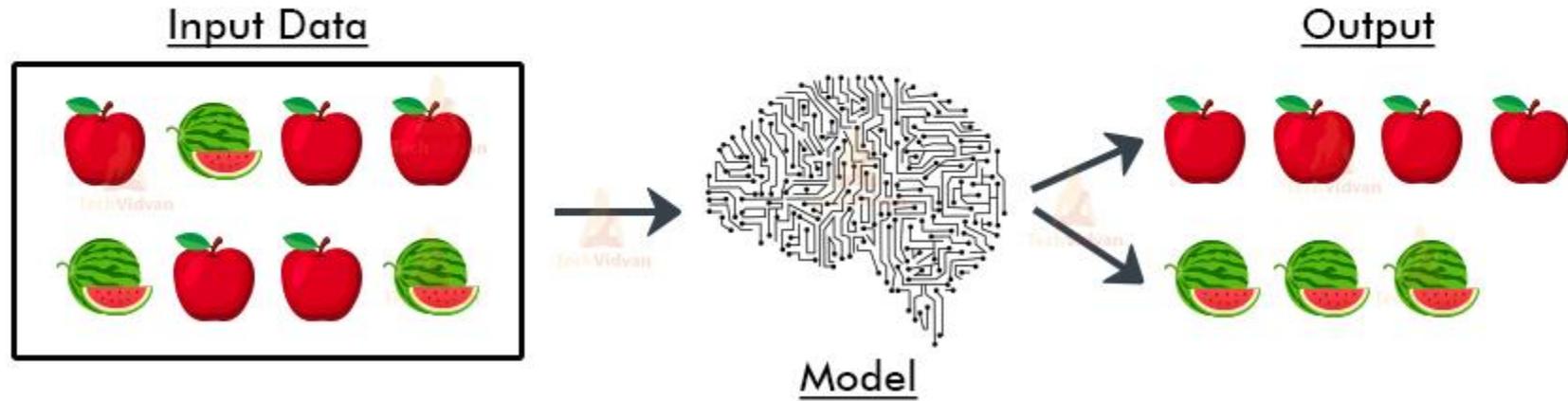


Unsupervised Learning in ML



Clustering

Clustering is the process of dividing the dataset into groups, consisting of similar data-points

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

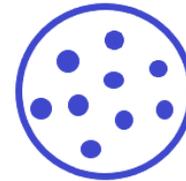
Periodic Table of the Elements

1 1IA 11A	2 IIA 2A											13 IIIA 3A	14 IVA 4A	15 VA 5A	16 VIA 6A	17 VIIA 7A	18 VIIIA 8A	
1 H Hydrogen 1.0079													5 B Boron 10.811	6 C Carbon 12.011	7 N Nitrogen 14.00674	8 O Oxygen 15.9994	9 F Fluorine 18.998403	10 Ne Neon 20.1797
3 Li Lithium 6.941	4 Be Beryllium 9.01218												13 Al Aluminum 26.981539	14 Si Silicon 28.0855	15 P Phosphorus 30.973762	16 S Sulfur 32.066	17 Cl Chlorine 35.4527	18 Ar Argon 39.948
11 Na Sodium 22.989768	12 Mg Magnesium 24.305	3 IIIB 3B	4 IVB 4B	5 VB 5B	6 VIB 6B	7 VIIB 7B	8	9 VIII 8	10	11 IB 1B	12 IIB 2B		31 Ga Gallium 69.723	32 Ge Germanium 72.64	33 As Arsenic 74.9216	34 Se Selenium 78.96	35 Br Bromine 79.904	36 Kr Krypton 83.80
19 K Potassium 39.0983	20 Ca Calcium 40.078	21 Sc Scandium 44.95591	22 Ti Titanium 47.88	23 V Vanadium 50.9415	24 Cr Chromium 51.9961	25 Mn Manganese 54.938	26 Fe Iron 55.847	27 Co Cobalt 58.9332	28 Ni Nickel 58.6934	29 Cu Copper 63.546	30 Zn Zinc 65.38		49 In Indium 114.818	50 Sn Tin 118.71	51 Sb Antimony 121.750	52 Te Tellurium 127.6	53 I Iodine 126.90447	54 Xe Xenon 131.29
37 Rb Rubidium 85.4678	38 Sr Strontium 87.62	39 Y Yttrium 88.90585	40 Zr Zirconium 91.224	41 Nb Niobium 92.90638	42 Mo Molybdenum 95.94	43 Tc Technetium 98.9062	44 Ru Ruthenium 101.07	45 Rh Rhodium 102.9055	46 Pd Palladium 106.42	47 Ag Silver 107.8682	48 Cd Cadmium 112.411		81 Tl Thallium 204.3833	82 Pb Lead 207.2	83 Bi Bismuth 208.98037	84 Po Polonium 209	85 At Astatine 208.9804	86 Rn Radon 222.0175
55 Cs Cesium 132.90545	56 Ba Barium 137.327	57-71 Lanthanide Series	72 Hf Hafnium 178.49	73 Ta Tantalum 180.94789	74 W Tungsten 183.85	75 Re Rhenium 186.207	76 Os Osmium 190.23	77 Ir Iridium 192.22	78 Pt Platinum 195.08	79 Au Gold 196.96655	80 Hg Mercury 200.59		113 Uut Ununtrium unknown	114 Uuq Ununquadium unknown	115 Uup Ununpentium unknown	116 Uuh Ununhexium unknown	117 Uus Ununseptium unknown	118 Uuo Ununoctium unknown
87 Fr Francium 223.0197	88 Ra Radium 226.0254	89-103 Actinide Series	104 Rf Rutherfordium [261]	105 Db Dubnium [262]	106 Sg Seaborgium [266]	107 Bh Bohrium [264]	108 Hs Hassium [265]	109 Mt Meitnerium [268]	110 Ds Darmstadtium [269]	111 Rg Roentgenium [272]	112 Cn Copernicium [277]							
			57 La Lanthanum 138.9055	58 Ce Cerium 140.116	59 Pr Praseodymium 140.90766	60 Nd Neodymium 144.24	61 Pm Promethium 144.9127	62 Sm Samarium 150.36	63 Eu Europium 151.965	64 Gd Gadolinium 157.25	65 Tb Terbium 158.92534	66 Dy Dysprosium 162.50	67 Ho Holmium 164.93032	68 Er Erbium 167.26	69 Tm Thulium 168.93421	70 Yb Ytterbium 173.04	71 Lu Lutetium 174.967	
			89 Ac Actinium 227.0287	90 Th Thorium 232.0381	91 Pa Protactinium 231.0362	92 U Uranium 238.02891	93 Np Neptunium 237.0482	94 Pu Plutonium 244.0642	95 Am Americium 243.0614	96 Cm Curium 247.0772	97 Bk Berkelium 247.0772	98 Cf Californium 251.0772	99 Es Einsteinium [252]	100 Fm Fermium [257]	101 Md Mendelevium [258]	102 No Nobelium [259]	103 Lr Lawrencium [262]	
			Alkali Metal	Alkaline Earth	Transition Metal	Basic Metal	Semimetals	Nonmetals	Halogens	Noble Gas	Lanthanides	Actinides						

Types of Clustering

1. Exclusive Clustering
2. Overlapping Clustering
3. Hierarchical Clustering

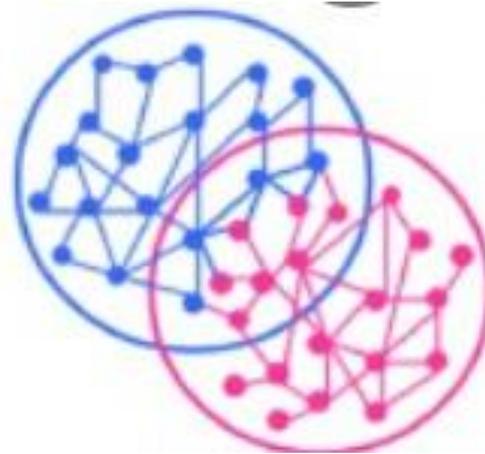
Exclusive Clustering: Exclusive Clustering is the hard clustering in which data point exclusively belongs to one cluster.
For example, K-Means Clustering.



Overlapping Clustering:

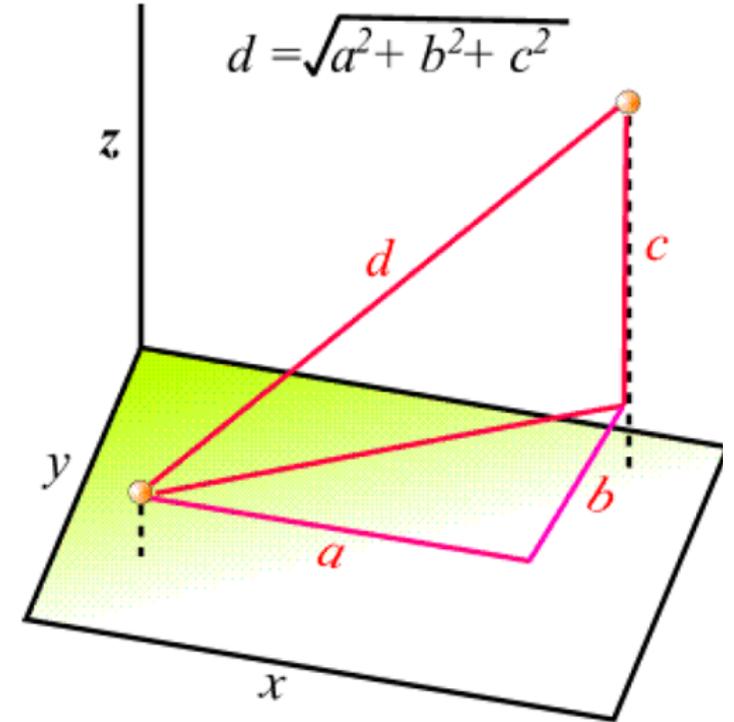
Overlapping clustering is the soft cluster in which data point belongs to multiple clusters.

For example, C-Means Clustering.



we can see that some of the blue data points and some of the pink data points are overlapped.

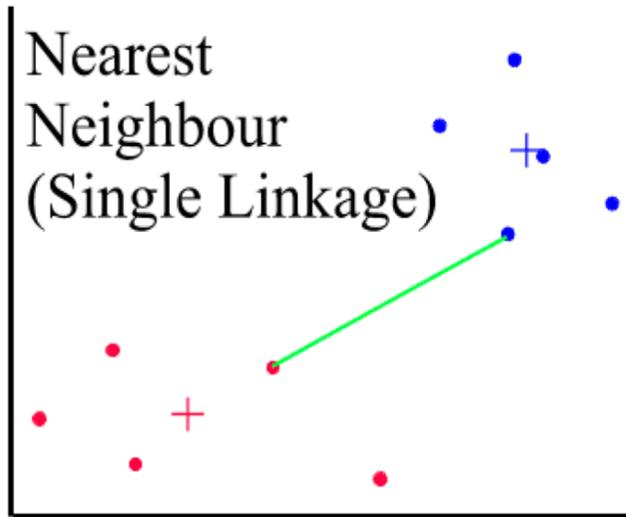
Euclidean Distance



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

Distances Between Clusters: 'single linkage' ('nearest neighbor')

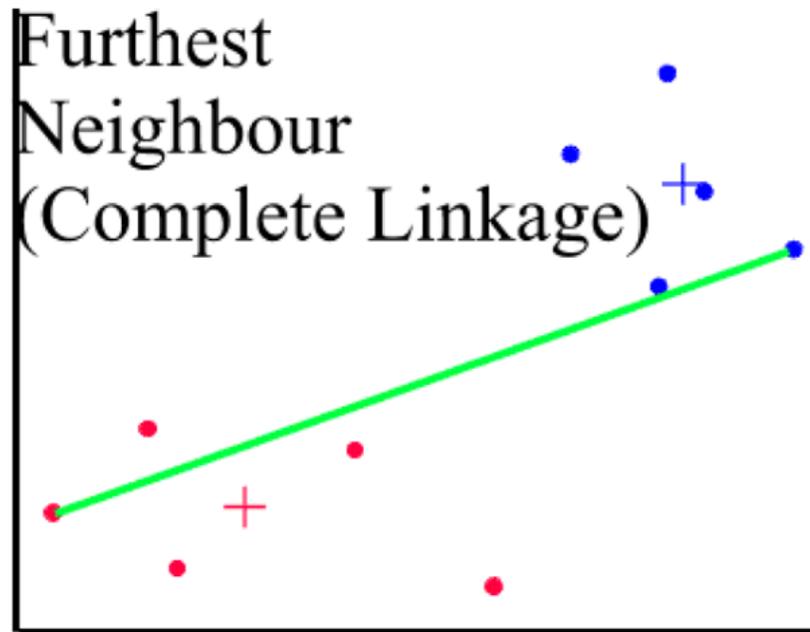
Distance between 2 clusters = **minimum distance** between members of the two clusters



$$\Delta(\mathbf{C}_\alpha, \mathbf{C}_\beta) = \min_{\mathbf{x} \in \mathbf{C}_\alpha, \mathbf{y} \in \mathbf{C}_\beta} \{ \Delta(\mathbf{x}, \mathbf{y}) \}$$

Distances Between Clusters: 'complete linkage' ('farthest neighbor')

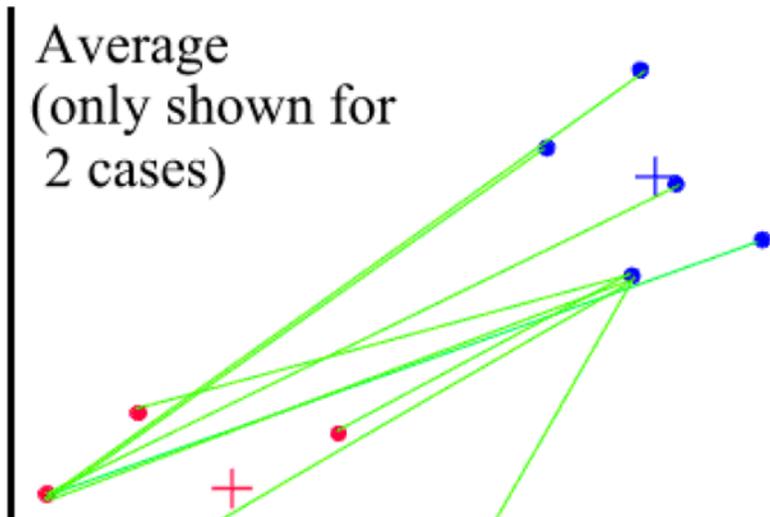
Distance between 2 clusters = **greatest distance** between members of the two clusters



$$\Delta(\mathbf{C}_\alpha, \mathbf{C}_\beta) = \max_{\mathbf{x} \in \mathbf{C}_\alpha, \mathbf{y} \in \mathbf{C}_\beta} \{ \Delta(\mathbf{x}, \mathbf{y}) \}$$

Distances Between Clusters: 'average linkage'

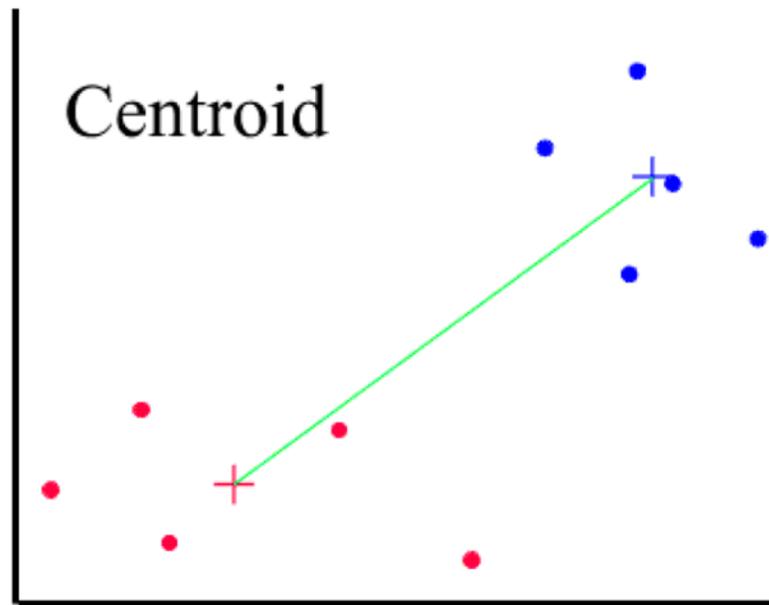
Distance between 2 clusters = **average** of all distances between members of the two clusters



$$\Delta(\mathbf{C}_\alpha, \mathbf{C}_\beta) = \frac{1}{|\mathbf{C}_\alpha| |\mathbf{C}_\beta|} \sum_{\mathbf{x} \in \mathbf{C}_\alpha} \sum_{\mathbf{y} \in \mathbf{C}_\beta} \Delta(\mathbf{x}, \mathbf{y})$$

Distances Between Clusters: 'centroid linkage'

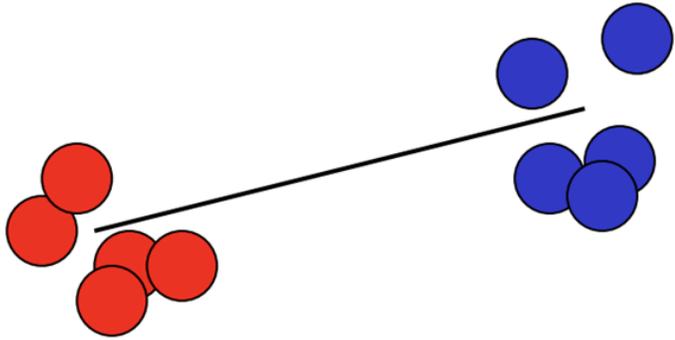
Distance between 2 clusters =
distance between their **centroids** (centers)



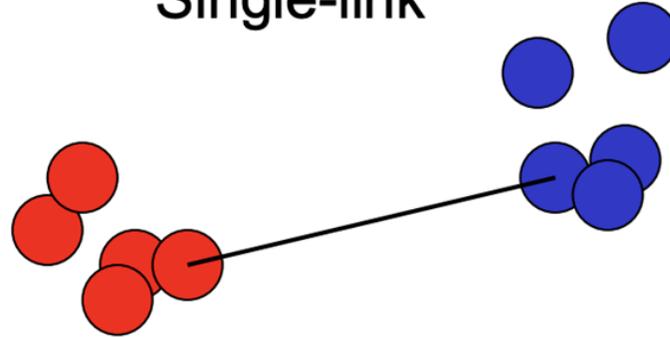
$$\Delta(\mathbf{C}_\alpha, \mathbf{C}_\beta) = \|\mathbf{m}_\alpha - \mathbf{m}_\beta\|$$

All linkage methods

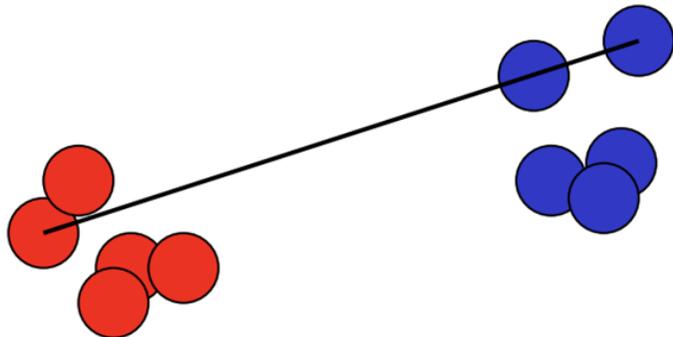
Distance between centroids



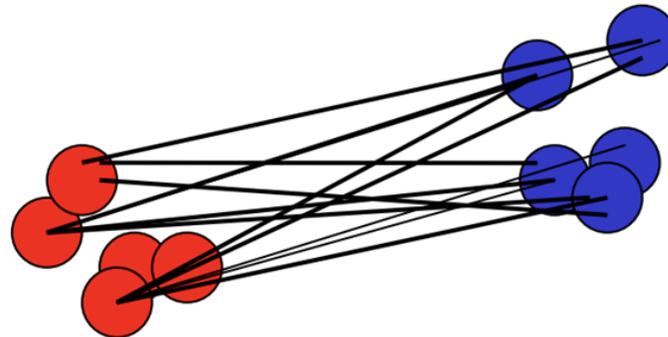
Single-link



Complete-link



Mean-link



Hierarchical clustering Technique:

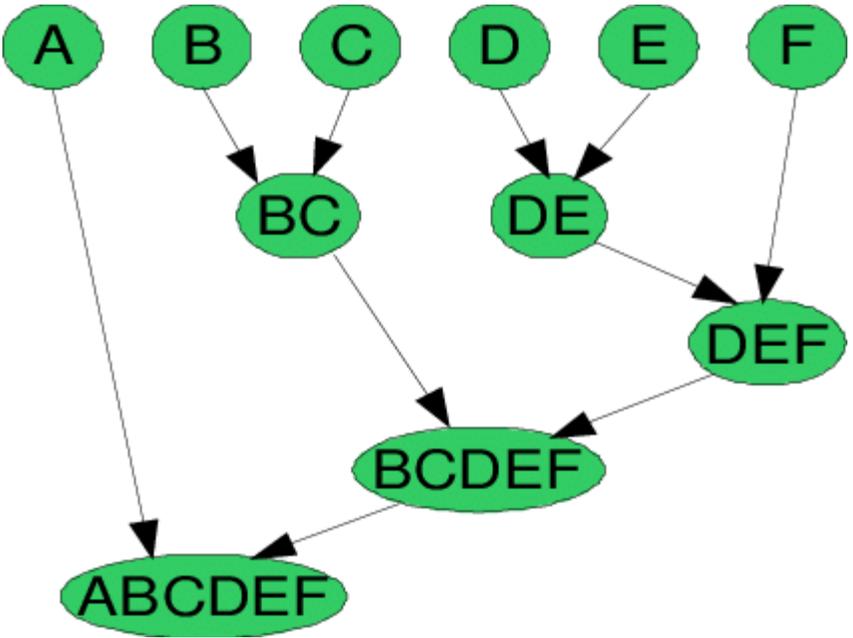
Hierarchical clustering is one of the popular and easy to understand clustering technique. This clustering technique is divided into two types:

1. Agglomerative
2. Divisive

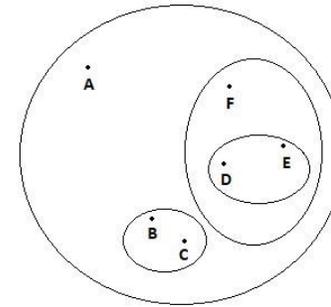
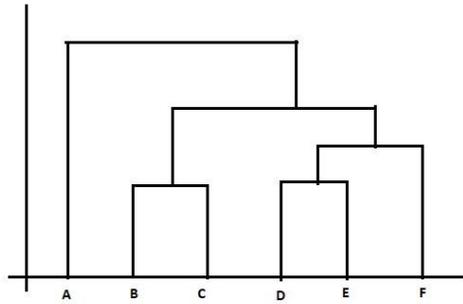
Agglomerative Hierarchical clustering Technique:

initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed.

Key operation is the computation of the proximity of two clusters
To understand better let's see a pictorial representation of the Agglomerative Hierarchical clustering Technique. Lets say we have six data points {A,B,C,D,E,F}.

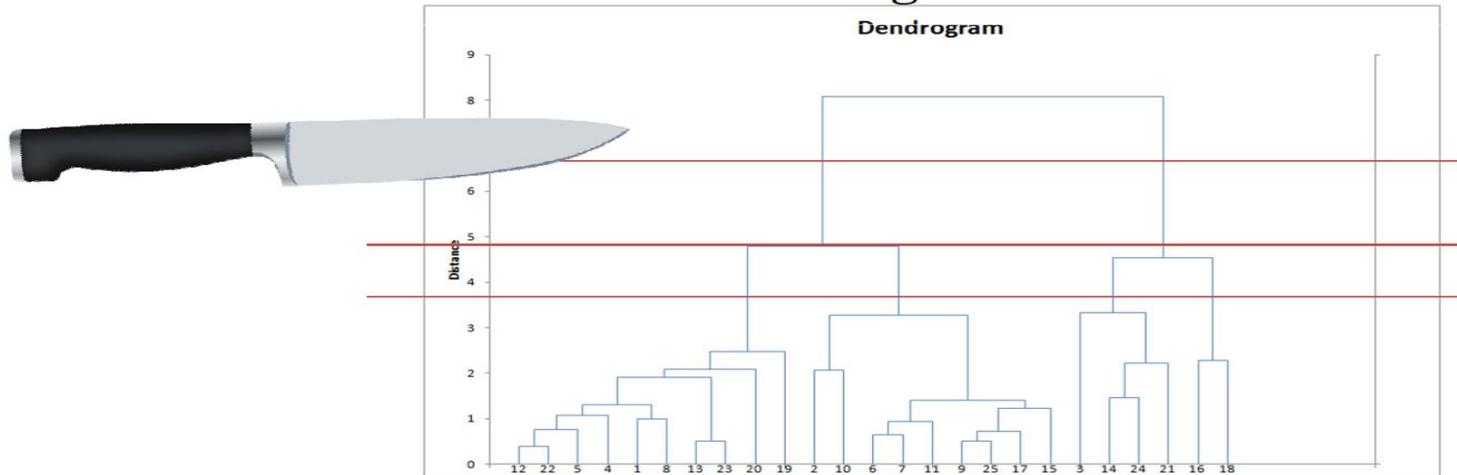


The Hierarchical clustering Technique can be visualized using a **Dendrogram**.
A **Dendrogram** is a tree-like diagram that records the sequences of merges or splits.



From Dendrograms to Clusters

- After dendrogram is obtained, **cut** it to create clusters. **How?**
- Examine *distance levels*
 - Cutpoint determines # clusters
 - Obtain statistics on resulting clusters



2. Divisive Hierarchical clustering Technique:

Since the Divisive Hierarchical clustering Technique is not much used in the real world, I'll give a brief of the Divisive Hierarchical clustering Technique.

In simple words, we can say that the Divisive Hierarchical clustering is exactly the opposite of the **Agglomerative Hierarchical clustering**. In Divisive Hierarchical clustering, we consider all the data points as a single cluster and in each iteration, we separate the data points from the cluster which are not similar. Each data point which is separated is considered as an individual cluster. In the end, we'll be left with n clusters.

As we're dividing the single clusters into n clusters, it is named as **Divisive Hierarchical clustering**. So, we've discussed the two types of the Hierarchical clustering Technique.

