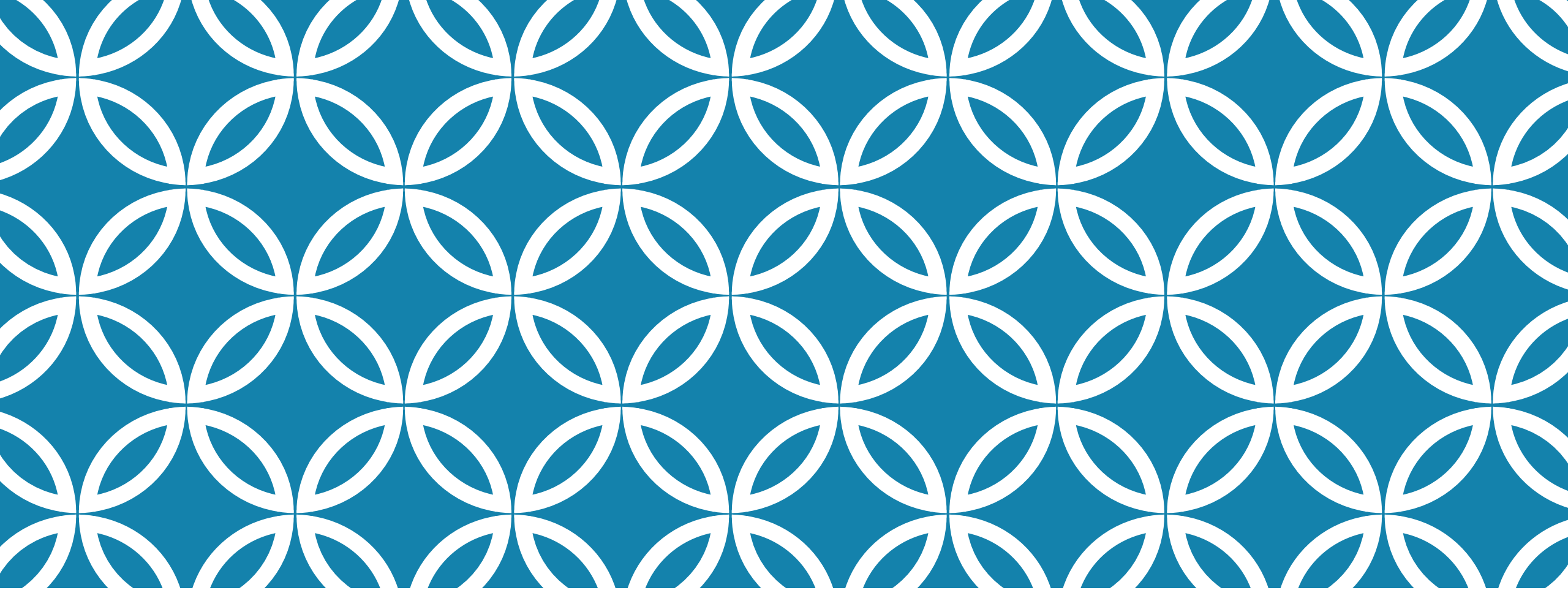


# CLUSTERING | K MEANS

ANSHU PANDEY

# CLUSTERING ANALYSIS



**How many clusters do you expect?**

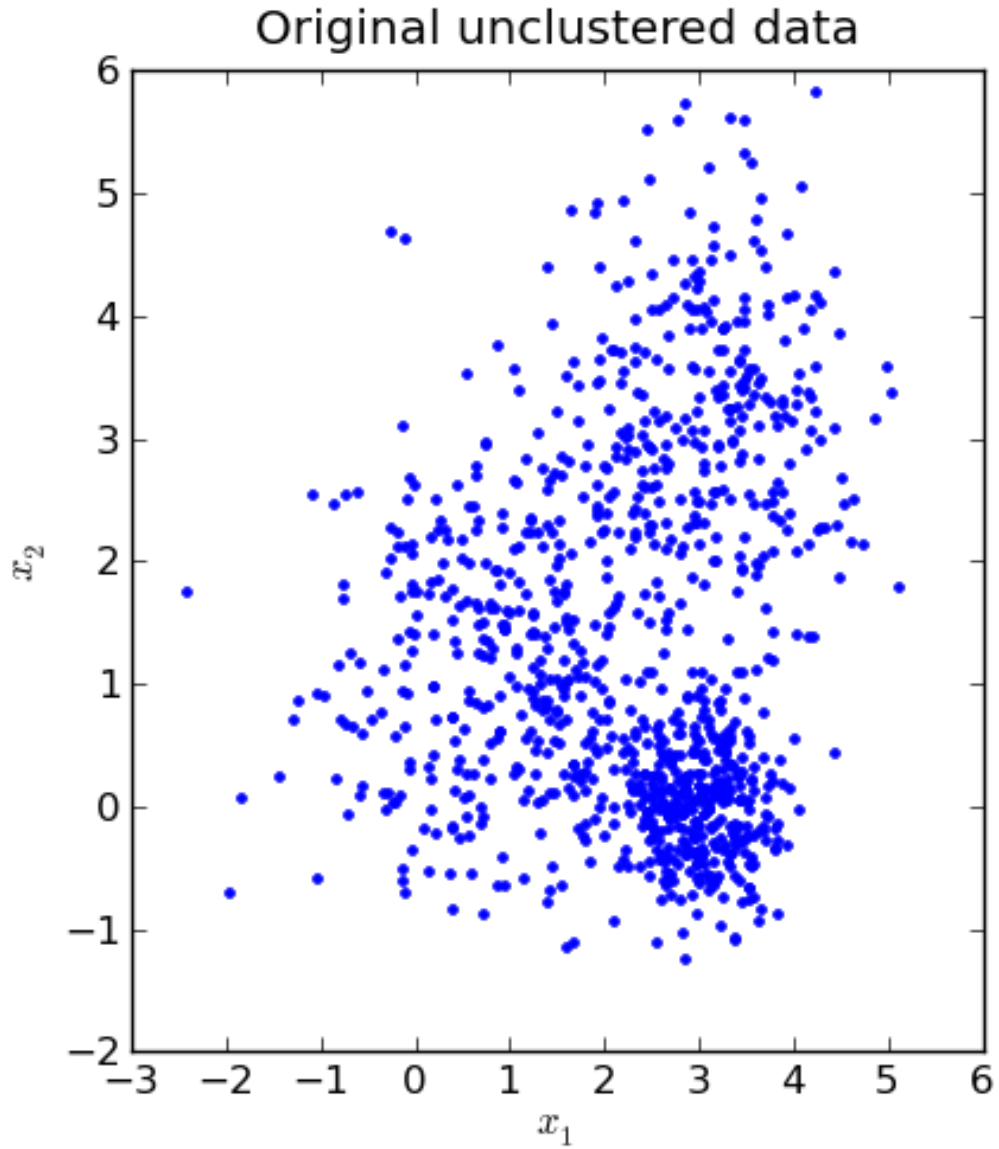

starshadow78/Flickr

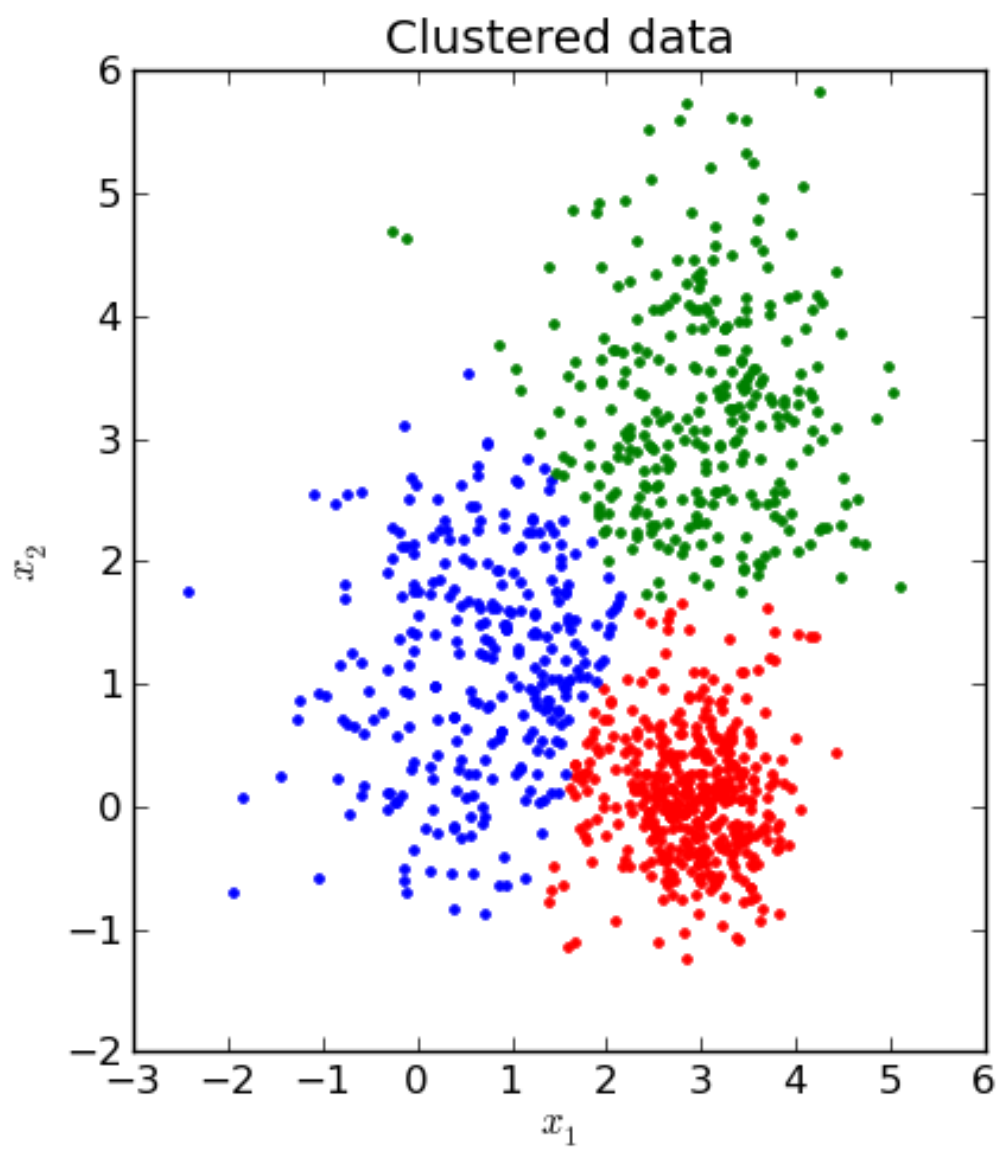

BIG
BLUE
BALL


BIG
BLUE
BALL

# CLUSTERING

Clustering is the categorisation of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

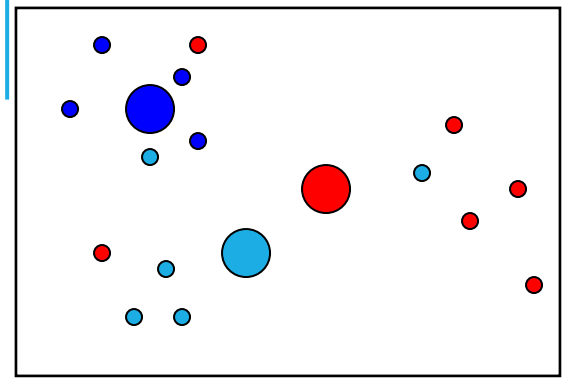

Original unclustered data
$x_1$
$x_2$


Clustered data
$x_1$
$x_2$

# K Means Clustering

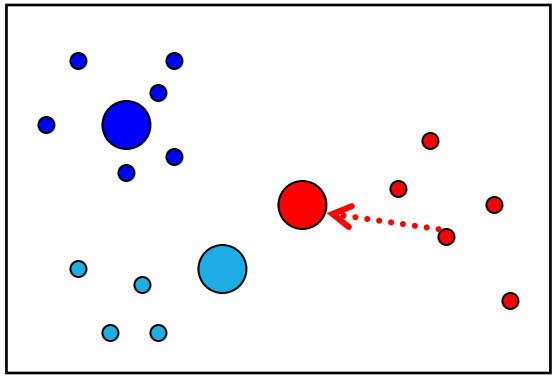
- Choose k initial centroids (center points).
- Each cluster is associated with a centroid.
- Each data object is assigned to closet centroid.
- The centroid of each cluster is then updated based on the data objects assignment to the cluster.
- Repeat the assignment and update steps until convergence.

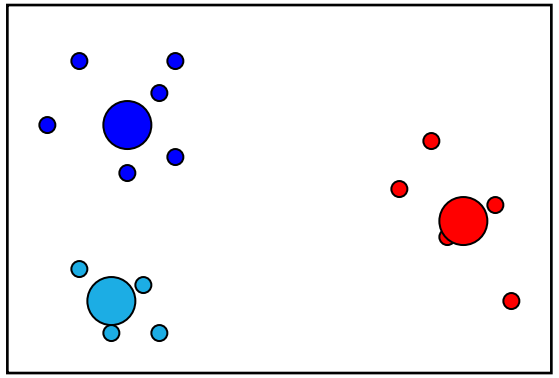
# K-MEANS: EXAMPLE, K = 3



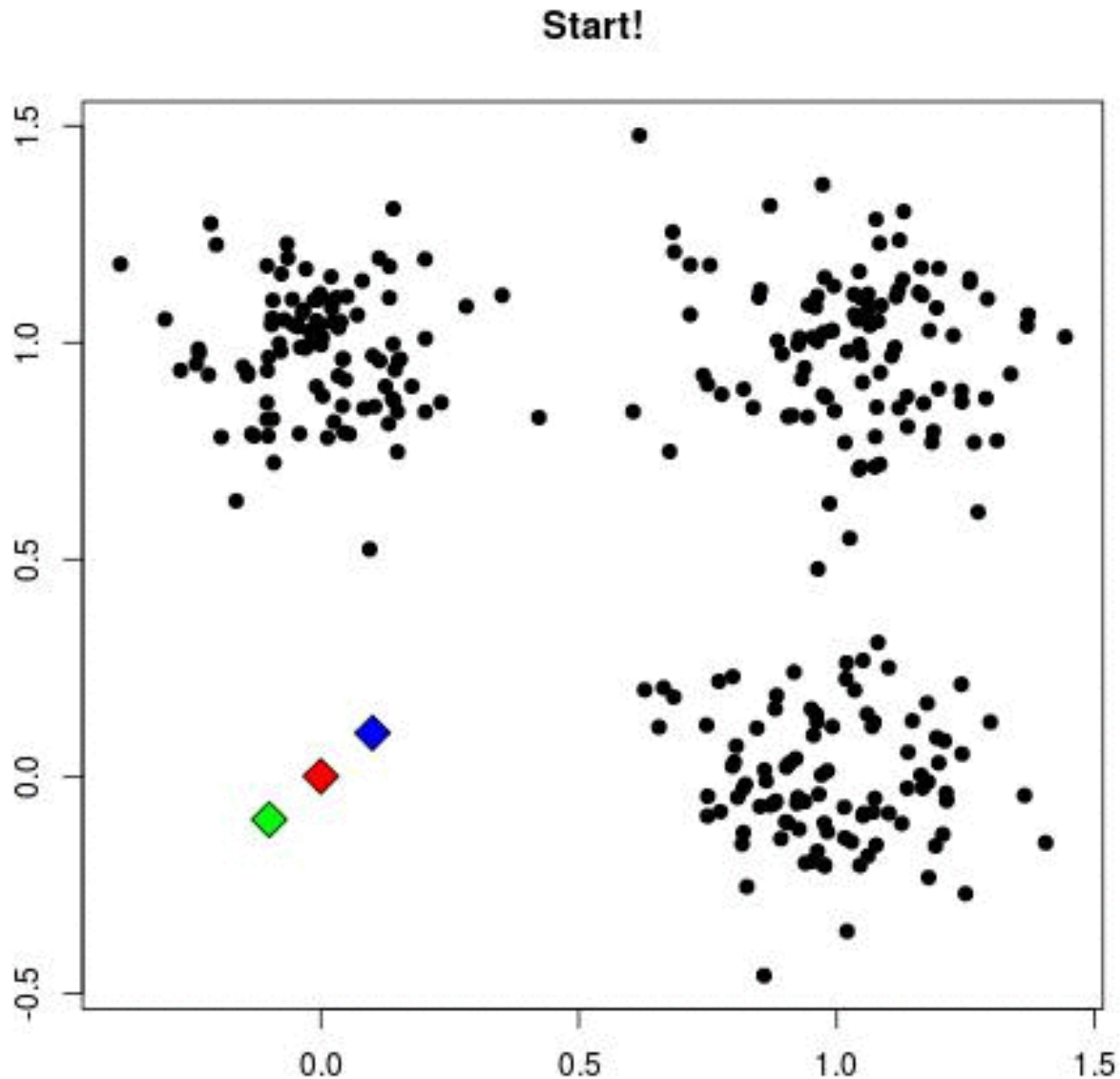
**Step 1:** Make random assignments and compute centroids (big dots)

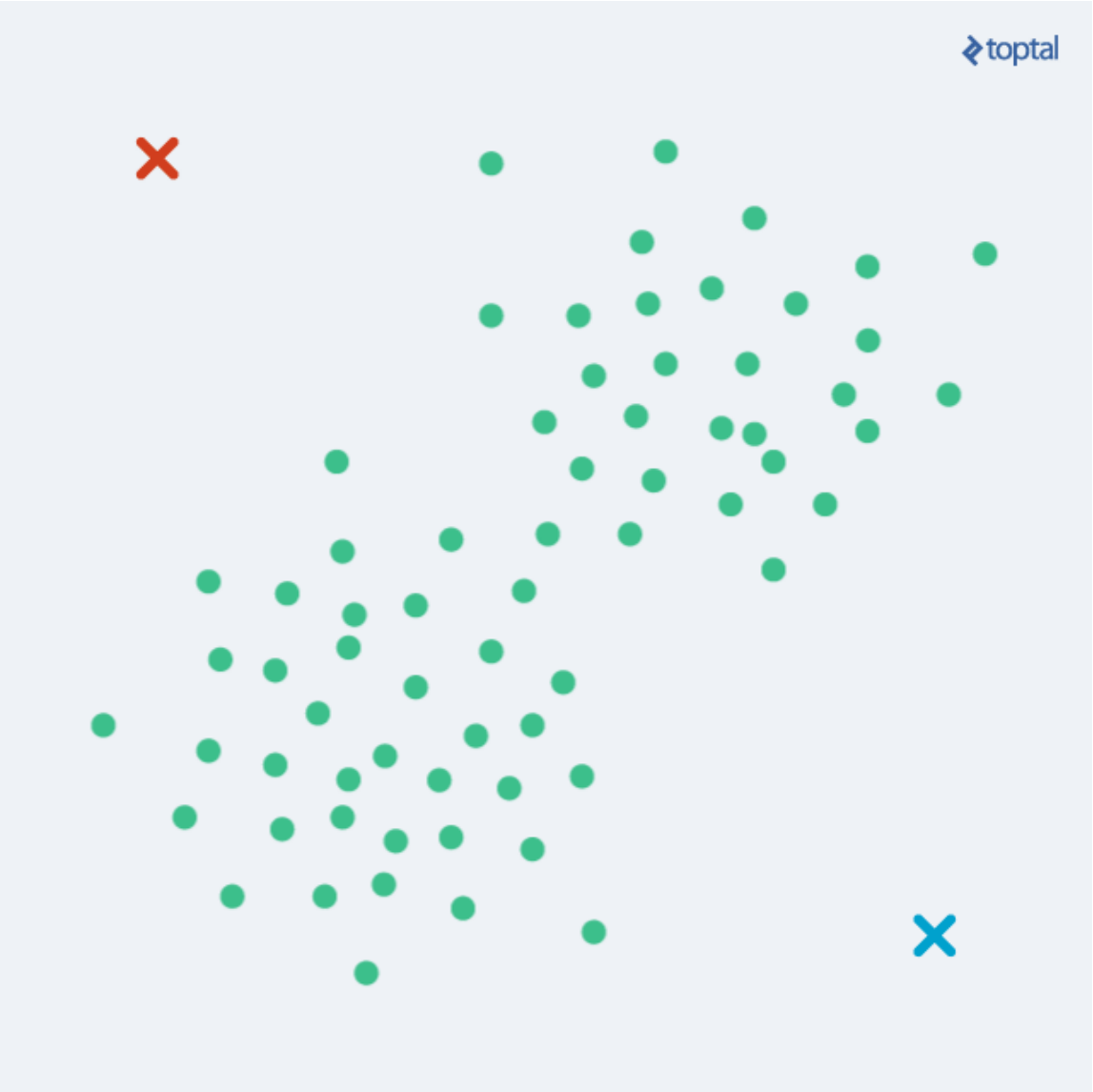
**Step 2:** Assign points to nearest centroids





**Step 3:** Re-compute centroids (in this example, solution is now stable)


Start!
1.5
1.0
0.5
0.0
-0.5
0.0
0.5
1.0
1.5


toptal

# *K*-MEANS ALGORITHM

For a given cluster assignment C of the data points, compute the cluster means $m_k$:

$$m_k = \frac{\sum_{i:C(i)=k} x_i}{N_k}, \; k = 1, \ldots, K.$$

For a current set of cluster means, assign each observation as:

$$C(i) = \arg\min_{1 \le k \le K} \| x_i - m_k \|^2, \; i = 1, \ldots, N$$

Iterate above two steps until convergence

# Applications of Clustering

- Data Mining
- Pattern recognition
- Image analysis
- Bioinformatics
- Voice mining
- Image processing
- Text mining
- Web cluster engines
- Weather report analysis

# IMPLEMENTATION OF K-MEANS ALGORITHM (K=2)

| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

# STEP 1 INITIALIZATION

Randomly we choose 2 centroids (k=2) for two clusters.

In this case the two centroids are m1=(1.0,1.0) and m2=(5.0,7.0)

| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

# STEP 2/1 FINDING THE NEAREST CENTROID FOR EVERY ELEMENT

For every element calculating its distance from center (Euclidian Distance)

| Individual | (v1,v2) | C1 | C2 |
|---|---|---|---|
| 1 | (1.0,1.0) | 0 | 7.12 |
| 2 | (1.5,2.0) | 1.12 | 6.10 |
| 3 | (3.0,4.0) | 3.61 | 3.61 |
| 4 | (5.0,7.0) | 7.21 | 0 |
| 5 | (3.5,5.0) | 4.72 | 2.5 |
| 6 | (4.5,5.0) | 5.31 | 2.06 |
| 7 | (3.5,4.5) | 4.30 | 2.92 |

# STEP 2/2 ASSIGNING ELEMENTS TO ANY OF THE CLUSTERS

Thus we obtain 2 clusters containing {1,2,3} and {4,5,6,7}

New centroids are

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

| Individual | (v1,v2) | C1 | C2 |
|---|---|---|---|
| 1 | (1.0,1.0) | 0 | 7.12 |
| 2 | (1.5,2.0) | 1.12 | 6.10 |
| 3 | (3.0,4.0) | 3.61 | 3.61 |
| 4 | (5.0,7.0) | 7.21 | 0 |
| 5 | (3.5,5.0) | 4.72 | 2.5 |
| 6 | (4.5,5.0) | 5.31 | 2.06 |
| 7 | (3.5,4.5) | 4.30 | 2.92 |

# STEP 3 ASSIGNING ELEMENTS TO NEW CLUSTERS ACCORDING TO DISTANCE

Now using these centroids m1=(1.83,2.33) & m2=(4.12,5.38) we compute the Euclidean distance of each object, as shown in table.

| Individual | (v1,v2) | C1 | C2 |
|---|---|---|---|
| 1 | (1.0,1.0) | 1.57 | 5.38 |
| 2 | (1.5,2.0) | 0.47 | 4.28 |
| 3 | (3.0,4.0) | 2.04 | 1.78 |
| 4 | (5.0,7.0) | 5.64 | 1.84 |
| 5 | (3.5,5.0) | 3.15 | 0.73 |
| 6 | (4.5,5.0) | 3.78 | 0.54 |
| 7 | (3.5,4.5) | 2.74 | 1.08 |

# STEP 3 ASSIGNING ELEMENTS TO NEW CLUSTERS ACCORDING TO DISTANCE

Therefore, new clusters are:

{1,2} and {3,4,5,6,7}

Next centroids are:
m1=(1.25,1.5) and m2 = (3.9,5.1)

| Individual | (v1,v2) | C1 | C2 |
|---|---|---|---|
| 1 | (1.0,1.0) | 1.57 | 5.38 |
| 2 | (1.5,2.0) | 0.47 | 4.28 |
| 3 | (3.0,4.0) | 2.04 | 1.78 |
| 4 | (5.0,7.0) | 5.64 | 1.84 |
| 5 | (3.5,5.0) | 3.15 | 0.73 |
| 6 | (4.5,5.0) | 3.78 | 0.54 |
| 7 | (3.5,4.5) | 2.74 | 1.08 |

## STEP 4

Now using these centroids m1=(1.25,1.5) & m2 = (3.9,5.1) we compute the Euclidean distance of each object, as shown in table.

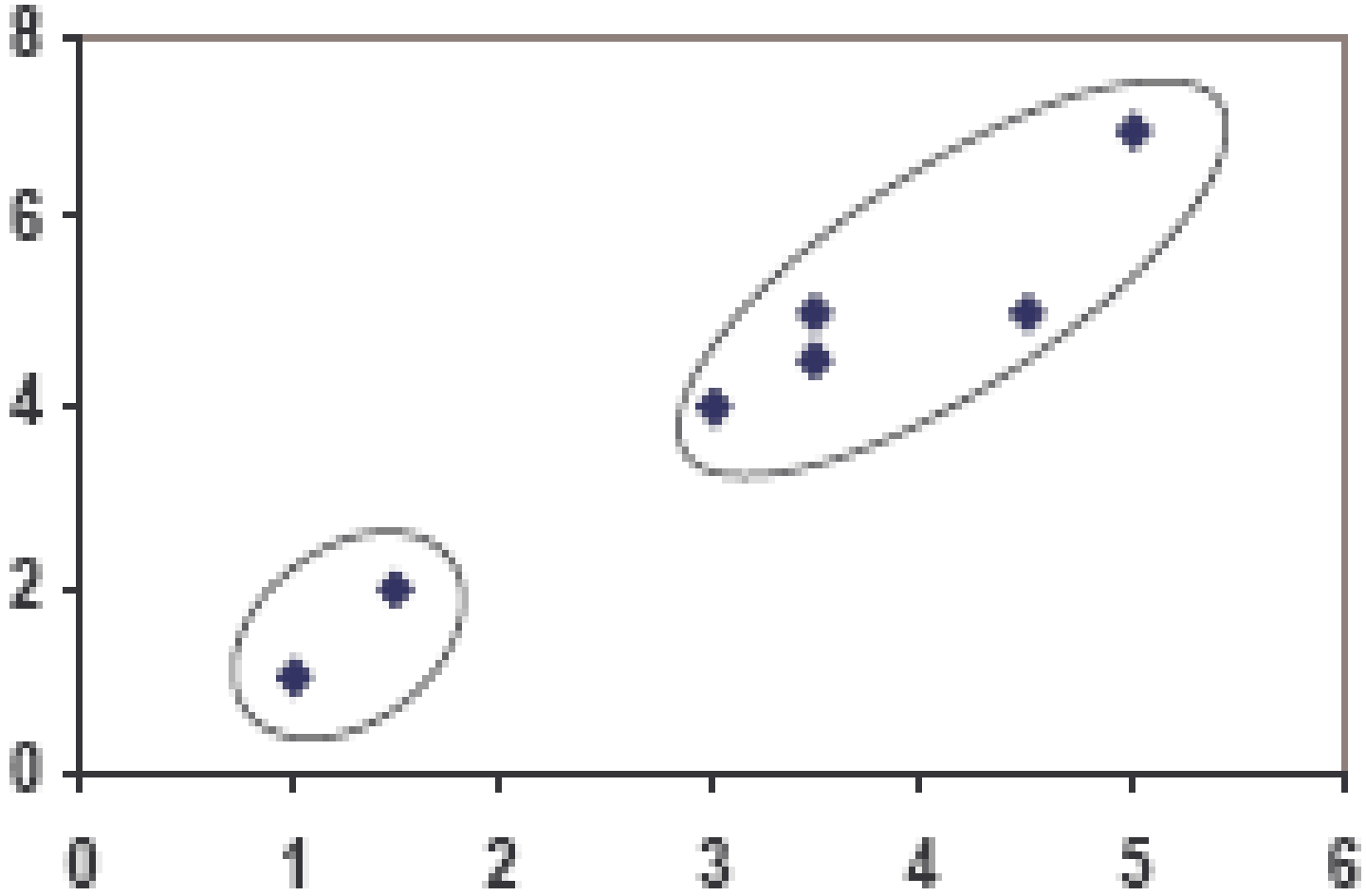
Therefore, new clusters are:

{1,2} and {**3**,4,5,6,7}

Therefore, there is no change in the cluster.

| Individual | (v1,v2) | C1 | C2 |
|---|---|---|---|
| 1 | (1.0,1.0) | 0.56 | 5.02 |
| 2 | (1.5,2.0) | 0.56 | 3.92 |
| 3 | (3.0,4.0) | 3.05 | 1.42 |
| 4 | (5.0,7.0) | 6.66 | 2.20 |
| 5 | (3.5,5.0) | 4.16 | 0.41 |
| 6 | (4.5,5.0) | 4.78 | 0.61 |
| 7 | (3.5,4.5) | 3.75 | 0.72 |

Thus, algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.


8
6
4
2
0
0
1
2
3
4
5
6

# HOW TO CHOOSE K?

Use another clustering method, like EM.

Run algorithm on data with several different values of K.

Use the prior knowledge about the characteristics of the problem.

# APPLICATIONS OF K-MEAN CLUSTERING

It is relatively efficient and fast. It computes result at **O(tkn),** where n is number of objects or points, k is number of clusters and t is number of iterations.

k-means clustering can be applied to machine learning or data mining

Used on acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization or Image Segmentation).

Also used for choosing color palettes on old fashioned graphical display devices and Image Quantization.

# CONCLUSION

K-means algorithm is useful for undirected knowledge discovery and is relatively simple.

K-means has found wide spread usage in lot of fields, ranging from unsupervised learning of neural network, Pattern recognitions, Classification analysis, Artificial intelligence, image processing, machine vision, and many others.