

Matrix Visualization: a review and perspective

Han-Ming Wu¹ (吳漢銘)

and

Chun-houh Chen² (陳君厚)

¹Department of Statistics, National Taipei University,
Taiwan

²Institute of Statistical Science, Academia Sinica, Taiwan

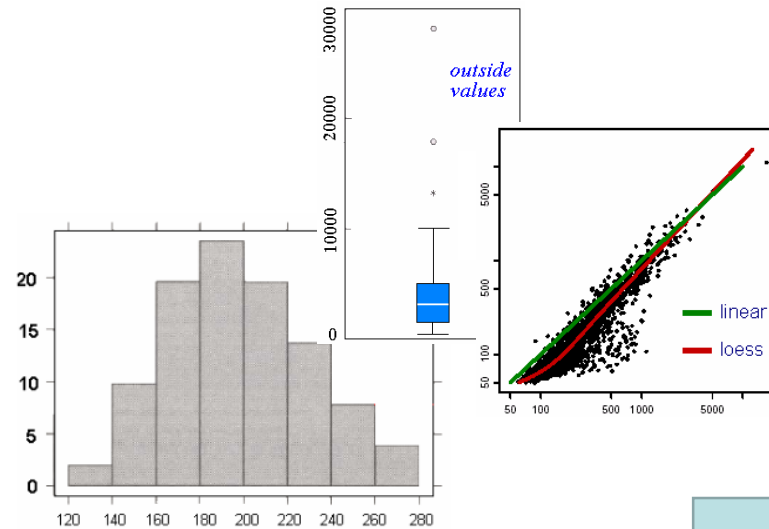


矩陣視覺化: 回顧與展望

Outlines

- Why Data Visualization?
- Heatmaps (i.e., Matrix Visualization)
- The Basic Principles of Matrix Visualization
(GAP (Generalized Association Plots) Approach)
 - Presentation of Raw Data Matrix
 - Seriation of Proximity Matrices and Raw Data Matrix
- Literature Review:
 - Applications/Software/Review/Point of View/Methods
- Related Works of MV
- Perspective

Data/Information Visualization



information

- 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**.

Visualization = Graphing for Data + Fitting + Graphing for Model



Why Data Visualization?

- It is not about "**infographics**", the beautiful, heavily customized products of expert graphic designers.
- Data visualization can provide clear understanding of patterns in data, detect hidden structures in data, condense information.
- **Anscombe's quartet** comprises four datasets. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.
- Four datasets have nearly identical simple statistical properties, yet appear very different when graphed.

	I		II		III		IV	
	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

Mean of *x* in each case: 9 (exact)

Sample variance of *x* in each case: 11 (exact)

Mean of *y* in each case: 7.50 (to 2 decimal places)

Sample variance of *y* in each case: 4.122 or 4.127 (to 3 decimal places)

Correlation between *x* and *y* in each case: 0.816 (to 3 decimal places)

Linear regression line in each case: $y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

https://en.wikipedia.org/wiki/Anscombe's_quartet

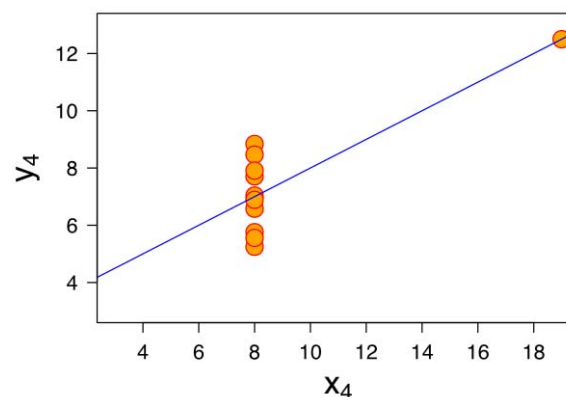
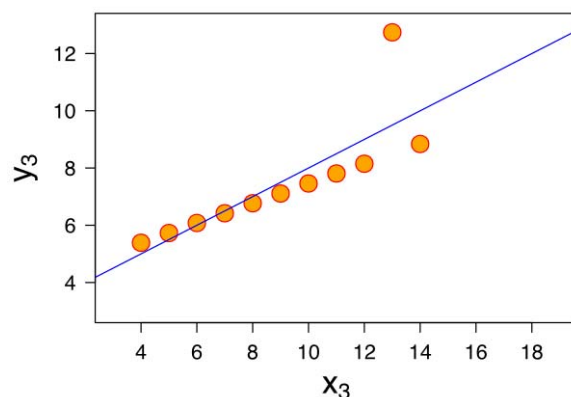
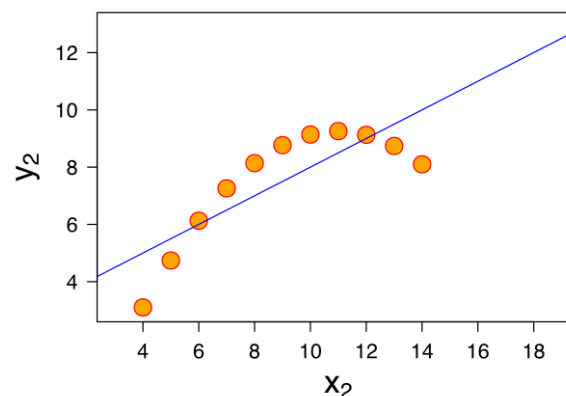
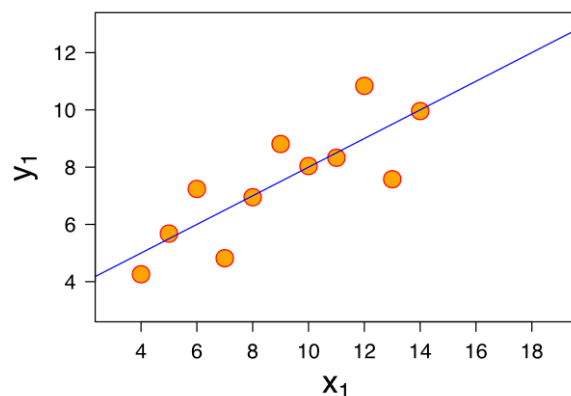
<http://ryanwomack.com/IASSIST/DataViz/>

Anscombe's Quartet



5/76

- Mean of x in each case: 9 (exact)
- Sample variance of x in each case: 11 (exact)
- Mean of y in each case: 7.50 (to 2 decimal places)
- Sample variance of y in each case: 4.122 or 4.127 (to 3 decimal places)
- Correlation between x and y in each case: 0.816 (to 3 decimal places)
- Linear regression line in each case: $y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

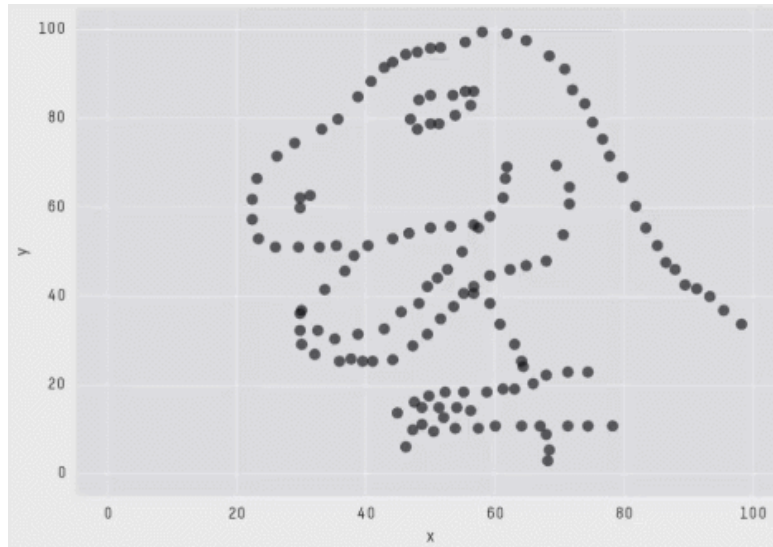


The Datasaurus Dozen

```
install.packages("datasauRus")
```

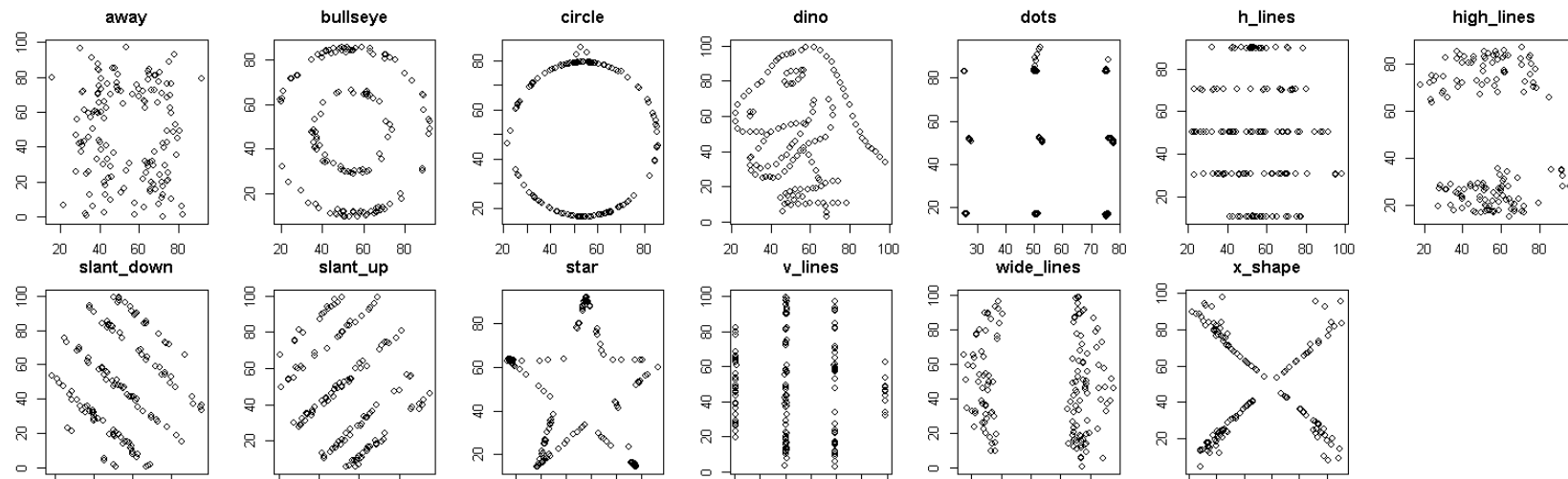


6/76

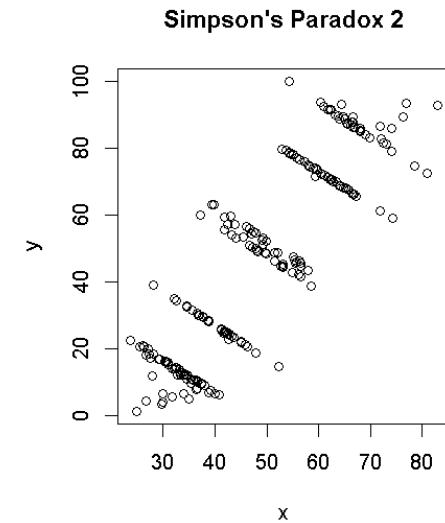
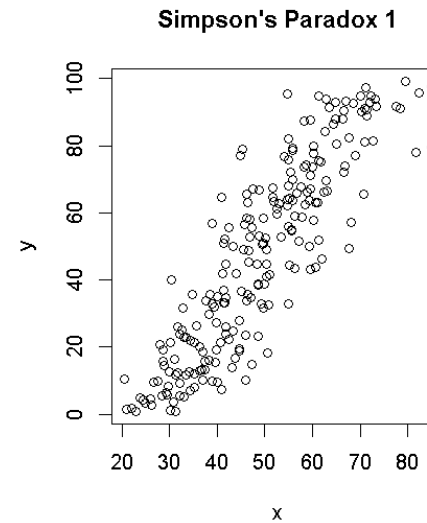
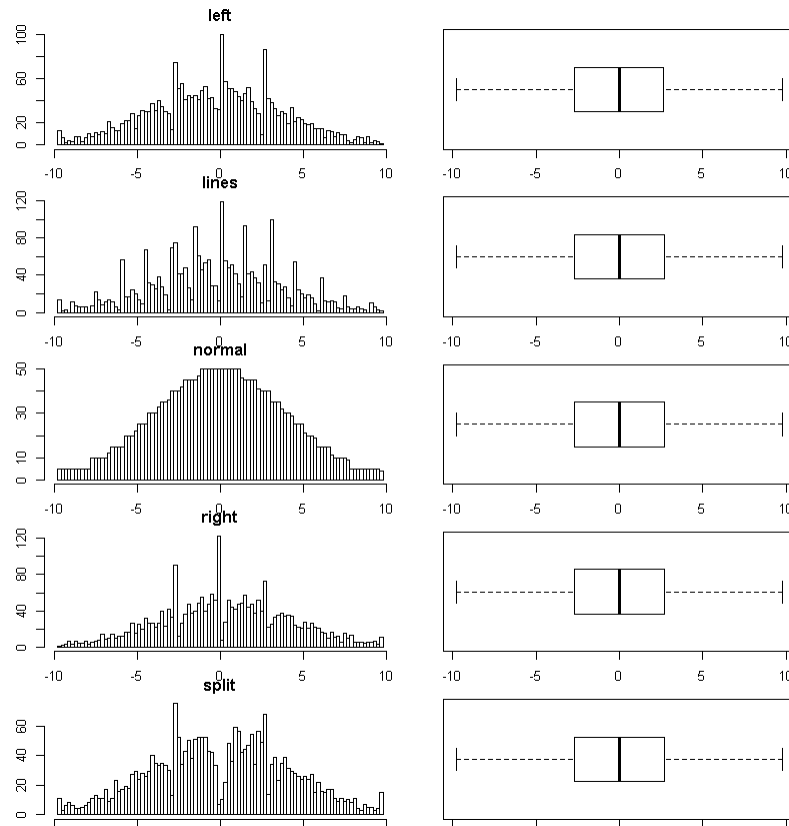


X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Justin Matejka and George Fitzmaurice, Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing.



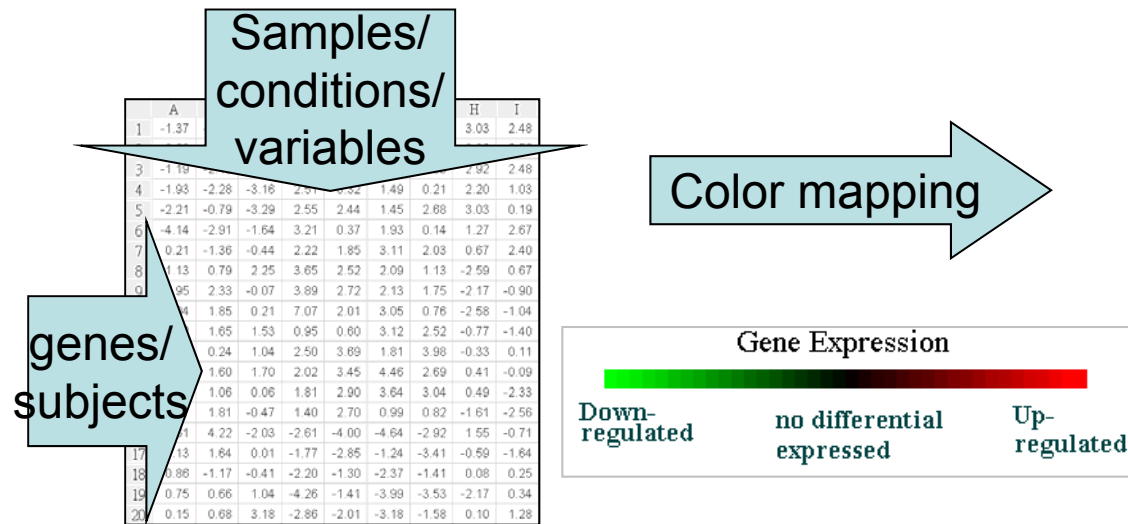
The Datasaurus Dozen: More examples



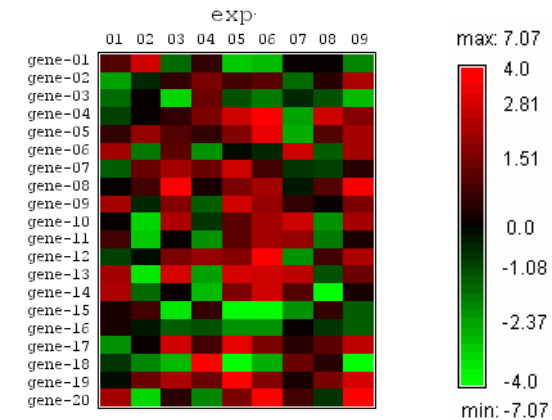


8/76

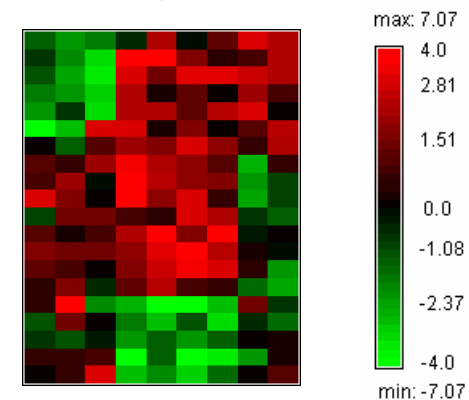
Heatmaps



Without ordering



Ordering/Seriation/
Clustering



- Heatmaps represent two-dimensional tables of **numbers** as shades of **colors**.
- The dense and intuitive display makes heatmaps well-suited for presentation of **high-throughput data**.
- Heatmaps rely fundamentally on **color encoding** and on **meaningful reordering** of the rows and columns.



Table 1. Using frequency of heatmap.

Journal	Num. of papers	Num. of heatmaps ^a	Per. ^b
<i>Nature Biotechnology</i>	81	19	23.46%
<i>Cancer Cell</i>	106	40	37.74%
<i>Genome Research</i>	144	58	40.28%
<i>Genome Biology</i>	92	26	28.26%
<i>Molecular & Cellular Proteomics</i>	241	59	24.48%

To estimate how many papers contain heatmaps, we went through all original research papers (excluding reviews and other articles) published in 2012 of five leading journals as below.

^aNum. of Heatmaps, the number of papers containing with at least one heatmap figure;

^bPer., the percentiles.

doi:10.1371/journal.pone.0111988.t001

Table 2. The summarization of the meth journals.

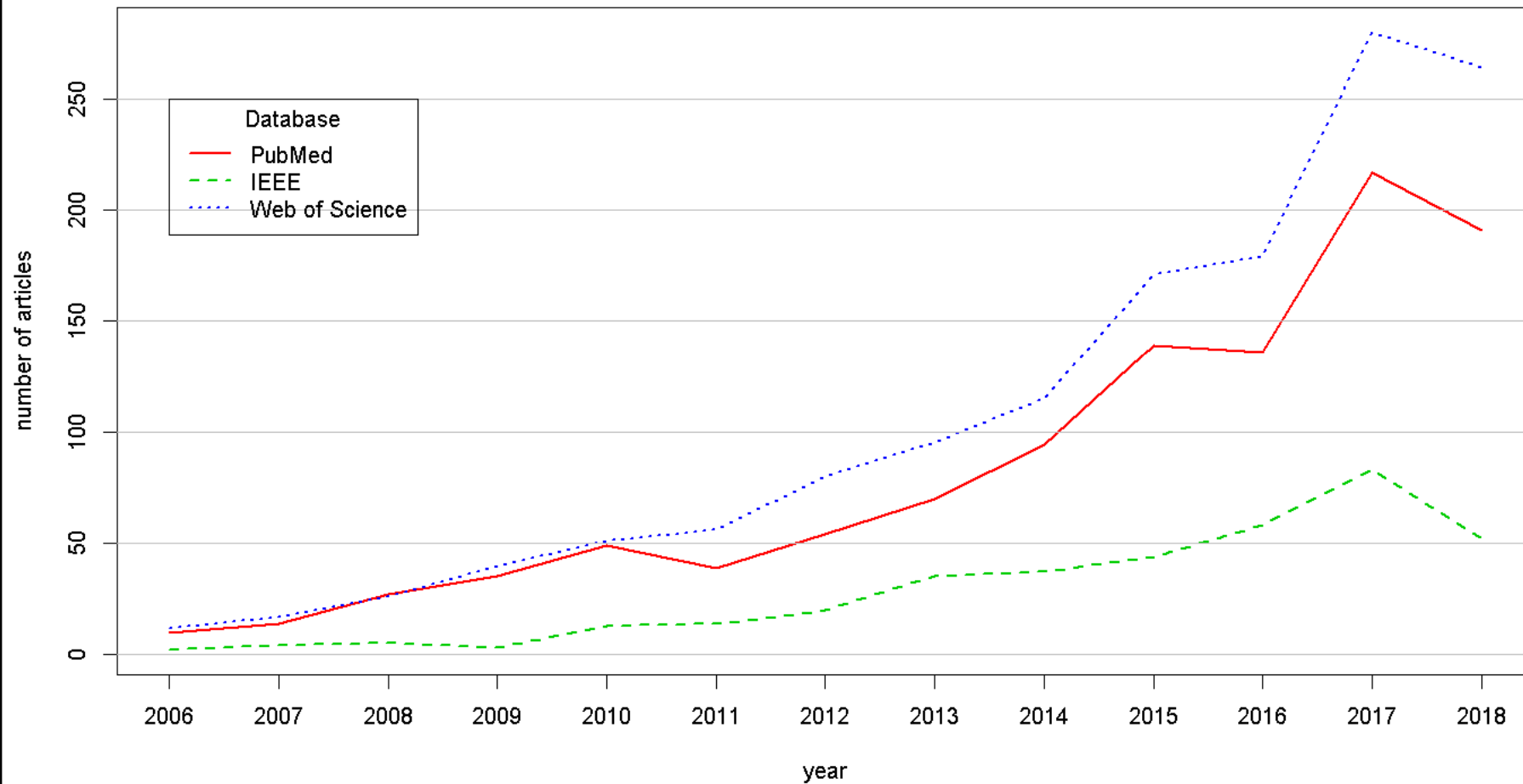
Tools ^a	Num. ^b
R	31
Java Treeview	16
MATLAB	7
SPSS	4
GeneSpring	2

Tools ^a	Num. ^b
MultiExperiment Viewer	2
Cytobank	1
Heatmap Builder	1
Integrative Genomics Viewer	1
Matrix2png	1
Mayday	1
Processing	1
N/A ^d	134
Total	202

Search "heatmap" (title/abstract) in the academic databases

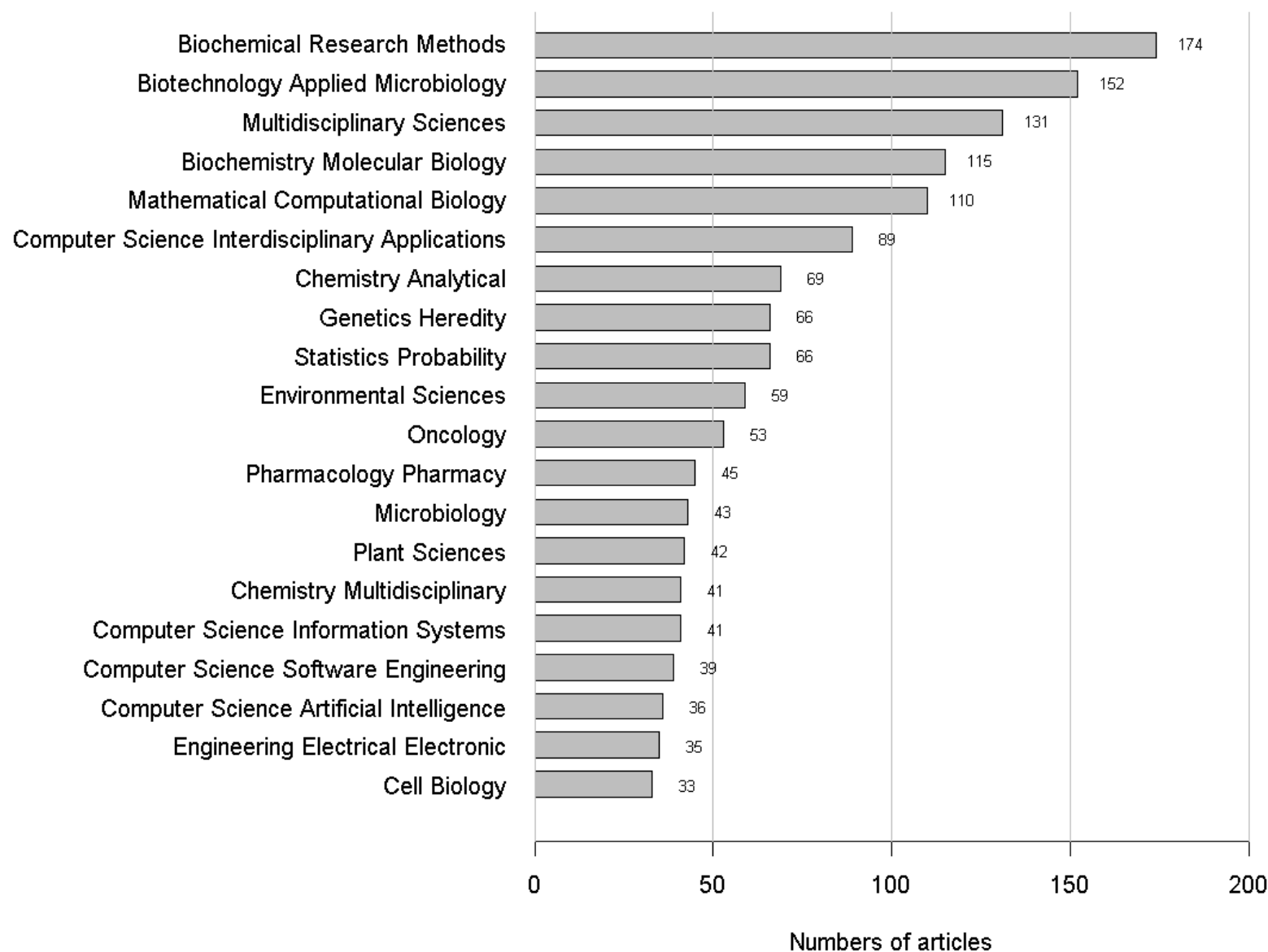


Heatmaps researches/applications





Top 20 Areas (Web of Science)

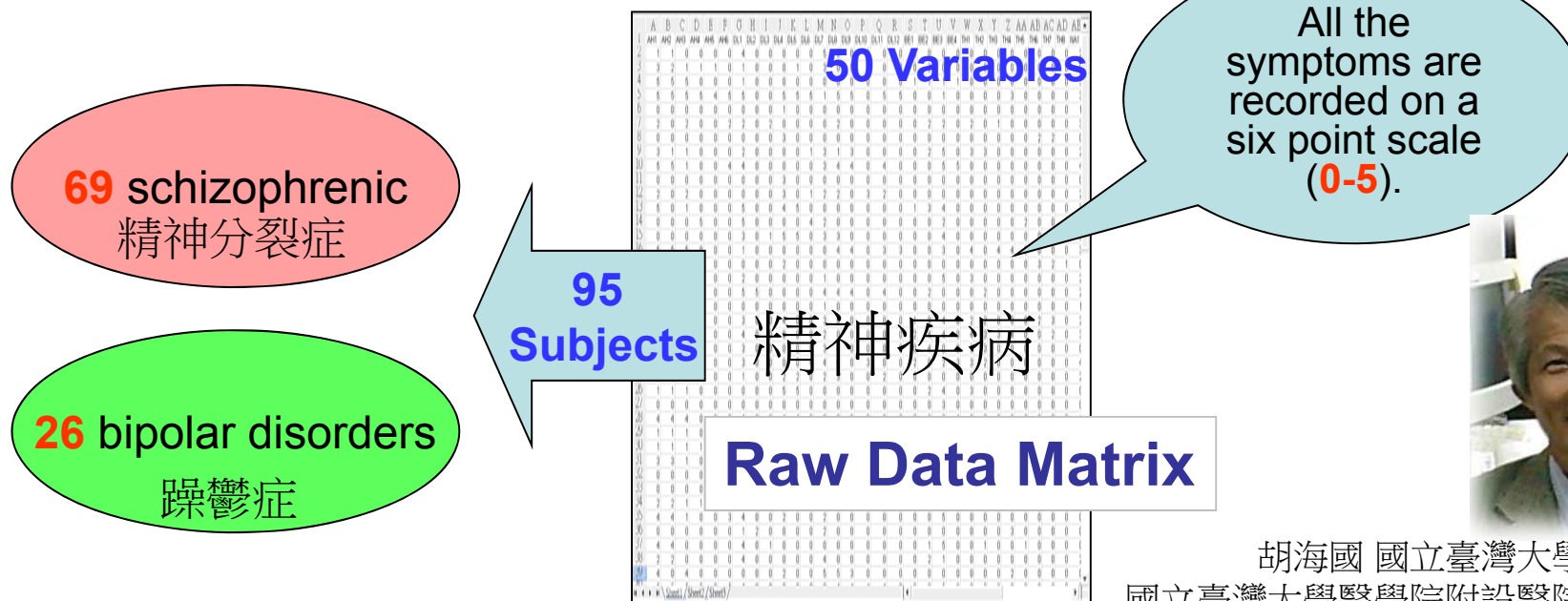
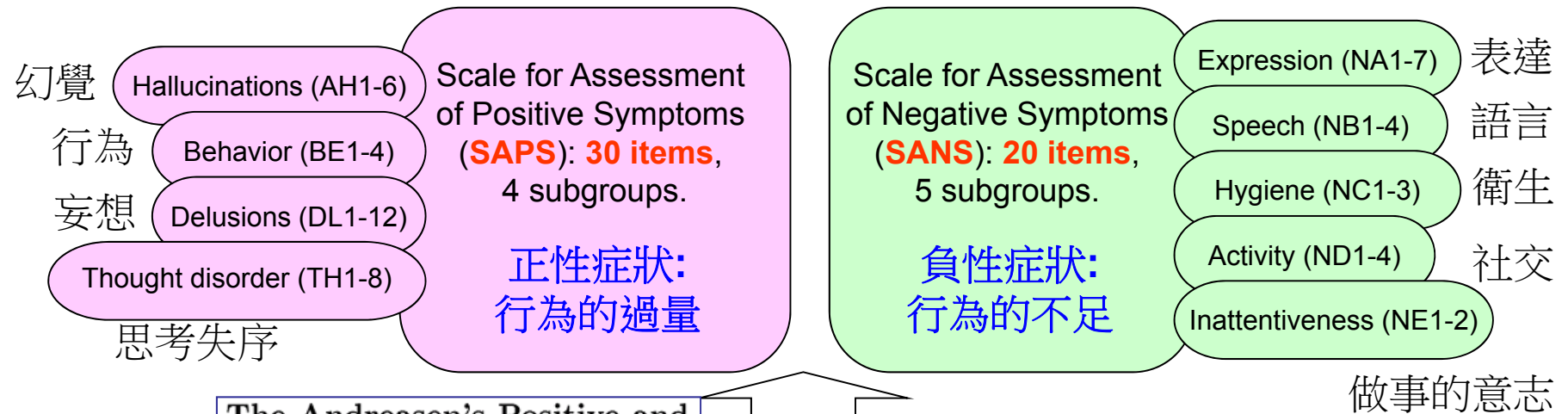


(1) The Basic Principles of Matrix Visualization

Presentation of Raw/Proximity Data Matrix

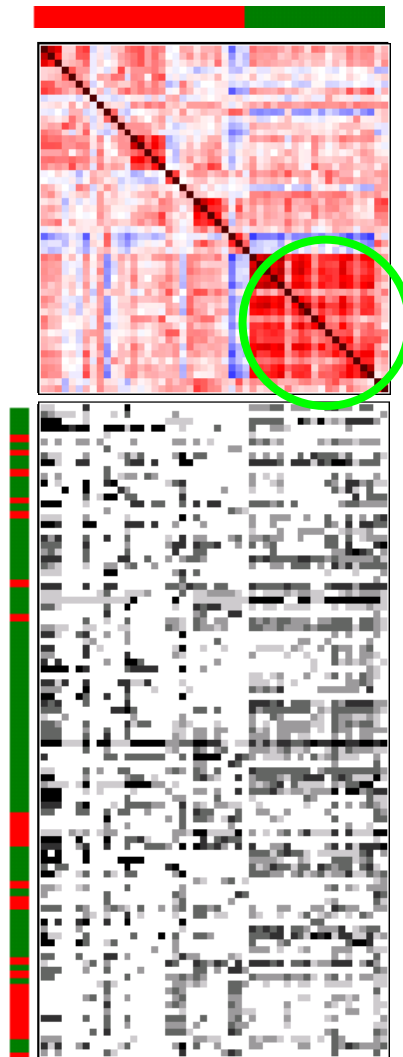
- Data Transformation
- Selection of Proximity Measures
- Color Spectrum
- Display Condition

Psychosis Disorder Data (Chen 2002)

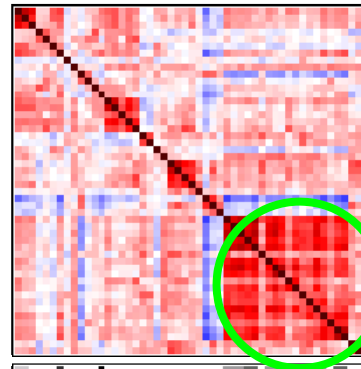


胡海國 國立臺灣大學 精神科教授
國立臺灣大學醫學院附設醫院 精神部主任

Presentation of Raw Data Matrix: Psychosis Disorder Data

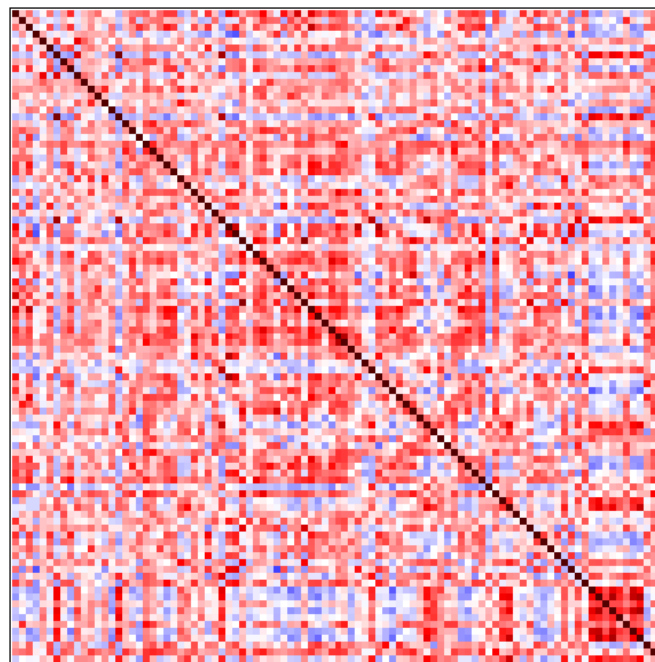


Raw Data Matrix



Correlation Matrix
for Variables

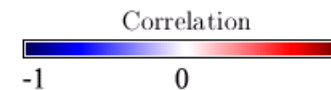
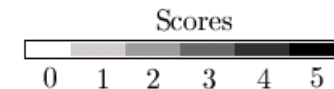
Correlation Matrix for Subjects



(1) Selection of Proximity Measures

Pearson Correlation Coefficient

(2) Color Spectrum



Symptoms

SAPS

SANS

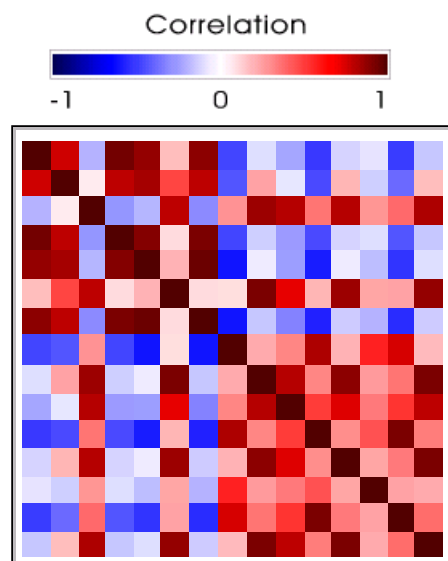
Patients

Schizophrenic

Bipolar disorders

(3) Range Matrix Condition

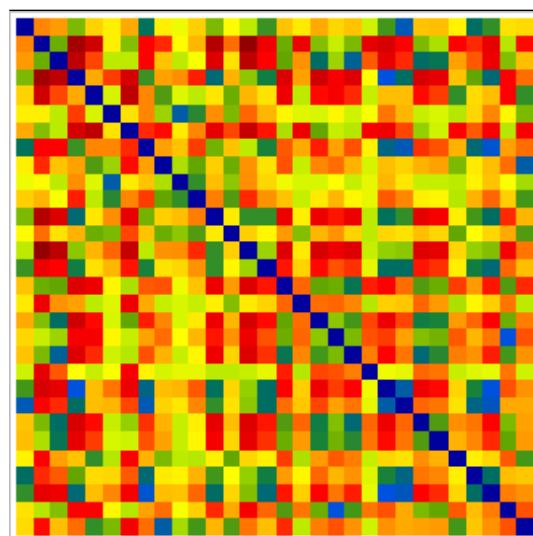
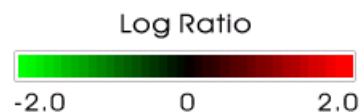
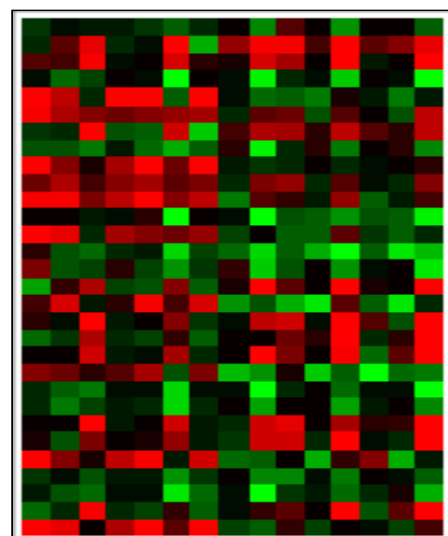
Selection of Proximity Measures



Proximity Matrix for Columns

Pearson Correlation Coefficient

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$



Euclidean Distance

$$d_{xy} = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$$

Other Similarity/Dissimilarity Measures

Proximity Matrix for Rows

The Fundamentals of Constructing and Interpreting Heat Maps

Nathaniel M. Vacanti

16/76

Sarah-Maria Fendt and Sophia Y. Lunt (eds.), Metabolic Signaling: Methods and Protocols, Methods in Molecular Biology, vol. 1862, pp279-291.

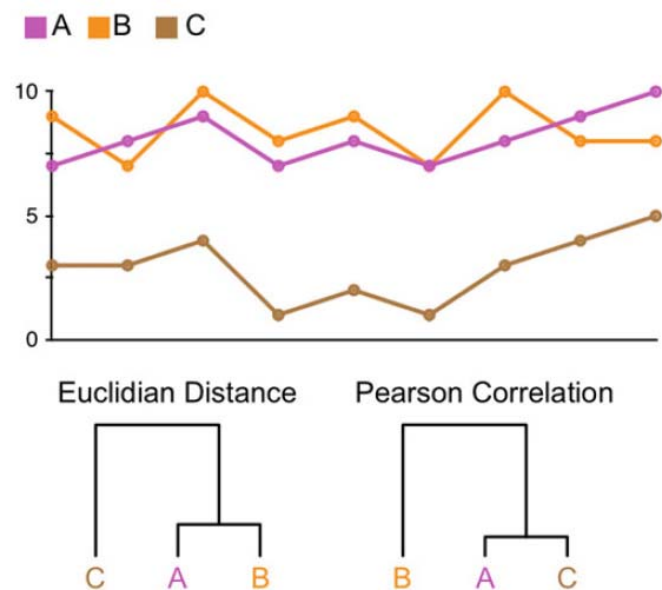


Fig. 5 An illustration of applying Euclidian distance and correlation-based distance methods. Values are fictional and provided for illustrative purposes

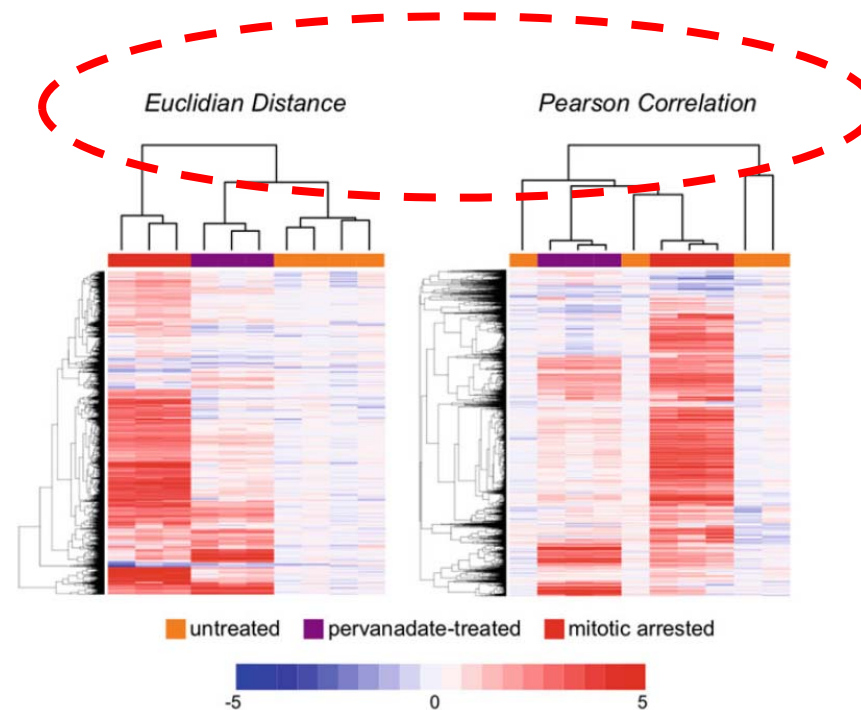
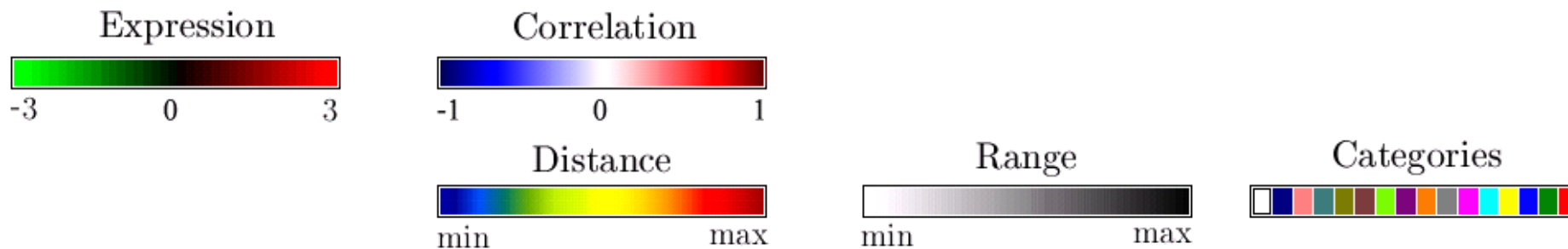
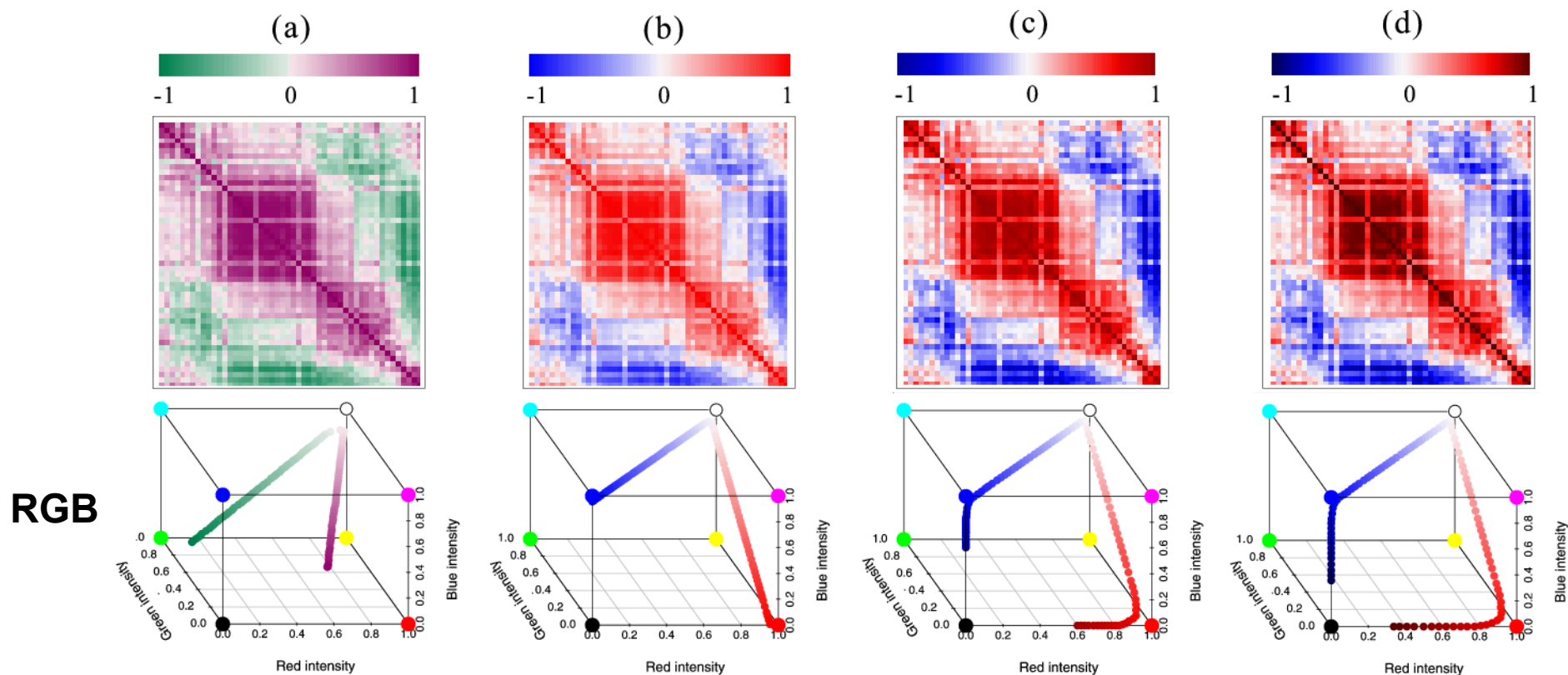


Fig. 6 Heat maps displaying the relative quantified phosphoproteome of HeLa cells under the specified conditions. Each column corresponds to a sample of HeLa cells and each row to a phosphorylated protein. Applied distance methods are provided above the respective heat maps. Complete linkage is applied. Values are normalized to untreated and log₂ transformed as described in [2]

Color Spectra



Correlation matrix map of 50 psychosis disorder variables

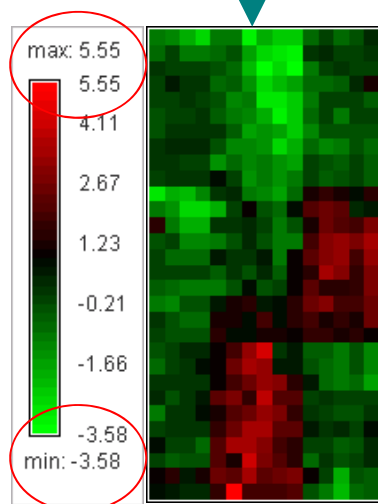


Display Conditions

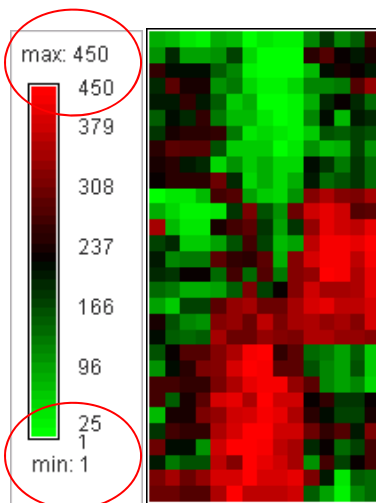
	A	B	C	D	E	F	G	H	I
1	-1.37	-2.30	-1.80	-0.55	2.45	-0.13	1.49	3.03	2.48
2	-0.68	-2.11	-3.42	4.67	4.57	1.75	0.61	0.92	2.52
3	-1.19	-2.49	-3.66	3.14	1.70	3.29	3.33	2.92	2.48
4	-1.93	-2.28	-3.16	2.51	0.32	1.49	0.21	2.20	1.03
5	-2.21	-0.79	-3.29	2.55	2.44	1.45	2.88	3.03	0.19
6	-4.14	-2.91	-1.64	3.21	0.37	1.93	0.14	1.27	2.67
7	0.21	-1.36	-0.44	2.22	1.85	3.11	2.03	0.67	2.40
8	1.13	0.79	2.25	3.65	2.52	2.09	1.13	-2.59	0.67
9	0.95	2.33	-0.07	3.89	2.72	2.13	1.75	-2.17	-0.90
10	3.04	1.85	0.21	7.07	2.01	3.05	0.76	-2.58	-1.04
11	-1.02	1.65	1.53	0.95	0.60	3.12	2.52	-0.77	-1.40
12	1.21	0.24	1.04	2.50	3.69	1.81	3.98	-0.33	0.11
13	1.74	1.60	1.70	2.02	3.45	4.46	2.69	0.41	-0.09
14	1.34	1.06	0.06	1.81	2.90	3.64	3.04	0.49	-2.33
15	0.57	1.81	-0.47	1.40	2.70	0.99	0.82	-1.61	-2.56
16	0.61	4.22	-2.03	-2.61	-4.00	-4.64	-2.92	1.55	-0.71
17	-1.13	1.64	0.01	-1.77	-2.85	-1.24	-3.41	-0.59	-1.64
18	-0.86	-1.17	-0.41	-2.20	-1.30	-2.37	-1.41	0.08	0.25
19	0.75	0.66	1.04	-4.26	-1.41	-3.99	-3.53	-2.17	0.34
20	0.15	0.68	3.18	-2.86	-2.01	-3.18	-1.58	0.10	1.28

Gene Expression

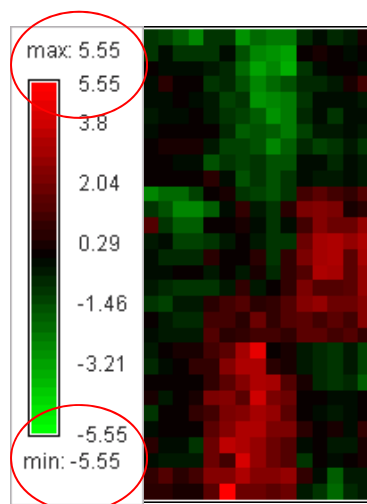
Down-regulated no differential expressed Up-regulated



Range Matrix Condition

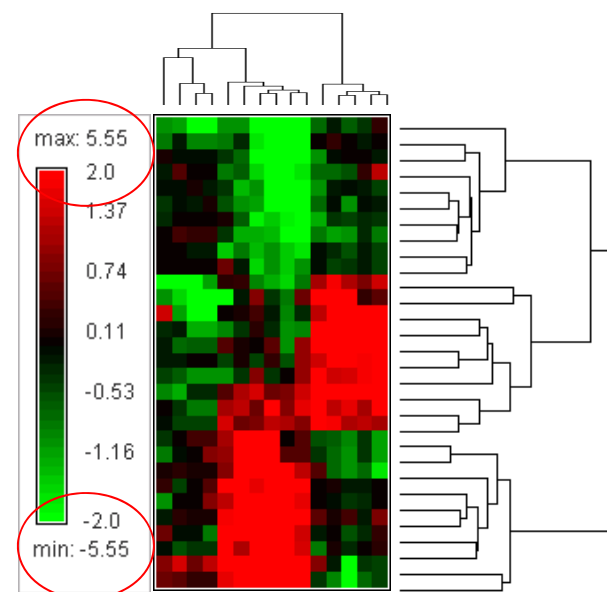


Rank Matrix Condition



Center Matrix Condition

range column condition
range row condition
center column condition
center row condition



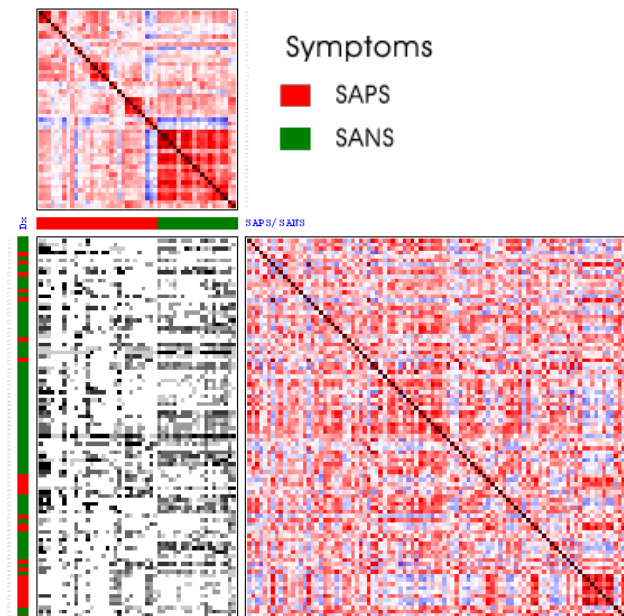
(2) The Basic Principles of Matrix Visualization

Seriation of Proximity Matrices and Raw Data Matrix

- Relativity of a Statistical Graph
- Global Criterion
 - Anti-Robinson Measurements
 - GAP Rank-Two Elliptical Seriation
- Local Criterion
 - Minimal Span Loss Function
 - Tree Seriation
 - Flipping of Tree Intermediate Nodes

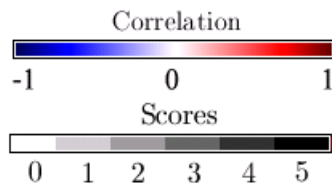
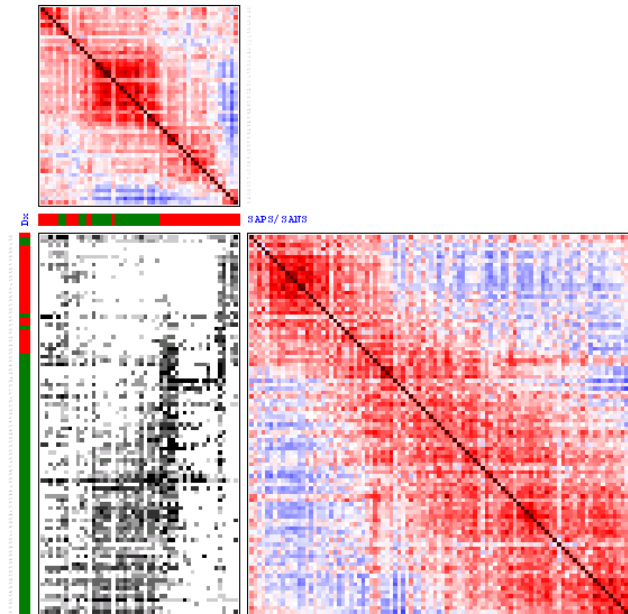
Relativity of a Statistical Graph

Placing similar objects at closer positions.
Placing different objects at distant positions.



**Seriation
Methods**

(1) Rank Two Ellipse Ordering
(Chen, 2002)

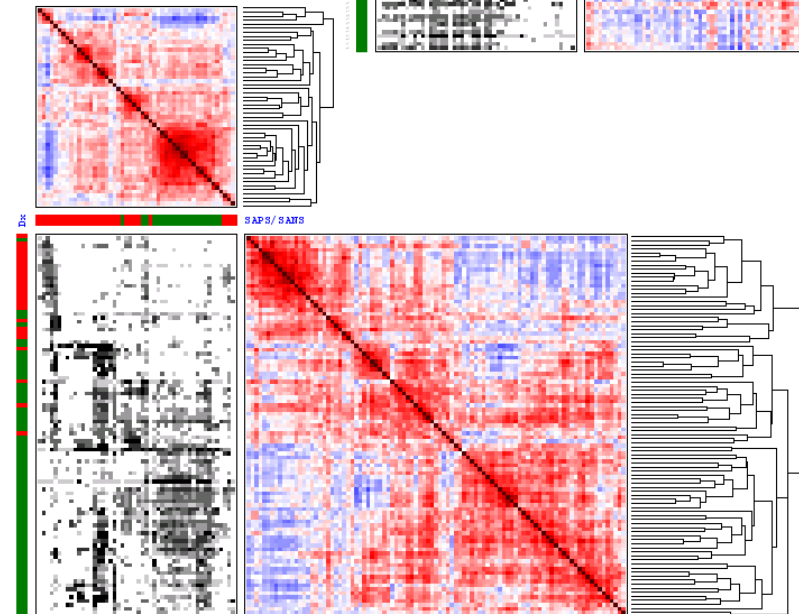


Patients

- Schizophrenic
- Bipolar disorder

**Seriation
Methods**

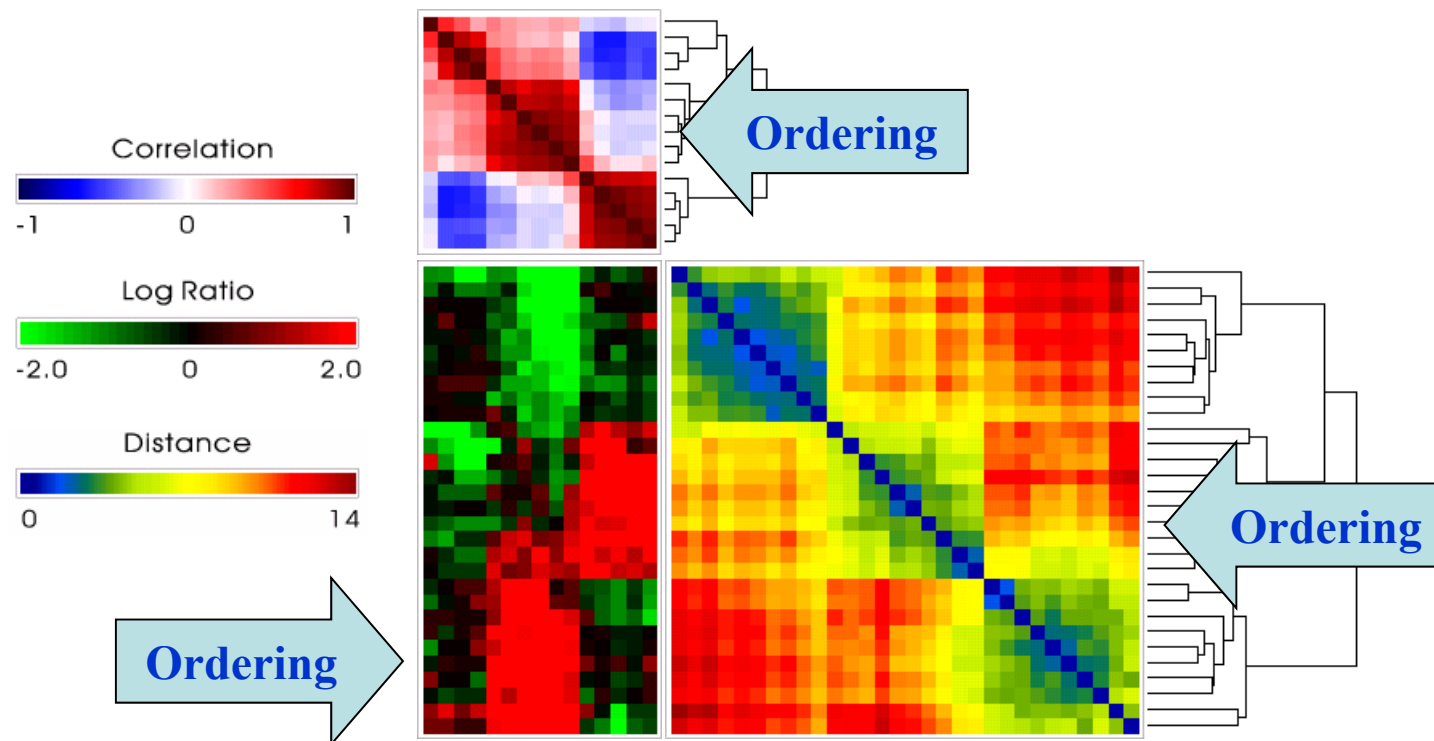
(2) Hierarchical Clustering
Tree (Average-Linkage)



Relativity of a Statistical Graph

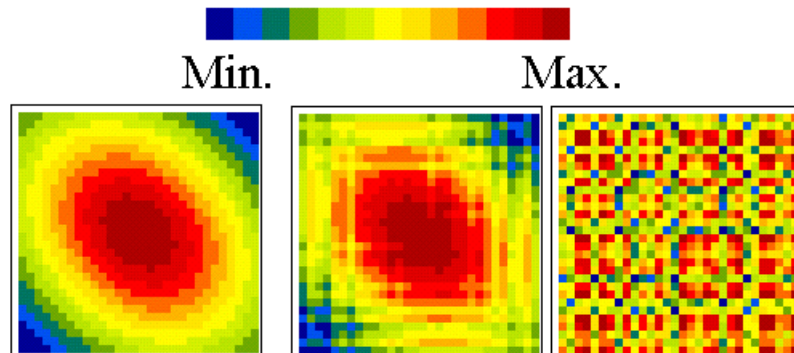
Placing similar (different) objects at closer (distant) positions

Without suitable permutations (orderings) of the variables and samples, matrix visualization is of no practical use in visually extracting information.



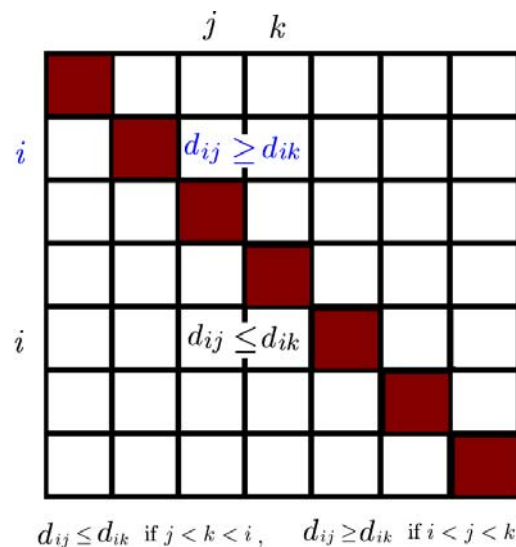
Criteria for a *good* Permutation

Global criterion: Anti-Robinson Measurements



Robinson pre-Robinson

When T is symmetric, we usually want T' to approximate a Robinson form (Robinson (1951)).

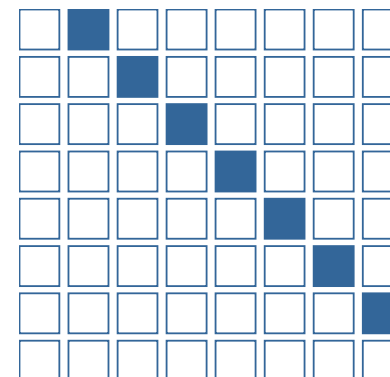


$$AR(i) = \sum_{i=1}^p \left[\sum_{j < k < i} I(d_{ij} < d_{ik}) + \sum_{i < j < k} I(d_{ij} > d_{ik}) \right],$$

$$AR(s) = \sum_{i=1}^p \left[\sum_{j < k < i} I(d_{ij} < d_{ik}) \cdot |d_{ij} - d_{ik}| + \sum_{i < j < k} I(d_{ij} > d_{ik}) \cdot |d_{ij} - d_{ik}| \right],$$

$$AR(w) = \sum_{i=1}^p \left[\sum_{j < k < i} I(d_{ij} < d_{ik}) |j - k| |d_{ij} - d_{ik}| + \sum_{i < j < k} I(d_{ij} > d_{ik}) |j - k| |d_{ij} - d_{ik}| \right].$$

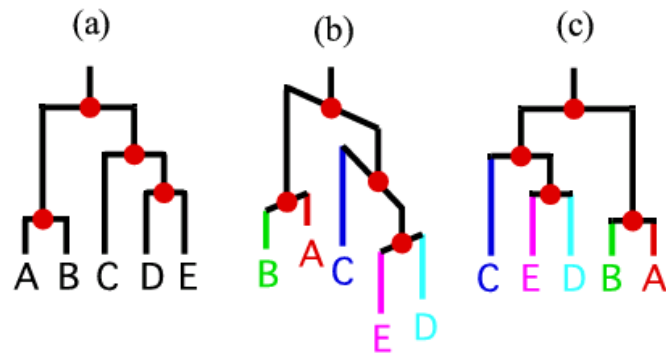
Local criterion: Minimal Span Loss Function



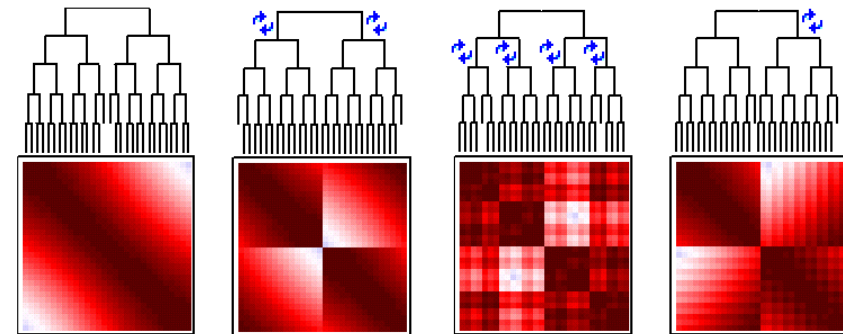
$$MS = \sum_{i=1}^{n-1} d_{i,i+1}$$

Different Seriations Generated from Identical Tree Structure

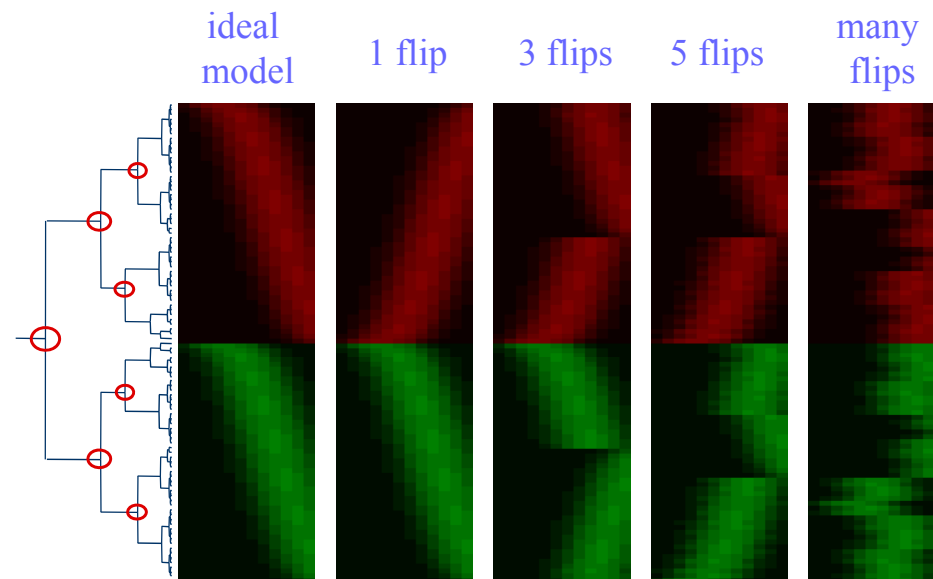
Tree seriation



Tree seriation for proximity matrices



Tree seriation for raw data matrices



Tien, Y. J., Lee, Y. S, Wu, H. M. and Chen, C. H.* (2008), Methods for Simultaneously Identifying Coherent **Local Clusters with Smooth Global Patterns** in Gene Expression Profiles. BMC Bioinformatics 9:155, 1-16.

Data: 517 genes by 13 arrays

GAP Rank-two elliptical seriation

Michael Eisen (1998) tree seriation

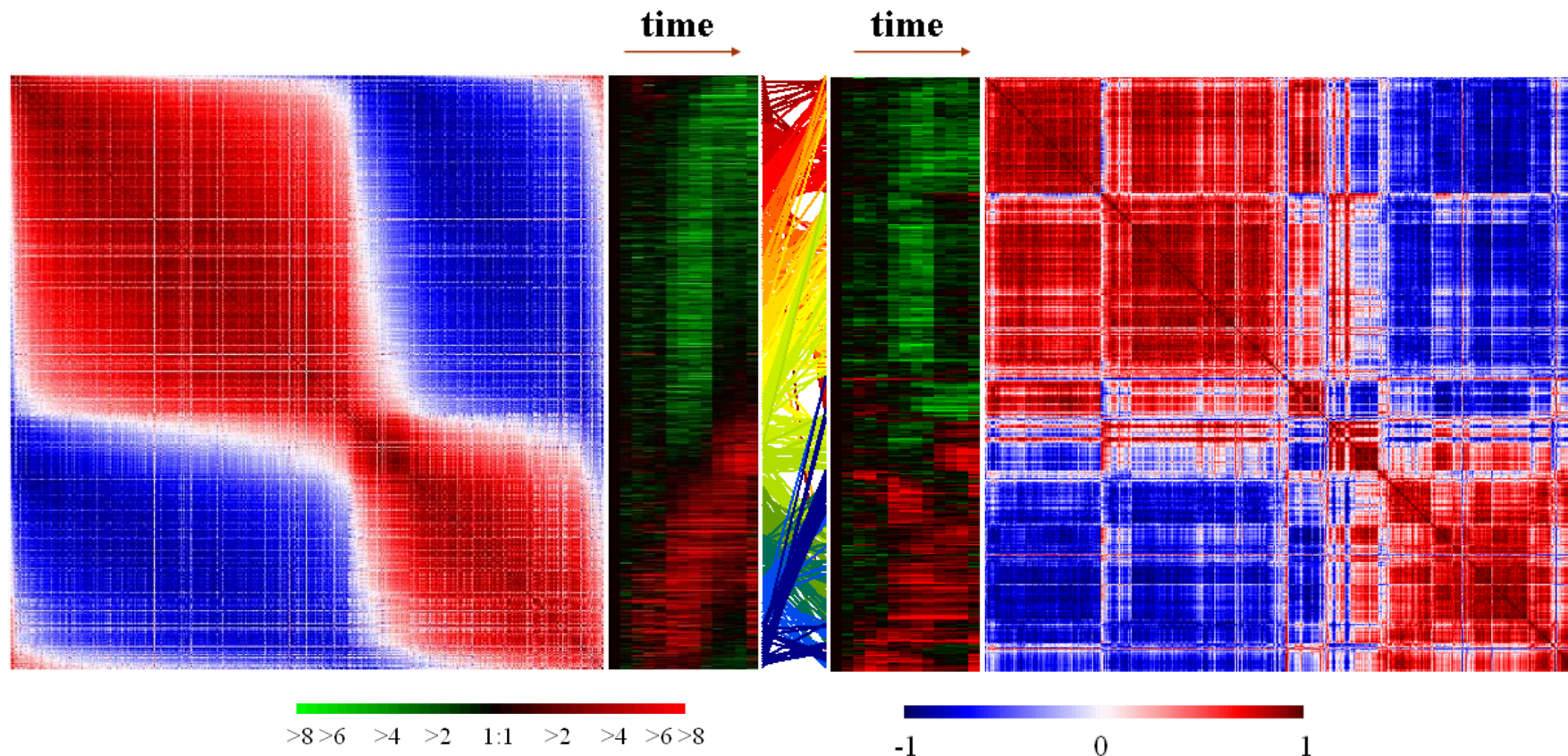
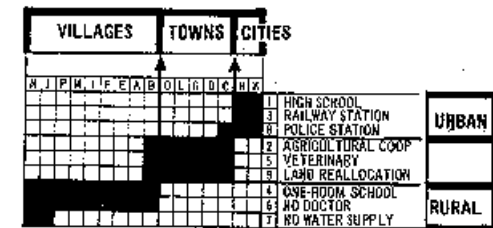
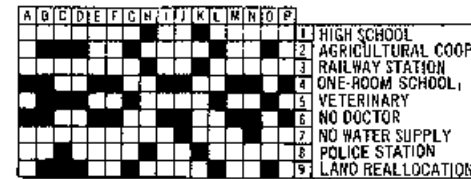


Image source: Dr. Chen Chun-houh's slide

Literature review (1)

Concept:

- Bertin (1967): reorderable matrix.
- Carmichael and Sneath (1969): taxometric maps.

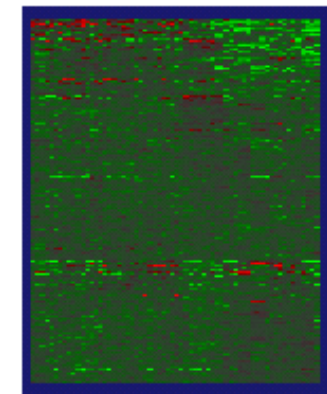


Clustering of data arrays:

- Hartigan (1972): direct clustering of a data matrix.
- Tibshirani (1999): block clustering.
- Lenstra (1974): traveling-salesman problem
- Slagle *et al.* (1975): shortest spanning path.

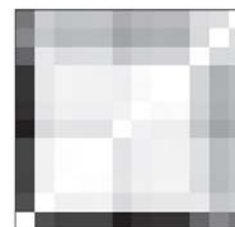
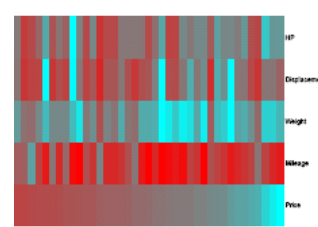
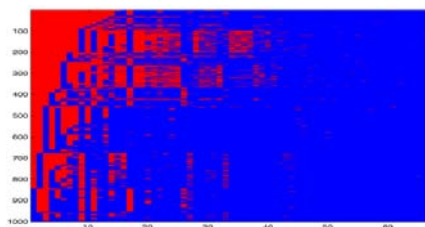
4. UN VOTES IN 1969-1970*

State	BASE	HUNG	CHINA	KOREA	SO AF	PAPUA
	1 2 3 4	5 6 7 8	9 10 11 12	13 14		
USA	1 1 1 1	3 1 2 3	3 1 3	2 2	1 3	
BGA	1 1 1 1	3 1 1 3	1 3	2 2	1 3	
YUG	1 3 3 3	3 1 1 3	1 2	3 1 1	2	
SYR	1 2 2 2	3 1 1 3	1 2	3 1 1	3	
UAR	1 3 3 3	3 1 1 3	2 2	3 1 1	3	
KEN	1 3 3 3	3 1 1 3	2 5	3 1 1	3	
TAN	1 2 2 2	3 1 1 3	2 5	3 1 1	3	
SEN	1 3 3 3	1 2 2 2	2 1	3 1 1	2	
DAH	1 3 3 3	1 3 1 3	1 3 1	2 3		
USA	1 3 3 3	1 3 3 1	3 1 1	3 3 1		
UNK	1 3 3 3	1 1 3 2	3 1 1	3 3 1		
FRA	1 3 3 3	3 1 2 3	3 1 1	3 2 2		
SWE	1 3 3 3	3 1 2 3	3 1 1	3 3 1		
NOR	1 3 3 3	3 1 3 2	3 1 1	3 3 1		
ATA	1 3 3 3	1 3 1 3	3 1 1	3 3 1		
NZ	1 3 3 3	1 3 1 1	3 1 1	3 3 1		
MEX	1 2 2 2	1 3 3 1	3 1 1	1 2		
VEN	1 2 2 2	1 3 3 1	3 1 1	1 1		
BRA	1 2 2 2	1 3 3 1	3 1 1	1 1		



Colour Representation:

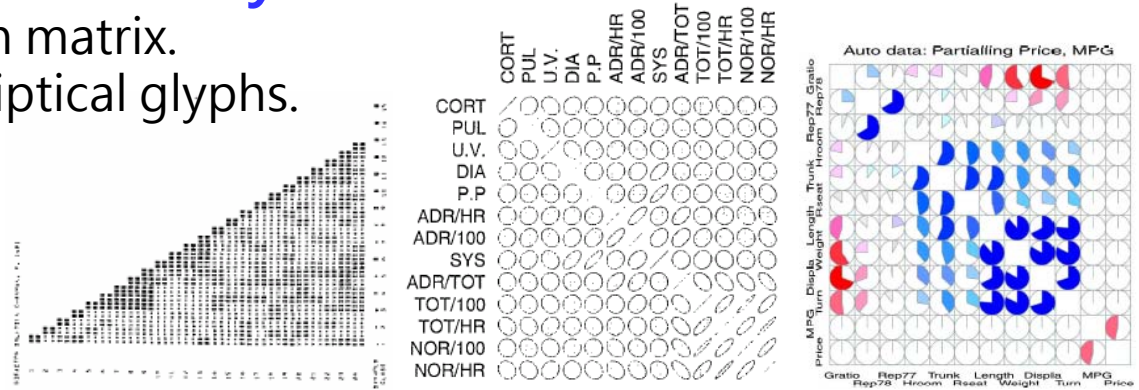
- Wegman (1990): colour histogram.
- Minnotte and West (1998): data image.
- Marchette and Solka (2003): outlier detection.



Literature review (2)

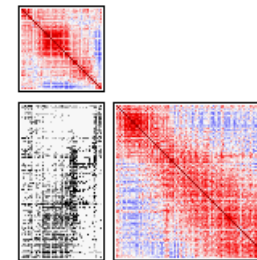
Exploring proximity matrices only:

- Ling (1973): shaded correlation matrix.
- Murdoch and Chow (1996): elliptical glyphs.
- Friendly (2002): corrgrams.



Integration of raw data matrix with two proximity matrices

- Chen (1996, 1999, and 2002): generalized association plots (GAP).



Reordering of variables and samples

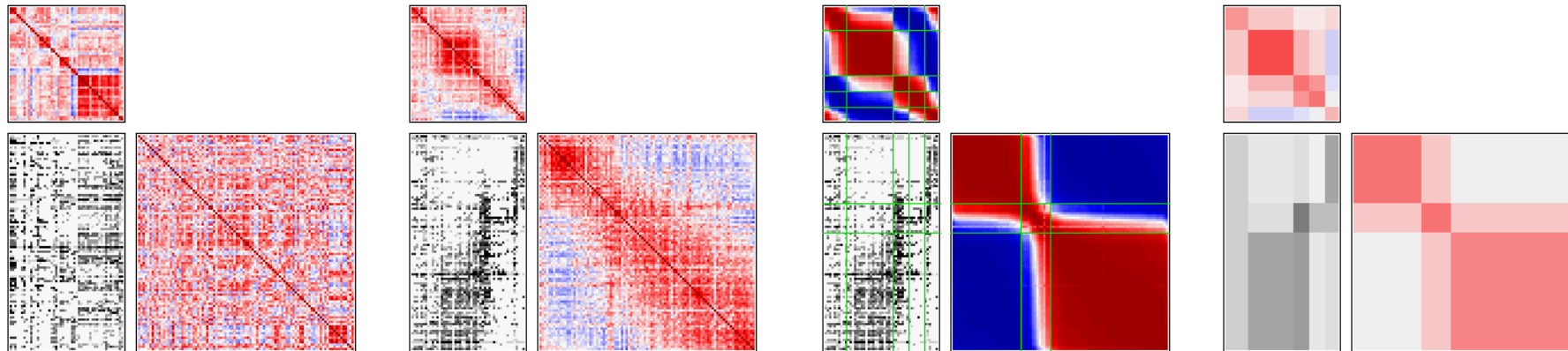
- Chen (2002): concept of relativity of a statistical graph.
- Friendly and Kwan (2003): effect ordering of data displays.
- Hurley (2004): placing interesting displays in prominent positions.

Matrix Visualization (MV): reorderable matrix, the heatmap, color histogram, data image and matrix visualization.

Generalized Association Plots (GAP)

(Chen, 2002)

Four Steps of Generalized Association Plots (GAP)



(1)
Presentation
呈現

(2)
Seriation
排序

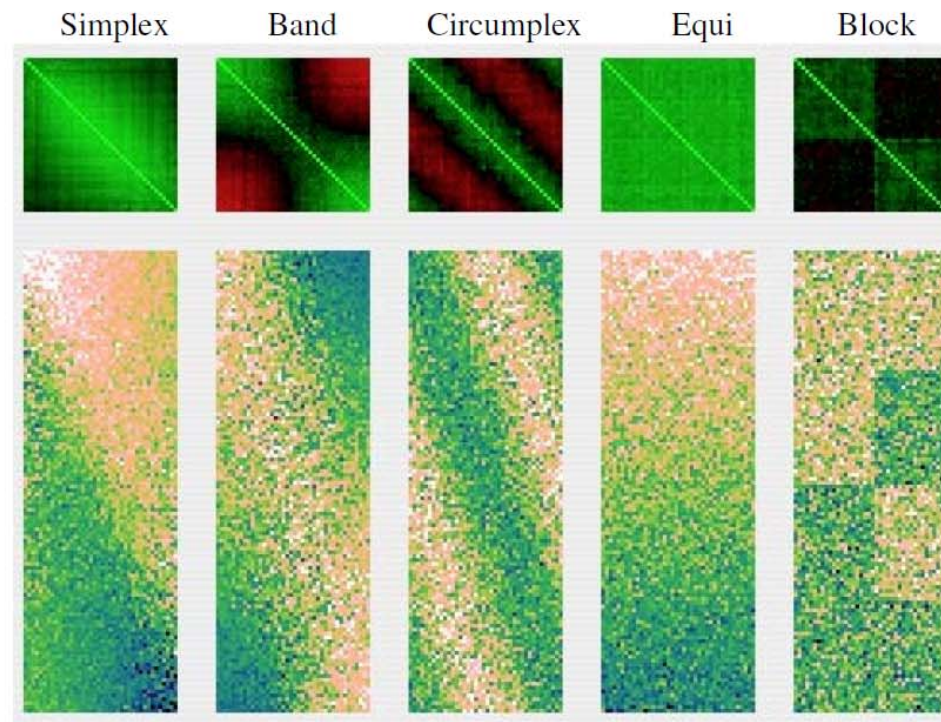
(3)
Partition
分割

(4)
Sufficient
充分

Clustering

Summarization

What kind of data have simple permutations?



These five patterns occur frequently in scientific datasets. Above each matrix is a heatmap of the covariance matrix on the columns (red for negative correlations, black for zero, green for positive). All except Block are (within error) topologically one-dimensional. Simplex maps to a spiral, band maps to a line, Circumplex maps to a circle, and Equi maps to a line. Block is two-dimensional.

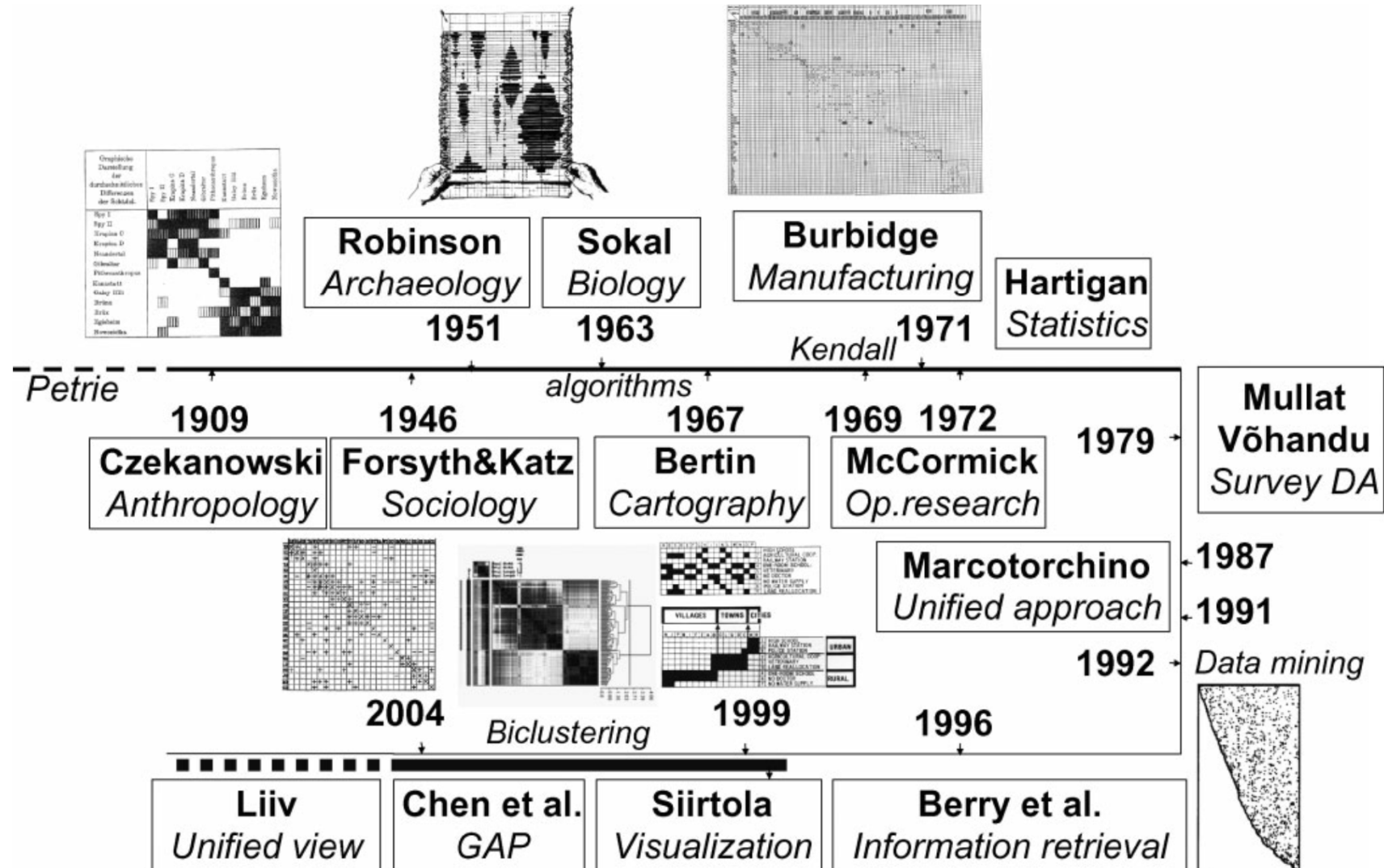
REVIEW

Seriation and Matrix Reordering Methods: An Historical Overview

Innar Liiv*

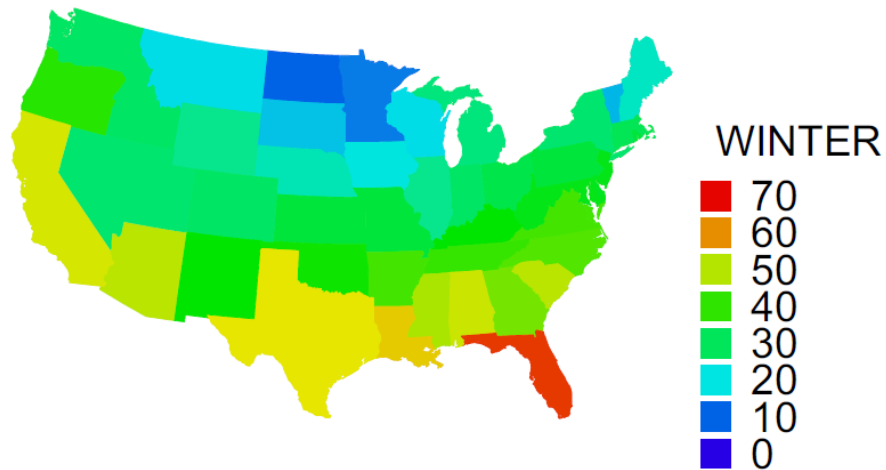
Statistical Analysis and Data Mining, Vol. 3 (2010)

Department of Informatics, Tallinn University of Technology, Tallinn, Estonia



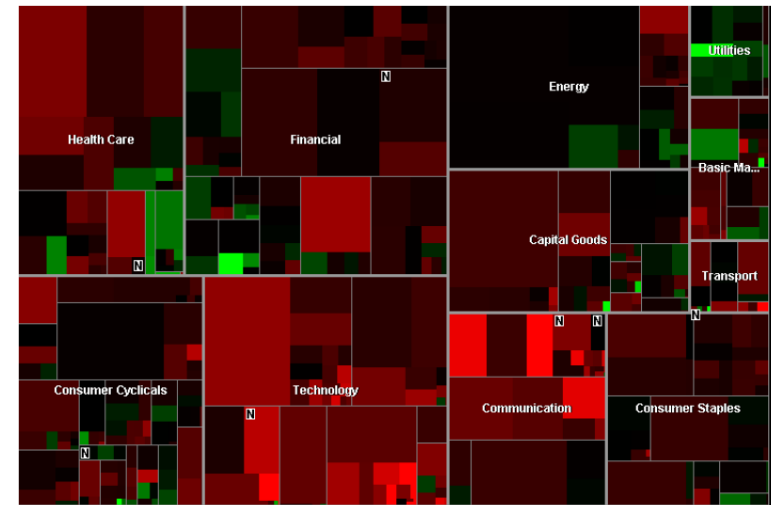
Applications: Other types of MV

Heat Map (Old, very old)



Geography's not a bad way to order the world. Aggregating over states, however, makes less sense for temperature than for electoral college data!

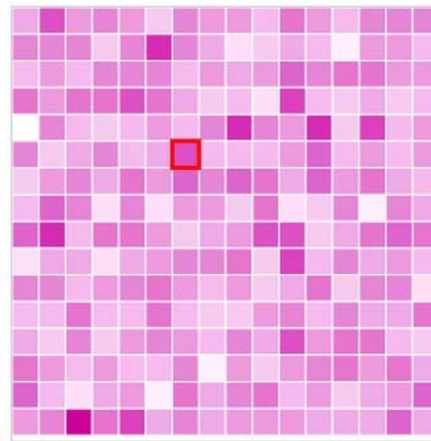
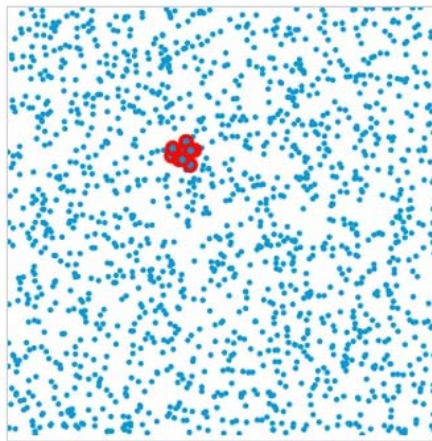
Treemap (Schneiderman, 1992)



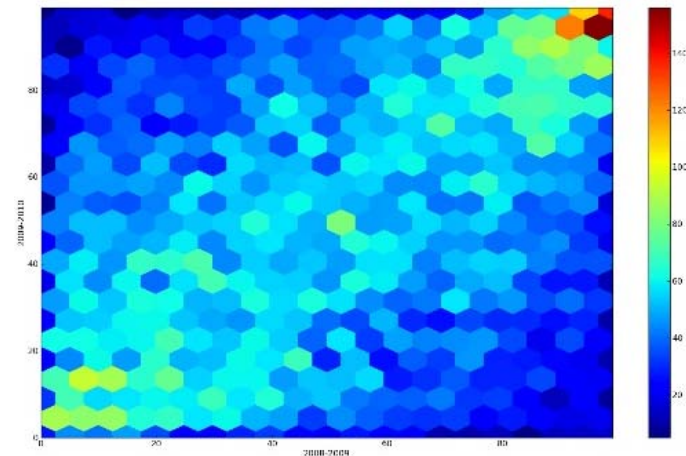
Applications: Binning Technique

- Binning is a technique of data aggregation used for grouping a dataset of N values into less than N discrete groups.
 - the XY plane is uniformly tiled with polygons (squares, rectangles or hexagons).
 - the number of points falling in each bin (tile) are counted and stored in a data structure.
 - the bins with count > 0 are plotted using a color range (heatmap) or varying their size in proportion to the count.

Rectangular binning



Hexagonal binning



<http://www.meccanismocomplesso.org/en/hexagonal-binning/>

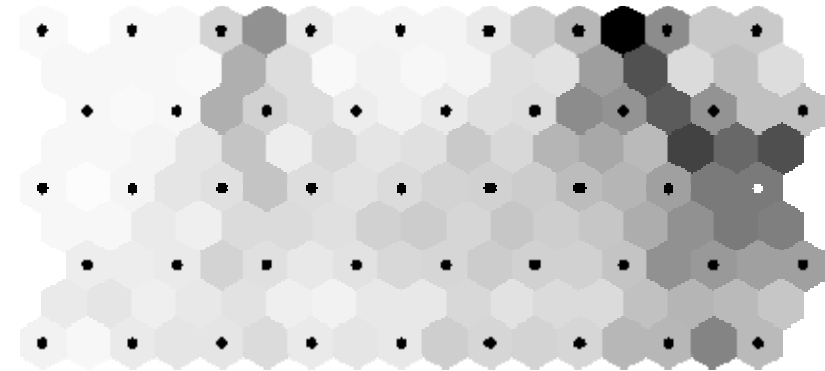
hexagonal heatmap in R

<https://www.visualcinnamon.com/2013/11/how-to-create-hexagonal-heatmap-in-r>

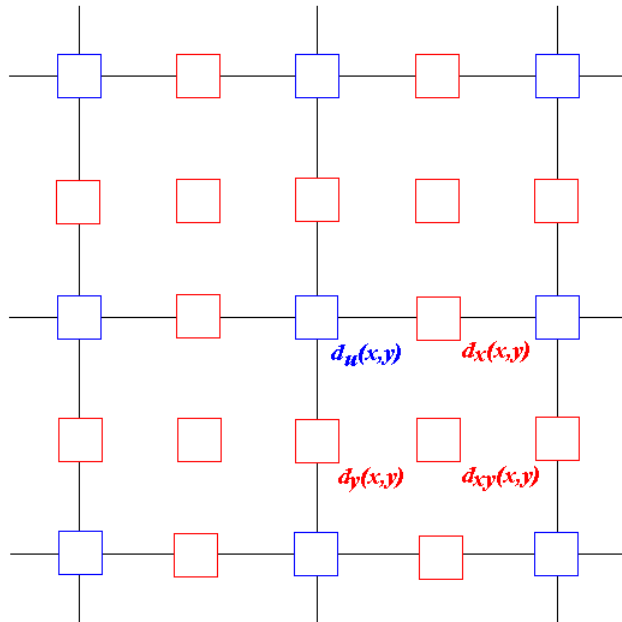
U-matrix: Unified Matrix Method

(Ultsch and Siemon 1989, Ultsch 1993)

U-matrix representation of SOM visualizes the distance between the neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent nodes.



U-matrix representation of the SOM



$b(x, y)$: matrix of neurons, of size $n_x \times n_y$.

$w_i(x, y)$: matrix of weights.

$u(x, y)$: U-matrix of size $(2n_x - 1) \times (2n_y - 1)$.

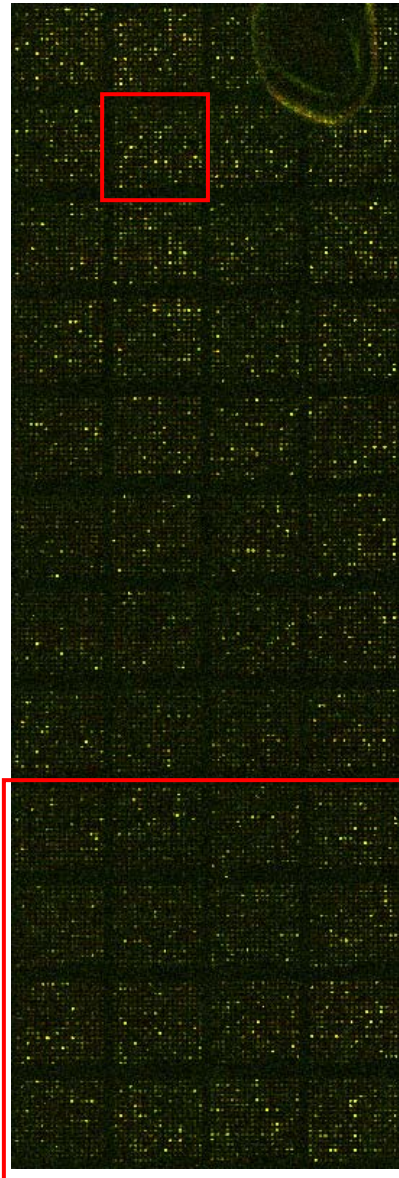
$$d_x(x, y): \|b(x, y) - b(x + 1, y)\| = \sqrt{\sum_i [w_i(x, y) - w_i(x + 1, y)]^2}$$

$$d_y(x, y): \|b(x, y) - b(x, y + 1)\| = \sqrt{\sum_i [w_i(x, y) - w_i(x, y + 1)]^2}$$

$$d_{xy}(x, y): \frac{1}{2} \left[\frac{\|b(x, y) - b(x + 1, y + 1)\|}{\sqrt{2}} + \frac{\|b(x, y + 1) - b(x + 1, y)\|}{\sqrt{2}} \right]$$

$d_u(x, y)$: the median of the surrounding elements.

Applications: Array Image



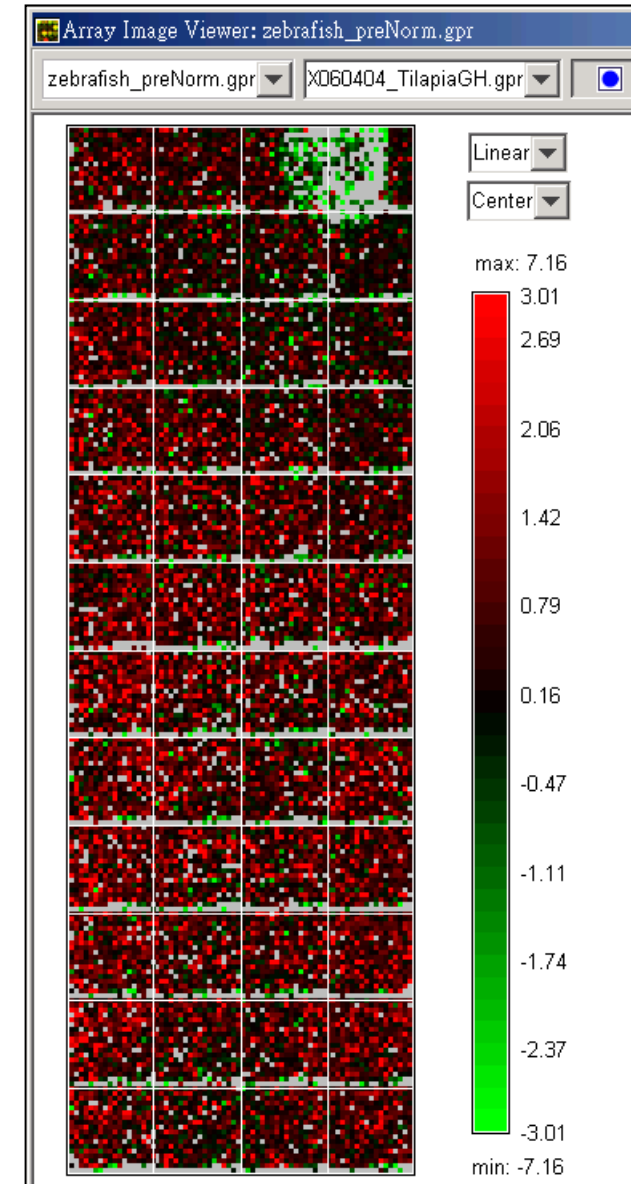
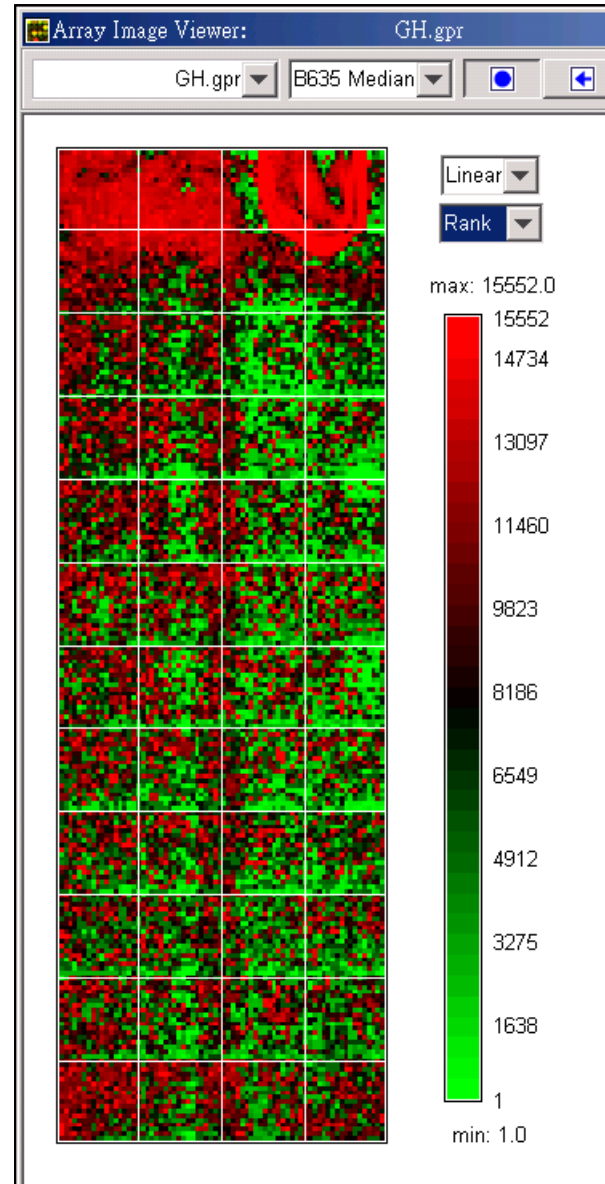
Blocks:
12 by 4

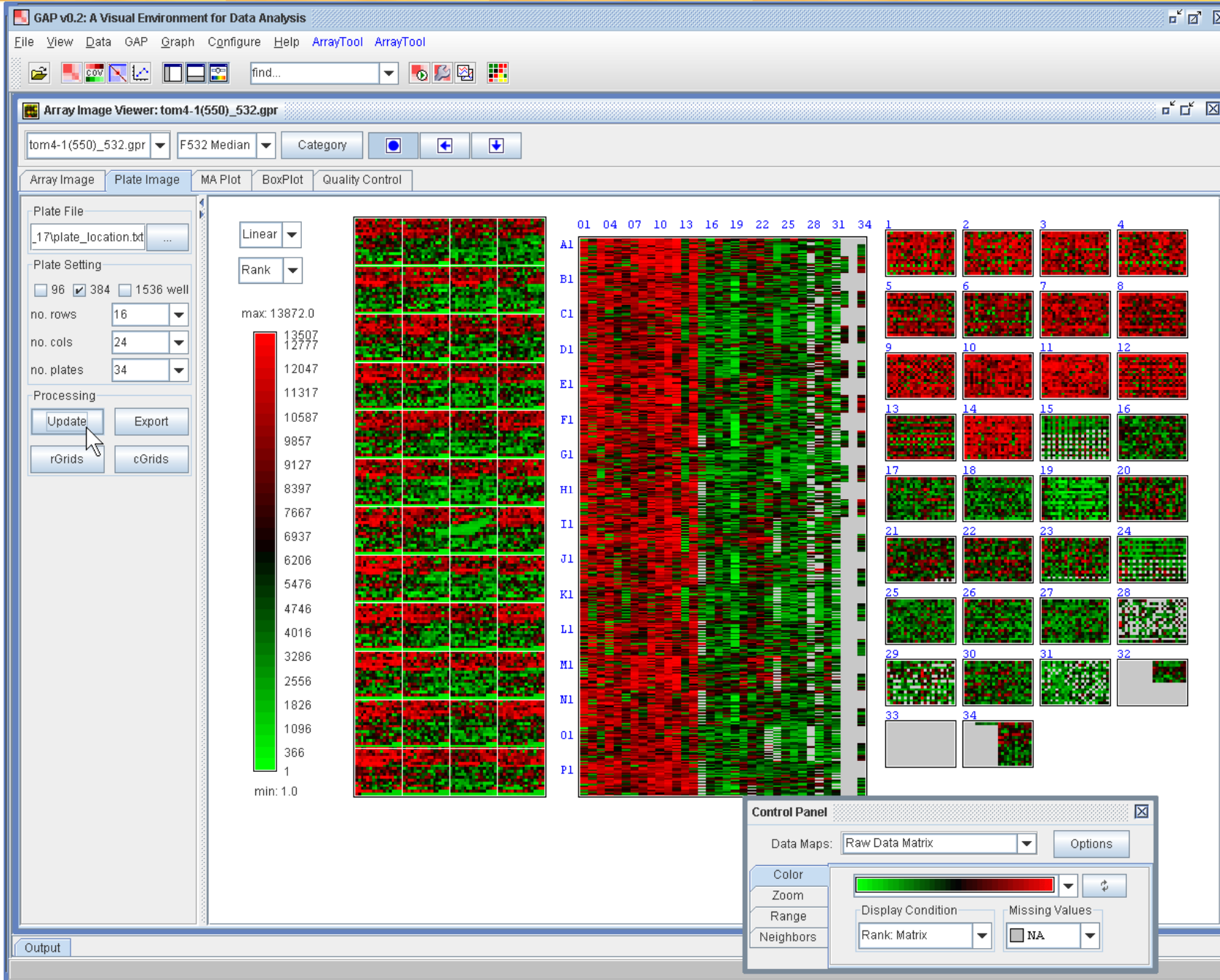
Features:
18 by 18

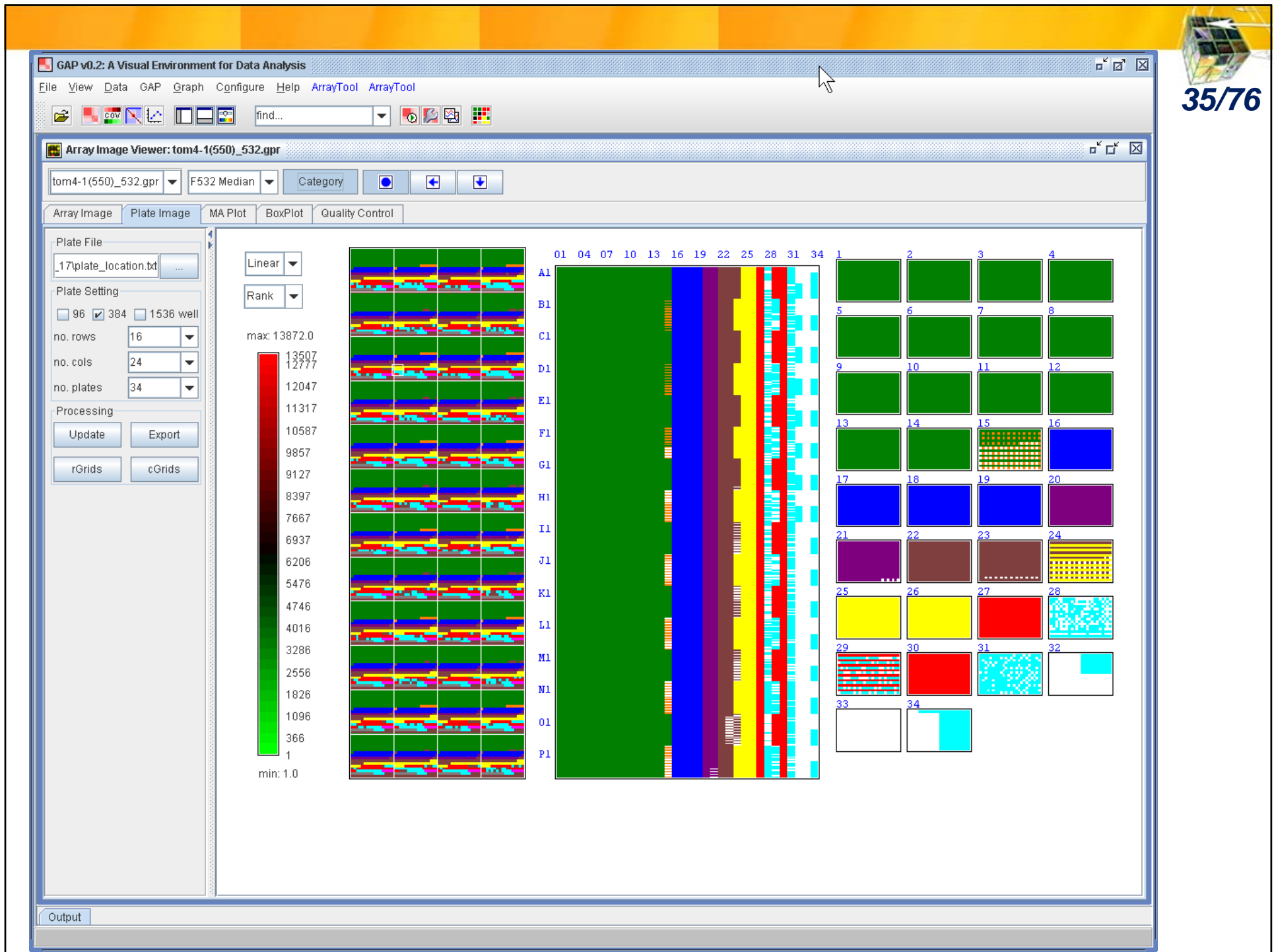
Signal
16-bit
0~65535

*.gpr

GAL

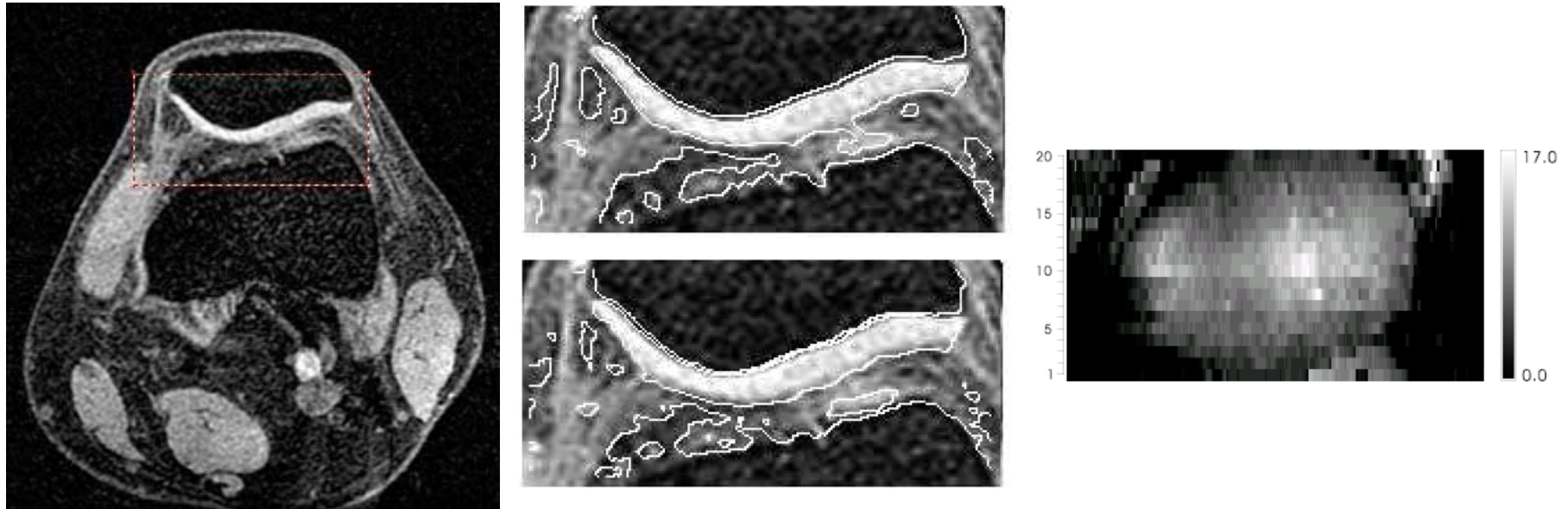






Applications: Image Reconstruction

Medical images (fMRI) of a knee



The cartilaginous tissues (the brighter part) is the object-of-interest.

Applications: Eye-tracking, mouse clicking

By Mike Bartels

eye-tracking research

Beyond the heat map

There is no doubt that eye-tracking research has begun. What was once a niche activity, comprised of a few esoteric tools used in many diverse applications, is now a widely used tool. New software have improved the accuracy of the data and new software has eased the process of analyzing the data. Coupled with a growing interest in eye-tracking, it has inspired a new generation of researchers. Of all of the industries that have been more enthusiastic than the fields of marketing, advertising, and web development, it has been the most.

Principles for interpreting eye-tracking data

See also: <https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/interpret/working-with-heat-maps-and-gaze-plots/>

How does this tool get us any closer to understanding our potential customers?



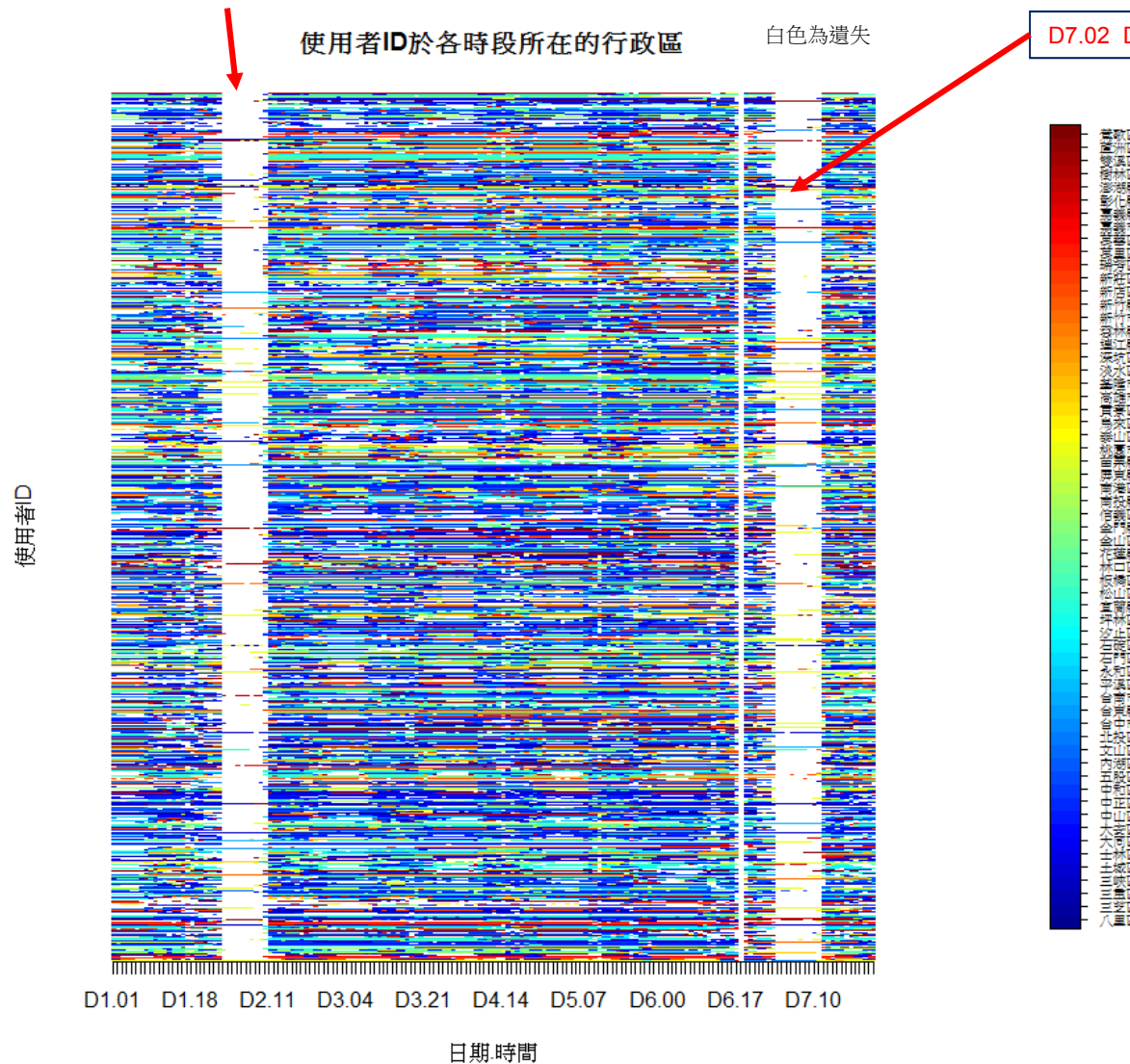
Applications

D2.01 D2.02 D2.03 D2.04 D2.05 D2.06 D2.07 D2.08 D2.09 D2.10

使用者ID於各時段所在的行政區

白色為遺失

D7.02 D7.03 D7.04 D7.05 D7.06 D7.07 D7.08 D7.09 D7.10 D7.11

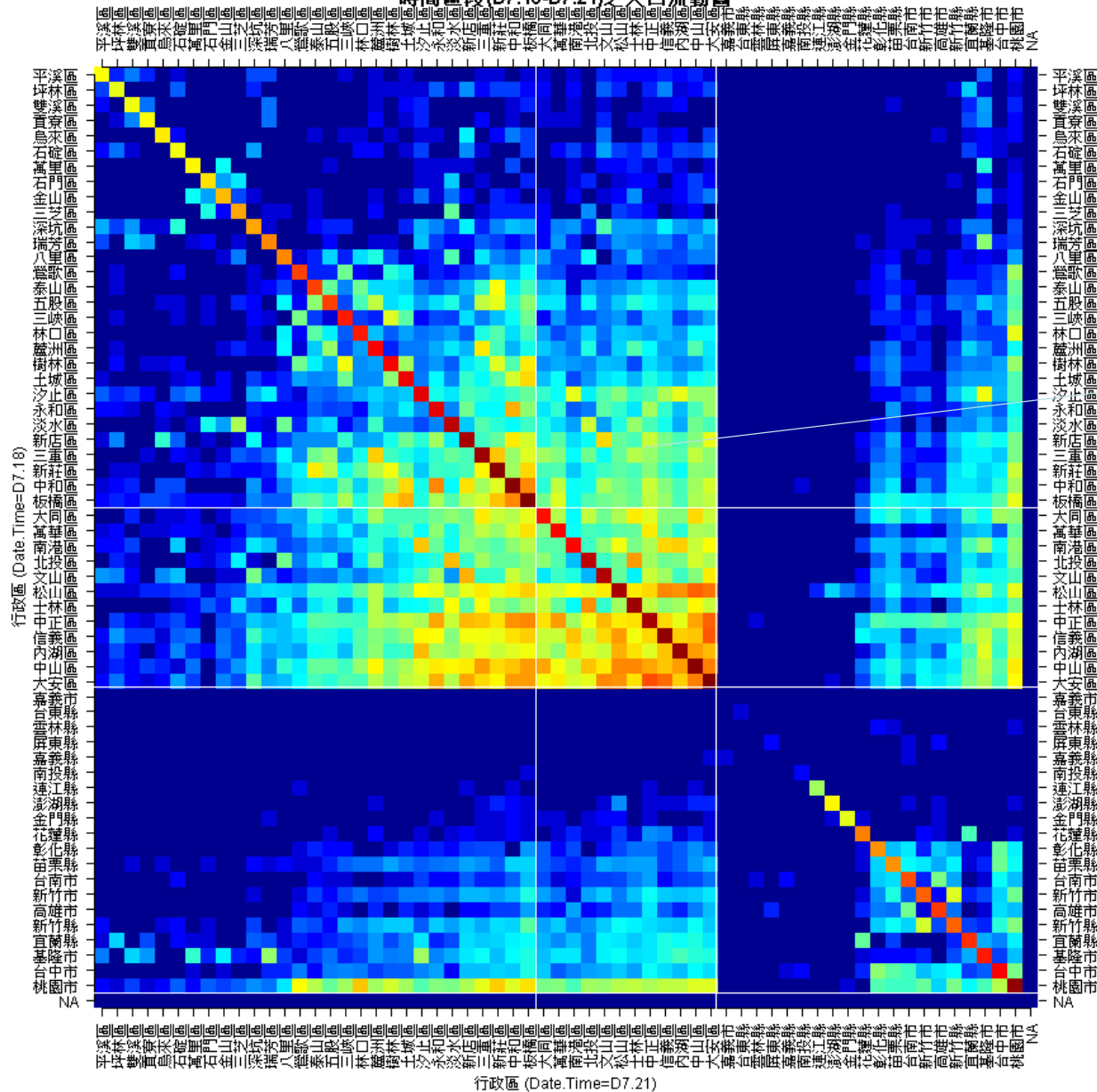


問題: 為何這兩天的時間
區段訊號，幾乎是遺失
的?

D2.01表示: 2018/06/05, 00:00(含)~01:00共
4個時間點(00:00, 00:15, 00:30, 00:45)之
停留地區, 取最多停留地區為此時段之停留
地區。D7.02表示: 2018/06/10, 01:00(含
)~02:00。



時間區段(D7.18-D7.21)之人口流動圖



總人次: 433743

Applications: **39/76**
Asymmetric matrix

中正區
(D7.21)

新店區
(D7.18) **624**

- 共624人從新店區(D7.18)移動到中正區(D7.21)。
- 對角線為停留在原區域之使用者人數。

出現人數

percentile: 99.5% 99.8% 99.7% 99.8% 99.9% 100%
number: 3990 5879 8117 9324 10820 15836

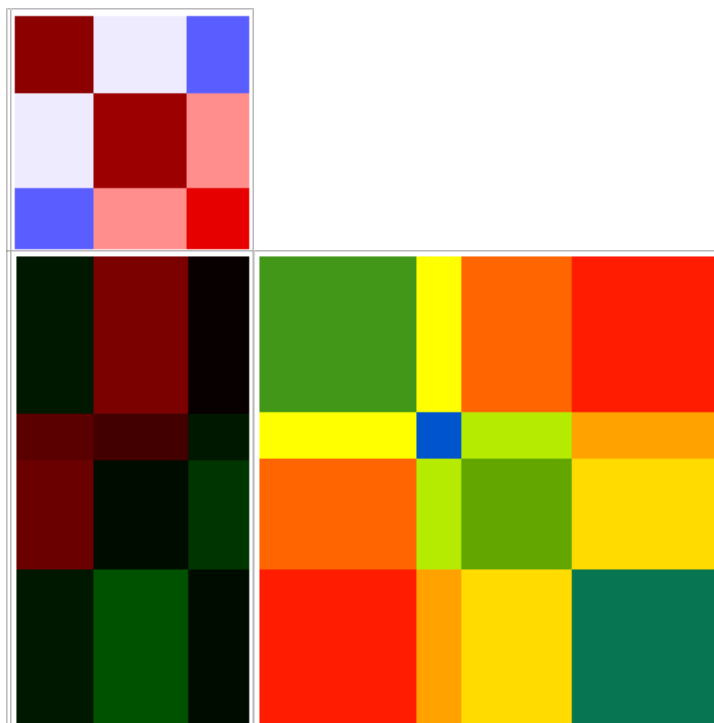
資料檔: move_map_2/
Admin_movement_D7.18-D7.21.csv

Sufficient Display (Chen, 2002)

(1) appropriate
permuted variables
and samples.

(2) carefully derived
partitions for variables
and samples.

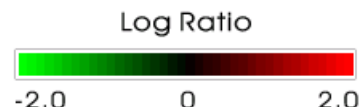
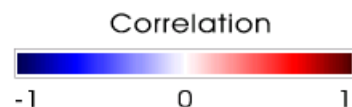
(3) representative
summary statistics
(means, medians or
Std.).



(1) subject-subject

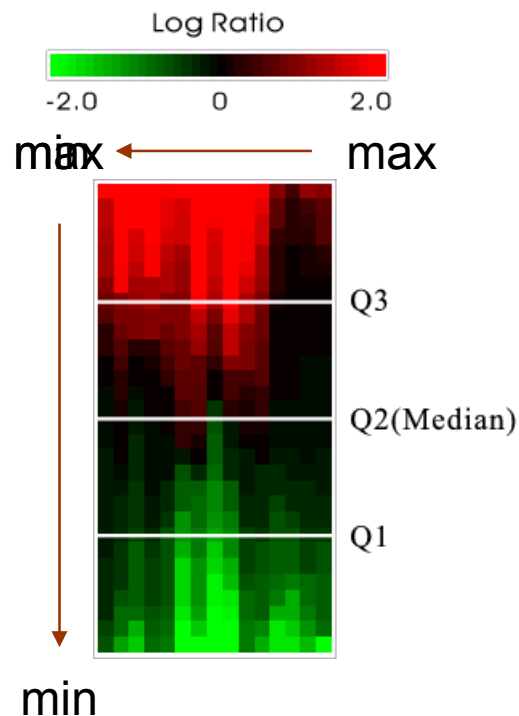
(2) variable-variable

(3) subject-variable



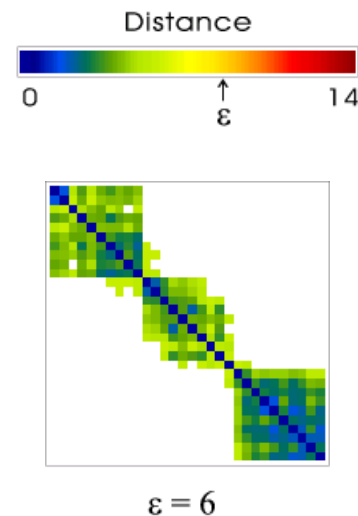
Generalization and Flexibility

Sediment Display



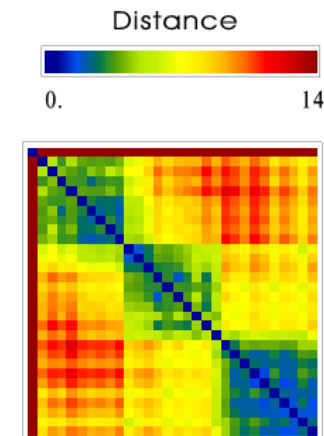
Similar information to that given by a boxplot when the color strips at the quartile positions are extracted.

Sectional Display



Display only those numerical values that satisfy certain conditions.

Restricted Display



Resolution of a Statistical Graph

Heatmaps in R

■ Static

- `image {graphics}`
- `heatmap {stats}`
- `pheatmap {pheatmap}` # pretty
- `heatmap.2 {gplots}` # Enhanced Heat Map
- `aheatmap {NMF}` # annotated heatmap
- `heatmap3 {heatmap3}`
- `annHeatmap2 {Heatplus}`, `heatmap_2 {Heatplus}`, `heatmap_plus {Heatplus}`
- `d3heatmap`
- `Heatmap {ComplexHeatmap}`
- `plot_ly {plotly}` # type = "heatmap"
- `heatmap.plus {heatmap.plus}`
- Heat map produced by `xyplot()` function
- `corrplot {corrplot}`
- `levelplot {lattice}`

■ Interactive

- `heatmaply`
- `fheatmap`
- `gapmap`
- `superheat`
- `shinyheatmap`: Ultra fast low memory heatmap web interface for big data genomics

■ Web Application

- A heatmap is created with the `geom_tile` geom from `ggplot`
- `Autoimage`

Heatmaps: Software-related Literature



- 2010, **neatmap** : non-clustering heat map alternatives in R
- 2011, **gitools**: analysis and visualisation of genomic data using interactive heat-maps
- 2014, advanced heat map and clustering analysis using **heatmap3**
- 2014, **hemi**: a toolkit for illustrating heatmaps
- 2014, **jheatmap** : an interactive heatmap viewer for the web
- 2015, an interactive cluster heat map to visualize and explore multidimensional metabolomic data
- 2015, **clustvis** : a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap
- 2016, **complex heatmaps** reveal patterns and correlations in multidimensional genomic data
- 2017, **Autoimage** : multiple heat maps for projected coordinates
- 2017, **clustergrammer** : a web-based heatmap visualization and analysis tool for high-dimensional biological data
- 2017, **shinyheatmap** : ultra fast low memory heatmap web interface for big data genomics
- 2017, a galaxy implementation of next-generation clustered heatmaps for interactive exploration of molecular profiling data
- 2018, **heatmaply** : an R package for creating interactive cluster heatmaps for online publishing
- 2018, **superheat**: an R package for creating beautiful and extendable heatmaps for visualizing complex data

Display of Genome-Wide Expression Patterns

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

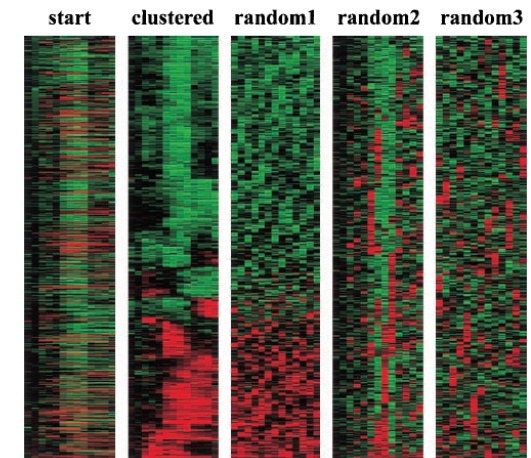
MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

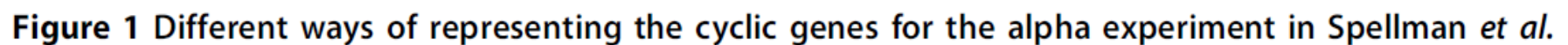
Software: Cluster and TreeView

FIG. 1. Clustered display of data from time course of serum stimulation of primary human fibroblasts. Experimental details are described elsewhere (11). Briefly, foreskin fibroblasts were grown in culture and were deprived of serum for 48 hr. Serum was added back and samples taken at time 0, 15 min, 30 min, 1 hr, 2 hr, 3 hr, 4 hr, 8 hr, 12 hr, 16 hr, 20 hr, 24 hr. The final datapoint was from a separate unsynchronized sample. Data were measured by using a cDNA microarray with elements representing approximately 8,600 distinct

human genes. All measurements are relative to time 0. Genes were selected for this analysis if their expression level deviated from time 0 by at least a factor of 3.0 in at least 2 time points. The dendrogram and colored image were produced as described in the text; the color scale ranges from saturated green for log ratios -3.0 and below to saturated red for log ratios 3.0 and above. Each gene is represented by a single row of colored boxes; each time point is represented by a single column. Five separate clusters are indicated by colored bars and by identical coloring of the corresponding region of the dendrogram. As described in detail in ref. 11, the sequence-verified named genes in these clusters contain multiple genes involved in (A) cholesterol biosynthesis, (B) the cell cycle, (C) the immediate-early response, (D) signaling and angiogenesis, and (E) wound healing and tissue remodeling. These clusters also contain named genes not involved in these processes and numerous uncharacterized genes. A larger version of this image, with gene names, is available at <http://rana.stanford.edu/clustering/serum.html>.

FIG. 3. To demonstrate the biological origins of patterns seen in Figs. 1 and 2, data from Fig. 1 were clustered by using methods described here before and after random permutation within rows (random 1), within columns (random 2), and both (random 3).





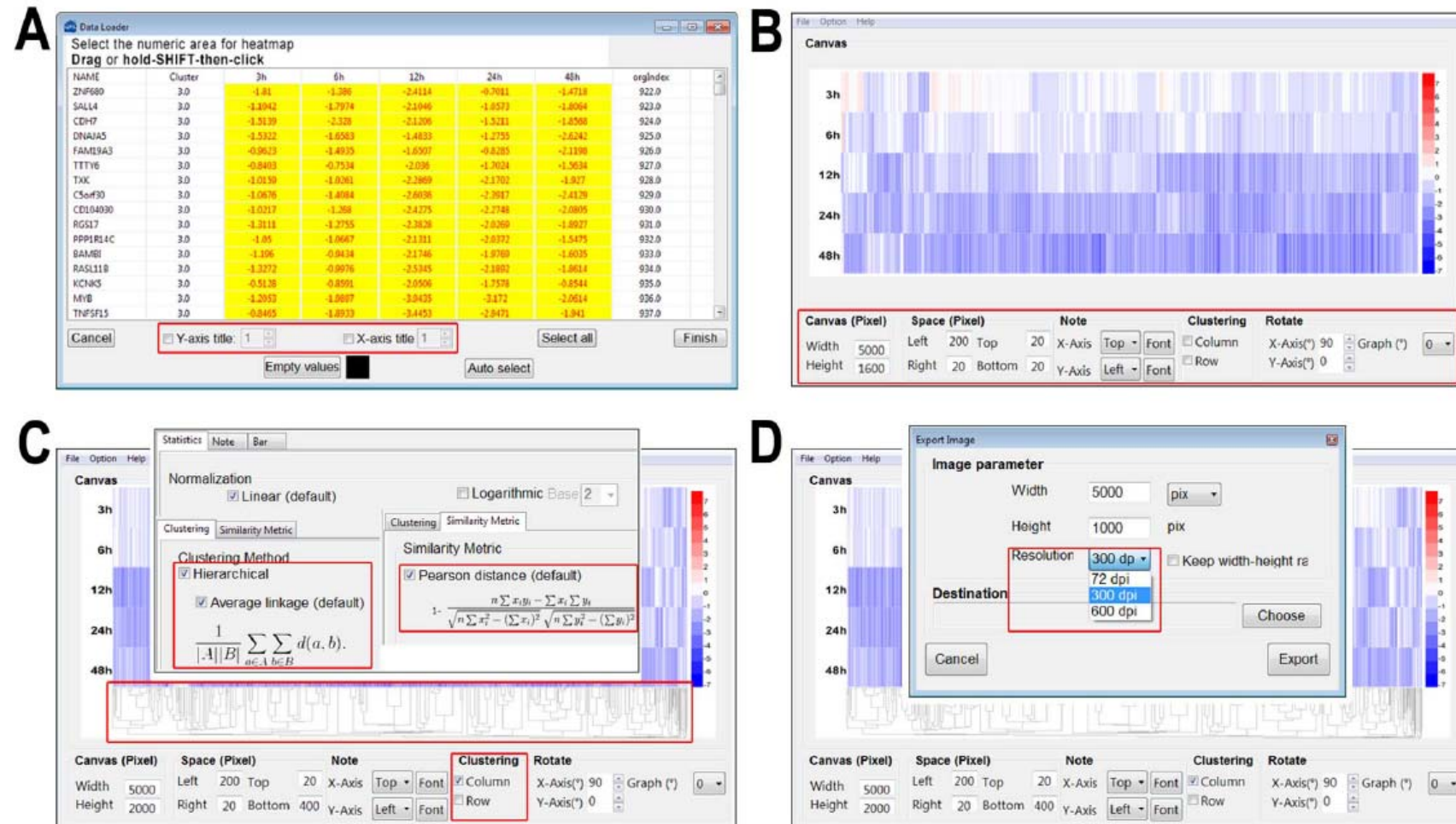


Figure 1. Usage of HemI 1.0. (A) The numerical data in one of three file formats can be directly loaded, whereas the data area can be selected by dragging or holding-SHIFT-then-click manipulations. Titles for X-axis or Y-axis can also be specified; (B) Multiple options for manipulating the heatmap; (C) The numeric data can be clustered for either or both of X-axis and Y-axis; (D) Publication-quality figures can be exported, and two figure formats were supported.
doi:10.1371/journal.pone.0111988.g001

- highly customizable legends and side annotation,
- a wider range of color selections,
- new labeling features which allow users to define multiple layers of phenotype variables, and
- Automatically conducted association tests based on the phenotypes provided,
- different agglomeration (clustering) methods for estimating distance between two samples

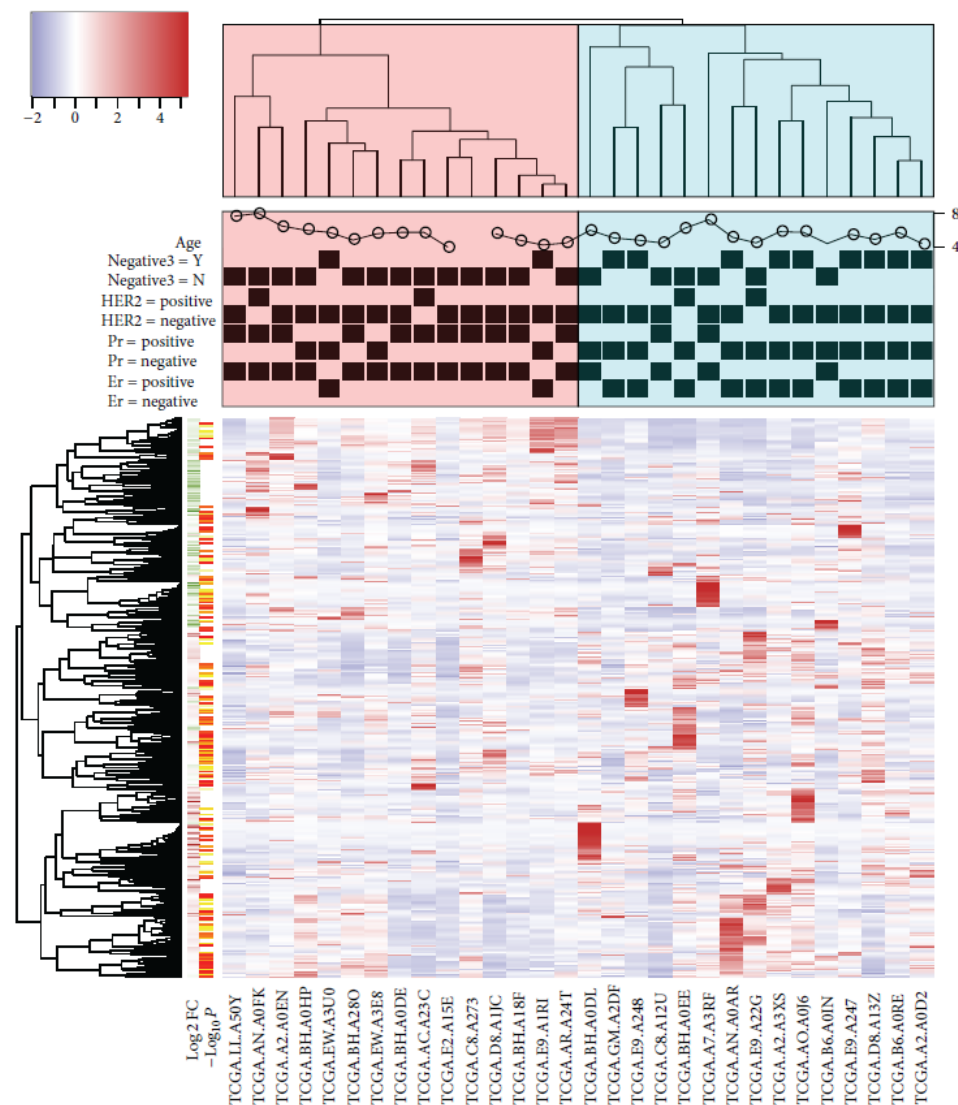


FIGURE 1: An example of "heatmap3" package. The heat map was generated based on 30 samples from TCGA BRCA dataset. The dendrogram of samples (top) was divided into two parts based on the correlation between samples' gene expression and then labeled, respectively. The categorical annotation bars (above heat map) demonstrate the annotation for age, TN, HER2, PR, and ER. The color bar on the left side demonstrates the log2 fold changes and negative log10 P values from comparison of triple negative patients versus nontriple negative patients.

Benton et al., 2015, an interactive cluster heat map to visualize and explore multidimensional metabolomic data, *Metabolomics* 11(4), pp1029-1034.

- A limitation of applying heat maps to global metabolomic data: the large number of ions that have to be displayed and the lack of information provided about important metabolomic parameters such as m/z and retention time.
- the interactive cluster heat map (XCMS Online): to process, statistically evaluate, and visualize mass-spectrometry

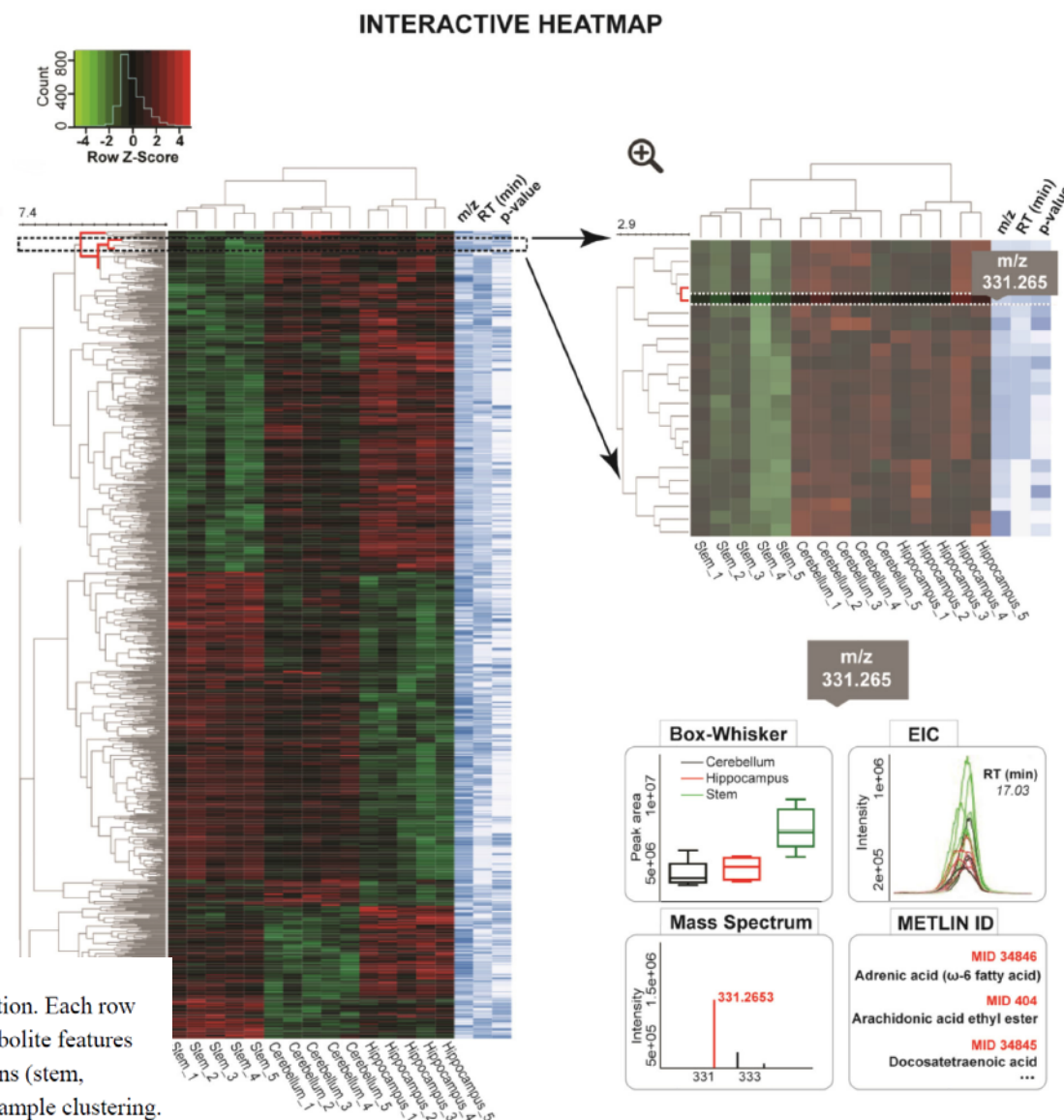


Figure 1.

Interactive, sortable heat map with customized metabolomic data visualization. Each row represents a metabolite feature and each column represents a sample. Metabolite features whose levels vary significantly ($p < 0.01$) across three different brain regions (stem, cerebellum and hippocampus) are projected on the heat map and used for sample clustering.

Metsalu, T. and Vilo, J., 2015, clustvis : a web tool for visualizing clustering of multivariate data using principal component analysis and heatmap, Nucleic Acids Research, 43. :W566-W570.



- ClustVis is written using **Shiny web** application framework

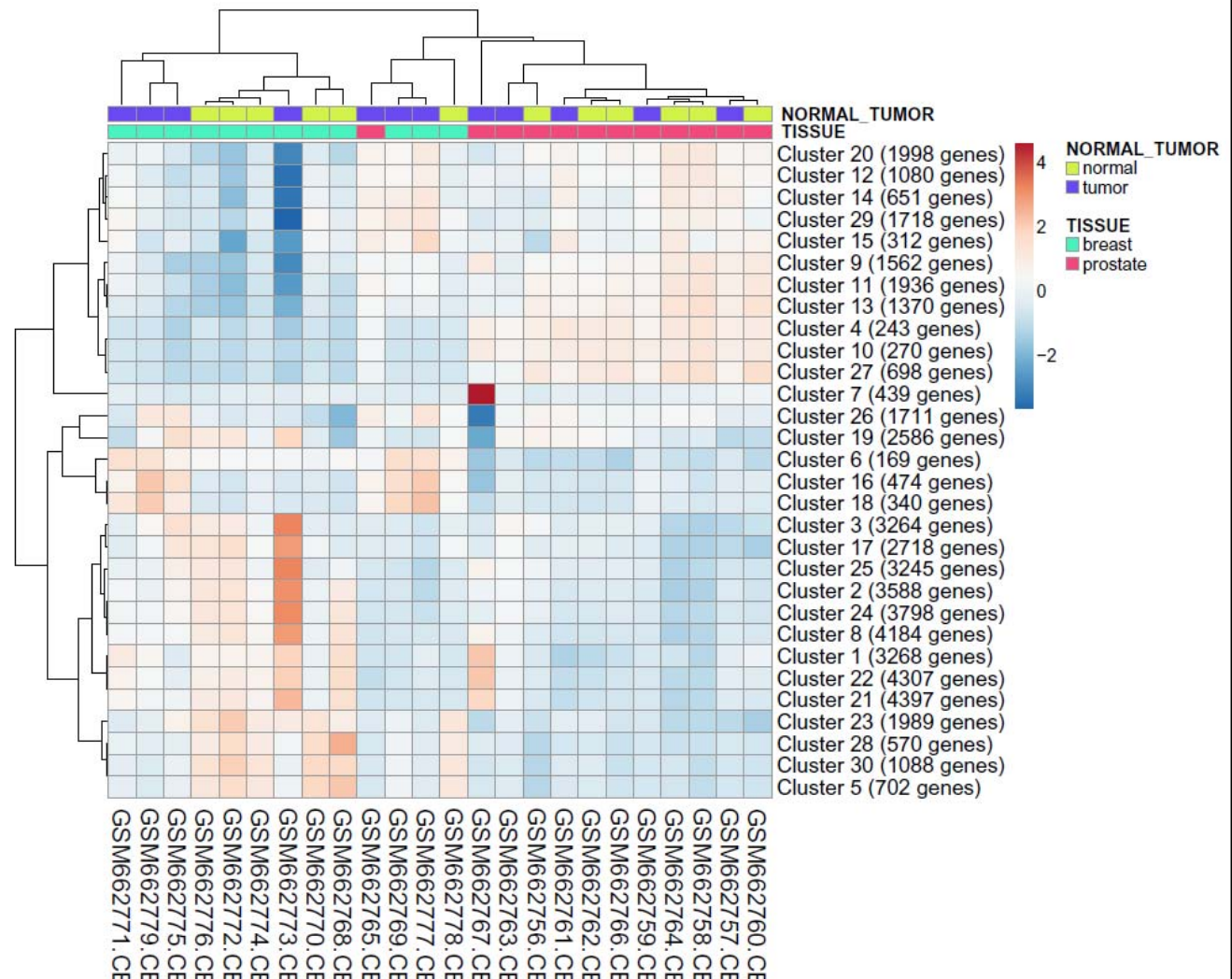
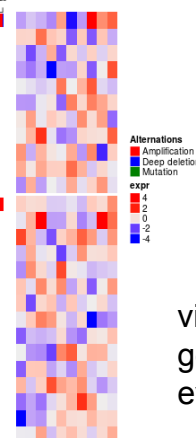
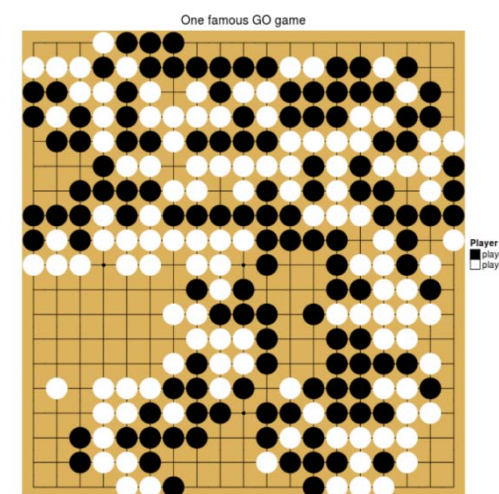
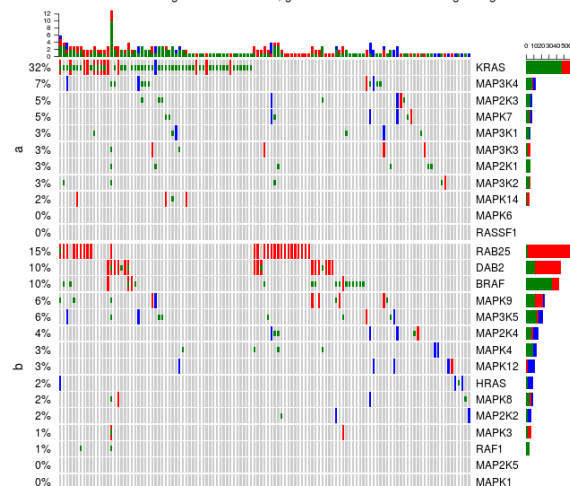
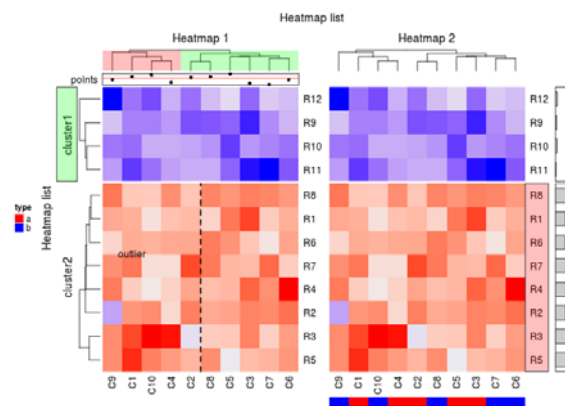


Figure 2. Heatmap of stromal molecular signatures of breast and prostate cancer samples. Annotations on top of the heatmap show clustering of the samples.



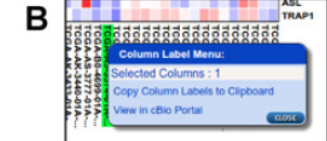
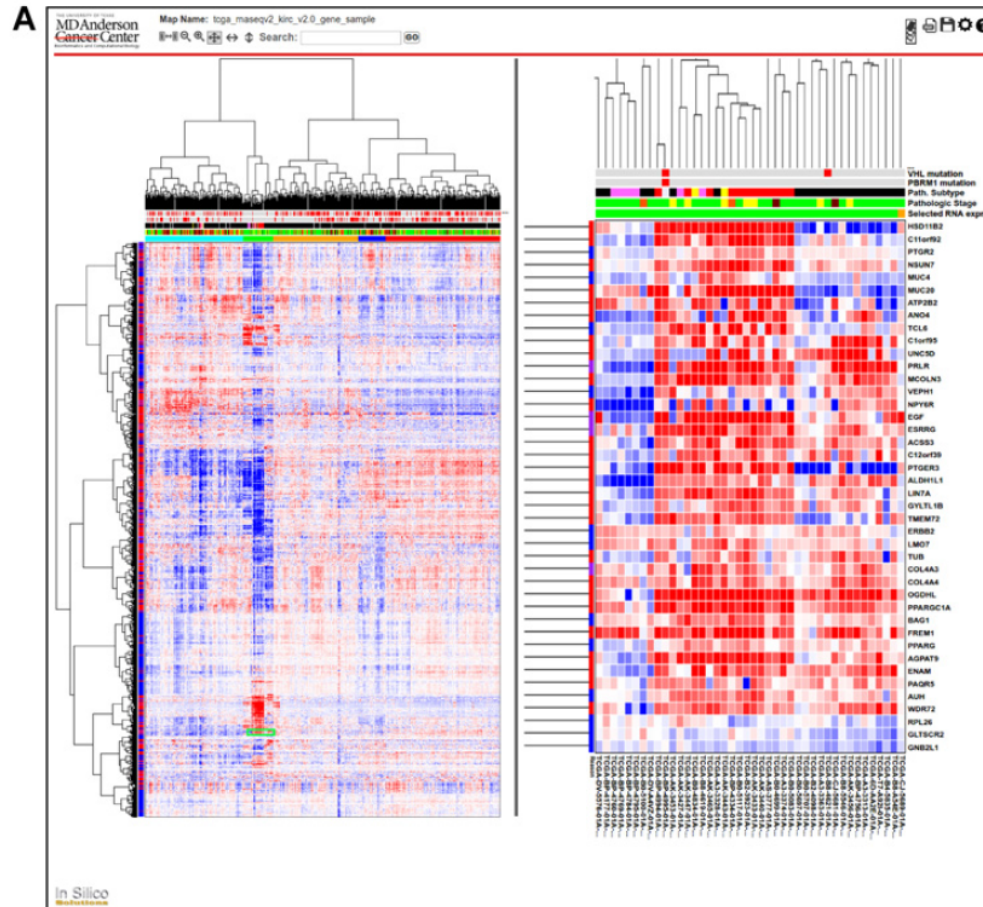
visualize multiple
genomic alteration
events by heatmap

Broom et al, 2017, a galaxy implementation of next-generation clustered heatmaps for interactive exploration of molecular profiling data, Cancer Res; 77(21); e23–26.



51/76

- Extreme zooming without loss of resolution for drill-down of large data matrices.
- Fluent navigation.
- Link-outs from labels to a variety of pertinent annotation resources, GeneCards, PubMed, Ontology, Google, and cBioPortal.
- Annotation with pathway maps.
- Flexible real-time recombination.
- Capture of all metadata necessary to reproduce chosen state of the map months or years later.
- High-resolution graphs meet the requirements of major journals.



Khomtchouk BB, Hennessy JR, Wahlestedt C (2017) **shinyheatmap**:
Ultra fast low memory heatmap web interface for big data
genomics. PLoS ONE 12(5): e0176334.

shinyheatmap

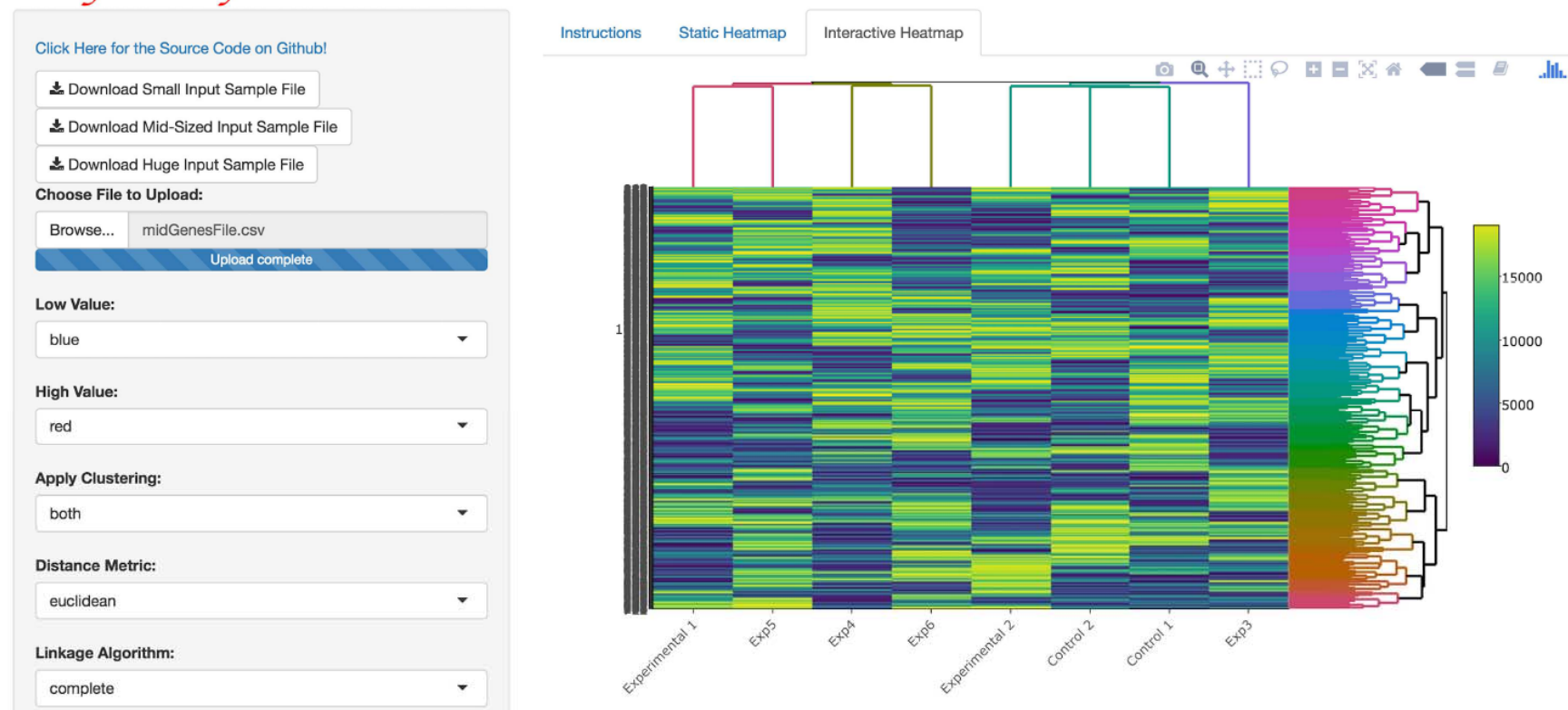


Fig 2. shinyheatmap interactive heatmap. shinyheatmap UI showcasing the visualization of an interactive heatmap generated from a large input dataset. An embedded panel that appears top right on-hover provides extensive download, zoom, pan, lasso and box select, autoscale, reset, and other features for interacting with the heatmap.

<https://doi.org/10.1371/journal.pone.0176334.g002>

Fernandez, N. F. et al. **Clustergrammer**, a web-based heatmap visualization and analysis tool for high-dimensional biological data. Sci. Data 4:170151 doi: 10.1038/sdata.2017.151 (2017).

- zooming, panning,
- filtering, reordering, sharing, performing enrichment analysis, and providing dynamic gene annotations.
- Clustergrammer can be used to generate shareable interactive visualizations by embedding Clustergrammer in Jupyter Notebooks.
- The
- Clustergrammer core libraries can also be used as a toolkit by developers to generate visualizations
- within their own applications.

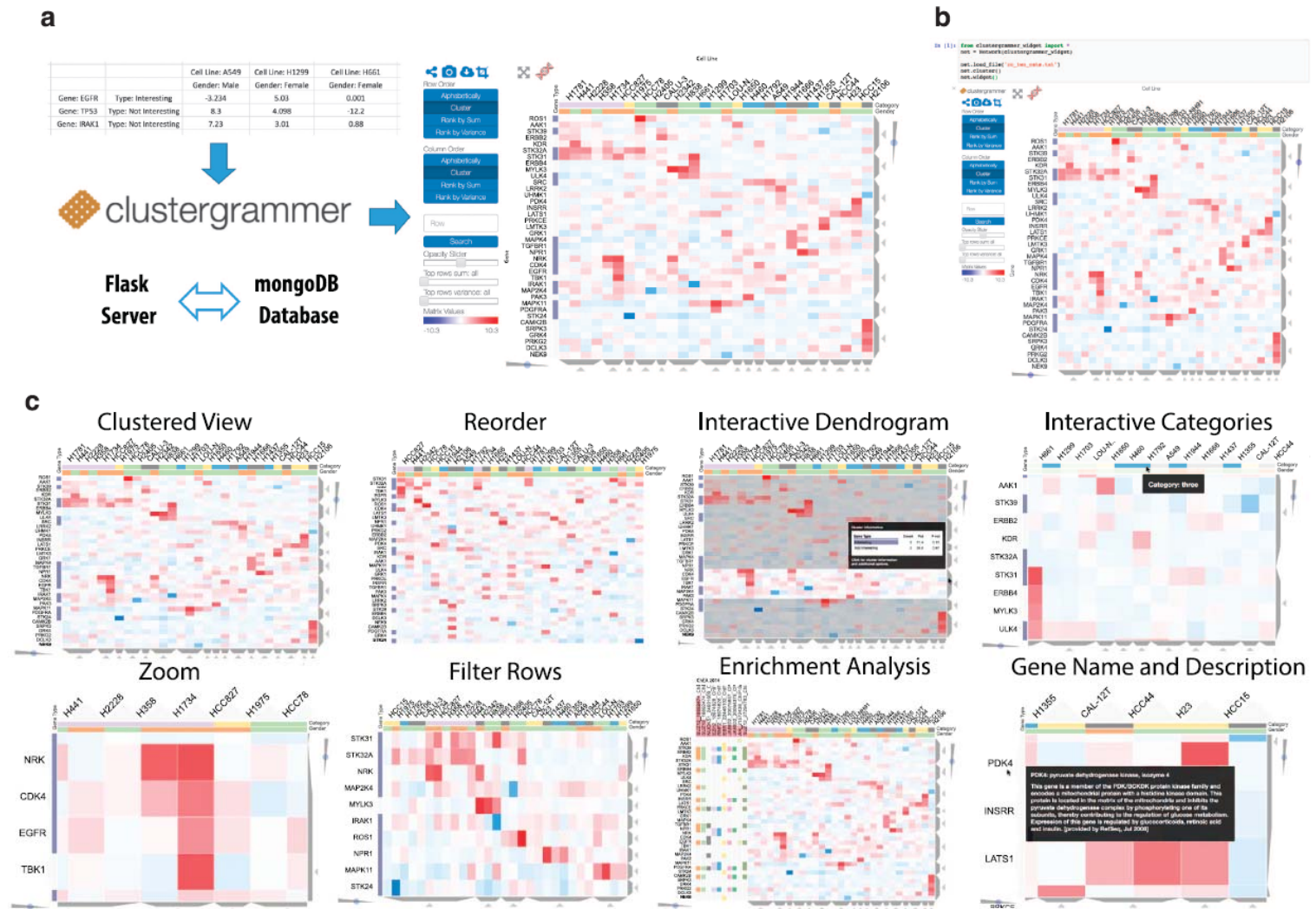


Figure 1. Clustergrammer web app, Jupyter widget, and interactivity. (a) Users can generate interactive and

- construction of heat maps for responses observed on regular or irregular grids, as well as non-gridded data,
- construction of heat maps with a common color scale, with individual color scales,
- projecting (Longitude and latitude) coordinates before plotting,
- easily adding geographic borders, points, and other features to the heat maps.

Two complicated maps

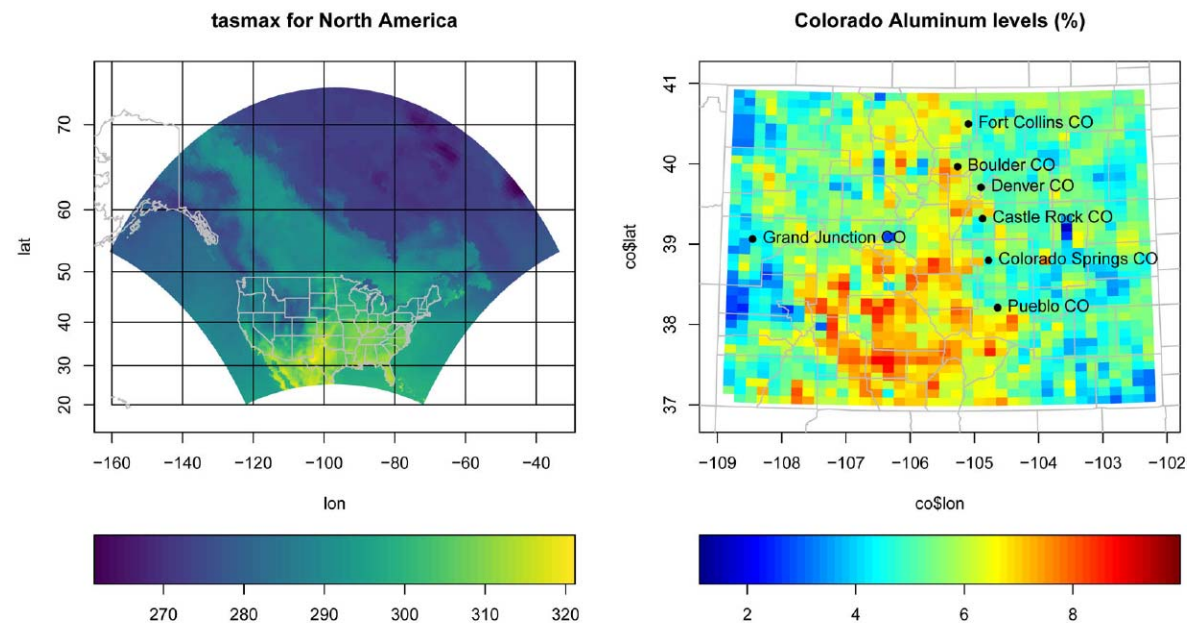


Figure 11.

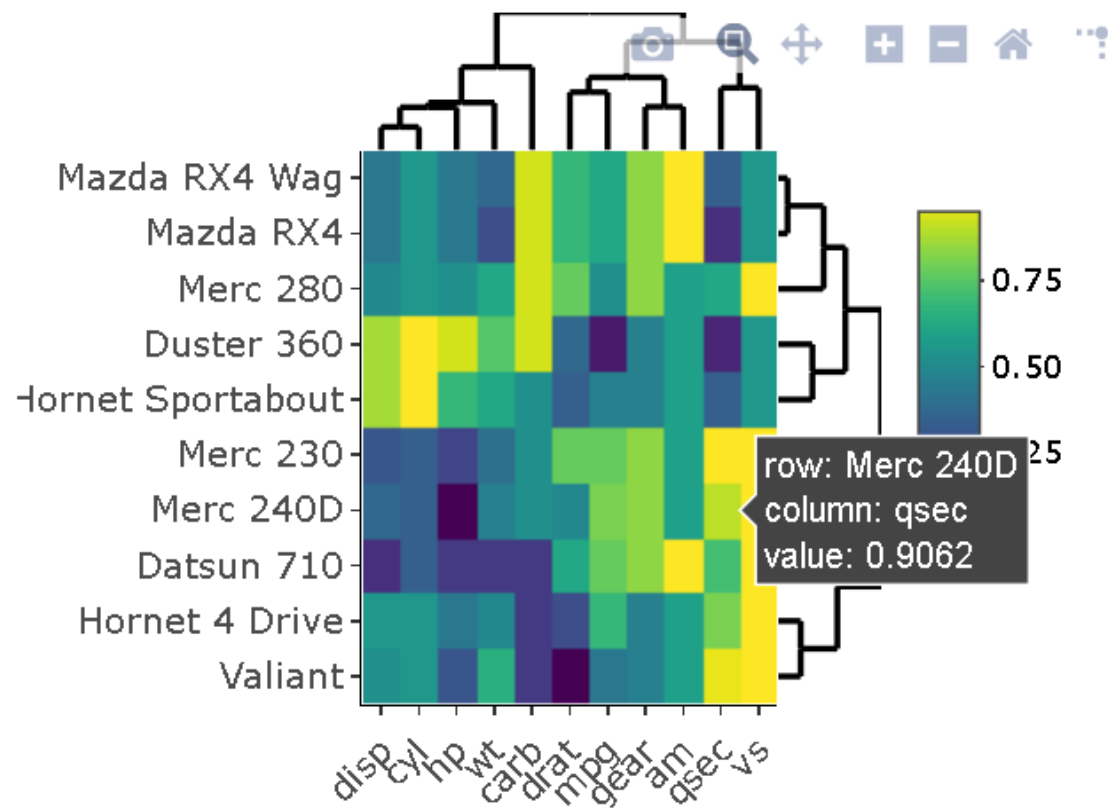
A complicated set of heat maps created using the `autolayout` and `autolegend` functions.

A heat map of the `tasmax` measurements for the first day of the `narccap` data using projected coordinates with an added geographic map of the continental U.S.

maximum daily surface air temperature (`tasmax`)

Galili et al., 2018, **heatmaply**: an R package for creating interactive cluster heatmaps for online publishing, Bioinformatics, 34(9), 2018, 1600–1602.

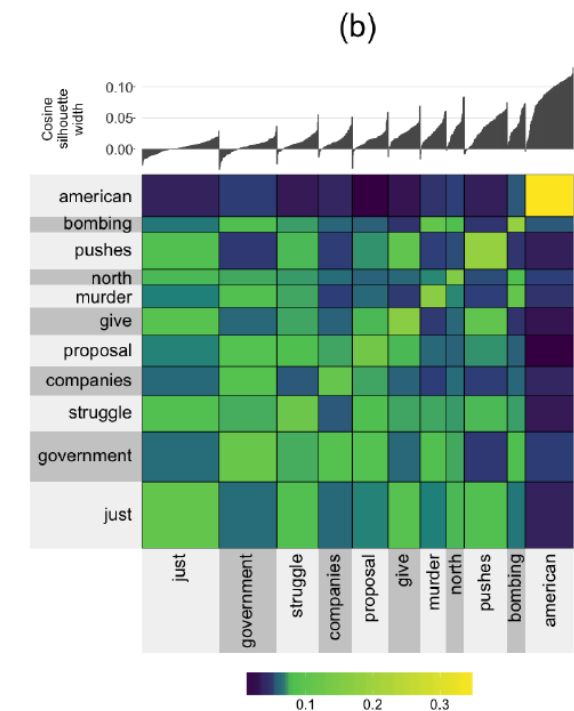
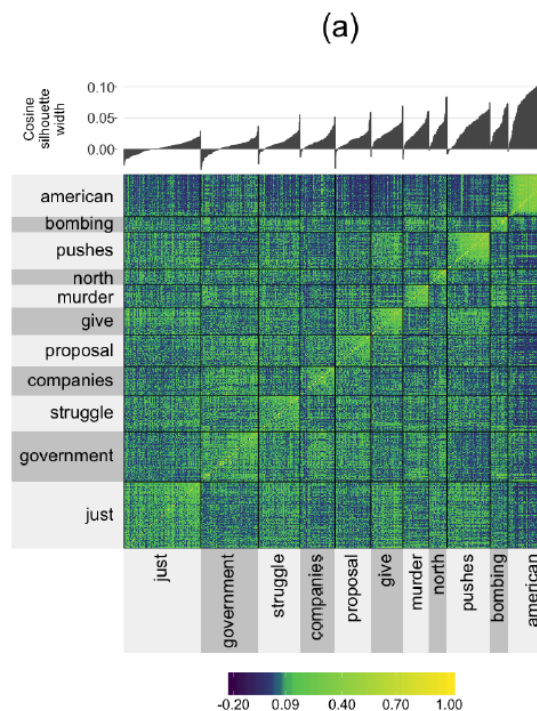
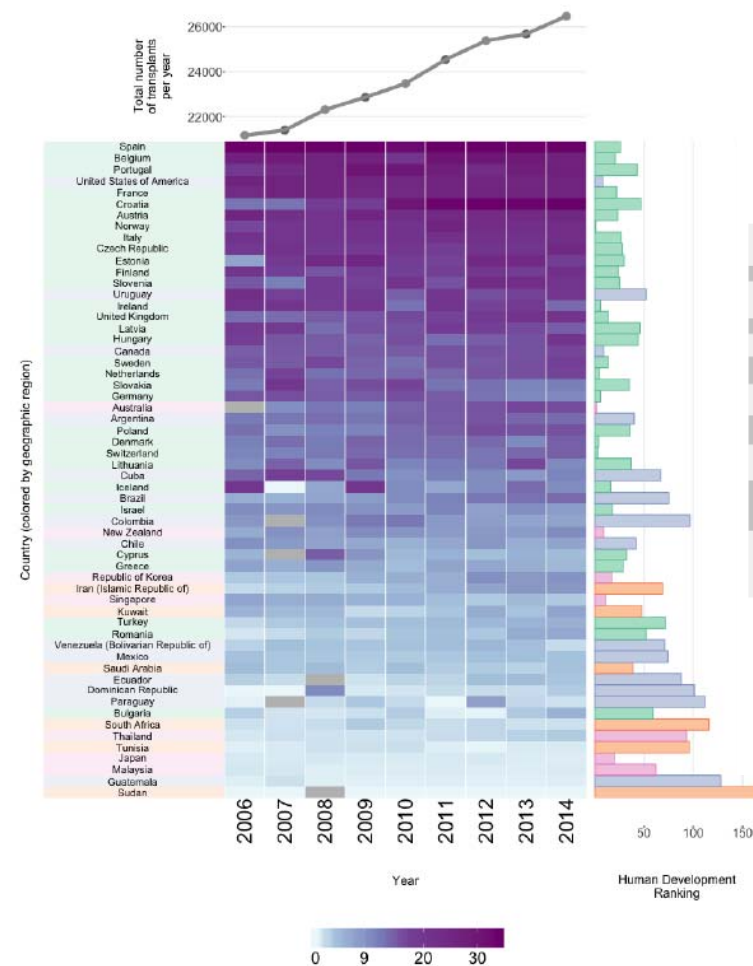
```
# the default by gplots::heatmap.2  
heatmaply(percentize(mtcars)[1:10,], margins = c(40, 130),  
          seriate = "mean")
```



Rebecca L. Barter , Bin Yu, 2017, **Superheat**: an R package for creating beautiful and extendable heatmaps for visualizing complex data, Journal of Computational and Graphical Statistics, <https://doi.org/10.1080/10618600.2018.1473780>



56/76



Generalized Association Plots

- Input Data Type: continuous or binary.
- Various seriation algorithms and **clustering analysis**.
- Various display conditions



■ Modules:

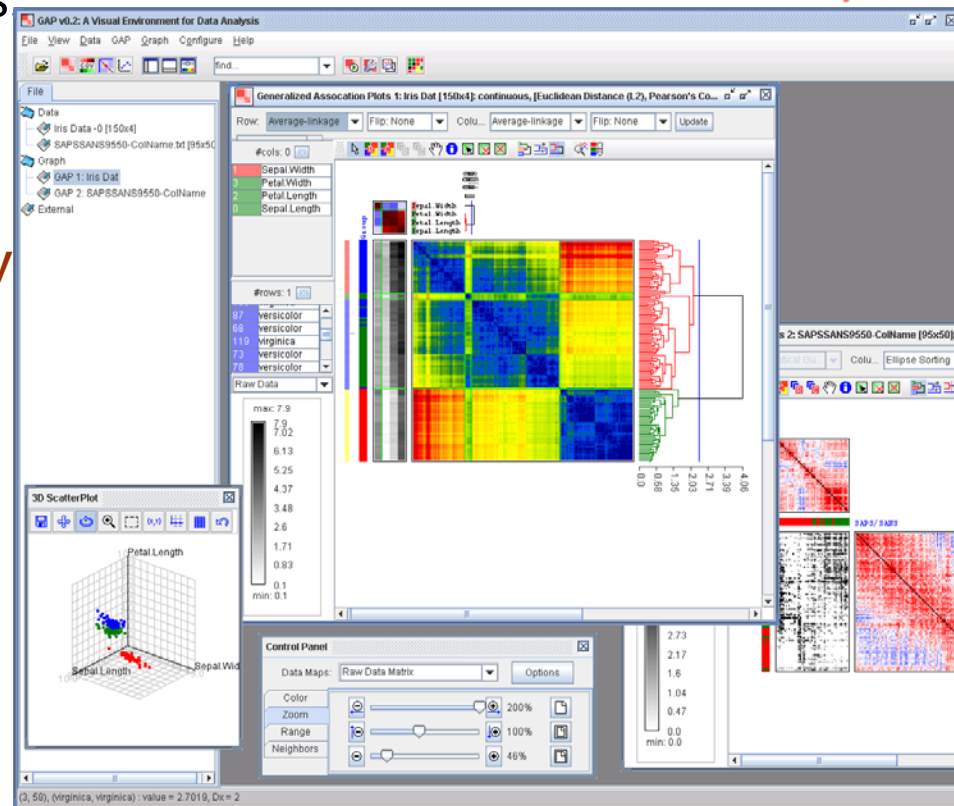
- **Covariate Adjusted.**
- **Proximity Modelling.**
- **Nonlinear Association Analy**
- **Missing Value Imputation.**

Statistical Plots

- 2D Scatterplot,
- 3D Scatterplot (Rotatable)

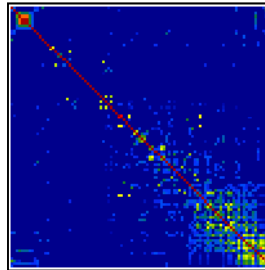
Download

<http://gap.stat.sinica.edu.tw/Software/GAP>

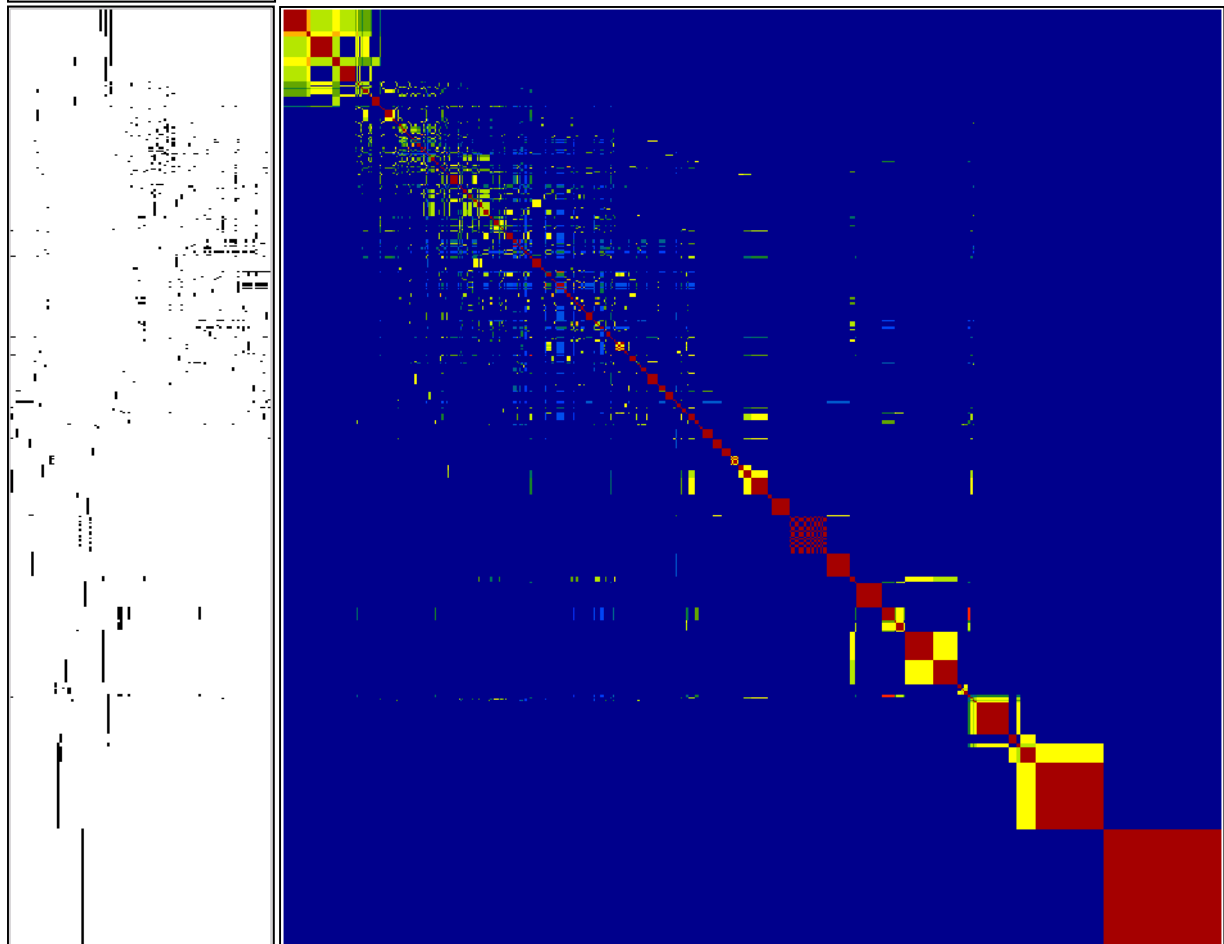
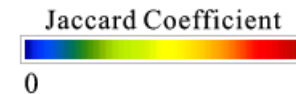


Wu, H. M., Tien, Y. J. and Chen, C. H.* (2010). GAP: A Graphical Environment for Matrix Visualization and Cluster Analysis, Computational Statistics and Data Analysis, 54, 767-778.

- KEGG (Kyoto Encyclopedia of Genes and genomes) metabolism pathways for yeast.
- **1177 related genes** involved in **100 metabolism pathway** of S. c. yeast.
- $(i, j) = 1$: i th gene is involved in j th pathway activities.



1-Jaccard distance coefficient
Elliptical serialiations



MV for Nominal/Categorical Data

- **Color-coding:** color version of relativity of a statistical graph still holds.

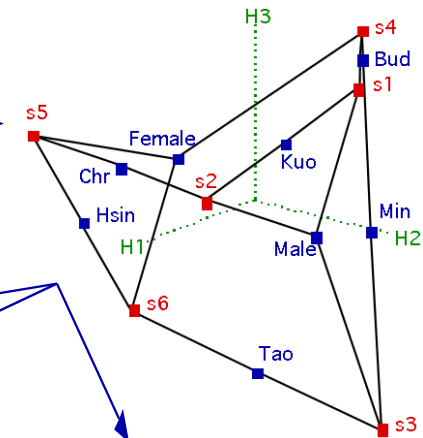
Concept of Categorical GAP with Gifi-Homals

Toy Data Set

Subject	Gender	Reli	Poli
s1	Male	Bud	Kuo
s2	Male	Chr	Kuo
s3	Male	Tao	Min
s4	Female	Bud	Min
s5	Female	Chr	Hsin
s6	Female	Tao	Hsin

1 1 1
1 2 1
1 3 2
2 1 2
2 2 3
2 3 3

Obtain the Homals' 3 Dimensional Dual Space Solution



- **Proximity:** for variables for subjects

Homals

(Gifi, 1990; Michailidis and De Leeuw, 1999)

⇒ **Categorical GAP** (Chen, 1999; Chang *et al.*, 2002)

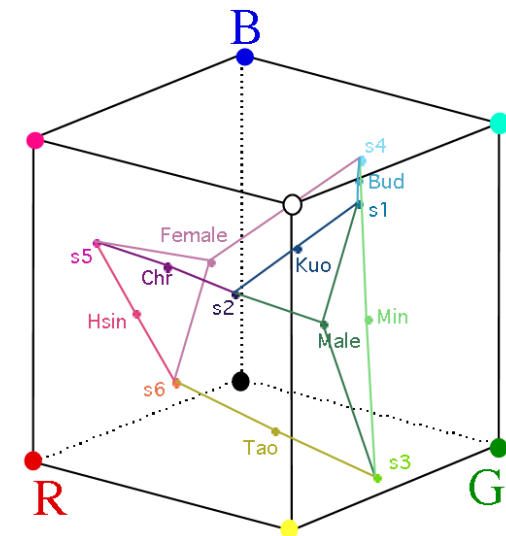
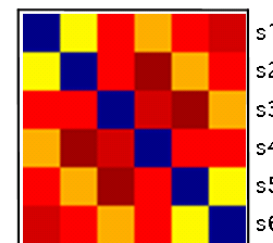
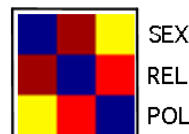
⇒ **Cartography GAP** (Chen *et al.*, 2005)

(3) Compute the Proximity for 2 Variables as the Sum of Weighted 3D Euclidean Distance between Corresponding Categories for the 2 Variables from the Homals' 3 Dimensional Dual Space.

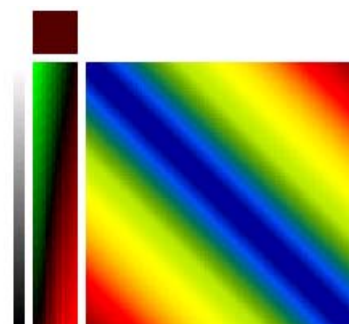
(2) Compute the Proximities for 2 Subjects as the 3D Euclidean Distances for the 2 subjects from the Homals' 3 Dimensional Dual Space.

(1) Scale the Homals' 3 Dimensional Dual Space into the RGB Cube

Close Distant

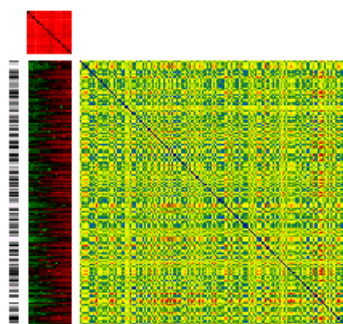


(a) model data



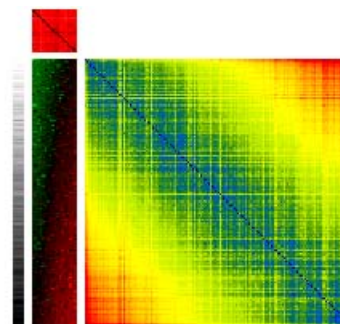
Continuous pattern (Cx)

(b) noisy data



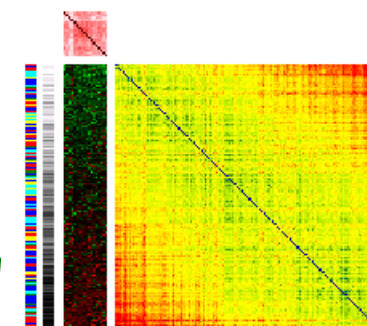
Cn

(c) sorted data

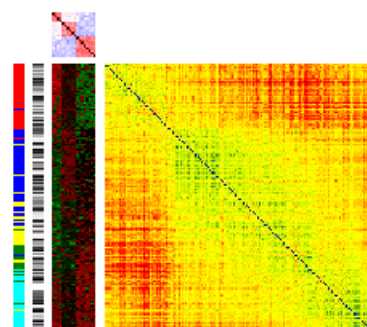


Cn

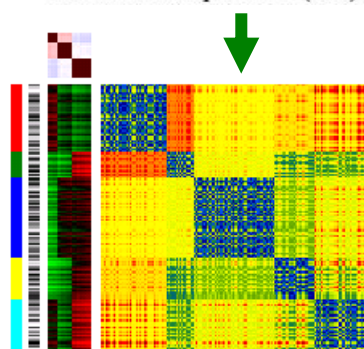
(d) covariate adjusted and sorted data



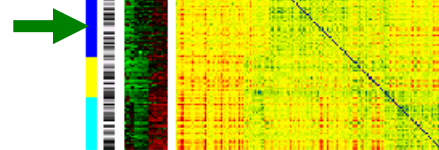
adjusted for discrete pattern



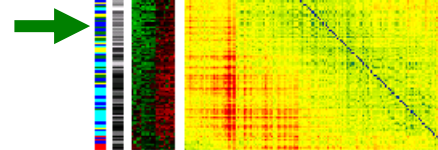
adjusted for continuous pattern



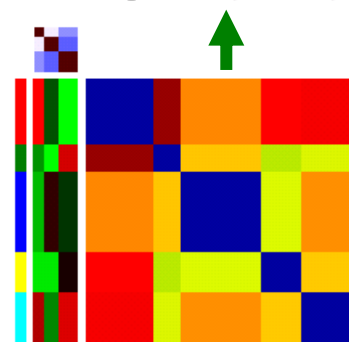
Mixed pattern (Cx+Dx)



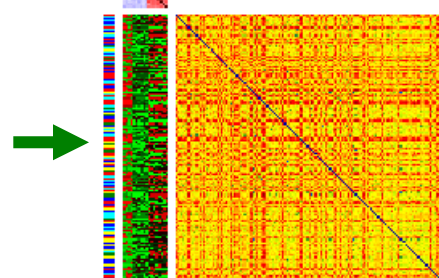
Cn+Dn



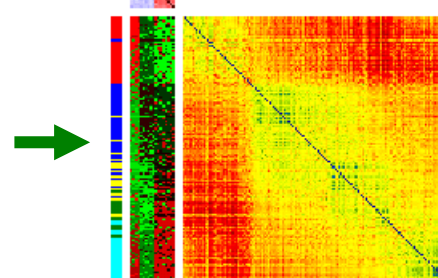
Cn+Dn



Discrete pattern (Dx)



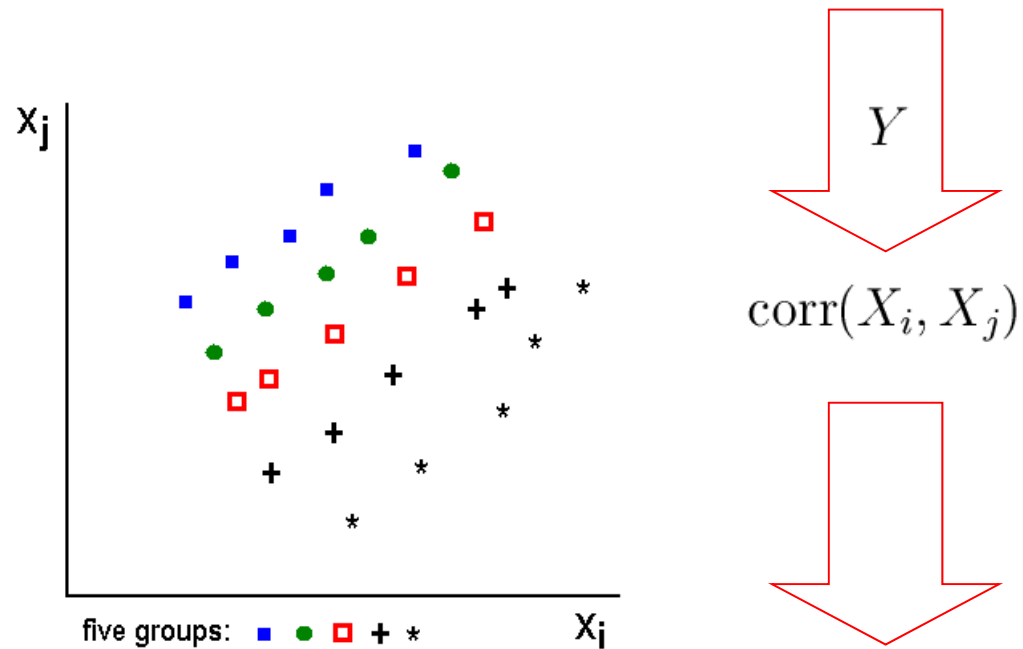
Dn



Dn

Motivation: Covariate-adjusted

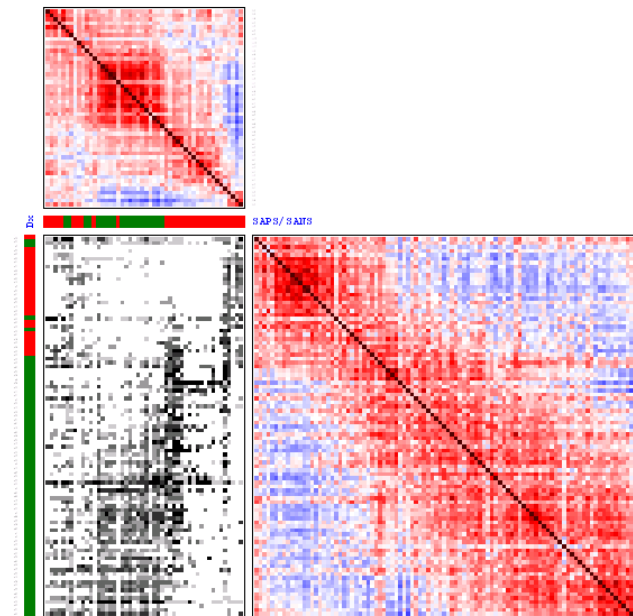
Psychosis Disorder Data



Y

$\text{corr}(X_i, X_j)$

Conditional Correlation

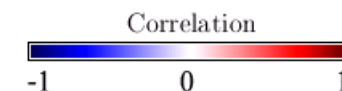


Symptoms

- SAPS
- SANS

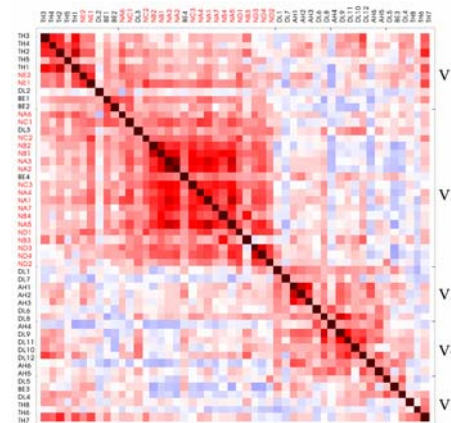
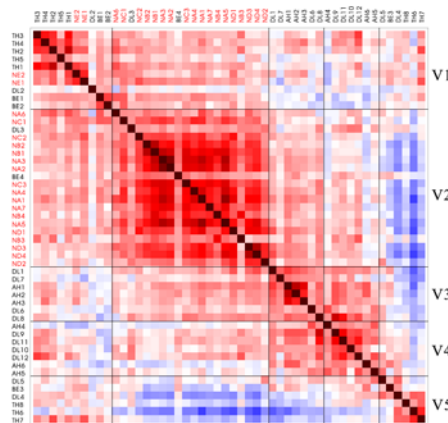
Patients

- Schizophrenic
- Bipolar disorder

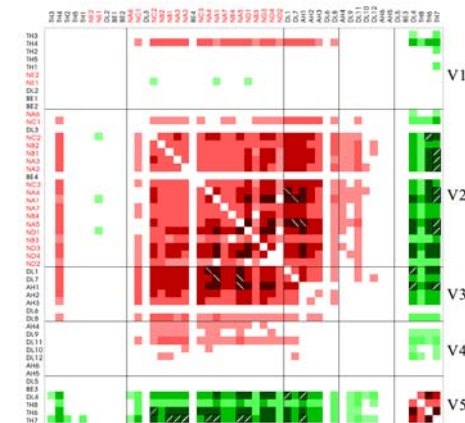
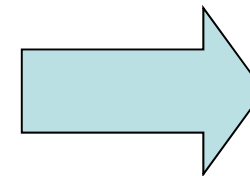


z-score Significant Map

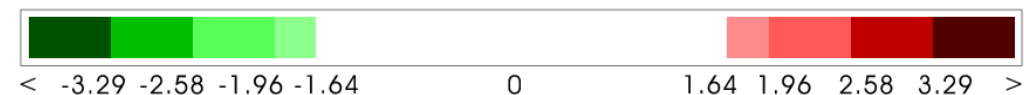
- This z-score significant map is helpful identifying variable pairs with the most **significant differences** in correlation before and after a covariate adjustment.

 R
 R^{adj}
 Z


Dunn and
Clark's
z test

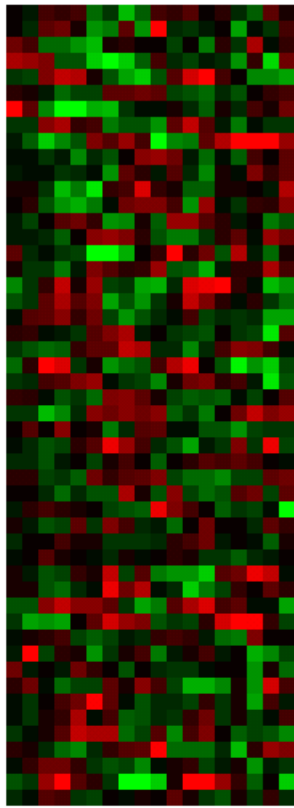


z Score

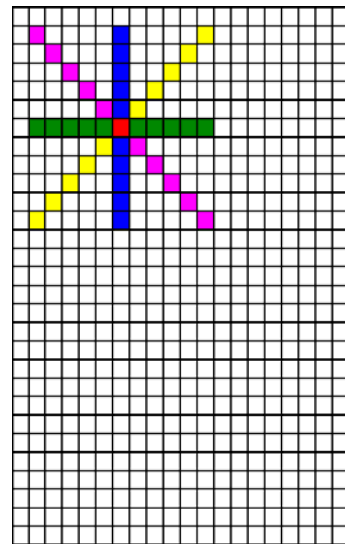


Elliptical Imputation of Missing Values

Step 4 Evaluation

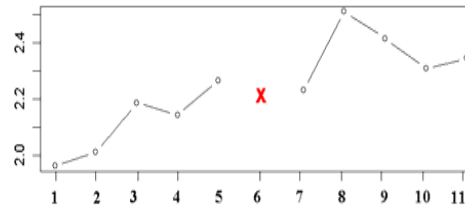


no. directions = 4
no. elements = 10

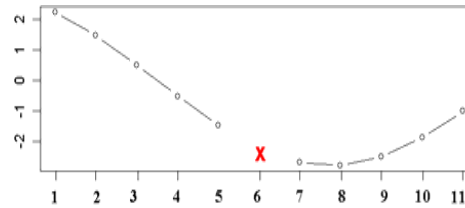


■ Missing Values

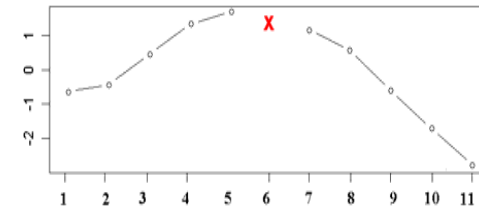
■ Horizontal



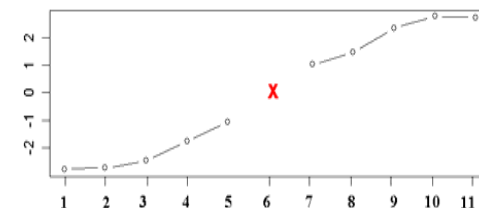
■ Vertical



■ Positive



■ Negative



(1) Fit Regression

$$\hat{Z}_d, d = 1, \dots, 4.$$

(2) Calculate weights

$$\text{slope}_d[i] = y_d[i + 1] - y_d[i], i = 1, \dots, 9.$$

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

(3) Impute values

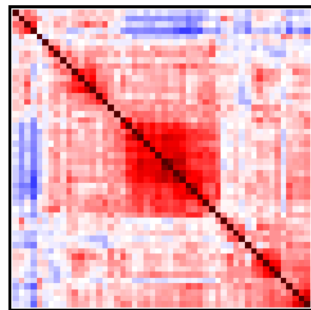
$$\text{ImputedValue} = \frac{\sum_{d=1}^4 w_d \hat{Z}_d}{\sum_{d=1}^4 w_d}$$

Interactive Diagnostic System for Hierarchical Clustering Tree with Matrix Visualization



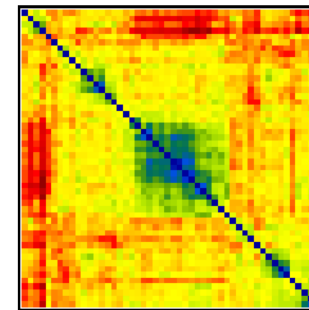
64/76

(1) Input **Proximity** Matrix



(e.g., Pearson's Correlation)

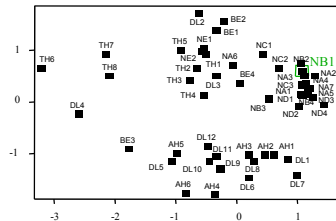
(2) Transformed **Disparity** Matrix



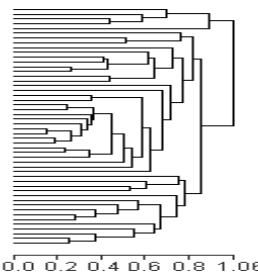
(e.g., Distance)

Statistical Modeling

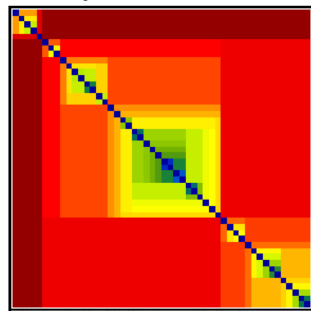
Multidimensional Scaling
(MDS)



Hierarchical Clustering Tree
(HCT)

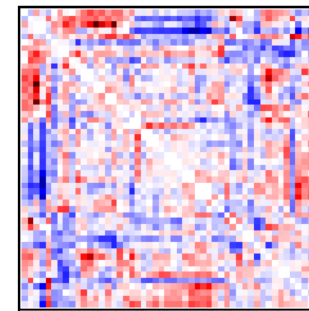


(e.g., Cophenetic Matrix)

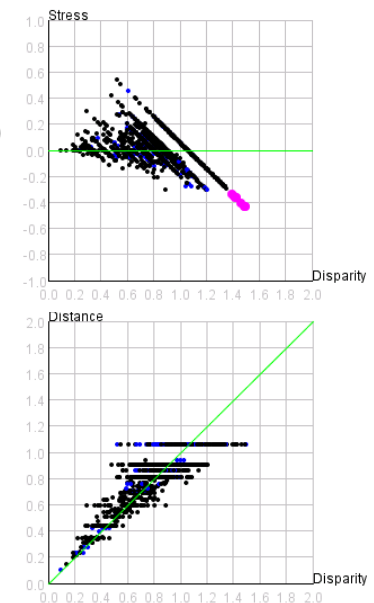
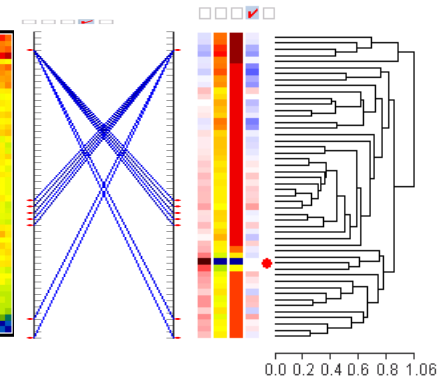


(3) Output **Distance** Matrix

(e.g., Residual Matrix)



(4) **Stress** Matrix

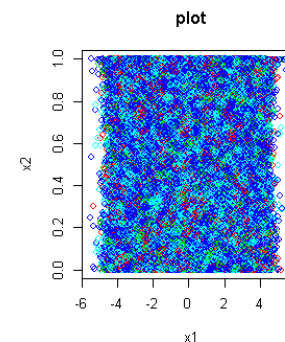
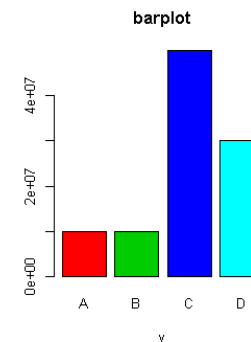
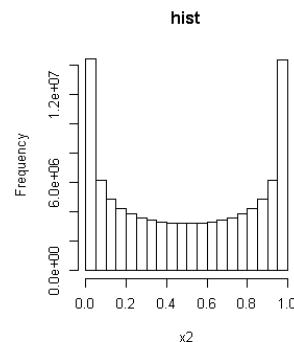
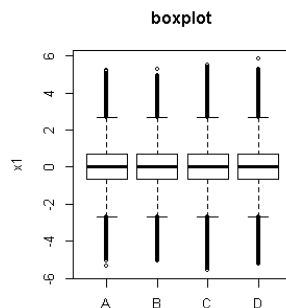
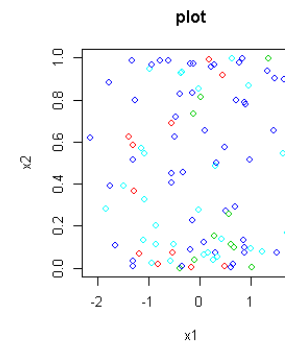
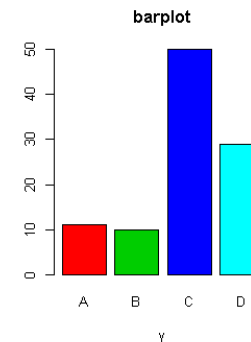
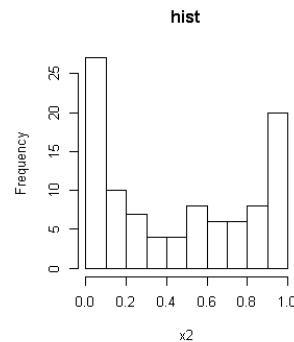
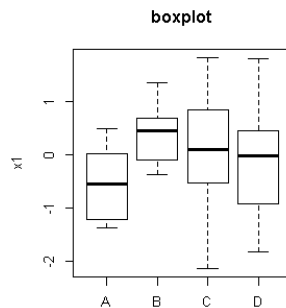


The Challenge of Visualizing Big Data

`> n <- 1e+02`

a large p?

`> n <- 1e+08`



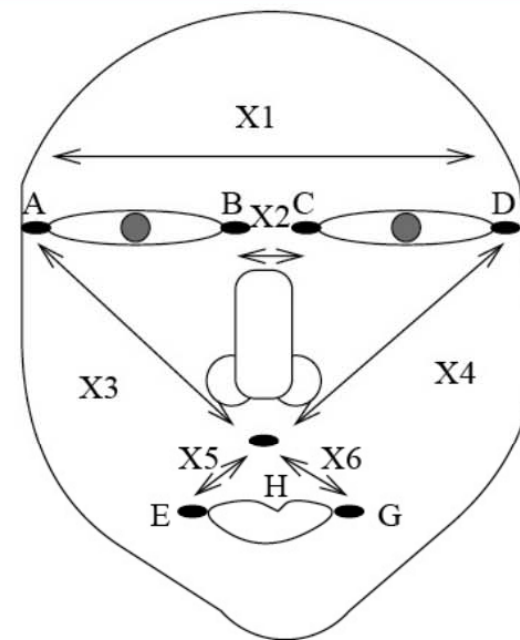
```
> n <- 1e+02
> y <- as.factor(sample(LETTERS[1:4], n, replace=T, prob=c(0.1, 0.1, 0.5, 0.3)))
> x1 <- rnorm(n)
> x2 <- rbeta(n, 0.5, 0.5)
> xydata <- data.frame(y, x1, x2)
> par(mfrow=c(1,4))
> boxplot(x1~y, data=xydata, ylab="x1", main="boxplot")
> hist(x2, xlab="x2", main="hist")
> barplot(table(y), xlab="y", col = 2:5, main="barplot")
> plot(x1, x2, main="plot", col=as.integer(y)+1)
```

Two principles:
Look at Less Data;
or Look at Data Faster

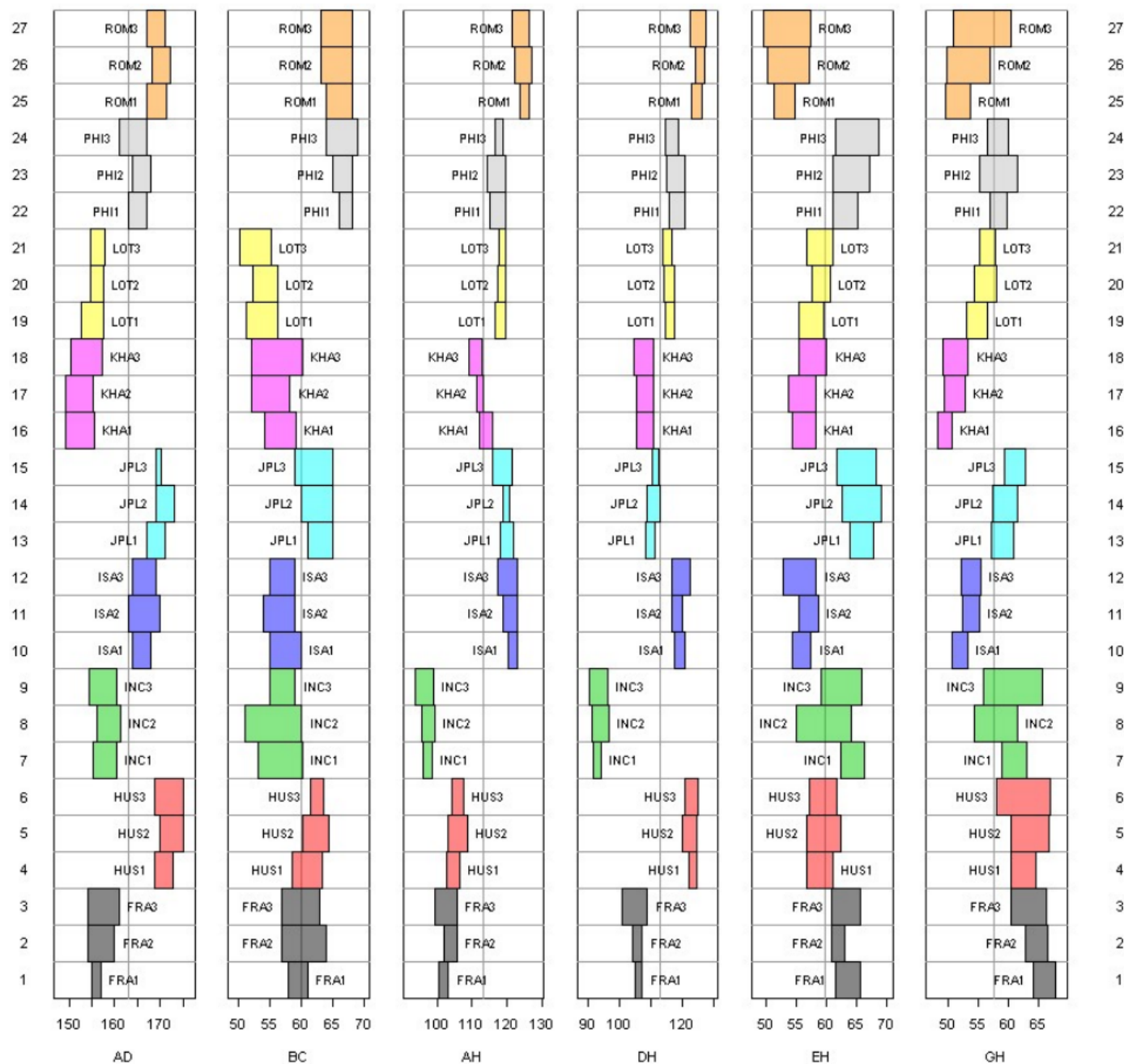
The interval-valued symbolic data

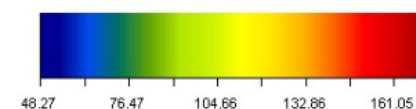
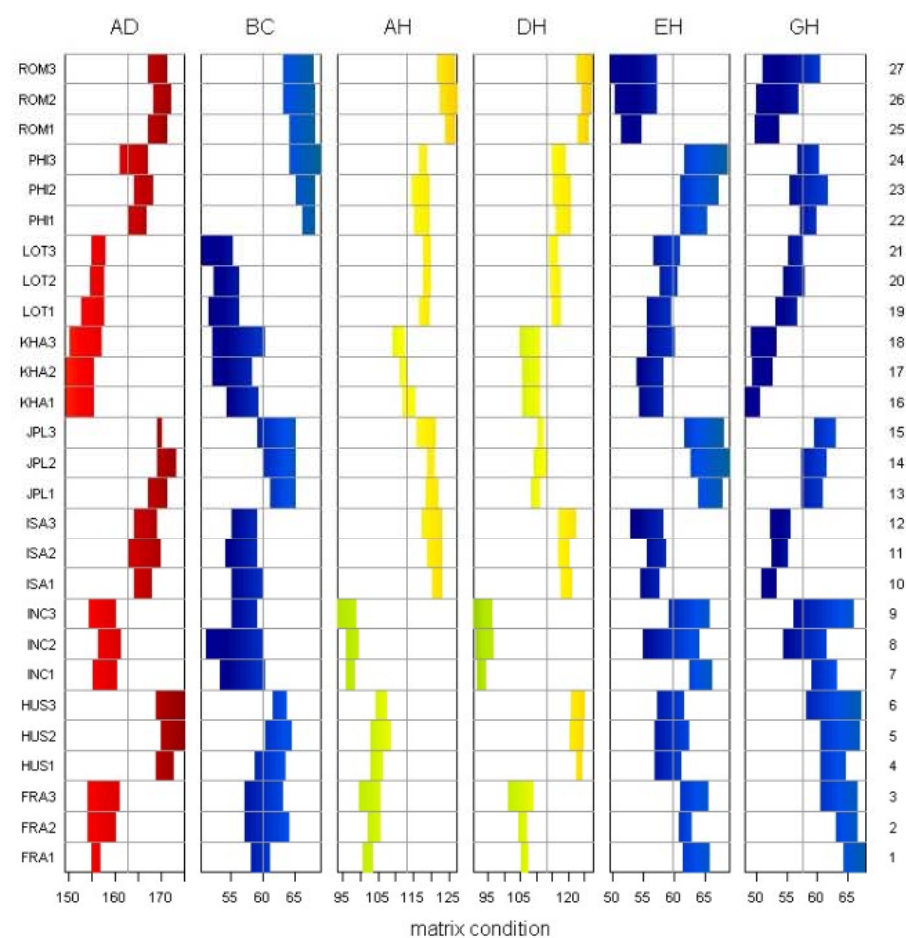
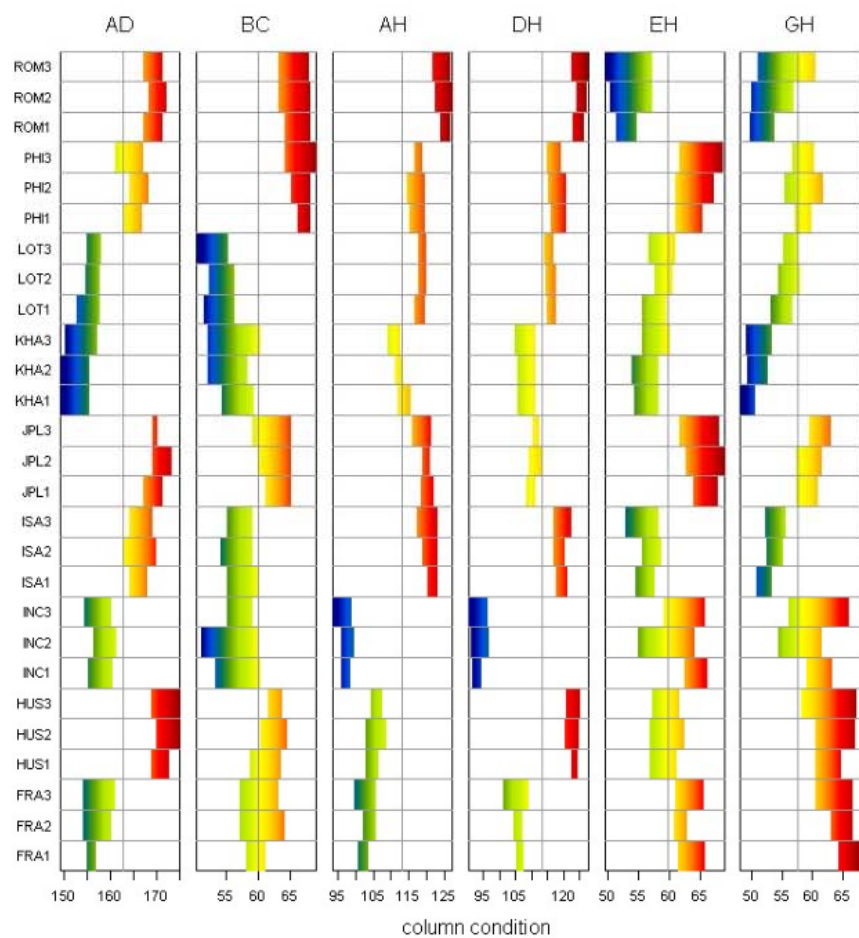
FACE RECOGNITION DATA

- 1 Details: Leroy et al. (1996),
Douzal-Chouakria, Billard and Diday (2011),
Le-Rademacher and Billard (2012)
- 2 The dataset gives **six face measurements** of **nine** men, each with **three** observations, resulting in a total **27 observations**. The measurements for each observation came from a sequence of images.

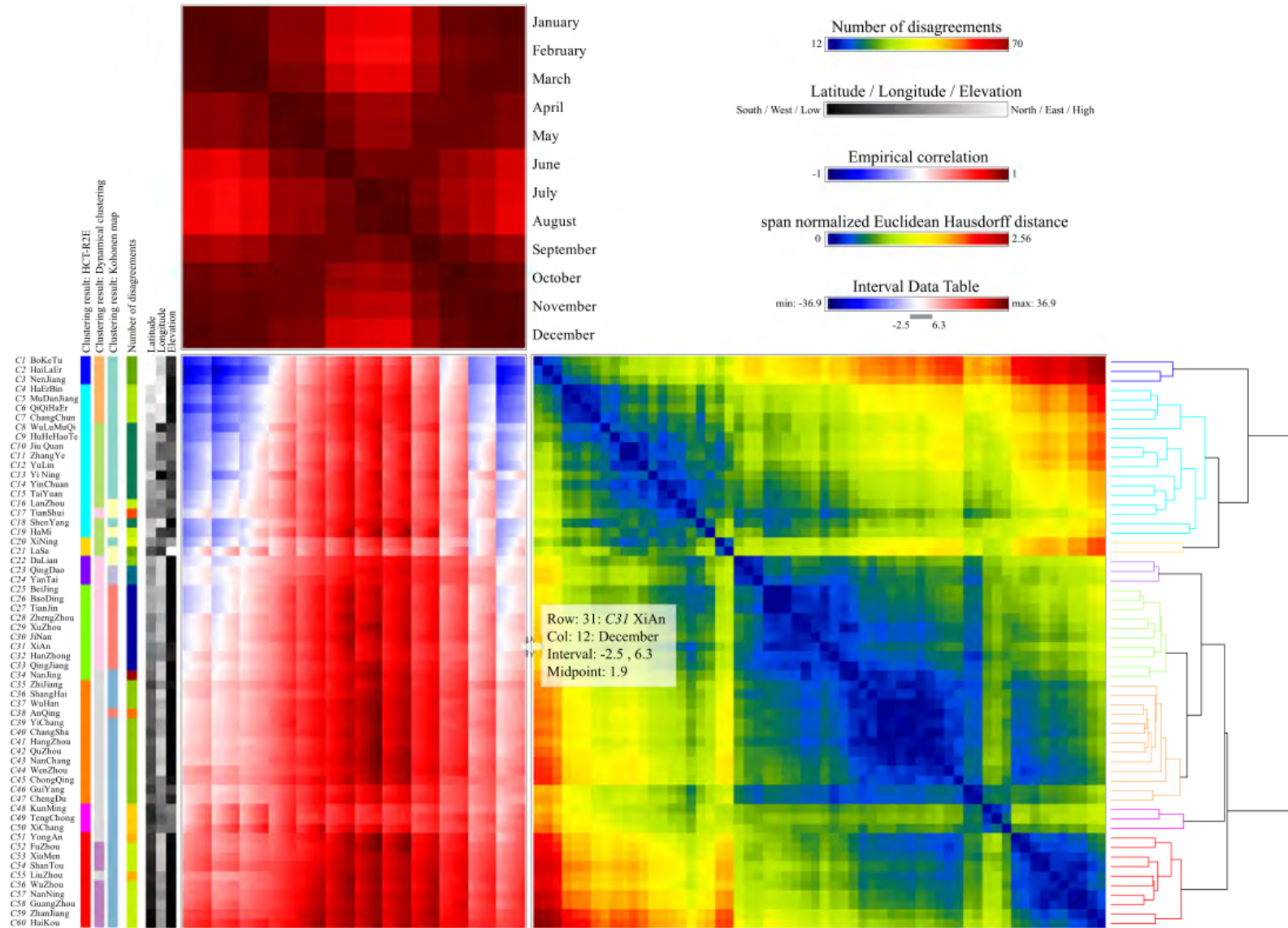


name	AD	BC	AH	DH	EH	GH
FRA1	(155, 157)	(58, 61.01)	(100.45, 103.28)	(105, 107.3)	(61.4, 65.73)	(64.2, 67.8)
FRA2	(154, 160.01)	(57, 64)	(101.98, 105.55)	(104.35, 107.3)	(60.88, 63.03)	(62.94, 66.47)
FRA3	(154.01, 161)	(57, 63)	(99.36, 105.65)	(101.04, 109.04)	(60.95, 65.6)	(60.42, 66.4)
HUS1	(168.86, 172.84)	(58.55, 63.39)	(102.83, 106.53)	(122.38, 124.52)	(56.73, 61.07)	(60.44, 64.54)
⋮						
ROM3	(167.11, 171.19)	(63.13, 68.03)	(121.62, 126.57)	(122.58, 127.78)	(49.41, 57.28)	(50.99, 60.46)





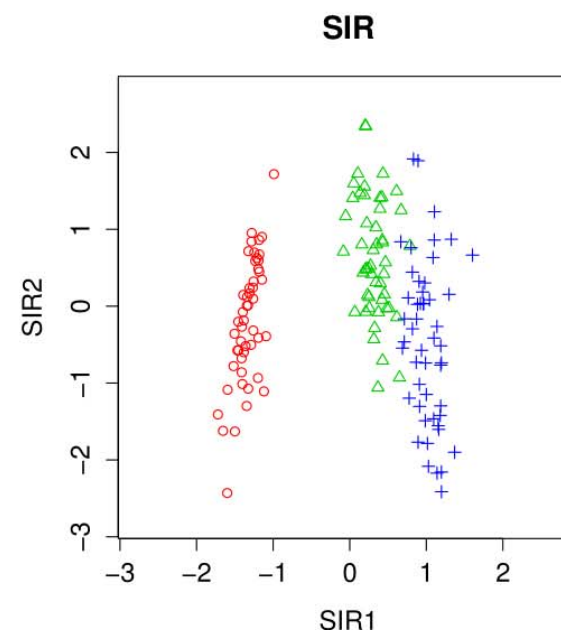
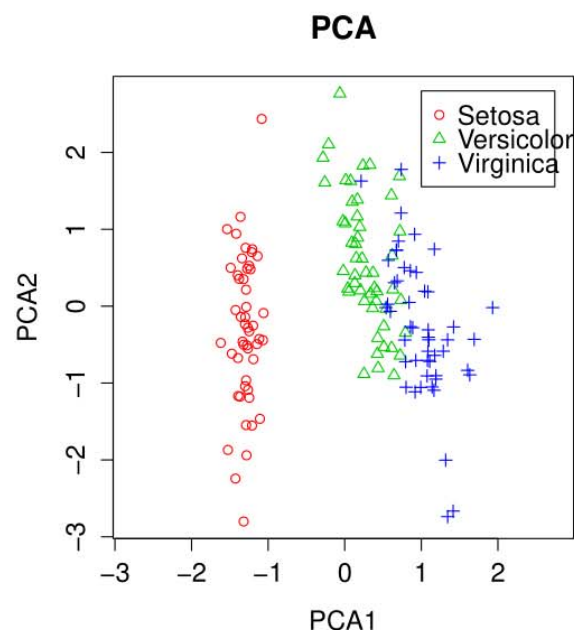
Kao, C.H., Nakano J., Shieh, S.H., Tien, Y.J., Wu, H.M., Yang, C.K., and Chen, C.H.* (2014), Exploratory data analysis of interval-valued symbolic data with matrix visualization, Computational Statistics & Data Analysis, 79, 14-29.



The histogram-valued symbolic data

EXAMPLE: IRIS DATA (150×4 , $y=(50, 50, 50)$)

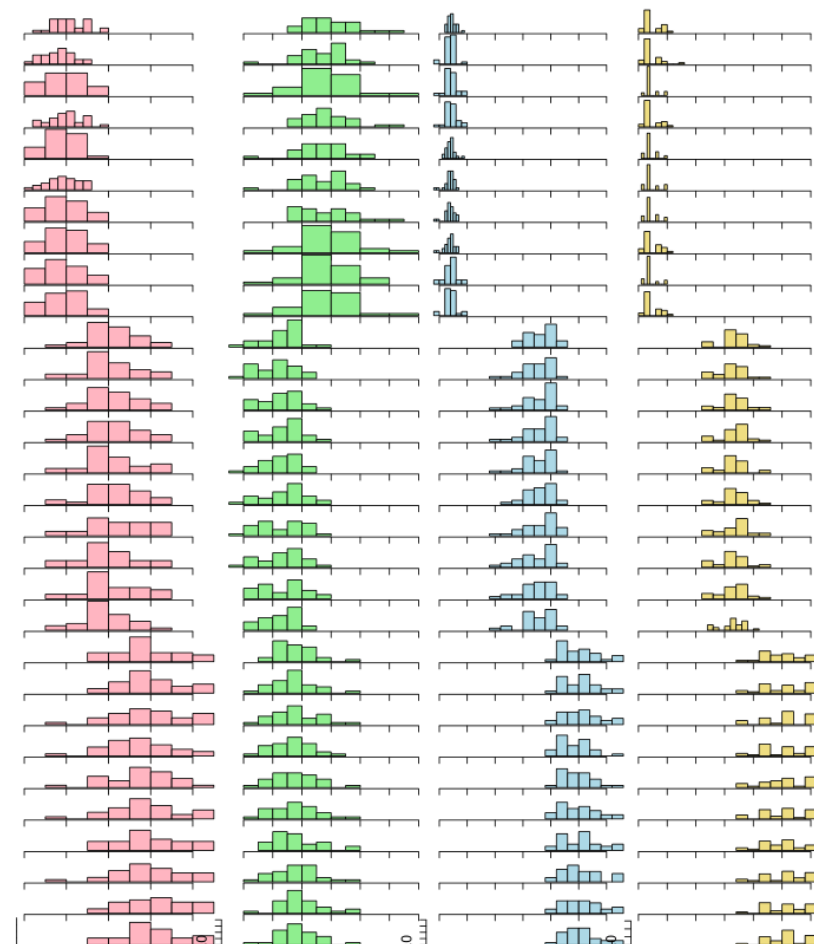
The iris data (Fisher, 1936) consists of 50 samples from each of three species of Iris (Setosa, Virginica and Versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.



The histogram-valued symbolic data

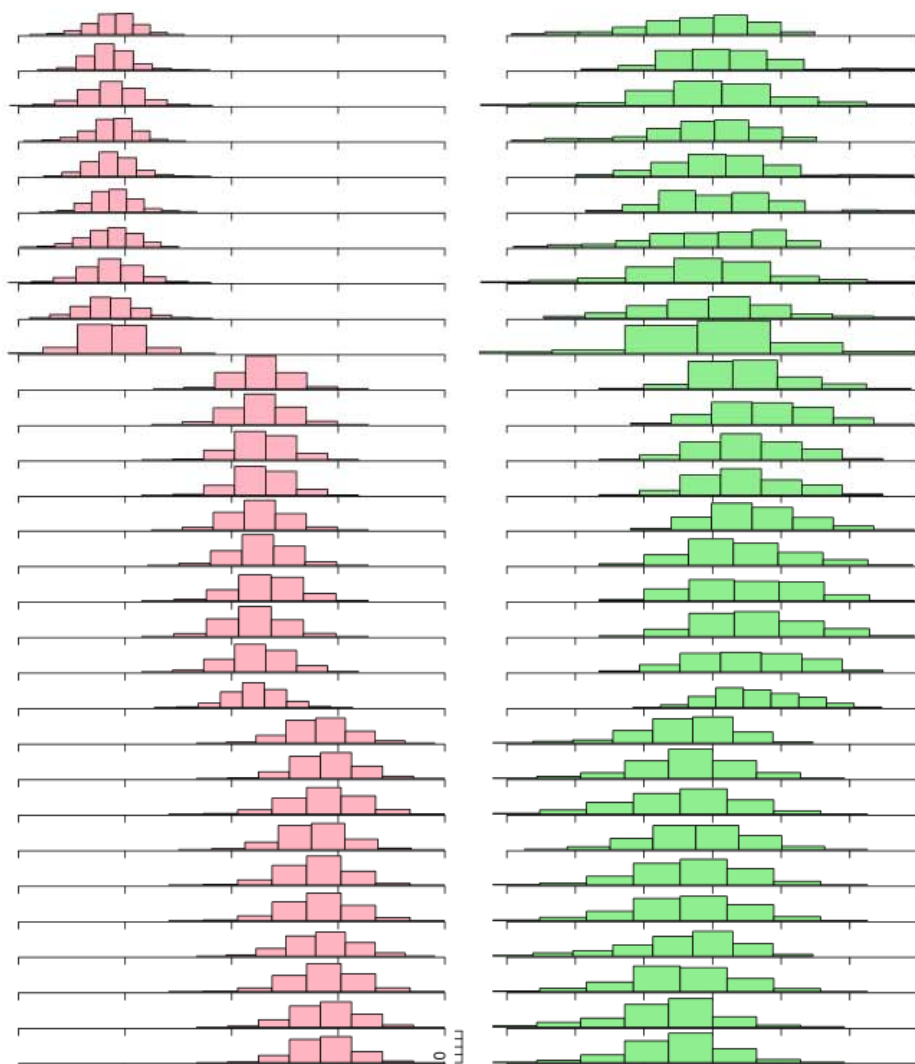
IRIS DATA: GENERATE HISTOGRAMS

- 1 Within each class, 30 observations were randomly sampled to generate histograms for four variables.
- 2 Repeat 10 times.

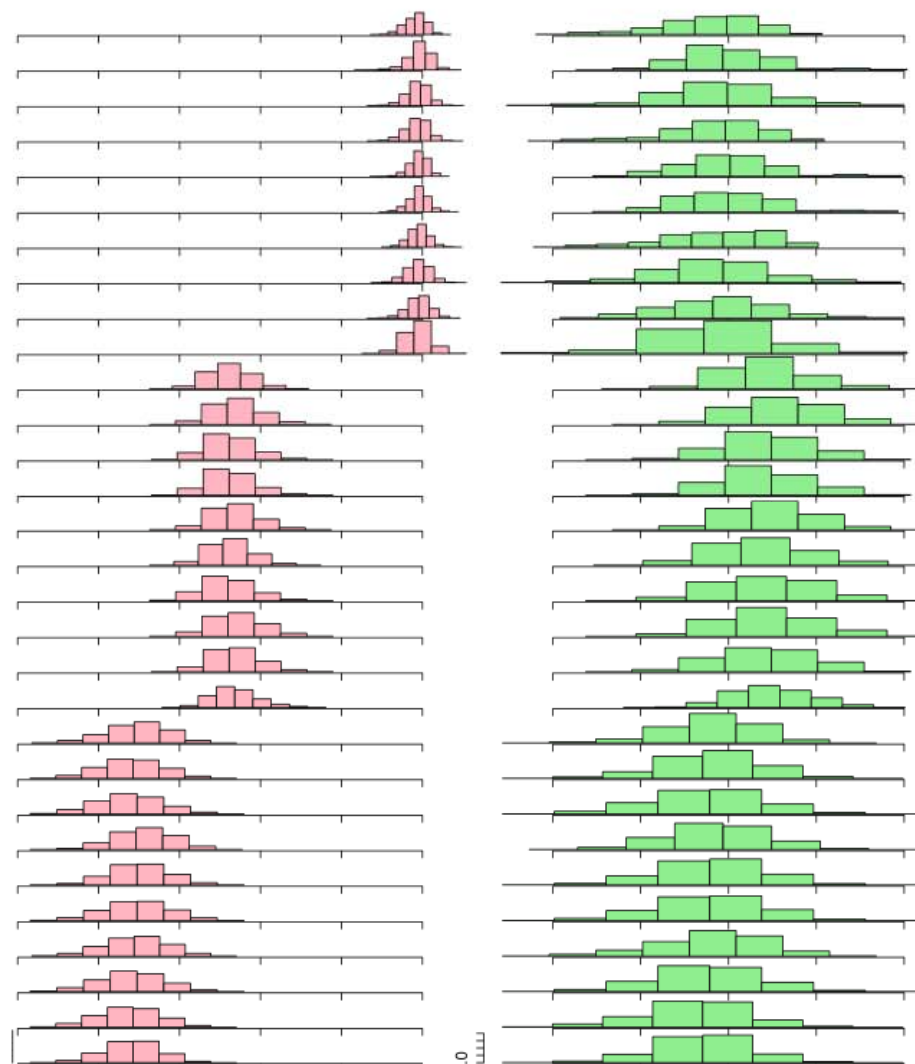




Histogram-PCA

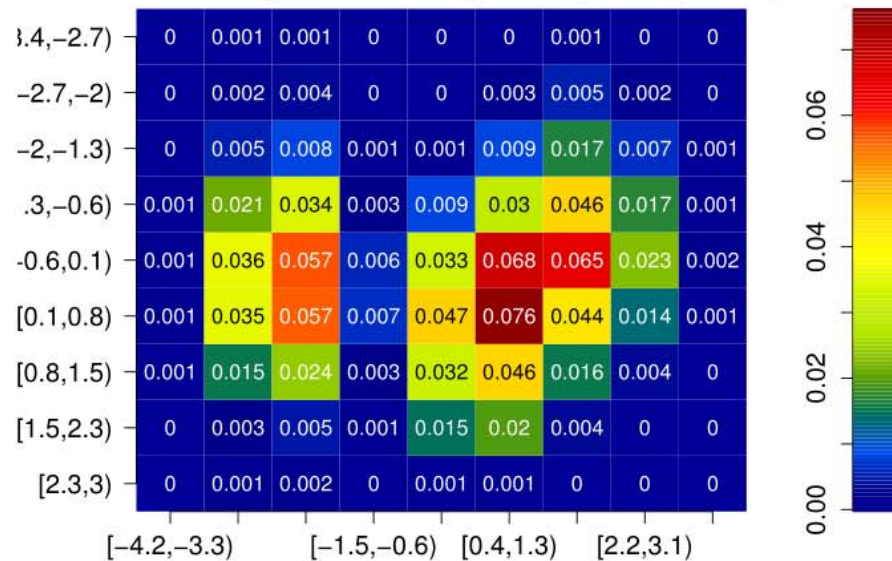


Histogram-SIR



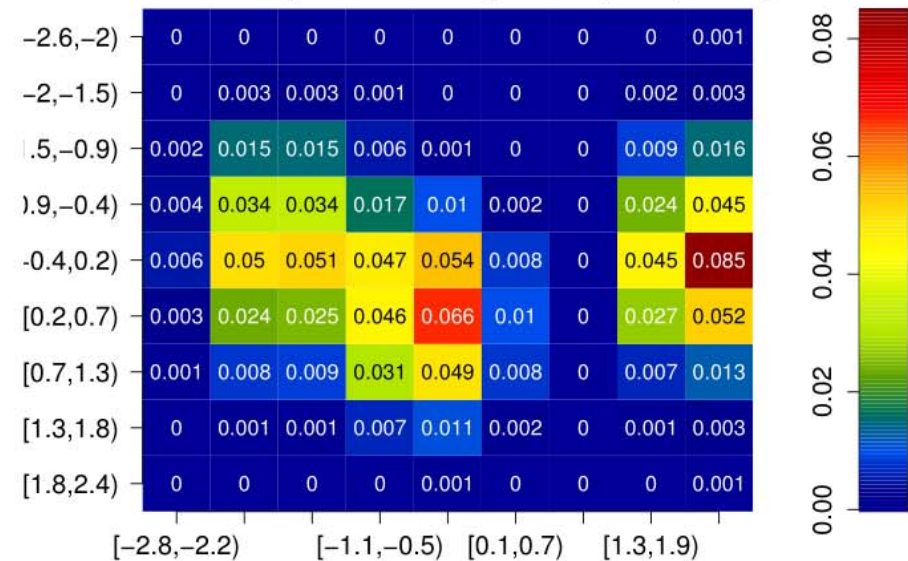
Histogram-PCA

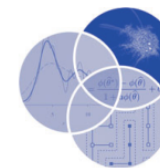
Joint Histogram of Histograms: (PCA1, PCA2)



Histogram-SIR

Joint Histogram of Histograms: (SIR1, SIR2)





Thinking by classes in data science: the symbolic data analysis paradigm

Edwin Diday*

How to cite this article:




WIREs Comput Stat 2016, 8:172–205. doi: 10.1002/wics.1384

Standard data table

Players	Y_1	Y_j	
ind_1			
ind_i		Y_{ij}	
ind_n			

A number
(Messi age)
or a
Category
(Messi
nationality)

Symbolic data table

	Y'_1	Y'_j	
Cl_1			
Cl_i			
Cl_k			

A symbolic
data
describing
Messi team

Age
interval

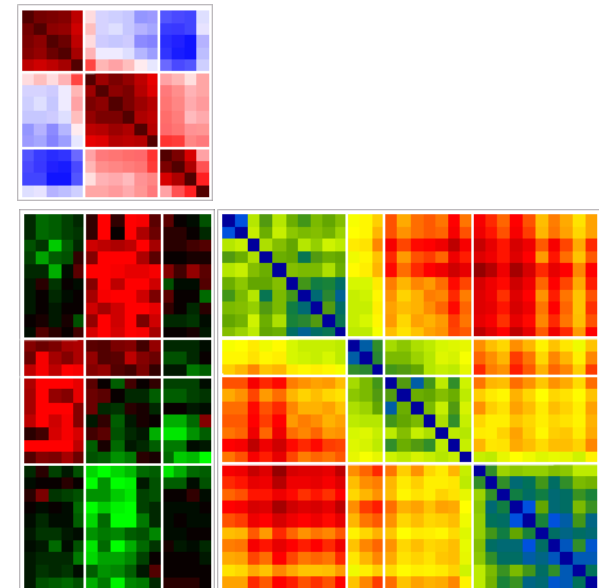
Weight
histogram

Nationalities
barshart

FIGURE 1 | From a standard data table (X, Y) describing a set of individuals X by a set of standard variables Y , to a symbolic data table (X', Y') describing a set of teams X' by a set of symbolic variables Y' .

Conclusion

- MV is the color order-based representation of data matrices.
- MV is suggested as a preliminary step in modern exploratory data analysis and is a continuing and active topic of research and application.
- MV has the opportunity to become one of the new generation of exploratory data analysis (EDA) tool for various data types.



Thanks for listening!

76/76



Han-Ming Wu 吳漢銘
<http://www.hmwu.idv.tw>

國立臺北大學 統計學系
Department of Statistics, National Taipei University

[Home](#) [About Me](#) [Photo Gallery](#) [Facebook](#) [Links](#) [Contact Me](#) [NTPU-107\(下\)課程](#) [【作業考試上傳區】](#)

NTPU-107(下)課程

- ✓ 微積分
- ✓ 電腦概論與程式設計
- 【作業考試上傳區】
- 【歷年課程】

827627

Today	807
This Week	2160
This Month	12240
All days	827627

Your IP: 1.34.216.123
2019-05-14 18:34
[Visitors Counter](#)

Home



2019 ISI Young Statisticians Workshop (YS-ISI2019)
Aug 18, 2019, Kuala Lumpur, Malaysia
<http://www.ys-isi.org/ys-isi2019>



iasc isi
International Association for Statistical Computing



IASC-ARS
The Asian Regional Section of
The International Association for Statistical Computing



The Young Statisticians Group in
The International Association for Statistical Computing
<http://www.ysg-iasc.org>



The Young Statisticians
International Statistical Institute
<http://www.ys-isi.org>



<http://www.hmwu.idv.tw>

R Software (R統計軟體教學)



Exploratory Symbolic Data Analysis

Teaching 教學

- Courses (歷年課程)
- R Software (R統計軟體教學)

Research 研究

- Publication (發表)
- Research (研究)

Service 服務

- Journal Referee (學術審稿)
- Alumni Math (數學系系友會)