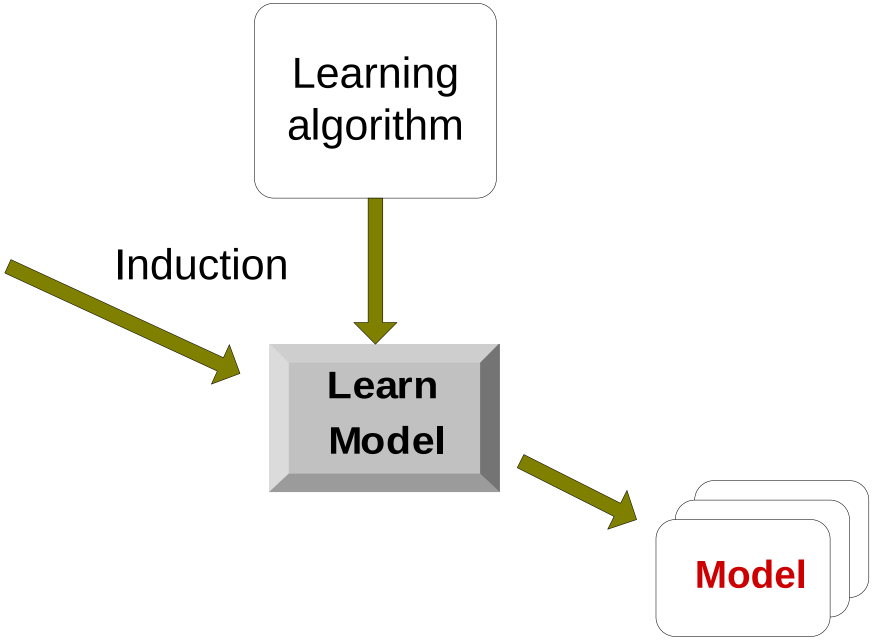
Classification

Model Training

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



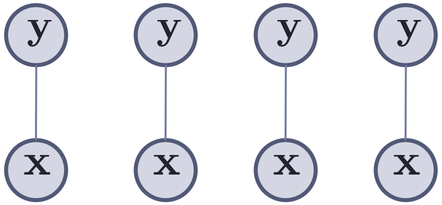
Classification

Imagine an agent or machine which experiences a series of sensory inputs:

$x_1, x_2, x_3, x_4, \ldots$

Supervised learning（监督学习）:

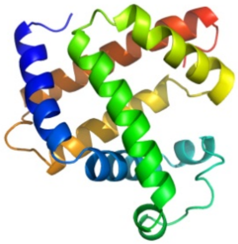
The machine is also given desired outputs $y_1, y_2, \ldots,$ and its goal is to learn to produce the correct output given a new input.

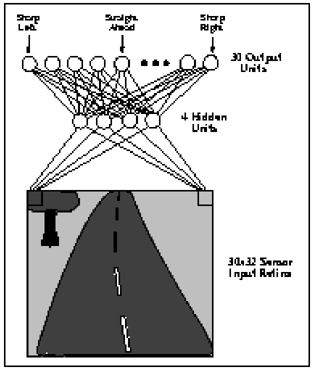


Supervised learning has many successes

- Document classification
- Protein prediction
- Face recognition
- Speech recognition
- Vehicle steering

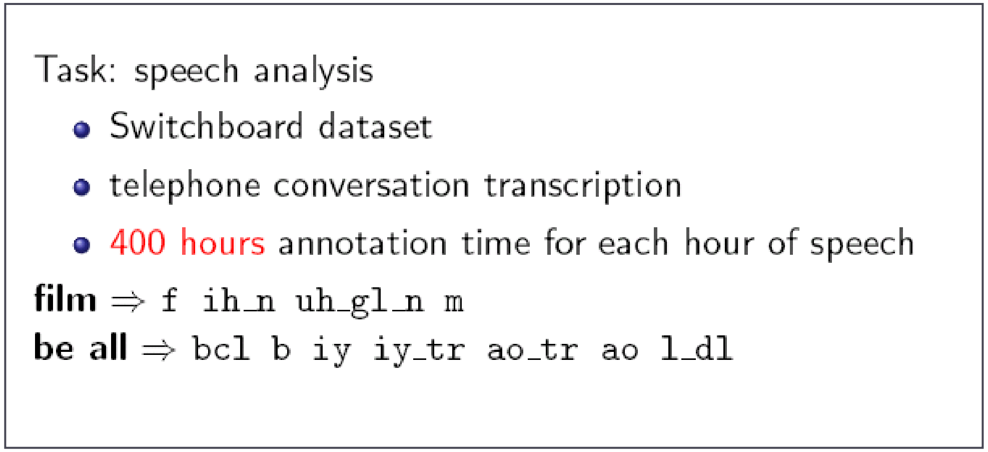
etc.





However…

- Labeled data can be rare or expensive in many real applications
  - Speech
  - Medical data
  - Protein
  - …



Task: speech analysis

- Switchboard dataset
- telephone conversation transcription
- 400 hours annotation time for each hour of speech

film ⇒ `f ih_n uh_gl_n m`

be all ⇒ `bcl b iy iy_tr ao_tr ao l_dl`

- Unlabeled data is much cheaper and abundant

Question: Can we use unlabeled data to help?

Can we use unlabeled data to help?

- Unlabeled data is missing important information…

- But maybe still has useful regularities that we can use.

Unsupervised learning

Web Content Mining: Clustering

From supervised to unsupervised classification

Outline

- Introduction to clustering

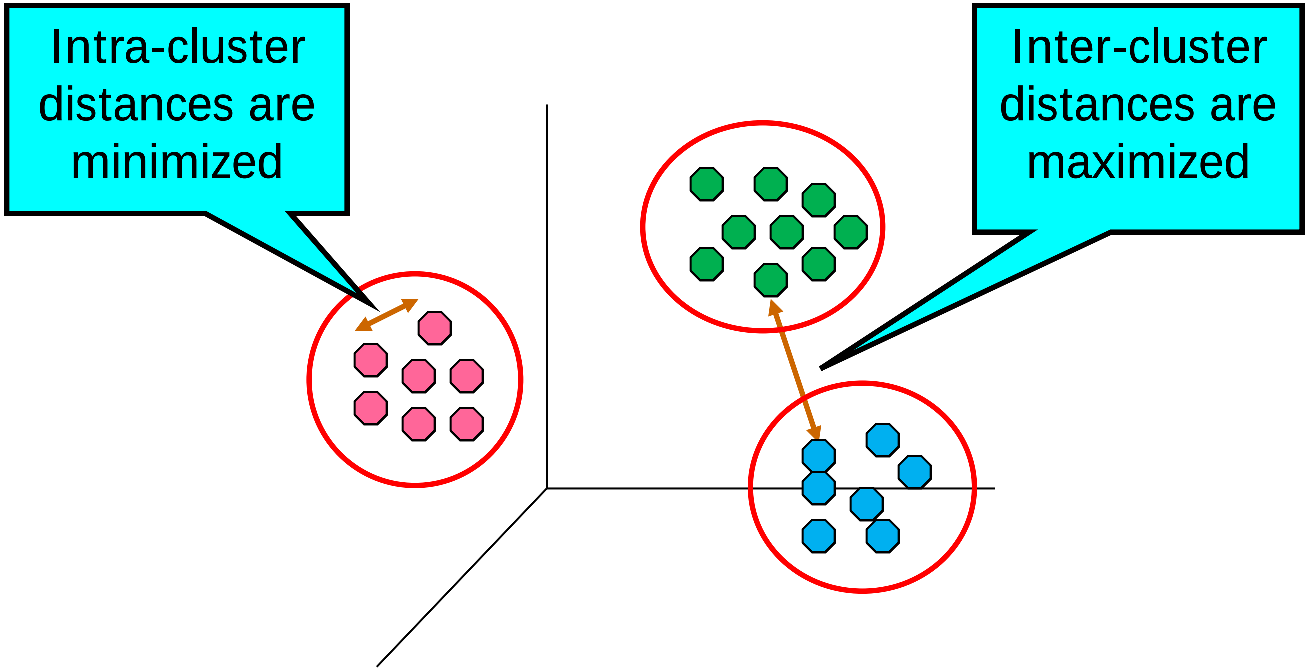
- Clustering methods
  - Hierarchical clustering
  - K-means

- Evaluation of clustering

What is clustering?

- Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects
  - It is the most common form of unsupervised learning
  - A common and important task that finds many applications in Web mining and other places

Clustering



Intra-cluster distances are minimized

Inter-cluster distances are maximized

Are there any “groups” in the data?

What is each group?

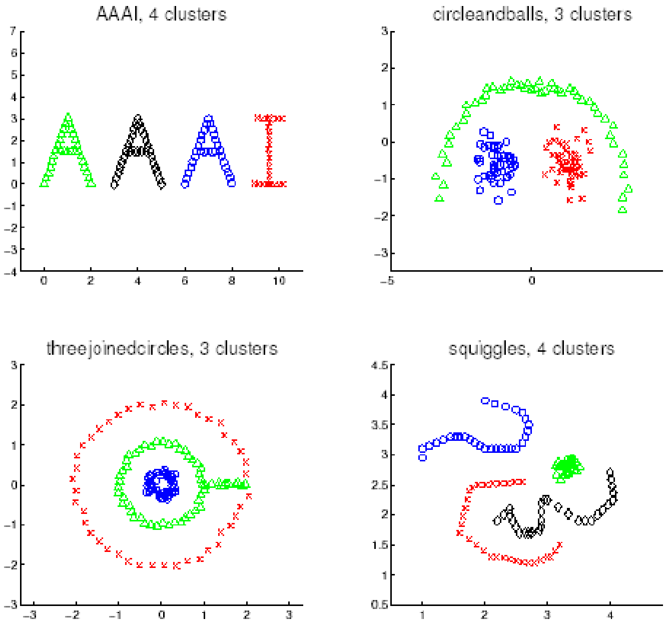
How many?

How to identify them?

Clustering

- Group the data objects into subsets or "clusters":
  - High similarity within clusters
  - Low similarity between clusters

- A common and important task that finds many applications in Science, Engineering, information Science, and other places
  - Group genes that perform the same function
  - Group individuals that has similar political view
  - Categorize documents of similar topics
  - Identify similar objects from pictures



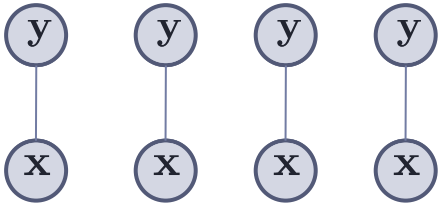
Classification vs. Clustering

Imagine an agent or machine which experiences a series of sensory inputs:

$x_1, x_2, x_3, x_4, \ldots$

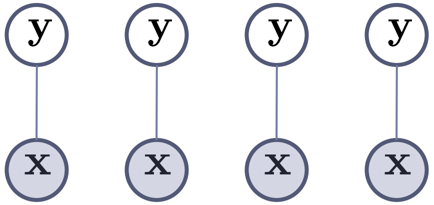
Supervised learning（监督学习）：

The machine is also given desired outputs $y_1, y_2, \ldots$, and its goal is to learn to produce the correct output given a new input.



Unsupervised learning（无监督学习）：

outputs $y_1, y_2, \ldots$ Not given, the agent still wants to build a model of x that can be used for reasoning, decision making, predicting things, communicating etc.



Why cluster on the Web?

- Whole corpus analysis/navigation
  - Better user interface
- For improving recall in search applications
  - Better search results
- For speeding up vector space retrieval
  - Faster search

Navigating document collections

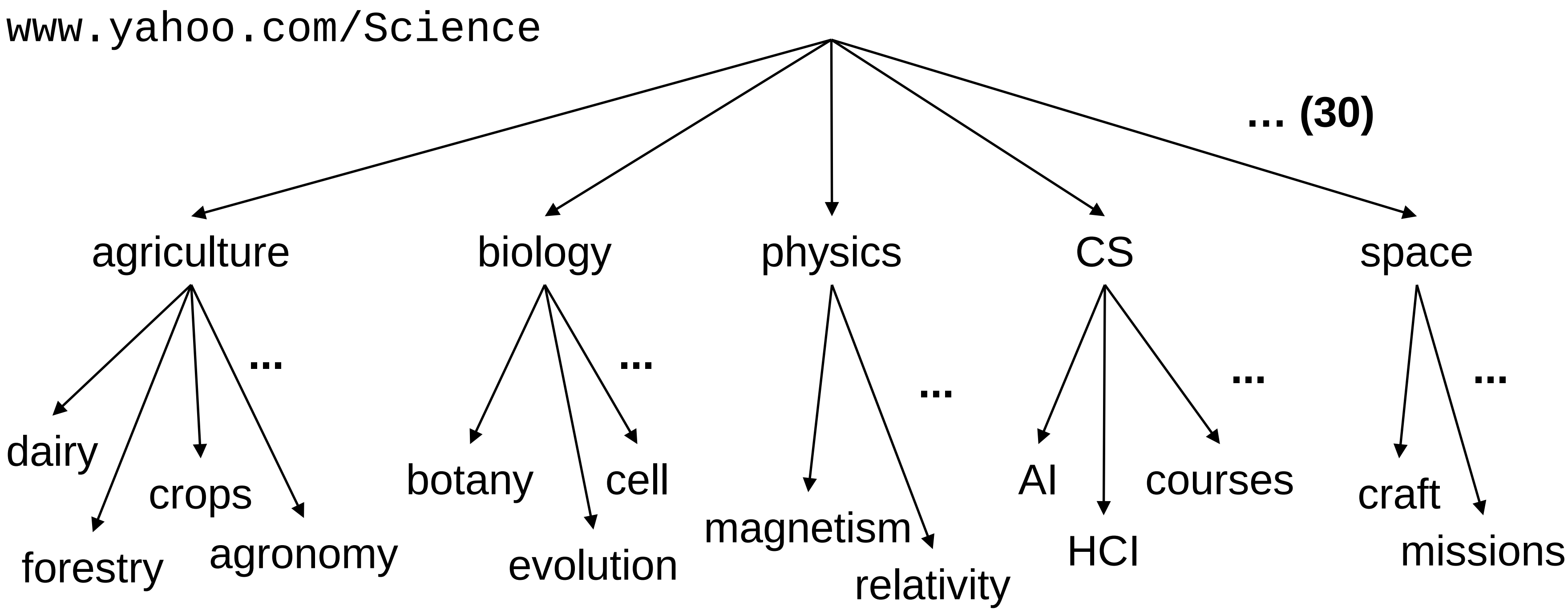
- Standard IR is like a book index
- Document clusters are like a table of contents
- People find having a table of contents useful

Index
Aardvark, 15
Blueberry, 200
Capricorn, 1, 45-55
Dog, 79-99
Egypt, 65
Falafel, 78-90
Giraffes, 45-59
…

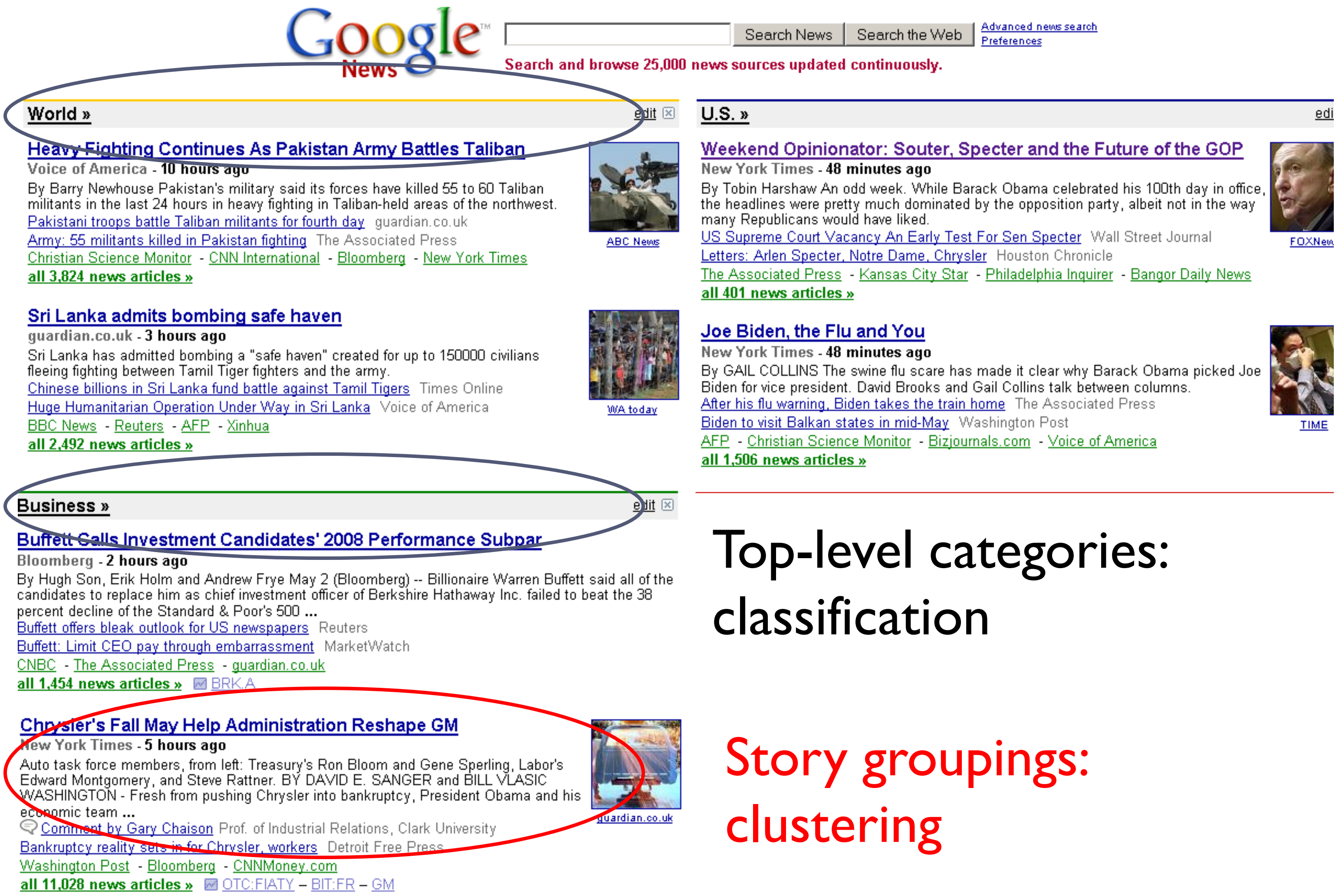
Table of Contents

1. Science of Cognition
    - 1.a. Motivations
        - 1.a.i. Intellectual Curiosity
        - 1.a.ii. Practical Applications
    - 1.b. History of Cognitive Psychology
2. The Neural Basis of Cognition
    - 2.a. The Nervous System
    - 2.b. Organization of the Brain
    - 2.c. The Visual System
3. Perception and Attention
    - 3.a. Sensory Memory
    - 3.b. Attention and Sensory Information Processing

Yahoo! Hierarchy *isn't* clustering but *is* the kind of output you want from clustering



Google News: automatic clustering gives an effective news presentation metaphor



Search and browse 25,000 news sources updated continuously.

World » edit

Heavy Fighting Continues As Pakistan Army Battles Taliban
Voice of America - **10 hours ago**
By Barry Newhouse Pakistan's military said its forces have killed 55 to 60 Taliban militants in the last 24 hours in heavy fighting in Taliban-held areas of the northwest.
Pakistani troops battle Taliban militants for fourth day guardian.co.uk
Army: 55 militants killed in Pakistan fighting The Associated Press
Christian Science Monitor - CNN International - Bloomberg - New York Times
all 3,824 news articles »

ABC News

Sri Lanka admits bombing safe haven
guardian.co.uk - **3 hours ago**
Sri Lanka has admitted bombing a "safe haven" created for up to 150000 civilians fleeing fighting between Tamil Tiger fighters and the army.
Chinese billions in Sri Lanka fund battle against Tamil Tigers Times Online
Huge Humanitarian Operation Under Way in Sri Lanka Voice of America
BBC News - Reuters - AFP - Xinhua
all 2,492 news articles »

WA today

Business » edit

Buffett Calls Investment Candidates' 2008 Performance Subpar
Bloomberg - **2 hours ago**
By Hugh Son, Erik Holm and Andrew Frye May 2 (Bloomberg) -- Billionaire Warren Buffett said all of the candidates to replace him as chief investment officer of Berkshire Hathaway Inc. failed to beat the 38 percent decline of the Standard & Poor's 500 **...**
Buffett offers bleak outlook for US newspapers Reuters
Buffett: Limit CEO pay through embarrassment MarketWatch
CNBC - The Associated Press - guardian.co.uk
all 1,454 news articles » BRK.A

Chrysler's Fall May Help Administration Reshape GM
New York Times - **5 hours ago**
Auto task force members, from left: Treasury's Ron Bloom and Gene Sperling, Labor's Edward Montgomery, and Steve Rattner. BY DAVID E. SANGER and BILL VLASIC WASHINGTON - Fresh from pushing Chrysler into bankruptcy, President Obama and his economic team **...**
Comment by Gary Chaison Prof. of Industrial Relations, Clark University
Bankruptcy reality sets in for Chrysler, workers Detroit Free Press
Washington Post - Bloomberg - CNNMoney.com
all 11,028 news articles » OTC:FIATY – BIT:FR – GM

guardian.co.uk

U.S. » edi

Weekend Opinionator: Souter, Specter and the Future of the GOP
New York Times - **48 minutes ago**
By Tobin Harshaw An odd week. While Barack Obama celebrated his 100th day in office, the headlines were pretty much dominated by the opposition party, albeit not in the way many Republicans would have liked.
US Supreme Court Vacancy An Early Test For Sen Specter Wall Street Journal
Letters: Arlen Specter, Notre Dame, Chrysler Houston Chronicle
The Associated Press - Kansas City Star - Philadelphia Inquirer - Bangor Daily News
all 401 news articles »

FOXNew

Joe Biden, the Flu and You
New York Times - **48 minutes ago**
By GAIL COLLINS The swine flu scare has made it clear why Barack Obama picked Joe Biden for vice president. David Brooks and Gail Collins talk between columns.
After his flu warning, Biden takes the train home The Associated Press
Biden to visit Balkan states in mid-May Washington Post
AFP - Christian Science Monitor - Bizjournals.com - Voice of America
all 1,506 news articles »

TIME

Top-level categories: classification

Story groupings: clustering

Why cluster on the Web?

- Whole corpus analysis/navigation
  - Better user interface
- For improving recall in search applications
  - Better search results
- For speeding up vector space retrieval
  - Faster search

For improving search recall

- *Cluster hypothesis* - Documents in the same cluster behave similarly with respect to relevance to information needs
- Therefore, to improve search recall:
  - Cluster docs in corpus a priori
  - When a query matches a doc *D*, also return other docs in the cluster containing *D*
- Hope if we do this: The query "car" will also return docs containing *automobile*
  - Because clustering grouped together docs containing *car* with those containing *automobile.*

Why might this happen?

Why cluster on the Web?

- Whole corpus analysis/navigation
  - Better user interface
- For improving recall in search applications
  - Better search results
- For speeding up vector space retrieval
  - Faster search

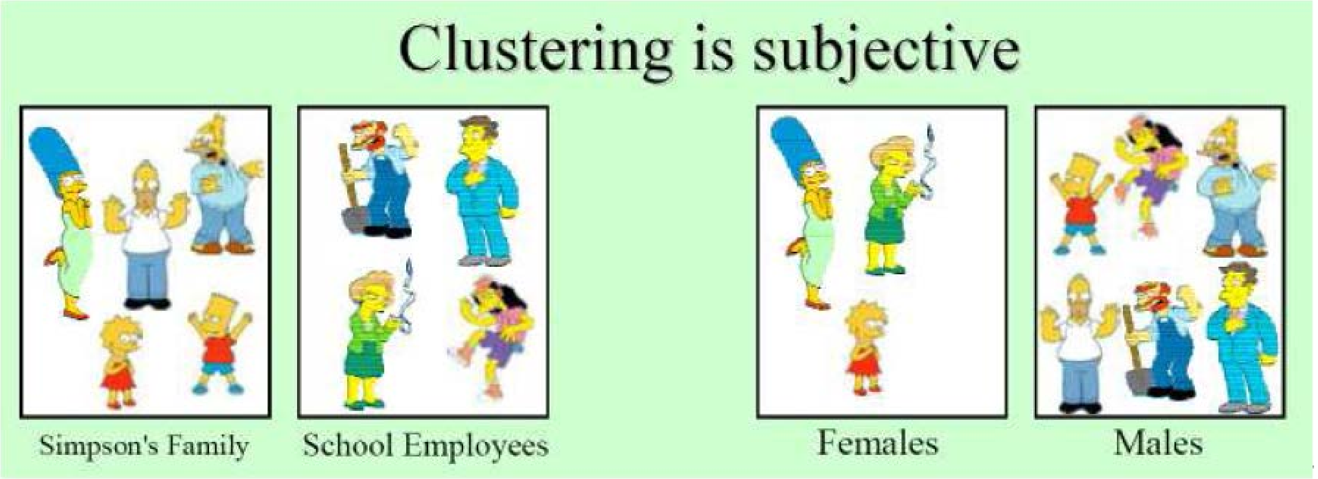
For speeding up vector space retrieval

- In vector space retrieval, we must find nearest doc vectors to query vector
- This entails finding the similarity of the query to every doc – slow (for some applications)
- By clustering docs in corpus a priori
  - find nearest docs in cluster(s) close to query
  - inexact but avoids exhaustive similarity computation

Issues for Clustering

- What is a natural grouping among these objects?
  - Definition of "groupness"
- Similarity measures
  - Using different measures for clustering can yield different clusters
- Number of clusters
  - Open problem

What is a natural grouping among these objects?



What is Similarity?

Hard to define!
But *we know it when we see it*

- The real meaning of similarity is a philosophical question. We will take a more pragmatic（实用的） approach

- Depends on representation and algorithm. For many rep./alg., easier to think in terms of a distance (rather than similarity) between vectors.

How many clusters?

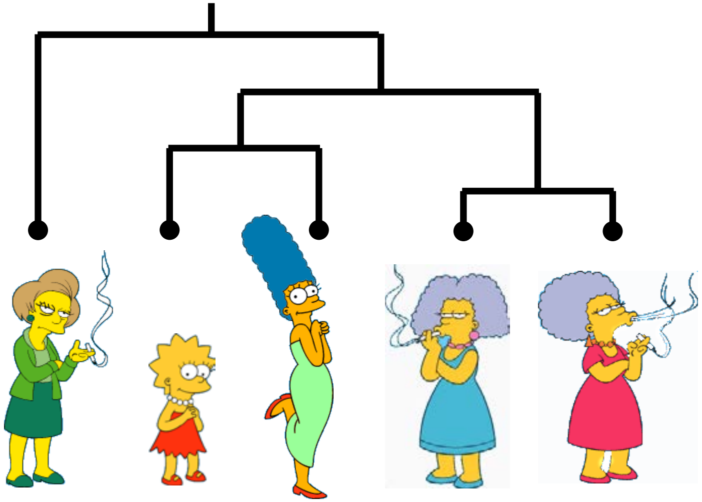
Open problem

- Fixed a priori?
- Completely data driven?
  - Avoid "trivial" clusters - too large or small
    - □ If a cluster's too large, then for navigation purposes you've wasted an extra user click without whittling down the set of documents much.

Two Types of Clustering

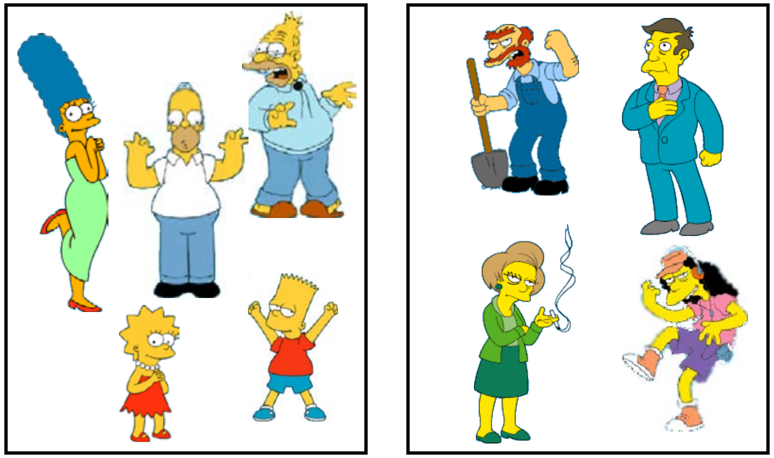
Hierarchical clustering

Cluster the data objects in a hierarchical way



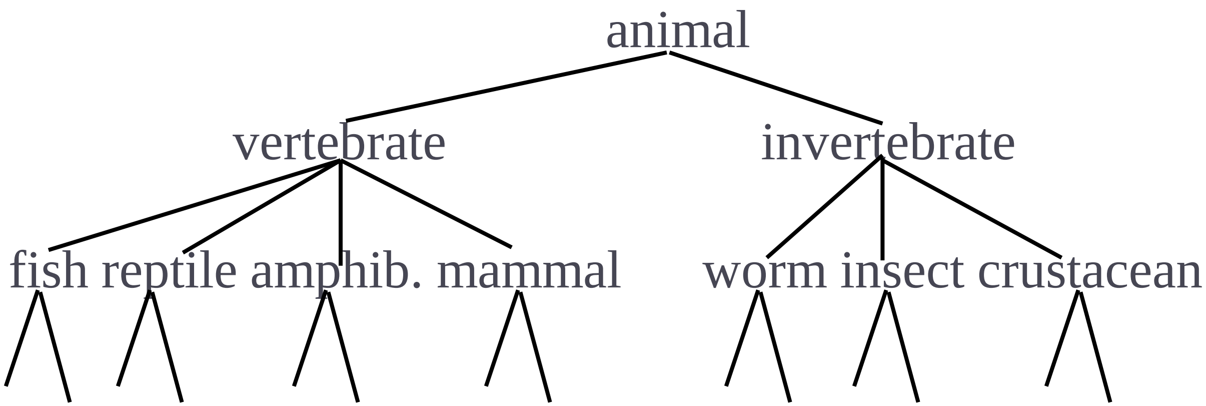
Partition-based clustering

Group the data directly into a flat clustering, usually by optimizing some criterion



Hierarchical Clustering

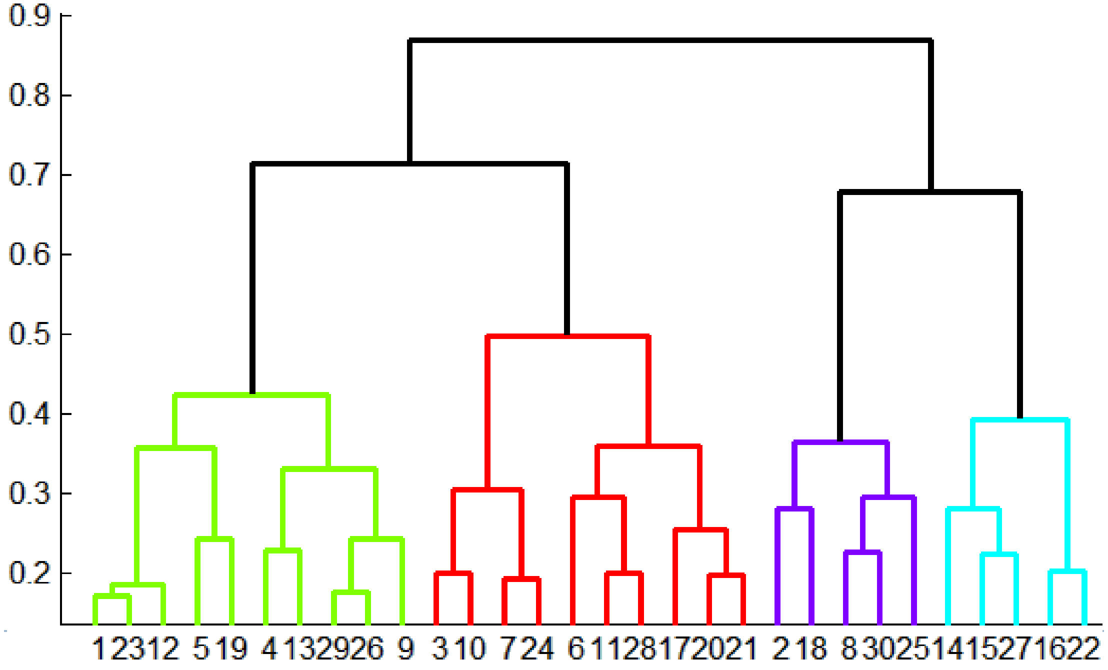
- Build a tree-based hierarchical taxonomy（分类系统） from a set of documents.



- Note that hierarchies are commonly used to organize information, for example in a web portal（网络门户）.
  - Yahoo! hierarchy is manually created, we will focus on automatic creation of hierarchies in data mining.

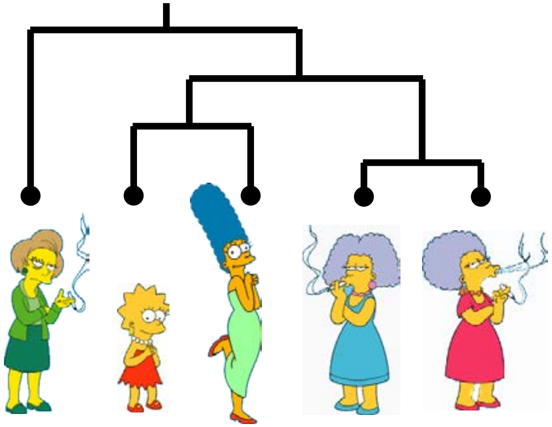
Dendrogram（系统树图）

- A useful tool for summarizing similarity measurement
  - The similarity between two objects in a dendrogram is represented as the height of the lowest internal node they share.
- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.



Hierarchical Clustering 层次聚类

Bottom-Up agglomerative clustering（合并式聚类）



- Starts with each obj 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.

Top-Down divisive clustering（分裂式聚类）

- Starting with all the data in a single cluster,
- Consider every possible way to divide the cluster into two. Choose the best division
- And recursively operate on both sides.

- Does not require the number of clusters *k* to be known in advance
  - But it does need a cutoff or threshold parameter condition

- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique

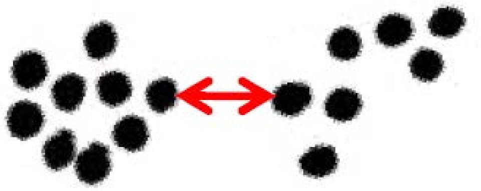
- Basic algorithm is straightforward

```
1. Compute the distance matrix
2. Let each data point be a cluster
3. **Repeat**
4.         Merge the two closest clusters
5.         Update the distance matrix
6. **Until** only a single cluster remains
```

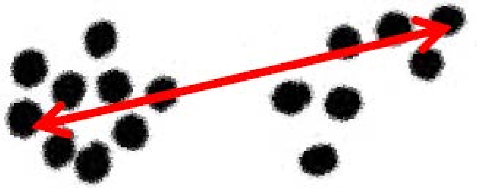
- Key operation is the computation of the distance of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

Closest Pair of Clusters

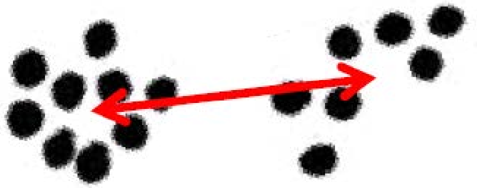
- The distance between two clusters is defined as the distance between
  - Single-Link
    - Nearest Neighbor: their closest members



  - Complete-Link
    - Furthest Neighbor: their furthest members



  - Centroid
    - Centers of gravity



  - Average-Link
    - Average of all cross-cluster pairs

Single Link Agglomerative Clustering

- Use maximum similarity of pairs

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

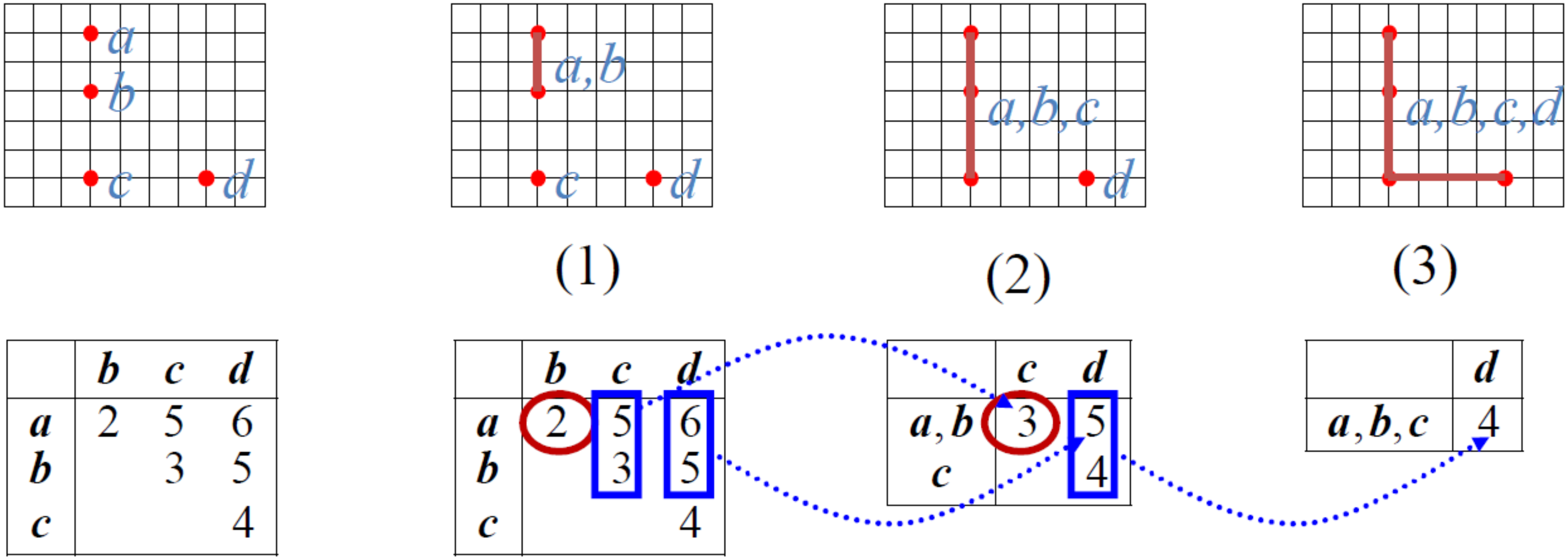
- Can result in "straggly" (long and thin) clusters due to chaining effect.

- 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 Agglomerative Clustering

Euclidean Distance



	b	*c*	*d*
a	2	5	6
b		3	5
c			4

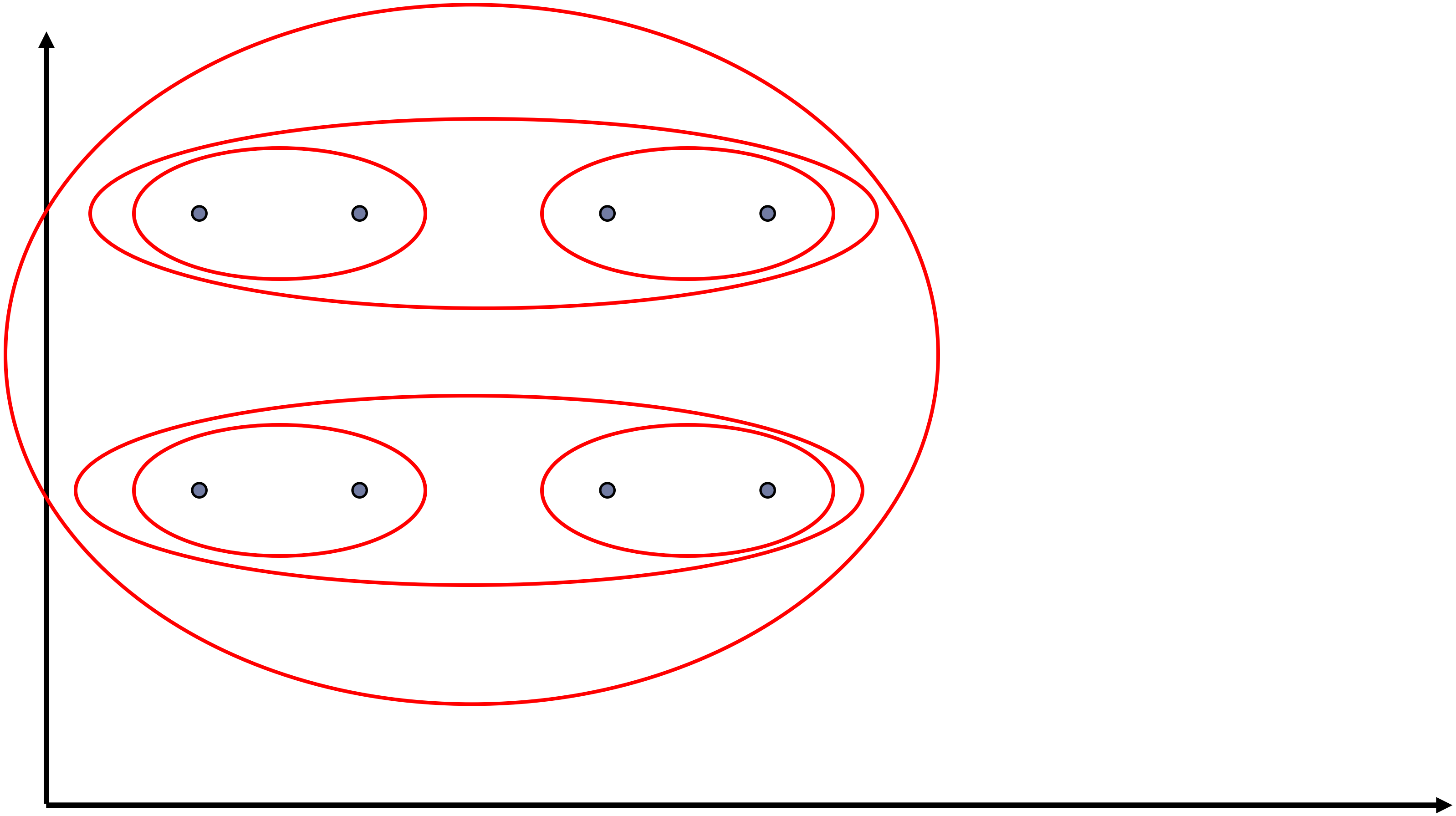
	b	*c*	*d*
a	2	5	6
b		3	5
c			4

	c	*d*
a,b	3	5
c		4

	d
a,b,c	4

Distance Matrix

Single Link Example



Complete Link Agglomerative Clustering

- Use minimum similarity of pairs

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

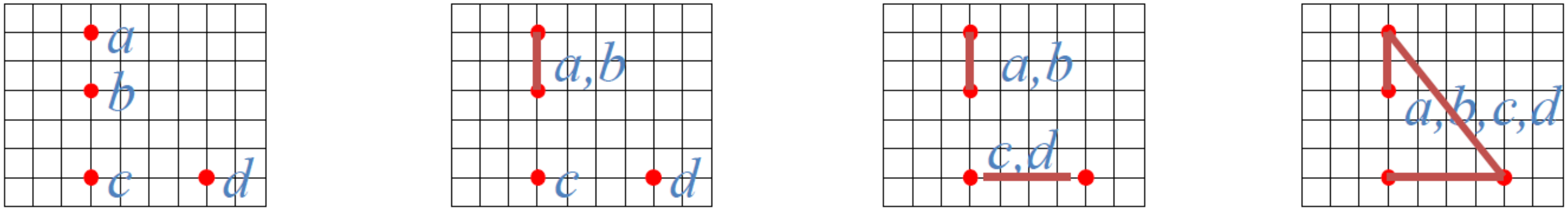
- Makes "tighter," spherical clusters that are typically preferable.

- 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 Method

Euclidean Distance



(1) (2) (3)

	b	c	d
a	2	5	6
b		3	5
c			4

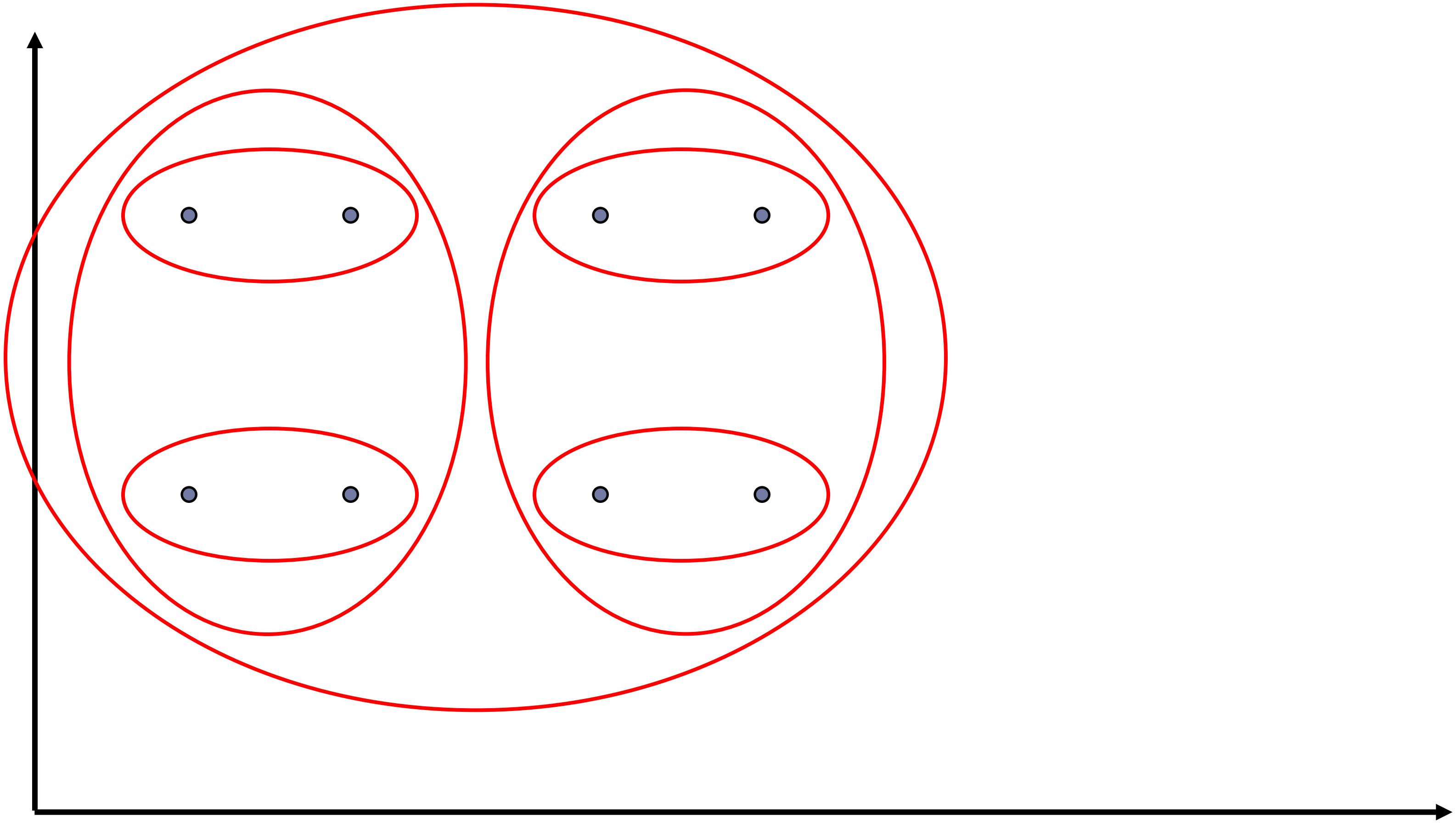
	b	c	d
a	2	5	6
b		3	5
c			4

	c	d
a,b	5	6
c		4

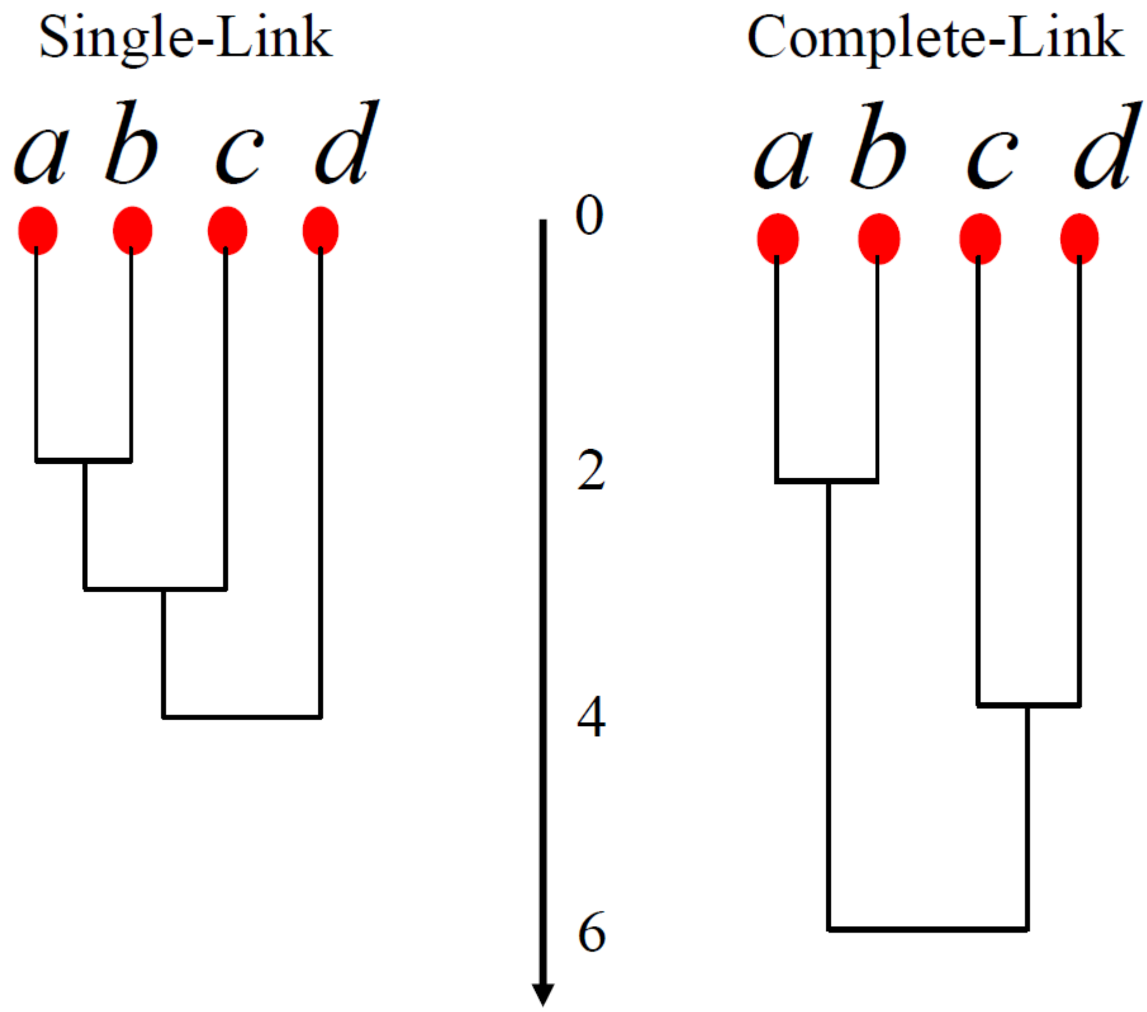
	c,d
a,b	6

Distance Matrix

Complete Link Example



Dendrograms



Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendrogram at the proper level

- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

Outline

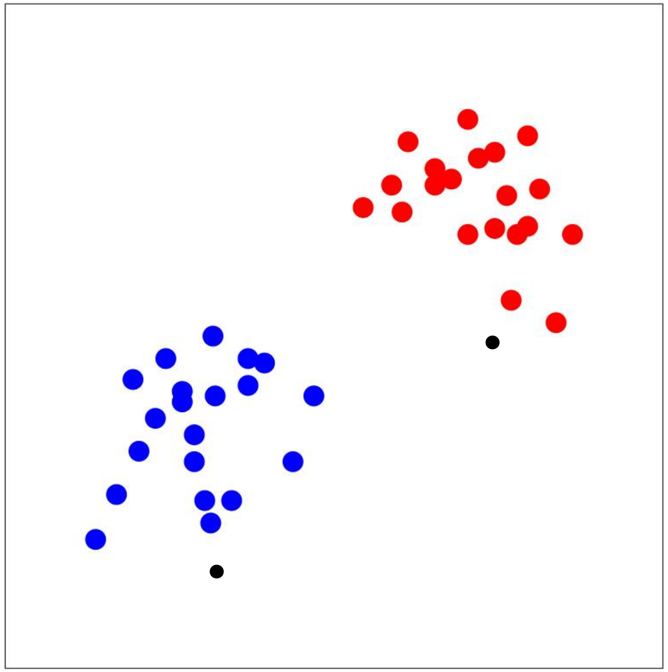
- Introduction to clustering

- Clustering methods
  - Hierarchical clustering
  - K-means

- Evaluation of clustering

K-Means Clustering

Create centers and assign points to centers to minimize sum of squared distance



K-Means Clustering

- Assumes documents are real-valued vectors.
- Clusters based on *centroids* (aka the *center of gravity* or mean) of points in a cluster, *c*:

$$\mu(\mathrm{c}) = \frac{1}{|c|}\sum_{x \in c} x$$

- Reassignment of instances to clusters is based on distance to the current cluster centroids.
  - (Or one can equivalently phrase it in terms of similarities)

K-Means Clustering

Algorithm *k-means*

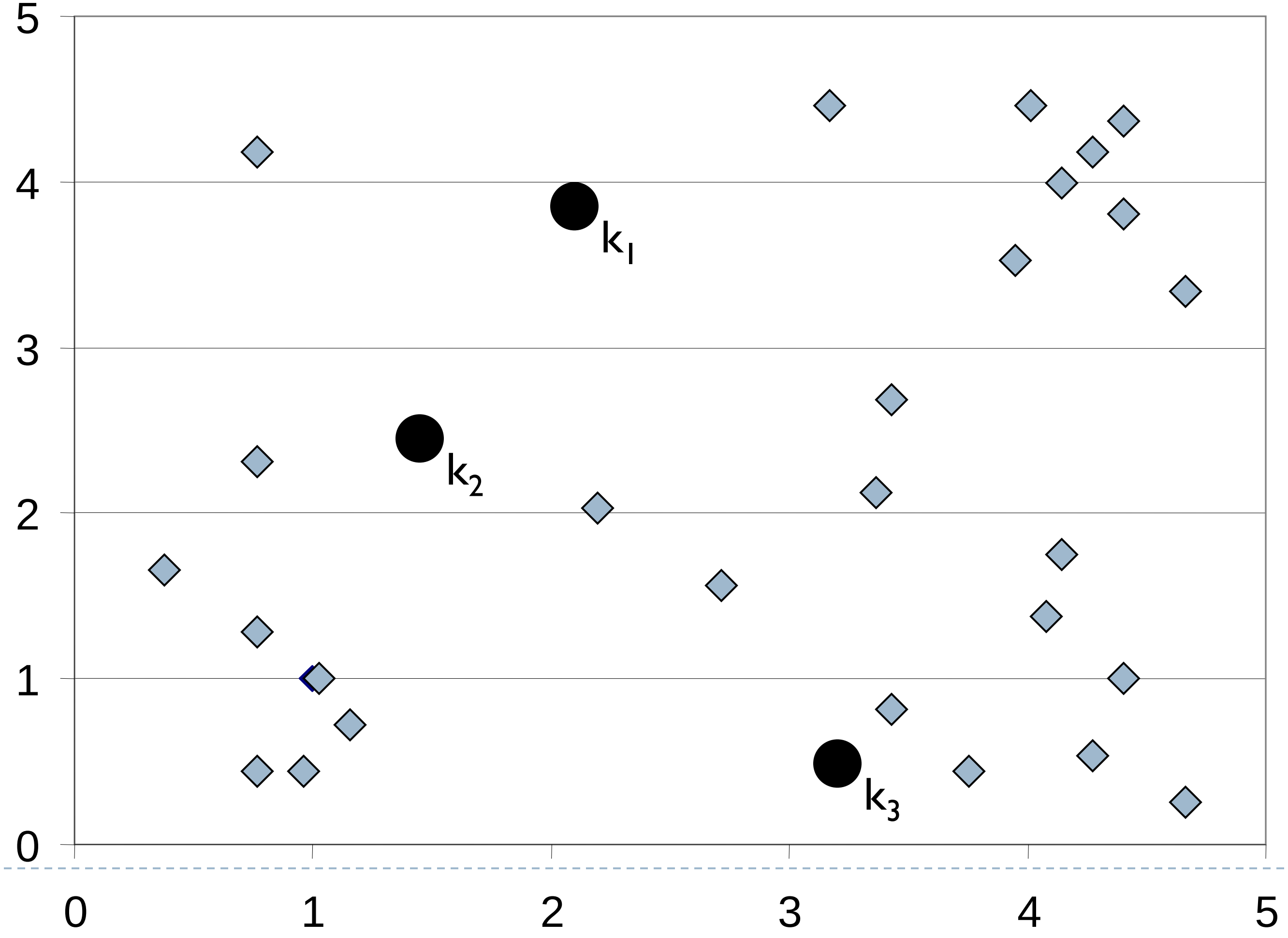
1. Decide on a value for *k*.
2. Initialize the *k* cluster centers (randomly, if necessary).
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.
4. Re-estimate the *k* cluster centers, by assuming the memberships found above are correct.

$$\mu_k = \sum_{i \in \text{cluster k}} \frac{1}{|C_k|} \mathbf{x}_i$$

5. If none of the *n* objects changed membership in the last iteration, exit. Otherwise goto 3.

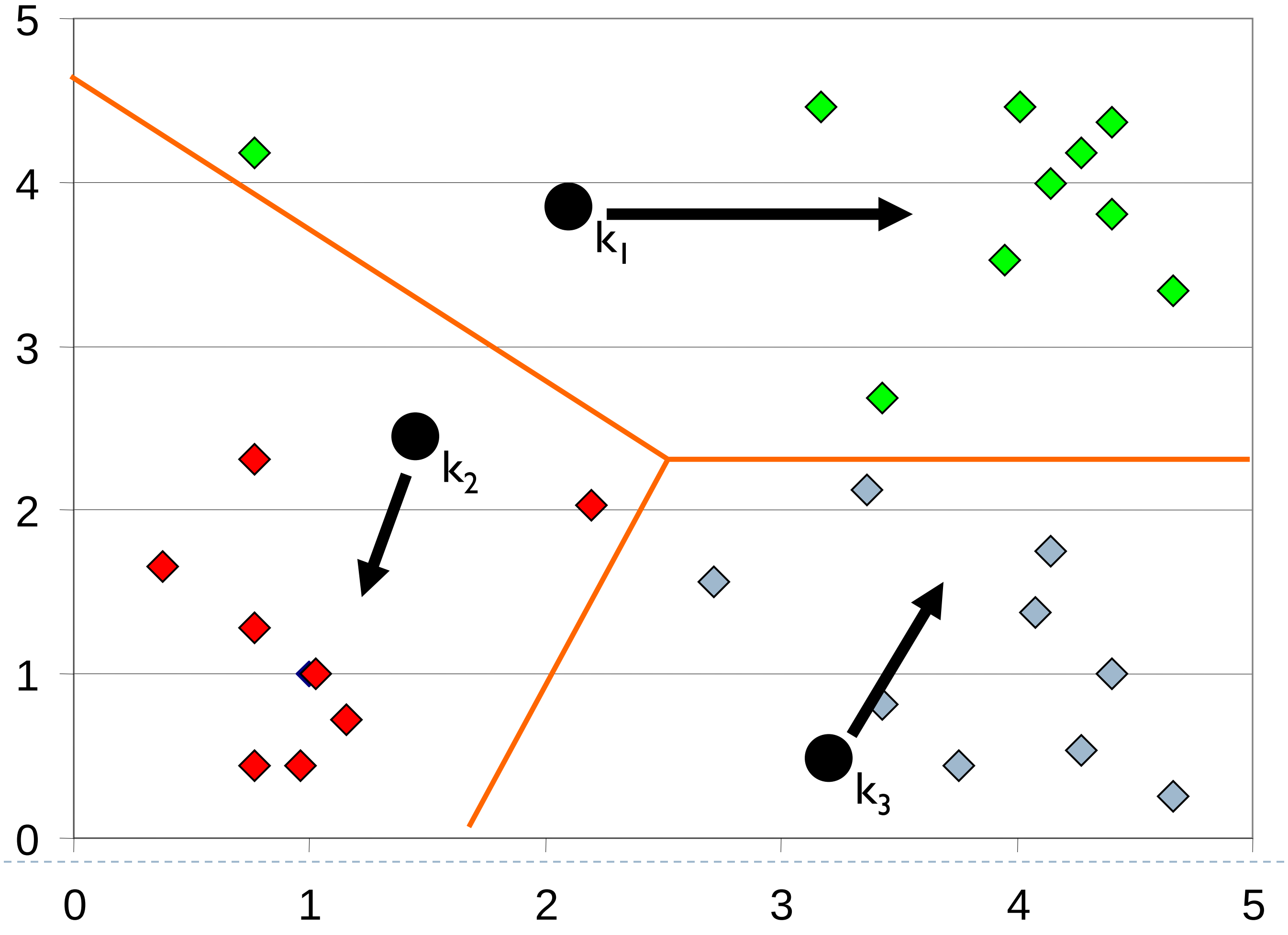
K-Means Clustering: Step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



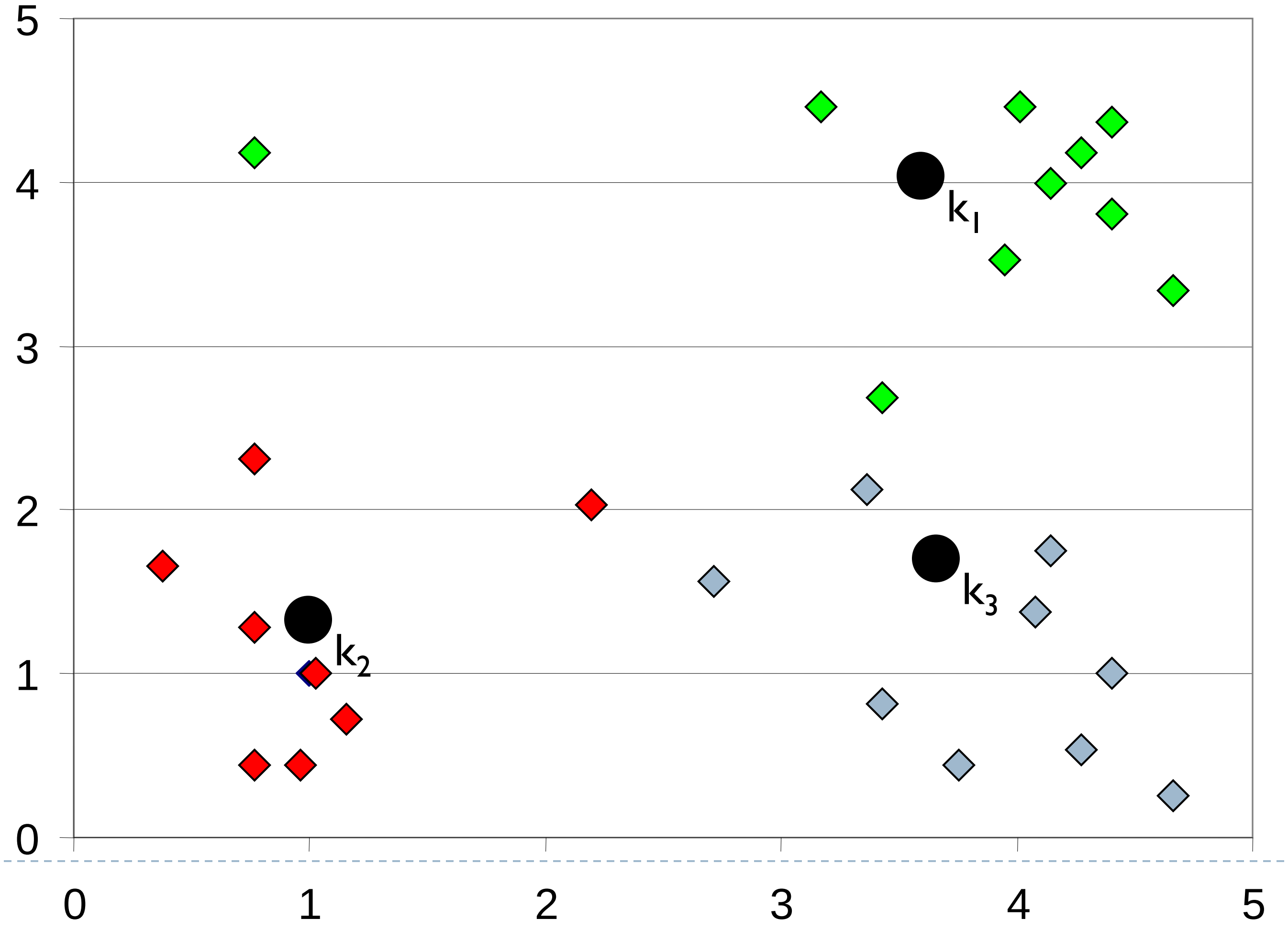
K-Means Clustering: Step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



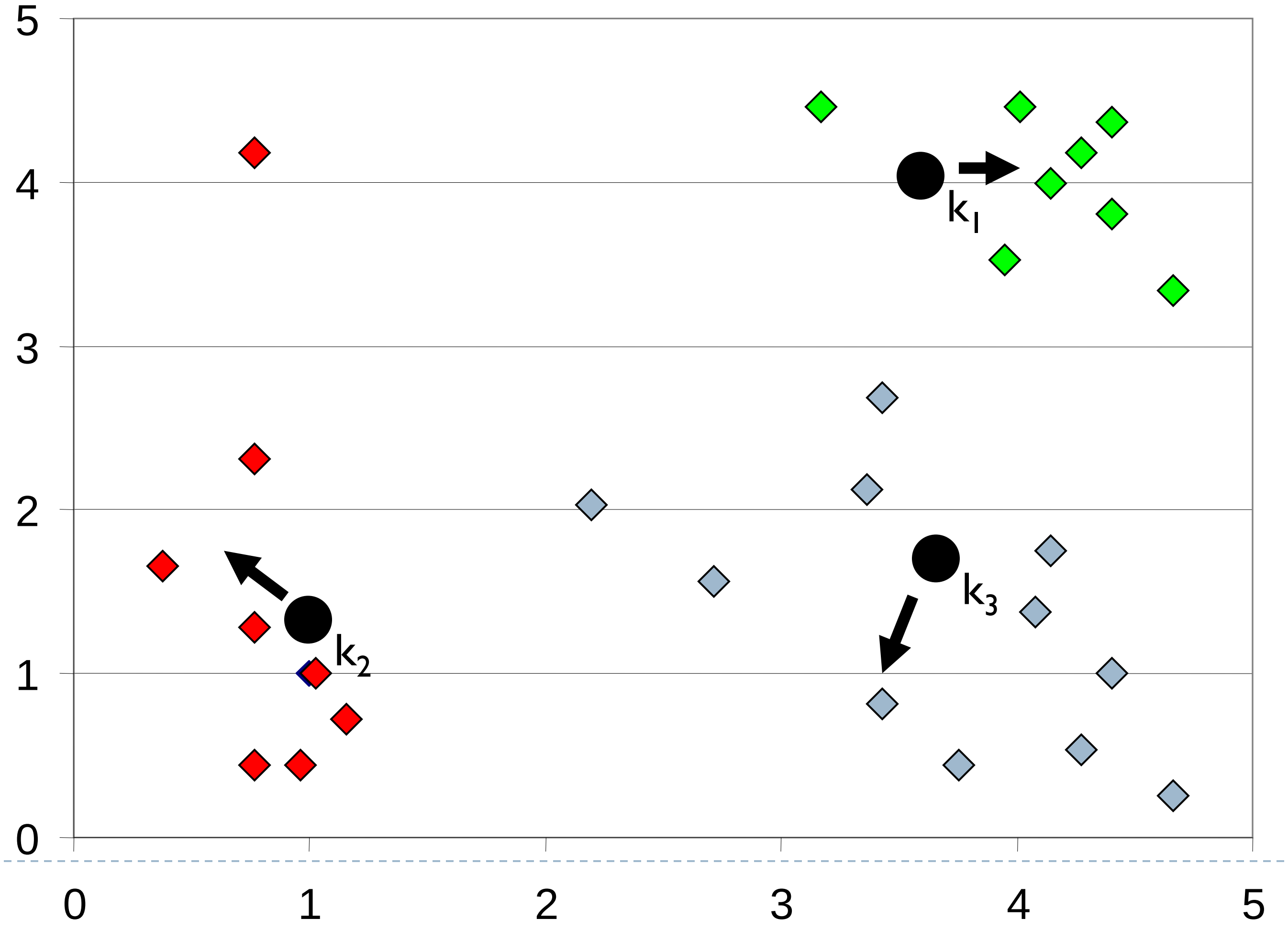
K-Means Clustering: Step 3

Algorithm: k-means, Distance Metric: Euclidean Distance

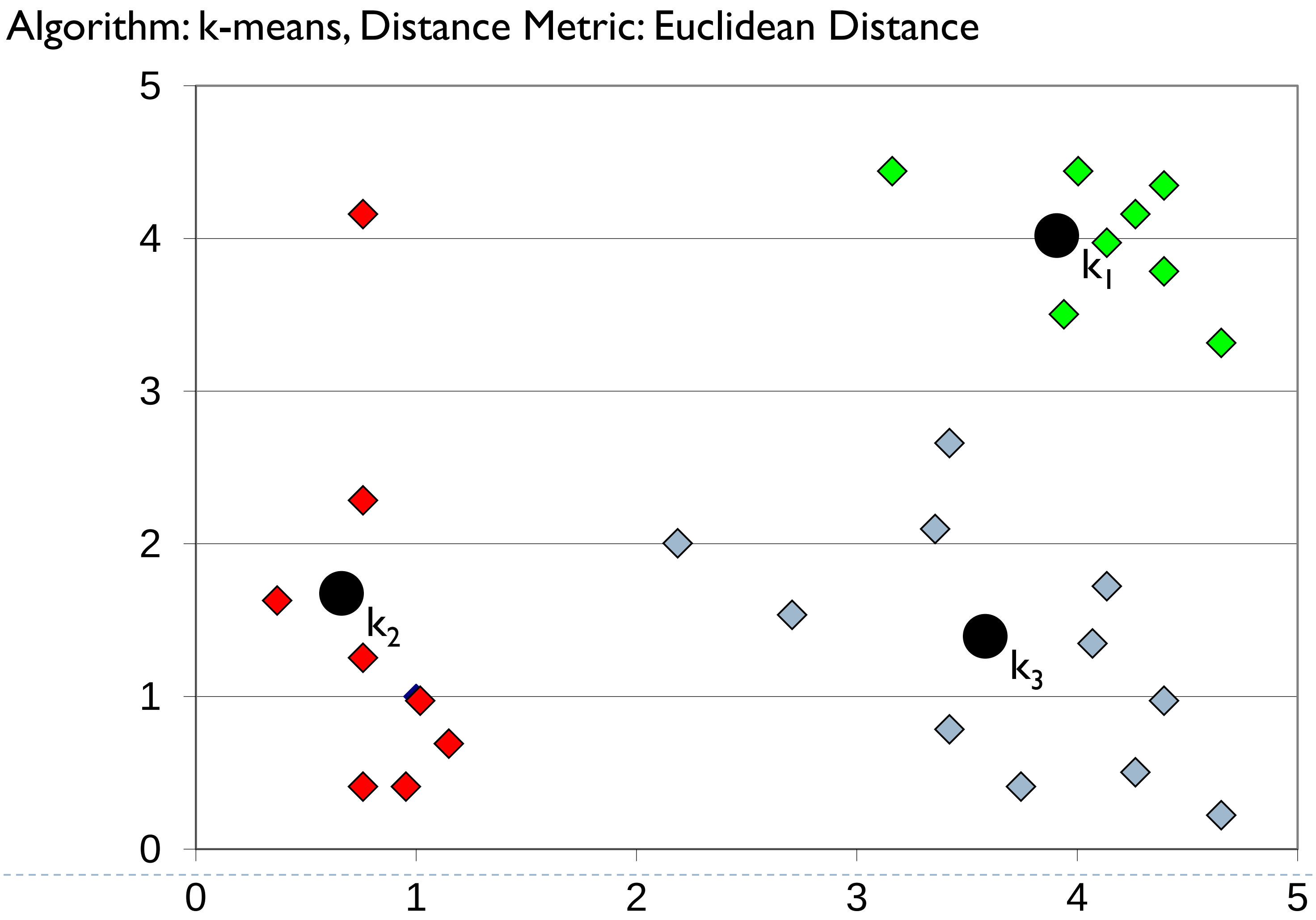


K-Means Clustering: Step 4

Algorithm: k-means, Distance Metric: Euclidean Distance



K-Means Clustering: Step 5



Computational Complexity

At each iteration,

- Computing distance between each of the n objects and the *k* cluster centers is O(*knd*), if *d* is the number of dimensions.
- Computing centroids: Each object gets added once to some centroid: O(*nd*).

Assume these two steps are each done once for *L* iterations:

O(*Lknd*).

Is k-means guaranteed to converge?

What is k-means optimizing?

K-Means recap

- Randomly initialize k centers

$$\mu^{(0)} = \mu_1^{(0)}, ..., \mu_k^{(0)}$$

- Classify: Assign each point j=1,…,n to nearest center:

$$C^{(t)}(j) \leftarrow \arg\min_i \|\mu_i - \mathbf{x}_j\|^2$$

- Recenter: Re-estimate the *k* cluster centers μ_i

$$\mu_i^{(t+1)} \leftarrow \arg\min_{\mu} \sum_{j \in \text{cluster i}} \|\mu - \mathbf{x}_j\|^2 \rightarrow \mu_i^{(t+1)} = \sum_{j \in \text{cluster i}} \frac{1}{C_i} \mathbf{x}_j$$

What is *K*-Means optimizing?

- Potential function $F(\mu, C)$ of centers μ and point allocations（分配） C:

$$F(\mu, C) = \sum_{j=1}^{n} \|\mu_{C(j)} - \mathbf{x}_j\|^2$$

- Optimal k-means

$$\min_{\mu} \min_{C} F(\mu, C)$$

K-Means algorithm

Optimize potential function $F(\mu, C)$

$$\min_{\mu} \min_{C} \sum_{j=1}^{n} \|\mu_{C(j)} - \mathbf{x}_j\|^2$$

(1) Fix μ , optimize C

$$\min_{C(1),C(2),\ldots,C(n)} \sum_{j=1}^{n} \|\mu_{C(j)} - \mathbf{x}_j\|^2$$

Exactly the first step-assign each point to the nearest cluster center

(2) Fix C, optimize

$$\min_{\mu(1),\mu(2),\ldots,\mu(k)} \sum_{j=1}^{n} \|\mu_{C(j)} - \mathbf{x}_j\|^2$$

Solution: average of points in cluster i, exactly second step (re-center)

More on K-Means objective

$$\boxed{\min_{Z:Z\in\{0,1\}^{t\times k}, Z\mathbf{1}=\mathbf{1}} \; \min_{U} \; tr\left((X - ZU)(X - ZU)'\right)}$$

- solving for U yields: $U = Z^{\dagger}X = (Z'Z)^{-1}Z'X$

$$\begin{aligned}
Z'X &= \begin{bmatrix} \text{sum of cluster 1 rows in } X \\ \vdots \\ \text{sum of cluster } k \text{ rows in } X \end{bmatrix} \\
Z'Z &= \begin{bmatrix} \#\text{ cluster 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \#\text{ cluster } k \end{bmatrix} \\
(Z'Z)^{-1} &= \begin{bmatrix} 1/(\#\text{ cluster 1}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/(\#\text{ cluster } k) \end{bmatrix}
\end{aligned}$$

More on K-Means objective

$$\boxed{\min_{Z:Z\in\{0,1\}^{t\times k},Z\mathbf{1}=\mathbf{1}} \min_{U} tr\left((X-ZU)(X-ZU)'\right)}$$

- solving for U yields: $U = Z^{\dagger}X = (Z'Z)^{-1}Z'X$

$$U = (Z'Z)^{-1}Z'X \quad = \begin{bmatrix} \text{mean of cluster 1 rows in } X \\ \vdots \\ \text{mean of cluster } k \text{ rows in } X \end{bmatrix}$$

$$X - ZU \quad = \begin{bmatrix} \text{row 1 of } X & - & \text{mean of row 1 cluster} \\ & \vdots & \\ \text{row } t \text{ of } X & - & \text{mean of row } t \text{ cluster} \end{bmatrix}$$

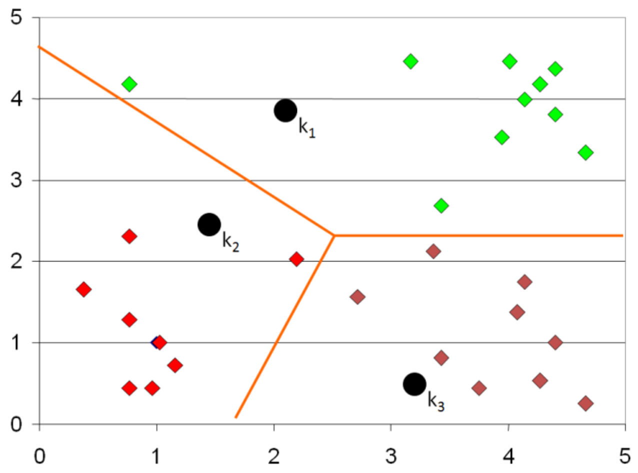
A few facts about *K*-Means

- Simple and efficient
- Always converges
    - Why?

- But…
    - K-means problem is **NP-hard**
    - No global solution
    - Not robust to noise and outliers

Seed Choice

Results can vary based on random seed selection.



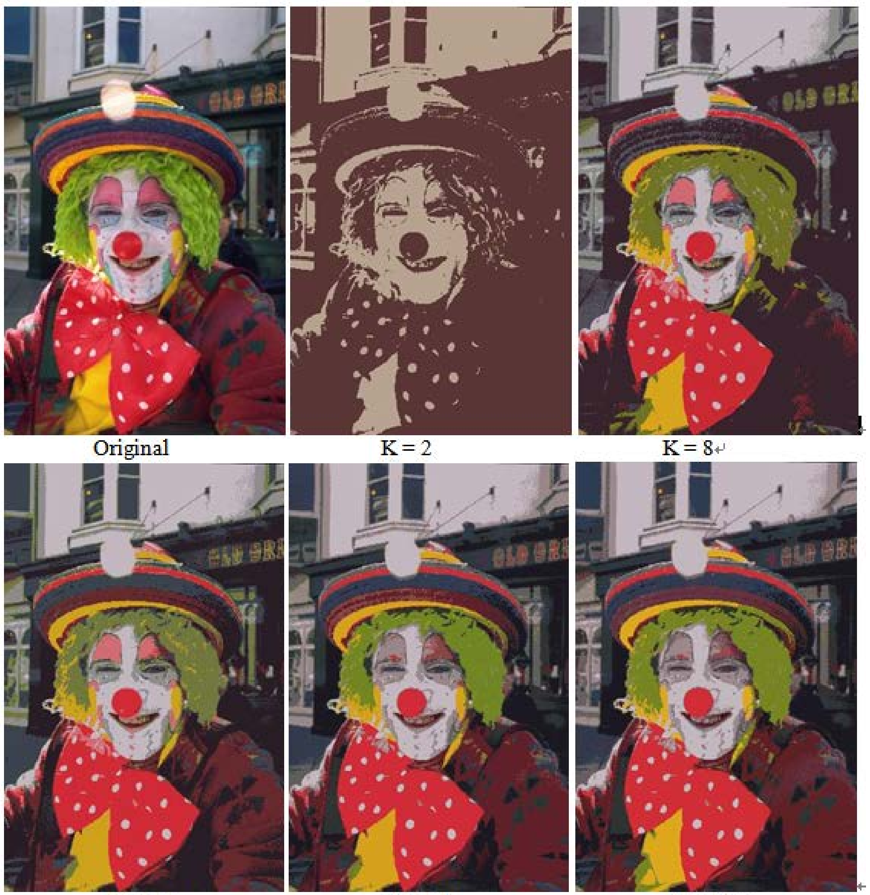
Some seeds can result in poor convergence rate, or convergence to suboptimal clustering.

- Try out multiple starting points
- Initialize with the results of another method.

Efficiency: Medoid As Cluster Representative

- The centroid does not have to be a document (typically won't be)
- Medoid: A cluster representative that is one of the documents
- For example: the document closest to the centroid
- One reason this is useful
  - Consider the representation of a large cluster (>1000 documents)
  - The centroid of this cluster will be a dense vector
  - The medoid of this cluster will be a sparse vector

K-Means for Image Segmentation



Original K = 2 K = 8

K = 11 K = 14 K = 15

Outline

- Introduction to clustering
- Clustering methods
  - Hierarchical clustering
  - K-means
- Evaluation of clustering

Clustering Evaluation: Hard Problem

The quality of a clustering is very hard to evaluate because

- We do not know the correct clusters

Evaluation Based on Internal Information

- **Intra-cluster cohesion** (compactness):
  - Cohesion measures how near the data points in a cluster are to the cluster centroid.
  - Sum of squared error (SSE) is a commonly used measure.

$$SSE = \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} dist^2(\mu_i, \mathbf{x})$$

- **Inter-cluster separation** (isolation):
  - Separation means that different cluster centroids should be far away from one another.
- In most applications, expert judgments are still the key.

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 （真实信息） … requires *labeled data*

- Assume documents with m true clusters $C=\{c_1, c_2, \ldots, c_m\}$, while our clustering algorithms produce K clusters, $\Omega=\{\omega_1, \omega_2, \ldots, \omega_K\}$ with n_i members.

Homework2: Evaluation of Clustering

- Accuracy of Clustering
- (Normalized) Mutual Information
- Purity

…

Indirect Evaluation

- In some applications, clustering is not the primary task, but used to help perform another task.
- We can use the performance on the primary task to compare clustering methods.
- For instance, in an application, the primary task is to provide recommendations on book purchasing to online shoppers.
  - If we can cluster books according to their features, we might be able to provide better recommendations.
  - We can evaluate different clustering algorithms based on how well they help with the recommendation task.
  - Here, we assume that the recommendation can be reliably evaluated.

Final Remarks on Clustering

- In clustering, clusters are inferred from the data without human input (unsupervised learning)
- However, in practice, it's a bit less clear: there are many ways of influencing the outcome of clustering: number of clusters, similarity measure, representation of documents, . . .

Hard vs. Soft Clustering

- Hard clustering: Each document belongs to exactly one cluster
  - More common and easier to do
- Soft clustering: A document can belong to more than one cluster.
  - Makes more sense for applications like creating browsable hierarchies
  - You may want to put a pair of sneakers in two clusters: (i) sports apparel and (ii) shoes
  - You can only do that with a soft clustering approach.

Summary

- Two types of clustering
  - Hierarchical, agglomerative clustering
  - Flat, partition-based clustering
- Key issues
  - Representation of data points
  - Similarity/distance measure
  - How many clusters?
- K-means: the basic partition-based algorithm
- Clustering evaluation