Proc du Jour: PROC TREE

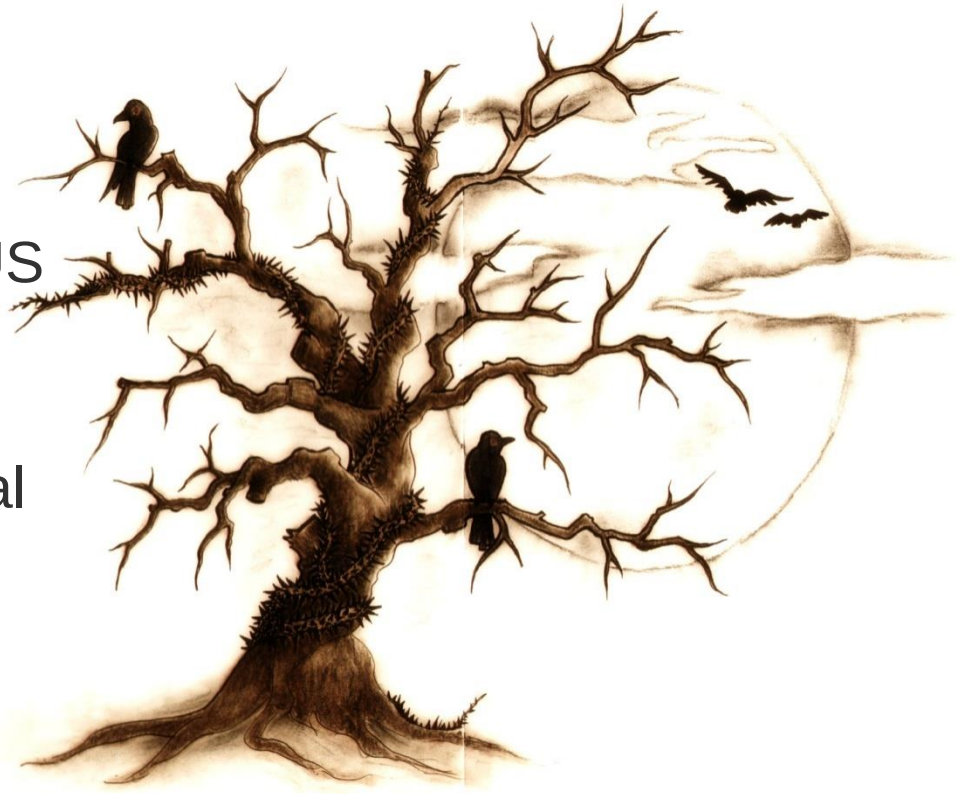
GHSUG: October 30, 2015

Lesley Harschnitz
lesley.harschnitz@arcelormittal.com

PROC TREE

- This procedure creates a DENDROGRAM from the output of either PROC CLUSTER or PROC VARCLUS

- PROC CLUSTER and PROC VARCLUS perform a statistical analysis of data to determine clusters



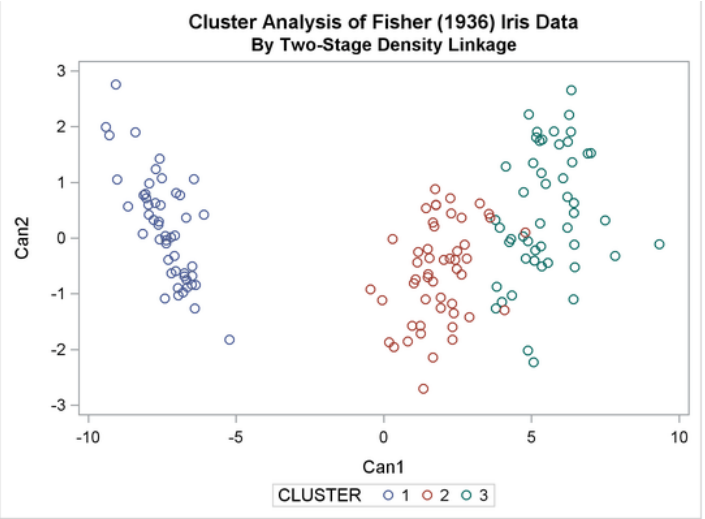
SAS/Stat® = Statistics ☺!

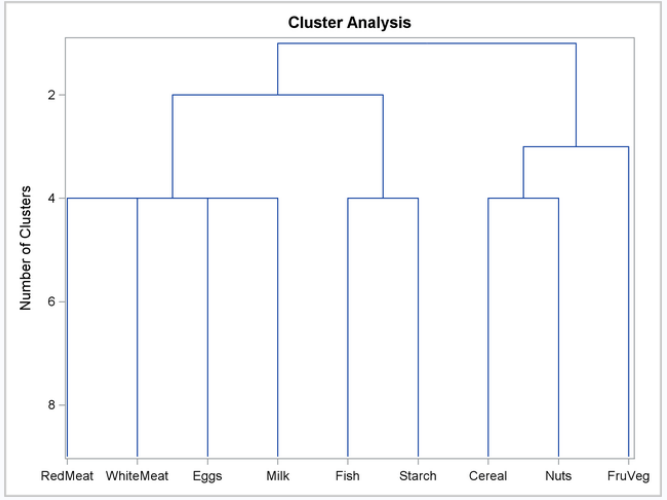
Note: All examples and charts in this presentation are taken from the on-line **SAS/STAT® User's Guide**

Clusters (Statistical)

- Based on the idea that members of one **cluster** are more similar to each other than to members of another cluster.

- **Hierarchical clustering** starts with the closest pairs and gradually adds in the next closest until all objects are connected. Clusters are identified by the maximum distance needed to connect parts of the cluster. Hierarchical clustering is what is used to create a DENDROGRAM or TREE diagram.



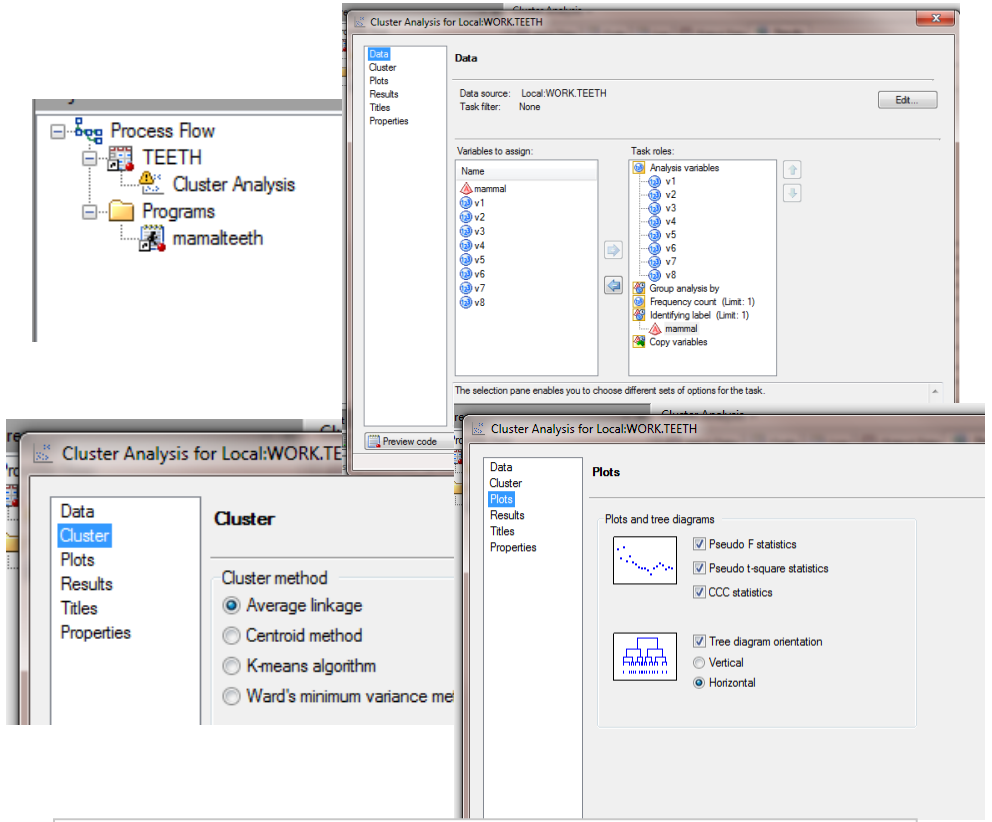


As is usual with statistics, there are a large number of mathematical ways to cluster, and it is important to understand and remember the assumptions involved.

The Code ---------- Enterprise Guide

```
ods graphics on;
proc cluster
   method=average std
   pseudo noeigen
   outtree=tree;
  id mammal;
  var v1-v8;
run;

proc tree horizontal;
run;
```

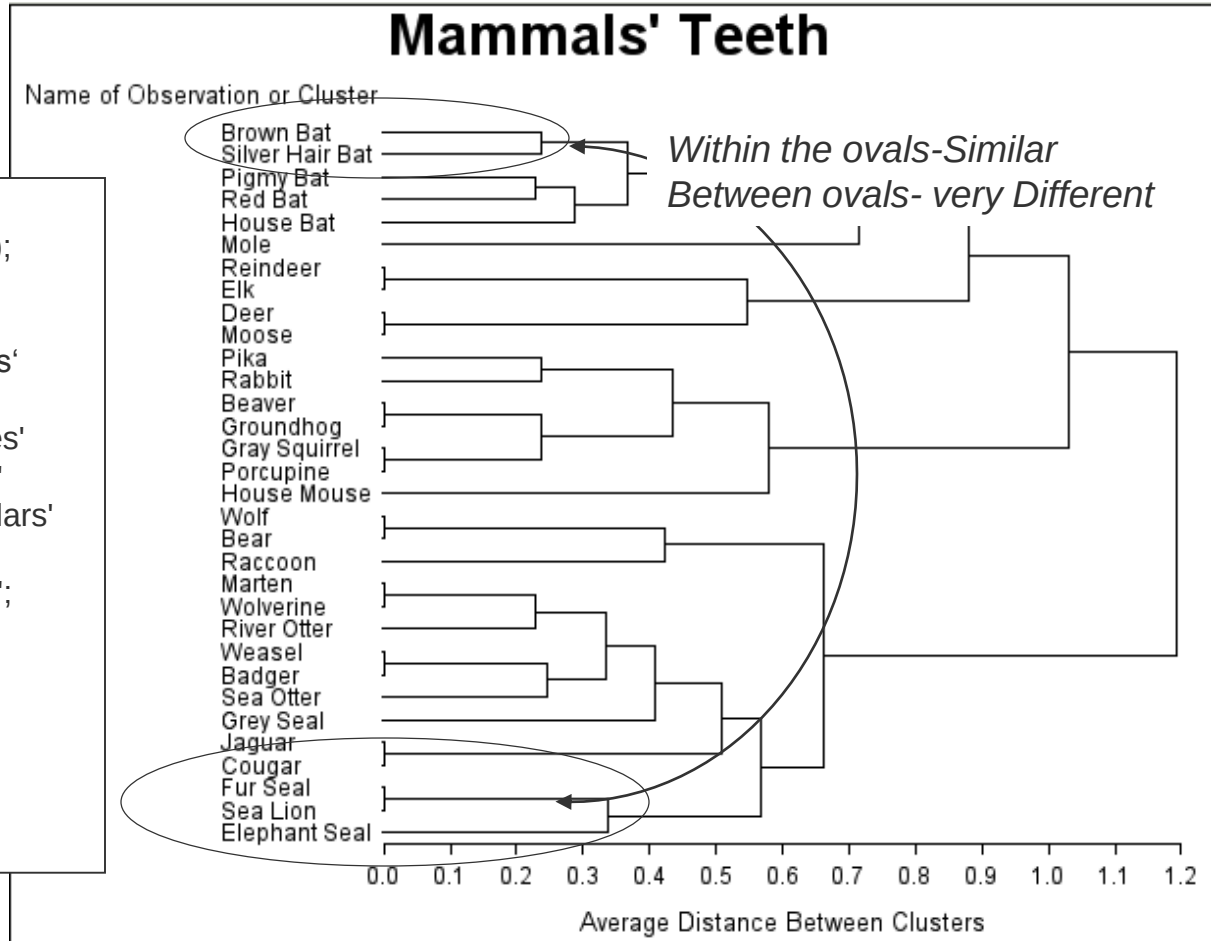


Tasks – Multivariate – Cluster Analysis

Reading a Tree Diagram

```
data teeth; title 'Mammals'' Teeth';
 input mammal $ 1-16 @21 (v1-v8) (1.);

 label          V1='Right Top Incisors'
                V2='Right Bottom Incisors‘
                V3='Right Top Canines'
                V4='Right Bottom Canines'
                V5='Right Top Premolars'
                V6='Right Bottom Premolars'
                V7='Right Top Molars'
                V8='Right Bottom Molars';
datalines;
Brown Bat                23113333
Mole                     32103333
Silver Hair Bat          23112333
Pigmy Bat                23112233
House Bat                23111233
```



Practical Uses of Hierarchical Clustering

- Grouping Customers for market analysis and to find outliers
- Grouping Products to reduce the number of processing models
- As an alternate to regression modeling

Cluster Analysis: Who/what is alike, who/what is different, who/what stands alone

ArcelorMittal