

MLDA@EEE

AY2022/23



What we do



Listen to our workshops

A stack of various workshop promotional cards from MLDA, including:

- Neural Network**: Introduction to neural networks.
- Robot Operating System**: MLDA PROUDLY PRESENTS.
- EE0005 CRASH REVISION**: Crash revision for EE0005 course.
- Transfer Learning: How to use state-of-the-art**: Workshop on transfer learning.
- MLDA Deep Learning Week INTRODUCTION TO CNNs**: Introduction to Convolutional Neural Networks (CNNs).
- Reinforcement Learning**: Workshop on reinforcement learning.
- NVIDIA Multiple Datatype Workshop**: Workshop on NVIDIA multiple datatype usage.

Each card includes details like date, time, venue, and a QR code for registration.

Do projects that belong to yourselves

A collage of various AI-related event posters and banners, including:

- AI TOURNAMENT BATTLE**: MLDA AI Football Cup 2021.
- ARSC 2021**: AI Research Student Conference.
- MLDA AI Reinforcement Learning through Child Bomber**: Reinforcement learning workshop.
- PRIZES**: Prizes offered for tournaments.
- Portable Hard Disk**: Prize for portable hard disk.
- Workshop**: MLDA AI Reinforcement Learning through Child Bomber.
- Workshop**: MLDA AI Football Cup 2021.
- Prizes**: Prizes for tournament participants.
- Workshop**: MLDA AI Reinforcement Learning through Child Bomber.

Hear from the industry

Two main banners for industry events:

- MLDA Deep Learning Week 2021 Opening Ceremony**: Hosted by Nanyang Technological University, Singapore. Includes a panel discussion on "AI-driven Future of Singapore".
- ARSC 2021**: AI Research Student Conference. Includes sections for keynote speeches, panel discussions, and tutorials.

The ARSC 2021 banner also lists companies attending the virtual career booth, including Shopee, Micron, and Jabil.

What we have



Academic facilities

- GPUs
 - RTX 3090, GTX 2080 Ti, GTX 1080 Ti
- Equipments
 - Robot Master *2
 - 3D printer *1
 - Drones *1
 - ...



Like-minded Students in NTU

- Project Partner / Research Partner
- Friends in similar path of personal development



Industrial Connections

- Sponsors & Partners

Follow us to stay tuned!



<https://www.ntu.edu.sg/eee/student-life/mlda>



Machine Learning and Data Analytics at EEE NTU



Machine Learning and Data Analytics Lab at NTU EEE

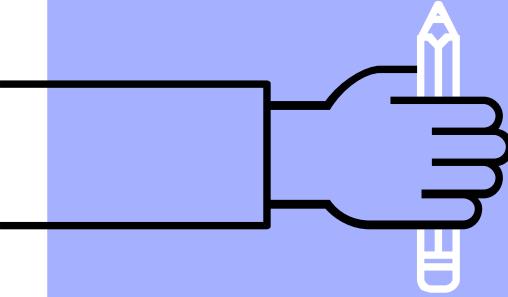
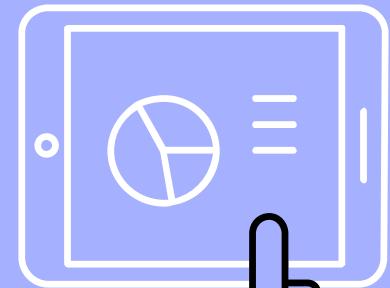
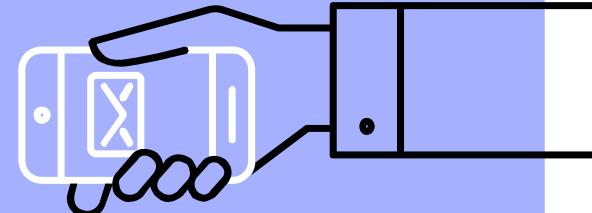


@mlda_at_eee_ntu



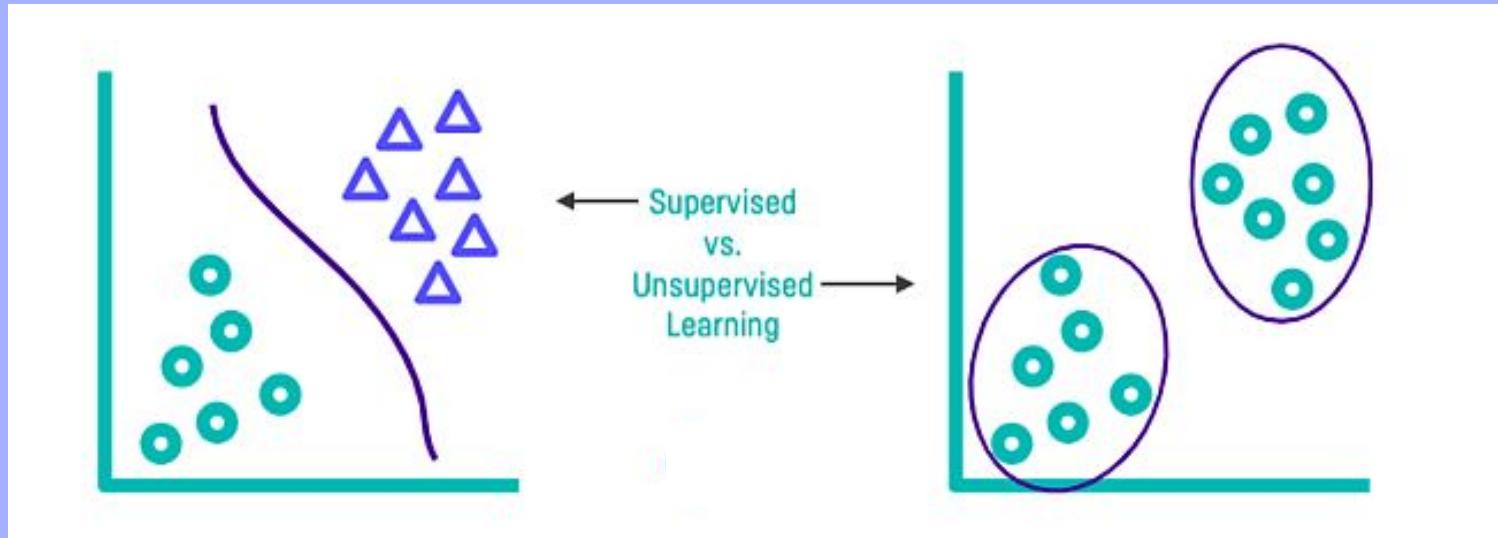
MLDA-NTU

Introduction to Clustering Analysis



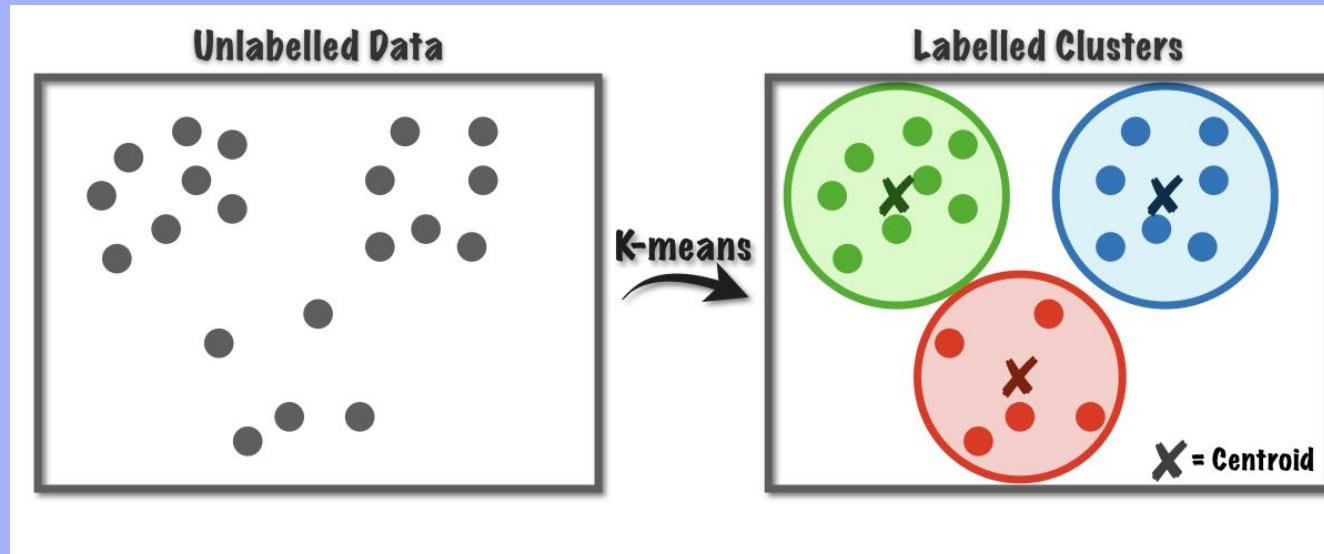
What is Clustering?

Unsupervised Machine Learning algorithm that is performed on unlabelled data



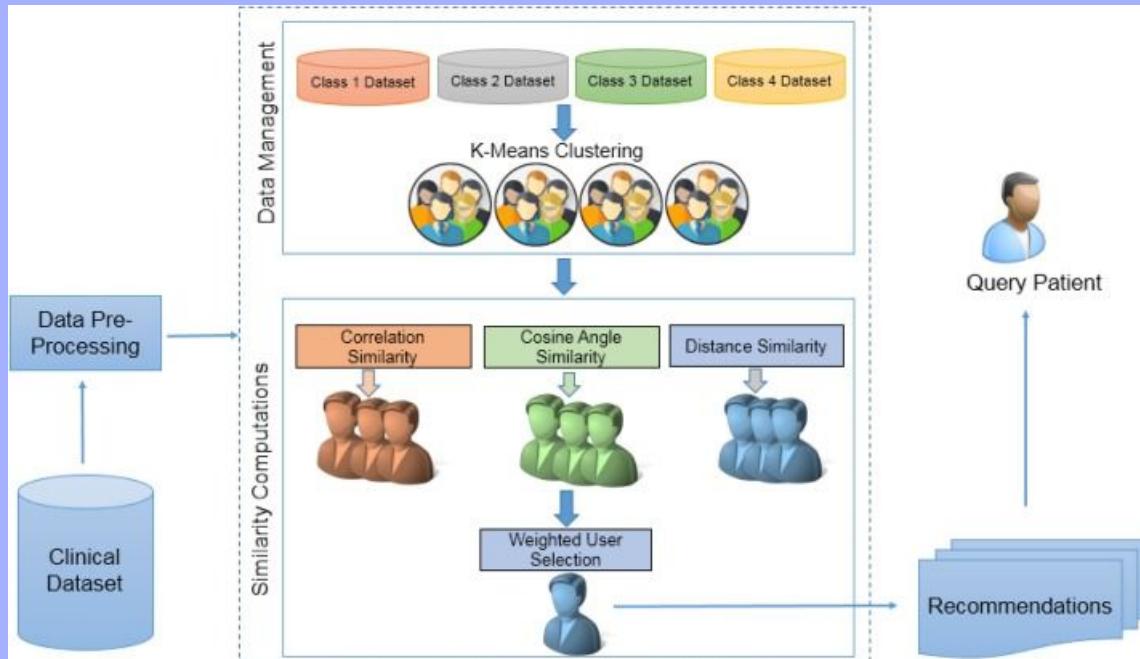
What is Clustering?

Process of grouping objects with similar characteristics in a group (or cluster)



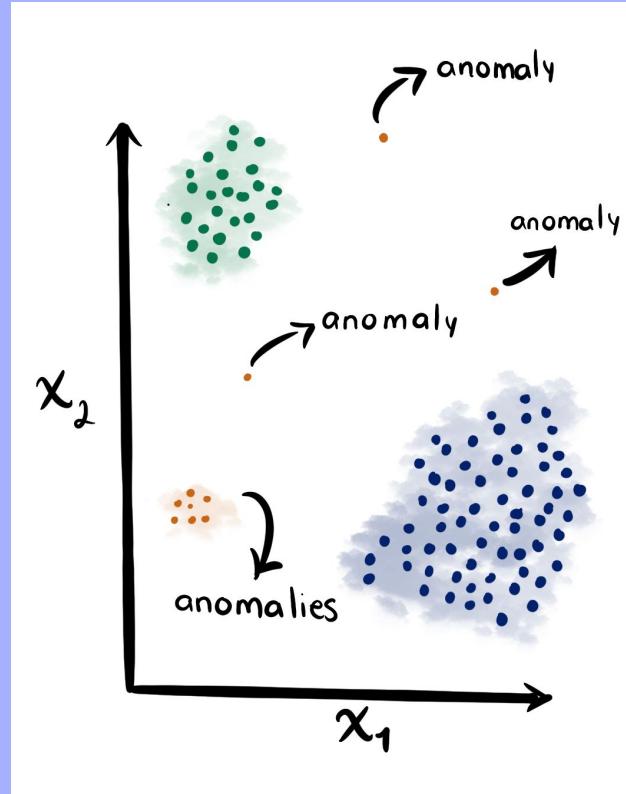
Where is Clustering used?

Recommender Systems



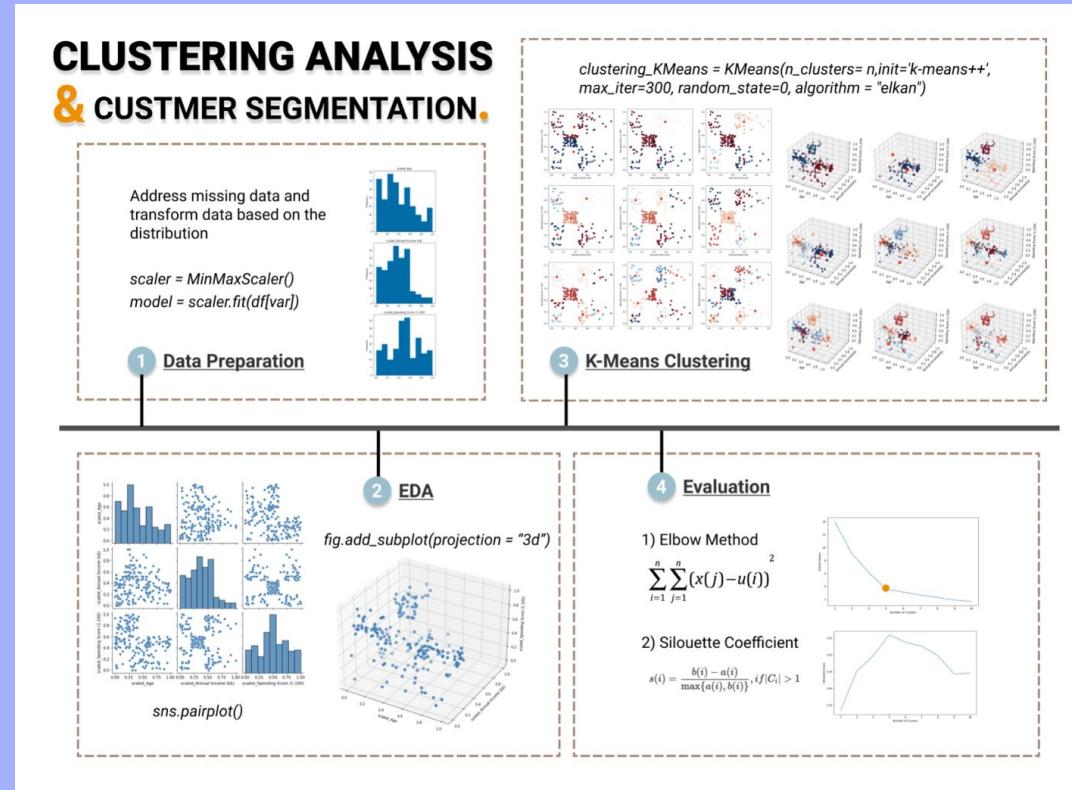
Where is Clustering used?

Anomaly Detection



Where is Clustering used?

Customer Segmentation



Clustering Process



Clustering Models

Centroid Model - K means

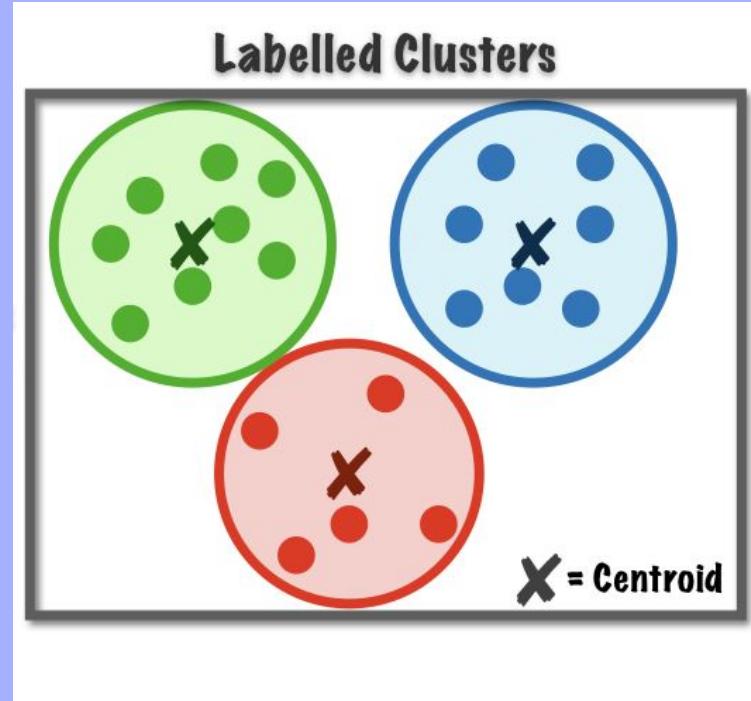
Step 1: Chooses a group of randomly selected centroids (centre points)

These centre points are the beginning points of clusters.

Step 2: Performs iterative (repetitive) calculations to assign a data point to nearest cluster based on its distance from the centroid. Then it will optimize the positions of centroids

Step 3: Halts iterations when

- Centroids have stabilized
- Intra-cluster variance cannot be minimized further



Euclidean distance in K means

K means uses euclidean distance to repeatedly assign data points to the closest centroid

Euclidean Distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

The diagram illustrates the formula for Intra-cluster variance:

$$J = \sum_{j=1}^k \sum_{i=1}^n \|x_i^{(j)} - c_j\|^2$$

Annotations explain the components:

- number of clusters: points to k
- number of cases: points to n
- case i : points to $x_i^{(j)}$
- centroid for cluster j : points to c_j
- Intra-cluster variance: points to the left side of the equation J .
- Distance function: points to the term $\|x_i^{(j)} - c_j\|^2$.

Then the algorithm calculates the centroid or mean of all data points in the cluster. Then it repeats the process until the intra cluster variance or sum of squared error cannot be minimized further.

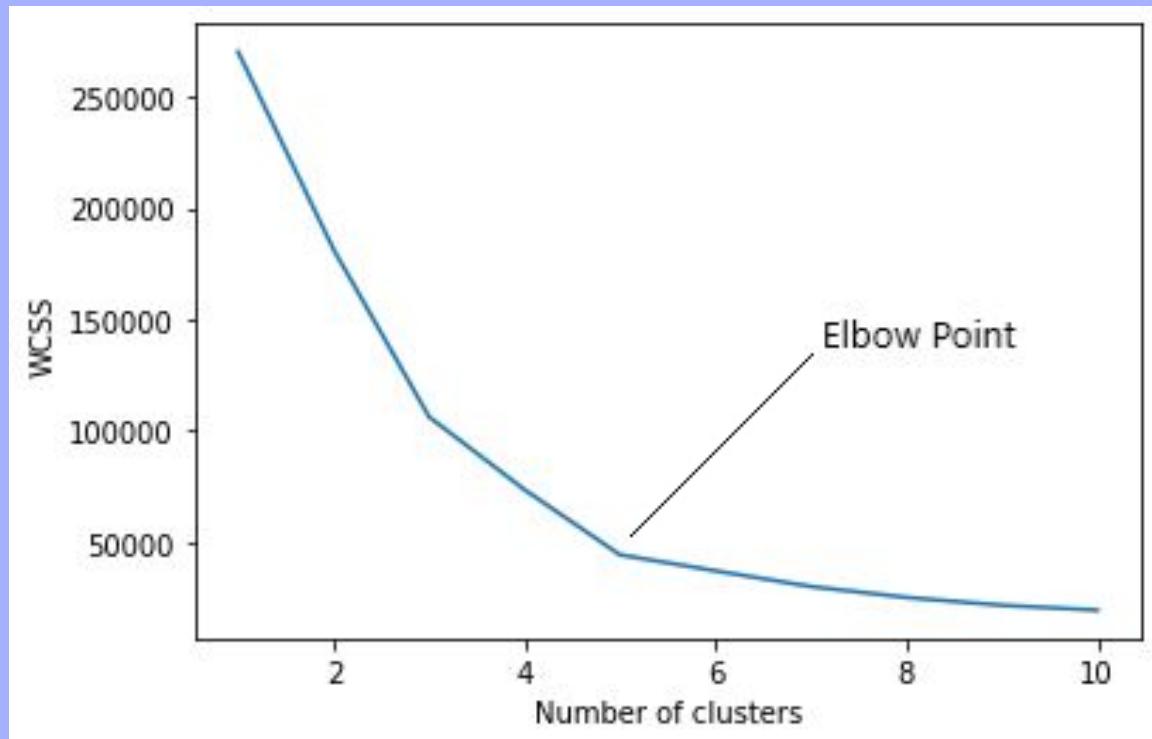
Identifying Optimal Number of Clusters

Elbow Method

Goal: Minimise WCSS

Within Cluster Sum of Square
(WCSS) → Intra-cluster
Variance

Elbow point → Optimal
Number of Clusters



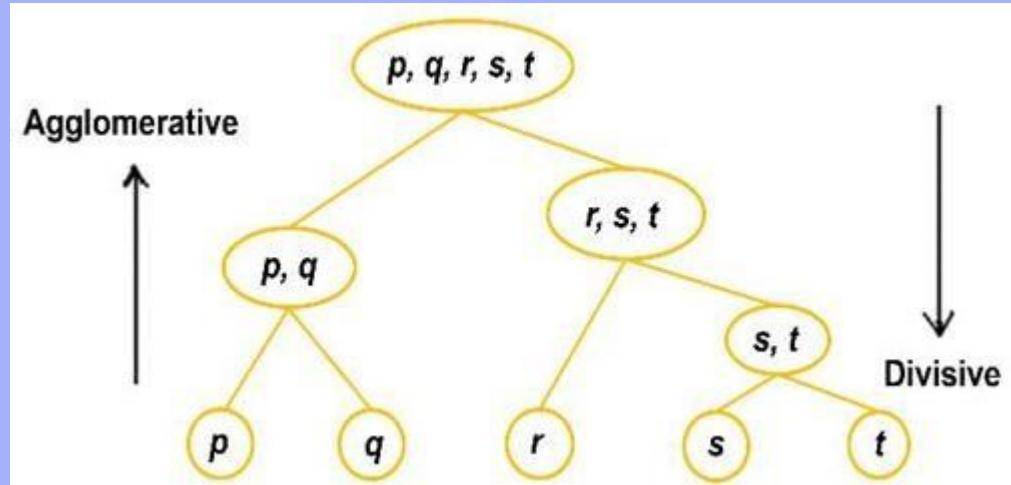
Connectivity Model - Hierarchical Clustering

Finds the nearest similar data point to another point based on how close they are in the data space

2 Types:

Agglomerative - Bottom-up approach

Divisive - Top-down approach



Agglomerative Hierarchical Clustering

Step 1: Compute distance (e.g. Euclidean distance) between each point and store them in a proximity matrix

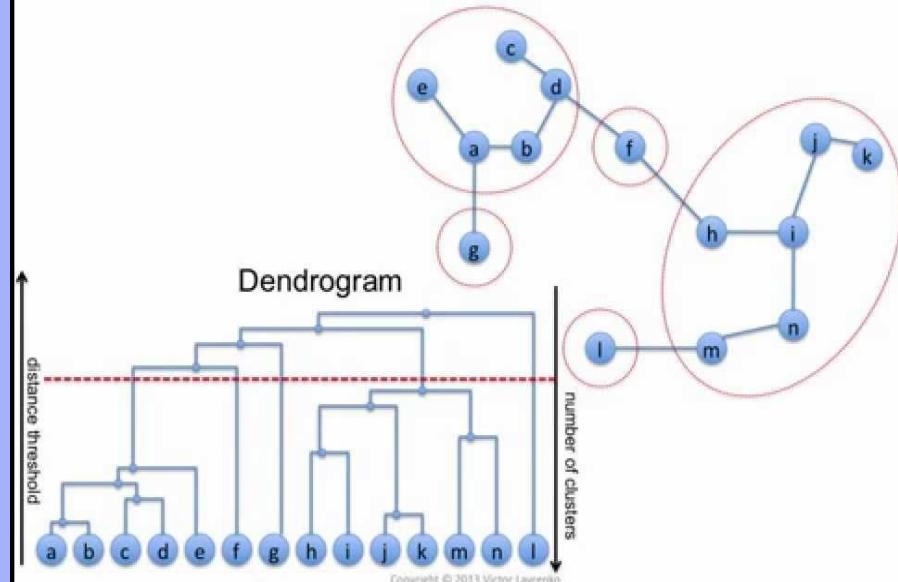
Step 2: Assign each data point as an individual cluster

Step 3: Repetitively merge two closest clusters and update proximity matrix with the distance between each cluster

There are different ways to calculate the distance between each cluster

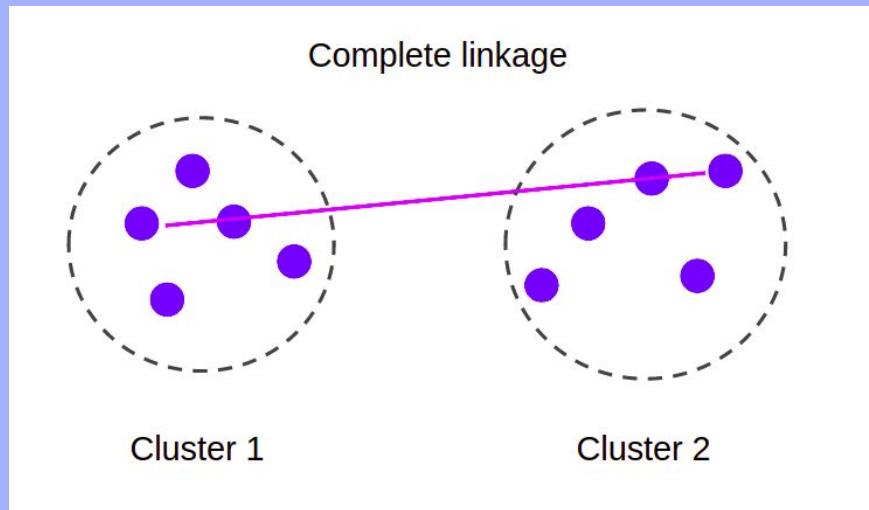
Step 4: Repeat until a single cluster remains

Agglomerative clustering: example



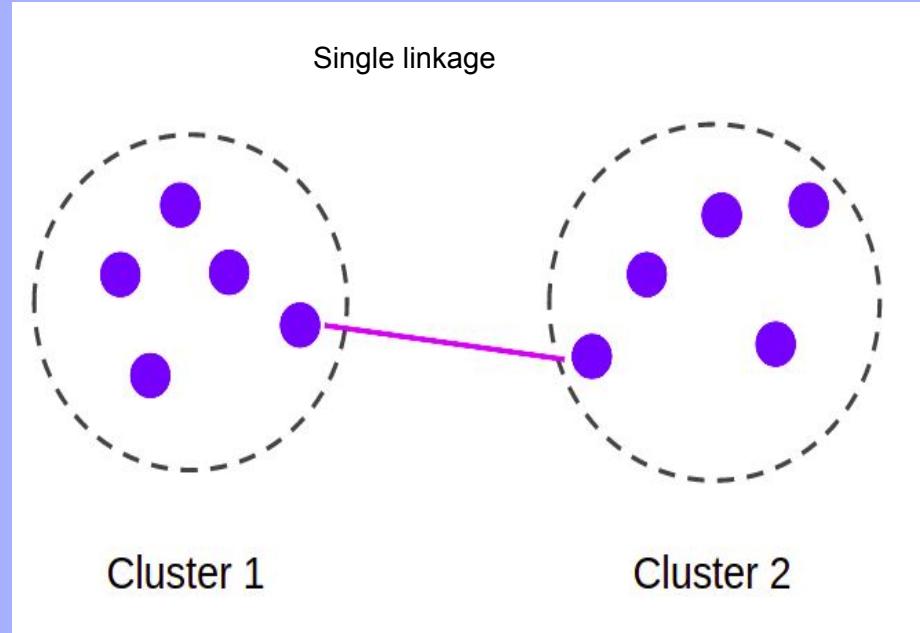
Calculate distance between each cluster

Complete-linkage: calculates the maximum distance between clusters before merging.



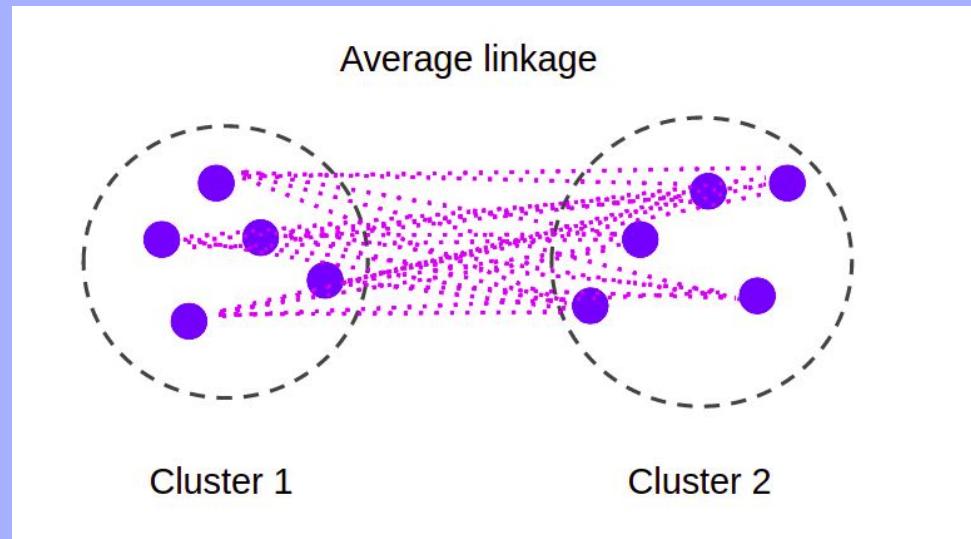
Calculate distance between each cluster

Single-linkage: calculates the minimum distance between the clusters before merging.



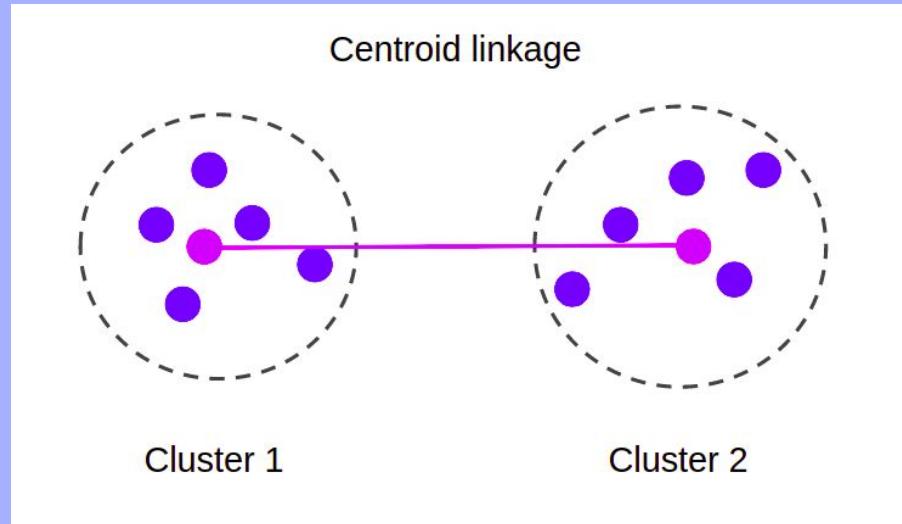
Calculate distance between each cluster

Average-linkage: calculates the average distance between clusters before merging.



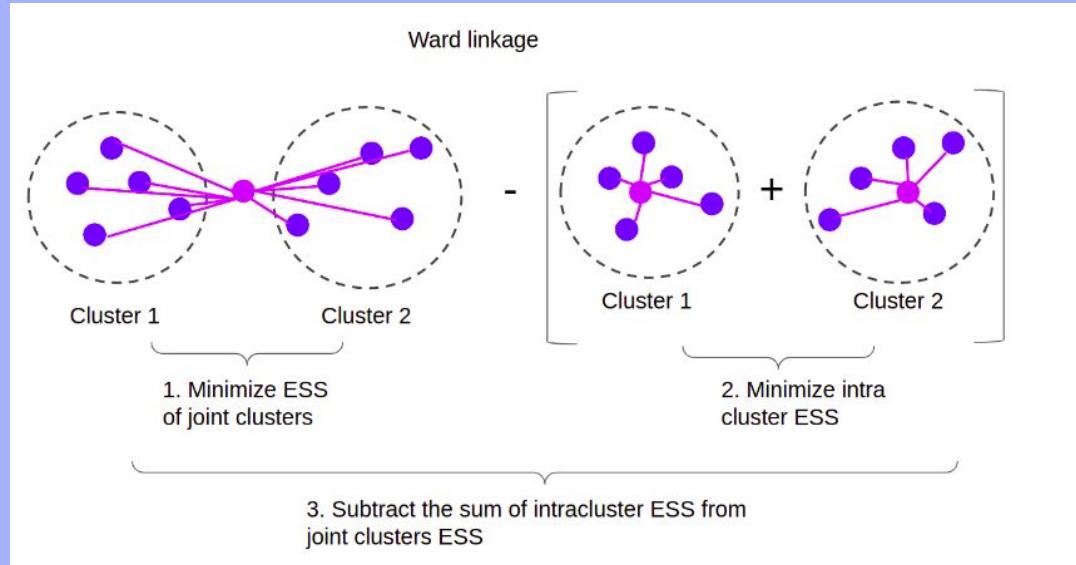
Calculate distance between each cluster

Centroid-linkage: finds centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two before merging.



Calculate distance between each cluster

Ward: specifies the distance between two clusters, computes the intra-cluster variance or sum of squares error (ESS), and successively chooses the next clusters based on the smaller ESS. Ward's Method seeks to minimize the increase of ESS at each step



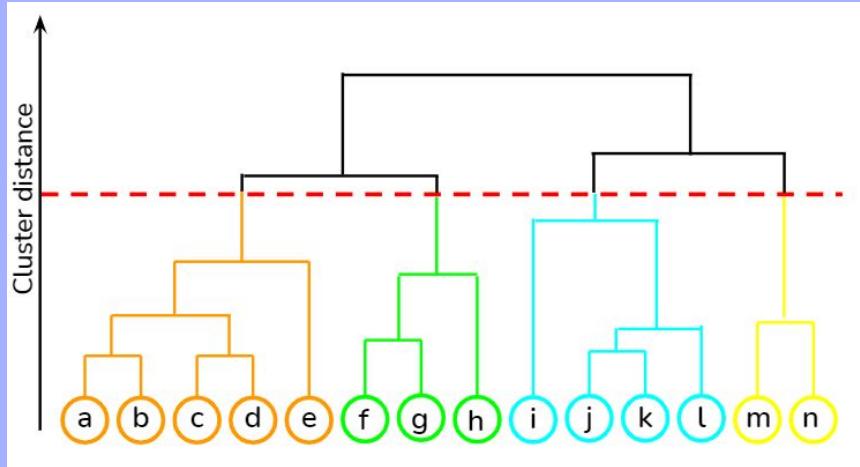
Dendrogram

Branching diagram that represents relationships between similarities among a group of data points

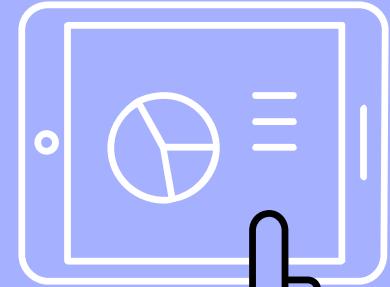
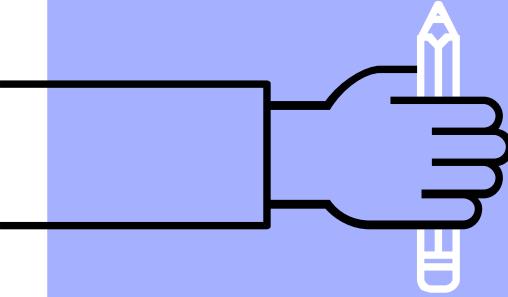
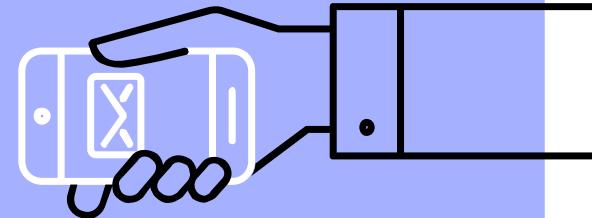
Y-axis: Distance between Clusters
X-axis: Data Points

Identify longest vertical lines in dendrogram

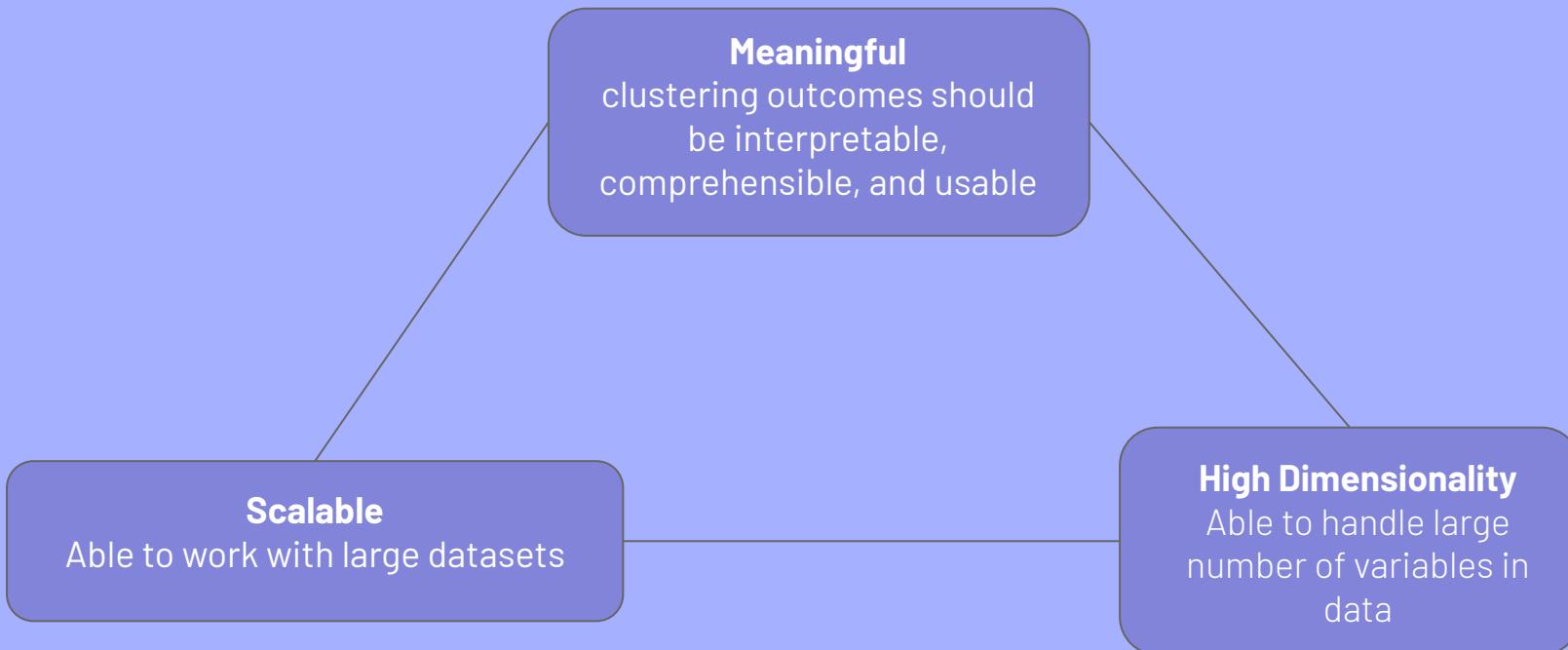
- Imply that data points or clusters are much further away and thus no so similar
- May not be optimal to merge clusters or data points that are too far away



Try it out
yourself!



Clustering Algorithm Properties



THANKS!

Any questions?

We value your feedback :)

Do help us to fill up the
feedback form:

[https://forms.office.com/
r/q09uzNtHr0](https://forms.office.com/r/q09uzNtHr0)

