Hierarchical Clustering

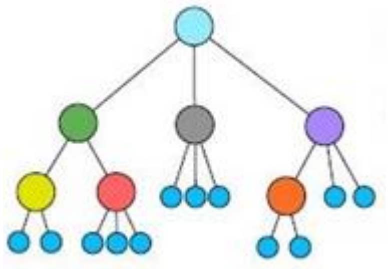
Dr. Uzair Ahmad

Clustering

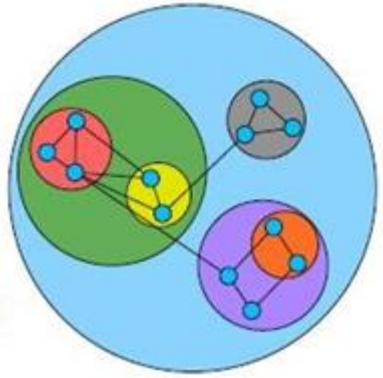
- Clustering
  - – Put similar things together

Hierarchical Clustering

- Agglomerative/Merge [Bottom-Up] Approach
  - Every point is a cluster
  - Pair-wise distances
  - Dendrogram
  - At least quadratic in data points



- Divisive [Top-down] Approach
  - Recursively split a cluster
  - Until individual datapoints are reached
  - Linear in data points



Hierarchical Clustering

1. Compute Distance between all pairs of clusters
    - NxN Similarity Matrix **C**
2. Merge nearest points into one cluster
    - N-1 Steps
3. Update row-columns of **C**

Hierarchical Clustering

- Requirements
  - Closeness/Distance Measure
  - Merging Measure
- Output
  - A Tree [Dendrogram]

Example

Data: Monthly Average Temperature (US)

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	YEARS
Average Temp (F)	31.9	32.3	35	52	60.8	68.7	73.3	72.1	65.2	54.8	40	38	59

		JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
		31.9	32.3	35	52	60.8	68.7	73.3	72.1	65.2	54.8	40	38
JAN	31.9	0											
FEB	32.3	0.4	0										
MAR	35	3.1	2.7	0									
APR	52	20.1	19.7	17	0								
MAY	60.8	28.9	28.5	25.8	8.8	0							
JUN	68.7	36.8	36.4	33.7	16.7	7.9	0						
JUL	73.3	41.4	41	38.3	21.3	12.5	4.6	0					
AUG	72.1	40.2	39.8	37.1	20.1	11.3	3.4	1.2	0				
SEP	65.2	33.3	32.9	30.2	13.2	4.4	3.5	8.1	6.9	0			
OCT	54.8	22.9	22.5	19.8	2.8	6	13.9	18.5	17.3	10.4	0		
NOV	40	8.1	7.7	5	12	20.8	28.7	33.3	32.1	25.2	14.8	0	
DEC	38	6.1	5.7	3	14	22.8	30.7	35.3	34.1	27.2	16.8	2	0

Example

Data: Monthly Average Temperature (US)

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	YEARS
Average Temp (F)	31.9	32.3	35	52	60.8	68.7	73.3	72.1	65.2	54.8	40	38	59

		JANFEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
		32.1	35	52	60.8	68.7	73.3	72.1	65.2	54.8	40	38
JANFEB	32.1	0										
MAR	35	2.9	0									
APR	52	19.9	17	0								
MAY	60.8	28.7	25.8	8.8	0							
JUN	68.7	36.6	33.7	16.7	7.9	0						
JUL	73.3	41.2	38.3	21.3	12.5	4.6	0					
AUG	72.1	40	37.1	20.1	11.3	3.4	1.2	0				
SEP	65.2	33.1	30.2	13.2	4.4	3.5	8.1	6.9	0			
OCT	54.8	22.7	19.8	2.8	6	13.9	18.5	17.3	10.4	0		
NOV	40	7.9	5	12	20.8	28.7	33.3	32.1	25.2	14.8	0	
DEC	38	5.9	3	14	22.8	30.7	35.3	34.1	27.2	16.8	2	0

Example

Data: Monthly Average Temperature (US)

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	YEARS
Average Temp (F)	31.9	32.3	35	52	60.8	68.7	73.3	72.1	65.2	54.8	40	38	59

		JANFEB	MAR	APR	MAY	JUN	JULAUG	SEP	OCT	NOV	DEC
		32.1	35	52	60.8	68.7	72.7	65.2	54.8	40	38
JANFEB	32.1	0									
MAR	35	2.9	0								
APR	52	19.9	17	0							
MAY	60.8	28.7	25.8	8.8	0						
JUN	68.7	36.6	33.7	16.7	7.9	0					
JULAUG	72.7	40.6	37.7	22.7	11.9	4	0				
SEP	65.2	33.1	30.2	13.2	4.4	3.5	7.5	0			
OCT	54.8	22.7	19.8	2.8	6	13.9	17.9	10.4	0		
NOV	40	7.9	5	12	20.8	28.7	32.7	25.2	14.8	0	
DEC	38	5.9	3	14	22.8	30.7	34.7	27.2	16.8	2	0

Example

Data: Monthly Average Temperature (US)

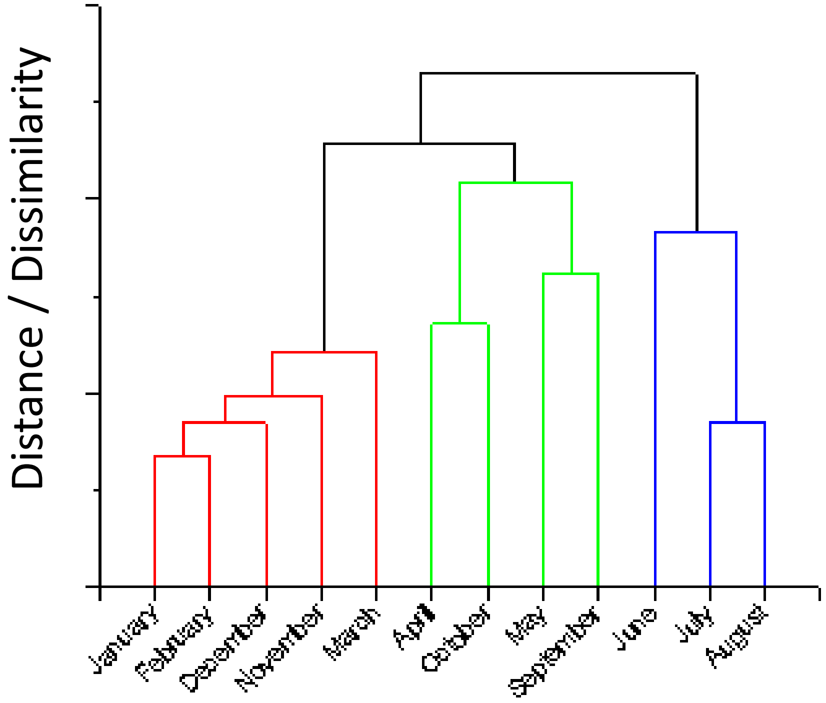
	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	YEARS
Average Temp (F)	31.9	32.3	35	52	60.8	68.7	73.3	72.1	65.2	54.8	40	38	59

		JANFEB	MAR	APR	MAY	JUN	JULAUG	SEP	OCT	NOV	DEC
		32.1	35	52	60.8	68.7	72.7	65.2	54.8	40	38
JANFEB	32.1	0									
MAR	35	2.9	0								
APR	52	19.9	17	0							
MAY	60.8	28.7	25.8	8.8	0						
JUN	68.7	36.6	33.7	16.7	7.9	0					
JULAUG	72.7	40.6	37.7	22.7	11.9	4	0				
SEP	65.2	33.1	30.2	13.2	4.4	3.5	7.5	0			
OCT	54.8	22.7	19.8	2.8	6	13.9	17.9	10.4	0		
NOV	40	7.9	5	12	20.8	28.7	32.7	25.2	14.8	0	
DEC	38	5.9	3	14	22.8	30.7	34.7	27.2	16.8	2	0

Example

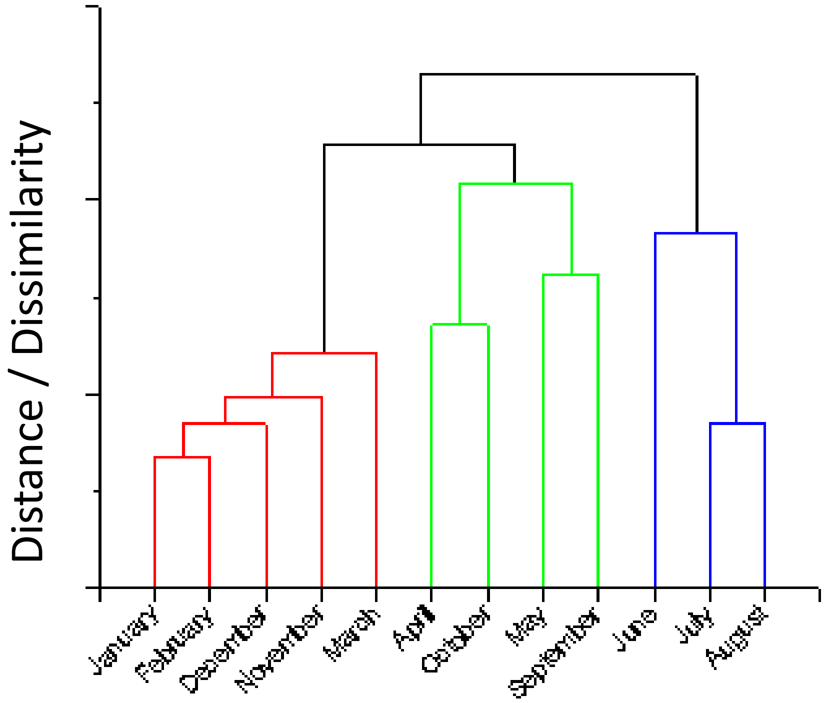
Data: Monthly Average Temperature (US)

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
Average Temp (F)	31.9	32.3	42.4	52	60.8	68.7	73.3	72.1	65.2	54.8	36.5	35.5



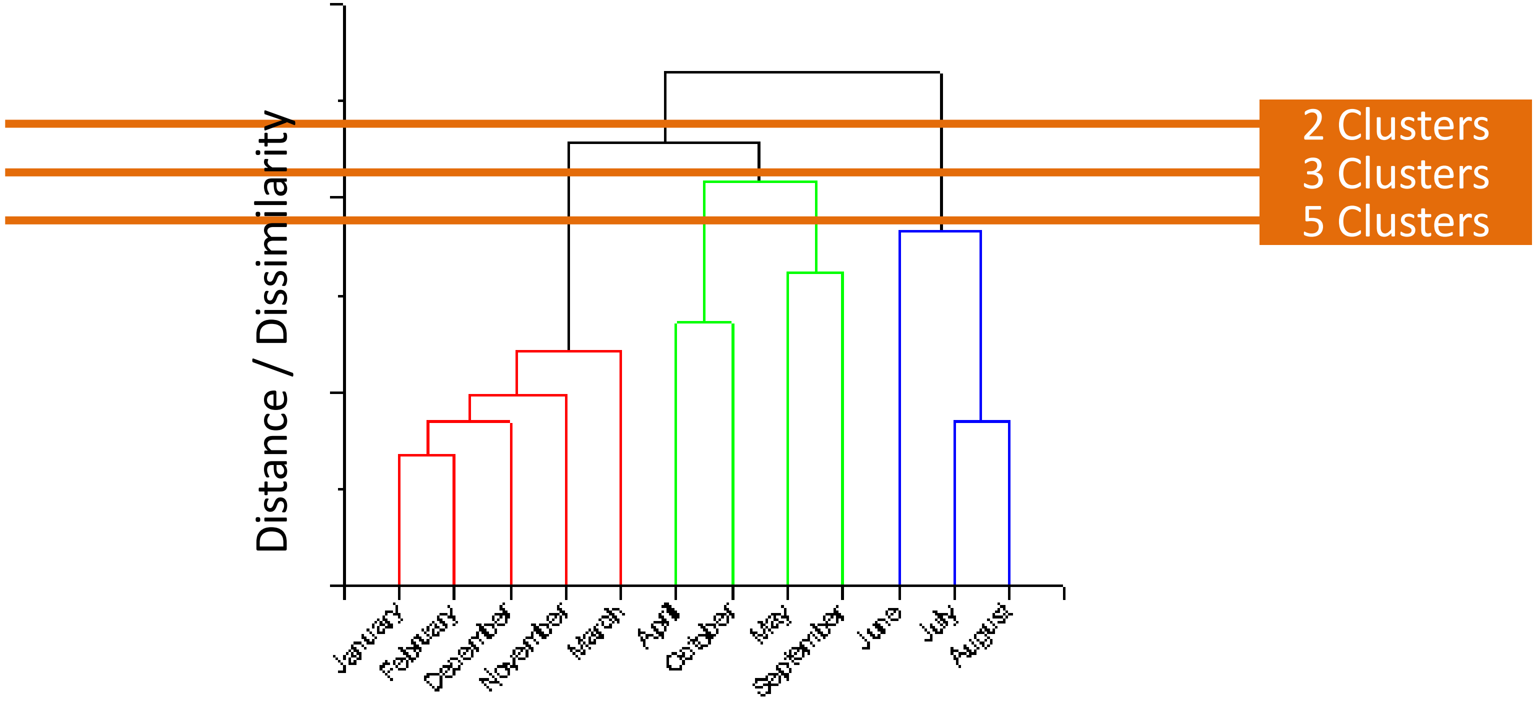
Hierarchical Clustering

- A Dendrogram explains
  1. How dissimilar two points are from each other
  2. When a cluster is formed



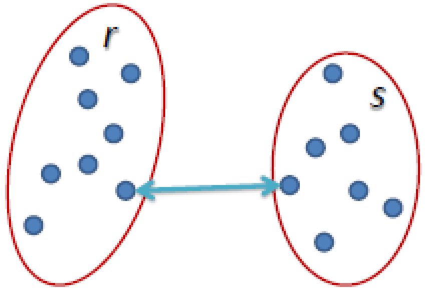
Hierarchical Clustering

- A Dendrogram does not explain
  - How many clusters are there in the data
  - Cut the Tree to see the clusters



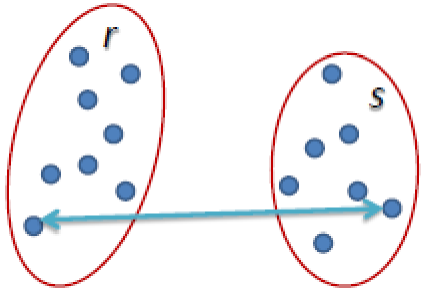
Merge criterion

- Nearest neighbor (Single Linkage)
  - $D_{\min(C_i,C_j)} = \min_{x \in C_i, y \in C_j} \|x - y\|^2.$



$L(r,s) = \min(D(x_{ri}, x_{sj}))$

- Farthest neighbor (Complete Linkage)
  - $D_{\max(C_i,C_j)} = \max_{x \in C_i, y \in C_j} \|x - y\|^2.$



- Centroid
  - $D_{\text{means}(C_i,C_j)} = \|\mu_i - \mu_j\|^2$

Summary

- Intuitive but subjective
- No need to estimate K
- Complexity > $O(N^2)$