Biochip informatics-(II) : Unsupervised Learning

Ju Han Kim, M.D., Ph.D.
Juhan@snu.ac.kr
SNUBI: SNUBiomedical Informatics
http://www.snubi.org/

Unsupervised ~~machine~~ learning

Topics will include

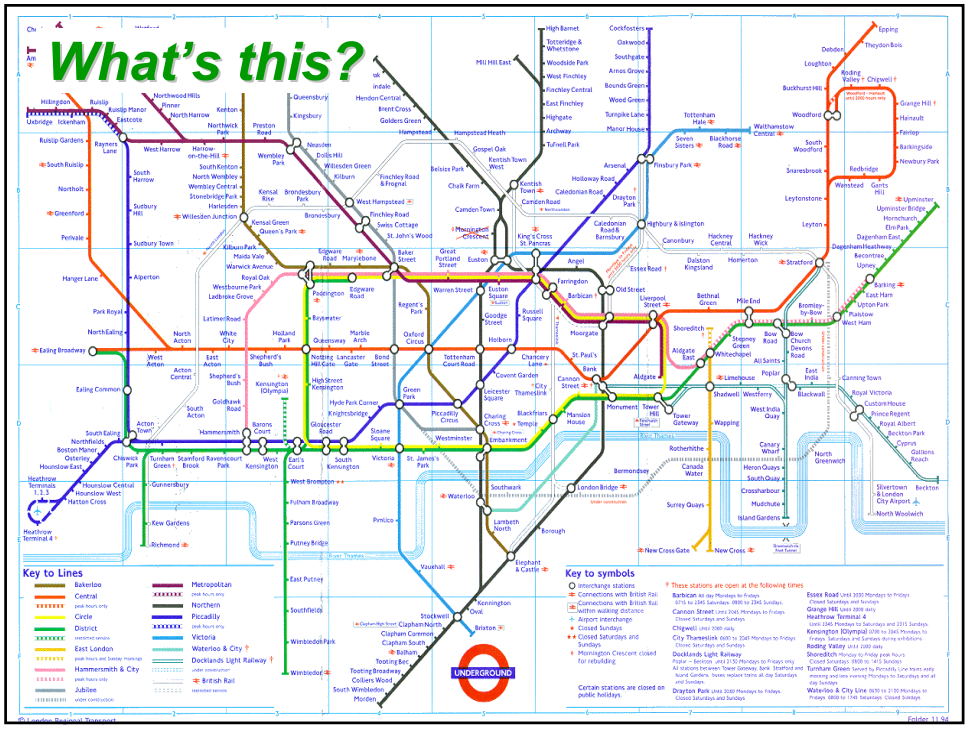
- *Technology perspectives*
- *Cognitive science perspectives*
- *Implementation/Application issues*
- *Biochip informatics*

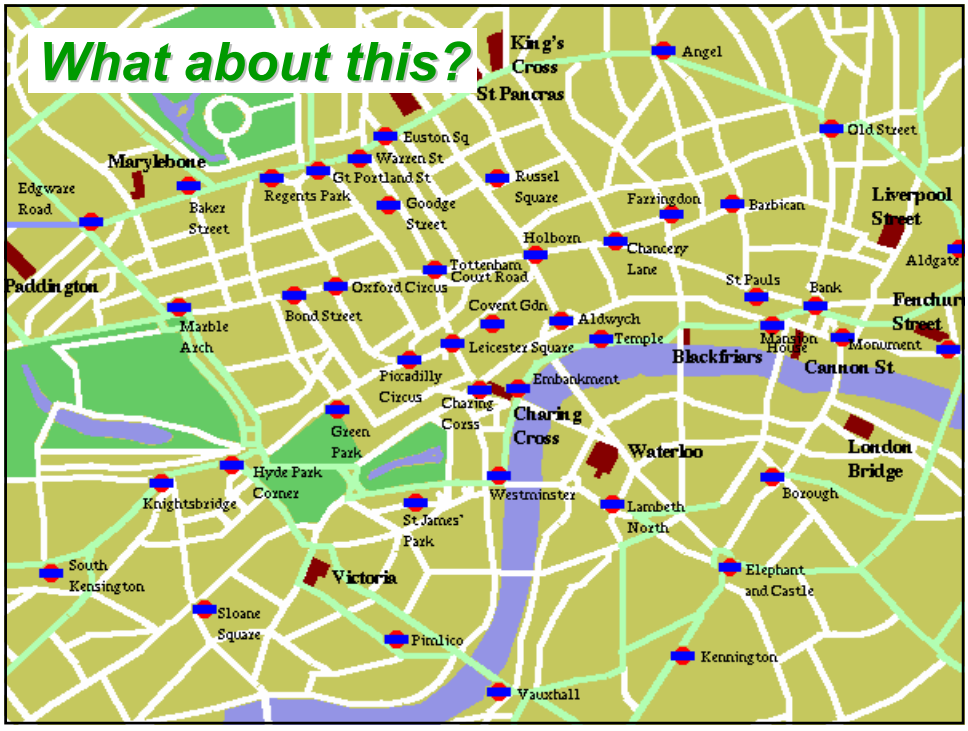
Unsupervised learning

- Prologue – Traveling London
- Observing and organizing complex data
- Inputs and measures
- Data preprocessing and projection
- Multidimensional scaling
- Hierarchical & partitional clustering
- Similarity
- An implementation
- Evaluation & recent developments in clustering
- Epilogue – Traveling Cambridge
- Clustering ideas

Prologue

- Traveling London -


What's this?
Key to Lines
Bakerloo
Central
Circle
District
Victoria
East London
Hammersmith & City
Jubilee
Metropolitan
Northern
Piccadilly
Waterloo & City
Docklands Light Railway
British Rail
Key to symbols
UNDERGROUND


What about this?
King's Cross St Pancras
Angel
Old Street
Euston Sq
Warren St
Gt Portland St
Regents Park
Russel Square
Goodge Street
Marylebone
Edgware Road
Baker Street
Farringdon
Barbican
Liverpool Street
Holborn
Chancery Lane
Aldgate
Paddington
Tottenham Court Road
Oxford Circus
St Pauls
Bank
Fenchurch Street
Covent Gdn
Aldwych
Temple
Marble Arch
Bond Street
Leicester Square
Mansion House
Monument
Cannon St
Blackfriars
Piccadilly Circus
Embankment
Charing Corss
Charing Cross
Green Park
Waterloo
London Bridge
Hyde Park Corner
Knightsbridge
Westminster
Borough
Lambeth North
St James' Park
South Kensington
Victoria
Elephant and Castle
Sloane Square
Pimlico
Kennington
Vauxhall

What about these?

Isn't this transformation obvious?

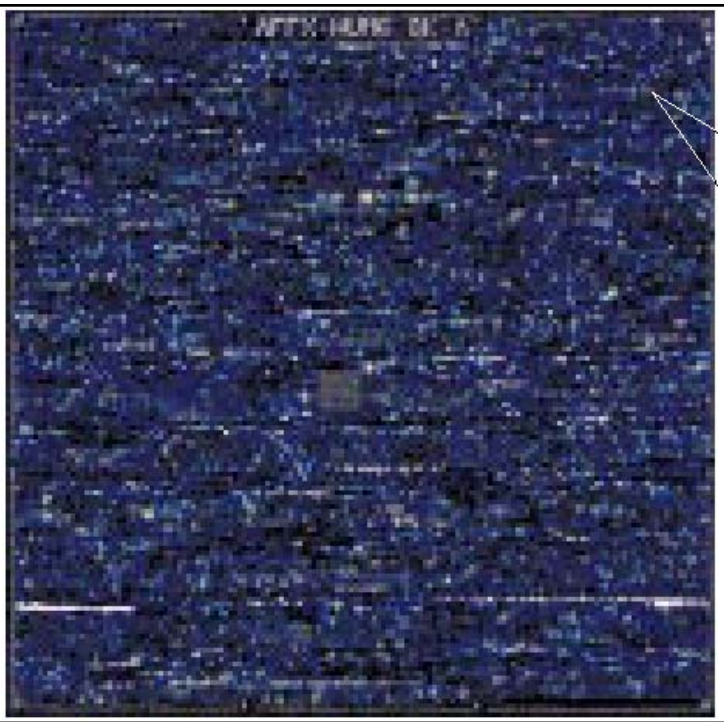
But history says…

Some Questions

- *Is it obvious? Are you sure?*
- *Which one has more information?*
- *Which one do you see more frequently?*
- *Which one do you prefer for your guide?*
- *Is it just about simple vs. complex?*
- *Which one has more valuable information with respect to… what ?*

Astronomer's Learning

Babylonians created the map of starry sky.
and Astronomy started then…



Observations and Clustering

- *Organizing complex data into meaningful structures*
- *Clustering is fundamentally an exploration of data structure*

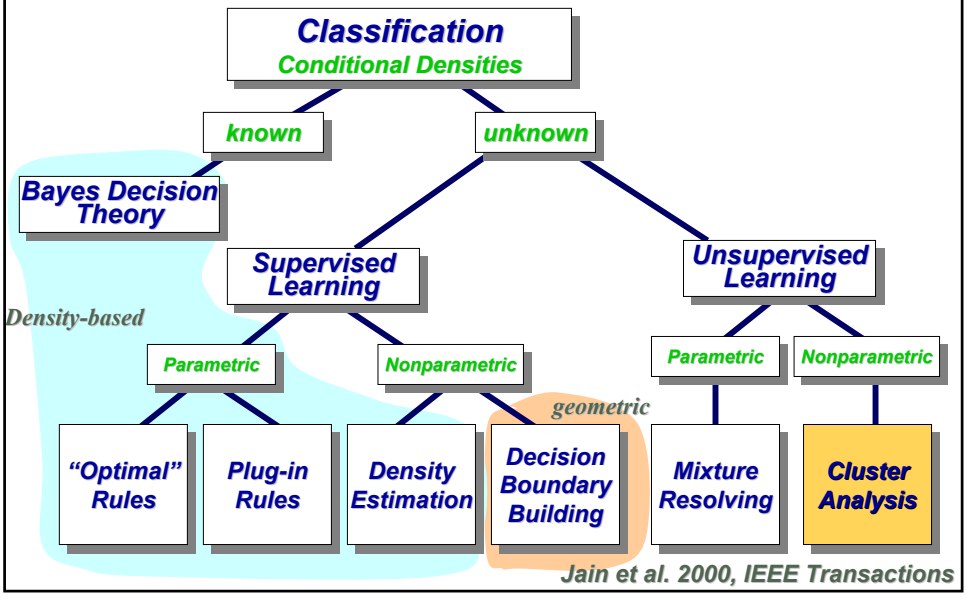
"Exploratory Data Analysis"

involving decision making
in non-parametric and unsupervised mode
and visualizing and projecting multidimensional data

Supervised vs. Unsupervised Classifications

- **Supervised Learning**
  - Artificial Neural Network
  - Classification Tree
  - Bayesian Belief Network
  - Boolean Network
  - Rough Set
  - Reinforcement Learning
  - Support Vector Machine
- **Unsupervised Learning**
  - Multi-dimensional Scaling
  - Graphic Representation of Multivariate Data
  - Clustering
  - Mixture resolving
  - Self-Orgnizing Feature Maps

Supervised vs. Unsupervised Classifications



Jain et al. 2000, IEEE Transactions

Why now?

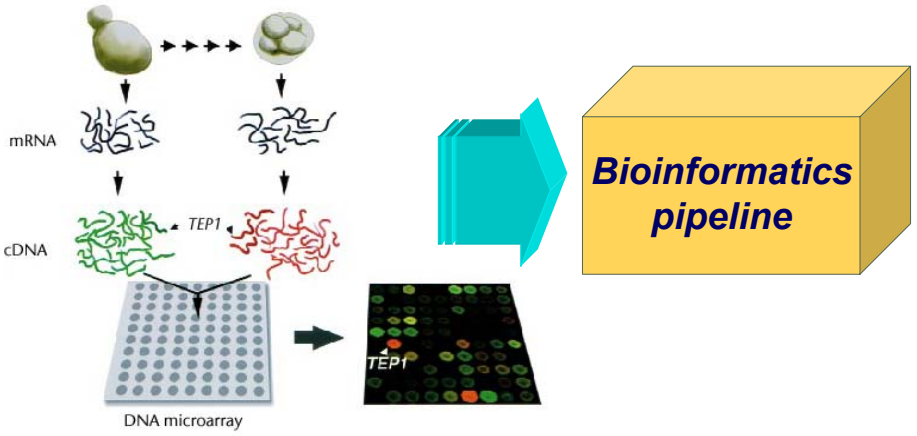
Computational power & Data Explosion

- Bioinformatics
- Data mining
- Speech recognition
- Multimedia data retrieval and analysis
- Biometry
- the Internet

Genomics and Clustering

	Measures	Algorithms
• Spellman et al.	Shapes	Eyes/Hands
• Eisen et al.	Correl	Hierarchical Tree
• Alon et al.	Correl	Simulated Annealing
• Tamayo et al.	Distance	SOM
• Butte & Kohane	M.I.	Relevance Networks
• Sharan & Shamir		Graph theory

Biochip basics



Biochip, Core competency

- They are the genes!
- We have the map!
- Natural measure of quantification.
- Literally, INFINITE # of states
- Dynamic series on time & space
- Don't need to extract bio-molecules.
- Now systemic perturbations!

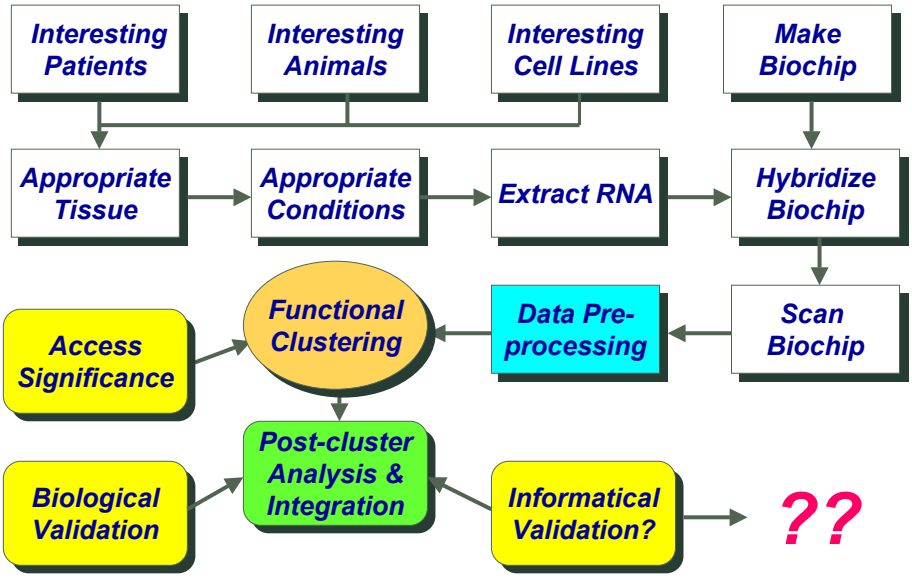
Streamlining & automation of the process

put abstraction barrier

Do biology in silico! **********

********** Say NO to lab bench!

A Biochip Informatics Strategy



Biochip informatics: challenges

- **Pre-processing:**
  - ✓ technology variation
  - ✓ noise & data filtering
  - ✓ missing / negative values / P & A calls
  - ✓ data scaling
  - ✓ Can I assume normality?
  - ✓ chip quality, other artifacts
- **Functional Clusters:**
  - ✓ clustering quality, consistency, & robustness
- **Statistical Issues:**
  - ✓ study design / # of replicates / multiple testing
- **Integrative Biochip Informatics**
  - ✓ Can we get more out of it?

Input Data

Data representaion

- pattern matrix
- proximity matirx

Data types

- degree of quantization (Anderberg, 1973)
- binary / discrete / continuous

Data scales

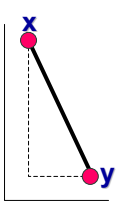
- Qualitative: nominal / ordinal
- Quantitative: interval / ratio

Proximity Measures

Minkowski metric

$$d(i,k) = \left(\sum_{j=1}^{d} | x_{ij} - x_{kj} |^{r} \right)^{1/r}$$

- Euclidean distance
- Manhattan (taxicab, city block)
- Chebychev ('sup')



Power distance

$$d(i,k) = \left(\sum_{j=1}^{d} | x_{ij} - x_{kj} |^{p} \right)^{1/r}$$

Correlation

Mutual Information

Symbol String distance

- Percent disagreement: (Hamming/Manhattan)
- Levenstein (Edit) distance
  LD(A, B) = min{a(l) + b(l) + c(l)}
  dynamic programming

Data Preprocessing

- Normalization
- Filtering
- Plotting
  - ✓ linear projection
    - Eigenvector projection
    - Mean square error
    - Discriminant function
  - ✓ non-linear projection
    - Graphical
    - Iterative mapping

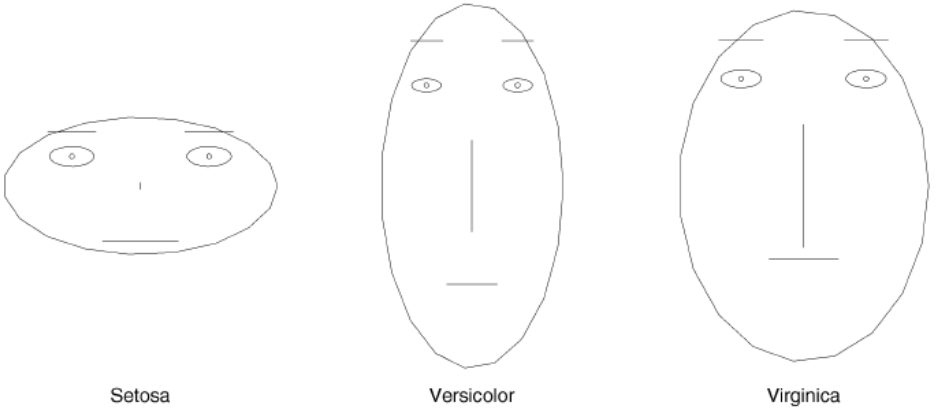
Feature selection & Feature extraction

- How many features?
- In what combination?

- The peaking phenomena, the curse of dimensionality, and the intrinsic dimensionality problem

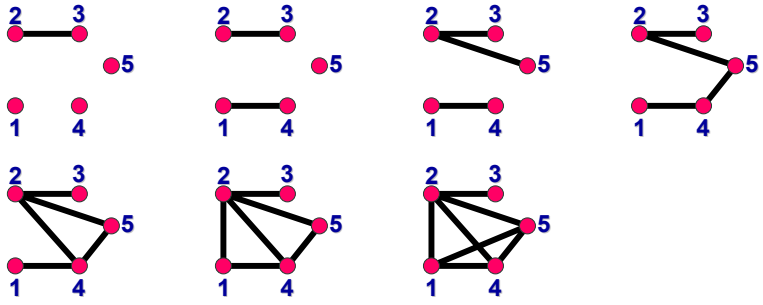
$$m_1 = (1, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{3}}, \cdots, \frac{1}{\sqrt{d}}) \text{ and}$$
$$m_2 = (-1, -\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{3}}, \cdots, -\frac{1}{\sqrt{d}}).$$

Graphical Representation of Multivariate Data



Chernoff Faces corresponding to the mean vectors of Iris Setosa, Iris Versicolor, and Iris Virginica

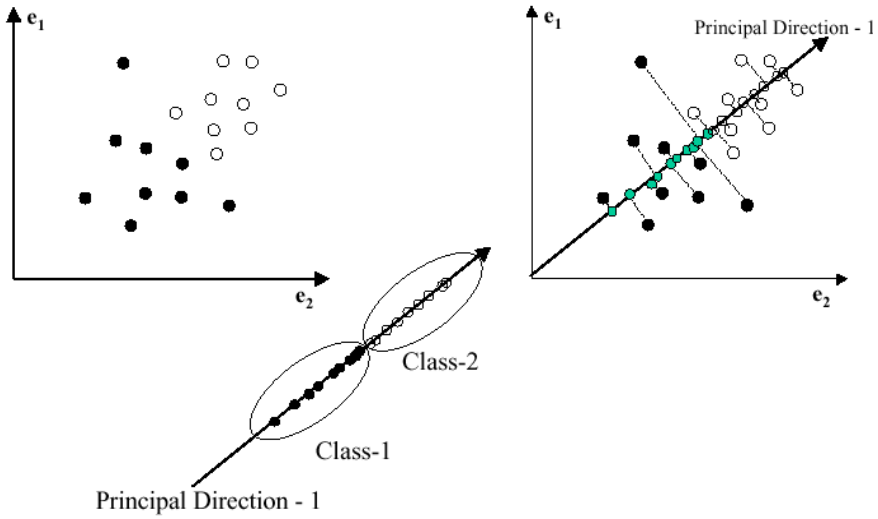
Threshold Graph, Single-link and complete-link hierarchical clusterings



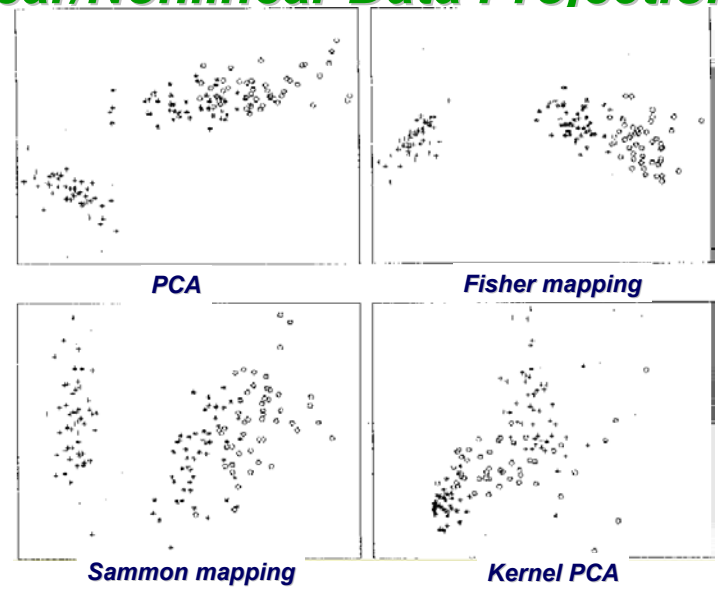
single-link *complete-link*

Principal component analysis

Basic Idea

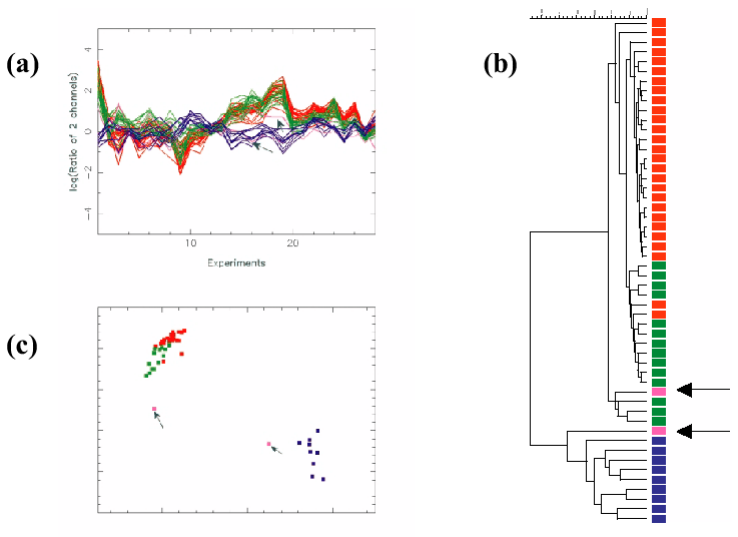


Linear/Nonlinear Data Projections



Two-dimensional Mappings of the Iris data set (Jain et al., 2000)

Nonlinear Sammon Mapping



Multi-dimensional Scaling (MDS)

A similarity structure analysis by visual configuration.

- ***Classical (metric, one matrix, unweigthed)***
- ***Metric: Torgerson***
  Non-metric: Sephard, Kruskal
  $s_{ij}=f(d_{ij})$***, monotonic regression, stress***
- ***Replicated***
- ***Weighted***

Multi-dimensional Scaling

Table 1 Flying Mileages Between 10 American Cities

Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington, DC	
0	587	1212	701	1936	604	748	2139	2182	543	Atlanta
587	0	920	940	1745	1188	713	1858	1737	597	Chicago
1212	920	0	879	831	1726	1631	949	1021	1494	Denver
701	940	879	0	1374	968	1420	1645	1891	1220	Houston
1936	1745	831	1374	0	2339	2451	347	959	2300	Los Angeles
604	1188	1726	968	2339	0	1092	2594	2734	923	Miami
748	713	1631	1420	2451	1092	0	2571	2408	205	New York
2139	1858	949	1645	347	2594	2571	0	678	2442	San Francisco
2182	1737	1021	1891	959	2734	2408	678	0	2329	Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington, DC

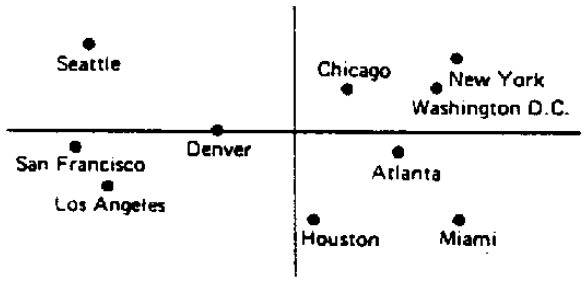
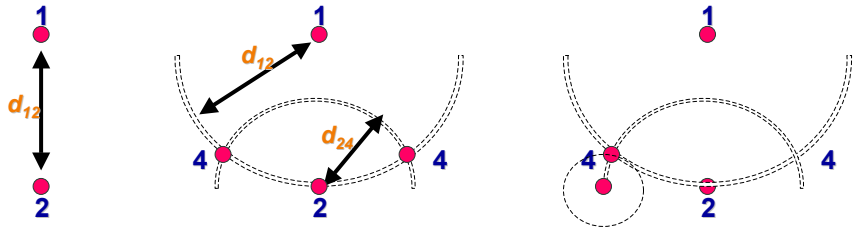
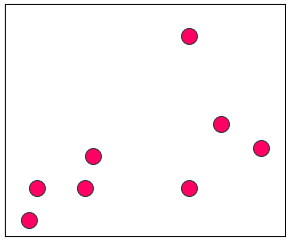


Figure 1 CMDS of flying mileages between 10 American cities.

Young, 1985. Encyclopedia of Statistical Sciences

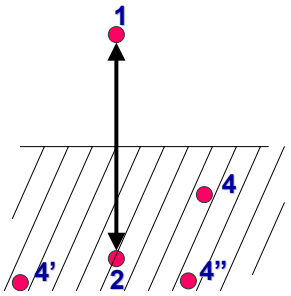
Multi-dimensional Scaling (MDS)

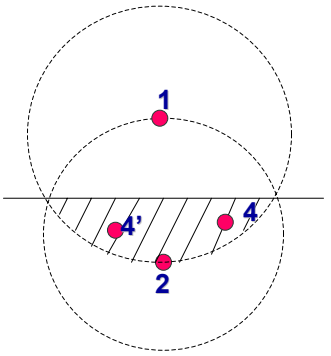




Multi-dimensional Scaling (MDS)

Space for ordinal data

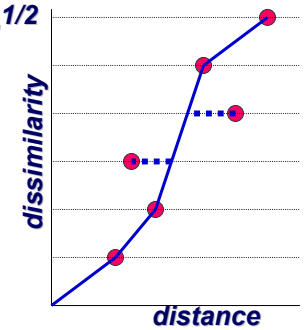




Non-metric MDS

- *Ordination:*

  translating ordinal scale to set of ratio scale
- *Shepard diagram*
- *Stress* = $\{\Sigma(d_{ij}-\delta_{ij})^2 / \Sigma d_{ij}^{\,2}\}^{1/2}$



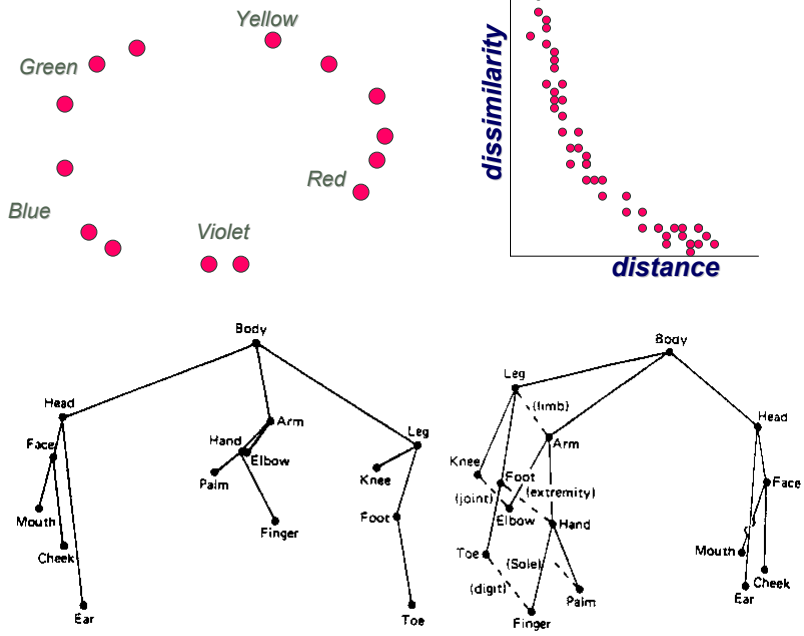
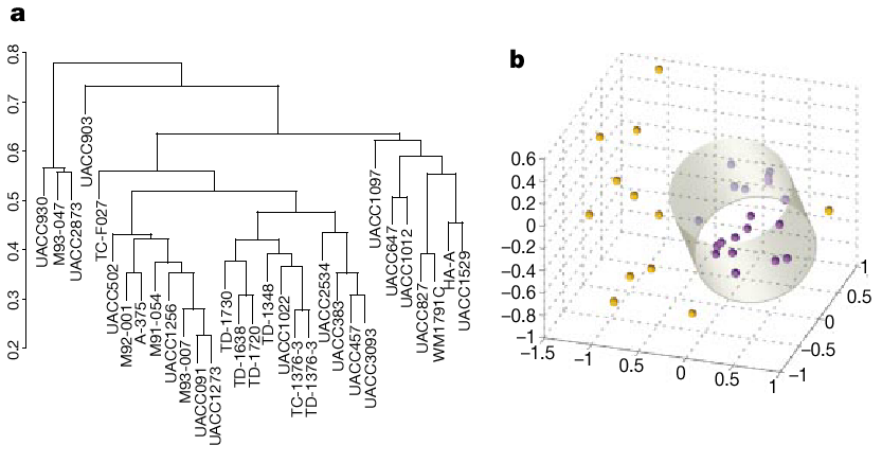


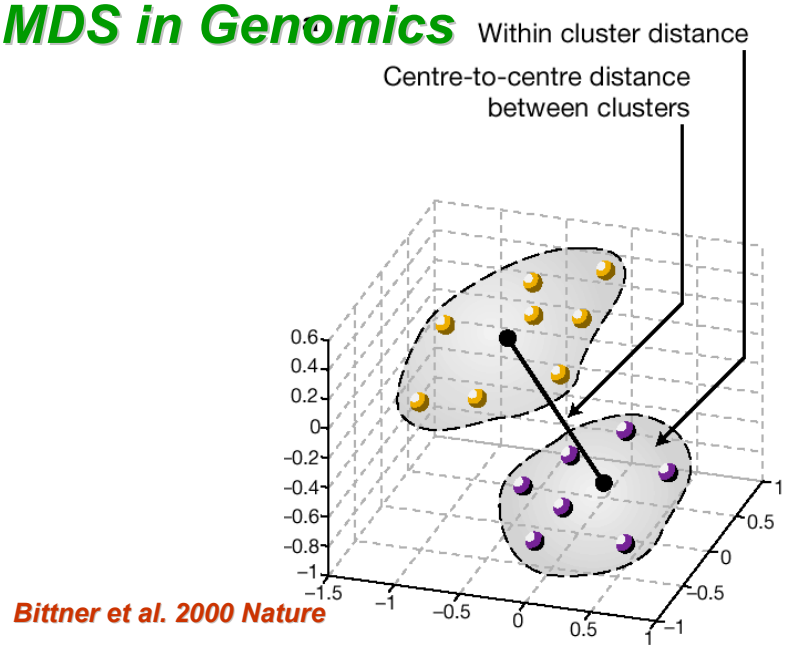
Figure 2 (a) RMDS of children's similarity judgments about 15 body parts; (b) RMDS of adults' similarity judgments about 15 body parts.

MDS in Genomics



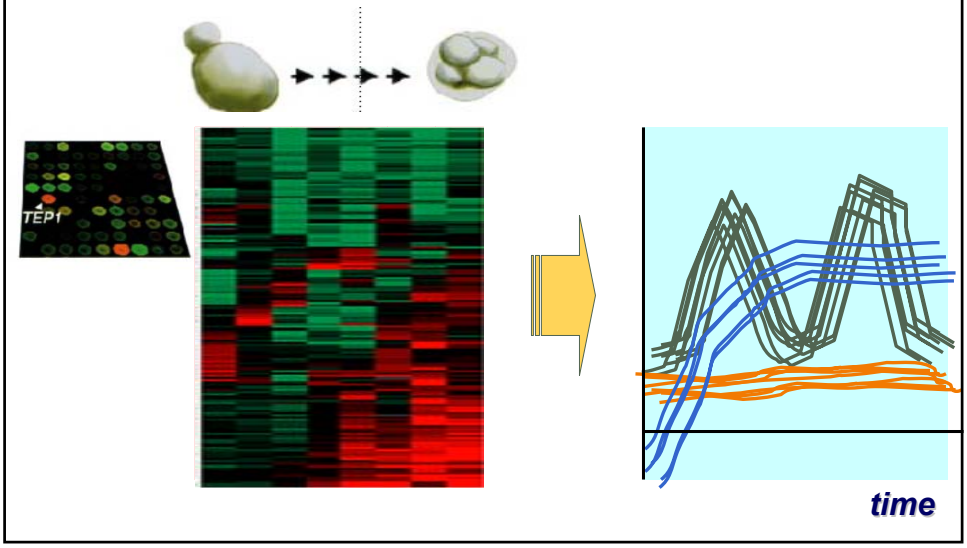
Bittner et al. 2000 Nature

MDS in Genomics

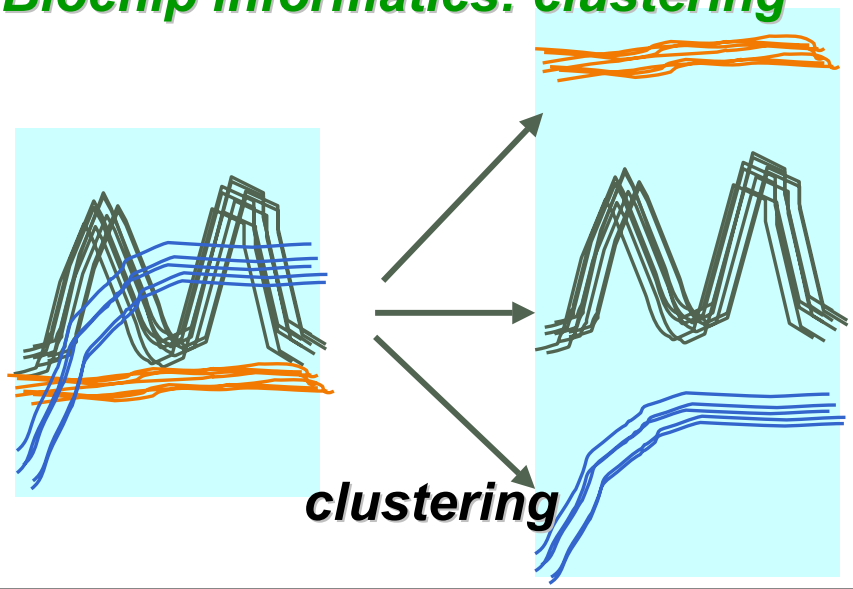


Bittner et al. 2000 Nature

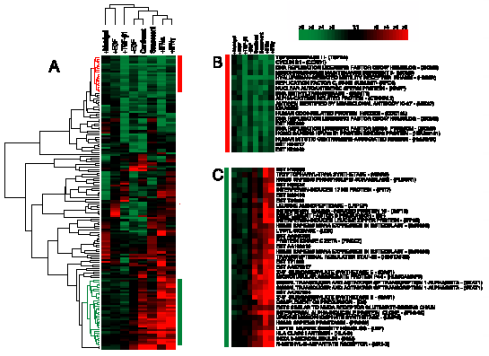
Biochip informatics: clustering



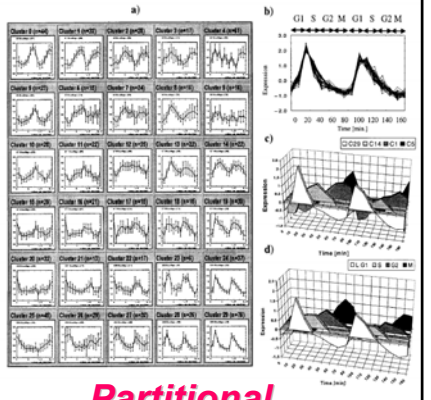
Biochip informatics: clustering



Hierarchical & Partitional Clustering



Hierarchical



Partitional

Dichotomies of Clustering Algorithms

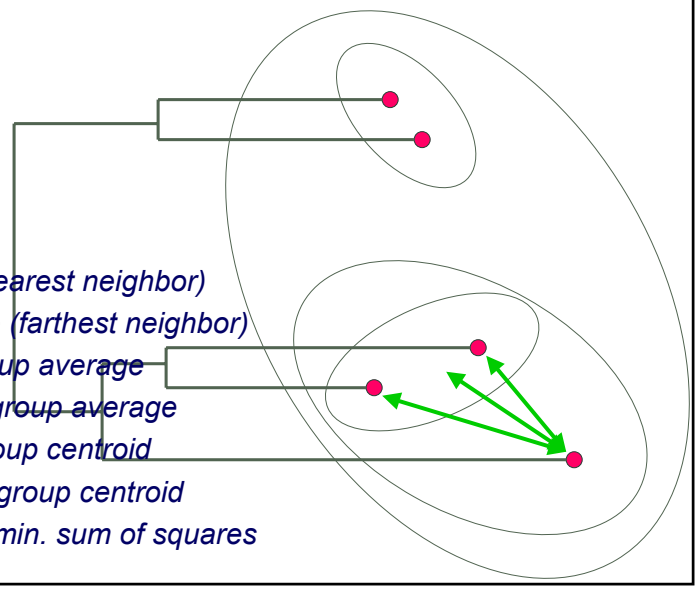
- *Hierarchical vs. Partitional*
- *Divisive vs. Agglomerative*
- *Graph theory vs. Matrix Algebra*
- *Exclusive vs. Non-exclusive*
- *Serial vs. Simultaneous*
- *Parametric vs. Non-parametric*

Most commonly,

agglomerative hierarchical clustering (dendrogram)

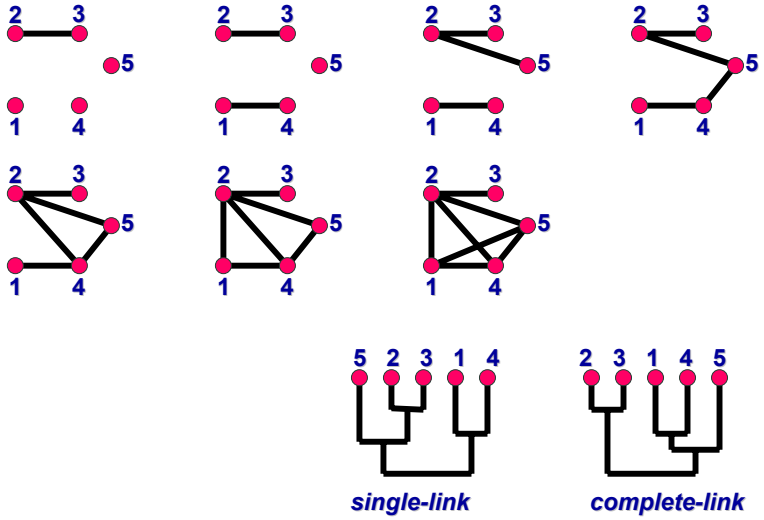
iterative square-error partitioning (K-means)

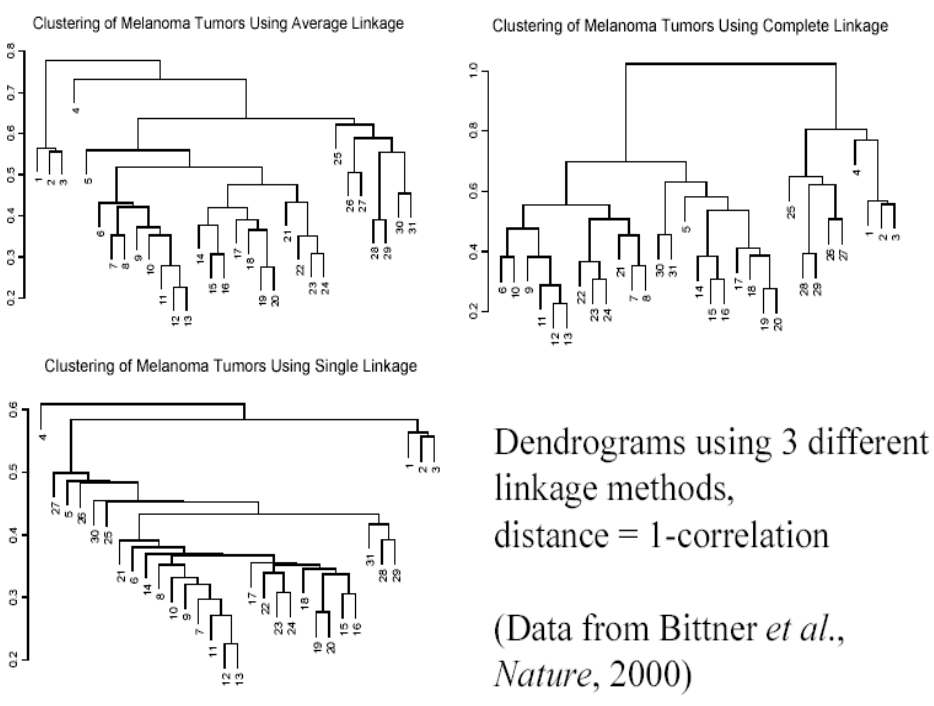
Hierarchical clustering in Genomics



- *single-linkage (nearest neighbor)*
- *complete-linkage (farthest neighbor)*
- *weighed pair-group average*
- *unweighed pair-group average*
- *weighted pair-group centroid*
- *unweighted pair-group centroid*
- *Ward's method: min. sum of squares*

Threshold Graph, Single-link and complete-link hierarchical clusterings

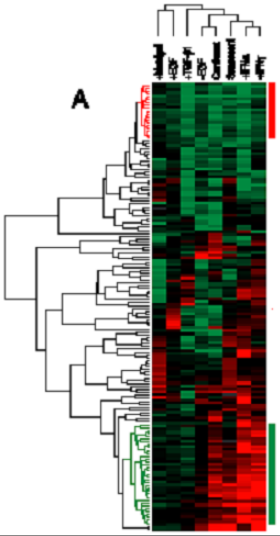




Dendrograms using 3 different linkage methods, distance = 1-correlation

(Data from Bittner *et al.*, *Nature*, 2000)

Hierarchical clustering in Genomics



- *deterministic by heuristics*
- *linear optimization: 2^{N-1}*
- *arbitrary cluster thresholding*
- *sensitive to a small perturbation*
- *difficulty in comparing diff. trees*
- *good for data the structure of which is inherently hierarchical*

Hierarchical clustering in Genomics

Fast optimal leaf ordering

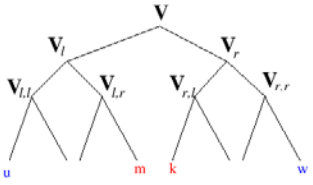
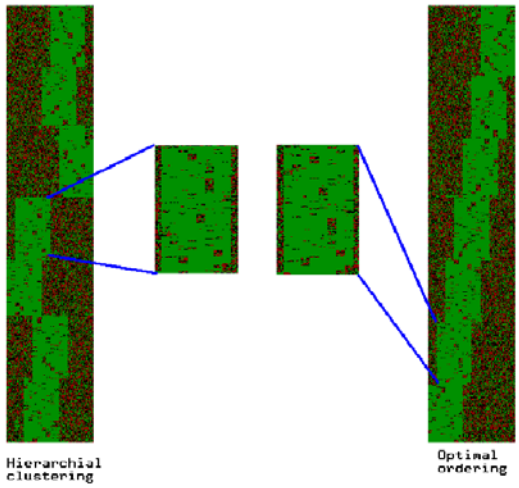


Fig. 2. For every pair of leaves $u \in v_l$ and $w \in v_r$ our algorithm computes $M(v, u, w)$ which is the optimal linear ordering when the leftmost leaf of v is u and the rightmost leaf of v is w. Note that when computing $M(v, u, w)$ we must have a leaf $m \in v_{l,r}$ as the rightmost leaf of v_l and $k \in v_{r,l}$ as the leftmost leaf of v_r.



K-means Algorithm

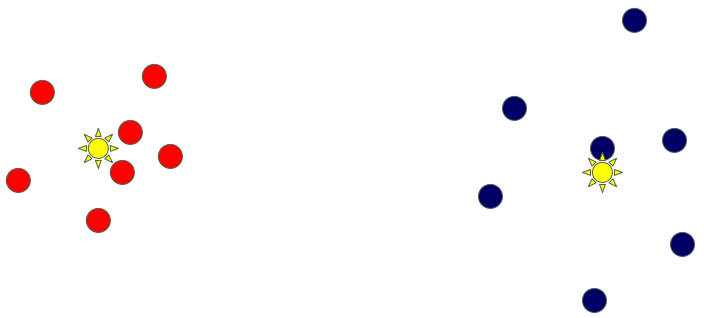
- 1. Select an initial partition with K clusters.

Repeat 2 & 3 until the cluster membership stabilizes

- 2. Calculate cluster center
- 3. Update partition by assigning each pattern to its closest cluster center

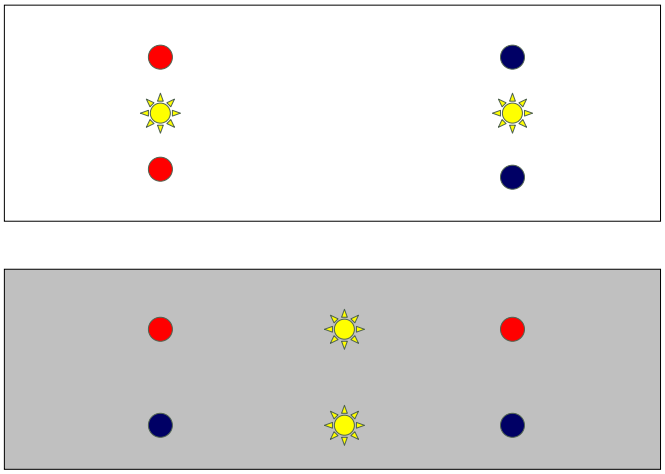
- 4. Adjust the number of clusters…

K-means Algorithm (K=2)



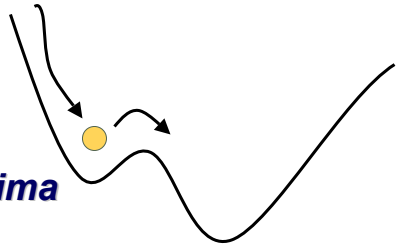
Convergence!!

K-means Algorithm : local search



Meta-heuristics

- Simulated Annealing
- Genetic Algorithms
- Tabu Search
- Hybrid Approach



* The long-narrow valley problem

K-means Algorithm

- Convergence (Selim and Ismail, 1984)
- Needs number of clusters (K)
- Center-based
- Iterative square-error-based partitioning
  (ANOVA in reverse)
- But greedy
- Sensitive to the initial partition
- Stopping criterion
- Updating the partition
- Assumption? about the data distribution

Variants of K-means Algorithm

- Fuzzy K-means
- Genetic algorithms
- Simulated/Deterministic annealing
- Tabu search
- Mapping it onto a neural network

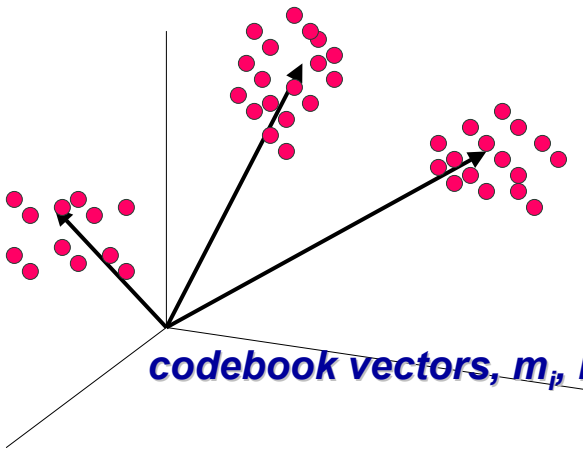
  .....

Vector Quantization (VQ)

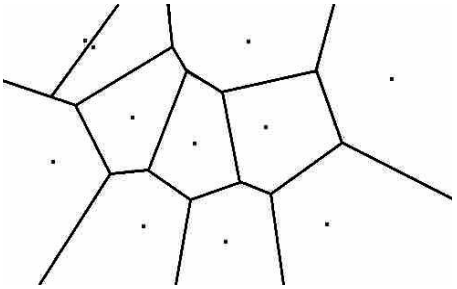
VQ is a classical signal-approximation method that usually forms a quantized approximation to the distribution of the input vectors using a finite number of so-called codebook vectors, usually in the Euclidean metric.

Once the codebook is chosen, the approximation of an input vector is to find the codebook vector close to the input vector.

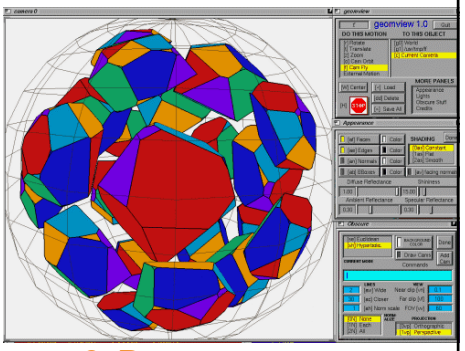
Vector Quantization (VQ)



Voroni Tessellation



2-D



3-D

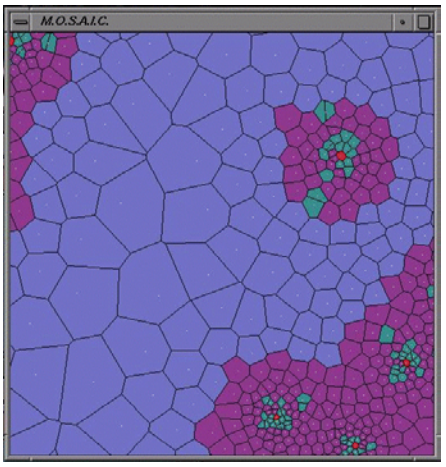
Useful illustration of VQ

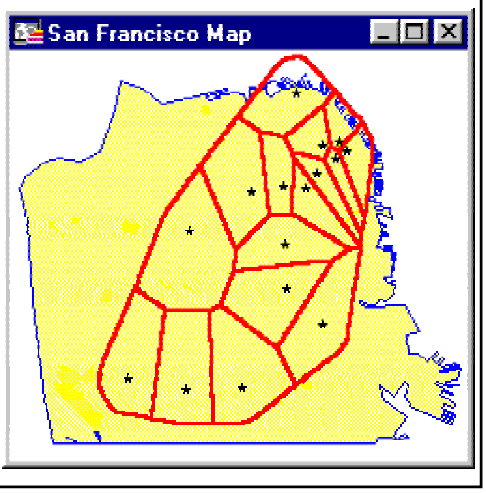
Voroni Tessellation

Not only in mathematics but also in Nature

Voroni Tessellation

… and in computations





Sequential K-means algorithm

Especially when the nature of the problem changes over time

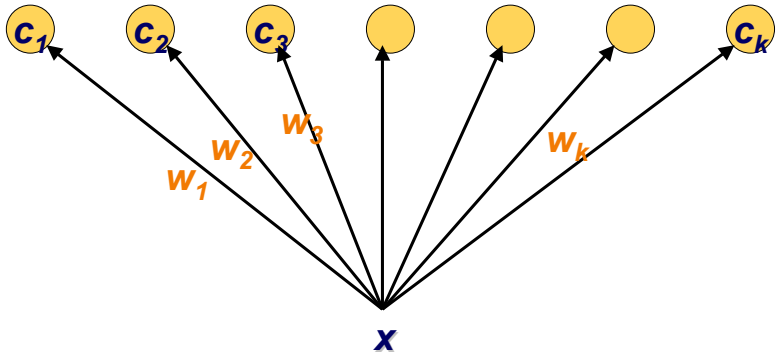
1. Set initial guess of centers, $\{c_1, c_2, \ldots, c_k\}$
2. Set counts, $\{n_1, n_2, \ldots, n_k\}$, to 0

Until interrupted

3. Get next example x
4. Get the closest center c_i to x
5. n_i++
6. Update $c_i = c_i + (1/n_1) * (x - c_i)$

Wait

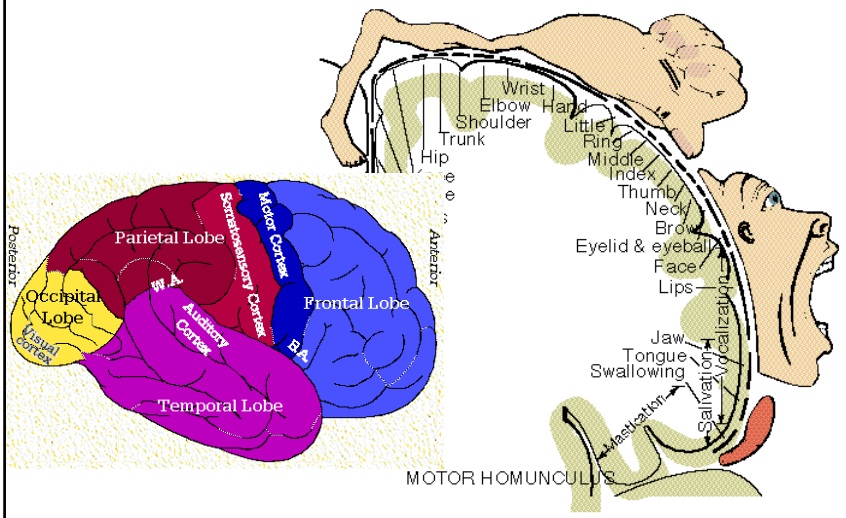
Interpreting clustering by ANN



- *Developing neural "memory" of a typical pattern*
- *By adjusting the "firing" rate of the nearest unit to a newly input pattern*

What if we can preserve the intervening patterns?

Somatotopic Maps



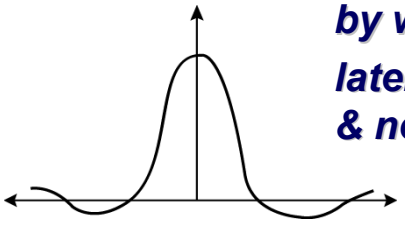
Kohonen Maps

- *" we should not update only c_i, but that we we should also update other cluster centers in the neighborhood of c_i along the layout."*

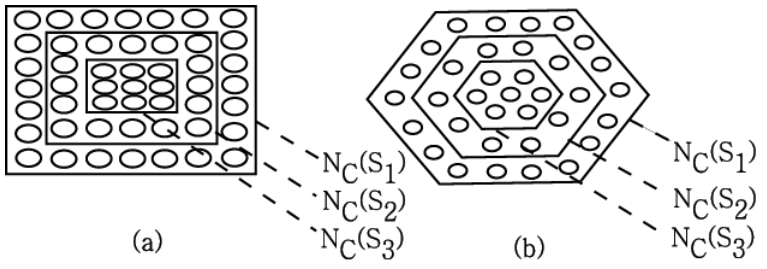
- *"the size of the neighborhood should be programmed to shrink as time goes on."*

Preserving the structure of clustering

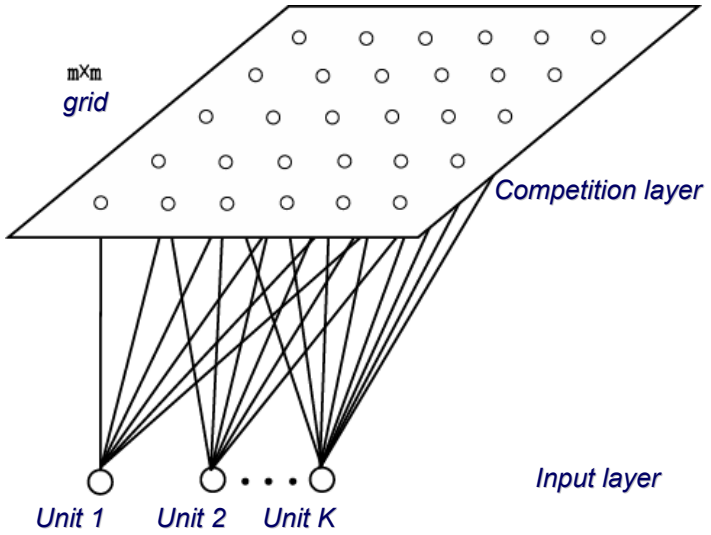
Competitive Learning



by way of
lateral inhibition (sombrero)
& neighbor competition



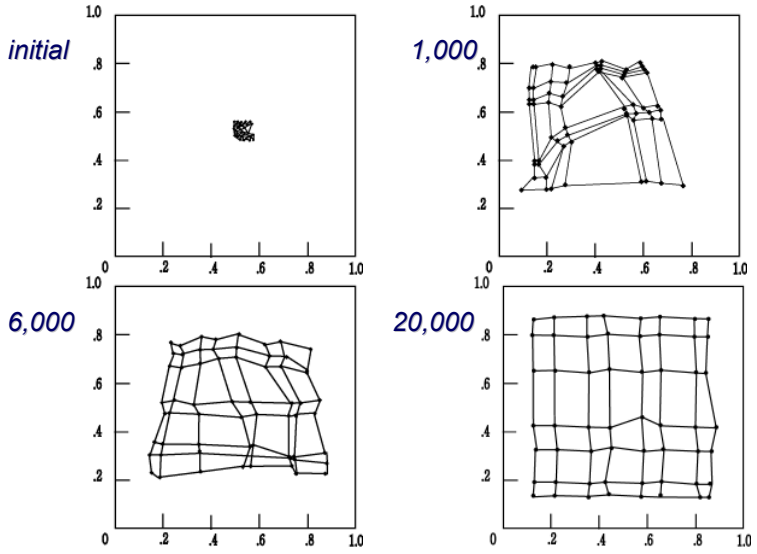
Kohonen Maps



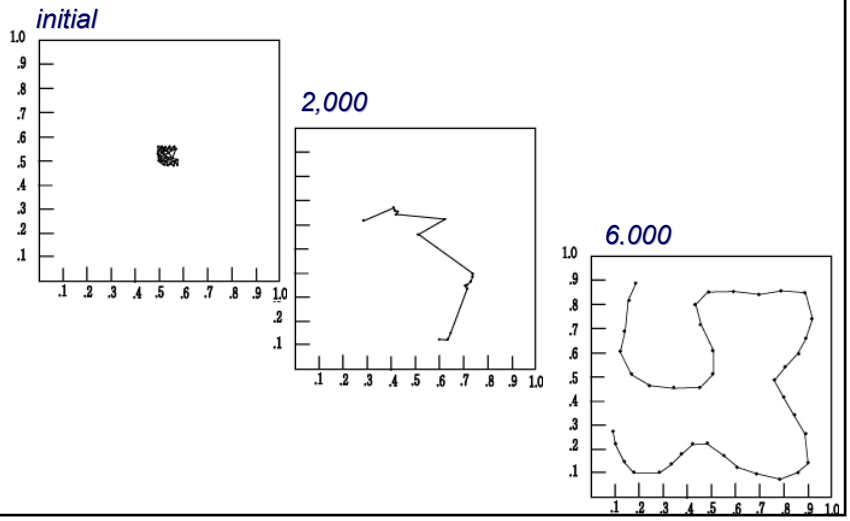
"Topology Preserving" Self-Organizing Feature Maps

- single feedforward: no back propagation
- non-hierarchical
- competitive:
  - winner takes it all
  - lateral inhibition, sombrero
- cooperative
  - winner and neighbor
- $w_{new} = w_{old} + \alpha (x - w_{old})$
- sequential learning
- fast (real time?) learning

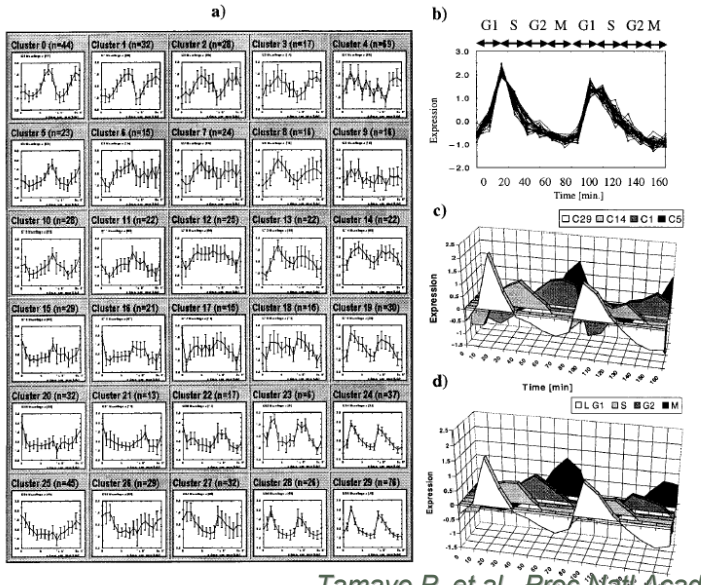
SOM Training Visualization



Linear self-organization of 2-D vector



SOM in Genomics



Tamayo P, et al., Proc Natl Acad Sci USA

Mixture Decomposition & EM Algorithm

Formal approaches to unsupervised classification

Mixture distribution: $p(y|\theta)=\Sigma\, \alpha_m\, p_m\,(y|\theta_m)$

- K random samples with density function $p_m(y|\theta_m)$,

- Each time a sample is to be generated, we choose one of these sources, with probabilities, $\{\alpha_1, \alpha_2, .., \alpha_k\}$

- estimating the parameters and the # of clusters

EM (expectation-maximization) algorithm

MCMC (Markov Chain Monte-Carlo) method

Similarity

Similarity

Similarity

Similarity

similarity

What does "similar" really mean?

Learning is impossible without ASSUMPTIONS

- ***Watanabe's Ugly Duckling theorem***
- ***Wolpert's No Free Lunch theorem***
- ***Michelle's Version Spaces***
- ***Schaffer's Conservation Law***

Classification is impossible without some sort of bias

Ugly Duckling Theorem

Satosi Watanabe, 1969

- *If we have no preconceived ideas or (inductive) bias about what sorts of categories are "natural" or "normal" and what aren't.*
- *It is possible to make two arbitrary patterns similar by encoding them with a sufficiently large number of redundant features.*
- *Because we have to consider all possible classes and possible ways of making sets out of the n objects.*

No Free Lunch Theorem

1. ***Theorem:*** *For any two algorithms, A and B, there exist datasets for which algorithm A will outperform algorithm B in prediction accuracy on unseen instances.*
2. ***Proof:*** *Take any Boolean concept. If A outperforms B on unseen instances, reverse the labels and B will outperform A.*
3. ***Extension:*** *For discrete spaces, the number of concepts for A outperform B in prediction accuracy is equal to the number for which B will outperform A.*

Useful Assumptions & Properties

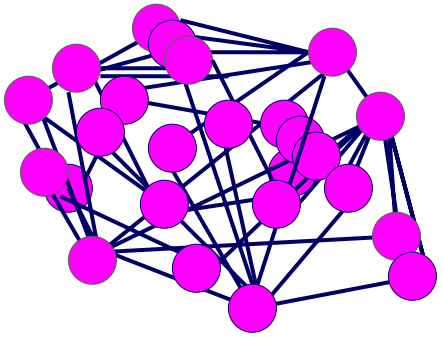
- *Smoothness assumption*
  - ✓ *Statisticians have used this for years.*
- *Few attributes assumption*
  - ✓ *Curse of dimensionality and peaking phenomena.*
    - ▫ *everything is far in high dimensions*
  - ✓ *Ockham's razor:*
    - ▫ *among the competing theories, the simplest is preferred of the more complex.*
- *Properties*
  - ✓ *test drive / accuracy / loss / error / ROC curve*
  - ✓ *comprehensiveness / interpretability/practicality*

Partitioning is…

- *the simple process of*
  - *selecting a criterion,*
  - *evaluating it for all possible partitions,*
  - *and selecting the partition that optimizes the criterion.*
- *However, … is the astronomical Stirling number of the second kind.*
  - *S(19, 4) = 11,259,666,000*
  - *$S(n, 2) = 2^{n-1}$*

Where would you cut?

into meaningful substructures!



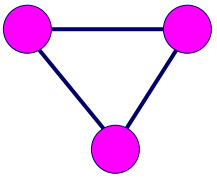
***Metric:** How do you define the best incision?*
***Algorithm:** How do you find and evaluate it?*

Where would you cut?

into meaningful substructures!

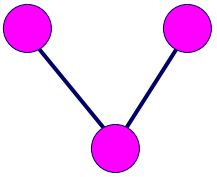
Where would you cut?

into meaningful substructures!



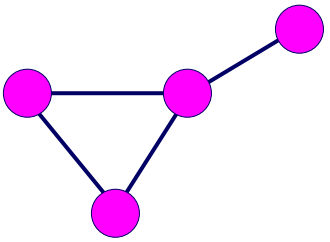
Where would you cut?

into meaningful substructures!



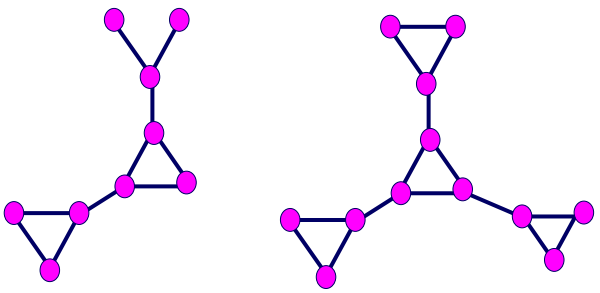
Where would you cut?

into meaningful substructures!

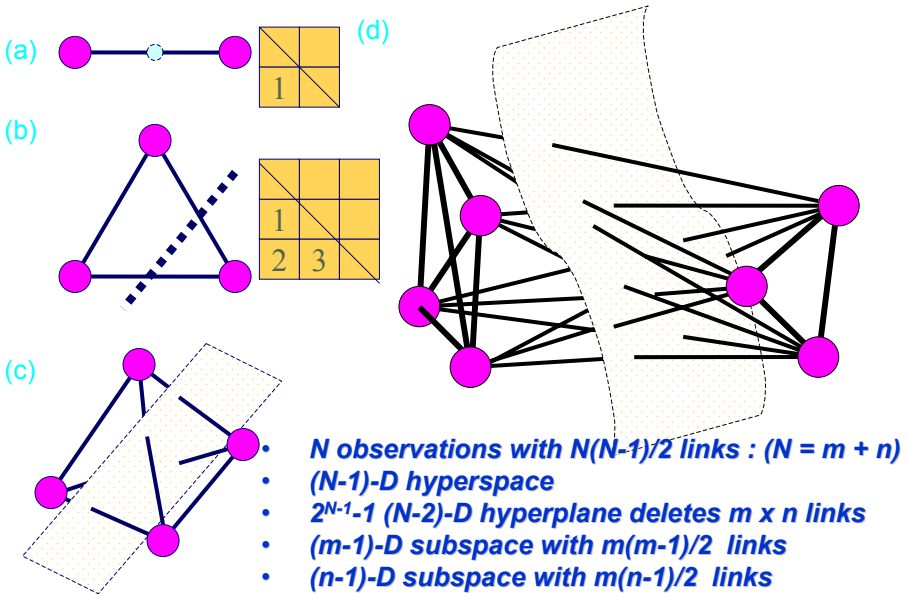


Where would you cut?

into meaningful substructures!

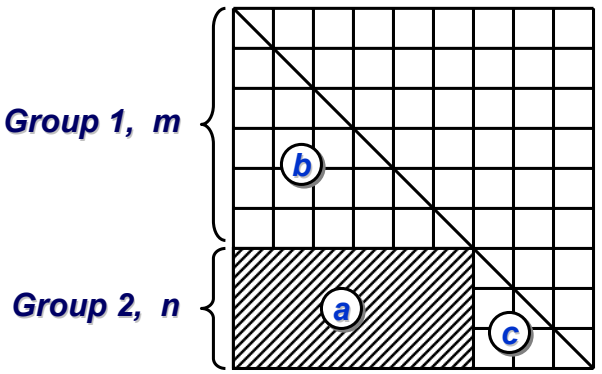


Data space and Incisional hyperplanes



- *N observations with N(N-1)/2 links : (N = m + n)*
- *(N-1)-D hyperspace*
- *2^{N-1}-1 (N-2)-D hyperplane deletes m x n links*
- *(m-1)-D subspace with m(m-1)/2 links*
- *(n-1)-D subspace with m(n-1)/2 links*

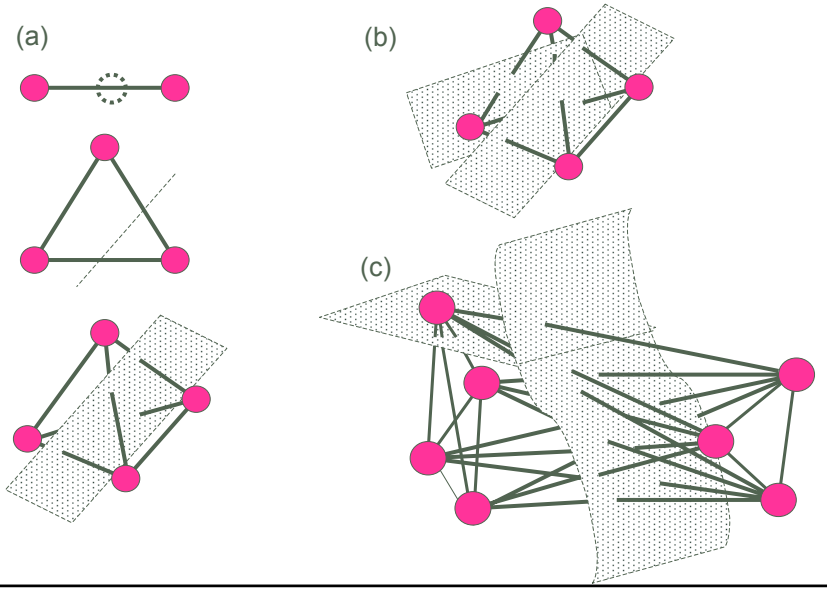
Matrix Representation of Data Space and The Matrix Incision Index (MII)



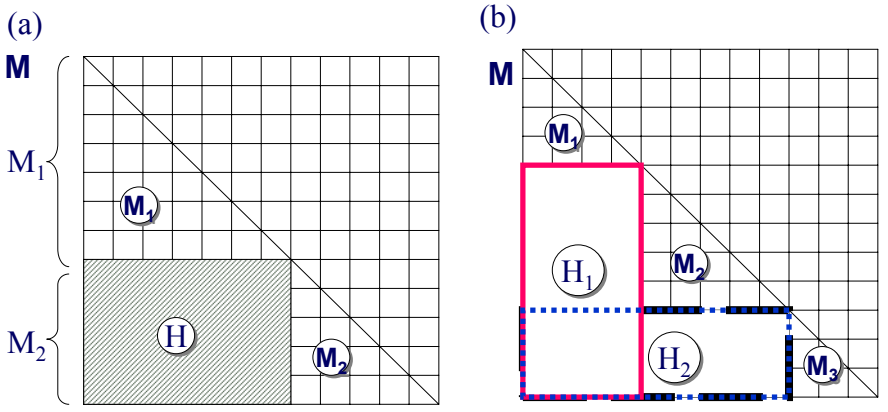
*MII = {(m / (n+m)) * b + (n / (n+m)) * c} / a*

a : loss of average link strength by incision
b : average link strength of group 1
c : average link strength of group 2

Data Hyperspace and Incisional Hyperplanes



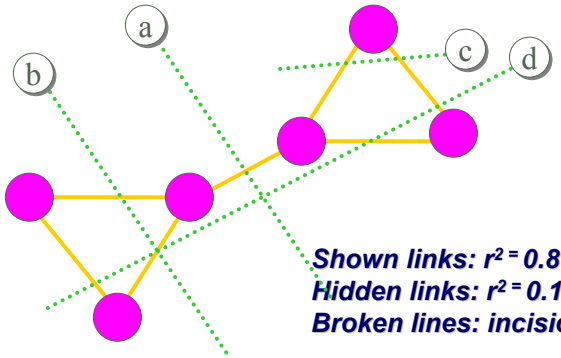
Matrix representation & Matrix incision index



$$MII_{\max} = \frac{1}{S_{cluster}(M)} \sum_i \frac{|E(M_i)|}{|E(M)|} S_{cluster}(M_i)$$

MII captures intuition…

of geometric space decomposition

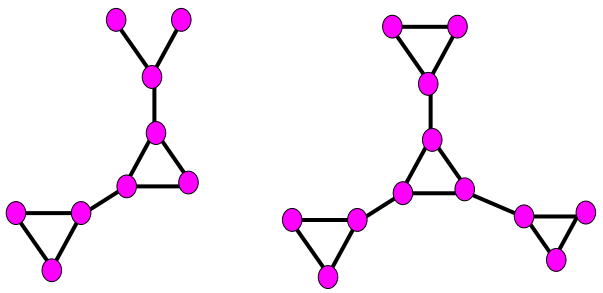


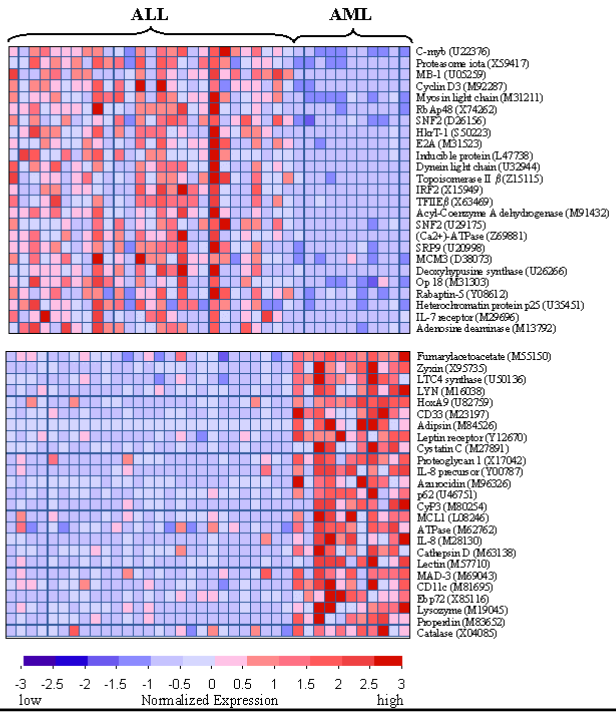
Shown links: $r^2 = 0.8$
Hidden links: $r^2 = 0.1$
Broken lines: incisional hyperplanes

	average loss	weighted avg. link strength	MII
a.	(0.8 + 0.8)/9 = 0.18	0.5(0.8) + 0.5(0.8) = 0.8	4.4
b.	(1.6 + 0.6)/8 = 0.275	0.33(0.8) + 0.67(0.54) = 0.5	1.8
c.	(1.6 + 0.3)/5 = 0.38	0.2(1) + 0.8(0.45) = 0.38	1.0
d.	(3.2 + 0.4)/8 = 0.45	0.33(0.1) + 0.67(0.45) = 0.28	0.64

MII captures intuition…

of geometric decomposition





Result: Leukemia Data Set(Golub et al.)

Training	ALL	AML
Cluster 1	25	0
Cluster 2	2	11

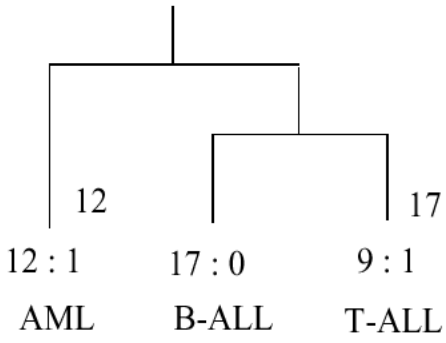
Test	ALL	AML
Cluster 1	19	1
Cluster 2	1	13

Result: Leukemia Data Set(Golub et al.)

AML : 28 - 38

T-ALL : 2, 3, 6, 9-11, 14, 23

B-ALL : 1, 4, 5, 7, 8, 12, 13, 15-22, 24-27



Result: Leukemia Data Set(Golub et al.)

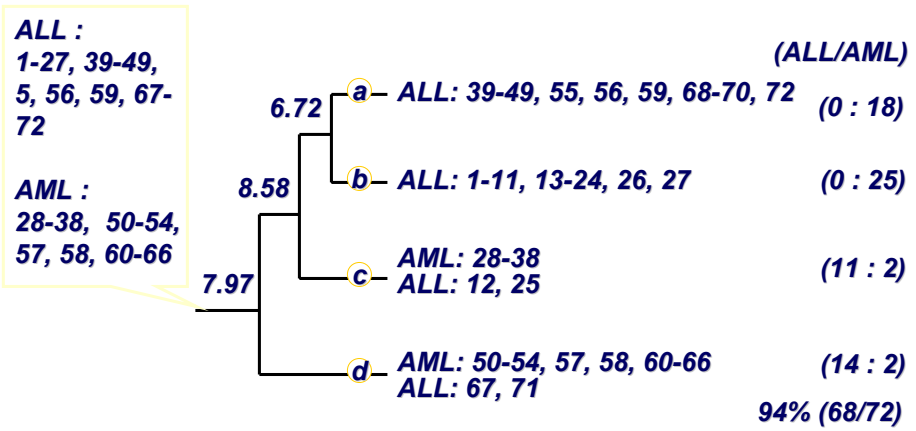
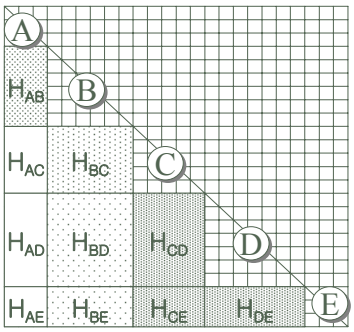


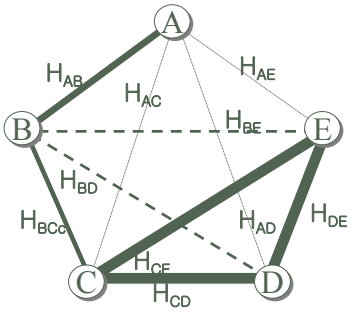
Table 1. Comparison between the actual class information and the cluster memberships created by the MITree-K algorithm in leukemia gene expression data set (Golub et al., 1999).

K			Classes	Members	
2	3	4			
M_1	M_1	M_1	26 ALL	1-11, 13-27	Training set
	M_2	M_2	11 AML 1 ALL	28-38 12	
M_2		M_3	13 AML	50-54, 57, 58, 60-65	Test set
	M_3	M_4	20 ALL 1 AML	39-49, 55, 56, 59, 67-72 66	

K = number of clusters; M_K = Clusters (or sub-matrices); AML = Acute Myeloblastic Leukemia; ALL = Acute Lymphoblastic Leukemia.

Systematic matrix decomposition & reconstruction method





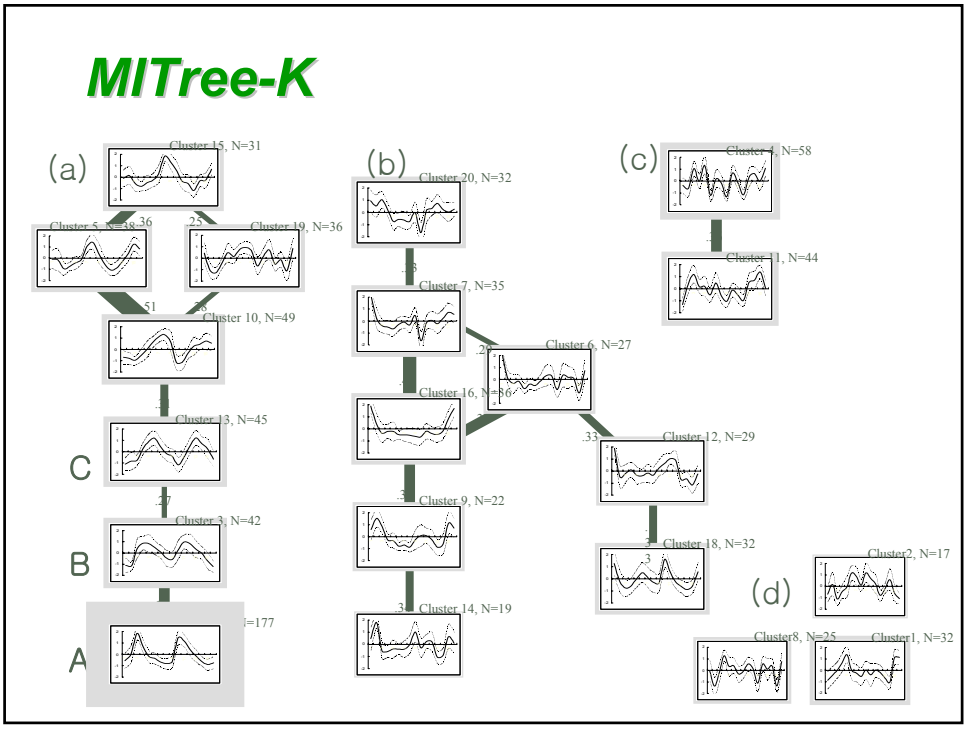
$P(M) = H_1(M) \cup H_2(M) \cup H_3(M) \cup H_4(M)$

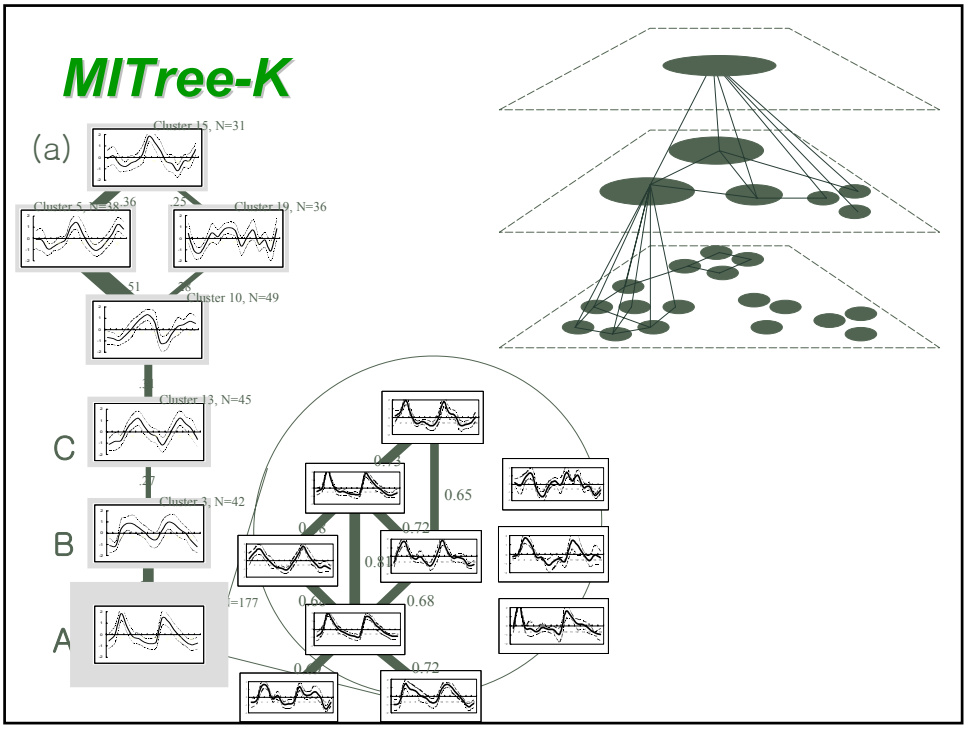
$H_1(M) = H_{AB}(A \cup B) \cup H_{AC}(A \cup C) \cup H_{AD}(A \cup D) \cup H_{AE}(A \cup E)$

$H_2(M) = H_{AC}(A \cup C) \cup H_{BC}(B \cup C) \cup H_{AD}(A \cup D) \cup H_{BD}(B \cup D) \cup H_{AE}(A \cup E) \cup H_{BE}(B \cup E)$

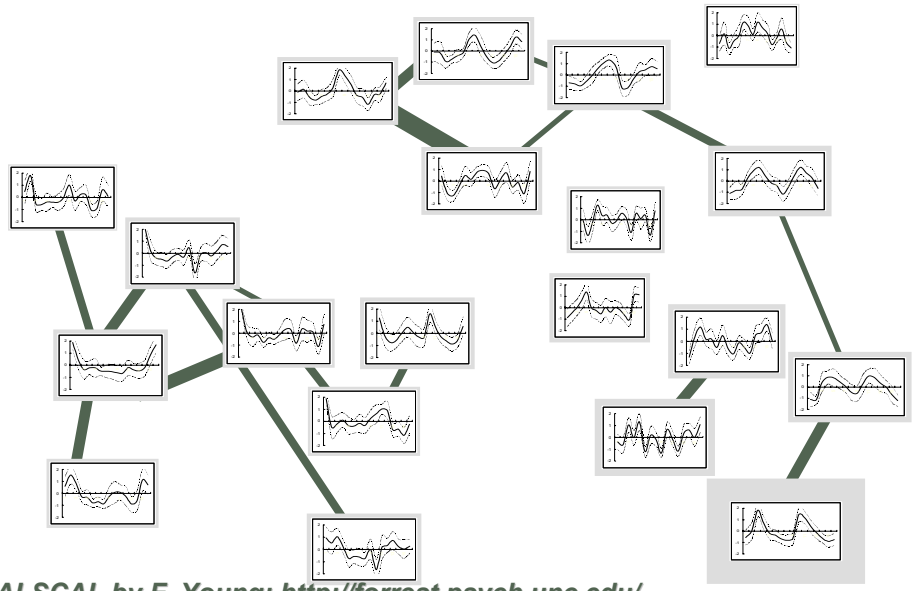
$H_3(M) = H_{AD}(A \cup D) \cup H_{BD}(B \cup D) \cup H_{CD}(C \cup D) \cup H_{AE}(A \cup E) \cup H_{BE}(B \cup E) \cup H_{CE}(C \cup E)$

$H_4(M) = H_{AE}(A \cup E) \cup H_{BE}(B \cup E) \cup H_{CE}(C \cup E) \cup H_{DE}(D \cup E)$


MITree-K
(a)
Cluster 15, N=31
Cluster 5, N=38
Cluster 19, N=36
Cluster 10, N=49
Cluster 13, N=45
Cluster 3, N=42
C
B
A
(b)
Cluster 20, N=32
Cluster 7, N=35
Cluster 6, N=27
Cluster 16, N=36
Cluster 12, N=29
Cluster 9, N=22
Cluster 18, N=32
Cluster 14, N=19
(c)
Cluster 4, N=58
Cluster 11, N=44
(d)
Cluster2, N=17
Cluster8, N=25
Cluster1, N=32


MITree-K
(a)
Cluster 15, N=31
Cluster 5, N=38
Cluster 19, N=36
Cluster 10, N=49
Cluster 13, N=45
Cluster 3, N=42
C
B
A
0.65
0.72
0.68
0.72

MDS of Clustering Structure



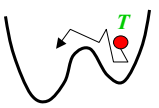
ALSCAL by F. Young: http://forrest.psych.unc.edu/

Algorithmic approaches

Deterministic Annealing

$$P(\text{old} \rightarrow \text{new}) = \begin{cases} \exp[-\Delta f / T] & \text{if } \Delta f > 0 \\ 1 & \text{else} \end{cases}$$

where $\Delta f = f(\text{new}) - f(\text{old})$



Evolutionary Strategy

- *Global optimization strategy by simulation of biological evolution*
- *Evolutionary Strategy : Real-value representation of each object*
- *Genetic Algorithm : Binary-value representation of each object*

Deterministic Annealing

$\langle WL_0 \rangle = \Sigma_{i<j} L_{ij} (1-x_i)(1-x_j) / \Sigma_{i<j} (1-x_i)(1-x_j)$

$\langle WL_1 \rangle = \Sigma_{i<j} L_{ij} x_i x_j / \Sigma_{i<j} x_i x_j$

$\langle BL \rangle = \Sigma_{i<j} L_{ij}[x_i (1-x_j)+x_j (1-x_i)] / \Sigma_{i<j} [x_i (1-x_j)+x_j (1-x_i)]$

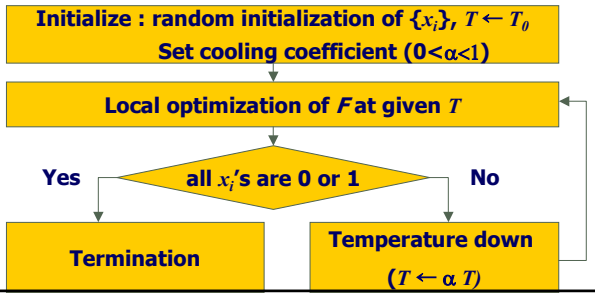
$n_0 = \Sigma_i (1-x_i), \quad n_1 = \Sigma_i (1-x_i)$

L_{ij} : similarity between i and j

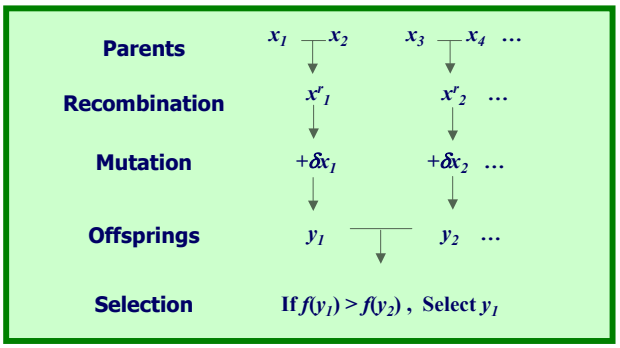
$F = -MII - TS$ (T : Temperature scheduled to decrease)

where $S = -\Sigma_i [x_i \log x_i + (1-x_i) \log (1-x_i)]$

S : Shannon Entropy (Measure of Randomness)

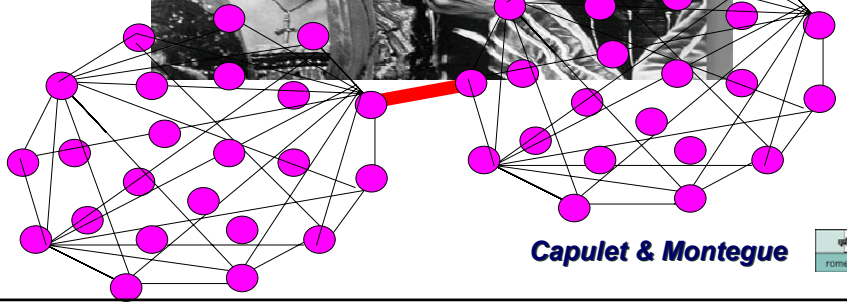


Evolutionary Strategy



Birds eye view? Capturing group effect

Romeo & Juliet effect

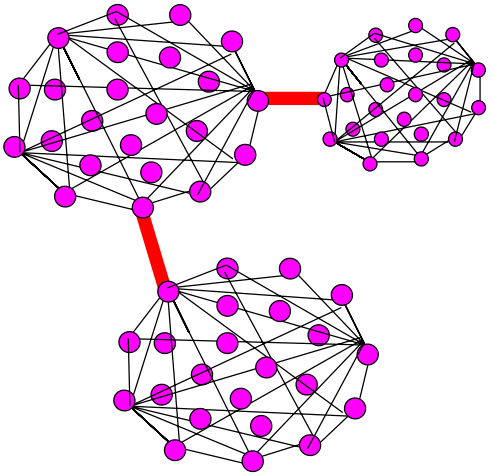


Capulet & Montegue

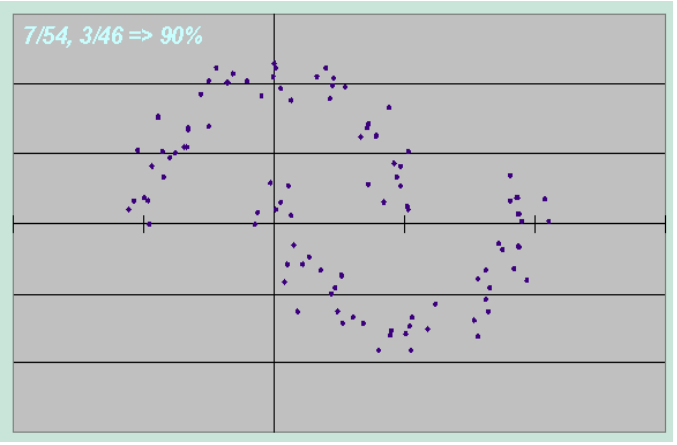
romeo.mid

What are you looking for?

You see what you want to see.



Geometric partitioning

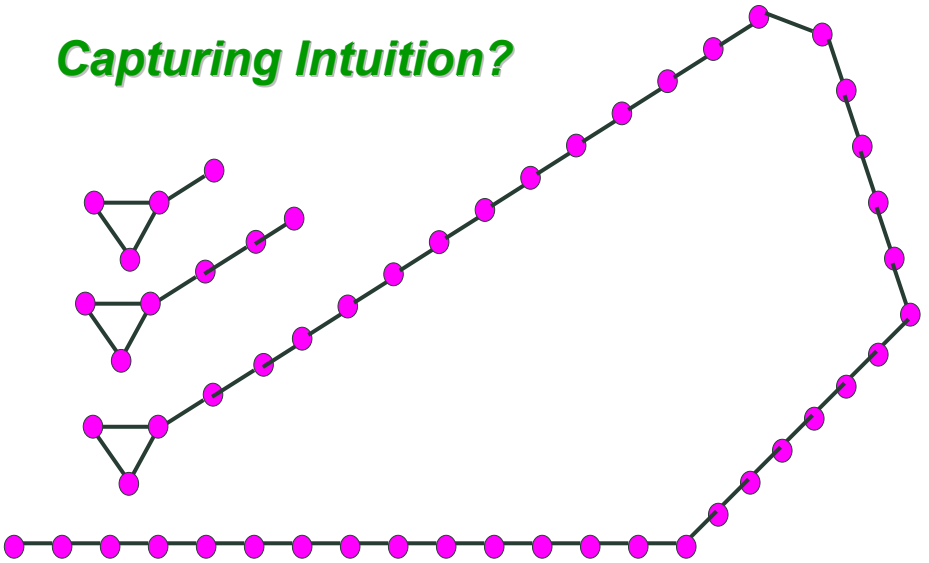


Distance^n may be the correlation coefficient^2?
Density function and physical integrity?

Reliability?

Where would you cut or join?

Capturing Intuition?



Discussions on MITree, MITree-K

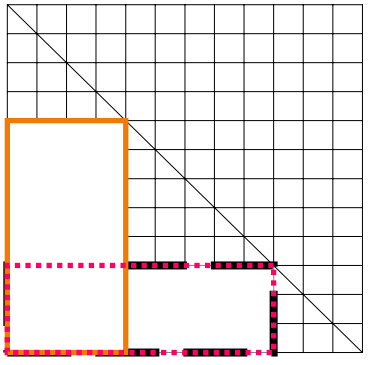
- clustering as a search and optimization
- no assumption on data distribution.
  - It depends only on the unique observation under consideration
- works both in hierarchical & partitional manner
- quantitative visualization of clustering structure
- capturing clustering structure
- global optimization
- consistency & quality issues
- supervised vs. unsupervised machine learning
- independent on measurement, goal, or algorithm
- three-tired model of clustering
  - measurement
  - principle
  - algorithm
- capturing (geometric) intuition

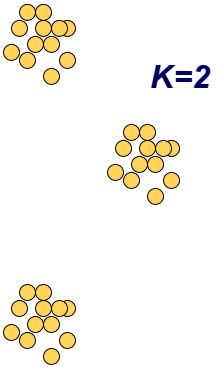
Other Issues in Clustering

- cluster consistency
- cluster robustness
- cluster qualities
- cluster comparisons

Clustering Consistency & Quality

Consistency





Clustering Consistency & Quality

Clustering consistency vs cluster robustness

Clustering Consistency & Quality

Consistency: Rand Index

$L_{concordant}$ = {edges equally connected or disconnected in both clustering solutions}

$L_{discordant}$ = {edges connected in one solution and disconnected in the other}

$$C(P_i, P_j) = \frac{|L_{concordant}|}{|L_{concordant}| + |L_{discordant}|}$$

$$C_N = \frac{2}{N(N-1)} \sum_{i<j} C(P_i, P_j)$$

Adjusted Rand index

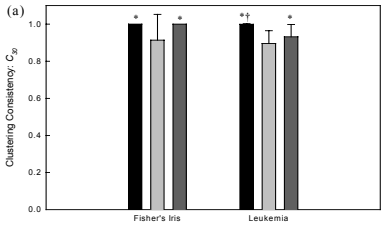
Clustering Consistency & Quality

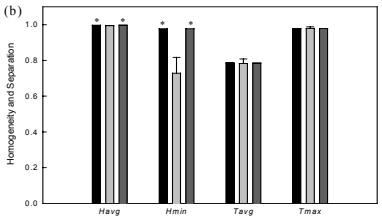
Table 2. Comparison of clustering consistency and quality measures of three clustering algorithms applied to Fisher's Iris and Golub's leukemia data sets (Golub et al., 1999) after 30 trials of creating three and four clusters, respectively.

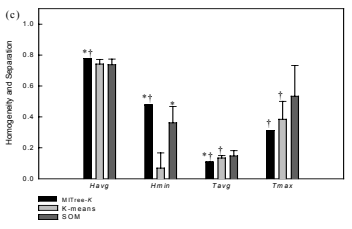
Index	MITree-K	K-means	SOM
Fisher's Iris Data Set			
C_{30}‡	1.000±0.00*	0.914±0.14	1.000±0.00*
H_{avg}‡	0.997±0.00*	0.994±0.00	0.997±0.00*
H_{min}‡	0.977±0.00*	0.728±0.10	0.977±0.00*
T_{avg}	0.786±0.00	0.783±0.03	0.785±0.00
T_{max}	0.978±0.00	0.979±0.01	0.979±0.00
Golub's Leukemia Data Set			
C_{30}‡	0.999±0.00*†	0.896±0.69	0.932±0.66*
H_{avg}‡	0.778±0.00*†	0.742±0.03	0.738±0.04
H_{min}‡	0.478±0.00*†	0.069±0.10	0.361±0.11*
T_{avg}‡	0.108±0.00*†	0.135±0.02†	0.147±0.03
T_{max}‡	0.312±0.00†	0.384±0.12†	0.534±0.20

Clustering Consistency & Quality

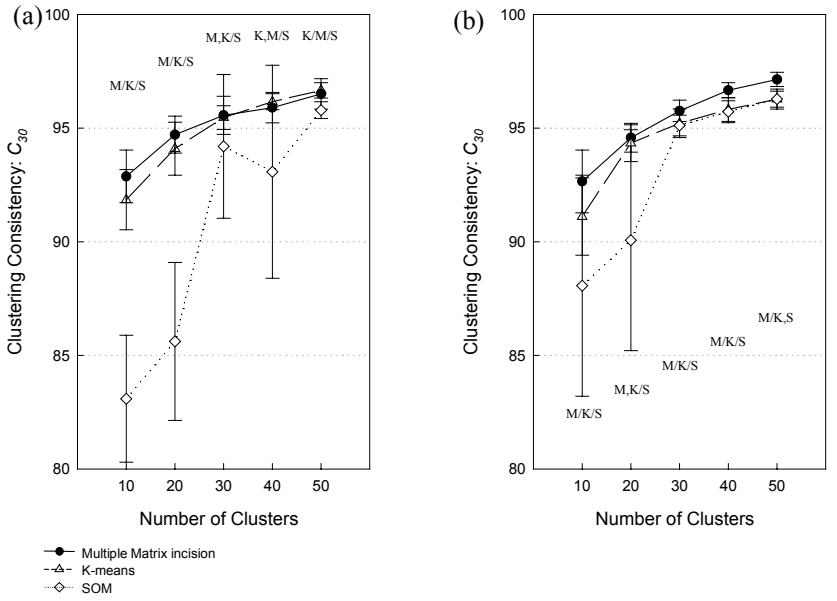
Test data sets







Clustering Consistency: Genomic Data sets



Clustering Consistency & Quality

Homogeneity

$$H_{avg} = \frac{1}{|O(M)|} \sum_{x \in O(M)} Correl(F(x), F(M)))$$

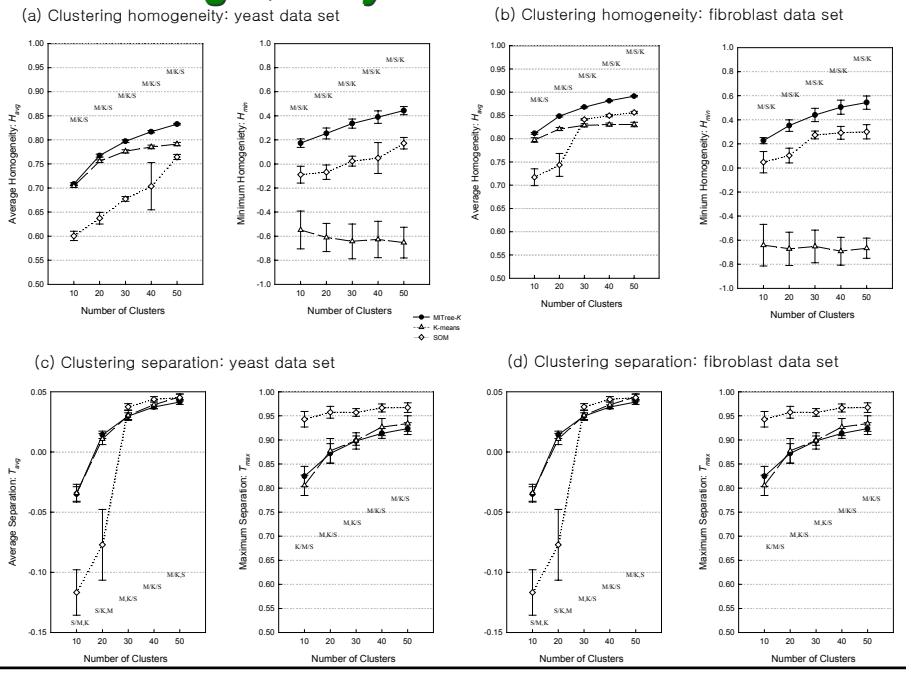
$$H_{\min} = \min_{x \in O(M)} Correl(F(x), F(M))$$

Separation

$$T_{avg} = \frac{1}{\sum_{i<j} |O(M_i)||O(M_j)|} \sum_{i<j} |O(M_i)||O(M_j)| Correl(F(M_i), F(M_j))$$

$$T_{\max} = \max_{i<j} Correl(F(M_i), F(M_j))$$

Clustering Quality: Genomic Data Sets



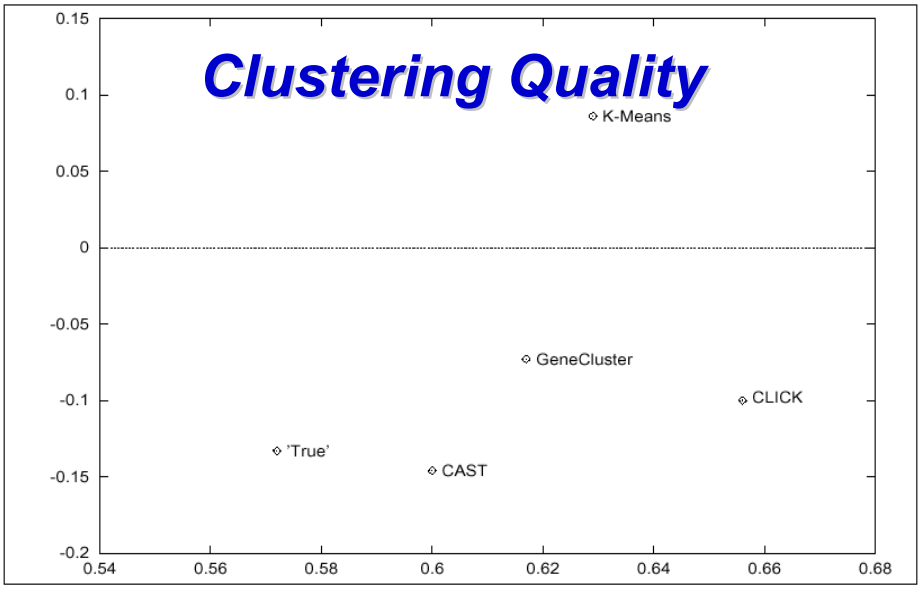


Figure 16: A comparison of homogeneity (x-axis) and separation (y-axis) values for all solutions. Recall that a solution improves if homogeneity increases or separation decreases.

Keyword Clustering

- *@ http://bio.ifom-firc.it/KW_CLUST/index.html*
- *clusters together sequence identifiers sharing common keywords.*
- *represented by a vector of keywords*
- *a hierarchical clustering algorithm is applied*
- *The tree can also be obtained in PHYLIP format for parsing with a tree viewer program such as ATV (http://www.genetics.wustl.edu/eddy/atv/).*

Trajectory Clustering

```
for i = number of time points down to 3
        row = 1; col = i;
        for 1:length of diag of Matrix(i-1)
            if H_row,col == 1
                if C_row,col-1,row+1,col >> 0
                    H_row,col-1 = 1
                    if H_row+1,col != 1;
                        H_row+1,col = 0;
                    end
                elseif C_row,col-1,row+1,col << 0
                    if H_row,col-1 != 1
                        H_row,col-1 = 0
                    end
                    H_row+1,col = 1;
                elseif C_row,col-1,row+1,col ~= 0
                    H_row,col-1 = 1 and H_row+1,col = 1;
                end
            end
            row = row + 1;
            col = col + 1;
        end
    end
```

Figure 1 Pseudocode for iterative clustering of more than 3 time points. diag(i) is the ith diagonal of the matrix counting from the main diagonal; it is equivalent to Matlab's diag function.

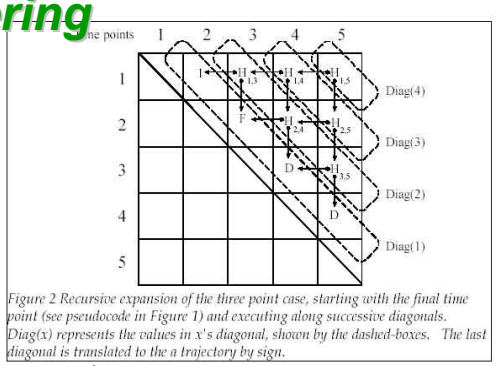


Figure 2 Recursive expansion of the three point case, starting with the final time point (see pseudocode in Figure 1) and executing along successive diagonals. Diag(x) represents the values in x's diagonal, shown by the dashed-boxes. The last diagonal is translated to the a trajectory by sign.

row	col
1	5
1	4
2	5
1	3
2	4
3	5

mammary gland example data traversal

Trajectory Clustering

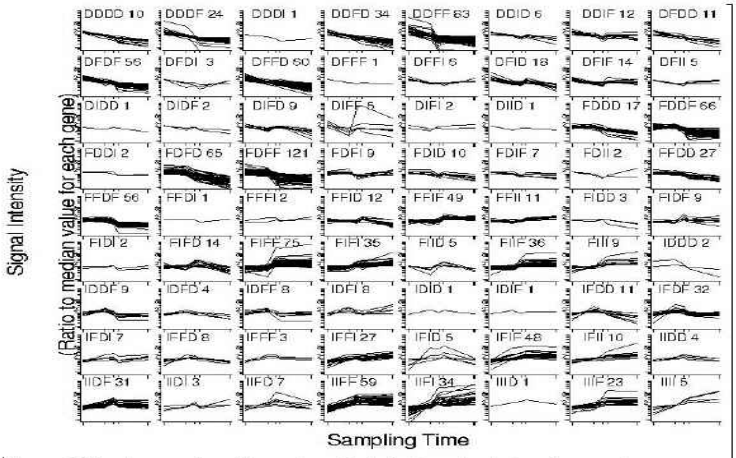


Figure 3 Clusters produced by automatic trajectory clustering for secretory activation in the mammary gland. Four replicates at each of five time points, Pregnancy days 12 and 17 and lactation days 1, 2 and 9, are represenedt in each plot. Intensities were normalized to the median for each gene and plotted on a log scale.

Analysing SNPs

$$s = H(C) - H(C|G)$$

$$p\text{-value}(l, s) = Prob(S \geq s)$$

Labels:	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−
Locus 1:	AA	AA	AA	AA	AA	Aa	Aa	Aa	Aa	aa	aa	aa	aa	aa	aa
Locus 2:	BB	BB	BB	Bb	Bb	Bb	bb	bb	bb	BB	BB	Bb	Bb	bb	bb

Figure 1. Illustration of the mutual information score for two classes: one with 9 individuals labels by '+', the other one with 6 individuals labelled by '-'. At locus 1, all people in the first class have genotypes AA or Aa, and all people in the second class have genotypes aa. This locus is informative and has the score of 0.97 (p-value=0.0002). At locus 2, there is no difference between genotype frequencies in different classes, the mutual information score for this locus is 0 (p-value=1).

Analysing SNPs

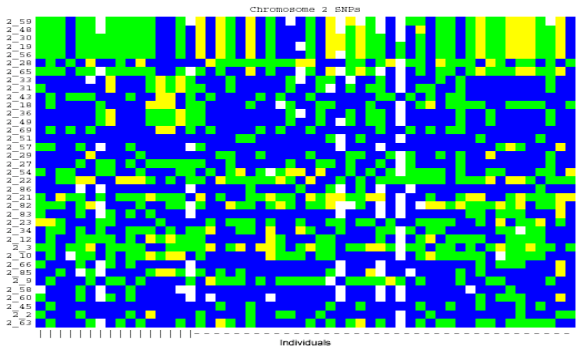


Figure 3. Graphical representation of the highest scoring SNPs from chromosome 2 and 'linked'/'not-linked' partition. Each column represents all genotypes for a given person; each row represents all genotypes for a given SNP. Blue corresponds to homozygous genotype for common allele, yellow corresponds to homozygous genotype for rare allele and green corresponds to heterozygous genotype. White corresponds to missing data. Loci are ordered with respect to mutual information score. Columns marked by '|' on the x-axis correspond to patients from 'not linked' group, columns marked by '-' on the x-axis correspond to patients from 'linked' group.

Analysing SNPs

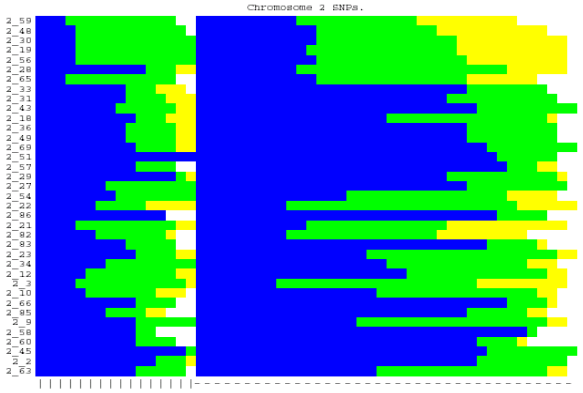
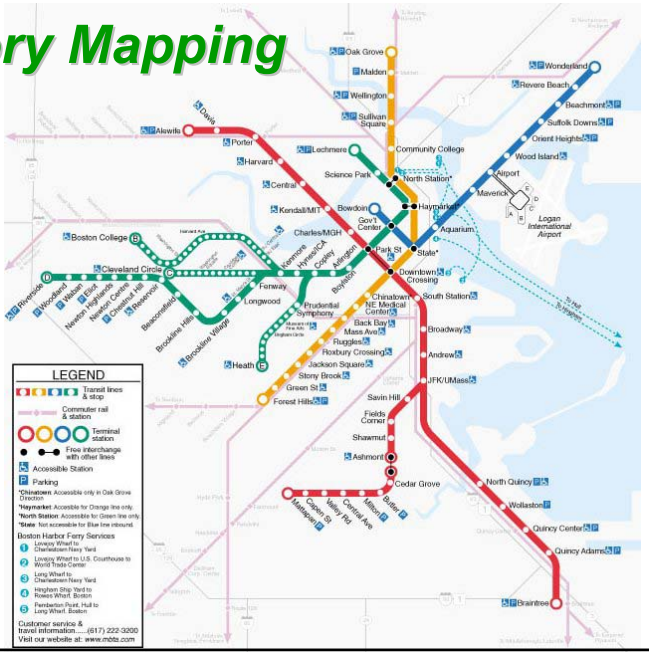
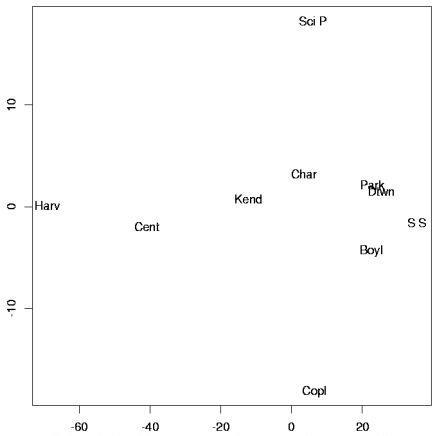


Figure 4. Same data as in Figure 3, now data in each row is sorted by genotypes within each group. This plot helps to visually assess mutual information score, since it is clear that the top 5 SNPs got high scores because individuals in 'not-linked' group do not have homozygous genotypes for the rare alleles at these SNPs. Note that columns no longer correspond to a particular individual.

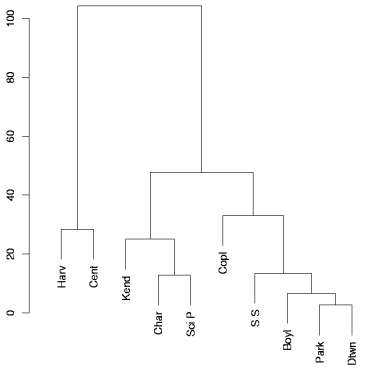
Trajectory Mapping



Trajectory Mapping



Multidimensional scaling



Hierarchical clustering

Gilbert 1997

Trajectory Mapping

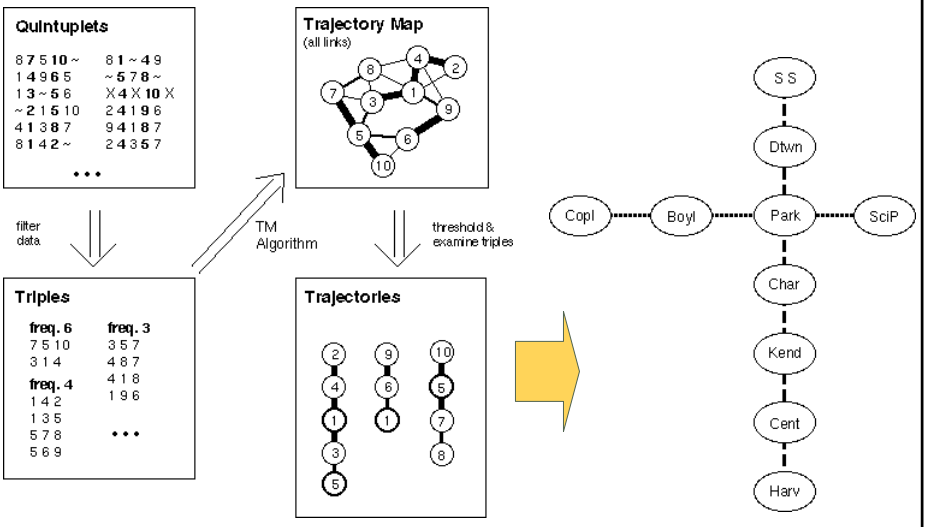
Quintuplets of Boston subway stations

ex	A	int	B	ex
—	S S	|	Copl	—
—	Cent	|	Boyl	—
Harv	Cent	*	Kend	Park
Kend	Char	Park	Boyl	Copl
—	SciP	Park	Copl	—
Harv	Cent	Kend	Char	Dtwn
Harv	Kend	Char	Park	S S
Harv	Kend	Park	S S	—
S S	Dtwn	Park	SciP	—
Cent	Kend	*	Char	Park
—	Harv	Char	Dtwn	S S
—	Harv	*	Cent	Kend
Kend	Char	Park	SciP	—
—	Harv	Kend	Char	S S
Kend	Char	*	Park	Dtwn
Harv	Cent	Char	Dtwn	S S
—	S S	|	SciP	—
—	Copl	*	Boyl	Park
—	Harv	|	Copl	—
SciP	Park	*	Boyl	Copl
SciP	Park	Boyl	Copl	—
—	Harv	Kend	Park	S S
—	Kend	|	Copl	—

ex	A	int	B	ex
—	Harv	Park	S S	—
—	Cent	|	SciP	—
—	Char	|	Copl	—
—	Kend	|	Boyl	—
Harv	Kend	Park	Dtwn	S S
Harv	Cent	Kend	Park	S S
S S	Dtwn	Park	Boyl	Copl
Kend	Char	Park	Dtwn	S S
—	Kend	|	SciP	—
—	Harv	|	Boyl	—
—	S S	|	Boyl	—
—	SciP	Park	Boyl	Copl
Char	Park	*	Dtwn	S S
—	Harv	|	SciP	—
Harv	Cent	Char	S S	—
Park	Dtwn	*	S S	—
Kend	Char	Park	S S	—
—	Cent	|	Copl	—
—	Dtwn	|	Copl	—
Boyl	Park	*	SciP	—
—	Harv	Cent	Kend	Park
Kend	Park	Dtwn	S S	—

— dead end * feasible, no sample | infeasible

Trajectory Mapping

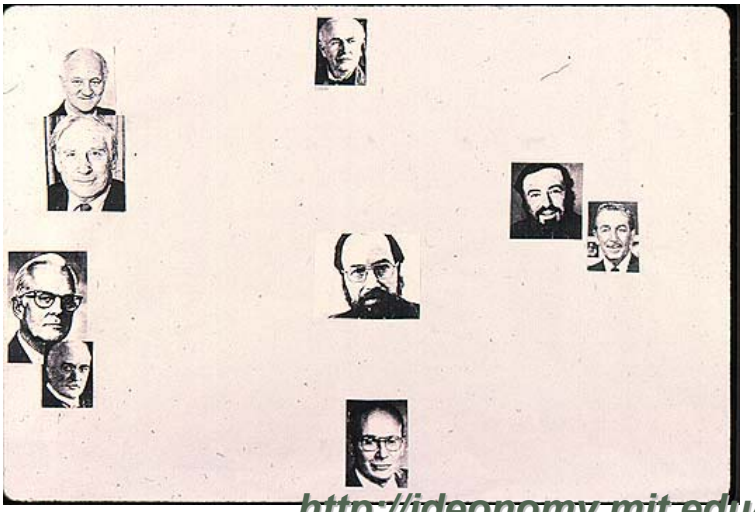


Algorithm using quintuplets and SA.

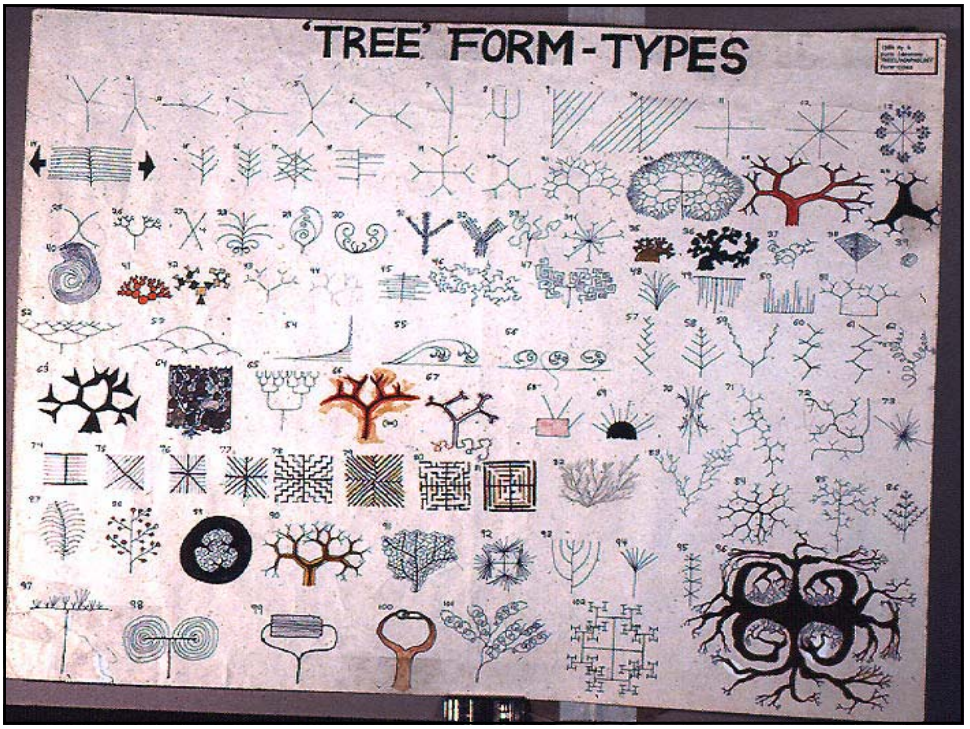
Gilbert 1997

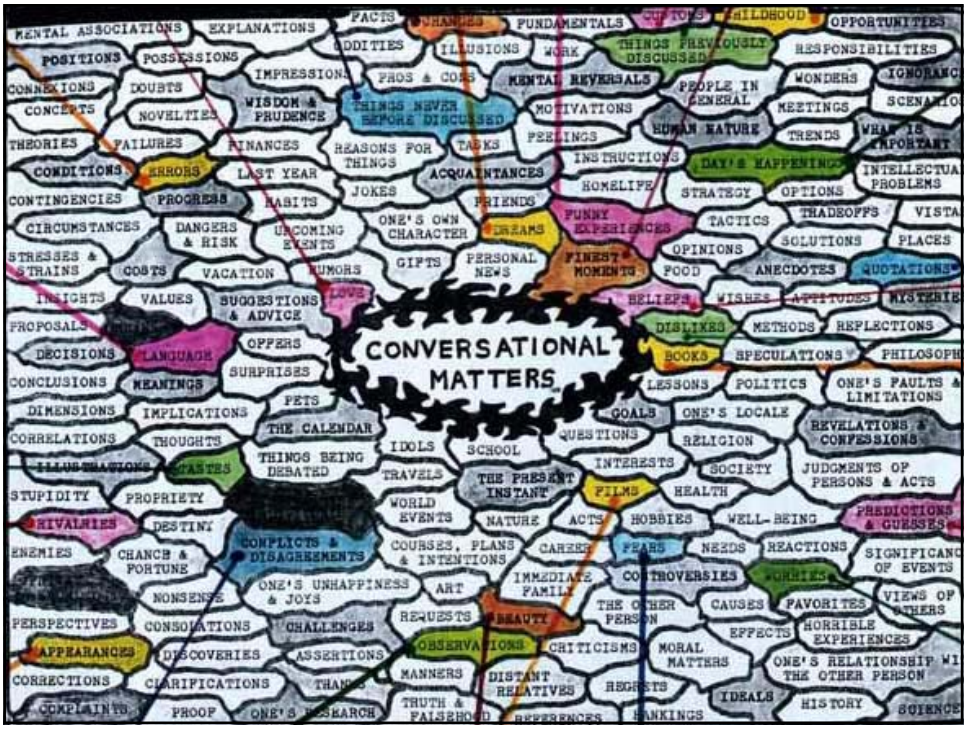
Ideonomy

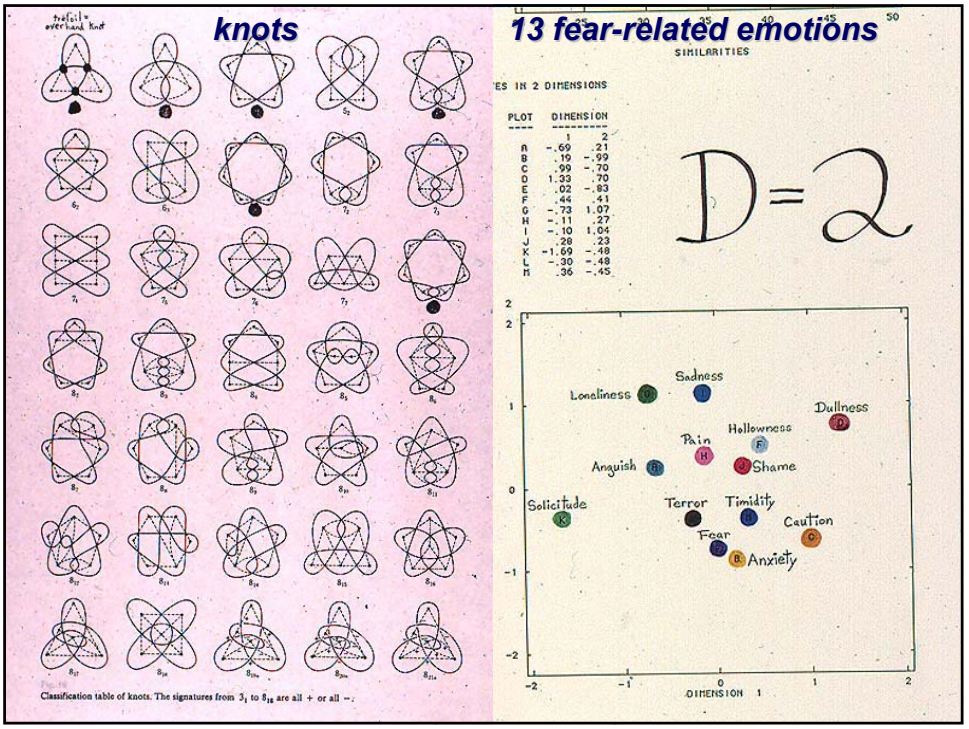
Science of Ideas -- *Patrick Gunkel*

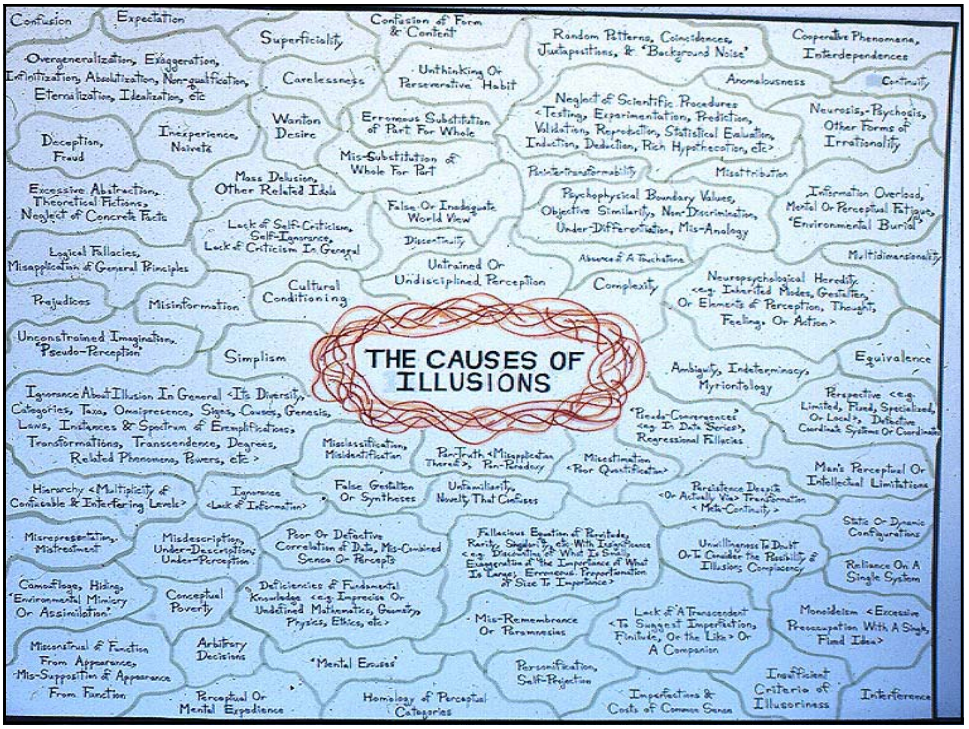


http://ideonomy.mit.edu/




CONVERSATIONAL MATTERS
MENTAL ASSOCIATIONS
EXPLANATIONS
FACTS
FUNDAMENTALS
CHILDHOOD
OPPORTUNITIES
POSITIONS
POSSESSIONS
ODDITIES
ILLUSIONS
WORK
THINGS PREVIOUSLY DISCUSSED
RESPONSIBILITIES
CONNEXIONS
DOUBTS
IMPRESSIONS
PROS & CONS
MENTAL REVERSALS
PEOPLE IN GENERAL
WONDERS
CONCEPTS
NOVELTIES
WISDOM & PRUDENCE
THINGS NEVER BEFORE DISCUSSED
MOTIVATIONS
MEETINGS
THEORIES
FAILURES
FINANCES
REASONS FOR THINGS
TASKS
FEELINGS
HUMAN NATURE
TRENDS
CONDITIONS
ERRORS
LAST YEAR
ACQUAINTANCES
INSTRUCTIONS
DAY'S HAPPENINGS
CONTINGENCIES
PROGRESS
HABITS
JOKES
FRIENDS
HOMELIFE
STRATEGY
OPTIONS
CIRCUMSTANCES
DANGERS & RISK
UPCOMING EVENTS
ONE'S OWN CHARACTER
DREAMS
FUNNY EXPERIENCES
TACTICS
THADEOFFS
VISTA
STRESSES & STRAINS
COSTS
VACATION
RUMORS
GIFTS
PERSONAL NEWS
FINEST MOMENTS
OPINIONS
SOLUTIONS
PLACES
INSIGHTS
VALUES
SUGGESTIONS & ADVICE
LOVE
FOOD
ANECDOTES
QUOTATIONS
PROPOSALS
BELIEFS
WISHES
ATTITUDES
MYSTERIES
DECISIONS
LANGUAGE
OFFERS
DISLIKES
METHODS
REFLECTIONS
CONCLUSIONS
MEANINGS
SURPRISES
BOOKS
SPECULATIONS
PHILOSOPHY
DIMENSIONS
IMPLICATIONS
PETS
LESSONS
POLITICS
ONE'S FAULTS & LIMITATIONS
CORRELATIONS
THOUGHTS
THE CALENDAR
GOALS
ONE'S LOCALE
REVELATIONS & CONFESSIONS
ILLUSTRATIONS
TASTES
THINGS BEING DEBATED
IDOLS
SCHOOL
QUESTIONS
RELIGION
STUPIDITY
PROPRIETY
TRAVELS
THE PRESENT INSTANT
INTERESTS
SOCIETY
JUDGMENTS OF PERSONS & ACTS
RIVALRIES
DESTINY
WORLD EVENTS
FILMS
HEALTH
PREDICTIONS & GUESSES
ENEMIES
CHANCE & FORTUNE
CONFLICTS & DISAGREEMENTS
NATURE
ACTS
HOBBIES
WELL-BEING
COURSES, PLANS & INTENTIONS
CAREER
FEARS
NEEDS
REACTIONS
SIGNIFICANCE OF EVENTS
NONSENSE
ONE'S UNHAPPINESS & JOYS
ART
IMMEDIATE FAMILY
CONTROVERSIES
WORRIES
CAUSES
FAVORITES
VIEWS OF OTHERS
PERSPECTIVES
CONSOLATIONS
CHALLENGES
REQUESTS
BEAUTY
THE OTHER PERSON
EFFECTS
HORRIBLE EXPERIENCES
APPEARANCES
DISCOVERIES
ASSERTIONS
OBSERVATIONS
CRITICISMS
MORAL MATTERS
ONE'S RELATIONSHIP WITH THE OTHER PERSON
CORRECTIONS
CLARIFICATIONS
THANKS
MANNERS
DISTANT RELATIVES
REGRETS
IDEALS
HISTORY
COMPLAINTS
PROOF
ONE'S RESEARCH
TRUTH & FALSEHOOD
REFERENCES
RANKINGS
SCIENCE


knots
13 fear-related emotions
Classification table of knots. The signatures from 3_1 to 8_{16} are all + or all −.
SIMILARITIES
ES IN 2 DIMENSIONS
PLOT DIMENSION
1 2
A -.69 .21
B .19 -.99
C .99 -.70
D 1.33 .70
E .02 -.83
F .44 .41
G -.73 1.07
H -.11 .27
I -.10 1.04
J .28 .23
K -1.69 -.48
L -.30 -.48
M .36 -.45
D=2
Sadness
Loneliness
Dullness
Hellowness
Pain
Anguish
Shame
Solicitude
Terror
Timidity
Caution
Fear
Anxiety
DIMENSION 1


THE CAUSES OF ILLUSIONS
Confusion
Expectation
Superficiality
Confusion of Form & Content
Random Patterns, Coincidences, Juxtapositions, & 'Background Noise'
Cooperative Phenomena, Interdependences
Overgeneralization, Exaggeration, Intensification, Absolutization, Non-qualification, Eternalization, Idealization, etc
Carelessness
Unthinking Or Perseverative Habit
Anomalousness
Continuity
Wanton Desire
Erroneous Substitution of Part For Whole
Neglect of Scientific Procedures <Testing, Experimentation, Prediction, Validation, Reproducibility, Statistical Evaluation, Induction, Deduction, Rich Hypothecation, etc>
Neurosis, Psychosis, Other Forms of Irrationality
Deception, Fraud
Inexperience, Naivete
Mis-Substitution of Whole For Part
Mass Delusion, Other Related Idols
Excessive Abstraction, Theoretical Fictions, Neglect of Concrete Facts
Lack of Self-Criticism, Self-Ignorance, Lack of Criticism In General
False Or Inadequate World View
Psychophysical Boundary Values, Objective Similarity, Non-Discrimination, Under-Differentiation, Mis-Analogy
Information Overload, Mental Or Perceptual Fatigue, 'Environmental Burden'
Logical Fallacies, Misapplication of General Principles
Discontinuity
Untrained Or Undisciplined Perception
Complexity
Multidimensionality
Prejudices
Misinformation
Cultural Conditioning
Neuropsychological Heredity <e.g. Inherited Modes, Gestalten, Or Elements of Perception, Thought, Feeling, Or Action>
Unconstrained Imagination, 'Pseudo-Perception'
Simplism
Ambiguity, Indeterminacy, Myriontology
Equivalence
Ignorance About Illusion In General <Its Diversity, Categories, Taxa, Omnipresence, Signs, Causes, Genesis, Laws, Instances & Spectrum of Exemplifications, Transformations, Transcendence, Degrees, Related Phenomena, Powers, etc>
'Pseudo-Consequences' <e.g. In Data Series>, Regressional Fallacies
Perspective <e.g. Limited, Fixed, Specialized, Or Local>, Defective Coordinate Systems Or Coordinates
Misclassification, Misidentification
Pan-Truth <Misapplication Thereof>, Pan-Paradoxy
Misestimation <Poor Quantification>
Man's Perceptual Or Intellectual Limitations
Hierarchy <Multiplicity of Confusable & Interfering Levels>
Ignorance <Lack of Information>
False Gestalten Or Syntheses
Unfamiliarity, Novelty That Confuses
Persistence Despite <Or Actual, Via> Transformation < Meta-Continuity >
State Or Dynamic Configurations
Misrepresentation, Mistreatment
Misdescription, Under-Description, Under-Perception
Poor Or Defective Correlation of Data, Mis-Combined Sense Or Percepts
Fallacious Equation of Positive, Rarity, Singularity, etc With Importance <e.g. Proximity of What Is Small, Exaggeration of the Importance of What Is Large, Erroneous Proportionation of Size To Importance>
Unwillingness To Doubt Or To Consider the Possibility & Illusory Complacency
Reliance On A Single System
Camouflage, Hiding, Environmental Mimicry Or Assimilation
Conceptual Poverty
Deficiencies & Fundamental Knowledge <e.g. Imprecise Or Undefined Mathematics, Geometry, Physics, Ethics, etc>
Mis-Remembrance Or Paramnesias
Lack of A Transcendent <To Suggest Imperfection, Finitude, Or the Like> Or A Companion
Monoideism <Excessive Preoccupation With A Single, Fixed Idea>
Misconstrual of Function From Appearance, Mis-Supposition of Appearance From Function
Arbitrary Decisions
'Mental Excuses'
Personification, Self-Projection
Insufficient Criteria of Illusoriness
Interference
Perceptual Or Mental Expedience
Homology of Perceptual Categories
Imperfections & Costs of Common Sense


ROBOT RULES
Thank you!