ifis

Institut für Informationssysteme
Technische Universität Braunschweig

# Information Retrieval and Web Search Engines

## Lecture 7: Document Clustering

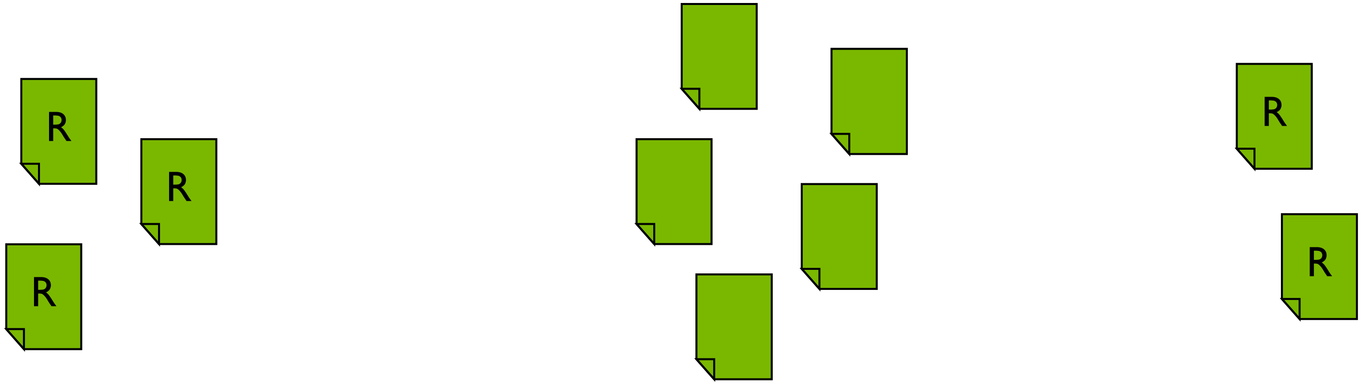
Wolf-Tilo Balke

Muhammad Usman

Institut für Informationssysteme
Technische Universität Braunschweig

# The Cluster Hypothesis

- The **Cluster Hypothesis** states:
  “Closely associated documents tend
  to be relevant to the same requests”
- **“Closely associated”** usually means **“similar”**
  (with respect to some kind of **similarity measure**)

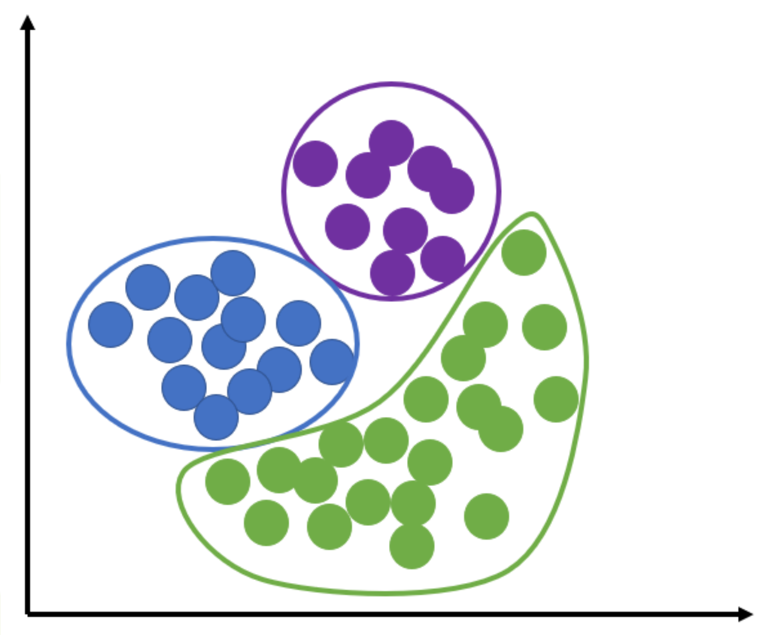


# The Cluster Hypothesis

- Experimental **validation** of the Cluster Hypothesis?
  - Proved to be problematic
  - Seems to be highly collection-specific
- Also depends on:
  - Representation of documents
  - Similarity measures
  - Queries
- **But:** It sounds reasonable and holds "often enough"
- In addition, real-world collections usually have a clear **cluster structure**
- Can we **exploit** clustering for information retrieval?

# Document Clustering

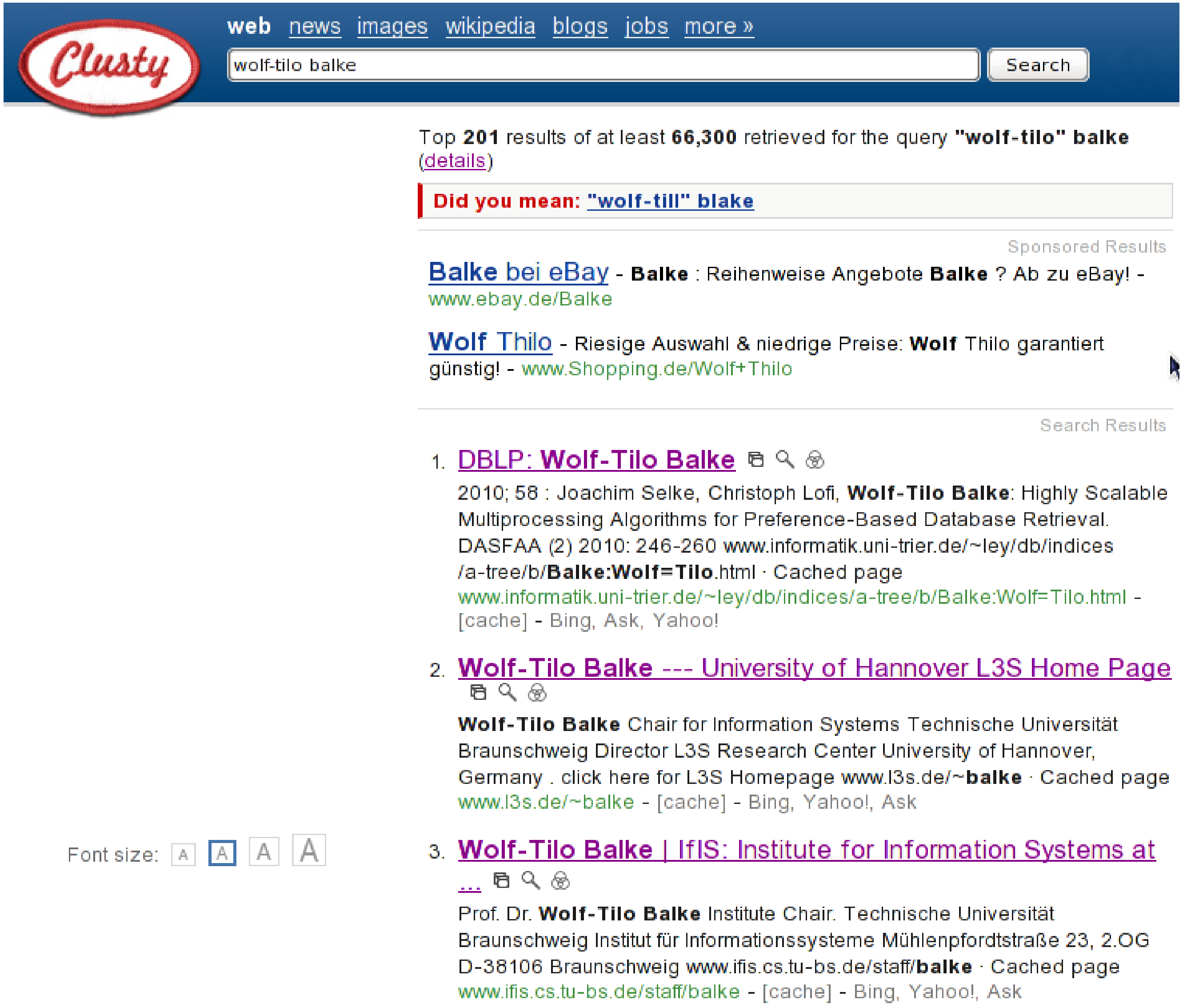
1. **Applications**
2. Issues in Clustering
3. Flat Clustering
4. Hierarchical Clustering



# I. Search Result Clustering

- In IR, results are typically presented by means of **ranked lists**

- What about **clusters?**



Top **201** results of at least **66,300** retrieved for the query **"wolf-tilo" balke** (details)

**Did you mean: "wolf-till" blake**

Sponsored Results

Balke bei eBay - **Balke** : Reihenweise Angebote **Balke** ? Ab zu eBay! - www.ebay.de/Balke

Wolf Thilo - Riesige Auswahl & niedrige Preise: **Wolf** Thilo garantiert günstig! - www.Shopping.de/Wolf+Thilo

Search Results

1. DBLP: Wolf-Tilo Balke
   2010; 58 : Joachim Selke, Christoph Lofi, **Wolf-Tilo Balke**: Highly Scalable Multiprocessing Algorithms for Preference-Based Database Retrieval. DASFAA (2) 2010: 246-260 www.informatik.uni-trier.de/~ley/db/indices /a-tree/b/**Balke:Wolf=Tilo**.html · Cached page
   www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Balke:Wolf=Tilo.html - [cache] - Bing, Ask, Yahoo!

2. Wolf-Tilo Balke --- University of Hannover L3S Home Page
   **Wolf-Tilo Balke** Chair for Information Systems Technische Universität Braunschweig Director L3S Research Center University of Hannover, Germany . click here for L3S Homepage www.l3s.de/~**balke** · Cached page
   www.l3s.de/~balke - [cache] - Bing, Yahoo!, Ask

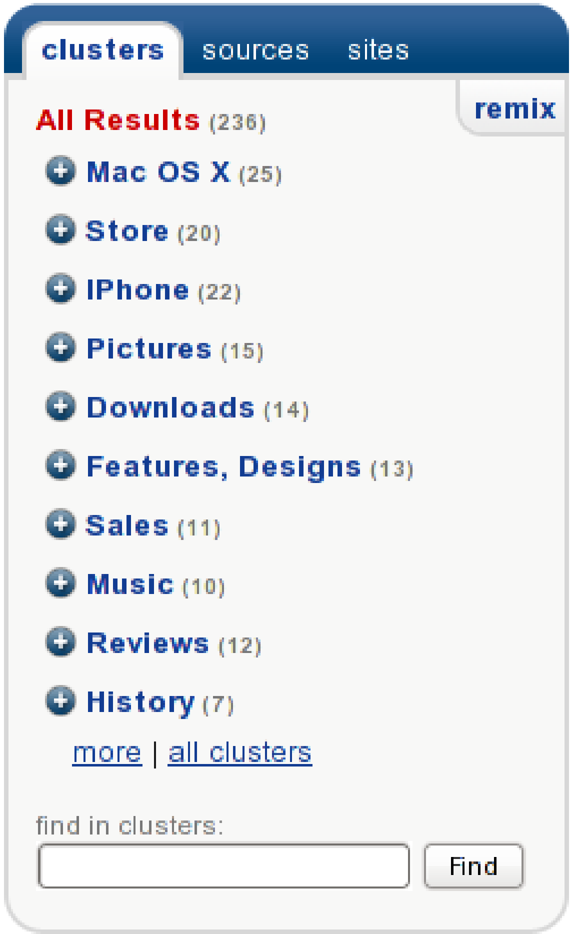
3. Wolf-Tilo Balke | IfIS: Institute for Information Systems at ...
   Prof. Dr. **Wolf-Tilo Balke** Institute Chair. Technische Universität Braunschweig Institut für Informationssysteme Mühlenpfordtstraße 23, 2.OG D-38106 Braunschweig www.ifis.cs.tu-bs.de/staff/**balke** · Cached page
   www.ifis.cs.tu-bs.de/staff/balke - [cache] - Bing, Yahoo!, Ask

# 1. Search Result Clustering

- Advantages:
  - Scanning **a few coherent groups** often is easier than scanning **many individual documents**
  - The cluster structure gives you an **impression** of what the result set looks like
- Disadvantages:
  - Finding **informative labels** for clusters is difficult
  - “Good” clusterings are **hard to find** (example on the next slide)

# 1. Search Result Clustering

- Cluster structure found for **query "apple"**:



All Results (236)

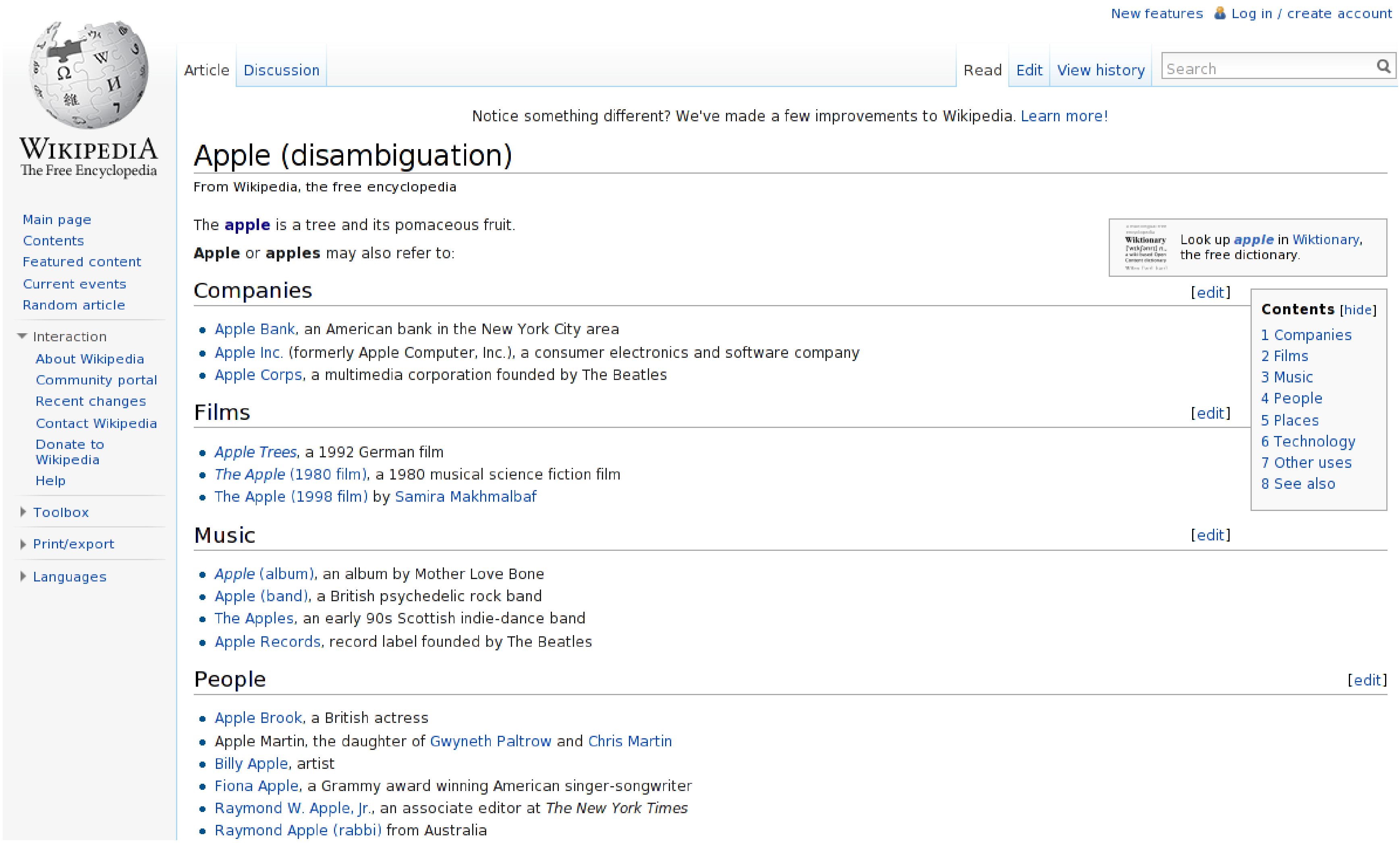
- Mac OS X (25)
- Store (20)
- IPhone (22)
- Pictures (15)
- Downloads (14)
- Features, Designs (13)
- Sales (11)
- Music (10)
- Reviews (12)
- History (7)

more | all clusters

find in clusters:

# 1. Search Result Clustering

- **Ideally,** a clustering should look like this:



Notice something different? We've made a few improvements to Wikipedia. Learn more!

## Apple (disambiguation)

From Wikipedia, the free encyclopedia

The **apple** is a tree and its pomaceous fruit.

**Apple** or **apples** may also refer to:

Look up ***apple*** in Wiktionary, the free dictionary.

**Contents** [hide]

1 Companies
2 Films
3 Music
4 People
5 Places
6 Technology
7 Other uses
8 See also

### Companies

- Apple Bank, an American bank in the New York City area
- Apple Inc. (formerly Apple Computer, Inc.), a consumer electronics and software company
- Apple Corps, a multimedia corporation founded by The Beatles

### Films

- *Apple Trees*, a 1992 German film
- *The Apple* (1980 film), a 1980 musical science fiction film
- The Apple (1998 film) by Samira Makhmalbaf

### Music

- *Apple* (album), an album by Mother Love Bone
- Apple (band), a British psychedelic rock band
- The Apples, an early 90s Scottish indie-dance band
- Apple Records, record label founded by The Beatles

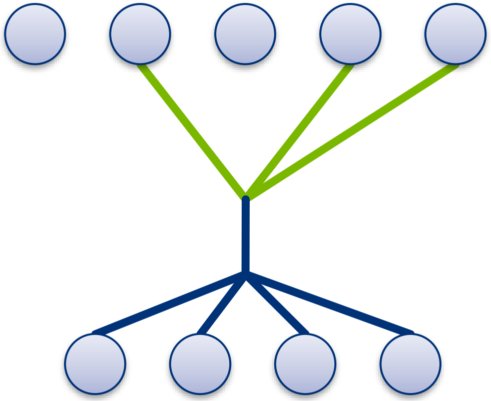
### People

- Apple Brook, a British actress
- Apple Martin, the daughter of Gwyneth Paltrow and Chris Martin
- Billy Apple, artist
- Fiona Apple, a Grammy award winning American singer-songwriter
- Raymond W. Apple, Jr., an associate editor at *The New York Times*
- Raymond Apple (rabbi) from Australia

# 2. Scatter–Gather

- Scatter-Gather is a **navigational user interface**
- Search without typing!

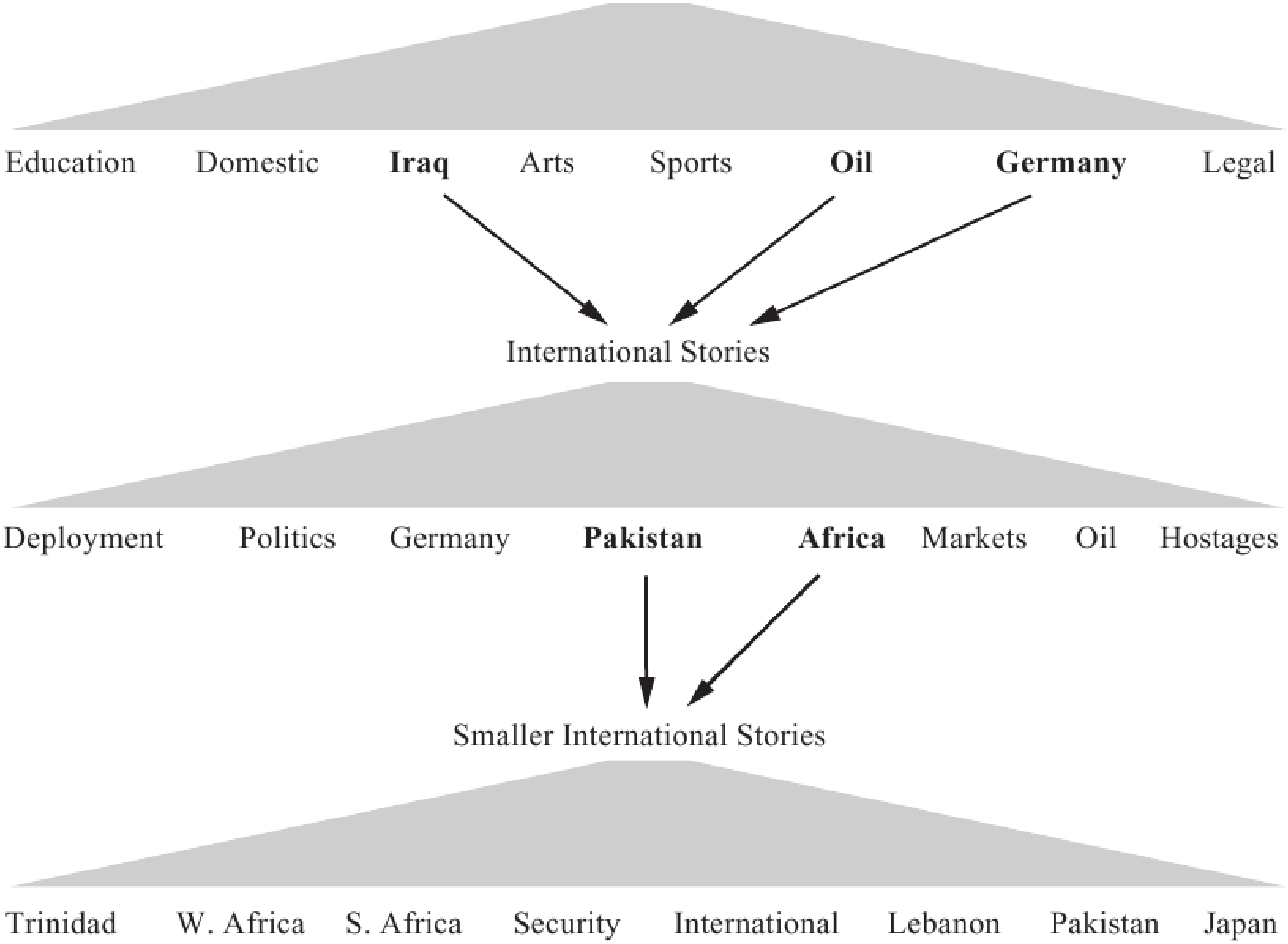
- **Idea:**
  1. Cluster the whole document collection into a **small number of clusters**
  2. Users formulate queries by **selecting** one or more of these **clusters**
  3. Selected clusters are **merged and clustered again**
  4. **Return to step 2** if not finished



# 2. Scatter–Gather

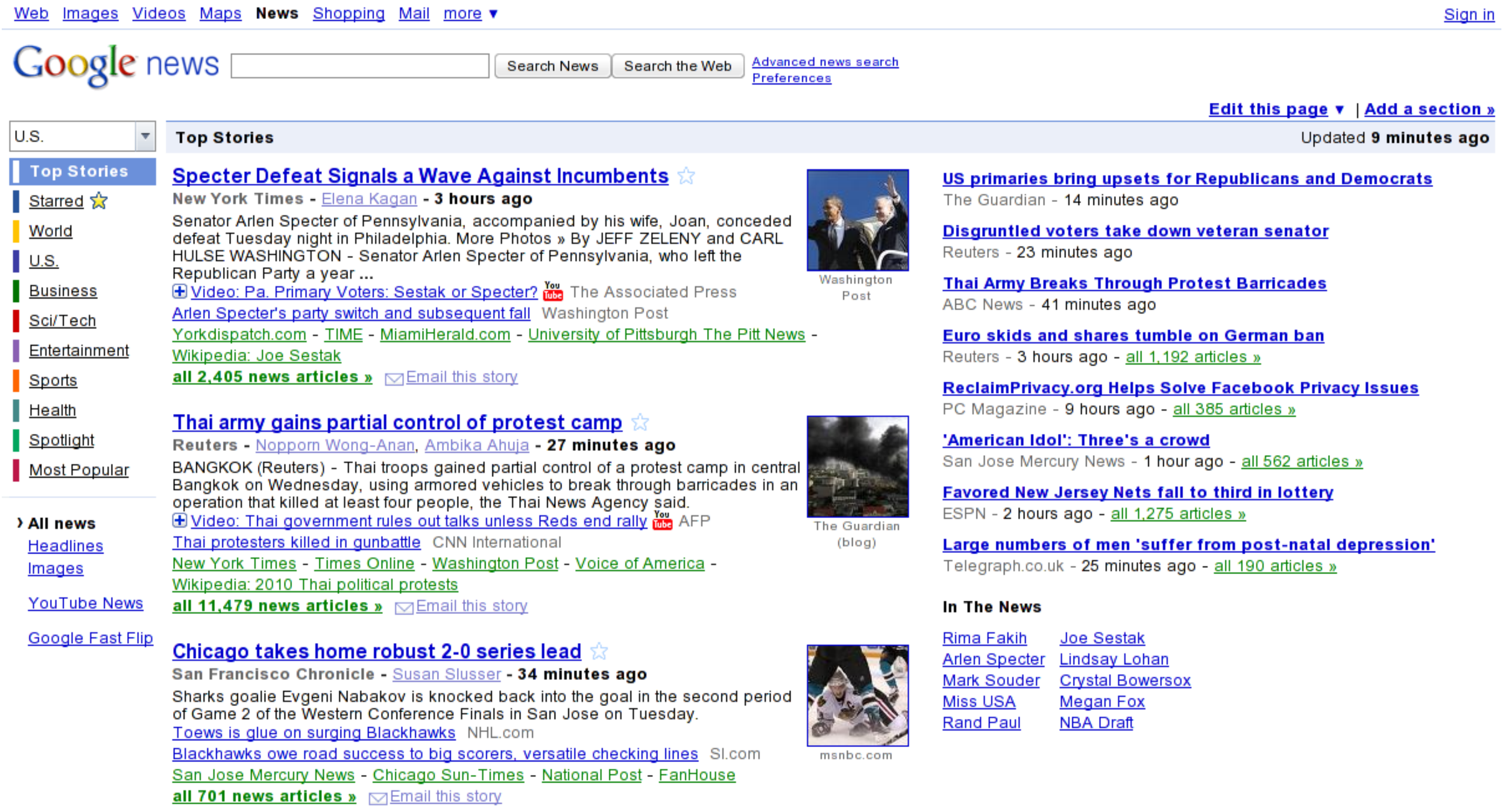
- Example from (Manning *et al.*, 2008):

**Collection:**
*New York Times* news stories



# 3. Collection Clustering

- Sometimes it makes sense to cluster the whole document collection hierarchically:



Web Images Videos Maps **News** Shopping Mail more ▾ Sign in

Google news [ ] Search News | Search the Web | Advanced news search | Preferences

Edit this page ▾ | Add a section »

U.S. ▾

**Top Stories** Updated **9 minutes ago**

**Top Stories**
Starred ☆
World
U.S.
Business
Sci/Tech
Entertainment
Sports
Health
Spotlight
Most Popular

› **All news**
Headlines
Images

YouTube News

Google Fast Flip

**Specter Defeat Signals a Wave Against Incumbents** ☆
**New York Times** - Elena Kagan - **3 hours ago**
Senator Arlen Specter of Pennsylvania, accompanied by his wife, Joan, conceded defeat Tuesday night in Philadelphia. More Photos » By JEFF ZELENY and CARL HULSE WASHINGTON - Senator Arlen Specter of Pennsylvania, who left the Republican Party a year ...
⊞ Video: Pa. Primary Voters: Sestak or Specter? The Associated Press
Arlen Specter's party switch and subsequent fall Washington Post
Yorkdispatch.com - TIME - MiamiHerald.com - University of Pittsburgh The Pitt News - Wikipedia: Joe Sestak
**all 2,405 news articles »** ✉ Email this story

Washington Post

**Thai army gains partial control of protest camp** ☆
**Reuters** - Nopporn Wong-Anan, Ambika Ahuja - **27 minutes ago**
BANGKOK (Reuters) - Thai troops gained partial control of a protest camp in central Bangkok on Wednesday, using armored vehicles to break through barricades in an operation that killed at least four people, the Thai News Agency said.
⊞ Video: Thai government rules out talks unless Reds end rally AFP
Thai protesters killed in gunbattle CNN International
New York Times - Times Online - Washington Post - Voice of America - Wikipedia: 2010 Thai political protests
**all 11,479 news articles »** ✉ Email this story

The Guardian (blog)

**Chicago takes home robust 2-0 series lead** ☆
**San Francisco Chronicle** - Susan Slusser - **34 minutes ago**
Sharks goalie Evgeni Nabakov is knocked back into the goal in the second period of Game 2 of the Western Conference Finals in San Jose on Tuesday.
Toews is glue on surging Blackhawks NHL.com
Blackhawks owe road success to big scorers, versatile checking lines SI.com
San Jose Mercury News - Chicago Sun-Times - National Post - FanHouse
**all 701 news articles »** ✉ Email this story

msnbc.com

**US primaries bring upsets for Republicans and Democrats**
The Guardian - 14 minutes ago

**Disgruntled voters take down veteran senator**
Reuters - 23 minutes ago

**Thai Army Breaks Through Protest Barricades**
ABC News - 41 minutes ago

**Euro skids and shares tumble on German ban**
Reuters - 3 hours ago - all 1,192 articles »

**ReclaimPrivacy.org Helps Solve Facebook Privacy Issues**
PC Magazine - 9 hours ago - all 385 articles »

**'American Idol': Three's a crowd**
San Jose Mercury News - 1 hour ago - all 562 articles »

**Favored New Jersey Nets fall to third in lottery**
ESPN - 2 hours ago - all 1,275 articles »

**Large numbers of men 'suffer from post-natal depression'**
Telegraph.co.uk - 25 minutes ago - all 190 articles »

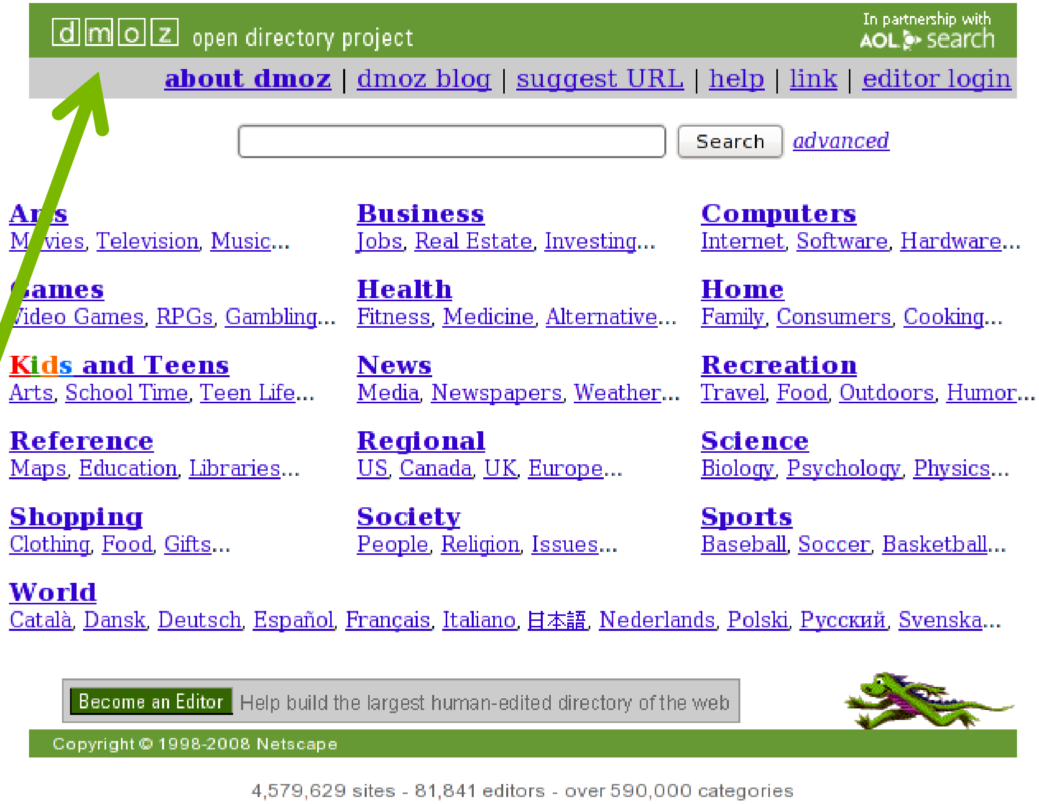
**In The News**

Rima Fakih
Arlen Specter
Mark Souder
Miss USA
Rand Paul

Joe Sestak
Lindsay Lohan
Crystal Bowersox
Megan Fox
NBA Draft

# 3. Collection Clustering

- Collection clustering is especially useful if…
  - The collections contains only a **small number of topics**
  - Each topic is covered by many documents in a similar fashion
- Advantages:
  - Enables exploratory browsing
  - Can be helpful even if users are unsure about which query terms to use



dmoz open directory project

In partnership with AOL search

about dmoz | dmoz blog | suggest URL | help | link | editor login

Search advanced

**Arts**
Movies, Television, Music...

**Games**
Video Games, RPGs, Gambling...

**Kids and Teens**
Arts, School Time, Teen Life...

**Reference**
Maps, Education, Libraries...

**Shopping**
Clothing, Food, Gifts...

**Business**
Jobs, Real Estate, Investing...

**Health**
Fitness, Medicine, Alternative...

**News**
Media, Newspapers, Weather...

**Regional**
US, Canada, UK, Europe...

**Society**
People, Religion, Issues...

**Computers**
Internet, Software, Hardware...

**Home**
Family, Consumers, Cooking...

**Recreation**
Travel, Food, Outdoors, Humor...

**Science**
Biology, Psychology, Physics...

**Sports**
Baseball, Soccer, Basketball...

**World**
Català, Dansk, Deutsch, Español, Français, Italiano, 日本語, Nederlands, Polski, Русский, Svenska...

Become an Editor Help build the largest human-edited directory of the web

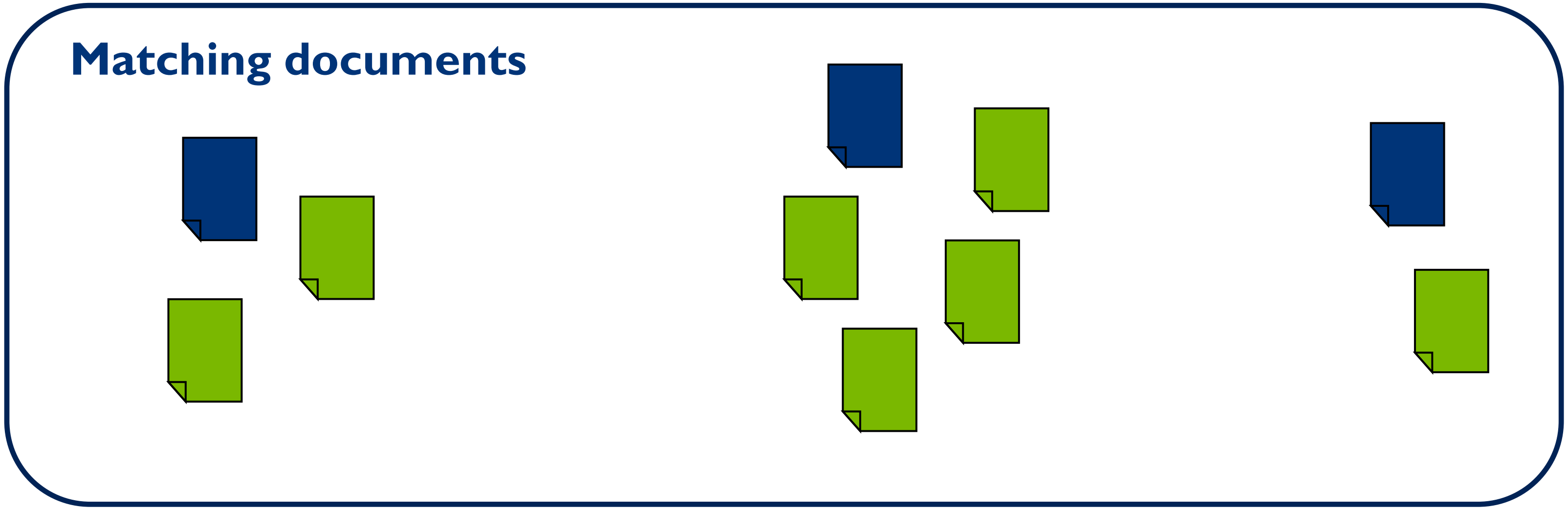
Copyright © 1998-2008 Netscape

4,579,629 sites - 81,841 editors - over 590,000 categories

**There's no clustering here! But dmoz is an example of using a global hierarchy for navigation**

# 4. Language Modeling

- Collection clustering can also be used to **extend small result lists**
- If there is only a small number of documents matching the query, **add similar documents** from the clusters containing the matching documents



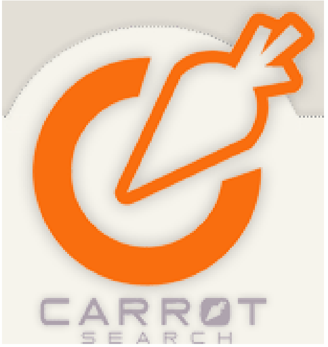
# 5. Cluster-based Retrieval

- Also interesting:
  Use collection clustering to **speed-up retrieval**
- **Idea:**
  - Cluster the whole collection
  - Represent each cluster by a (possibly virtual) document, e.g., a typical or average document contained in the cluster
  - Speed-up query processing by first finding the clusters having best-matching representatives and then doing retrieval only on the documents in these clusters
    1. Find best-matching clusters
    2. Build the set of documents contained in these clusters
    3. Find best-matching documents

# Cluster Based Retrieval

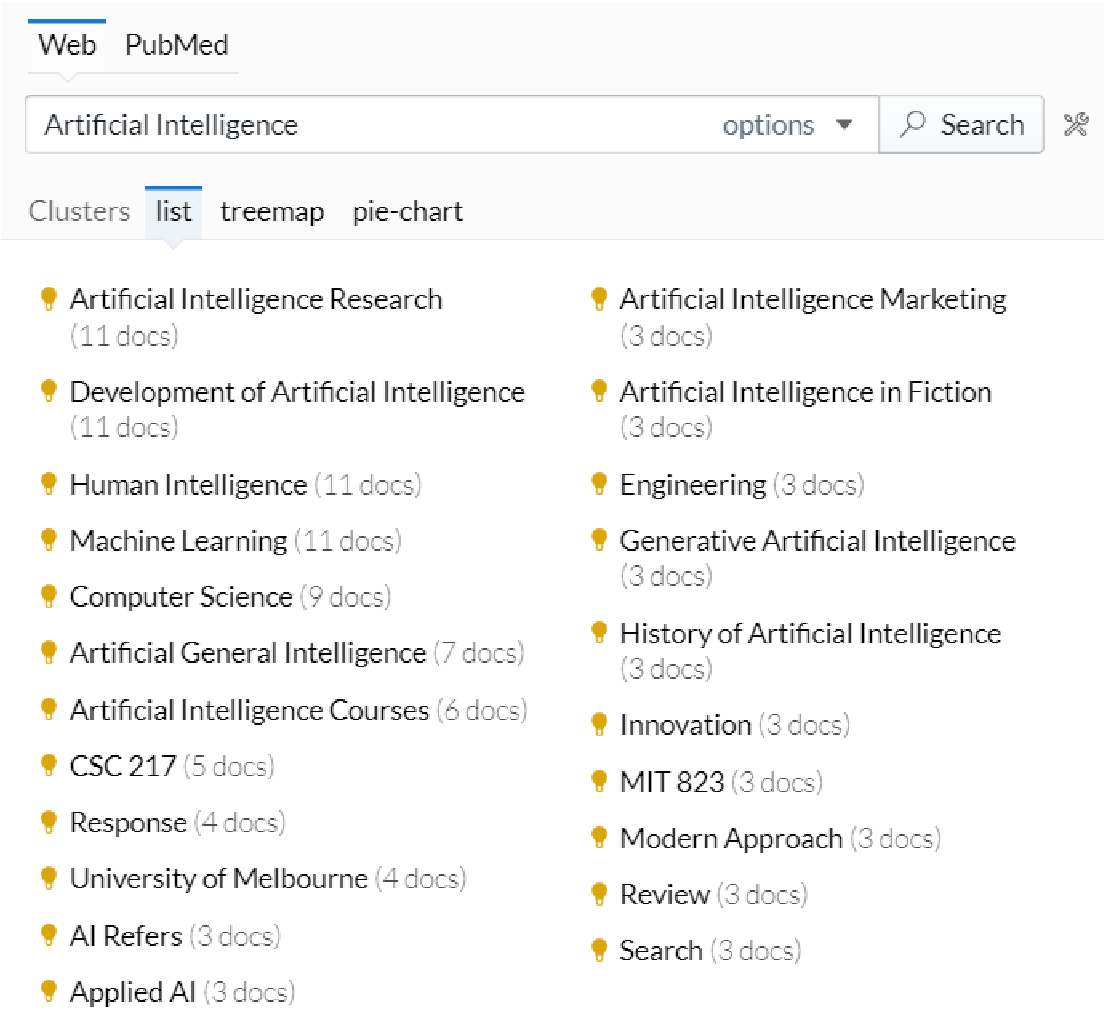
Detour

- Carrot2
  - Open source!
  - Cluster search results into thematic groups
  - http://search.carrot2.org



# Cluster Based Retrieval

Detour

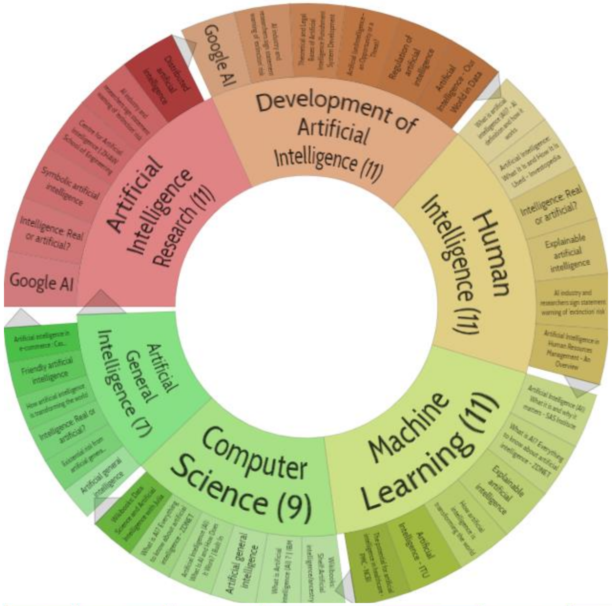


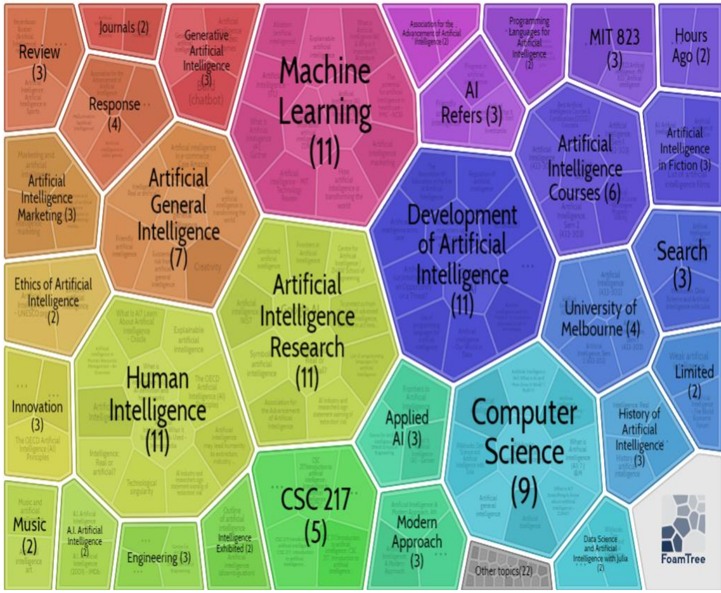
Web PubMed

Artificial Intelligence | options ▾ | Search

Clusters | list | treemap | pie-chart

- Artificial Intelligence Research (11 docs)
- Development of Artificial Intelligence (11 docs)
- Human Intelligence (11 docs)
- Machine Learning (11 docs)
- Computer Science (9 docs)
- Artificial General Intelligence (7 docs)
- Artificial Intelligence Courses (6 docs)
- CSC 217 (5 docs)
- Response (4 docs)
- University of Melbourne (4 docs)
- AI Refers (3 docs)
- Applied AI (3 docs)
- Artificial Intelligence Marketing (3 docs)
- Artificial Intelligence in Fiction (3 docs)
- Engineering (3 docs)
- Generative Artificial Intelligence (3 docs)
- History of Artificial Intelligence (3 docs)
- Innovation (3 docs)
- MIT 823 (3 docs)
- Modern Approach (3 docs)
- Review (3 docs)
- Search (3 docs)





# Cluster Based Retrieval

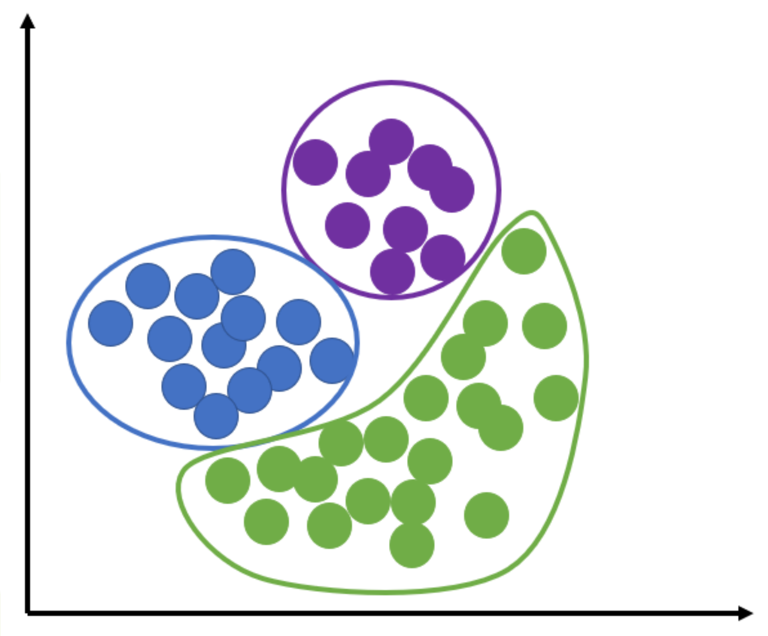
Detour

How are clusters formed?

- Document Representation:
  - TF-IDF, Bag of Words, Word embedding
- Similarity Computation:
  - Cosine similarity, Euclidean distance, or Jaccard similarity
- Clustering Algorithm:
  - k-means, hierarchical clustering, or density-based clustering

# Document Clustering

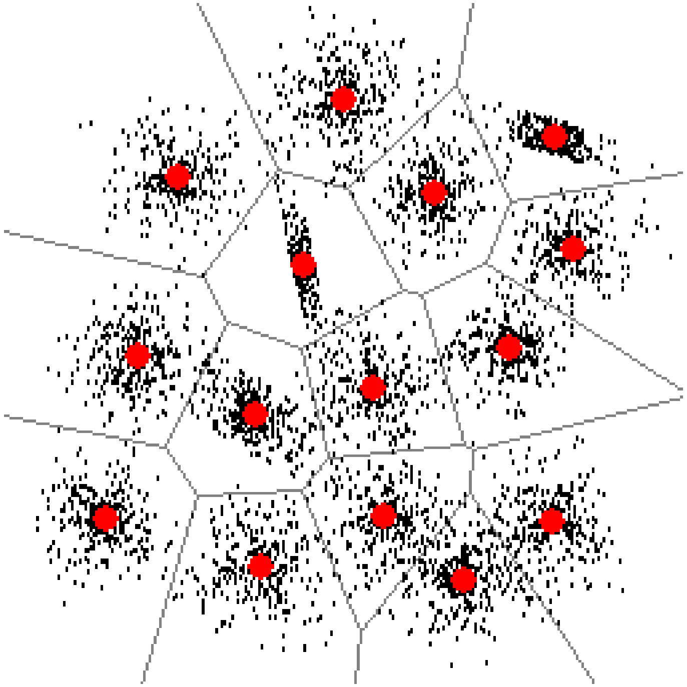
1. Applications
2. **Issues in Clustering**
3. Flat Clustering
4. Hierarchical Clustering



# Issues in Clustering

- Clustering is more difficult than you might think

1. **How many clusters?**
2. **Flat** or **hierarchical?**
3. **Hard** or **soft?**
4. What's a **good** clustering?
5. How to **find** it?

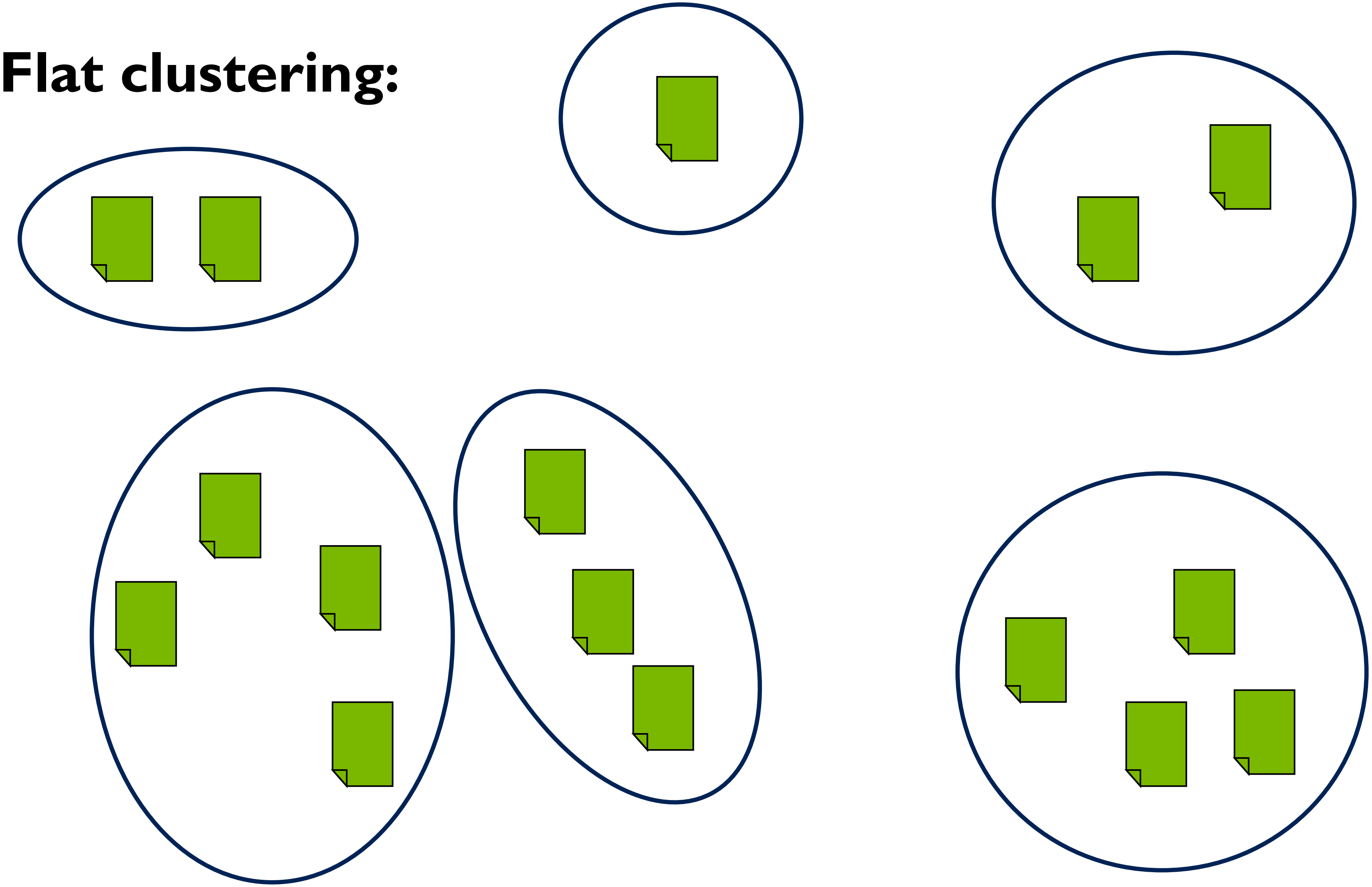


# 1. How Many Clusters?

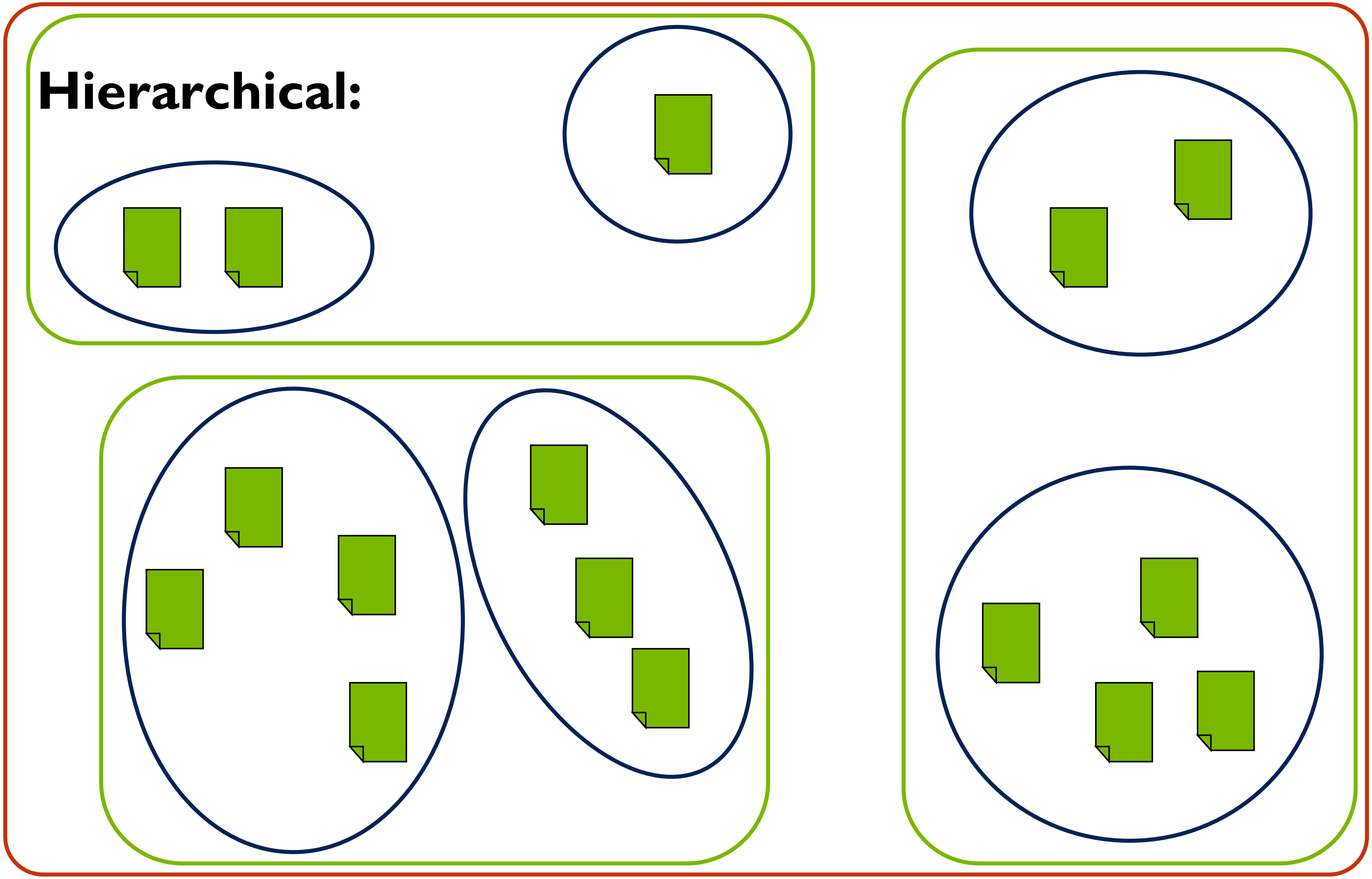
- Let ***k*** denote the **number of clusters** from now on
- Basically, there are two different approaches regarding the choice of *k*
  - Define *k* before searching for a clustering, then only consider clusterings having exactly *k* clusters
  - Do not define a fixed *k*, i.e., let the number of clusters depend on some measure of clustering quality to be defined
- The "right" choice depends on the problem you want to solve…

# 2. Flat or Hierarchical?

**Flat clustering:**



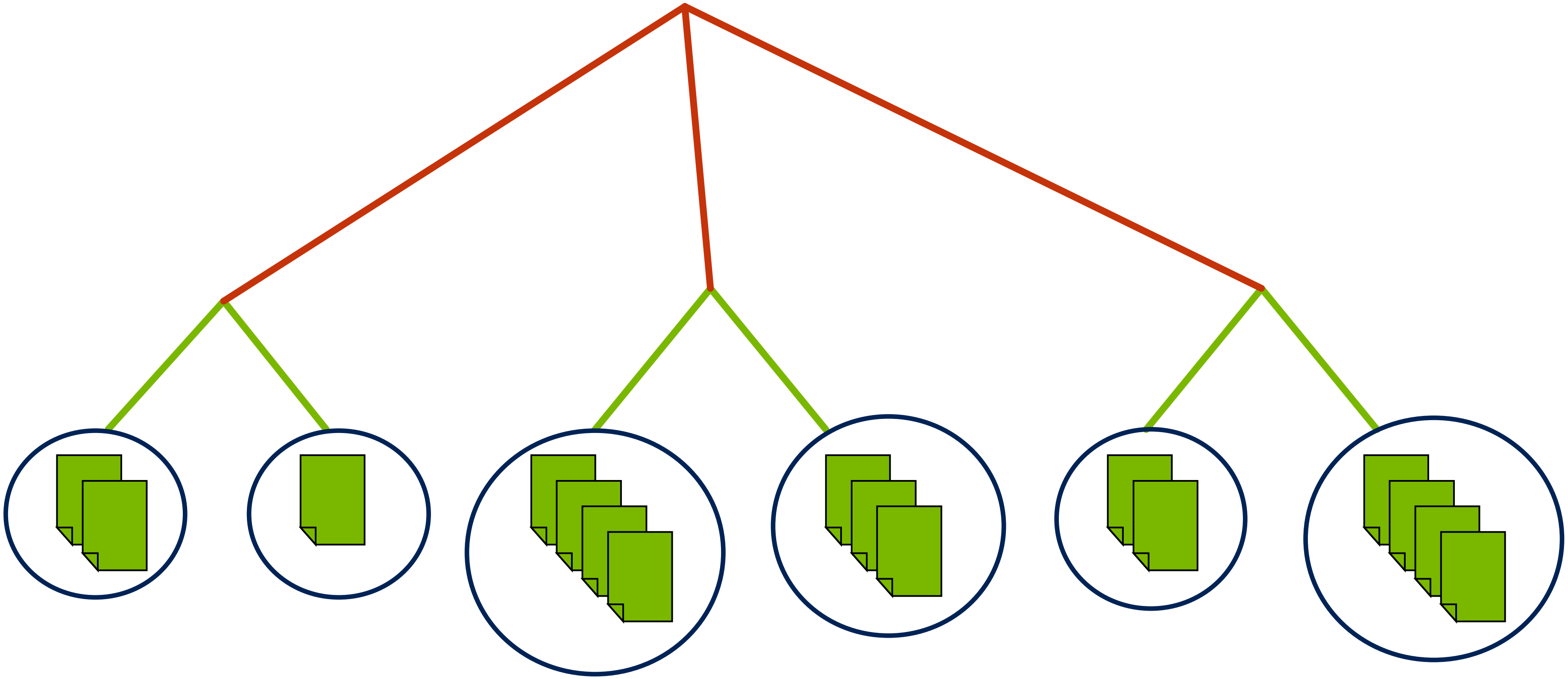
# 2. Flat or Hierarchical?



**Hierarchical:**

# 2. Flat or Hierarchical?

## Hierarchical:



# 3. Hard or Soft?

- **Hard clustering:**
  - Every document is assigned to exactly one cluster (at the lowest level, if the clustering is hierarchical)
  - More common and easier to do

- **Soft clustering:**
  - A document's assignment is a **distribution** over all clusters (fuzzy, probabilistic, or something else)
  - Better suited for creating browsable hierarchies (a knife can be a weapon as well as a tool)
  - Example: **LSI** (*k* clusters/topics)

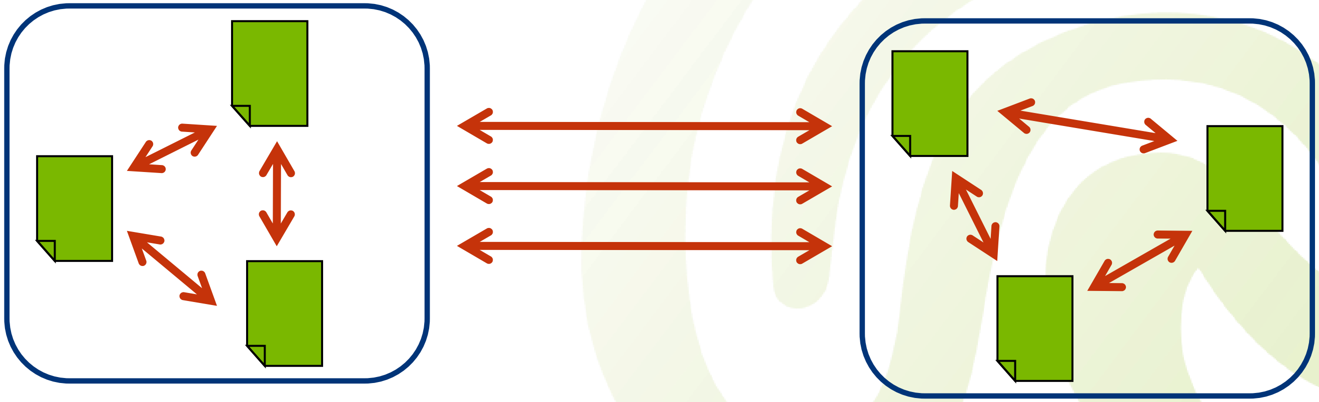
# 4. What's a Good Clustering?

- **Abstract Problem Statement:**
  - **Given:**
    - A **collection** of *n* documents
    - The type of clustering to be found (see previous slides)
    - An **objective function** *f* that assigns a number to any possible clustering of the collection
  - **Task:**
    Find a clustering that minimizes the objective function (or maximizes, respectively)

- Let's exclude a nasty special case:
  We don't want empty clusters!

# 4. What's a Good Clustering?

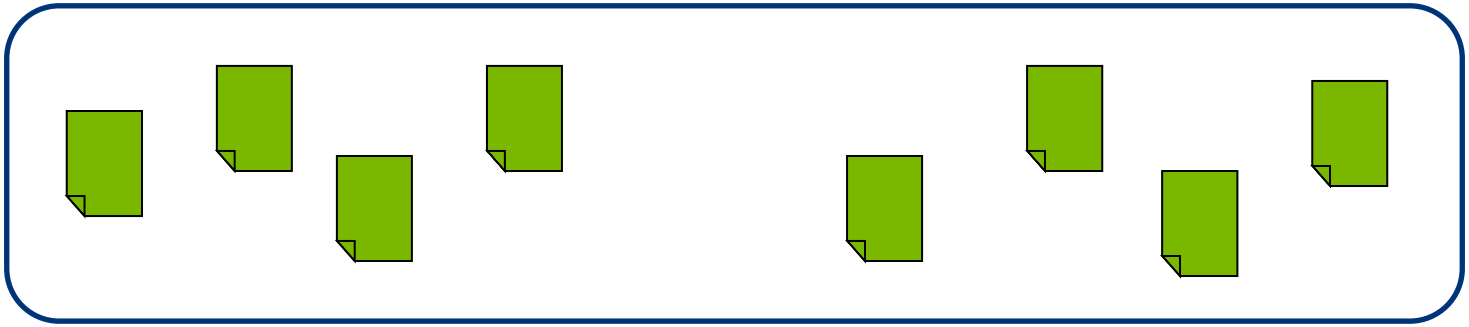
- The **overall quality** of a clustering is measured by *f*
- Usually, *f* is closely related to a **measure of distance** between documents (e.g. cosine similarity)
- Popular **primary goals:**
  - Low inter-cluster similarity, i.e. documents from different clusters should be dissimilar
  - High intra-cluster similarity, i.e. all documents within a cluster should be mutually similar



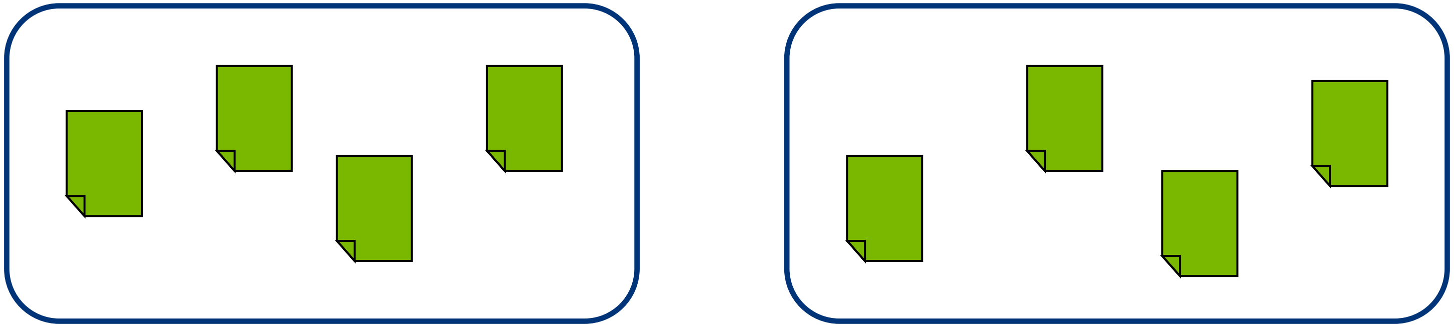
# 4. What’s a Good Clustering?

Inter-cluster similarity and intra-cluster similarity:

**BAD:**



**GOOD:**



# 4. What's a Good Clustering?

- **Common secondary goals:**
  - Avoid very small clusters
  - Avoid very large clusters
  - …
- All these goals are **internal (structural) criteria**
- **External criterion:**
  Compare the clustering
  against a hand-crafted reference clustering (later)

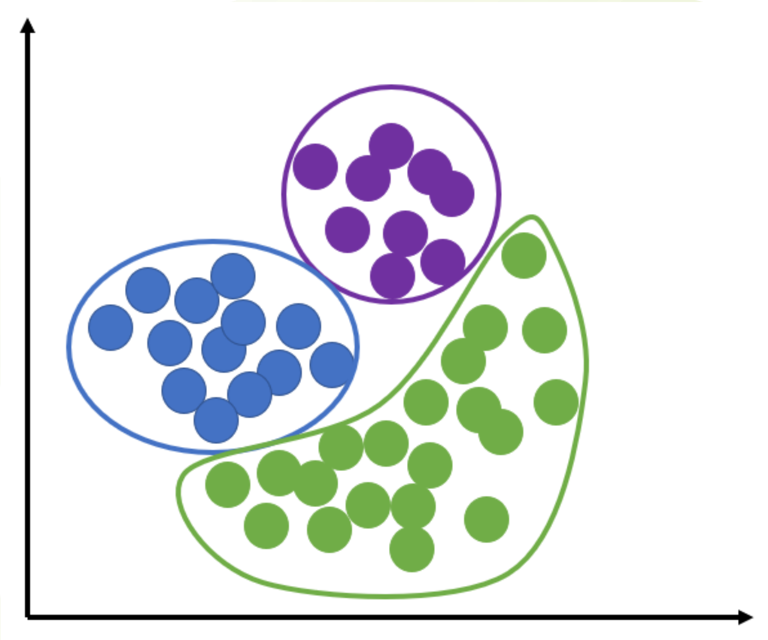
# 5. How to Find a Good Clustering?

- Naïve approach:
  - Try all possible clusterings
  - Choose the one minimizing/maximizing *f*
- Hmm, how many different clusterings are there?
  - There are S(*n*, *k*) distinct hard, flat clusterings of a *n*-element set into exactly *k* clusters
  - S(·, ·) are the **Stirling numbers of the second kind**
  - Roughly: S(*n*, *k*) is exponential in *n*

- The naïve approach fails miserably…
- Let’s use some heuristics…

# Document Clustering

1. Applications
2. Problem Statement
3. **Flat Clustering**
4. Hierarchical Clustering



# K-Means Clustering

- K-means clustering:
  - The most important **(hard) flat clustering** algorithm, i.e., every cluster is a set of documents
  - The number of clusters *k* is defined in advance
  - Documents usually are represented as **unit vectors**
  - **Objective:**
    Minimize the average distance from cluster centers!

- Let's work out a more precise definition of the objective function…

# K-Means Clustering

- **Centroid** of a cluster:
  - Let $A = \{d_1, \ldots, d_m\}$ be a document cluster (a set of unit vectors)
  - The **centroid** of $A$ is defined as:

$$\mu(A) = \frac{1}{m} \sum_{i=1}^{m} d_i$$

- **RSS** of a cluster:
  - Again, let $A$ be a document cluster
  - The **residual sum of squares** (RSS) of $A$ is defined as:

$$\mathrm{RSS}(A) = \sum_{i=1}^{m} \left\| d_i - \mu(A) \right\|^2$$

# K-Means Clustering

$$\mu(A) = \frac{1}{m}\sum_{i=1}^{m} d_i \qquad \mathrm{RSS}(A) = \sum_{i=1}^{m} \left\| d_i - \mu(A) \right\|^2$$

- In k-means clustering, the **quality of the clustering** into (disjoint) clusters $A_1, \ldots, A_k$ is measured by:

$$\mathrm{RSS}(A_1, \ldots, A_k) = \sum_{j=1}^{k} \mathrm{RSS}(A_j)$$

- K-means clustering tries to **minimize this value**
- Minimizing RSS($A_1, \ldots, A_k$) is equivalent to **minimizing the average squared distance** between each document and its cluster's centroid

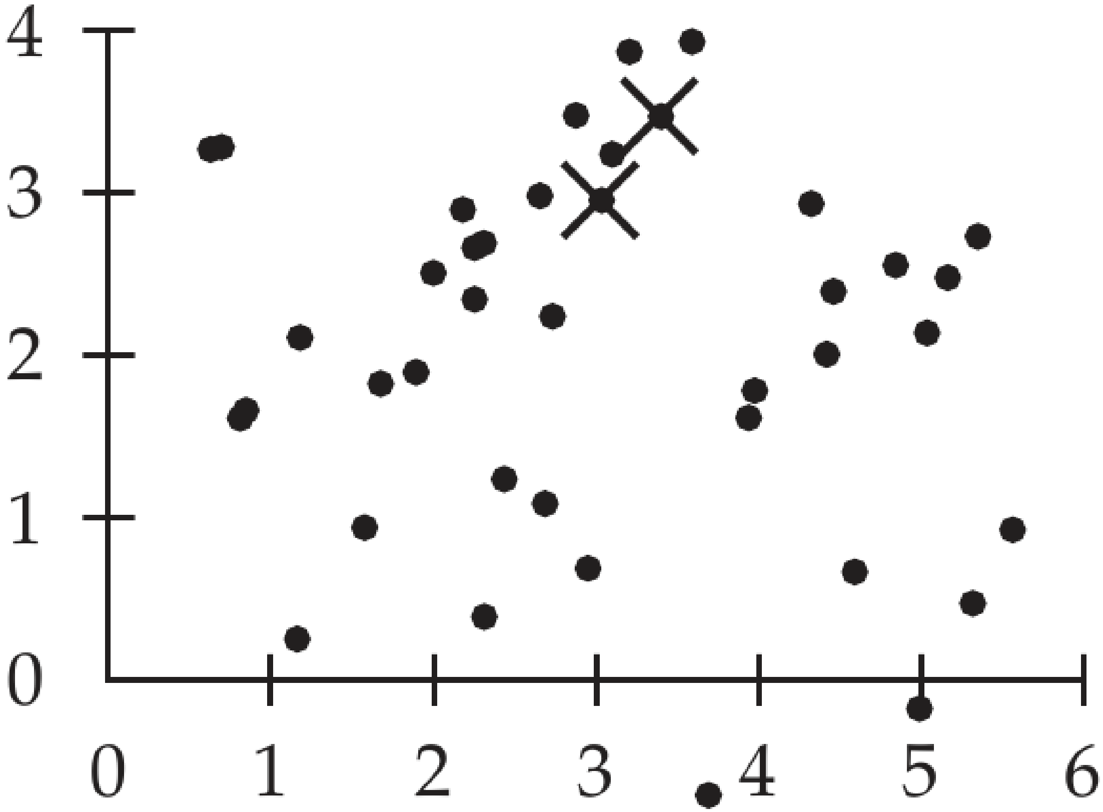
# K-Means Clustering

- The **k-means algorithm (aka Lloyd's algorithm):**
  1. Randomly select *k* documents as **seeds** (= initial centroids)
  2. Create *k* **empty clusters**
  3. Assign exactly one centroid to each cluster
  4. Iterate over the whole document collection:
     Assign each document to the cluster with the nearest centroid
  5. Recompute cluster centroids based on contained documents
  6. Check if clustering is "good enough"; return to (2) if not
- What's "good enough"?
  - Small change since previous iteration
  - Maximum number of iterations reached
  - RSS "small enough"

# K-Means Clustering

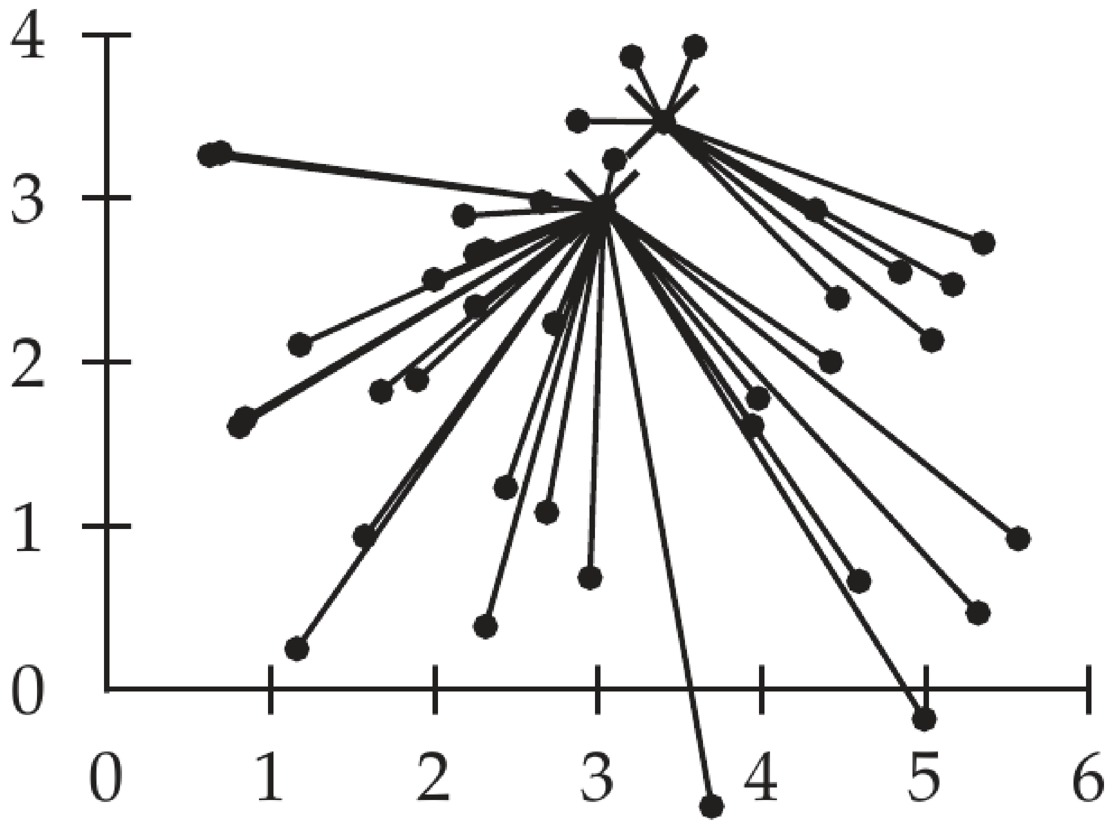
- Example from (Manning *et al.*, 2008):

1. Randomly select *k* = 2 seeds (initial centroids)



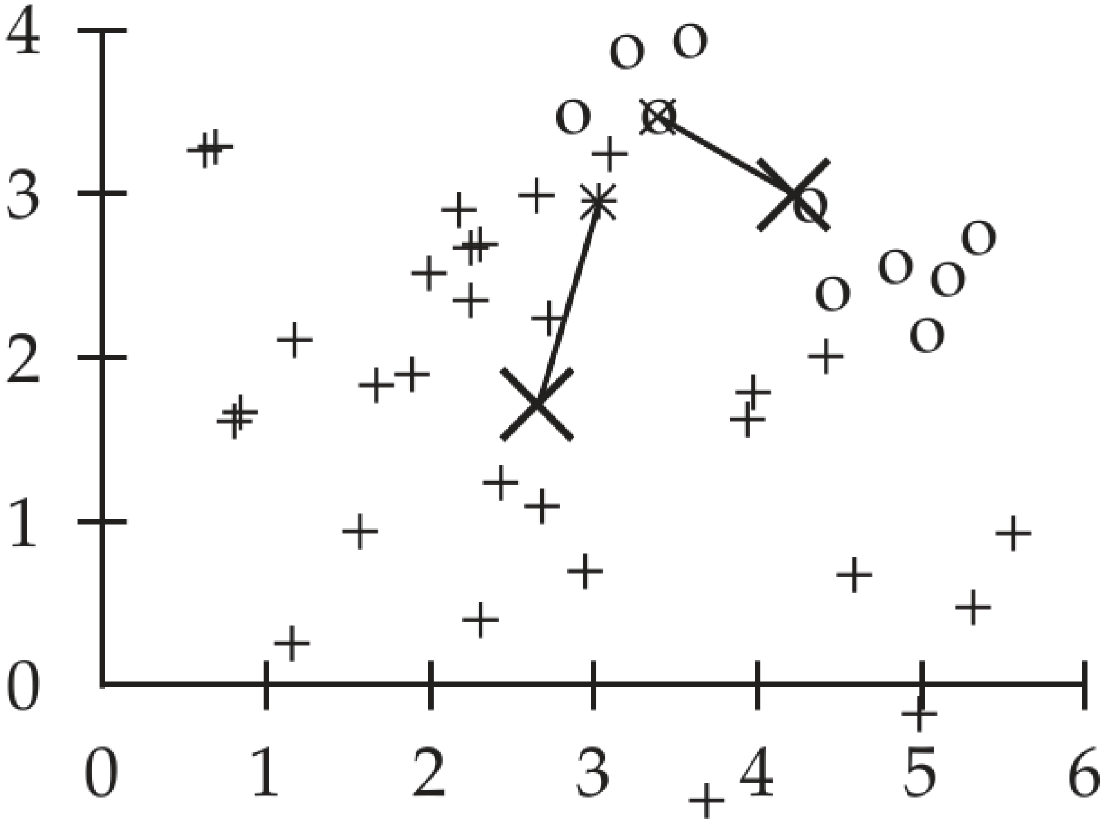
# K-Means Clustering

4. Assign each document to the cluster having the nearest centroid



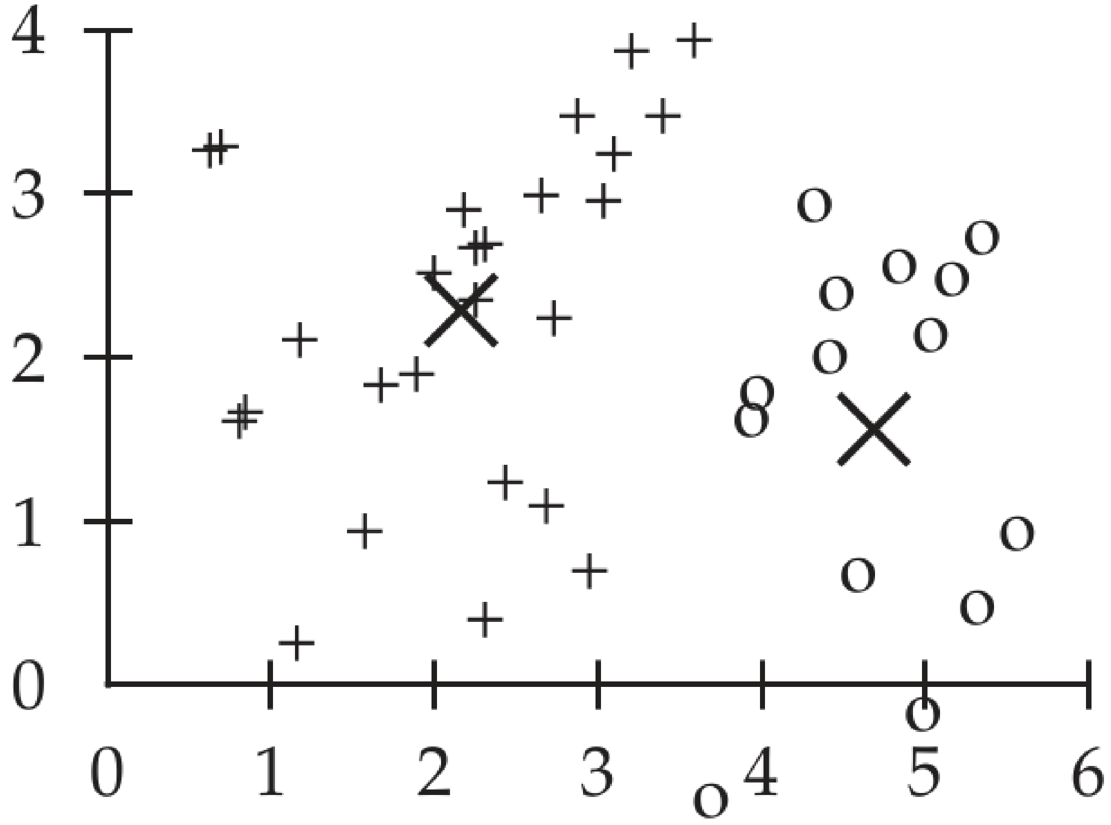
# K-Means Clustering

5. Recompute centroids



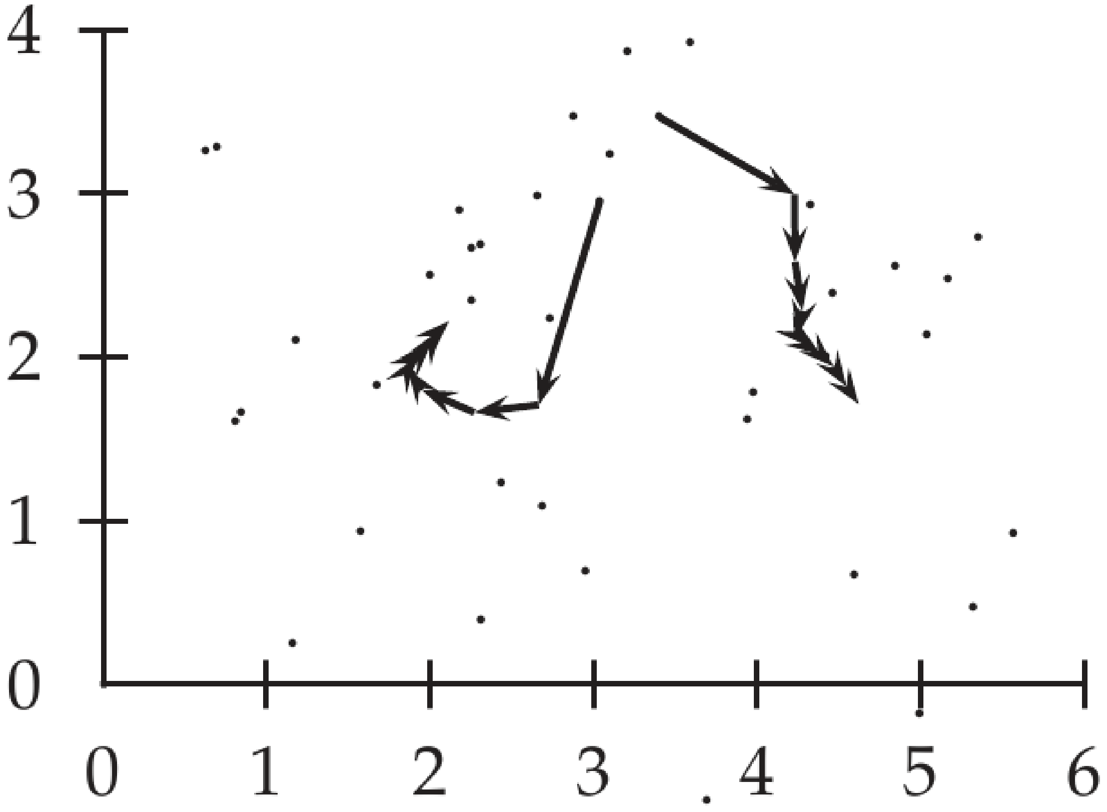
# K-Means Clustering

Result after
9 iterations:



# K-Means Clustering

Movement of centroids in 9 iterations:

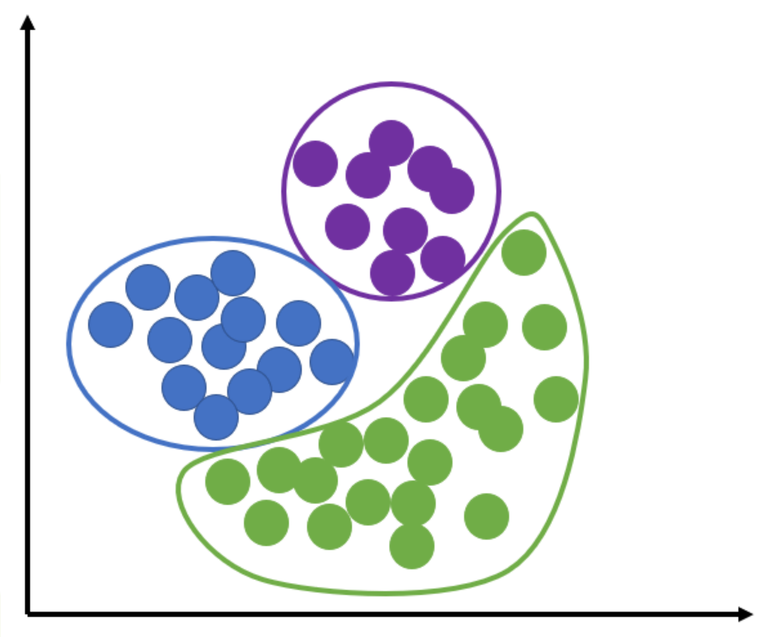


# Variants and Extensions of K-Means

- K-means clustering is a popular representative of the class of **partitional clustering algorithms**
  - Start with an initial guess for *k* clusters, update cluster structure iteratively
- Similar approaches:
  - **K-medoids:** Use document lying closest to the centroid instead of centroid
  - **Fuzzy c-means:** Similar to k-means but soft clustering
  - **Model-based clustering:** Assume that data has been generated randomly around *k* unknown "source points"; find the *k* points that most likely have generated the observed data **(maximum likelihood)**

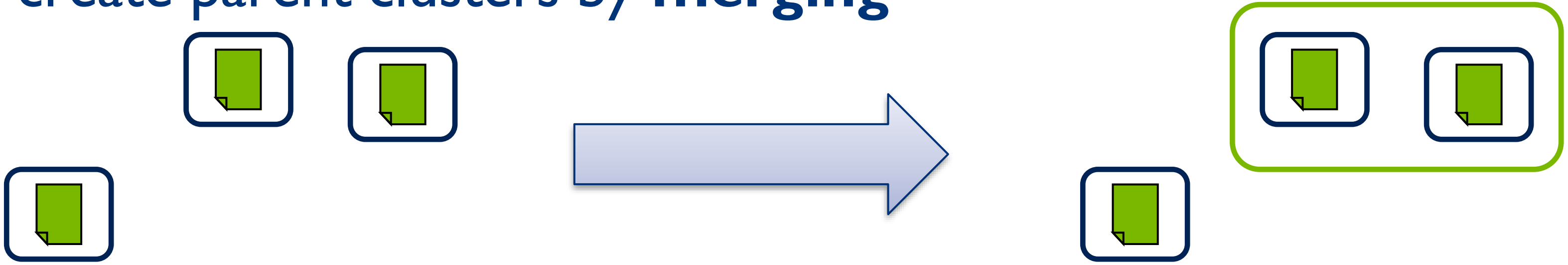
# Document Clustering

1. Applications
2. Issues in Clustering
3. Flat Clustering
4. **Hierarchical Clustering**

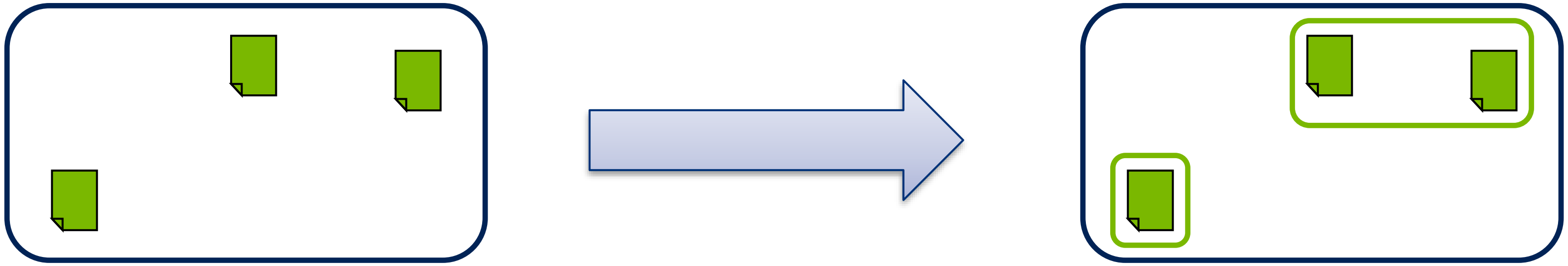


# Hierarchical Clustering

- Two major approaches:
  - **Agglomerative** (bottom-up):
    Start with individual documents as initial clustering, create parent clusters by **merging**



  - **Divisive** (top-down):
    Start with an initial large cluster containing all documents, create child clusters by **splitting**

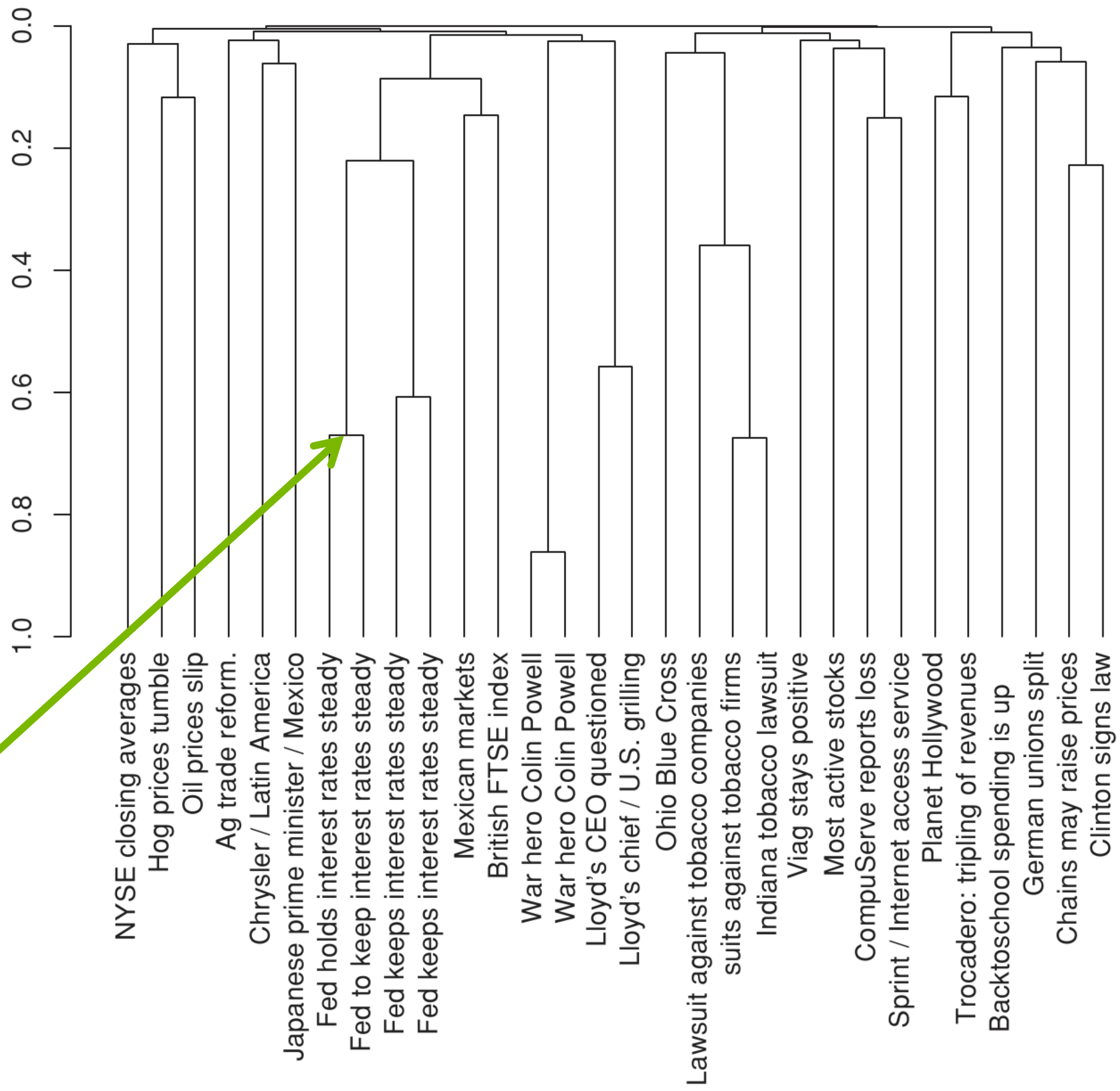


# Agglomerative Clustering

- Assume that we have some measure of similarity between **clusters**
- A simple agglomerative clustering algorithm:
  1. For each document:
     Create a new cluster containing only this document
  2. Compute the similarity between every pair of clusters
     (if there are *m* clusters, we get an *m* × *m* **similarity matrix**)
  3. **Merge** the two clusters having **maximal similarity**
  4. If there is more than one cluster left, go back to (2)

# Agglomerative Clustering

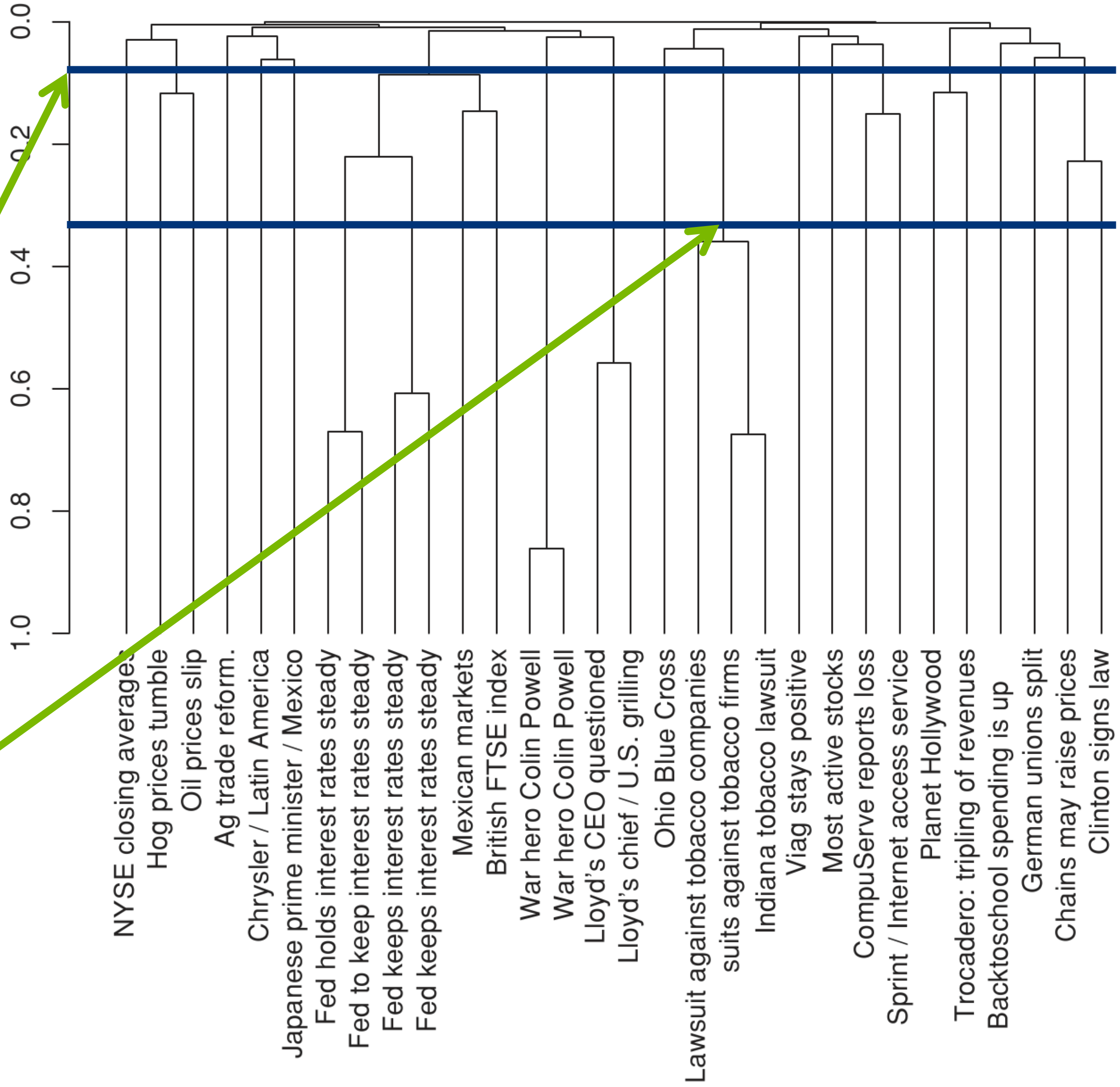
- **Dendrogram** from (Manning *et al.*, 2008):
  - Documents from Reuters-RCV1 collection
  - Cosine similarity



**Cosine similarity of "Fed holds…" and "Fed to keep…" is around 0.68**

# Agglomerative Clustering

- Get non-binary splits by cutting the dendrogram at prespecified levels of similarity



**Gives 17 clusters**

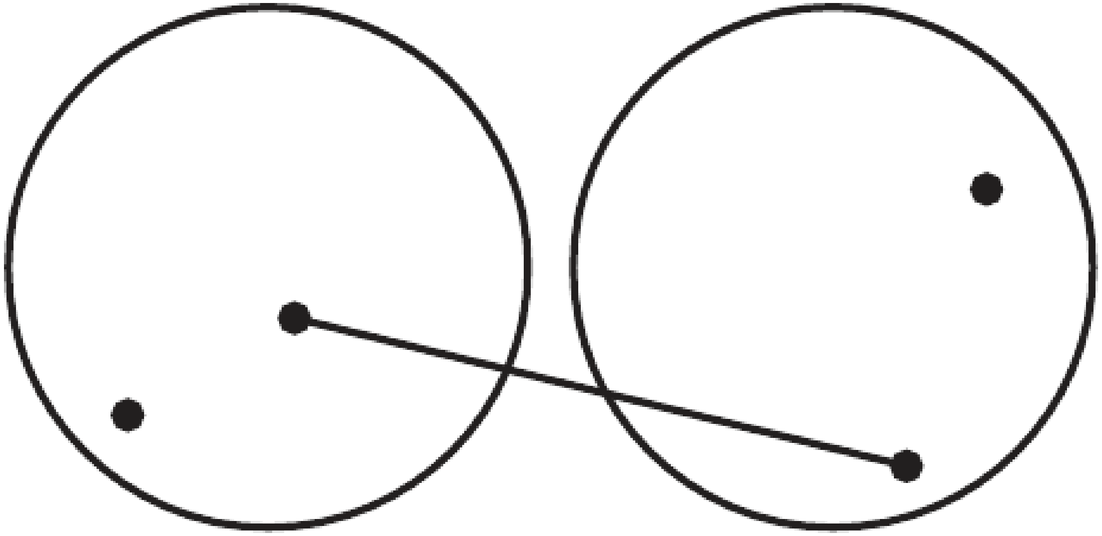
**Gives a cluster of size 3**

# Similarity of Clusters

- We just assumed that we can measure similarity between clusters… But how to do it?
- Typically, measures of **cluster similarity** are derived from some measure of **document similarity** (e.g. Euclidean distance)
- There are several popular definitions of cluster similarity:
  - Single link
  - Complete link
  - Centroid
  - Group average

# Similarity of Clusters

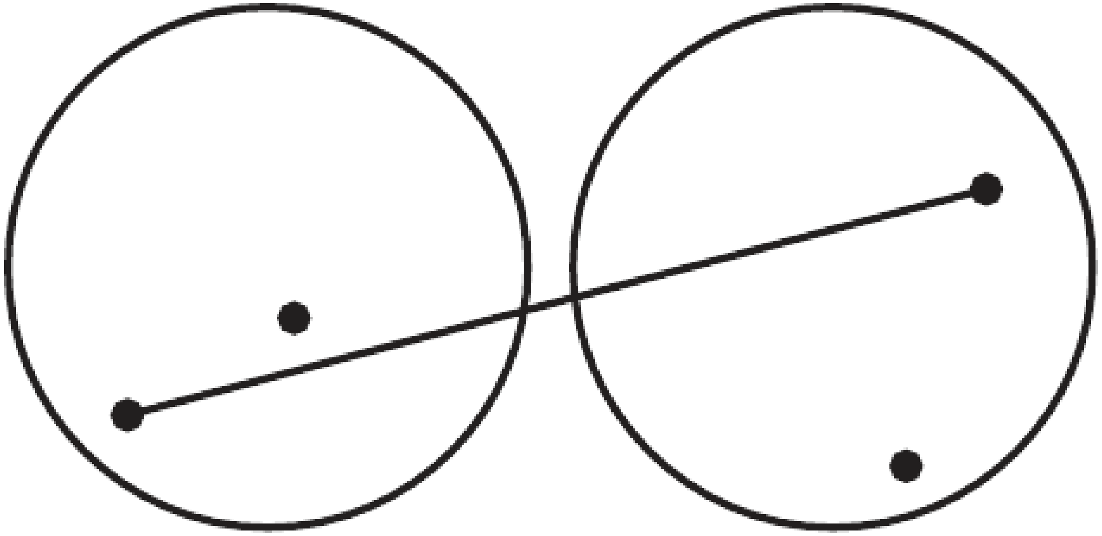
- **Single-link clustering:**
  Similarity of two clusters
  = similarity of their most similar members



- **Problem:**
  Single-link clustering often produces **long chains**

# Similarity of Clusters

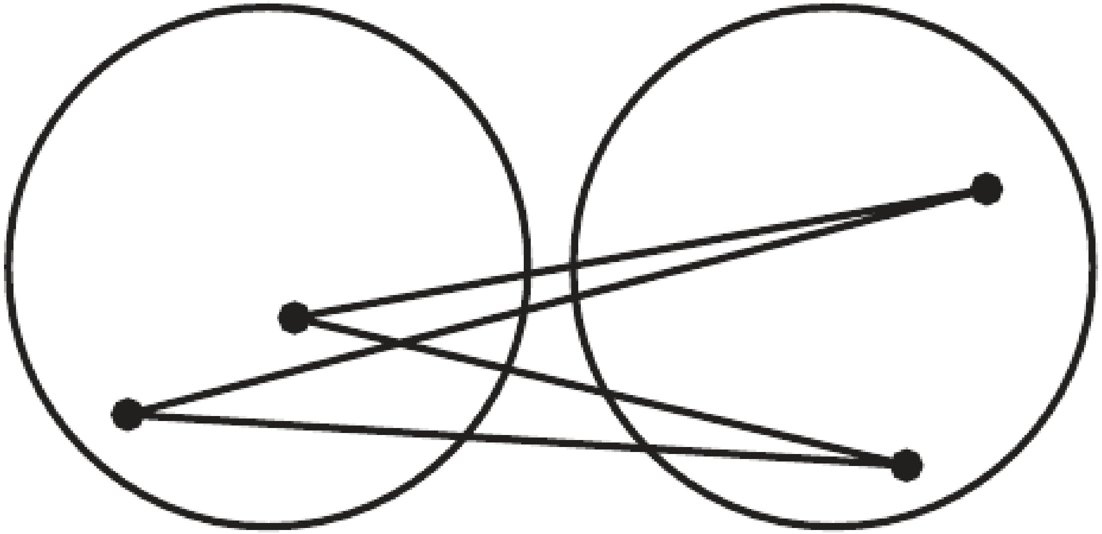
- **Complete-link clustering:**
  Similarity of two clusters
  = similarity of their most dissimilar members



- **Problem:**
  Complete-link clustering is sensitive to outliers

# Similarity of Clusters

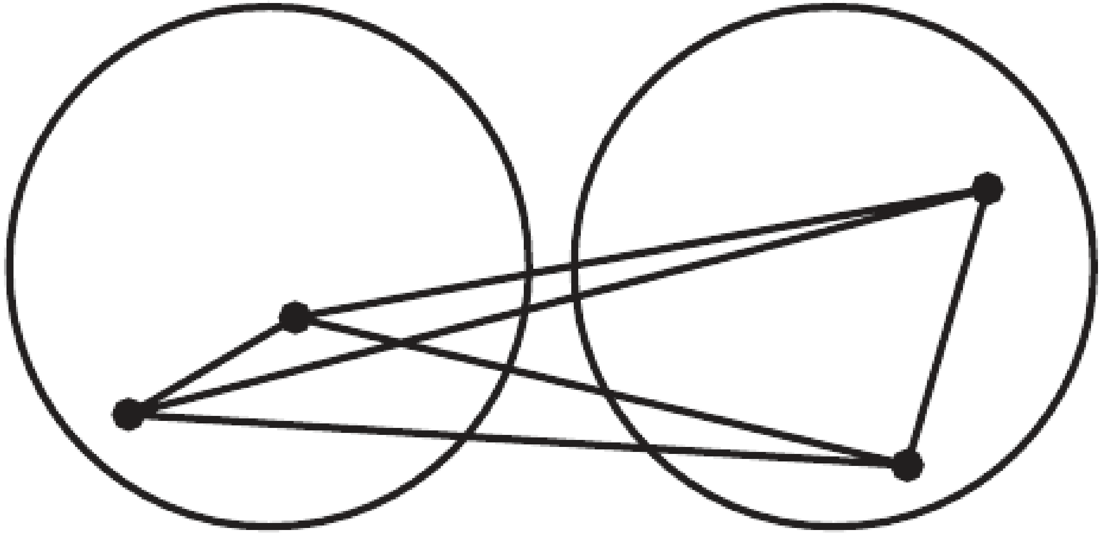
- **Centroid clustering:**
  Similarity of two clusters
  = average inter-similarity (= similarity of centroids)



- **Problem:**
  Similarity to other clusters can improve by merging
  (leads to overlaps in dendrogram)

# Similarity of Clusters

- **Group average clustering:**
  Similarity of two clusters = average of all similarities



- **Problem:**
  Computation is expensive

# Divisive Clustering

- How does **divisive clustering** work?
- We won’t go into details here
- But there is a simple method:
  - Use a flat clustering algorithm as a subroutine to split up clusters (e.g. 2-means clustering)
- Again, there might be **constraints** on clustering quality:
  - Avoid very small clusters
  - Avoid splitting into clusters of extremely different cardinalities
  - …

# Evaluation

- Finally, how to evaluate clusterings?
- We already used **internal criteria**
  (e.g. the total centroid distance for k-means clustering)
- Compare against a manually built reference clustering involves **external criteria**
- **Example:** The **Rand index**
  - Look at all pairs of documents!
  - What **percentage of pairs** are in **correct** relationship?
    - True positives: The pair is correctly contained in the same cluster
    - True negatives: The pair is correctly contained in different clusters
    - False positives: The pair is wrongly contained in the same cluster
    - False negatives: The pair is wrongly contained in different clusters

# Next Lecture

- Relevance Feedback
- Classification