Data Science for Business
Lecture #3
Limitations of k-Means

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

Outline

Influence of Scaling Data on Clustering

Influence of Initial Starting Points

Limitations of k-Means in Detecting Patterns

Influence of Scaling Data on Clustering

Structuring the data for k-Means

The basic objective of k-Means is to identify a grouping of observations that allows us to reduce the amount of information that must be considered. For k-Means to be useful we really want it to be robust, so that different people should reach the same conclusion with the same data.

k-Means requires the analyst to make several choices:

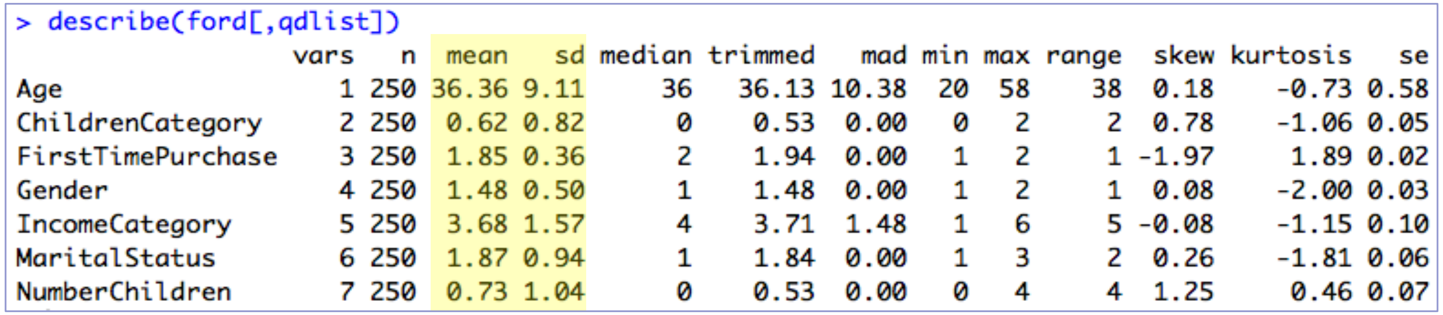
1. What data is to be used?
2. What is the value of k?
3. Select an initial partitioning of the data from which the algorithm works

What is the effect of using "scaled" versus "unscaled" data for the k-Means analysis of the Ford Ka demographic data?

Compare the Scaled and Unscaled Ford Ka Demographic Data

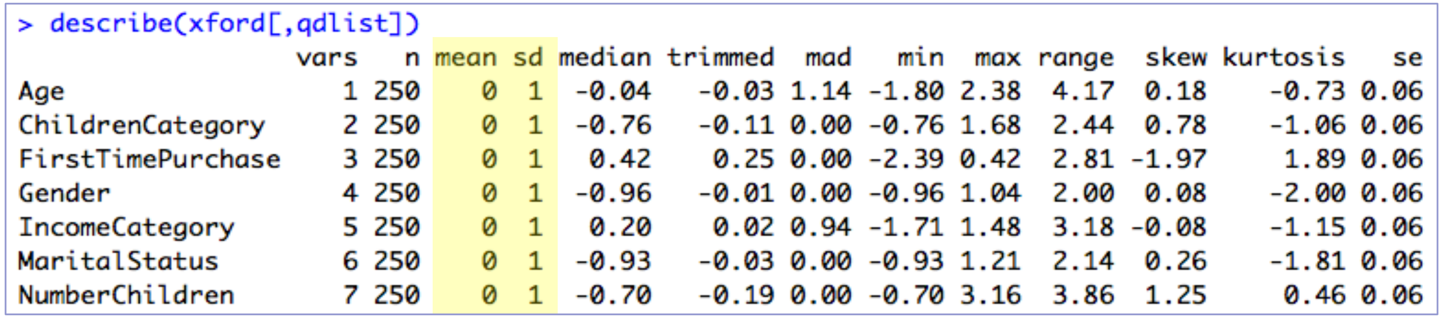
What is the difference in information between the two datasets?

Unscaled data (e.g., “Natural Units”)



```
> describe(ford[,qdlist])
                  vars   n  mean   sd median trimmed   mad min max range  skew kurtosis   se
Age                  1 250 36.36 9.11     36   36.13 10.38  20  58    38  0.18    -0.73 0.58
ChildrenCategory     2 250  0.62 0.82      0    0.53  0.00   0   2     2  0.78    -1.06 0.05
FirstTimePurchase    3 250  1.85 0.36      2    1.94  0.00   1   2     1 -1.97     1.89 0.02
Gender               4 250  1.48 0.50      1    1.48  0.00   1   2     1  0.08    -2.00 0.03
IncomeCategory       5 250  3.68 1.57      4    3.71  1.48   1   6     5 -0.08    -1.15 0.10
MaritalStatus        6 250  1.87 0.94      1    1.84  0.00   1   3     2  0.26    -1.81 0.06
NumberChildren       7 250  0.73 1.04      0    0.53  0.00   0   4     4  1.25     0.46 0.07
```

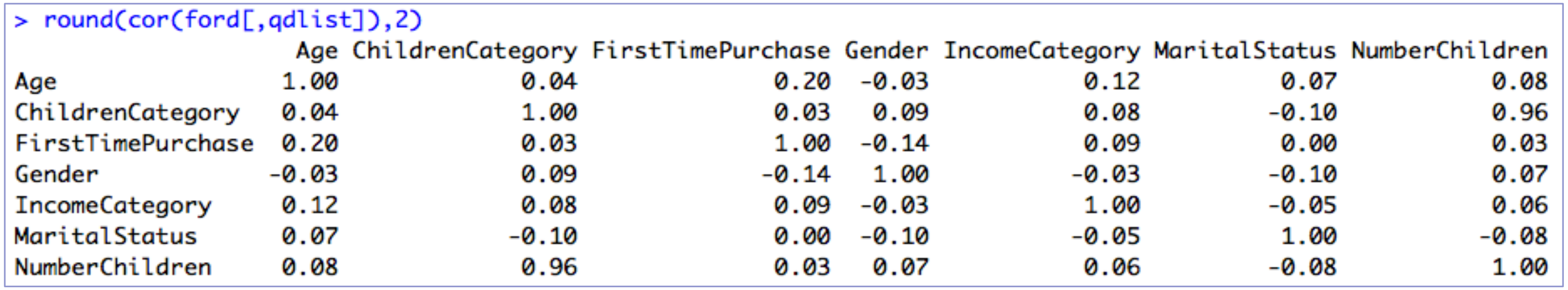
Scaled data (e.g., mean standardized to 0 and standard deviation to 1)



```
> describe(xford[,qdlist])
                  vars   n mean sd median trimmed  mad   min  max range  skew kurtosis   se
Age                  1 250    0  1  -0.04   -0.03 1.14 -1.80 2.38  4.17  0.18    -0.73 0.06
ChildrenCategory     2 250    0  1  -0.76   -0.11 0.00 -0.76 1.68  2.44  0.78    -1.06 0.06
FirstTimePurchase    3 250    0  1   0.42    0.25 0.00 -2.39 0.42  2.81 -1.97     1.89 0.06
Gender               4 250    0  1  -0.96   -0.01 0.00 -0.96 1.04  2.00  0.08    -2.00 0.06
IncomeCategory       5 250    0  1   0.20    0.02 0.94 -1.71 1.48  3.18 -0.08    -1.15 0.06
MaritalStatus        6 250    0  1  -0.93   -0.03 0.00 -0.93 1.21  2.14  0.26    -1.81 0.06
NumberChildren       7 250    0  1  -0.70   -0.19 0.00 -0.70 3.16  3.86  1.25     0.46 0.06
```

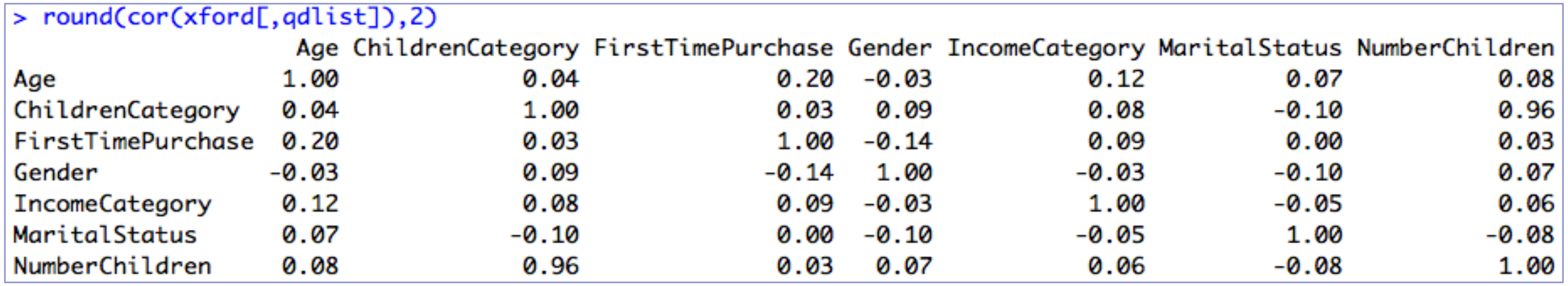
Compare the correlation patterns

Unscaled data (e.g., “Natural Units”)



```
> round(cor(ford[,qdlist]),2)
                    Age ChildrenCategory FirstTimePurchase Gender IncomeCategory MaritalStatus NumberChildren
Age                1.00             0.04              0.20  -0.03           0.12          0.07           0.08
ChildrenCategory   0.04             1.00              0.03   0.09           0.08         -0.10           0.96
FirstTimePurchase  0.20             0.03              1.00  -0.14           0.09          0.00           0.03
Gender            -0.03             0.09             -0.14   1.00          -0.03         -0.10           0.07
IncomeCategory     0.12             0.08              0.09  -0.03           1.00         -0.05           0.06
MaritalStatus      0.07            -0.10              0.00  -0.10          -0.05          1.00          -0.08
NumberChildren     0.08             0.96              0.03   0.07           0.06         -0.08           1.00
```

Scaled data (e.g., mean standardized to 0 and standard deviation to 1)

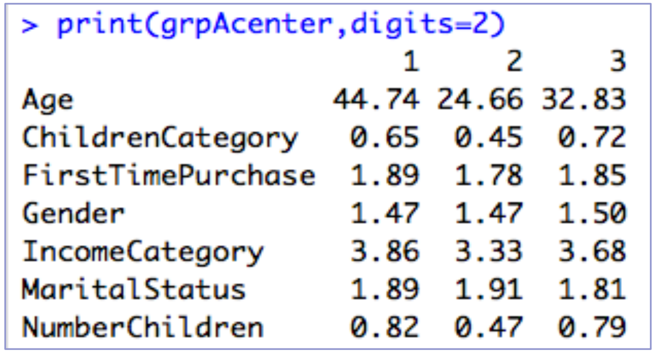


```
> round(cor(xford[,qdlist]),2)
                    Age ChildrenCategory FirstTimePurchase Gender IncomeCategory MaritalStatus NumberChildren
Age                1.00             0.04              0.20  -0.03           0.12          0.07           0.08
ChildrenCategory   0.04             1.00              0.03   0.09           0.08         -0.10           0.96
FirstTimePurchase  0.20             0.03              1.00  -0.14           0.09          0.00           0.03
Gender            -0.03             0.09             -0.14   1.00          -0.03         -0.10           0.07
IncomeCategory     0.12             0.08              0.09  -0.03           1.00         -0.05           0.06
MaritalStatus      0.07            -0.10              0.00  -0.10          -0.05          1.00          -0.08
NumberChildren     0.08             0.96              0.03   0.07           0.06         -0.08           1.00
```

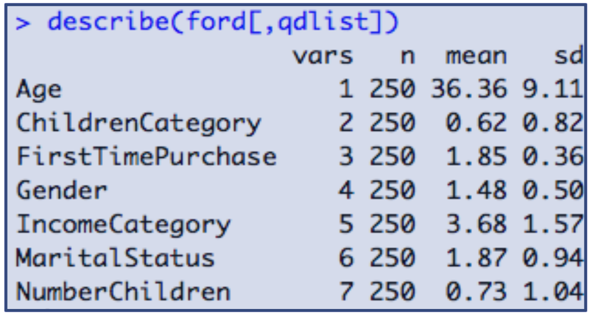
Compare the k-Means solution (with 3 clusters)

What interpretation would you give for each cluster solution?

Unscaled data (e.g., "Natural Units")

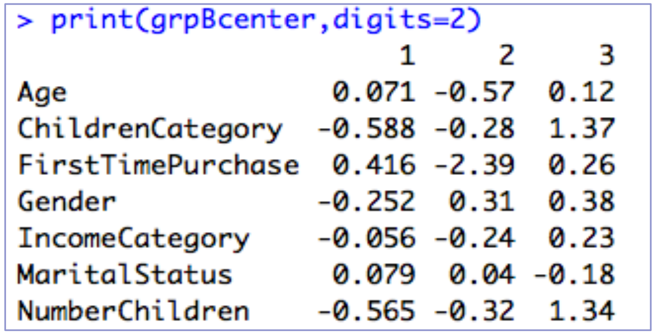


```
> print(grpAcenter,digits=2)
                      1     2     3
Age               44.74 24.66 32.83
ChildrenCategory   0.65  0.45  0.72
FirstTimePurchase  1.89  1.78  1.85
Gender             1.47  1.47  1.50
IncomeCategory     3.86  3.33  3.68
MaritalStatus      1.89  1.91  1.81
NumberChildren     0.82  0.47  0.79
```

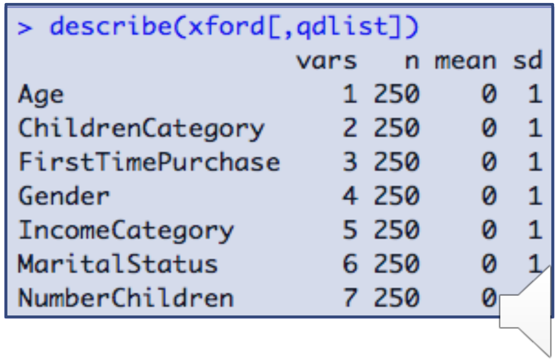


```
> describe(ford[,qdlist])
                  vars   n  mean   sd
Age                  1 250 36.36 9.11
ChildrenCategory     2 250  0.62 0.82
FirstTimePurchase    3 250  1.85 0.36
Gender               4 250  1.48 0.50
IncomeCategory       5 250  3.68 1.57
MaritalStatus        6 250  1.87 0.94
NumberChildren       7 250  0.73 1.04
```

Scaled data (e.g., mean standardized to 0 and standard deviation to 1)



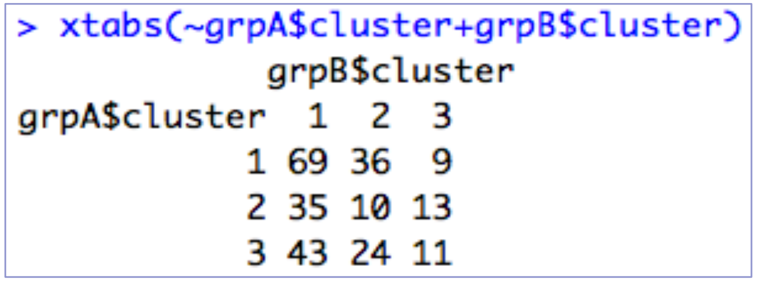
```
> print(grpBcenter,digits=2)
                       1     2     3
Age                0.071 -0.57  0.12
ChildrenCategory  -0.588 -0.28  1.37
FirstTimePurchase  0.416 -2.39  0.26
Gender            -0.252  0.31  0.38
IncomeCategory    -0.056 -0.24  0.23
MaritalStatus      0.079  0.04 -0.18
NumberChildren    -0.565 -0.32  1.34
```



```
> describe(xford[,qdlist])
                  vars   n mean sd
Age                  1 250    0  1
ChildrenCategory     2 250    0  1
FirstTimePurchase    3 250    0  1
Gender               4 250    0  1
IncomeCategory       5 250    0  1
MaritalStatus        6 250    0  1
NumberChildren       7 250    0
```

Compare the k-Means solution (with 3 clusters)

Compare the cluster assignments from the Unscaled and Scaled data



```
> xtabs(~grpA$cluster+grpB$cluster)
             grpB$cluster
grpA$cluster  1  2  3
           1 69 36  9
           2 35 10 13
           3 43 24 11
```

What was the impact of standardization on the solution?

Summary

k-Means is attempting to find the groups of observations that explain as much of the variation as possible.

Scaling changes the variation of the data. If we standardize then all the variables have equal weights. Sometimes this is useful, when we really want all the variables to have about the same weight, but bad if there are some variables that we know are really important (say age and gender for ad buys or profits)

Lesson. *The scale of the data matters. The variables you use and how you transform them can dramatically alter the solutions you find.*

"Let the data speak" or "Let me speak to the data"

Influence of Initial Starting Points for k-Means

Robustness of k-Means

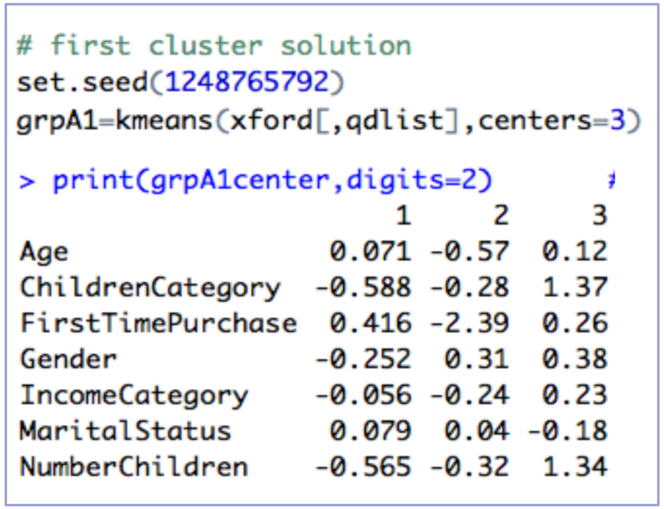
The basic objective of k-Means is to identify a grouping of observations that allows us to reduce the amount of information that must be considered. For k-Means to be useful we really want it to be robust, so that different people should reach the same conclusion with the same data.

Remember that k-Means requires the analyst to make several choices:

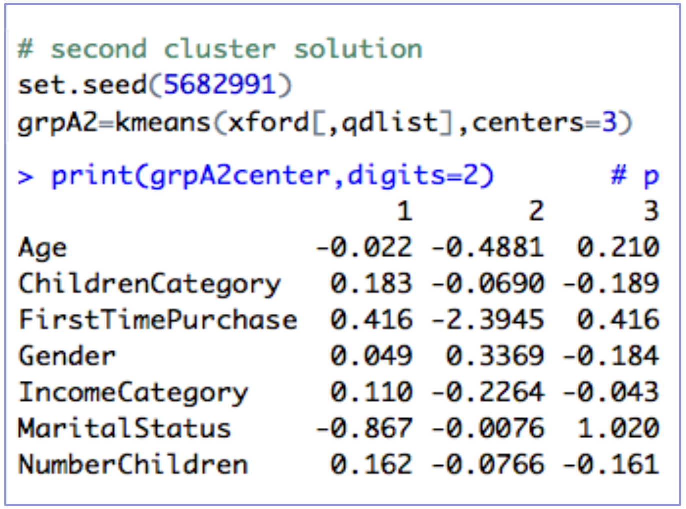
1. What data is to be used?
2. What is the value of k?
3. Select an initial partitioning of the data from which the algorithm works

Typically we focus on the first two questions, but let’s consider the 3rd question of the initial point.

Compare k-Means with two different starting points

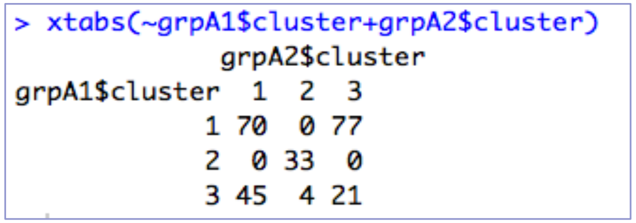


```
# first cluster solution
set.seed(1248765792)
grpA1=kmeans(xford[,qdlist],centers=3)

> print(grpA1center,digits=2)          #
                       1     2     3
Age                0.071 -0.57  0.12
ChildrenCategory  -0.588 -0.28  1.37
FirstTimePurchase  0.416 -2.39  0.26
Gender            -0.252  0.31  0.38
IncomeCategory    -0.056 -0.24  0.23
MaritalStatus      0.079  0.04 -0.18
NumberChildren    -0.565 -0.32  1.34
```



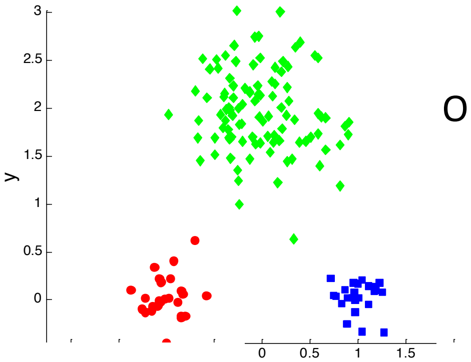
```
# second cluster solution
set.seed(5682991)
grpA2=kmeans(xford[,qdlist],centers=3)

> print(grpA2center,digits=2)        # p
                       1       2      3
Age               -0.022 -0.4881  0.210
ChildrenCategory   0.183 -0.0690 -0.189
FirstTimePurchase  0.416 -2.3945  0.416
Gender             0.049  0.3369 -0.184
IncomeCategory     0.110 -0.2264 -0.043
MaritalStatus     -0.867 -0.0076  1.020
NumberChildren     0.162 -0.0766 -0.161
```

What is the difference in cluster solutions?

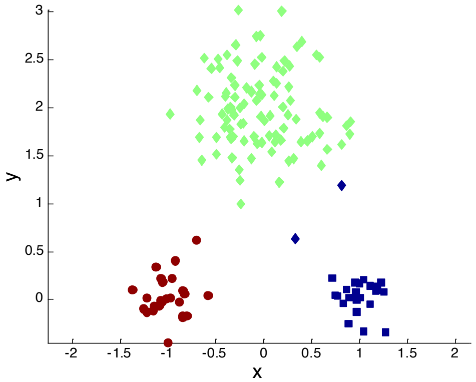


```
> xtabs(~grpA1$cluster+grpA2$cluster)
             grpA2$cluster
grpA1$cluster  1  2  3
            1 70  0 77
            2  0 33  0
            3 45  4 21
```

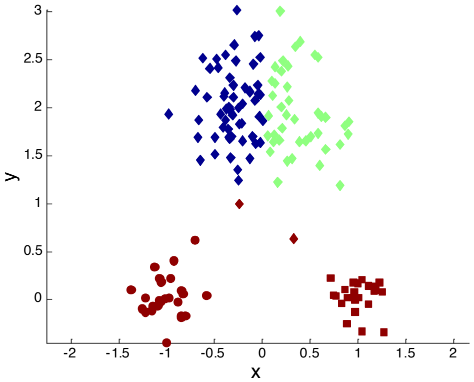
Two different k-Means Clusterings



Original Points

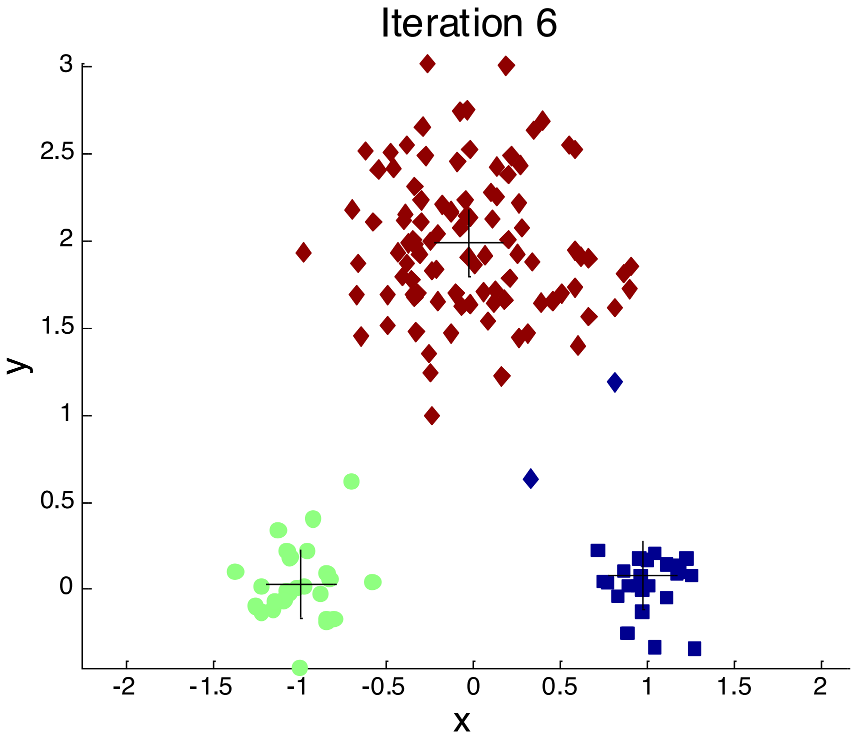


Optimal Clustering

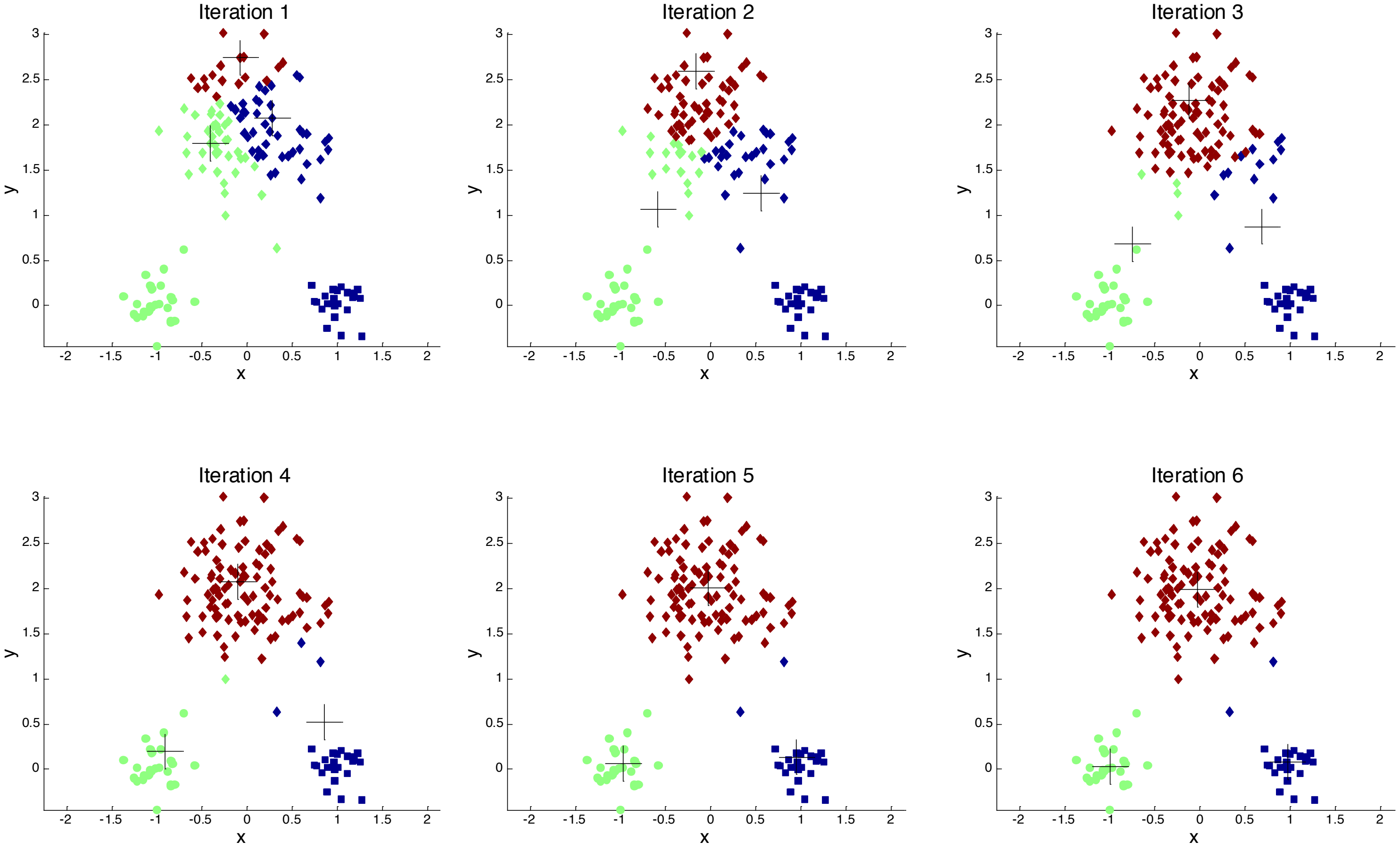


Sub-optimal Clustering

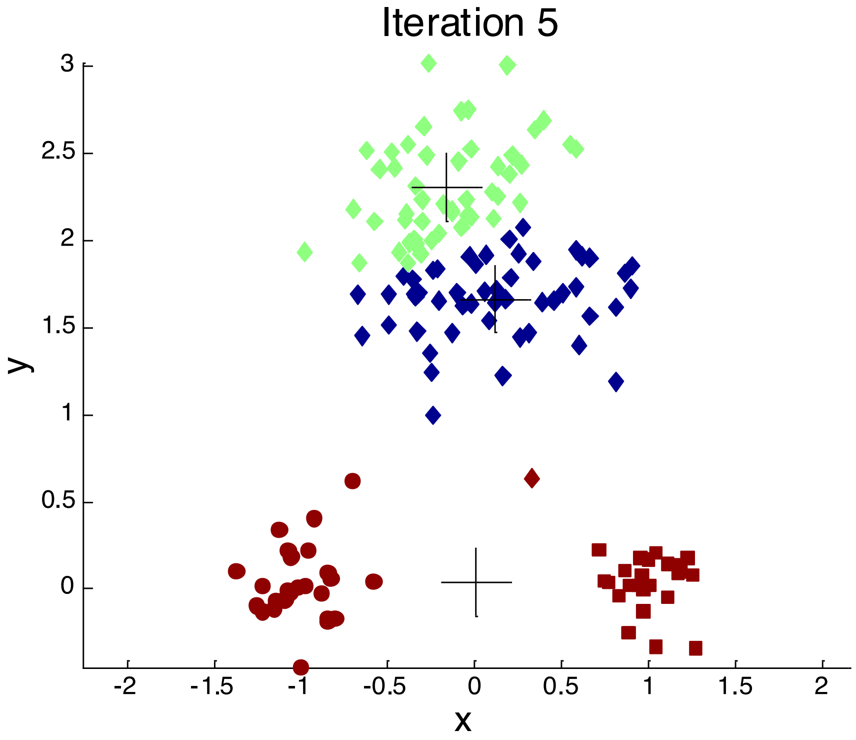
Importance of Choosing Initial Centroids



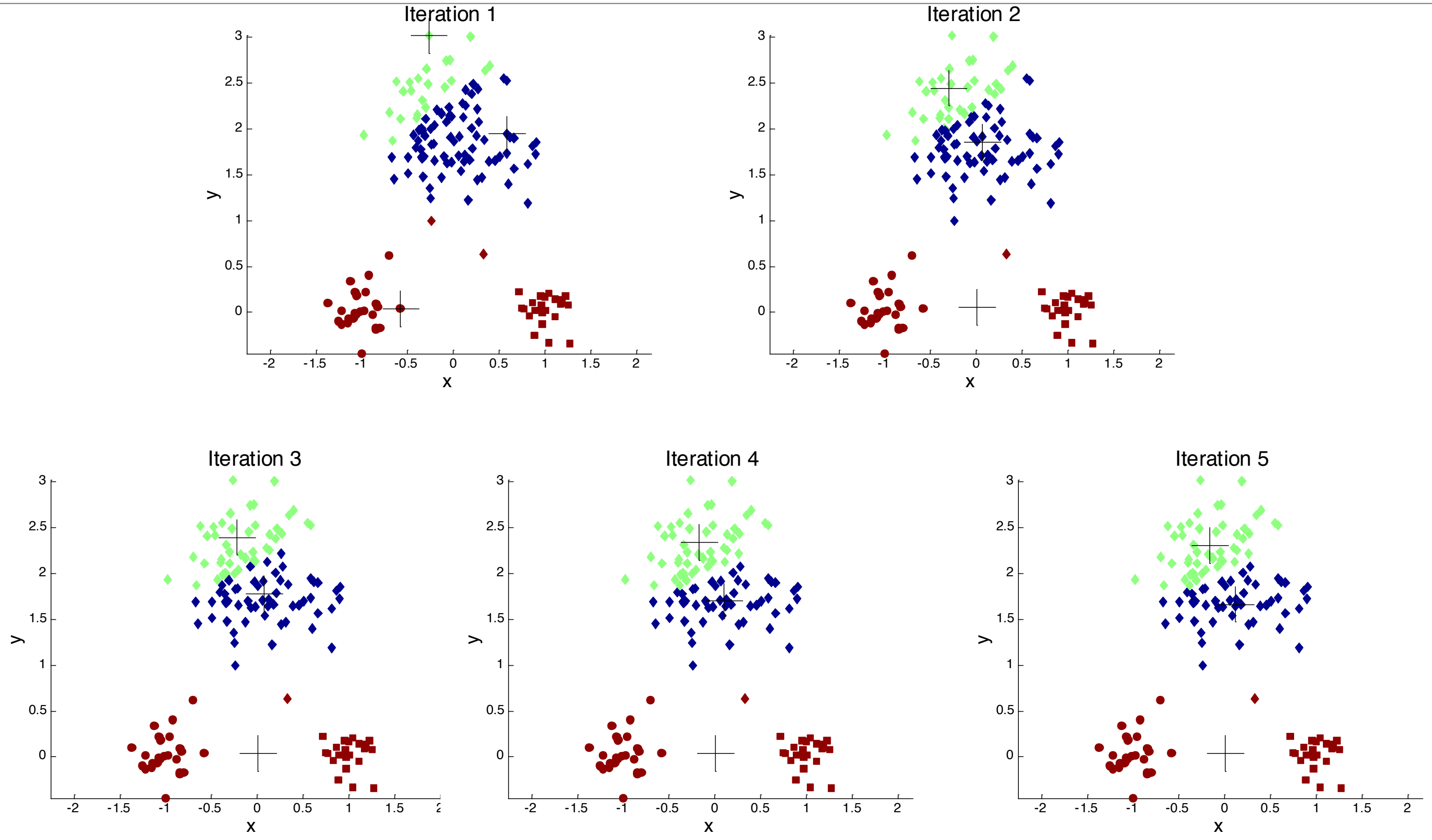
Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids …



Importance of Choosing Initial Centroids …



Problem with Random Selection of Initial Centroids

If there are K ‘real’ clusters then the chance of selecting one centroid from each cluster is small.

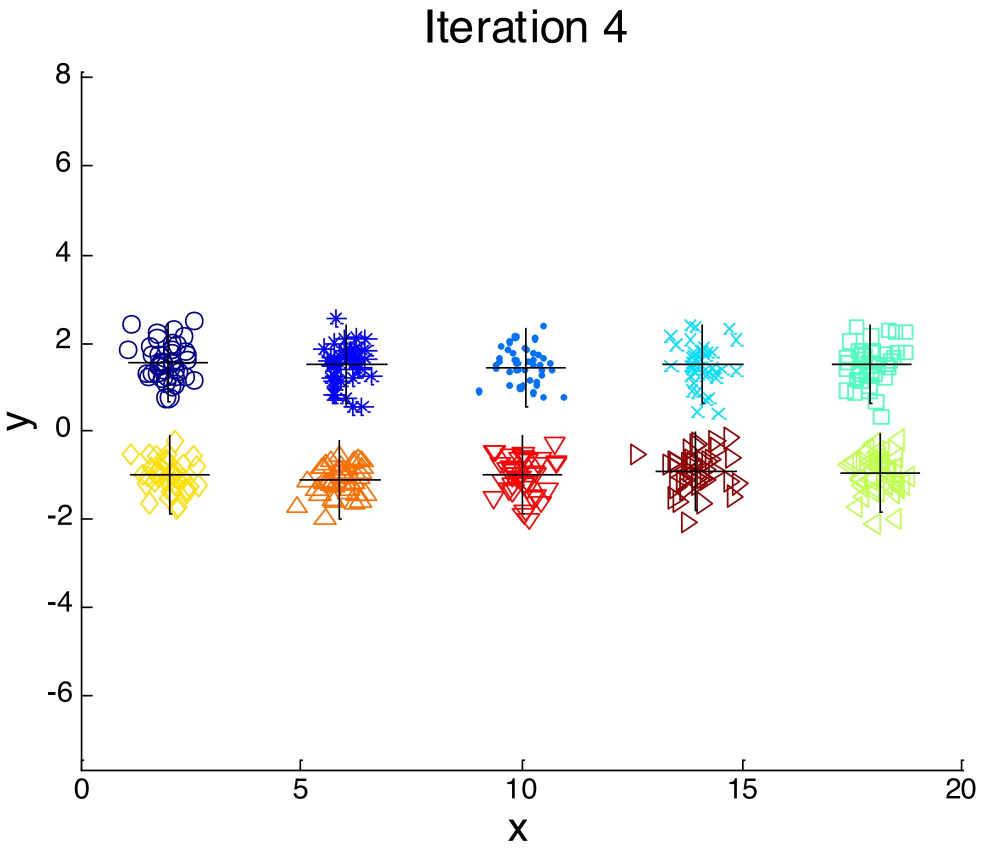
- Chance is relatively small when K is large
- If clusters are the same size, n, then

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

- For example, if K = 10, then probability = 10!/1010 = 0.00036
- Sometimes the initial centroids will readjust themselves in ‘right’ way, and sometimes they don’t
- Consider an example of five pairs of clusters

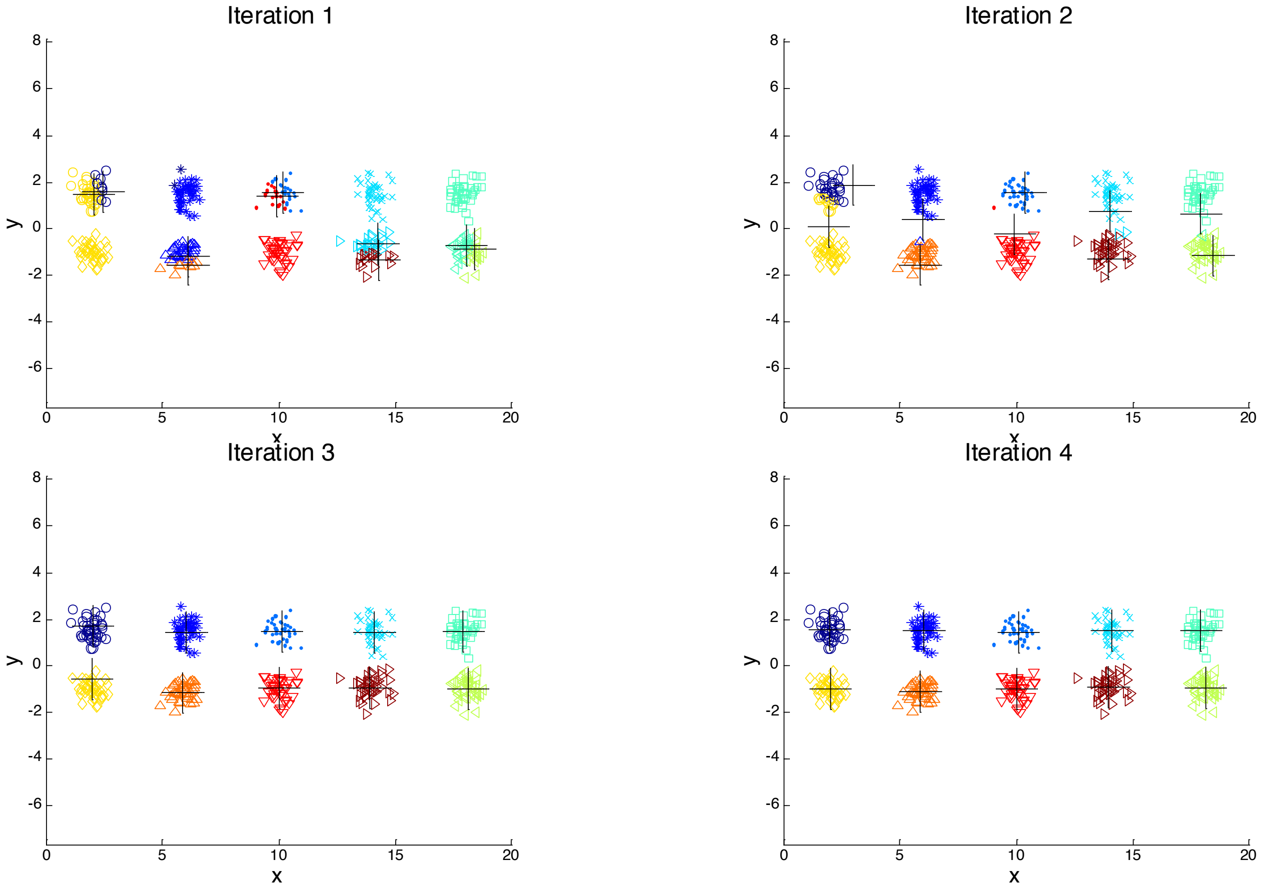
10 Clusters Example

Starting with two initial centroids in one cluster of each pair of clusters



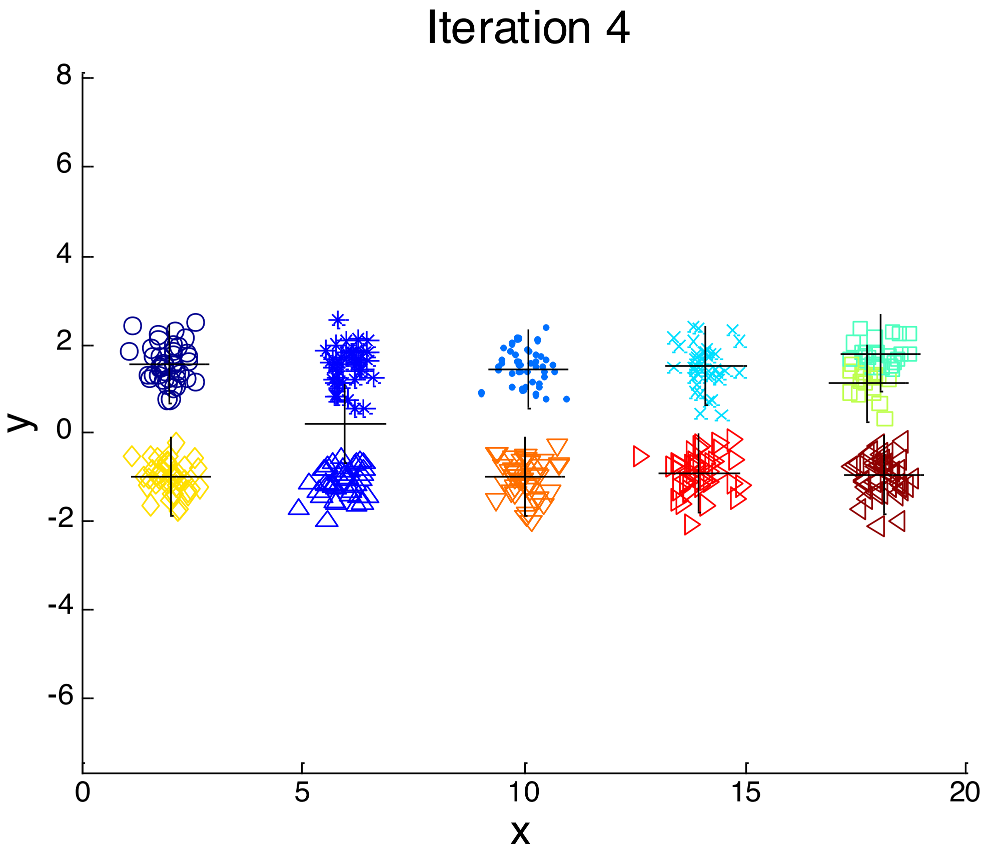
10 Clusters Example

Starting with two initial centroids in one cluster of each pair of clusters



10 Clusters Example with an Alternative Initialization

Starting with some pairs of clusters having three initial centroids, while other have only one.



10 Clusters Example with an Alternative Initialization

Starting with some pairs of clusters having three initial centroids, while other have only one.

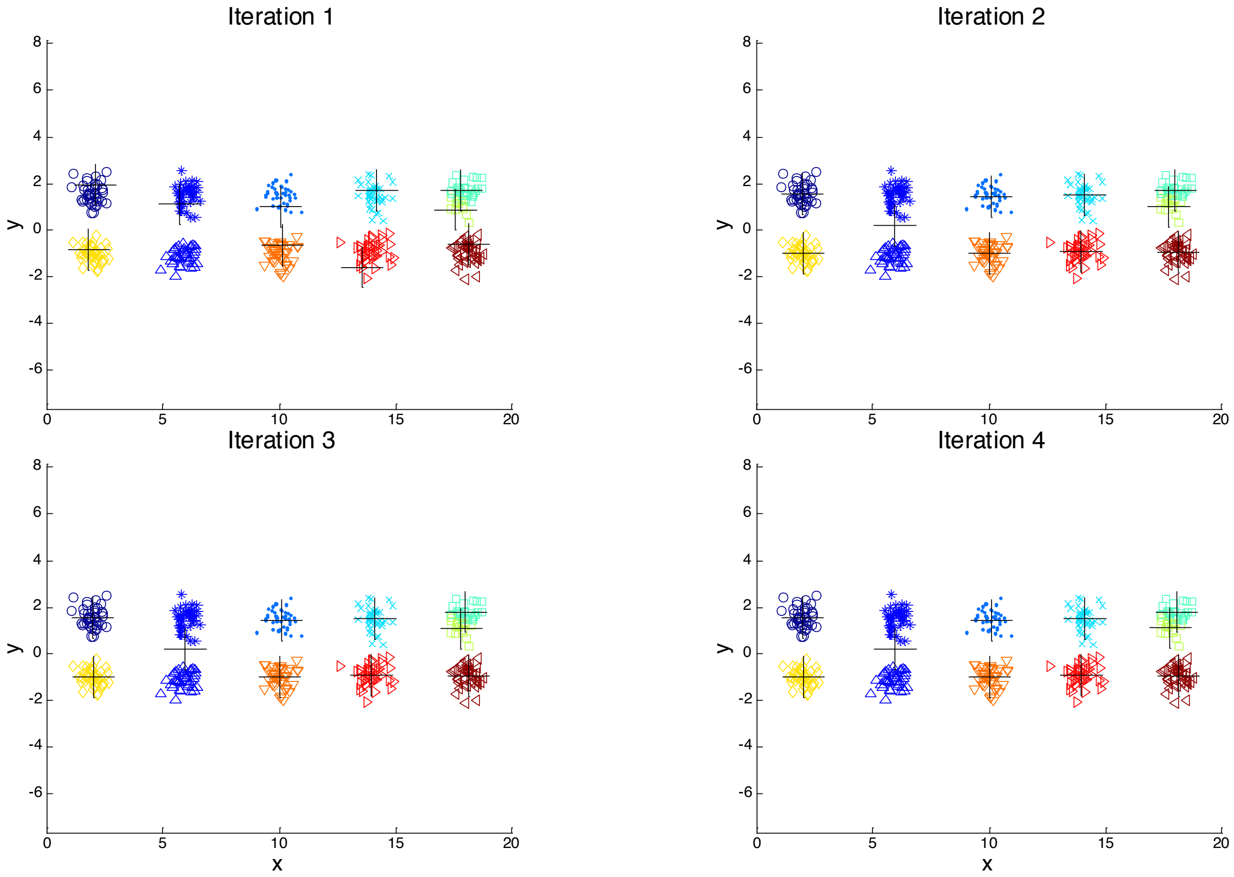
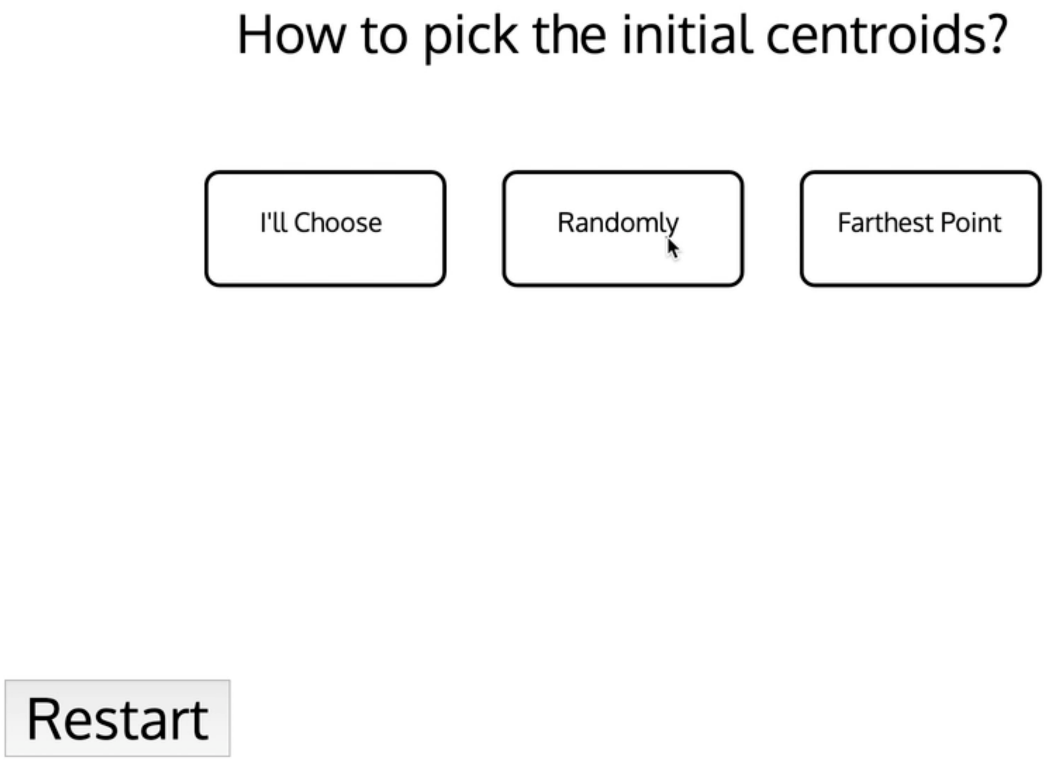


Illustration of k-Means using Packed Circles
Moral: Bad initializations can yield bad solutions

https://www.naftaliharris.com/blog/visualizing-k-Means-clustering/

Illustration of k-Means using Packed Circles
Moral: Bad initializations can yield bad solutions



How to pick the initial centroids?

https://www.naftaliharris.com/blog/visualizing-k-Means-clustering/

Solutions to Initialization Problem

Multiple runs

- Helps, but probability is not on your side

Sample and use hierarchical clustering to determine initial centroids

Select more than k initial centroids and then select among these initial centroids

- Select most widely separated

Postprocessing

Bisecting k-Means

- Not as susceptible to initialization issues

Limitations of k-Means in Detecting Patterns

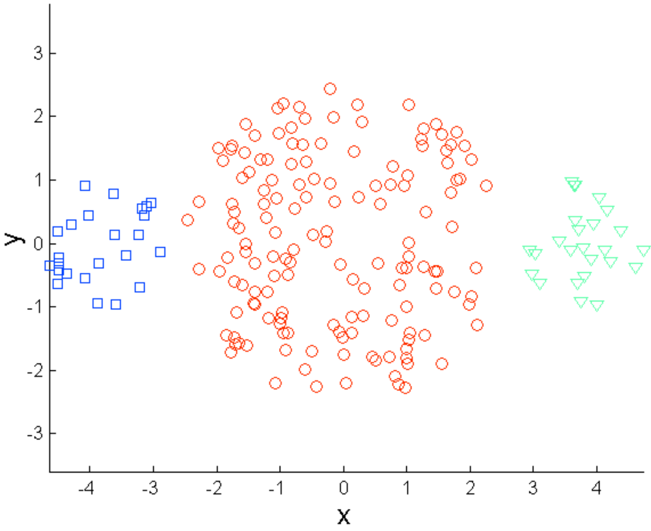
Limitations of k-Means

k-Means has problems when clusters are of differing

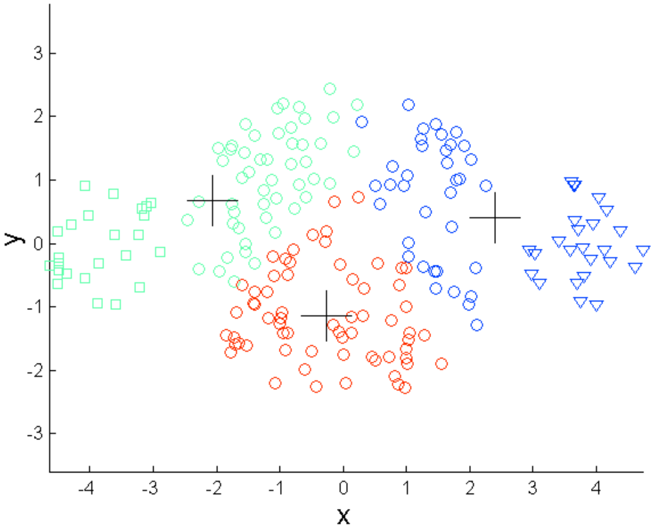
- Sizes
- Densities
- Non-globular shapes

k-Means has problems when the data contains outliers.

Limitations of k-Means: Differing Sizes

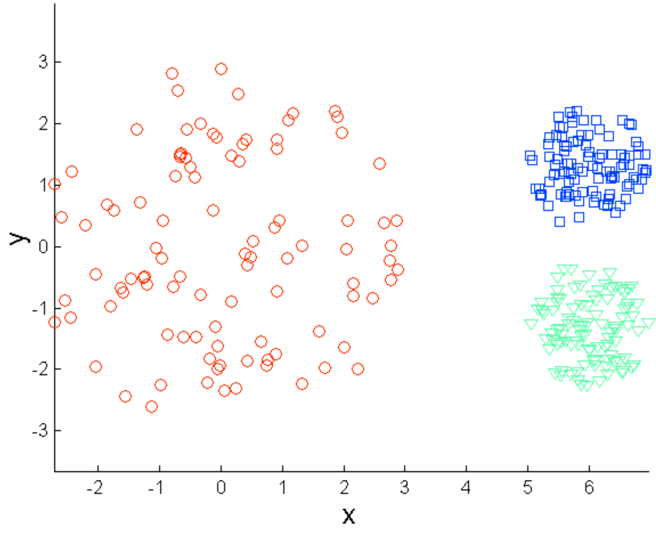


Original Points

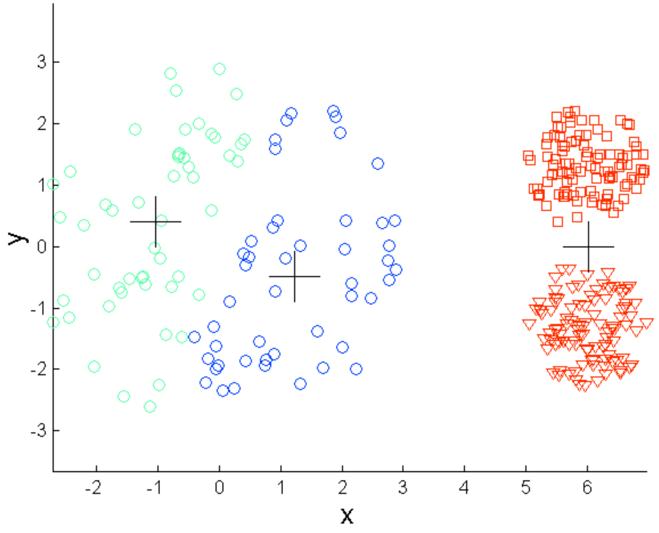


k-Means (3 Clusters)

Limitations of k-Means: Differing Density

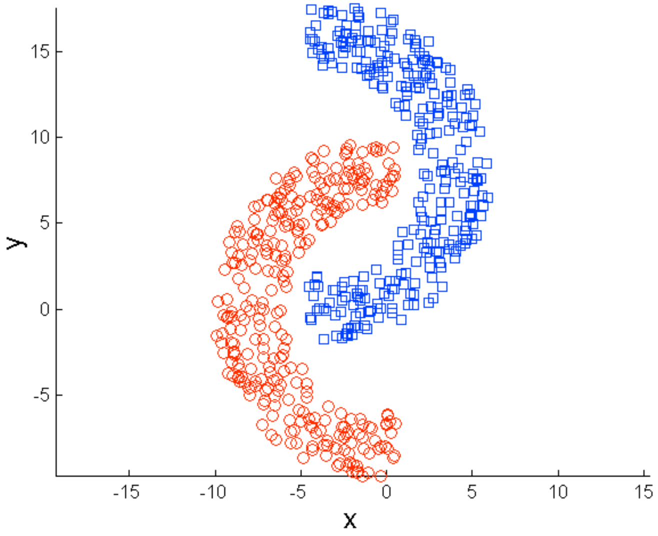


Original Points

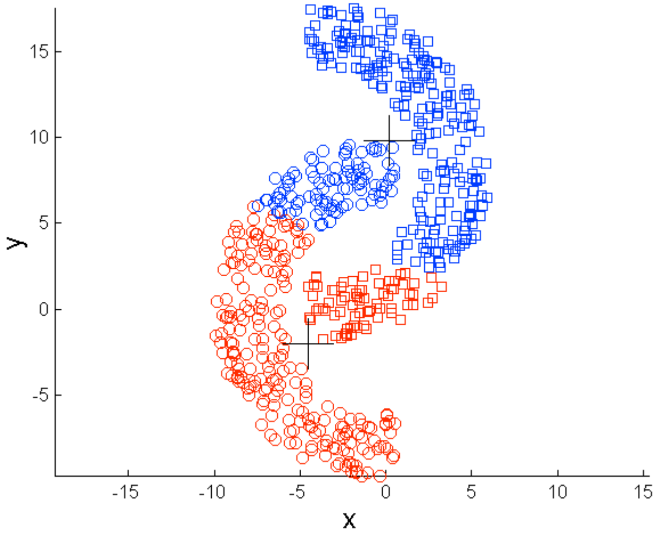


k-Means (3 Clusters)

Limitations of k-Means: Non-globular Shapes

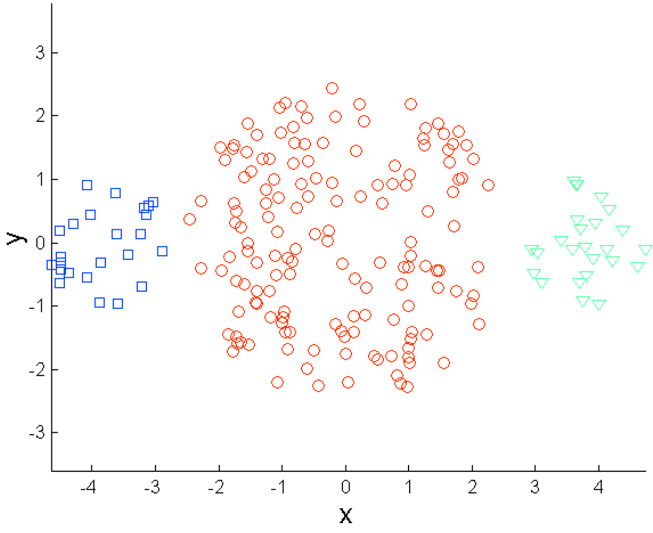


Original Points

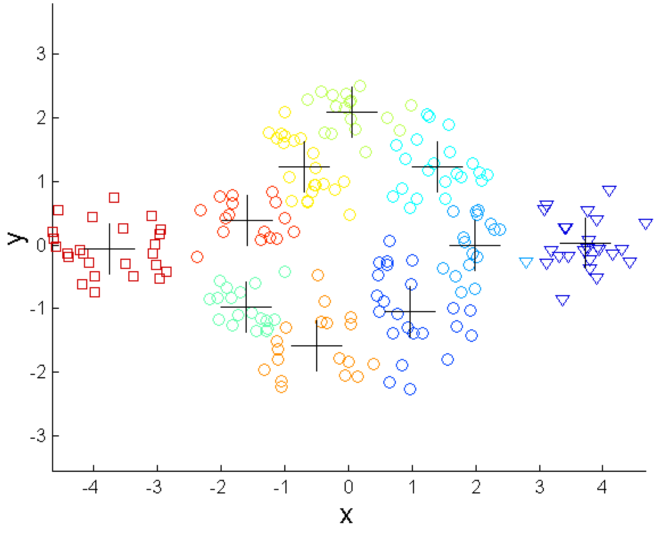


k-Means (2 Clusters)

Overcoming k-Means Limitations



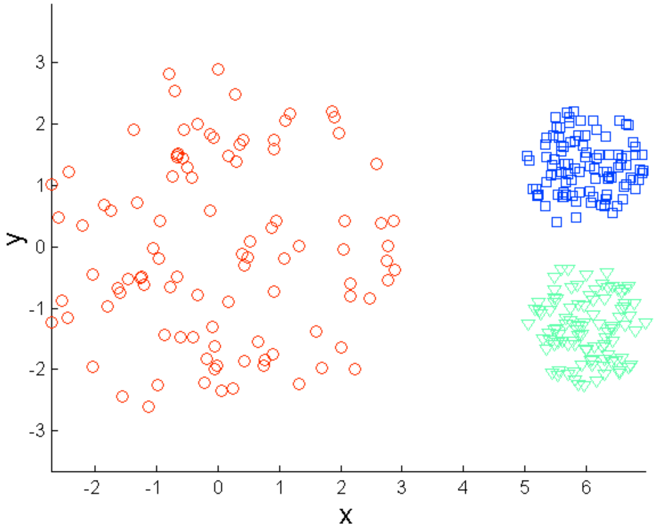
Original Points



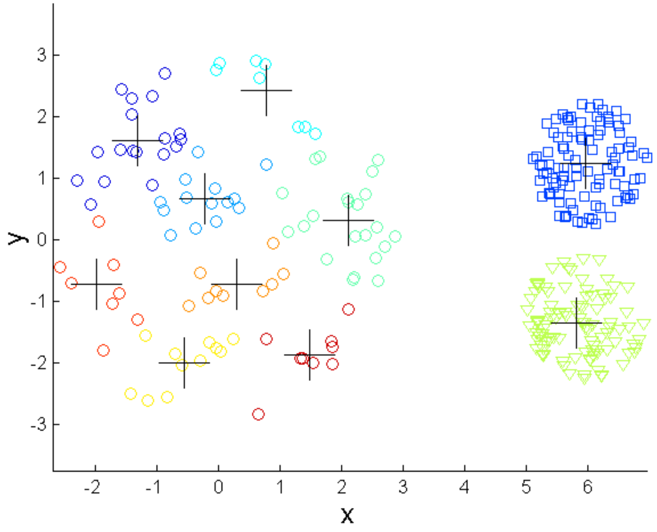
k-Means Clusters

One solution is to use many clusters.
Find parts of clusters, but need to put together.

Overcoming k-Means Limitations

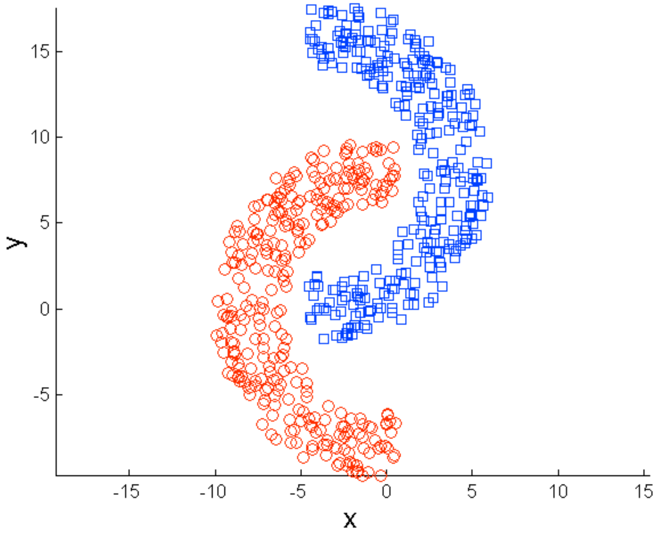


Original Points

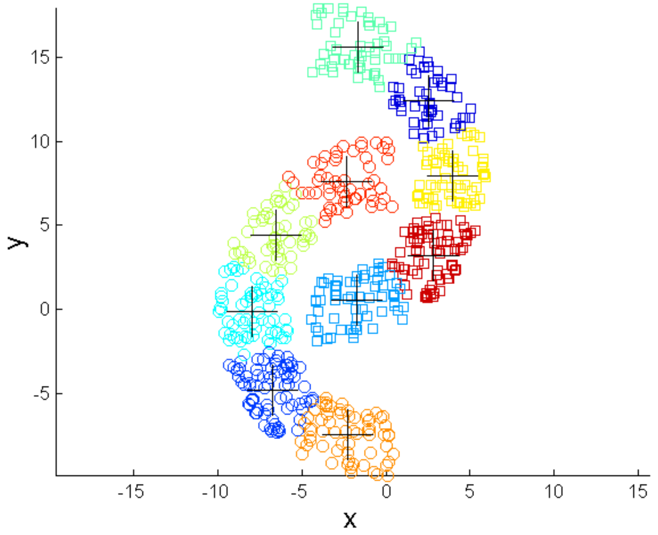


k-Means Clusters

Overcoming k-Means Limitations



Original Points

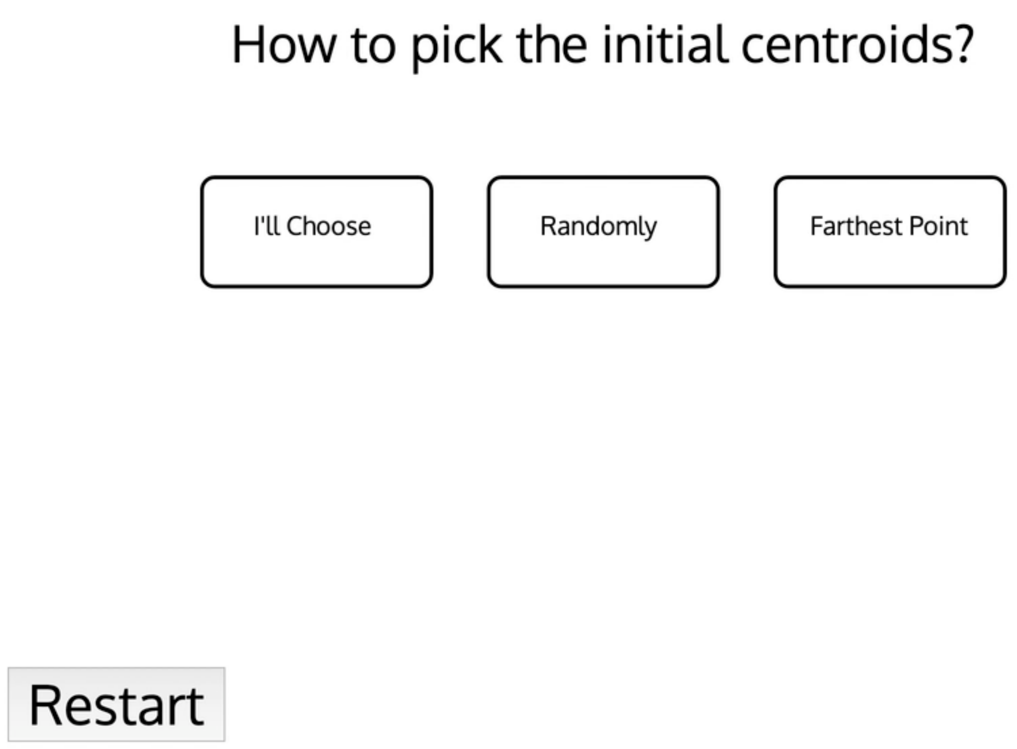


k-Means Clusters

Illustration of k-Means using a Smiley Face
Moral: Overcome limitations by increasing k

https://www.naftaliharris.com/blog/visualizing-k-Means-clustering/

Illustration of k-Means using a Smiley Face
Moral: Overcome limitations by increasing k



How to pick the initial centroids?

https://www.naftaliharris.com/blog/visualizing-k-Means-clustering/

Conclusions

Discussion

Strengths:

Easy, explainable

Widely applicable

Easily updated, automatically adapts to changing data

Weaknesses:

Need entire dataset to classify new instance: difficult to search and to store

Heavy reliance on distance function or scaling of original data

What Cluster Analysis is Not…

Supervised classification

- Have class label information

Simple segmentation

- Dividing students into different registration groups alphabetically, by last name

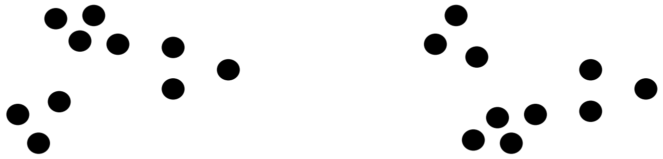
Results of a query

- Groupings are a result of an external specification

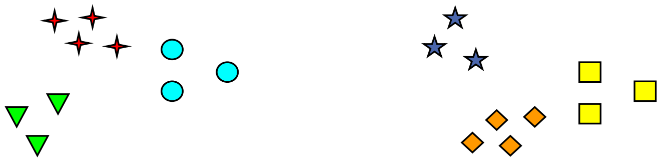
Graph partitioning

- Some mutual relevance and synergy, but areas are not identical

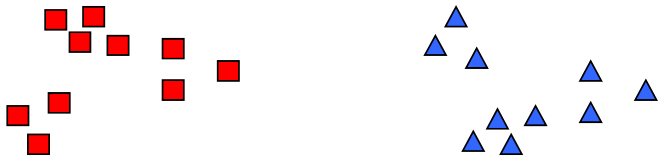
Notion of a Cluster can be Ambiguous



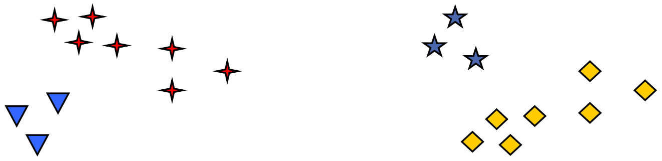
How many clusters?



Six Clusters



Two Clusters



Four Clusters

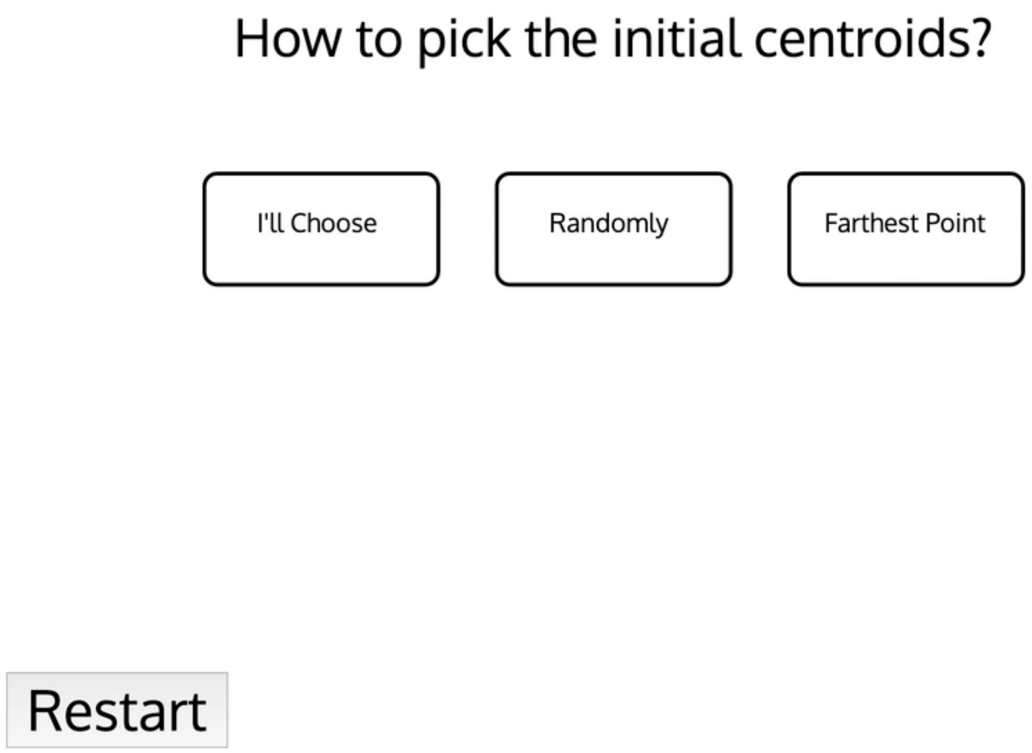
Illustration of k-Means using Uniform Data

Moral: k-Means finds a structure even if your data lacks one

https://www.naftaliharris.com/blog/visualizing-k-Means-clustering/

Illustration of k-Means using Uniform Data

Moral: k-Means finds a structure even if your data lacks one



How to pick the initial centroids?

https://www.naftaliharris.com/blog/visualizing-k-Means-clustering/

Clustering

The science part of cluster analysis is the algorithm, the art of applying cluster analysis applying the various methods to find a scheme that yields “good” clusters

Cluster Analysis reduces large datasets with millions of records to a small number of prototypical records

Clustering is unsupervised, since there is not “solution” the appropriateness is largely at the discretion of the analyst

Final Comment on Cluster Validity

“The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.”

Algorithms for Clustering Data, Jain and Dubes