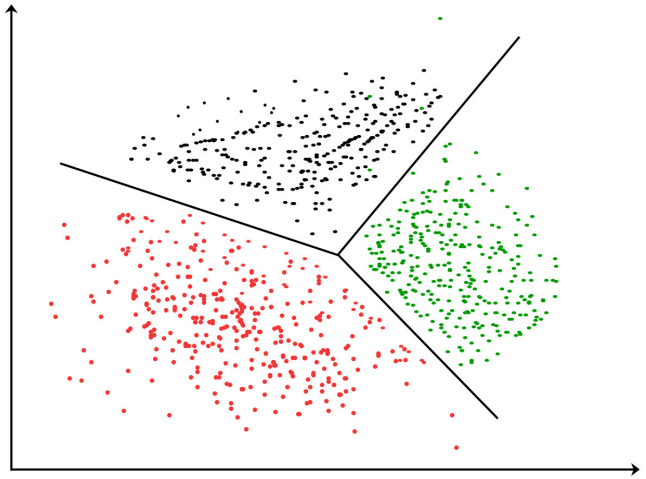
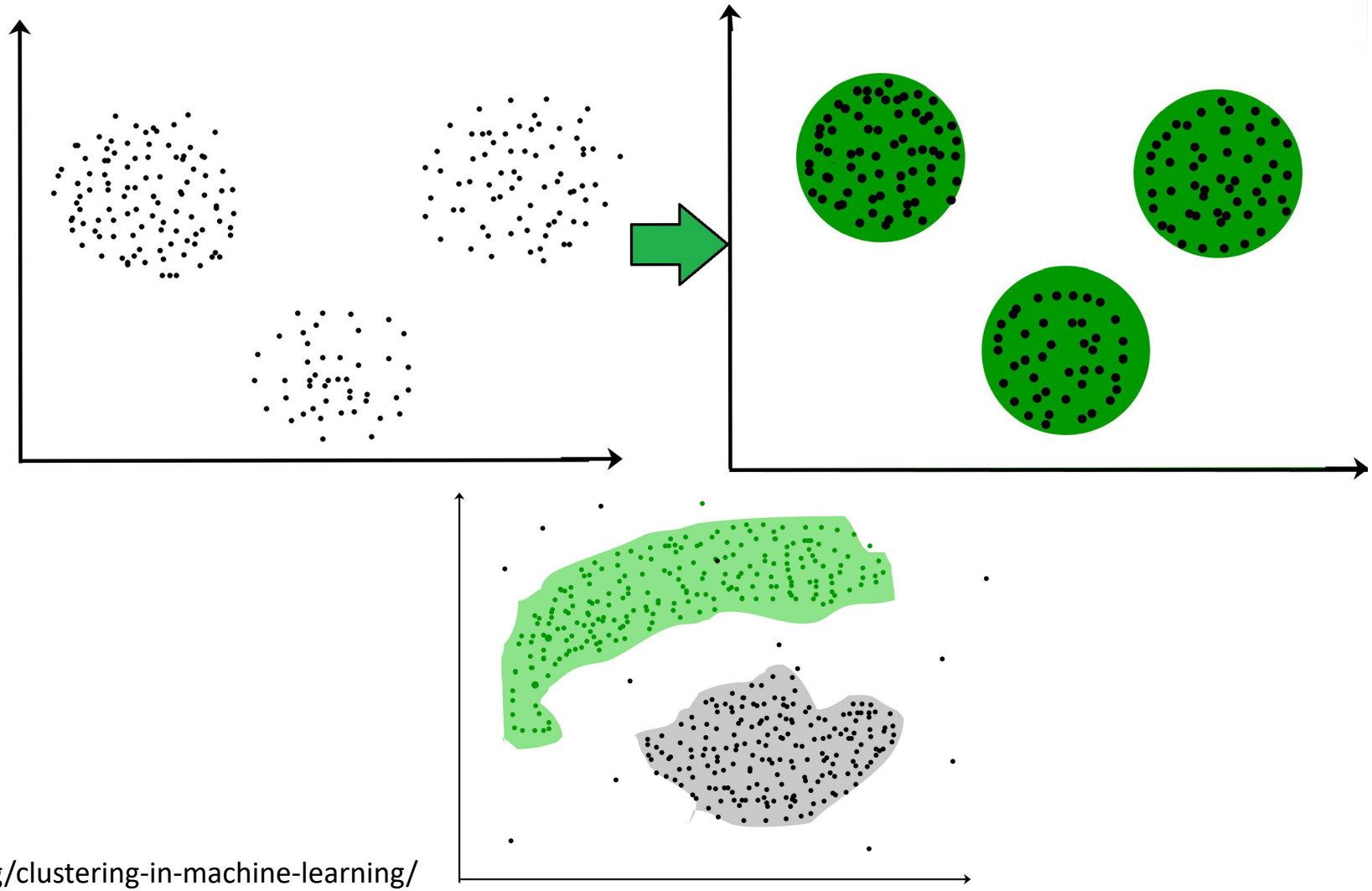
Clustering and Clustering Algorithms



What is Clustering?



geeksforgeeks.org/clustering-in-machine-learning/

What is Clustering?

Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

Important Points

- We are often given a (very) large set of data
- We would like to group the data into clusters so that elements in a cluster are similar to other elements in the same cluster but dissimilar to elements in other clusters.
- We assign labels to the clusters after the clusters are formed
- As CS folks, it is often our task to form the clusters and let a subject matter expert assign labels

Why Clustering?

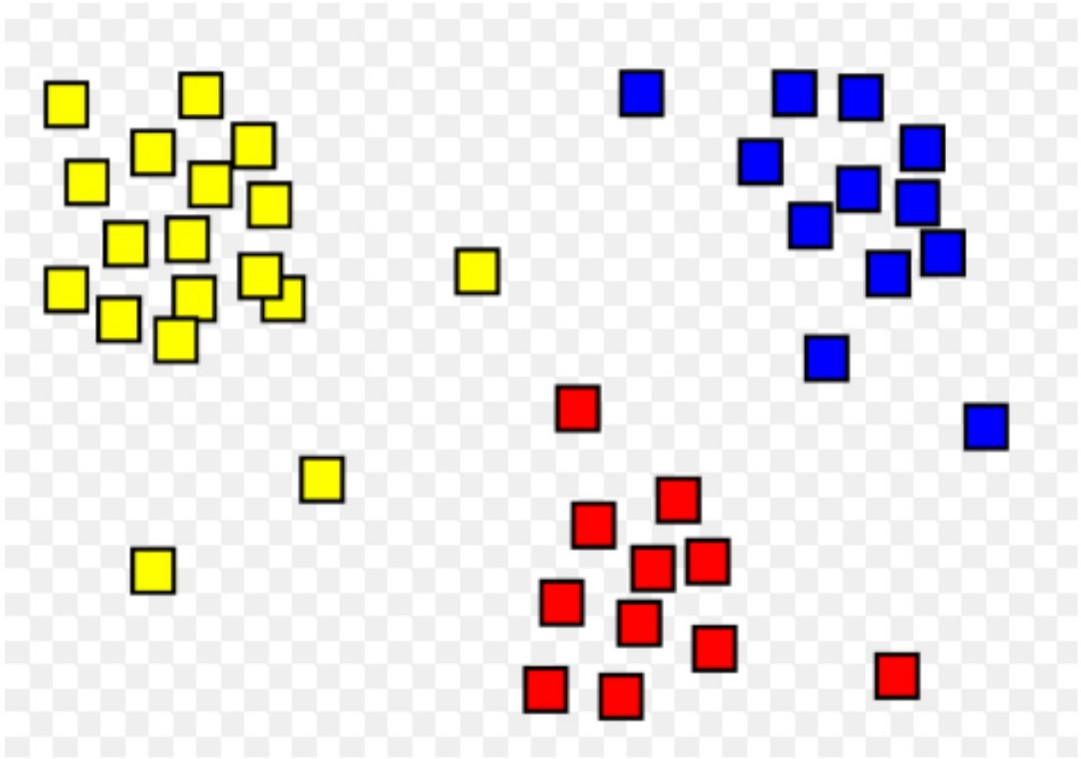
Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.

https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

An Example

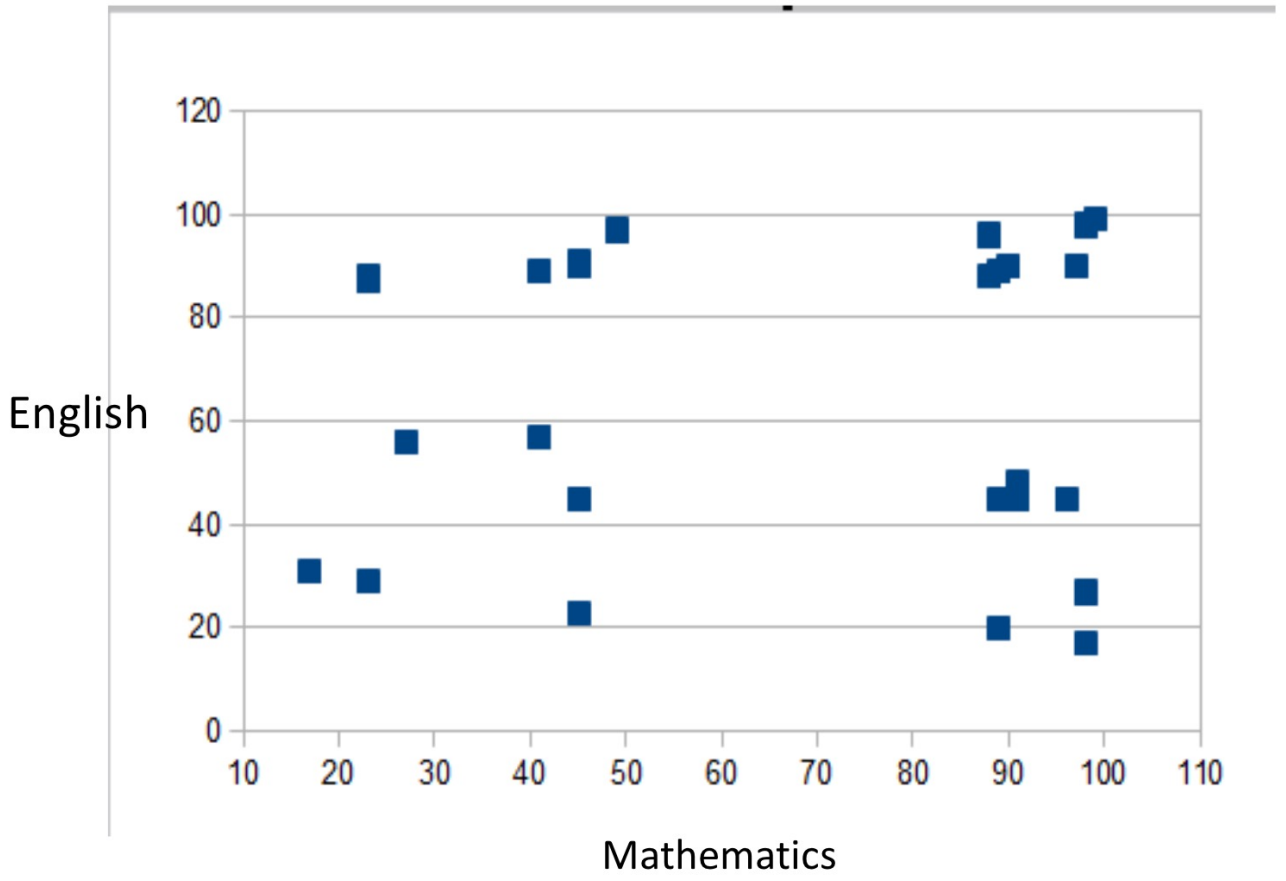
- Suppose that a hospital collects two medical tests on its patients and that each test some how is translated into a number.
- We can do a scatter plot in the x-y plane of the lab results
- A subject matter expert may be able to look at the cluster formed and assign labels (diseases, conditions, etc) to the clusters



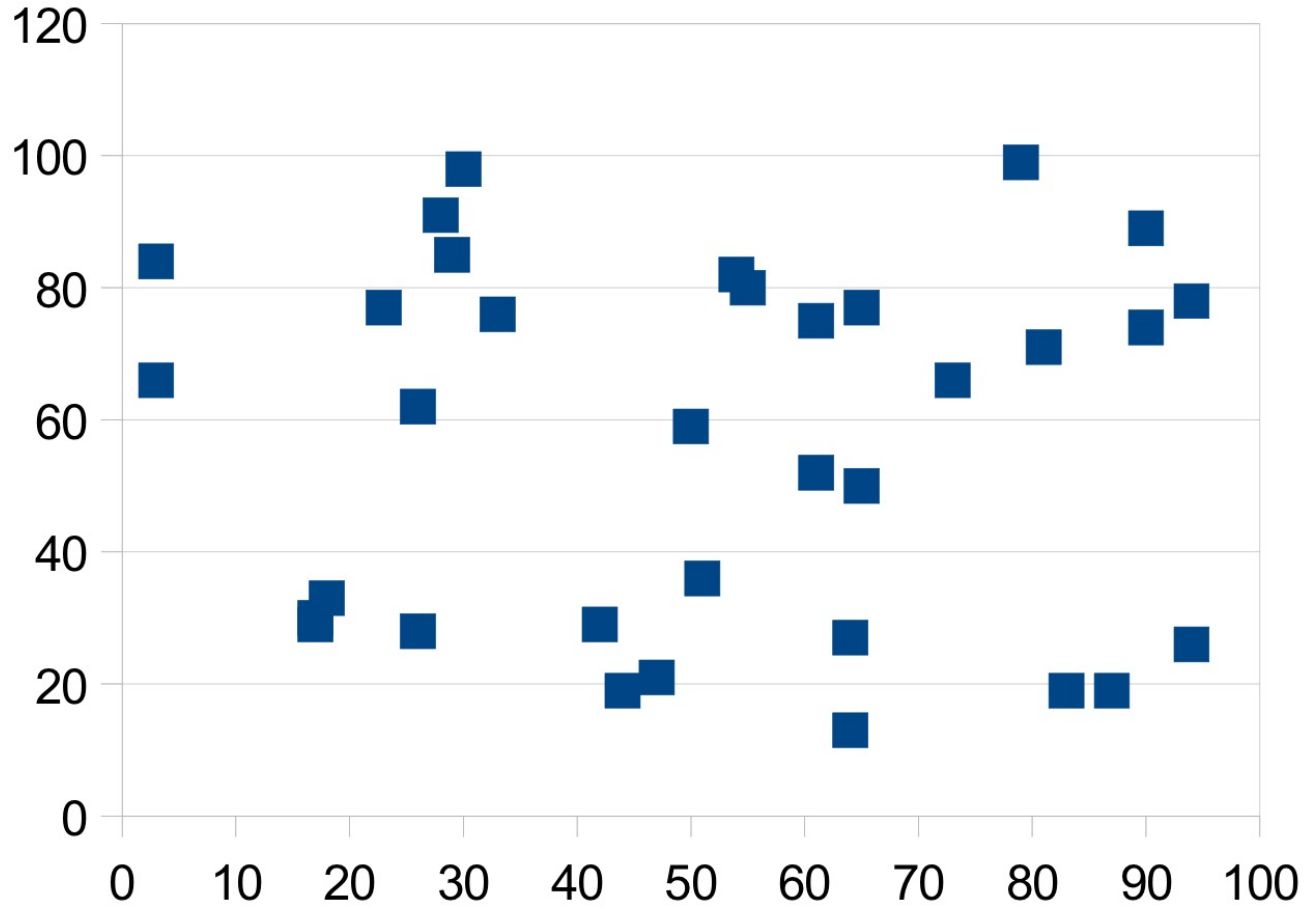
An Example (Continued)

- **The example shows that clustering takes two steps**
  - **Formation of the clusters. This is done by well known mathematical techniques and is usually done by technical people like you**
  - **Interpreting the clusters. This is usually done by subject matter experts. For example**
    - **On the previous slide, a doctor or nurse might say**
      - **All the yellow patients have diabetes**
      - **All the red patients have high blood pressure**
      - **All the blue patients have liver disease**

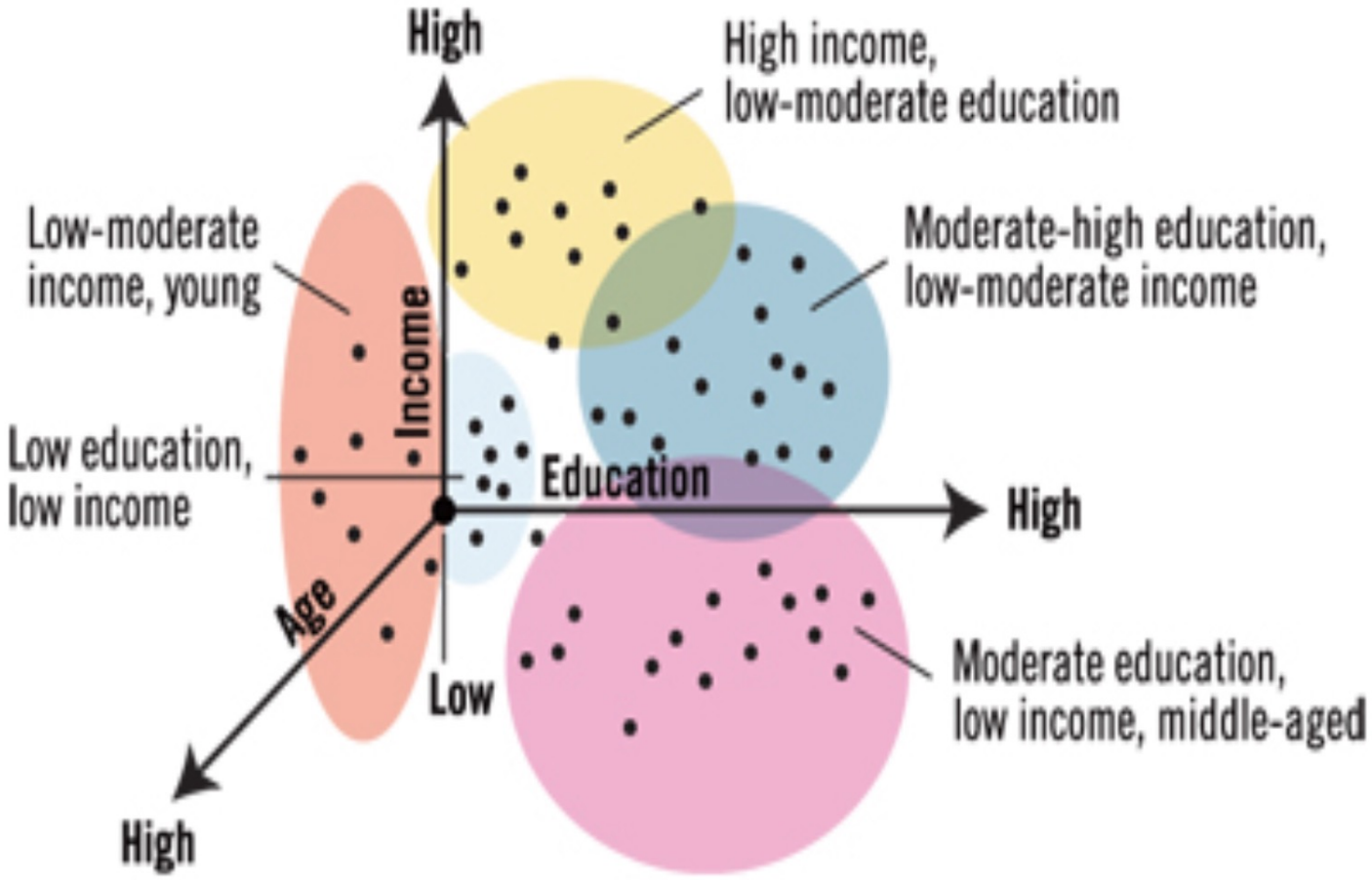
English and Math Scores



Not So Obvious



Income, Age, and Education.



Requirements of clustering

The following points throw light on why clustering is required in data mining −

- **Scalability** − We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

Clustering Video 1

- In the following video, the presenter discusses clustering in general.
- Later we will be concerned with implementing cluster algorithms in a high level language.
- https://www.youtube.com/embed/2QTeuO0C-fY

Important Points Video 1

- The presenter defined clustering
- She presented several diverse applications of clustering
- She presented several types of clustering including
  - Hierarchical
  - Partitioning
- **Hierarchical** clustering is implemented either **bottom up** or **top down**

- In bottom up, each data point begins as its own cluster. We combine clusters until we achieve the desired number of clusters
- In top down clustering, all data points belong to a single cluster. This cluster is repeatedly divided intil a desired number of clusters is chieved.
- **Partitioning** allows a specified number of clusters to be constructed by methods such as k-means
- In CS 620, we will only consider hierarchical and partitioning

Bottom Up and Top Down

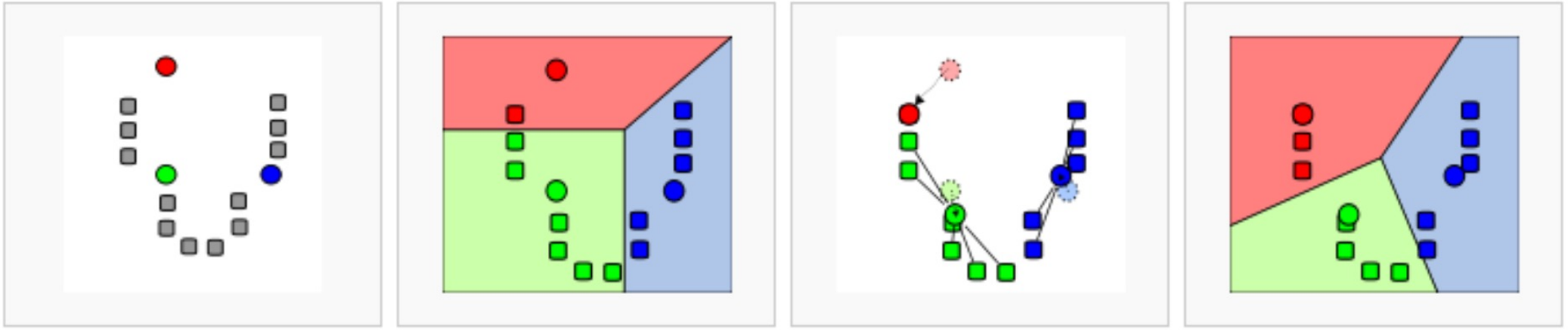
- **In these two types of clustering we have n data points and either**
  - **Start with one cluster and continue until each element is its own cluster**
  - ***OR* start with n clusters and continue until we have one cluster**
  - **Clearly n clusters and 1 cluster are not informative!!**
- **Most often we would**
  - **Stop when a specified number of clusters had been reached or**
  - **Stop when the clusters "make sense"**
- **The *art* of clustering appears again!!**

k-means : a Partitioning Algorithm

- One of the partitioning algorithms is called the k-means algorithm
- The idea is to divide the data into k clusters
- https://www.youtube.com/watch?v=_aWzGGNrcic

Standard algorithm – thanks to Wikipedia

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

Example with playing cards

- In the next video, the presenter demonstrates k-means with k = 3
- He will
  - Take a subset of a deck of cards as the space to cluster
  - Randomly set the initial centroids by selecting 3 more cards from the deck
  - Use the number of spots on the card as the clustering criteria. The distance from a card to a centroid is the difference in the number of spots on each card

https://www.youtube.com/watch?v=zHbxbb2ye3E

Choice of Centroids affect on Clusters

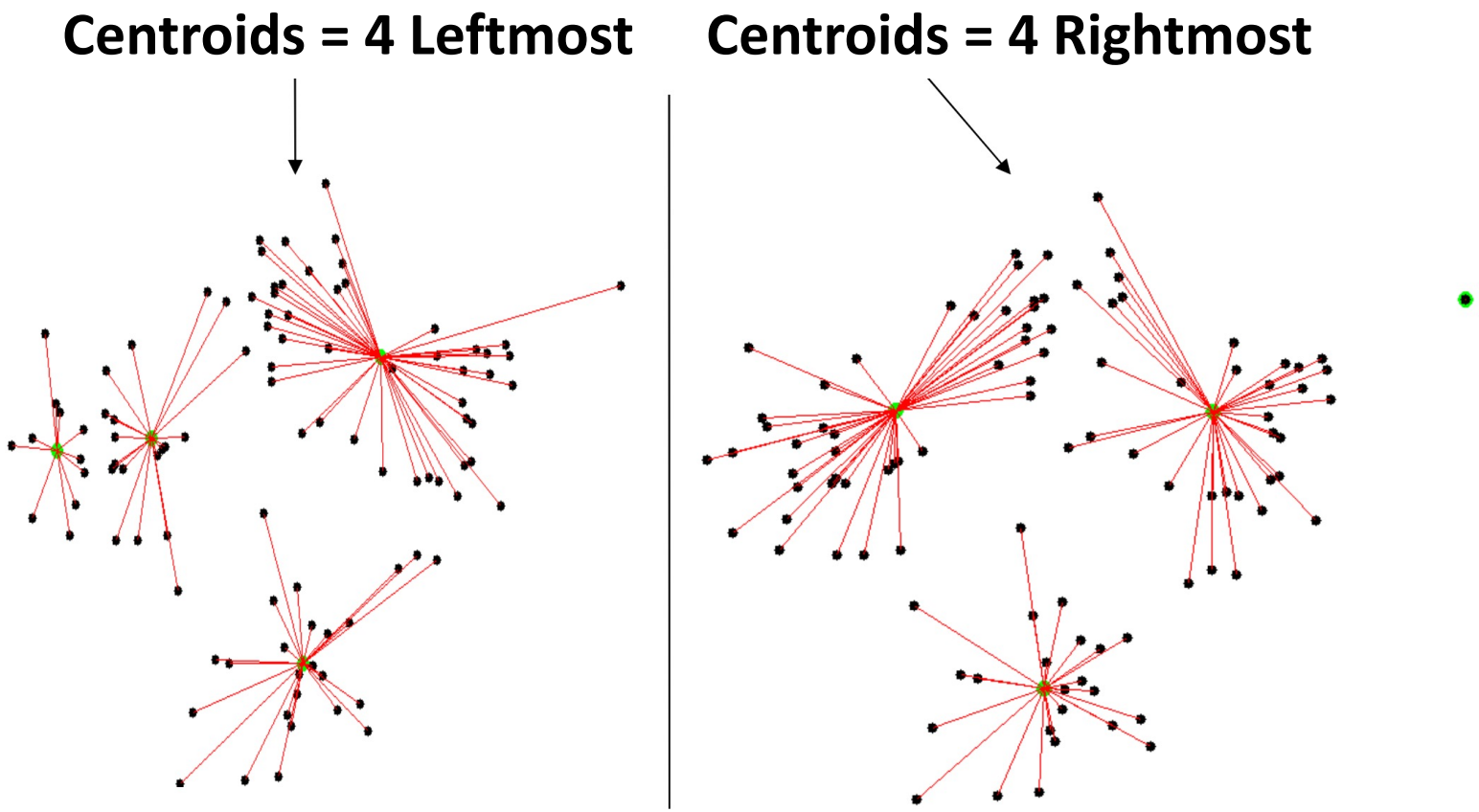
An animation to show how the choice of initial centroids affects the final clusters

- The animation below is provided by Andrey A. Shabalin, a faculty member at Virginia Commonwealth University
- In the video, we will see how a set of 2D points is clustered into k clusters if he chooses the initial 4 centroids to be
  - The 4 leftmost points and then the 4 right most points
  - The 4 topmost points and then the 4 bottom most points

http://shabal.in/visuals/kmeans/1.html

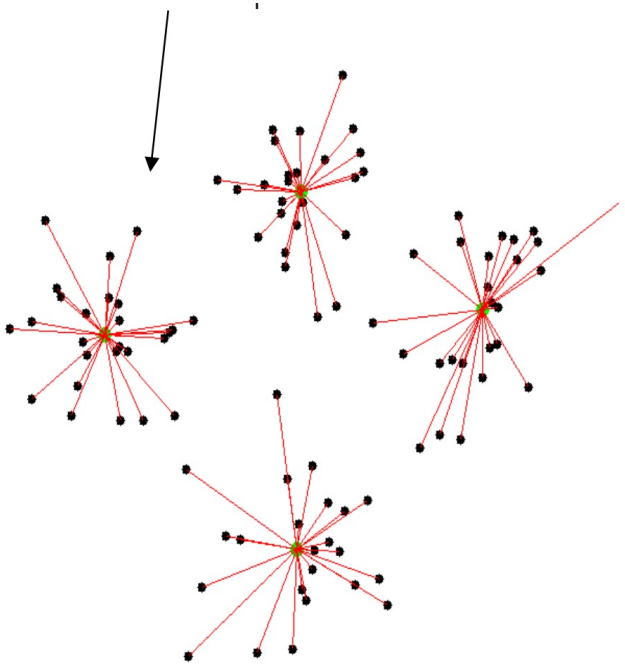
http://shabal.in/visuals.html

Final Clusters, Left and Right

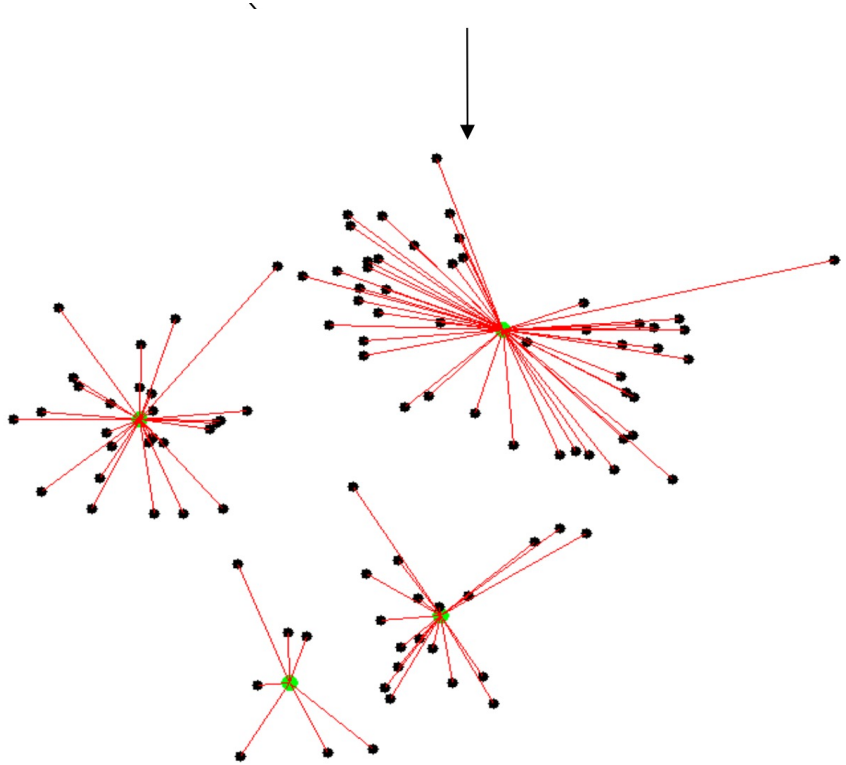


Final Clusters, Top and Bottom

Centroids = 4 Topmost

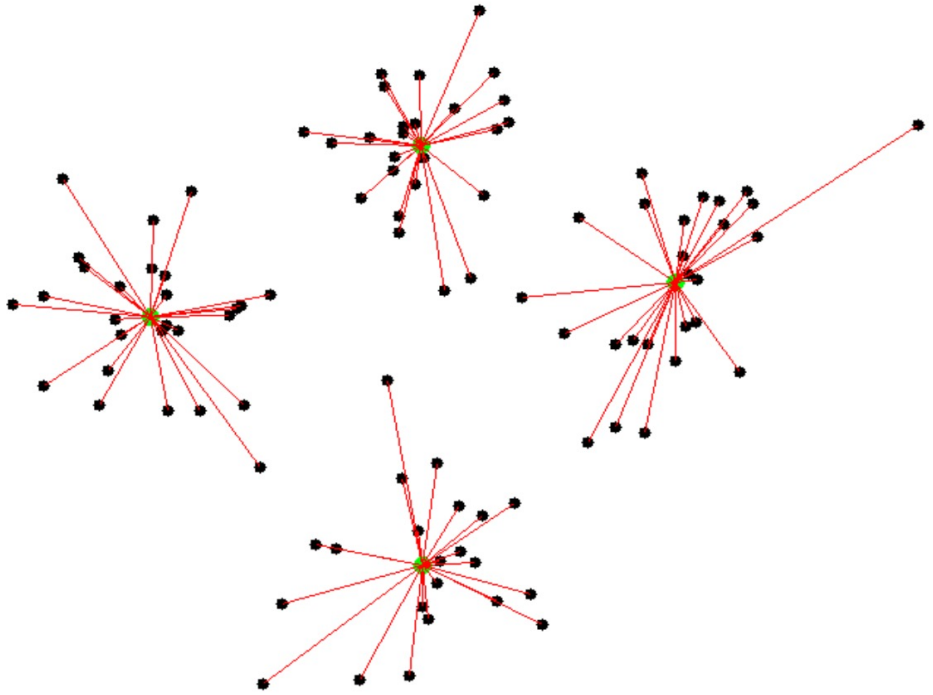


Centroids = 4 Bottommost



Final Clusters, Random

Centroids Chosen Randomly



Centroid Selection

As the examples show, the choice of initial centroids can greatly affect the final result

Conventional wisdom is that we should choose random starting centroids that are representative of the data

Example:

Suppose that our data consists of pairs of numbers (x,y)

Suppose that we inspect the data and observe that x ranges from xmin to xmax and y ranges from ymin to ymax

We could choose the initial centroids so that

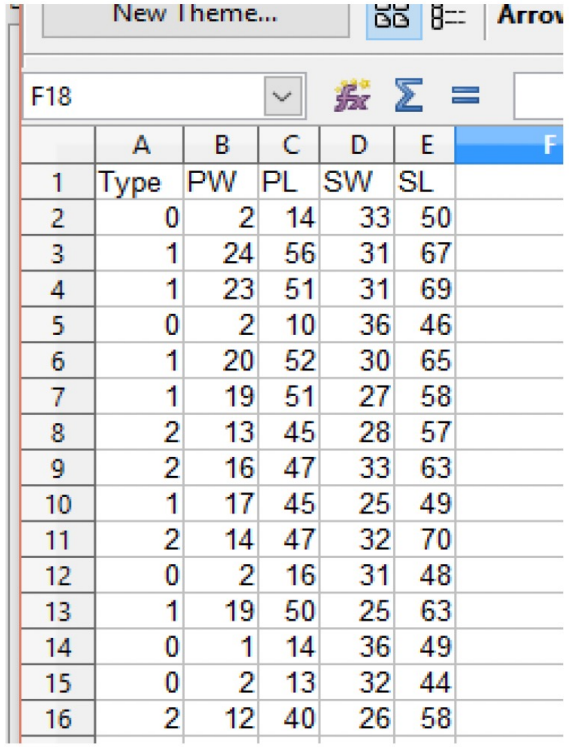
– each x is randomly chosen in the interval [xmin, xmax]

– each y is randomly chosen in the interval [ymin, ymax]

k-means in higher dimensions

- Much data used in clustering has more than just x and y coordinates. In fact, we can envision data sets where each data element has scores or even hundreds of coordinates.
- A well known data set used in statistics, AI, and other places is called Fisher's Iris Data
- This data contains 150 observations and each observations has 5 coordinates
- For each observation, the following are recorded
  - coordinate[0]: iris type (0,1, or 2) (not used in clustering)
  - coordinates[1] and [2]: petal length and petal width
  - coordinates[3] and [4]: sepal length and sepal width

16 of the 150 observations from Fisher's Iris Data



	A	B	C	D	E
1	Type	PW	PL	SW	SL
2	0	2	14	33	50
3	1	24	56	31	67
4	1	23	51	31	69
5	0	2	10	36	46
6	1	20	52	30	65
7	1	19	51	27	58
8	2	13	45	28	57
9	2	16	47	33	63
10	1	17	45	25	49
11	2	14	47	32	70
12	0	2	16	31	48
13	1	19	50	25	63
14	0	1	14	36	49
15	0	2	13	32	44
16	2	12	40	26	58

- Column A is a numeric code for the iris type. 0 is type Setosa, 1 is type Versicolor, and 2 is type Virginica
- Columns B,C,D, and E contained the measured data of petal width (PW), petal length (PL), sepal width (SW) and sepal length (SL)
- We would use columns B,C,D, and E to cluster the data.
- Column A would help us name the clusters
- If we are ***successful*** in our clustering, we would expect all or most of the Setosa irises to be in a cluster, all or most of the Versicolor irises to be in a second cluster and all or most of the Virginica irises to be in a 3rd cluster

Output of clustering using SAS (Statistical Analysis System)

Table of CLUSTER by Species				
CLUSTER	Species			
	Setosa	Versicolor	Virginica	Total
1	0	49	15	64
2	0	1	35	36
3	50	0	0	50
Total	50	50	50	150

SAS is a widely used statistical and business analytics software. R is also widely used for the same things.

Using 3 clusters, a subject matter expert might reasonably conclude

- Cluster 3 represents Setosa
- Cluster 2 represents Virginica
- Cluster 1 is largely Versicolor

The subject matter expert might also reasonably conclude that it is reasonably accurate to identify irises by the four measurements indicated in the earlier slides.

K-means Clustering in C/Java/Python

To perform k-means clustering in C/Java/Python, we would follow these steps

- Decide how many clusters the user wants, say k
- Obtain n data points with m coordinates and one classification.
  - This is the set of data we want to cluster
- Initialize the centroid of each of the k clusters
  - We have discussed at least two methods to do this
- Iterate: while done == false
  - Save old centroids
  - for each element x of the input data set
    - Find the closest clusterd to x – say cluster j
    - Add x to the list[j] of data elements closest to cluster j
  - Re-compute centroids for i=0,1,2,...,k-1
  - if new centroids = old centroids, done = true

To demonstrate final clustering

For each j = 0,1,2,...,k-1

print classification of each element in list[j]

- If we obtain a clustering such as the on the right from SAS, we might feel that the clustering is informative.

Table of CLUSTER by Species				
	Species			
CLUSTER	Setosa	Versicolor	Virginica	Total
1	0	49	15	64
2	0	1	35	36
3	50	0	0	50
Total	50	50	50	150

Using Clusters

A previous Instructure had used clustering like k-means in the field of target recognition (Very old research!!). The following is a simplified description of how the clusters were used in a research project.

- Given
  - A set of 1000 images of aircraft. With about 330 images of three types of aircraft
    - Fighter
    - Transport
    - Bomber

From the set of photographs

- Image processing was used and three features were extracted for each of the thousand photographs.
- The data were clustered into three groups using a method similar to k-means
- Suppose that (for argument's sake) that the final cluster centroids were

Fighter	25	45	80
Bomber	10	10	5
Transport	45	65	40

- Suppose that a photograph of an unknown aircraft was presented, and its three features were [27, 48, 90]
- What is a reasonable aircraft type to assign to this unknown aircraft?

Patient Data

- Suppose that a hospital clusters patients based on 10 lab readings and 5 clusters are formed
  - Healthy
  - Cancer
  - Diabetes
  - Heart Disease
  - Kidney Disease

- If a new patient arrives and his/her lab readings, what is a reasonable to use the clustering to obtain a quick diagnosis of his/her problem?

- How could you use the iris data and clustering to identify irises of unknown type?