DATA SCIENCE
DREAM JOB

Unsupervised Learning

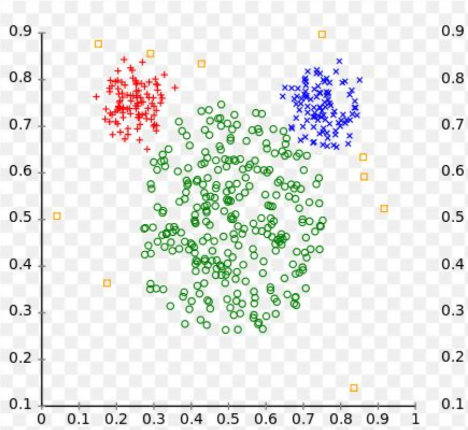
K-MEANS CLUSTERING

Table of Contents

What is Clustering?

K-Means Clustering is a clustering algorithm and is considered to be an ***Unsupervised Learning*** technique.

It is used to divide a ***group*** of data points into clusters, where each point in the cluster is similar to each other.



What is Clustering Used For?

Finding Groups:

- Types of Customers
- Types of Complaints
- Types of Consumer Behaviors
- The list GOES ON...

Data Reduction:

- Summarization of data
- Compression (e.g. Image Processing: vector quantization)

Finding Anomalies:

- Fraud
- Security

(Note - Anomalies can be considered clusters that are small and are points that are very far away from any centroid)

GOALS

- Summarize your data
- Partition your data
- Explore your data
- Find patterns in your data

Find **groups** of data that are all **similar**

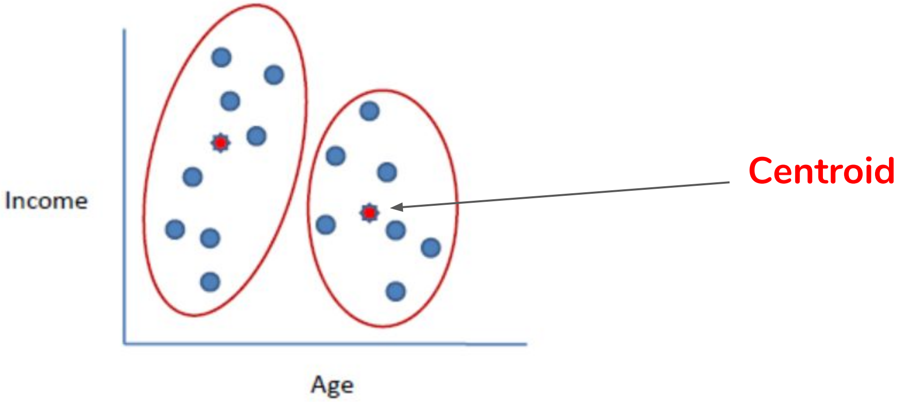
Important Terminologies

1. Centroid
2. Distance Measure

Centroids

Simple: the center point of a cluster.

If **K=3** (we want to find 3 clusters, then we would have **3 centroids**, or centers, one for each cluster



Distance

Distance measures how **similar** two elements are and will influence the shape of the clusters.

To achieve accurate clustering, you need to:

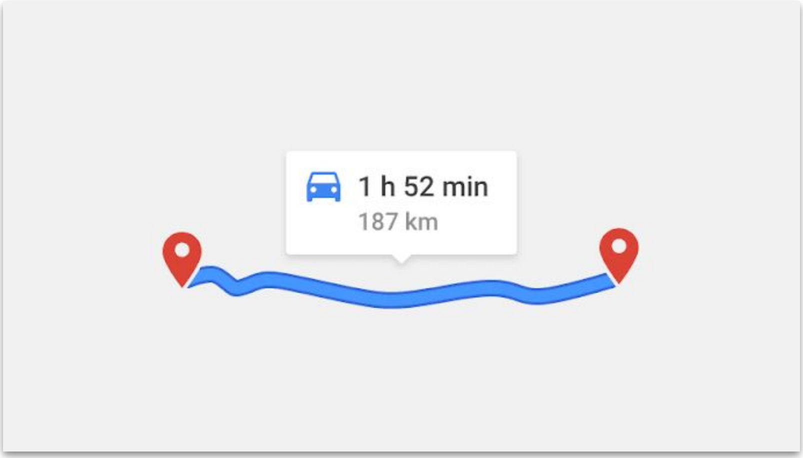
1. Choose the right distance metric
2. Have good intuition behind your data

Distance

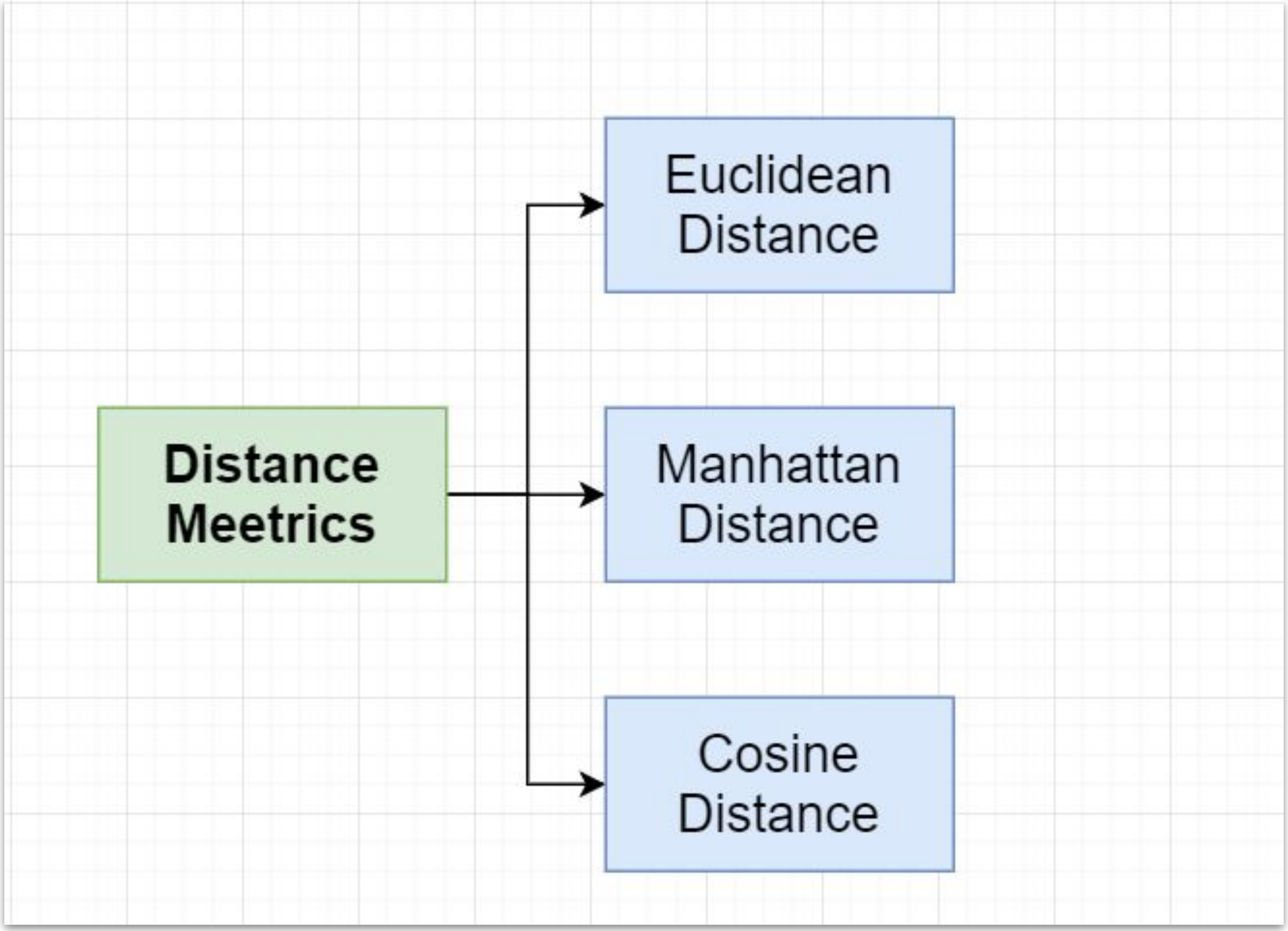
Distance → Dissimilarity

Smaller the distance → More Similarity

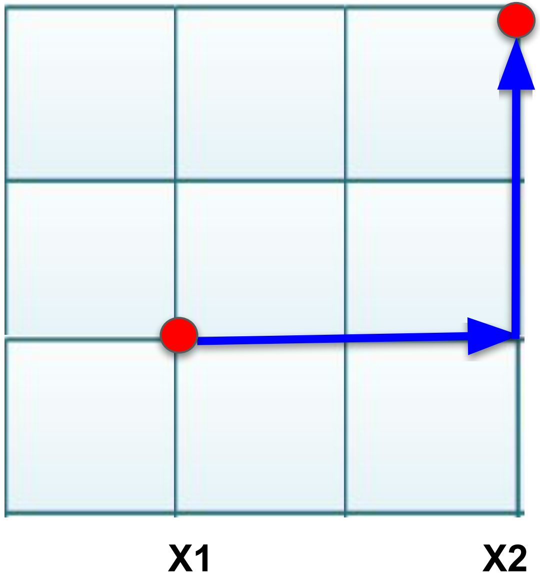
BIGGER the distance → Less Similarity



Common Distance Metrics



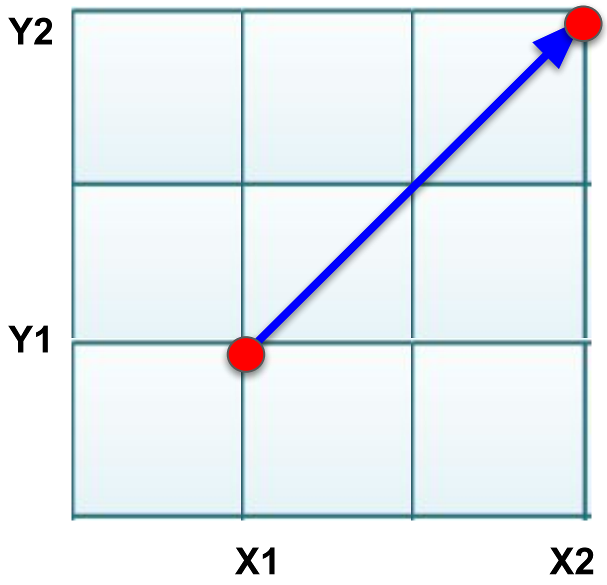
Manhattan Distance (L1)



The distance between two points is the sum of the (absolute) differences of their coordinates.

$$|x_1 - x_2| + |y_1 - y_2|$$

Euclidean Distance (L2)



The distance can be defined as a straight line between 2 points.

(Most common distance metric)

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Objective/Cost Function

Objective: To minimize total intra-cluster variance (e.g. Sum of Squared Error SSE)

Minimize Error: The error is the distance of each observation to the nearest cluster

number of clusters

number of cases

case i

centroid for cluster j

objective function $\leftarrow$

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

Goals for Clustering

We want:

- ❑ **Seperation:** Observations in different clusters are **dissimiliar** to each other
- ❑ **Homogeneity:** Observations in the same cluster are **similar** to each other
- ❑ Find natural groupings

What are the Inputs & Outputs?

Inputs

- ❑ A set of numerical inputs (normally scaled)

Outputs

- ❑ A set of labels, one for each observation
- ❑ A set of centroids, one for each cluster

Basic Steps for Clustering

- Preprocessing
  - Normalization/Standardization
- Distance/Similarity Measure
  - Similarity of two feature vectors
- Clustering Criterion
  - Based on cost function
- Clustering Algorithms
  - Based on clustering algorithm
- Validation/Interpretation

Preprocessing

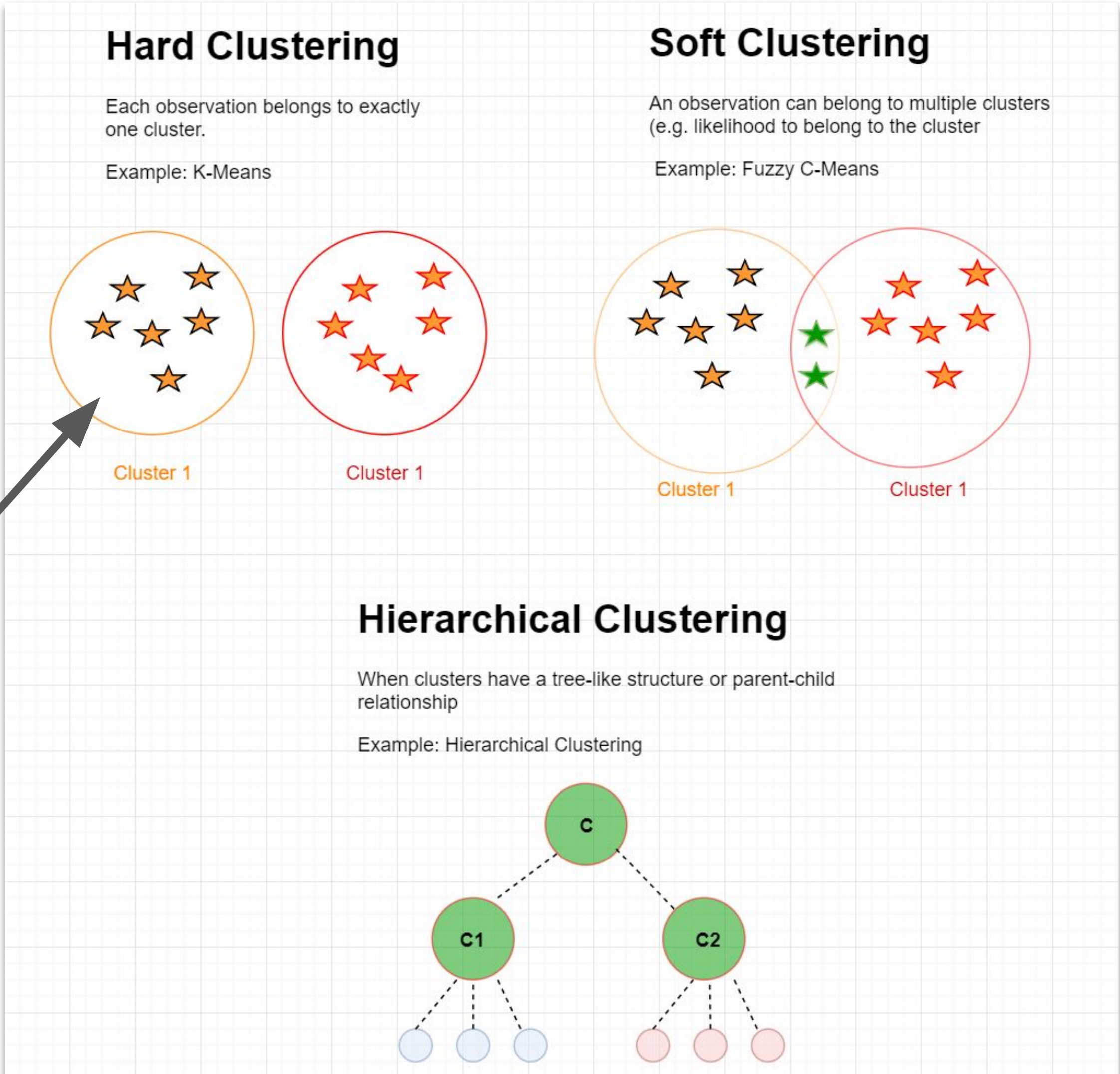
A. Normalization & Standardization

Scaling is important because remember, we’re using distance as our metric of similarity

B. Remove Outliers

Types of Clustering

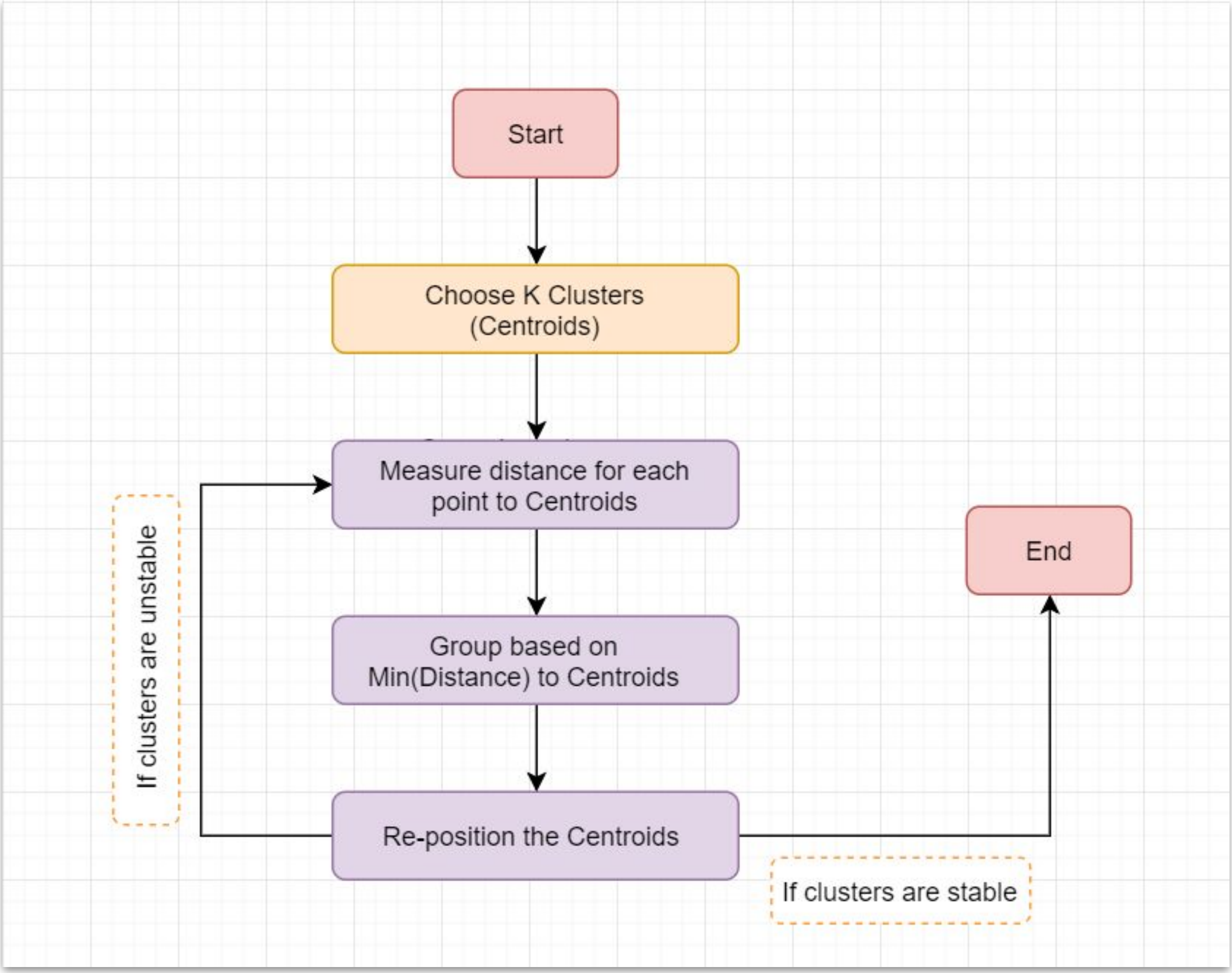
- We're going to focus on **K-Means**



K-Means Breakdown

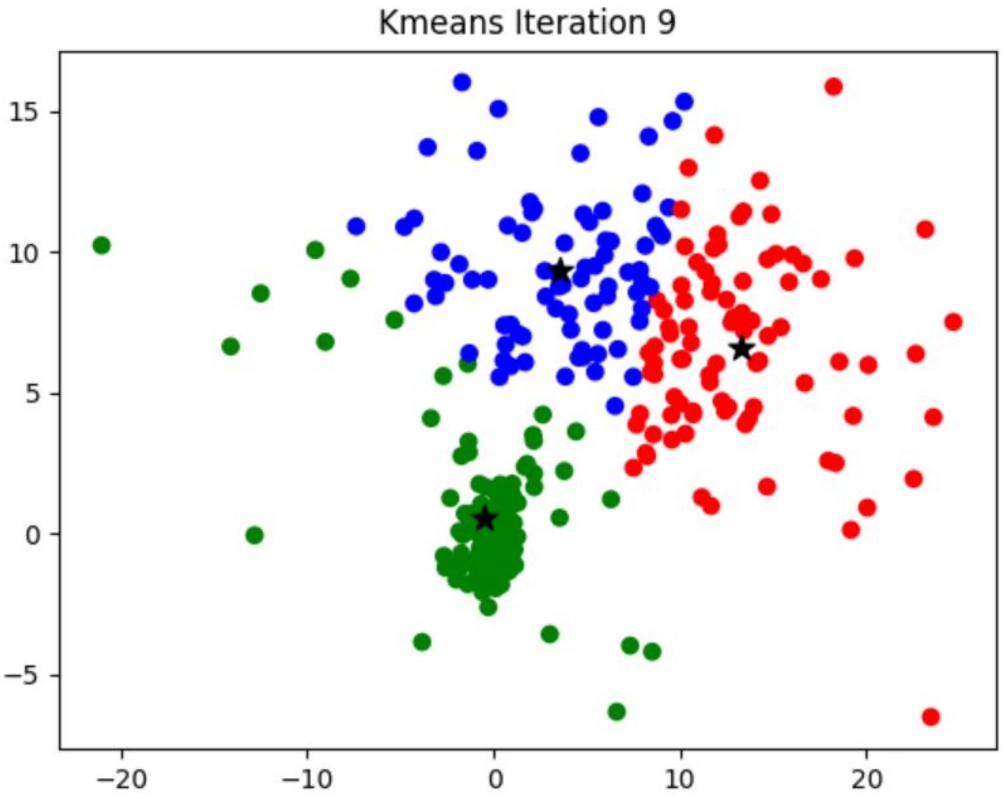
Iterate until clusters are stable:

1. Determine centroid locations
2. Determine distance of each observation to centroids
3. Group objects based on minimum distance (find closest centroid)



K-Means Breakdown

WATCH CLIP
Source: https://sandipanweb.files.wordpress.com/2017/03/kmeans5.gif?w=640&zoom=2



Performance

- ➔ *So how do we choose the* ***right*** *amount of clusters?*
- ➔ *How do we know which cluster is better than one or the other?*

This question makes evaluating clusters one of the most ***trickiest*** *parts of K-Means.*

Clustering is Subjective

*Example: How would you group these pokemons into **clusters**?*

How to Choose "Right" Amount of Clusters

- ➔ *Heuristic Criteria*
- ➔ *Elbow Method*
- ➔ *MANY MANY more…*

*Remember, clustering is very **subjective** and it also depends on your **problem**.*

Finding K - Heuristic Criteria

A. Your boss wants you to identify 7 groups of customer phone calls that your call center receives

Then, K = 7 :)

B. You want to separate a population into 3 shirt sizes.

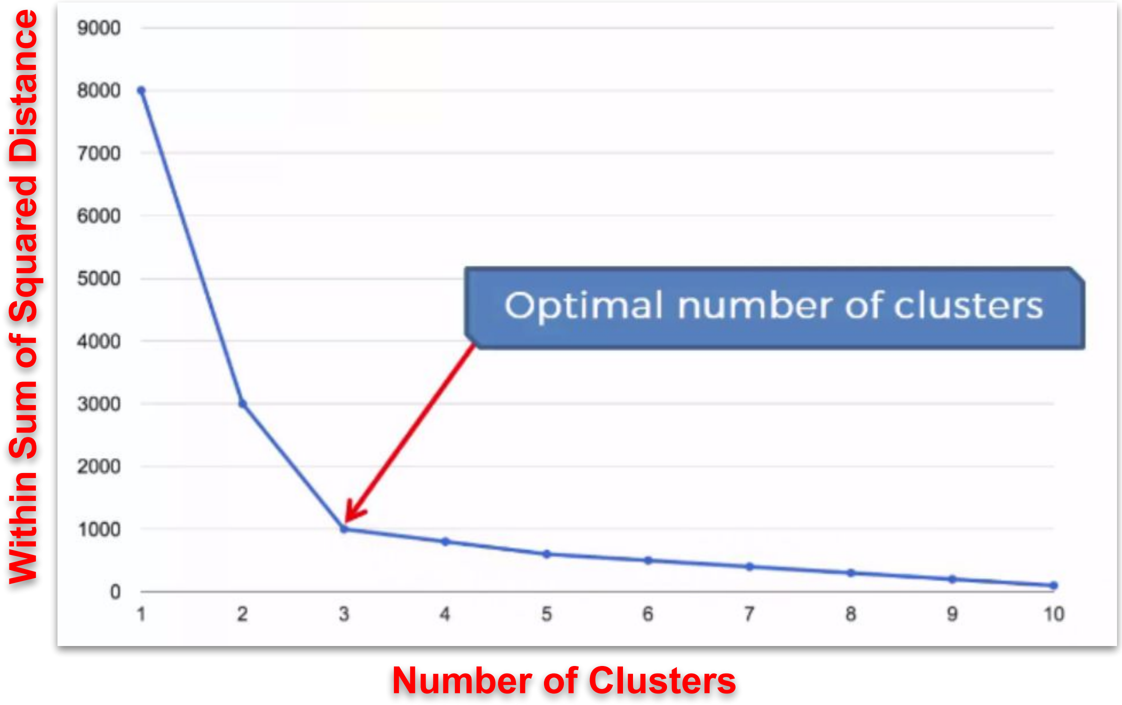
Then, K = 3 :)

Finding K - Elbow Method

GOAL: To identify when the set of clusters explains **"most"** of the variance in the data.

X - Number of Clusters

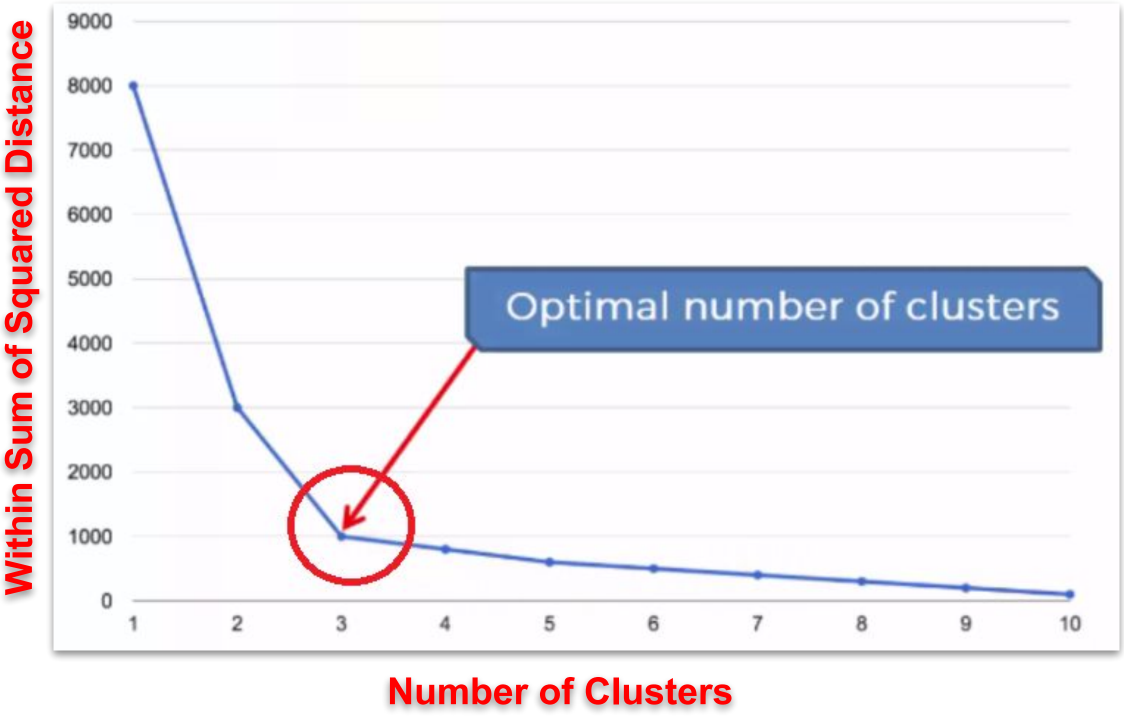
Y - Within SSE (Cumulative variance explained)



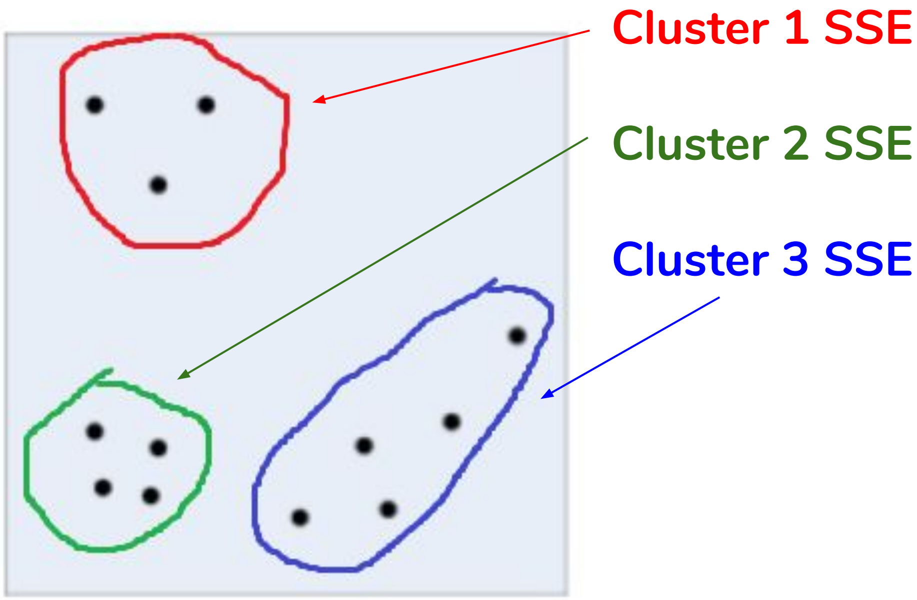
Finding K - Elbow Method

Since increasing K will always decrease our metric, the **"elbow point"** will allow us to see where the rate of decrease **sharply shifts**

We want to find the point where the distortion remains constant, even if we increase K further



Within Sum of Squared Distance



$k = 3$

$$\underbrace{\sum_{i=1}^{k}}_{\text{For each cluster}} \underbrace{\sum_{p \in C_i}}_{\text{For each point in the cluster}} \underbrace{||p - \mu_i||^2}_{\text{Distance from point to centroid}}$$

Finding K - Elbow Method

In short...

The **“ELBOW”** is where the cumulative variance starts to **FLATTEN OUT.**

And *adding* in new clusters beyond this point only yields relatively *small increase* in variance.