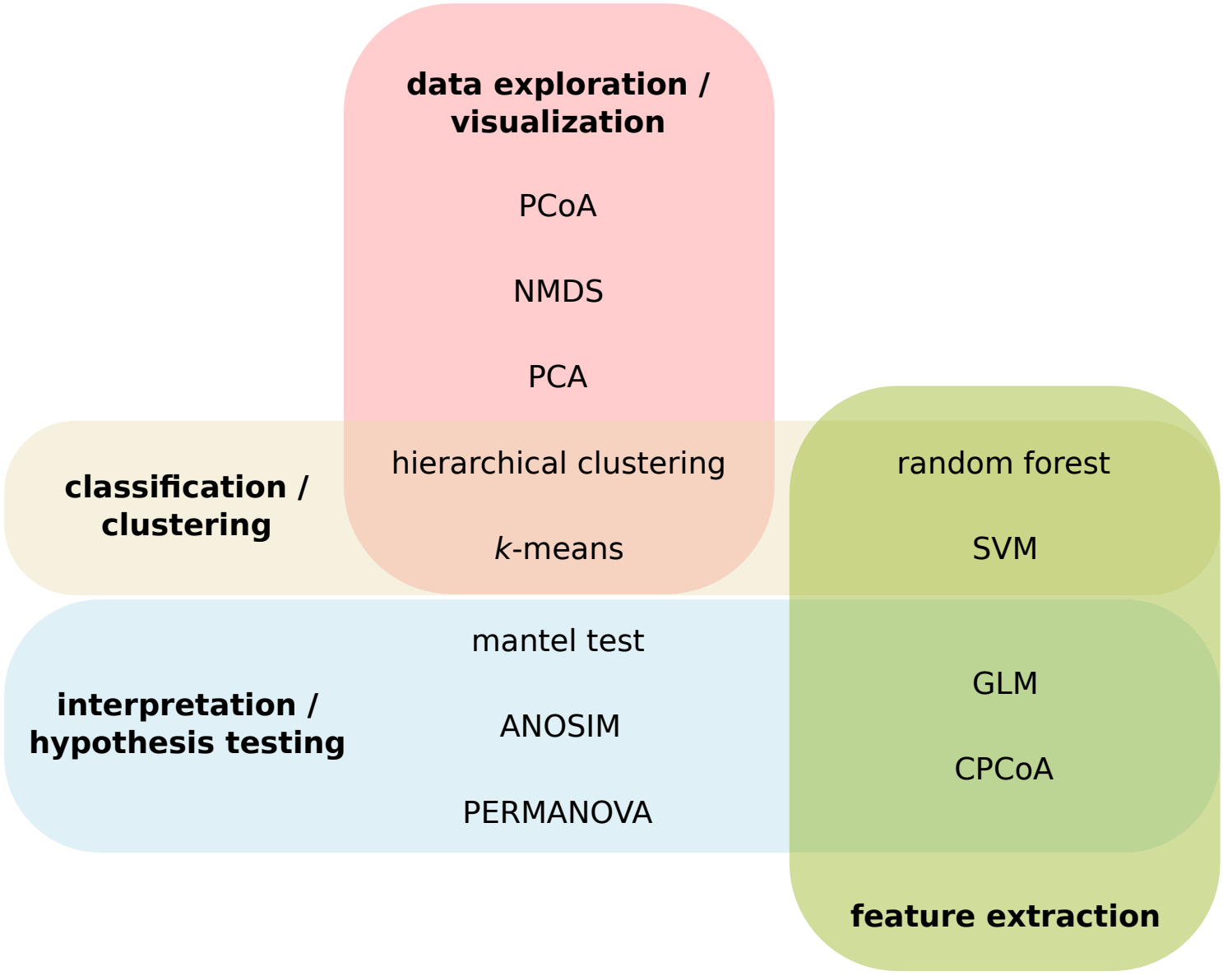
# Diversity analyses and community data visualization

Ruben Garrido-Oter
Max Planck Institute for Plant Breeding Research


DECRyPT bioinformatics workshop - October 2019

# Overview of common multivariate analysis methods in microbial ecology



# Ordination methods in microbial ecology

**PCA (Principal Component Analysis; Pearson, 1901)**

consists on rotating the original system of coordinates to maximize dispersion
input are coordinates of datapoints in a high-dimensional space
most widely used and simple (fast) ordination method
R function: prcomp (stats)

**PCoA (Principal Coordinate Analysis; Gower, 1966)**

similar to PCA but first transforms distances into coordinates in a new space
input are pairwise distances between datapoints
popular in microbial ecology because it allows employing various distances
R function: cmdscale (stats)

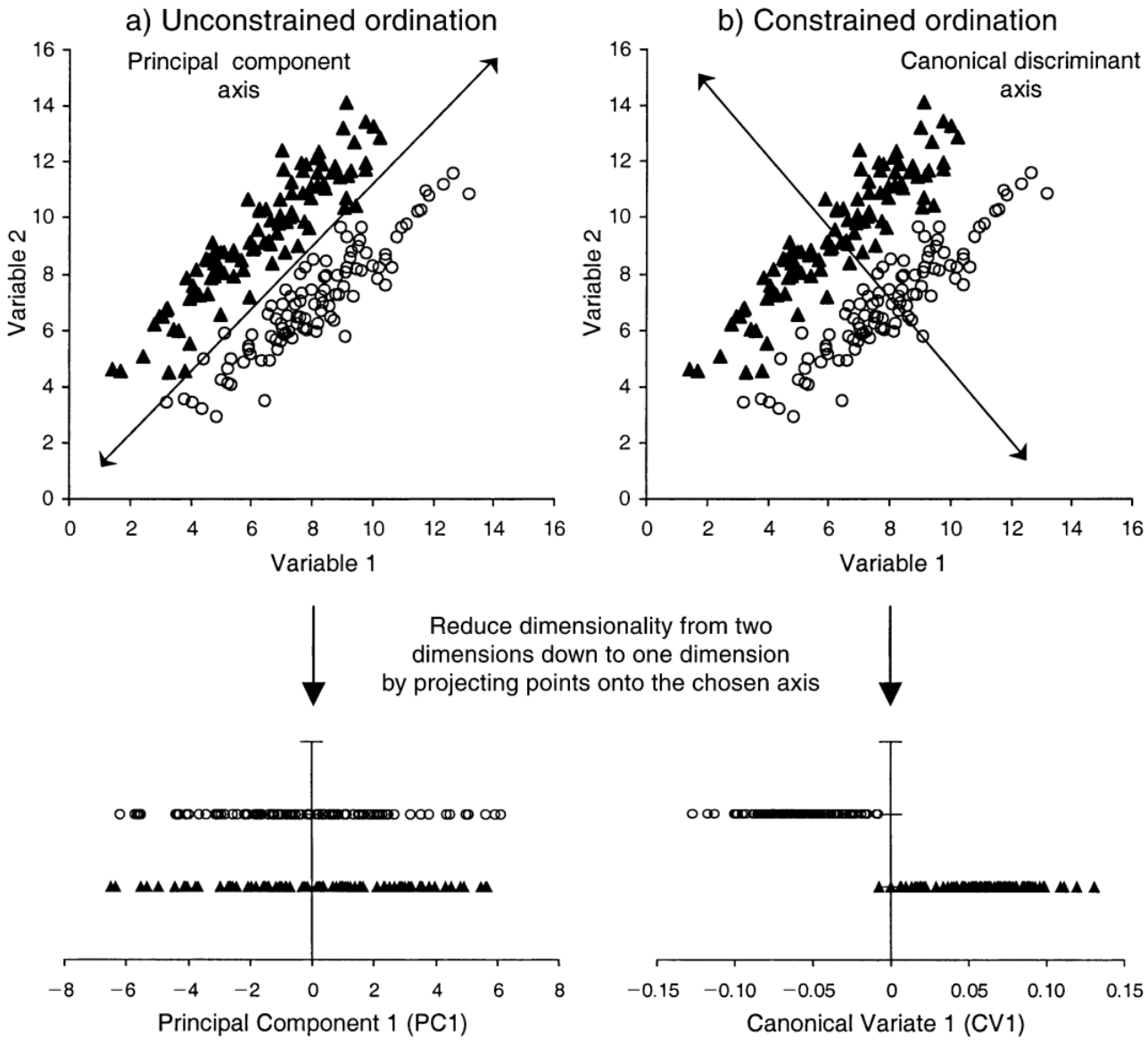
**NMDS (Non-metric Multidimensional Scaling; Kruskal et al., 1964)**

numerical rather than analytical method (slow(er), non-deterministic)
number of dimensions *k* are chosen *a priori*
all variance of the data is used to distribute points in a *k*-dimensional space
(Euclidean) distances in the new space are monotonically related to original distances
R function: isoMDS (MASS)

**CPCoA (Constrained Principal Coordinate Analysis; Legendre and Legendre, 1998)**

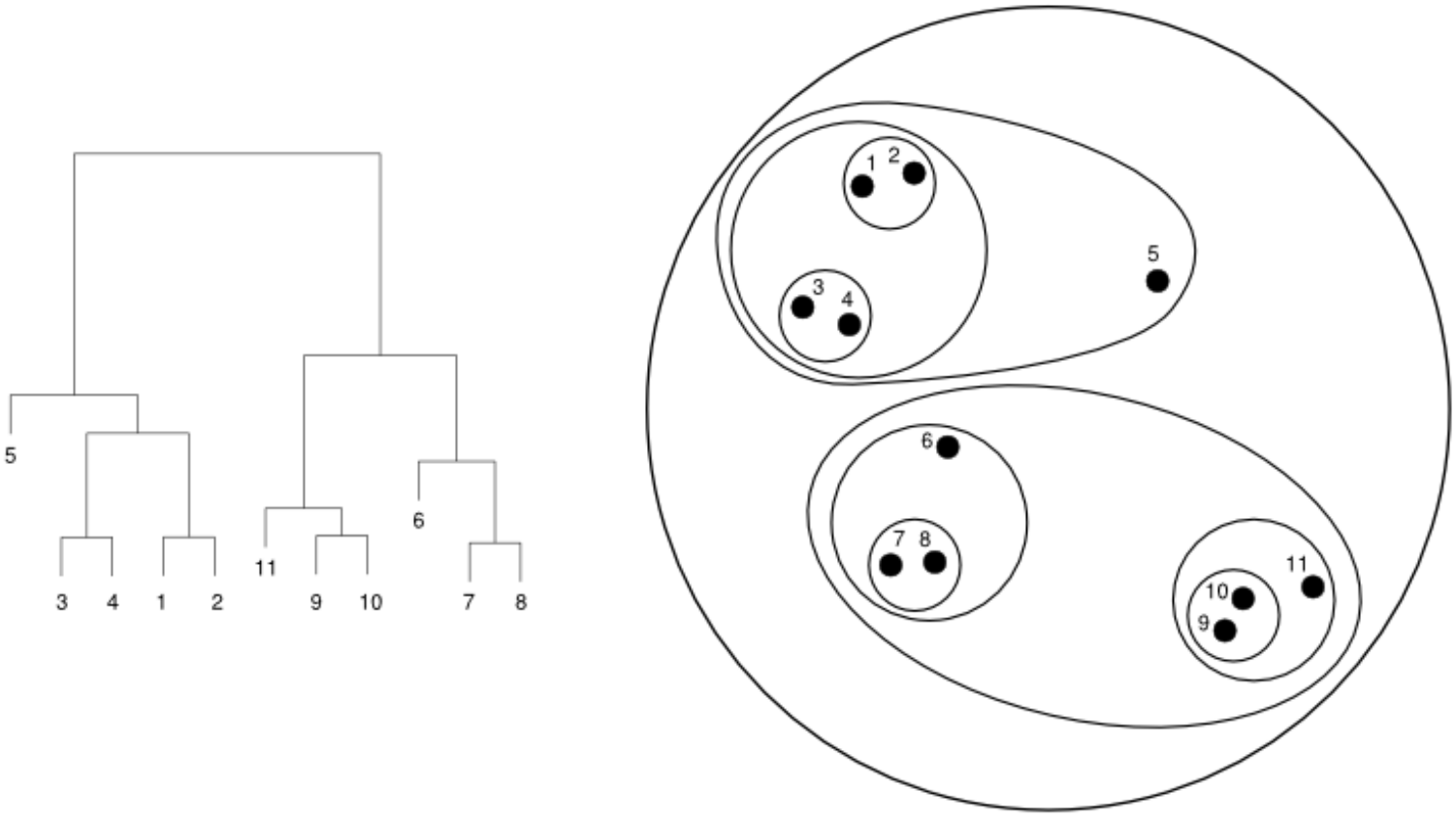
similar to PCoA but attempts to maximize separation between groups (env. variables)
used to address specific hypotheses (e.g. significant differences among groups)
statistical test of hypothesis by permutation procedures
R function: capscale (vegan)

# Constrained vs. unconstrained ordination




Anderson and Willis, 2003

# Hierarchical clustering



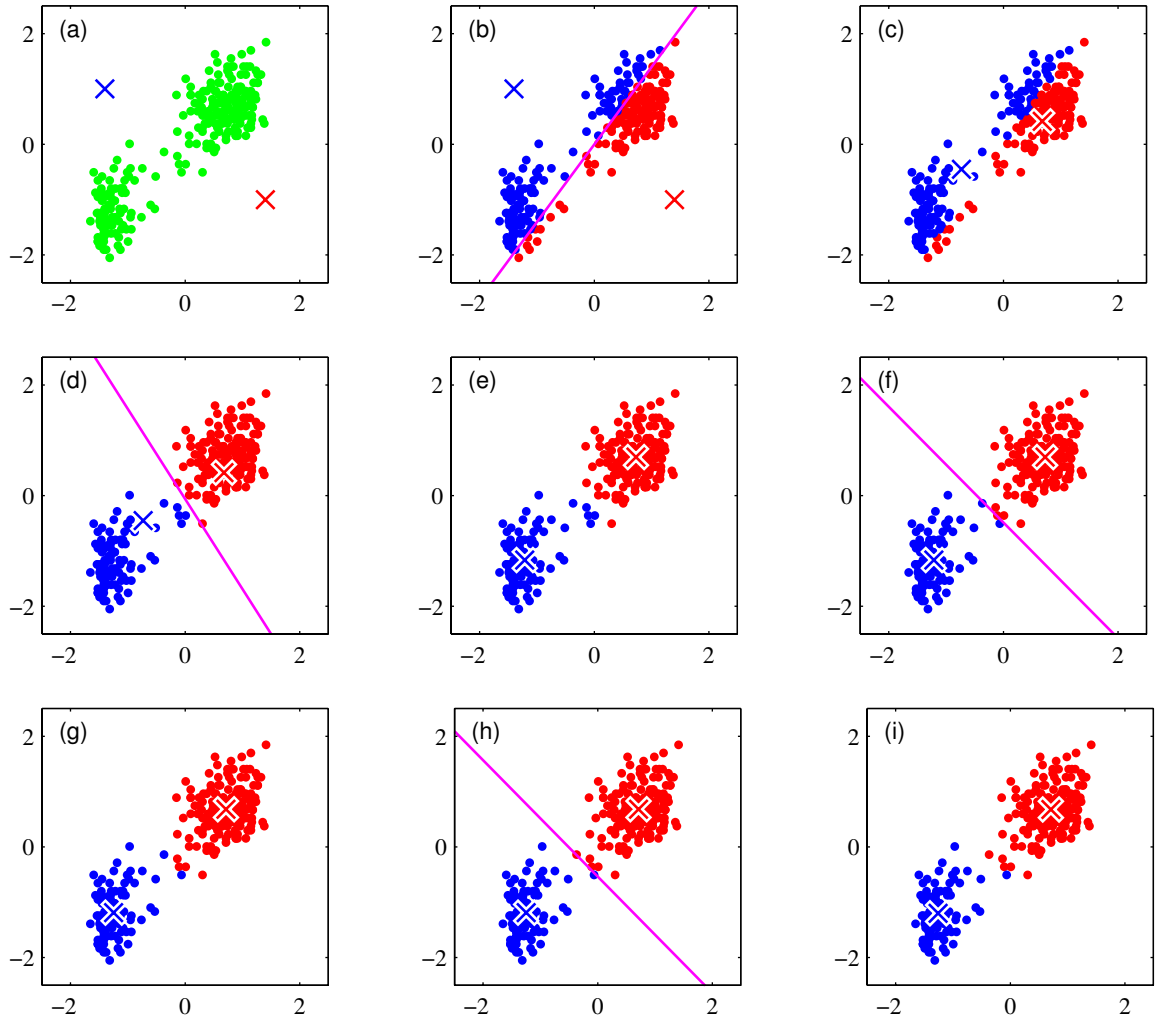

unsupervised clustering algorithm that can be used for data visualization

there are many variants (single linkage, average linkage, UPGMA, etc.)

fast and robust, can capture non-linear groups of datapoints

difficult to choose the number of clusters *k*

# *k*-means clustering



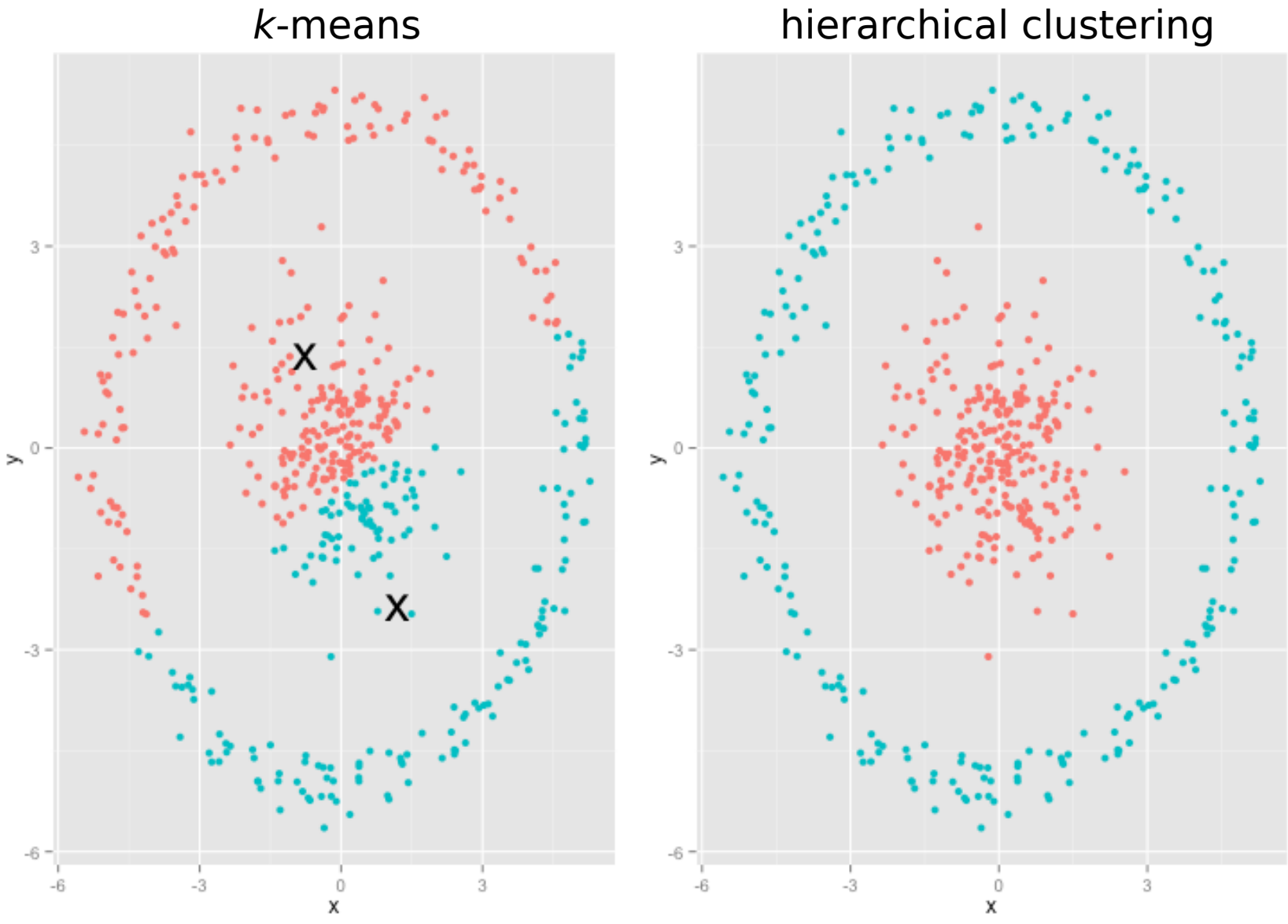

widely used, fast and robust unsupervised clustering method

depends on initialization of centroids

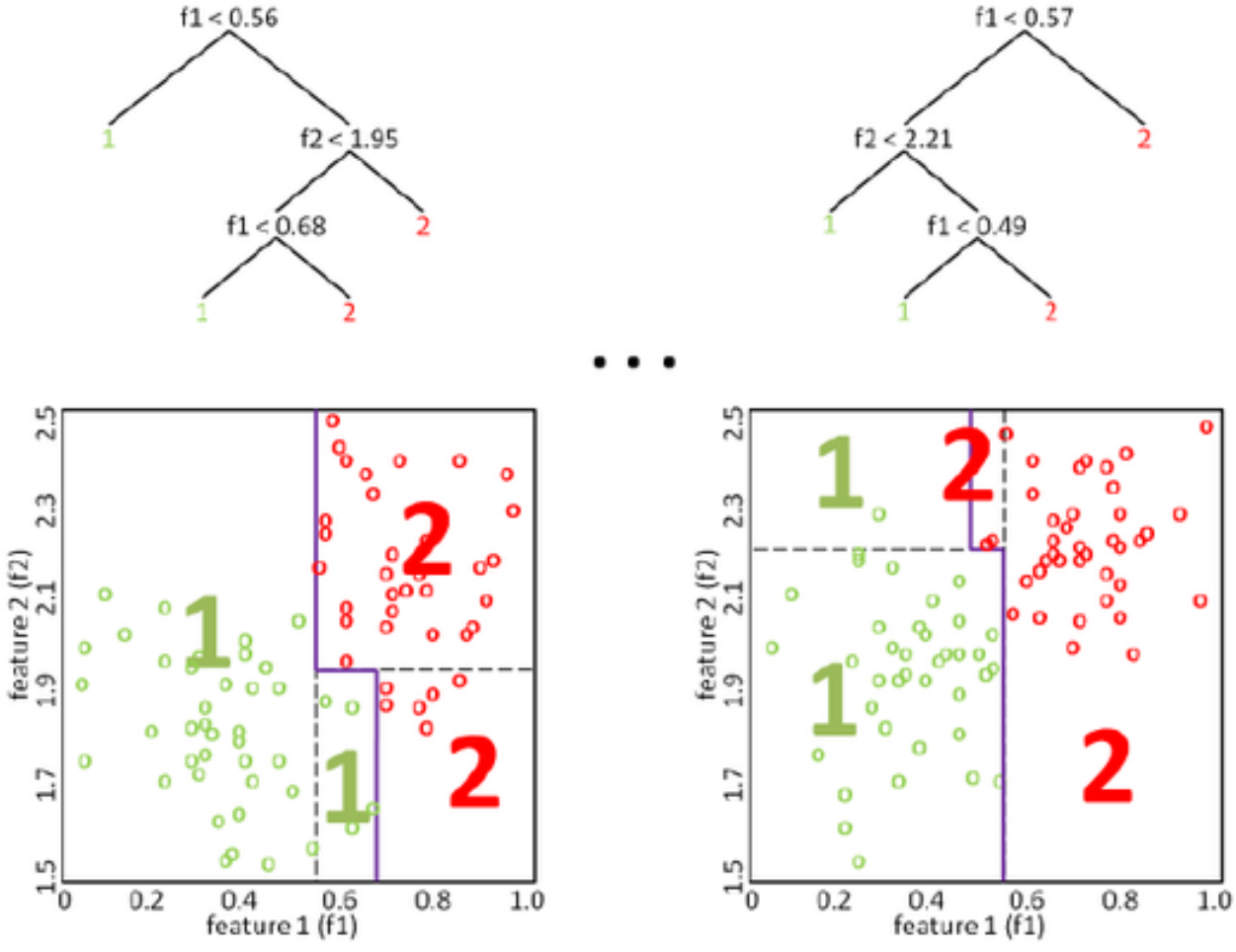
difficult to choose the number of clusters $k$

linear boundaries for classification

# Linear vs. non-linear classification boundaries



# Random Forests



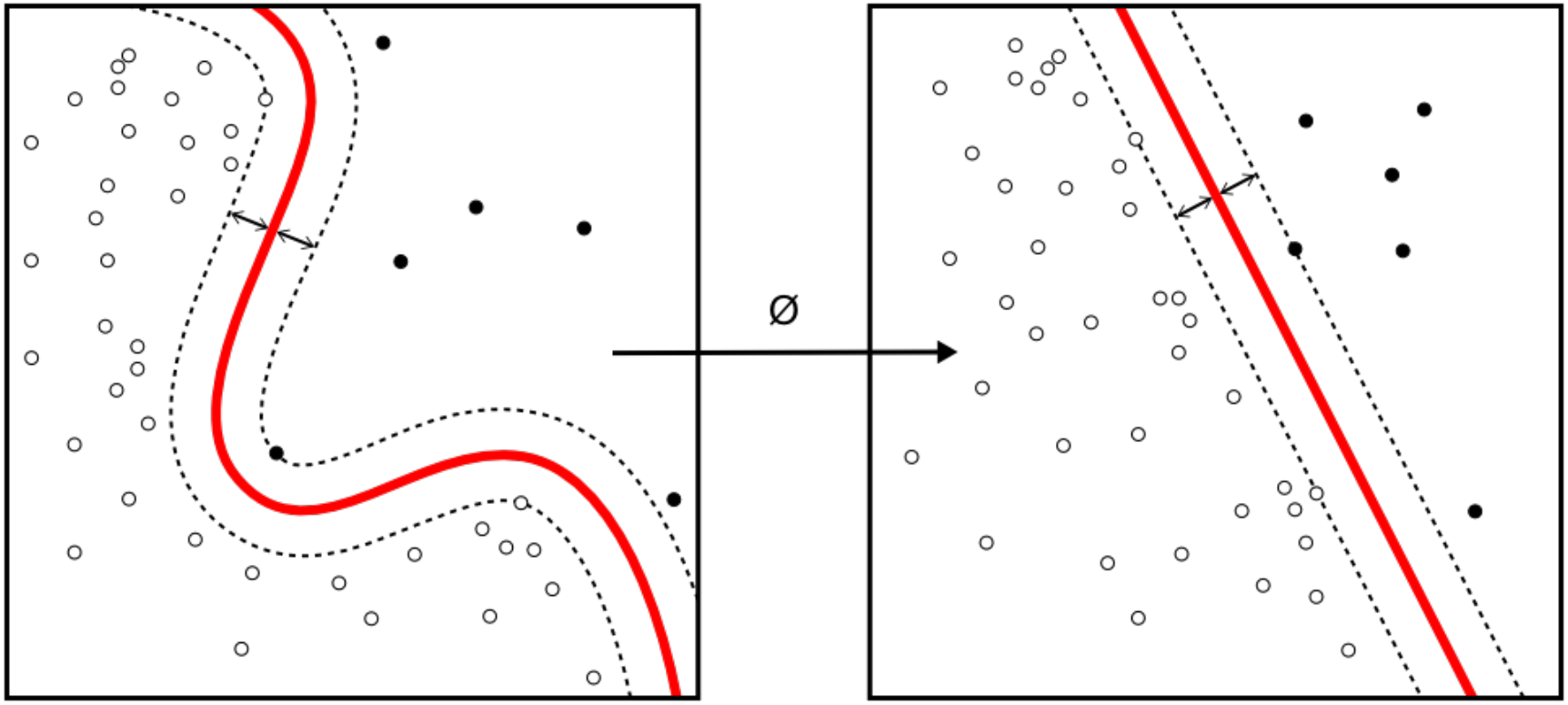

supervised clustering method that learns how to discriminate groups based on decision trees

classify observations into large groups based on predictor values (OTU abundances)

high classification accuracy, can be used for feature extraction (predictor OTUs)

prone to over-fitting, requires cross-validation

# Support Vector Machines



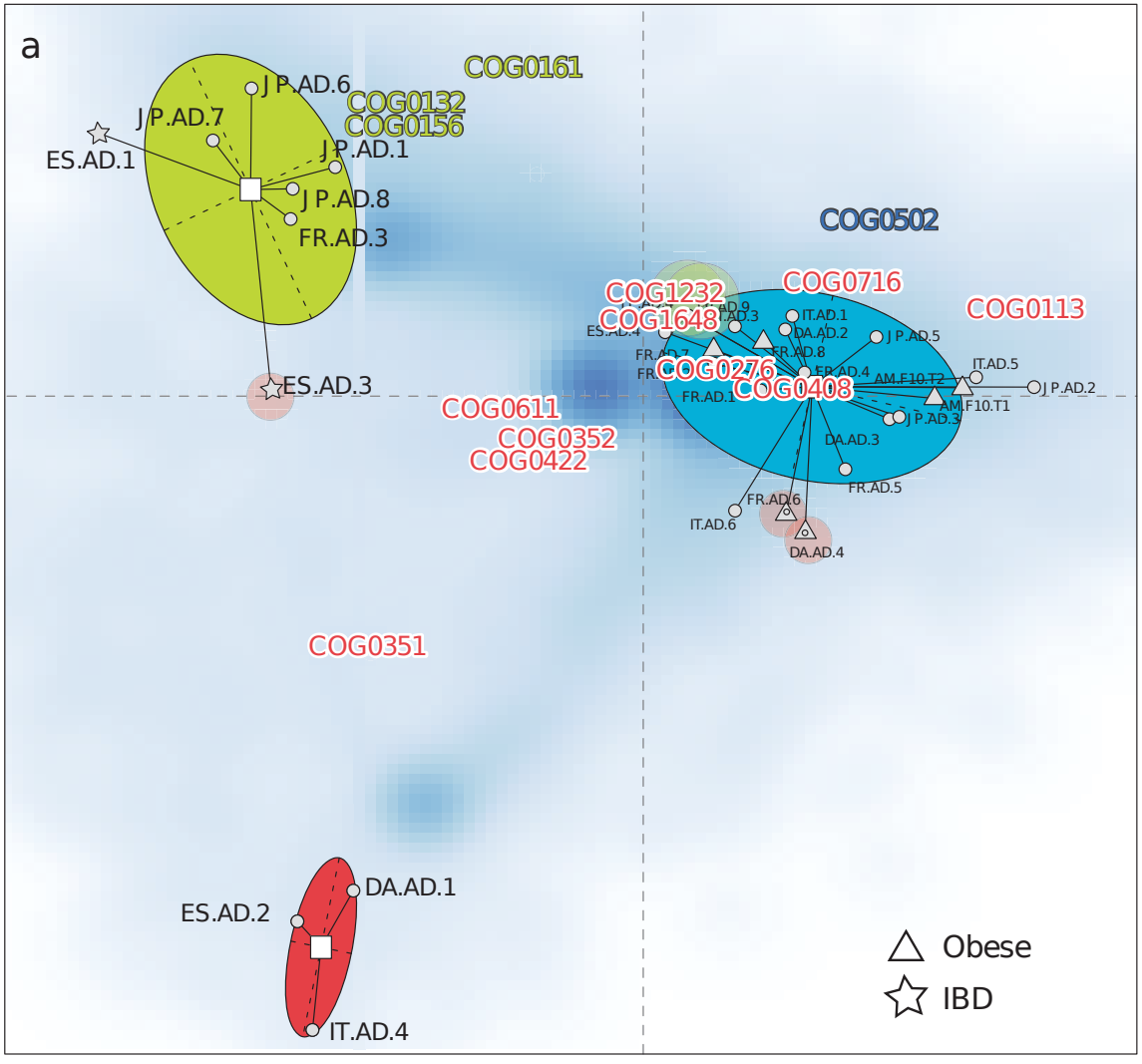

supervised clustering algorithm that attempts to separate groups by maximizing distances to a boundary (margin)

high accuracy; can be used to extract reliable features (predictor OTUs)

prone to over-fitting, requires cross-validation

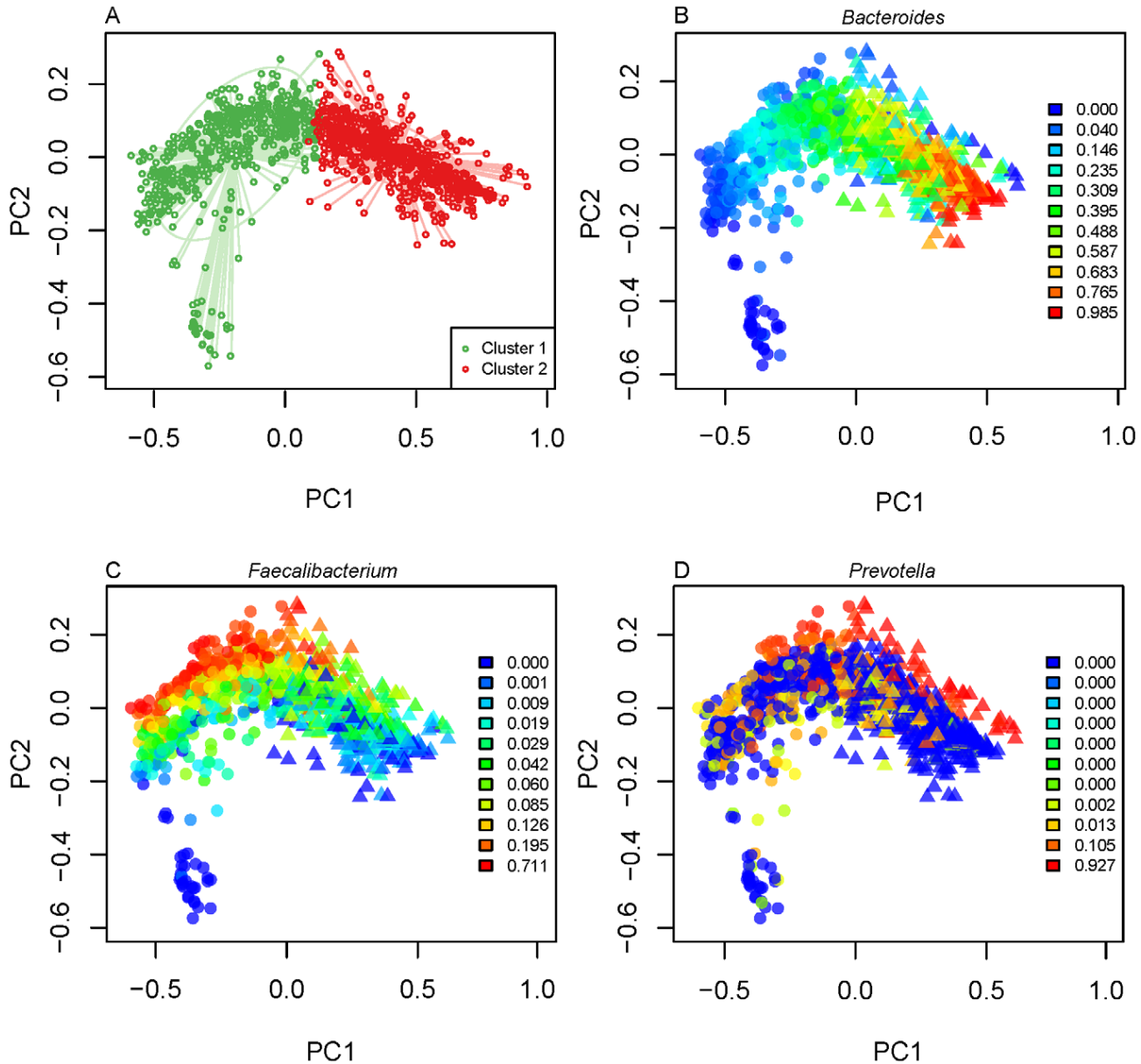
not yet widely employed in microbial ecology

# Data exploration vs. classification




Enterotypes of the human gut microbiome
(Arumugam *et al.*, 2011)

# Data exploration vs. classification




Enterotypes of the human gut microbiome
(Koren *et al.*, 2013)