CSC 390
Topics in Artificial Intelligence

"Unsupervised Machine Learning"

Fall 2016

Prof. Sara Mathieson
Smith College

Outline: 10/13

- Today: Clustering Paper Discussion
- Reminders:
  - Office hours today 4-5pm (Ford 355)
  - Homework 4 due Monday Oct 17
- Grace Hopper next week
  - Tuesday: continue dimensionality reduction
  - Thursday: special topic
- Midterm (take-home)
  - out Tuesday Oct 18
  - due Tuesday Oct 25
  - all material through today
- Homework 5: in-class presentations (start Tues Nov 1)

Paper 1

"Data clustering: 50 years beyond K-means"

by Anil K. Jain

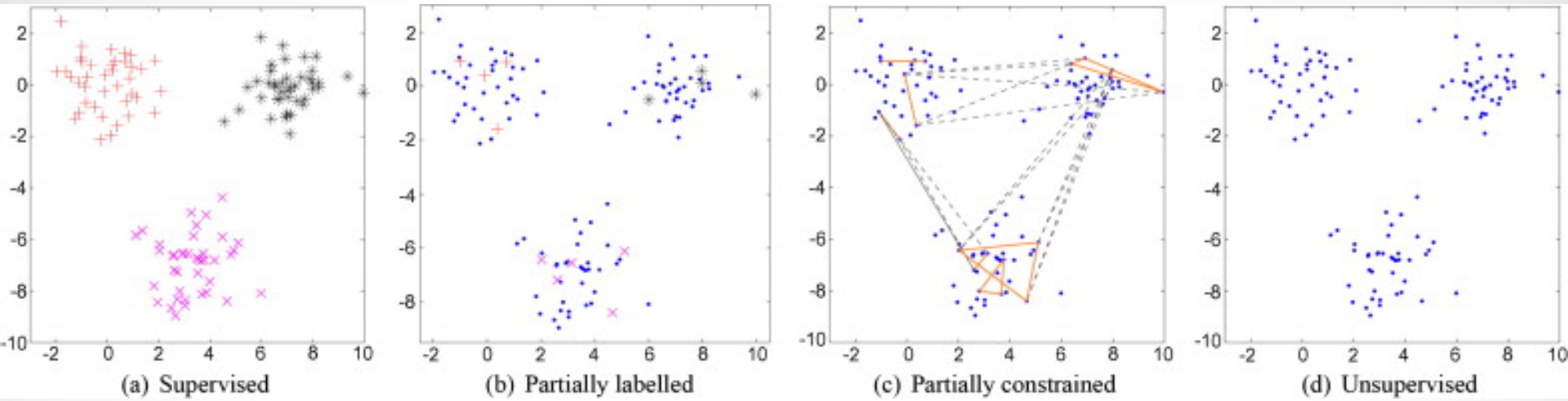


Fig. 1.
Learning problems: dots correspond to points without any labels. Points with labels are denoted by plus signs, asterisks, and crosses. In (c), the must-link and cannot-link constraints are denoted by solid and dashed lines, respectively (figure taken from Lange et al. (2005).

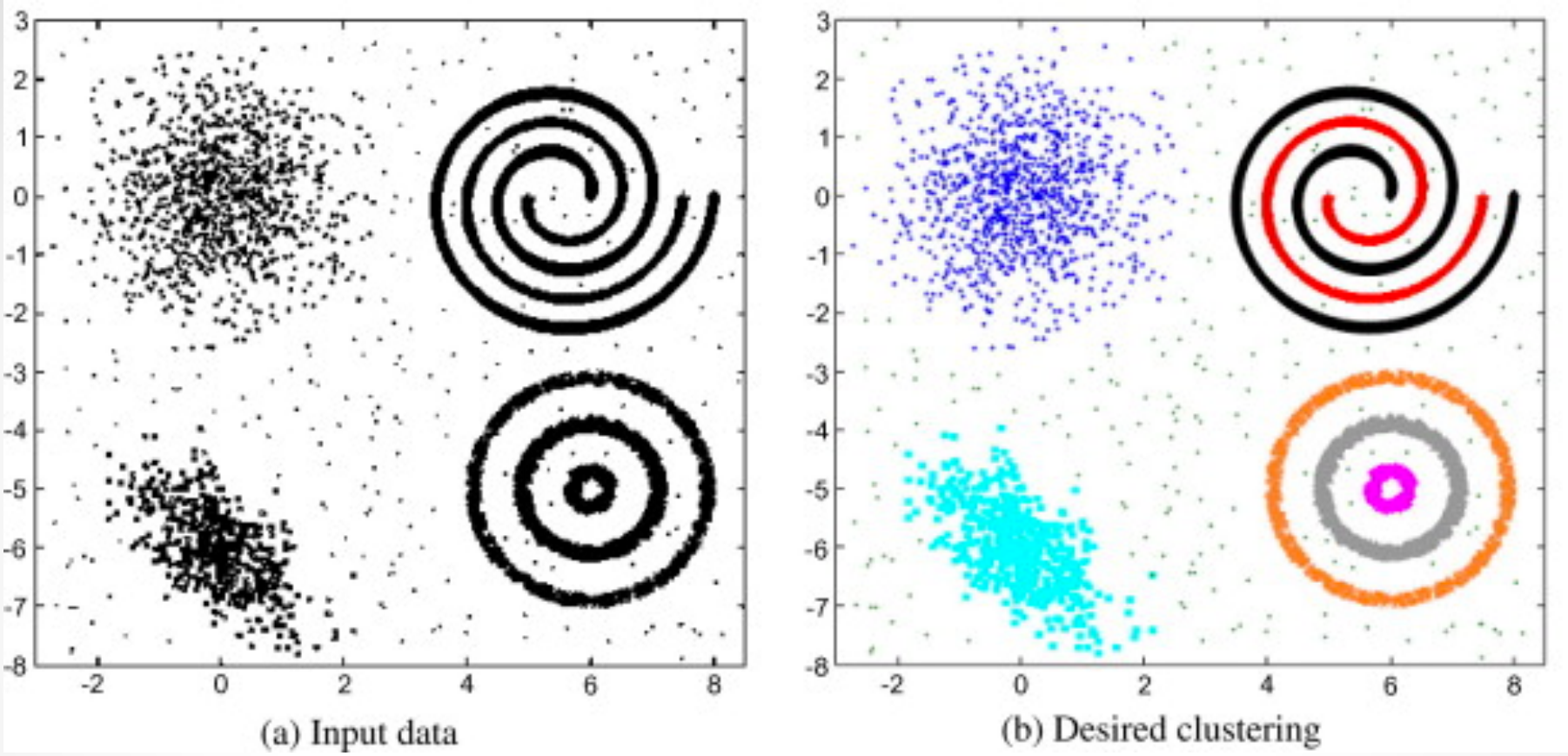


Fig. 2.
Diversity of clusters. The seven clusters in (a) (denoted by seven different colors in 1(b)) differ in shape, size, and density. Although these clusters are apparent to a data analyst, none of the available clustering algorithms can detect all these clusters.

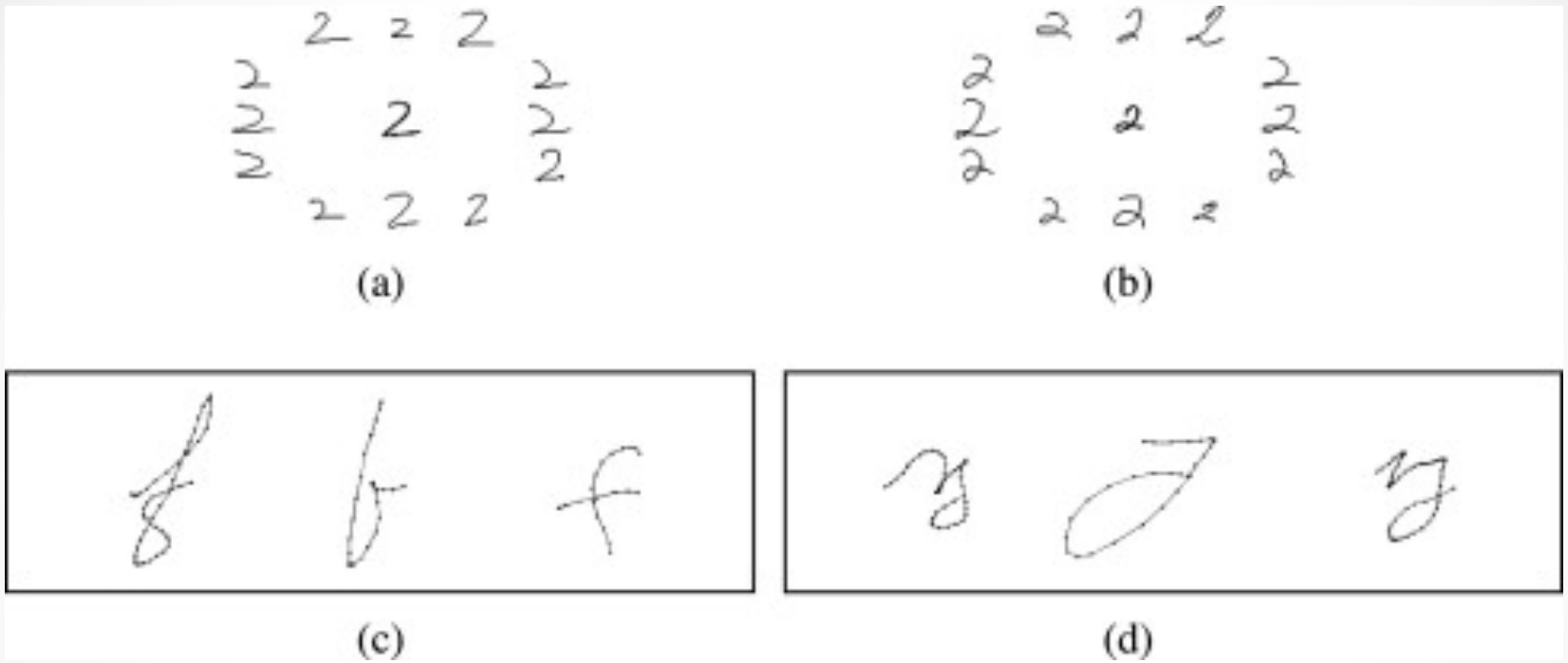


Fig. 3.
Finding subclasses using data clustering. (a) and (b) show two different ways of writing the digit 2; (c) three different subclasses for the character ‘f’; (d) three different subclasses for the letter ‘y’.

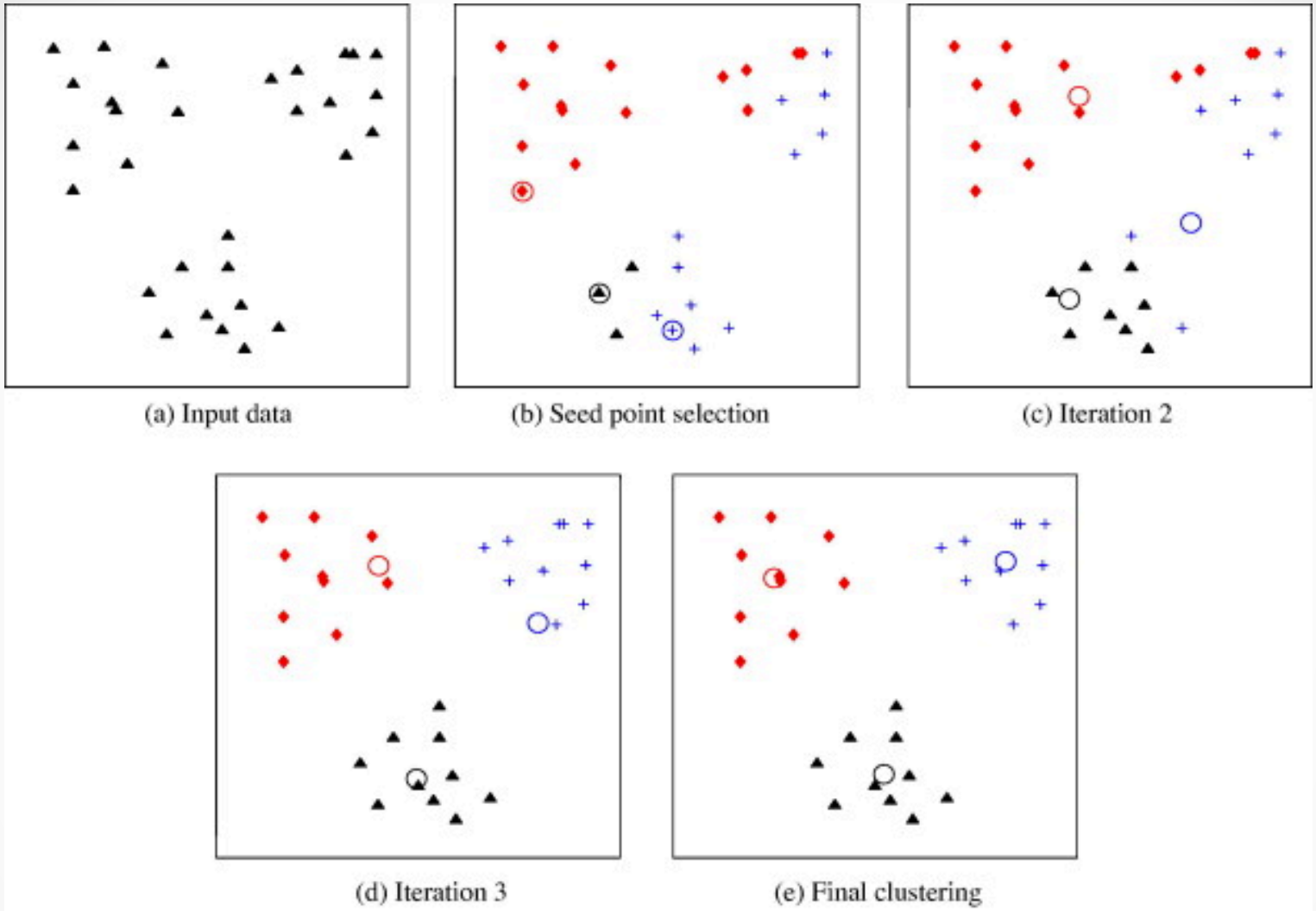


Fig. 4.
Illustration of K-means algorithm. (a) Two-dimensional input data with three clusters; (b) three seed points selected as cluster centers and initial assignment of the data points to clusters; (c) and (d) intermediate iterations updating cluster labels and their centers; (e) final clustering obtained by K-means algorithm at convergence.

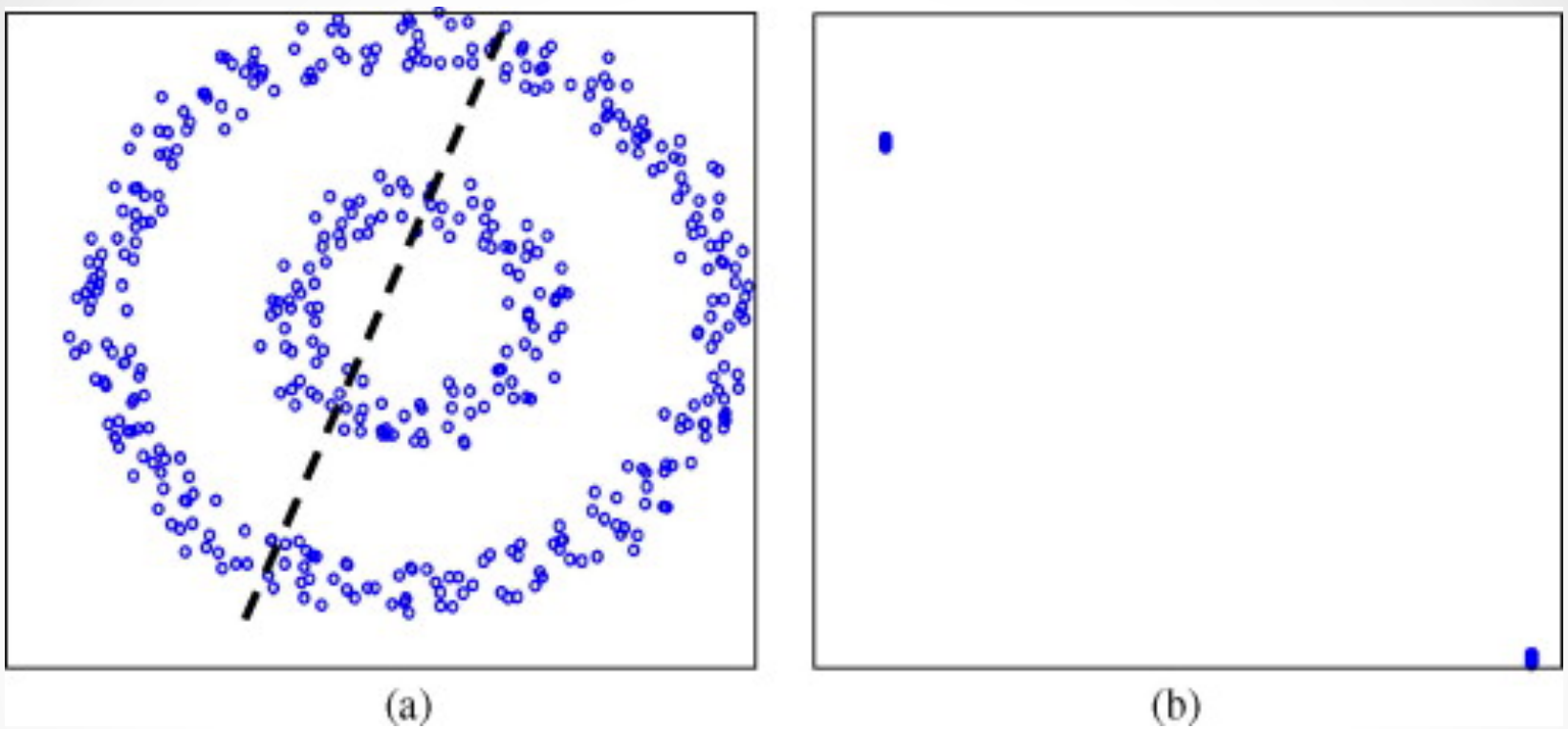


Fig. 5.
Importance of a good representation. (a) “Two rings” dataset where K-means fails to find the two “natural” clusters; the dashed line shows the linear cluster separation boundary obtained by running K-means with K = 2. (b) a new representation of the data in (a) based on the top 2 eigenvectors of the graph Laplacian of the data, computed using an RBF kernel; K-means now can easily detect the two clusters.

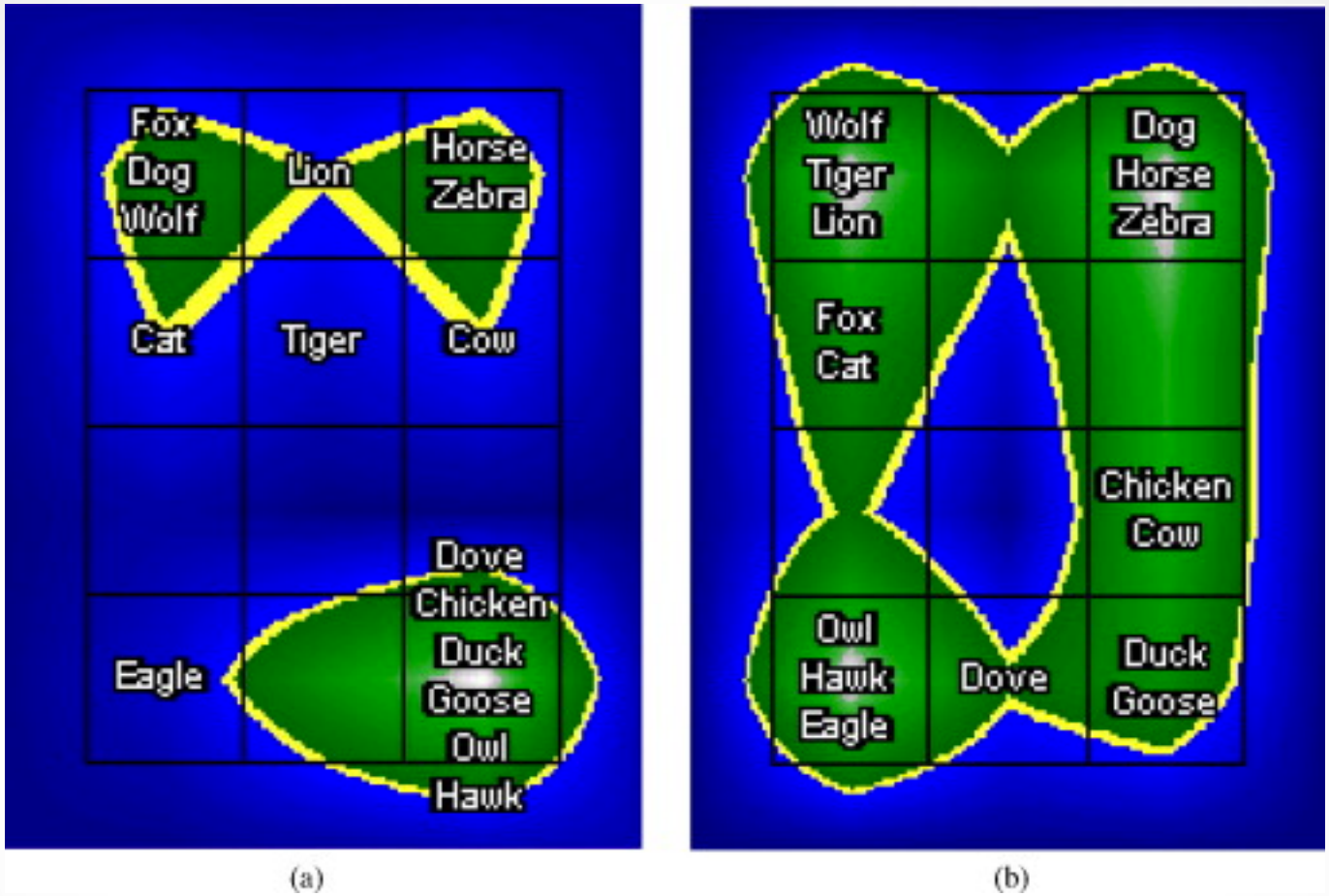


Fig. 6.
Different weights on features result in different partitioning of the data. Sixteen animals are represented based on 13 Boolean features related to appearance and activity. (a) Partitioning with large weights assigned to the appearance based features; (b) a partitioning with large weights assigned to the activity features. The figures in (a) and (b) are excerpted from Pampalk et al. (2003), and are known as "heat maps" where the colors represent the density of samples at a location; the warmer the color, the larger the density.

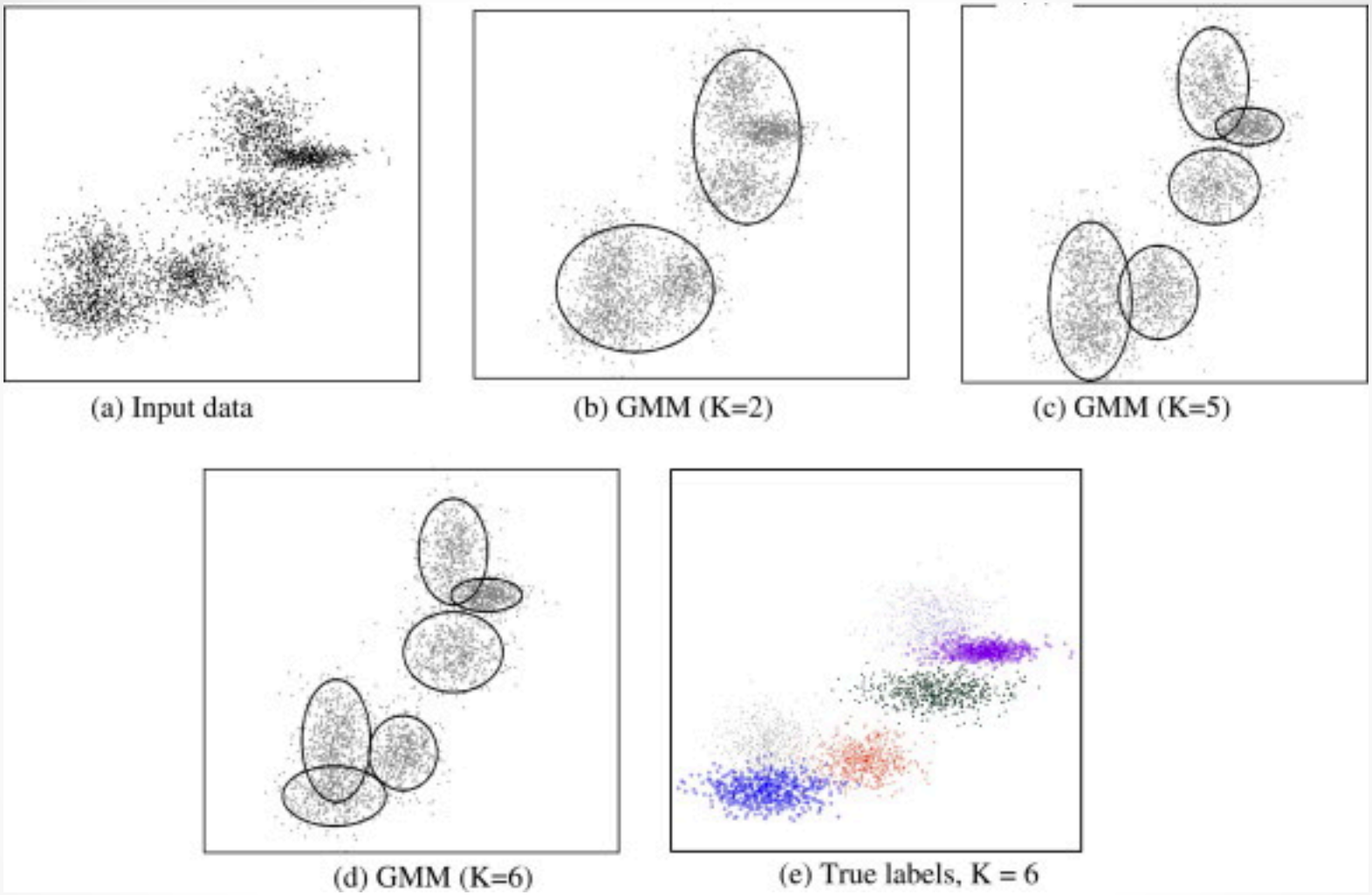


Fig. 7.
Automatic selection of number of clusters, K. (a) Input data generated from a mixture of six Gaussian distributions; (b)–(d) Gaussian mixture model (GMM) fit to the data with 2, 5, and 6 components, respectively; and (e) true labels of the data.0

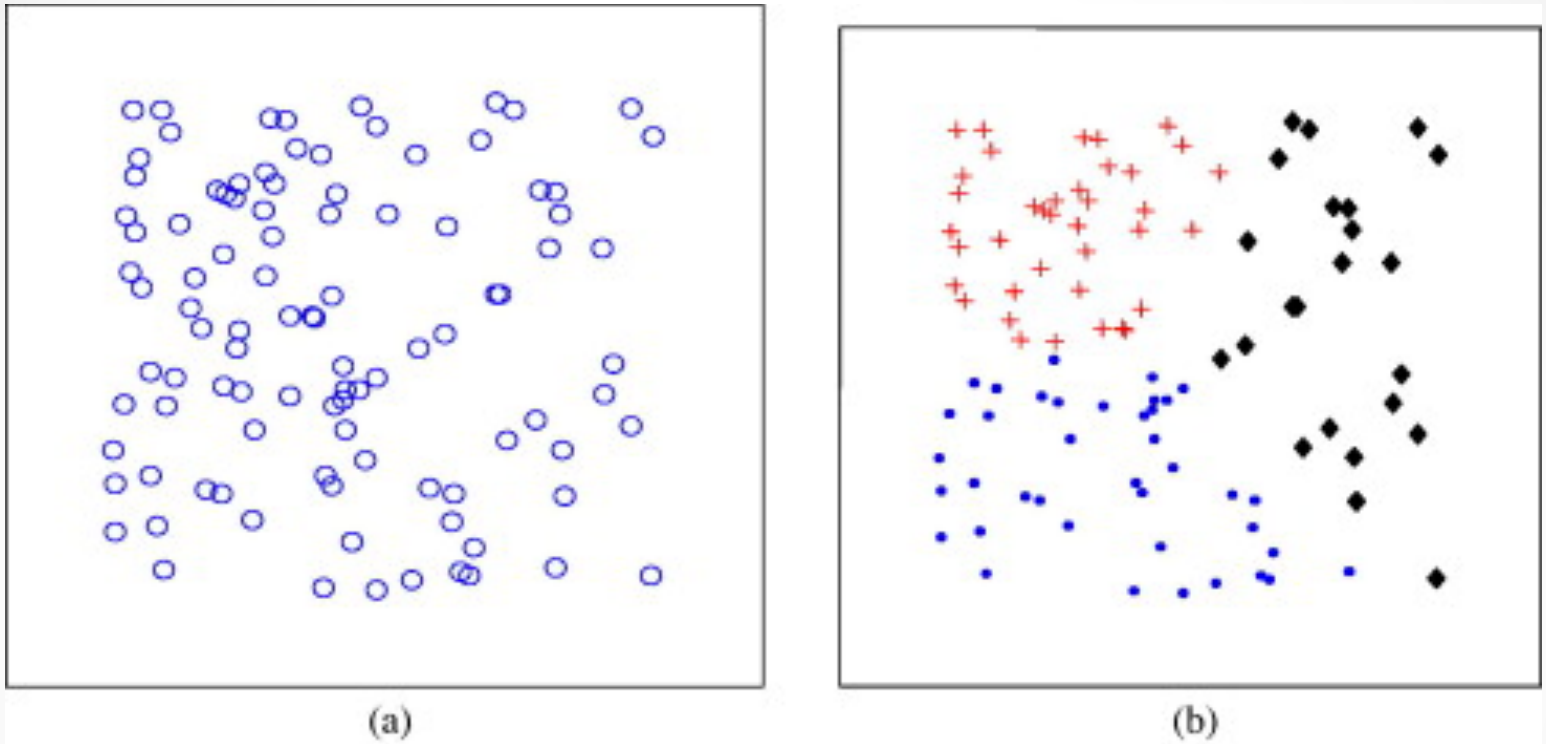


Fig. 8.
Cluster validity. (a) A dataset with no “natural” clustering; (b) K-means partition with K = 3.

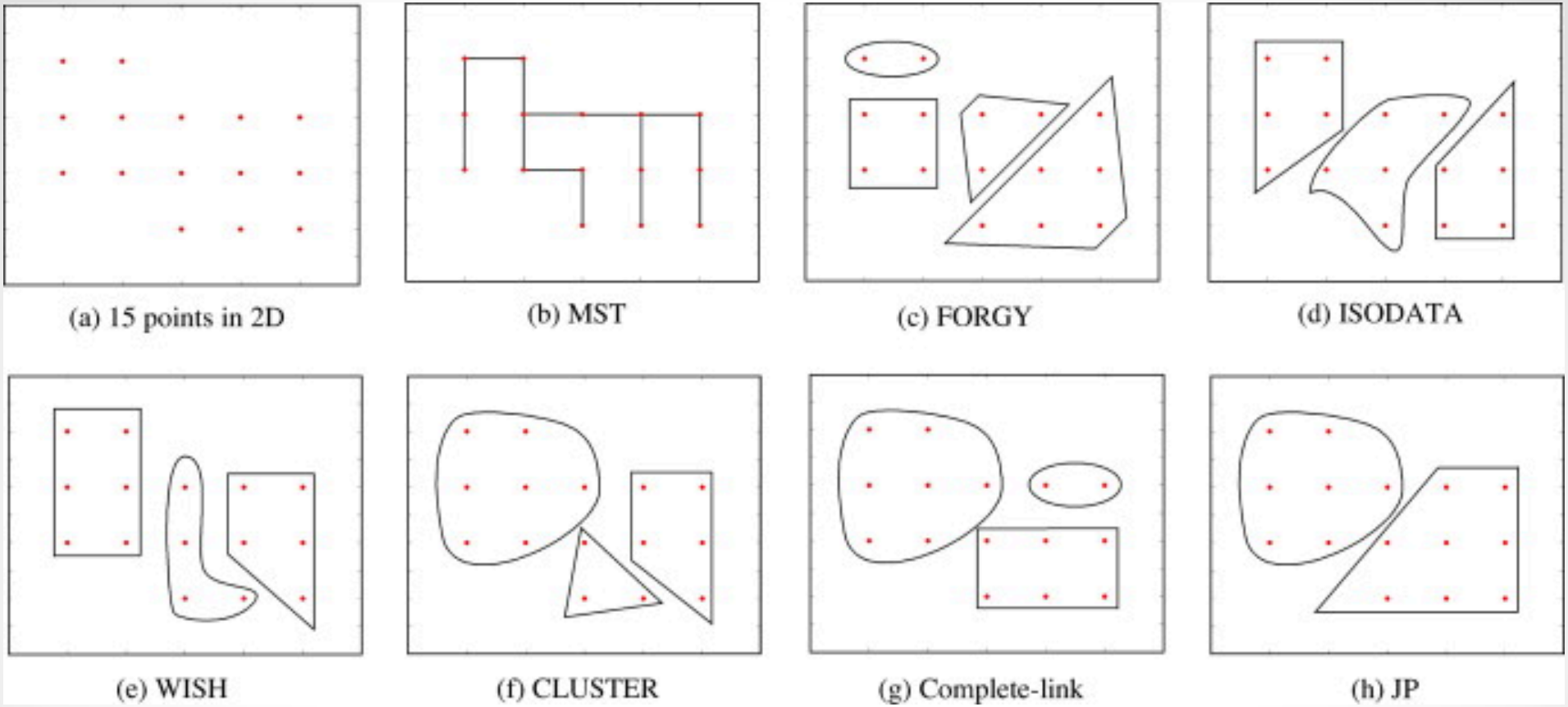


Fig. 9.
Several clusterings of fifteen patterns in two dimensions: (a) fifteen patterns; (b) minimum spanning tree of the fifteen patterns; (c) clusters from FORGY; (d) clusters from ISODATA; (e) clusters from WISH; (f) clusters from CLUSTER; (g) clusters from complete-link hierarchical clustering; and (h) clusters from Jarvis-Patrick clustering algorithm. (Figure reproduced from Dubes and Jain (1976).)

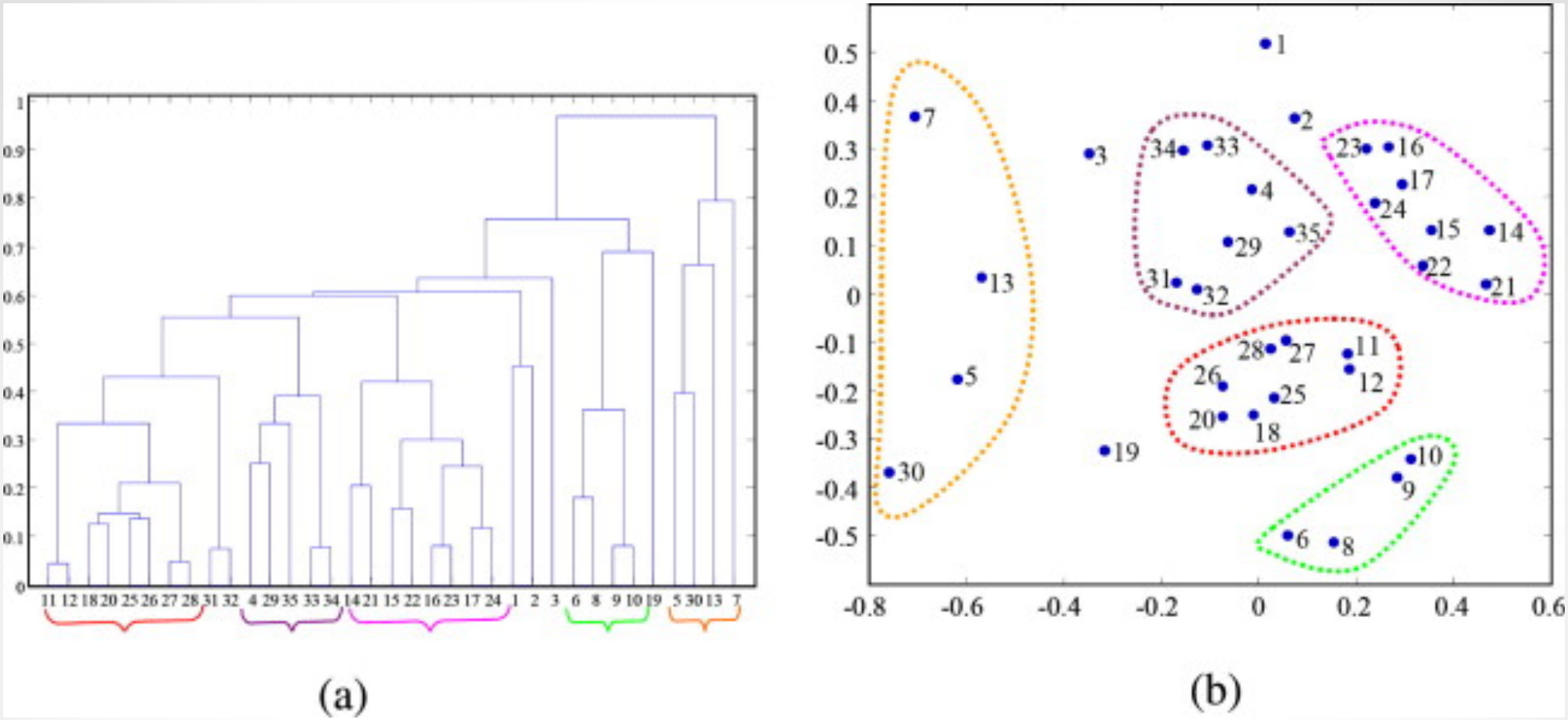


Fig. 10.
Clustering of clustering algorithms. (a) Hierarchical clustering of 35 different algorithms; (b) Sammon's mapping of the 35 algorithms into a two-dimensional space, with the clusters highlighted for visualization. The algorithms in the group (4, 29, 31–35) correspond to K-means, spectral clustering, Gaussian mixture models, and Ward's linkage. The algorithms in group (6, 8–10) correspond to CHAMELEON algorithm with different objective functions.

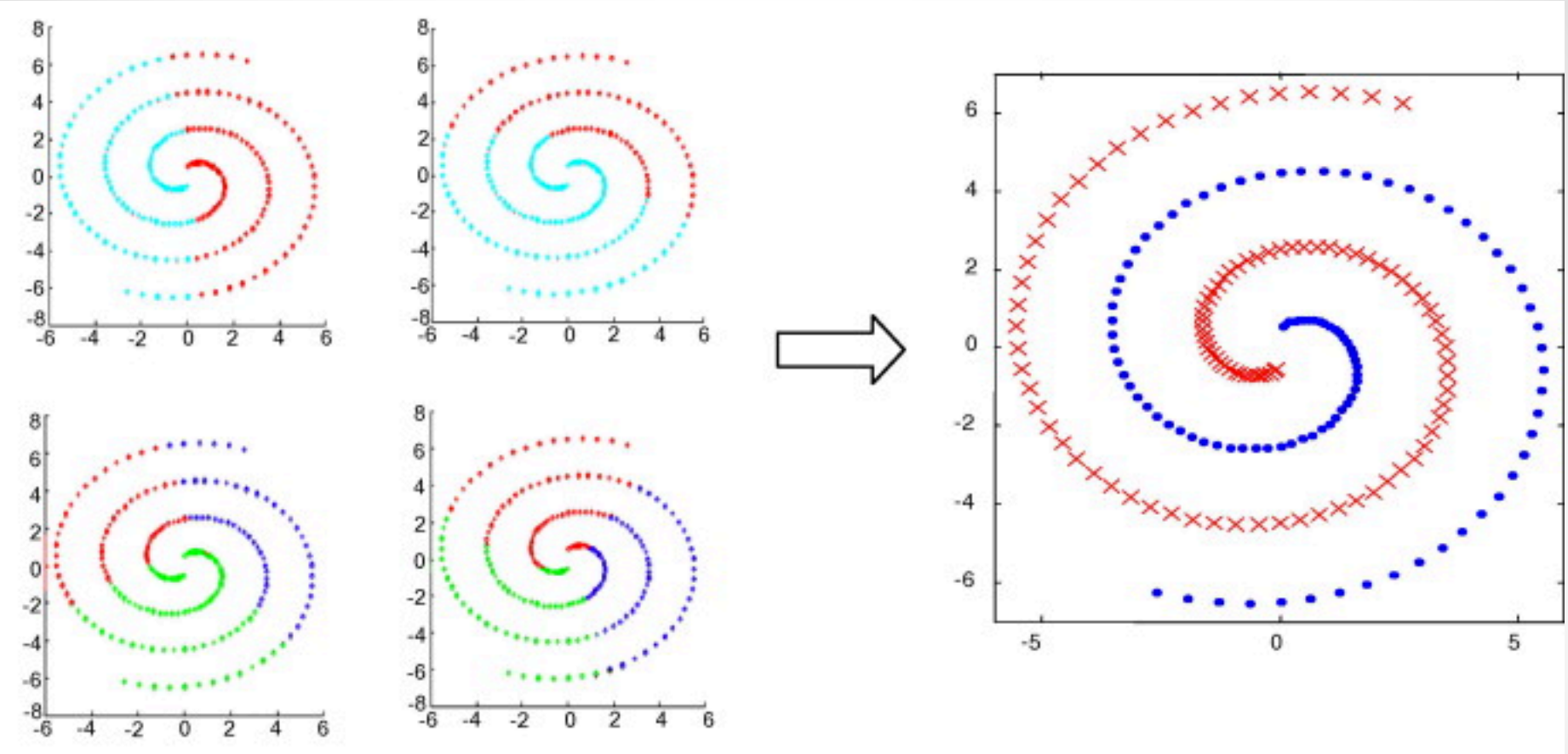
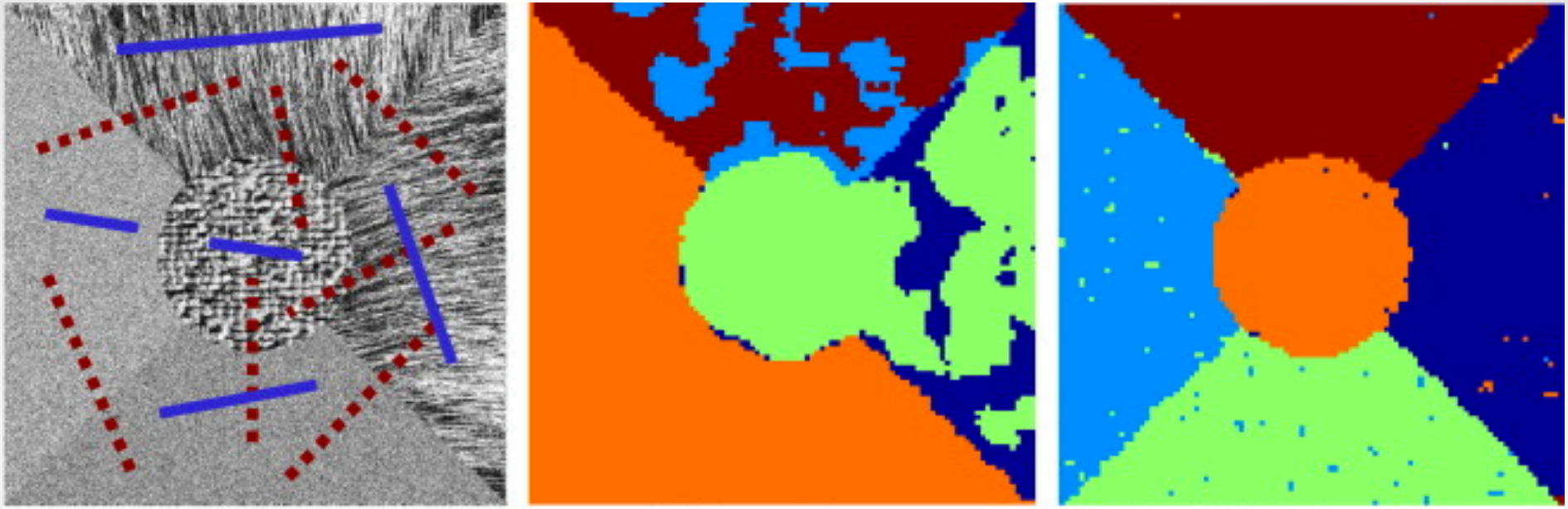


Fig. 11.
Clustering ensembles. Multiple runs of K-means are used to learn the pair-wise similarity using the “co-occurrence” of points in clusters. This similarity can be used to detect arbitrary shaped clusters.



(a) Input image and constraints (b) No constraints (c) 10% pixels in constraints

Fig. 12.
Semi-supervised learning. (a) Input image consisting of five homogeneous textured regions; examples of must-link (solid blue lines) and must not link (broken red lines) constraints between pixels to be clustered are specified. (b) 5-Cluster solution (segmentation) without constraints. (c) Improved clustering (with five clusters) with 10% of the data points included in the pair-wise constraints (Lange et al., 2005).

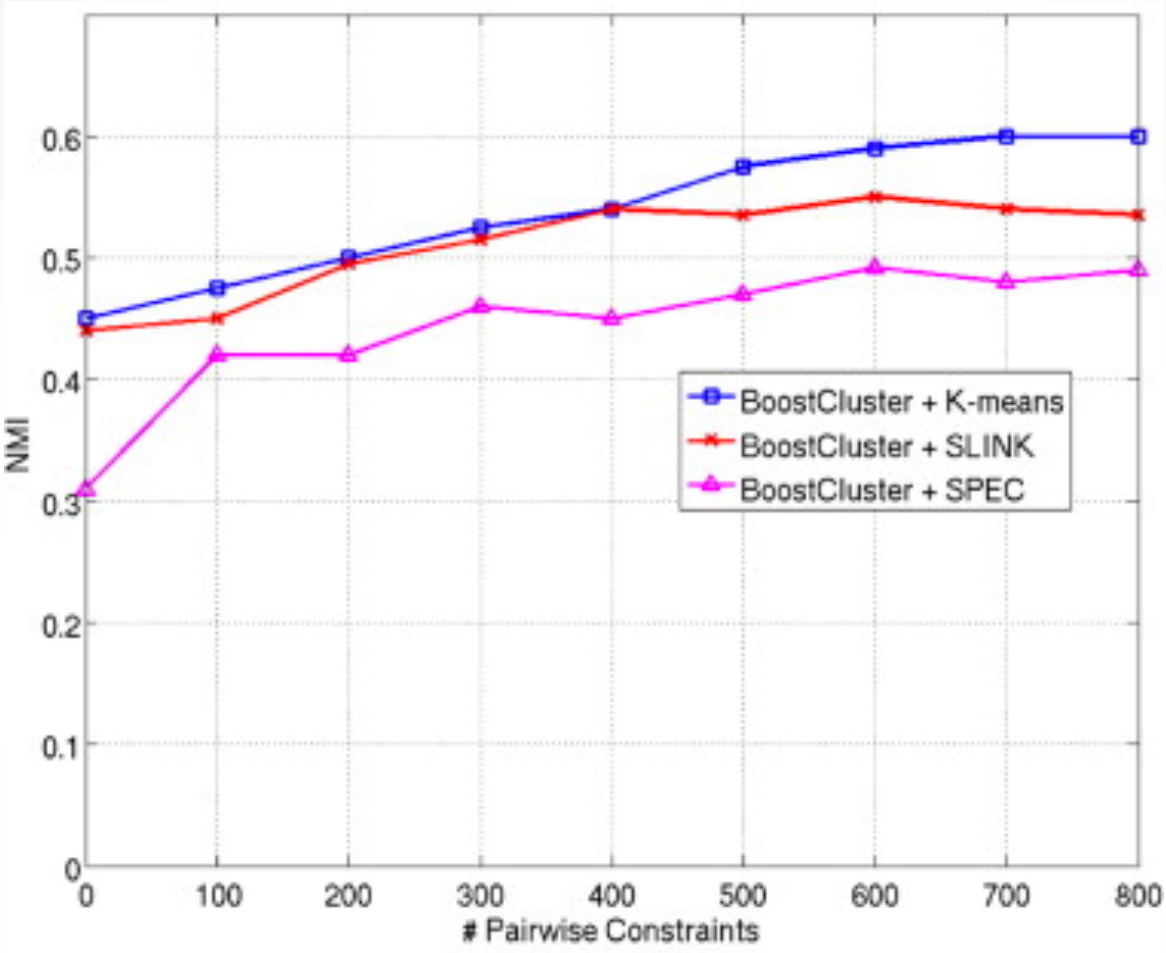


Fig. 13.
Performance of BoostCluster (measured using Normalized Mutual Information (NMI)) as the number of pair-wise constraints is increased. The three plots correspond to boosted performance of K-means, Single-Link (SLINK), and Spectral clustering (SPEC).

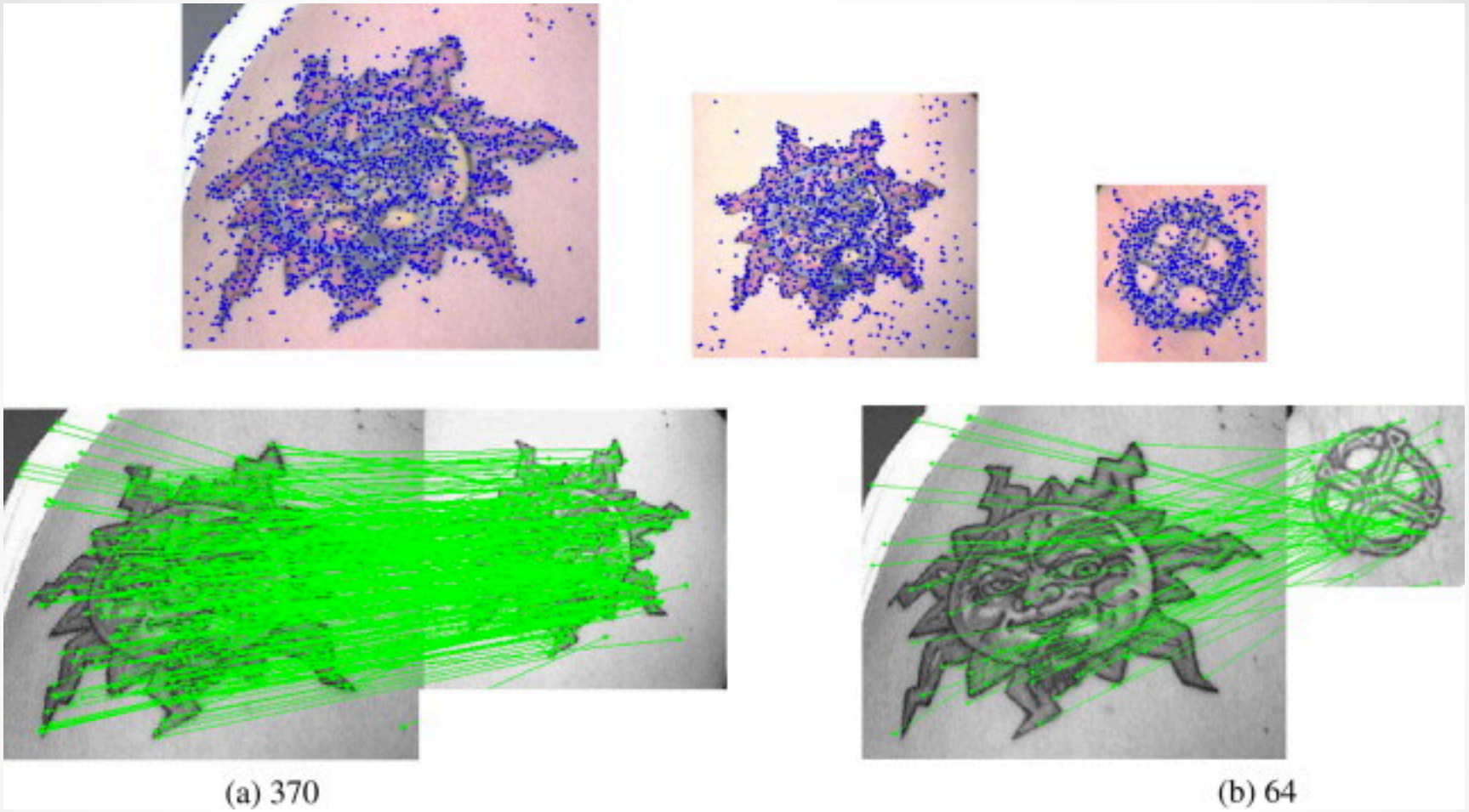


Fig. 14.
Three tattoo images represented using SIFT key points. (a) A pair of similar images has 370 matching key points; (b) a pair of different images has 64 matching key points. The green lines show the matching key-points between the images (Lee et al., 2008).

Paper 2

"Dynamic hierarchical algorithms for document clustering"

by Reynaldo Gil-García, Aurora Pons-Porrata

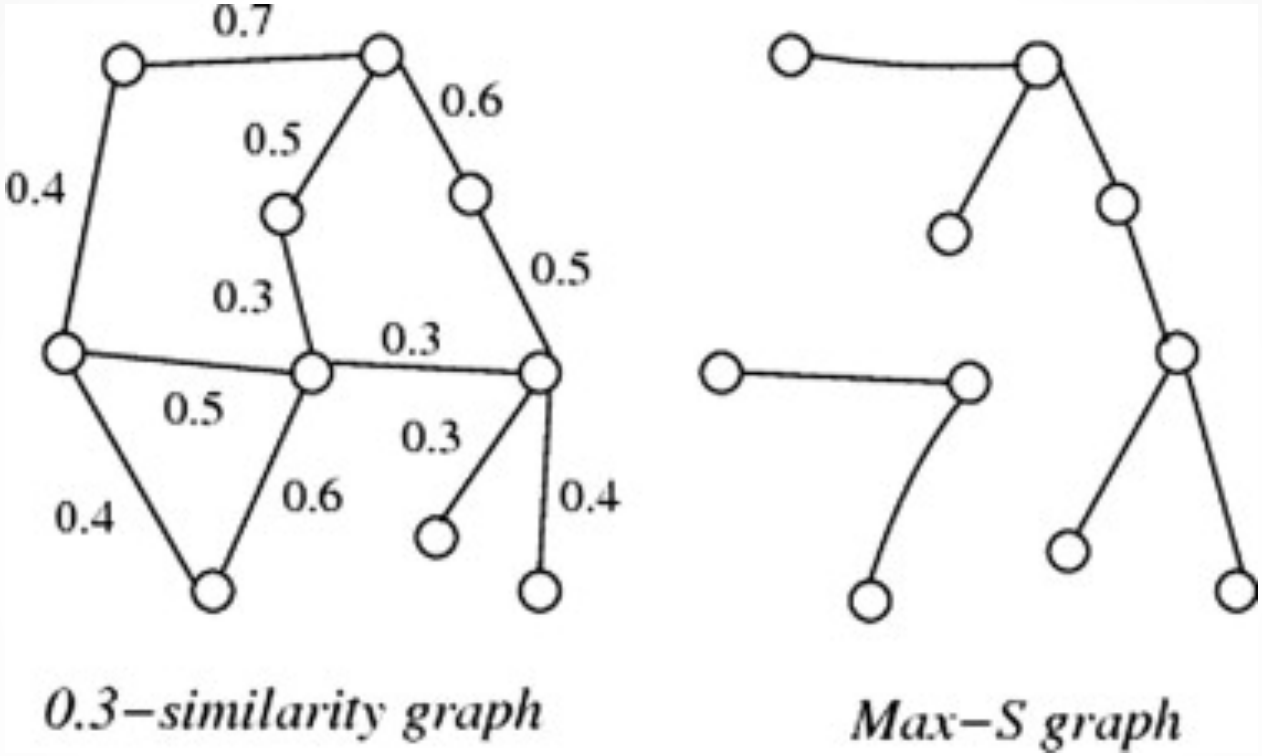


Fig. 1.
Graphs based on β-similarity.

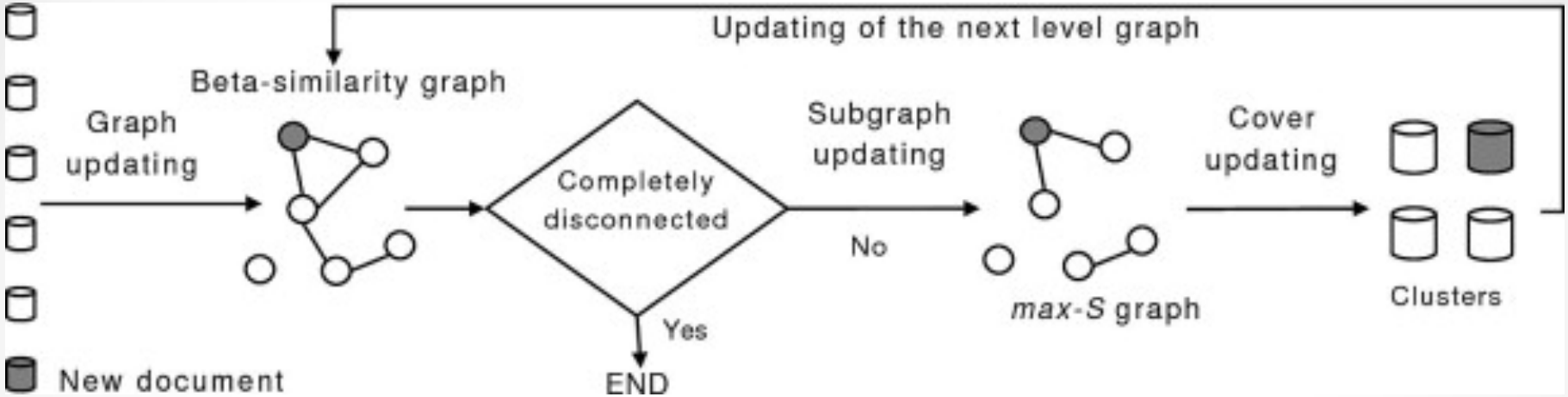


Fig. 2.
Dynamic hierarchical agglomerative framework.

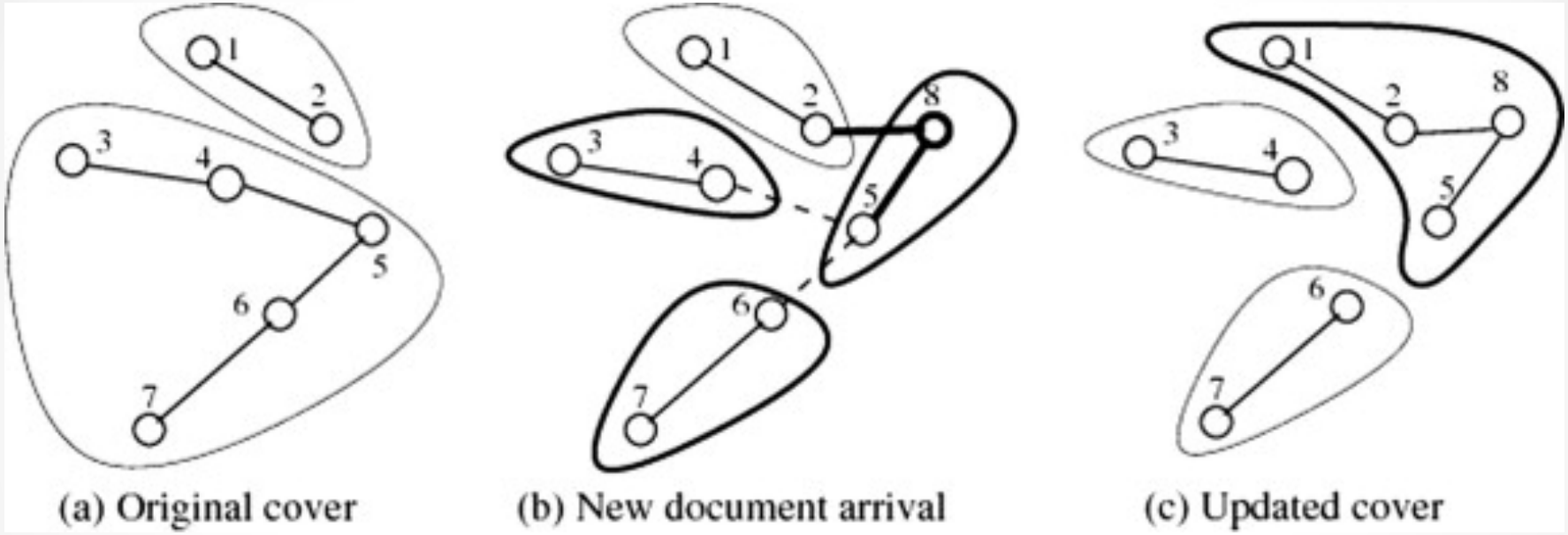


Fig. 3.
Updating of the connected component cover.

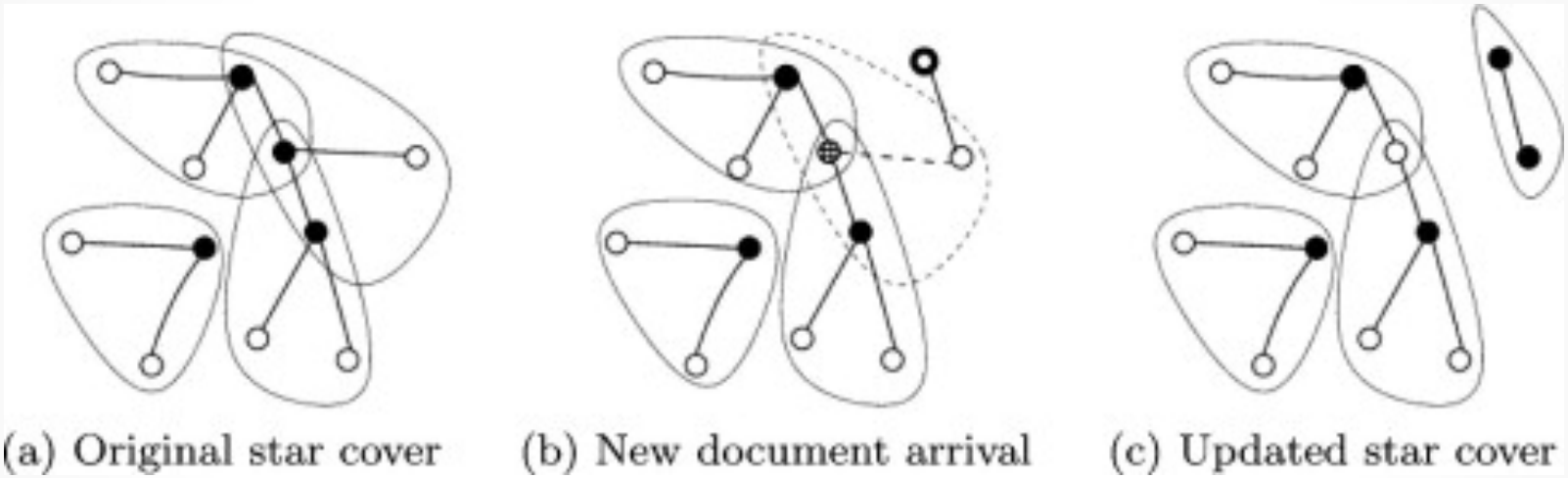


Fig. 4.
Star cover updating (black circles represent the stars).

Table 2.
Overall F1 results for document collections (x = not scalable to run).

Data	Method	C	F1	Ov.	Method	C	F1	Ov.
afp	UPGMA	693	0.89	17.95	DHC	172	0.85	3.00
	BKM	693	0.86	9.61	DHS	468	0.85	6.07
eln	UPGMA	5828	0.56	52.09	DHC	521	0.47	5.00
	BKM	5828	0.55	12.65	DHS	4023	0.54	13.19
tdt	UPGMA	9821	0.90	35.61	DHC	2571	0.84	5.00
	BKM	9821	0.90	14.21	DHS	6776	0.87	10.14
reu	UPGMA	10368	0.68	50.51	DHC	1904	0.55	5.00
	BKM	10368	0.64	14.03	DHS	7023	0.58	13.08
reu2	UPGMA	x	x	x	DHC	7298	0.39	6.00
	BKM	23147	0.47	15.05	DHS	16146	0.44	12.22
re0	UPGMA	1502	0.63	24.19	DHC	450	0.64	5.00
	BKM	1502	0.62	10.83	DHS	1059	0.61	8.97
re1	UPGMA	1655	0.63	23.39	DHC	478	0.62	4.00
	BKM	1655	0.63	10.52	DHS	1170	0.62	8.47
wap	UPGMA	1558	0.62	45.37	DHC	474	0.56	4.00
	BKM	1558	0.67	12.24	DHS	1060	0.58	8.63
la12	UPGMA	6277	0.56	57.42	DHC	1393	0.56	4.00
	BKM	6277	0.67	13.30	DHS	4276	0.63	12.40
k1a	UPGMA	2338	0.61	52.25	DHC	741	0.54	5.00
	BKM	2338	0.64	13.18	DHS	1633	0.57	8.74
k1b	UPGMA	2338	0.90	52.25	DHC	741	0.90	5.00
	BKM	2338	0.89	13.18	DHS	1639	0.89	9.71
Ohscal	UPGMA	11161	0.42	67.41	DHC	1868	0.30	5.00
	BKM	11161	0.44	13.78	DHS	7834	0.39	12.82
Reviews	UPGMA	4067	0.79	48.74	DHC	867	0.67	4.00
	BKM	4067	0.72	12.47	DHS	2828	0.77	10.17
Hitech	UPGMA	2299	0.54	35.40	DHC	532	0.52	4.00
	BKM	2299	0.53	11.31	DHS	1531	0.55	9.73
Sports	UPGMA	8579	0.77	73.65	DHC	1662	0.81	6.00
	BKM	8579	0.74	13.6	DHS	5968	0.83	12.13

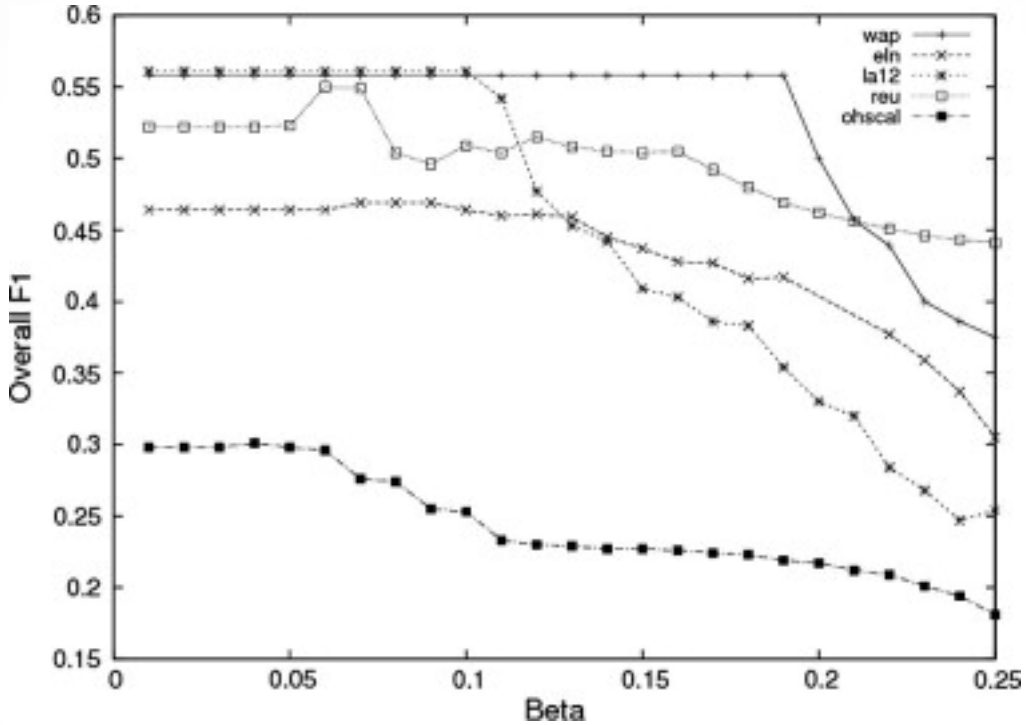


Fig. 5.
DHC sensitivity to β.

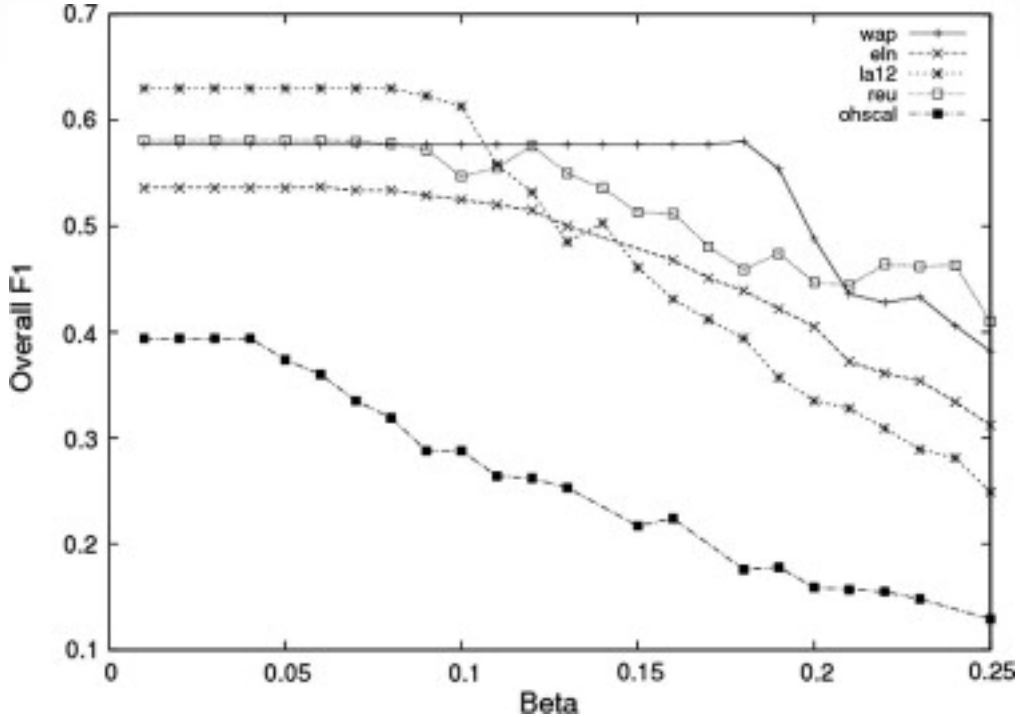


Fig. 6.
DHS sensitivity to β.

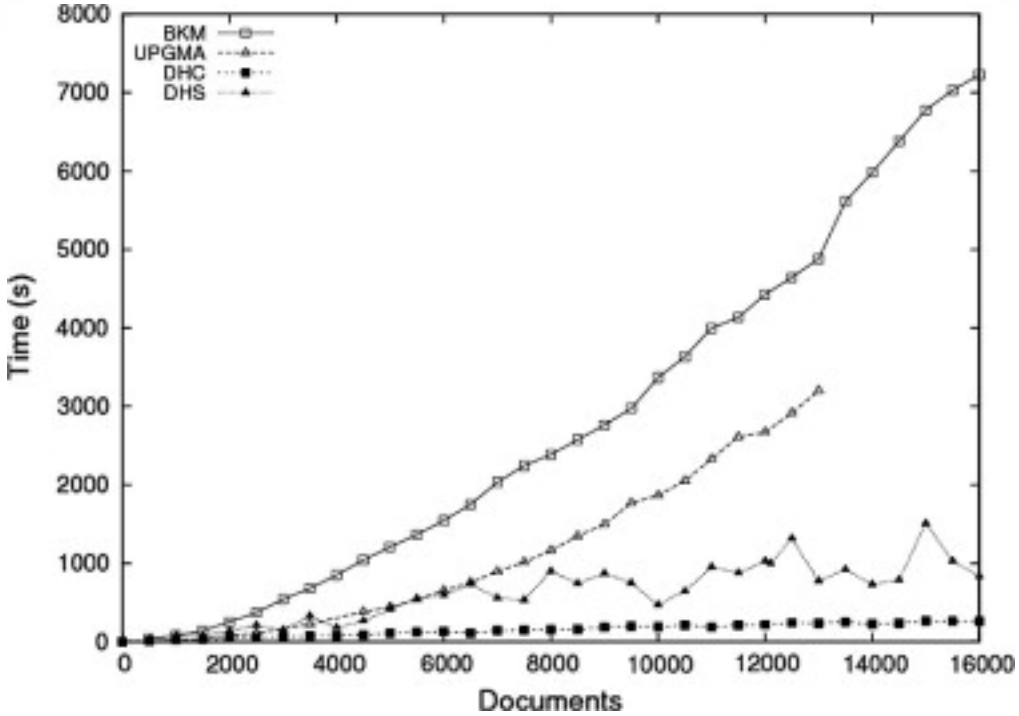


Fig. 7.
Time performance.

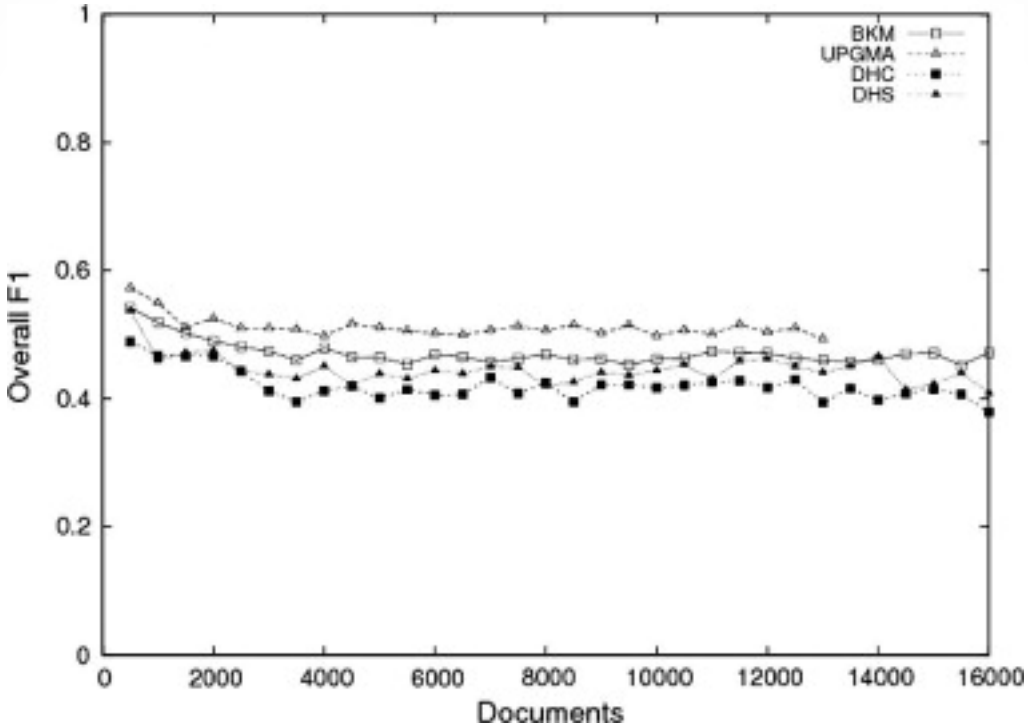


Fig. 8.
Clustering quality in a dynamic environment.