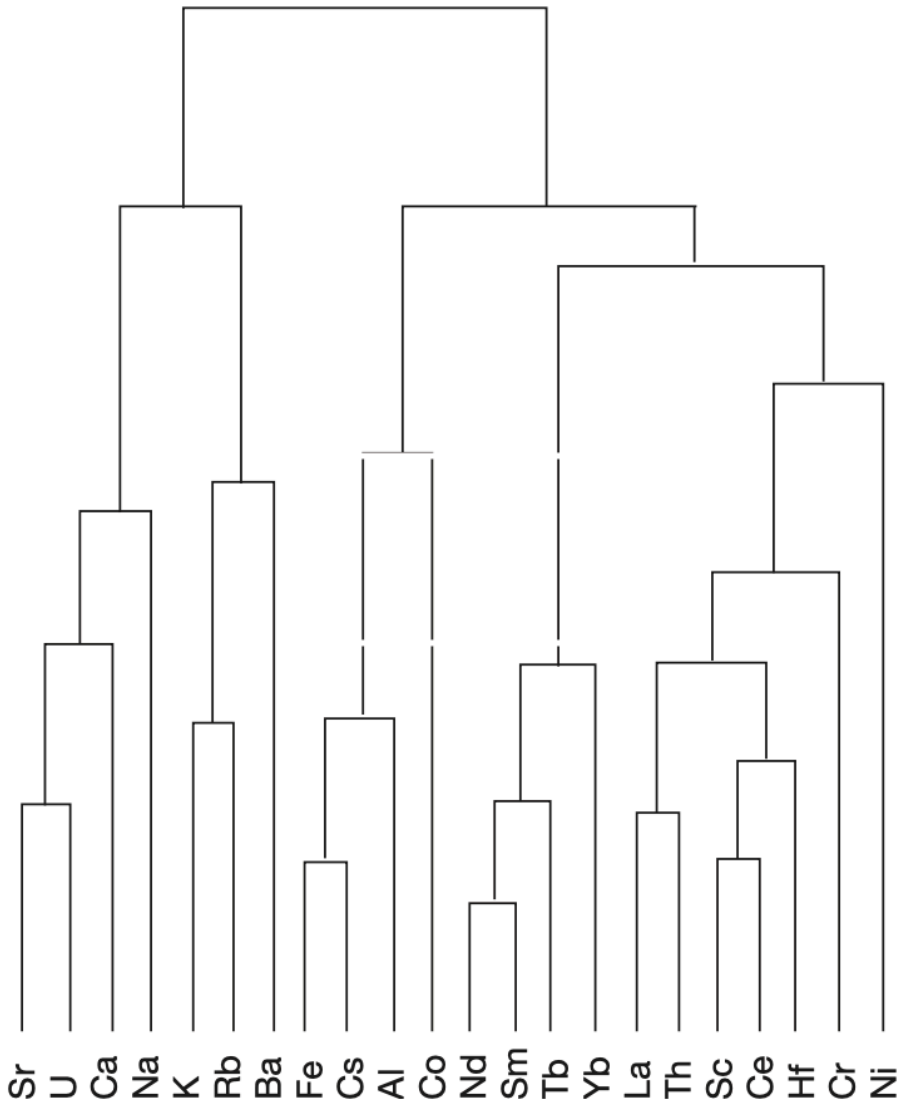
Dendrogram: Hierarchical Clustering



Affinity groupings of elements (Dupre et al., 1996)

Clustering
Clustering is used to help uncover structure in data, or to uncover structure in the samples represented by the data. Basically, cluster analysis is separating sample points into groups in data through a mathematical model.

Human's and Machine's Point of View
As humans, we depend on our ability to process visual imagery and then to group inputs into useful categories for specific purposes. Within a group, individuals resemble each other but are different in some respects from individuals in other groups. However, Machine, or algorithms, use mathematics to group samples into clusters. Clusters are an artefact of a mathematical method.

Ground Truth Labels and Clusters
We, as humans, sometimes, provide physical (ground truth) labels during the collection process. Consequently, data provided with ground truth labels by are called "labelled classes", as opposed to "clusters" by a clustering model /algorithm that detects structure in data. Understanding the difference between the human's and machine's points of view is a great help when you are trying to make a computer find clusters.

Distance between Points
Machines identify clusters of observations which may be present in data by quantifing how close individuals are to each other. A quantitative measure of closeness is "distance". Two points are close when their similarity is large, or their distance is small. Here we use Euclidean distance.

Distance between Clusters
There are several ways to measure distance between clusters:

1) **Single linkage**: smallest distance between an element in one cluster and an element in the other. $d(Ci,Cj) = \min\{d(x,y) \mid x \in Ci, y \in Cj\}$
2) **Complete linkage**: the opposite of single linkage, the largest distance between an element in one cluster and an element in the other. $d(Ci,Cj) = \max\{d(x,y) \mid x \in Ci, y \in Cj\}$
3) **Average linkage**: average distance between an element in one cluster and an element in the other. $d(Ci,Cj) = \sum x \in Ci \sum y \in Cj\ d(x,y) / (n_i \times n_j)$
4) Centroid linkage: distance between the centers of two clusters. $d(Ci,Cj) = d(u_i, u_j)$
5) **Ward linkage**: The distance between two clusters is defined as the increase in sum of the squared errors SSE when the two clusters are merged. Ward's distance = Weighted centroid distance. $d(Ci,Cj) = (n_i n_j/(n_i+n_j))(||u_i-u_j||^2)$

Hierarchial Clustering
Hierarchial clustering is an unsupervised learning clustering algorithm. Here's how the hierarchical clustering algorithm works:

1) Initially, each point is treated as a singleton cluster.
2) If two (or more) points are close, put them in the same cluster.
3) Repeat step 2
4) Eventually, all points (items) are in one cluster and the process ends.

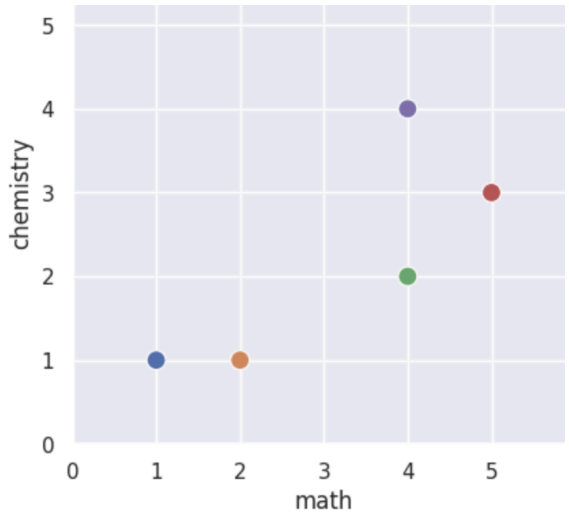
An Hands-on Example with a Toy Dataset

Let's look at a two-dimensional example; but the same applies to high-dimensional data, as you know. Here's a toy dataset with 5 samples (points) and 2 features (variables). We'll see how sample points are grouped step by step and how the dendrogram is generated during the clustering process.

	math	chemistry	labels
0	1	1	A
1	2	1	B
2	4	2	C
3	5	3	D
4	4	4	E

We have Five points:
A = (1,1)
B = (2,1)
C = (4,2)
D = (5,3)
E = (4,4)

The five points can be seen in a plot, as in the following



Step 1: Initial Clustering
{A}
{B}
{C}
{D}
{E}

Initially, we take each sample point as a cluster, and we have five clusters.

We compute the distance matrix, and have the distance matrix:

```
AB = sqrt((2-1)^2+(1-1)^2)=1
AC = sqrt((4-1)^2+(2-1)^2)=sqrt(10)=3.16
AD = sqrt((5-1)^2+(3-1)^2)=sqrt(20)=4.47
AE = sqrt((4-1)^2+(4-1)^2)=sqrt(18)=4.24
BC = sqrt((4-2)^2+(2-1)^2)=sqrt(5)= 2.24
BD = sqrt((5-2)^2+(3-1)^2)=sqrt(13)= 3.61
BE = sqrt((4-2)^2+(4-1)^2)=sqrt(13)= 3.61
CD = sqrt((5-4)^2+(3-2)^2)=sqrt(2)= 1.41
CE = sqrt((4-4)^2+(4-2)^2)=sqrt(4)= 2
DE = sqrt((4-5)^2+(4-3)^2)=sqrt(2)= 1.41
```

	A	B	C	D	E
A	0				
B	1	0			
C	3.16	2.24	0		
D	4.47	3.61	1.41	0	
E	4.24	3.61	2	1.41	0

Then, we find the closest clusters, and joined them to form a two-point cluster {AB}.

Step 2: We have four clusters:
{AB}
{C}
{D}
{E}

We compute distance matrix based on the four clusters and have the new distance matrix:

```
ABC = min{AC,BC} = min{3.16, 2.24} = 2.24
ABD = min{AD,BD} = min{4.47, 3.61} = 3.61
ABE = min{AE,BE} = min{4.24, 3.61} = 3.61
```

	AB	C	D	E
AB	0			
C	2.24	0		
D	3.61	1.41	0	
E	3.61	2	1.41	0

Closest clusters are merged as {CD}.

Step 3: Then we have three clusters:
{AB}
{CD}
{E}

We compute the distance matrix and get the new distance matrix:

```
ABCD = min{AC,AD,BC,BD} = min{3.16, 4.47, 2.24, 3.61} = 2.24
ABE = min{AE,BE} = min{4.24, 3.61} = 3.61
CDE = min{CE,DE} = min{2, 1.41} = 1.41
```

	AB	CD	E
AB	0		
CD	2.24	0	
E	3.61	1.41	0

Closest clusters are merged as {CDE}.

Step 4: Then we have two clusters:
{AB}
{CDE}

We compute the new distance matrix-

```
ABCDE = min{AC,AD,AE,BC,BD,BE} = min{3.61,4.47,4.24,2.24,3.61,3.61} = 2.24
```

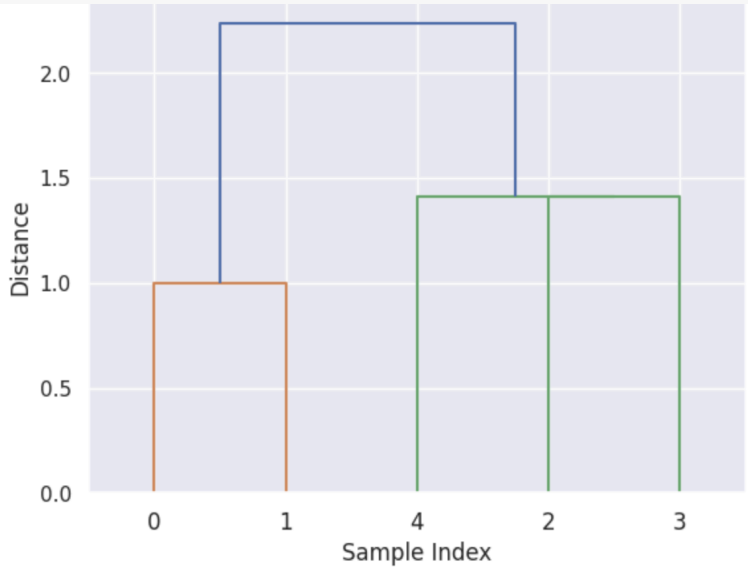
	AB	CDE
AB	0	
CDE	2.24	0

Step 5: Finally, we have one cluster containing all elements {ABCDE}.

Dendrogram

A dendrogram is a tree structure used to represent the process of hierarchical clustering.

```
Z = linkage(X, 'single')
dendrogram(Z)
```



If we use ward linkage,

```
Z = linkage(X, method='ward')
dendrogram(Z)
```

we get the following: