MACHINE LEARNING I

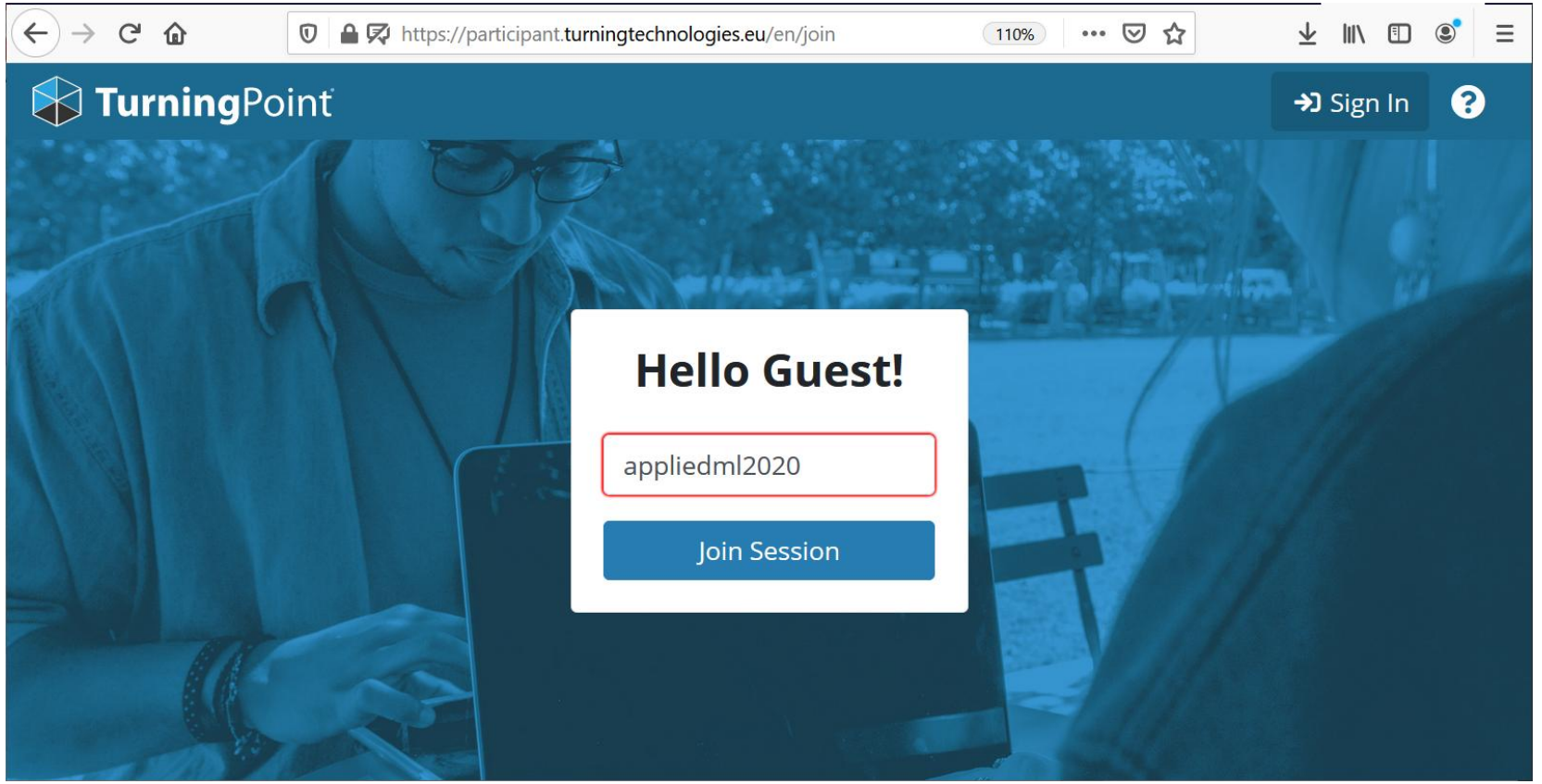
Clustering

Interactive Lecture

Launch polling system

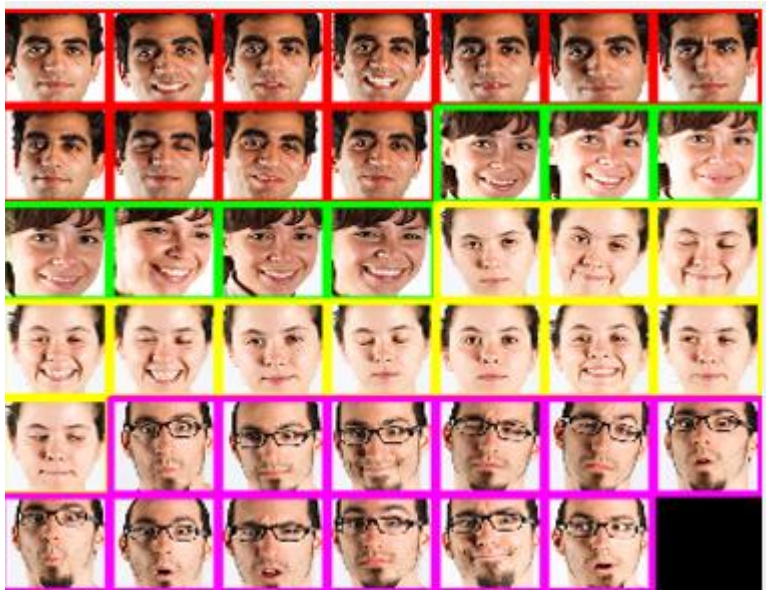
https://participant.turningtechnologies.eu/en/join

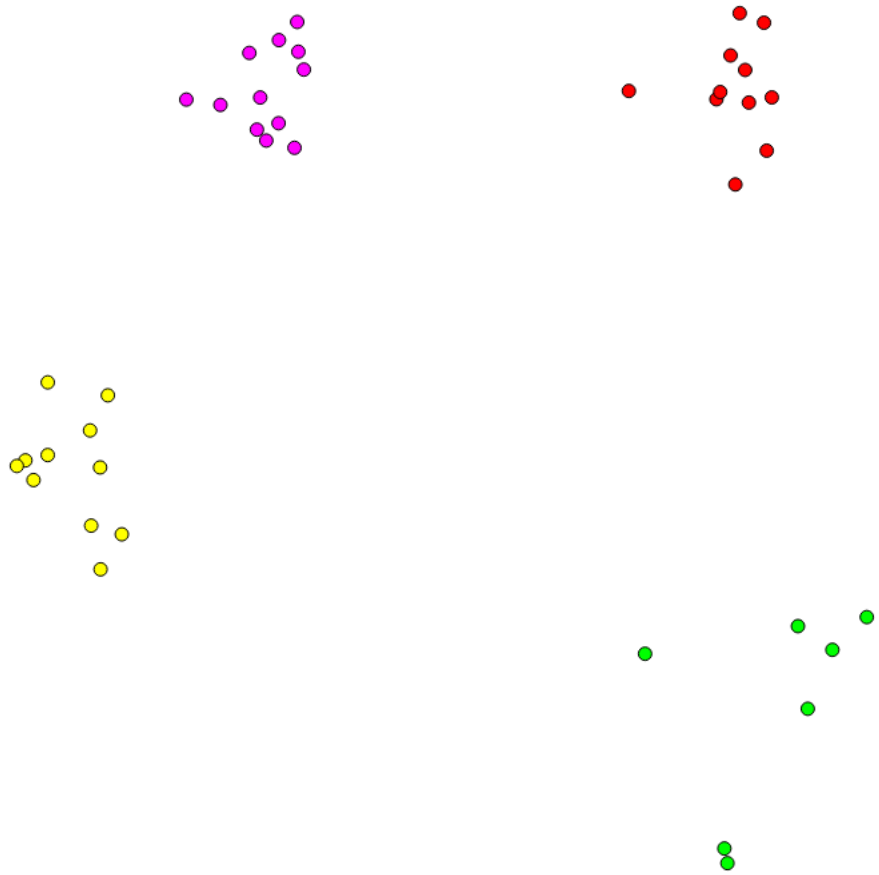
Acces as GUEST and enter the session id: *appliedml2020*



Clustering Principle

- Clustering methods know neither the number of clusters, nor how the points cluster.
- Easy if groups are well separated.



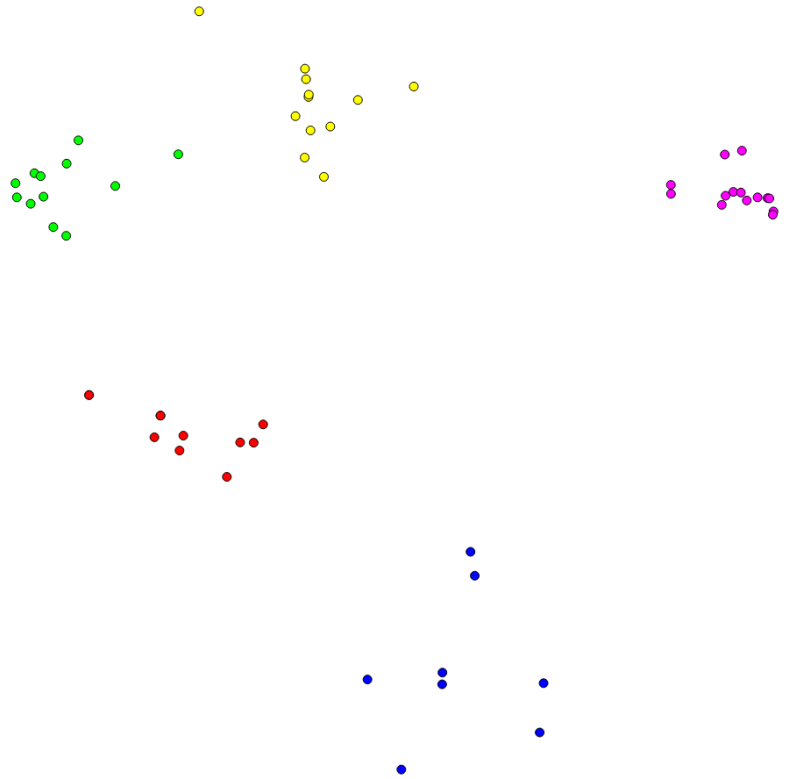


Semi-Supervised Clustering Principle

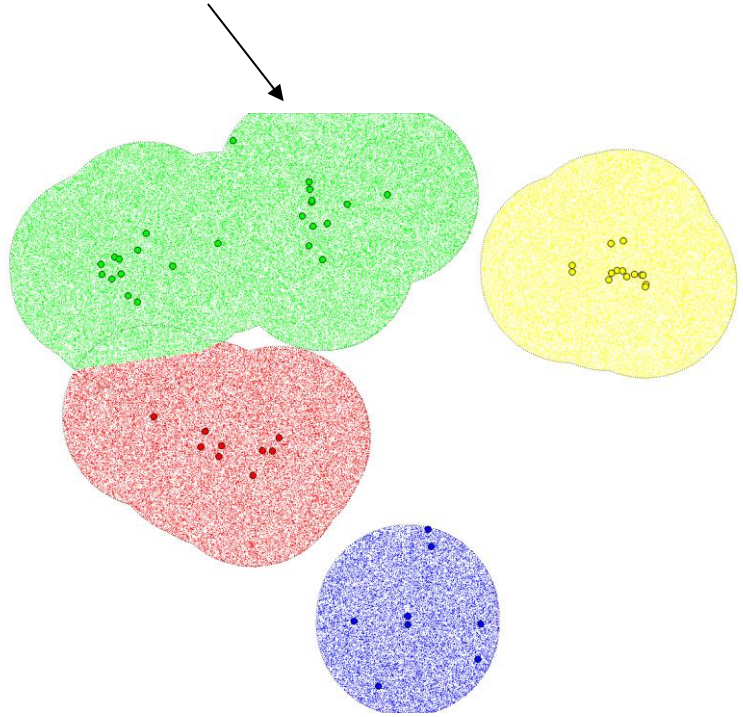
- When groups are not easily separable, one can use *semi-supervised* clustering.
- Semi-supervised clustering consists of labelling only a subset of the datapoints

→ number of clusters is known!

5 classes, less separable



Wrong classification of 2 tight groups



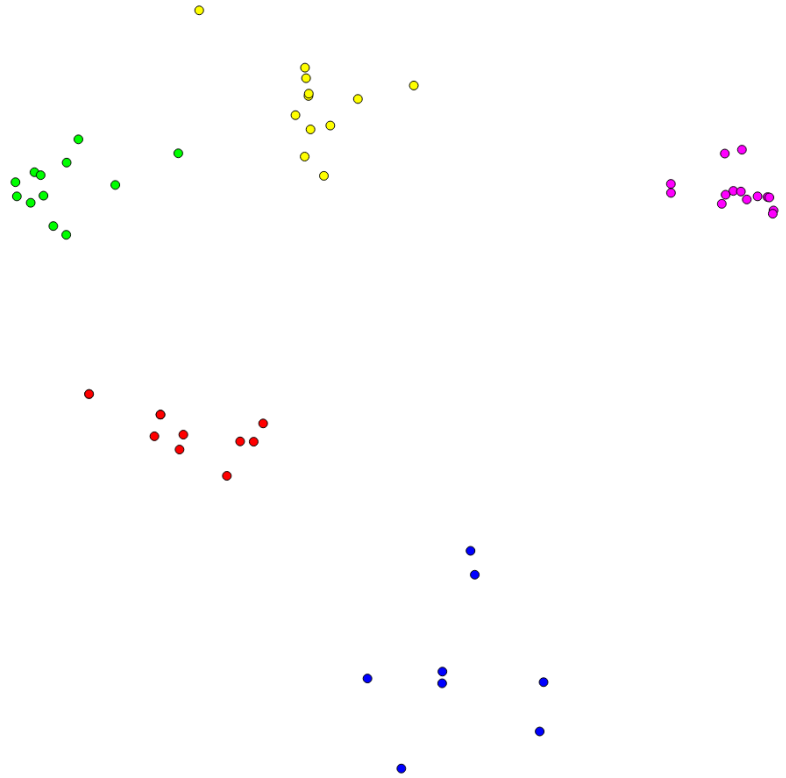
Unsupervised clustering result

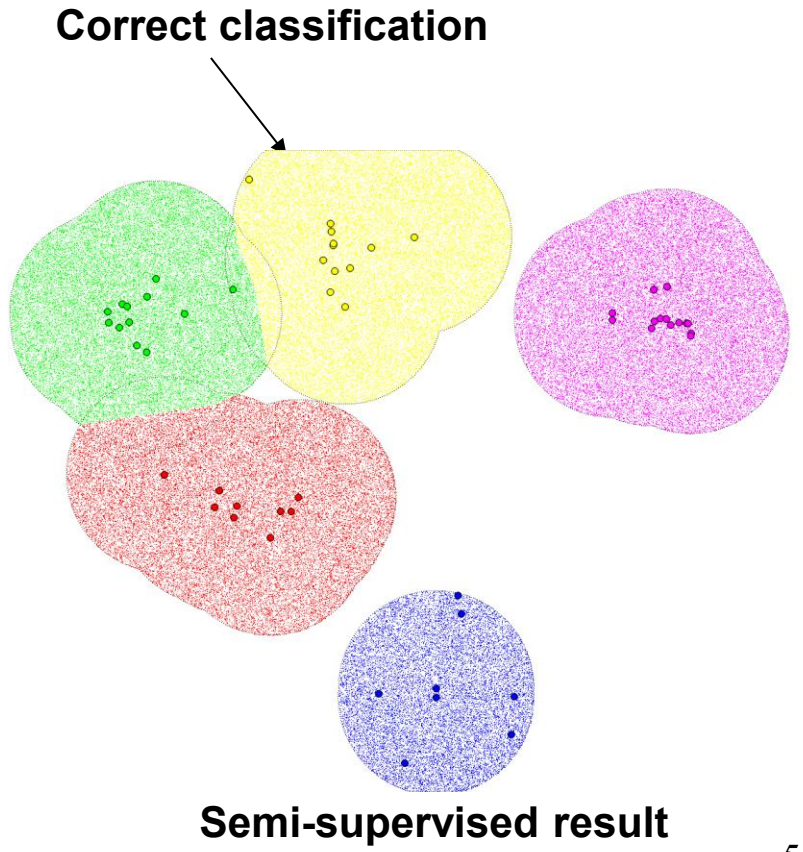
Semi-Supervised Clustering Principle

- When groups are not easily separable, one can use *semi-supervised* clustering.
- Semi-supervised clustering consists of labelling only a subset of the datapoints

→ number of clusters is known!

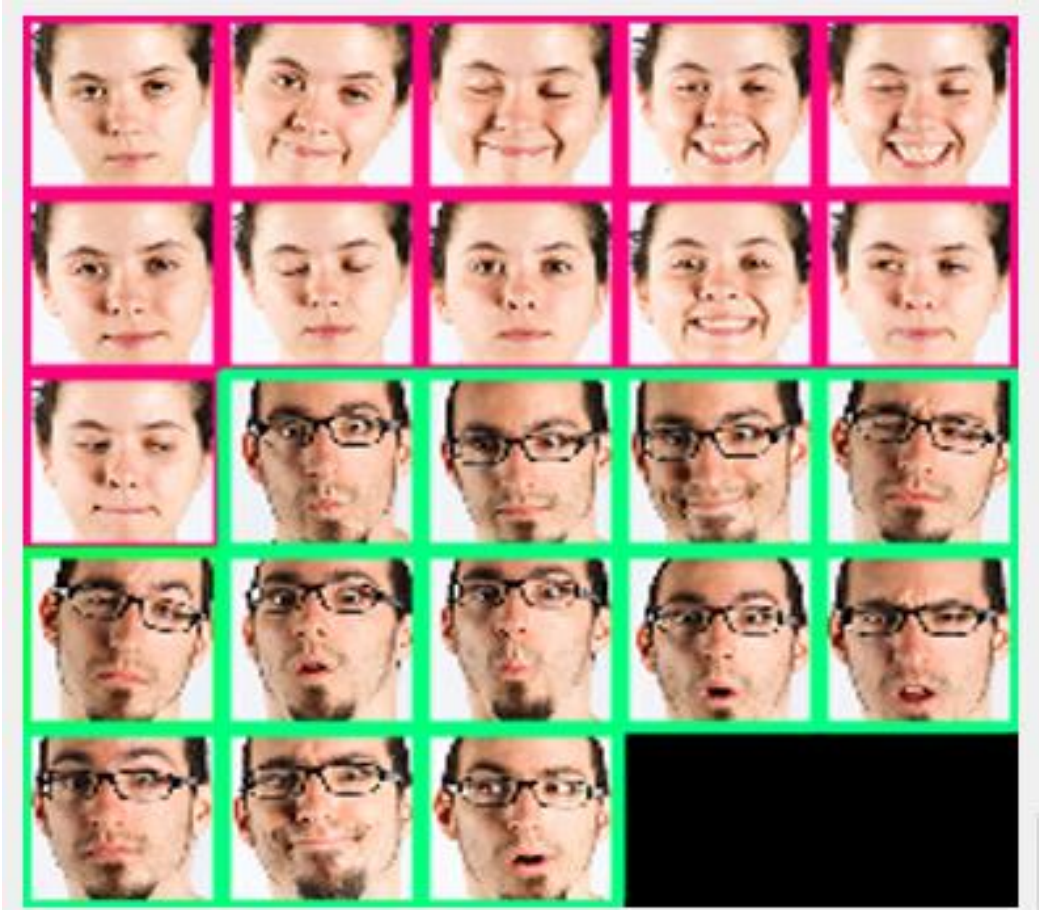
5 classes, less separable





Semi-supervised result

Clustering Metrics: Measure of similarity



Which subgroups of pictures are similar and why?

Measure of Similarity

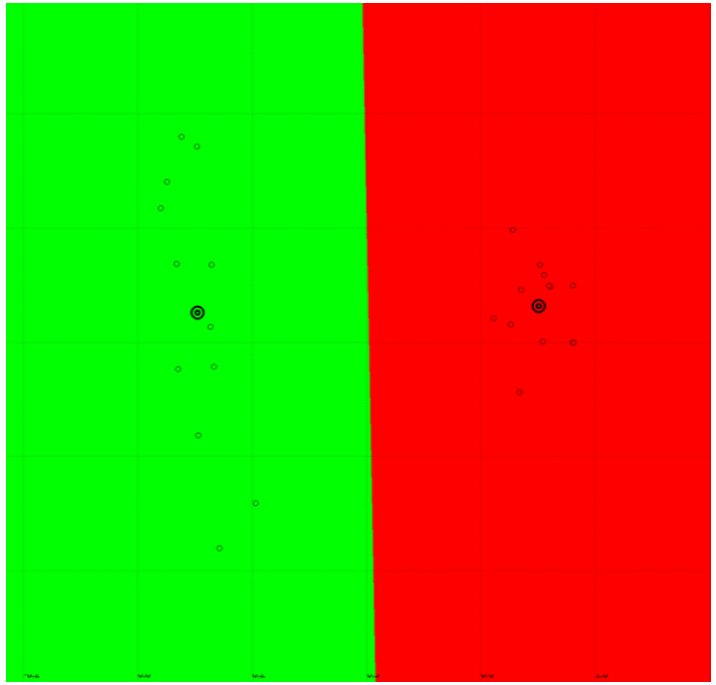
Groups of points are said to belong to the same cluster if they are similar enough.

Each clustering methods come with predefined metric(s).

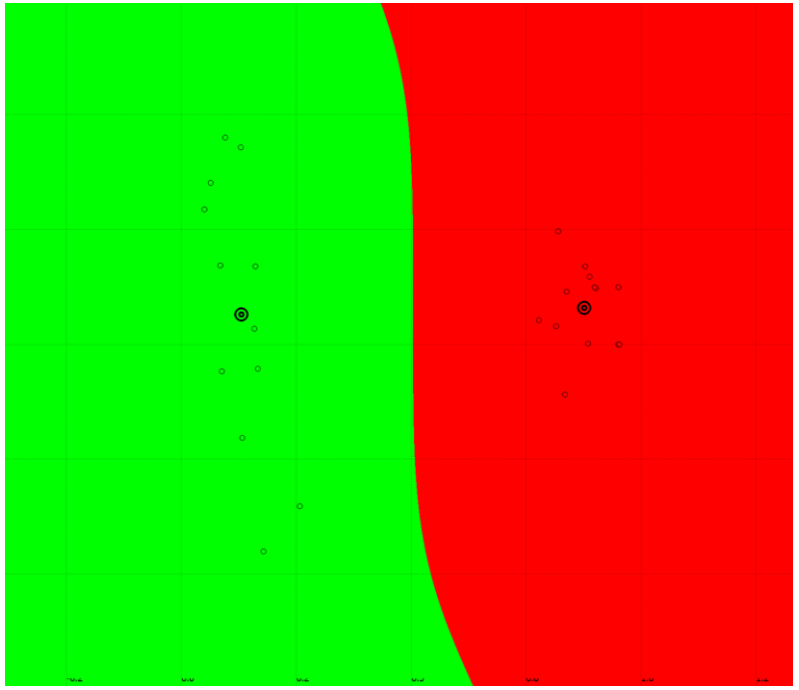
Measure of Similarity: K-means

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

K-means minimizes $J\left(\mu^1,...,\mu^K\right)=\sum_{k=1}^{K}\sum_{x^i\in c_k} d\left(x^i,\mu^k\right)$ with $d\left(x^i,\mu^k\right)=\sqrt[p]{\sum_{i=1}^{N}\left|x_i^i-\mu_i\right|^p}$



p=2



p=5

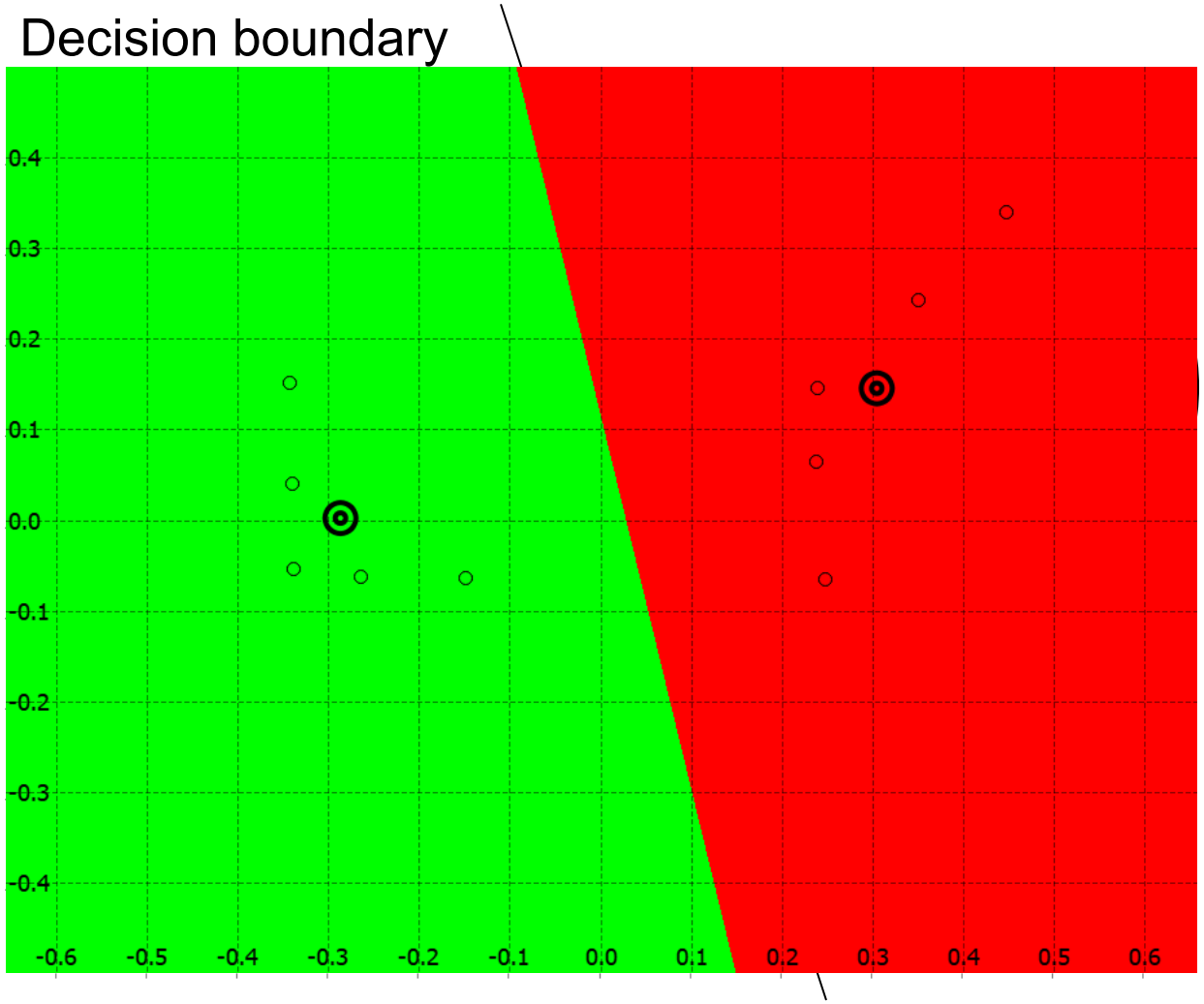
Measure of Similarity: DBSCAN

DBSCAN uses two criteria to decide on cluster assignment:

a) All points must be within a ball in norm-2 from one another

b) A cluster must contain a minimum number of datapoints.

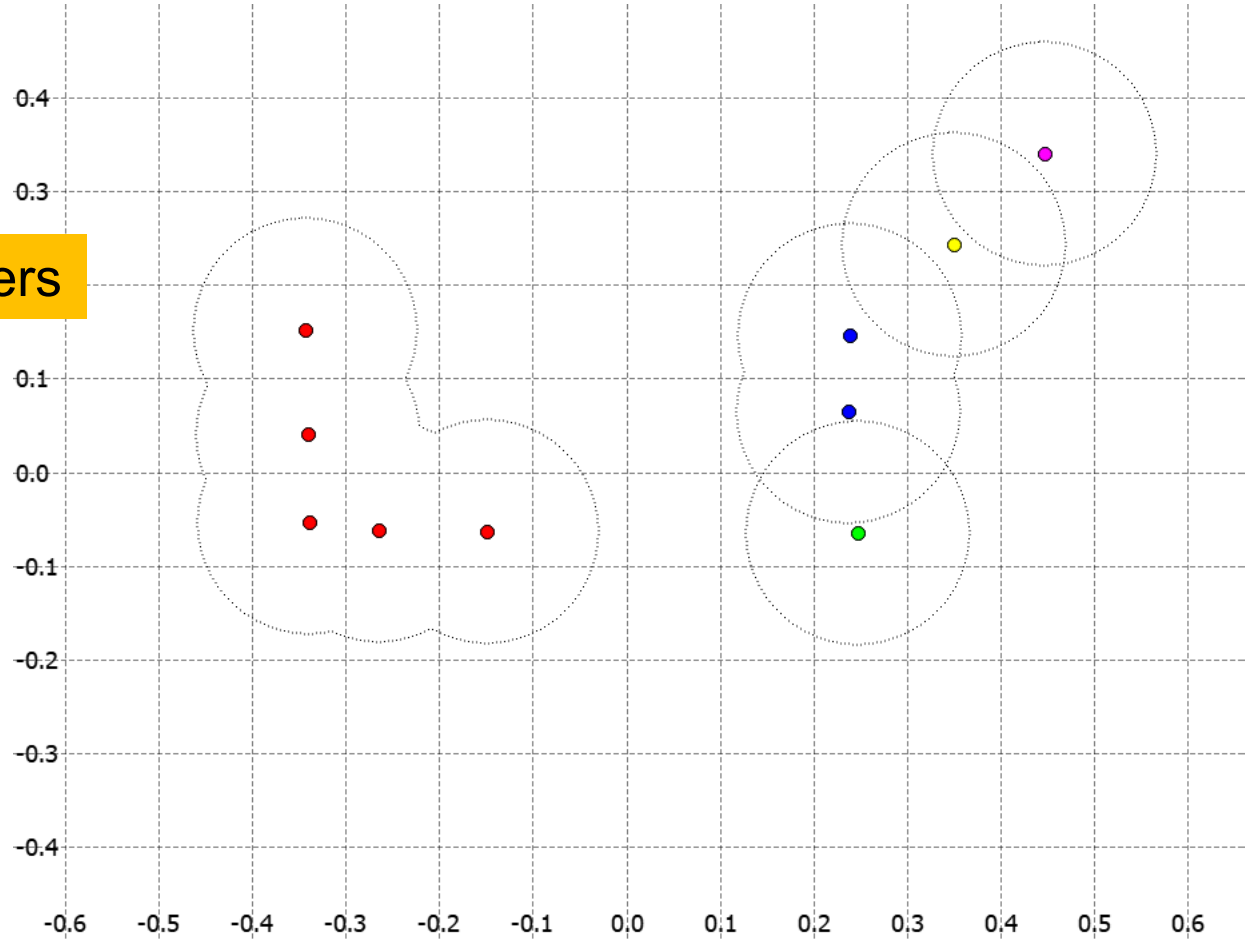
Measure of Similarity: K-means



Metrics is Norm-2, two equidistant balls surrounding the clusters

Measure of Similarity: DBSCAN

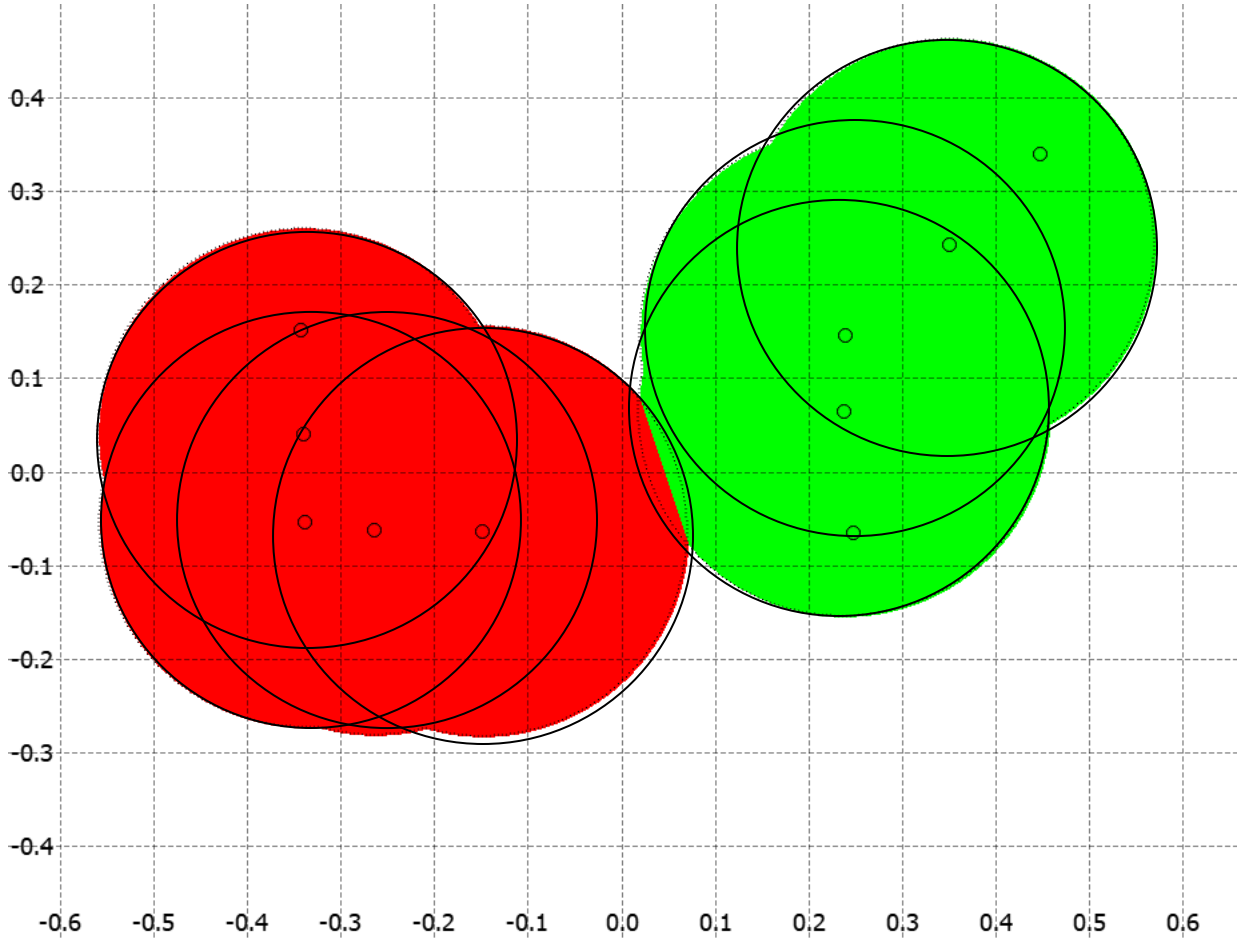
Generates tighter clusters



Metrics is Norm-2, at least 1 pt in each cluster

Measure of Similarity: DBSCAN

Generates larger balls



Metrics is Norm-2, at least 4 pts in each cluster

Clustering Principle

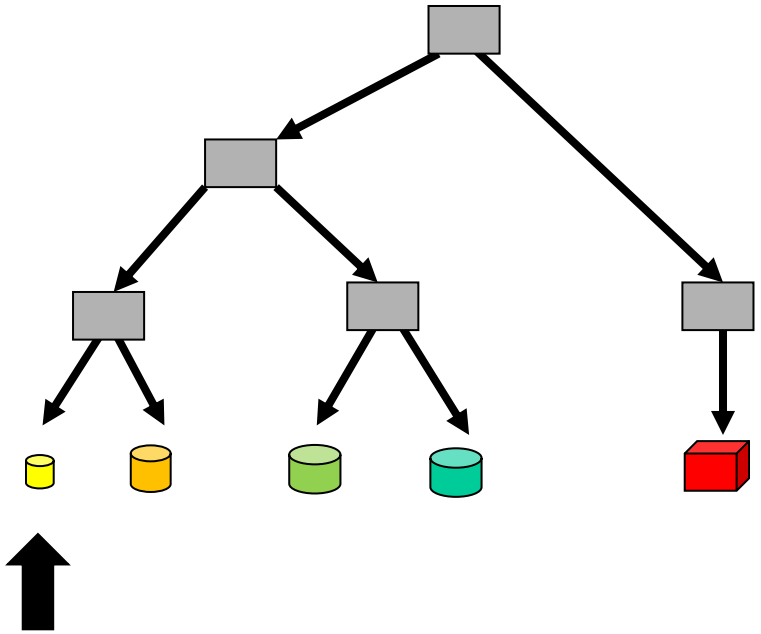
Clustering can be used as:

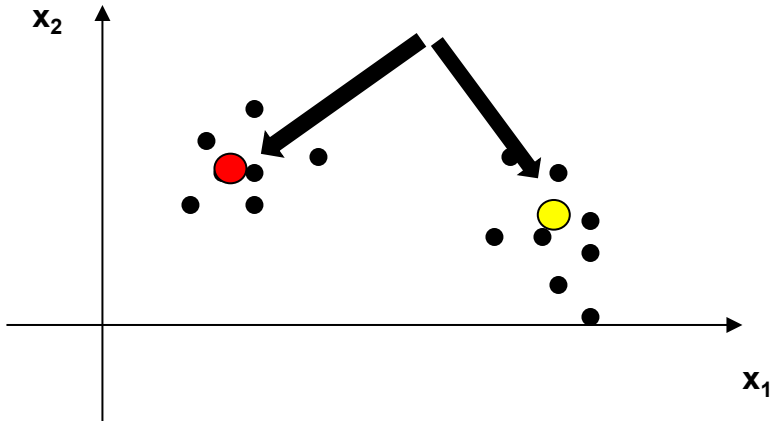
- *Feature extraction method*: for identifying underlying structure in data and salient features, best visualized through **cluster prototypes**.
- *Compression method*: for organizing the data and summarizing it through **cluster prototypes.**

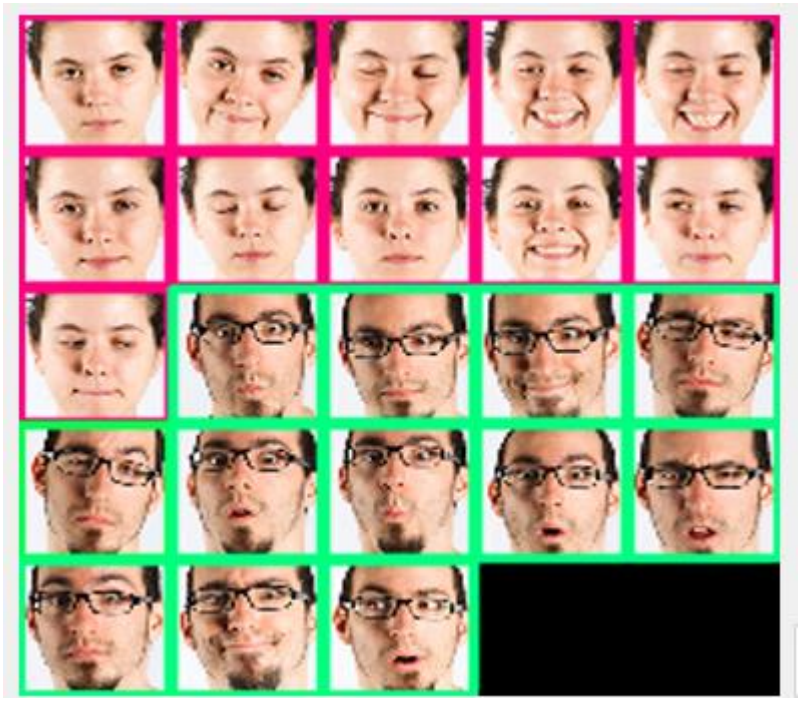
Clustering Prototypes

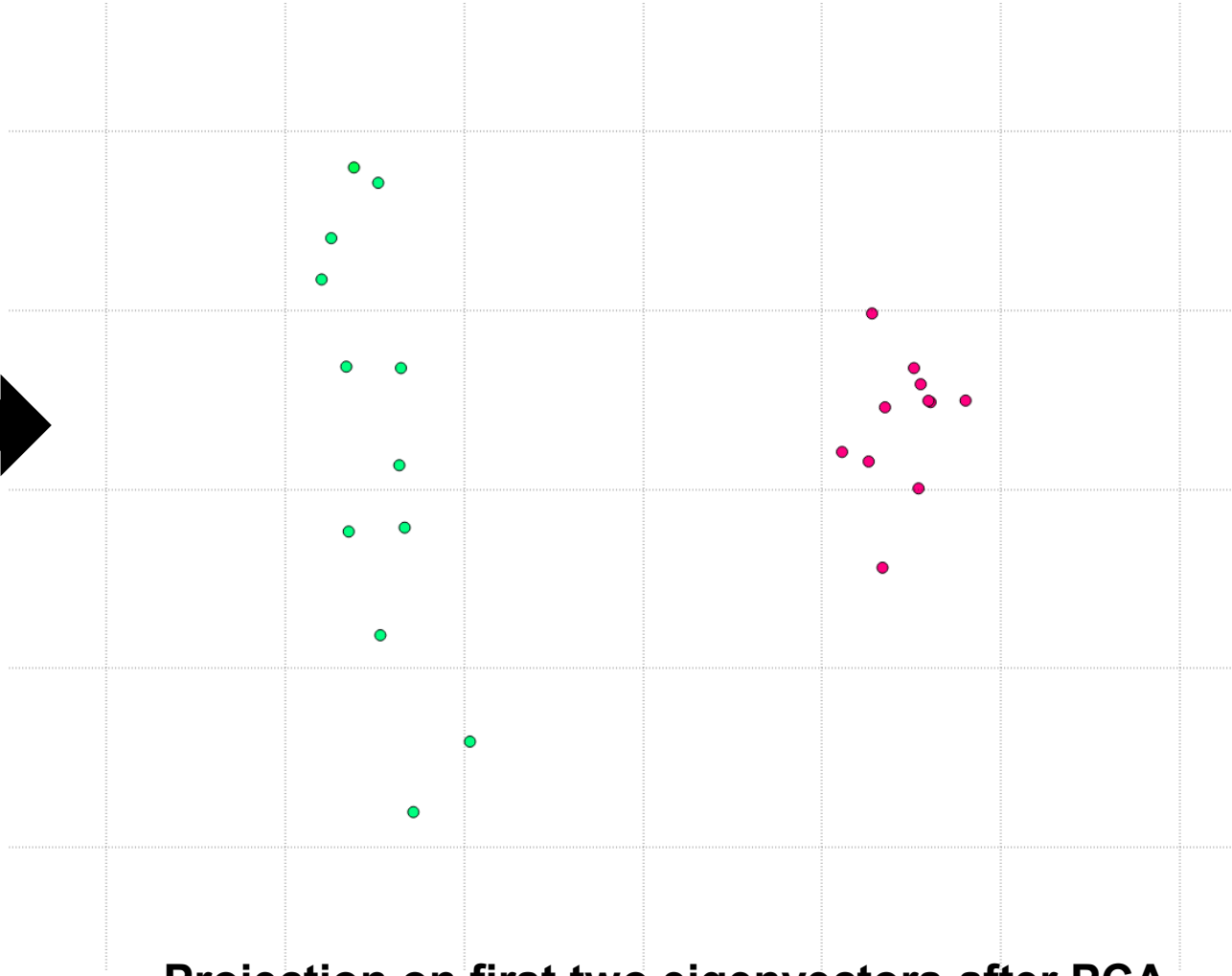
A cluster prototype can be:

- A typical datapoint, best representative of the cluster (one item at the end of a branch in hierarchical clustering)
- The average of the datapoints in the cluster (e.g. centroid in K-means)



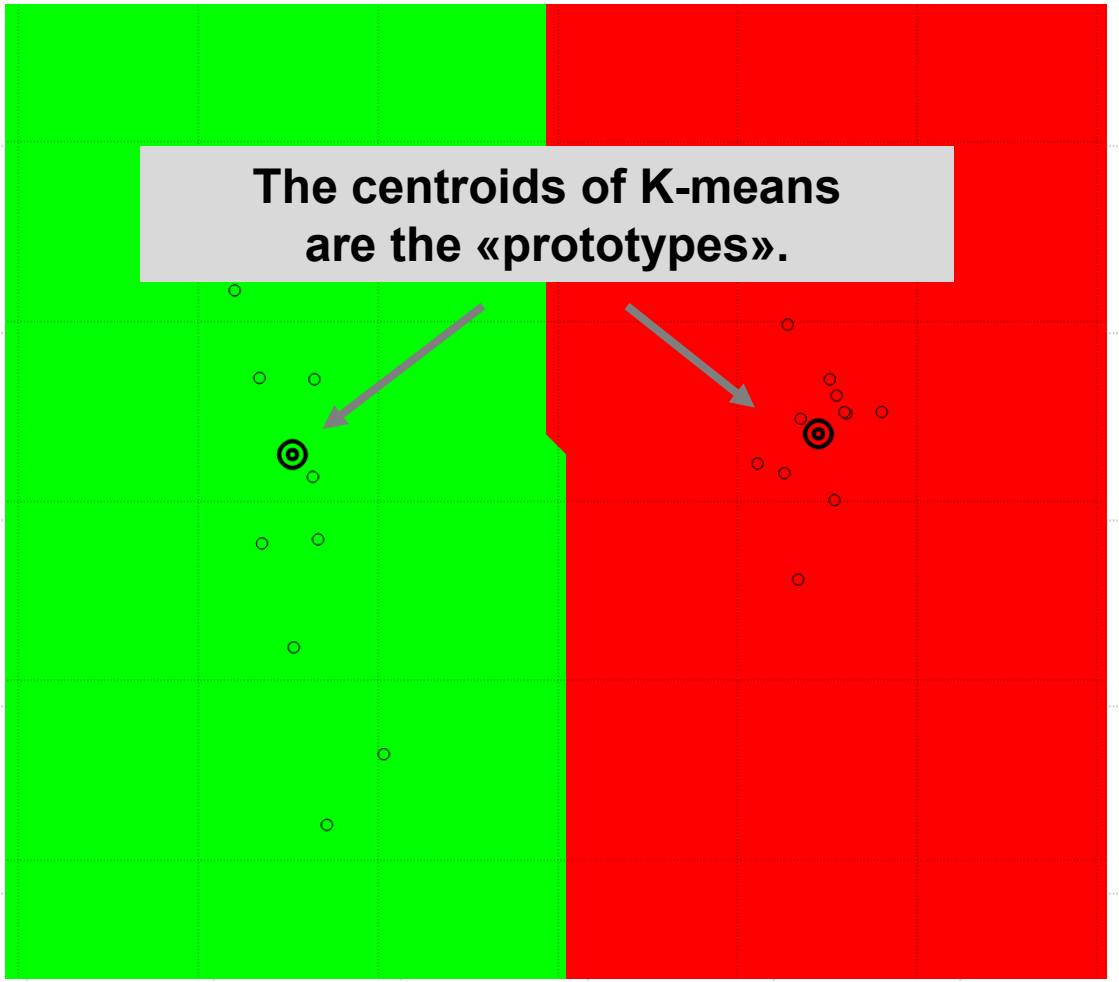






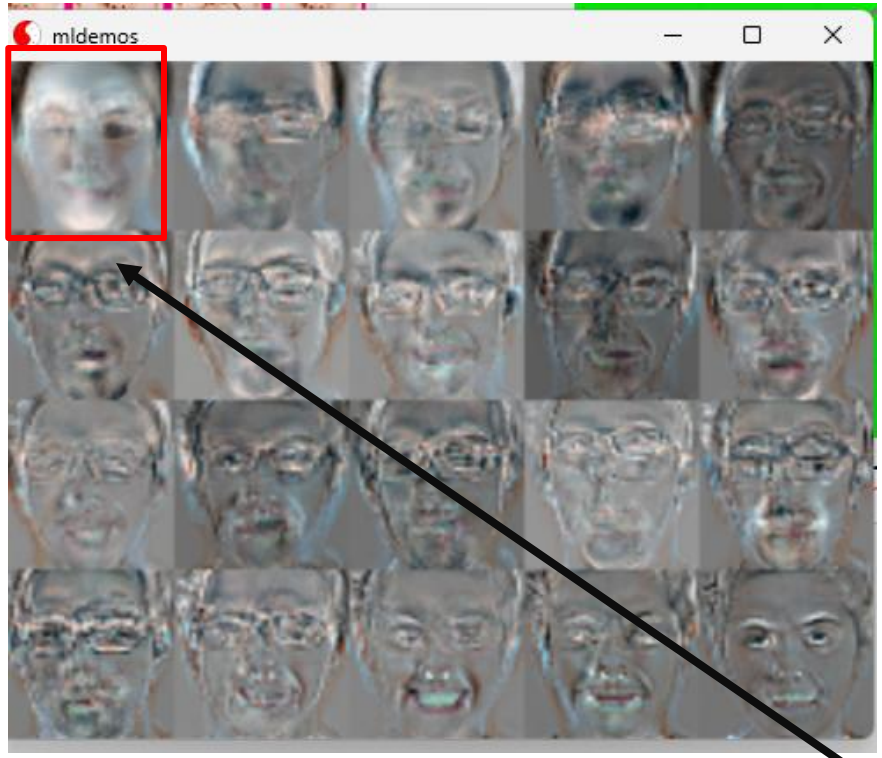
- Start with a set of datapoints.
- Algorithm does not know the true labels.
- It knows neither the number of groups nor how to regroup datapoints.

Projection on first two eigenvectors after PCA



Clustering with K-means, K=2, norm-1

- If K-means was run on the full dimension of the dataset (not on the 1st 2 PCA projections), the prototypes would be images.
- The prototypes are usually different from any of the dataset images. They are a sort of average of all images of each person.



The «prototypes» indicate how the two groups of images project on the eigenvectors.

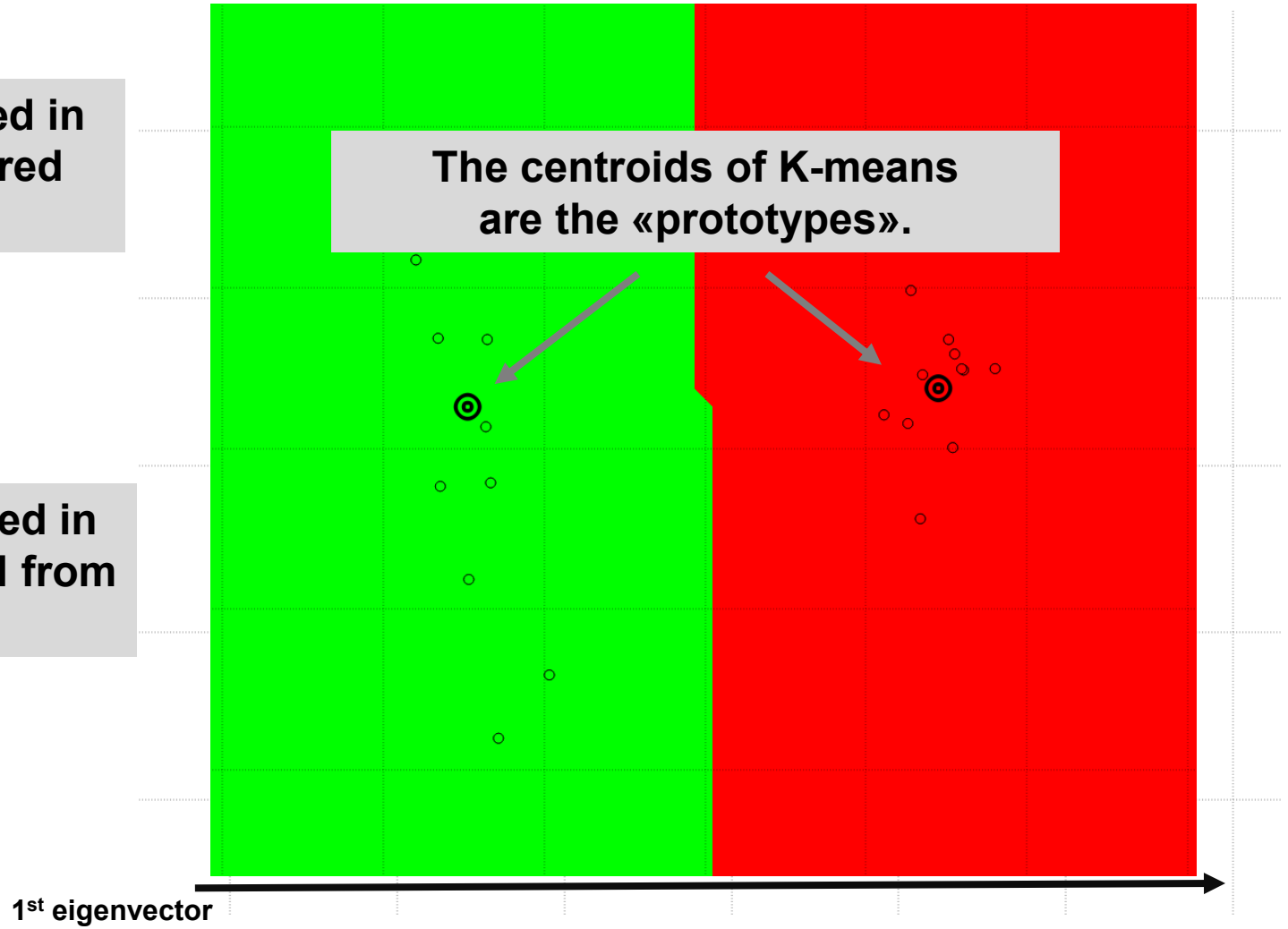
The centroids of K-means are the «prototypes».

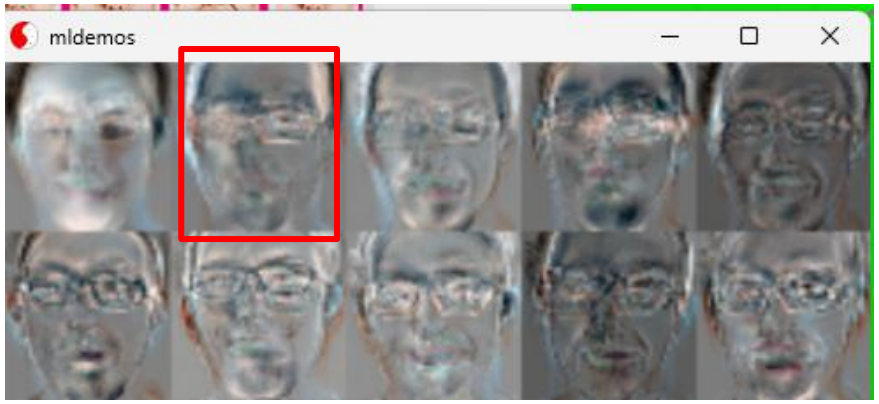
The projection on 1st eigenvector separate the two groups perfectly, whereas on the 2nd eigenvectors they are superimposed.

1st eigenvector

The prototypes of the two persons are mixed in the first eigenvectors. They can be recovered from running ICA on 1st eigenvector

The prototypes of the two persons are mixed in the first eigenvector. They can be recovered from running ICA on 1st eigenvector

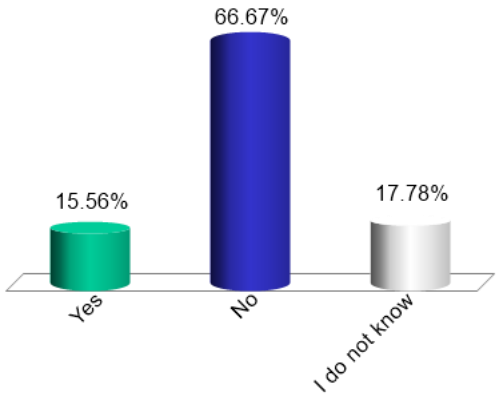


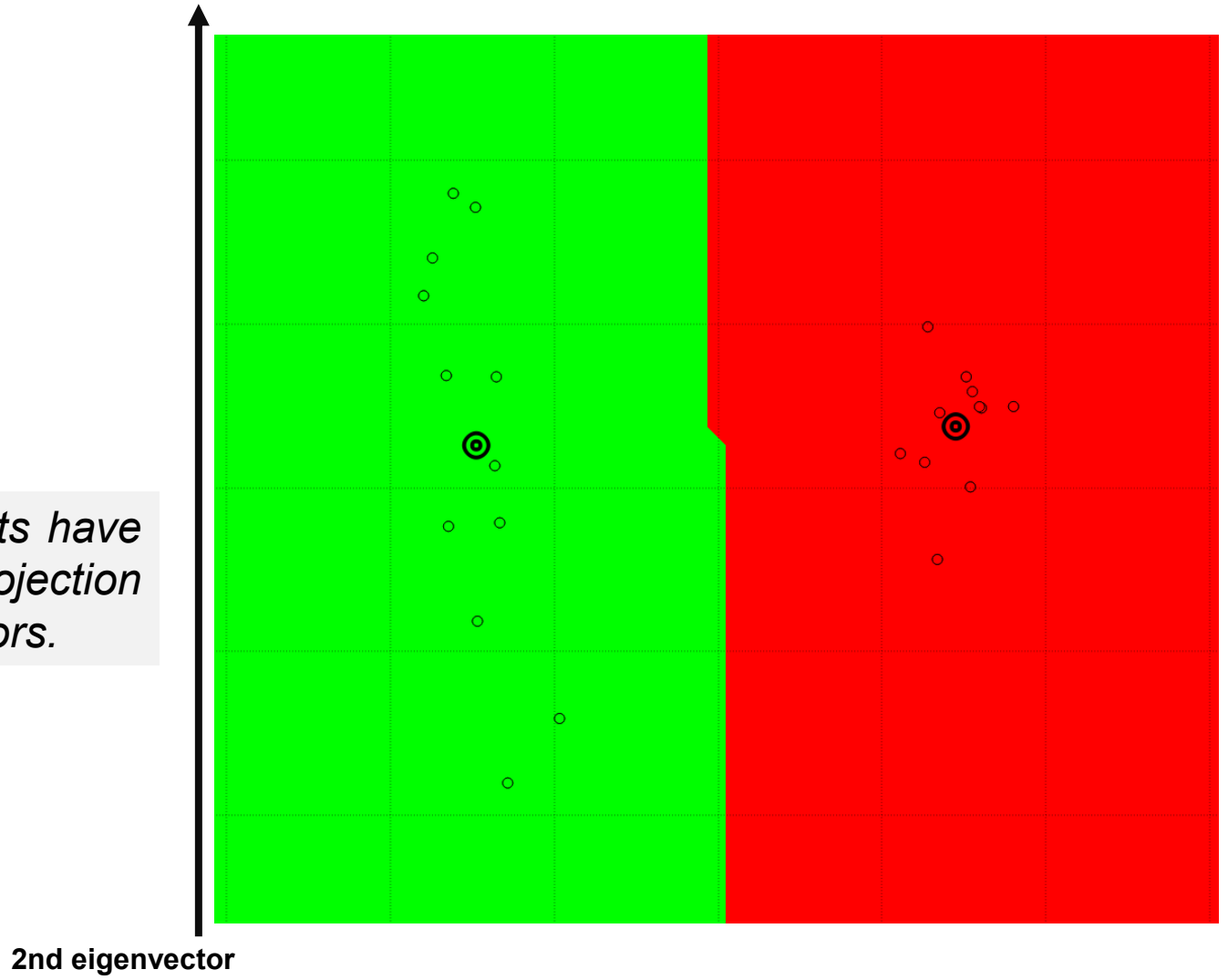


Could ICA be used on the 2nd eigenvector to recover the two groups?

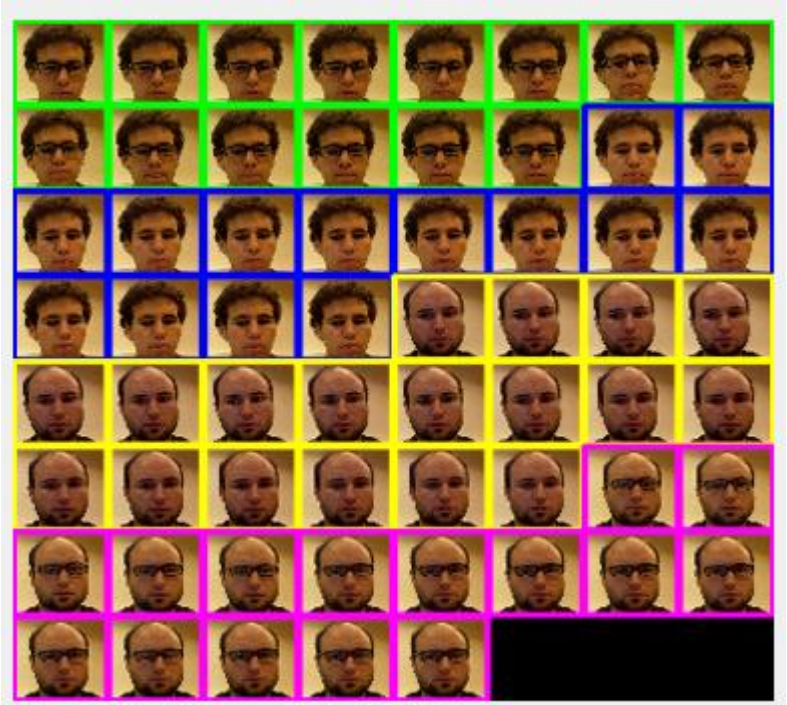
A. Yes
B. No ✓
C. I do not know

As the two datapoints have (quasi) identical projection on the 2nd eigenvectors.

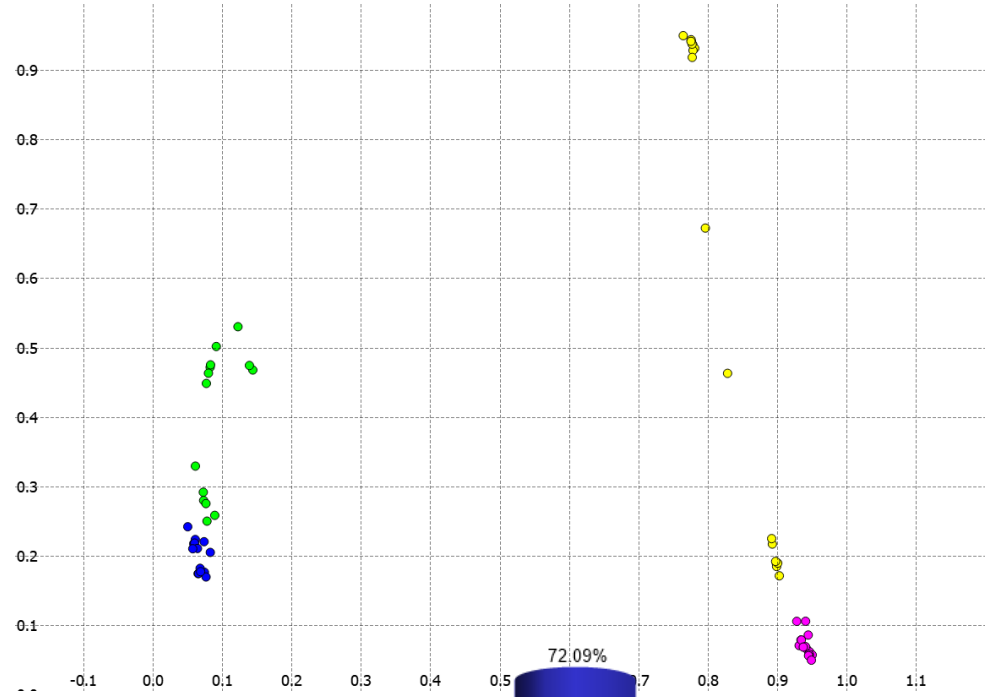




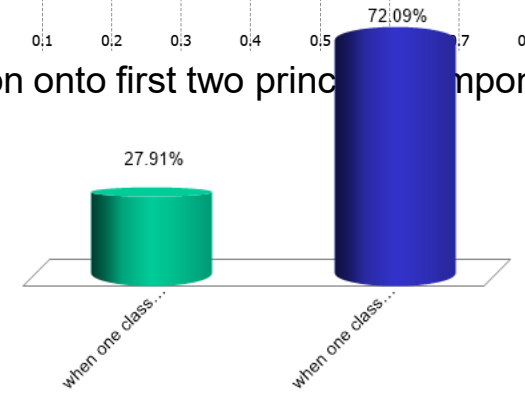
High intra-class similarity is necessary for achieving a good clustering



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



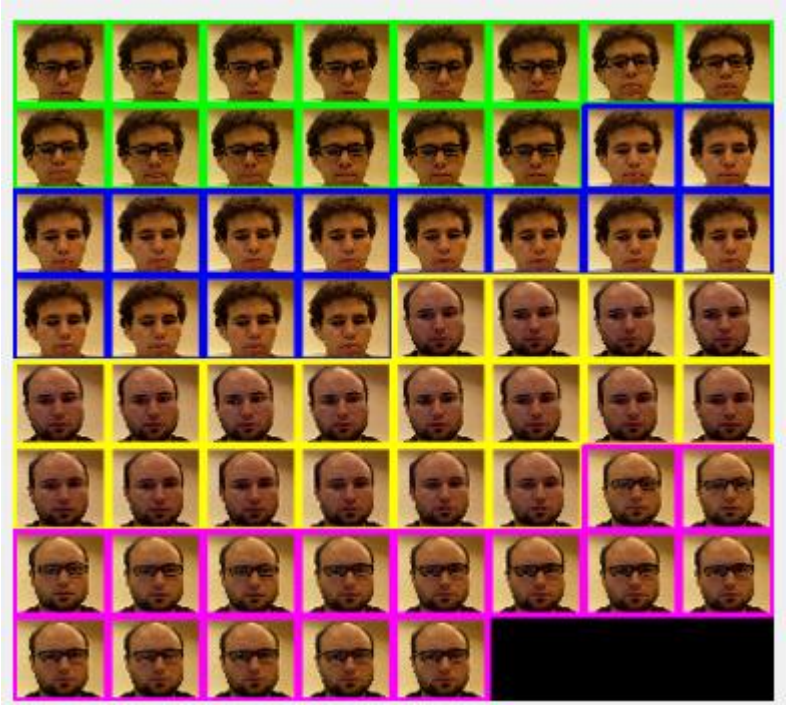
Projection onto first two princ mponents after PCA



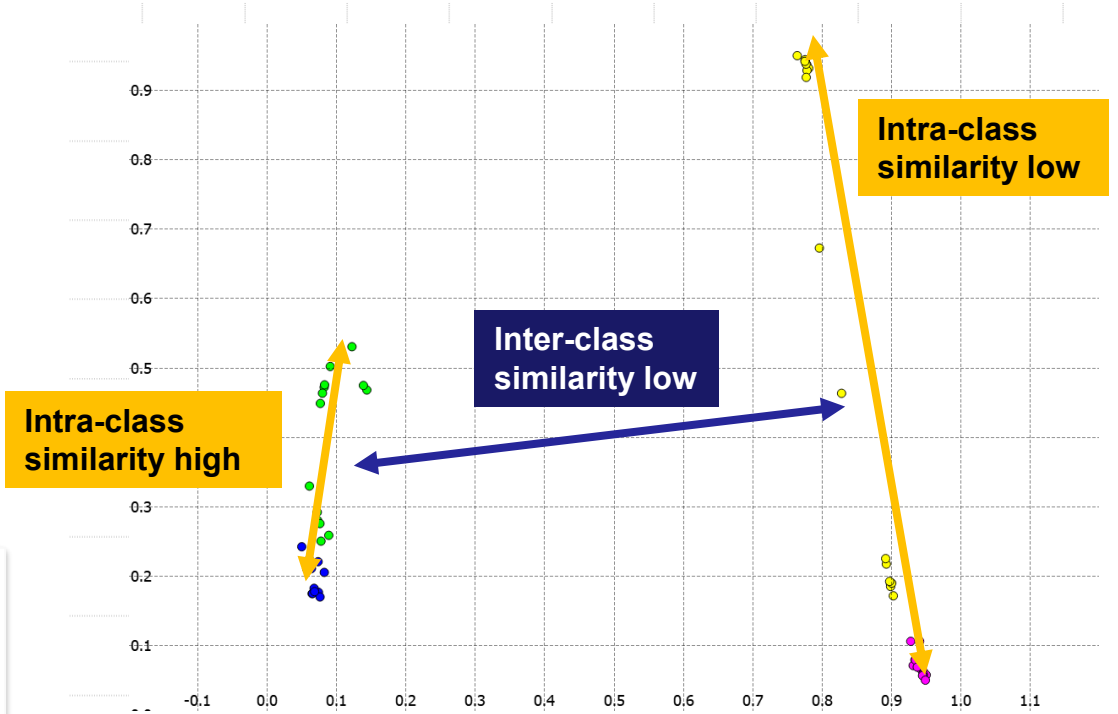
When is intra-class similarity the highest ?

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

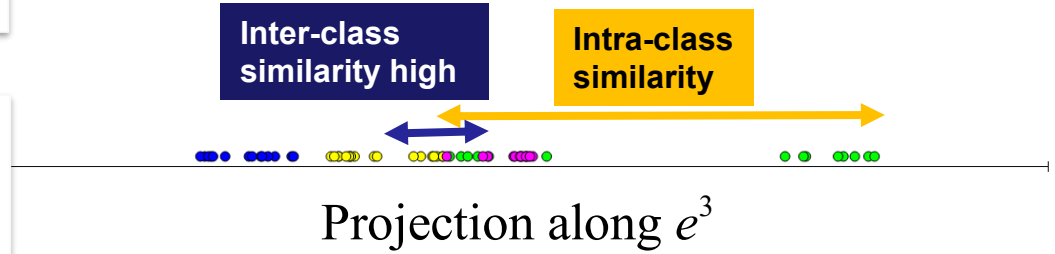
B. when one classifies images of person1 against person2.



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



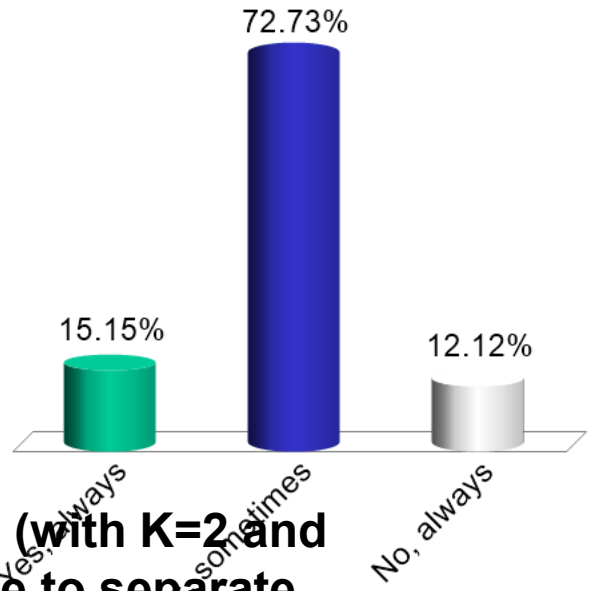
Projection onto first two principal components after PCA



Projection along e^3

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

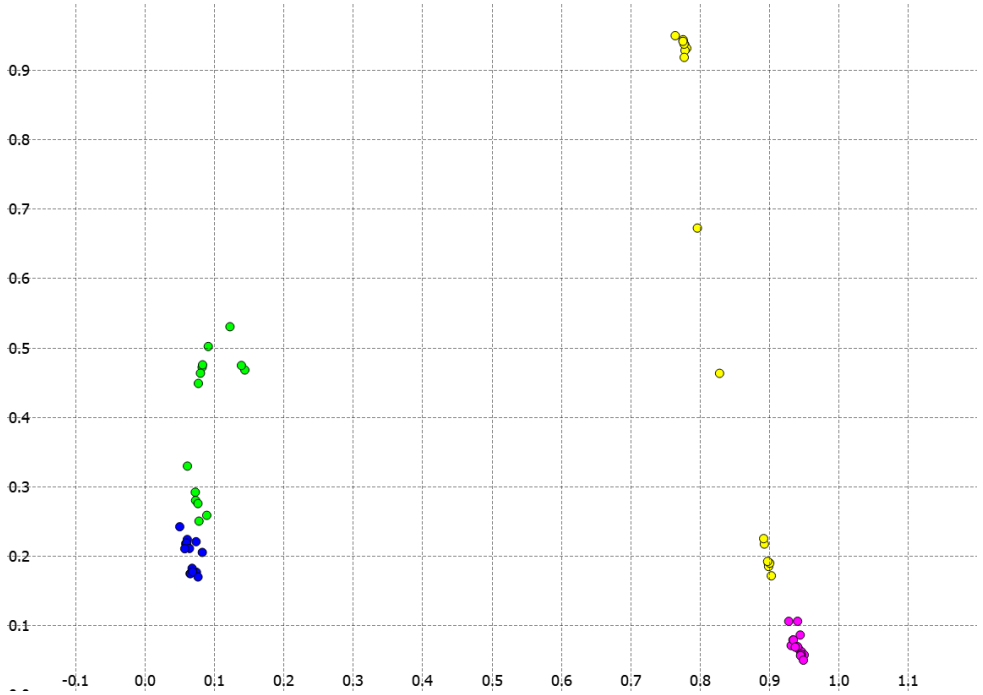
Intra-class similarity is low when classifying images of people with glasses vs images of people without glasses, especially for person2.



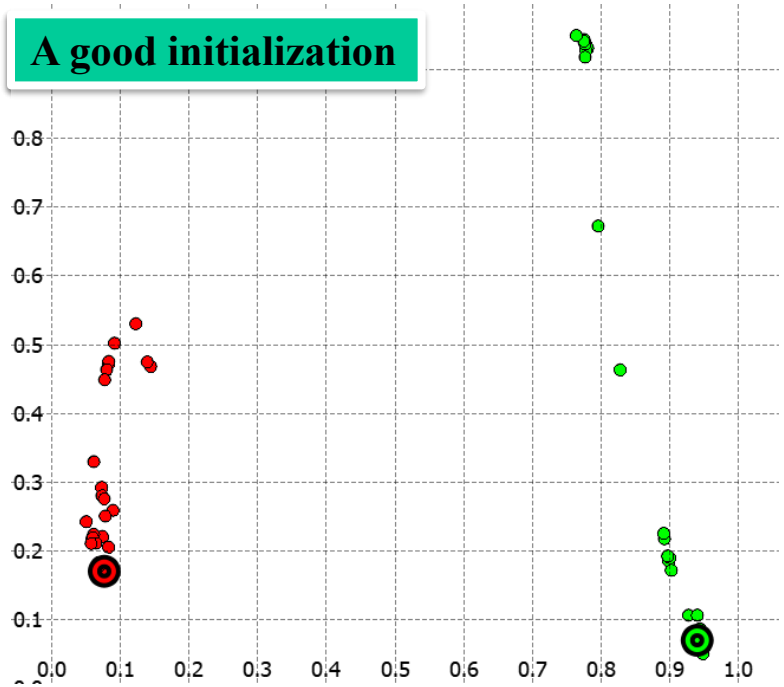
Would K-means (with K=2 and norm-2) be able to separate the two persons correctly?

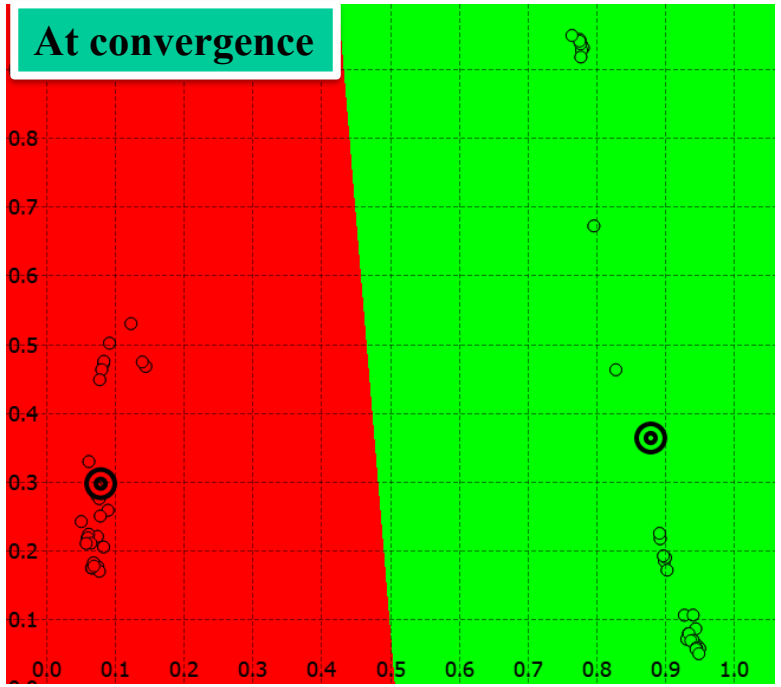
A. Yes, always
B. Yes, sometimes
C. No, always

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



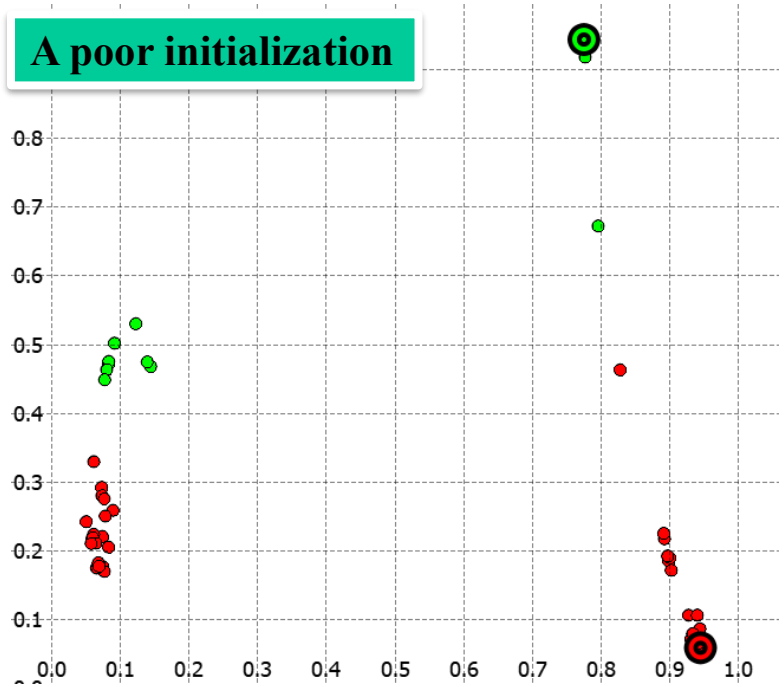
Projection onto first two principal components after PCA

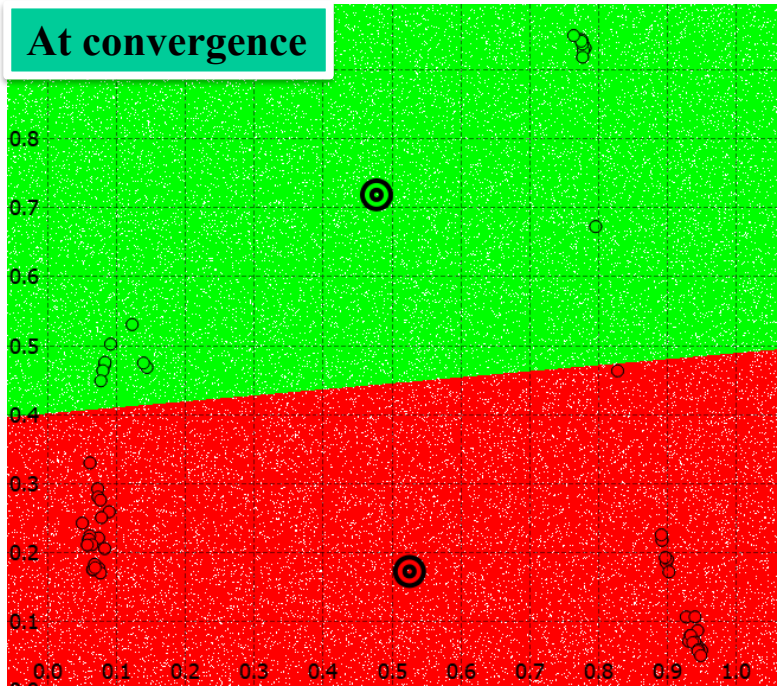




Would K-means (with K=2 and norm-2) be able to separate the two persons correctly?

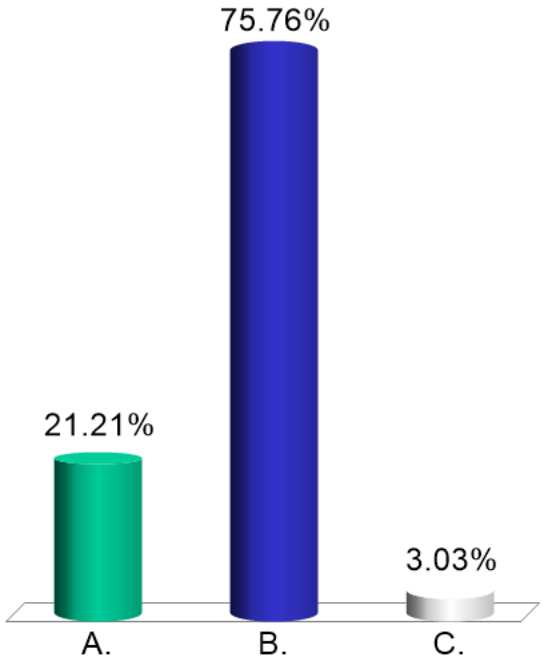
Yes, sometimes



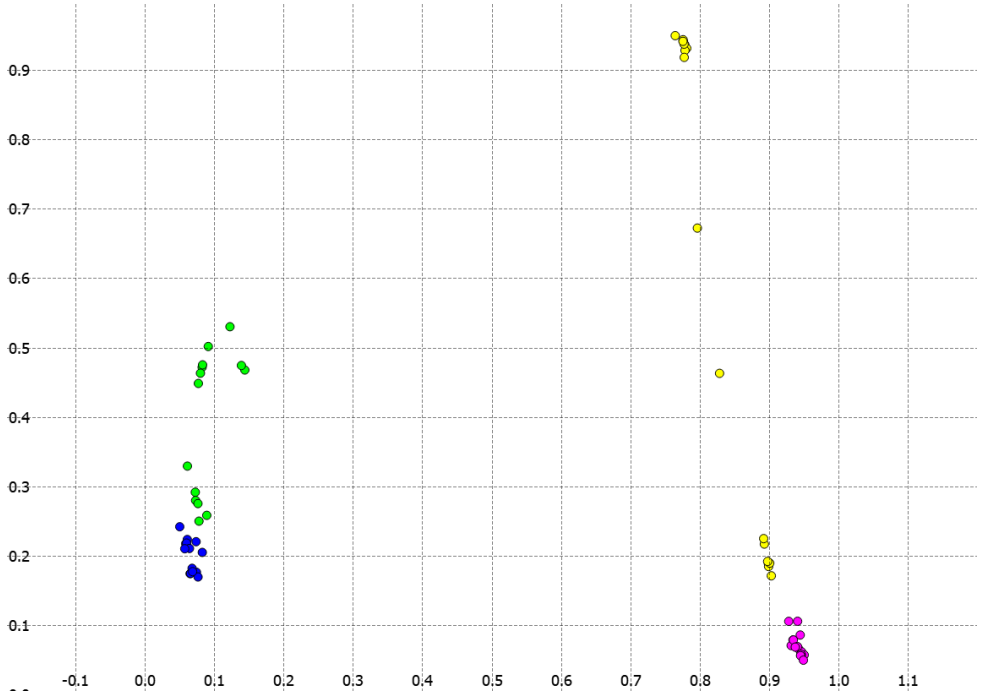


Would K-means (with K=2 and norm-2) be able to separate the two persons correctly?

Yes, sometimes

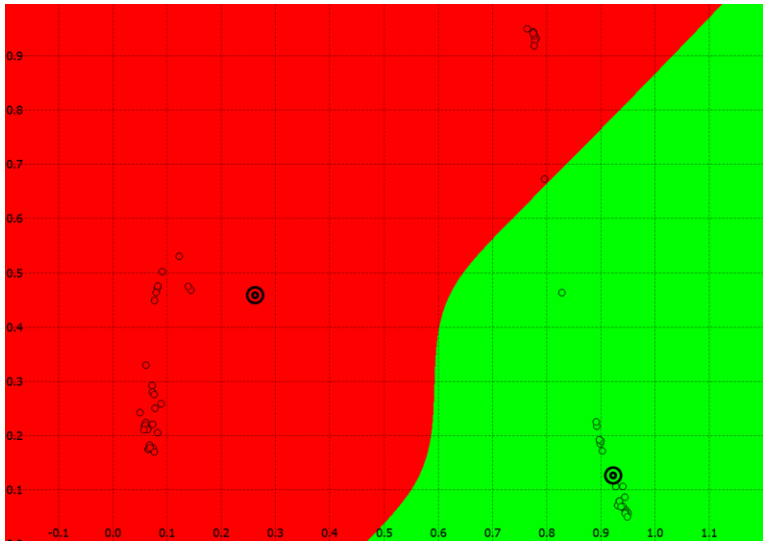


- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses

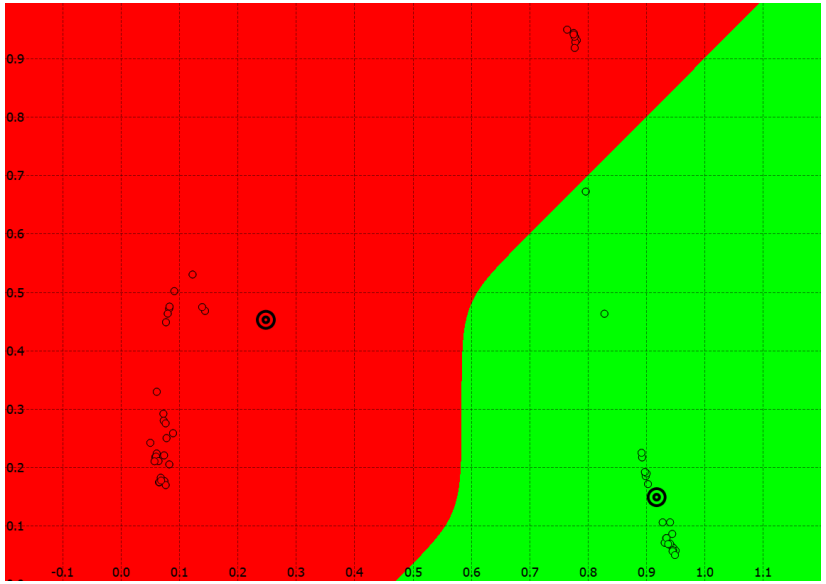


Would K-means (K=2) with other L-p norms be able to separate the two persons correctly always?

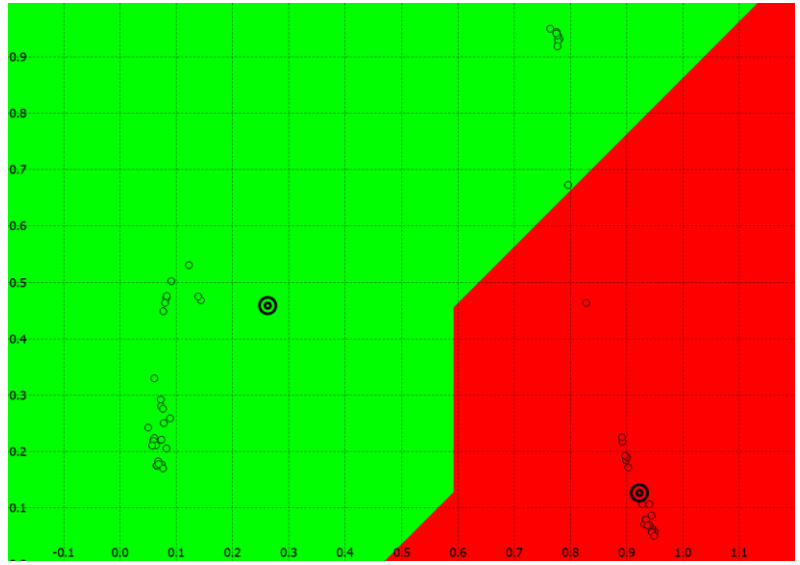
A. Yes
B. No
C. I do not know



P=6



P=9

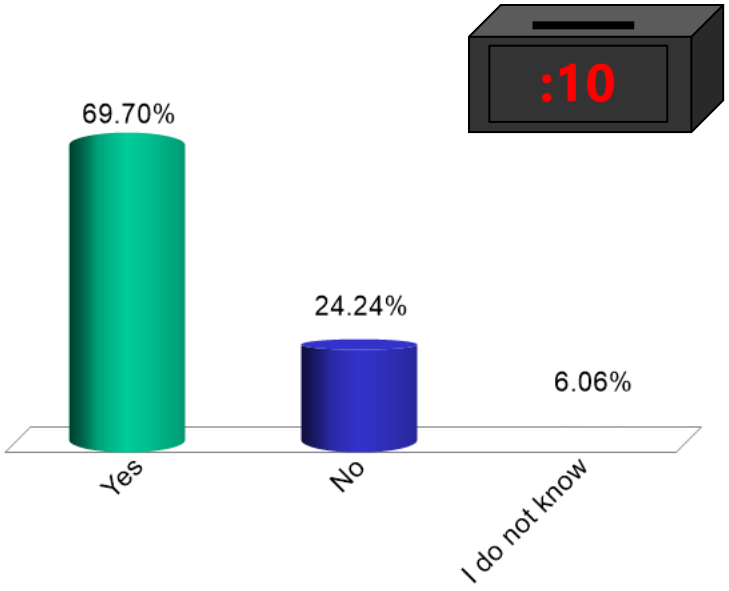


L-inf

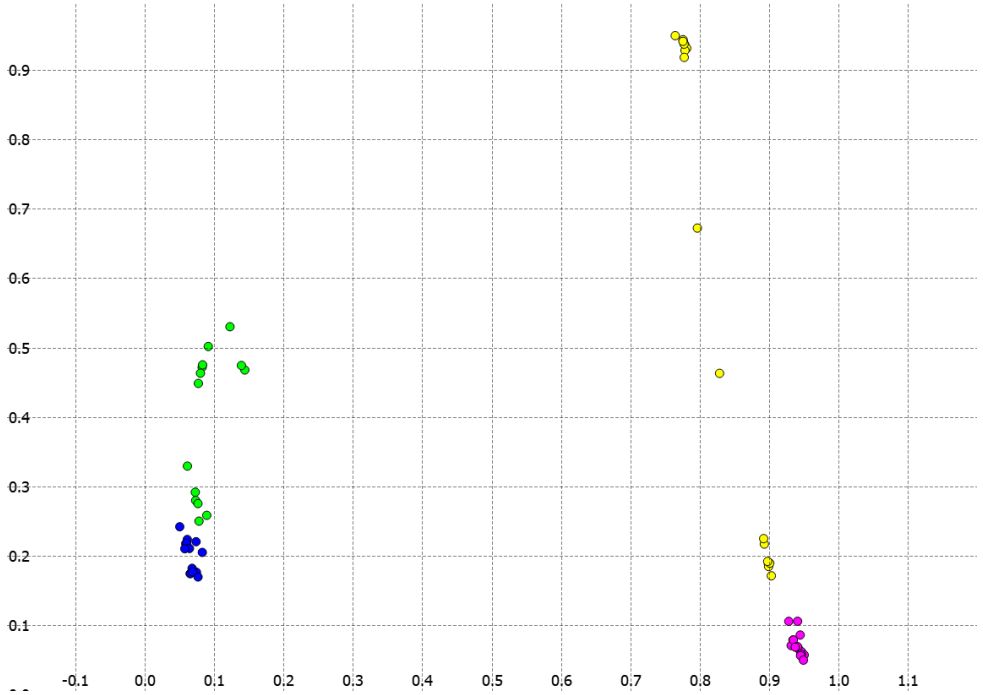
Would K-means (K=2) with other L-p norms be able to separate the two persons correctly always?

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.



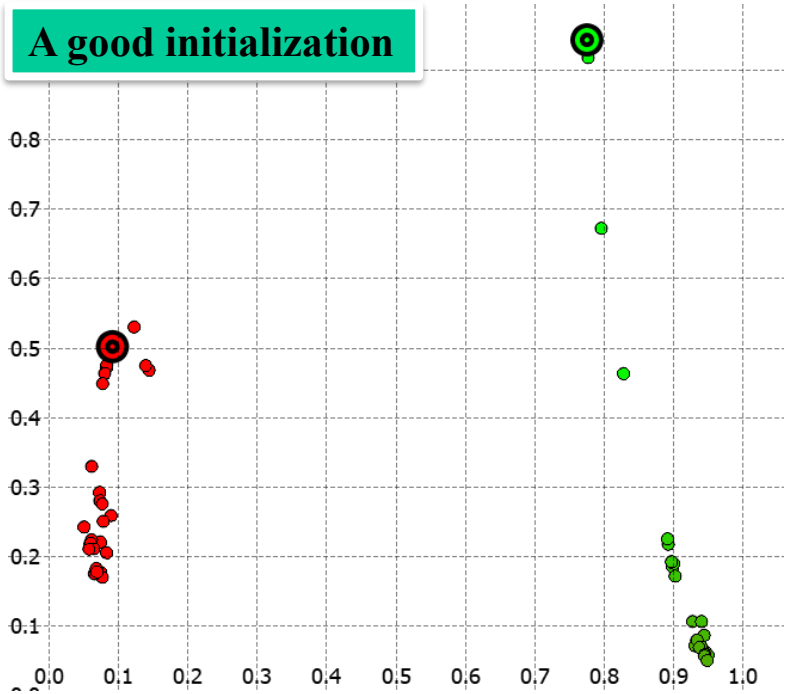
- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses

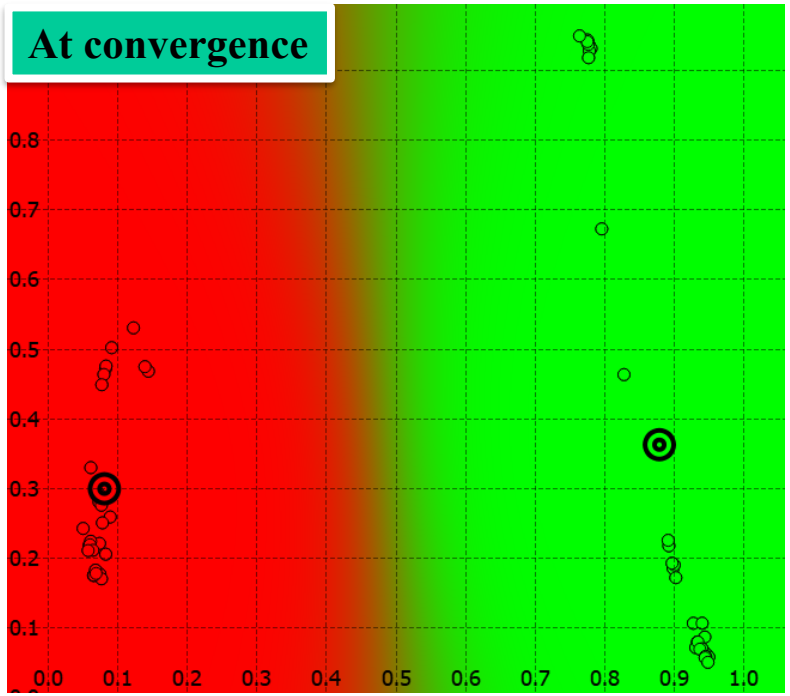


Would soft K-means be able to separate the two persons correctly with a large β for a good initialization ?

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

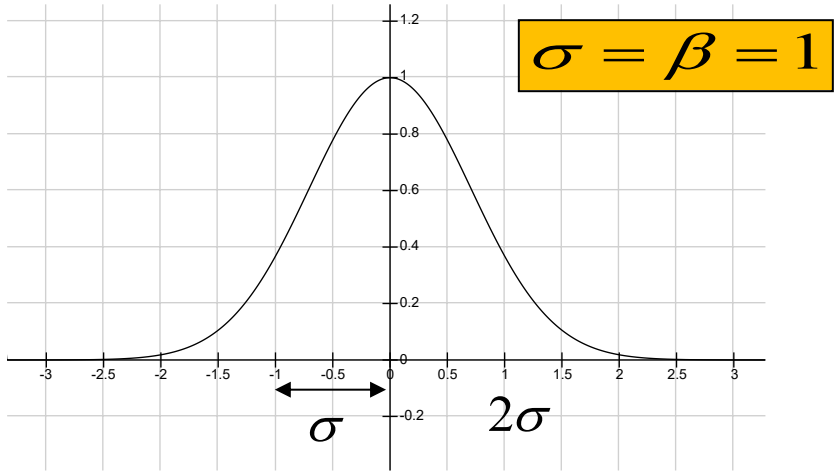
$$r_i^k = \frac{e^{\left(-\beta \cdot d\left(\mu^k, x^i\right)\right)}}{\sum_{k'} e^{\left(-\beta \cdot d\left(\mu^{k'}, x^i\right)\right)}}$$

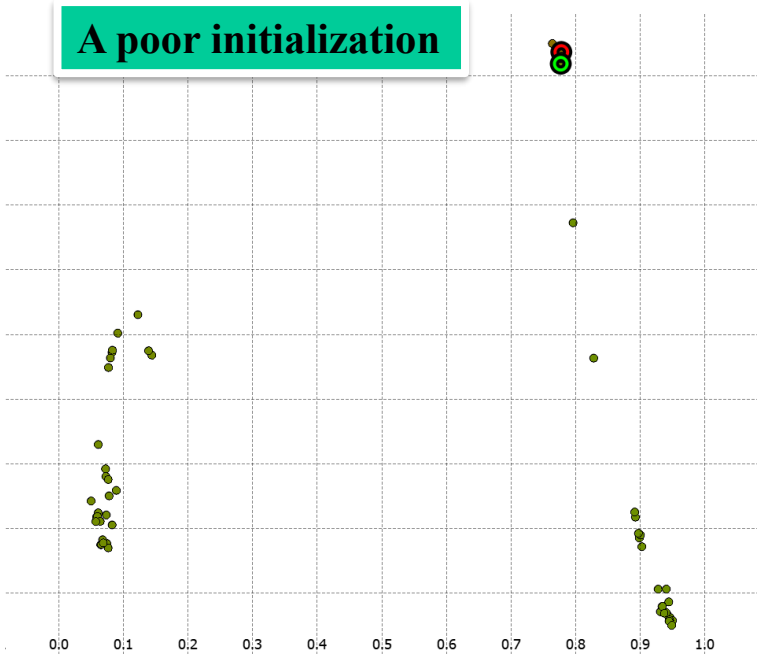


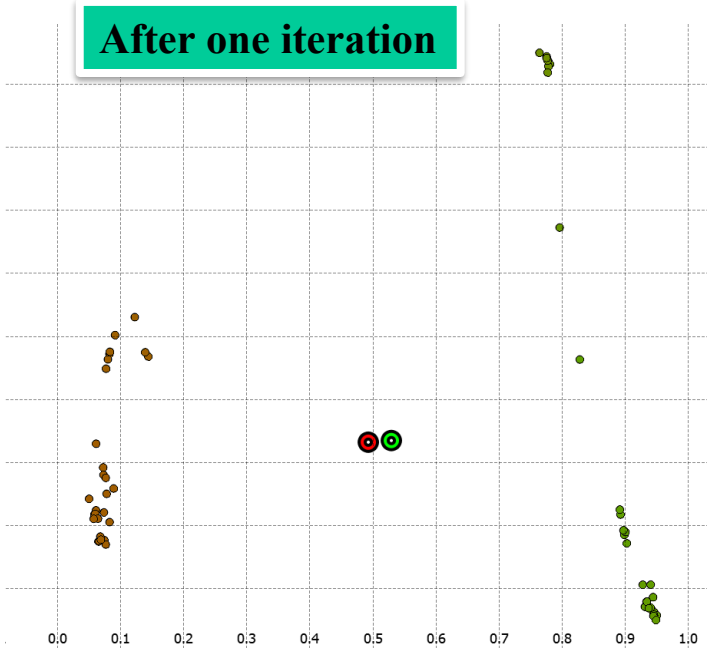


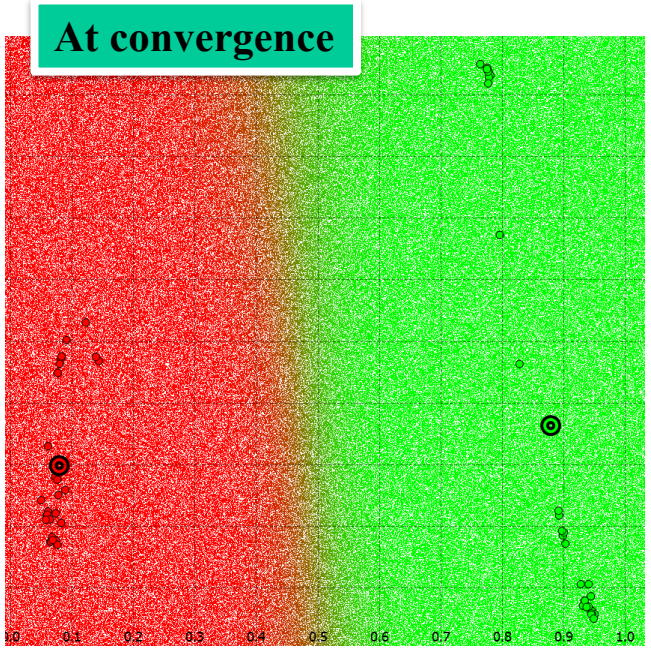
Would soft K-means be able to separate the two persons correctly with a large β for a good initialization?

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





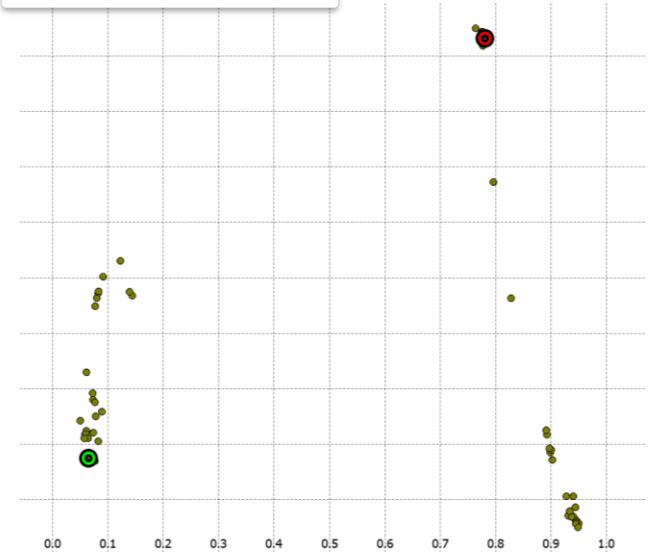




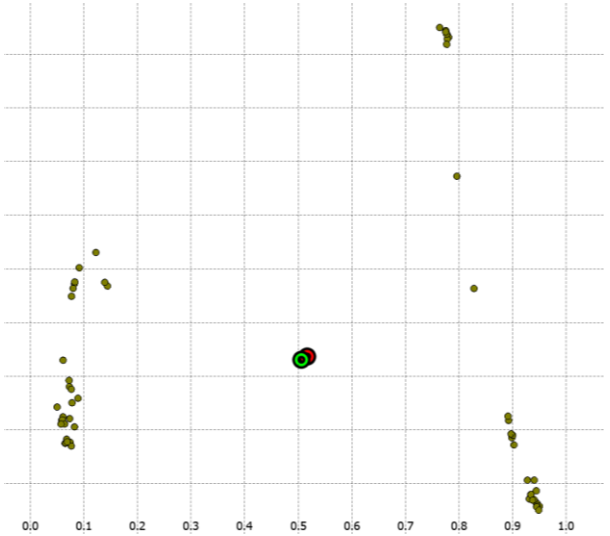
Would soft K-means be able to separate the two persons correctly with a large β with a poor initialization of the centroids?

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.

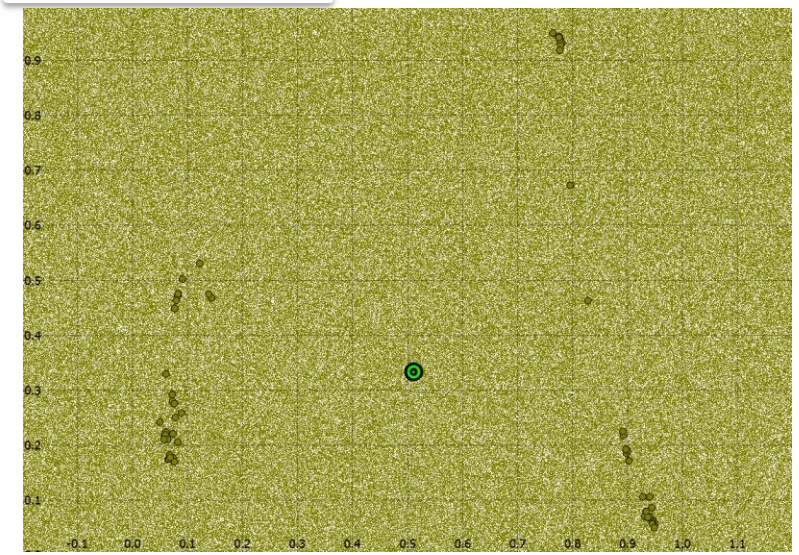
At initialization



After one iteration

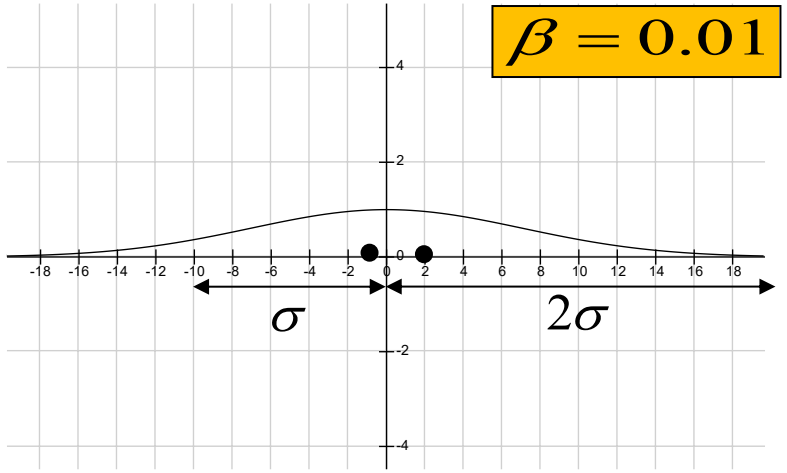


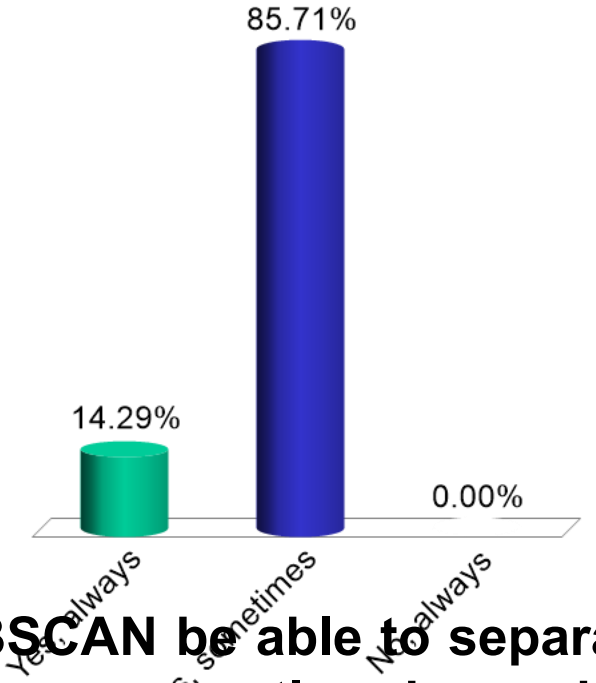
At convergence



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.

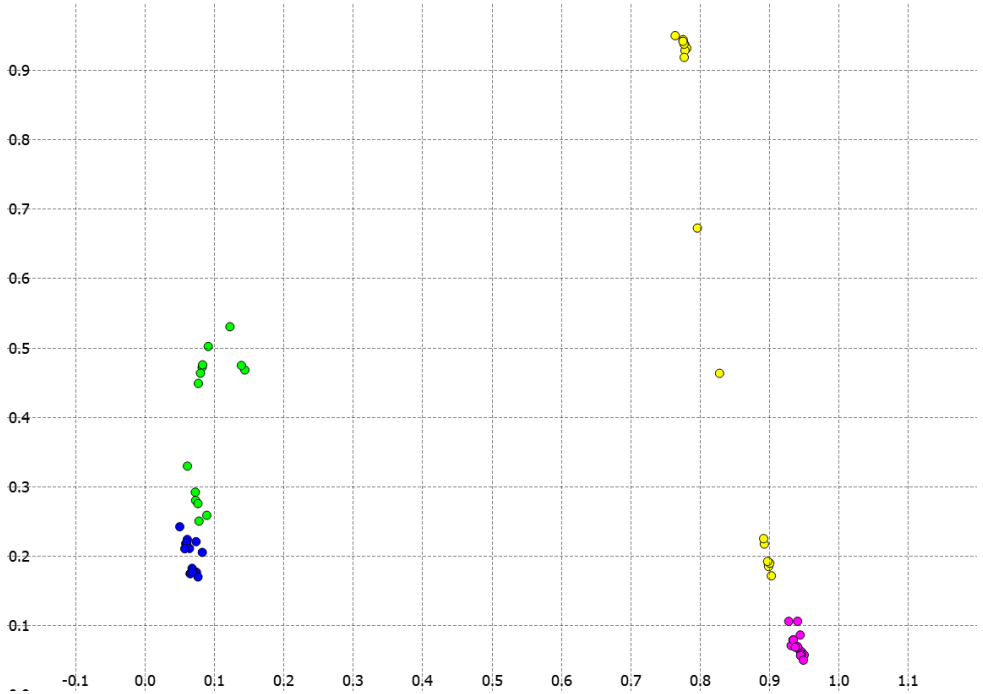




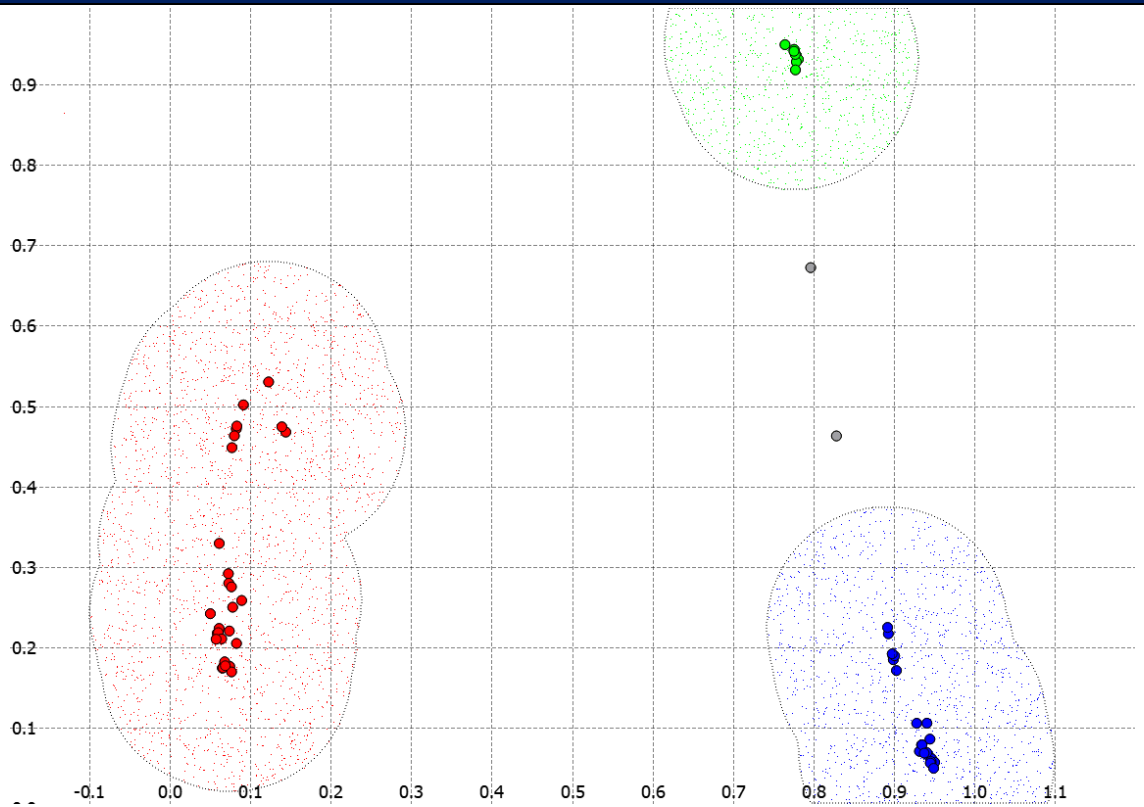
Would DBSCAN be able to separate the two persons correctly when minimum number of data points in cluster is 2 and for any radius ε?

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

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

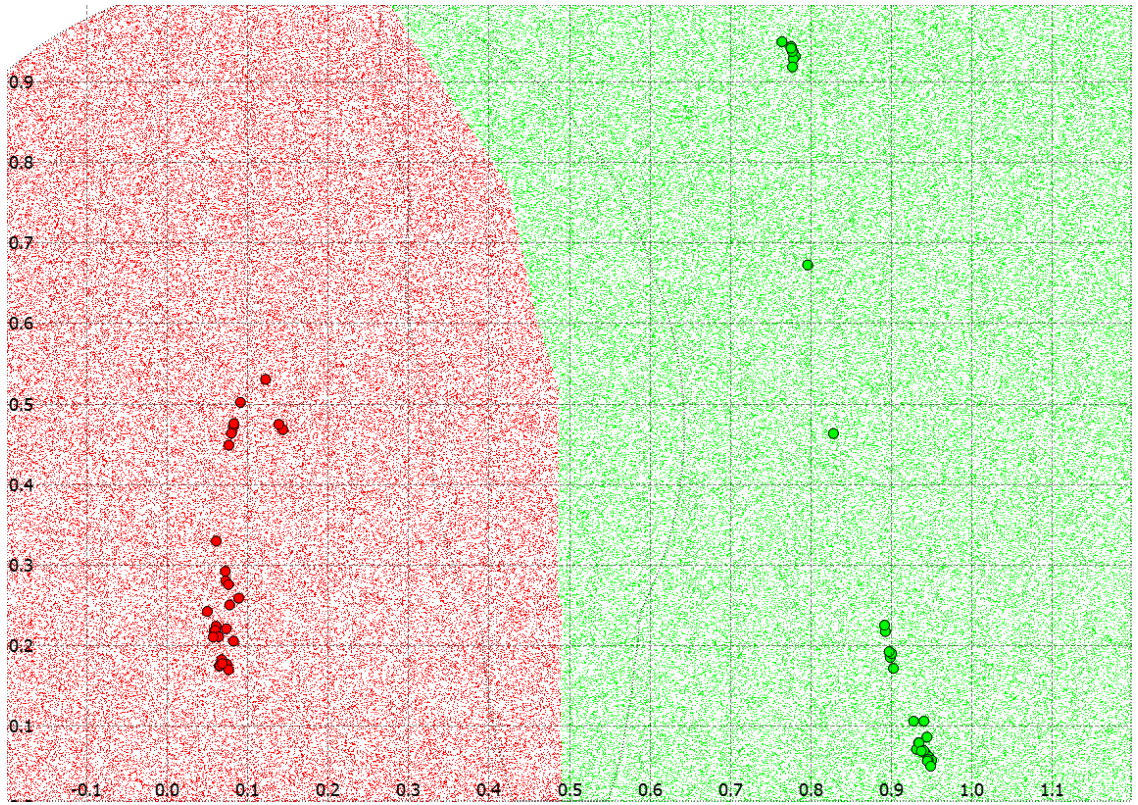


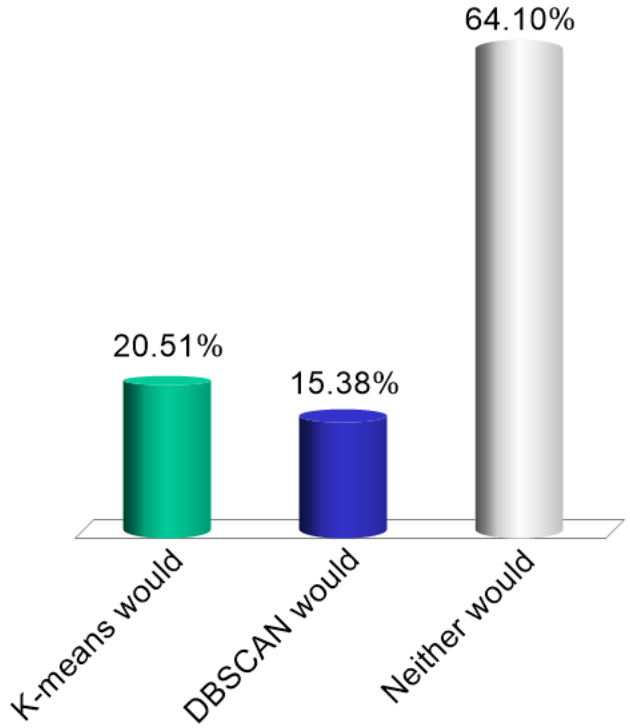
Projection onto first two principal components after PCA



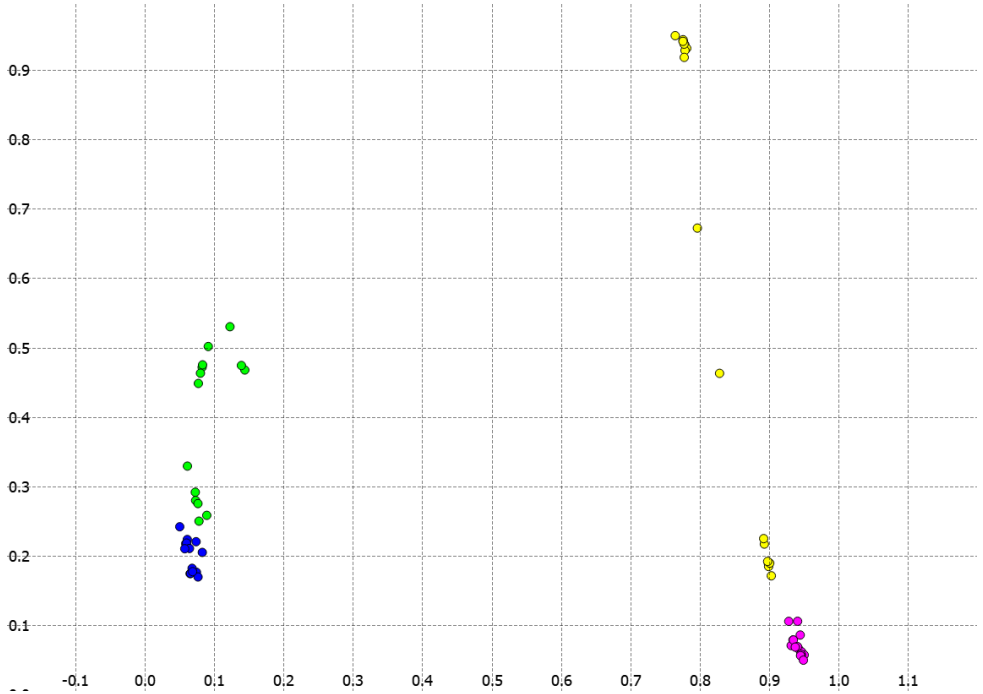
No, when the ε is too small

yes, when the ε is large enough



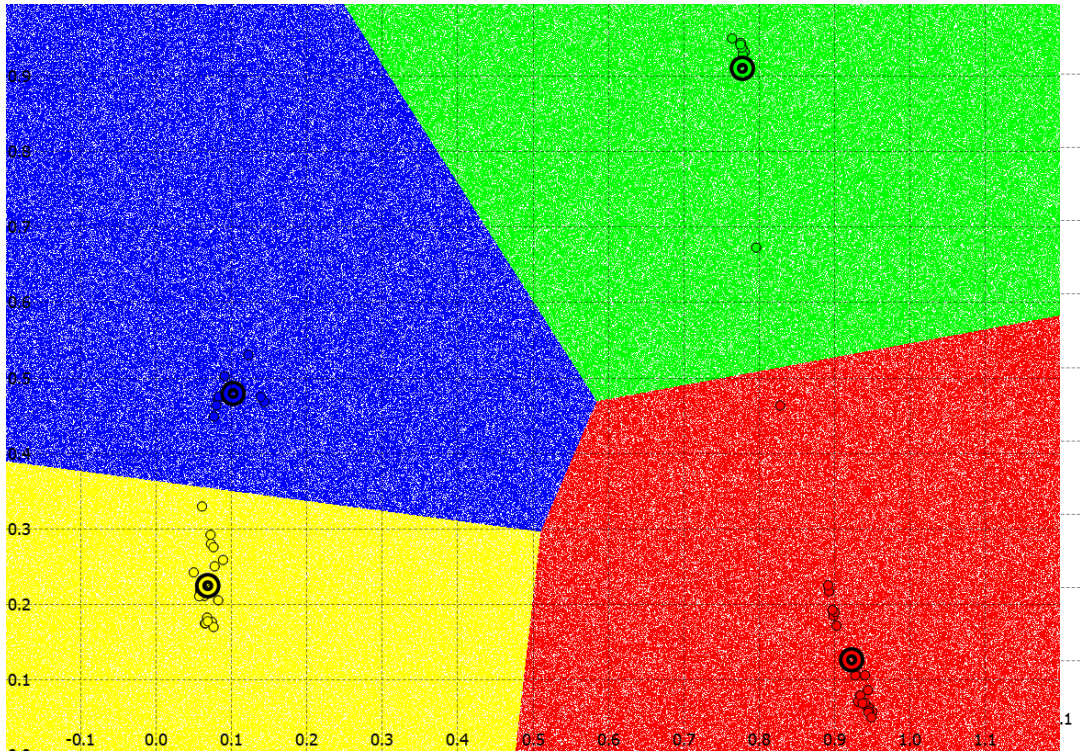


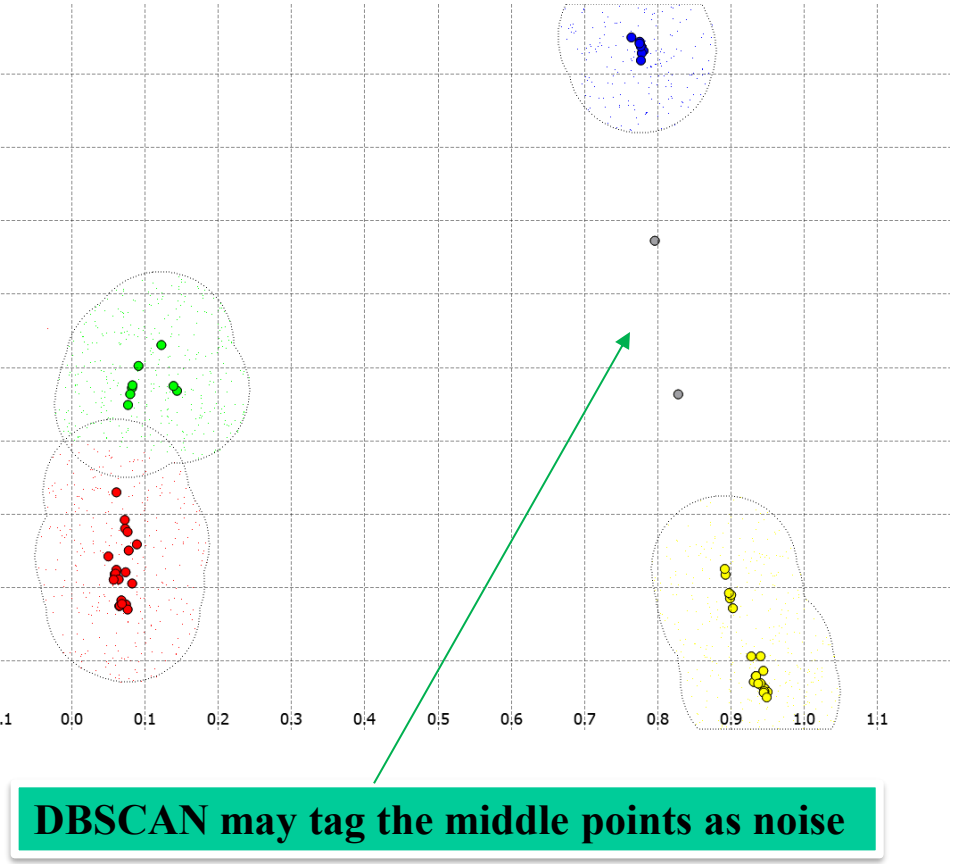
- Person1 with glasses
- Person1 without glasses
- Person2 without glasses
- Person2 with glasses



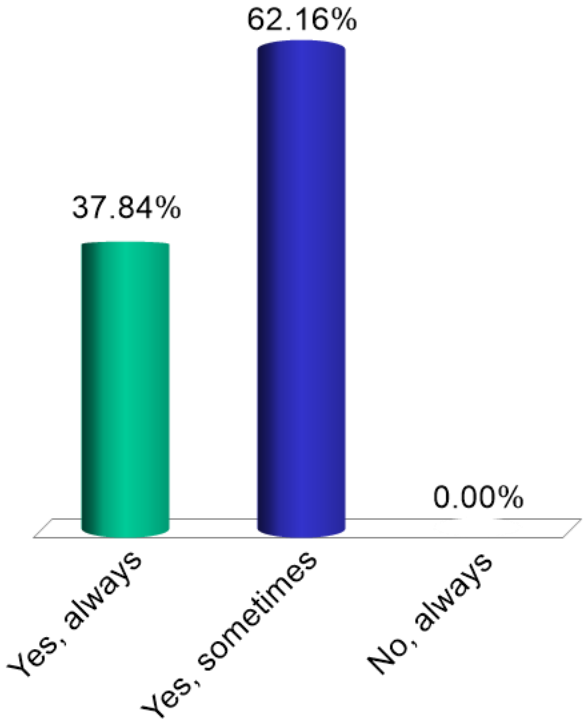
Would K-means or DBSCAN be able to separate the 4 classes correctly?

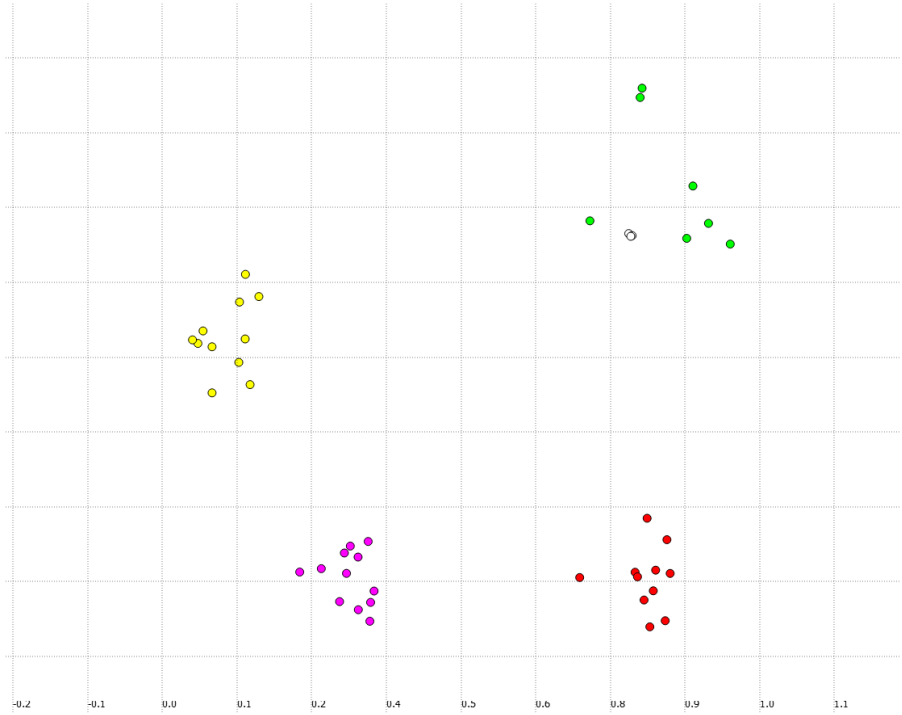
A. K-means would
B. DBSCAN would
C. Neither would





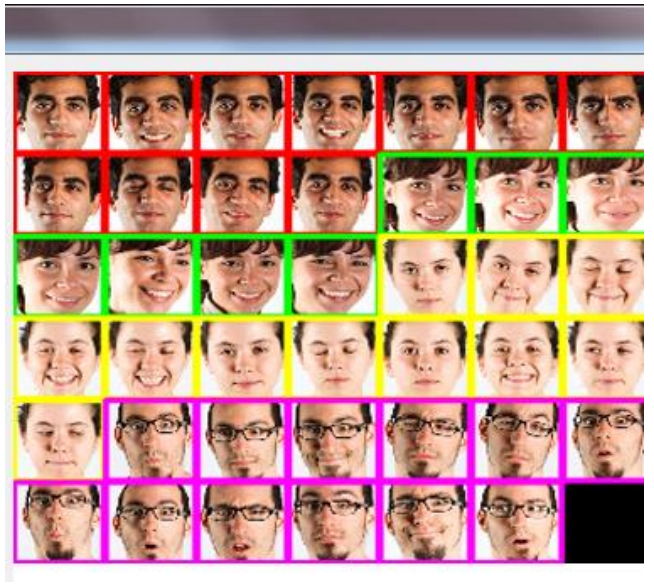
Neither of the two methods can cluster the 4 clusters correctly as the distance within clusters is bigger than across clusters for the glasses/no-glasses groups.

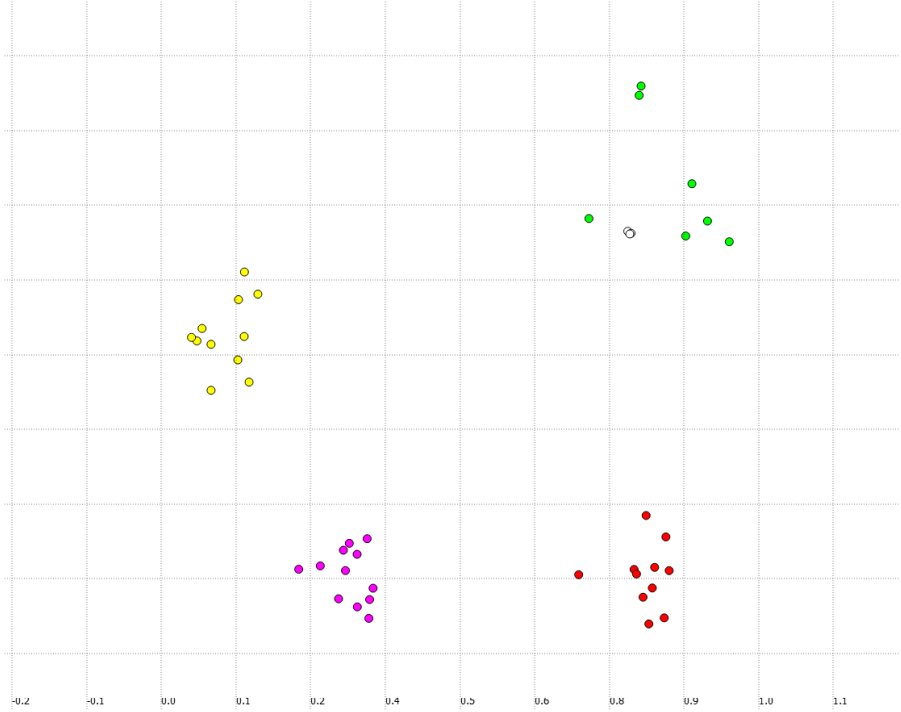




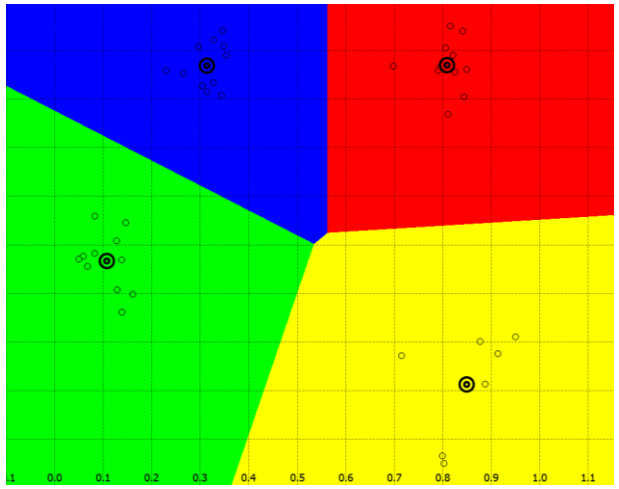
Would K-means be able to separate the 4 classes correctly?

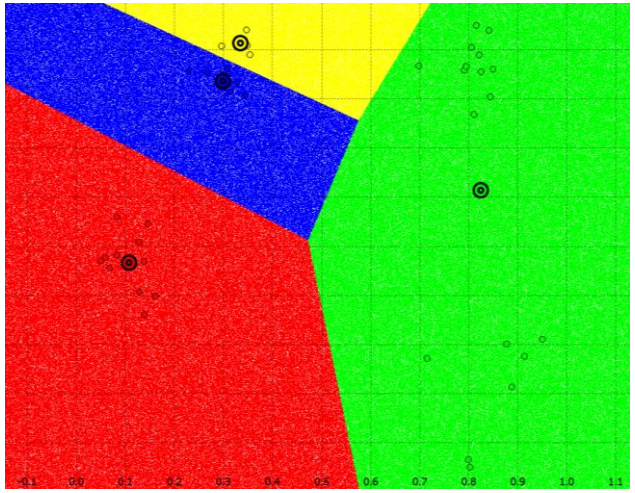
A. Yes, always
B. Yes, sometimes
C. No, always

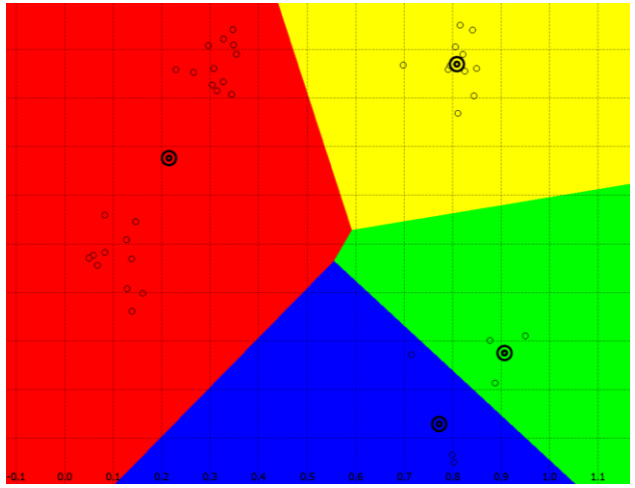


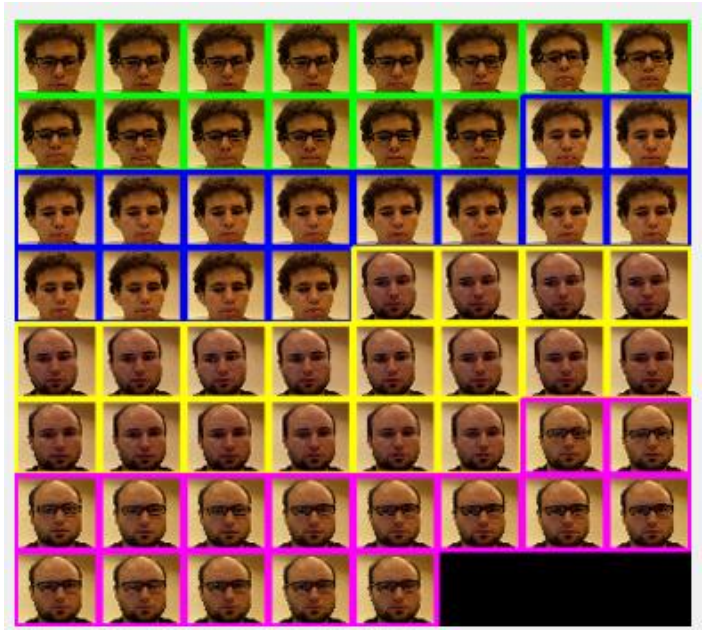


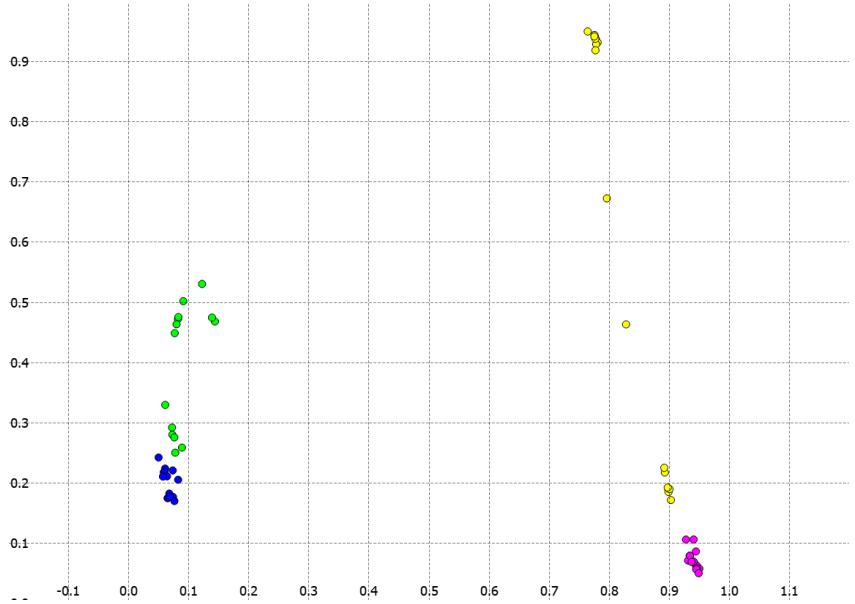
Three solutions, depending on initialization





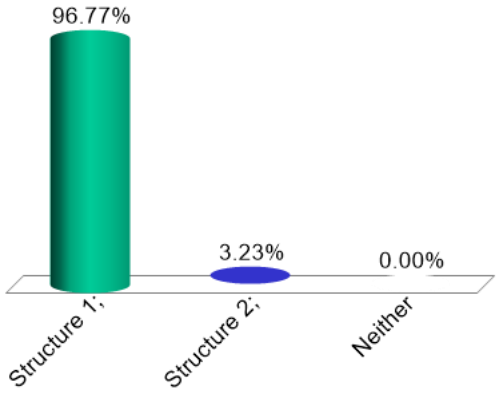






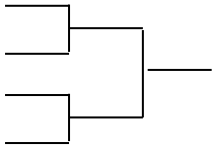
Which of the following structure would hierarchical clustering lead to when using Euclidean distance?

A. Structure 1;
B. Structure 2;
C. Neither ✓



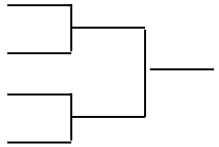
Structure 1

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



Structure 2

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



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

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

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

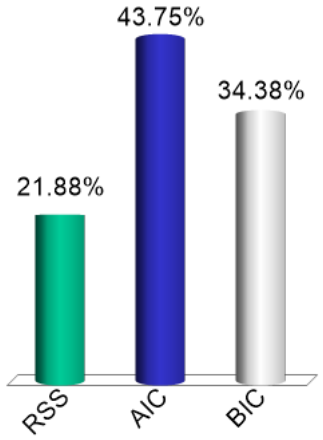
Number of free parameters *B=(K*N)*
K: # clusters
N: # dimensions

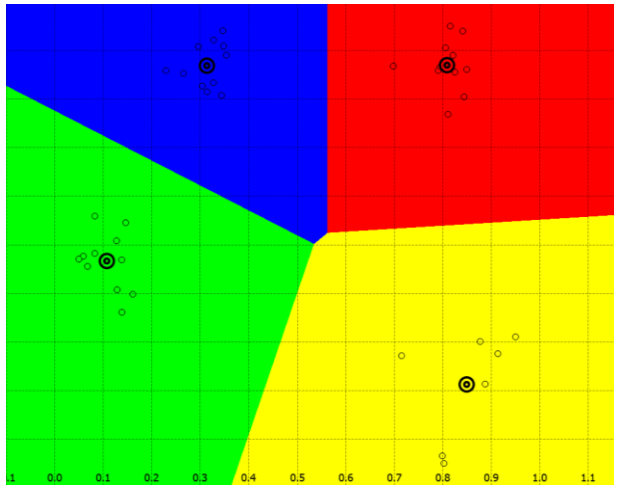
$$\text{RSS} = \sum_{k=1}^{K} \sum_{x \in C_k} \left| x - \mu^k \right|^2$$

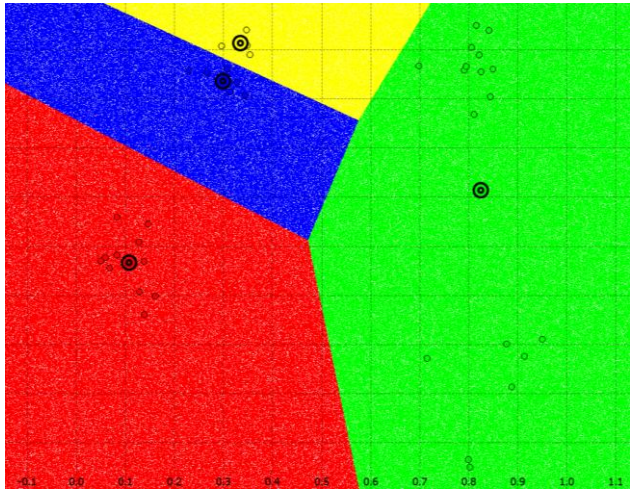
$$BIC_{RSS} = RSS + \ln(M)\, B$$

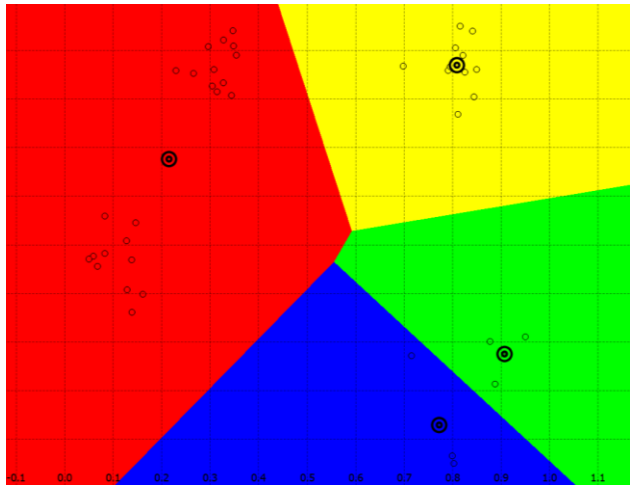
Which of the three metrics (AIC, BIC and RSS) would be most informative to determine the best solution across the 3 solutions below?

A. RSS
B. AIC
C. BIC







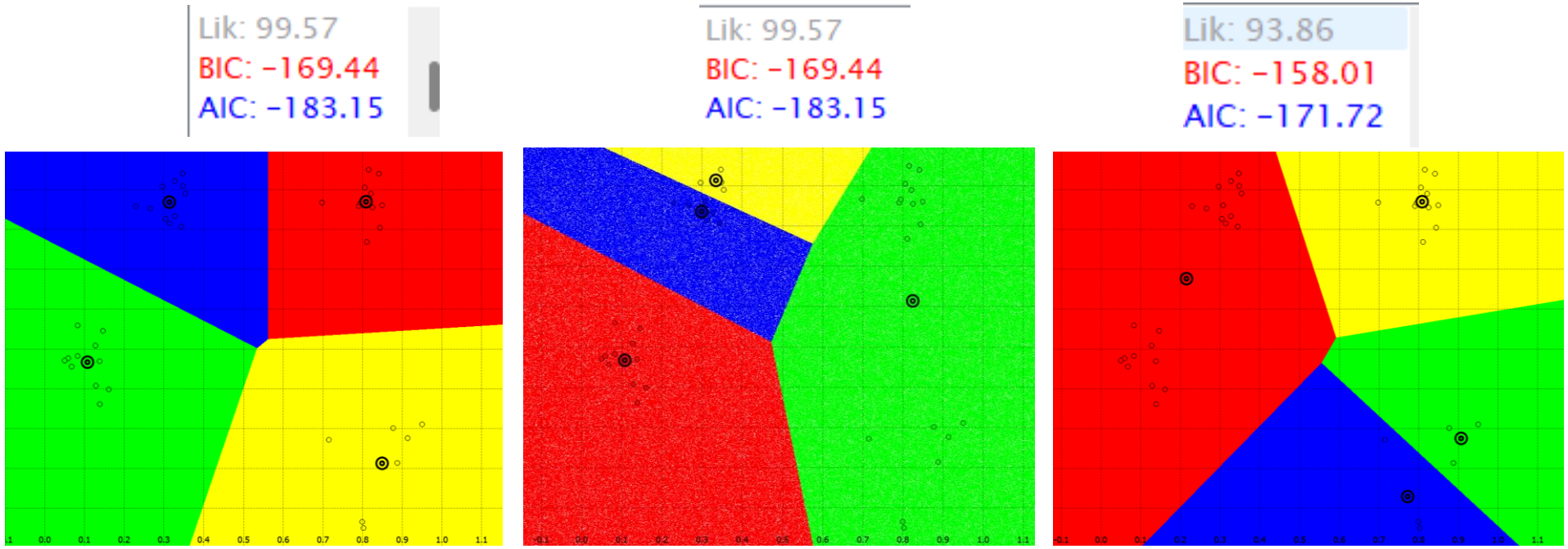


As the number of parameters (here K) remains the same, the BIC and AIC measures are affected only by the RSS measure. All three metrics will hence convey the same information.

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

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

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



RSS for K-Means:

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

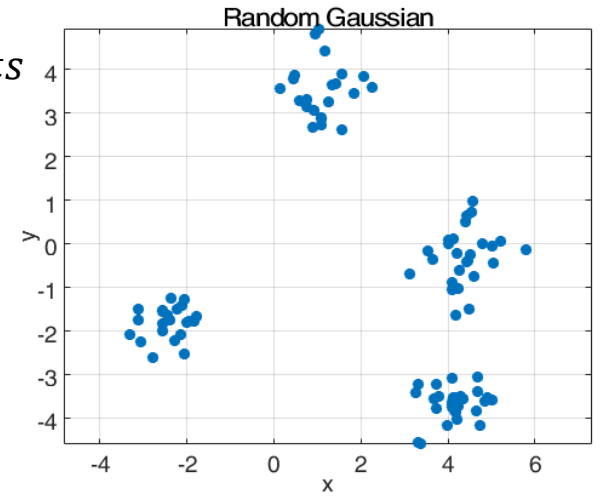
$$\text{RSS}=\sum_{k=1}^{K}\sum_{x^i \in c_k}\left\|x^i-\mu^k\right\|_2$$

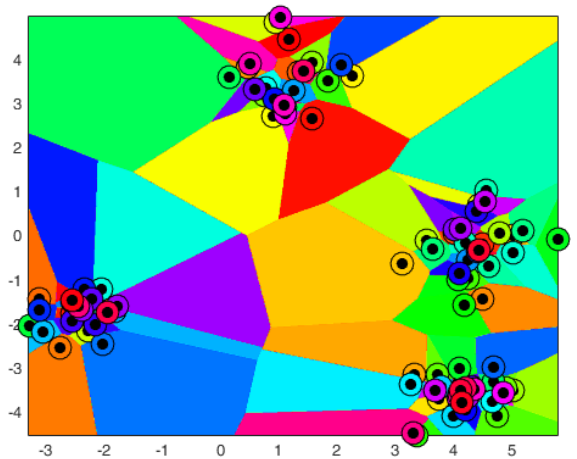
← *Measure of Distortion*

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*



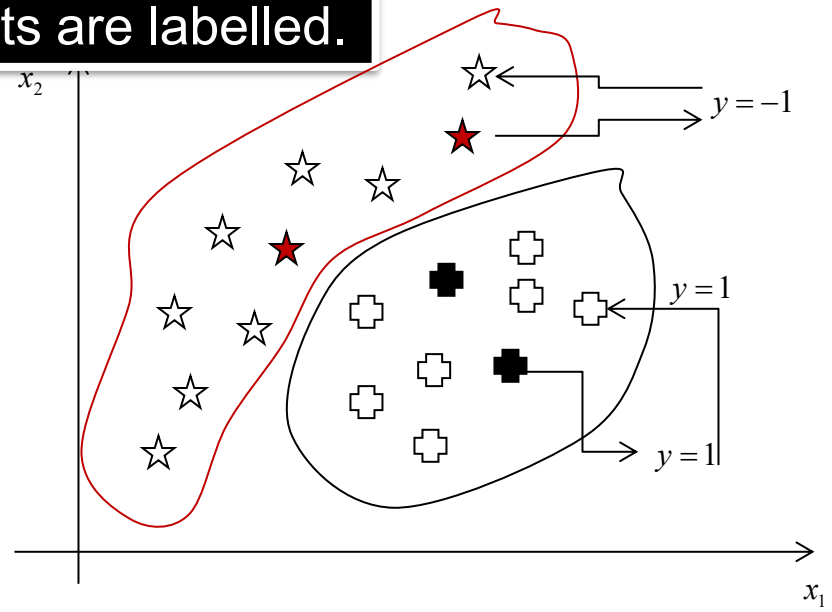


RSS: 0
K: M *clusters*

➢ 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.



Cluster 1

Cluster 2

Class 1 y=+1

Class 2 y=-1

Clustering F1-Measure :

F_1 provides a measure of how good the clustering is:

$F_1 \in [0,1]$

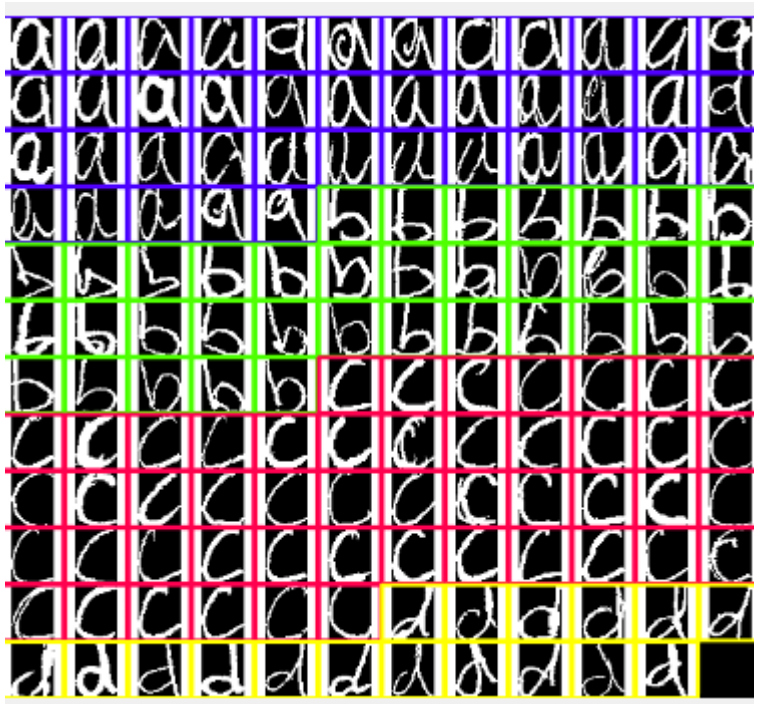
$F_1 = 1$ is the optimum.

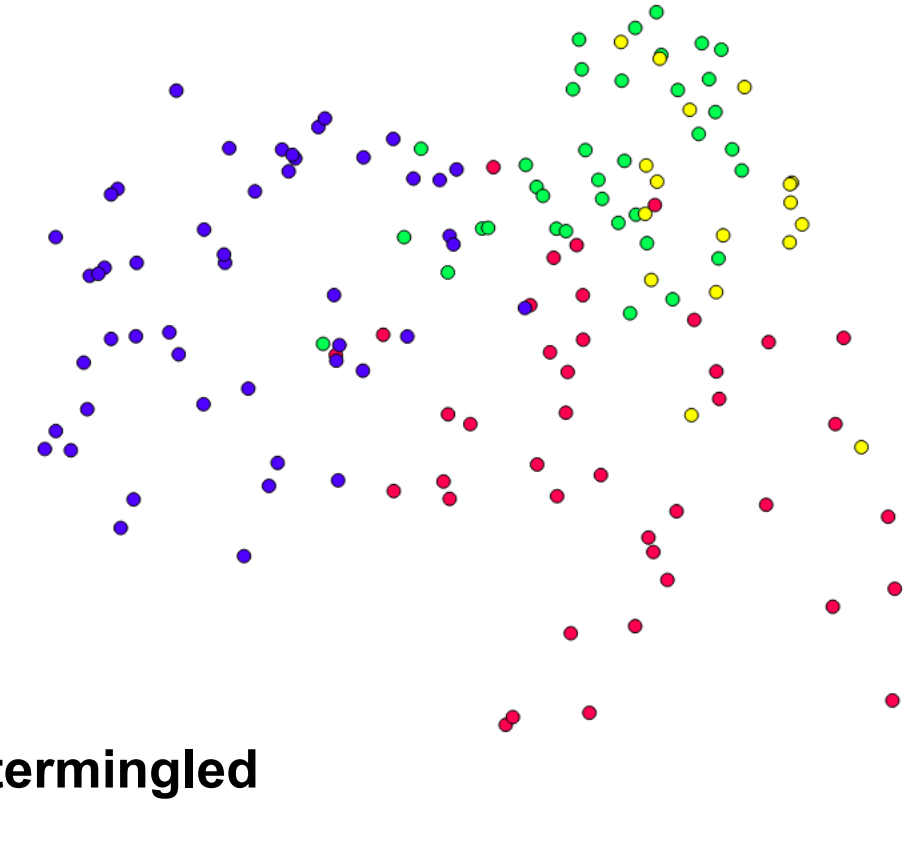
Tradeoff between clustering correctly all datapoints of the same class in the same cluster and making sure that each cluster contains points of only one class.

Semi-Supervised Clustering Principle

- When groups are not easily separable, one can use *semi-supervised* clustering.
- Semi-supervised clustering consists of labelling only a subset of the datapoints

→ number of clusters is known!





4 classes, highly intermingled

Semi-Supervised Clustering Principle

- ❑ When groups are not easily separable, one can use semi/supervised clustering
- ❑ Semi-supervised clustering consists of labelling only a subset of the datapoints

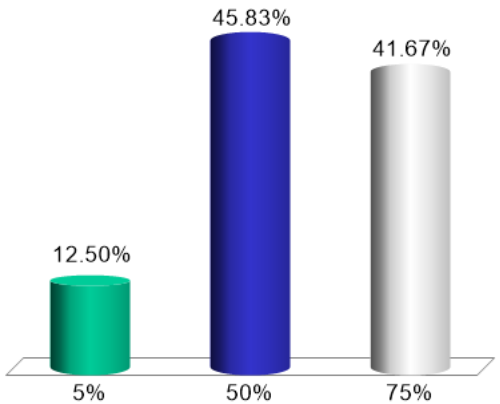
→ number of clusters is known!

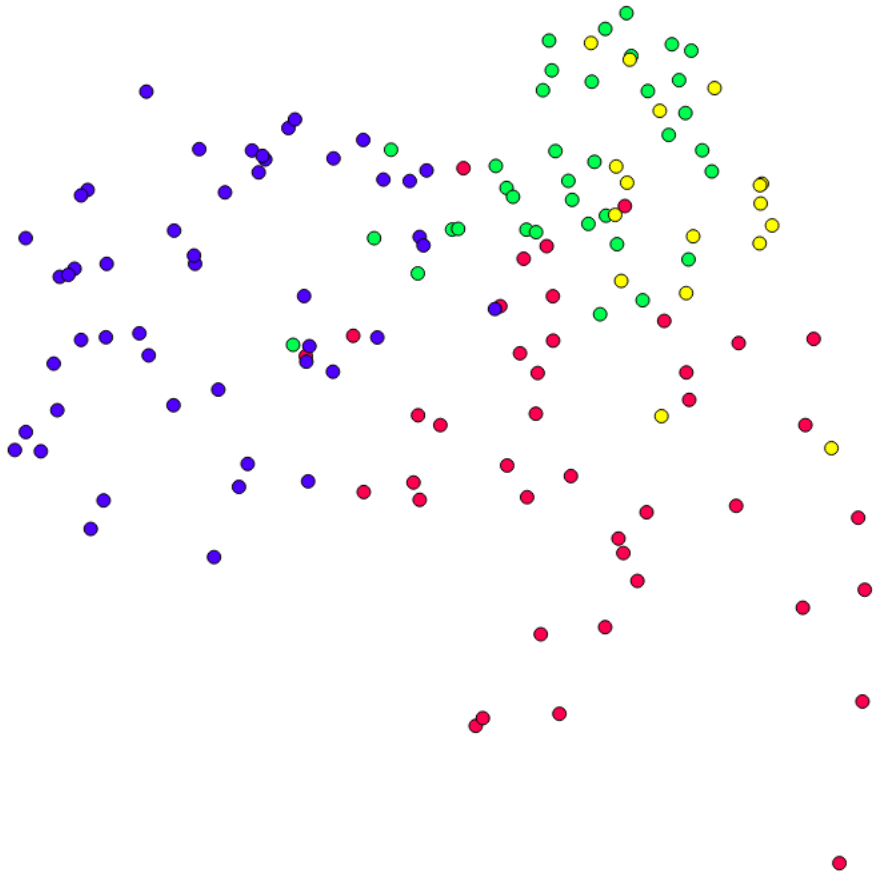
How many datapoints need to be labelled for correct clustering?

A. 5%

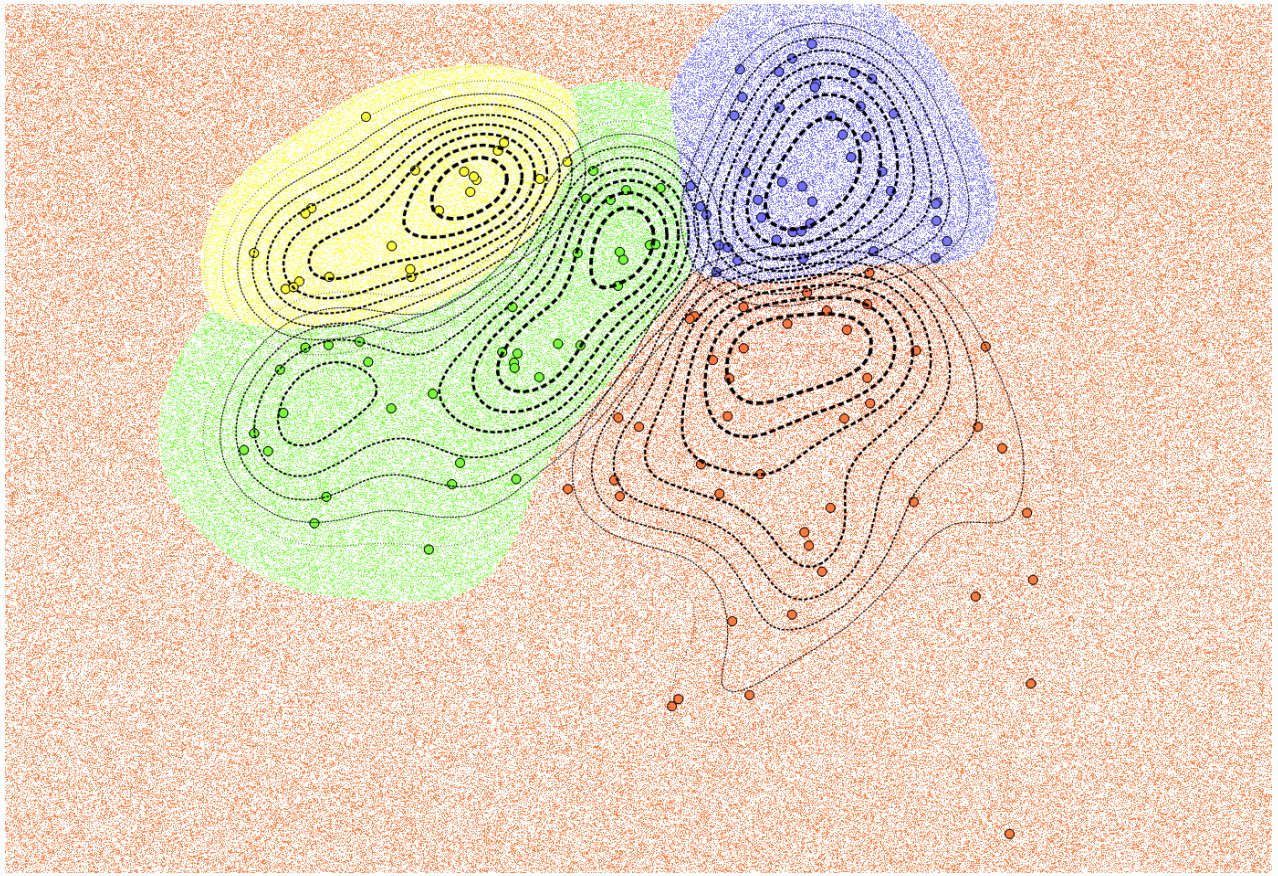
B. 50%

C. 75%

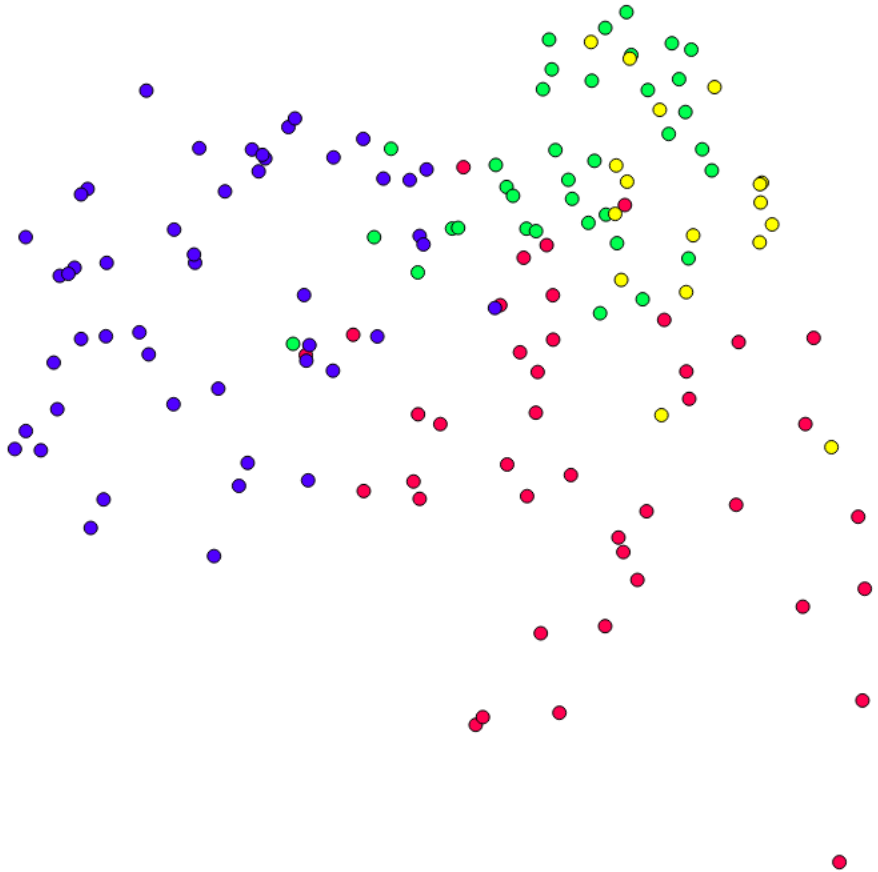




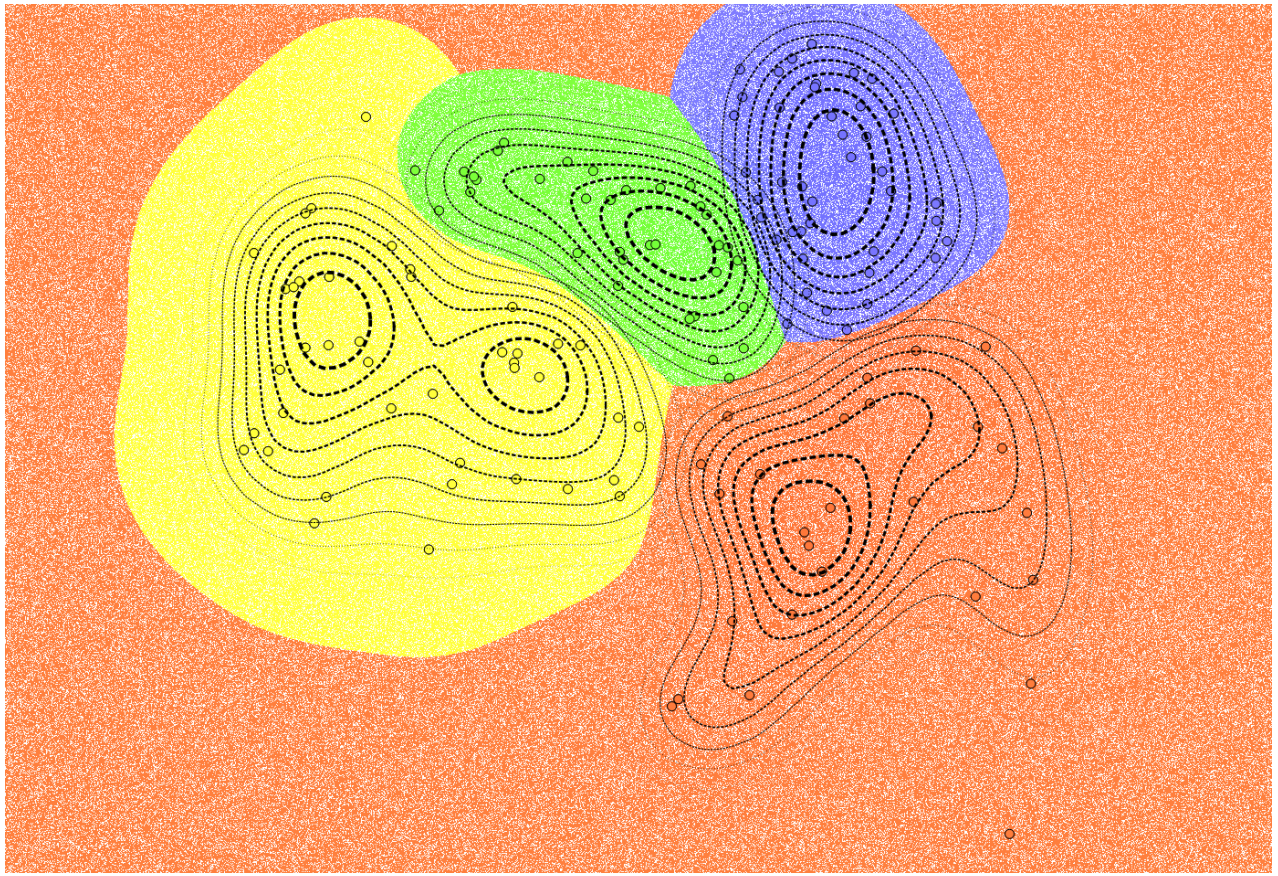
Semi-Supervised Clustering: Fraction Labelled

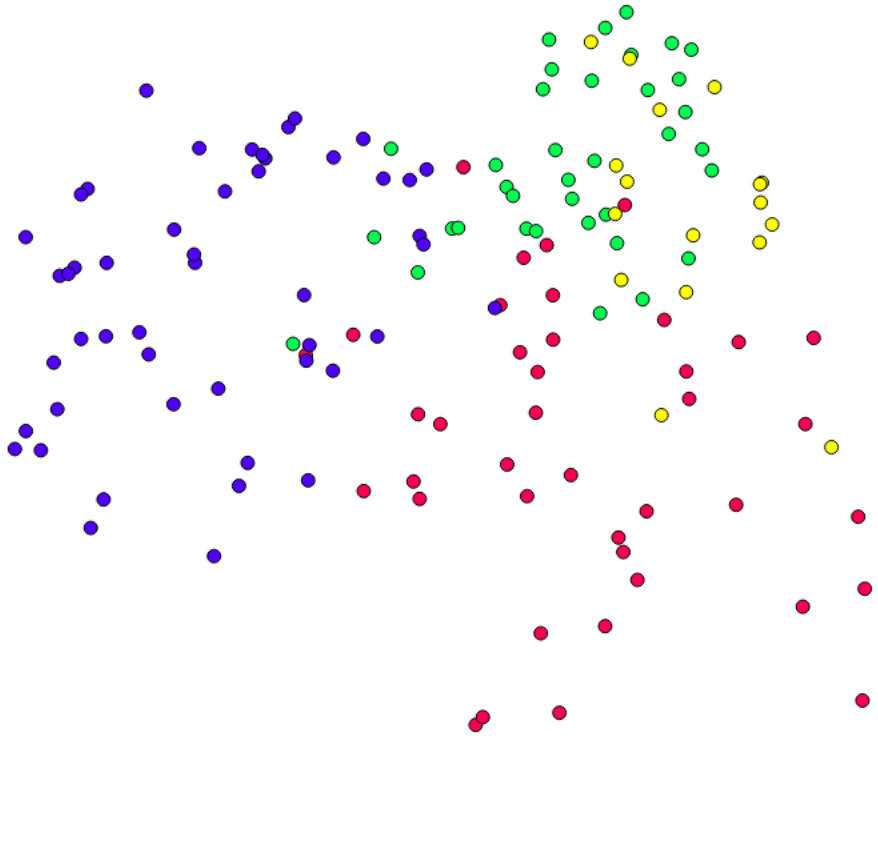


5% data labelled, F1=0.49



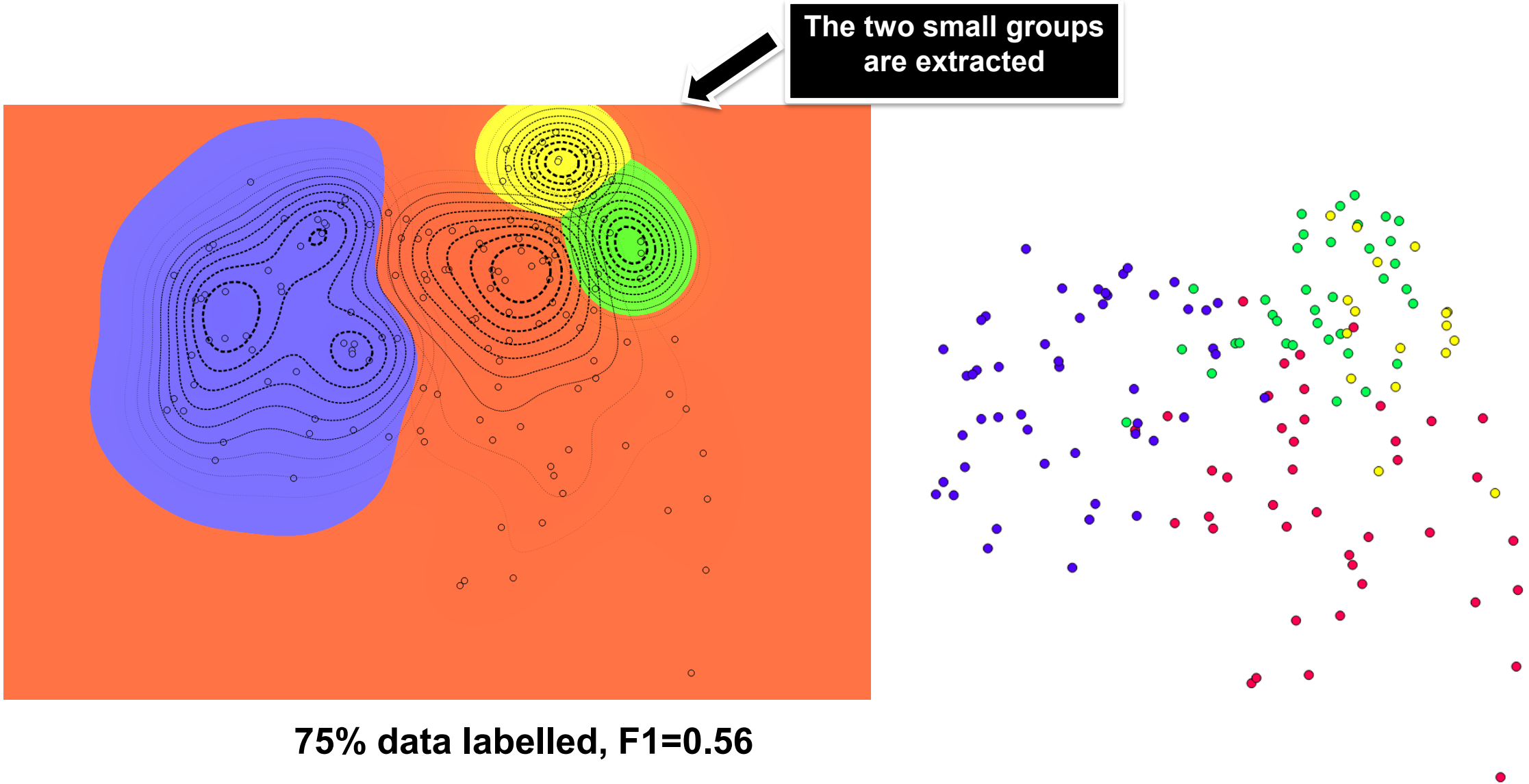
Semi-Supervised Clustering: Fraction Labelled





50% data labelled, F1=0.52

Semi-Supervised Clustering: Fraction Labelled

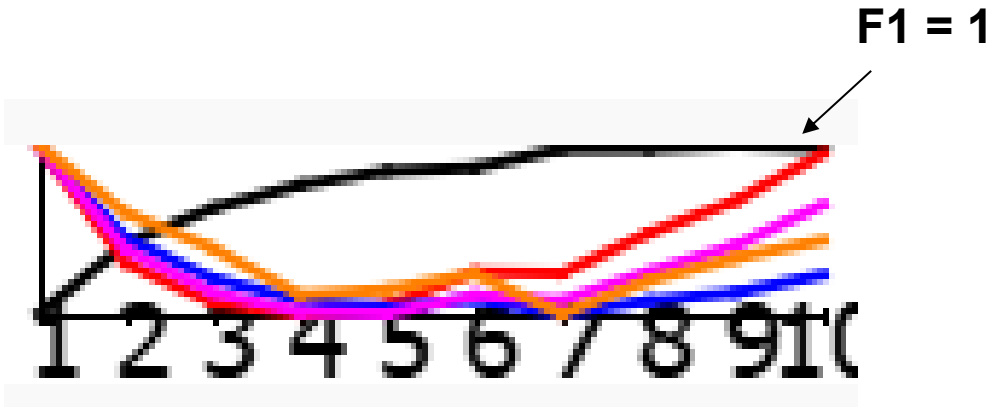


75% data labelled, F1=0.56

F1-measure and other metrics:

20% datapoints labelled

RSS
AIC
BIC
F1

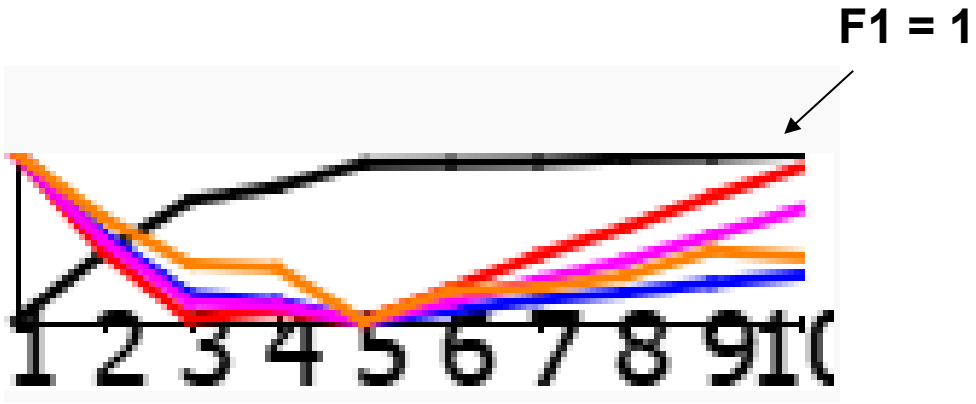


Which is the correct number of clusters?

F1-measure and other metrics:

50% datapoints labelled

RSS
AIC
BIC
F1

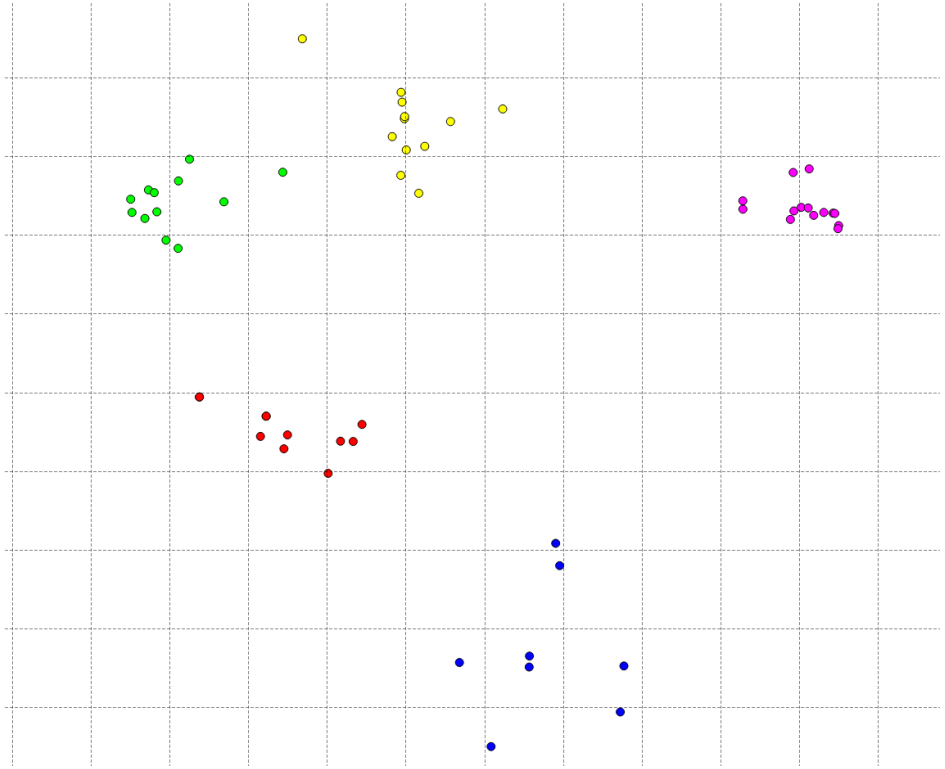


Which is the correct number of clusters?

Careful: result for 1 single run!

F1-measure and other metrics:

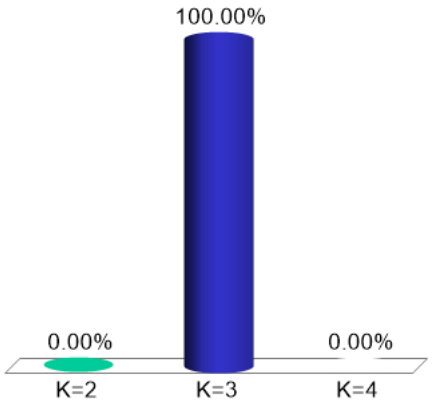
Projections onto two first eigenvectors



True number of clusters was 5.

Which is the best solution?

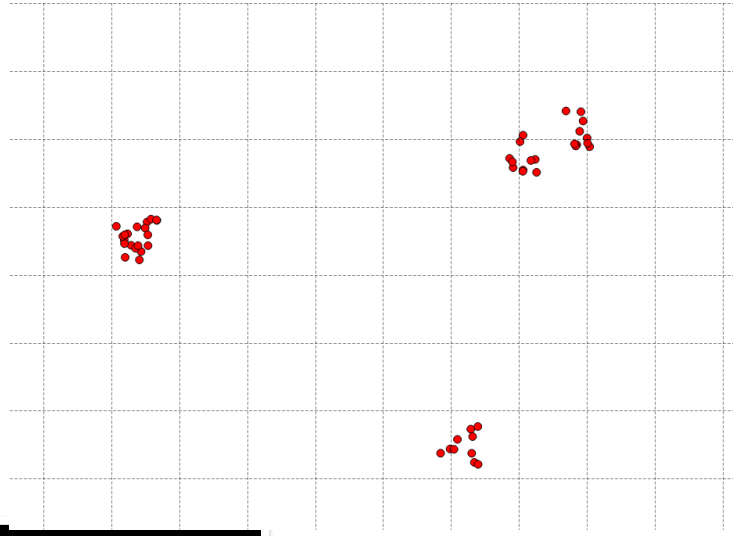
A. K=2
B. K=3
C. K=4



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

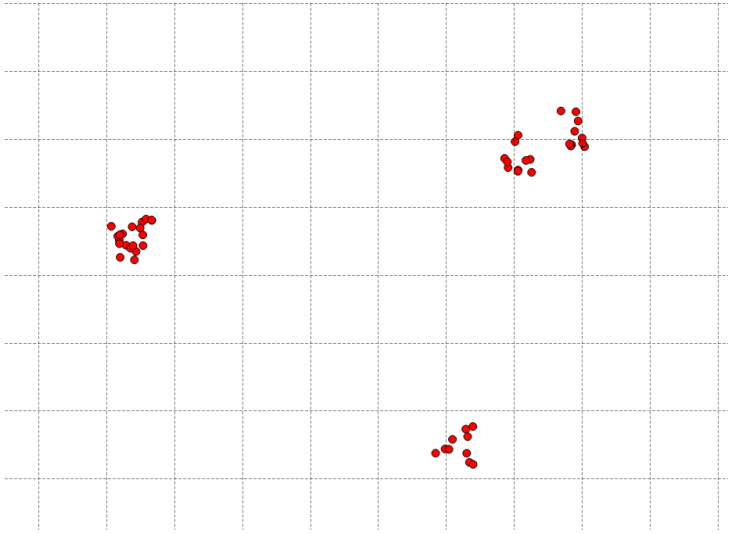
Careful: results given for 1 single run!



Original dataset
True value: K=3

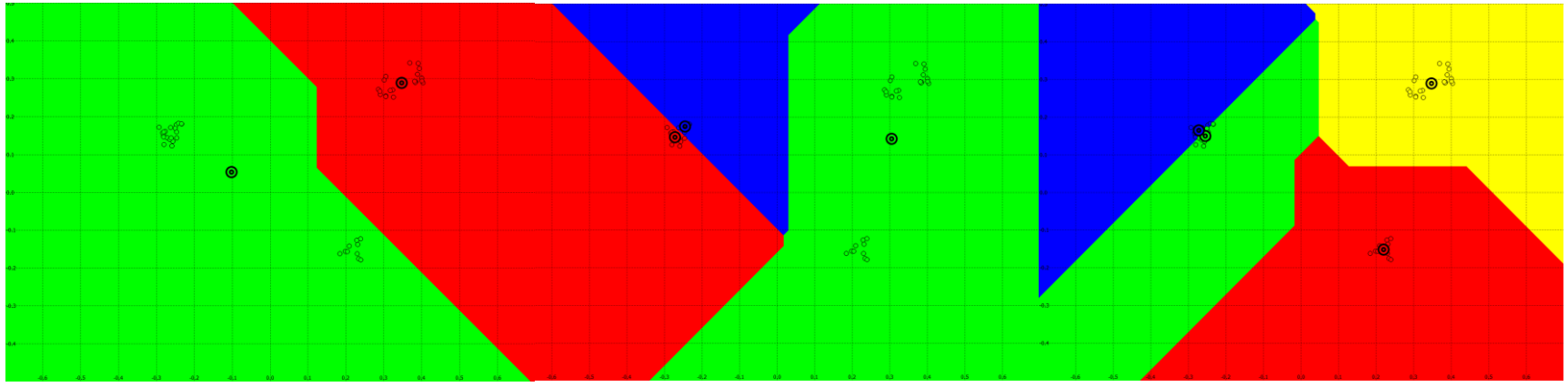
Trust F1 with more datapoints labelled
Optimum on all 3 metrics.

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



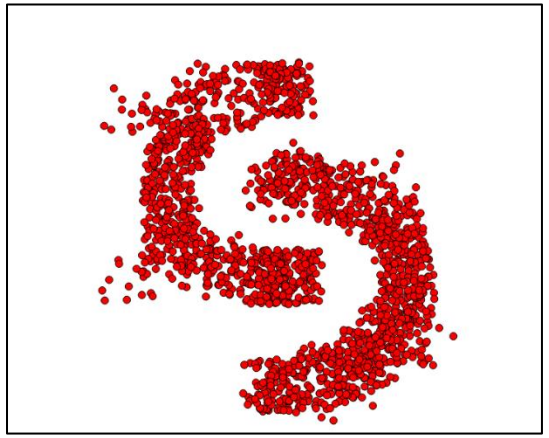
Original dataset
True value: K=3

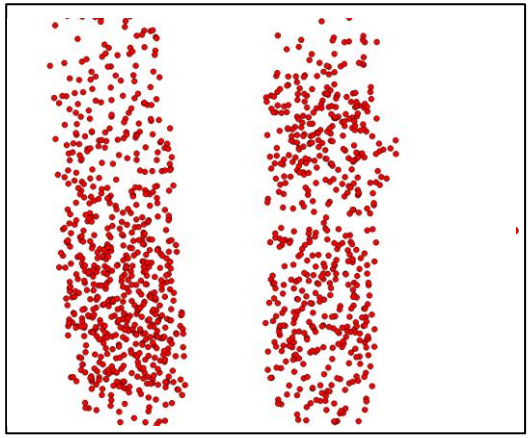
K-means is sensitive to initialization. Make sure to repeat and take best run when comparing results in RSS, AIC and BIC.

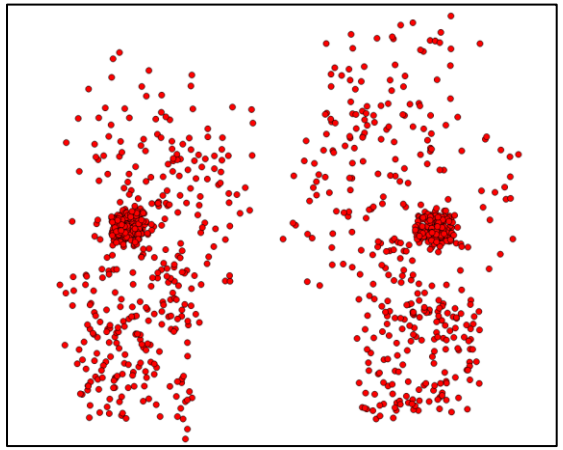


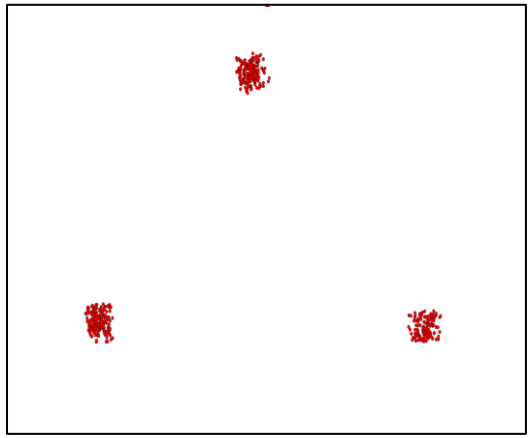
Clustering methods -- exercise

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



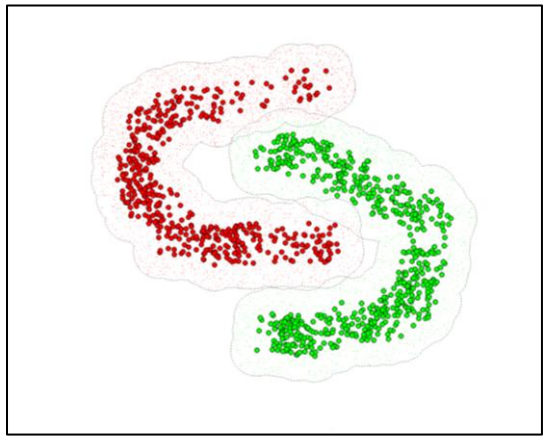




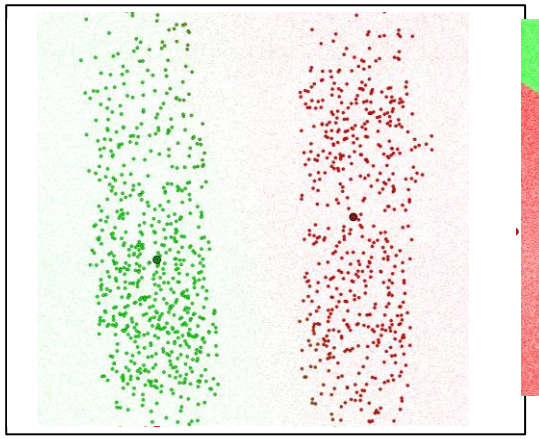


Use the cheapest clustering method (in computational costs) whenever possible.

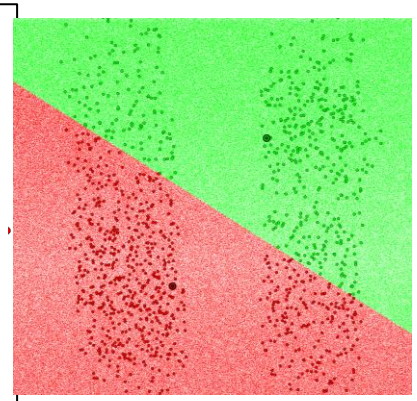
The cheapest is hard K-means, followed by soft K-means (computing an exponential is more costly than computing norm 2) and then DBSCAN.



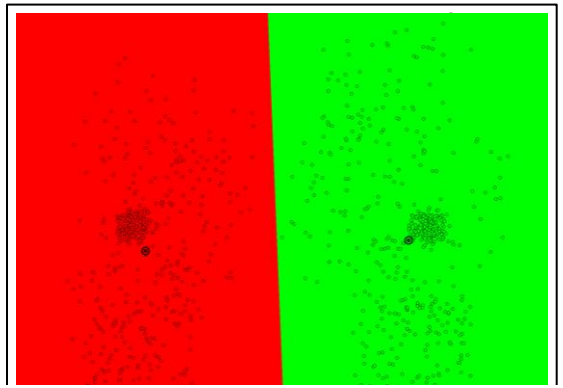
DBSCAN



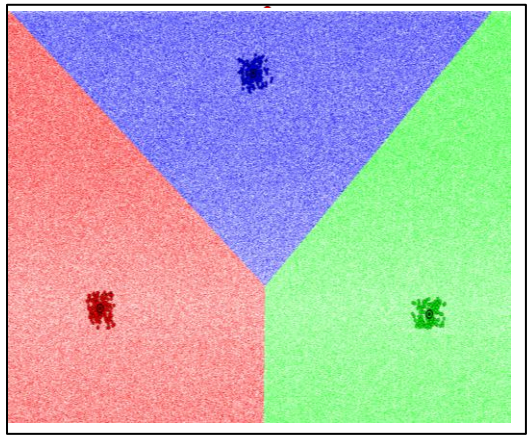
Soft K-means



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