CSC 390
Topics in Artificial Intelligence

"Unsupervised Machine Learning"

Fall 2016

Prof. Sara Mathieson
Smith College

Outline: 10/18

- Finish Paper Discussion
- More about how to read a scientific paper
- Begin: probability and statistics background

- Grace Hopper
  - Thursday: special topic: machine learning in population genetics

- Midterm (take-home)
  - out Tuesday Oct 18
  - due Wednesday Oct 25 at 5pm

- Homework 5: in-class presentations (start Tues Nov 1)

- Feedback

Strategies for reading a paper

- If you're not sure you want to read the paper… read the abstract

Strategies for reading a paper

- If you're not sure you want to read the paper… read the abstract
- If you're new to the field… read the introduction in detail, then skim conclusions

Strategies for reading a paper

- If you're not sure you want to read the paper… read the abstract
- If you're new to the field… read the introduction in detail, then skim conclusions
- After that, move on to the methods if you're interested in the details

Strategies for reading a paper

- If you're not sure you want to read the paper… read the abstract
- If you're new to the field… read the introduction in detail, then skim conclusions
- After that, move on to the methods if you're interested in the details
- If you're experienced in the field… skim conclusions first, then go to the methods

Strategies for reading methods and results

- Spend a lot of time on the setup (notation, etc), since without this the rest can be very confusing

Strategies for reading methods and results

- Spend a lot of time on the setup (notation, etc), since without this the rest can be very confusing

- Spend a lot of time on the initial method figures (especially if they have examples, etc)

Strategies for reading methods and results

- Spend a lot of time on the setup (notation, etc), since without this the rest can be very confusing
- Spend a lot of time on the initial method figures (especially if they have examples, etc)
- Endless variants of their algorithms can sometimes be skimmable

Strategies for reading methods and results

- Spend a lot of time on the setup (notation, etc), since without this the rest can be very confusing
- Spend a lot of time on the initial method figures (especially if they have examples, etc)
- Endless variants of their algorithms can sometimes be skimmable
- Results: spend time on the key evaluation metrics, some later material can be skimmed if it's long

Impact of the review process

- Often reviewers will request lots of additional comparisons (vary the parameters, compare with more methods, etc)

Impact of the review process

- Often reviewers will request lots of additional comparisons (vary the parameters, compare with more methods, etc)

- Can make the results long, but also valuable

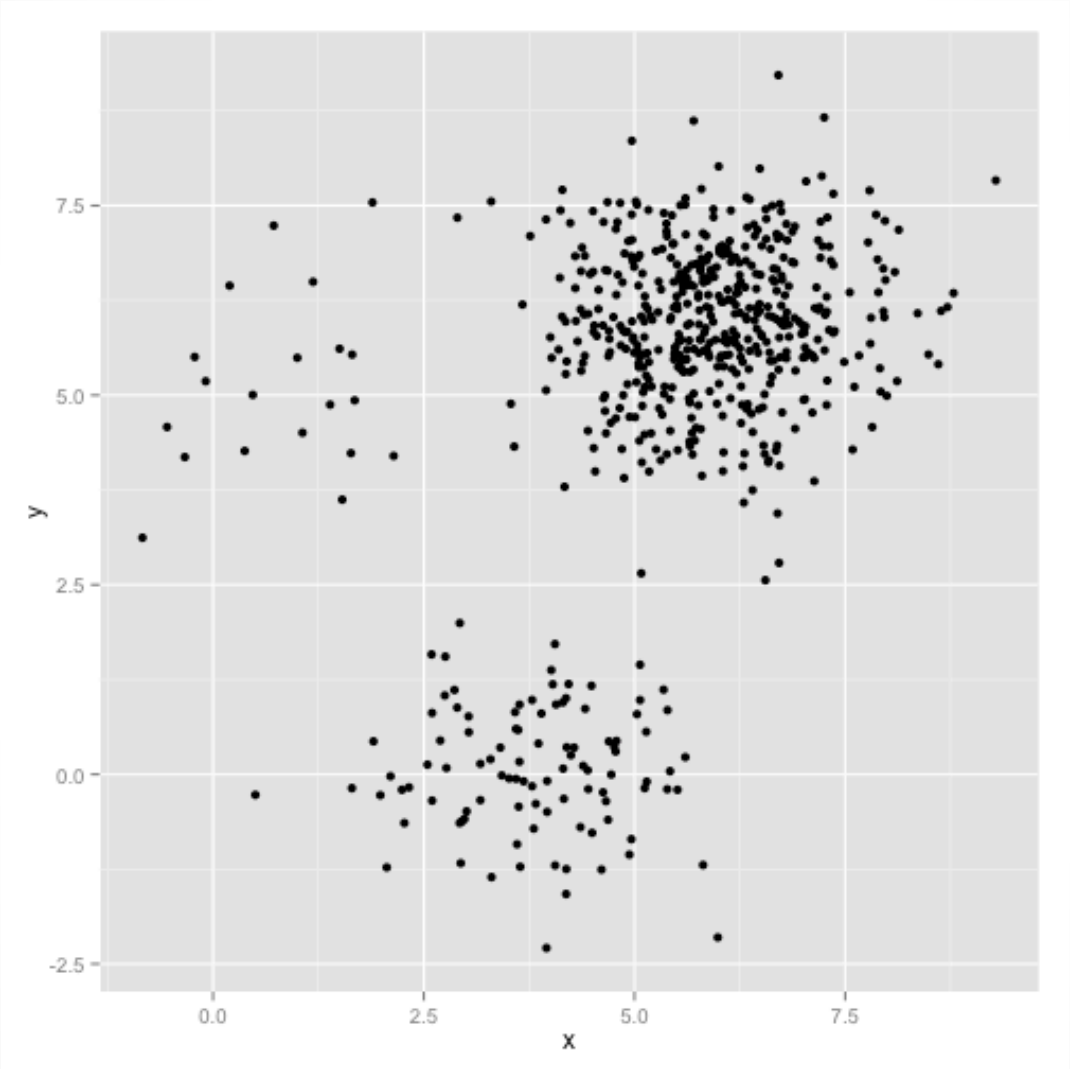
Impact of the review process

- Often reviewers will request lots of additional comparisons (vary the parameters, compare with more methods, etc)
- Can make the results long, but also valuable
- Use methods and conclusions to decide whether you want to run the algorithm on your data

Impact of the review process

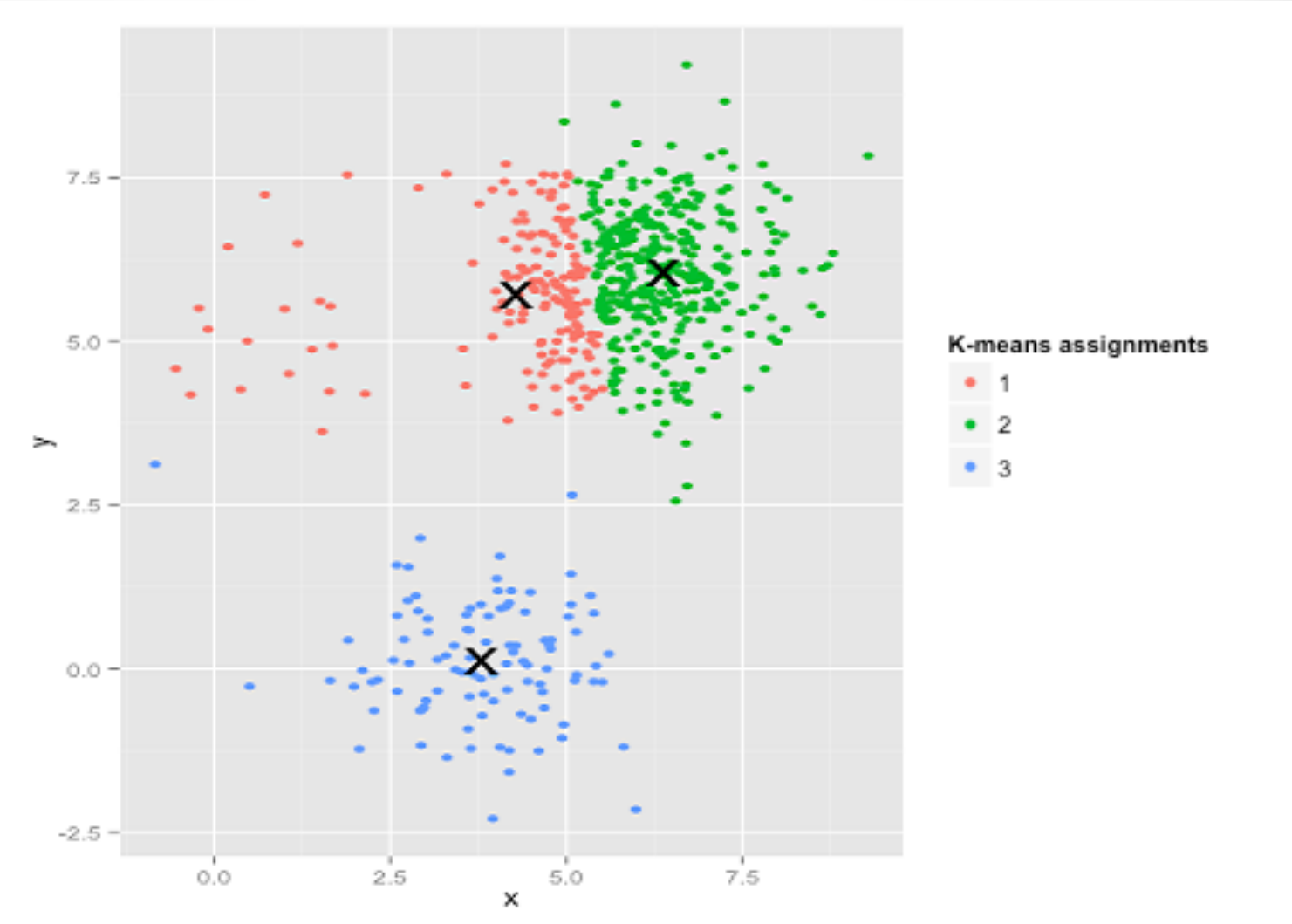
- Often reviewers will request lots of additional comparisons (vary the parameters, compare with more methods, etc)
- Can make the results long, but also valuable
- Use methods and conclusions to decide whether you want to run the algorithm on your data
- Results on one dataset may not be that meaningful for another

Example where k-means produces undesirable results



- http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Example where k-means produces undesirable results



- http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Example where k-means produces undesirable results

- Larger issue: k-means is biased toward equally sized clusters

- http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

Terms from these two papers

- Strategies:
  - Google term along with the word "clustering"
  - Use context or try to find the first time the term was used
  - Mark and come back if it ends up being an important term
  - Relate back to a non-scientific definition of the term

Terms from these two papers

- Monotone
- Disjoint
- Chaining effect
- Document taxonomy
- Similarity matrix
- Distributions
- Agglomerative, divisive
- Overlapping
- Cardinality
- Cosine similarity
- Centroid
- Partition
- Data representation
- Topics

Paper 1

"Data clustering: 50 years beyond K-means"

by Anil K. Jain

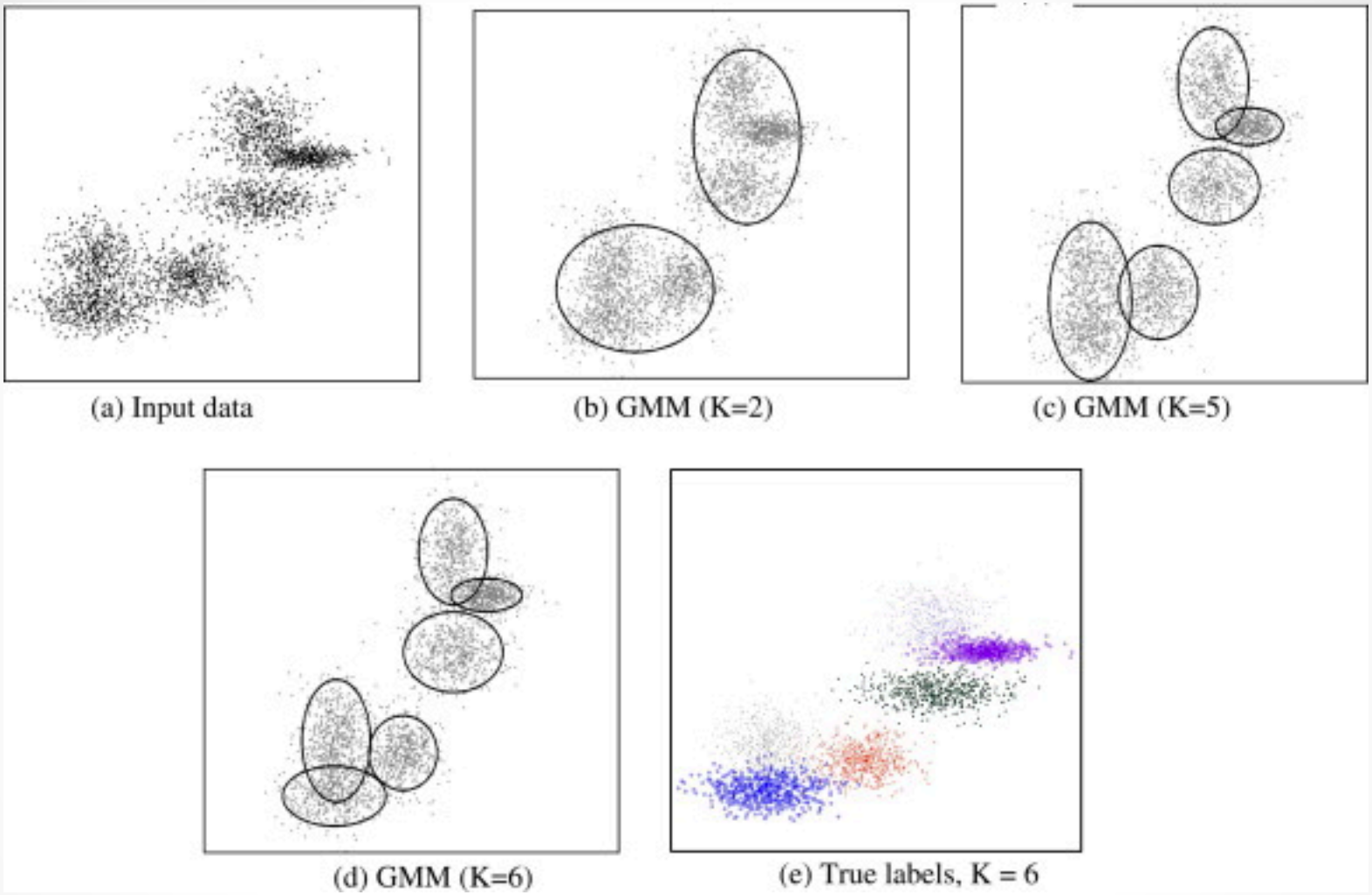


Fig. 7.
Automatic selection of number of clusters, K. (a) Input data generated from a mixture of six Gaussian distributions; (b)–(d) Gaussian mixture model (GMM) fit to the data with 2, 5, and 6 components, respectively; and (e) true labels of the data.0

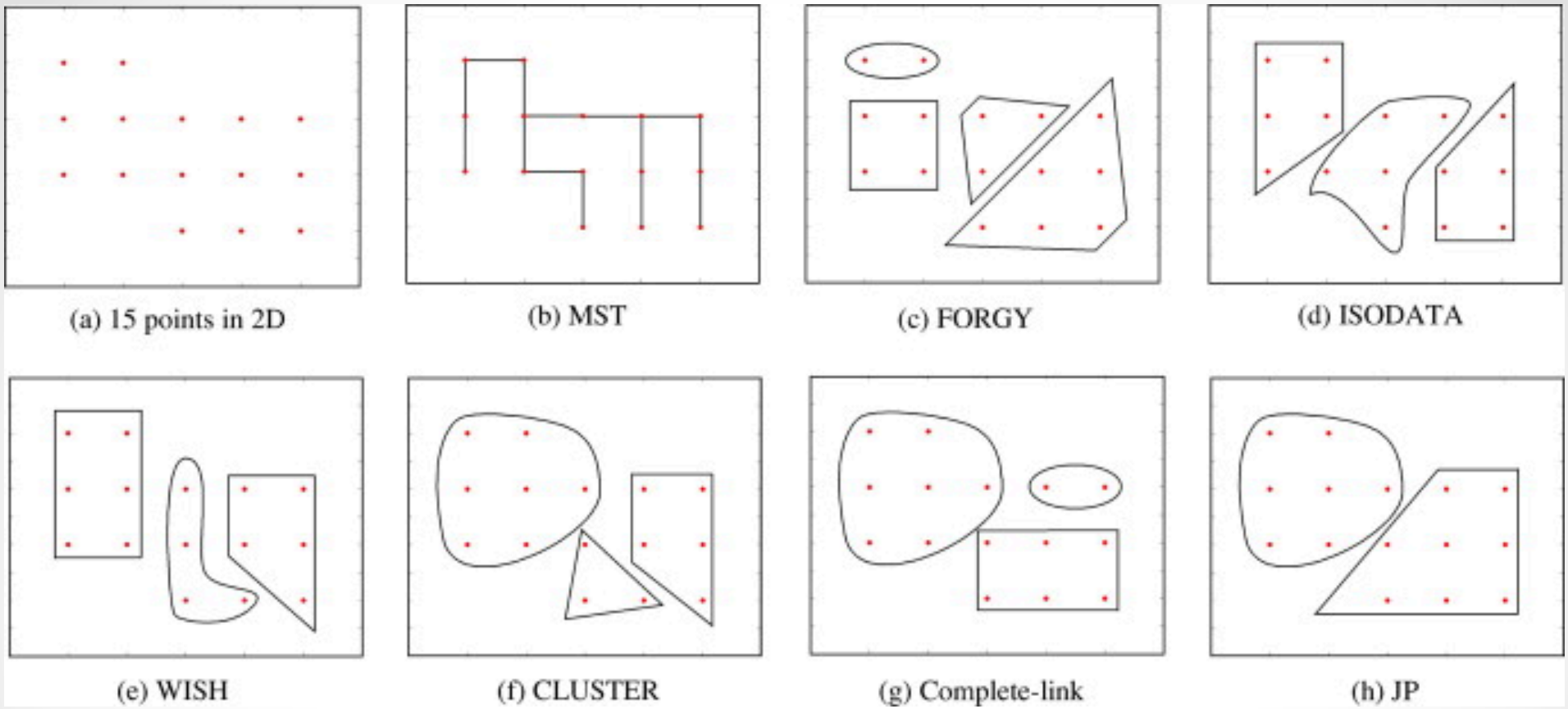


Fig. 9.
Several clusterings of fifteen patterns in two dimensions: (a) fifteen patterns; (b) minimum spanning tree of the fifteen patterns; (c) clusters from FORGY; (d) clusters from ISODATA; (e) clusters from WISH; (f) clusters from CLUSTER; (g) clusters from complete-link hierarchical clustering; and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

Paper 2

"Dynamic hierarchical algorithms for document clustering"

by Reynaldo Gil-García, Aurora Pons-Porrata

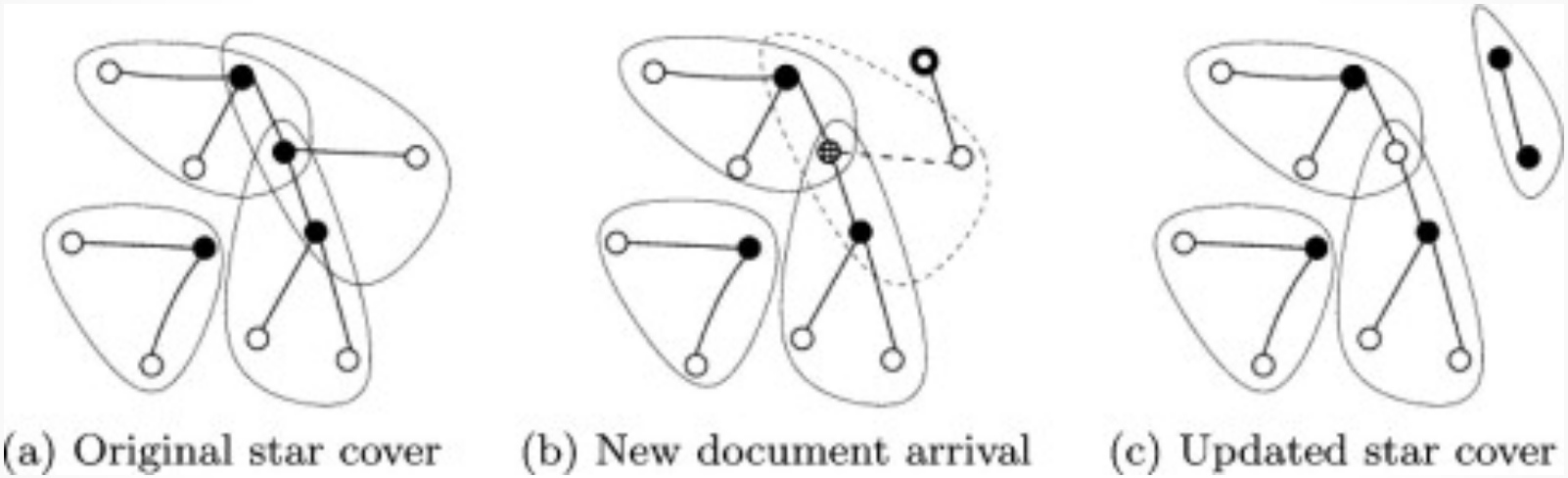


Fig. 4.
Star cover updating (black circles represent the stars).

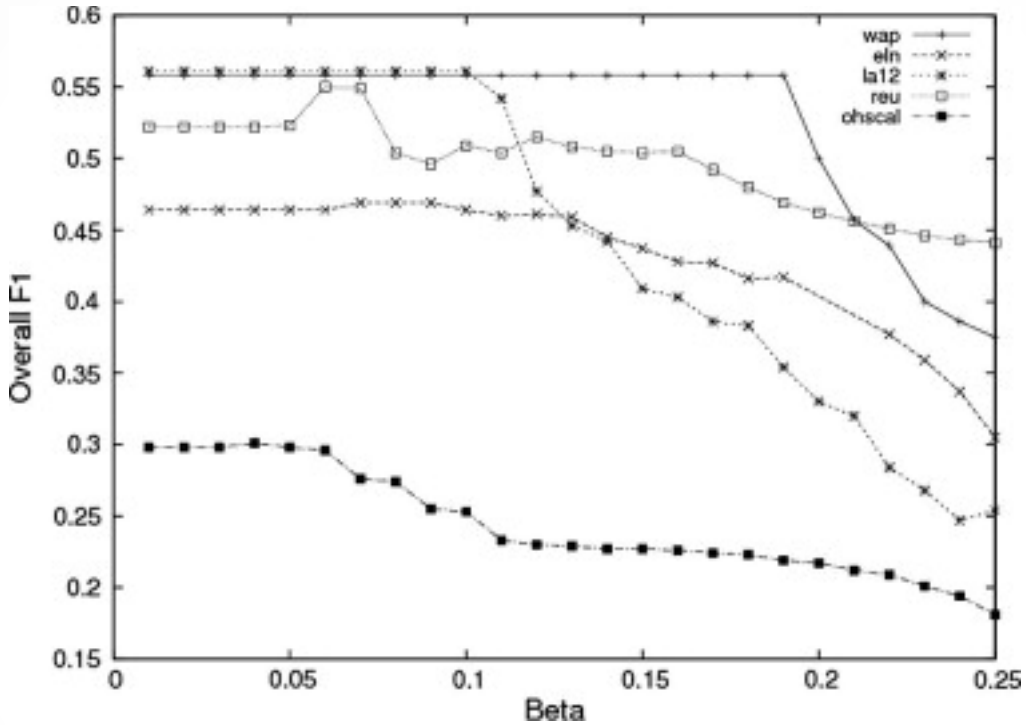


Fig. 5.
DHC sensitivity to β.

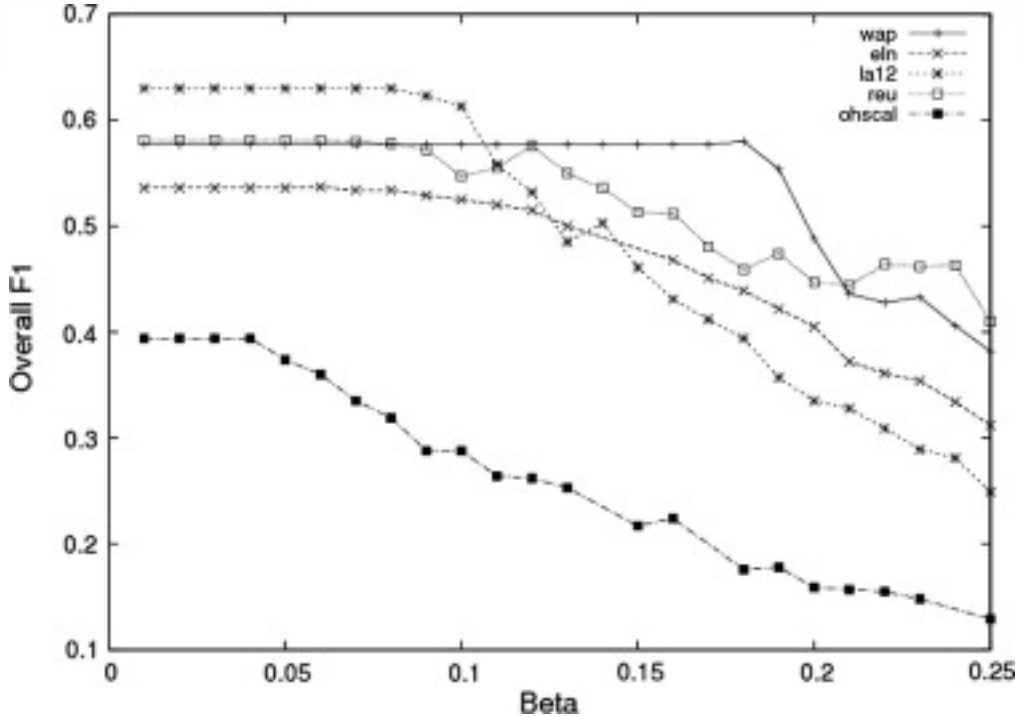


Fig. 6.
DHS sensitivity to β.

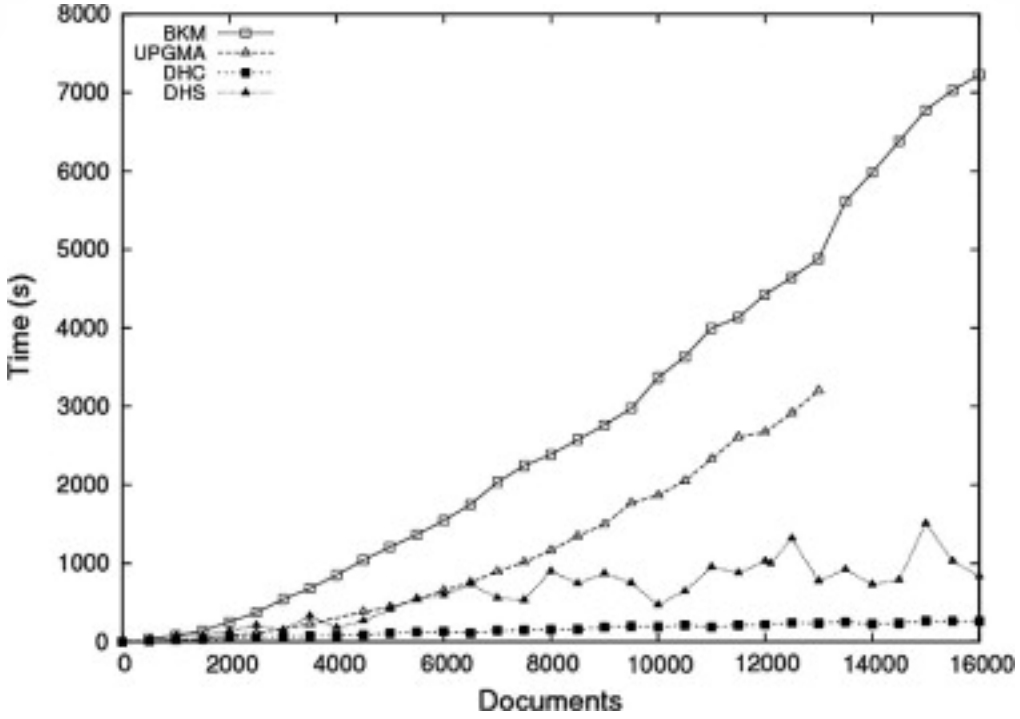


Fig. 7.
Time performance.

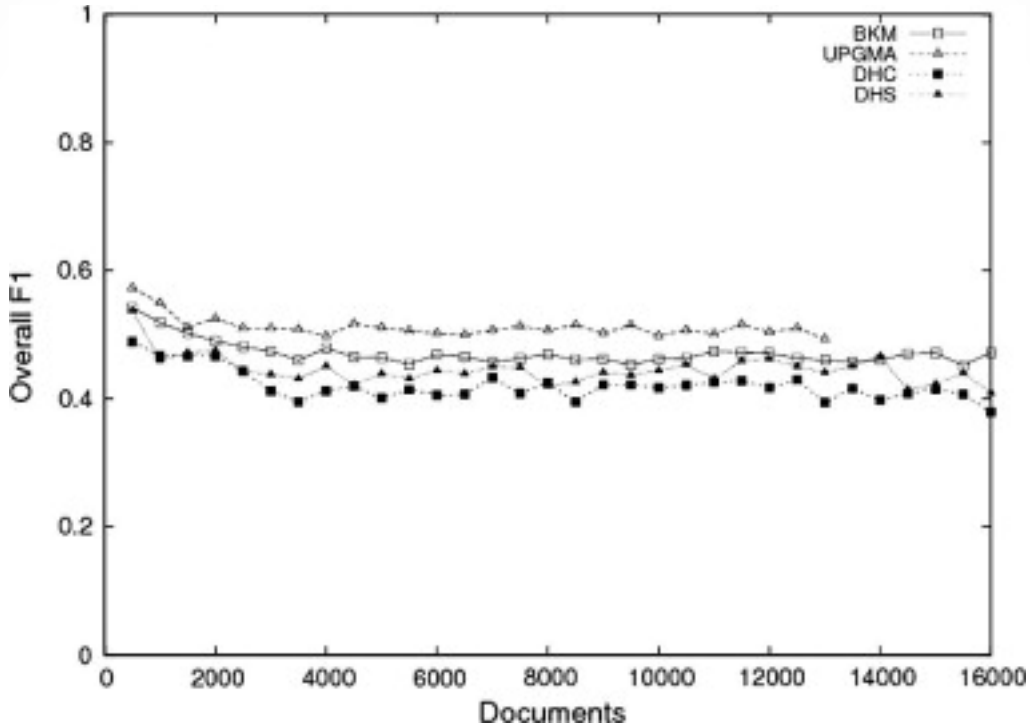


Fig. 8.
Clustering quality in a dynamic environment.