POLITECNICO DI MILANO

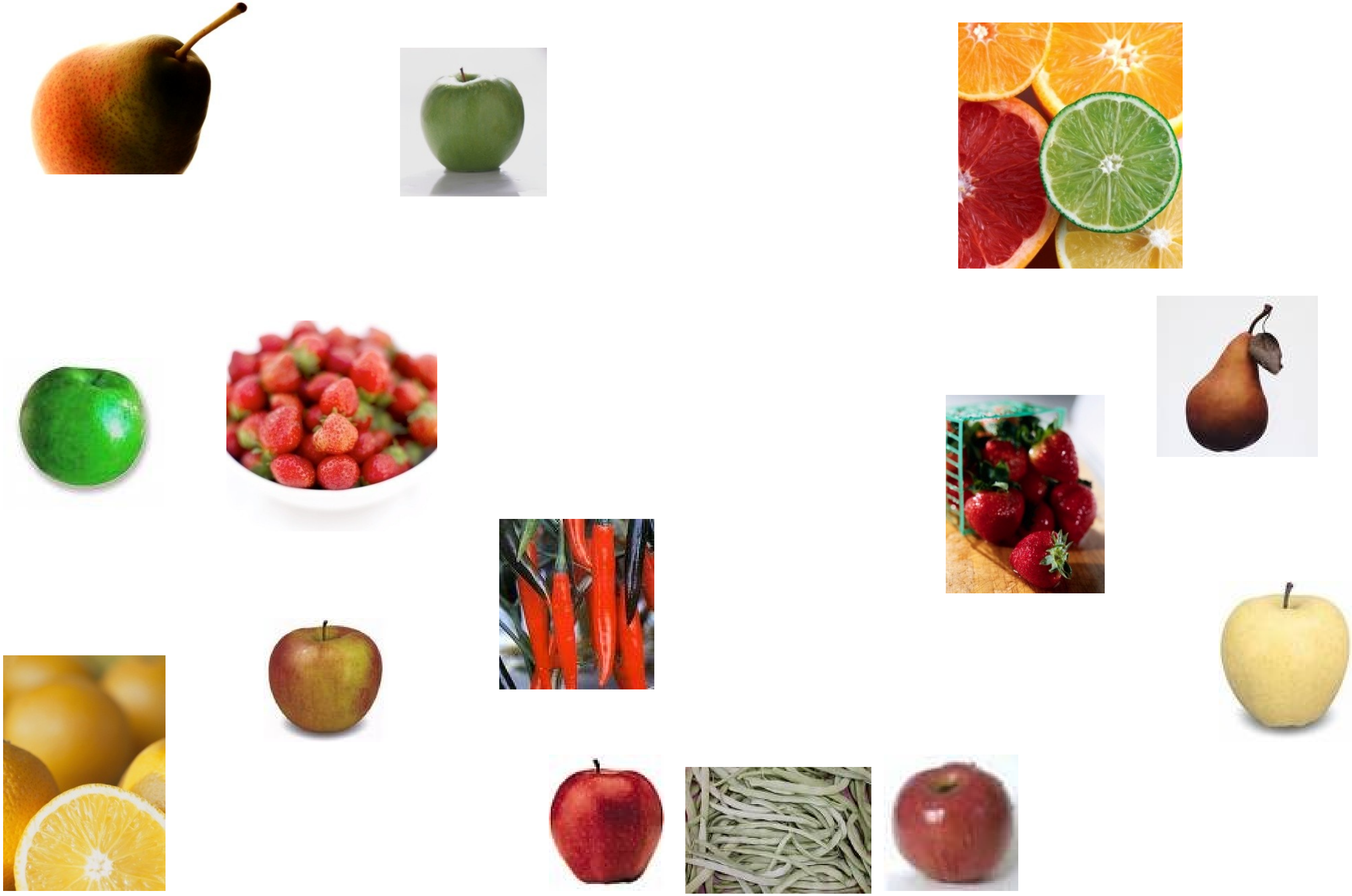
Clustering: Hierarchical

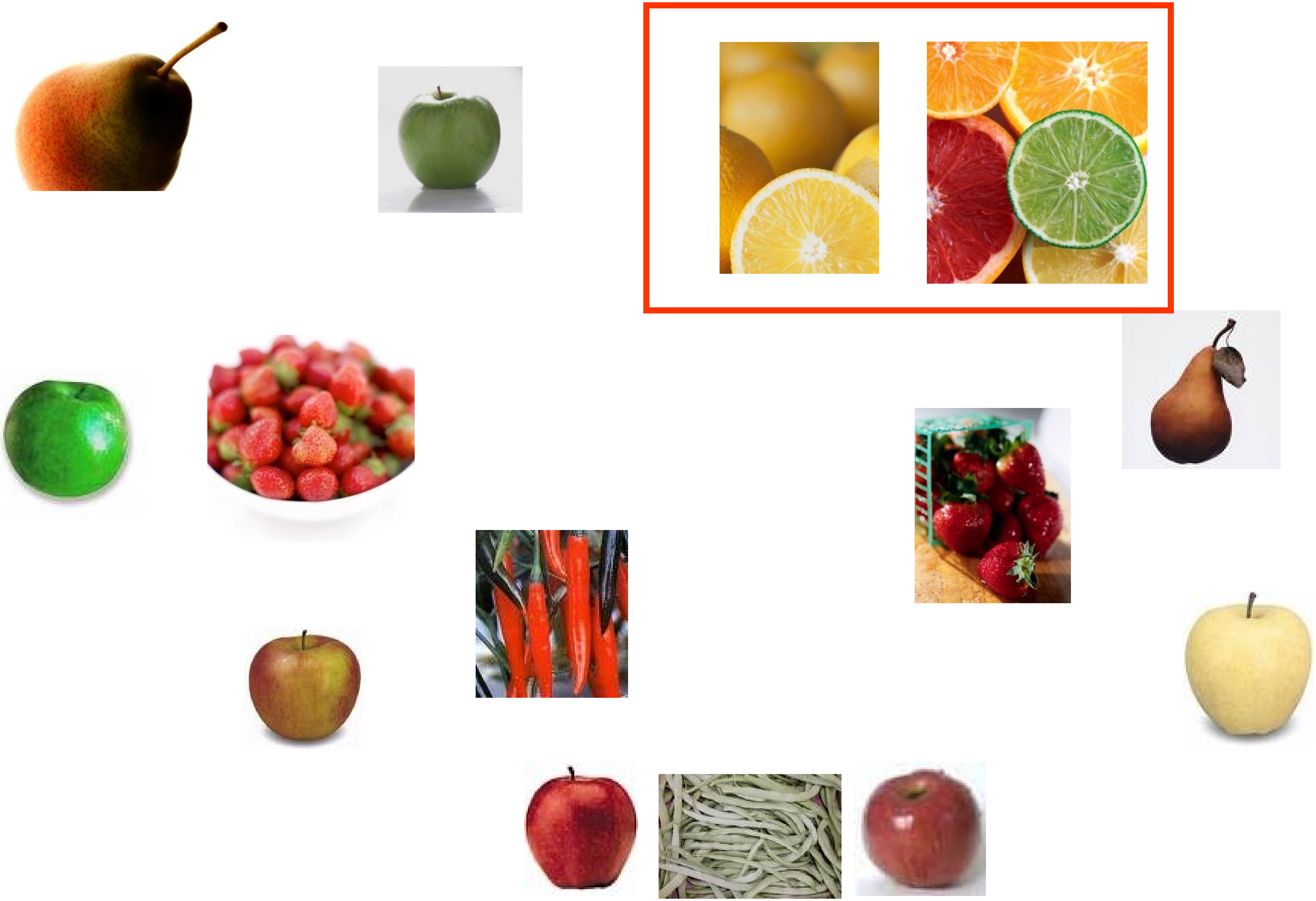
Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

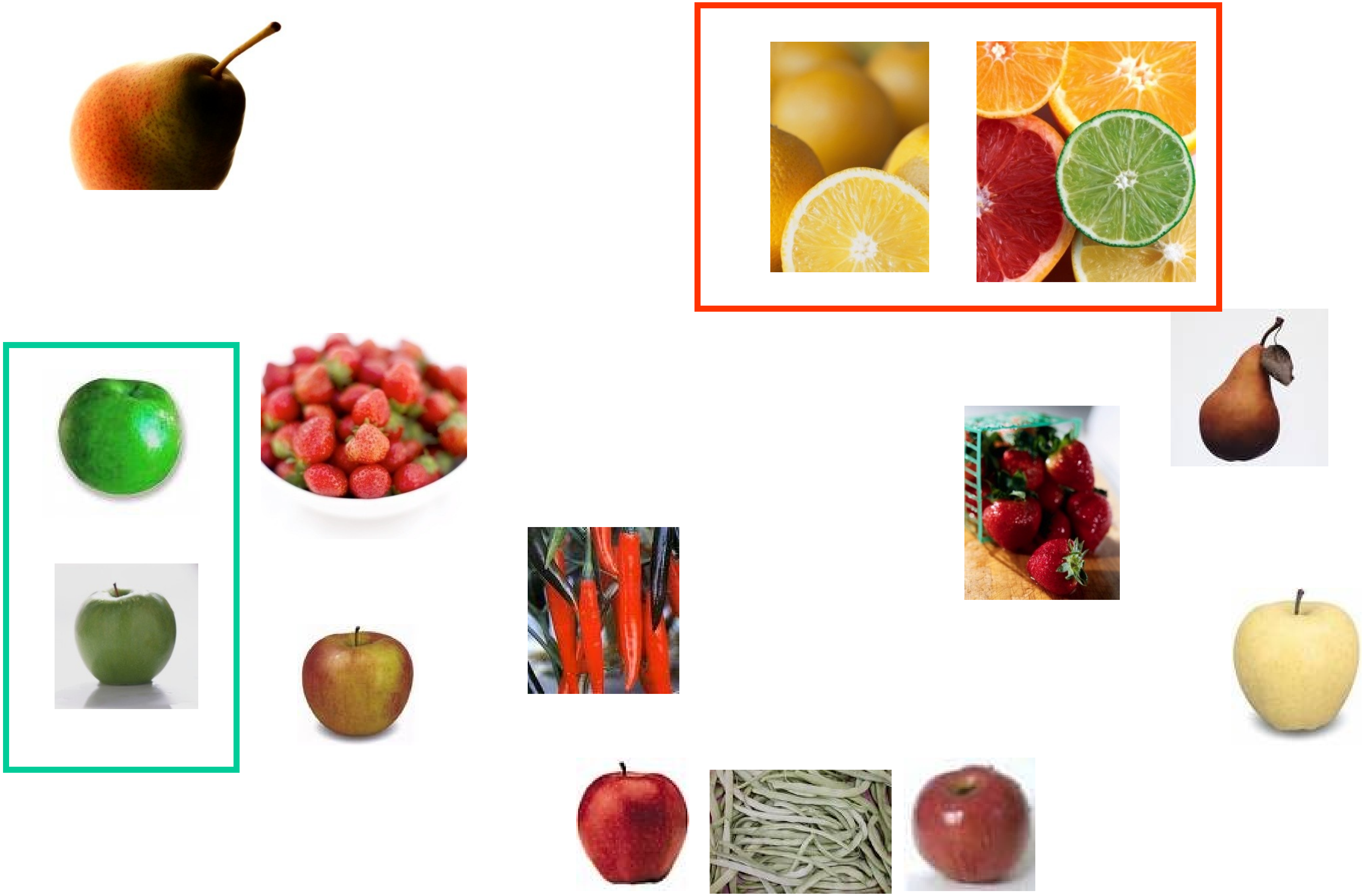
Lecture outline

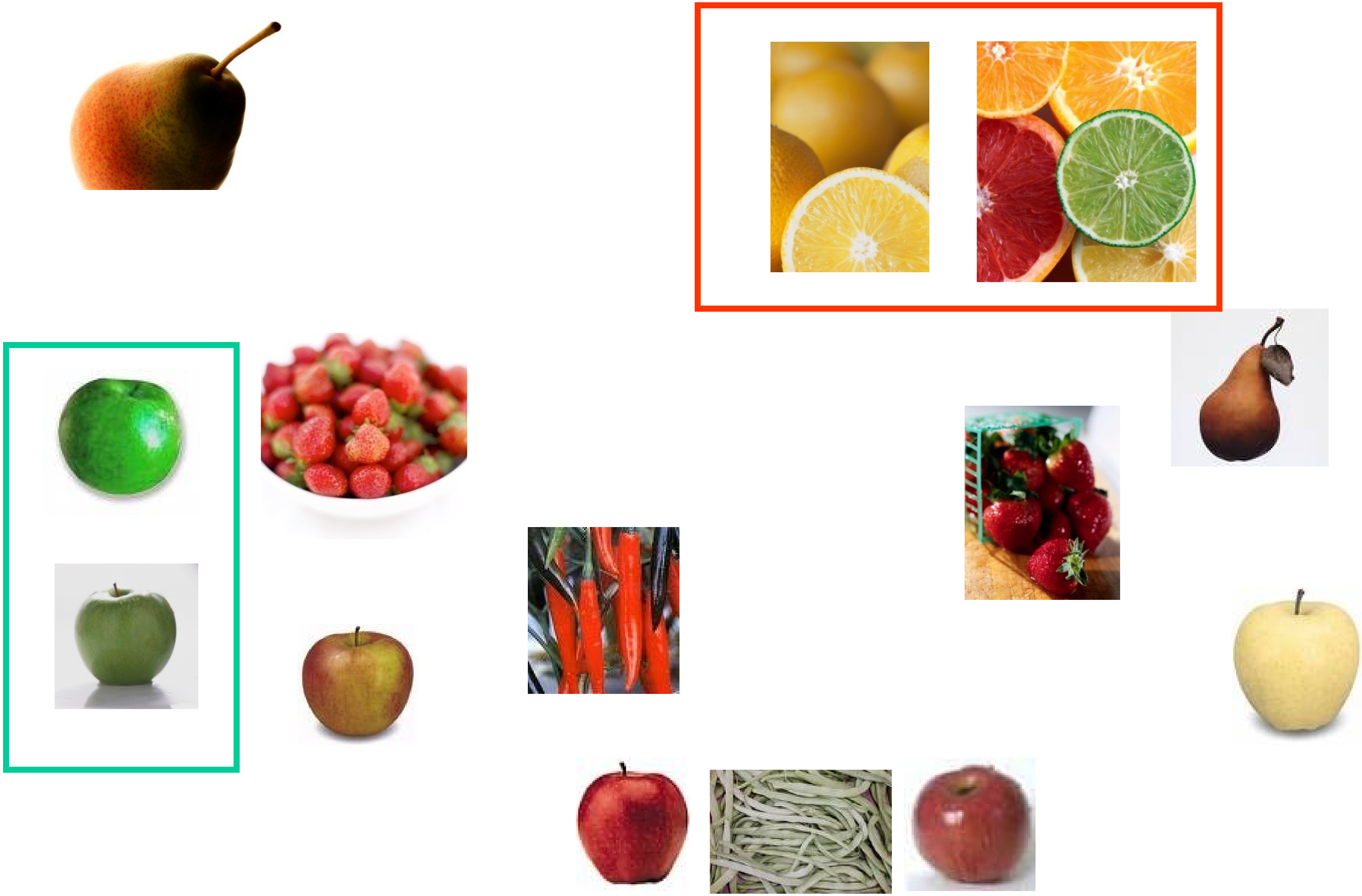
- Building hierarchies
- What is hierarchical clustering?
- Agglomerative clustering
- Issues and limitations
- Algorithms for hierarchical clustering

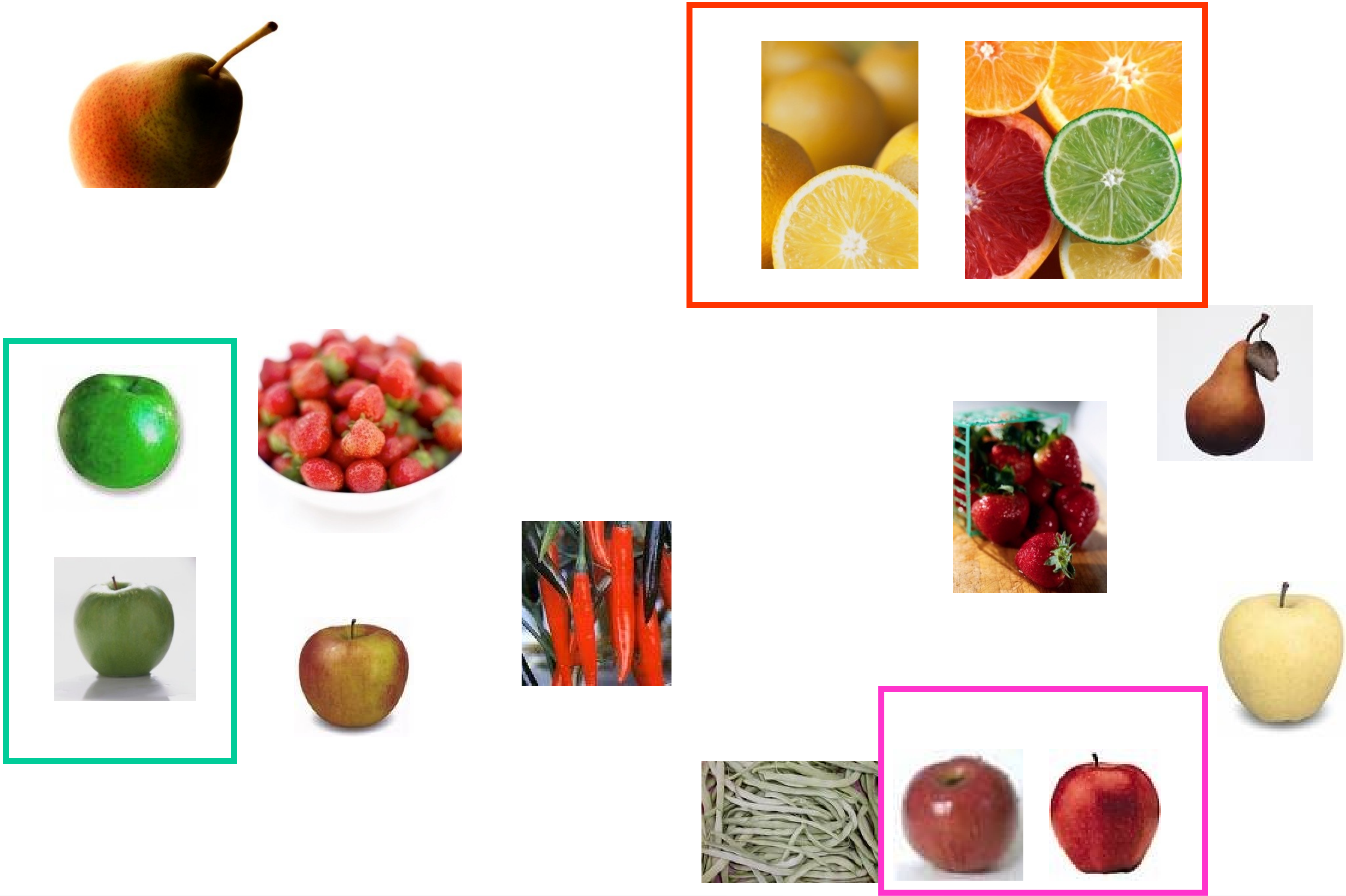
Building hierarchies…

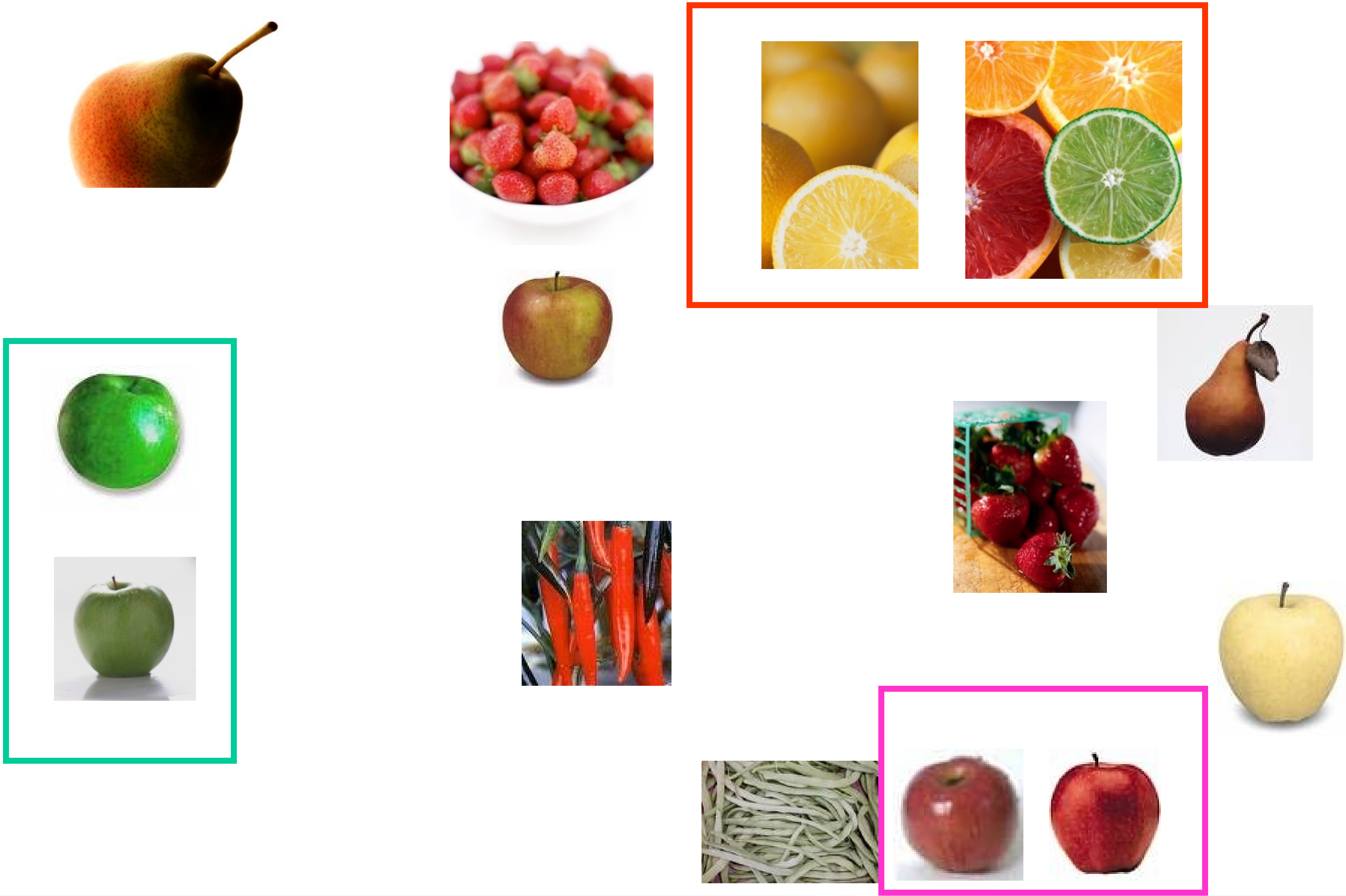


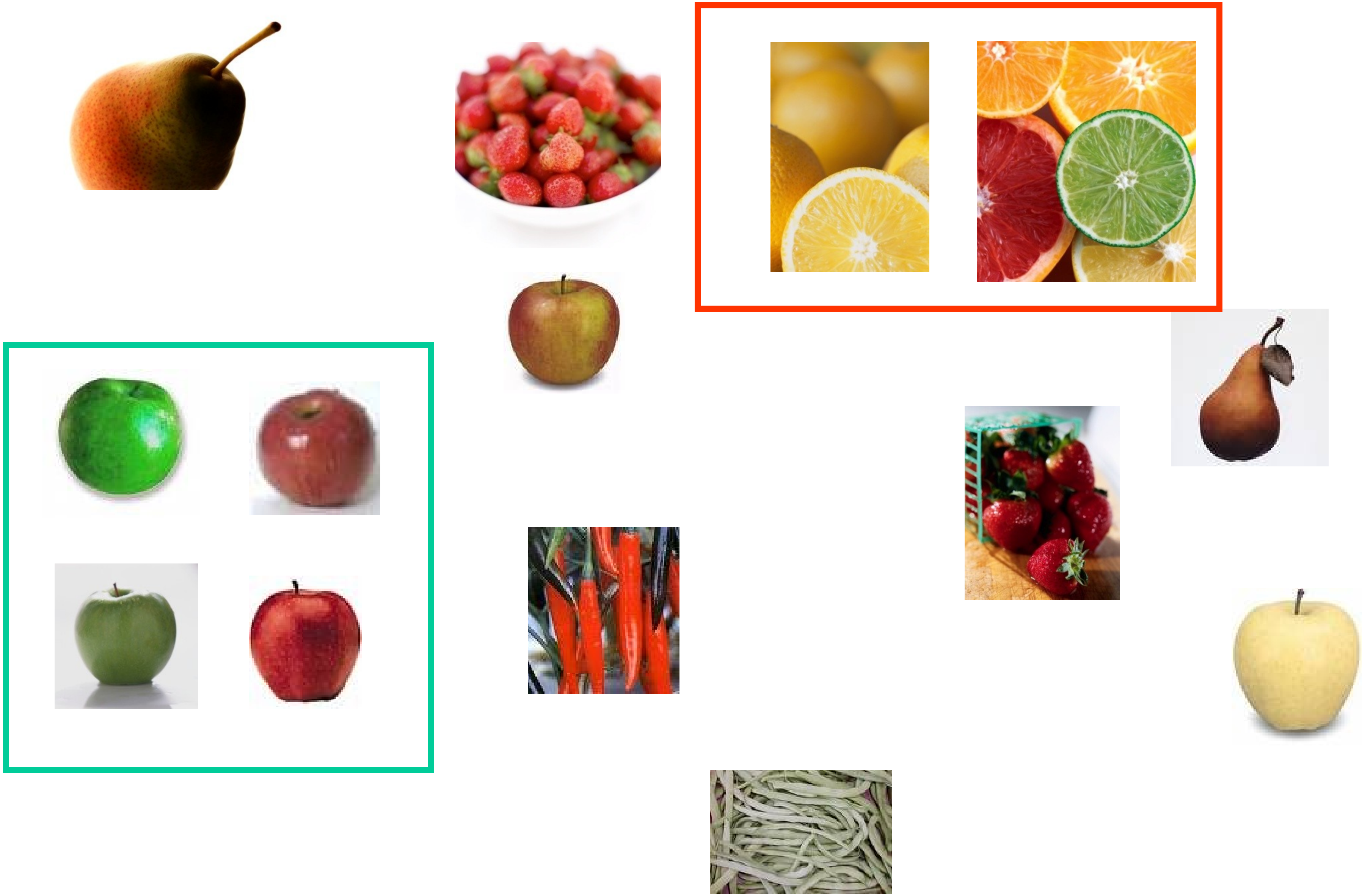


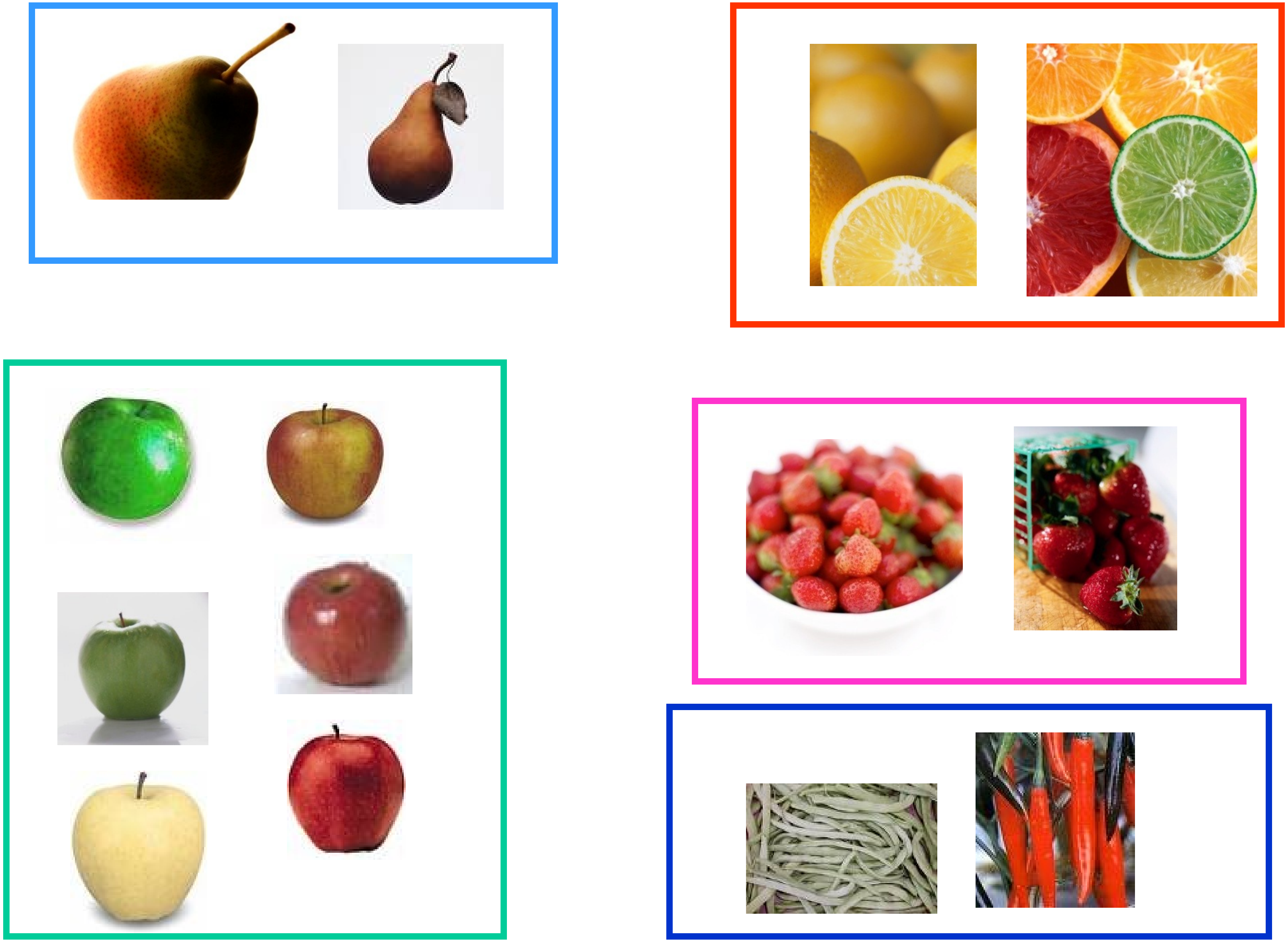








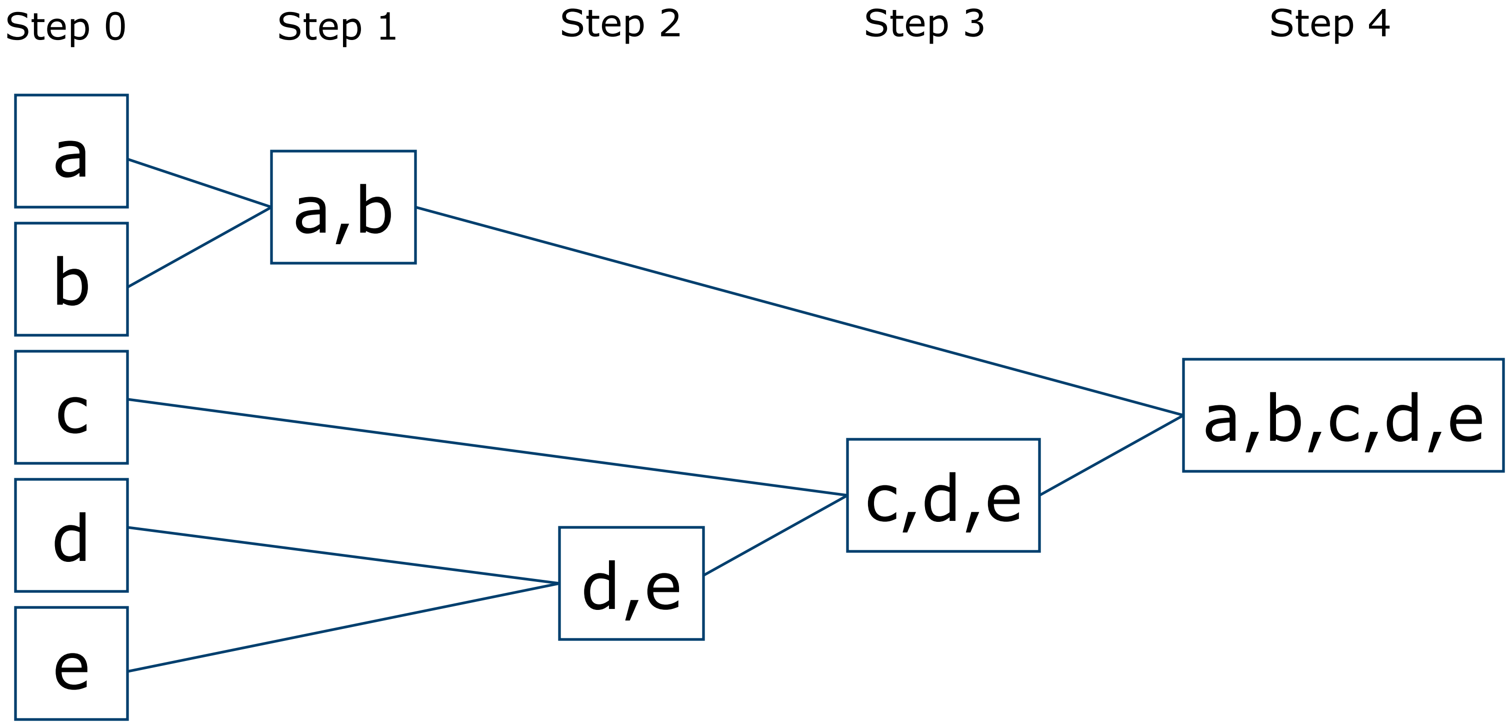




Hierarchical clustering…

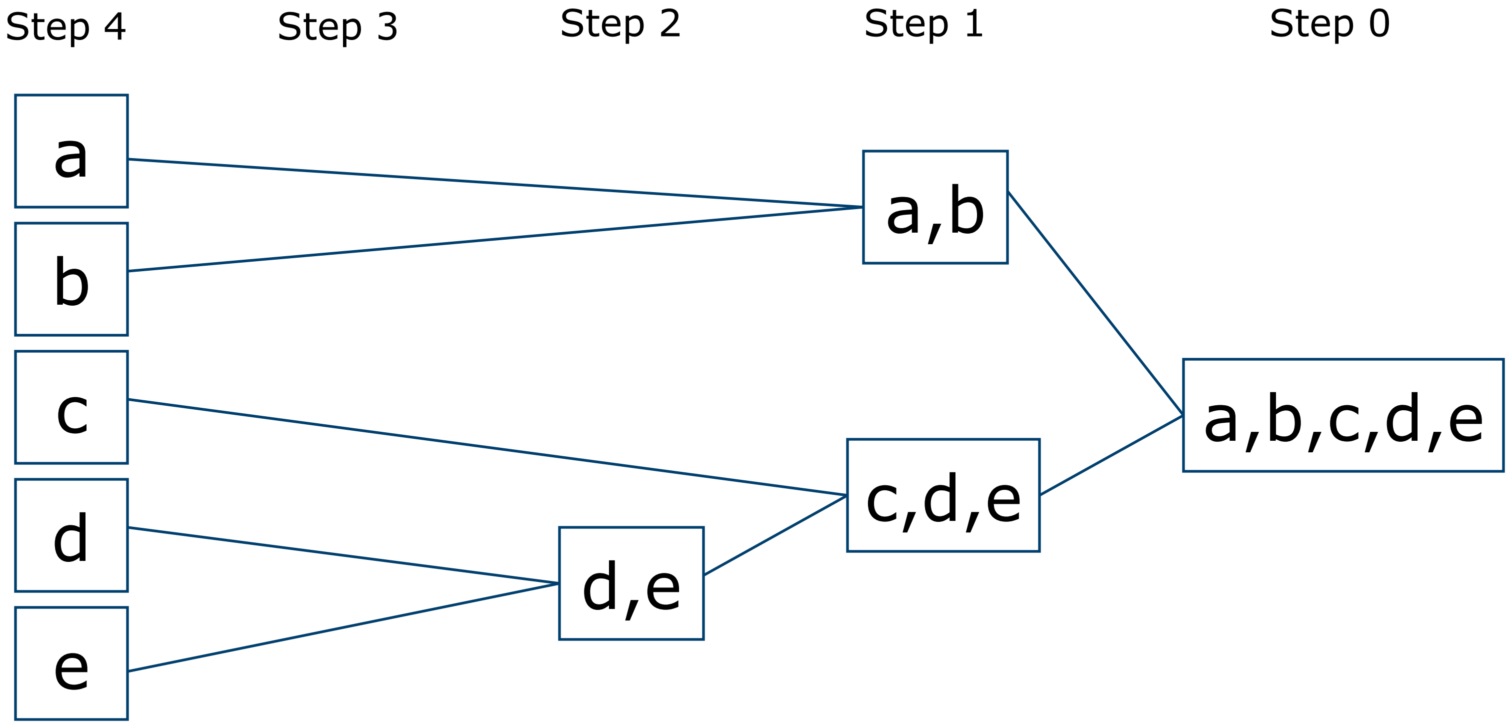
What is hierarchical clustering?

- Suppose we have five items, a, b, c, d, and e.
- Initially, we consider one cluster for each item
- Then, at each step we merge together the most similar clusters, until we generate one cluster



What is hierarchical clustering?

- Alternatively, we start from one cluster containing the five elements
- Then, at each step we split one cluster to improve intracluster similarity, until all the elements are contained in one cluster



Step 4 Step 3 Step 2 Step 1 Step 0

a

b

c

d

e

d,e

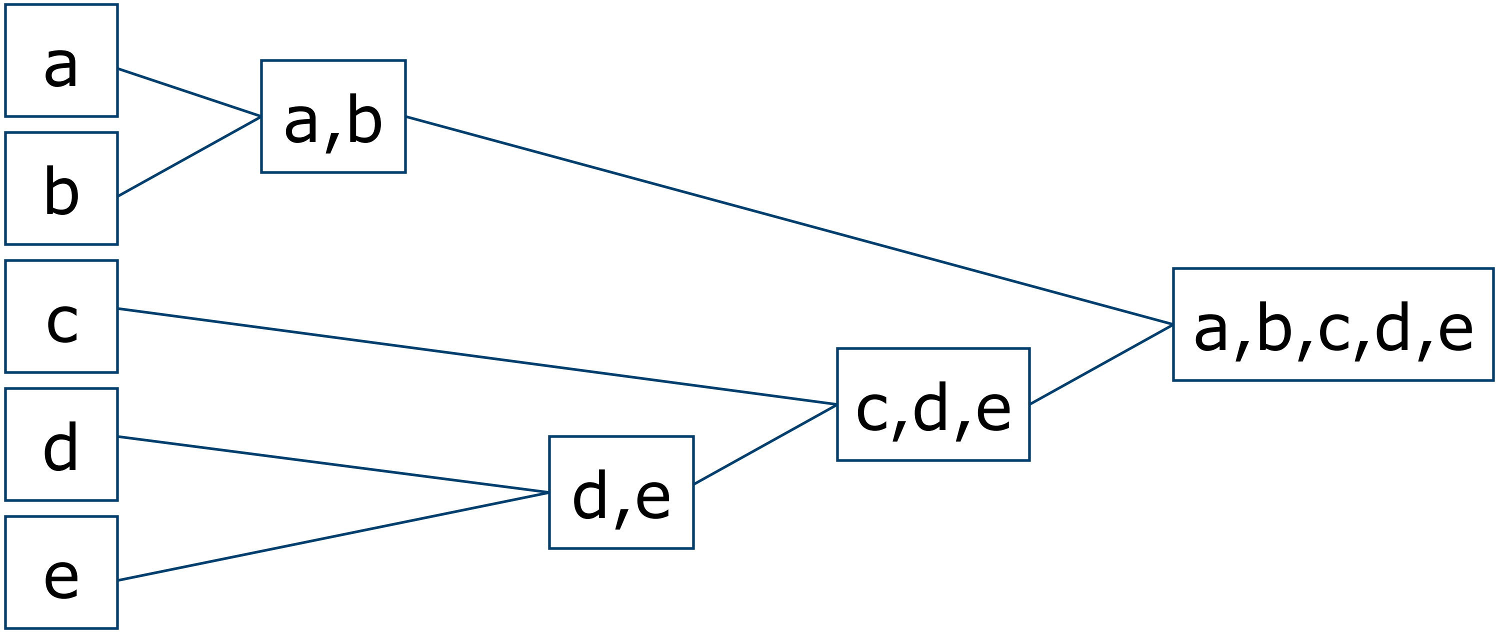
a,b

c,d,e

a,b,c,d,e

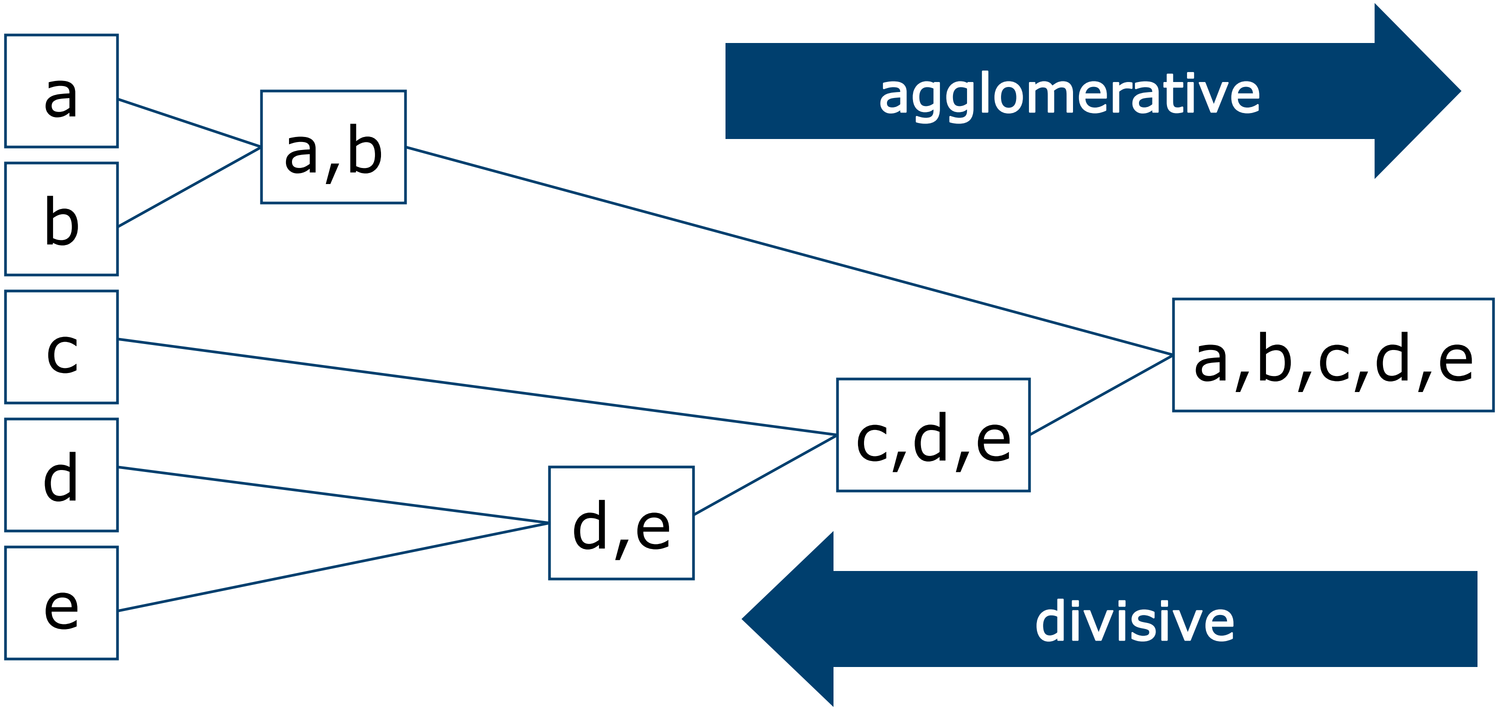
What is hierarchical clustering?

- By far, it is the most common clustering technique
- Produces a hierarchy of nested clusters
- The hiearchy be visualized as a dendrogram: a tree like diagram that records the sequences of merges or splits



What are the approaches?

- Agglomerative
  - Start individual clusters, at each step, merge the closest pair of clusters until only one cluster (or k clusters) left
- Divisive
  - Start with one cluster, at each step, split a cluster until each cluster contains a point (or there are k clusters)



Strengths of Hierarchical Clustering

- No need 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, …)

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

Agglomerative clustering…

Agglomerative Clustering Algorithm

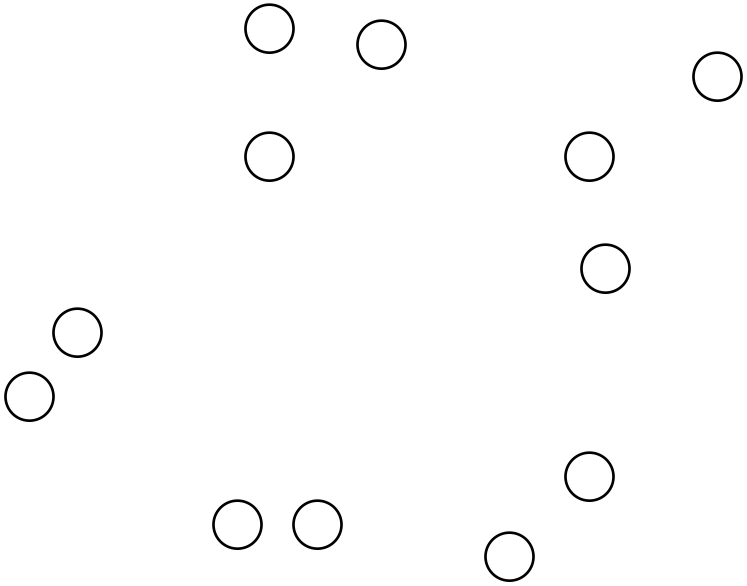
- More popular hierarchical clustering technique

- Basic algorithm
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. Repeat
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. 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

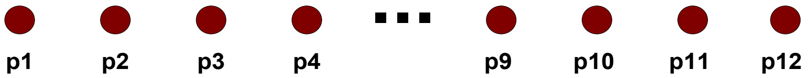
Starting Situation

- Start with clusters of individual points and a proximity matrix



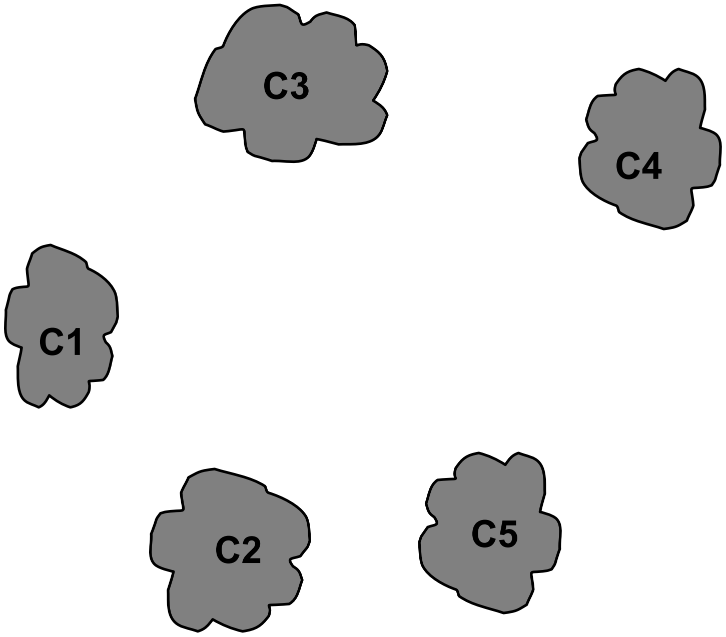
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
⋮						

Proximity Matrix



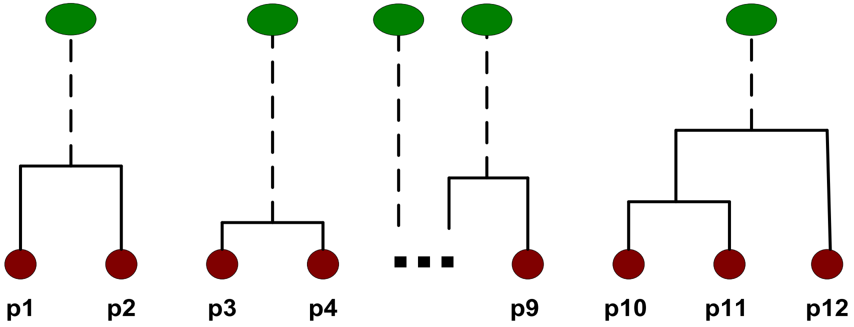
Intermediate Situation

- After some merging steps, we have some clusters



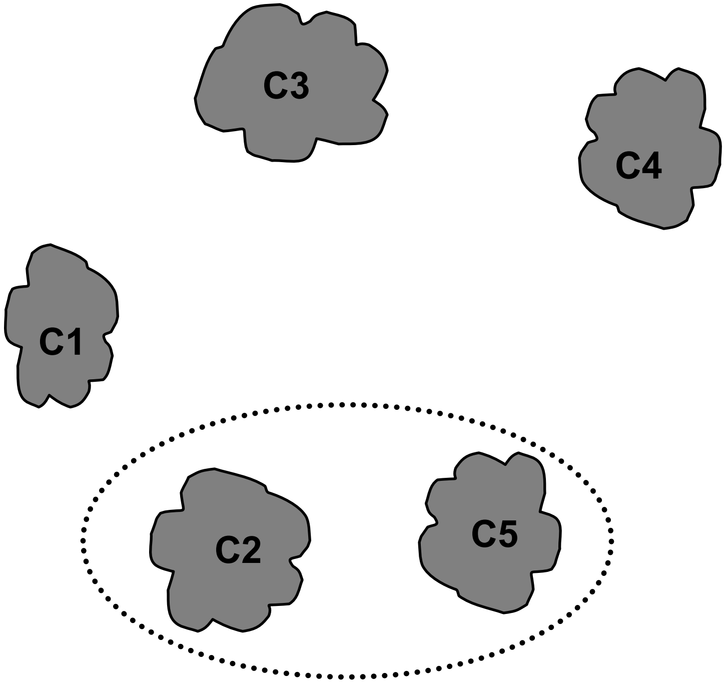
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

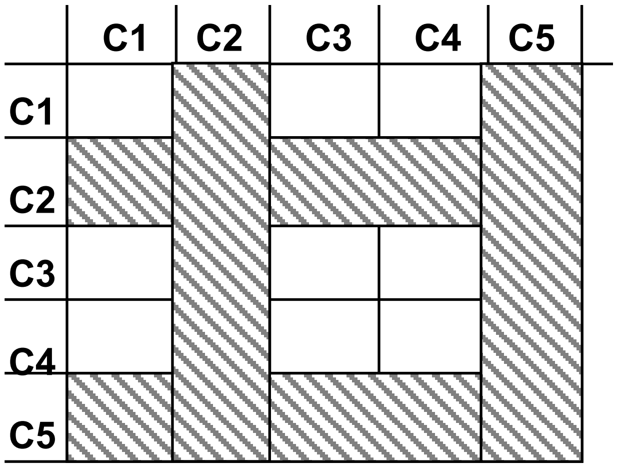
Proximity Matrix



Intermediate Situation

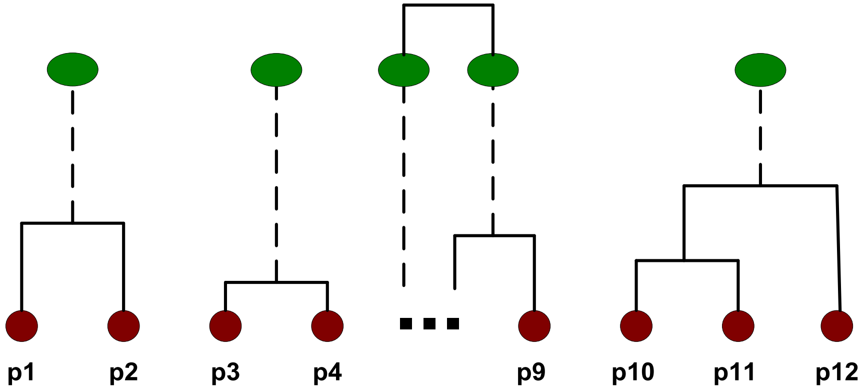
- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.





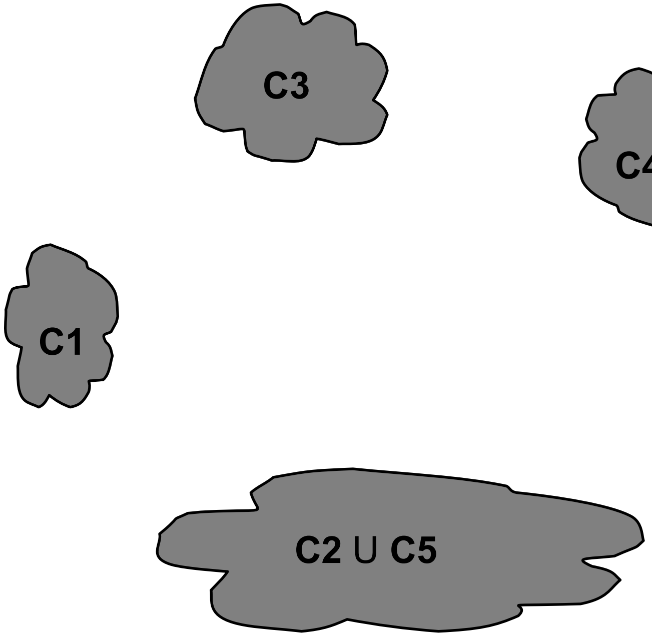
	C1	C2	C3	C4	C5
C1		▨			▨
C2	▨	▨	▨	▨	▨
C3		▨			▨
C4		▨			▨
C5	▨	▨	▨	▨	▨

Proximity Matrix



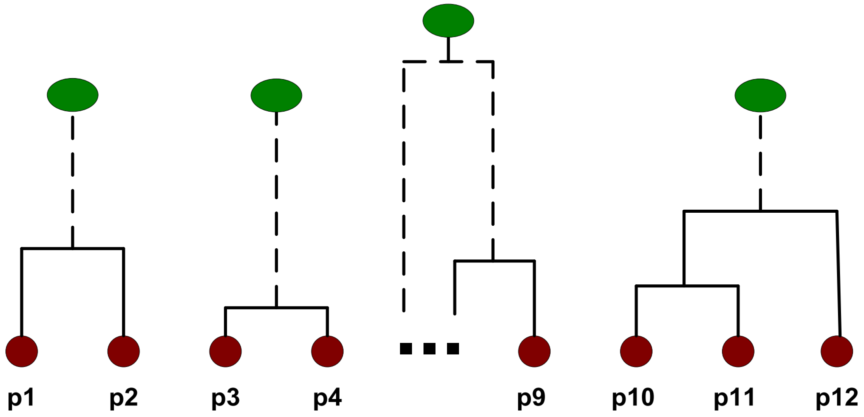
After Merging

- The question is "How do we update the proximity matrix?"



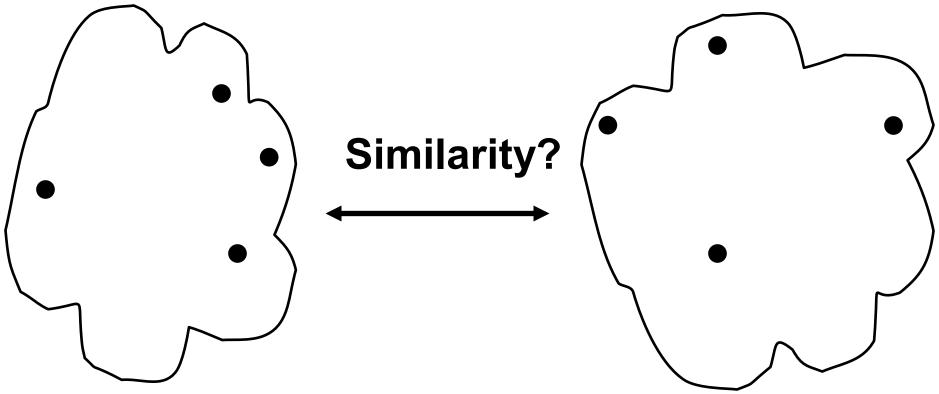
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



Similarity between clusters…

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

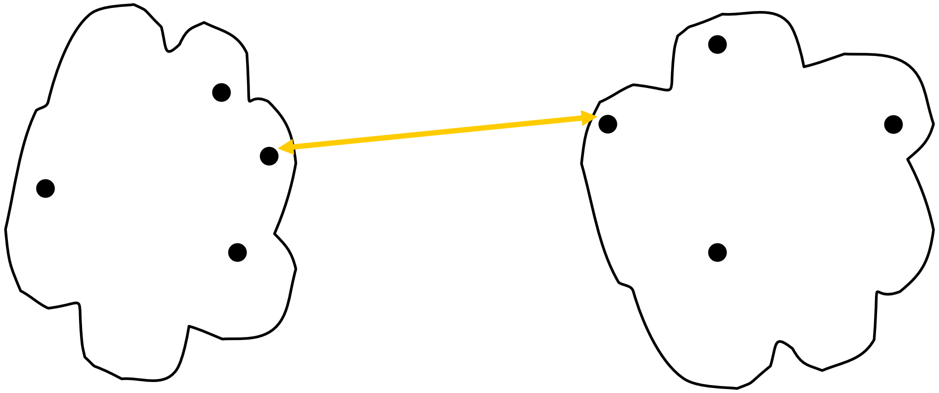
.

.

Proximity Matrix

- ❑ MIN
- ❑ MAX
- ❑ Group Average
- ❑ Distance Between Centroids
- ❑ Other methods driven by an objective function
  - Ward’s Method uses squared error

How to Define Inter-Cluster Similarity

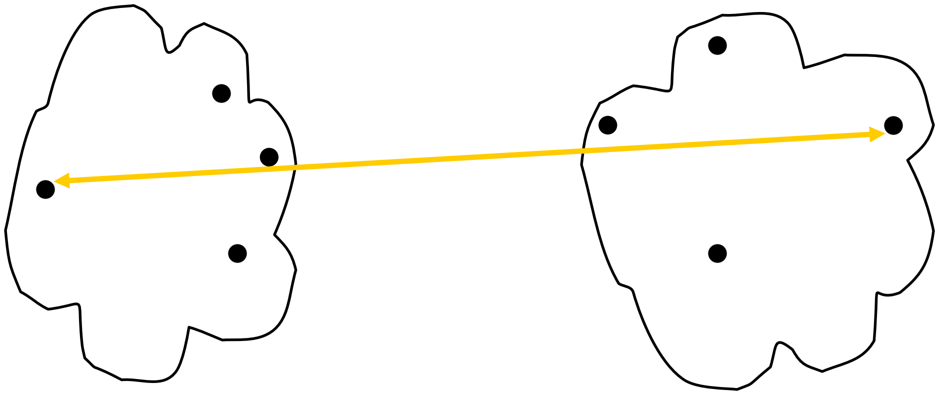


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

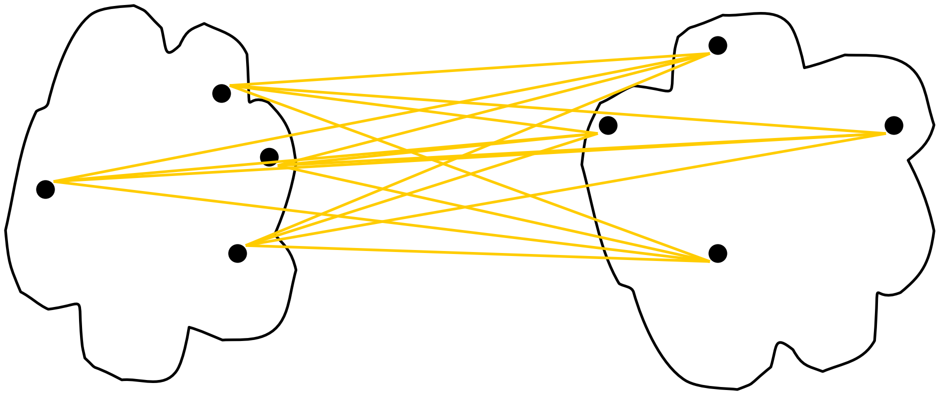


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

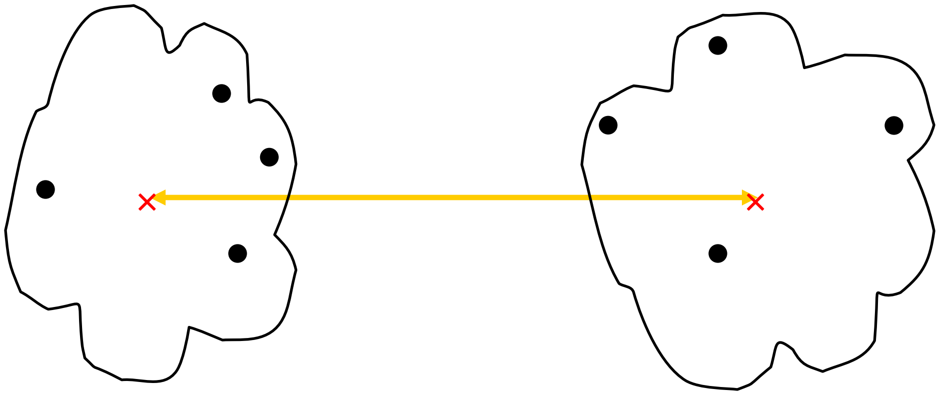


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward’s Method uses squared error

How to Define Inter-Cluster Similarity



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

.

.

Proximity Matrix

- ❑ MIN
- ❑ MAX
- ❑ Group Average
- ❑ Distance Between Centroids
- ❑ Other methods driven by an objective function
  - Ward’s Method uses squared error

Hierarchical Clustering: Time and Space requirements

- $O(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.

- $O(N^3)$ time in many cases
  - There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
  - Complexity can be reduced to $O(N^2 \log(N))$ time for some approaches

An example…

Example

- Suppose we have five items, a, b, c, d, and e.
- We wanto to perform hierarchical clustering on five instances following an agglomerative approach
- First: we compute the distance or similarity matrix
- D_{ij} is the distance between instancce “i” and “j”

$$D = \begin{pmatrix} 0.0 & & & & \\ 2.0 & 0.0 & & & \\ 6.0 & 5.0 & 0.0 & & \\ 10.0 & 9.0 & 4.0 & 0.0 & \\ 9.0 & 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}$$

Example

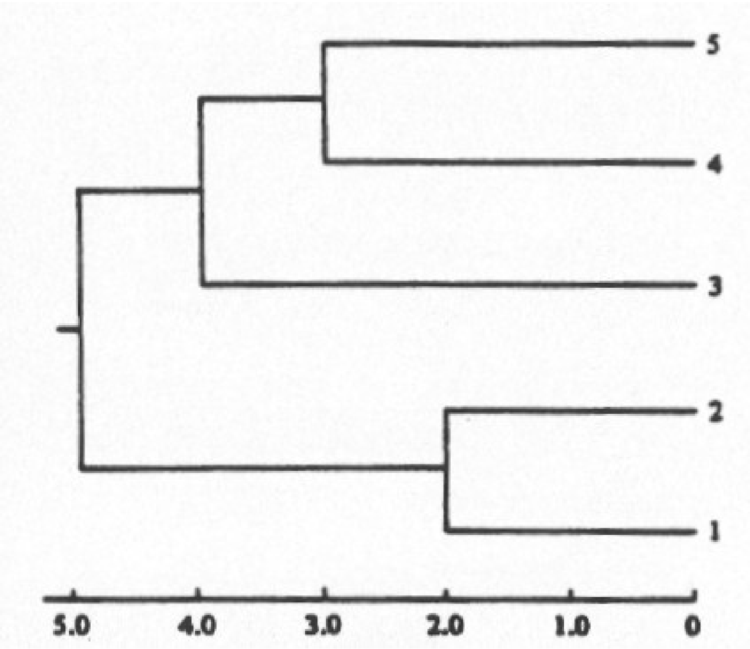
- Group the two instances that are closer
- In this case, a and b are the closest items ($D_{2,1}$=2)
- Compute again the distance matrix, and start again.
- Suppose we apply single-linkage (MIN), we need to compute the distance between the new cluster {1,2} and the others
  - $d_{(12)3} = \min[d_{13},d_{23}] = d_{23} = 5.0$
  - $d_{(12)4} = \min[d_{14},d_{24}] = d_{24} = 9.0$
  - $d_{(12)5} = \min[d_{15},d_{25}] = d_{25} = 8.0$

Example

- The new distance matrix is,

$$D = \begin{pmatrix} 0.0 & & & \\ 5.0 & 0.0 & & \\ 9.0 & 4.0 & 0.0 & \\ 8.0 & 5.0 & 3.0 & 0.0 \end{pmatrix}$$

- At the end, we obtain the following dendrogram

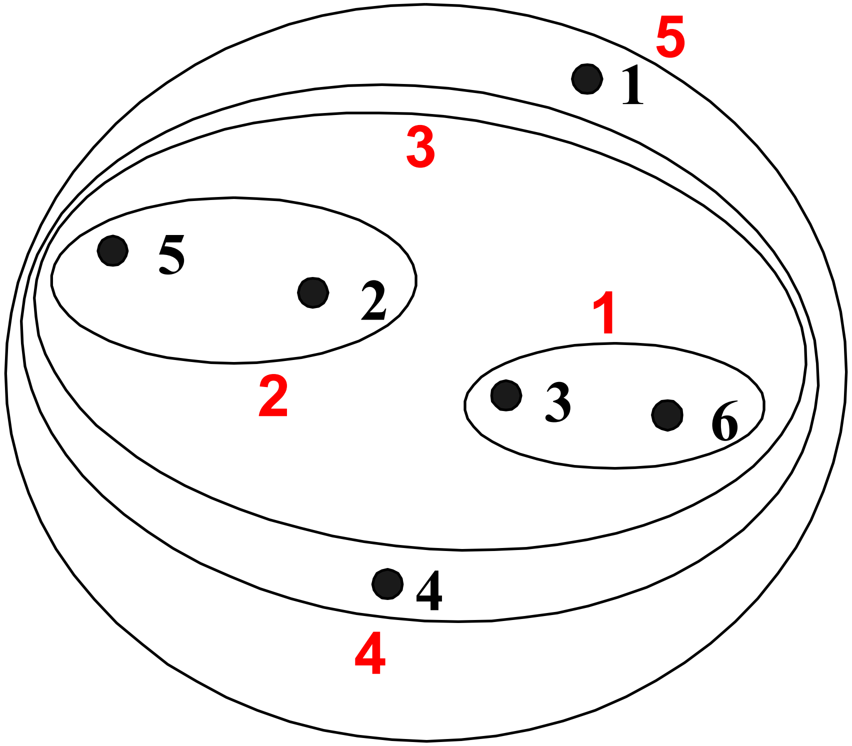


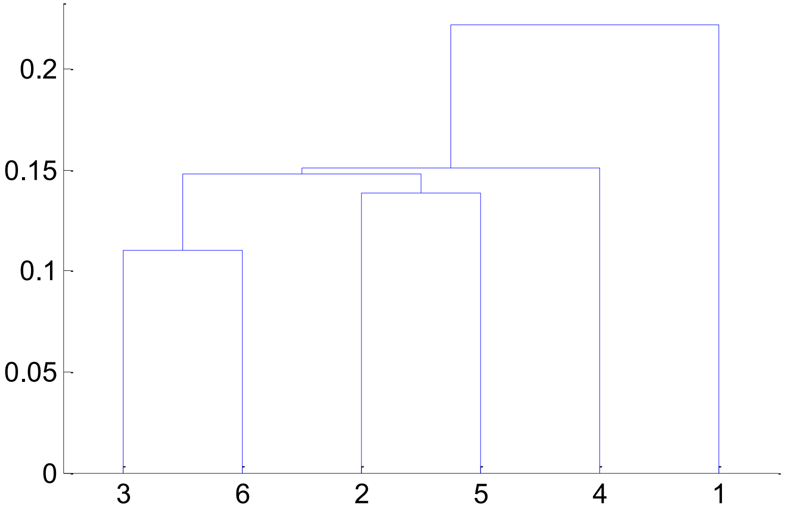
Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
- Determined by one pair of points, i.e., by one link in the proximity graph.

	P1	P2	P3	P4	P5	P6
P1	0.0					
P2	0.24	0.0				
P3	0.22	0.15	0.00			
P4	0.37	0.20	0.15	0.00		
P5	0.34	0.14	0.28	0.29	0.00	
P6	0.23	0.25	0.11	022	0.39	0.00

Hierarchical Clustering: MIN





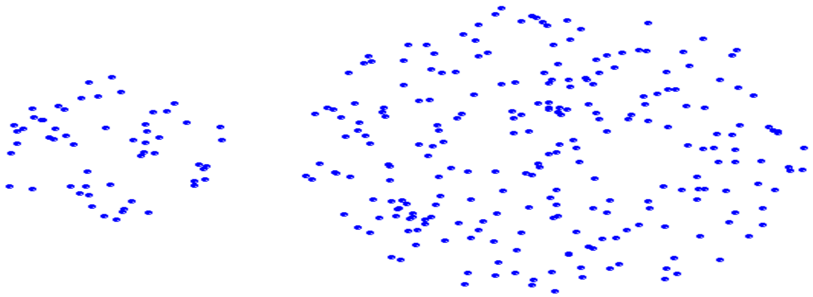
Nested Clusters

Dendrogram

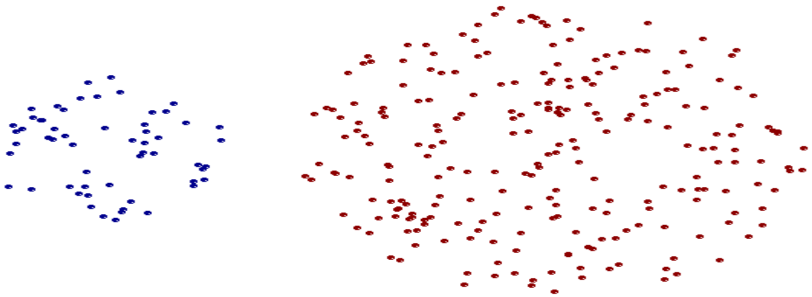
Strengths and limitations…

Strength of MIN

- Can handle non-elliptical shapes



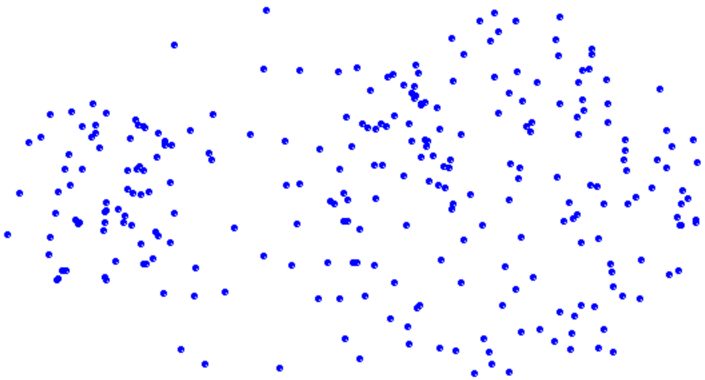
Original Points



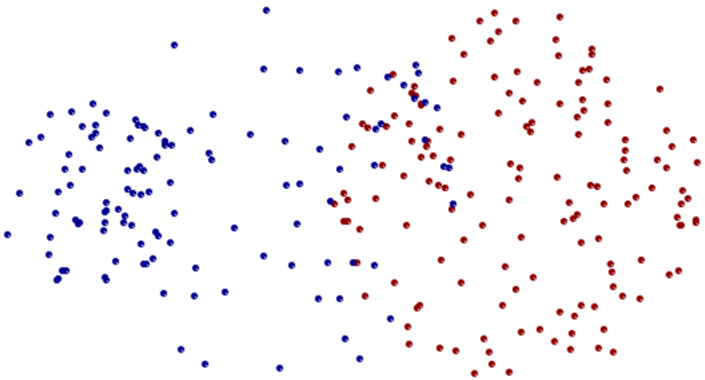
Two Clusters

Limitations of MIN

- Sensitive to noise and outliers



Original Points

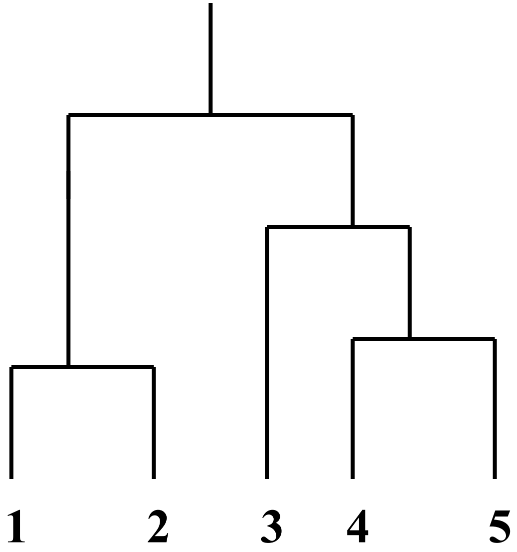


Two Clusters

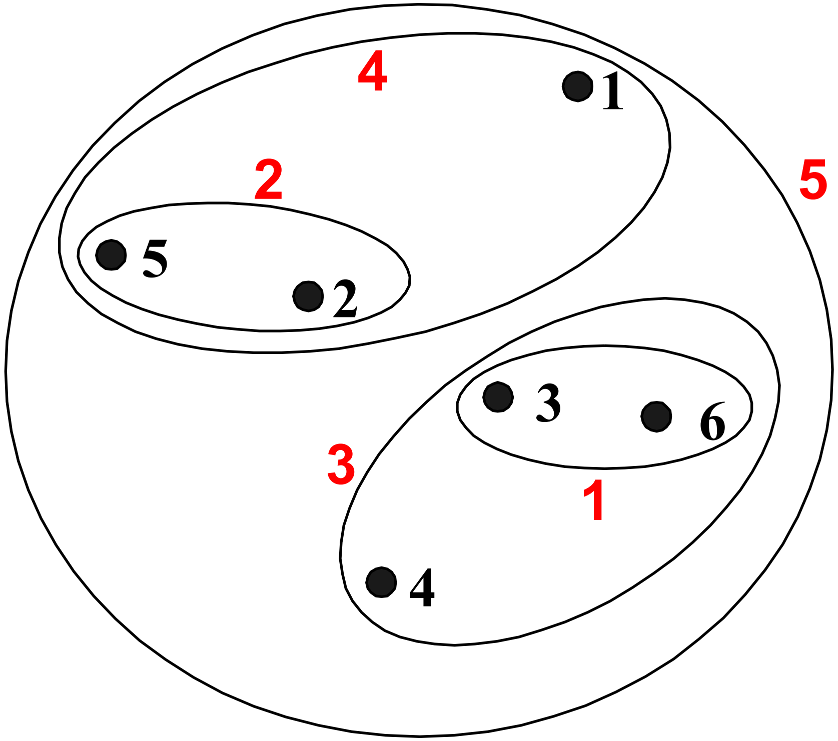
Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

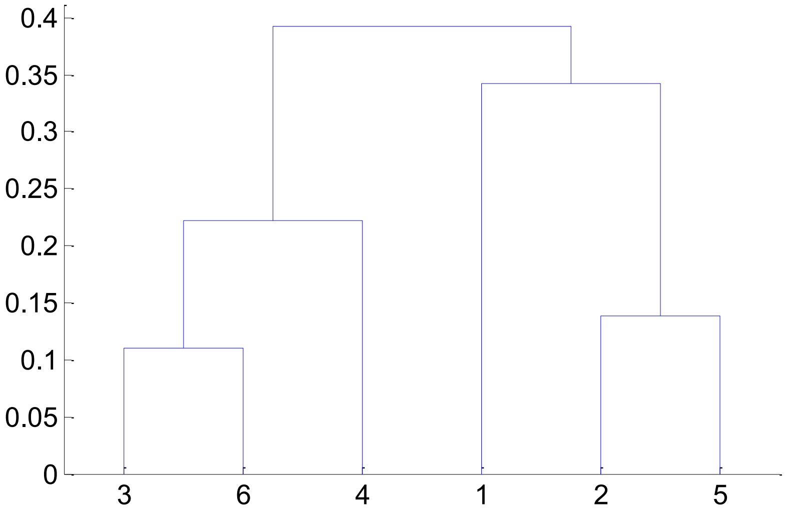
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: MAX



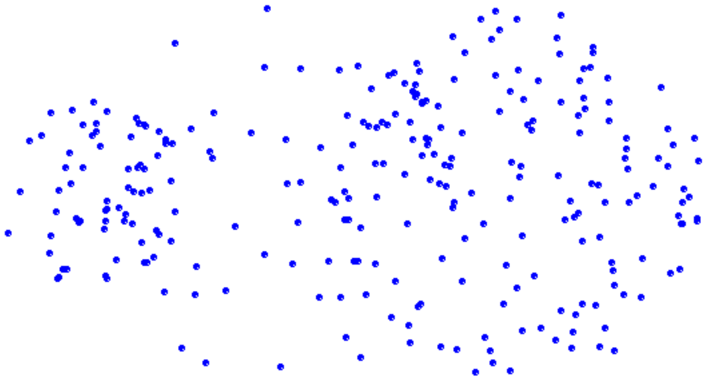
Nested Clusters



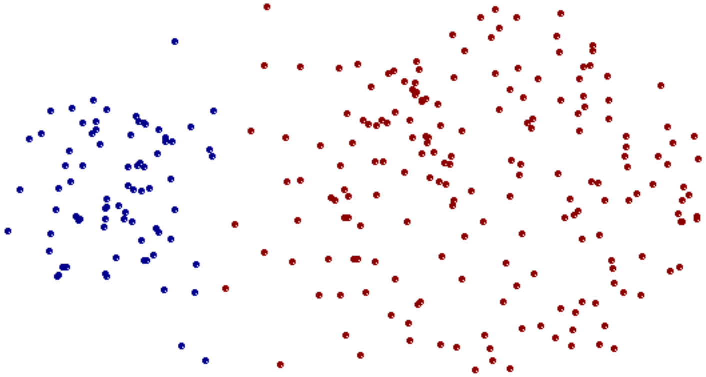
Dendrogram

Strength of MAX

- Less susceptible to noise and outliers



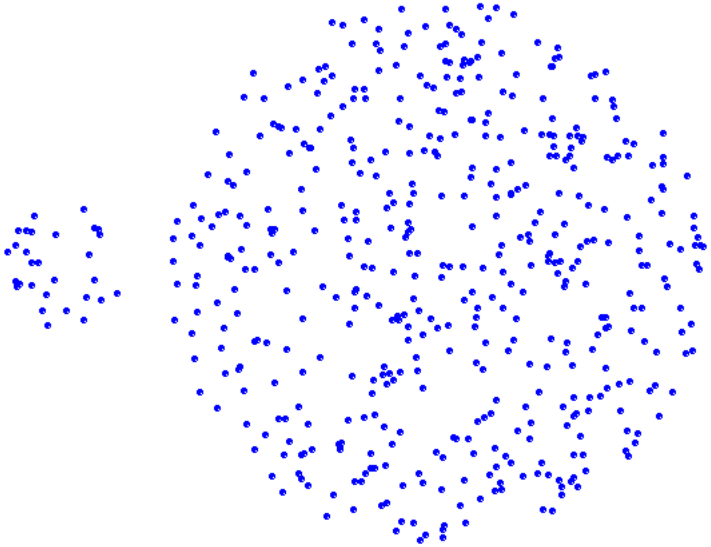
Original Points



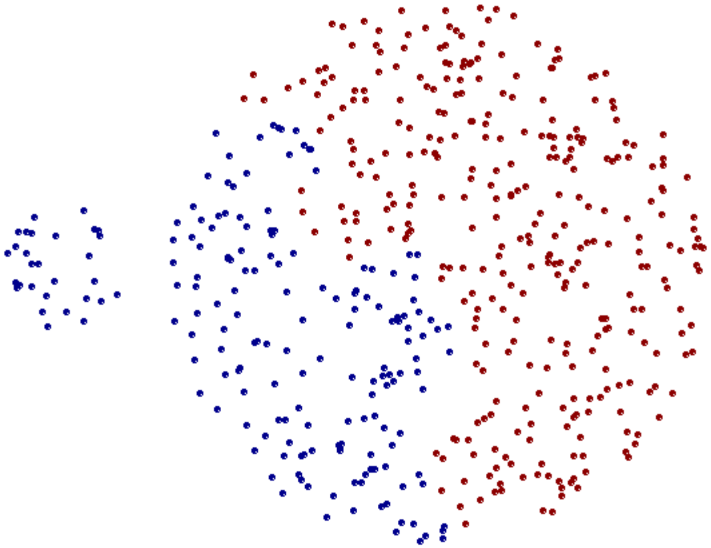
Two Clusters

Limitations of MAX

- Tends to break large clusters
- Biased towards globular clusters



Original Points



Two Clusters

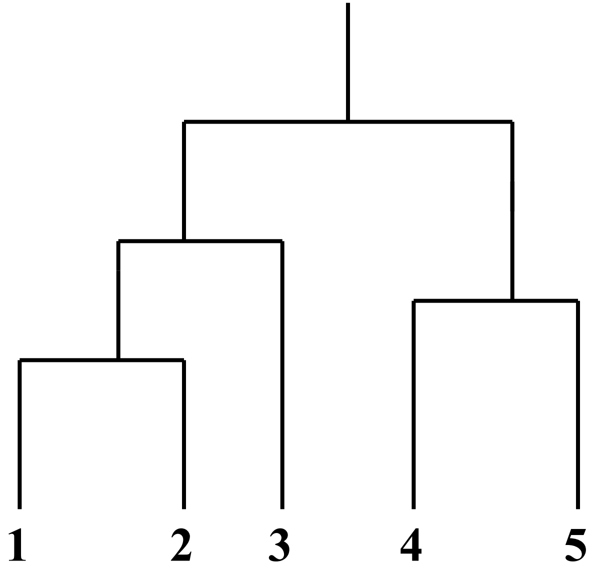
Cluster Similarity: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

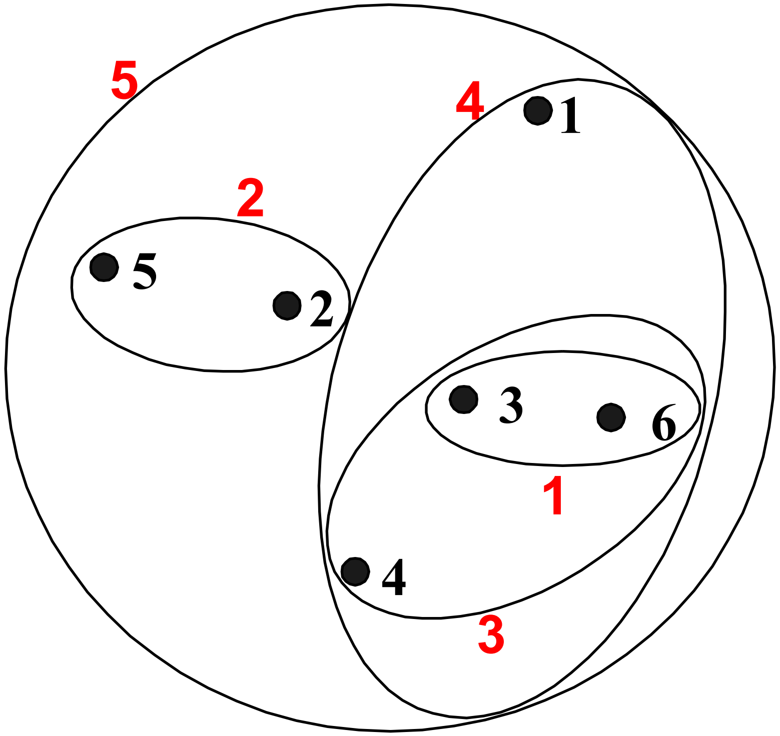
$$\mathbf{proximity}(\mathbf{Cluster}_i,\mathbf{Cluster}_j) = \frac{\sum\limits_{\substack{p_i \in \mathbf{Cluster}_i \\ p_j \in \mathbf{Cluster}_j}} \mathbf{proximity}(p_i,p_j)}{|\mathbf{Cluster}_i| * |\mathbf{Cluster}_j|}$$

- Need to use average connectivity for scalability since total proximity favors large clusters

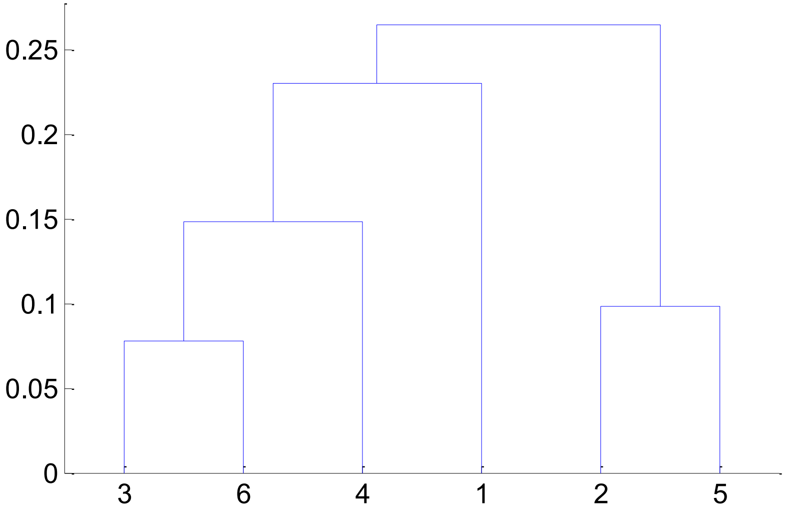
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: Group Average



Nested Clusters



Dendrogram

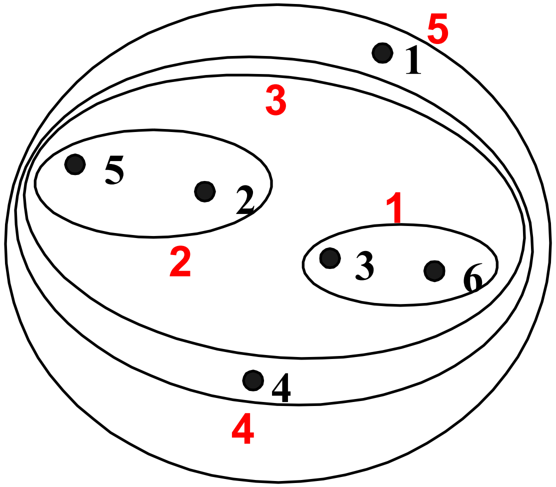
Hierarchical Clustering: Group Average

- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

Cluster Similarity: Ward's Method

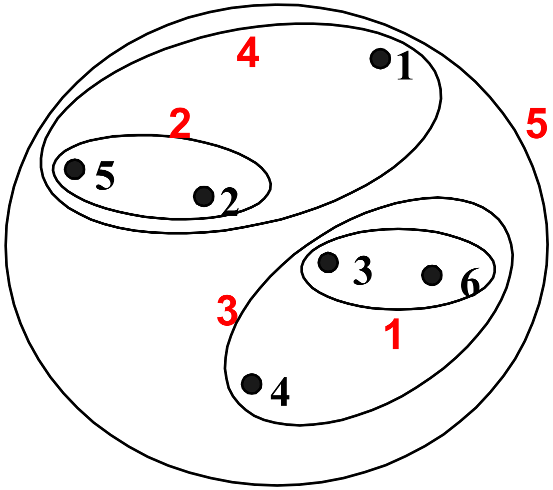
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

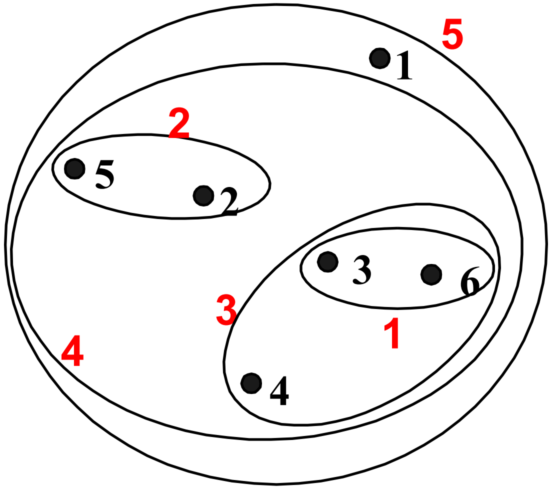
Hierarchical Clustering: Comparison



MIN

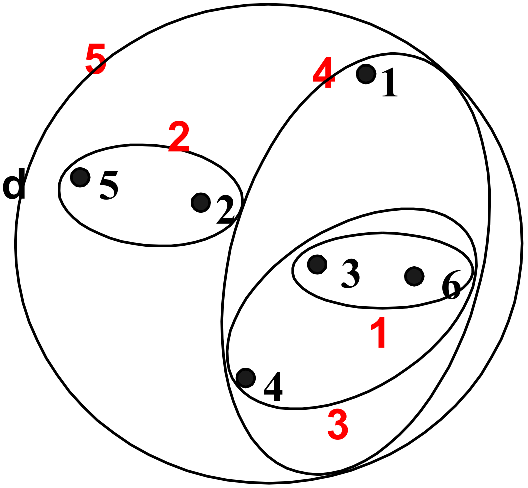
MAX





Group Average

Ward's Method



Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

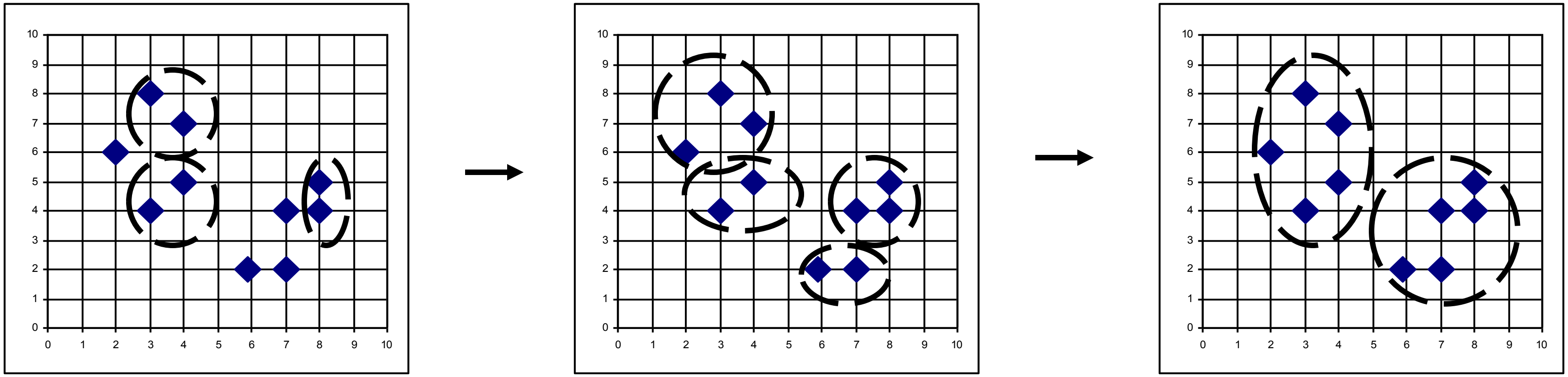
Measure the Quality of Clustering

- Dissimilarity/Similarity metric
- Similarity is expressed in terms of a distance function, which is typically metric: d(i, j)
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, Boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough"
    - the answer is typically highly subjective.

Algorithms…

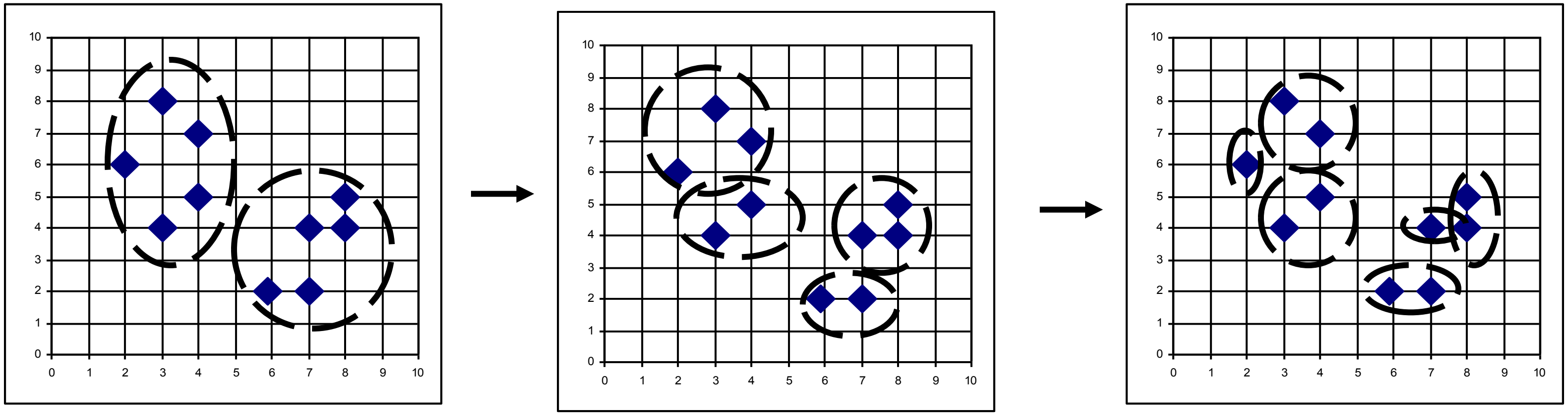
AGNES (Agglomerative Nesting)

- ❑ Introduced in Kaufmann and Rousseeuw (1990)
- ❑ Implemented in statistical analysis packages, e.g., Splus
- ❑ Use the Single-Link method and the dissimilarity matrix.
- ❑ Merge nodes that have the least dissimilarity
- ❑ Go on in a non-descending fashion
- ❑ Eventually all nodes belong to the same cluster



DIANA (Divisive Analysis)

- ❑ Introduced in Kaufmann and Rousseeuw (1990)
- ❑ Implemented in statistical analysis packages, e.g., Splus
- ❑ Inverse order of AGNES
- ❑ Eventually each node forms a cluster on its own



Recent Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
  - do not scale well: time complexity of at least O(n^2), where n is the number of total objects
  - can never undo what was done previously

- Integration of hierarchical with distance-based clustering
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
  - ROCK (1999): clustering categorical data by neighbor and link analysis
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling

BIRCH (1996)

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies (Zhang, Ramakrishnan & Livny, SIGMOD'96)
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering
  - Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
  - Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree
- Scales linearly: finds a good clustering with a single scan and improves the quality with a few additional scans
- Weakness: handles only numeric data, and sensitive to the order of the data record.

CHAMELEON: Hierarchical Clustering Using Dynamic Modeling (1999)

- CHAMELEON: by G. Karypis, E.H. Han, and V. Kumar’99
- Measures the similarity based on a dynamic model
  - Two clusters are merged only if the interconnectivity and closeness (proximity) between two clusters are high relative to the internal interconnectivity of the clusters and closeness of items within the clusters
  - Cure ignores information about interconnectivity of the objects, Rock ignores information about the closeness of two clusters
- A two-phase algorithm
  - Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters
  - Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters

Summary

Summary

- ❑ Hierarchical clustering builds a hierarchy from data working either bottom-up (agglomerative) or top-down (divisive)
- ❑ No need to assume any particular number of clusters
- ❑ Any desired number of clusters can be obtained by ‘cutting’ the hierarchy at the proper level
- ❑ Hierarchies may correspond to meaningful taxonomies
- ❑ Distance function to measure similarity between items
- ❑ Inter-cluster similarity defined based on the distance (MIN, MAX, Group Average, etc.)