Universitas Carolina
Charles University in Prague

Matrix Visualization:
Approaching Statistics and Statistical Approach

Lecture 1: General Introduction

Chun-houh Chen

Institute of Statistical Science Academia Sinica

April 10, 2013

Matrix Visualization:

Approaching Statistics and Statistical Approach

Lecture 1: General Introduction

Chun-houh Chen

Institute of Statistical Science Academia Sinica

April 10, 2013

Jaromír ANTOCH

Jaromír ANTOCH

COMPSTAT 2004, 16th Symposium of IASC
PRAGUE, August 23-27, 2004

Matrix Visualization and Information Mining

Chun-houh Chen

Institute of Statistical Science
Academia Sinica
Taipei, Taiwan

COMPSTAT 2004, 16th Symposium of IASC
PRAGUE, August 23-27, 2004

Generalized Association Plots with a Covariate Adjustment

Han-Ming Wu and Chun-Houh Chen
Taipei, Taiwan

Abstract

Generalized association plots (GAP)(Chen, 2002) is a dimension-free information visualization environment for exploring the association of subjects/variables embedded in the data. When effects of covariates such as gender or age are of interest in the analysis, covariates adjustment has to be taken into account. This article focuses on extending the framework by incorporating covariate variables through the estimation of conditional correlations.

For a discrete covariate, we can estimate the conditional correlation directly or alternatively perform the within and between analysis (WABA) (Dansereau, Alutto and Yammarino, 1984) to investigate the conditional association structure. The WABA procedure decomposes a correlation matrix into the within- and between-component matrices. The contribution of the covariate effects can then be studied from the between-component relative to the original correlation matrix. While the within components act as residuals.

When the covariate is of continuous nature, the conditional correlation is equivalent to the partial correlation under the assumption of joint normal distribution (Kurowicka and Cooke, 2000). Another possibility is to discretize the continuous variable into several groups and treat it as a discrete one. Simulation and empirical studies of the proposed methods will be reported and discussed.

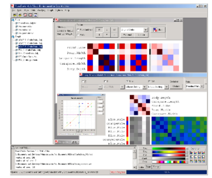
Data Visualization

Visualization = Graphing (for DATA)
+ Fitting
+ Graphing (for MODEL)

- Exploiting the human visual system to extract information from data.
- Provides an overview of complex data sets.
- Identifies structure, patterns, trends, anomalies, and relationships in data.
- Assists in identifying the areas of interest.

Software: *VisualEnvir*

A Visual Environment for Data Analysis



VisualEnvir a java-designed software for exploratory data analysis.

Main Features
- Partial correlation analysis.
- Within And Between Analysis (WABA).
- Sliced Inverse Regression (SIR).
- Principle Components Analysis (PCA).
- Generalized association plots with ellipse seriation and various display conditions.

Han-Ming Wu
hmwu@stat.sinica.edu.tw
http://www.sinica.edu.tw/~hmwu/
Institute of Statistical Science, Academia Sinica
中央研究院 統計科學研究所

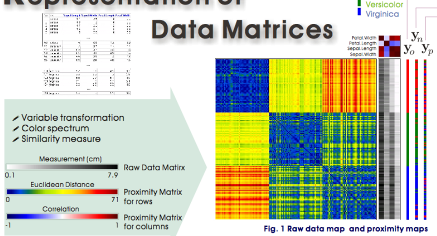
Fisher's Iris Data

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis.

The sepal length, sepal width, petal length, and petal width are measured in centimeters on fifty iris specimens from each of three species, Iris setosa, I. versicolor, and I. Virginica.

Images source: http://www.stat.auckland.ac.nz/~ihaka/120

Representation of Data Matrices



- Variable transformation
- Color spectrum
- Similarity measure

Fig. 1 Raw data map and proximity maps

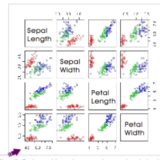


Fig. 3 Colored by observed species labels

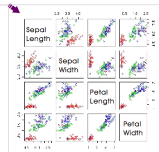


Fig. 4 Colored by Noise species labels

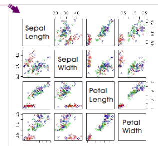


Fig. 5 Colored by permuted species labels

ScatterPlot Matrices of four variables with colors by observed group labels (Y_o), noise group labels (Y_n) and permuted group labels (Y_p).

It is informative to visualize the distribution of species covariate by imposing distinguishable colors (Fig 3). However, when the groups labels are perturbed with noise or permutation, the clusters information would be distorted for the visualization aspect (Fig 4 & 5).

Generalized Association Plots

(Chen 2002)

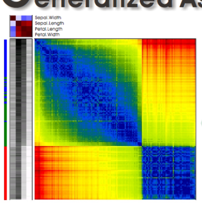


Fig. 2 Sorted data map and proximity maps

FOUR STEPS

1. **Raw data map** and proximity maps with suitable color projection.
2. **Sorted data map** and proximity maps with principle of geometry.
3. **Partitioned data map** and proximity maps with near stationary iterations.
4. **Sufficient graphs** With three linkage for a multivariate data set.

Fig 2. shows the sorted data matrix and proximity matrices by ellipse seriation. Iris setosa is linearly separable from the other 2 while the latter are not linearly separable from each other.

Methods

The covariate adjustment is implemented by the estimate of the conditional correlation. The main idea is working on the residuals instead of raw data. The total correlation matrix ($\rho(X_i, X_j)$) can be decomposed into model and residual component matrices. The contribution of the covariate effects can then be studied from the model component relative to the original correlation matrix.

Simultaneously, the raw data matrix can decomposed into fitted data matrix and residual data matrix. The conditional correlation matrix is the correlation of the model data. The covariate adjustment is achieved by calculating the correlation matrix of the residual data.

For a discrete covariate, we can estimate the conditional correlation by group means. This is equivalent to perform the within and between analysis (WABA) (Dansereau, Alutto and Yammarino, 1984). The between component is the model component and the within component acts as residuals.

For a continuous covariate, conditional correlation can be modeled by simple linear regression. The partial correlation is then a part of residual component.

Decomposition of Pearson Product-Moment Correlation

$$\rho(X_i,X_j) = \frac{Cov(X_i,X_j)}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}$$
$$= \frac{\sqrt{Var(E[X_i|Y])}\sqrt{Var(E[X_j|Y])}}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}\frac{Cov(E[X_i|Y]),E[X_j|Y])}{\sqrt{Var(E[X_i|Y])}\sqrt{Var(E[X_j|Y])}}$$
$$+ \frac{\sqrt{E[Var(X_i|Y)]}\sqrt{E[Var(X_j|Y)]}}{\sqrt{Var(X_i)}\sqrt{Var(X_j)}}\frac{E[Cov(X_i|Y,X_j|Y)]}{\sqrt{E[Var(X_i|Y)]}\sqrt{E[Var(X_j|Y)]}}$$

Component	Data	=	Model	+	Residuals
Correlation	$\rho(X_i,X_j)$		$\rho(E[X_i\|Y],E[X_j\|Y])$		$\rho(X_i-E[X_i\|Y],X_j-E[X_j\|Y])$

Discrete Y: Model by group means.

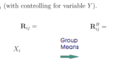
Component	$\mathbf{R}_{ij}$	=	$\eta_i^B\ \eta_j^B\ \mathbf{R}_{ij}^B$	+	$\eta_i^W\ \eta_j^W\ \mathbf{R}_{ij}^W$
Correlation	$\mathbf{R}_{ij}$		$\mathbf{R}_{ij}^B$		$\mathbf{R}_{ij}^W$

Continuous Y: Model by $X_i=\beta_{0i}+\beta_{1i}Y+\epsilon_i,\ i=1,\cdots,p.$

Component	$\mathbf{R}_{ij}$	=	$\sqrt{\mathbf{R}_{iy}^2}\sqrt{\mathbf{R}_{jy}^2}\mathbf{R}_{\pm1}$	+	$\sqrt{1-\mathbf{R}_{iy}^2}\sqrt{1-\mathbf{R}_{jy}^2}\mathbf{R}_{ij:y}$
Correlation	$\mathbf{R}_{ij}$		$\mathbf{R}_{\pm1}$		$\mathbf{R}_{ij:y}$

where
- $\mathbf{R}_{\pm1}\equiv \mathrm{sgn}(\mathbf{R}_{iy}\mathbf{R}_{jy})1$.
- $\mathbf{R}_{ij}$: the total correlation for variables X_i and X_j.
- $\mathbf{R}^B$: the between-group correlation matrix with elements $\mathbf{R}_{ij}^B$.
- $\mathbf{R}^W$: the within-group correlation matrix with elements $\mathbf{R}_{ij}^W$.
- η_i^B: the between η correlation for variable X_i.
- η_i^W: the within η correlation for variable X_i.
- $\mathbf{R}_{ij:y}$: the partial correlation of X_i and X_j (with controlling for variable Y).

Properties
- If K=1, then $\mathbf{R}_{ij}=\eta_i^W\ \eta_j^W\ \mathbf{R}_{ij}^W$.
- If K=N, then $\mathbf{R}_{ij}=\eta_i^B\ \eta_j^B\ \mathbf{R}_{ij}^B$.
- If $\mathbf{R}_{iy}^2, \mathbf{R}_{jy}^2\to 1$, then $\mathbf{R}_{ij}\to\mathbf{R}_{\pm1}$.
- If $\mathbf{R}_{iy}^2, \mathbf{R}_{jy}^2\to 0$, then $\mathbf{R}_{ij}\to\mathbf{R}_{ij:y}$.



Intuitive Explanation of Between-group Correlation

For discrete case, if the number of groups equal to 1, the total correlation of Xi and Xj will reduce to within component, that means there are no group effects. If the number of groups equal to number of observations, the total correlation will reduce to between component, that means all the variations can be explained by group covariate and with no within variations left.
For continuous case, if the correlations of Xi and Y and Xj and Y approximate to zeros, the total correlation will be reduced to partial correlation. Therefore we could investigate the performance of the model component to measure the effect of the covariate.

Results

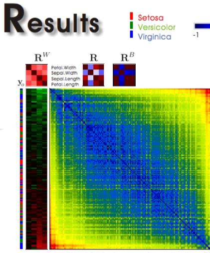


Fig. 6 Sorted maps after adjusting Y_o

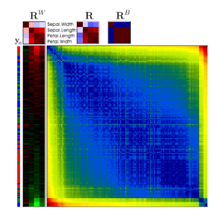


Fig. 7 Sorted maps after adjusting Y_n

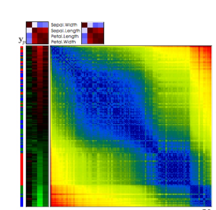
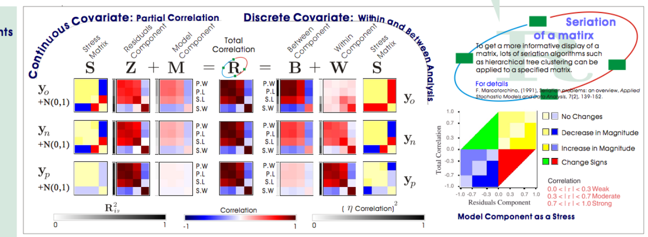


Fig. 8 Sorted maps after adjusting Y_p

Fig. 6~8 show the sorted residual data maps and the corresponding proximity maps after adjusting the observed, noise and permuted species class labels, respectively. The residual data maps are displayed using a green-black-red color spectrum. If the model (group means) fits data well, the most of residuals will display near black (Fig. 6). The sepal width are negative correlated with sepal length, petal length and petal width in Fig. 2. After removing the effect of observed species classes, four measurements became positive correlated and the resulting Euclidean distance matrix for the residuals have no separable groups structure. Moreover the ordering of the observed species class labels is more random. The result suggests that the species class labels could represent the main structure embedded in the data. When the species class labels are added a little noise or permuted, then most of variation will remain in the residuals and the within component can reflect most effects. The Euclidean distance matrices for residuals are getting close to the un-adjusted map in Fig. 2.

Analysis of Model and Residual Components



Conclusion

The conditional correlation can provide the association between two variables while the conditional correlation.

The partial is the result of holding constant a third variable via residuals. Conditional correlation is equivalent to partial correlation under some assumptions (K

The contribution of the covariate effects can then be studied from the between-com

When the covariate is of continuous nature, the conditional correlation is equivalent Cooke, 2000). Another possibility is to discretize the continuous variable into several

Questions

When effects of a covariate such as species in iris data are of interest in the analysis, **covariates adjustment** has to be considered. Therefore we extend the GAP framework by incorporating covariate variable through the estimation of conditional correlations and try to answer the following questions:

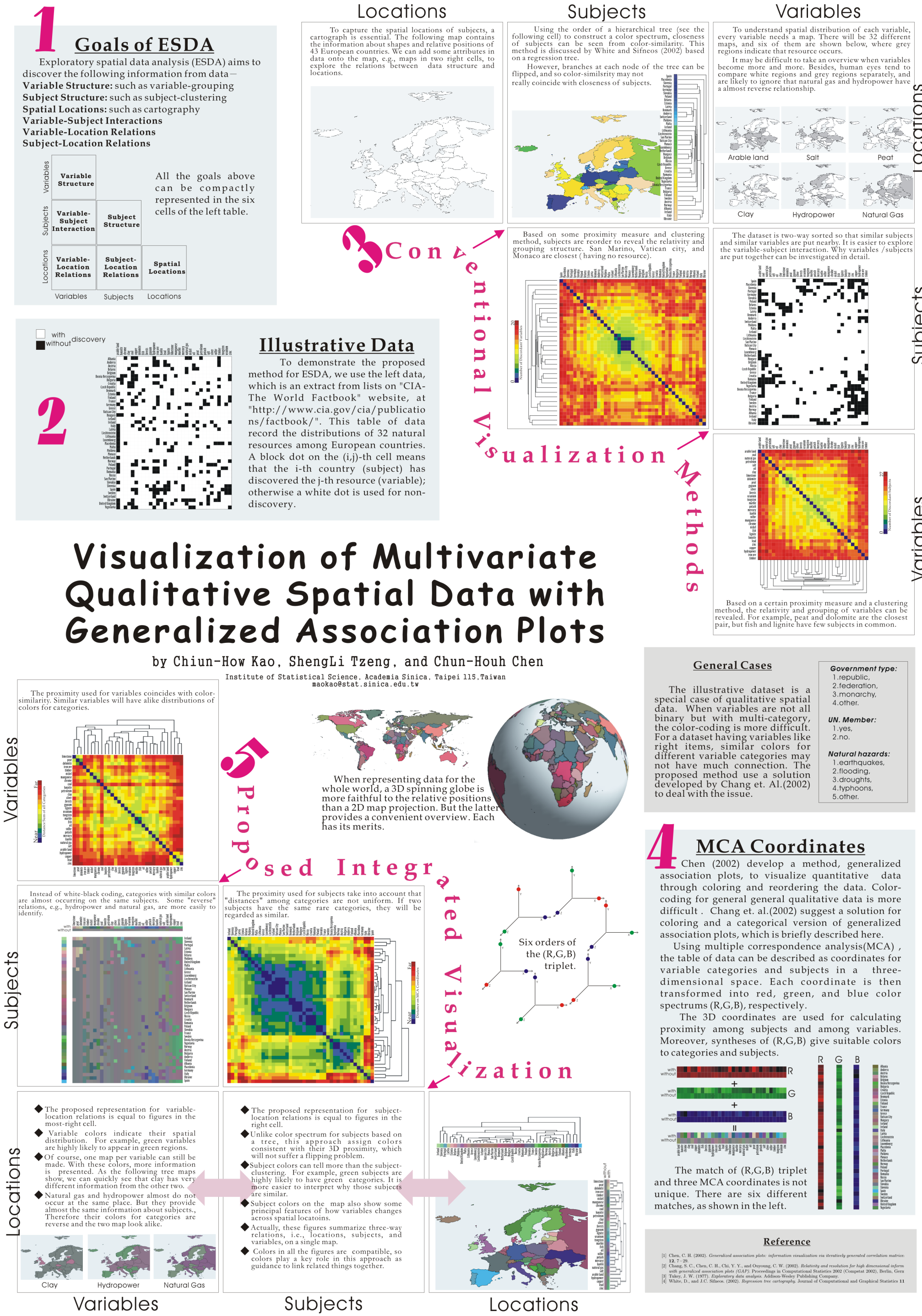
1. How to measure the contribution of effects of a covariate in calculation of correlation?
2. How to adjust the effects of a covariate and visualize the adjusted results?
3. What if a covariate is discrete or continuous case?

[1] Dansereau, F., Alutto, J. A., and Yammarino, F. J. (1984). *Theory testing in organizational behavior: The varient approach.* Englewood Cliffs, NJ: Prentice-Hall.

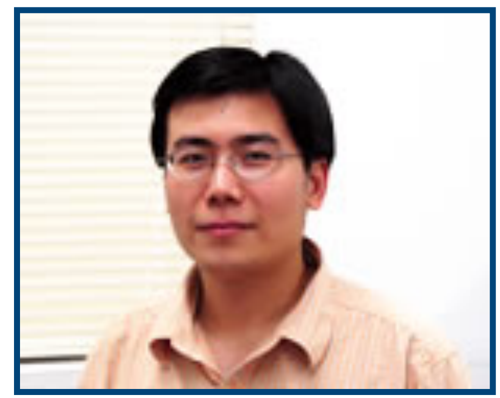
[2] Chen, C. H. (2002). Generalized association plots: information visualization via iteratively generated correlation matrices. *Statistica Sinica* **12**, 7–29.

[3] Kurowicka, D. and Cooke, R. (2000). Conditional and partial correlation for graphical uncertainty models. *In Proceedings of the 2nd International Conference on Mathematical Methods in Reliability, Recent Advances in Reliability Theory*, 259–276.

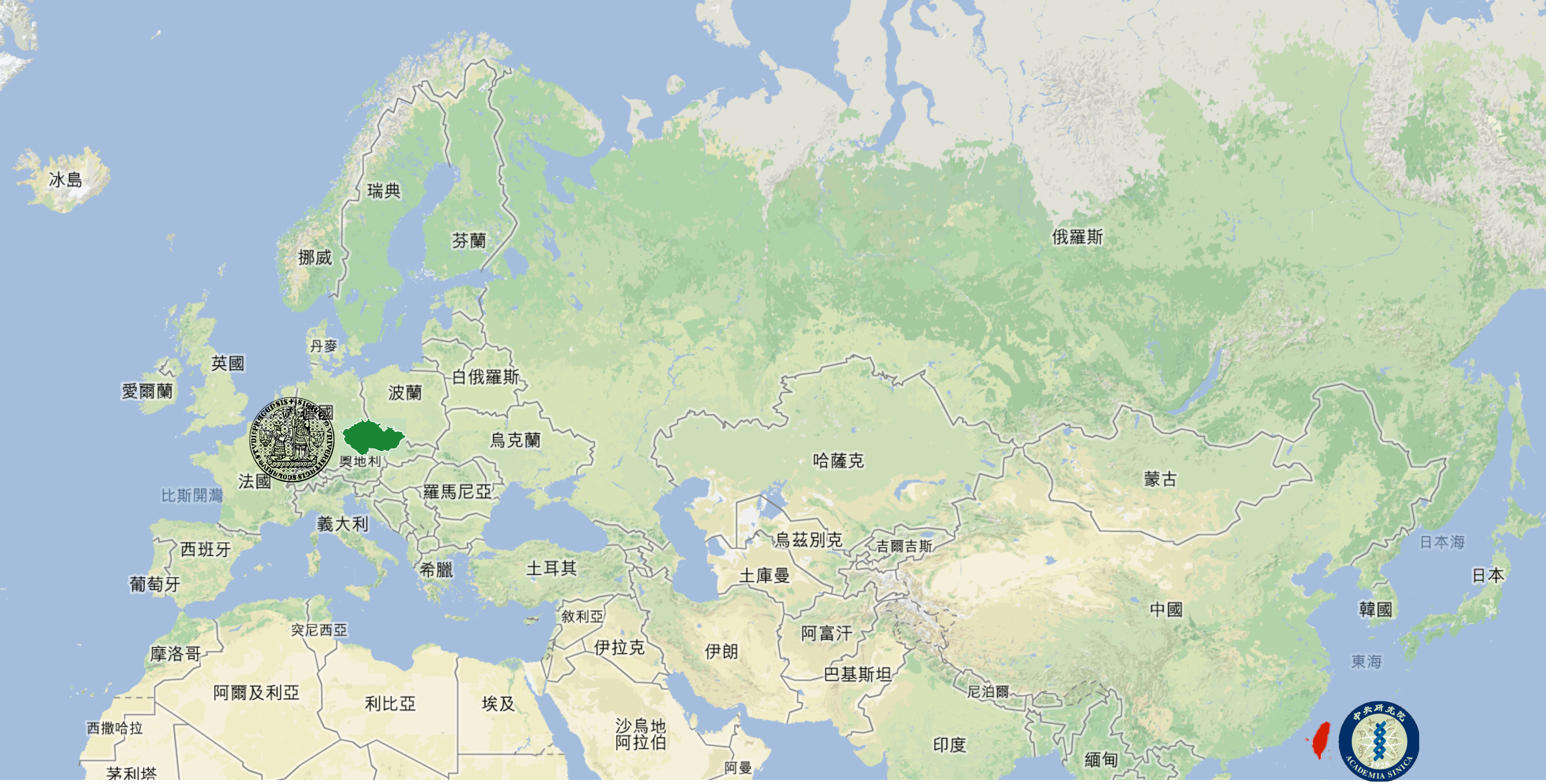
Dr. Wu, Han-Ming
(Poster, Mon. Tue.)



Mr. Kao, Chiun-How
(Poster, Mon. Tue.)



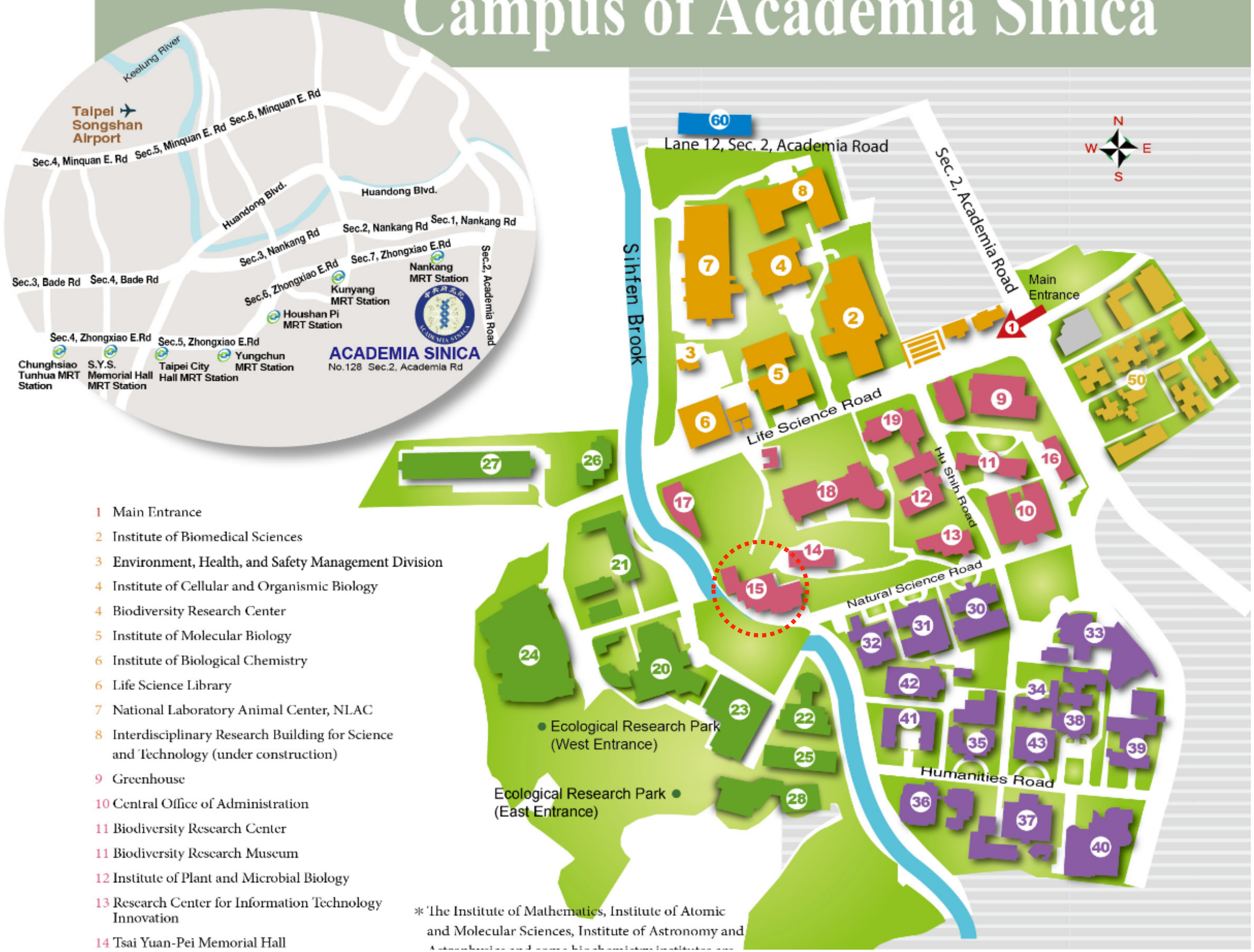
Mr. Tzeng, ShengLi



Czech Republic
Area: 78,866 km2
Pop: 10,513,209
Den: 134/km2

Taiwan (ROC)
Area: 36,193 km2
Pop: 23,315,822
Den: 643/km2

Campus of Academia Sinica



1 Main Entrance
2 Institute of Biomedical Sciences
3 Environment, Health, and Safety Management Division
4 Institute of Cellular and Organismic Biology
4 Biodiversity Research Center
5 Institute of Molecular Biology
6 Institute of Biological Chemistry
6 Life Science Library
7 National Laboratory Animal Center, NLAC
8 Interdisciplinary Research Building for Science and Technology (under construction)
9 Greenhouse
10 Central Office of Administration
11 Biodiversity Research Center
11 Biodiversity Research Museum
12 Institute of Plant and Microbial Biology
13 Research Center for Information Technology Innovation
14 Tsai Yuan-Pei Memorial Hall

* The Institute of Mathematics, Institute of Atomic and Molecular Sciences, Institute of Astronomy and

15 Institute of Statistical Science


LINGUISTIC DATA

Institute of Statistical Science

Strengthened by your nurturing and participation since 1982, the Institute of Statistical Science, Academia Sinica now enters its age of standing firm.

子曰：“吾十有五而志於學，三十而立， 四十而不惑，五十而知天命，六十而耳順，七十而從心所欲，不踰矩。”

Confucius's own account of his gradual progress and attainments. The Master said, "At 15, I had my mind bent on learning. At 30, I stood firm. At 40, I had no doubts. At 50, I knew the decrees of Heaven. At 60, my ear was an obedient organ for the reception of truth. At 70, I could follow what my heart desired, without transgressing what was right."

Chi-Huey Wong,
President of Academia Sinica
2006 ~ Present

Ph.D. in Chemistry,
Massachusetts Institute of Technology in 1982.
Postdoctoral fellow: Harvard University
Assistant Professor: Texas A&M University in 1983, Professor and Ernest W. Hahn Chair:
Scripps Research Institute (1989-2006)
Director of the Genomics Research Center at
Academia Sinica, Taipei (2003-2006).

Yuan T. Lee,
President of Academia Sinica
1994 ~ 2006

Ph.D. in Chemistry,
University of California, Berkeley in 1965
Postdoctoral Fellow: Berkeley (1965~1967)
Assistant Professor: University of Chicago in 1968
Professor: University of California, Berkeley ('74~'94)

Nobel Prize laureate, Chemistry in 1986
(with John C. Polanyi and Dudley R. Herschbach)
President, International Council for Science Units (ICSU) (2011 ~ 2014)

中央研究院

TIGP

TAIWAN INTERNATIONAL GRADUATE PROGRAM

Academia Sinica

We currently offer nine interdisciplinary Ph.D. programs:

1. Chemical Biology and Molecular Biophysics
2. Molecular Science and Technology
3. Molecular and Biological Agricultural Sciences
4. Bioinformatics
5. Molecular and Cell Biology
6. Nano Science and Technology
7. Molecular Medicine
8. Computational Linguistics and Chinese Language Processing
9. Earth System Science
10. Biodiversity

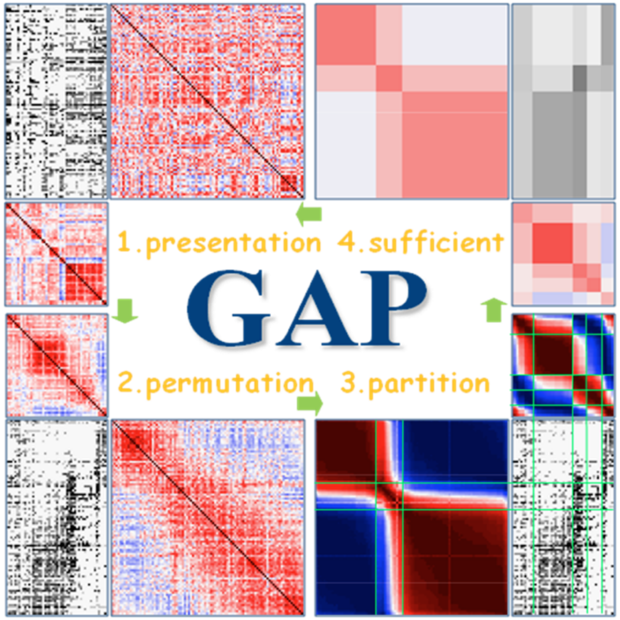
Matrix Visualization: *Approaching Statistics and Statistical Approach*

Chun-houh Chen
Institute of Statistical Science, Academia Sinica, Taiwan

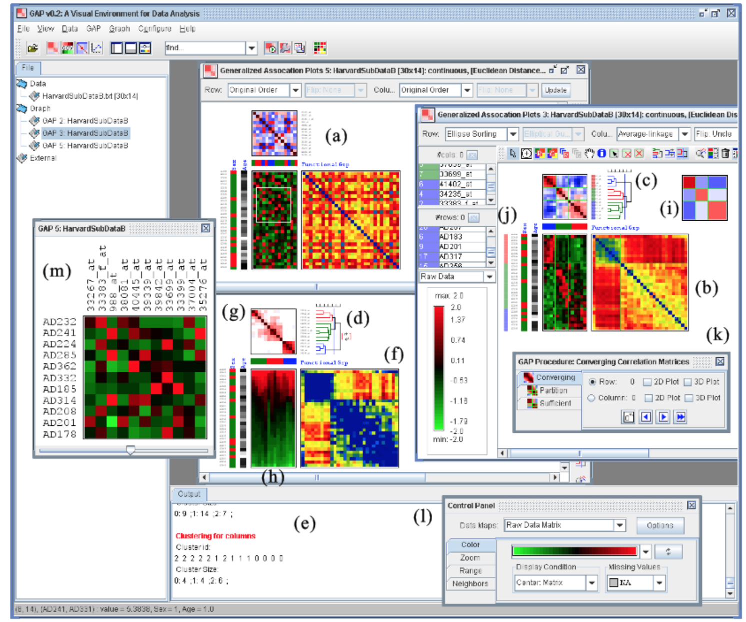
"*It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it*" (John Tukey, 1977). Data analysts and statistics practitioners nowadays are facing difficulties in understanding higher and higher dimensional data with more and more complex nature while conventional graphics/visualization tools do not answer the needs. It is statisticians' responsibility for coming up with graphics/visualization environment that can help users really understand what one CAN DO for complex data generated from modern techniques and sophisticated experiments.

In this lecture I'll summarize our works on matrix visualization for interpreting statistics and statistical approach for implementing matrix visualization. We create matrix visualization environment (GAP: Generalized Association Plots) for conducting statistical analyses, from descriptive statistics, model fitting, inference, to diagnosing. On the other end, we also introduce statistical concepts into matrix visualization environment for visualizing more versatile and complex data structure. With these two matrix-visualization procedures interact with each other we hope a good statistics solution can be achieved.

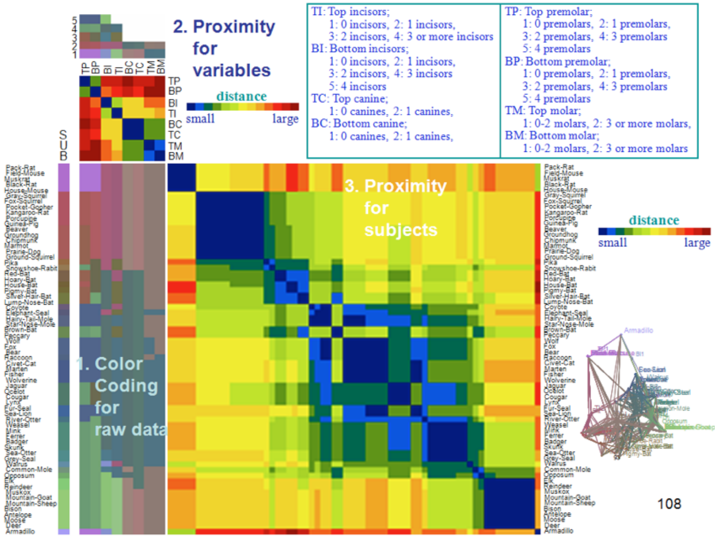
Generalized Association Plots (GAP) http://gap.stat.sinica.edu.tw/Software/



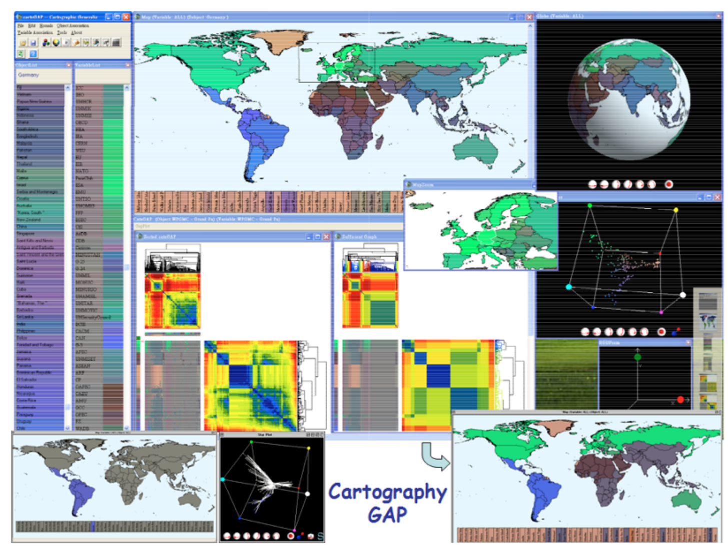
1. GAP basic concept



2. GAPsoftware



3. GAP for categorical data



4. GAP for cartography data

Matrix Visualization by Chun-houh Chen at Charles University, Spring 2013

Week 1 (April 10, 2013)

Introduction and Overview of MV
GAP software

Week 2 (April 17, 2013)

MV for continuous data
MV for binary data

Week 3 (April 24, 2013)

MV for nominal data
MV for cartography data

Week 4 (May 1, 2013)

Interactive Diagnostic Plots for Multidimensional Scaling
MV for proximity matrix modeling

Week 5 (May 8, 2013)

MV with covariate-adjustment
MV for ANOVA data

Week 6 (May 15, 2013)

MV for SDA data
Miscellaneous issues: MV with missing values,

Week 7 (May 15, 2013)

Student presentations

Lab 309 (???) for Information Visualization

Mr. C.H. Kao
Ph.D. student

Dr. Gary Tien
Postdoc. Fellow

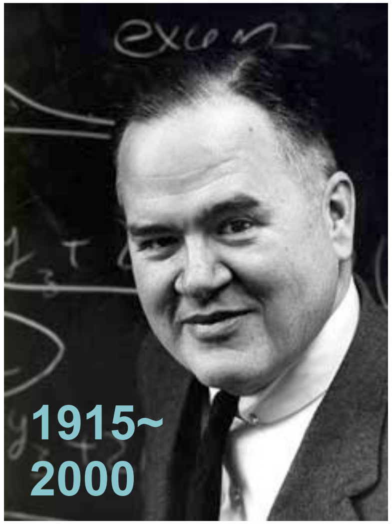
Prof. H.M. Wu
Dept. Math.
Tamkang U.

Prof. S.Y. Shiu
Dept. Stat.
Nat'l Taipei U.

Dr. Mirrian Ho
Postdoc. Fellow

張勝傑
張文宗
陳柏旭
鐘雅齡
黃建勳
林香誼
劉勝宗
曾聖澧
葉紫君
吳怡真
林倩如
歐陽智聞
. . .

Exploratory Data Analysis
EDA, John Tukey (1977)



It is important to understand what you **CAN DO** before you learn to measure how **WELL** you seem to have **DONE** it.

allow the **data to speak** for themselves before standard assumptions or formal modeling

The greatest value of a picture is when it ***forces*** us to notice what we **never expected to see**.

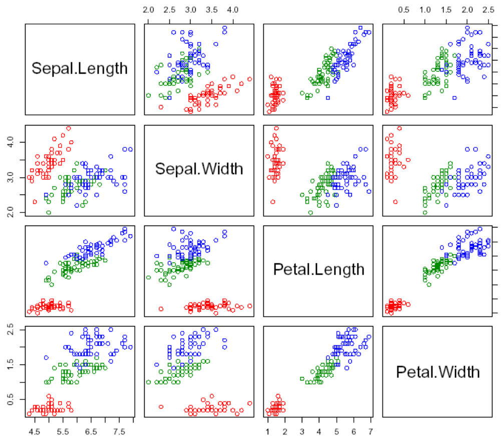
Matrix Visualization as an EDA tool for assisting formal mathematical modeling

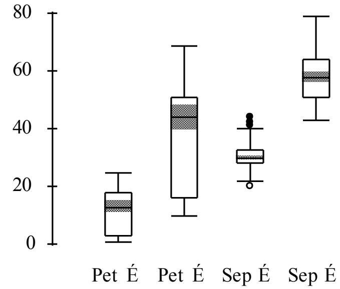
Graphics/Visualization for high dimensional data?
P>5 p>10 p>100 p>10000

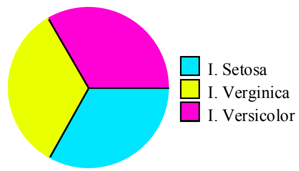
The Iris Data

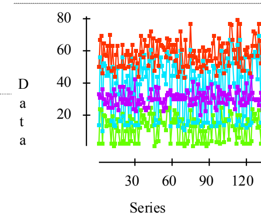
Sepal Length	Sepal Width	Petal Length	Petal Width
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
4.7	3.2	1.3	0.2
4.6	3.1	1.5	0.2
5.0	3.6	1.4	0.2
⋮	⋮	⋮	⋮

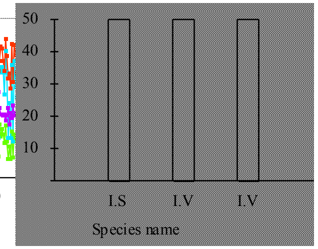
Scatterplot Matrices

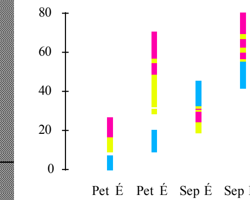


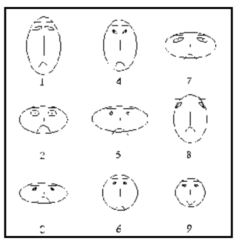


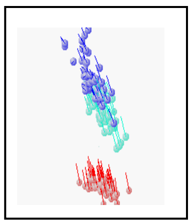


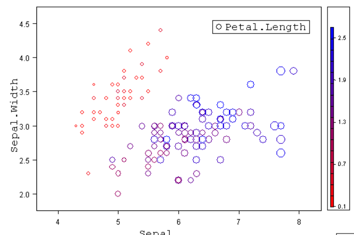


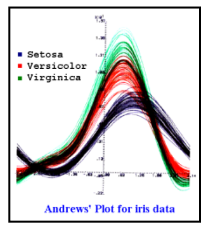


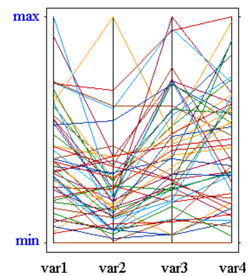


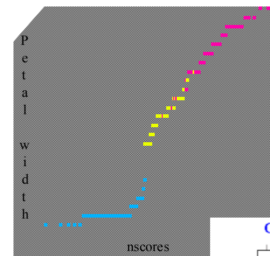


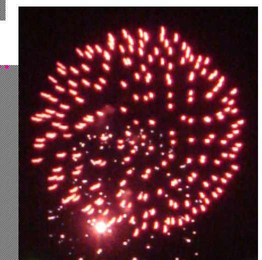


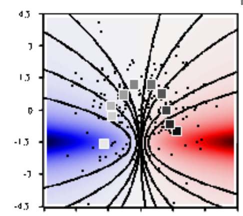




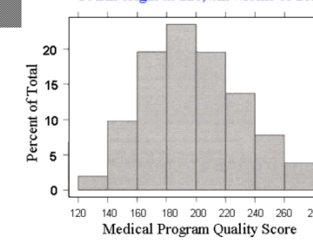




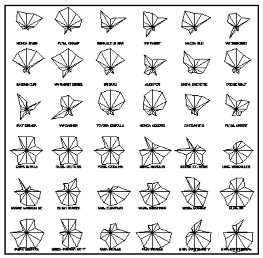


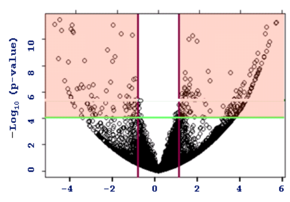


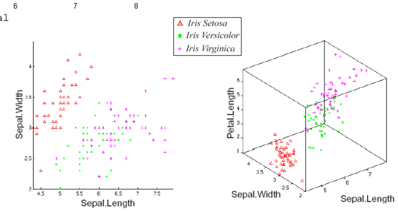
O. Bin origin at 120, bin widths of 20.

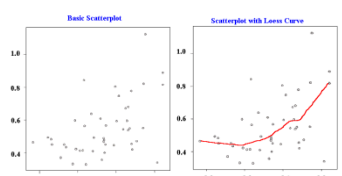


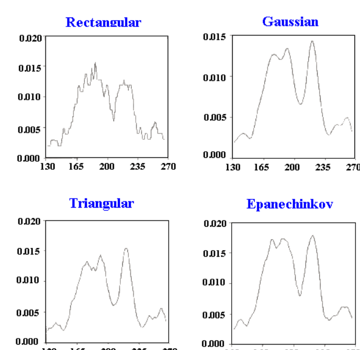
Star plot of Automobile Data

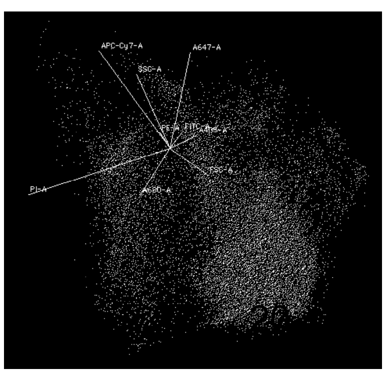




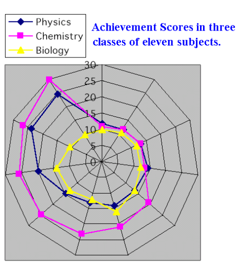


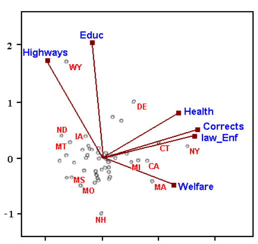




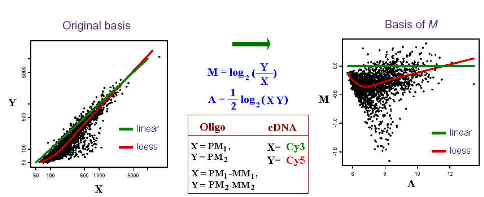


Achievement Scores in three classes of eleven subjects.





Biplot of 1992 State Policy Spending



What can we (statisticians) do for data/information visualization

1. Same as information scientist:
   to create effective graphical/visualization tools/environments

2. To bring in more statistical sense/concept into graphical/visualization tools/environments

   Our approach:
   MV (Matrix Visualization)

   Our tools/environment:
   GAP (Generalized Association Plots)

Approaching Statistics

We create matrix visualization environment for conducting statistical analyses:

✓ **descriptive statistics**:

continuous data: proximity measure, color coding
binary data: proximity measure, black-white (for monary data)
nominal data: Homogeneity Analysis (Dual Scaling, Multiple Corresp. Analysis)
clustering:
nonlinear data structure
symbolic data analysis
Huge Data Sets

✓ **model fitting**:

data with cartography link (here or descriptive)
statistical genetics

✓ **inference**:

missing value
 missing mechanism identification, estimation
EDA for identifying/formulating better hypotheses
MANCOVA
Covariate-Adjusted MV

✓ **diagnosing**:

Interactive Diagnosing System for Statistical methods modeling proximity matrices:
 HCT (Hierarchical Clustering Tree), MDS (Multidimensional Scaling)
 FA (Factor Analysis)

Statistical Approach

We also introduce statistical concepts into matrix visualization environment for visualizing more versatile and complex data structure.

Homogeneity Analysis (Dual Scaling, Multiple Correspondence Analysis)
Nonlinear data structure: use isomap proximity measure
MANCOVA
Covariate-Adjusted MV
Symbolic Data Analysis for handling data with dependent structure:
- Clustered (non-independent) Data
- Repeated Measures Analysis –
- Genetic Familial Data
- Huge Data Sets
  - Large n
  - Large p
  - Large n & p

Other Types of Symbolic Data

Recent Review Articles for MV

The History of the Cluster Heat Map
Leland WILKINSON and Michael FRIENDLY

The American Statistician,
May 2009, Vol. 63, No. 2 179

REVIEW
Seriation and Matrix Reordering Methods: An Historical Overview by Innar Liiv

Statistical Analysis and Data Mining
3: 70–91, 2010

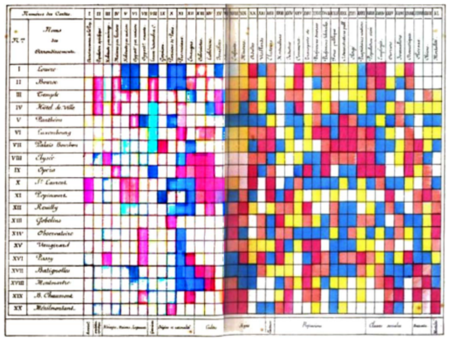


Figure 2. Shaded matrix display from Loua (1873), available online at *http://books.google.com/books/. This was designed as a summary of* 40 separate maps of Paris, showing the characteristics (e.g., national origin, professions, age, social classes) of 20 districts, using a color scale ranging from white (low) through yellow and blue to red (high).

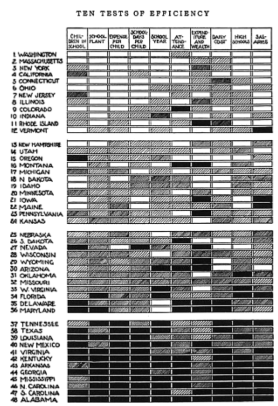
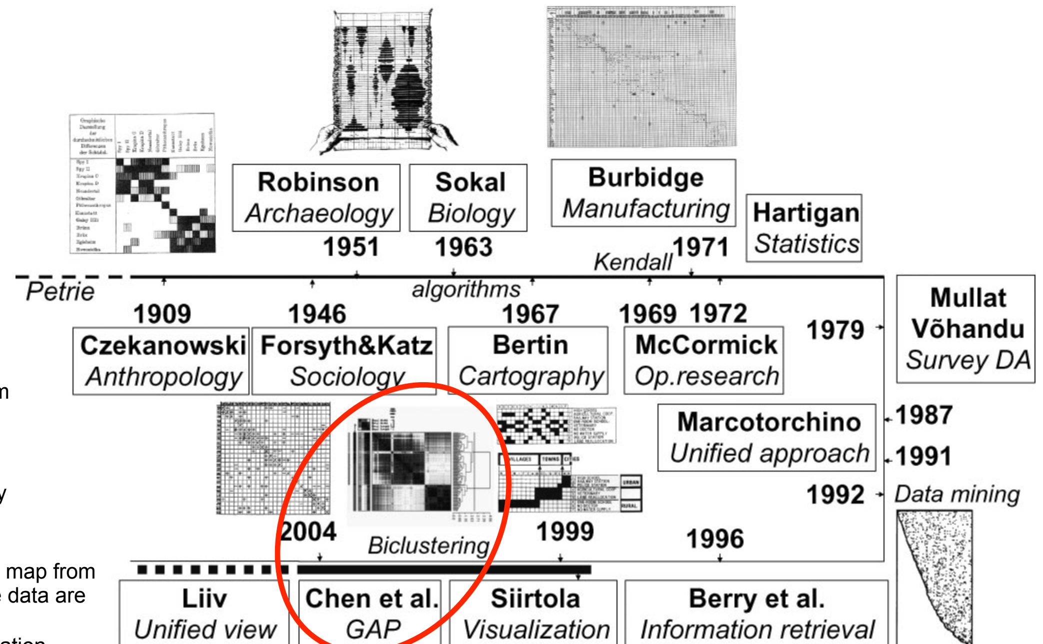


Figure 3. Sorted shaded display from Brinton (1914). The data are ranks of U.S. states on each of 10 educational features assessed in 1910. The matrix has been sorted by the row-marginal ranks.



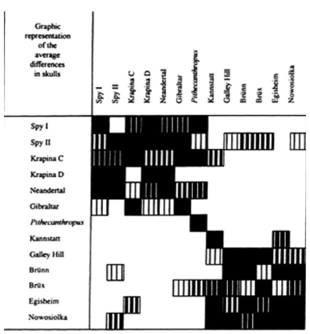


Figure 5. Sorted shaded display from Czekanowski (1909), reproduced in Hage and Harary (1995).

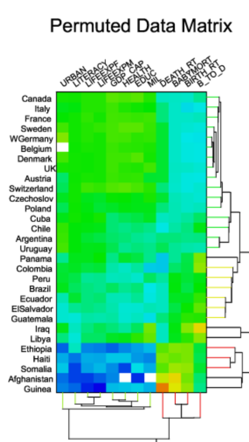


Figure 9. Cluster heat map from Wilkinson (1994). The data are social statistics (i.e., urbanization, literacy, life expectancy for females, GDP, health expenditures, educational expenditures, military expenditures, death rate, infant mortality, birth rate, and ratio of birth to death rate) from a United Nations survey of world countries. The variables were standardized before the hierarchical clustering was performed.

Matrix Visualization (MV): reorderable matrix, heatmap, color histogram, data image

Data

Taiwan **Multidimensional Psychopathological** Group Research Project, (**MPGRP**) Part I: Schizophrenia

Project period: July 1, 1993 to June 30, 1998

Patients (**95 subjects**):

95 First-Admission Psychosis Patients

Schizophrenia (69) + Bipolar Disorder(26)

Rating scales (**50 variables**):

SAPS: Scale for Assessment of **Positive Symptom (30)**

SANS: Scale for Assessment of **Negative Symptom (20)**

SAPS (Scale for Assessment of Positive Symptom)

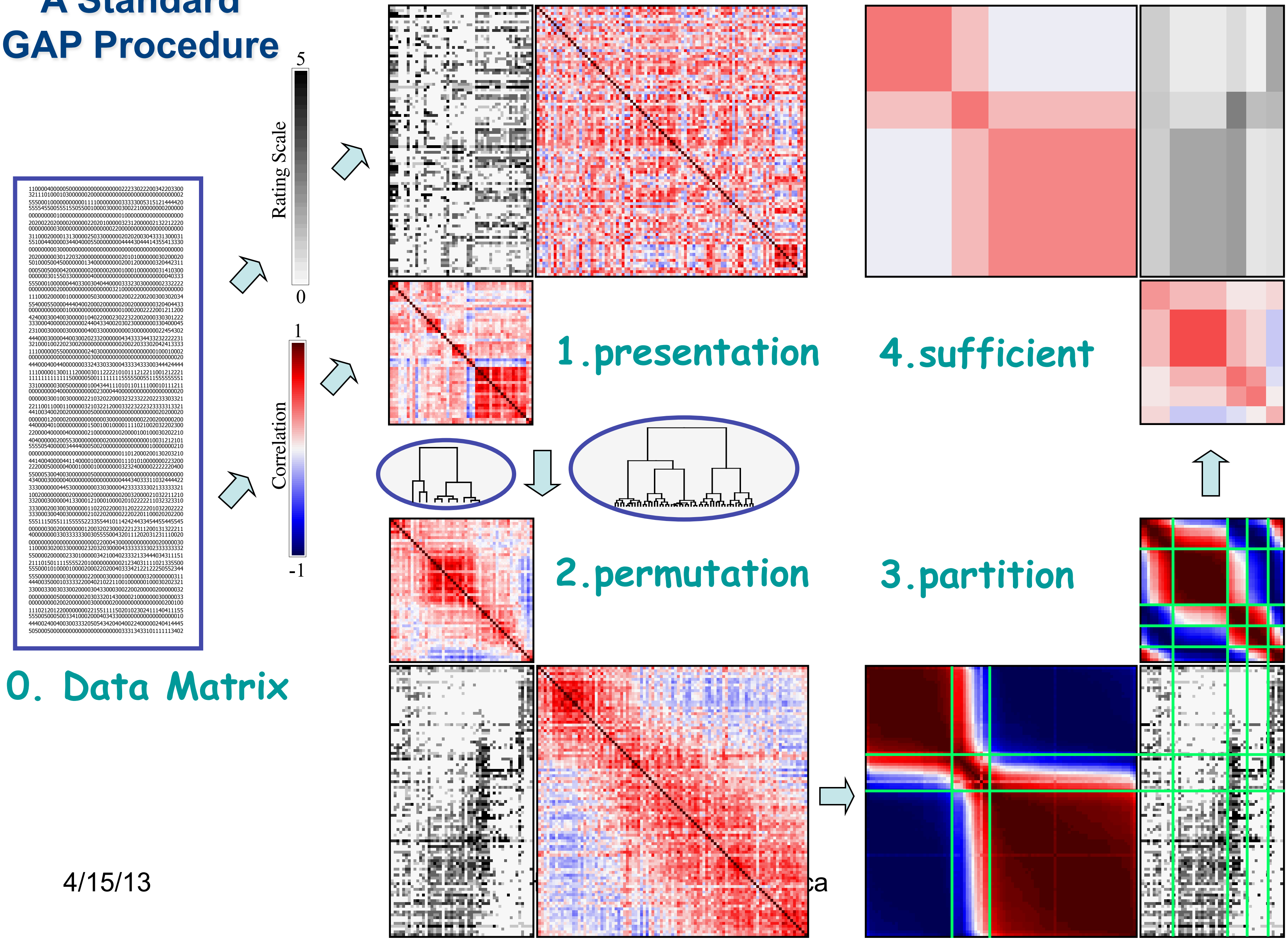
AH1	Auditory Hallucinations
AH2	Voices Commenting
AH3	Voices Conversing
AH4	Somatic or Tactile Hallucinations
AH5	Olfactory Hallucinations
AH6	Visual Hallucinations
DL1	Persecutory Delusions
DL2	Delusion of Jealousy
DL3	Delusion of Sin or Guilt
DL4	Grandiose Delusions
DL5	Religious Delusions
DL6	Somatic Delusions
DL7	Ideas and Delusions of Reference
DL8	Delusions of Being Controlled
DL9	Delusions of Mind Reading
DL10	Thought Broadcasting
DL11	Thought Insertion
DL12	Thought Withdrawal
BEH1	Clothing and Appearance
BEH2	Social and Sexual Behavior
BEH3	Aggressive and Agitated Behavior
BEH4	Repetitive or Stereotyped Behavior
TH1	Derailment
TH2	Tangentiality
TH3	Incoherence
TH4	Illogicality
TH5	Circumstantiality
TH6	Pressure of Speech
TH7	Distractible Speech
TH8	Clanging

SANS (Scale for Assessment of Negative Symptom)

NA1	Unchanging Facial Expression
NA2	Decreased Spontaneous Movements
NA3	Paucity of Expressive Gestures
NA4	Poor Eye Contact
NA5	Affective Nonresponsivity
NA6	Inappropriate Affect
NA7	Lack of Vocal Inflections
NB1	Poverty of Speech
NB2	Poverty of Content of Speech
NB3	Blocking
NB4	Increased Latency of Response
NC1	Grooming and Hygiene
NC2	Impersistence at Work or School
NC3	Physical Anergia
ND1	Recreational Interest and Activities
ND2	Sexual Interest and Activity
ND3	Ability to Feel Intimacy and Closeness
ND4	Relation With Friends and Peers
NE1	Social Inattentiveness
NE2	Inattentiveness During MSE

Approaching Statistics & Statistical Approach

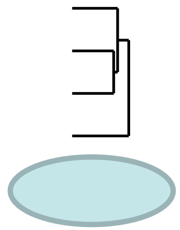
A Standard GAP Procedure



Some essential elements in a GAP MV procedure

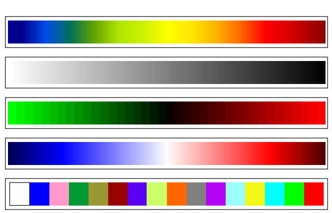
3. Proximity (Variable *p* * *p*)
Continuous
Ordinal
Binary
Nominal

4. Permutation (variable)



1. Data Matrix (*n* * *p*)

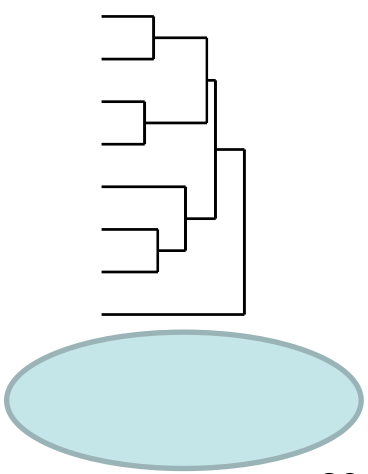
(w/ Color coding)
Continuous
Ordinal
Binary
Nominal



2. Proximity Matrix for Subject (*n* * *n*)

Continuous
Ordinal
Binary
Nominal

4. Permutation (subject)



Continuous GAP Example

Academia Sinica Weekly Newsletter

中央研究院 發行 73年11月01日創刊 99年2月4日出版 院內刊物/非賣品 第1258期

Are there **major** and **minor** institutes in Academia Sinica?
an example of **matrix visualization**

Knowledge World

陳君厚副研究員(統計科學研究所)

大或小、人員多或少；如果你蒐集了許多變數(variable)，則需要多變量統計方法去分析資料。筆者在此介紹一套"看"高維度資料的方法: 矩陣視覺化(matrix visualization: MV)。為了介紹MV，我們以本院31個所(處)中心為樣本蒐集20個變數(表一：17數值變數、3共變數(covariate))；資料之蒐集以公開及方便性為主。讀者對這20個變數的選擇當然有所疑慮--約聘僱人員與院外計畫等變數未納入、某些變數可能資料時間太短(如前瞻計畫)、某些變數可能應使用相對數值(如年輕著作獎)、人事變數比例是否過高等。筆者強調此資料之蒐集以方法介紹為主，非以資料分析為目的。我們將原始資料(人數、件

3

Div. of Life Sci. (5+3)

- Plant-Microb Bio
- Cellul-Organ Bio
- Bio Chemistry
- Molecular Bio
- Biomedical Sci
- Agric Biot (Ctr)
- Genomics (Ctr)
- Biodiversity (Ctr)

Div. of Math-Physi Sci (8+3)

- Mathematics
- Physics
- Chemistry
- Earth
- Information
- Statistics
- Atomic-Molecular
- Astron-Astrophy
- Applied Sci (Ctr)
- Envir Change (Ctr)
- Inf Tech Innov (Ctr)

Div. of Hum-Social Sci (11+1)

- Hist-Philol
- Ethnology
- Mod Hist
- Economic
- Europ-Ame
- Chi Liter-Phil
- Taiwan Hist
- Sociology
- Linguistics
- Political Sci
- Iurisprudentiae
- Hum-Soc Sci (Ctr)

Are there major and minor institutes in Academia Sinica?

A. Personnel Variables
1. Research Fellow
2. Senior RF (%)
3. Female RF (%)
4. Research Scientist
5. Other Research
6. Administrative Staff
7. Total Personnel

B. Project Variables
8. Thematic Project
9. Investigator Project
10. Career Project
11. Postdoc (AS)
12. Postdoc (Regular)

C. Other Variables
13. Junior Award
14. Patent

D. Budget Variables
15. Operating Expense
16. Equipment Expense

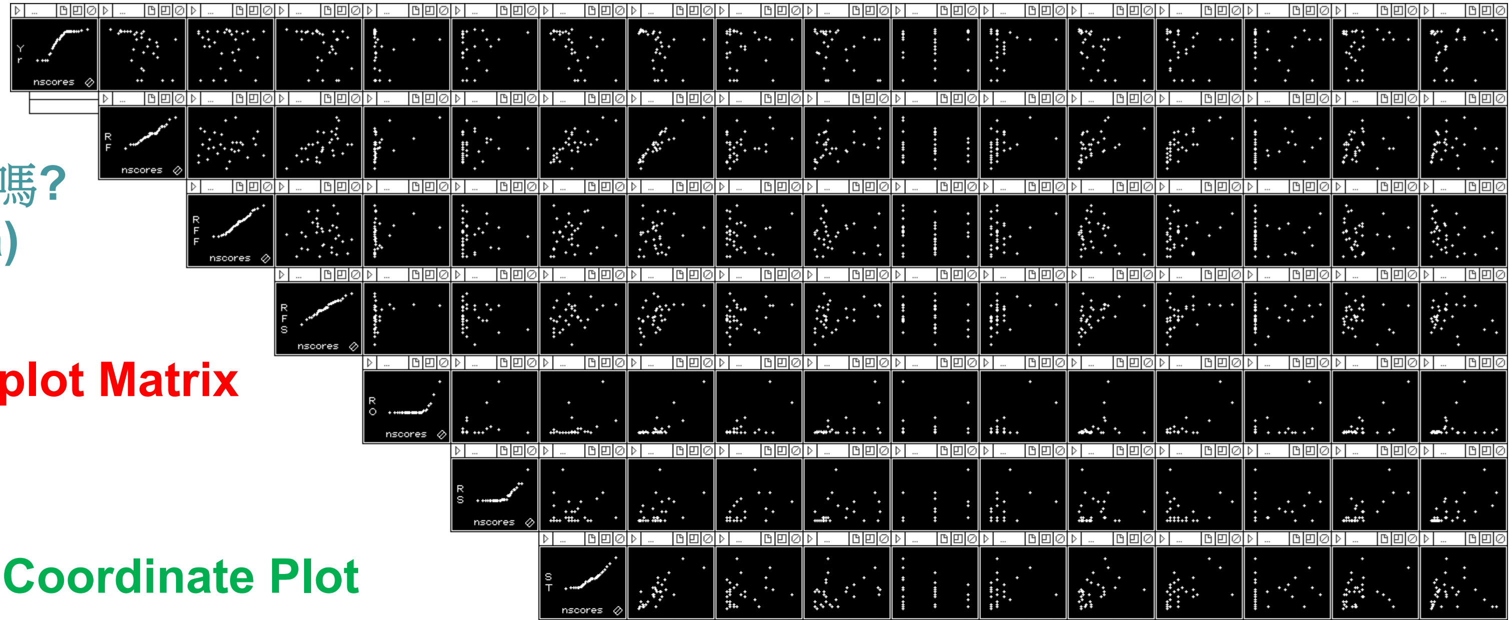
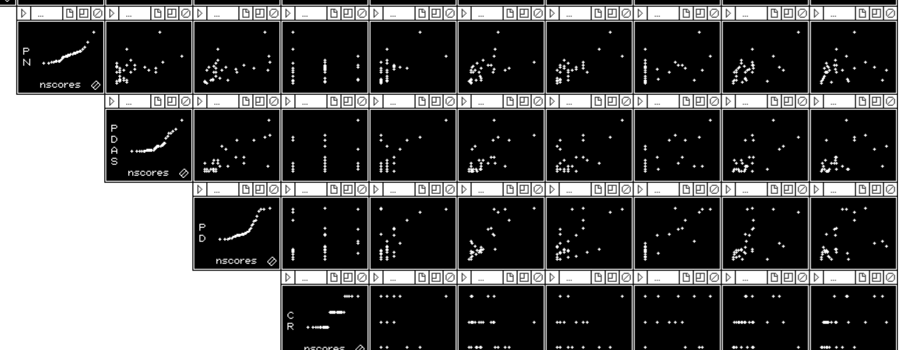
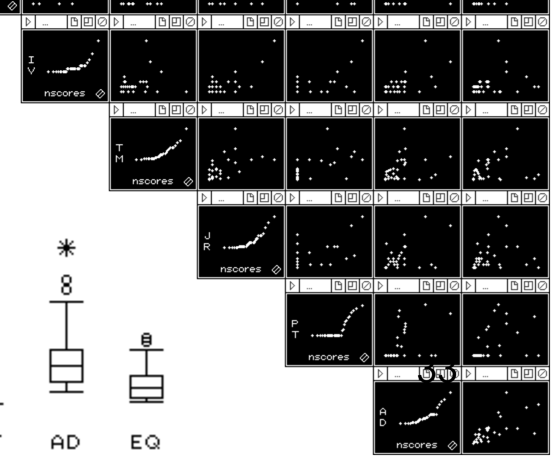
E. Institute Variables
17. Year of Establish

Covariate
18. Division
19. Research Center
20. Preparatory Office

Are there major and minor institutes in Academia Sinica? (original data)

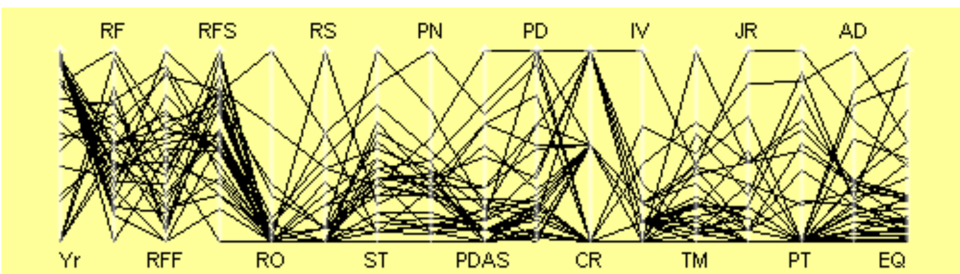
Name	Div	Cen	Pre	Yr	RF	RFS	RFF	RS	OR	AD	PN	TM	IV	CR	PDA	PD	JR	PT	OP	EQ
Math	0	0	0	1947	26	0.769231	0.192308	1	0	7	34	2	1	1	1	22	1	0	45539	38375
Phys	0	0	0	1928	40	0.6	0.025	3	0	9	52	30.5	0	1	13	153	7	20	120657	99572
Chem	0	0	0	1928	23	0.434783	0.130435	0	3	10	36	17	0	0	2	150	4	38	95299	51506
Eart	0	0	0	1982	30	0.533333	0.1	7	4	16	57	7	1	1	4	31	2	0	69640	36377
Info	0	0	0	1982	37	0.540541	0.054054	3	0	6	46	10.7	2	1	5	94	7	25	127149	38979
Stat	0	0	0	1987	36	0.444444	0.138889	0	1	8	45	2	2	1	0	22	2	0	83410	19495
Atom	0	0	0	1995	26	0.730769	0.038462	0	0	14	40	19	4	0	11	117	7	27	124889	77647
Astr	0	0	1	1993	22	0.181818	0.090909	9	2	2	37	3	1	1	9	41	0	0	270658	144166
Appl	0	1	0	2004	16	0.125	0	1	0	2	19	1	0	2	1	19	1	9	85299	62790
Envi	0	1	0	2004	12	0.25	0.083333	0	0	2	14	1	0	1	3	2	1	0	73687	28500
Inno	0	1	0	2007	4	0	0.25	3	0	8	15	0	0	1	1	0	0	0	192877	64885
Plan	1	0	0	1944	27	0.37037	0.37037	6	3	11	47	9	2	2	3	92	1	22	121440	35152
Cell	1	0	0	1959	19	0.421053	0.315789	4	0	10	33	11.5	0	1	11	76	1	16	86000	17405
BiCm	1	0	0	1977	20	0.55	0.35	1	10	8	39	12	2	1	5	74	5	8	109008	37542
MoBi	1	0	0	1993	31	0.612903	0.483871	6	20	7	64	12	10	2	18	152	17	43	222961	60006
Biom	1	0	0	1993	49	0.591837	0.326531	9	33	18	109	14	6	0	12	142	14	36	339270	77403
Agri	1	1	0	2006	11	0.363636	0.272727	5	0	2	18	1	1	2	2	55	3	33	94698	137182
Geno	1	1	0	2003	22	0.227273	0.227273	16	0	5	43	1.5	3	2	8	50	1	0	250841	119556
Biod	1	1	0	2004	18	0.666667	0.166667	0	1	3	22	6	1	1	2	37	0	0	87864	11882
Hist	2	0	0	1928	47	0.617021	0.234043	1	8	22	78	12.5	1	0	3	31	11	0	121339	39163
Ethn	2	0	0	1928	26	0.269231	0.5	1	1	12	40	6	0	1	1	27	2	0	56133	13394
Mode	2	0	0	1965	37	0.378378	0.513514	0	1	13	51	4	1	0	0	28	3	0	70360	17630
Econ	2	0	0	1970	34	0.588235	0.205882	0	1	8	43	1.5	2	1	0	10	5	0	57589	20839
Euro	2	0	0	1972	27	0.518519	0.296296	0	1	9	37	1	1	0	1	7	3	0	42942	23016
Chin	2	0	0	2002	26	0.307692	0.384615	0	0	7	33	5	1	1	4	22	8	0	36622	15271
Taiw	2	0	0	2004	14	0.357143	0.428571	0	1	3	18	3.5	0	0	1	16	0	0	35569	9258
Soci	2	0	0	2000	21	0.428571	0.285714	0	0	4	25	5.5	1	0	1	28	3	0	42830	12468
Ling	2	0	0	2004	16	0.5625	0.5625	0	0	2	18	5.83	1	0	3	21	1	0	42733	8557
Poli	2	0	1	2002	10	0.4	0	0	0	1	11	1	0	0	0	6	0	0	27766	5823
Iuri	2	0	1	2004	11	0.181818	0.272727	0	0	4	15	0	1	0	0	2	0	0	30024	6178
Huma	2	1	0	2004	27	0.37037	0.333333	1	0	6	34	1	0	0	0	26	3	0	118930	20249

中研院有大小所嗎?
(original data)

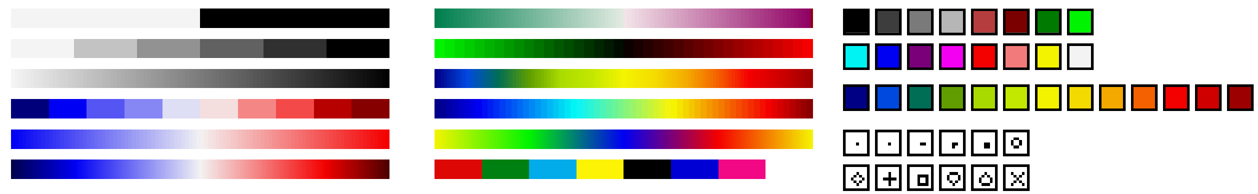
Scatter-plot Matrix

Parallel Coordinate Plot

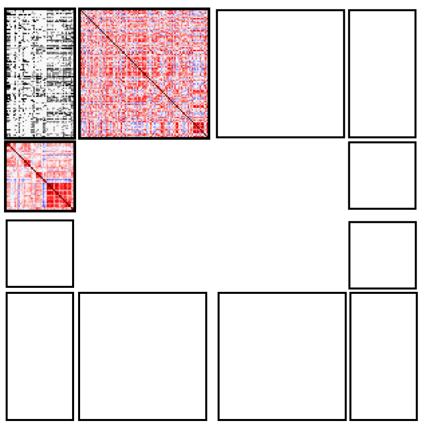


Side-by-side Box-Plot

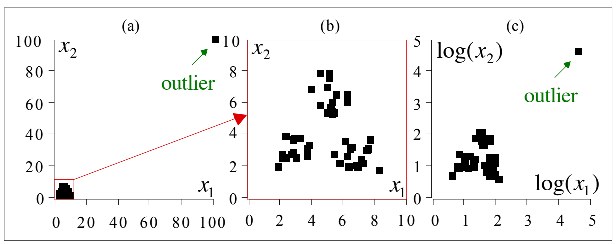
1. Selection of suitable color spectrum

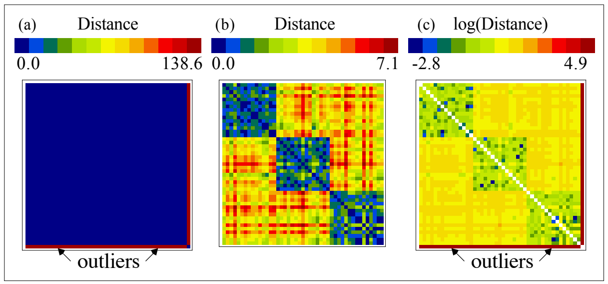


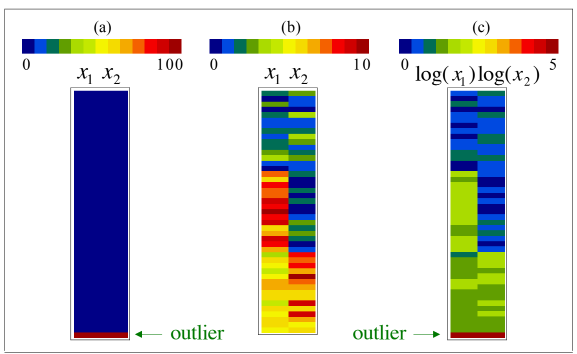
1.presentation



2. Transformation/Standardization of data? (“Resolution” of a Statistical Graph)





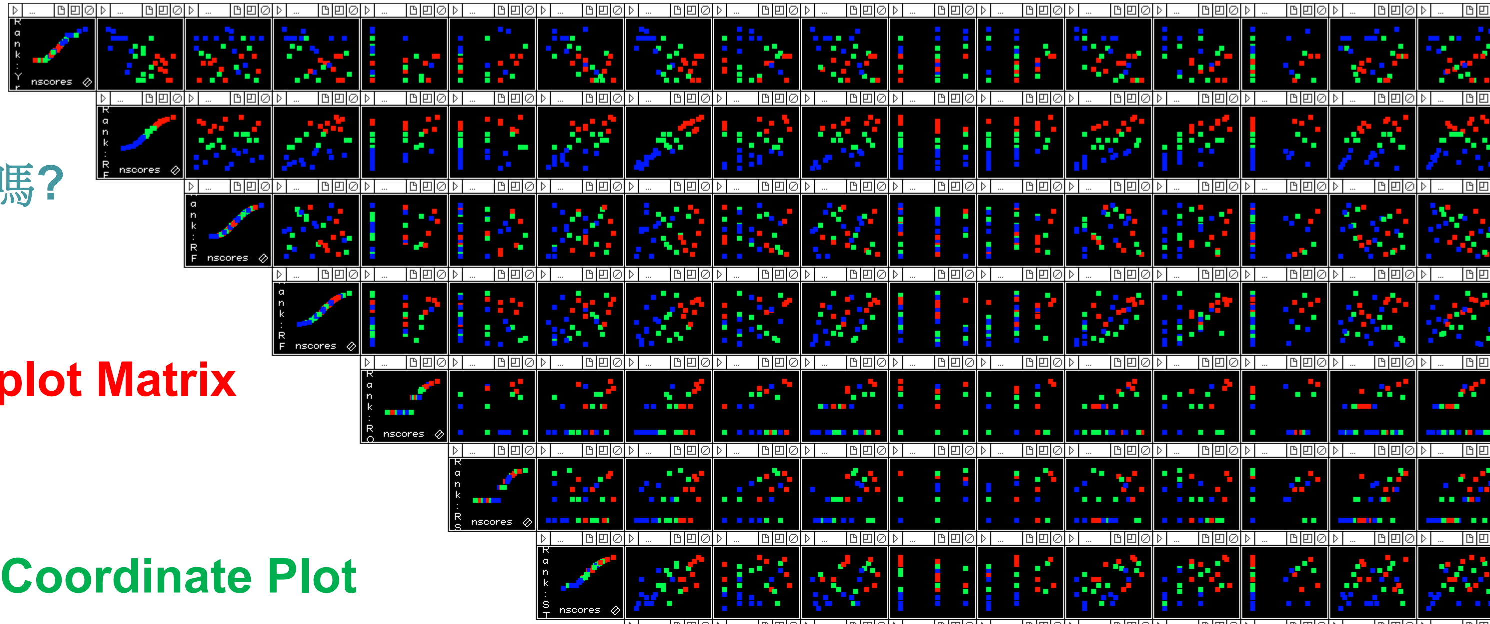


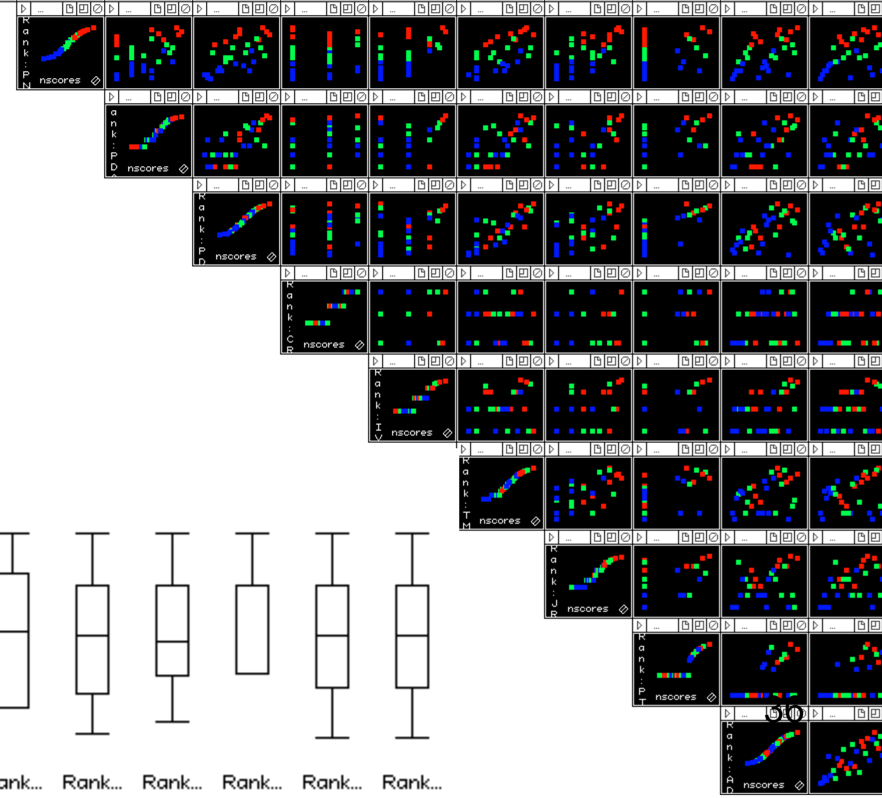
3. Selection of proximities for variable/sample correlation/covariance/distance/ . . .

Are there major and minor institutes in Academia Sinica? (rank data)

Name	Div	Cen	Pre	Yr	RF	RFS	RFF	RS	OR	AD	PN	TM	IV	CR	PDA	PD	JR	PT	OP	EQ
Math	0	0	0	6	17.5	31	12	17.5	8.5	15	13.5	11.5	16.5	19.5	10	11	10	10.5	8	19
Phys	0	0	0	2.5	29	26	3	22	8.5	21.5	27	31	5.5	19.5	30	31	26	24	22	28
Chem	0	0	0	2.5	15	17	9	7.5	25.5	23.5	15	29	5.5	6.5	15	29	22	30	19	22
Eart	0	0	0	12.5	23	20	8	28	27	29	28	21	16.5	19.5	21.5	17.5	15	10.5	11	17
Info	0	0	0	12.5	27.5	21	5	22	8.5	12.5	24	23	25	19.5	23.5	26	26	26	26	20
Stat	0	0	0	14	26	18	10	7.5	20	18.5	23	11.5	25	19.5	3.5	11	15	10.5	14	11
Atom	0	0	0	18	17.5	30	4	7.5	8.5	28	19.5	30	29	6.5	27.5	27	26	27	25	27
Astr	0	0	1	16	13.5	3.5	7	29.5	24	4	16.5	13	16.5	19.5	26	20	3.5	10.5	30	31
Appl	0	1	0	26	7.5	2	1.5	17.5	8.5	4	8	5.5	5.5	29	10	8	10	22	15	24
Envi	0	1	0	26	5	6	6	7.5	8.5	4	2	5.5	5.5	19.5	18.5	2.5	10	10.5	13	15
Inno	0	1	0	31	1	1	16	22	8.5	18.5	3.5	1.5	5.5	19.5	10	1	3.5	10.5	27	25
Plan	1	0	0	5	21	11.5	25	26.5	25.5	25	25	22	25	29	18.5	25	10	25	24	16
Cell	1	0	0	7	10	15	21	24	8.5	23.5	11.5	24	5.5	19.5	27.5	24	10	23	16	9
BiCm	1	0	0	11	11	22	24	17.5	29	18.5	18	25.5	25	19.5	23.5	23	23.5	21	20	18
MoBi	1	0	0	16	24	27	28	26.5	30	15	29	25.5	31	29	31	30	31	31	28	23
Biom	1	0	0	16	31	25	22	29.5	31	30	31	28	30	6.5	29	28	30	29	31	26
Agri	1	1	0	30	3.5	10	17.5	25	8.5	4	6	5.5	16.5	29	15	22	19	28	18	30
Geno	1	1	0	22	13.5	5	14	31	8.5	11	21.5	9.5	28	29	25	21	10	10.5	29	29
Biod	1	1	0	26	9	29	11	7.5	20	7.5	9	19.5	16.5	19.5	15	19	3.5	10.5	17	5
Hist	2	0	0	2.5	30	28	15	17.5	28	31	30	27	16.5	6.5	18.5	17.5	29	10.5	23	21
Ethn	2	0	0	2.5	17.5	7	29	17.5	20	26	19.5	19.5	5.5	19.5	10	14	15	10.5	9	7
Mode	2	0	0	8	27.5	13	30	7.5	20	27	26	15	16.5	6.5	3.5	15.5	19	10.5	12	10
Econ	2	0	0	9	25	24	13	7.5	20	18.5	21.5	9.5	25	19.5	3.5	6	23.5	10.5	10	13
Euro	2	0	0	10	21	19	20	7.5	20	21.5	16.5	5.5	16.5	6.5	10	5	19	10.5	7	14
Chin	2	0	0	20.5	17.5	8	26	7.5	8.5	15	11.5	16	16.5	19.5	21.5	11	28	10.5	4	8
Taiw	2	0	0	26	6	9	27	7.5	20	7.5	6	14	5.5	6.5	10	7	3.5	10.5	3	4
Soci	2	0	0	19	12	16	19	7.5	8.5	9.5	10	17	16.5	6.5	10	15.5	19	10.5	6	6
Ling	2	0	0	26	7.5	23	31	7.5	8.5	4	6	18	16.5	6.5	18.5	9	10	10.5	5	3
Poli	2	0	1	20.5	2	14	1.5	7.5	8.5	1	1	5.5	5.5	6.5	3.5	4	3.5	10.5	1	1
Iuri	2	0	1	26	3.5	3.5	17.5	7.5	8.5	9.5	3.5	1.5	16.5	6.5	3.5	2.5	3.5	10.5	2	2
Huma	2	1	0	26	21	11.5	23	17.5	8.5	12.5	13.5	5.5	5.5	6.5	3.5	13	19	10.5	21	12

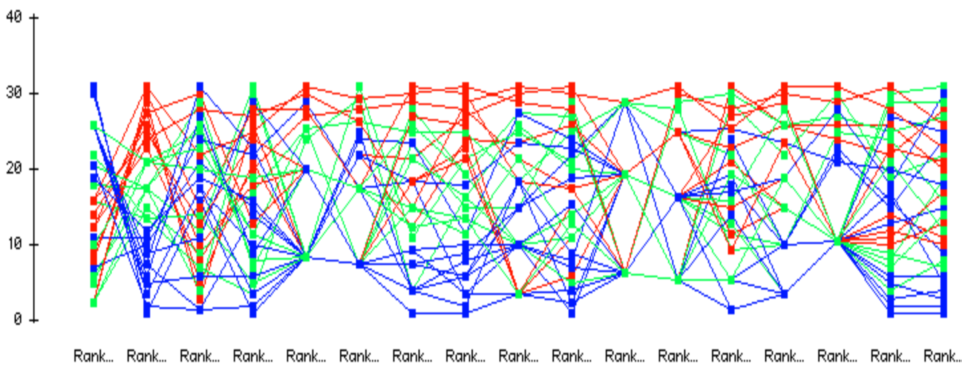
中研院有大小所嗎?
(rank data)



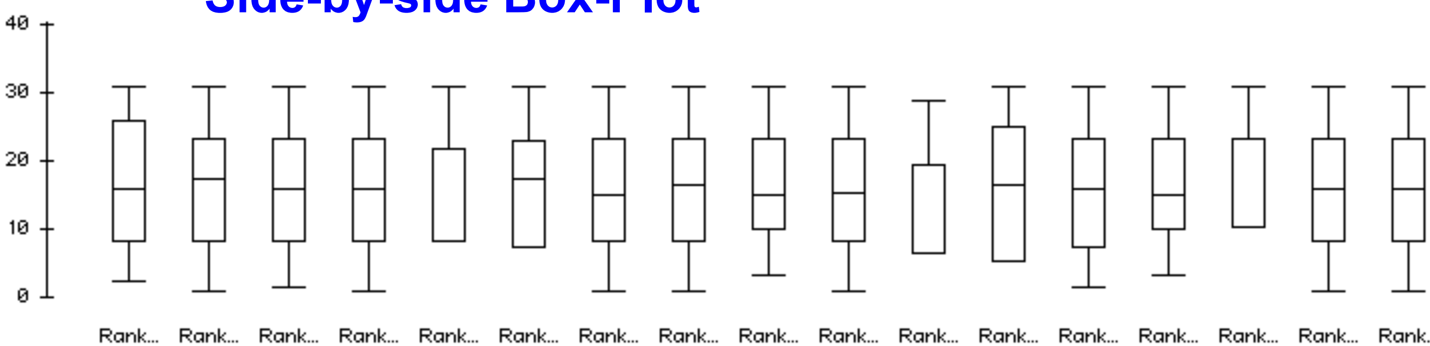


Scatter-plot Matrix

Parallel Coordinate Plot



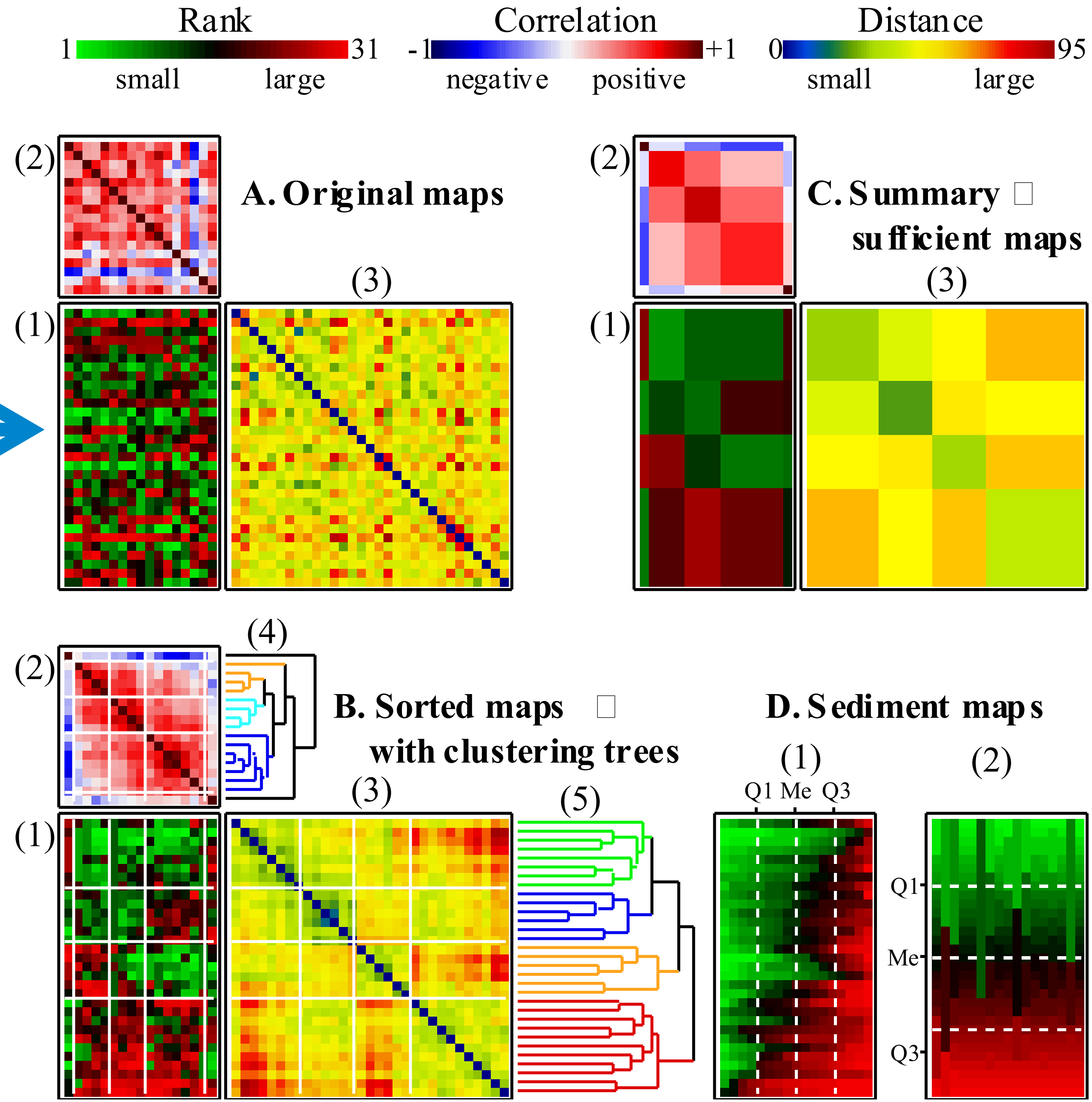
Side-by-side Box-Plot

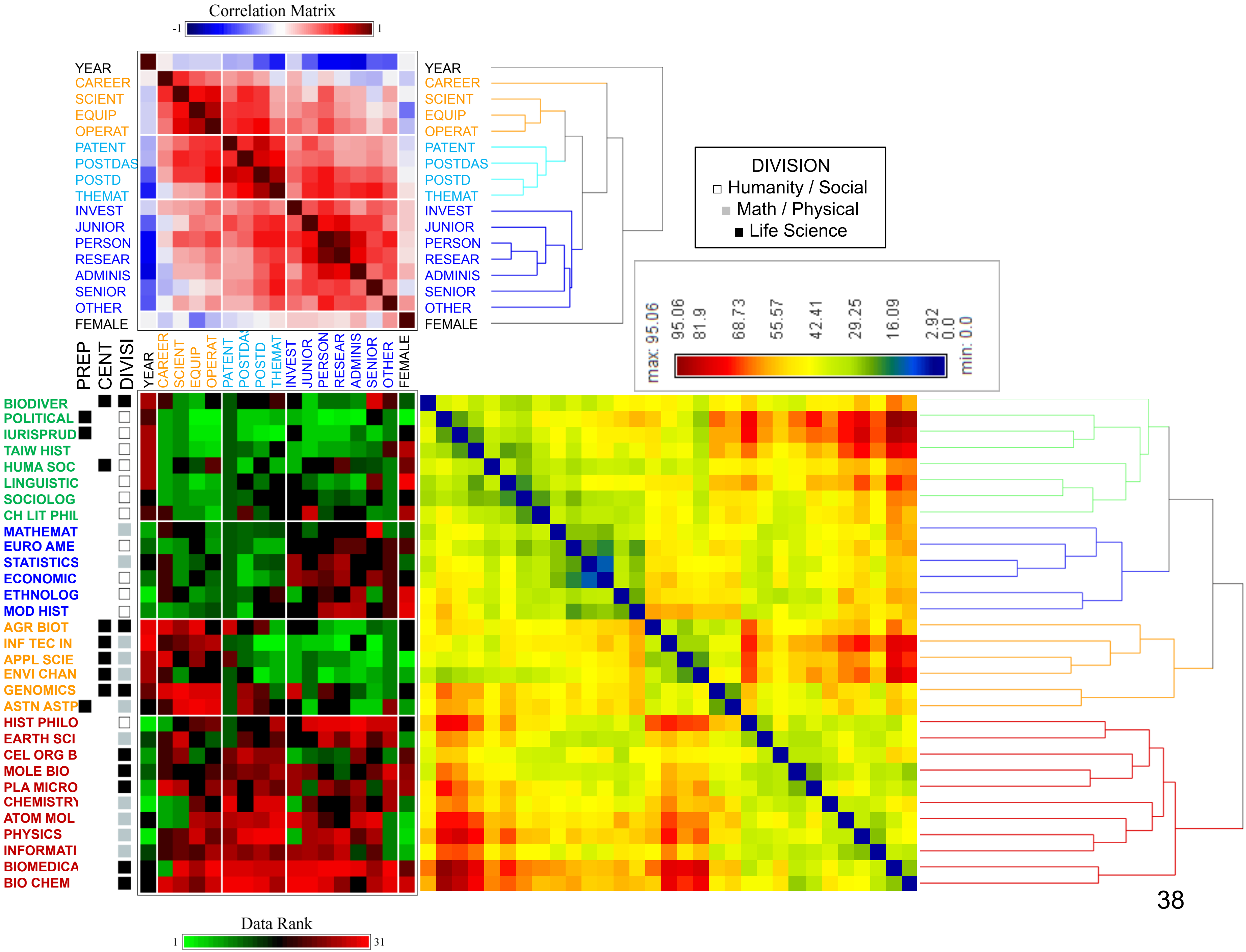


Are there major and minor institutes in Academia Sinica?

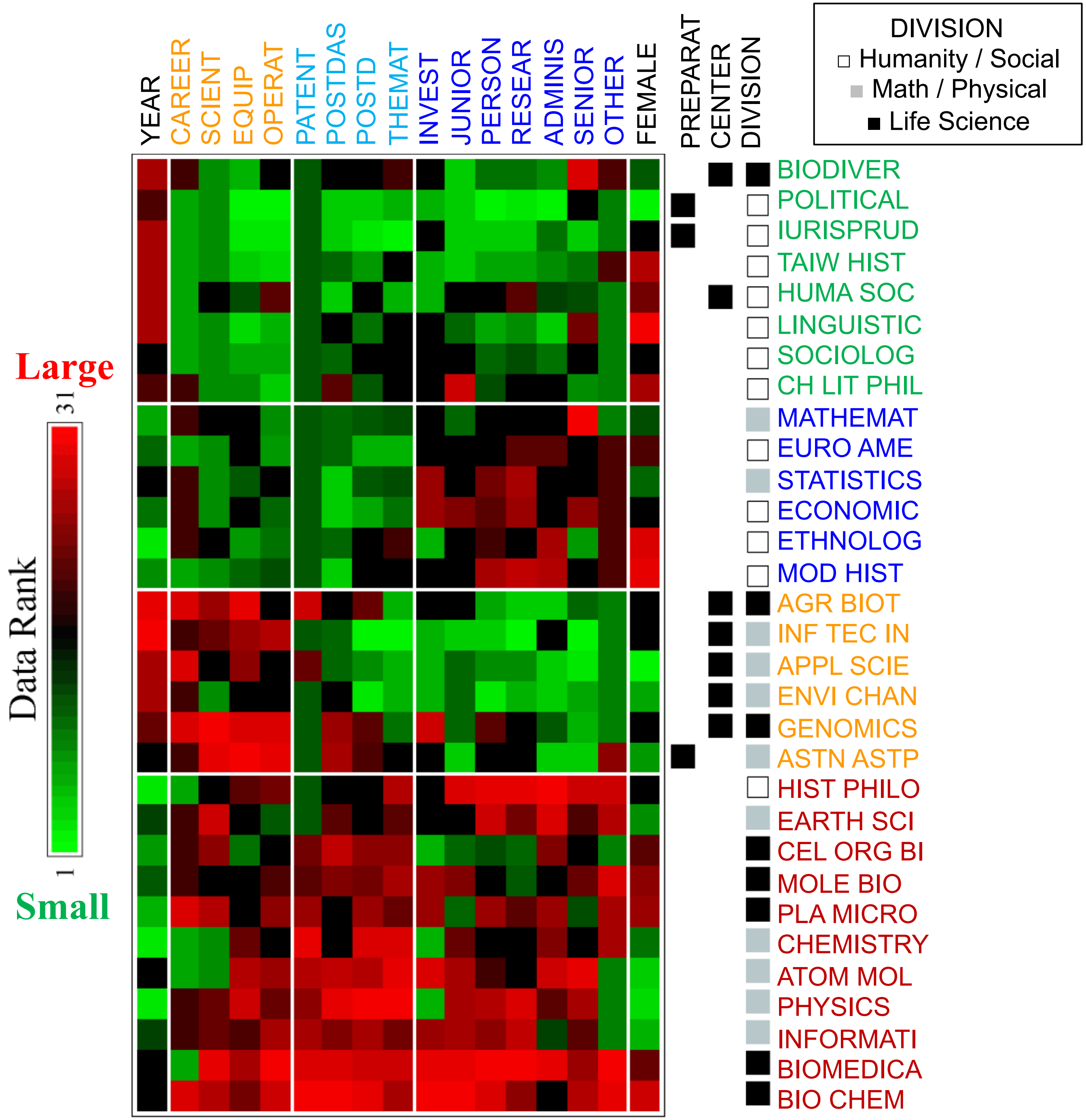
Rank Transformed Data

6	18	31	12	18	9	15	14	12	17	20	10	11	10	11	8	19
3	29	26	3	22	9	22	27	31	6	20	30	31	26	24	22	28
3	15	17	9	8	26	24	15	29	6	7	15	29	22	30	19	22
13	23	20	8	28	27	29	28	21	17	20	22	18	15	11	11	17
13	28	21	5	22	9	13	24	23	25	20	24	26	26	26	26	20
14	26	18	10	8	20	19	23	12	25	20	4	11	15	11	14	11
18	18	30	4	8	9	28	20	30	29	7	28	27	26	27	25	27
16	14	4	7	30	24	4	17	13	17	20	26	20	4	11	30	31
26	8	2	2	18	9	4	8	6	6	29	10	8	10	22	15	24
26	5	6	6	8	9	4	2	6	6	20	19	3	10	11	13	15
31	1	1	16	22	9	19	4	2	6	20	10	1	4	11	27	25
5	21	12	25	27	26	25	25	22	25	29	19	25	10	25	24	16
7	10	15	21	24	9	24	12	24	6	20	28	24	10	23	16	9
11	11	22	24	18	29	19	18	26	25	20	24	23	24	21	20	18
16	24	27	28	27	30	15	29	26	31	29	31	30	31	31	28	23
16	31	25	22	30	31	30	31	28	30	7	29	28	30	29	31	26
30	4	10	18	25	9	4	6	6	17	29	15	22	19	28	18	30
22	14	5	14	31	9	11	22	10	28	29	25	21	10	11	29	29
26	9	29	11	8	20	8	9	20	17	20	15	19	4	11	17	5
3	30	28	15	18	28	31	30	27	17	7	19	18	29	11	23	21
3	18	7	29	18	20	26	20	20	6	20	10	14	15	11	9	7
8	28	13	30	8	20	27	26	15	17	7	4	16	19	11	12	10
9	25	24	13	8	20	19	22	10	25	20	4	6	24	11	10	13
10	21	19	20	8	20	22	17	6	17	7	10	5	19	11	7	14
21	18	8	26	8	9	15	12	16	17	20	22	11	28	11	4	8
26	6	9	27	8	20	8	6	14	6	7	10	7	4	11	3	4
19	12	16	19	8	9	10	10	17	17	7	10	16	19	11	6	6
26	8	23	31	8	9	4	6	18	17	7	19	9	10	11	5	3
21	2	14	2	8	9	1	1	6	6	7	4	4	4	11	1	1
26	4	4	18	8	9	10	4	2	17	7	4	3	4	11	2	2
26	21	12	23	18	9	13	14	6	6	7	4	13	19	11	21	12





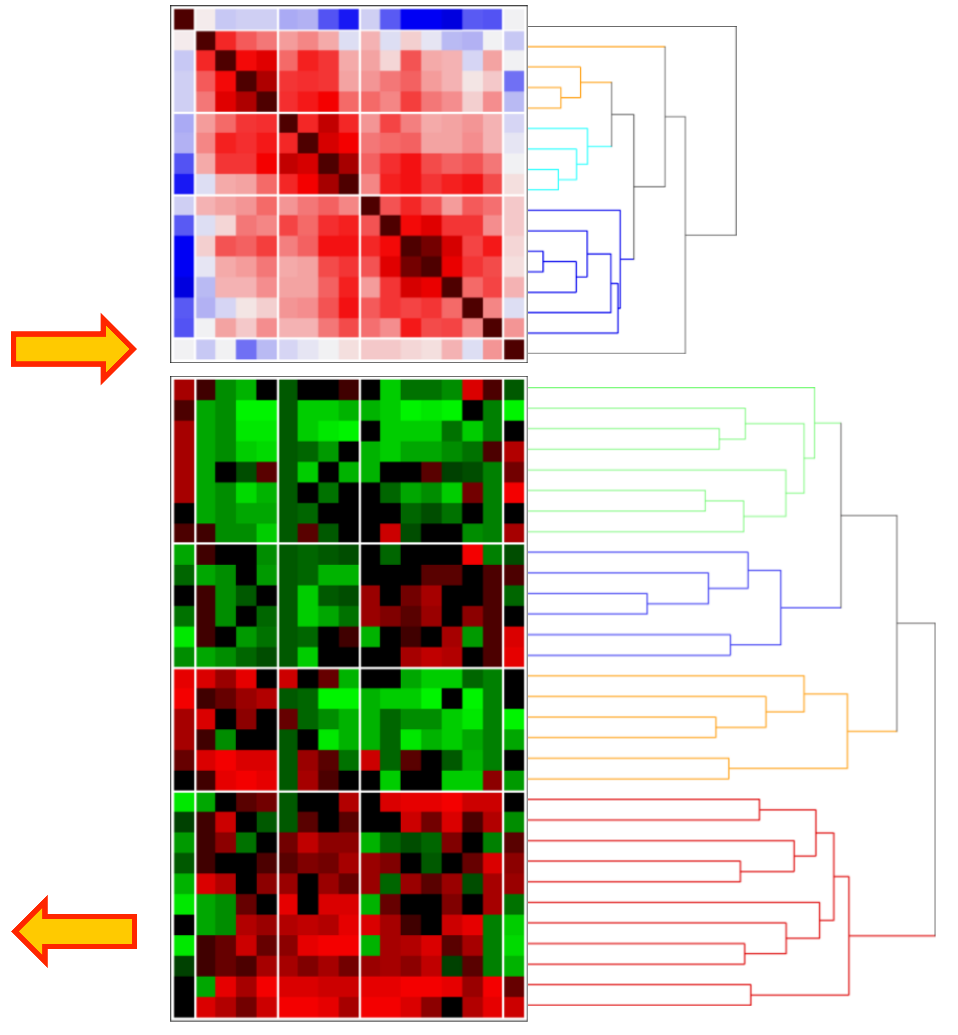
Correlation Matrix
-1
1
YEAR
CAREER
SCIENT
EQUIP
OPERAT
PATENT
POSTDAS
POSTD
THEMAT
INVEST
JUNIOR
PERSON
RESEAR
ADMINIS
SENIOR
OTHER
FEMALE
DIVISION
Humanity / Social
Math / Physical
Life Science
max: 95.06
95.06
81.9
68.73
55.57
42.41
29.25
16.09
2.92
0.0
min: 0.0
PREP
CENT
DIVISI
BIODIVER
POLITICAL
IURISPRUD
TAIW HIST
HUMA SOC
LINGUISTIC
SOCIOLOG
CH LIT PHIL
MATHEMAT
EURO AME
STATISTICS
ECONOMIC
ETHNOLOG
MOD HIST
AGR BIOT
INF TEC IN
APPL SCIE
ENVI CHAN
GENOMICS
ASTN ASTP
HIST PHILO
EARTH SCI
CEL ORG B
MOLE BIO
PLA MICRO
CHEMISTRY
ATOM MOL
PHYSICS
INFORMATI
BIOMEDICA
BIO CHEM
Data Rank
1
31



YEAR
CAREER
SCIENT
EQUIP
OPERAT
PATENT
POSTDAS
POSTD
THEMAT
INVEST
JUNIOR
PERSON
RESEAR
ADMINIS
SENIOR
OTHER
FEMALE
PREPARAT
CENTER
DIVISION
DIVISION
Humanity / Social
Math / Physical
Life Science
Large
31
Data Rank
1
Small
BIODIVER
POLITICAL
IURISPRUD
TAIW HIST
HUMA SOC
LINGUISTIC
SOCIOLOG
CH LIT PHIL
MATHEMAT
EURO AME
STATISTICS
ECONOMIC
ETHNOLOG
MOD HIST
AGR BIOT
INF TEC IN
APPL SCIE
ENVI CHAN
GENOMICS
ASTN ASTP
HIST PHILO
EARTH SCI
CEL ORG BI
MOLE BIO
PLA MICRO
CHEMISTRY
ATOM MOL
PHYSICS
INFORMATI
BIOMEDICA
BIO CHEM

Are there major and minor Instit. in Academia Sinica?

6	18	31	12	18	9	15	14	12	17	20	10	11	10	11	8	19
3	29	26	3	22	9	22	27	31	6	20	30	31	26	24	22	28
3	15	17	9	8	26	24	15	29	6	7	15	29	22	30	19	22
13	23	20	8	28	27	29	28	21	17	20	22	18	15	11	11	17
13	28	21	5	22	9	13	24	23	25	20	24	26	26	26	26	20
14	26	18	10	8	20	19	23	12	25	20	4	11	15	11	14	11
18	18	30	4	8	9	28	20	30	29	7	28	27	26	27	25	27
16	14	4	7	30	24	4	17	13	17	20	26	20	4	11	30	31
26	8	2	2	18	9	4	8	6	6	29	10	8	10	22	15	24
26	5	6	6	8	9	4	2	6	6	20	19	3	10	11	13	15
31	1	1	16	22	9	19	4	2	6	20	10	1	4	11	27	25
5	21	12	25	27	26	25	25	22	25	29	19	25	10	25	24	16
7	10	15	21	24	9	24	12	24	6	20	28	24	10	23	16	9
11	11	22	24	18	29	19	18	26	25	20	24	23	24	21	20	18
16	24	27	28	27	30	15	29	26	31	29	31	30	31	31	28	23
16	31	25	22	30	31	30	31	28	30	7	29	28	30	29	31	26
30	4	10	18	25	9	4	6	6	17	29	15	22	19	28	18	30
22	14	5	14	31	9	11	22	10	28	29	25	21	10	11	29	29
26	9	29	11	8	20	8	9	20	17	20	15	19	4	11	17	5
3	30	28	15	18	28	31	30	27	17	7	19	18	29	11	23	21
3	18	7	29	18	20	26	20	20	6	20	10	14	15	11	9	7
8	28	13	30	8	20	27	26	15	17	7	4	16	19	11	12	10
9	25	24	13	8	20	19	22	10	25	20	4	6	24	11	10	13
10	21	19	20	8	20	22	17	6	17	7	10	5	19	11	7	14
21	18	8	26	8	9	15	12	16	17	20	22	11	28	11	4	8
26	6	9	27	8	20	8	6	14	6	7	10	7	4	11	3	4
19	12	16	19	8	9	10	10	17	17	7	10	16	19	11	6	6
26	8	23	31	8	9	4	6	18	17	7	19	9	10	11	5	3
21	2	14	2	8	9	1	1	6	6	7	4	4	4	11	1	1
26	4	4	18	8	9	10	4	2	17	7	4	3	4	11	2	2
26	21	12	23	18	9	13	14	6	6	7	4	13	19	11	21	12



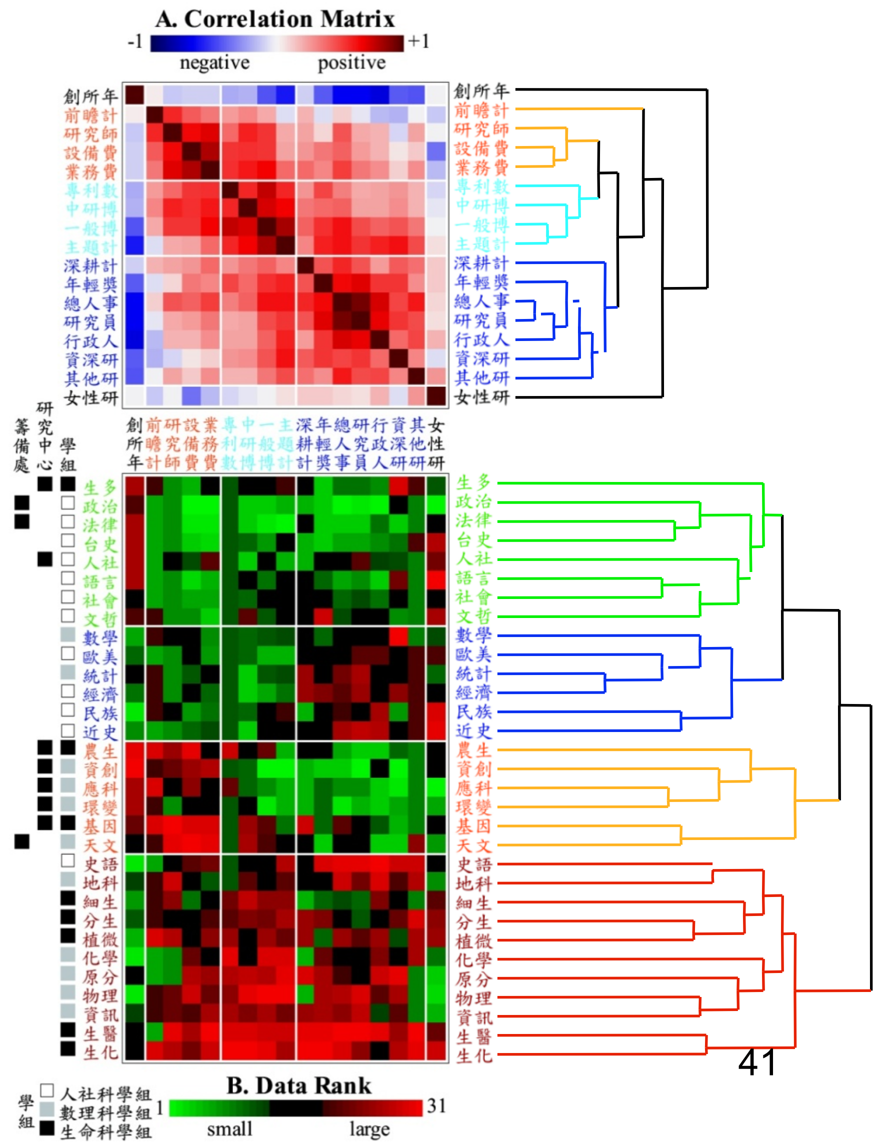
Formulate more appropriate Hypothesis for answering that question.

矩陣視覺化一個例子: 中研院有大小所嗎?

It is important to understand what you **CAN DO** before you learn to measure how **WELL** you seem to have **DONE** it.

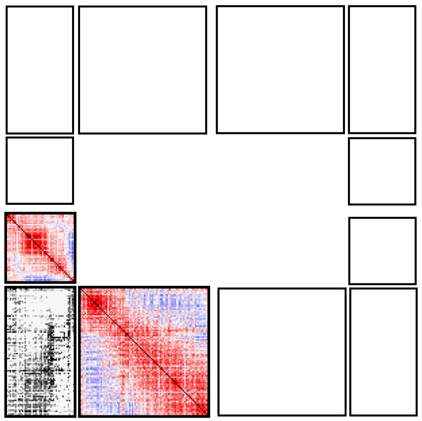
The greatest value of a picture is when it ***forces*** us to notice what we **never expected to see**.

Name	Div	Cen	Pre	Yr	RF	RFF	RFS	RO	RS	ST	PN	PD	PD	CR	IV	TM	JR	PT	AD	EQ
Math	0	0	0	6	17.5	12	31	8.5	17.5	15	13.5	10	11	19.5	16.5	11.5	10	10.5	8	19
Phys	0	0	0	2.5	29	3	26	8.5	22	21.5	27	30	31	19.5	5.5	31	26	24	22	28
Chem	0	0	0	2.5	15	9	17	25.5	7.5	23.5	15	15	29	6.5	5.5	29	22	30	19	22
[illegible]																				17
[illegible]																				20
[illegible]																				11
[illegible]																				27
[illegible]																				31
[illegible]																				24
[illegible]																				15
[illegible]																				25
[illegible]																				16
[illegible]																				9
[illegible]																				18
[illegible]																				23
[illegible]																				26
[illegible]																				30
[illegible]																				29
[illegible]																				5
[illegible]																				21
[illegible]																				7
[illegible]																				10
[illegible]																				13
[illegible]																				14
[illegible]																				8
[illegible]																				4
[illegible]																				6
[illegible]																				3
Poli	2	0	1	20.5	2	1.5	14	8.5	7.5	1	1	3.5	4	6.5	5.5	5.5	3.5	10.5	1	1
Iuri	2	0	1	26	3.5	17.5	3.5	8.5	7.5	9.5	3.5	3.5	2.5	6.5	16.5	1.5	3.5	10.5	2	2
Huma	2	1	0	26	21	23	11.5	8.5	17.5	12.5	13.5	3.5	13	6.5	5.5	5.5	19	10.5	21	12



allow the **data to speak** for themselves before standard assumptions or formal modeling

2.permutation



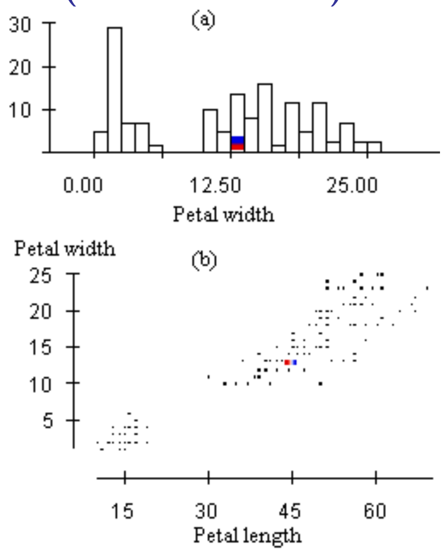
Approaching Statistics

Purpose of permutations

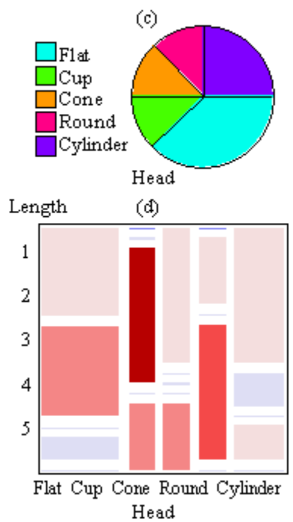
Relativity of a statistical graph (Chen 2002)

Concept of placing similar (different) objects at closer (distant) positions

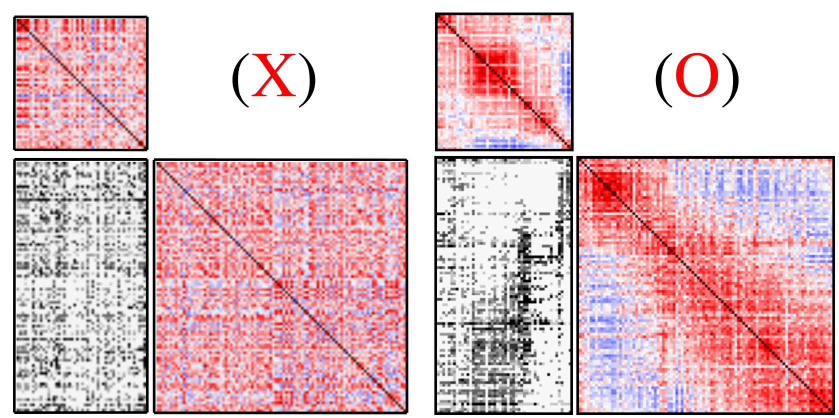
1. Continuous (V) (Fisher / Iris)



2. Categorical (X) (Hartigan / Tools)



3. Matrix Visualization (X) (MPGRP)



Friendly & Kwan (2003): effect-ordered data display
Hurley (2004): interesting displays in prominent positions

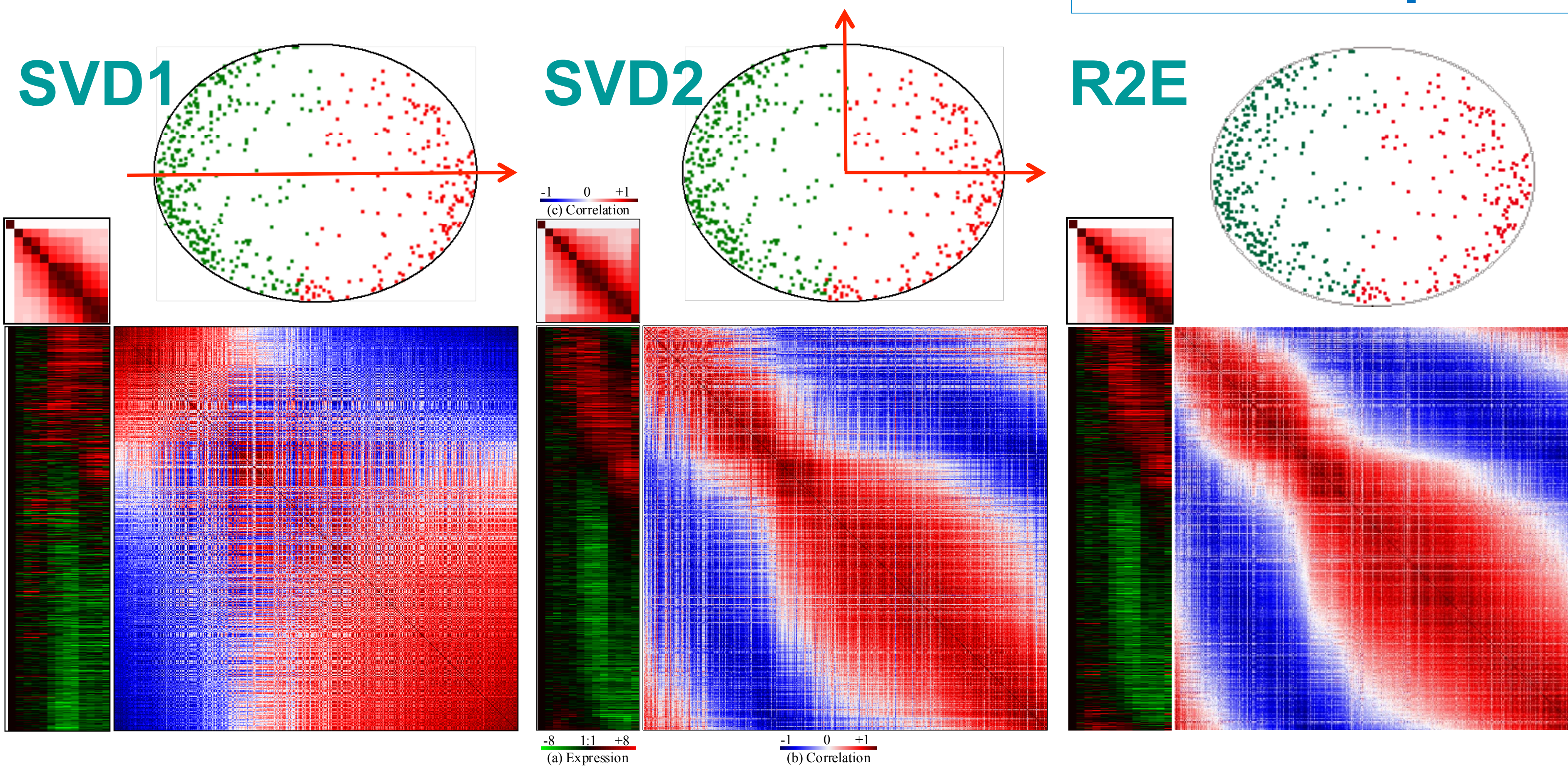
Statistical Approach

Identify Global Trend: Singular Value Decomposition

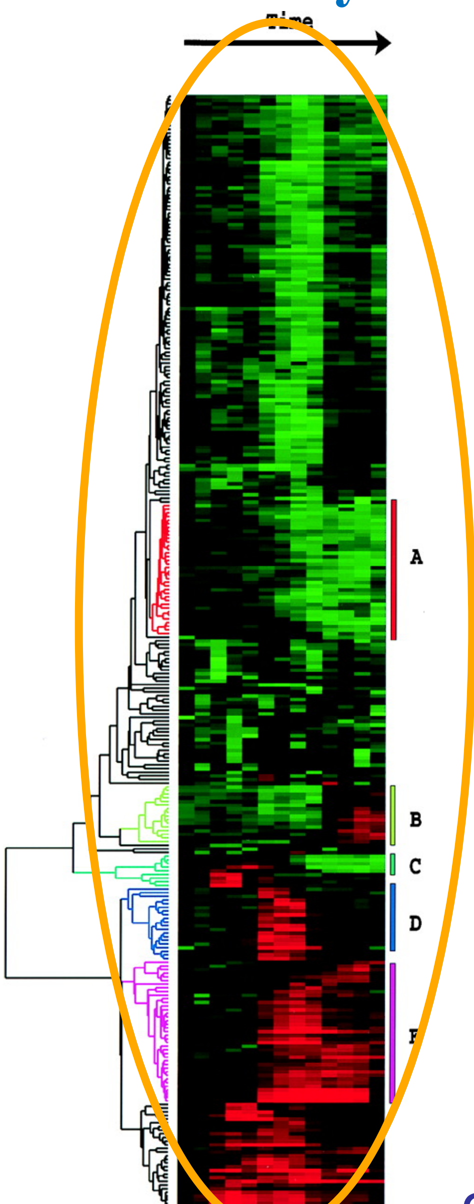
SVD

Alter O. et al 2000, PNAS

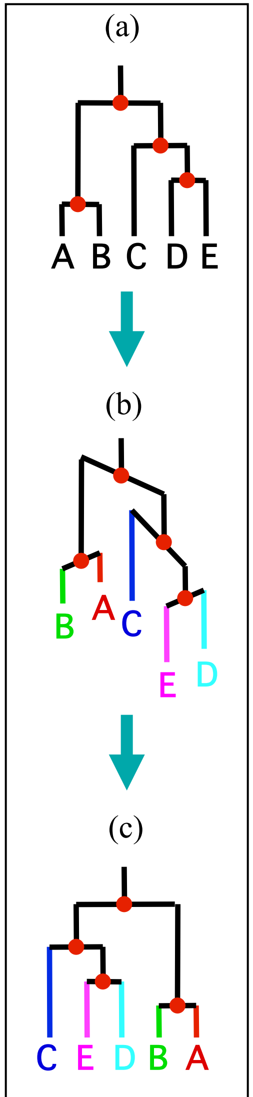
Chen 2002, Statistica Sinica
Rank 2 Elliptical



Statistical Approach:
Identify Local Clusters



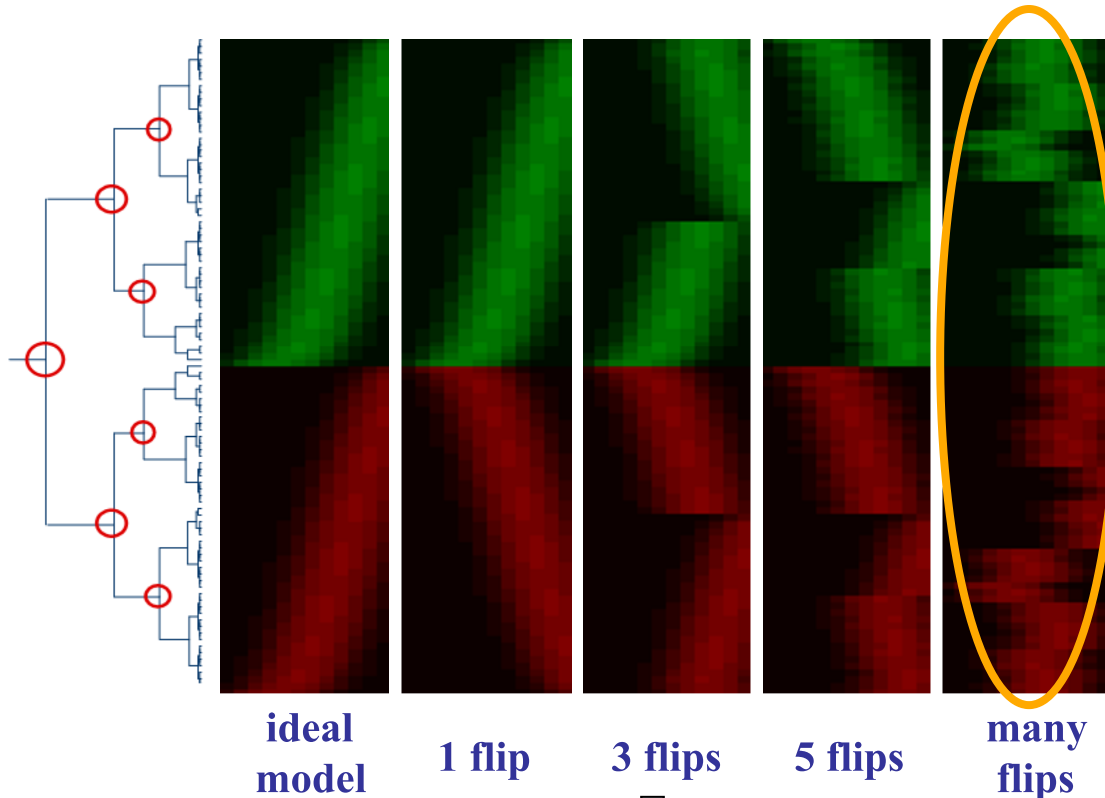
Eisen et al. (1998)



$2^{n-1}=2^{5-1}=16$

Tree seriation & flipping of intermediate nodes

Different Seriations (Ordering of Terminal Nodes or Leaves) Generated from Identical Tree Structure



external and internal references for guiding flipping mechanism

Approaching Statistics & Statistical Approach

BMC Bioinformatics

BioMed Central

Methodology article

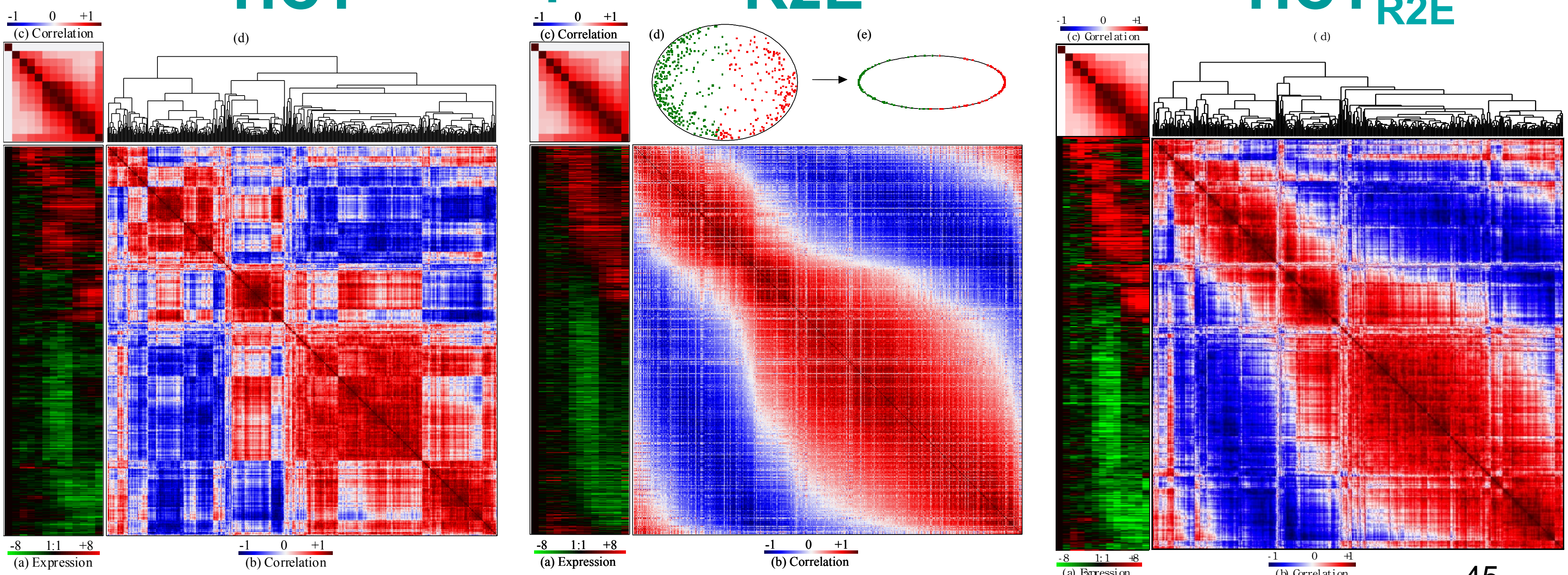
Open Access

Methods for simultaneously identifying coherent local clusters with smooth global patterns in gene expression profiles

Yin-Jing Tien[1], Yun-Shien Lee[2,3], Han-Ming Wu[4] and Chun-Houh Chen*[5]

Address: [1]Institute of Statistics, National Central University, Tao-Yuan, 32001, Taiwan, [2]Genomic Medicine Research Core Laboratory, Chang Gung Memorial Hospital (CGMH), Tao-Yuan, 33305, Taiwan, [3]Department of Biotechnology, Ming Chuan University, Tao-Yuan, 33348, Taiwan, [4]Department of Mathematics, Tamkang University, Tamsui 25137, Taiwan and [5]Institute of Statistical Science, Academia Sinica, Taipei, 11529, Taiwan

HCT + R2E = HCT_{R2E}



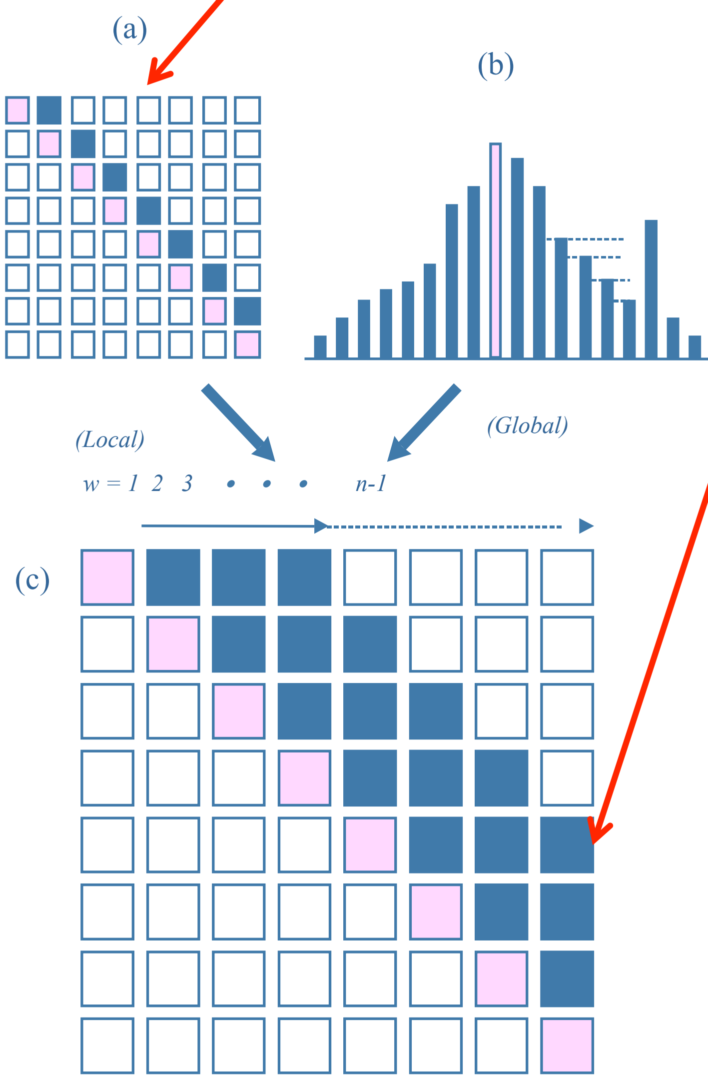
Hierarchical Tree Seriation | GAP Elliptical (R2E) Seriation | Tree guided by (R2E)

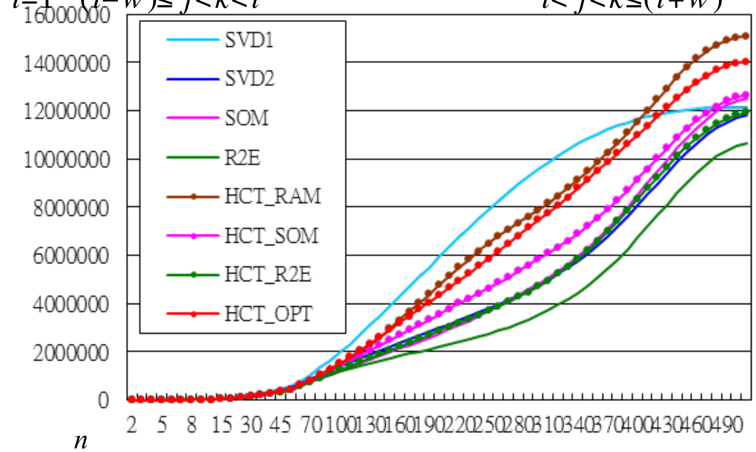
Statistical Approach: Evaluation of permutation algorithms

The Generalized anti-Robinson (GAR) criterion

$$AR = \sum_{i=1}^{n}[\sum_{j<k<i} I(d_{ij} < d_{ik}) + \sum_{i<j<k} I(d_{ij} > d_{ik})]$$

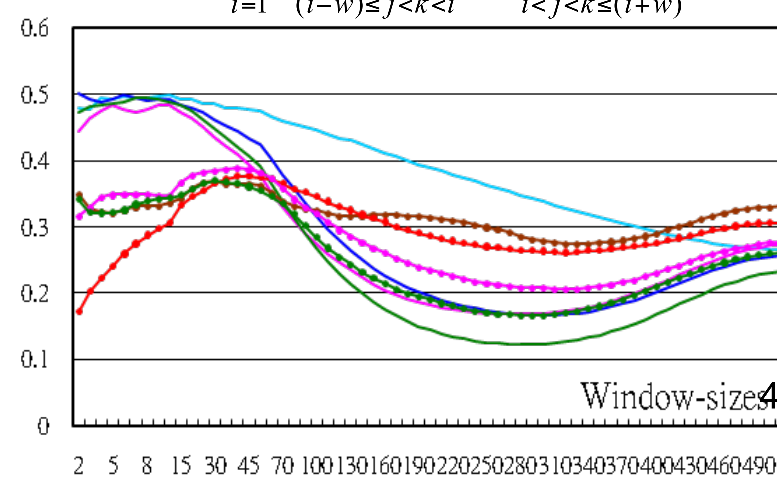
$$GAR = \sum_{i=1}^{n}[\sum_{(i-w)\le j<k<i} I(d_{ij} < d_{ik}) + \sum_{i<j<k\le(i+w)} I(d_{ij} > d_{ik})]$$





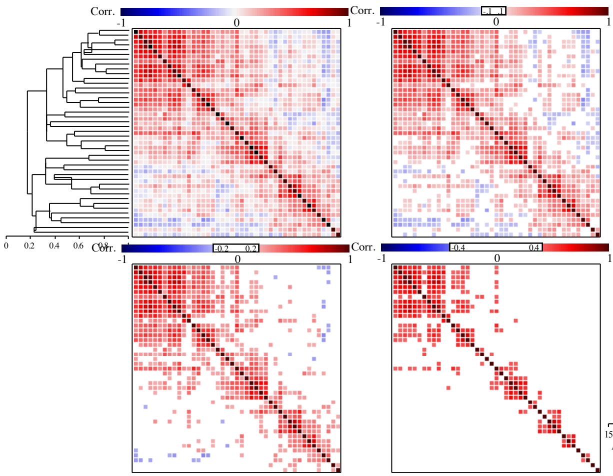
Relative ***GAR***

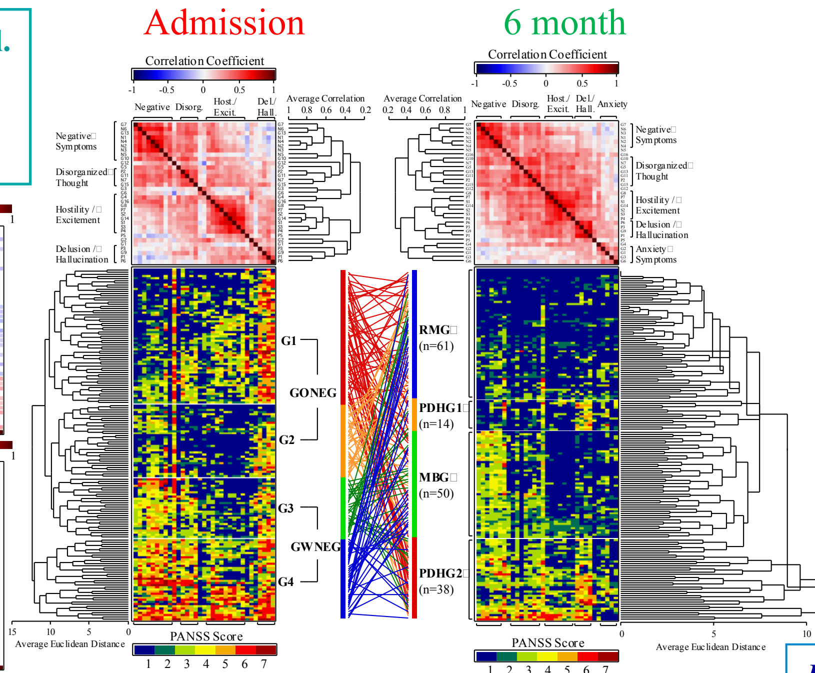
$$RGAR = \frac{\sum_{i=1}^{n}[\sum_{(i-w)\le j<k<i} I(d_{ij} < d_{ik}) + \sum_{i<j<k\le(i+w)} I(d_{ij} > d_{ik})]}{\sum_{i=1}^{n}[\sum_{(i-w)\le j<k<i} 1 + \sum_{i<j<k\le(i+w)} 1]}$$



GAP for Heritable (Genetic) Disease: Schizophrenia (National Taiwan University)

Psychiatry Research (1998) **Lin, Chen et al.**
Psychopathological Dimensions in Schizophrenia: A Correlational Approach to Items of the SANS and SAPS





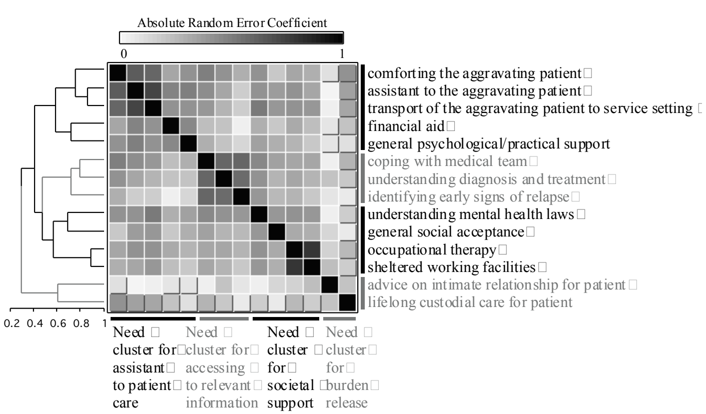
Admission **Hwu et al.**
Schizophrenia Research (2002)
Symptom Patterns and Subgrouping of Schizophrenic Patients: Significance of Negative Symptoms Assessed on Admission

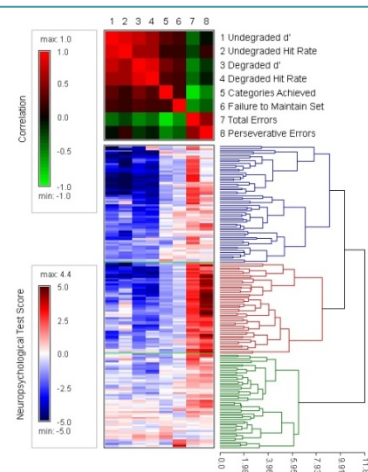
6 month **Liu et al.**
J. of the Formosan Med. Ass.
Validity of a 3-Subtype Model of Schizophrenia: Symptomatology, Social Function, and Neuropsychological Impairment

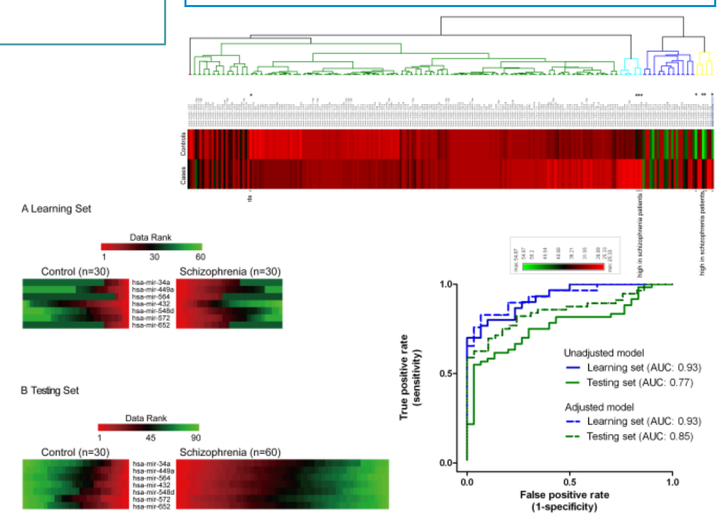
PLoS ONE (2011) **Lai et al.**
MicroRNA expression aberration as potential peripheral blood biomarkers for schizophrenia

J. of the Formosan Med. Ass. (2008) **Yeh et al.**
Factors Related to Perceived Needs of Chief Caregivers of Patients with Schizophrenia

Genes, Brain and Behavior (2009) **Lin et al.**
Clustering by neurocognition for fine-mapping of the schizophrenia susceptibility loci on chromosome 6p



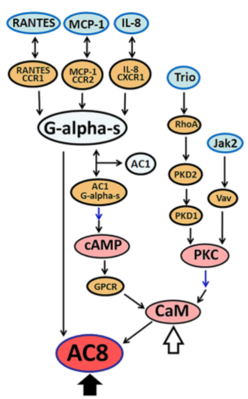


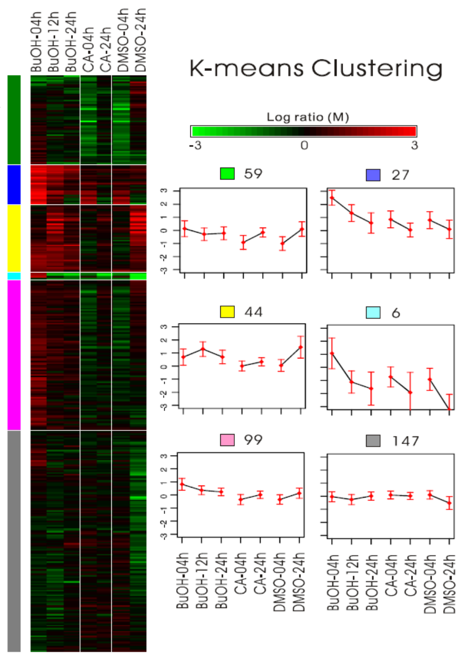


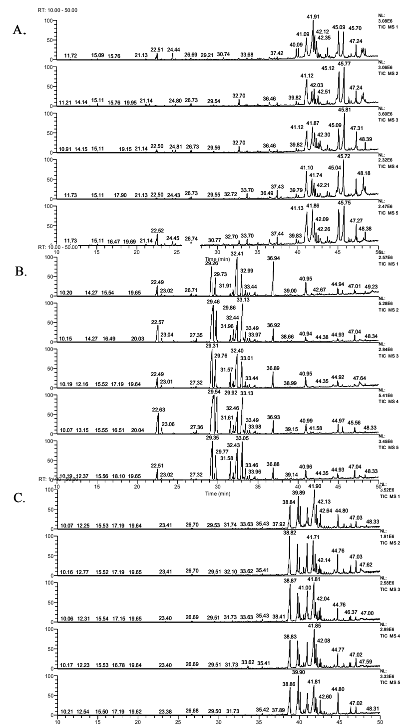
GAP for Comparative Metabolome: Chinese Herbal Medicine

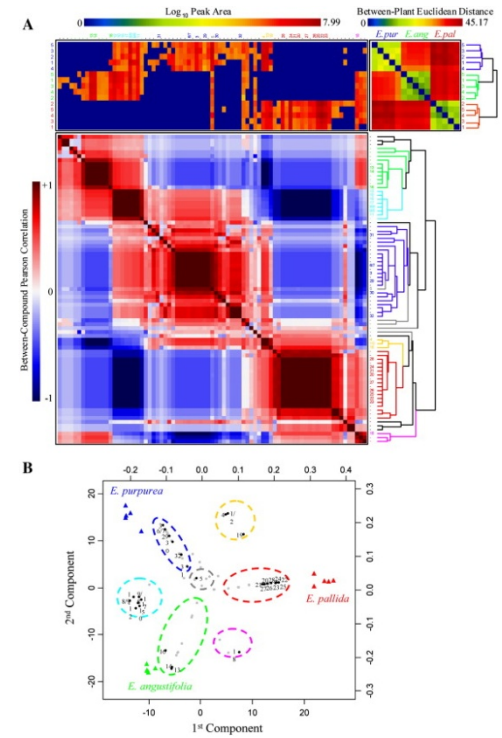
Drs. Ning-Sun Yang, Lie-Fen Shyur, Wen-Chin Yang
Agricultural Biotechnology Research Center (ABRC) of Academia Sinica

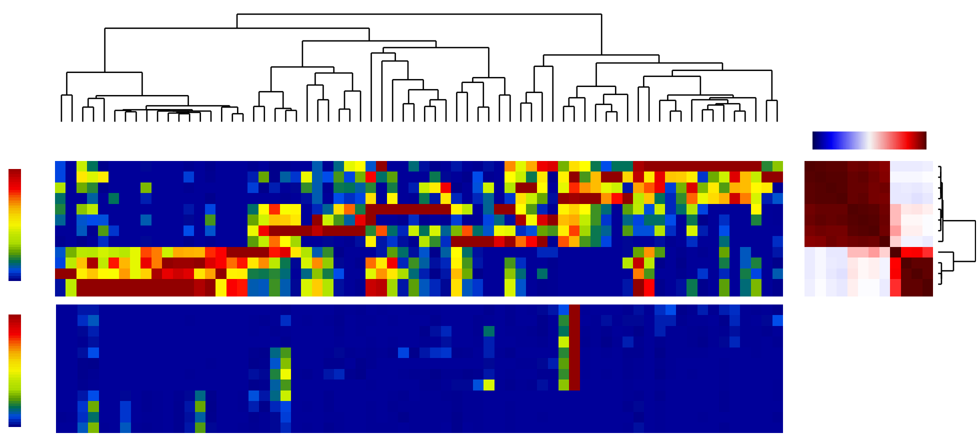
BMC Genomics **9** (2008)
Genomics and proteomics of immune modulatory effects of a butanol fraction of Echinacea purpurea in human dendritic cells **Wang et al.**









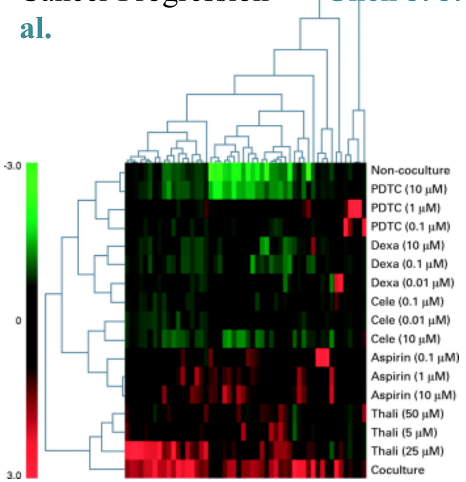


Phytochemistry **70** (2009)
Anti-diabetic properties of three common Bidens pilosa variants in Taiwan
Chien et al.

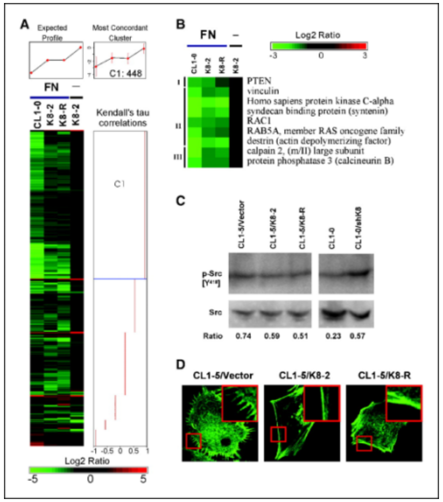
Journal of Nutritional Biochemistry (2010)
Comparative metabolomics approach coupled with cell- and gene-based assays for species classification and anti-inflammatory bioactivity validation of Echinacea plants
Hou et al.

GAP for Cancer Study: Non–Small Cell Lung Cancer (National Taiwan University)

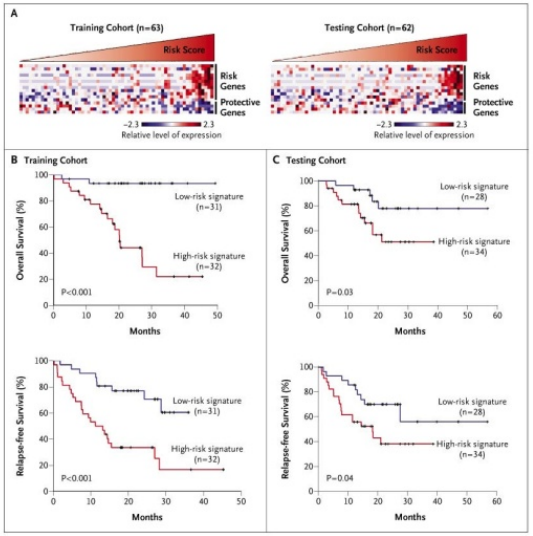
Journal of Clinical Oncology **23** (2005)
Tumor-Associated Macrophages in Cancer Progression **Chen J. J. et al.**



Cancer Research **66** (2006)
Non–Small Cell Lung Cancer with Tumor Cell Invasiveness **Sher Y. P. et al.**

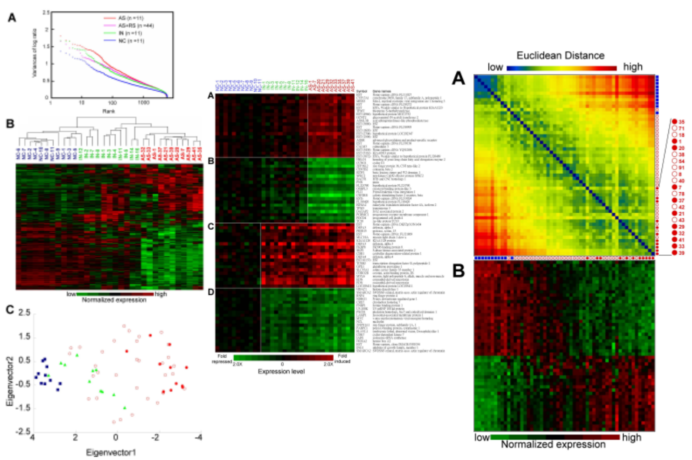


The New England Journal of Medicine **356** (2007) A Five-Gene Signature and Clinical Outcome in Non–Small-Cell Lung Cancer **Chen H. Y. et al.**



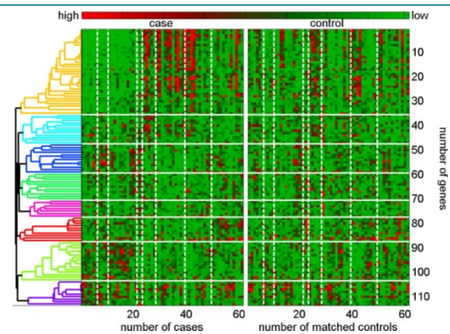
GAP for Infectious Disease: SARS (Chang Gung Memorial Hospital)

BMC Genomics **6** (2005)Molecular signature of clinical severity in recovering patients with (SARS-CoV) **Lee Y. S. et al. (Chang Gung Hospital)**

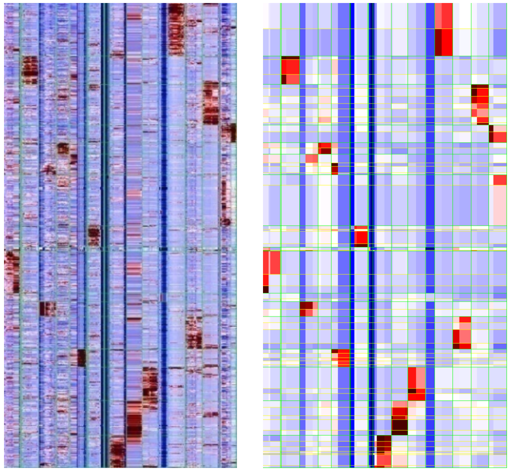


GAP for Endophenotypess/Nutrition (Academia Sinica)

Genetic Epidemiology **30** (2006)
Using endophenotypes for pathway clusters to map complex disease genes
Pan W. H. et al. (Academia Sinica)

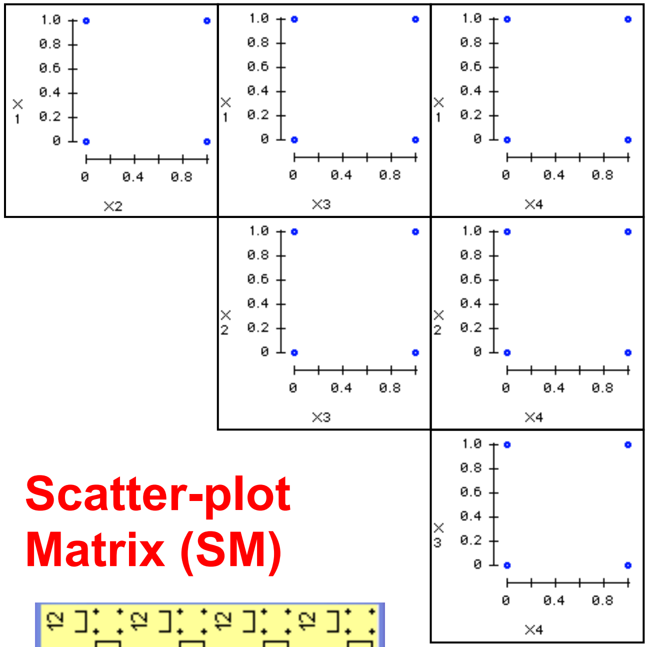


Nutritional Sciences Journal **30** (2006)
Evaluating the DOH Food Guide Based on Taiwanese Food Choices **Pan W. H. et al.**

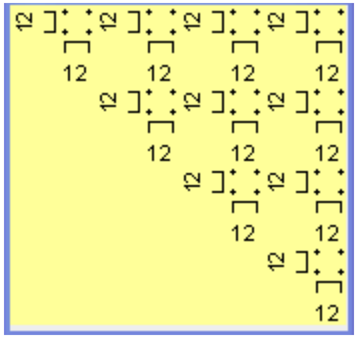


Matrix visualization of binary data

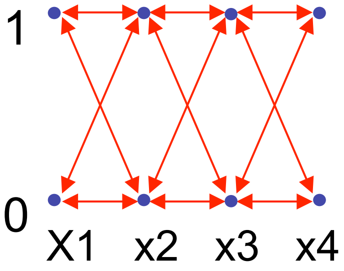
Graphic tools for high-dimensional non continuous data visualization w/o dimension reduction

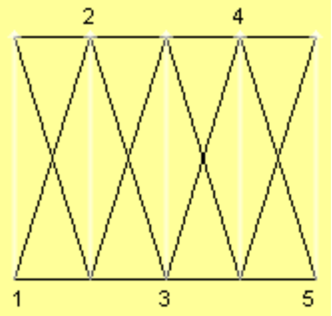


Scatter-plot Matrix (SM)

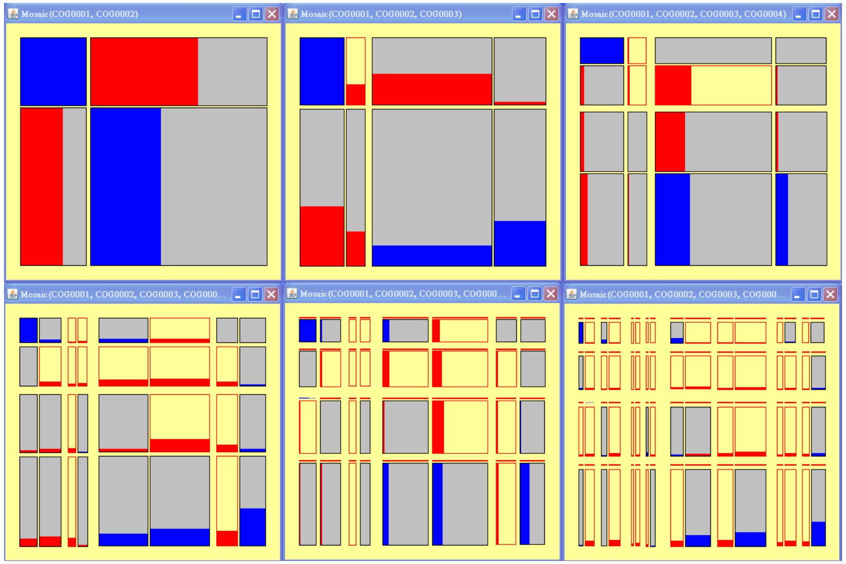


Parallel Coordinates Plot (PCP)





Mosaic Plot (Display)



Approaching Statistics

Essential elements in a GAP MV procedure?

Continuous

3. Variable Proximity

Correlation
Covariance
polychoric
Correlation . . .

3. Variable Proximity

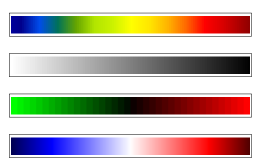
Binary

		Object B		
		1	0	
Object A	1	a	b	$(a+b)$
	0	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	$(a+b+c+d)$

2. Subject Proximity

Euclidean Distance
Manhattan Distance
Correlation …

1. Data Matrix



1. Data Matrix

?

2. Subject Proximity

Statistical Approach

Selection of Proximity Measures for Matrix Visualization of Binary Data

Tzeng, S. L., Wu, H. M., and Chen, C. H. (2009)

Proc. 2009 2nd Int'l Conf. on BioMed. Engin. & Info. (BMEI 2009), Tianjin, China

(available in *IEEE Xplore Digital Library*)

Table 1. Commonly used similarity coefficients for binary data.

		Object B: 1	Object B: 0	
Object A	1	a	b	$(a+b)$
	0	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	$(a+b+c+d)$

Similarity	Formula
Braun	$\frac{a}{\max(a+b,\ a+c)}$
Dice	$\frac{2a}{2a+b+c}$
Hamman	$\frac{a+d-(b+c)}{a+b+c+d}$
Jaccard	$\frac{a}{a+b+c}$
Kappa	$\left(1+\frac{(b+c)(a+b+c+d)}{2ad-2bc}\right)^{-1}$
Kulczynskl	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$

Similarity	Formula
Ochiai	$\frac{a}{\sqrt{((a+b)(a+c))}}$
Phi	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
Rao	$\frac{a}{a+b+c+d}$
Rogers	$\frac{a+d}{a+2b+2c+d}$
simple match	$\frac{a+d}{a+b+c+d}$
Simpson	$\frac{a}{\min(a+b,\ a+c)}$
Sneath	$\frac{a}{a+2b+2c}$
Yule	$\frac{ad-bc}{ad+bc}$

Two issues for selecting similarity measures for binary:

I. Symmetric or Asymmetric :

		Object B 1	Object B 0	
Object A	1	a	b	$(a+b)$
	0	c	d	$(c+d)$
		$(a+c)$	$(b+d)$	$(a+b+c+d)$

SYMMETRIC: if both of its categories are equally important, i.e., there is no preference on which outcome should be coded as 0 or 1. Gender is an typical example of symmetric binary variable. (♀/ ♂, Bioif/Biost, ….)

Symmetric binary variables should be treated as nominal ones.

Similarity measures: often a function of both the co-occurrence and co-absence frequencies between two variables

e.g., simple matching $\frac{a+d}{a+b+c+d}$ Rogers $\frac{a+d}{a+2b+2c+d}$ Hamman $\frac{a+d-(b+c)}{a+b+c+d}$

ASYMMETRIC if the outcomes of the two states are not equally important, such as the positive and negative outcomes of a disease diagnosis. Conventionally the most important outcome, which is usually the uncommon one is coded by 1 and the other by 0.

Therefore, asymmetric binary variables are often considered "monary" (as if there is only one state)

Similarity measures: a function of co-occurrence frequencies,

e.g. $\frac{2a}{2a+b+c}$ $\frac{a}{a+b+c}$ $\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$ $\frac{a}{\sqrt{((a+b)(a+c))}}$ $\frac{a}{\max(a+b,\ a+c)}$ $\frac{a}{\min(a+b,\ a+c)}$ $\frac{a}{a+2b+2c}$

Both symmetric and asymmet $\left(1+\frac{(b+c)(a+b+c+d)}{2ad-2bc}\right)^{-1}$ Phi $\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$ Yule $\frac{ad-bc}{ad+bc}$

Binary GAP Example

CGMIM *Online*

http://www.bccrc.ca/ccr/CGMIM/

CGMIM performs automated text-mining of OMIM to identify genetically-related cancers

Online Mendelian In Man (OMIM) is a computerized database of information about genes and heritable traits in human populations

OMIM is maintained on the Internet by the National Center for Biotechnology Information at the US National Institutes of Health

CGMIM considers 21 anatomic sites based on the major cancers identified by the National Cancer Institute of Canada

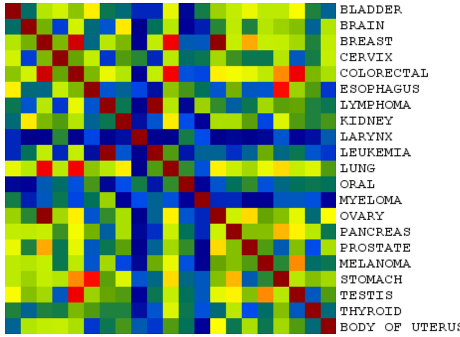
CGMIM compares each OMIM entry name and alternative name with a list of gene names assigned by HUGO (HUman Genome Organization).

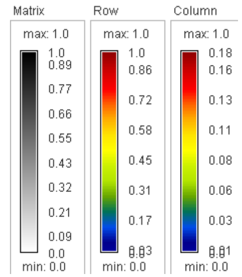
CGMIM produces the number of genes for which an OMIM entry mentions each pair of cancers, as well as a ratio of the observed ... genes for the combination

BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

21 Cancer Sites

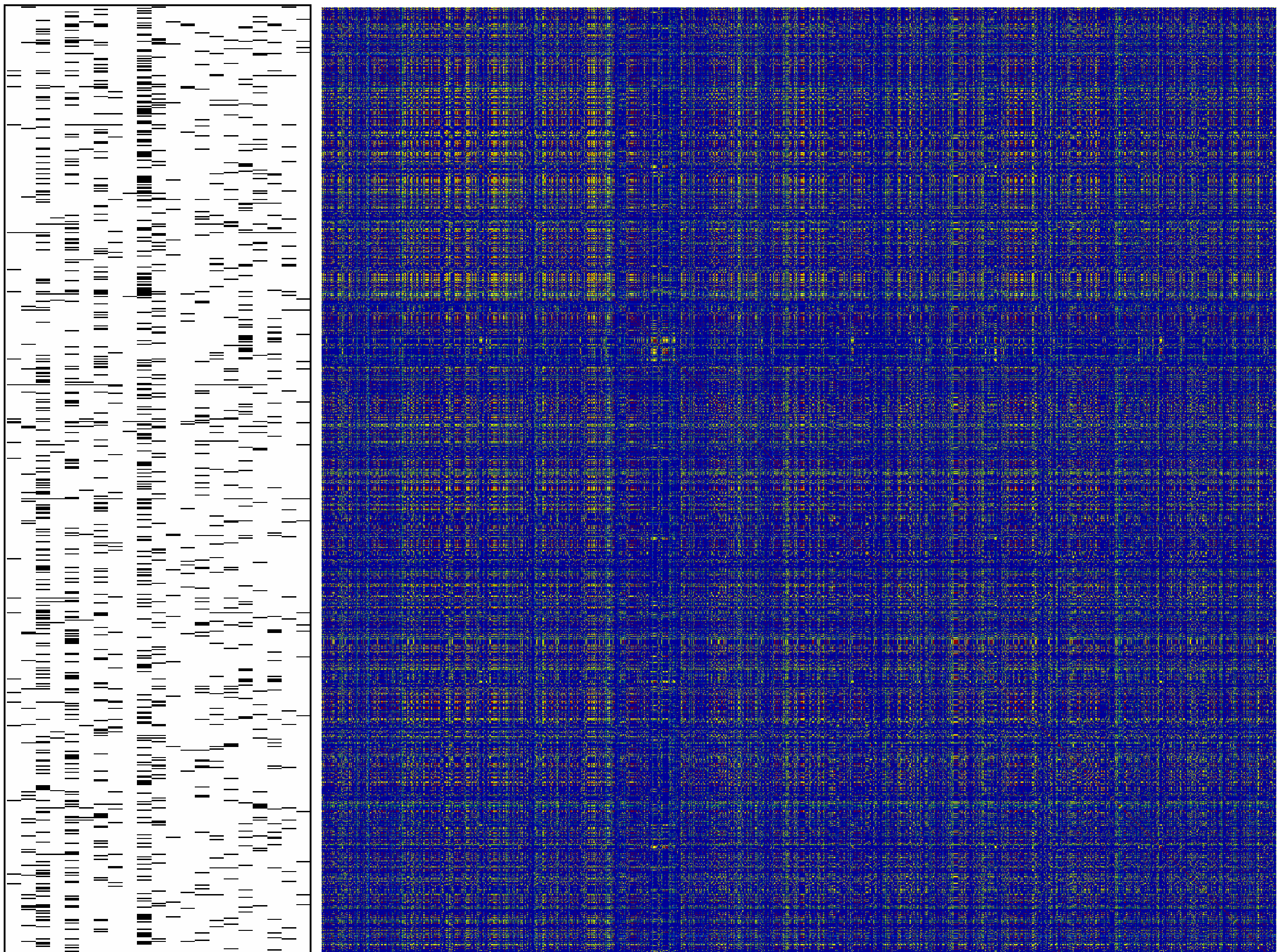
BLADDER
BRAIN
BREAST
CERVIX
COLORECTAL
ESOPHAGUS
LYMPHOMA
KIDNEY
LARYNX
LEUKEMIA
LUNG
ORAL
MYELOMA
OVARY
PANCREAS
PROSTATE
MELANOMA
STOMACH
TESTIS
THYROID
BODY OF UTERUS





CGMIM
All Data (1948 genes * 21 Sites)
Original Order

Jaccard: a/(a+b+c)



1948 Related Genes

CGMIM

All Data (1948 genes * 21 Sites)

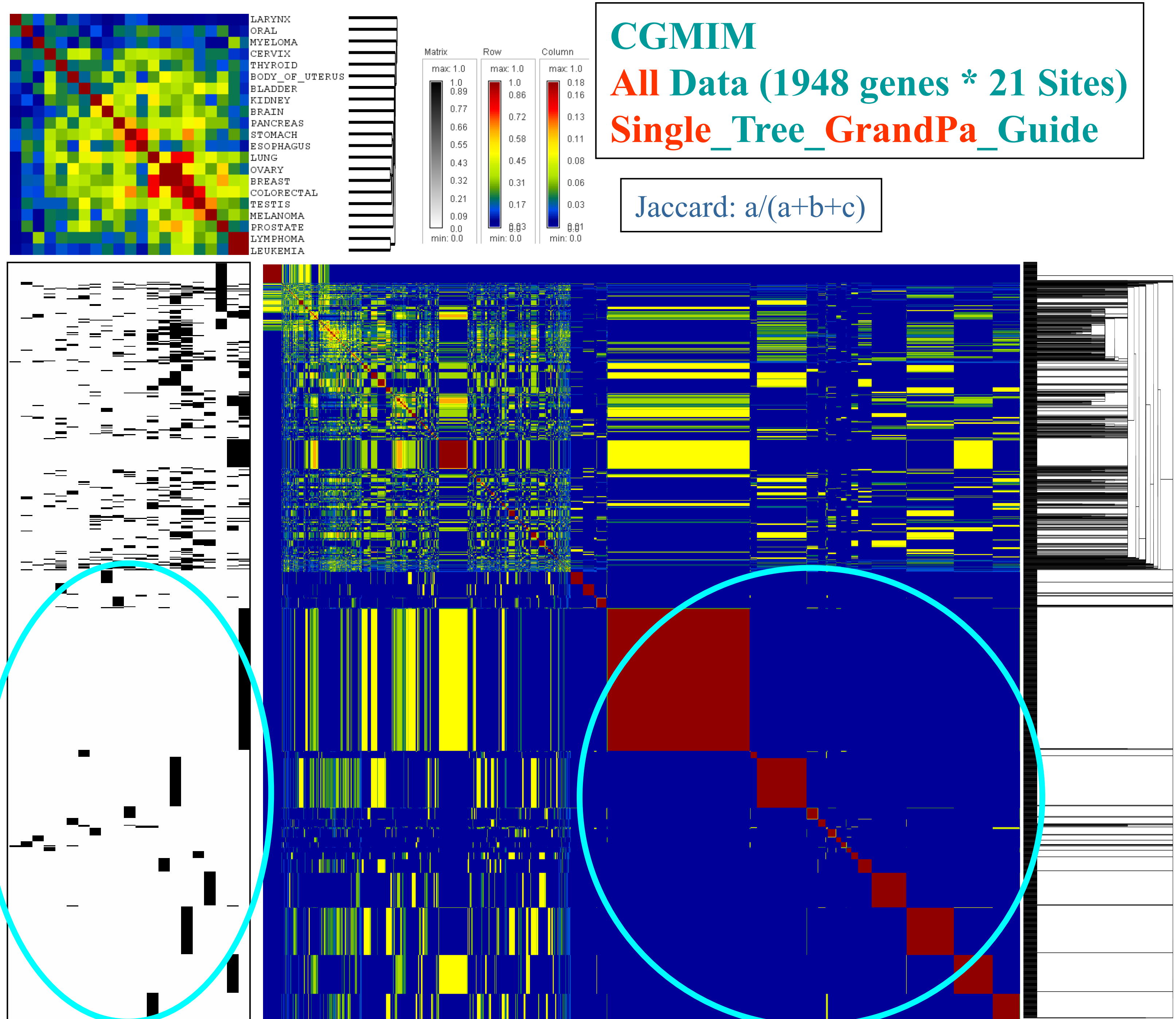
Single_Tree_GrandPa_Guide

Jaccard: a/(a+b+c)

21 Cancer Sites

LARYNX
ORAL
MYELOMA
CERVIX
THYROID
BODY_OF_UTERUS
BLADDER
KIDNEY
BRAIN
PANCREAS
STOMACH
ESOPHAGUS
LUNG
OVARY
BREAST
COLORECTAL
TESTIS
MELANOMA
PROSTATE
LYMPHOMA
LEUKEMIA

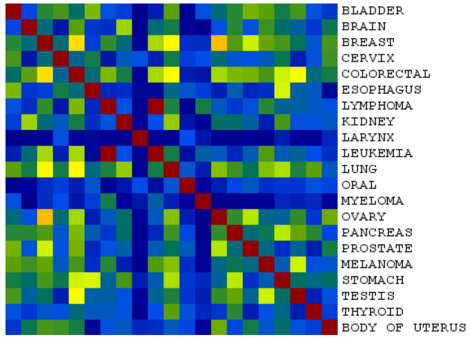
1948 Related Genes

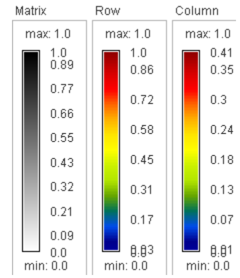


21 Cancer Sites

BLADDER
BRAIN
BREAST
CERVIX
COLORECTAL
ESOPHAGUS
LYMPHOMA
KIDNEY
LARYNX
LEUKEMIA
LUNG
ORAL
MYELOMA
OVARY
PANCREAS
PROSTATE
MELANOMA
STOMACH
TESTIS
THYROID
BODY OF UTERUS

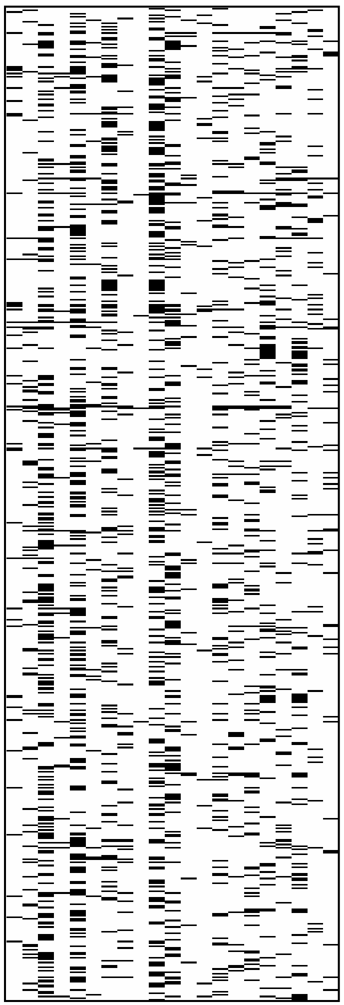
768 Related Genes

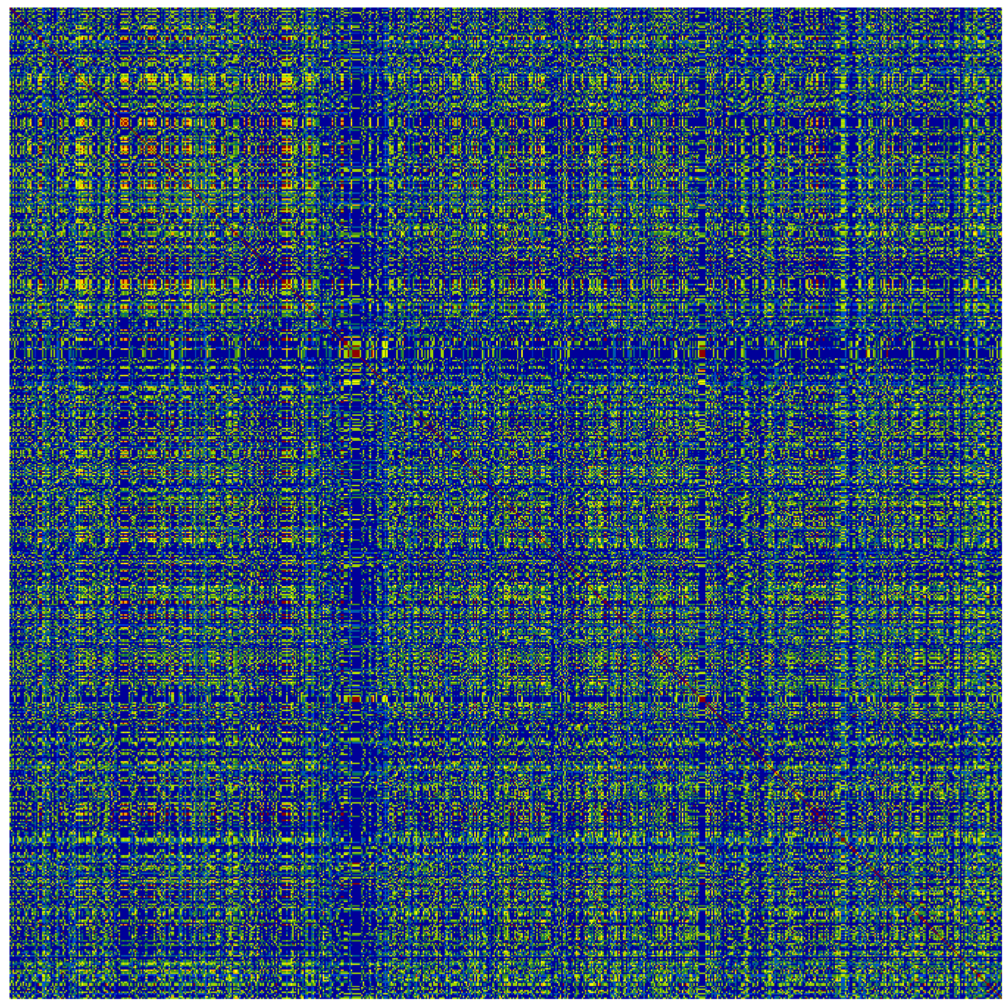




CGMIM
768 genes at least at 2 Sites
Original Order

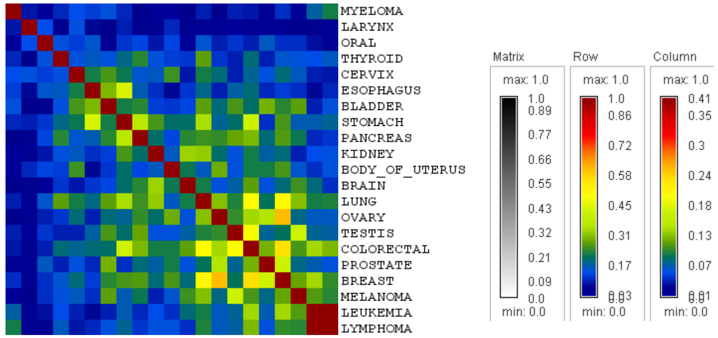
Jaccard: a/(a+b+c)





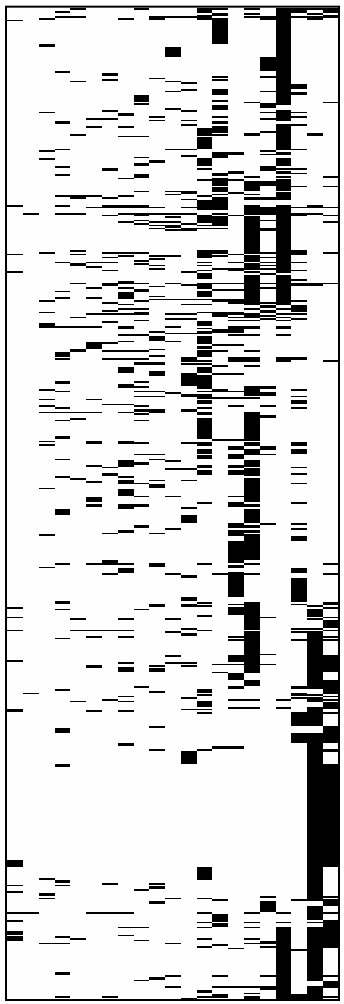
21 Cancer Sites

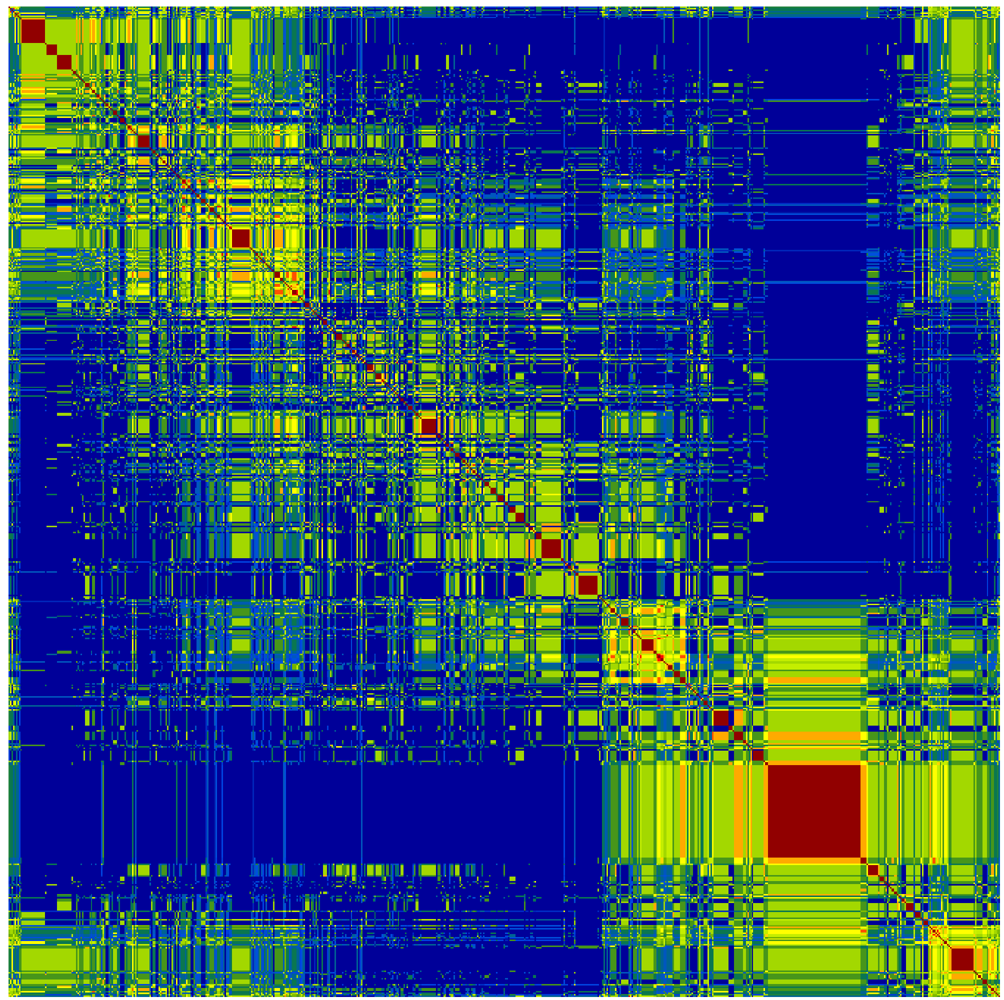
MYELOMA
LARYNX
ORAL
THYROID
CERVIX
ESOPHAGUS
BLADDER
STOMACH
PANCREAS
KIDNEY
BODY_OF_UTERUS
BRAIN
LUNG
OVARY
TESTIS
COLORECTAL
PROSTATE
BREAST
MELANOMA
LEUKEMIA
LYMPHOMA



CGMIM
768 genes at least at 2 Sites
GAP_Elliptical_Order

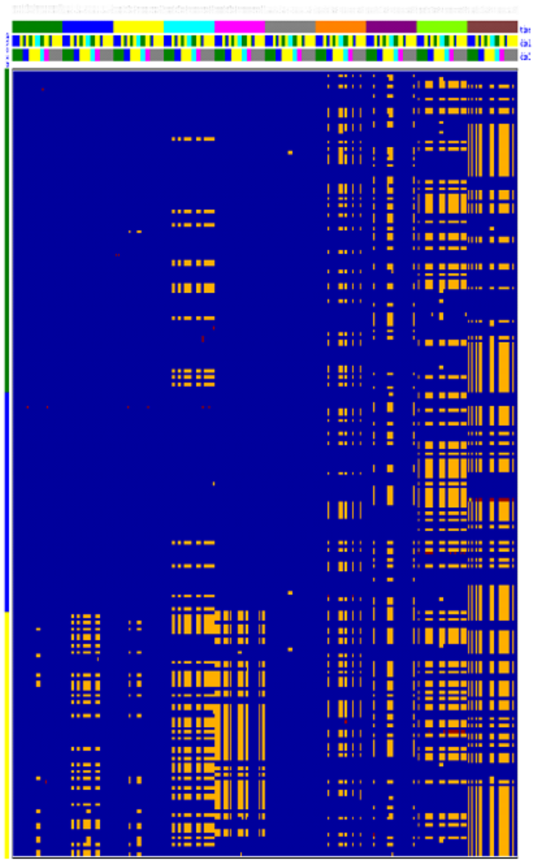
Jaccard: a/(a+b+c)



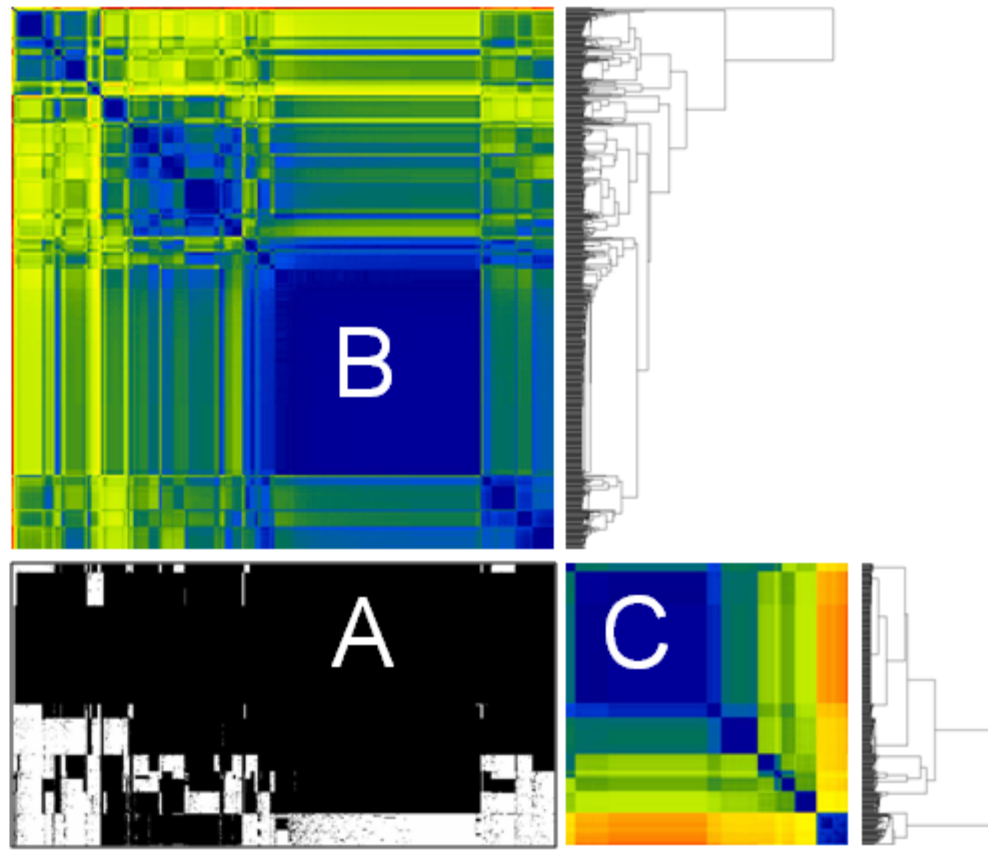


768 Related Genes

Matrix Visualization for data quality control and missing pattern exploration



Drop-out structure of schizophrenia **patients** in a 5-year follow-up study



Missing and multi-level **stratification** structure of individual **SNPs profiles** for patients with a certain disease (simulation data generated from parameters estimated from a real data).

Matrix visualization of nominal data (GAP approach)

Example:

Classification of Animals Data

Shizuhiko Nishisato 2006

A typical nominal data

Shizuhiko Nishisato, 2006

Classification of Animals

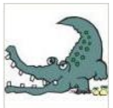
35 animals were sorted into piles of similar animals by 15 variables (Genotypes / Phenotypes ?)

What about 3500 samples 1500 variables ?

Animal/Subject	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
Alligator	8	1	6	9	6	1	3	4	1	4	3	4	3	6	4
Bear	6	3	2	6	6	1	3	4	4	5	5	1	4	2	2
Camel	4	3	9	3	1	4	3	5	4	2	5	1	7	7	8
Cat	6	3	7	4	0	1	1	2	3	3	1	1	6	3	5
Cheetah	3	3	7	4	0	1	3	5	4	3	6	1	6	2	2
Chiken	7	2	4	1	2	5	7	1	5	1	1	3	8	4	1
Chimpanzee	5	3	5	7	5	2	4	4	2	2	6	4	3	6	6
Cow	1	3	9	6	1	1	3	5	3	4	1	1	4	5	8
Crane	7	2	4	5	2	5	5	1	5	1	2	3	8	4	1
Crow	7	2	4	5	2	5	5	1	5	1	2	3	8	4	1
Dog	6	3	7	10	0	2	1	2	3	3	1	1	4	3	5
Duck	7	2	4	1	2	5	5	1	5	1	2	3	8	4	1
Elephant	4	3	6	3	1	4	3	5	4	5	3	1	7	7	2
Fox	6	3	7	4	0	1	6	2	3	3	3	1	4	3	5
Frog	8	1	3	2	3	3	2	3	1	4	4	2	1	1	3
Giraffe	1	3	8	3	1	4	3	5	4	2	5	1	7	7	8
Goat	3	3	9	6	1	4	6	5	3	3	1	1	5	3	5
Hawk	7	2	4	5	2	5	5	1	5	1	3	3	8	4	1
Hippopotamus	4	3	6	6	6	4	3	3	4	4	5	1	7	7	2
Horse	6	3	9	6	1	2	3	5	3	3	1	1	5	5	8
Leopard	1	3	7	4	0	1	3	5	4	3	3	1	6	2	2
Lion	5	3	7	4	6	1	3	5	4	3	3	1	7	2	2
Lizard	2	1	3	2	3	3	2	3	1	4	4	2	2	1	3
Monkey	6	3	5	7	5	2	4	4	2	2	6	4	3	6	6
Ostrich	3	2	4	1	2	5	3	1	5	1	5	3	8	7	8
Pig	1	3	9	6	1	1	6	5	3	3	1	1	5	5	5
Pigeon	7	2	4	5	2	5	5	1	5	1	2	1	8	4	1
Rabbit	6	3	1	6	0	4	6	2	3	3	1	1	5	3	5
Racoon	6	3	7	10	4	1	6	2	3	3	3	1	4	3	5
Rhinoceros	4	3	5	6	6	4	3	5	4	4	5	1	7	7	2
Snake	8	1	3	9	6	3	2	3	1	4	4	2	2	1	3
Sparrow	7	2	4	5	2	5	5	1	5	2	2	3	8	4	1
Tiger	5	3	7	4	0	1	3	5	4	3	3	1	6	2	2
Tortoise	8	1	3	9	3	3	2	3	1	5	4	2	1	1	3
Turkey	7	2	4	1	2	5	7	1	5	1	1	3	8	4	1

Alligator
Bear
Camel
Cat
Cheetah
Chiken
Chimpanzee
Cow
Crane
Crow
Dog
Duck
Elephant
Fox
Frog
Giraffe
Goat

Hawk
Hippopotamus
Horse
Leopard
Lion
Lizard
Monkey
Ostrich
Pig
Pigeon
Rabbit
Racoon
Rhinoceros
Snake
Sparrow
Tiger
Tortoise
Turkey


Alligator

Bear

Camel

Cat

Cheetah

Chicken

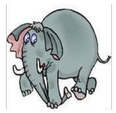
Cow

Crane

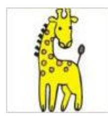
Chimpanzee

Crow

Dog

Duck

Elephant

Fox

Frog

Giraffe

Goat

Hawk

Hippopotamus

Horse

Leopard

Lion

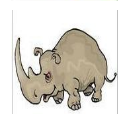
Lizard

Ostrich

Pig

Pigeon

Rabbit

Racoon

Rhinoceros

Snake

Sparrow

Tiger

Tortoise

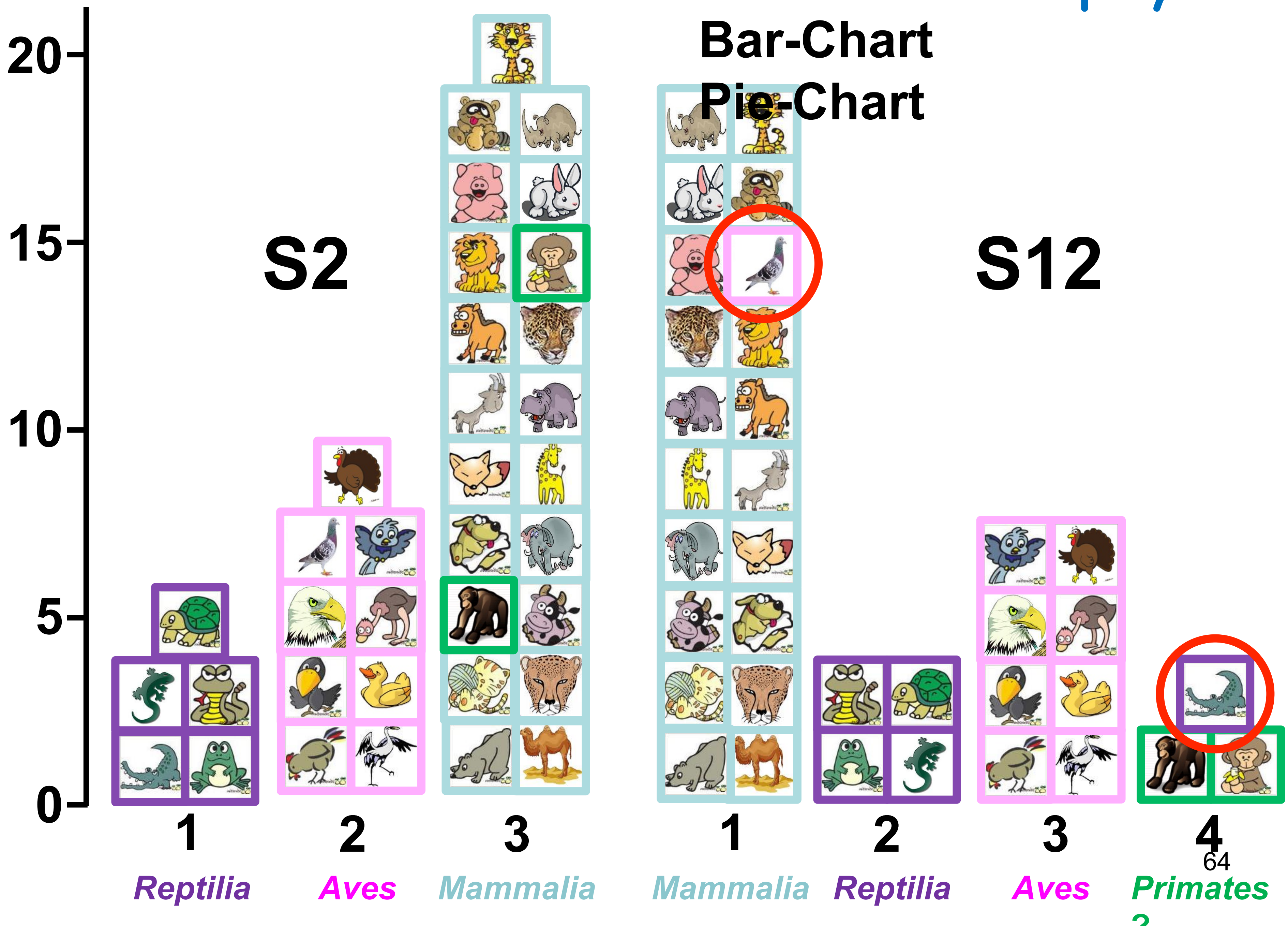
Turkey

Monkey

Uni-variate Display

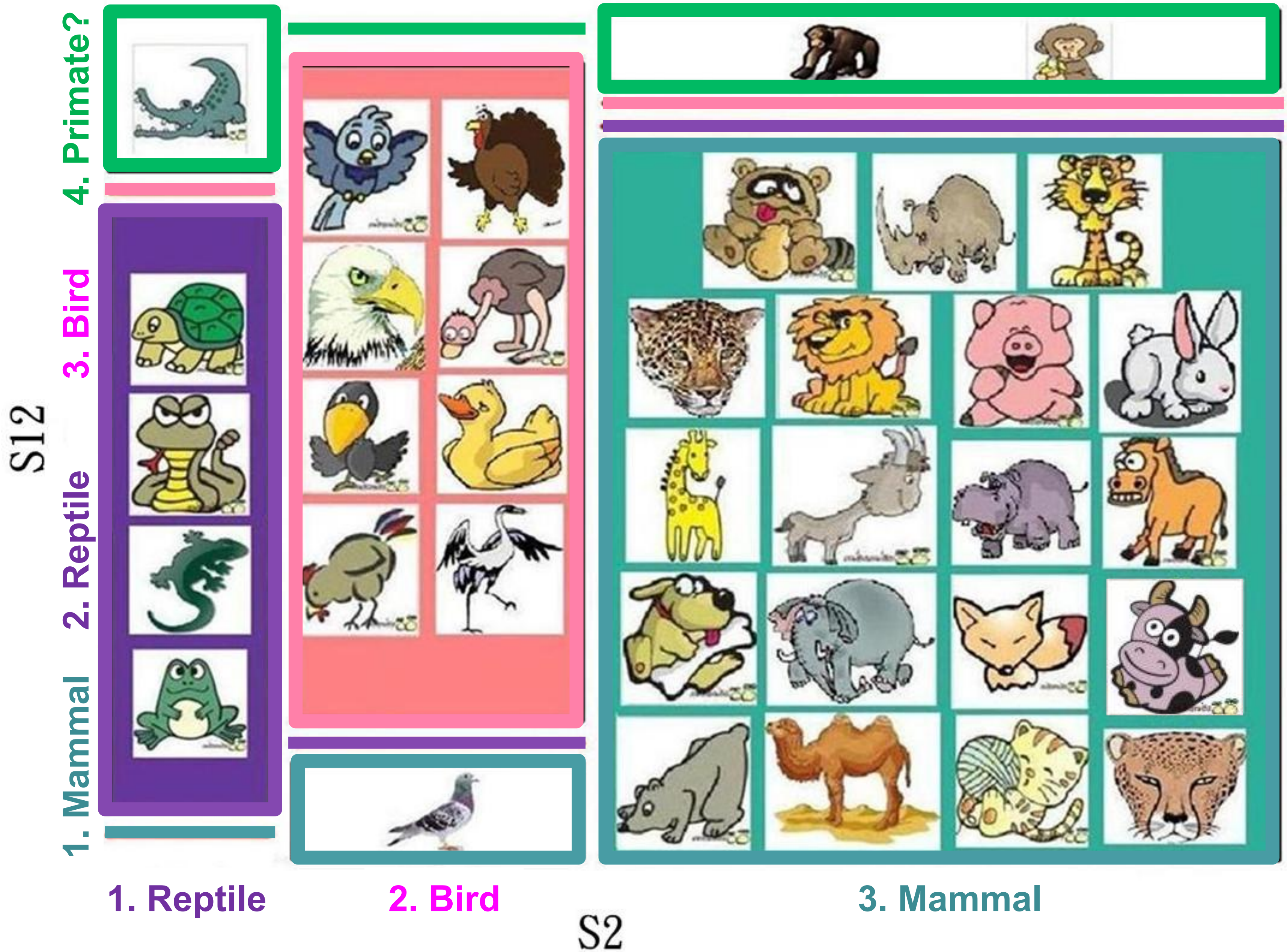
Bar-Chart

Pie-Chart



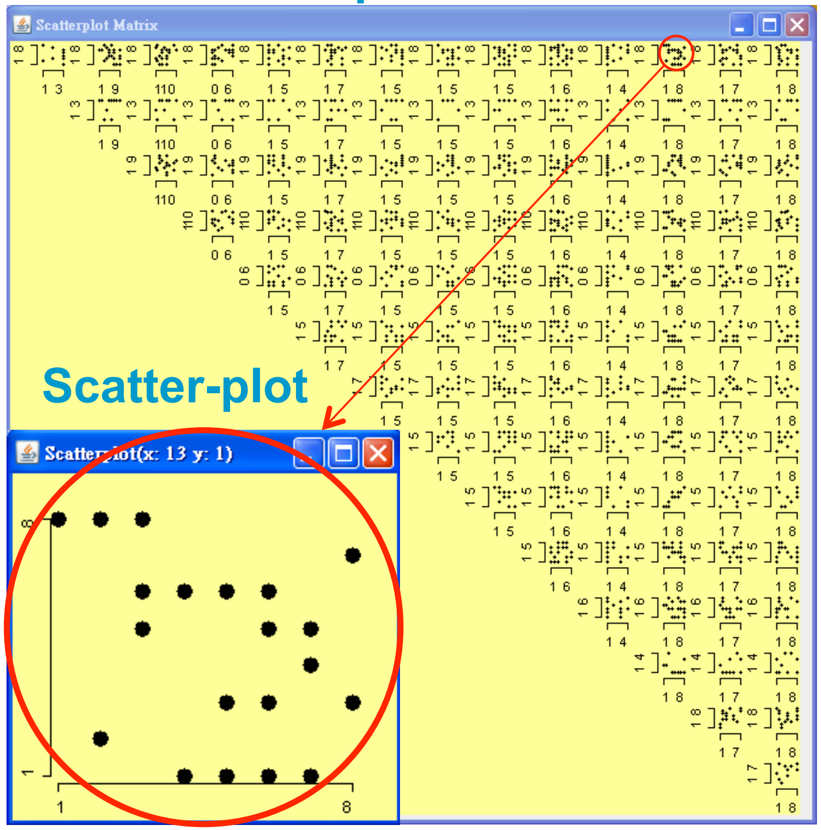
Bi-variate Display

Mosaic Display

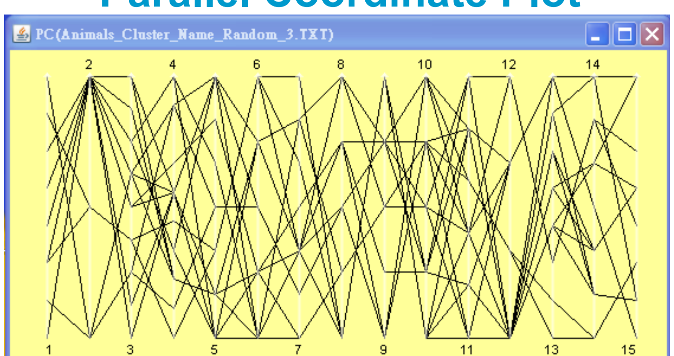


Conventional statistical visualization for this data

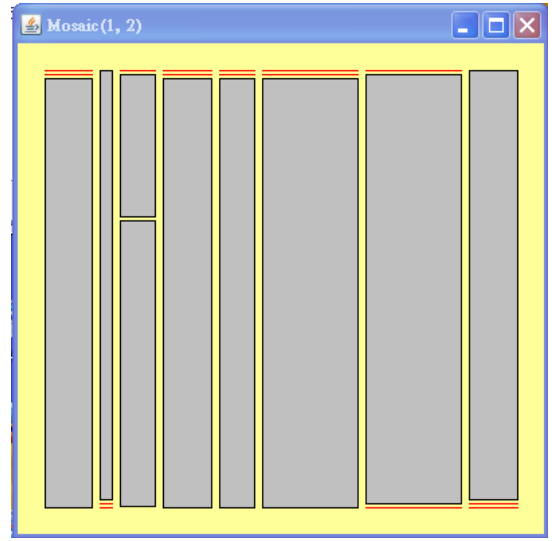
Scatter-plot Matrix



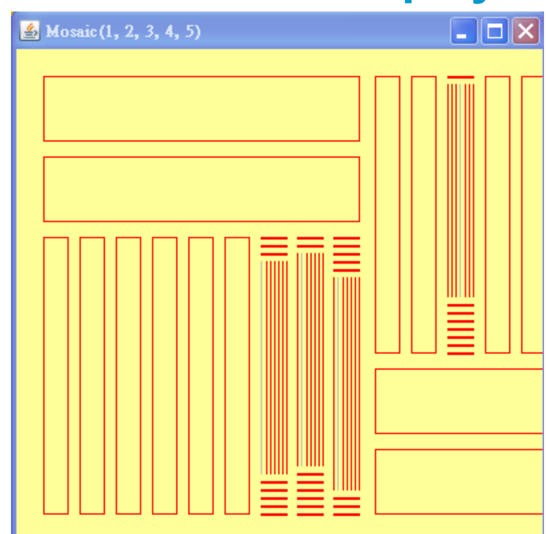
Parallel Coordinate Plot



2D Mosaic Display



5D Mosaic Display



Approaching Statistics & Statistical Approach

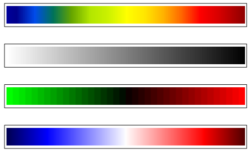
Essential elements in a GAP MV procedure?

Continuous

3. Variable Proximity

Correlation
Covariance
polychoric
Correlation . . .

1. Data Matrix



2. Subject Proximity

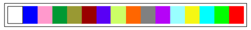
Euclidean Distance
Manhattan Distance
Correlation …

Nominal

3. Variable Proximity

χ^2 type measurements ?

1. Data Matrix



?

2. Subject Proximity

Matching
proportion ?

Is there a natural way of taking care of all 3 problems?

Statistical Approach: **Dual Scaling/Homogeneity Analysis/MCA**

Early Works:

Richardson & Kuder (1933)
Hirschfeld (1935)
Horst (1935)
Edgerton & Kolbe (1936)
Hotelling (1936)
Wilks (1938)
Fisher (1940)
Maung (1941)
Guttman (1941, 1946)
Hayashi (1950, 1952)
Bock (1956, 1960)

Major Groups:

Hayashi school (1950-)
Benzecri school (1960-)
Gifi group (1967-)
de Leeuw & others
Toronto group (1969-)
Nishisato & others

PCA for categorical Data

Aliases:

Method of Reciprocal Average
Simultaneous Linear Regression
Appropriate Scoring, Additive Scoring
Hayashi's Theory of Quantification
Principal Component Analysis of Qualitative Data
Optimal Scaling
Analyse Factrorielle des Correspondances
Homogeneity Analysis
Correspondence Analysis
Correspondence Factor Analysis
Basic Structure Content Scaling
Dual Scaling
Descriptive Multivariate Analysis
Nonlinear Multivariate Analysis

Gifi, A (1990) Nonlinear Multivariate Analysis
Michailidis G, and de Leeuw, J. (1998), "The Gifi System of Descriptive Multivariate Analysis," *Statistical Science*, **13**, 307-336.
Nishisato, S. (1996), "Gleaning in the field of dual scaling," *Psychometrika*, **61**, 559-599.
Nishisato, S. (2006), Multidimensional Nonlinear Descriptive Analysis

Mammals Dentition Example

The data for this example are taken from Hartigan (1975) (also discussed in Michailidis and De Leeuw,1999). Dental characteristics are used in the classification of 66 different kinds of mammals. Mammals' teeth are divided into four groups: incisors, canines, premolars, and molars.

Description for Variables

TI: Top incisors;
  1: 0 incisors, 2: 1 incisors,
  3: 2 incisors, 4: 3 or more incisors
BI: Bottom incisors;
  1: 0 incisors, 2: 1 incisors,
  3: 2 incisors, 4: 3 incisors
  5: 4 incisors
TC: Top canine;
  1: 0 canines, 2: 1 canines,
BC: Bottom canine;
  1: 0 canines, 2: 1 canines,
TP: Top premolar;
  1: 0 premolars, 2: 1 premolars,
  3: 2 premolars, 4: 3 premolars
  5: 4 premolars
BP: Bottom premolar;
  1: 0 premolars, 2: 1 premolars,
  3: 2 premolars, 4: 3 premolars
  5: 4 premolars
TM: Top molar;
  1: 0-2 molars, 2: 3 or more molars,
BM: Bottom molar;
  1: 0-2 molars, 2: 3 or more molars

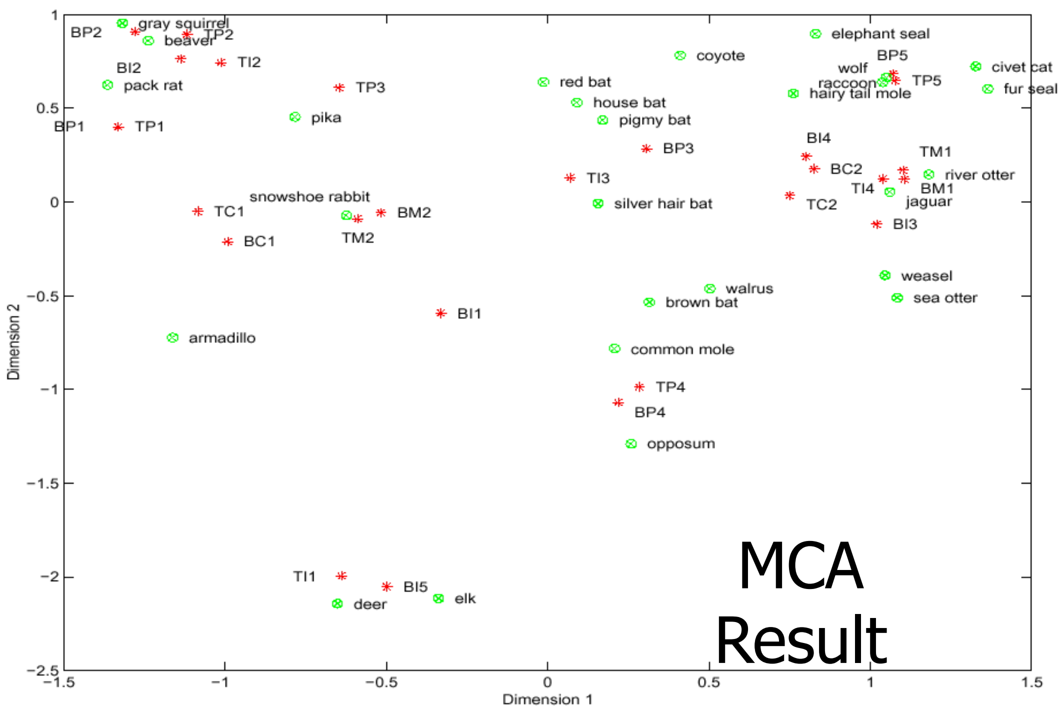
TI	BI	TC	BC	TP	BP	TM	BM	
4	5	2	2	4	4	2	2	Opposum
4	4	2	2	5	5	2	2	Hairy-Tail-Mole
4	3	2	1	4	4	2	2	Common-Mole
4	4	2	2	5	5	2	2	Star-Nose-Mole
3	4	2	2	4	4	2	2	Brown-Bat
3	4	2	2	3	4	2	2	Silver-Hair-Bat
3	4	2	2	3	3	2	2	Pigmy-Bat
3	4	2	2	2	3	2	2	House-Bat
2	4	2	2	3	3	2	2	Red-Bat
2	4	2	2	3	3	2	2	Hoary-Bat
3	4	2	2	3	4	2	2	Lump-Nose-Bat
1	1	1	1	1	1	2	2	Armadillo
3	2	1	1	3	3	2	2	Pika
3	2	1	1	4	3	2	2	Snowshoe-Rabit
2	2	1	1	3	2	2	2	Beaver
2	2	1	1	3	2	2	2	Marmot
2	2	1	1	3	2	2	2	Groundhog
2	2	1	1	3	2	2	2	Prairie-Dog
2	2	1	1	3	2	2	2	Ground-Squirrel
2	2	1	1	3	2	2	2	Chipmunk
2	2	1	1	2	2	2	2	Gray-Squirrel
2	2	1	1	2	2	2	2	Fox-Squirrel
2	2	1	1	2	2	2	2	Pocket-Gopher
2	2	1	1	2	2	2	2	Kangaroo-Rat
2	2	1	1	1	1	2	2	Pack-Rat
2	2	1	1	1	1	2	2	Field-Mouse
2	2	1	1	1	1	2	2	Muskrat
2	2	1	1	1	1	2	2	Black-Rat
2	2	1	1	1	1	2	2	House-Mouse
2	2	1	1	2	2	2	2	Porcupine
2	2	1	1	2	2	2	2	Guinea-Pig
2	4	2	2	5	5	2	2	Coyote
4	4	2	2	5	5	1	2	Wolf
4	4	2	2	5	5	1	2	Fox
4	4	2	2	5	5	1	2	Bear
4	4	2	2	5	5	1	1	Civet-Cat
4	4	2	2	5	5	2	1	Raccoon
4	4	2	2	5	5	1	1	Marten
4	4	2	2	5	5	1	1	Fisher
4	4	2	2	4	4	1	1	Weasel
4	4	2	2	4	4	1	1	Mink
4	4	2	2	4	4	1	1	Ferrer
4	4	2	2	5	5	1	1	Wolverine
4	4	2	2	4	4	1	1	Badger
4	4	2	2	4	4	1	1	Skunk
4	4	2	2	5	4	1	1	River-Otter
4	3	2	2	4	4	1	1	Sea-Otter
4	4	2	2	4	3	1	1	Jaguar
4	4	2	2	4	3	1	1	Ocelot
4	4	2	2	4	3	1	1	Cougar
4	4	2	2	4	3	1	1	Lynx
4	3	2	2	5	5	1	1	Fur-Seal
4	3	2	2	5	5	1	1	Sea-Lion
2	1	2	2	4	4	1	1	Walrus
4	3	2	2	4	4	1	1	Grey-Seal
3	2	2	2	5	5	1	1	Elephant-Seal
3	4	2	2	4	4	2	2	Peccary
1	5	2	1	4	4	2	2	Elk
1	5	1	1	4	4	2	2	Deer
1	5	1	1	4	4	2	2	Moose
1	5	2	1	4	4	2	2	Reindeer
1	5	1	1	4	4	2	2	Antelope
1	5	1	1	4	4	2	2	Bison
1	5	1	1	4	4	2	2	Mountain-Goat
1	5	1	1	4	4	2	2	Muskox
1	5	1	1	4	4	2	2	Mountain-Sheep

Multiple Correspondence Analysis (MCA)

TI	BI	TC	BC	TP	BP	TM	BM	
4	5	2	2	4	4	2	2	Opposum
4	4	2	2	5	5	2	2	Hairy-Tail-Mole
4	3	2	1	4	4	2	2	Common-Mole
4	4	2	2	5	5	2	2	Star-Nose-Mole
3	4	2	2	4	4	2	2	Brown-Bat
3	4	2	2	3	4	2	2	Silver-Hair-Bat
3	4	2	2	3	3	2	2	Pigmy-Bat
3	4	2	2	2	3	2	2	House-Bat
2	4	2	2	3	3	2	2	Red-Bat
2	4	2	2	3	3	2	2	Hoary-Bat
3	4	2	2	3	4	2	2	Lump-Nose-Bat
1	1	1	1	1	1	2	2	Armadillo
3	2	1	1	3	3	2	2	Pika
3	2	1	1	4	3	2	2	Snowshoe-Rabit
2	2	1	1	3	2	2	2	Beaver
2	2	1	1	3	2	2	2	Marmot
2	2	1	1	3	2	2	2	Groundhog
2	2	1	1	3	2	2	2	Prairie-Dog
2	2	1	1	3	2	2	2	Ground-Squirrel
2	2	1	1	3	2	2	2	Chipmunk
2	2	1	1	2	2	2	2	Gray-Squirrel
2	2	1	1	2	2	2	2	Fox-Squirrel
2	2	1	1	2	2	2	2	Pocket-Gopher
2	2	1	1	2	2	2	2	Kangaroo-Rat
2	2	1	1	1	1	2	2	Pack-Rat
2	2	1	1	1	1	2	2	Field-Mouse
2	2	1	1	1	1	2	2	Muskrat
2	2	1	1	1	1	2	2	Black-Rat
2	2	1	1	1	1	2	2	House-Mouse
2	2	1	1	2	2	2	2	Porcupine
2	2	1	1	2	2	2	2	Guinea-Pig
2	4	2	2	5	5	2	2	Coyote
4	4	2	2	5	5	1	2	Wolf
4	4	2	2	5	5	1	2	Fox
4	4	2	2	5	5	1	2	Bear
4	4	2	2	5	5	1	1	Civet-Cat
4	4	2	2	5	5	2	1	Raccoon
4	4	2	2	5	5	1	1	Marten
4	4	2	2	5	5	1	1	Fisher
4	4	2	2	4	4	1	1	Weasel
4	4	2	2	4	4	1	1	Mink
4	4	2	2	4	4	1	1	Ferrer
4	4	2	2	5	5	1	1	Wolverine
4	4	2	2	4	4	1	1	Badger
4	4	2	2	4	4	1	1	Skunk
4	4	2	2	5	4	1	1	River-Otter
4	3	2	2	4	4	1	1	Sea-Otter
4	4	2	2	4	3	1	1	Jaguar
4	4	2	2	4	3	1	1	Ocelot
4	4	2	2	4	3	1	1	Cougar
4	4	2	2	4	3	1	1	Lynx
4	3	2	2	5	5	1	1	Fur-Seal
4	3	2	2	5	5	1	1	Sea-Lion
2	1	2	2	4	4	1	1	Walrus
4	3	2	2	4	4	1	1	Grey-Seal
3	2	2	2	5	5	1	1	Elephant-Seal
3	4	2	2	4	4	2	2	Peccary
1	5	2	1	4	4	2	2	Elk
1	5	1	1	4	4	2	2	Deer
1	5	1	1	4	4	2	2	Moose
1	5	2	1	4	4	2	2	Reindeer
1	5	1	1	4	4	2	2	Antelope
1	5	1	1	4	4	2	2	Bison
1	5	1	1	4	4	2	2	Mountain-Goat
1	5	1	1	4	4	2	2	Muskox
1	5	1	1	4	4	2	2	Mountain-Sheep

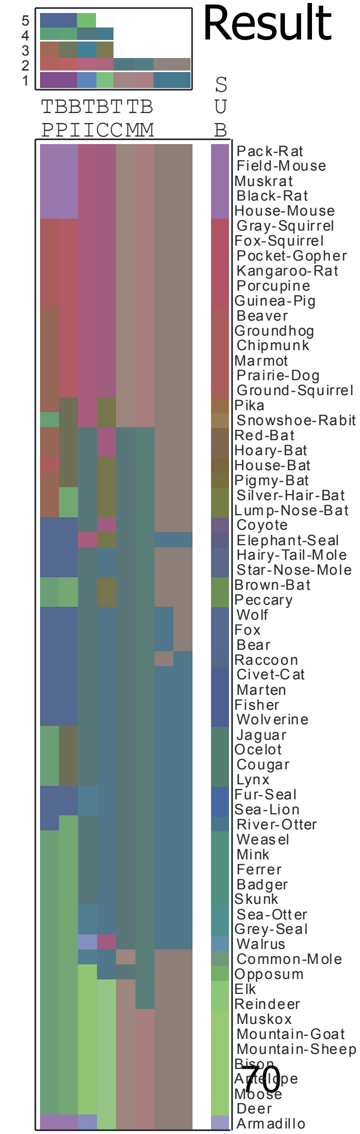
Description for Variables

- TI: Top incisors;
  1: 0 incisors, 2: 1 incisors,
  3: 2 incisors, 4: 3 or more incisors
- BI: Bottom incisors;
  1: 0 incisors, 2: 1 incisors,
  3: 2 incisors, 4: 3 incisors
  5: 4 incisors
- TC: Top canine;
  1: 0 canines, 2: 1 canines,
- BC: Bottom canine;
  1: 0 canines, 2: 1 canines,
- TP: Top premolar;
  1: 0 premolars, 2: 1 premolars,
  3: 2 premolars, 4: 3 premolars
  5: 4 premolars
- BP: Bottom premolar;
  1: 0 premolars, 2: 1 premolars,
  3: 2 premolars, 4: 3 premolars
  5: 4 premolars
- TM: Top molar;
  1: 0-2 molars, 2: 3 or more molars,
- BM: Bottom molar;
  1: 0-2 molars, 2: 3 or more molars



MCA Result

CateGAP Result



UCI Machine Learning Repository

Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one.

Mushroom Data Set

Download: Data Folder, Data Set Description

Abstract: From Audobon Society Field Guide; mushrooms described in terms of physical characteristics; classification: poisonous or edible

Data Set Characteristics:	Multivariate	**Number of Instances:**	8124	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	22	**Date Donated**	1987 04/27
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	48017

Origin:
Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf
Donor: Jeff Schlimmer (Jeffrey.Schlimmer '@' a.gp.cs.cmu.edu)



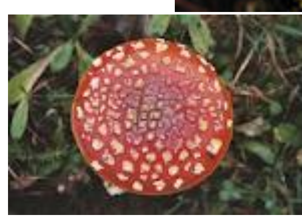




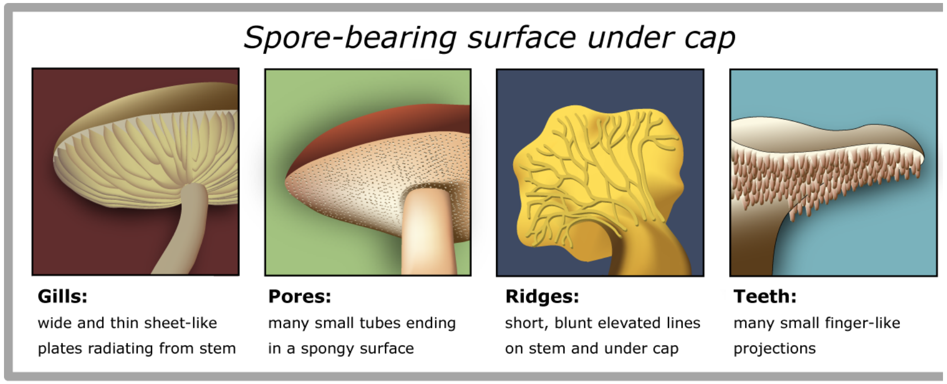


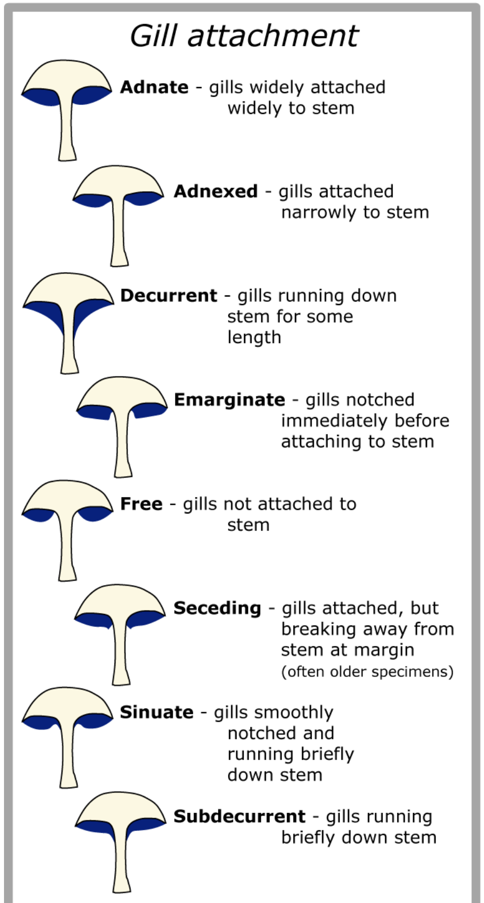

ZAPAŁKI

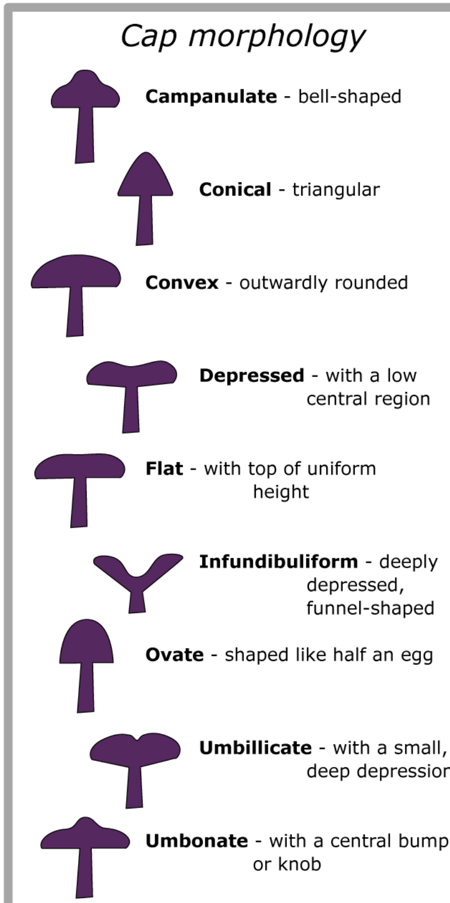


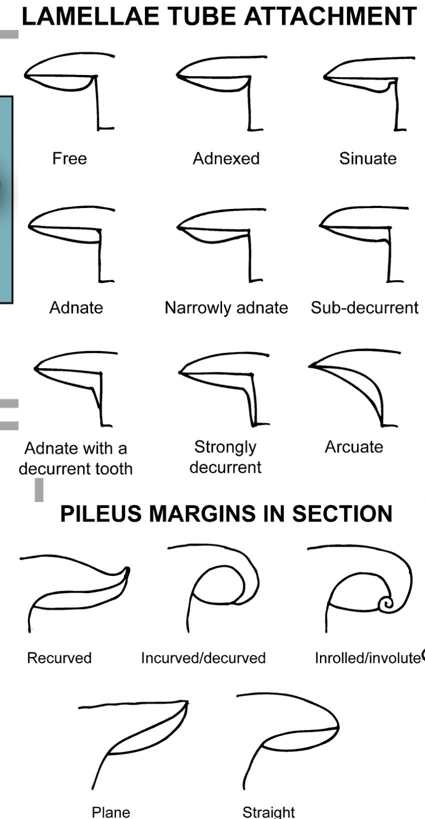



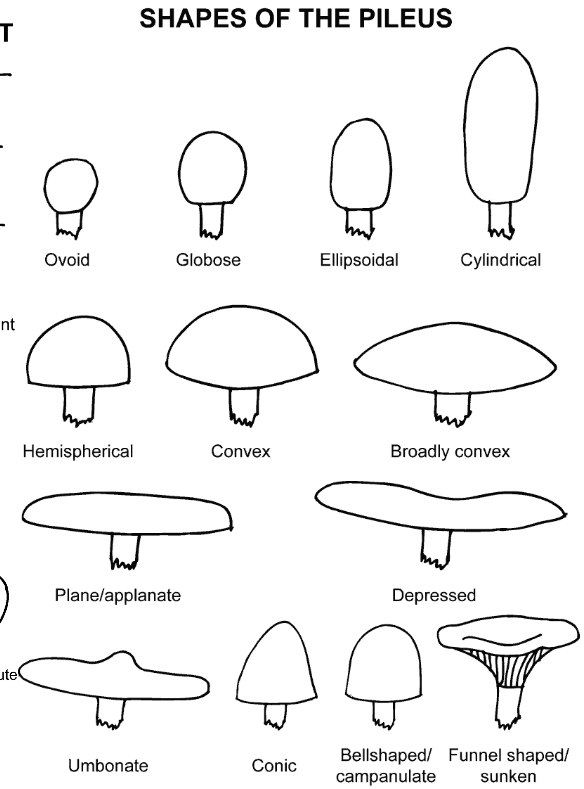


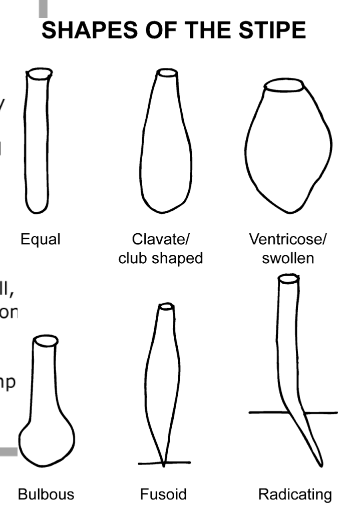

Spore-bearing surface under cap
Gills:
wide and thin sheet-like plates radiating from stem
Pores:
many small tubes ending in a spongy surface
Ridges:
short, blunt elevated lines on stem and under cap
Teeth:
many small finger-like projections

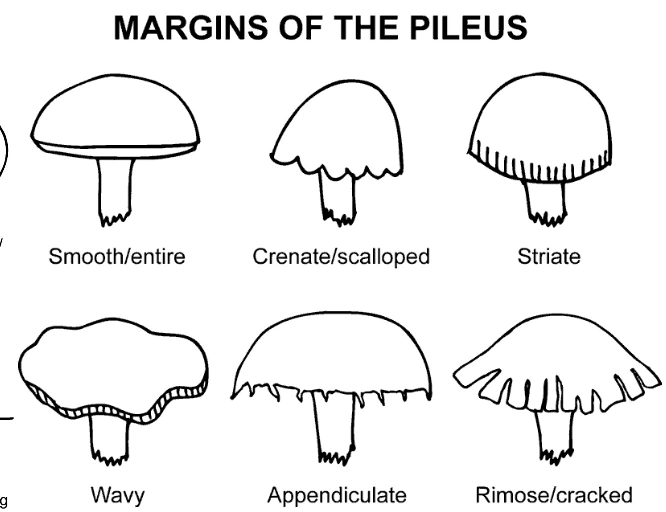

Gill attachment
Adnate - gills widely attached widely to stem
Adnexed - gills attached narrowly to stem
Decurrent - gills running down stem for some length
Emarginate - gills notched immediately before attaching to stem
Free - gills not attached to stem
Seceding - gills attached, but breaking away from stem at margin (often older specimens)
Sinuate - gills smoothly notched and running briefly down stem
Subdecurrent - gills running briefly down stem


Cap morphology
Campanulate - bell-shaped
Conical - triangular
Convex - outwardly rounded
Depressed - with a low central region
Flat - with top of uniform height
Infundibuliform - deeply depressed, funnel-shaped
Ovate - shaped like half an egg
Umbillicate - with a small, deep depression
Umbonate - with a central bump or knob


LAMELLAE TUBE ATTACHMENT
Free
Adnexed
Sinuate
Adnate
Narrowly adnate
Sub-decurrent
Adnate with a decurrent tooth
Strongly decurrent
Arcuate
PILEUS MARGINS IN SECTION
Recurved
Incurved/decurved
Inrolled/involute
Plane
Straight


SHAPES OF THE PILEUS
Ovoid
Globose
Ellipsoidal
Cylindrical
Hemispherical
Convex
Broadly convex
Plane/applanate
Depressed
Umbonate
Conic
Bellshaped/ campanulate
Funnel shaped/ sunken


SHAPES OF THE STIPE
Equal
Clavate/ club shaped
Ventricose/ swollen
Bulbous
Fusoid
Radicating


MARGINS OF THE PILEUS
Smooth/entire
Crenate/scalloped
Striate
Wavy
Appendiculate
Rimose/cracked

Attribute Information:

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s
3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,
   pink=p,purple=u,red=e,white=w,yellow=y
4. bruises?: bruises=t,no=f
5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
6. gill-attachment: attached=a,descending=d,free=f,notched=n
7. gill-spacing: close=c,crowded=w,distant=d
8. gill-size: broad=b,narrow=n
9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g,
   green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
10. stalk-shape: enlarging=e,tapering=t
11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?
12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s
13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s
14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,
    pink=p,red=e,white=w,yellow=y
15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,
    pink=p,red=e,white=w,yellow=y
16. veil-type: partial=p,universal=u
17. veil-color: brown=n,orange=o,white=w,yellow=y
18. ring-number: none=n,one=o,two=t
19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,
    none=n,pendant=p,sheathing=s,zone=z
20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,
    orange=o,purple=u,white=w,yellow=y
21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Approaching Statistics & Statistical Approach

Matrix Visualization with cartography links

160 international organization

Data:

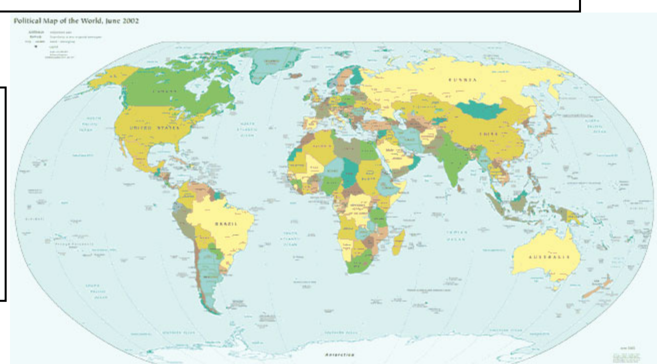
160 international organization membership pattern (variables) for 230 countries/regions (subjects)

0. non-member □ 1. member ■

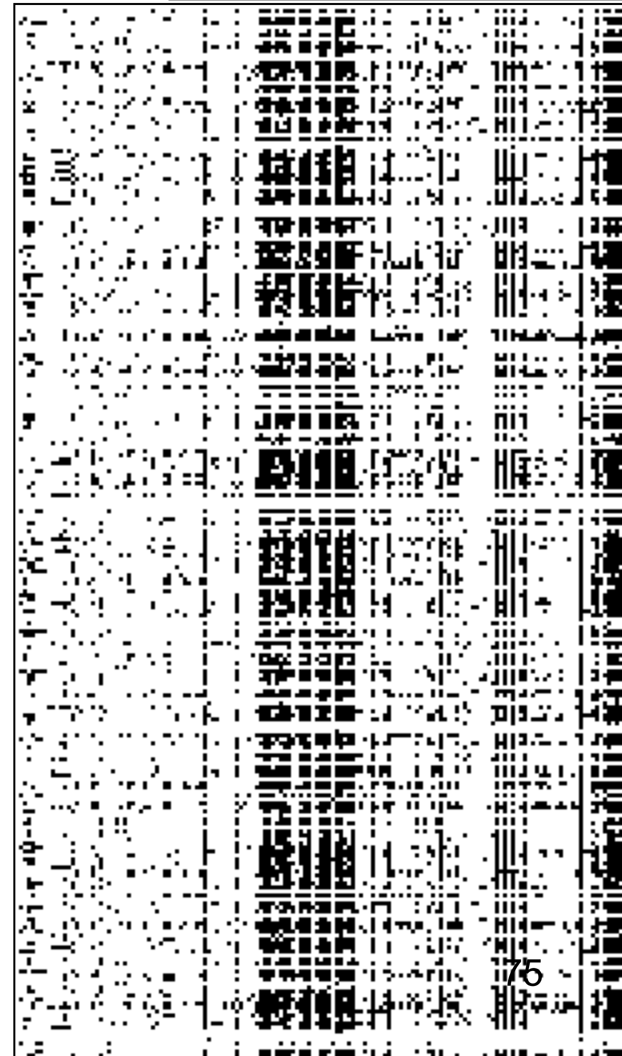
2. observer 3. associate member

4. guest 5. dialogue partner

230 countries (regions)

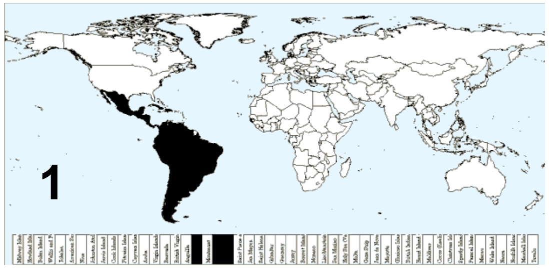


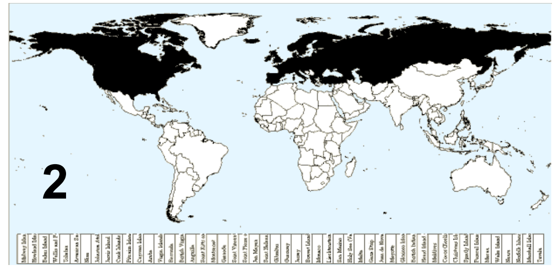
CIA Political Map of the World

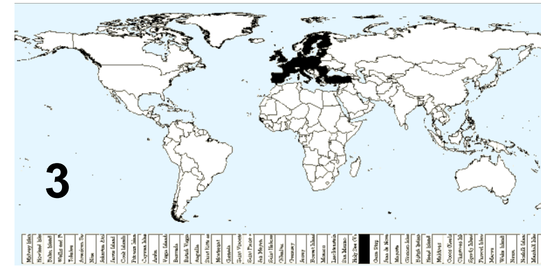


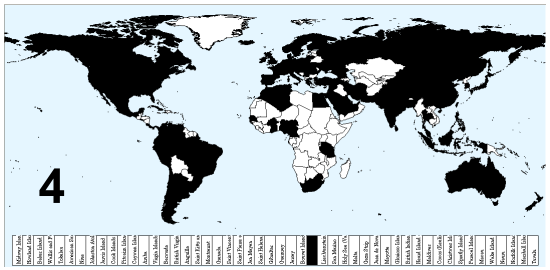
http://www.faqs.org/docs/factbook/index.html

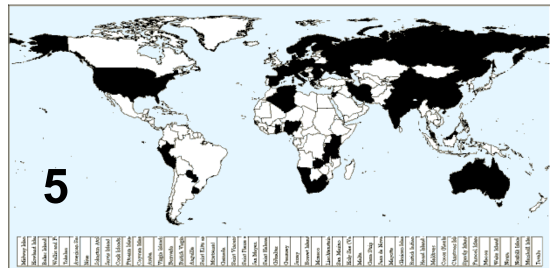
Draw one membership map for each organization (variable)?

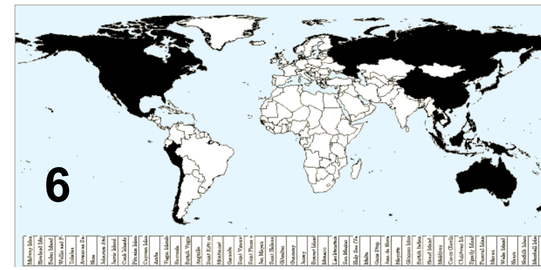


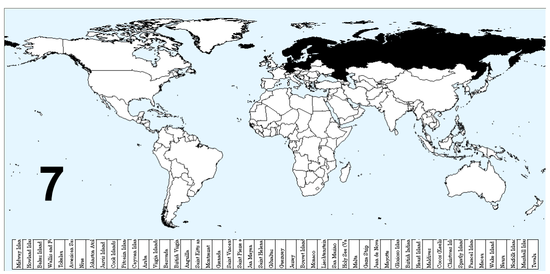


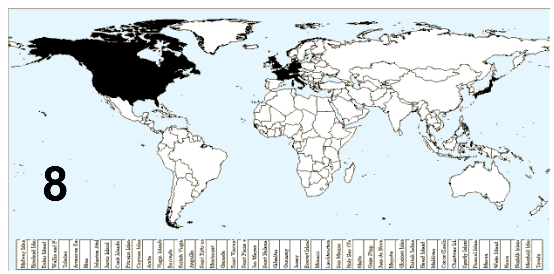


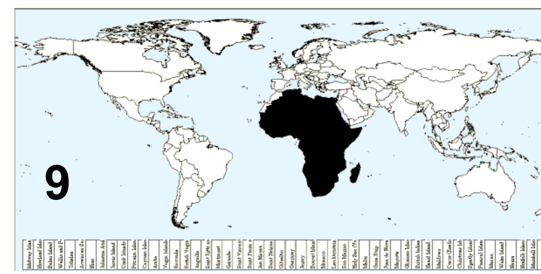




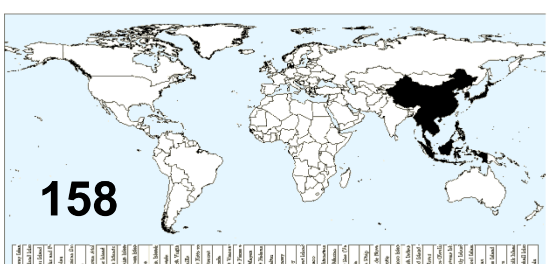


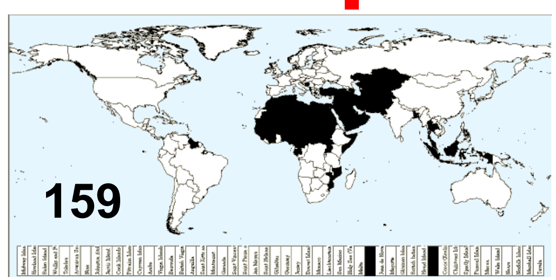


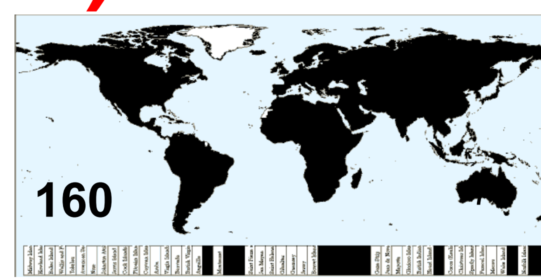




· · · 160 maps (?) · · ·

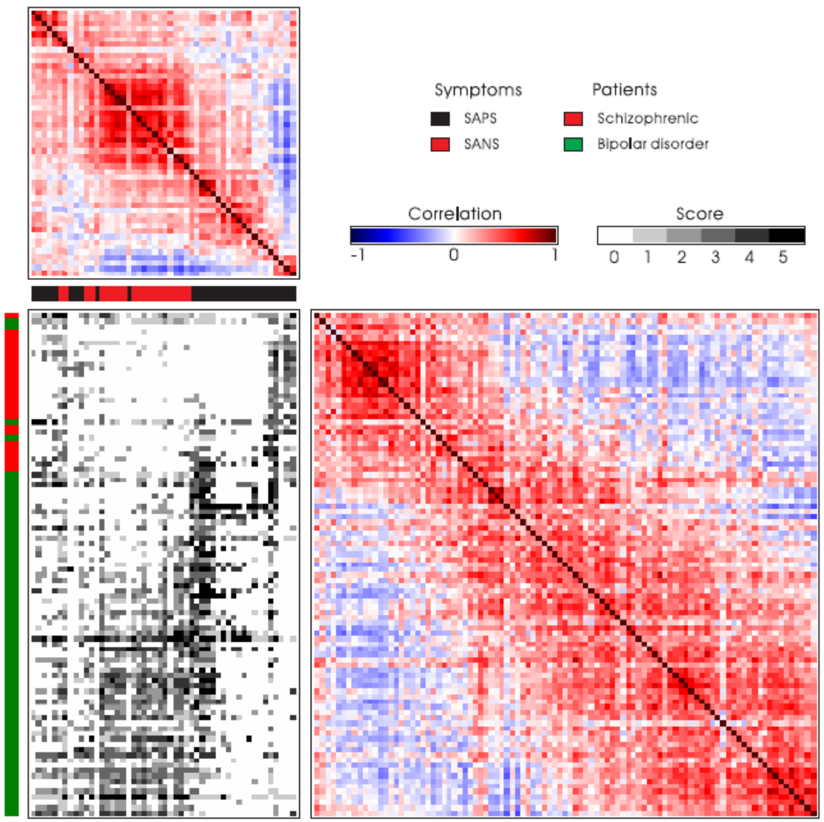




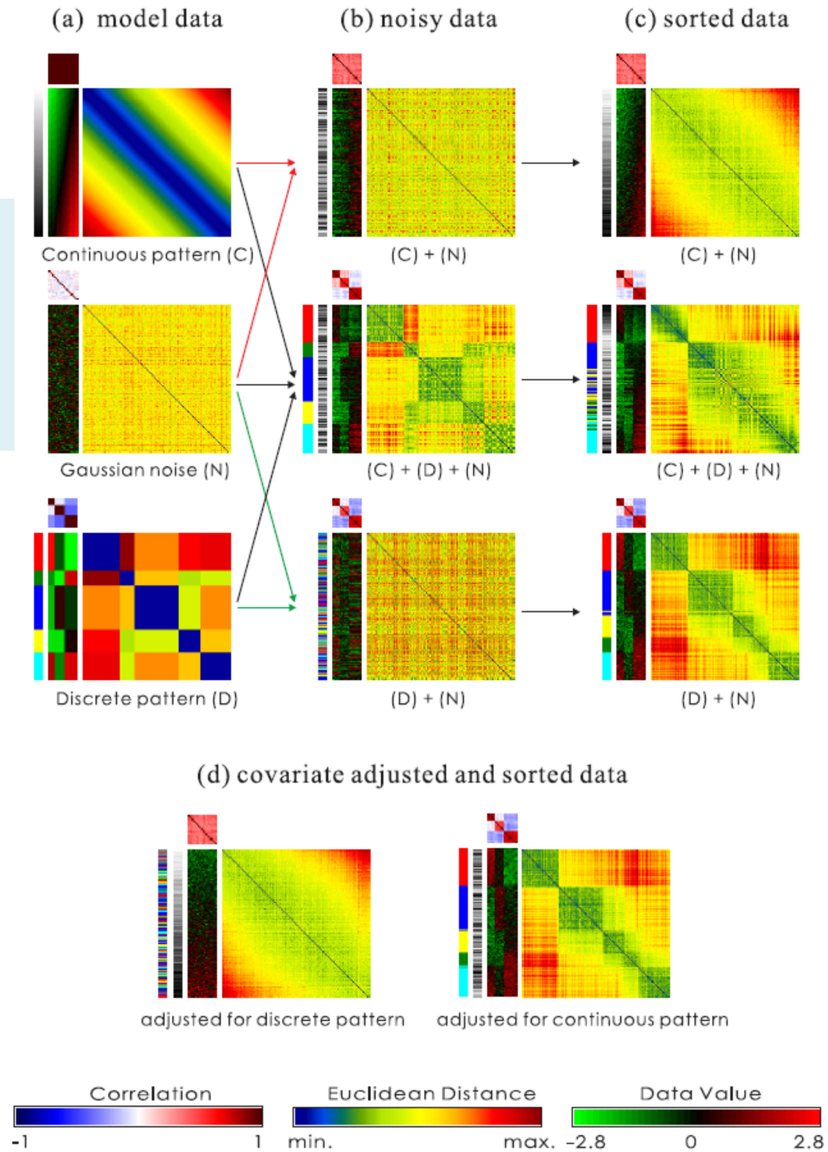


Approaching Statistics & Statistical Approach

Covariate-adjusted Matrix Visualization

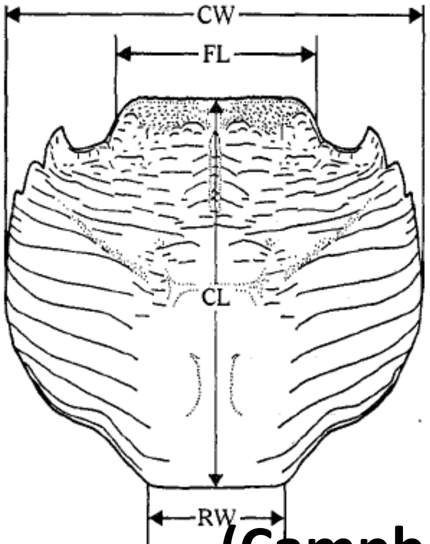


Psychosis disorder data with covariates



A simulation data;
(a) the model data sets, (b) noisy data sets,
(c) sorted data sets and
(d) covariate adjusted and sorted data sets

Morphological Measurements on Leptograpsus Crabs



The crabs data (Campbell and Mahon 1974) in MASS package in R has **200 crabs** with **5** morphological **measurements** (frontal lobe size (FL), rear width (RW), carapace length (CL), carapace width (CW), and body depth (BD)) on **50 crabs each** of **two species** (**blue** (**B**) and **orange** (**O**)) and both **sexes**.

sp	sex	FL	RW	CL	CW	BD
B	M	8.1	6.7	16.1	19	7
B	M	8.8	7.7	18.1	20.8	7.4
B	M	9.2	7.8	19	22.4	7.7
B	M	9.6	7.9	20.1	23.1	8.2
B	M	9.8	8	20.3	23	8.2
			⋮			
B	F	7.2	6.5	14.7	17.1	6.1
B	F	9	8.5	19.3	22.7	7.7
B	F	9.1	8.1	18.5	21.6	7.7
B	F	9.1	8.2	19.2	22.2	7.7
B	F	9.5	8.2	19.6	22.4	7.8
			⋮			
O	M	9.1	6.9	16.7	18.6	7.4
O	M	10.2	8.2	20.2	22.2	9
O	M	10.7	8.6	20.7	22.7	9.2
O	M	11.4	9	22.7	24.8	10.1
O	M	12.5	9.4	23.2	26	10.8
			⋮			
O	F	11.4	9.2	21.7	24.1	9.7
O	F	12.5	10	24.1	27	10.9
O	F	12.6	11.5	25	28.1	11.5
O	F	12.9	11.2	25.8	29.1	11.9
O	F	14	11.9	27	31.4	12.6

Matrix Visualization for MANCOVA modeling

Y. J. Tien, H. M. Wu,
Y. S. Lee, and C. H. Chen (2010)

MANOVA model examples

Model 1: one-factor fixed effect model

$$Y_{(i)jk} = \mu_{(i)} + \rho_{(i)j} + \varepsilon_{(i)jk}$$

$i = 1, 2, \cdots, p$ $j = 1, 2, \cdots, n$ $k = 1, 2, \cdots, n_j$ $E(\varepsilon_{(i)jk}) = 0$ $Var(\varepsilon_{(i)jk}) = \sigma^2$

$$H_0 : \rho_{(i)j} = 0$$

Overall

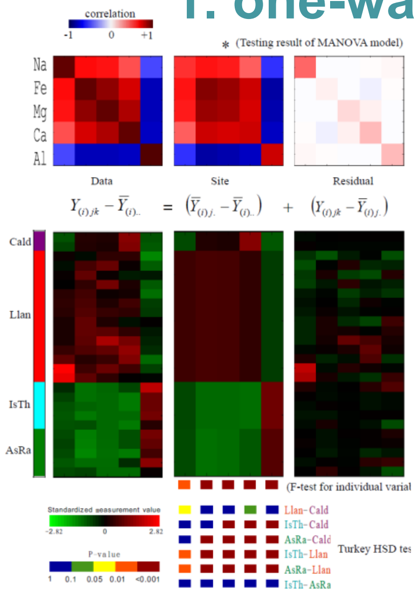
Individual variable

Post-hoc analysis (multiple comparison)

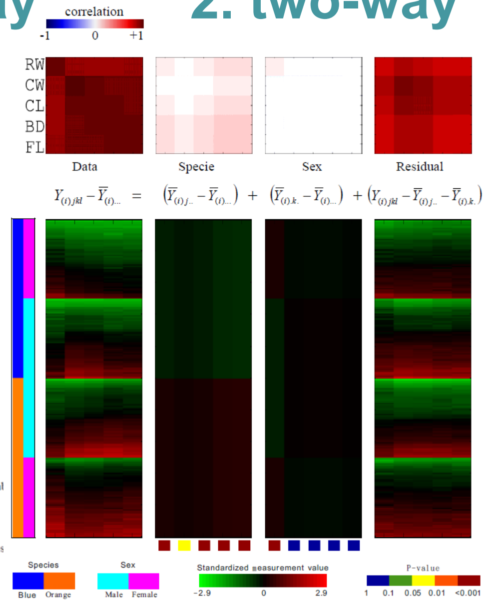
Extensions of MANOVA MV

Approaching Statistics Statistical Approach

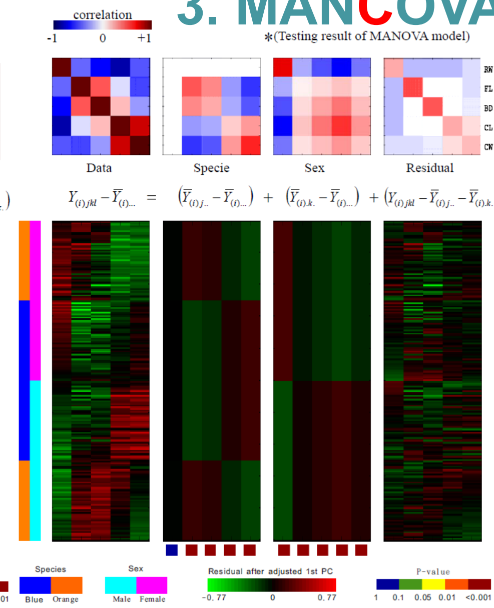
1. one-way



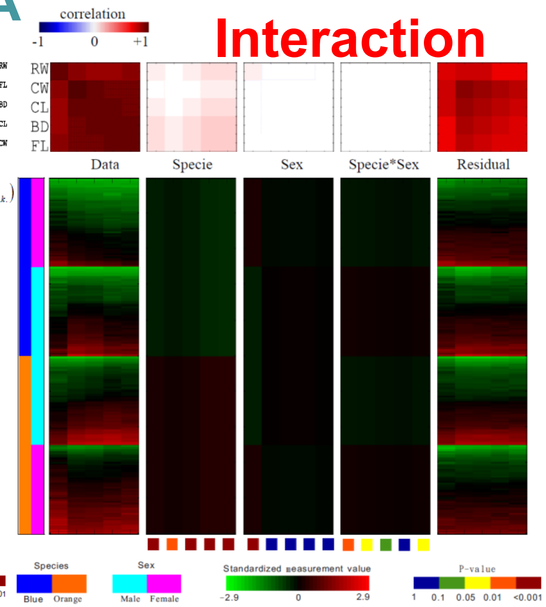
2. two-way



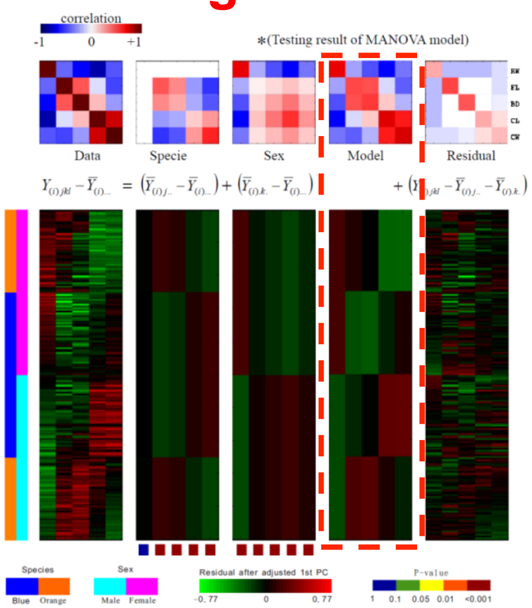
3. MANCOVA



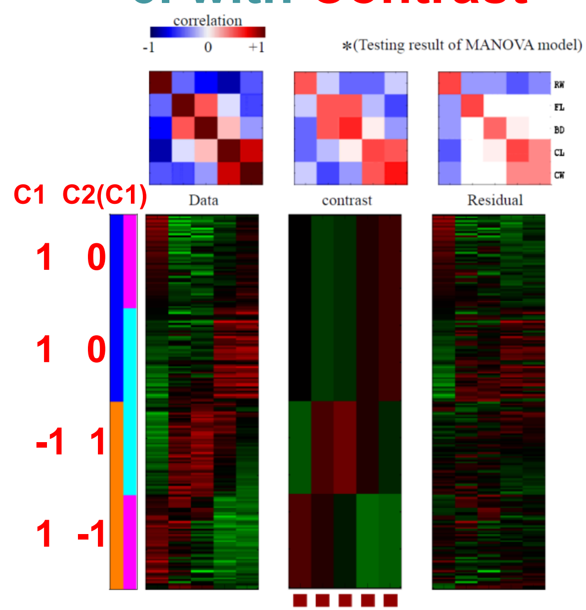
4. with Interaction



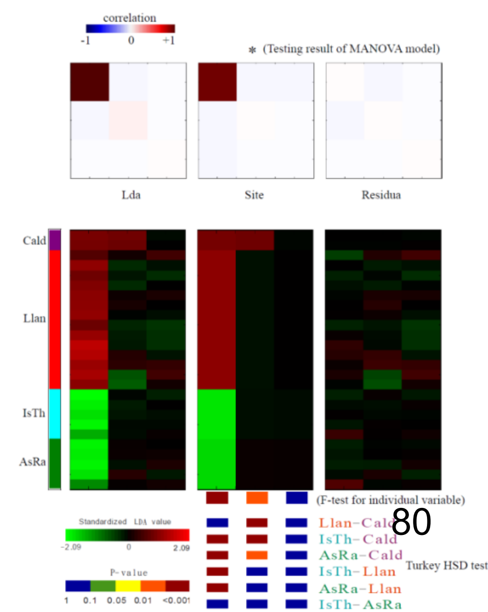
5. with Regression Model



6. with Contrast



7. with Reduced rank



Extension of Matrix Visualization for Symbolic Data (Analysis):

The GAP Approach (with Junji Nakano)

Clustered (non-independent) Data

1. Hierarchical (multi-level) Data

2. Genetic Familial Data

Huge Data Sets

1. Large n

2. Large p

3. Large n & p

Other Types of Symbolic Data

1.1 Symbolic Data Analysis (SDA) and 1.2 Matrix Visualization (MV)

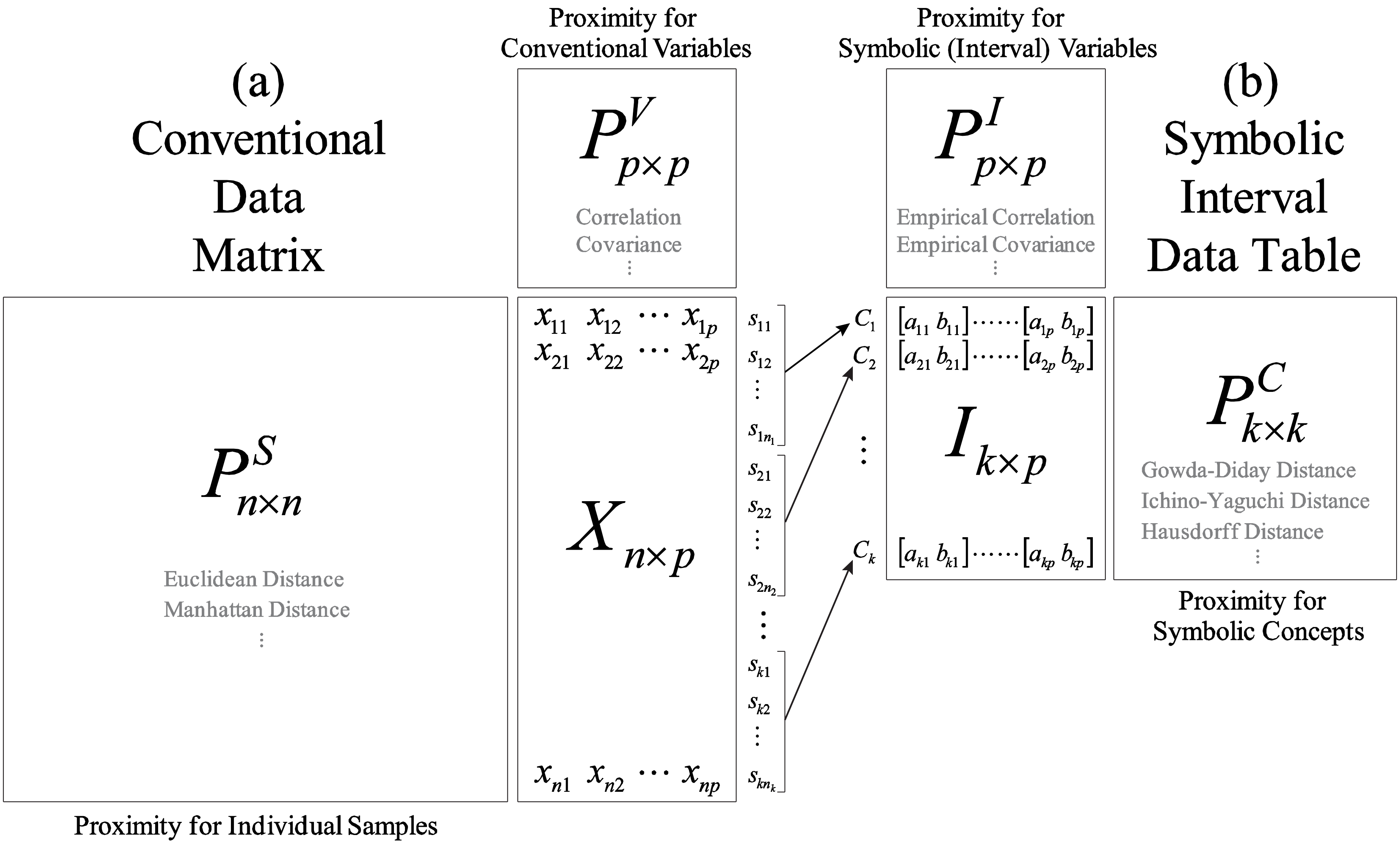


Fig. 1. Diagram for related conventional data matrix and symbolic (interval type) data table with their corresponding proximity matrices for samples/concepts and variables.

2.1 Proximity matrix for interval (range) variables

The empirical covariance function between *I_i and I_j* is given by

$$Cov(I_i, I_j) = \frac{1}{4k}\sum_{c=1}^{k}[(a_{ci} + b_{ci})(a_{cj} + b_{cj)}]$$
$$-\frac{1}{4k^2}[\sum_{c=1}^{k}(a_{ci} + b_{ci})][\sum_{c=1}^{k}(a_{cj} + b_{cj})].$$

The empirical correlation coefficient between *I_i and I_j* is given by

$$r(I_i, I_j) = \frac{Cov(I_i, I_j)}{S_{Z_i}S_{Z_j}},$$

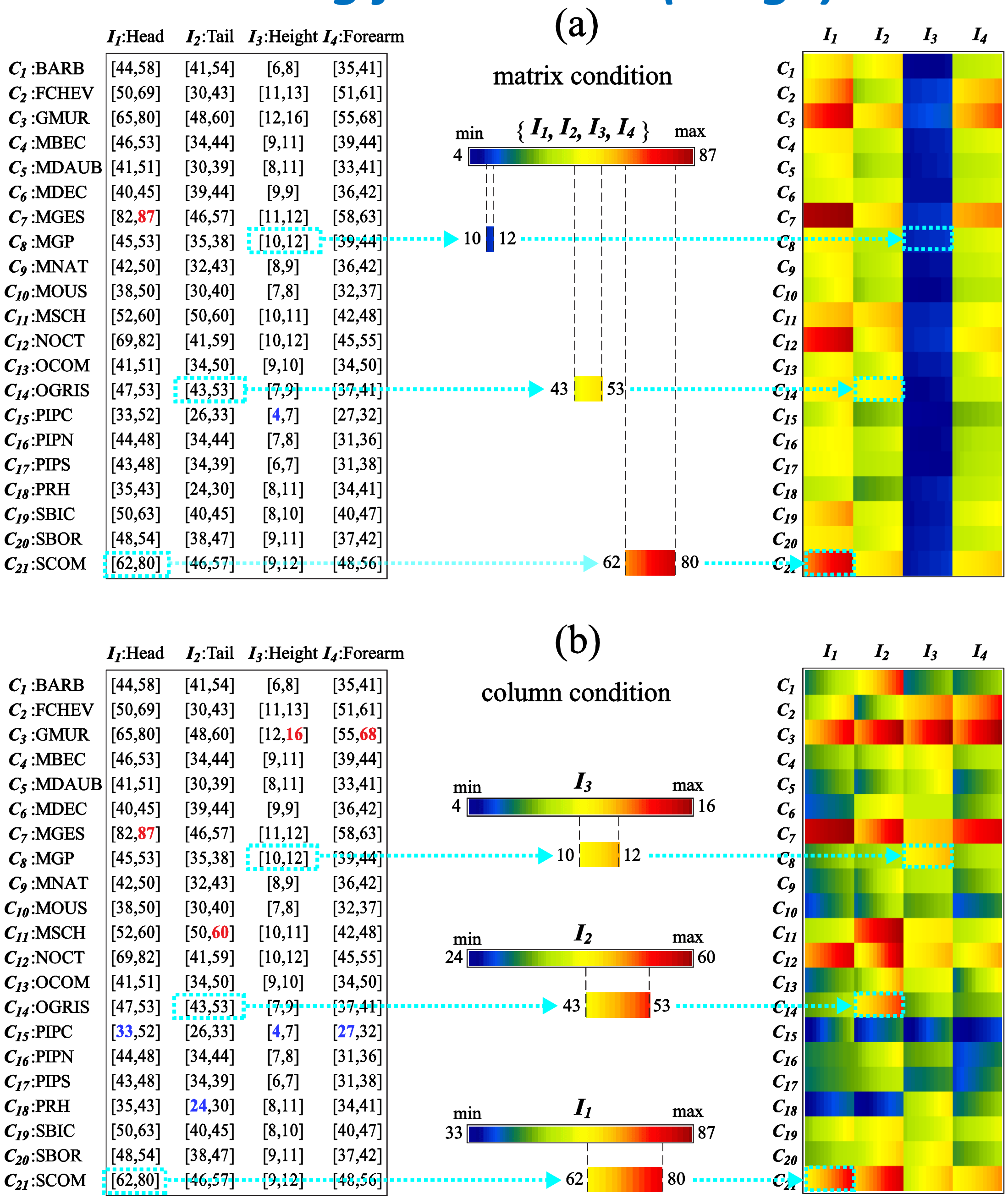
$$S^2_{Z_i} = \frac{1}{3k}\sum_{c=1}^{k}(b^2_{ci} + b_{ci}a_{ci} + a^2_{ci}) - \frac{1}{4k^2}[\sum_{c=1}^{k}(b_{ci} + a_{ci})]^2.$$

2.2 Proximity matrix for concepts with interval variables

Table 1. Distance measures for interval type symbolic data proposed in Billard and Diday (2006).

Measure Name	Formula	Component detail
The Gowda-Diday distance (Gowda and Diday, 1991)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \frac{\lvert a_{ir}-a_{jr}\rvert}{\lvert \max_c b_{cr} - \min_c a_{cr}\rvert} + \frac{\lvert\lvert b_{ir}-a_{ir}\rvert - \lvert b_{jr}-a_{jr}\rvert\rvert}{max(b_{ir},b_{jr})-min(a_{ir},a_{jr})} + \frac{\lvert b_{ir}-a_{ir}\rvert + \lvert b_{jr}-a_{jr}\rvert - 2I_r}{max(b_{ir},b_{jr})-min(a_{ir},a_{jr})}$ where $I_r = \lvert max(a_{ir}, a_{jr}) - min(b_{ir}, b_{jr})\rvert$
The Ichino-Yaguchi distance (Ichino and Yaguchi, 1994)	$\sqrt[q]{\sum_{r=1}^{p} D(I_{ir}, I_{jr})^q}$	$D(I_{ir}, I_{jr}) = \lvert[a_{ir}, b_{ir}] \cup [a_{jr}, b_{jr}]\rvert - \lvert[a_{ir}, b_{ir}] \cap [a_{jr}, b_{jr}]\rvert + \gamma(2\lvert[a_{ir}, b_{ir}] \cap [a_{jr}, b_{jr}]\rvert - \lvert[a_{ir}, b_{ir}]\rvert - \lvert[a_{jr}, b_{jr}]\rvert)$ where $0 \leq \gamma \leq 0.5$
The L_1 distance	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \lvert \frac{a_{ir}+b_{ir}}{2} - \frac{a_{jr}+b_{jr}}{2} \rvert$
The L_2 distance (de Carvalho et al., 2006)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = (\frac{a_{ir}+b_{ir}}{2} - \frac{a_{jr}+b_{jr}}{2})^2$
The City-Block distance (de Souza and de Carvalho, 2004)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = \lvert a_{ir} - a_{jr}\rvert + \lvert b_{ir} - b_{jr}\rvert$
The Hausdorff distance (Chavent and Lechevallier, 2002)	$\sum_{r=1}^{p} D(I_{ir}, I_{jr})$	$D(I_{ir}, I_{jr}) = max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$
The Euclidean Hausdorff distance	$\sqrt[2]{\sum_{r=1}^{p} D(I_{ir}, I_{jr})^2}$	$D(I_{ir}, I_{jr}) = max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$
The normalized Euclidean Hausdorff distance	$\sqrt[2]{\sum_{r=1}^{p} [\frac{D(I_{ir}, I_{jr})}{H_r}]^2}$	$D(I_{ir}, I_{jr}) = max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$ $H_r^2 = \frac{1}{2k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} D(I_{ir}, I_{jr})^2$
The span normalized Euclidean Hausdorff distance	$\sqrt[2]{\sum_{r=1}^{p} [\frac{D(I_{ir}, I_{jr})}{\lvert R_r\rvert}]^2}$	$D(I_{ir}, I_{jr}) = max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$ $\lvert R_r\rvert = \max_c b_{cr} - \min_c a_{cr}$

2.3 Color coding for interval (range) data table



	I_1:Head	I_2:Tail	I_3:Height	I_4:Forearm
C_1 :BARB	[44,58]	[41,54]	[6,8]	[35,41]
C_2 :FCHEV	[50,69]	[30,43]	[11,13]	[51,61]
C_3 :GMUR	[65,80]	[48,60]	[12,16]	[55,68]
C_4 :MBEC	[46,53]	[34,44]	[9,11]	[39,44]
C_5 :MDAUB	[41,51]	[30,39]	[8,11]	[33,41]
C_6 :MDEC	[40,45]	[39,44]	[9,9]	[36,42]
C_7 :MGES	[82,87]	[46,57]	[11,12]	[58,63]
C_8 :MGP	[45,53]	[35,38]	[10,12]	[39,44]
C_9 :MNAT	[42,50]	[32,43]	[8,9]	[36,42]
C_{10}:MOUS	[38,50]	[30,40]	[7,8]	[32,37]
C_{11}:MSCH	[52,60]	[50,60]	[10,11]	[42,48]
C_{12}:NOCT	[69,82]	[41,59]	[10,12]	[45,55]
C_{13}:OCOM	[41,51]	[34,50]	[9,10]	[34,50]
C_{14}:OGRIS	[47,53]	[43,53]	[7,9]	[37,41]
C_{15}:PIPC	[33,52]	[26,33]	[4,7]	[27,32]
C_{16}:PIPN	[44,48]	[34,44]	[7,8]	[31,36]
C_{17}:PIPS	[43,48]	[34,39]	[6,7]	[31,38]
C_{18}:PRH	[35,43]	[24,30]	[8,11]	[34,41]
C_{19}:SBIC	[50,63]	[40,45]	[8,10]	[40,47]
C_{20}:SBOR	[48,54]	[38,47]	[9,11]	[37,42]
C_{21}:SCOM	[62,80]	[46,57]	[9,12]	[48,56]

	I_1:Head	I_2:Tail	I_3:Height	I_4:Forearm
C_1 :BARB	[44,58]	[41,54]	[6,8]	[35,41]
C_2 :FCHEV	[50,69]	[30,43]	[11,13]	[51,61]
C_3 :GMUR	[65,80]	[48,60]	[12,16]	[55,68]
C_4 :MBEC	[46,53]	[34,44]	[9,11]	[39,44]
C_5 :MDAUB	[41,51]	[30,39]	[8,11]	[33,41]
C_6 :MDEC	[40,45]	[39,44]	[9,9]	[36,42]
C_7 :MGES	[82,87]	[46,57]	[11,12]	[58,63]
C_8 :MGP	[45,53]	[35,38]	[10,12]	[39,44]
C_9 :MNAT	[42,50]	[32,43]	[8,9]	[36,42]
C_{10}:MOUS	[38,50]	[30,40]	[7,8]	[32,37]
C_{11}:MSCH	[52,60]	[50,60]	[10,11]	[42,48]
C_{12}:NOCT	[69,82]	[41,59]	[10,12]	[45,55]
C_{13}:OCOM	[41,51]	[34,50]	[9,10]	[34,50]
C_{14}:OGRIS	[47,53]	[43,53]	[7,9]	[37,41]
C_{15}:PIPC	[33,52]	[26,33]	[4,7]	[27,32]
C_{16}:PIPN	[44,48]	[34,44]	[7,8]	[31,36]
C_{17}:PIPS	[43,48]	[34,39]	[6,7]	[31,38]
C_{18}:PRH	[35,43]	[24,30]	[8,11]	[34,41]
C_{19}:SBIC	[50,63]	[40,45]	[8,10]	[40,47]
C_{20}:SBOR	[48,54]	[38,47]	[9,11]	[37,42]
C_{21}:SCOM	[62,80]	[46,57]	[9,12]	[48,56]

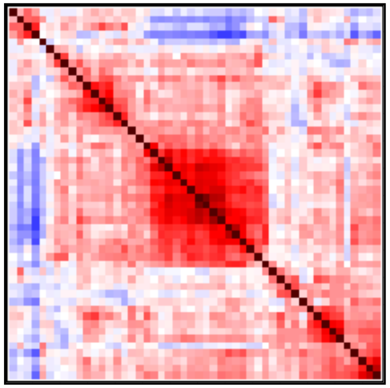
Fig. 2. Color-coding scheme for interval-valued SDA data using the Bats example.
(a) matrix condition;
(b) column condition.

Ongoing / Future Directions for GAP_MV:

1. MV for diagnosing **proximity** matrix modeling
2. MV with **covariate** adjustment
3. MV with dependent (**clustered**) structure
4. MV for data with **missing** values
5. MV with **nonlinear** proximity measurement
6. MV for **longitudinal** multivariate data
7. MV for **multi-conditioned** multivariate data
8. MV with **dependent** variable
9. MV for **mixed** data
10. MV for **huge** data set
11. MV for **time** series data
12. MV for color-**blind** people

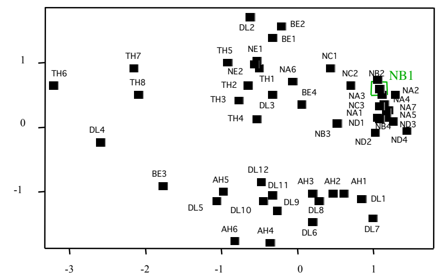
Matrix Visualization
Statistical Modeling of Proximity Matrix

Input
Proximity Matrix

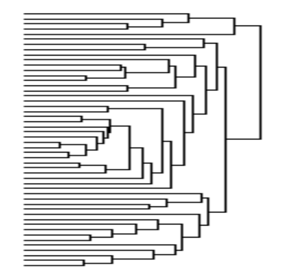


Statistical Models:

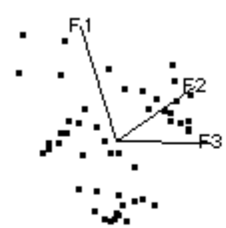
1. Multidimensional Scaling (MDS)



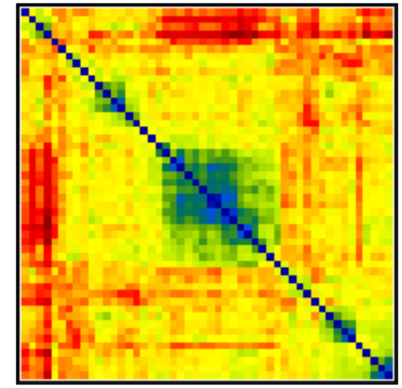
2. Hierarchical Clustering Tree (HCT)

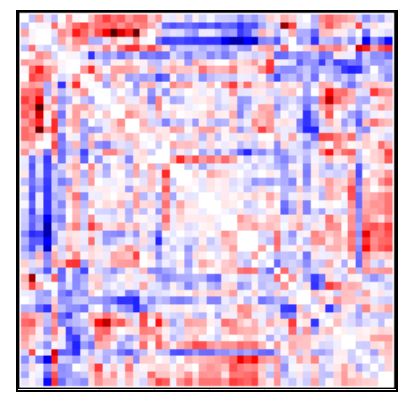


3. Factor Analysis (FA)

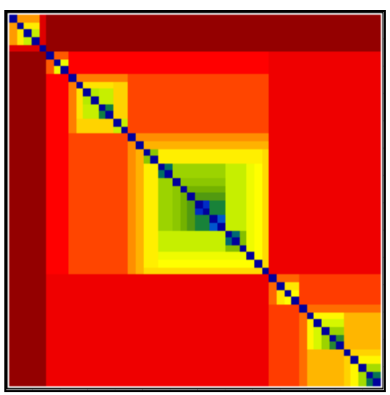


Transformed
Disparity Matrix

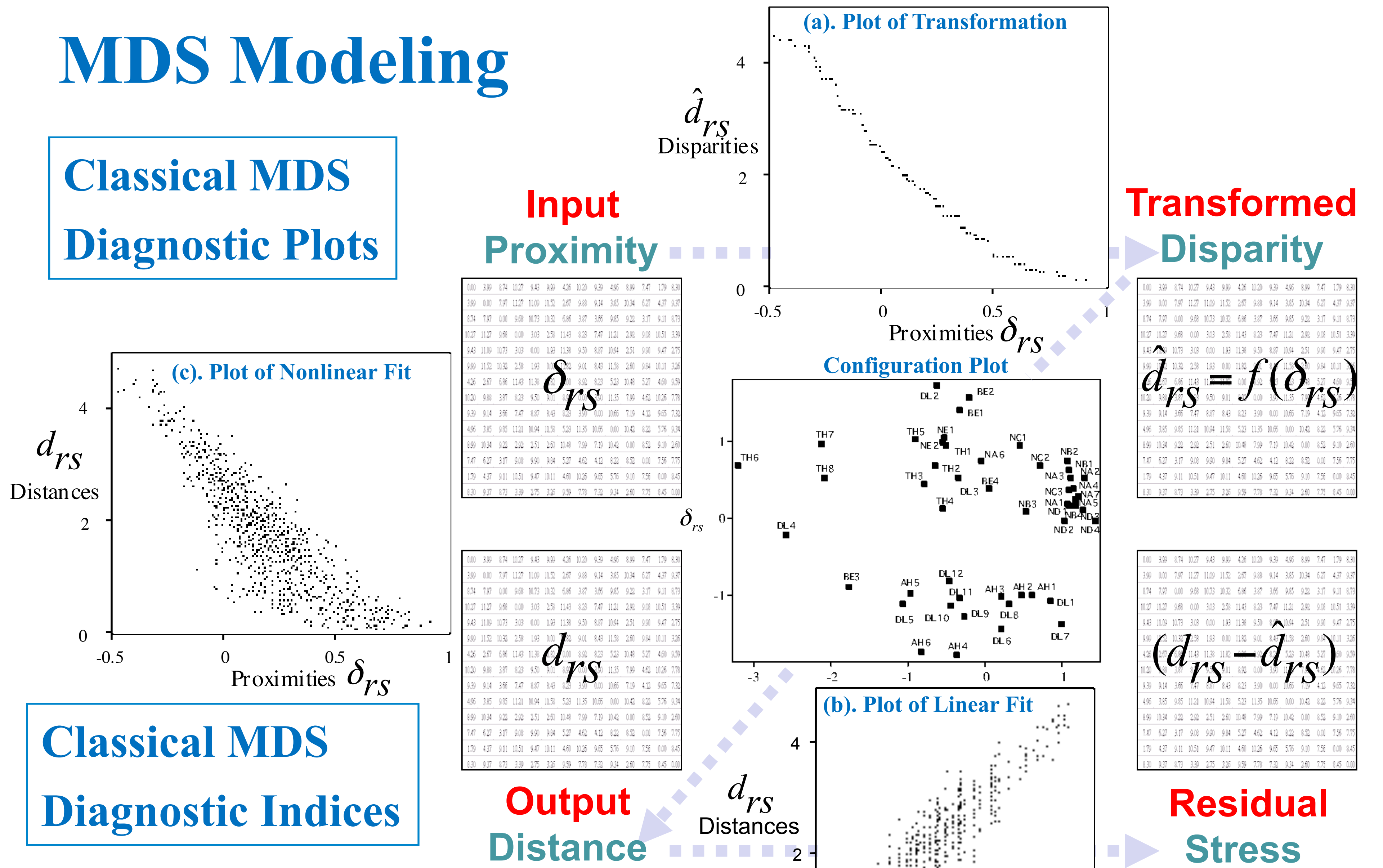




Output
Distance Matrix



Residual
Stress Matrix



$$STRESS(q) = \left\{ \frac{\sum_r \sum_s (d_{rs}^q - \hat{d}_{rs}^q)^2}{\sum_r \sum_s (d_{rs}^q)^2} \right\}^{1/2} \qquad SSTRESS(q) = \sum_r \sum_s ((d_{rs}^q)^2 - (\hat{d}_{rs}^q)^2)^2$$

cophenetic correlation (cc) = correlation of d_{rs} and $\hat{d}_{rs} = 0.883$

3. MV with dependent (clustered) structure

Family 1

Father1
Mother1
Sib11
Sib12

Family 2

Father2
Mother2
Sib21
Sib22

.
.
.

. . .

Family k

Fatherk
Motherk
Sibk1
Sibk2

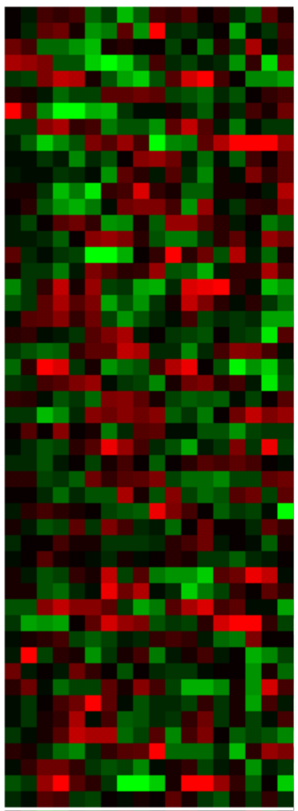
How to:

Compute proximity
for individuals?
for clusters?

Display matrices
for individuals?
for clusters?

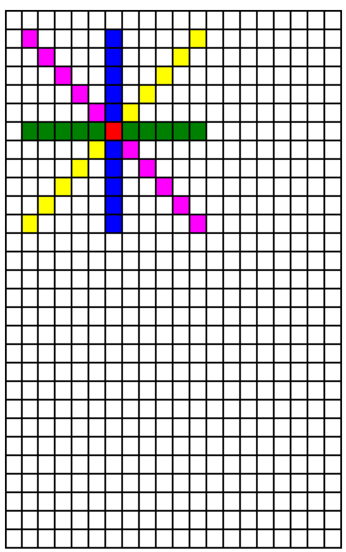
4. MV for data with missing values

Step 4
Evaluation



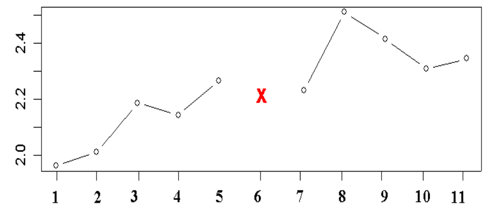
no. directions = 4

no. elements = 10

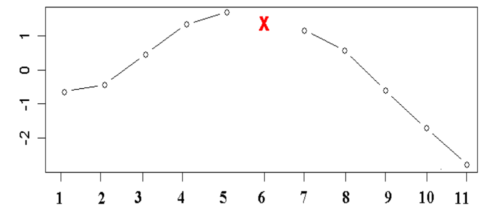


Missing Values

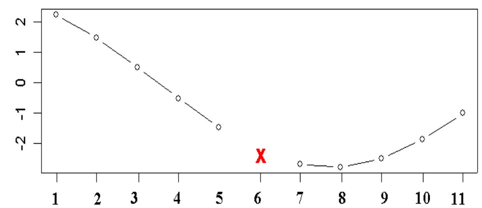
Horizontal



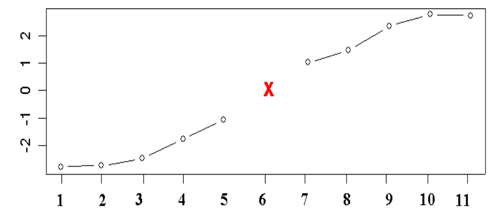
Positive



Vertical



Negative



(1) Fit Regression $\hat{Z}_d, \; d = 1, \cdots, 4.$

(2) Calculate weights

$$slope_d[i] = y_d[i+1] - y_d[i], \;\; i = 1, \cdots, 9.$$

$$w_d = \frac{1}{\text{var}(slope_d)}$$

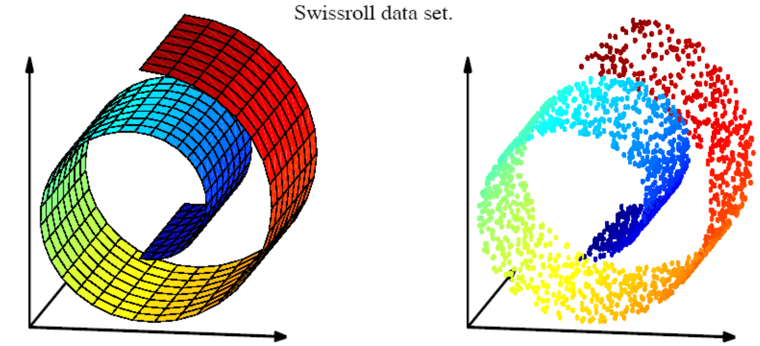
(3) Impute values

$$ImputedValue = \frac{\sum_{d=1}^{4} w_d \hat{Z}_d}{\sum_{d=1}^{4} w_d}$$

Approaching Statistics & Statistical Approach

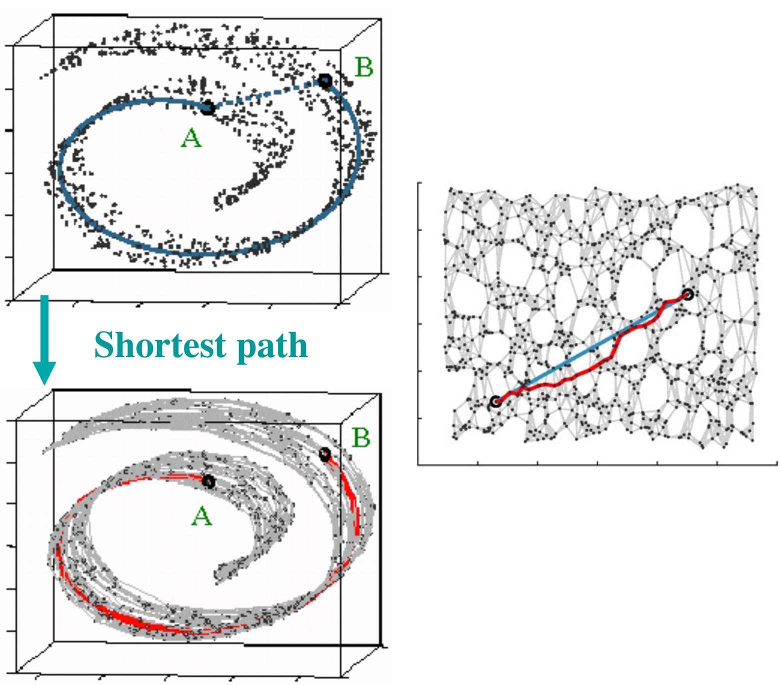
5. MV with nonlinear proximity measurement

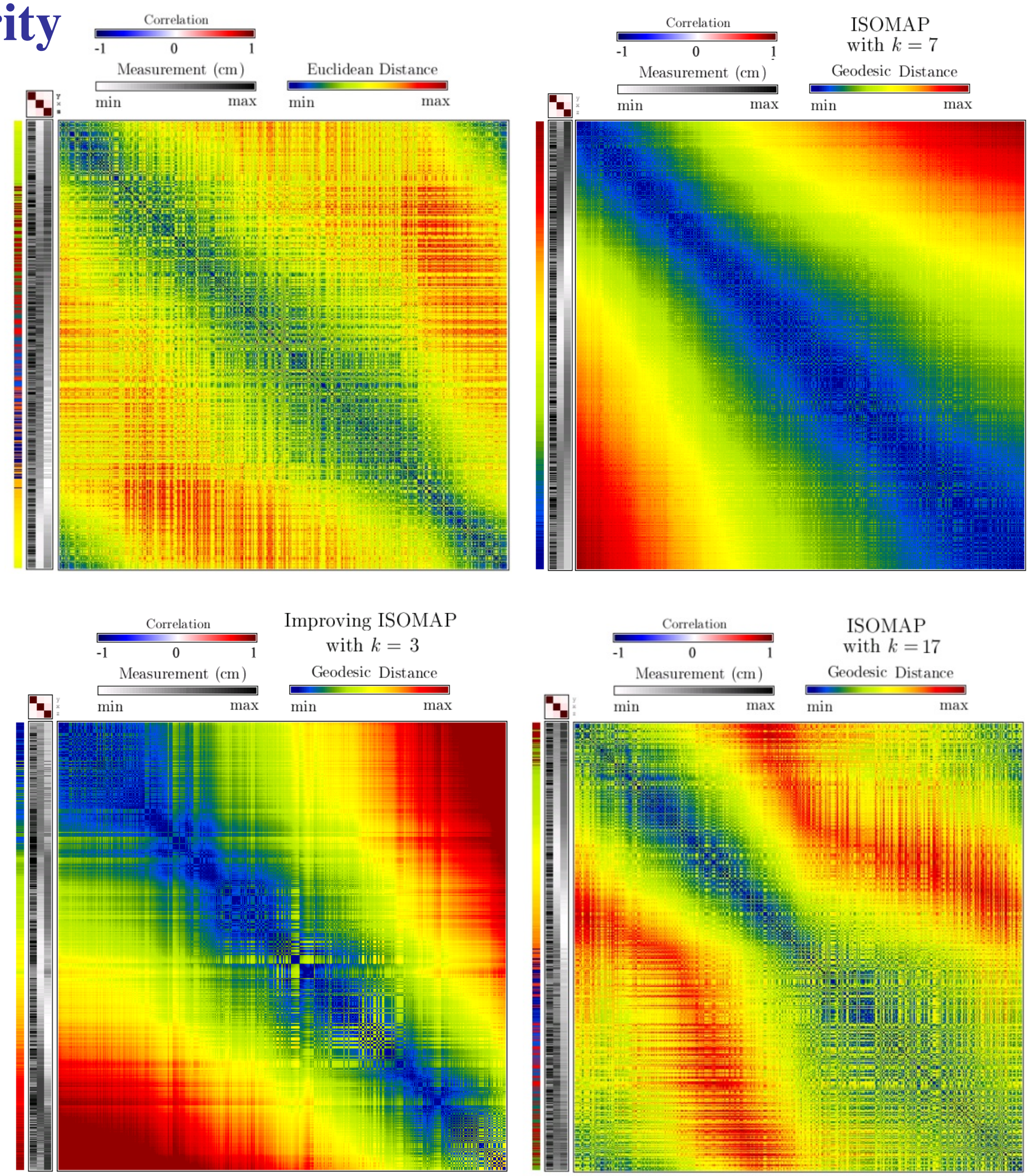
Concept of Manifolds and Nonlinearity



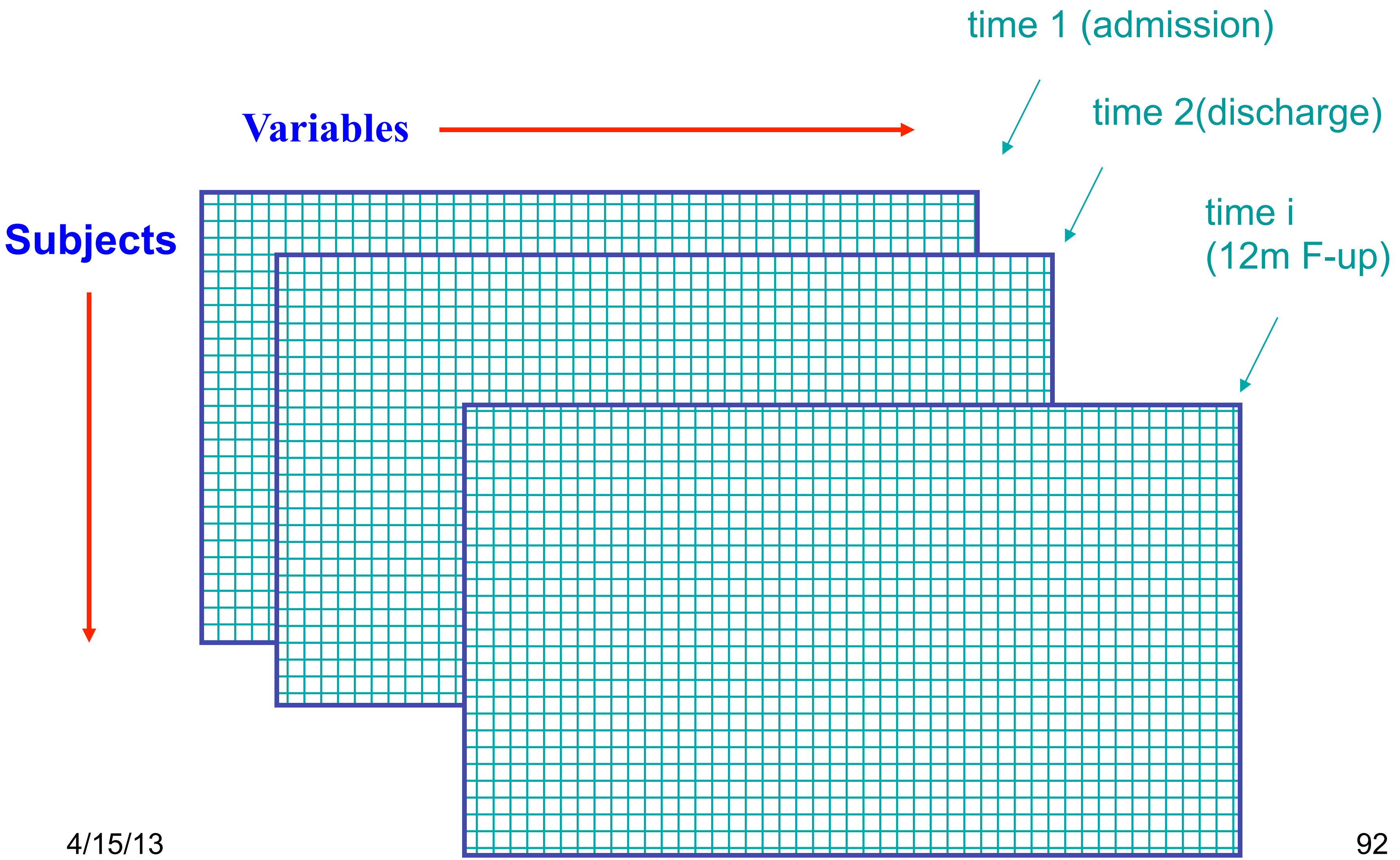
(a) manifold (b) sample data

Isometric Mapping (isomap)





6. MV for longitudinal multivariate data



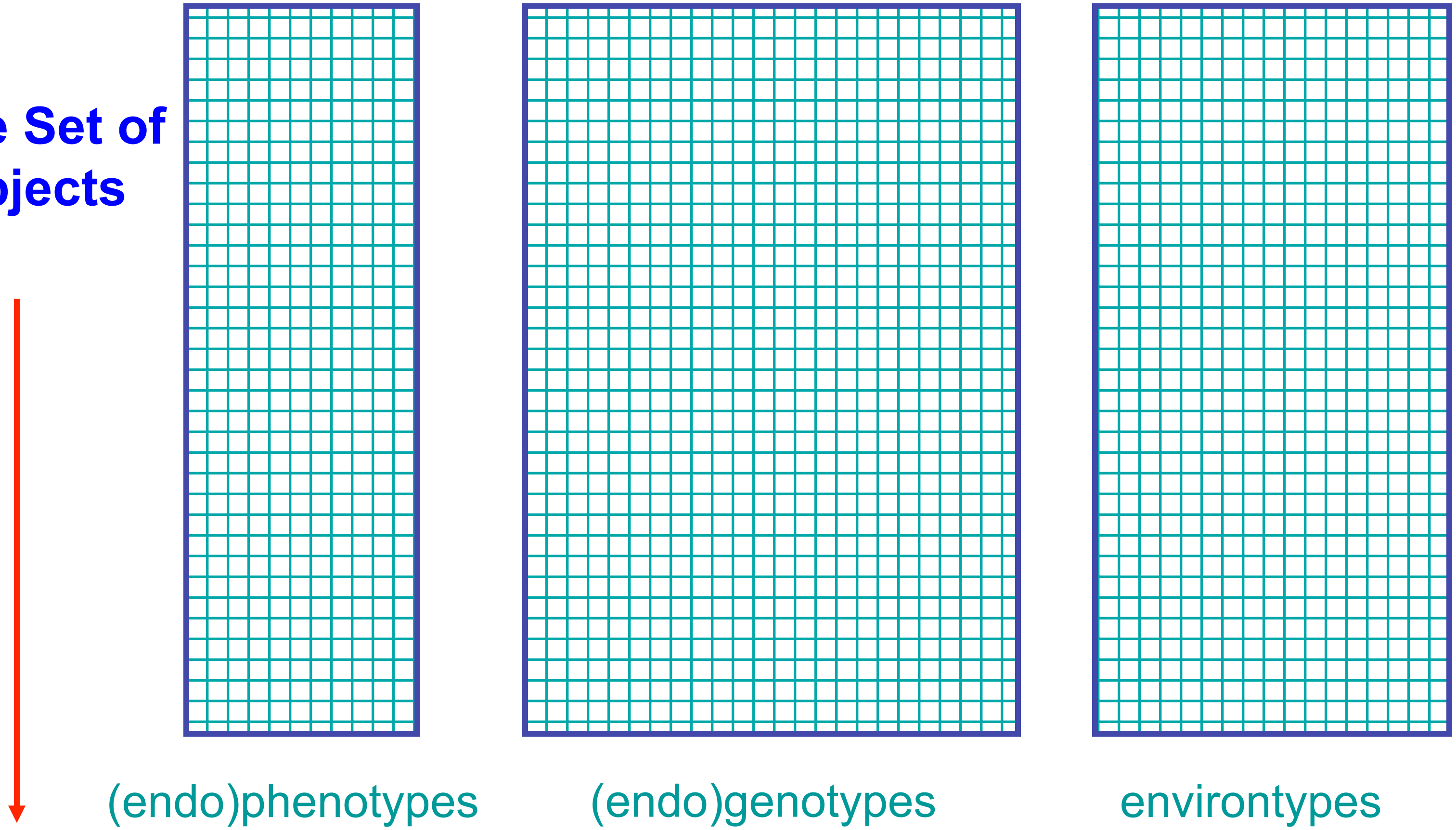
7. MV for multi-conditioned multivariate data

Set X
(CPT)
Raw Scores / Dprime / LnBeta

Set Y
(BCDS)
A / B / C / D …

Set Z
NPT
WAIS-R / WMS-R / Trail-A / Trail-B

Same Set of Subjects



(endo)phenotypes (endo)genotypes environtypes

8. MV with dependent variable(s)

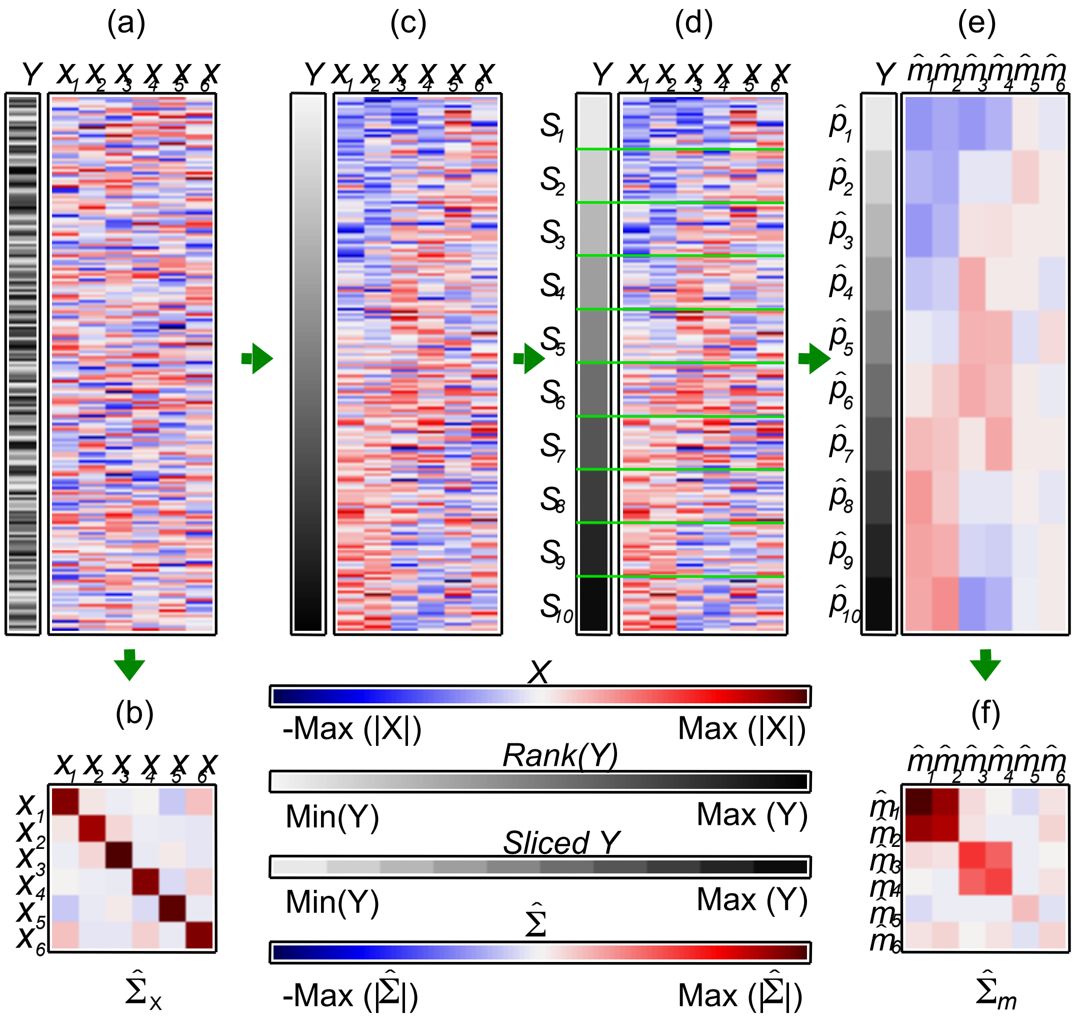


Figure 4. Matrix map of the raw data matrix $(Y, \mathbf{x})$ with a PCA analysis and the SIR algorithm. (a). original (unsorted) matrix map; (b). sample covariance matrix of $\mathbf{x}$ in (a), $\hat{\Sigma}_{\mathbf{x}}$; (c). sorted (by rank of Y) map; (d). sliced sorted map; (e) map for sliced mean matrix, $\hat{m}$; (f). sample covariance matrix of sliced mean matrix in (e), $\hat{\Sigma}_m$.

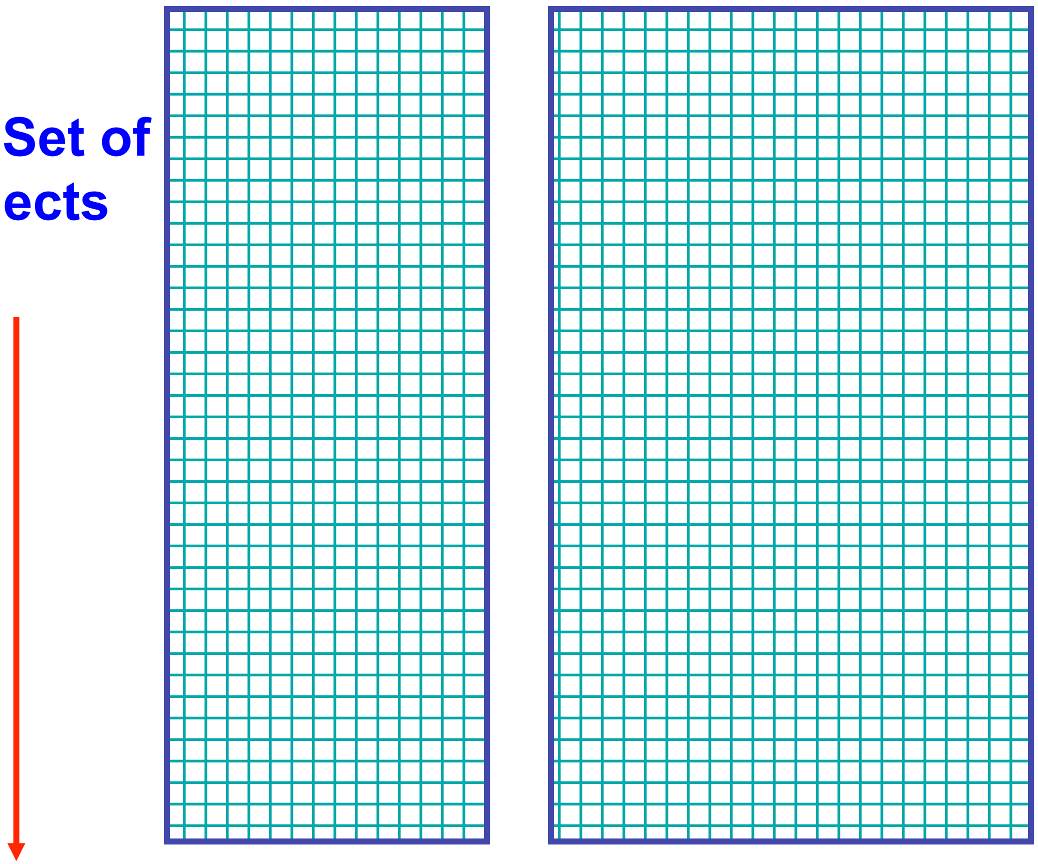
MV for a regression context with dependent variables is similar but not identical to MV with adjusting covariates.

Sliced inverse regression (SIR) Li (1991) is a natural staring point.

9. MV for mixed data

Continuous Categorical

Same Set of Subjects



1. Calculation of **proximity** matrices for variables and subjects

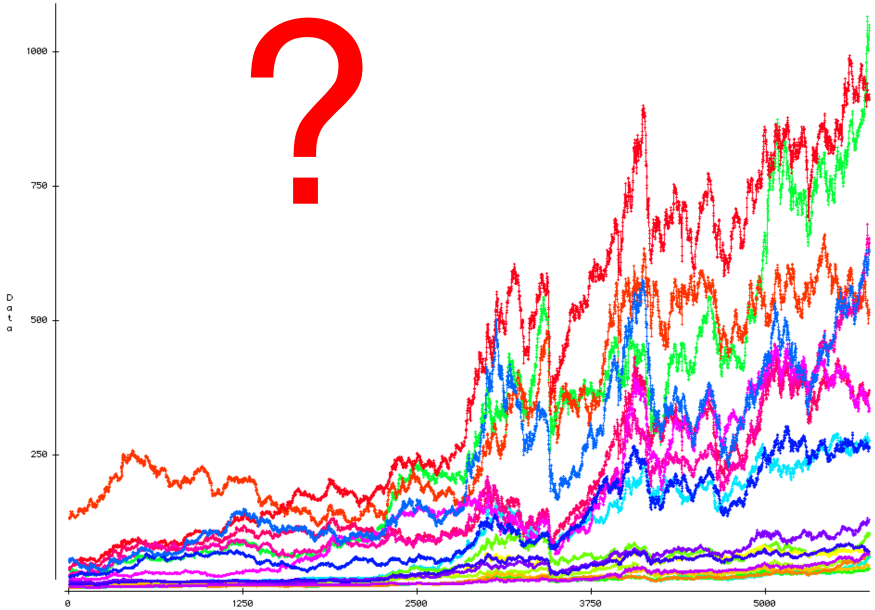
General similarity coefficients Gower (1971)

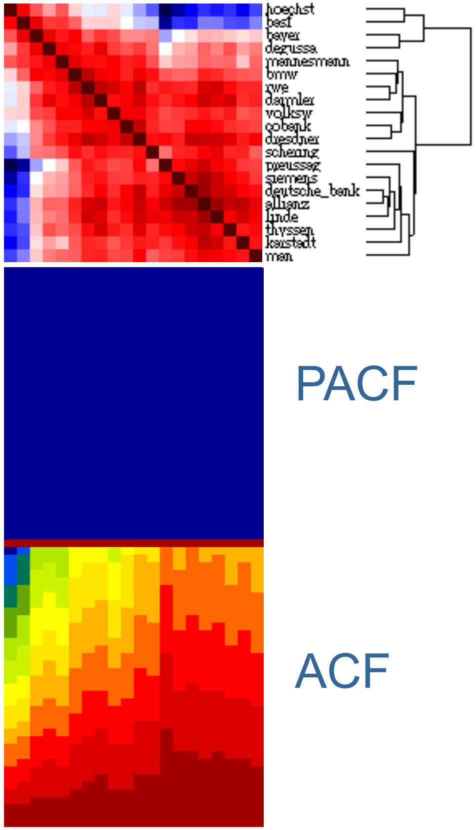
General weighted two-way dissimilarity coefficients introduced Cox and Cox (2000)

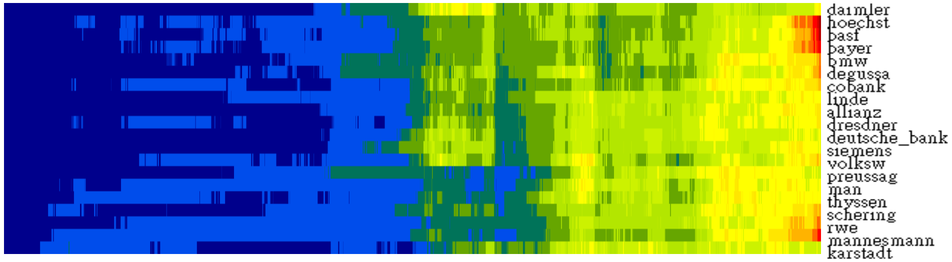
2. **Color coding** for a data matrix with mixed data is a more difficult task.

Approaching Statistics & Statistical Approach

11. MV for multiple time series data

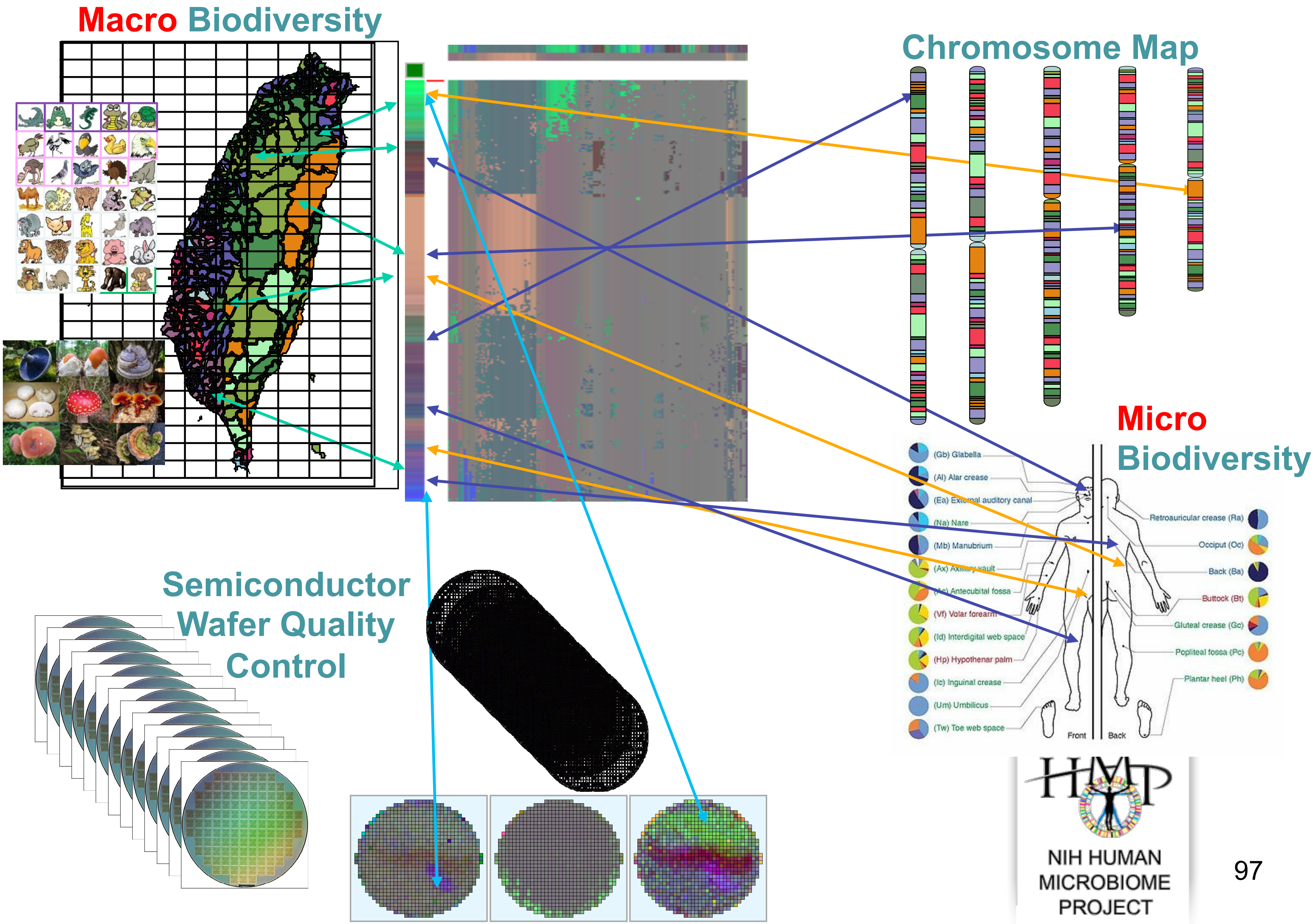






Data provided by Professor WOLFGANG HÄRDLE

From physical maps to conceptual maps



Approaching Statistics & Statistical Approach

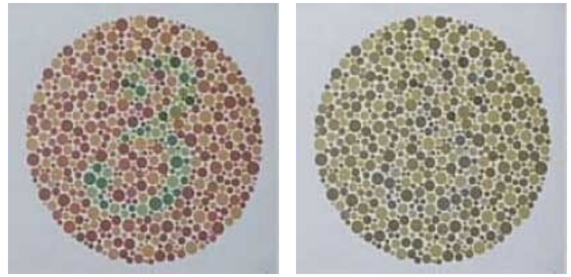
12. MV for Color Blind people

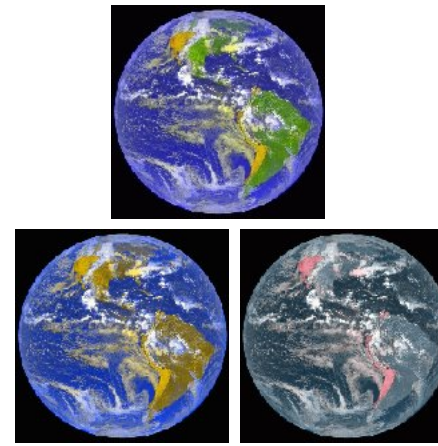
Vischeck

http://www.vischeck.com/examples/

Types of color blind

Monochromacy
Dichromacy
Protanopia and deuteranopia
Hereditary tritanopia
Anomalous Trichromacy



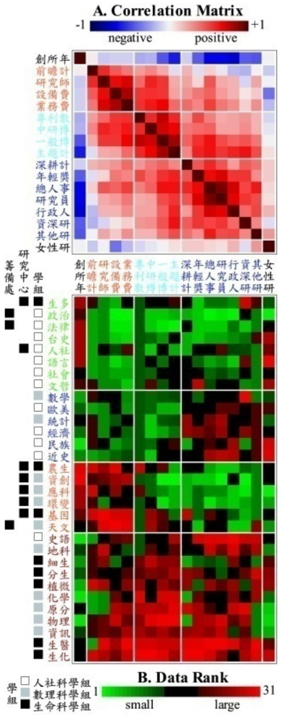


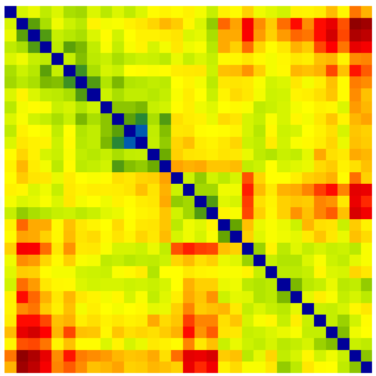
To act passively to prevent from using color systems that are difficult for color blind people to understand. or
To work actively in assisting people with visual impairments to have better visualization of data/information.

"I believe there are more mathematics/statistics blind people than color blind people"

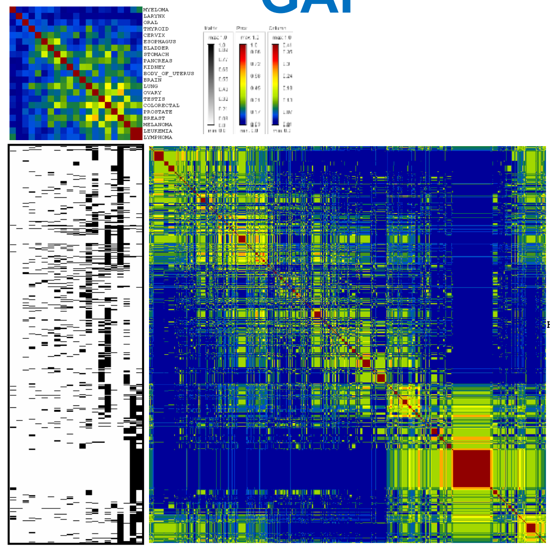
Approaching Statistics & Statistical Approach

Continuous GAP

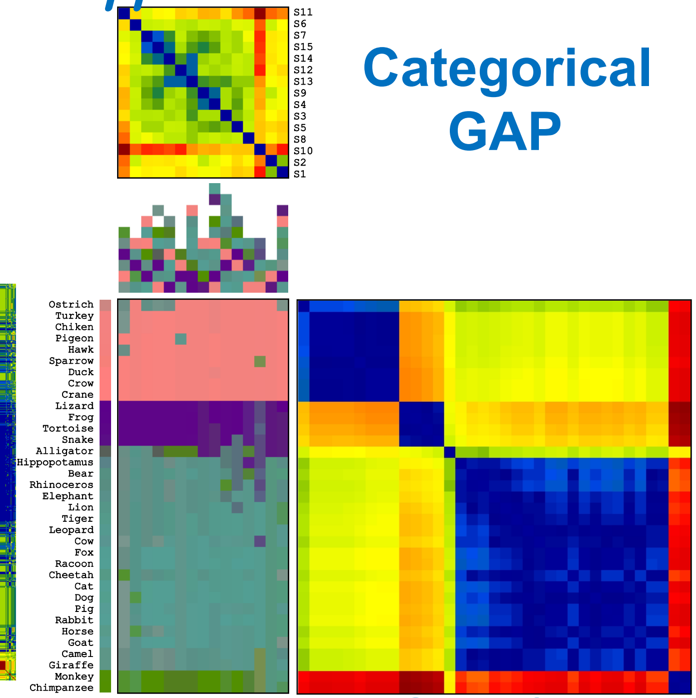




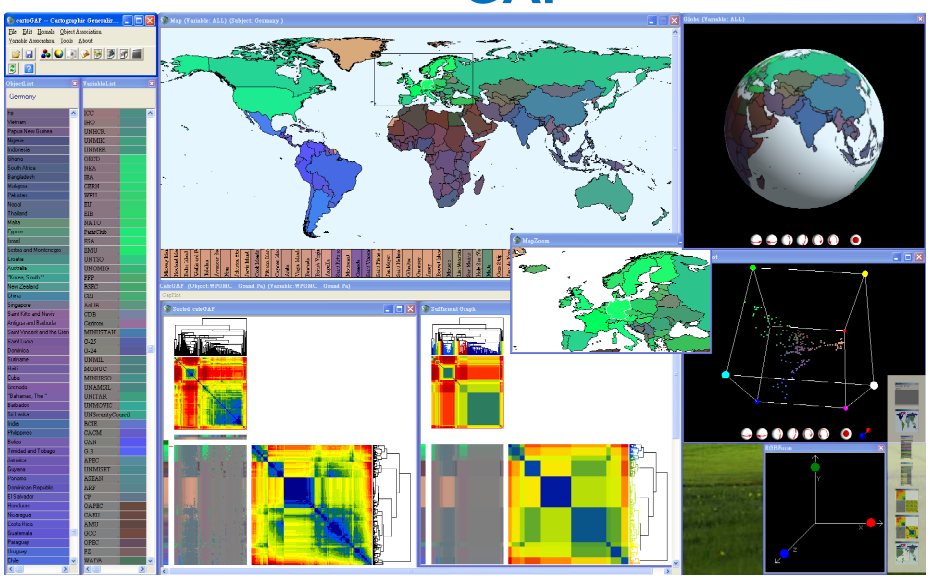
Binary GAP



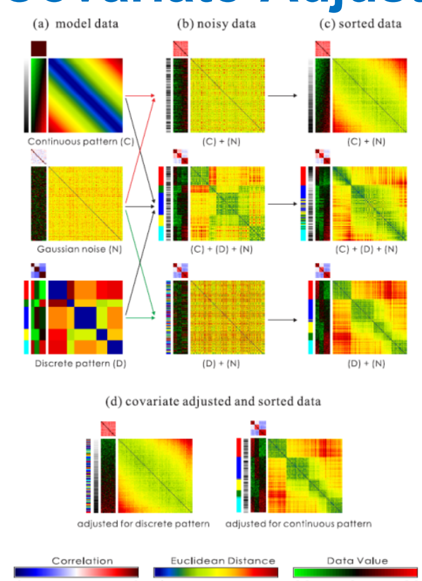
Categorical GAP



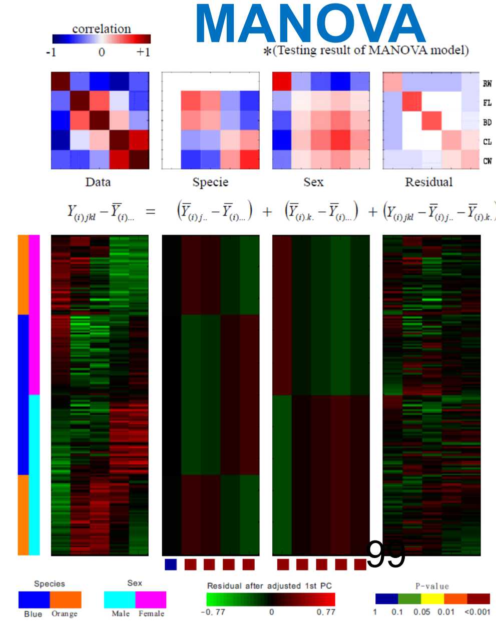
Cartography GAP



GAP with Covariate-Adjust



GAP for MANOVA



$$Y_{(i)jkl} - \bar{Y}_{(i)...} = (\bar{Y}_{(i)j..} - \bar{Y}_{(i)...}) + (\bar{Y}_{(i).k.} - \bar{Y}_{(i)...}) + (Y_{(i)jkl} - \bar{Y}_{(i)j..} - \bar{Y}_{(i).k.})$$

Thank You!