Machine Learning

By Ghazal Lalooha

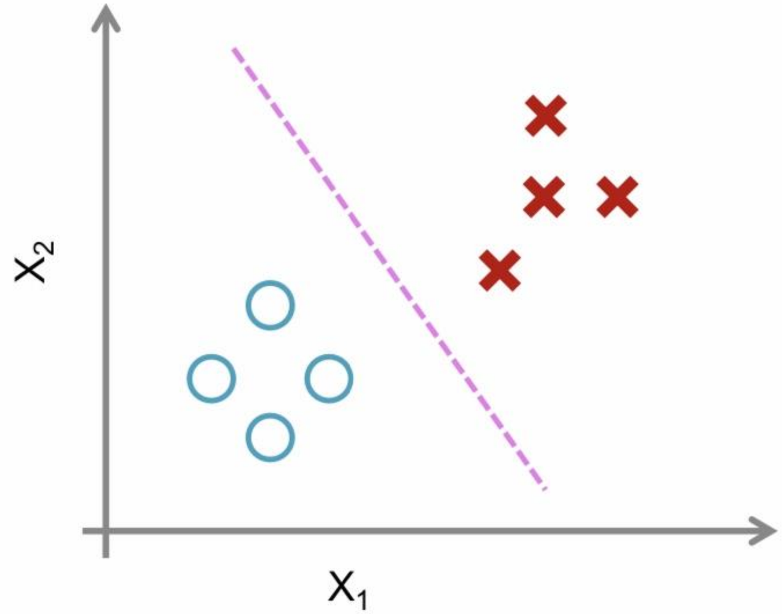
Unsupervised Learning

Table of Contents

- Unsupervised learning
- applications
- Clustering
- K-means algorithm
- Improved clustering
- Two-part generator algorithm
- Hierarchical clustering

Supervised Learning

- Supervised Learning: For each example, the correct answer is given.



Training set

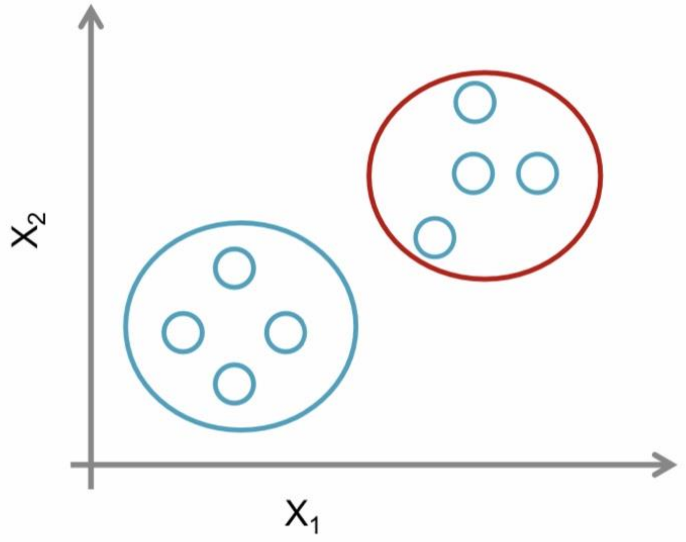
$$\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),(x^{(3)},y^{(3)}),\cdots,(x^{(m)},y^{(m)})\}$$

Unsupervised Learning

- Unsupervised learning: Not knowing the correct answers.

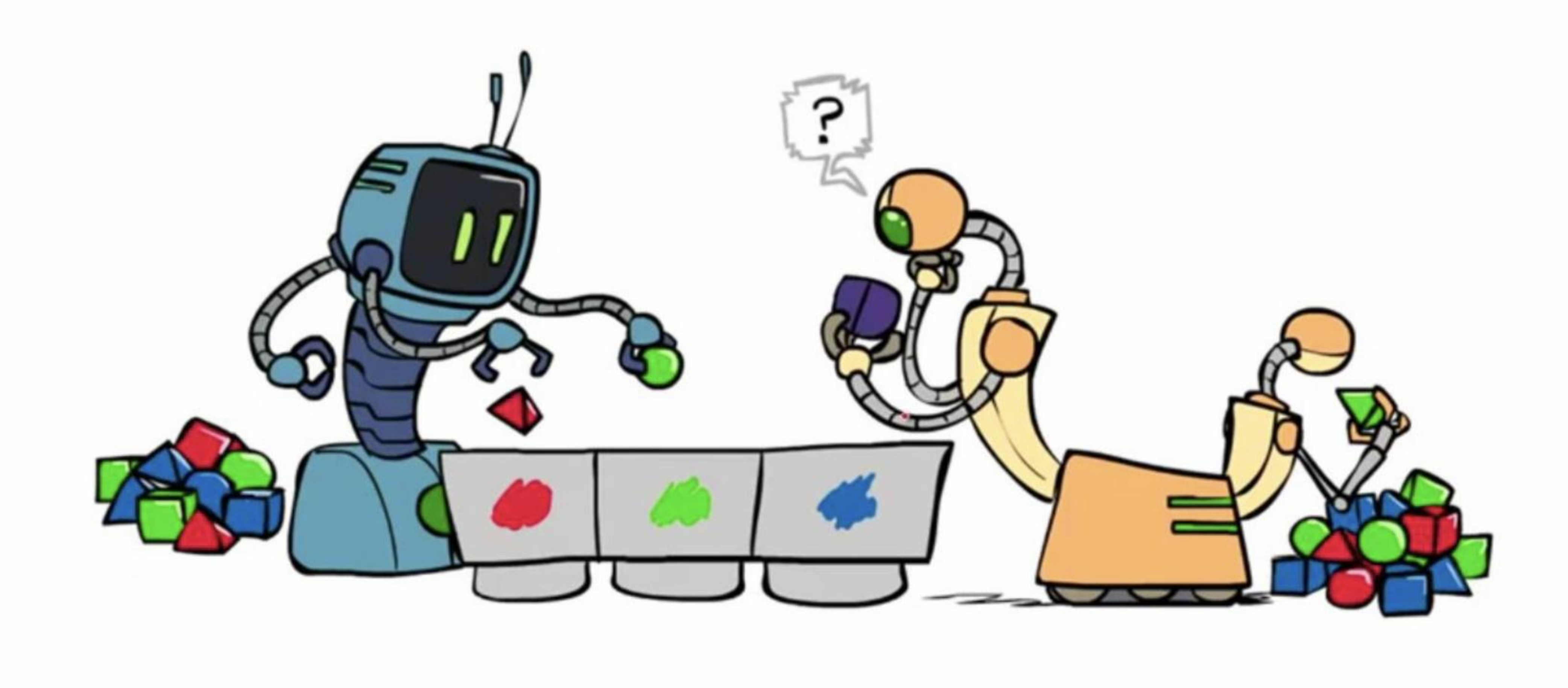
$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

Training set

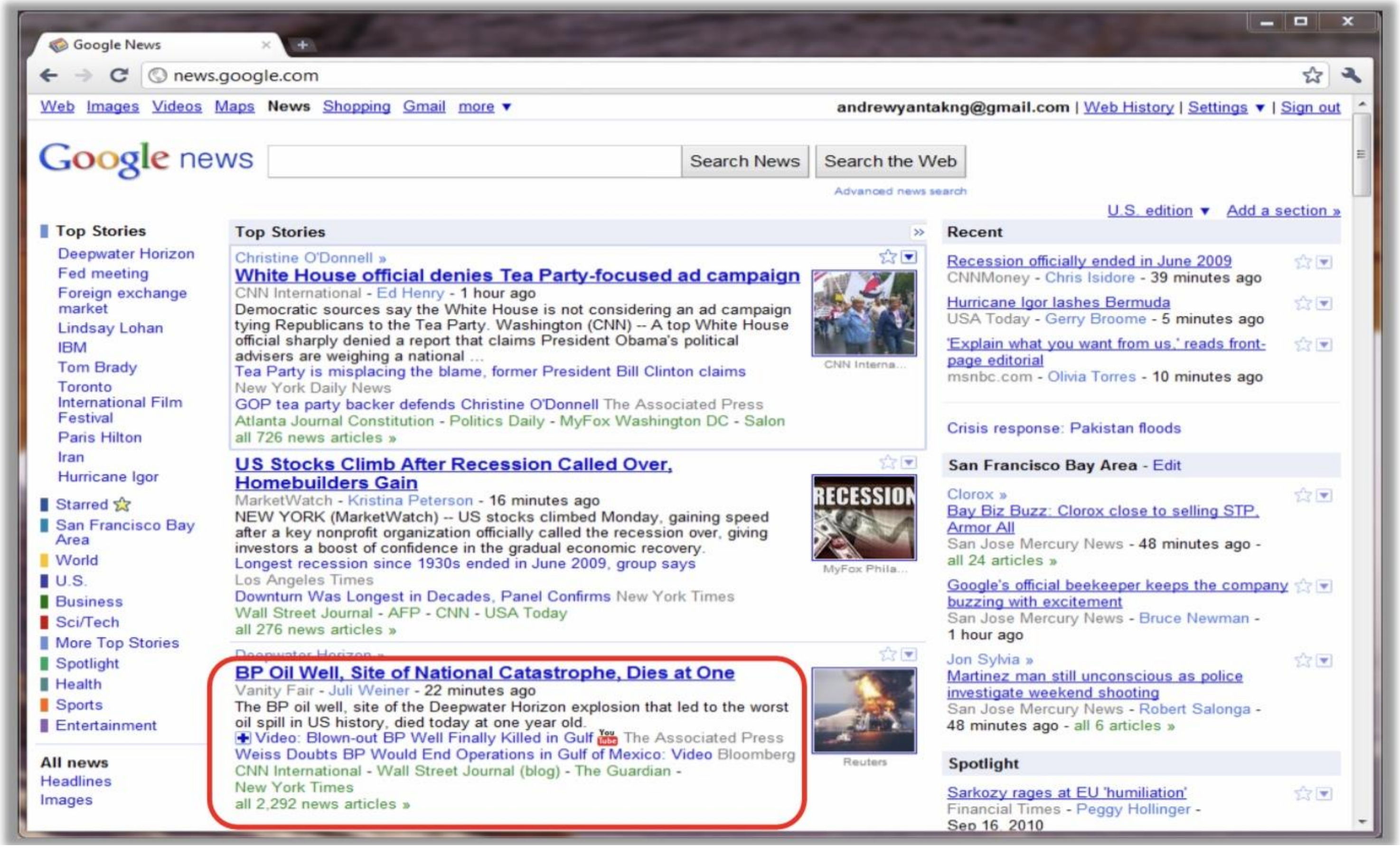


- Goal: recognizing the structure in the input data (grouping similar data)

Clustering



Application of clustering: grouping related news


Google News
news.google.com
Web Images Videos Maps News Shopping Gmail more ▾
andrewyantakng@gmail.com | Web History | Settings ▾ | Sign out
Google news
Search News
Search the Web
Advanced news search
U.S. edition ▾ Add a section »

Top Stories
Deepwater Horizon
Fed meeting
Foreign exchange market
Lindsay Lohan
IBM
Tom Brady
Toronto International Film Festival
Paris Hilton
Iran
Hurricane Igor

Starred ☆
San Francisco Bay Area
World
U.S.
Business
Sci/Tech
More Top Stories
Spotlight
Health
Sports
Entertainment

All news
Headlines
Images

Top Stories

Christine O'Donnell »

White House official denies Tea Party-focused ad campaign

CNN International - Ed Henry - 1 hour ago

Democratic sources say the White House is not considering an ad campaign tying Republicans to the Tea Party. Washington (CNN) -- A top White House official sharply denied a report that claims President Obama's political advisers are weighing a national ...

Tea Party is misplacing the blame, former President Bill Clinton claims New York Daily News

GOP tea party backer defends Christine O'Donnell The Associated Press

Atlanta Journal Constitution - Politics Daily - MyFox Washington DC - Salon

all 726 news articles »

CNN Interna...

US Stocks Climb After Recession Called Over, Homebuilders Gain

MarketWatch - Kristina Peterson - 16 minutes ago

NEW YORK (MarketWatch) -- US stocks climbed Monday, gaining speed after a key nonprofit organization officially called the recession over, giving investors a boost of confidence in the gradual economic recovery.

Longest recession since 1930s ended in June 2009, group says Los Angeles Times

Downturn Was Longest in Decades, Panel Confirms New York Times

Wall Street Journal - AFP - CNN - USA Today

all 276 news articles »

MyFox Phila...

Deepwater Horizon »

BP Oil Well, Site of National Catastrophe, Dies at One

Vanity Fair - Juli Weiner - 22 minutes ago

The BP oil well, site of the Deepwater Horizon explosion that led to the worst oil spill in US history, died today at one year old.

Video: Blown-out BP Well Finally Killed in Gulf The Associated Press

Weiss Doubts BP Would End Operations in Gulf of Mexico: Video Bloomberg

CNN International - Wall Street Journal (blog) - The Guardian - New York Times

all 2,292 news articles »

Reuters

Recent

Recession officially ended in June 2009
CNNMoney - Chris Isidore - 39 minutes ago

Hurricane Igor lashes Bermuda
USA Today - Gerry Broome - 5 minutes ago

'Explain what you want from us,' reads front-page editorial
msnbc.com - Olivia Torres - 10 minutes ago

Crisis response: Pakistan floods

San Francisco Bay Area - Edit

Clorox »
Bay Biz Buzz: Clorox close to selling STP, Armor All
San Jose Mercury News - 48 minutes ago -
all 24 articles »

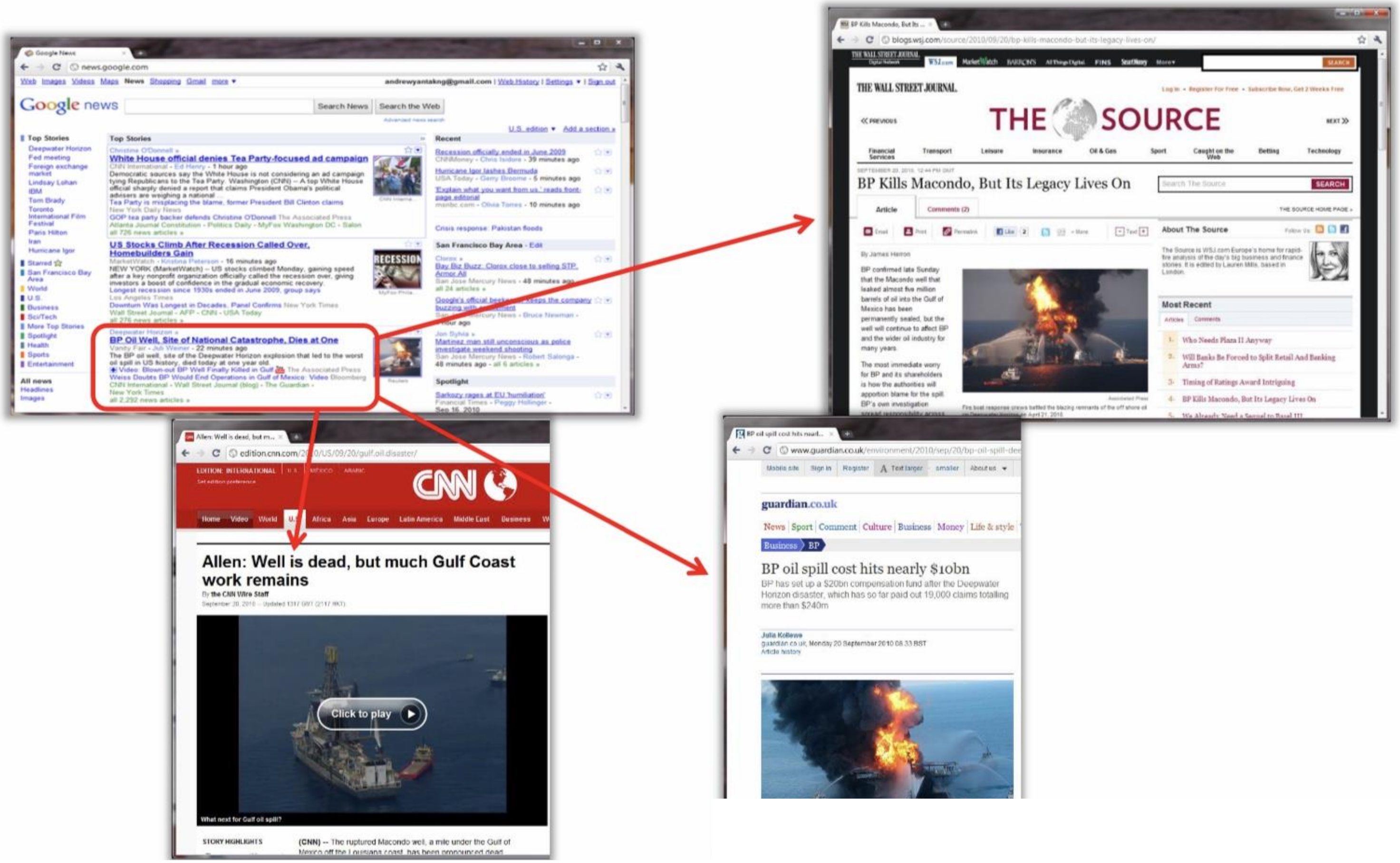
Google's official beekeeper keeps the company buzzing with excitement
San Jose Mercury News - Bruce Newman - 1 hour ago

Jon Sylvia »
Martinez man still unconscious as police investigate weekend shooting
San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »

Spotlight

Sarkozy rages at EU 'humiliation'
Financial Times - Peggy Hollinger -
Sep 16, 2010

Application of clustering: grouping related news



Google news

BP Oil Well, Site of National Catastrophe, Dies at One

THE WALL STREET JOURNAL. THE SOURCE

BP Kills Macondo, But Its Legacy Lives On

CNN

Allen: Well is dead, but much Gulf Coast work remains

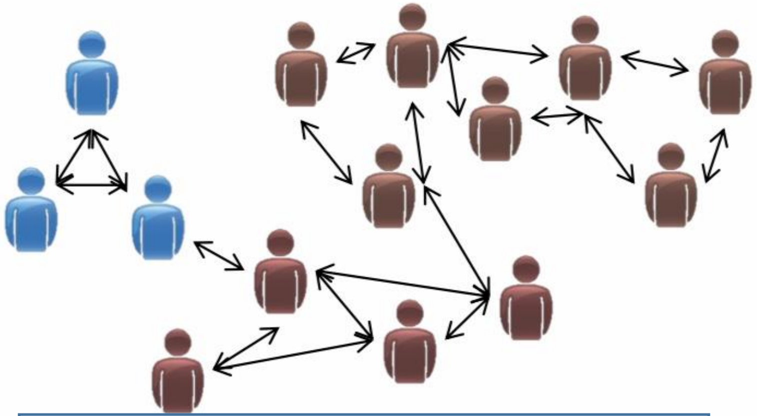
guardian.co.uk

BP oil spill cost hits nearly $10bn

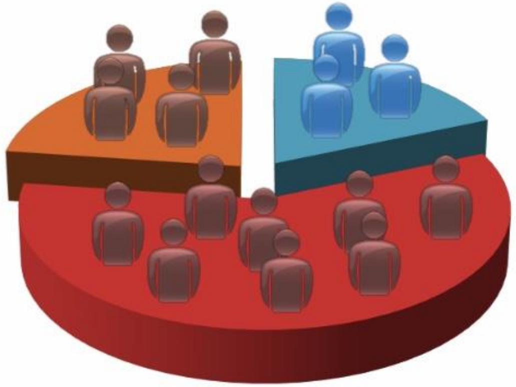
BP has set up a $20bn compensation fund after the Deepwater Horizon disaster, which has so far paid out 19,000 claims totalling more than $240m

Some other applications of unsupervised learning

Organization of computing clusters (data center)



Social networks analysis

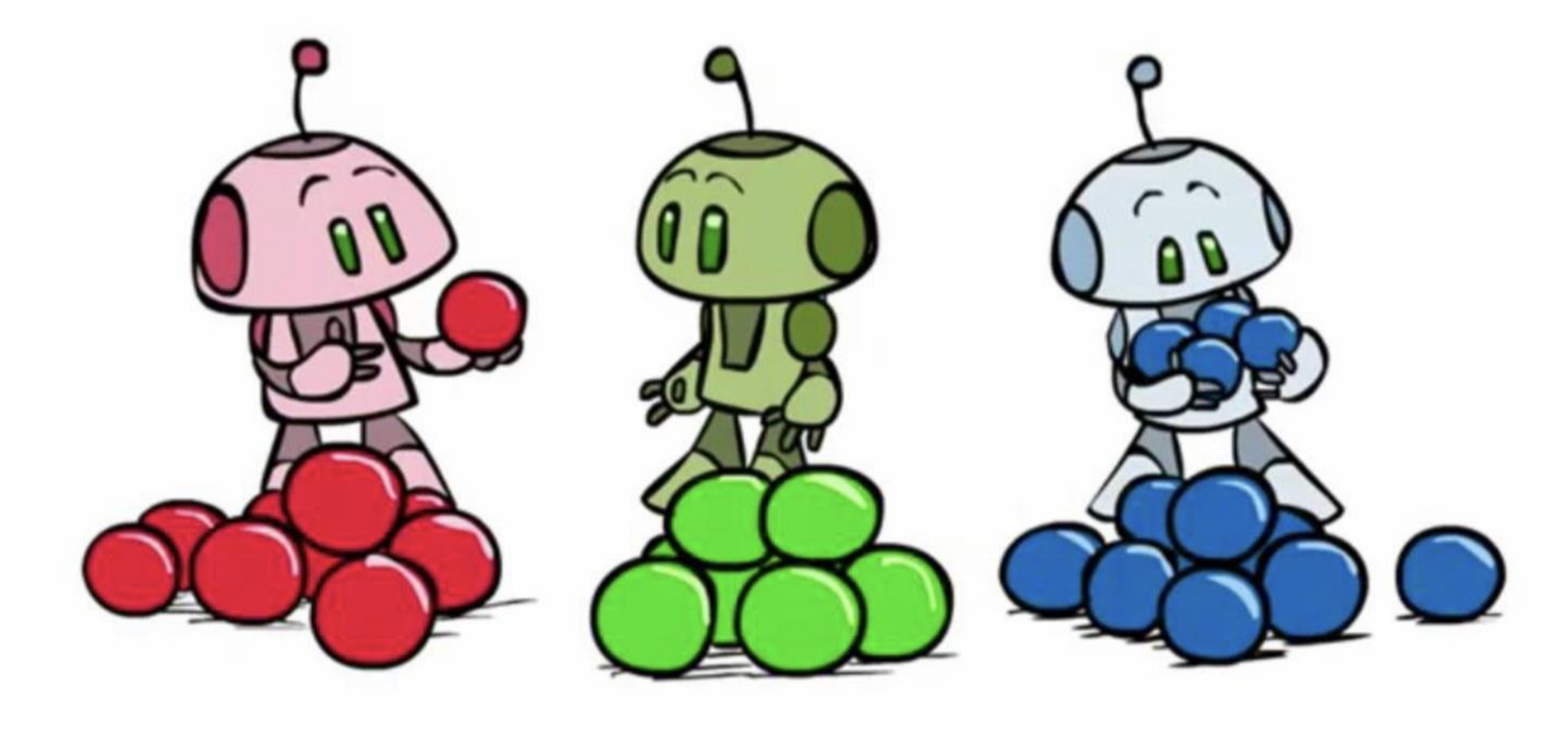


Market segmentation

Image credit: NASA/JPL-Caltech/E. Churchwell (Univ. of Wisconsin, Madison)

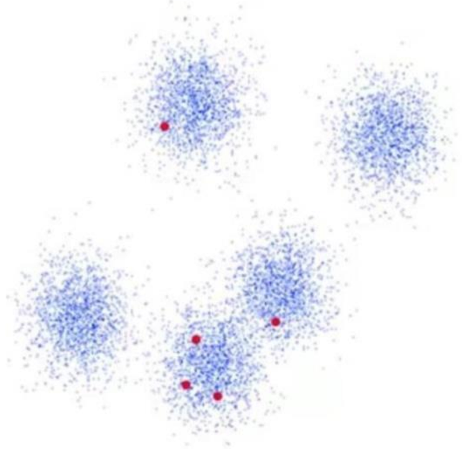
Analysis of astronomical data (how galaxies form)

K-means clustering algorithm



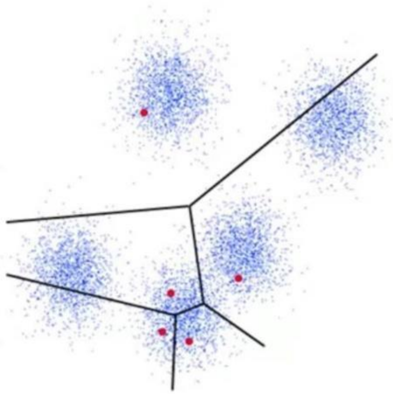
K-means clustering algorithm

- An iterative clustering algorithm:
  - Choose K points randomly as centers of clusters.
  - Repeat the following steps:
    - Assign each data to a cluster with the closest center.
    - Update the center of each cluster by averaging the data assigned to that cluster.
  - Stop: when no data changes its cluster in an iteration.



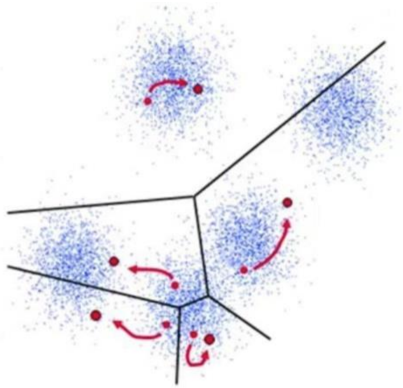
K-means clustering algorithm

- An iterative clustering algorithm:
  - Choose K points randomly as centers of clusters.
  - Repeat the following steps:
    - Assign each data to a cluster with the closest center.
    - Update the center of each cluster by averaging the data assigned to that cluster.
  - Stop: when no data changes its cluster in an iteration.

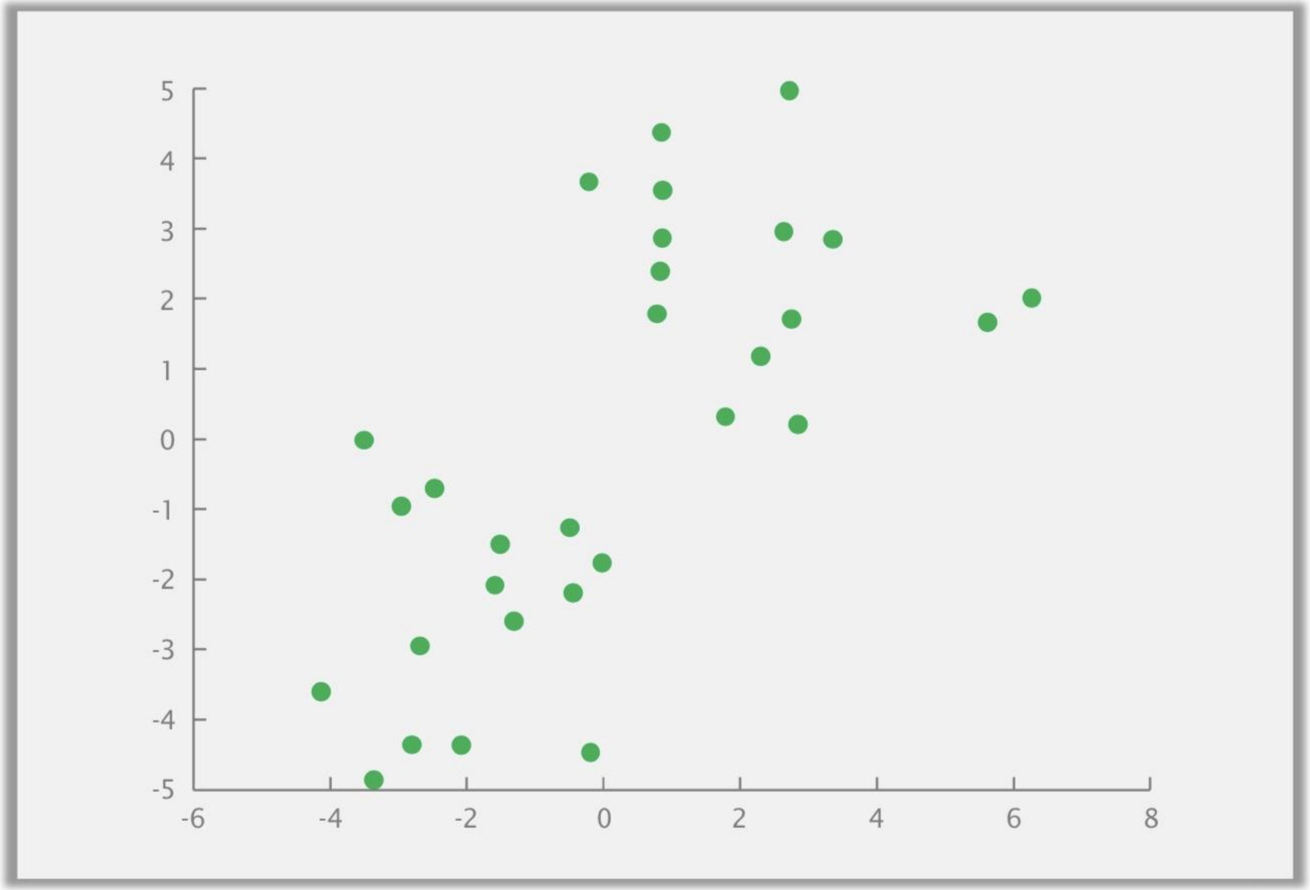


K-means clustering algorithm

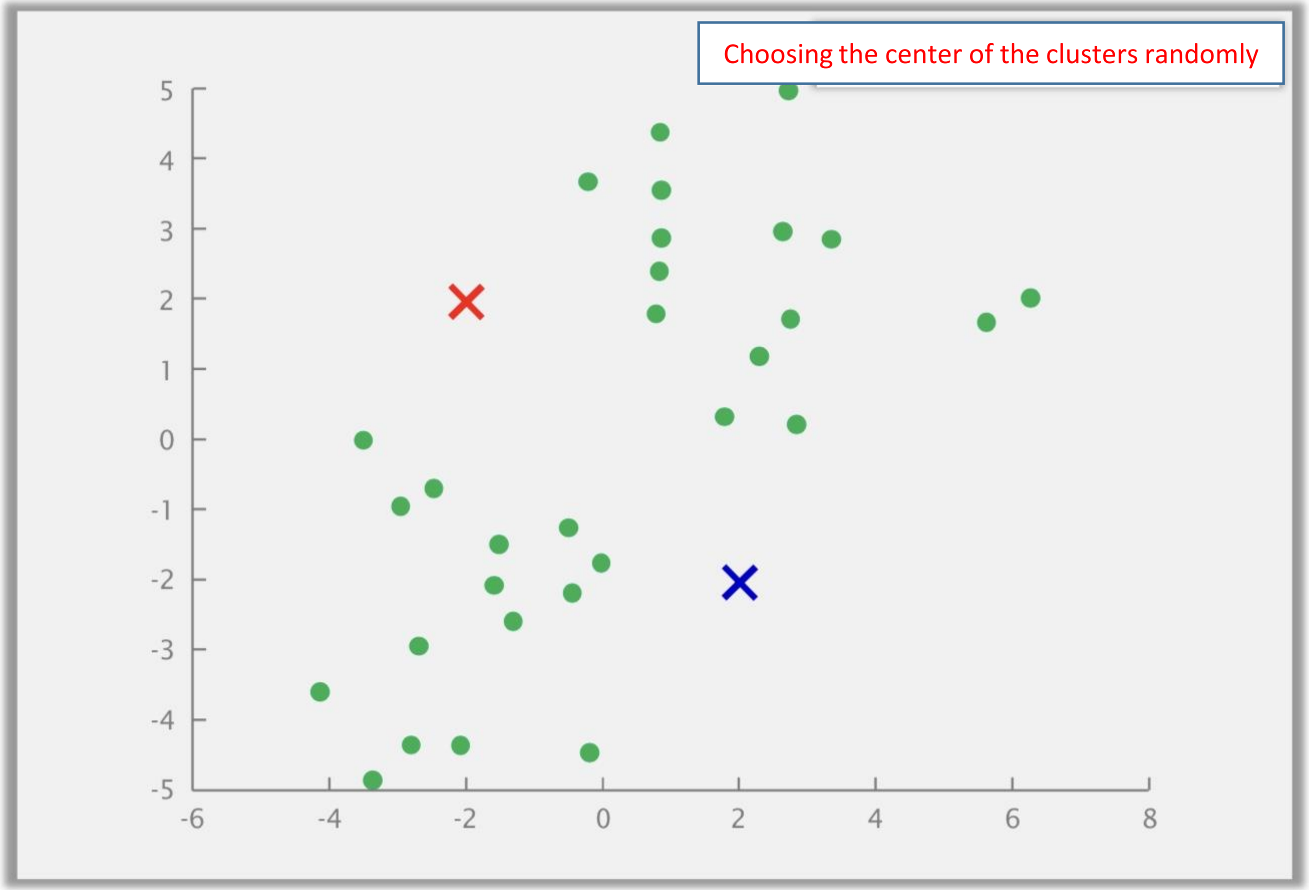
- An iterative clustering algorithm:
  - Choose K points randomly as centers of clusters.
  - Repeat the following steps:
    - Assign each data to a cluster with the closest center.
    - Update the center of each cluster by averaging the data assigned to that cluster.
  - Stop: when no data changes its cluster in an iteration.



K-means algorithm: demonstration implementation

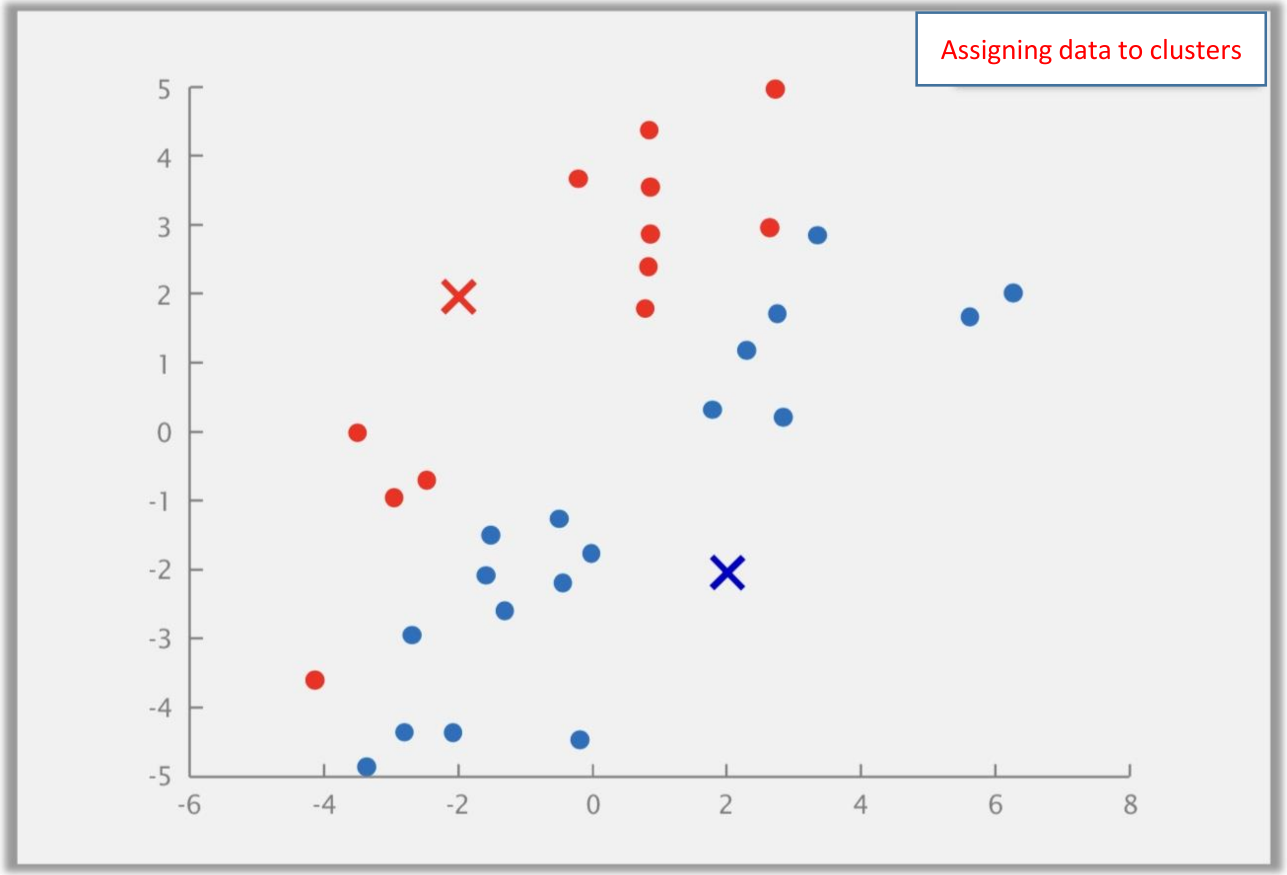


K-means algorithm: demonstration implementation

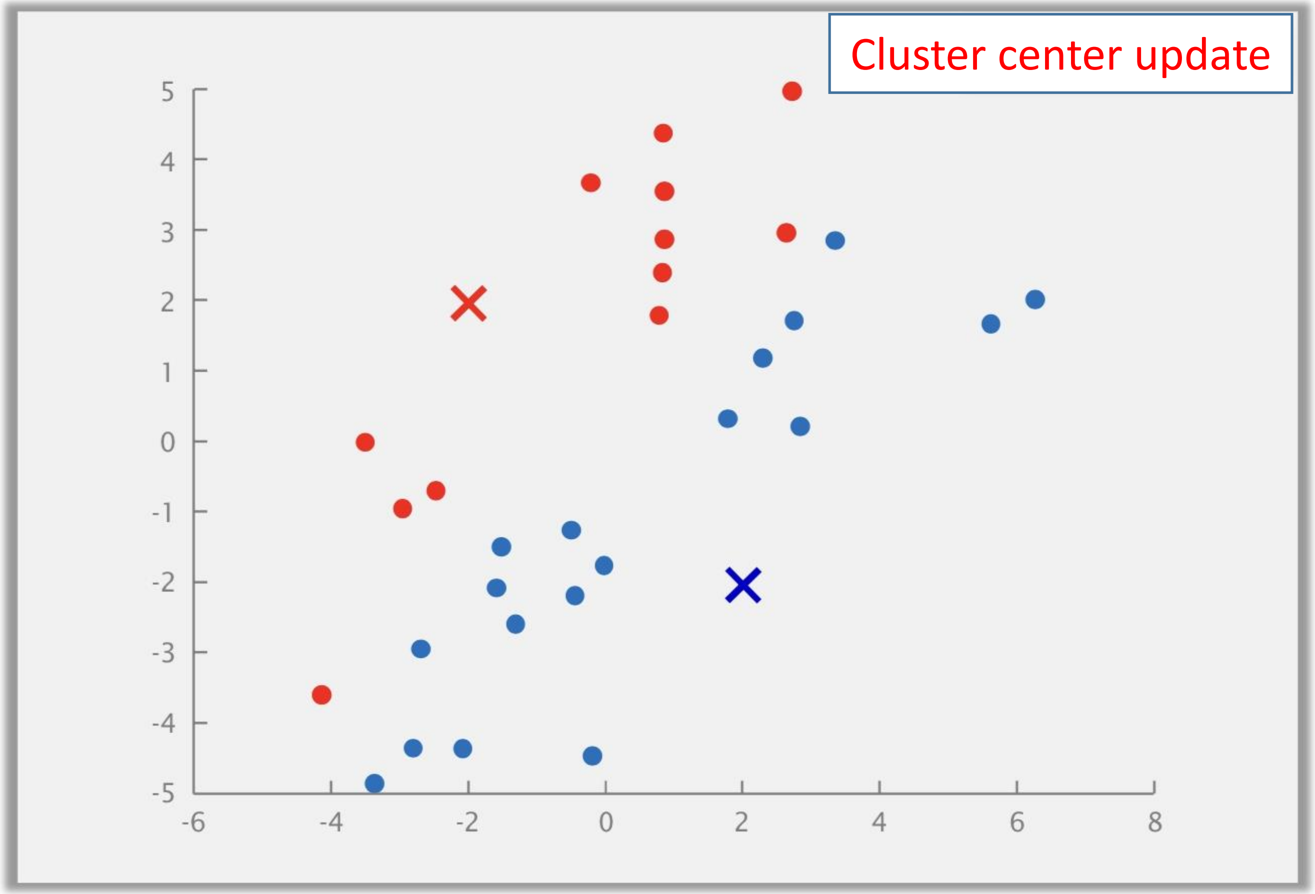


Choosing the center of the clusters randomly

K-means algorithm: demonstration implementation

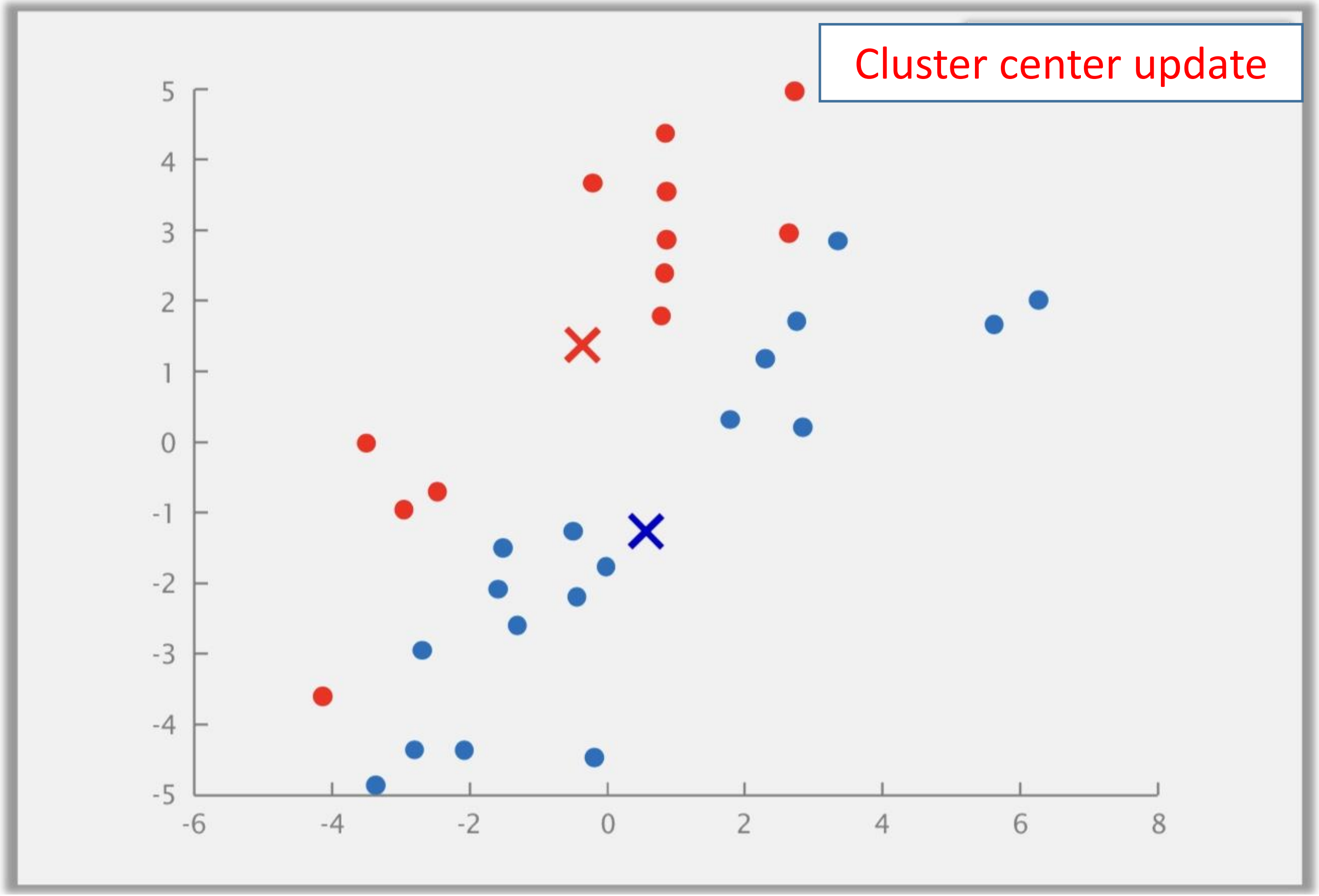


K-means algorithm: demonstration implementation

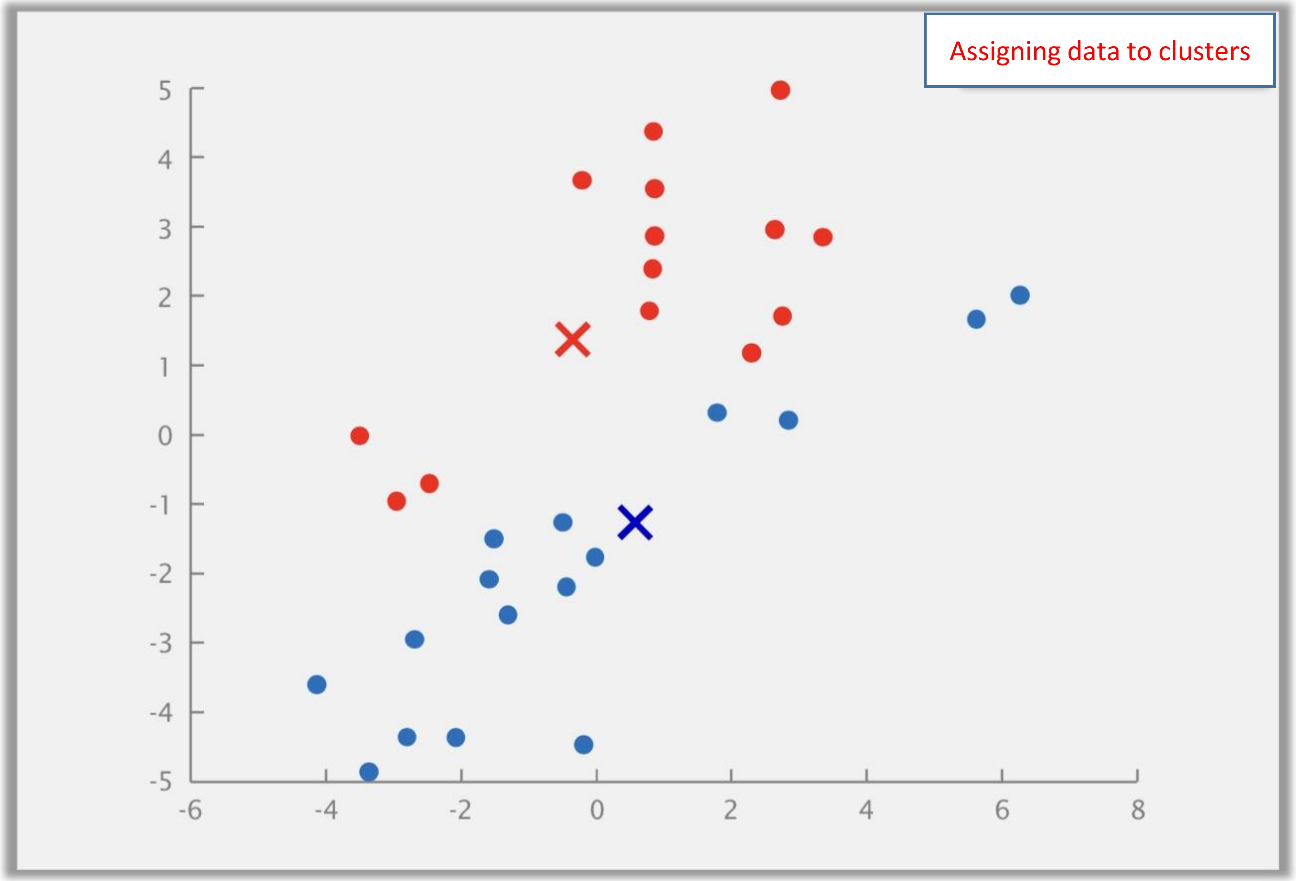


Cluster center update

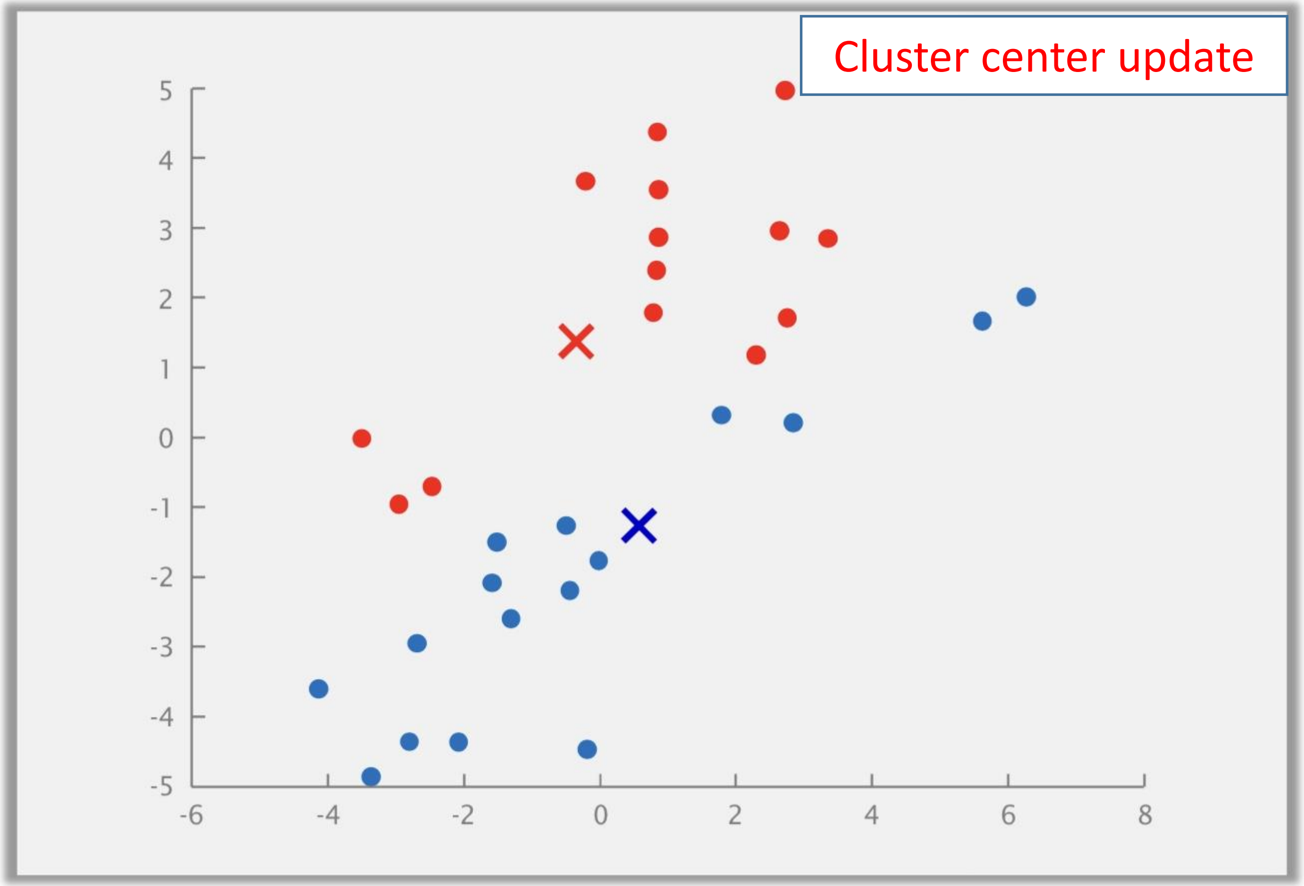
K-means algorithm: demonstration implementation



K-means algorithm: demonstration implementation

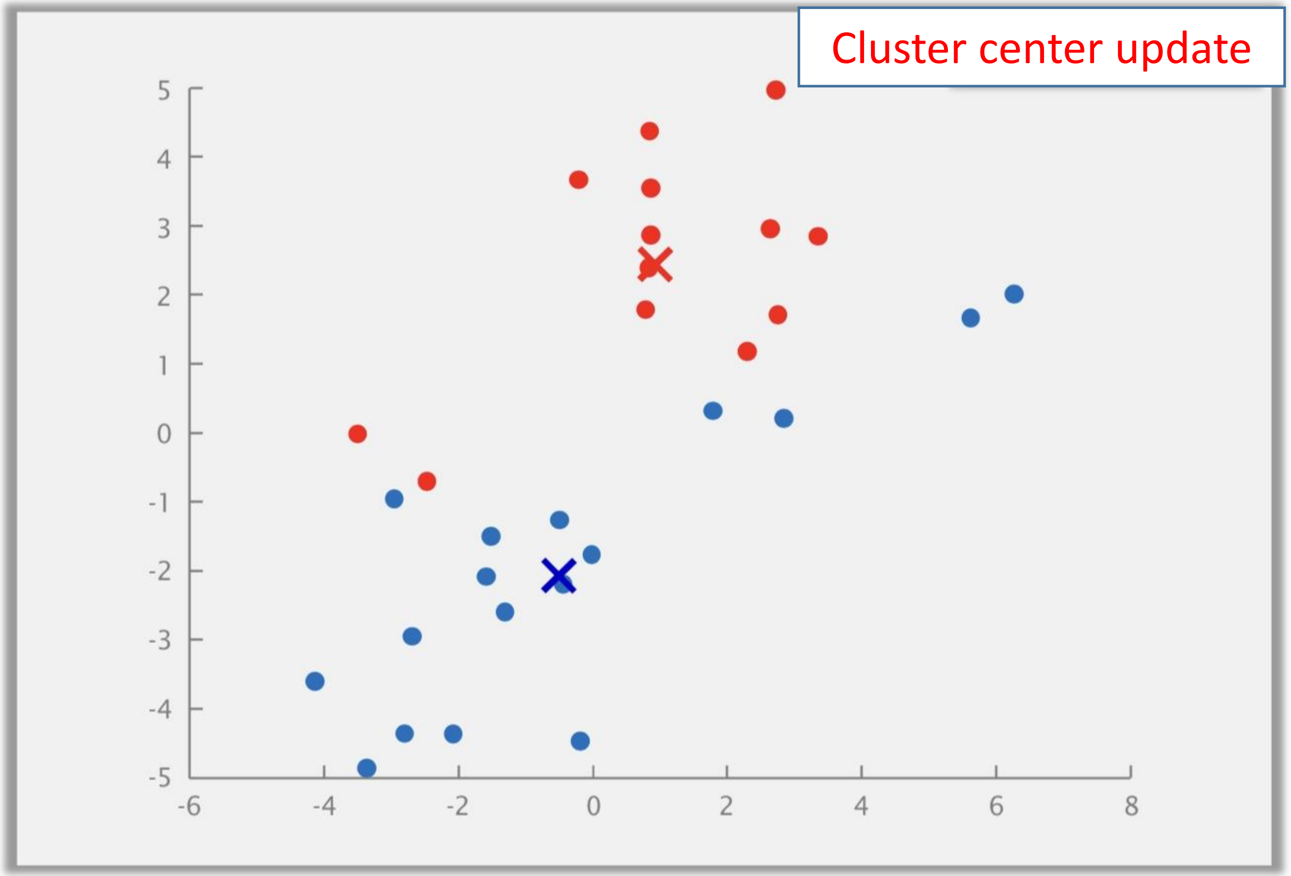


K-means algorithm: demonstration implementation



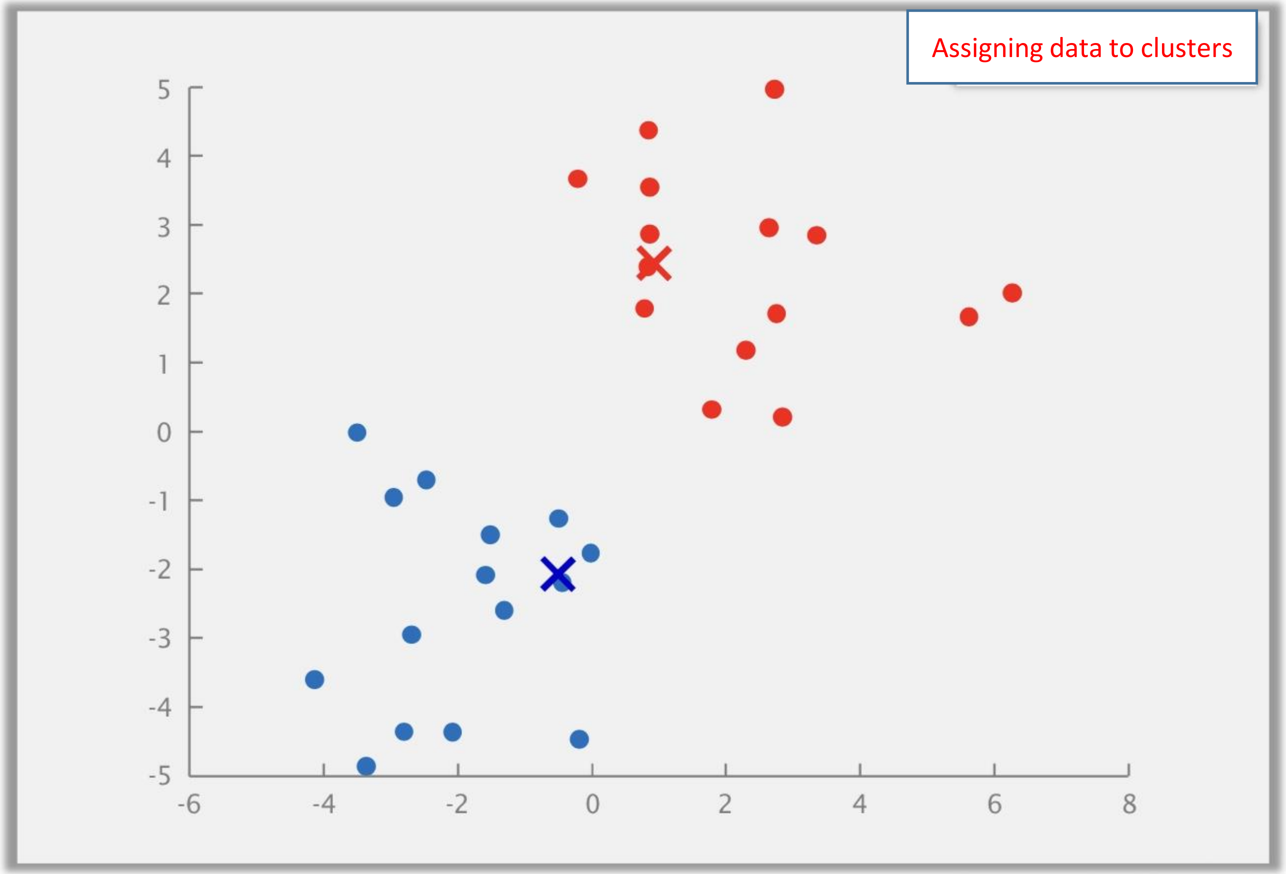
Cluster center update

K-means algorithm: demonstration implementation

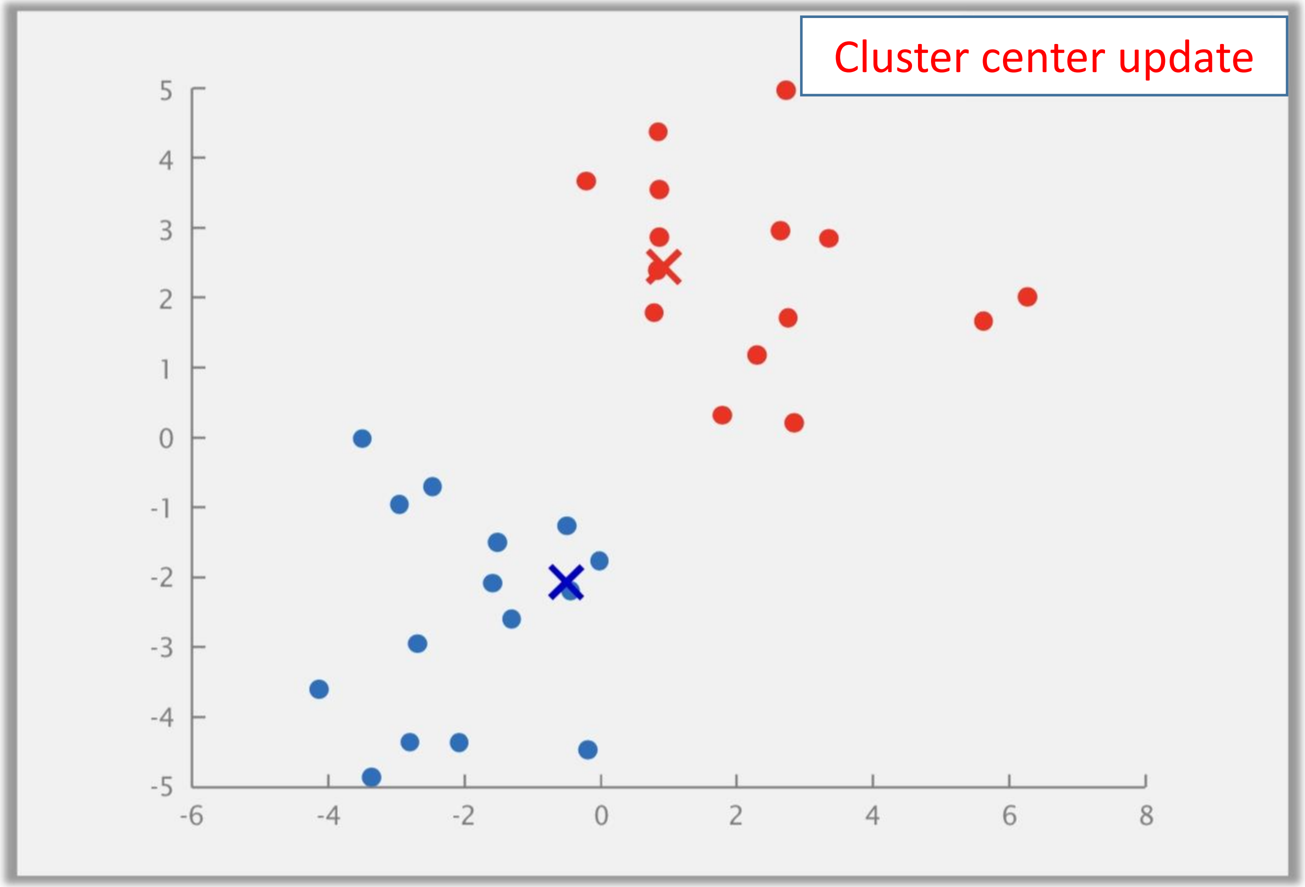


Cluster center update

K-means algorithm: demonstration implementation

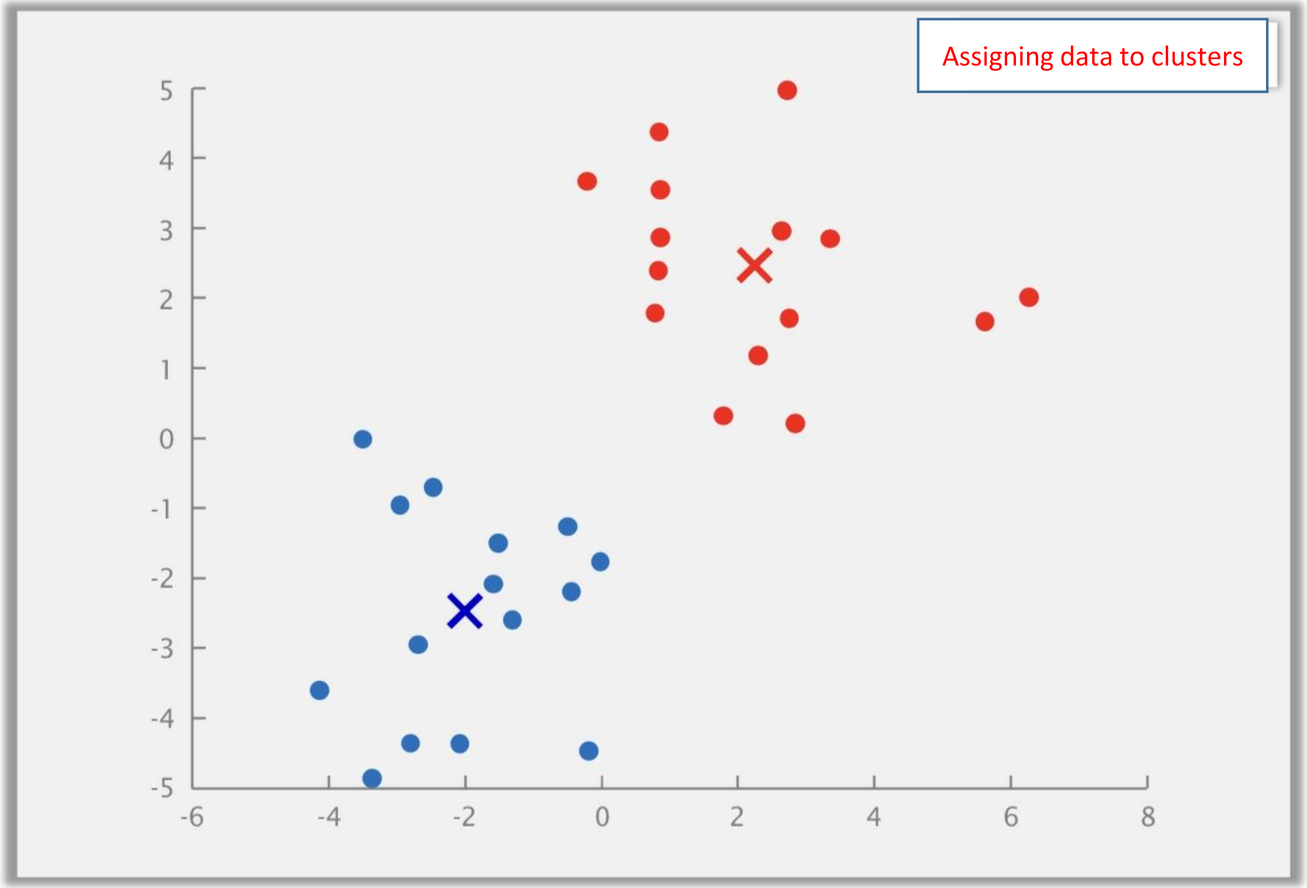


K-means algorithm: demonstration implementation

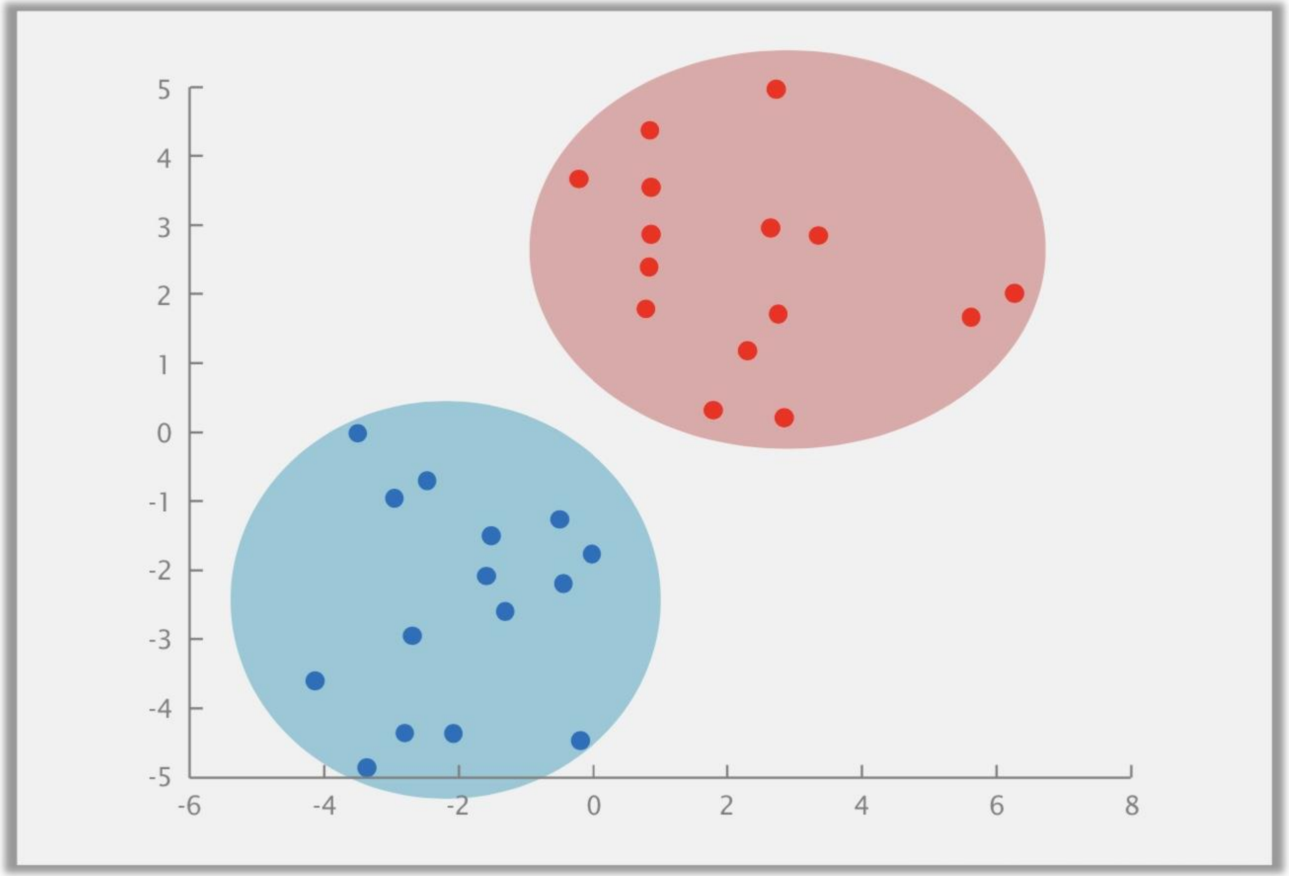


Cluster center update

K-means algorithm: demonstration implementation



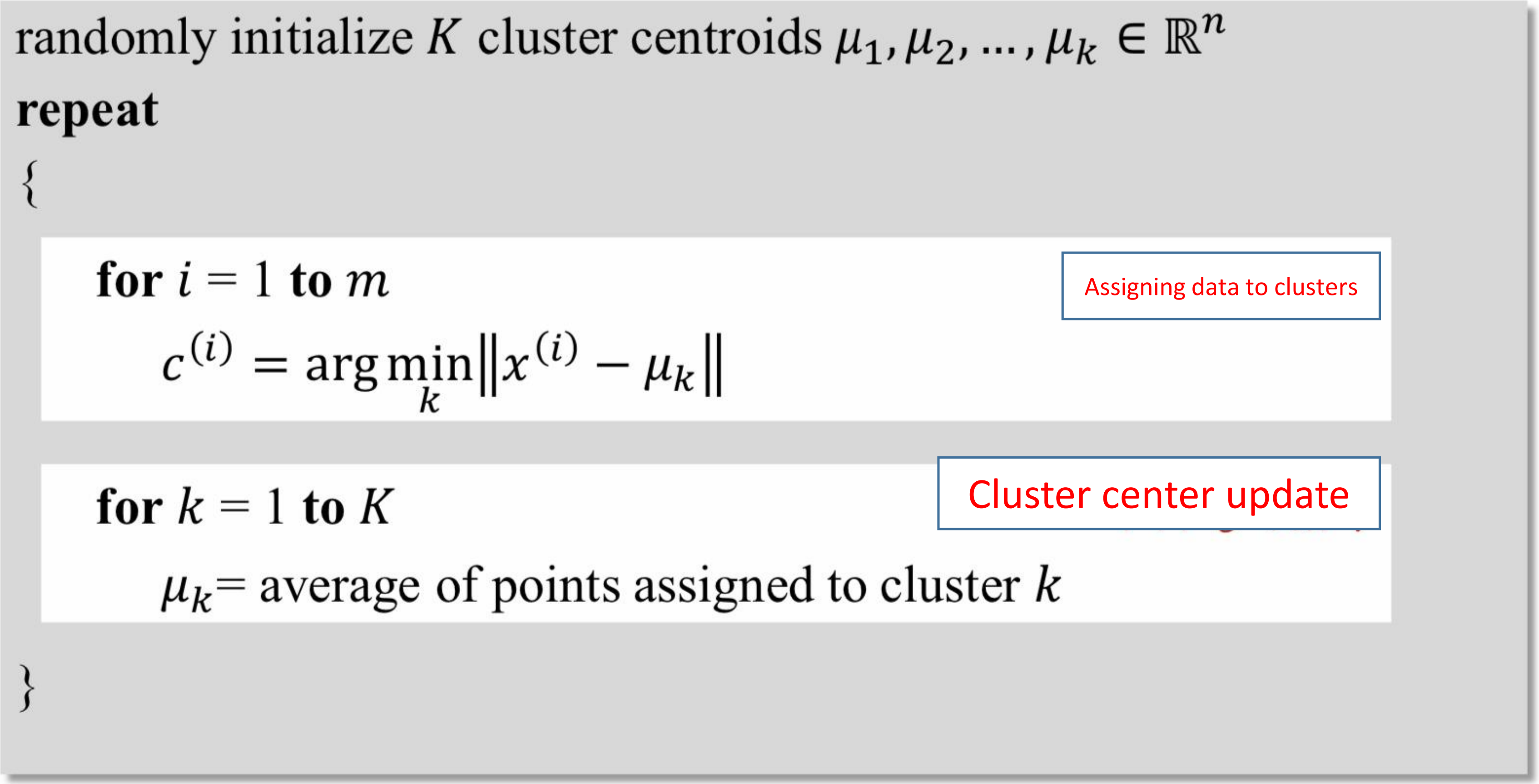
K-means algorithm: demonstration implementation



K-means algorithm

- Entrance:
  - Number of clusters: k
  - Training set: {x $^{(1)}$, x $^{(2)}$, ... , x $^{(m)}$}
- Note: In the training set, no label is assigned to the data.
- Note: There is no need to add the attribute x_0 = 1 in clustering.

K-means algorithm



```
randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$
repeat
{
    for $i = 1$ to $m$
        $c^{(i)} = \arg\min_k \|x^{(i)} - \mu_k\|$
    for $k = 1$ to $K$
        $\mu_k$= average of points assigned to cluster $k$
}
```

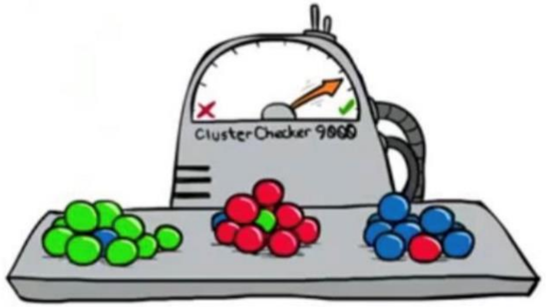
Assigning data to clusters

Cluster center update

Clustering: objective function

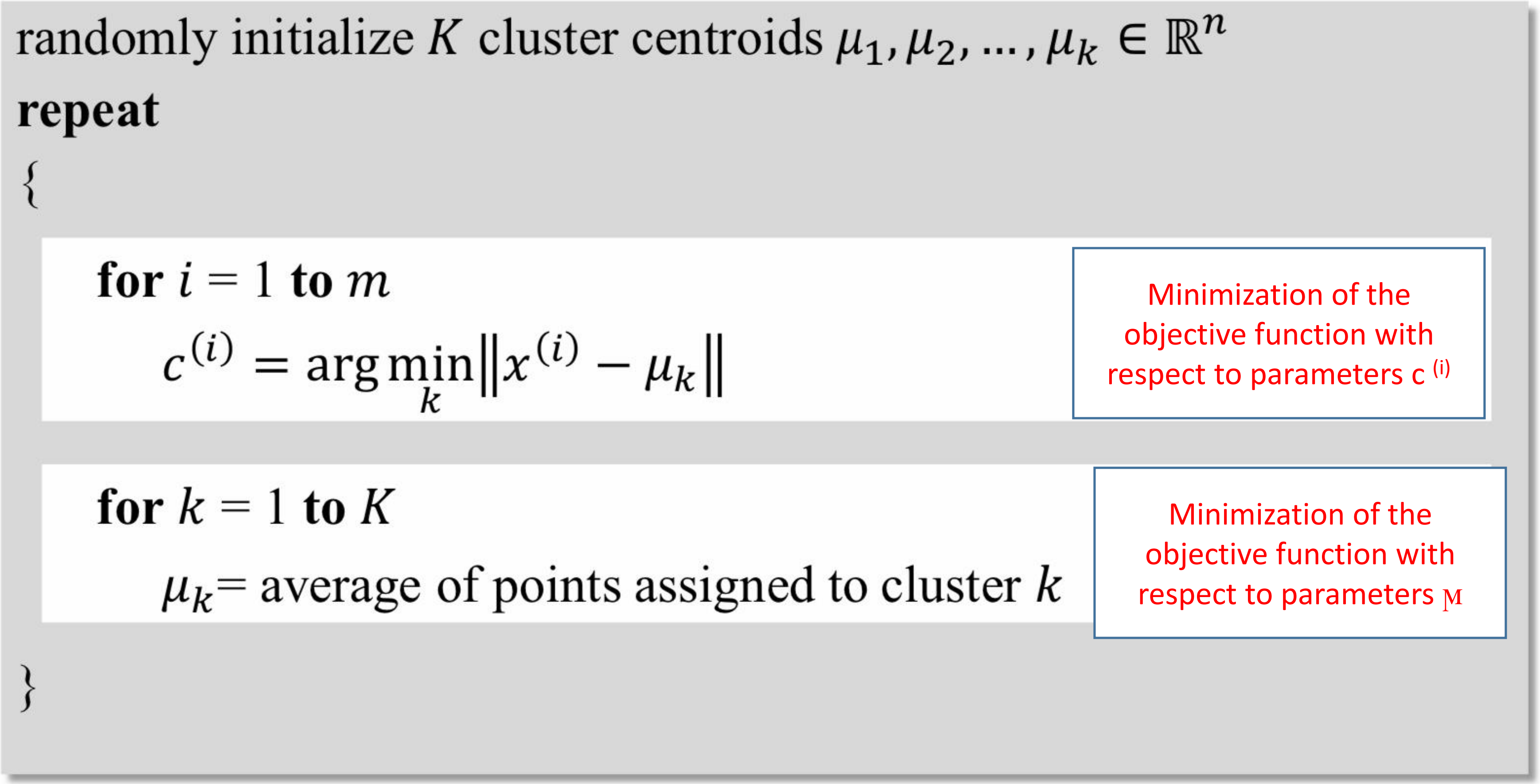
objective function

- symbols:
  - M_k: cluster center k
  - c $^{(i)}$: the number of the cluster assigned to the data x $^{(i)}$
  - M_c $^{(i)}$: the center of the cluster assigned to the data x $^{(i)}$
- The objective function



$$J\left(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k\right) = \frac{1}{m} \sum_{i=1}^{m} \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

K-means algorithm



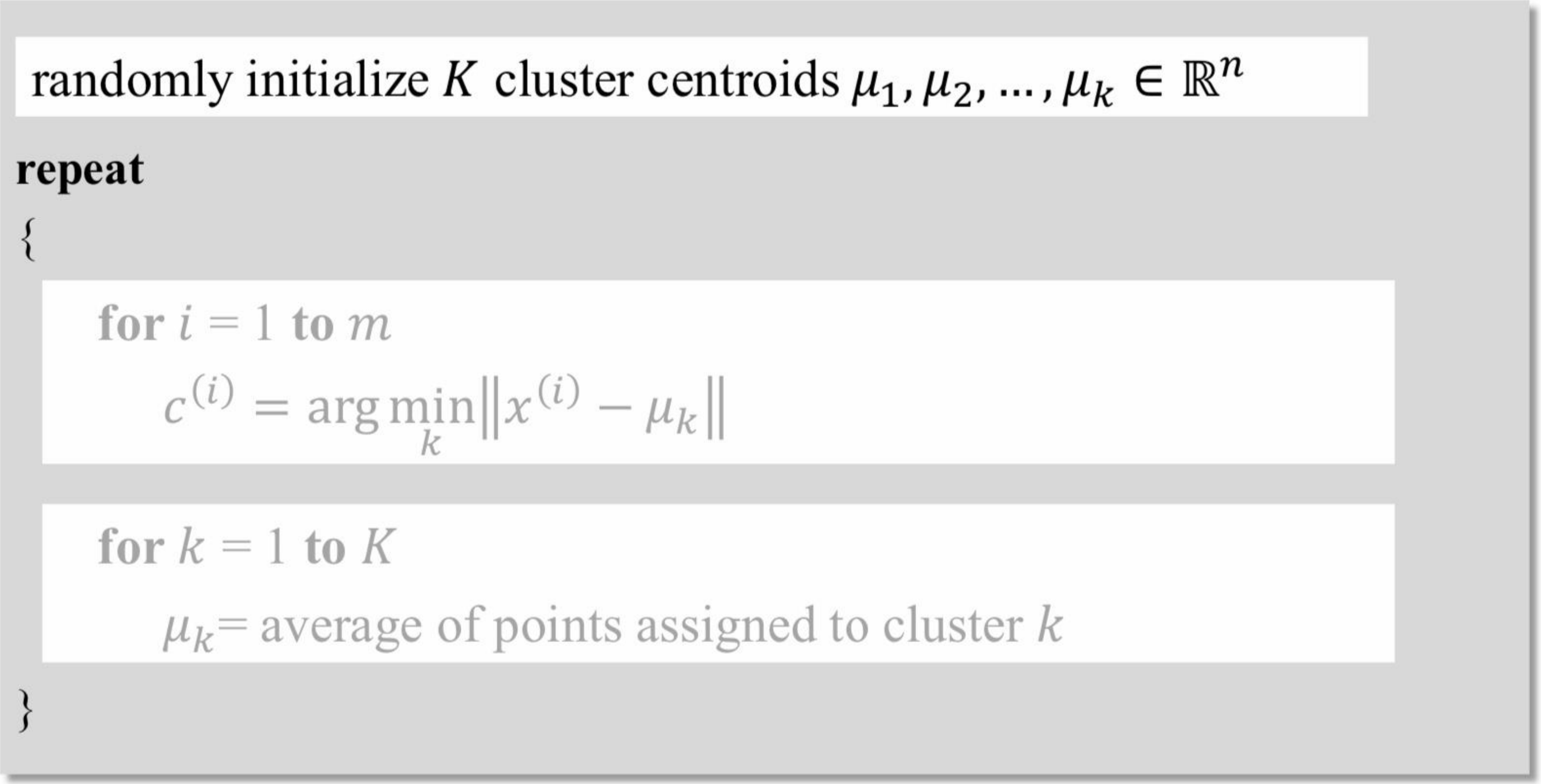
```
randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$
repeat
{
    for $i = 1$ to $m$
        $c^{(i)} = \arg\min_k \|x^{(i)} - \mu_k\|$

    for $k = 1$ to $K$
        $\mu_k$= average of points assigned to cluster $k$
}
```

Minimization of the objective function with respect to parameters c $^{(i)}$

Minimization of the objective function with respect to parameters M

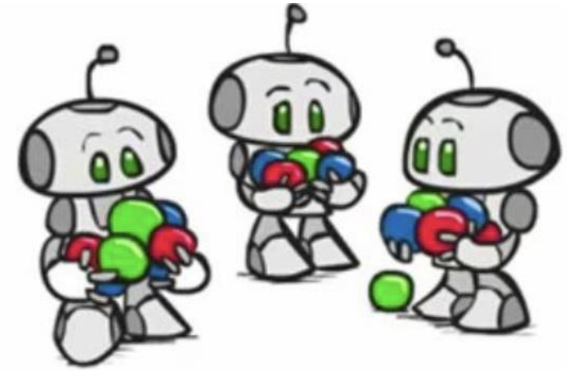
Clusters’ centers initializing

K-means algorithm

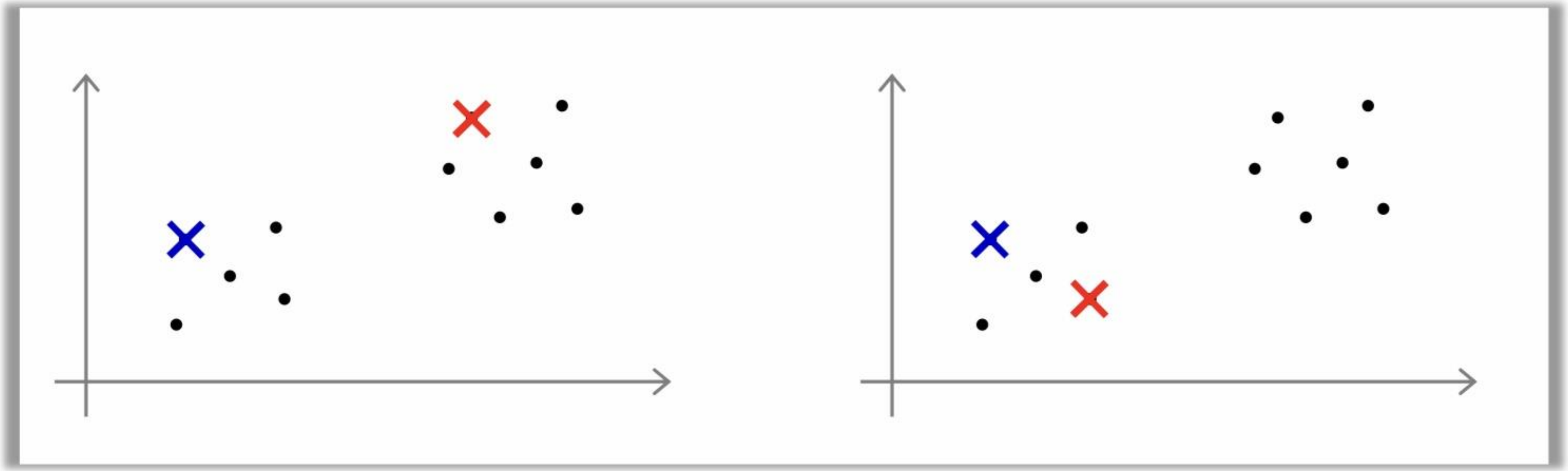


```
randomly initialize $K$ cluster centroids $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$
repeat
{
    for $i = 1$ to $m$
        $c^{(i)} = \arg\min_{k} \|x^{(i)} - \mu_k\|$
    for $k = 1$ to $K$
        $\mu_k$= average of points assigned to cluster $k$
}
```

Clusters’ centers initializing

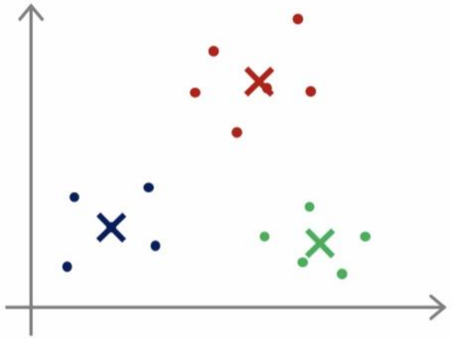


Initial initialization (K $\leq$ m):
Selection of K training sample randomly
Assigning cluster centers to K selected samples

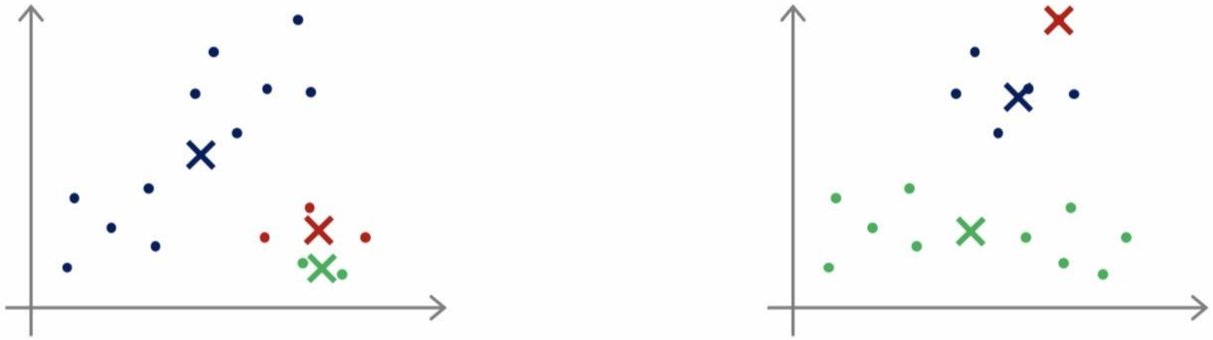


Local Optimum

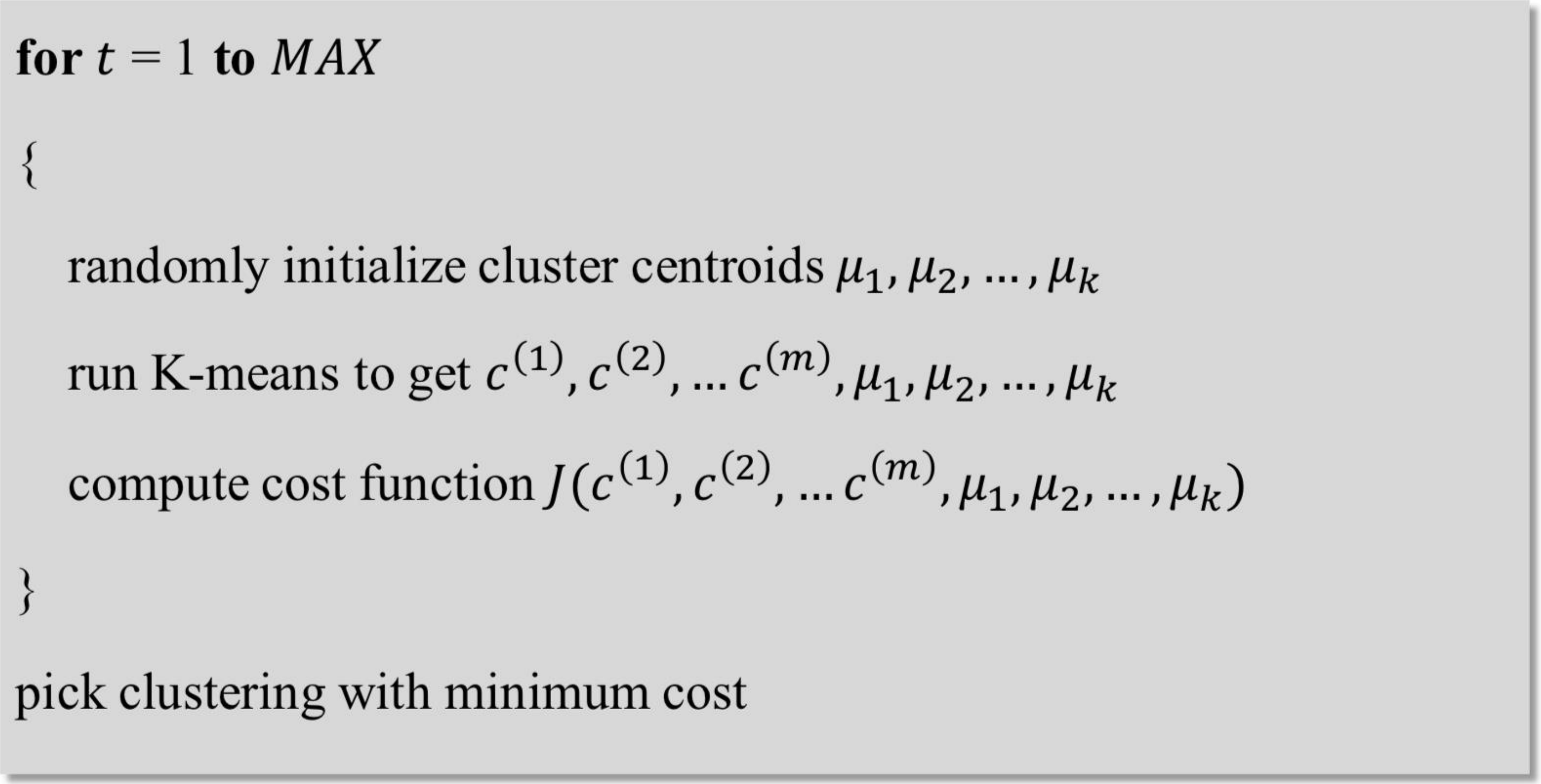
- Global Optimum



- Local Optimum



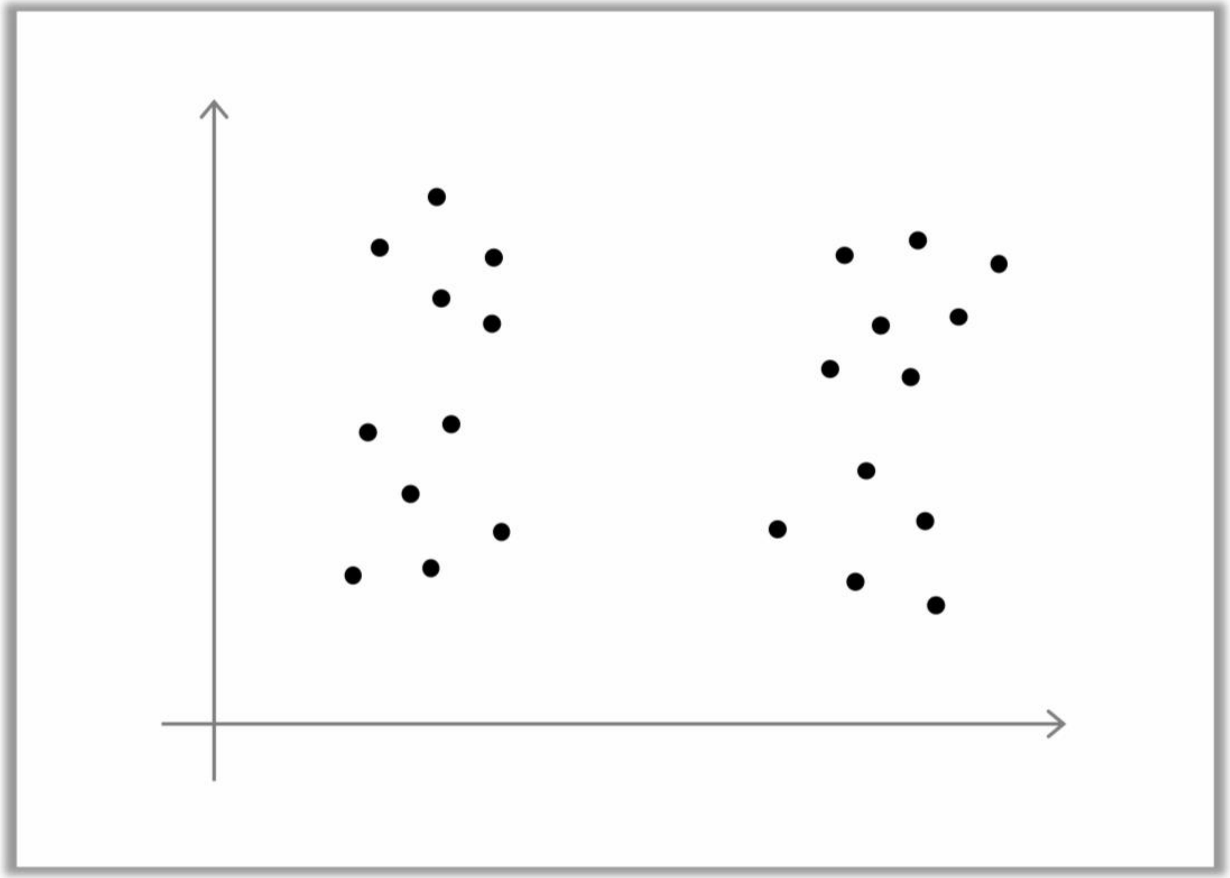
Local Optimum Avoidance



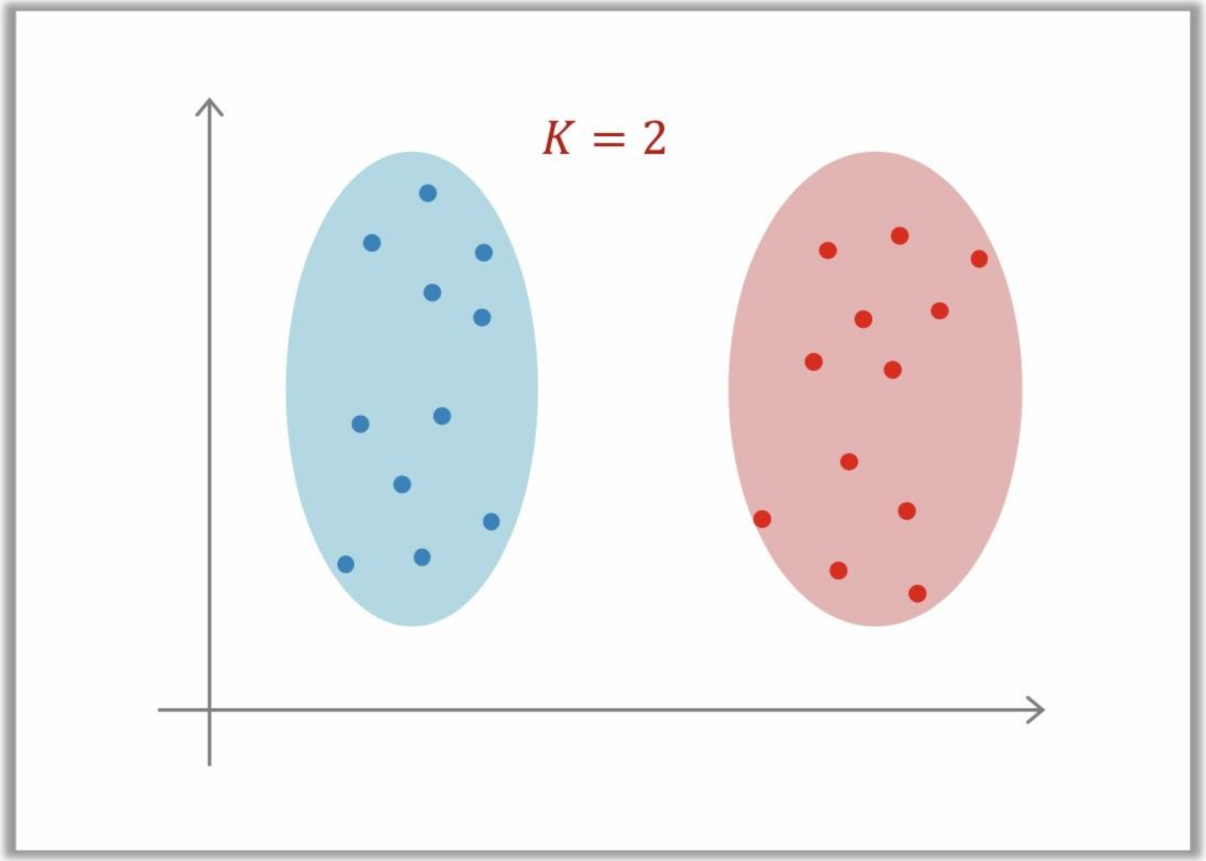
```
for $t = 1$ to $MAX$
{
    randomly initialize cluster centroids $\mu_1, \mu_2, \dots, \mu_k$
    run K-means to get $c^{(1)}, c^{(2)}, \dots c^{(m)}, \mu_1, \mu_2, \dots, \mu_k$
    compute cost function $J(c^{(1)}, c^{(2)}, \dots c^{(m)}, \mu_1, \mu_2, \dots, \mu_k)$
}
pick clustering with minimum cost
```

Determine the number of clusters

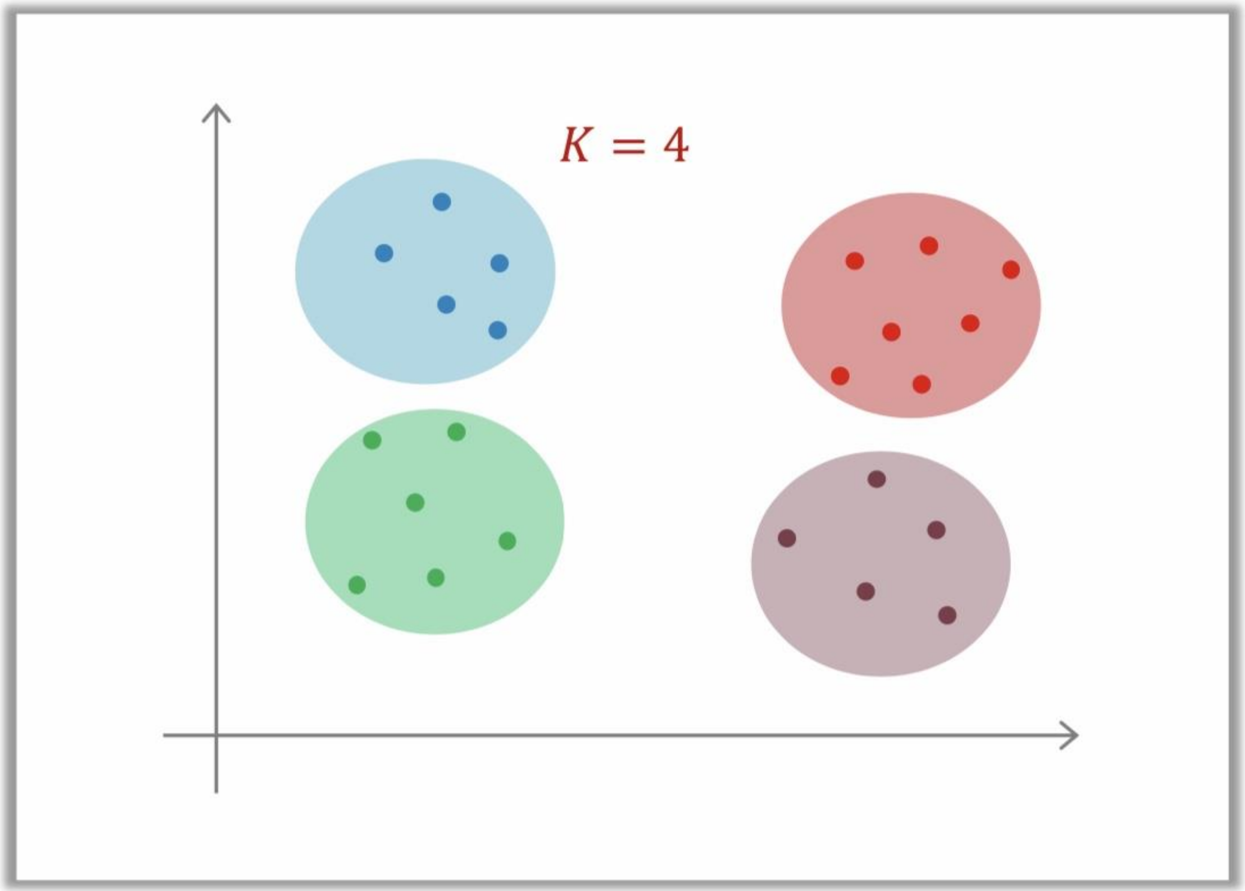
What is the right value for K?



What is the right value for K?

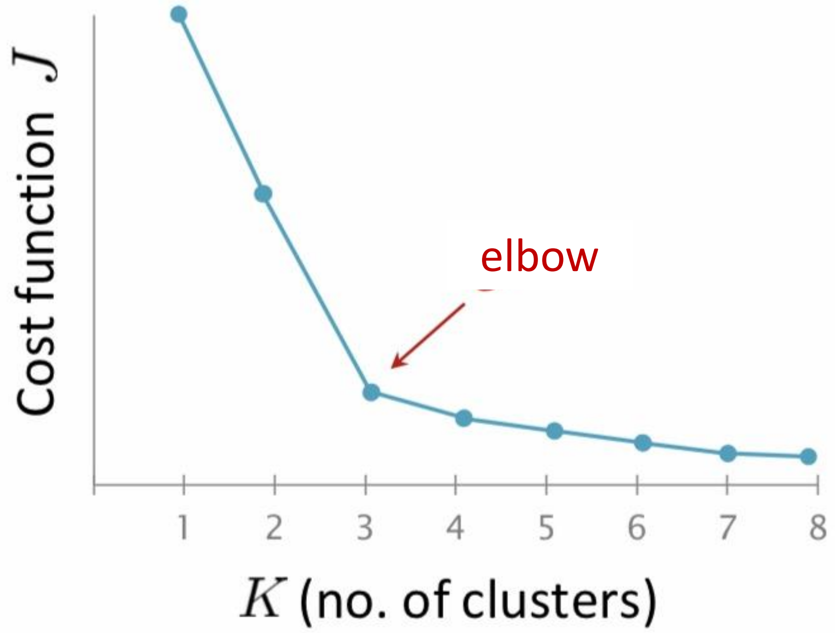


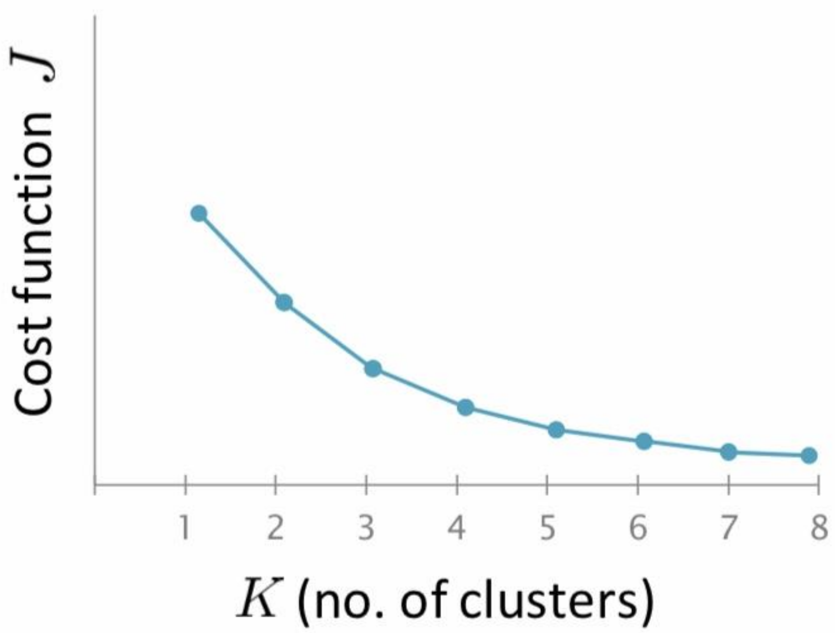
What is the right value for K?



Determine the number of clusters

- “elbow” method:





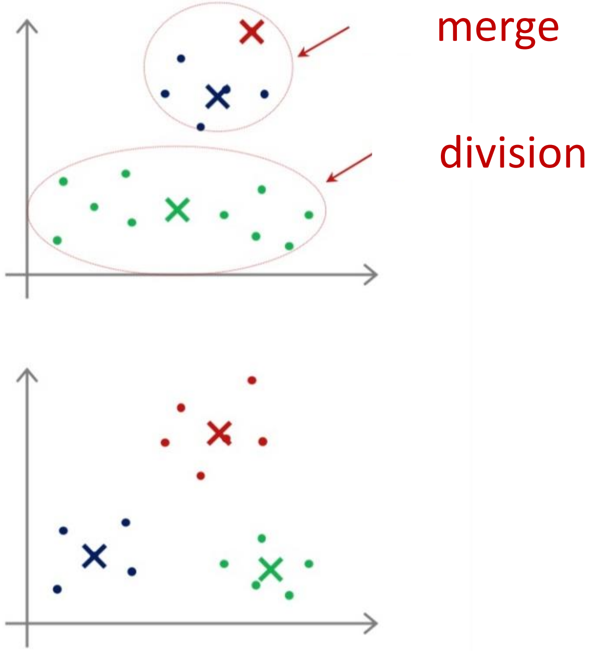
Clustering improvement

Clustering improvement by clusters post-processing

By running K-means on the data of this cluster with a value of K = 2

- Division:
  - Splitting a cluster with the highest error into two clusters

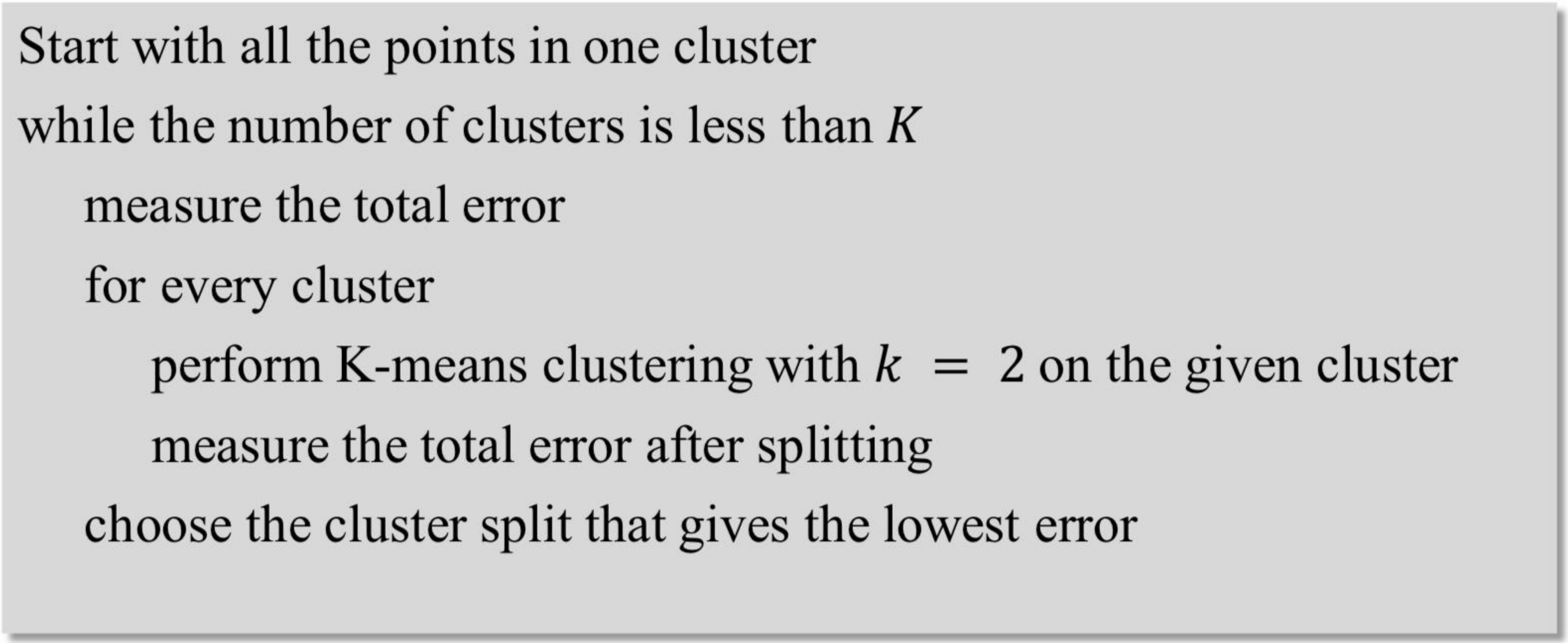
- merge:
  - Merge the two closest clusters
  - Merging two clusters with minimal increase in total error



Two-part K-means algorithm

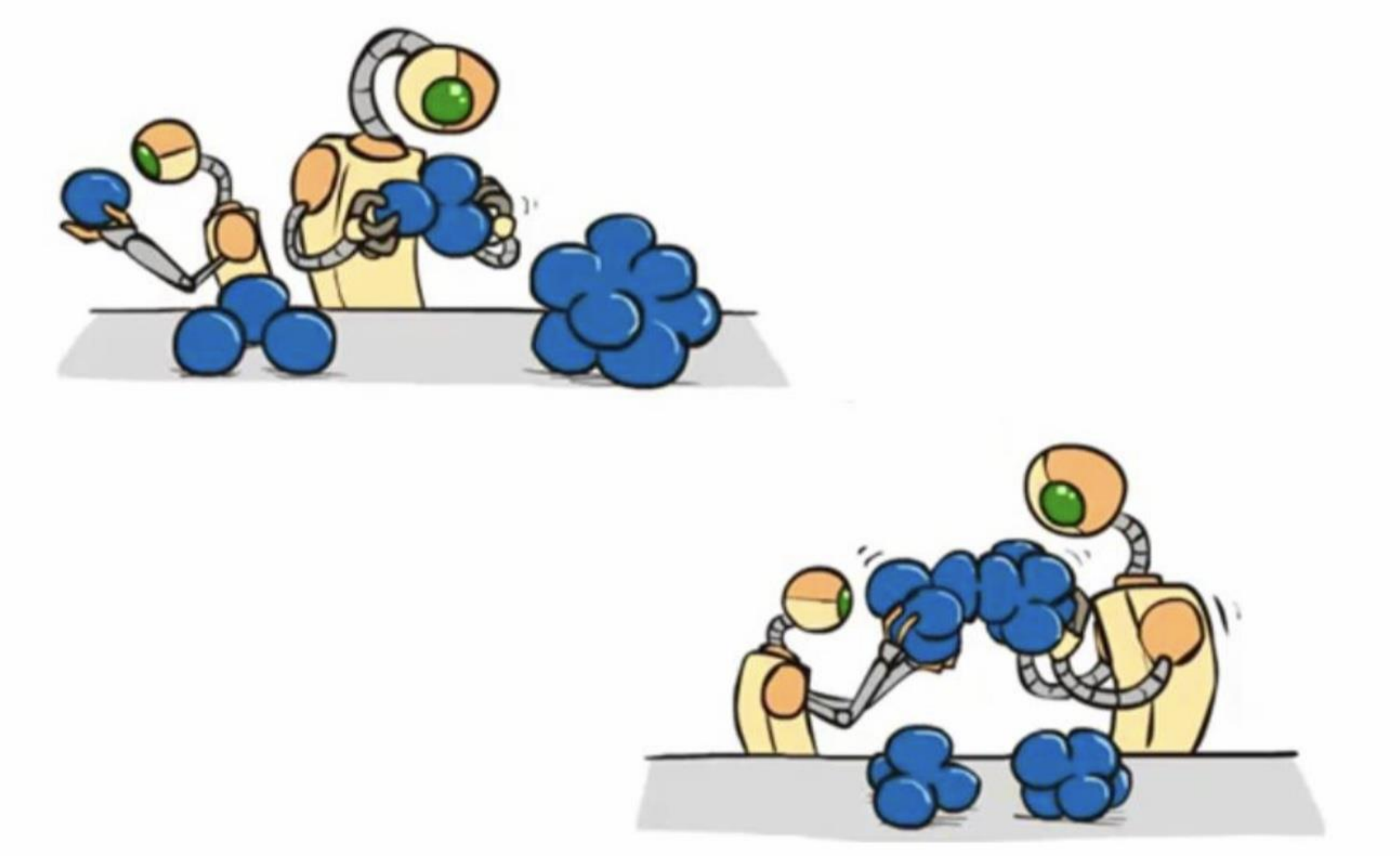
- Algorithm of two parts:
    - Start with a cluster containing all the data
    - Choose one cluster at a time:
        - Divide the selected cluster into two clusters using the K-means algorithm.
        - Calculate the total clustering error.
        - Choose the clustering with the least error.
    - Repeat the above process until you reach the desired number of clusters.

Two-part K-means algorithm



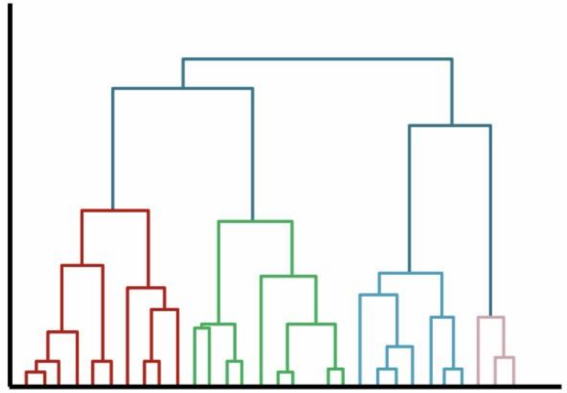
```
Start with all the points in one cluster
while the number of clusters is less than $K$
    measure the total error
    for every cluster
        perform K-means clustering with $k = 2$ on the given cluster
        measure the total error after splitting
    choose the cluster split that gives the lowest error
```

Hierarchical clustering



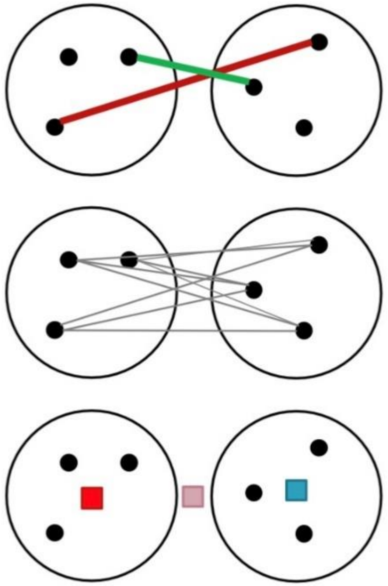
Hierarchical clustering

- Hierarchical clustering:
  - First, merge very similar data.
  - Gradually create larger clusters by merging smaller clusters.
- Algorithm:
  - At first, each data represents a cluster.
  - Repeat the following steps:
    - Choose the two closest clusters each time.
    - Merge those two clusters into a new cluster.
    - Stop: when there is only one cluster left.
- Create a tree diagram containing a wide range of clusters.



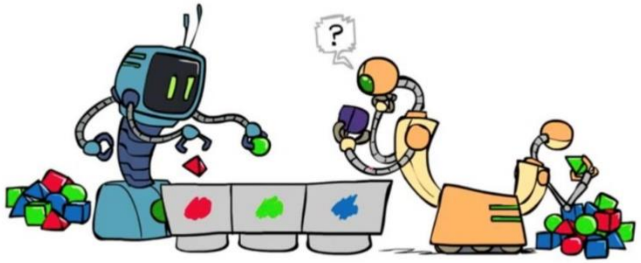
Hierarchical clustering

- Question: How to define the closest two clusters?
- Criteria for determining the similarity of clusters:
  - Nearest pair (one-link clustering)
  - Farthest pair (all-link clustering)
  - Average distance of all pairs
  - "WARD" method (least dispersion like K-means)
- Different criteria create different clusters.



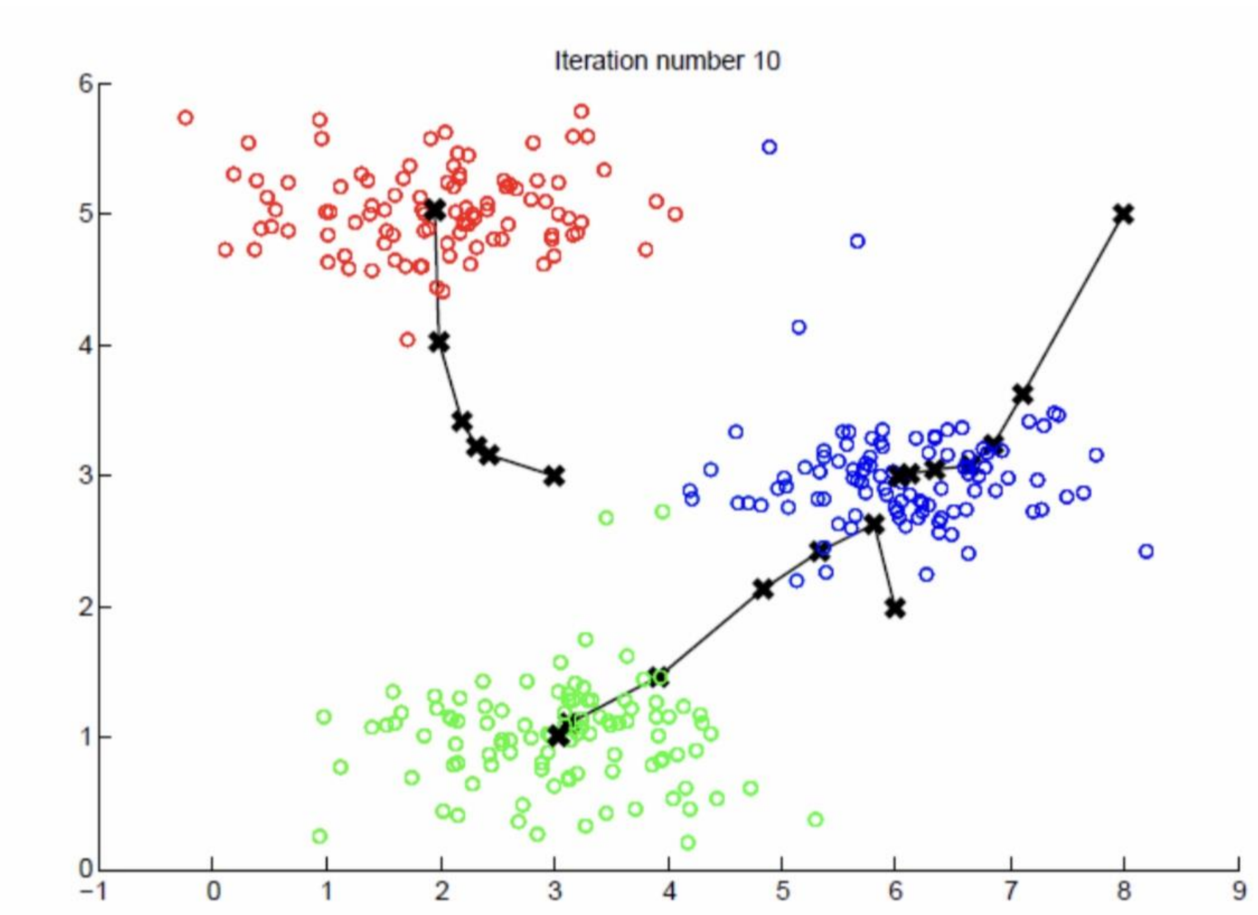
Summary

- Unsupervised learning: finding structure in data
- Clustering: grouping similar data
  - K-means clustering algorithm
    - Easy implementation
    - Slow for very large data sets
    - Possibility of getting stuck in the local optimum
  - Post-processing of clusters: splitting and merging of clusters
  - Two-part K-means algorithm
    - Better clustering than K-means algorithm
  - Hierarchical clustering algorithms



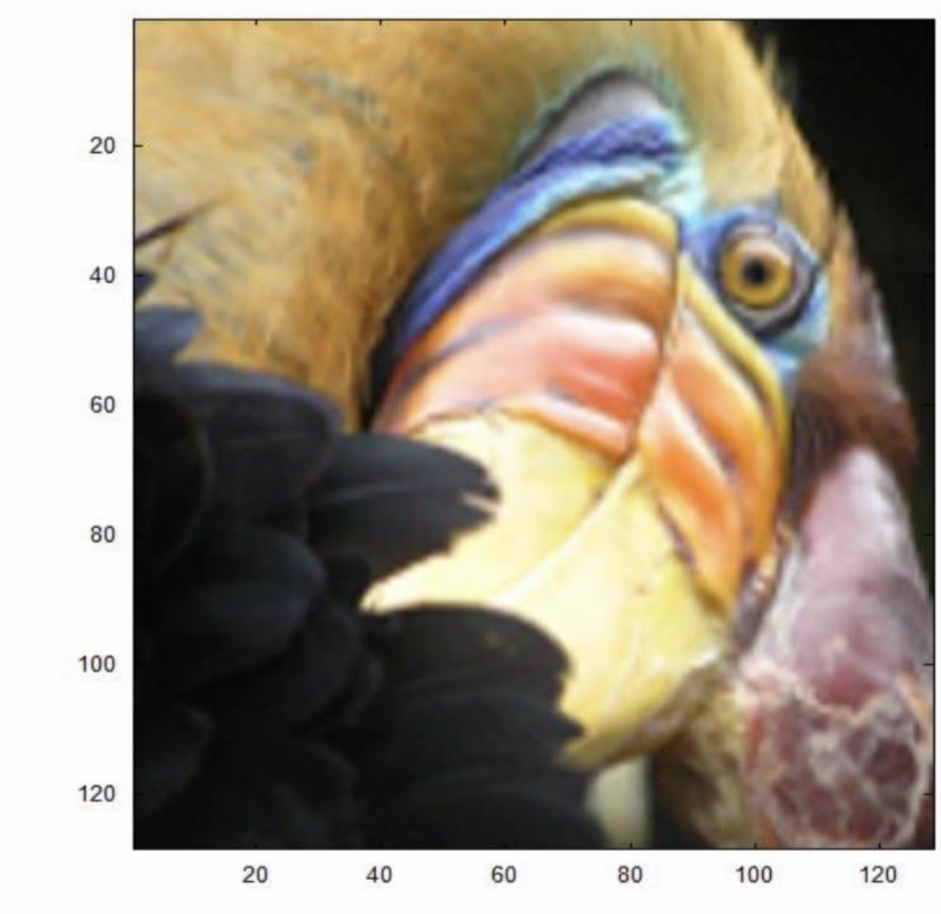
Practices

Practice1: K-means algorithm implementation



Exercise 2: Image compression using K-means

Main image



compressed image (16 colors)