APPLIED MACHINE LEARNING

Clustering

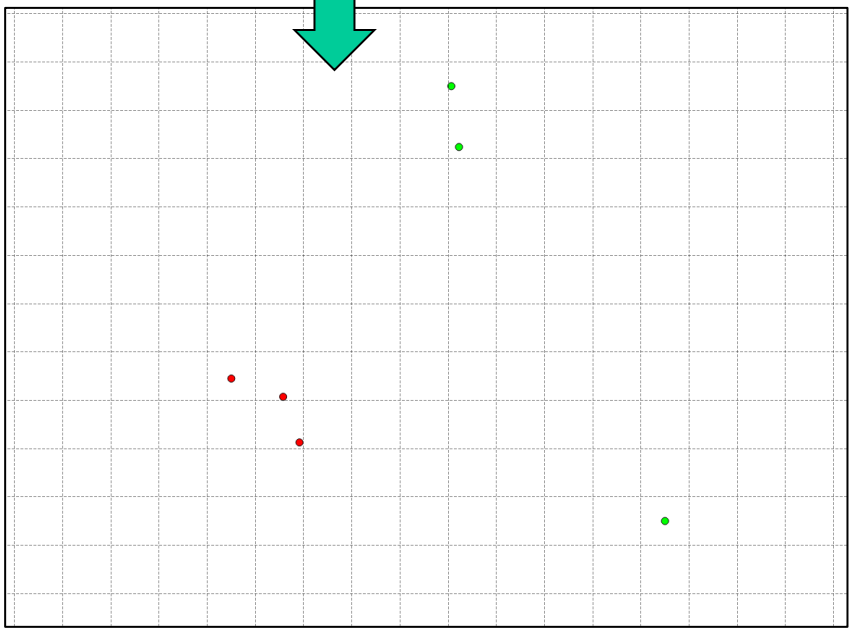
Interactive Lecture

Clustering Principle

Start with a set of datapoints

Algorithm does not know the true labels

It knows neither the number of groups nor what regroup datapoints



After PCA projection,
Groups are easier to tell apart

Clustering methods will automatically find how to regroup points.

Clustering Principle

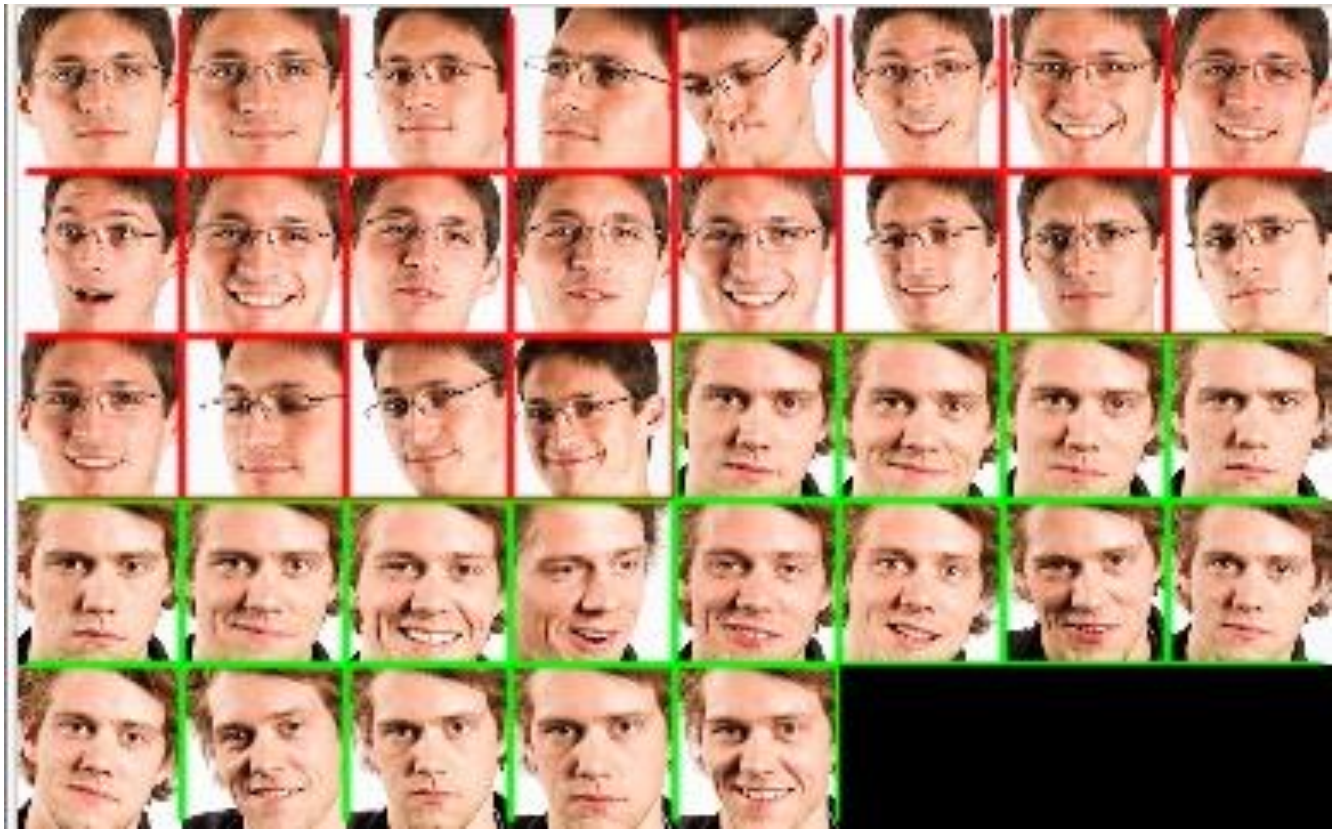
Clustering can be used as:

- *Feature extraction method*: for identifying underlying structure in data and salient features, best visualized through **cluster prototype**.
- *Compression method*: for organizing the data and summarizing it through **cluster prototypes.**

A cluster prototype can be:

- A typical datapoint, best representative of the cluster
- The average (centroid) of the datapoints in the cluster

Clustering, features, metrics



Which subgroups of pictures are similar and why?

High intra-class similarity is necessary for achieving a good clustering

Clustering Principle

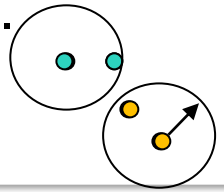
Groups of points are said to belong to the same cluster
if they are similar enough.

→ measure of similarity.

K-Means and soft-K-means minimize a measure of distance of all datapoints attached to the cluster to its centroid, using norm-p.

$$\textbf{\textit{Measure of distance}}: d(x^i, \mu^k) = \sqrt[p]{\sum_{i=1}^{N} |(x_i^i - \mu_i)|^p}$$

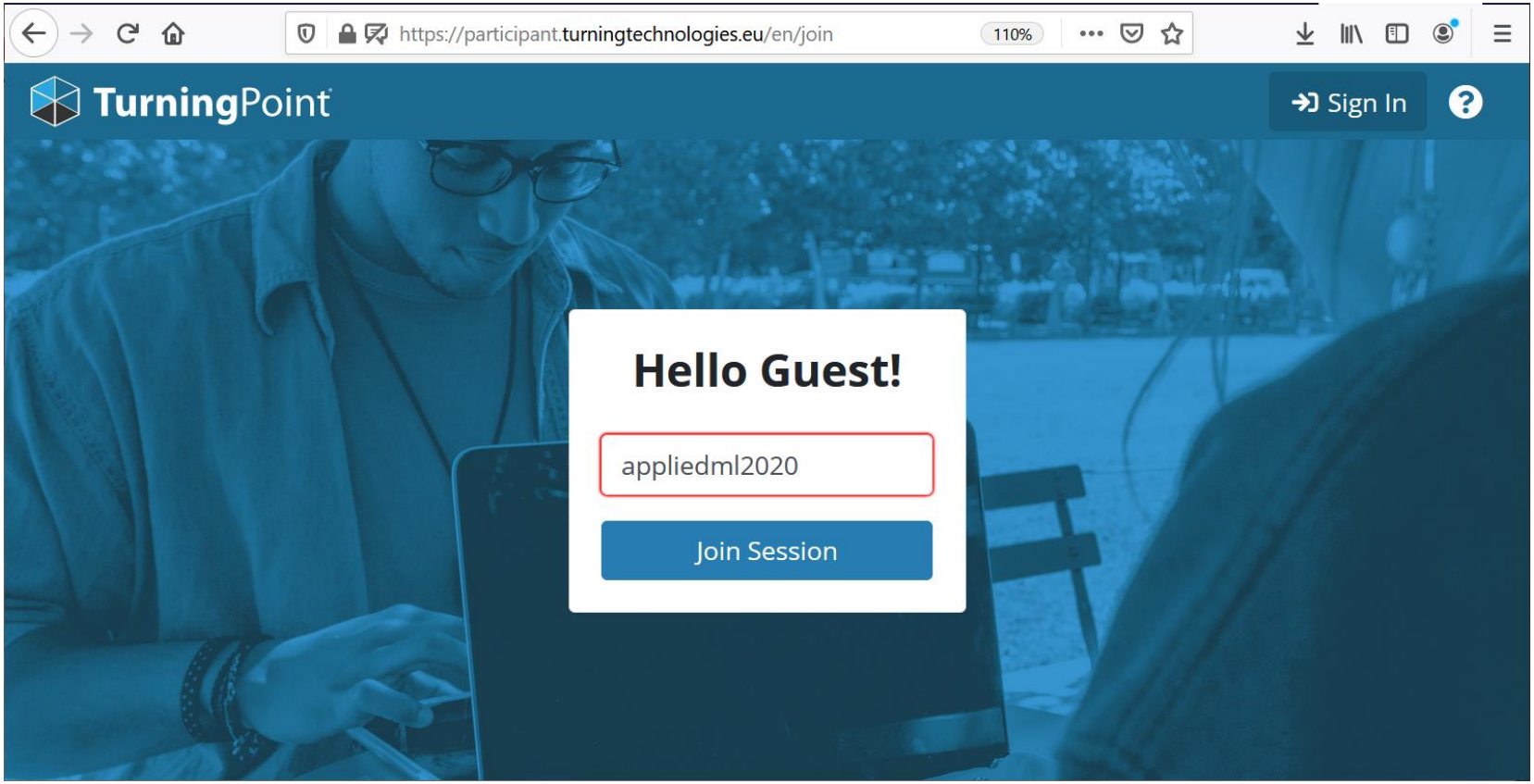
DBSCA uses a lower bound on norm-2 (size of ball) and on number of datapoints to decide on cluster assignment.

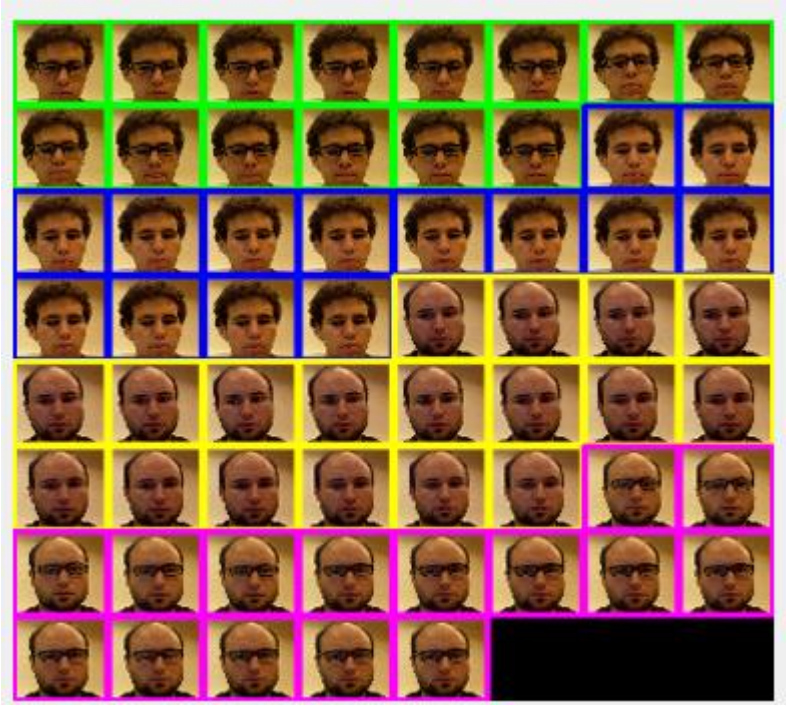


Launch polling system

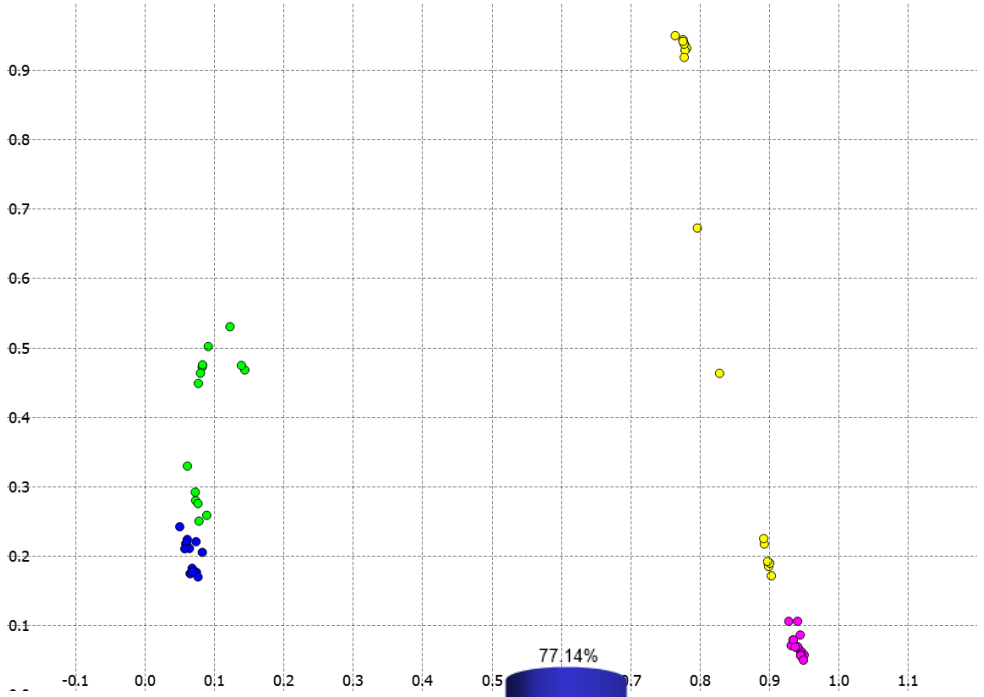
https://participant.turningtechnologies.eu/en/join

Acces as GUEST and enter the session id: *appliedml2020*





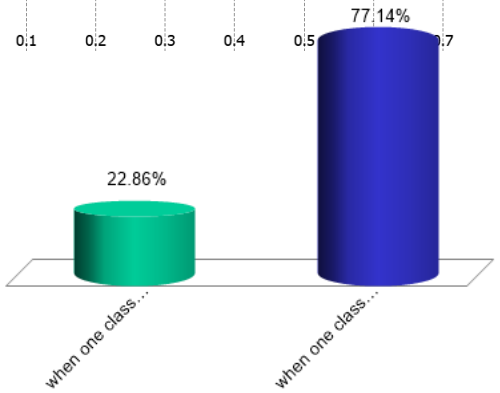
- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses

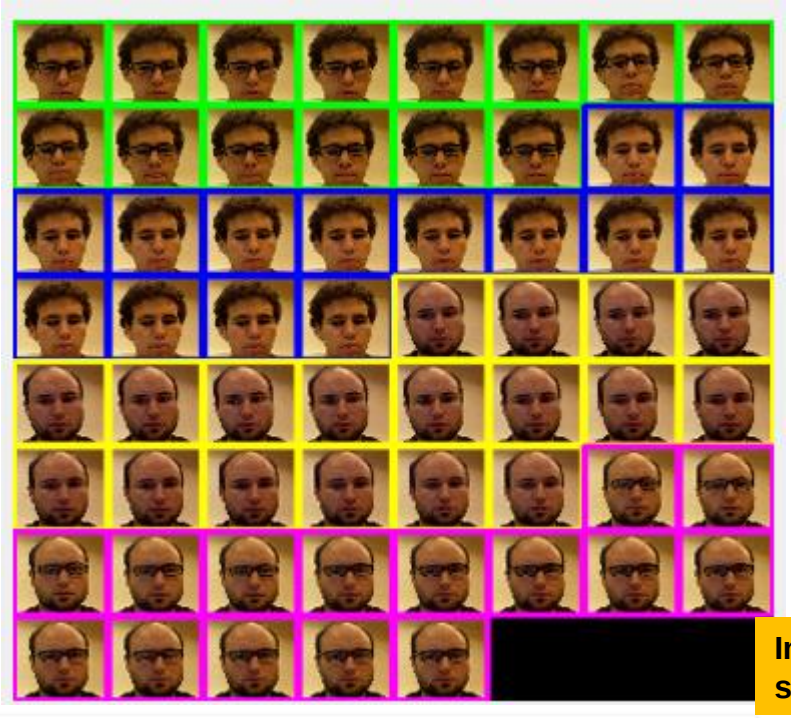


When is intra-class similarity the highest ?

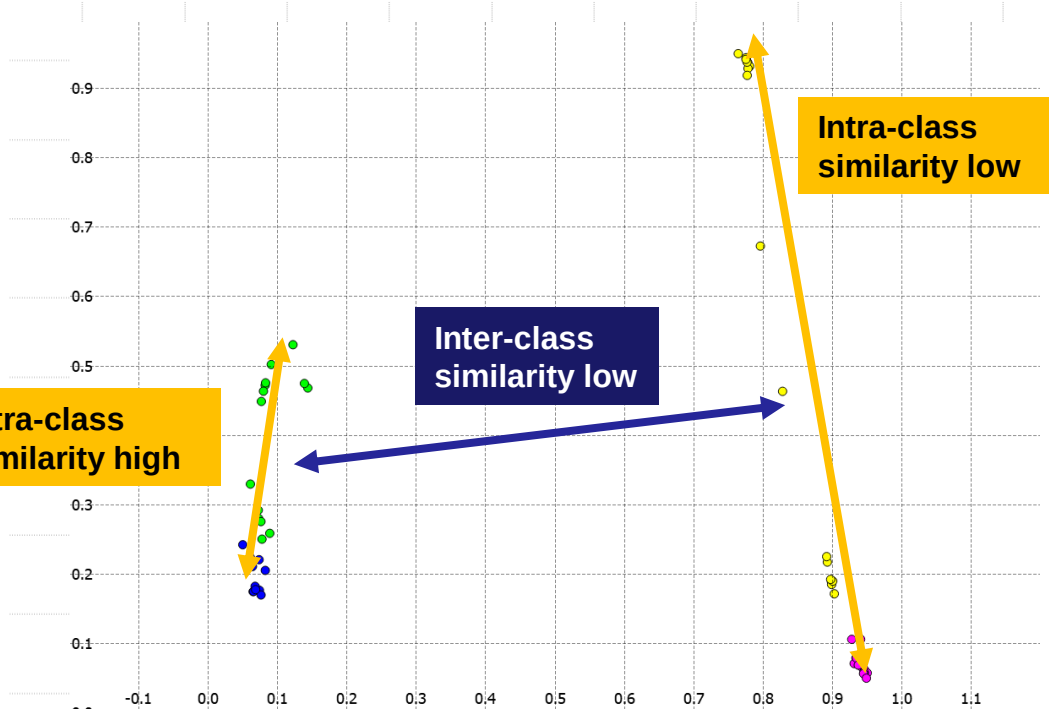
A. when one classifies images of faces with and without glasses;

B. when one classifies images of person1 against person2.





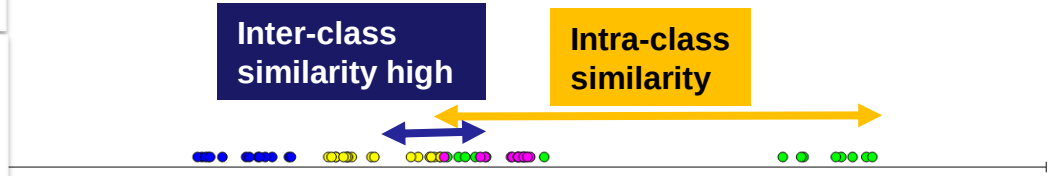
- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses



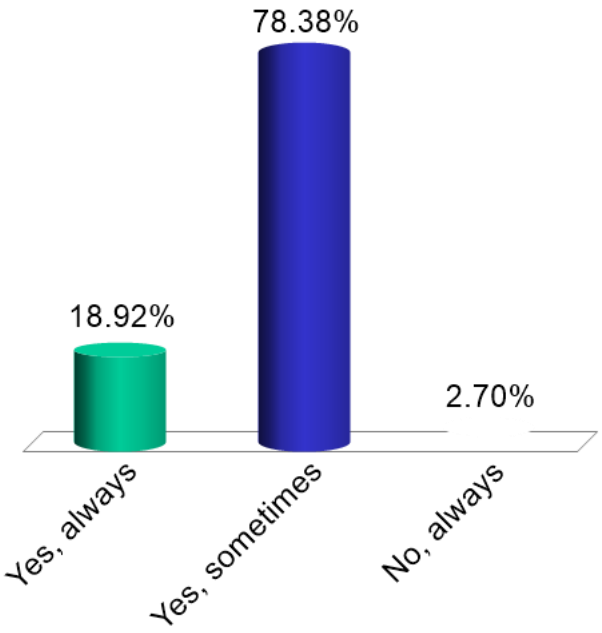
Projection onto first two principal components after PCA

Inter-class similarity is lower than intra-class similarity when one classifies images of person1 against person2, for one of the 2 persons.

Intra-class similarity is low when classifying persons with glasses vs persons without glasses, especially for person2.



Projection along e^3



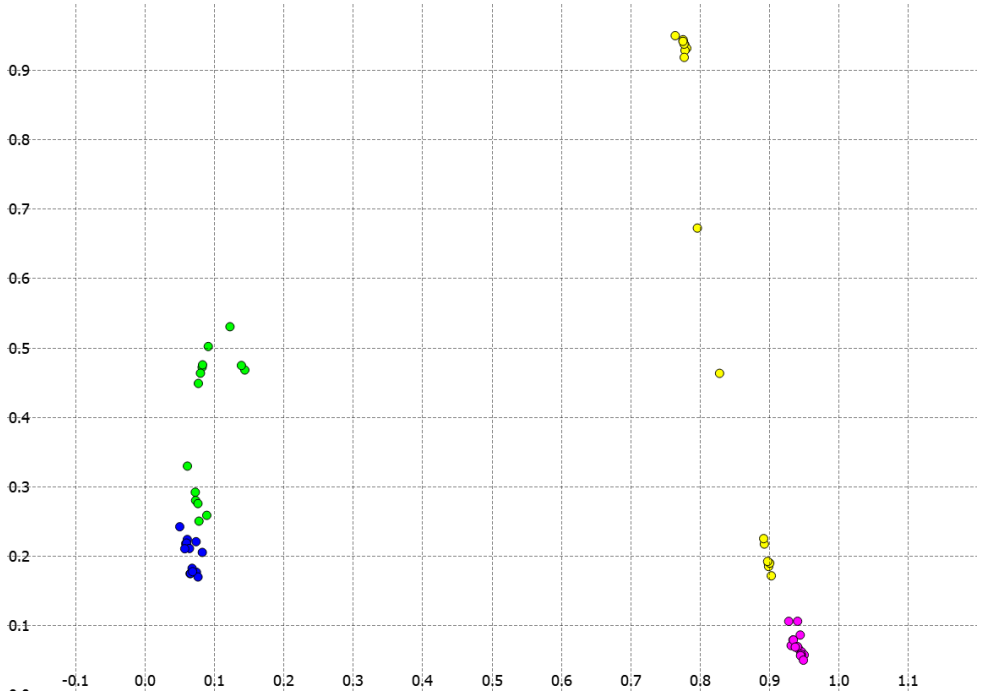
Would K-means (with K=2 and norm-2) be able to separate the two persons correctly?

A. Yes, always

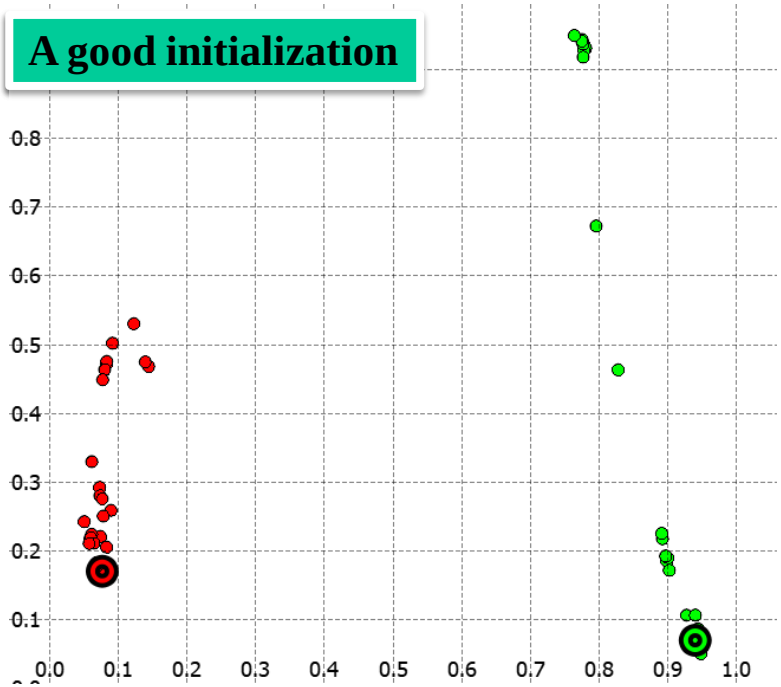
B. Yes, sometimes

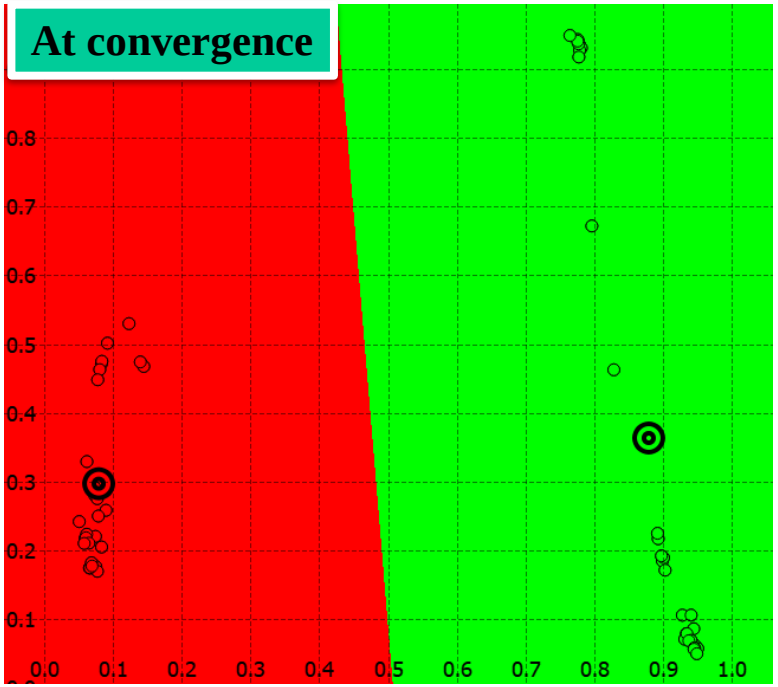
C. No, always

- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses



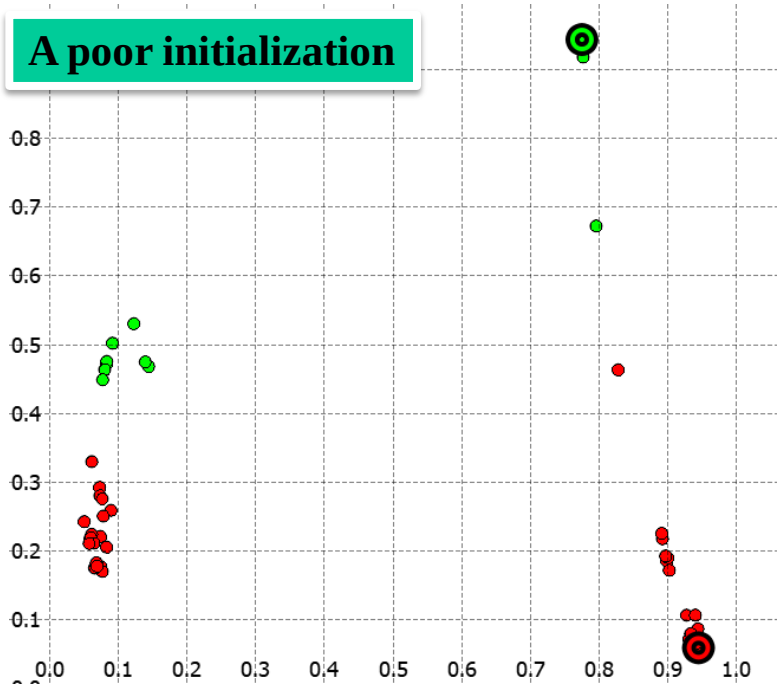
Projection onto first two principal components after PCA

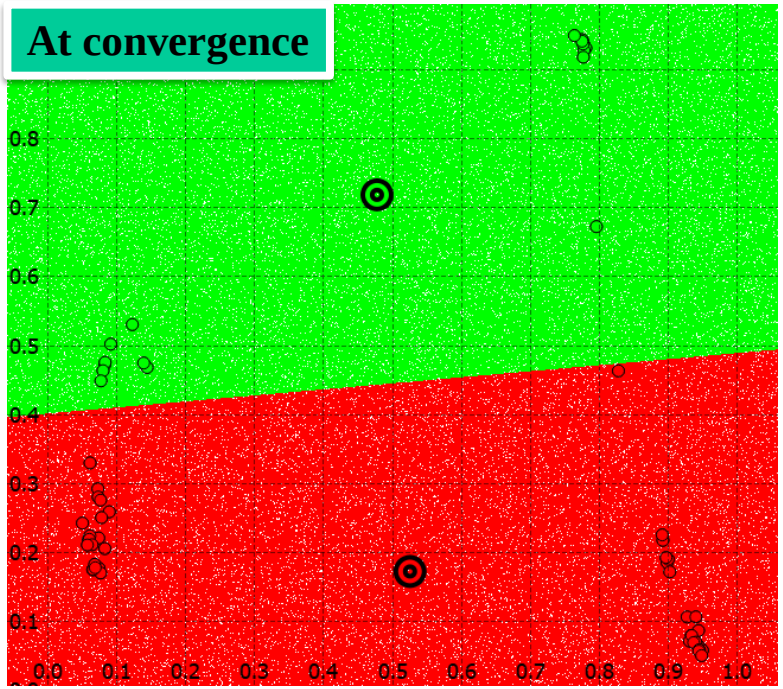




Would K-means (with K=2 and norm-2) be able to separate the two persons correctly?

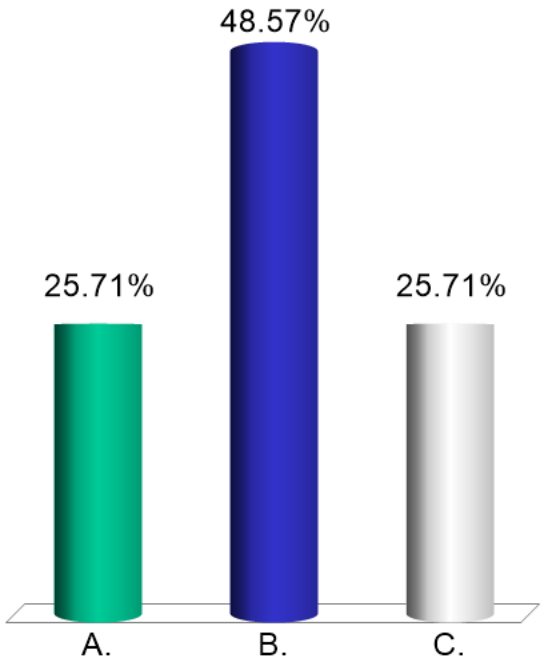
Yes, sometimes



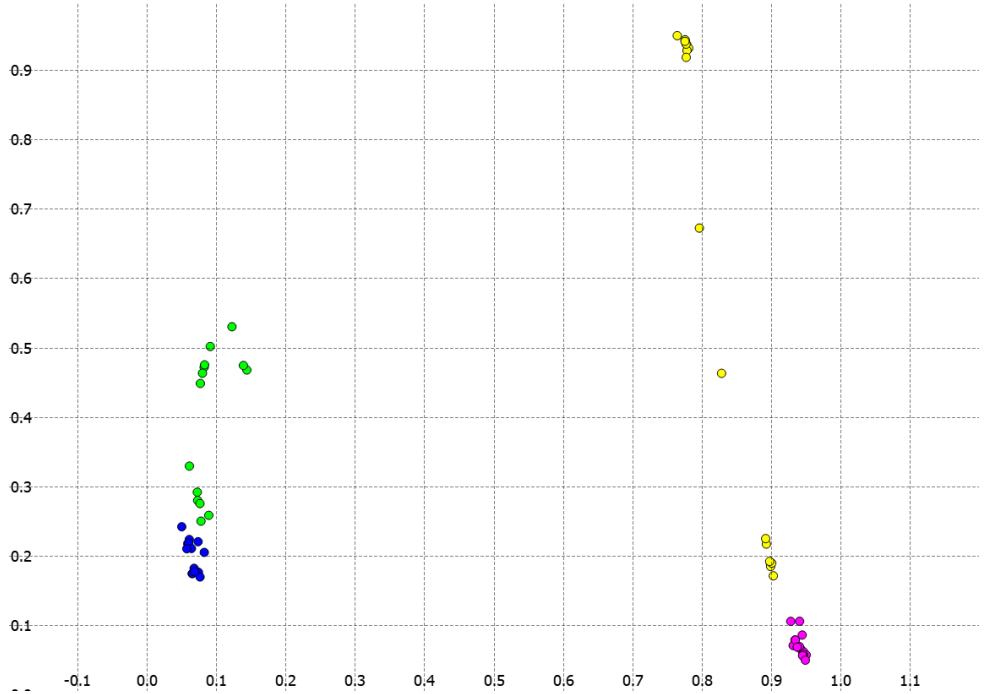


Would K-means (with K=2 and norm-2) be able to separate the two persons correctly?

Yes, sometimes

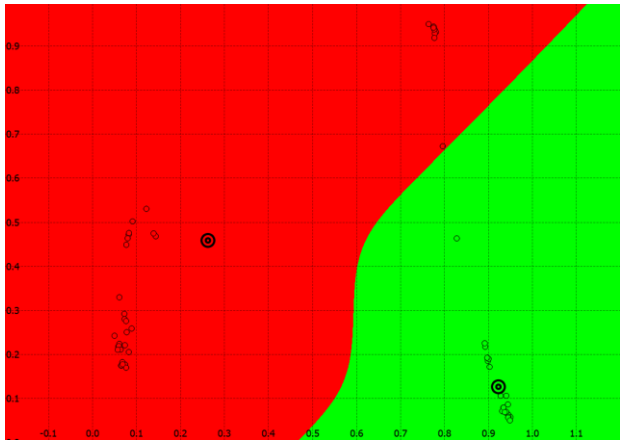


- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses

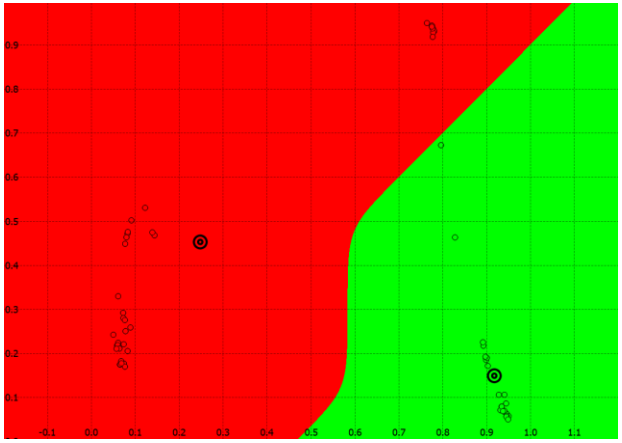


Would K-means (K=2) with other L-p norms be able to separate the two persons correctly always?

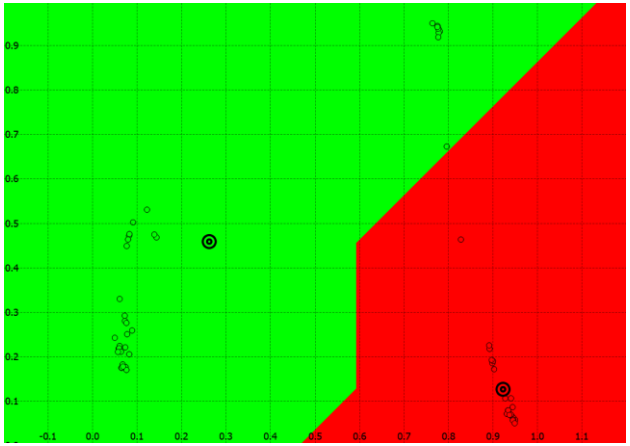
A. Yes
B. No
C. I do not know



P=6



P=9

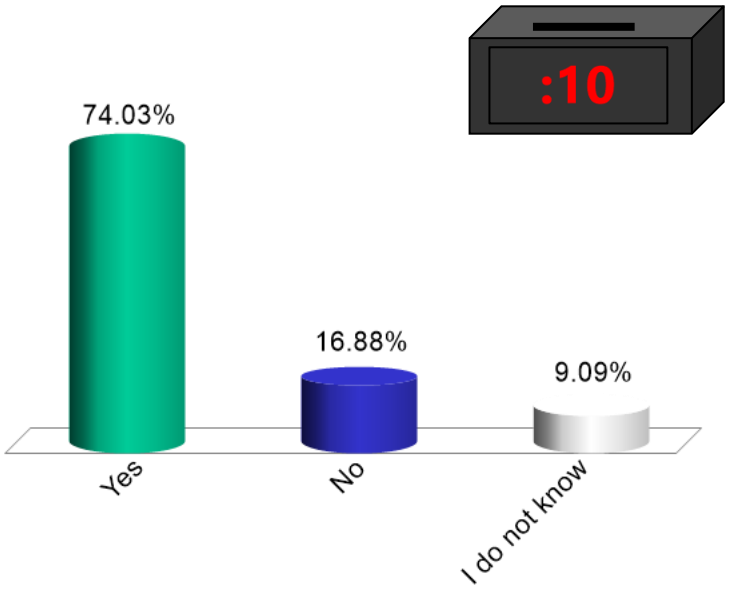


L-inf

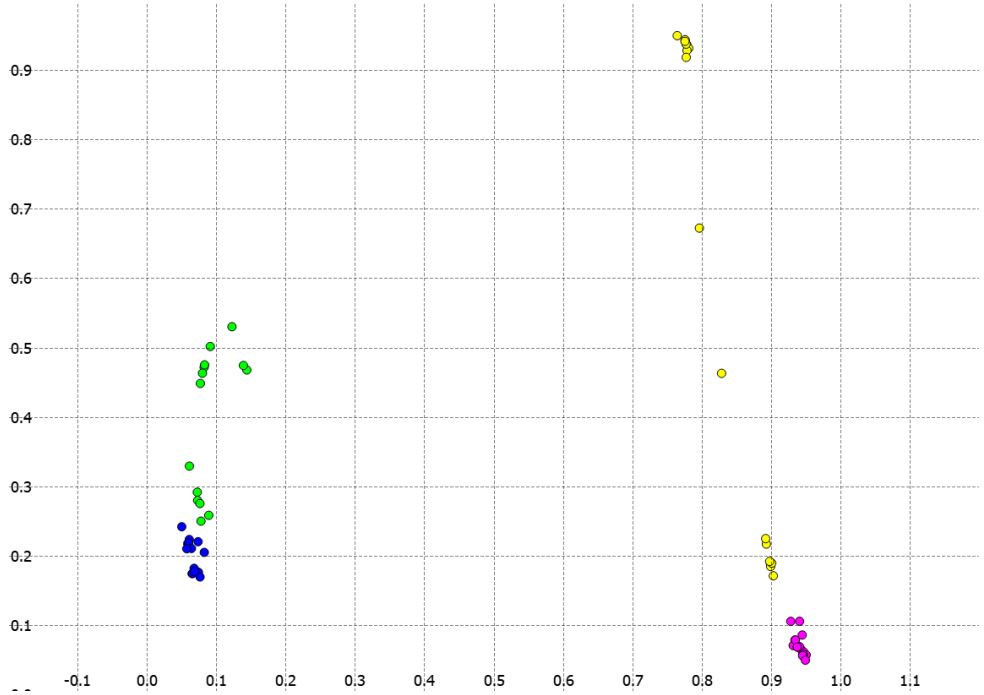
Would K-means (K=2) with other L-p norms be able to separate the two persons correctly always?

In general, NO. The decision boundary is determined by the positioning of the centroids, which are influenced by a) the ratio across intra-cluster distance / inter-cluster distance and b) their position at initialization.

The p of the norm changes the softness of the boundary.



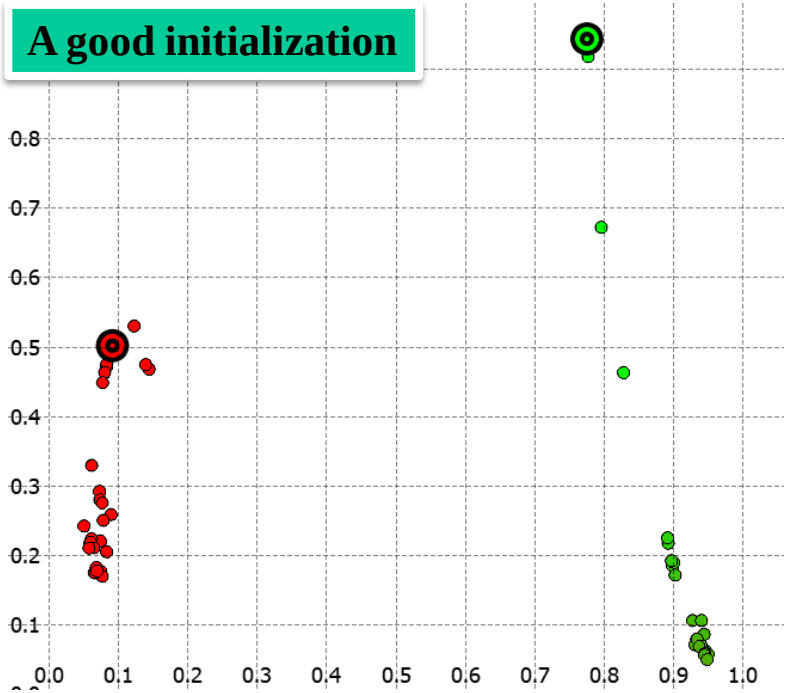
- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses

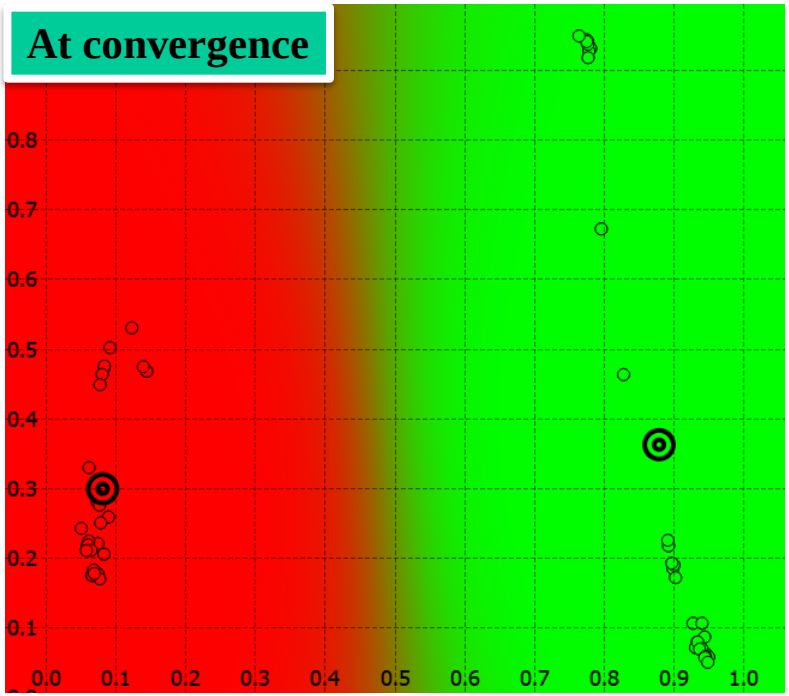


Would soft K-means be able to separate the two persons correctly with a large β for a good initialization ?

A. Yes
B. No
C. I do not know

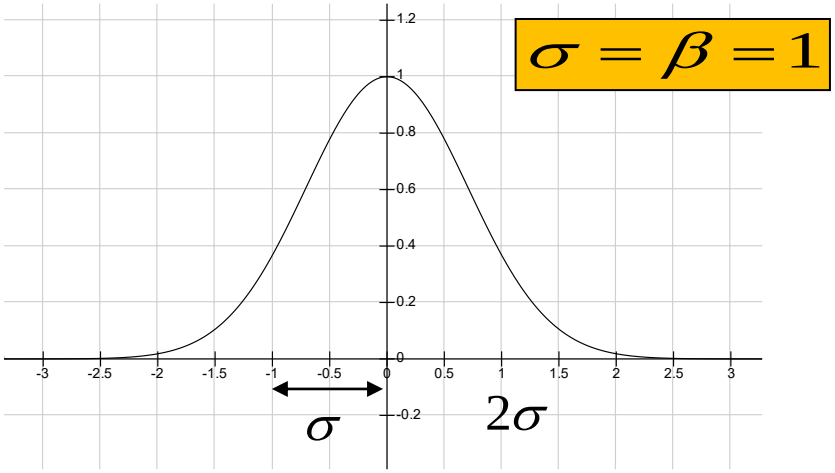
$$r_i^k = \frac{e^{\left(-\beta \cdot d\left(\mu^k, x^i\right)\right)}}{\sum_{k'} e^{\left(-\beta \cdot d\left(\mu^{k'}, x^i\right)\right)}}$$

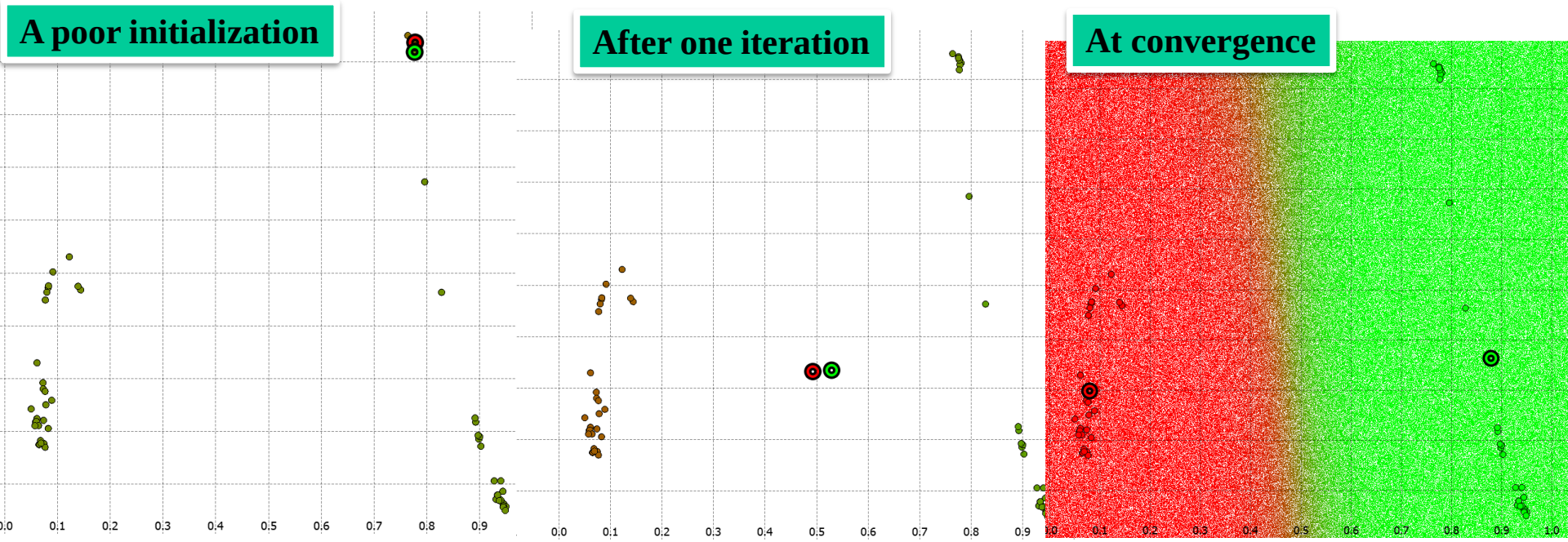




Would soft K-means be able to separate the two persons correctly with a large β for a good initialization?

Yes. It takes into account close-by datapoints, and discards influence of datapoints far away.

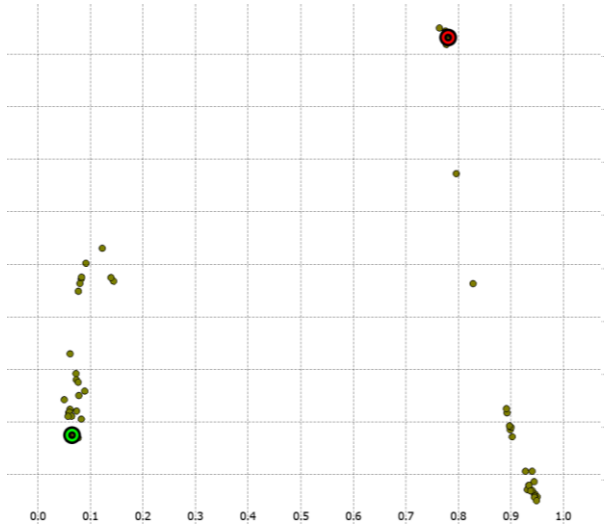




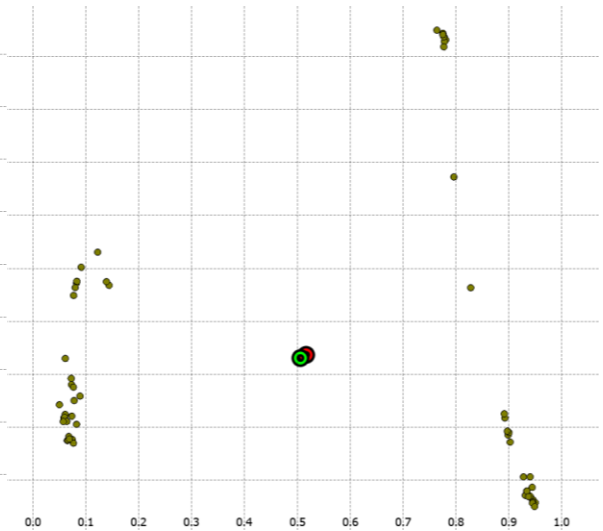
Would soft K-means be able to separate the two persons correctly with a large β with a poor initialization of the centroids?

Yes, even when the two centroids are initialized close to one another and in one region of the space. The centroids are quickly attracted by either of the two groups.

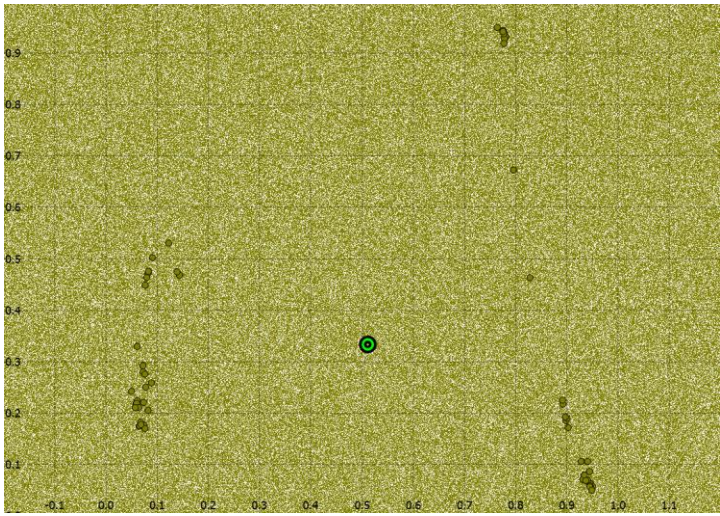
At initialization



After one iteration

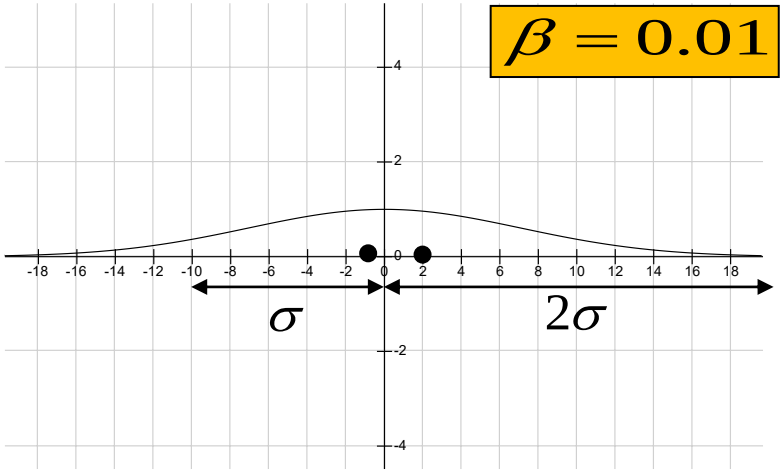


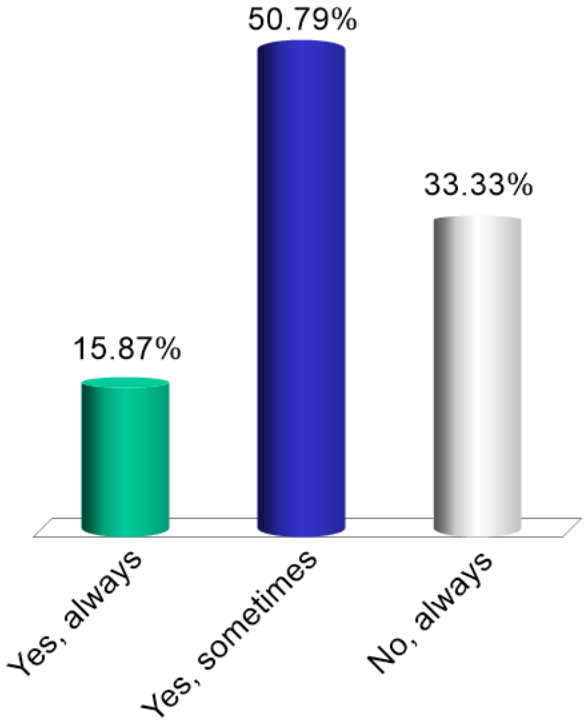
At convergence



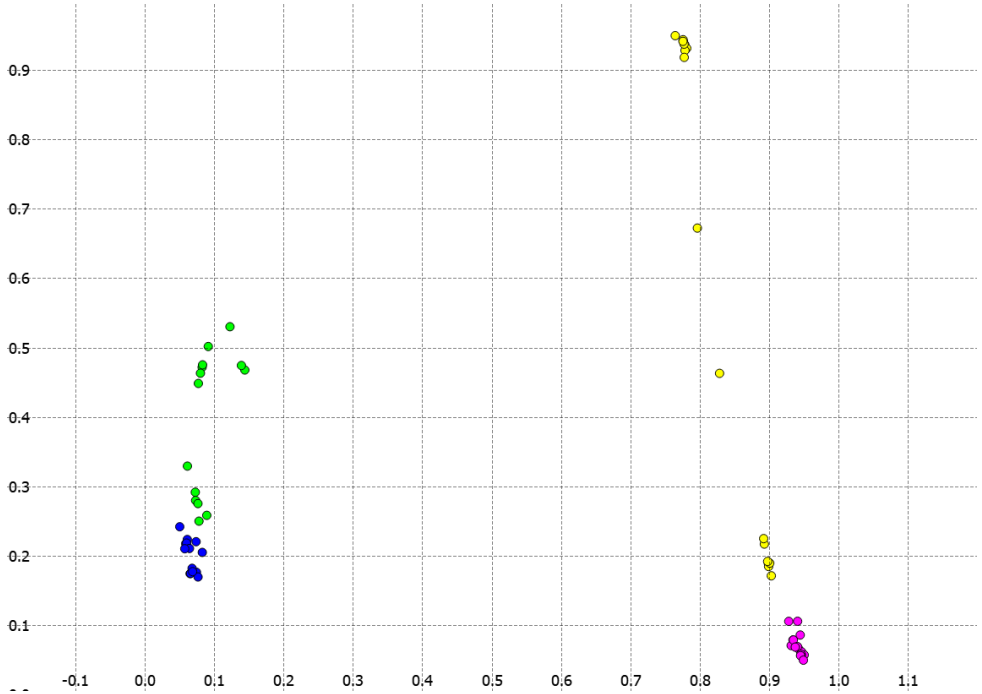
Would soft K-means be able to separate the two persons correctly with a small β with poor initialization of the centroids?

No. With a small β, all centroids are to the mean of the dataset and end up superimposed to one another.





- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses



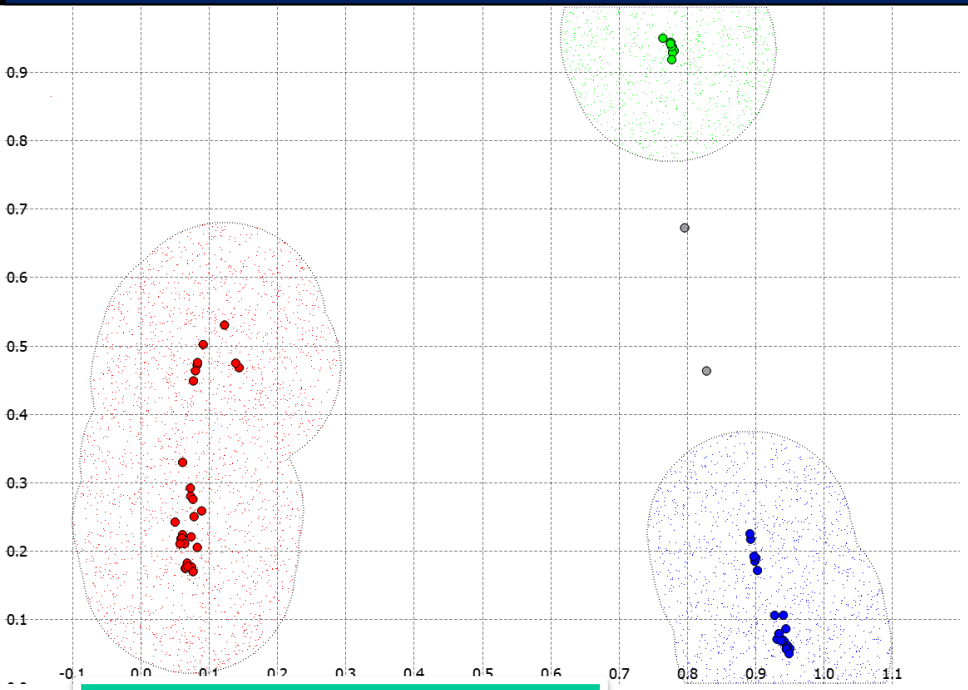
Projection onto first two principal components after PCA

Would DBSCAN be able to separate the two persons correctly when m_{data}=1?

A. Yes, always

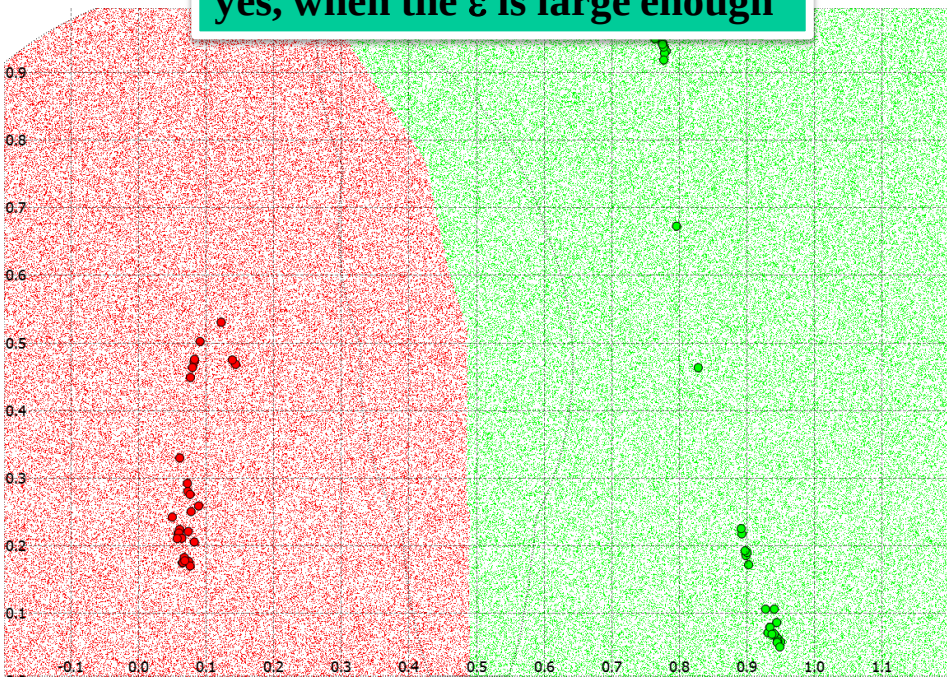
B. Yes, sometimes

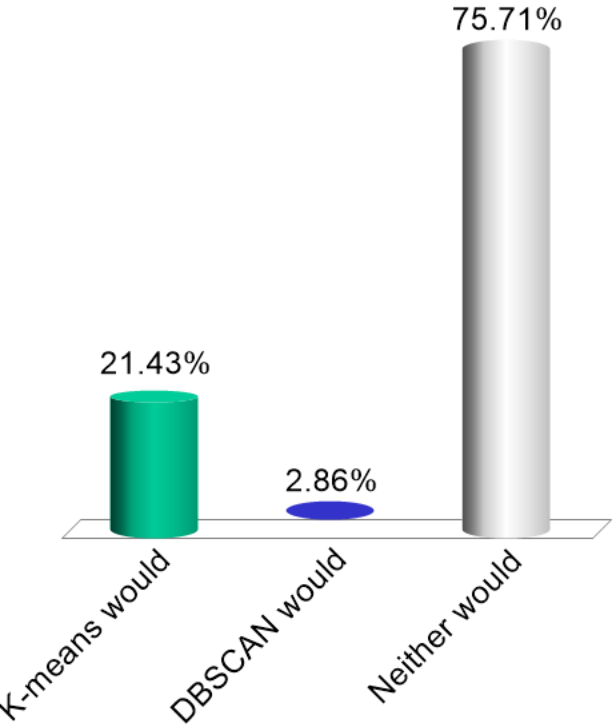
C. No, always



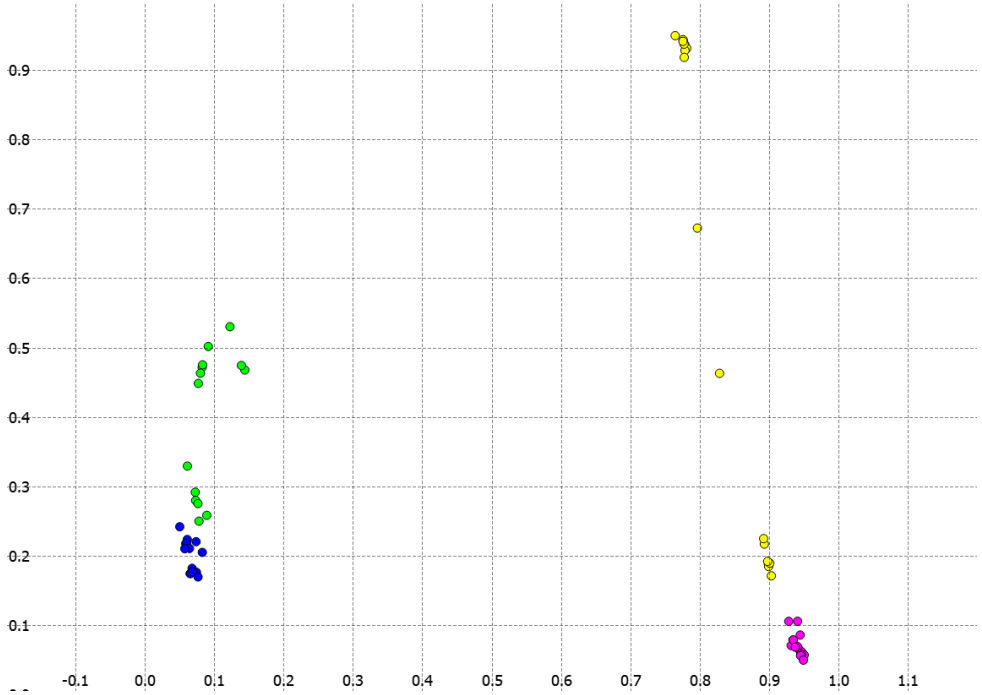
No, when the ε is too small

yes, when the ε is large enough



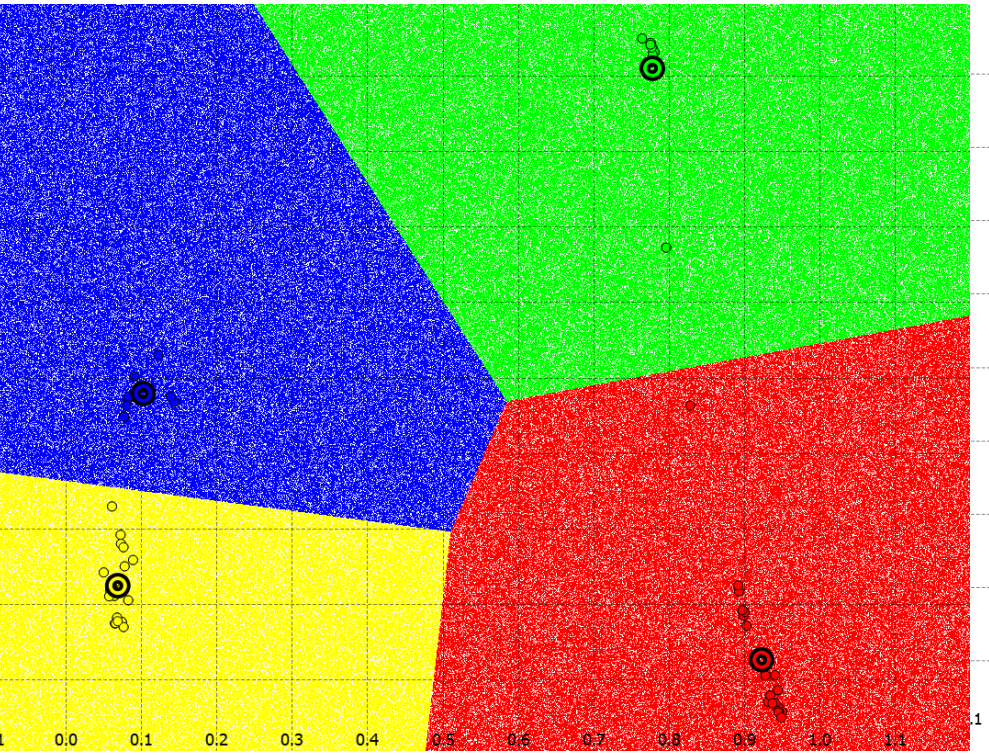


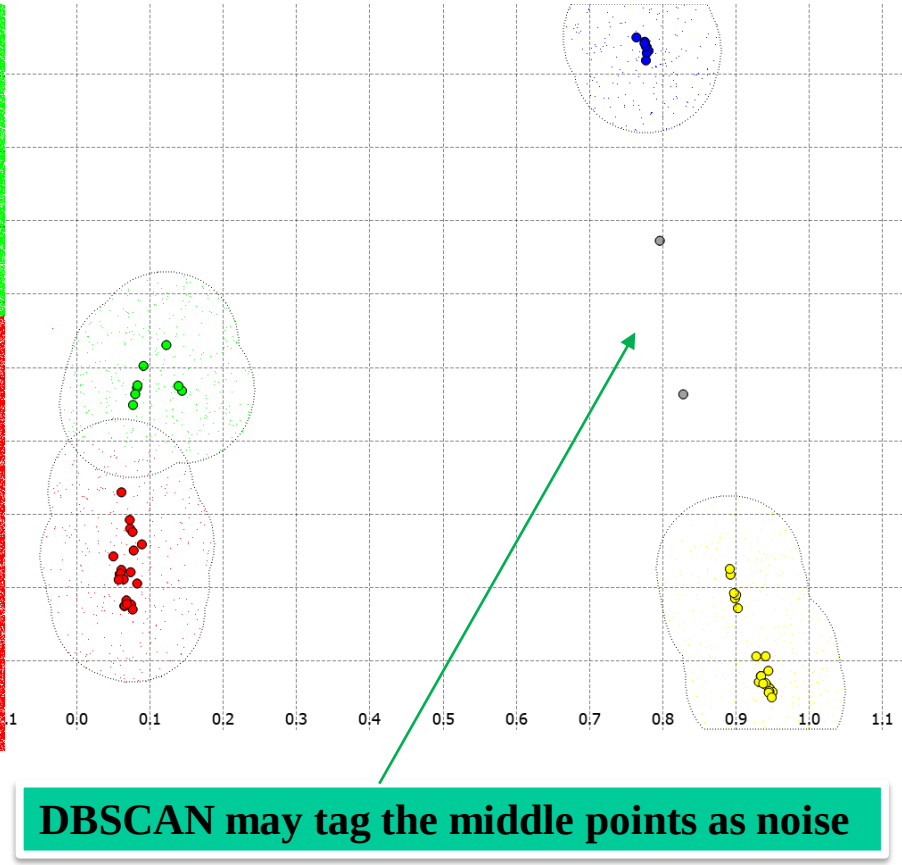
- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses



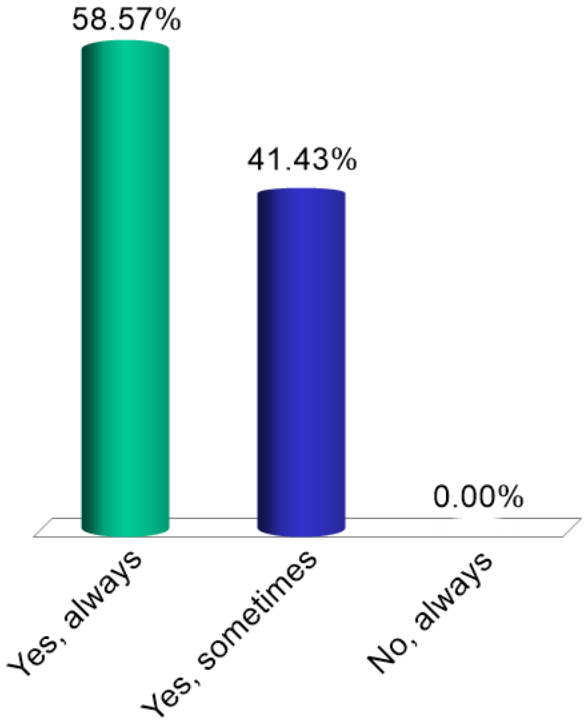
Would K-means or DBSCAN be able to separate the 4 classes correctly?

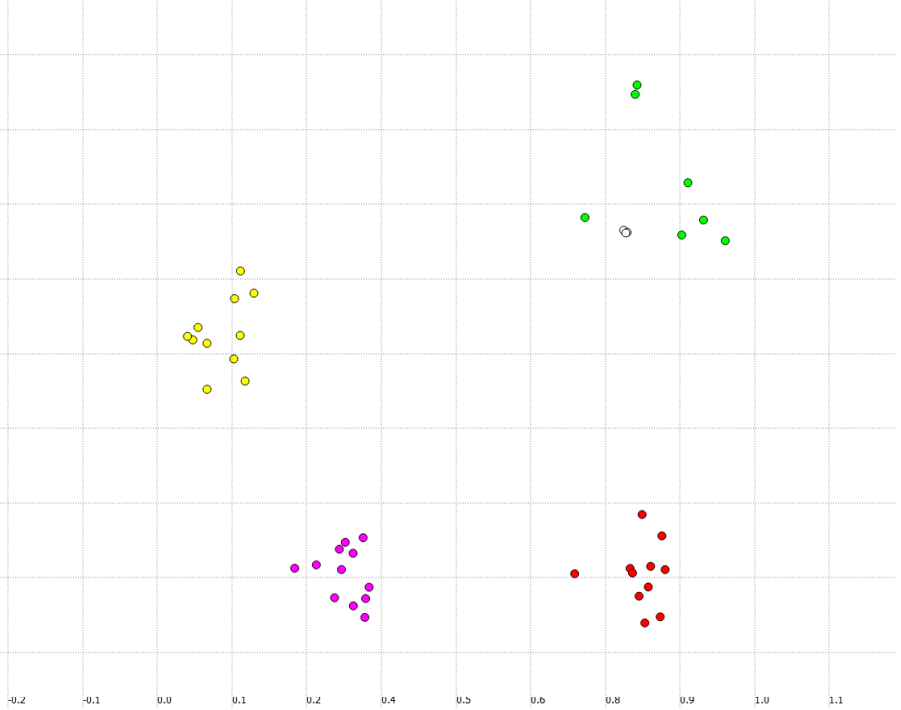
A. K-means would
B. DBSCAN would
C. Neither would





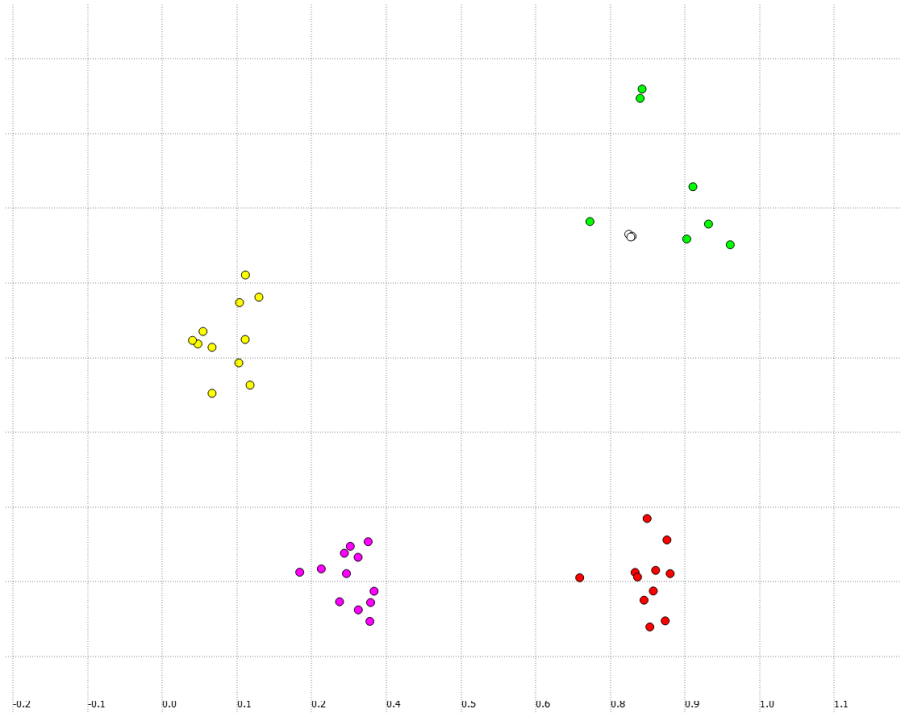
Neither of the two methods can cluster the 4 clusters correctly as the distance within clusters is bigger than across clusters for the glasses/no-glasses groups.



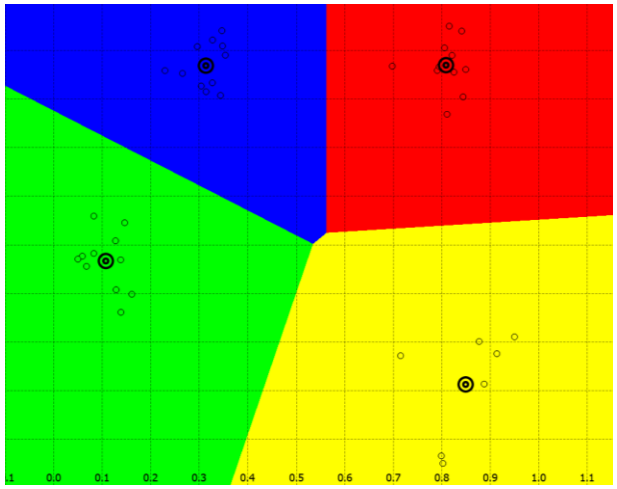


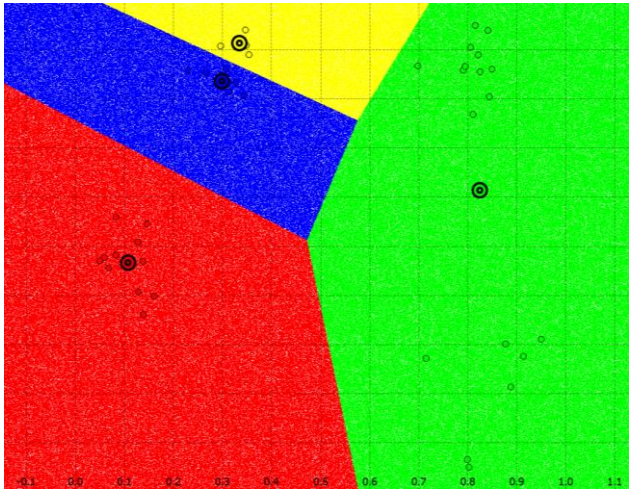
Would K-means be able to separate the 4 classes correctly?

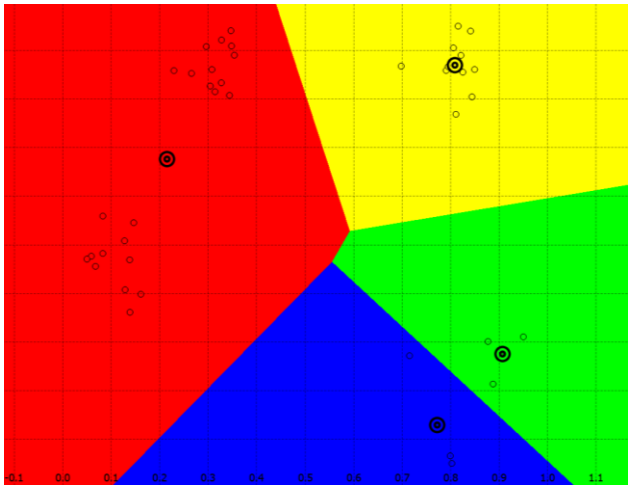
A. Yes, always
B. Yes, sometimes
C. No, always



Three solutions, depending on initialization







Evaluation of Clustering Methods

Two types of measures: Internal versus external measures

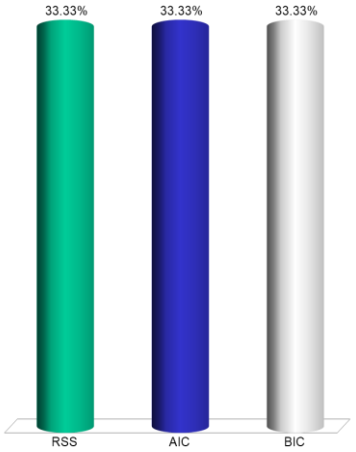
Internal measures rely on datapoints only and on a good choice of measure of similarity:

Examples: RSS, BIC and AIC

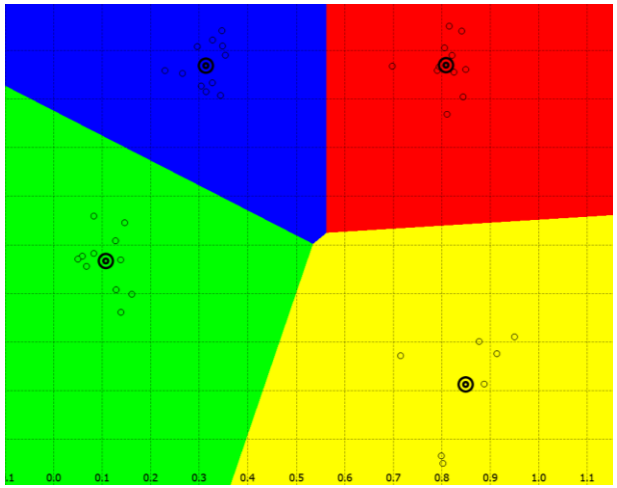
External measures rely on ground truth (class labels):

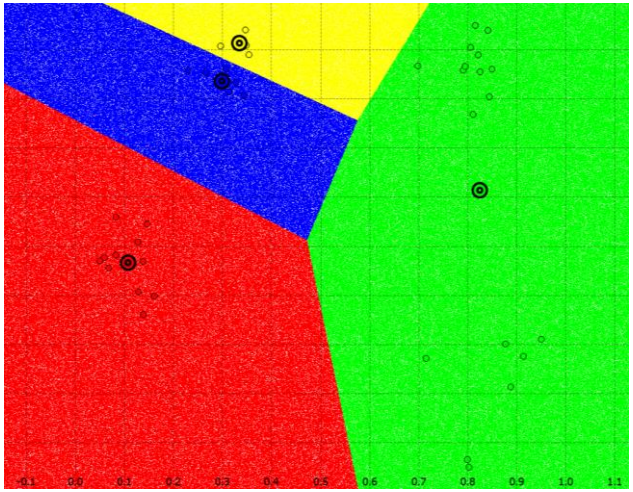
- Example: F1-measure

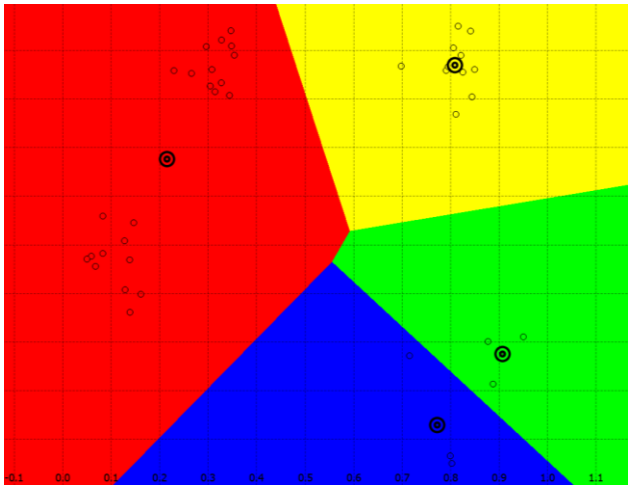
Which of the three metrics (AIC, BIC and RSS) would be most informative to determine the best solution across the 3 solutions below?



A. RSS
B. AIC
C. BIC







AIC, BIC, RSS measures of performances for K-Means

As the number of parameters (here K) remains the same, the BIC and AIC measures are affected only by the RSS measure. All three metrics will hence convey the same information.

$$AIC_{RSS} = RSS + B$$

Number of free parameters *B=(K*N)*
K: # clusters
N: # dimensions

$$\text{RSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \left| x - \mu^k \right|^2$$

$$BIC_{RSS} = RSS + \ln(M)\, B$$

RSS for K-Means:

Goal of K-means is to find cluster centers μ^k which minimize distortion.

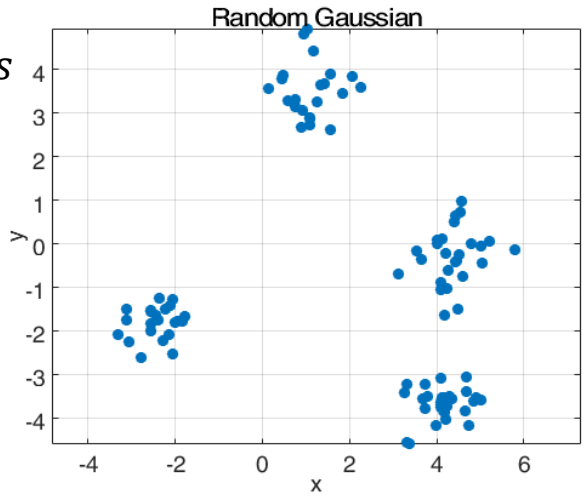
$$\text{RSS}=\sum_{k=1}^{K}\sum_{x^i \in c_k}\left\|x^i-\mu^k\right\|_2$$

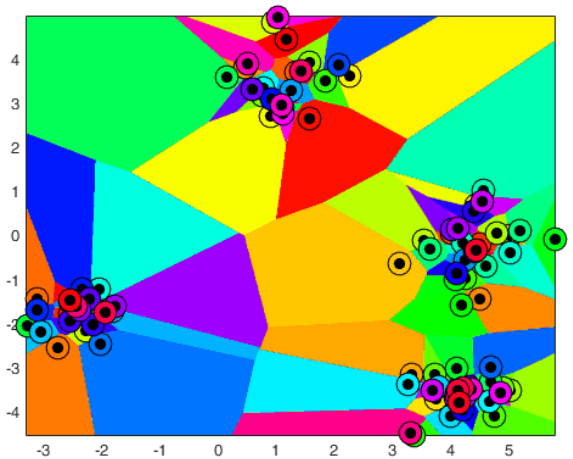
← *Measure of Distortion*

By $\uparrow K$ we $\downarrow RSS$, what is the optimal K such that $RSS \to 0$?

➢ $RSS = 0$ when $K = M$. One has as many clusters as datapoints!

M: 100 *datapoints*
N: 2 *dimensions*



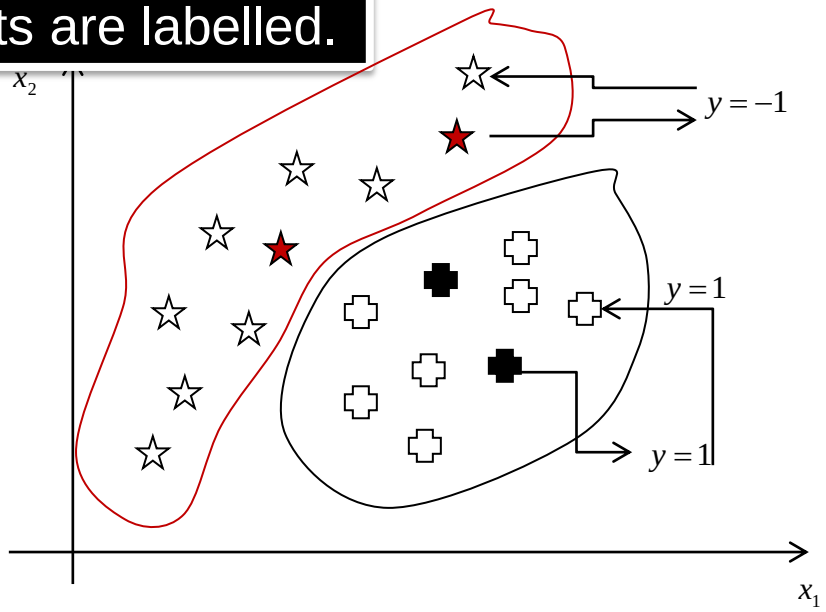


RSS: 0
K: M *clusters*

➢ However, it can still be used to determine an 'optimal' K by *monitoring the slope of the decrease of the measure* as K increases.

Semi-supervised clustering

A subset of the date points are labelled.



Clustering F1-Measure :

F_1 provides a measure of how good the clustering is:

$F_1 \in [0,1]$

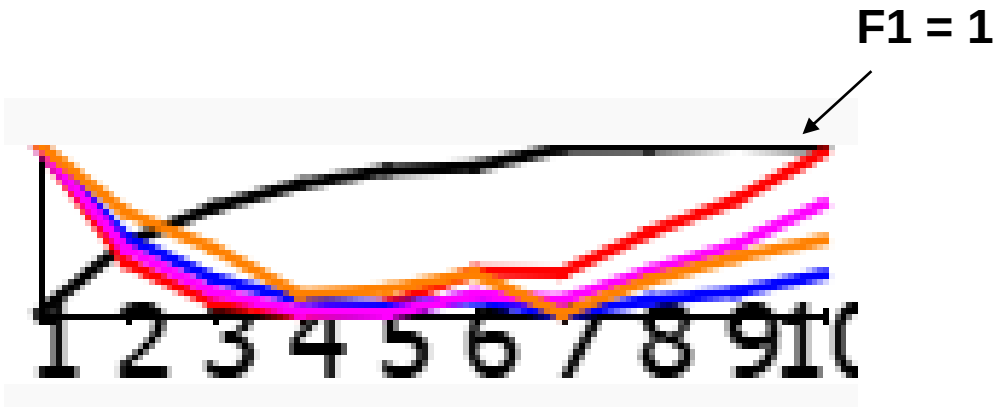
$F_1 = 1$ is the optimum.

Tradeoff between clustering correctly all datapoints of the same class in the same cluster and making sure that each cluster contains points of only one class.

F1-measure and other metrics:

20% datapoints labelled

RSS
AIC
BIC
F1

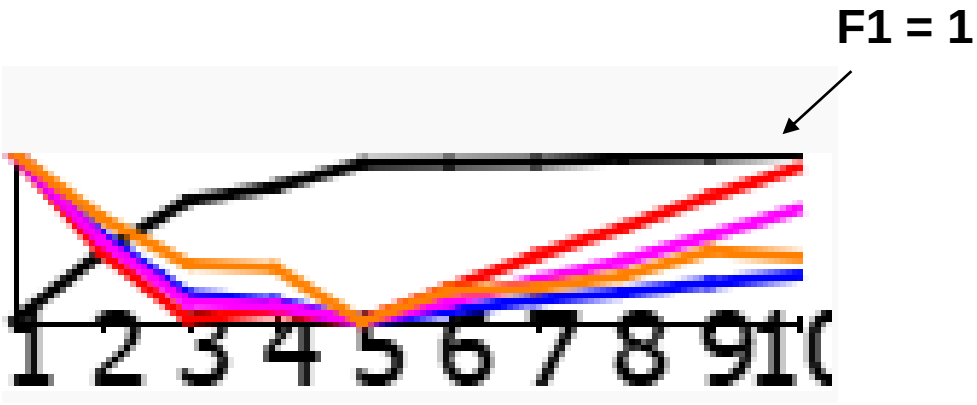


Which is the correct number of clusters?

F1-measure and other metrics:

50% datapoints labelled

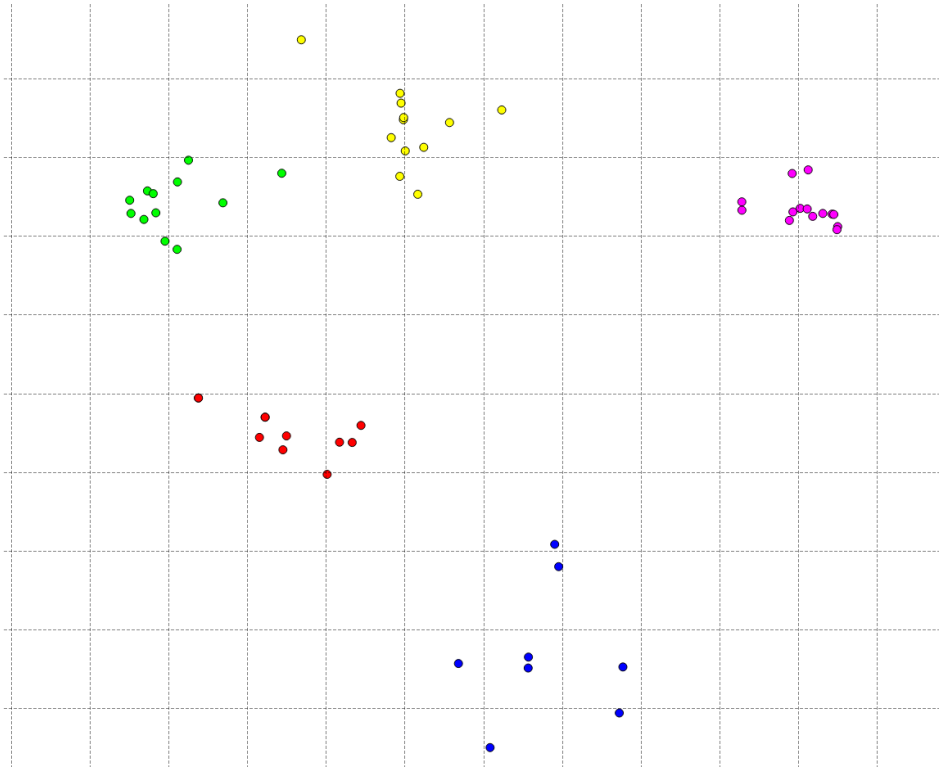
RSS
AIC
BIC
F1



Which is the correct number of clusters?

F1-measure and other metrics:

Projections onto two first eigenvectors



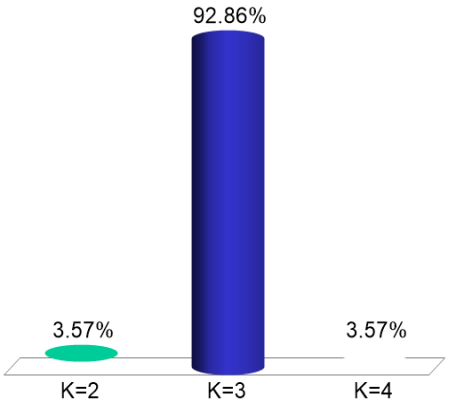
True number of clusters was 5.

Which is the best solution?

A. K=2

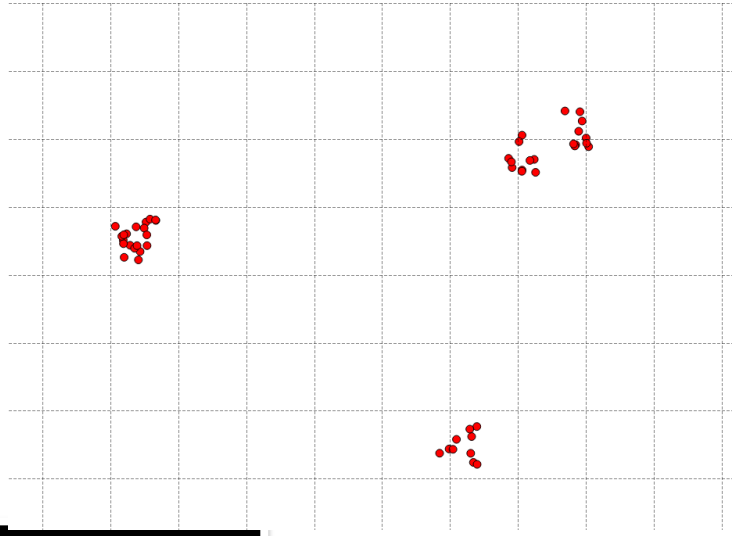
B. K=3

C. K=4



K	BIC	AIC	F1-measure (computed on 20% labelled datapoints)
2	401	356	0.5
3	252	256	0.61
4	297	283	0.72

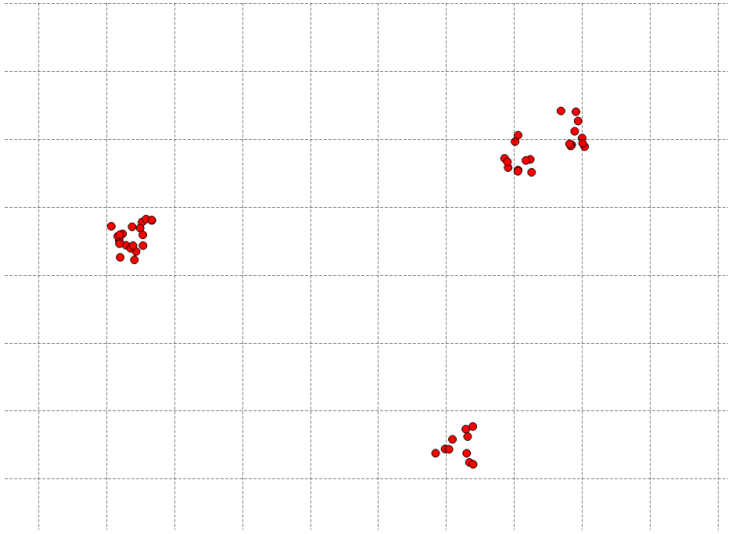
K	BIC	AIC	F1-measure (computed on 50% labelled datapoints)
2	258	265	0.62
3	252	266	0.75
4	275	290	0.52



Original dataset
True value: K=3

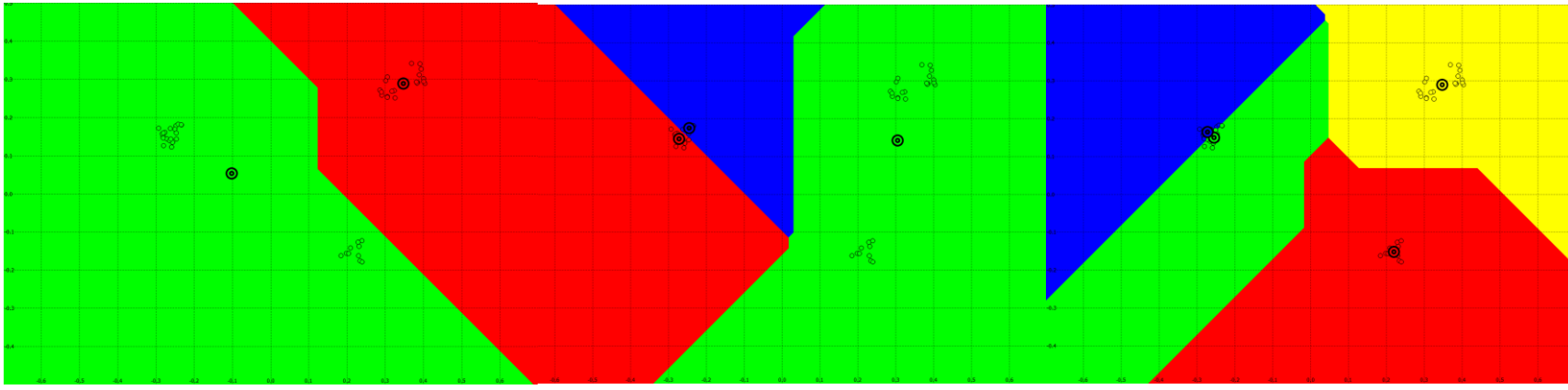
Trust F1 with more datapoint labelled
Optimum on all 3 metrics.

		AIC	F1-measure (computed on 50% labelled datapoints)
2	258	265	0.62
3	252	266	0.75
4	275	290	0.52



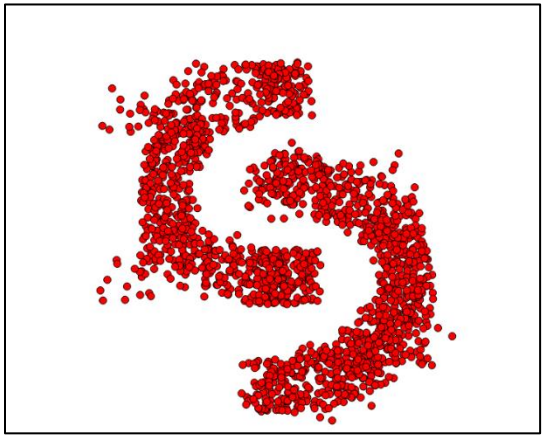
Original dataset
True value: K=3

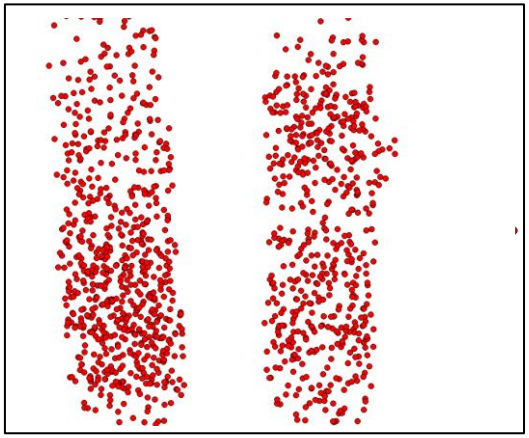
K-means is sensitive to initialization. Make sure to repeat and take best run when comparing results in RSS, AIC and BIC.

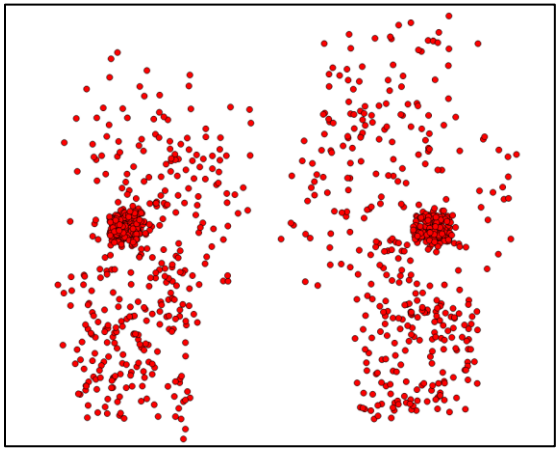


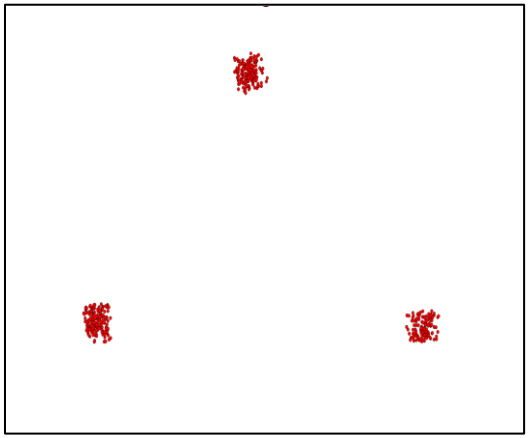
Clustering methods -- exercise

Which clustering method (Hard/Soft k-means, DBSCAN) would you use to cluster each of the following datasets and why?



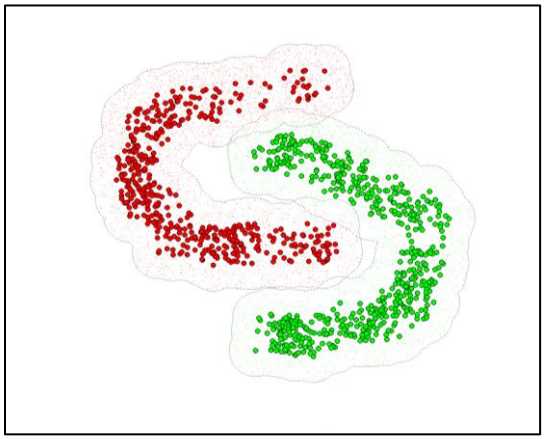




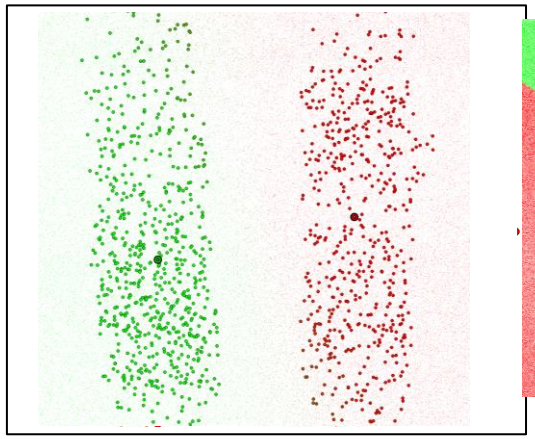


Use the cheapest clustering method (in computational costs) whenever possible.

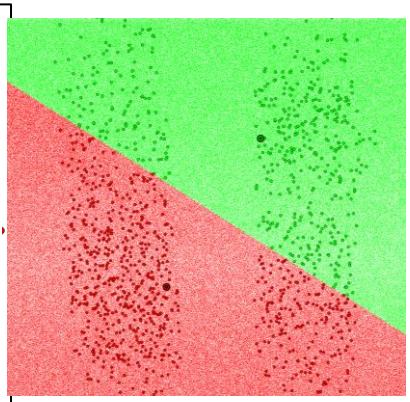
The cheapest is hard K-means, followed by soft K-means (computing an exponential is more costly than computing norm 2) and then DBSCAN.



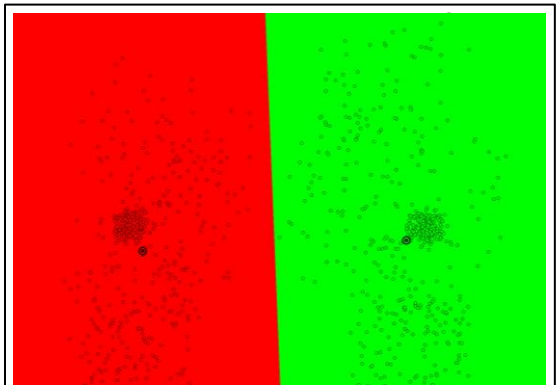
DBSCAN



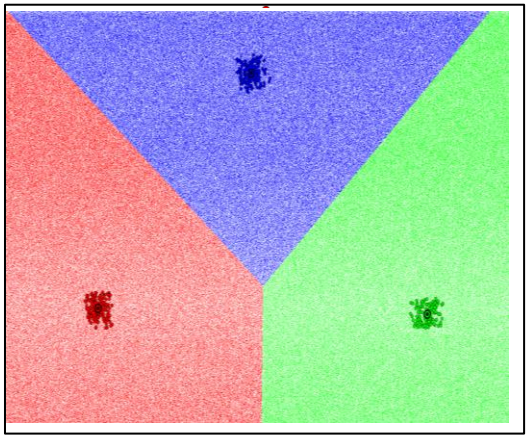
Soft K-means



Hard K-means is possible but may find wrong solution because intra-cluster distance is smaller than inter-cluster disance



Hard K-means and soft Kmeans both possible; The large group helps to find the correct solution irrespective of initialization



Hard K-means