Introduction to Data Science

Lecture 4: Unsupervised Learning

Clustering, K-means

Data Science and Engineering Department
Faculty of Informatics
ELTE University

Zakarya Farou zakaryafarou@inf.elte.hu

Supervised Learning vs Unsupervised Learning

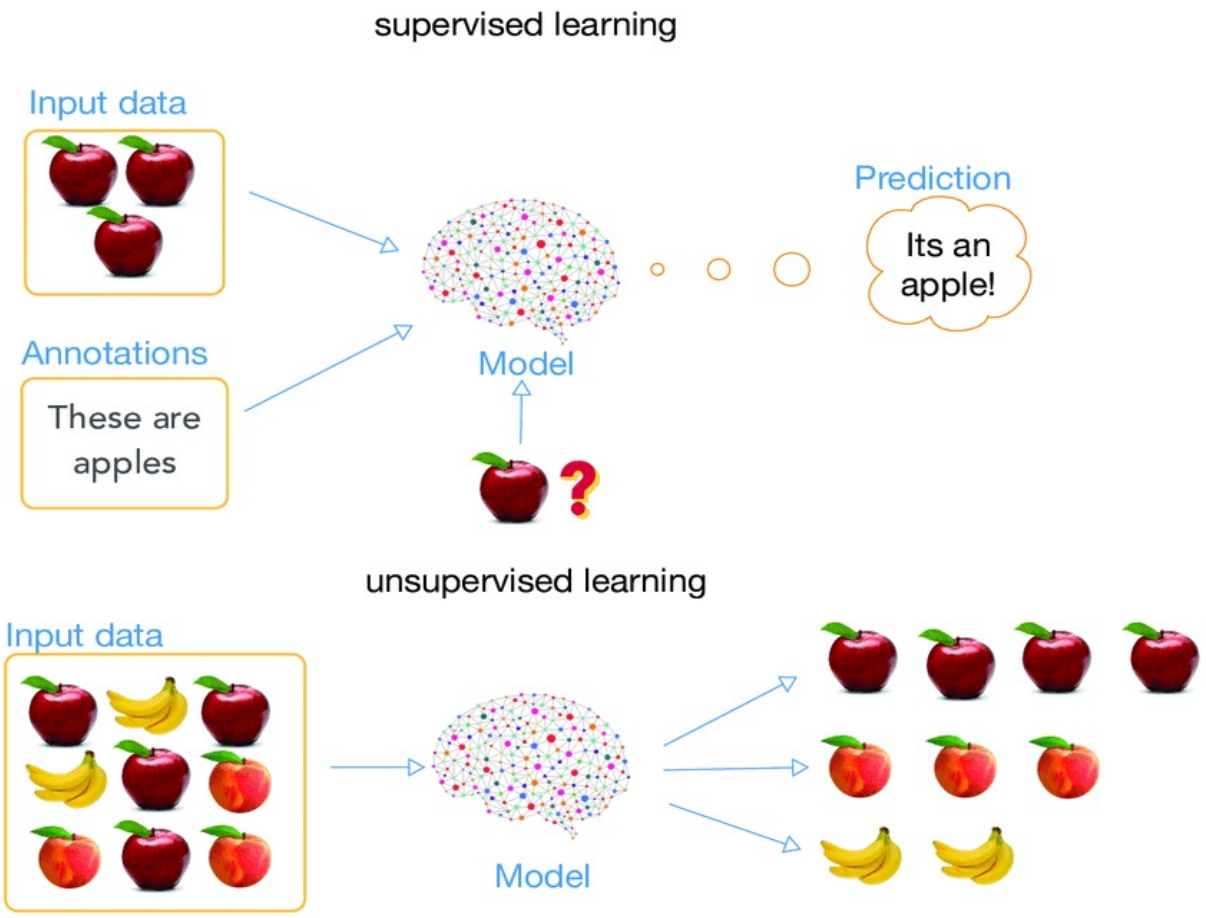


image source

https://www.researchgate.net/figure/Supervised-learning-and-unsupervised-learning-Supervised-learning-uses-annotation_fig1_329533120

What is Clustering? Why do it?

Consider these data:

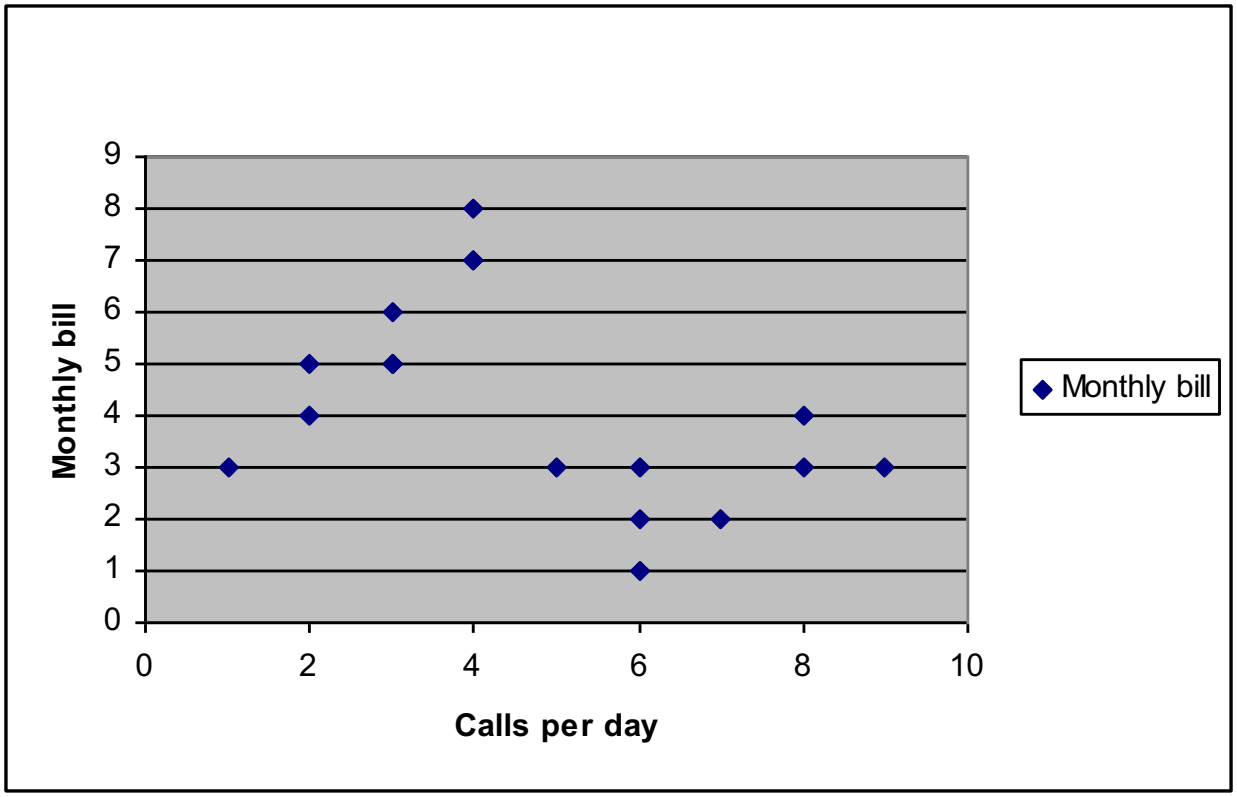
Made up; maybe they are
17 subscribers to a mobile phone services company, and show the mean calls per day and the mean monthly bill for each customer

Do you spot any patterns or Structure ??

Subscriber	Calls per day	Monthly bill
1	1	3
2	4	7
3	4	8
4	3	5
5	6	1
6	9	3
7	3	5
8	7	2
9	4	7
10	6	3
11	2	5
12	8	4
13	5	3
14	6	2
15	2	4
16	3	6
17	8	3

What is Clustering? Why do it?

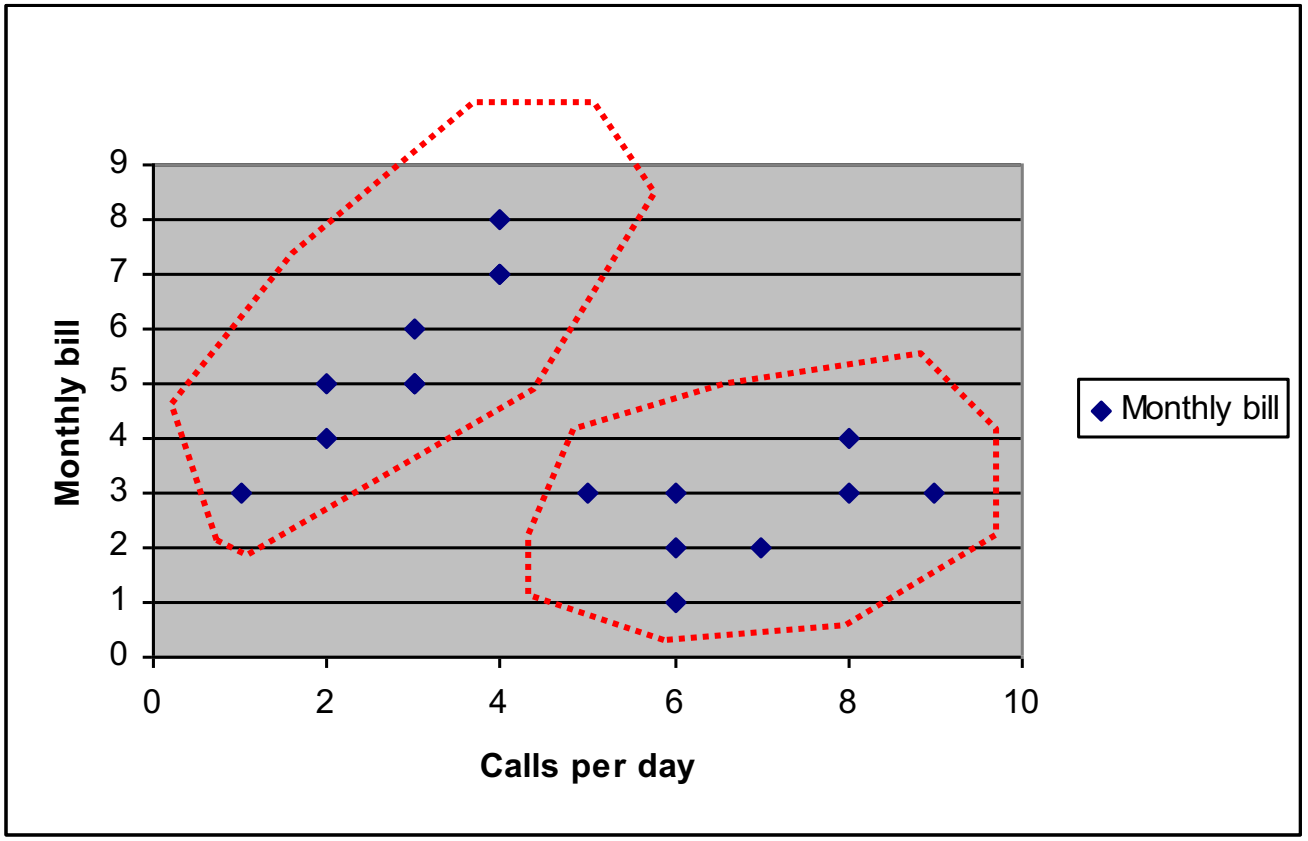
Here is a plot of the data, with calls as X and bills as Y



Now, do you spot any patterns or structures??

What is Clustering? Why do it?

Clearly there are two clusters -- two distinct *types* of customer



Top left: few calls but highish bills; bottom right: many calls, low bills

So, clustering is all about plotting/visualising and noting distinct groups by eye, right?

Not really, because:

- We can only spot patterns by eye (i.e. with our brains) if the data is 1D, 2D or 3D. Most data of interest is much higher dimensional – e.g. 10D, 20D, 1000D.
- Sometimes, the clusters are not so obvious as a bunch of data all in the same place – we will see examples.

- So: we need automated algorithms that can do what you just did (find distinct groups in the data), but which can do this for any number of dimensions and for perhaps more complex kinds of groups.

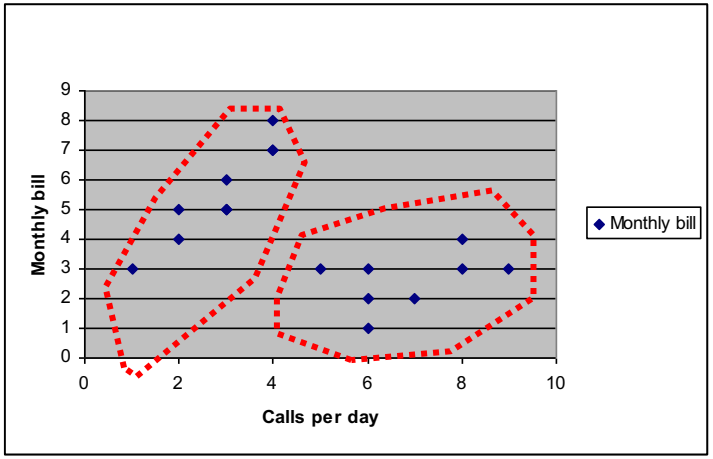
Clustering is about:

- Finding the natural groupings in a data set
- Often called ‘cluster analysis’;
- A key tool in ‘exploratory analysis’ or ‘data exploration.’

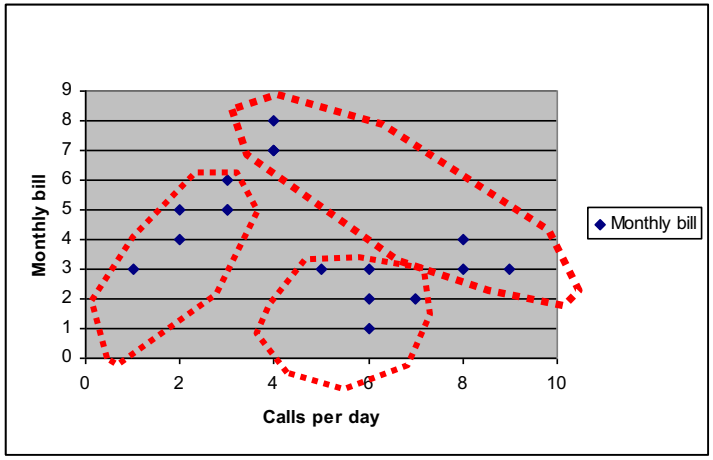
- Inspection of the results helps us learn useful things about our data – e.g. if we are doing this with supermarket baskets, each group is a collection of typical baskets, which may relate to “general housekeeping”, “late night dinner”, “quick lunchtime shopper”, and perhaps other types that we are not expecting

Quality of a clustering?

Why is this:



Better than this?



Quality of a clustering?

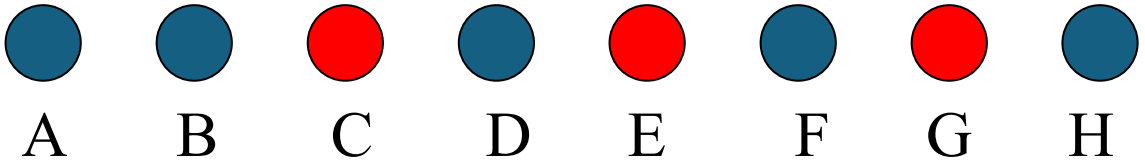
A `good clustering' has the following properties:

- Items in the same cluster tend to be **close** to each other
- Items in different clusters tend to be **far** from each other

It is not hard to come up with a metric – an easily calculated value – that can be used to give a score to any clustering. There are many such metrics. E.g.

S = the mean distance between pairs of items in the same cluster
D = the mean distance between pairs of items in different clusters
Measure of cluster quality is: **D/S** -- the higher the better.

Let's try that

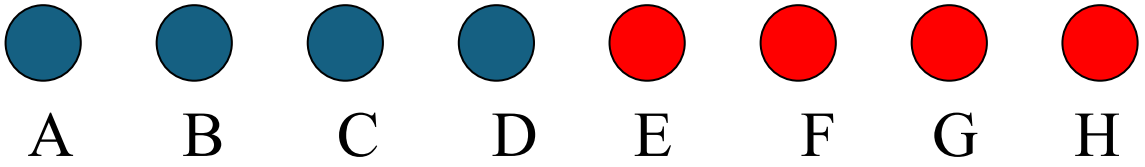


S = [AB + AD + AF + AH + BD + BF + BH + DF + DH + FH + CE + CG + EG] / 13 = 44/13 = 3.38

D = [AC + AE + AG + BC + BE + BG + DC + DE + DG + FC + FE + FG + HC + HE + HG]/15 = 40/15 = 2.67

Cluster Quality = D/S = 0.77

Let's try that again

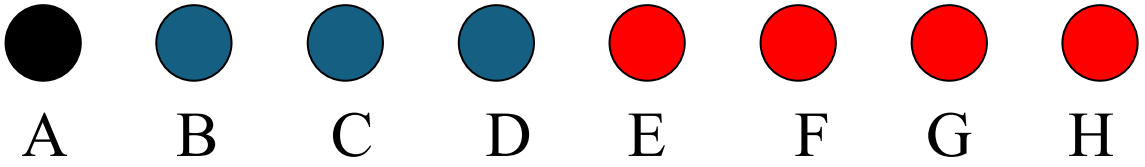


S = [AB + AC + AD + BC+ BD + CD + EF + EG + EH + FG + FH + GH] / 12 = 20/12 = 1.67

D = [AE + AF + AG + AH + BE + BF + BG + BH + CE + CF + CG + CH + DE + DF + DG + DH]/16 = 68/16 = 4.25

Cluster Quality = D/S = 2.54

But what about this?



S = [AB + CD + EF + EG + EH + FG + FH + GH] / 8 = 12/8 = 1.5

D = [AC + AD + AE + AF + AG + AH + BC + BD + BE + BF + BG + BH + CE + CF + CG + CH + DE + DF + DG + DH]/20 = 72/20=3.6

Cluster Quality = D/S = 2.40

Some important notes

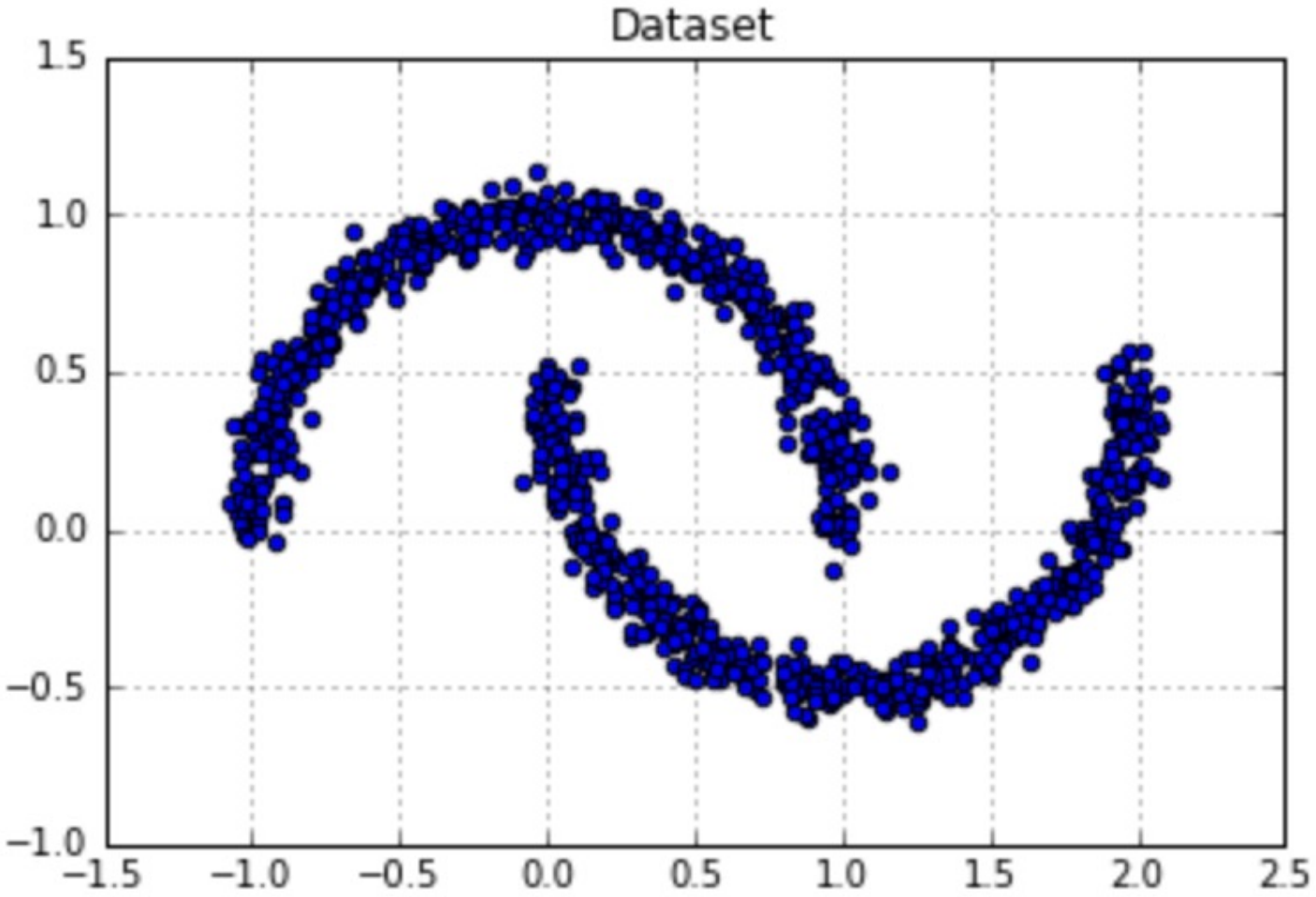
- There is usually no `correct' clustering.

- Clustering algorithms (whether or not they work with cluster quality metrics) *always use some kind of* ***distance*** *or* ***similarity*** *measure* -- the result of the clustering process will depend on the chosen distance measure.

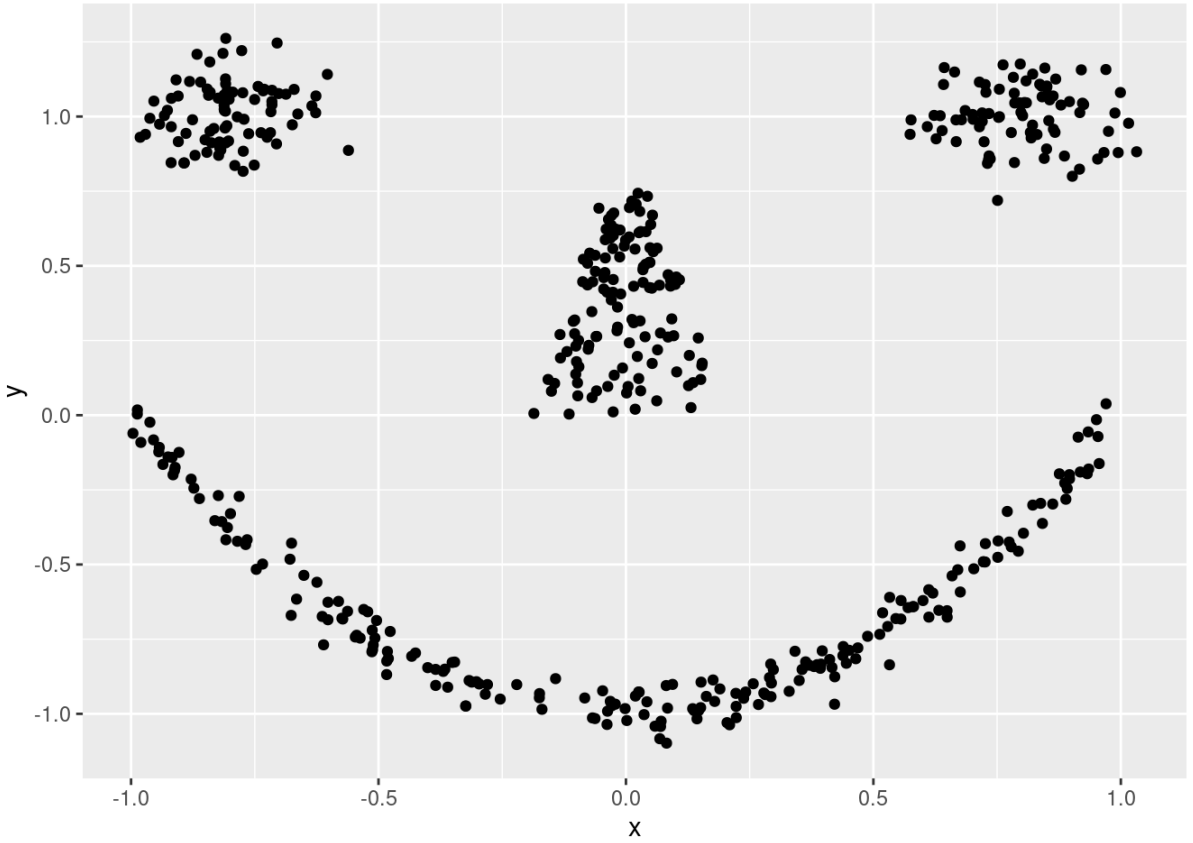
- The choice of the algorithm and/or distance measure will depend on the kind of cluster shapes you might expect in the data.

- Our D/S measure for cluster quality will not work well in lots of cases

Examples: sometimes groups are not simple to spot, even in 2D



Examples: sometimes groups are not simple to spot, even in 2D

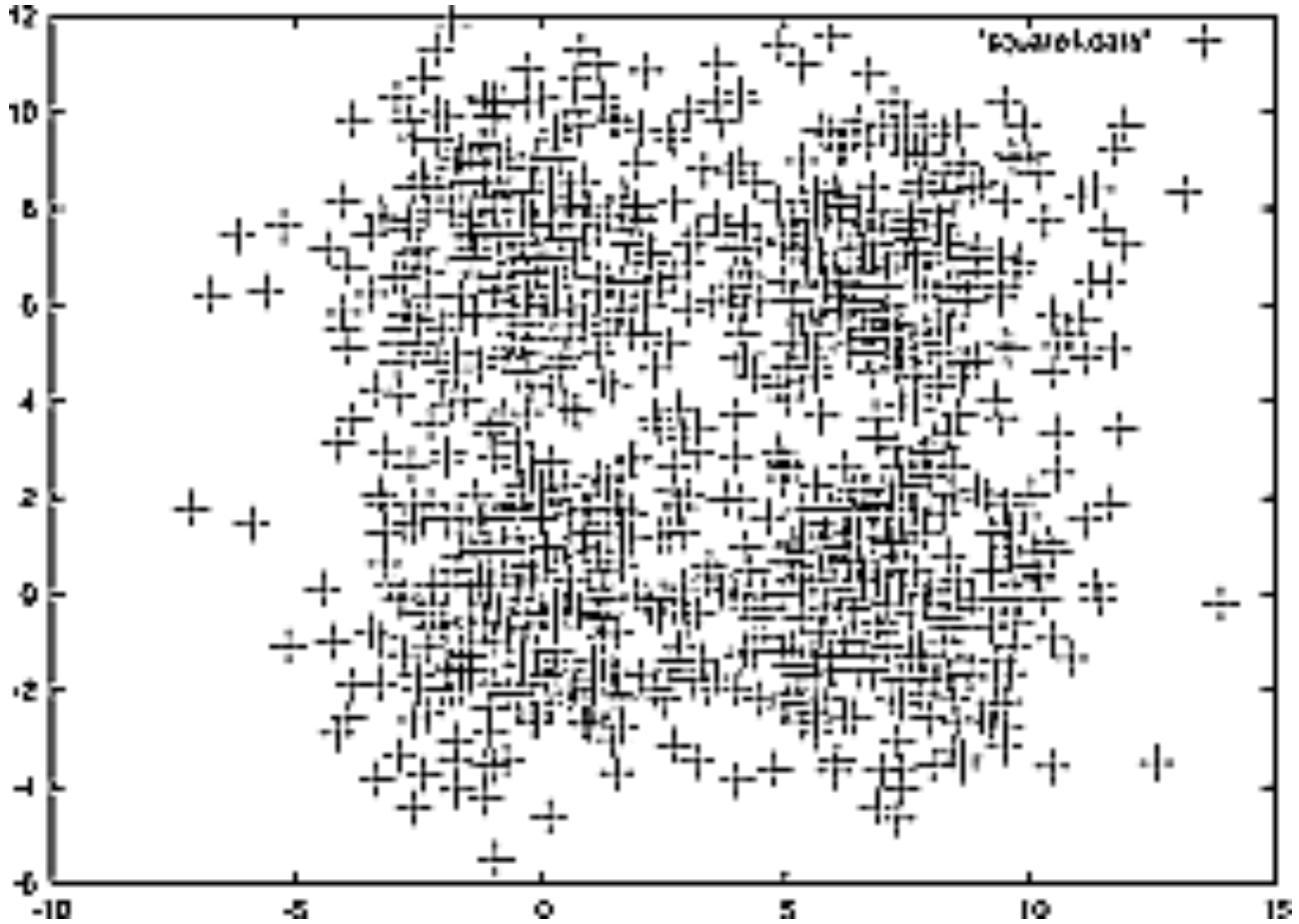


Brain Training

- Think about why D/S is not a useful cluster quality measure in the general case.

- Try to design a cluster quality metric that will work well in the cases of the previous slides (not very difficult)

In many problems the clusters are more `conventional' – but maybe fuzzy and unclear



How to do clustering

- The most commonly used methods:
  - ✦K-Means
  - ✦DBSCAN
  - ✦Hierarchical Agglomerative Clustering

Different kinds of clustering can be done, which avoids the issue of deciding the number of clusters in advance!

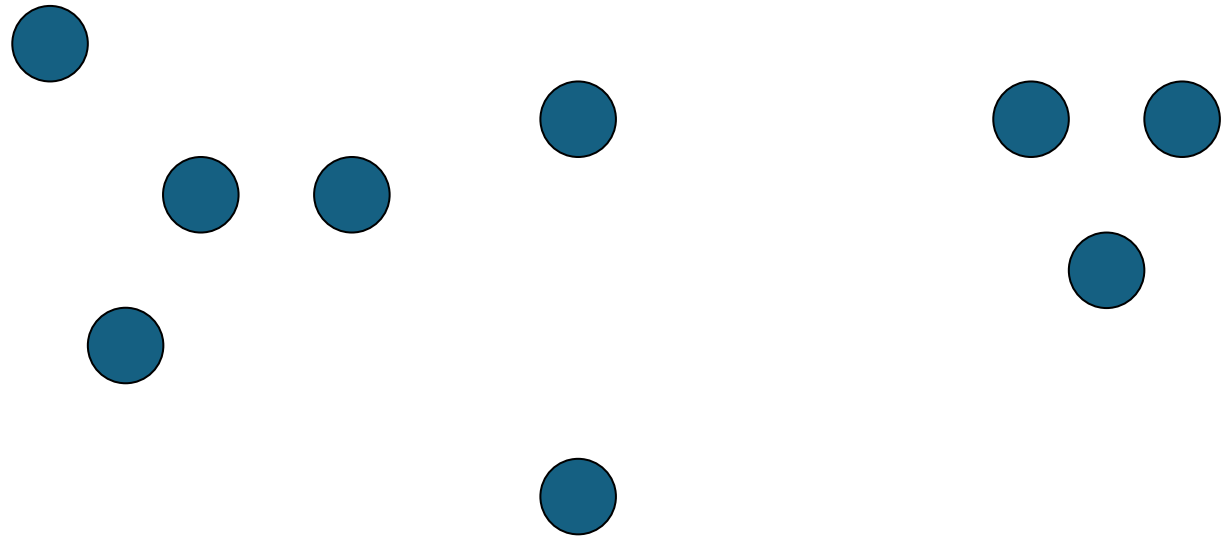
K-Means

Kmeans algorithm iteratively tries to partition the dataset into K distinct non-overlapping clusters. Each data point belongs to only one group. It tries to minimize the sum of the squared distance between the data points and the cluster's centroid.

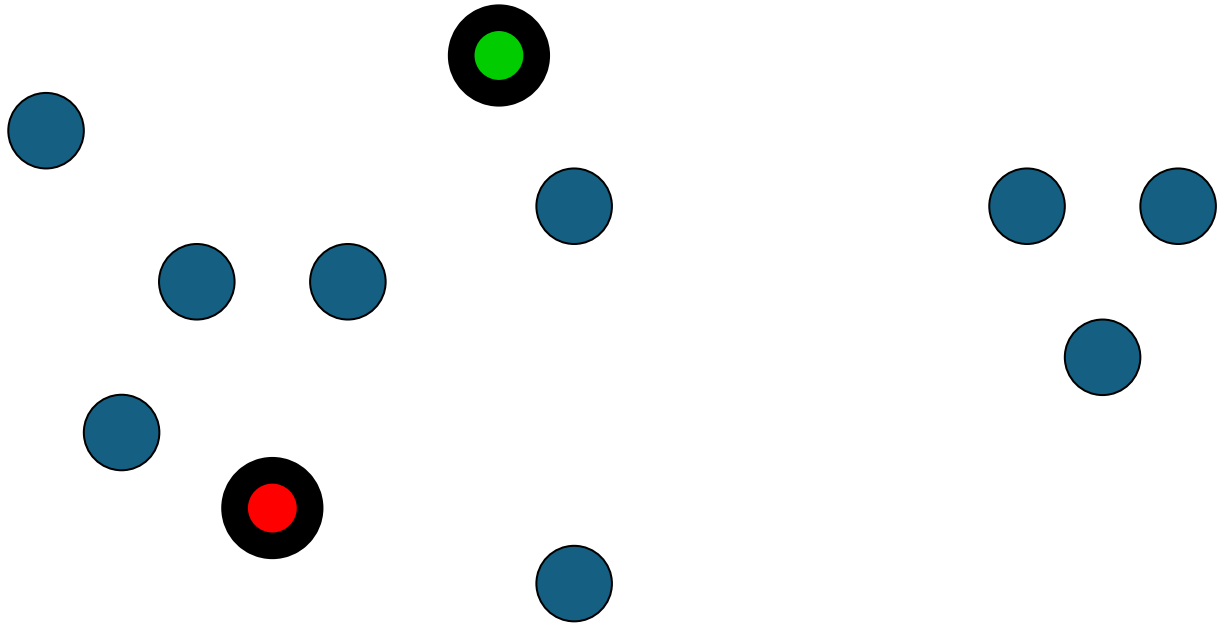
```
for k = 1 to K do
    μ_k ← some random location          // randomly initialize mean for kth cluster
end for
repeat
    for n = 1 to N do
        z_n ← argmin_k ||μ_k − x_n||    // assign example n to closest center
    end for
    for k = 1 to K do
        μ_k ← MEAN({ x_n : z_n = k })   // re-estimate mean of cluster k
    end for
until converged
return z                                // return cluster assignments
```

Let’s see it

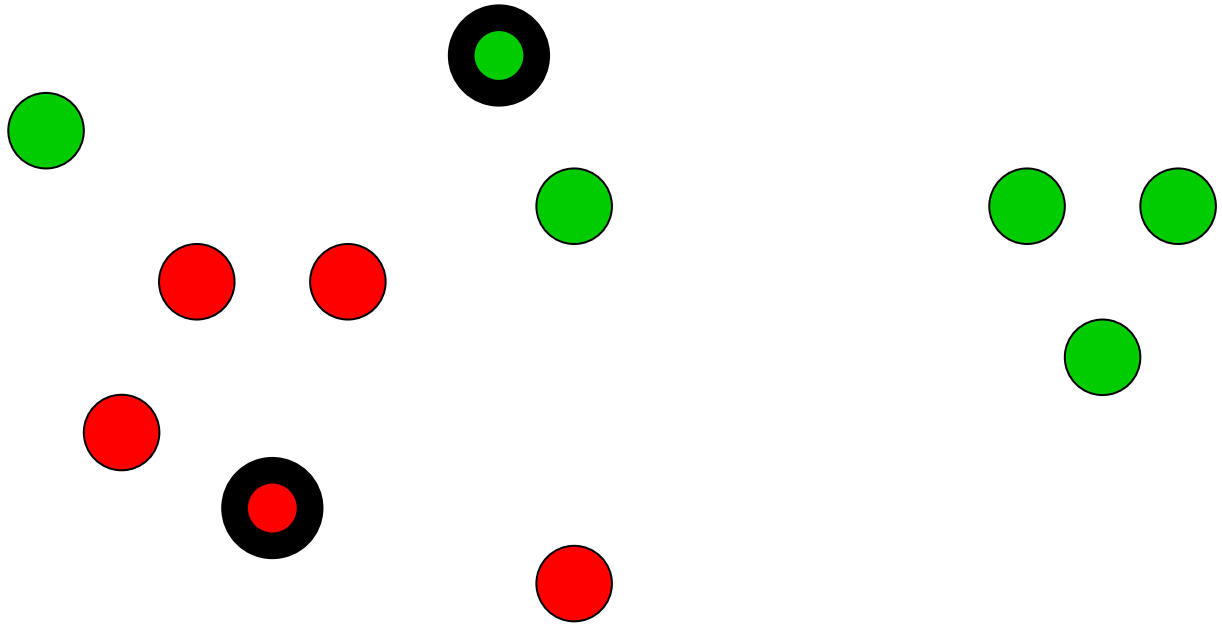
- Here is the data; we choose k = 2 and run 2-means



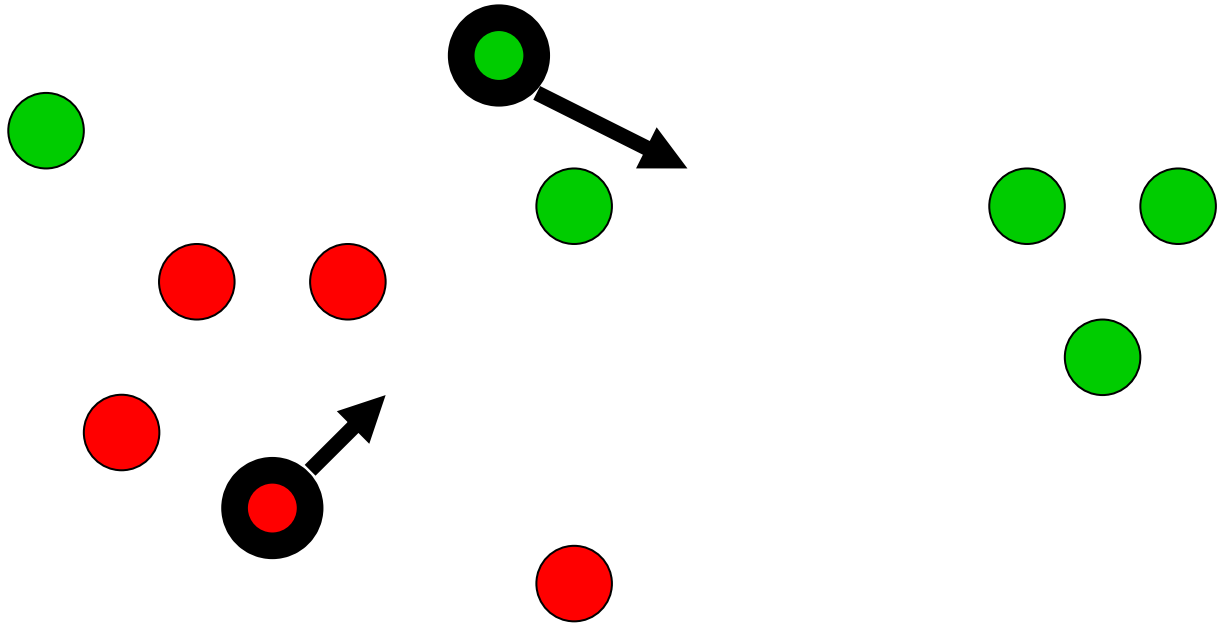
We choose two cluster centres -- randomly



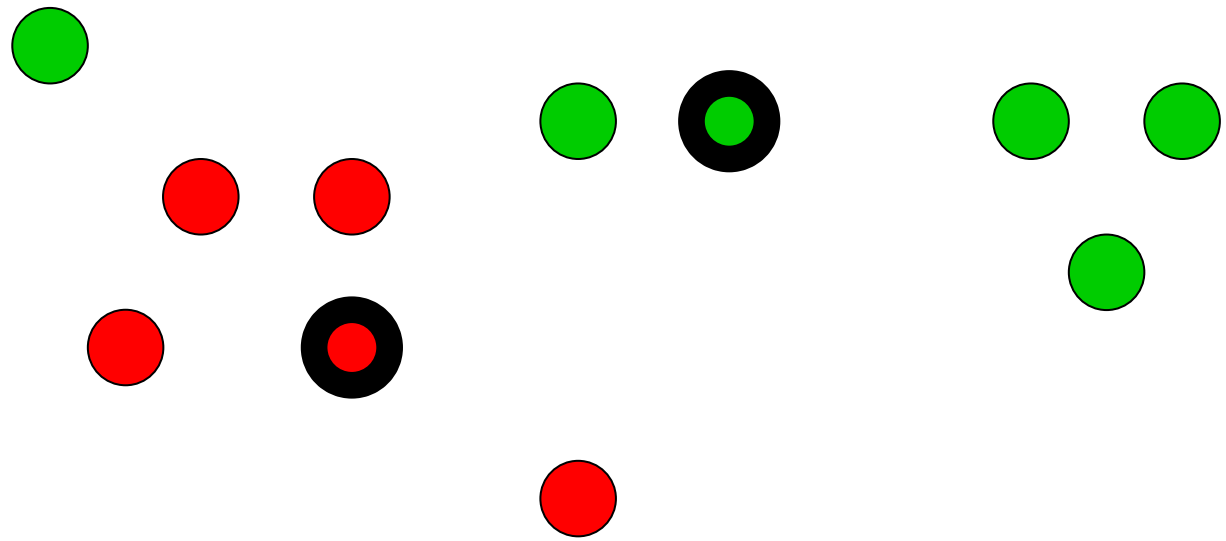
Step 1: decide which cluster each point is in – the one whose centre is closest



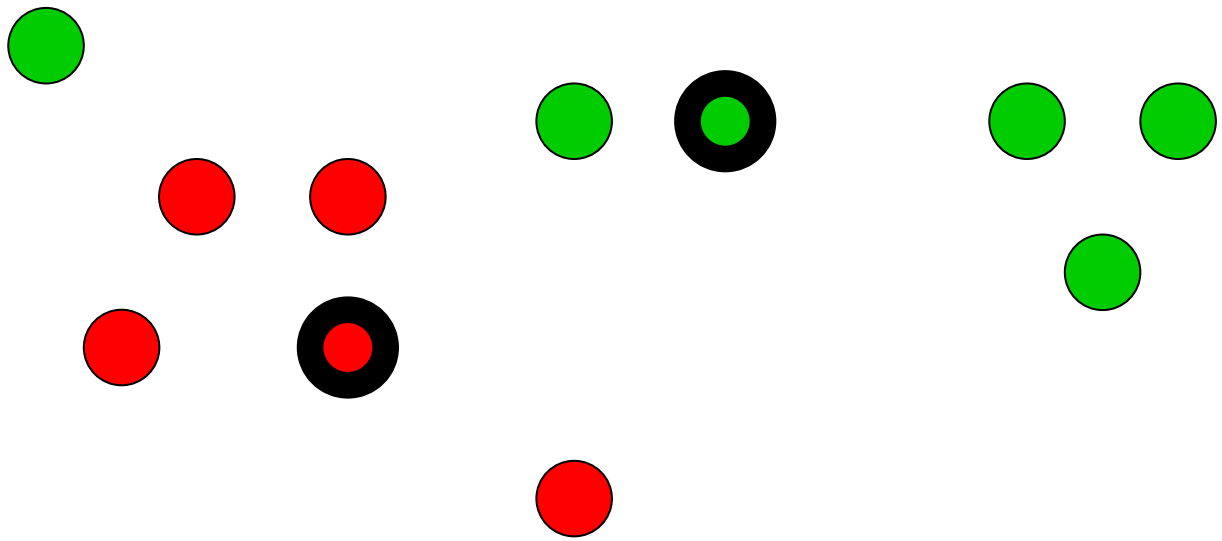
Step 2: We now have two clusters – recompute the centre of each cluster



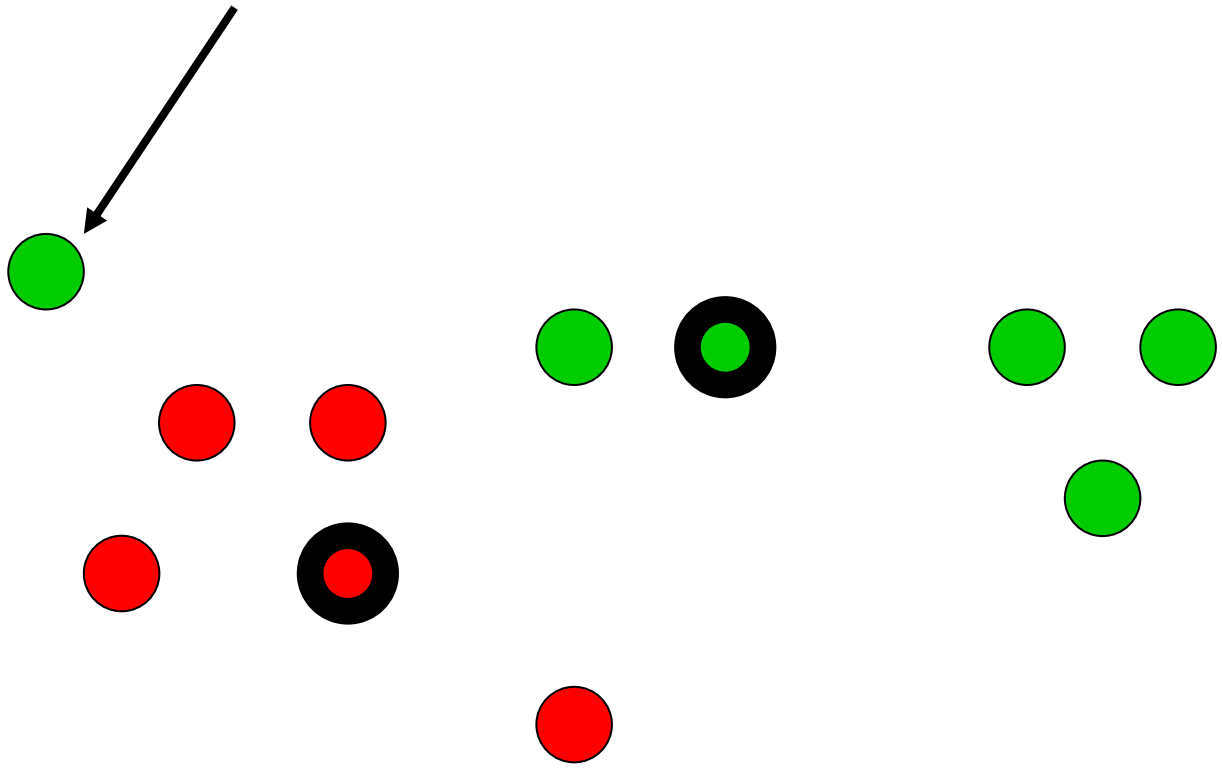
These are the new centres



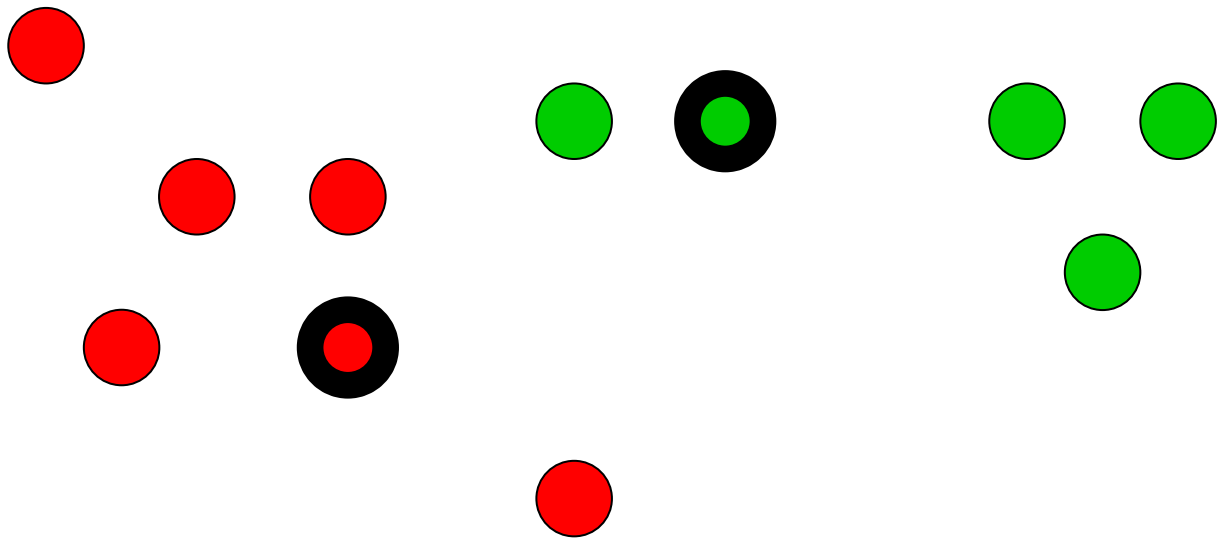
Step 1: decide which cluster each point is in – the one whose centre is closest



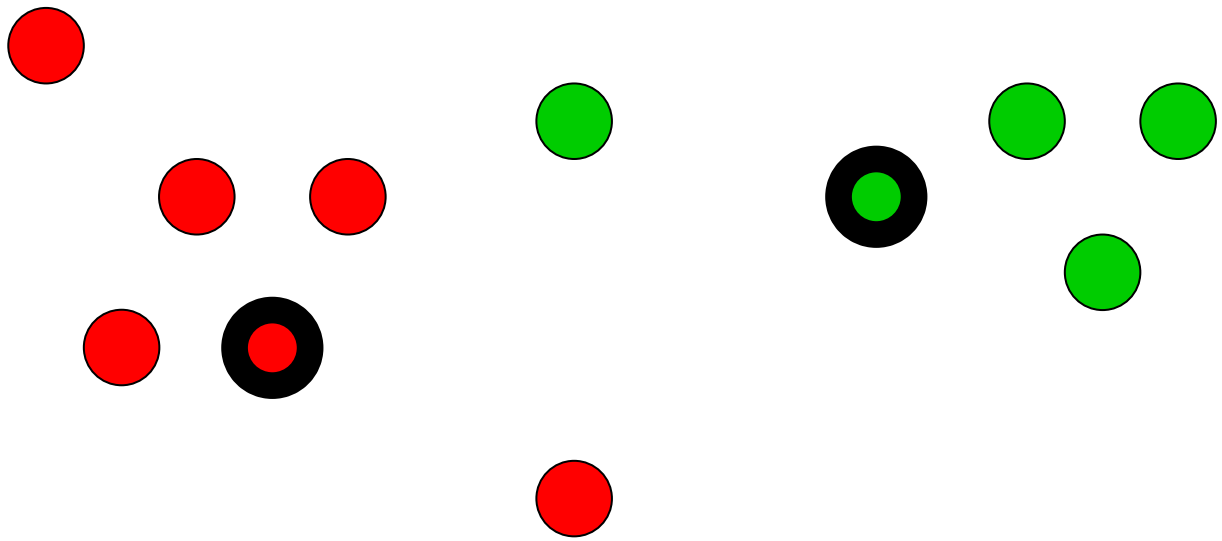
This one has to be reassigned:



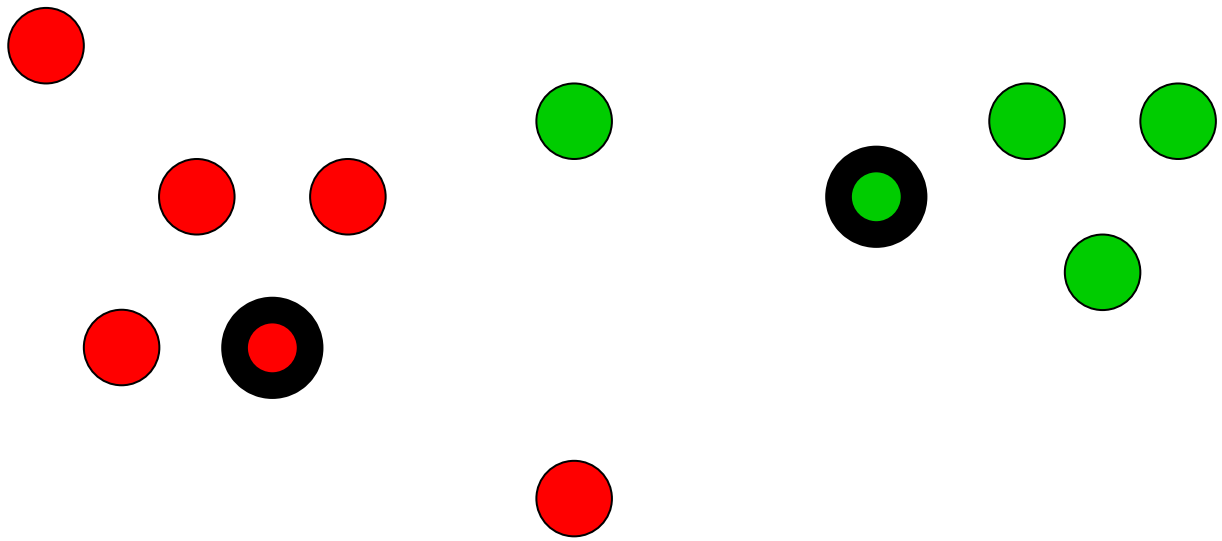
Step 2: We now have two clusters – recompute the centre of each cluster



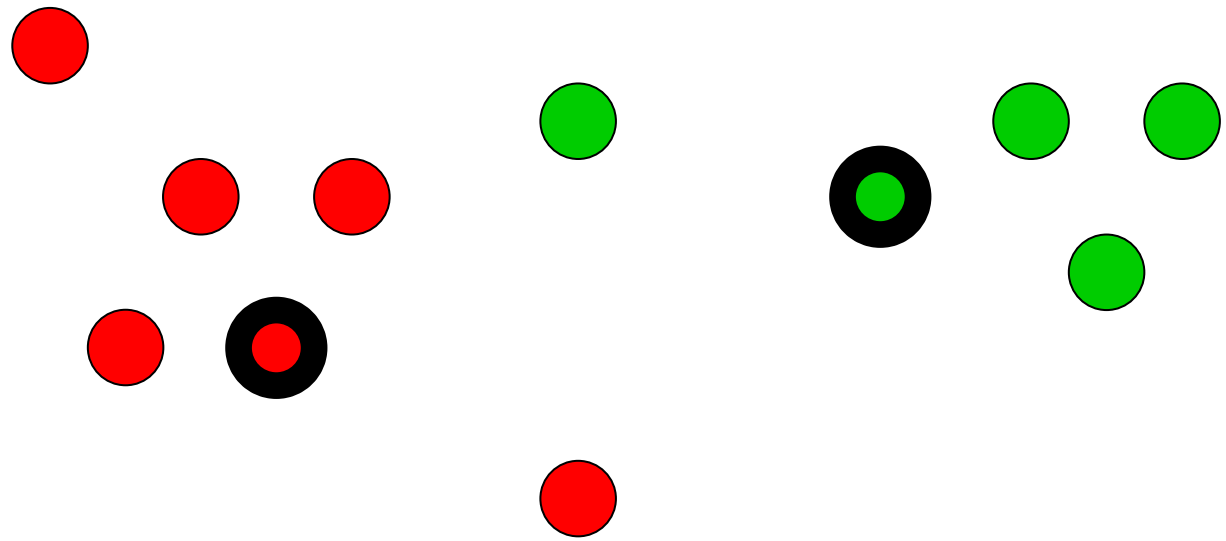
Centres now slightly moved



Step 1: decide which cluster each point is in – the one whose centre is closest



In this case, nothing gets reassigned to a new cluster – so the algorithm is finished



Properties of K–means algorithm

- It is guaranteed to converge in a finite number of iterations.

- Running time per iteration:
    1. Assign data points to the closest cluster center

       O(KN) time

    2. Change the cluster center to the average of its assigned points

       O(N)

Kmeans Convergence

Objective

$$\min_{\mu}\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

1. Fix μ, optimize C:

Step 1 of kmeans

$$\min_{C} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2 = \min_{c} \sum_{i}^{n} \left|x_i - \mu_{x_i}\right|^2$$

2. Fix C, optimize μ:

$$\min_{\mu} \sum_{i=1}^{k} \sum_{x \in C_i} |x - \mu_i|^2$$

   – Take partial derivative of μ_i and set to zero, we have

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Step 2 of kmeans

Kmeans takes an alternating optimization approach, each step is guaranteed to decrease the objective – thus guaranteed to converge

Example: K-Means for Segmentation

K=2

The goal of Segmentation is to partition an image into regions, each of which has a reasonably homogenous visual appearance.

Original

Example: K-Means for Segmentation

K=2

K=3

K=10

Original

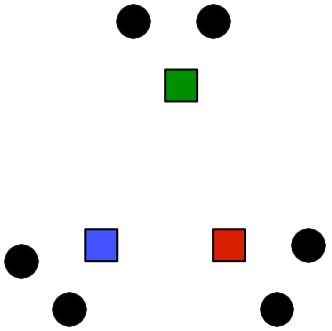
4%

8%

17%

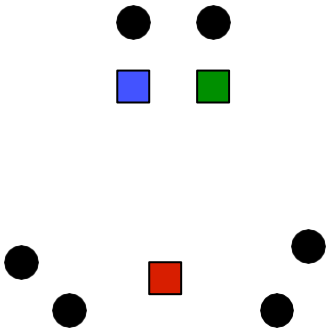
Initialization

- K-means **algorithm** is a heuristic
  - Requires initial means
  - It does matter what you pick!

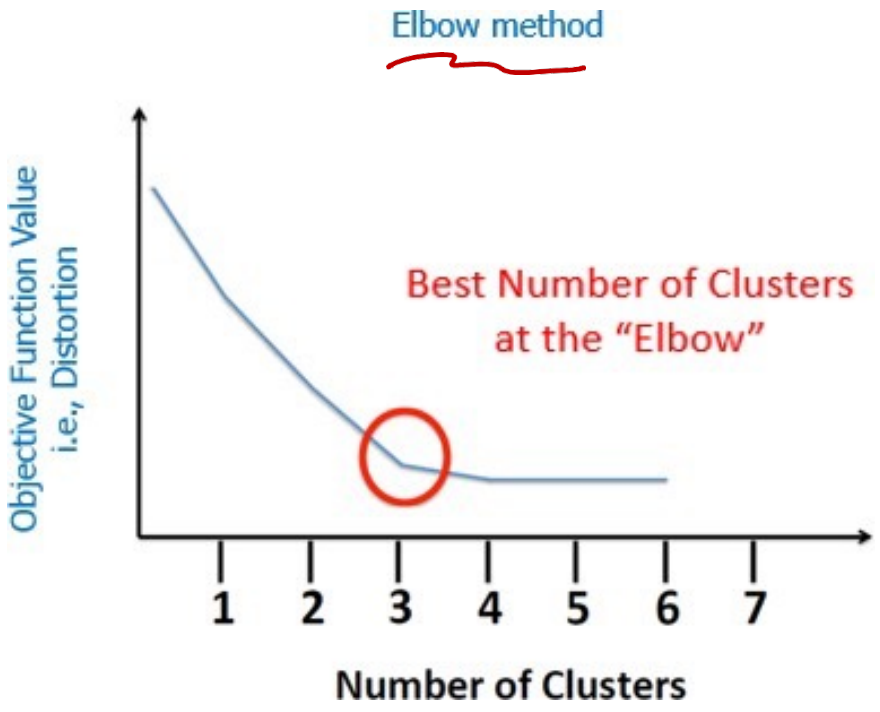


  - What can go wrong?

  - Various schemes for preventing this kind of thing: variance-based split / merge, initialization heuristics



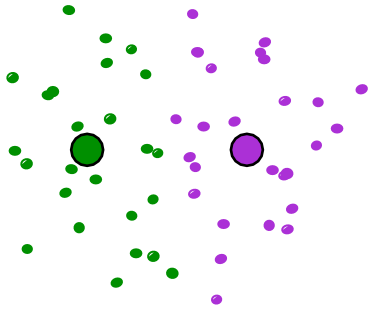
How to Choose K?



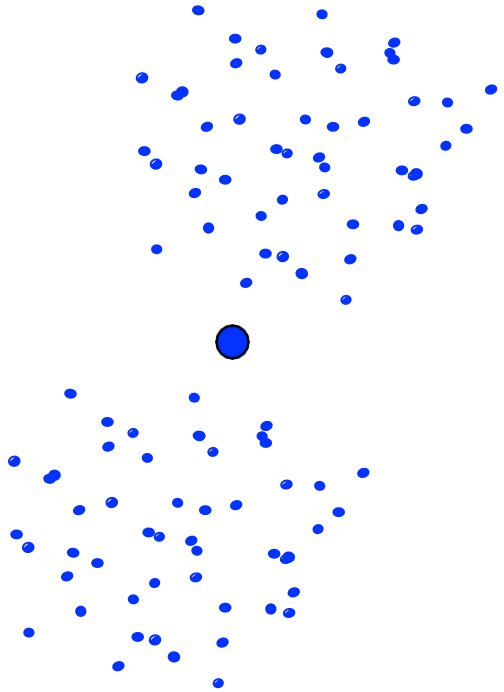
Distortion score: computing the sum of squared distances from each point to its assigned center

K-Means Getting Stuck

A local optimum:

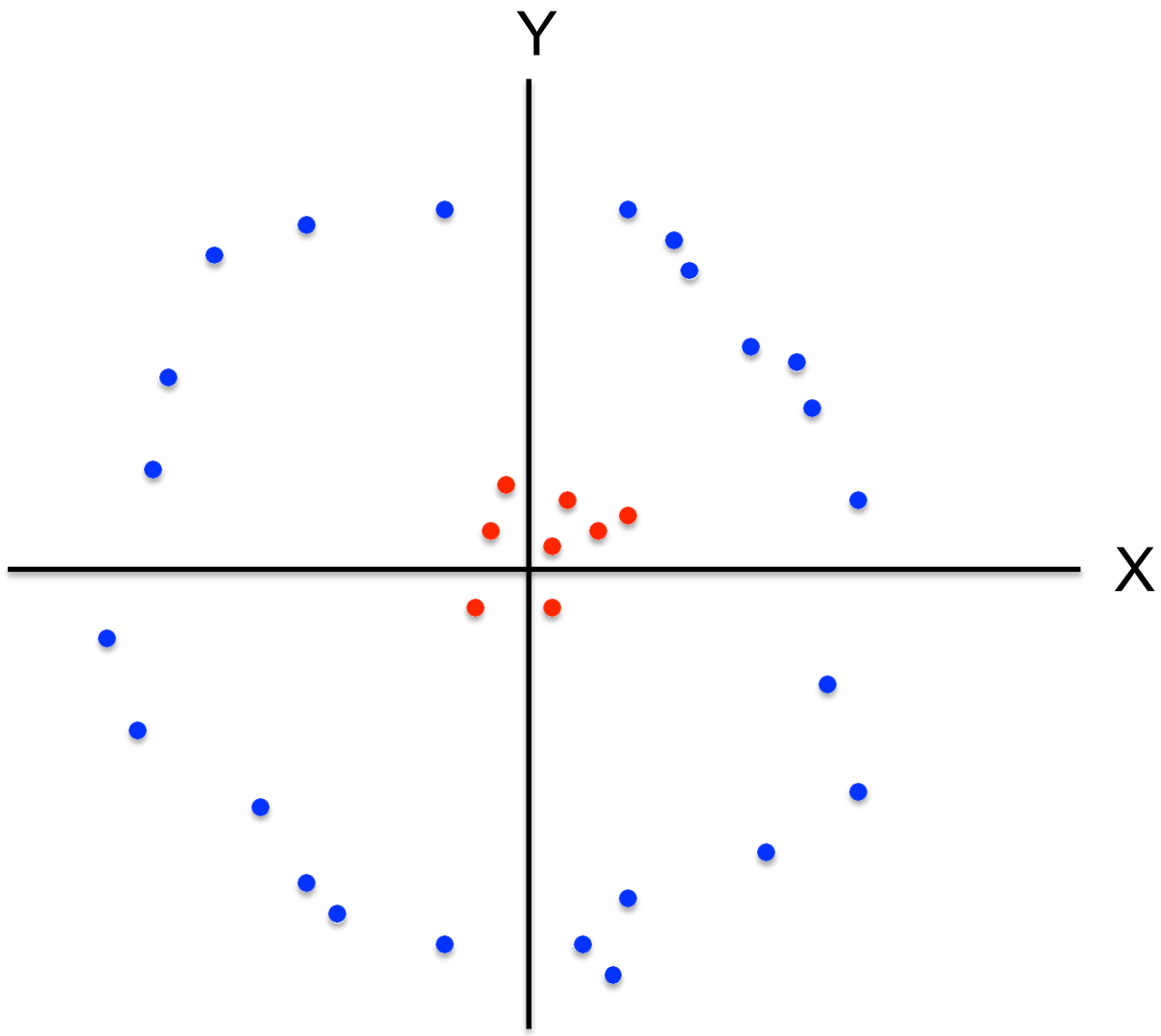


Would be better to have
one cluster here



… and two clusters here

K-means not able to properly cluster



Changing the features can help