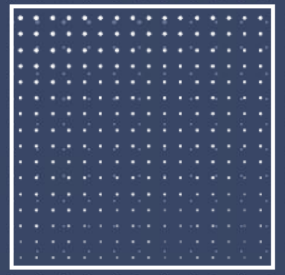
ESKWELABS

Intro to Machine Learning

January 18, 2021



ESKWELABS

Machine Learning

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.” - Arthur Samuel (1959)

Machine Learning is the study of algorithms that
- improve their performance P
- at some task T
- with experience E.

Machine Learning

“Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.” - Arthur Samuel (1959)

Machine Learning is the study of algorithms that
- improve their performance P Percentage of dogs classified correctly
- at some task T Recognizing photos of dogs
- with experience E. Database of images labeled as dogs and non-dogs

Traditional Programming

Machine Learning

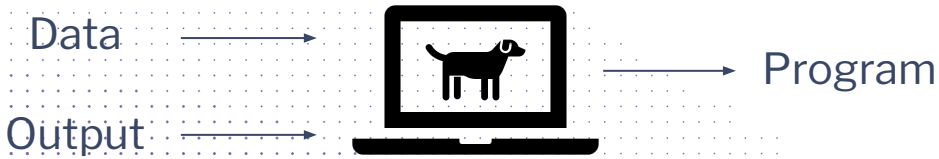


Image Processing

Natural Language Processing

Credit Risk Analysis

Recommendation Engines

Classification

THE DATA SCIENCE PROCESS

- Business Understanding
  - Data Understanding
    - Data Preparation
      - Modelling
        - Evaluation
          - Deployment

*"Machine Learning experts often didn't build their work around the final objective – **deriving business value**."*

https://hbr.org/2016/12/why-youre-not-getting-value-from-your-data-science

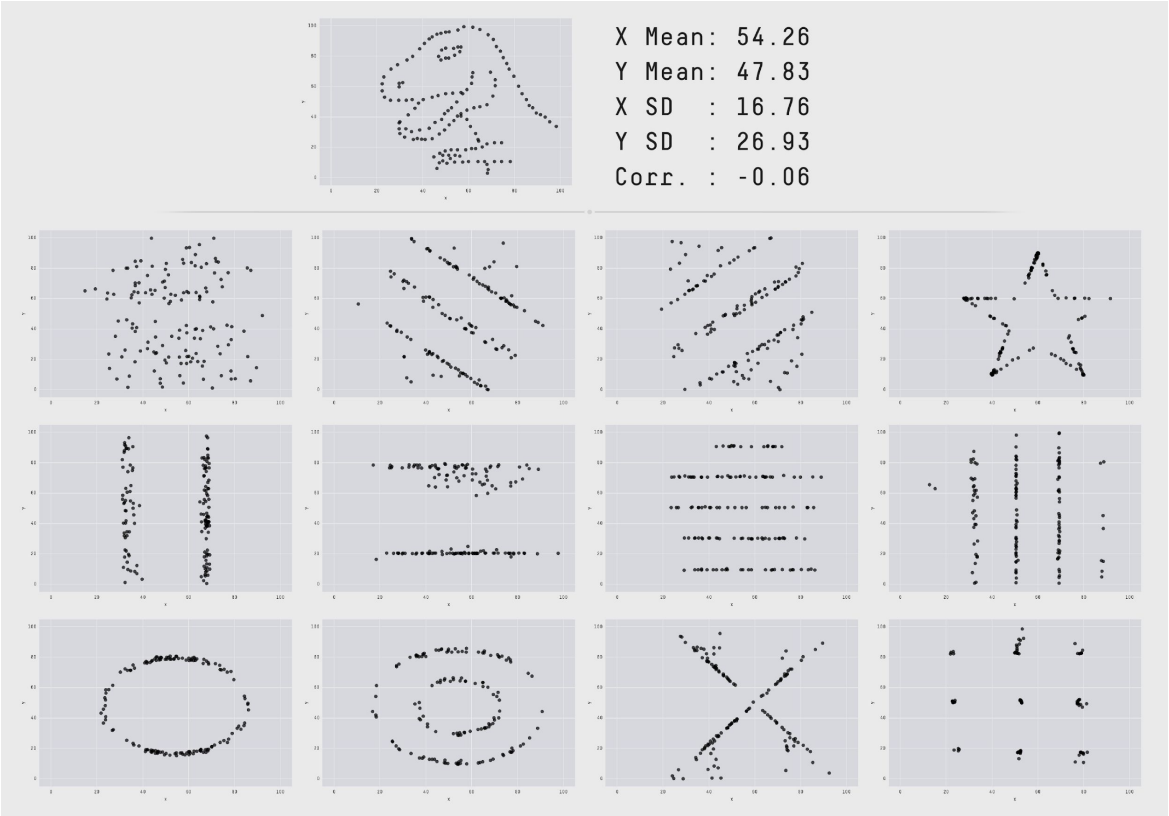
- *What is the purpose of the model?*
- *What problem will it solve?*
- *What do we gain from building this model?*

“Numbers don’t lie… but humans do.”

- Ernie Lindsey
American novelist

The Datasaurus Dozen

13 datasets each have the same summary statistics to two decimal places, while being drastically different in appearance.



https://www.autodesk.com/research/publications/same-stats-different-graphs

VARIABLE LABELS

HHID	"Case Identification"
/HVIDX	"Line number"
/HV000	"Country code and phase"
/HV001	"Cluster number"
/HV002	"Household number"
/HV003	"Respondent's line number (answering Household questionnaire)"
/HV004	"Ultimate area unit"
/HV005	"Household sample weight (6 decimals)"
/HV006	"Month of interview"
/HV007	"Year of interview"
/HV008	"Date of interview (CMC)"
/HV008A	"Date of interview Century Day Code (CDC)"
/HV009	"Number of household members"
/HV010	"Number of eligible women in household"
/HV011	"NA - Number of eligible men in household"
/HV012	"Number of de jure members"
/HV013	"Number of de facto members"
/HV014	"Number of children 5 and under (de jure)"
/HV015	"Result of household interview"
/HV016	"Day of interview"
/HV017	"Number of visits"
/HV018	"Interviewer identification"
/HV019	"NA - Keyer identification"
/HV020	"Ever-married sample"
/HV021	"Primary sampling unit"
/HV022	"Sample strata for sampling errors"
/HV023	"Stratification used in sample design"

VALUE LABELS

HV003

0 "Incomplete household"

/HV015

1 "Completed"
2 "No Household member/no competent member at home"
3 "Entire Household absent for extended period of time"
4 "Postponed"
5 "Refused"
6 "Dwelling vacant or address not a dwelling"
7 "Dwelling destroyed"
8 "Dwelling not found"
9 "Other"

/HV020

0 "All woman sample"
1 "Ever married sample"

/HV022

1 "BUCAY, ABRA"
2 "BUTUAN CITY (CAPITAL), AGUSAN DEL NORTE"
3 "CITY OF CABADBARAN, AGUSAN DEL NORTE (EXCLUDING BUTUAN CITY)"
4 "TALACOGON, AGUSAN DEL SUR"
5 "NUMANCIA, AKLAN"
6 "BACACAY, ALBAY"
7 "TIBIAO, ANTIQUE"
8 "CONNER, APAYAO"
9 "DIPACULAO, AURORA"
10 "CITY OF LAMITAN, BASILAN"
11 "DINALUPIHAN, BATAAN"

VARIABLE LABELS

HHID "Case Identification"
/HVIDX "Line number"
/HV000 "Country code and phase"
/HV001 "Cluster number"
/HV002 "Household number"
/HV003 "Respondent's line number (answering Household questionnaire)"
/HV004 "Ultimate area unit"
/HV005 "Household sample weight (6 decimals)"
/HV006 "Month of interview"
/HV007 "Year of interview"
/HV008 "Date of interview (CMC)"
/HV008A "Date of interview Century Day Code (CDC)"
/HV009 "Number of household members"
/HV010 "Number of eligible women in household"
/HV011 "NA - Number of eligible men in househ
/HV012 "Number of de jure members"
/HV013 "Number of de facto members"
/HV014 "Number of children 5 and under (de jur
/HV015 "Result of household interview"
/HV016 "Day of interview"
/HV017 "Number of visits"
/HV018 "Interviewer identification"
/HV019 "NA - Keyer identification"
/HV020 "Ever-married sample"
/HV021 "Primary sampling unit"
/HV022 "Sample strata for sampling errors"
/HV023 "Stratification used in sample design"

VALUE LABELS

HV003
0 "Incomplete household"
/HV015
1 "Completed"
2 "No Household member/no competent member at home"
3 "Entire Household absent for extended period of time"
4 "Postponed"
5 "Refused"
6 "Dwelling vacant or address not a dwelling"
7 "Dwelling destroyed"
8 "Dwelling not found"
9 "Other"
/HV020
0 "All woman sample"

Know your data dictionary

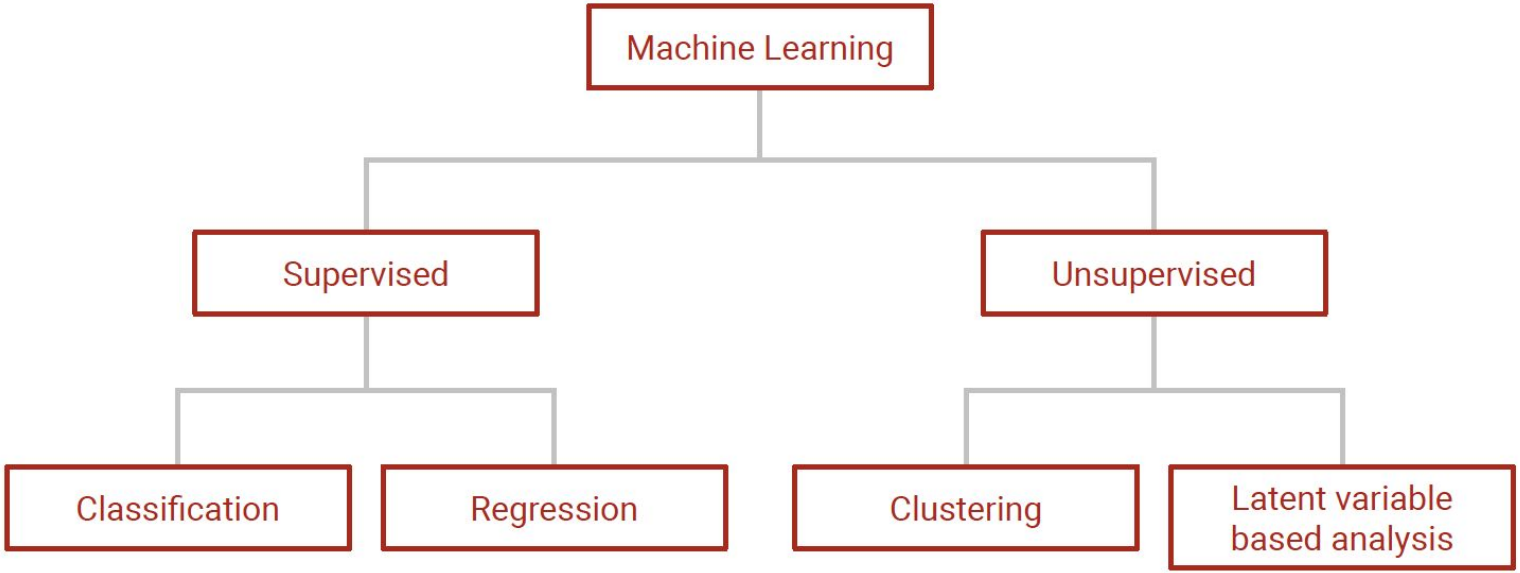
3 "CITY OF CABADBARAN, AGUSAN DEL NORTE (EXCLUDING BUTUAN CITY)"
4 "TALACOGON, AGUSAN DEL SUR"
5 "NUMANCIA, AKLAN"
6 "BACACAY, ALBAY"
7 "TIBIAO, ANTIQUE"
8 "CONNER, APAYAO"
9 "DIPACULAO, AURORA"
10 "CITY OF LAMITAN, BASILAN"
11 "DINALUPIHAN, BATAAN"

Business Understanding | Data Understanding | Data Preparation | Modeling | Evaluation | Deployment

Performing data checks

- *Are there missing values?*
- *Are there illogical data?*
- *Are there outliers?*

- ***How do we handle messy data?***


Machine Learning
Supervised
Unsupervised
Classification
Regression
Clustering
Latent variable based analysis

Supervised Learning

- Finds patterns for a prediction task
- Algorithms are trained using labeled data
- Input and output variable is given
- Can be evaluated using various metrics

Unsupervised Learning

- Finds patterns in data without a specific prediction task in mind
- Algorithms uses data which are not labeled
- Only input variable is given
- Difficult to interpret because there is no gold standard and no single objective
- Can be a useful pre-processor for supervised learning

Performance and Interpretability

Predictive Metrics:

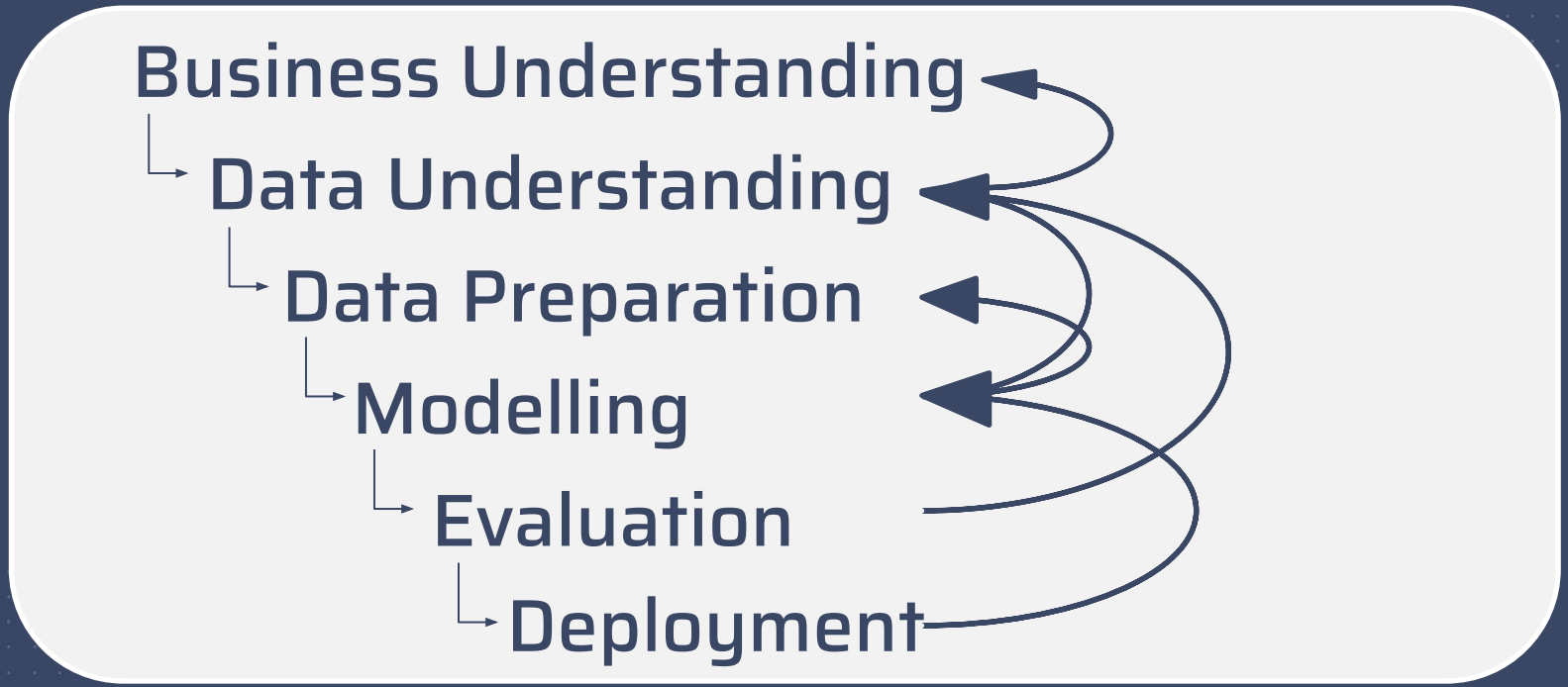
- Accuracy
- Precision
- Recall
- RMSE

- What is the basis for the results?
- How can the results be applied to the initial problem?

Cool model... Now, what?

The most successful data science projects are those that are used by people.

THE DATA SCIENCE PROCESS



Unsupervised Machine Learning

Supervised Learning

- Finds patterns for a prediction task
- Algorithms are trained using labeled data
- Input and output variable is given
- Can be evaluated using various metrics

Unsupervised Learning

- Finds patterns in data without a specific prediction task in mind
- Algorithms uses data which are not labeled
- Only input variable is given
- Difficult to interpret because there is no gold standard and no single objective
- Can be a useful pre-processor for supervised learning

Clustering

Unsupervised Machine Learning

- Clustering refers to a very broad set of techniques **finding subgroups, or clusters**, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other,
- To make this concrete, **we must define what it means for two or more observations to be similar or different**.
- This is often a domain-specific consideration that must be made based on knowledge of the data being studied.

Sprint 1 Objective:

Cluster Analysis of Public Schools in the Philippines

Sprint 1 Objective:

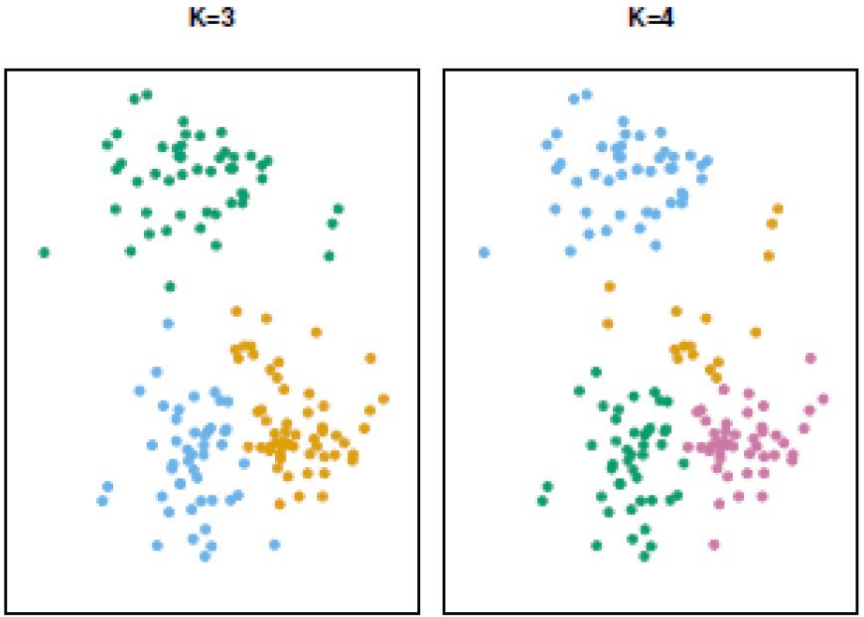
We will use cluster analysis to group schools along with others of similar capacity metrics to determine a capacity building strategy for each specific cluster.

Any Questions?

K-Means Clustering

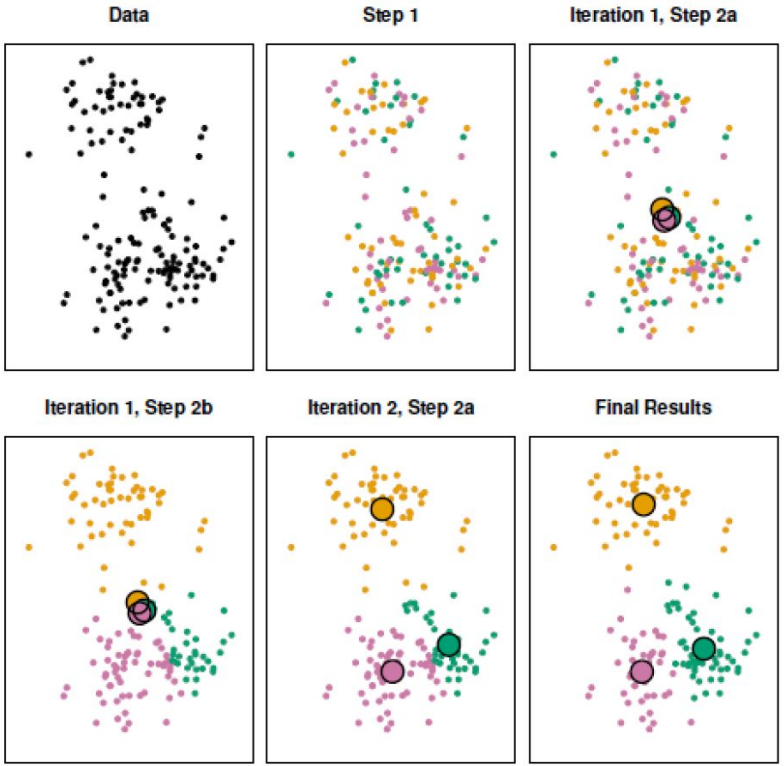
K-Means Clustering

- an iterative algorithm that partition the dataset into distinct subgroups
- each data point belongs to only one group.
- makes the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible



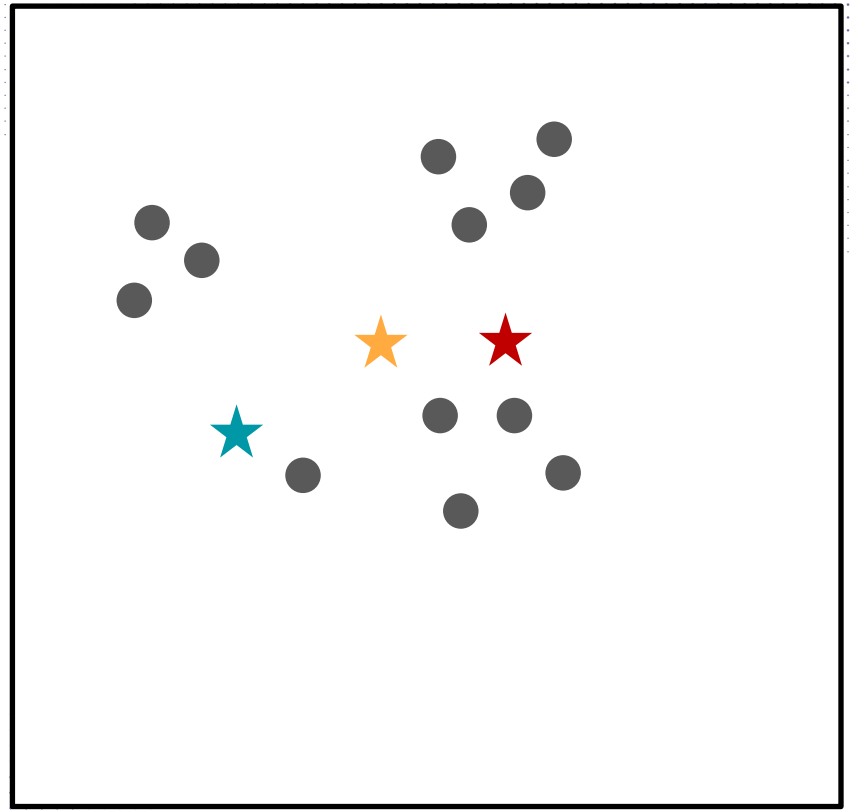
K-Means Clustering

1. Randomly assign a number from 1 to K to each of the observations as initial cluster assignments
2. Iterate until the cluster assignments stop changing:
   - For each of the K clusters, compute the cluster centroid
   - Re-assign each observation to the cluster whose centroid is closest



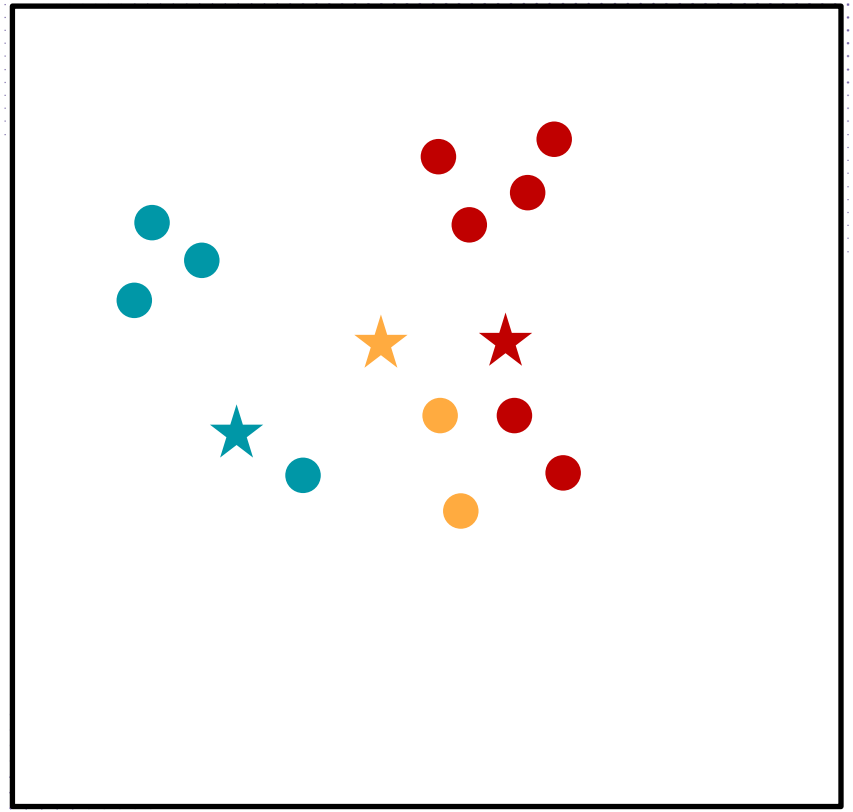
K-Means Clustering

1. Randomly assign centroids



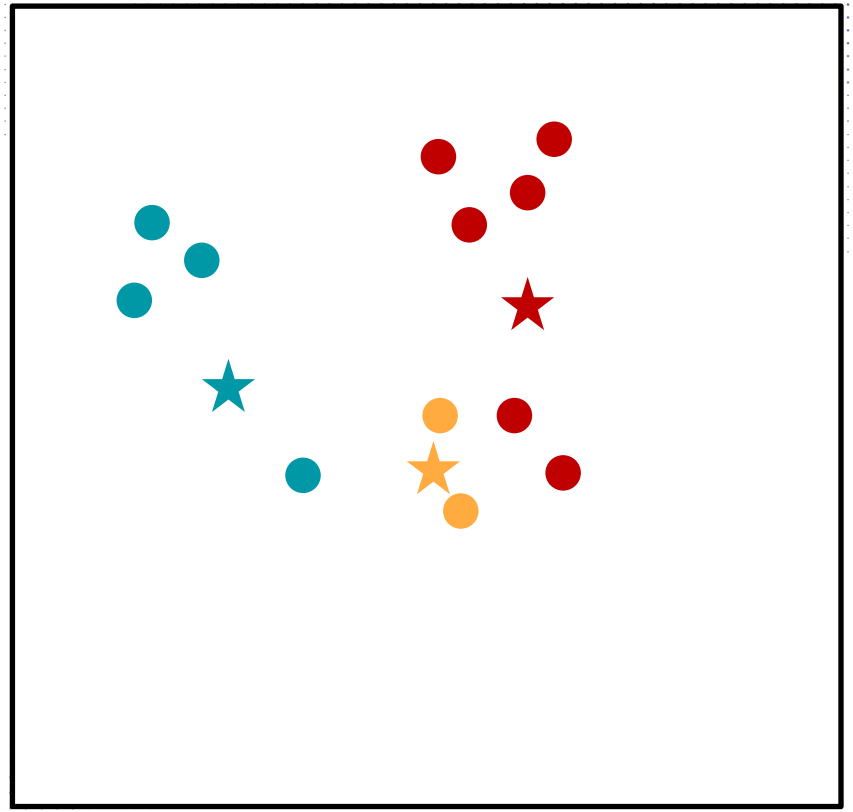
K-Means Clustering

2. Re-assign each observation to the cluster whose centroid is closest



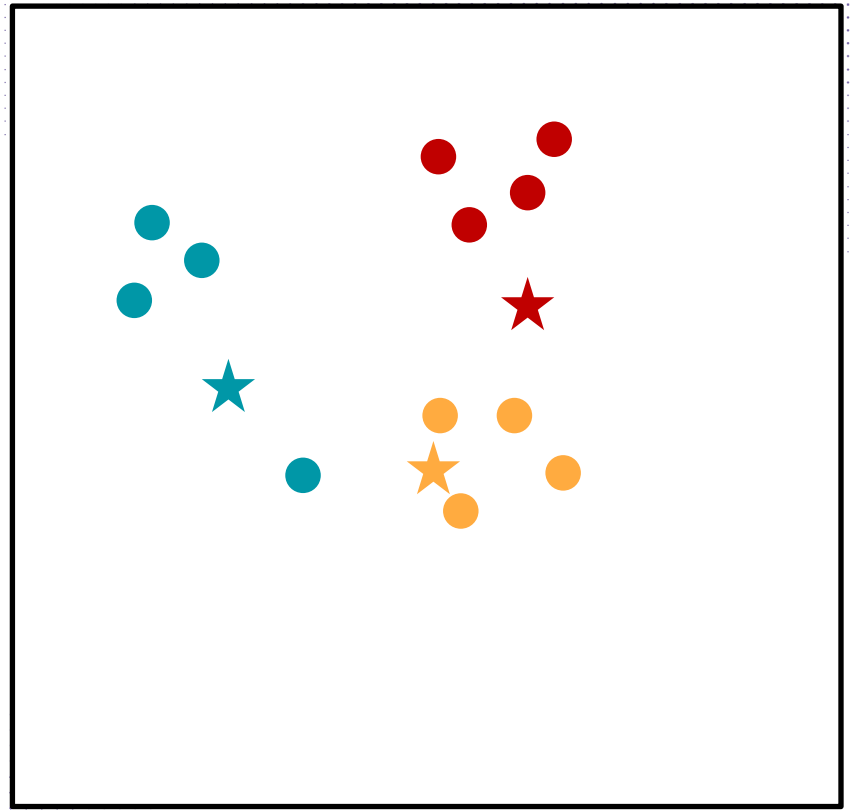
K-Means Clustering

3. Iterate until the cluster assignments stop changing



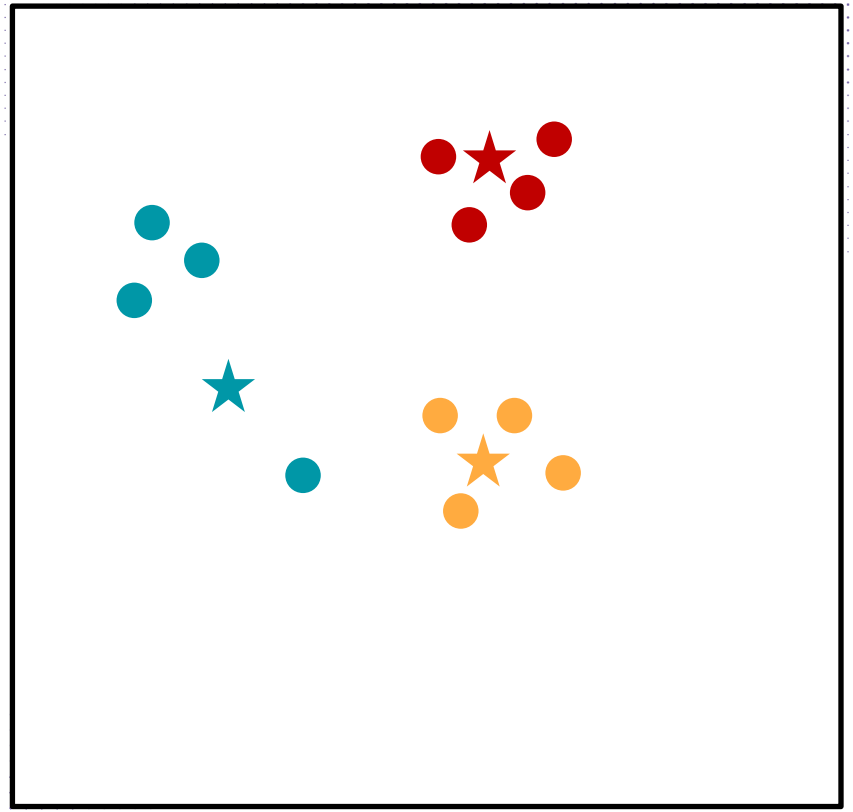
K-Means Clustering

3. Iterate until the cluster assignments stop changing



K-Means Clustering

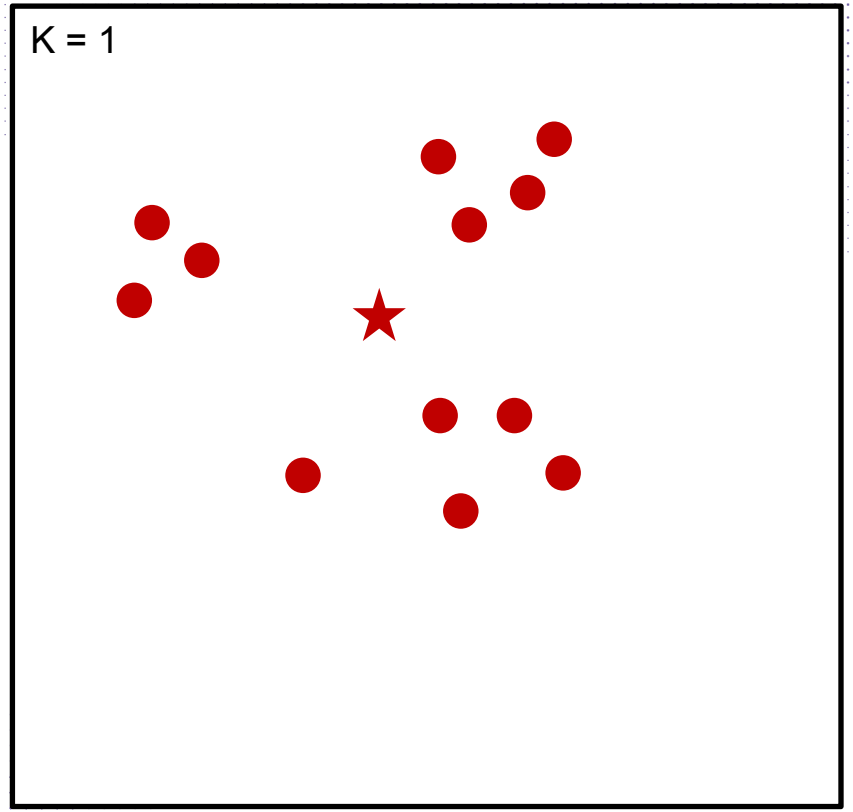
3. Iterate until the cluster assignments stop changing



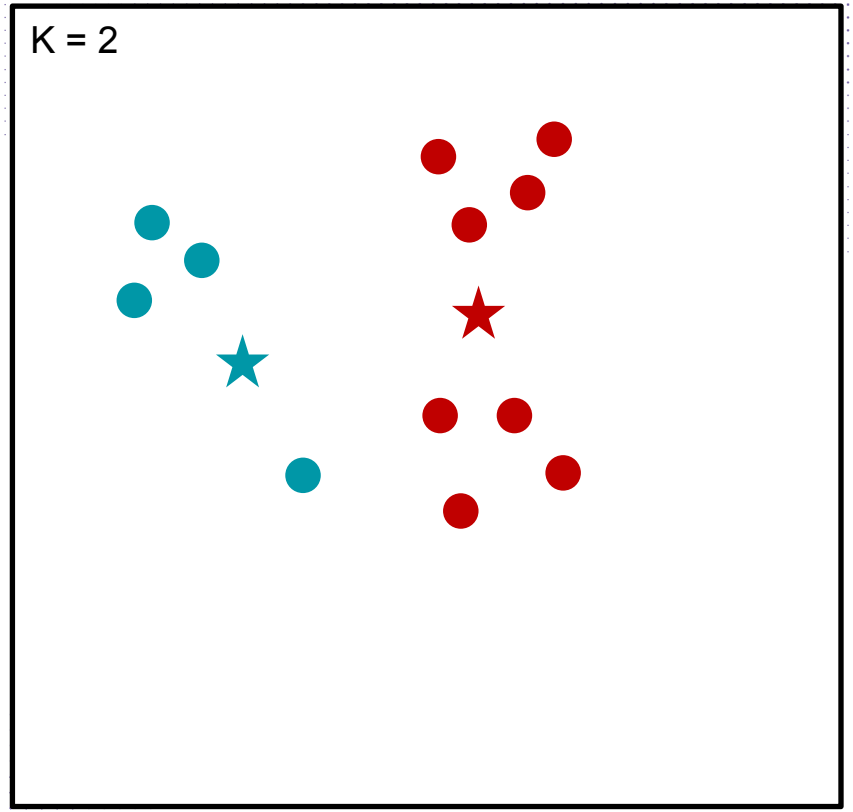
The K-means algorithm aims to choose centroids that minimize the **inertia**, or **within-cluster sum-of-squares** criterion.

Inertia: tells how far away the points within a cluster are.

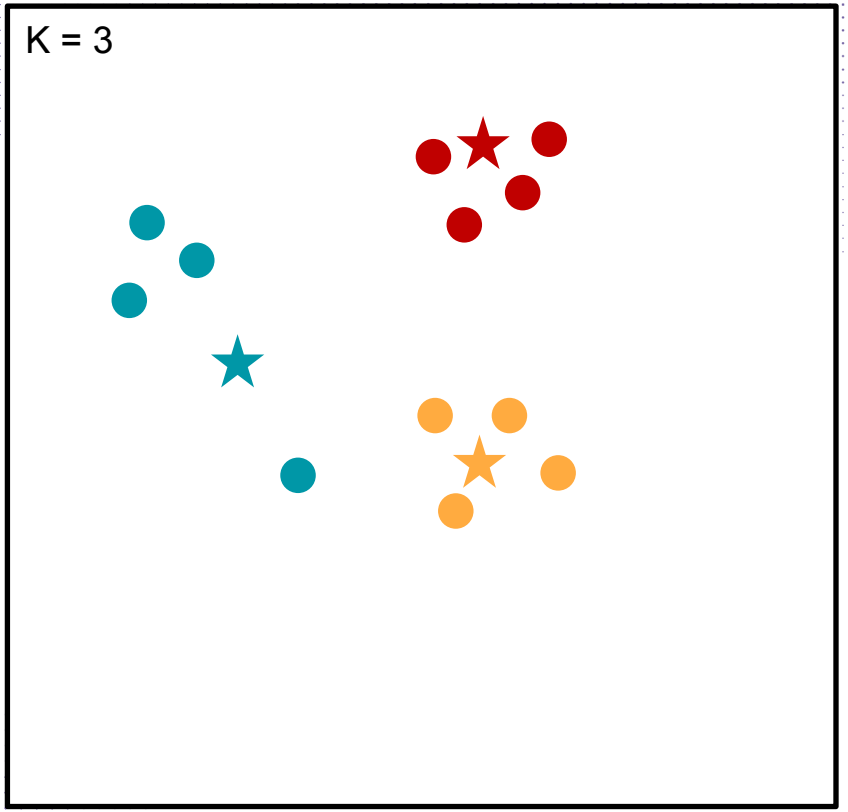
The Elbow Method: minimizing inertia



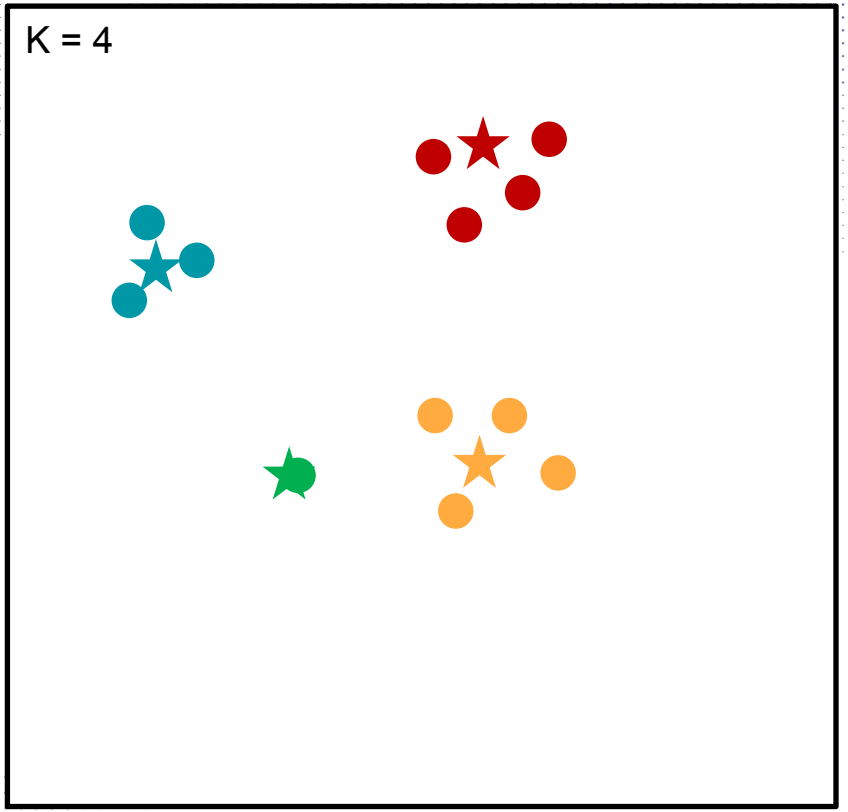
The Elbow Method: minimizing inertia



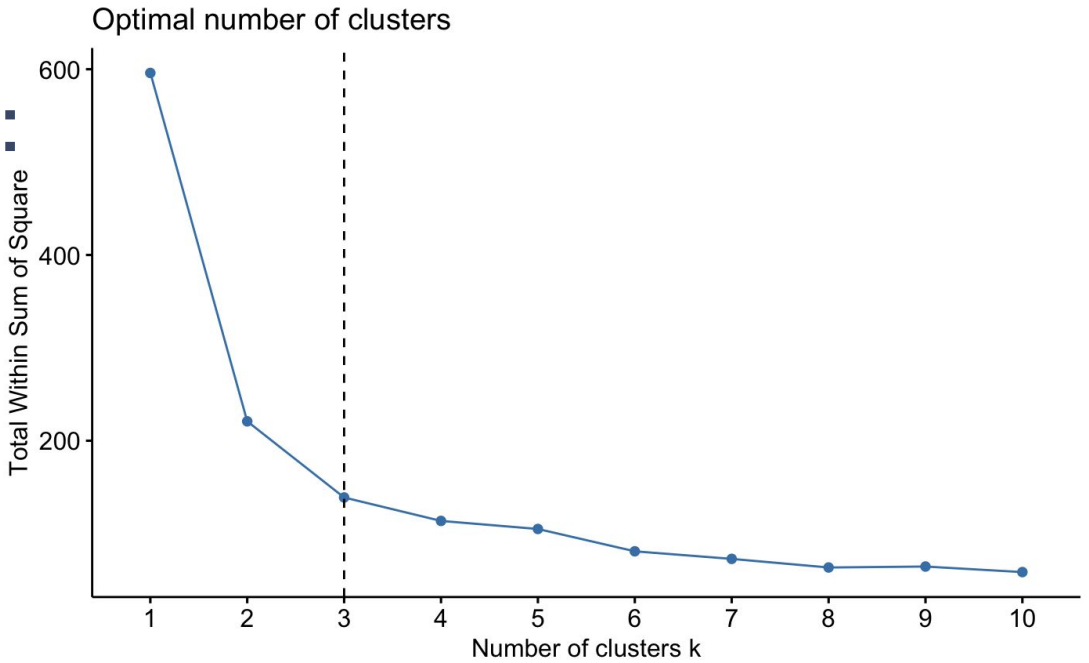
The Elbow Method: minimizing inertia



The Elbow Method: minimizing inertia



The Elbow Method: minimizing inertia



Silhouette score

Tells how far away the datapoints in one cluster are, from the datapoints in another cluster.

The range of silhouette score is from **-1 to 1.** Score should be closer to 1 than -1.

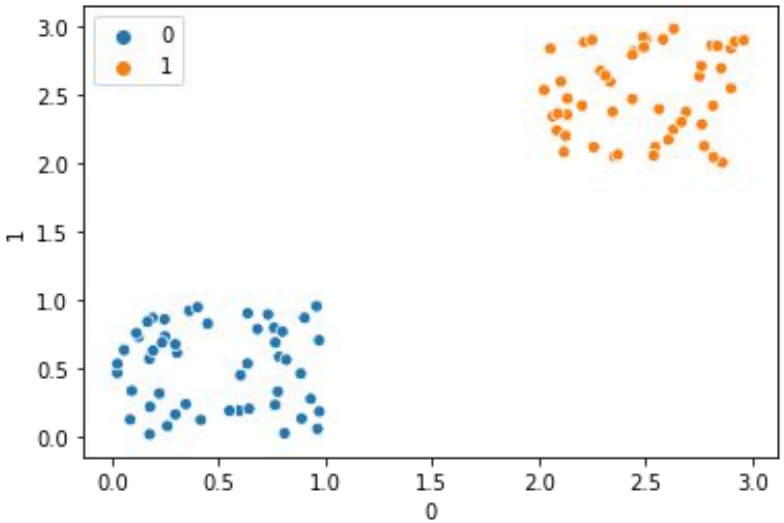
Silhouette score = (x-y)/ max(x,y)

y= average intra-cluster distance (i.e. the average distance between each point within a cluster.)

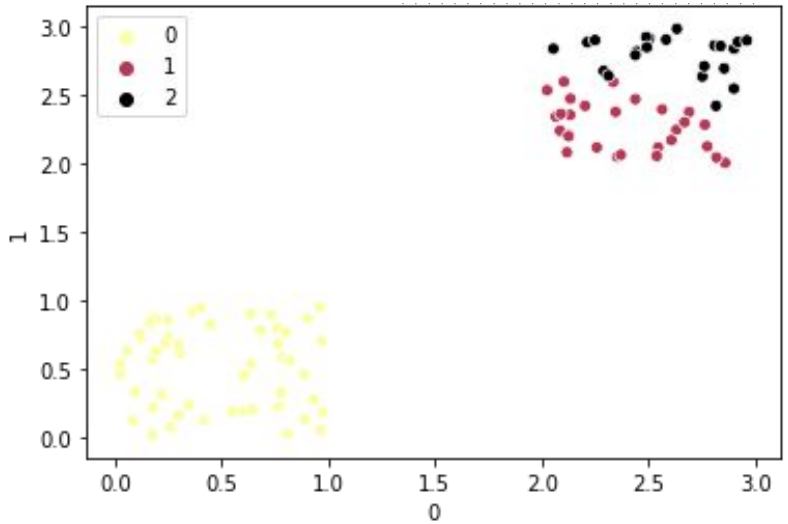
x= average inter-cluster distance (i.e. the average distance between all clusters.)

Silhouette score = (x-y)/ max(x,y)

Silhouette Score(n=2):
0.8062146115881652



Silhouette Score(n=3):
0.5969732708311737



https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

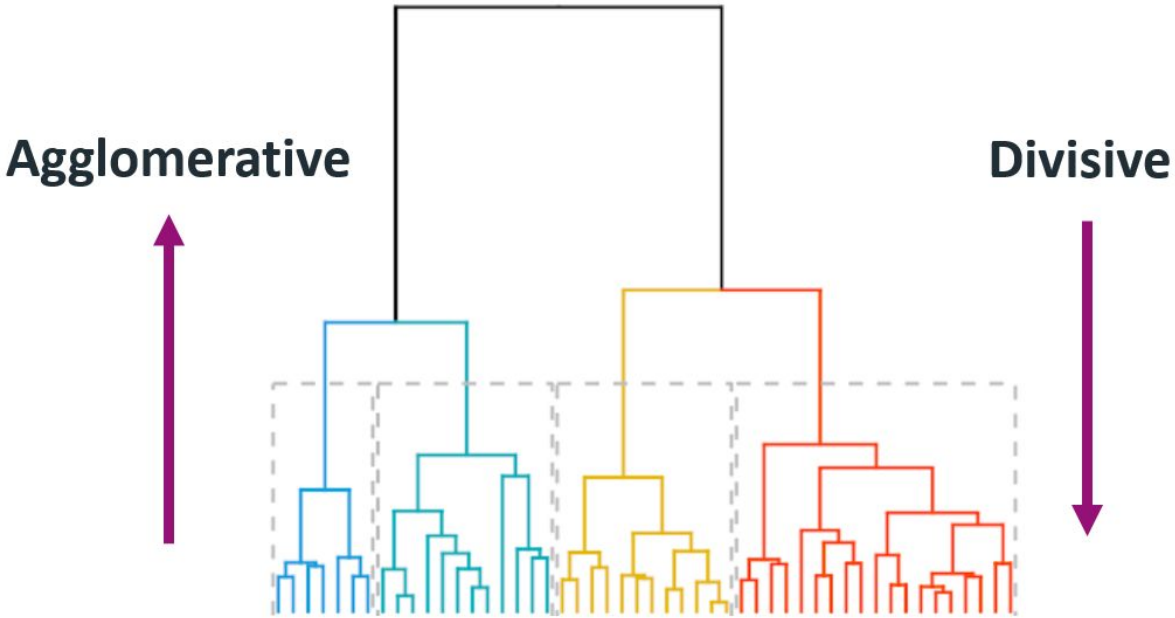
Hierarchical Clustering

Types of Hierarchical Clustering:

1. Divisive: Starts with one cluster that is iteratively split until each point forms its own cluster.

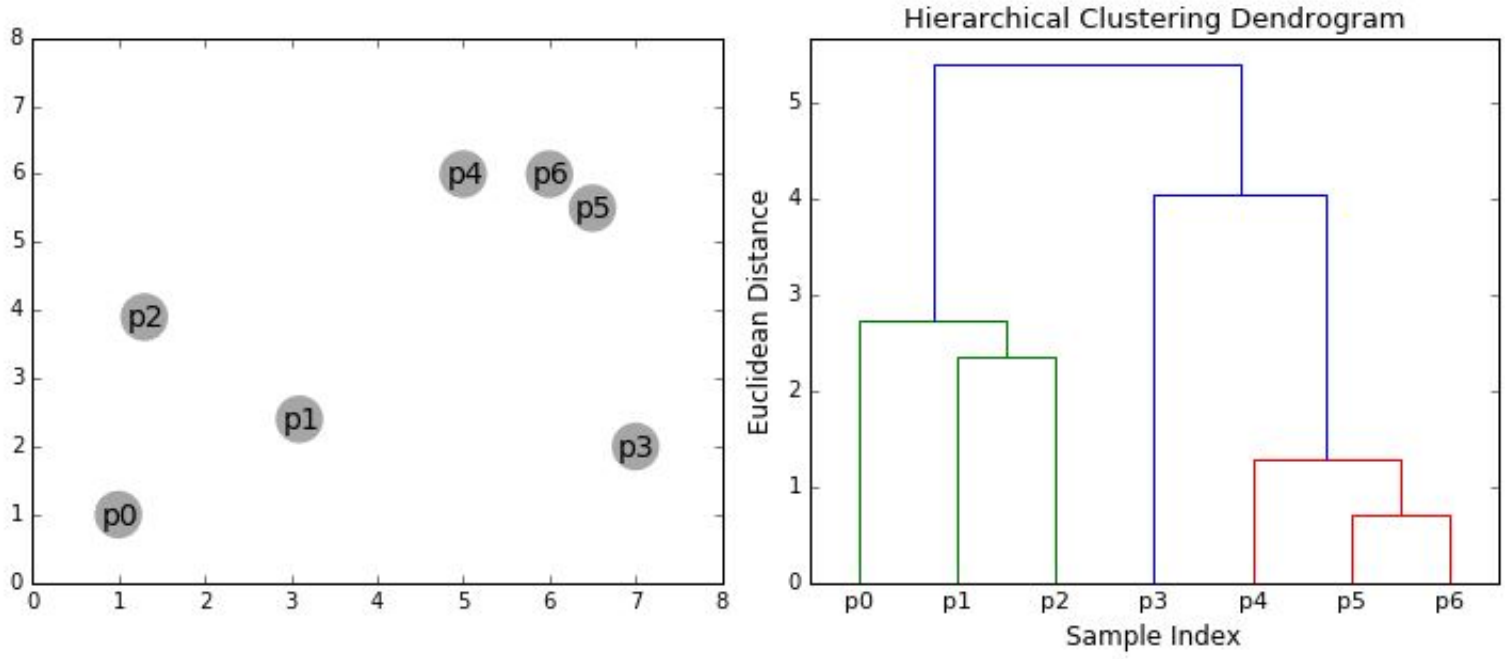
2. Agglomerative: Individual points are iteratively combined until all points belong to the same cluster.

Types of Hierarchical Clustering:



https://www.mygreatlearning.com/blog/hierarchical-clustering/

Agglomerative Clustering



https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Types of Linkages

- **Single Linkage**

  $D(c_1,c_2) = \min D(x_1,x_2)$

  Minimum distance or distance between closest elements in clusters

- **Complete Linkage**

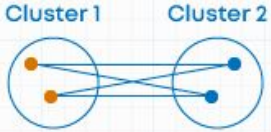
  $D(c_1,c_2) = \max D(x_1,x_2)$

  Maximum distance between elements in clusters

- **Average Linkage**

  $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\Sigma\Sigma D(x_1,x_2)$

  Average of the distances of all pairs



- **Centroid Method**

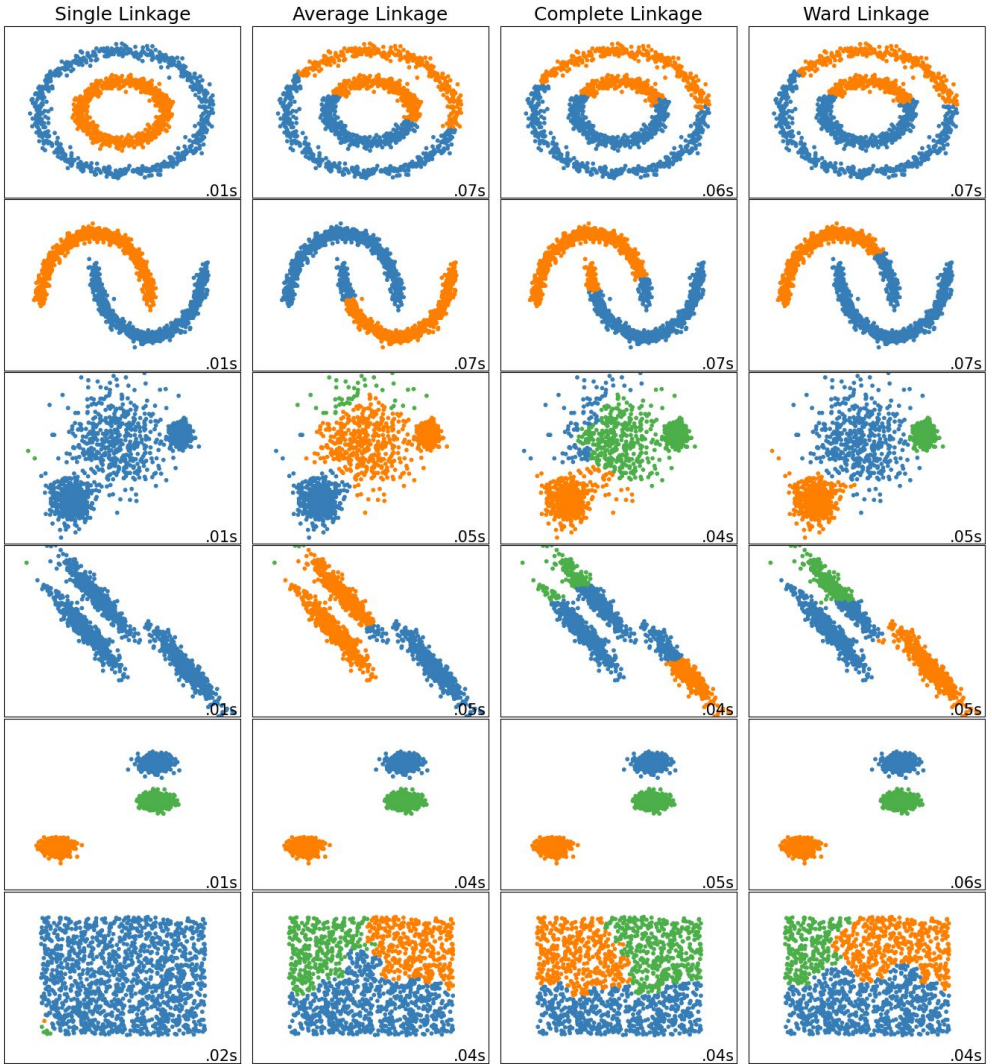
  Combining clusters with minimum distance between the centroids of the two clusters

- **Ward's Method**
  - Combining clusters where increase in within cluster variance is to the smallest degree.

  - Objective is to minimize the total within cluster vairance

https://www.mygreatlearning.com/blog/hierarchical-clustering/

Types of Linkages



https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html

Any Questions?