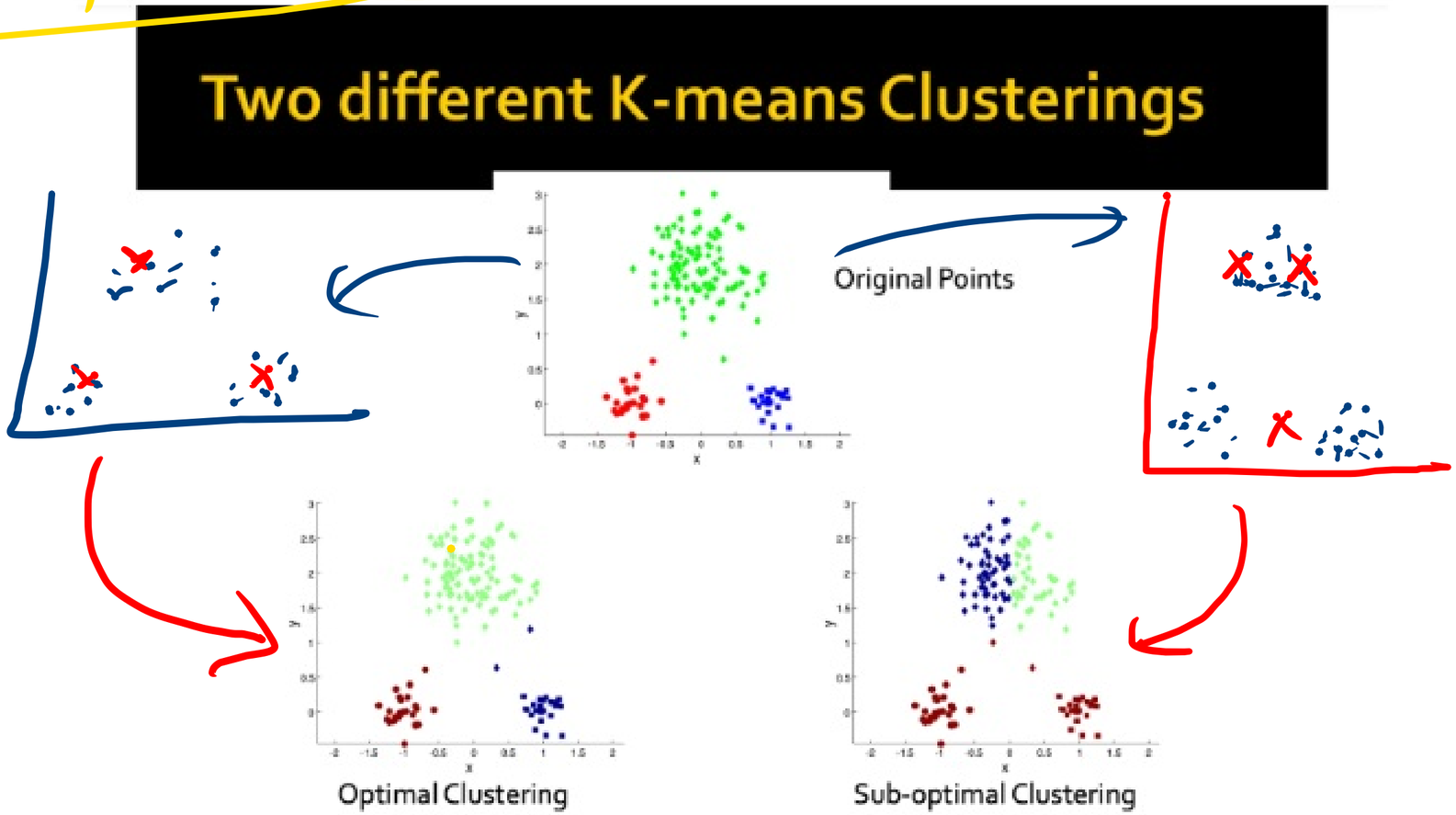
Disadvantage of K-means



→ Initialization of cluster

① Run Kmeans multiple times with random initialization of centroids

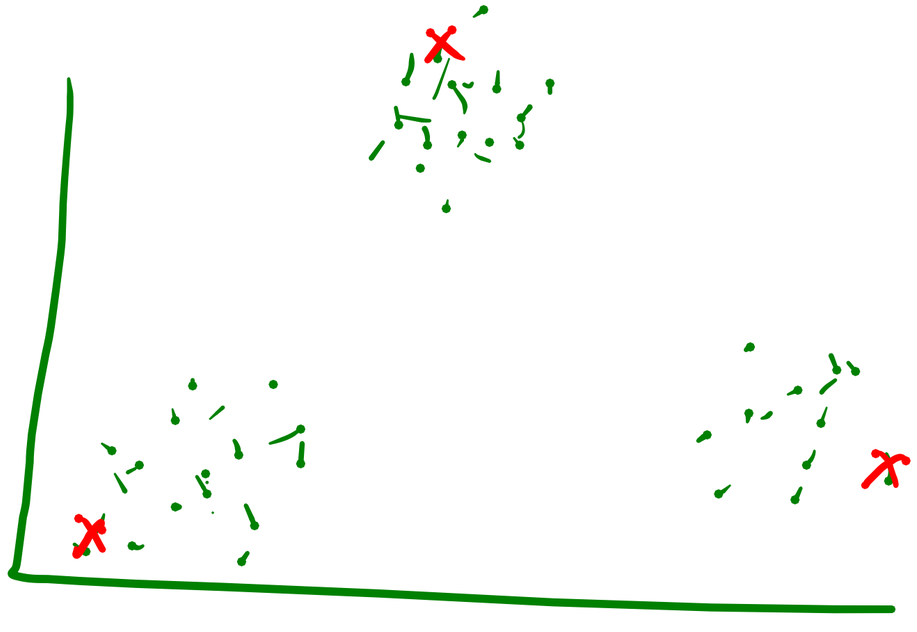
② Kmeans ++

Initialization ↳ Select 1st centroid from data point (randomly)

↳ Select 2nd centroid which is farthest from 1st one

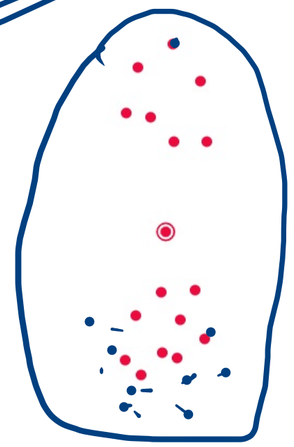
↳ Select 3rd centroid which is farthest from both 1st & 2nd centroid

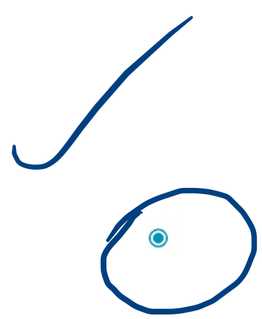
⟶ Run Kmeans



→ Outlier

Case 1



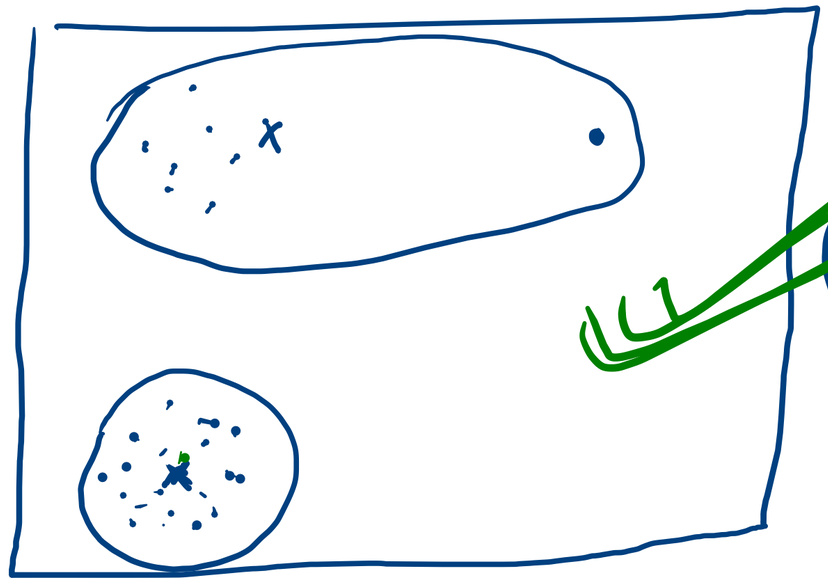


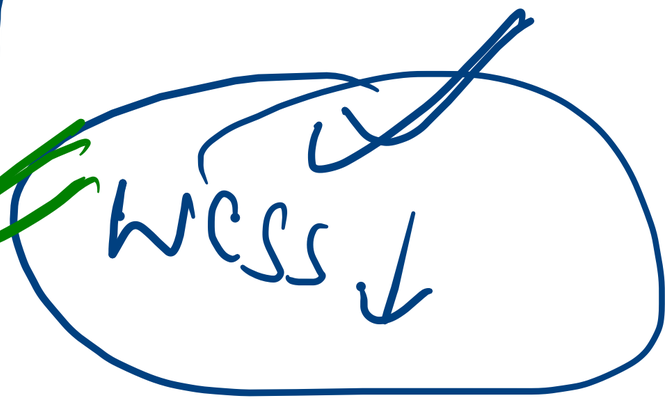
WCSS ↑

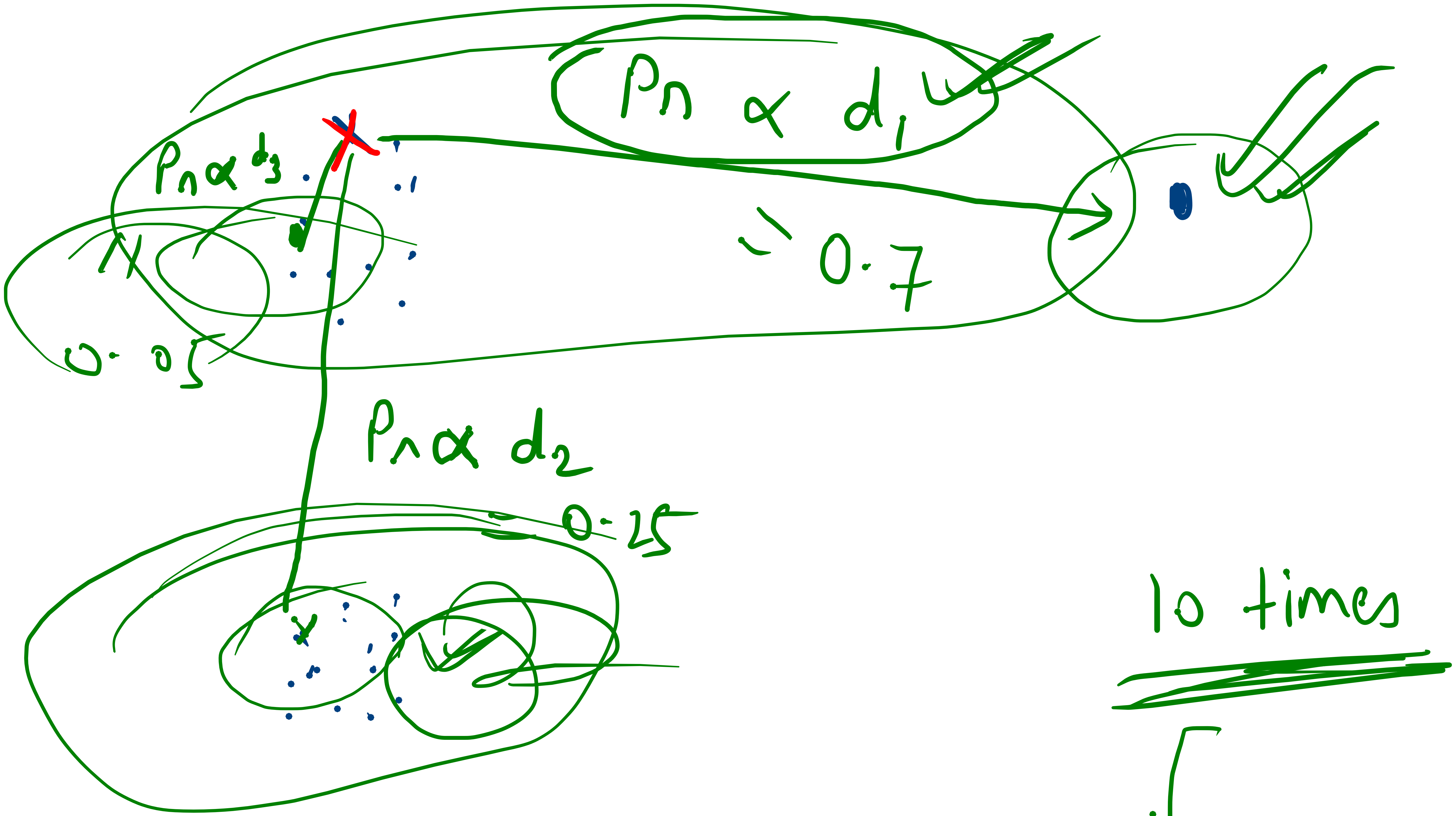
→ lowest WCSS

↳ Best cluster

Case 2

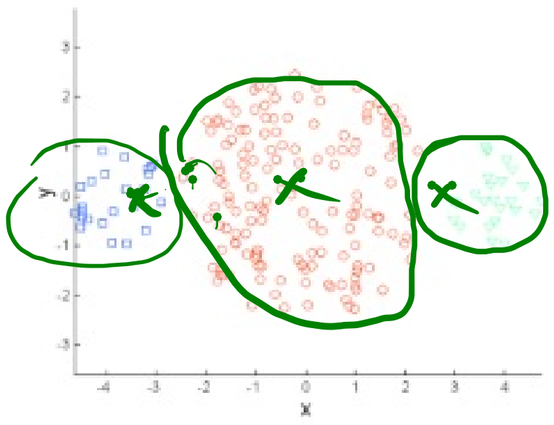




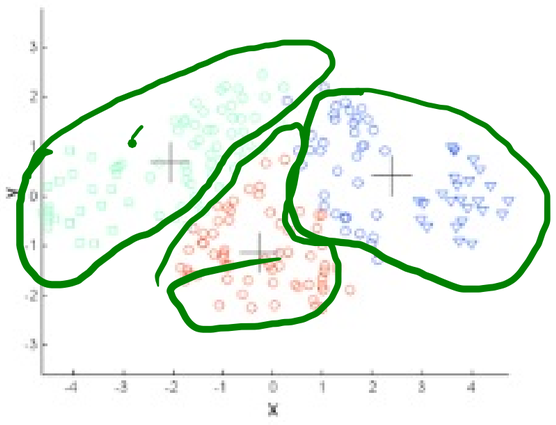


10 times

Limitations of K-means: Differing Sizes

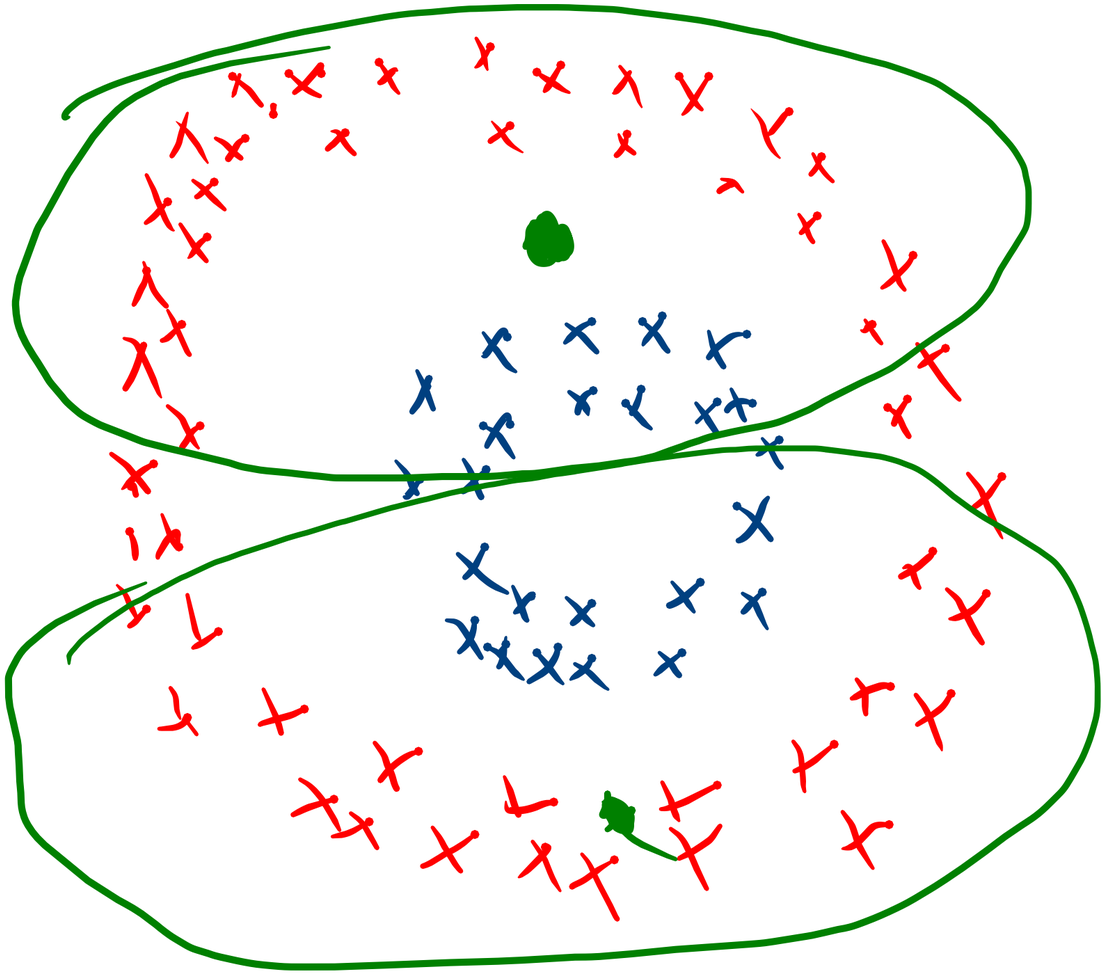


Original Points

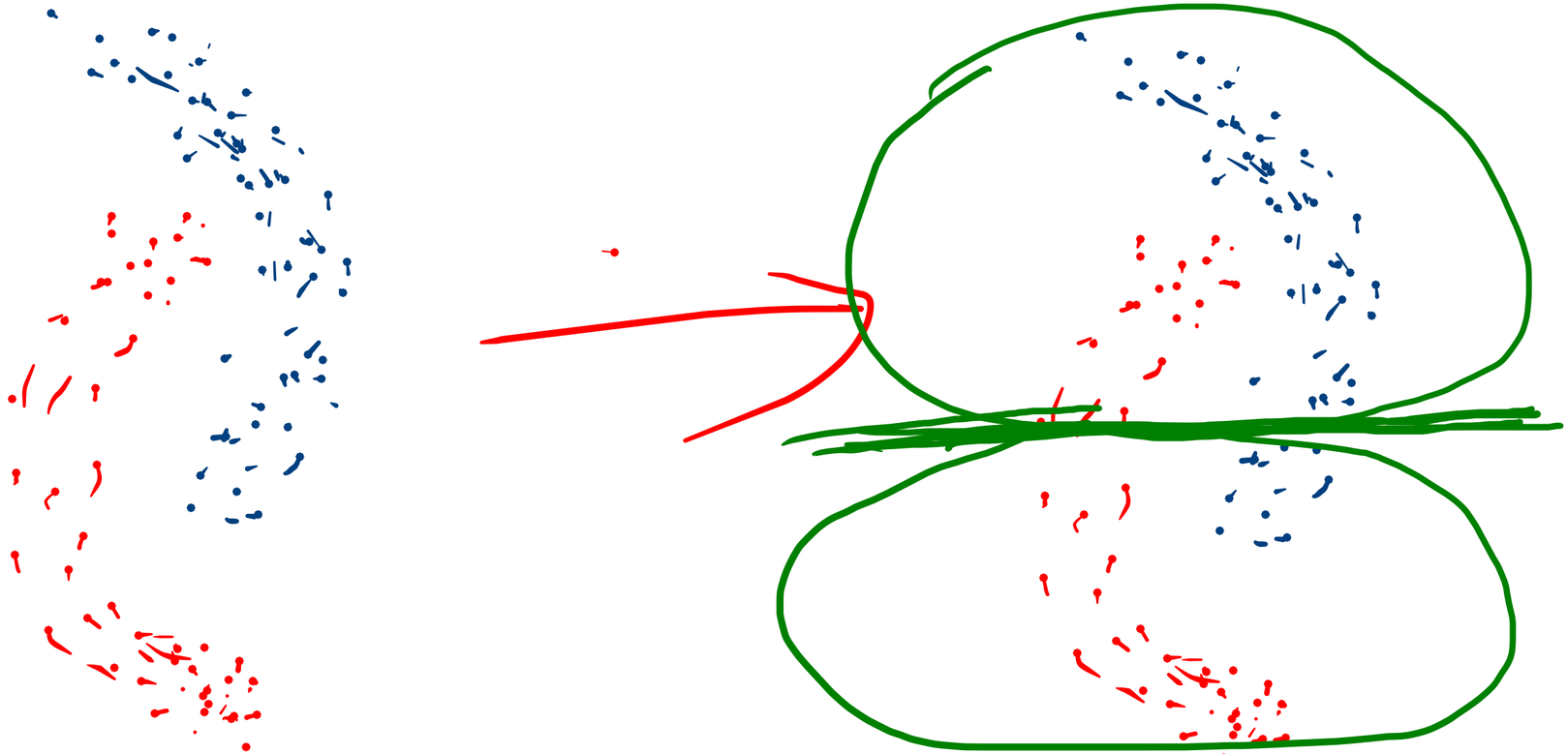


K-means (3 Clusters)

Works
Clusters
and
Similar
Size



K-means always form spherical clusters



linear decision boundary

K median → less sensitive to outlier

→ Intialize k cluster centre

→ Repeat

→ Assignment

→ Update → $\frac{1}{n_K} \sum_{i \in C_i} x_i$

Median of point

K-means

→ Number of clusters need to be decided initially

Does the centroid in K-means / K-median need to be a data point ??

Always guaranteed that medoid is a datapoint

K-medoids (less sensitive to outlies)

- Select K points as medoids
- Repeat
  - Assign each point to cluster which is closest and calculate total sum of distance (dissimilarity) ← dold
  - Swap a medoid with non-medoid and calculate total sum of distance (dissimilarity) ← dnew
  - S = dnew − dold
  - S < 0 → Select this point as better medoid

Eg Datapoints → Manhattan

(2,6), (3,4), (3,8), (4,7) (6,2) (6,4)

(7,3) (7,4) (8,5) (7,6)

	Dist from (3,4)	Dist from (7,4)
(2,6)	|3-2| + |4-6| = 3	|7-2| + |4-6| = 7
(3,8)	7	8
(4,7)	4	6
(6,2)	5	3
(6,4)	3	1
(7,3)	5	1
(8,5)	6	2
(7,6	6	2

old dist

Dissimilanity = 3 + 4 + 4 + 3 + 1 + 1 + 2 + 2

= 20

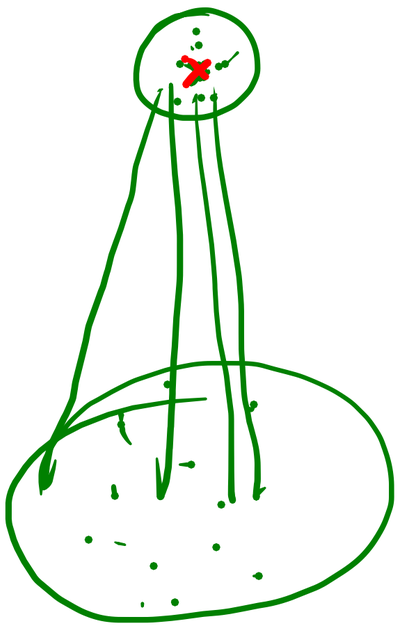
	Distance from (3,4)	Distance from (7,3)
(2,6)	|3-2|+|4-6| = 3	|2-7| + |6-3| = 8
(3,8)	4	9
(4,7)	4	7
(6,2)	5	2
(6,4)	3	2
(7,4)	4	1
(8,5)	6	3
(7,6)	6	3

New dist = Dissimilarity (total dist) = 3+4+4+2+2+1+3+3

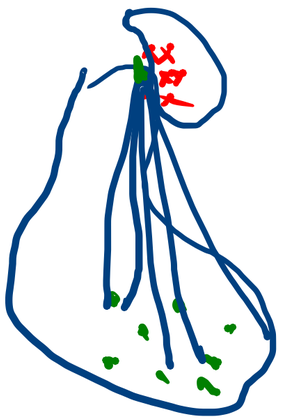
= 22

S = New - Old = 22 - 20 = 2

$$S.H = \frac{b-a}{\max(b,a)} = \frac{b-0}{b} = 1$$



$b \gg a$



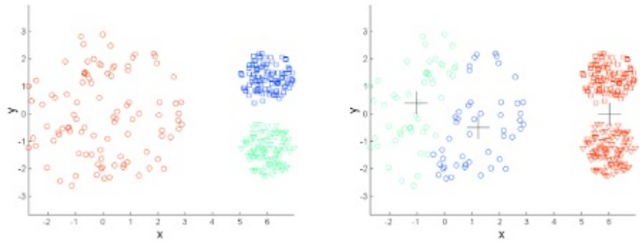
$a \gg b$

$$\frac{0-a}{\max(0,a)}$$

$$= \frac{-a}{a}$$

$$= -1$$

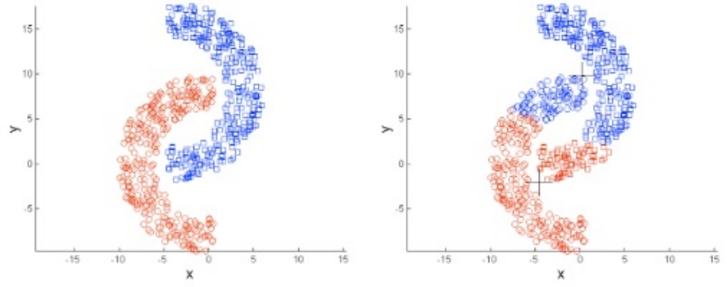
Limitations of K-means: Differing Density



Original Points

K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

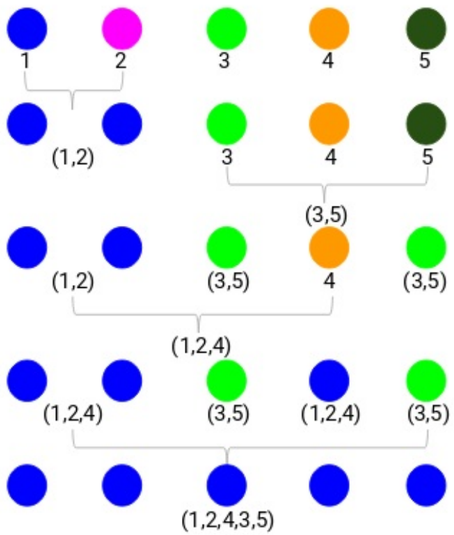


Student_ID	Marks
1	10
2	7
3	28
4	20
5	35

ID	1	2	3	4	5
1	0	3	18	10	25
2	3	0	21	13	28
3	18	21	0	8	7
4	10	13	8	0	15
5	25	28	7	15	0

Student_ID	Marks
(1,2)	10
3	28
4	20
5	35

ID	(1,2)	3	4	5
(1,2)	0	18	10	25
3	18	0	8	7
4	10	8	0	15
5	25	7	15	0


1
2
3
4
5
(1,2)
3
4
5
(3,5)
(1,2)
(3,5)
4
(3,5)
(1,2,4)
(1,2,4)
(3,5)
(1,2,4)
(3,5)
(1,2,4,3,5)