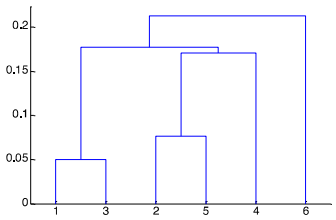
Cluster Analysis

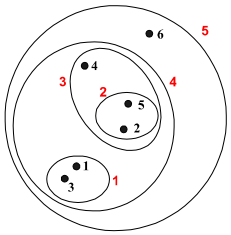
K-means vs. Hierarchical Clustering

- K-means uses the intra cluster distance to find clusters
- Hierarchical clustering uses the inter cluster distances.

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree





Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - 'cut' the dendogram at the proper level to have a certain number of clusters
- They may correspond to meaningful taxonomies
  - Example in biological sciences e.g.,
    - animal kingdom,
    - phylogeny reconstruction,
    - …

Hierarchical Clustering

Algorithm

Let each data point be a cluster

Compute the proximity matrix

Repeat

Merge **the two closest** clusters

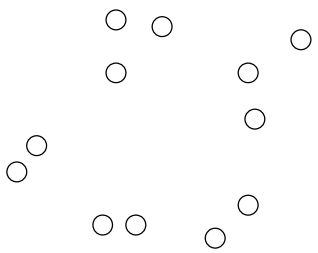
Update the proximity matrix

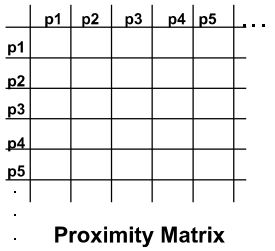
Until only a single cluster remains

- Key operation is the computation of *cluster closeness*

Starting Situation

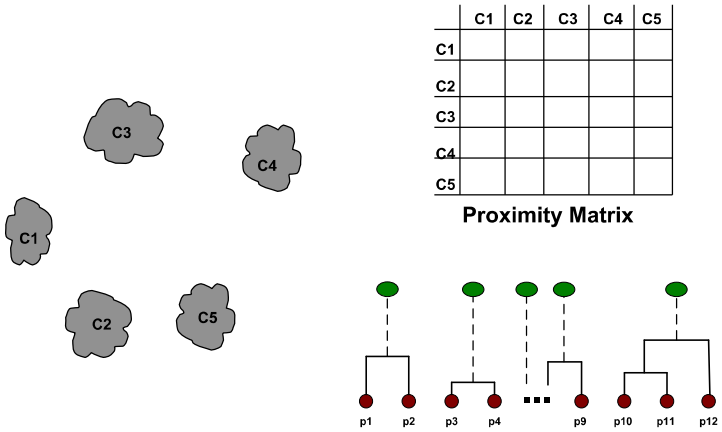
- Start with clusters of individual points and a proximity matrix





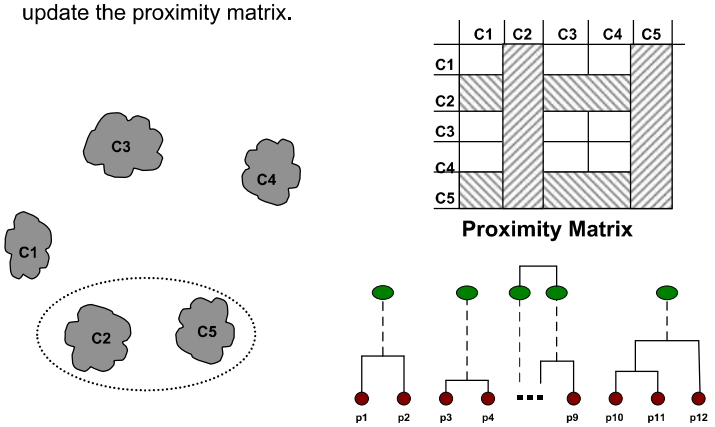
Intermediate Situation

- After some merging steps, we have some clusters



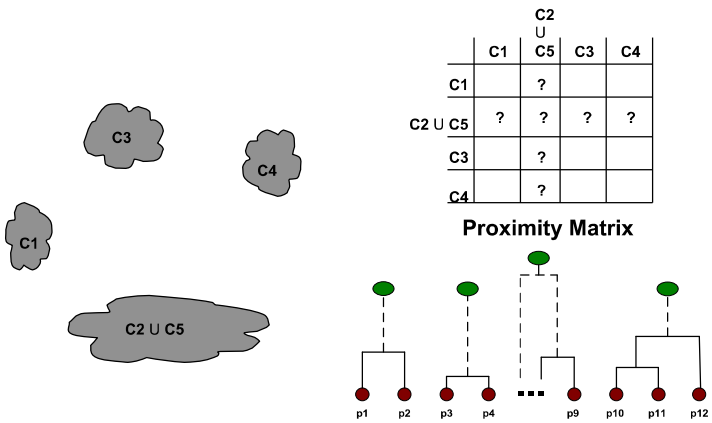
Intermediate Situation

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

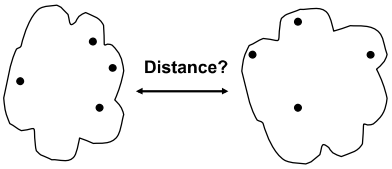


After Merging

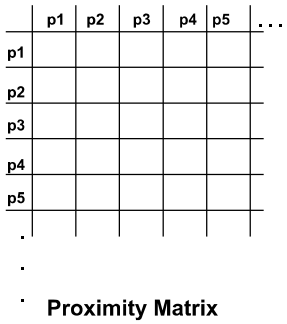
- The question is "How do we update the proximity matrix?"



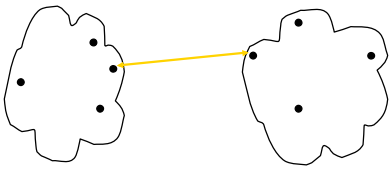
First Define Inter-Cluster Similarity



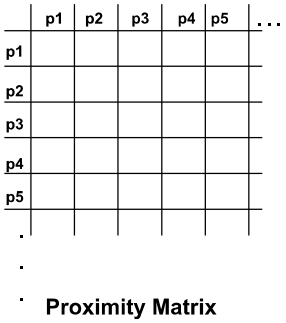
- MIN
- MAX
- Group Average



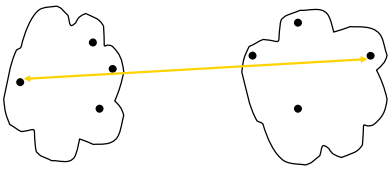
How to Define Inter-Cluster Similarity



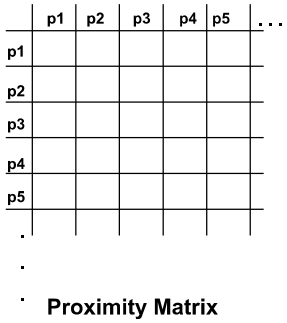
- MIN
- MAX
- Group Average



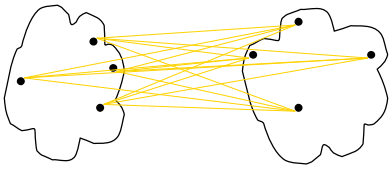
How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average



How to Define Inter-Cluster Similarity



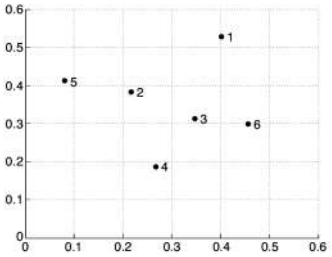
- MIN
- MAX
- Group Average

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

Proximity Matrix

Cluster Similarity: MIN

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points

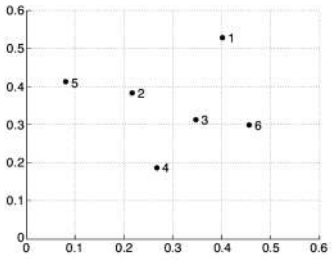


```
        1       2       3       4       5       6
[1[ 0.  ,  0.24,  0.22,  0.37,  0.34,  0.23],
 2[ 0.24,  0.  ,  0.15,  0.2 ,  0.14,  0.25],
 3[ 0.22,  0.15,  0.  ,  0.16,  0.28,  0.11],
 4[ 0.37,  0.2 ,  0.16,  0.  ,  0.29,  0.22],
 5[ 0.34,  0.14,  0.28,  0.29,  0.  ,  0.39],
 6[ 0.23,  0.25,  0.11,  0.22,  0.39,  0.  ]]
```

```
        1       2     3&6       4       5
[1[ 0.  ,  0.24,  0.22,  0.37,  0.34],
 2[ 0.24,  0.  ,  0.15,  0.2 ,  0.14],
3&6[ 0.22,  0.15,  0.  ,  0.16,  0.28],
 4[ 0.37,  0.2 ,  0.16,  0.  ,  0.29],
 5[ 0.34,  0.14,  0.28,  0.29,  0.  ],]
```

Cluster Similarity: MIN

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points

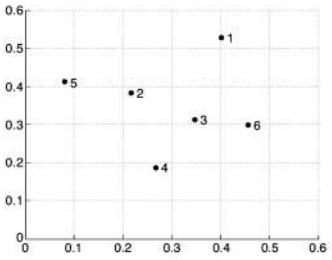


```
        1       2     3&6       4       5
[1[ 0.  ,  0.24,  0.22,  0.37,  0.34],
 2[ 0.24,  0.  ,  0.15,  0.2 ,  0.14],
3&6[ 0.22,  0.15,  0.  ,  0.16,  0.28],
 4[ 0.37,  0.2 ,  0.16,  0.  ,  0.29],
 5[ 0.34,  0.14,  0.28,  0.29,  0.  ],]
```

```
        1     2&5     3&6       4
[1[ 0.  ,  0.24,  0.22,  0.37],
2&5[ 0.24,  0.  ,  0.15,  0.2 ],
3&6[ 0.22,  0.15,  0.  ,  0.16],
 4[ 0.37,  0.2 ,  0.16,  0.  ],]
```

Cluster Similarity: MIN

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points

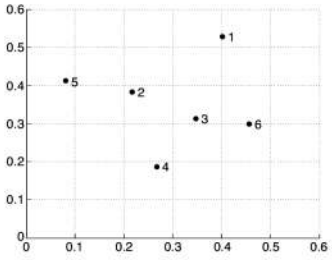


```
        1     2&5     3&6       4
[1[ 0.  ,  0.24,  0.22,  0.37],
2&5[ 0.24,  0.  ,  0.15,  0.2 ],
3&6[ 0.22,  0.15,  0.  ,  0.16],
 4[ 0.37,  0.2 ,  0.16,  0.  ],]
```

```
        1   2&3&5&6       4
[1[ 0.  ,  0.22,  0.37],
2&3&5&6[ 0.22,  0.  ,  0.16 ],
 4[ 0.37,  0.16,  0.  ],]
```

Cluster Similarity: MIN

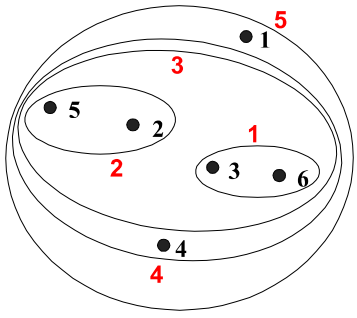
- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points



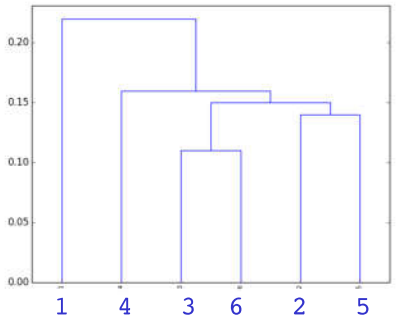
```
        1   2&3&5&6       4
[1[ 0.  ,  0.22,  0.37],
2&3&5&6[ 0.22,  0.  ,  0.16 ],
 4[ 0.37,  0.16,  0.  ],]
```

```
        1   2&3&4&5&6
[1[ 0.  ,  0.22],
2&3&5&6[ 0.22,  0.  ]]
```

Hierarchical Clustering: MIN

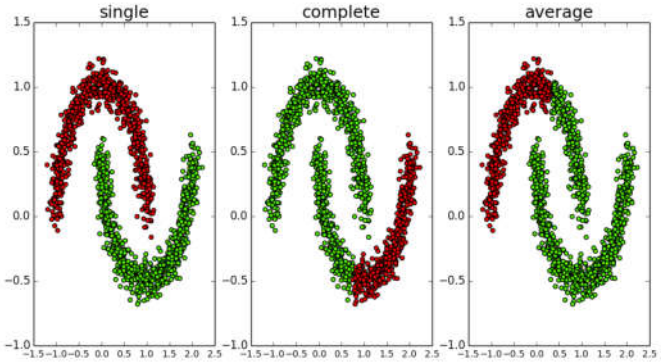


Nested Clusters



Dendrogram

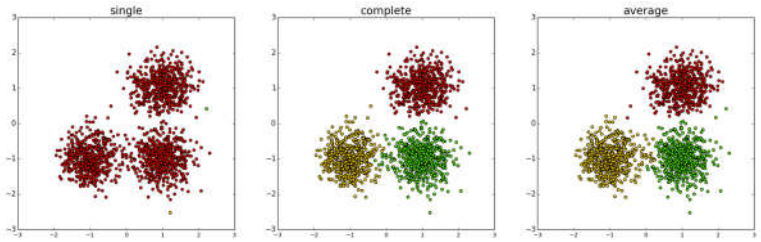
Strength of MIN (single)



Can handle non-globular shapes

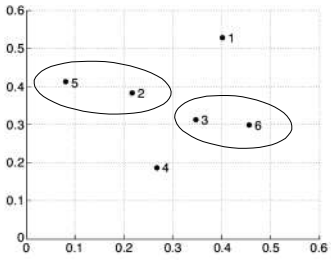
Limitations of MIN (single)

Sensitive to noise and outliers



Cluster Similarity: MAX

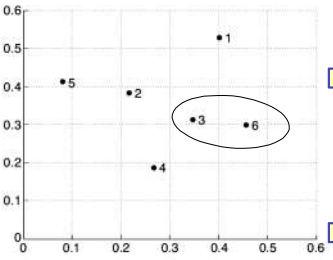
- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters



```
[[ 0.  ,  0.24,  0.22,  0.37,  0.34,  0.23],
 [ 0.24,  0.  ,  0.15,  0.2 ,  0.14,  0.25],
 [ 0.22,  0.15,  0.  ,  0.16,  0.28,  0.11],
 [ 0.37,  0.2 ,  0.16,  0.  ,  0.29,  0.22],
 [ 0.34,  0.14,  0.28,  0.29,  0.  ,  0.39],
 [ 0.23,  0.25,  0.11,  0.22,  0.39,  0.  ]]
```

Cluster Similarity: MAX

- 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

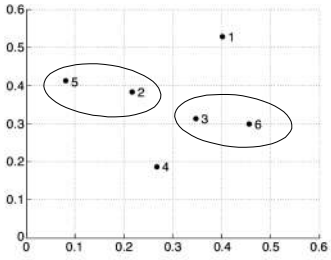


```
[[ 0.  ,  0.24,  0.22,  0.37,  0.34,  0.23],
 [ 0.24,  0.  ,  0.15,  0.2 ,  0.14,  0.25],
 [ 0.22,  0.15,  0.  ,  0.16,  0.28,  0.11],
 [ 0.37,  0.2 ,  0.16,  0.  ,  0.29,  0.22],
 [ 0.34,  0.14,  0.28,  0.29,  0.  ,  0.39],
 [ 0.23,  0.25,  0.11,  0.22,  0.39,  0.  ]]
```

```
          1     2   3&6     4     5
[ 1[ 0.  ,  0.24,  0.23,  0.37,  0.34],
  2[ 0.24,  0.  ,  0.25,  0.2 ,  0.14],
3&6[ 0.23,  0.25,  0.  ,  0.22,  0.39],
  4[ 0.37,  0.2 ,  0.22,  0.  ,  0.29],
  5[ 0.34,  0.14,  0.39,  0.29,  0.  ]]
```

Cluster Similarity: MAX

- 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

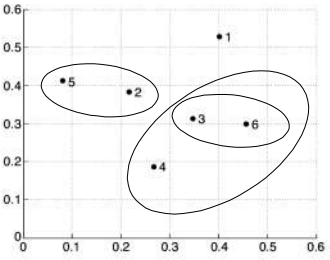


```
          1     2   3&6     4     5
[ 1[ 0.  ,  0.24,  0.23,  0.37,  0.34],
  2[ 0.24,  0.  ,  0.25,  0.2 ,  0.14],
3&6[ 0.23,  0.25,  0.  ,  0.22,  0.39],
  4[ 0.37,  0.2 ,  0.22,  0.  ,  0.29],
  5[ 0.34,  0.14,  0.39,  0.29,  0.  ]]
```

```
          1   2&5   3&6     4
[ 1[ 0.  ,  0.34,  0.23,  0.37],
2&5[ 0.34,  0.  ,  0.39,  0.29],
3&6[ 0.23,  0.39,  0.  ,  0.22],
  4[ 0.37,  0.29,  0.22,  0.  ]]
```

Cluster Similarity: MAX

- 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

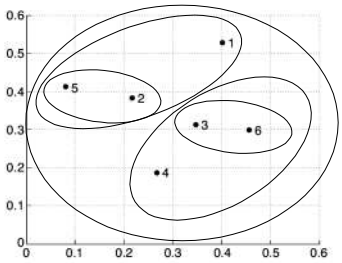


```
          1   2&5   3&6     4
[ 1[ 0.  ,  0.34,  0.23,  0.37],
2&5[ 0.34,  0.  ,  0.39,  0.29],
3&6[ 0.23,  0.39,  0.  ,  0.22],
  4[ 0.37,  0.29,  0.22,  0.  ]]
```

```
          1   2&5 3&4&6
[ 1[ 0.  ,  0.34,  0.37],
2&5[ 0.34,  0.  ,  0.39],
3&6[ 0.37,  0.39,  0.  ]]
```

Cluster Similarity: MAX

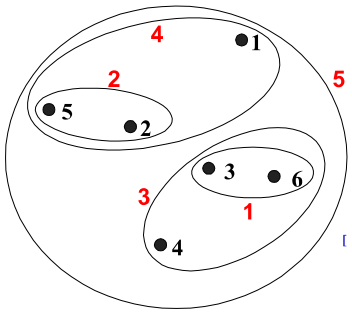
- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters
-



	1	2&5	3&4&6
1	0.	0.34	0.37
2&5	0.34	0.	0.39
3&4&6	0.37	0.39	0.

	1&2&5	3&4&6
1&2&5	0.	0.39
3&6	0.39	0.

Hierarchical Clustering: MAX

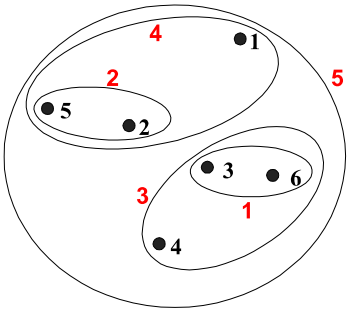


Nested Clusters

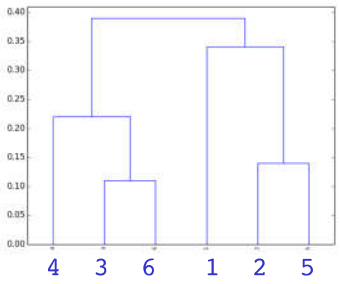
```
[[ 0.  , 0.24, 0.22, 0.37, 0.34, 0.23],
 [ 0.24, 0.  , 0.15, 0.2 , 0.14, 0.25],
 [ 0.22, 0.15, 0.  , 0.16, 0.28, 0.11],
 [ 0.37, 0.2 , 0.16, 0.  , 0.29, 0.22],
 [ 0.34, 0.14, 0.28, 0.29, 0.  , 0.39],
 [ 0.23, 0.25, 0.11, 0.22, 0.39, 0.  ]]
```

Note: Algorithmically, you need to consider *all pairs* of clusters. Here, we use the figure to avoid some of the silly comparisons (e.g. when we considered merging 4 with 3&6 we didn't consider merging with 1)

Hierarchical Clustering: MAX

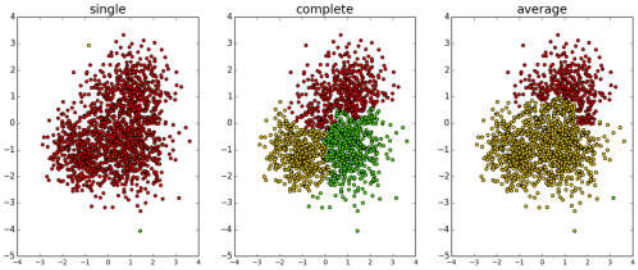


Nested Clusters



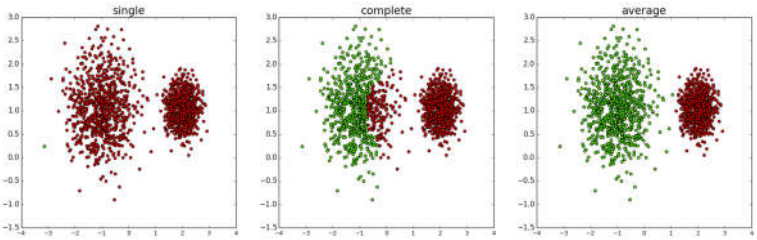
Dendrogram

Strengths of MAX (complete)



Less susceptible with respect to noise and outliers

Limitations of MAX (complete)



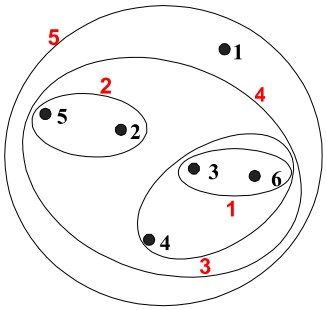
Tends to break large clusters

Cluster Similarity: Group Average

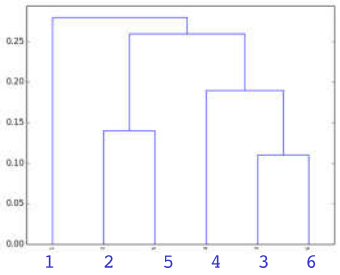
- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

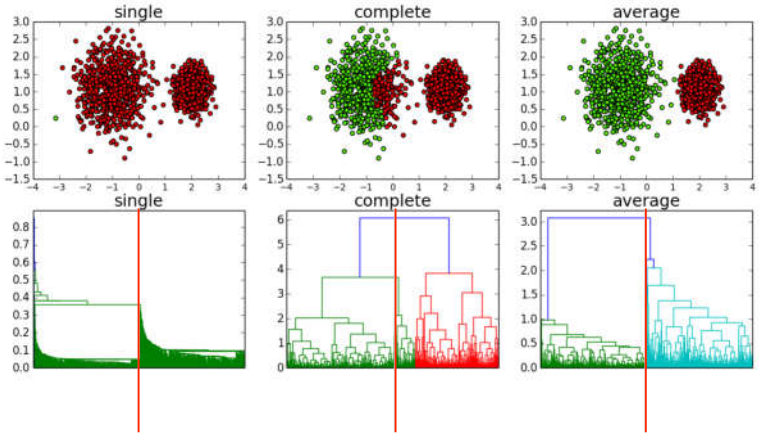
Hierarchical Clustering: Group Average



Nested Clusters

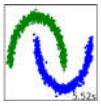


Dendrogram

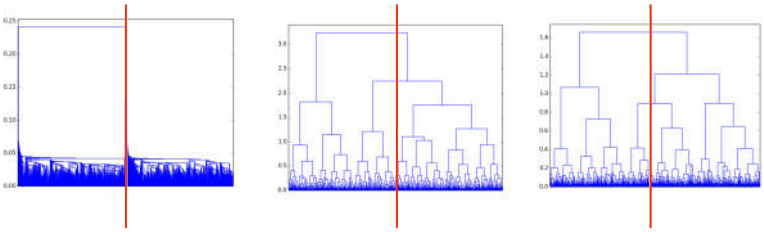


Self test

- Which dendrogram can create this clustering?



separation of two clusters



- Which merging method is it?