Introduction to Computational Ecology

Gail Rosen

Outline

- Brief History
- Diversity Measures
  - Richness
  - Evenness
  - Diversity Indices
- Visualization techniques
  - Dimension reduction through distances
  - Scaling
- Statistical Significance

An (abbreviated) history

Numerical ecology

phenetics and statistical analysis of organismal counts

macroecology

16S rRNA gene era

sequence analysis as a surrogate for counting

mapping of marker to taxonomy

WGS (Whole genome shotgun):
synthesis of phylogenomics, functional genomics, and numerical ecology

Metaranscriptomics: functional genomics, gene expression phylogenetics, numerical ecology

This article is part of the series The Genomic Standards Consortium and beyond: best practice in genomics research.

Technical Note

Open Access

The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome

Daniel McDonald[1], **Jose C Clemente**[2], **Justin Kuczynski**[3], **Jai Ram Rideout**[4], **Jesse Stombaugh**[2], **Doug Wendel**[2], **Andreas Wilke**[5], **Susan Huse**[6], **John Hufnagle**[6], **Folker Meyer**[5], **Rob Knight**[1,2,7] and **J Gregory Caporaso**[4,5]*

* Corresponding author: J G Caporaso gregcaporaso@gmail.com

▸ Author Affiliations

1 Biofrontiers Institute, University of Colorado, Boulder, CO, USA

2 Department of Chemistry & Biochemistry, University of Colorado, Boulder, CO, USA

3 Second Genome, San Bruno, CA, USA

4 Department of Computer Science, Northern Arizona University, Flagstaff, AZ, USA

5 Argonne National Laboratory, Argonne, IL, USA

6 Marine Biological Laboratory, Woods Hole, MA, USA

7 Howard Hughes Medical Institute, Boulder, CO, USA

For all author emails, please log on.

GigaScience 2012, **1**:7 doi:10.1186/2047-217X-1-7

The electronic version of this article is the complete one and can be found online at:

Biological Observation Matrix

BIOM file format (MacDonald et al. 2012)

- Standard recognized by EMP, MG-RAST, VAMPS
- Based on JSON data interchange format
  - Computational structure in multiple languages

"facilitates the efficient handling and storage of large, sparse biological contingency tables"

Encapsulates metadata and contingency table (e.g., OTU table) in one file

http://biom-format.org/documentation/format_versions/biom-1.0.html#example-biom-files

Our Observations

Samples

Species/Gene/Protein Family/Metabolic Pathway

Ecological data matrix.

	Descriptors						
Objects	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	···	$\mathbf{y}_j$	···	$\mathbf{y}_p$
$\mathbf{x}_1$	y_{11}	y_{12}	y_{13}	···	y_{1j}	···	y_{1p}
$\mathbf{x}_2$	y_{21}	y_{22}	y_{23}	···	y_{2j}	···	y_{2p}
$\mathbf{x}_3$	y_{31}	y_{32}	y_{33}	···	y_{3j}	···	$3p$
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
$\mathbf{x}_i$	y_{i1}	y_{i2}	y_{i3}	···	y_{ij}	···	y_{ip}
.	.	.	.		.		.
.	.	.	.		.		.
.	.	.	.		.		.
$\mathbf{x}_n$	y_{n1}	y_{n2}	y_{n3}	···	y_{nj}	···	y_{np}

Calculate Distances between samples

Sites	Species		
	y_1	y_2	y_3
x_1	0	1	1
x_2	1	0	0
x_3	0	4	4

From these data, the following distances are calculated between sites:

Sites	Sites		
	x_1	x_2	x_3
x_1	0	1.732	4.243
x_2	1.732	0	5.745
x_3	4.243	5.745	0

Alpha Diversity: Hill's generalized diversity index

cases. For a community consisting of S species with relative abundances $p_1, p_2, \ldots, p_S$, the Hill diversities are defined by

$$D_\alpha = \left(\sum_{i=1}^{s} p_i^\alpha \right)^{\frac{1}{1-\alpha}}.$$

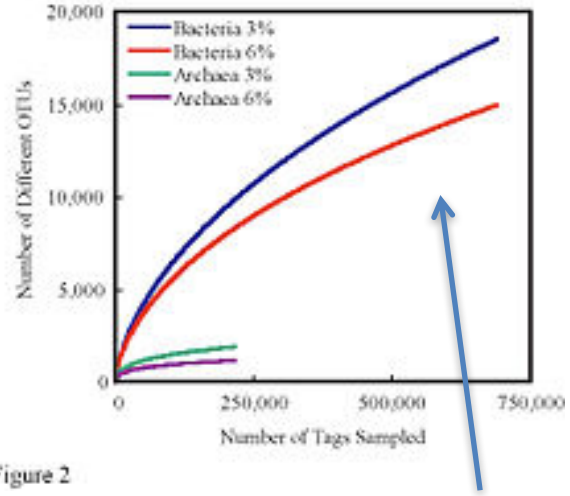


Figure 2

α =0, then D_0= Species Richness

D_0 =S (Each species weighted equally)

Alpha>0, then rare species contribute less

Harder to get all species so larger alpha is easier to estimate

http://www.nature.com/ismej/journal/v7/n6/full/ismej201310a.html

Chao1 estimates lower bound on D_0 (Assume unobserved frequency similar to rarest observed)

$$S_{chao1} = S_{obs} + \frac{n_1 (n_1 - 1)}{2 (n_2 + 1)}$$

where,

S_{chao1} = the estimated richness

S_{obs} = the observed number of species

n_1 = the number of OTUs with only one sequence (i.e. "singletons")

n_2 = the number of OTUs with only two sequences (i.e. "doubletons")

D_2 and Simpson's index

$$D_2 = \left(\sum_i^S p_i^2 \right)^{-1} = \frac{1}{\sum_i^S p_i^2}$$

Simpson's (1949) original index ($1/D_2$) is a measure of dominance rather than diversity

The complement of Simpson's index of dominance is

$$Diversity = 1 - \sum_i^S p_i^2$$

and is a measure of diversity. It is the likelihood that two randomly chosen individuals will be different species.

D_1 and Shannon-Wiener index

If alpha = 1 then D_1 is a nonsense equation because the exponent is 1/0. But if we use limits to define D_1 as *alpha* approaches 1 then

$$D_1 = \lim \alpha \rightarrow 1 \left(D_a \right)$$

$$D_1 = \log^{-1} \left(- \sum_{i}^{S} p_i \log p_i \right)$$

The logarithmic form of D_1 is the **Shannon-Wiener** index (H'), which measures the "information content" of a sample unit:

$$H' = \log(D_1) = -\sum_{i}^{S} p_i \log p_i$$

The units for H' are the number of "bits" encoded with species of various relative abundances. (H' is highest for equal abundances)

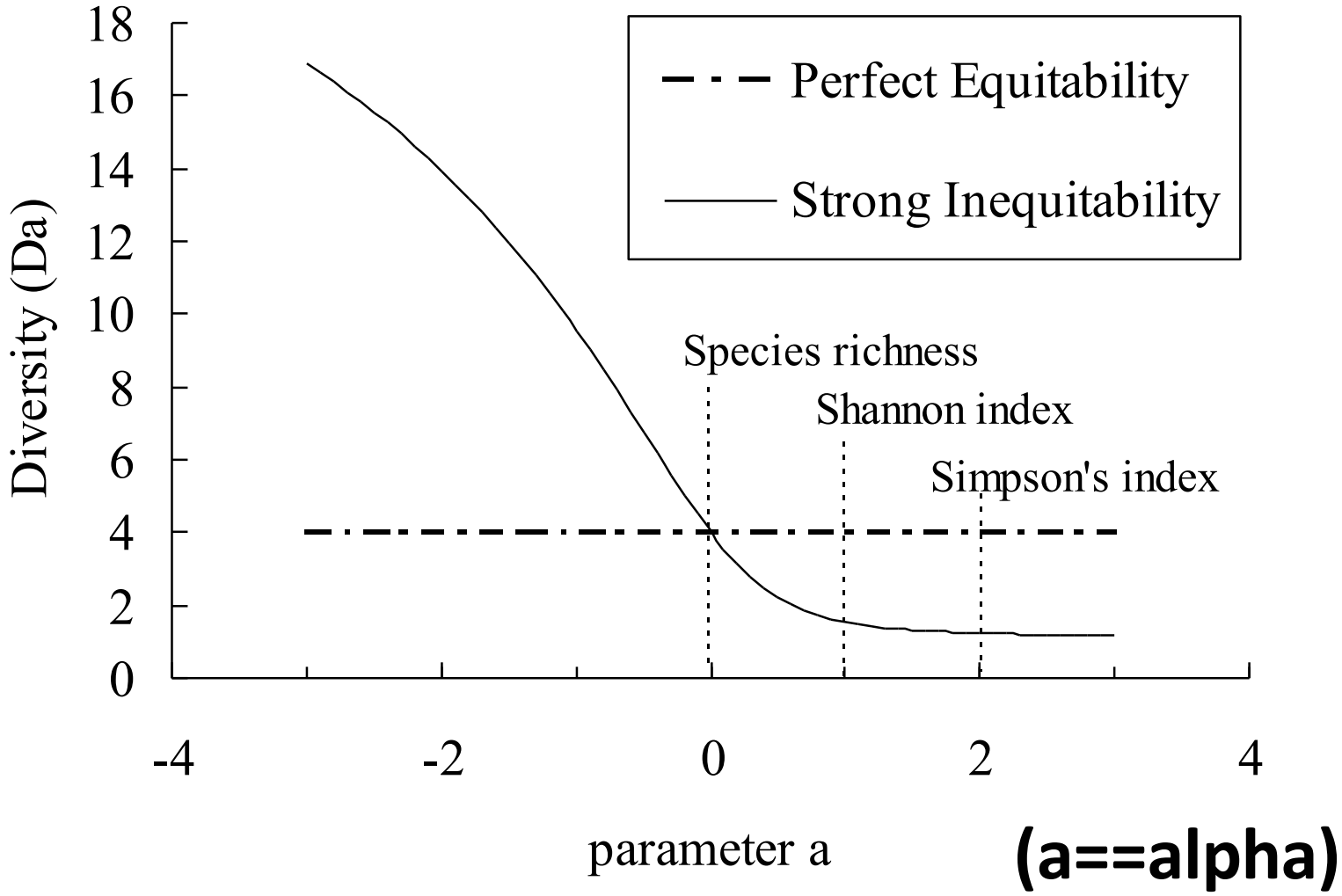


Figure 4.1. Influence of equitability on Hill's (1973a) generalized diversity index. Diversity is shown as a function of the parameter *a* for two cases: a sample unit with strong inequitability in abundance and a sample unit with perfect equitability in abundance (all species present have equal abundance; see Table 4.1).

Evenness

An easy-to-use measure (Pielou 1966, 1969) is "**Pielou's *J***"

$$J = \frac{H'}{\log S}$$

where

H′ is the Shannon-Wiener diversity measure

S is the average species richness.

If there is perfect equitability then log(*S*) = *H′* and *J* = 1.

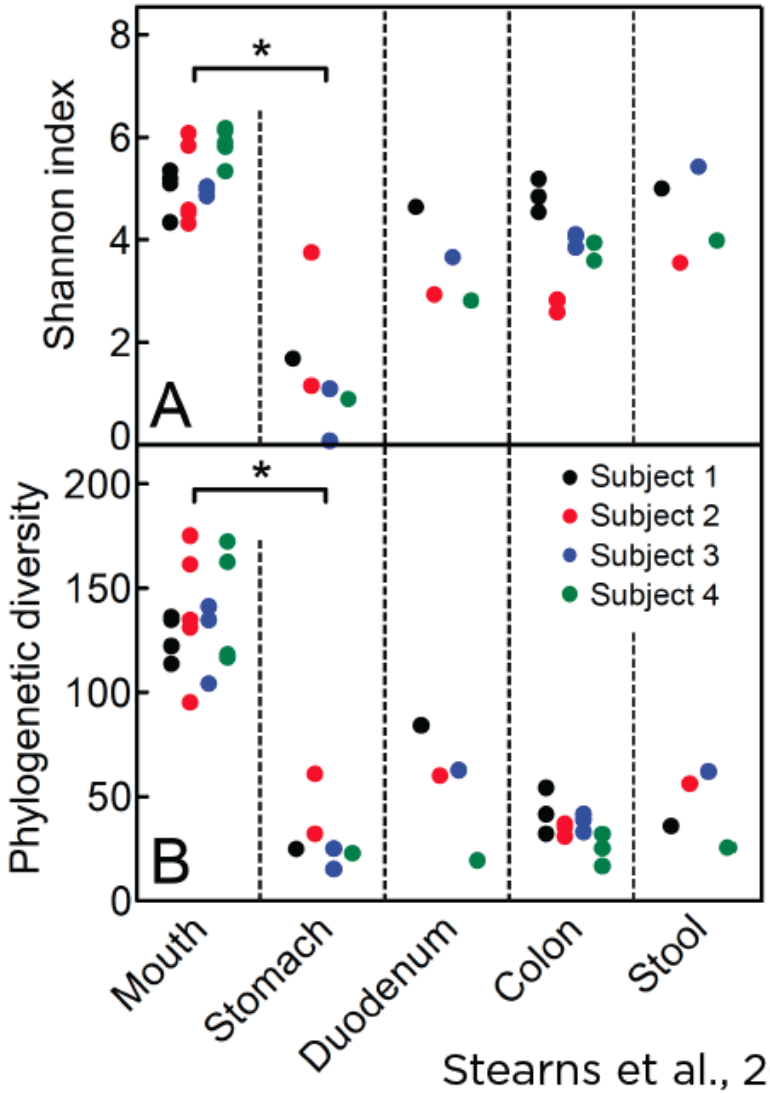
Who lives with whom, and why, and where?

Data reduction is essential for:
a) summarizing large numbers of observations into manageable numbers
b) visualizing many interconnected variables in a compact manner

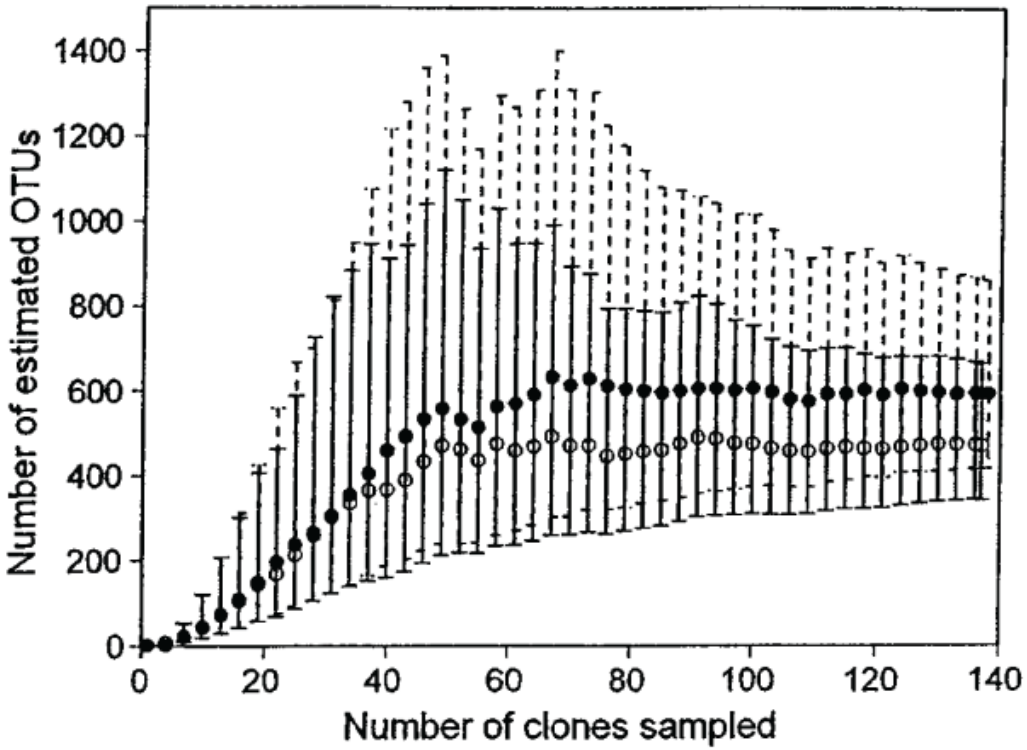
Alpha diversity: species richness (and evenness) within a single sample
Beta diversity: change in species composition across a collection of samples
Gamma diversity: total species richness across an environmental gradient

α-diversity: Richness and Evenness

Shannon index (H'), Estimators (Chao1, ACE), Phylogenetic Diversity



Stearns et al., 2011



Hughes et al., 2001

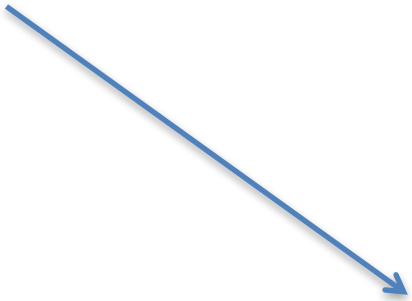
Shannon index (H'): richness and evenness
Estimators: richness
Faith's PD: phylogenetic richness

Visualization (ordination)

- Complementary to data clustering
  - looks for discontinuities
- Ordination extracts main trends as continuous axes
  - analysis of the square matrix derived from the OTU table
- Non-parametric, unconstrained ordination methods most widely used (and best suited)
  - methods that can work directly on a square matrix
- An appropriate metric is required to derive this square matrix
  - many options...

Dimension Reduction

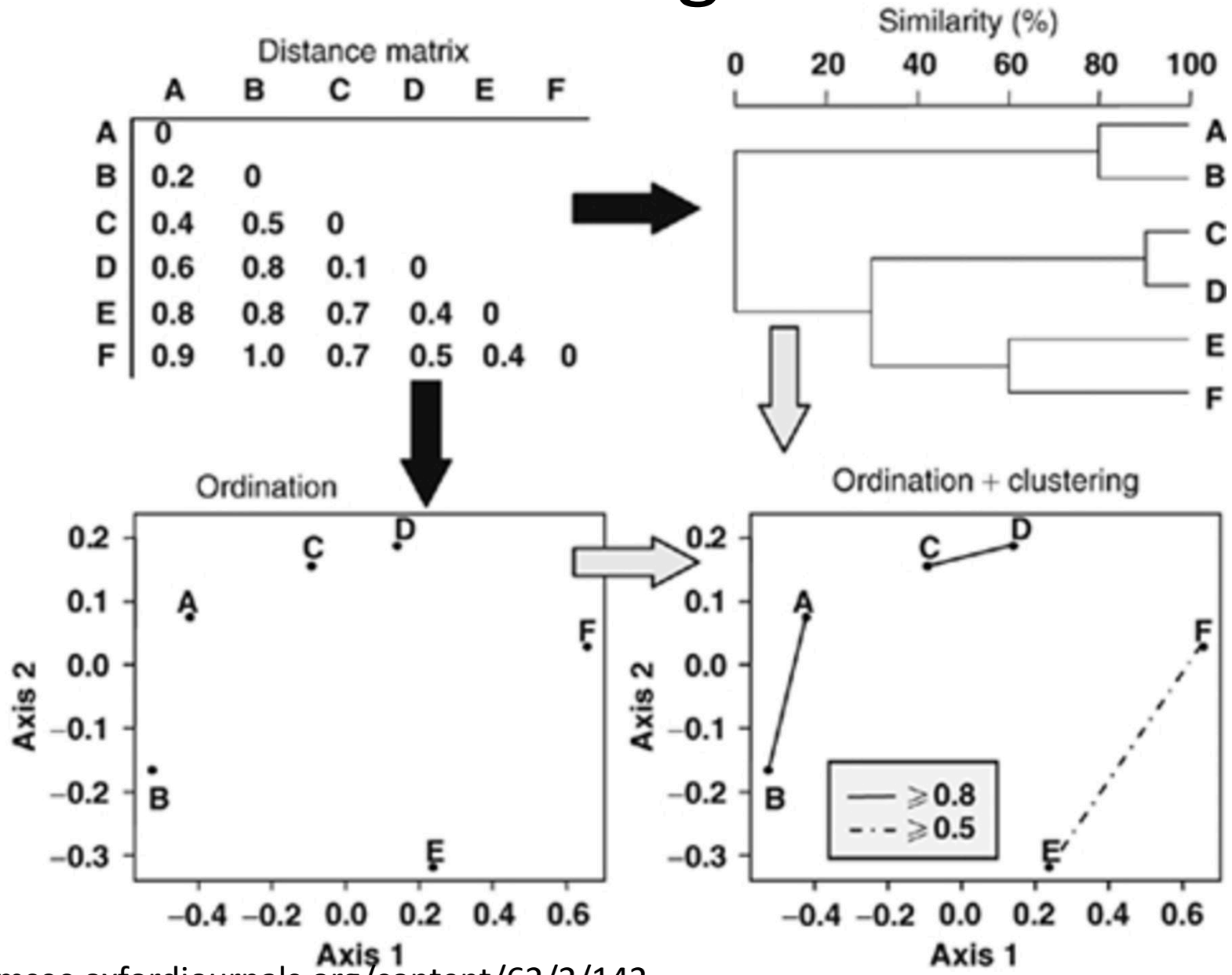
OTU	SampleA	SampleB	SampleC
OTU_1	0	12	8
OTU_2	1	22	0
OTU_3	6	0	2



Distance Matrix

	A	B	C
A	0		
B	0.2	0	
C	0.4	0.5	0

Scaling



http://femsec.oxfordjournals.org/content/62/2/142

Distance measures used

- Euclidean
- Hellinger

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

$$D(X, Y) = \sqrt{\sum_{i=1}^{n} \left(\sqrt{\frac{x_i}{\hat{x}}} - \sqrt{\frac{y_i}{\hat{y}}} \right)^2}, \text{ with } \hat{y}$$

$$= \sum_{i=1}^{n} y_i \qquad (5)$$

- Bray-Curtis

$$D(X, Y) = 1 - 2 \frac{\sum_{i=1}^{n} \min(x_i, y_i)}{\sum_{i=1}^{n} (x_i + y_i)} \qquad (3)$$

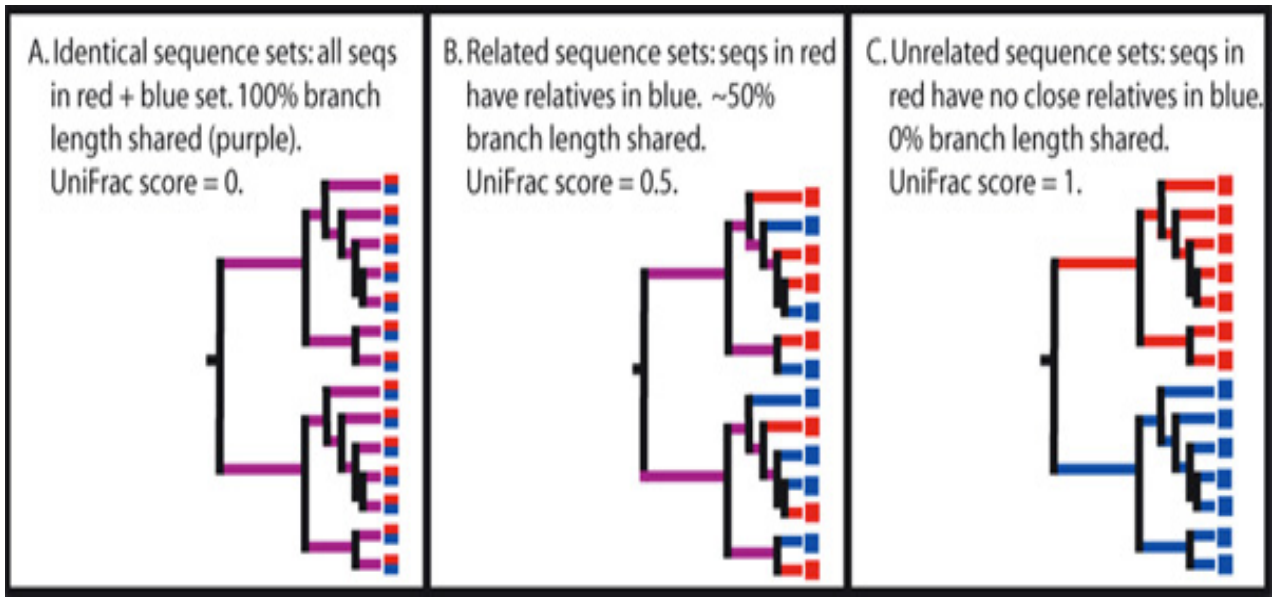
Not a true distance if does not satisfy triangle inequality

http://www.nature.com/ismej/journal/v4/n10/full/ismej201051a.html

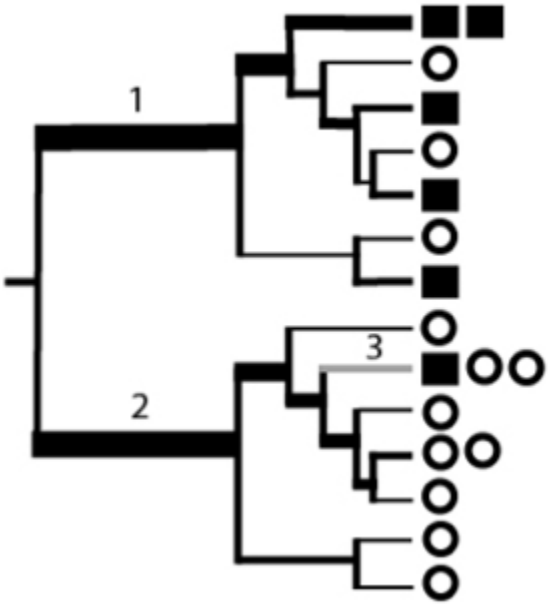
Game Changer – Distance based on Phylogeny

- Unifrac

Unweighted



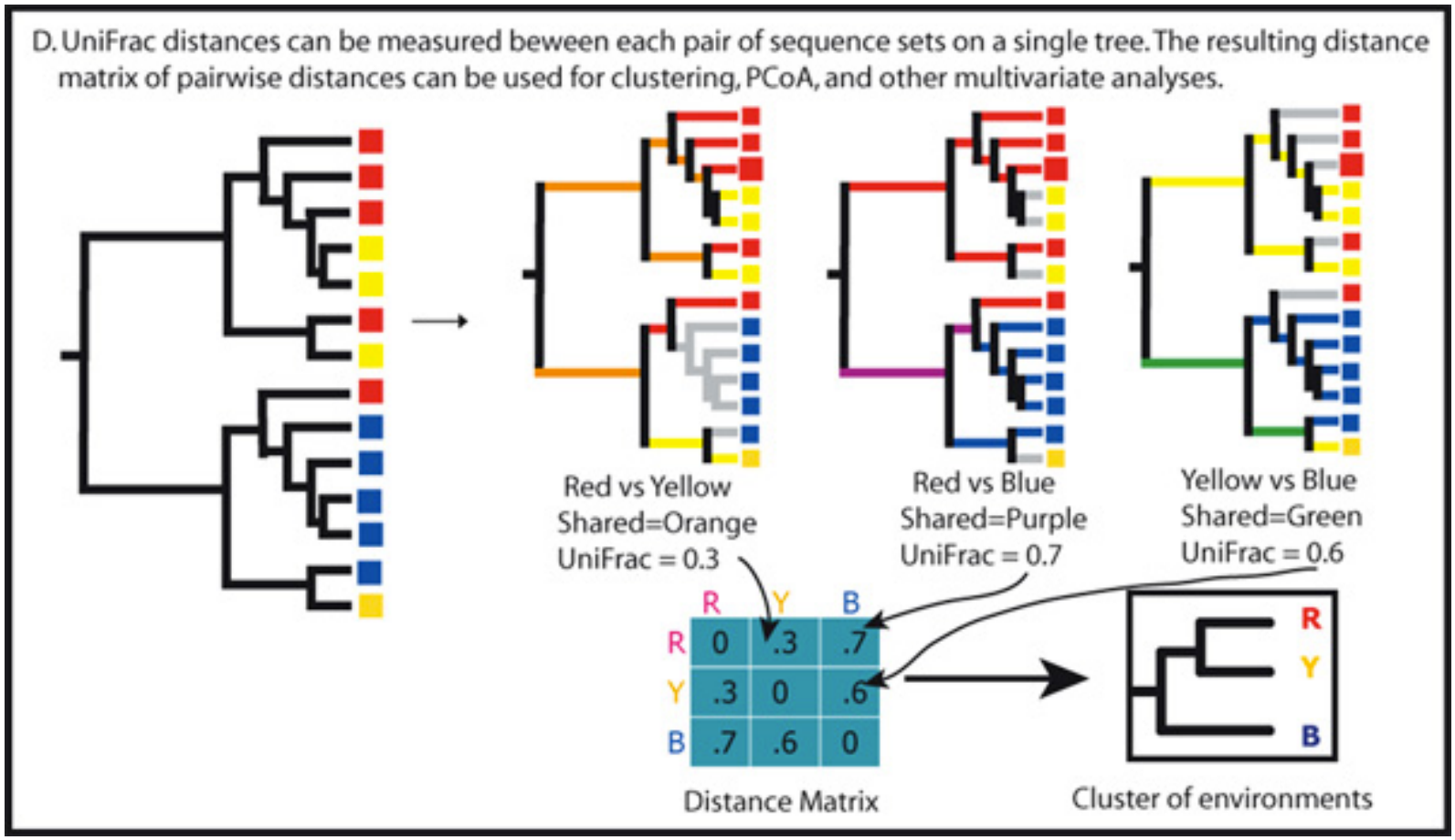
Weighted



Related to Kantorovich –Rubinstein Earth-mover distance:

https://liorpachter.wordpress.com/2013/09/18/unifrac-revealed/

Unifrac


D. UniFrac distances can be measured beween each pair of sequence sets on a single tree. The resulting distance matrix of pairwise distances can be used for clustering, PCoA, and other multivariate analyses.
Red vs Yellow
Shared=Orange
UniFrac = 0.3
Red vs Blue
Shared=Purple
UniFrac = 0.7
Yellow vs Blue
Shared=Green
UniFrac = 0.6
R
Y
B
R 0 .3 .7
Y .3 0 .6
B .7 .6 0
Distance Matrix
R
Y
B
Cluster of environments

Ordination example 1 (of many):

Principal Coordinates Analysis

Classical Multidimensional Scaling (MDS; Gower 1966)

Procedure:

- based on eigenvectors
- position objects in low-dimensional space while preserving distance relationships as well as possible

highly flexible

- can choose among many association measures

In microbial ecology, used for visualizing phylogenetic or count-based distances

Consistent visual output for given distance matrix

Include variance explained (%) on Axis 1 and 2

Ordination example 2 (of many):

Non-metric Multidimensional Scaling

Ordination not based on eigenvectors

Does not preserve exact distances among objects

- attempts to preserve ordering of samples ("ranks")

Procedure:

- iterative, tries to position the objects in a few (2-3) dimensions in such a way that minimizes the "stress"
- how well does the new ranked distribution of points represent the original distances in the association matrix? Can express as R^2 on axes 1 and 2.
- the adjustment goes on until the stress value reaches a local minimum (heuristic solution)

NMDS often represents distance relationships better than PCoA in the same number of dimensions

Susceptible to the "local minimum issue", and therefore should have strong starting point (e.g., PCoA) or many permutations

- You won't get the same result each time you run the analysis. Try several runs until you are comfortable with the result.

T-SNE

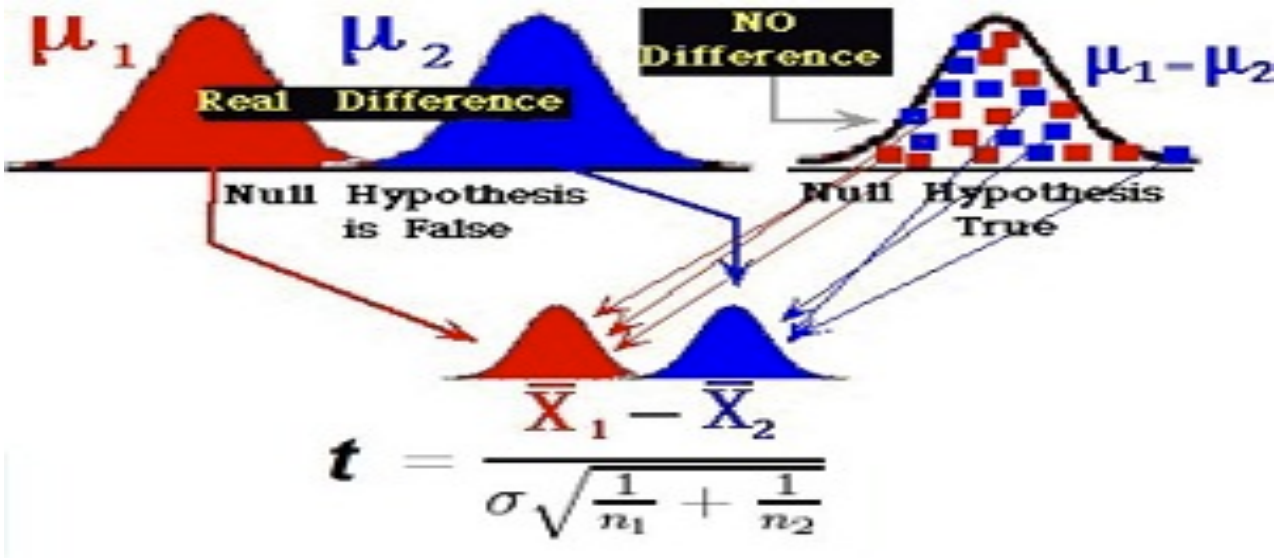
- Another Game Changer
- http://www.cs.toronto.edu/~hinton/absps/tsne.pdf

- http://distill.pub/2016/misread-tsne/

Statistical Tests

- T-test
- ANOVA
- MANOVA
- Permanova/Anosim

T-test



As an example, in the one-sample *t*-test $t = = \frac{Z}{(s/\sqrt{n})} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{(s/\sqrt{n})}$,

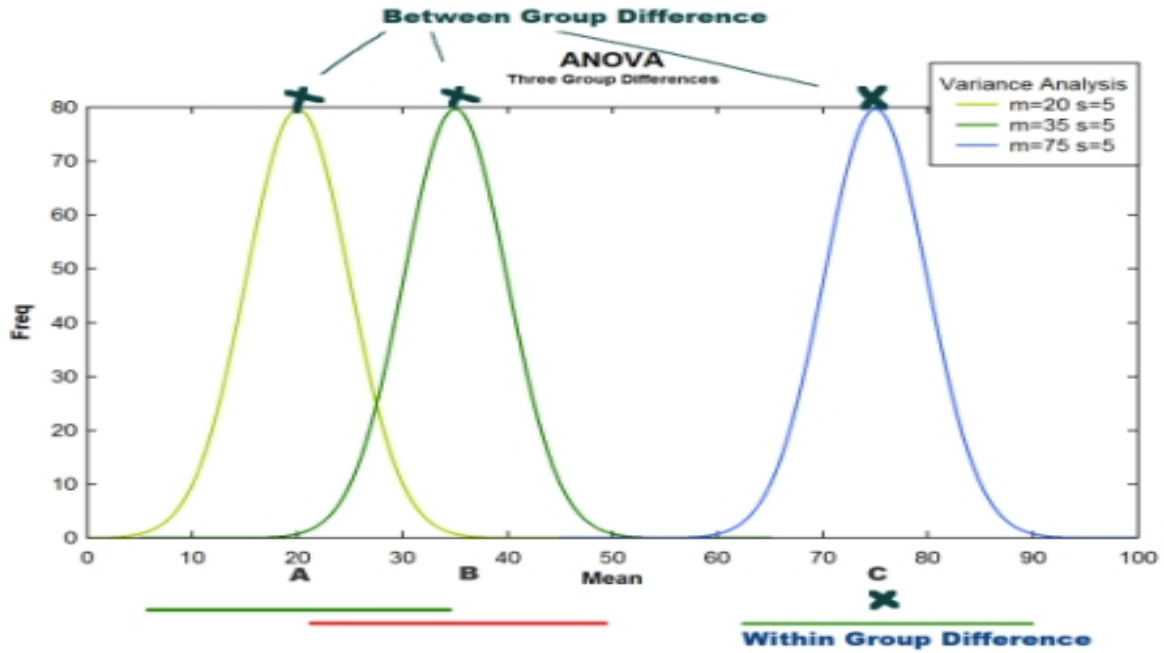
where $\bar{X}$ is the sample mean from a sample $X_1, X_2, \ldots, X_n$, of size n, and s is the sample standard deviation. σ is the population standard deviation of the data.

The assumptions underlying a *t*-test are that

- X follows a normal distribution with mean μ and variance σ^2
- s^2 follows a χ^2 distribution with p degrees of freedom under the null hypothesis, where p is a positive constant
- Z and s are independent.

ANOVA

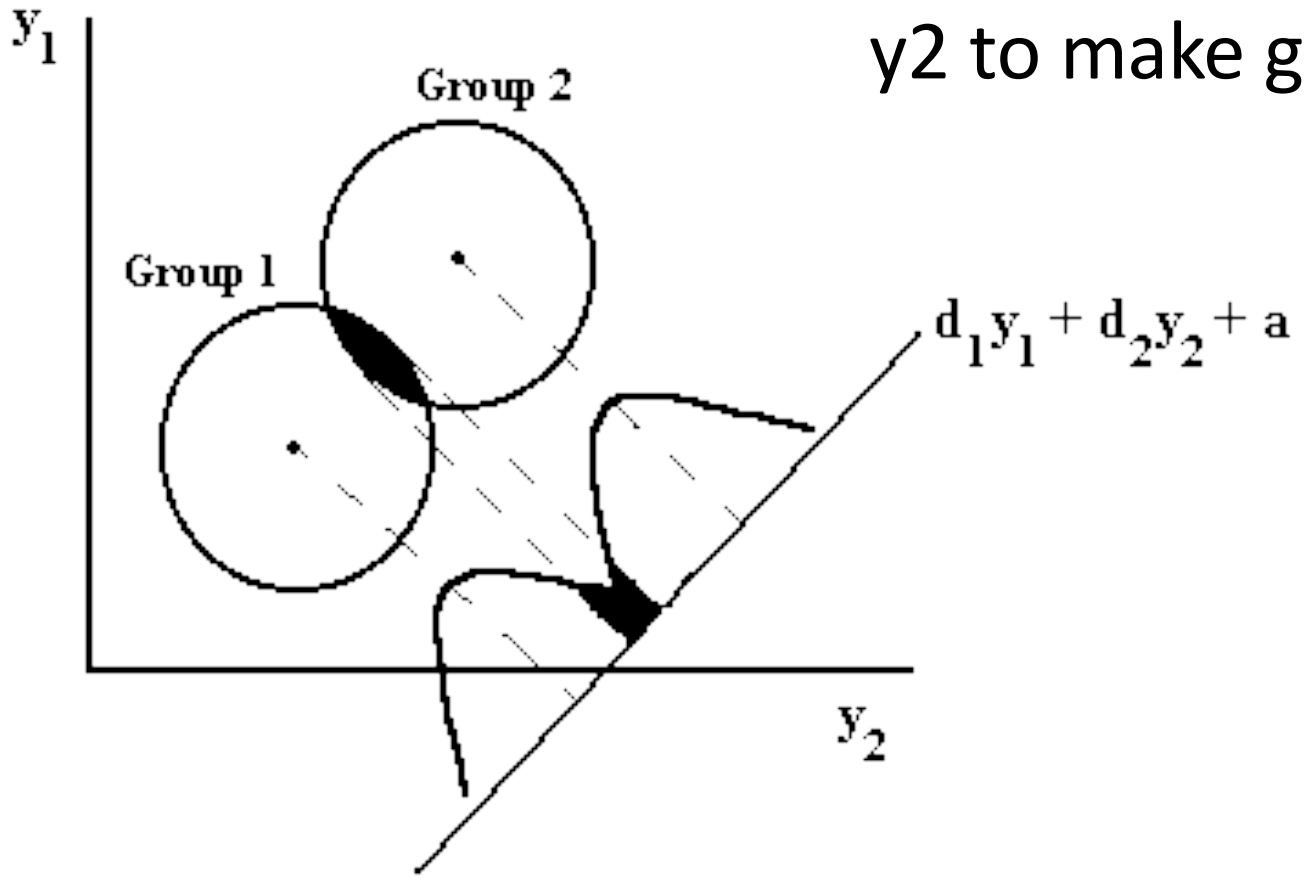
- T-test is only for two distributions



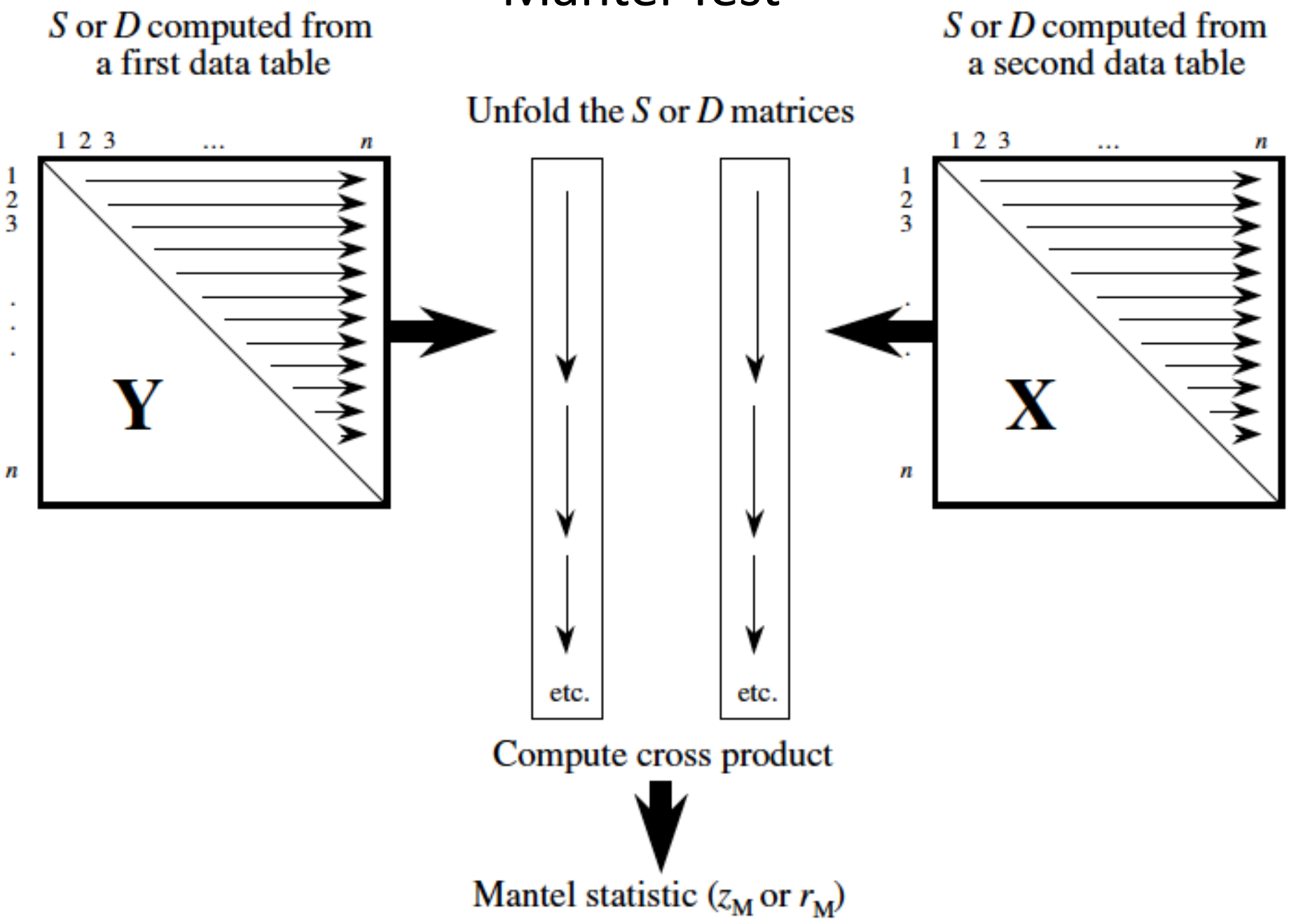
Source of Variation	d.f.	SS	MS	F_0
Factor A (between groups)	a-1	$SSA = \sum_{i=1}^{a} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$	$MSA = \frac{SSA}{(a-1)}$	$\frac{MSA}{MSE}$
Factor B (between groups)	b-1	$SSB = \sum_{j=1}^{b} n_j \left(\bar{y}_{.j} - \bar{y}_{..}\right)^2$	$MSB = \frac{SSB}{(b-1)}$	$\frac{MSB}{MSE}$
Error (within groups)	(a-1)(b-1)	$SSE = SST - SSA - SSB$	$MSE = \frac{SSE}{(a-1)(b-1)}$	
Total	N-1	$SST = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(y_{ij} - \bar{y}_{..}\right)^2$		

MANOVA

* ANOVA with dependent variables
* Combination of y1 and y2 to make groups



Mantel Test

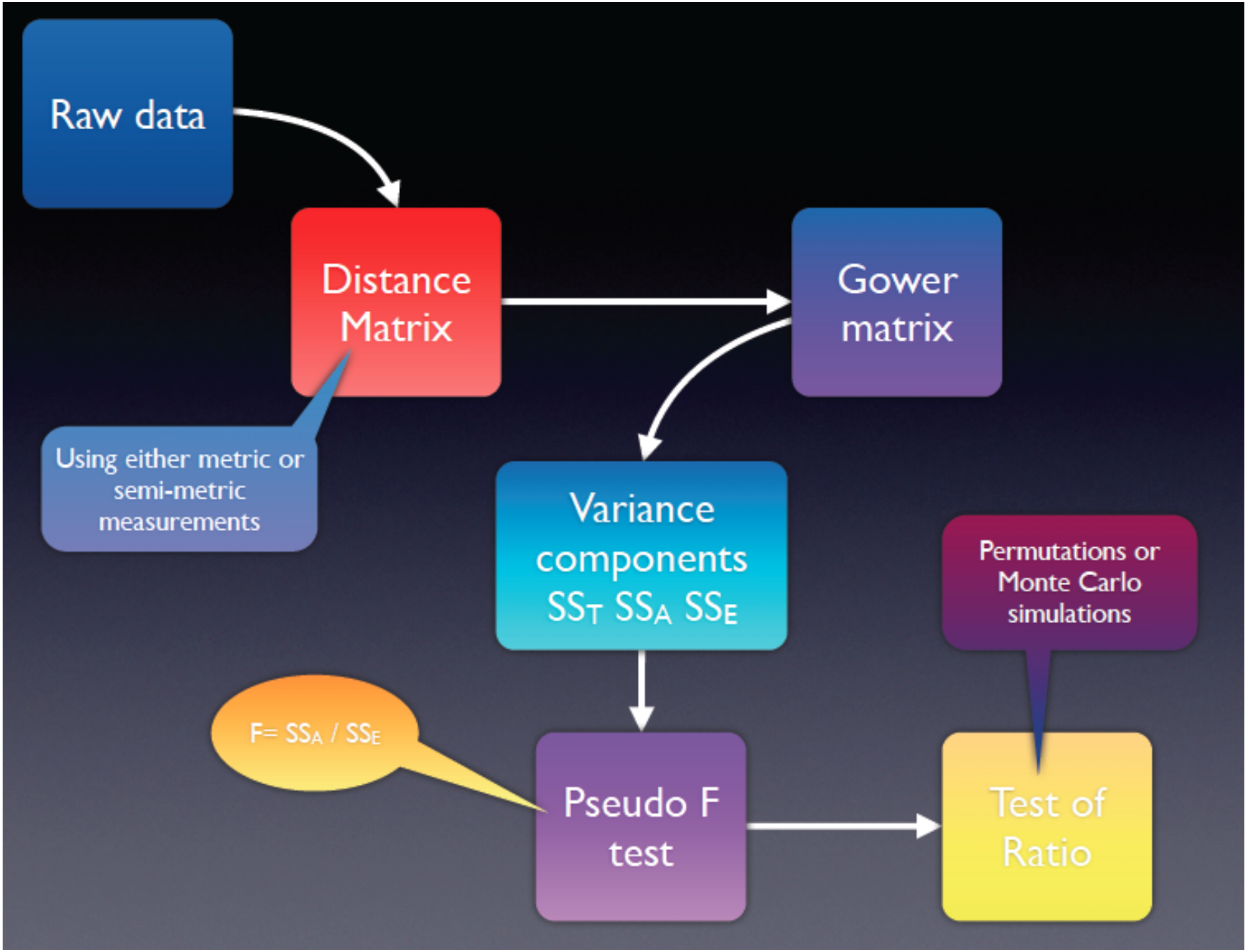

S or D computed from a first data table
S or D computed from a second data table
Unfold the S or D matrices
1 2 3 ... n
1
2
3
n
Y
X
etc.
etc.
Compute cross product
Mantel statistic (z_M or r_M)

What alternatives are out there?

- ***Parametric***
  - MANOVA
  - MANCOVA
- ***Non-parametric***
  - ANOSIM
  - Mantel test
  - Multiple response permutation procedures
  - Distance based redundancy analysis (dbRDA)
  - Multivariate analysis of variance using distance measurements (PERMANOVA)

PERMANOVA

- Name – means “Multivariate Analysis of Variance using Distance Metrics”
- ANOVA/MANOVA is parametric and assumes distributions – PERMANOVA is non-parametric
- Variation can be paritioned into FACTORS
- Distribution-free
- Sensitive to within-group dispersion


Raw data
Distance Matrix
Gower matrix
Using either metric or semi-metric measurements
Variance components SS_T SS_A SS_E
Permutations or Monte Carlo simulations
$F = SS_A / SS_E$
Pseudo F test
Test of Ratio

ANOSIM

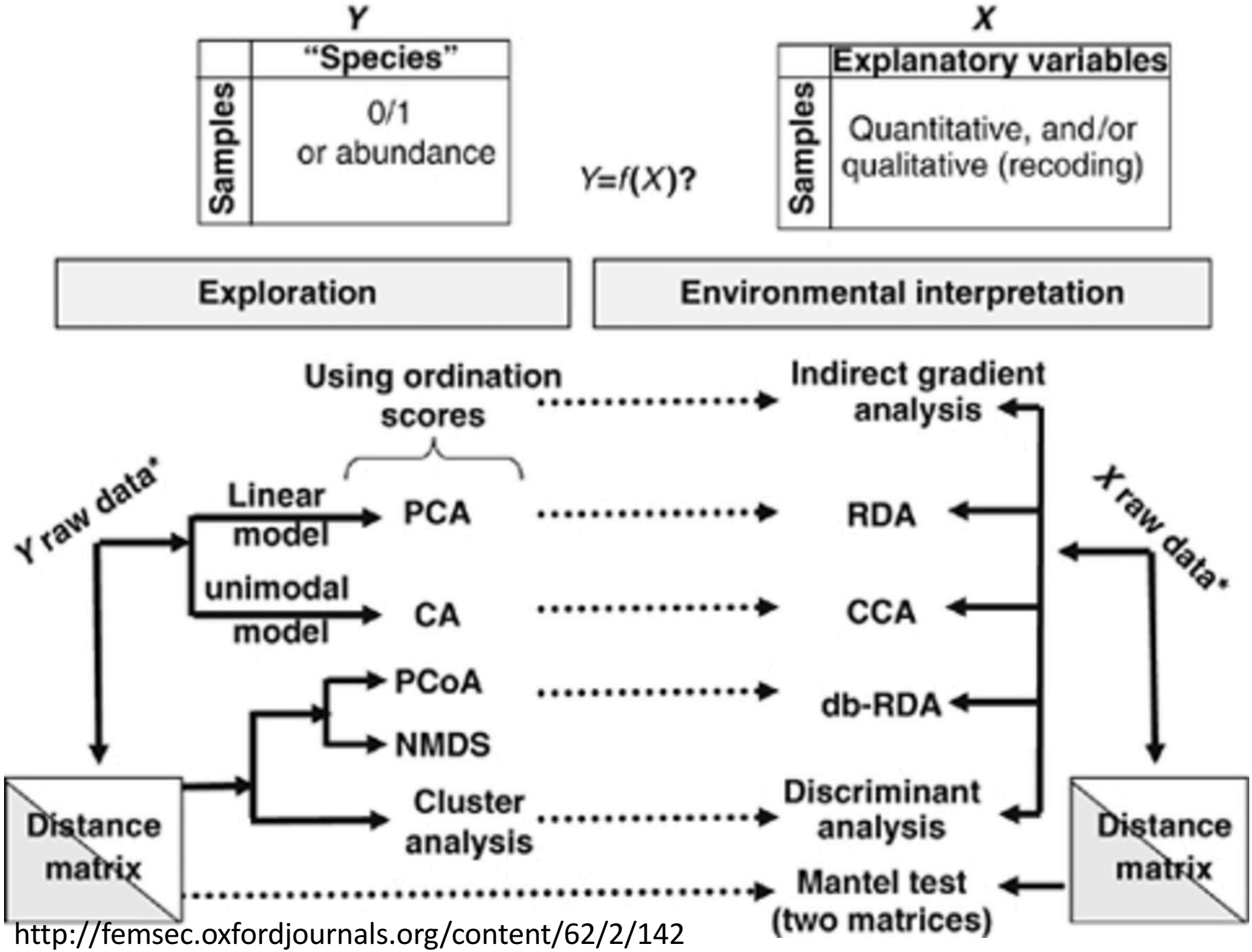
(a) **X** = ranked distances

	Group 1		Group 2		
D	5	6	7	8	9
5					
6	2				
7	4	8.5			
8	6.5	5	1		
9	8.5	10	3	6.5	

$$R = \frac{\bar{r}_B - \bar{r}_W}{n\,(n-1)\,/4} \qquad \textbf{(10.28)}$$

where $\bar{r}_B$ is the mean of the ranks in the *between*-group submatrix (i.e. the rectangle, in Fig. 10.22a, crossing groups 1 and 2), $\bar{r}_W$ is the mean of the ranks in all *within*-group submatrices (i.e. the two triangles in the Figure), and n is the total number of objects. In the present example, $\bar{r}_B = 7.083$ and $\bar{r}_W = 3.125$, so that $R = 0.79167$

Summary



http://femsec.oxfordjournals.org/content/62/2/142

Quantitative descriptors	Semiquantitative descriptors	Qualitative descriptors	Descriptors of mixed precision
Difference between two samples:			
Hotelling T^2	---	Log-linear models	---
Difference among several samples:			
MANOVA	---	Log-linear models	MANOVALS
db-RDA, CCA	---	db-RDA, CCA	db-RDA
Scatter diagram	Rank diagram	Multiway contingency table	Quantitative-rank diagram
Association coefficients R:			
Covariance	---	Information, X^2	---
Pearson *r*	Spearman *r* Kendall τ	Contingency	---
Partial *r*	Partial τ		
Multiple *R*	Kendall *W*		
Species diversity:			
Diversity measures	Diversity measures	Number of species	---
Association coeff. Q	Association coeff. Q	Association coeff. Q	Association coeff. Q
Clustering	Clustering	Clustering	Clustering
Ordination:			
Principal component a. Correspondence a. Principal coordinate a. Nonmetric multi-dimensional scaling	---	Correspondence a. HOMALS	PRINCALS PRINCIPALS Principal coordinate a. Nonmetric multi-dimensional scaling ALSCAL, GEMSCAL
Factor analysis	---	---	FACTALS
Regression simple linear (I and II) multiple linear polynomial partial linear nonlinear, logistic smoothing (splines, LOWESS) multivariate; see also canonical a.	Regression nonparametric	Correspondence	Regression logistic dummy MORALS
Path analysis	---	Log-linear models Logit models	PATHALS
Canonical analysis:			
Redundancy analysis (RDA)			CORALS, OVERALS
Canonical correspondence a. (CCA)		CCA	db-RDA
Canonical correlation a. (CCorA)			
Discriminant analysis	---	Discrete discriminant a. Log-linear models	CRIMINALS Logistic regression