

Data Clustering 50 Years Beyond K-means

Anil K. Jain
Department of Computer Science
Michigan State University



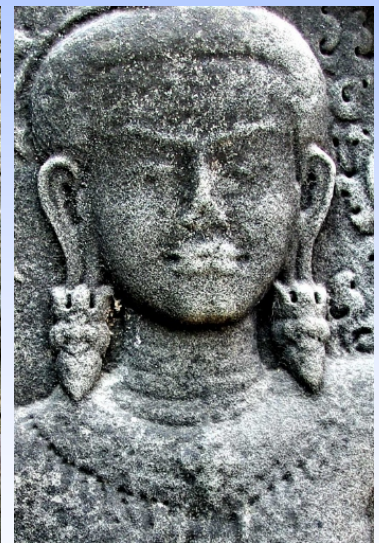
Angkor Wat



Hindu temple built by a Khmer king ~1,150AD;
Khmer kingdom declined in the 15th century; French
explorers discovered the hidden ruins in late 1800's

Apsaras of Angkor Wat

- Angkor Wat contains the most unique gallery of ~2,000 women depicted by detailed full body portraits
- What **facial types** are represented in these portraits?



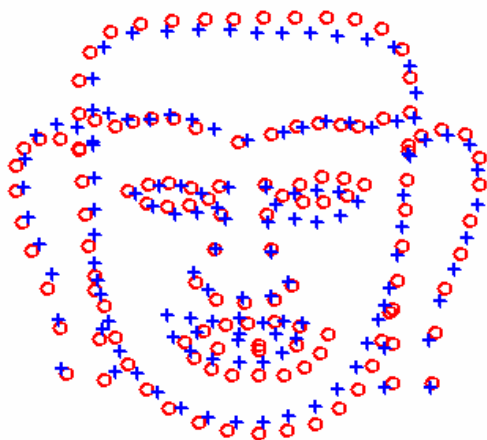
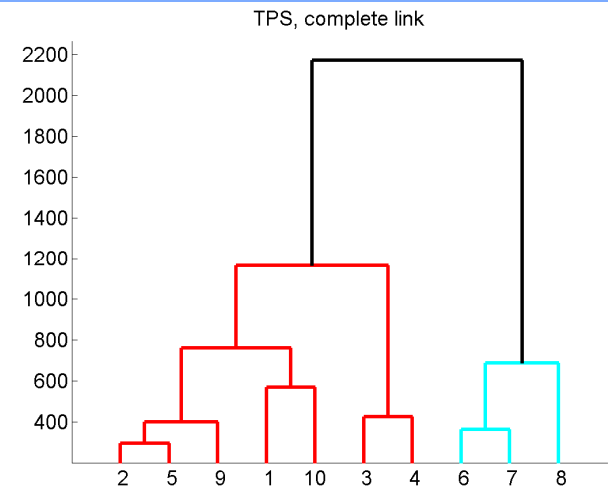
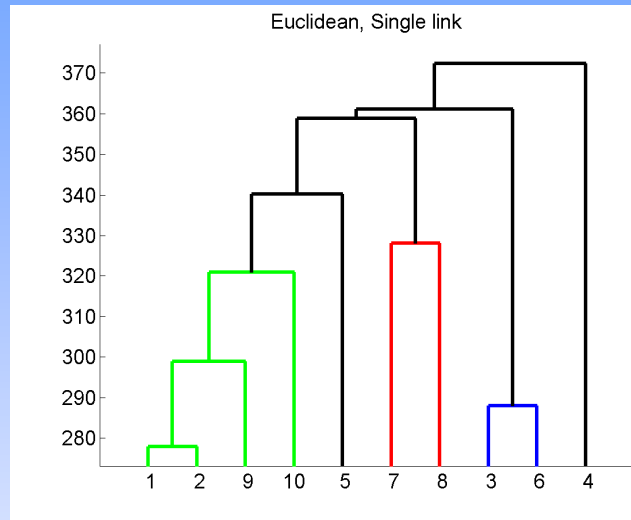
Kent Davis, "Biometrics of the Godeless", DatAsia, Aug 2008

S. Marchal, "Costumes et Parures Khmers: D'apres les devata D'Angkor-Vat", 1927

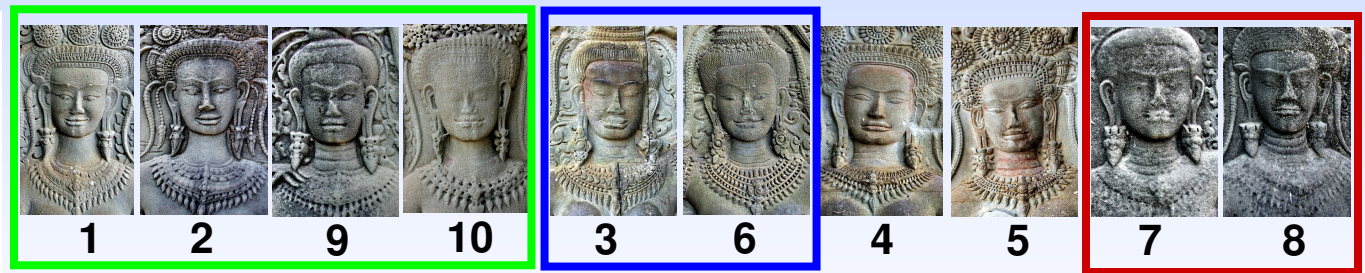
Clustering of Apsara Faces



127 landmarks



Shape alignment



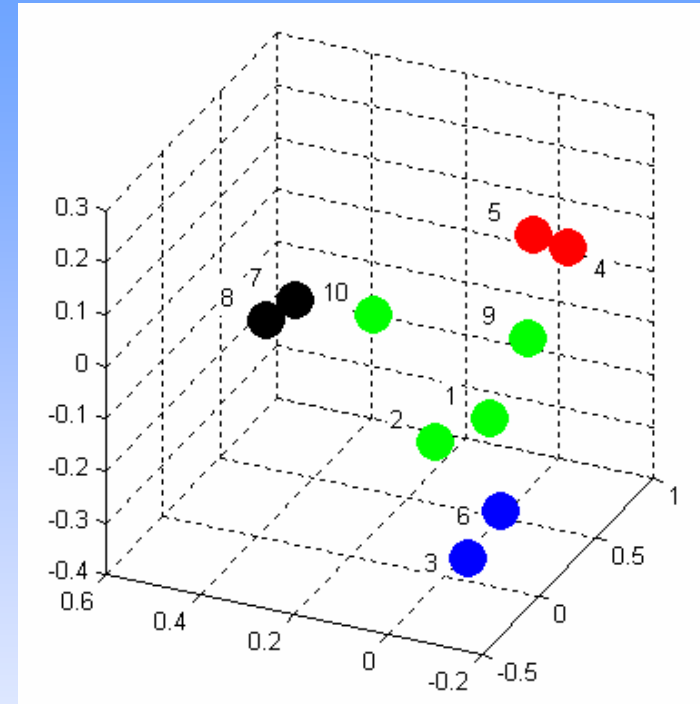
Single Link clusters

An ethnologist needs to validate the groups

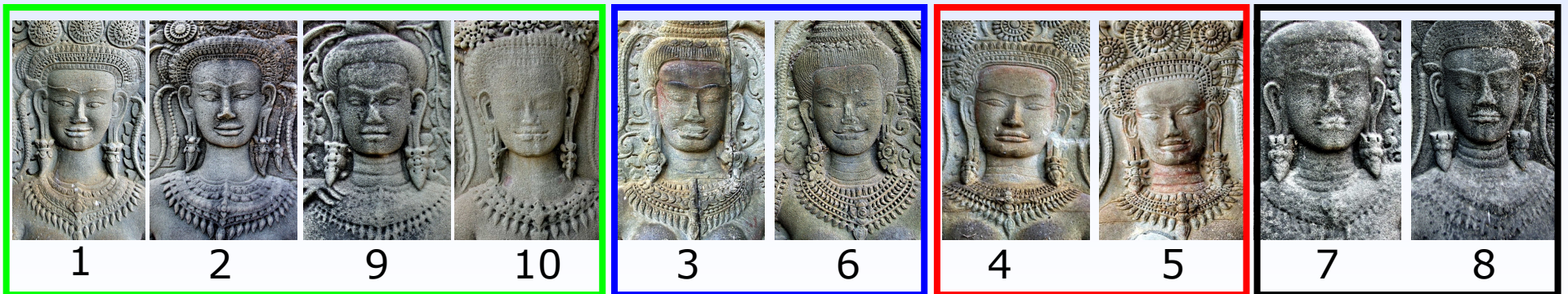
Clustering of Apsara Faces

.51	.72	.96	.70	.73	.89	.78	.61	.61
.67	1	.72	.70	.80	.66	.55	.55	.59
.87	.78	.53	.73	.77	.71	.83		
.69	.78	.87	.94	.88	.97			
.72	.74	.75	.62	.71				
.72	.76	.69	.78					
.60	.75	.86						
.72	.67							
0	.68							

Dissimilarity matrix



4 clusters with K-means in 3D feature space



Data Explosion

- The digital universe was ~281 exabytes (281 billion gigabytes) in 2007
- By 2011, the digital universe will be 10 times the size it was in 2006
- Images and video, captured by over one billion devices (camera phones), are the major source
- To archive and effectively use this data, we need tools for data visualization & categorization

<http://eon.businesswire.com/releases/information/digital/prweb509640.htm>

<http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>

Exploratory Data Analysis

- A collection of techniques to gain insight into data, **uncover underlying structure**, generate hypotheses, detect anomalies, and identify important measurements (Tukey, 1977)
- Does not require assumptions common in **confirmatory data analysis** (hypothesis testing or discriminant analysis)
- Graphical techniques, visualization, outlier detection, multidimensional scaling, **clustering**

Clustering

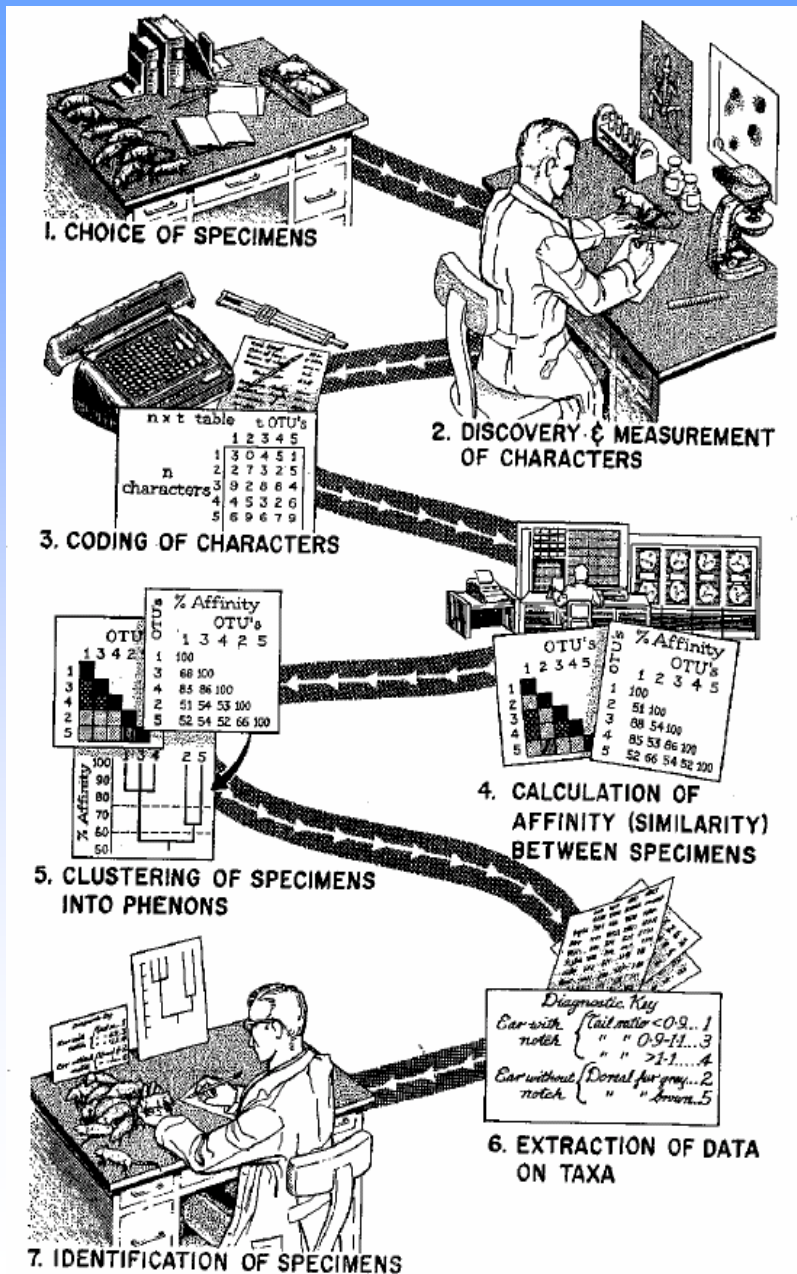
“A statistical classification technique for **discovering** whether the **individuals** of a population fall into different **groups** by making quantitative comparisons of multiple **characteristics**” - Webster's

- Q-analysis, typology, grouping, clumping, taxonomy, unsupervised learning
- Given a **representation** of n **objects**, find K **clusters** based on the given measure of **similarity**

A.K. Jain and R. C. Dubes, algorithms for Clustering Data, Prentice Hall, 1988

http://www.cse.msu.edu/~jain/Clustering_Jain_Dubes.pdf

Numerical Taxonomy



Michener (1957) makes a distinction between hierarchies of categories for

- **Convenience**: as a method for organizing data
- **Natural classification**: based on phylogenetic relationship or degree of similarity among *forms*

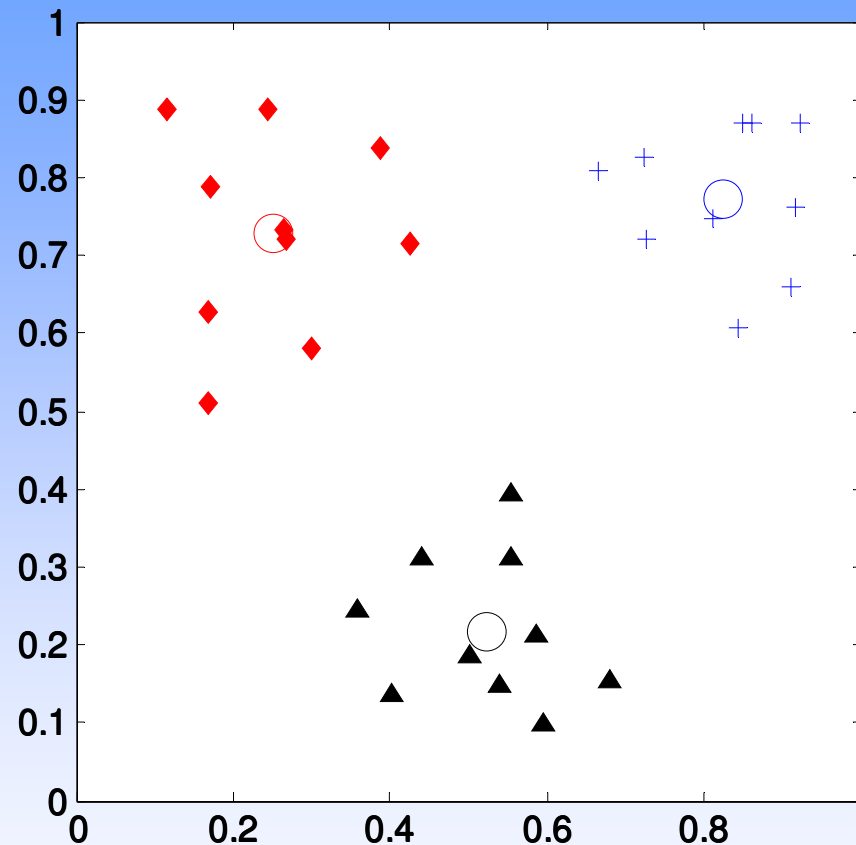
Historical Developments

- **Cluster analysis** first appeared in the title of a 1954 article analyzing anthropological data (JSTOR)
- Hierarchical Clustering: Sneath (1957), Sorensen (1957)
- **K-Means: Steinhaus¹ (1956), Lloyd² (1957), Cox³ (1957), Ball & Hall⁴ (1967), MacQueen⁵ (1967)**
- Mixture models (Wolfe, 1970)
- Graph-theoretic methods (Zahn, 1971)
- K Nearest neighbors (Jarvis & Patrick, 1973)
- Fuzzy clustering (Bezdek, 1973)
- Self Organizing Map (Kohonen, 1982)
- Vector Quantization (Gersho and Gray, 1992)

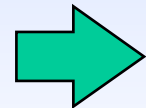
¹Acad. Polon. Sci., ²Bell Tel. Report, ³JASA, ⁴Behavioral Sci., ⁵Berkeley Symp. Math Stat & Prob.



K-Means Algorithm



- Initialization
- Value of K
- Distance metric



Bisecting K-means (Karypis et al.); X-means (Pelleg and Moore); K-means with constraints (Davidson); scalable K-means (Bradley et al.)

Beyond K-Means

- ~155 papers on clustering in ML conf. (2006-07); Google Scholar: 1,560 papers with “data clustering” in 2007 alone!
- Methods differ on **choice of objective function, generative models and heuristics**
- Density-based (Ether et al., 1996)
- Subspace (Agrawal et al., 1998)
- Spectral (Hagen & Kahng, 1991; Shi & Malik, 2000)
- Dirichlet Process (Ferguson, 1973; Rasmussen, 2000)
- Information bottleneck (Tishby et al., 1999)
- Non-negative matrix factorization (Lee & Seung, 1999)
- Ensemble (Strehl & Ghosh, 2002; Fred & Jain, 2002)
- Semi-supervised (Wagstaff et al., 2003; Basu et al., 2004)
- Overlapping (Segal et al., 2003; Banerjee et al., 2005)
- Maximum margin (Xu et al., 2005)
- Discriminative (Bach & Harchaoui, 2007; Ye et al., 2007)

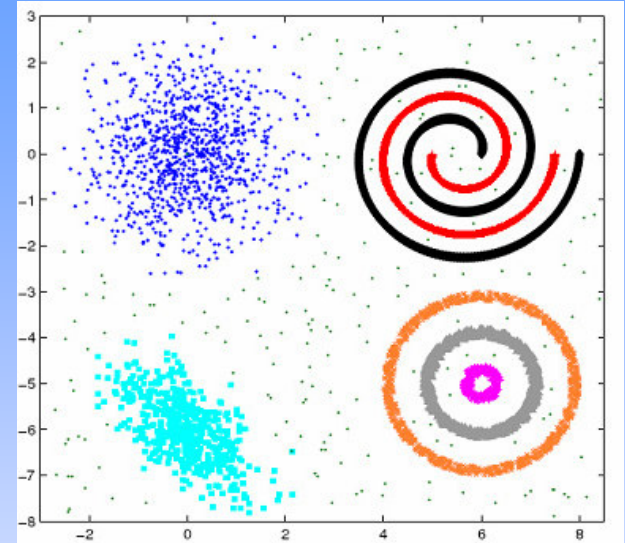
User's Dilemma!

- What features and normalization scheme to use?
- How to define pair-wise similarity?
- How many clusters?
- Which clustering method?
- How to choose algorithmic parameters?
- Does the data have any clustering tendency?
- Are the discovered clusters & partition valid?
- How to visualize, interpret & evaluate clusters?

Dubes and Jain, "Clustering Techniques: User's Dilemma", *Pattern Recognition*, 1976

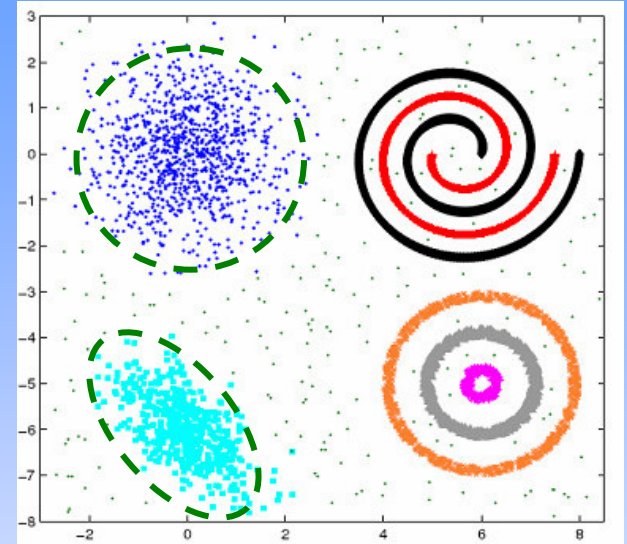
What is a Cluster?

- A set of entities which are **alike**; entities from different clusters are not alike



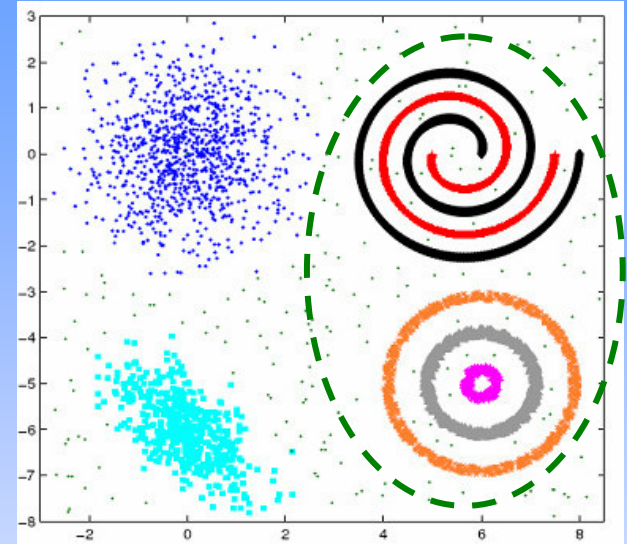
What is a Cluster?

- A set of entities which are **alike**; entities from different clusters are not alike
- Compact clusters
 - within-cluster **distance** < between-cluster distance



What is a Cluster?

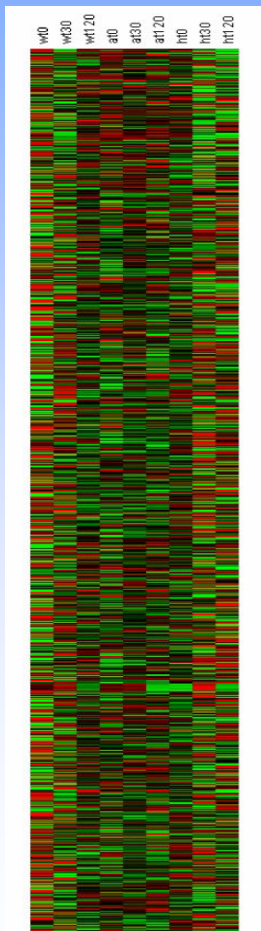
- A set of entities which are **alike**; entities from different clusters are not alike
- Compact clusters
 - within-cluster **distance** < between-cluster distance
- Connected clusters
 - within-cluster **connectivity** > between-cluster connectivity
- Ideal cluster: **compact** and **isolated**



Representation

Objects: pixels, images, time series, documents

Representation: features, similarity



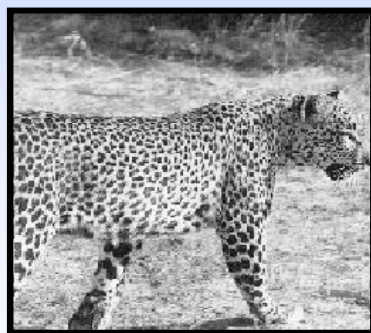
Gene Expressions



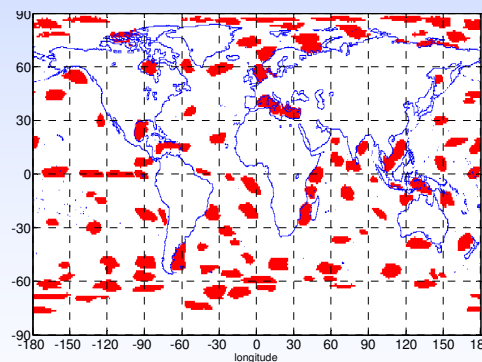
Image retrieval



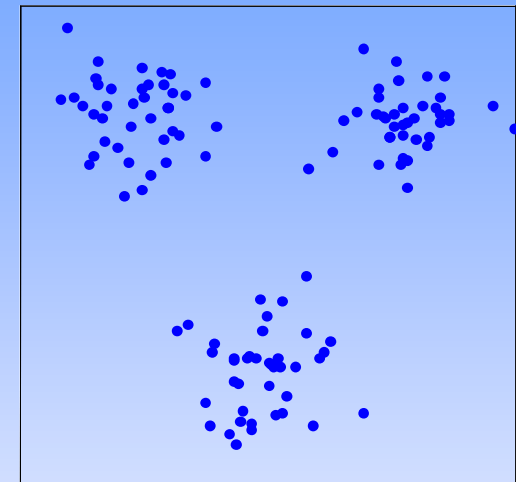
Handwritten digits



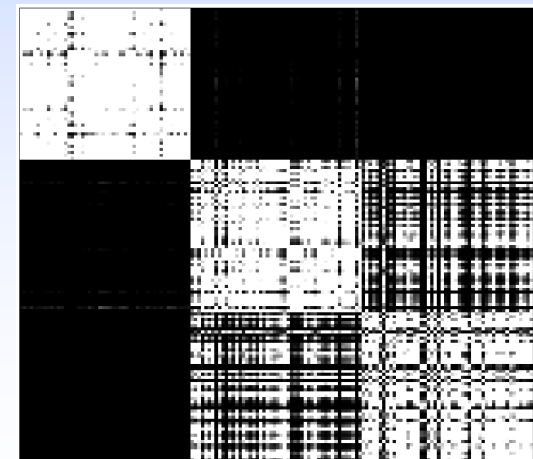
Segmentation



Sea-surface temperature
time series



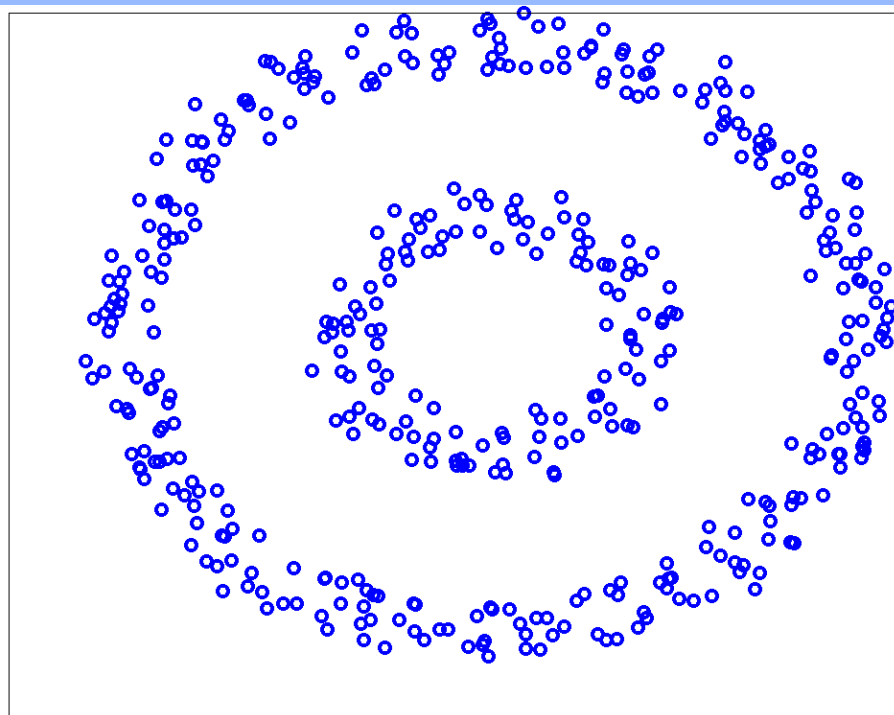
nxd pattern matrix



nxn similarity matrix

Good Representation

A good representation leads to compact & isolated clusters



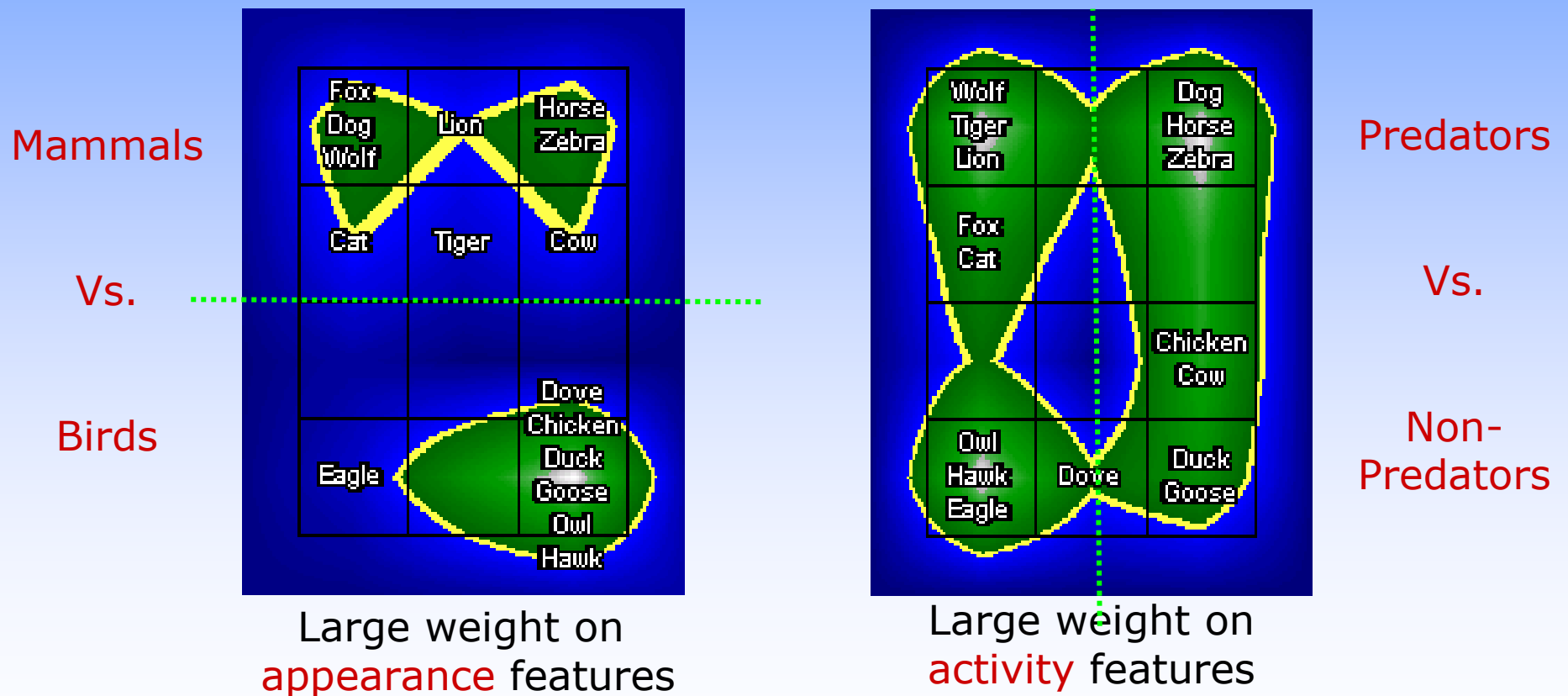
Points in given 2D space



Representation based on eigenvectors of RBF kernel

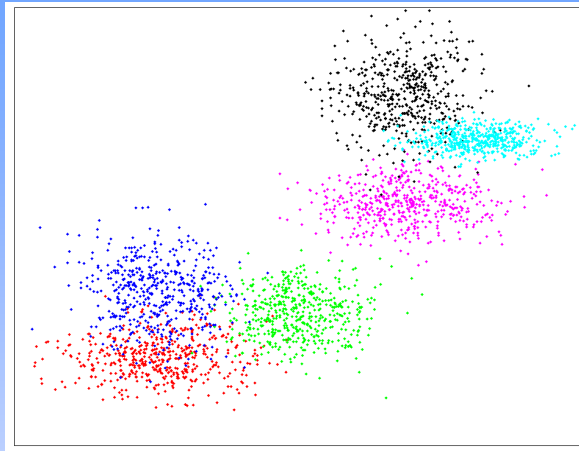
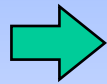
Purpose of Grouping

Two different meaningful groupings of 16 animals based on 13 Boolean features (appearance & activity)

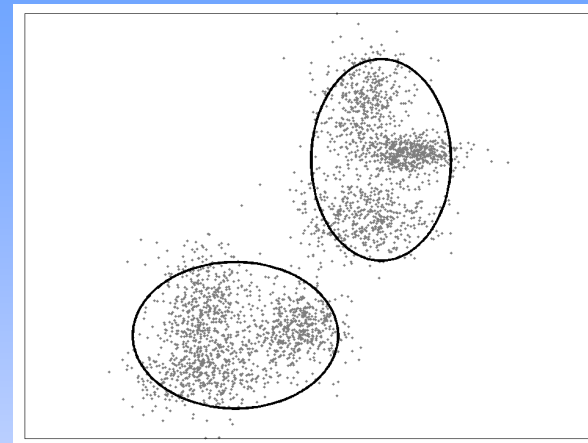


<http://www.ofai.at/~elias.pampalk/kdd03/animals/>

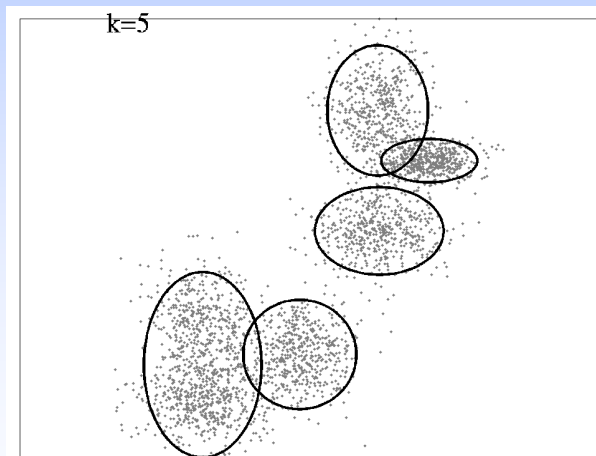
Number of Clusters



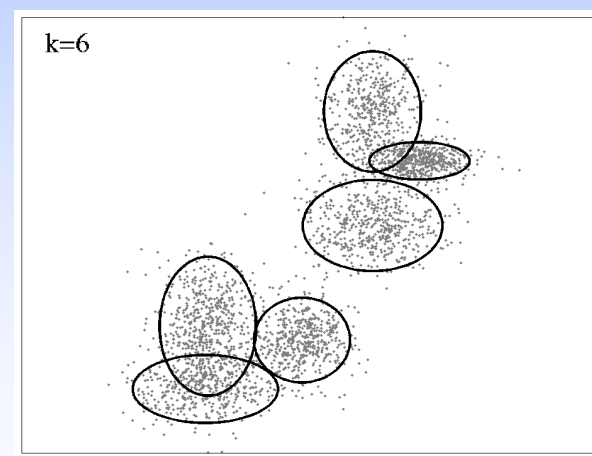
Original data
Training data = 6



Clustering with $K = 2$



Clustering with $K = 5$

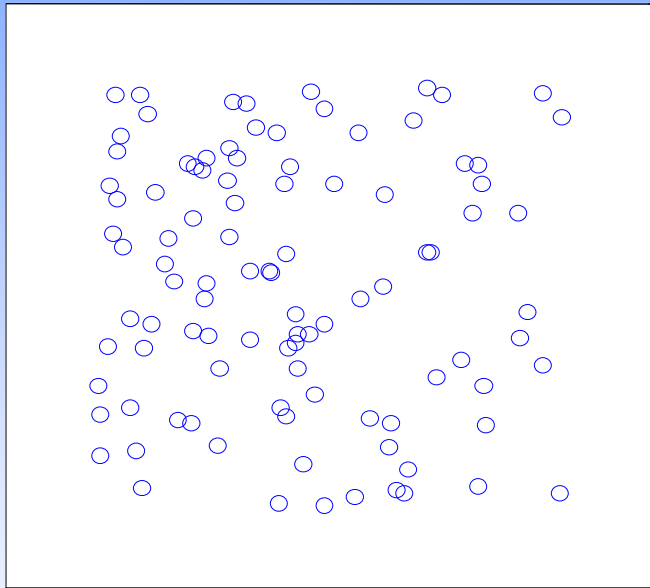


Clustering with $K = 6$

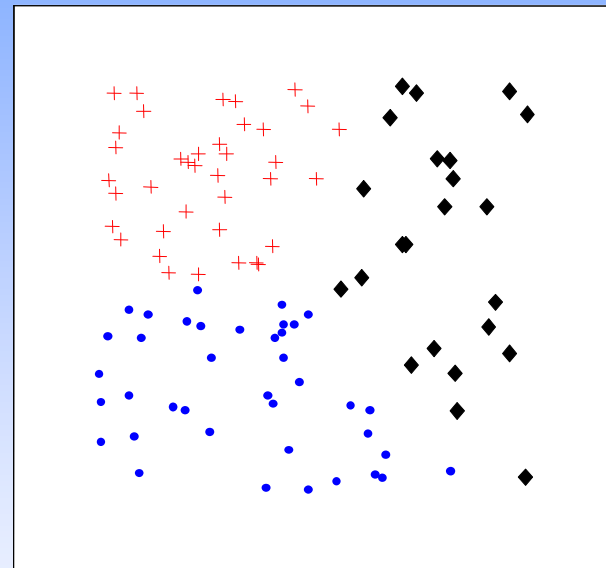
Clustering is in the eyes of the beholder

Cluster Validity

- Clustering algorithms find clusters, even if there are no **natural** clusters in the data!



100 2D uniform data points

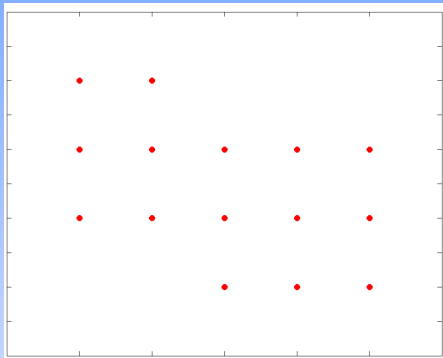


K-Means with $K=3$

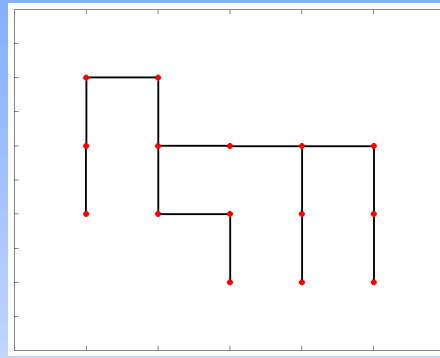
- Cluster stability** (Lange et. al, 2004)

Comparing Clustering Methods

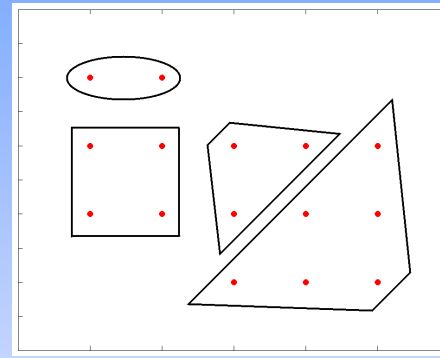
Which clustering algorithm is the best?



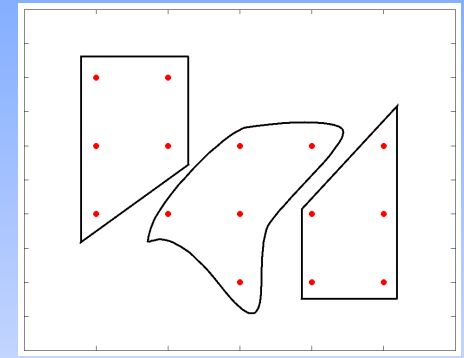
15 Data points



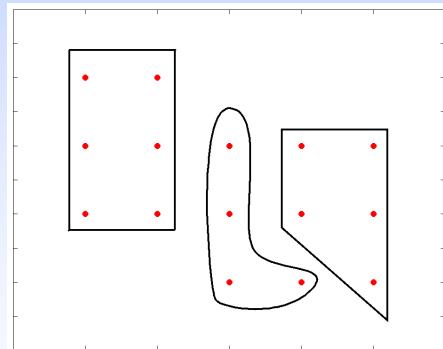
MST



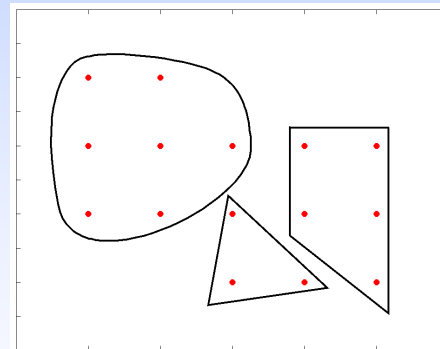
FORGY



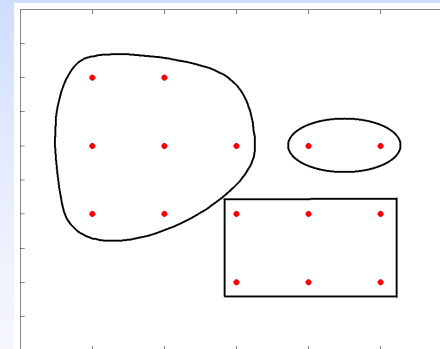
ISODATA



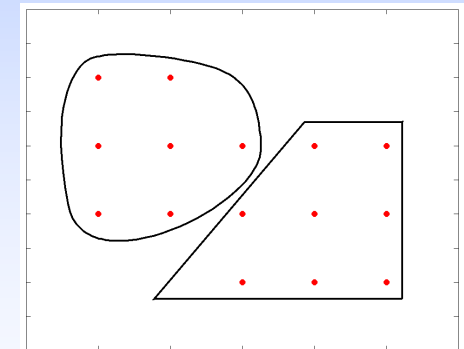
WISH



CLUSTER



Complete Link

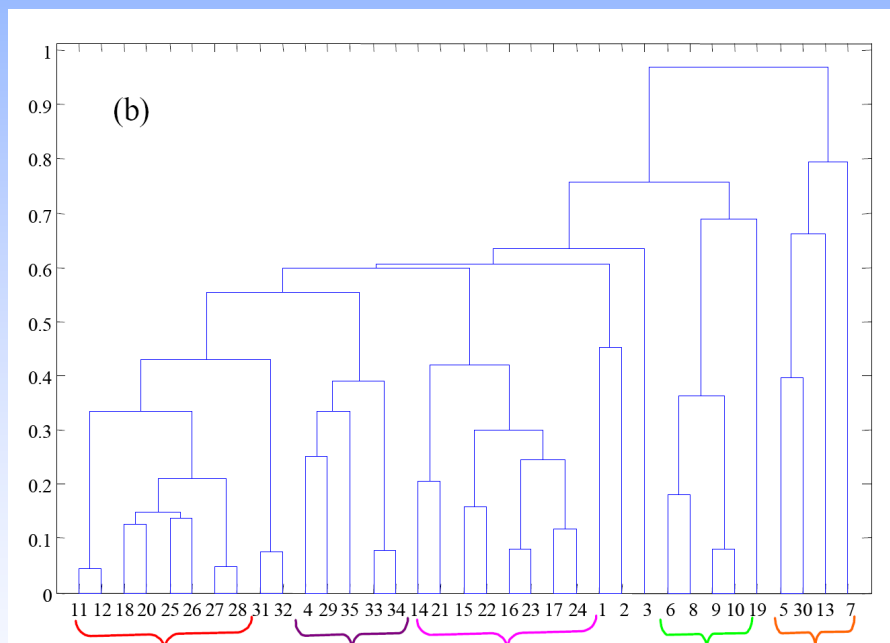


JP

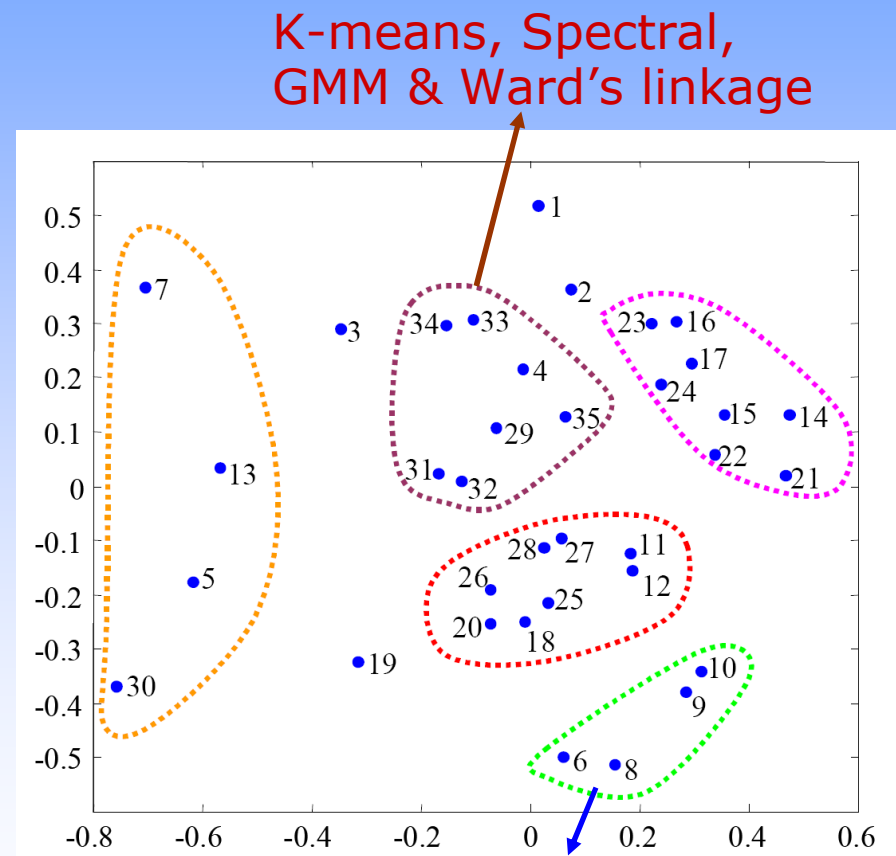
Dubes and Jain, "Clustering Techniques: User's Dilemma", *Pattern Recognition*, 1976

Grouping of Clustering Algorithms

Clustering **method** vs. clustering **algorithm**



Hierarchical clustering of 35 different algorithms (evaluated on 12 datasets)

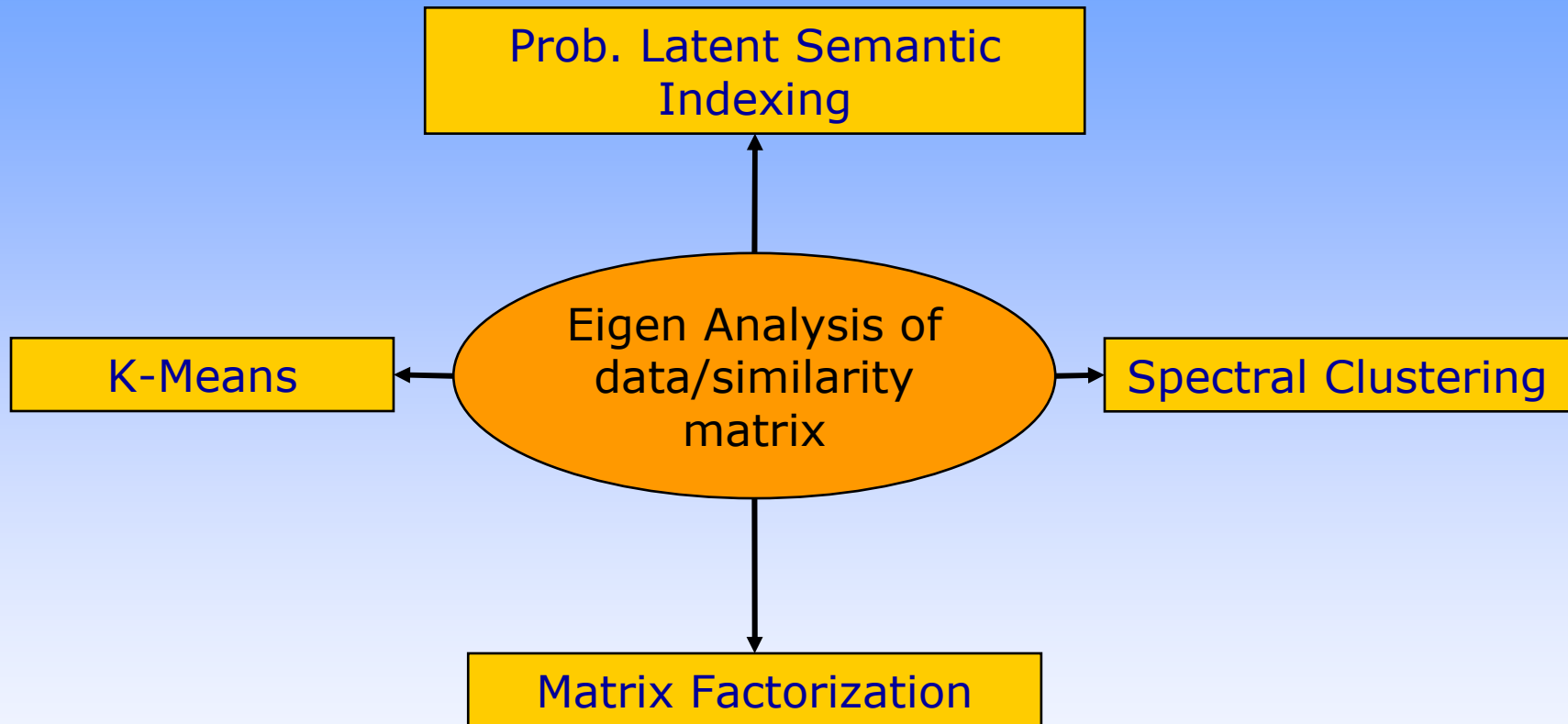


K-means, Spectral,
GMM & Ward's linkage

Chameleon variants

A. K. Jain, A. Topchy, M. Law, J. Buhmann, "Landscape of Clustering Algorithms", ICPR, 2004

Mathematical & Statistical Links



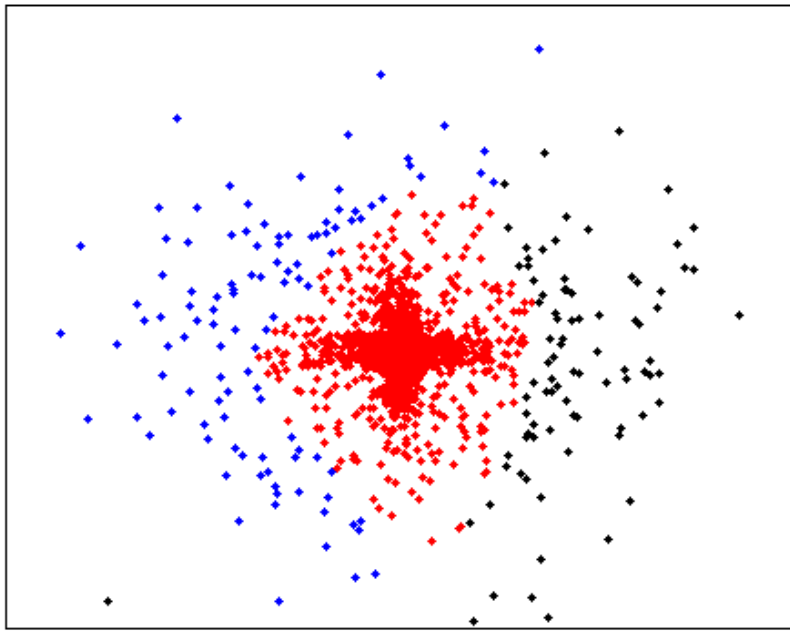
Zha et al., 2001; Dhillon et al., 2004; Gaussier et al., 2005, Ding et al., 2006; Ding et al., 2008

Admissibility Criteria

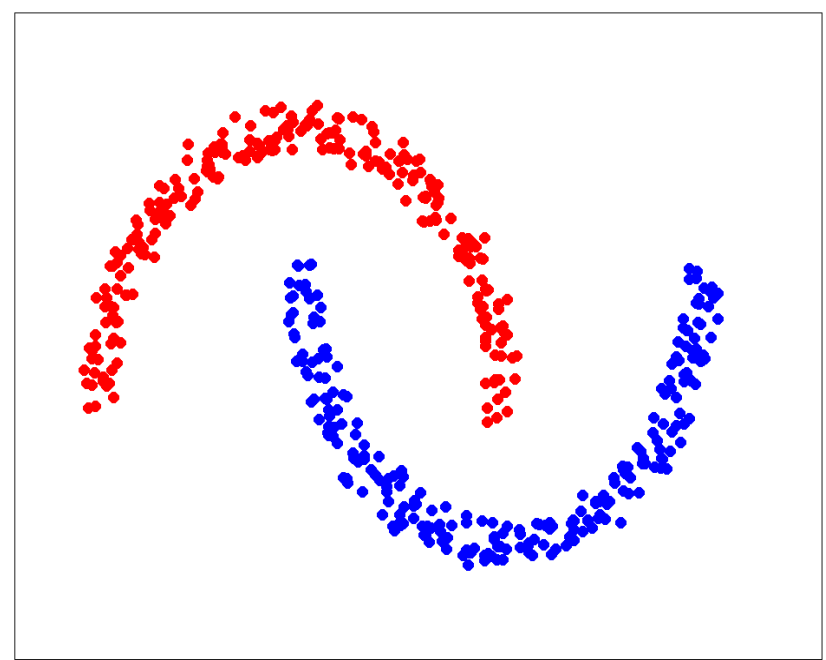
- A technique is **P-admissible** if it satisfies a **desirable** property P (Fisher & Van Ness, Biometrika, 1971)
- Properties that test sensitivity w.r.t. changes that do not alter the essential structure of data: **Point & cluster proportion, cluster omission, monotone**
- Impossibility theorem (Kleinberg, NIPS 2002); no clustering function satisfies **scale invariance, richness and consistency** properties
- Difficulty in **unifying** the informal concept of clustering and inherent **tradeoffs**

No Best Clustering algorithm

Each algorithm, implicitly or explicitly, **imposes a structure** on the data; if the match is “good”, algorithm is successful



Scatter plot with $K=3$
MSO algorithm with $K=3$



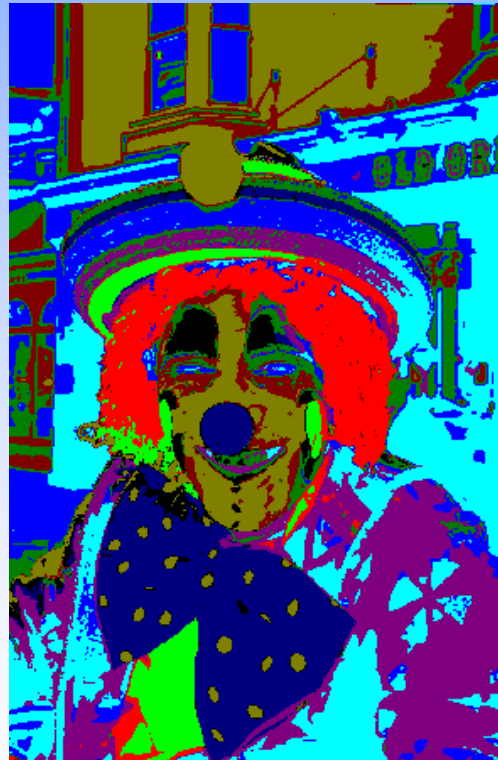
Scatter plot with $K=2$
MSO algorithm with $K=2$

Data Compression

- Pixels with similar attributes and spatial location are clustered to find segments (Leeser et al., '98)
- Each segment indexed to its mean attribute value



Input image



Segmentation

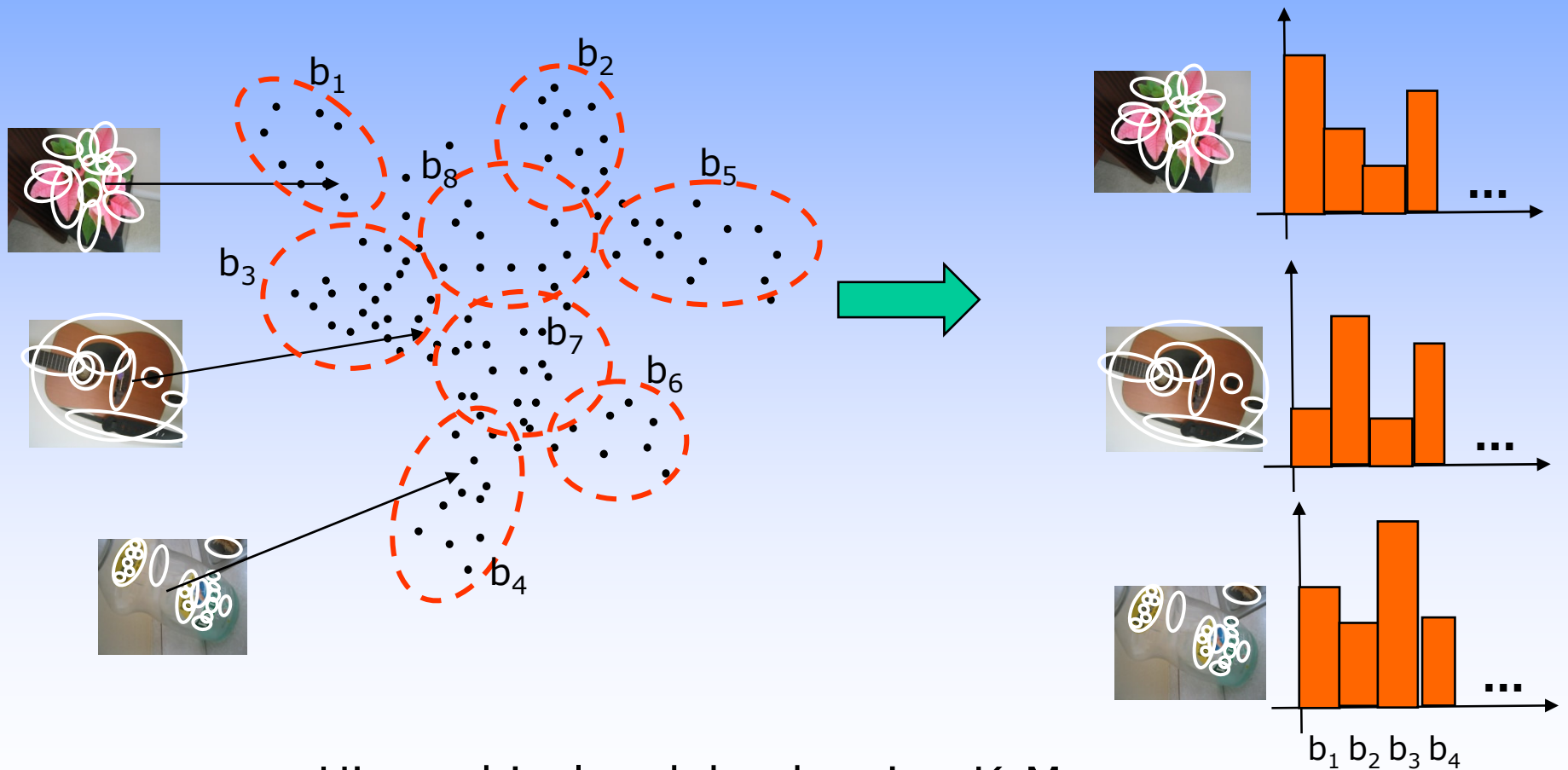


Reconstruction

<http://www.ece.neu.edu/groups/rpl/projects/kmeans/>

Object Recognition

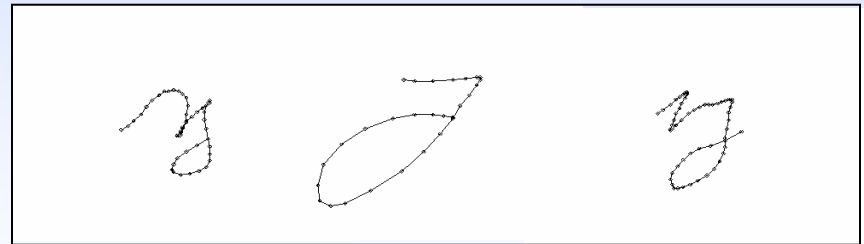
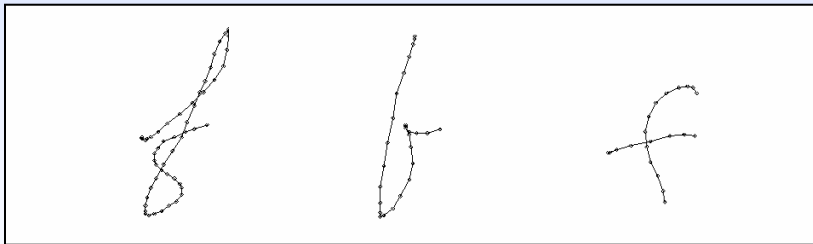
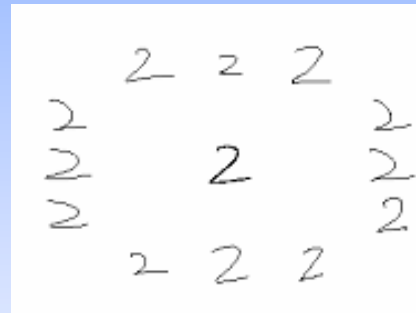
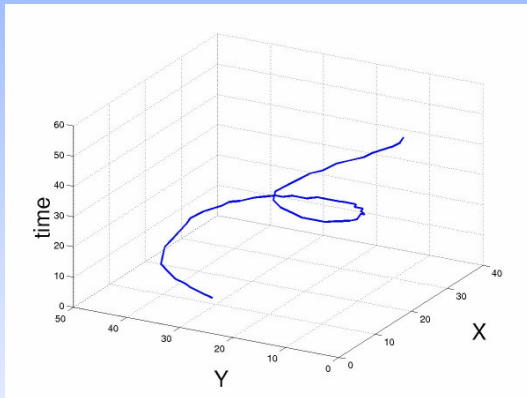
Local descriptors are hierarchically quantized in a vocabulary tree (Nister et al., CVPR, 2006)



Hierarchical codebook using K-Means

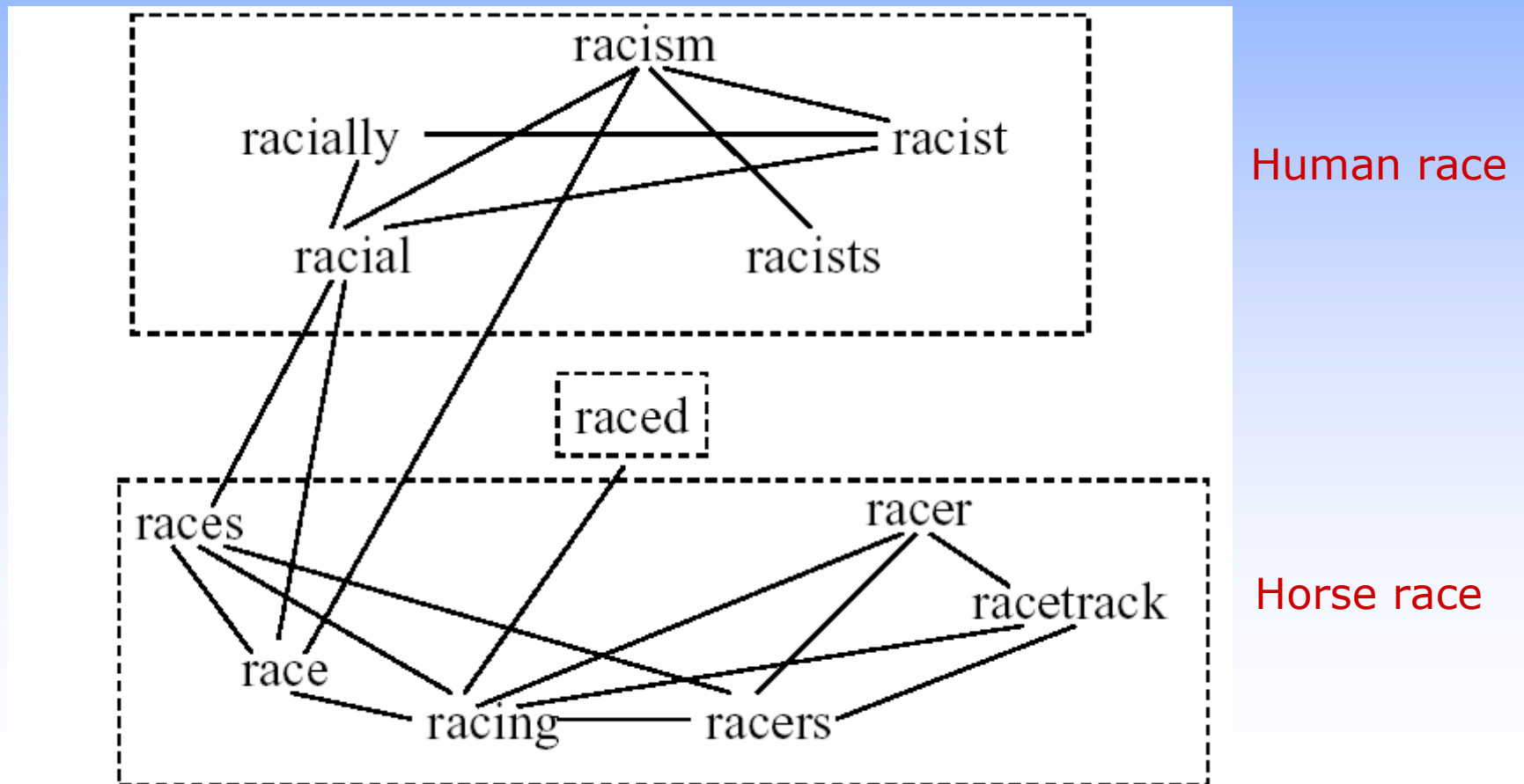
Finding Lexemes

- Find **subclasses** in handwritten “online” characters (122,000 characters written by 100 writers)
- Performance improves by modeling subclasses



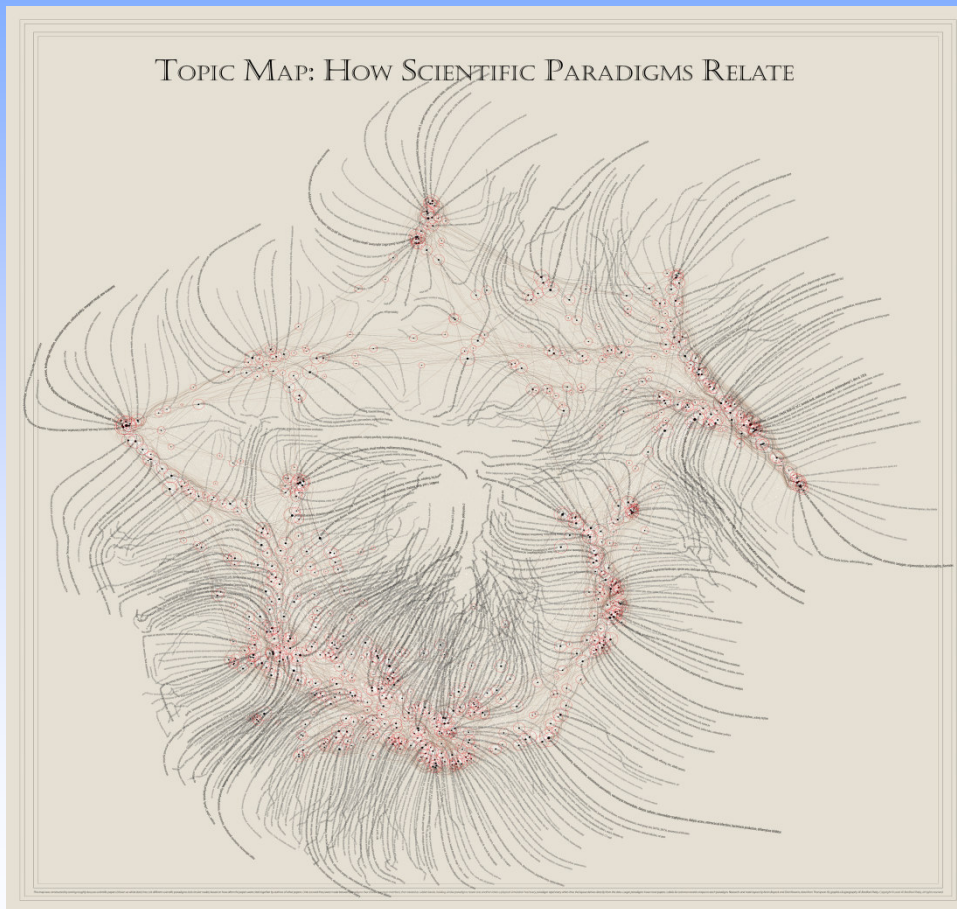
Information Retrieval

Xu & Croft (ACM TOIS, 1998) used corpus analysis based on word co-occurrence to **refine the large equivalence classes** generated by a stemmer



Map of Science

Clustering of network (relational) data



800,000 scientific papers
clustered into 776
scientific paradigms based
on how often the papers
were cited together by
authors of other papers

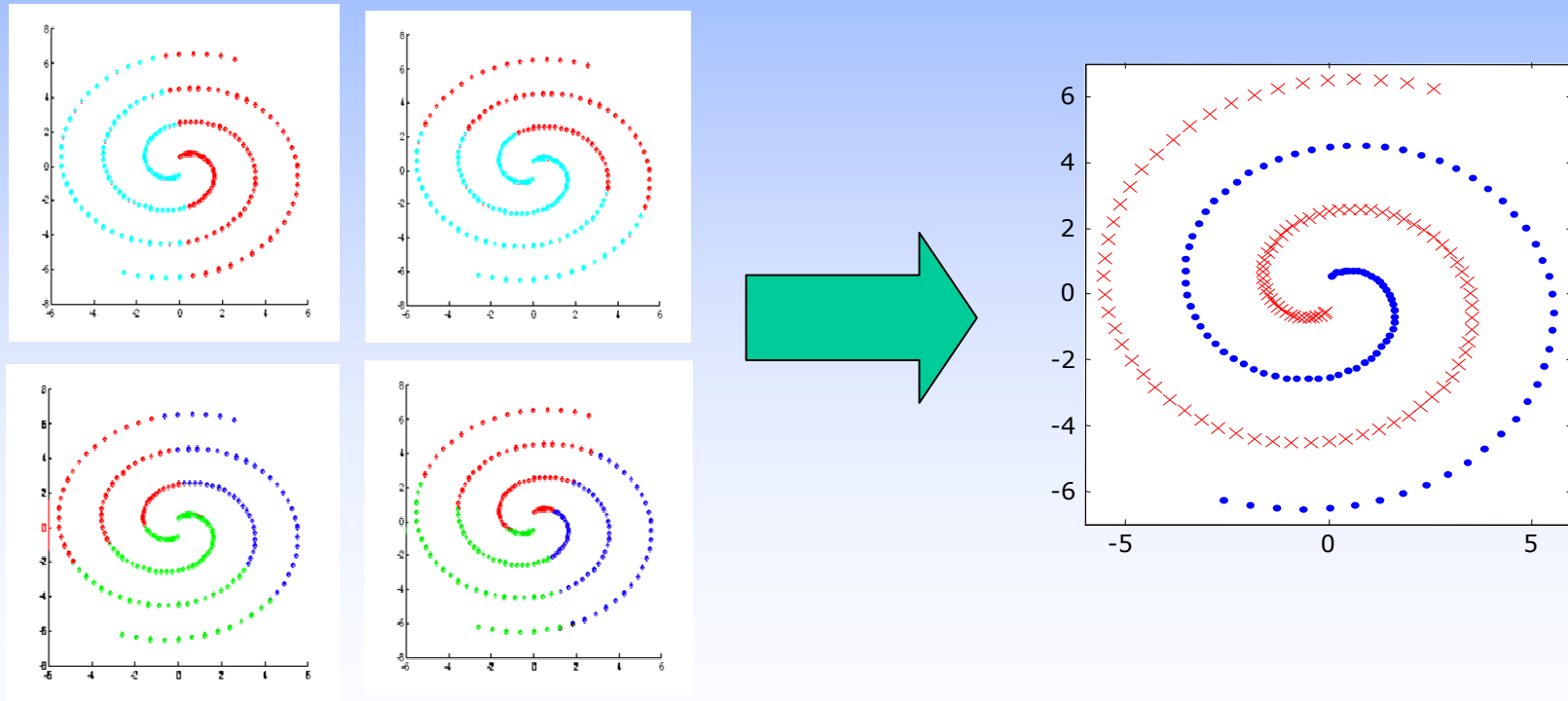
Nature (2006)

Some Trends

- Large-scale data
 - Clustering of 1.5B images into 50M clusters; 10 hours on 2000 CPUs (Liu et al., WACV 2007)
- Evidence Accumulation
 - Multi-way clustering (documents/words/authors)
 - Multi-modal data (clustering genes based on expression levels and text literature, Yang et al., CSB 2007)
- Domain Knowledge
 - How to acquire & incorporate domain knowledge? Pairwise constraints, feature constraints (e.g., WordNet)
- Complex Data Types
 - Dynamically evolving data (cluster maintenance)
 - Networks/graphs (How to define kernel/similarity matrix?)

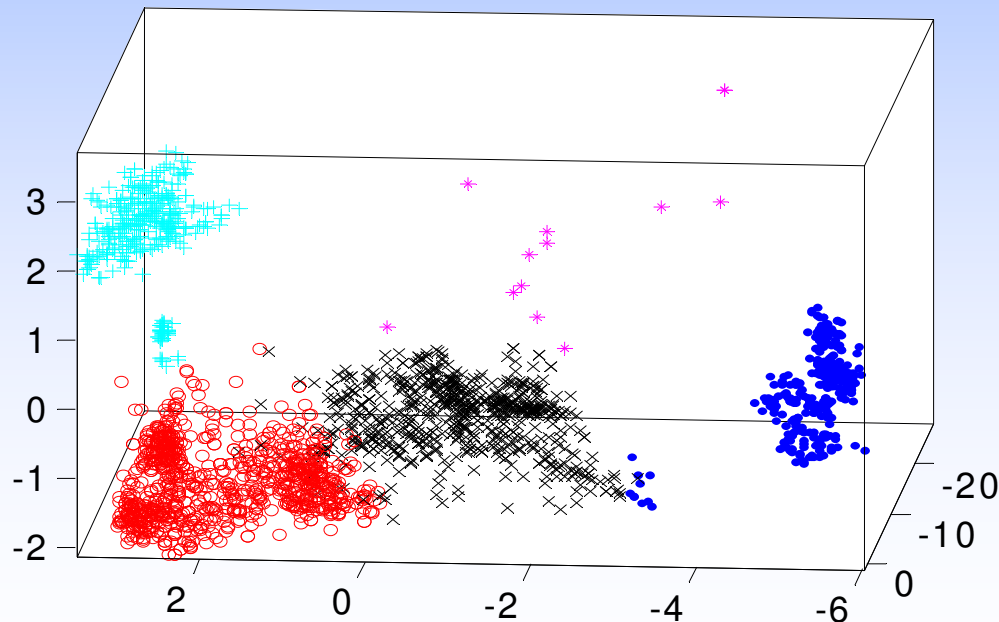
Clustering Ensemble

- Combine many “weak” partitions of a data to generate a better partition (Strehl & Ghosh, 2002; Fred & Jain, 2002)
- Pairwise co-occurrences from different K-Means partitions



Multiobjective Clustering

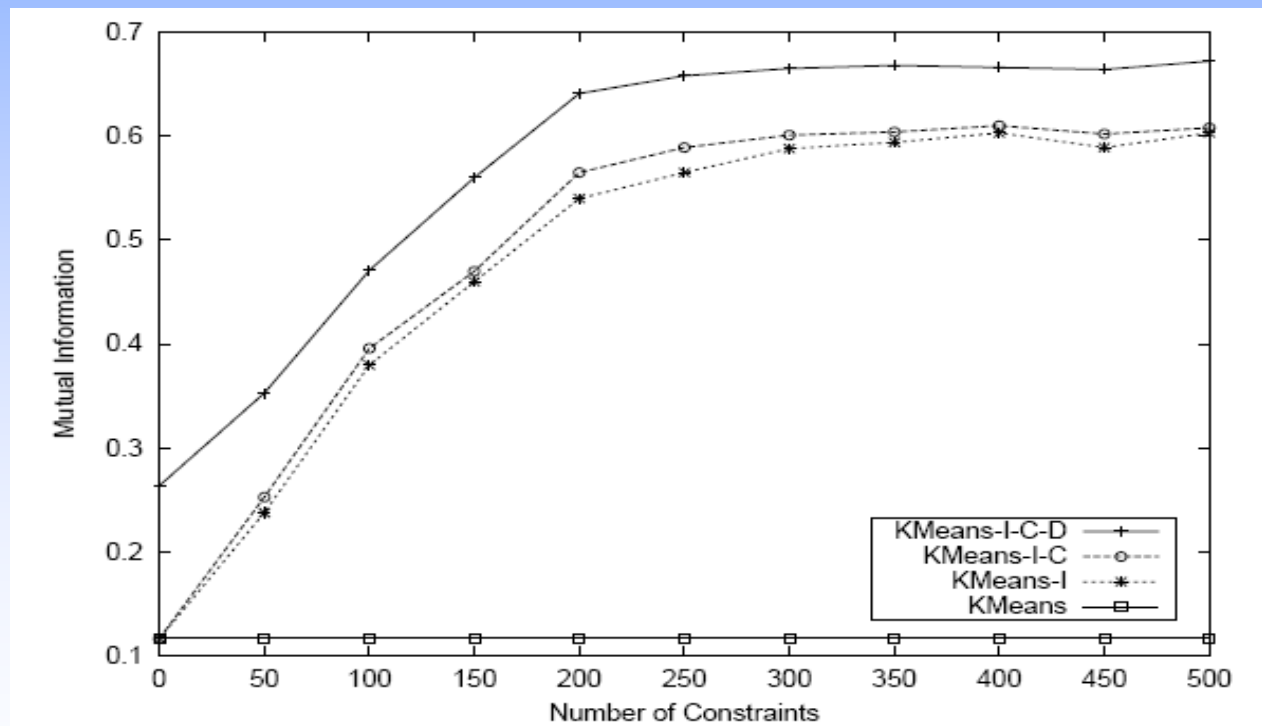
- Different clusters in data may have different shapes and densities; difficult for a single criterion
- Find **stable clusters** from different algorithms
- Four stable clusters identified in *image segmentation* data using GMM, Single-link, K-means and spectral



Law, Topchy and Jain, CVPR 2004

Semi-supervised Clustering

Clustering with side information: modify the objective function of a given algorithm or design a new algorithm to utilize **pairwise** constraints

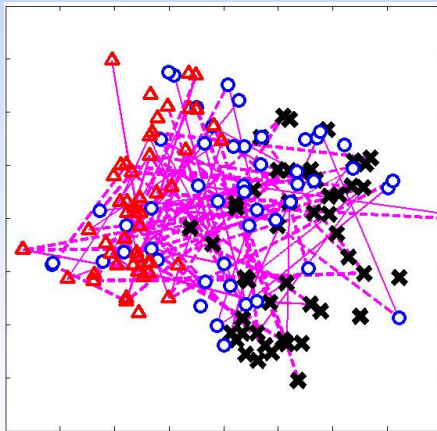


I: initialization, C: constraints, D: distance learning

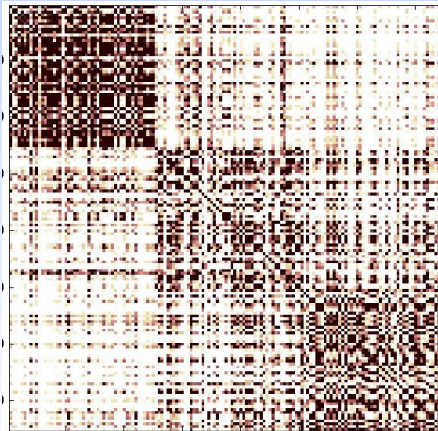
Basu et al., KDD'04

BoostCluster

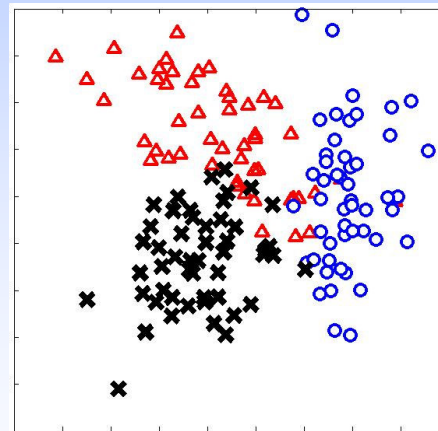
- Can we improve **any** generic clustering algorithm in the presence of “constraints”?
- **BoostCluster**: an unsupervised boosting algorithm to iteratively update the similarity matrix given the constraints & clustering output



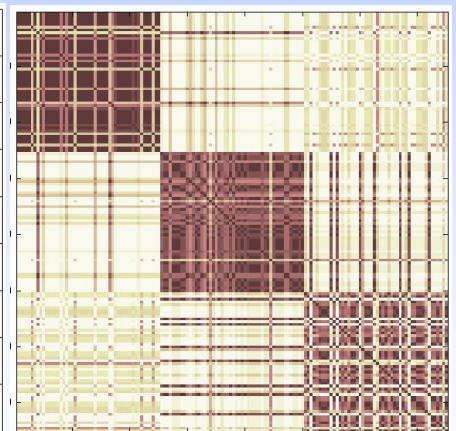
Original Data



Similarity Matrix

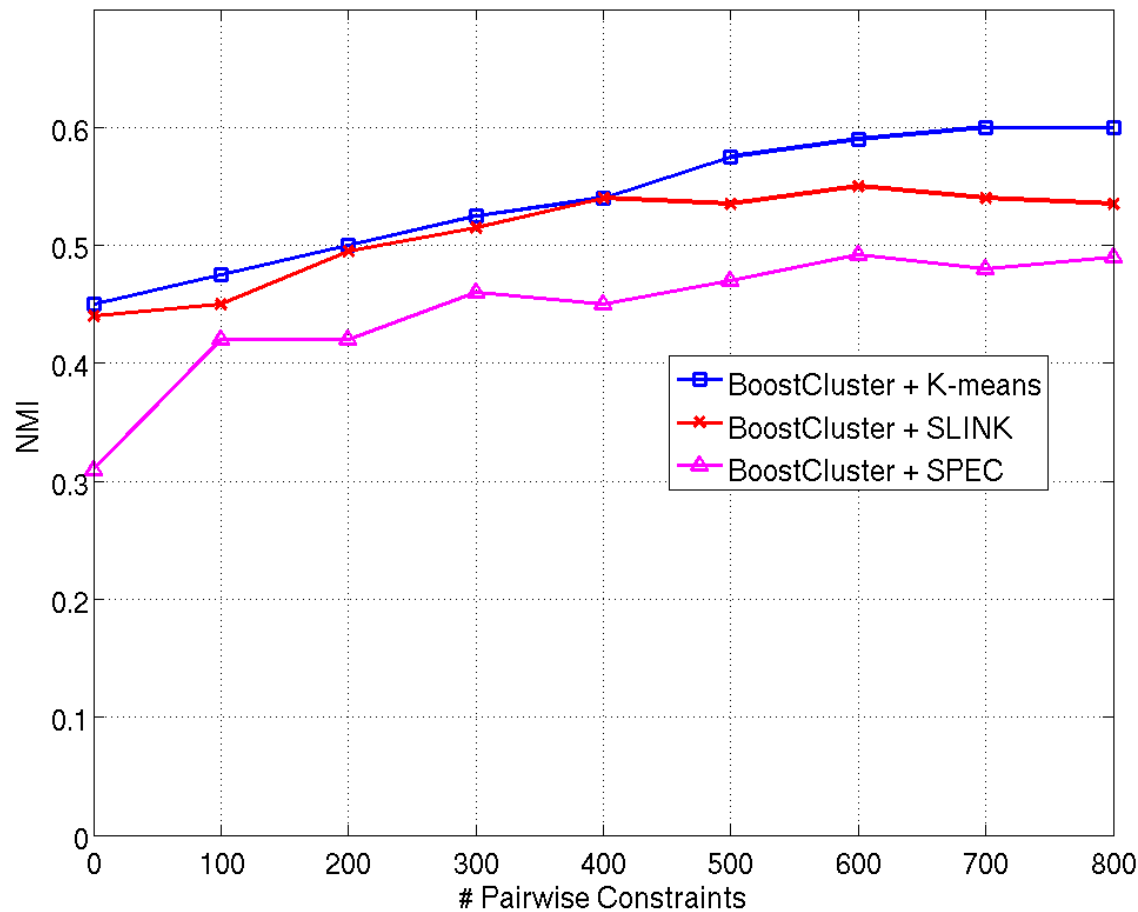


New representation



Similarity matrix

Performance of BoostCluster



Handwritten digit (UCI); 4,000 points in 256 dimensions; 10 clusters

Summary

- It is **natural** to seek clustering methods to group a heterogeneous set of objects based on similarity
- Objective should not be to choose the **best** clustering technique– it would be fruitless & contrary to the exploratory nature of clustering
- **Enough** clustering algorithms known to uncover specific data structures are available; **representation** is critical
- Future research: **rational basis** for comparing clustering methods, **quick-look** procedures for very large databases, taking **multiple looks** at the same data and incorporating **domain knowledge**