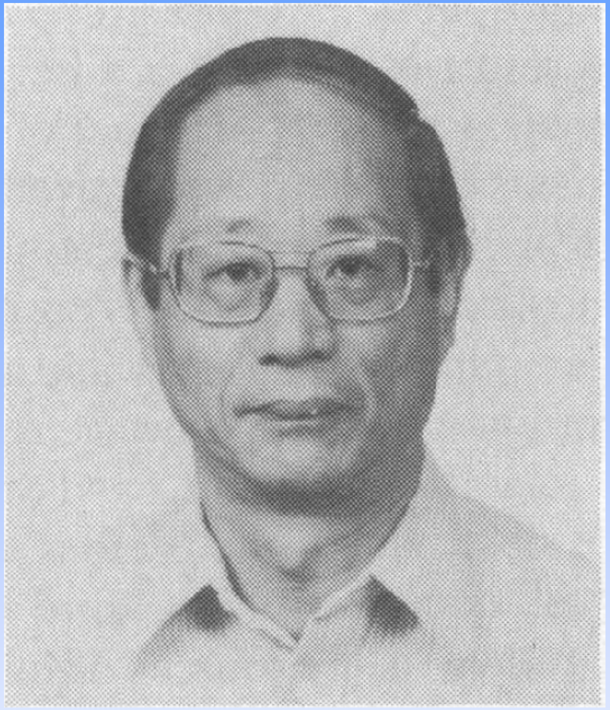
Data Clustering:
50 Years Beyond K-means

Anil K. Jain
Department of Computer Science
Michigan State University

King-Sun Fu



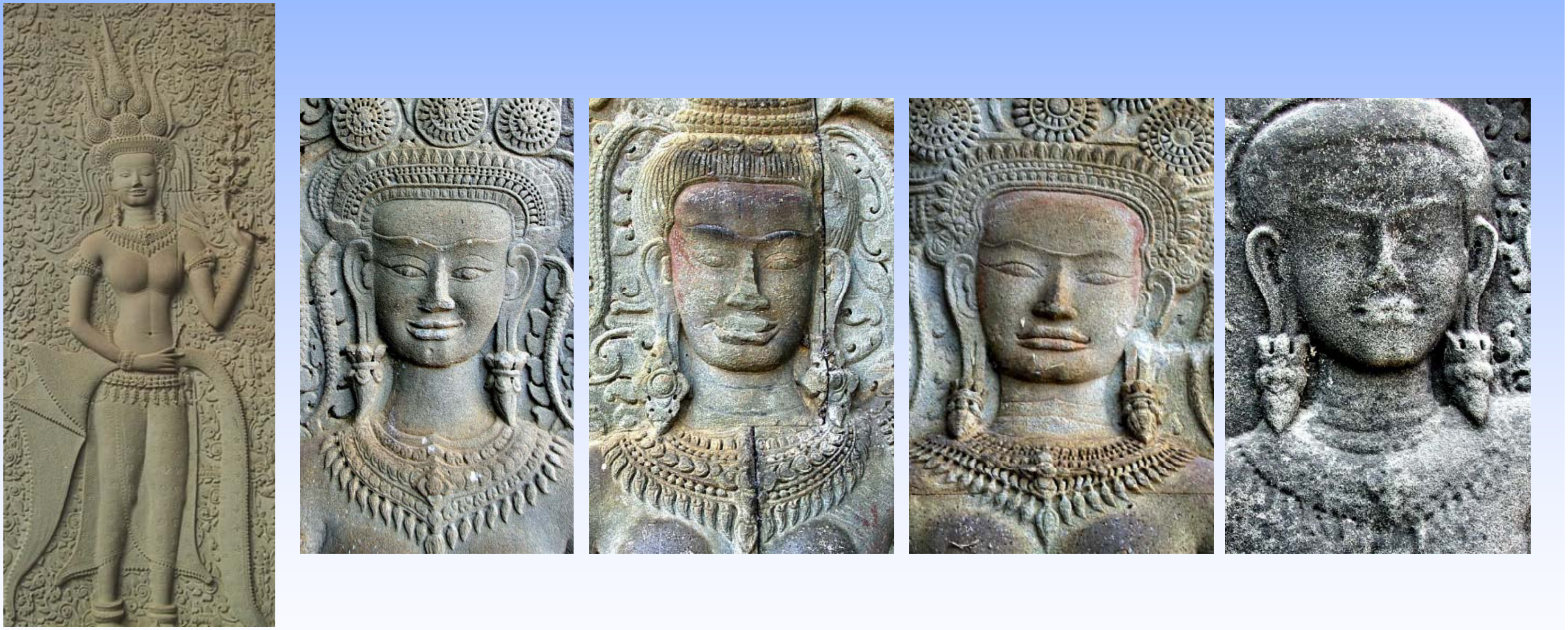
King-Sun Fu (1930-1985), a professor at Purdue was instrumental in the founding of IAPR, served as its first president, and is widely recognized for his extensive contributions to pattern recognition. *(Wikipedia)*

Angkor Wat, Siem Reap

Hindu temple built by a Khmer king ~1150 AD; Khmer kingdom declined in the 15th century; French explorers discovered the hidden ruins in 1860 (Angelina Jolie alias “Lora Croft” in *Tomb Raider* thriller)

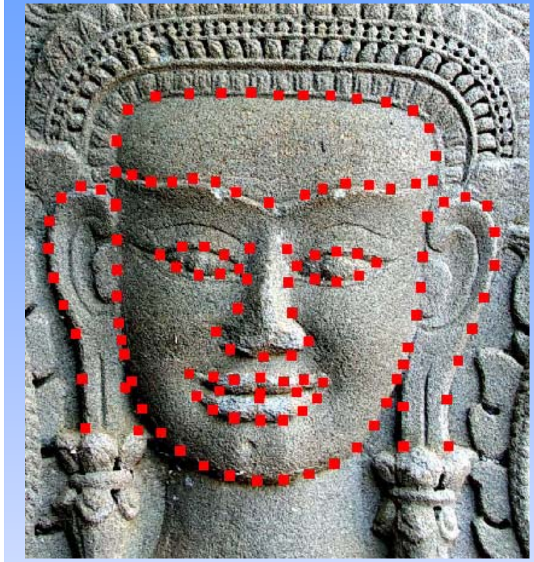
Apsaras of Angkor Wat

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

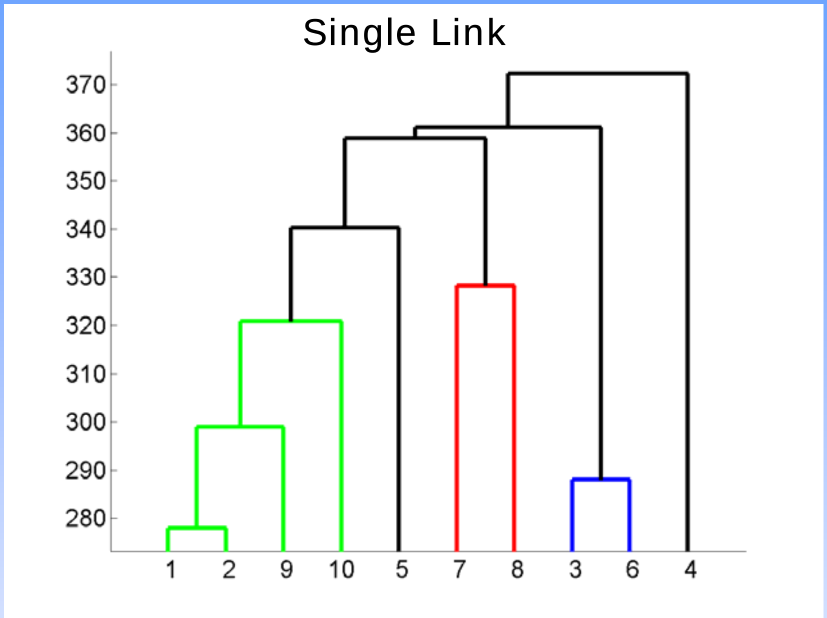


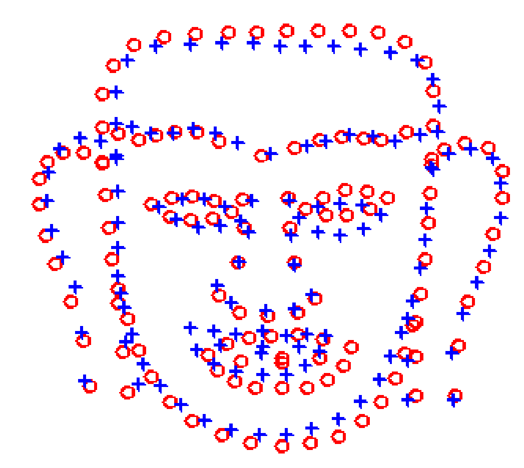
Kent Davis, Biometrics of the Godesess, DatAsia, Aug 2008
S. Marchal, Costumes et Parures Khmers: D'apres les devata D'Angkor-Vat, 1927

Clustering of Apsara Faces

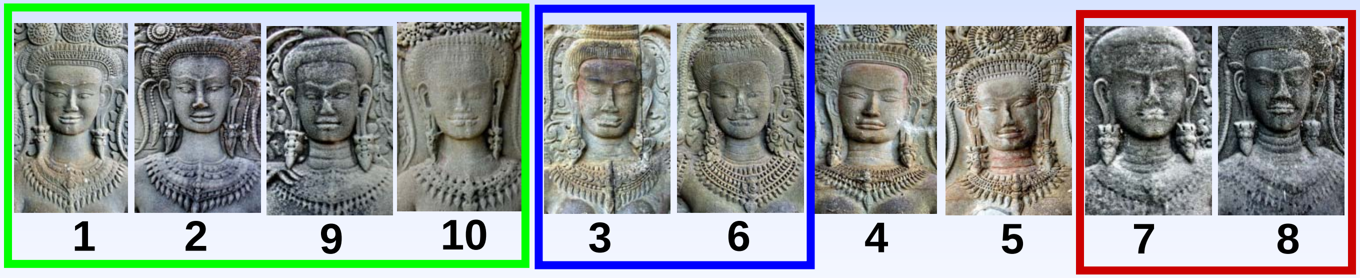


127 landmarks





Shape alignment



Single Link clusters

How do we validate the groups?

Ground Truth

Khmer Cultural Center

Data Explosion

- The digital universe was ~281 exabytes (281 billion gigabytes) in 2007; it would grow 10 times by 2011
- 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 categorization

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

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

Data Clustering

- Grouping of objects into meaningful categories
- Classification vs. clustering
- Unsupervised learning, exploratory data analysis, grouping, clumping, taxonomy, typology, Q-analysis
- Given a representation of n objects, find K clusters based on a measure of similarity
- Partitional vs. hierarchical

A. K. Jain and R. C. Dubes. Algorithms for Clustering Data, Prentice Hall, 1988. (available for download at: http://dataclustering.cse.msu.edu/)

Why Clustering?

- Natural classification: degree of similarity among forms (phylogenetic relationship or taxonomy)
- Data exploration: discover underlying structure, generate hypotheses, detect anomalies
- Compression: method for organizing data
- Applications: any scientific field that collects data! Astronomy, biology, marketing, engineering,…..

Google Scholar: ~1500 clustering papers in 2007 alone!

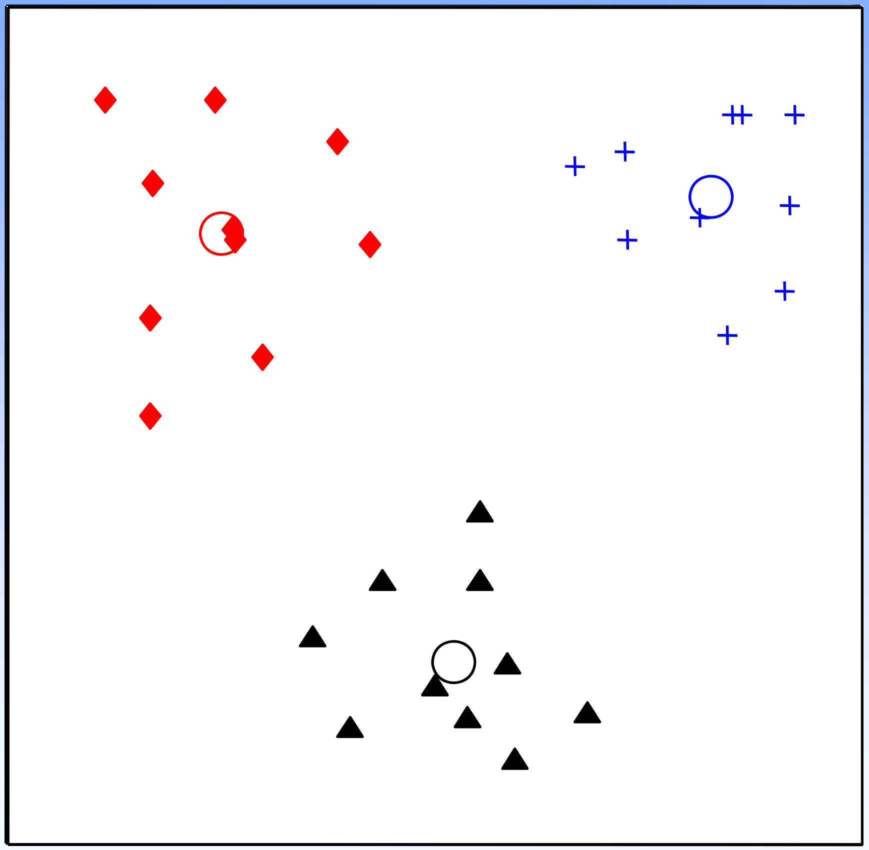
Historical Developments

- Cluster analysis first appeared in the title of a 1954 article analyzing anthropological data *(JSTOR)*
- Hierarchical Clustering: *Sneath (1957), Sorensen (1957)*
- K-Means: independently discovered *Steinhaus[1] (1956), Lloyd[2] (1957), Cox[3] (1957), Ball & Hall[4] (1967), MacQueen[5] (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)*

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

K-Means Algorithm

Minimize the squared error; Initialize K means; assign points to closest mean; update means; iterate



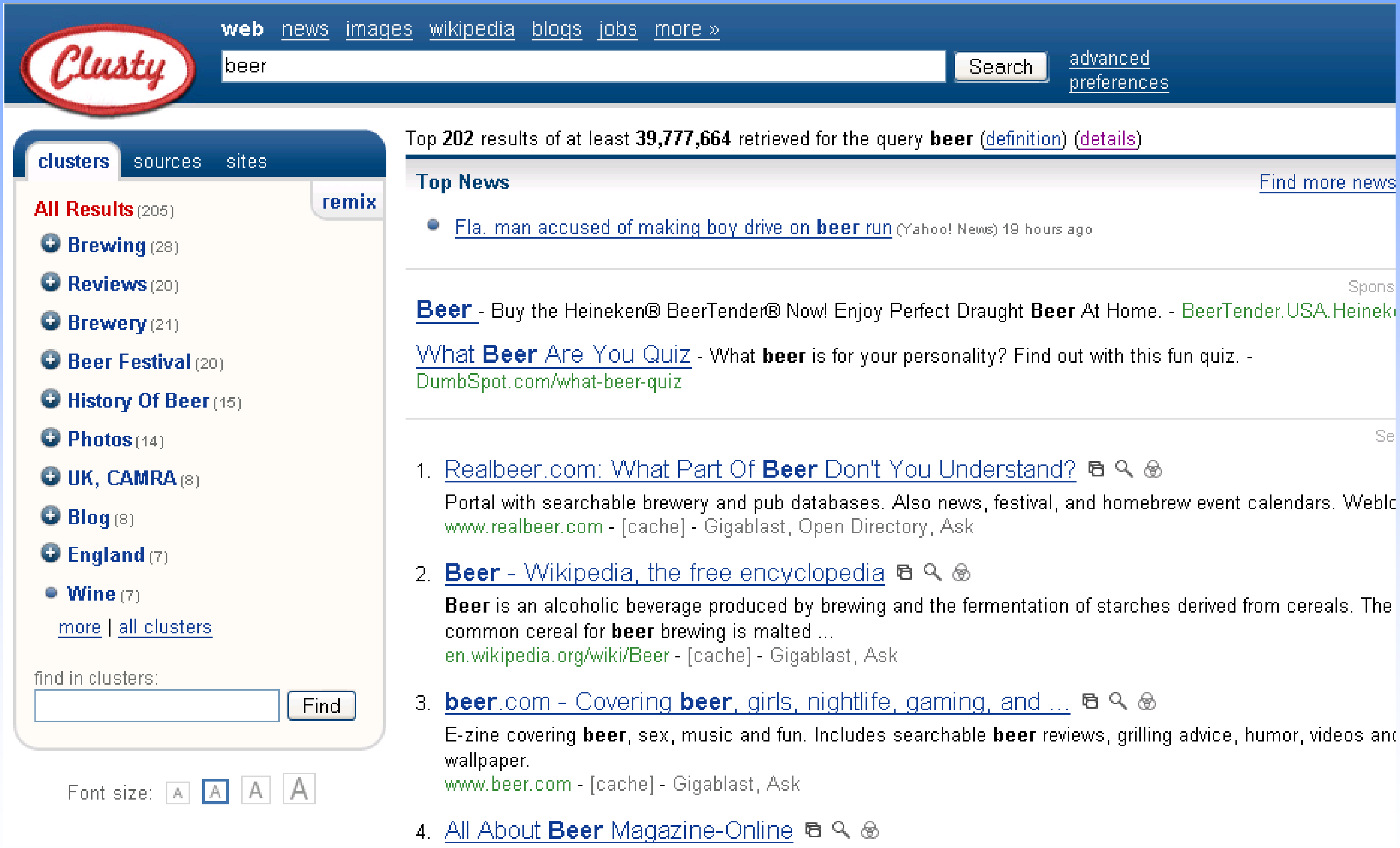
Bisecting K-means *(Karypis et al.)*; X-means *(Pelleg and Moore)*; Constrained K-means *(Davidson)*; Scalable K-means *(Bradley et al.)*

Beyond K-Means

- Developments in Data Mining and Machine Learning
  - Bayesian models, kernel methods, association rules (subspace clustering), graph mining, large scale clustering
- Choice of models, objective functions, and heuristics
- Density-based *(Ester et al., 1996)*
- Spectral *(Hagen & Kahng, 1991; Shi & Malik, 2000)*
- Information bottleneck *(Tishby et al., 1999)*
- Non-negative matrix factorization *(Lee & Seung, 1999)*
- Ensemble *(Fred & Jain, 2002; Strehl & Ghosh, 2002)*
- Semi-supervised *(Wagstaff et al., 2003; Basu et al., 2004)*

Structure Discovery

Cluster web retrieved documents



All Results (205)
- Brewing (28)
- Reviews (20)
- Brewery (21)
- Beer Festival (20)
- History Of Beer (15)
- Photos (14)
- UK, CAMRA (8)
- Blog (8)
- England (7)
- Wine (7)

more | all clusters

find in clusters:

Top **202** results of at least **39,777,664** retrieved for the query **beer** (definition) (details)

Top News Find more news

- Fla. man accused of making boy drive on **beer** run (Yahoo! News) 19 hours ago

Spons

Beer - Buy the Heineken® BeerTender® Now! Enjoy Perfect Draught **Beer** At Home. - BeerTender.USA.Heinek

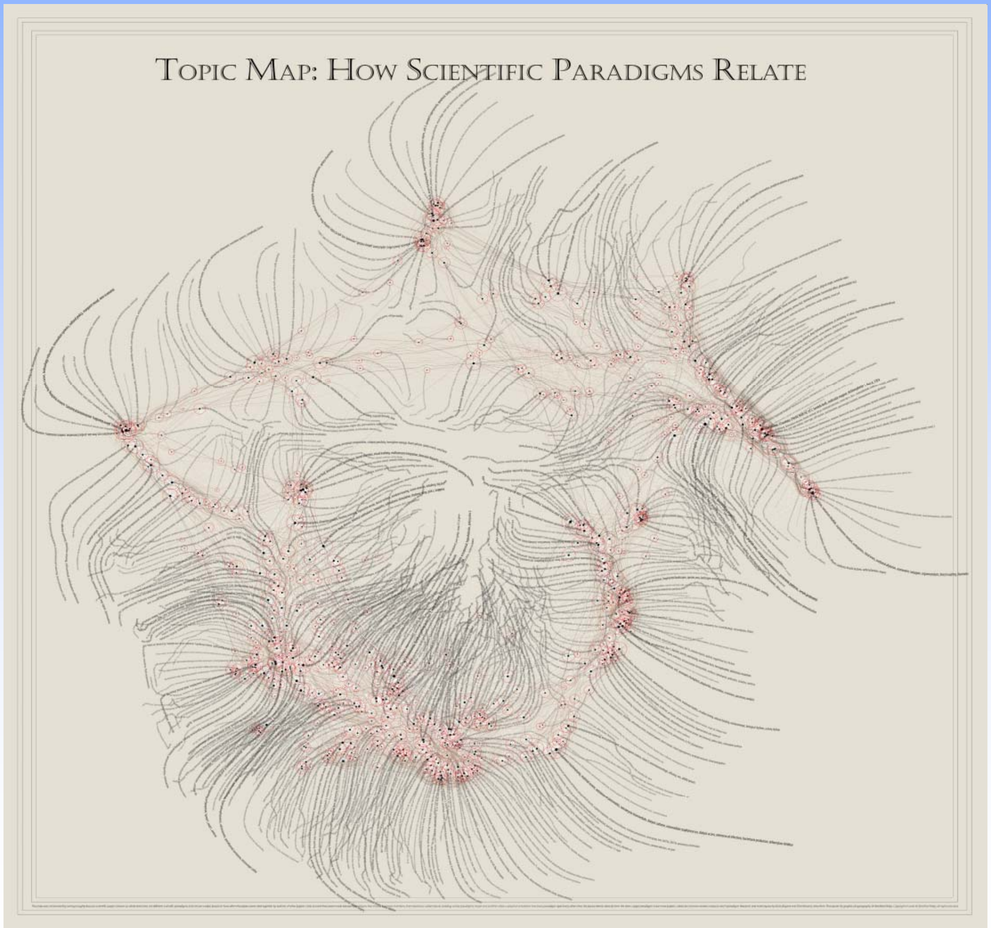
What **Beer** Are You Quiz - What **beer** is for your personality? Find out with this fun quiz. - DumbSpot.com/what-beer-quiz

Se

1. Realbeer.com: What Part Of **Beer** Don't You Understand?
   Portal with searchable brewery and pub databases. Also news, festival, and homebrew event calendars. Weblo
   www.realbeer.com - [cache] - Gigablast, Open Directory, Ask
2. **Beer** - Wikipedia, the free encyclopedia
   Beer is an alcoholic beverage produced by brewing and the fermentation of starches derived from cereals. The common cereal for **beer** brewing is malted ...
   en.wikipedia.org/wiki/Beer - [cache] - Gigablast, Ask
3. **beer**.com - Covering **beer**, girls, nightlife, gaming, and ...
   E-zine covering **beer**, sex, music and fun. Includes searchable **beer** reviews, grilling advice, humor, videos and wallpaper.
   www.beer.com - [cache] - Gigablast, Ask
4. All About **Beer** Magazine-Online

Topic Discovery

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



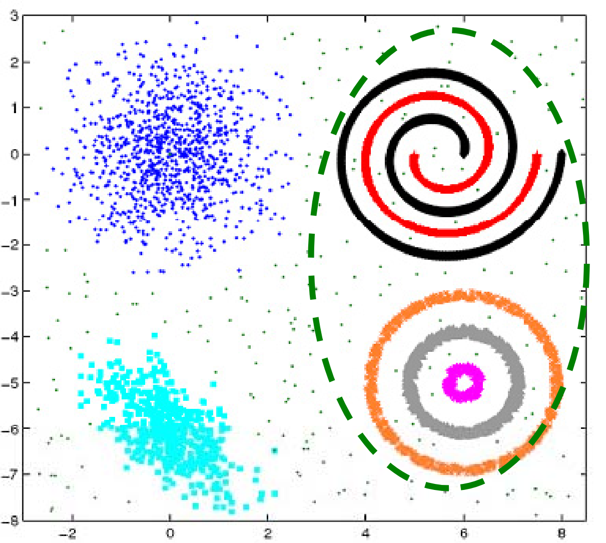
Map of Science, *Nature (2006)*

User’s Dilemma!

- What is a cluster?
- Which features and normalization scheme?
- How to define pair-wise similarity?
- How many clusters?
- Which clustering method?
- Does the data have any clustering tendency?
- Are the discovered clusters & partition valid?

R. Dubes and A.K. Jain, Clustering Techniques: User’s Dilemma, *Pattern Recognition*, 1976

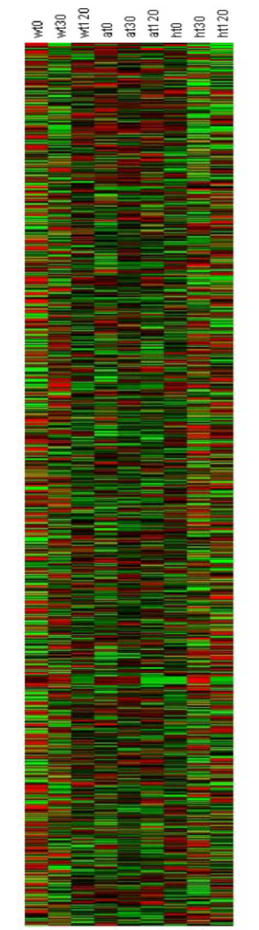
Cluster



- A set of similar entities; entities in different clusters are not alike
- How do we define similarity?
- Compact clusters
  - within-cluster distance < between-cluster distance
- Connected clusters
  - within-cluster connectivity > between-cluster connectivity
- Ideal cluster: compact and isolated

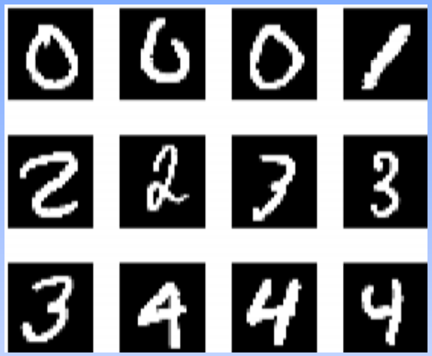
Representation

No universal representation; domain dependent

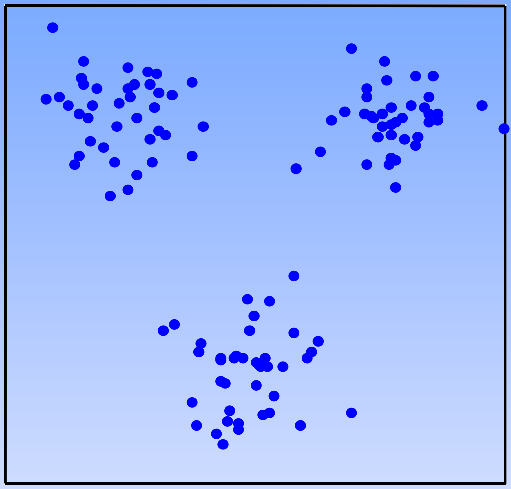


Gene Expressions

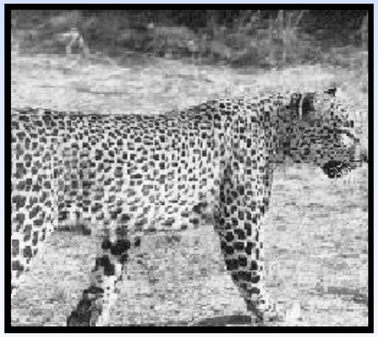
Image retrieval



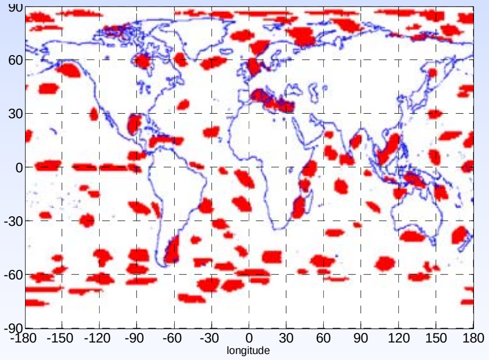
Handwritten digits



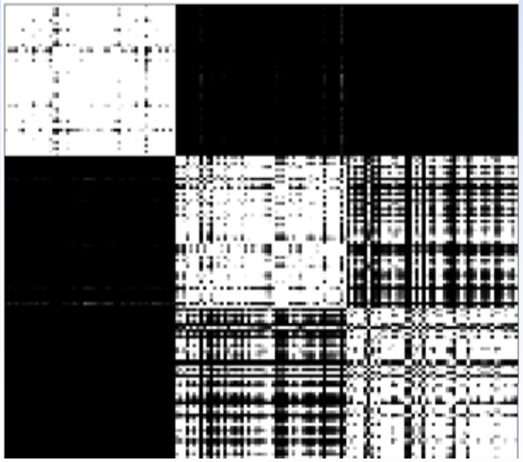
nxd pattern matrix



Segmentation



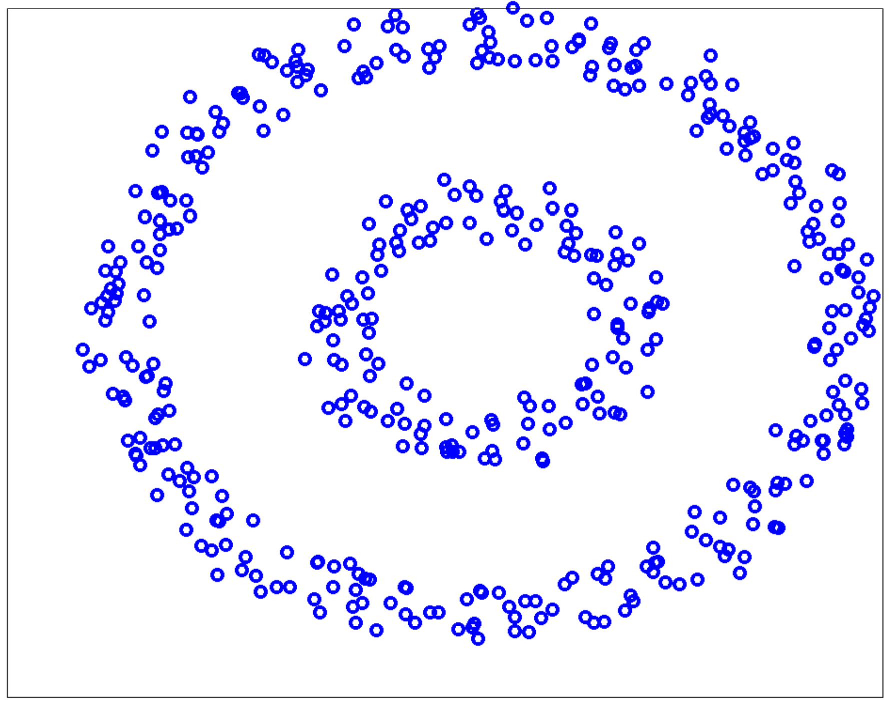
Time series (sea-surface temp)



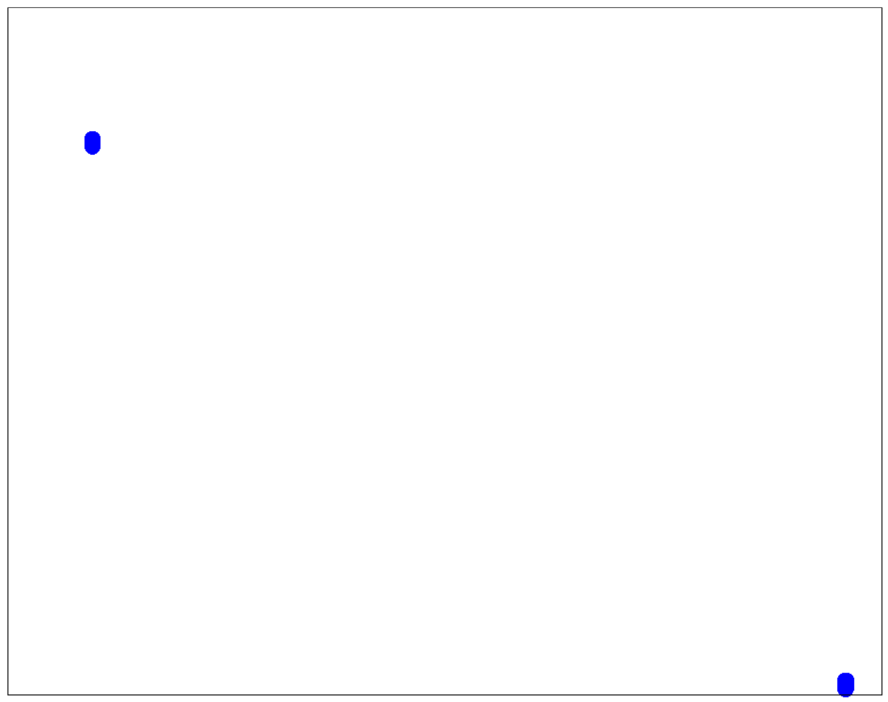
nxn similarity matrix

Good Representation

Good representation => compact & isolated clusters



Points in given 2D space



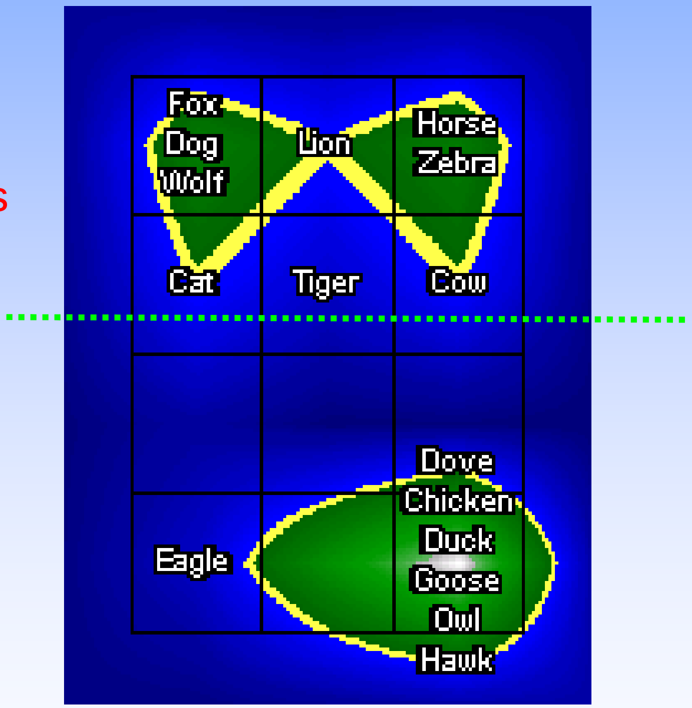
Eigenvectors of RBF kernel

Feature Weighting

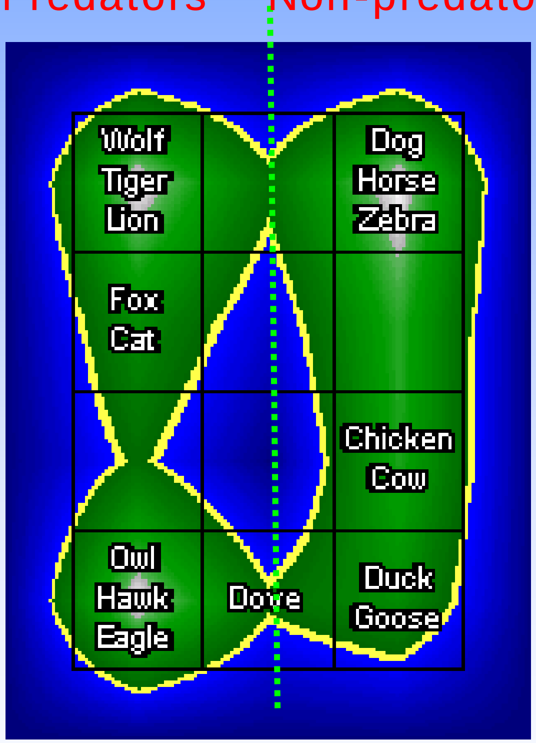
Two different meaningful groupings of 16 animals based on 13 Boolean features (appearance & activity)

Mammals

Birds



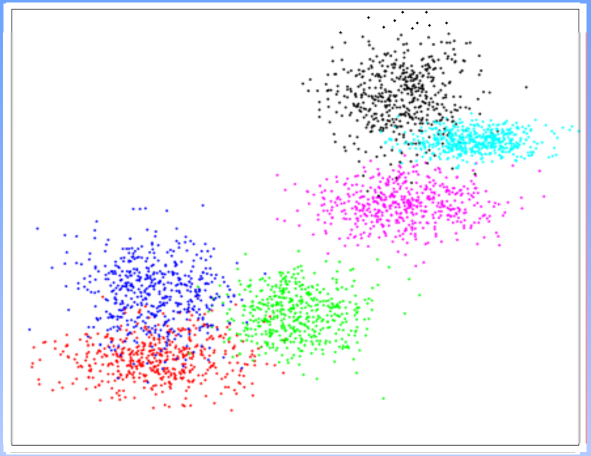
Predators Non-predators



Large weight on appearance features Large weight on activity features

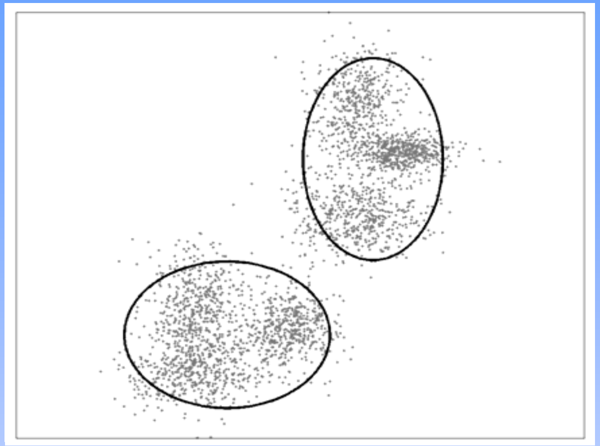
http://www.ofai.at/~elias.pampalk/kdd03/animals/

Number of Clusters

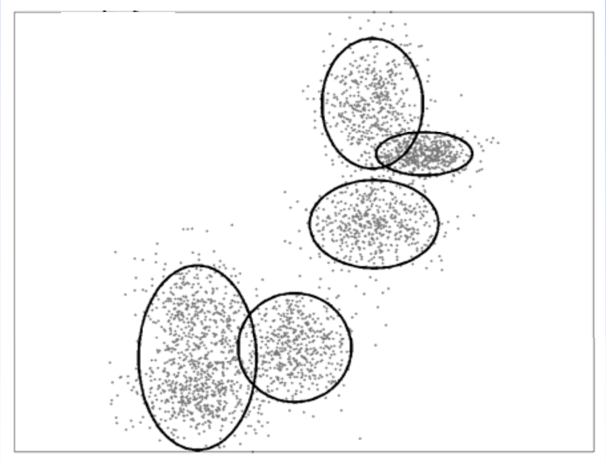


True labels, K = 6

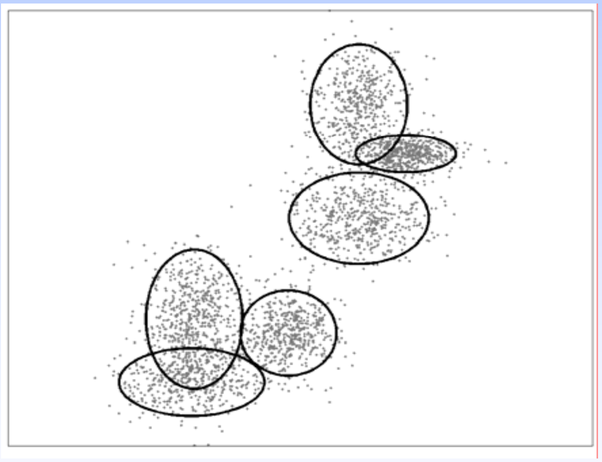
Input data



GMM (K=2)



GMM (K=5)

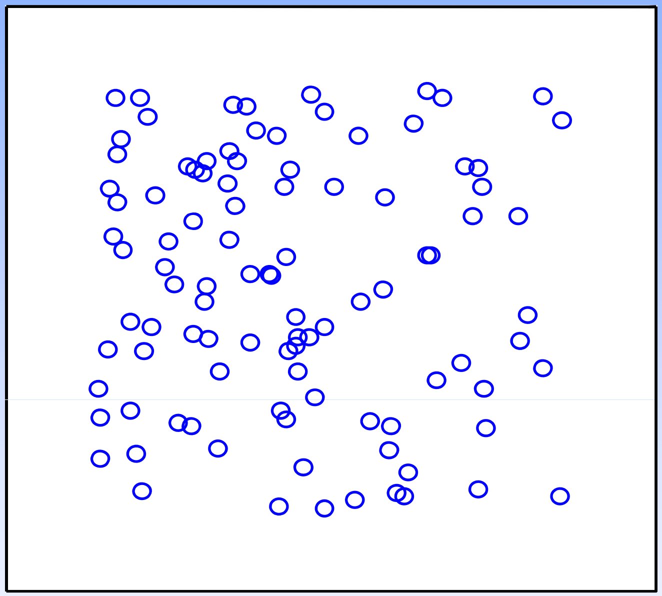


GMM (K=6)

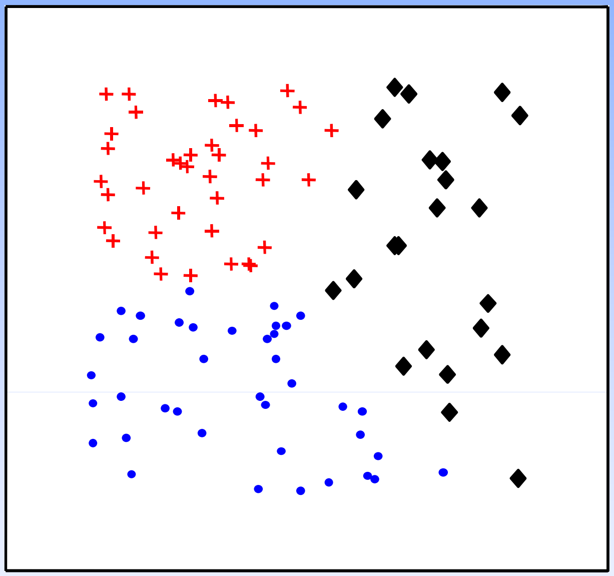
M. Figueiredo and A.K. Jain, Unsupervised Learning of Finite Mixture Models, *IEEE PAMI*, 2002

Cluster Validity

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



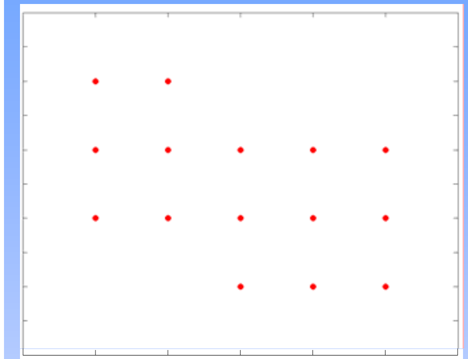
100 2D uniform data points



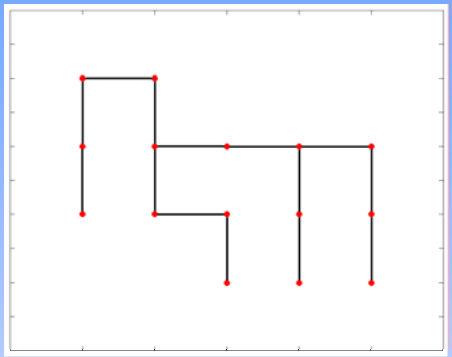
K-Means; K=3

- Easy to design new methods, difficult to validate
- Cluster stability *(Jain & Moreau, 1989; Lange et. al, 2004)*

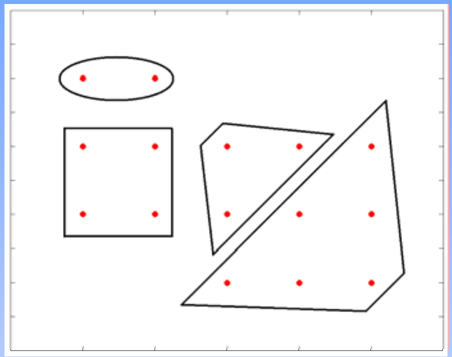
Comparing Clustering Algorithms



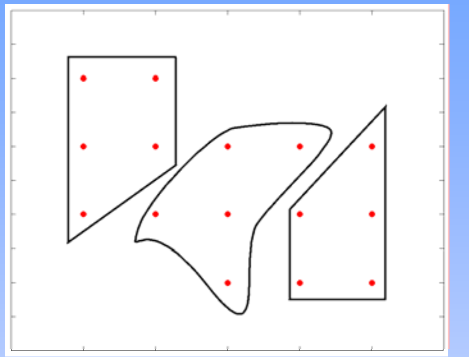
15 points in 2D



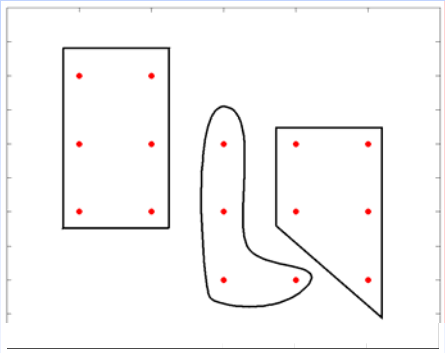
MST



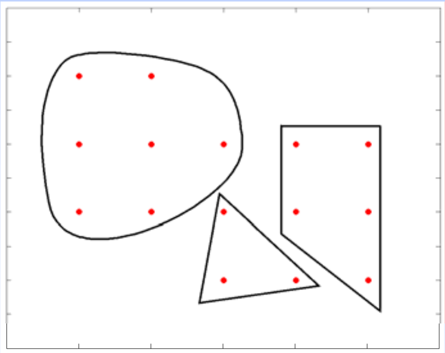
FORGY



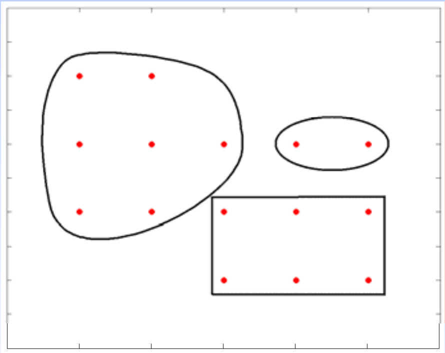
ISODATA



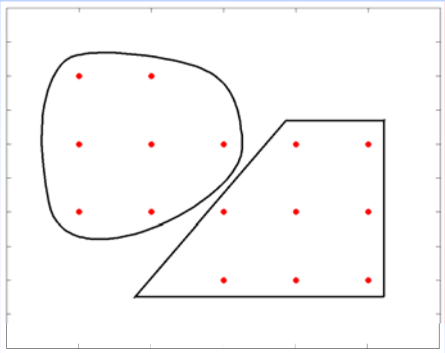
WISH



CLUSTER



Complete-link



JP

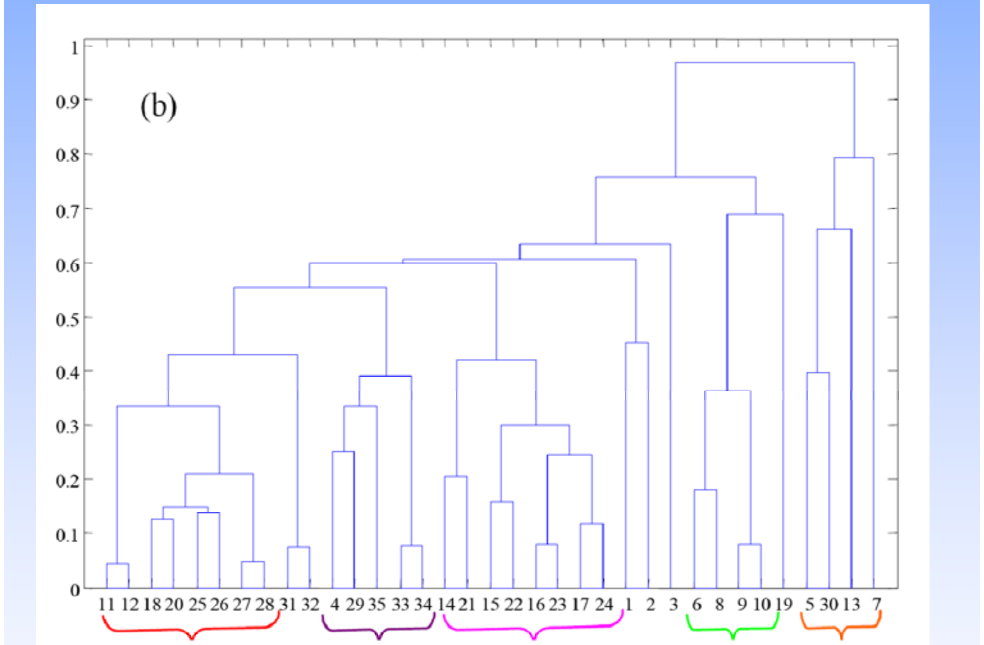
FORGY, ISODATA, WISH, CLUSTER are all MSE algorithms

R. Dubes and A.K. Jain, Clustering Techniques: User's Dilemma, *Pattern Recognition*, 1976

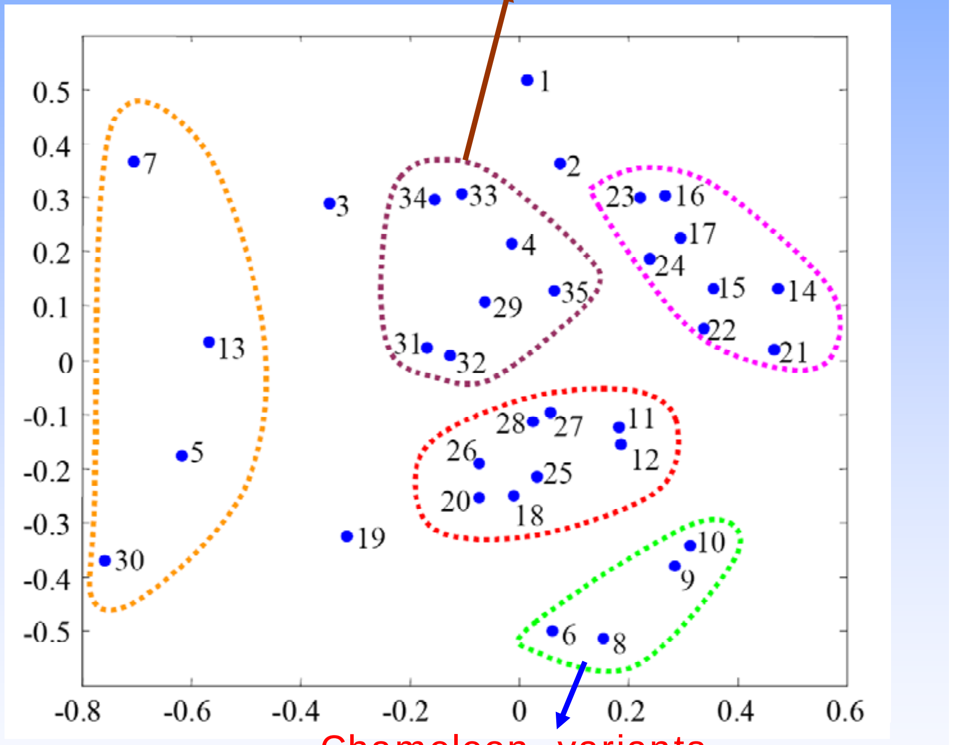
Grouping of Clustering Algorithms

Clustering method vs. clustering algorithm

K-means, Spectral, GMM, Ward’s linkage



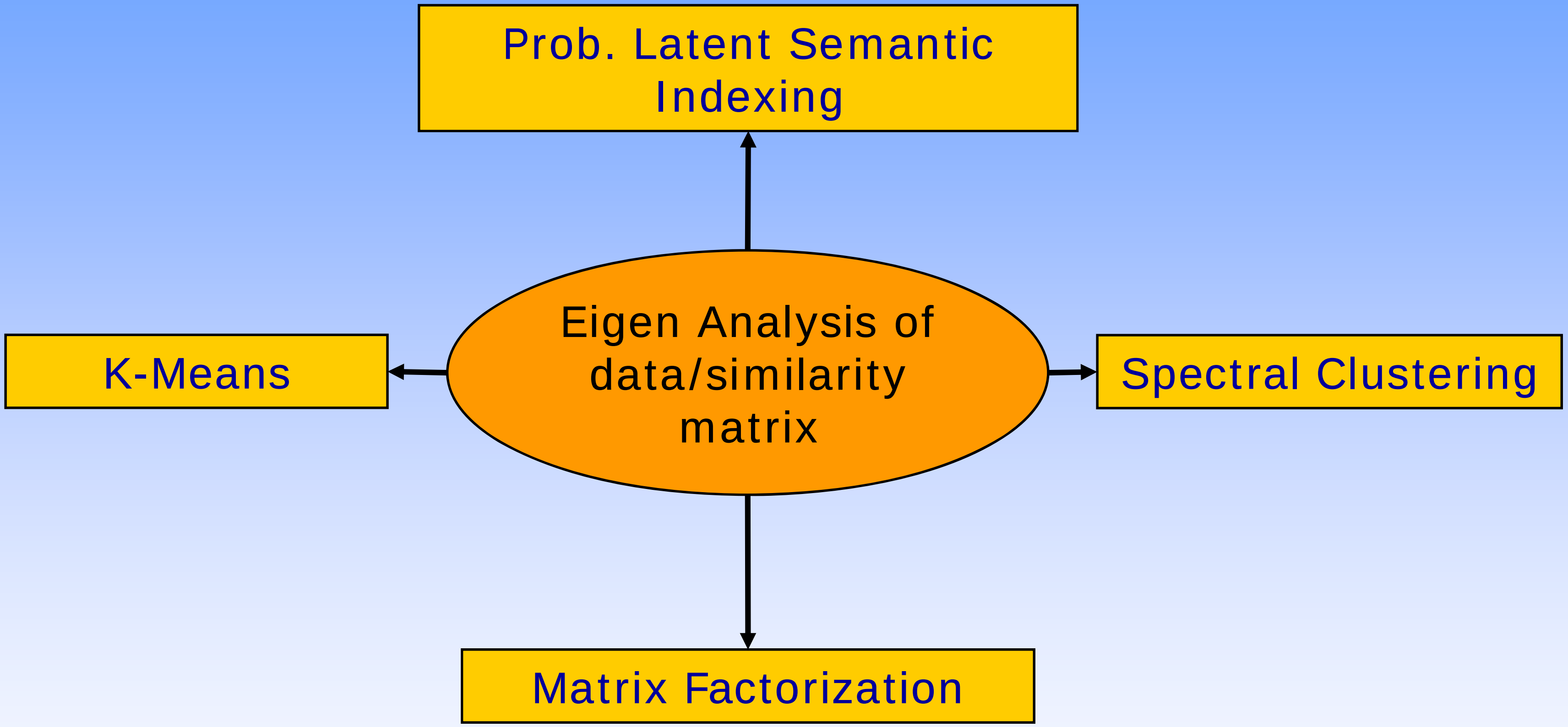
Hierarchical clustering of 35 different algorithms



Chameleon variants

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

Mathematical & Statistical Links


Prob. Latent Semantic Indexing
K-Means
Eigen Analysis of data/similarity matrix
Spectral Clustering
Matrix Factorization

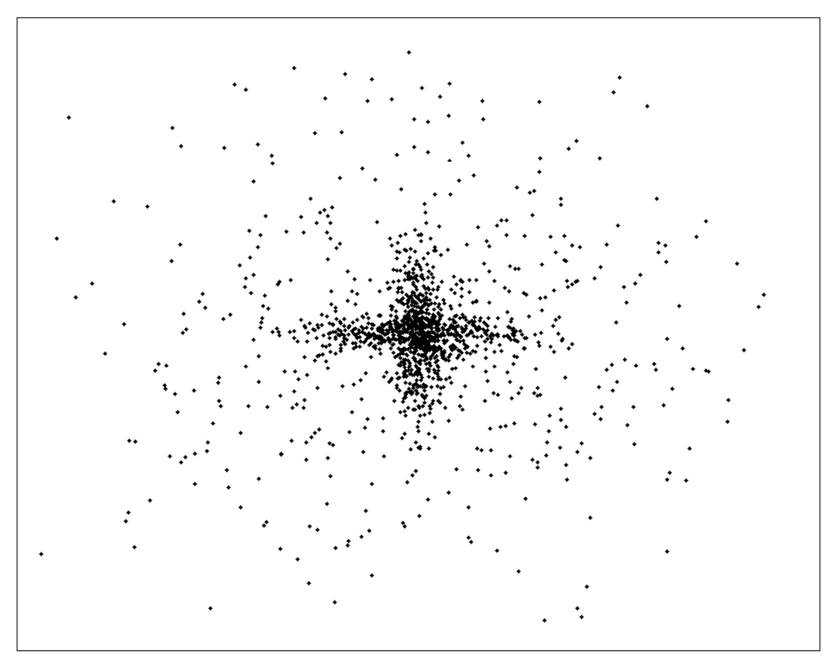
Zha et al., 2001; Dhillon et al., 2004; Gaussier et al., 2005, Ding et al., 2006; Ding et al., 2008

Admissibility Criteria

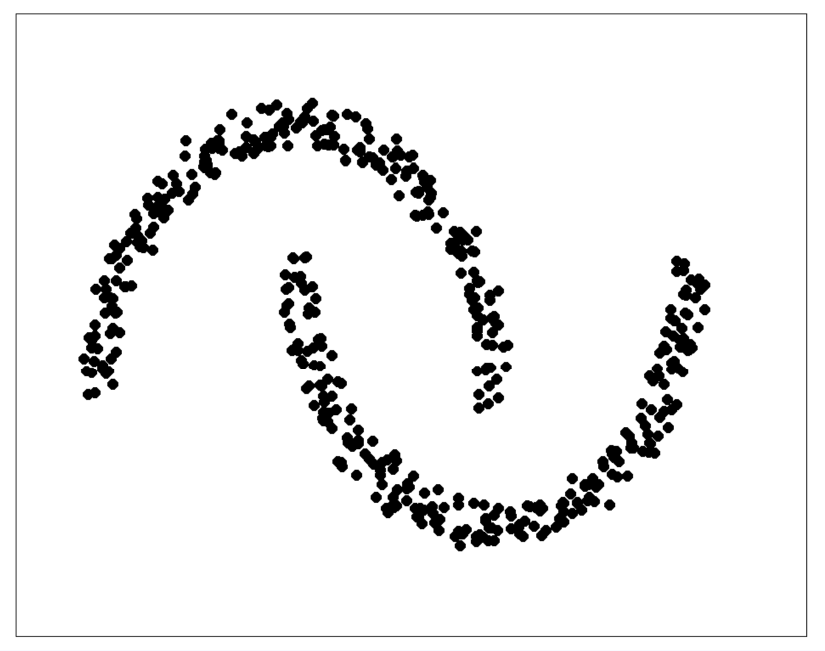
- A technique is P-admissible if it satisfies a desirable property P (*Fisher & Van Ness, Biometrika, 1971*)
- Properties that test sensitivity w.r.t. changes that do not alter the essential structure of data: point & cluster proportion, cluster omission, monotone
- Could be used to eliminate obviously bad methods
- Impossibility theorem (*Kleinberg, NIPS 2002*); no clustering function satisfies all three properties: scale invariance, richness and consistency

No Best Clustering algorithm!

- Each algorithm imposes a structure on data
- Good fit between model & data => success



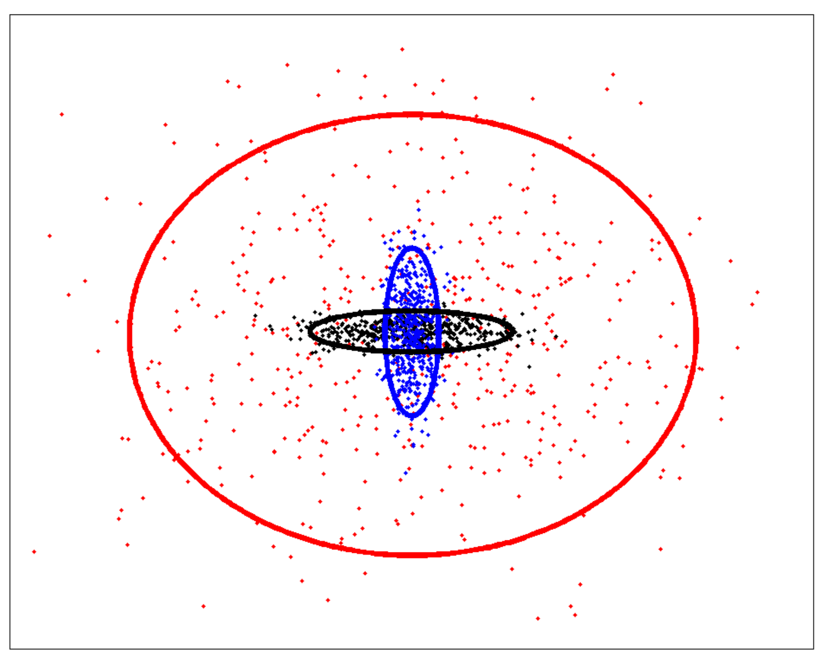
Mixture of 3 Gaussians



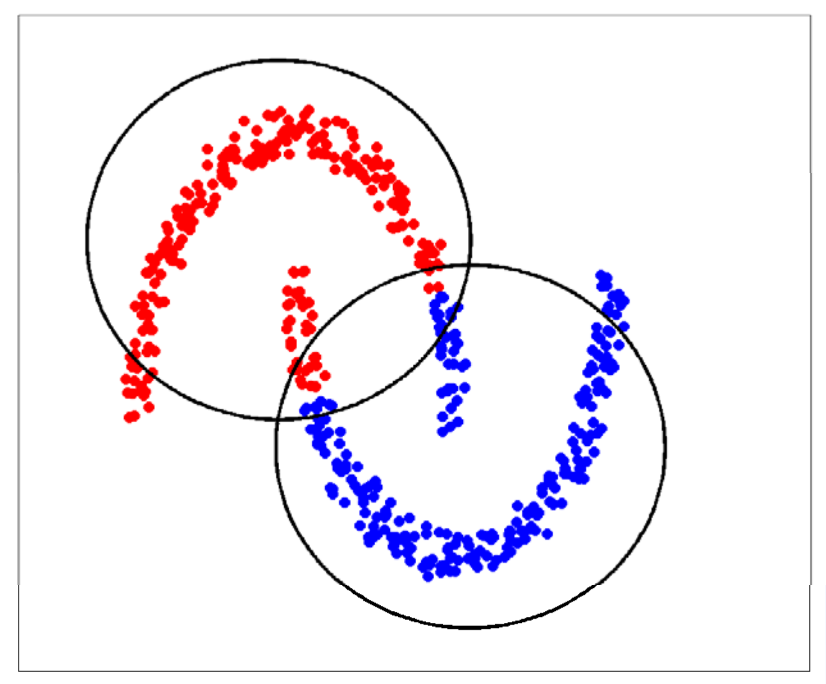
Two “half rings”

No Best Clustering algorithm!

- Each algorithm imposes a structure on data
- Good fit between model & data => success



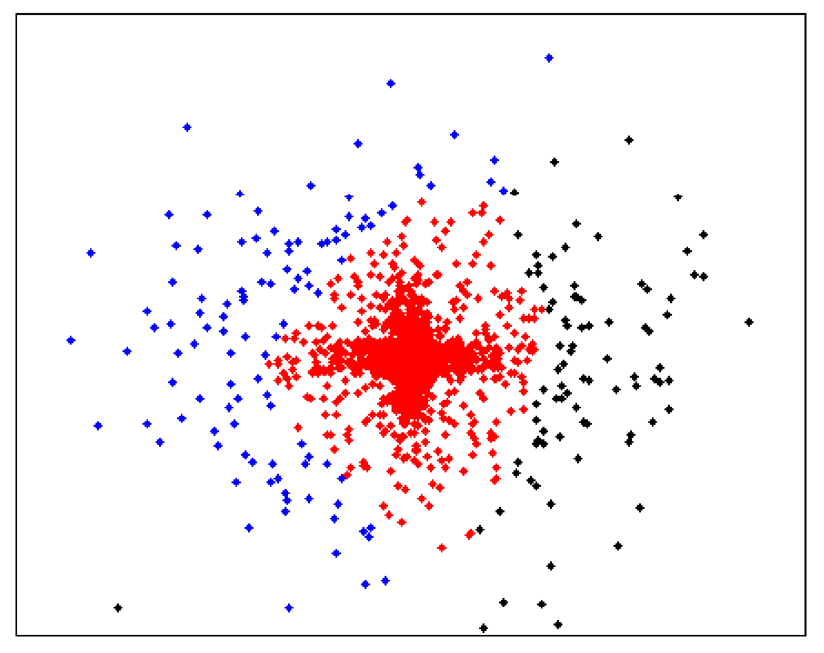
GMM; K = 3



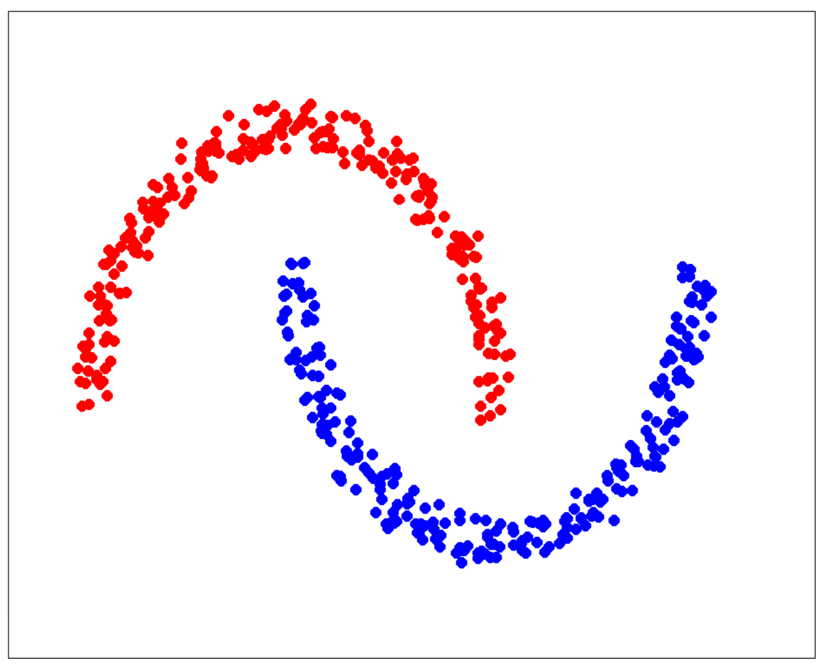
GMM; K = 2

No Best Clustering algorithm!

- Each algorithm imposes a structure on data
- Good fit between model & data => success



Spectral; K = 3



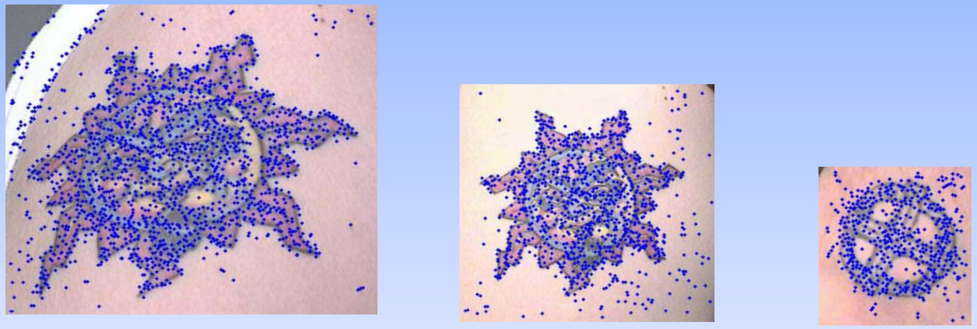
Spectral; K = 2

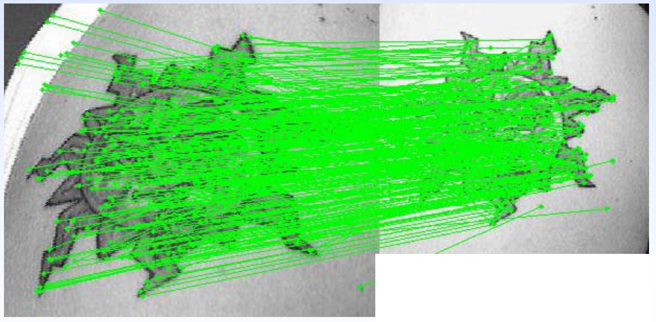
Some Trends

- Large-scale data
  - Clustering of 1.5B images into 50M clusters *(Liu et al., WACV 2007)*
- Evidence Accumulation
  - Combining multiple partitions (different algorithms, parameters, representations)
- Domain Knowledge
  - Pair-wise constraints, feature constraints (e.g., WordNet)
- Multi-way clustering
  - Simultaneously cluster documents, words and authors
- Complex Data Types
  - Dynamically evolving data (data streams)
  - Networks/graphs/tree (similarity matrix for structured data?)

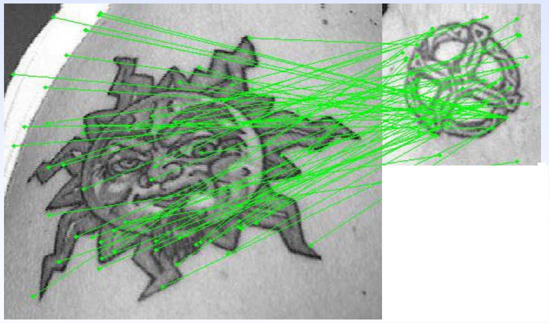
Content-based Image Retrieval

- Given a query image, retrieve visually similar images
- Key-point based CBIR: Image similarity based on the number of matching SIFT key points; ~1000 key points/image





370 matching points



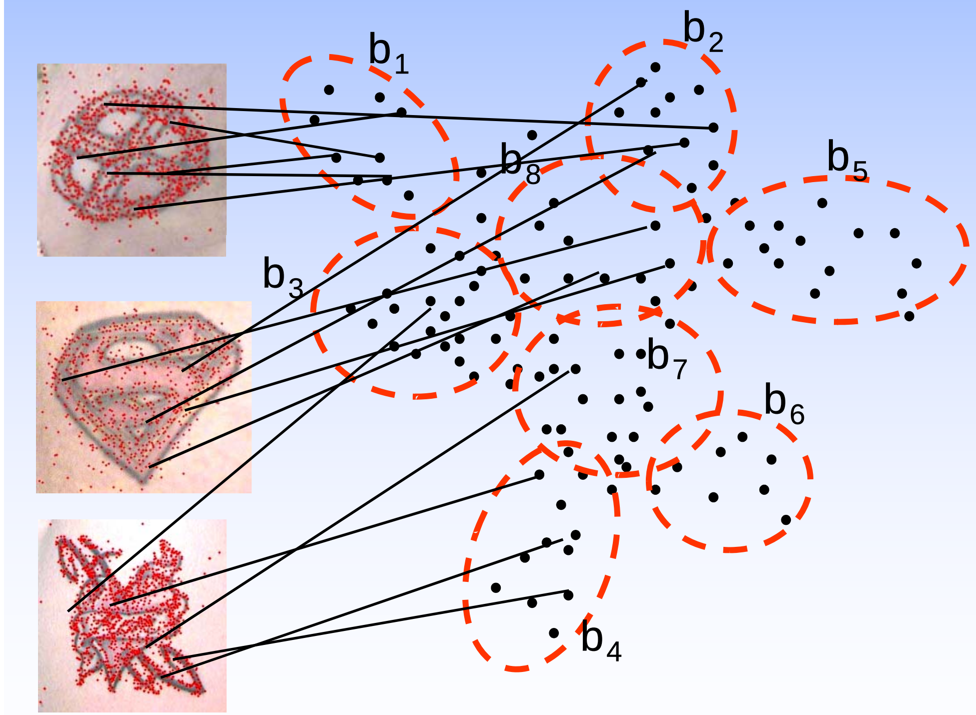
64 matching points

Large Image Database: Challenges

- A database with 10 million images
- Matching between two images ~ 10 msec.
- Linear scanning: 30 hours to answer one query!
- Text retrieval is much more efficient
  - – 0.1 sec. to search 10 billion docs in Google
- Solution: convert CBIR to text retrieval problem *(Sivic & Zisserman, ICCV 2003)*

Text Retrieval for CBIR

- Key points → visual words
  - Group key points from all the images into a number of clusters
  - Each cluster is a visual word
- Bag-of-words representation for images



Bag of key points

Visual word	b_1	b_2	b_3	b_4	b_5	...
	5	2	0	0	0	…
	0	1	3	0	0	…
	0	0	1	4	0	…
…	…	…	…	…	…	…

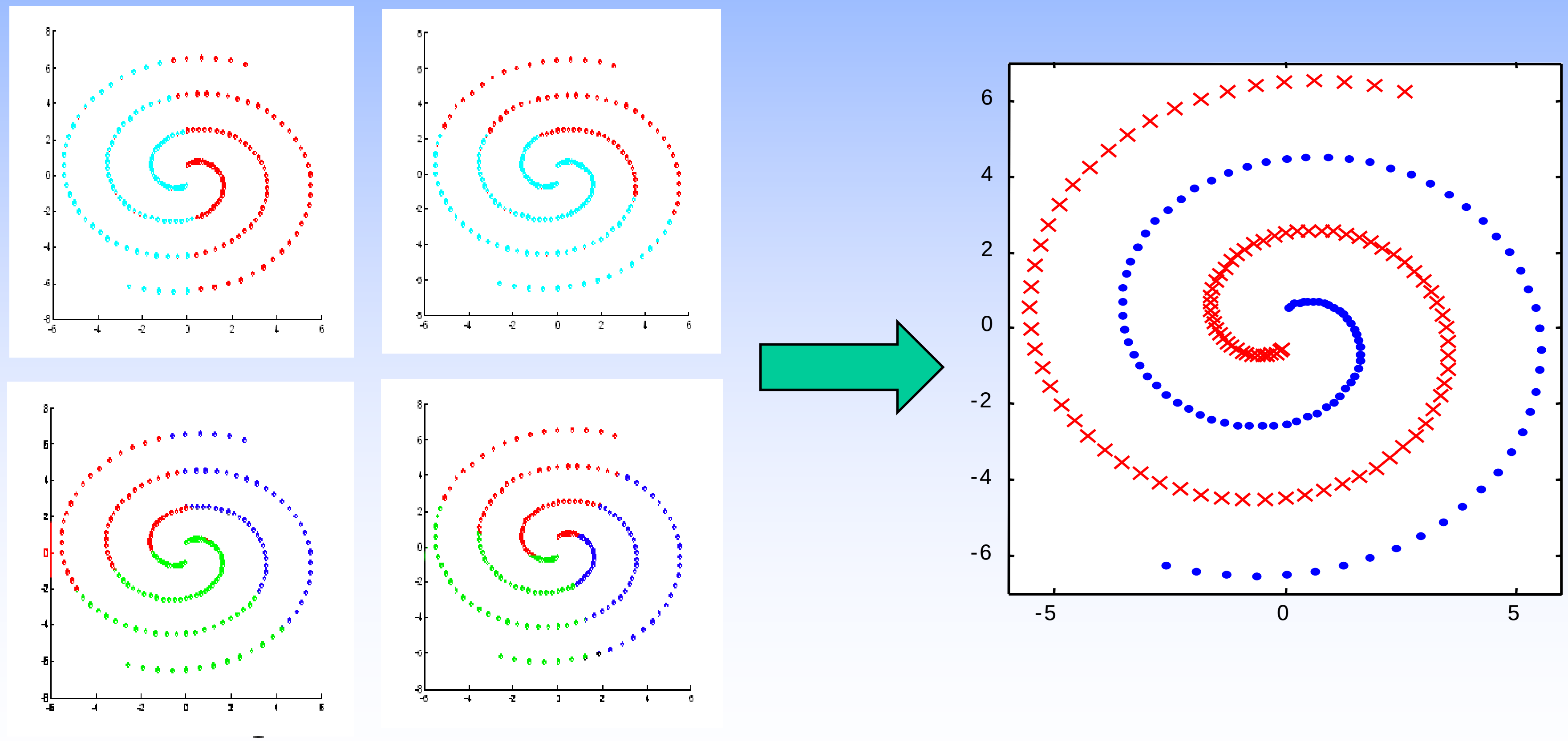
Bag of words

Large-scale Clustering

- Challenges in clustering key points
  - Very large number of key points: 10 million images x 1000 key points → 10 billion key points!
  - Very large number of clusters: 100K ~ 1 million clusters
  - Requires efficient clustering algorithms
- Efficient K-means clustering
  - Find the closest cluster center efficiently
  - Large no. of key points by KD-tree *(Moore, NIPS 1998)*
  - Large no. of clusters by KD-tree *(Philbin et al., CVPR 2007)*

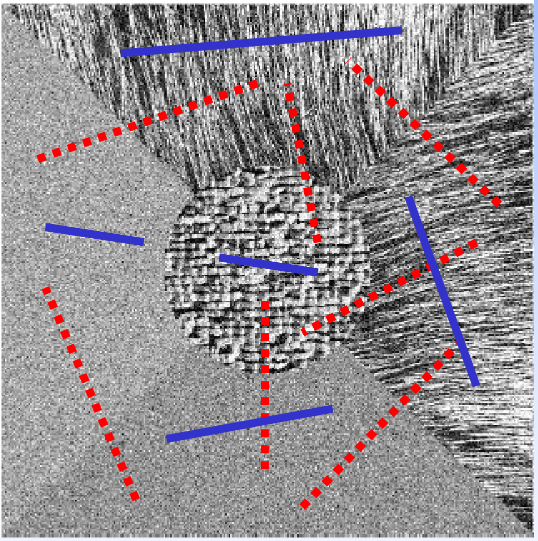
Clustering Ensemble

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



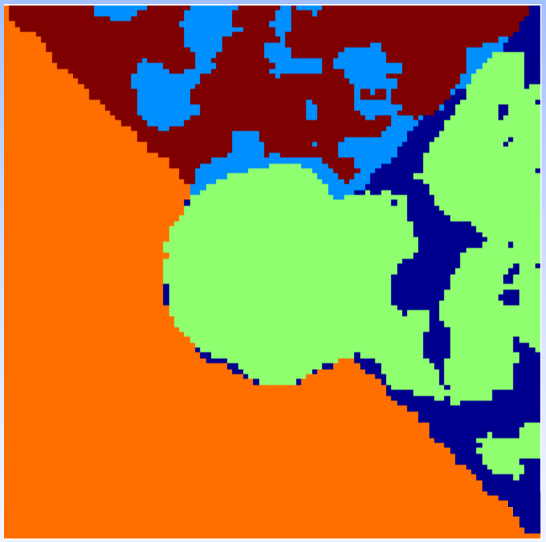
Semi-supervised Clustering

- Improve the partition given domain knowledge
- Side information: pair-wise constraints

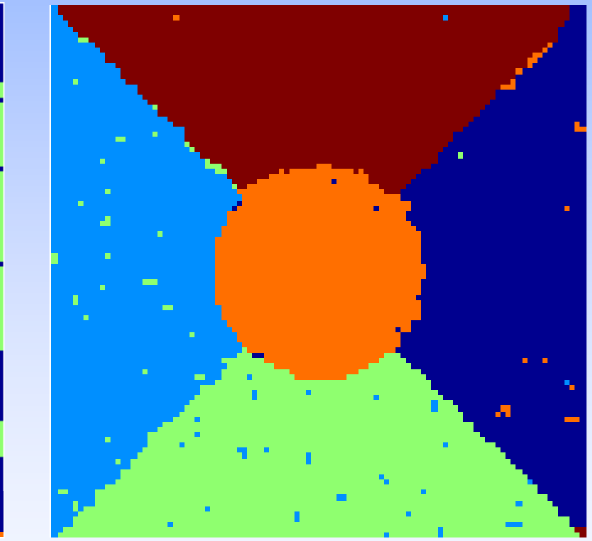


Input Image & constraints

▪▪▪▪ Must-not link ▬ Must link



No constraints

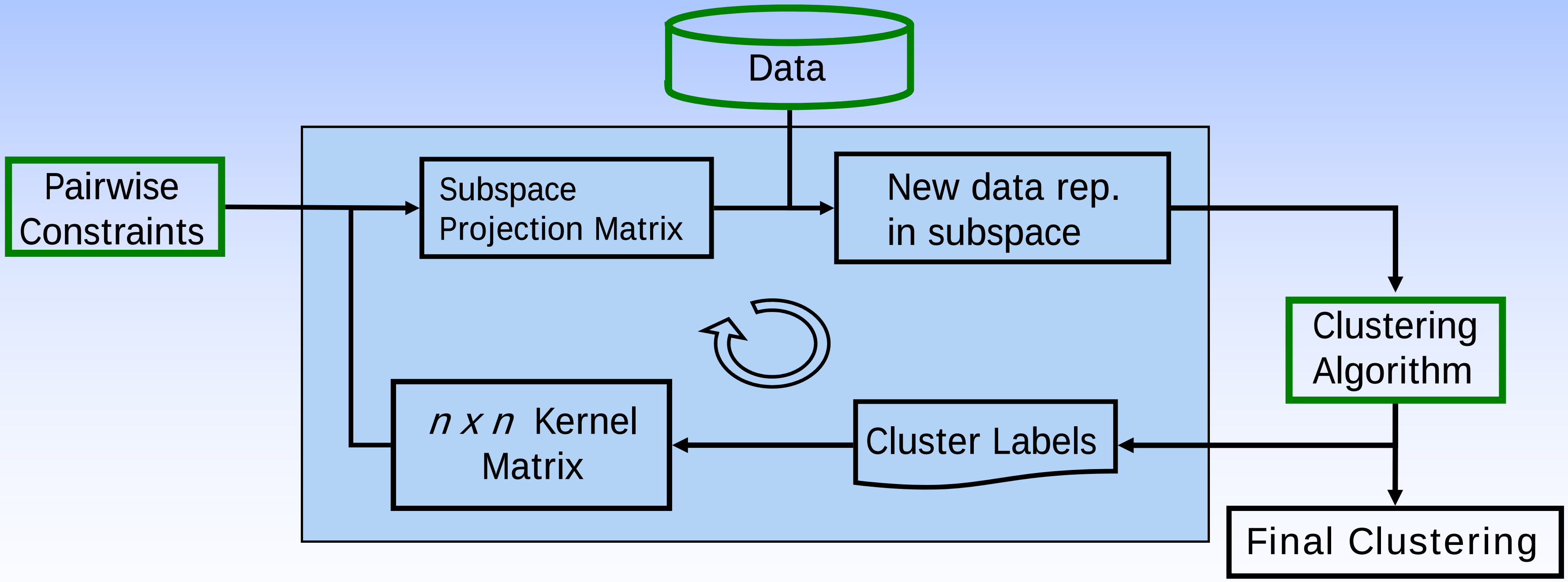


10% pixels in constraints

Lange, Law, Jain & Buhman, CVPR, 2005

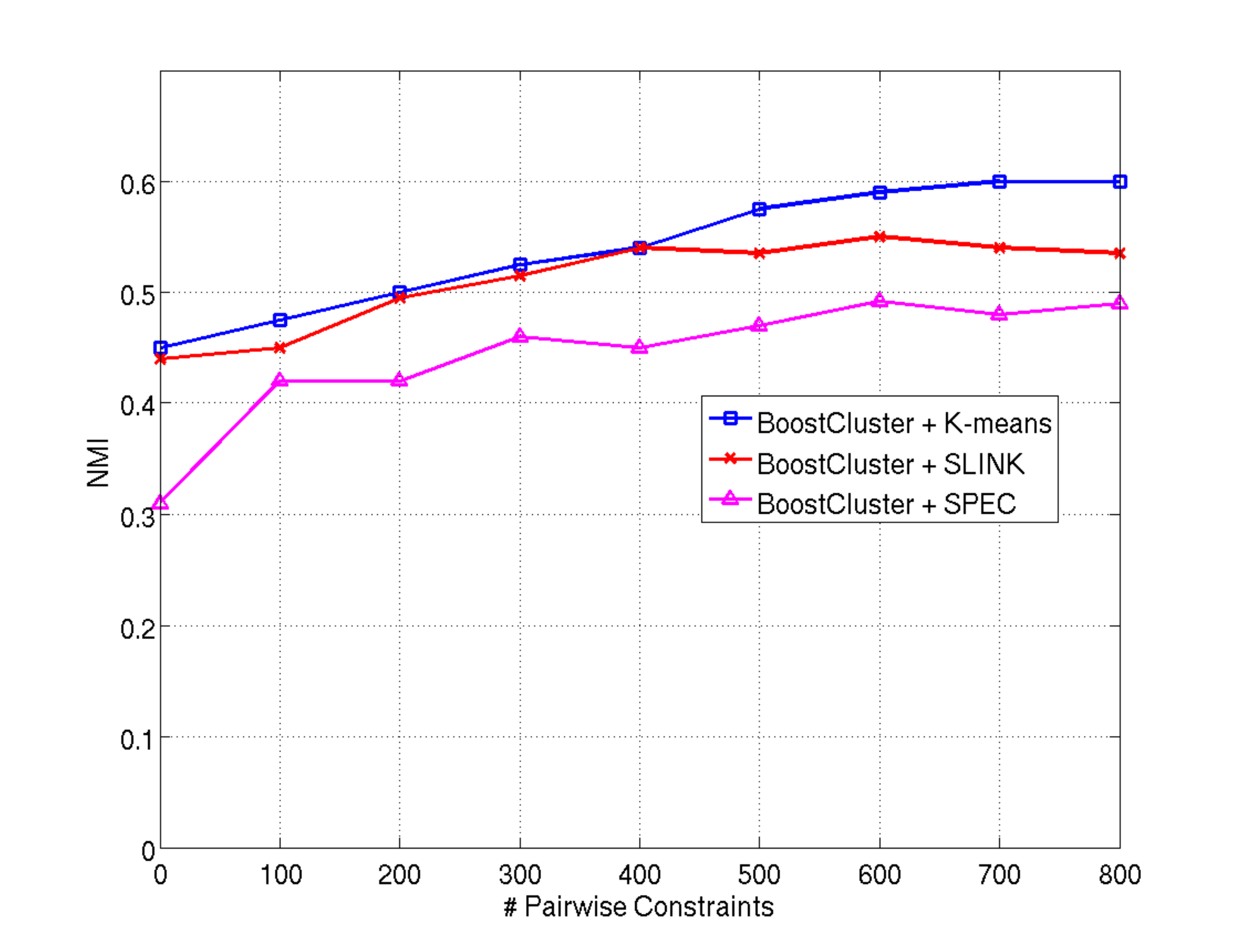
BoostCluster

- Instead of designing new objective fn. improve any given clustering algorithm
- Unsupervised boosting algorithm iteratively updates the similarity matrix input to clustering



Liu, Jin & Jain, BoostCluster: Boosting Clustering by Pairwise Constraints, *KDD*, 2007

Performance of BoostCluster



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

Summary

- Organizing data into sensible groupings arises naturally in many fields
- Cluster analysis is an exploratory tool
- Thousand of algorithms; no best algorithm
- Challenges: representation & similarity; domain knowledge; validation; rational basis for comparing methods, large databases, multiple looks at the same data
- K-means continues to be popular & admissible
- No Silver Bullet!

Acknowledgements

- Richard Dubes, B. Chandrasekaran, Laveen Kanal, Eric Backer, Ana Fred, Mario Figueiredo, Rong Jin, M. Narasimha Murthy, Joachim Buhmann, Robert Duin, Tin Ho, Theo Pavlidis, Josef Kittler, Jake Aggarwal, George Nagy
- My current & former students, in particular Steve Smith, J.C. Mao, Patrick Flynn, Vincent Moreau, Martin Law, Pavan Mallapragada