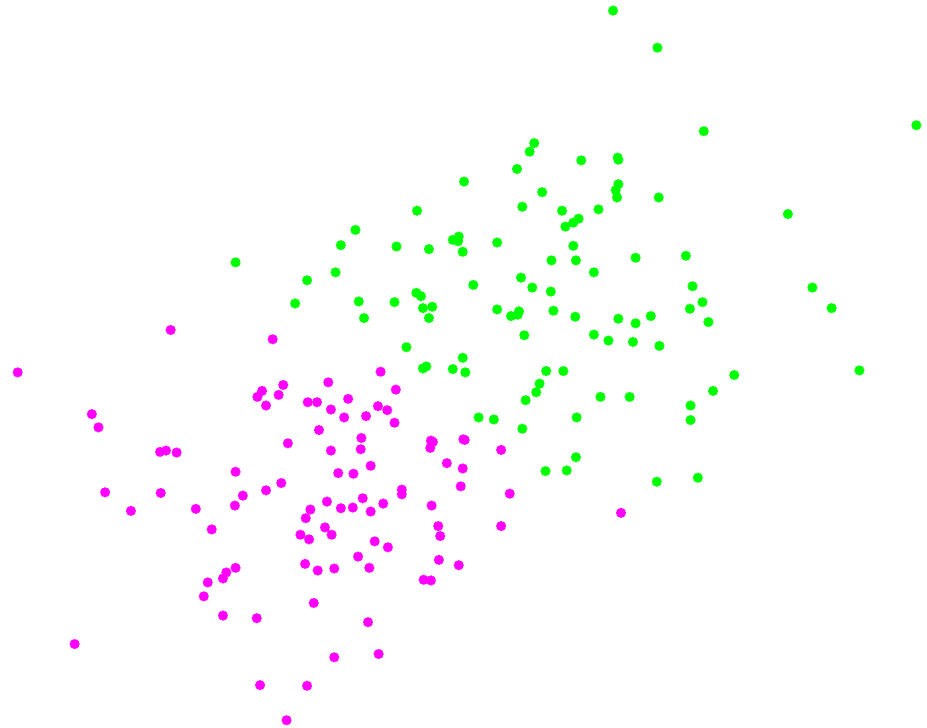
# Machine learning

## TM the Great Unread

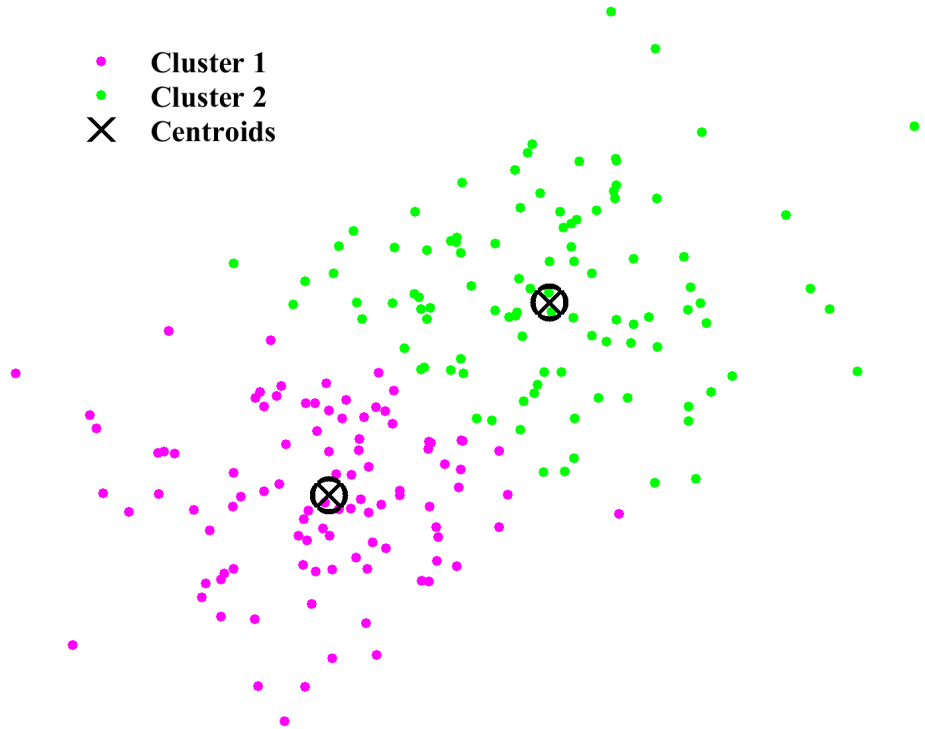
DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

July 29, 2016

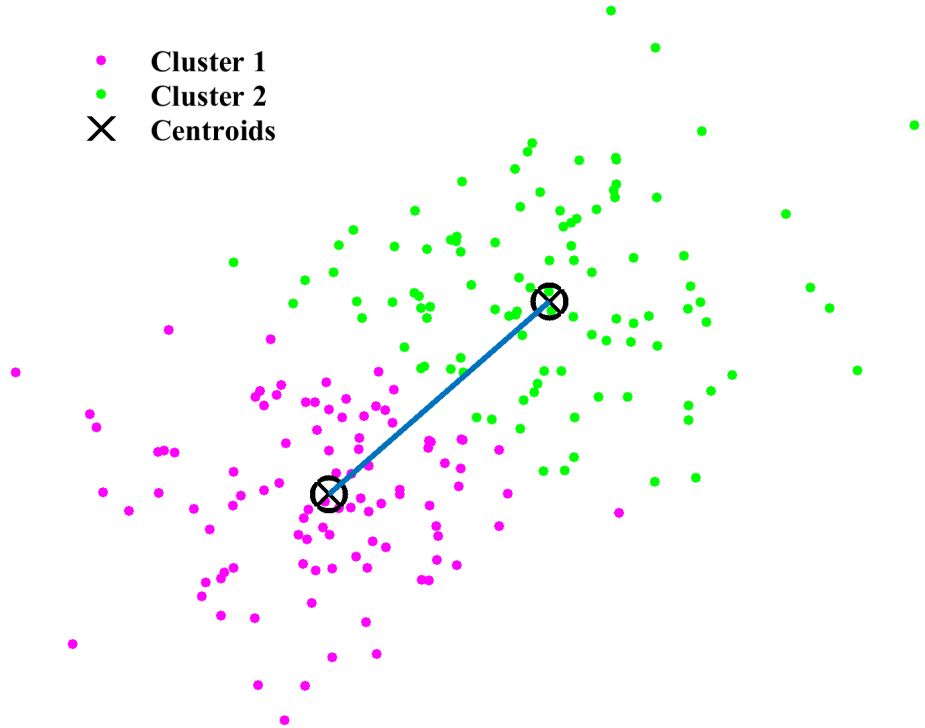

Implicit assumption that we study differences in variables (e.g., terms) between homogeneous objects (e.g., documents)



Systematic differences between objects result in non-random subsets that are often ignored




Cluster analysis: partitions data into homogeneous subsets using inter-object similarity/distance measures




Minimize distance between the centroid and points within each cluster
Maximize distance centroids and points between clusters

| | $t_1$ | $t_2$ | ... | $t_k$ |
|---|---|---|---|---|
| $d_1$ | $f_{d_1,t_1}$ | | | |
| $d_2$ | | $f_{d_2,t_2}$ | | |
| ... | | | ... | |
| $d_n$ | | | | $f_{d_n,t_n}$ |

$\Rightarrow$

| | $d_1$ | $d_2$ | ... | $d_n$ |
|---|---|---|---|---|
| $d_1$ | 0 | | | |
| $d_2$ | | 0 | | |
| ... | | | ... | |
| $d_n$ | | | | 0 |

$$C = \{d_{1,C_1}, d_{2,C_2}, ..., d_{n,C_k}\} \text{ where } k \leq n$$

Convert our matrix of $n$ documents measured on $k$ terms to a matrix of inter-document similarity and then apply a clustering method to the similarity/distance matrix

Either because we want conceptually meaningful groups of documents (or terms) that share common characteristics *or* because we want useful groups that abstract from the individual documents (summarization or compression)

Clustering for understanding or utility

# $k$-means clustering

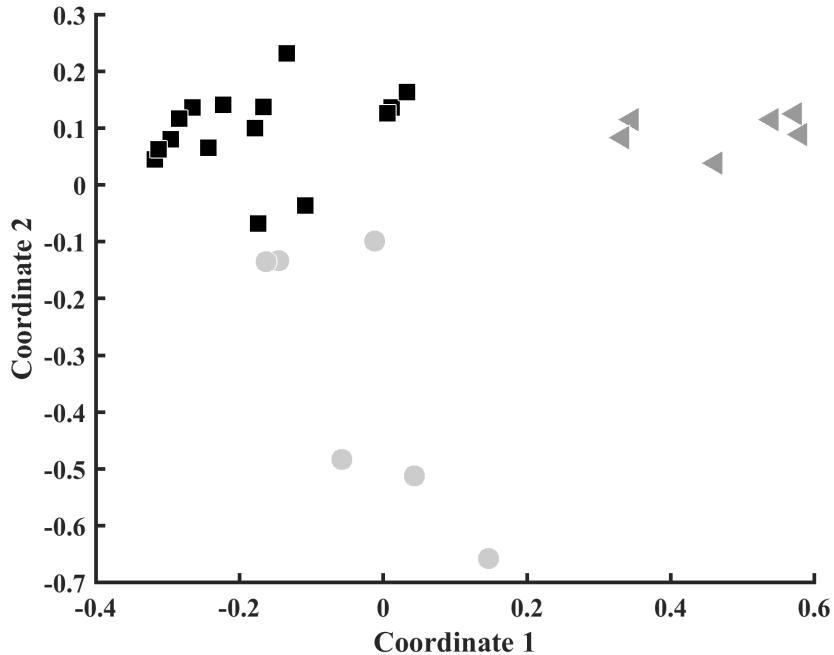
Widely used clustering methods that partitions $n$ documents (or terms) in $k$ clusters

Clusters are non-overlapping, so a document belong exclusively to one cluster

| | |
|---|---|
| 1. | select $k$ points as initial centroids |
| 2. | **repeat** |
| 3. | form $k$ clusters by assigning each point to its closest centroid |
| 4. | recompute the centroid of each cluster |
| 5. | **until** centroid do not change |

$k$-means is a prototyped-based clustering method that finds a centroid (mean) of all the points in a cluster and minimizes the distance (within-cluster sum of squares) of each point to centroid

# bibleNLP




Principle Component Analysis of the DTM is often used for visualization purpose

# Agglomerative hierarchical clustering

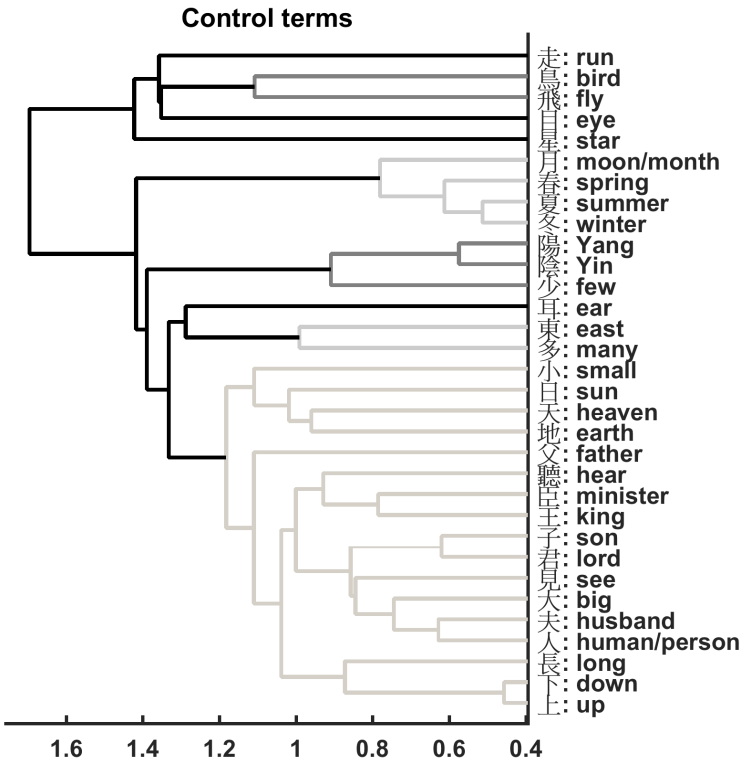
Set of clustering methods that starts with each document as a single cluster and then repeatedly merge the two closest clusters until a single, all encompassing cluster remains (alternate methods use divisive clustering)

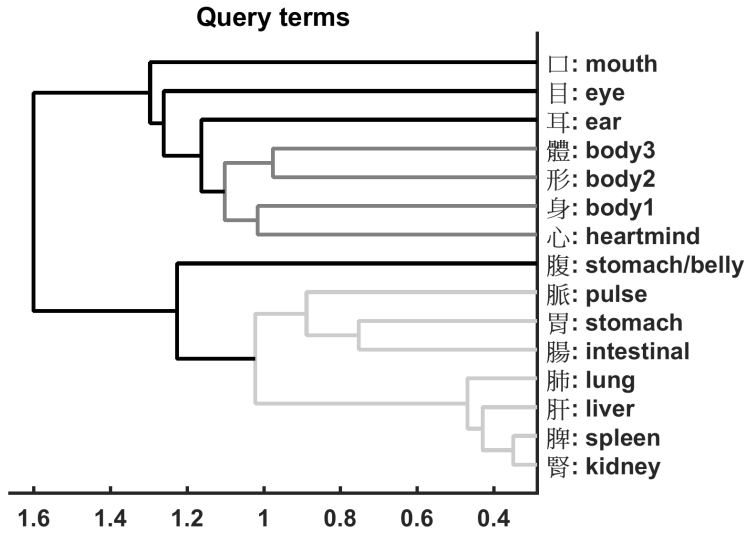
Hierarchical clustering produce nested clusters that are organized in a tree-like structure (visualized with a dendrogram)

| | |
|---|---|
| 1. | compute proximity matrix |
| 2. | **repeat** |
| 3. | merge the closest two clusters |
| 4. | update the proximity matrix to reflect the distance between the new clusters and the original clusters |
| 5. | **until** only on cluster remains |

To compute the proximity between groups of data points a particular technique is chosen (e.g. MIN, MAX, group average)

# Dualism in Ancient Chinese Litt.




With hierarchical clustering you cut or prune the tree at some level to define clusters.