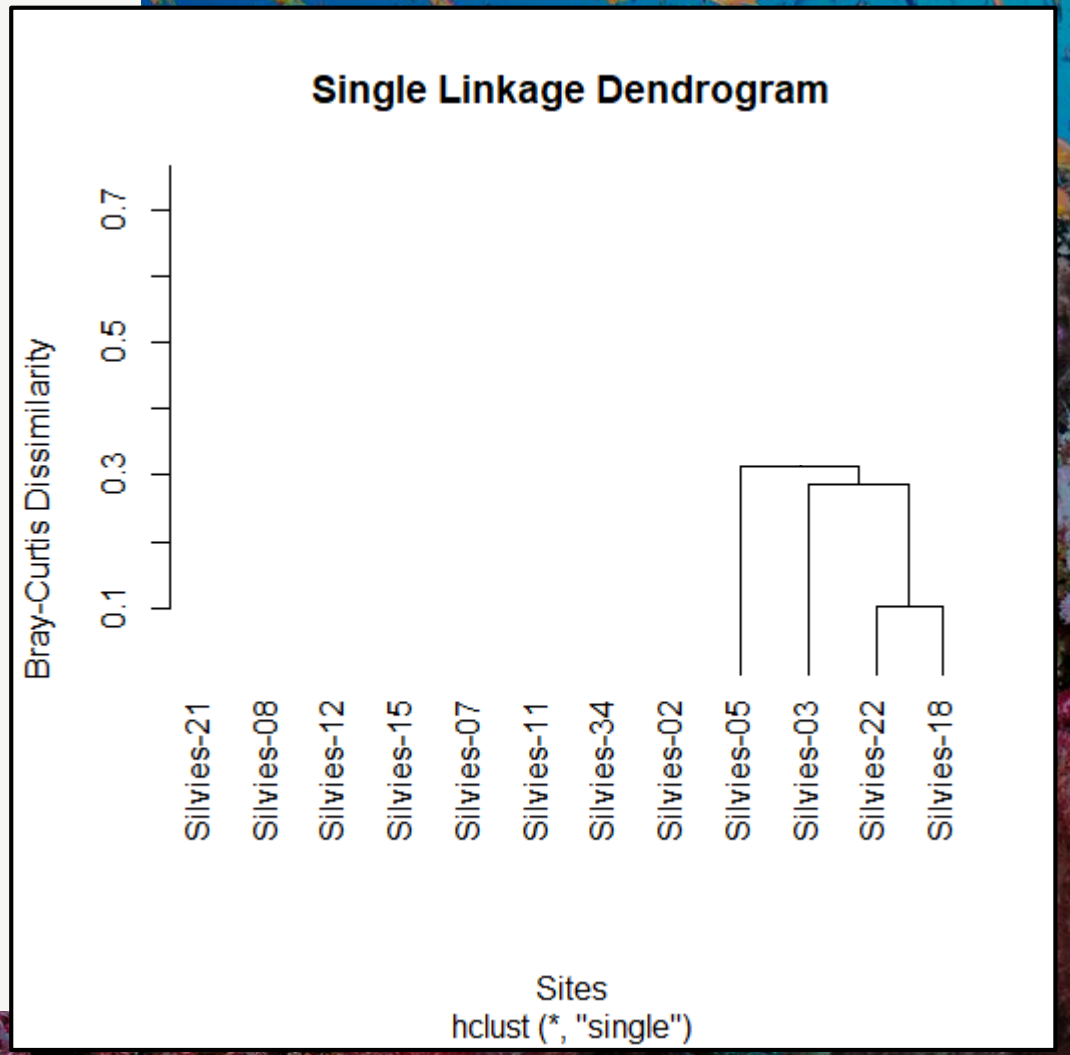
Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-02
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-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	
Silvies-22/18/03	1.00	0.73	0.93	0.79	0.70	0.80	0.73	0.31	0.34



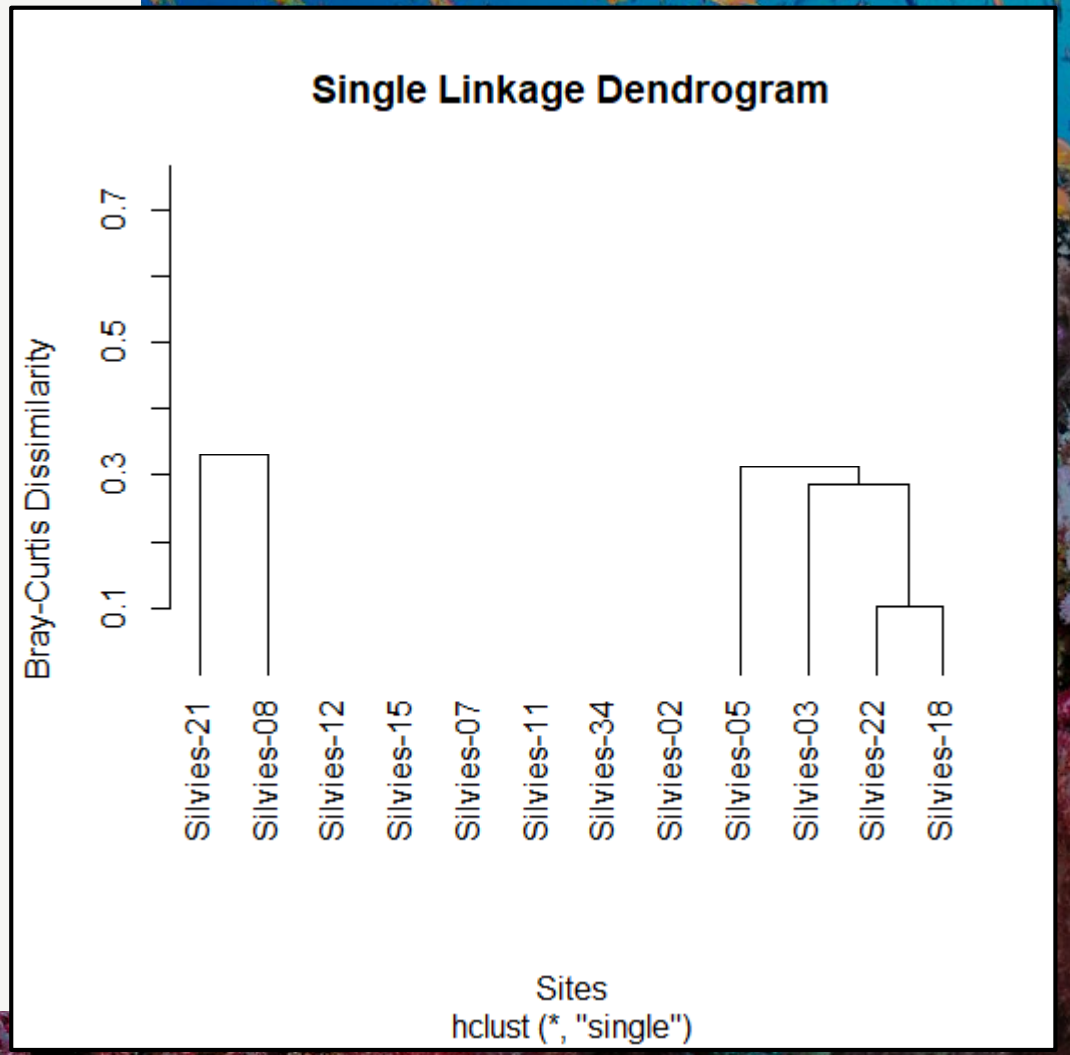
Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-05	S-02
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-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	0.42	
Silvies-22/18/03	1.00	0.73	0.93	**0.79**	**0.70**	**0.80**	**0.73**	0.31	**0.34**



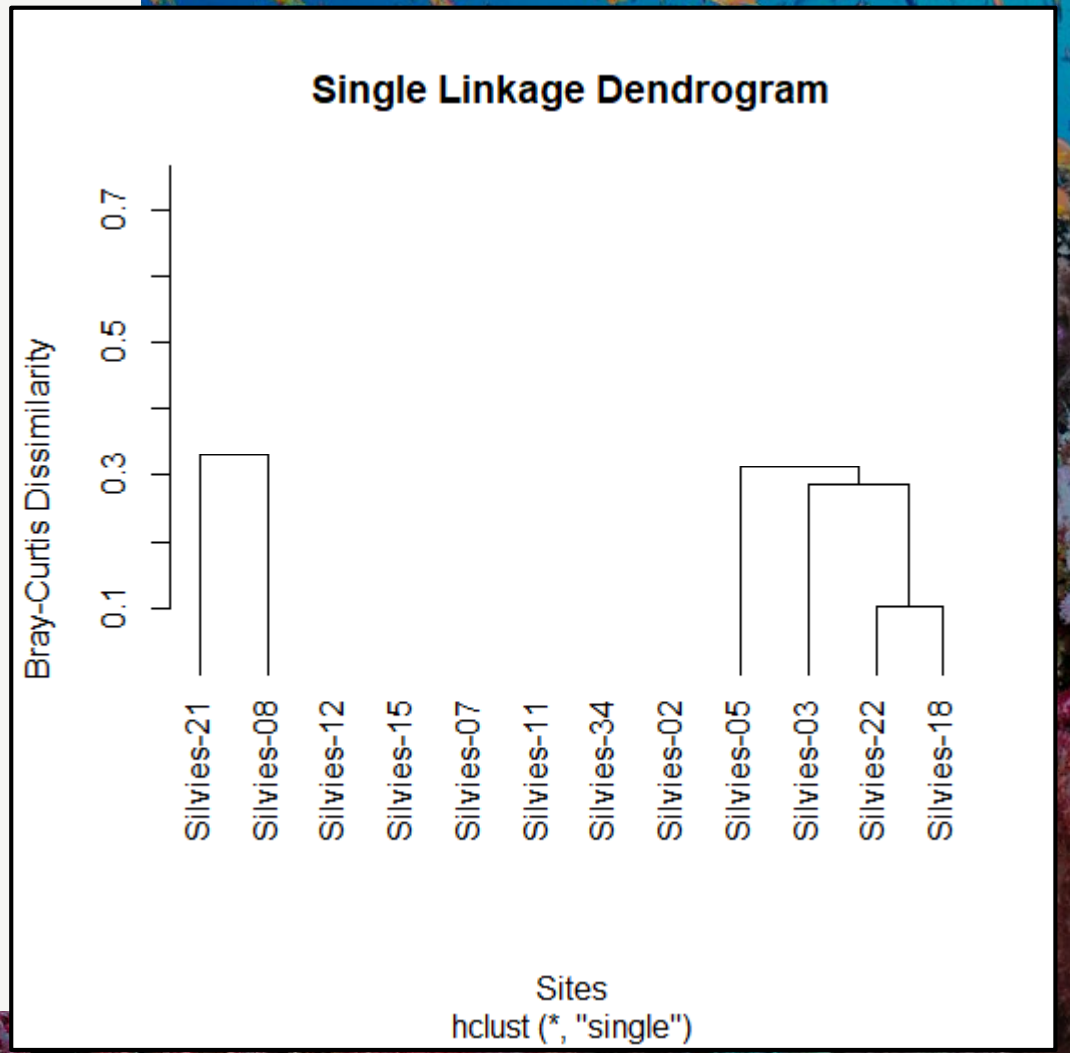
Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-02
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-02	1.00	0.64	0.95	0.50	0.58	0.86	0.74	
Silvies-22/18/03/05	0.75	0.72	0.76	0.79	0.70	0.80	0.73	0.34



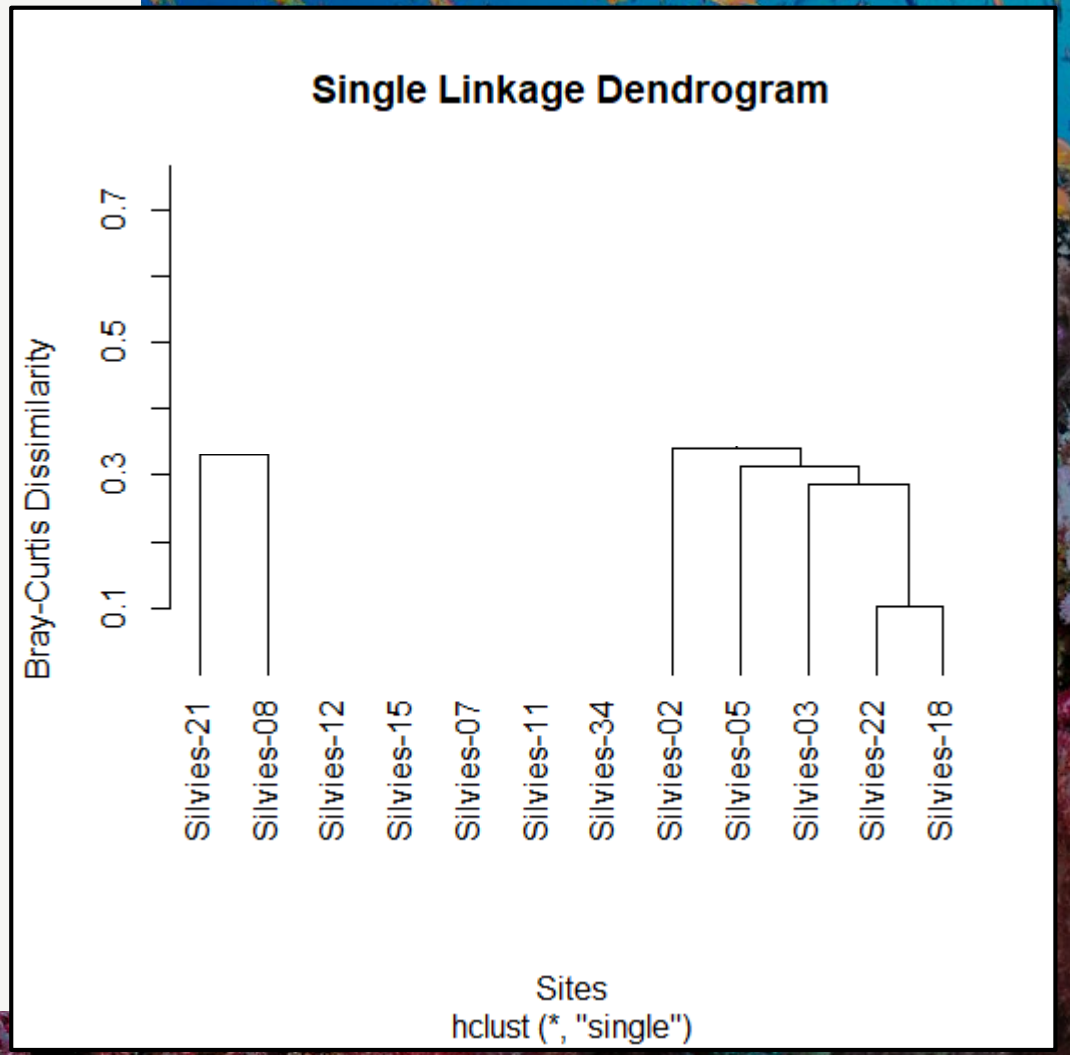
Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-21	S-08	S-02
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-02	1.00	0.64	0.95	0.50	0.58	0.86	**0.74**	
Silvies-22/18/03/05	0.75	0.72	0.76	0.79	0.70	0.80	**0.73**	0.34



Agglomerative Hierarchical Clustering Methods: Single Linkage

	S-15	S-07	S-12	S-34	S-11	S-08/21	S-02
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-08/21	1.00	0.87	0.88	0.93	1.00		
Silvies-02	1.00	0.64	0.95	0.50	0.58	0.74	
Silvies-22/18/03/05	0.75	0.72	0.76	0.79	0.70	0.73	0.34



Agglomerative Hierarchical Clustering Methods: Single Linkage

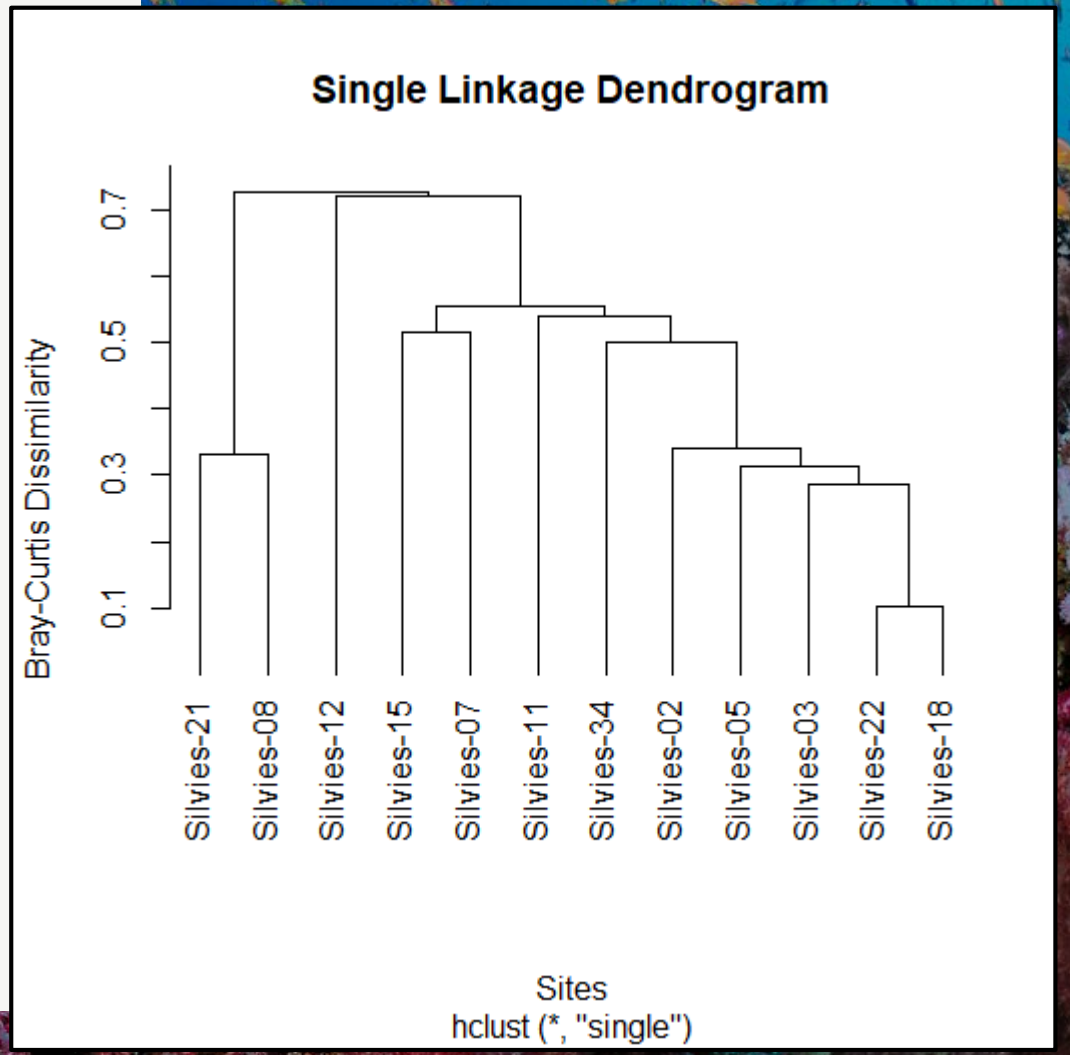
Single linkage clustering merges clusters based on the *minimum* distance between points.

Advantages:

- Simple and intuitive
- Captures elongated or irregularly shaped clusters

Disadvantages:

- Sensitive to noise and outliers
- Can lead to chaining effect



Agglomerative Hierarchical Clustering Methods: Single Linkage

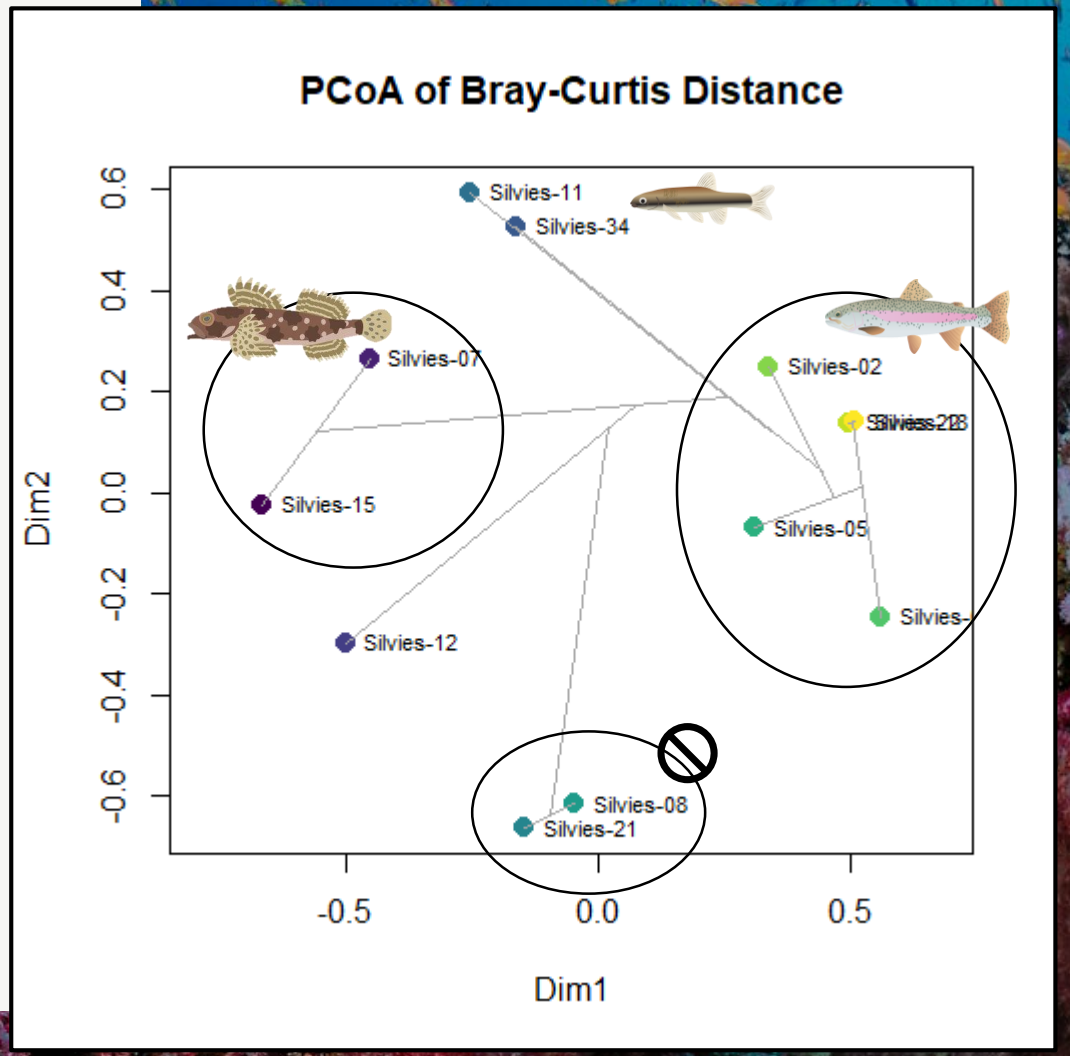
Single linkage clustering merges clusters based on the *minimum* distance between points.

Advantages:

- Simple and intuitive
- Captures elongated or irregularly shaped clusters

Disadvantages:

- Sensitive to noise and outliers
- Can lead to chaining effect



Agglomerative Hierarchical Clustering Methods: Complete Linkage

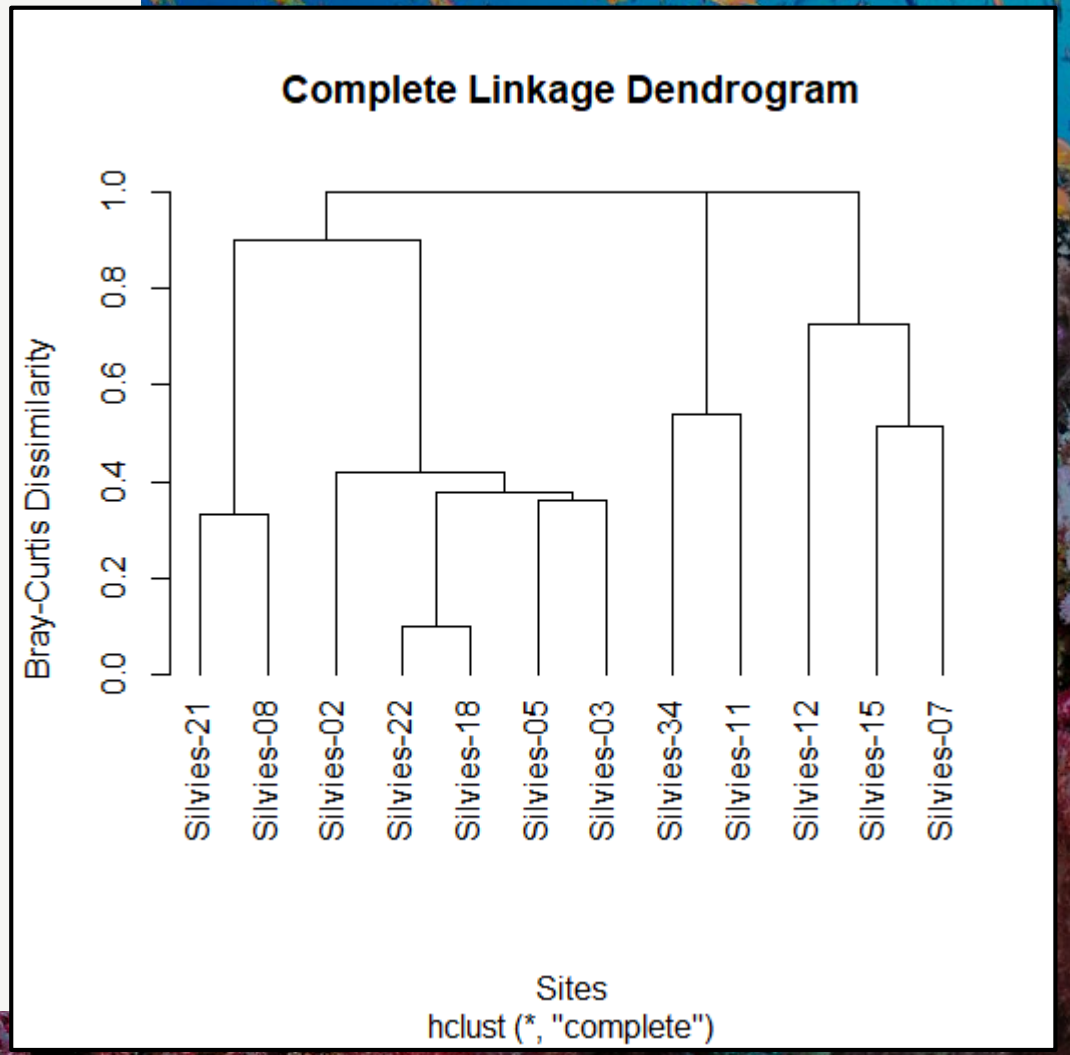
Complete linkage clustering merges clusters based on the *maximum* distance between points.

Advantages:

- Simple and intuitive
- Produces compact and spherical clusters
- Less prone to chaining effect

Disadvantages:

- Sensitive to noise and outliers



Agglomerative Hierarchical Clustering Methods: Complete Linkage

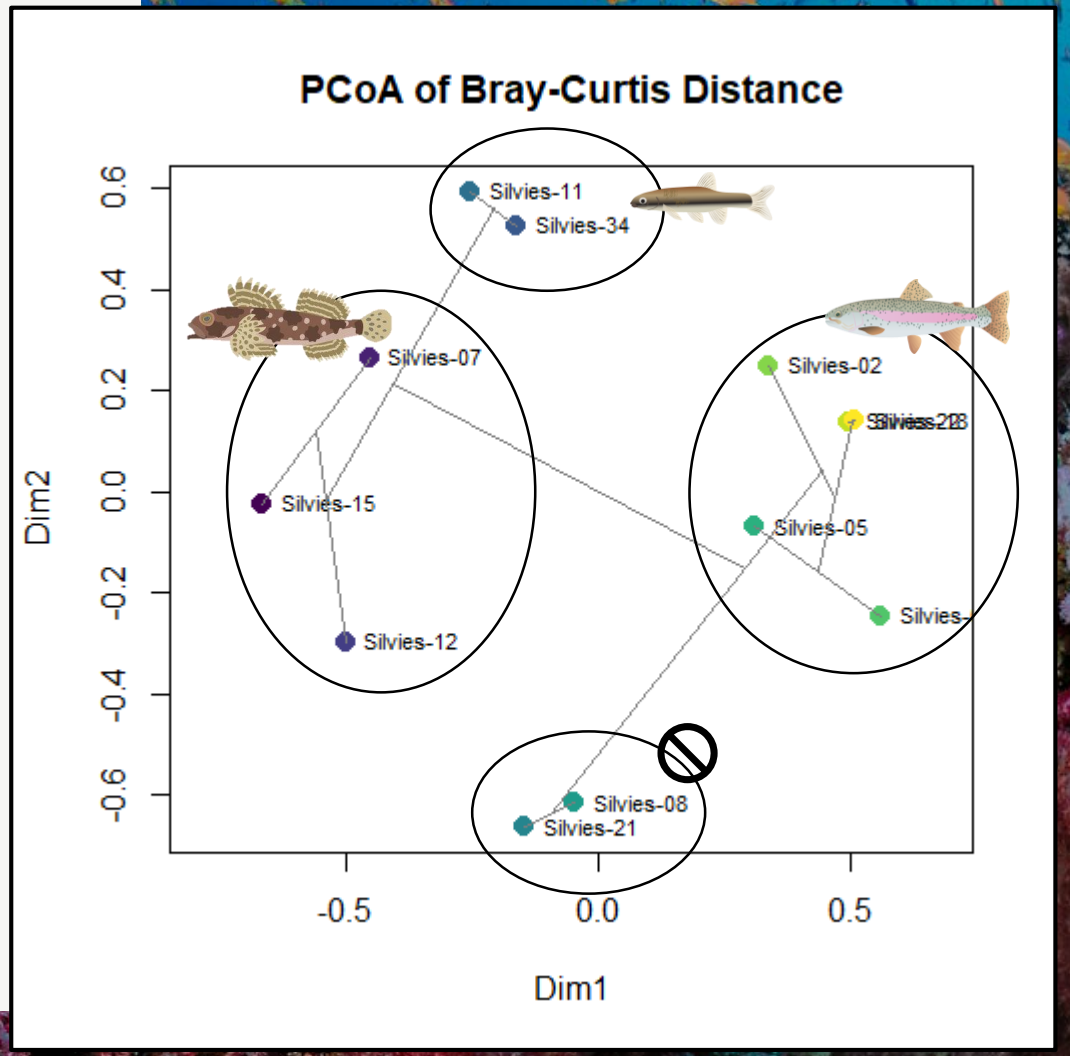
Complete linkage clustering merges clusters based on the *maximum* distance between points.

Advantages:

- Simple and intuitive
- Produces compact and spherical clusters
- Less prone to chaining effect

Disadvantages:

- Sensitive to noise and outliers

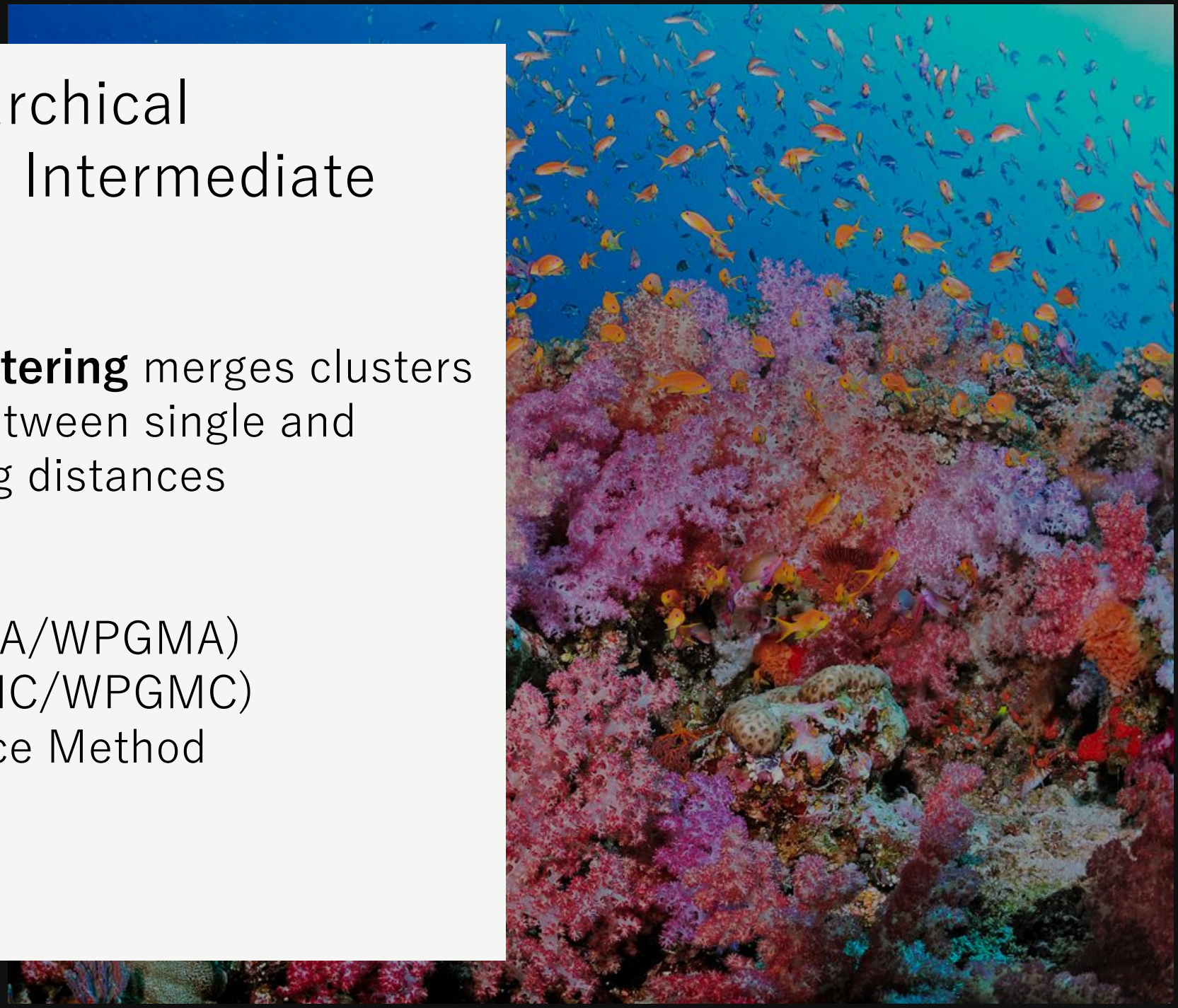


Agglomerative Hierarchical Clustering Methods: Intermediate Linkage

Intermediate linkage clustering merges clusters based on a compromise between single and complete linkage clustering distances

Variants include:

- Average Linkage (UPGMA/WPGMA)
- Centroid Linkage (UPGMC/WPGMC)
- Ward's Minimum Variance Method



Agglomerative Hierarchical Clustering Methods: Average Linkage (UPGMA)

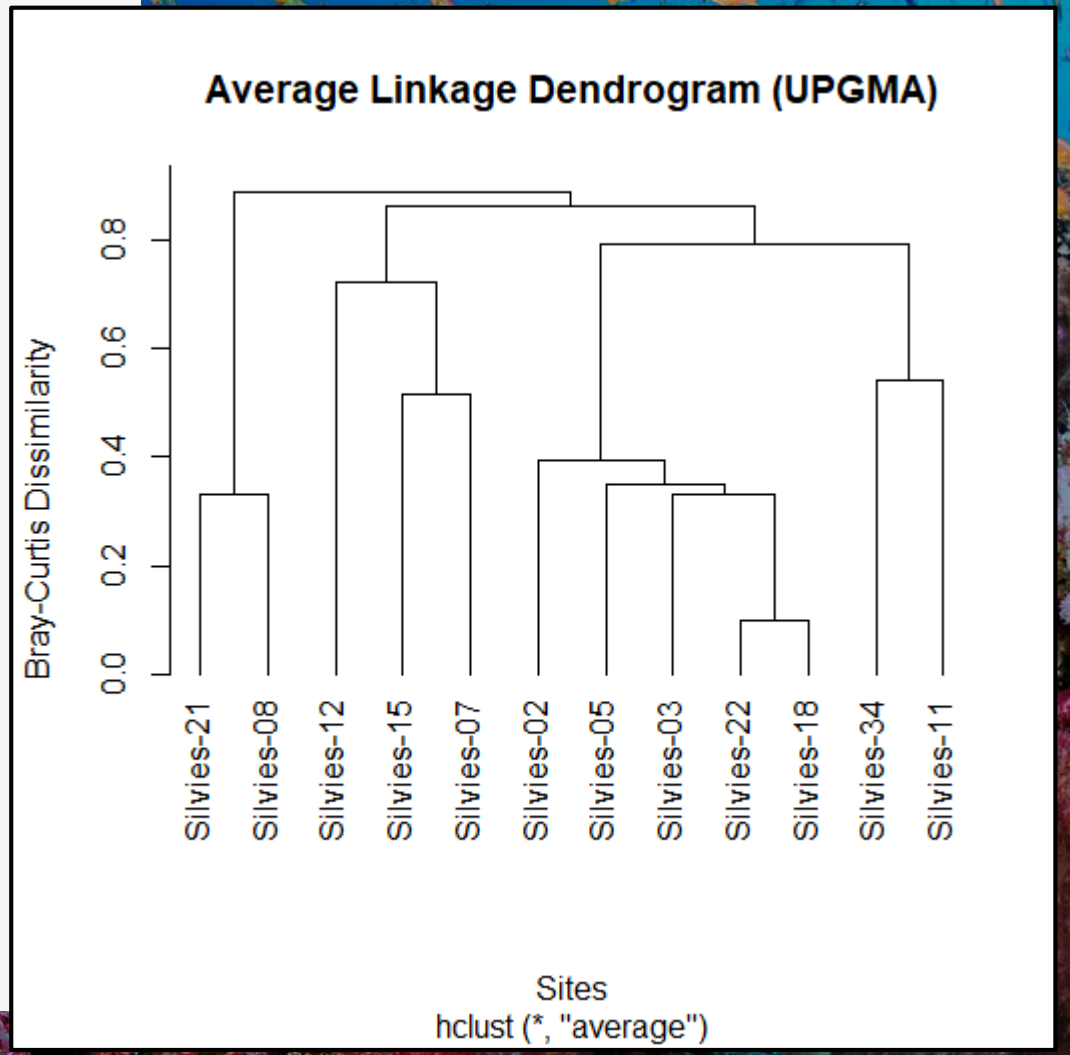
Average linkage clustering merges clusters based on the *average* distance between points.

Advantages:

- Simple to understand and implement
- Produces dendrograms that represent the hierarchy of clusters

Disadvantages:

- Sensitive to noise and outliers
- Assumes a constant rate of evolution



Agglomerative Hierarchical Clustering Methods: Average Linkage (UPGMA)

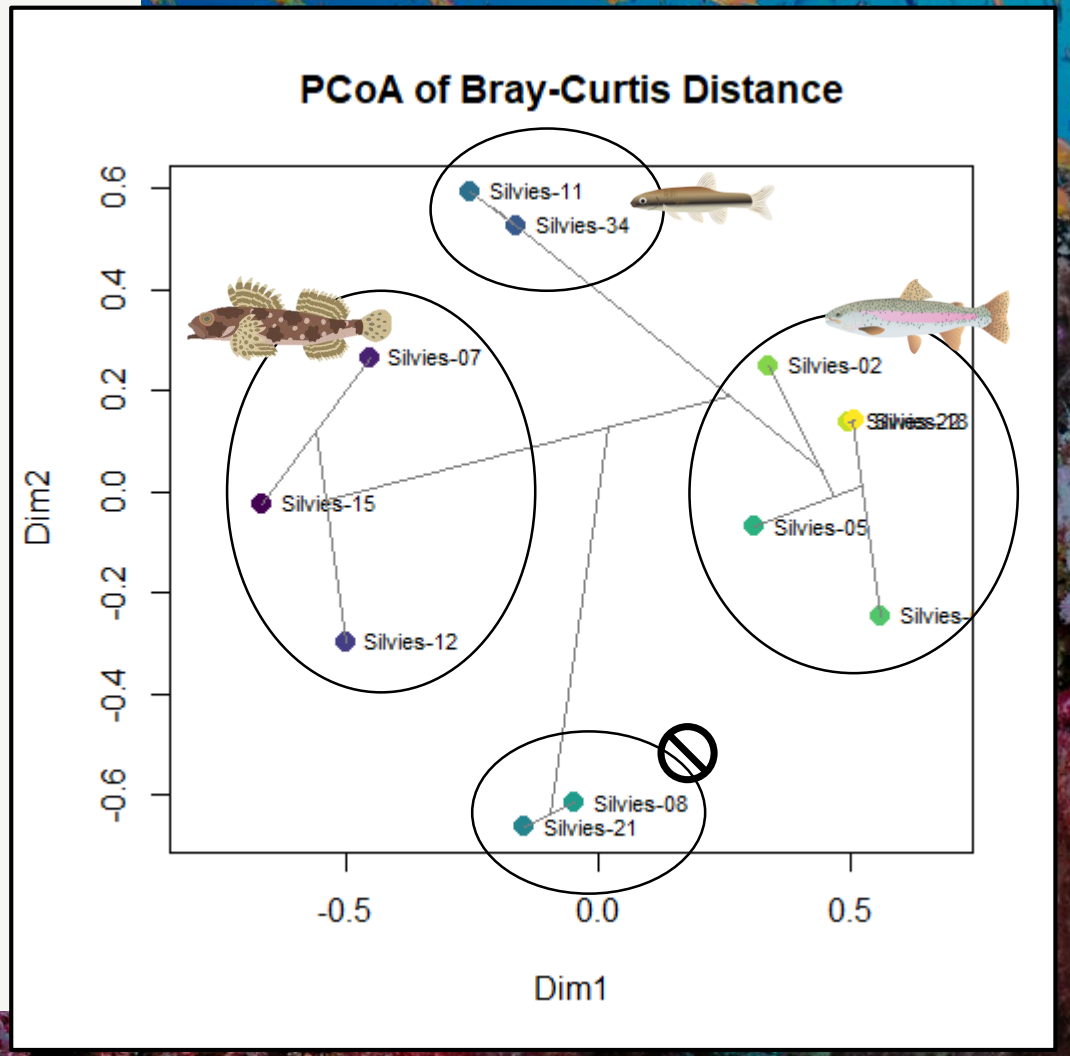
Average linkage clustering merges clusters based on the *average* distance between points.

Advantages:

- Simple to understand and implement
- Produces dendrograms that represent the hierarchy of clusters

Disadvantages:

- Sensitive to noise and outliers



Agglomerative Hierarchical Clustering Methods: Weighted Average (WPGMA)

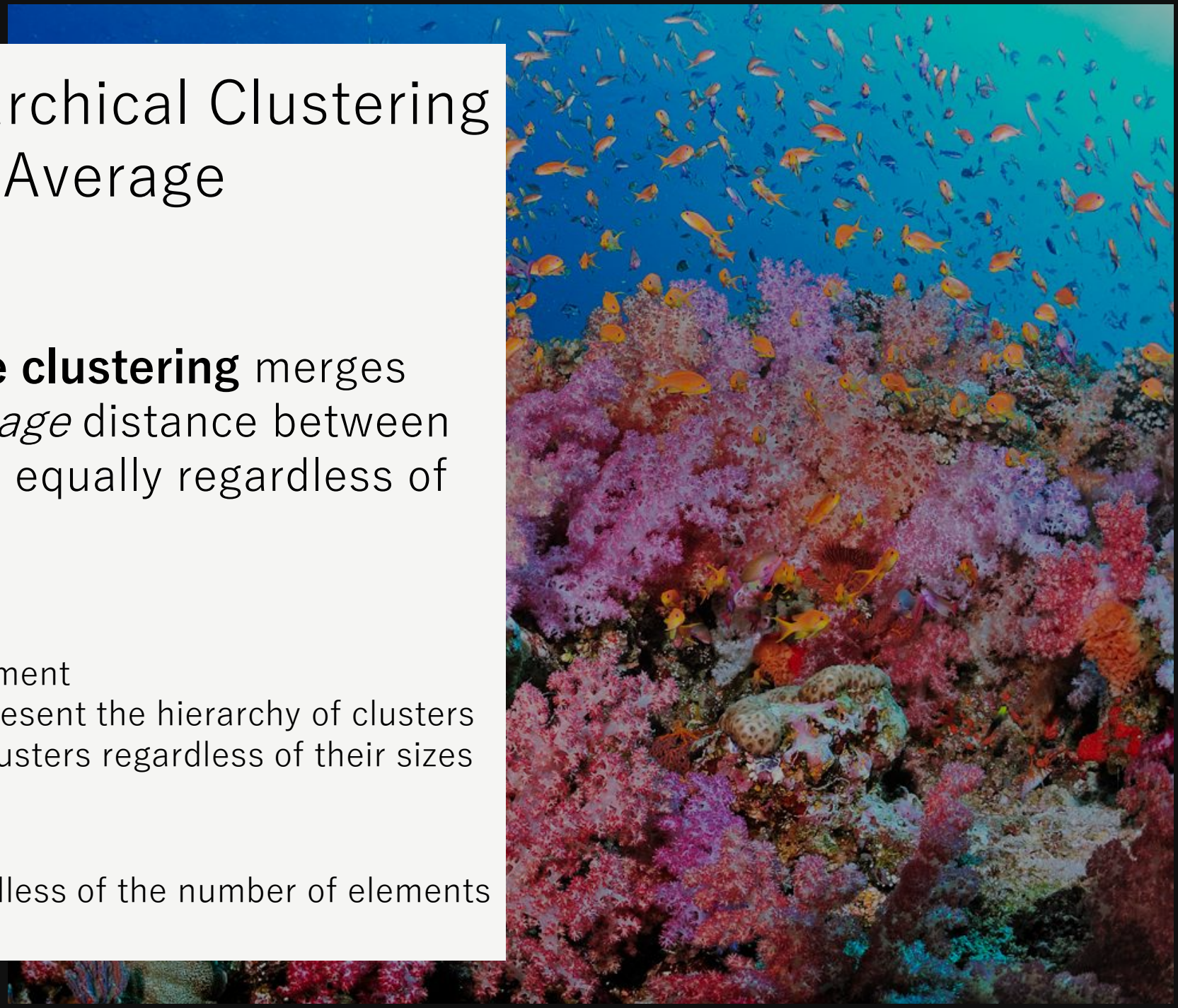
Weighted average linkage clustering merges clusters based on the *average* distance between points, but weighs clusters equally regardless of their size.

Advantages:

- Simple to understand and implement
- Produces dendrograms that represent the hierarchy of clusters
- Maintains equal weighting for clusters regardless of their sizes

Disadvantages:

- Treats all clusters equally regardless of the number of elements
- Sensitive to noise and outliers



Agglomerative Hierarchical Clustering Methods: Unweighted Centroid (UPGMC)

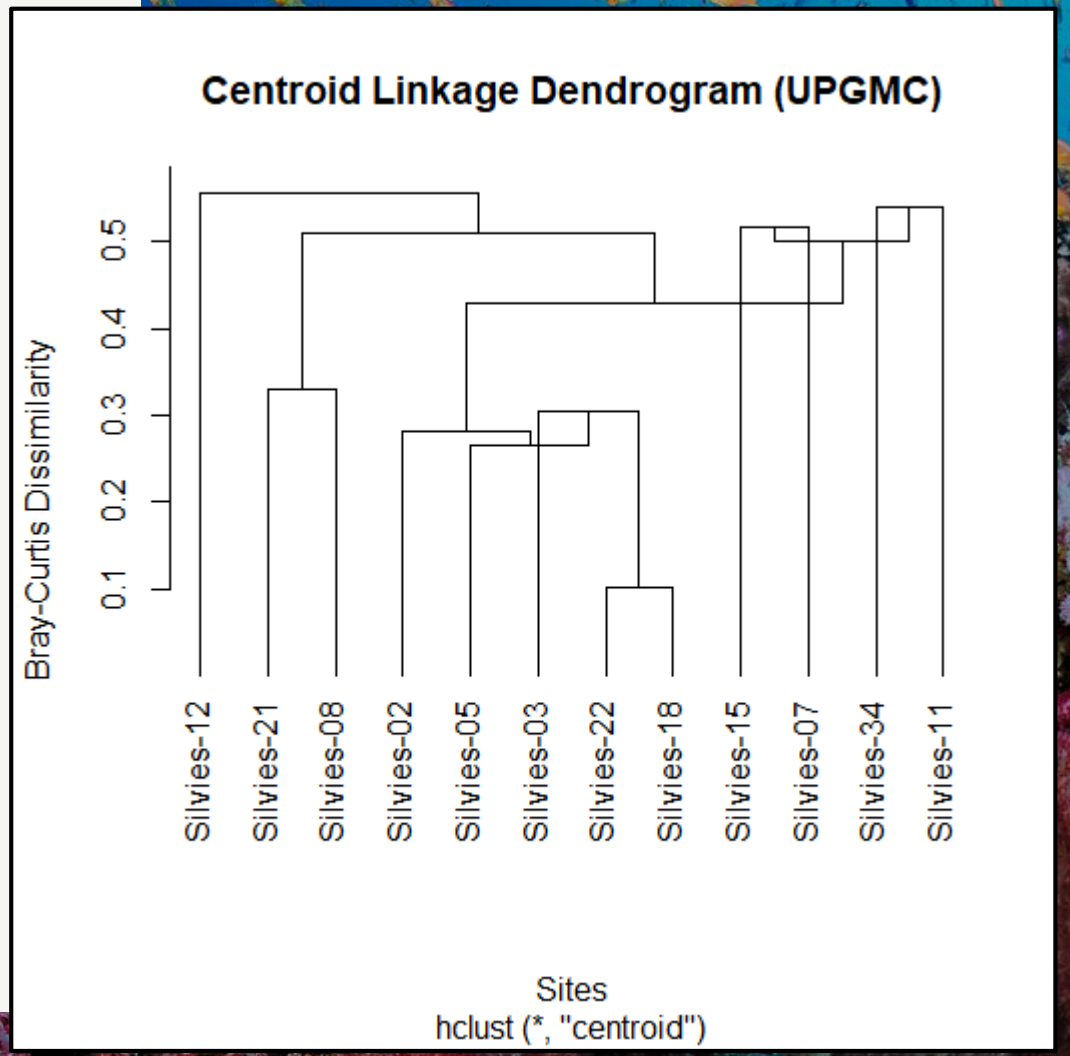
Unweighted centroid clustering merges clusters based on the *distance between their centroids* (geometric center).

Advantages:

- Produces clusters based on geometric center, which can be more representative of the cluster's overall location
- Less sensitive to outliers

Disadvantages:

- Sensitive to the shape and density of clusters



Agglomerative Hierarchical Clustering Methods: Unweighted Centroid (UPGMC)

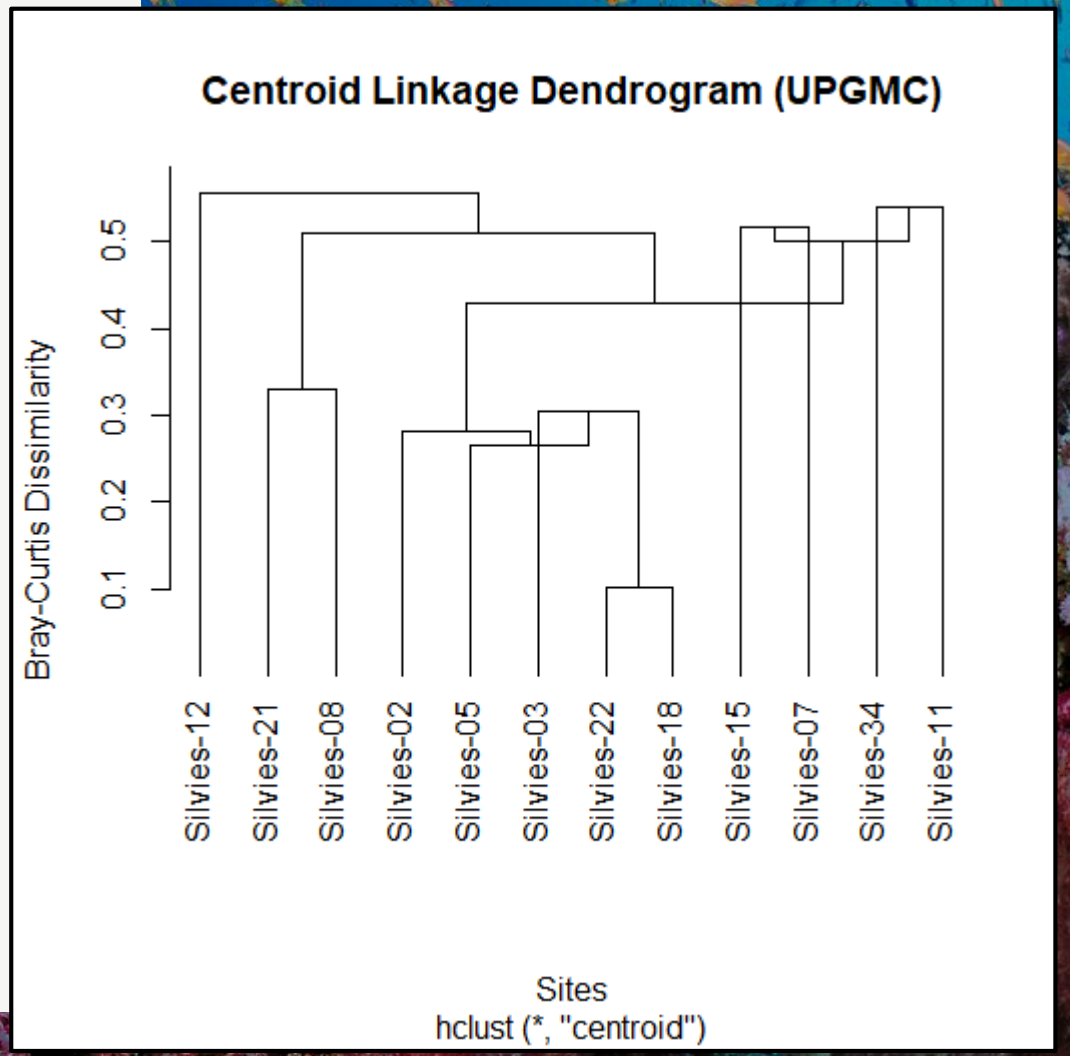
Unweighted centroid clustering merges clusters based on the *distance between their centroids* (geometric center).

Advantages:

- Produces clusters based on geometric center, which can be more representative of the cluster's overall location
- Less sensitive to outliers

Disadvantages:

- Sensitive to the shape and density of clusters
- Can result in **reversals**



Agglomerative Hierarchical Clustering Methods: Unweighted Centroid (UPGMC)

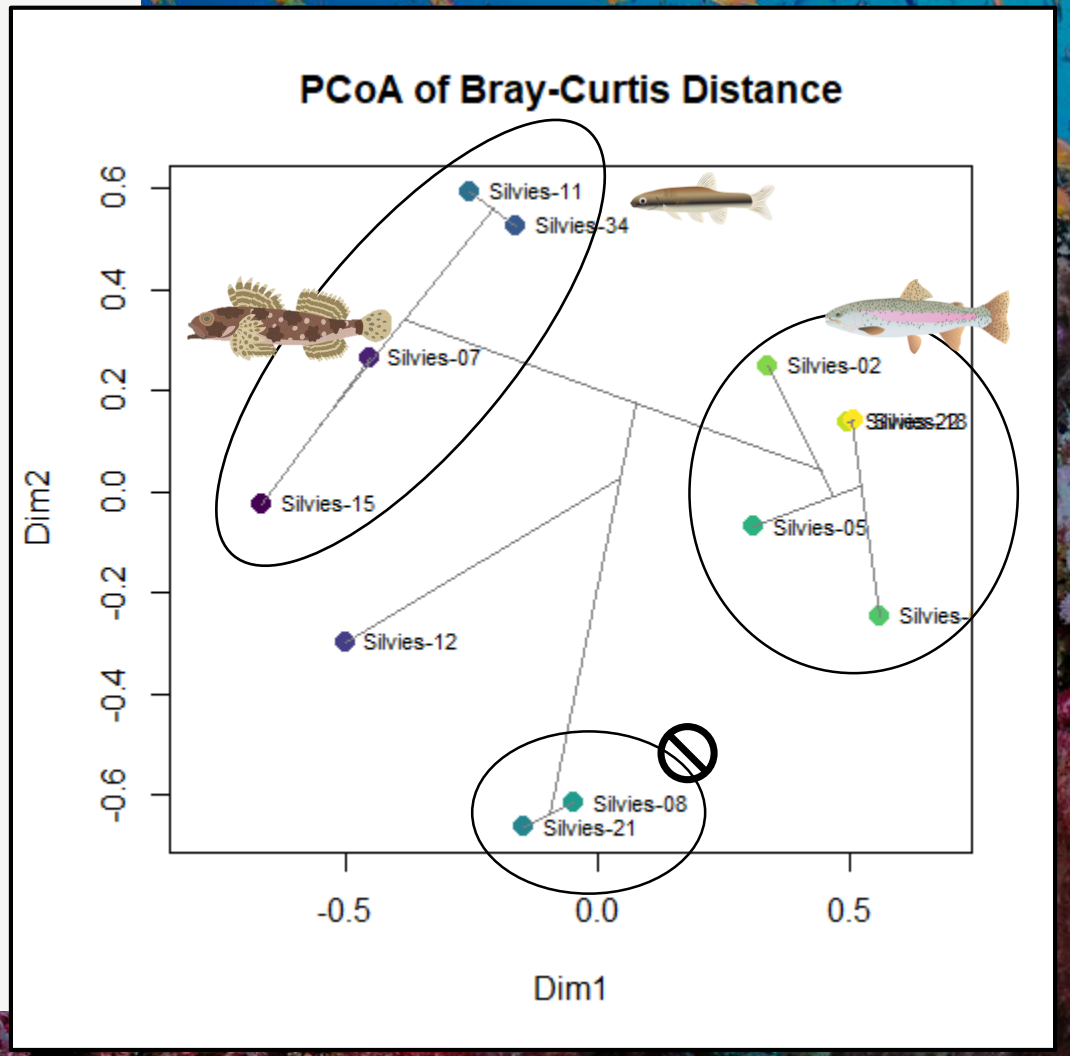
Unweighted centroid clustering merges clusters based on the *distance between their centroids* (geometric center).

Advantages:

- Produces clusters based on geometric center, which can be more representative of the cluster's overall location
- Less sensitive to outliers

Disadvantages:

- Sensitive to the shape and density of clusters
- Can result in reversals



Agglomerative Hierarchical Clustering Methods: Weighted Centroid (WPGMC)

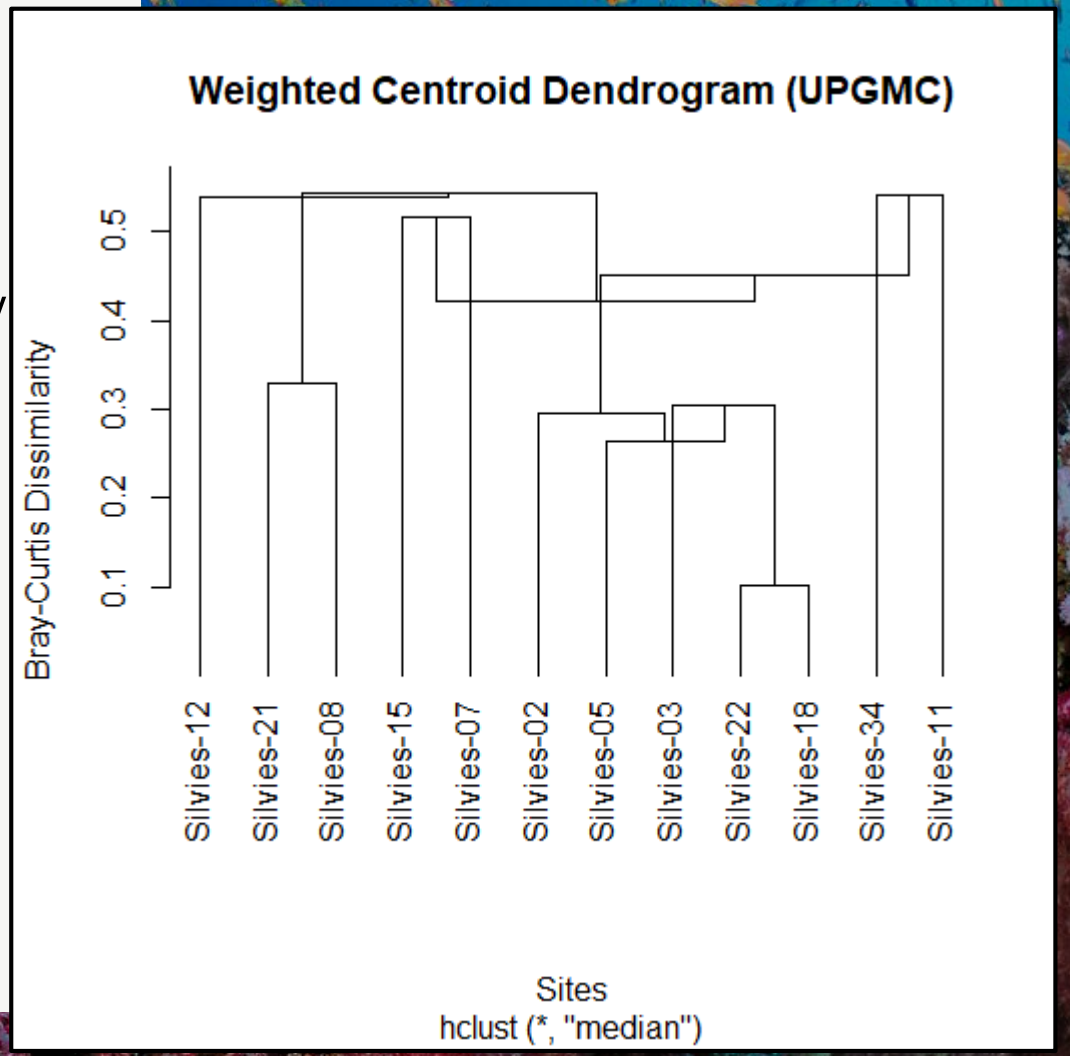
Weighted centroid clustering merges clusters based on the *distance between their centroids,* where clusters are weighted equally regardless of their size.

Advantages:

- Produces clusters based on geometric center
- Less sensitive to outliers

Disadvantages:

- Sensitive to the shape and density of clusters
- Can produce clusters with varying sizes
- Can result in reversals



Agglomerative Hierarchical Clustering Methods: Weighted Centroid (WPGMC)

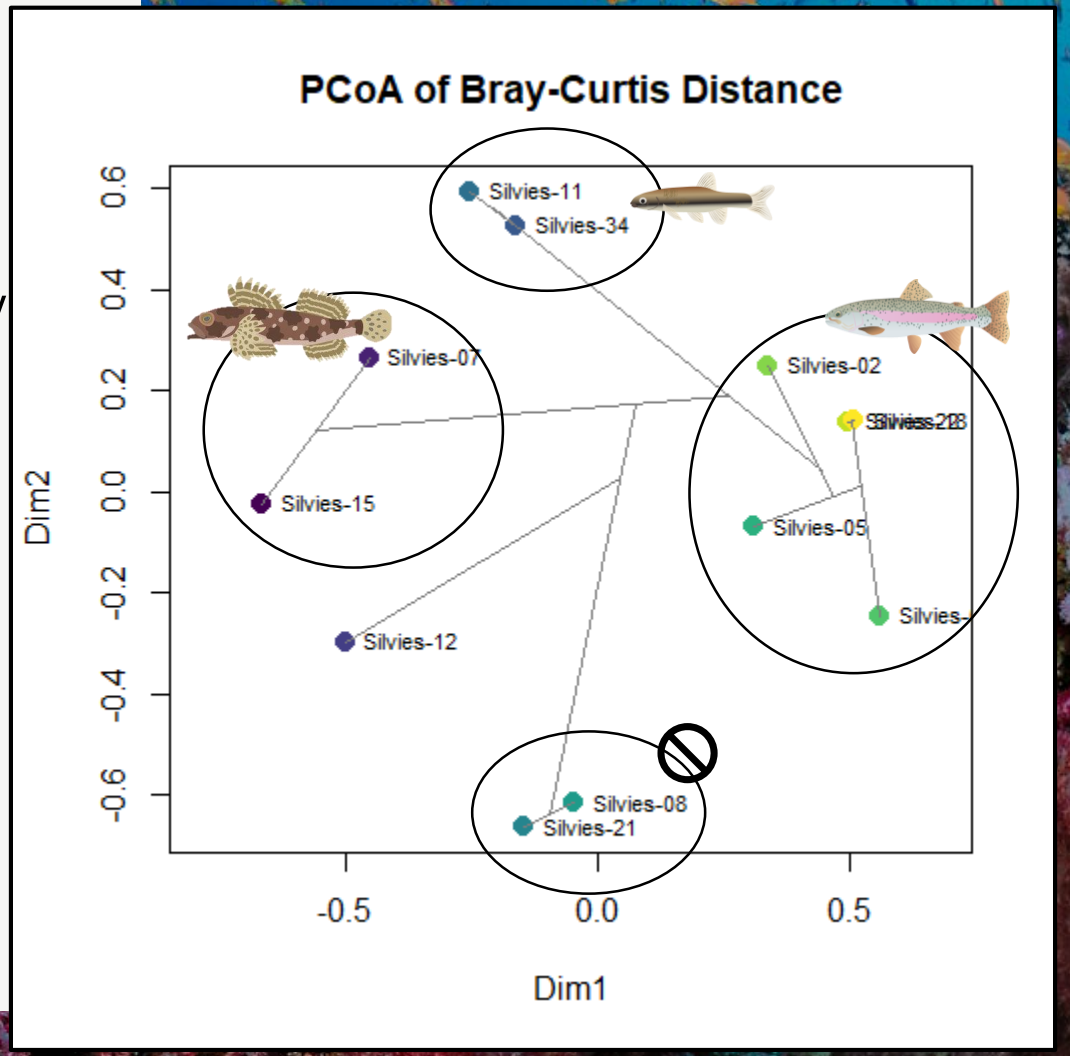
Weighted centroid clustering merges clusters based on the *distance between their centroids,* where clusters are weighted equally regardless of their size.

Advantages:

- Produces clusters based on geometric center
- Less sensitive to outliers

Disadvantages:

- Sensitive to the shape and density of clusters
- Can produce clusters with varying sizes
- Can result in reversals



Agglomerative Hierarchical Clustering Methods: Ward's Minimum Variance

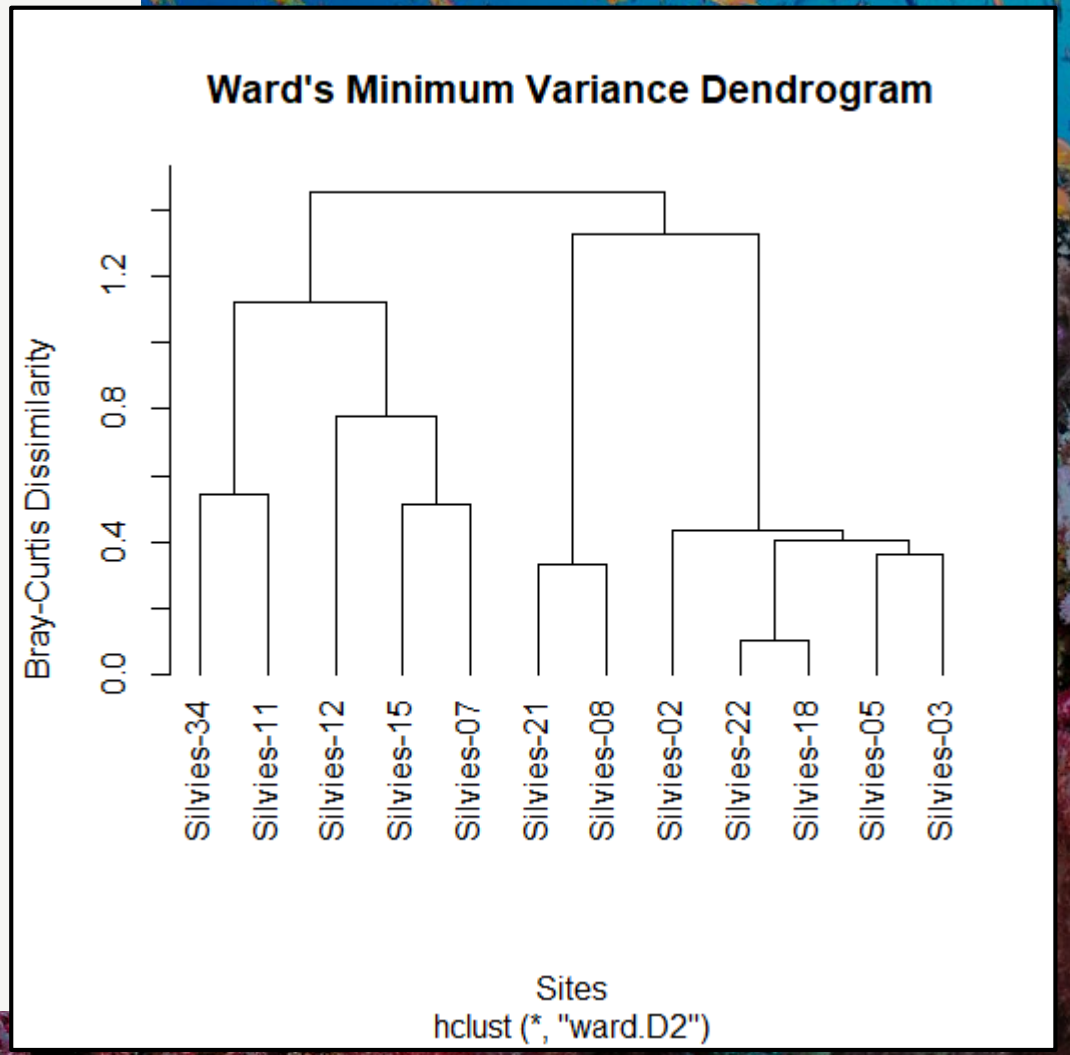
Ward's minimum variance merges clusters to achieve *the smallest possible increase in the sum of squared distances* within each cluster.

Advantages:

- Produces clusters with minimum variance, leading to compact and spherical clusters
- Often results in more balanced and interpretable clusters

Disadvantages:

- Computationally intensive for large datasets
- Sensitive to outliers



Agglomerative Hierarchical Clustering Methods: Ward's Minimum Variance

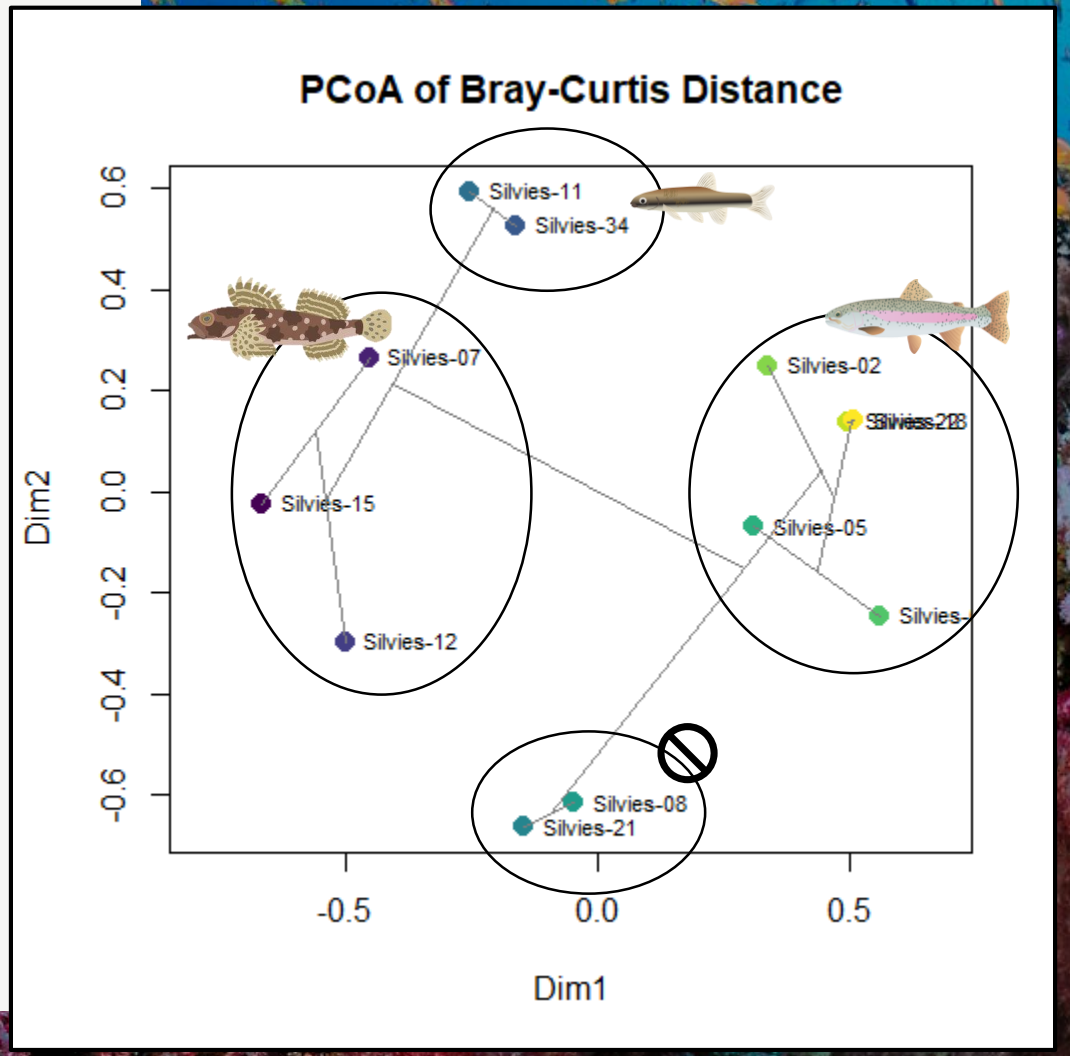
Ward's minimum variance merges clusters to achieve *the smallest possible increase in the sum of squared distances* within each cluster.

Advantages:

- Produces clusters with minimum variance, leading to compact and spherical clusters
- Often results in more balanced and interpretable clusters

Disadvantages:

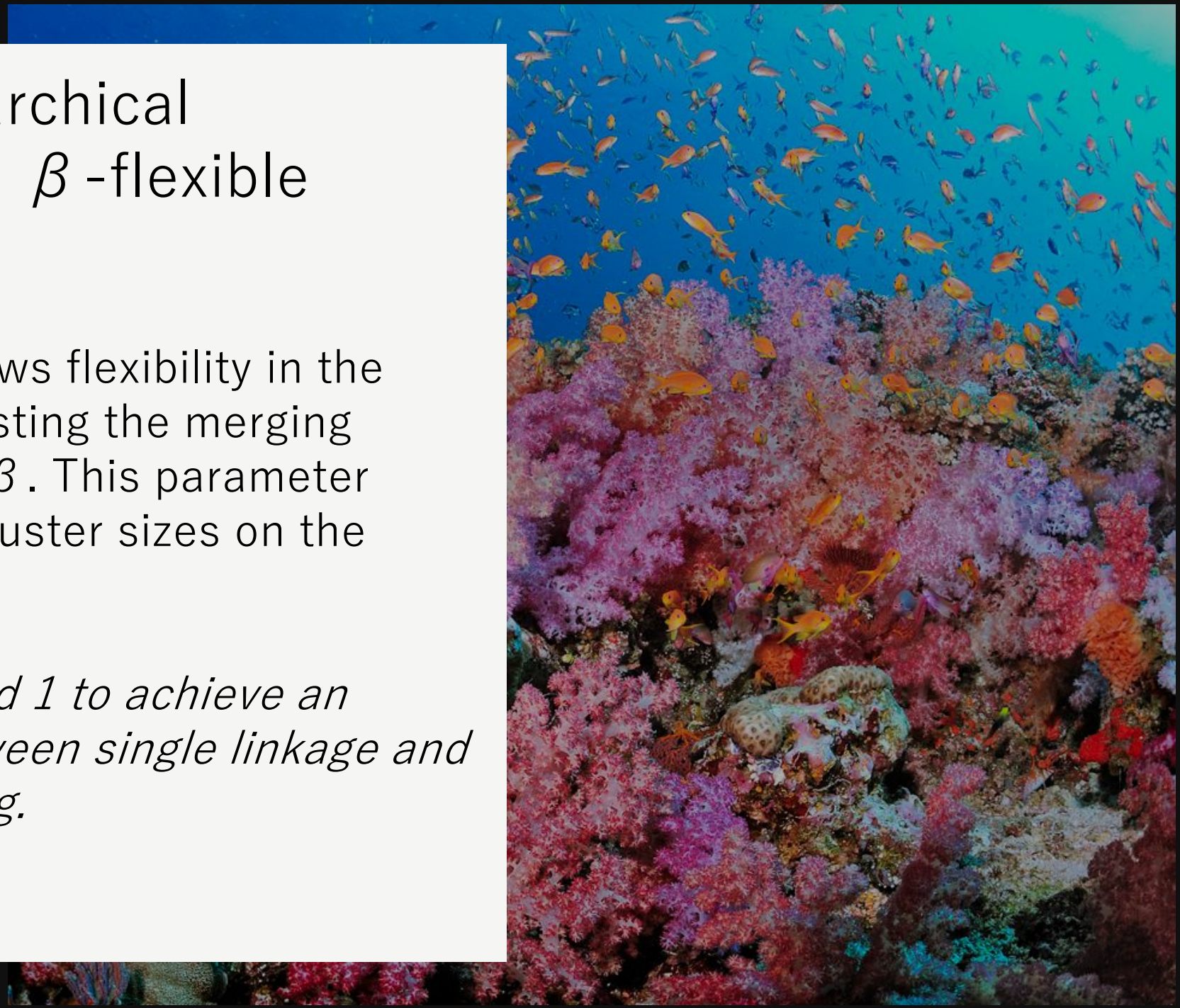
- Computationally intensive for large datasets
- Sensitive to outliers



Agglomerative Hierarchical Clustering Methods: β -flexible Clustering

β -flexible Clustering allows flexibility in the clustering process by adjusting the merging criteria with a parameter, β . This parameter controls the influence of cluster sizes on the distance calculation.

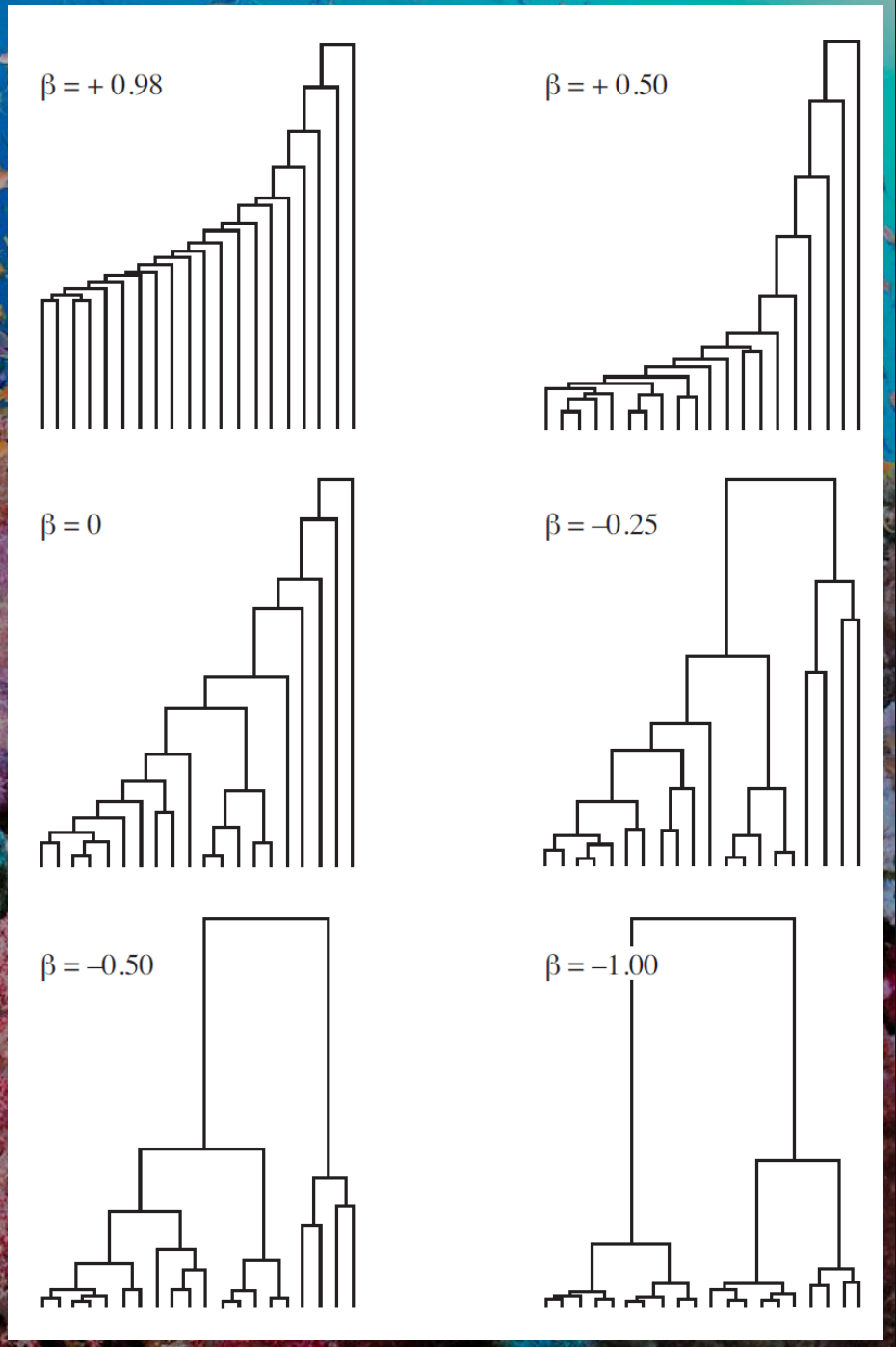
β is varied between -1 and 1 to achieve an intermediate solution between single linkage and complete linkage clustering.



Agglomerative Hierarchical Clustering Methods: β -flexible Clustering

β -flexible Clustering allows flexibility in the clustering process by adjusting the merging criteria with a parameter, β . This parameter controls the influence of cluster sizes on the distance calculation.

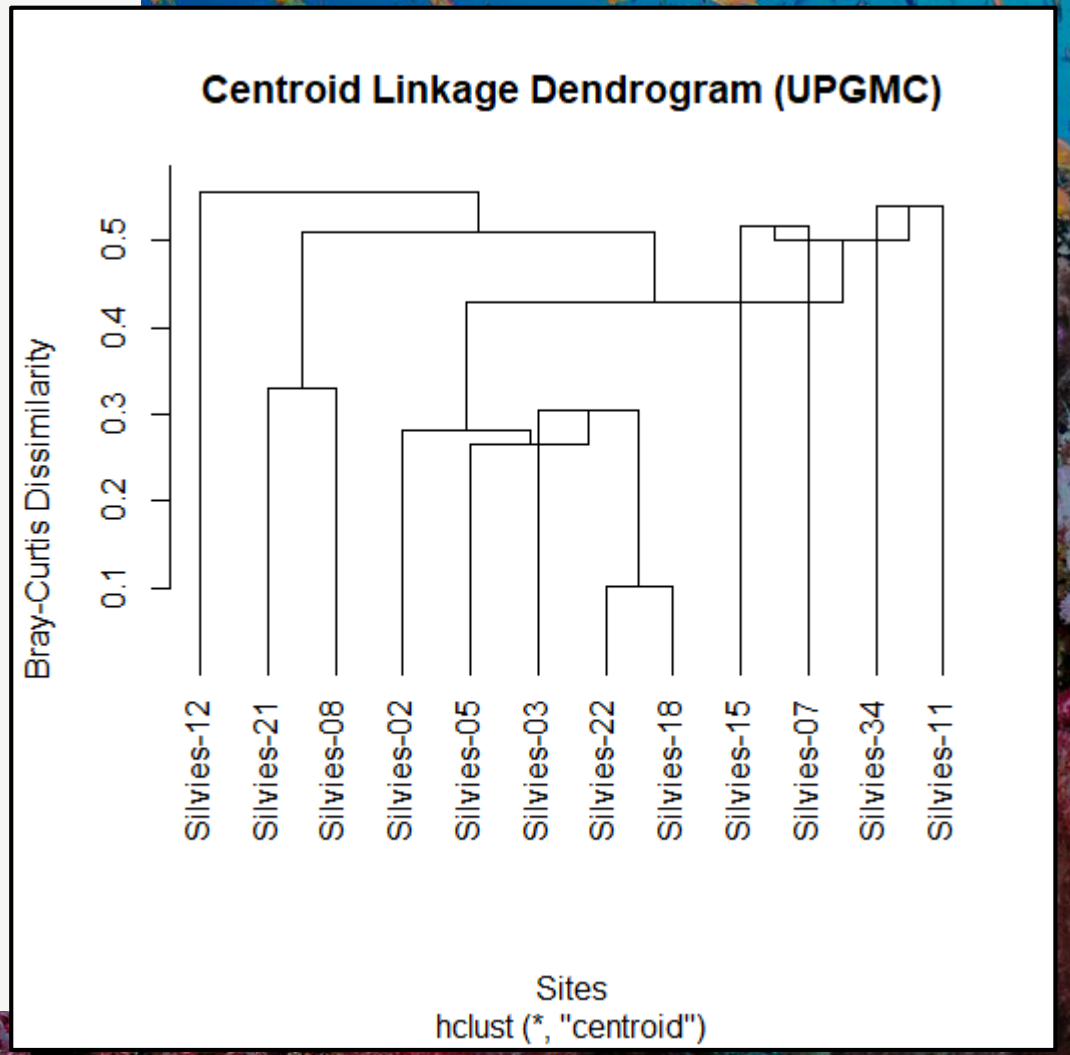
β is varied between -1 and 1 to achieve an intermediate solution between single linkage and complete linkage clustering.



Legendre & Legendre Figure 8.14

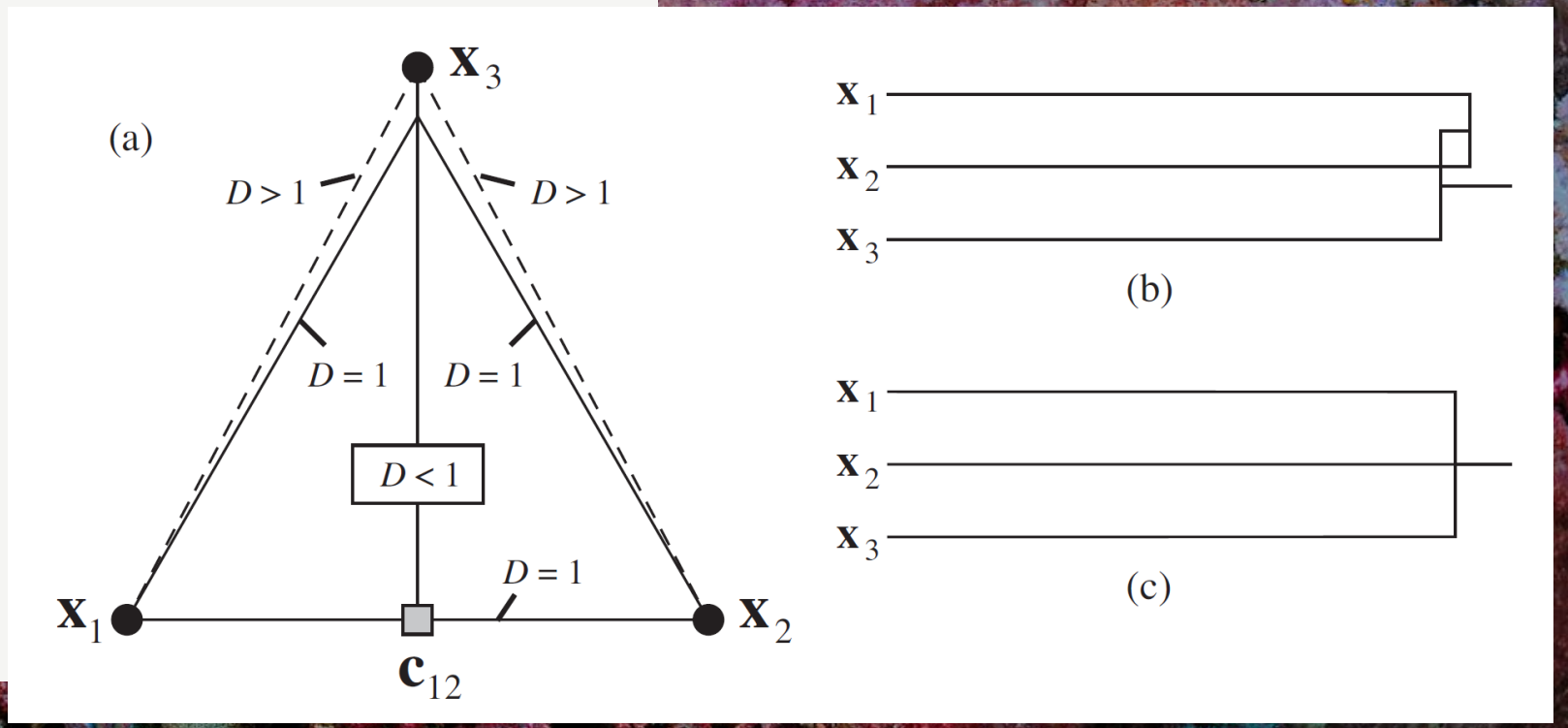
Agglomerative Hierarchical Clustering Methods: Reversals

A **reversal** occurs when $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first, even when the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller than the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$.



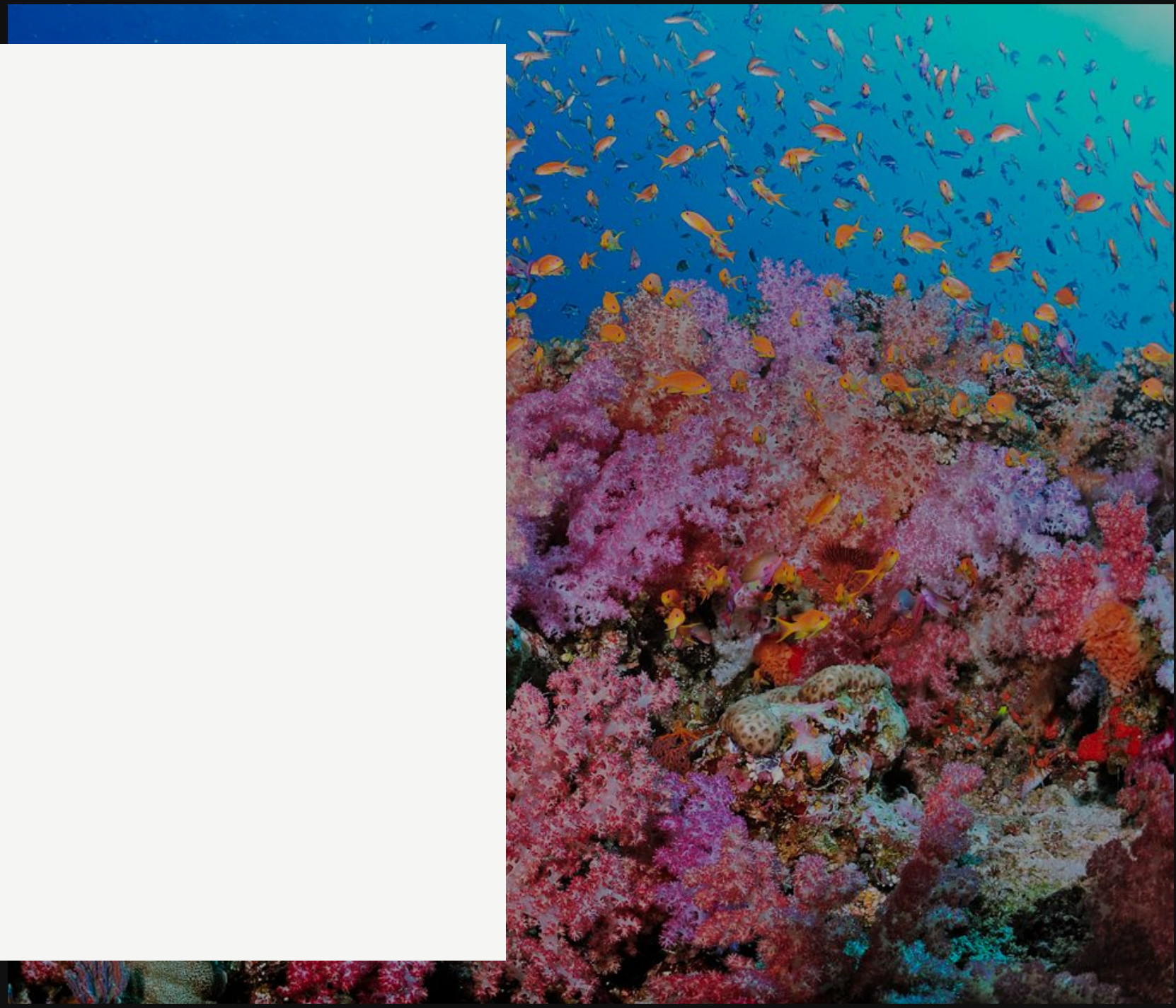
Agglomerative Hierarchical Clustering Methods: Reversals

A **reversal** occurs when $\mathbf{x}_1$ and $\mathbf{x}_2$ cluster first, even when the distance from $\mathbf{x}_3$ to the centroid $\mathbf{c}_{12}$ is smaller than the distance from $\mathbf{x}_1$ to $\mathbf{x}_2$.



Legendre & Legendre Figure 8.16

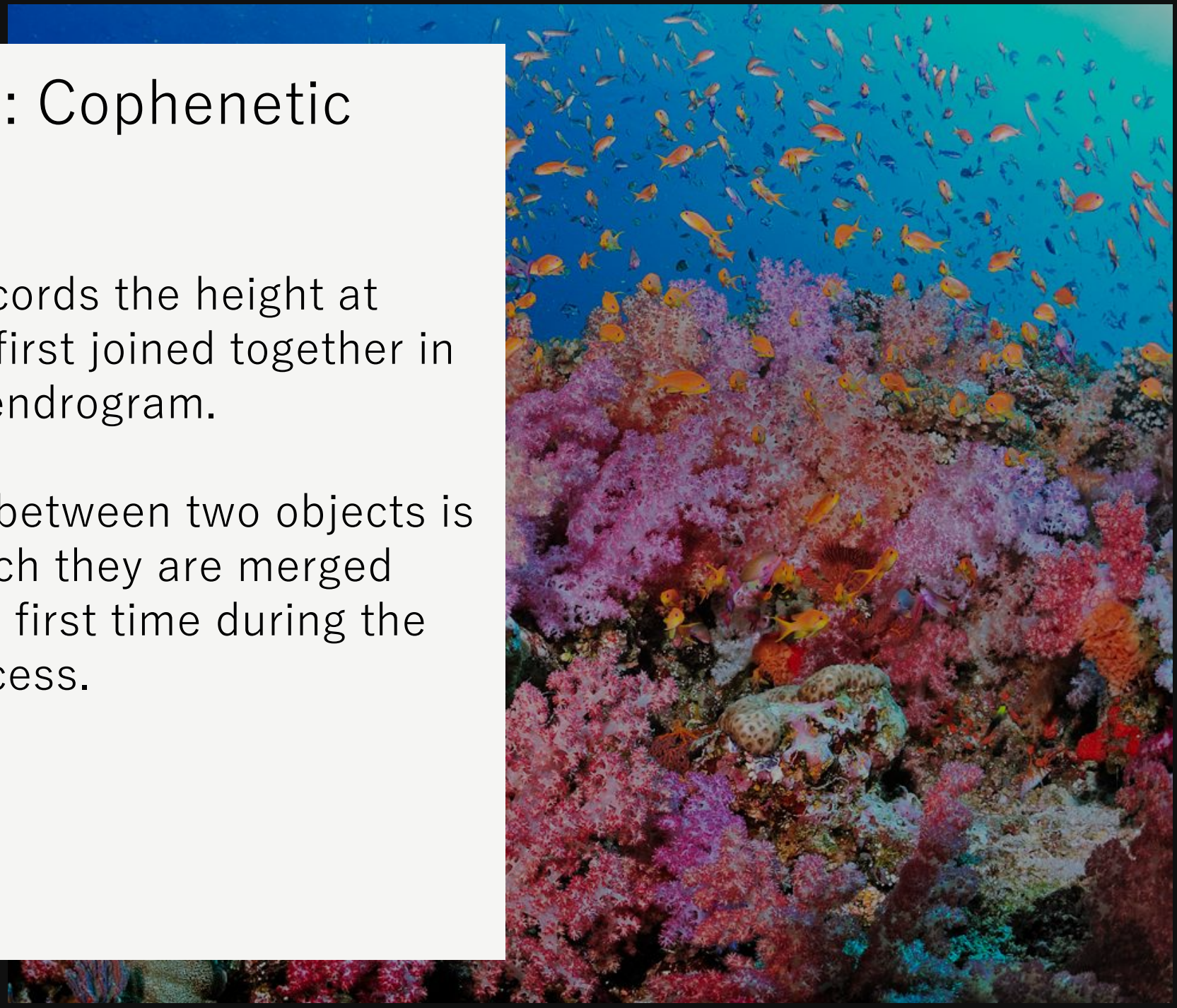
Clustering Statistics



Clustering Statistics: Cophenetic Matrix

The **cophenetic matrix** records the height at which pairs of objects are first joined together in a hierarchical clustering dendrogram.

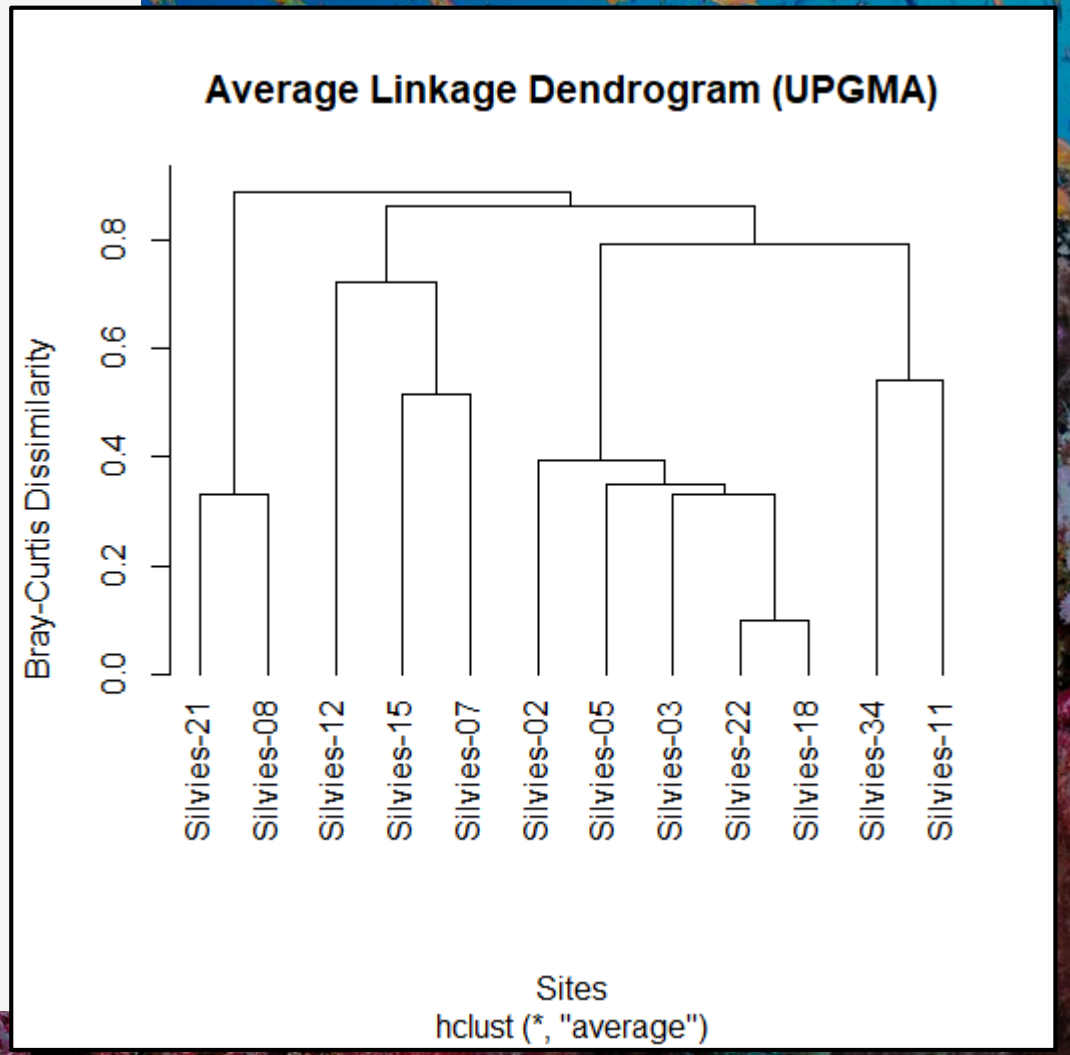
The **cophenetic distance** between two objects is the level (or height) at which they are merged into a single cluster for the first time during the hierarchical clustering process.



Clustering Statistics: Cophenetic Matrix

For each pair of objects (i, j), find the height of the dendrogram at which these objects are first merged into the same cluster.

This height is the **cophenetic distance** $d_c(i,j)$.

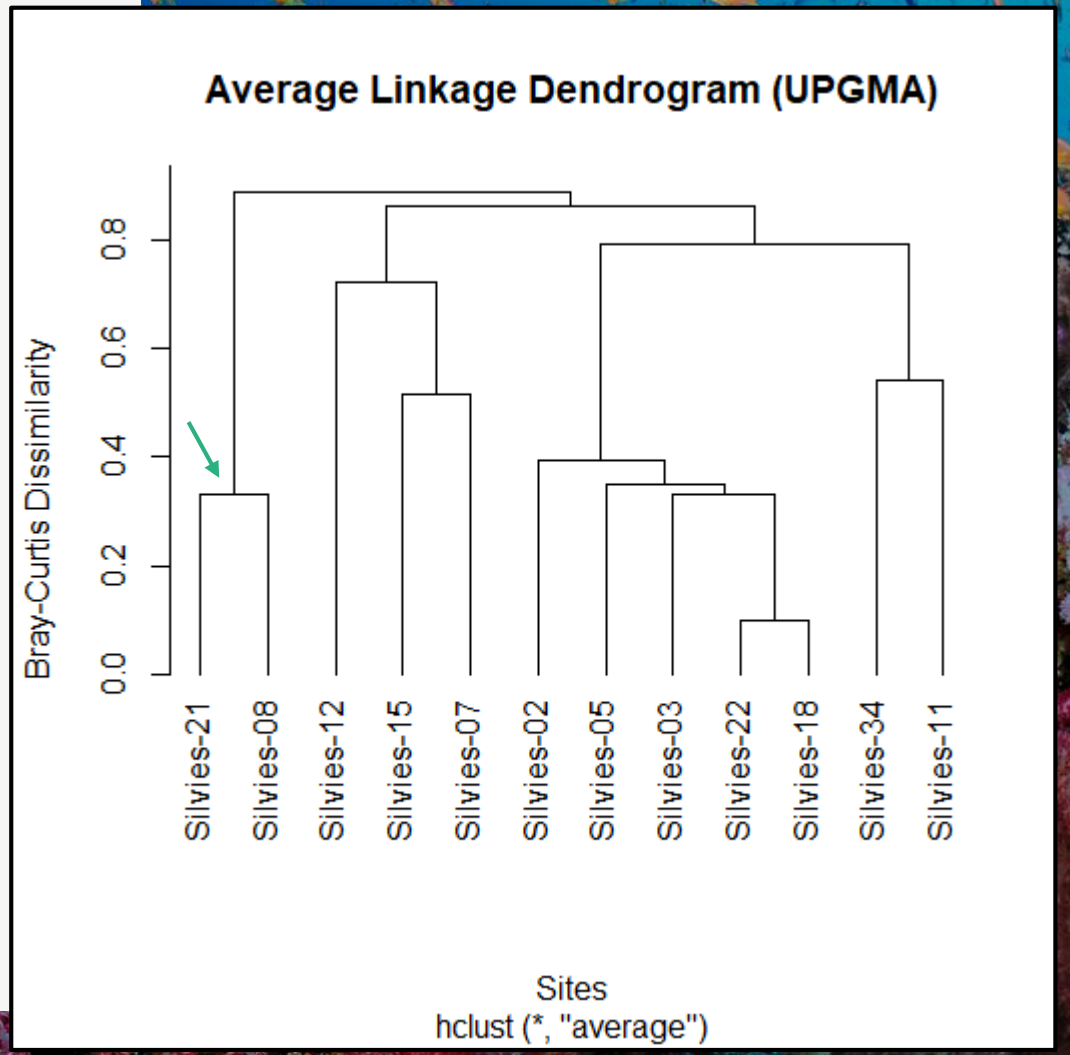


Clustering Statistics: Cophenetic Matrix

For each pair of objects (i, j), find the height of the dendrogram at which these objects are first merged into the same cluster.

This height is the **cophenetic distance** $d_c(i,j)$.

For example, d_c for S-21 and S-08 is 0.33—the same as this pair's Bray-Curtis distance…



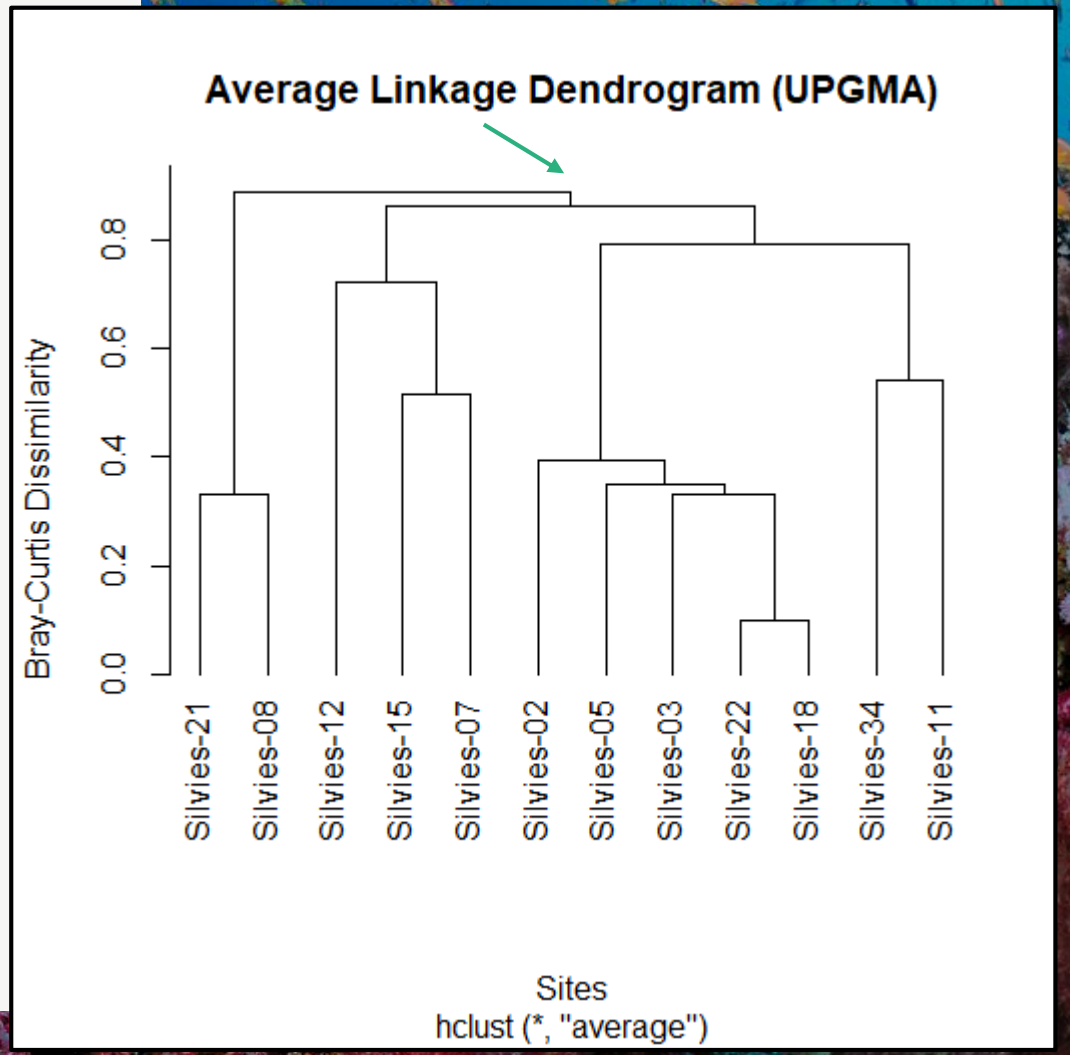
Clustering Statistics: Cophenetic Matrix

For each pair of objects (i, j), find the height of the dendrogram at which these objects are first merged into the same cluster.

This height is the **cophenetic distance** $d_c(i,j)$.

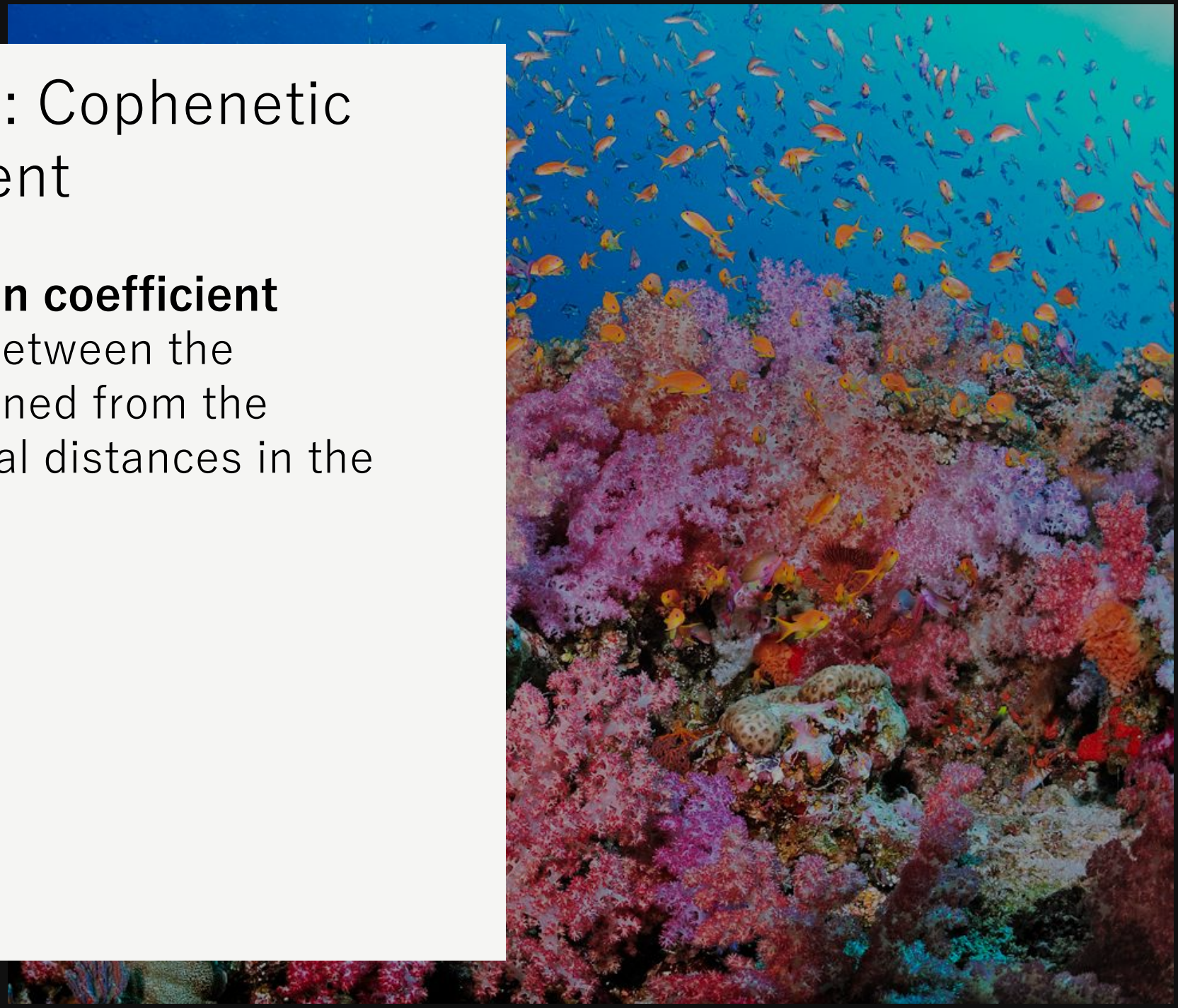
For example, d_c for S-21 and S-08 is 0.33—the same as this pair's Bray-Curtis distance…

BUT, d_c for S-21/S-08 and all other sites is 0.89—notably larger than some of the Bray-Curtis values!



Clustering Statistics: Cophenetic Correlation Coefficient

The **cophenetic correlation coefficient** measures the correlation between the cophenetic distances obtained from the dendrogram and the original distances in the distance matrix.

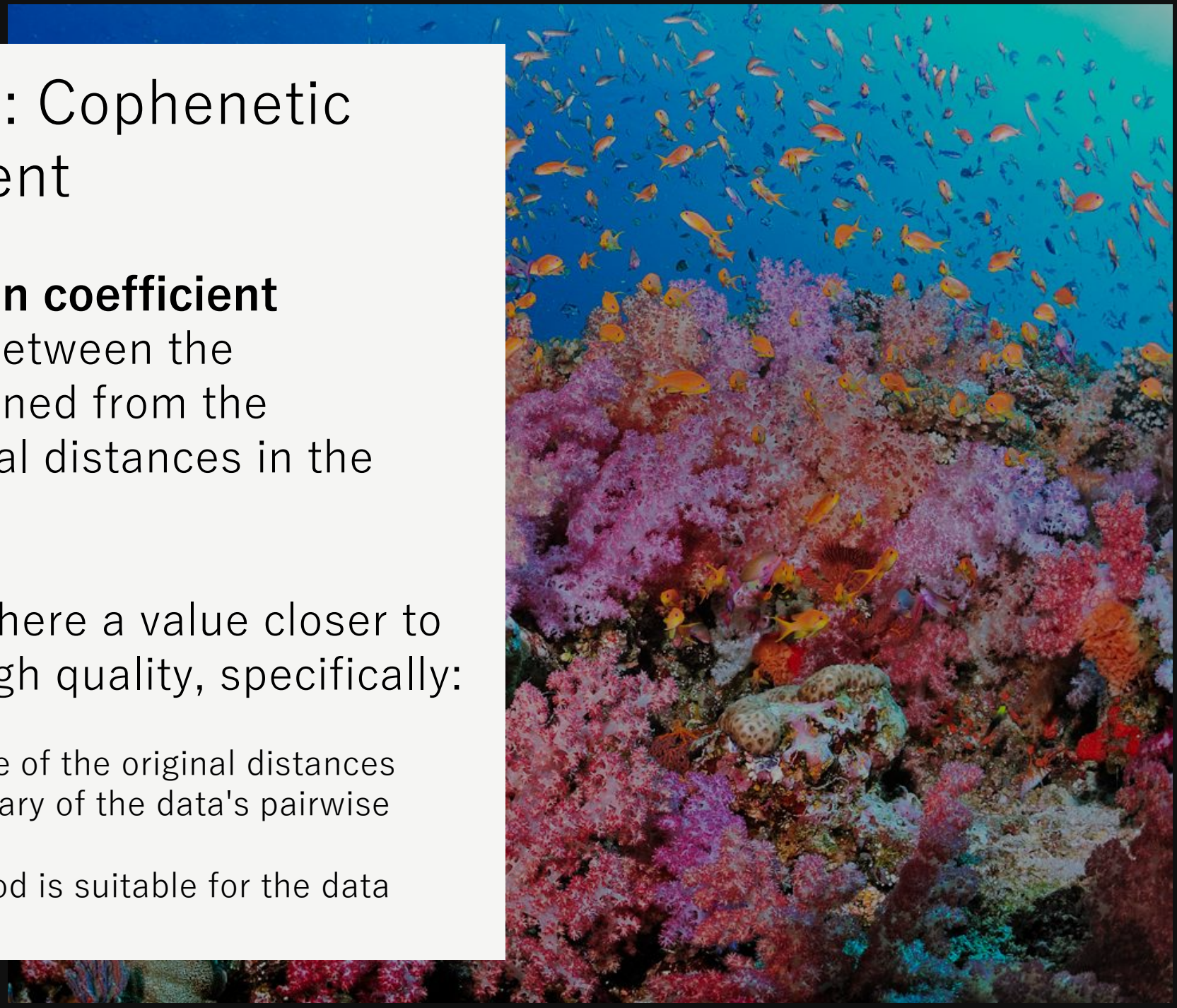


Clustering Statistics: Cophenetic Correlation Coefficient

The **cophenetic correlation coefficient** measures the correlation between the cophenetic distances obtained from the dendrogram and the original distances in the distance matrix.

Values range from 0 – 1, where a value closer to 1 indicates a solution of high quality, specifically:

- Clusters are highly representative of the original distances
- The dendrogram is a good summary of the data's pairwise relationships
- The hierarchical clustering method is suitable for the data

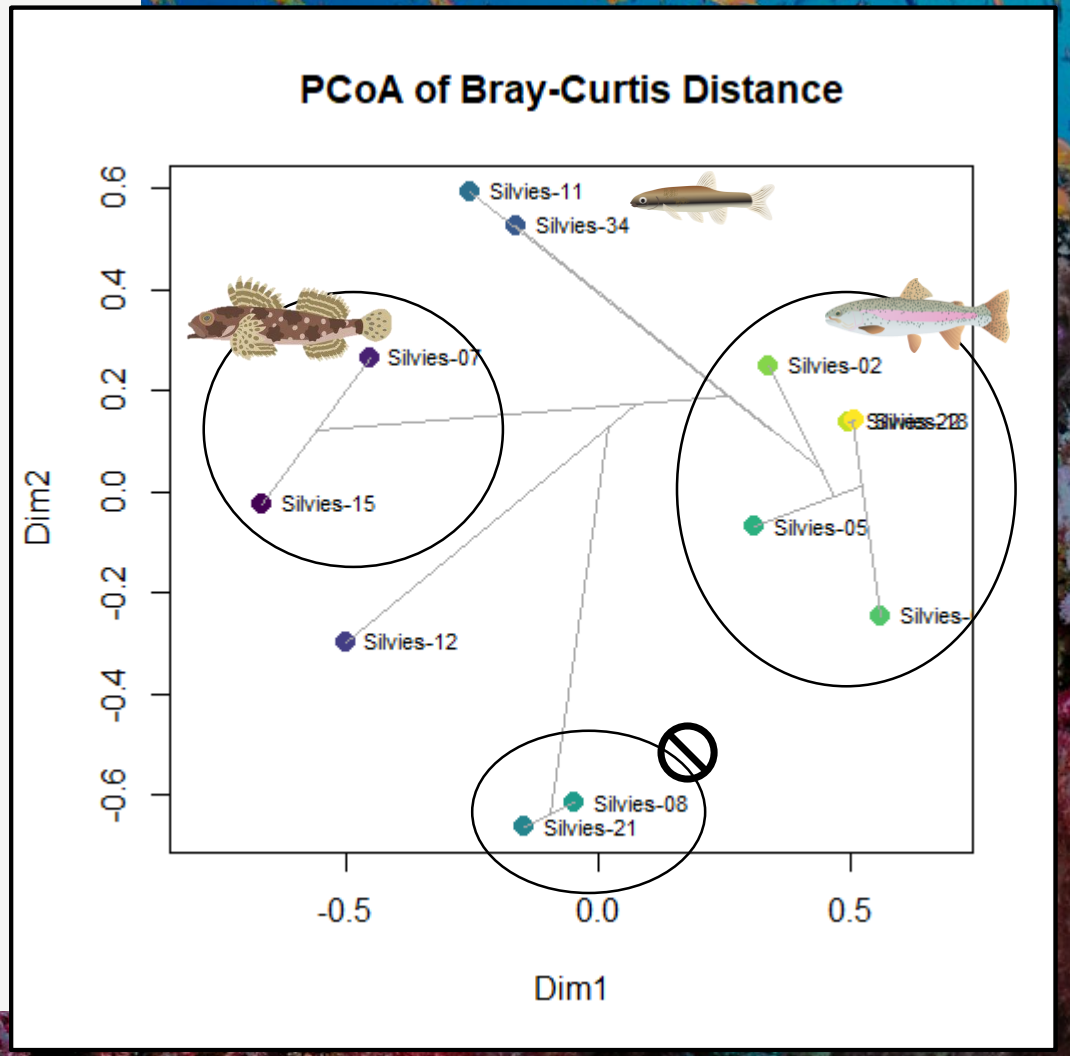


Clustering Statistics: Cophenetic Correlation Coefficient

The **cophenetic correlation coefficient** measures the correlation between the cophenetic distances obtained from the dendrogram and the original distances in the distance matrix.

Values range from 0 – 1, where a value closer to 1 indicates a solution of high quality.

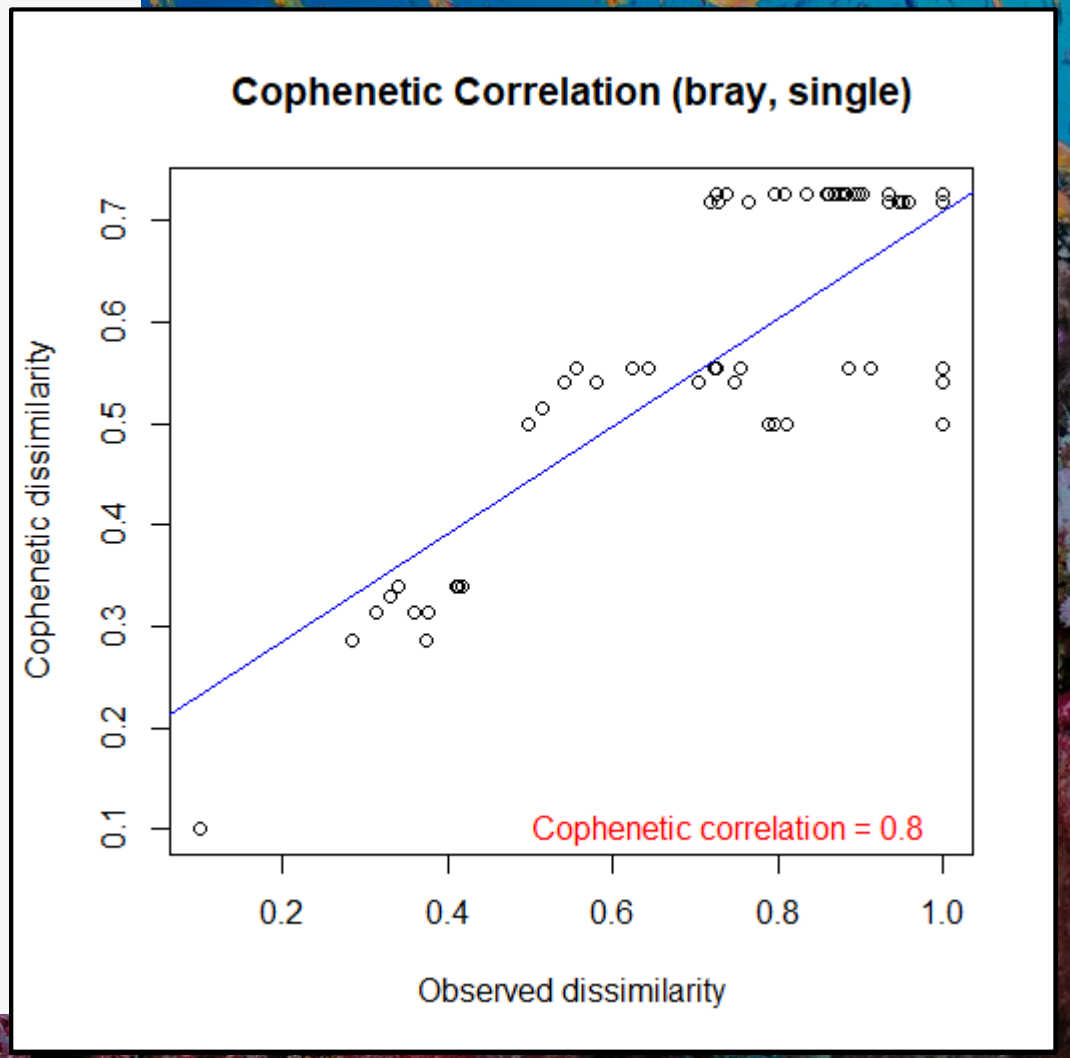
Single linkage cophenetic coefficient = 0.80



Clustering Statistics: Cophenetic Correlation Coefficient

The **cophenetic correlation coefficient** measures the correlation between the cophenetic distances obtained from the dendrogram and the original distances in the distance matrix.

Values range from 0 – 1, where a value closer to 1 indicates a solution of high quality.



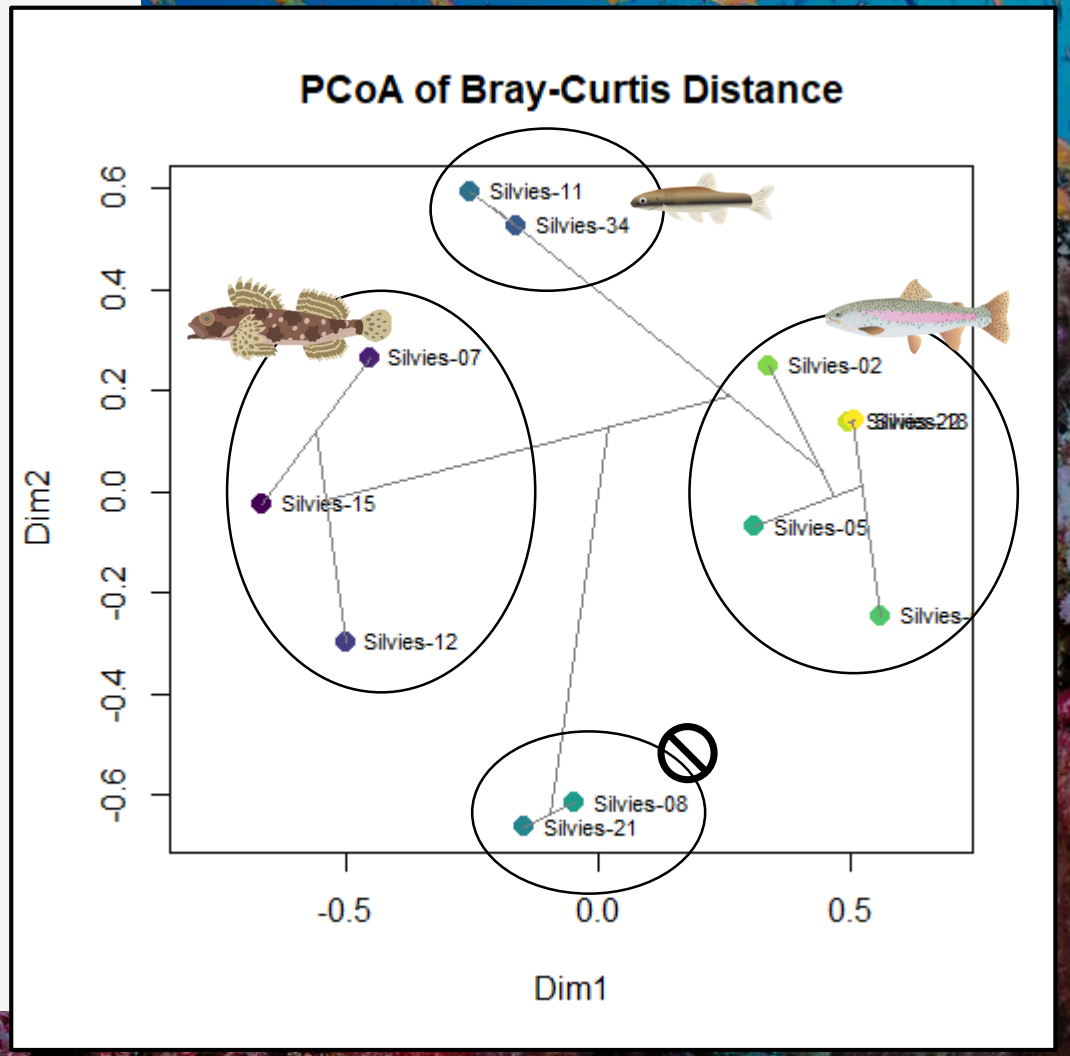
Single linkage cophenetic coefficient = 0.80

Clustering Statistics: Cophenetic Correlation Coefficient

The **cophenetic correlation coefficient** measures the correlation between the cophenetic distances obtained from the dendrogram and the original distances in the distance matrix.

Values range from 0 – 1, where a value closer to 1 indicates a solution of high quality.

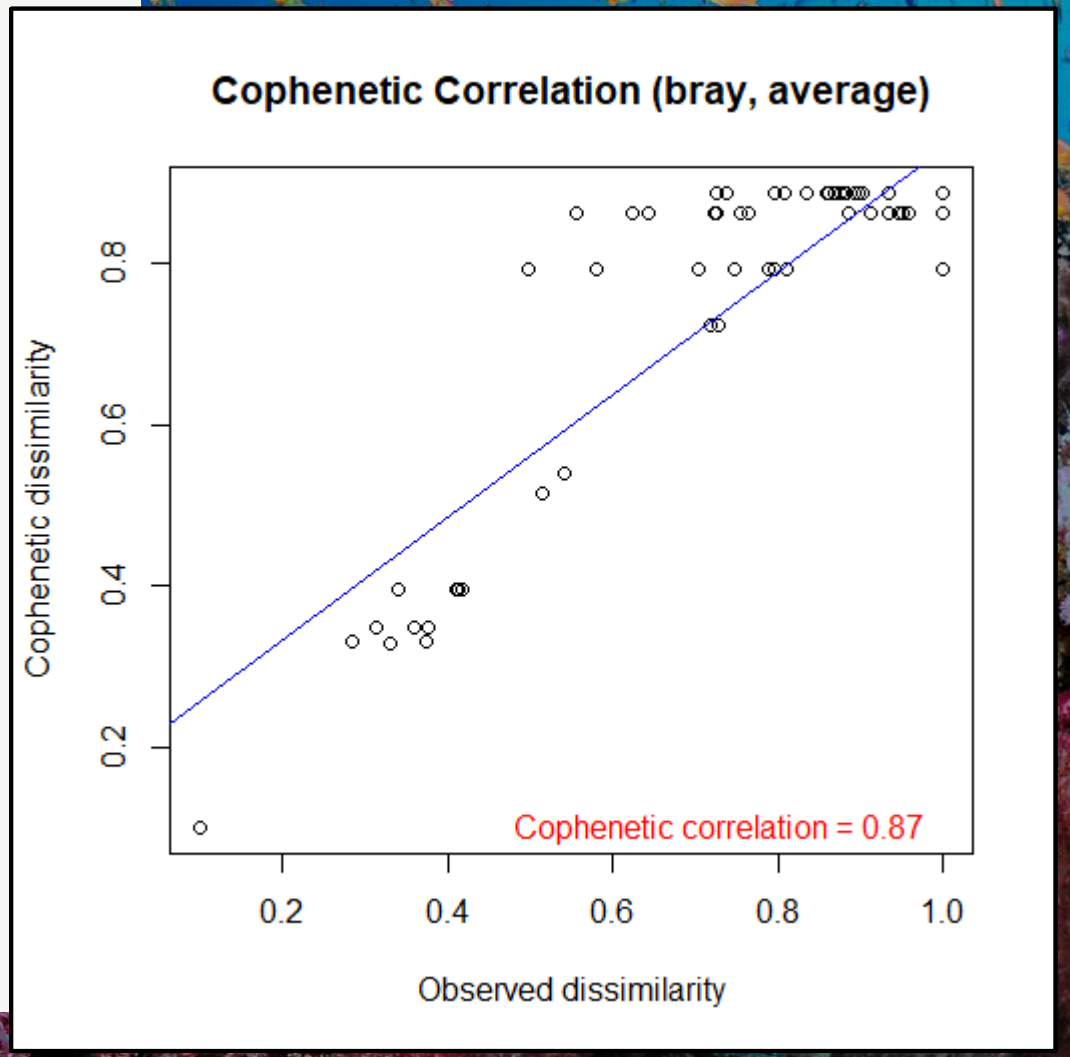
Average linkage cophenetic coefficient = 0.87



Clustering Statistics: Cophenetic Correlation Coefficient

The **cophenetic correlation coefficient** measures the correlation between the cophenetic distances obtained from the dendrogram and the original distances in the distance matrix.

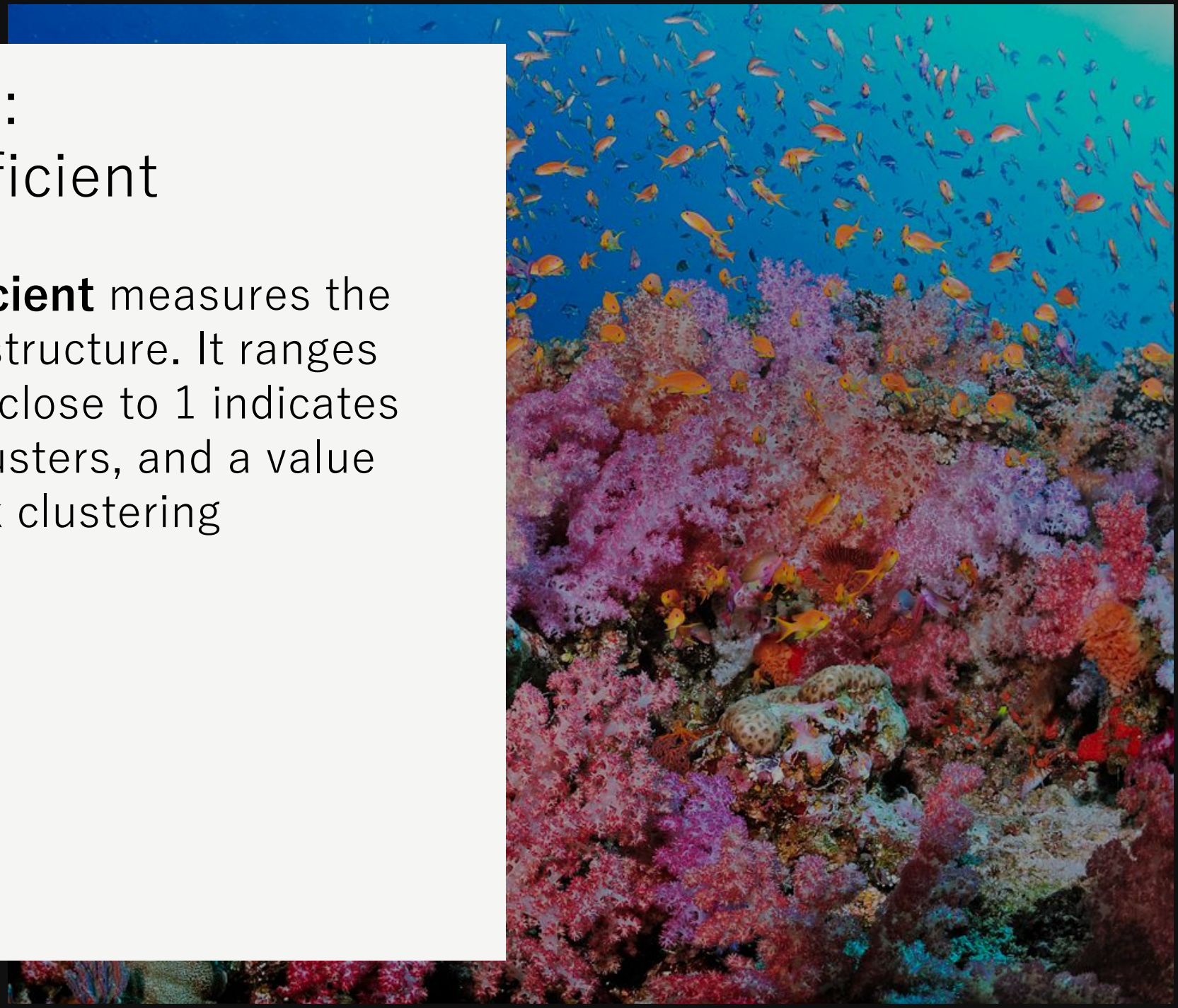
Values range from 0 – 1, where a value closer to 1 indicates a solution of high quality.



Average linkage cophenetic coefficient = 0.87

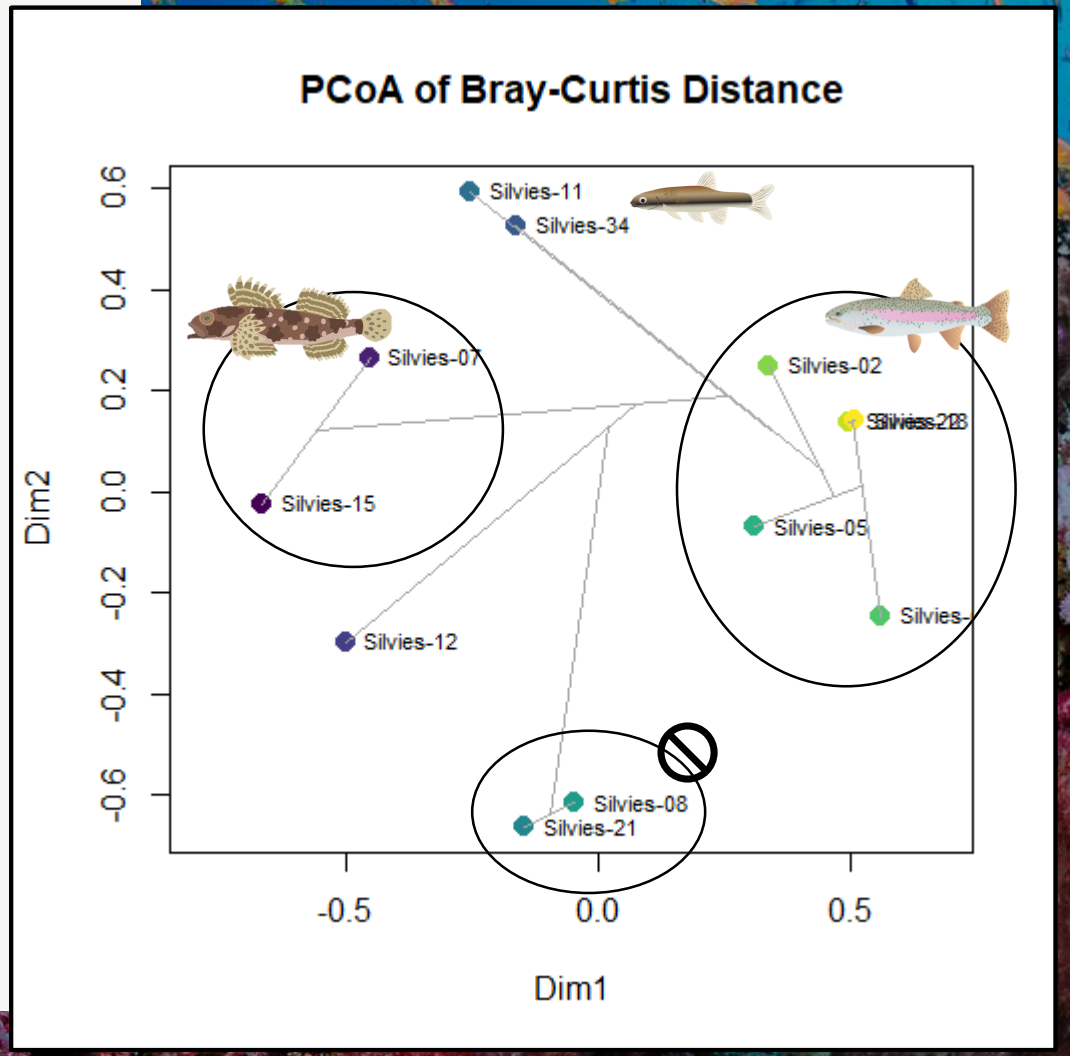
Clustering Statistics: Agglomerative Coefficient

The **agglomerative coefficient** measures the strength of the clustering structure. It ranges from 0 to 1, where a value close to 1 indicates distinct, well-separated clusters, and a value close to 0 suggests a weak clustering structure.



Clustering Statistics: Agglomerative Coefficient

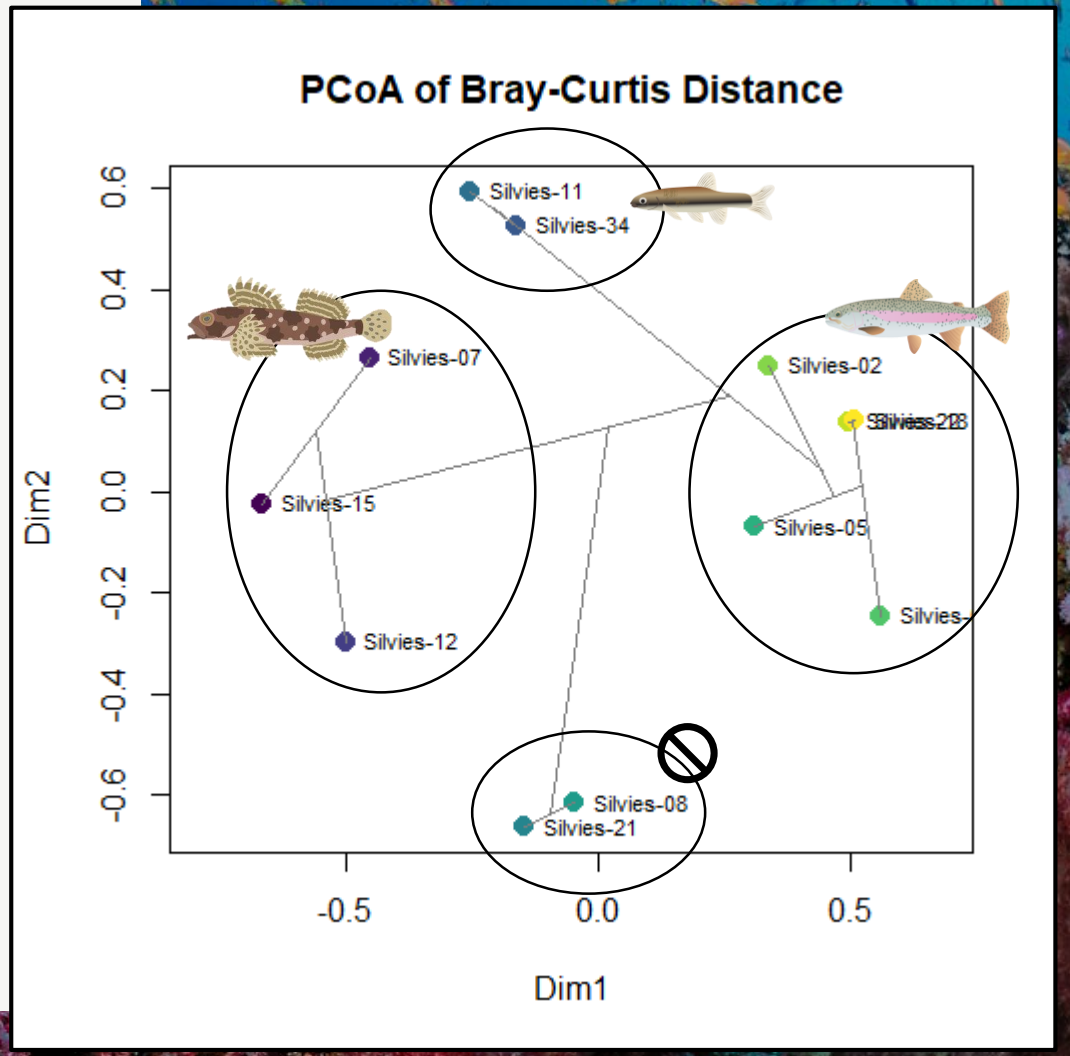
The **agglomerative coefficient** measures the strength of the clustering structure. It ranges from 0 to 1, where a value close to 1 indicates distinct, well-separated clusters, and a value close to 0 suggests a weak clustering structure.



Single linkage agglomerative coefficient = 0.47

Clustering Statistics: Agglomerative Coefficient

The **agglomerative coefficient** measures the strength of the clustering structure. It ranges from 0 to 1, where a value close to 1 indicates distinct, well-separated clusters, and a value close to 0 suggests a weak clustering structure.

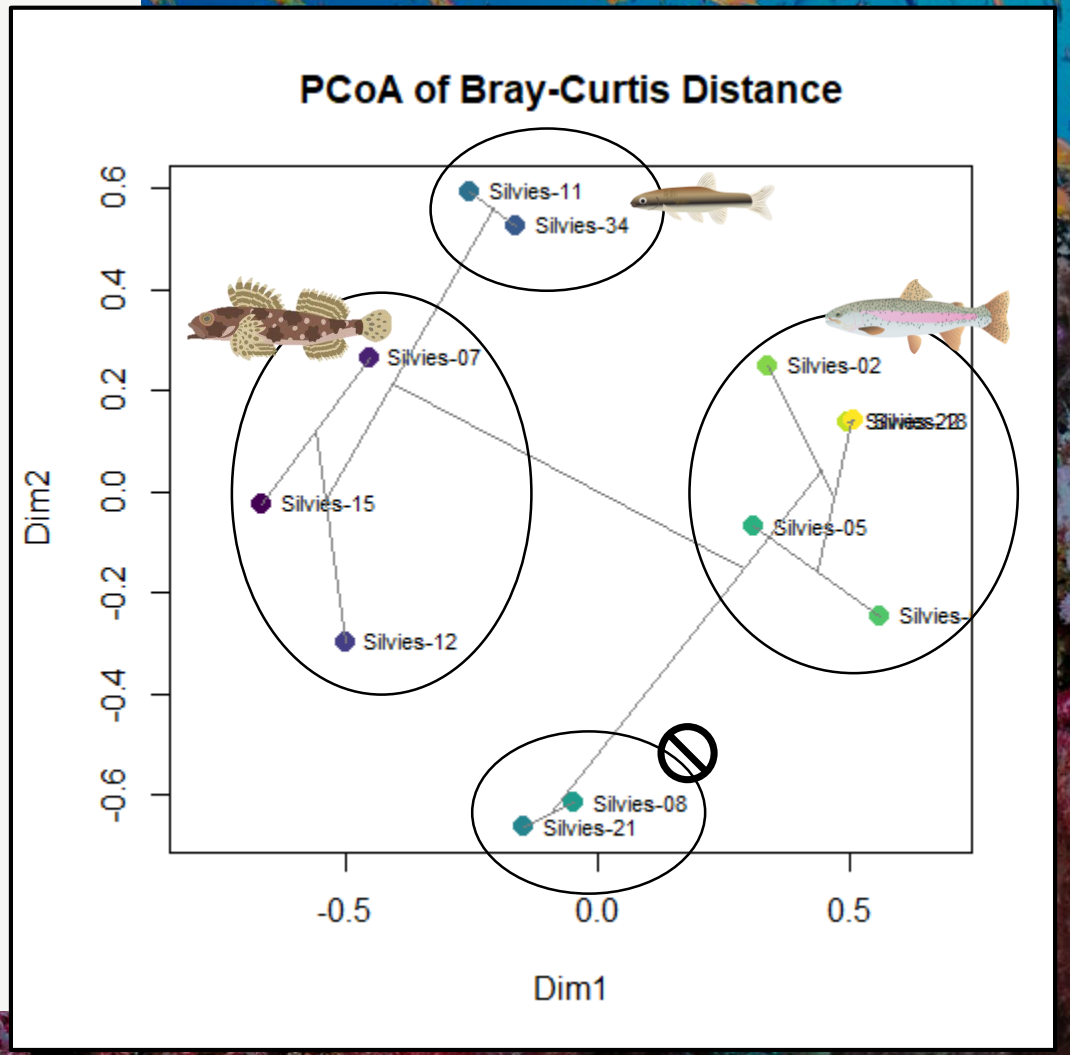


Average linkage agglomerative coefficient = 0.56

Clustering Statistics: Agglomerative Coefficient

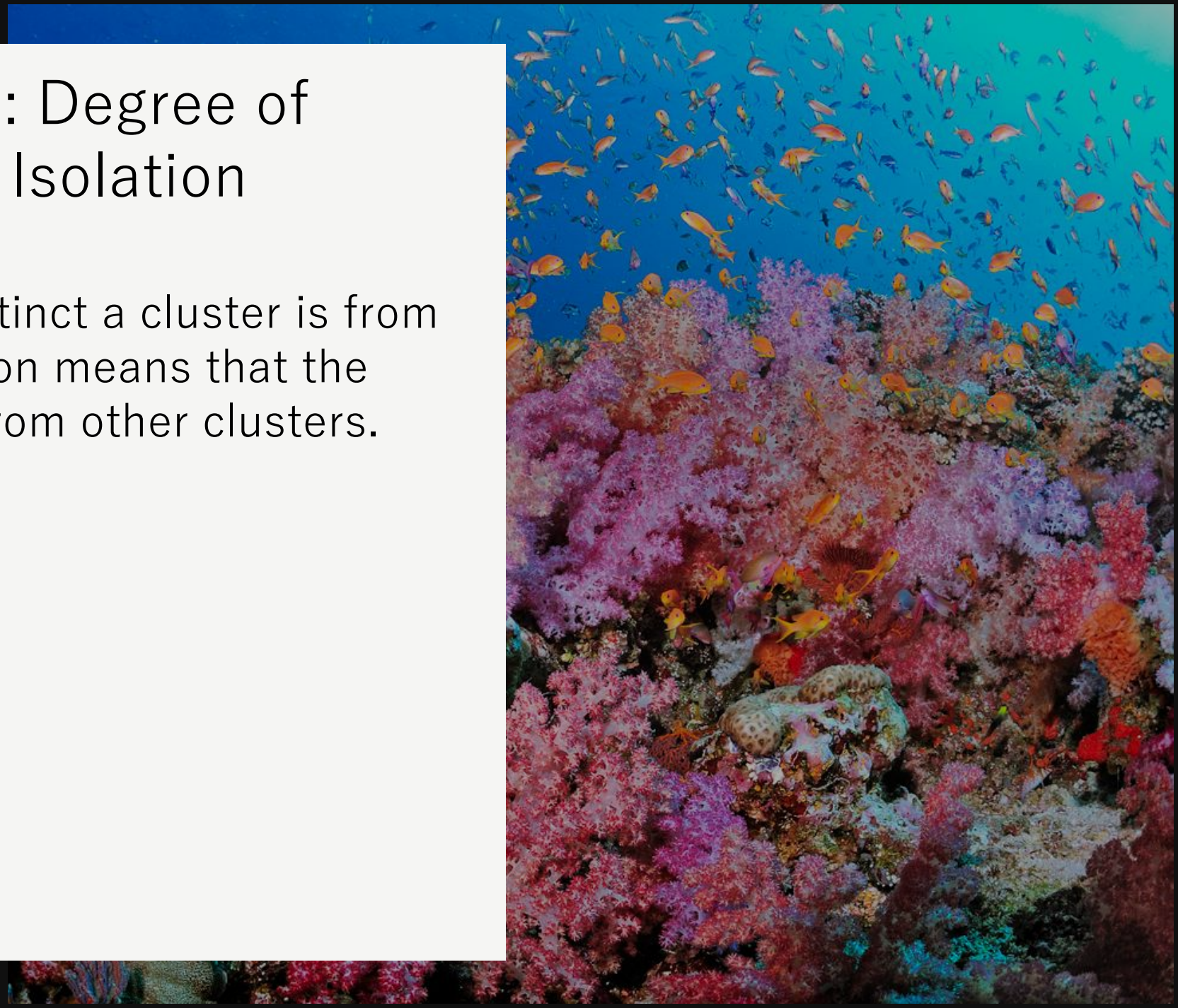
The **agglomerative coefficient** measures the strength of the clustering structure. It ranges from 0 to 1, where a value close to 1 indicates distinct, well-separated clusters, and a value close to 0 suggests a weak clustering structure.

Ward's agglomerative coefficient = 0.72



Clustering Statistics: Degree of Connectedness and Isolation

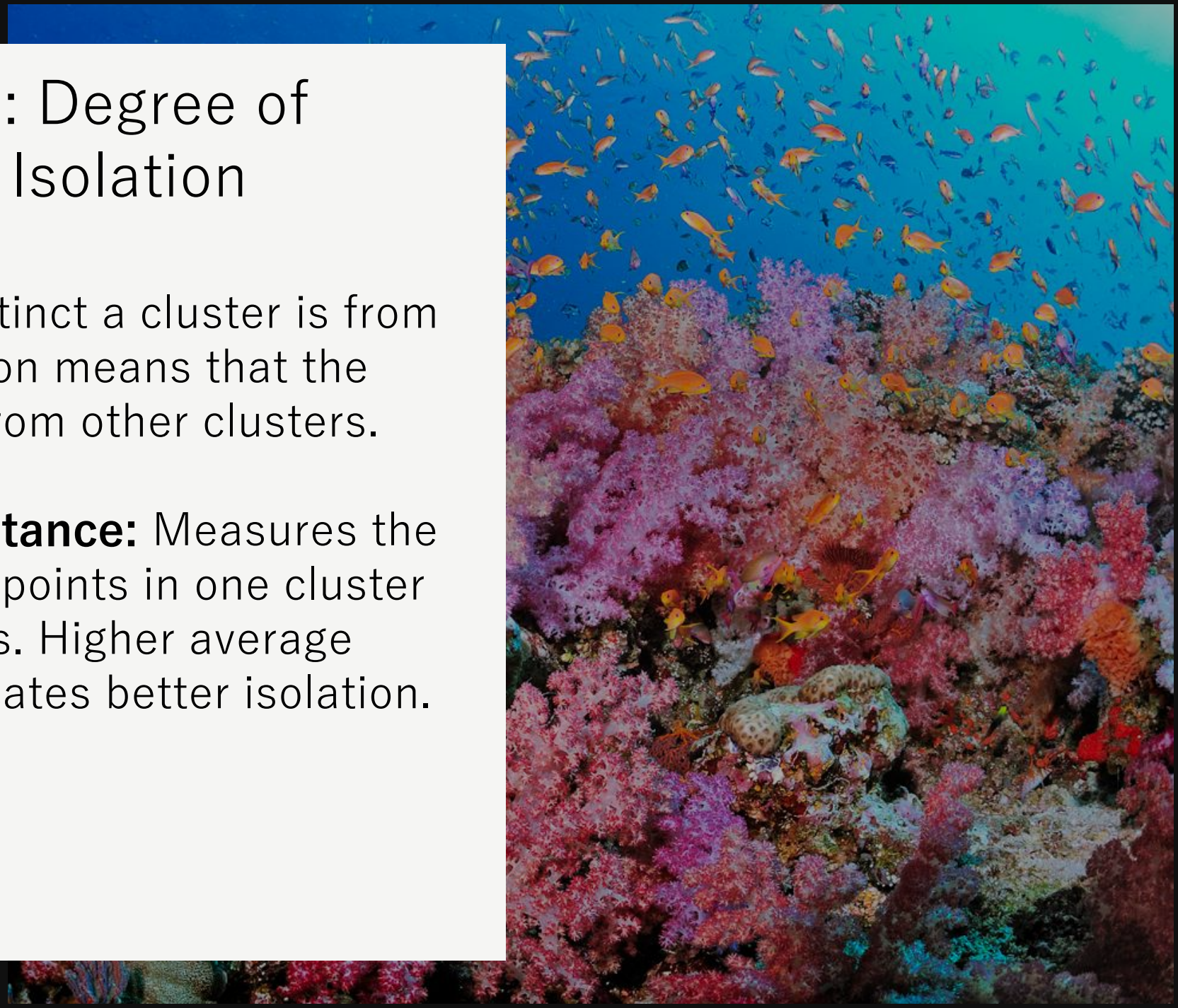
Isolation refers to how distinct a cluster is from other clusters. High isolation means that the cluster is well-separated from other clusters.



Clustering Statistics: Degree of Connectedness and Isolation

Isolation refers to how distinct a cluster is from other clusters. High isolation means that the cluster is well-separated from other clusters.

Average Inter-Cluster Distance: Measures the average distance between points in one cluster and points in other clusters. Higher average inter-cluster distance indicates better isolation.

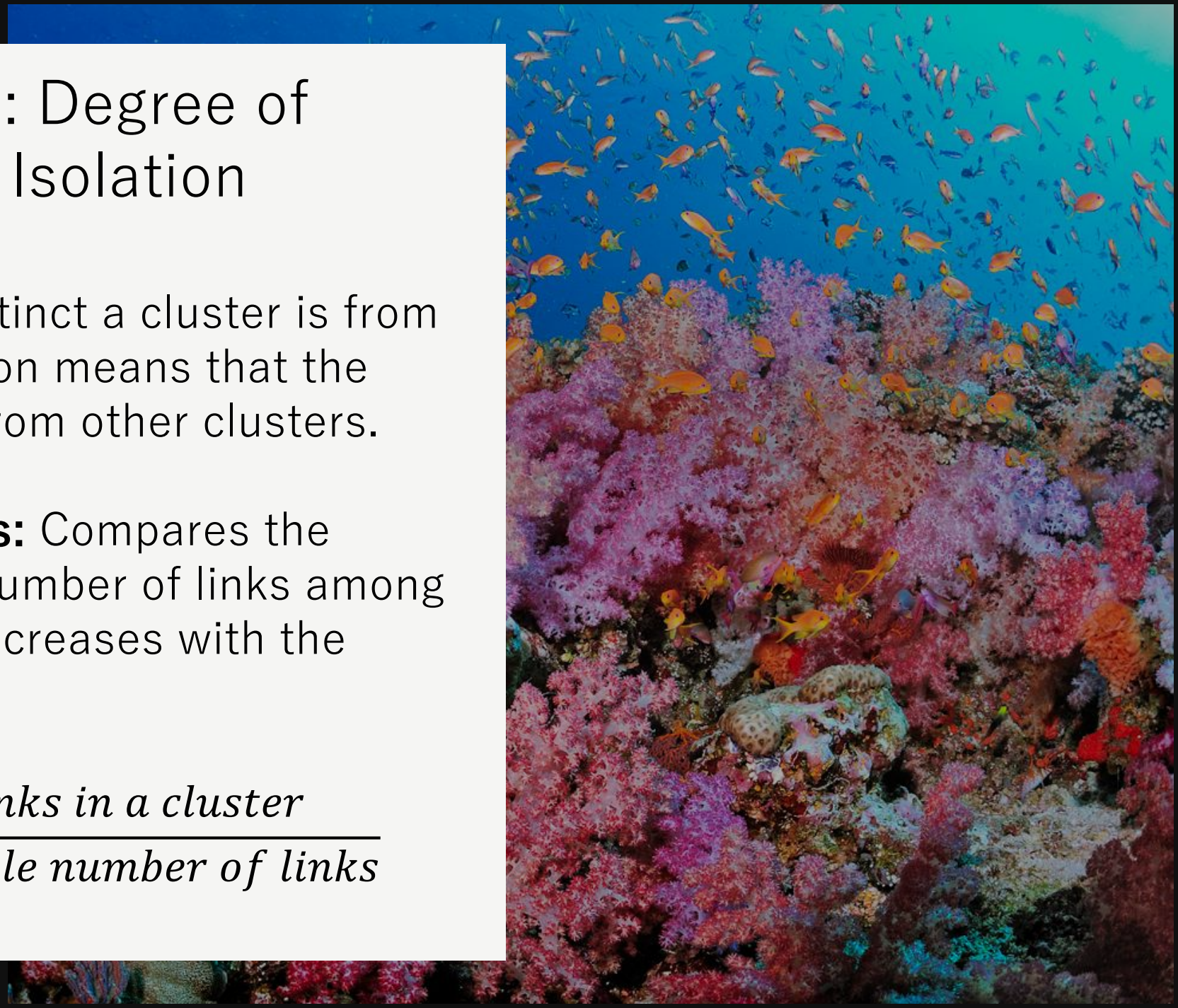


Clustering Statistics: Degree of Connectedness and Isolation

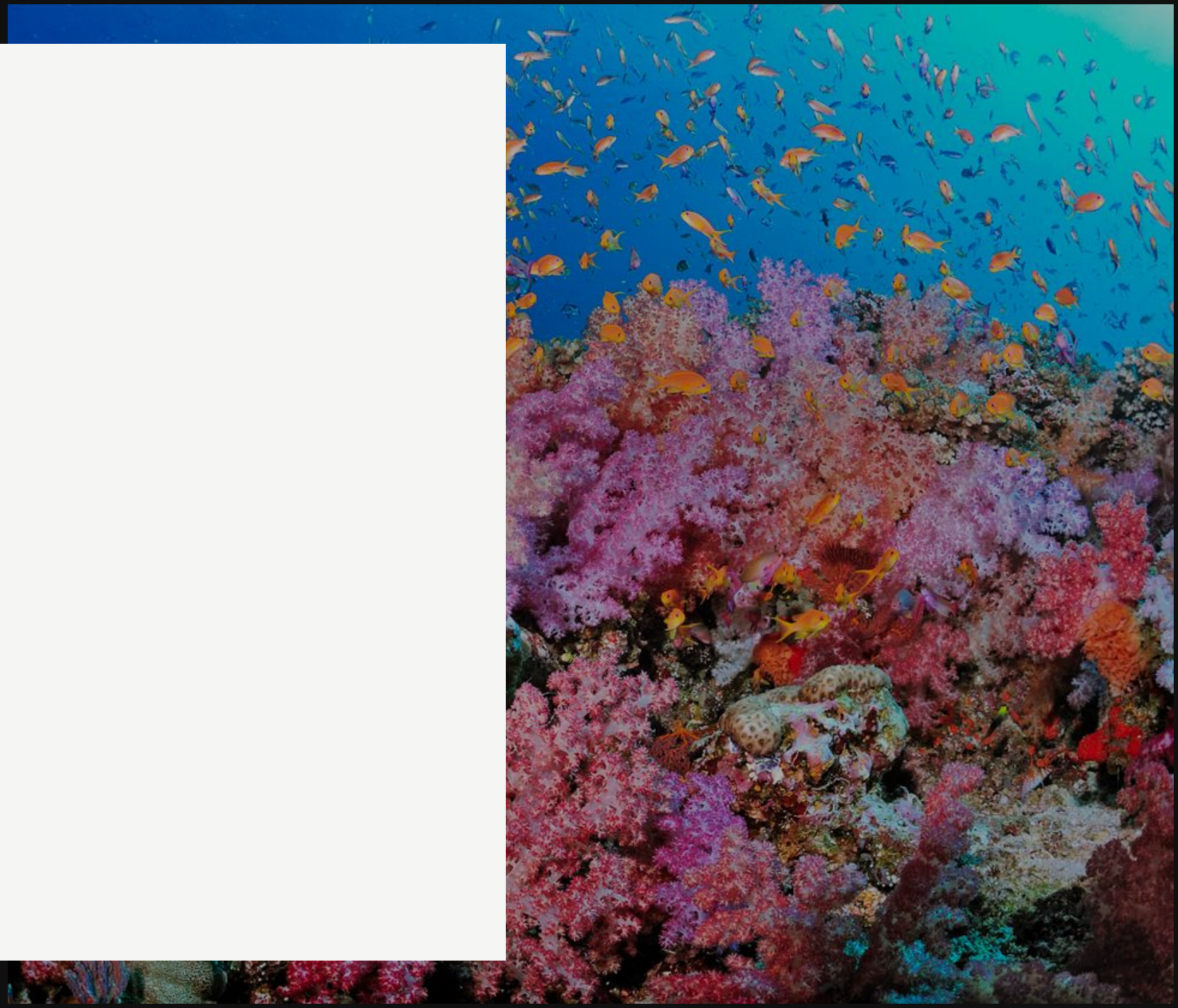
Isolation refers to how distinct a cluster is from other clusters. High isolation means that the cluster is well-separated from other clusters.

Degree of Connectedness: Compares the number of objects to the number of links among them, where link density increases with the degree of connectedness.

$$Co = \frac{number\ of\ links\ in\ a\ cluster}{maximum\ possible\ number\ of\ links}$$

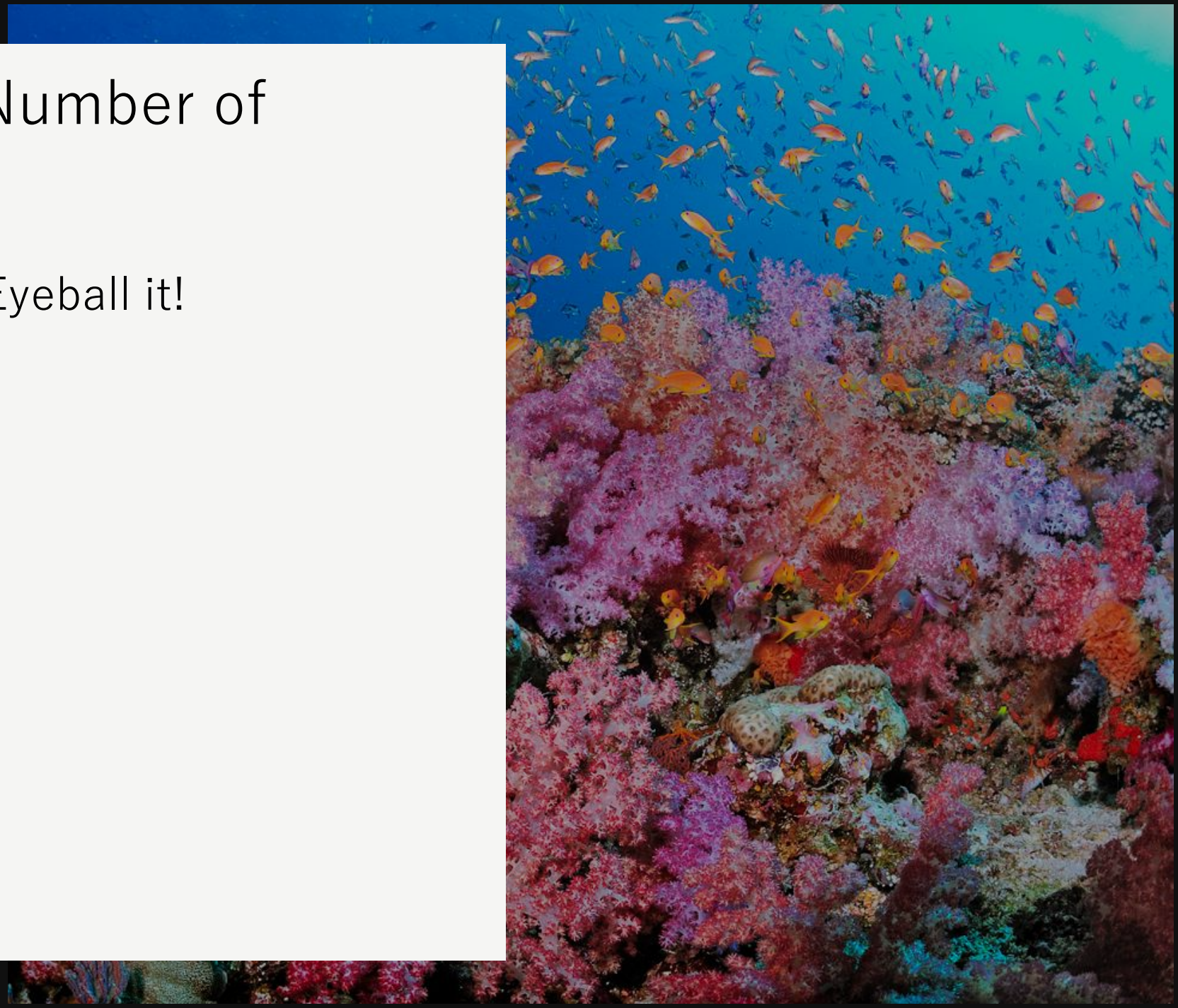


Cluster Validation



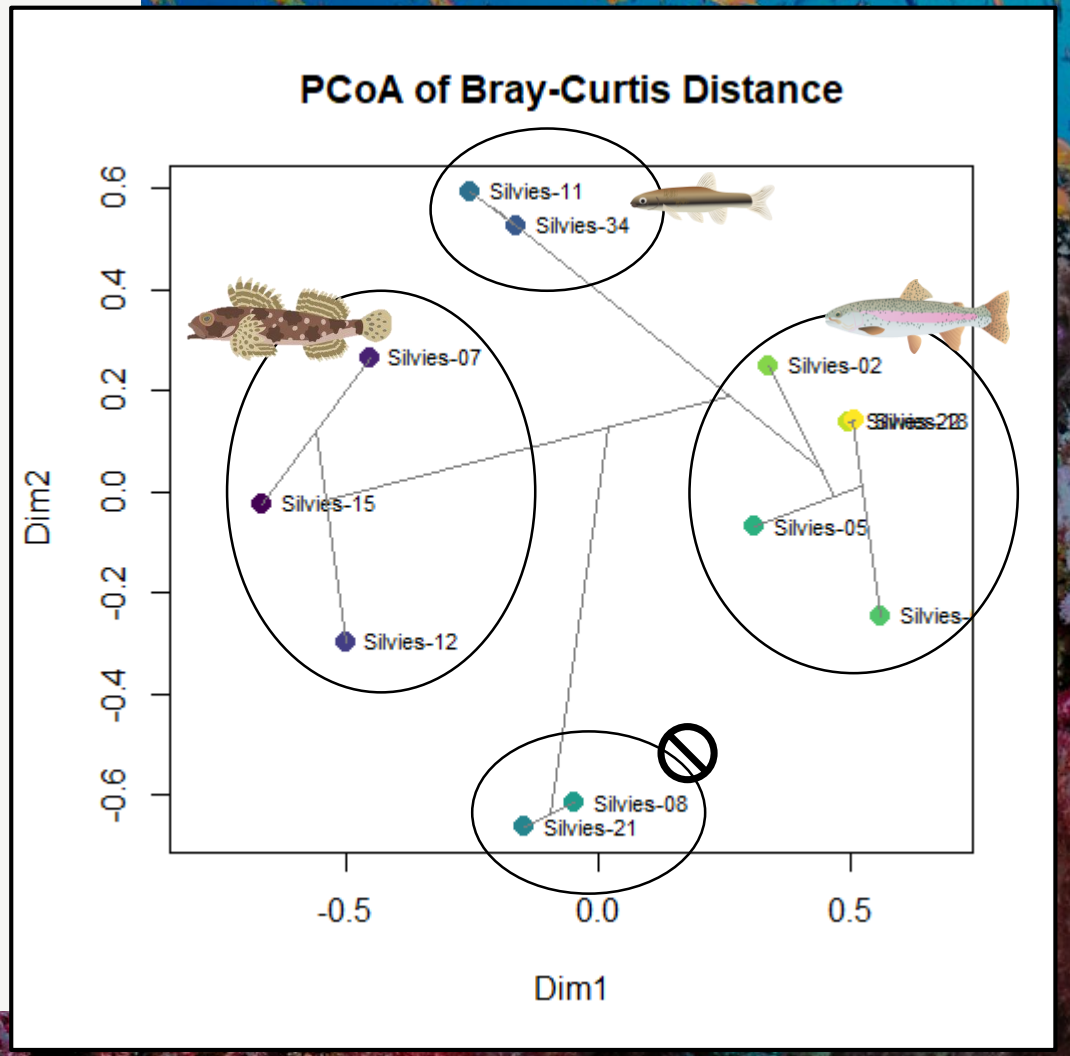
Cluster Validation: Number of Clusters

Nonstatistical methods: Eyeball it!



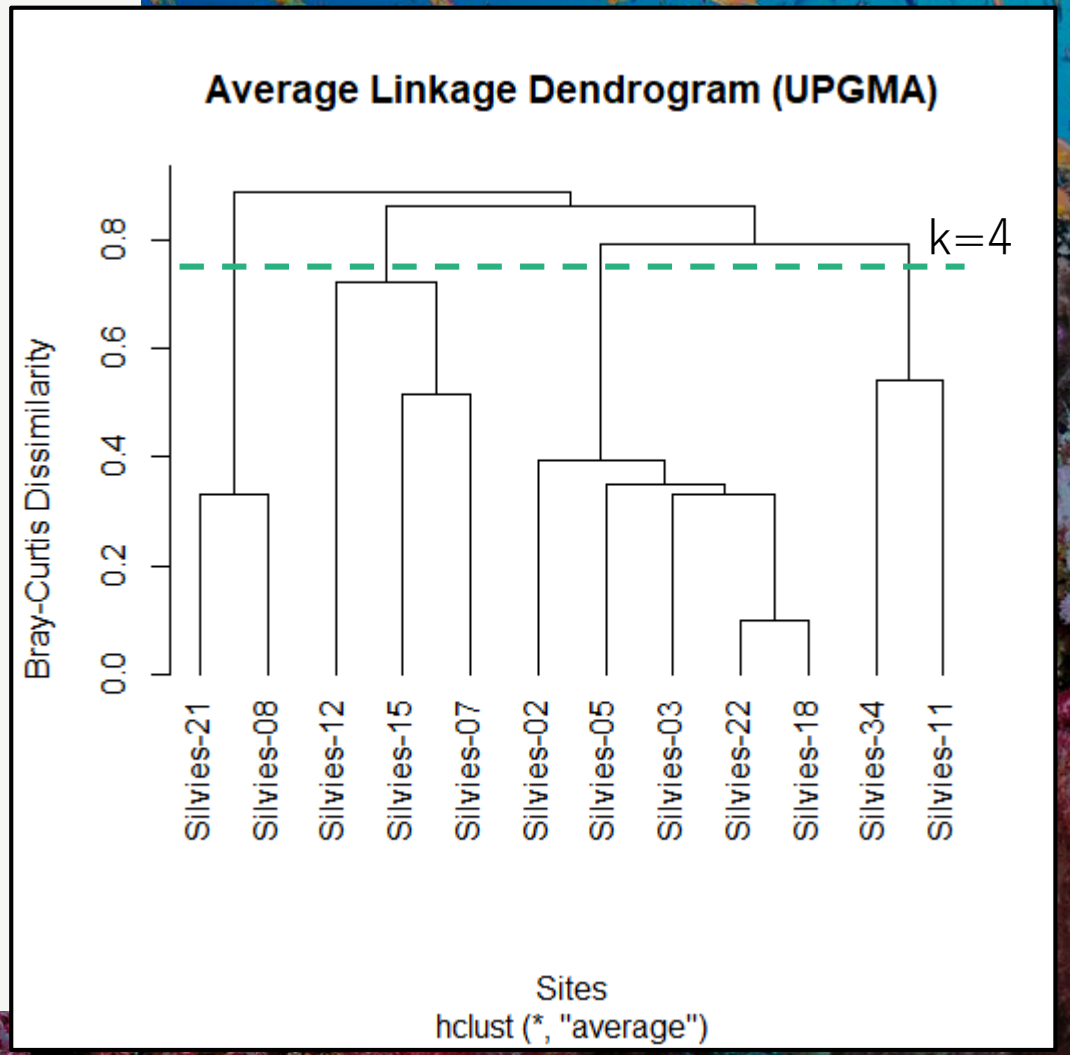
Cluster Validation: Number of Clusters

Nonstatistical methods: Eyeball it!



Cluster Validation: Number of Clusters

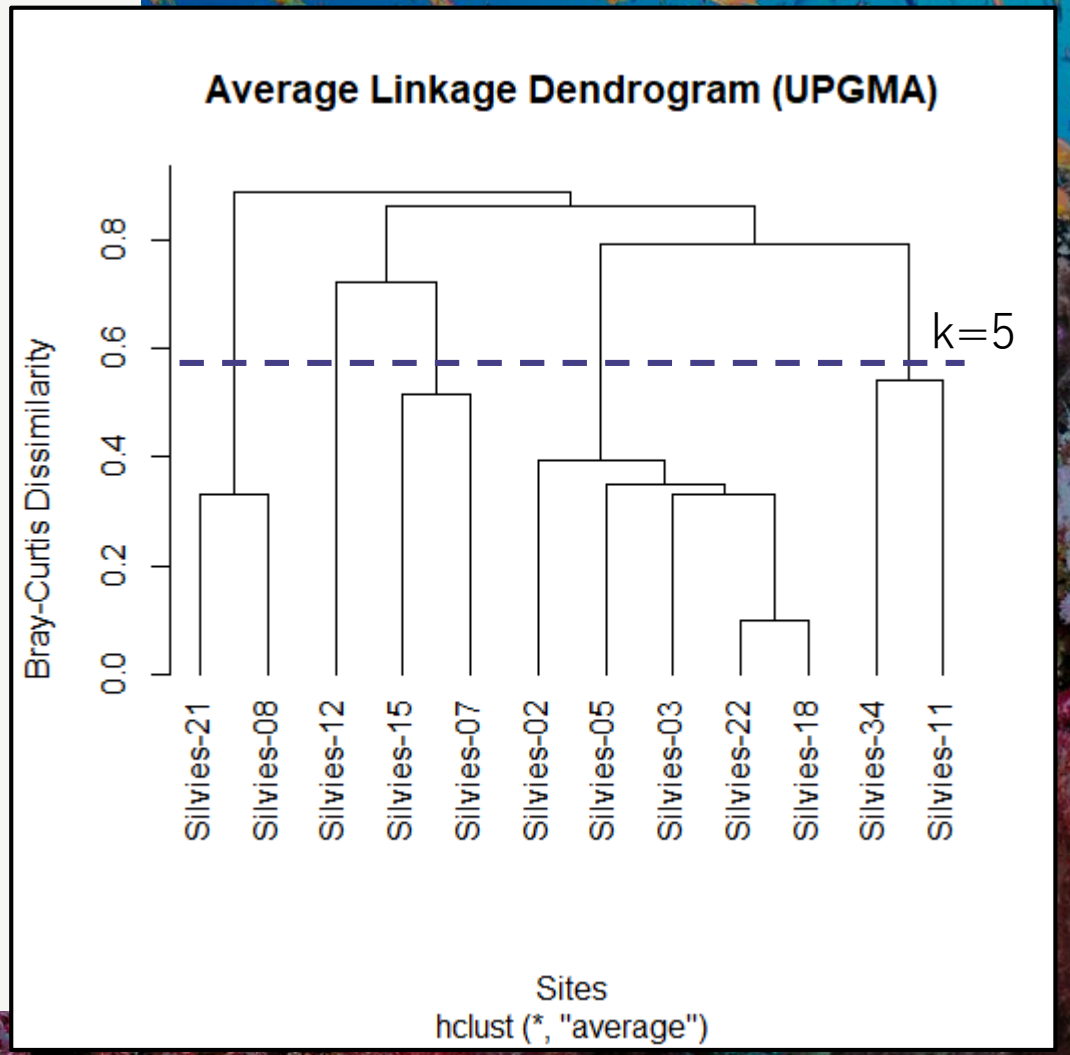
Nonstatistical methods: Eyeball it!



Cluster Validation: Number of Clusters

Nonstatistical methods: Eyeball it!

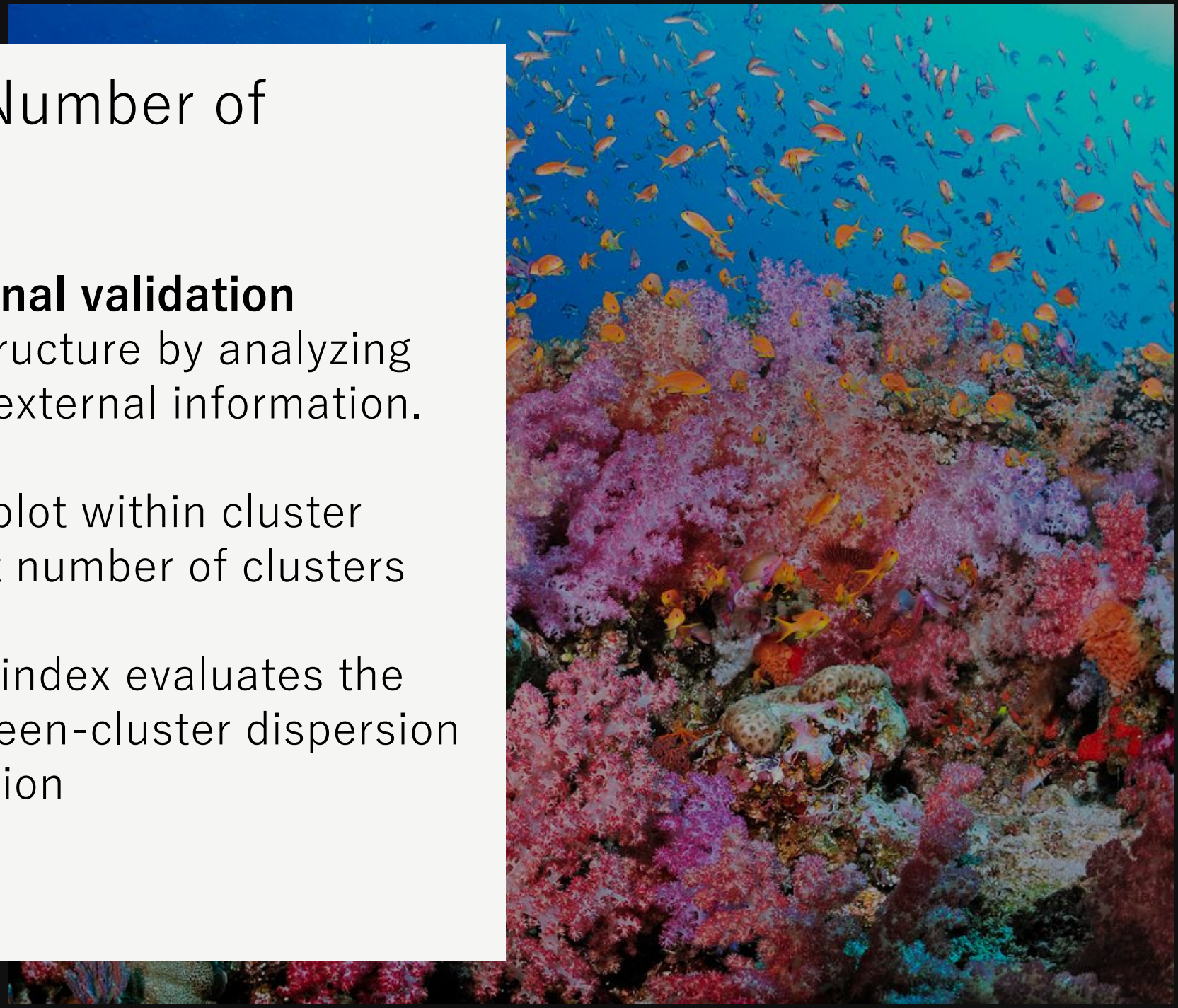
Avoid “orphan” clusters if possible



Cluster Validation: Number of Clusters

Statistical methods: Internal validation evaluates the clustering structure by analyzing the dataset itself, without external information.

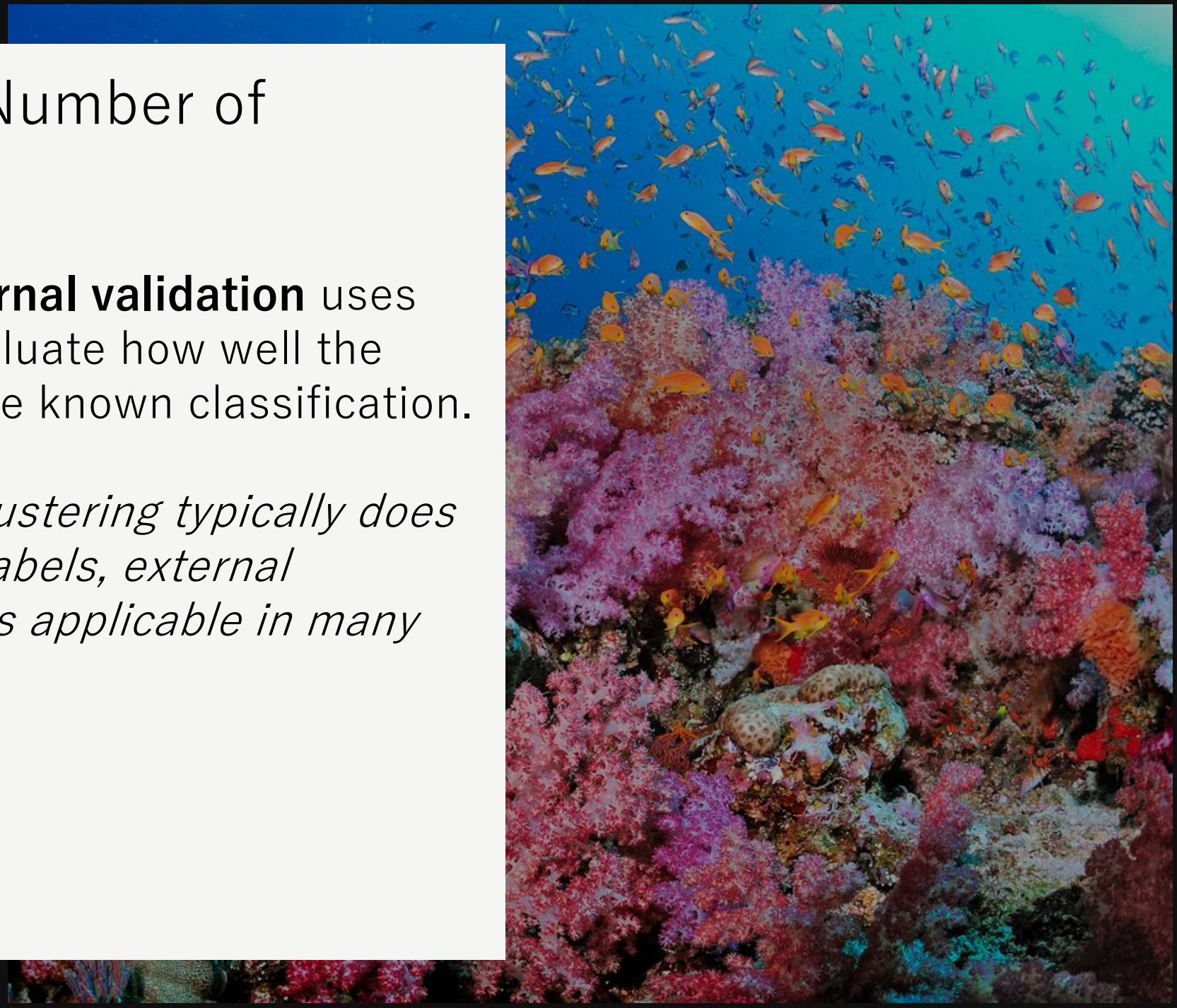
- The “**Elbow method**” – plot within cluster sums of squares against number of clusters
- The **Calinski-Harabasz** index evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion



Cluster Validation: Number of Clusters

Statistical methods: External validation uses external information to evaluate how well the clustering results match the known classification.

Note: Since hierarchical clustering typically does not start with predefined labels, external validation methods are less applicable in many cases.

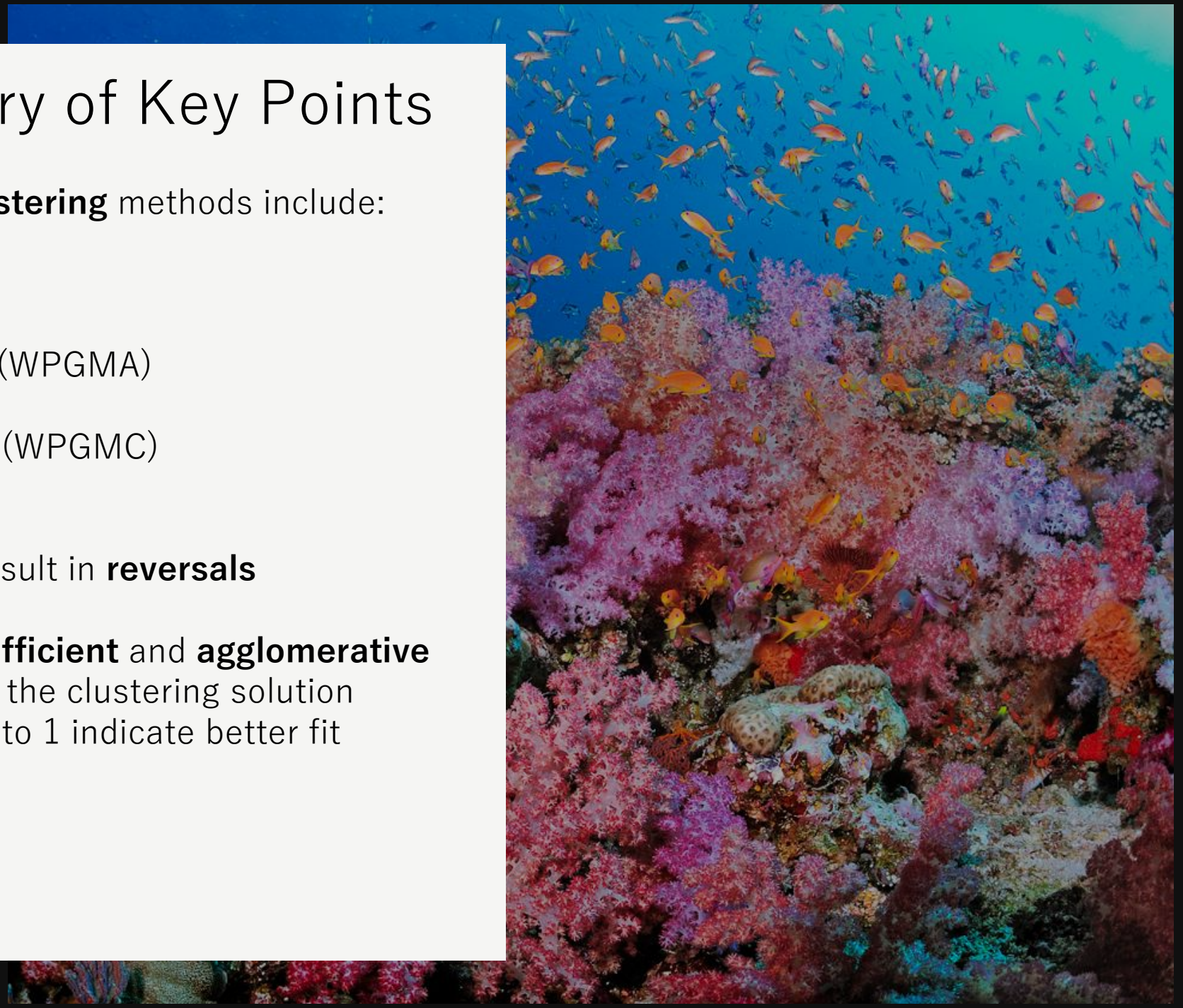


Conclusion: Summary of Key Points

- **Agglomerative hierarchical clustering** methods include:
  - **Single linkage**
  - **Complete linkage**
  - **Average linkage** (UPGMA)
  - **Weighted average linkage** (WPGMA)
  - **Centroid linkage** (UPGMC)
  - **Weighted centroid linkage** (WPGMC)
  - **Ward's minimum variance**

- Centroid linkage methods can result in **reversals**

- The **cophenetic correlation coefficient** and **agglomerative coefficient** are used to evaluate the clustering solution
  - In both cases, values closer to 1 indicate better fit



Questions?