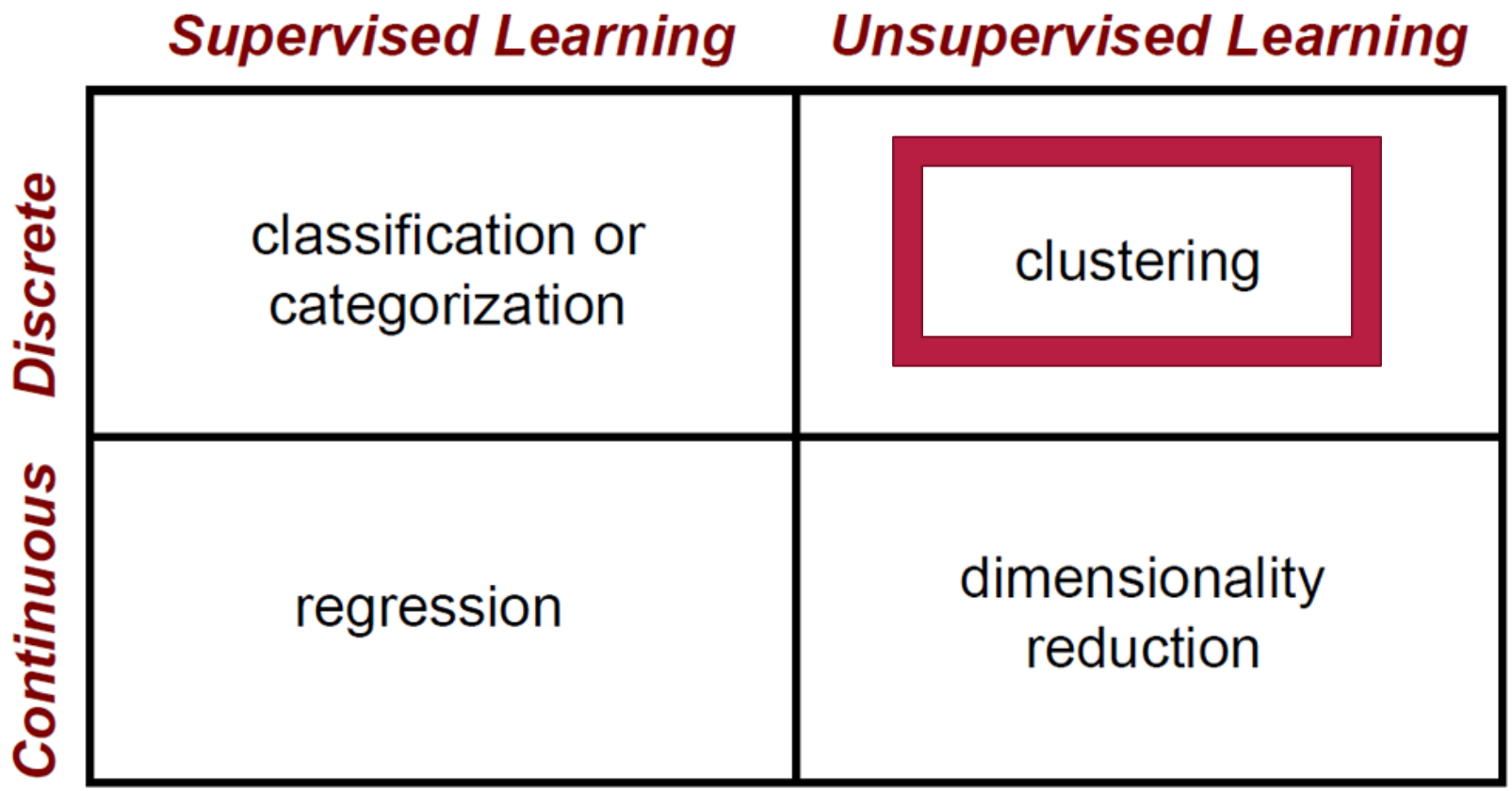
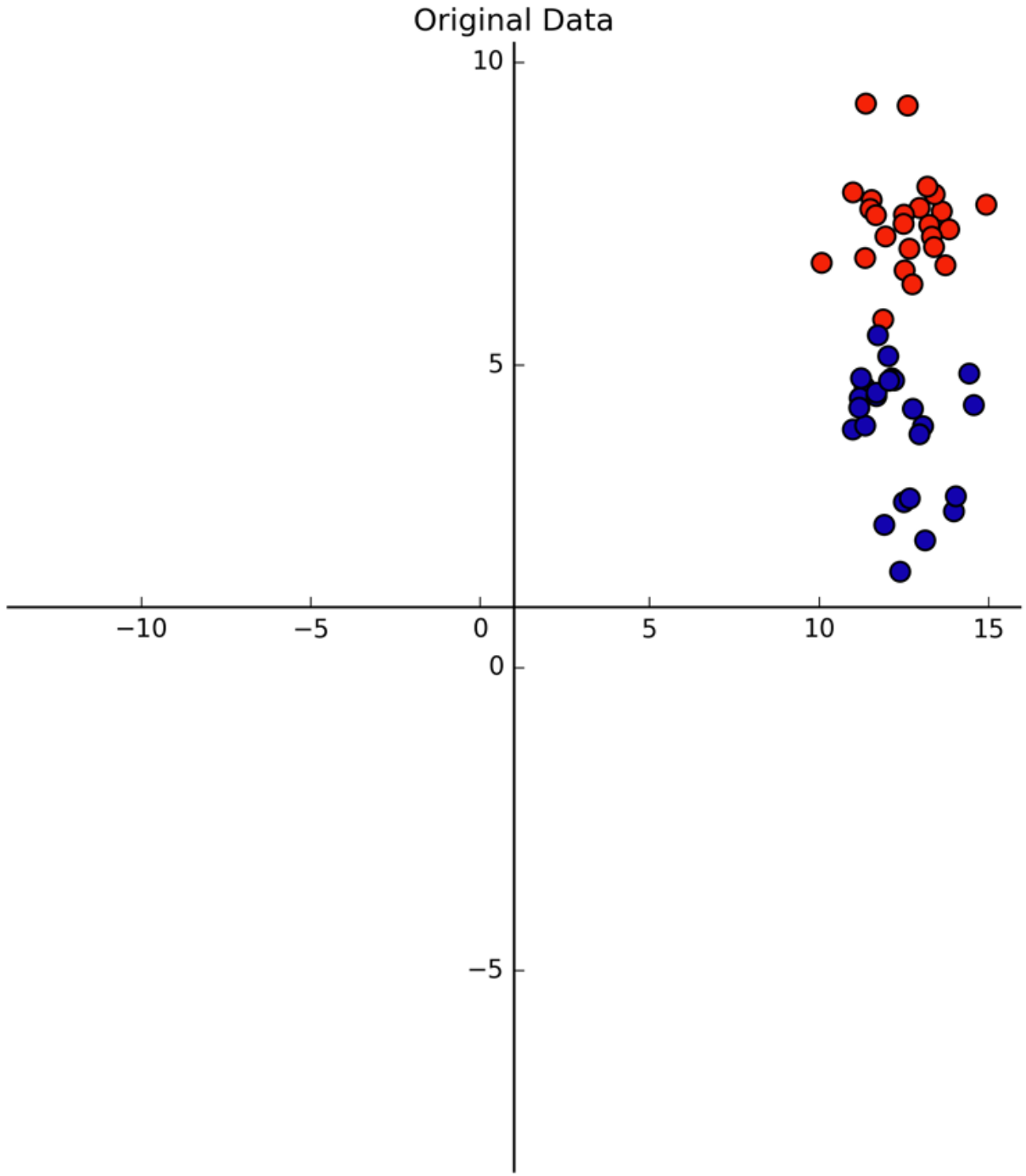
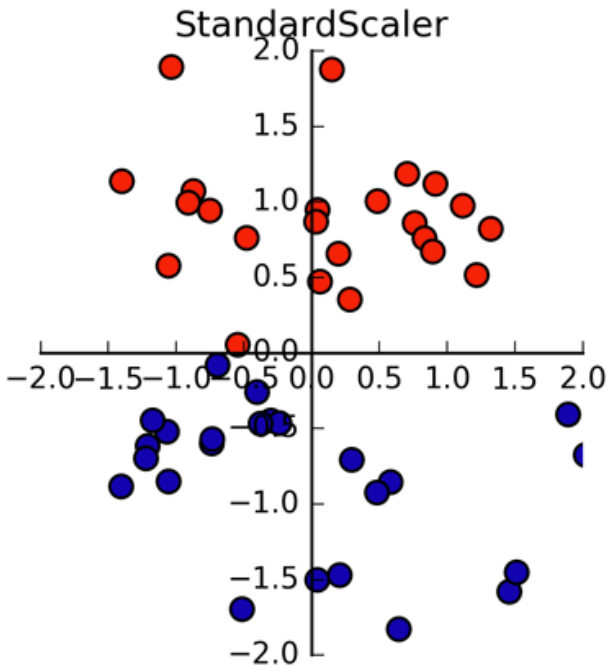
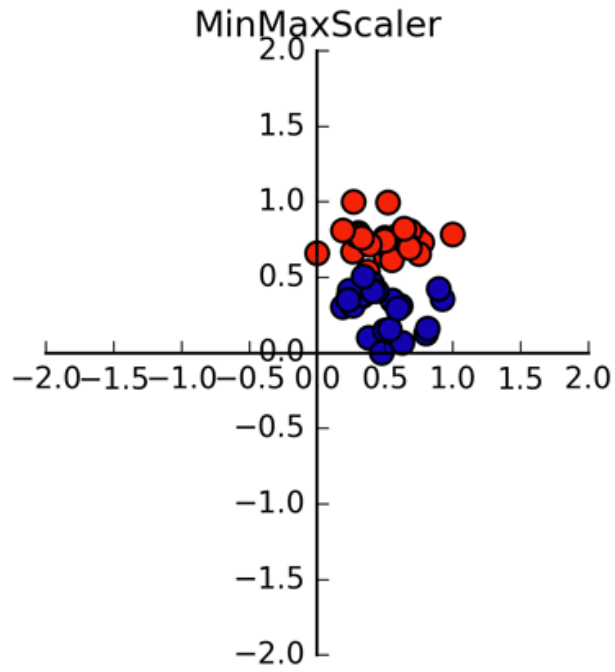
BDA502 WEEK-5 INTRODUCTION TO MACHINE LEARNING

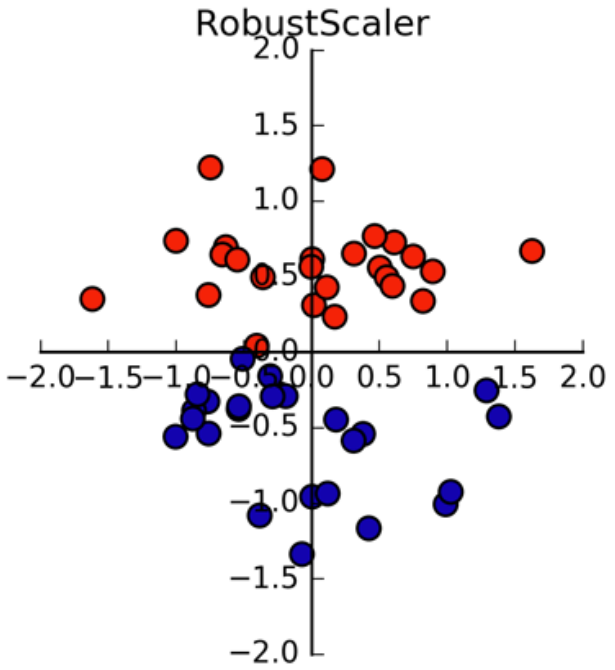
Machine Learning Problems

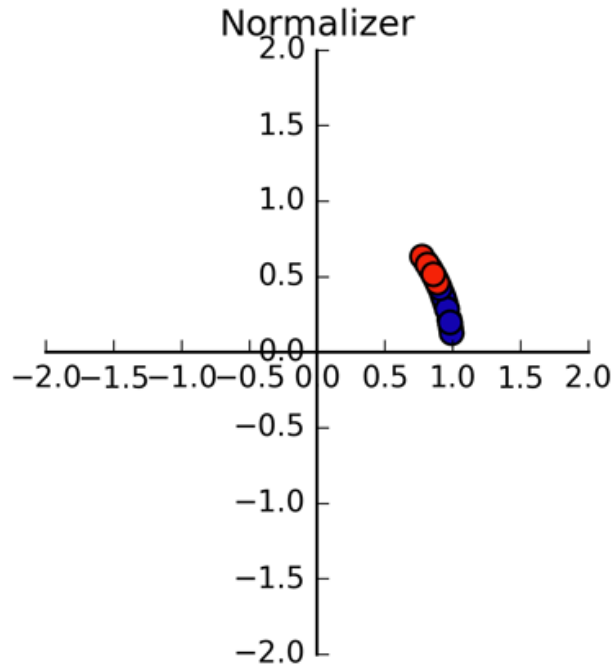



Original Data
10
5
0
−5
−10
−5
0
5
10
15


StandardScaler
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0
0.5
1.0
1.5
2.0


MinMaxScaler
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0
0.5
1.0
1.5
2.0


RobustScaler
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0
0.5
1.0
1.5
2.0


Normalizer
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0
−1.5
−1.0
−0.5
0
0.5
1.0
1.5
2.0

CLUSTERING STRATEGIES

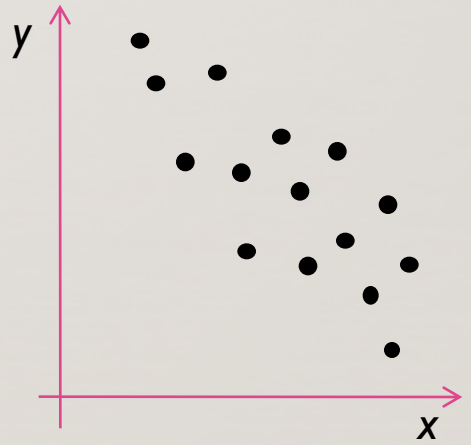
- K-means
  - Iteratively re-assign points to the nearest cluster center
- Agglomerative clustering
  - Start with each point as its own cluster and iteratively merge the closest clusters
- Mean-shift clustering
  - Estimate modes of pdf
- Spectral clustering
  - Split the nodes in a graph based on assigned links with similarity weights

As we go down this chart, the clustering strategies have more tendency to transitively group points even if they are not nearby in feature space

GRAPHING PAIRED DATA SETS

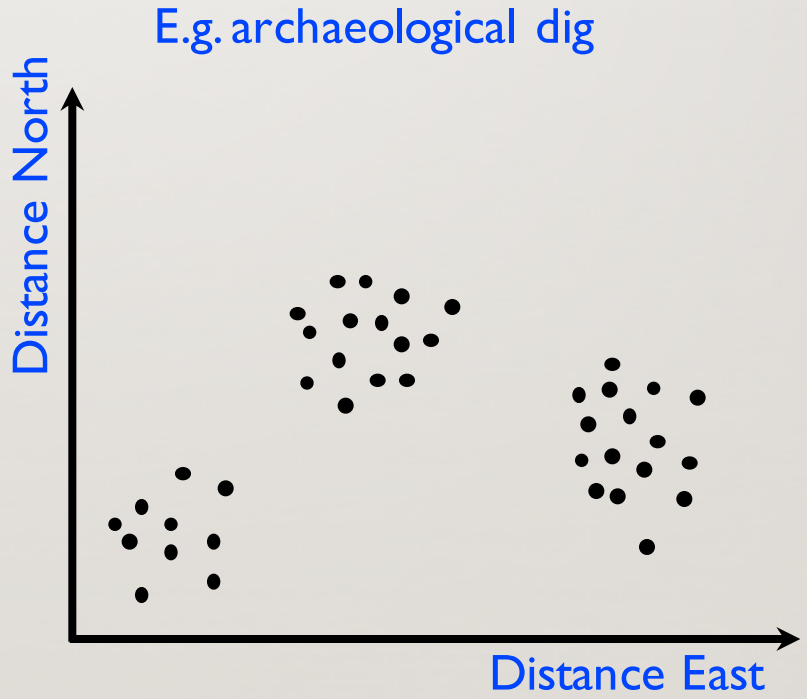
Paired Data Sets

- Each entry in one data set corresponds to one entry in a second data set.
- Graph using a **scatter plot.**
  - The ordered pairs are graphed as points in a coordinate plane.
  - Used to show the relationship between two quantitative variables.



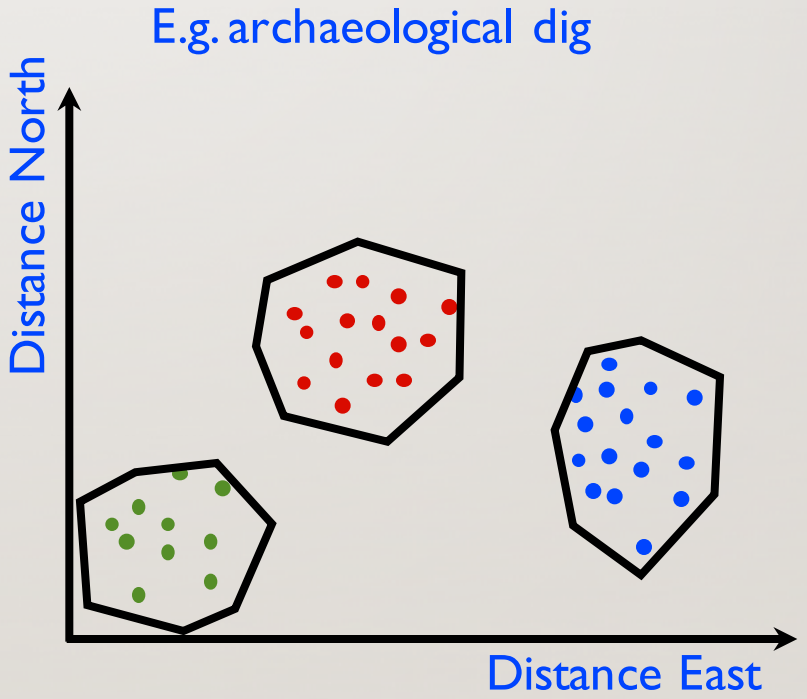
Clustering

Grouping data according to similarity



Clustering

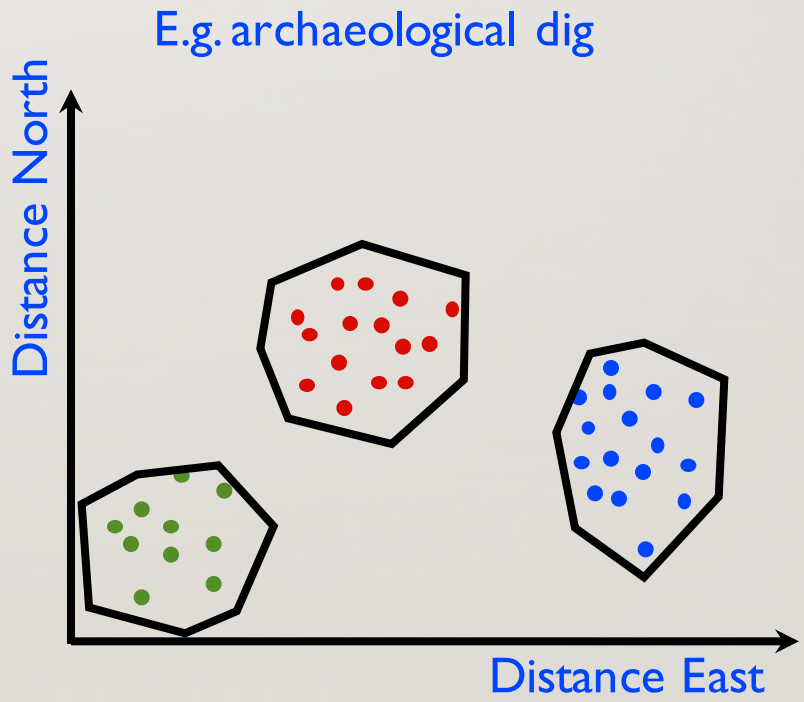
Grouping data according to similarity



K-Means Algorithm

Benefits

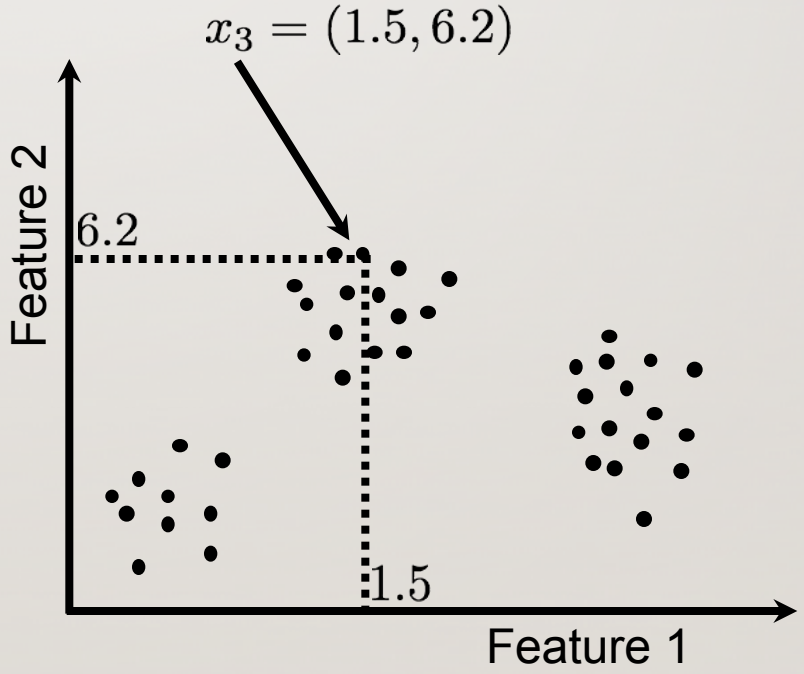
- Popular
- Fast
- Conceptually straightforward



K-Means: preliminaries

Data: Collection of values

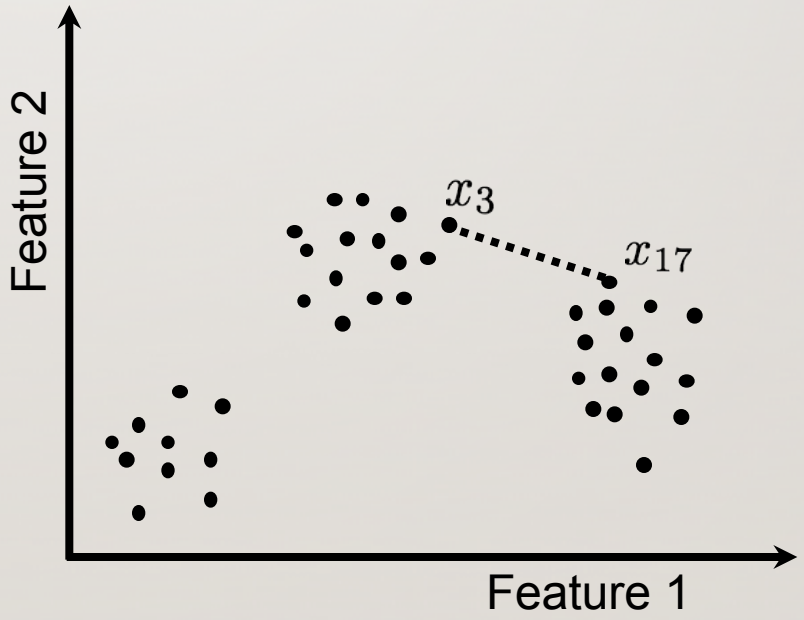
```
data = lines.map(line=>
    parseVector(line))
```



K-Means: preliminaries

Dissimilarity:
Squared Euclidean distance

```
dist = p.squaredDist(q)
```



K-Means: Preliminaries

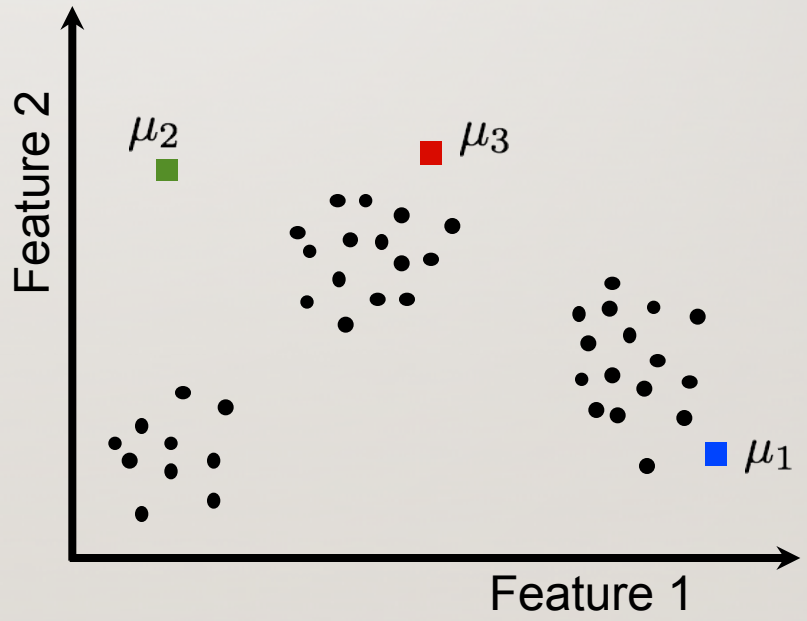
K = Number of clusters

$\mu_1, \mu_2, \ldots, \mu_K$

Data assignments to clusters

$S_1, S_2, \ldots, S_K$

Cluster central point or centroid or cluster center.



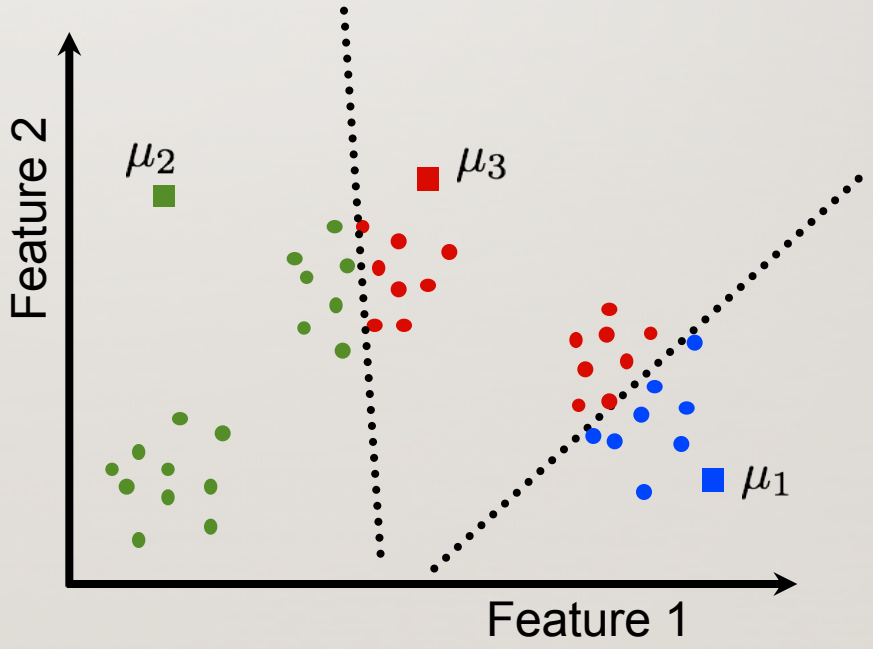
K-Means: preliminaries

K = Number of clusters

$\mu_1, \mu_2, \ldots, \mu_K$

Data assignments to clusters

$S_1, S_2, \ldots, S_K$



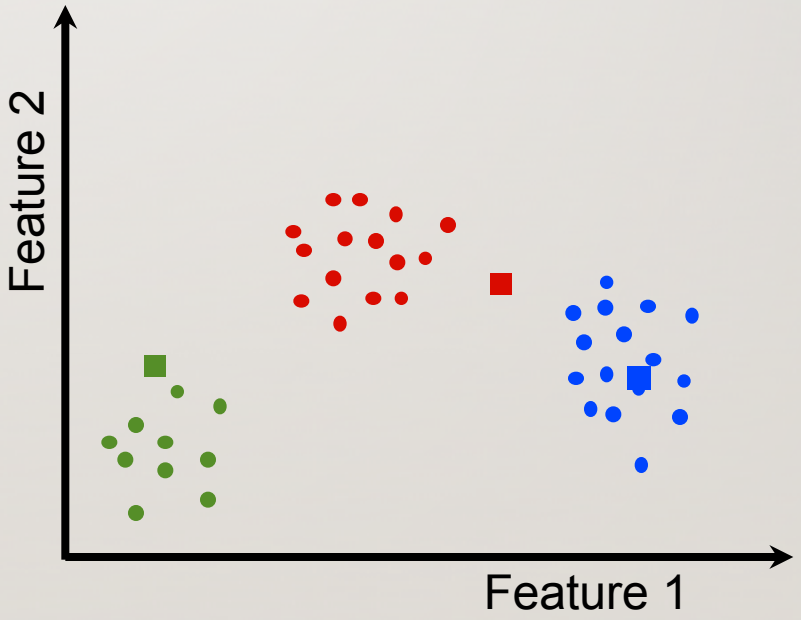
K-Means Algorithm

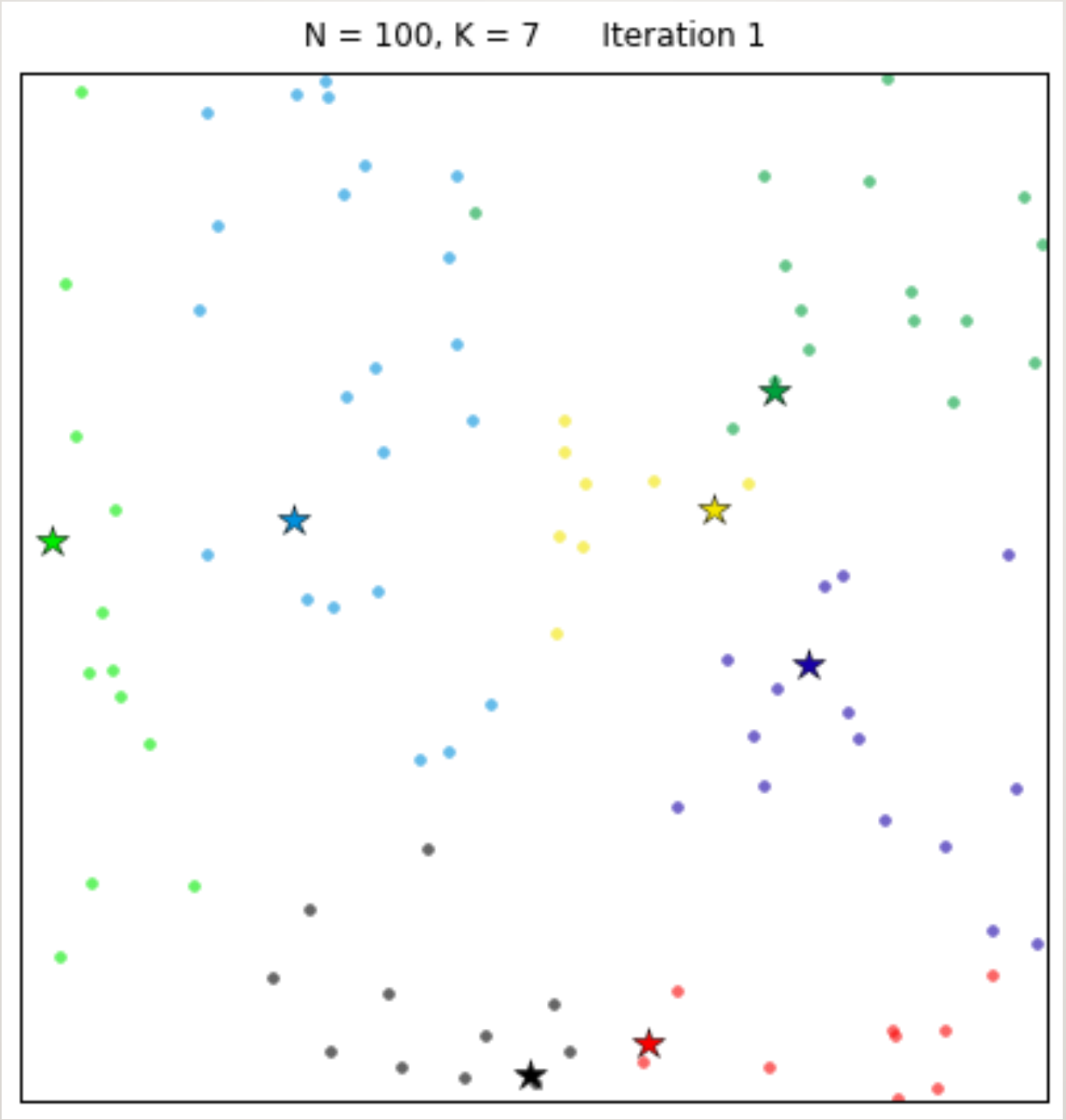
- Initialize K cluster centers

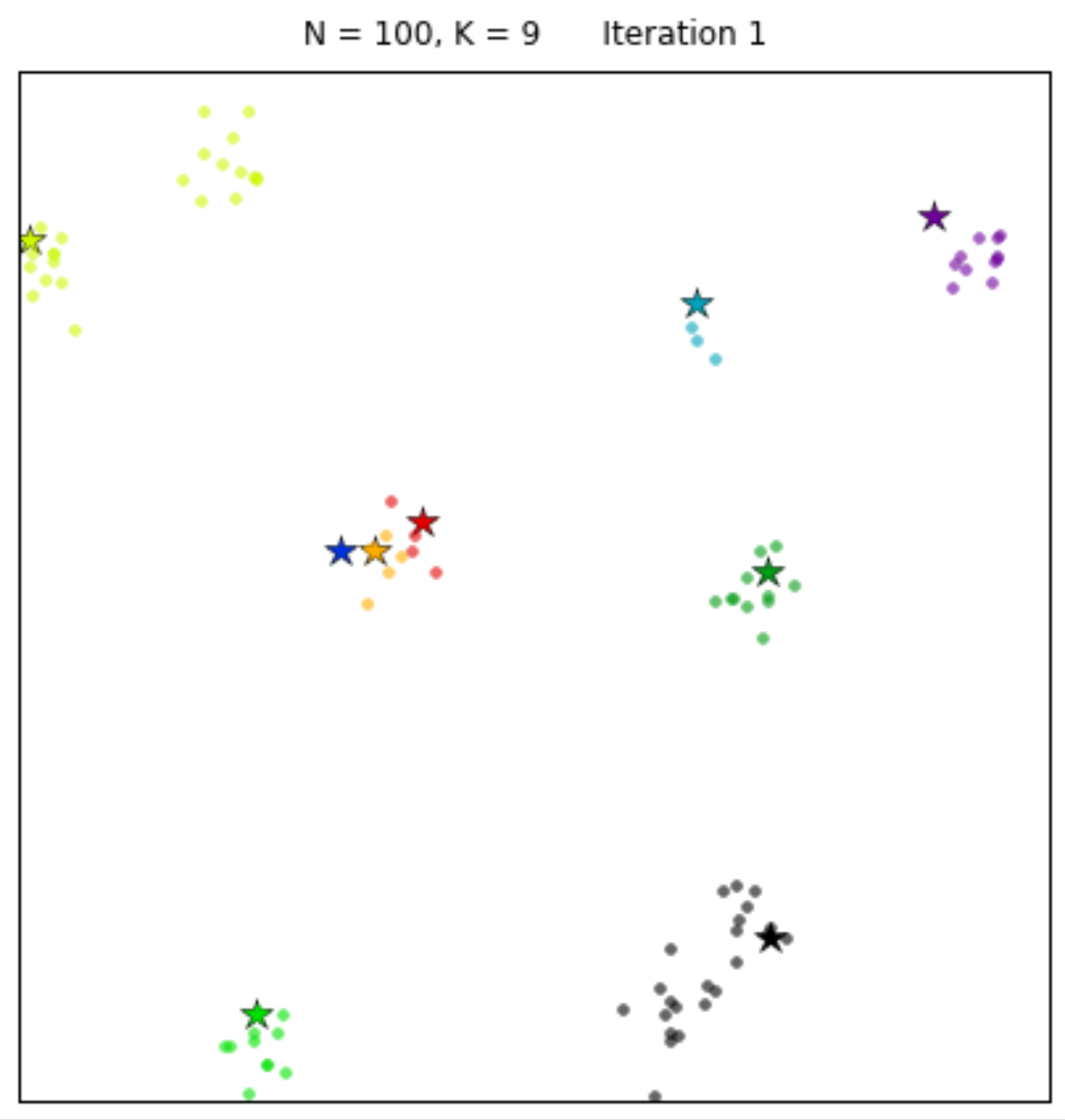
```
centers = data.takeSample(
    false, K, seed)
```

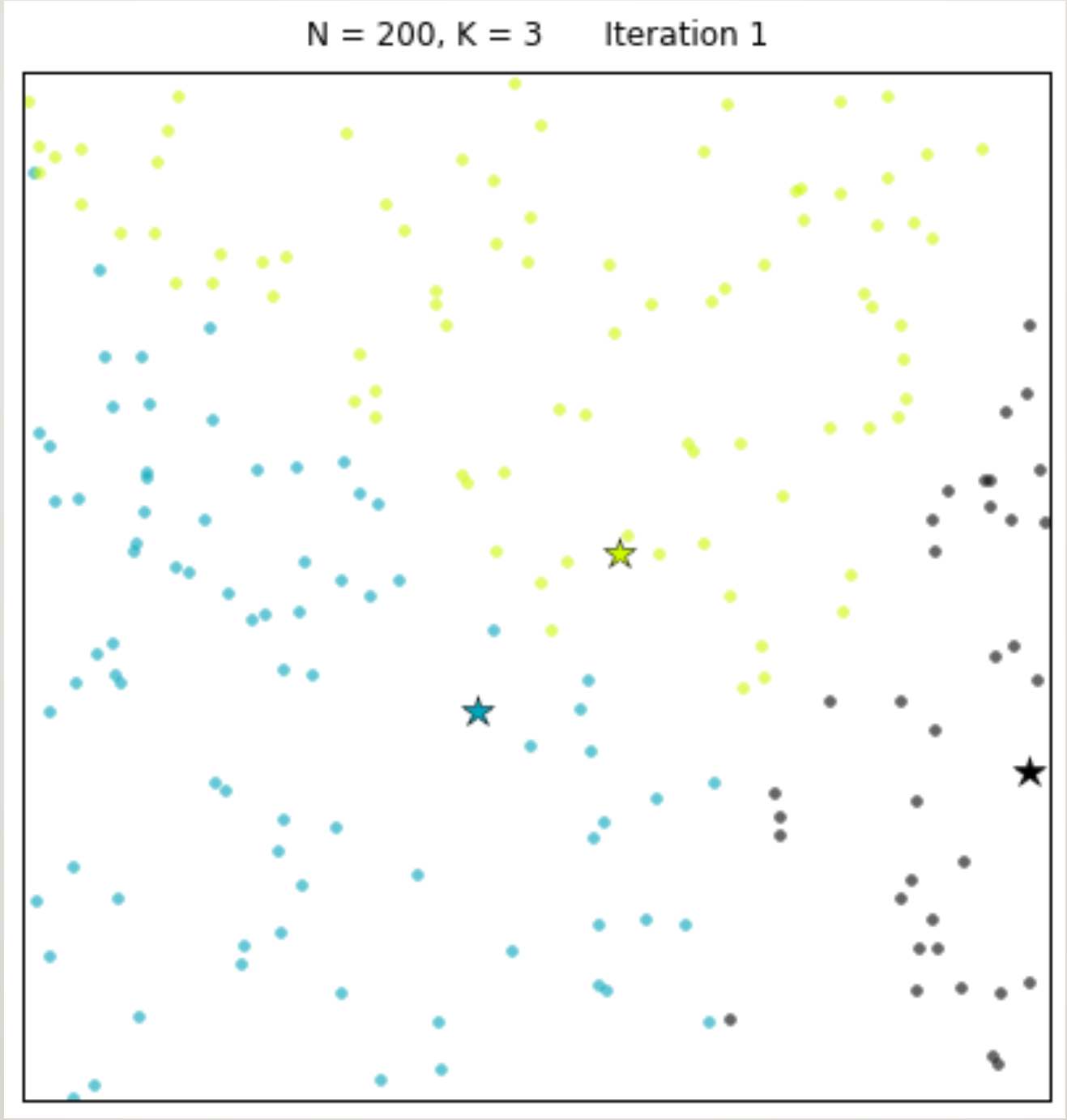
- Repeat until convergence:

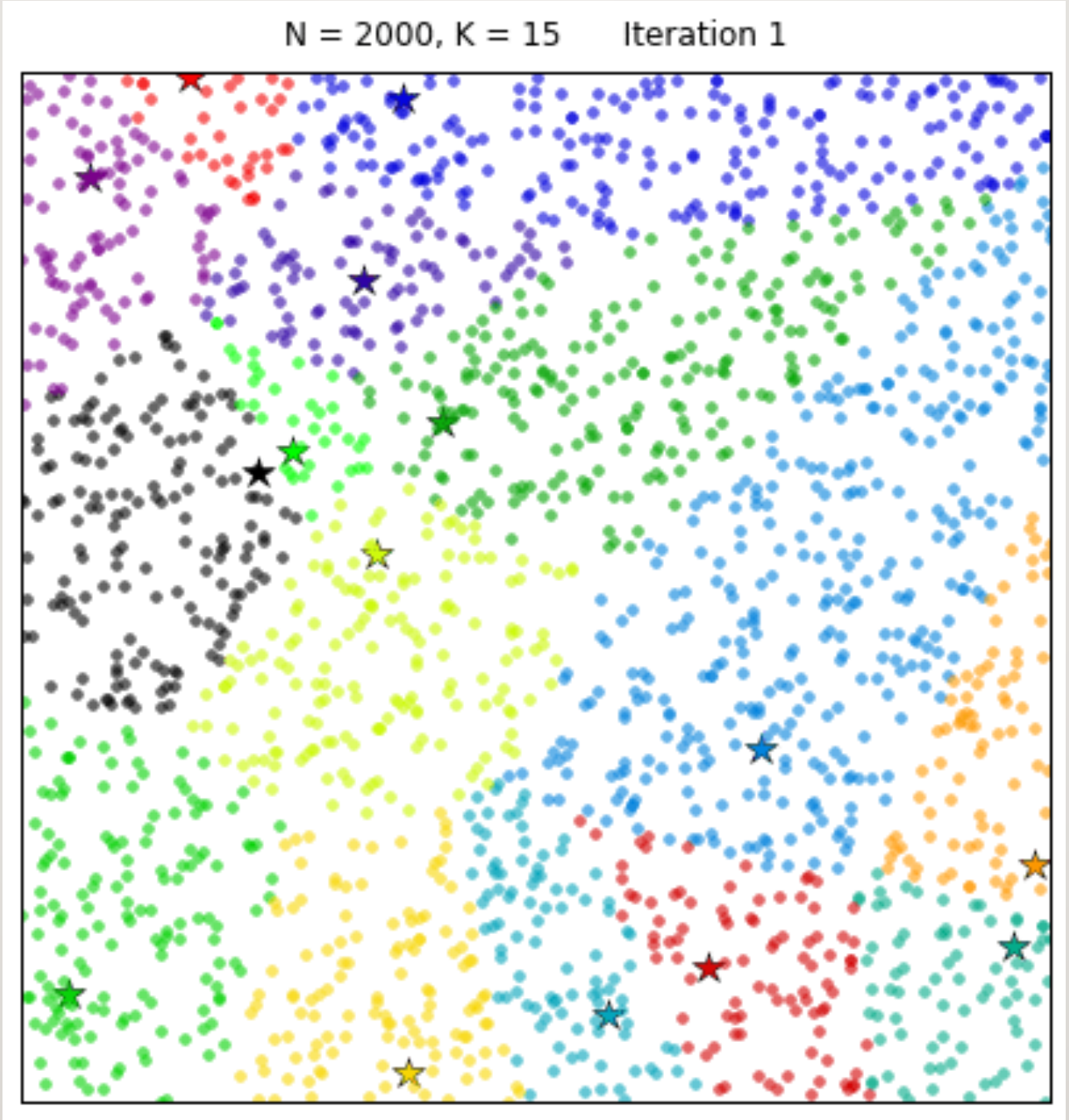
```
while (dist(centers, newCenters) > ε)
  closest = data.map(p =>

  (closestPoint(p,centers),p))

  pointsGroup =
      closest.groupByKey()

newCenters =pointsGroup.mapValues(
    ps => average(ps))
```




N = 100, K = 7
Iteration 1


N = 100, K = 9
Iteration 1


N = 200, K = 3
Iteration 1


N = 2000, K = 15
Iteration 1

CRUCIAL ISSUES

- Determining the number of clusters: Requires trial and error. Analysts should evaluate different clustering solutions (domain-knowledge is crucial)
- Analysts should examine and evaluate the revealed cluster solution and assess, among other things, the number and relative size of the clusters, their cohesion, and separation.
- A good clustering solution contains tightly cohesive and highly separated clusters. More specifically, the solution, through the use of descriptive statistics and specialized measures, should be examined.

EXAMINATION AND EVALUATION

- **THE NUMBER OF CLUSTERS AND THE SIZE OF EACH CLUSTER**

  A rich but manageable number of clusters (large versus small)

- **COHESION OF THE CLUSTERS**

  Dense concentrations of records around their centroids (pooled standard deviations, the cluster radius, average sum of squares error (SSE))

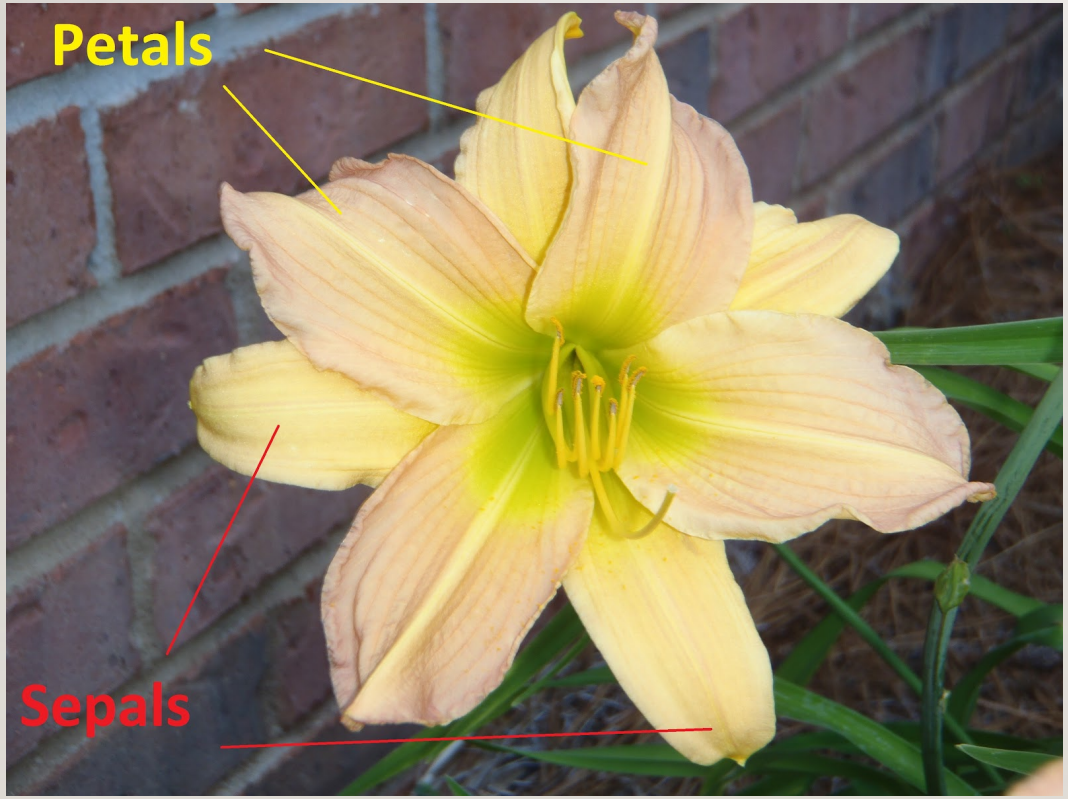
- **SEPARATION OF THE CLUSTERS**

  A proximity matrix with the distances between the cluster centroids

  Distances of each cluster's centroid to the overall centroid of the whole population

→ Silhouette coefficient: internal cohesion and the external separation of a clustering solution

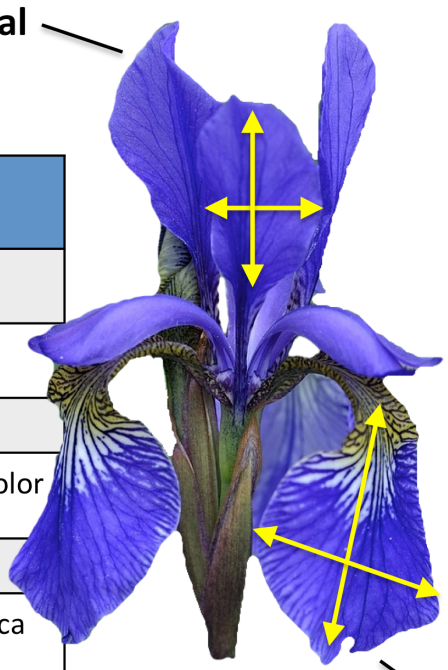
EXAMPLE: INTERPRETING A SCATTER PLOT



The British statistician Ronald Fisher introduced a famous data set called Fisher's Iris data set. This data set describes various physical characteristics, such as petal length and petal width (in millimeters), for three species of iris. The petal lengths form the first data set and the petal widths form the second data set. *(Source: Fisher, R. A., 1936)*

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica



Features
(attributes, measurements, dimensions)

Class labels
(targets)

Classes	**3**
Samples per class	**50**
Samples total	**150**
Dimensionality	**4**
Features	**real, positive**

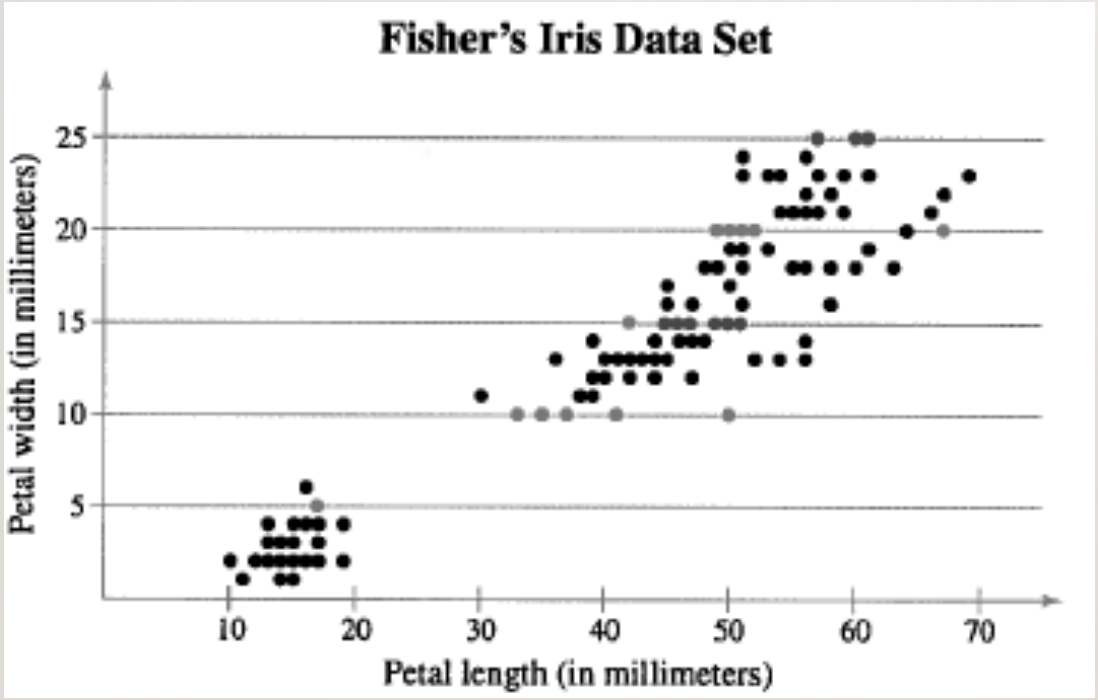
Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
7.0	3.2	4.7	1.4	*I. versicolor*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.5	2.8	4.6	1.5	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
5.8	2.7	5.1	1.9	*I. virginica*
7.1	3.0	5.9	2.1	*I. virginica*
6.3	2.9	5.6	1.8	*I. virginica*
6.5	3.0	5.8	2.2	*I. virginica*

Wikipedia, 2017

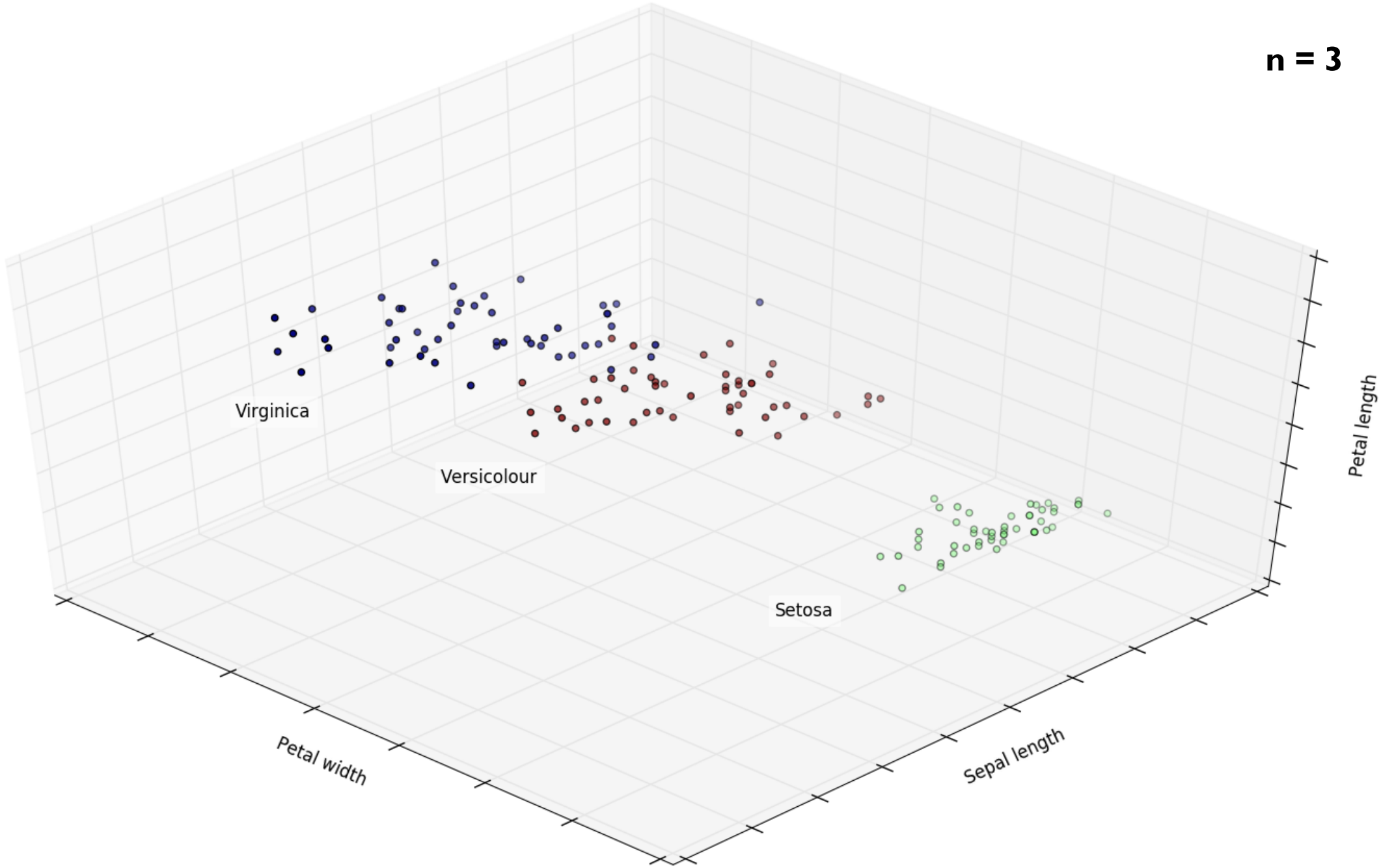
INTERPRETING A SCATTER PLOT

As the petal length increases, what tends to happen to the petal width?

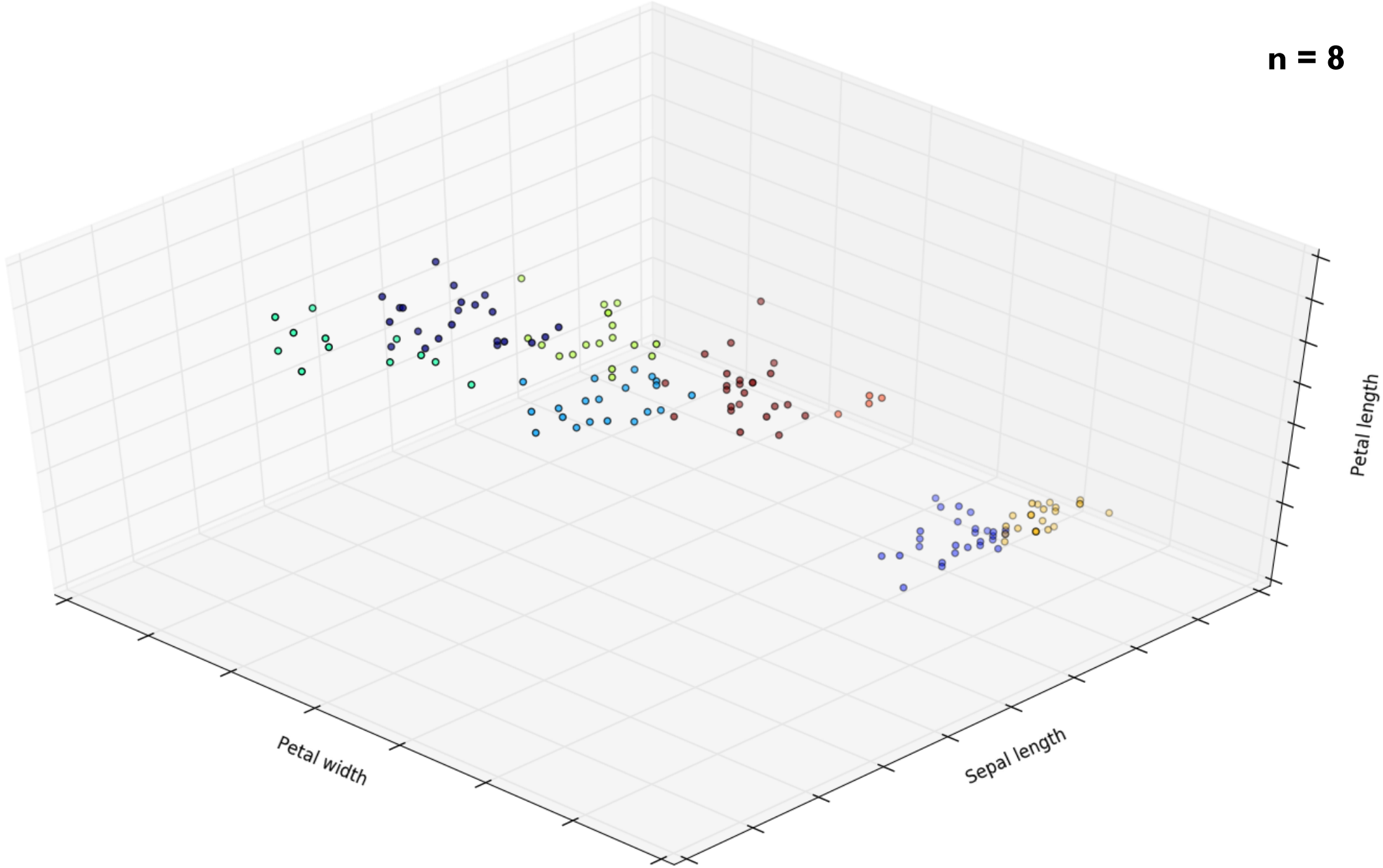


Each point in the scatter plot represents the petal length and petal width of one flower.

n = 3


Virginica
Versicolour
Setosa
Petal width
Sepal length
Petal length

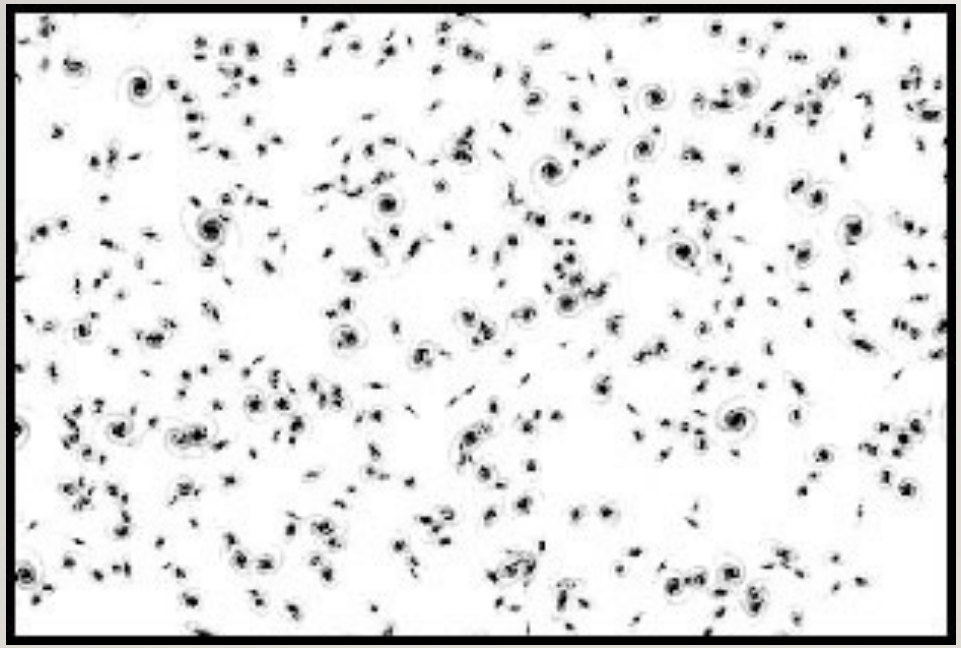
n = 8



KEY TERMS FOR ANALYZING THE OUTPUTS

- Homogeneity score (homo)
- Completeness Score (compl)
- V-measure (v-meas)
- Adjusted Rand Index (ARI)
- Adjusted Mutual Information (AMI)
- Silhouette Coefficient (silhouette)

HOMOGENEITY SCORE (HOMO)



- *A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.*
- *This metric is independent of the absolute values of the labels: a permutation of the class or cluster label values won't change the score value in any way.*

COMPLETENESS SCORE

- All members of the same class are in the same cluster.
- Completeness metric of a cluster labeling given a ground truth.
- A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster.

V-MEASURE (V-MEAS)

- V-measure cluster labeling given a ground truth.
- The V-measure is the harmonic mean between homogeneity and completeness:

v = 2 * (homogeneity * completeness) / (homogeneity + completeness)

ADJUSTED RAND INDEX (ARI)

- The Rand Index computes a similarity measure between two clusters by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clusters.
- The raw RI score is then "adjusted for chance" into the ARI score.

ADJUSTED MUTUAL INFORMATION (AMI)

- Adjusted Mutual Information (AMI) is an adjustment of the Mutual Information (MI) score to account for chance.
- It accounts for the fact that the MI is generally higher for two clusterings with a larger number of clusters, regardless of whether there is actually more information shared.
- For two clusterings U and V, the AMI is given as:

$$E\{MI(U,V)\} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{ij}}{N}\log\left(\frac{N\cdot n_{ij}}{a_i b_j}\right) \times \frac{a_i!b_j!(N-a_i)!(N-b_j)!}{N!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(N-a_i-b_j+n_{ij})!}$$

-1 0 +1
SILHOUETTE ANALYSIS

- Silhouette analysis can be used to study the separation distance between the resulting clusters.
- The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually.

SILHOUETTE COEFFICIENT

$$\mathrm{Si} = [b(i) - a(i)] / \max\{a(i), b(i)\}$$

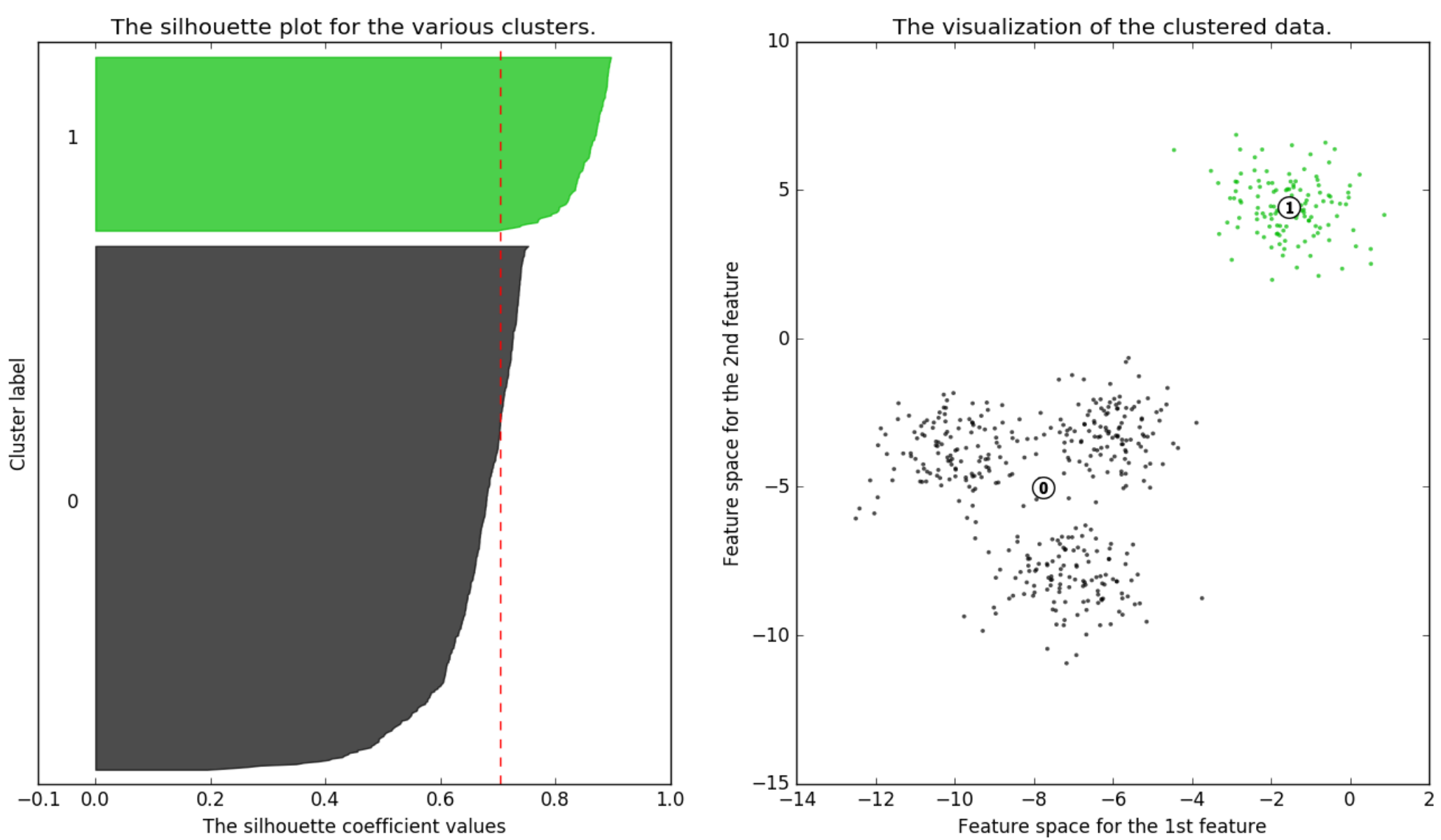
(1) The average (Euclidean) distance to all other records in the same cluster. This value indicates how well a specific record fits a cluster: $a(i)$

(2) The average (Euclidean) distance of the record to all the members of the neighboring cluster: $b(i)$

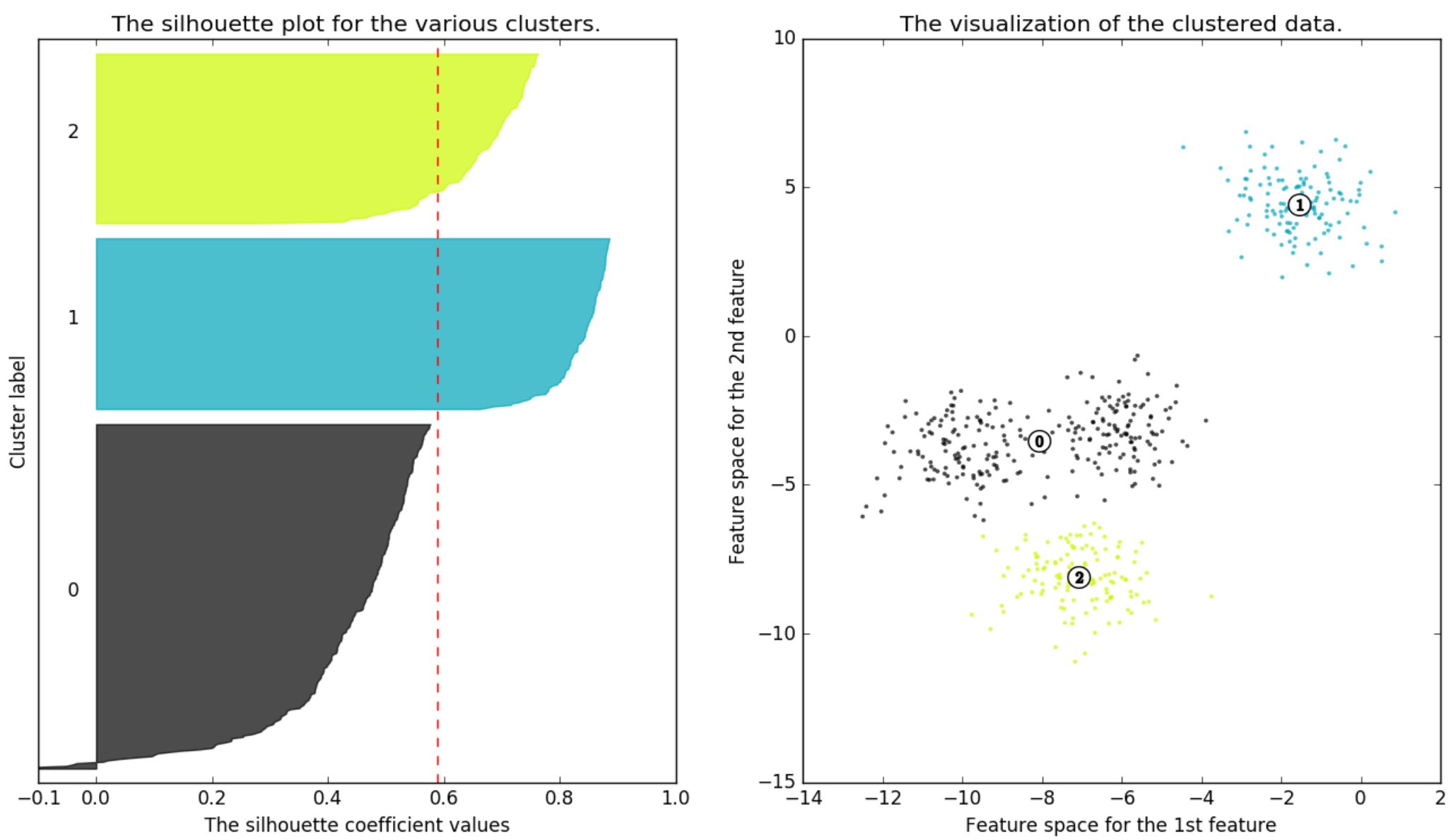
(3) The silhouette coefficient varies between −1 and 1. Analysts hope for positive coefficient values, ideally close to 1, as this would indicate $a(i)$ values close to 0 and perfect internal homogeneity.

(4) By averaging over the cases of a cluster we can calculate its average silhouette coefficient. The overall silhouette coefficient is a measure of the goodness of the clustering solution and can be calculated by taking the average over all records/data points.

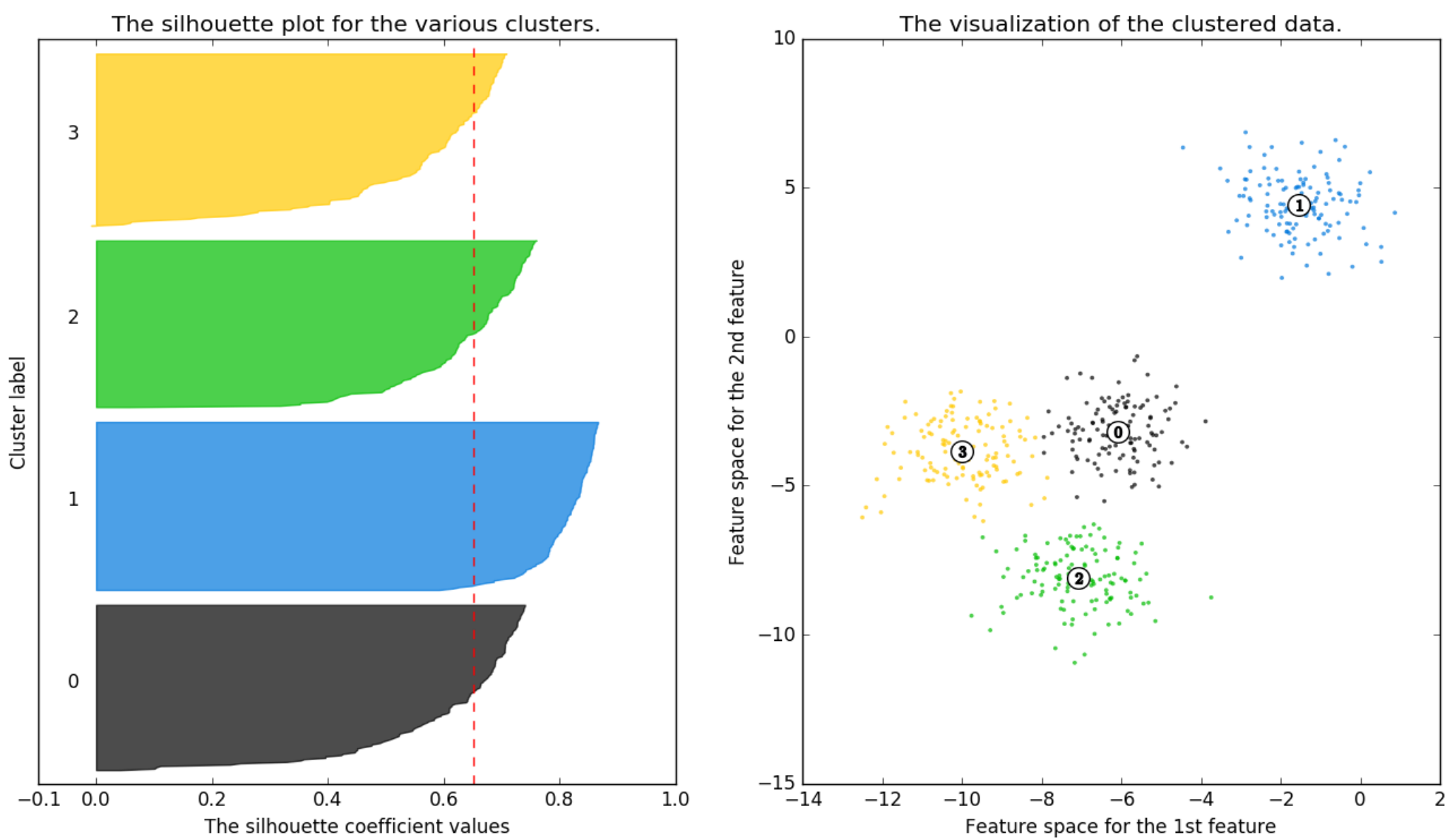
→ An average silhouette coefficient greater than 0.5 indicates reasonable partitioning, whereas a coefficient less than 0.2 denote a problematic solution.



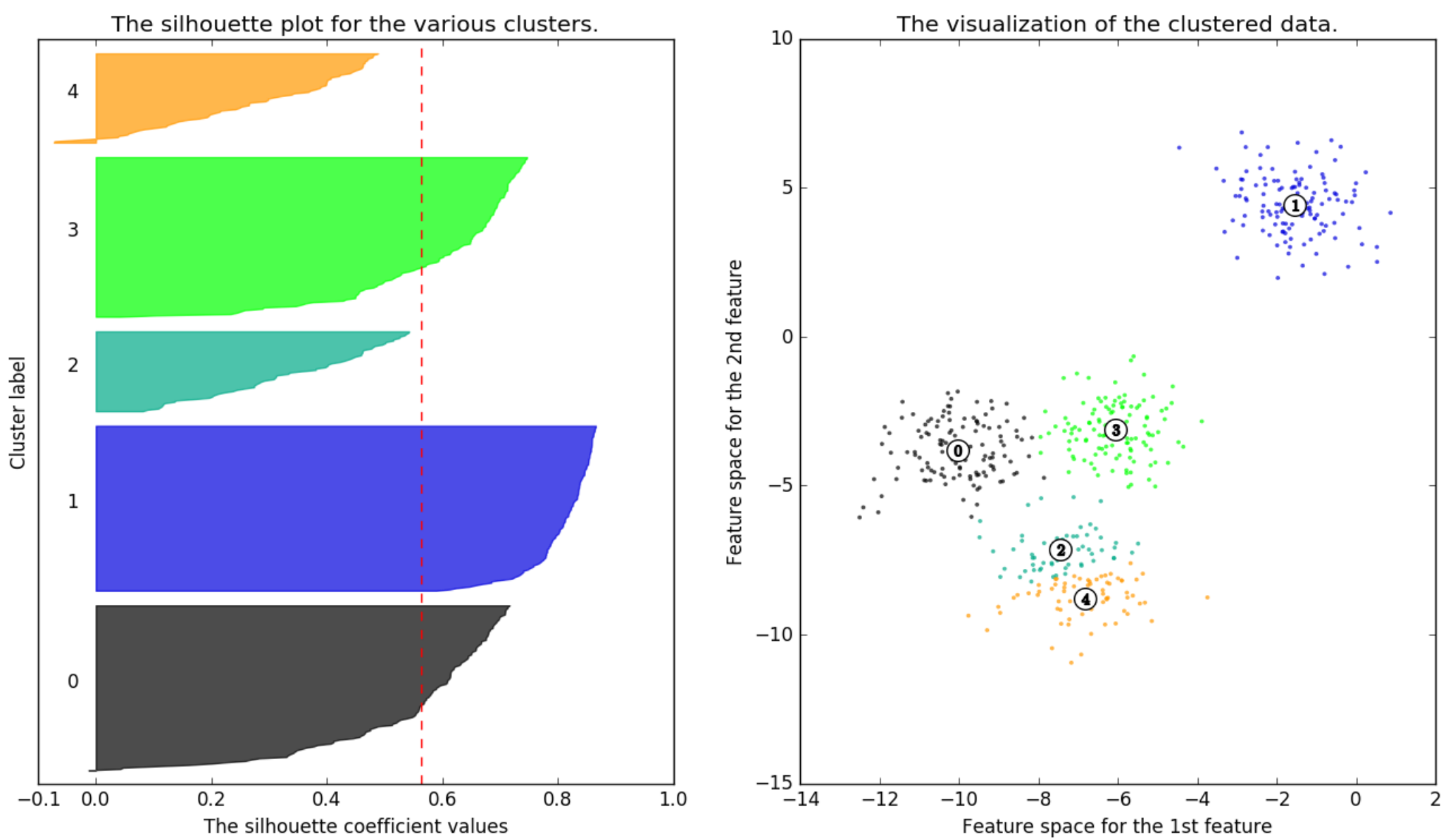
For no. clusters = 2: The average silhouette score is : 0.704978749608



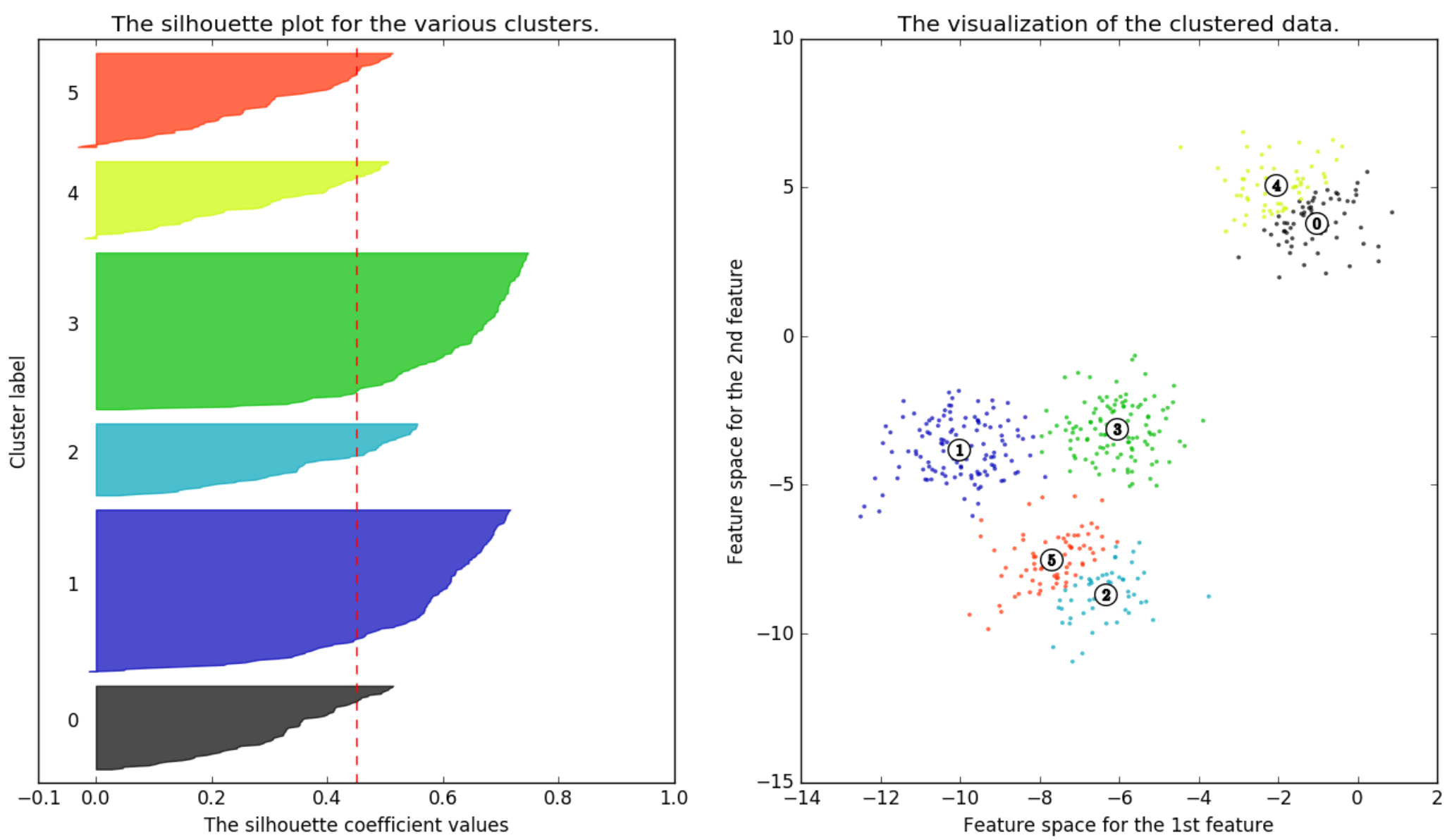
For no. clusters = 3: The average silhouette_score is : 0.588200401213



For no. clusters = 4: The average silhouette score is : 0.650518663273



For no. clusters = 5: The average silhouette score is : 0.563764690262



For no. clusters = 6: The average silhouette score is : 0.450466629437