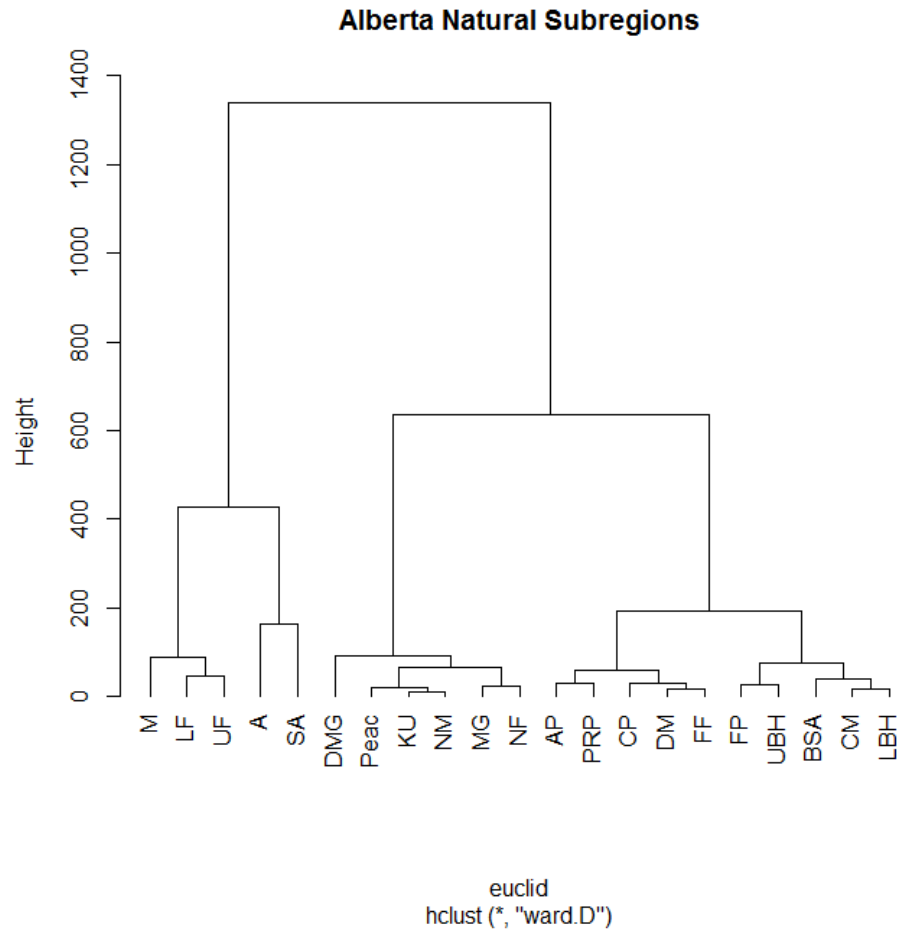


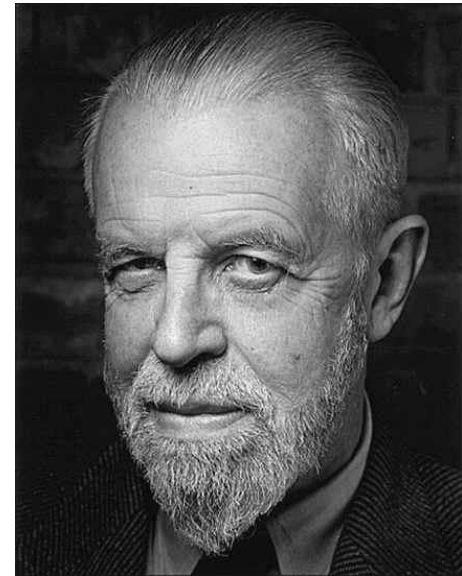
## Cluster Analysis



**Objective:** Group data points into classes of similar points based on a series of variables

Useful to find the true groups that are assumed to really exist, BUT if the analysis generates unexpected groupings it could inform new relationships you might want to investigate

Also useful for data reduction by finding which data points are similar and allow for subsampling of the original dataset without losing information



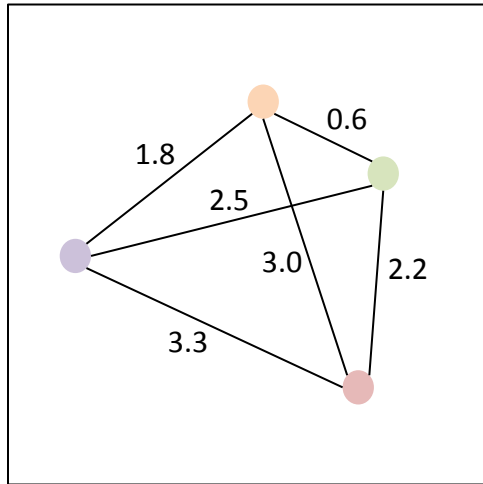
Alfred Louis Kroeber (1876-1961)

# The math behind cluster analysis

Once we calculate a distance matrix between points we use that information to build a *tree*

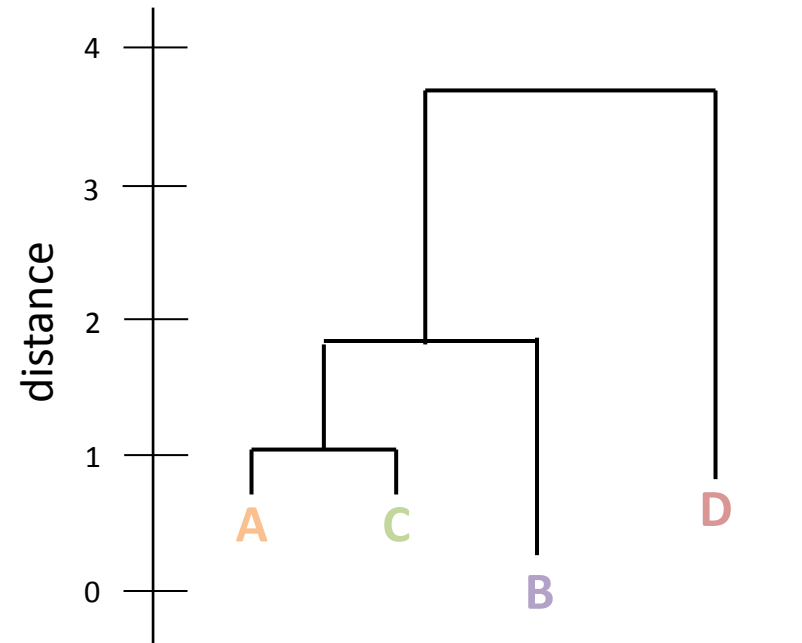
**Ordination** – visualizes the information in the distance calculations

	A	B	C	D	...
A	0	1.8	0.6	3.0	
B	1.8	0	2.5	3.3	
C	0.6	2.5	0	2.2	
D	3.0	3.3	2.2	0	
...					



If distances are not equal between points we can draw a “hanging tree” to illustrate distances

The result of a cluster analysis is a tree or dendrogram



# Building trees & creating groups

1. **Nearest Neighbour Method** – create groups by starting with the smallest distances and build branches

In effect we keep asking data matrix “Which plot is my nearest neighbour?” to add branches

2. **Centroid Method** – creates a group based on smallest distance to group centroid rather than group member

First creates a group based on small distance then uses the centroid of that group to find which additional points belong in the same group

3. **Wards Method** – creates groups such that variance is minimized within clusters

Looks for spherical clusters

The process starts with all points in individual clusters (bottom up) and the process repeatedly merges a pair of clusters such that when merged there is a minimum increase in total within-cluster variance This process continues until a single group including all points (the top of the tree) is defined

# Building trees & creating groups

One of the problems with Cluster Analysis is that different methods may produce different results – generally no accepted *best* method

**Good News:** If your data really has clear groups all methods will find them and give you similar results

Therefore it is best to try multiple algorithms and see what groups logically make sense

If you have a dummy dataset with pre-determined groups you can use it to see which algorithm best recreates what you expect

# Cluster analysis in R

Distance matrix of your data rows based on your predictor variables  
You need to calculate this before running the cluster analysis

## CA in R:

```
hclust(distMatrix, method) (stats package)
```

We create distance matrices in Lab 5

What type of algorithm should be used to cluster points and define groups

"ward.D" = Ward's minimum variance method

"ward.D2" = Ward's minimum variance method – however dissimilarities are squared before clustering

"single" = Nearest neighbours method

"complete" = distance between two clusters is defined as the maximum distance between an observation in one cluster and an observation in the other cluster

"average" = distance between two clusters is defined as the mean distance between an observation in one cluster and an observation in the other cluster

"mcquitty" = when two clusters are be joined, the distance of the new cluster to any other cluster is calculated as the average of the distances of the soon to be joined clusters to that other cluster

"median" = uses group median

"centroid" = uses group centroid

# Cluster analysis in R

```
E:\LauraG\3. Teaching\7. RENR 690\1.Winter 2016\5. Distance - Cluster & NMDS (2-03)\R & SAS\Lab5.r - R Editor

##### Cluster analysis using various distances #####
tree=hclust(euclid, method="ward.D")
plot(tree, hang=-1, main="Alberta Natural Subregions")
```

From the output plot we can compare the groups cluster analysis has generated

Each collection of *branches* could be considered a group

It is up to you to decide how far down the tree you want to specify your groups

**We want small distances and groups that logically make sense**

