Computational Intelligence in Clustering: A Survey and Applications

Donald Wunsch

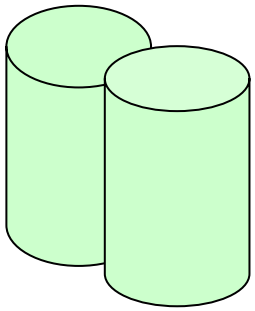
Applied Computational Intelligence Laboratory

University of Missouri - Rolla

Acknowledgements

- Funding
  - **NSF**
  - DARPA / BBN Inc.
  - Sandia
  - Boeing
  - **MK Finley Missouri Professorship**
- Senior Personnel
  - Danil Prokhorov *
  - Hu Xiao *
  - Alexander Novokhodko *
  - **Sam Mulder** *
  - Frank Harary
- Personnel
  - Xindi Cai C
  - Rohit Dua C
  - **Rui Xu** C

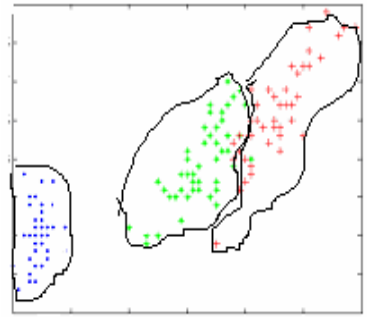
Cluster Analysis



Data Samples

Feature Selection or Extraction

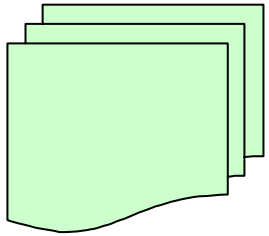
Clustering Algorithm Design or Selection



Clusters

Clusters Validation

Results Interpretation



Knowledge

Clustering Algorithms

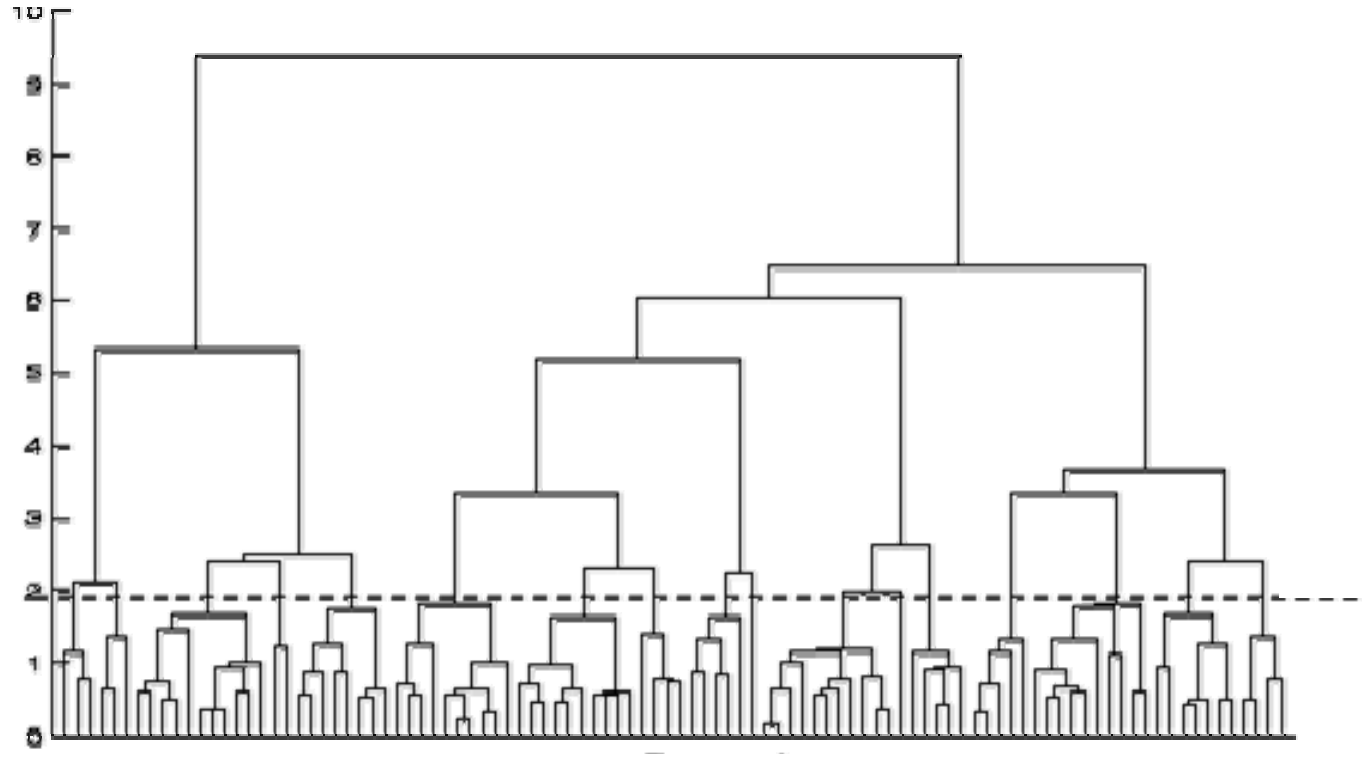
- Hierarchical
  - Agglomerative: Single linkage, complete linkage, group average linkage, median linkage, centroid linkage, ward's method, BIRCH, CURE, ROCK…
  - Divisive: DIANA, MONA…
- Squared Error-Based (Vector Quantization)
  - *K*-means, ISODATA, GKA, PAM…
- pdf Estimation via Mixture Densities
  - MCLUST, GMDD, AutoClass…
- Graph Theory-Based
  - Chameleon, DTG, HCS, CLICK, CAST…
- Combinatorial Search Techniques-Based
  - GGA, TS clustering, SA clustering…
- Fuzzy
  - FCM, MM, PCM, FCS…
- Neural Networks-Based
  - LVQ, SOFM, ART, SART, HEC, SPLL…
- Kernel-Based
  - Kernel *K*-means, SVC…
- Sequential Data
  - Sequence Similarity
  - Indirect sequence clustering
  - Statistical sequence clustering
- Large-Scale Data Sets
  - CLARA, CURE, CLARANS, BIRCH, DBSCAN, DENCLUE, WaveCluster, FC, ART…
- Data visualization and High-dimensional Data
  - PCA, ICA, Projection pursuit, Isomap, LLE, CLIQUE, OptiGrid, ORCLUS…

Proximity

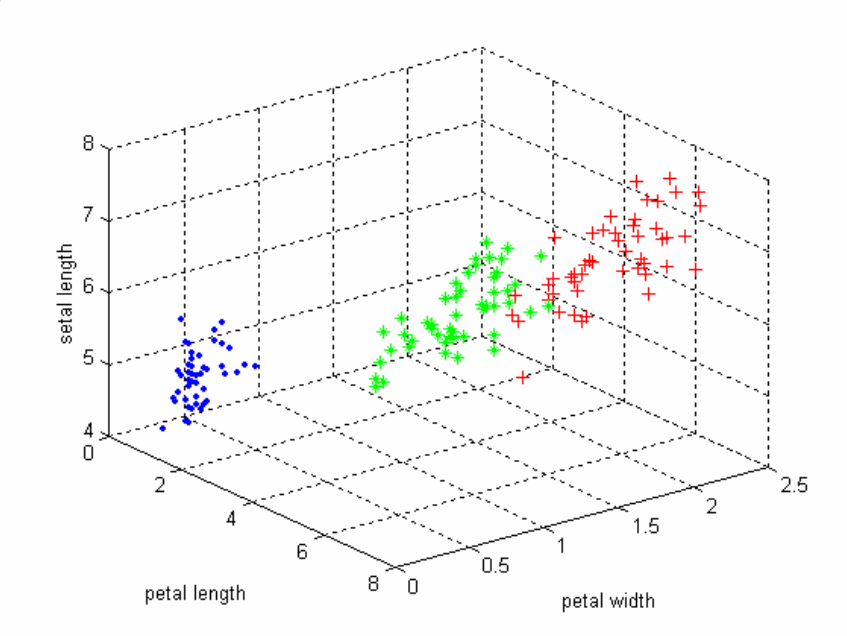
TABLE 1
SIMILARITY AND DISSIMILARITY MEASURE FOR QUANTITATIVE FEATURES

Measures	Forms	Comments	Examples and Applications
Minkowski distance	$D_{ij} = \left(\sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{n} \right)^{1/n}$	Metric. Invariant to any translation and rotation only for $n = 2$ (Euclidean distance). Features with large values and variances tend to dominate over other features.	Fuzzy c-means with measures based on Minkowski family [130].
Euclidean distance	$D_{ij} = \left(\sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{2} \right)^{1/2}$	The most commonly used metric. Special case of Minkowski metric at $n = 2$. Tend to form hyperspherical clusters.	K-means algorithm [191]
City-block distance	$D_{ij} = \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert$	Special case of Minkowski metric at $n = 1$. Tend to form hyperrecutangular clusters.	Fuzzy ART [57]
Sup distance	$D_{ij} = \max_{1 \le l \le d} \lvert x_{il} - x_{jl} \rvert$	Special case of Minkowski metric at $n \to \infty$.	Fuzzy c-means with sup norm [39].
Mahalanobis distance	$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{S}$ is the within-group covariance matrix.	Invariant to any nonsingular linear transformation. $\mathbf{S}$ is calculated based on all objects. Tend to form hyperellipsoidal clusters. When features are not correlated, squared Mahalanobis distance is equivalent to squared Euclidean distance. May cause some computational burden.	Ellipsoidal ART [13], Hyperellipsoidal clustering algorithm [194].
Pearson correlation	$D_{ij} = (1 - r_{ij})/2$, where $r_{ij} = \dfrac{\sum_{l=1}^{d} (x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{d} (x_{il} - \bar{x}_i)^2 \sum_{l=1}^{d} (x_{jl} - \bar{x}_j)^2}}$	Not a metric. Derived from correlation coefficient. Unable to detect the magnitude of differences of two variables.	Widely used as the measure for analyzing gene expression data [80].
Point symmetry distance	$D_{ir} = \min_{\substack{j=1,\ldots,N \\ and\ j \neq i}} \dfrac{\lVert (\mathbf{x}_i - \mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r) \rVert}{\lVert (\mathbf{x}_i - \mathbf{x}_r) \rVert + \lVert (\mathbf{x}_j - \mathbf{x}_r) \rVert}$	Not a metric. Compute the distance between an object $\mathbf{x}_i$ and a reference point $\mathbf{x}_r$. D_{ir} is minimized when a symmetric pattern exists.	SBKM (Symmetry-based K-means) [264].
Cosine similarity	$S_{ij} = \cos\alpha = \dfrac{\mathbf{x}_i^T \mathbf{x}_j}{\lVert \mathbf{x}_i \rVert \lVert \mathbf{x}_j \rVert}$	Independent of vector length. Invariant to rotation, but not to linear transformations.	The most commonly used measure in document clustering [261].

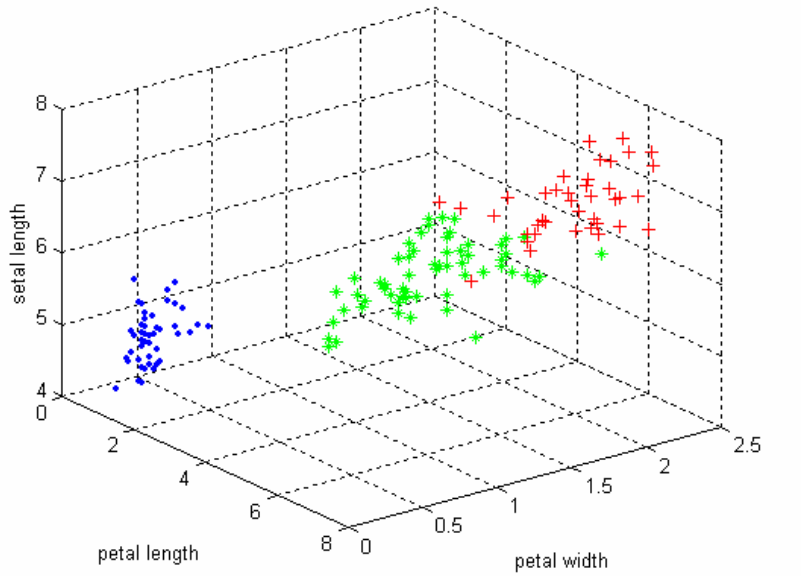
Hierarchical Clustering



Partitional Clustering

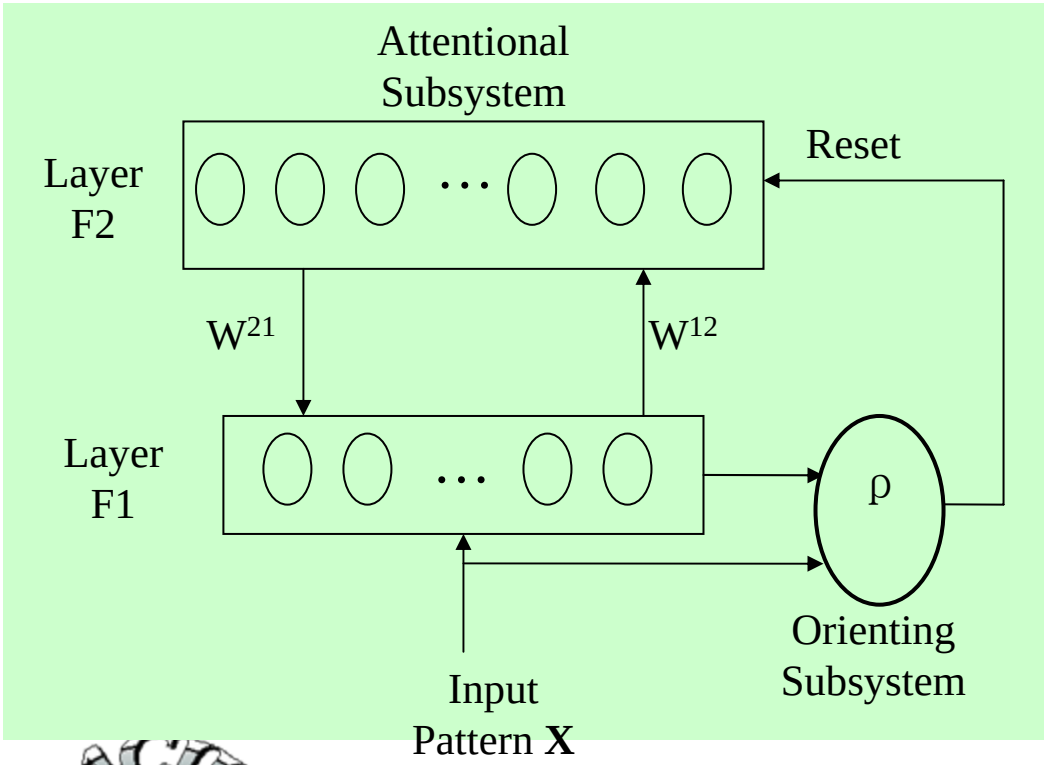


IRIS data set



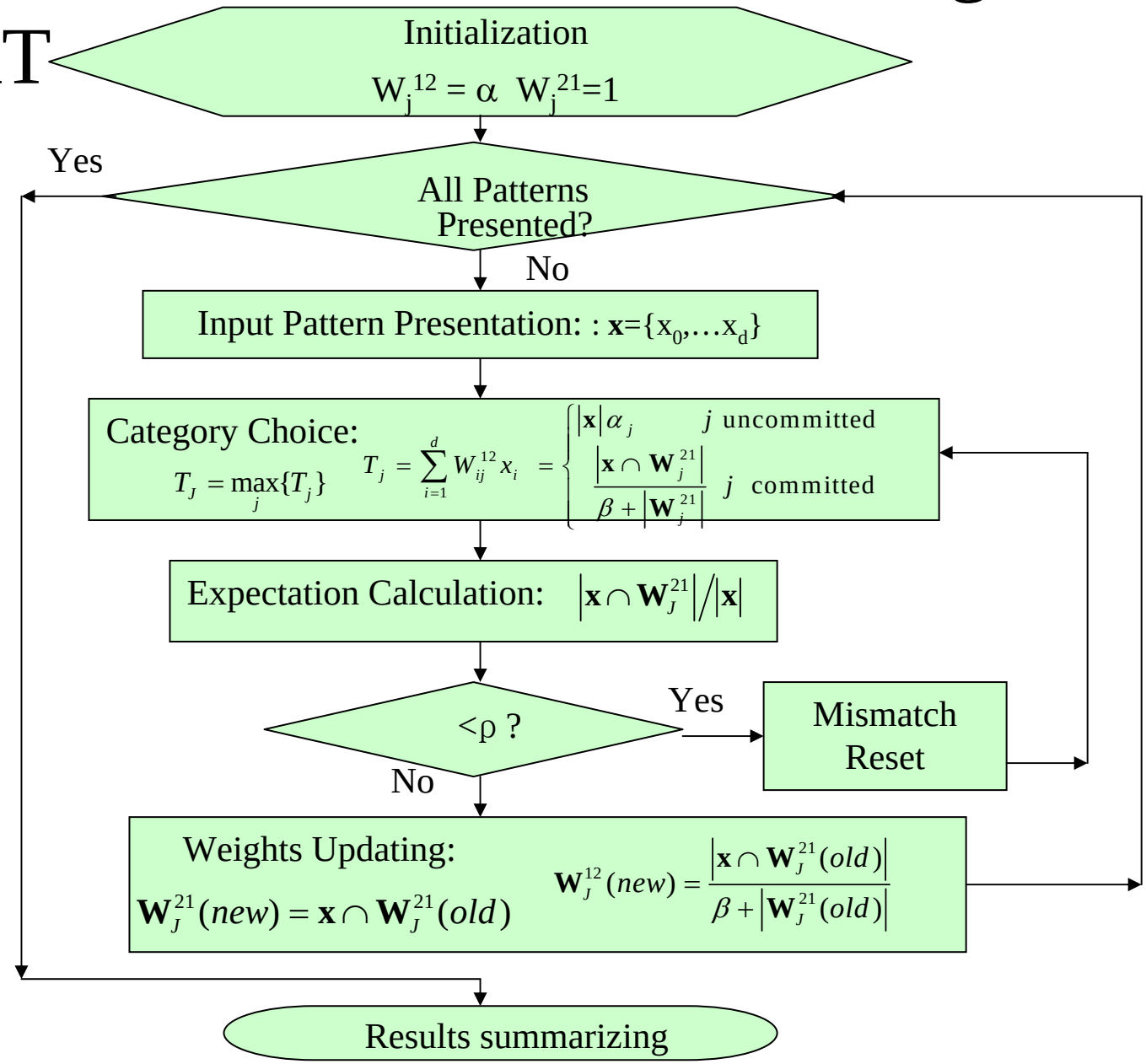
K-Means clustering result

Neural Network Based Clustering - ART



- ART family
  - ART1, binary inputs
  - ART2, analog inputs
  - ARTMAP
  - Fuzzy ART, hyperrectangular clusters
  - Ellipsoid ART, hyperellipsoid clusters

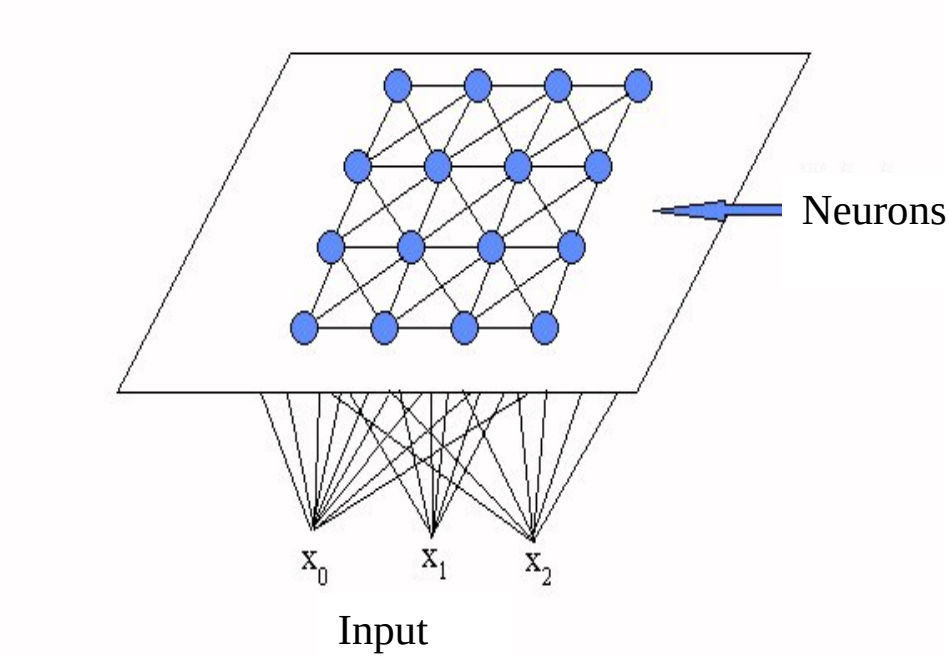
Neural Network Based Clustering - ART



Kernel Based Clustering

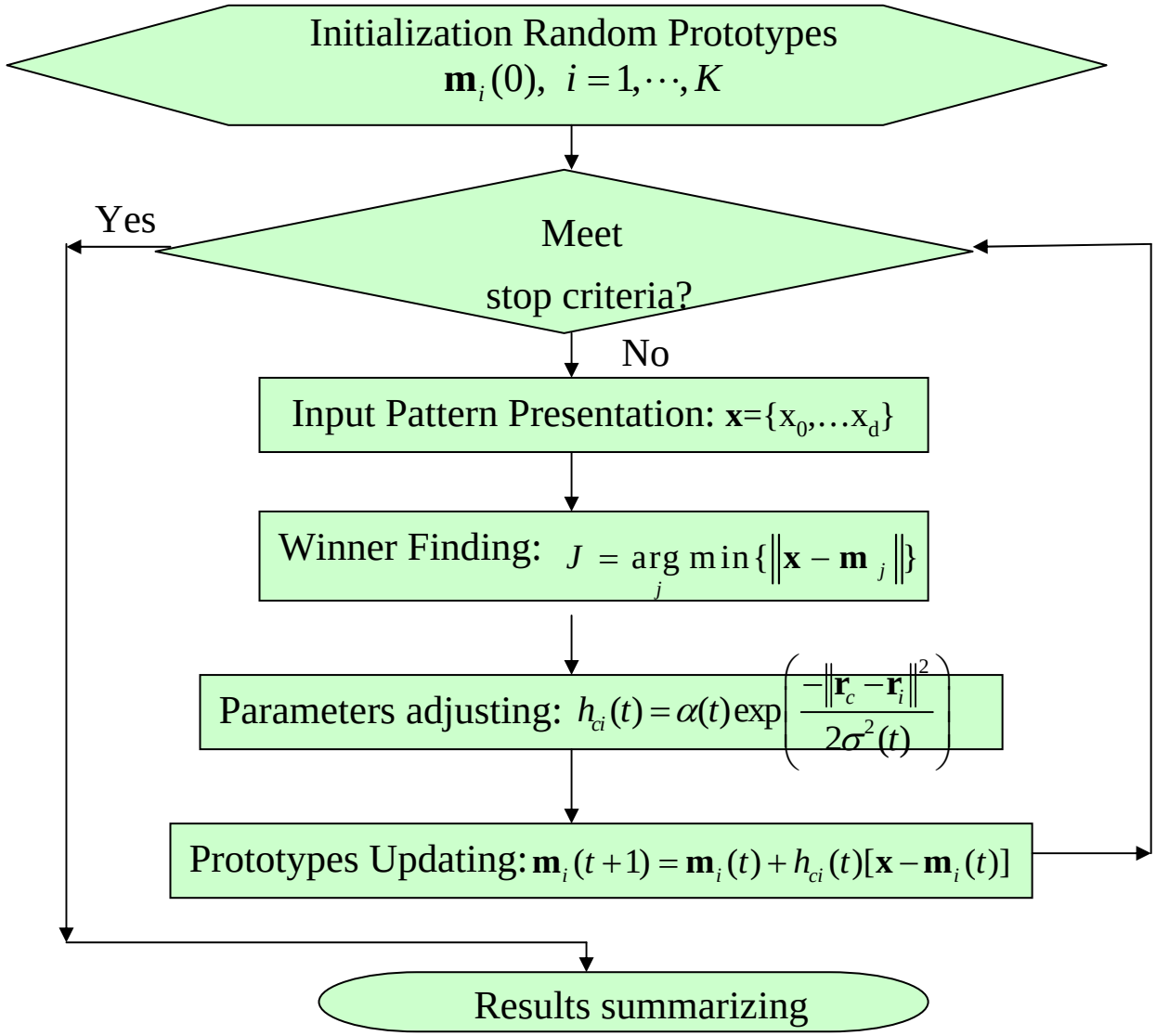
- More possible to obtain a linearly separable hyperplane in a high-dimensional feature space - Cover's theorem
- Kernel inner product - Mercer's theorem
- Reformulate more powerful nonlinear versions for existing linear algorithms
  - Kernel PCA
  - Kernel K-means
- Generate arbitrary clustering shapes other than hyperellipsoid and hypersphere
- Capable of dealing with noise and outliers

Neural Network Based Clustering - SOFM



- Prototype vectors represent high-dimensional patterns
- Good visualization (usually 2-D)
- Inputs are connected to every neuron
- Winner-take-all learning

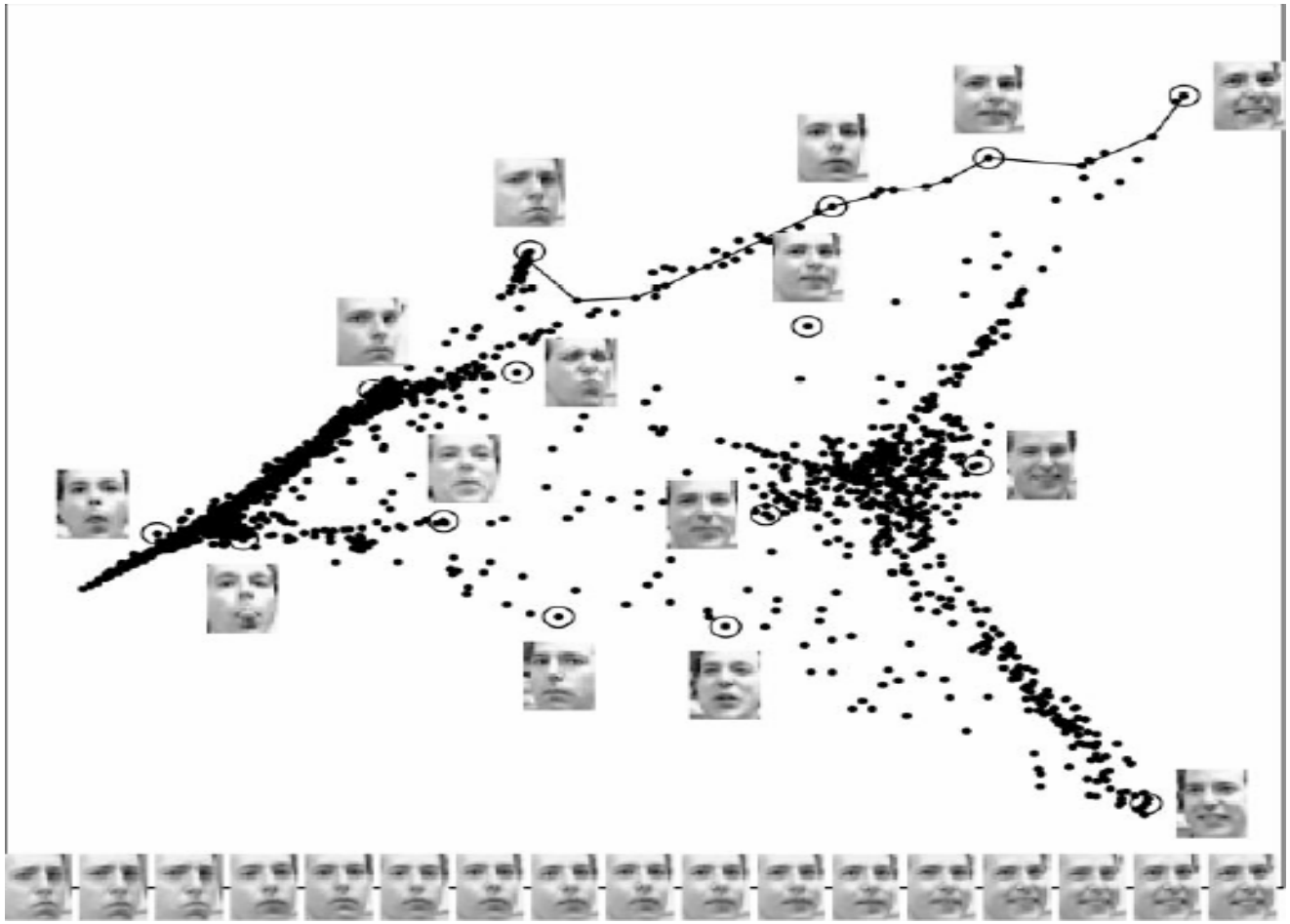
Neural Network Based Clustering - SOFM



Clustering Large-Scale Data

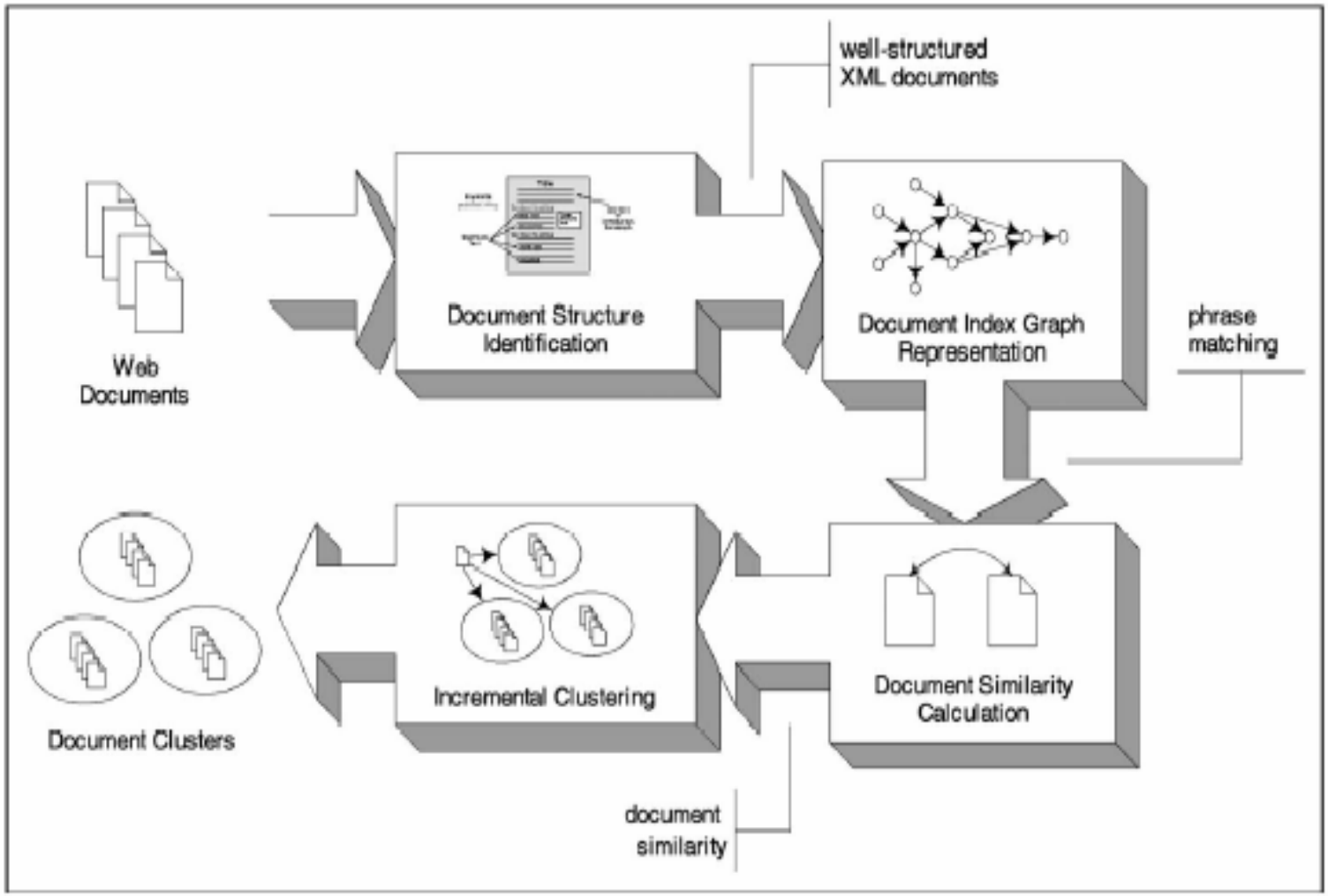
Cluster algorithm	Complexity	Capability of tackling high dimensional data
K-means	$O(NKd)$ (time) $O(N+K)$ (space)	No
Fuzzy c-means	Near $O(N)$	No
Hierarchical clustering*	$O(N^2)$ (time) $O(N^2)$ (space)	No
CLARA	$O(K(40+K)^2+K(N-K))^*$ (time)	No
CLARANS	Quadratic in total performance	No
BIRCH	$O(N)$ (time)	No
DBSCAN	$O(N \log N)$ (time)	No
CURE	$O(N_{sample}^2 \log N_{sample})$ (time) $O(N_{sample})$ (space)	Yes
WaveCluster	$O(N)$ (time)	No
DENCLUE	$O(N \log N)$ (time)	Yes
FC	$O(N)$ (time)	Yes
CLIQUE	Linear with the number of objects, Quadratic with the number of dimensions	Yes
OptiGrid	Between $O(Nd)$ and $O(Nd \log N)$	Yes
ORCLUS	$O(K_0^3+K_0Nd+K_0^2d^3)$ (time) $O(K_0d^2)$ (space)	Yes

Application – Human Face Pose Recognition



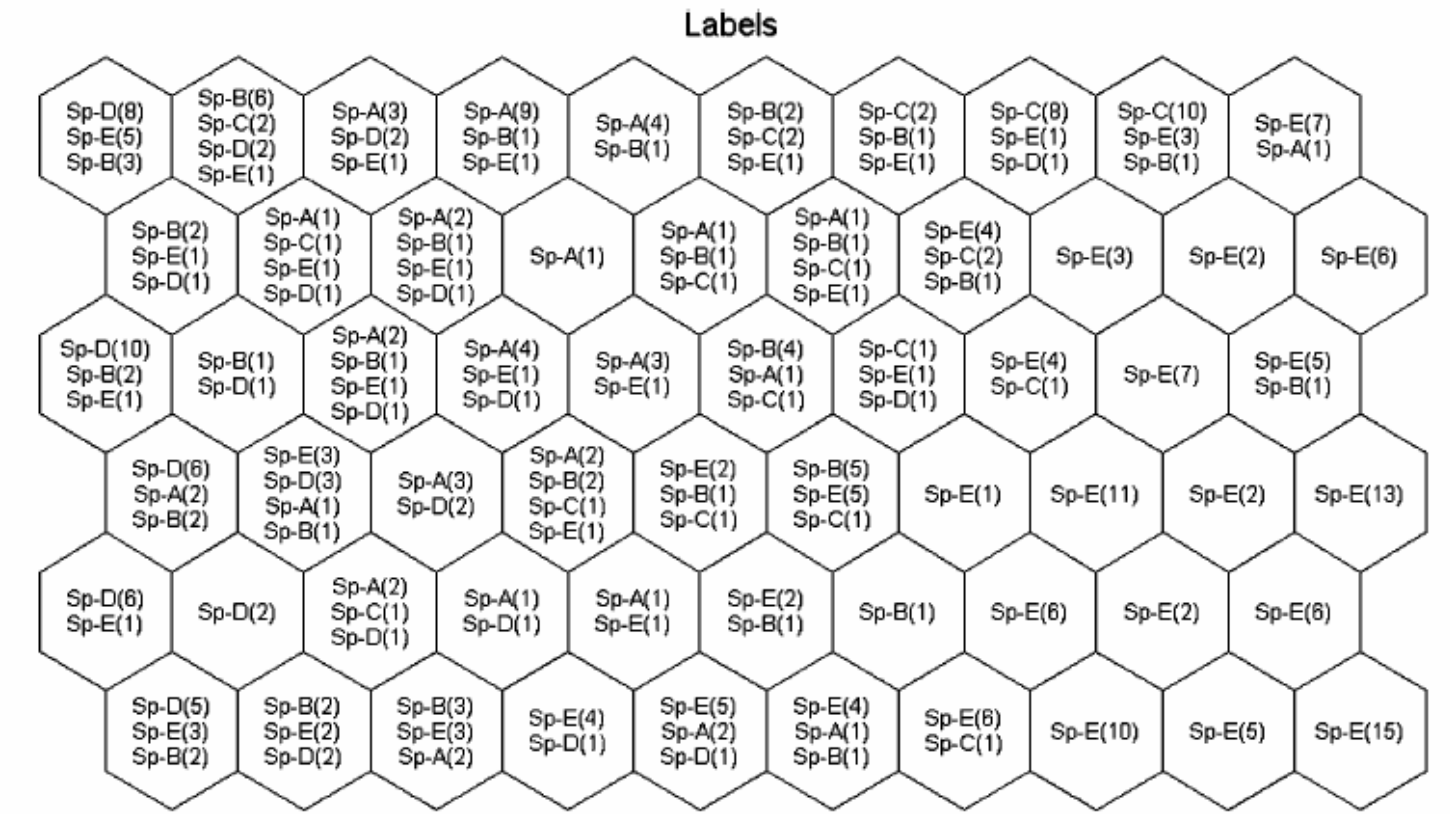
S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science*, vol. 290, no. 5500, pp. 2323-2326, 2000

Application – Web Document Clustering System



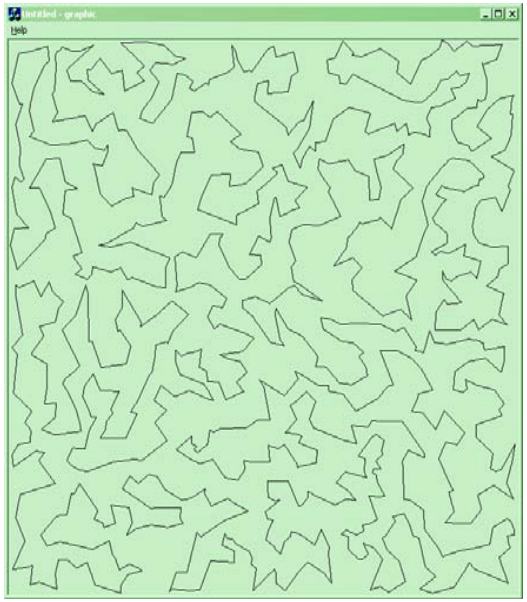
Hammouda, et. al., Efficient phrase-based document indexing for web document clustering,, IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 10, pp. 1279 - 1296 ,2004.

Application – Text Clustering by SOFM

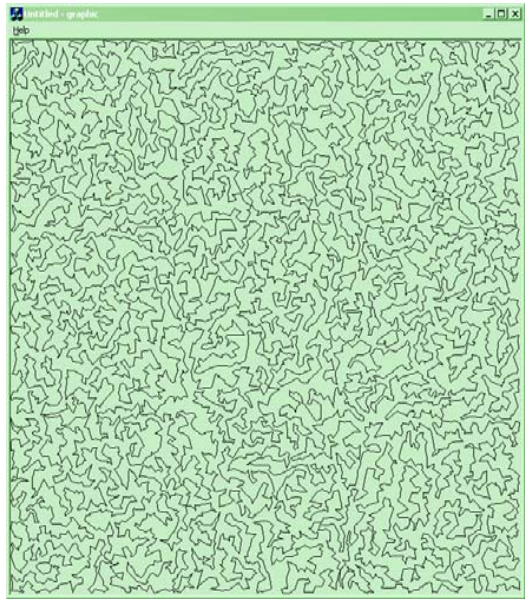


G. Tambouratzis, "Assessing the Effectiveness of Feature Groups in Author Recognition Tasks With the SOM Model," IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews : Accepted for future publication vol. PP, no. 99, pp.1 – 11, 2005 .

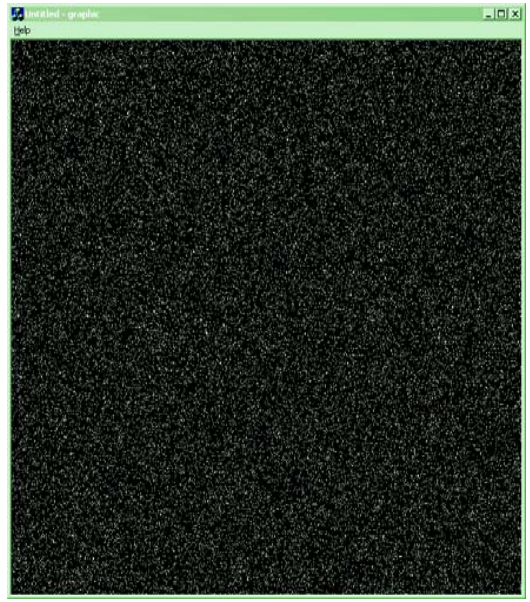
Application – Traveling Salesman Problem



1K cities



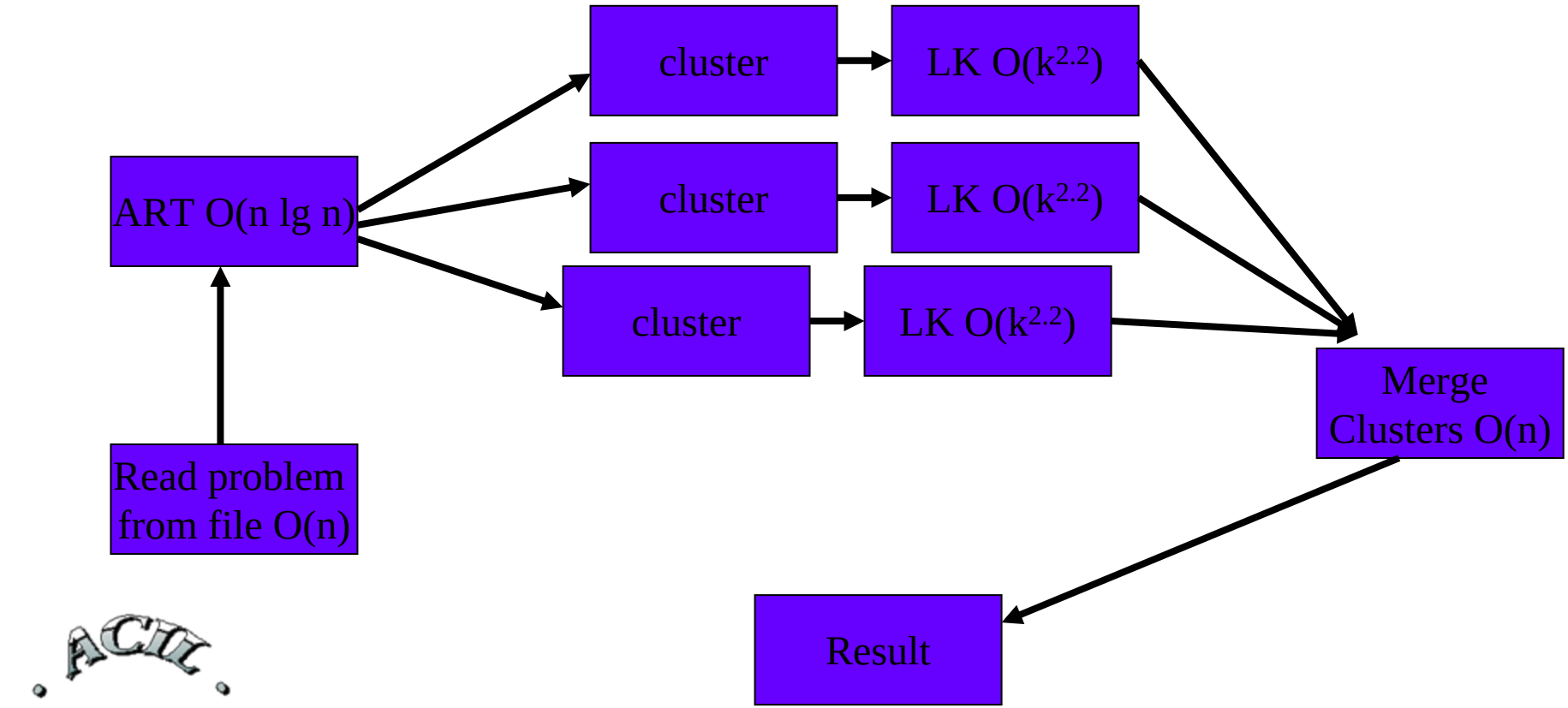
10K cities



1M cities

S. Mulder and D. Wunsch, “Million city traveling salesman problem solution by divide and conquer clustering with adaptive resonance neural networks,” *Neural Networks*, vol. 16, pp. 827-832, 2003.

Algorithm Overview



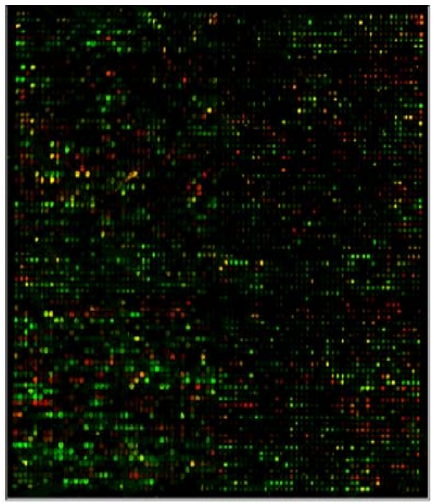
#cities	Tour Length	1P Time	2P Time	Vig factor	% off	Speedup
1000	2.58E+07	0.422	0.281	0.7	10.40%	1.50
2000	3.61E+07	1.031	0.672	0.7	10.64%	1.53
8000	7.14E+07	8.328	4.281	0.72	10.97%	1.95
10000	7.97E+07	11.359	7.297	0.75	10.57%	1.56
20000	1.12E+08	24.641	14.406	0.8	10.53%	1.71
250000	4.00E+08	315.078	209.687	0.92	11.64%	1.50
1000000	7.94E+08	1468.165	986.48	0.97	11.03%	1.49
10000000	2.52E+09	10528.7		0.98	1.27%	
CONCORDE						
1000	2.34E+07	1.670				
2000	3.26E+07	3.500				
8000	6.43E+07	26.570				
10000	7.20E+07	37.620				
20000	1.01E+08	84.830				
250000	3.58E+08	1379.540				
1000000	7.15E+08	9013.53				
10000000	2.495E+09	43630.7				

Even better news…

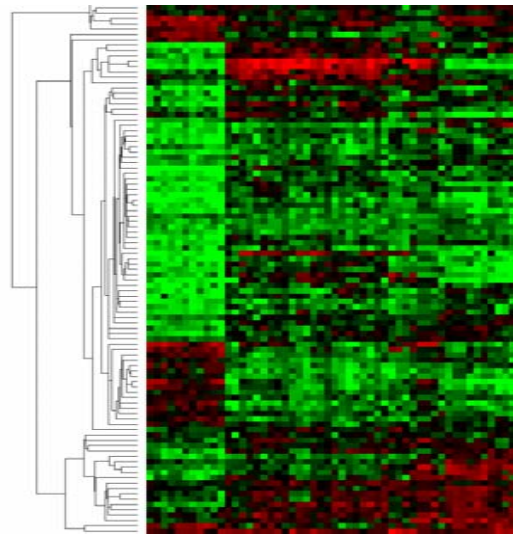
- Continued Scaling Results
- Parallelizability
- Memory Management

# Cities	Concorde	Clustering Divide & Conquer
250 K	68 MB	7 MB
1 M	261 MB	26 MB

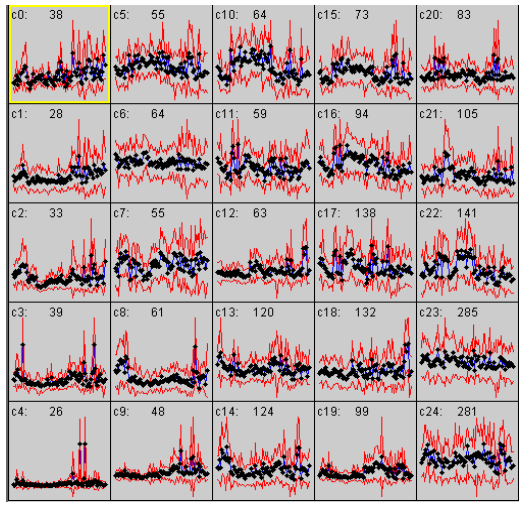
Application - Gene Expression Data Analysis: Gene Clustering



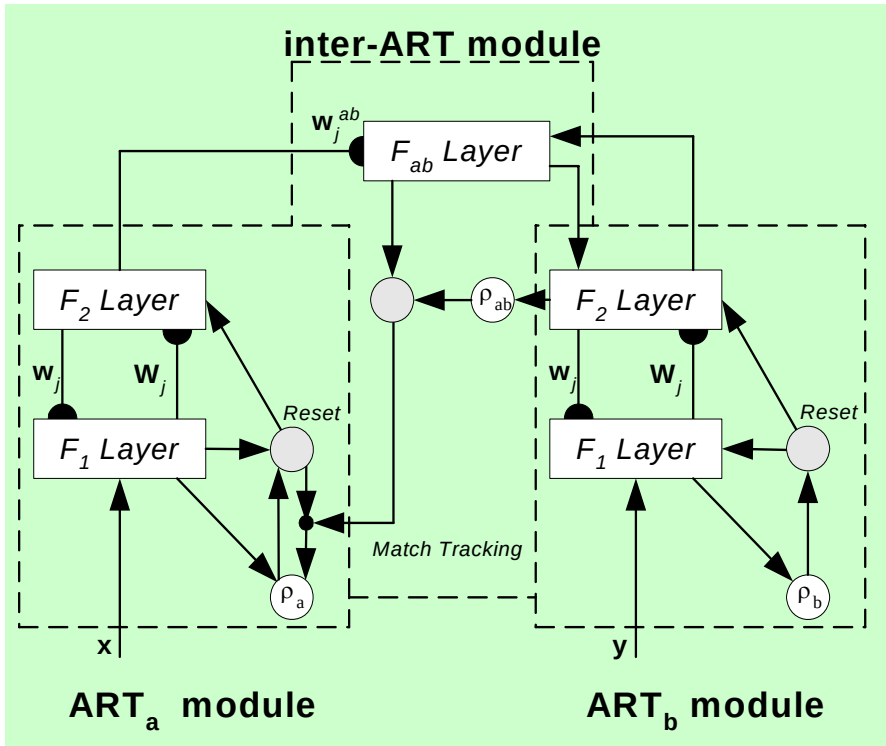
Hierarchical Clustering (CLUSTER and TreeView software, M. Eisen)

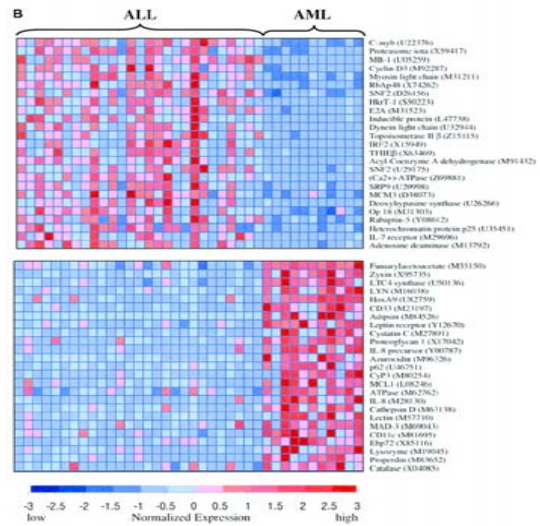


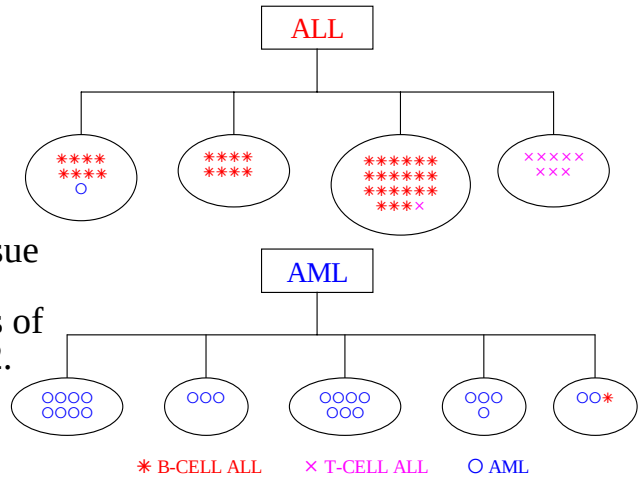
SOFM Clustering
(GeneCluster software, Whitehead Institute/MIT Center for Genome Research)



Application - Gene Expression Data Analysis: Cancer Identification



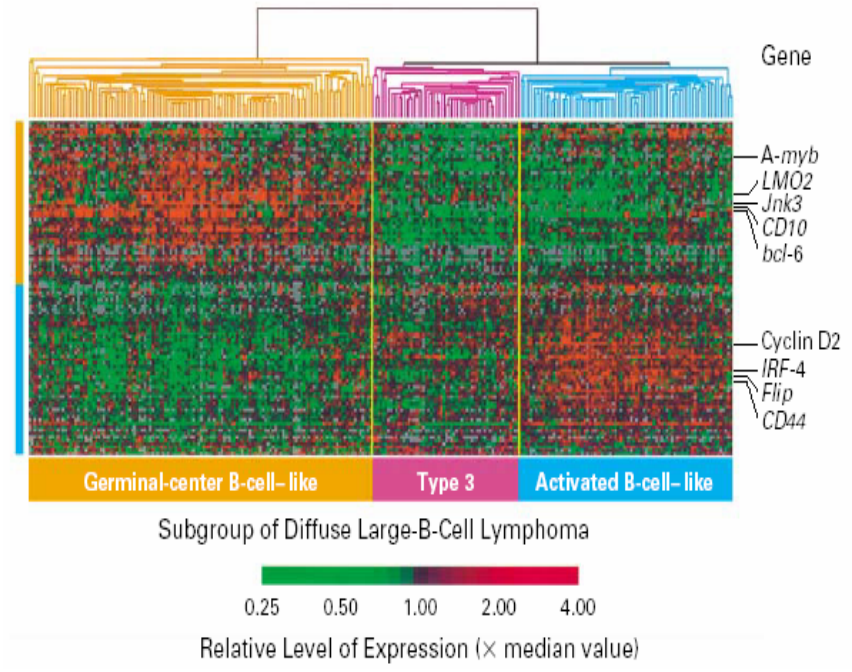


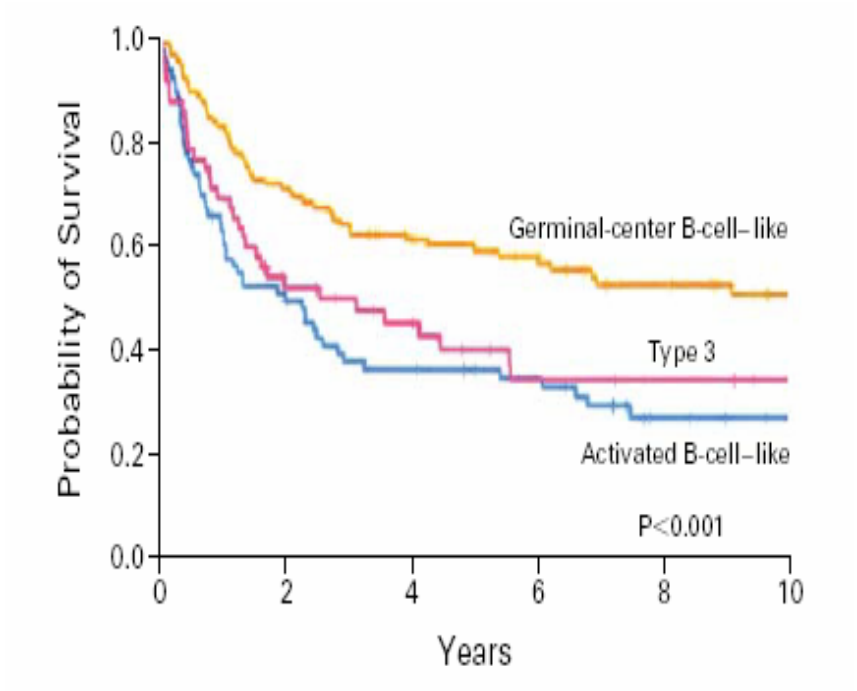


Rui Xu, G. Anagnostopoulos and Donald C. Wunsch II, "Tissue Classification Through Analysis of Gene Expression Data Using A New Family of ART Architectures", Proceedings of IJCNN 2002, vol.1, pp.300 - 304, Honolulu, Hawaii, 2002.

T. Golub, et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, 286: 531-537,1999.

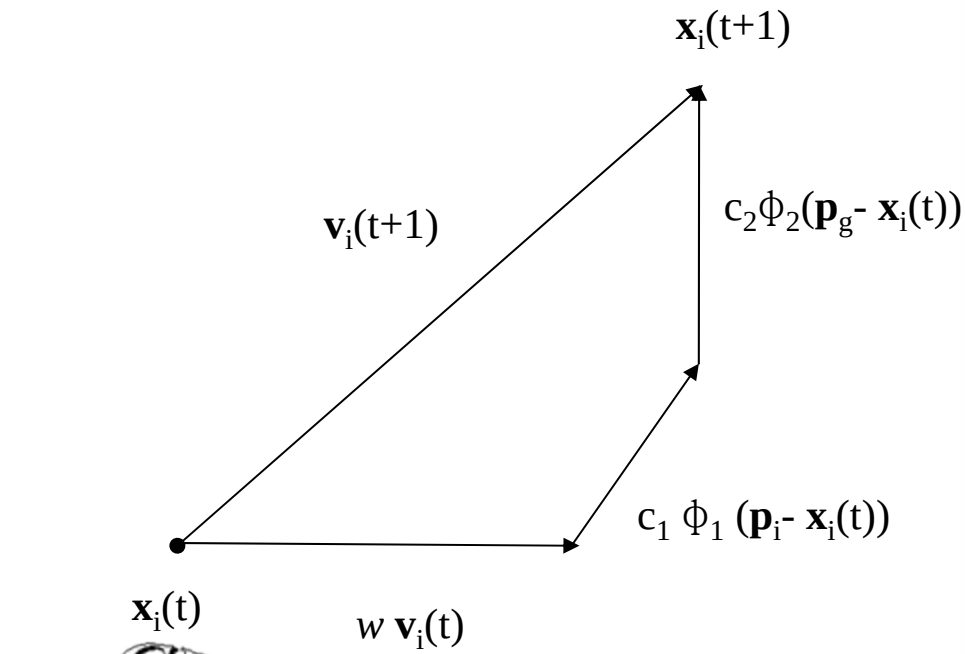
Application - Gene Expression Data Analysis: Survival Analysis

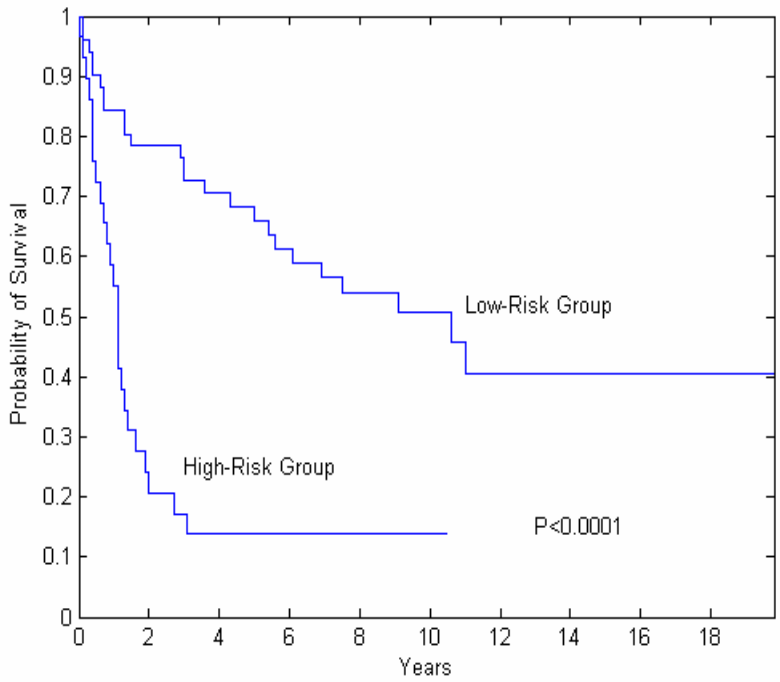




A. Rosenwald, et. al., "The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma," *The New England Journal of Medicine*, vol. 346, no. 25, pp. 1937-1947, 2002.

Application - Gene Expression Data Analysis: Dimension Reduction by Particle Swarm Optimization



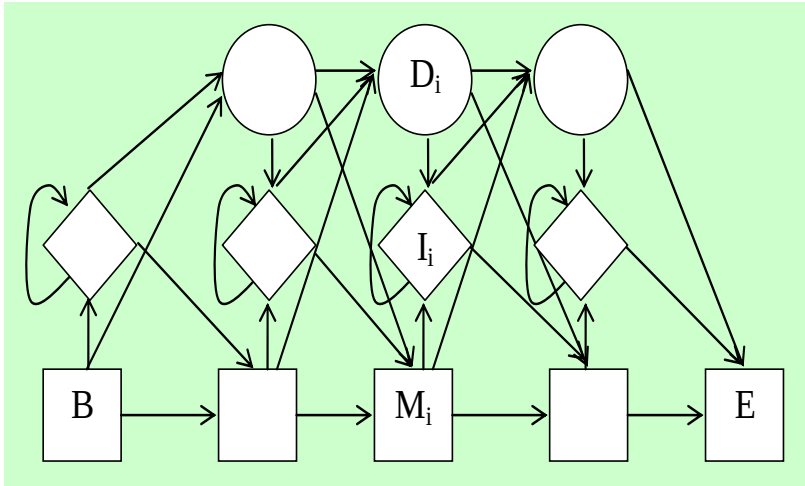


Kaplan-Meier curves show significant differences in survival of the high-risk and low-risk group in the DLBCL data set of Resenwald et al..

Application - DNA or Protein Sequence Clustering

```
                                10                  20                  30
HBA_HUMAN       ..VLSPADKTNVKAAWGKVGAHAGEYGAEALERHF.....
HBA_ORNAN       ..MLTDAEKKEVTALWGKAAGHGEEYGAEALERLF.....
HBAD_ANAPL      ..MLTAEDKKLITQLWEKVAGHQEEFGSEALQRMF.....
HBA_CARAU       ..SLSDKDKAVVKALWAKIGSRADEIGAEALGRML....
HBB_HUMAN       v.HLTPEEKSAVTALWGKVNV—DEVGGEALGRLL……
```

DNA or protein sequences to be clustered



DNA or protein families

Hidden Markov Model (A. Krogh, et. al, Hidden Markov models in computational biology: Applications to protein modeling, *Journal of Molecular Biology*, 1994.)

Challenges

- Generate arbitrary shapes of clusters
- Handle large volume of data as well as high-dimensional features with acceptable time and storage complexities
- Detect and remove possible outliers and noise
- Decrease the reliance of algorithms on user-dependent parameters
- Provide some insight for the number of potential clusters without prior knowledge
- Have the capability of dealing with newly occurring data without re-learning from scratch
- Be immune to the effects of order of input patterns
- Show good data visualization and provide users with results that can simplify further analysis
- Be capable of handling both numerical and nominal data or be easily adaptable to some other data type