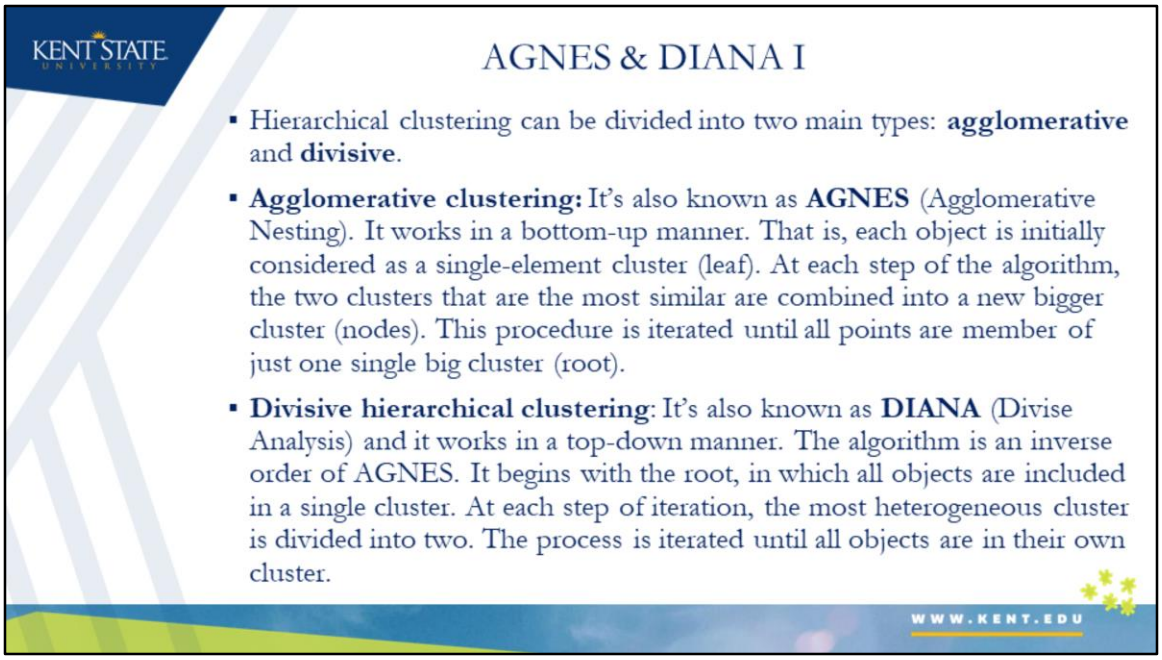
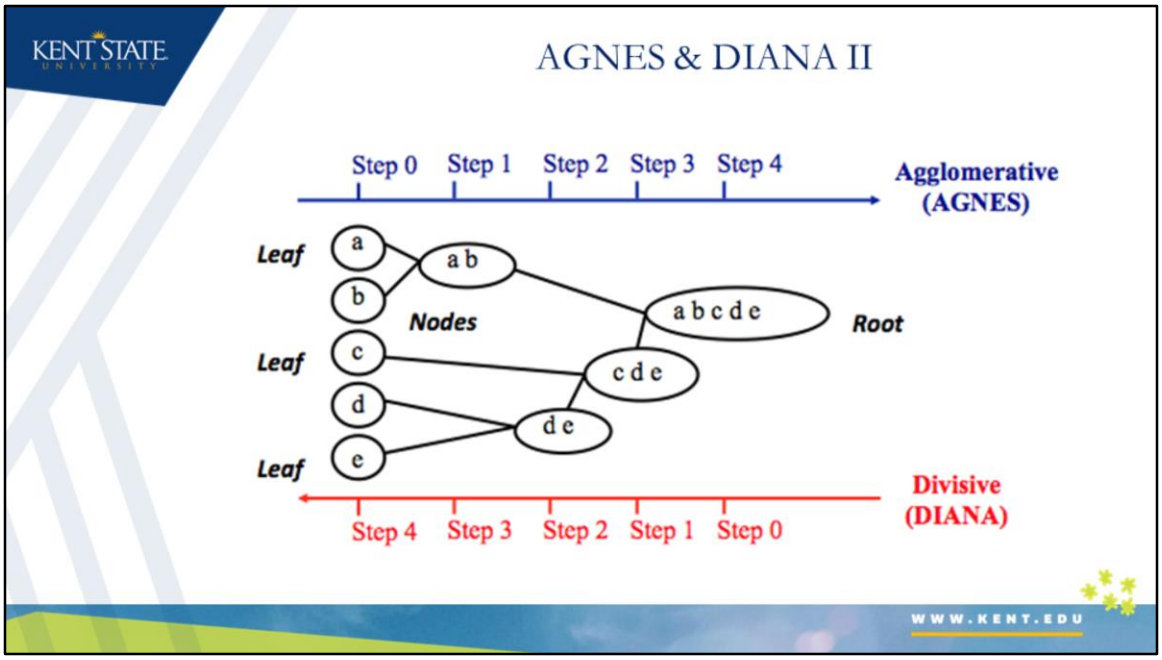
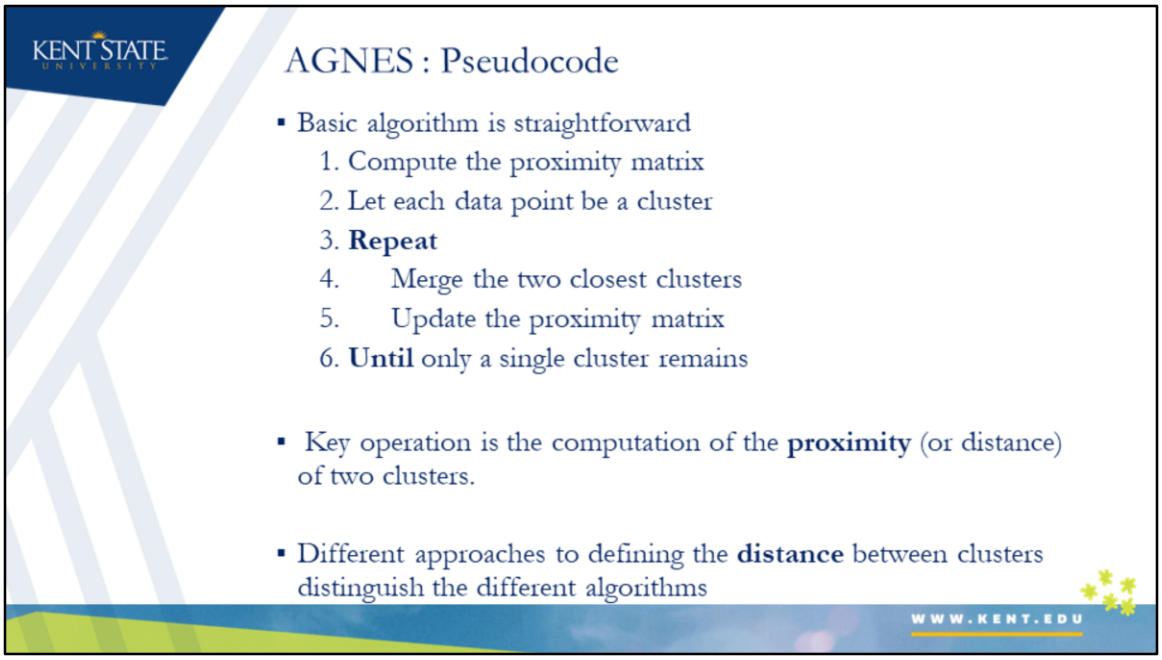
In this module, we discuss Agnes and Diana. The former provides a way to do Agglomerative Clustering, while the latter, Divisive Clustering.



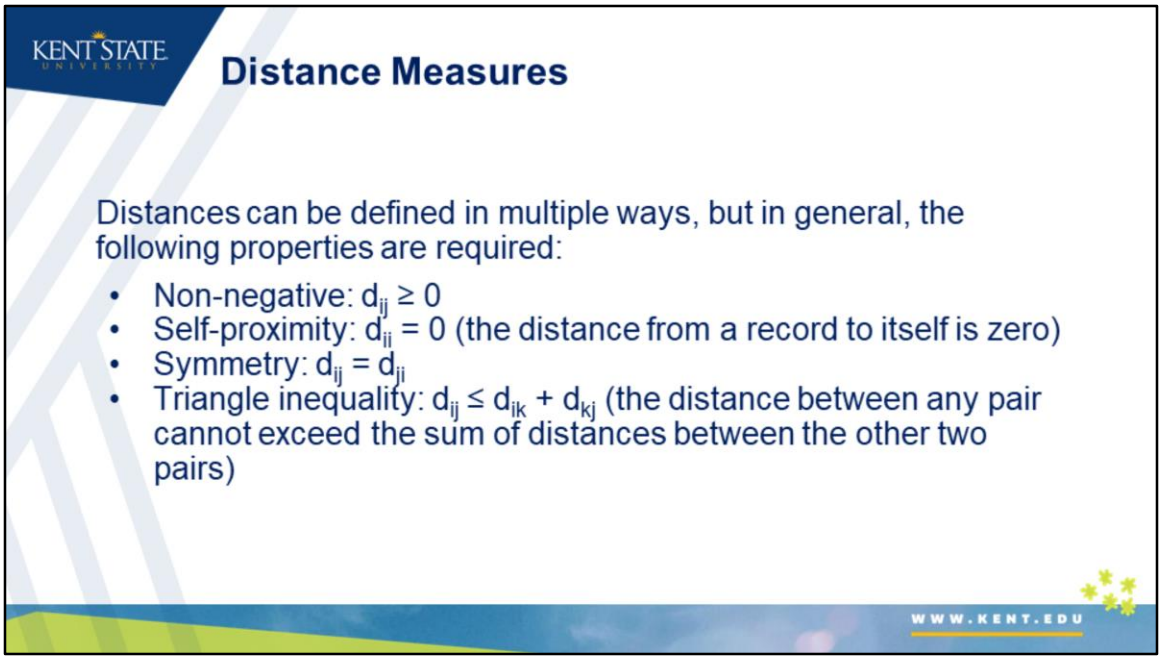
As we discussed earlier, Hierarchical clustering can either be Agglomerative, i.e., a bottoms up approach, or a top-down approach, called Divisive. Here we will look at both approaches.



This diagram depicts both approaches. While both approaches can lead to the same solution, the time complexity may vary depending on how the splitting (divisive), or merging (agglomerative) approaches are iimplemented.



The process is simple. Merge clusters based on their proximity to other clusters. Initially, all data points are in a cluster of size 1. Proximity is a measure of distance, and that can be defined in any fashion, including using well-known methods for distance.



All distances follow some basic properties. These include, non-negativity, self proximity, symmetry, and the triangle inequality.

KENT STATE UNIVERSITY

Distance Measures - Cont.

- **Euclidean Distance** $d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$
  - Scale dependent
  - Ignores relationship between variables
  - Sensitive to outliers
- **Manhattan Distance** $d_{ij} = \sum_{m=1}^{p} |x_{im} - x_{jm}|$
  - Useful in presence of outliers
- **Maximum Coordinate Distance**
  - Looks only at the measurement on which records *i* and *j* deviate the most

$$d_{ij} = \max_{m = 1, 2, ..., p} |x_{im} - x_{jm}|$$

WWW.KENT.EDU

While the Euclidean distance is the most commonly used one, other distance measures do exist. The City Block, or Manhattan Distance is also widely used.

KENT STATE UNIVERSITY

Similarity Measures

- **Correlation Based Similarity**
  - A popular similarity measure is the square of the Pearson correlation coefficient, r^2. Such measures can be converted to distance measures as
    $$d_{ij} = 1 - r_{ij}^2$$
- **Mahalanobis Distance (Statistical Distance)**
  - This metric has an advantage over the other metrics mentioned in that it takes into account the correlation between measurements. With this metric, measurements that are highly correlated with other measurements do not contribute as much as those that are uncorrelated or mildly correlated.
  - S is the covariance matrix.

$$d_{ij} = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

WWW.KENT.EDU

In the previous measures of distance, no consideration is given when values are correlated. It is possible to define distances that specifically account for this correlation. The Mahalanobis Distance is commonly used in Statistical Methods, and specifically accounts for the correlation between observations. These measures can be used as part of HC.

Distance Measures for Categorical Data

In the case of measurements with binary values, it is more intuitively appealing to use similarity measures than distance measures.

- Matching coefficient (a + d) / n

		Record j: 0	Record j: 1	
Record i	0	a	b	$a+b$
	1	c	d	$c+d$
		$a+c$	$b+d$	n

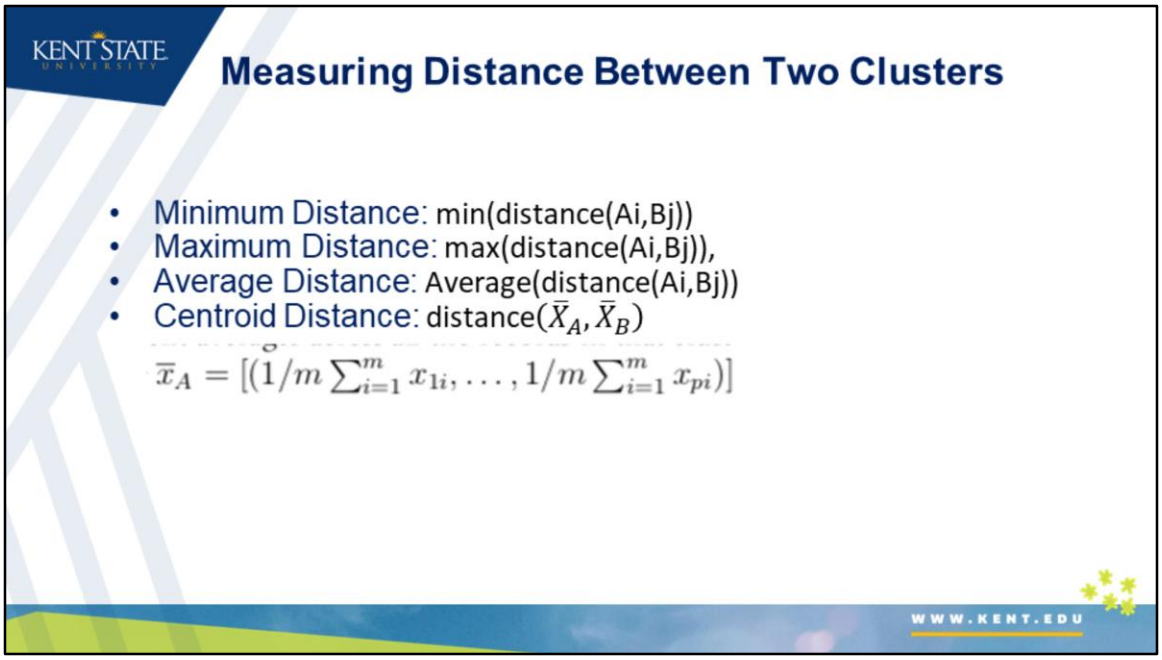
- Jacquard's coefficient d / (b + c + d)
  - This coefficient ignores zero matches. This is desirable when we do not want to consider two people to be similar simply because a large number of characteristics are absent in both.
- Mixed Data?

WWW.KENT.EDU

So far, the distances are all based on numerical measures. What about for categorical variables. In this case, we typically want to calculate a similarity index. The Matching coefficient and Jacquard's coefficients are two approaches for calculating the similarity index.

When the measurements are mixed (some continuous and some binary), a similarity coefficient suggested by Gower is very useful. Gower's similarity measure is a weighted average of the distances computed for each variable, after scaling each variable to a [0,1] scale.

.



We define a cluster as a set of one or more records. How do we measure distance between clusters? The idea is to extend measures of distance between records into distances between clusters. Consider cluster A, which includes the m records A1,A2,...,Am and cluster B, which includes n records B1,B2,...,Bn. The most widely used measures of distance between clusters are:

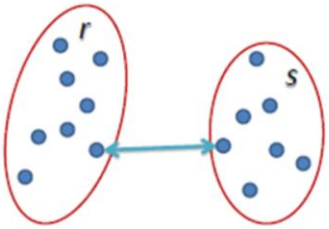
Minimum Distance: The distance between the pair of records Ai and Bj that are closest: min(distance(Ai,Bj))
Maximum Distance: The distance between the pair of records Ai and Bj that are farthest: max(distance(Ai,Bj))
Average Distance: The average distance of all possible distances between records in one cluster and records in the other cluster: Average(distance(Ai,Bj))
Centroid Distance: The distance between the two cluster centroids. A cluster centroid is the vector of measurement averages across all the records in that cluster.

Each of these distances (minimum, maximum, average, and centroid distance) can be implemented in the hierarchical scheme as described as follows.

KENT STATE UNIVERSITY

Single (Min) Linkage

- In single (min) linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.
- For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



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

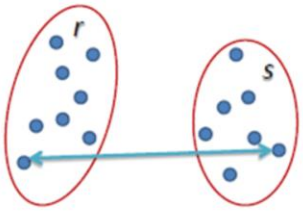
WWW.KENT.EDU

In single linkage clustering, the distance measure that we use is the minimum distance (the distance between the nearest pair of records in the two clusters, one record in each cluster). This method has a tendency to cluster together at an early stage records that are distant from each other because of a chain of intermediate records in the same cluster. Such clusters have elongated sausage-like shapes when visualized as objects in space.

KENT STATE UNIVERSITY

Complete (Max) Linkage

- In complete (max) linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster.



- For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.

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

WWW.KENT.EDU

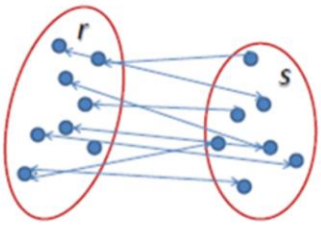
In complete linkage clustering, the distance between two clusters is the maximum distance (between the farthest pair of records).

This method tends to produce clusters at the early stages with records that are within a narrow range of distances from each other. If we visualize them as objects in space, the records in such clusters would have roughly spherical shapes.

KENT STATE UNIVERSITY

Average Linkage

- In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.



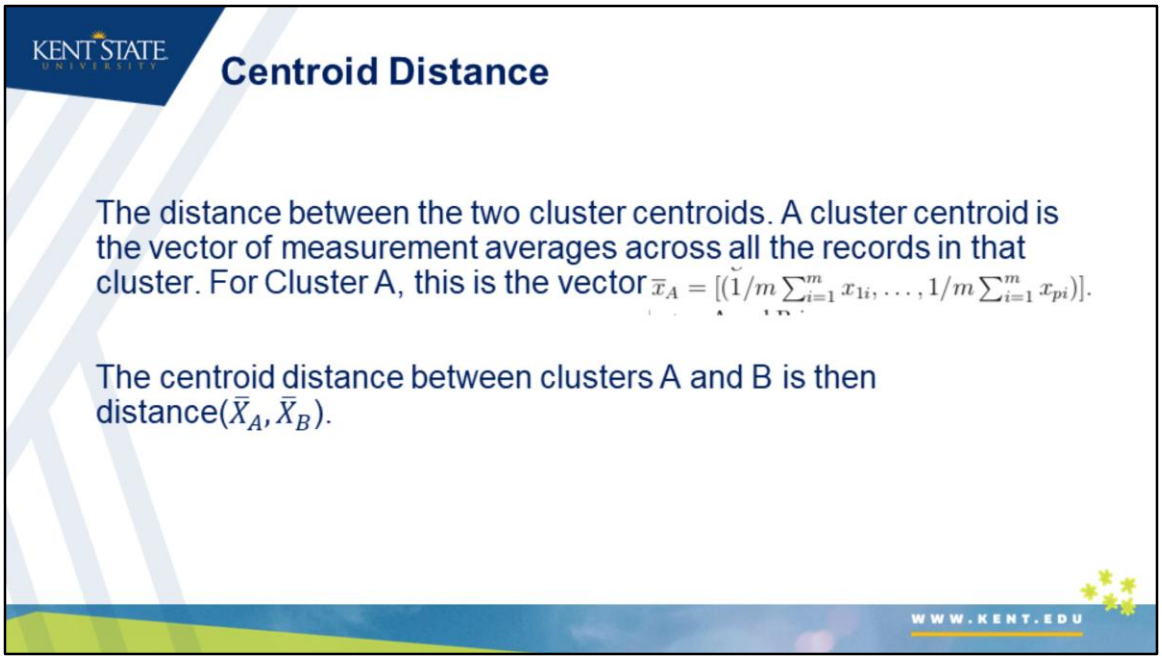
- For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

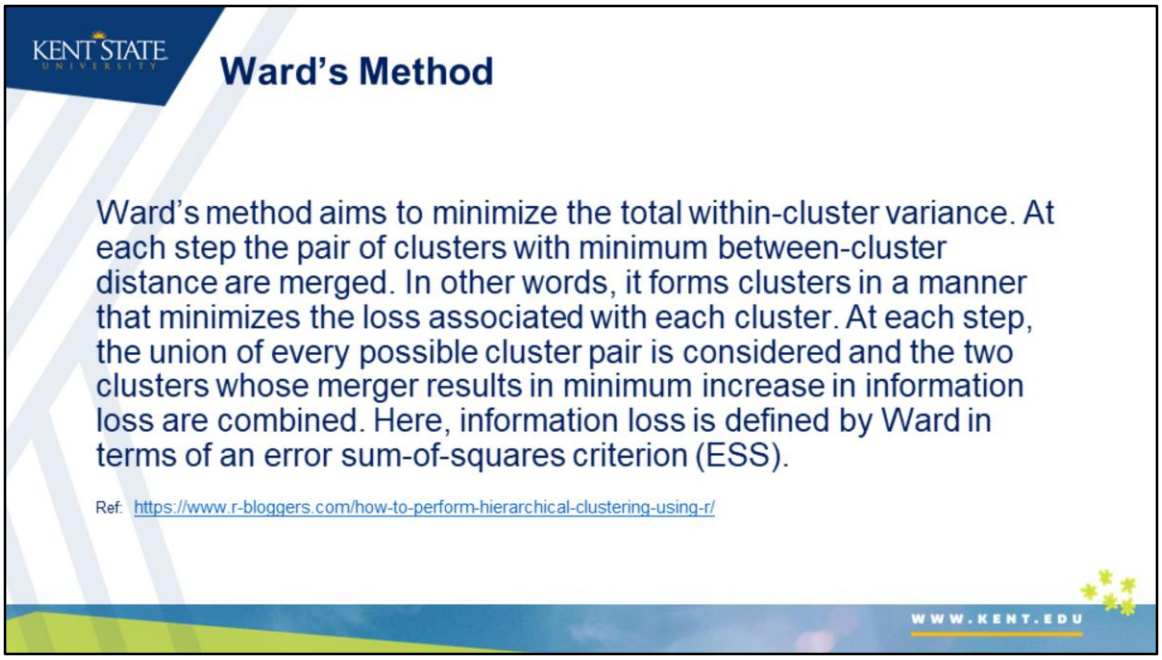
WWW.KENT.EDU

Average linkage clustering is based on the average distance between clusters (between all possible pairs of records).

Note that unlike average linkage, the results of the single and complete link- age methods depend only on the ordering of the inter-record distances. Linear transformations of the distances (and other transformations that do not change the ordering) do not affect the results.



Centroid linkage clustering is based on centroid distance, where clusters are represented by their mean values for each variable, which forms a vector of means. The distance between two clusters is the distance between these two vectors. In average linkage, each pairwise distance is calculated, and the average of all such distances is calculated. In contrast, in centroid distance clustering, just one distance is calculated: the distance between group means. This method is also called Unweighted Pair-Group Method using Centroids (UPGMC).



Ward's method is also agglomerative, in that it joins records and clusters together progressively to produce larger and larger clusters, but operates slightly differently from the general approach described previously. Ward's method considers the "loss of information" that occurs when records are clustered together. When each cluster has one record, there is no loss of information and all individual values remain available. When records are joined together and represented in clusters, information about an individual record is replaced by the information for the cluster to which it belongs. To measure loss of information, Ward's method employs a measure "error sum of squares" (ESS) that measures the difference between individual records and a group mean.

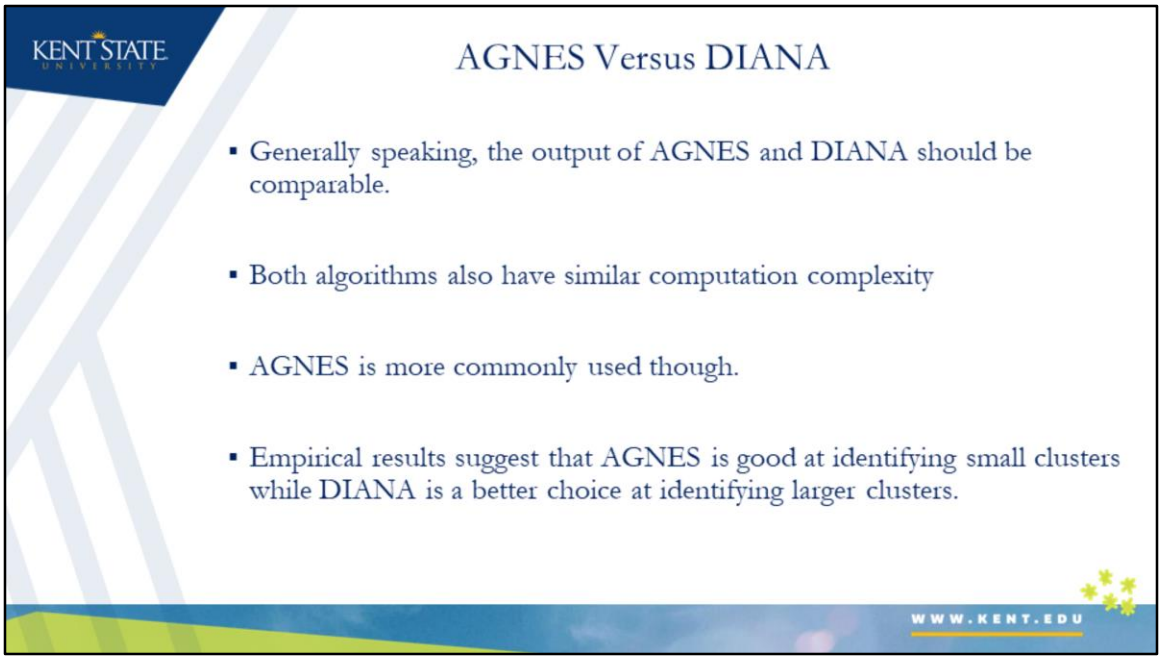
For example, consider the values (2, 6,5,6,2,2,2,2,0,0,0) with a mean of 2.5. Their ESS is equal to

$(2-2.5)2 +(6-2.5)2 +(5-2.5)2 +...+(0-2.5)2 = 50.5.$

The loss of information associated with grouping the values into a single group is therefore 50.5. Now group the records into four groups: (0, 0, 0), (2, 2, 2, 2), (5), (6, 6). The loss of information is the sum of the ESS's for each group, which is 0 (each record in each group is equal to the mean for that group, so

the ESS for each group is 0). Thus, clustering the 10 records into 4 clusters results in no loss of information, and this would be the first step in Ward's method.

Ward's method tends to result in convex clusters that are of roughly equal size, which can be an important consideration in some applications (e.g., in establishing meaningful customer segments).



This summarizes the choice of using Agnes or Diana. Both approaches produce similar results, though Agnes is usually more commonly used.