CSC 1002 Week 13

Machine Learning - kMeans & kNN

Q & A

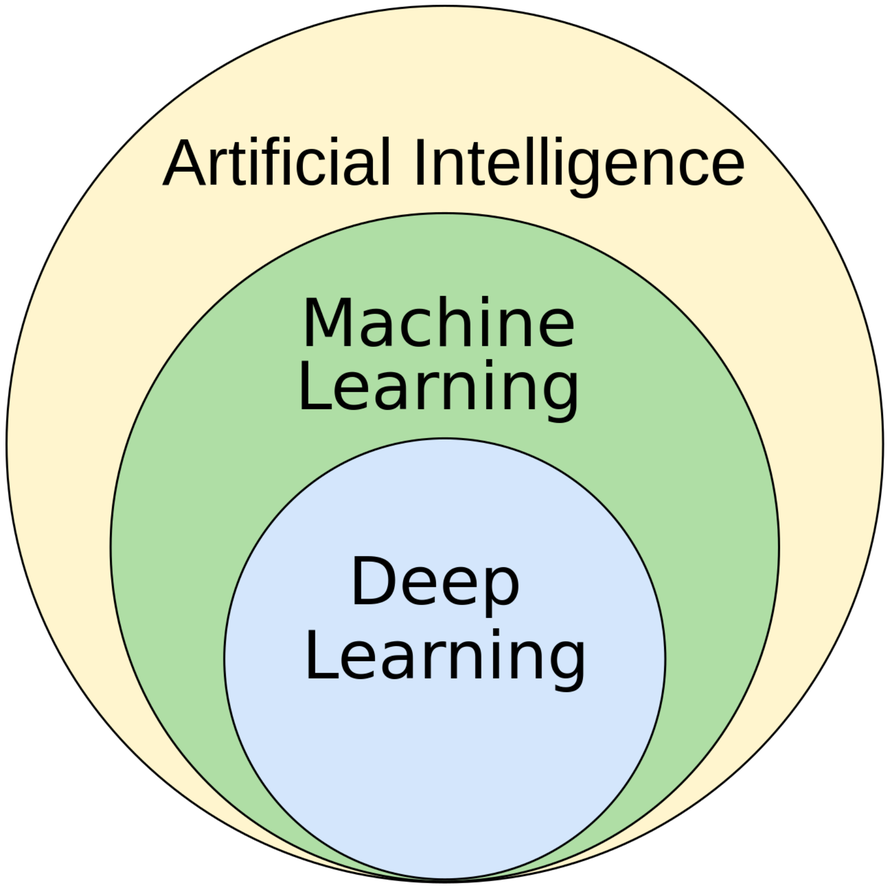
Assignment 3 - Snake

Machine Learning

kNN & kMeans

Machine Learning - Wikipedia

- “Wikipedia – …. ML gives computers the ability to perform tasks without explicit instructions …. using certain algorithms …”
- Popular Fields: Natural Language Processing (nlp), computer vision, speech recognition
- Approaches (high-level)
  - Supervised - learn from sample data and make prediction on new data. Ex: kNN
  - Unsupervised - learn on its own to discover trends and structures in the data. Ex: kMeans
  - Reinforcement - learn on feedback and rewards



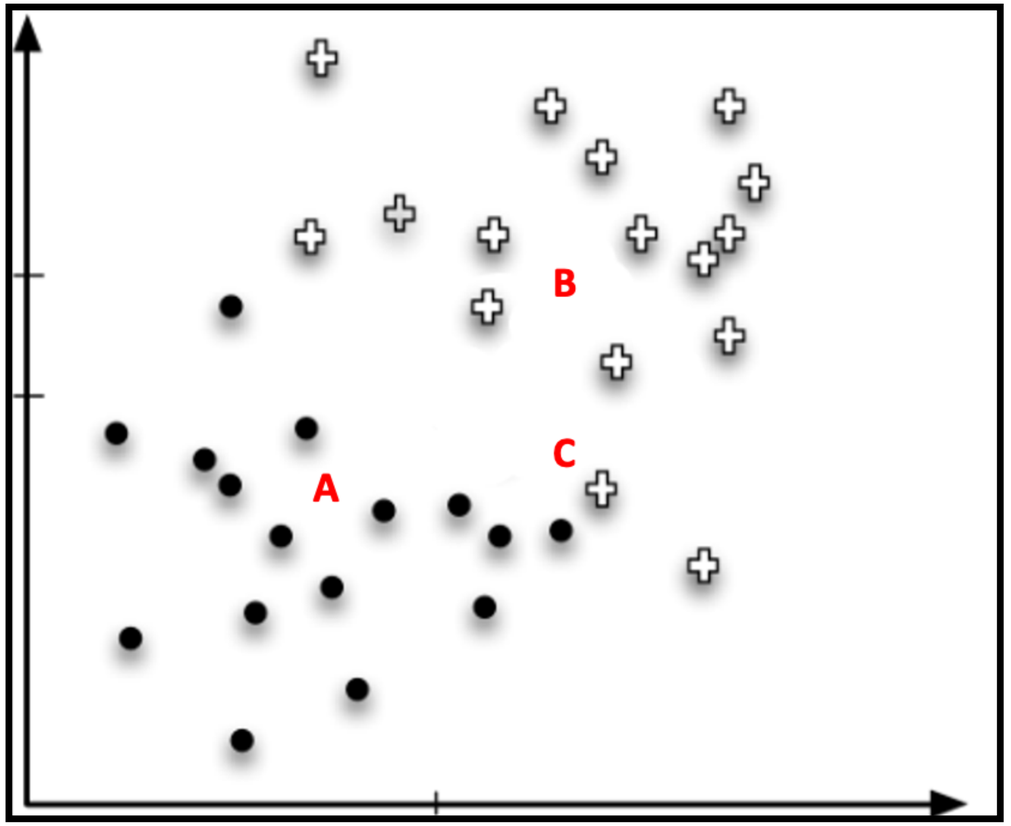
kNN

Supervised Learning

Supervised Learning

kNN - k Nearest Neighbors

- make a prediction based on neighbors
- find the k nearest neighbors
- neighbor is determined by simple notion of distance
- predict result based on majority vote of neighbors
- simple but expensive in computation

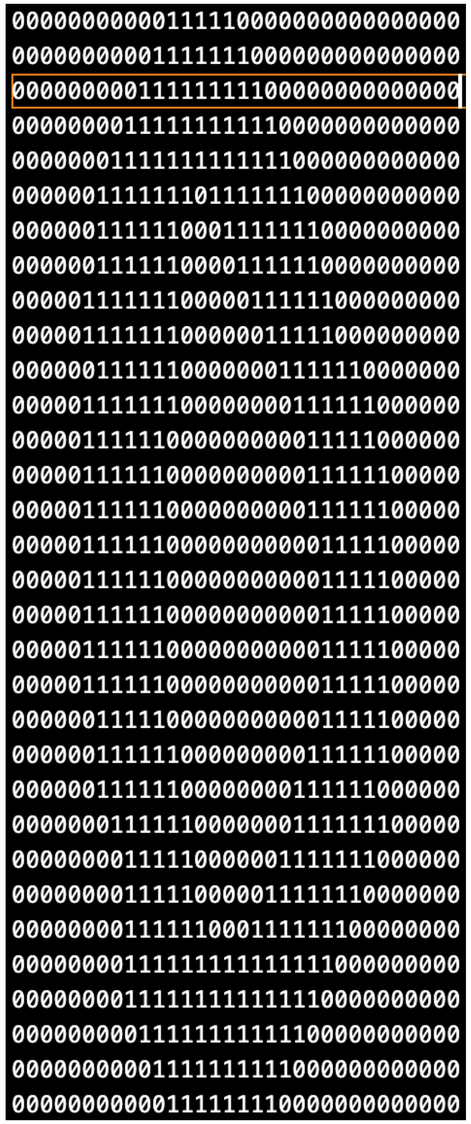


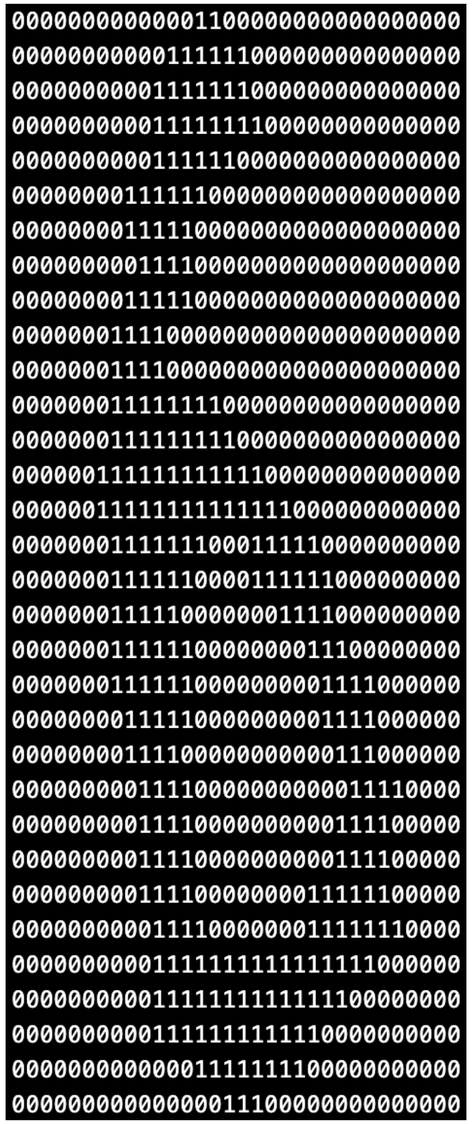
Supervised Learning - hand-written digit

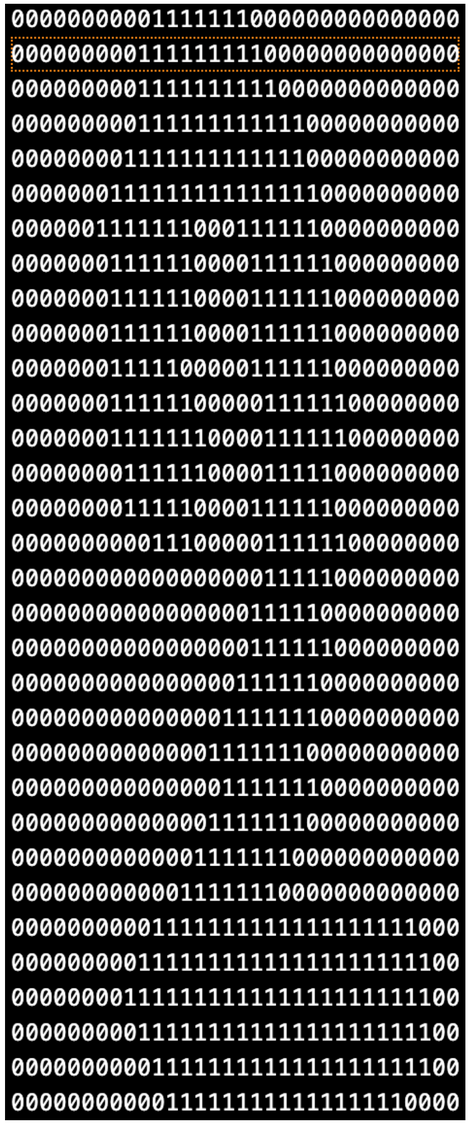
kNN - k Nearest Neighbors

Online demo : https://henryjin.dev/demo/mnist/

Sample Data


00000000000011111000000000000000
00000000001111111000000000000000
00000000011111111110000000000000
00000001111111111111000000000000
00000011111110111111100000000000
00000111111100011111110000000000
00000111111000011111110000000000
00001111111000001111111000000000
00001111111000000111111000000000
00001111111000000011111100000000
00001111111000000001111110000000
00001111111000000000111111000000
00001111111000000000111111000000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00001111111000000000011111100000
00000111111000000000111111000000
00000111111000000001111111000000
00000011111100000011111110000000
00000001111111111111111000000000
00000000111111111111110000000000
00000000011111111111100000000000
00000000001111111111000000000000
00000000000111111110000000000000

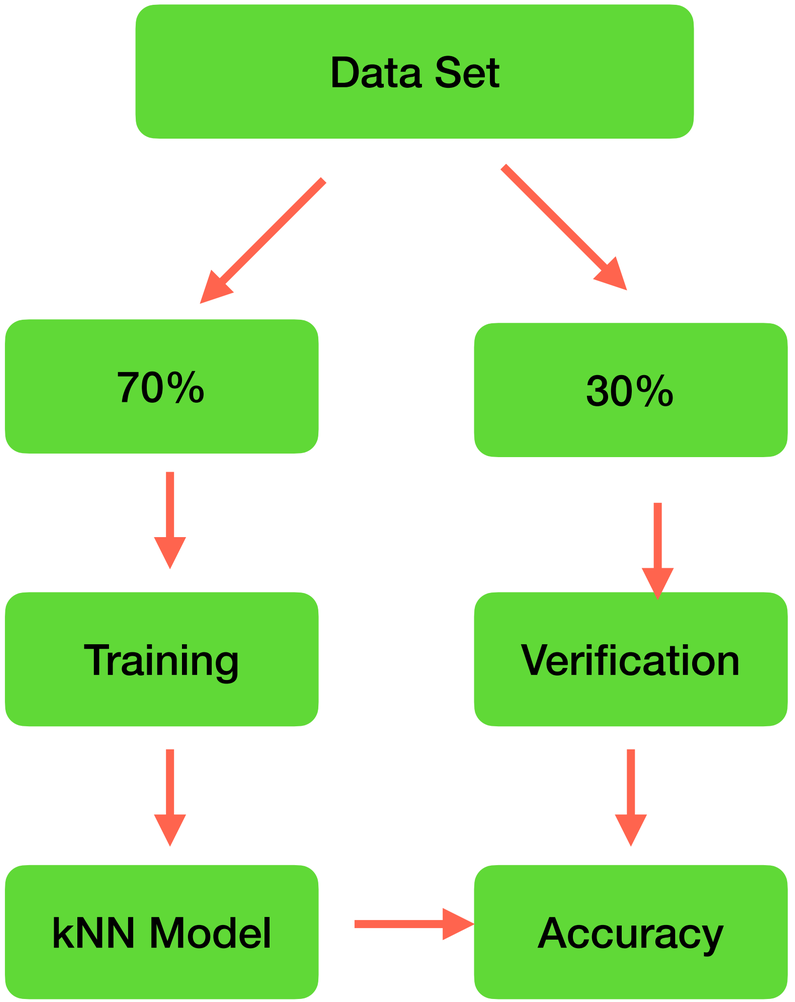

00000000000001100000000000000000
00000000000011111000000000000000
00000000001111111000000000000000
00000000011111111000000000000000
00000000111111110000000000000000
00000001111111000000000000000000
00000001111110000000000000000000
00000011111100000000000000000000
00000011111000000000000000000000
00000111111000000000000000000000
00000111110000000000000000000000
00001111111000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00001111110000000000000000000000
00000111110000000000000000000000
00000111110000000000000000000000
00000011111000000000000000000000
00000001111111111111111000000000
00000000111111111111110000000000
00000000011111111111100000000000
00000000001111111111000000000000
00000000000111111110000000000000
00000000000011111100000000000000
00000000000001110000000000000000


00000000000111111100000000000000
00000000001111111110000000000000
00000000011111111110000000000000
00000000111111111111000000000000
00000001111111111111100000000000
00000011111111111111110000000000
00000111111100011111110000000000
00000111111000011111110000000000
00001111111000001111111000000000
00001111111000000111111000000000
00001111111000000011111100000000
00001111111000000001111110000000
00001111111000000000111111000000
00001111111000000000111111000000
00000111111000000000011111100000
00000111111000000000011111100000
00000011111000000000011111100000
00000000000000000001111100000000
00000000000000000011111000000000
00000000000000000111110000000000
00000000000000001111110000000000
00000000000000011111100000000000
00000000000000111111000000000000
00000000000001111110000000000000
00000000000011111100000000000000
00000000000111111000000000000000
00000000001111111000000000000000
00000000011111111111111111111000
00000000111111111111111111111100
00000000111111111111111111111100
00000000111111111111111111111100
00000000111111111111111111111100
00000000011111111111111111110000

Supervised Learning - Training Data Set

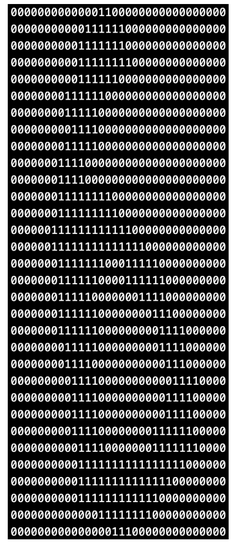
kNN - k Nearest Neighbors

Training




```
19  ----------------------------------------
20                Testing Info
21  ----------------------------------------
22                0 = 187,   2,  99%
23                1 = 192,   6,  97%
24                2 = 194,   1,  99%
25                3 = 197,   2,  99%
26                4 = 176,  10,  95%
27                5 = 177,  10,  95%
28                6 = 193,   2,  99%
29                7 = 198,   3,  99%
30                8 = 164,  16,  91%
31                9 = 181,  23,  89%
32  ----------------------------------------
33          Accuracy = 96.12%
34    Correct/Total = 1859/1934
35  ----------------------------------------
36  End of Training @ 2017-01-23 19:59:12
```

kNN - Prediction



Given Input

Find 10 closest

kNN Model

Predict

6

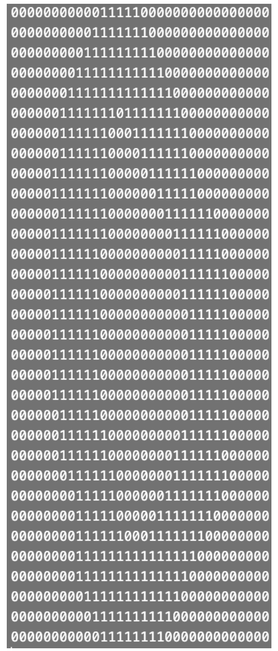
6 x 6 times, 5 x 4 times

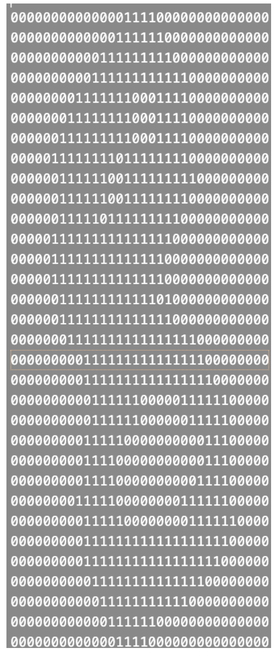
Distance Computation

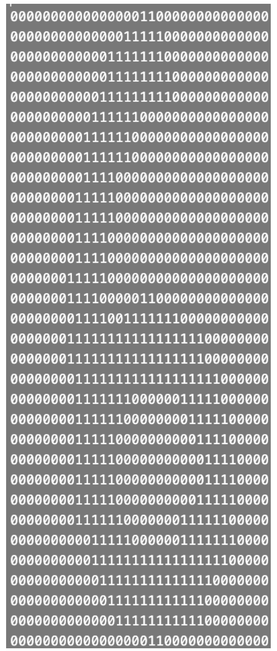
000000000000000011111001010101101010101011111111111111110000

001111110000000011111001010101101010101011111111111111110000

- Based on distance.
  - standard vector equation
  - XOR
- Based on overlapping image
  - And







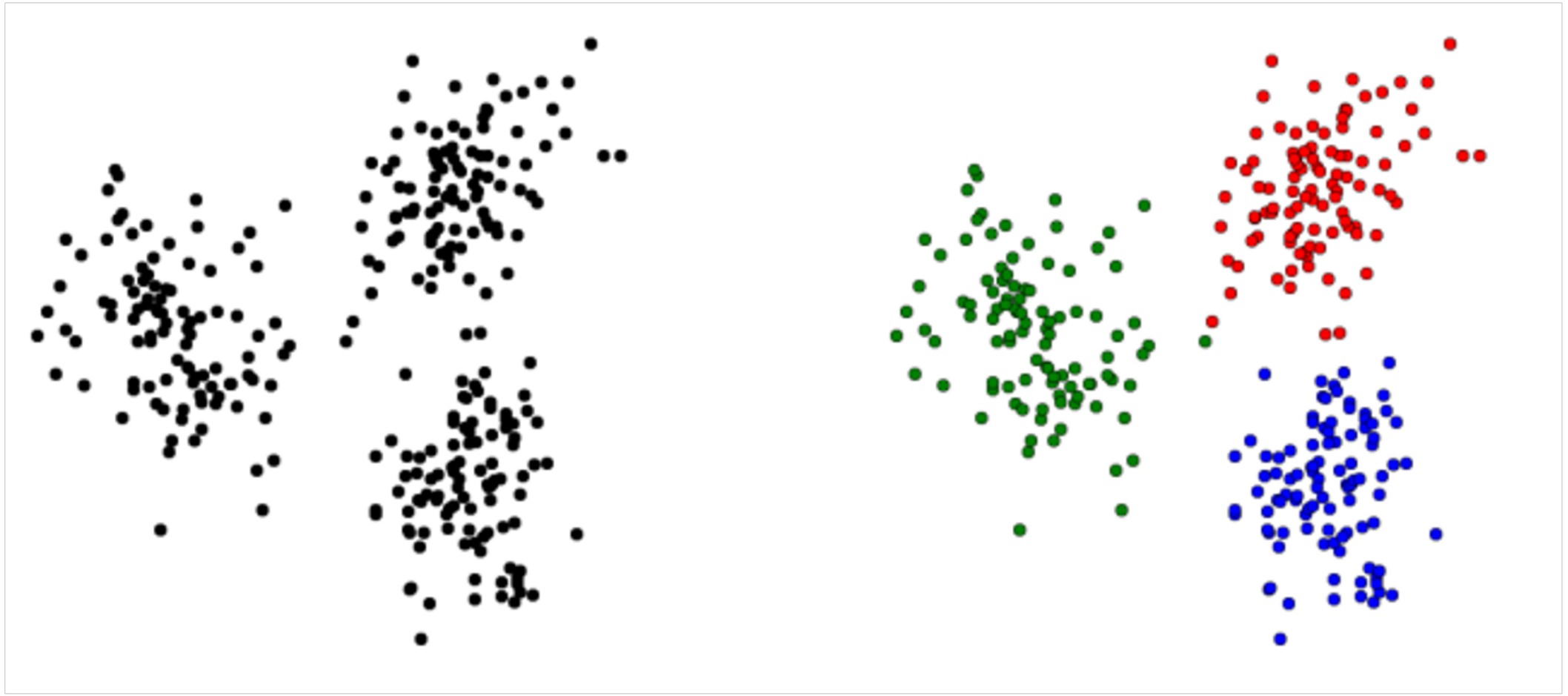
Show Demo with numbers

kMeans

Unsupervised Learning - Clustering

Unsupervised Learning - Clustering

kMeans - partition data into k groups based on similarities



kMeans - Definition

- K-means is an unsupervised learning algorithm for clustering data points
- The variable K represents the number of clusters that need to be created
- It classifies unlabeled data into K clusters based on similarities
- Clusters are non-overlapping
- The algorithm iteratively divides data points into clusters by minimizing the variance in each cluster
-

kMeans - Simple Example - 2 Clusters

Data Set = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Cluster	Data	Mean	Variance
1	1,2,3,4,5	3.0	2.5
2	6,7,8,9,10	8.0	2.5

Cluster	Data	Mean	Variance
1	1,2,3,9,10	5.0	17.5
2	4,5,6,7,8	6.0	2.5

Cluster	Data	Mean	Variance
1	1,3,5,7,9	5.0	10.0
2	2,4,6,8,10	6.0	10.0

kMeans - Image Compression

- kMeans can be used to reduce the size of an image file while maintaining its visual quality.
- This technique involves clustering the pixels in an image into a smaller number of groups
- Then representing each group by its mean color.
- The resulting image will have fewer colors, which reduces the file size, but the overall appearance of the image is still preserved.

kMeans - Image Compression

kMeans - Iterative Approach

- step 0: randomly pick “k” sample data as the initial mean points
- step 1: assign closest mean-point to each sample data
- step 3: if no changes of point assignment, stop and process is done
- step 4: update each group’s mean, and continue from step 1

kMeans - Simple Example - 2 Clusters

Data Set = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Iter	Mean 1	Group 1	Mean 2	Group 2
0	3	1,2,3,4	5	5,6,7,8,9,10
1	2.5	1,2,3,4,5	7.5	6,7,8,9,10
2	3.0	1,2,3,4,5	8.0	6,7,8,9,10

Iter	Mean 1	Group 1	Mean 2	Group 2
0	7	1,2,3,4,5,6,7	9	8,9,10
1	4.0	1,2,3,4,5,6	9.0	7,8,9,10
2	3.5	1,2,3,4,5	8.5	6,7,8,9,10
3	3.0	1,2,3,4,5	8.0	6,7,8,9,10

kMeans - Simple Example - 2 Clusters

Data Set = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

Iter	Mean 1	Group 1	Mean 2	Group 2
0				
1				
2				

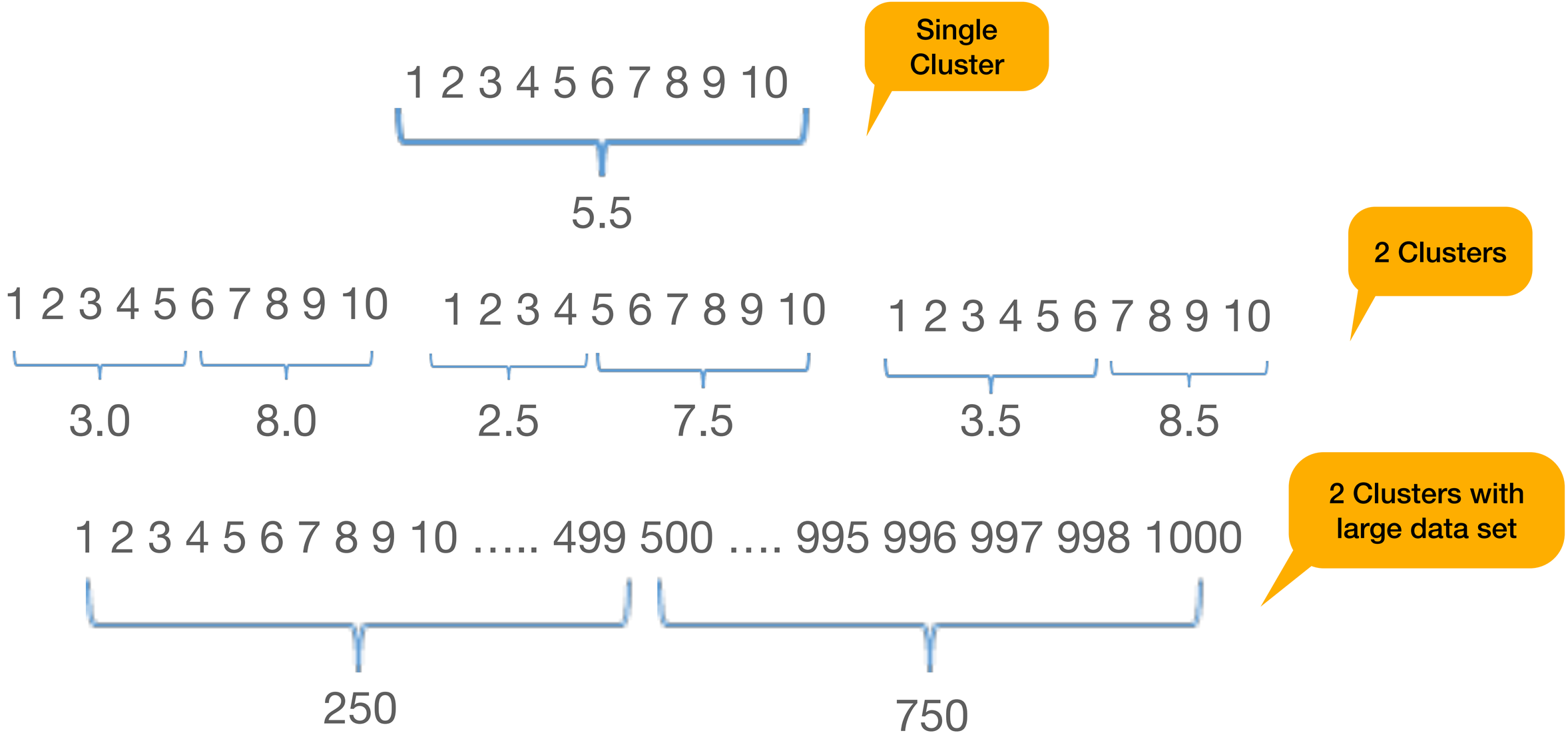
kMeans - Simple Example - 2 Clusters

Data Set = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100

Iter	Mean 1	Group 1	Mean 2	Group 2
0	3	1,2,3,4	5	5,6,7,8,9,10,100
1	2.5	1,2,3,4,5,6,7,8,9,10	20.71	100
2	5.0	1,2,3,4,5,6,7,8,9,10	100.0	100

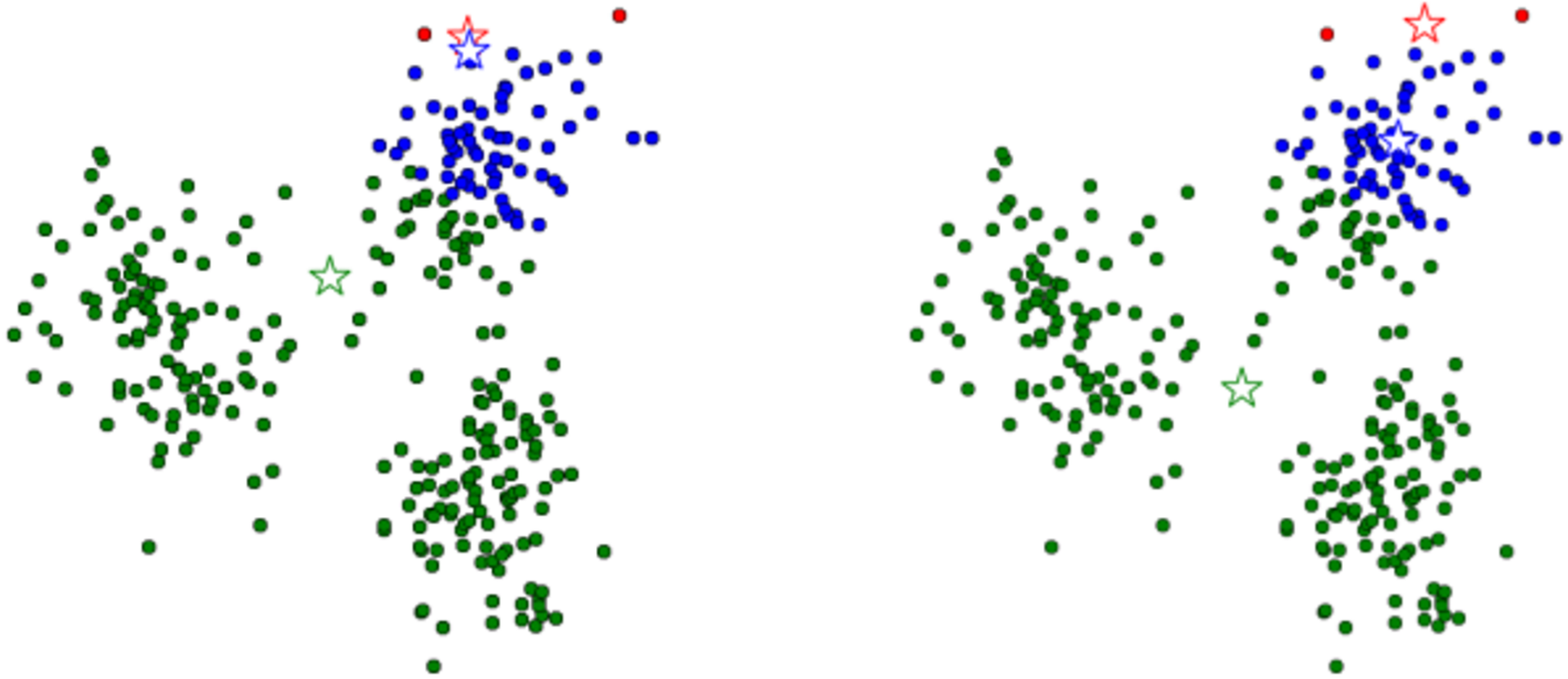
Show Demo with numbers

kMeans - Basic Example

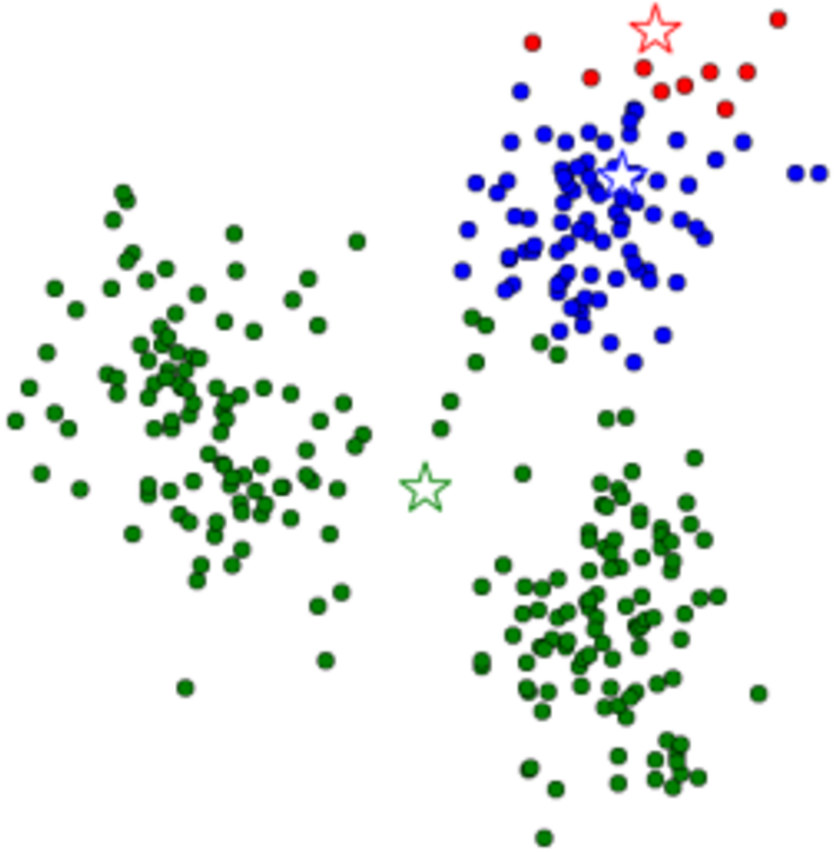


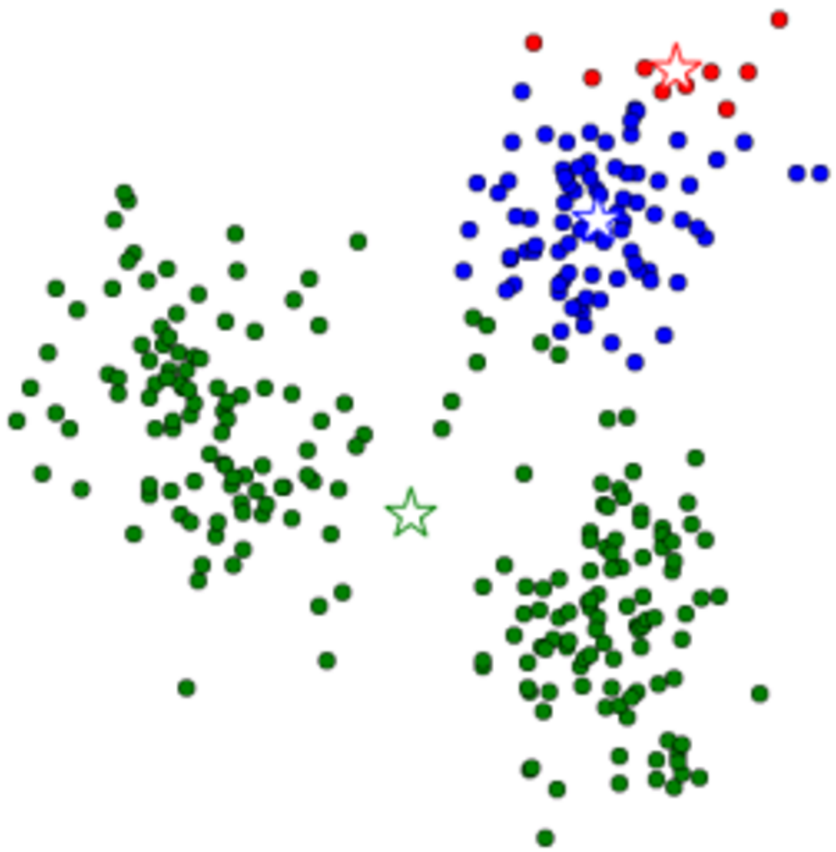
kMeans - Iteration 1 & 2

Initial Setup - Randomly pick 3 samples as the group mean

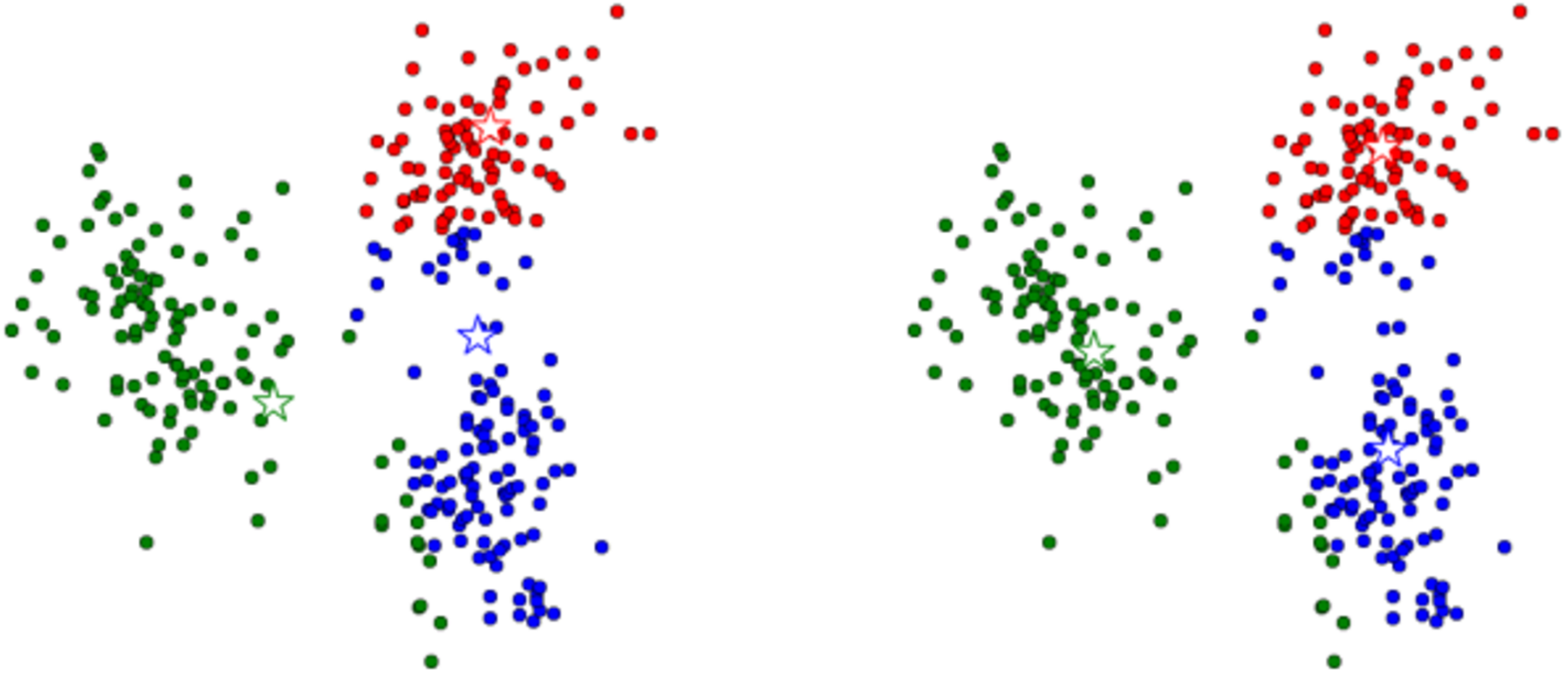


kMeans - Iteration 3 & 4





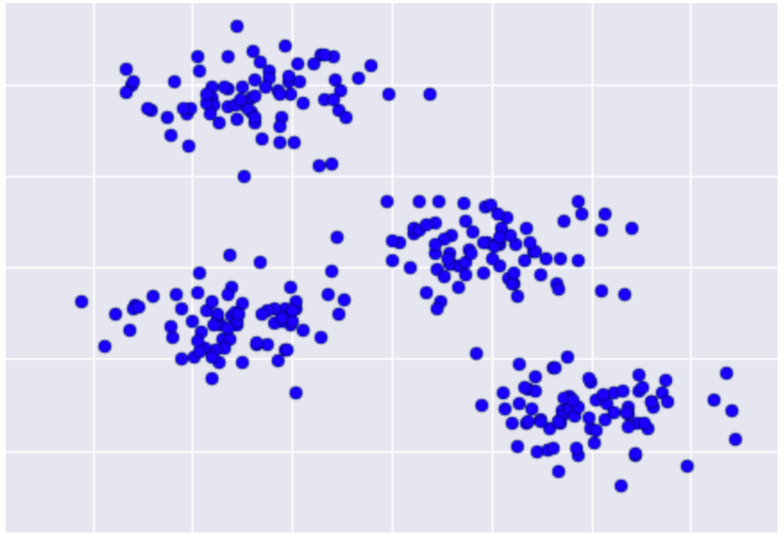
kMeans - Iteration 10

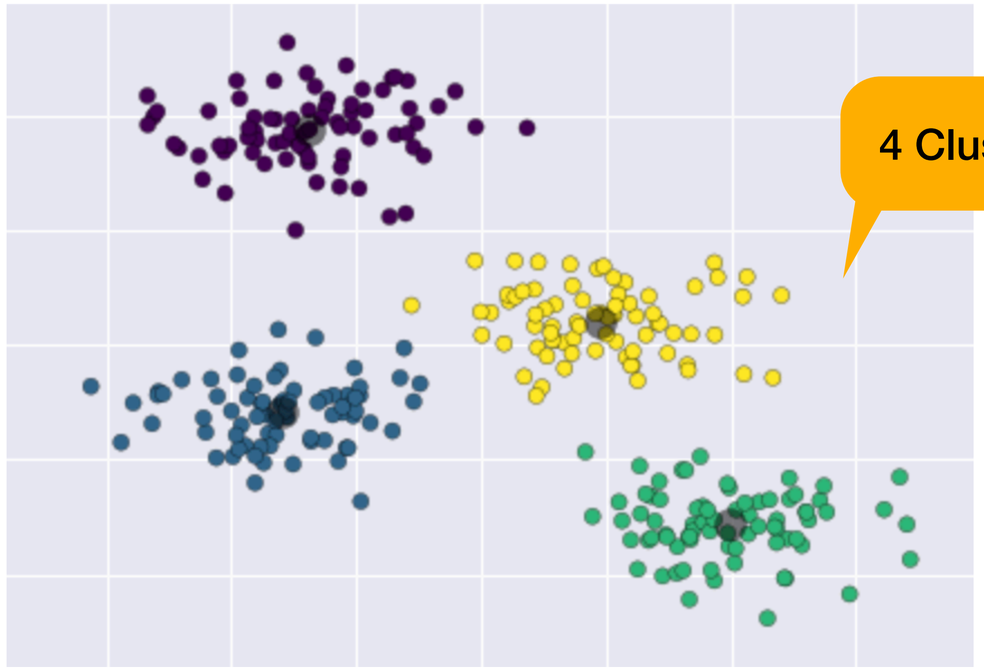


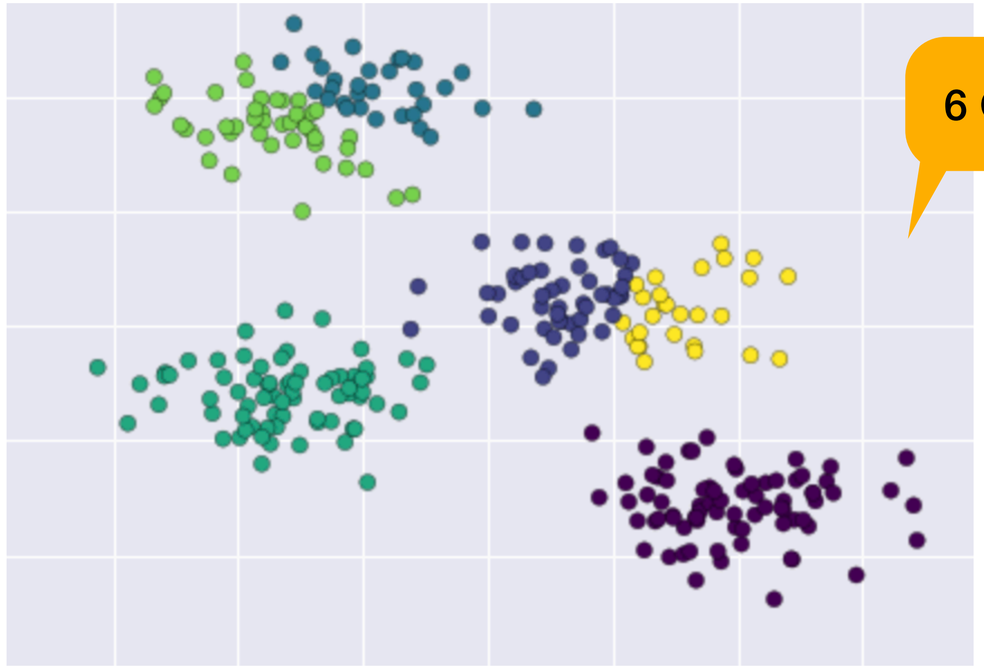
kMeans - Demo

kMeans - Limitation

Cluster Selection

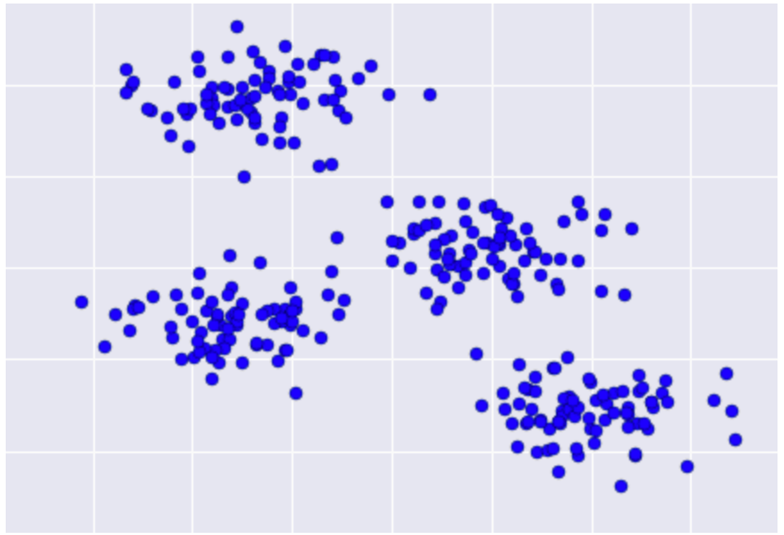


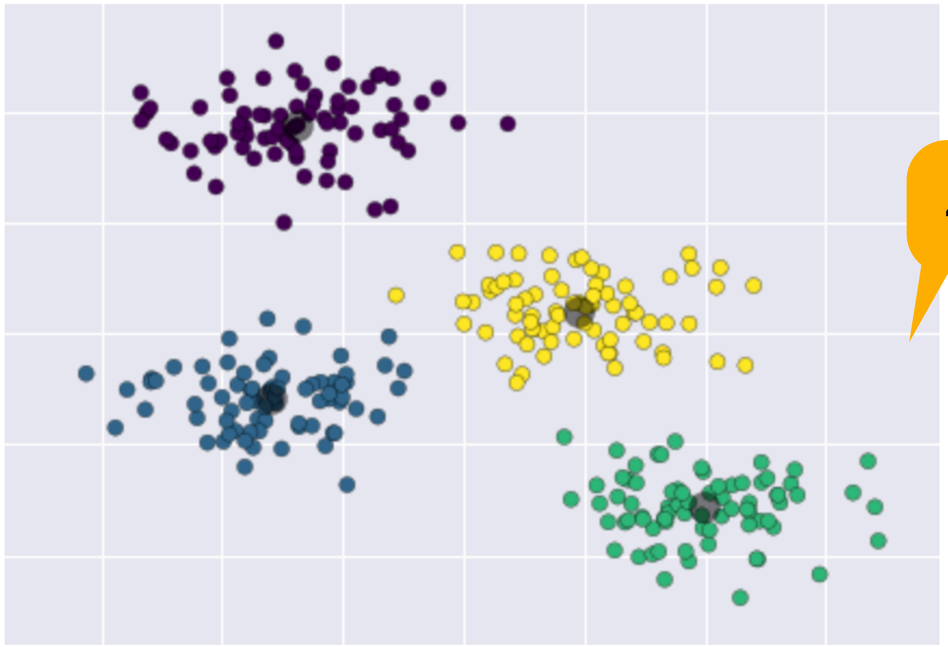




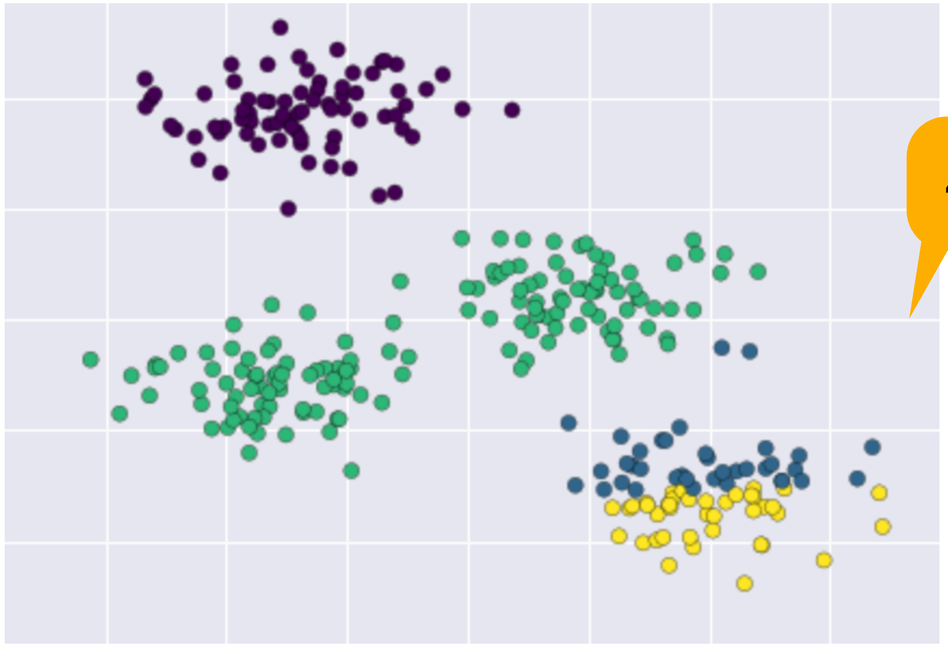
kMeans - Limitation

Globally Optimal Clusters





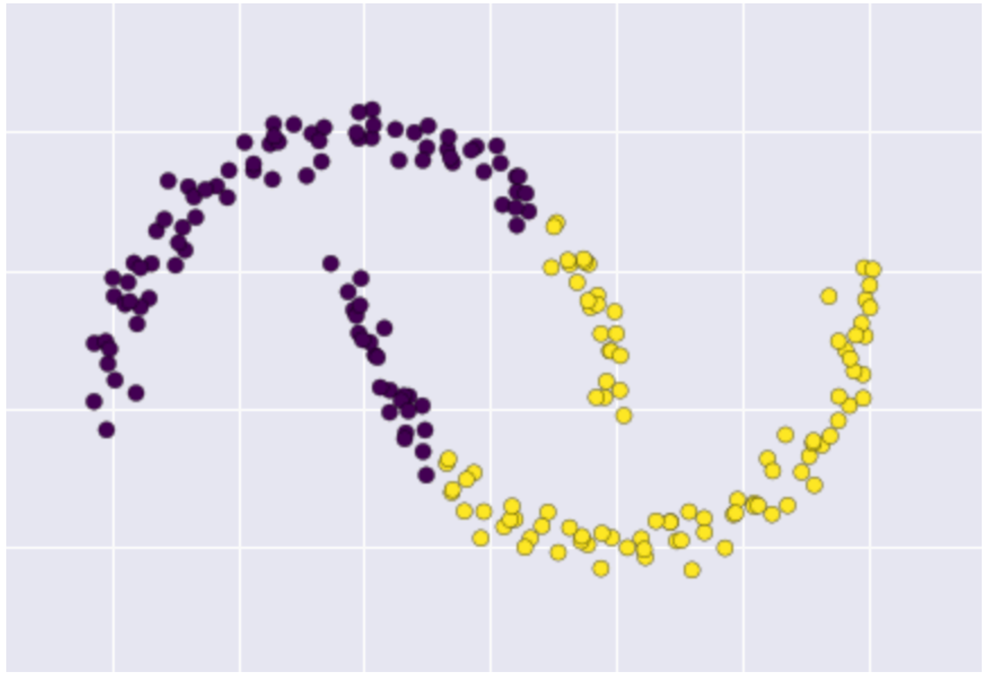
4 Clusters

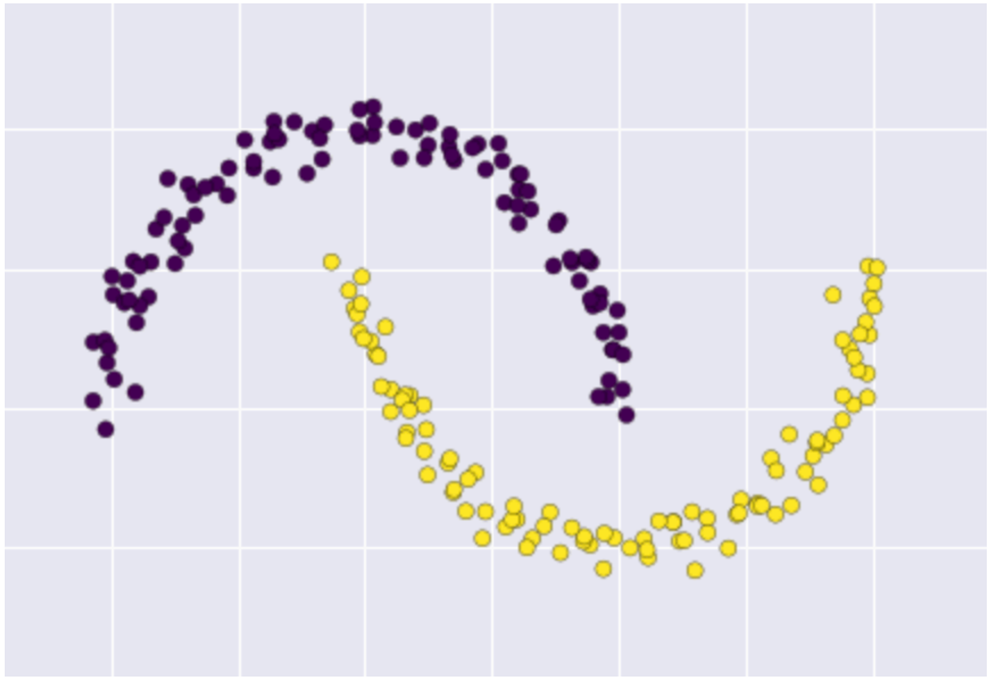


4 Clusters

kMeans - Limitation

Linear Boundaries





The End – Thank You & Good Bye