Applications of Clustering

- Market research
- Pattern recognition
- Data analysis
- Image processing
- Taxonomy generation
- Gene expression analysis
- Event detection
- …

- Outlier detection:
  - Network security (intrusions)
  - Credit card fraud detection
- Preprocessing:
  - Retrieval
  - Feature selection
  - Approximation & summarization
  - Classification

Common Similarity Measures

- Interval-scaled vectors:
  - Euclidean distance.

$$d(x,y) = \|x-y\|_2 = \sqrt{\sum_{j=1}^{n} \left(x_j - y_j\right)^2}$$

  - Manhattan (L_1) distance.

$$d(x,y) = \|x-y\|_1 = \sum_{j=1}^{n} \left|x_j - y_j\right|$$

- Interval-scaled vectors (continued):
  - Cosine measure (<u>not</u> a metric!).
    document clustering

$$s(x,y) = \frac{x \cdot y}{\|x\|\,\|y\|} = \frac{\sum_{j=1}^{n} x_j y_j}{\sqrt{\sum_{j=1}^{n} x_j^2}\,\sqrt{\sum_{j=1}^{n} y_j^2}}$$

Clustering Strategies

- Partitional (Centroid Based) clustering:
  - Given: target number of clusters *k*.
  - Goal: partition data set into exactly *k* clusters.
  - Each object must appear in exactly one cluster.
- (Connectivity Based) Hierarchical clustering:
  - Clustering formed by composition or decomposition.
  - History of composition / decomposition operations forms a hierarchical relationship.
- Agglomerative (bottom-up) approach:
  - Larger clusters formed by merging smaller clusters.
  - Usually terminates when all clusters merged (but earlier termination is possible).
- Divisive (top-down) approach:
  - Smaller clusters formed by splitting larger clusters.
  - Often terminates when leaf clusters contain exactly one element (but earlier termination is possible).

Clustering Strategies

- Density-based clustering:
  - Clusters grow into regions of high density.
  - Density usually computed over neighbourhoods of fixed size.
  - Connectivity constraints can be similar to those of agglomerative clustering.
  - Local criteria for growth $\rightarrow$ non-spherical clusters.
  - Minimum density criterion $\rightarrow$ noise & outlier elimination.
- (Distribution) Model-based clustering:
  - Guess a model explaining the data distribution.
  - Find the best fit of data to clusters as explained by the model.
  - Can lead to automatic determination of number of clusters.
  - Determination of noise & outliers according to the model.
  - Sometimes confused with classification when the model is learned from a training set.

Hierarchical Methods

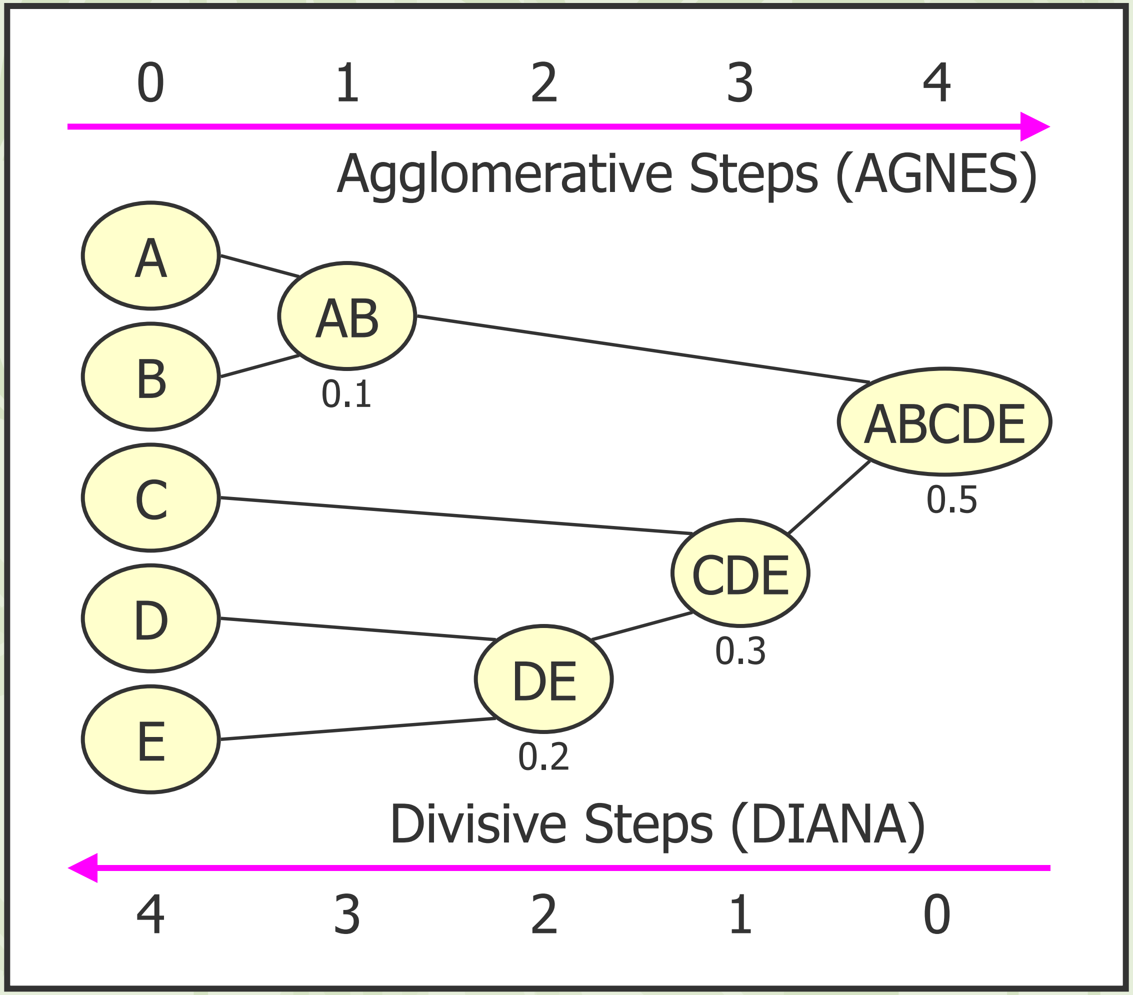
Agglomerative vs Divisive

- Agglomerative (bottom-up) approach:
  - Basic method: AGNES (AGlomerative NEsting), Kaufman & Rousseeuw, 1990.
  - Initially, each object in its own cluster.
  - At each step, two clusters are merged.
  - Choice of clusters according to distance criterion.
- Divisive (top-down) approach:
  - Basic method: DIANA (DIvisive ANAlysis), Kaufman & Rousseeuw, 1990.
  - Initially, all objects in a single cluster.
  - At each step, a cluster is split into two.
  - Choice of cluster according to a distance criterion between the two clusters generated by the split.

Dendrogram

- Dendrogram:
  - Tree structure describing merge / split history.
  - This example: split / merge according to closest pair of cluster members.
  - "Single-linkage" strategy.

$d(*,*)$	A	B	C	D	E
A	0	0.1	0.8	0.7	1.0
B	0.1	0	0.5	0.6	0.9
C	0.8	0.5	0	0.3	0.4
D	0.7	0.6	0.3	0	0.2
E	1.0	0.9	0.4	0.2	0



Inter-Cluster Distance

- Common measures:
  - Minimum distance (single linkage).

$$d_{\min}(A,B) = \min_{a \in A;\, b \in B} d(a,b)$$

  - Maximum distance (complete linkage).

$$d_{\max}(A,B) = \max_{a \in A;\, b \in B} d(a,b)$$

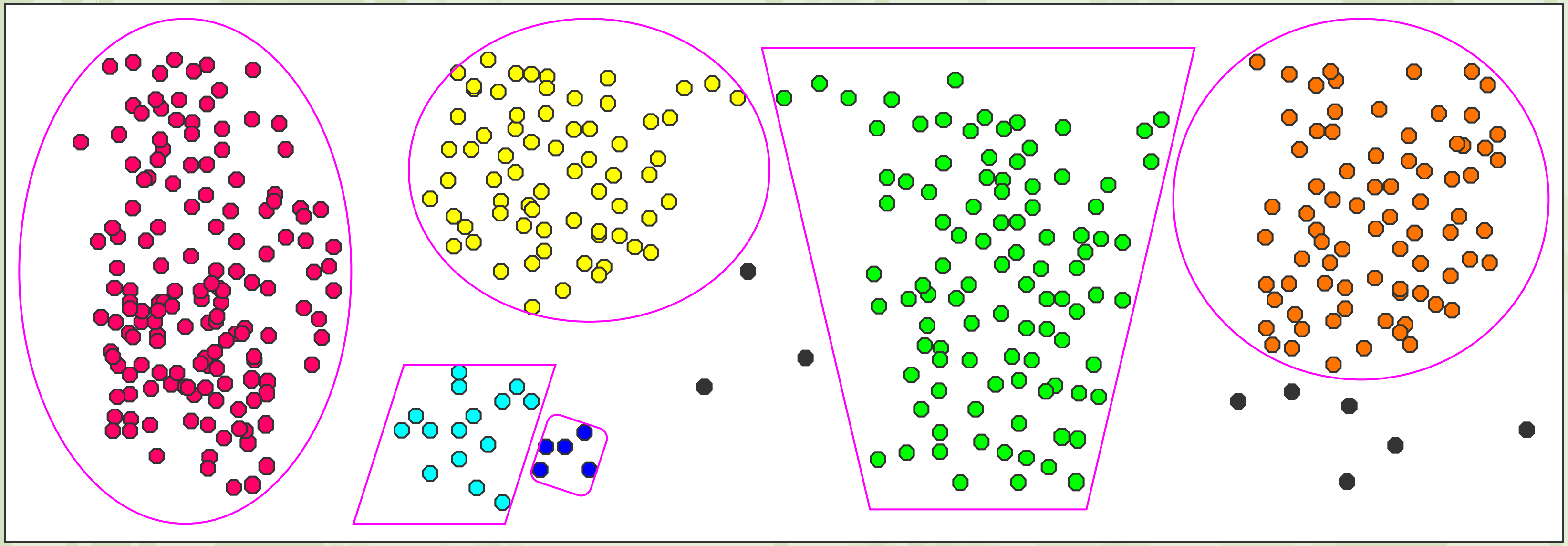
  - Average distance.

$$d_{\text{avg}}(A,B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

Merge / Split Strategies

- Single Linkage:
  - Also called Nearest Neighbour.
  - Mininum-distance measure.
  - Links determined by only two closest objects.
  - Repeated merges can lead to chaining.
  - Excessive chaining can produce incoherent clusters.

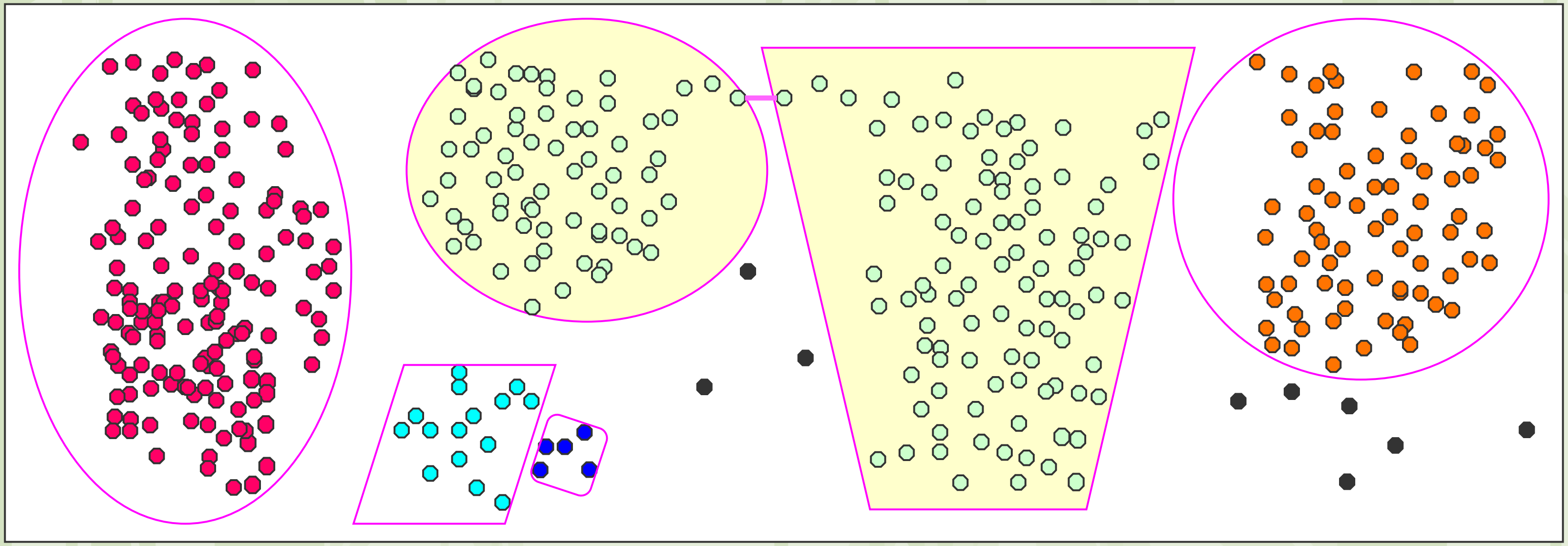
$$d_{\min}(A,B) = \min_{a \in A;\, b \in B} d(a,b)$$



Merge / Split Strategies

- Single Linkage:
  - Also called Nearest Neighbour.
  - Mininum-distance measure.
  - Links determined by only two closest objects.
  - Repeated merges can lead to chaining.
  - Excessive chaining can produce incoherent clusters.

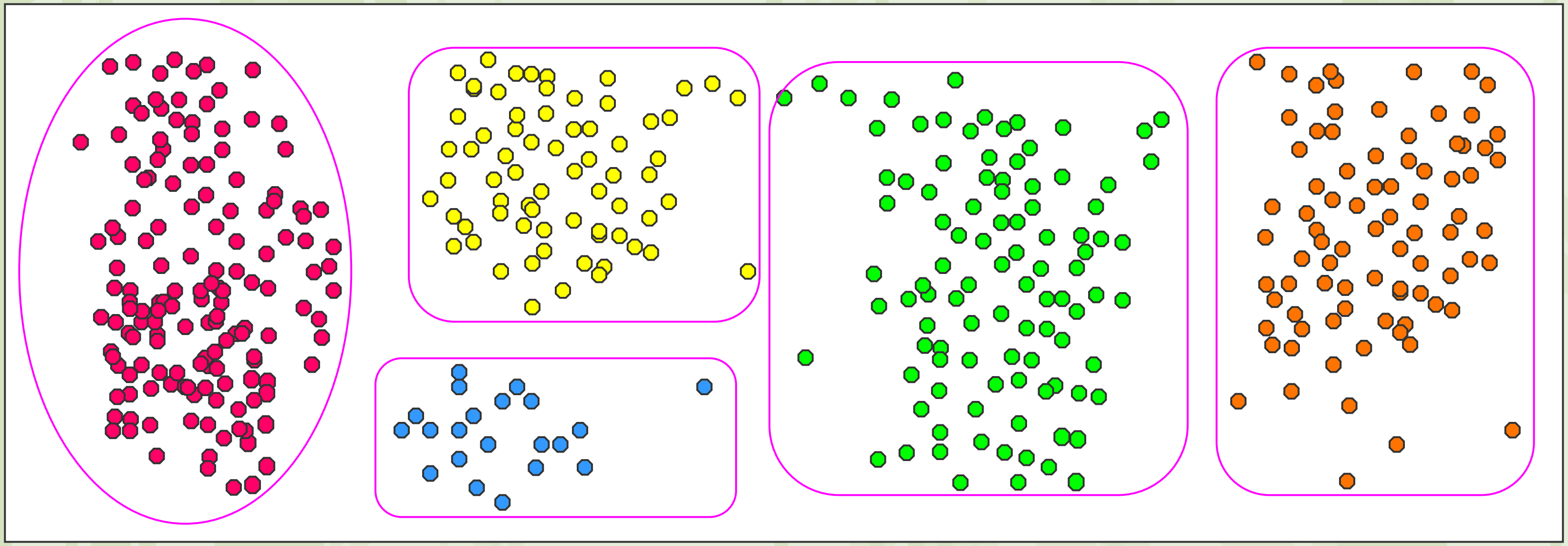
$$d_{\min}(A,B) = \min_{a \in A;\, b \in B} d(a,b)$$



Merge / Split Strategies

- Complete Linkage:
  - Also called Farthest Neighbour.
  - Maximum-distance measure.
  - Links determined by only two farthest objects.
  - Merge order highly influenced by noise.
  - Clusters produced are more rounded, compact.

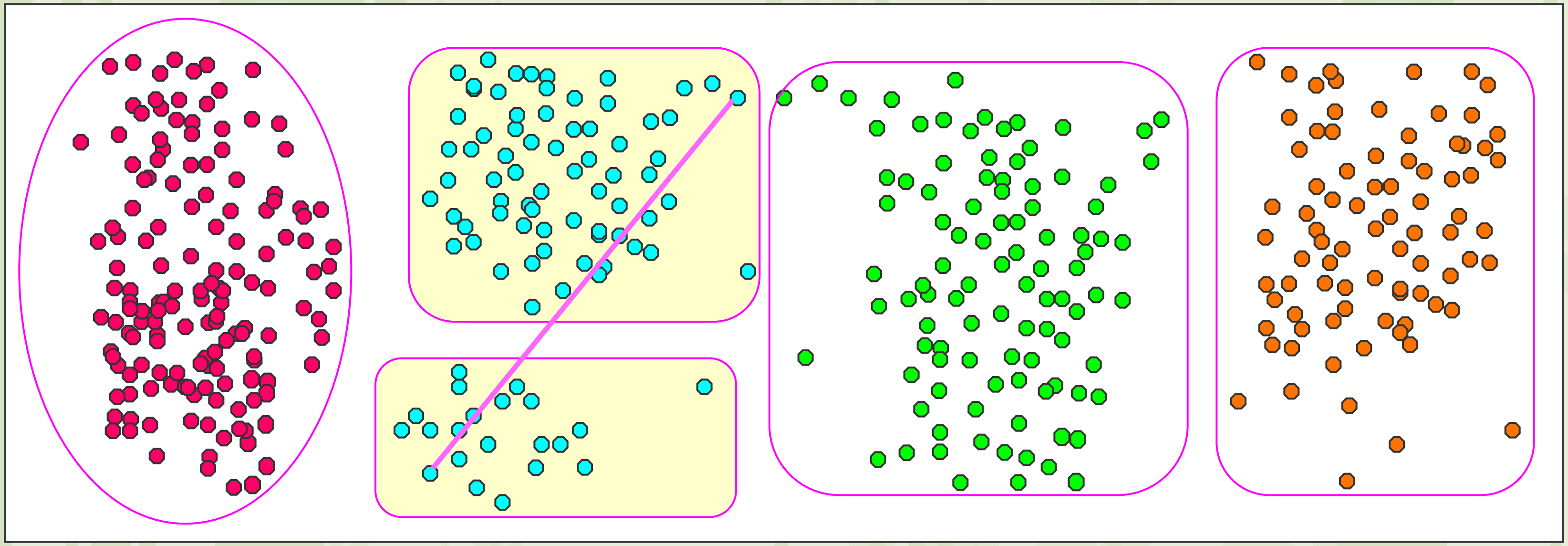
$$d_{\max}(A,B) = \max_{a \in A;\, b \in B} d(a,b)$$



Merge / Split Strategies

- Complete Linkage:
  - Also called Farthest Neighbour.
  - Maximum-distance measure.
  - Links determined by only two farthest objects.
  - Merge order highly influenced by noise.
  - Clusters produced are more rounded, compact.

$$d_{\max}(A,B) = \max_{a \in A;\, b \in B} d(a,b)$$



Merge / Split Strategies

- Average Linkage:
  - Compromise between minimum and maximum distance.
  - Quadratic number of distances computed.
  - Less affected by noise.
  - Less prone to chaining problems.

$$d_{\text{avg}}(A,B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a,b)$$

Example (1)

- Using the single linkage method

	A	B	C	D	E
A	**0**	**0.1**	**0.8**	**0.7**	**1.0**
B	**0.1**	**0**	**0.5**	**0.6**	**0.9**
C	**0.8**	**0.5**	**0**	**0.3**	**0.4**
D	**0.7**	**0.6**	**0.3**	**0**	**0.2**
E	**1.0**	**0.9**	**0.4**	**0.2**	**0**

$d(*,*)$	A	B	C	D	E
A	0	0.1	0.8	0.7	1.0
B	0.1	0	0.5	0.6	0.9
C	0.8	0.5	0	0.3	0.4
D	0.7	0.6	0.3	0	0.2
E	1.0	0.9	0.4	0.2	0

Example (1)

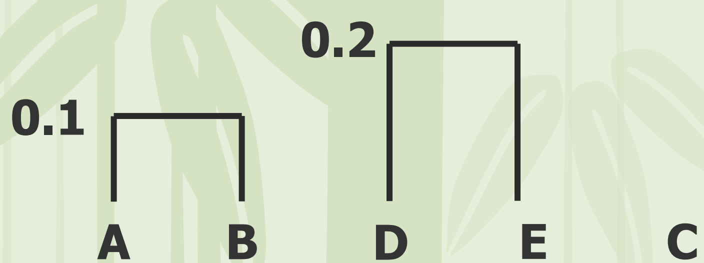
	A	B	C	D	E
A	0				
B	0.1	0			
C	0.8	0.5	0		
D	0.7	0.6	0.3	0	
E	1.0	0.9	0.4	0.2	0

	A,B	C	D	E
A,B	0			
C		0		
D			0	
E				0

Example (1)

	A,B	C	D	E
A,B	0			
C	0.5	0		
D	0.6	0.3	0	
E	0.9	0.4	0.2	0

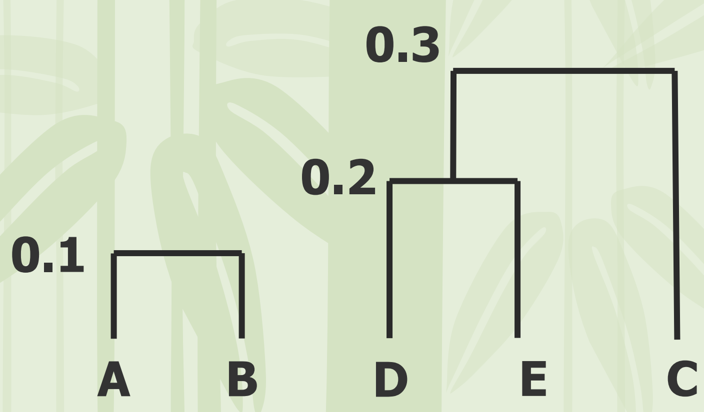
	A,B	C	D,E
A,B	0		
C		0	
D,E			0



Example (1)

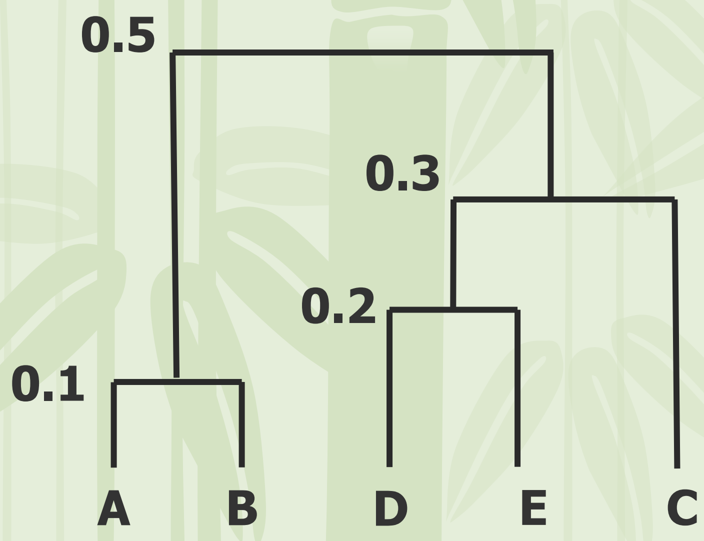
	A,B	C	D,E
A,B	0		
C	0.5	0	
D,E	0.6	0.3	0

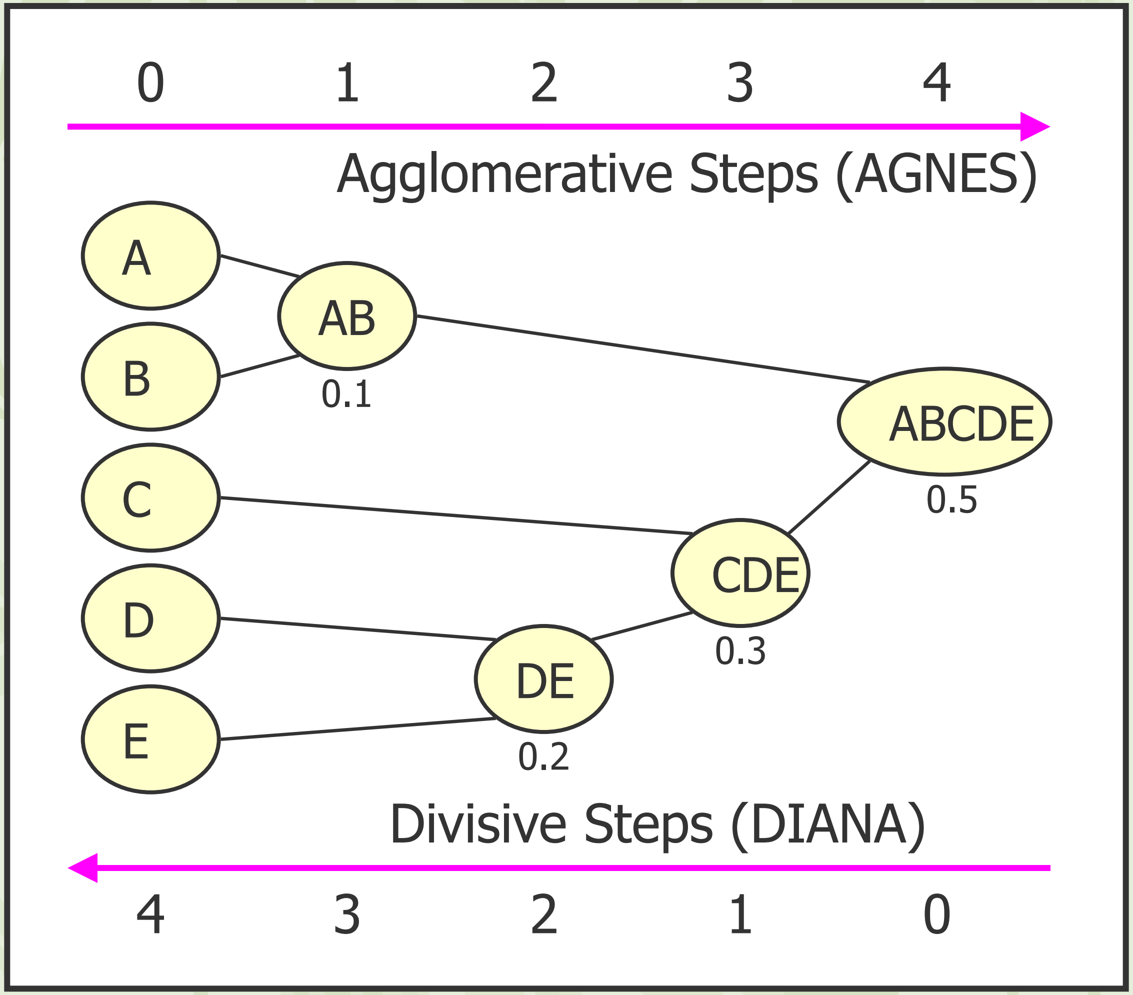
	A,B	C,D,E
A,B	0	
C,D,E	0.5	0



Example (1)

	A,B	C,D,E
A,B	0	
C,D,E	0.5	0





Example (2)

- Using the single linkage method

	X	Y
P1	0.40	0.53
P2	0.22	0.38
P3	0.35	0.32
P4	0.26	0.19
P5	0.08	0.41
P6	0.45	0.30

Example (2)

- Create the distance matrix, in this example we use the Euclidean distance as measure of distance.

	P1	P2	P3	P4	P5	P6
P1	0	0.23	0.22	0.37	0.34	0.23
P2	0.23	0	0.15	0.20	0.14	0.25
P3	0.22	0.15	0	0.15	0.28	0.11
P4	0.37	0.20	0.15	0	0.29	0.22
P5	0.34	0.14	0.28	0.29	0	0.39
P6	0.23	0.25	0.11	0.22	0.39	0

Example (2)

- Create the distance matrix, in this example we use the Euclidean distance as measure of distance.

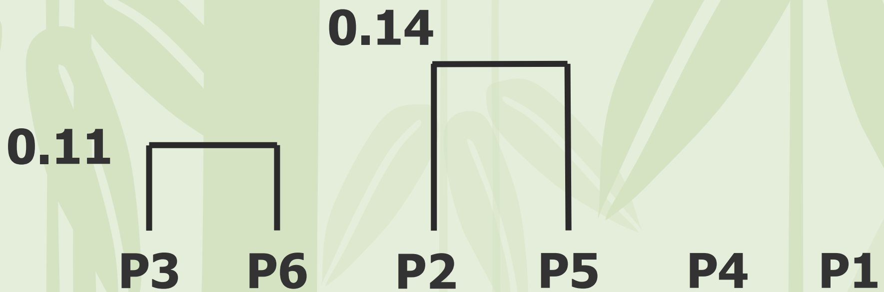
	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

	P1	P2	P3, P6	P4	P5
P1	0				
P2		0			
P3, P6			0		
P4				0	
P5					0

Example (2)

	P1	P2	P3, P6	P4	P5
P1	0				
P2	0.23	0			
P3, P6	0.22	0.15	0		
P4	0.37	0.20	0.15	0	
P5	0.34	0.14	0.28	0.29	0

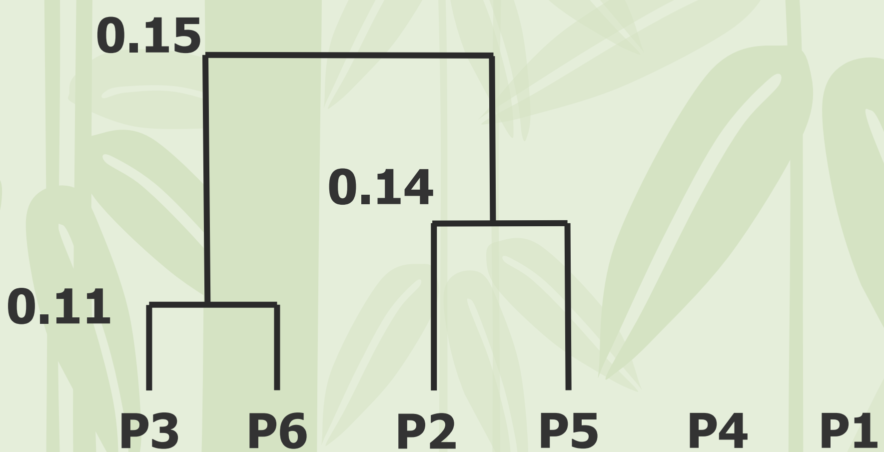
	P1	P2, P5	P3, P6	P4
P1	0			
P2, P5		0		
P3, P6			0	
P4				0



Example (2)

	P1	P2, P5	P3, P6	P4
P1	0			
P2, P5	0.23	0		
P3, P6	0.22	0.15	0	
P4	0.37	0.20	0.15	0

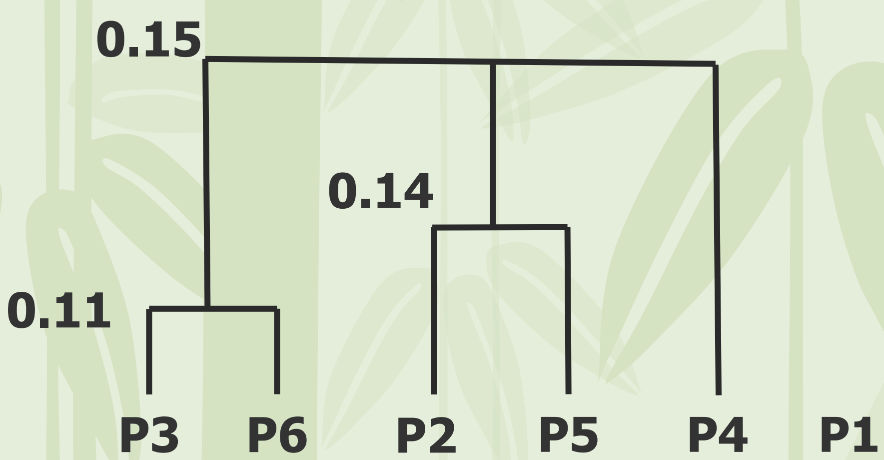
	P1	P2, P5, P3, P6	P4
P1	0		
P2, P5, P3, P6		0	
P4			0



Example (2)

	P1	P2, P5, P3, P6	P4
P1	0		
P2, P5, P3, P6	0.22	0	
P4	0.37	0.15	0

	P1	P2, P5, P3, P6, P4
P1	0	
P2, P5, P3, P6, P4		0



Example (2)

	P1	P2, P5, P3, P6, P4
P1	0	
P2, P5, P3, P6, P4	0.22	0