Warming up …

When applying a learning algorithm, some things are properties of the problem you are trying to solve, and some things are up to you to choose as the ML programmer.

Which of the following are properties of the problem?

- The data generating distribution
- The train/dev/test split
- The learning model
- The loss function

CMPS 460 – Spring 2022

MACHINE LEARNING

Tamer Elsayed

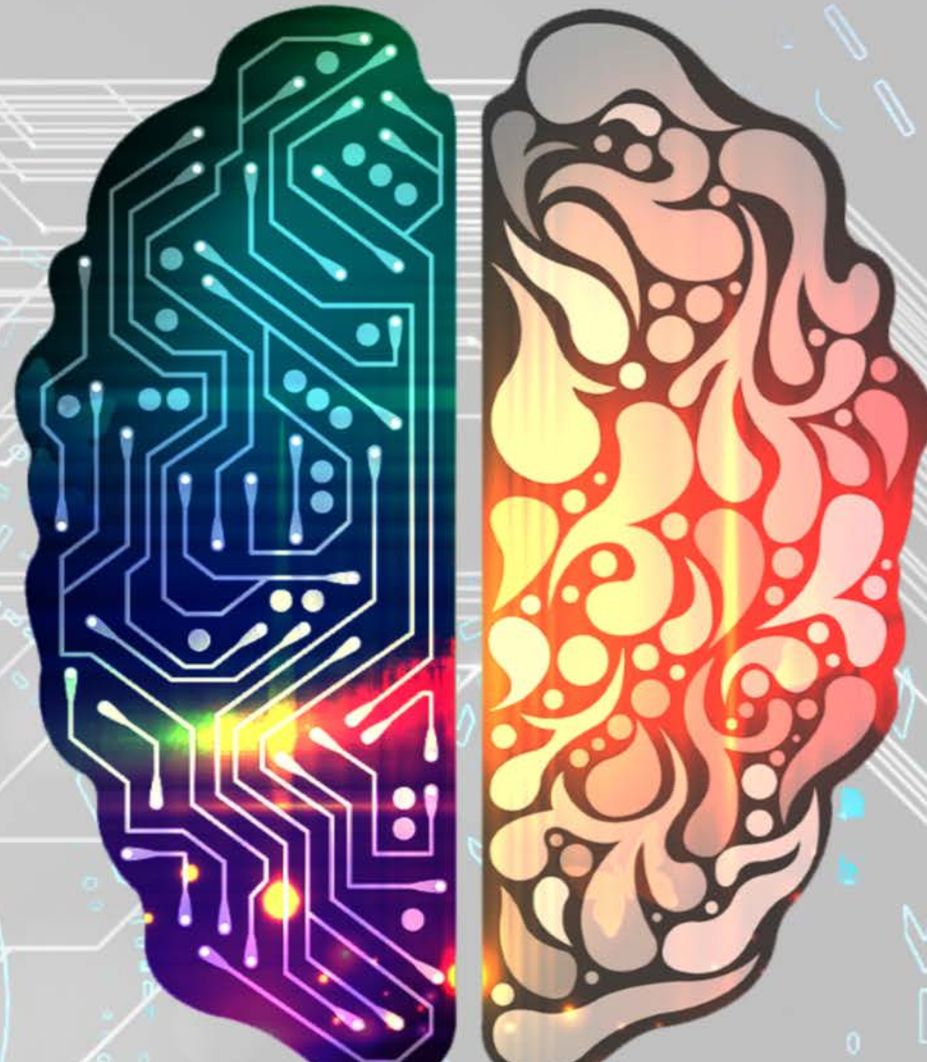


Image hosted by: WittySparks.com | Image source: Pixabay.com

3.b

K-Means Clustering

3.4-3.5

Roadmap …

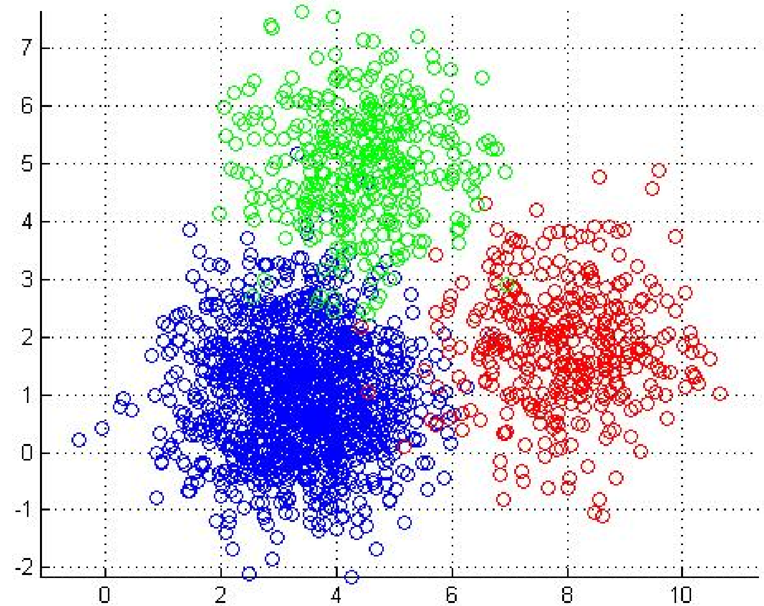
- A new algorithm
  - K-Means Clustering

- Fundamental Machine Learning Concepts
  - Unsupervised vs. supervised learning

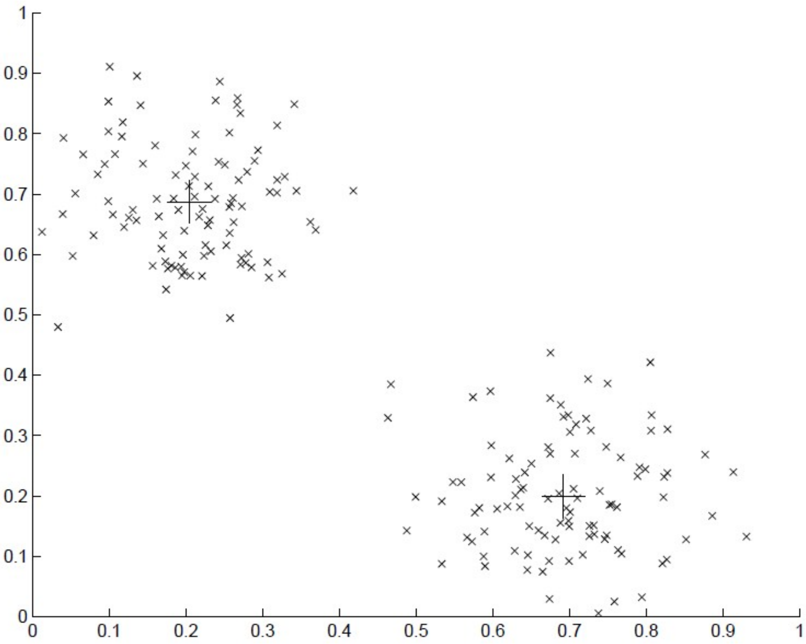
What is Clustering?

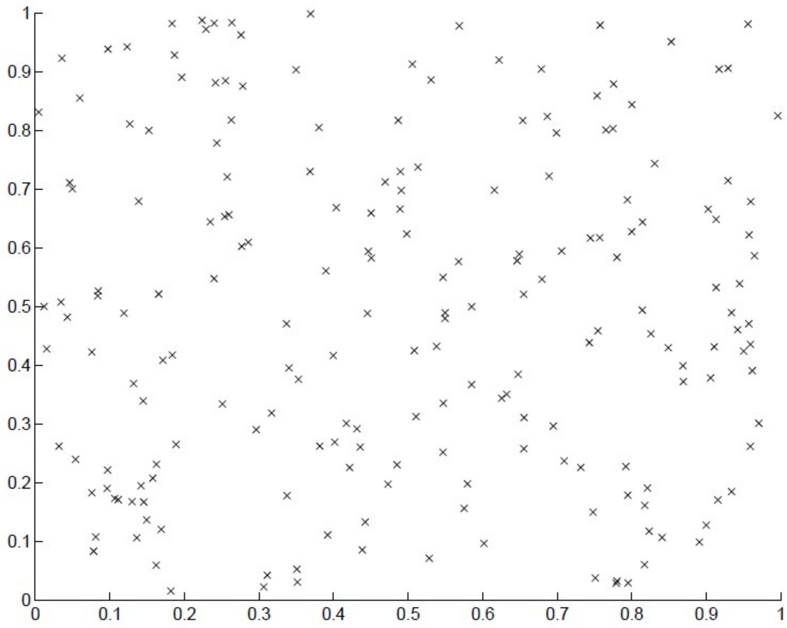
Clustering

- Goal: automatically partition examples into groups of similar examples.
  - find the similarities between objects according to the object attributes and group the similar objects into clusters.



Different Underlying Structures





Why Clustering?

- Often used for **exploratory analysis** of the data

- Automatically organizing data
- Understanding hidden structure in data
- Preprocessing for further analysis

What can we cluster in practice?

- news articles or web pages by topic
- protein sequences by function, or genes according to expression profile
- users of social networks by interest
- customers according to purchase history
- …

Clustering Setup

- **Input**
  - a set S of n points $\{x_1, x_2, ... x_n\}$ in feature space
  - a distance measure specifying distance $d(x_i,x_j)$ between pairs (x_i,x_j)
- **Output**
  - A partition $\{S_1,S_2, ... S_k\}$ of S

Supervised?

Unsupervised Learning

- Clustering is an example of unsupervised learning.
- We are not given examples of classes Y.
- There are **no predictions** made.
- Instead we have to discover clusters/structure in data.

Clustering

- *How do I group these documents by topic?*
- *How do I group my customers by purchase patterns?*

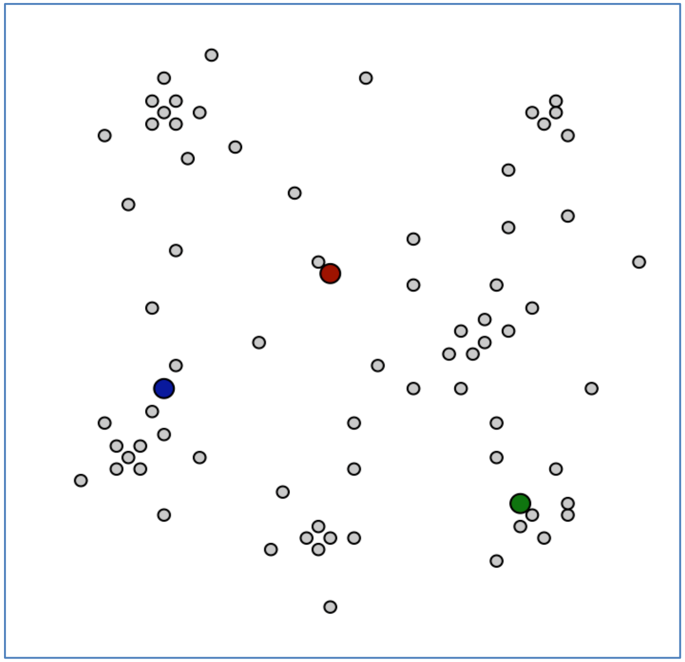
- Sort items into groups by similarity:
  - Items in a cluster are more similar to each other than they are to items in other clusters.
  - Need to detail the properties that characterize "similarity"
    - Or of distance, the "inverse" of similarity

- Our Example: K-means Clustering

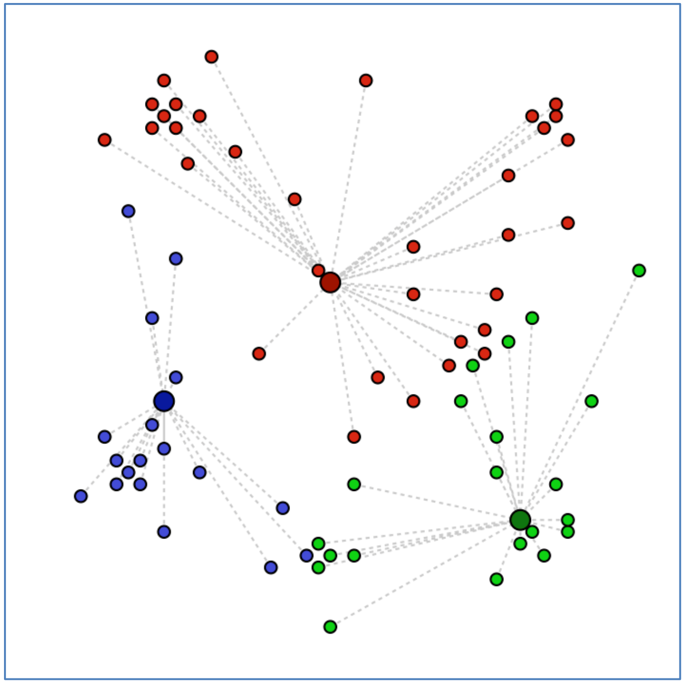
K-means Clustering

The Algorithm

1. Choose *K*; then select K random "centroids“
   - In our example, K=3

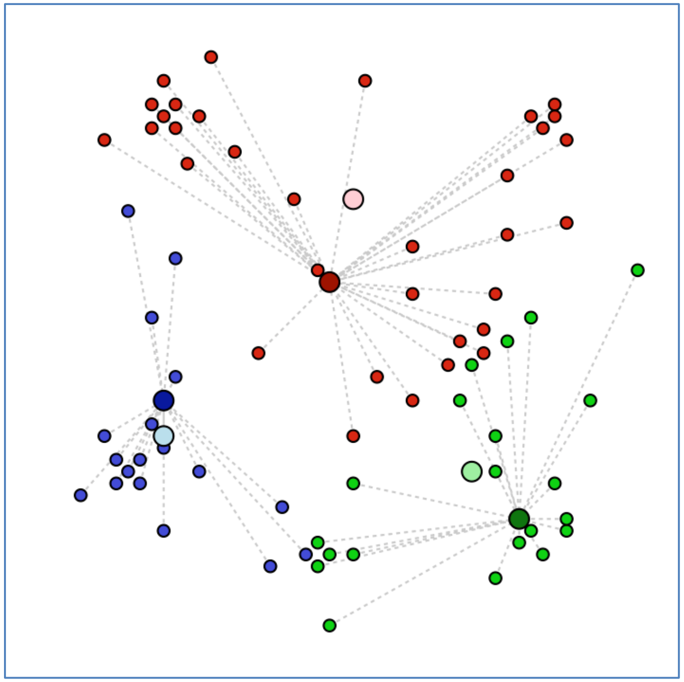


2. Assign each point to the cluster with the ***closest*** centroid

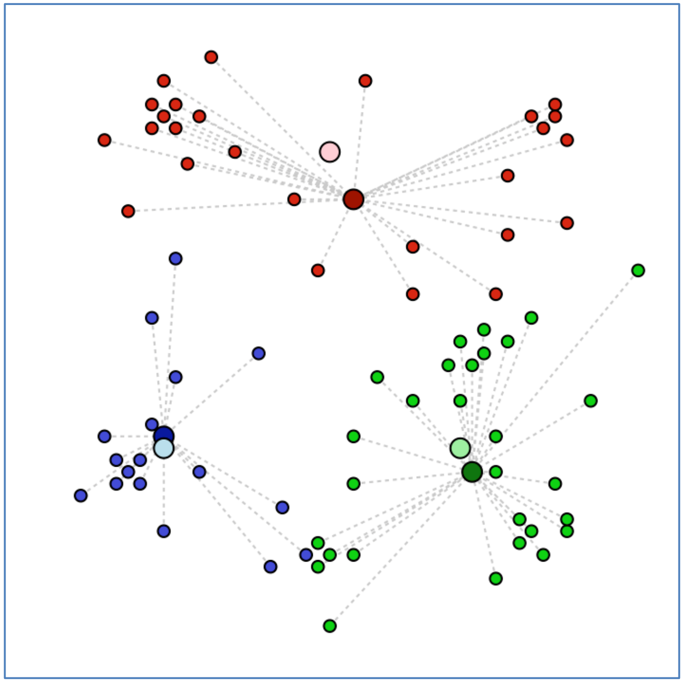


The Algorithm

3. Recalculate the resulting centroids



4. Repeat steps 2 & 3 until point assignments no longer change



K-means Clustering

Algorithm 4 K-MEANS($\mathbf{D}$, K)

for $k = 1$ **to** K **do**
    $\mu_k \leftarrow$ some random location // randomly initialize mean for kth cluster
end for
repeat
    for $n = 1$ **to** N **do**
        $z_n \leftarrow \text{argmin}_k \, ||\boldsymbol{\mu}_k - \boldsymbol{x}_n||$ // assign example n to closest center
    end for
    for $k = 1$ **to** K **do**
        $\mathbf{X}_k \leftarrow \{ \boldsymbol{x}_n : z_n = k \}$ // points assigned to cluster k
        $\mu_k \leftarrow \text{MEAN}(\mathbf{X}_k)$ // re-estimate mean of cluster k
    end for
until μs stop changing
return z // return cluster assignments

Visualization …

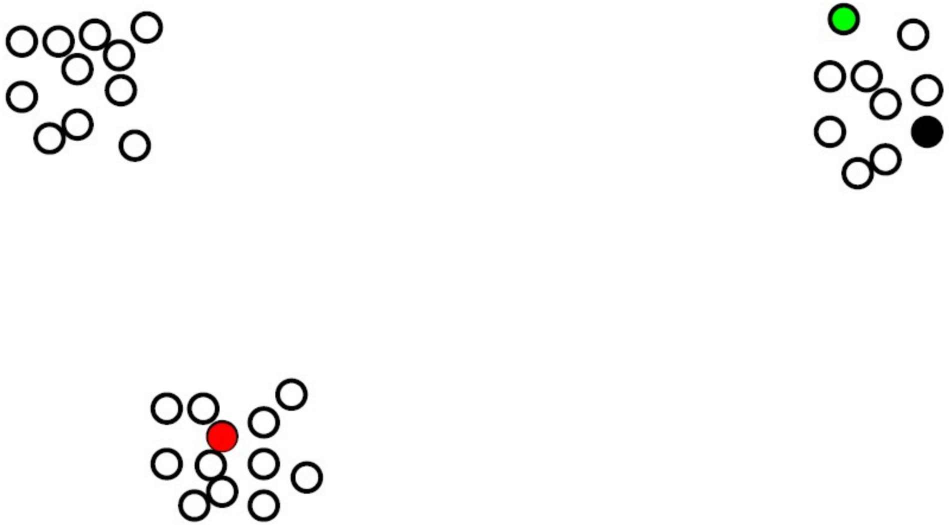
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-means Properties

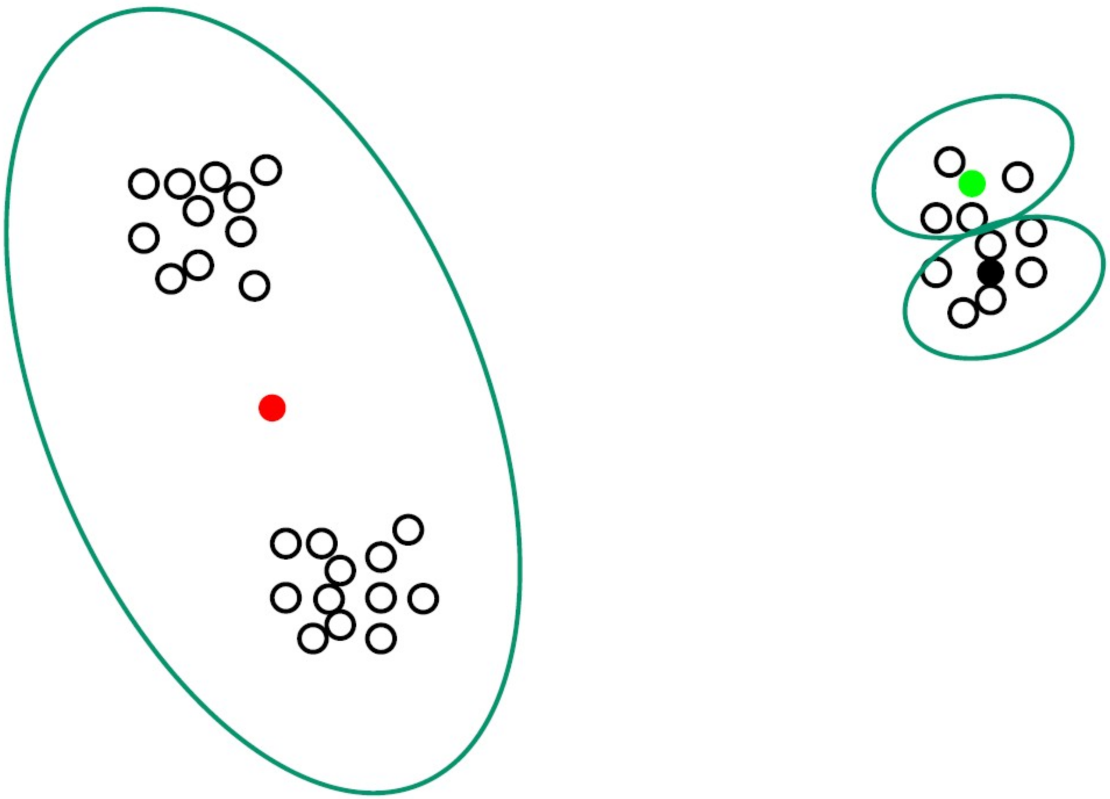
- Different initializations yield different results!
  - Doesn’t necessarily converge to best partition

- K is a hyper-parameter
  - Needs to be set in advance (or learned on dev set)

Impact of Initialization



Impact of Initialization

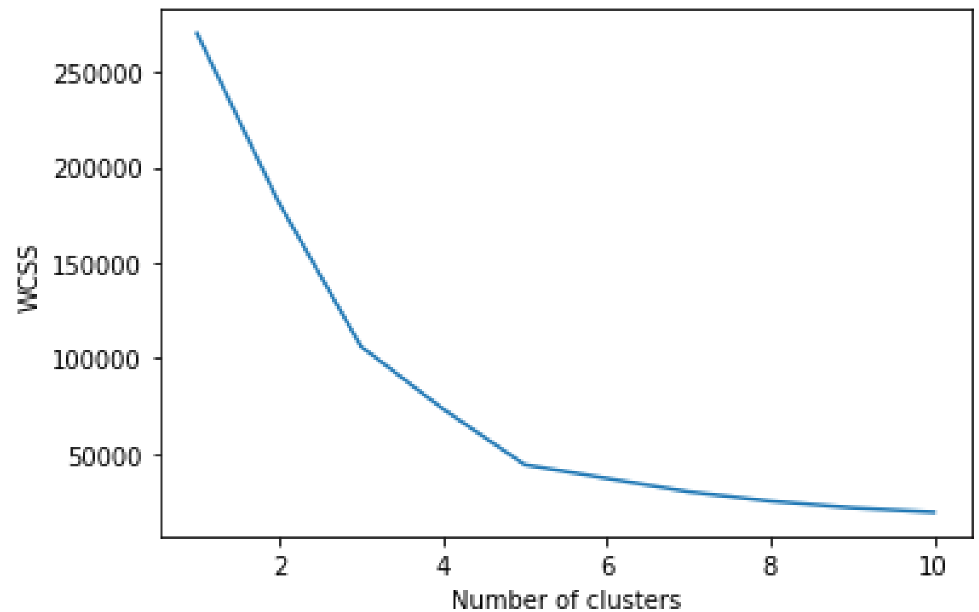


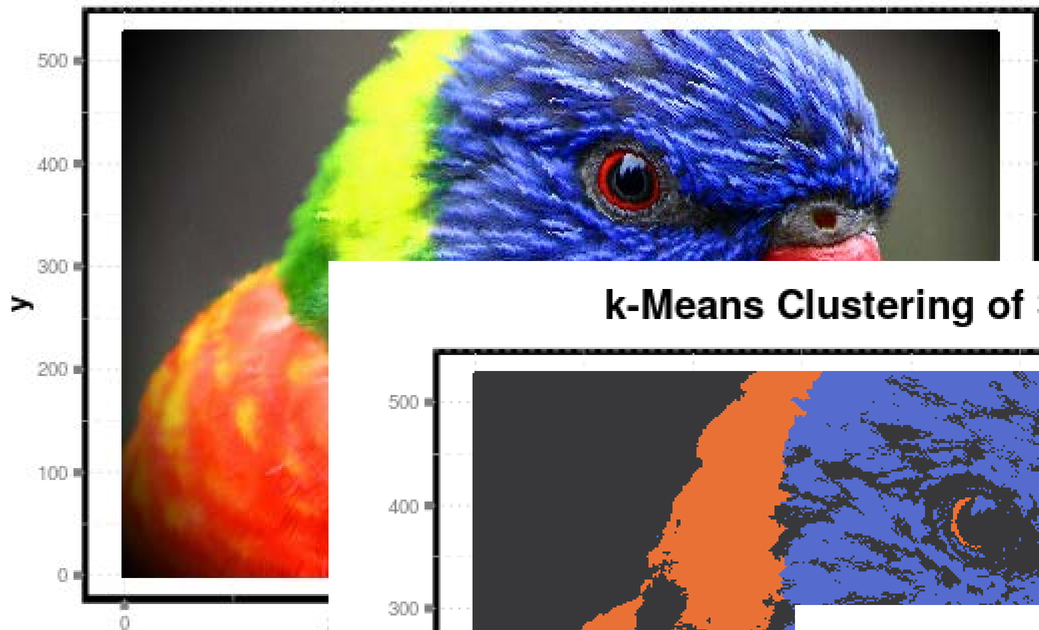
Picking K

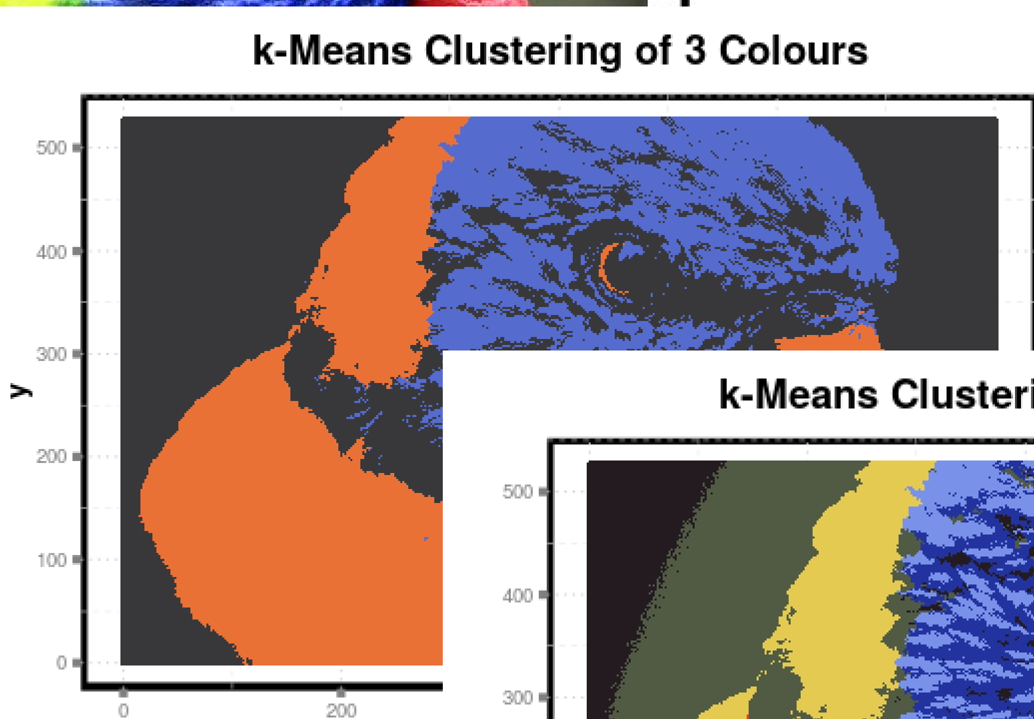
Heuristic: find the "elbow" of the **within-sum-of-squares** (wss) plot as a function of K.

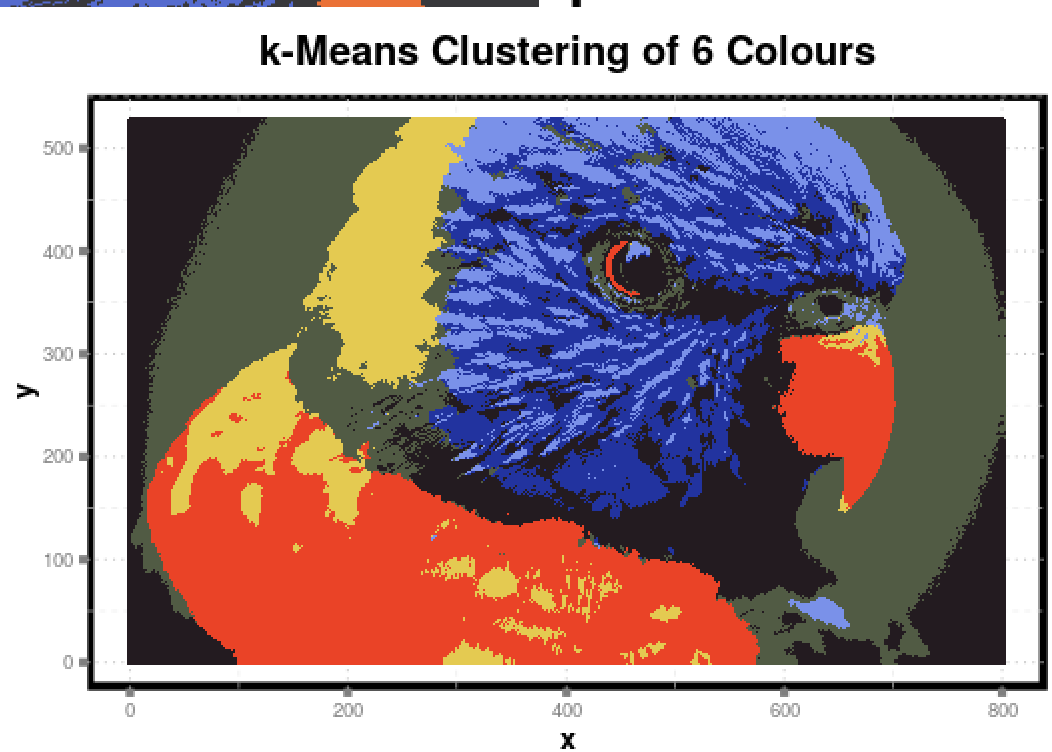
$$\text{wss} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - c_i|^2$$

k: # of clusters
n_i: # points in i^{th} cluster
c_i: centroid of i^{th} cluster
x_{ij}: j^{th} point of i^{th} cluster




Original Image: Colorful Bird
y
500
400
300
200
100
0
0


k-Means Clustering of 3 Colours
y
500
400
300
200
100
0
0
200


k-Means Clustering of 6 Colours
y
500
400
300
200
100
0
0
200
400
600
800
x

Use Cases

- Animals: height, weight and average lifespan
- Customers: household income, yearly purchase amount in dollars, family size
- Patients: record with measures of BMI, HBA1C, HDL
- Objects in image or video