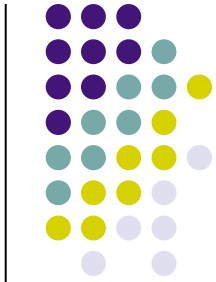
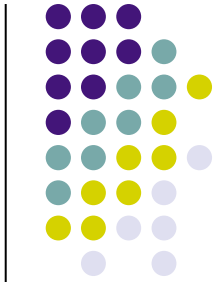


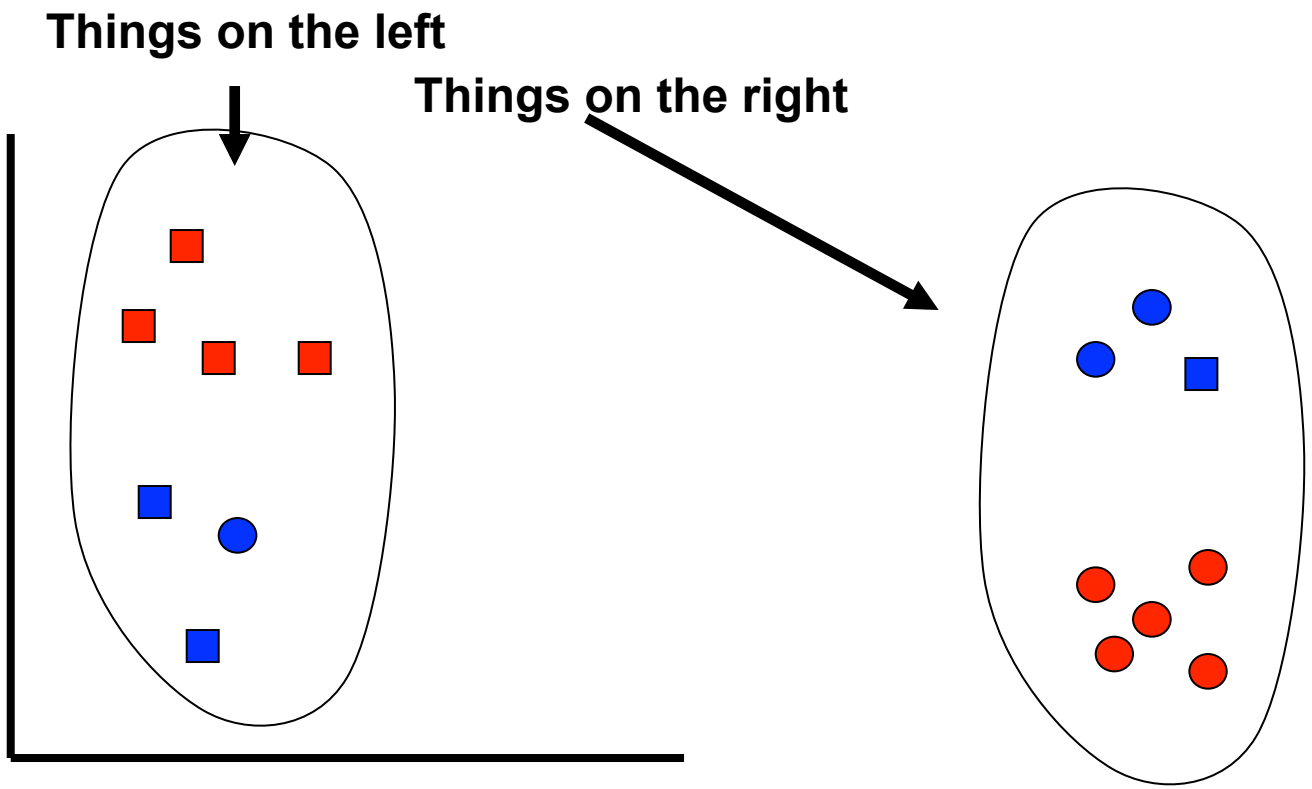
Hierarchical Clustering

adapted from:
http://www.stanford.edu/class/cs276/handouts/lecture17-clustering.ppt

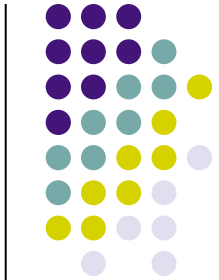
Clustering



- Grouping data into (hopefully useful) sets.

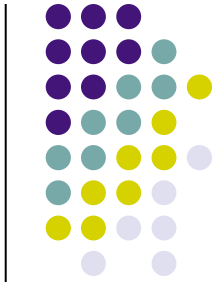


Steps in Clustering



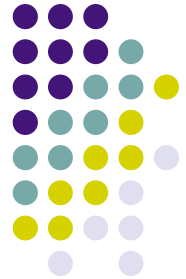
- Select Features
- Define a Proximity Measure
- Define Clustering Criterion
- Define a Clustering Algorithm
- Validate the Results
- Interpret the Results

Kinds of Clustering



- Sequential
  - Fast
  - Results depend on data order
- Cost Optimization
  - Fixed number of clusters (typically)
- Hierarchical
  - Start with many clusters
  - join clusters at each step

Hierarchical Clustering



- Cluster based on similarities/distances
- Distance measure between instances $\boldsymbol{x}^r$ and $\boldsymbol{x}^s$

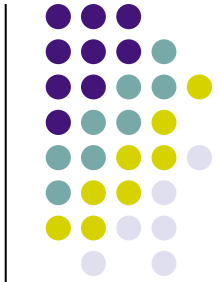
  Minkowski (L_p) (Euclidean for $p = 2$)

$$d_m\left(\mathbf{x}^r, \mathbf{x}^s\right) = \left[\sum_{j=1}^{d}\left(x_j^r - x_j^s\right)^p\right]^{1/p}$$

  City-block distance

$$d_{cb}\left(\boldsymbol{x}^r, \boldsymbol{x}^s\right) = \sum_{j=1}^{d}\left|x_j^r - x_j^s\right|$$

Agglomerative Clustering



- Start with *N* groups each with one instance
- Merging similar groups to form larger groups until there is a single one

Divisive Clustering

- Start with *a single group*
- Divide large groups into smaller groups until each group contains a single instance

Agglomerative Clustering

- Start with *N* groups each with one instance
- At each iteration, choose the 2 closest groups to merge
- Continue until all data is in 1 cluster
- Distance between two groups (clusters) $\mathcal{G}_i$ and $\mathcal{G}_j$:
  - Single-link:

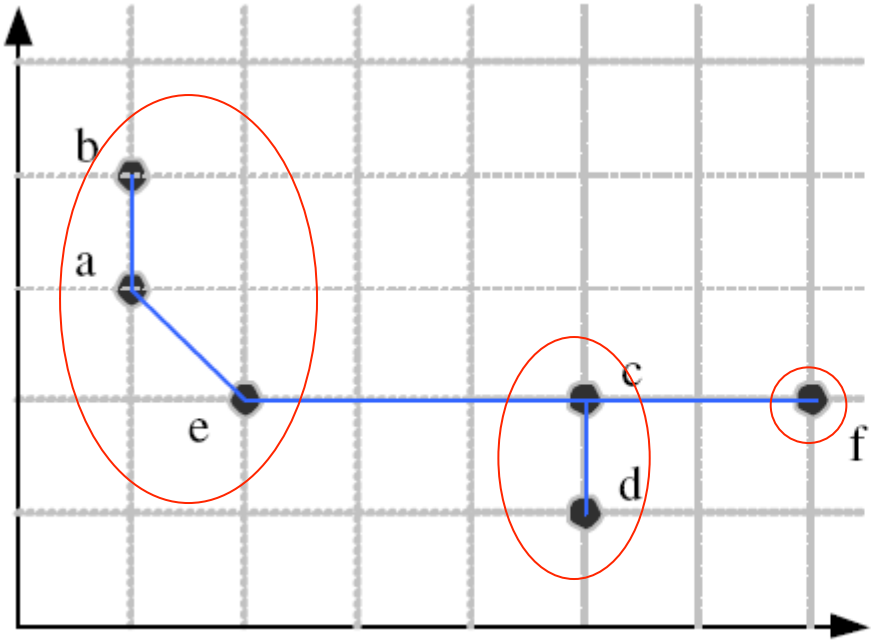
$$d(\mathcal{G}_i, \mathcal{G}_j) = \min_{\boldsymbol{x}^r \in \mathcal{G}_i, \boldsymbol{x}^s \in \mathcal{G}_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

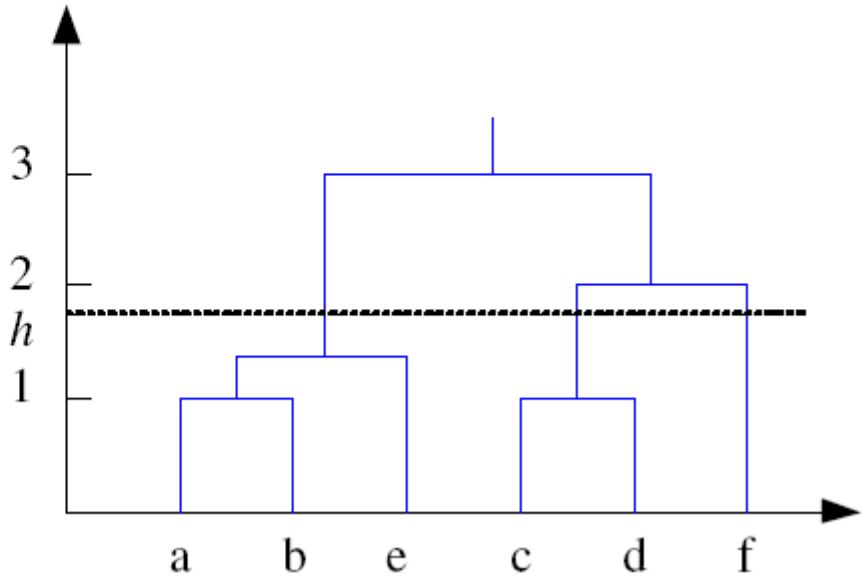
  - Complete-link:

$$d(\mathcal{G}_i, \mathcal{G}_j) = \max_{\boldsymbol{x}^r \in \mathcal{G}_i, \boldsymbol{x}^s \in \mathcal{G}_j} d(\boldsymbol{x}^r, \boldsymbol{x}^s)$$

  - Average-link
  - centroid

Example: Single-Link Clustering

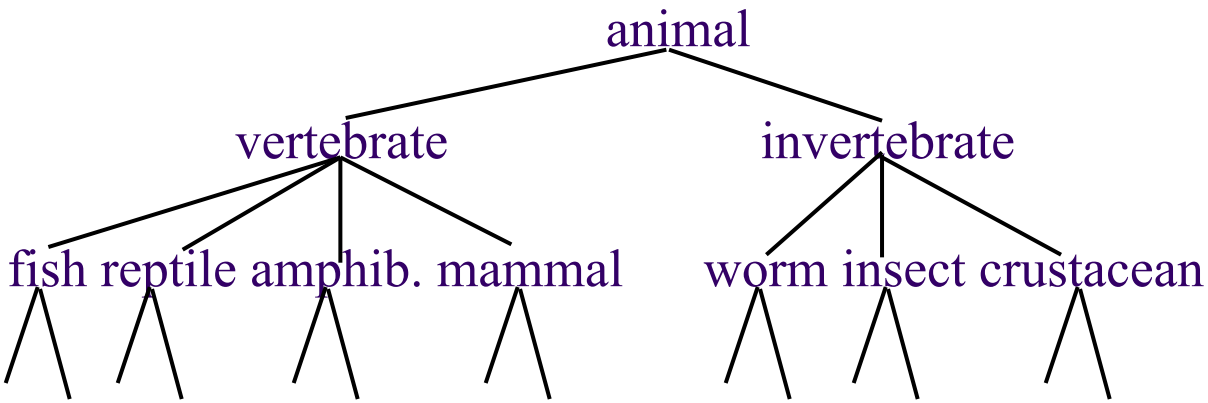




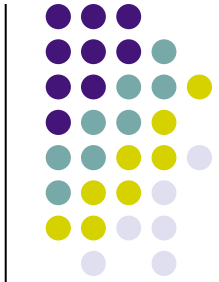
Dendrogram

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.



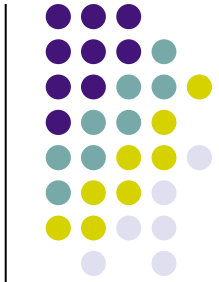
Cosine Similarity



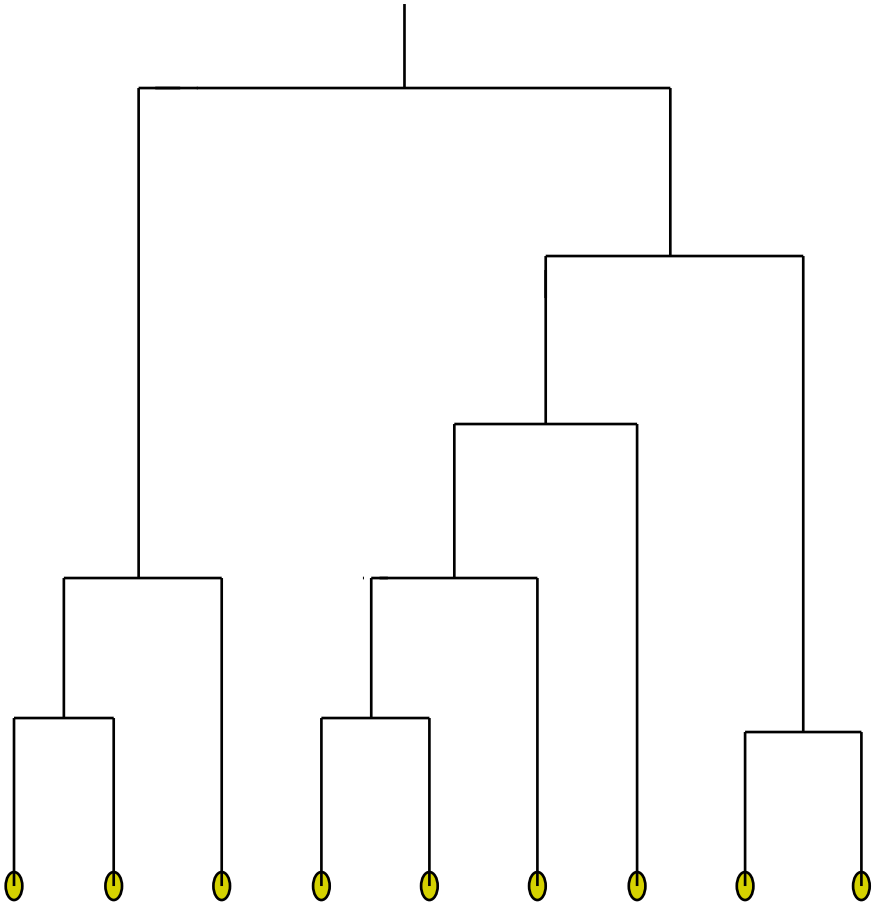
the *dot-product* (sum of products) of two normalized vectors is the cosine of the angle between them

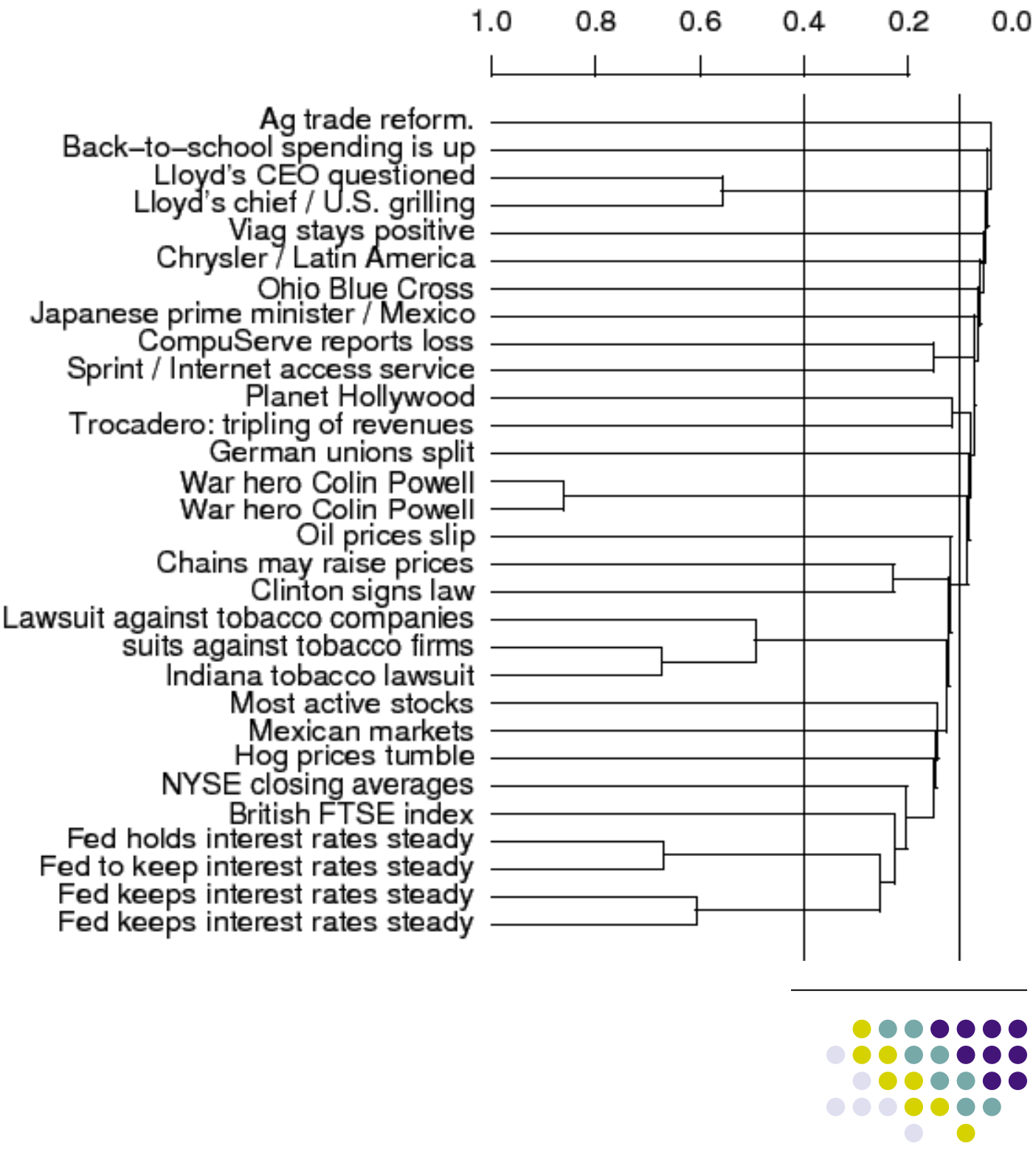
$$(d_j \cdot d_k)/(|d_j||d_k|) = \cos(\theta)$$

Dendrogram: Hierarchical Clustering

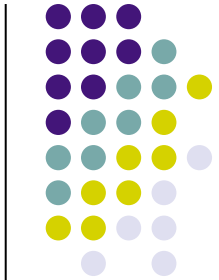


- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.




1.0
0.8
0.6
0.4
0.2
0.0
Ag trade reform.
Back–to–school spending is up
Lloyd's CEO questioned
Lloyd's chief / U.S. grilling
Viag stays positive
Chrysler / Latin America
Ohio Blue Cross
Japanese prime minister / Mexico
CompuServe reports loss
Sprint / Internet access service
Planet Hollywood
Trocadero: tripling of revenues
German unions split
War hero Colin Powell
War hero Colin Powell
Oil prices slip
Chains may raise prices
Clinton signs law
Lawsuit against tobacco companies
suits against tobacco firms
Indiana tobacco lawsuit
Most active stocks
Mexican markets
Hog prices tumble
NYSE closing averages
British FTSE index
Fed holds interest rates steady
Fed to keep interest rates steady
Fed keeps interest rates steady
Fed keeps interest rates steady

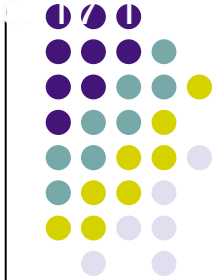
How do we choose the cut?



- Cut at a prespecified level of similarity
- Cut where the gap between two successive combination similarities is largest
- Prespecify k
- Residual sum of squares

$$K = \arg\min_{K'} [\mathrm{RSS}(K') + \lambda K']$$

Hierarchical Agglomerative Clustering (HAC)



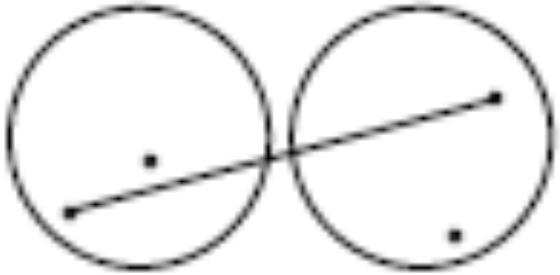
- Starts with each doc in a separate cluster
  - then repeatedly joins the *closest pair* of clusters, until there is only one cluster.
- The history of merging forms a binary tree or hierarchy.

```
SIMPLEHAC(d_1, ..., d_N)
 1  for n ← 1 to N
 2  do for i ← 1 to N
 3       do C[n][i] ← SIM(d_n, d_i)
 4       I[n] ← 1  (keeps track of active clusters)
 5  A ← []  (assembles clustering as a sequence of merges)
 6  for k ← 1 to N − 1
 7  do ⟨i, m⟩ ← arg max_{⟨i,m⟩: i≠m ∧ I[i]=1 ∧ I[m]=1} C[i][m]
 8       A.APPEND(⟨i, m⟩)  (store merge)
 9       for j ← 1 to N
10       do C[i][j] ← SIM(i, m, j)
11            C[j][i] ← SIM(i, m, j)
12       I[m] ← 0  (deactivate cluster)
13  return A
```

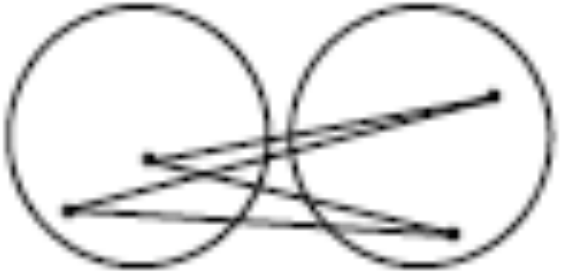
Closest pair of clusters

- Many variants to defining closest pair of clusters
- **Single-link**
  - Similarity of the *most* cosine-similar (single-link)
- **Complete-link**
  - Similarity of the "furthest" points, the *least* cosine-similar
- **Centroid**
  - Clusters whose centroids are the most cosine-similar
- **Average-link**
  - Average cosine between pairs of elements

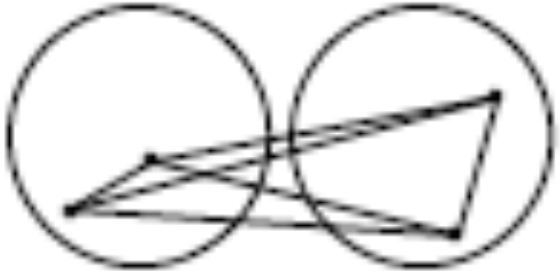
(a) single-link: maximum similarity



(b) complete-link: minimum similarity

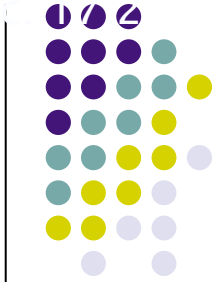


(c) centroid: average inter-similarity



(d) group-average: average of all similarities

Single Link Agglomerative Clustering



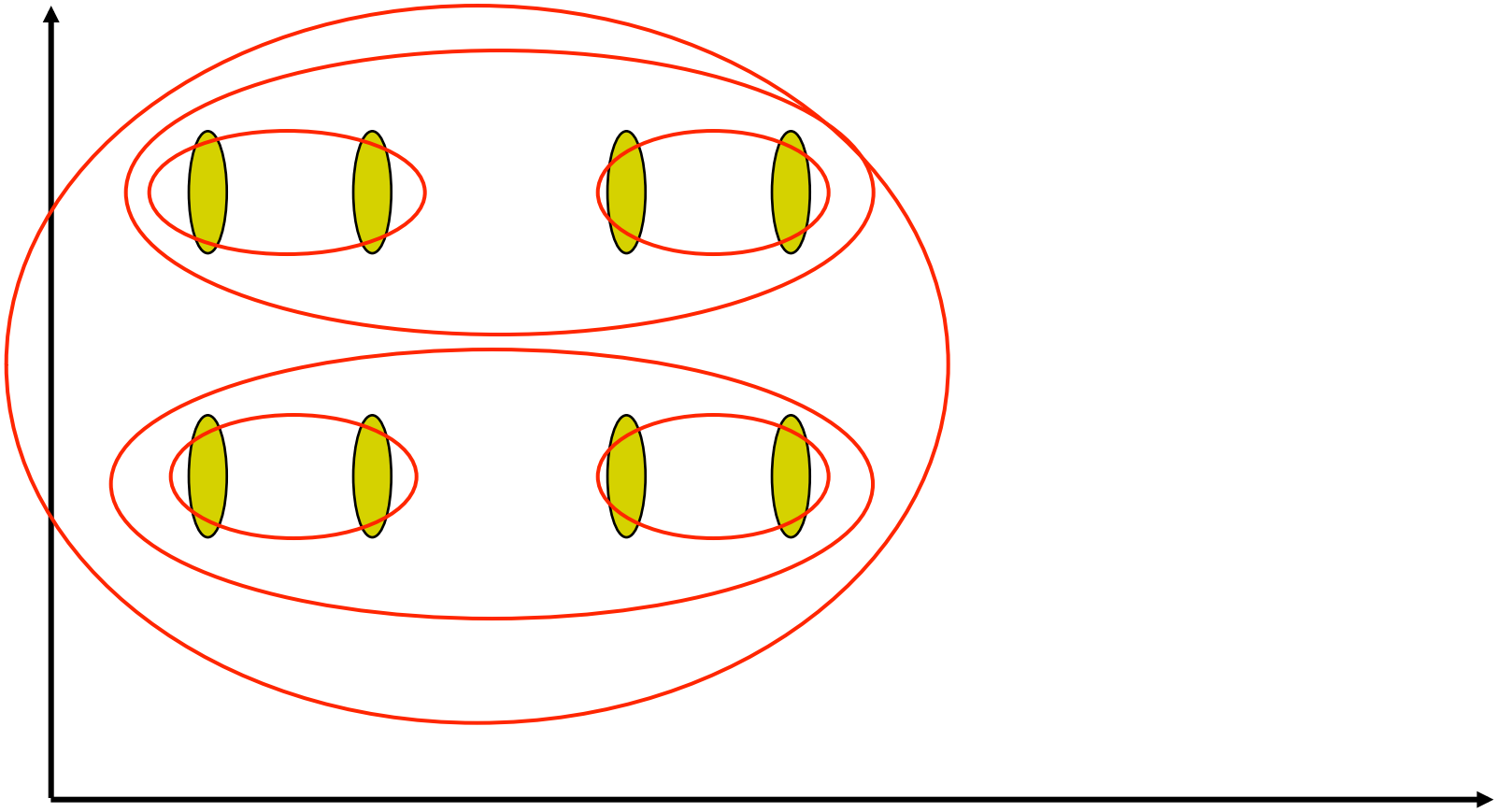
- Use maximum similarity of pairs:

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

- Can result in "straggly" (long and thin) clusters due to chaining effect.
- After merging c_i and c_j, the similarity of the resulting cluster to another cluster, c_k, is:

$$sim((c_i \cup c_j), c_k) = \max(sim(c_i, c_k), sim(c_j, c_k))$$

Single Link Example



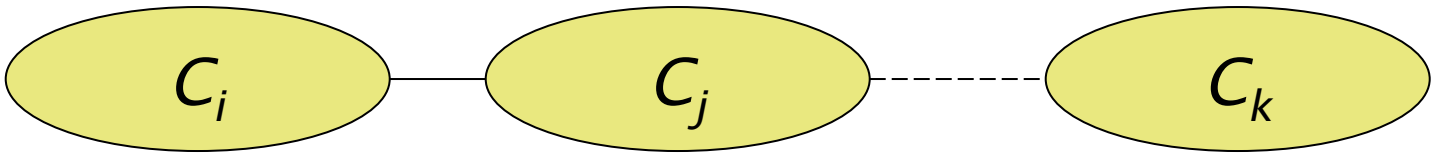
Complete Link

- Use minimum similarity of pairs:

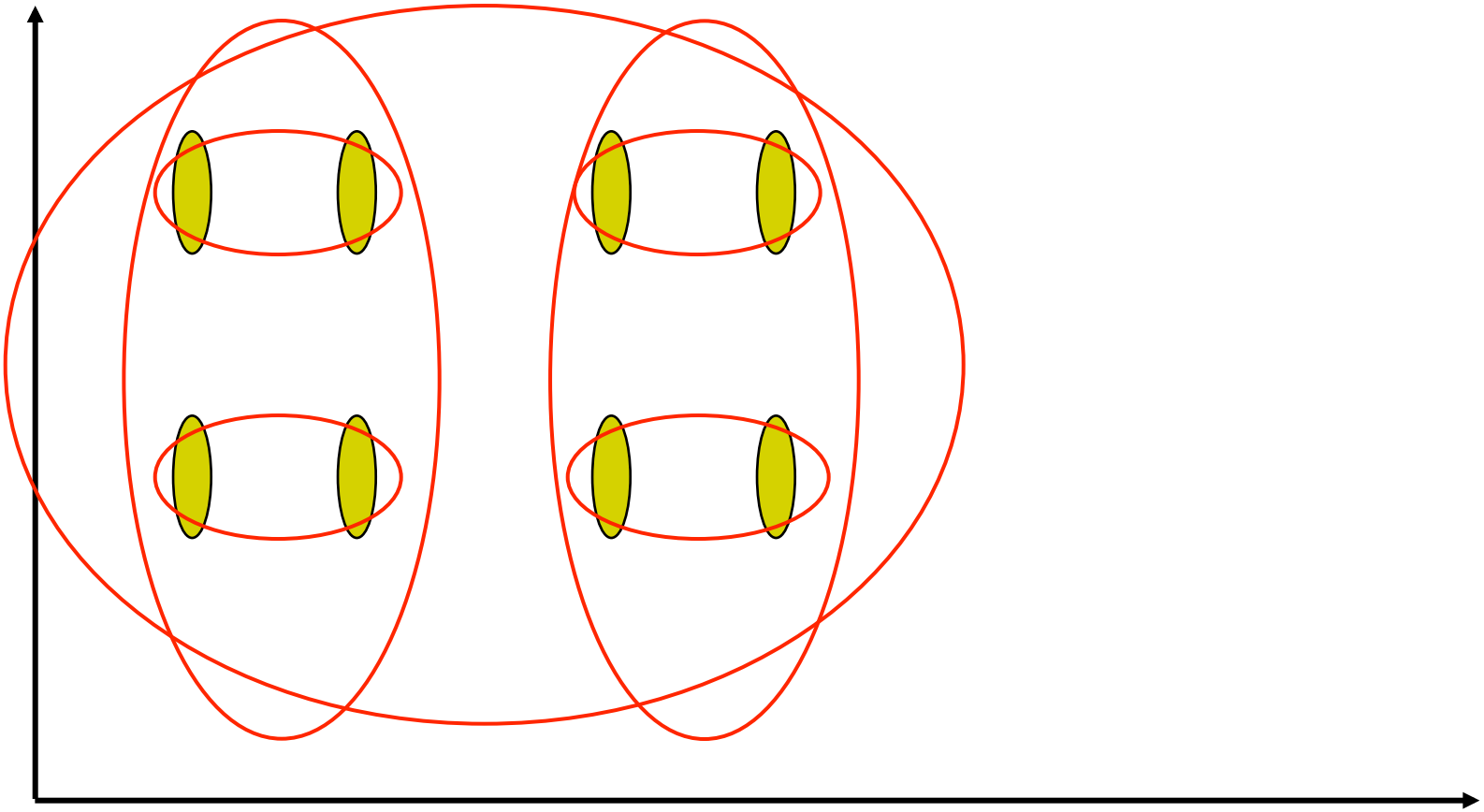
$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

- Makes "tighter," spherical clusters that are typically preferable.
- After merging c_i and c_j, the similarity of the resulting cluster to another cluster, c_k, is:

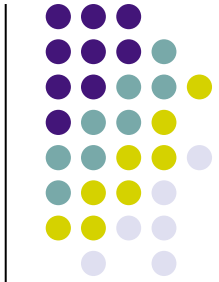
$$sim((c_i \cup c_j), c_k) = \min(sim(c_i, c_k), sim(c_j, c_k))$$



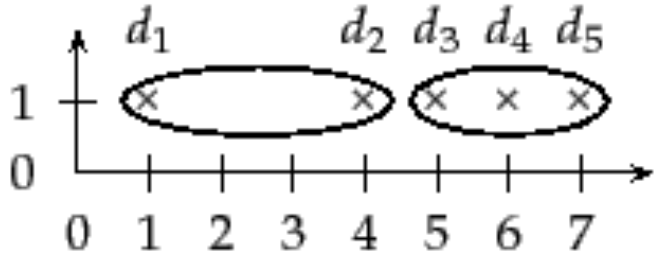
Complete Link Example

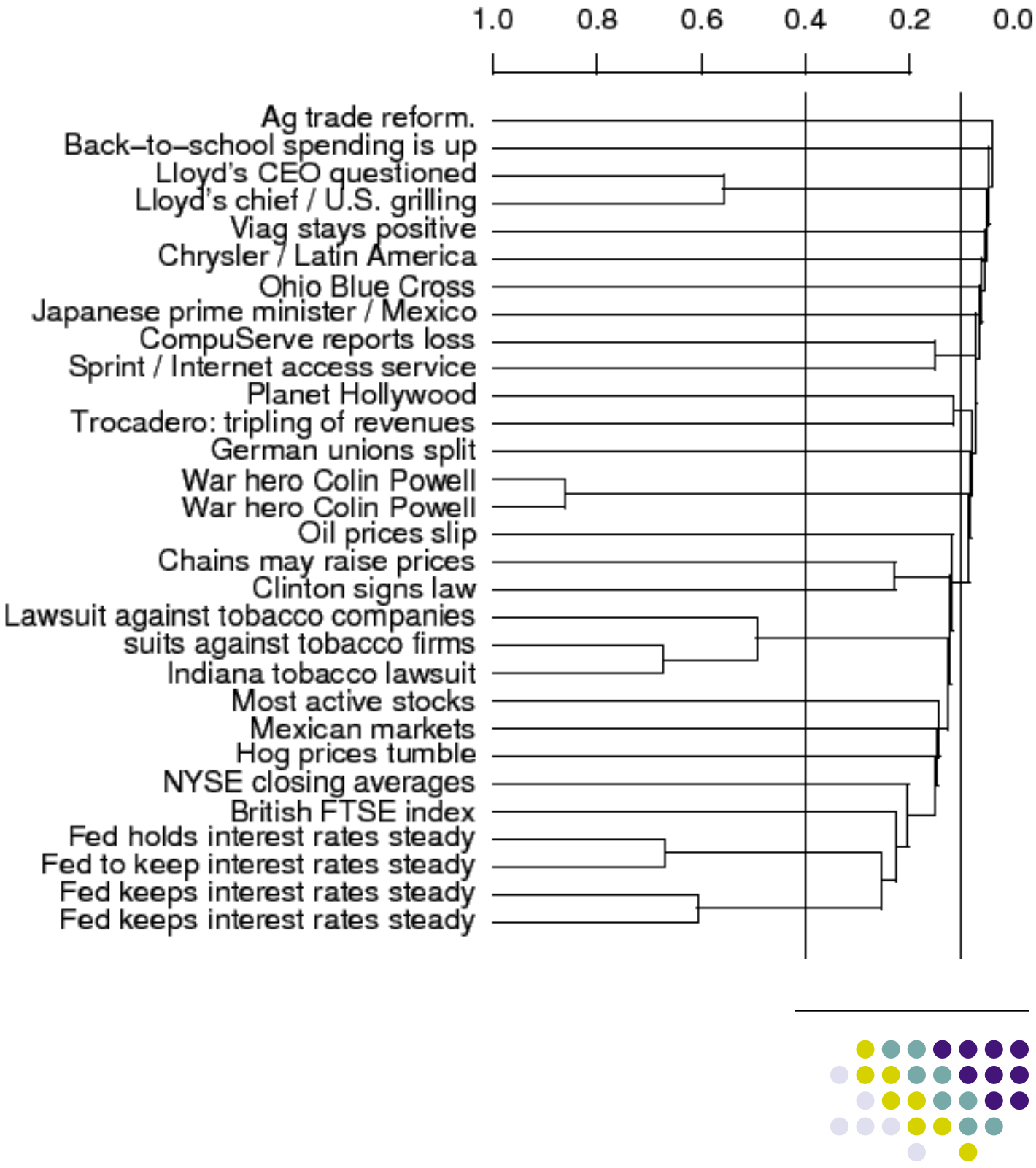


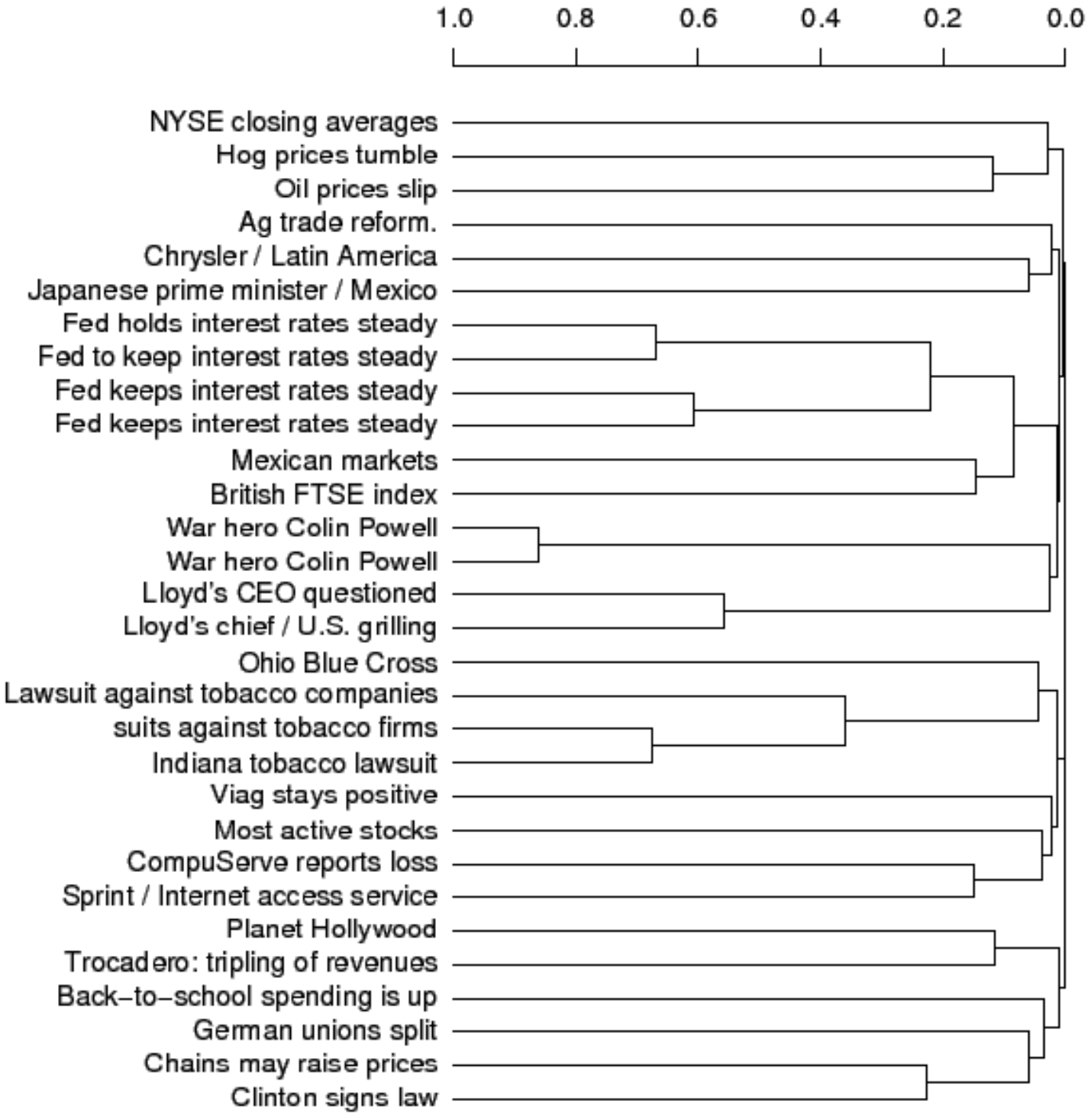
Single and Complete Link



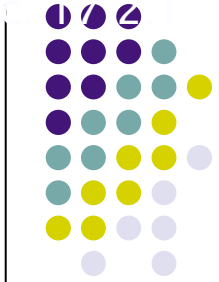
- Reduce cluster quality to similarity between a pair of docs
- Bad for single link -> chaining
- Bad for complete link -> too much attention to outliers


d_1
d_2
d_3
d_4
d_5
1
0
0 1 2 3 4 5 6 7


1.0
0.8
0.6
0.4
0.2
0.0
Ag trade reform.
Back–to–school spending is up
Lloyd's CEO questioned
Lloyd's chief / U.S. grilling
Viag stays positive
Chrysler / Latin America
Ohio Blue Cross
Japanese prime minister / Mexico
CompuServe reports loss
Sprint / Internet access service
Planet Hollywood
Trocadero: tripling of revenues
German unions split
War hero Colin Powell
War hero Colin Powell
Oil prices slip
Chains may raise prices
Clinton signs law
Lawsuit against tobacco companies
suits against tobacco firms
Indiana tobacco lawsuit
Most active stocks
Mexican markets
Hog prices tumble
NYSE closing averages
British FTSE index
Fed holds interest rates steady
Fed to keep interest rates steady
Fed keeps interest rates steady
Fed keeps interest rates steady


1.0
0.8
0.6
0.4
0.2
0.0
NYSE closing averages
Hog prices tumble
Oil prices slip
Ag trade reform.
Chrysler / Latin America
Japanese prime minister / Mexico
Fed holds interest rates steady
Fed to keep interest rates steady
Fed keeps interest rates steady
Fed keeps interest rates steady
Mexican markets
British FTSE index
War hero Colin Powell
War hero Colin Powell
Lloyd's CEO questioned
Lloyd's chief / U.S. grilling
Ohio Blue Cross
Lawsuit against tobacco companies
suits against tobacco firms
Indiana tobacco lawsuit
Viag stays positive
Most active stocks
CompuServe reports loss
Sprint / Internet access service
Planet Hollywood
Trocadero: tripling of revenues
Back–to–school spending is up
German unions split
Chains may raise prices
Clinton signs law

Computational Complexity



- In the first iteration, all HAC methods need to compute similarity of all pairs of N initial instances, which is $O(N^2)$.
- In each of the subsequent N-2 merging iterations, compute the distance between the most recently created cluster and all other existing clusters.
- In order to maintain an overall $O(N^2)$ performance, computing similarity to each other cluster must be done in constant time.
  - Often $O(N^3)$ if done in a naïve way
  - or $O(N^2 \log N)$ if done in a more clever way

Group Average

- Similarity of two clusters = average similarity of all pairs within merged cluster.

$$sim(c_i, c_j) = \frac{1}{|c_i \cup c_j|(|c_i \cup c_j| - 1)} \sum_{\vec{x} \in (c_i \cup c_j)} \sum_{\vec{y} \in (c_i \cup c_j): \vec{y} \neq \vec{x}} sim(\vec{x}, \vec{y})$$

- Compromise between single and complete link.
- Two options:
  - A) Averaged across all ordered pairs in the merged cluster
  - B) Averaged over all pairs *between* the two original clusters

Computing Group Average Similarity

$$sim(c_i, c_j) = \frac{1}{\left|c_i \cup c_j\right|\left(\left|c_i \cup c_j\right| - 1\right)} \sum_{\vec{x} \in (c_i \cup c_j)} \sum_{\vec{y} \in (c_i \cup c_j): \vec{y} \neq \vec{x}} sim(\vec{x}, \vec{y})$$

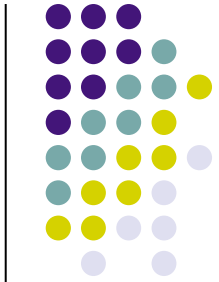
- Always maintain sum of vectors in each cluster.

$$\vec{s}(c_j) = \sum_{\vec{x} \in c_j} \vec{x}$$

- Compute similarity of clusters in constant time:

$$sim(c_i, c_j) = \frac{(\vec{s}(c_i) + \vec{s}(c_j)) \bullet (\vec{s}(c_i) + \vec{s}(c_j)) - (|c_i| + |c_j|)}{(|c_i| + |c_j|)(|c_i| + |c_j| - 1)}$$

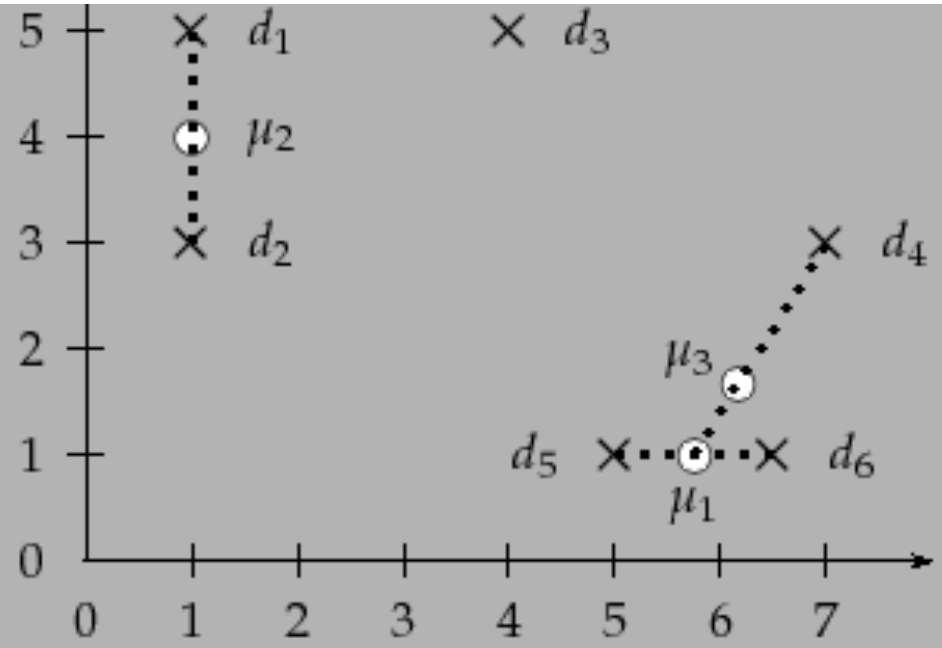
Centroid Clustering



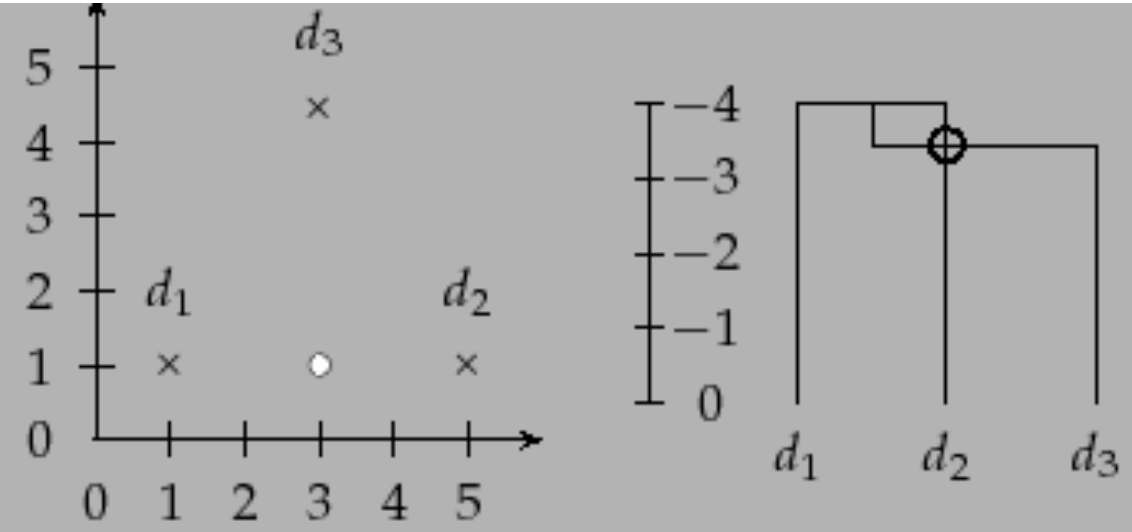
- Similarity of two clusters is defined as the similarity of their centroids

$$\begin{aligned}\text{SIM-CENT}(\omega_i, \omega_j) &= \vec{\mu}(\omega_i) \cdot \vec{\mu}(\omega_j) \\ &= \left(\frac{1}{N_i} \sum_{\vec{d}_m \in \omega_i} \vec{d}_m\right) \cdot \left(\frac{1}{N_j} \sum_{\vec{d}_n \in \omega_j} \vec{d}_n\right) \\ &= \frac{1}{N_i N_j} \sum_{\vec{d}_m \in \omega_i} \sum_{\vec{d}_n \in \omega_j} \vec{d}_m \cdot \vec{d}_n\end{aligned}$$

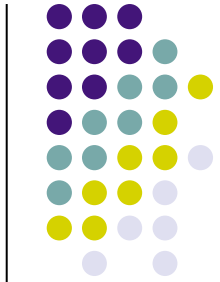
- Equivalent to the average similarity of all pairs of documents from different clusters



► **Figure 17.8** Three iterations of centroid clustering. Each iteration merges the two clusters whose centroids are closest.

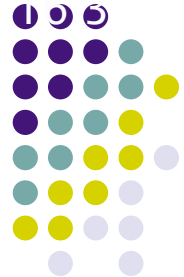


▶ **Figure 17.9** Centroid clustering is not monotonic. The documents d_1 at $(1+\epsilon, 1)$, d_2 at $(5, 1)$, and d_3 at $(3, 1+2\sqrt{3})$ are almost equidistant, with d_1 and d_2 closer to each other than to d_3. The non-monotonic inversion in the hierarchical clustering of the three points appears as an intersecting merge line in the dendrogram. The intersection is circled.



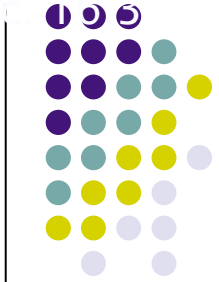
method	combination similarity	time compl.	optimal?	comment
single-link	max inter-similarity of any 2 docs	$\Theta(N^2)$	yes	chaining effect
complete-link	min inter-similarity of any 2 docs	$\Theta(N^2 \log N)$	no	sensitive to outliers
group-average	average of all sims	$\Theta(N^2 \log N)$	no	best choice for most applications
centroid	average inter-similarity	$\Theta(N^2 \log N)$	no	inversions can occur

What Is A Good Clustering?



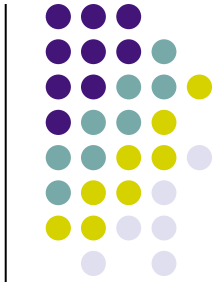
- Internal criterion: A good clustering will produce high quality clusters in which:
  - the intra-class (that is, intra-cluster) similarity is high
  - the inter-class similarity is low
  - The measured quality of a clustering depends on both the document representation and the similarity measure used

External criteria for clustering quality



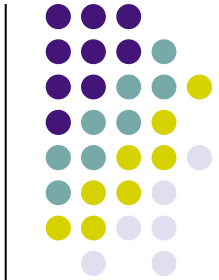
- Quality measured by its ability to discover some or all of the hidden patterns or latent classes in gold standard data
- Assesses a clustering with respect to ground truth … requires *labeled data*

Divisive Clustering

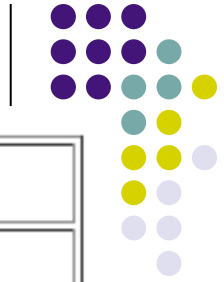


- Top down clustering
- ALG
  - Start with all documents in 1 cluster
  - Split using a flat clustering algorithm
  - Apply recursively until each doc is in its own cluster
- More efficient
- Benefits from complete info

Cluster Labeling



- Differential Cluster Labeling
- Cluster-internal Labeling



		labeling method		
	# docs	centroid	mutual information	title
4	622	oil plant mexico production crude **power 000 refinery gas** bpd	plant oil production **barrels** crude bpd mexico **dolly capacity petroleum**	MEXICO: Hurricane Dolly heads for Mexico coast
9	1017	police security **russian** people military peace killed told **grozny court**	police killed military security peace told **troops forces rebels** people	RUSSIA: Russia's Lebed meets rebel chief in Chechnya
10	1259	00 000 tonnes traders futures wheat prices **cents september** tonne	**delivery** traders futures tonne tonnes **desk** wheat prices 000 00	USA: Export Business - Grain/oilseeds complex