# Information Retrieval & Text Mining

## Document Clustering

**Dr. Saeed Ul Hassan**
**Information Technology University**

# Today’s Topic

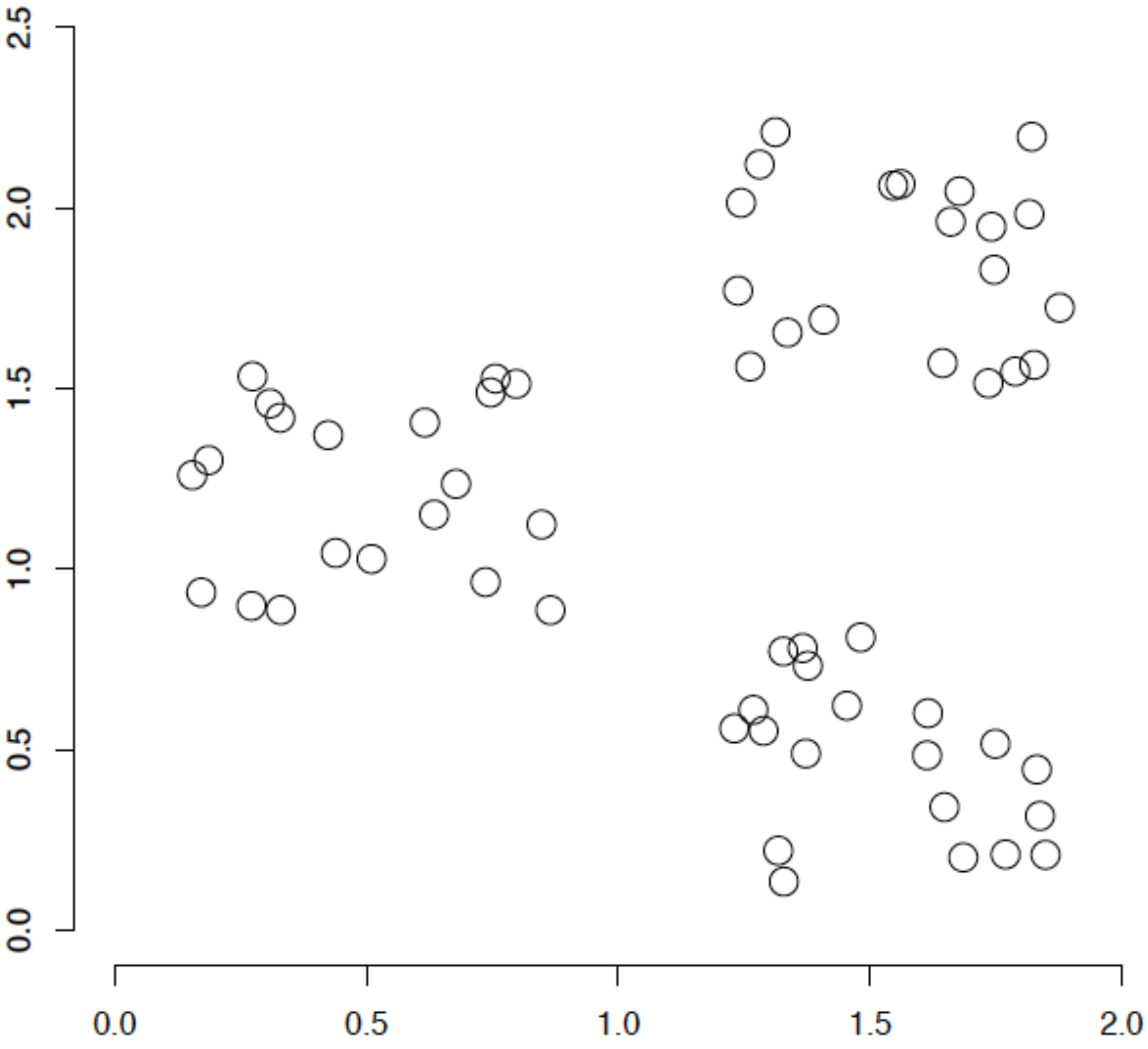
- **Document clustering**
  - Why we need it?
  - Meta-data representation
    - Textual, non textual (images, tables, algorithms, etc)
  - Evaluation
- **Clustering algorithms**
  - Partitioning algorithms
  - Hierarchical algorithms

# What is clustering?

- **Clustering**: *The process of grouping a set of objects into classes of similar objects.*
  - Documents within a cluster should be similar.
  - Documents from different clusters should be dissimilar.

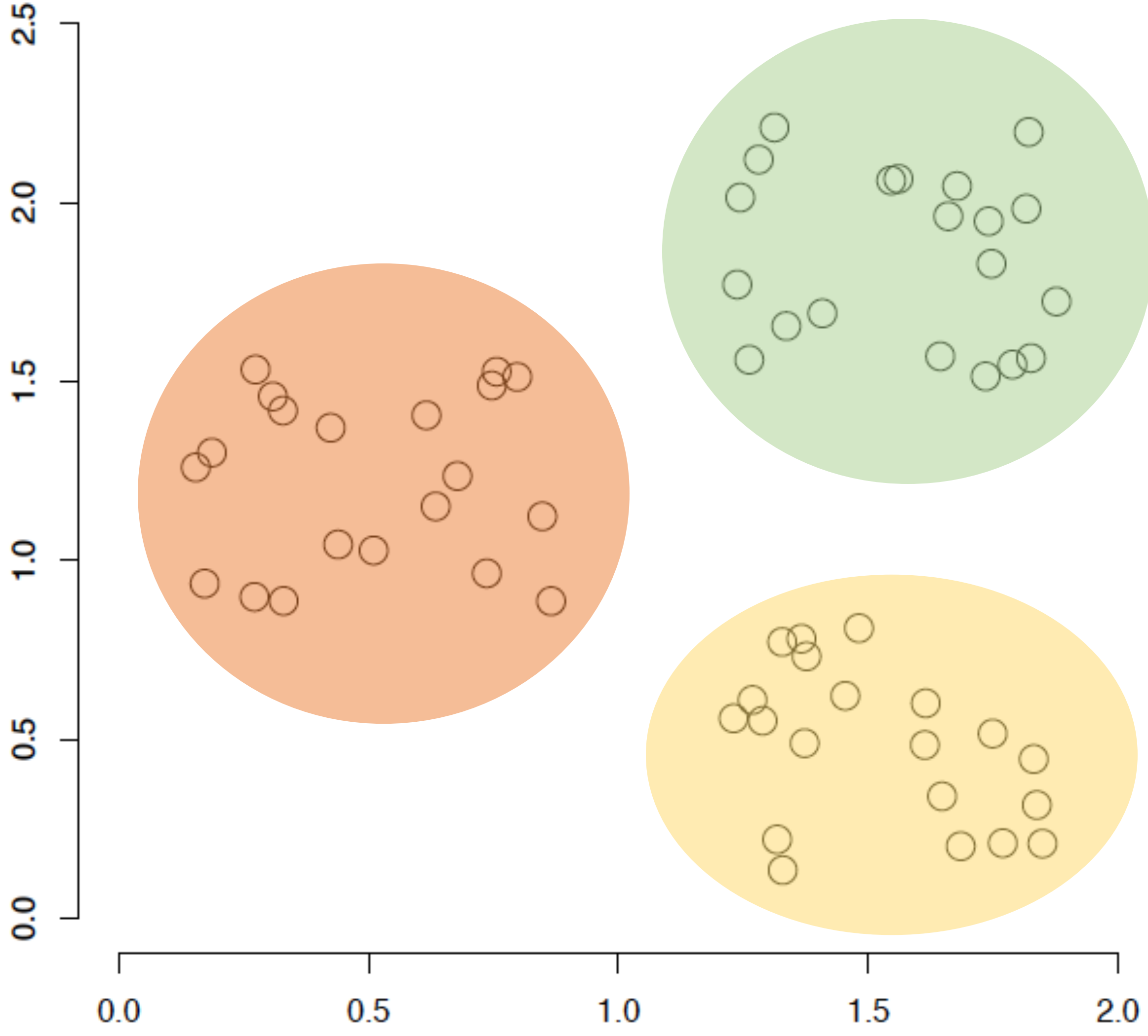
- **The commonest form of *unsupervised learning***
    - **Unsupervised learning** = learning from raw data, as opposed to **Supervised learning** = where a classification of examples are available.
  - A common and important task that finds many applications in IR and other domains.

# A data set with clear cluster structure



- How would you design an algorithm for finding the three clusters in this case?

# A data set with clear cluster structure

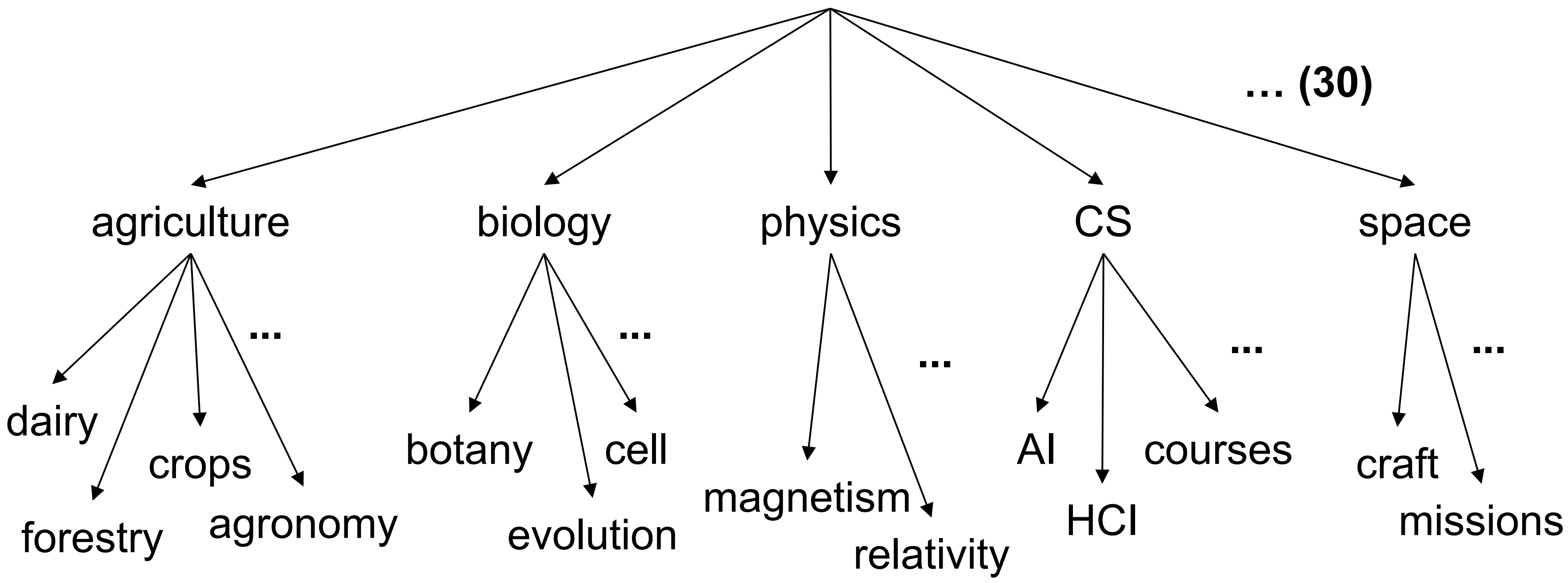


- 3 Clusters

# Applications of clustering in IR

- **Whole corpus analysis/navigation**
  - Better user interface: search without typing
- **For improving recall in search applications**
  - Better search results (like pseudo RF)
- **For better navigation of search results**
  - Effective “user recall” will be higher
- **For speeding up vector space retrieval**
  - Cluster-based retrieval gives faster search

# Hierarchy isn't clustering but is the kind of output you want from clustering



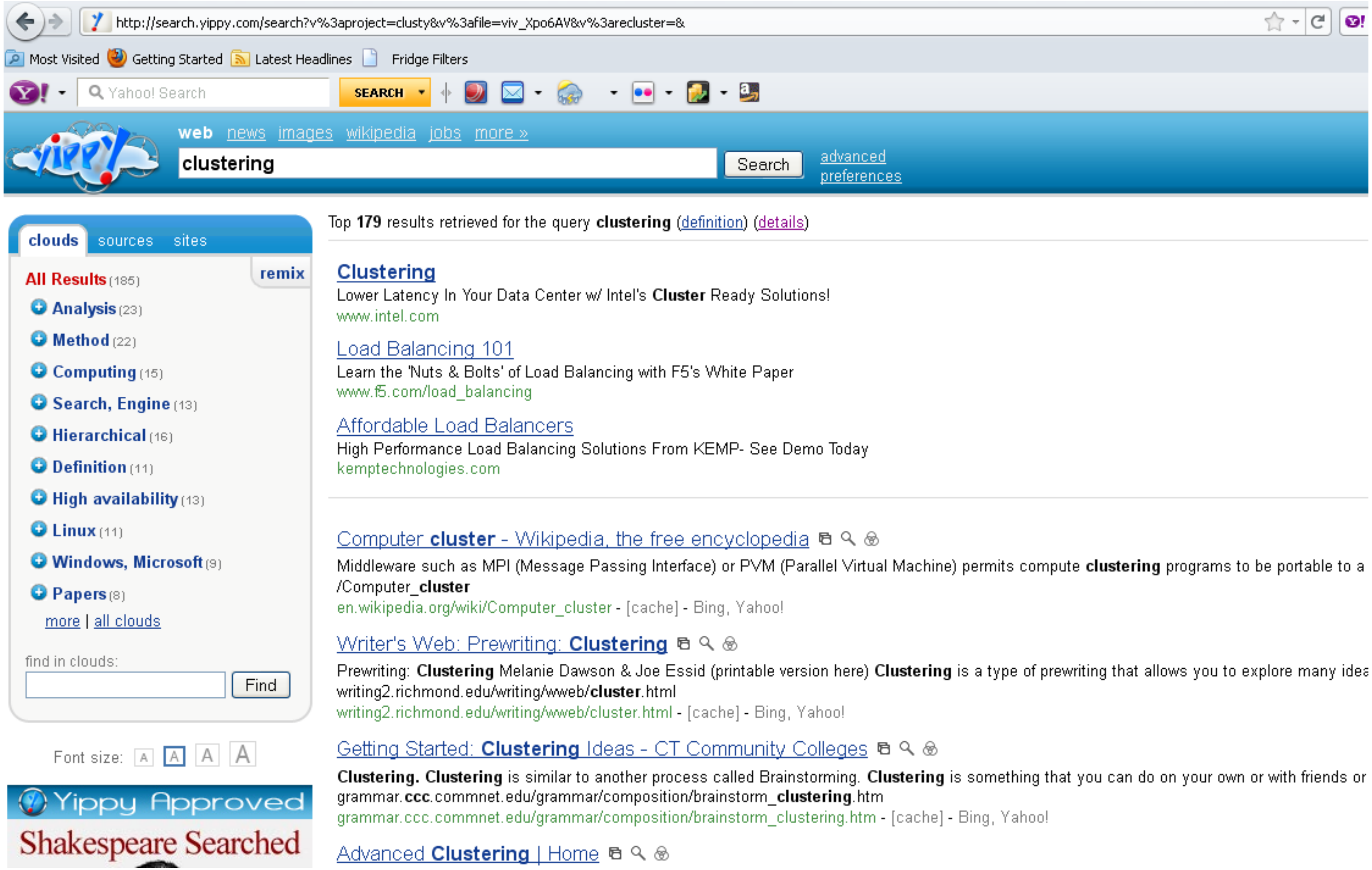
www.yahoo.com/Science

# Applications of clustering in IR

- **Whole corpus analysis/navigation**
  - Better user interface: search without typing
- **For improving recall in search applications**
  - Better search results (like pseudo Relevance Feedback)
- **For better navigation of search results**
  - Effective "user recall" will be higher
- **For speeding up vector space retrieval**
  - Cluster-based retrieval gives faster search

# Applications of clustering in IR

- **Whole corpus analysis/navigation**
  - Better user interface: search without typing
- **For improving recall in search applications**
  - Better search results (like pseudo Relevance Feedback)
- **For better navigation of search results**
  - Improved navigation of high recall
- **For speeding up vector space retrieval**
  - Cluster-based retrieval gives faster search



Top **179** results retrieved for the query **clustering** (definition) (details)

**Clustering**
Lower Latency In Your Data Center w/ Intel's **Cluster** Ready Solutions!
www.intel.com

Load Balancing 101
Learn the 'Nuts & Bolts' of Load Balancing with F5's White Paper
www.f5.com/load_balancing

Affordable Load Balancers
High Performance Load Balancing Solutions From KEMP- See Demo Today
kemptechnologies.com

Computer **cluster** - Wikipedia, the free encyclopedia
Middleware such as MPI (Message Passing Interface) or PVM (Parallel Virtual Machine) permits compute **clustering** programs to be portable to a /Computer_**cluster**
en.wikipedia.org/wiki/Computer_cluster - [cache] - Bing, Yahoo!

Writer's Web: Prewriting: **Clustering**
Prewriting: **Clustering** Melanie Dawson & Joe Essid (printable version here) **Clustering** is a type of prewriting that allows you to explore many idea writing2.richmond.edu/writing/wweb/**cluster**.html
writing2.richmond.edu/writing/wweb/cluster.html - [cache] - Bing, Yahoo!

Getting Started: **Clustering** Ideas - CT Community Colleges
**Clustering. Clustering** is similar to another process called Brainstorming. **Clustering** is something that you can do on your own or with friends or grammar.**ccc**.commnet.edu/grammar/composition/brainstorm_**clustering**.htm
grammar.ccc.commnet.edu/grammar/composition/brainstorm_clustering.htm - [cache] - Bing, Yahoo!

Advanced **Clustering** | Home

# Applications of clustering in IR

- **Whole corpus analysis/navigation**
  - Better user interface: search without typing
- **For improving recall in search applications**
  - Better search results (like pseudo Relevance Feedback)
- **For better navigation of search results**
  - Effective "user recall" will be higher
- **For speeding up vector space retrieval**
  - Cluster-based retrieval gives faster search
  - Match query with median of clusters

# Issues for clustering

- **Representation for clustering**
  - Document representation
  - Need a notion of similarity/distance

- **How many clusters?**
  - Number of clusters
  - Avoid too large or small number of clusters

# 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 browse-able 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.
- We will only do hard clustering today.

# Clustering Algorithms

- **Partitioning algorithms**
  - Usually start with a random (partial) partitioning
  - Refine it iteratively
    - *K* means clustering

- **Hierarchical algorithms**
  - Bottom-up
  - Top-down

# Partitioning Algorithms

- **Partitioning method**: Construct a partition of *n* documents into a set of *K* clusters
- **Given**: a set of documents and the number *K*
- **Find**: a partition of *K* clusters that optimizes the chosen partitioning criterion
  - Globally optimal
    - Intractable for many objective functions
  - Effective heuristic methods: *K*-means algorithm

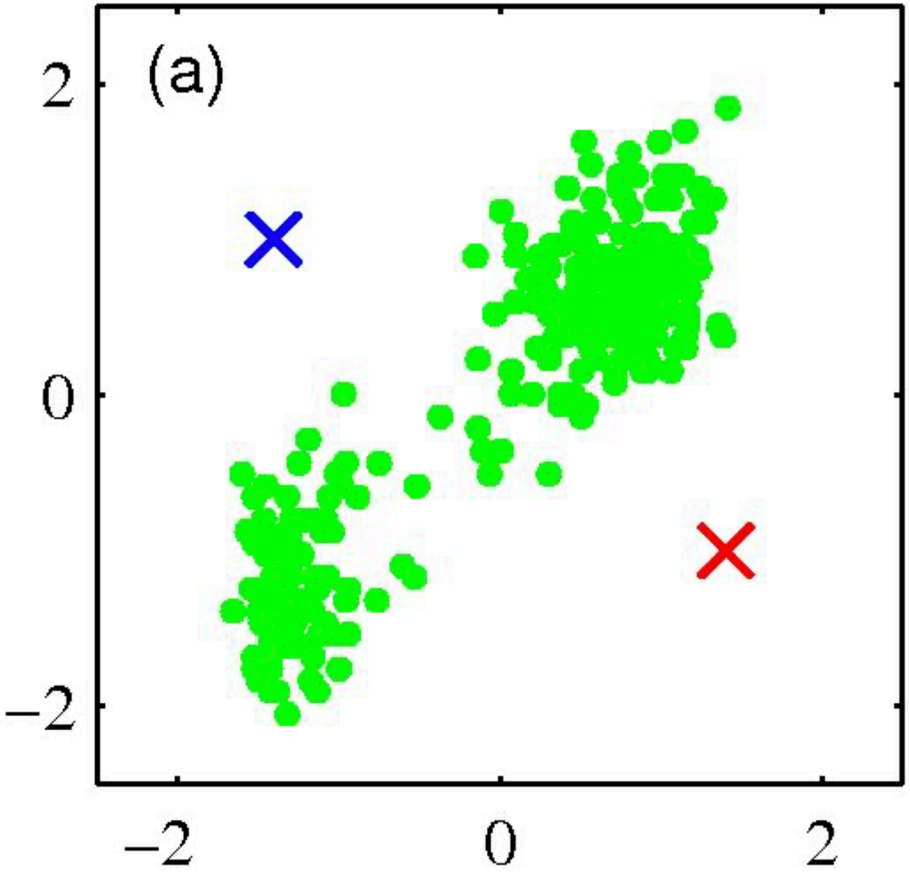
# K-Means

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

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

- Reassignment of instances to clusters is based on distance to the current cluster centroids.

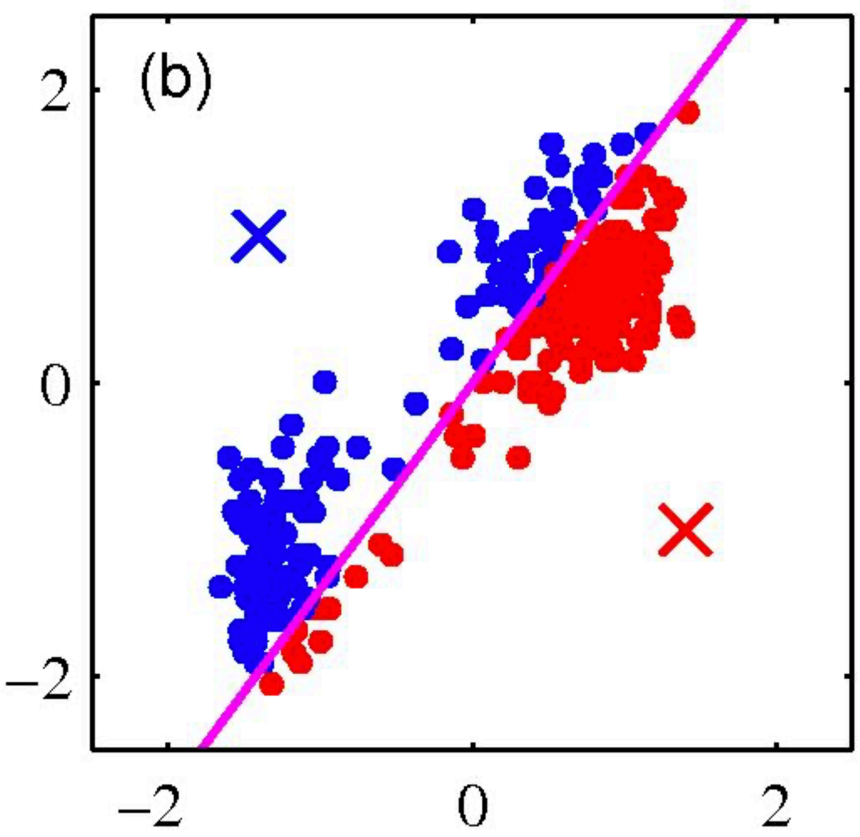
# K-Means



Pick *K* random points as cluster centres (means)

Shown here for *K*=2

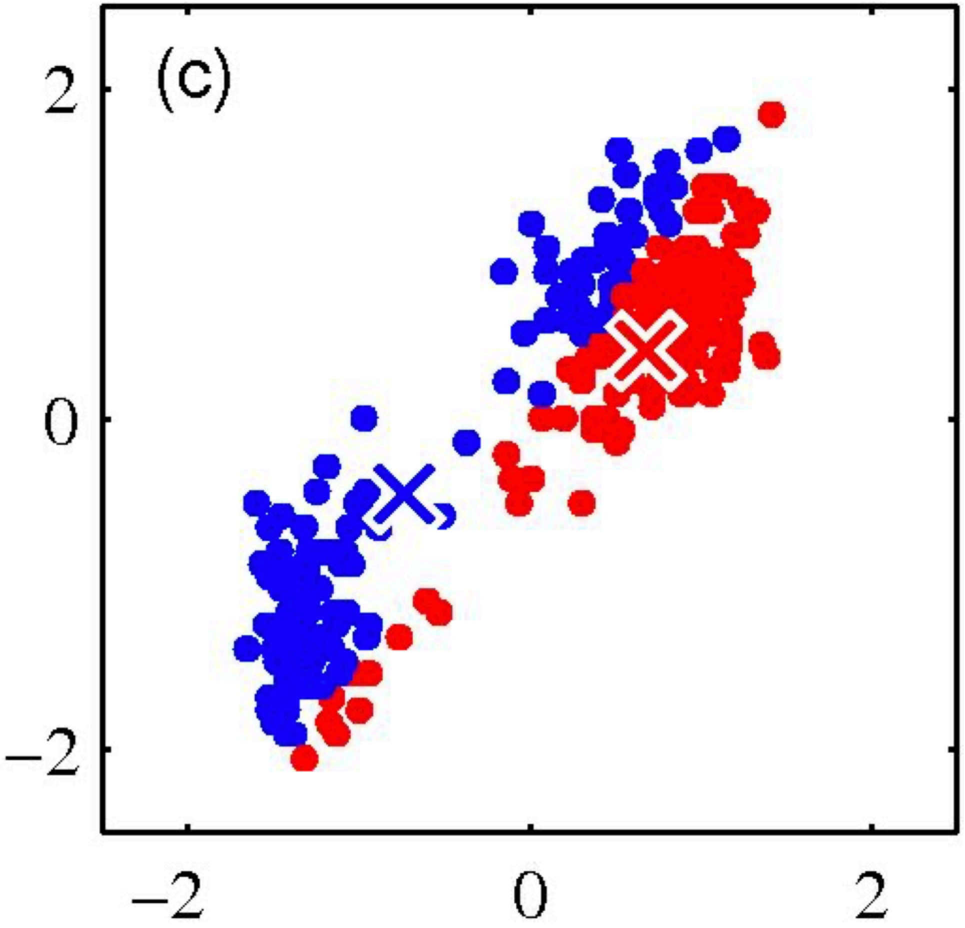
# K-Means



**Iterative Step 1**

- Assign data points to closest cluster canter

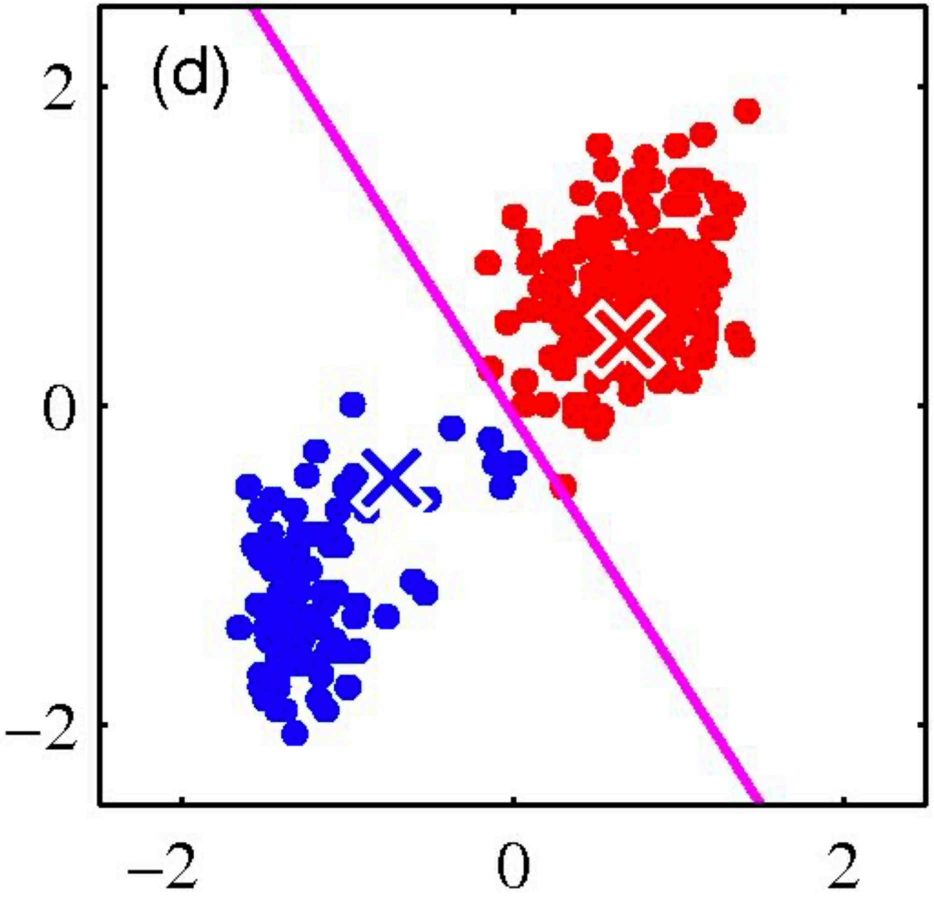
# K-Means



**Iterative Step 2**

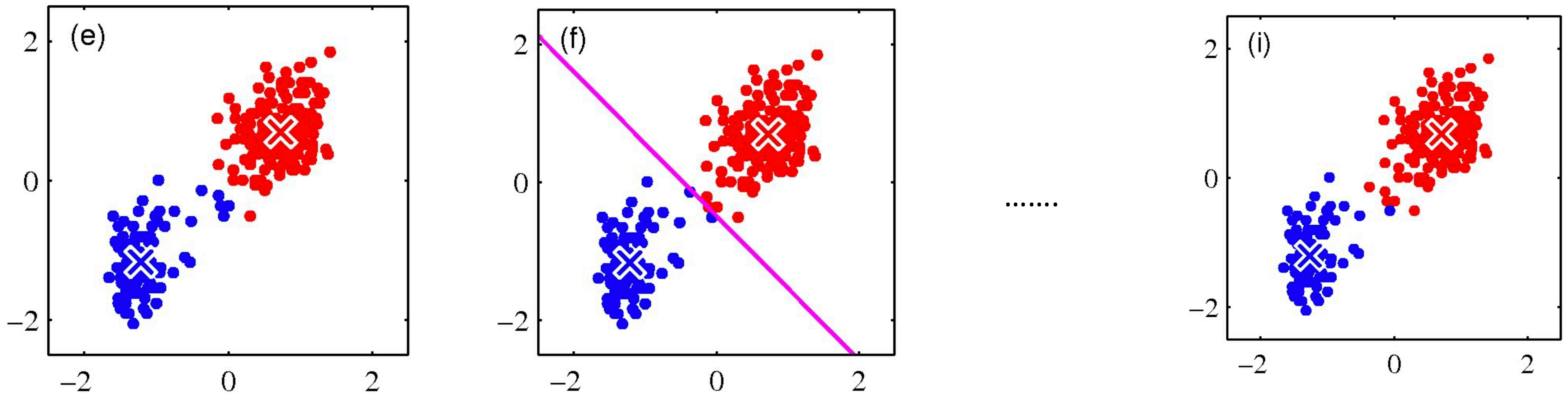
- Change the cluster center to the average of the assigned points

# K-Means



Repeat until convergence

# K-Means



.......

# Termination conditions

- Several possibilities, e.g.,
  - A fixed number of iterations.
  - Doc partition unchanged.

# How Many Clusters?

- Number of clusters *K* is given
  - Partition *n* docs into predetermined number of clusters

- Finding the "right" number of clusters is part of the problem
  - Given docs, partition into an "appropriate" number of subsets.
  - E.g., for query results - ideal value of *K* not known up

- Tradeoff between having more clusters (better focus within each cluster) and having too many clusters

# 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*

- Assume documents with $C$ gold standard classes, while our clustering algorithms produce $K$ clusters, $\omega_1$, $\omega_2$, …, $\omega_K$ with $n_i$ members.

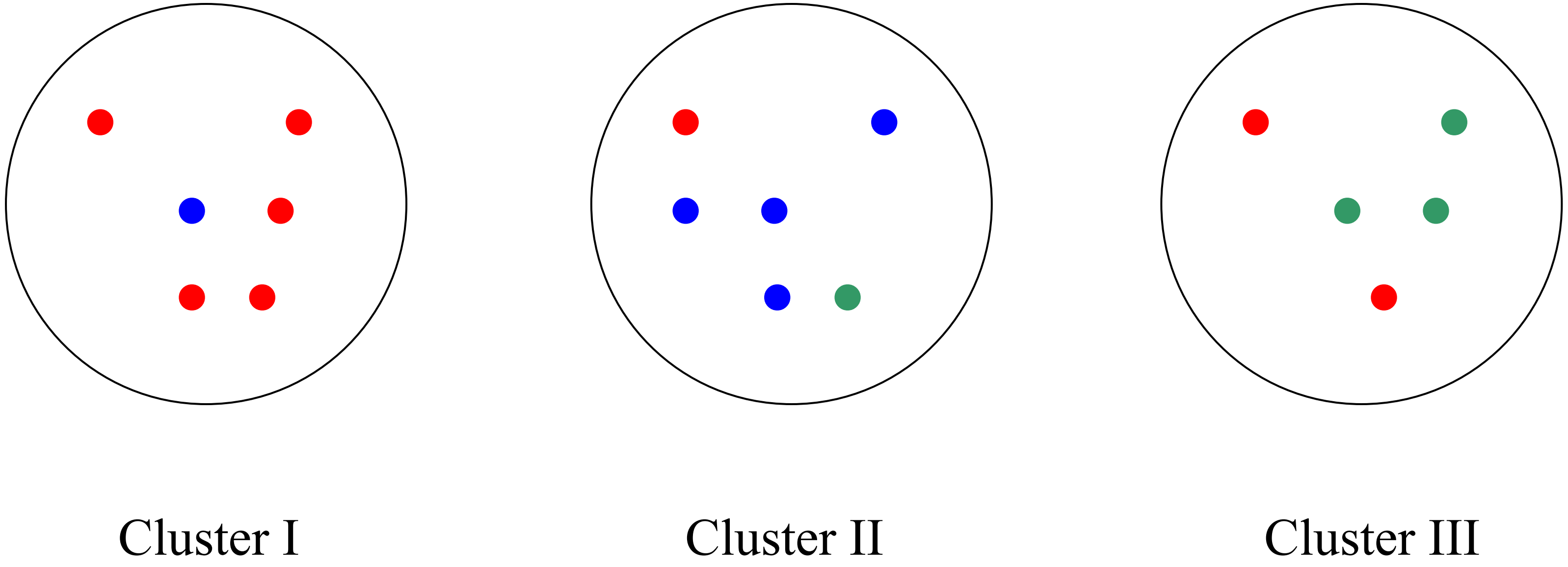
# External Evaluation of Cluster Quality

- Simple measure: purity, the ratio between the dominant class in the cluster $\omega_i$ and the size of cluster $\omega_i$

$$Purity(\omega_i) = \frac{1}{n_i} \max{}_j (n_{ij}) \quad j \in C$$

- Biased because having *n* clusters maximizes purity
- Others are entropy of classes in clusters (or mutual information between classes and clusters)

# Purity example



Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

# Final words

- 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, ...