WEEKS 5-9

Introduction to Machine Learning

Dr Mykola Gordovskyy

Week 8

- Unsupervised learning – background
- K-means clustering

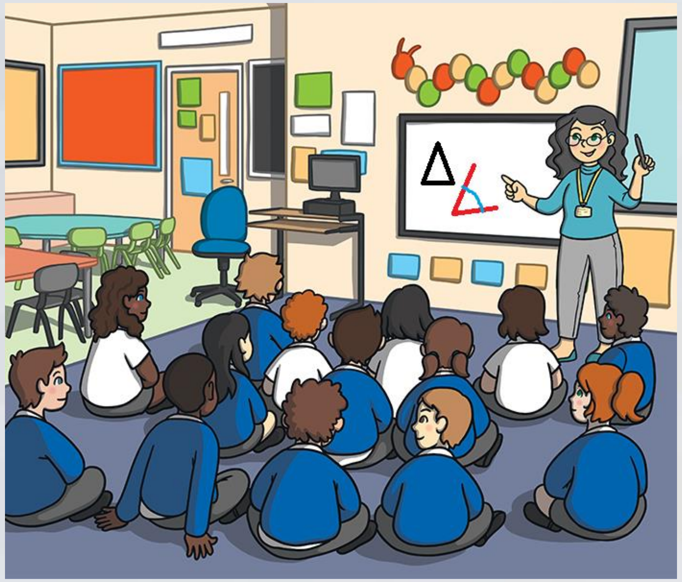
+

- Decision trees revisited - continuous independent variables (“features”)
- Support Vector Machines revisited - math

Unsupervised learning

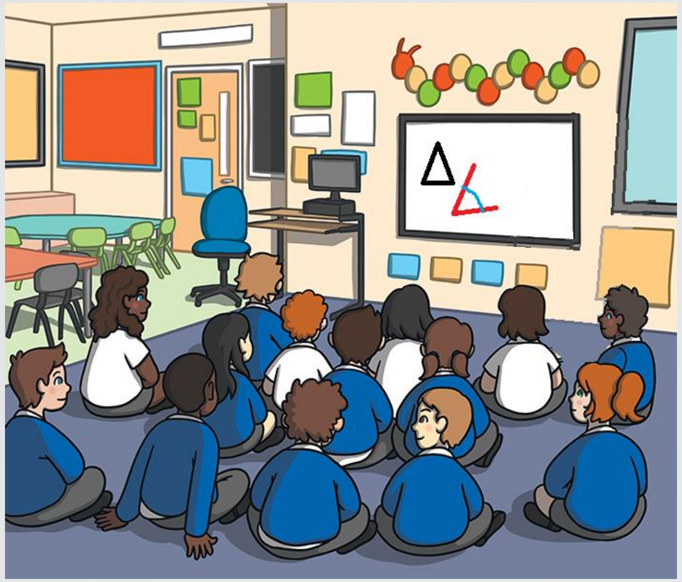
- Requires input data, but no labelling

Supervised learning



Sorting, grouping, predictions are done based on the labels we assign to training data

Unsupervised learning



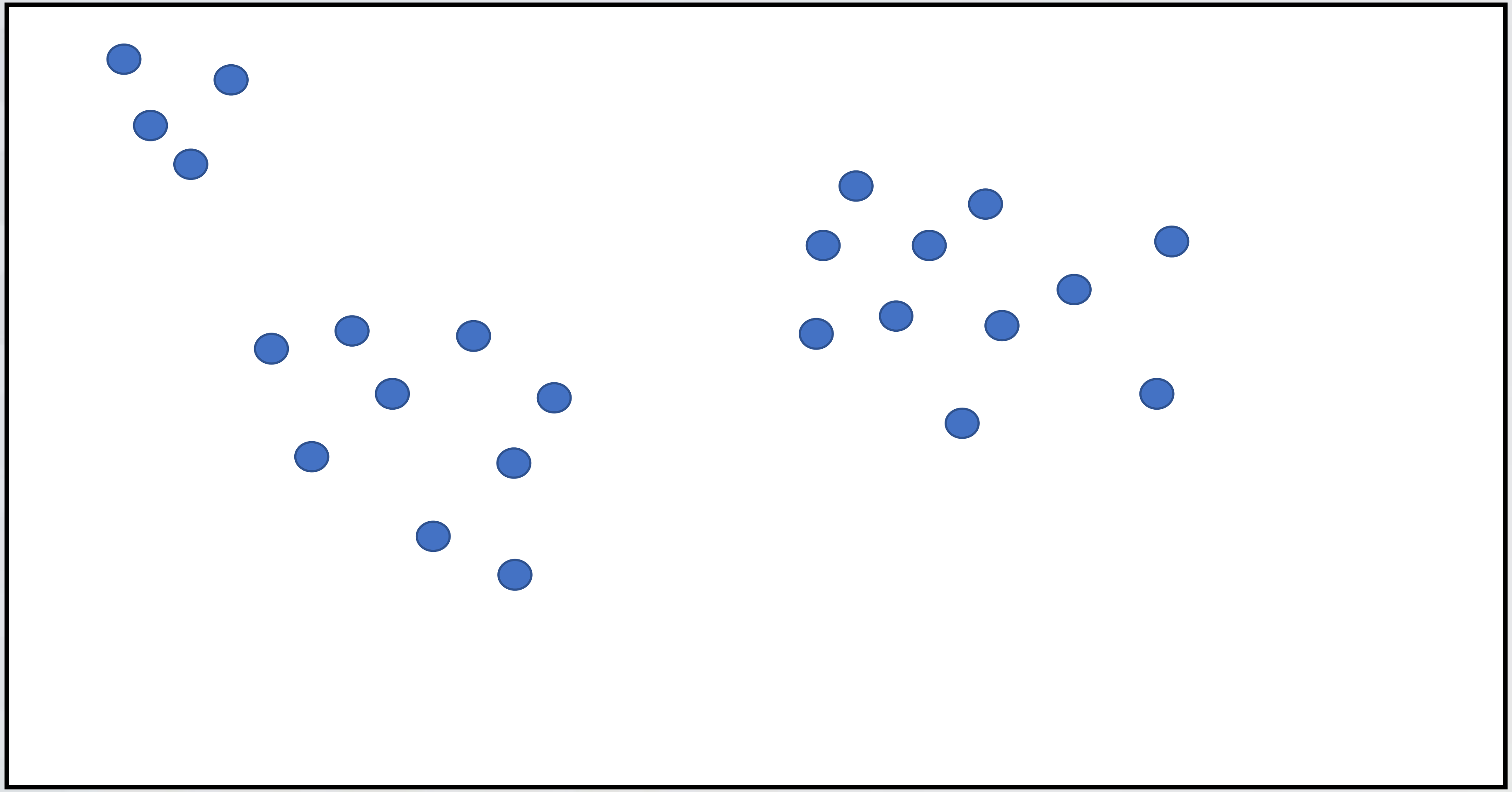
Sorting, grouping, predictions are done based on patterns within the data

Unsupervised learning

- Data clustering
- Detecting patterns and correlations within data
- … and abnormalities within the data
- Powerful approaches when you don't know what you are looking for
- … but within reasons, because it is possible to get "insane" results
- Hence, it is important to "know your data"

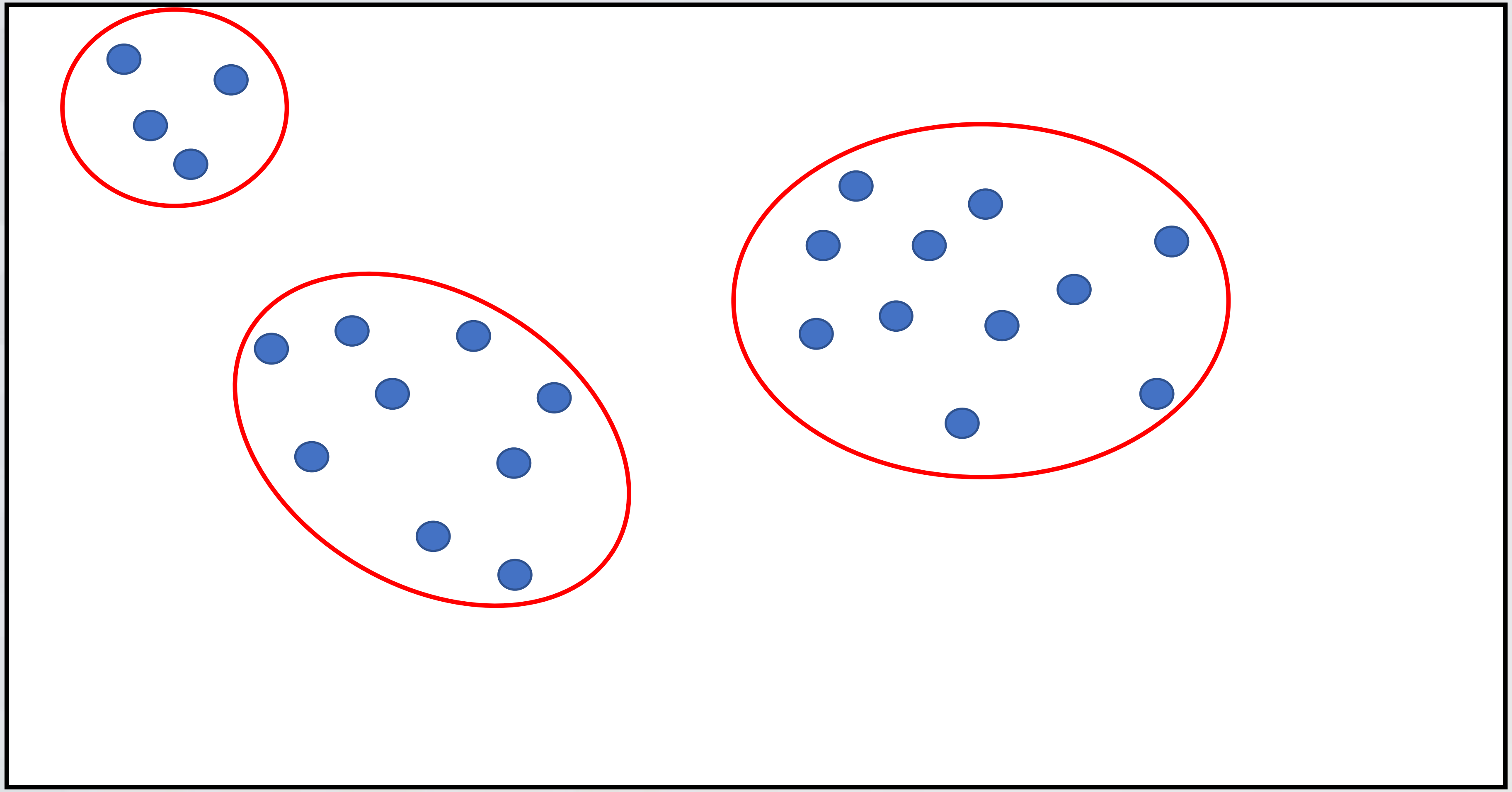
Clustering

- Sort data based on closeness or similarities within the data



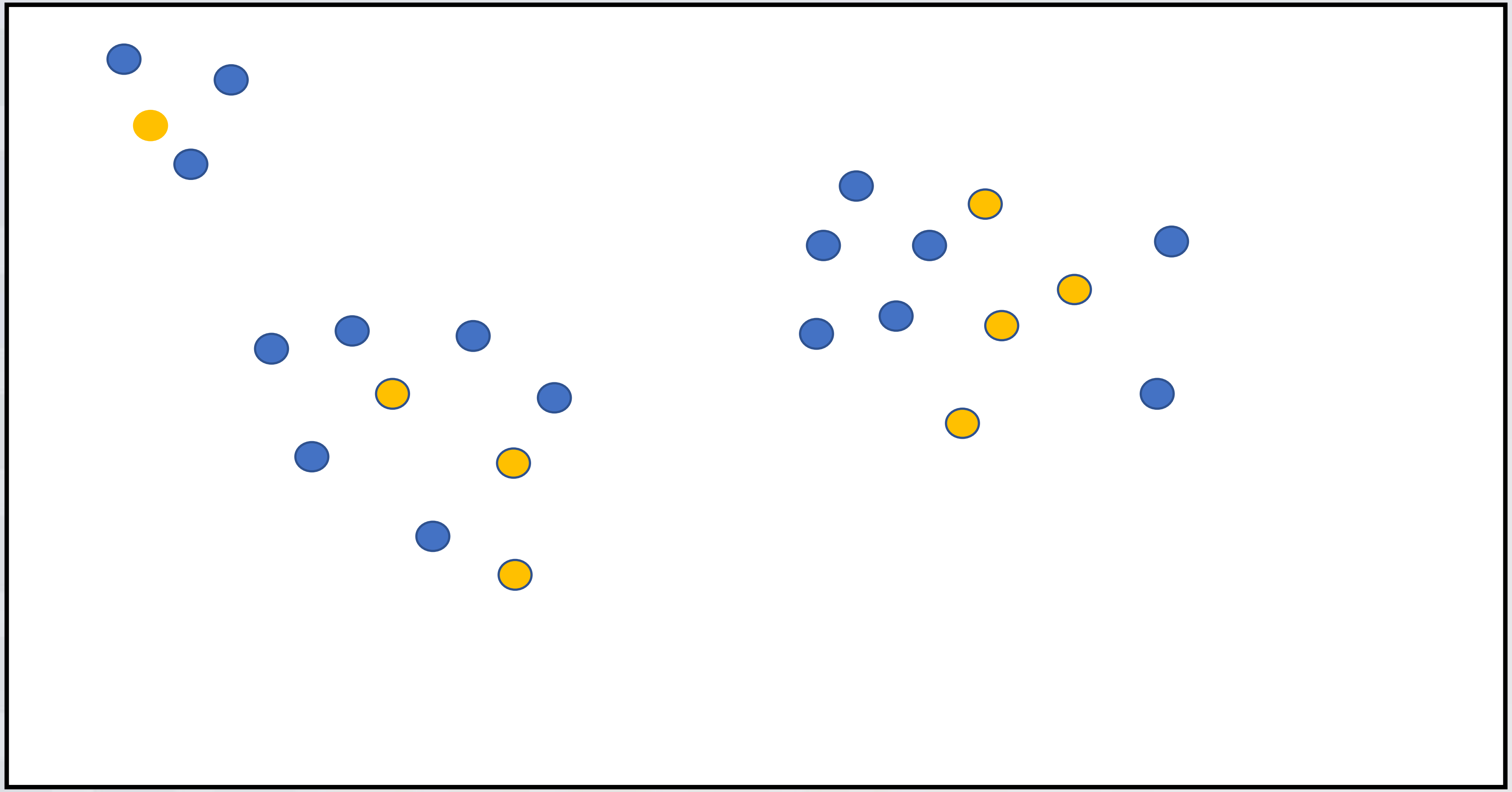
Clustering

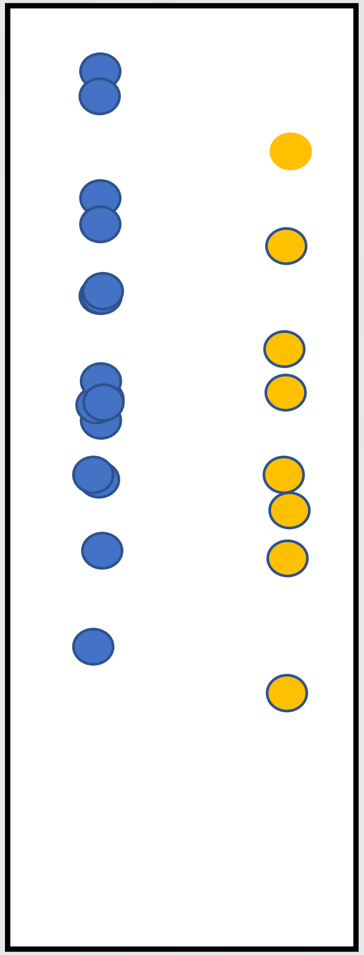
- Sort data based on closeness or similarities within the data



Clustering

- Sort data based on closeness or similarities within the data





Clustering

- Partition algorithms
  - k-Means
  - Spectral clustering
  - Gaussian approaches

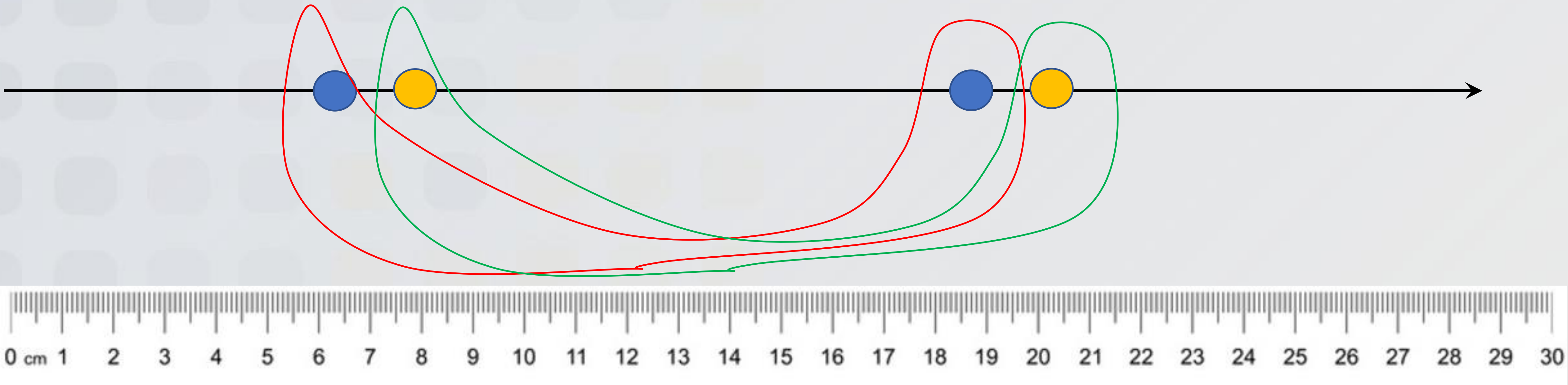
Main criteria: distance within the Euclidean space

- Hierarchical algorithms
  - Agglomerative (from bottom to top)
  - Divisive (top to bottom)

Clustering using different variables separately

Clustering

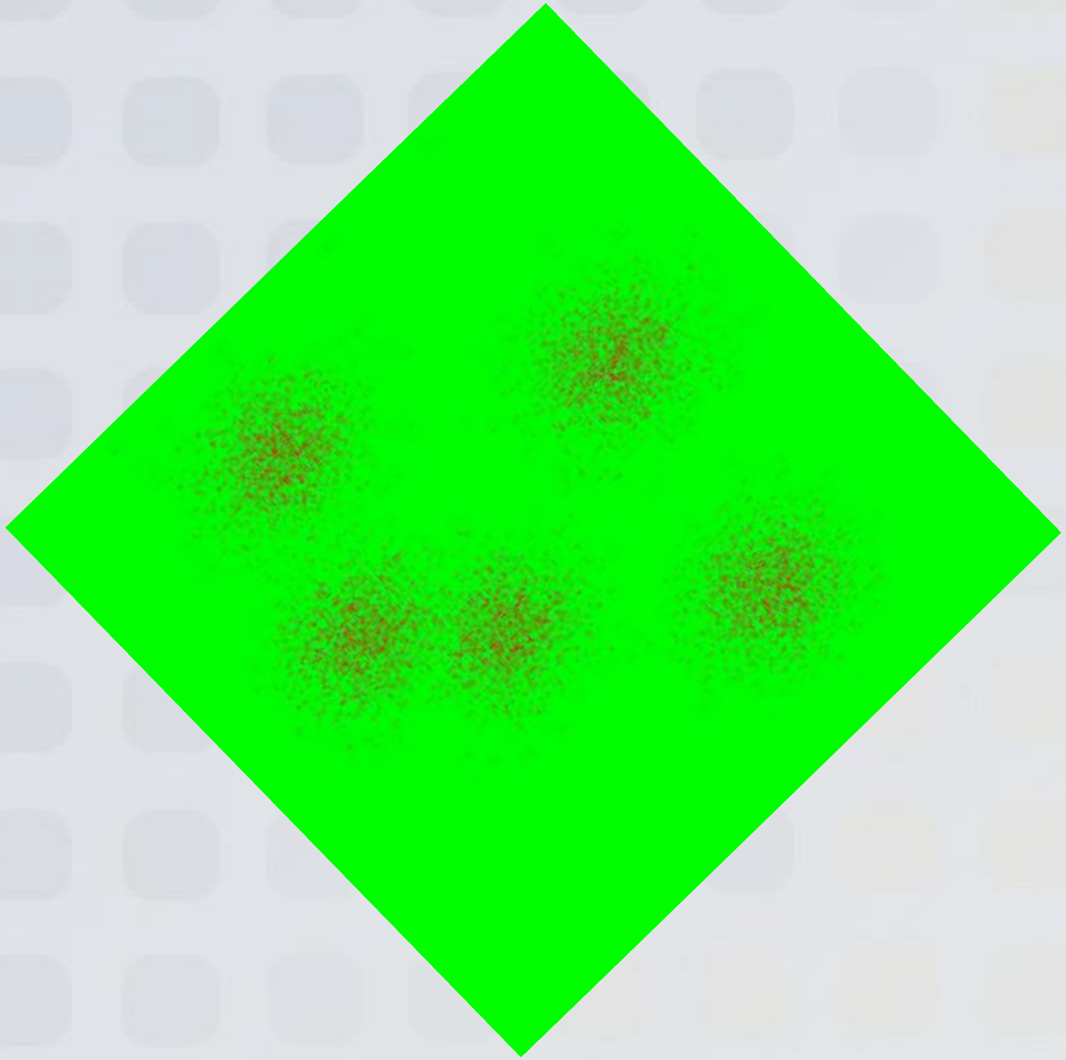
OR



Clustering

- Closeness or similarity?
- … - its really the same, similarity and closeness in Euclidean space
- Difference in qualitative characteristics -> distance?
- Clustering results are very sensitive to the measure of similarity

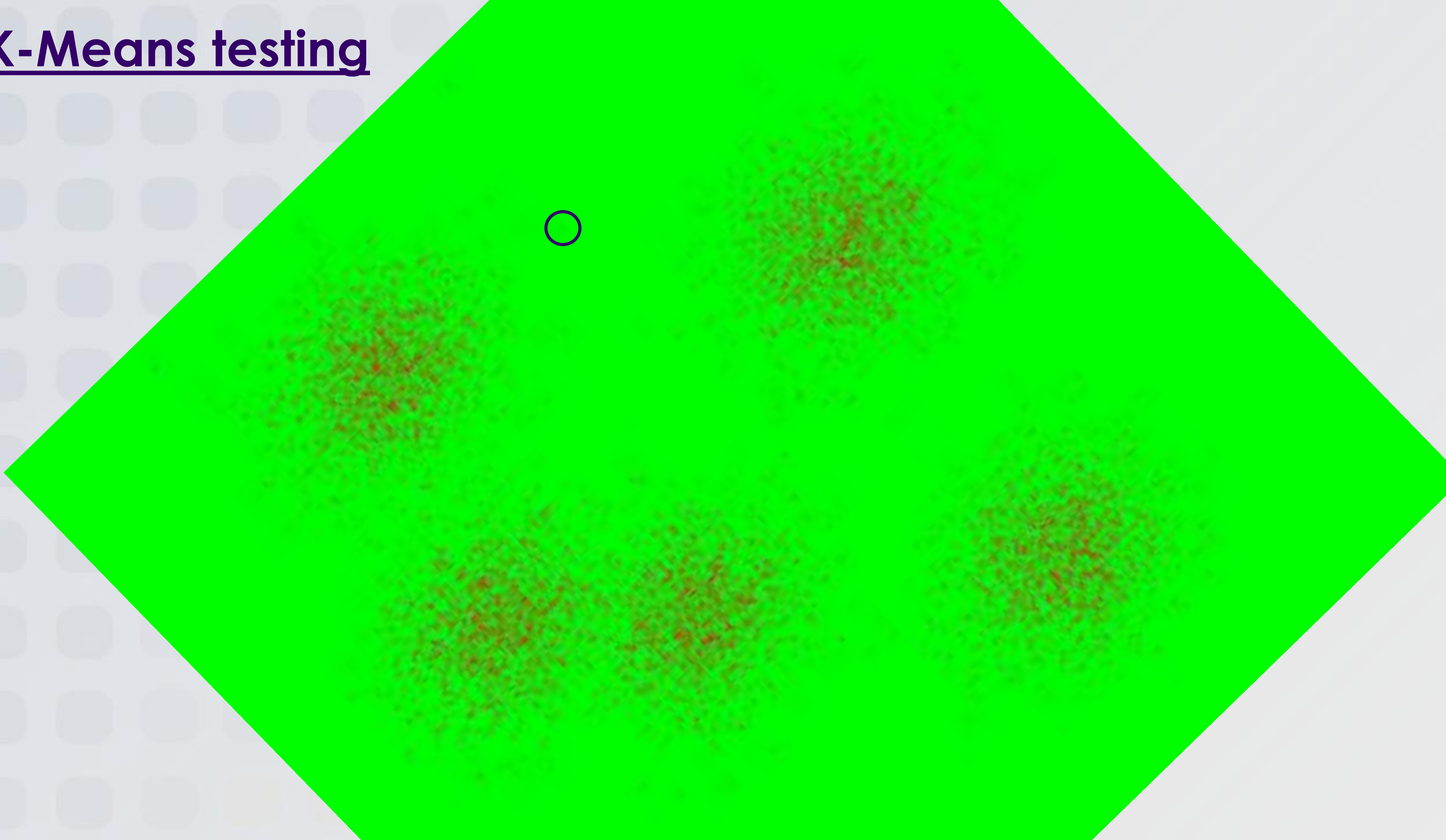
K-Means testing



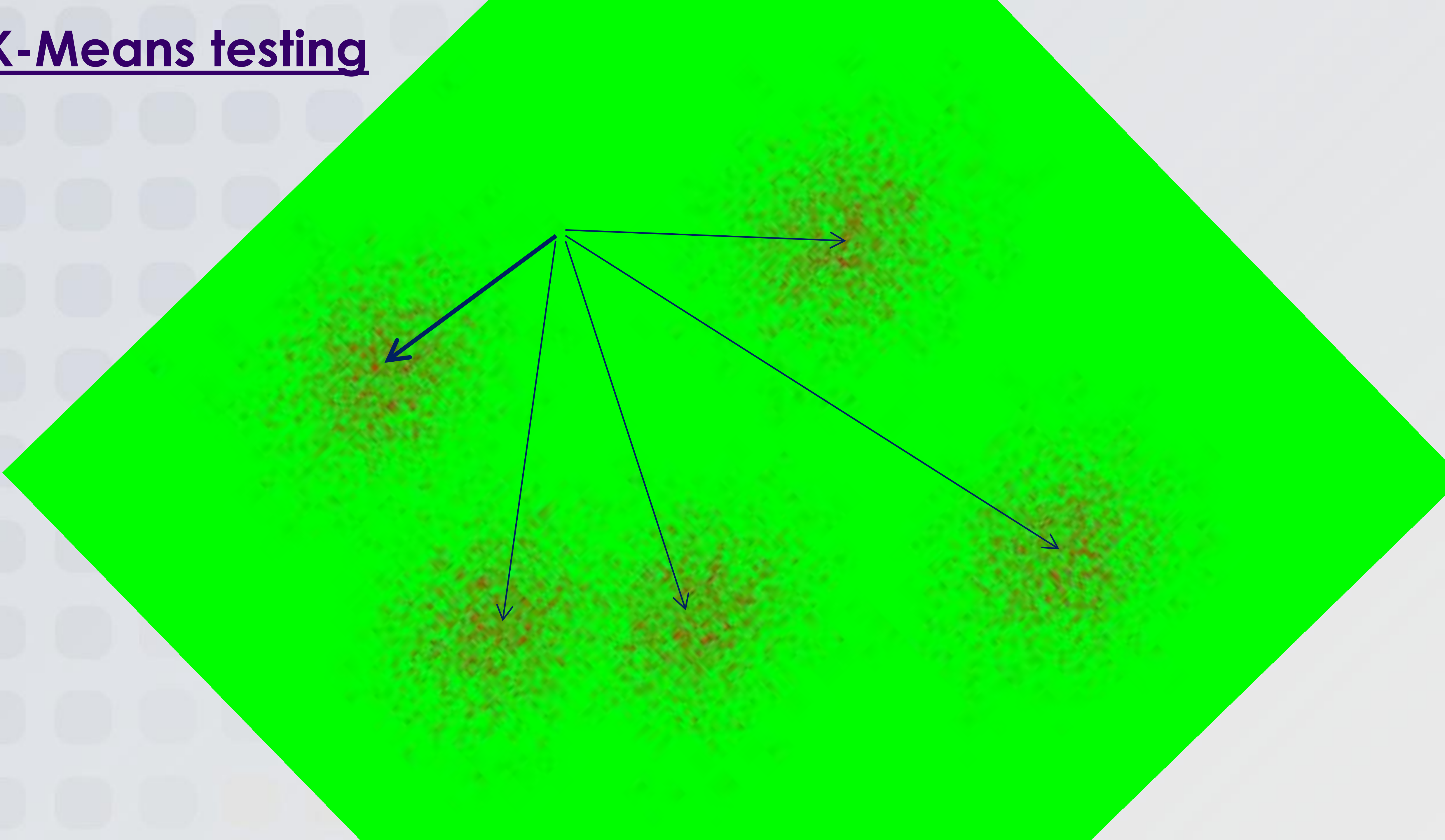
- (You!) Define the number of clusters in your data

- Assign each point to one of the clusters based on the distances between the point and the centers of the clusters. A point is assigned to the cluster whose center is closest to that point.

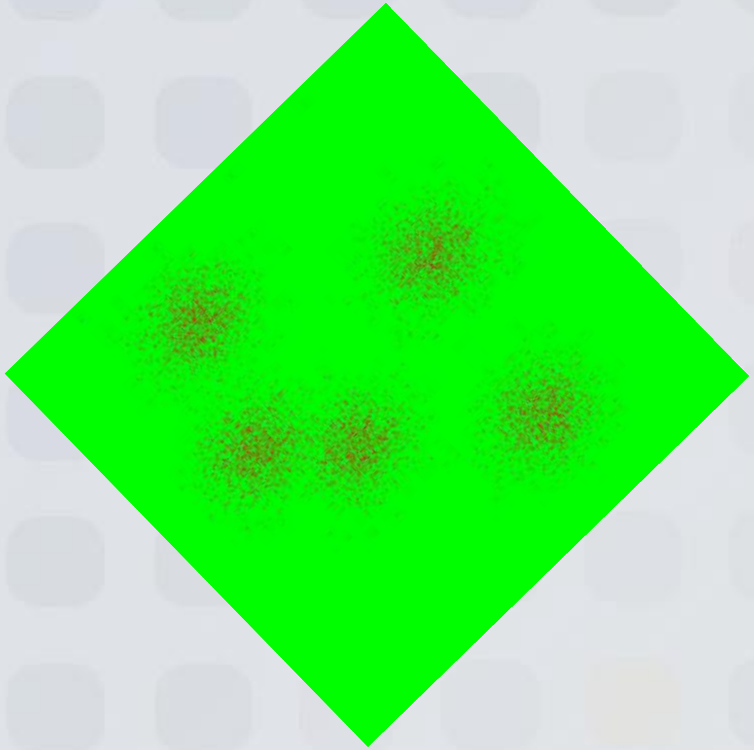
K-Means testing



K-Means testing



K-Means testing



- … but how do we find the centres? – Iterative procedure

0) Initialise

- Pick random points for the centre of each cluster

1) Iterate

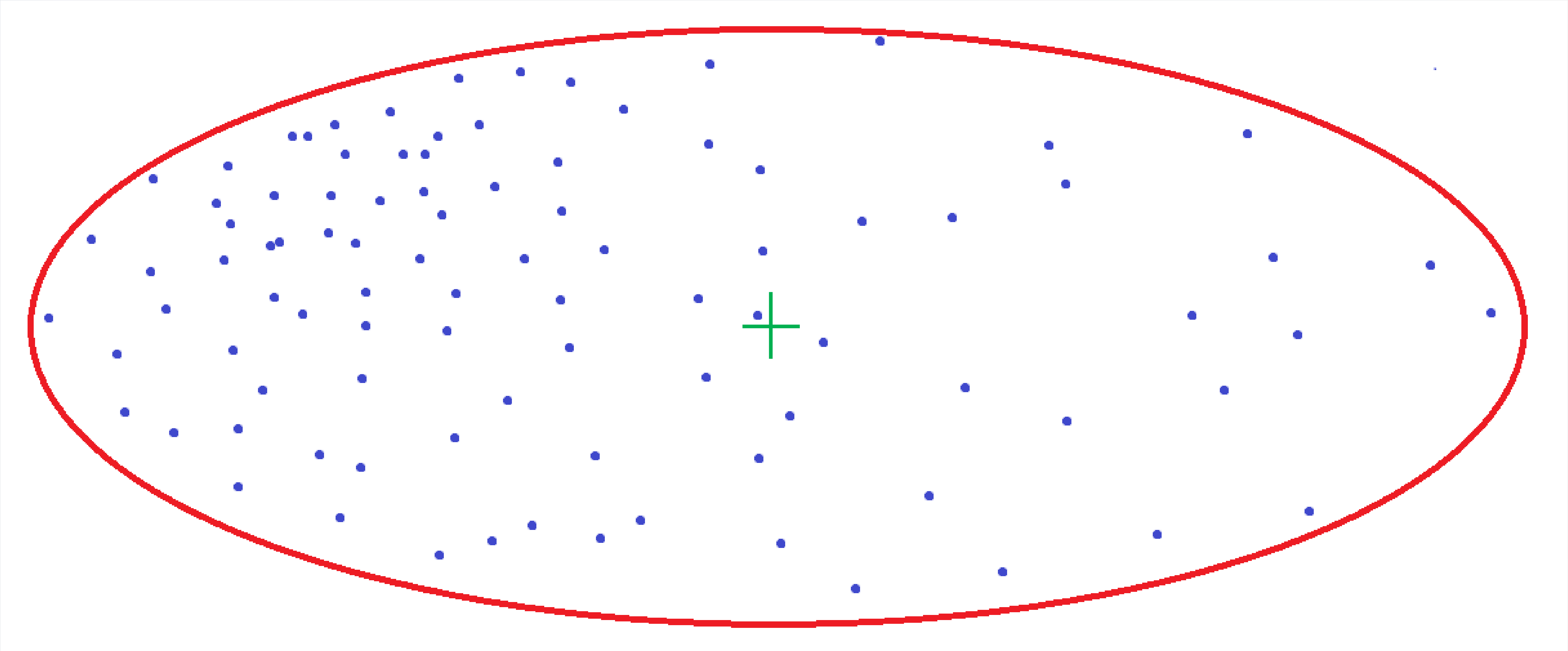
- Assign each data point to a cluster based on the closest distance to one of the centres
- Calculate new locations of cluster centres as an average position for each cluster (i.e. new centres of clusters = centres-of-mass of clusters)
- … repeat unless you can stop

2) Stop

- You can stop when the point labels stop changing

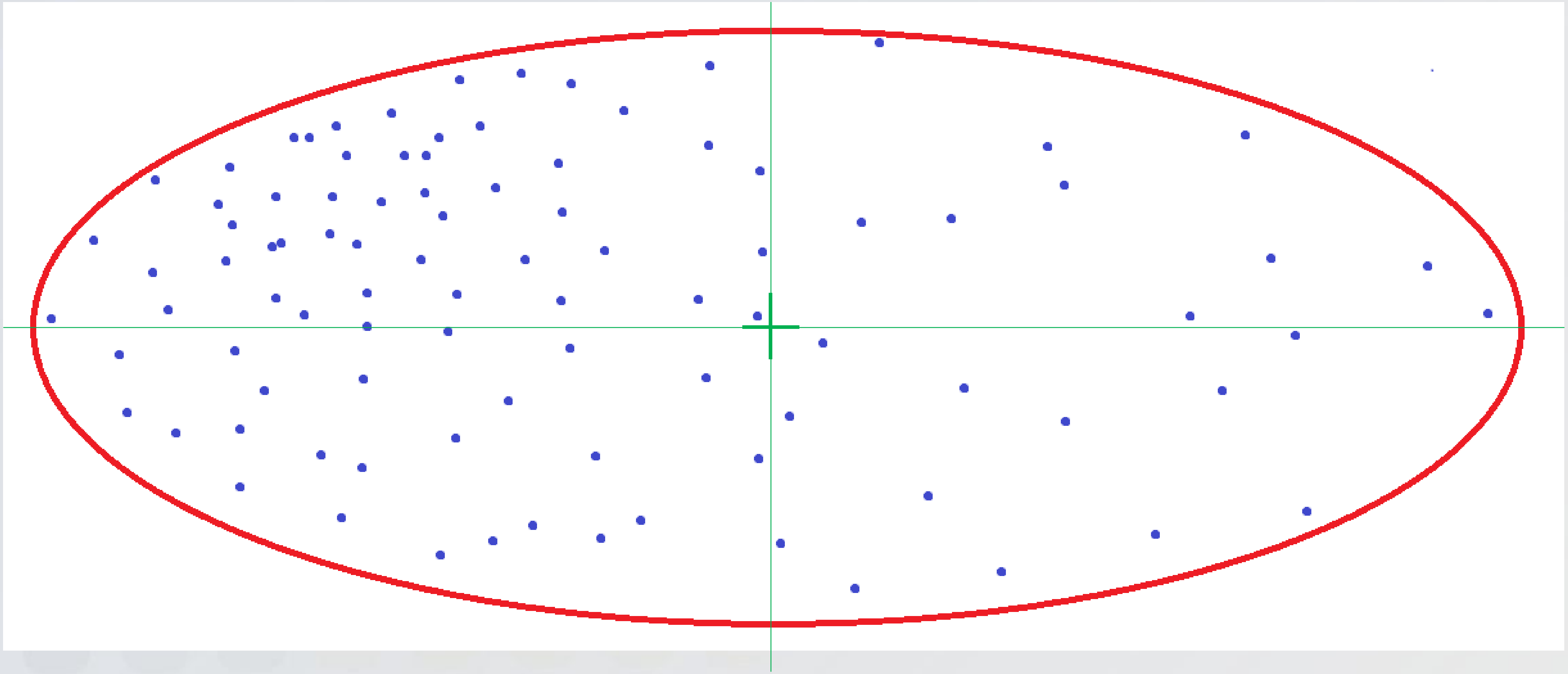
Centre of mass

N points with positions X_i, Y_i



Centre of mass

N points with positions X_i, Y_i



Centre of mass

N points with positions X_i, Y_i

$$X_c = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$Y_c = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

More general case when entries have weights w_i

$$X_c = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

$$Y_c = \frac{\sum_{i=1}^{N} w_i Y_i}{\sum_{i=1}^{N} w_i}$$

Centre of mass

N points with positions X_i, Y_i

$$X_c = \frac{1}{N} \sum_{i=1}^{N} X_i$$

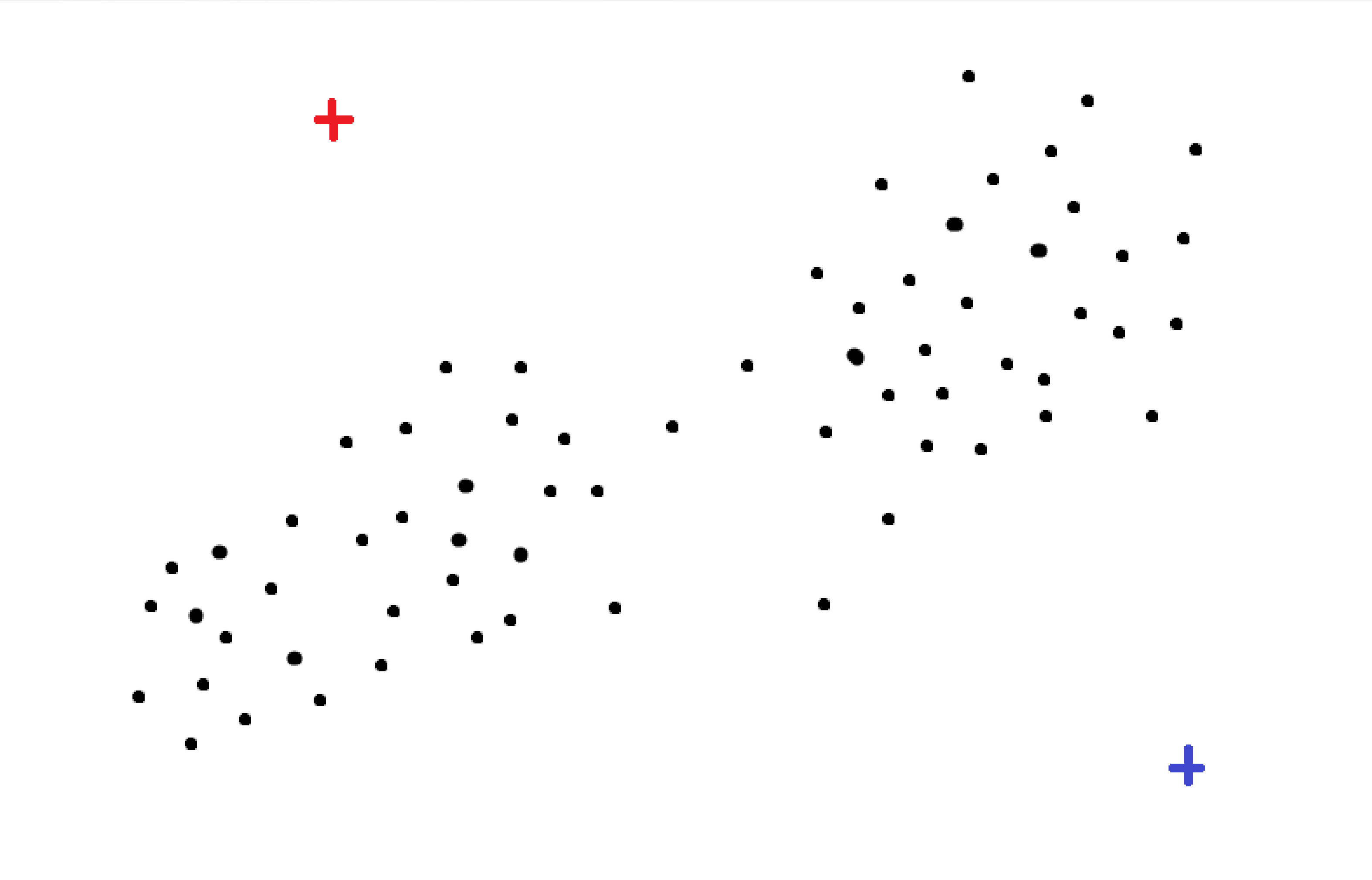
$$Y_c = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

More general case when entries have weights w_i

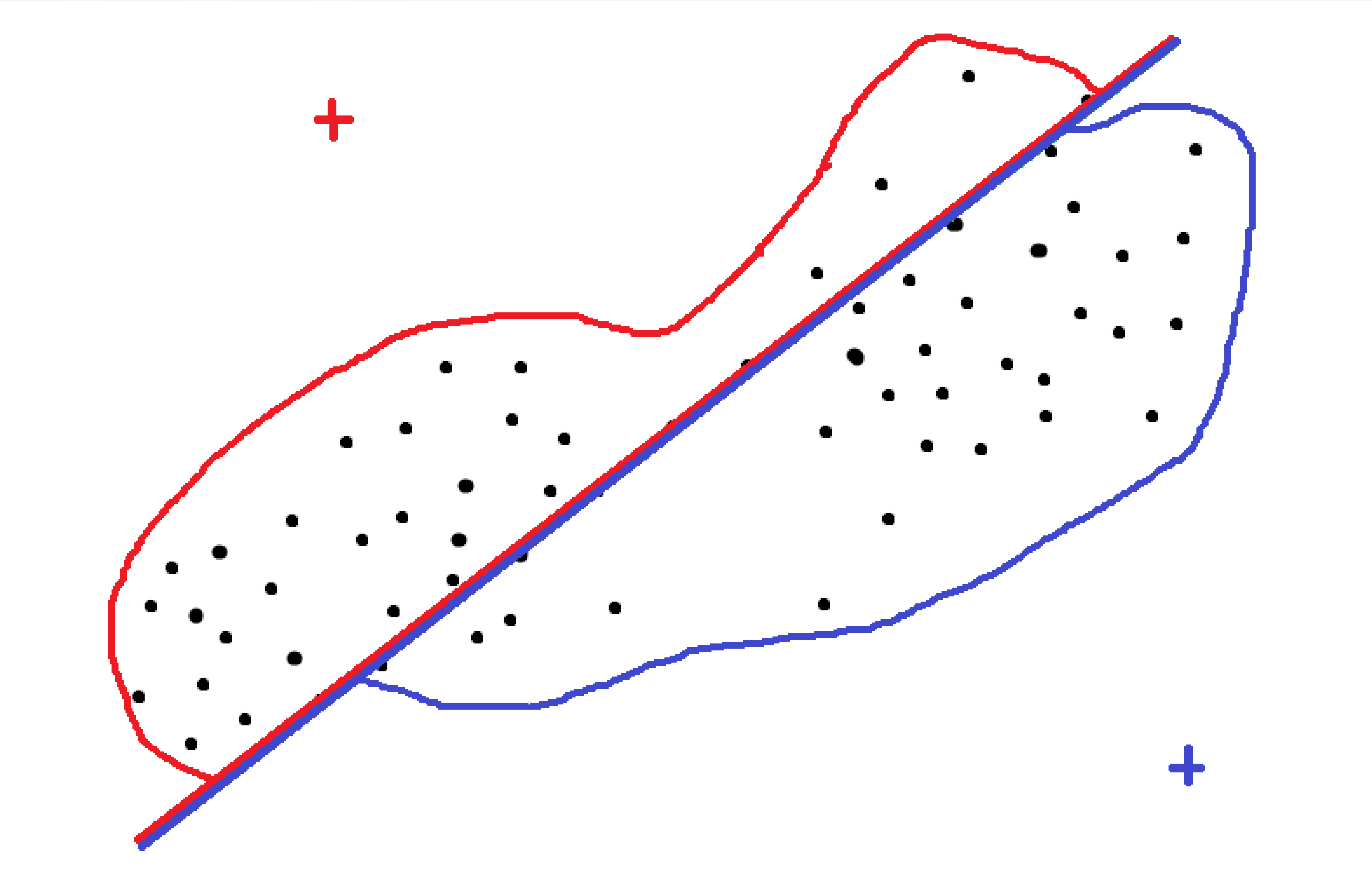
$$X_c = \frac{\sum_{i=1}^{N} w_i X_i}{\sum_{i=1}^{N} w_i}$$

$$Y_c = \frac{\sum_{i=1}^{N} w_i Y_i}{\sum_{i=1}^{N} w_i}$$

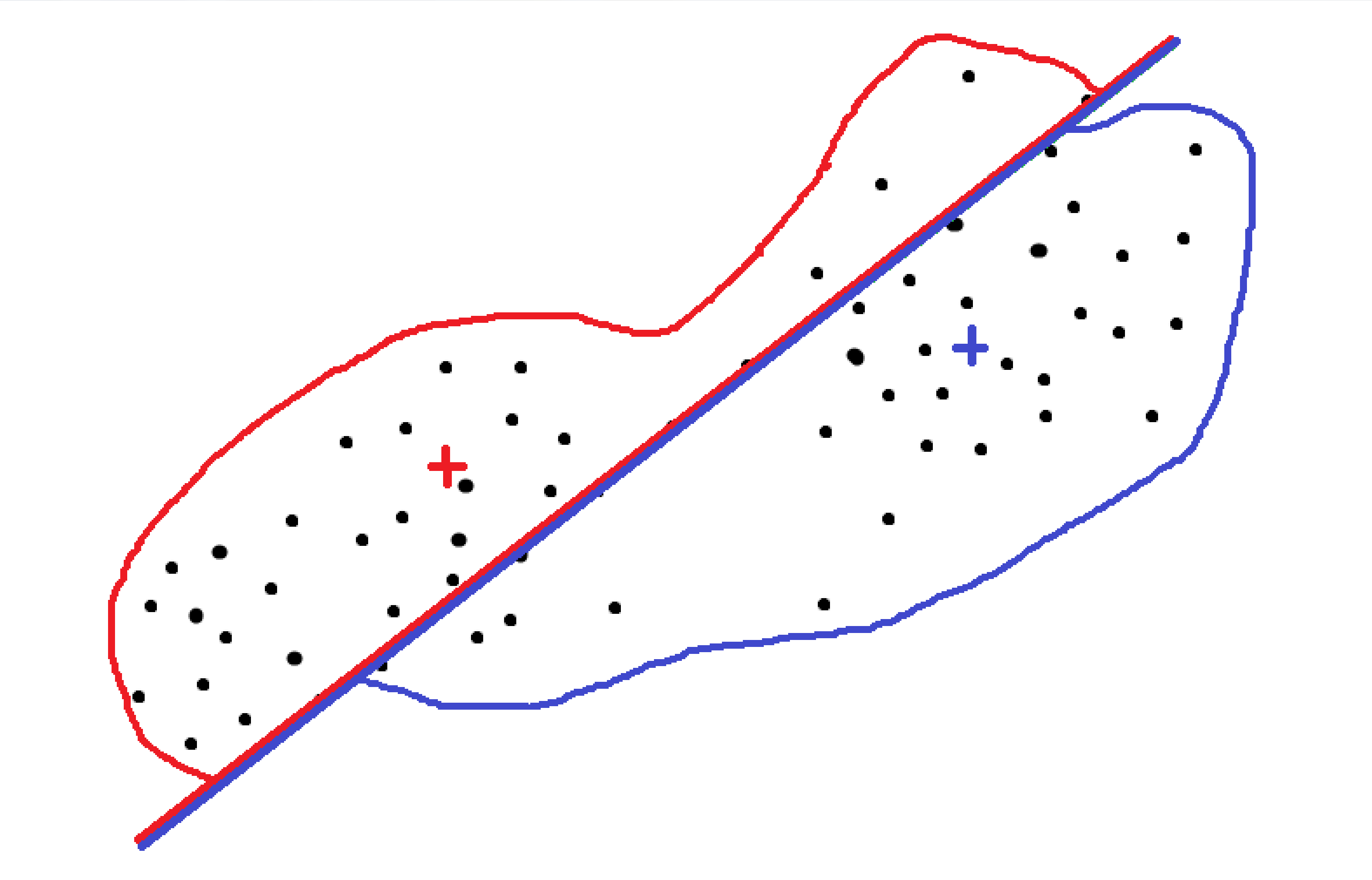
K-Means example



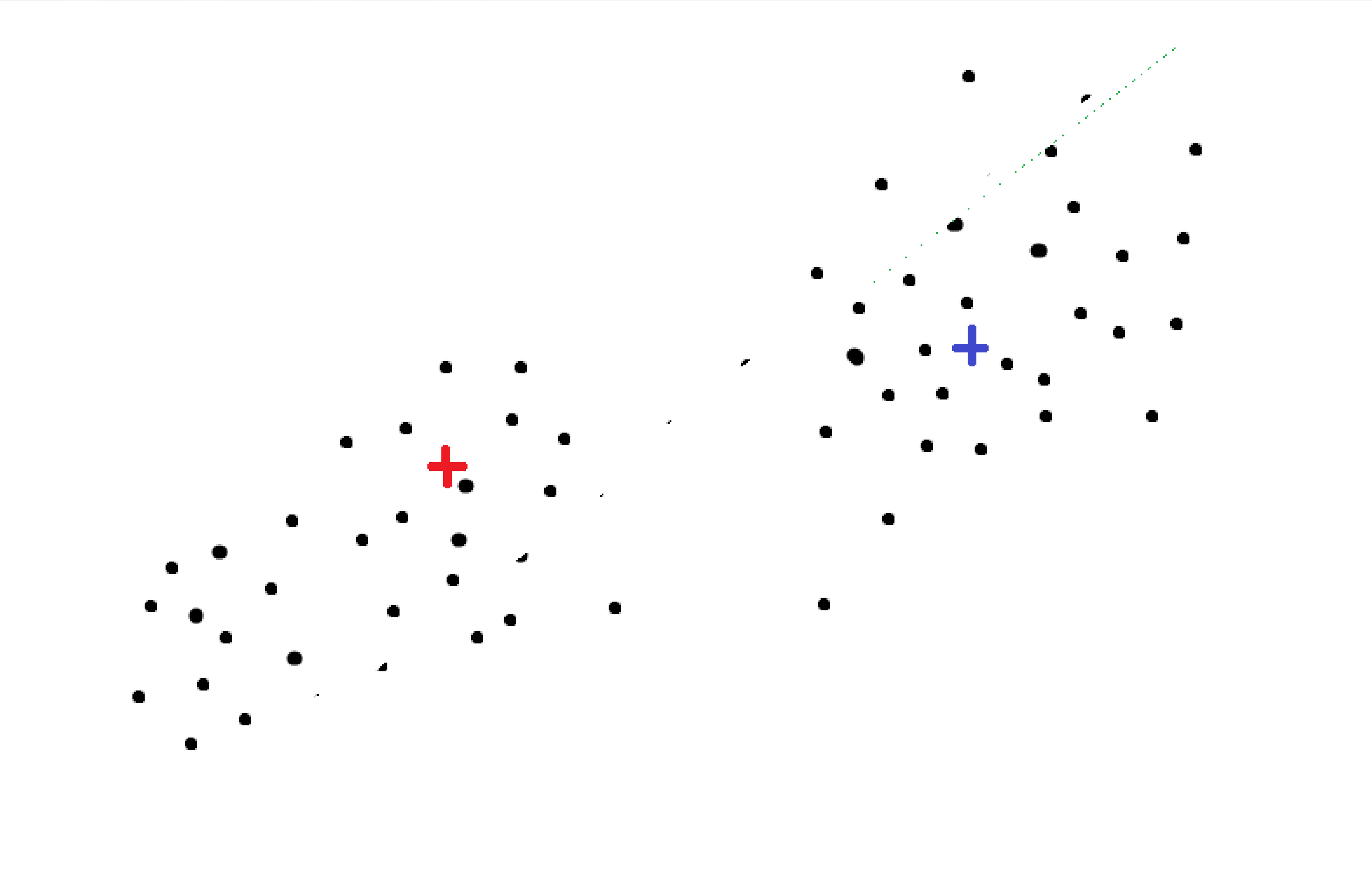
K-Means example



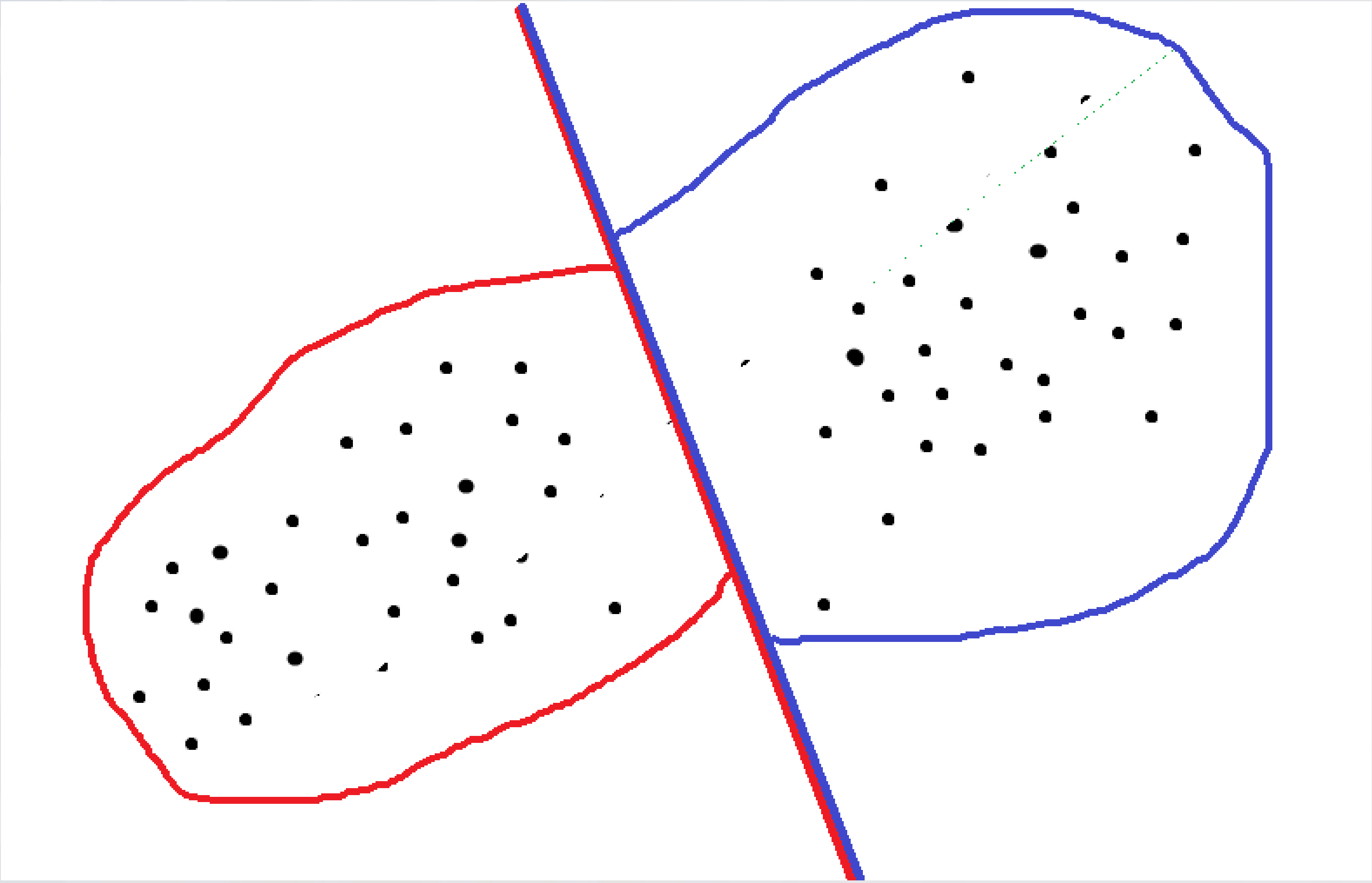
K-Means example



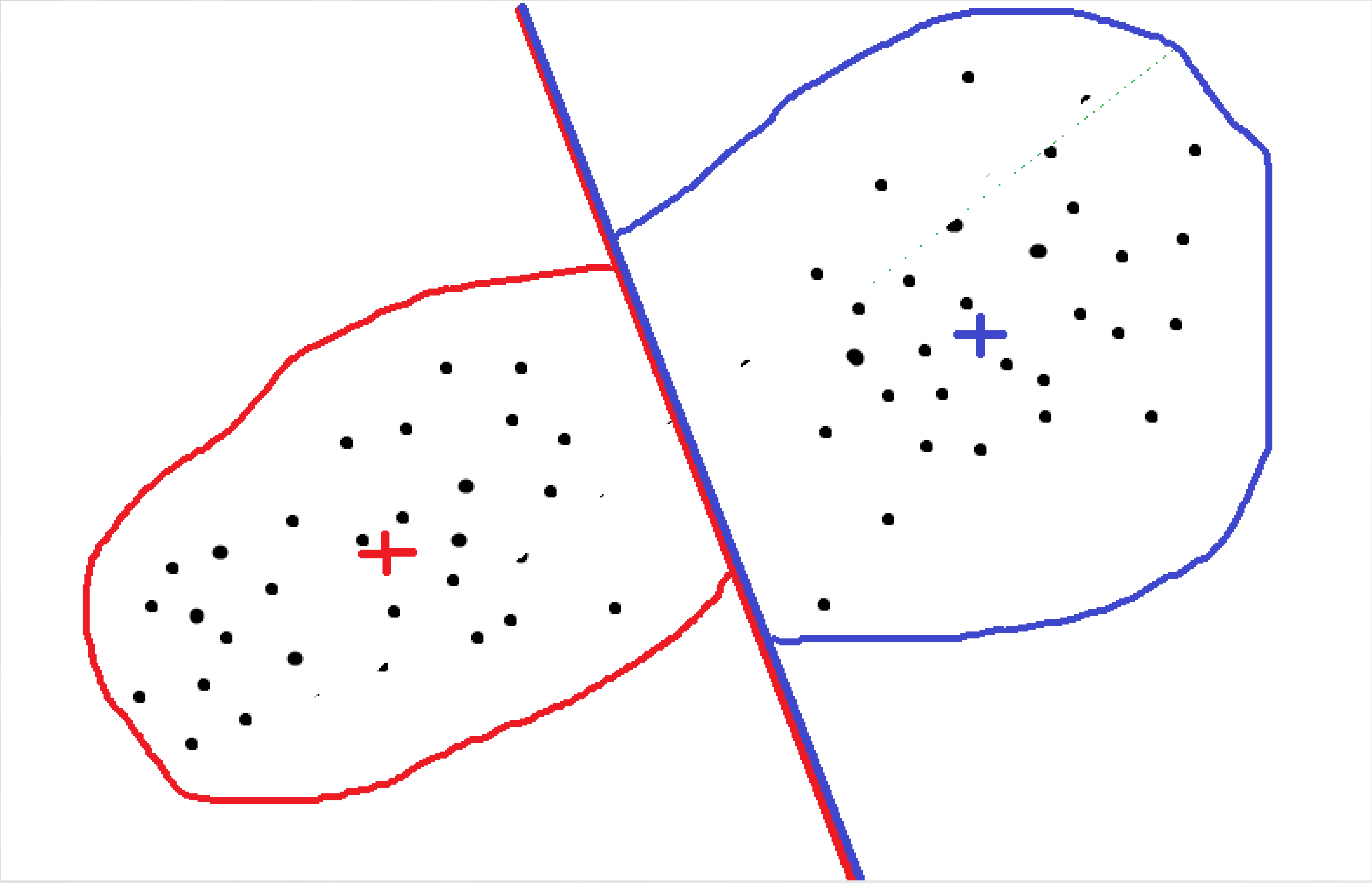
K-Means example



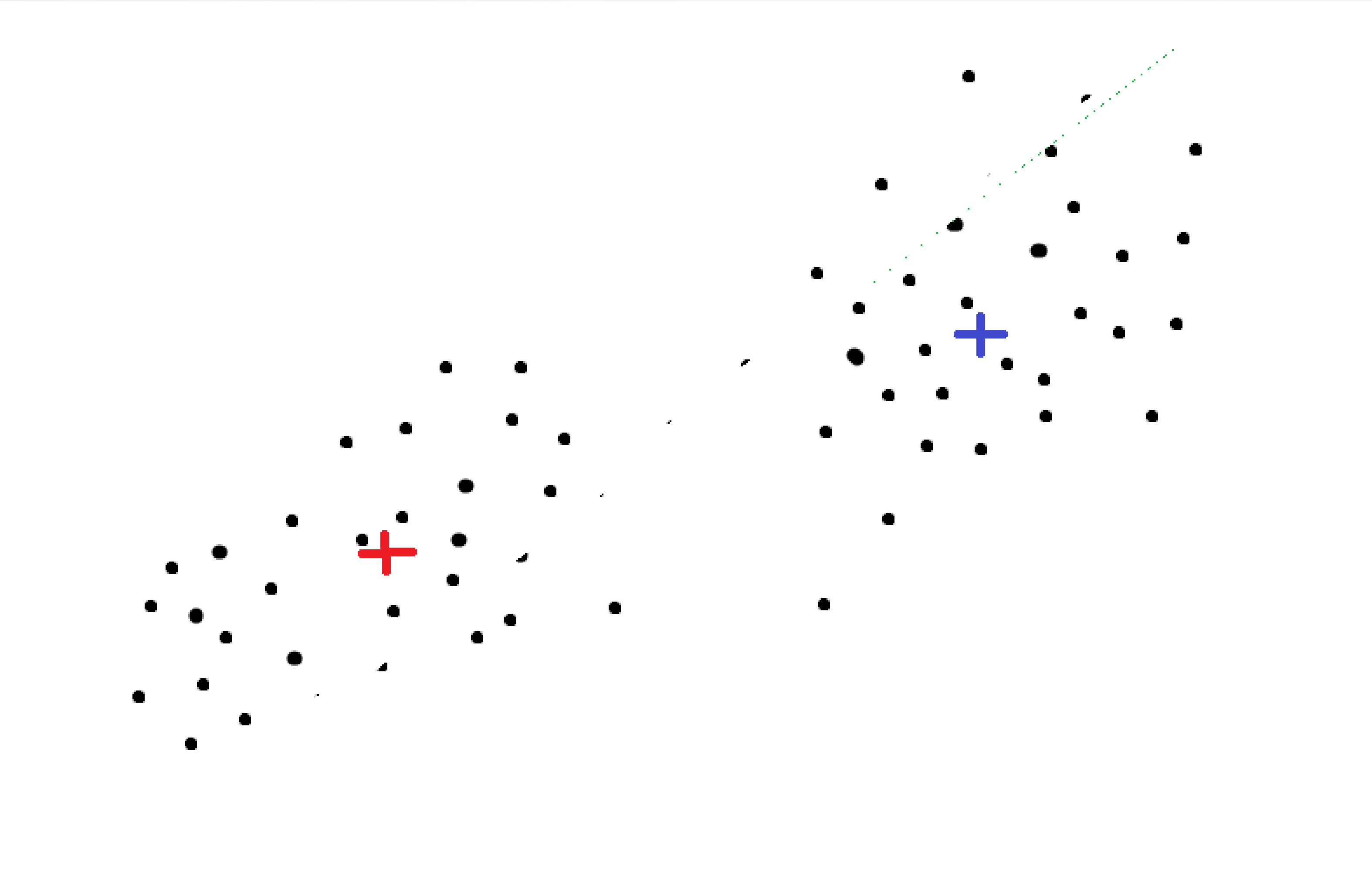
K-Means example



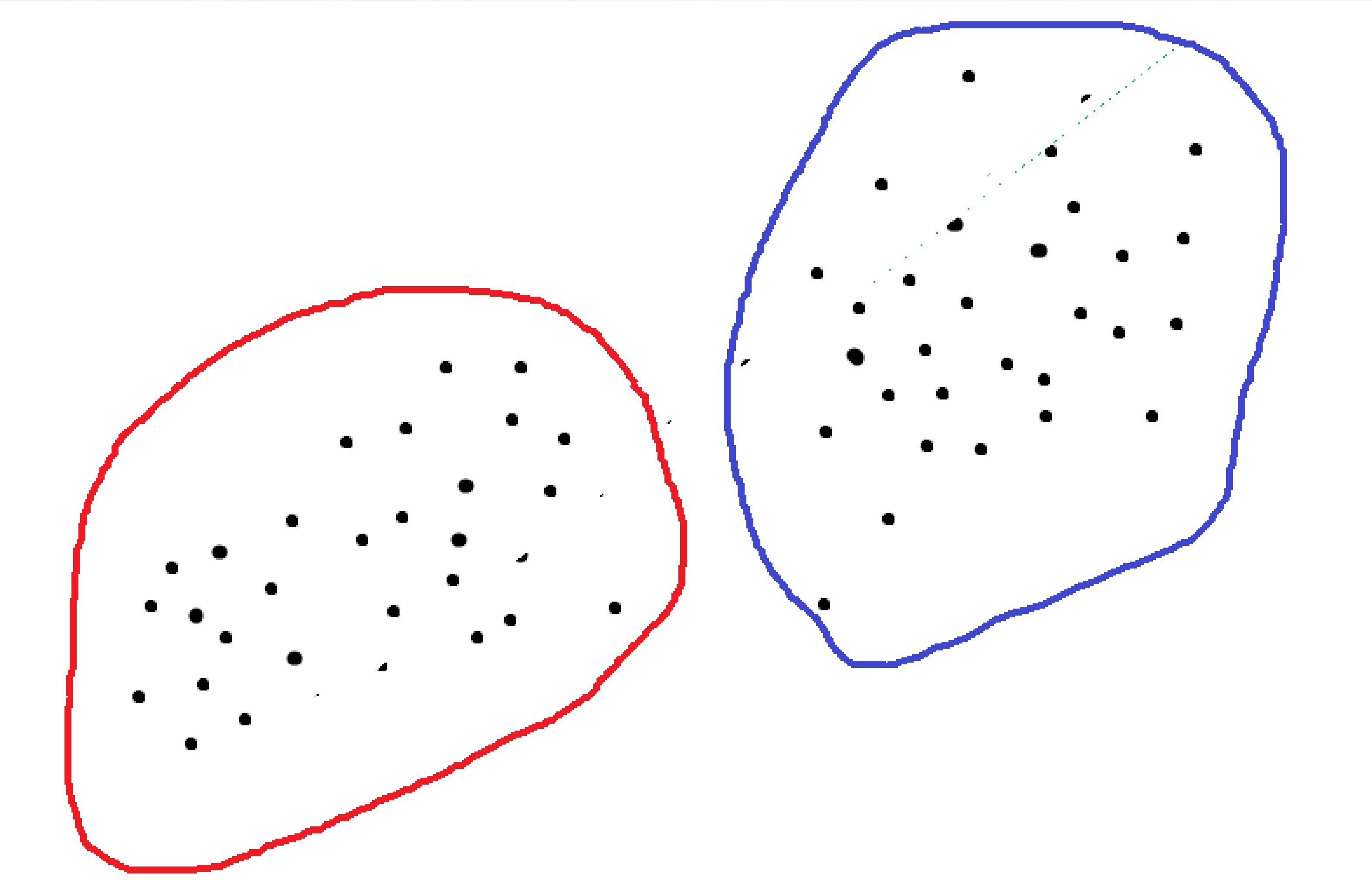
K-Means example



K-Means example



K-Means example



Properties of k-Means algorithm

- Will converge after a finite number of iterations

- Running time per iteration
  - Assigning all points with labels ~kN
  - Recalculate the positions of cluster centres ~N

- Euclidean space properties
  - Distance(A→B)=Distance(B→A)
  - Distances can be only positive
  - Entries at the same location should have the same label (belong to the same cluster)
  - “Triangle inequality”, i.e. Distance(A→B)+Distance(B→C)≥Distance(C→A)

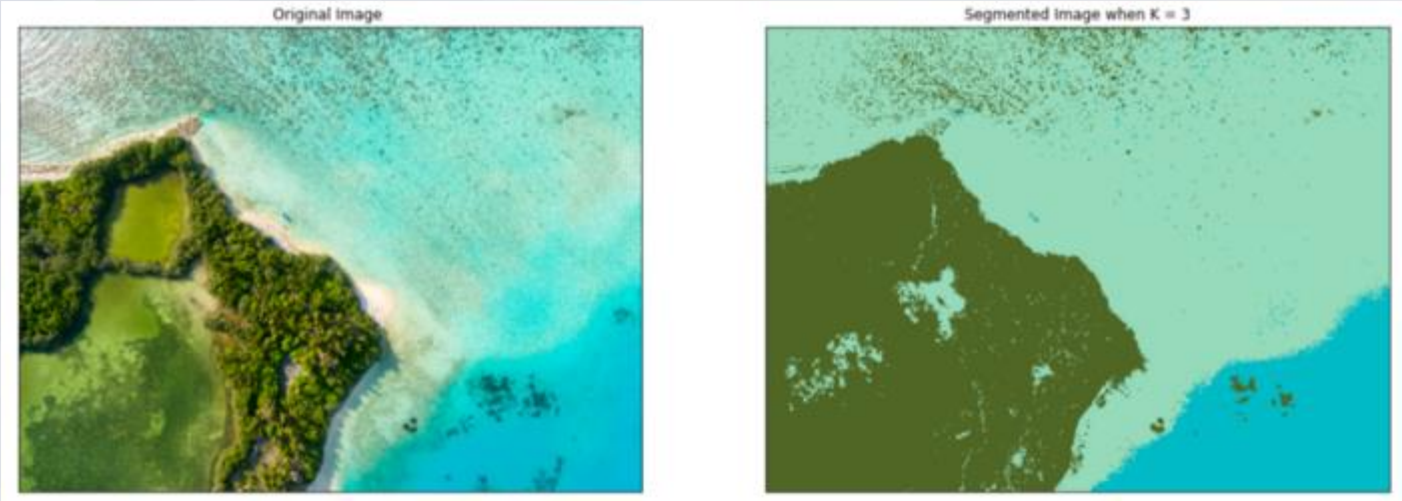
k-Means examples

Converting an image from 8 to 2 bit per pixel

R. A. Fisher (1890 − 1962), one of the parents of modern statistics

(from Hastie et al. 2009)

k-Means examples

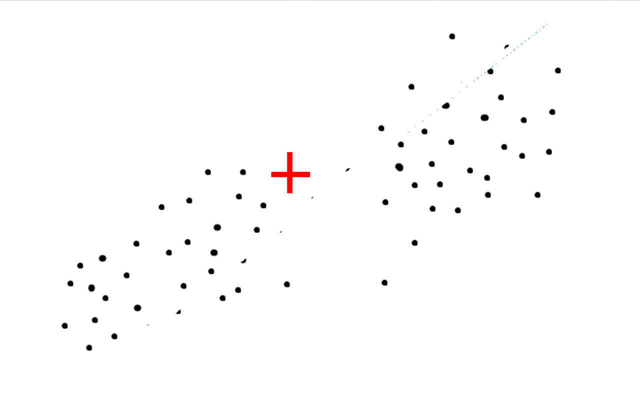


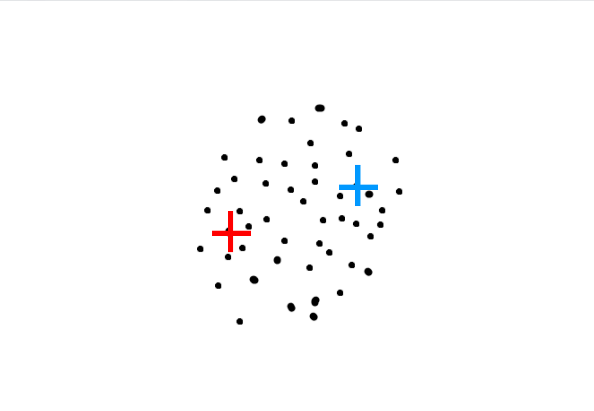
(from medium.com and kdnuggets.com)

k-Means algorithm

- It is a heuristic algorithm, and, hence, your input matters
  - how many clusters?
  - what variables to chose?

- What can go wrong here?

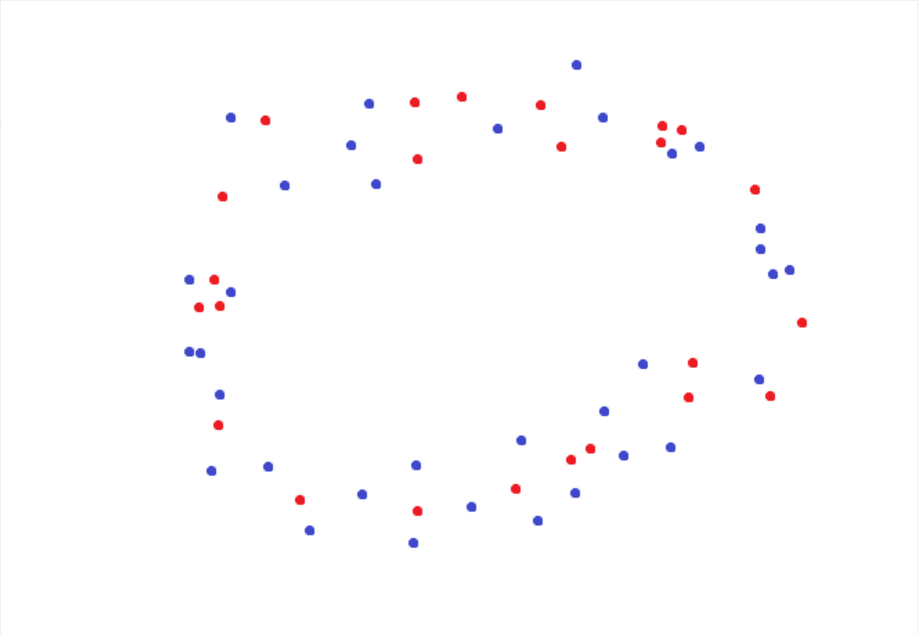




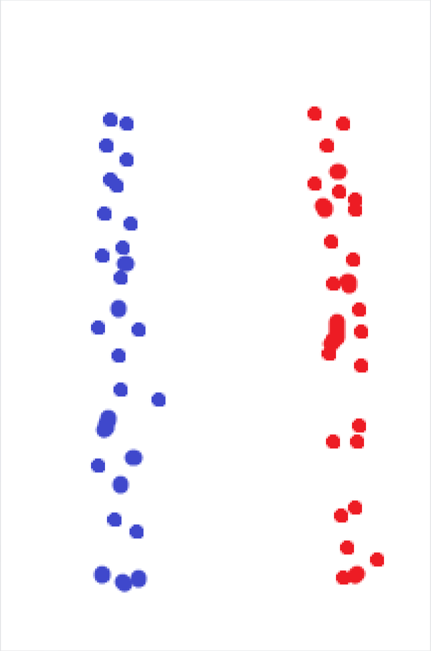
k-Means algorithm

- What can go wrong here?

X-Y plane

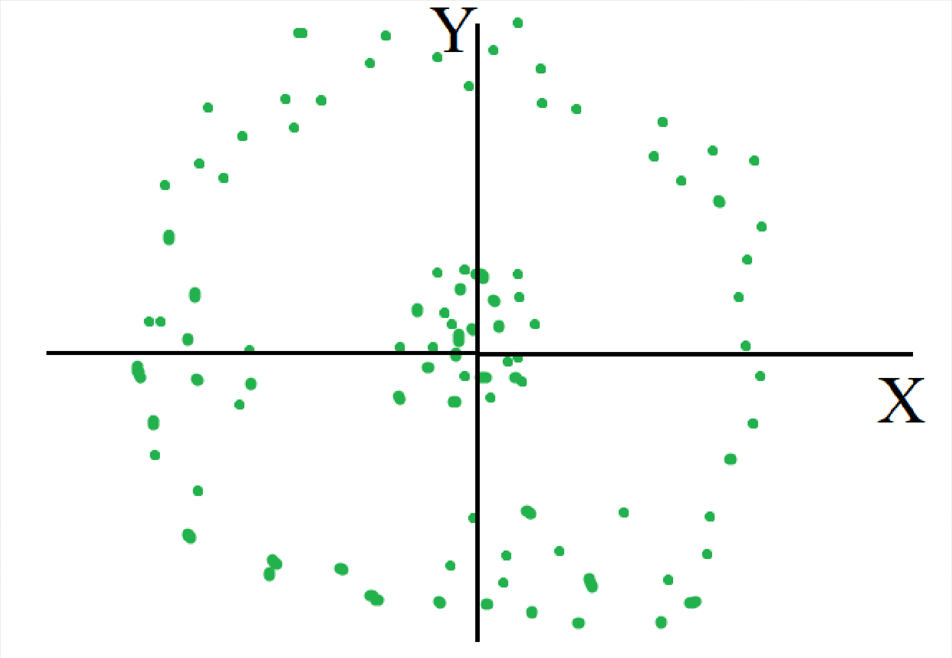


X-C plane



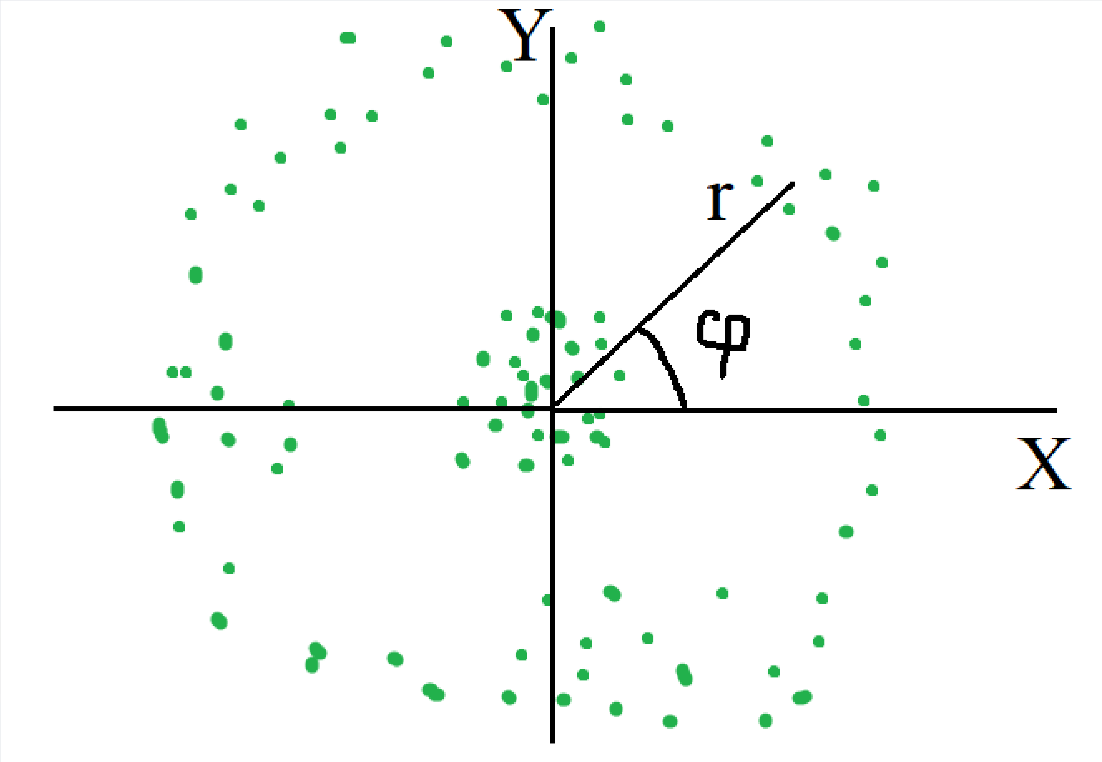
k-Means algorithm

- What can go wrong here?



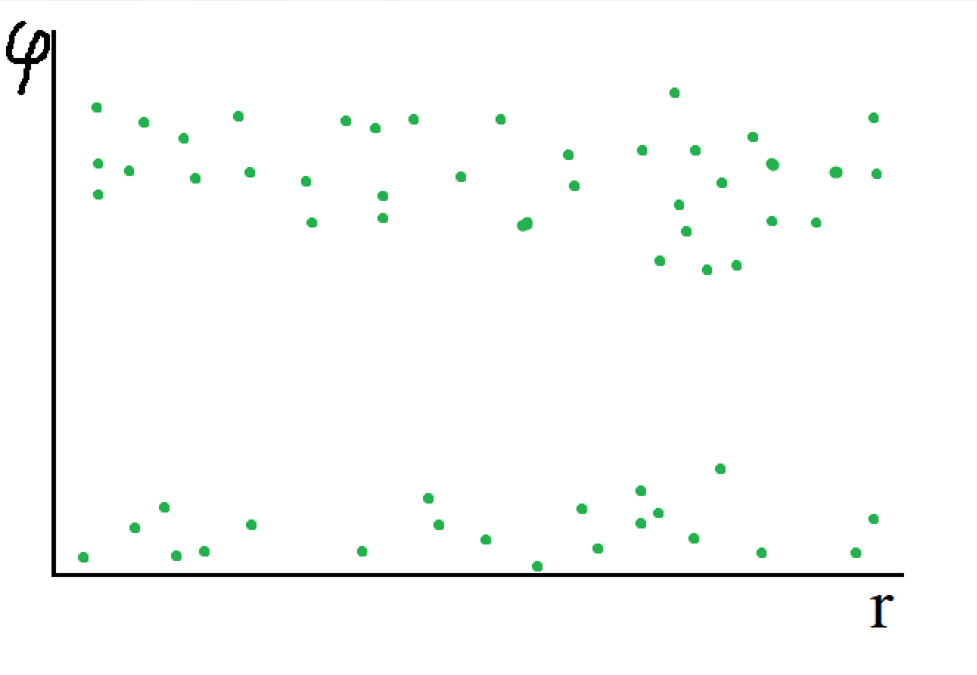
k-Means algorithm

- What can go wrong here?



k-Means algorithm

- What can go wrong here?



k-Means – Agglomerative clustering

- Start with clustering very similar entries
- Then create higher level clusters

- Algorithm

  Initiate:

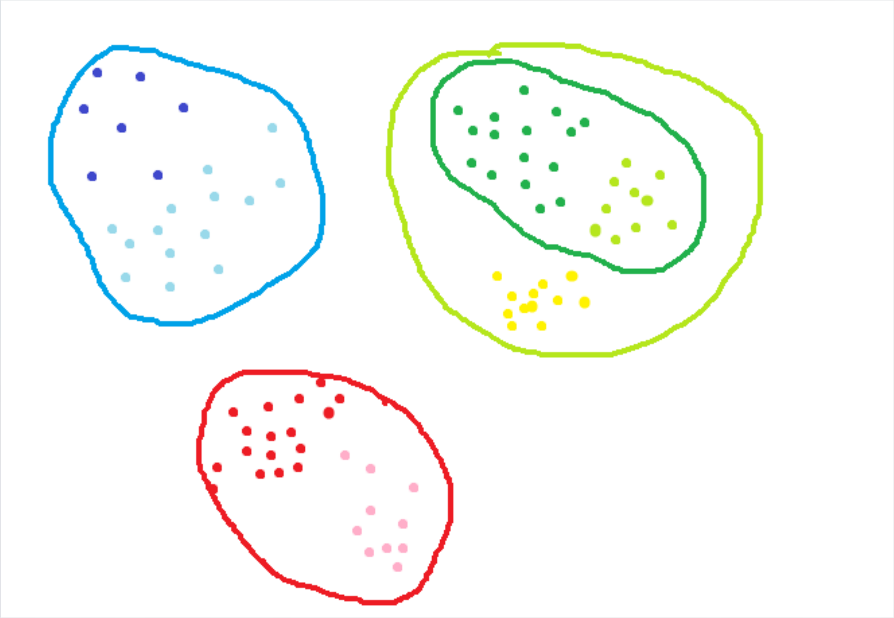
  - each entry is a cluster

  Iterations:

  - Take two closest clusters and merge them
  - Repeat until stop

  Stop:

  - When only one cluster left



- Produces not one clustering, but a family of clusterings represented by a dendrogram