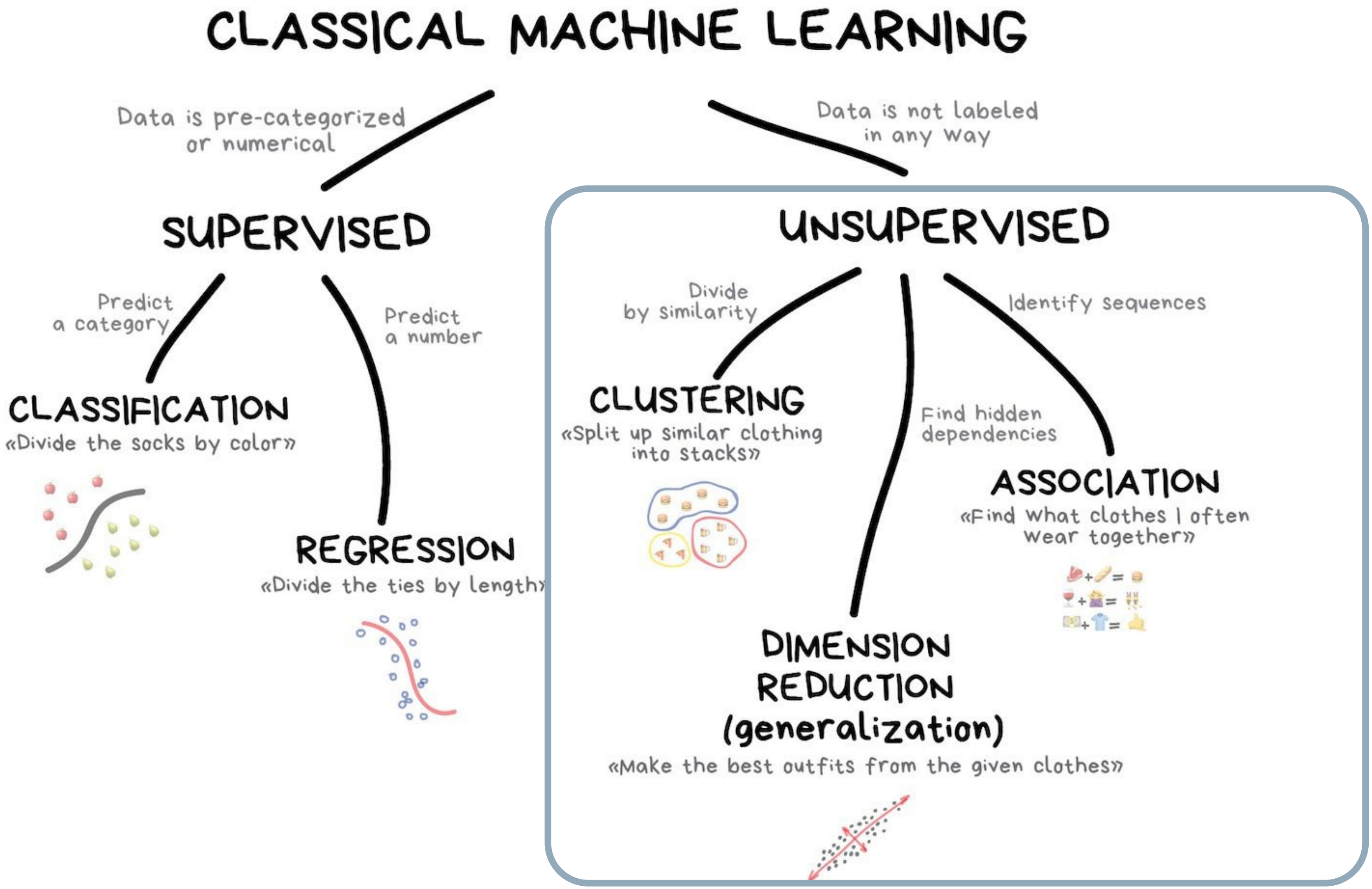
# CLUSTERING

Week11

# Type of Learning

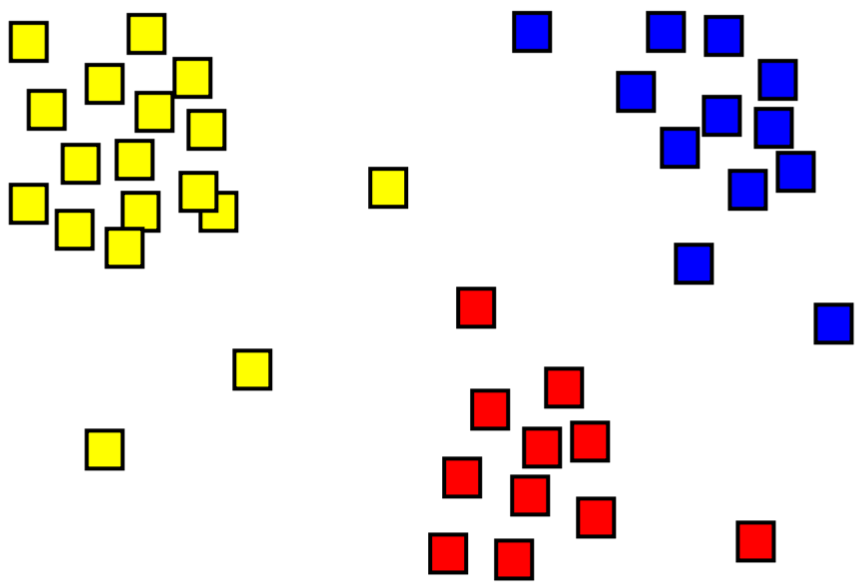


# Clustering

# Unsupervised Learning: Clustering

- [Remind] Unsupervised learning is learning with unlabeled data
  - No certain output to be estimated

- Clustering is to group a set of data points to satisfy following conditions as much as possible
  - Data points in the same group are more similar to each other than to data points in other groups



# Clustering

- Data points in the same group are $\boxed{\text{more similar}}$ to each other than to data points in other groups

**1. How to know**
**some points are more similar than others?**

**⬇**

**Using distance measure**

**2. How to group?**

**⬇**

**Determine certain rule to group**

# *k*-means Clustering

# $k$-means Clustering

- Objective function of clustering

$$\sum_i \min_j \left\| \mathbf{x}_i - \boldsymbol{\mu}_j \right\|^2$$

  - $j \in [1,2,\ldots,k]$
  - $\boldsymbol{\mu}_j$ is the centroid of j-th cluster

⬇

**Combinatorial Optimization Problem**

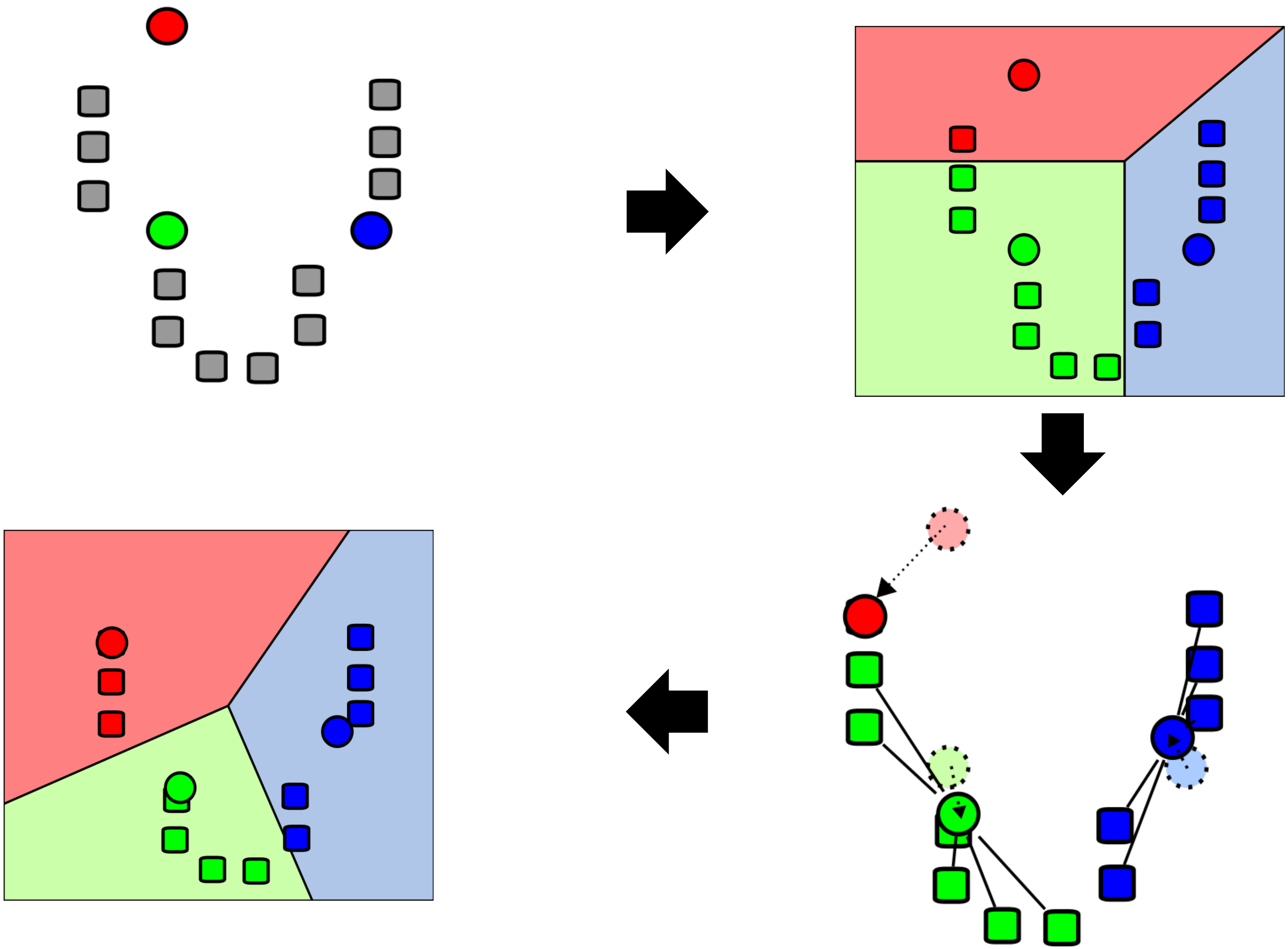
# $k$-means Clustering

- Procedure of $k$-means clustering

1. Set initial $k$ centroids
2. Assign each data point to $i$-th group whose centroid is the closest to data point among all centroids
3. Update centroids
4. Check terminal condition and if condition is not fulfilled, go to step 2

**Terminal conditions**
**[No change in centroids or the number of iteration is over the pre-specified threshold]**

# $k$-means Clustering



# How to Update Centroids

- Arithmetic mean

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

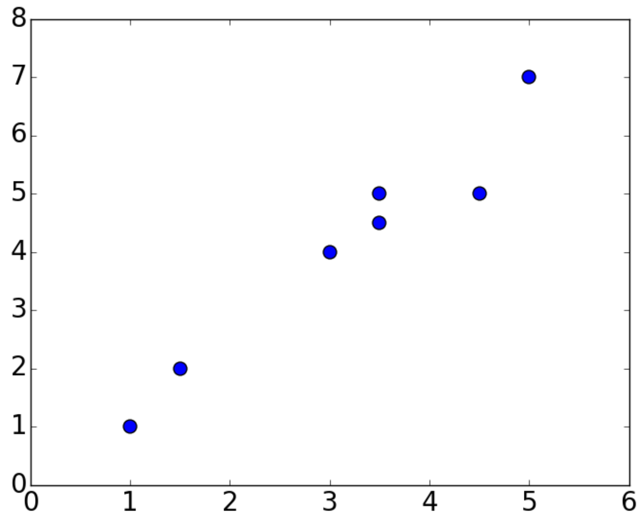
  - $t$ is iteration
  - $m_i$ is $i$-th group centroid
  - $S_i$ is a set of $i$-th group and $|S_i|$ is size of $S_i$

- Example
  - If (3,1), (2,2), (4,6) belong to group, updated centroid is

$$\left(\frac{3+2+4}{3}, \frac{1+2+6}{3}\right) = (3, 2)$$

# Question

- Clustering for 2D data set




| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $x$ | 1.0 | 1.5 | 3.0 | 5.0 | 3.5 | 4.5 | 3.5 |
| $y$ | 1.0 | 2.0 | 4.0 | 7.0 | 5.0 | 5.0 | 4.5 |

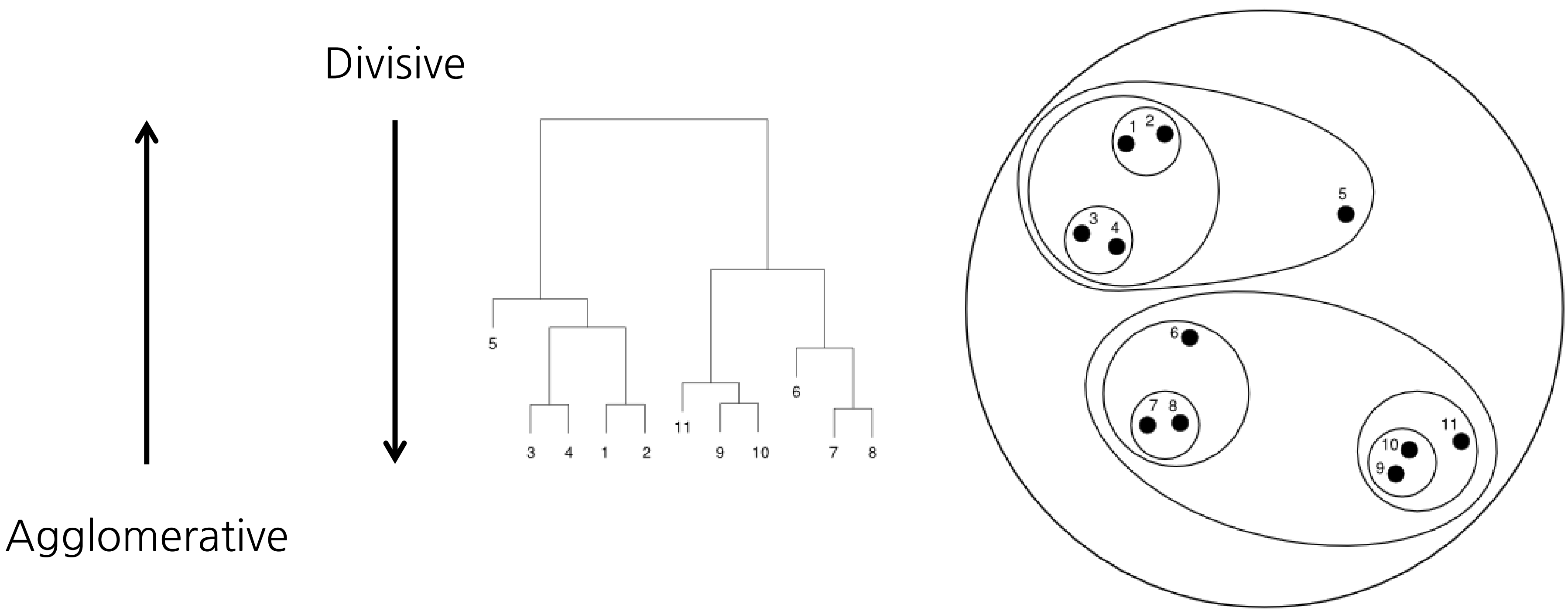
1) When $k = 2$ and initial centroids are (1.0, 1.0) and (5.0, 7.0), determine group of each data point

2) What are new centroids of two groups?

# Hierarchical Clustering

# Hierarchical Clustering

- Hierarchical clustering builds a hierarchy of clusters
  - Agglomerative: Bottom up approach, each data point starts in its own cluster and pairs of clusters are merged as one moves up the hierarchy
  - Divisive: Top down approach, all data points start in one cluster and splits are performed recursively as one moves down the hierarchy



# Linkage Criteria for Agglomerative Clustering

- Way to calculate similarity between two clusters $A, B$

| Type | Formula |
|---|---|
| Complete-linkage | $\max\{d(a,b): a \in A, b \in B\}$ |
| Single-linkage | $\min\{d(a,b): a \in A, b \in B\}$ |
| Mean linkage | $\frac{1}{\lvert A \rvert \lvert B \rvert} \sum_{a \in A} \sum_{b \in B} d(a,b)$ |
| Centroid linkage | $d(c_A, c_B)$ |
| Ward linkage | $\mathrm{Var}(A \cup B) - \mathrm{Var}(A) - \mathrm{Var}(B)$ |

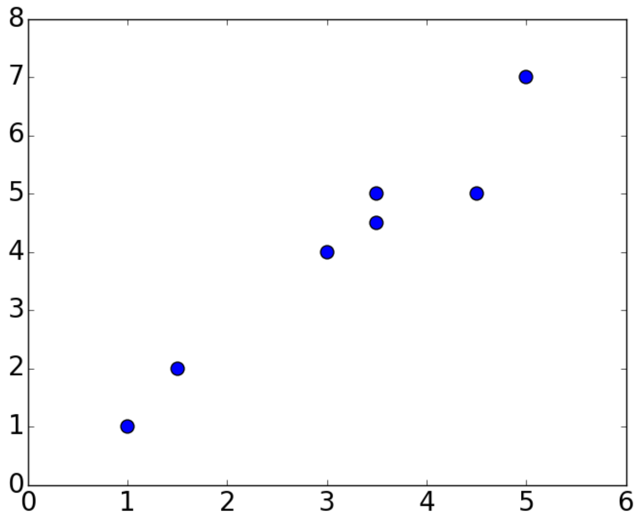
- $a$ belongs to $A$, $b$ belongs to $B$
- $\mathrm{Var}(X)$ is within-cluster variance (variance of cluster $X$)

$$Var(X) = \frac{1}{n_A} \sum_{i \in A} \lVert \mathbf{x}_i - \mu_A \rVert^2$$

- $d(a,b)$ is distance between two data points $a$ and $b$

# Question

- Clustering for 2D data set



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| $x$ | 1.0 | 1.5 | 3.0 | 5.0 | 3.5 | 4.5 | 3.5 |
| $y$ | 1.0 | 2.0 | 4.0 | 7.0 | 5.0 | 5.0 | 4.5 |
| $C$ | 1 | 1 | 2 | 2 | 2 | 2 | 2 |

1) Using complete-linkage, calculate linkage criterion of cluster 1 and 2

2) Using centroid-linkage, calculate linkage criterion of cluster 1 and 2

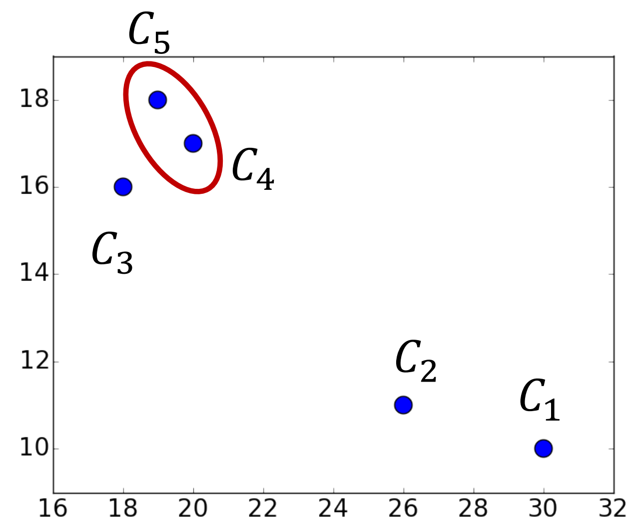
# Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
  - Start each data as own cluster
  - Distance measure between two points: Euclidean distance

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Distance $d(A, B) = \min\{d(a, b): a \in A, b \in B\}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 4.12 | 0 | | | |
| 3 | 15.23 | 11.18 | 0 | | |
| 4 | 12.21 | 8.48 | 4.12 | 0 | |
| 5 | 13.60 | 9.90 | 3.61 | **1.41** | 0 |



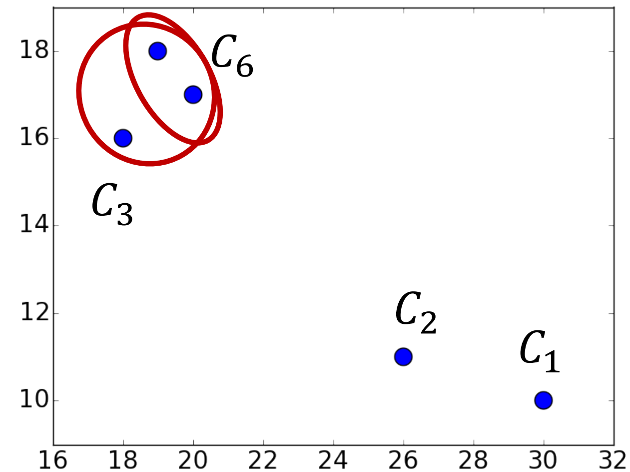
# Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
  - Merge cluster 4 and 5 to create new cluster

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Distance $d(A,B) = \min\{d(a,b): a \in A, b \in B\}$

| | 1 | 2 | 3 | 6 |
|---|---|---|---|---|
| 1 | 0 | | | |
| 2 | 4.12 | 0 | | |
| 3 | 15.23 | 11.18 | 0 | |
| 6 | 12.21 | 8.48 | **3.61** | 0 |



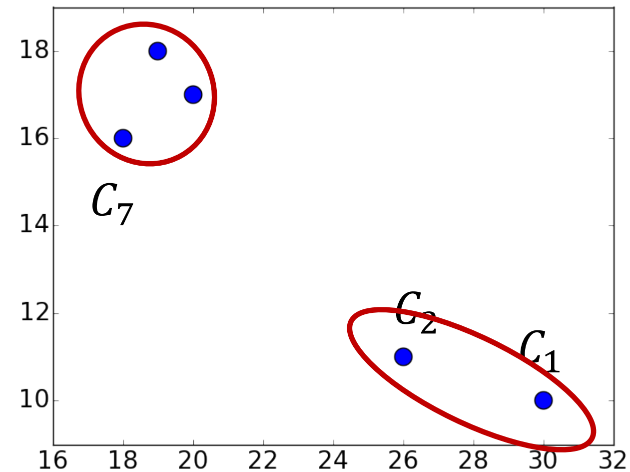
# Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
  - Merge cluster 3 and 6 to create new cluster

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Distance $d(A,B) = \min\{d(a,b): a \in A, b \in B\}$

| | 1 | 2 | 7 |
|---|---|---|---|
| 1 | 0 | | |
| 2 | **4.12** | 0 | |
| 7 | 12.21 | 8.48 | 0 |



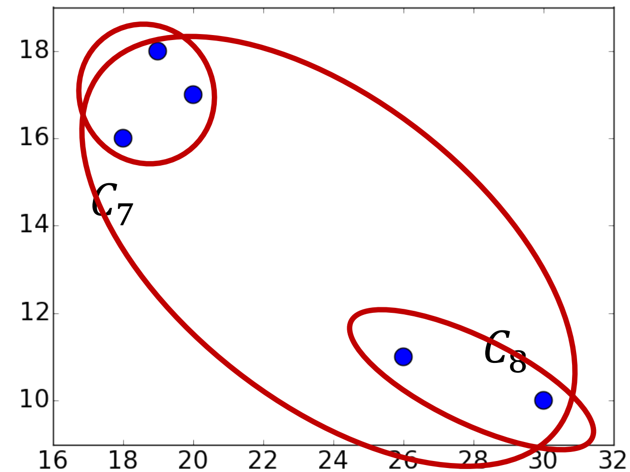
# Example: Single Linkage Clustering

- Find clusters through single linkage hierarchy clustering
  - Merge cluster 1 and 2 to create new cluster

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

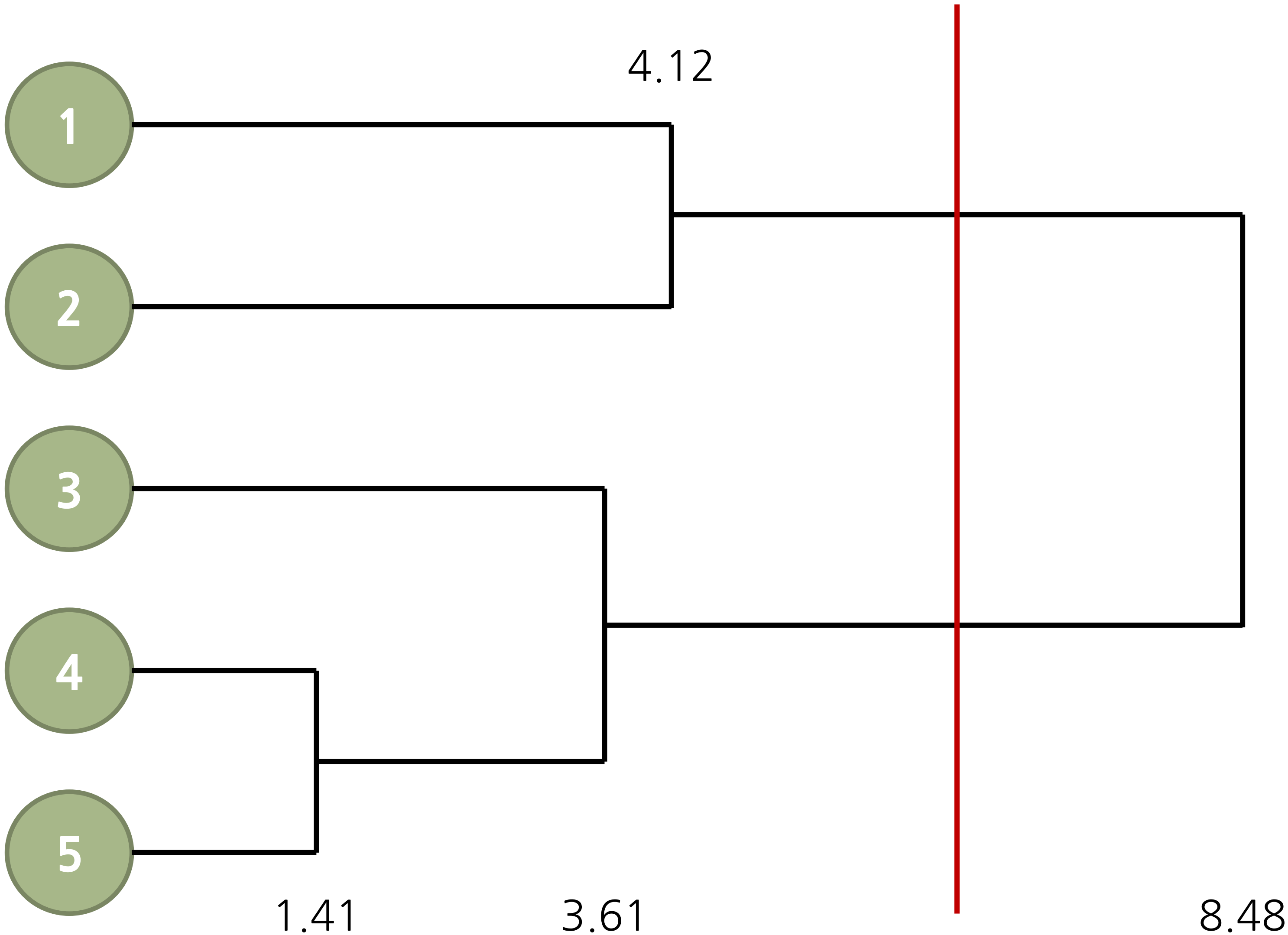
Distance $d(A, B) = \min\{d(a, b): a \in A, b \in B\}$

| | 7 | 8 |
|---|---|---|
| 7 | 0 | |
| 8 | 8.48 | 0 |



# Example: Single Linkage Clustering

- Dendrogram



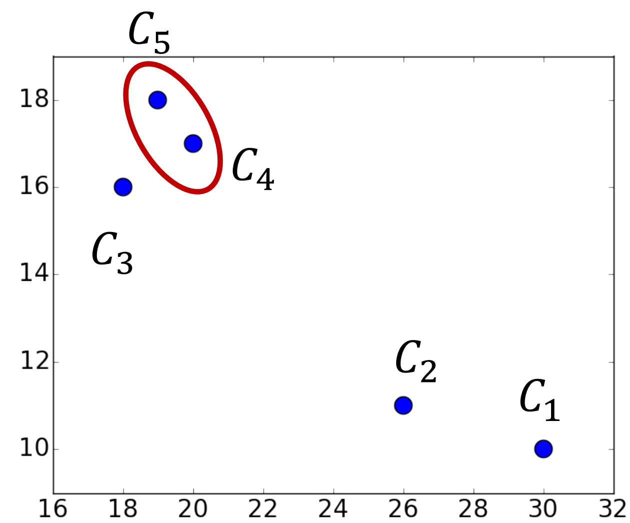
# Example: Complete Linkage Clustering

- Find clusters through complete linkage hierarchy clustering
  - Start each data as own cluster

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Distance $d(A, B) = \max\{d(a, b) : a \in A, b \in B\}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 4.12 | 0 | | | |
| 3 | 15.23 | 11.18 | 0 | | |
| 4 | 12.21 | 8.48 | 4.12 | 0 | |
| 5 | 13.60 | 9.90 | 3.61 | **1.41** | 0 |



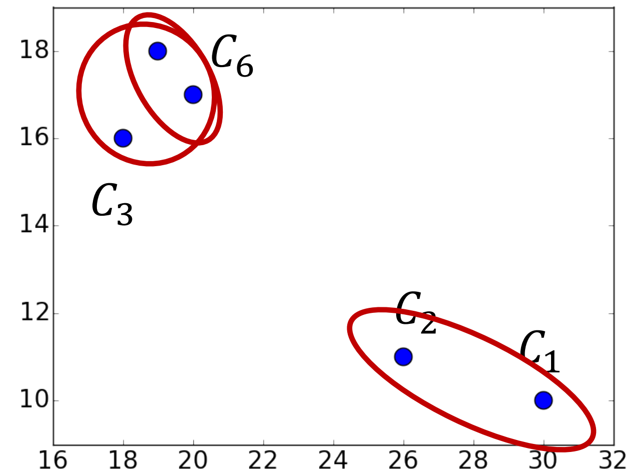
# Example: Complete Linkage Clustering

- Find clusters through complete linkage hierarchy clustering
  - Merge cluster 4 and 5 to create new cluster

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Distance $d(A,B) = \max\{d(a,b): a \in A, b \in B\}$

| | 1 | 2 | 3 | 6 |
|---|---|---|---|---|
| 1 | 0 | | | |
| 2 | **4.12** | 0 | | |
| 3 | 15.23 | 11.18 | 0 | |
| 6 | 13.60 | 9.90 | **4.12** | 0 |



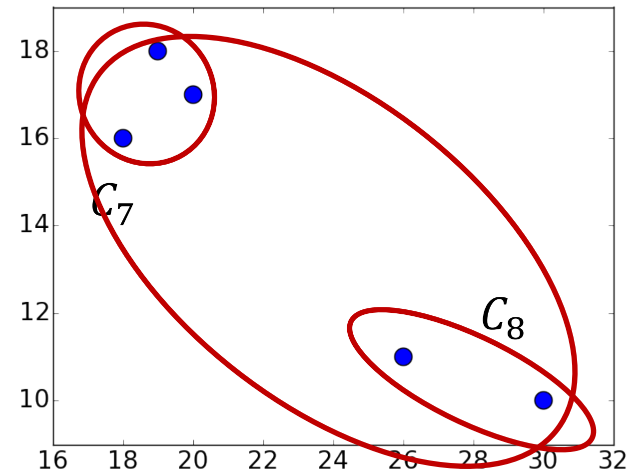
# Example: Complete Linkage Clustering

- Find clusters through complete linkage hierarchy clustering
  - Merge cluster 1 and 2 to create new cluster
  - Merge cluster 3 and 6 to create new cluster

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

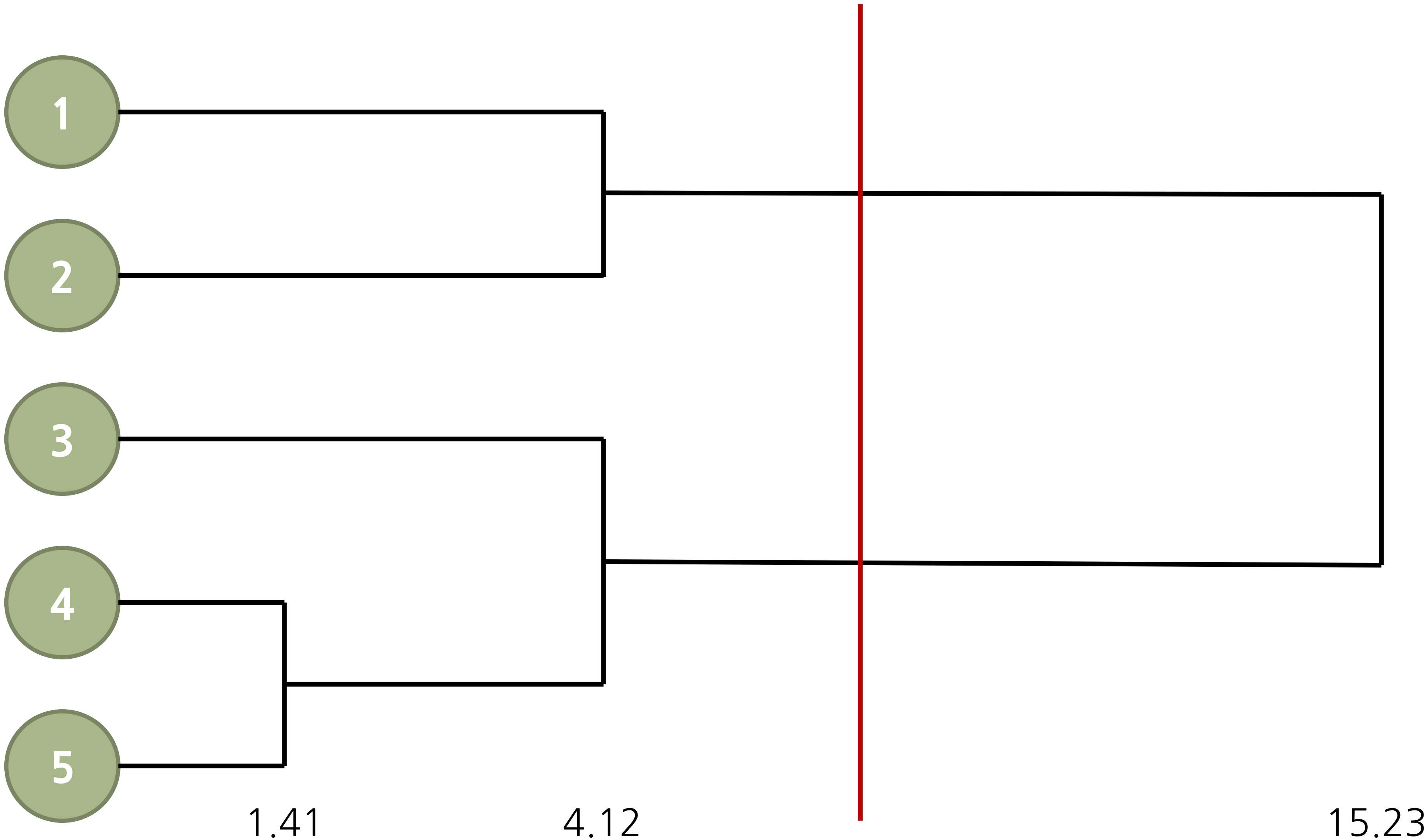
Distance $d(A, B) = \max\{d(a, b): a \in A, b \in B\}$

| | 7 | 8 |
|---|---|---|
| 7 | 0 | |
| 8 | 15.23 | 0 |



# Example: Complete Linkage Clustering

- Dendrogram



# Divisive Clustering - DIANA

- Divisive method starts with one cluster including all samples
  - At each step, divide cluster into two sub clusters until every cluster consists of one data point
  - This algorithm is based on the average distance between one object and the others

$$\bar{d}(i,C) = \begin{cases} \dfrac{1}{|C|-1} \displaystyle\sum_{j \in C, j \neq i} d(i,j), & if\ i \in C \\ \dfrac{1}{|C|} \displaystyle\sum_{j \in C} d(i,j), & if\ \ i \notin C \end{cases}$$

    - $i$ represent $i$-th object

# DIANA Algorithm

1. • Consider all samples as one cluster

2. • Select the cluster $C$ containing two objects with the longest distance

3. • Divide cluster $C$ into two as follows (At first, $C'$ is empty set($\phi$))
   - Find object $i$ with maximum $\bar{d}(i, C)$
   - $C \leftarrow C - \{i\}, C' \leftarrow C' \cup \{i\}$
   - If there exist the objects $j$ in $C$ whose $e(j) = \bar{d}(j, C) - \bar{d}(j, C') > 0$, select one of them with maximum $e(j)$, remove $j$ from $C$ and add $j$ into $C'$
   - If $e(j) < 0$ for all objects in $C$, finish this step

4. • Repeat step 2 and 3 until the number of clusters is the same as the number of samples

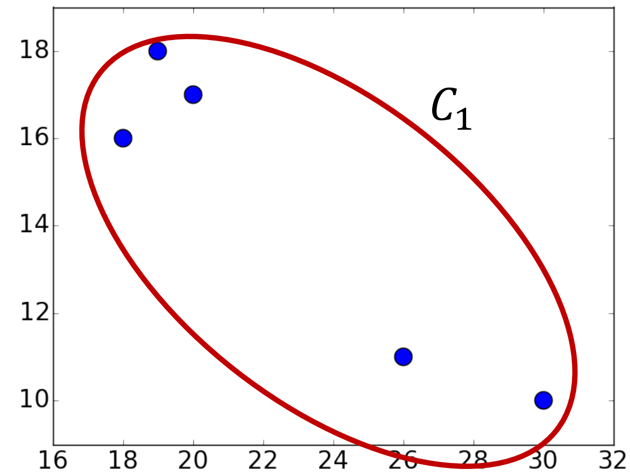
# Example: DIANA

- Find clusters through DIANA
  - Start with a cluster consisting of all objects
  - $C_1 = \{1,2,3,4,5\}$
  - $C_2 = \{\ \}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Step 2: Find pair of objects with the longest distance

| $d(i,j)$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | 4.12 | 0 | | | |
| 3 | **15.23** | 11.18 | 0 | | |
| 4 | 12.21 | 8.48 | 4.12 | 0 | |
| 5 | 13.60 | 9.90 | 3.61 | 1.41 | 0 |



# Example: DIANA

- Find clusters through DIANA
  - $C_1$ is the selected cluster

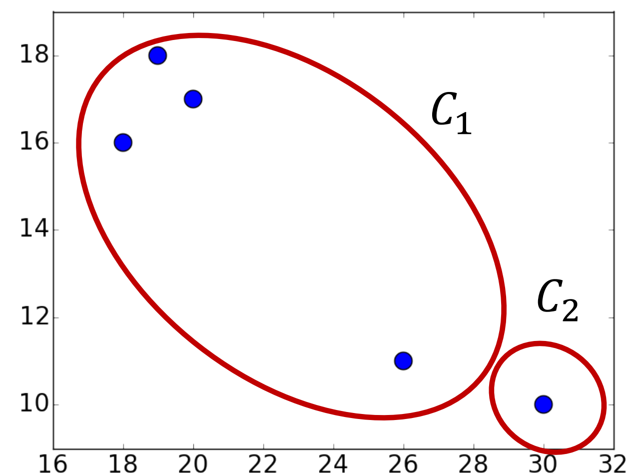
| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Step 3

| $d(i,j)$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | 4.12 | 15.23 | 12.21 | 13.60 |
| 2 | 4.12 | 0 | 11.18 | 8.48 | 9.90 |
| 3 | 15.23 | 11.18 | 0 | 4.12 | 3.61 |
| 4 | 12.21 | 8.48 | 4.12 | 0 | 1.41 |
| 5 | 13.60 | 9.90 | 3.61 | 1.41 | 0 |

↓ Average except 0

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\bar{d}\,(i, C_1)$ | **11.29** | 8.42 | 8.54 | 6.56 | 7.13 |



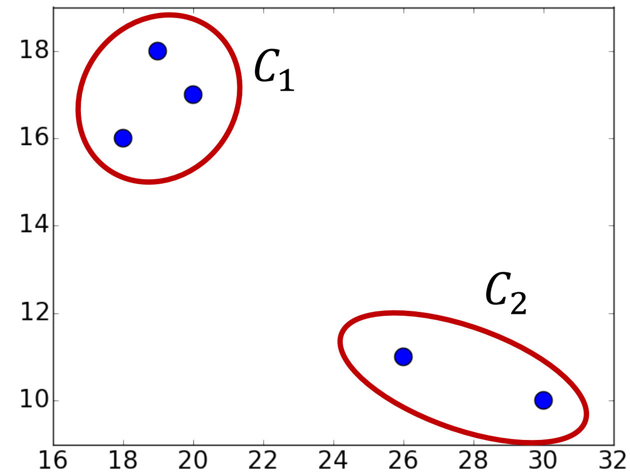
# Example: DIANA

- Find clusters through DIANA
  - $C_1 = \{2,3,4,5\}, C_2 = \{1\}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Step 3

| | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| $\bar{d}\ (i, C_1)$ | 9.85 | 6.30 | 4.67 | 4.97 |
| $\bar{d}\ (i, C_2)$ | 4.12 | 15.2 | 12.2 | 13.6 |
| $e(i)$ | **5.73** | -8.9 | -7.53 | -8.63 |



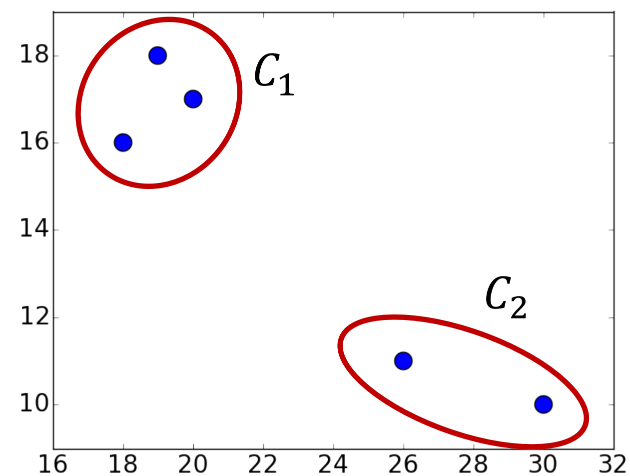
# Example: DIANA

- Find clusters through DIANA
  - $C_1 = \{3,4,5\}, C_2 = \{1,2\}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Step 3

| | 3 | 4 | 5 |
|---|---|---|---|
| $\bar{d}\,(i, C_1)$ | 3.87 | 2.77 | 2.51 |
| $\bar{d}\,(i, C_2)$ | 13.21 | 10.35 | 11.75 |
| $e(i)$ | -9.34 | -7.58 | -9.24 |



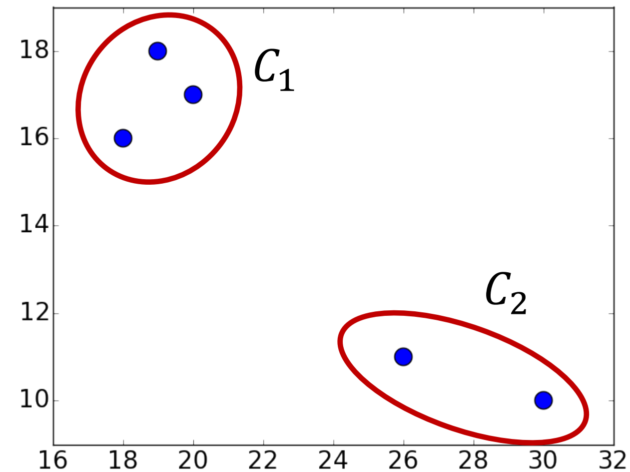
# Example: DIANA

- Find clusters through DIANA
  - $C_1 = \{3,4,5\}, C_2 = \{1,2\}$
  - Find pair of objects wit the longest distance

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Step 2: Find pair of objects with the longest distance

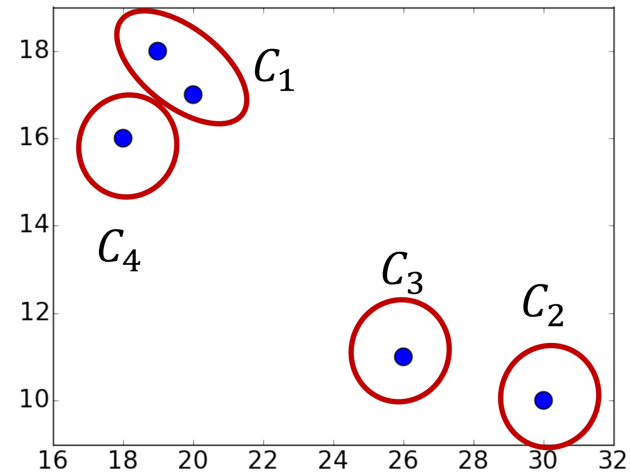
| $d(i,j)$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | | | | |
| 2 | **4.12** | 0 | | | |
| 3 | | | 0 | | |
| 4 | | | **4.12** | 0 | |
| 5 | | | 3.61 | 1.41 | 0 |



# Example: DIANA

- Find clusters through DIANA
  - Select $C_2$
  - $C_1 = \{1,2\}, C_3 = \{\ \}$
  - $C_2$ contains only two object, so divide $C_2$ into two clusters directly: $C_2 = \{1\}, C_3 = \{2\}$
  - Select $C_1$
  - $C_1 = \{3,4,5\}, C_4 = \{\ \}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |




Step 3

| | 3 | 4 | 5 |
|---|---|---|---|
| $\bar{d}(i, C_1)$ | **3.87** | 2.77 | 2.51 |

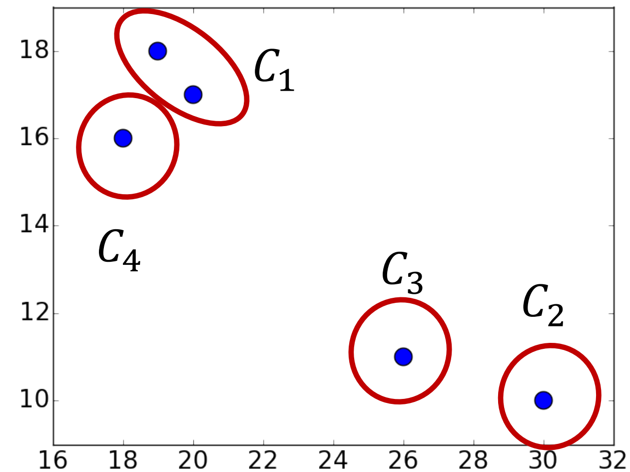
# Example: DIANA

- Find clusters through DIANA
  - $C_1 = \{4,5\}, C_4 = \{3\}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

Step 3

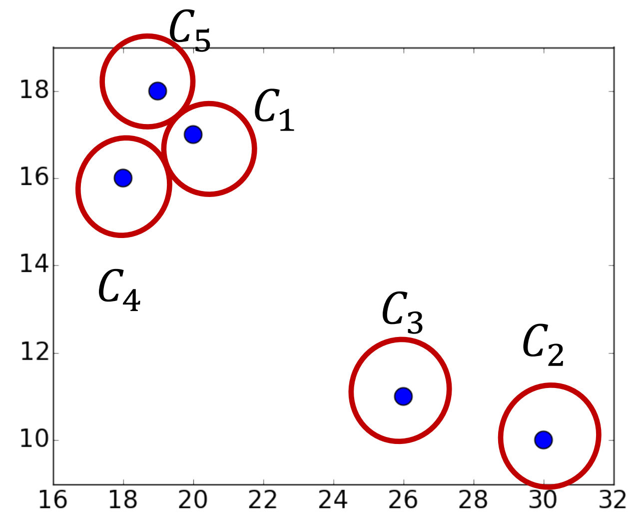
| | 4 | 5 |
|---|---|---|
| $\bar{d}\ (i, C_1)$ | 1.41 | 1.41 |
| $\bar{d}\ (i, C_4)$ | 4.12 | 3.61 |
| $e(i)$ | -2.71 | -2.20 |



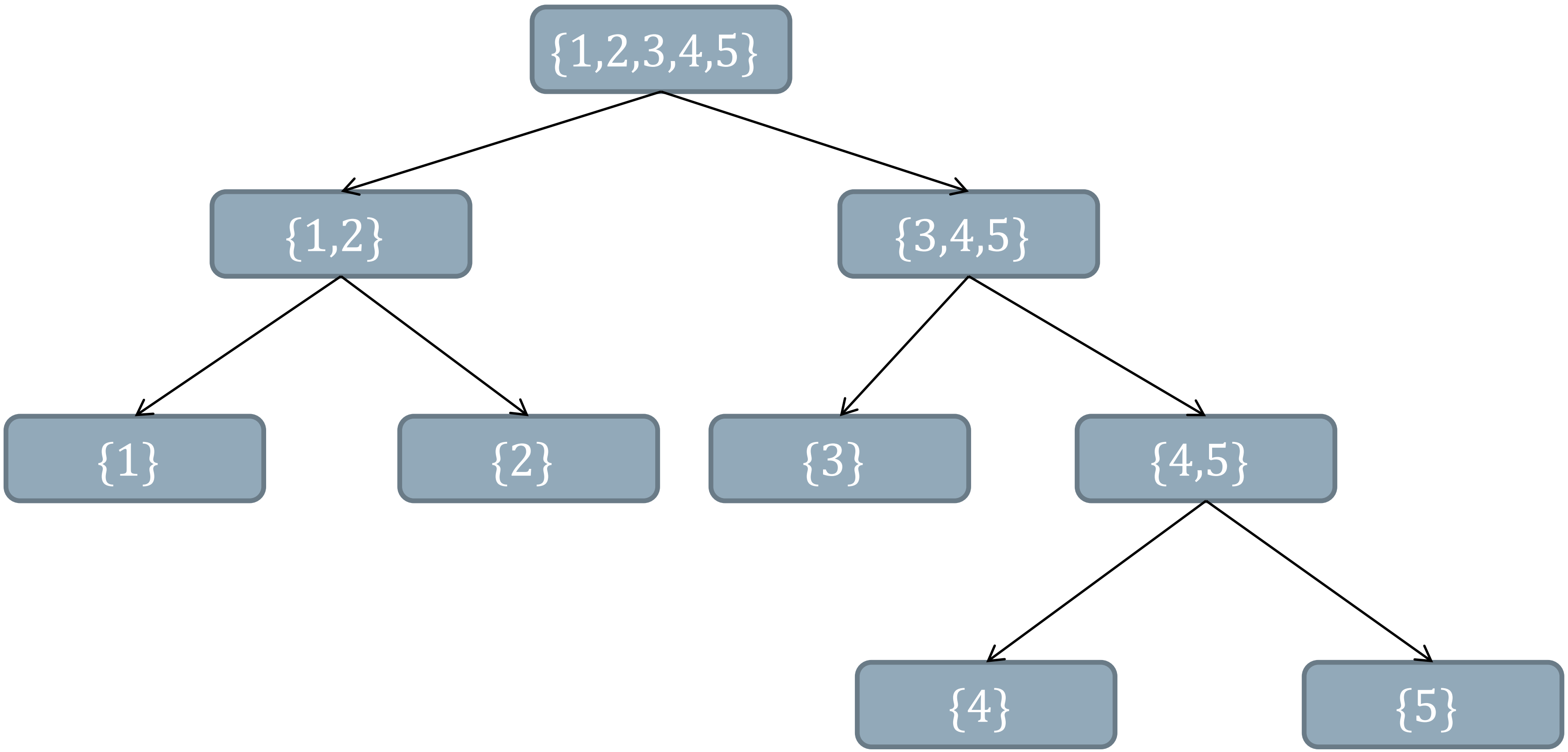
# Example: DIANA

- Find clusters through DIANA
  - Select $C_1$
  - $C_1 = \{4,5\}, C_5 = \{\ \}$
  - $C_1$ contains only two object, so divide $C_1$ into two clusters directly: $C_1 = \{4\}, C_5 = \{5\}$

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $x$ | 30 | 26 | 16 | 20 | 19 |
| $y$ | 10 | 11 | 16 | 17 | 18 |

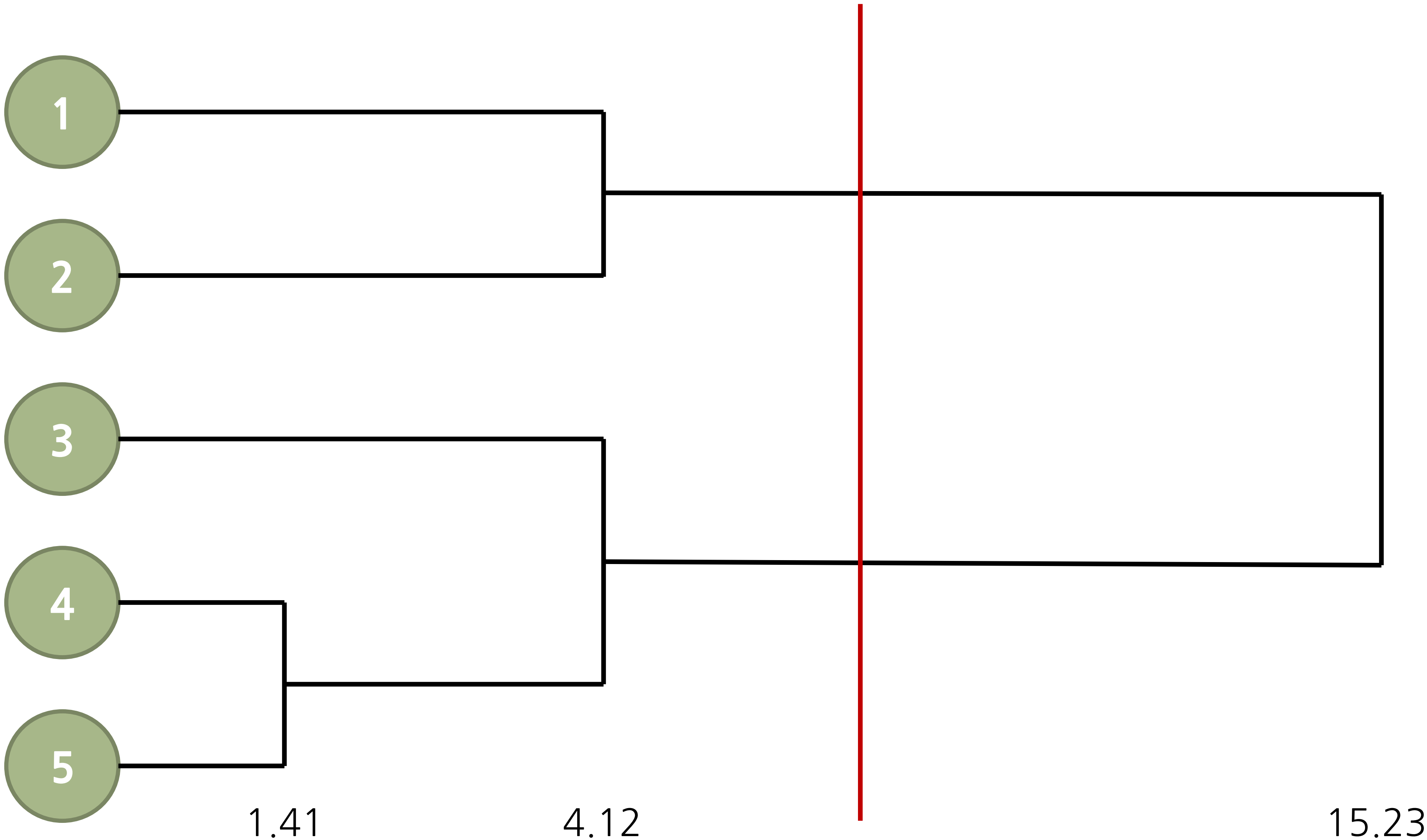


# Example: DIANA



# Example: DIANA

- Dendrogram

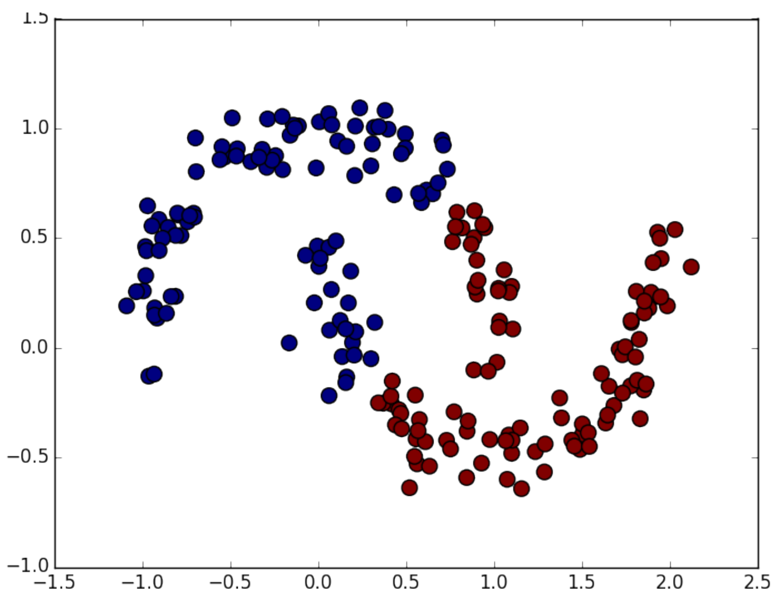


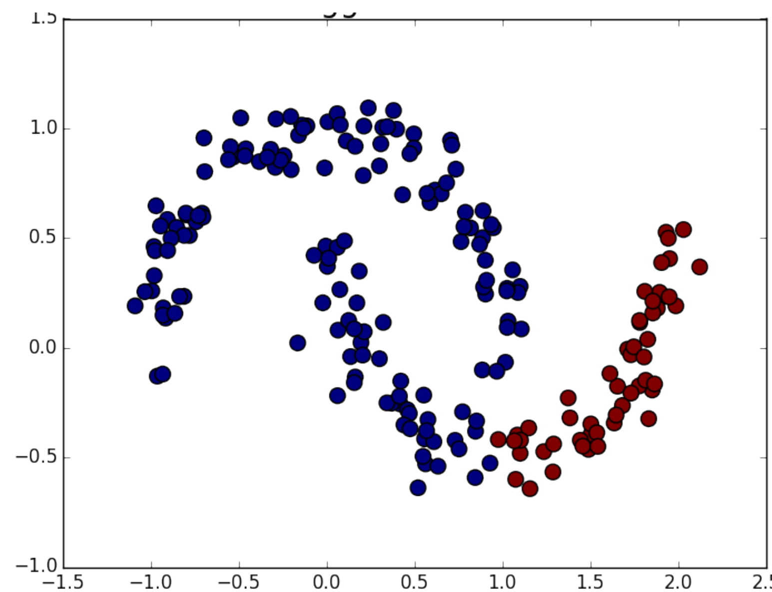
# Evaluation Measure

# How to Measure Clustering Quality

- Clustering problem is unsupervised problem
  - No explicit answer for learning
  - We need to define a method to measure quality of clustering

**Which one is better?**



# How to Measure Clustering Quality

- Measures that do not require ground truth labels
  - Inertia
    - Within-cluster sum-of-squares

$$\sum_{i=0}^{n} \min_{\mu_j \in C} \left\| x_j - \mu_i \right\|^2$$

  - Silhouette Coefficient
    - $s(i)$: Silhouette coefficient of $i$-th sample
    - $a(i)$: The mean distance between a sample and all other points in the same class
    - $b(i)$: The mean distance between a sample and all other points in the next nearest cluster

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$$-1 \le s(i) \le 1$$

    - Overall clustering quality can be obtained by averaging $s(i)$ for all samples

# How to Measure Clustering Quality

- Clustering performance evaluation measure
  - Homogeneity: each cluster contains only members of a single class

$$h = 1 - \frac{H(C|K)}{H(C)}$$

    - $H(C)$ is the entropy of the classes

$$H(C) = -\sum_{c=1}^{|C|} \frac{n_c}{n} \cdot \log\left(\frac{n_c}{n}\right)$$

    - $H(C|K)$ is the conditional entropy of the classes given the cluster assignments

$$H(C|K) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \log\left(\frac{n_{c,k}}{n_k}\right)$$

    - $n$ is the total number of samples, $n_c$ and $n_k$ are the number of samples respectively belonging to class $c$ and cluster $k$
    - $n_{c,k}$ is the number of samples from class $c$ assigned to cluster $k$
  - Completeness: all members of a given class are assigned to the same cluster

$$c = 1 - \frac{H(K|C)}{H(K)}$$

# How to Measure Clustering Quality

- Clustering performance evaluation measure
  - Adjusted Rand Index(ARI)
    - Given the knowledge of the ground truth class assignments and our clustering algorithm assignments of the same samples, the adjusted Rand index is a function that measures the similarity of the two assignments

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]}$$

    - $C$ is a ground truth class assignment, $K$ is the clustering
    - $a$ is the number of pairs of elements that are in the same set in $C$ and in the same set in $K$
    - $b$ is the number of pairs of elements that are in different sets in $C$ and in different sets in $K$
    - Raw Rand index $\mathrm{RI} = \frac{a+b}{C_2^n}$($C_2^n$ is the total number of possible pairs in the dataset)

$$C_2^n = \frac{n!}{2!\,(n-2)!}$$

# How to Measure Clustering Quality

- Contingency table

| | $K_1$ | $K_2$ | $\cdots$ | $K_s$ | $sums$ |
|---|---|---|---|---|---|
| $C_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1s}$ | $a_1$ |
| $C_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2s}$ | $a_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $C_r$ | $n_{r1}$ | $n_{r2}$ | $\cdots$ | $n_{rs}$ | $a_r$ |
| $sums$ | $b_1$ | $b_2$ | $\cdots$ | $b_s$ | $n$ |

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

$$\binom{n}{k} = {}_nC_k$$

# How to Measure Clustering Quality

- **[Drawbacks]**
  Homogeneity, Completeness, and ARI require knowledge of the ground truth classes while is almost never available in practice or require manual assignment by human annotators