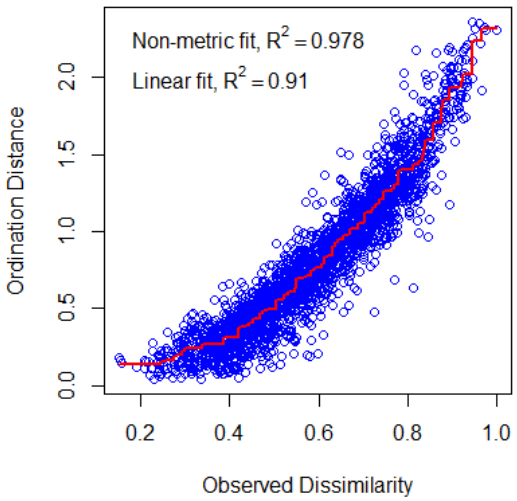
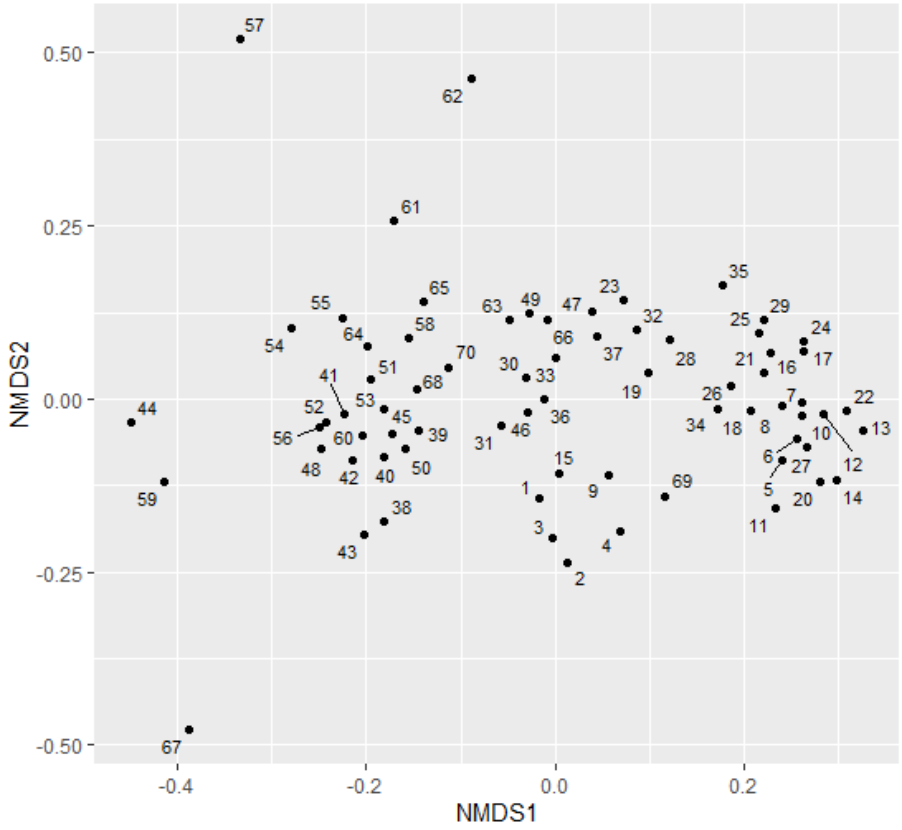
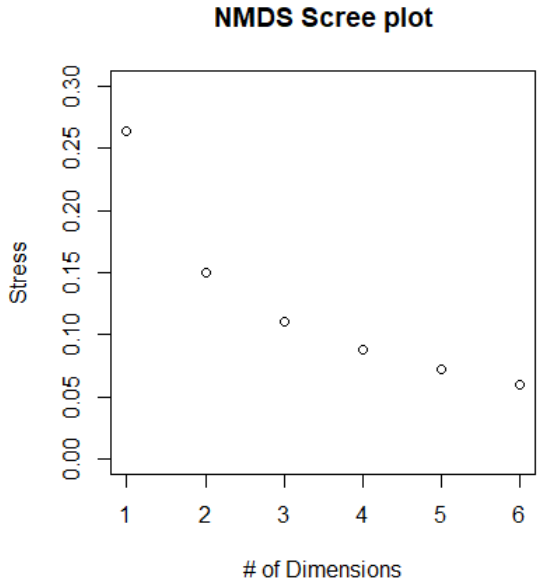
NMDS initial outputs



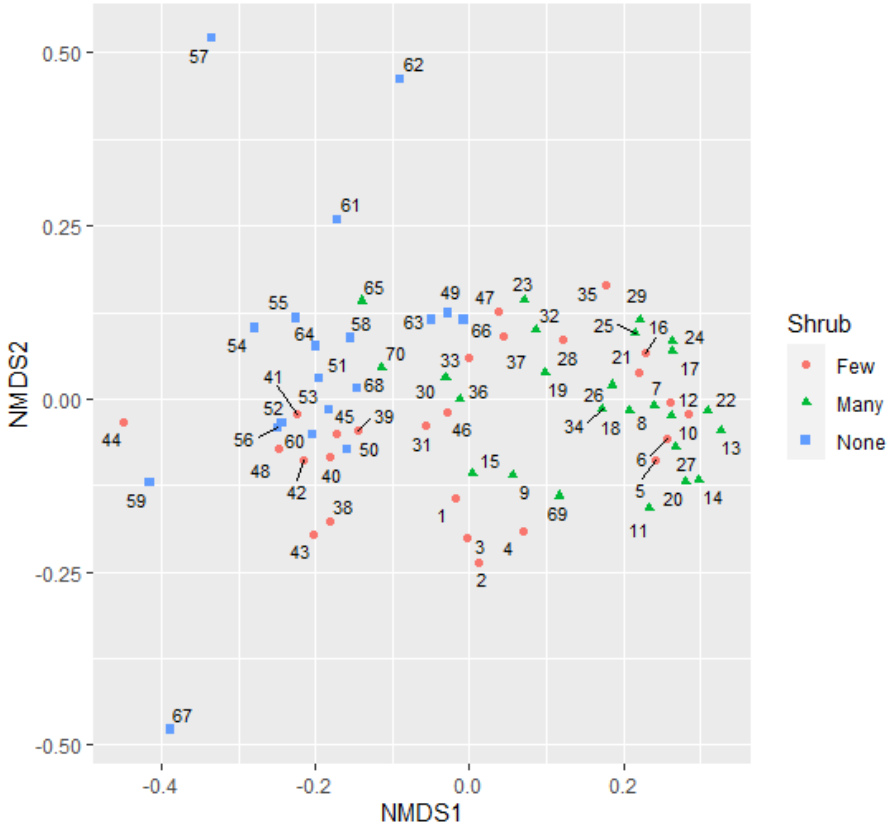




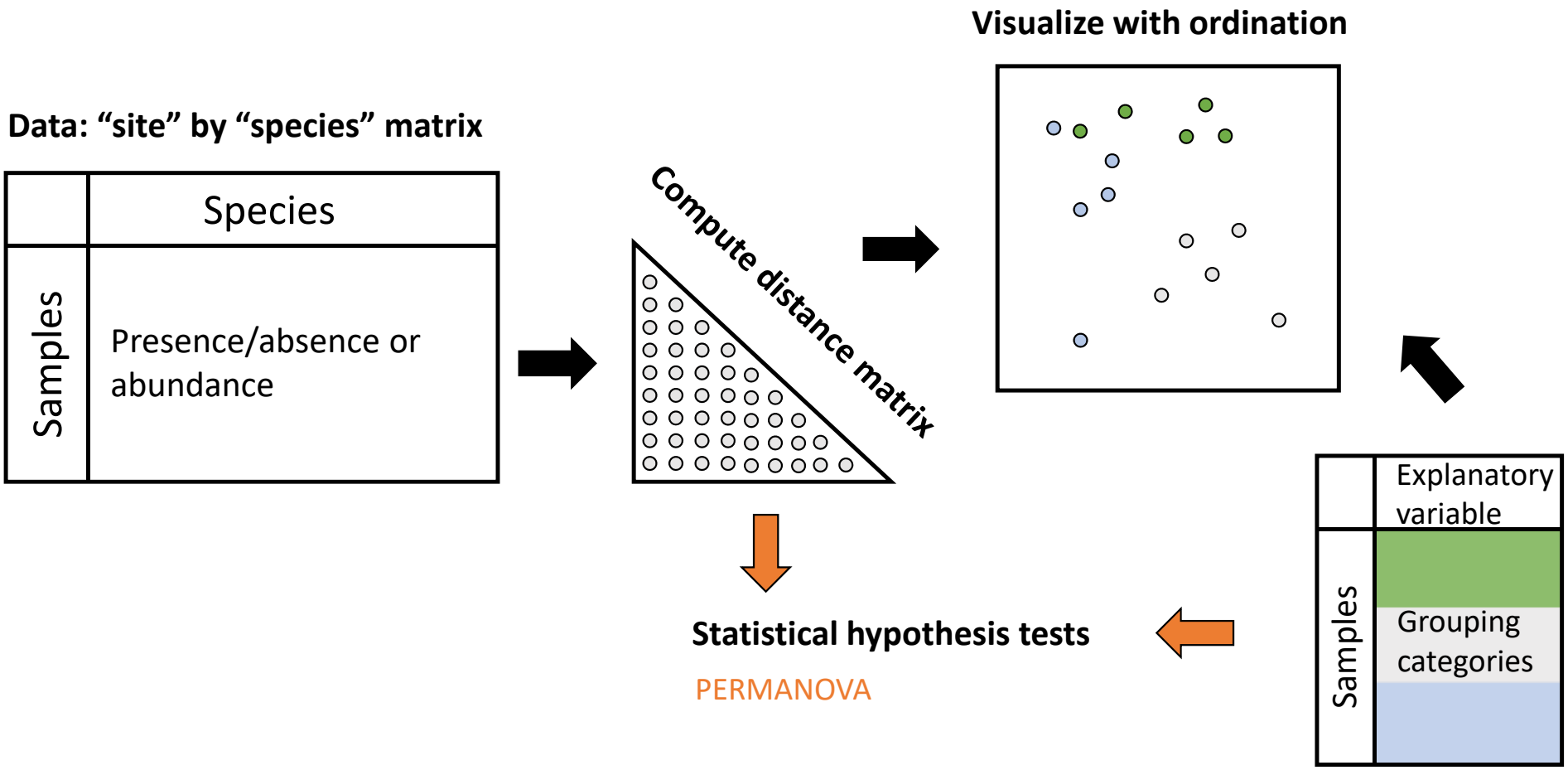
NMDS (final) output

It looks like the abundance of shrubs is an important factor in determining mite community composition

How can we test this hypothesis statistically?



NMDS: Non-metric Multi-Dimensional Scaling



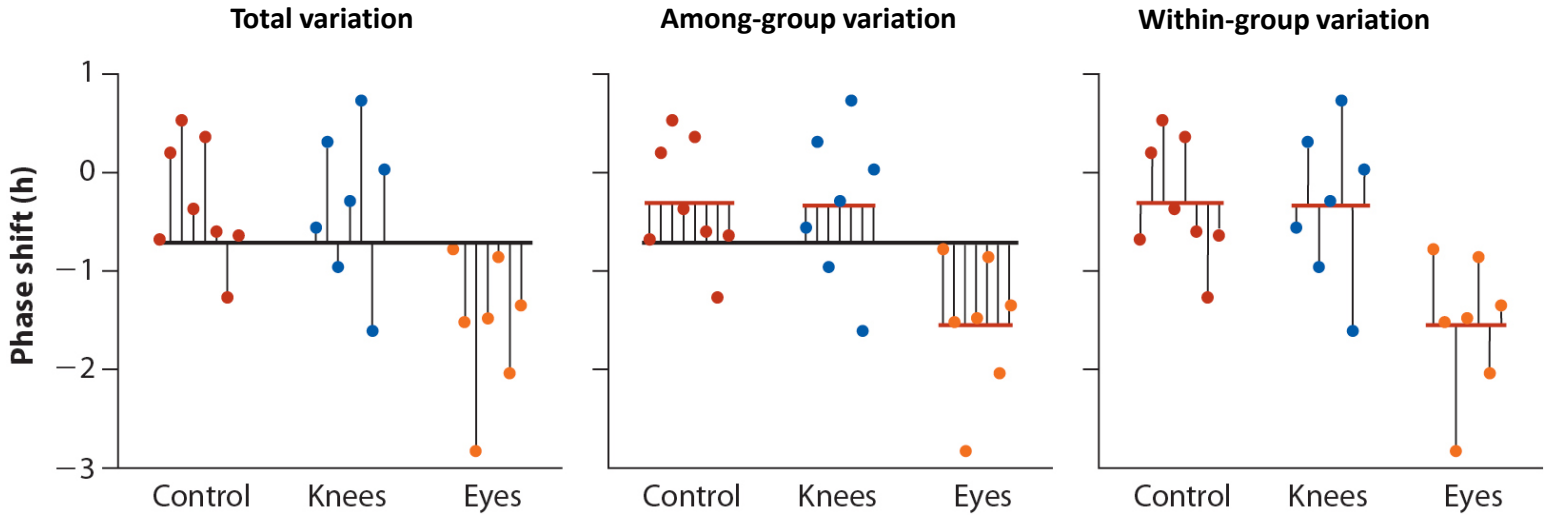
PERMANOVA is a distance-based MANOVA

- "Permutational" ANOVA
- Also called dbMANOVA (distance-based MANOVA) and npMANOVA (nonparametric MANOVA)
- A great solution for multivariate hypothesis testing where data do not meet the strict distributional requirements of a traditional MANOVA

The way PERMANOVA works is analogous to our familiar ANOVA: Recall…

Use variance ratio (*F*) to test the hypothesis that groups are from the same population:

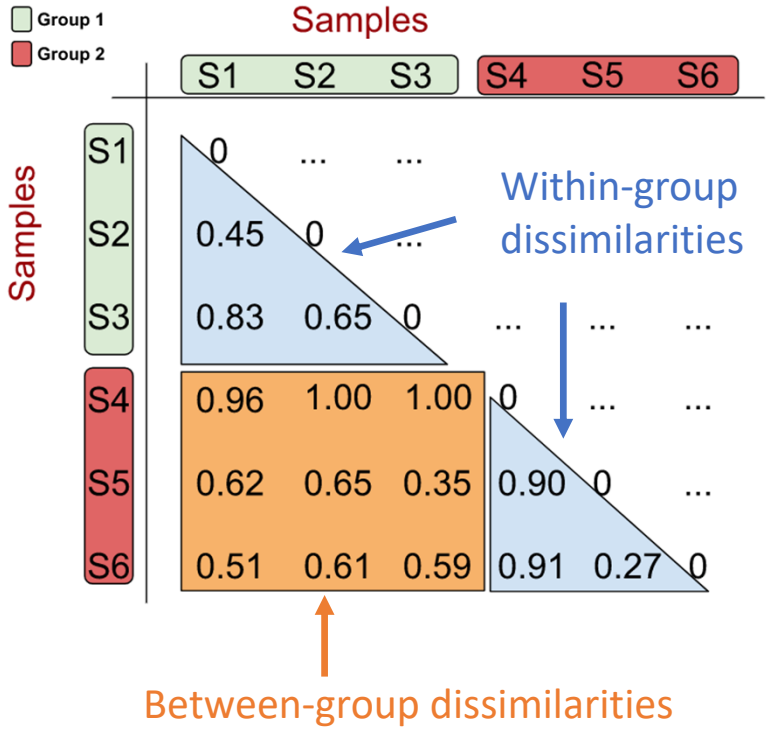
$$F = \frac{Mean\ square\ among\ groups}{Mean\ square\ within\ groups} = \frac{MS_{groups}}{MS_{errors/resids}}$$



(Figure modified from Whitlock and Schluter Fig. 15.1-2)

In a PERMANOVA, we analyze distance measures instead of the raw data values

Example Bray-Curtis distance matrix where samples can be organize into two groups, for which we'd like to compare their composition



In an ANOVA:

$$F = \frac{Mean\ square\ among\ groups}{Mean\ square\ within\ groups}$$

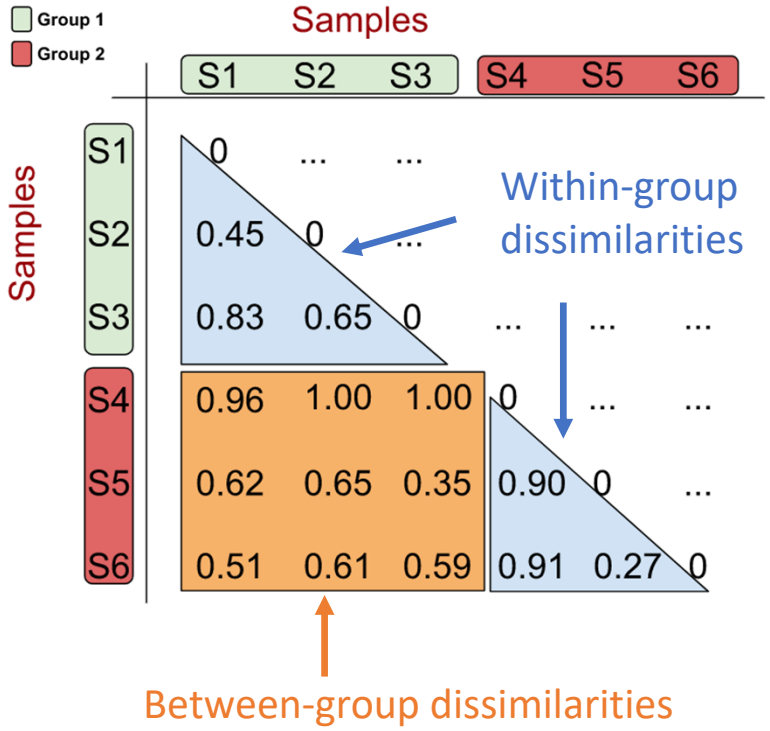
In a PERMANOVA:

$$pseudo\ F = \frac{Sum\ of\ squares\ among\ groups\ /_{factor\ df}}{Sum\ of\ squares\ within\ groups\ /_{residual\ df}}$$

(adapted from Buttigieg and Ramette 2014)

In a PERMANOVA, we analyze distance measures instead of the raw data values

Example Bray-Curtis distance matrix where samples can be organize into two groups, for which we'd like to compare their composition



$$pseudo\ F = \frac{Sum\ of\ squares\ among\ groups\ /\ factor\ df}{Sum\ of\ squares\ within\ groups\ /\ residual\ df}$$

Null hypothesis: No difference in dissimilarities among groups

Usually in an ANOVA, we would compare our *F* value to the null *F* distribution, and obtain the probability for observing this *F* or larger by chance alone (*p* value)

In a PERMANOVA, we use a permutational test to generate the null *F* distribution given our data, and obtain *p*.

(adapted from Buttigieg and Ramette 2014)

Example dataset

- mite_abund_matrix.csv: 70 samples and their mite composition
- mite_explain_var.csv: some information about the different properties of each sample that might explain difference in composition

Does mite composition vary by shrub cover and topography?

(Borcard and Legendre 1994)

To R!

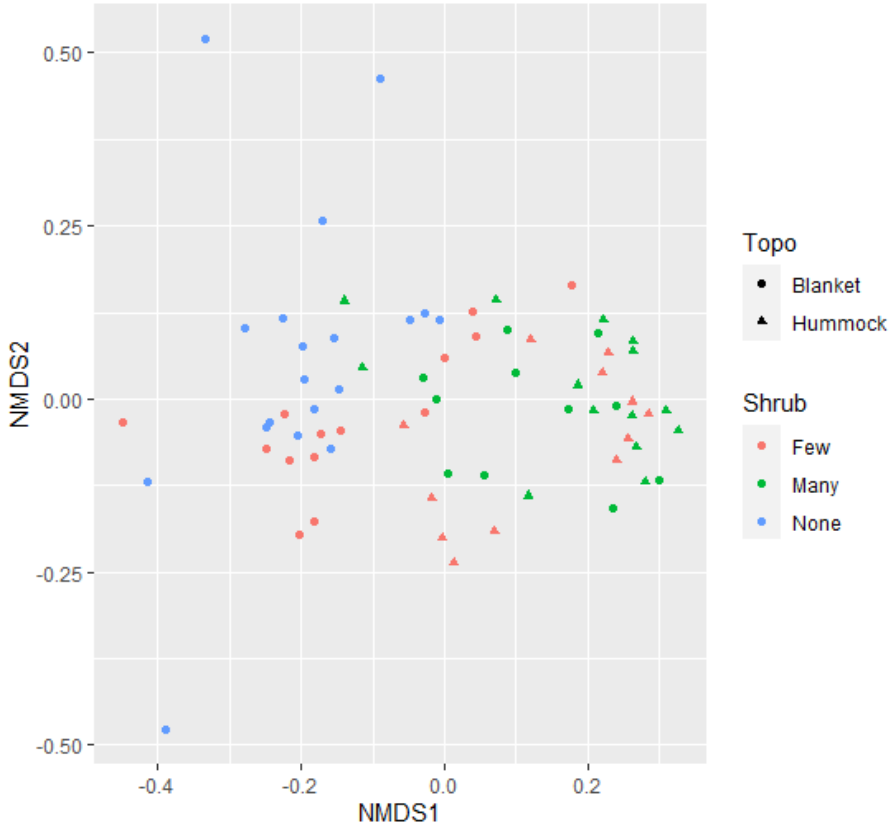
Interpret PERMANOVA table

```
Call: adonis(formula = mite.dist ~ Shrub * Topo, data = mite.info, permutations = 999)
Permutation: free Number of permutations: 999
Terms added sequentially (first to last)
           Df SumsOfSqs MeanSqs F.Model R2      Pr(>F)
Shrub       2  3.1221    1.56105 9.9836  0.21244 0.001 ***
Topo        1  0.8810    0.88102 5.6345  0.05995 0.001 ***
Shrub:Topo  1  0.5297    0.52971 3.3878  0.03604 0.013 *
Residuals  65 10.1635    0.15636 0.69157
Total      69 14.6963                    1.00000
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that a significant effect in PERMANOVA can be driven by both shifts in means and dispersions, so always a good idea to look at R2 as well to understand how much of total variation is explained by each term

We found that orbatid mite composition differed significantly among shrub cover classes ($F_{2,65}$ = 9.98, R^2 = 0.21, P = 0.001) and topographies ($F_{1,65}$ = 5.63, R^2 = 0.06, P = 0.001), as well as their interaction ($F_{1,65}$ = 9.98, R^2 = 0.04, P = 0.013). Shrub cover explained a much larger amount of total variation in mite composition (21%) compared to topography (6%).

Fig. 1 Non metric multidimensional scaling plot of orbatid mite composition. Each point corresponds to a single sample. Shapes indicate the topographical category of each sample location, and colors indicate the amount of shrub cover. The stress for this ordination solution is 0.15.



Some final words on NMDS and PERMANOVA

- Pros
  - Makes no distribution assumptions
  - Robust to outliers and missing data
- Cons
  - Iterative process to get an optimized solution instead of calculating unique solution (computationally intensive, solutions can be unstable)
  - NMDS axes are dimensionless and not directly interpretable

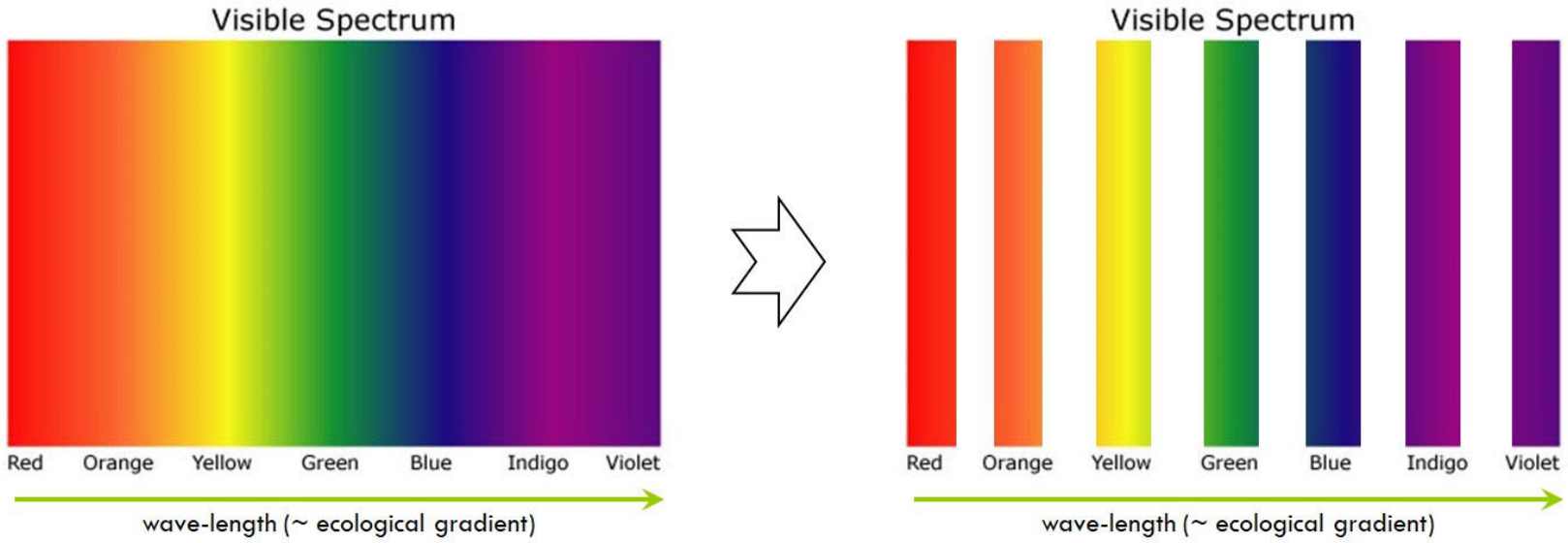
Break

What can we do with a distance matrix?

1. Ordination: visualize and find patterns
2. Clustering: group samples based on distances
3. Analysis and hypothesis testing:
   a) Are distances between groups greater than distances within groups? (PERMANOVA)
   b) Are distances between samples within groups homogenous among groups? (PERMDISP)

Clustering and classification

Goal: find discontinuities in multivariate data and group objects into classes/clusters that are more similar to each other than other groups



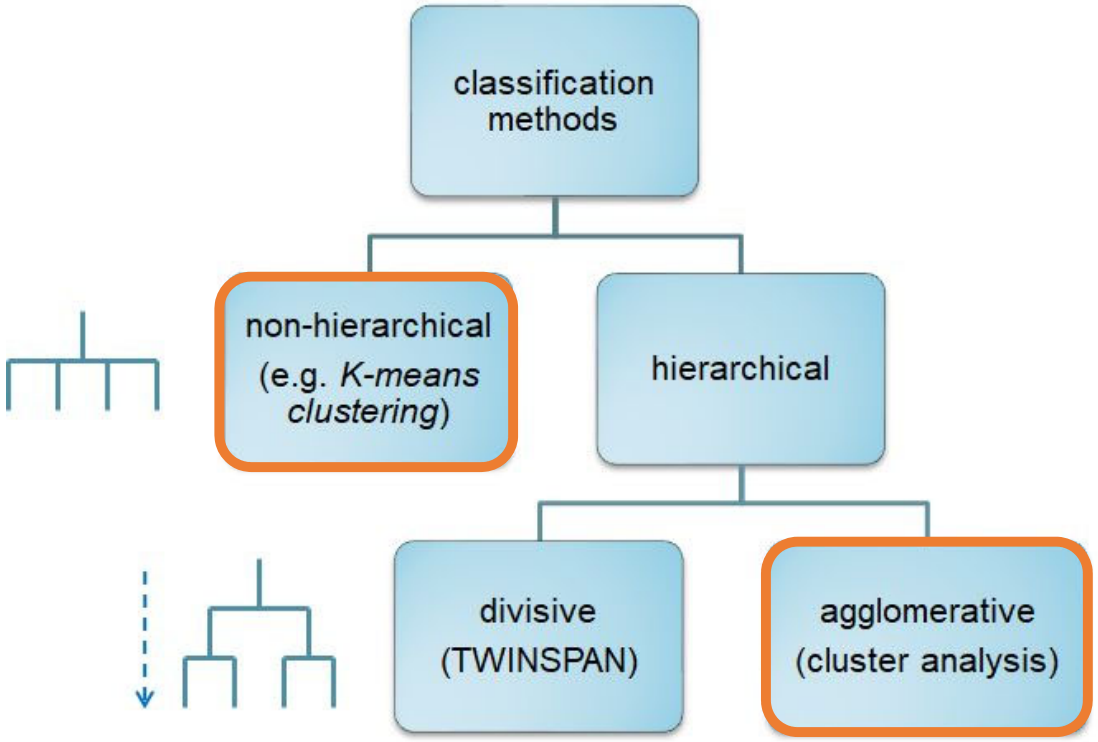
(from Zelený)

Important things to consider

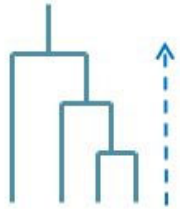
- What is the purpose of the analysis?
  - Supervised (uses external criteria) vs. Unsupervised (based purely on data at hand)
- What are the classifying/clustering criteria?
  - Choice of dis/similarity metric
- How to decide where to draw boundaries between groups?
  - Algorithm choice

- Clustering does not test a statistical hypothesis

Types of algorithms



Cluster analyses generally produce *dendrograms*

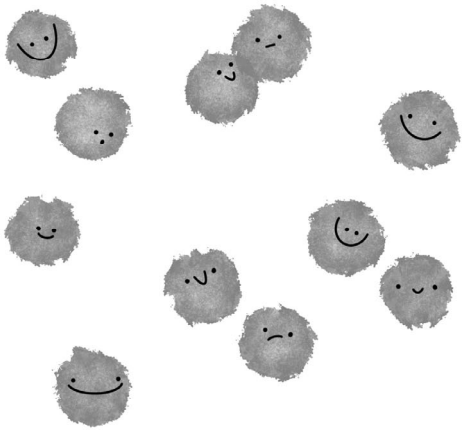


(from Zelený)

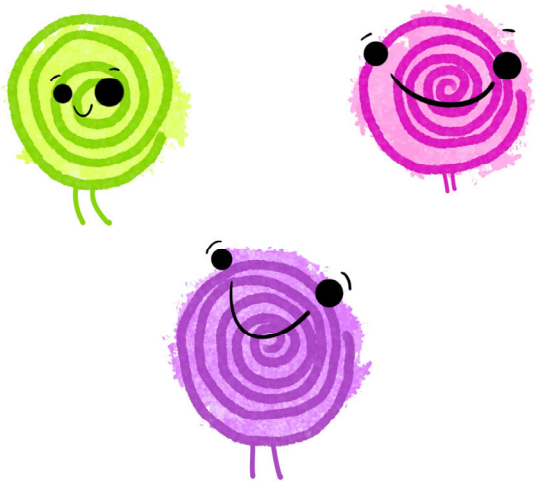
k-means clustering

: assign each observation to one of k clusters based on the nearest cluster centroid.

OBSERVATIONS

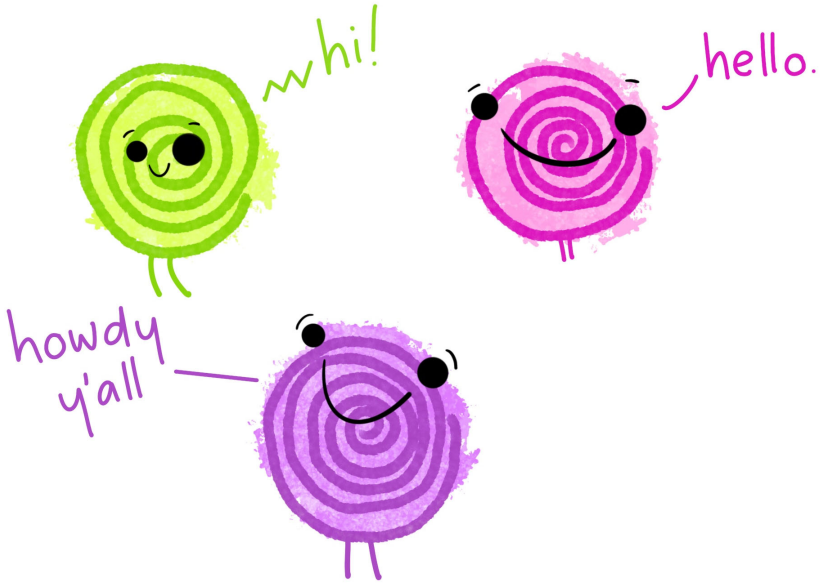


cluster CENTROIDS

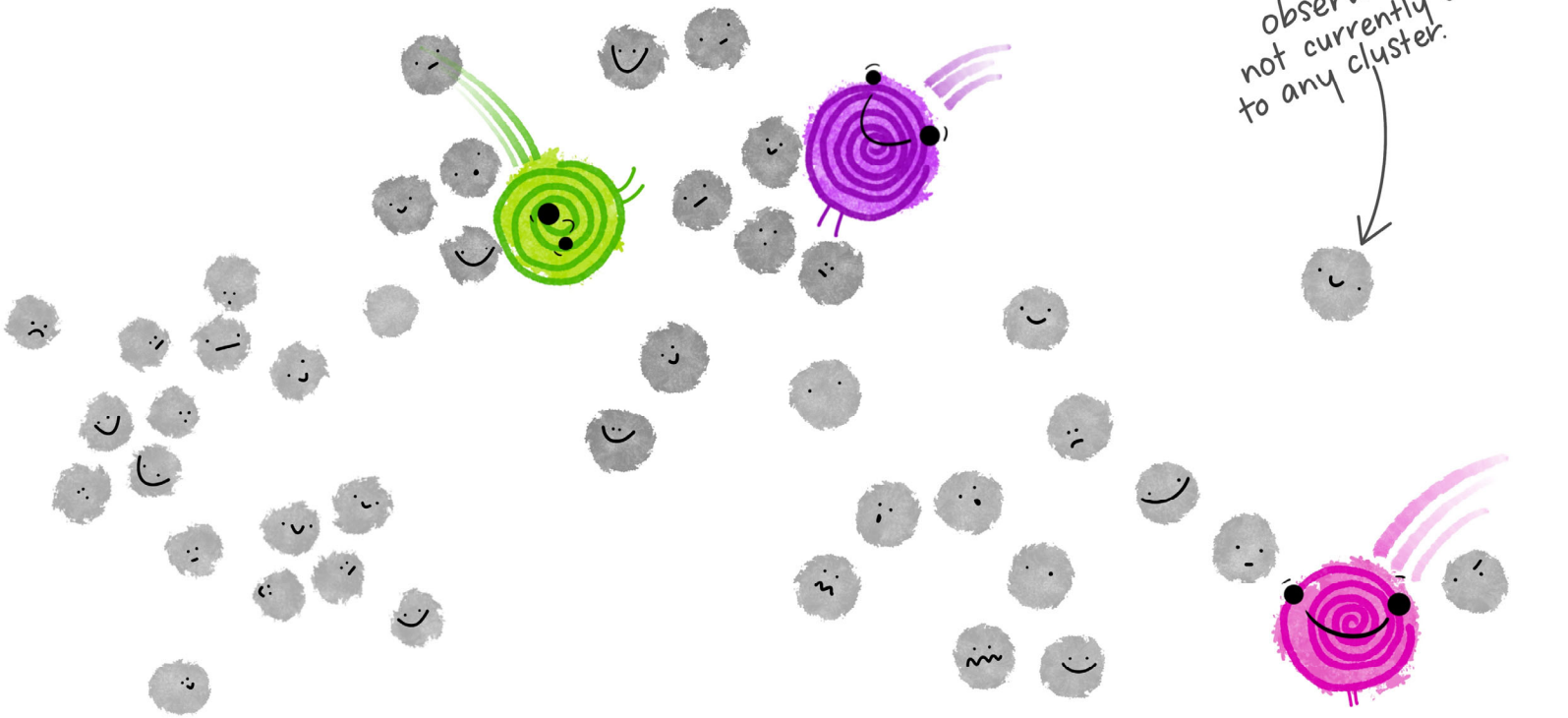


1 Specify the number of clusters (in this example, k = 3).

Then imagine k cluster centroids are created.



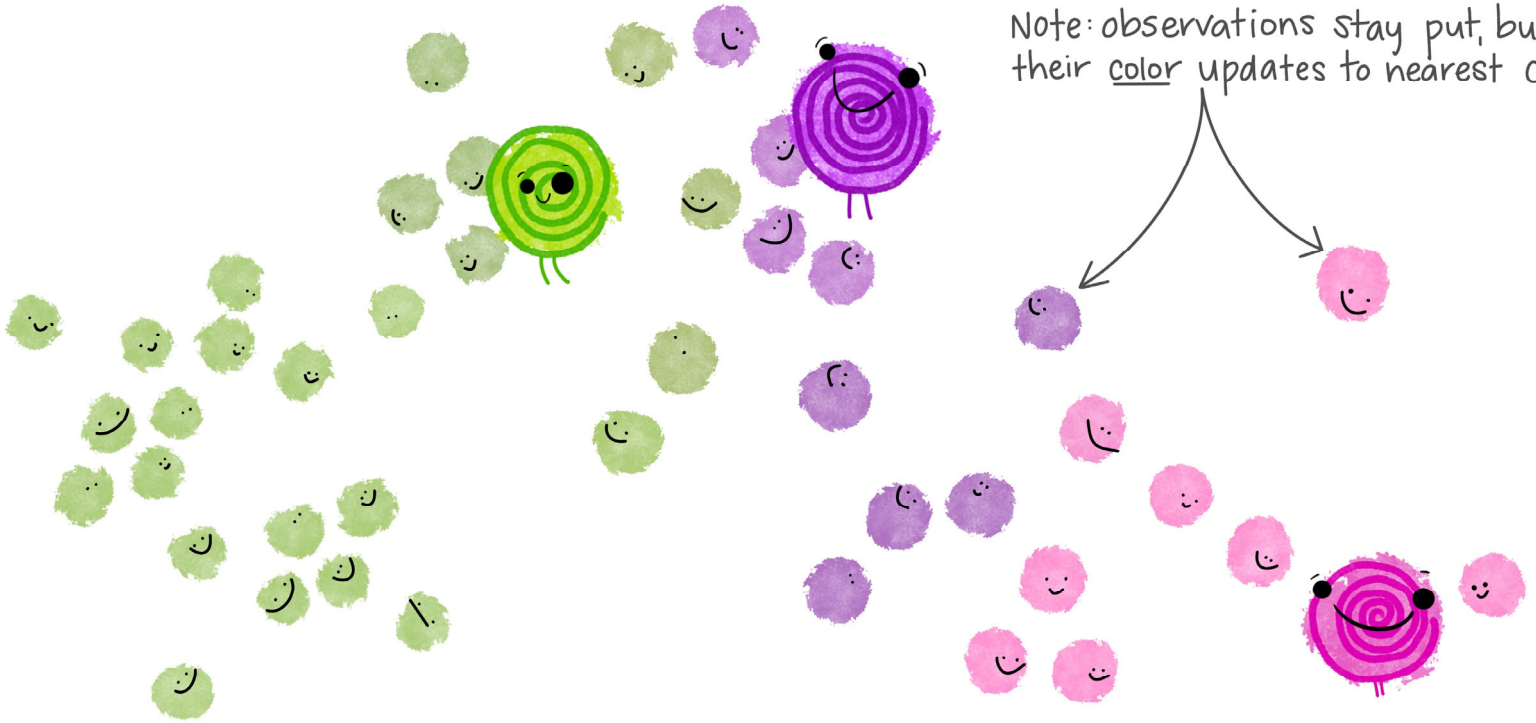
2 Those k centroids get randomly placed in your space.



Each observation gets temporarily "assigned" to its closest centroid.

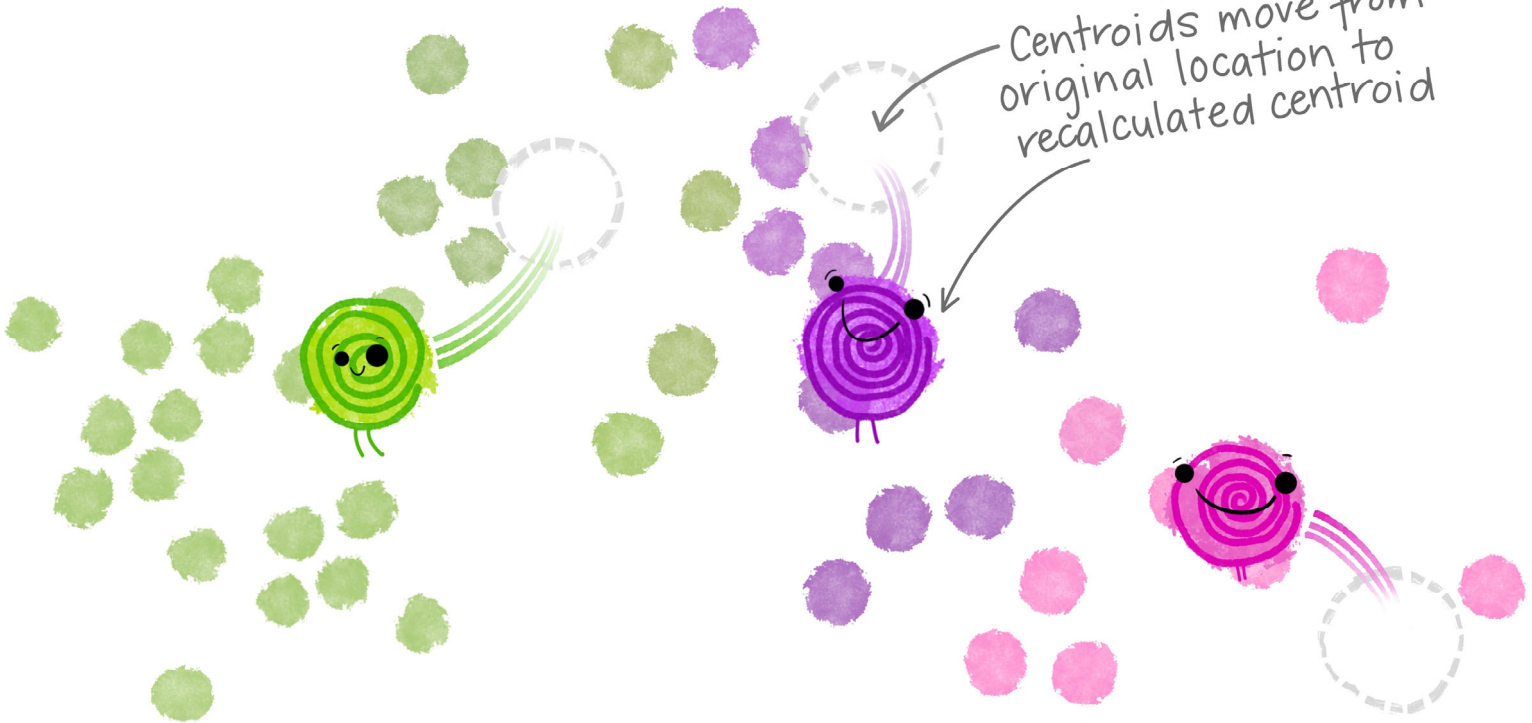
(e.g. by Euclidean distance)

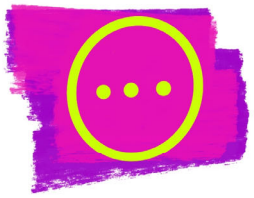
Note: observations stay put, but their color updates to nearest centroid!



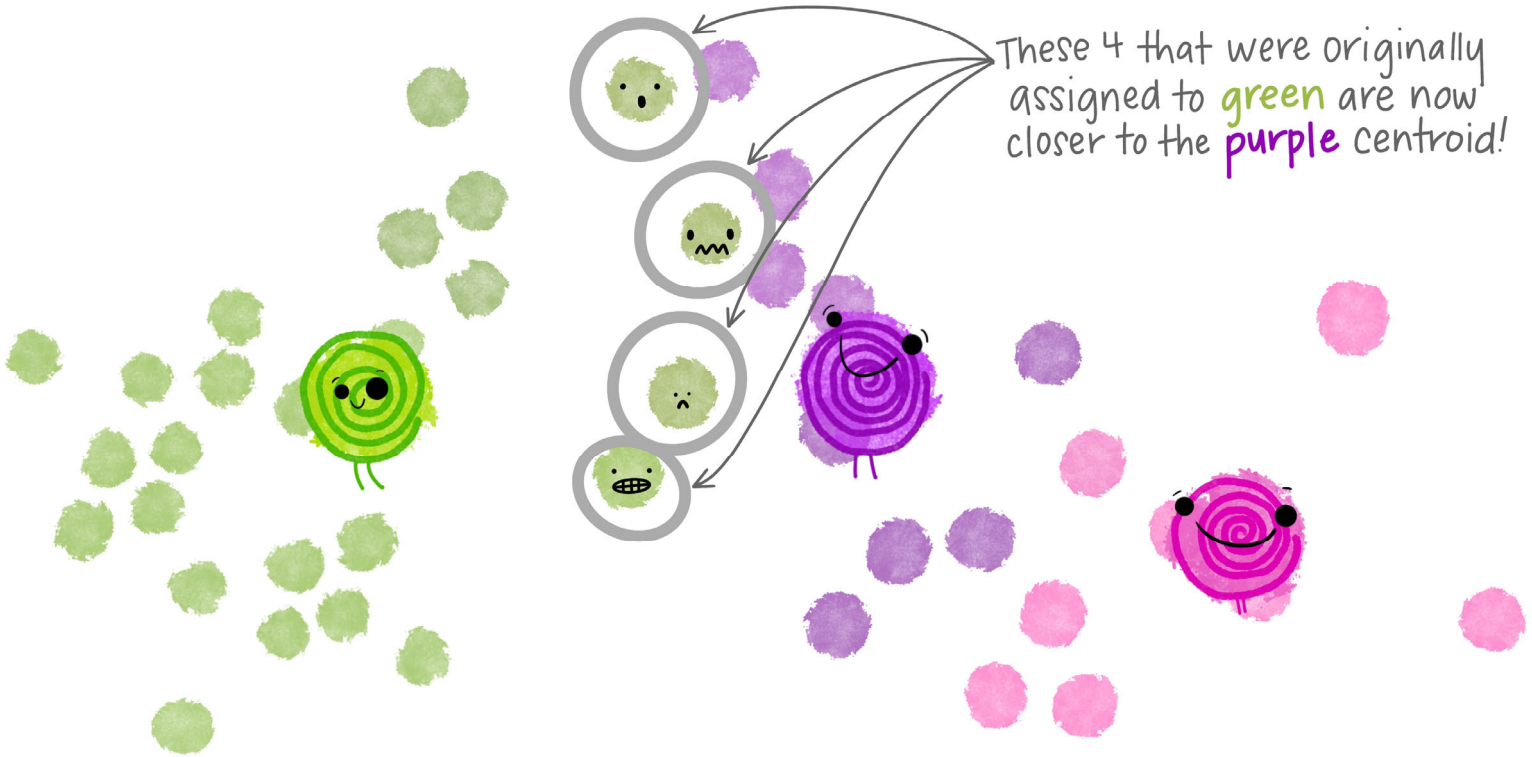
4

Then the centroid of each cluster is calculated based on all observations assigned to that cluster...



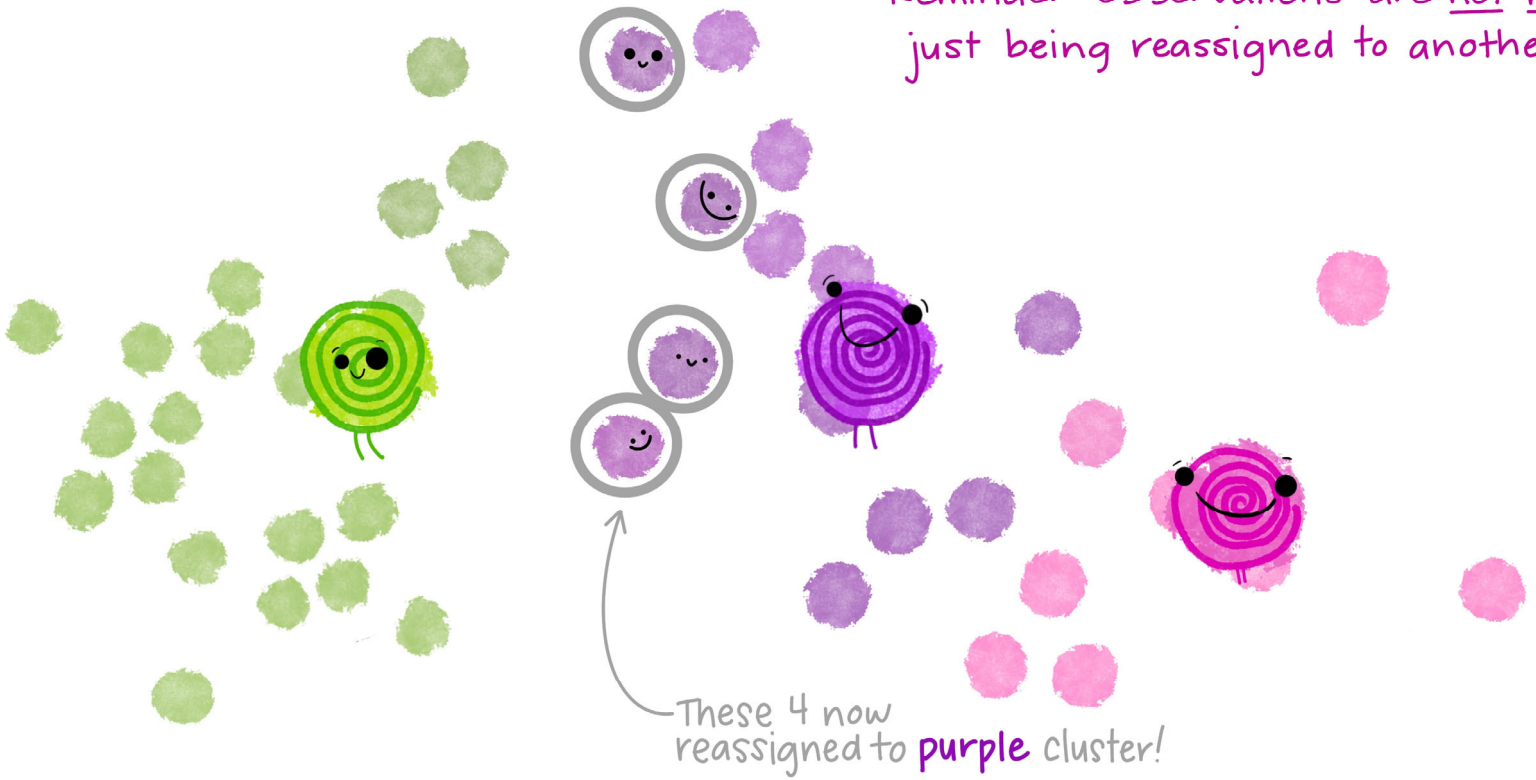


UH OH. Now that the cluster centroids have moved, some of the observations are now closer to a different centroid!



NO PROBLEM!
Observations get reassigned* to a different cluster based on the recalculated centroid.

*Reminder: observations are not moving, just being reassigned to another cluster.

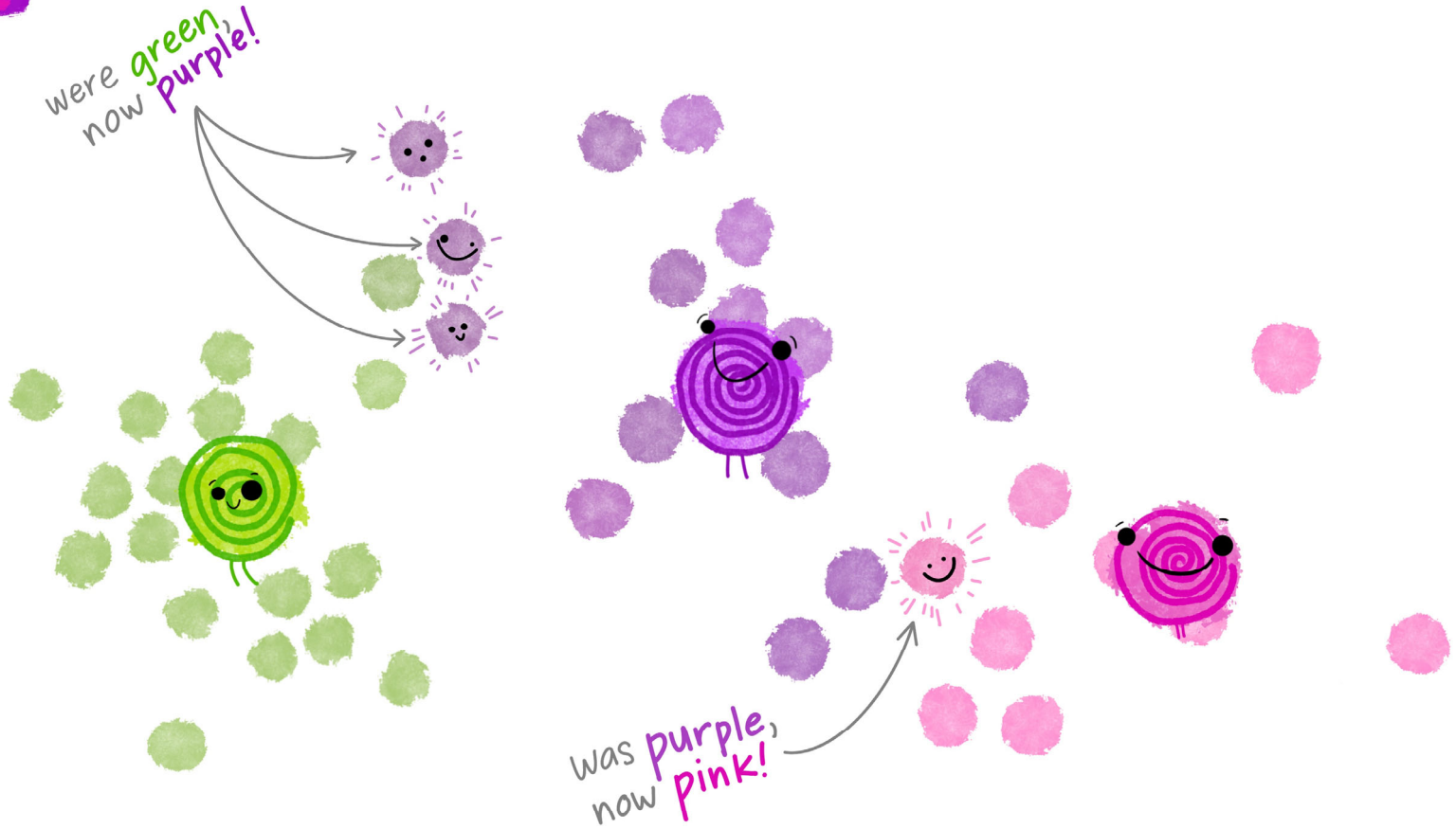


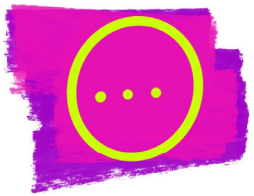
6

But now that observations have been reassigned, the centroids need to move again [recalculate centroids from updated clusters]

7

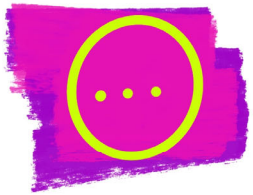
Again, now observations are reassigned as needed to the closest centroid.





Then the centroid for each cluster is recalculated...

...which means observations will be reassigned...



That iterative process of

Recalculate cluster centroids
↳ Reassign observations to nearest centroid
↳ Recalculate cluster centroids
↳ Reassign observations to nearest centroid
↳ Recalculate cluster centroids
↳ Reassign observations to nearest centroid

↓

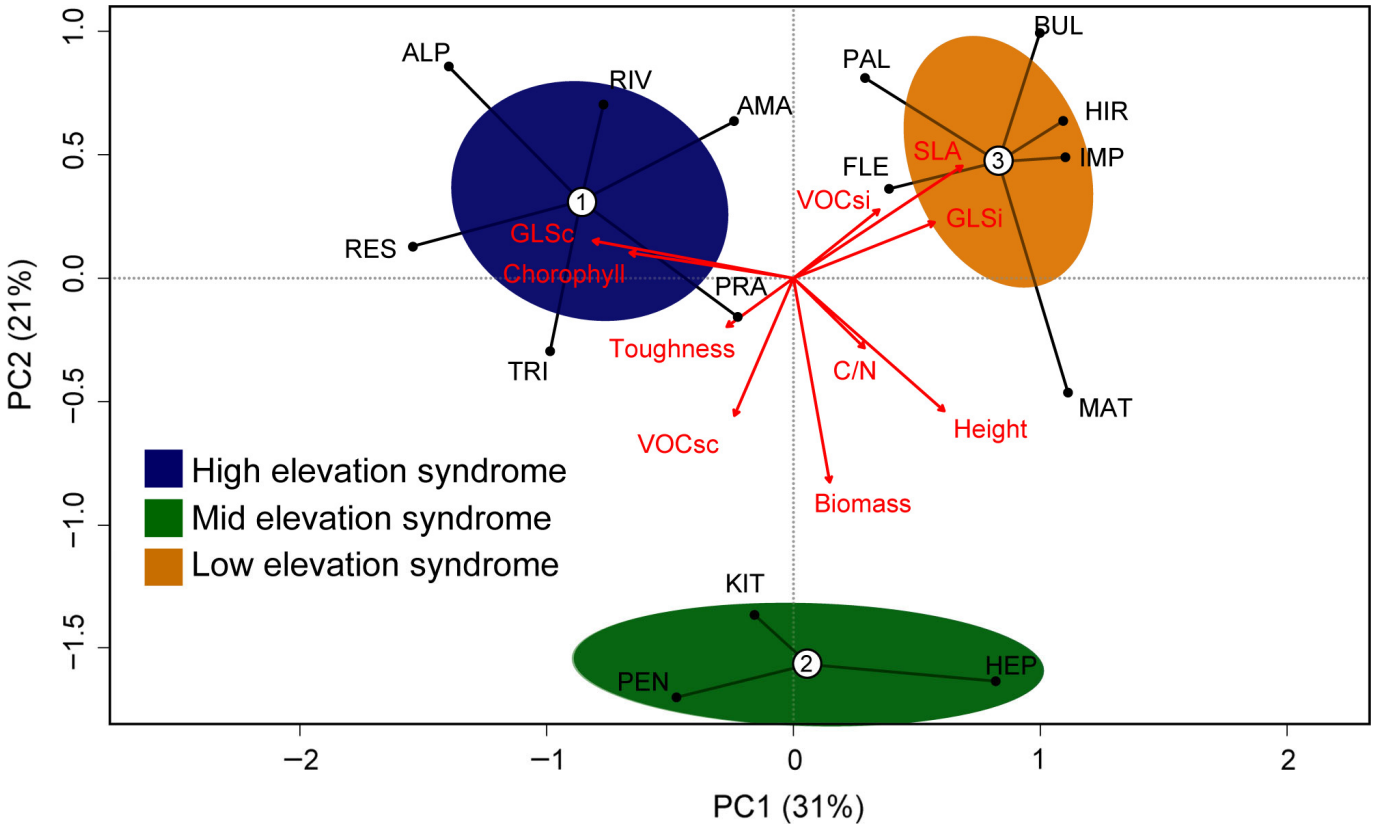
Continues until nothing is moving or being reassigned anymore!

Which means the iteration is done and each observation is assigned to its final cluster.

K means clustering recap

- This is a linear method using Euclidean distances, so not so great for zero-inflated, abundance/proportion data
- Iterative process
- User defines how many clusters desired in the solution

The unfolding of plant growth form-defense syndromes along elevation gradients (Defossez et al. 2018)



"The three clusters extracted from the k-means analysis altogether explained 51% of the total variability shared among traits involved in both plant growth form and defenses (Figure S3). These clusters represent three syndromes related to the growth-defense trade-offs, and within-traits correlations (Fig. 1), and which correspond to three characteristic zones along the studied elevation gradient (Figure S2)."

hierarchical clustering: single linkage

(Step-by-step: combine clusters with the smallest distance between elements)

also called “nearest neighbor”

elements

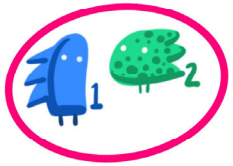
DISTANCE MATRIX

	1	2	3	4	5	6	7
1	0	10	30	40	60	85	82
2	10	0	24	38	55	87	90
3	30	24	0	16	26	50	63
4	40	38	16	0	21	52	67
5	60	55	26	21	0	41	58
6	85	87	50	52	41	0	32
7	82	90	63	67	58	32	0

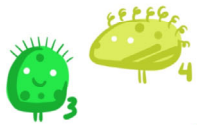
Treat each element as a cluster

- Find smallest distance between elements in 2 clusters
- Those clusters get merged.

elements



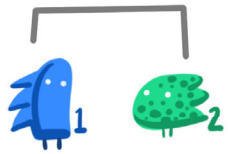
1 & 2 merged first



DISTANCE MATRIX

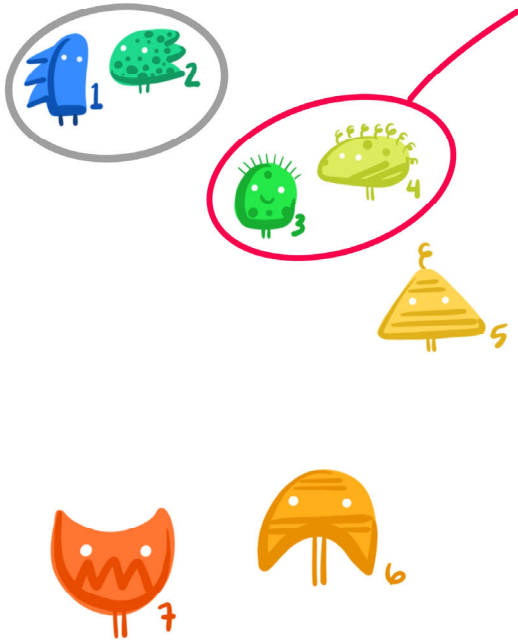
	1	2	3	4	5	6	7
1	0	10	30	40	60	85	82
2	10	0	24	38	55	87	90
3	30	24	0	16	26	50	63
4	40	38	16	0	21	52	67
5	60	55	26	21	0	41	58
6	85	87	50	52	41	0	32
7	82	90	63	67	58	32	0

build the DENDROGRAM



Now 1 & 2 are a single cluster.
Find the 2 clusters with smallest distance between elements,
then merge them.

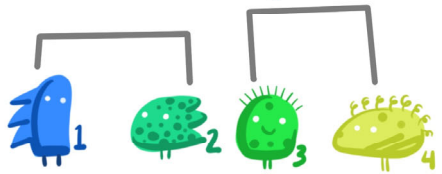
elements



DISTANCE MATRIX

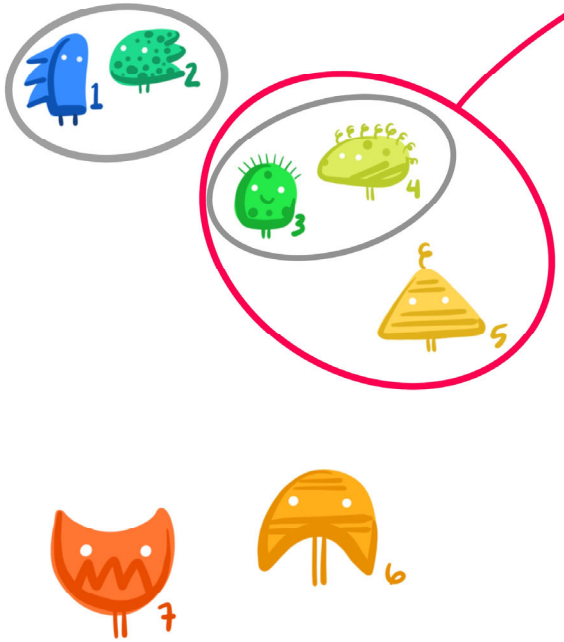
	1	2	3	4	5	6	7
1	0	~~10~~	30	40	60	85	82
2	~~10~~	0	24	38	55	87	90
3	30	24	0	16	26	50	63
4	40	38	16	0	21	52	67
5	60	55	26	21	0	41	58
6	85	87	50	52	41	0	32
7	82	90	63	67	58	32	0

build the DENDROGRAM



Repeat! Now the 2 clusters with the smallest distance between elements are the (3, 4) and 5 clusters, so we merge them!

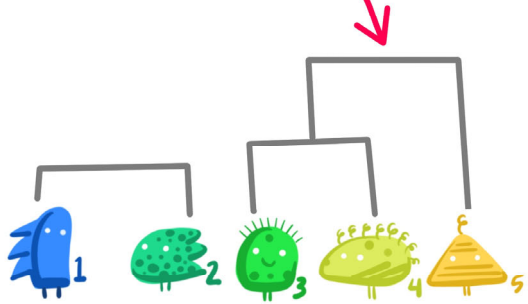
elements



DISTANCE MATRIX

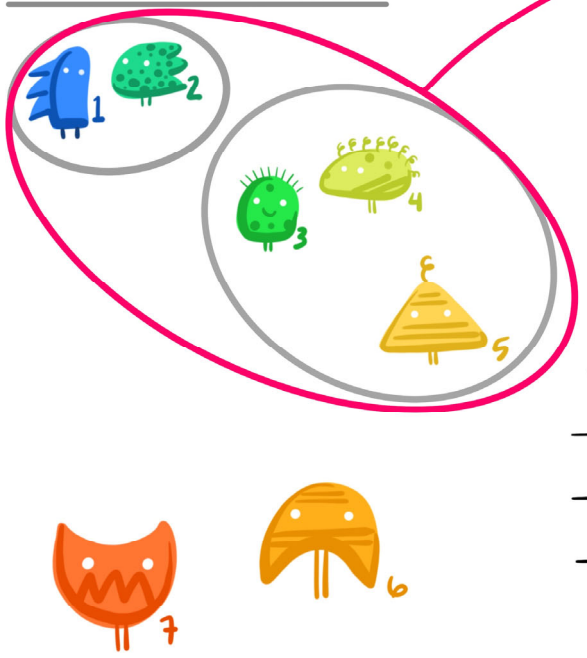
	1	2	3	4	5	6	7
1	0	~~10~~	30	40	60	85	82
2	~~10~~	0	24	38	55	87	90
3	30	24	0	~~16~~	26	50	63
4	40	38	~~16~~	0	21	52	67
5	60	55	26	21	0	41	58
6	85	87	50	52	41	0	32
7	82	90	63	67	58	32	0

build the DENDROGRAM



Yep, do it again! Now, the smallest distance between elements in two clusters is between 2 & 3, so we merge the clusters they're in!

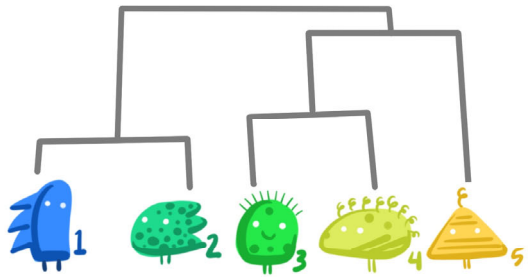
elements



DISTANCE MATRIX

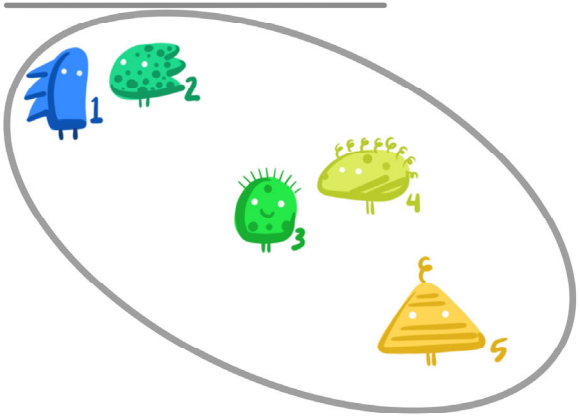
	1	2	3	4	5	6	7
1	0	~~10~~	30	40	60	85	82
2	~~10~~	0	24	38	55	87	90
3	30	24	0	~~16~~	26	50	63
4	40	38	~~16~~	0	~~21~~	52	67
5	60	55	26	~~21~~	0	41	58
6	85	87	50	52	41	0	32
7	82	90	63	67	58	32	0

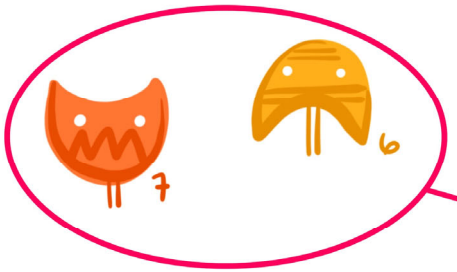
build the DENDROGRAM



The next smallest distance between elements in separate clusters is between 6 & 7, so we merge them...

elements

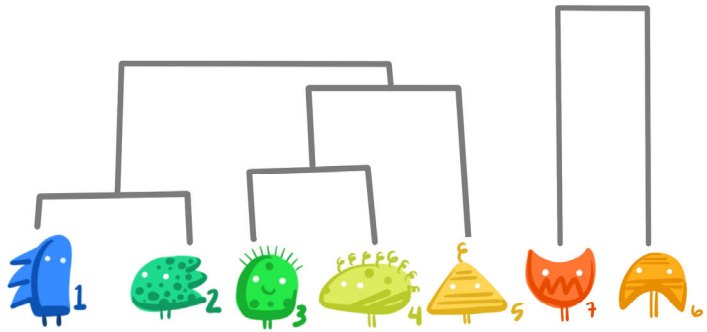




DISTANCE MATRIX

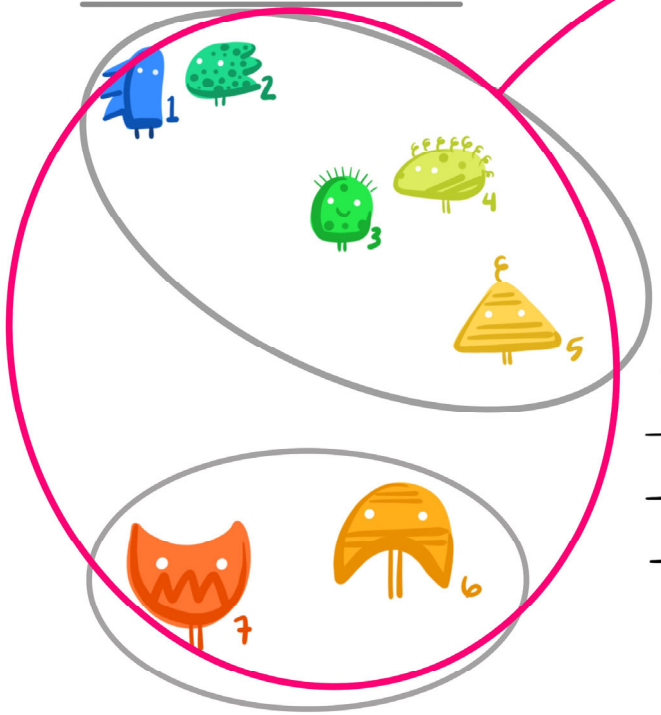
	1	2	3	4	5	6	7
1	0	~~10~~	~~30~~	~~40~~	~~60~~	85	82
2	~~10~~	0	~~24~~	~~38~~	~~55~~	87	90
3	~~30~~	~~24~~	0	~~16~~	~~26~~	50	63
4	~~40~~	~~38~~	~~16~~	0	~~21~~	52	67
5	~~60~~	~~55~~	~~26~~	~~21~~	0	41	58
6	85	87	50	52	41	0	32
7	82	90	63	67	58	32	0

build the DENDROGRAM



Now we only have two clusters, so they get merged!

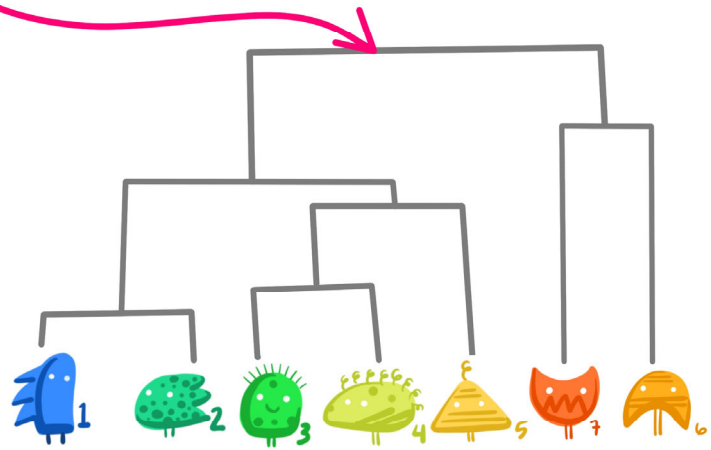
elements



DISTANCE MATRIX

	1	2	3	4	5	6	7
1	0	~~10~~	~~30~~	~~40~~	~~60~~	85	82
2	~~10~~	0	~~24~~	~~38~~	~~55~~	87	90
3	~~30~~	~~24~~	0	~~16~~	~~26~~	50	63
4	~~40~~	~~38~~	~~16~~	0	~~21~~	52	67
5	~~60~~	~~55~~	~~26~~	~~21~~	0	~~41~~	58
6	85	87	50	52	~~41~~	0	~~32~~
7	82	90	63	67	58	~~32~~	0

build the DENDROGRAM

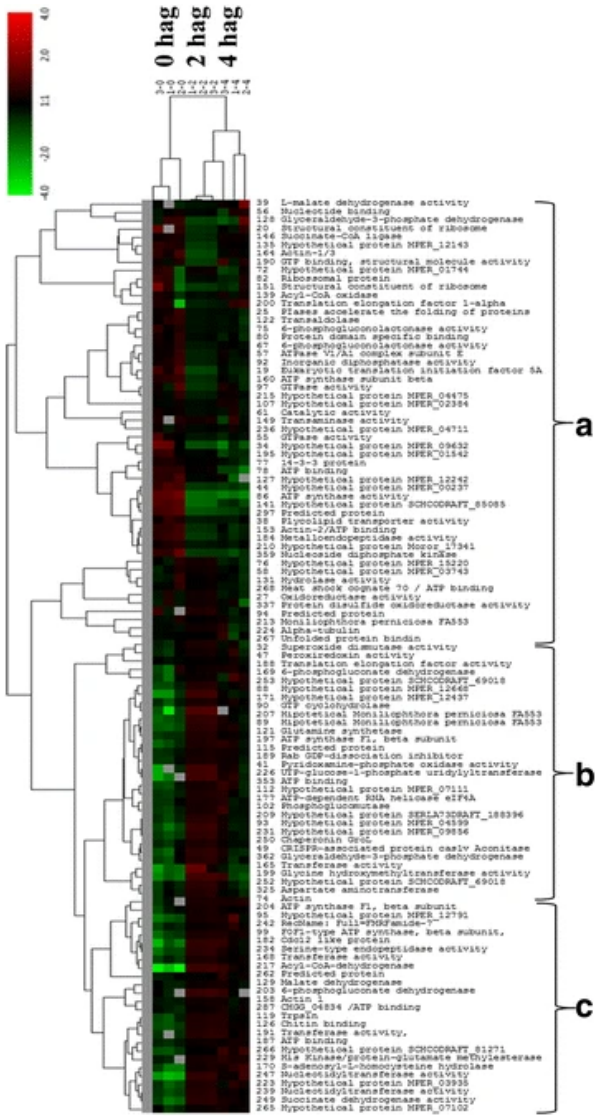


tada.

7/7

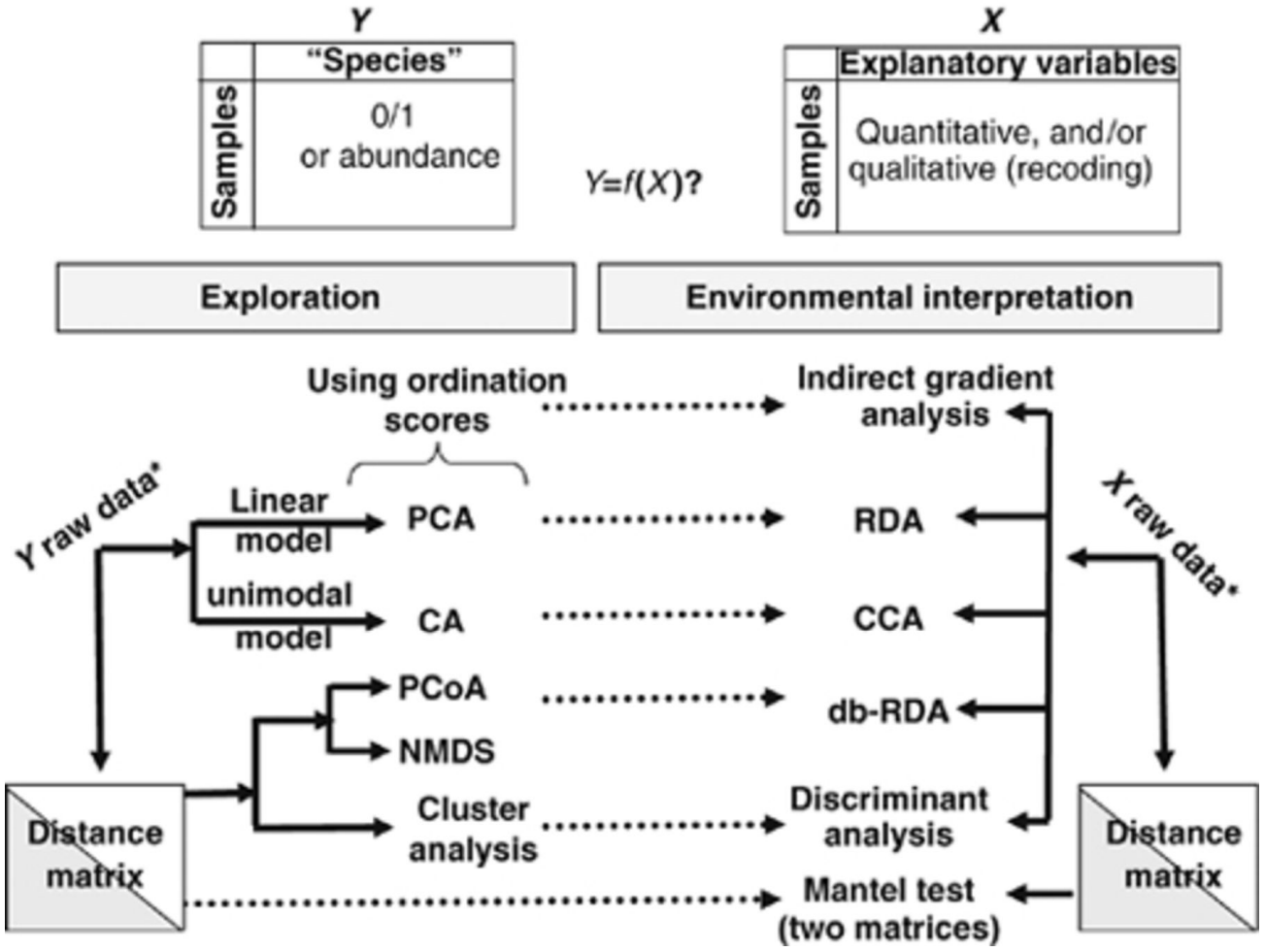
Hierarchical clustering recap

- Based on any distance matrix
- Can be conducted with different methods (e.g. single-linkage vs. complete-linkage)
- Distances on dendrogram roughly reflect the distances between samples
- Good at revealing hierarchical relationships and gradients
- The clusters may or may not be easily interpretable (not user-defined)



Proteomic analysis during of spore germination of *Moniliophthora perniciosa*, the causal agent of witches' broom disease in cacao (Mares et al. 2017)

"Bi-directional Hierarchical Clustering Analysis generated by Cluster 3.0 software showing the global profile of differential expression of proteins common to the three germination times. **a** Proteins repressed 4 h after germination. **b** proteins repressed 2 h after germination. **c** Proteins induced 4 h after germination"



(Ramette 2007)

This week

- Learn about distance-based ordination, hypothesis testing, and clustering methods
- Practice how to run an NMDS and PERMANOVA analysis
- Gain appreciation for the vast diversity of tools available to analyze multivariate data, and some of the important themes that guide decision-making

Next week

- No class!
- Lab 10 is the last lab! Not due until the week after.