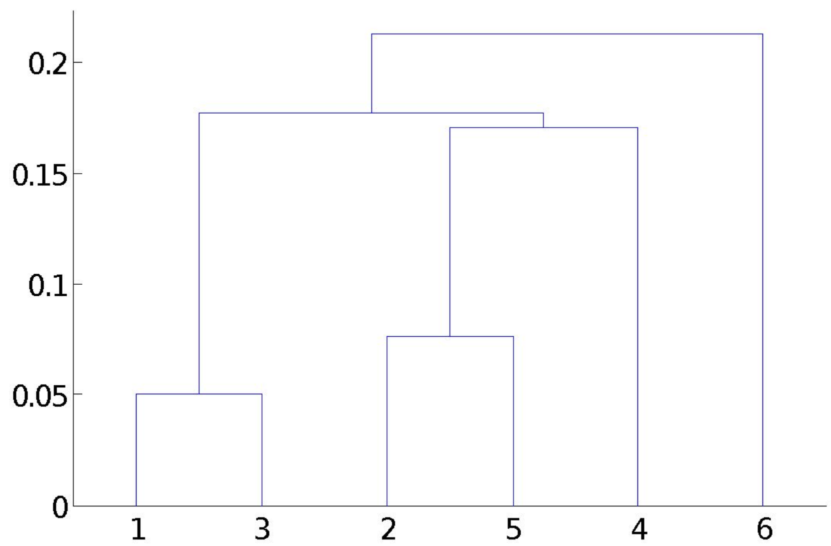
Birla Institute of Applied Sciences
बिरला इंस्टिट्यूट ऑफ़ अप्लाइड साइंसेस
Bhimtal, Distt: Nainital, Uttarakhand- 263136

Pattern Recognition

- S. S. Samant

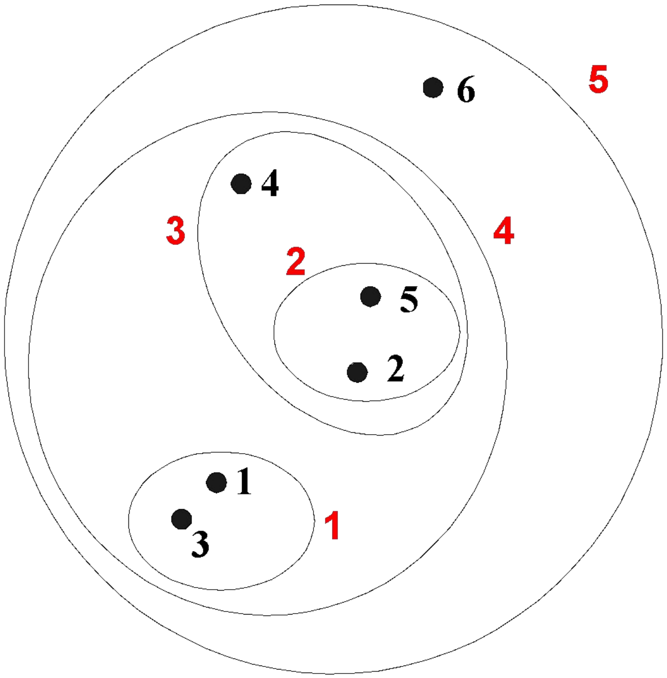
Agglomerative Hierarchical Clustering

- Starts with each point being a cluster, and at each step, merge the *closest* pair of clusters
- Displayed graphically using a **dendrogram – a tree like structure** (*dendro* "tree", *gramma* "drawing")



Agglomerative Hierarchical Clustering

- Starts with each point being a cluster, and at each step, merge the *closest* pair of clusters
- Can also be displayed graphically using a **nested cluster diagram**



Agglomerative Hierarchical Clustering

Basic algorithm

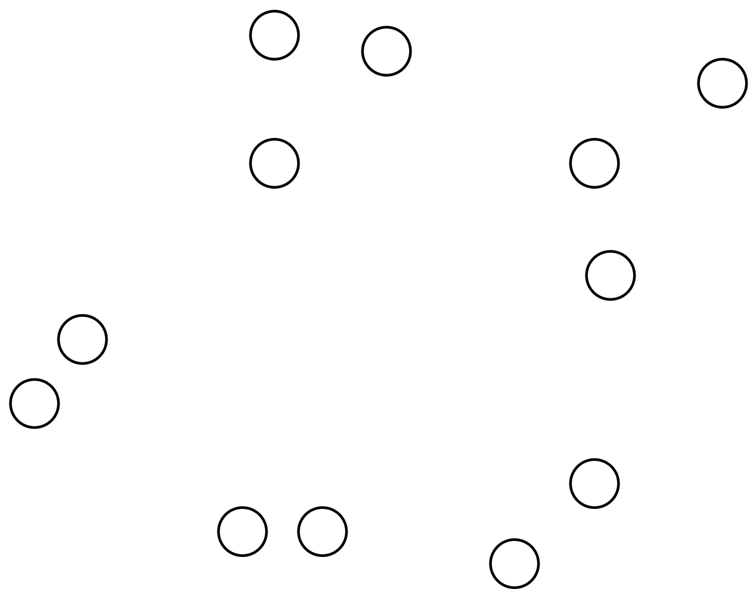
1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

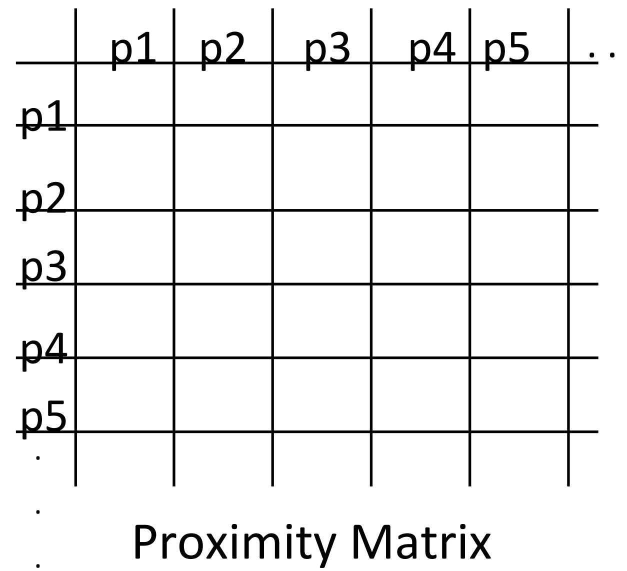
Key operation is the computation of the proximity of two clusters

- Different approaches to defining the distance between clusters distinguish the different algorithms

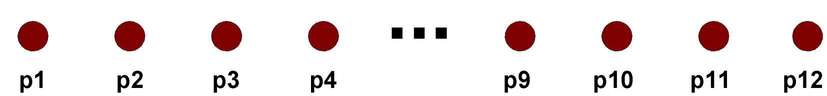
Starting Situation

- Start with clusters of individual points and a proximity matrix



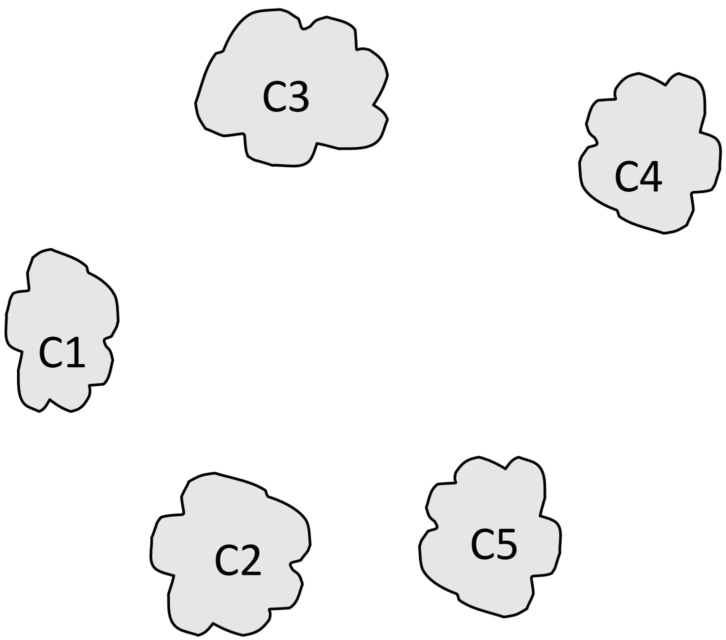


Proximity Matrix



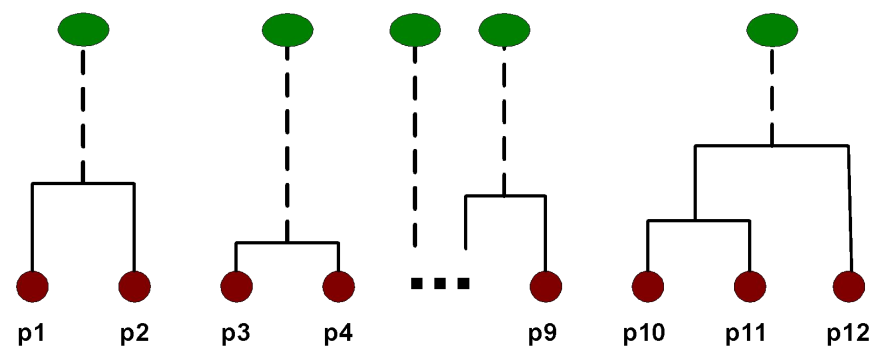
Intermediate Situation

- After some merging steps, we have some clusters



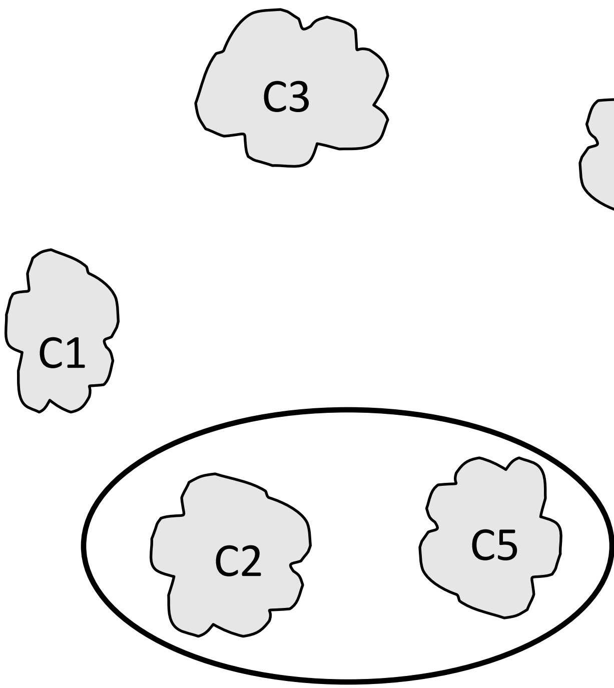
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

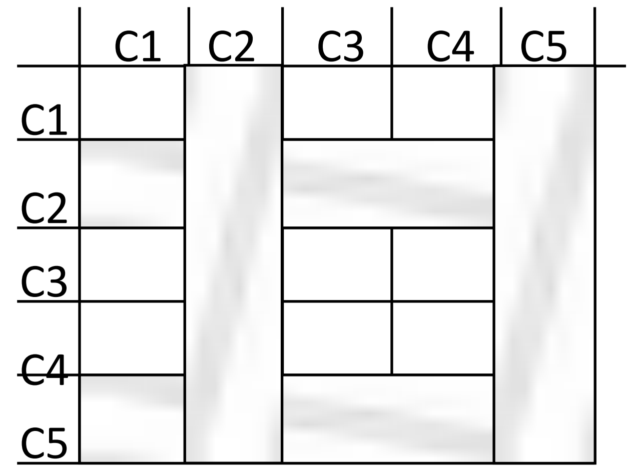
Proximity Matrix



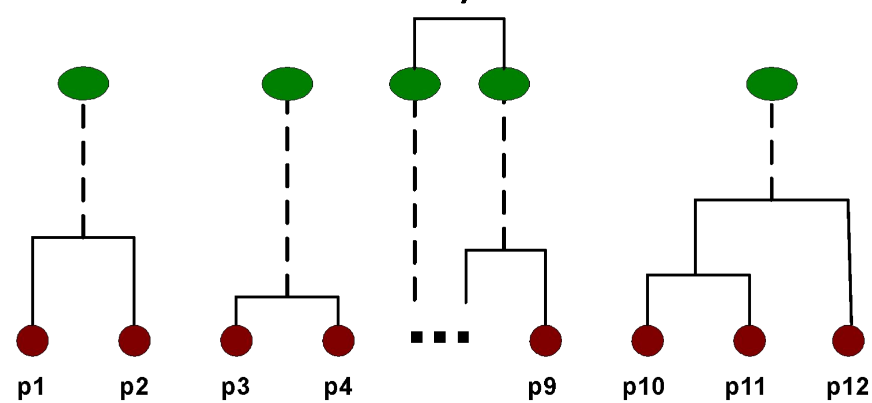
Intermediate Situation

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





Proximity Matrix

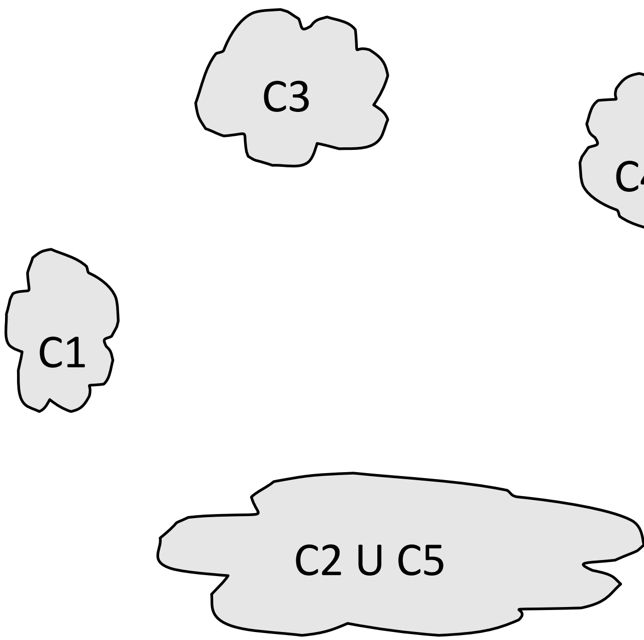


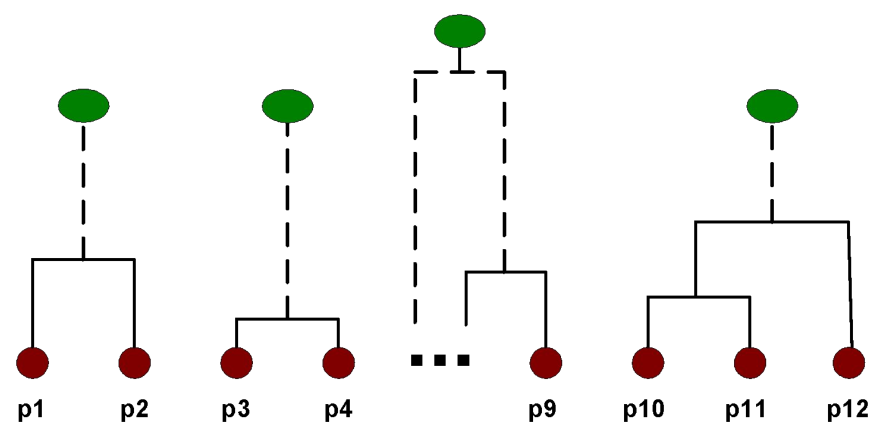
After Merging

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

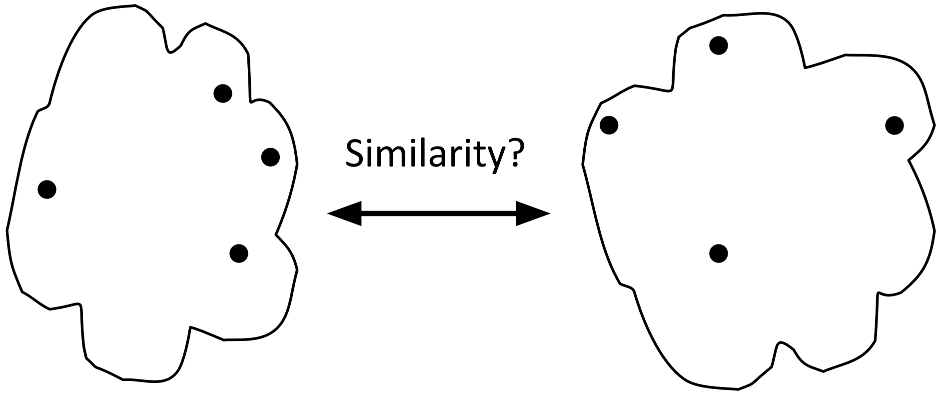
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix





How to Define Inter-Cluster Similarity

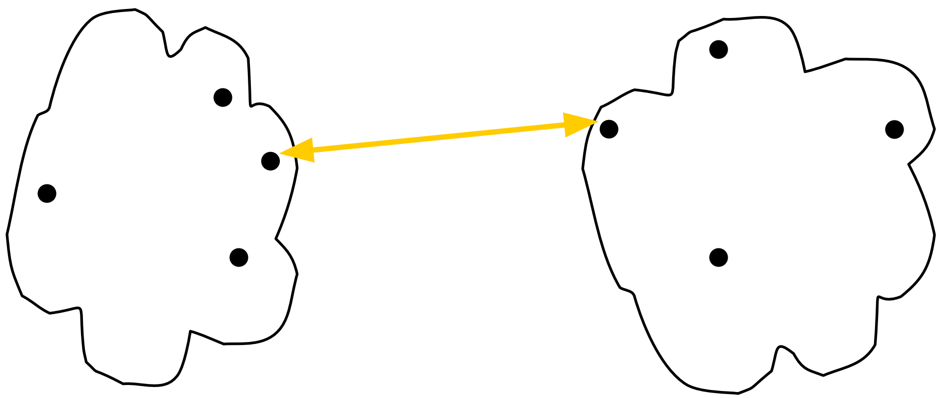


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity

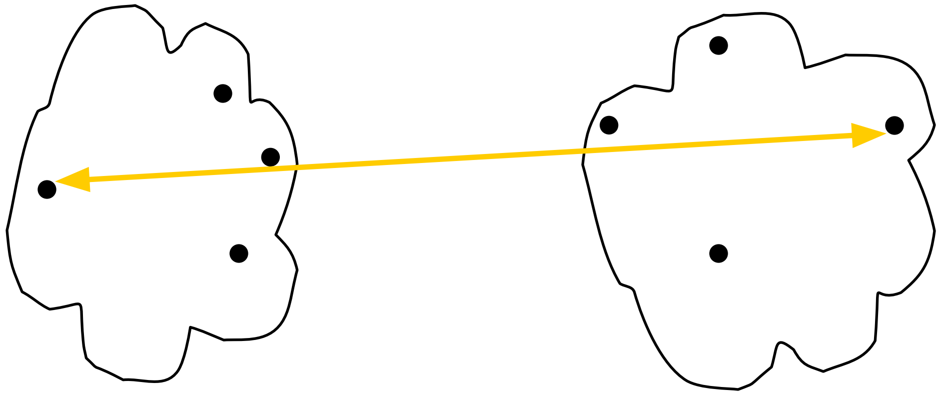


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity

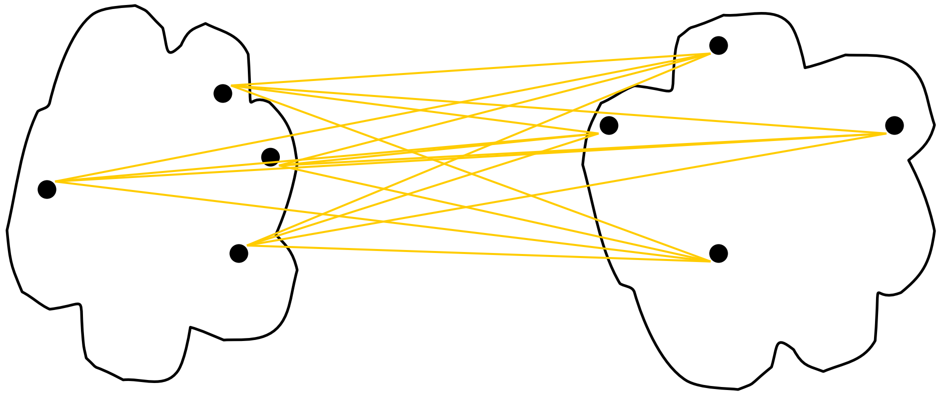


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity

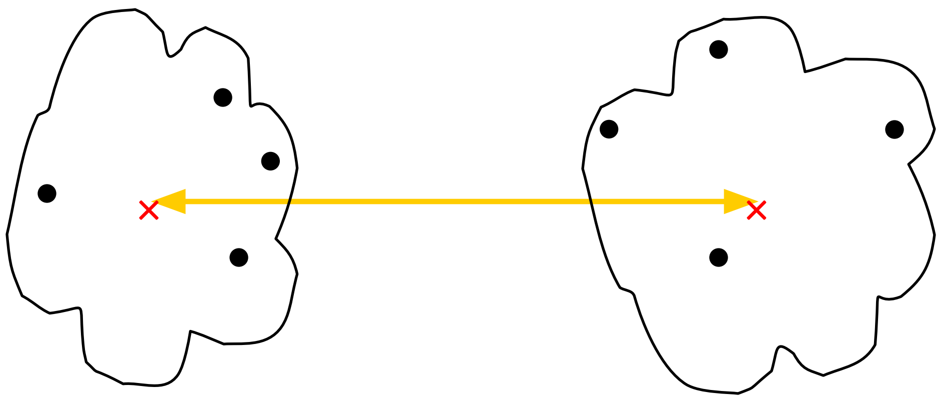


- MIN
- MAX
- Group Average
- Distance Between Centroids

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

Proximity Matrix

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

Cluster Similarity: MIN or Single Link

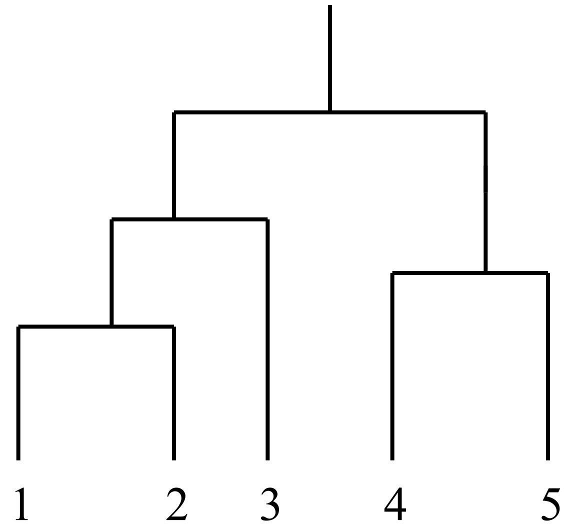
- Similarity of two clusters is based on the two closest points in the different clusters

	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

Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two closest points in the different 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



Cluster Similarity: MAX or Complete Linkage

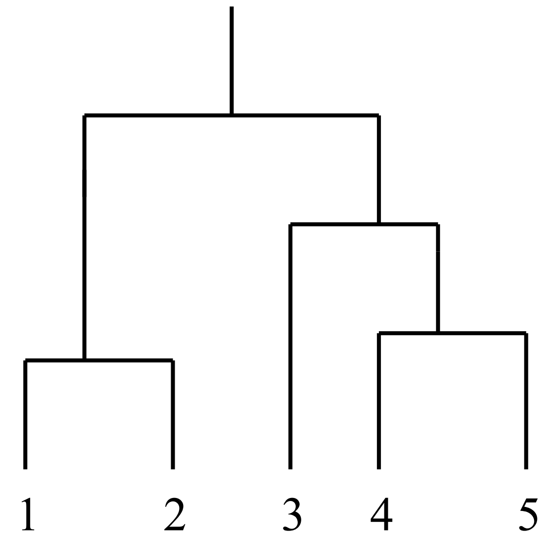
- Similarity of two clusters is based on the two farthest points in the different clusters

	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

Cluster Similarity: MAX or Complete Linkage

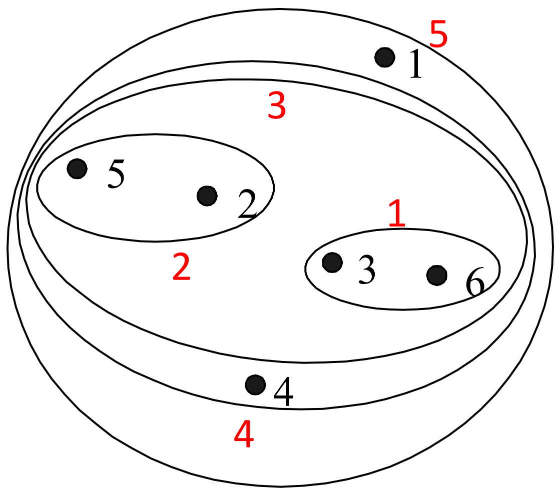
- Similarity of two clusters is based on the two farthest points in the different clusters

	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

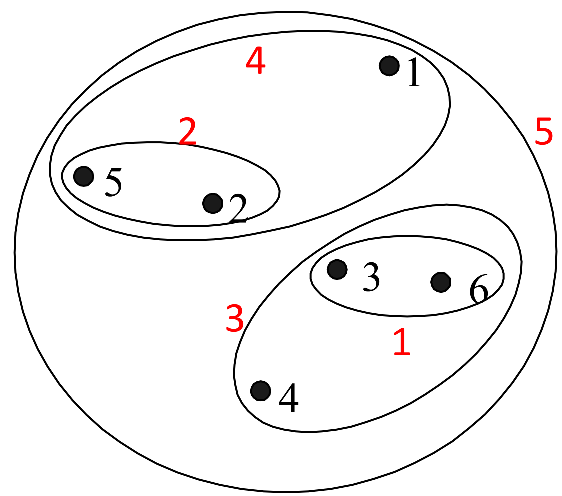


Hierarchical Clustering: Comparison

MIN



MAX



Examples

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Given the data above, perform single link and complete link hierarchical clustering. Draw dendrogram of your results.

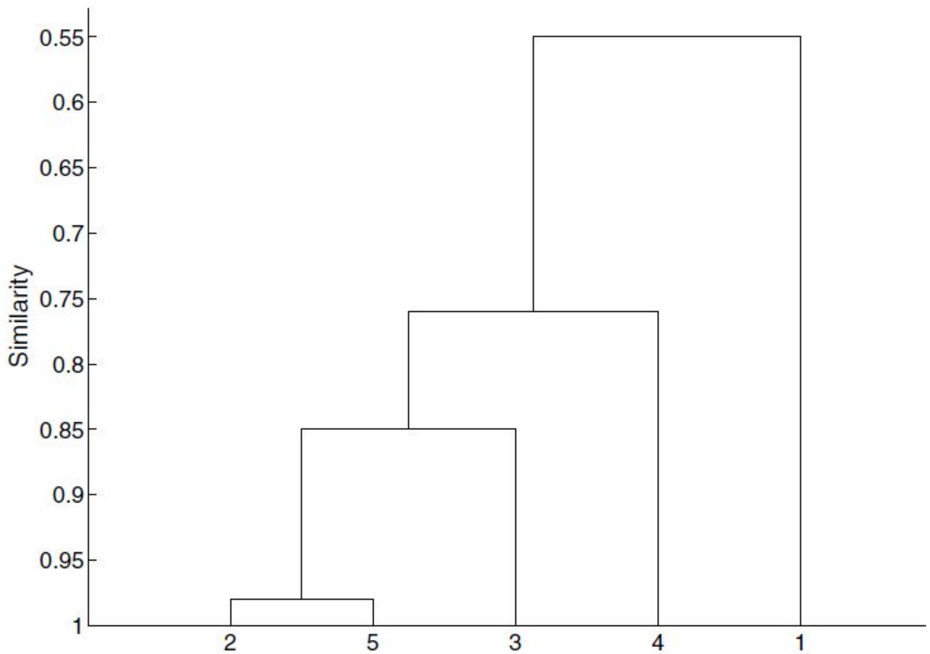
Single Link:

Examples

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Given the data above, perform single link and complete link hierarchical clustering. Draw dendrogram of your results.

Single Link:

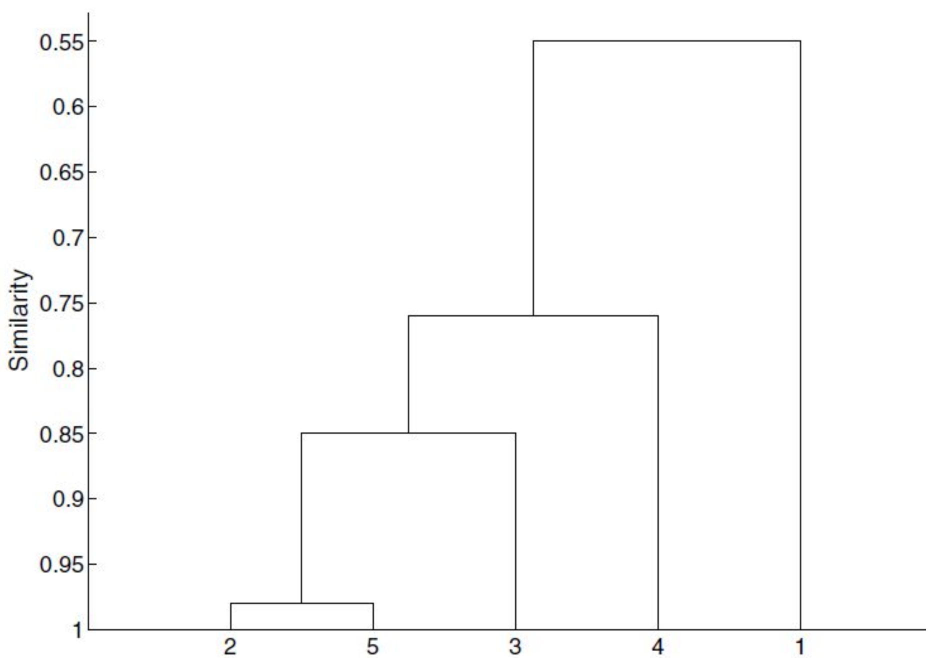


Examples

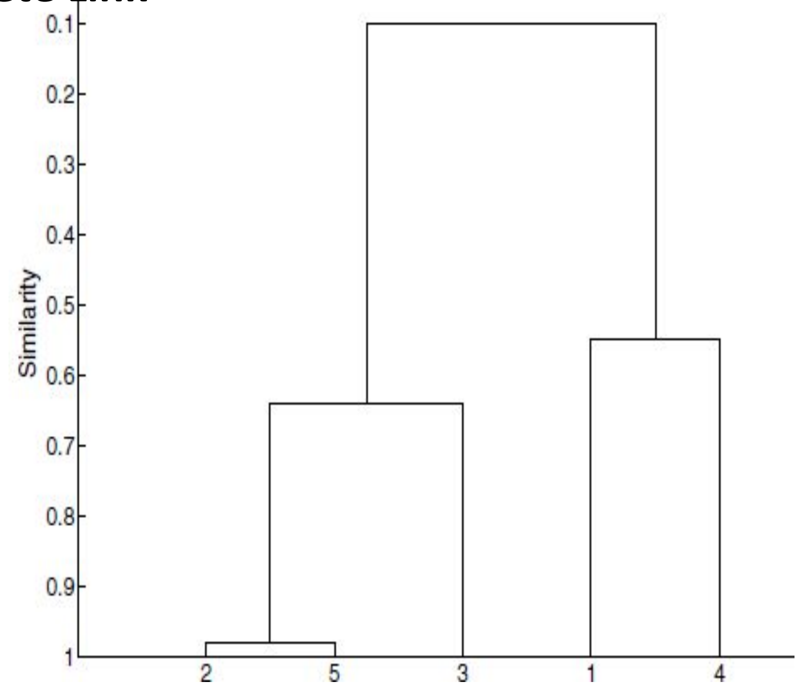
	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Given the data above, perform single link and complete link hierarchical clustering. Draw dendrogram of your results.

Single Link:



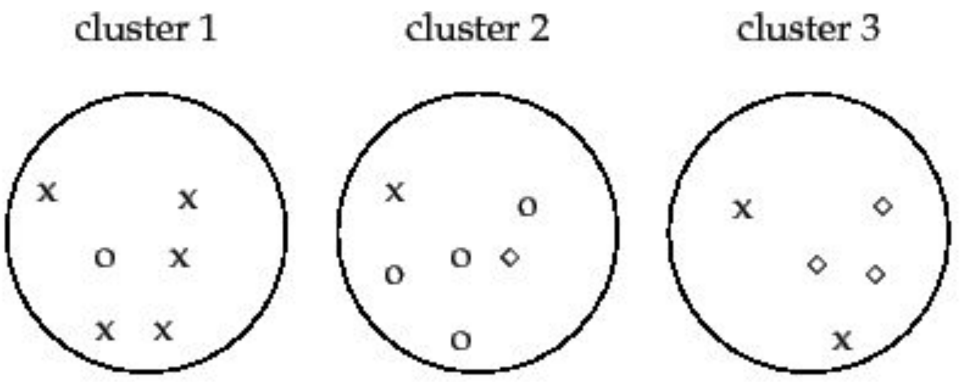
Complete Link



Example - HAC on Iris dataset

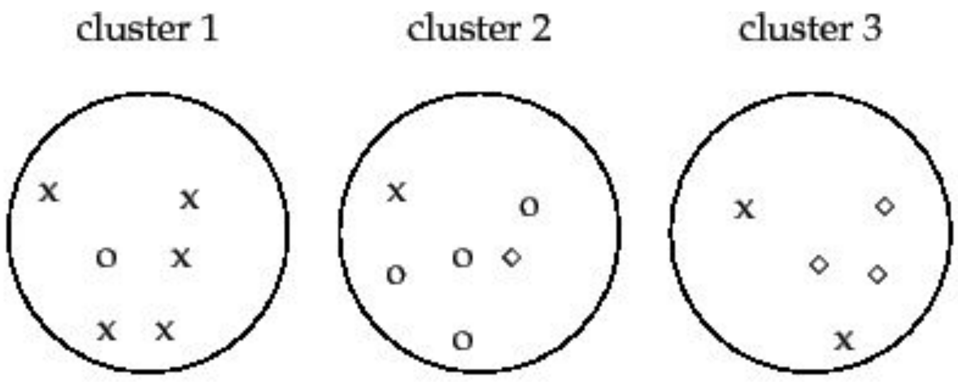
```
from sklearn import datasets
import matplotlib.pyplot as plt
from sklearn.cluster import AgglomerativeClustering
from sklearn import metrics

iris = datasets.load_iris()
X = iris.data
y = iris.target
plt.scatter(X[:,0], X[:,1], c=y, cmap='rainbow', s=10)
plt.title('Actual',fontsize=15, fontweight='bold')
plt.xlabel('Sepal Length',fontsize=15)
plt.ylabel('Petal Length',fontsize=15)
plt.figure()

cls = AgglomerativeClustering(n_clusters = 3, linkage='average')
cls.fit(X)

hac_labels = cls.labels_
print (metrics.silhouette_score(X, hac_labels))
plt.scatter(X[:,0], X[:,1],c=hac_labels, cmap='rainbow', s=10)
plt.xlabel('Sepal Length',fontsize=15)
plt.ylabel('Petal Length',fontsize=15)
plt.title('Predicted clusters',fontsize=15, fontweight='bold')
plt.show()
```

External Evaluation - Purity



What is the purity of the clustering?

External Evaluation - Purity

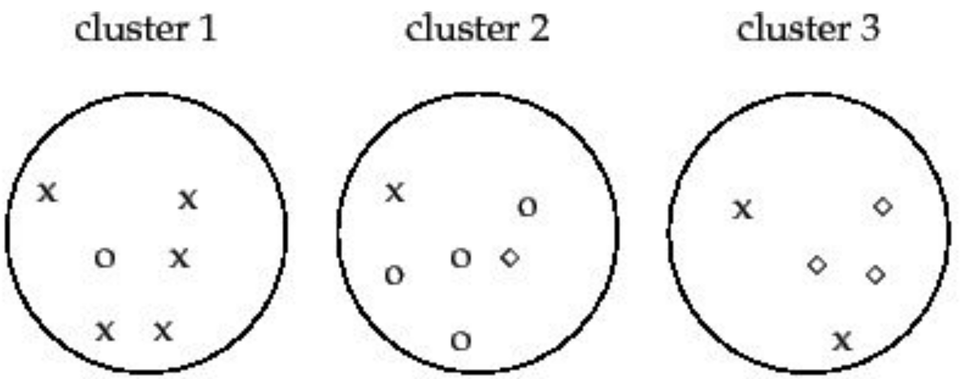


What is the purity of the clustering?

= (5+4+3)/17

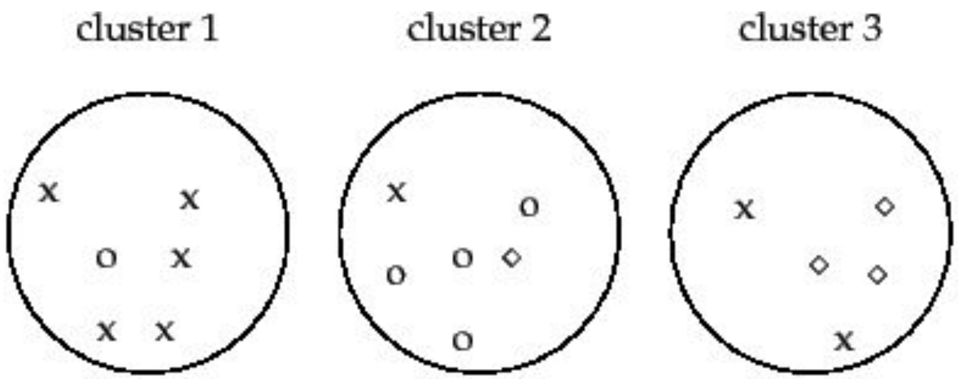
= 0.71

External Evaluation – **Rand Index**



- Look at the example in pairs
- If there are a N examples, then N(N-1)/2 pairs
- A good clustering assigns two similar examples to same cluster, and two dissimilar examples to different clusters. Everything else is bad!
- Let TP be the number of similar pairs assigned to the same cluster, TN be the number of dissimilar pairs assigned to different clusters, FP be the number of dissimilar pairs to same cluster, and FN be the no. of similar pairs assigned to different clusters

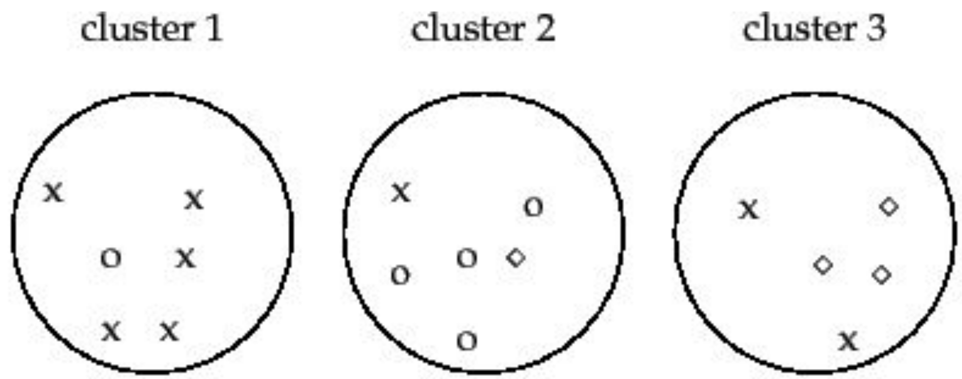
External Evaluation – Rand Index



- Look at the example in pairs
- If there are a N examples, then N(N-1)/2 pairs
- A good clustering assigns two similar examples to same cluster, and two dissimilar examples to different clusters. Everything else is bad!
- Let TP be the number of similar pairs assigned to same cluster, TN be the number of dissimilar pairs assigned to different clusters, FP be the number of dissimilar pairs to same cluster, and FN be the no. of similar pairs assigned to different clusters

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

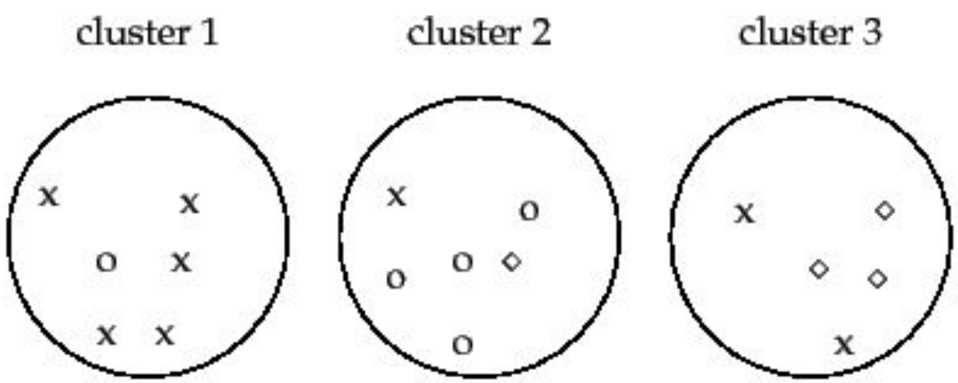
External Evaluation – Rand Index



- Look at the example in pairs
- If there are a N examples, then N(N-1)/2 pairs
- A good clustering assigns two similar examples to same cluster, and two dissimilar examples to different clusters. Everything else is bad!
- Let TP be the number of similar pairs assigned to same cluster, TN be the number of dissimilar pairs assigned to different clusters, FP be the number of dissimilar pairs to same cluster, and FN be the no. of similar pairs assigned to different clusters

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \qquad {}^{N}C_{2}$$

External Evaluation – Rand Index

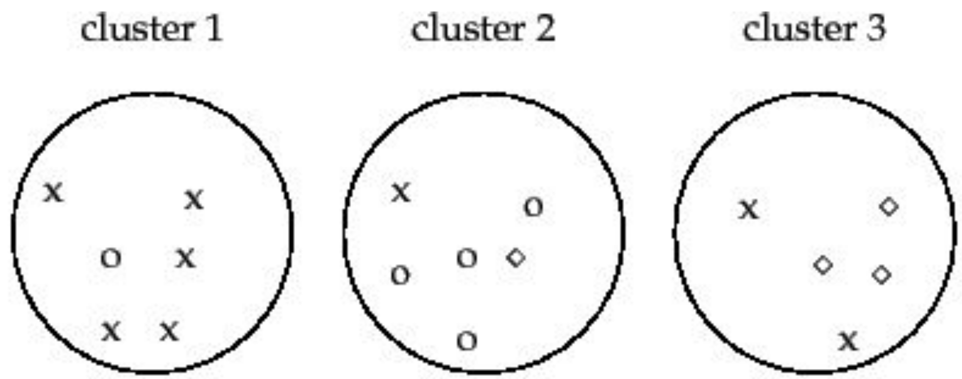


- Look at the example in pairs
- If there are a N examples, then N(N-1)/2 pairs
- A good clustering assigns two similar examples to same cluster, and two dissimilar examples to different clusters. Everything else is bad!
- Let TP be the number of similar pairs assigned to same cluster, TN be the number of dissimilar pairs assigned to different clusters, FP be the number of dissimilar pairs to same cluster, and FN be the no. of similar pairs assigned to different clusters

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

Find TP, TN, FP, FN

External Evaluation – Rand Index

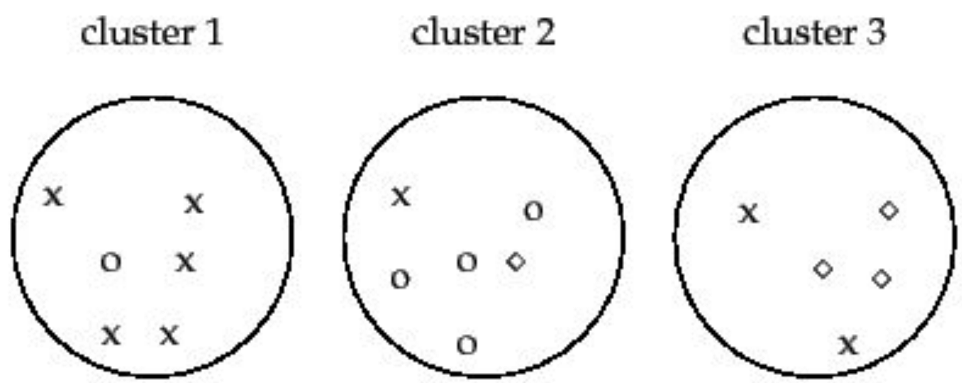


- Look at the example in pairs
- If there are a N examples, then N(N-1)/2 pairs
- A good clustering assigns two similar examples to same cluster, and two dissimilar examples to different clusters. Everything else is bad!
- Let TP be the number of similar pairs assigned to same cluster, TN be the number of dissimilar pairs assigned to different clusters, FP be the number of dissimilar pairs to same cluster, and FN be the no. of similar pairs assigned to different clusters

	Same cluster	Diff. clusters
Same class	20	24
Diff. class	20	72

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

External Evaluation – Rand Index



- Look at the example in pairs
- If there are a N examples, then N(N-1)/2 pairs
- A good clustering assigns two similar examples to same cluster, and two dissimilar examples to different clusters. Everything else is bad!
- Let TP be the number of similar pairs assigned to same cluster, TN be the number of dissimilar pairs assigned to different clusters, FP be the number of dissimilar pairs to same cluster, and FN be the no. of similar pairs assigned to different clusters

	Same cluster	Diff. clusters
Same class	20	24
Diff. class	20	72

RI = 0.68

Thank You!