iasc isi International Association for Statistical Computing

October IASC-AMG Webinar

Data Visualization and Exploratory Data Analysis with Matrix Visualization

Chun-houh Chen

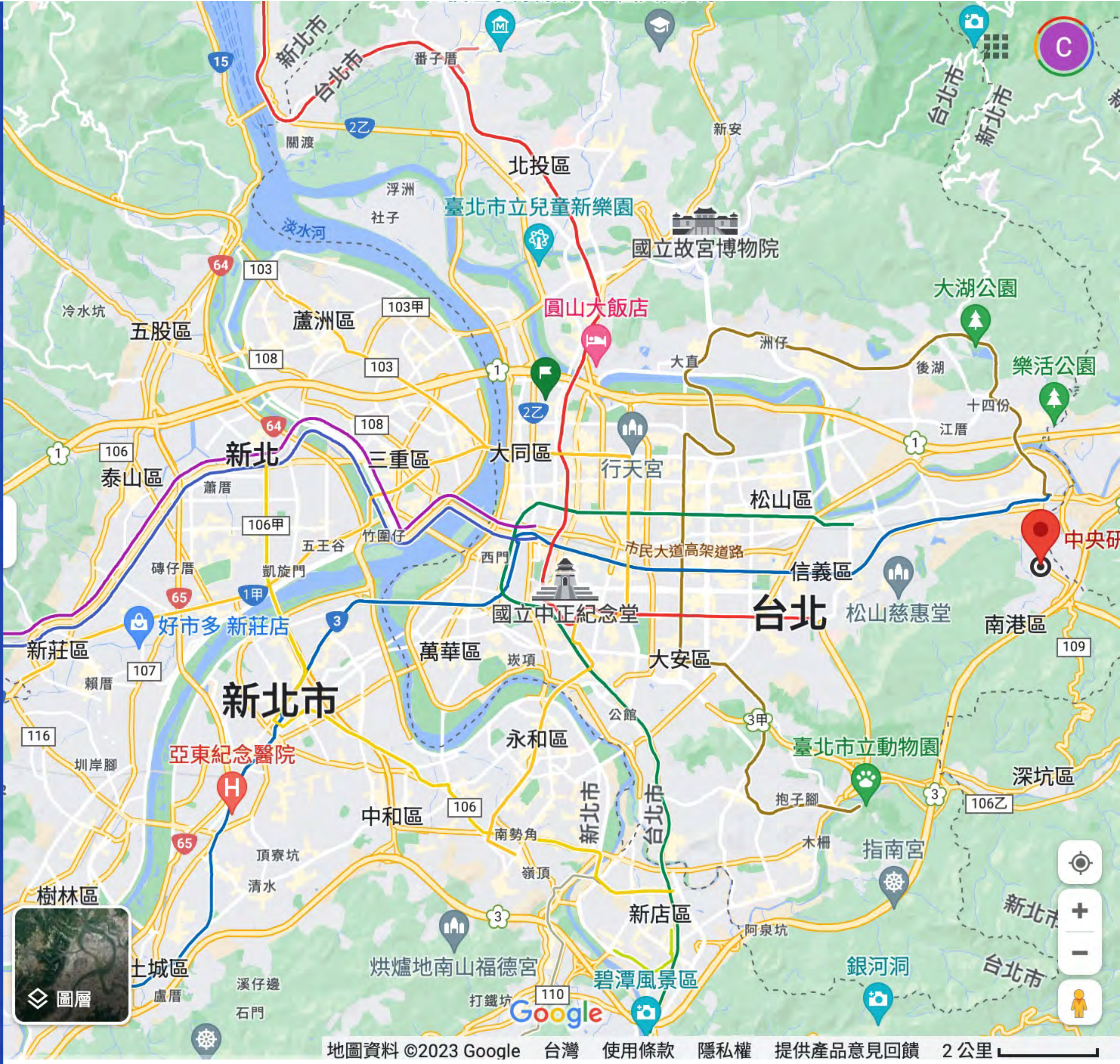
統計科學研究所
INSTITUTE OF STATISTICAL SCIENCE

Taiwan

WIKIPEDIA
The Free Encyclopedia

Taiwan
(Republic of China)

CHINA
120°
© Nations Online Project
122°
East China Sea
Sanzhi
Jinshan
Danshui
Wanli
Keelung City
Linkou
Taipei
Guanyin
Taoyuan
Xizhi
Fulong
25°
Zhongli
Jingtong
Hukou
Xindian
Bade
Yangmei
Toucheng
Wulai
Hsinchu City
Zhubei City
Yilan City
Guishan Is.
Zhudong
Sanwan
Suao
Tongxiao
Miaoli City
Dongao
Nanao
Xueshan Range
Aohua
Qinshui
Xueshan Snow Mountain (3,886 m)
Hemei
Taichung City
Taiping
Xincheng
Changhua City
Renai
24°
Puli
Hualien City
Nantou
Yuchi
Chechang
Sun Moon Lake
Shuilien
Lunbei
Douliu City
Xinyi
Fengbin
Shuilin
Zhuqi
Magong
Penghu Islands
(Pescadores Island)
Tropic of Cancer
Chiayi City
Alishan
Ruisui
Taibao City
Yushan Jade Mountain (3,952 m)
Shuishang
Budai
Yuli
Changbin
Xinying
Philippine Sea
Xuejia
Fuli
Madou
Yujing
Central Mountain Range
Hai'an Range
Laonong
Chenggong
Xinhua
23°
Anping
Tainan City
Dulan
Meinong
Gangshan
Wutai
Taitung City
Sandimen
Pingtung City
Kaohsiung City
Taimali
Green Island
Jinlun
Huatung Basin
Donggang
Fangliao
Dawu
Liuqiu Island
(Golden Lion Island)
Mudan
Xuhai
Checheng
Hengchun
Orchid Island
Kenting
22°
Maobitou
South China Sea
Luzon Strait
Taiwan Strait
Pescadores Strait
N
CHINA
Fuzhou
Matsu Islands
Putian
Quanzhou
Taiwan Strait
Taipei
TAIWAN
Xiamen
Quemoy Islands
(Kinmen Island)
TAIWAN
National capital
County capital
City, town, village
Major airport
International boundary
Highway
Main road
High Speed Rail
Railroad
0 20 40 60 80 km
0 10 20 30 40 50 mi

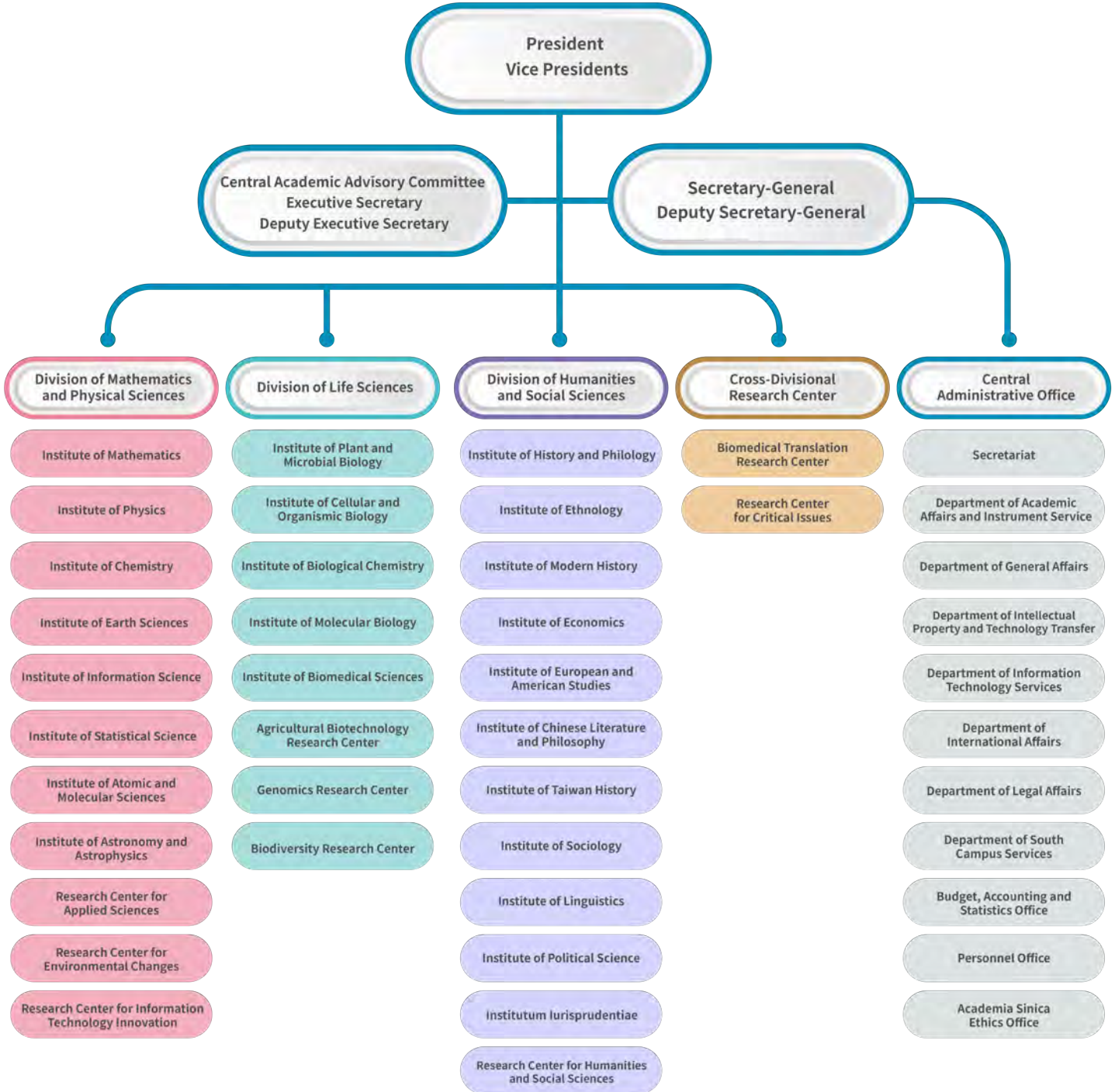


新北市
台北市
番子厝
關渡
新安
北投區
浮洲
社子
臺北市立兒童新樂園
國立故宮博物院
淡水河
冷水坑
五股區
蘆洲區
圓山大飯店
大湖公園
洲仔
大直
後湖
樂活公園
十四份
江厝
新北
三重區
大同區
行天宮
泰山區
蕭厝
松山區
五王台
竹圍仔
市民大道高架道路
西門
信義區
中央研
磚仔厝
凱旋門
好市多 新莊店
國立中正紀念堂
台北
松山慈惠堂
南港區
新莊區
萬華區
崁頂
大安區
賴厝
新北市
公館
永和區
臺北市立動物園
圳岸腳
亞東紀念醫院
深坑區
抱子腳
中和區
南勢角
木柵
指南宮
頂寮坑
嶺頂
清水
樹林區
新店區
阿泉坑
新北市
烘爐地南山福德宮
銀河洞
土城區
碧潭風景區
溪仔邊
打鐵坑
台北市
盧厝
石門
圖層
Google
地圖資料 ©2023 Google 台灣 使用條款 隱私權 提供產品意見回饋 2 公里

Overview of Academia Sinica

- Founded in 1928
- A staff of about 9,000, including approx. 900 PIs
- 24 research institutes and 9 research centers

- Missions:
  - ➢ To undertake in-depth research in sciences and humanities
  - ➢ To nurture young talents
  - ➢ To direct, coordinate and promote research activities to raise the academic standard in Taiwan


President
Vice Presidents
Central Academic Advisory Committee
Executive Secretary
Deputy Executive Secretary
Secretary-General
Deputy Secretary-General
Division of Mathematics and Physical Sciences
Institute of Mathematics
Institute of Physics
Institute of Chemistry
Institute of Earth Sciences
Institute of Information Science
Institute of Statistical Science
Institute of Atomic and Molecular Sciences
Institute of Astronomy and Astrophysics
Research Center for Applied Sciences
Research Center for Environmental Changes
Research Center for Information Technology Innovation
Division of Life Sciences
Institute of Plant and Microbial Biology
Institute of Cellular and Organismic Biology
Institute of Biological Chemistry
Institute of Molecular Biology
Institute of Biomedical Sciences
Agricultural Biotechnology Research Center
Genomics Research Center
Biodiversity Research Center
Division of Humanities and Social Sciences
Institute of History and Philology
Institute of Ethnology
Institute of Modern History
Institute of Economics
Institute of European and American Studies
Institute of Chinese Literature and Philosophy
Institute of Taiwan History
Institute of Sociology
Institute of Linguistics
Institute of Political Science
Institutum Iurisprudentiae
Research Center for Humanities and Social Sciences
Cross-Divisional Research Center
Biomedical Translation Research Center
Research Center for Critical Issues
Central Administrative Office
Secretariat
Department of Academic Affairs and Instrument Service
Department of General Affairs
Department of Intellectual Property and Technology Transfer
Department of Information Technology Services
Department of International Affairs
Department of Legal Affairs
Department of South Campus Services
Budget, Accounting and Statistics Office
Personnel Office
Academia Sinica Ethics Office

Life Sciences Boulevard

Statistical Science

Math. & Physical Sciences Boulevard

Humanities & Social Sciences Boulevard

Academia Sinica

Institute of Statistical Science

The Institute of Statistical Science was formally founded on August 3, 1987. The Institute currently has 31 research fellows, 26 postdoctoral fellows, 50 research assistants, and approximately 26 supporting staff members in administration and computing.

Institute of Statistical Science,

Academia Sinica

Statistica Sinica

-- one of the top ranked statistics journals.

Submission

410
No. of manuscripts received - 2023

19.3%
Acceptance rate - 2023

Speed

51
No. of days from submission to first decision - 2023.1-2023.12

516
No. of days from acceptance to publication - 2023.7-2024.6

Usage

419,894
No. of downloads - 2023.8-2024.7

653
No. of subscription - 2023.10-2024.9

Impact

1.5
IF - 2023

1.5
5 YR IF - 2023

31 / Q1
SJR - 2023
SCImago Journal Rankings - 2023

https://www3.stat.sinica.edu.tw/statistica/

OUTLINE

Exploratory Data Analysis (EDA)

Matrix Visualization (MV)

Generalized Association Plots (GAP)

MV for Continuous Data

MV for Binary Data

MV for Categorical Data

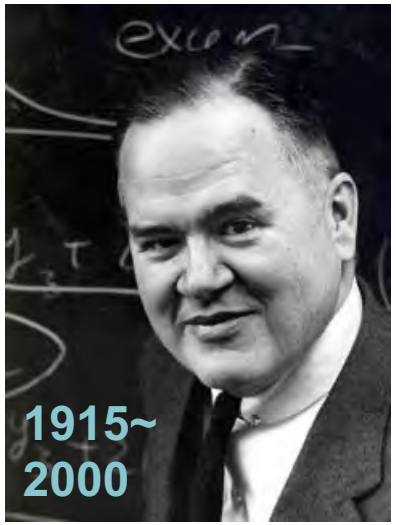
MV for Symbolic Data

MV for Cartography Data

MV for Big Data

Future Works?

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

Springer Handbooks of Computational Statistics

C. Chen
W. Härdle
A. Unwin
(Editors)

Handbook of Data Visualization

Springer

Visualizing the data is an essential part of any data analysis. Modern computing developments have led to big improvements in graphic capabilities and there are many new possibilities for data displays. This new volume in the series Springer Handbooks of Computational Statistics gives an overview of modern data visualization methods, both in theory and practice. There are definitive chapters on modern graphical tools such as mosaic plots, parallel coordinate plots and linked views. There are chapters dedicated to graphical methodology for particular areas of statistics, for example Bayesian analysis, genomic data and cluster analysis, as well as chapters on software for graphics. Specialists from all over the world have contributed papers on their areas of expertise.



9 783540 330363

› springer.com

Chen | Härdle | Unwin (Eds.)
Handbook of Data Visualization

International Statistical Review (2008)

Short Book Reviews

Editor: Simo Puntanen

Readership: Researchers and practitioners of almost any field.This handbook shows hundreds of ways to visualize data by using modern, high-quality statistical graphics. The articles of 66 authors reveal basics and backgrounds as well as details and dynamics of this fascinating area, which should be an essential part of any data analysis or statistical modelling.

The book includes over 500 examples of different graphs, such as parallel coordinate plots, grand tours, mosaic plots, matrix diagrams, micromap plots, and linked views. The interplay between multivariate statistical methods and various graphics is evident in several articles. In visualizing huge data sets efficiently, the advances in computer hardware and software have made totally new possibilities available. It is most enjoyable to see such a large number of specialists sharing their insights of these methods within one volume. This book really feeds the imagination of the reader. High-dimensionally recommended!

Handbook of Data Visualization

2nd edition ?




吳漢銘
Han-Ming Wu (Hank)
Tamkang University

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

GAP: A graphical environment for matrix visualization and cluster analysis

Han-Ming Wu [a], Yin-Jing Tien [b], Chun-houh Chen [c,*]

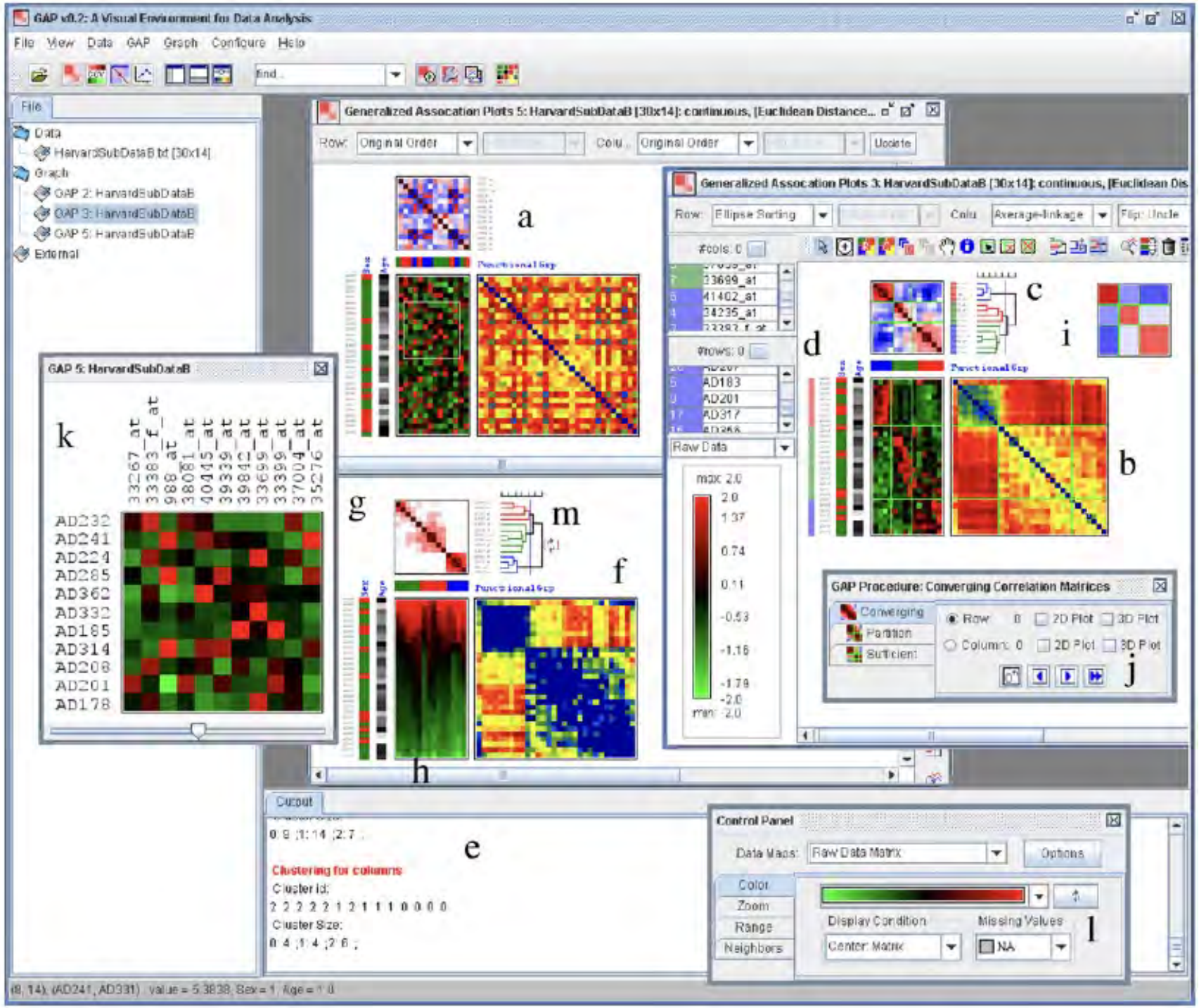
[a] *Department of Mathematics, Tamkang University, Taipei County 25137, Taiwan*
[b] *Institute of Statistics, National Central University, Taoyuan 32001, Taiwan*
[c] *Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan*

ARTICLE INFO

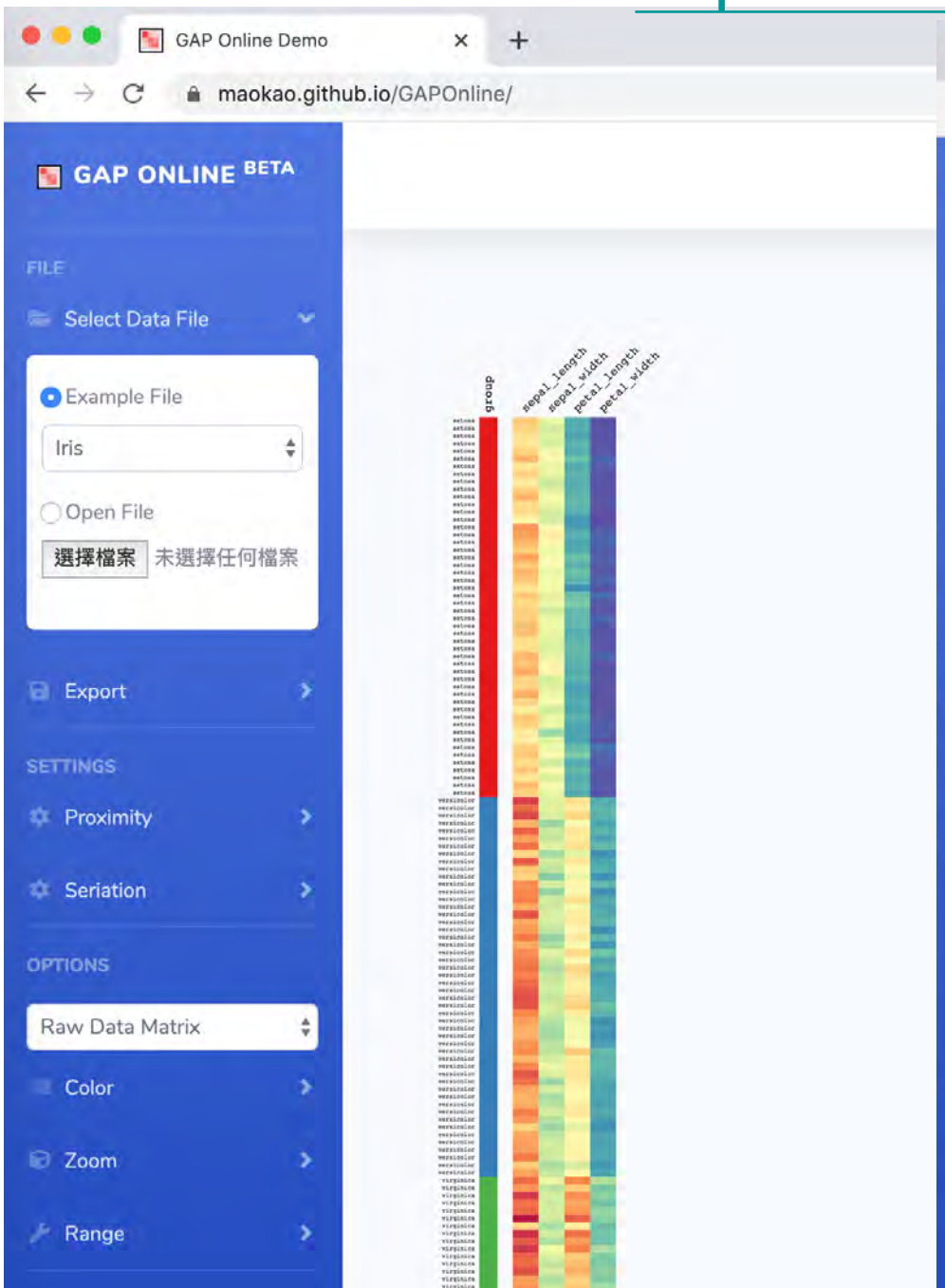
Article history:
Available online 8 October 2008

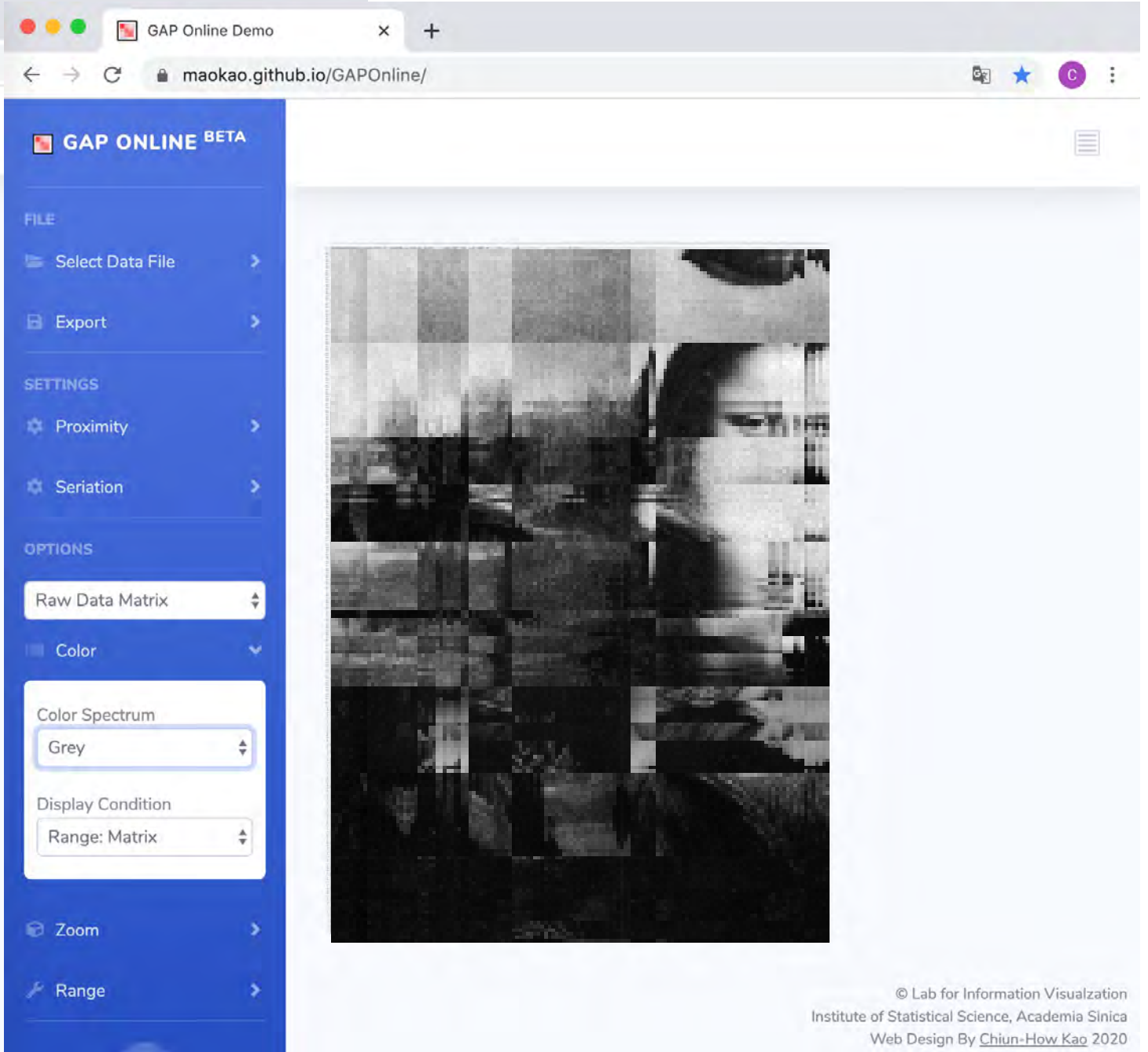
ABSTRACT

GAP is a Java-designed exploratory data analysis (EDA) software for matrix visualization (MV) and clustering of high-dimensional data sets. It provides direct visual perception for exploring structures of a given data matrix and its corresponding proximity matrices, for variables and subjects. Various matrix permutation algorithms and clustering methods with validation indices are implemented for extracting embedded information. GAP has a friendly graphical user interface for easy handling of data and proximity matrices. It is more powerful and effective than conventional graphical methods when dimension reduction techniques fail or when data is of ordinal, binary, and nominal type.


GAP v0.2: A Visual Environment for Data Analysis
File View Data GAP Graph Configure Help
Generalized Association Plots 5: HarvardSubDataB [30x14]: continuous, [Euclidean Distance...
Generalized Association Plots 3: HarvardSubDataB [30x14]: continuous, [Euclidean Dis
GAP 5: HarvardSubDataB
GAP Procedure: Converging Correlation Matrices
Control Panel
Data Matrix: Raw Data Matrix
Options
Clustering for columns
Cluster Id:
Cluster Size:
a
b
c
d
e
f
g
h
i
j
k
l
m

https://maokao.github.io/GAPOnline/





Topics：Statistical graphics and data visualization

PartI: Data Desk for Interactive Statistical graphics
PartI: GAP for high-dimensional data visualization

Preparation for the lecture：----------

1. Install Data Desk: Please visit the following website and install Data Desk 30 Day Trial
https://datadescription.com/trial/

---

2. Install GAP (Generalized Association Plots) ***GAP 64bit version***
http://gap.stat.sinica.edu.tw/Software/download.htm (click the link with mouse right-button)
With the iGAP (interval generalized association plots) software at:
http://gap.stat.sinica.edu.tw/Software/iGAP/index.htm (click the link with mouse right-button)
And here is the link to online version of GAP (for small size data demonstration):
https://maokao.github.io/GAPOnline/

---

3. Download demo files for Data Desk and GAP at:
http://gap.stat.sinica.edu.tw/Software/download.htm (click the link with mouse right-button)

---

4. download this handbook as your statistical graphics/visualization reference:
https://link.springer.com/book/10.1007%2F978-3-540-33037-0

Why Matrix Visualization?

a Taiwan PM2.5 example:

73 Environmental Protection Administration Stations

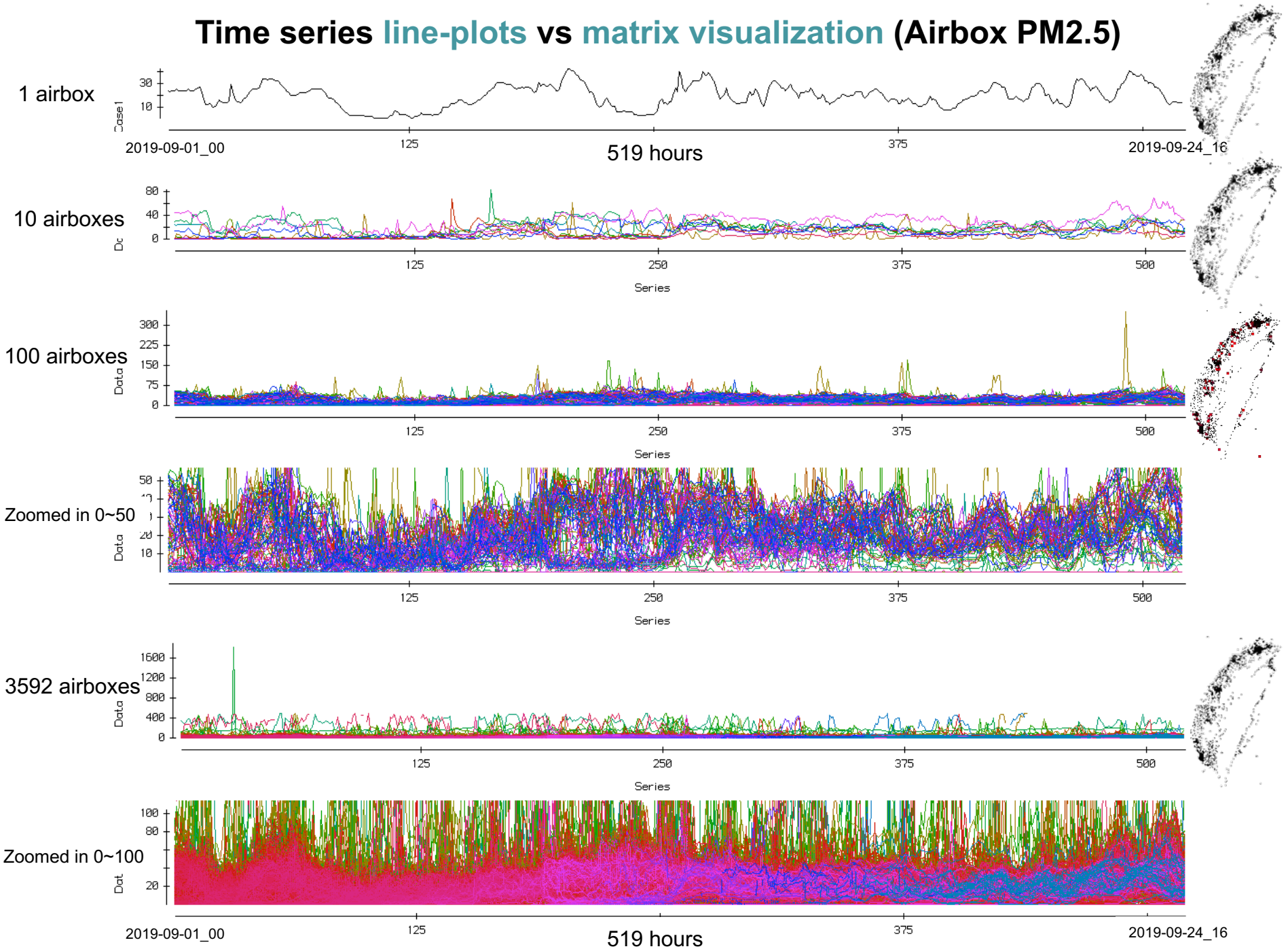
vs.

8000+ Airboxes

2019.Sep.01(Sun):00 ~ 2019.Sep.24(Tue):16 (519 hours)

Time series line-plots vs matrix visualization (Airbox PM2.5)

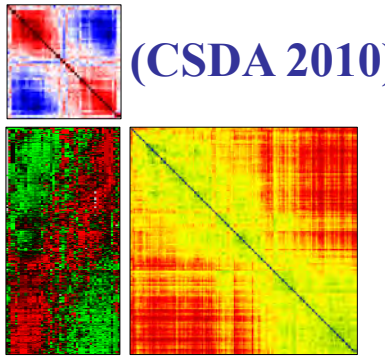
24-Hour Average ug/m3	Index Values	AQI Category
0.0 - 12.0	0 - 50	Good
12.1 - 35.4	51 - 100	Moderate
35.5 - 55.4	101 - 150	Unhealthy for Sesitive Groups
55.5 - 150.4	151 - 200	Unhealthy
150.5 - 250.4	201 - 300	Very Unhealthy
250.5 - 350.4	301 - 400	Hazardous
350.5 - 500	401 - 500	



Matrix visualization (MV) for Exploratory Data Analysis (EDA)

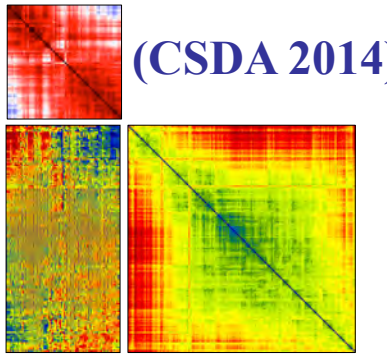
Modules of GAP for Matrix Visualization

Continuous X



(CSDA 2010)

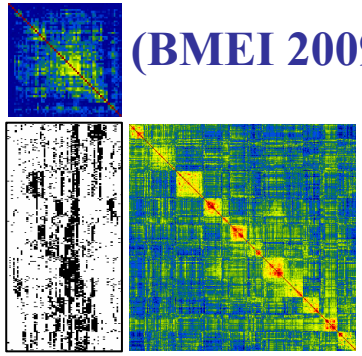
Symbolic X



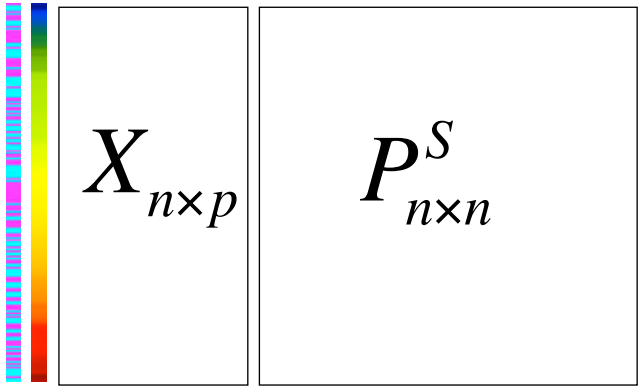
(CSDA 2014)

$P^{V}_{p\times p}$

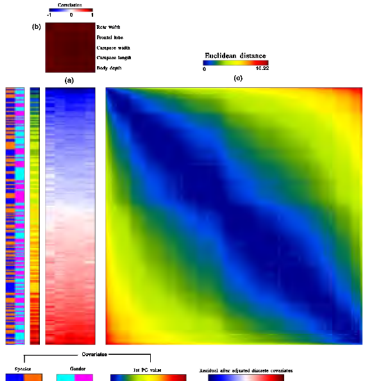
Binary X



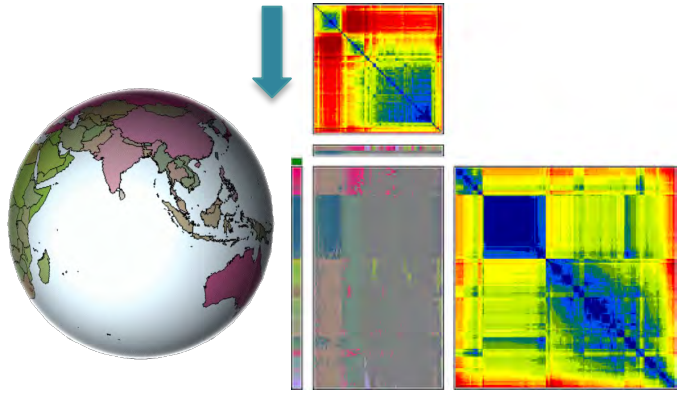
(BMEI 2009)



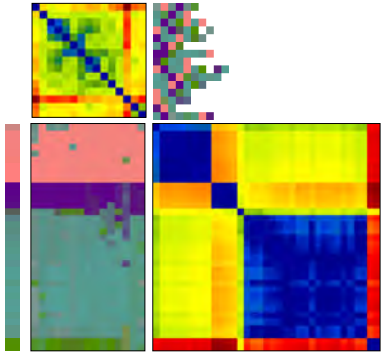
Covariate Adjusted X



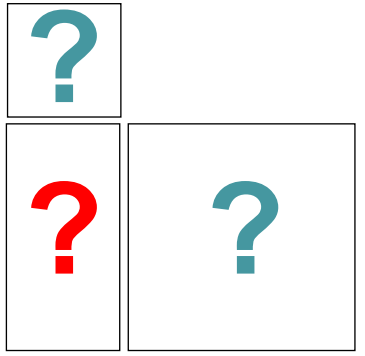
Cartographical X



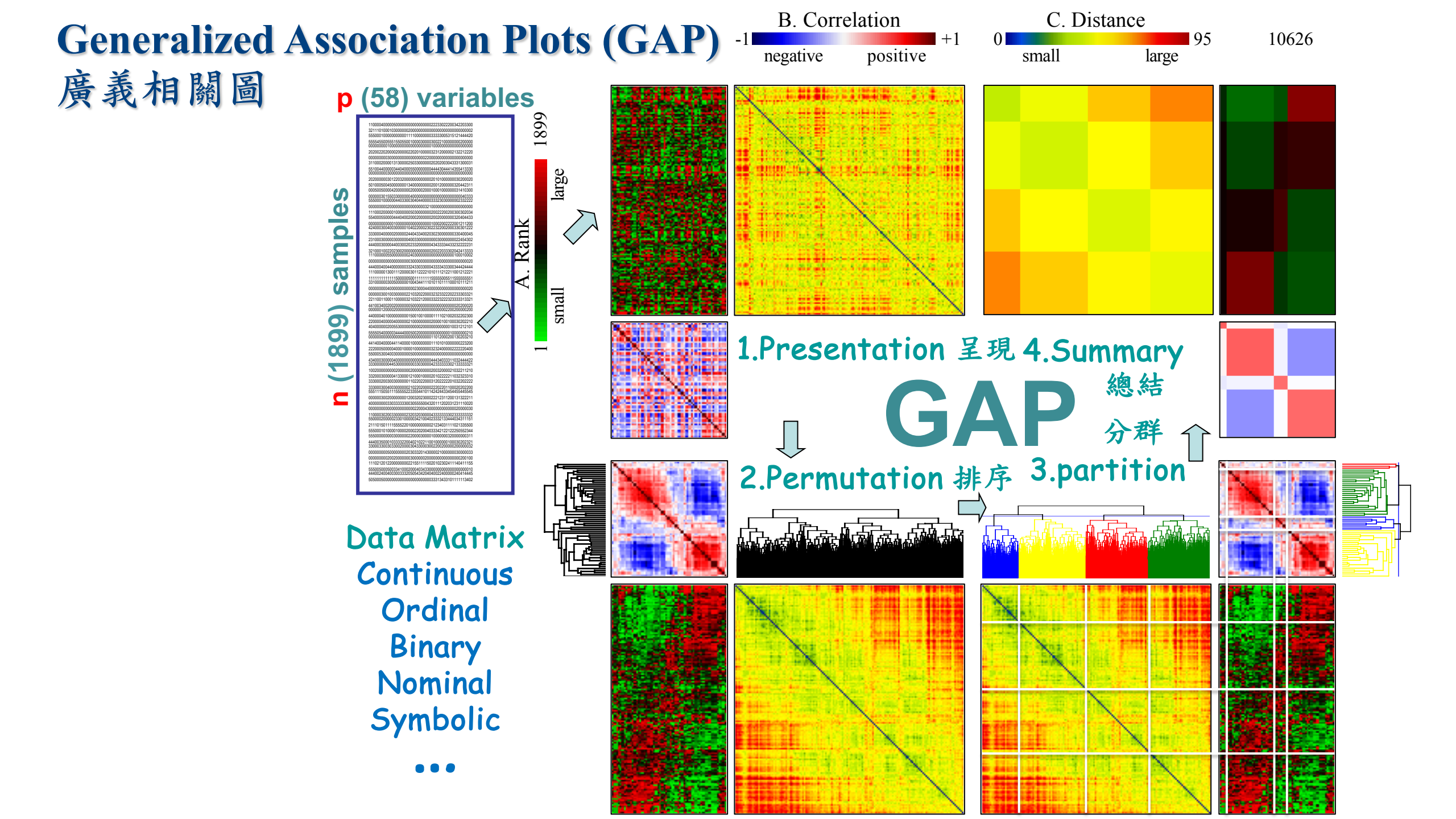
Categorical X



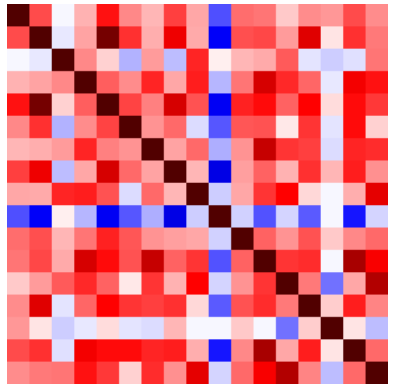
??????? X



Generalized Association Plots (GAP)
廣義相關圖

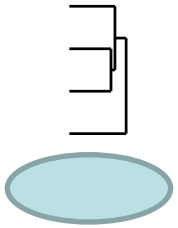


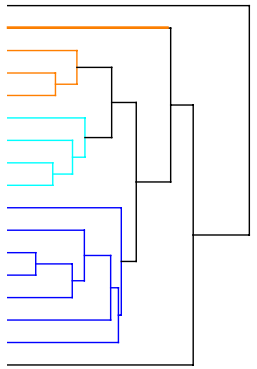
Some essential elements in a GAP MV procedure
一些 廣義相關圖 矩陣視覺化 的主要元素

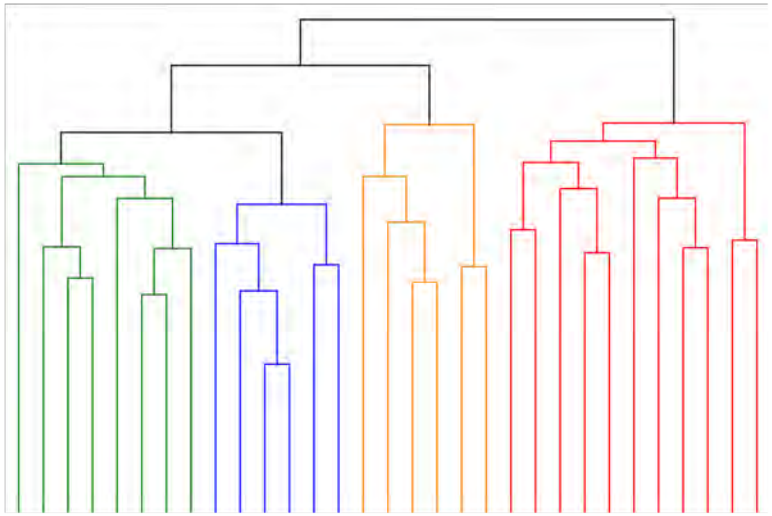


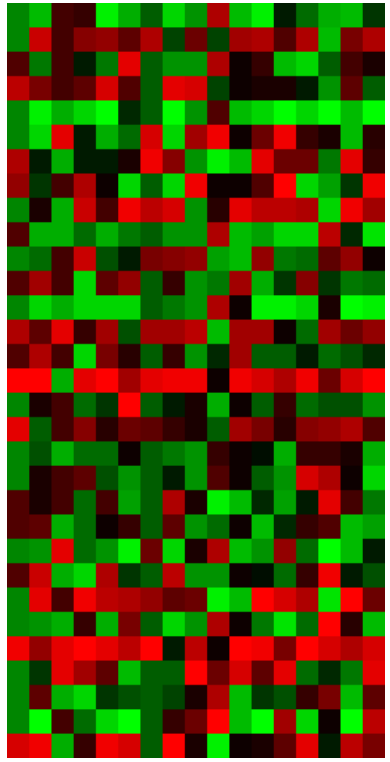
3. Proximity (Variable *p* * *p*)
變數間關係矩陣
Contin, Ordinal, Binary, Nominal

4. Permutation 排序 (variable)

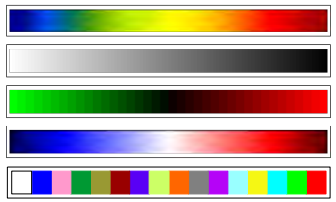






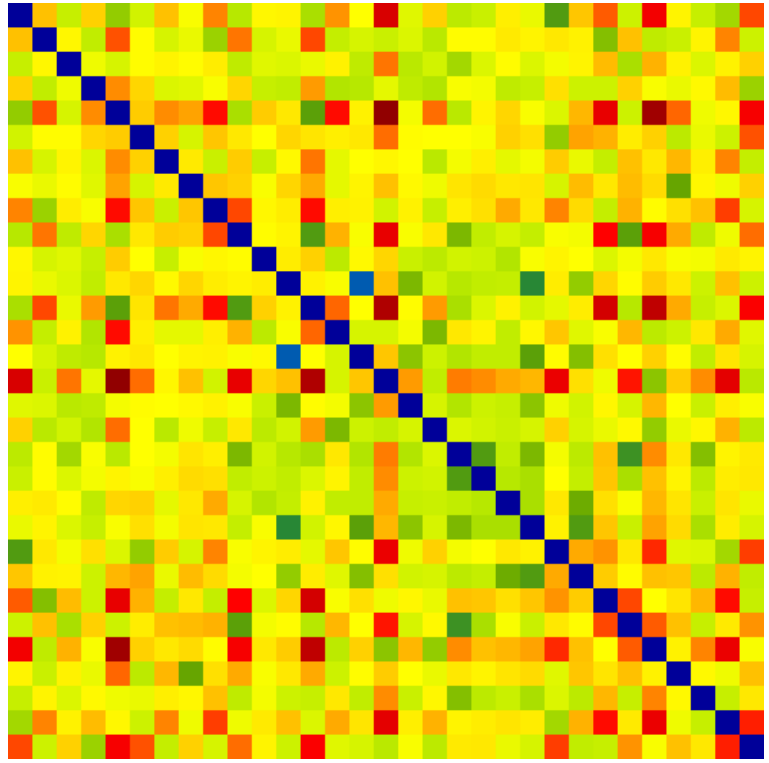


1. Data Matrix
資料矩陣
(*n* * *p*)
(w/ Color coding)
Continuous
Ordinal
Binary
Nominal



2. Proximity Matrix for Subject
樣本間之關係矩陣
(*n* * *n*)

連續型 Continuous
有序型 Ordinal
二元型 Binary
名目型 Nominal



Results on different seriation methods on rows and columns

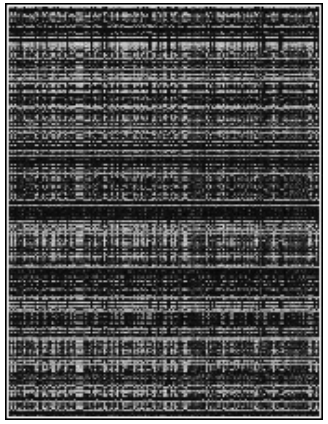
Treat the La Gioconda as a data matrix

Purpose of Permutation:

- present data structure
- allow the data to speak for themselves
- notice what we never expected to see
- place columns/rows with close proximity together with each other (relativity of a statistical graph, Chen 2002)

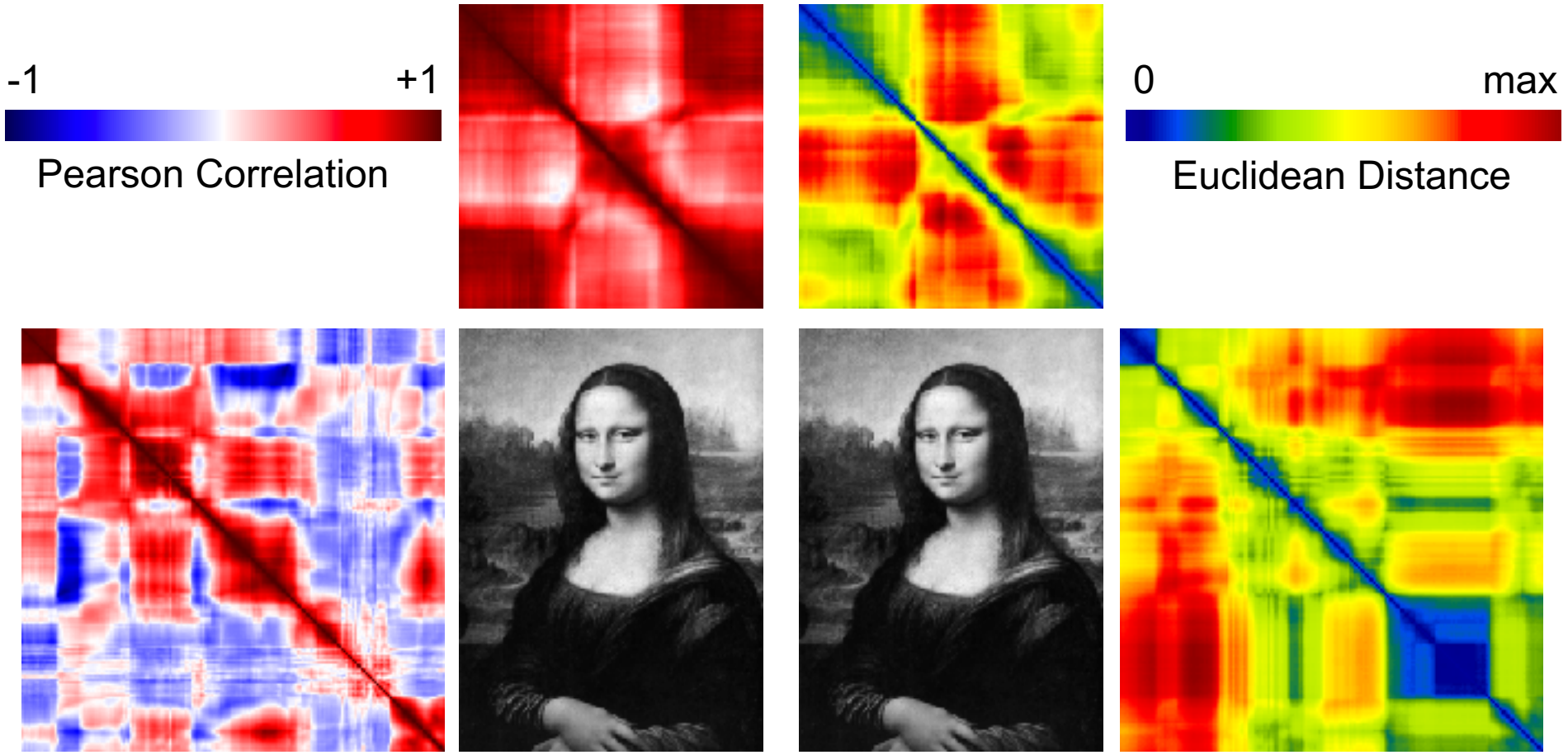
Original orders

Randomize Permutations

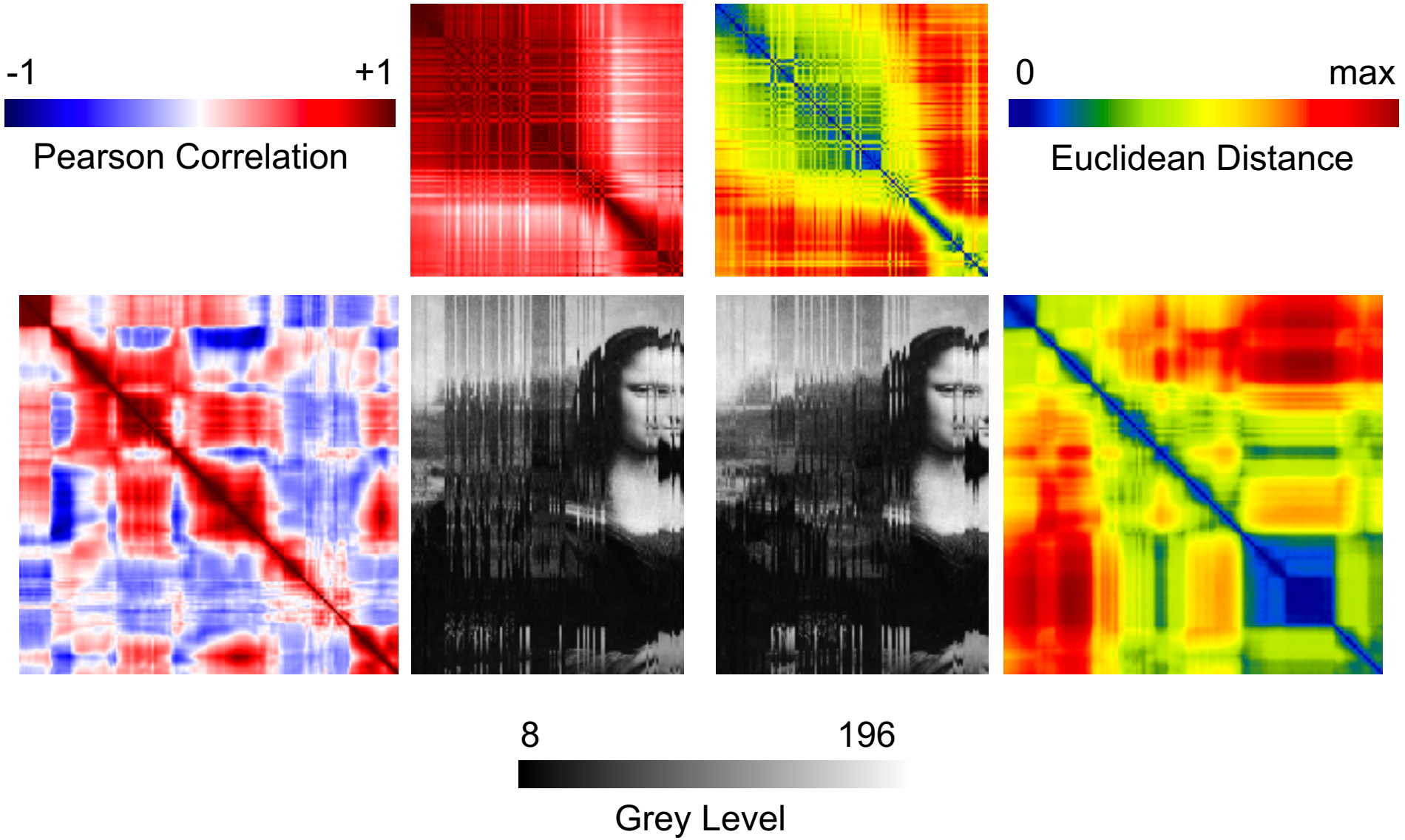


We shall call this concept of placing similar (distinct) objects at positions close to (far away from) each other in a plot for representing the association structure the concept of relativity of a statistical graph.

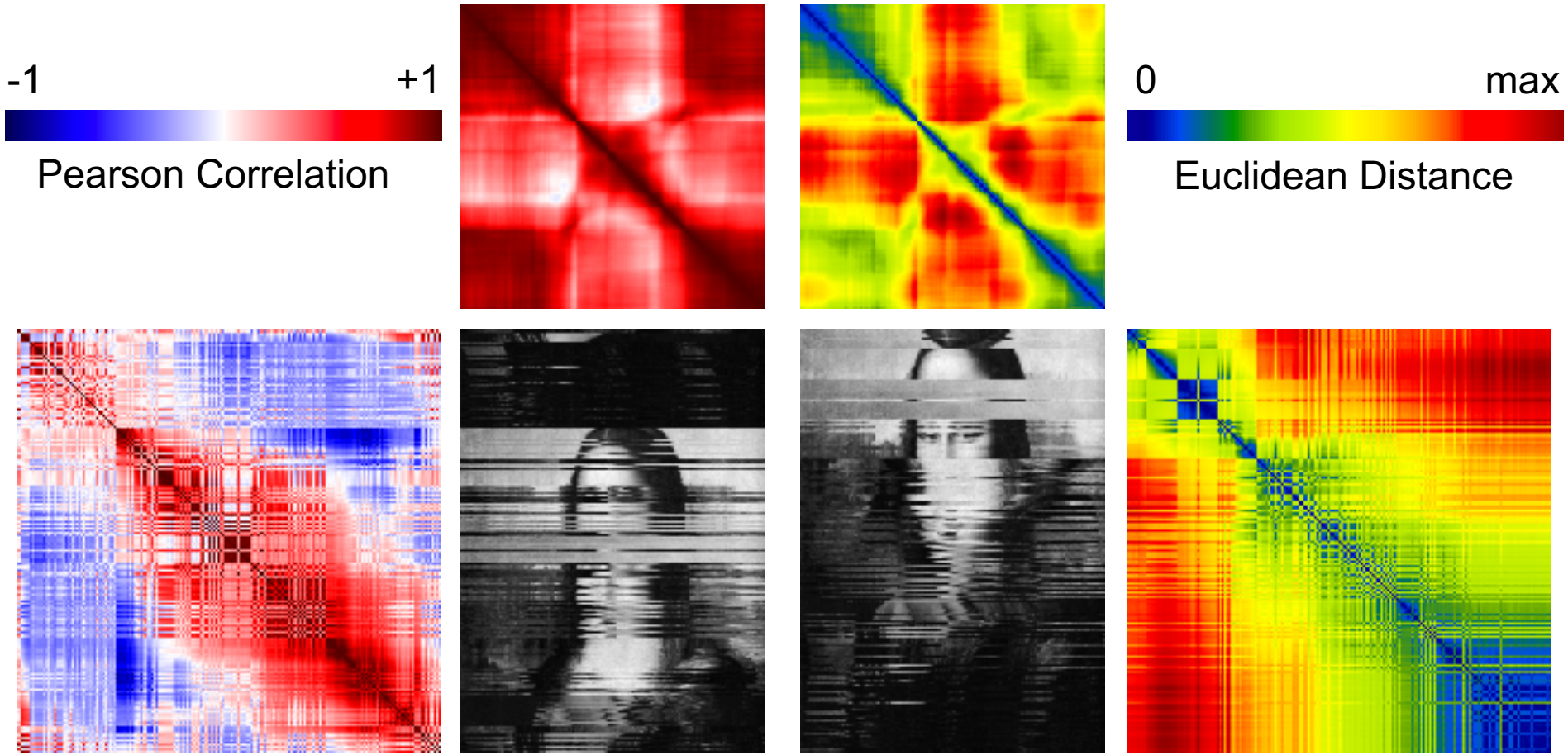
Original Orders



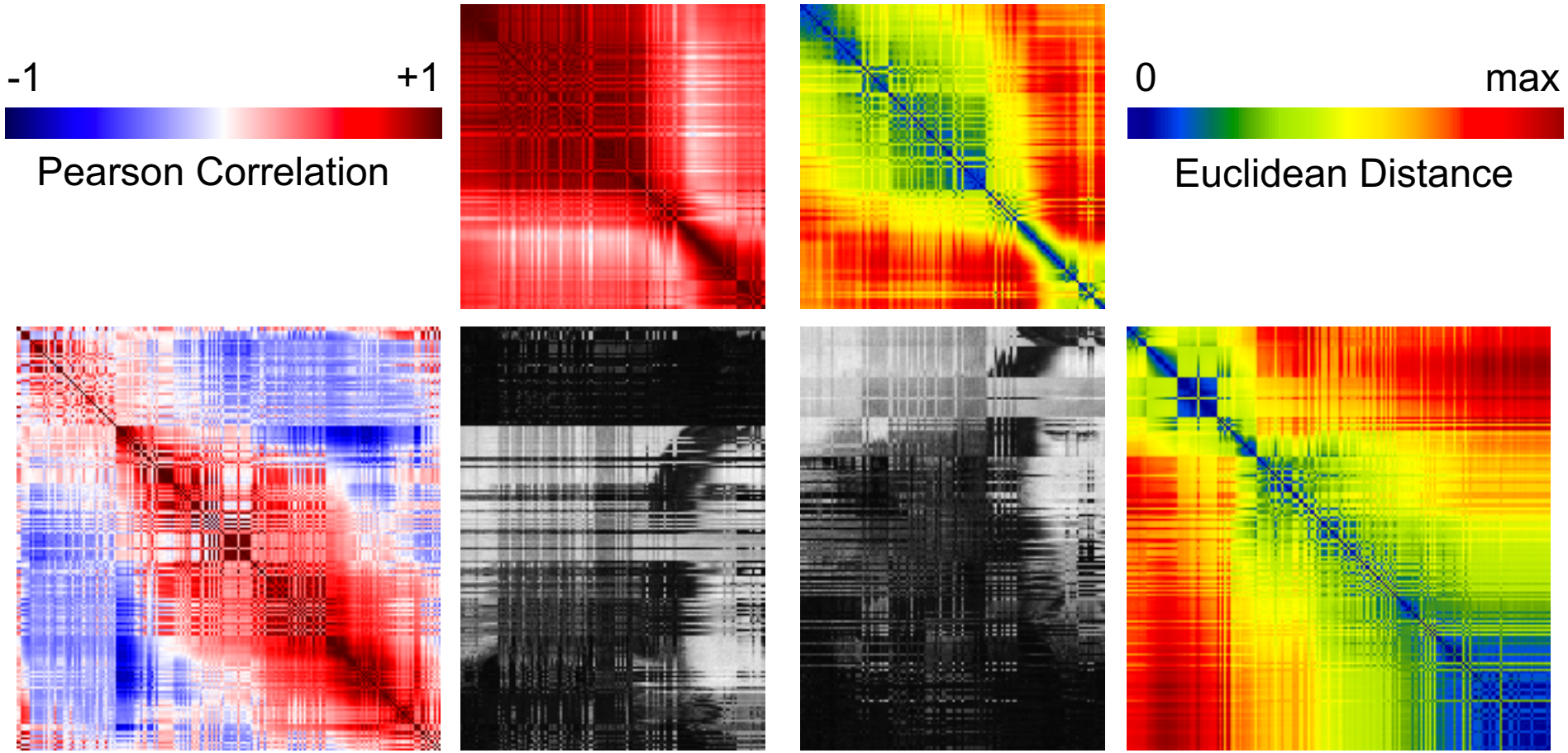
Rank 2 Elliptical Seriation col

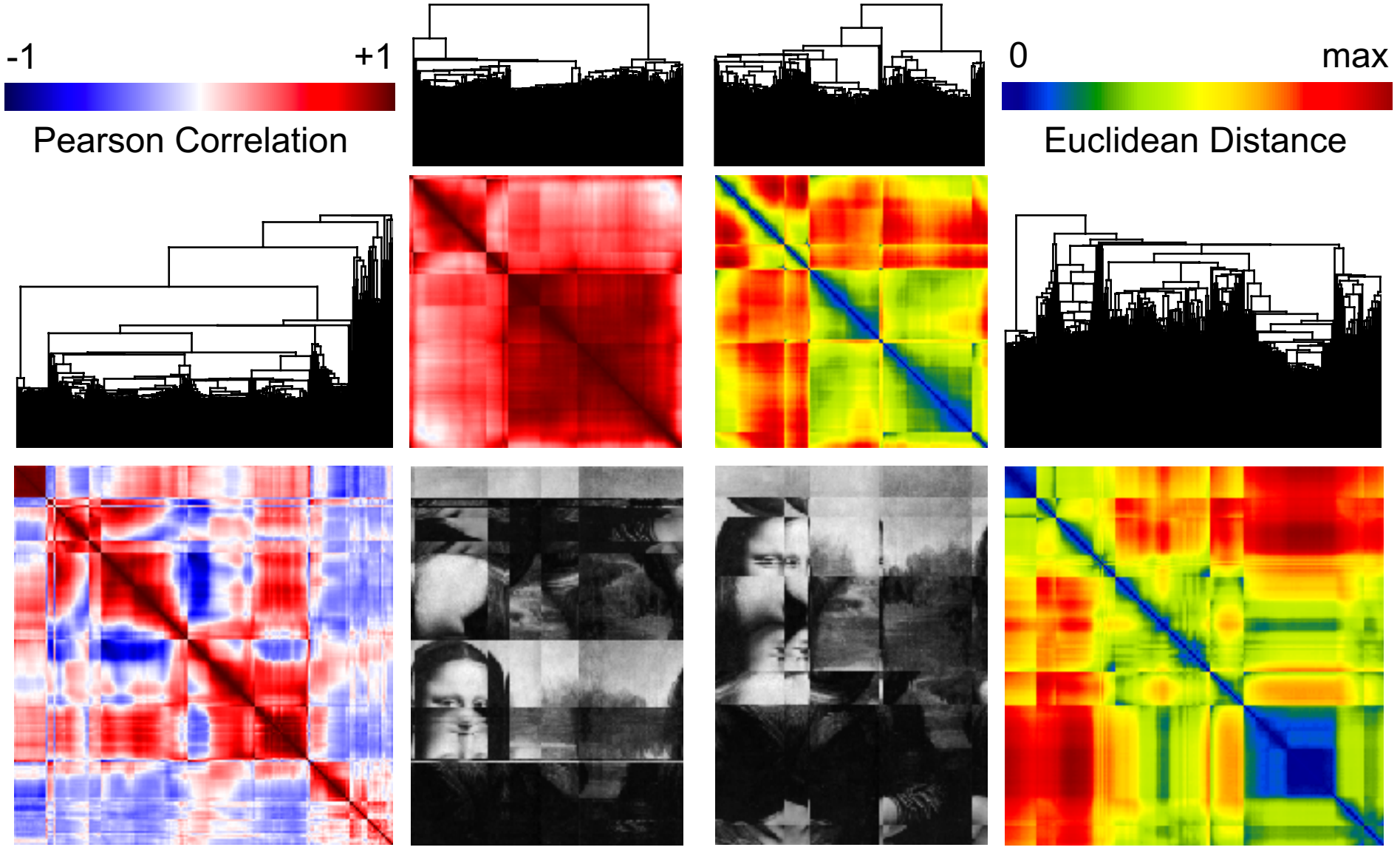


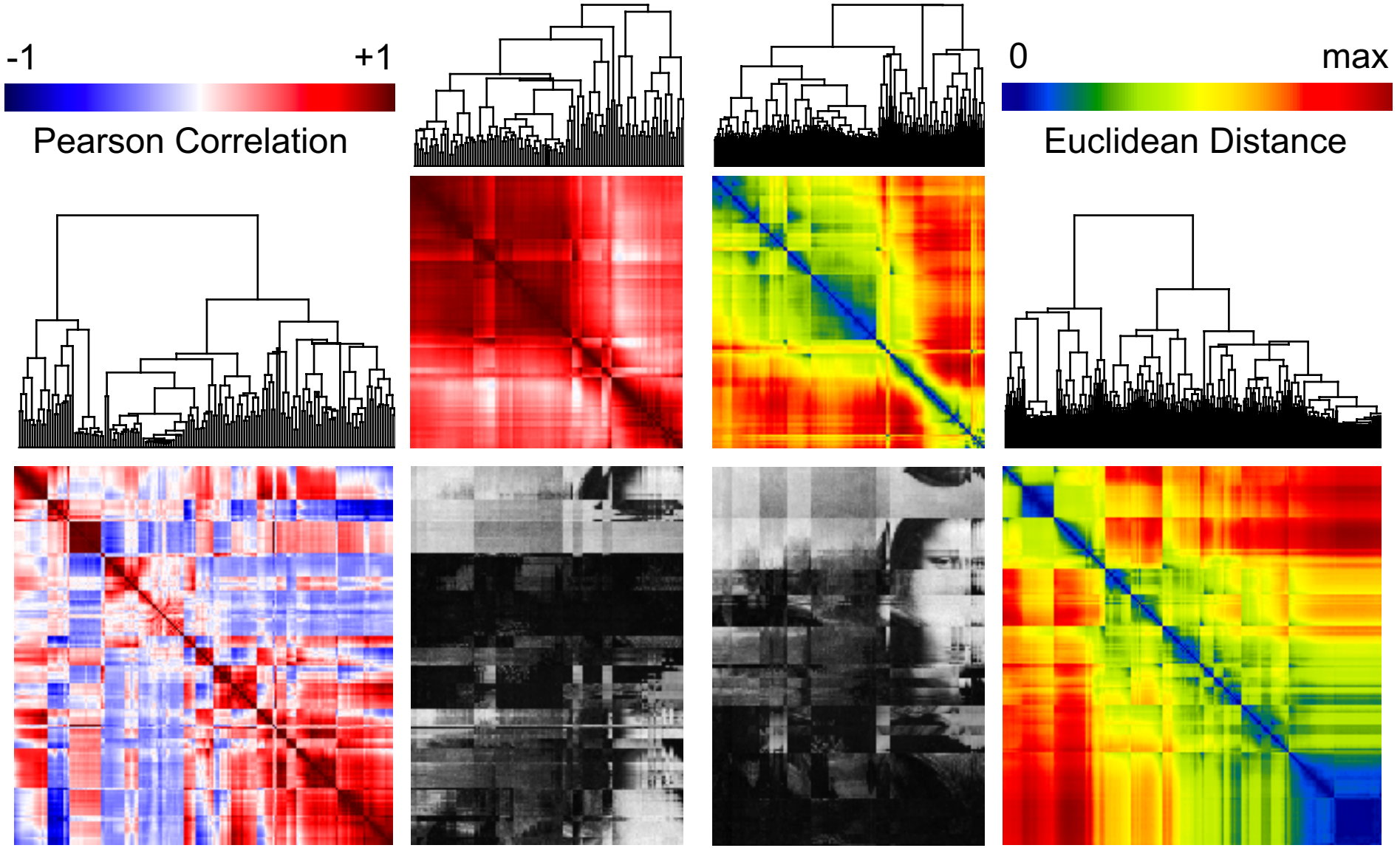
Rank 2 Elliptical Seriation row



Rank 2 Elliptical Seriation row/col




-1
+1
Pearson Correlation
0
max
Euclidean Distance
8
196
Grey Level
Single Linkage by Uncle



Centroid Linkage by R2E

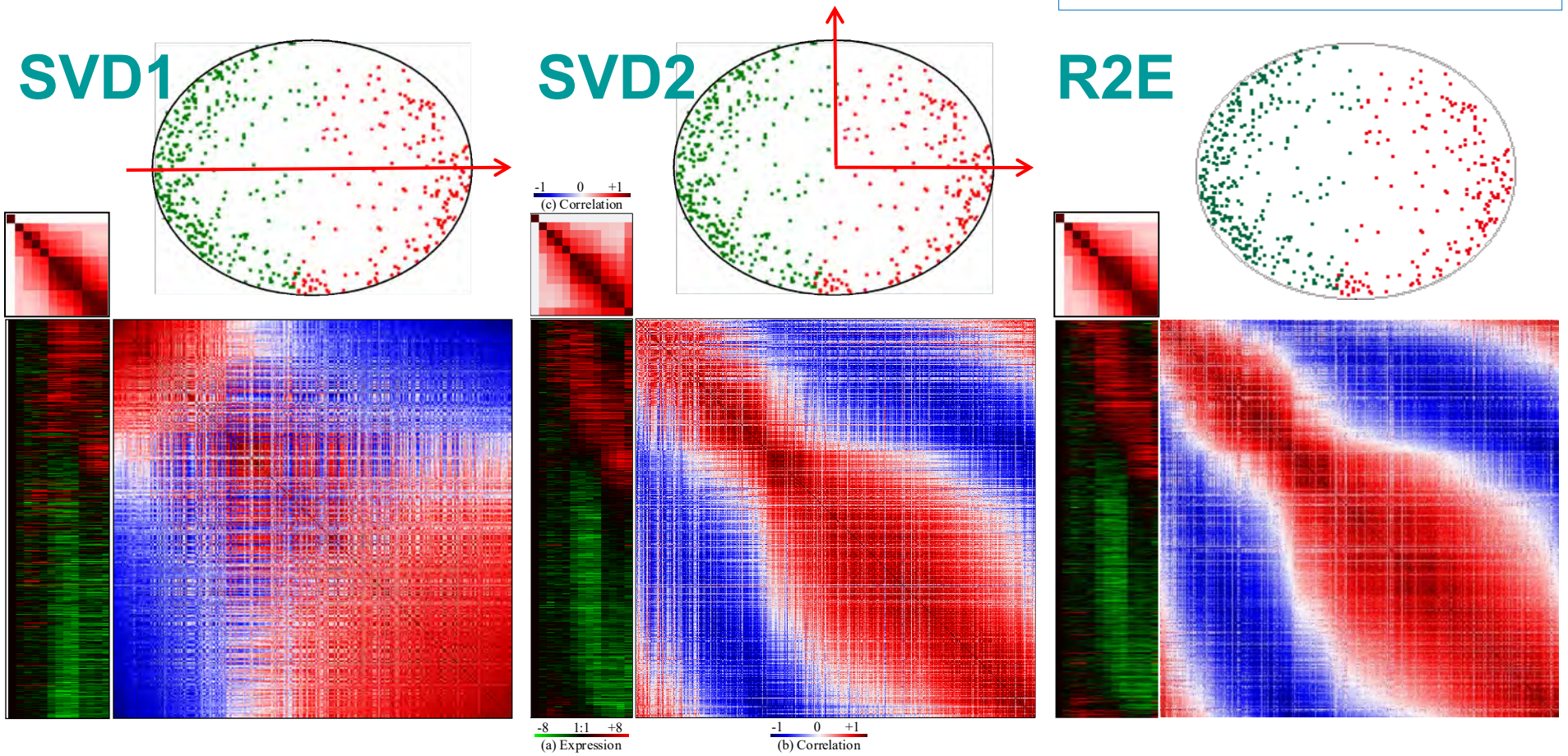
Statistical Approach

Identify Global Trend: Singular Value Decomposition

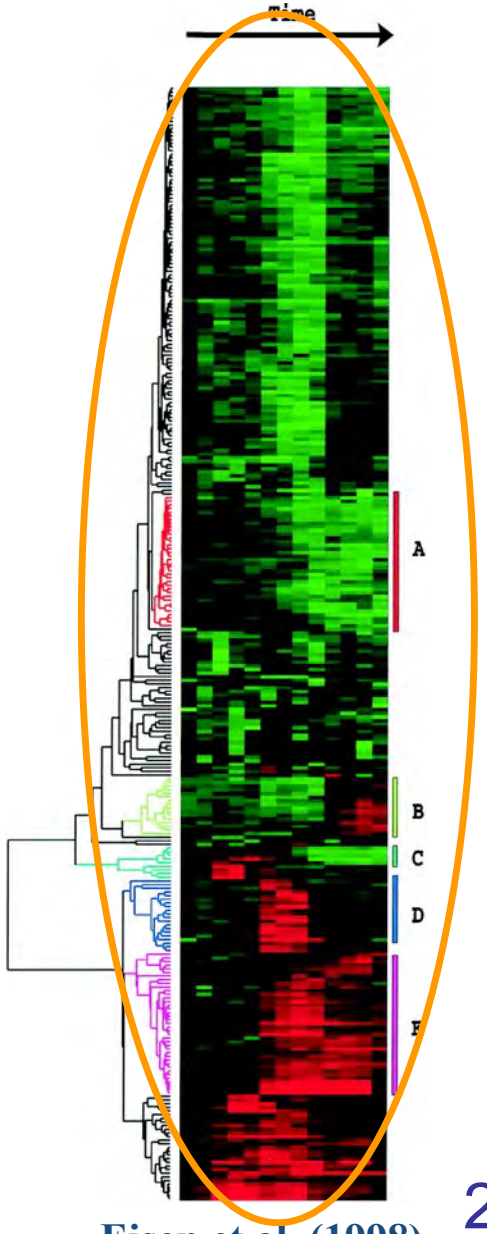
SVD

Alter O. et al 2000, PNAS

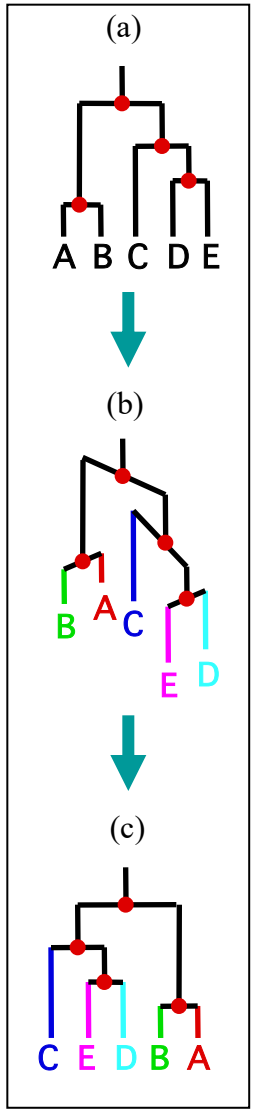
Chen 2002, Statistica Sinica Rank 2 Elliptical



Statistical Approach:
Identify Local Clusters



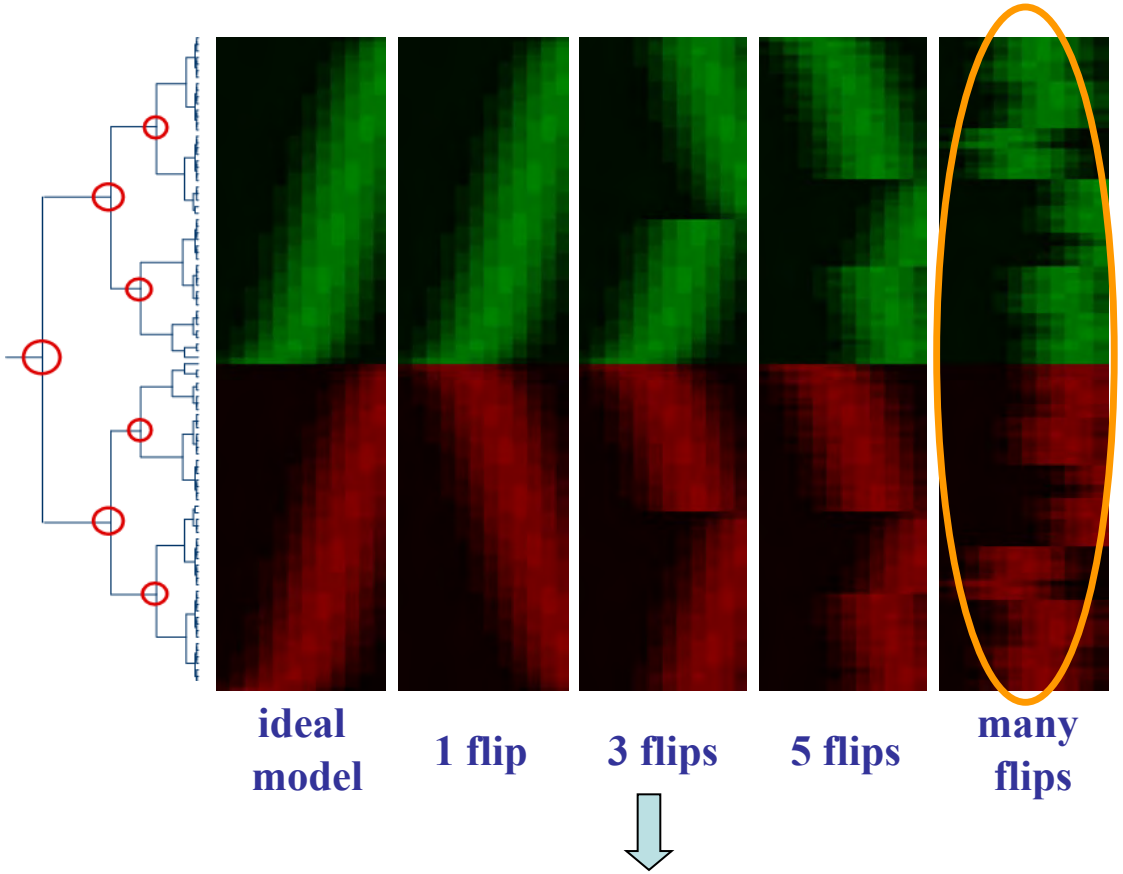
Eisen et al. (1998)



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

Tree seriation & flipping of intermediate nodes

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



ideal model | 1 flip | 3 flips | 5 flips | many flips

external and internal references for guiding flipping mechanism

Approaching Statistics & Statistical Approach

BMC Bioinformatics

BioMed Central

Methodology article

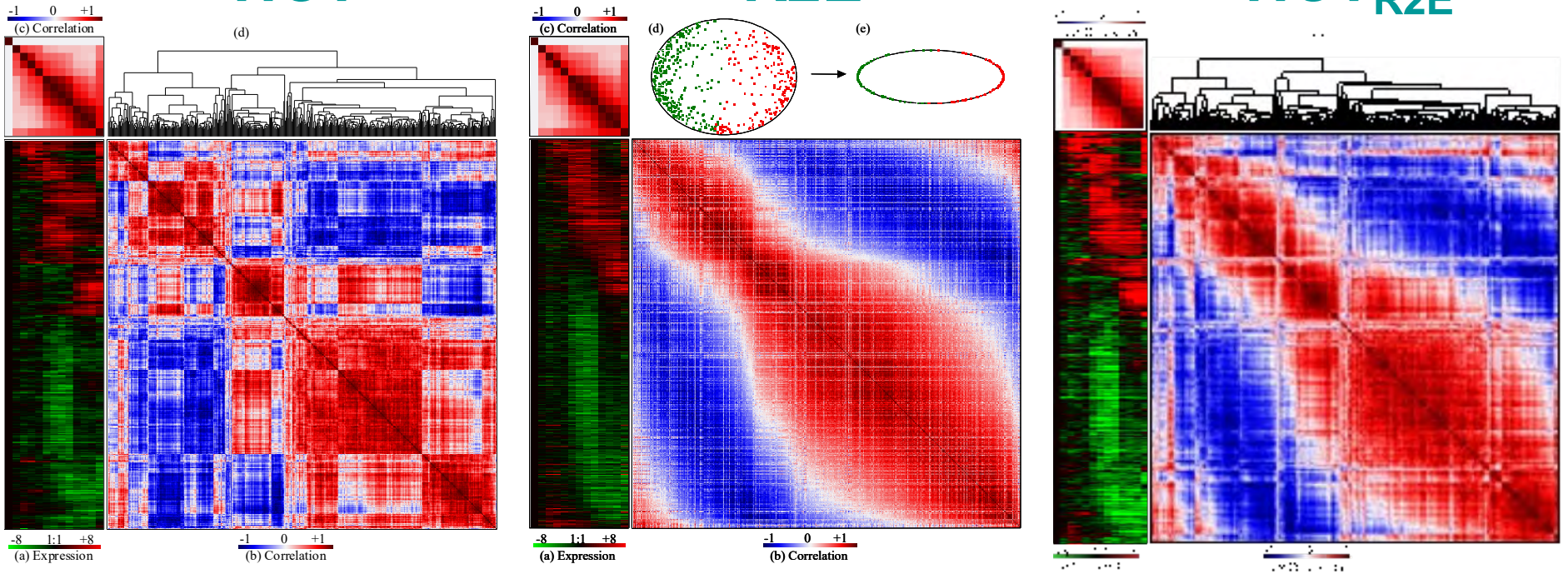
Open Access

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

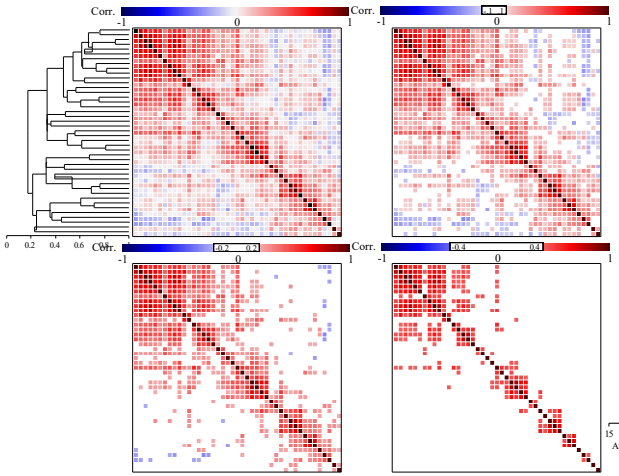
HCT + **R2E** = **HCT$_{R2E}$**



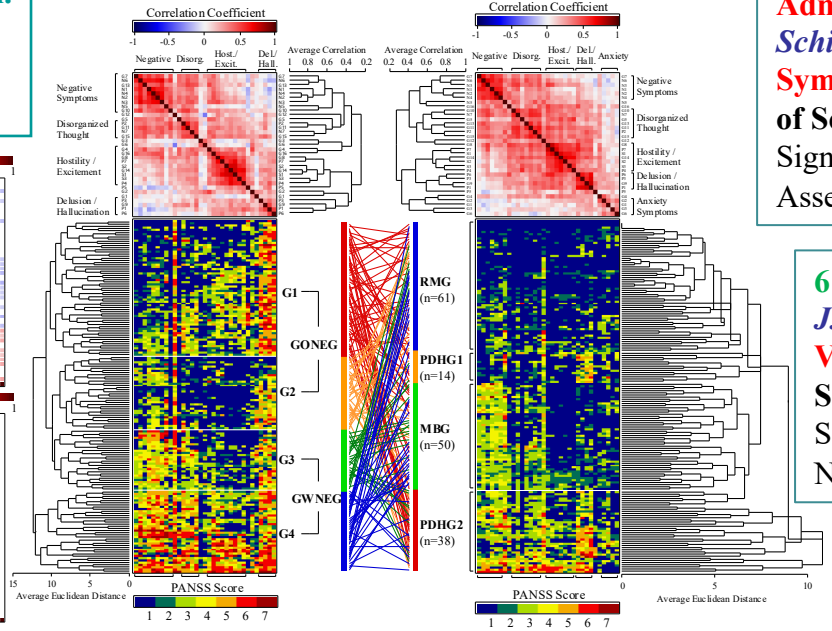
Hierarchical Tree Seriation **GAP Elliptical (R2E) Seriation** **Tree guided by (R2E)**

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 6 month

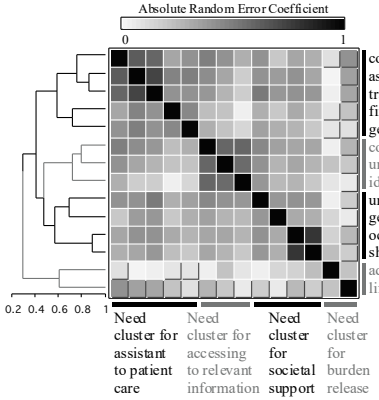


Admission **Hwu et al.**
Schizophrenia Research (2002)
Symptom Patterns and Subgrouping of Schizophrenic Patients: Significance of Negative Symptoms Assessed on Admission

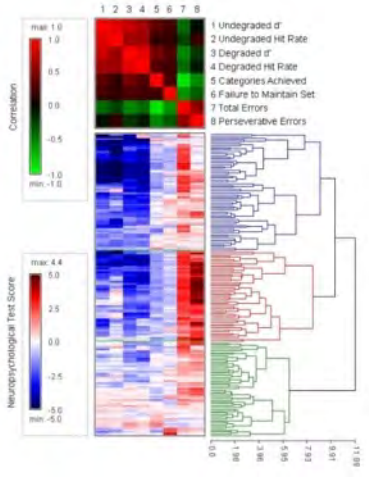
6 month **Liu et al.**
J. of the Formosan Med. Ass.
Validity of a 3-Subtype Model of Schizophrenia: Symptomatology, Social Function, and Neuropsychological Impairment

Schizophrenia Research (2013) **Liu et al.**
Development of a brief self-report questionnaire for screening putative pre-psychotic states.

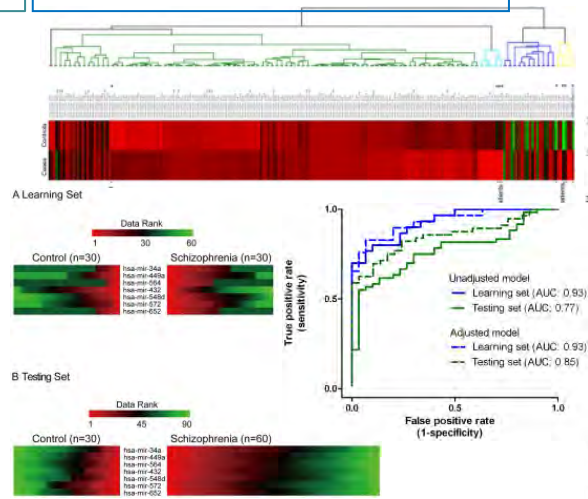
J. of the Formosan Med. Ass.(2008) **Yeh et al.** Factors Related to Perceived Needs of Chief Caregivers of Patients with Schizophrenia

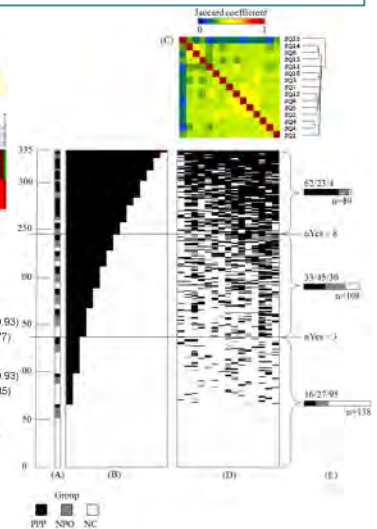


Genes, Brain and Behavior (2009) **Lin et al.** Clustering by neurocognition for fine-mapping of the schizophrenia susceptibility loci on chromosome 6p



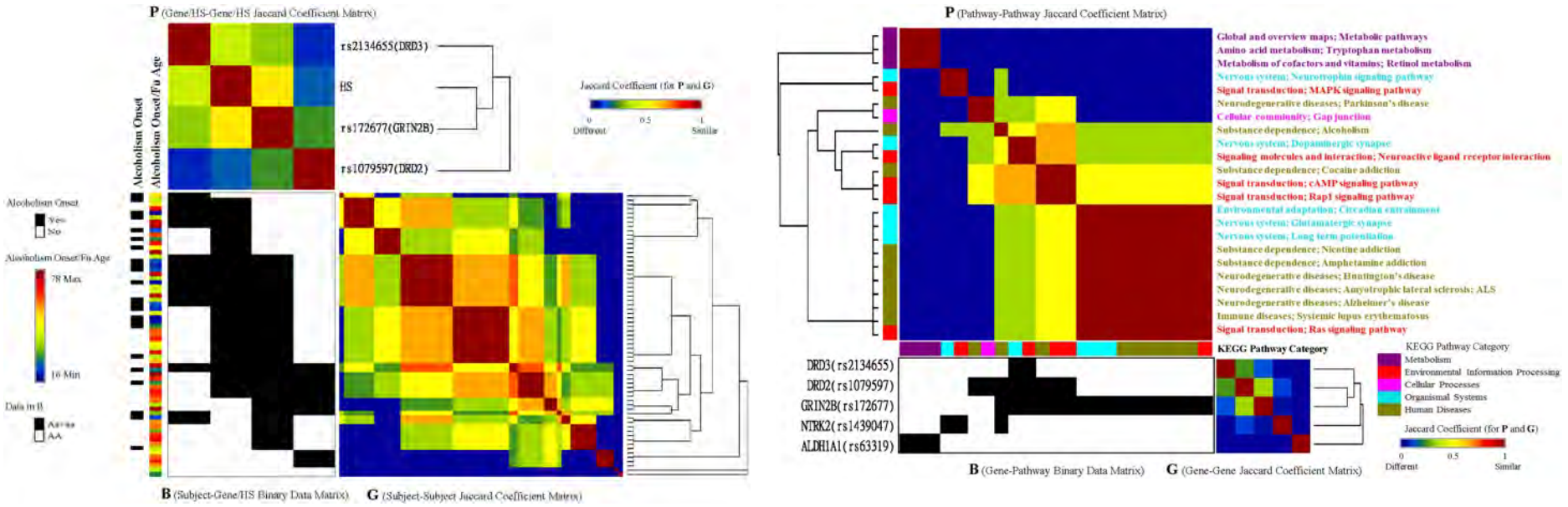
PLoS ONE (2011) **Lai et al.**
MicroRNA expression aberration as potential peripheral blood biomarkers for schizophrenia





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

Scientific Reports (2017) Yang, Chen et al.
Using an Event-History with Risk-Free Model to Study the Genetics of Alcoholism

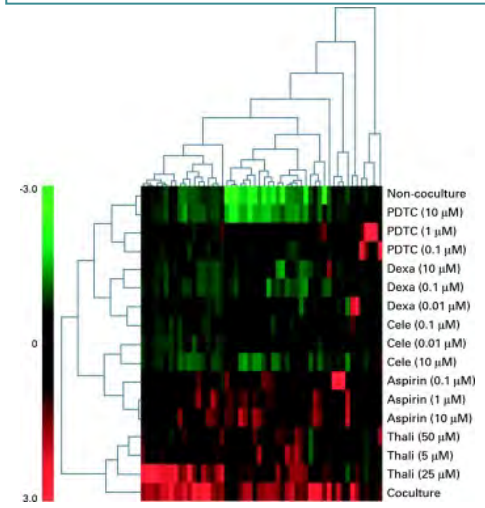


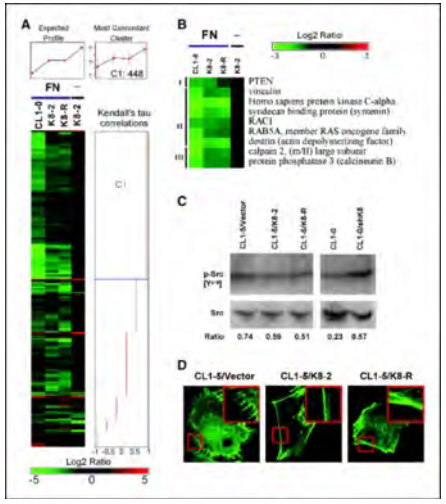
GAP for Cancer Study: Non–Small Cell Lung Cancer (National Taiwan University)

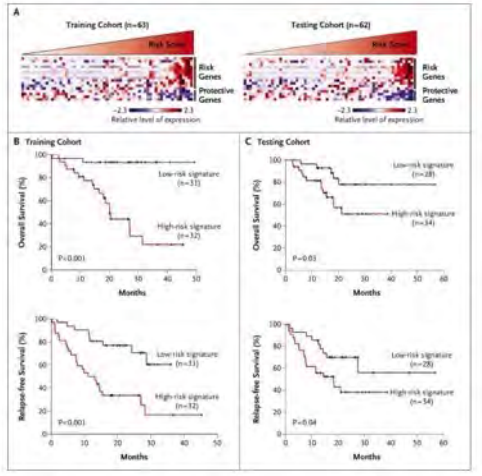
Journal of Clinical Oncology **23** (2005) Tumor-Associated Macrophages in Cancer Progression **Chen J. J. et al.**

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.**

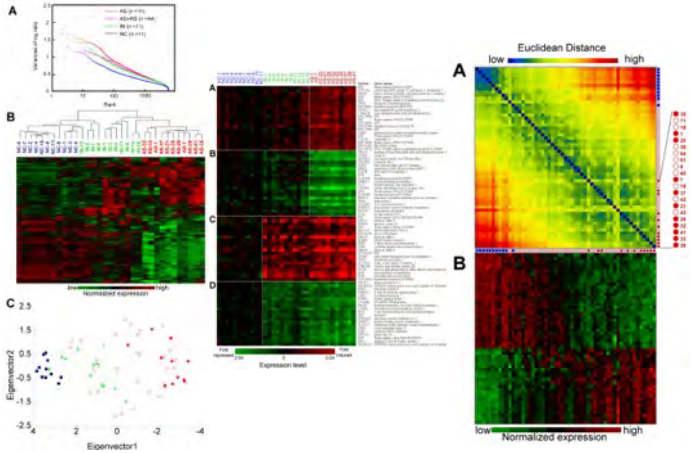






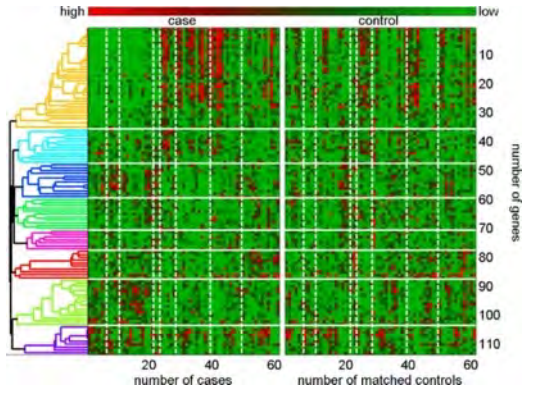
GAP for Infectious Disease: SARS

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

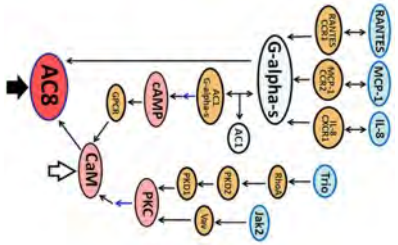
Genetic Epidemiology **30** (2006) Using endophenotypes for pathway clusters to map complex disease genes **Pan W. H. et al. (Academia Sinica)**

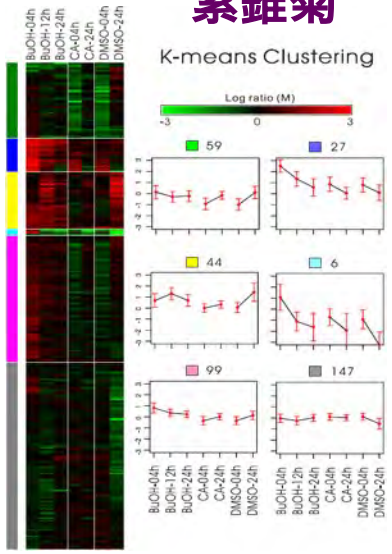


GAP for Comparative Metabolome: Chinese Herbal Medicine

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.





紫錐菊

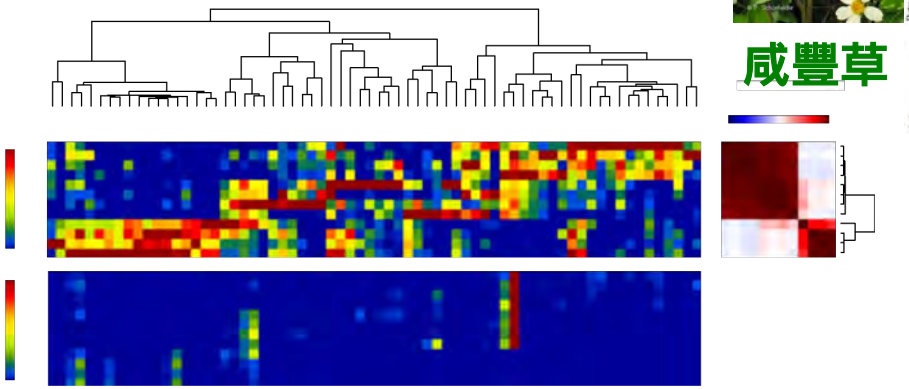
***Journal of Nutritional Biochemistry* 21** (2010) Comparative metabolomics approach coupled with cell- and gene-based assays for species classification and anti-inflammatory bioactivity validation of Echinacea plants
Hou et al.

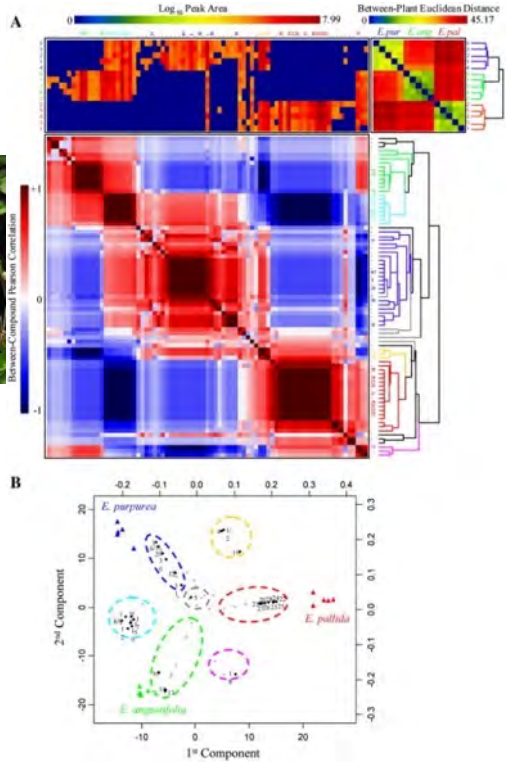
***BMC Complementary and Alternative Medicine* 13** (2013) Morus alba and active compound oxyresveratrol exert anti-inflammatory activity via inhibition of leukocyte migration involving MEK/ERK signaling.
Chen et al.

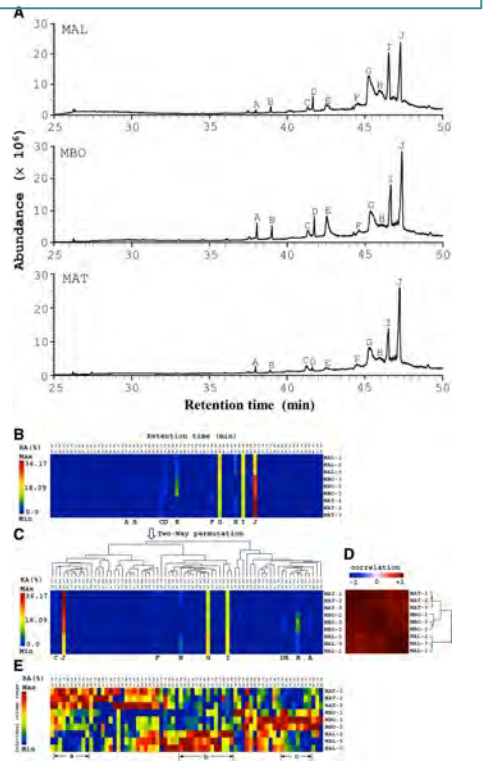
白桑

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

咸豐草

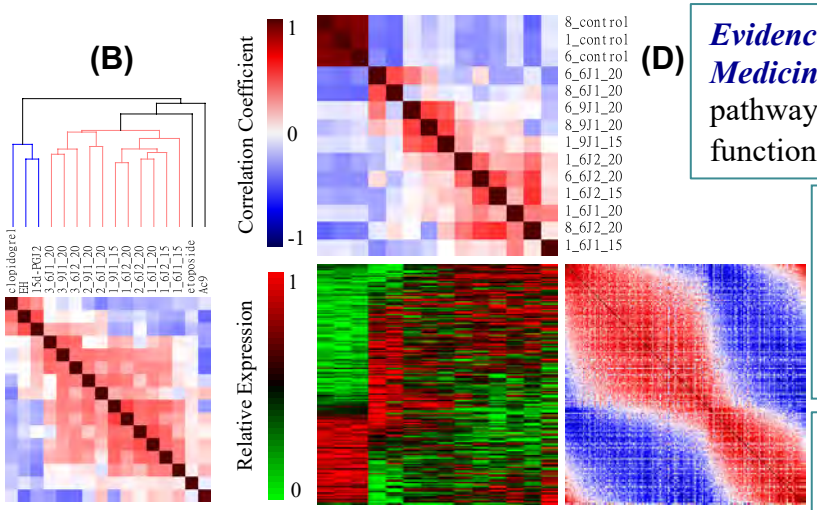






GAP for 1. Protein-Protien interaction in cross-talk pathways; 2. Therapeutic drug screening; 3. Chinese Herbal Medicine

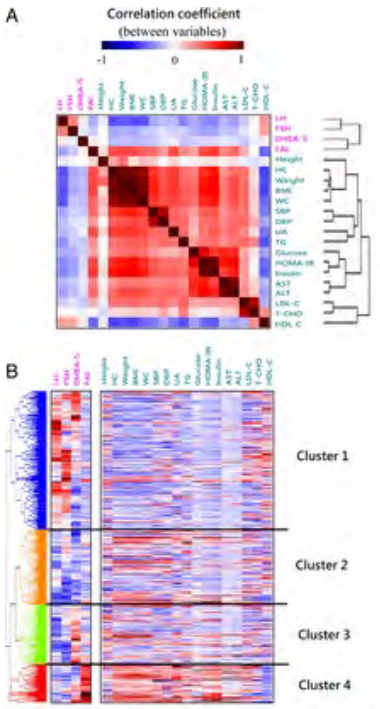
C-Y F. Huang, Nat'l Yang-Ming Univ.

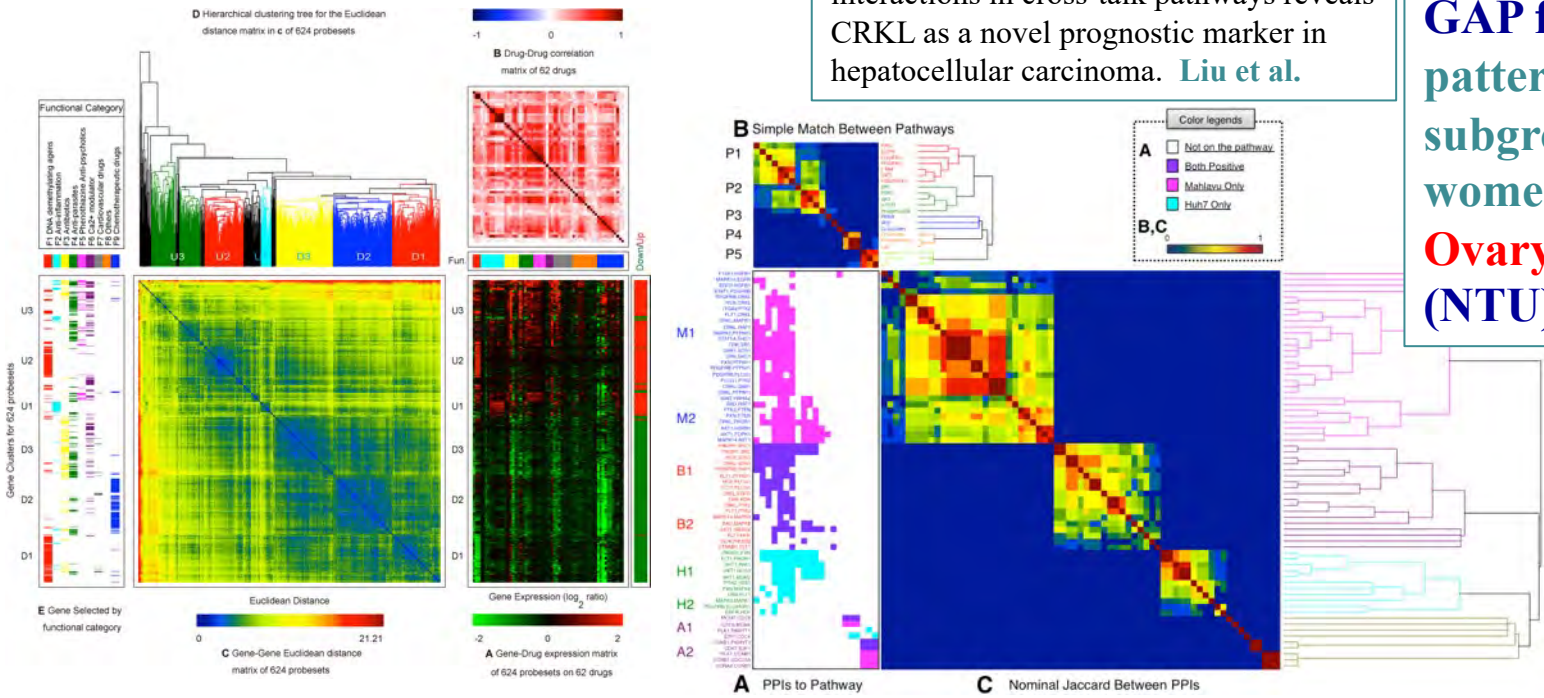


Evidence-Based Complementary and Alternative Medicine (eCAM) (2015) Gene expression profiling and pathway network analysis predicts a novel anti-tumor function for a botanical-derived drug, PG2. **Kuo et al.**

Open Access Scientific Reports **1** (2012) In silico Therapeutic Drug Screening for Reversing the Lung Adenocarcinoma Overexpressed Gene Signatures. **Kuo Y. L. et al. (Nat'l Yang-Ming Univ.)**

Molecular and Cellular Proteomics **12** (2013) An analysis of protein-protein interactions in cross-talk pathways reveals CRKL as a novel prognostic marker in hepatocellular carcinoma. **Liu et al.**





GAP for Symptom patterns & phenotypic subgrouping of women with Polycystic Ovary Syndrome (NTU)

Human Reproduction ***30*** (2015) Symptom patterns and phenotypic subgrouping of women with polycystic ovary syndrome: Association between endocrine characteristics and metabolic aberrations. Huang et al.

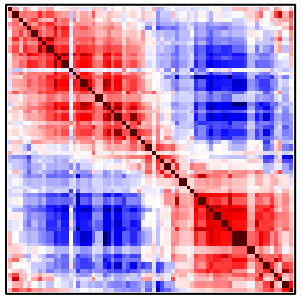
Matrix visualization of binary data (GAP approach)

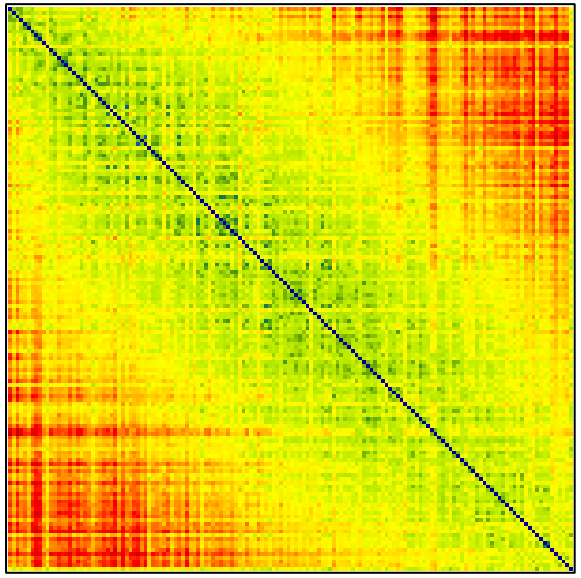
Essential elements in a GAP MV procedure?

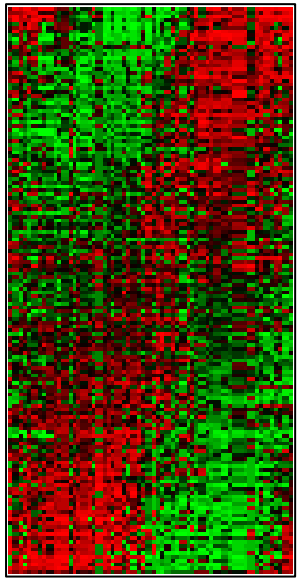
Continuous

Binary

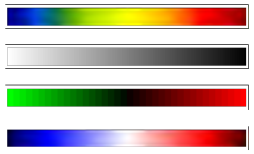
Correlation
Covariance
. . .







Euclidean Distance
Manhattan Distance
Correlation …



3. Variable Proximity

?

Object A

Object B

	1	0	
1	a	b	$(a+b)$
0	c	d	$(c+d)$
	$(a+c)$	$(b+d)$	$(a+b+c+d)$

1. Color Coding

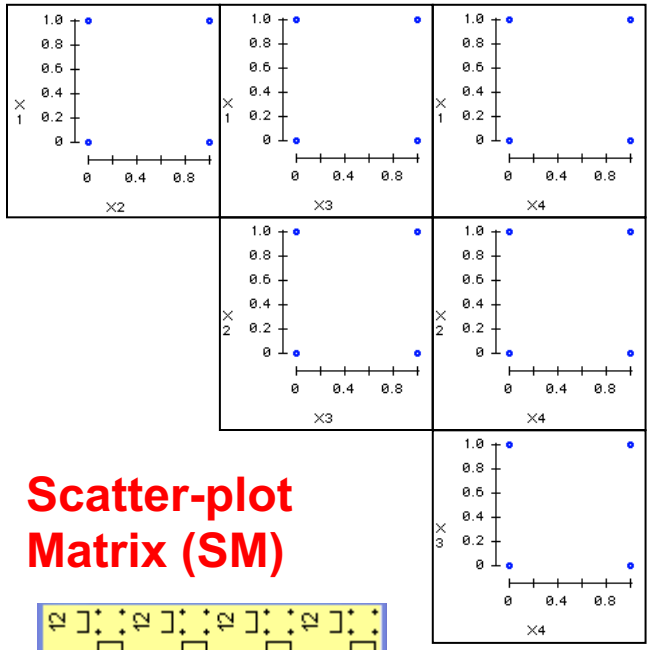
?

2. Subject Proximity

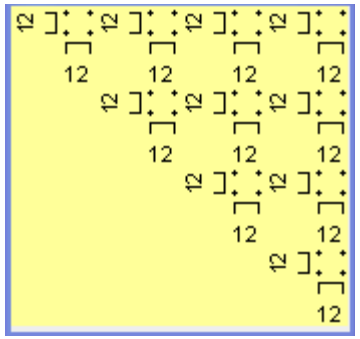
?

Why Matrix Visualization (binary)?

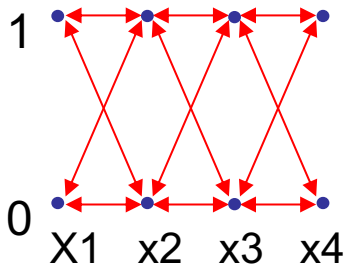
Graphic tools for high-dimensional non continuous data visualization w/o dimension reduction

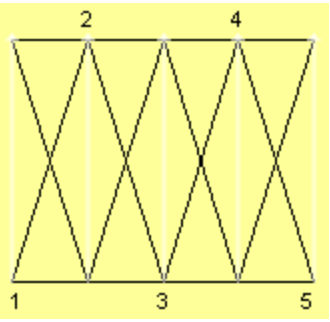


Scatter-plot Matrix (SM)

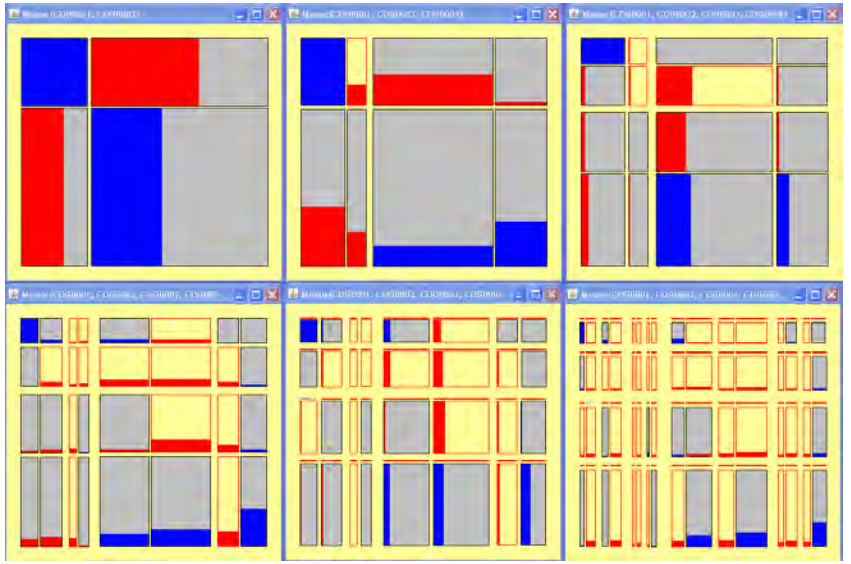


Parallel Coordinates Plot (PCP)





Mosaic Plot (Display)



Tzeng et al. (BMEI 2009) (*IEEE Xplore Digital Library*)

Commonly used similarity coefficients for binary data

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

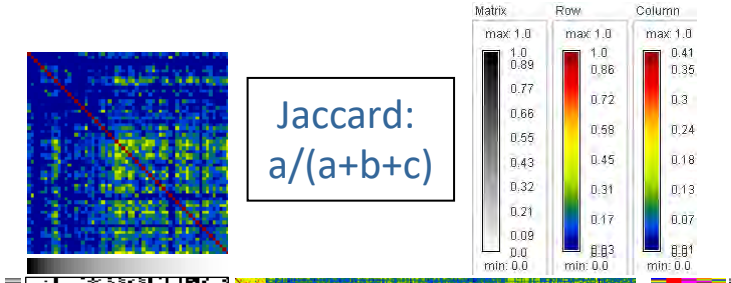
Similarity	Formula
Braun	$\frac{a}{\max(a+b,\ a+c)}$
Dice	$\frac{2a}{2a+b+c}$
Hamman	$\frac{a+d-(b+c)}{a+b+c+d}$
Jaccard	$\frac{a}{a+b+c}$
Kappa	$\left(1+\frac{(b+c)(a+b+c+d)}{2ad-2bc}\right)^{-1}$
Kulczynskl	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
Ochiai	$\frac{a}{\sqrt{((a+b)(a+c))}}$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

Identification of Salmonella Serotypes with Pulsed-Field Gel Electrophoresis (PFGE) Fingerprints

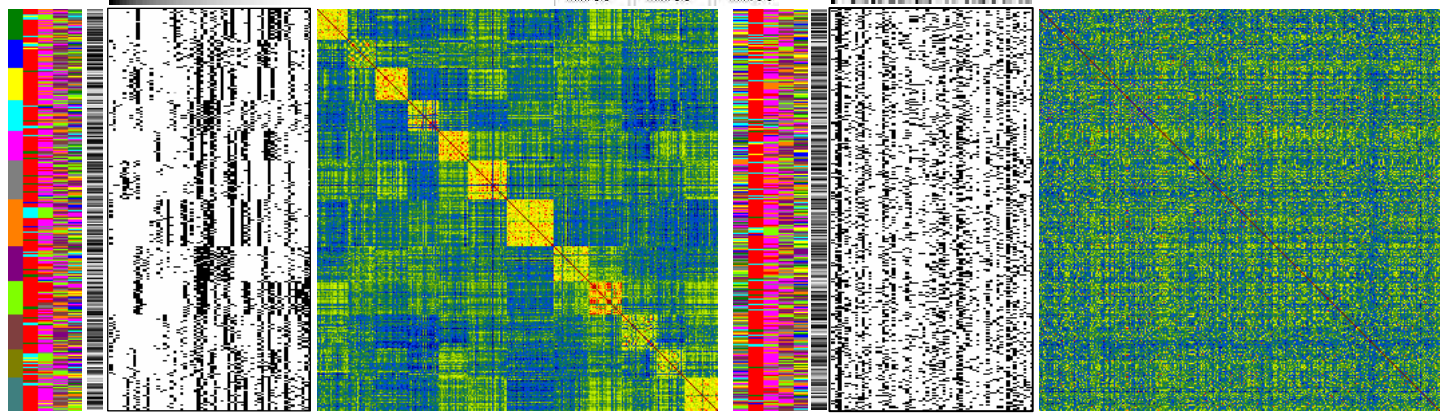
12 serotypes

Anatum
Bareilly
Berta
Derby
Hartford
Litchfield
Mbandaka
Panama
Paratyphi
Schwarze
Senftenbe
Stanley

Jaccard: a/(a+b+c)

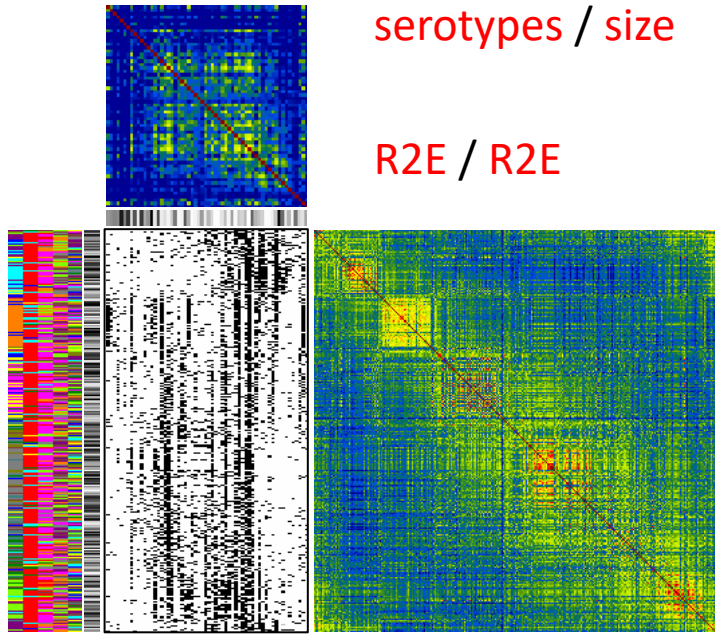


Wen Zou, NCTR 1390 patterns from 12 less frequent serotypes. CDC PulseNet from 2005~2010. 60 different sizes of bands (PFGE: Pulsed-field gel electrophoresis)



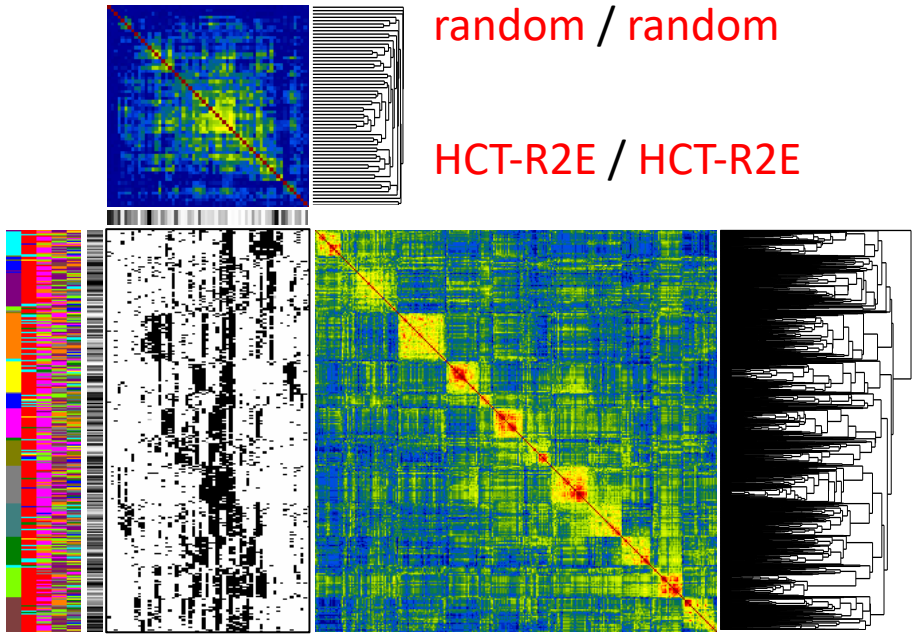
serotypes / size

R2E / R2E



random / random

HCT-R2E / HCT-R2E



Matrix visualization of nominal data

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 students (a university in Nishinomia, Japan)

	鱷	Alligator		狗	Dog		豹	Leopard
	熊	Bear		鴨	Duck		獅	Lion
	駱駝	Camel		象	Elephant		蜥蜴	Lizard
	貓	Cat		狐狸	Fox		鴕鳥	Ostrich
	獵豹	Cheetah		青蛙	Frog		豬	Pig
	雞	Chicken		長頸鹿	Giraffe		鴿	Pigeon
	乳牛	Cow		山羊	Goat		兔	Rabbit
	鶴	Crane		鷹	Hawk		浣熊	Racoon
	黑猩猩	Chimpanzee		河馬	Hippopotamus		犀牛	Rhinoceros
	烏鴉	Crow		馬	Horse		蛇	Snake

	麻雀	Sparrow
	虎	Tiger
	龜	Tortoise
	火雞	Turkey
	猴	Monkey

A zoo keeper


企鵝館
濕地生態池
溫馨花園
鸚鵡屋
兩棲爬蟲館
鳥園站
鳥園廣場
河馬廣場
光平隧道
水鳥觀察區
生命紀念碑
保育大道
酷Cool節能屋
動物醫療中心
飼料調配中心
野生動物收容中心
野生動物檢疫救傷中心
穿山甲館
大貓熊館
行政大樓
教育中心
無尾熊館
列車總站
線手指教室
作業道路
保育廊道
昆蟲館
保育公園
公務門
出口 入口
捷運動物園站
貓空纜車動物園站

A typical nominal data

Shizuhiko Nishisato, 2006

Classification of Animals

35 animals were sorted into piles of similar animals by 15 students (a university in Nishinomia, Japan)

How to proceed with data visualization ?

How to see data ?

What about

3500 samples?

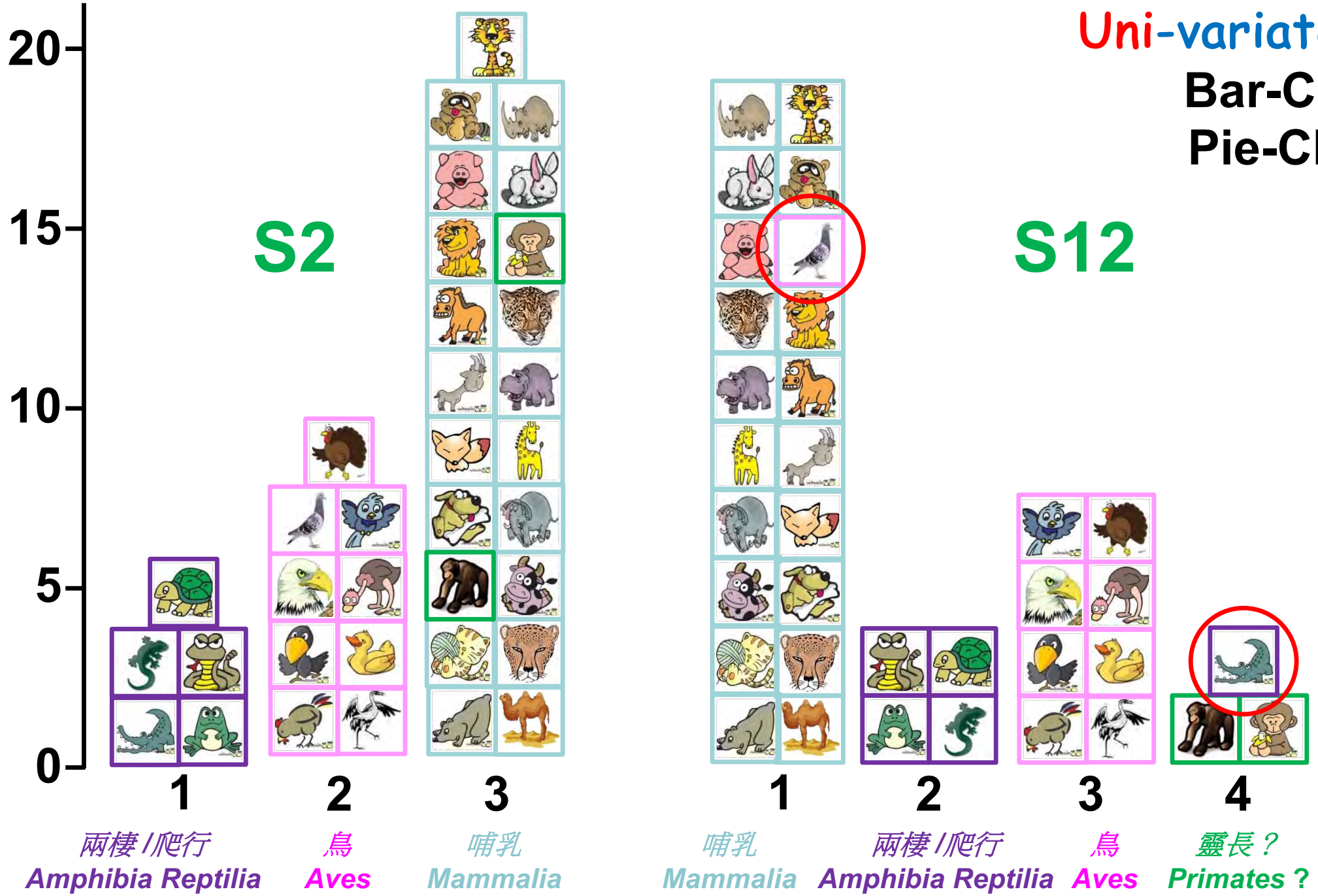
1500 variables?

動物 \ 學生		S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
鱷魚	Alligator	8	1	6	9	6	1	3	4	1	4	3	4	3	6	4
熊	Bear	6	3	2	6	6	1	3	4	4	5	5	1	4	2	2
駱駝	Camel	4	3	9	3	1	4	3	5	4	2	5	1	7	7	8
貓	Cat	6	3	7	4	0	1	1	2	3	3	1	1	6	3	5
獵豹	Cheetah	3	3	7	4	0	1	3	5	4	3	6	1	6	2	2
雞	Chiken	7	2	4	1	2	5	7	1	5	1	1	3	8	4	1
黑猩猩	Chimpanzee	5	3	5	7	5	2	4	4	2	2	6	4	3	6	6
乳牛	Cow	1	3	9	6	1	1	3	5	3	4	1	1	4	5	8
鶴	Crane	7	2	4	5	2	5	5	1	5	1	2	3	8	4	1
烏鴉	Crow	7	2	4	5	2	5	5	1	5	1	2	3	8	4	1
狗	Dog	6	3	7	10	0	2	1	2	3	3	1	1	4	3	5
鴨	Duck	7	2	4	1	2	5	5	1	5	1	2	3	8	4	1
象	Elephant	4	3	6	3	1	4	3	5	4	5	3	1	7	7	2
狐狸	Fox	6	3	7	4	0	1	6	2	3	3	3	1	4	3	5
青蛙	Frog	8	1	3	2	3	3	2	3	1	4	4	2	1	1	3
長頸鹿	Giraffe	1	3	8	3	1	4	3	5	4	2	5	1	7	7	8
山羊	Goat	3	3	9	6	1	4	6	5	3	3	1	1	5	3	5
鷹	Hawk	7	2	4	5	2	5	5	1	5	1	3	3	8	4	1
河馬	Hippopotamus	4	3	6	6	6	4	3	3	4	4	5	1	7	7	2
馬	Horse	6	3	9	6	1	2	3	5	3	3	1	1	5	5	8
豹	Leopard	1	3	7	4	0	1	3	5	4	3	3	1	6	2	2
獅	Lion	5	3	7	4	6	1	3	5	4	3	3	1	7	2	2
蜥蜴	Lizard	2	1	3	2	3	3	2	3	1	4	4	2	2	1	3
猴	Monkey	6	3	5	7	5	2	4	4	2	2	6	4	3	6	6
鴕鳥	Ostrich	3	2	4	1	2	5	3	1	5	1	5	3	8	7	8
豬	Pig	1	3	9	6	1	1	6	5	3	3	1	1	5	5	5
鴿	Pigeon	7	2	4	5	2	5	5	1	5	1	2	1	8	4	1
兔子	Rabbit	6	3	1	6	0	4	6	2	3	3	1	1	5	3	5
浣熊	Racoon	6	3	7	10	4	1	6	2	3	3	3	1	4	3	5
犀牛	Rhinoceros	4	3	5	6	6	4	3	5	4	4	5	1	7	7	2
蛇	Snake	8	1	3	9	6	3	2	3	1	4	4	2	2	1	3
麻雀	Sparrow	7	2	4	5	2	5	5	1	5	2	2	3	8	4	1
虎	Tiger	5	3	7	4	0	1	3	5	4	3	3	1	6	2	2
龜	Tortoise	8	1	3	9	3	3	2	3	1	5	4	2	1	1	3
火雞	Turkey	7	2	4	1	2	5	7	1	5	1	1	3	8	4	1

Uni-variate Display

Bar-Chart

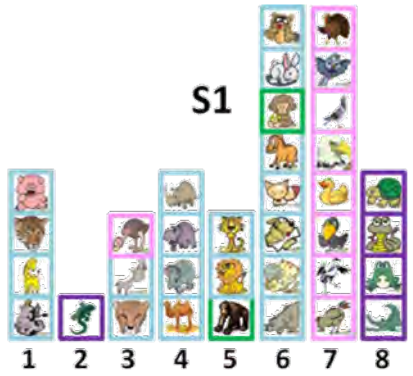
Pie-Chart

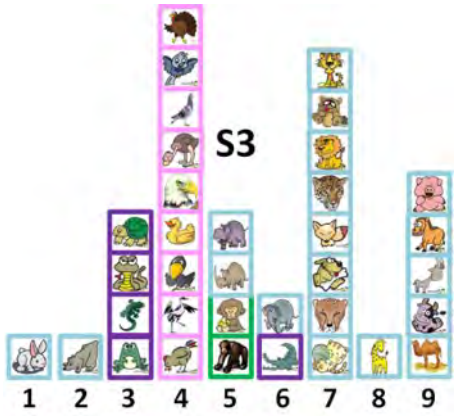


15 students

15 categorical variables

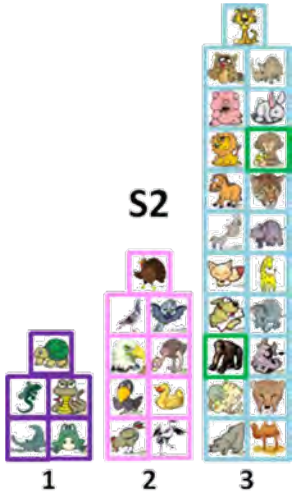
15 bar-charts





How to visualize the association between two categorical variables ?

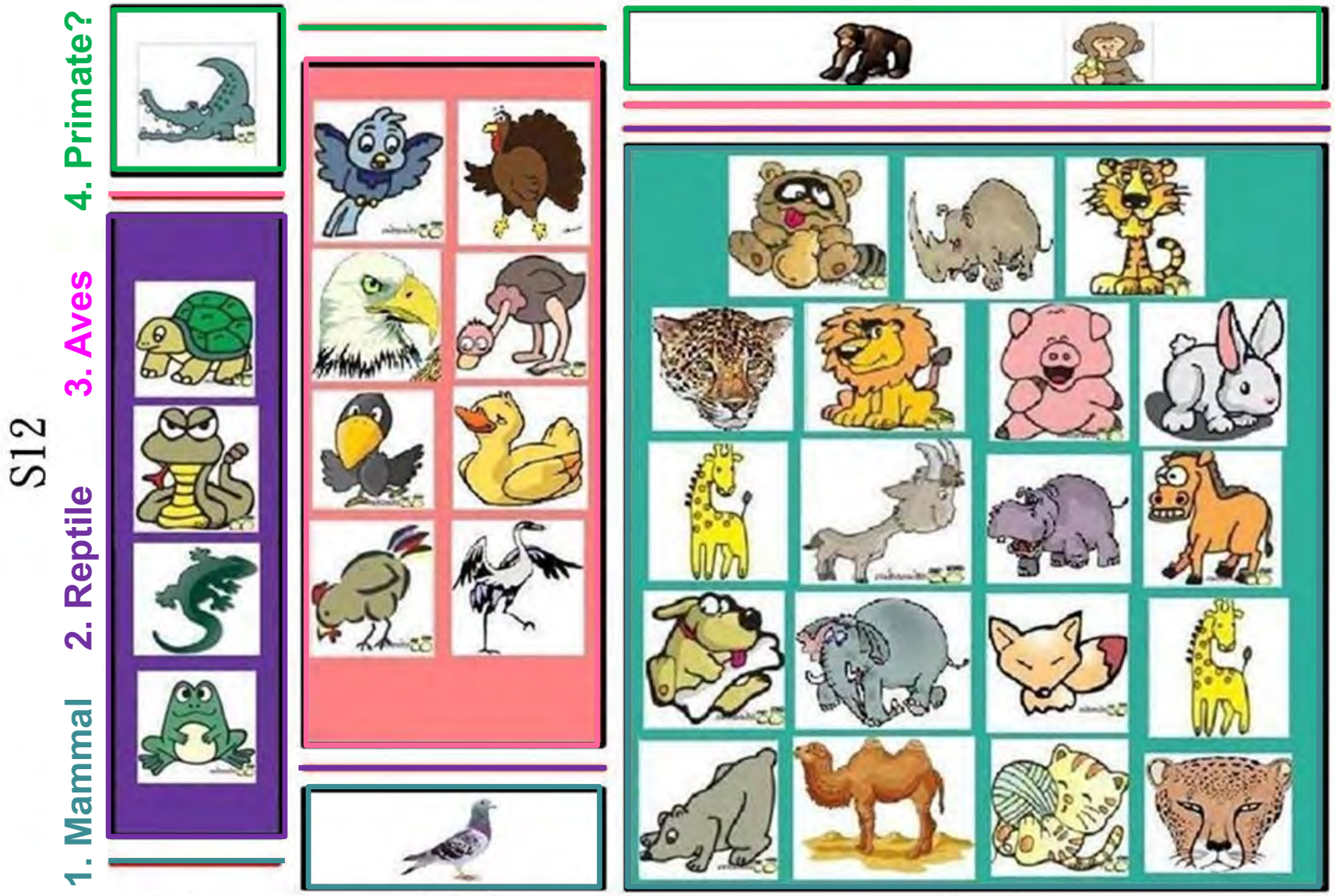
S2 (3)



S12 (4)

Bi-variate Display

Mosaic Display

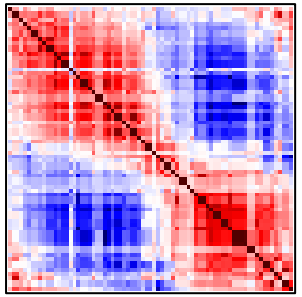


Some essential elements in a GAP MV procedure

Continuous

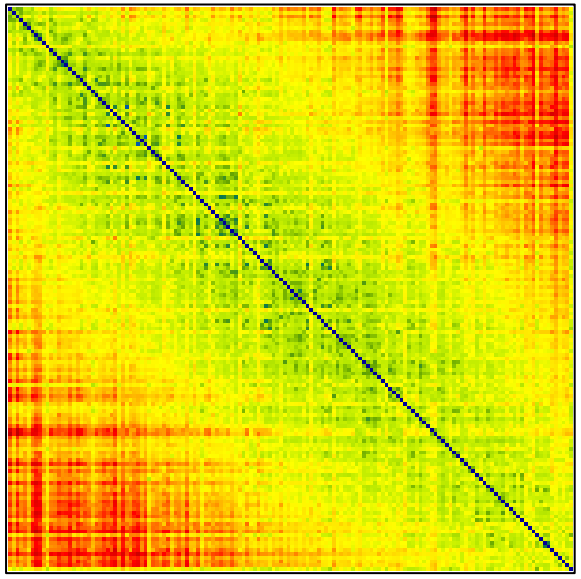
Nominal

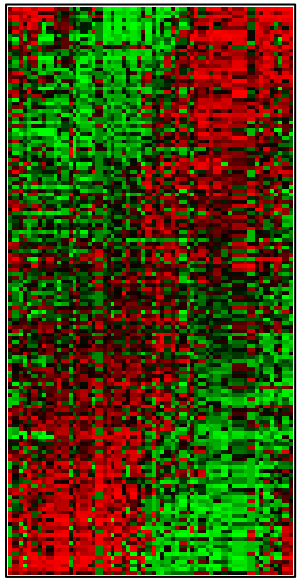
Correlation
Covariance
. . .



3. Variable Proximity

?





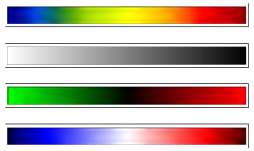
1. Color Coding

?

2. Subject Proximity

?

Euclidean Distance
Manhattan Distance
Correlation …



A0_Dog	B0_Rabbit	C0_Turkey	D0_Lion
A1_Alligator	B1_Frog	C1_Pig	D1_Elephant
A2_Chimpanzee	B2_Goat	C2_Crane	D2_Camel
A3_Cow	B3_Tiger	C3_Leopard	D3_Hawk
A4_Crow	B4_Rhinoceros	C4_Ostrich	D4_Fox
A5_Pigeon	B5_Giraffe	C5_Lizard	
A6_Cheetah	B6_Duck	C6_Horse	
A7_Chiken	B7_Sparrow	C7_Racoon	
A8_Bear	B8_Hippopotamus	C8_Tortoise	
A9_Cat	B9_Monkey	C9_Snake	

Matrix visualization of nominal data

Example:
Mushroom Data
UCI Machine Learning Repository

UCI Machine Learning Repository

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

Each species is identified as definitely **edible** (可以吃), definitely **poisonous** (有毒), or of **unknown** (不確定) edibility and not recommended. This latter class was combined with the poisonous one.

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)

Data Set Information:
This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525).

1. 菌傘形狀 cap shape
2. 菌傘表面 cap surface
3. 菌傘顏色 cap color
4. 瘀傷 bruises
5. 氣味 odor
6. 菌摺附著 gill attachment
7. 菌摺間距 gill spacing
8. 菌摺大小 gill size
9. 菌摺顏色 gill color
10. 柄形狀 stalk shape
11. 柄根 stalk root
12. 環上方柄表面 stalk surface abo
13. 環下方柄表面 stalk surface belo
14. 環上方柄顏色 stalk color above
15. 環下方柄顏色 stalk color below
16. 菌膜類型 veil type
17. 菌膜顏色 veil color
18. 環數 ring number
19. 環型 ring type
20. 孢印顏色 spore print color
21. 族群 Population
22. 棲息地 habitat

choice

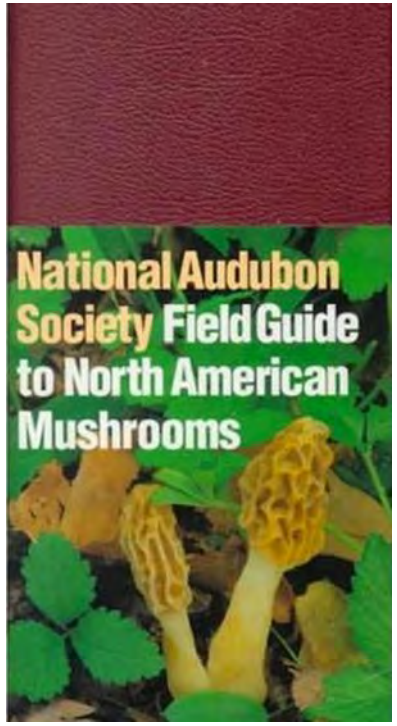
edible

deadly

deadly

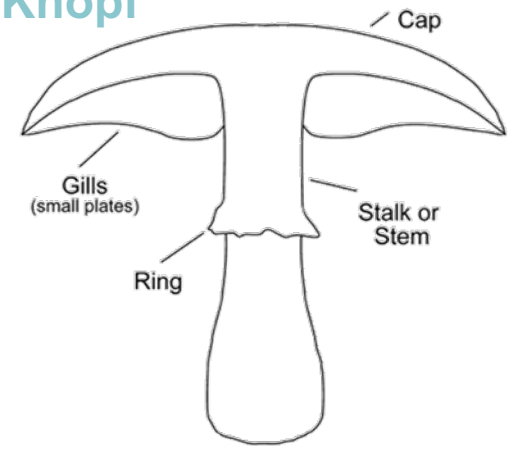
edible

choice



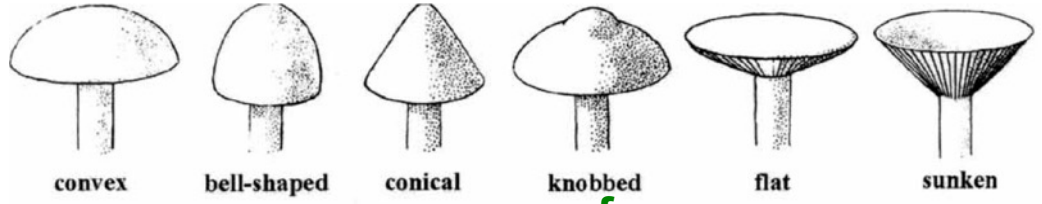
The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

http://www.english-online.at/biology/mushrooms/mushrooms-and-fungi.htm

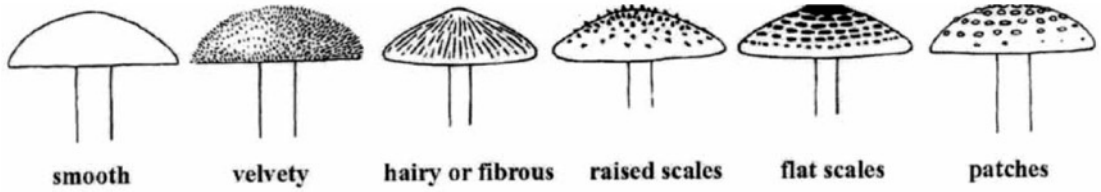


The Parts of a Mushroom

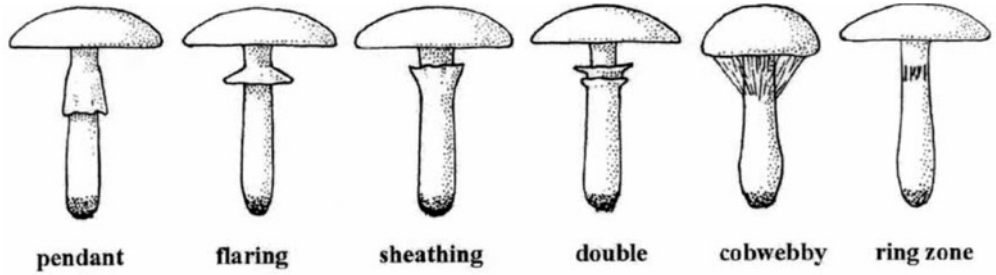
cap-shape



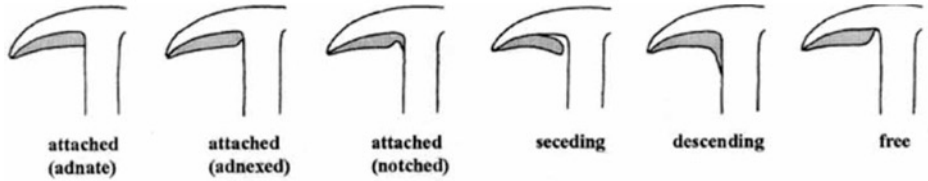
cap-surface



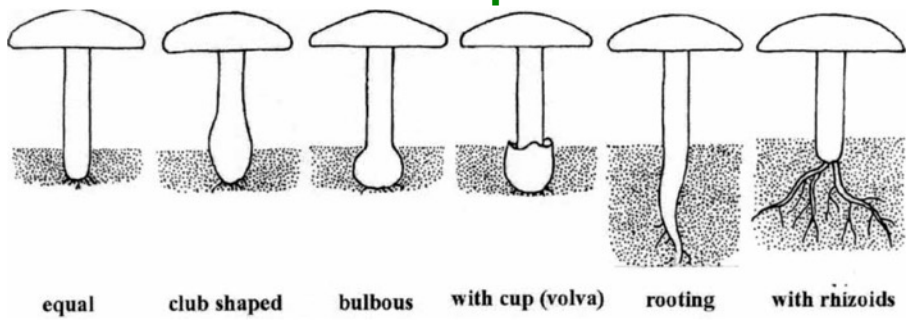
ring-type



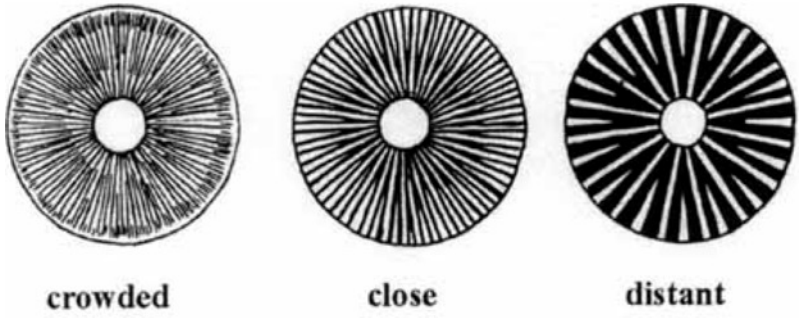
gill-attachment



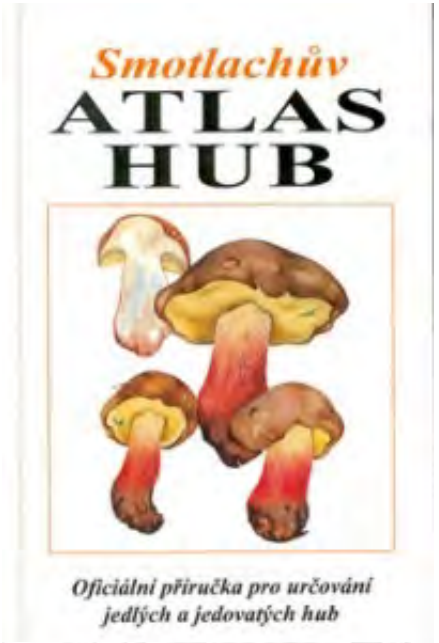
stalk-shape/root

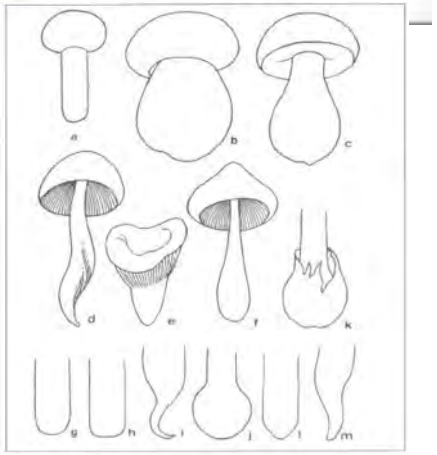


gill-spacing

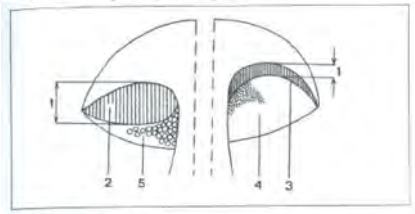


Kniha Smotlachův Atlas Hub (2002)



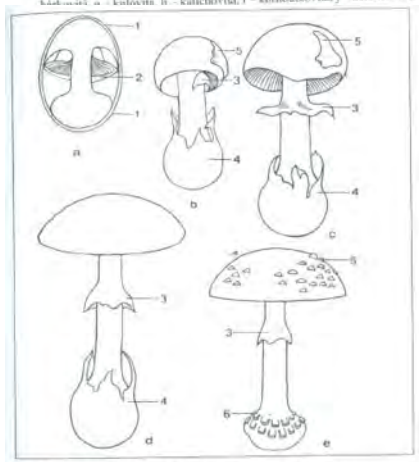


8. TYPY TŘENĚ
a – válcovitý, b – soudkovitý, c – hřebatý, d – vřetenovitý, e – kuželovitý, f – kyjovitý, ukončení báze třeně: g – zaokrouhlené, h – tupé, i – zahrocené, j – hlízovité, k – hlízovité s pochvou, l – kořenující, m – zúžené



5. ZNAKY TRUBEK
1 – výška trubek – trubky vysoké, vyklenuté ven, 3 – trubky nízké, vyklenuté dovnitř, 4 – drobná ústí trubek, 5 – široká ústí trubek

1. TYPY PLODNIC
a až d – kloboukaté plodnice (a – hřibovité houby, b – muchomůrky, c – smr. že, d – lupenaté houby), e až j – nekloboukaté plodnice (e – míškovité, f – po- houbovité, g – kyjovité, h – kalichovité, i – korkoutkovité, j – keříčkovité)



9. VÝVOJ PLODNICE MUCHOMŮRKY ZELENÉ (a až d)
a – průřez mladou plodnicí pokrytou ještě celkovou plachetkou, tzv. vajíčko, b, c – mladé plodnice, d – vzrostlá plodnice muchomůrky zelené, e – vzrostlá plodnice muchomůrky červené
1. celková plachetka, 2. částečná plachetka, 3. prsten (vzniká z částečné plachetky), 4. pochva (vzniká z celkové plachetky), 5. zbytky celkové plachetky z kloboučku, tzv. vločky nebo tečky, 6. zbytky celkové plachetky na bázi třeně muchomůrky červené, tzv. bradavky

BAREVNÁ ZOBRAZENÍ HUB
A POPISY ZOBRAZENÝCH DRU

Význam značek

Jedlost, jedovatost

delicious

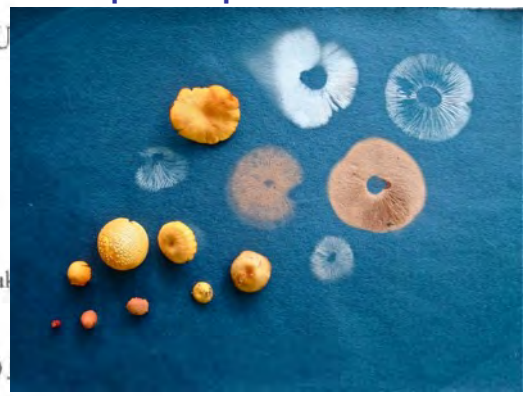
edible

deadly (smrtelně)

poisonous

not for food

spore-print-color



http://resurrectionfern.typepad.com/resurrection_fern/2009/08/not-all-mushrooms-are-alike.html

Místo růstu

Coniferous forest 針葉樹林

Broadleaf forest 闊葉樹林

houby rostoucí ve smíšených lesích

population / habitat

houby rostoucí na kmenech živých i odumřelých stromů, na trouchnivějícím dřevě apod.

houby rostoucí na loukách, pastvinách, v příkopech, podél cest, na polích apod. (mimo les)

month / season

1 cap shape
2 cap surface
3 cap color
4 bruises
5 odor
6 gill attachment
7 gill spacing
8 gill size
9 gill color
10 stalk shape
11 stalk root
12 stalk surface above ring
13 stalk surface below ring
14 stalk color above ring
15vstalk color below ring
16 veil type
17 veil color
18 ring number
19 ring type
20 spore print color
21 Population
22 habitat

Kumar P., et al. (2012) **A Benchmark to Select Data Mining Based Classification Algorithms for Business Intelligence and Decision Support Systems**.
IJDKP Vol.2, No.5.

Table1. Composition of Data Sets

S. No	Set	Total Size	Missing Value	Effective Size	Class	Total Attributes
1.	Mushroom	8124	2115	5609	2	23(Nominal valued)

Table 2. Predictive accuracy on M

Data Set / Classifier	Mushroom
CHAID	98.36 %
QUEST	86.57%
C4.5	96.0%
Neural N/W	100%
Logistic Regression	99.9%
k-means	85.29%
Genetic Algorithm	98%
SVM	100%
SVM-ABoostM1	100%

Table6. Comprehensibility c

Data Set / Classifier	Mushroom	
	Leaf Node	Depth
CHAID	5	2
QUEST	4	3
C4.5	13	5
Neural N/W	Nil	Nil
Logistic Regression	Nil	Nil
k-means	Nil	Nil
Genetic Algorithm	6	4
SVM	Nil	Nil
SVM-ABoostM1	Nil	Nil

Table3. Error Rate on Mushro

Data Set / Classifier	Mushroom
CHAID	1.6 %
QUEST	13.13%
C4.5	
Neura	
Logis Regre	
k-me	
Gene	
SVM	
SVM	

Table 4. Training Time on Mu

Data Set / Classifier	Mushroom
CHAID	60
QUEST	60
C4.5	60
Neural N/W	240
Logistic Regression	120
k-means	60

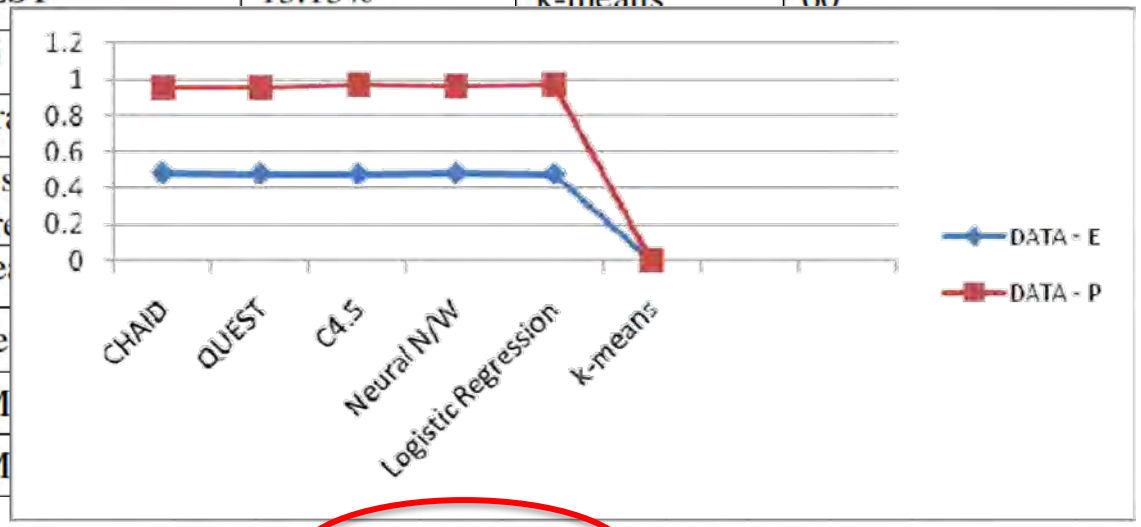


Figure4. Classification Index on Mushroom data set.

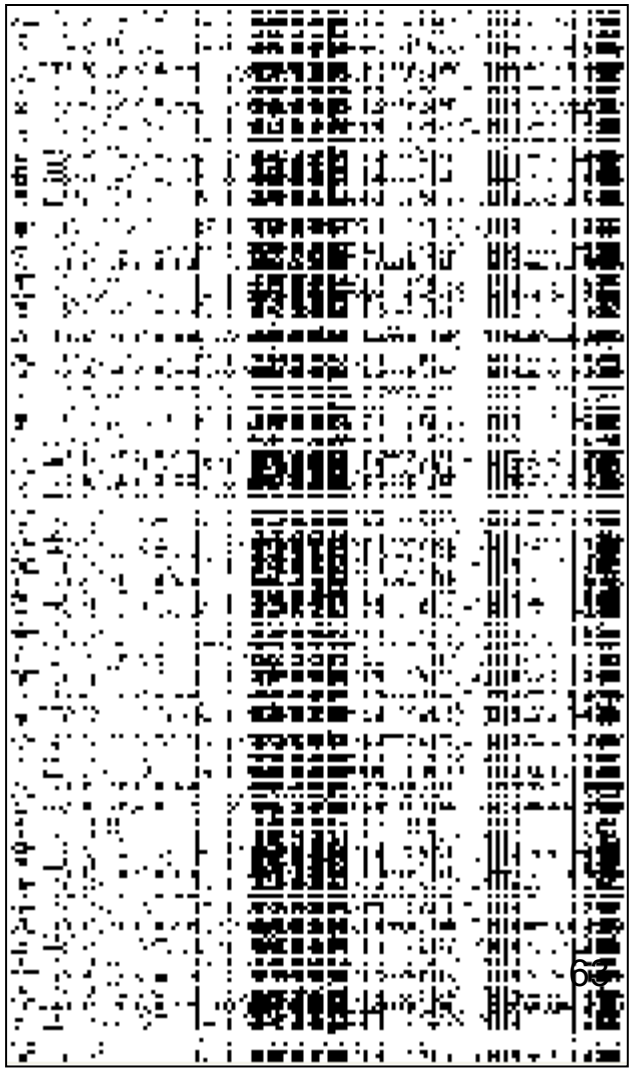
Matrix visualization for data with Cartography links

Example:
international organization membership pattern

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

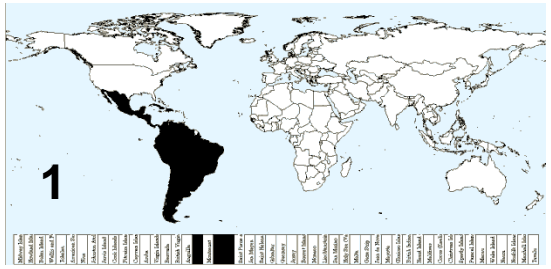


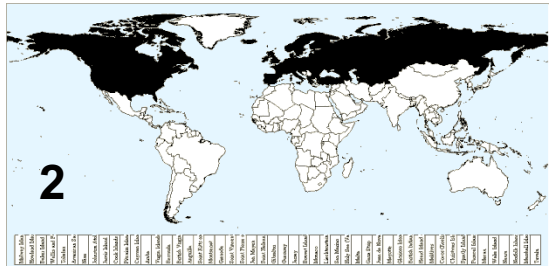
230 countries (regions)

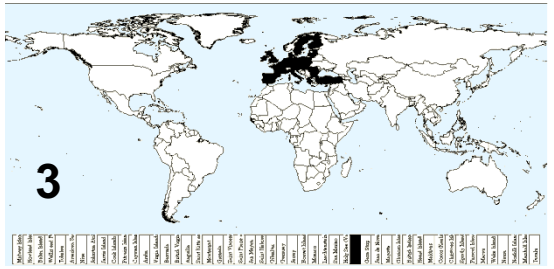
CIA Political Map of the World

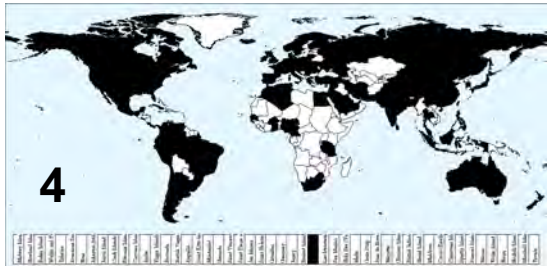
http://www.faqs.org/docs/factbook/index.html

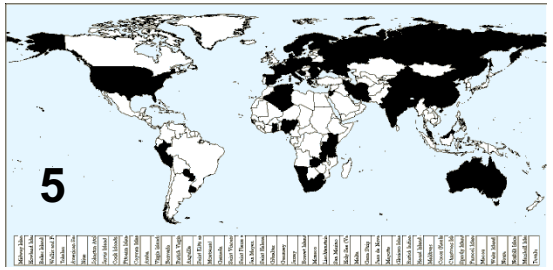
Draw one membership map for each organization (variable)?

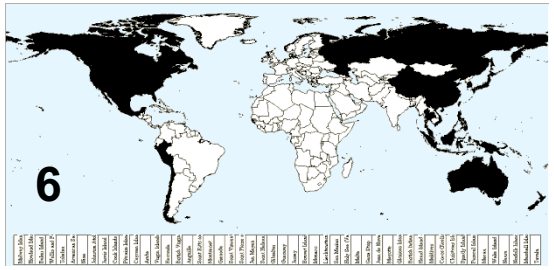


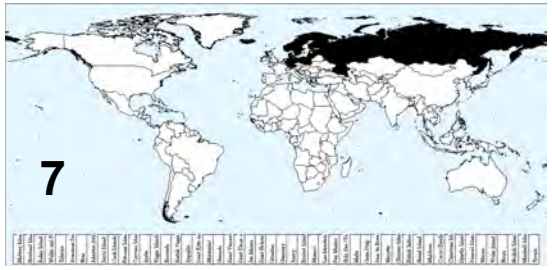


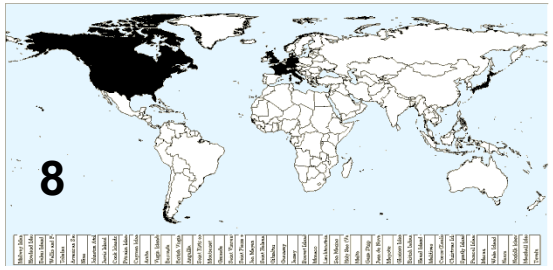


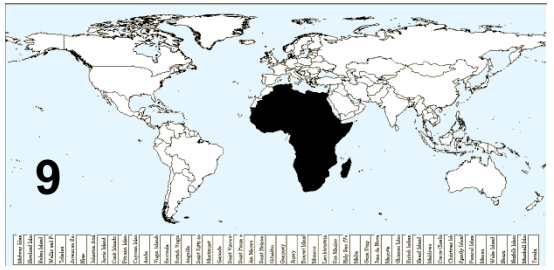




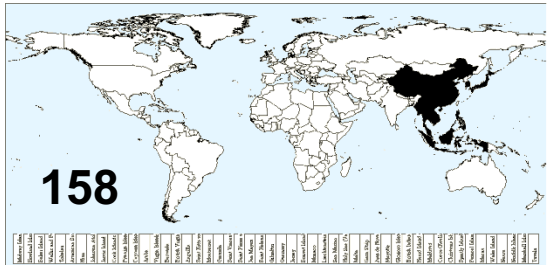


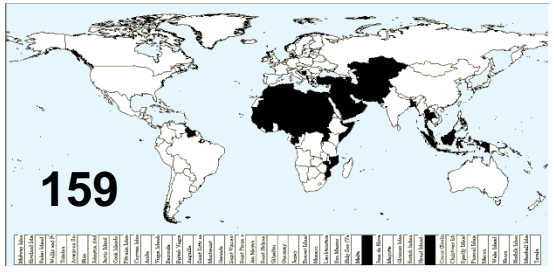


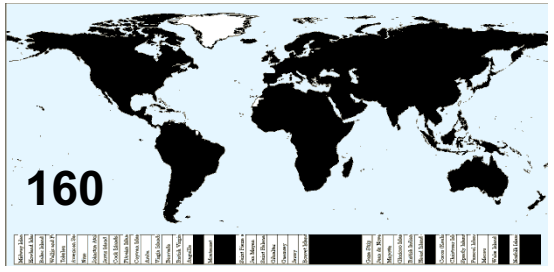


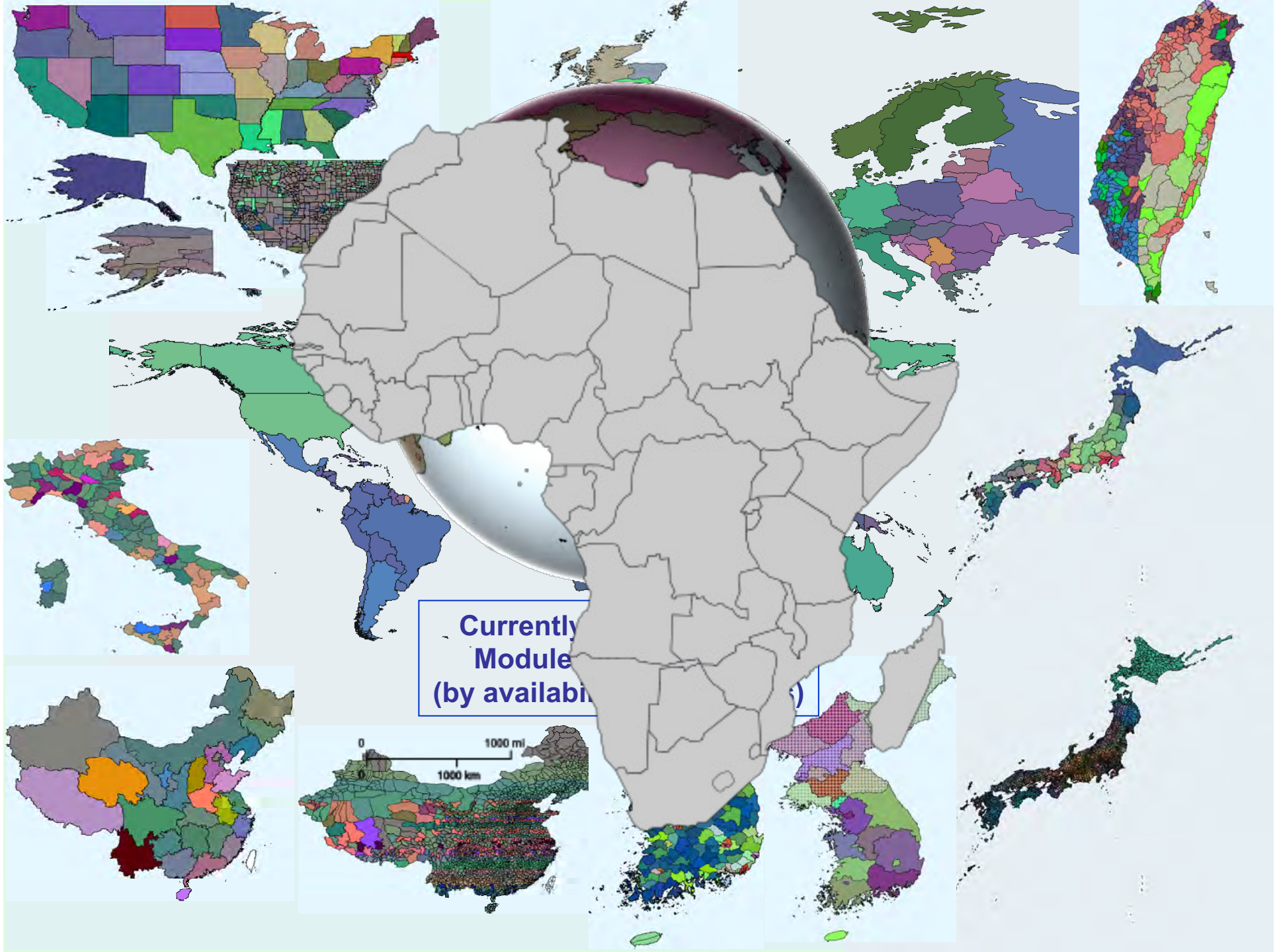


• • • 160 maps (?) • • •








Currently
Module
(by availabi
)
0
1000 mi
0
1000 km

Matrix visualization with Covariate adjustment

OXFORD ACADEMIC

Bioinformatics

Issues Advance articles Submit ▼ Purchase Alerts About ▼ All Bioinformatic

Article Contents

Covariate-adjusted heatmaps for visualizing biological data via correlation decomposition

Han-Ming Wu, Yin-Jing Tien, Meng-Ru Ho, Hai-Gwo Hwu, Wen-chang Lin, Mi-Hua Tao, Chun-Houh Chen ✉

Bioinformatics, bty335, https://doi.org/10.1093/bioinformatics/bty335

Published: 26 April 2018 **Article history** ▼

Matrix visualization with Covariate adjustment

Example:
Morphological measurements of Leptograpsus crabs

GAP with Covariate-adjustment using Correlation Decomposition

$$\text{total correlation } \mathbf{R} = \text{model component } \mathbf{M} + \text{residual component } \mathbf{E}$$

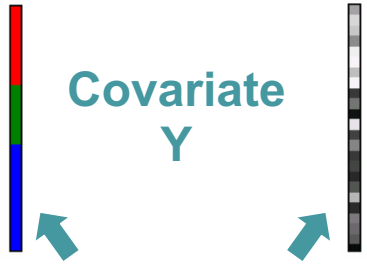
Discrete $\mathbf{R}_{ij} = \eta_i^B \, \eta_j^B \, \mathbf{R}_{ij}^B + \eta_i^W \, \eta_j^W \, \mathbf{R}_{ij}^W$

WABA (Within And Between Analysis)
(Dansereau et al.,1984)

Continuous $\mathbf{R}_{ij} = \sqrt{\mathbf{R}_{iy}^2}\sqrt{\mathbf{R}_{jy}^2}\mathbf{R}_{\pm 1} + \sqrt{1-\mathbf{R}_{iy}^2}\sqrt{1-\mathbf{R}_{jy}^2}\mathbf{R}_{ij\,;y}$

Partial Correlation
(Kurowicka, 2000)

group means $E[\mathbf{X}|Y] = \hat{\mathbf{X}}$



Covariate Y

simple linear regression

$$X_i = \beta_{0i} + \beta_{1i} Y + \epsilon_i$$

$$X_j = \beta_{0j} + \beta_{1j} Y + \epsilon_j$$

$\rho(E[X_i|Y], E[X_j|Y])$ $\quad$ $\rho(X_i - E[X_i|Y], X_j - E[X_j|Y])$

Correlations for columns

$\rho(X_i, X_j)$	$\rho(\hat{X}_i, \hat{X}_j)$	$\rho(\epsilon_i, \epsilon_j)$
$\mathbf{R}$	$\mathbf{R}^B$	$\mathbf{R}^W$

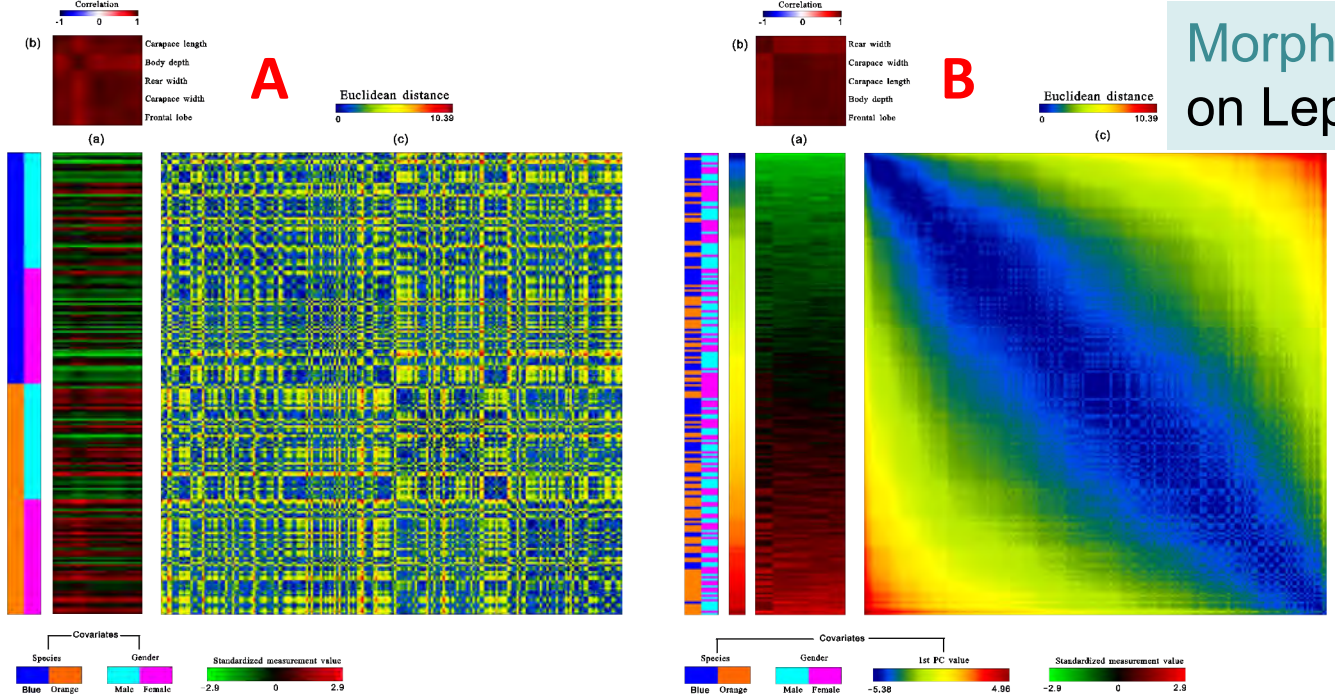
$$\mathbf{X} = \hat{\mathbf{X}} + \mathcal{E}$$

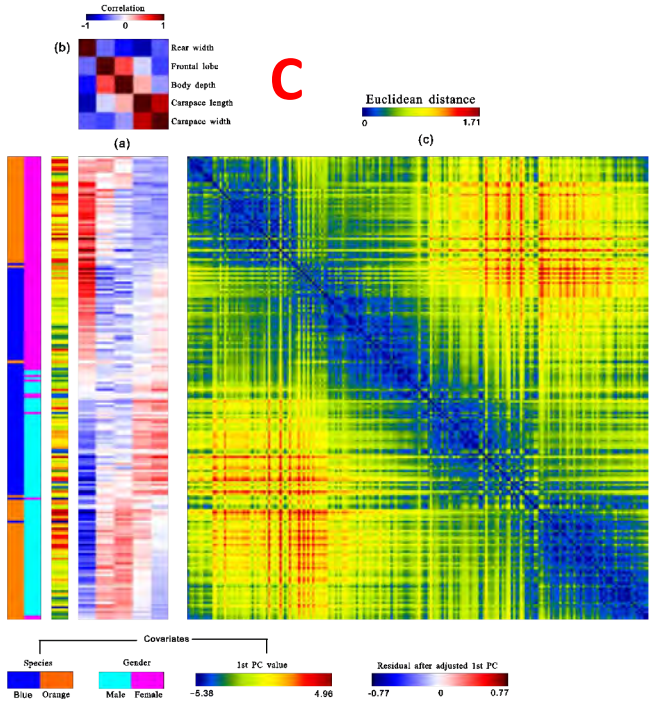
Raw Data $\quad$ Fitted Data $\quad$ Residuals

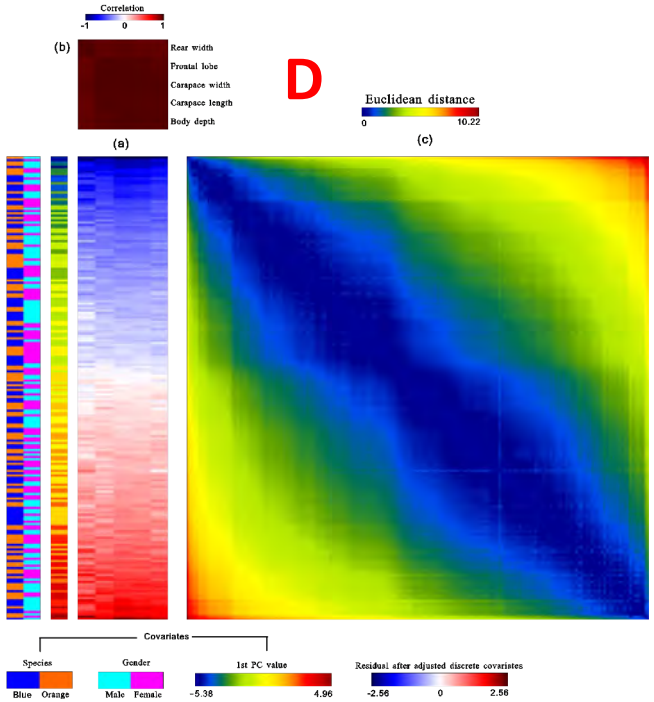
Correlation (Distance) for rows based on

(1) raw data
(2) fitted data
(3) residual data

Morphological Measurements on Leptograpsus Crabs







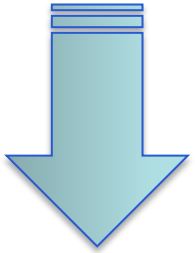
A. MV for the crabs data with covariates.

B. MV With R2E seriation. with 1st PC attached

C. MV adjusted for continuous COV 1st PC with R2E

D. MV adjusted for discrete COV (species, sex) with R2E

Matrix Visualization for Symbolic Data (Analysis)

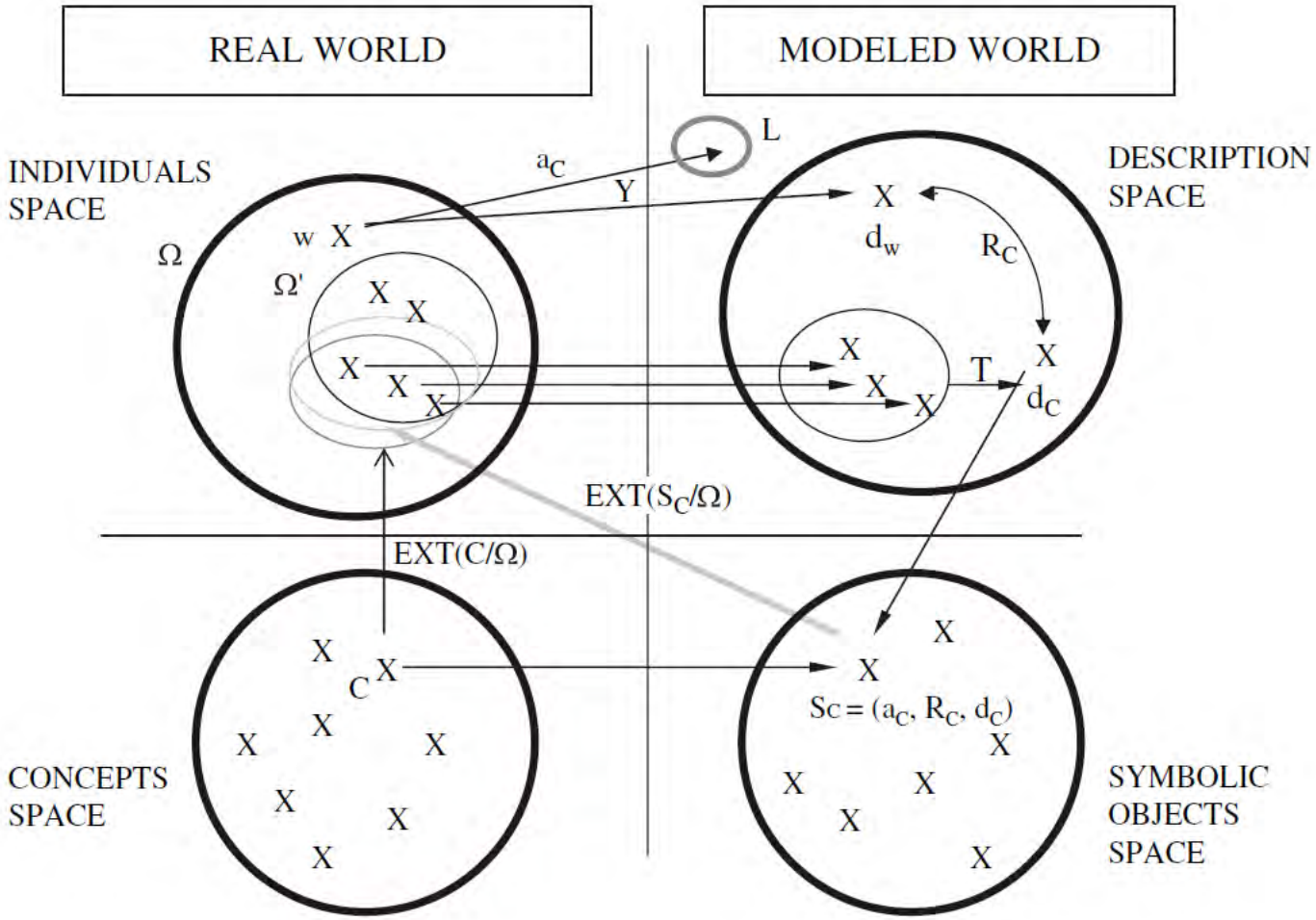


for Big Data?

Edwin Diday

Edwin Diday died on April 28, 2023 at the age of 83. He was elected ISI member in 1985.

Symbolic Data Analysis



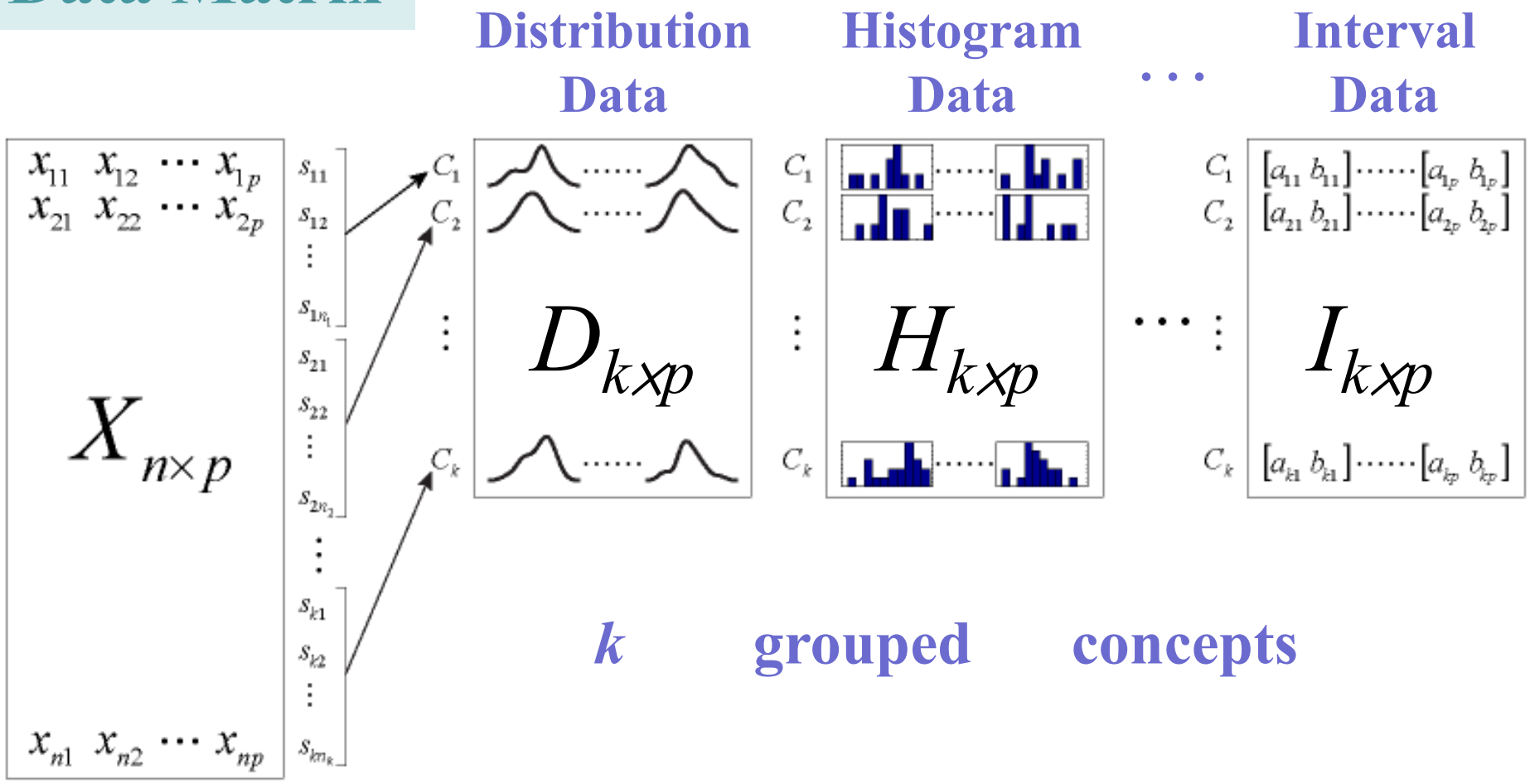
Diday, E., 1987. The symbolic approach in clustering and related methods of Data Analysis. In H.-H. Bock (ed.), Classification and Related Methods of Data Analysis. Amsterdam, North-Holland, 673-684.

Bock, H.-H., Diday, E. (Eds.), 2000. Analysis of symbolic data. Springer-Verlag, Berlin, New York.

Billard, L., Diday, E., 2003. From the statistics of data to the statistics of knowledge: symbolic data analysis. Journal of the American Statistical Association, 98, 470-487.

Conventional Data Matrix

Symbolic Data Matrix



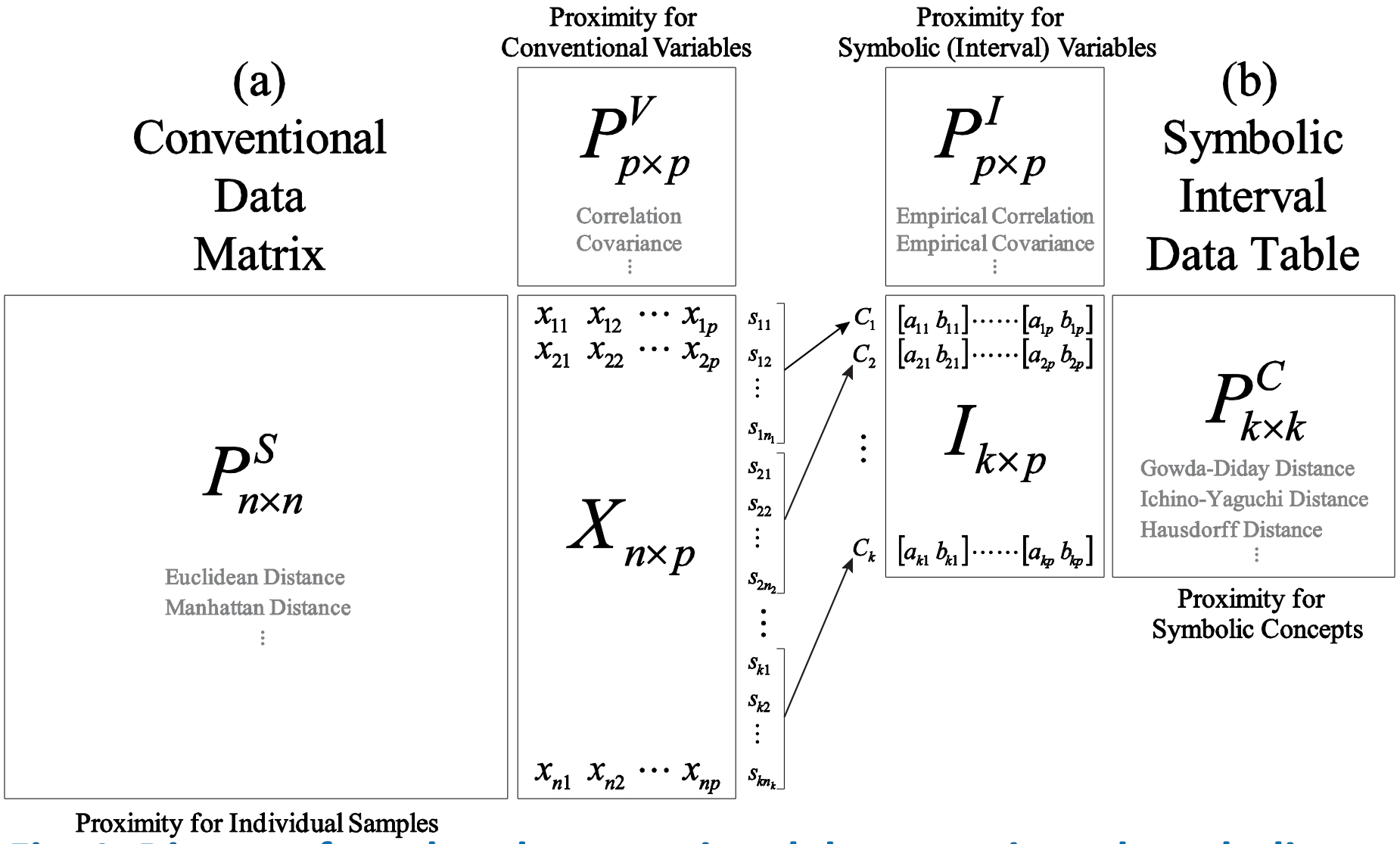
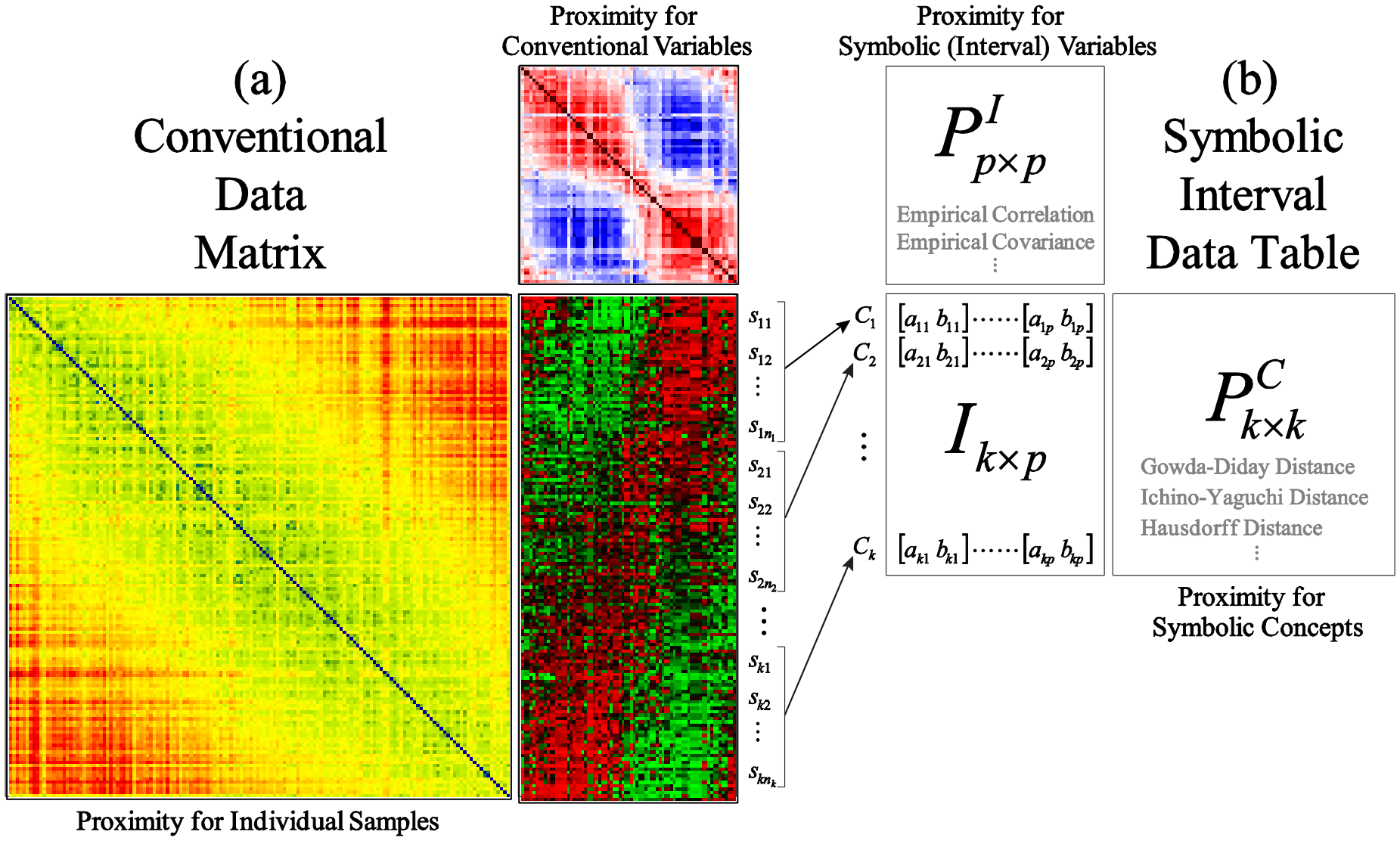


Fig. 1. Diagram for related conventional data matrix and symbolic (interval type) data table with their corresponding proximity matrices for samples/concepts and variables.

Essential elements in a GAP MV procedure?



Proximity matrix for interval (range) variables

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

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

$$\begin{array}{l|c} C_1 & [a_{11}\ b_{11}]\cdots\cdots[a_{1p}\ b_{1p}] \\ C_2 & [a_{21}\ b_{21}]\cdots\cdots[a_{2p}\ b_{2p}] \\ \vdots & I_{k\times p} \\ C_k & [a_{k1}\ b_{k1}]\cdots\cdots[a_{kp}\ b_{kp}] \end{array}$$

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

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

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

	I1:Head	*I2*:Tail	*I3*:Height	*I4*:Forearm
C1:BARB	[44,58]	[41,54]	[6,8]	[35,41]
C2:FCHEV	[50,69]	[30,43]	[11,13]	[51,61]
C3:GMUR	[65,80]	[48,60]	[12,16]	[55,68]
C4:MBEC	[46,53]	[34,44]	[9,11]	[39,44]
C5:MDAUB	[41,51]	[30,39]	[8,11]	[33,41]
C6:MDEC	[40,45]	[39,44]	[9,9]	[36,42]
C7:MGES	[82,87]	[46,57]	[11,12]	[58,63]
C8:MGP	[45,53]	[35,38]	[10,12]	[39,44]
C9:MNAT	[42,50]	[32,43]	[8,9]	[36,42]
C10:MOUS	[38,50]	[30,40]	[7,8]	[32,37]
C11:MSCH	[52,60]	[50,60]	[10,11]	[42,48]
C12:NOCT	[69,82]	[41,59]	[10,12]	[45,55]
C13:OCOM	[41,51]	[34,50]	[9,10]	[34,50]
C14:OGRIS	[47,53]	[43,53]	[7,9]	[37,41]
C15:PIPC	[33,52]	[26,33]	[4,7]	[27,32]
C16:PIPN	[44,48]	[34,44]	[7,8]	[31,36]
C17:PIPS	[43,48]	[34,39]	[6,7]	[31,38]
C18:PRH	[35,43]	[24,30]	[8,11]	[34,41]
C19:SBIC	[50,63]	[40,45]	[8,10]	[40,47]
C20:SBOR	[48,54]	[38,47]	[9,11]	[37,42]
C21:SCOM	[62,80]	[46,57]	[9,12]	[48,56]

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

Proximity matrix for concepts with interval variables

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 $0 \le \gamma \le 0.5$
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	$\sqrt[2]{\sum_{r=1}^{p} [\frac{D(I_{ir}, I_{jr})}{\lvert R_r\rvert}]^2}$	$D(I_{ir}, I_{jr}) = max(\lvert a_{ir} - a_{jr}\rvert, \lvert b_{ir} - b_{jr}\rvert)$ $\lvert R_r\rvert = \max_c b_{cr} - \min_c a_{cr}$

$$\begin{array}{c|c} C_1 & [a_{11}\ b_{11}]\cdots\cdots[a_{1p}\ b_{1p}] \\ C_2 & [a_{21}\ b_{21}]\cdots\cdots[a_{2p}\ b_{2p}] \\ & \\ & I_{k\times p} \\ & \\ C_k & [a_{k1}\ b_{k1}]\cdots\cdots[a_{kp}\ b_{kp}] \end{array}$$

	I1:Head	*I2*:Tail	*I3*:Height	*I4*:Forearm
C1:BARB	[44,58]	[41,54]	[6,8]	[35,41]
C2:FCHEV	[50,69]	[30,43]	[11,13]	[51,61]
C3:GMUR	[65,80]	[48,60]	[12,16]	[55,68]
C4:MBEC	[46,53]	[34,44]	[9,11]	[39,44]
C5:MDAUB	[41,51]	[30,39]	[8,11]	[33,41]
C6:MDEC	[40,45]	[39,44]	[9,9]	[36,42]
C7:MGES	[82,87]	[46,57]	[11,12]	[58,63]
C8:MGP	[45,53]	[35,38]	[10,12]	[39,44]
C9:MNAT	[42,50]	[32,43]	[8,9]	[36,42]
C10:MOUS	[38,50]	[30,40]	[7,8]	[32,37]
C11:MSCH	[52,60]	[50,60]	[10,11]	[42,48]
C12:NOCT	[69,82]	[41,59]	[10,12]	[45,55]
C13:OCOM	[41,51]	[34,50]	[9,10]	[34,50]
C14:OGRIS	[47,53]	[43,53]	[7,9]	[37,41]
C15:PIPC	[33,52]	[26,33]	[4,7]	[27,32]
C16:PIPN	[44,48]	[34,44]	[7,8]	[31,36]
C17:PIPS	[43,48]	[34,39]	[6,7]	[31,38]
C18:PRH	[35,43]	[24,30]	[8,11]	[34,41]
C19:SBIC	[50,63]	[40,45]	[8,10]	[40,47]
C20:SBOR	[48,54]	[38,47]	[9,11]	[37,42]
C21:SCOM	[62,80]	[46,57]	[9,12]	[48,56]

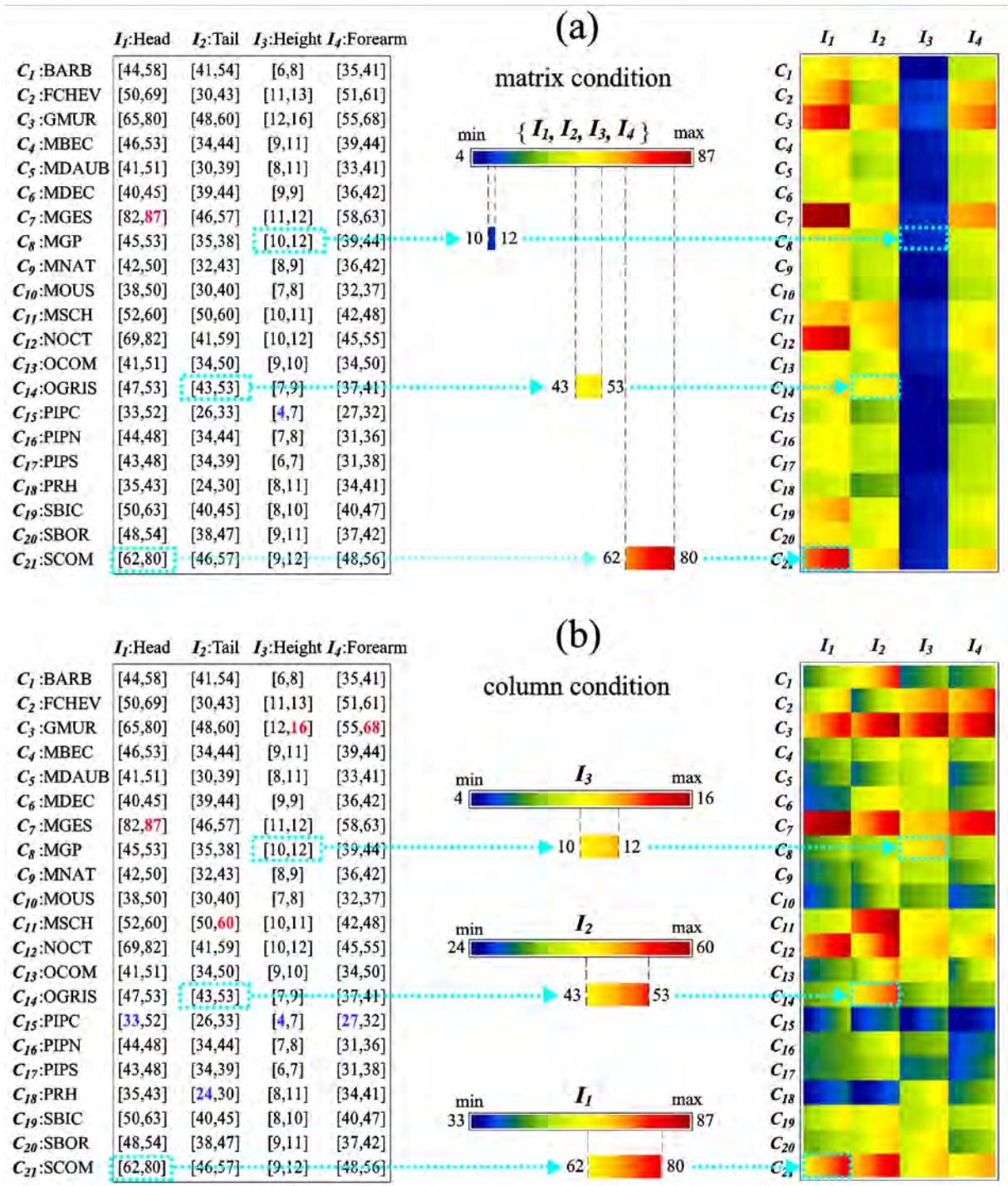
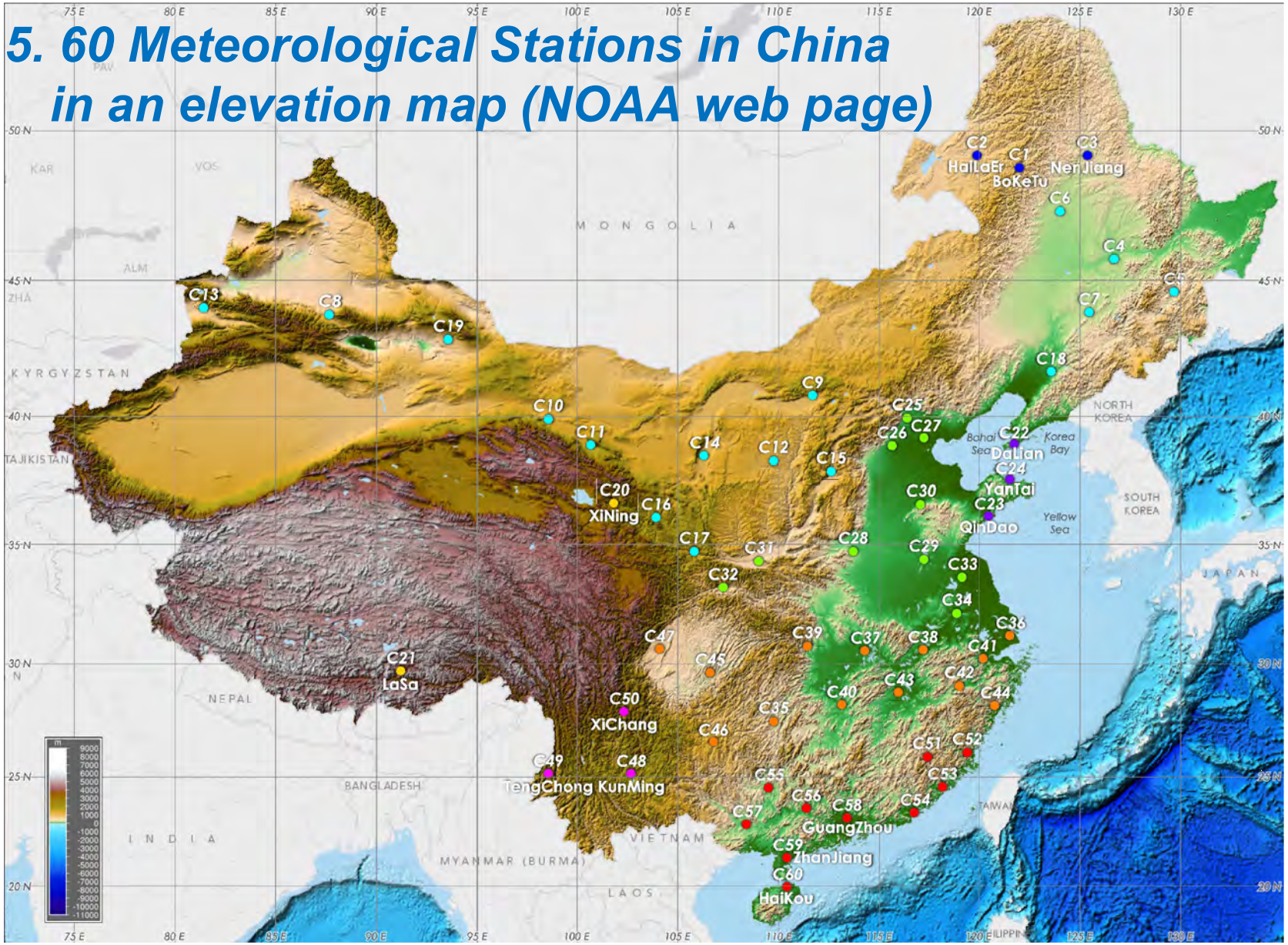


Fig. 3.
Color-coding scheme for interval-valued symbolic data using the Bats example.
(a) Matrix condition.
(b) Column condition.

Fig. 5. 60 Meteorological Stations in China in an elevation map (NOAA web page)



Colors for representing related clusters of stations identified from tree structure in Fig. 6(a) are used to code each of the individual stations and white outer circle for those stations with number of disagreements ≥48.

WIKIPEDIA
The Free Encyclopedia

Abuja

Federal capital city and local government area


Maitama skyline


Abuja National Mosque


Zuma Rock


Fountain in Millennium Park


City Gate


Central Bank headquarters

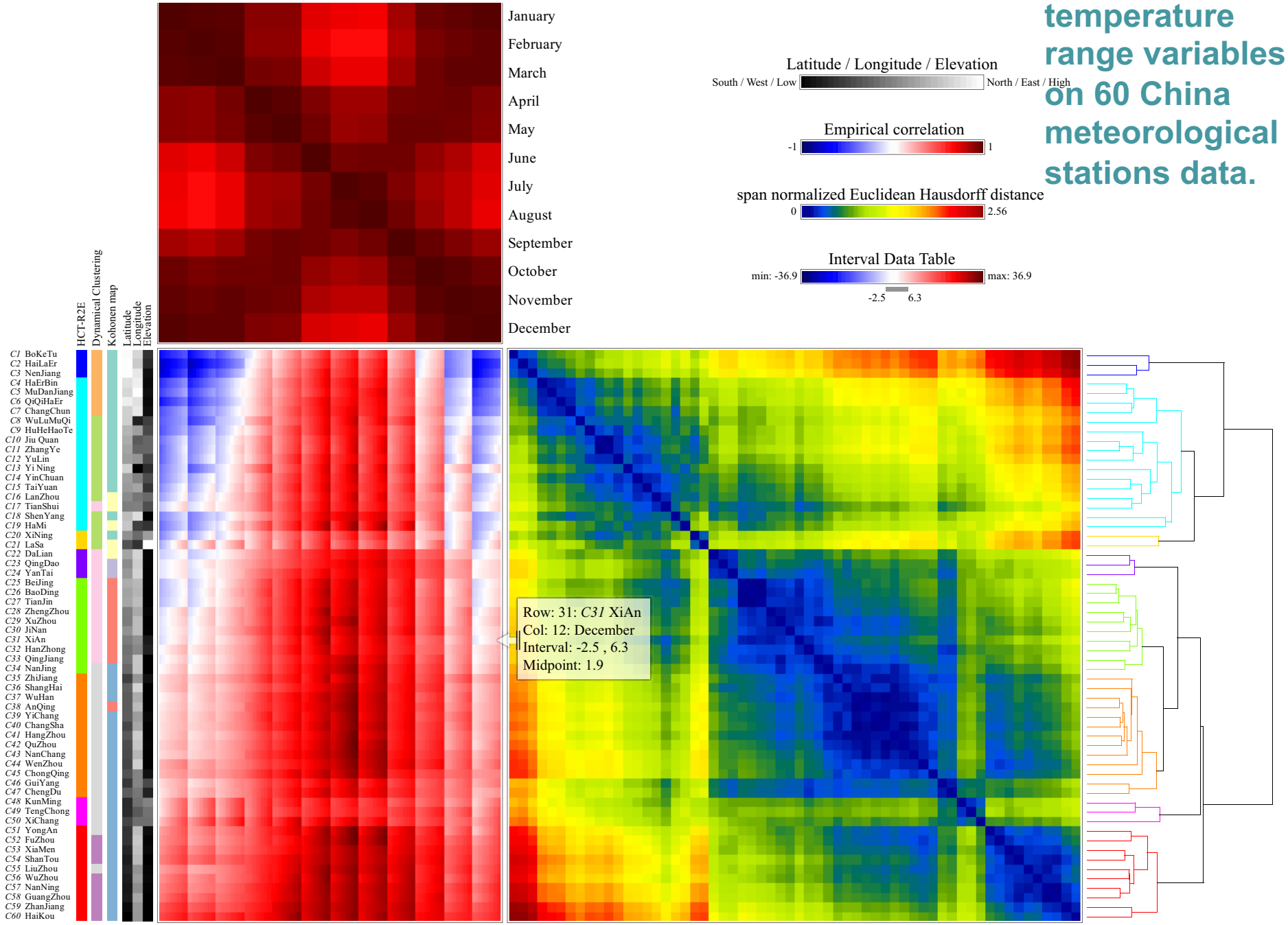

Central Business District skyline

Climate data for Abuja, Nigeria (1991–2020) [hide]

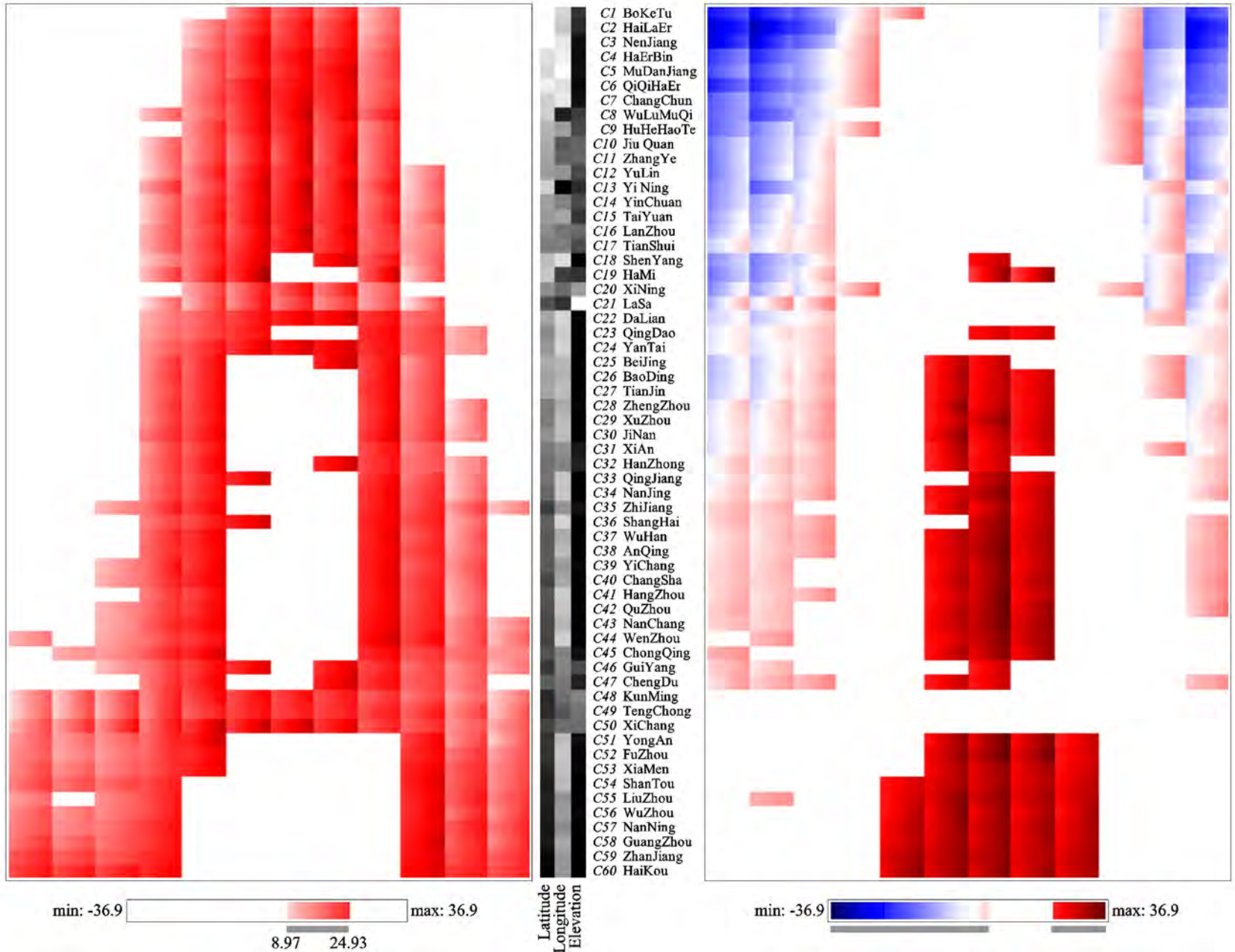
Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °C (°F)	40.0 (104.0)	42.0 (107.6)	42.0 (107.6)	41.0 (105.8)	39.7 (103.5)	37.8 (100.0)	34.7 (94.5)	33.5 (92.3)	34.0 (93.2)	40.0 (104.0)	38.1 (100.6)	39.0 (102.2)	42.0 (107.6)
Mean daily maximum °C (°F)	35.2 (95.4)	36.9 (98.4)	37.3 (99.1)	35.6 (96.1)	32.9 (91.2)	30.8 (87.4)	29.5 (85.1)	28.7 (83.7)	29.9 (85.8)	31.3 (88.3)	34.2 (93.6)	35.0 (95.0)	33.1 (91.6)
Daily mean °C (°F)	26.9 (80.4)	29.3 (84.7)	30.7 (87.3)	30.0 (86.0)	28.1 (82.6)	26.6 (79.9)	25.7 (78.3)	25.2 (77.4)	25.7 (78.3)	26.5 (79.7)	26.9 (80.4)	26.3 (79.3)	27.3 (81.1)
Mean daily minimum °C (°F)	18.5 (65.3)	21.6 (70.9)	24.1 (75.4)	24.4 (75.9)	23.3 (73.9)	22.3 (72.1)	22.0 (71.6)	21.8 (71.2)	21.6 (70.9)	21.6 (70.9)	19.7 (67.5)	17.7 (63.9)	21.5 (70.7)
Record low °C (°F)	11.0 (51.8)	13.7 (56.7)	15.0 (59.0)	14.0 (57.2)	15.0 (59.0)	17.2 (63.0)	16.0 (60.8)	15.0 (59.0)	18.0 (64.4)	16.5 (61.7)	13.0 (55.4)	8.9 (48.0)	8.9 (48.0)
Average precipitation mm (inches)	0.8 (0.03)	5.9 (0.23)	22.4 (0.88)	72.8 (2.87)	156.7 (6.17)	194.4 (7.65)	249.8 (9.83)	308.3 (12.14)	229.4 (9.03)	169.5 (6.67)	9.7 (0.38)	1.3 (0.05)	1,421.1 (55.95)
Average precipitation days (≥ 1 mm)	0.1	0.4	1.9	6.0	11.3	12.2	15.0	17.3	16.0	13.0	0.9	0.0	94.0
Average relative humidity (%)	44.4	44.9	56.1	71.5	80.9	84.7	86.8	88.0	86.8	83.4	65.8	50.9	70.3

Source: NOAA[82]

Fig. 6. (a) Three MV maps sorted by HCT–R2E dendrograms for 12 monthly temperature range variables on 60 China meteorological stations data.



(b) Midpoint condition range display for 60 China meteorological stations data
Left panel: only temperature intervals with midpoint within the range of (9–25 ˚C);
Right panel: only intervals with midpoint outside the range of (9–25 ˚C).



Matrix Visualization for Big Data ?

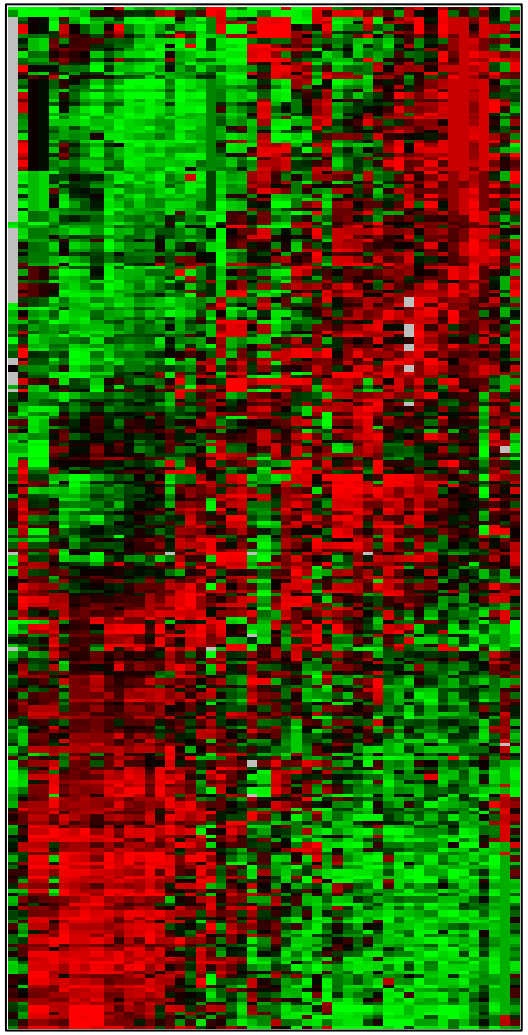
全民健康保險研究資料庫
National Health Insurance Research Database

99.9% of Taiwan‘s population (23,894,394) were enrolled in **National Health Insurance Program**. Foreigners in Taiwan are also eligible for this program.

The database of this program contains **registration** files and **original claim** data for reimbursement.

GAP for Big Data?

100,000 variables



23,894,394 samples

?

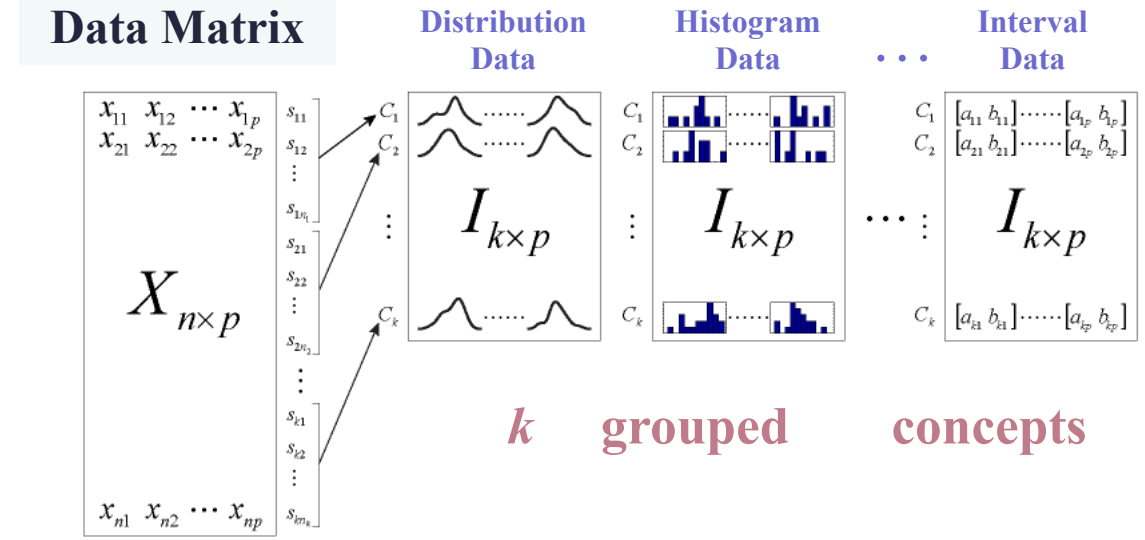
?

?

- Scaling
- Sampling
- Dimension reduction/
  Variable selection
- Clustering
  - Summary
  - (Sufficient Display)
    - Single value
    - Interval-type value

Conventional Data Matrix

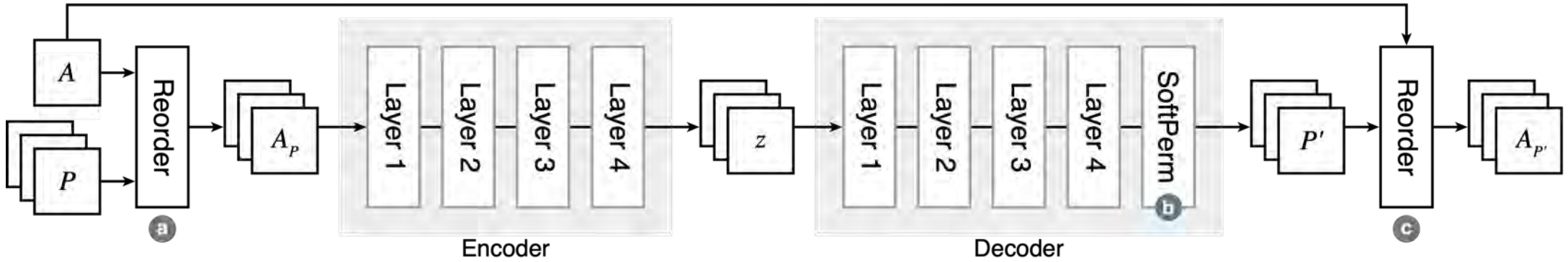
Symbolic Data Matrix



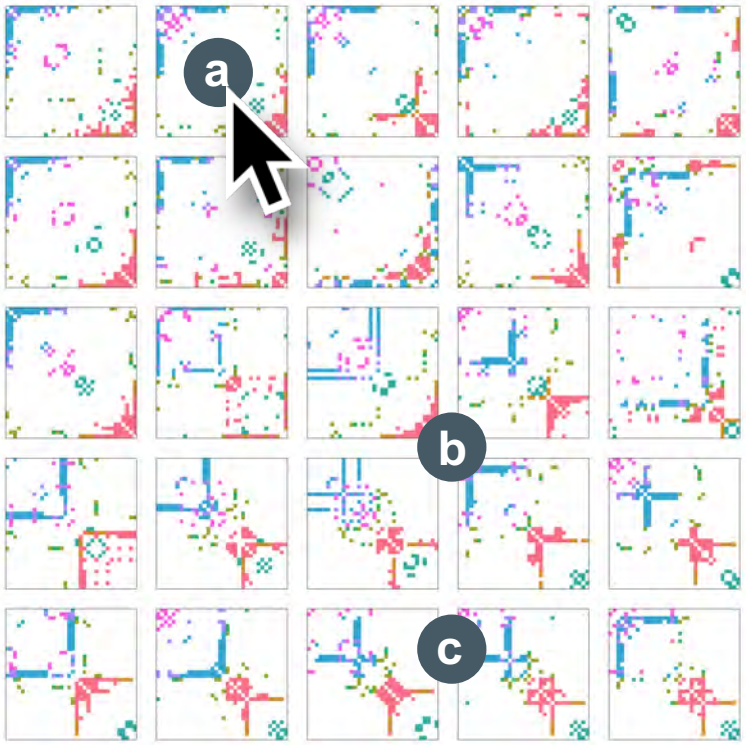
n individual samples

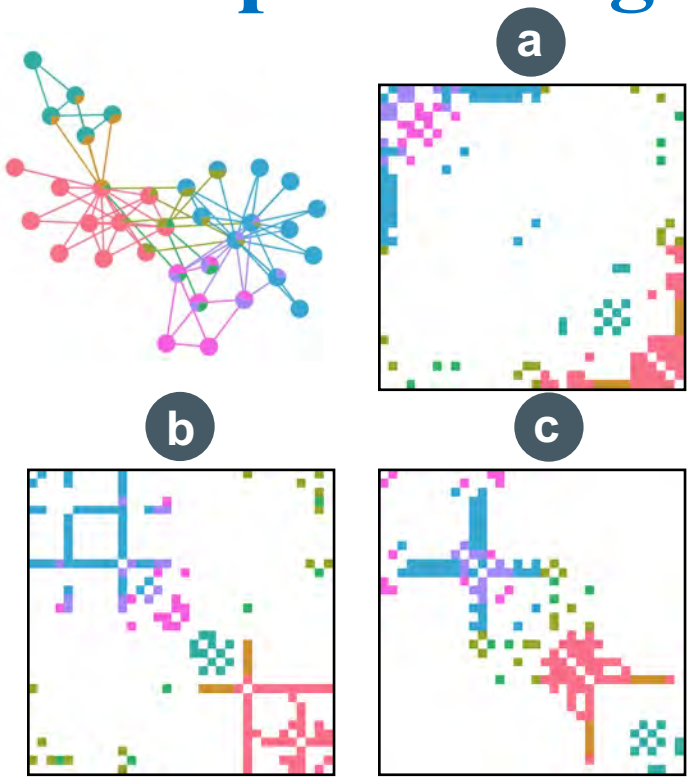
k grouped concepts

A Deep Generative Model for Reordering Adjacency Matrices IEEE Transactions on Visualization and Computer Graphics (2021) Oh-Hyun Kwan, Kwan-Liu Ma, Chiun-How Kao, Chun-houh Chen

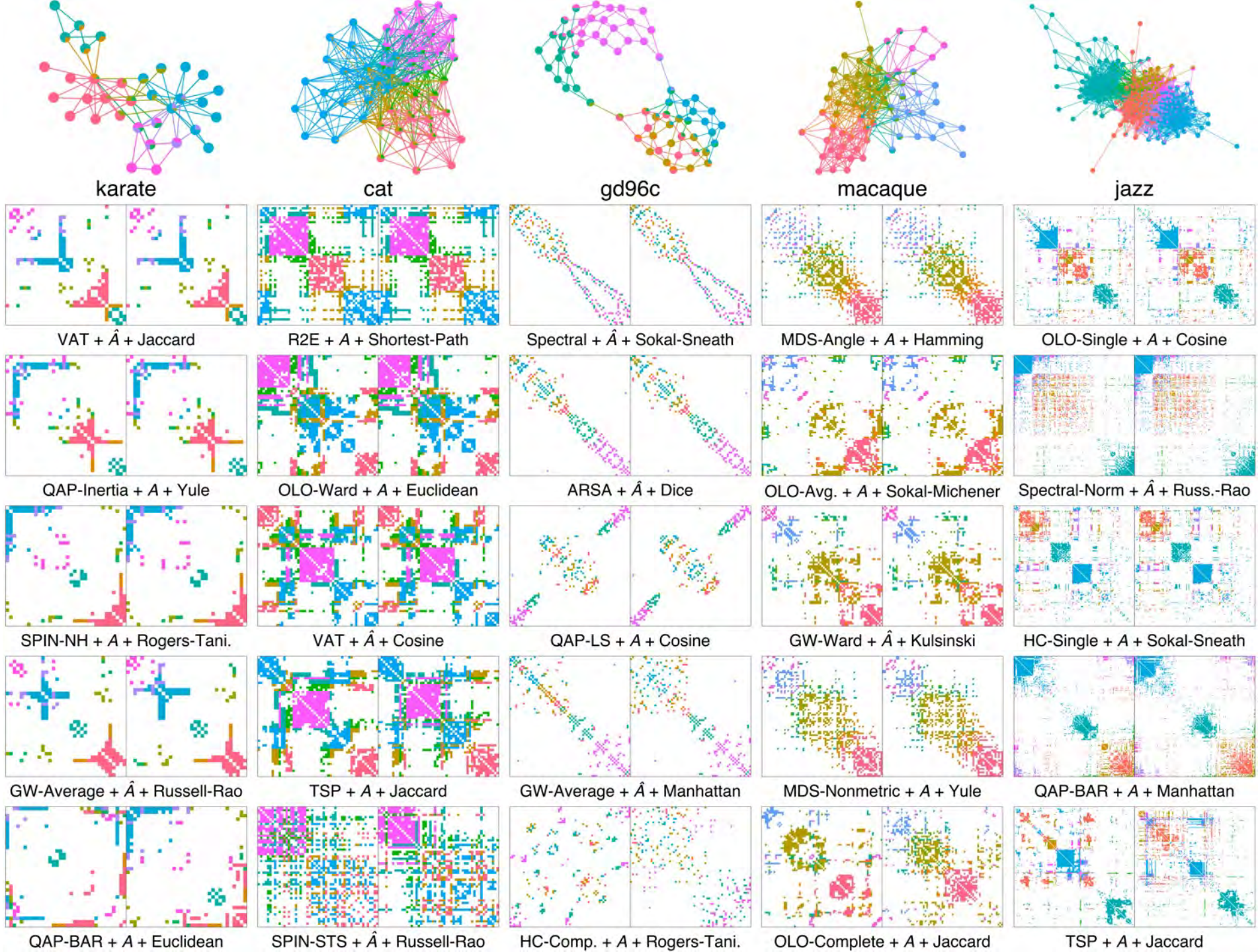


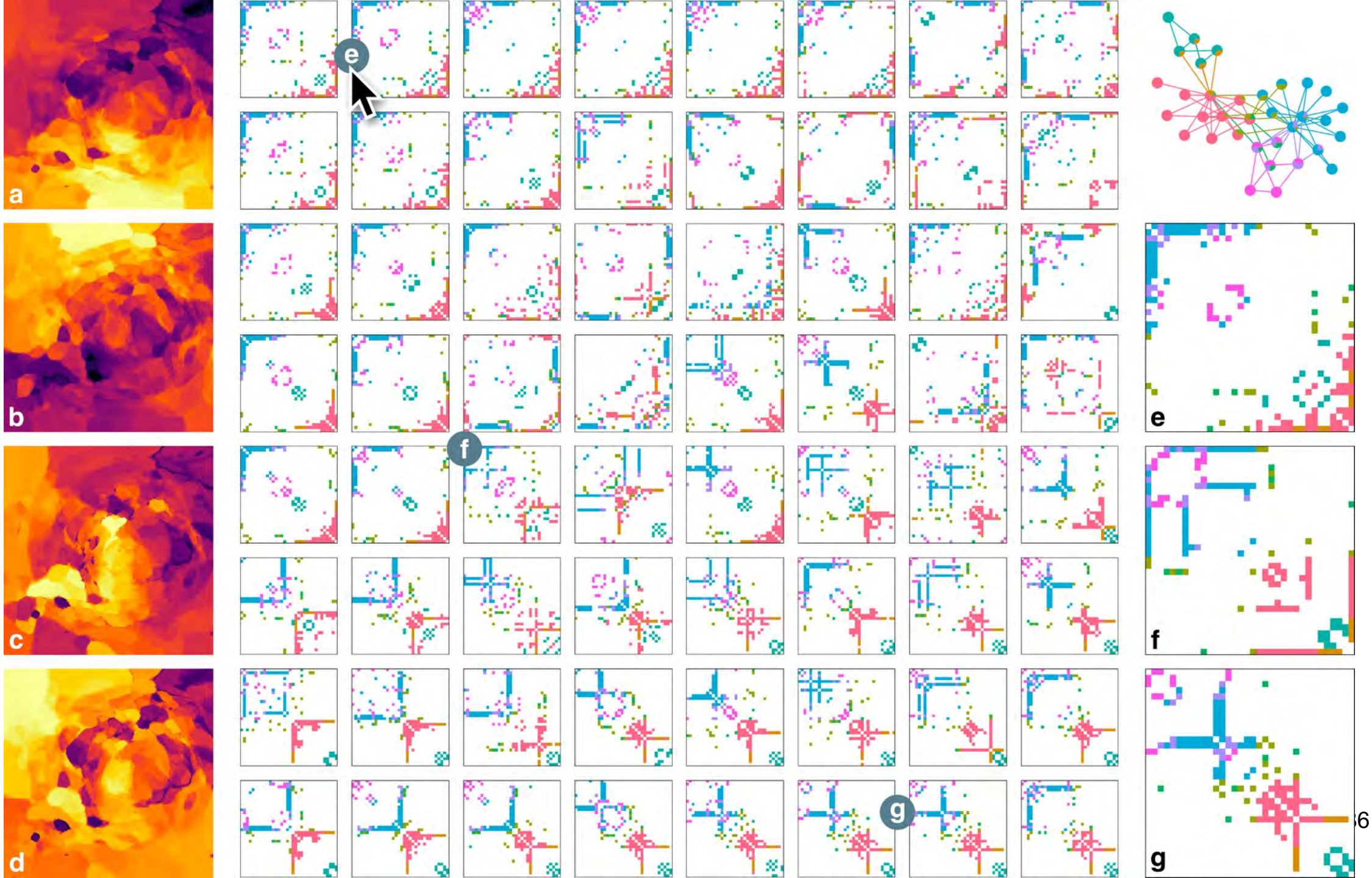
GAP modules with deep learning





- Design an encoder-decoder architecture that learns a generative model for matrix reordering;
- The model learns a latent space of diverse matrix reorderings;
- Build WYSIWYG Interface with the Learned Latent Space.


karate
cat
gd96c
macaque
jazz
VAT + Â + Jaccard
R2E + A + Shortest-Path
Spectral + Â + Sokal-Sneath
MDS-Angle + A + Hamming
OLO-Single + A + Cosine
QAP-Inertia + A + Yule
OLO-Ward + A + Euclidean
ARSA + Â + Dice
OLO-Avg. + A + Sokal-Michener
Spectral-Norm + Â + Russ.-Rao
SPIN-NH + A + Rogers-Tani.
VAT + Â + Cosine
QAP-LS + A + Cosine
GW-Ward + Â + Kulsinski
HC-Single + A + Sokal-Sneath
GW-Average + Â + Russell-Rao
TSP + A + Jaccard
GW-Average + Â + Manhattan
MDS-Nonmetric + A + Yule
QAP-BAR + A + Manhattan
QAP-BAR + A + Euclidean
SPIN-STS + Â + Russell-Rao
HC-Comp. + A + Rogers-Tani.
OLO-Complete + A + Jaccard
TSP + A + Jaccard


a
b
c
d
e
f
g
e
f
g

RESEARCH ARTICLE

Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations

Hsin-Chou Yang, Chun-houh Chen, Jen-Hung Wang, Hsiao-Chi Liao, Chih-Ting Yang, Chia-Wei Chen, Yin-Chun Lin, Chiun-How Kao, Mei-Yeh Jade Lu, and James C. Liao

PNAS first published November 12, 2020; https://doi.org/10.1073/pnas.2007840117

Contributed by James C. Liao, October 7, 2020 (sent for review April 24, 2020; reviewed by George Michailidis and Chiara Sabatti)

Article Figures & SI Info & Metrics PDF

Significance

Article Alerts
Email Article
Citation Tools
Request Permissions
Share
Tweet
讚 8
Mendeley

Current Issue
Submit

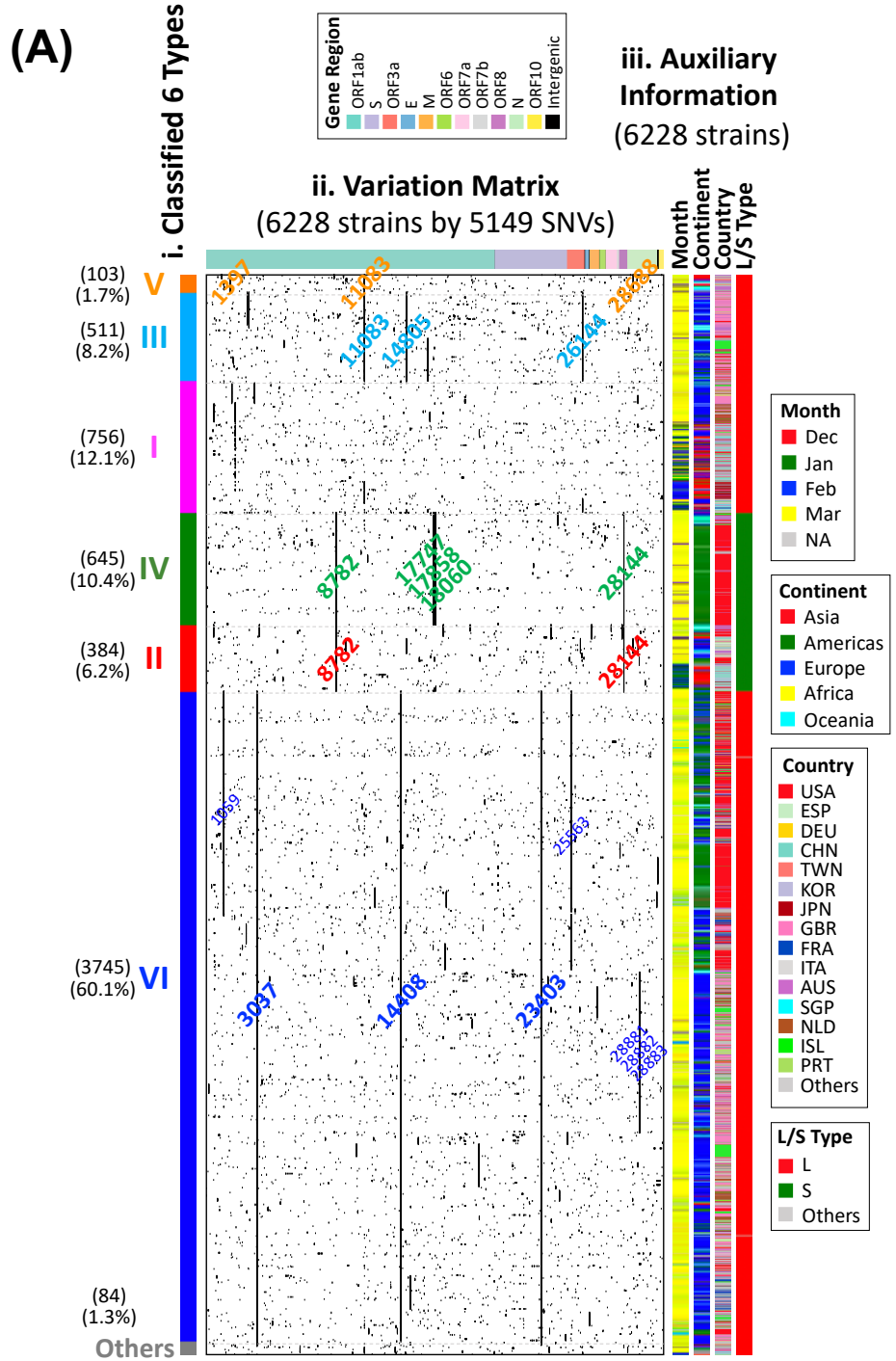
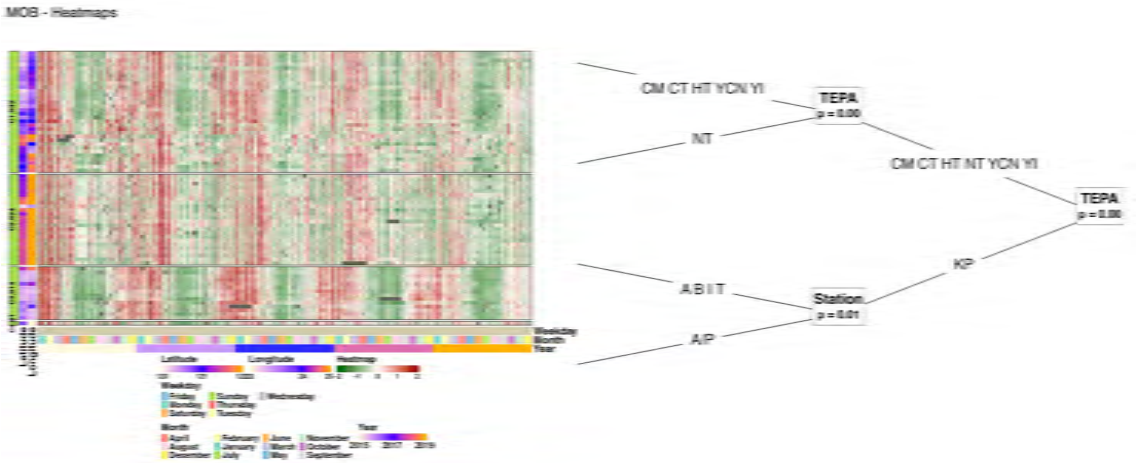


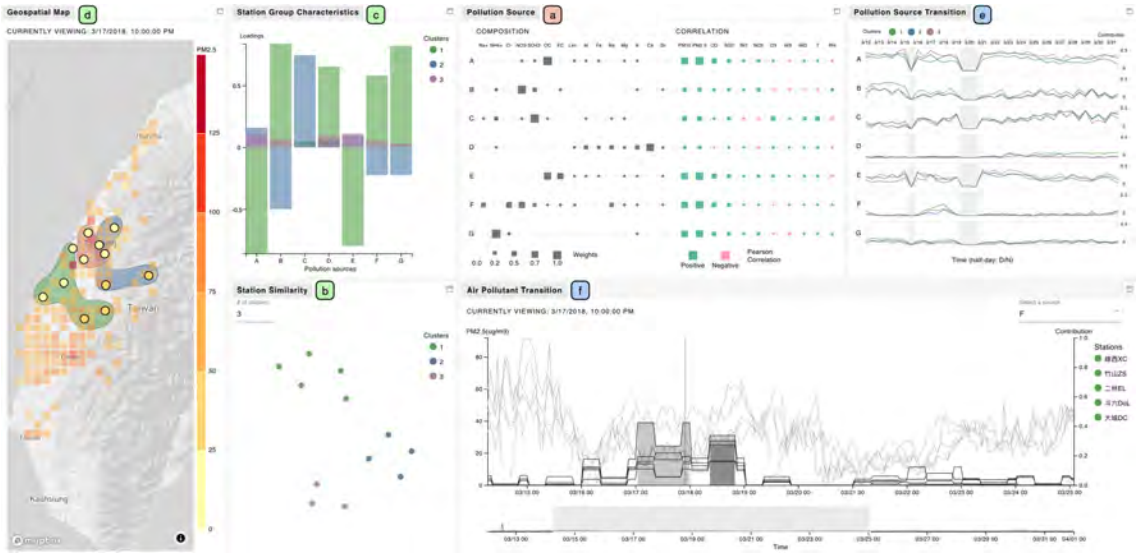
Fig. 1. Variation matrix map and viral strain type. (A) Classified six types with variation matrix map for 6,228 validated SARS-CoV-2 viral strains. (i) Color bar for classified six types for 6,228 validated strains. (ii) Variation matrix map for 6,228 strains with 5,643 variation sites. Strains are sorted by classified strain types in the order of V-III-I-IV-II-VI-Others from the maximum parsimony dendrogram for 1,932 strains. Nucleotides are listed by relative positions in the genome with color bands indicating their corresponding genome regions on the top panel. All 5,643 nucleotides have at least one variation among 6,228 strains. Signature variations for each type (and subtypes for Type VI) are labeled with corresponding type colors. (iii) Auxiliary information for each virus strain on month of data collection, continent, country, two strain types (L, S) defined by Tang et al. (2020, National Science Review).

GAP Applications for Air Pollution ($PM_{2.5}$) data

Ashouri et al., (2023) An interactive clustering-based visualization tool for air quality data analysis, ***Aerosol and Air Quality Research***

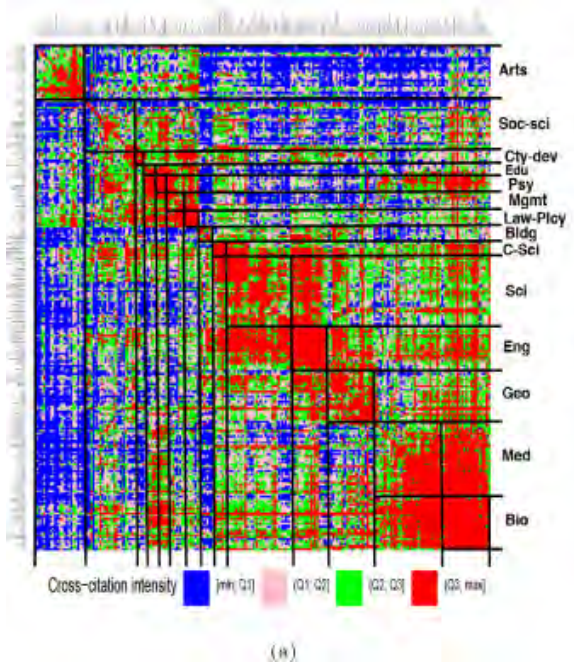


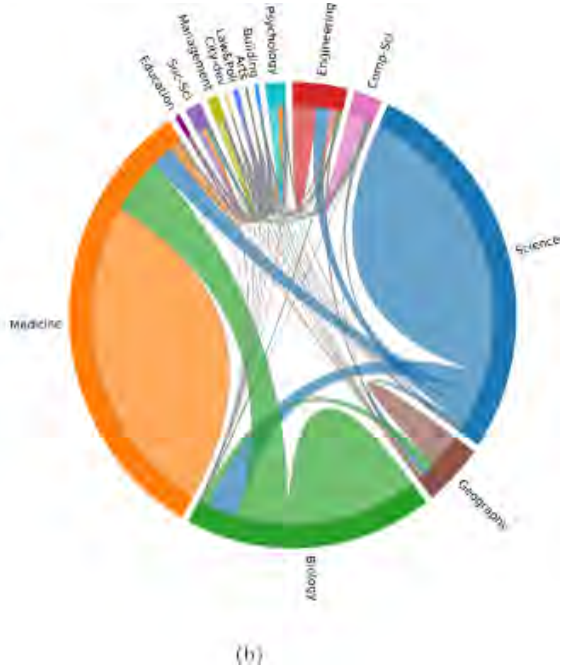
Kuo, et al., (2022), Visual Analytics of Air Pollution Data with Machine-Learning-Aided Analysis Workflows, ***IEEE PacificVis***



GAP Application for Article's Scientific Prestige (ASP) metric

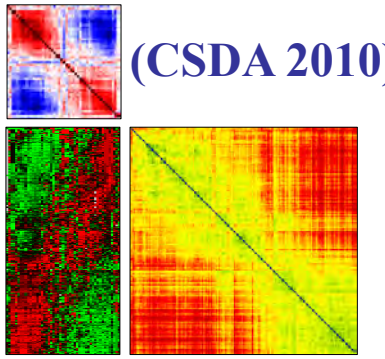
Chen et al., (2023) Article's scientific prestige: Measuring the impact of individual articles in the web of science, ***Journal of Informetrics***





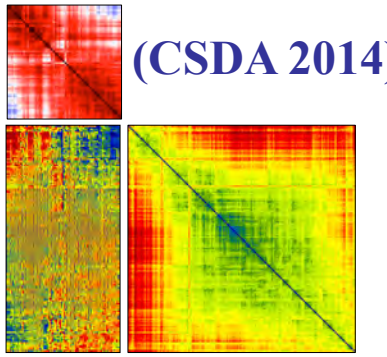
Modules of GAP for Matrix Visualization

Continuous X



(CSDA 2010)

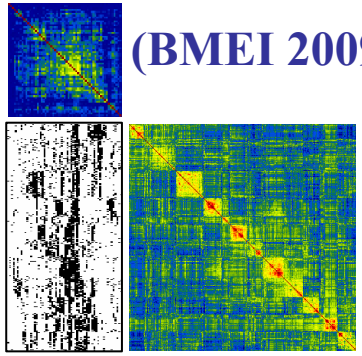
Symbolic X



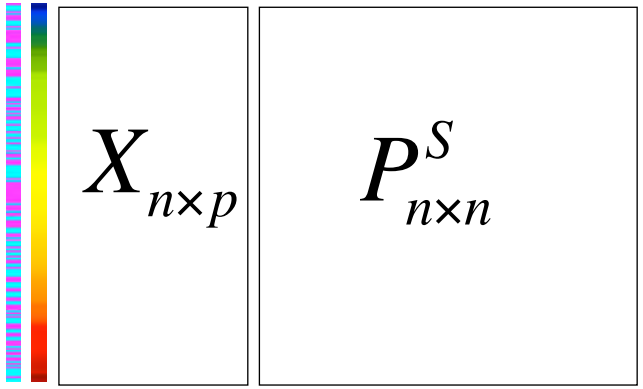
(CSDA 2014)

$P^{V}_{p\times p}$

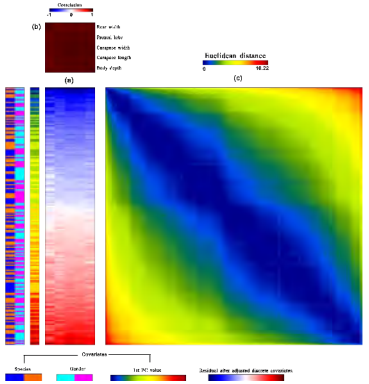
Binary X



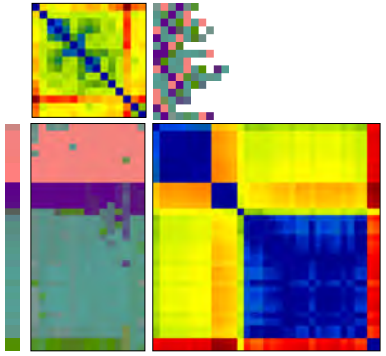
(BMEI 2009)



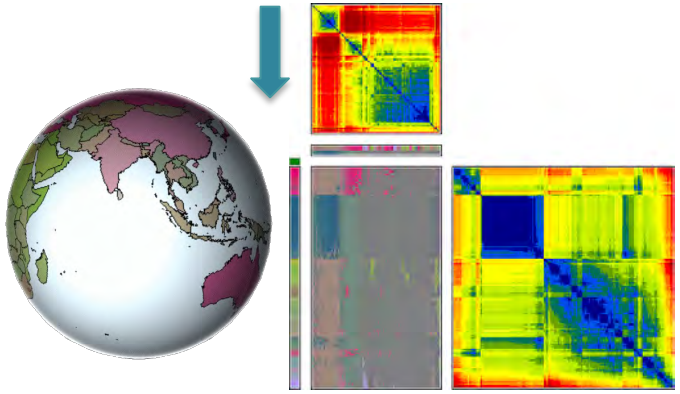
Covariate Adjusted X



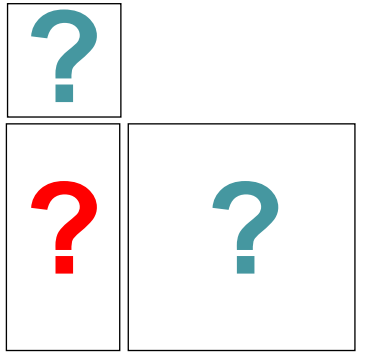
Categorical X



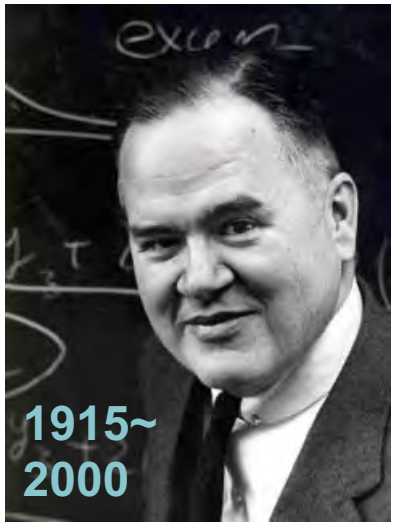
Cartographical X



??????? X



Exploratory Data Analysis
EDA, John Tukey (1977)

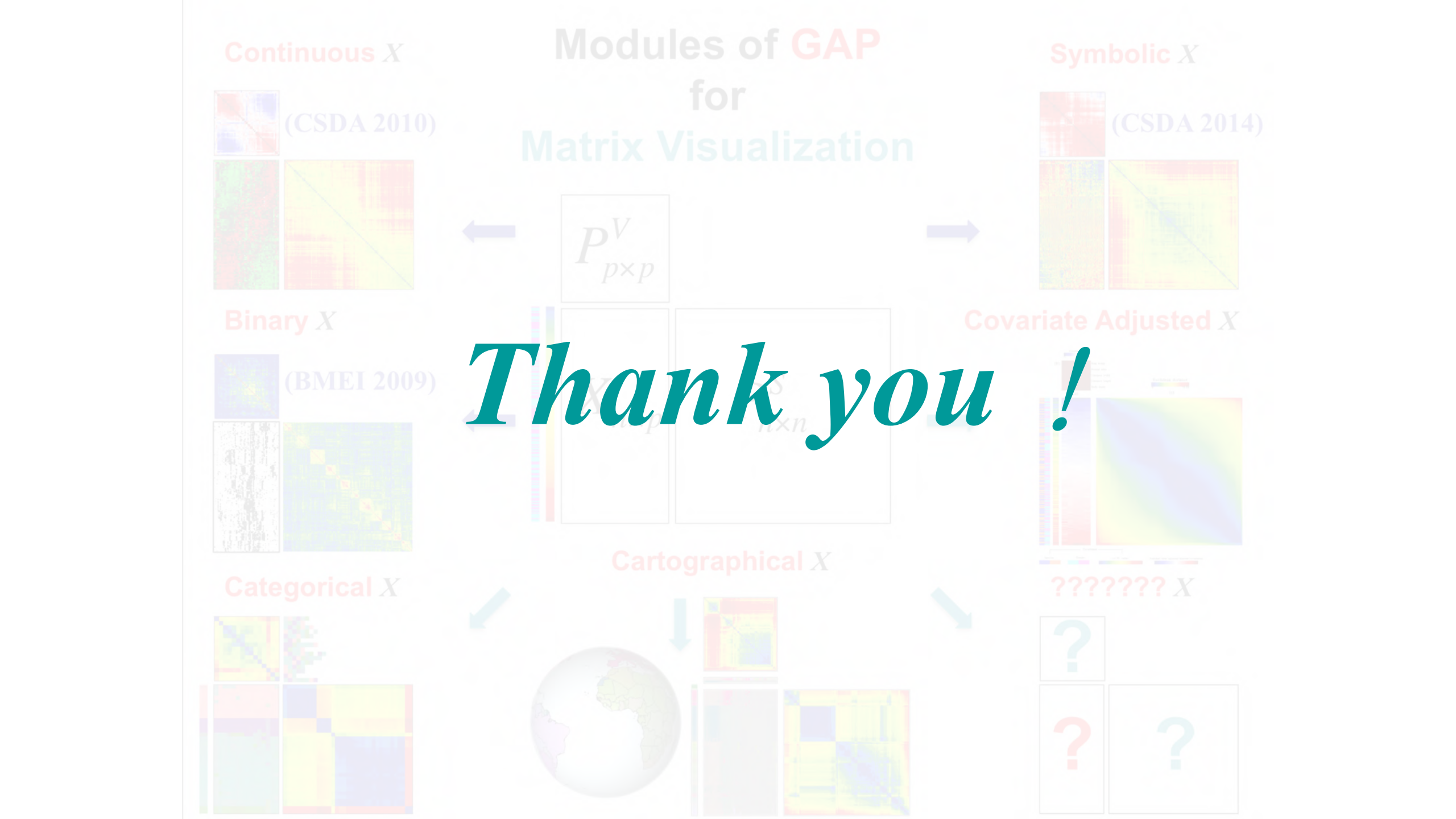


It is important to understand what you **CAN DO** before you learn to measure how **WELL** you seem to have **DONE** it.

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


Modules of GAP
for
Matrix Visualization
Continuous X
(CSDA 2010)
Symbolic X
(CSDA 2014)
Binary X
(BMEI 2009)
Covariate Adjusted X
Cartographical X
Categorical X
??????? X
Thank you !