AI6122 Text Data Management & Analysis

Topic: Clustering

Roadmap

- Document clustering
  - Document representations
  - Distance/Similarity
- Clustering algorithms
  - Partitional
  - Hierarchical
  - DBSCAN
- Clustering Evaluation
  - Purity
  - Rand Index

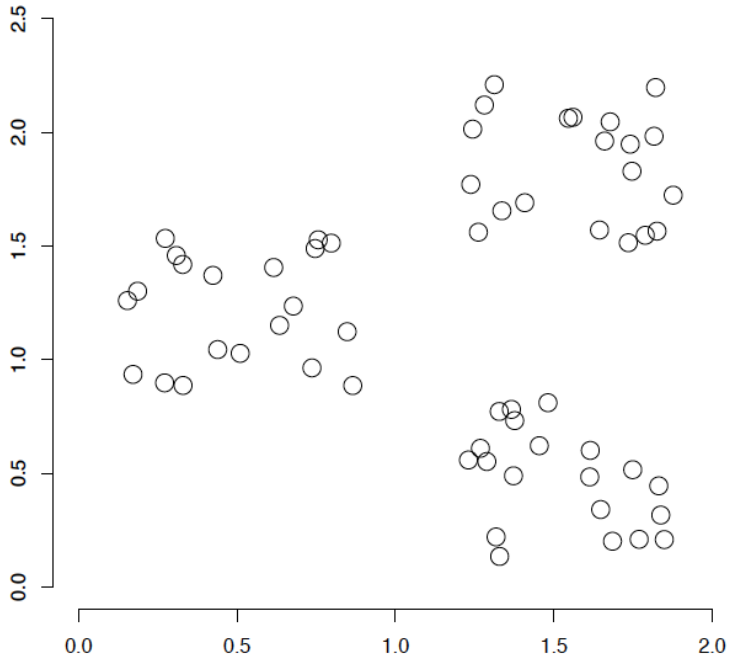
What is clustering?

- 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.
  - Classes are not pre-defined

- The commonest form of unsupervised learning
  - Unsupervised learning = learning from data without labels
  - Supervised learning: learning from data with annotations
  - A common and important task that finds many applications in document man agent

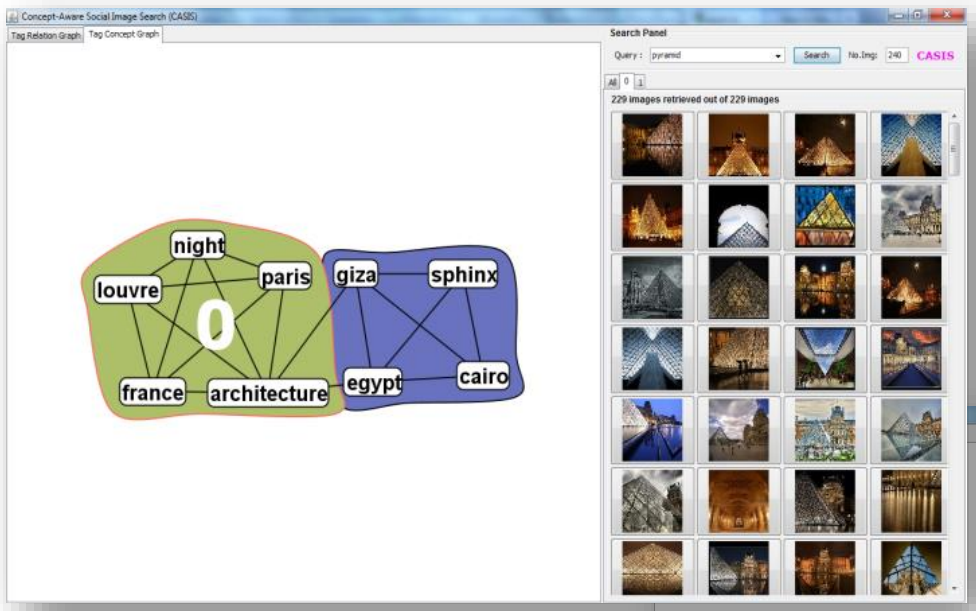
A data set with clear cluster structure

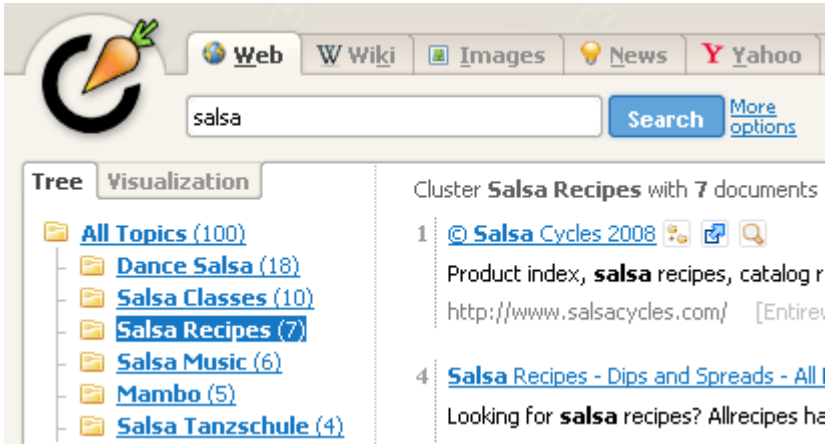
- How would you design an algorithm for finding the three clusters in this case?

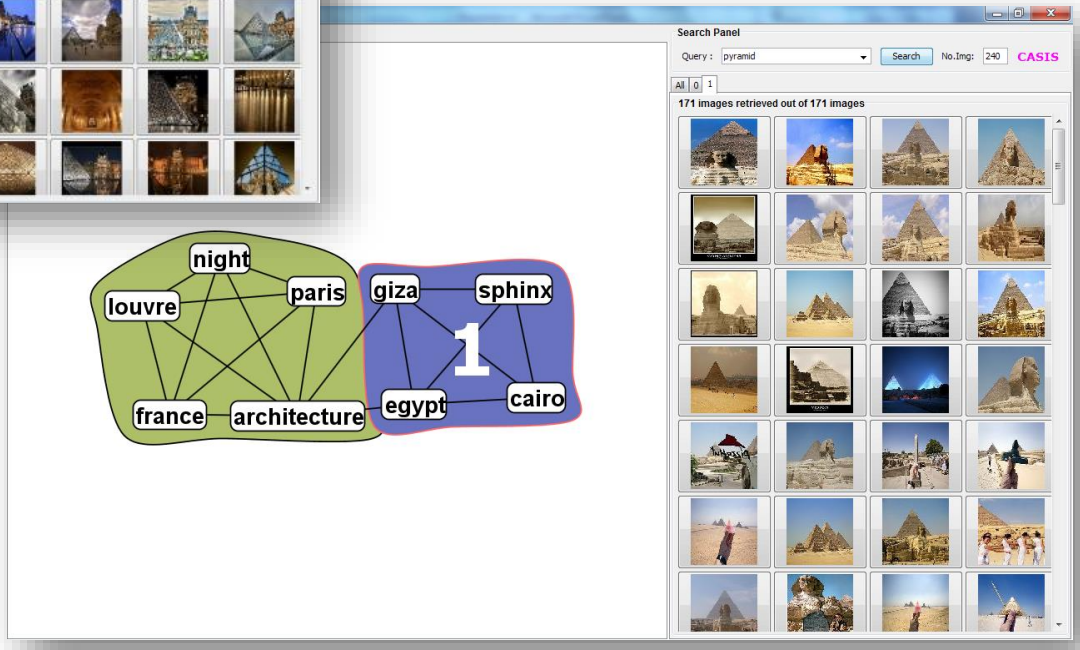


- Unsupervised learning
  - Clusters are inferred from the data without human input.
  - However, there are many ways of influencing the outcome of clustering: number of clusters, similarity measure, representation of documents, . . .

Some applications of clustering



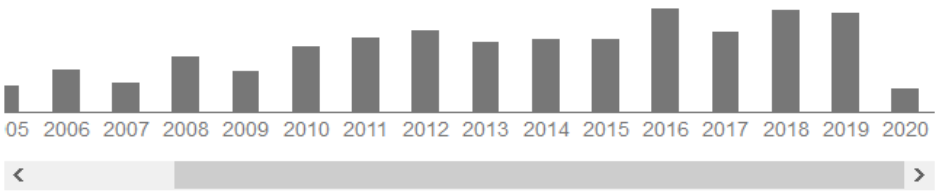




It might be hard to find perfect clustering

An impossibility theorem for clustering

Authors	Jon M Kleinberg
Publication date	2003
Conference	Advances in neural information processing systems
Pages	463-470
Description	Although the study of clustering is centered around an intuitively compelling goal, it has been very difficult to develop a unified framework for reasoning about it at a technical level, and profoundly diverse approaches to clustering abound in the research community. Here we suggest a formal perspective on the difficulty in finding such a unification, in the form of an impossibility theorem: for a set of three simple properties, we show that there is no clustering function satisfying all three. Relaxations of these properties expose some of the interesting (and unavoidable) trade-offs at work in well-studied clustering techniques such as single-linkage, sum-of-pairs, k-means, and k-median.
Total citations	Cited by 646



Scholar articles
An impossibility theorem for clustering
JM Kleinberg - Advances in neural information processing systems, 2003
Cited by 646 Related articles All 33 versions

http://papers.nips.cc/paper/2340-an-impossibility-theorem-for-clustering.pdf

Notion of similarity/distance

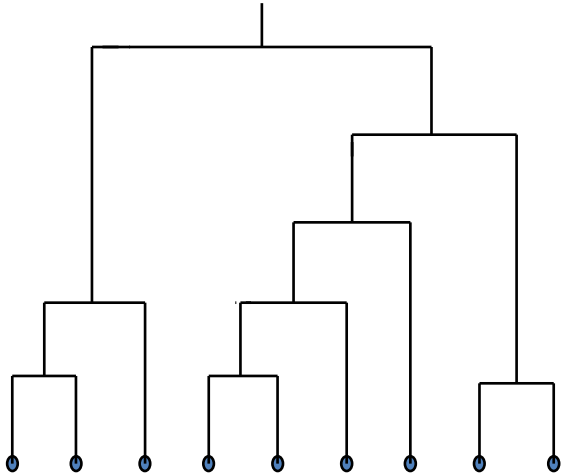
- Distance or Similarity
  - Ideal: semantic similarity.
  - Practical: term-statistical similarity

- Documents as vectors.
  - For many algorithms, easier to think in terms of a distance (rather than similarity) between docs.
  - But real implementations use cosine similarity

Clustering Algorithms

- Flat algorithms: K-means clustering
  - Input: a set of documents and the number K
  - Output: a partition of K flat clusters

- Hierarchical algorithms
  - Input: a set of documents
  - Output: a (binary) hierarchy
  - Approach
    - Bottom-up, agglomerative
    - (Top-down, divisive)



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

K-Means

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

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

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

K-Means Algorithm

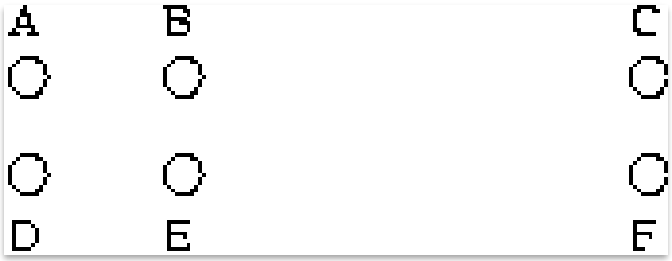
- Select k random documents $\{ s_1, s_2, \ldots s_k \}$ as seeds

- Until clustering converges (or other stopping criterion)
  - For each document d_i
    - Assign d_i to the cluster c_j such that $dist(d_i, s_j)$ is minimal
  - For each cluster c_j, $s_j = \mu(c_j)$ //update the seeds to the centroid of each cluster

- Several possibilities for termination
  - A fixed number of iterations.
  - Document partition unchanged.
  - Centroid positions don't change
    - Does this mean that the documents in a cluster are unchanged?

Issues with k-Means

- Seed Choice
  - Results can vary based on random seed selection.
  - Try with B and E
  - Try with D and F

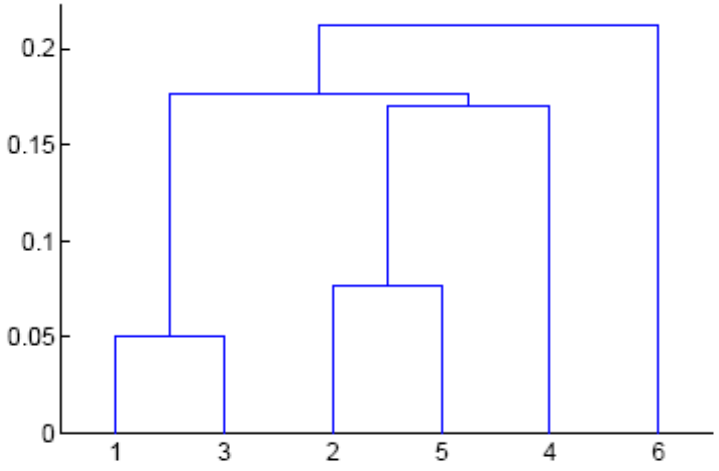


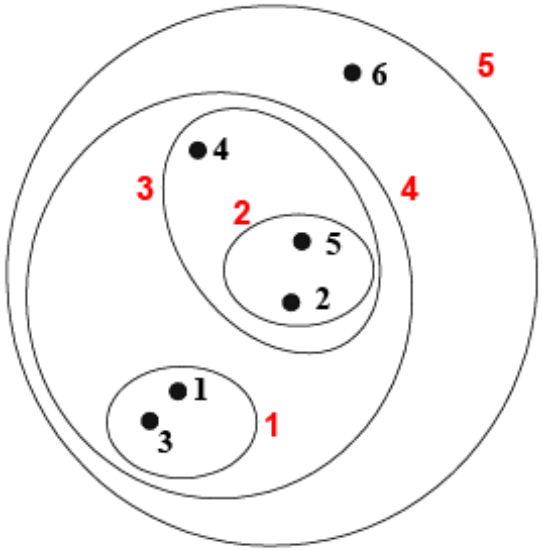
- Select good seeds using a heuristic
  - Document least similar to any existing mean
  - Try out multiple starting points
  - Initialize with the results of another method

- Number of clusters *K*
  - Fixed a priori?
  - Finding the “right” number of clusters is part of the problem

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.
  - A tree like diagram that records the sequences of merges or splits
  - One approach: recursive application of a partitional clustering algorithm.

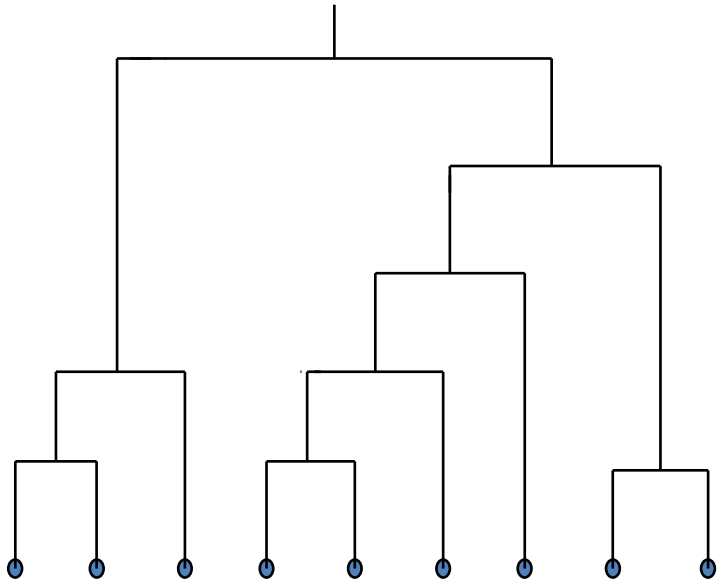




Strengths of Hierarchical Clustering

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

- They may correspond to meaningful taxonomies
  - E.g. Yahoo ! Directory



Two Types of Hierarchical Clustering

- Agglomerative: More popular
  - Start with the points as individual clusters, e.g., each document is one cluster
  - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive:
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

HAC: Algorithm

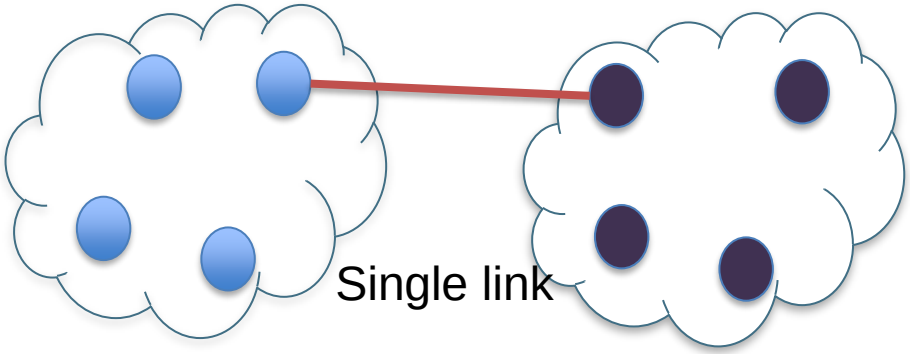
- Major steps:
  - Compute the proximity matrix
  - Let each data point be a cluster
  - Repeat
    - Merge the two closest clusters
    - Update the proximity matrix
  - Until only a single cluster remains

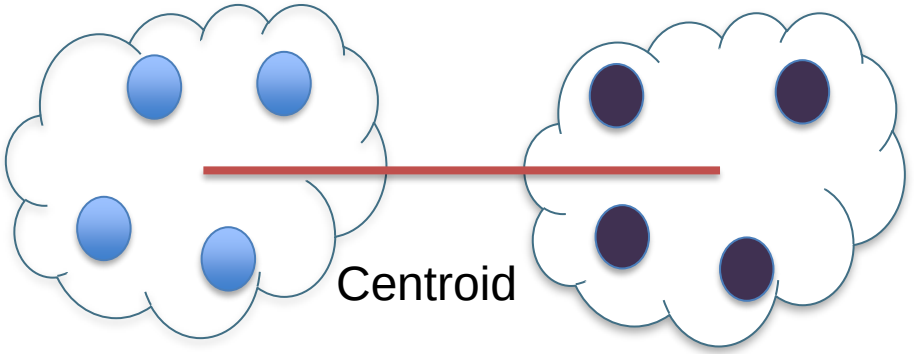
- Key operation is the computation of the **proximity of two clusters**:
  - Different approaches to defining the distance between clusters distinguish the different algorithms

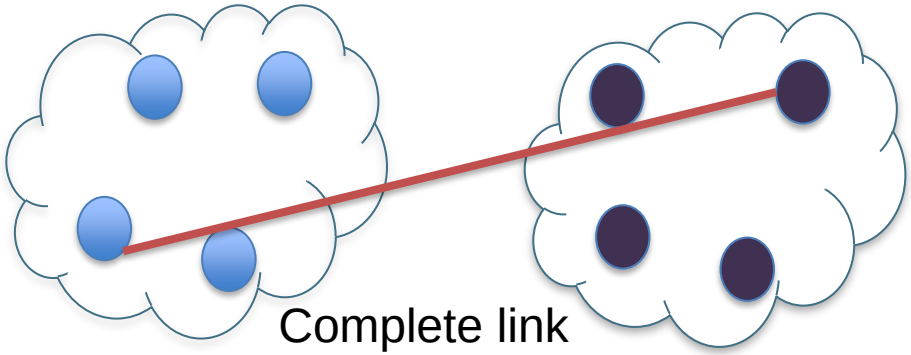
Closest pair of clusters (Many variants)

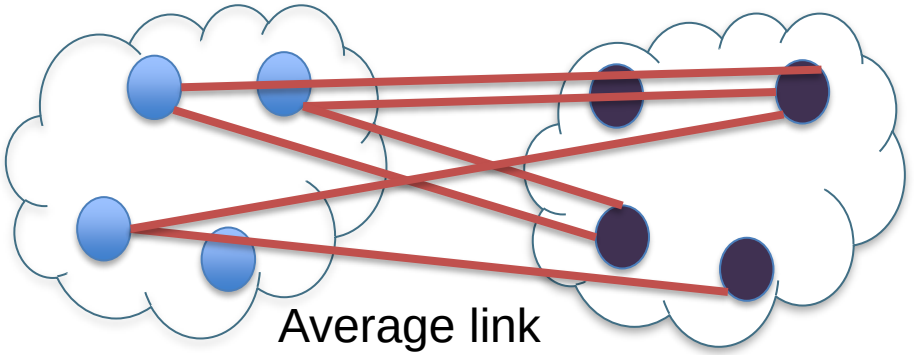
- Many variants to defining closest pair of clusters
- Single-link
  - Similarity of the most cosine-similar (single-link)
- Complete-link
  - Similarity of the “furthest” points, the least cosine-similar
- Centroid
  - Clusters whose centroids are the most cosine-similar
- Average-link
  - Average cosine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

Cluster Similarity



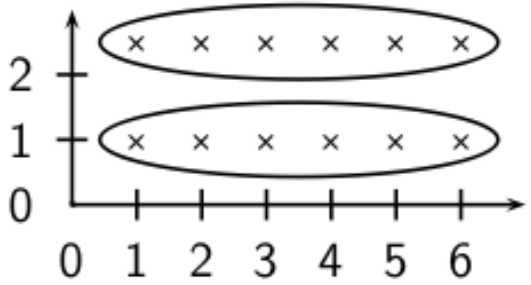




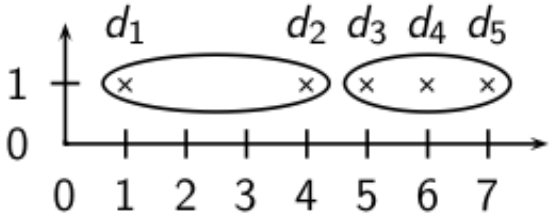


Potential issues

- Single-link clustering often produces long, straggly clusters.
  - For most applications, these are undesirable.

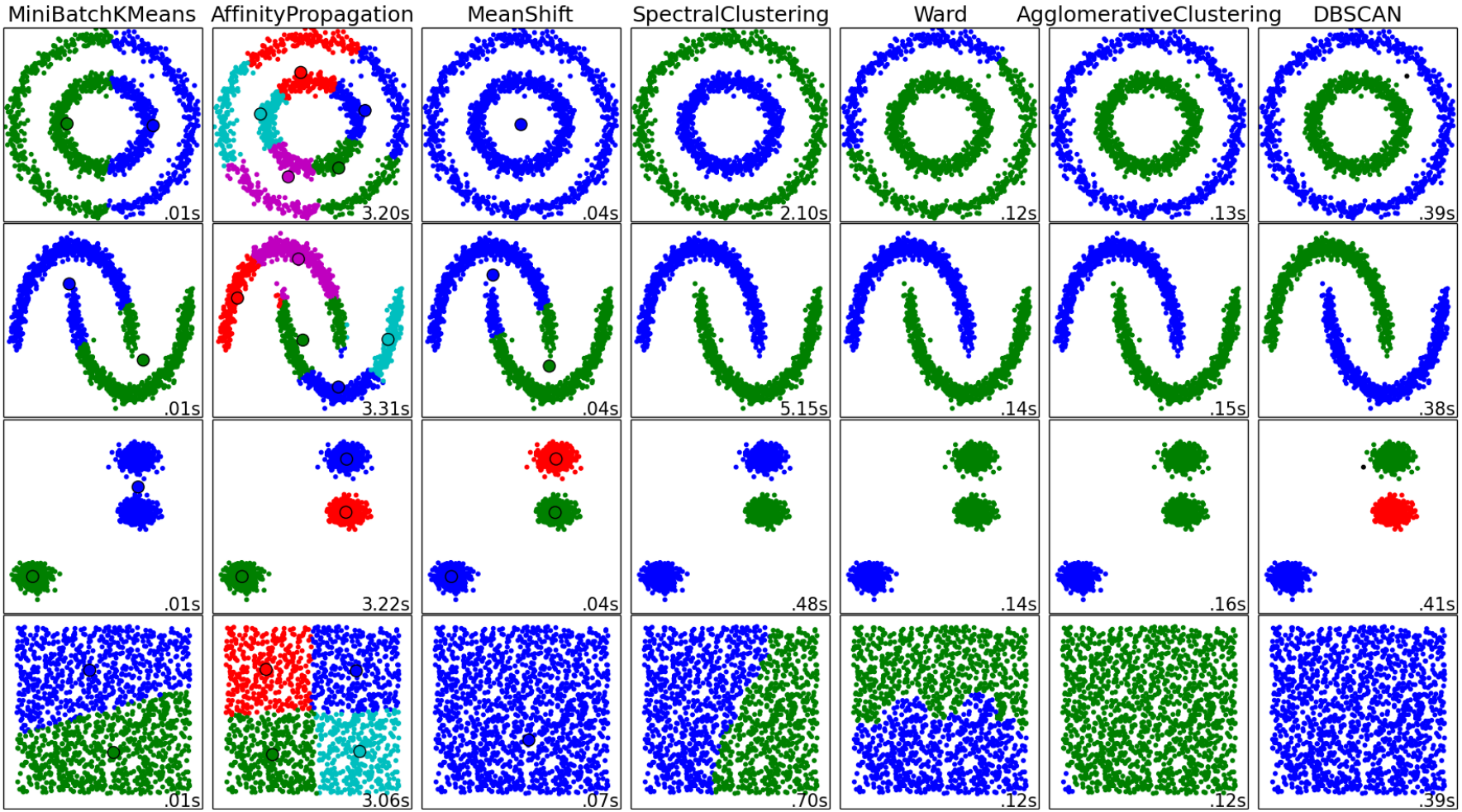


- Complete-link clustering is sensitive to outliers
  - In this example, d_2 is splited from its right neighbors because $sim(d_1, d_2)$ is slightly larger than $sim(d_2, d_5)$
  - Clearly this is undesirable.
  - The reason is the outlier d_1



19

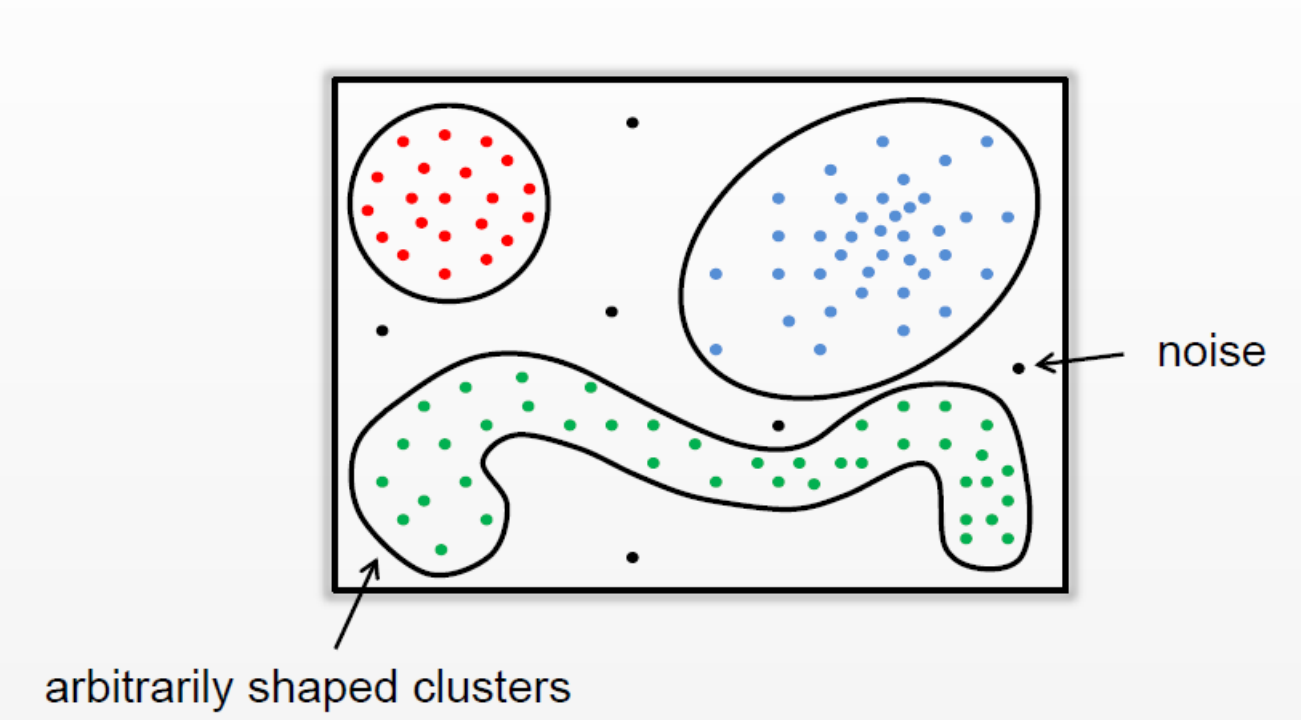
There are many different kinds of clustering



Source: https://scikit-learn.org/0.15/auto_examples/cluster/plot_cluster_comparison.html

DBSCAN

- Density-based spatial clustering of applications with noise (DBSCAN)
- Discovers clusters of arbitrary shape in spatial databases with noise
  - A cluster is a region of *high density*, *Noise* points lie in regions of *low* density



Adopted from Jan-Willem van de Meent's lecture

DBSCAN

- DBSCAN is a **density-based** algorithm.
  - Density = number of points within a specified radius (Eps)

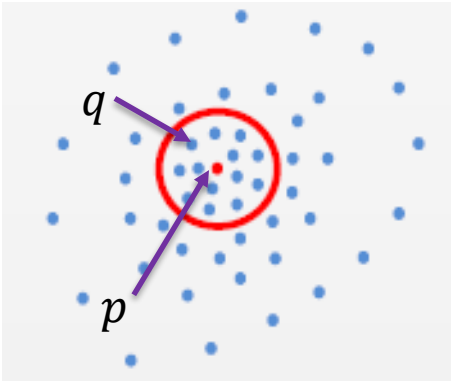
- Three kinds of points:
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

Eps-neighbourhood of a point p

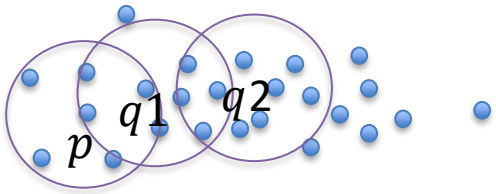
$$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$$

Density-reachability

- Directly density-reachable
  - A point q is directly density-reachable from point p if
    - p is a core point, and
    - q is in p's Eps-neighbourhood
  - Density-reachability is asymmetric



- Density-reachable (directly and indirectly):
  - A point $q1$ is directly density-reachable from p
  - $q2$ is directly density-reachable from $q1$
  - Then p → $q1$ → $q2$ form a chain
  - Point $q2$ is indirectly reachable from p



- All points not reachable from any other point are *outliers* or *noise points*.

DBSCAN

- Start with an arbitrary point p from the data and retrieve all points density-reachable from p with regard to Eps and MinPts.
  - If p is a core point, the procedure yields a cluster with regard to Eps and MinPts, and the point is classified.
  - If p is a border point, no points are density-reachable from p and DBSCAN visits the next unclassified point
- DBSCAN visits each point of the data, possibly multiple times
- DBSCAN is not entirely deterministic
  - border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data are processed.

An online demo: https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

DBSCAN

- Pros
  - Resistant to Noise
  - Can handle clusters of different shapes and sizes
- Cons
  - Cannot handle varying densities
  - Sensitive to parameters, hard to determine the correct set of parameters

A density-based algorithm for discovering clusters in large spatial databases with noise.

Authors	Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu
Publication date	1996/8/2
Journal	Kdd
Volume	96
Issue	34
Pages	226-231
Description	Clustering algorithms are attractive for the task of class identification in spatial databases. However, the application to large spatial databases rises the following requirements for clustering algorithms: minimal requirements of domain knowledge to determine the input parameters, discovery of clusters with arbitrary shape and good efficiency on large databases. The well-known clustering algorithms offer no solution to the combination of these requirements. In this paper, we present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. We performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data of the SEQUOIA 2000 benchmark. The results of our experiments demonstrate that (1) DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS, and that (2) DBSCAN outperforms CLARANS by factor of more than 100 in terms of efficiency.
Total citations	Cited by 17118

Clustering evaluation

- 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

External Evaluation of Cluster Quality

- Assume documents with C gold standard classes,
  - Labeled classes $C = \{c_1, c_2, \ldots, c_j\}$
  - Our clustering algorithms produce K clusters, $\Omega = \{\omega 1, \omega 2, \ldots, \omega_k\}$

- Simple measure: purity

$$Purity\ (\Omega, C) = \frac{1}{N} \sum_{k} \max_{j} |w_k \cap c_j|$$

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

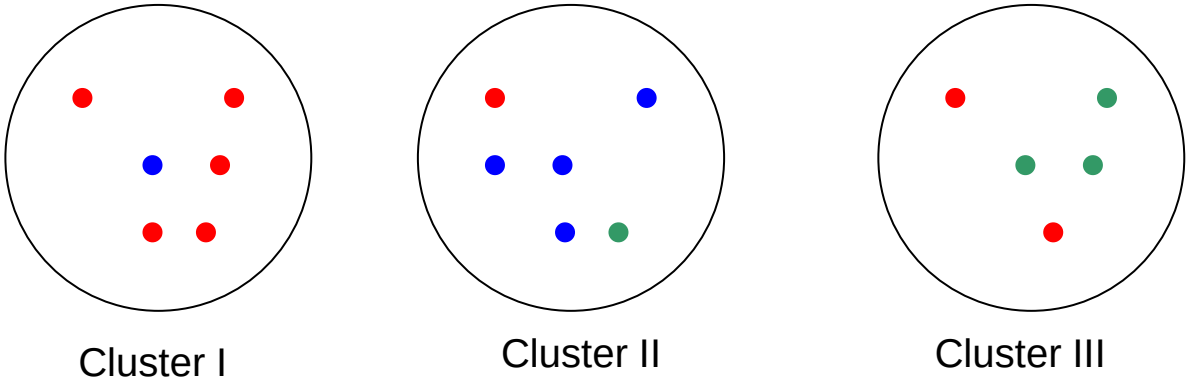
Another measure: Rand Index

- There are total N(N-1)/2 possible pairs in a document collection of N documents
- A true-positive decision (TP) assigns two similar documents in the same cluster, similarly we can define FP, TN, FN

$$\text{Rand Index: } RI = \frac{TP+TN}{TP+FP+FN+TN}$$

Number of pairs	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	20	24
Different classes in ground truth	20	72

Purity and Rand Index (Example)



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

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

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

Purity = (5+4+3)/17 = 0.71

Rand Index:

$$TP+FP=\binom{6}{2}+\binom{6}{2}+\binom{5}{2}=40$$

$$TP=\binom{5}{2}+\binom{4}{2}+\binom{3}{2}+\binom{2}{2}=20$$

$$FP=40-20=20$$

$$FN=5\times1+5\times2+1\times2+1\times4+1\times3=24$$

$$TN=\binom{17}{2}-TP-FP-FN=72$$

Summary

- Document clustering
  - Document representations
  - Distance/Similarity
- Clustering algorithms
  - Partitional
  - Hierarchical
  - DBSCAN
- Clustering Evaluation
  - Purity
  - Rand Index