Clustering

Informal goal

- Given set of objects and measure of similarity between them, group similar objects together

- What mean by “similar”?
- What is good grouping?
- Computation time / quality tradeoff

General types of clustering

- “Soft” versus “hard” clustering
  - Hard: partition the objects
    - each object in exactly one partition
  - Soft: assign degree to which object in cluster
    - view as probability or score
- flat versus hierarchical clustering
  - hierarchical = clusters within clusters

Applications:

Many

- biology
- astronomy
- computer aided design of circuits
- information organization
- marketing
- …

Clustering in information search and analysis

- Group information objects
  - ⇒ discover topics
  - ? other groupings desirable
- Clustering versus classifying
  - classifying: have pre-determined classes with example members
  - clustering:
    - get groups of similar objects
    - added problem of labeling clusters by topic
      - e.g. common terms within cluster of docs.

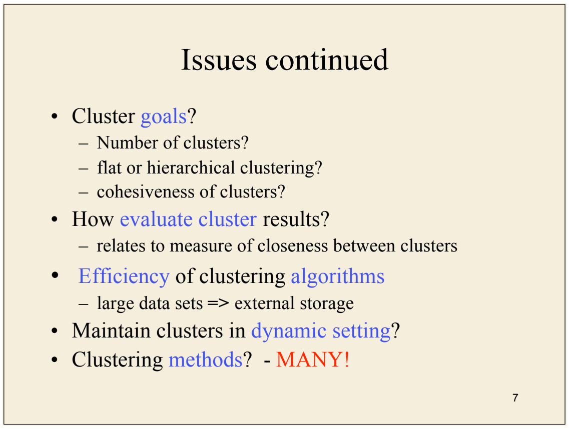
Example applications in search

- Query evaluation: cluster pruning (§7.1.6)
  - cluster all documents
  - choose representative for each cluster
  - evaluate query w.r.t. cluster reps.
  - evaluate query for docs in cluster(s) having most similar cluster rep.(s)
- Results presentation: labeled clusters
  - cluster only query results
  - e.g. Clusty.com (metasearch)

hard / soft? flat / hier?

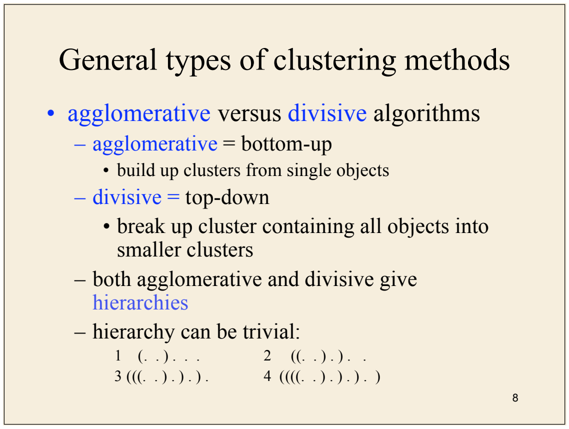
Issues

- What attributes represent items for clustering purposes?
- What is measure of similarity between items?
  - General objects and matrix of pairwise similarities
  - Objects with specific properties that allow other specifications of measure
    - Most common: Objects are d-dimensional vectors
      - » Euclidean distance
      - » cosine similarity
- What is measure of similarity between clusters?



Issues continued

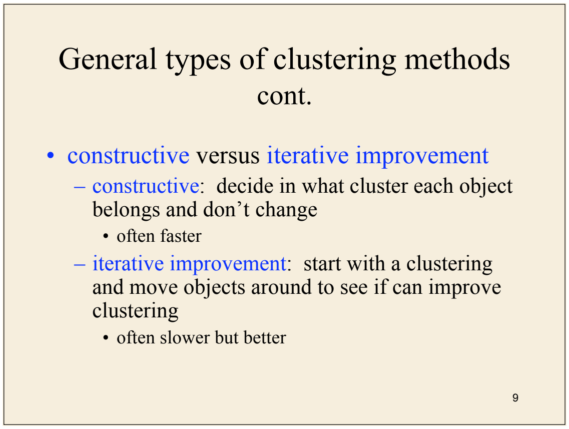
- Cluster goals?
  - Number of clusters?
  - flat or hierarchical clustering?
  - cohesiveness of clusters?
- How evaluate cluster results?
  - relates to measure of closeness between clusters
- Efficiency of clustering algorithms
  - large data sets => external storage
- Maintain clusters in dynamic setting?
- Clustering methods? - MANY!



General types of clustering methods

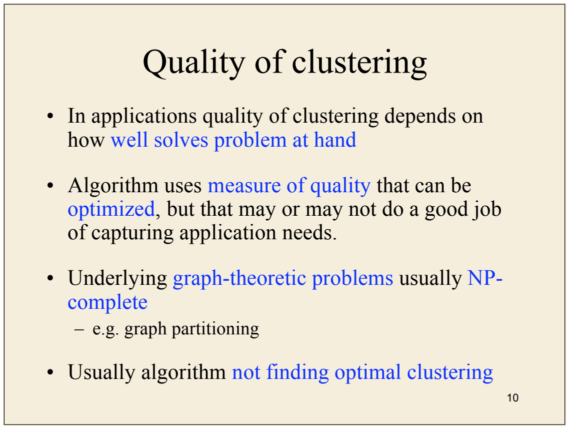
- agglomerative versus divisive algorithms
  - agglomerative = bottom-up
    - build up clusters from single objects
  - divisive = top-down
    - break up cluster containing all objects into smaller clusters
  - both agglomerative and divisive give hierarchies
  - hierarchy can be trivial:

```
1  (. . ) . . .        2  ((. . ) . ) .  .
3 (((. . ) . ) . ) .   4 ((((. . ) . ) . ) . )
```



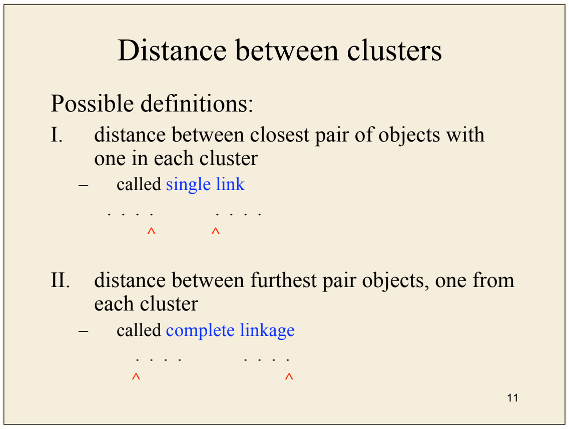
General types of clustering methods cont.

- constructive versus iterative improvement
  - constructive: decide in what cluster each object belongs and don't change
    - often faster
  - iterative improvement: start with a clustering and move objects around to see if can improve clustering
    - often slower but better



Quality of clustering

- In applications quality of clustering depends on how well solves problem at hand
- Algorithm uses measure of quality that can be optimized, but that may or may not do a good job of capturing application needs.
- Underlying graph-theoretic problems usually NP-complete
  - e.g. graph partitioning
- Usually algorithm not finding optimal clustering



Distance between clusters

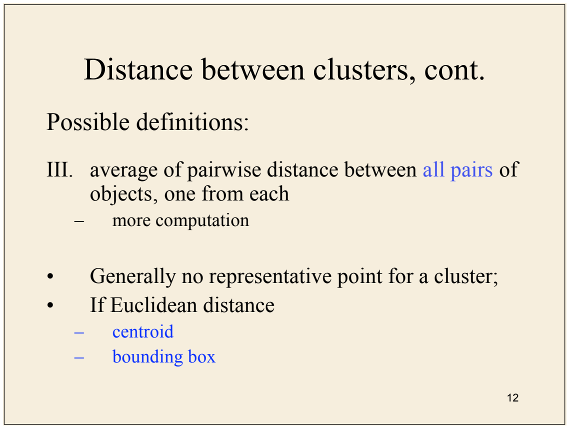
Possible definitions:

I. distance between closest pair of objects with one in each cluster
   - called single link

```
. . . .      . . . .
      ^         ^
```

II. distance between furthest pair objects, one from each cluster
   - called complete linkage

```
  . . . .       . . . .
 ^                    ^
```



Distance between clusters, cont.

Possible definitions:

III. average of pairwise distance between all pairs of objects, one from each
   - more computation

- Generally no representative point for a cluster;
- If Euclidean distance
  - centroid
  - bounding box