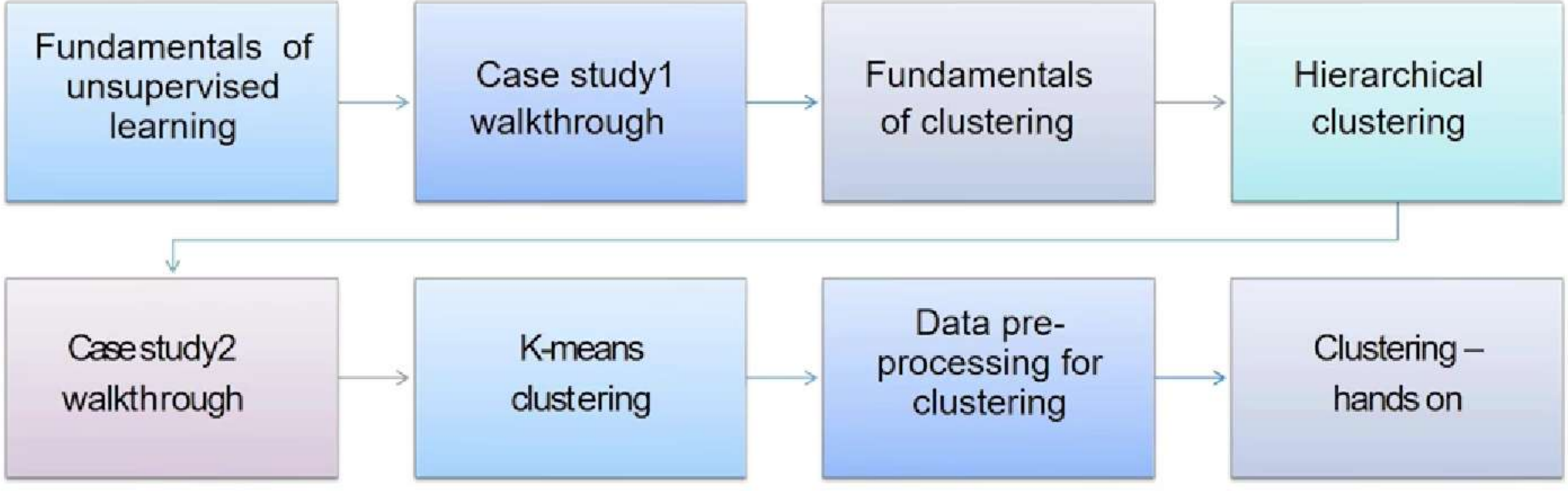
Agenda - Clustering



What is unsupervised learning?

No defined dependent and independent variables.

Patterns in the data are used to identify / group similar observations

Supervised vs unsupervised learning

Supervised learning

- Clearly defined X and Y variables
- Predict a continuous response (Regression)
- categorical response (classification)

Unsupervised learning

- Unlabelled data
- Emerging patterns based on similarity identified
- Clustering
- Association rules (market basket analysis)

What is clustering?

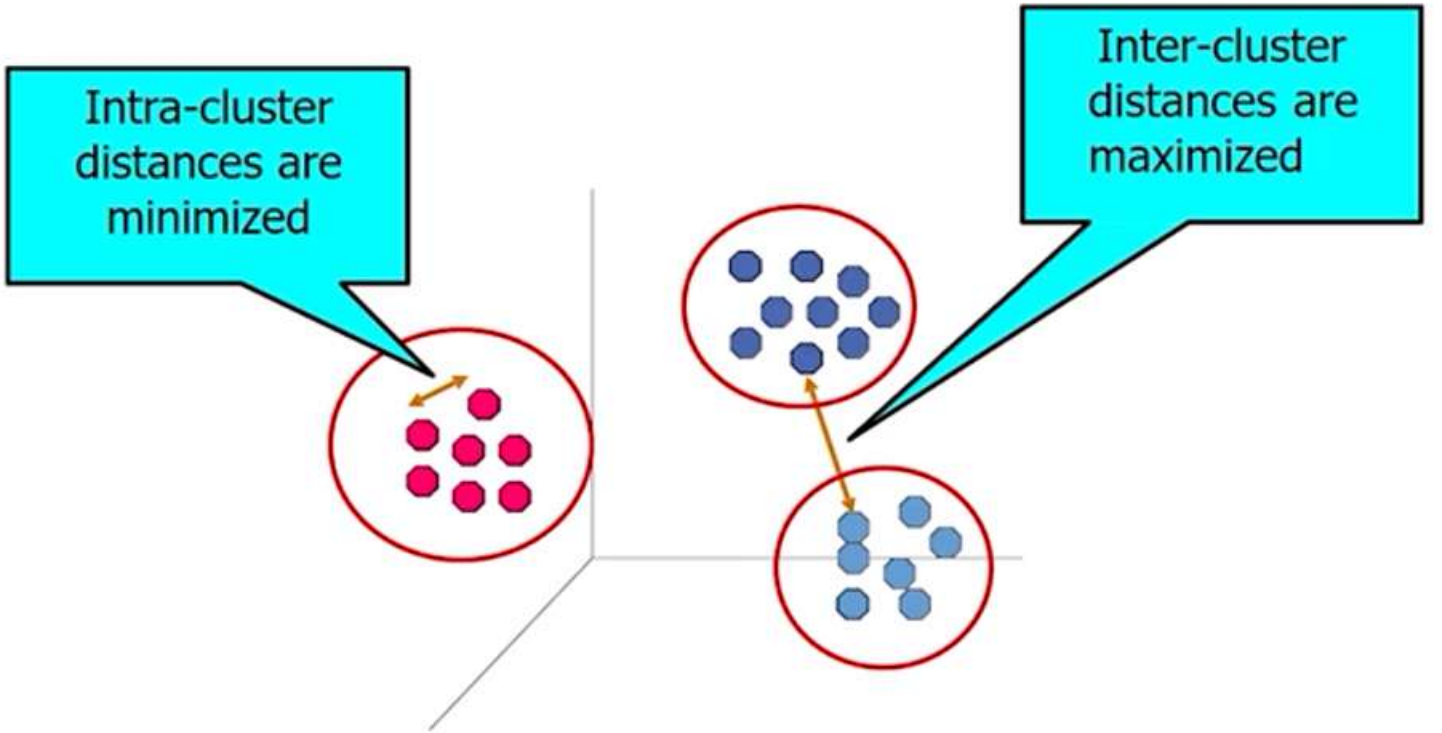
Grouping objects

Heterogeneity between groups

Homogeneity within groups

SSB > SSW

What is clustering?



This is classic case of Unsupervised learning

Why do wecluster?

Group records such that

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

Clustering results are used:

- As a stand-alone tool to get insight into data distribution
- Visualization of clusters may unveil important information
- As a preprocessing step for other algorithms

ClusterAnalysis – Use cases

Image processing	Web	Market Segmentation	Market structure analysis	Finance
• cluster images based on their visual content	• Cluster groups of users based on their access patterns on webpages • Cluster webpages based on their content	• customers are segmented based on demographic and transaction history information, and a marketing strategy is tailored for each segment	• identifying groups of similar products according to competitive measures of similarity	• cluster analysis can be used for creating *balanced portfolios*

Clustering vs PCA

PCA – grouping variables that relate to each other

Clustering – Segment variables according to the distance between them.

		COMP1			COMP2				COMP		
	AID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
CLUSTER1	1	2.51	9.19	4.45	5.33	7.27	0.7	5.85	4.01	1.34	6.1
	2	7.51	1.77	2.01	9.31	6.61	7.69	3.29	8.85	0.2	6.35
	3	2.52	2.61	5.65	1.24	0.97	2.85	9.87	3.14	3.7	5.17
	4	6.56	5.9	1.65	6.69	8.04	0.8	1.91	7.42	8.02	1.43
	5	6.91	7.78	5.63	3.84	8.99	1.56	0.13	7.29	6.45	9.58
CLUSTER2	6	2.63	3.16	1.39	0.55	9.85	4.58	0.97	5.89	0.04	3.88
	7	3.78	9.9	5.07	5.41	3.27	4.04	2.11	9.47	4.98	0.32
	8	5.63	6.86	9.24	4.47	5.46	7.05	7.7	9.21	7.99	9.51
	9	6.09	8.36	1.03	1.81	0.58	2.02	9.86	8.2	0.81	0.25
	10	2.26	3.48	7.69	0.9	6.07	0.74	2.31	6.48	0.45	6.78
CLUSTER3	11	3.79	2.52	2.93	1.92	7.12	4.22	2.07	6.73	1.35	6.64
	12	6.37	5.13	4.09	1.39	3.74	3.67	5.46	4.17	1.6	0.92
	13	3.9	8.14	8.91	4.7	8.73	8.5	5.75	6.76	0.17	5.08
	14	2.07	3.23	2.8	0.43	8.51	0.48	2.52	8.83	0.01	0.37
	15	1.39	8.66	3.57	6.68	2.54	4.89	7.27	2.75	7.43	9.89

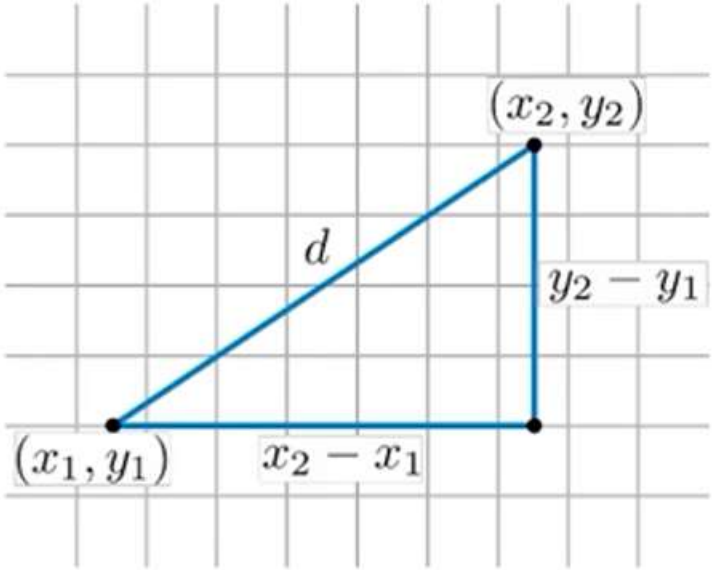
Measuring similarity - Distances

- Eucledian distance
- Manhattan distance
- Chebyshev distance

Euclidean distance

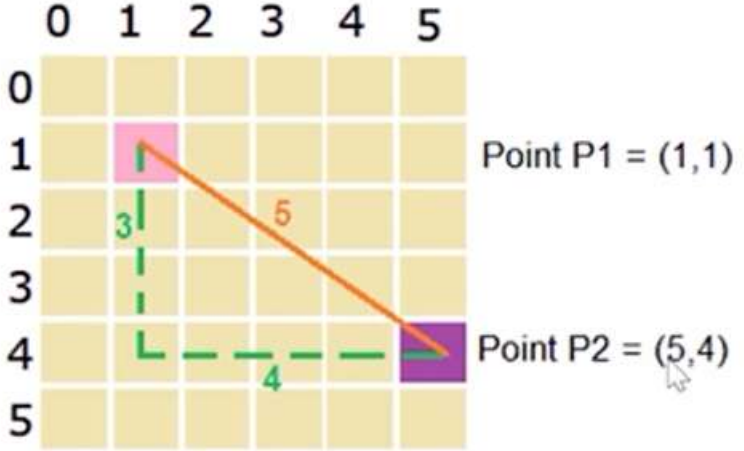
$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

Where x_1 to x_p are the independent variables of i and j



Manhattan distance (city –block distance)

- Distance between the projection of points on the axis.



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

Chebyshev Distance (chessboard distance)

- Max (|x1-x2|, |y1-y2|, |z1-z2|, ….)

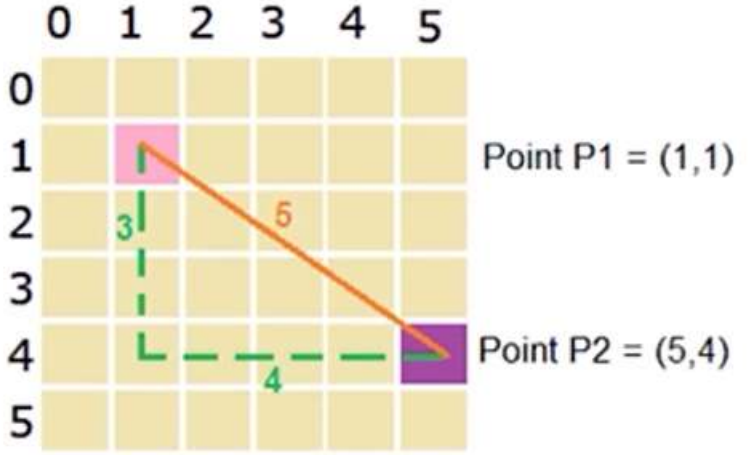
Minkowski Distance

Mathematical formula: $\left(\sum_{i=1}^{m}|x_i - y_i|^p\right)^{1/p}$

- If p=2, then the above equation resembles the equation of Euclidean Distance.
- If p=1, then the above equation resembles the equation of Manhattan Distance.

Manhattan distance (city –block distance)

- Distance between the projection of points on the axis.



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

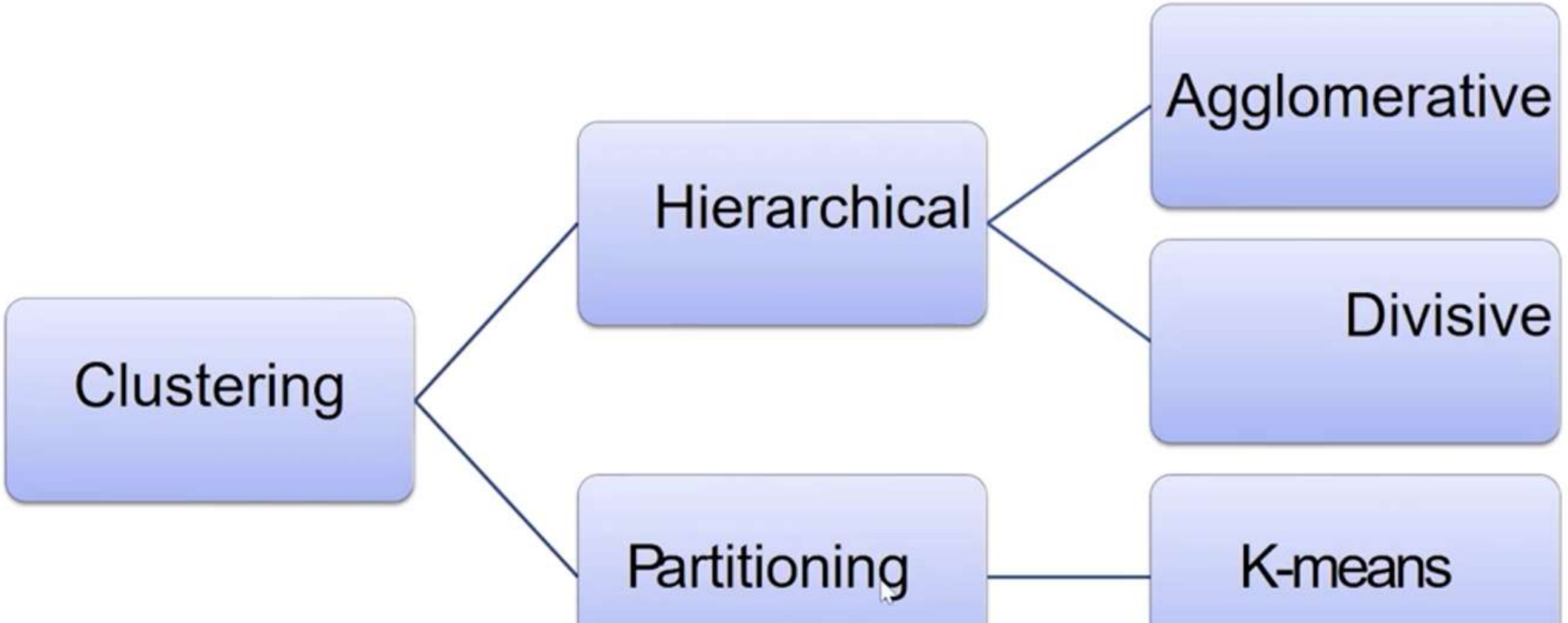
Manhattan distance = $|5-1| + |4-1| = 7$

Minkowski Distance

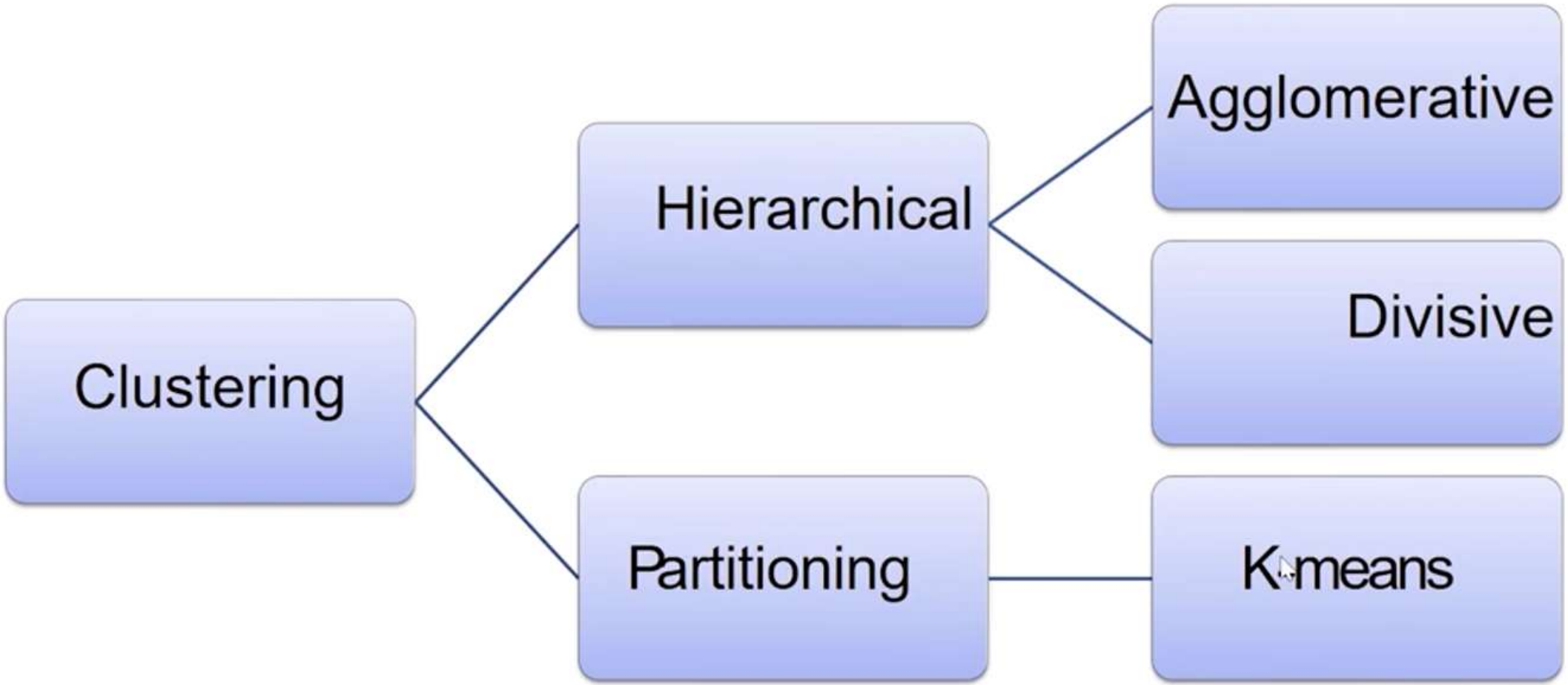
Mathematical formula: $\left(\sum_{i=1}^{m}|x_i - y_i|^p\right)^{1/p}$

- If p=2, then the above equation resembles the equation of Euclidean Distance.
- If p=1, then the above equation resembles the equation of Manhattan Distance.

Types of clustering



Types of clustering



Clustering types

Agglomerative clustering

- Bottom up approach
- start with each object forming a separate group
- It keeps on merging the objects or groups that are close to one another

Divisive approach

- Top down approach
- start with all of the objects in the same cluster
- a cluster is split up into smaller clusters

Partitioning

- constructs ‘k’ partition of data
- Each partition will represent a cluster and $k \leq n$