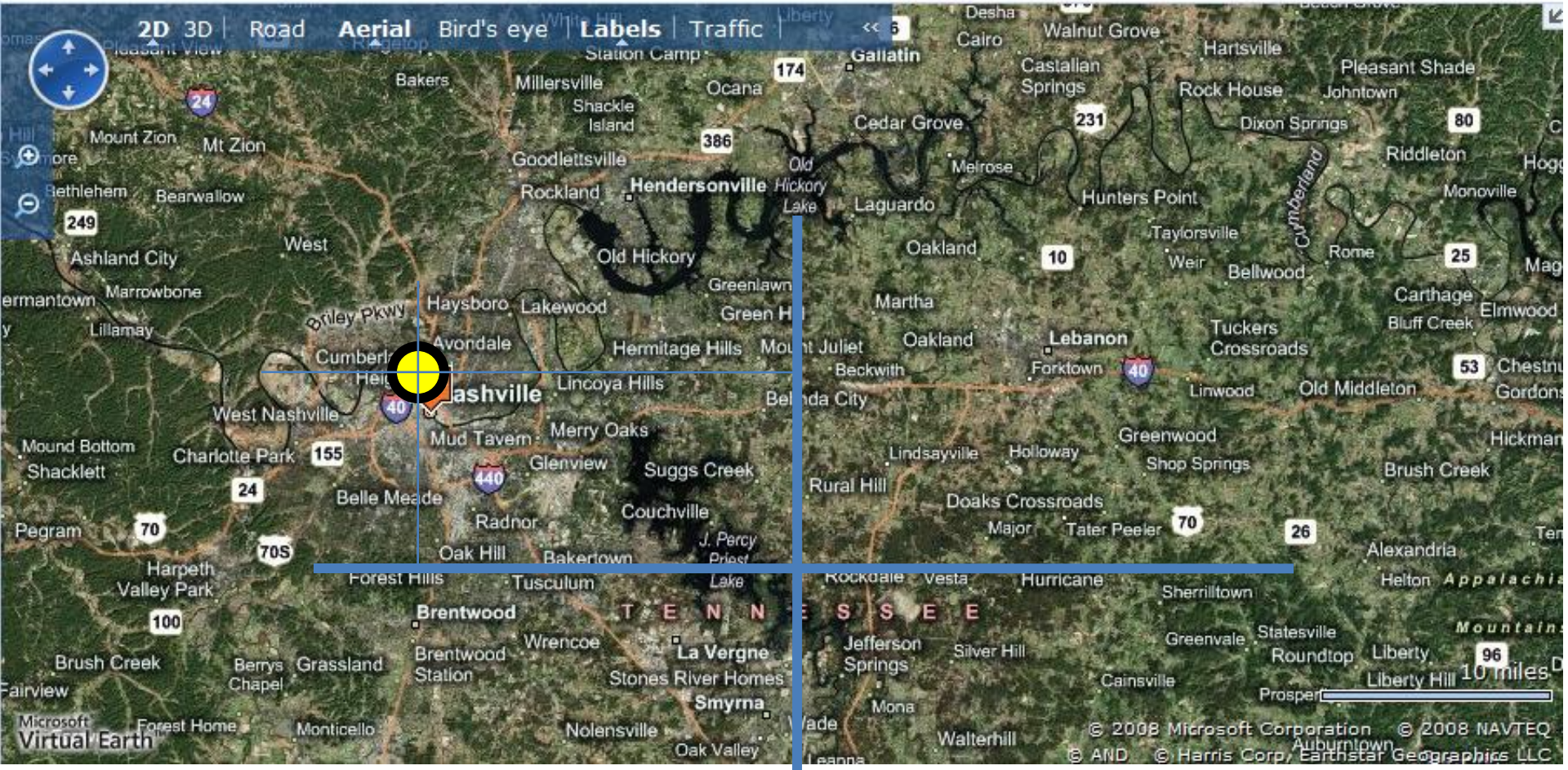
Understanding Data Mining

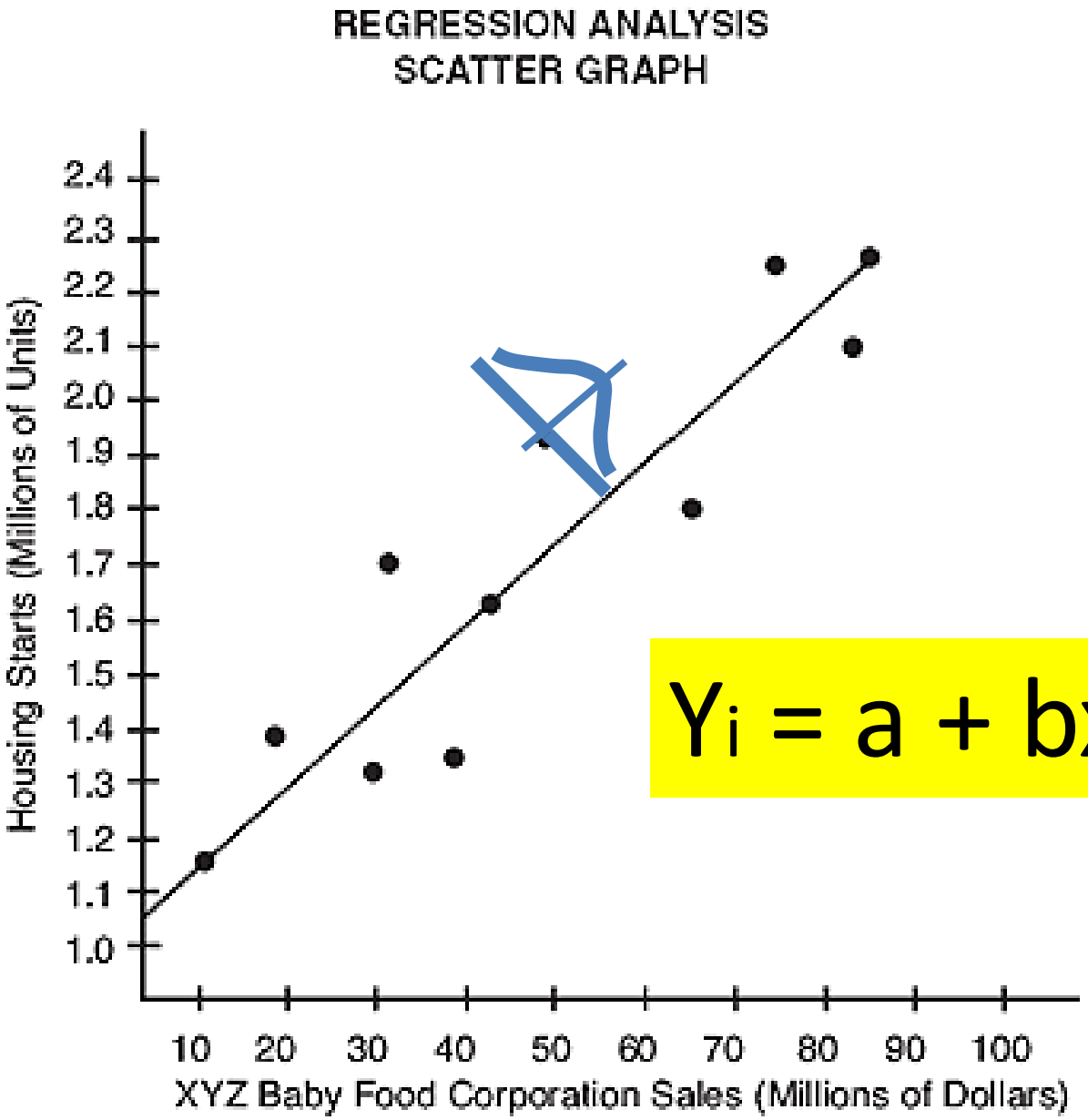
Craig A. Stevens, PMP, CC
craigastevens@westbrookstevens.com
www.westbrookstevens.com

Examples of Classical Statistical Methods

Latitude 36.19N and *Longitude* -86.78W

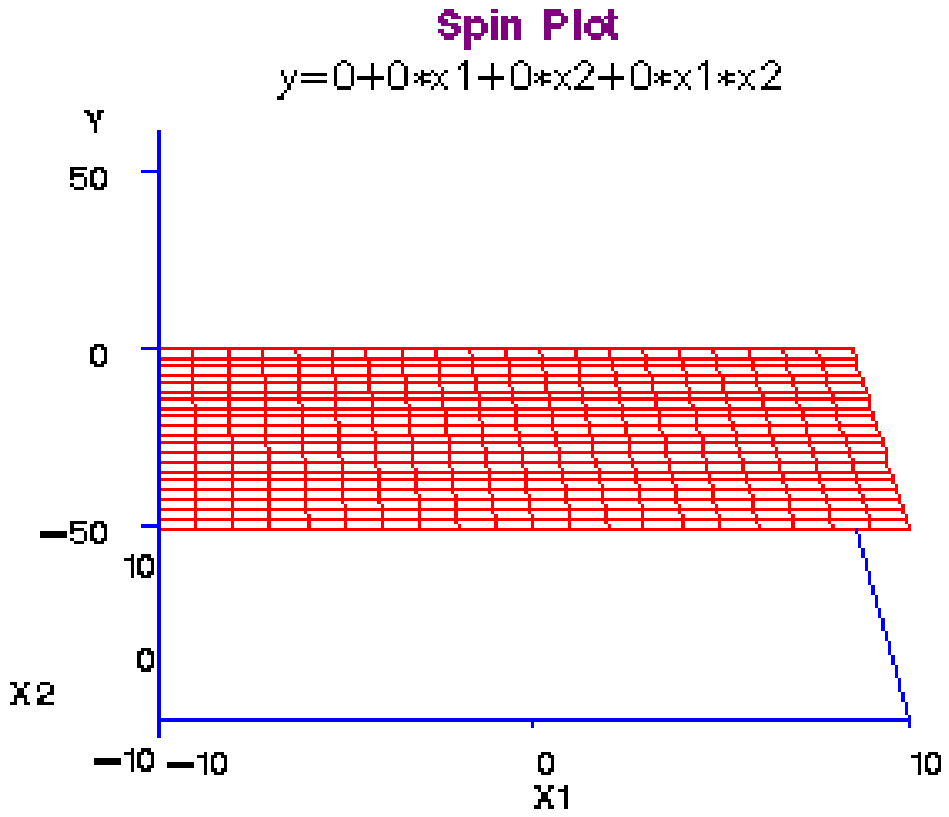


Nashville, TN, USA


REGRESSION ANALYSIS
SCATTER GRAPH
Housing Starts (Millions of Units)
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
10
20
30
40
50
60
70
80
90
100
XYZ Baby Food Corporation Sales (Millions of Dollars)

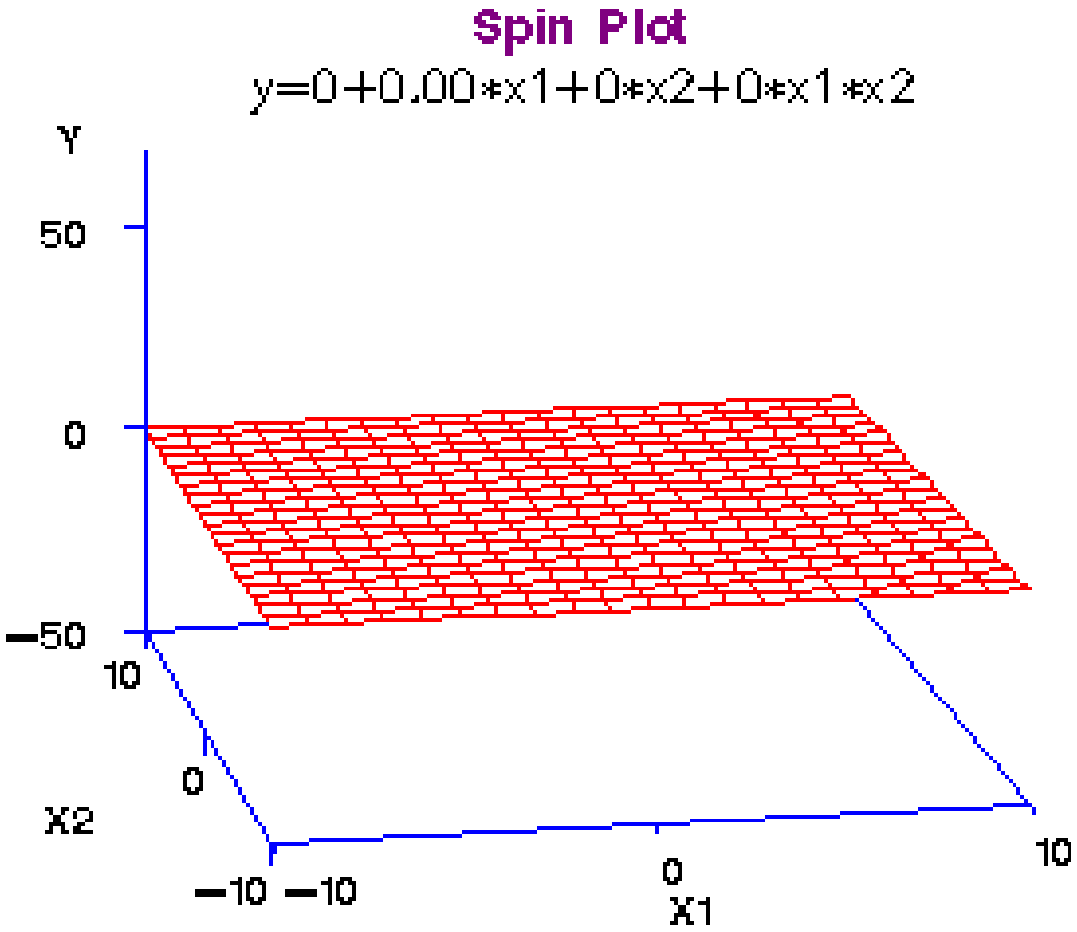
$$Y_i = a + bx_i + e$$

Multiple Regression



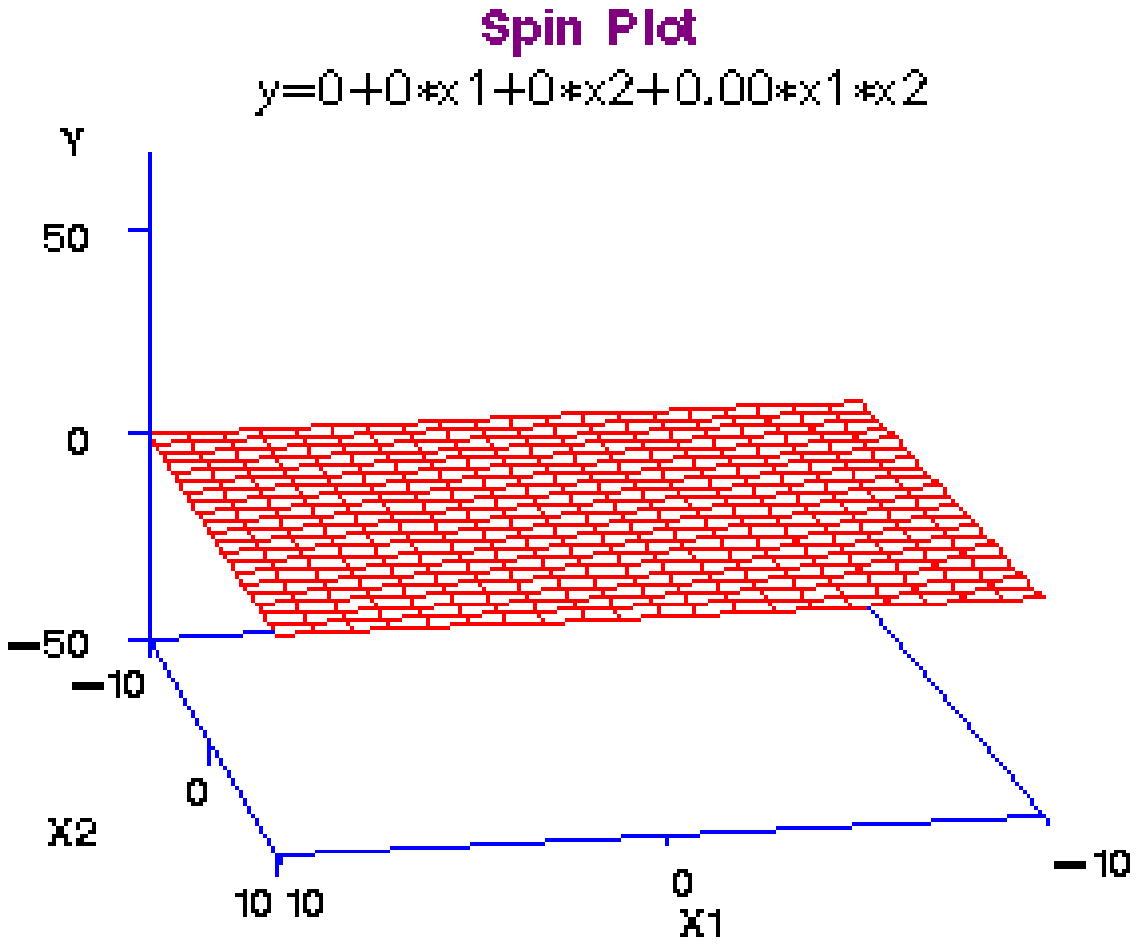
http://www.ats.ucla.edu/stat/sas/faq/spplot/reg_int_cont.htm

Multiple Regression



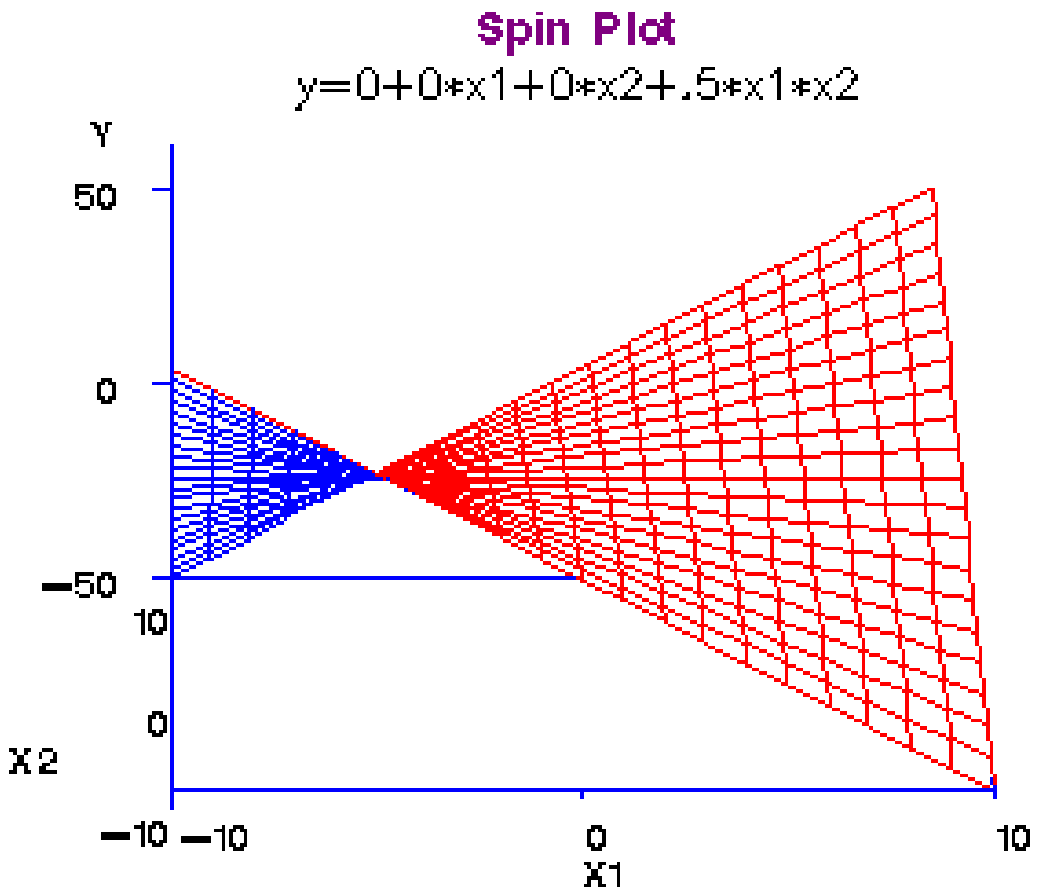
http://www.ats.ucla.edu/stat/sas/faq/spplot/reg_int_cont.htm

Multiple Regression



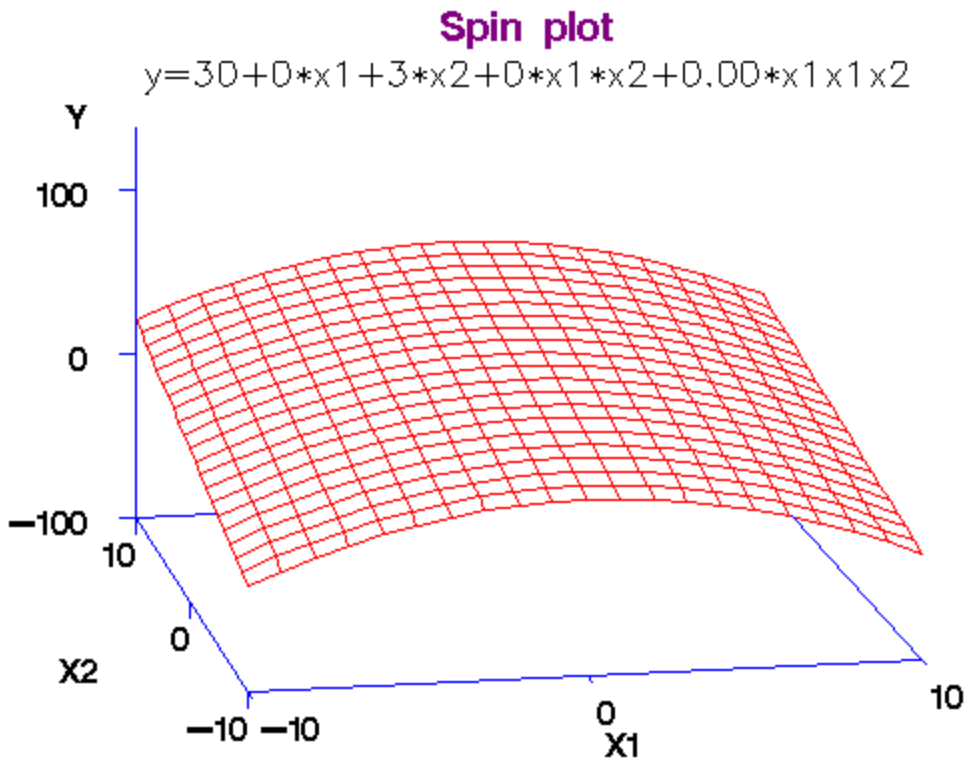
http://www.ats.ucla.edu/stat/sas/faq/spplot/reg_int_cont.htm

Multiple Regression



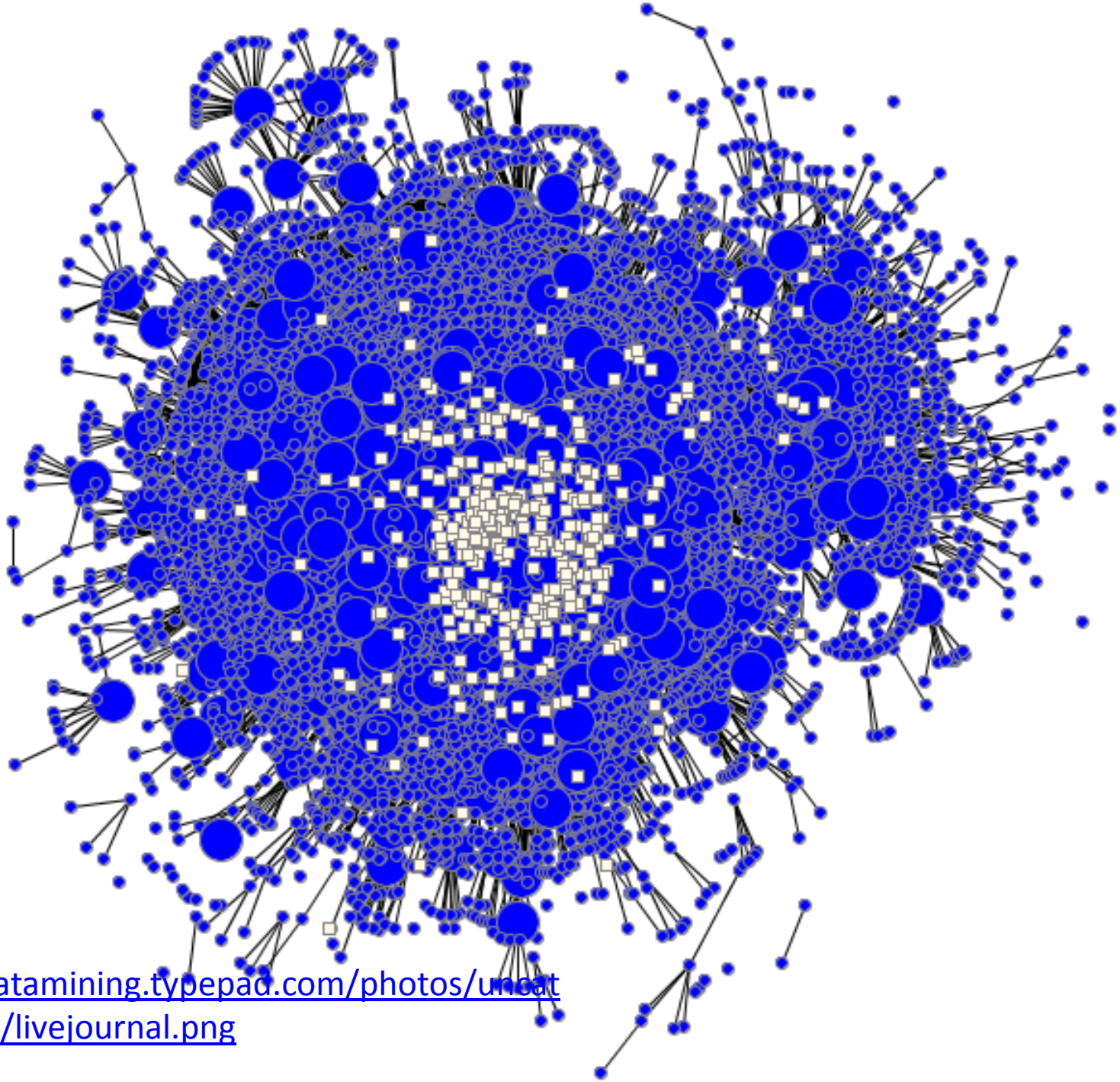
http://www.ats.ucla.edu/stat/sas/faq/spplot/reg_int_cont.htm

Multiple Regression



http://www.ats.ucla.edu/stat/sas/faq/spplot/reg_int_cont.htm

Data Mining

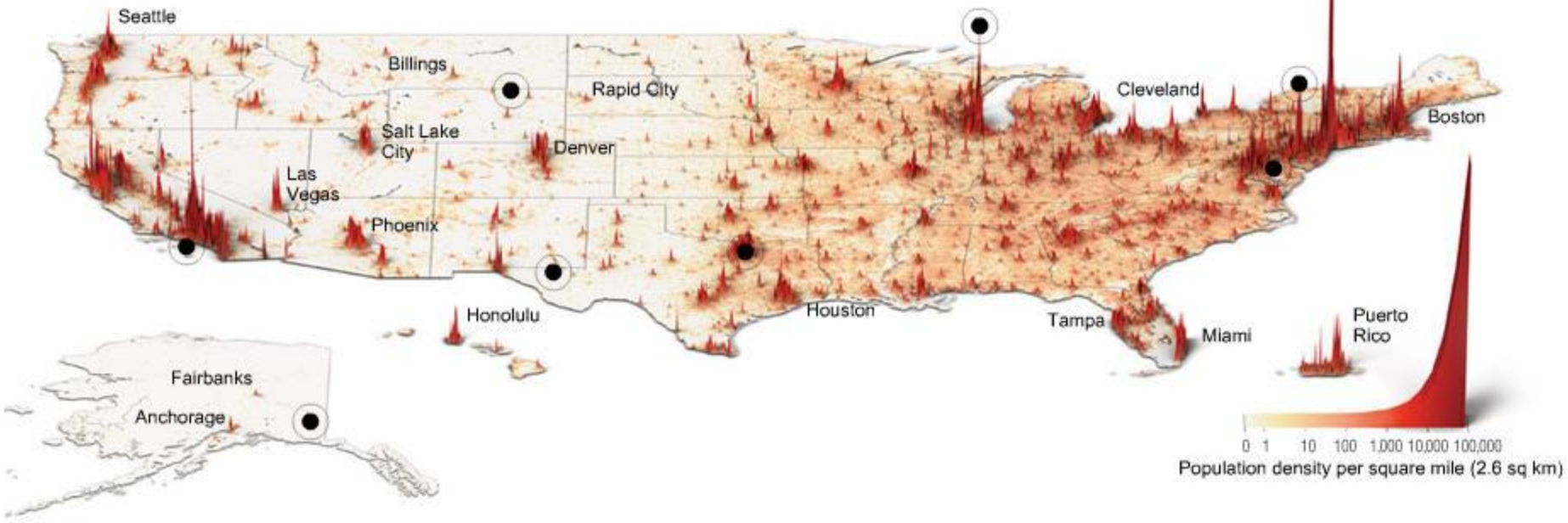


http://datamining.typepad.com/photos/uncategorized/livejournal.png

This Is Where We Live

Take a look at America by the numbers.

Roll over the map to learn more»



What is Data Mining?

- The process of identifying hidden patterns, trends, and relationships in large quantities of data.

Why Do Data Mining?

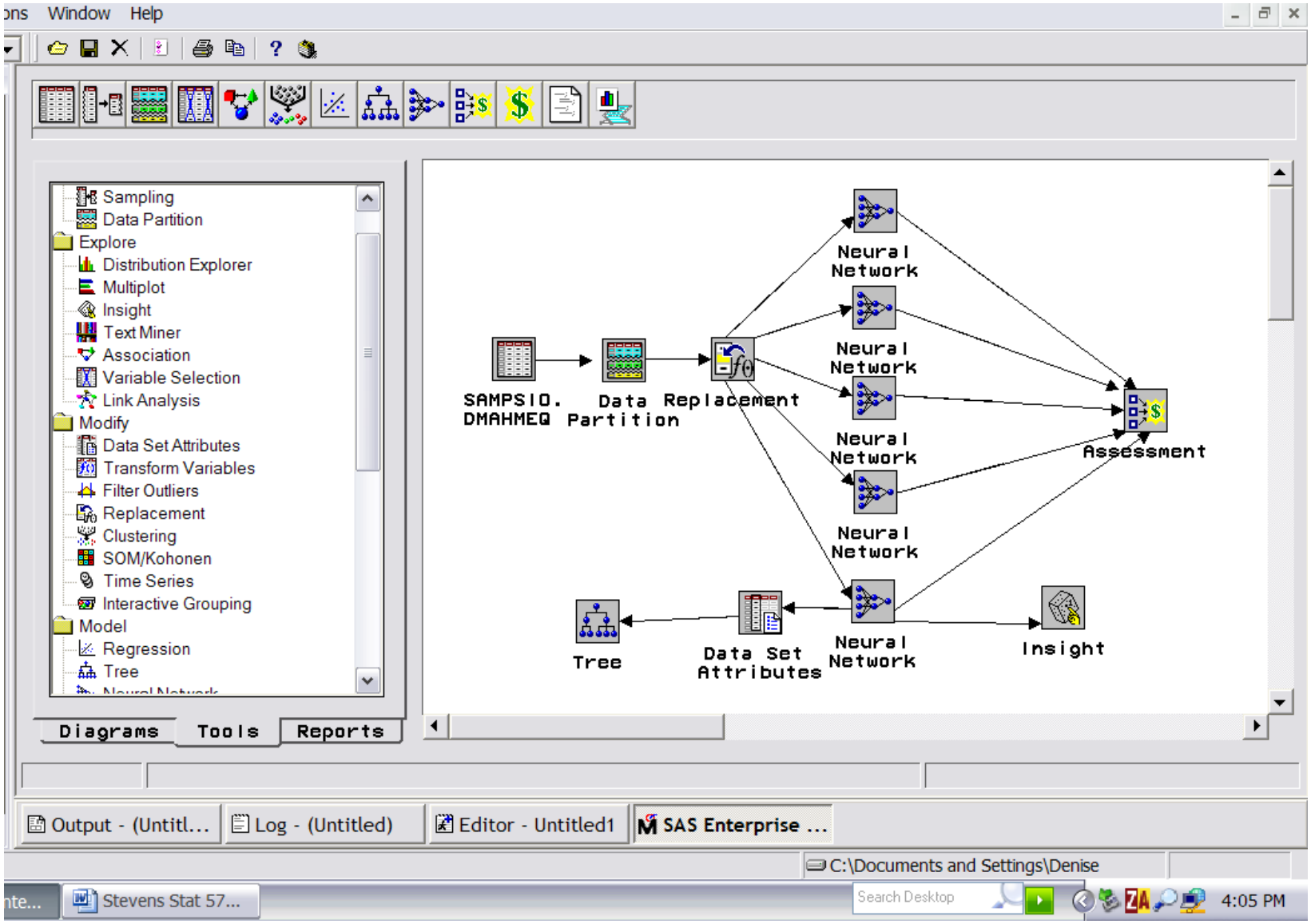
- To discover useful information for making decisions.
- Too many variables for Classical Statistical methods to work.
  - Large Number of Records 10^8 - 10^{12}
    - Gigabyte – Terabyte
  - High Dimensional Data
    - Lots of Variables (10 – 10^4 attributes)

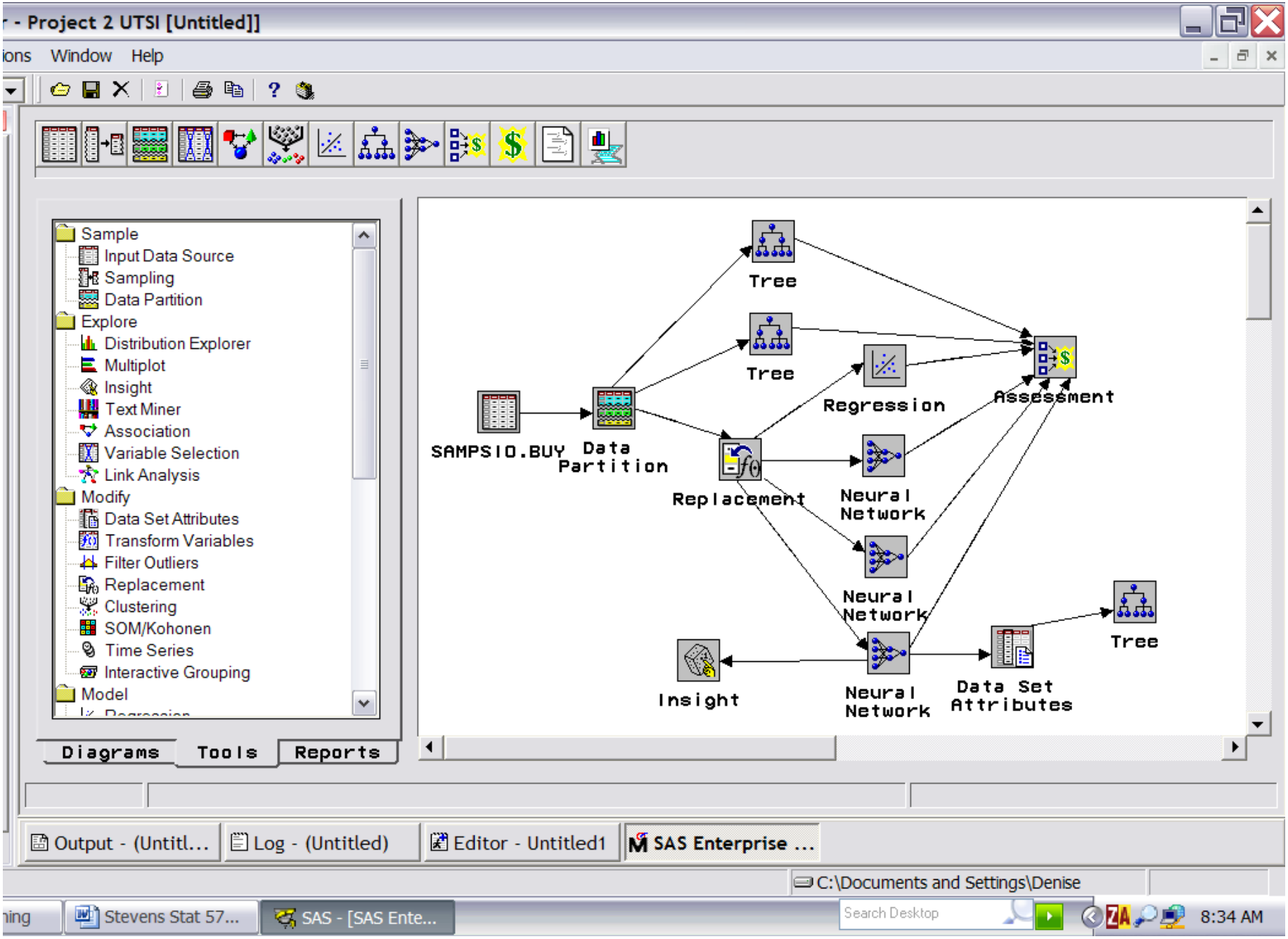
The Huber-Wegman Taxonomy of Data Set Sizes

Descriptor	Data Set Size in Bytes	Storage Mode
Tiny	10^2	Piece of Paper
Small	10^4	A few Pieces of Paper
Medium	10^6	A Floppy Disk
Large	10^8	Hard Disk
Huge	10^10	Multiple Hard Disks
Massive	10^12	Robotic Magnetic Tape Storage Silos
Super Massive	10^15	Distributed Data Archives

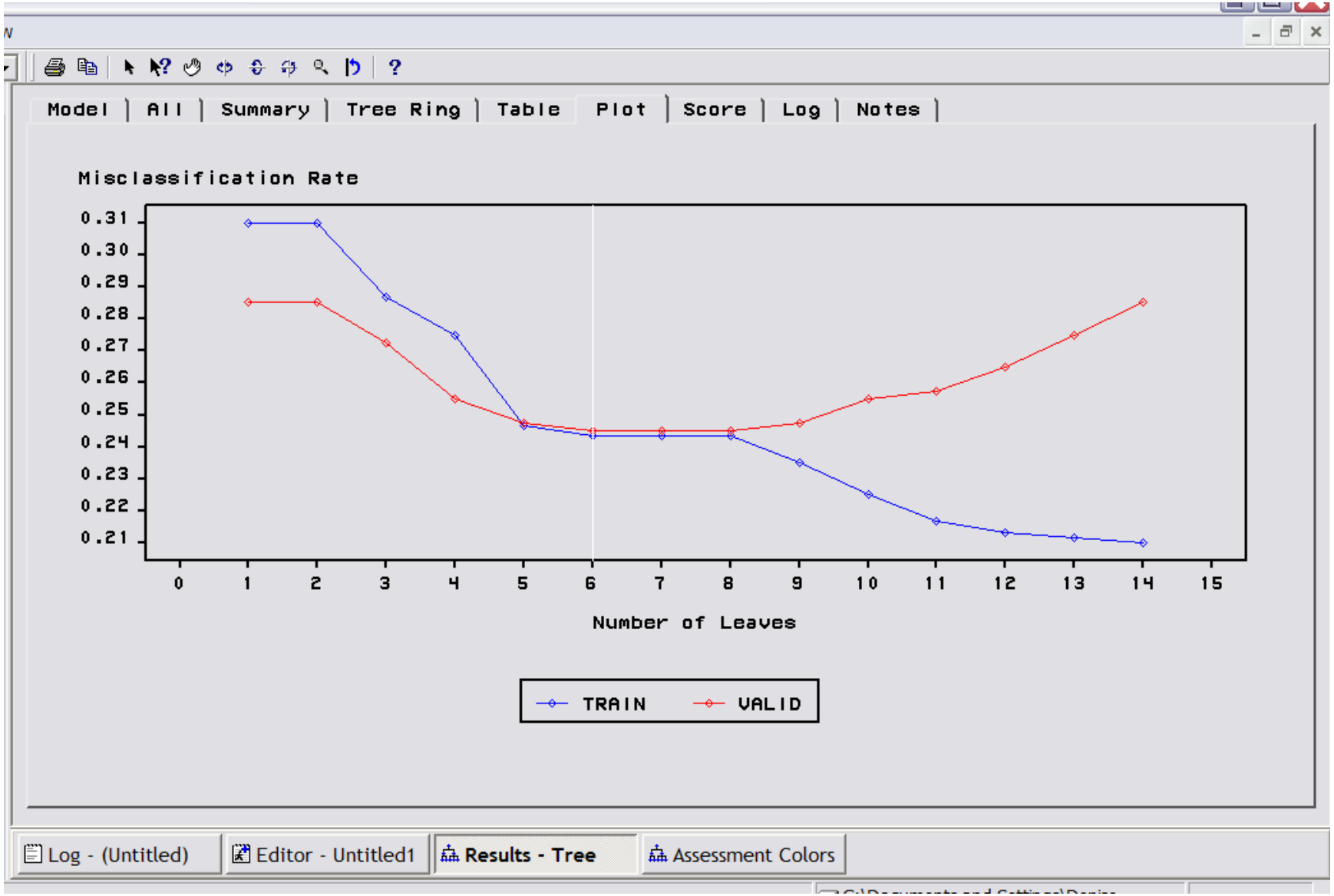
Name	Model Role	Measurement Level	Description
BAD	Target	Binary	1=client defaulted on loan 0=loan repaid
CLAGE	Input	Interval	Age of oldest trade line in months
CLNO	Input	Interval	Number of trade lines
DEBTINC	Input	Interval	Debt-to-income ratio
DELINQ	Input	Interval	Number of trade lines
DEROG	Input	Interval	Number of major derogatory reports
JOB	Input	Nominal	Six occupational categories
LOAN	Input	Interval	Amount of the loan request
MORTDUE	Input	Interval	Amount due on existing mortgage
NINQ	Input	Interval	Number of recent credit inquiries
REASON	Input	Binary	DebtCon=debt consolidation, HomeImp=home improvement
VALUE	Input	Interval	Value of current property
YOJ	Input	Interval	Years at present job

SAS Enterprise Miner Objects

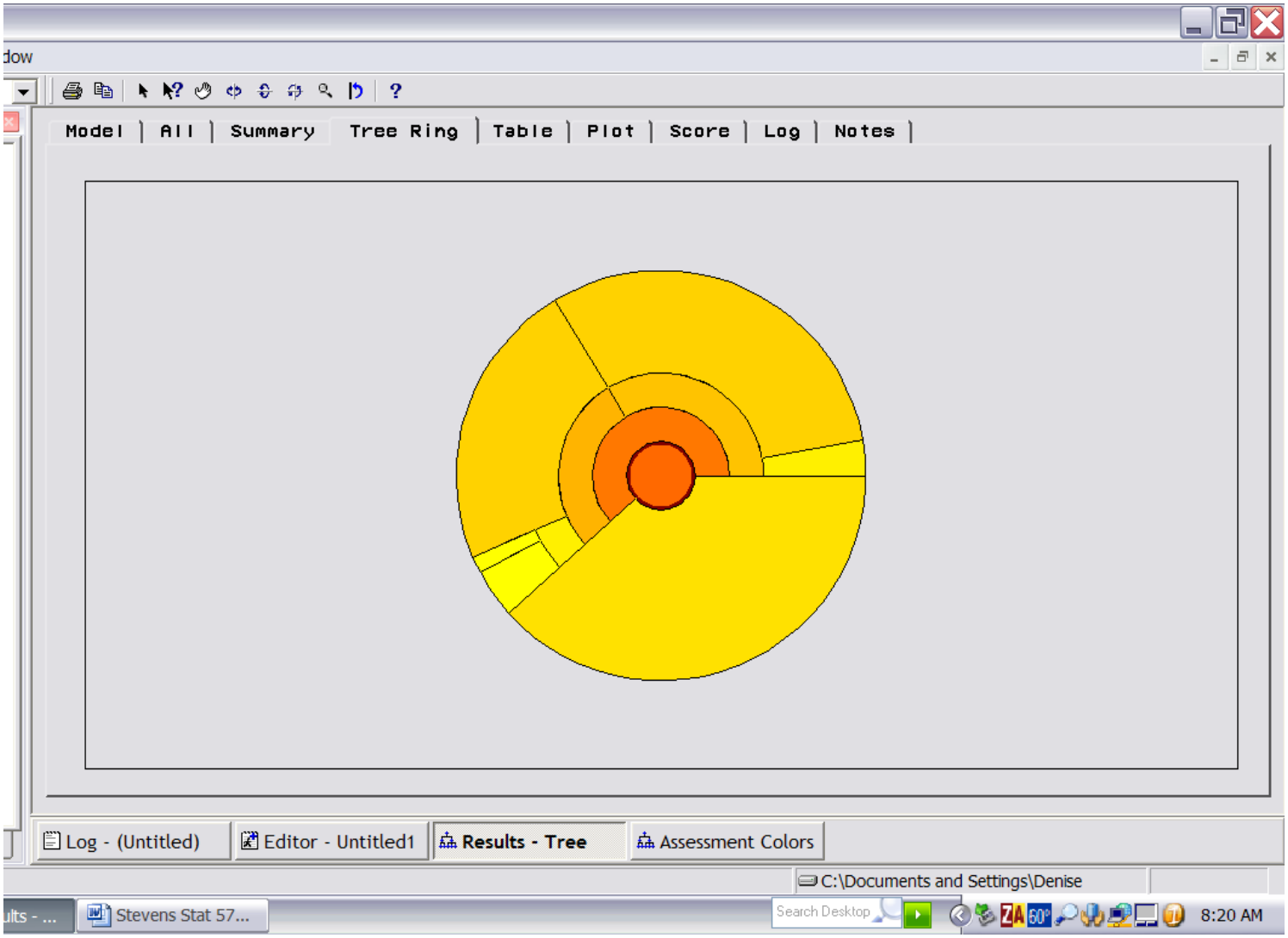



- Project 2 UTSI [Untitled]]
ions Window Help
Sample
Input Data Source
Sampling
Data Partition
Explore
Distribution Explorer
Multiplot
Insight
Text Miner
Association
Variable Selection
Link Analysis
Modify
Data Set Attributes
Transform Variables
Filter Outliers
Replacement
Clustering
SOM/Kohonen
Time Series
Interactive Grouping
Model
Diagrams
Tools
Reports
Tree
Tree
Regression
Assessment
SAMPSIO.BUY
Data Partition
Replacement
Neural Network
Neural Network
Insight
Neural Network
Data Set Attributes
Tree
Output - (Untitl...
Log - (Untitled)
Editor - Untitled1
SAS Enterprise ...
C:\Documents and Settings\Denise
Stevens Stat 57...
SAS - [SAS Ente...
Search Desktop
8:34 AM

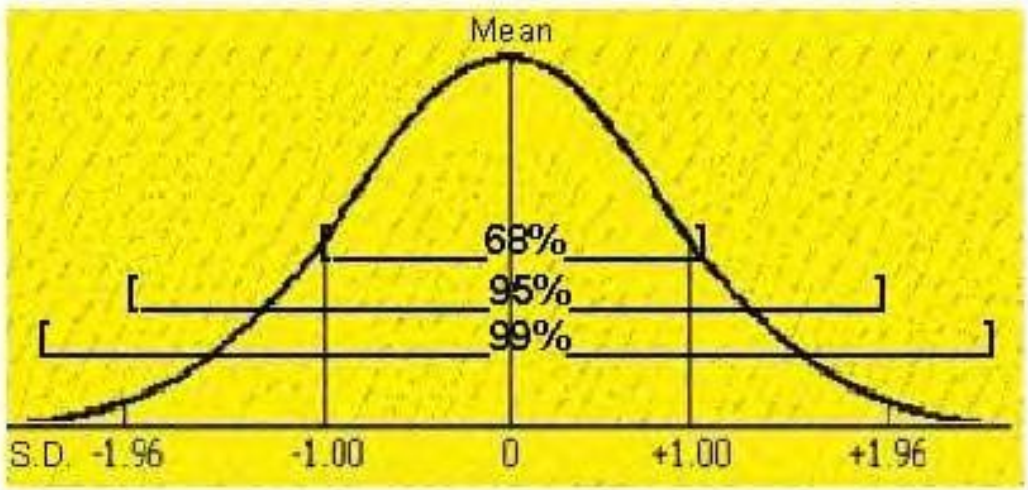
Shows the Cut off Point is 6 Variables



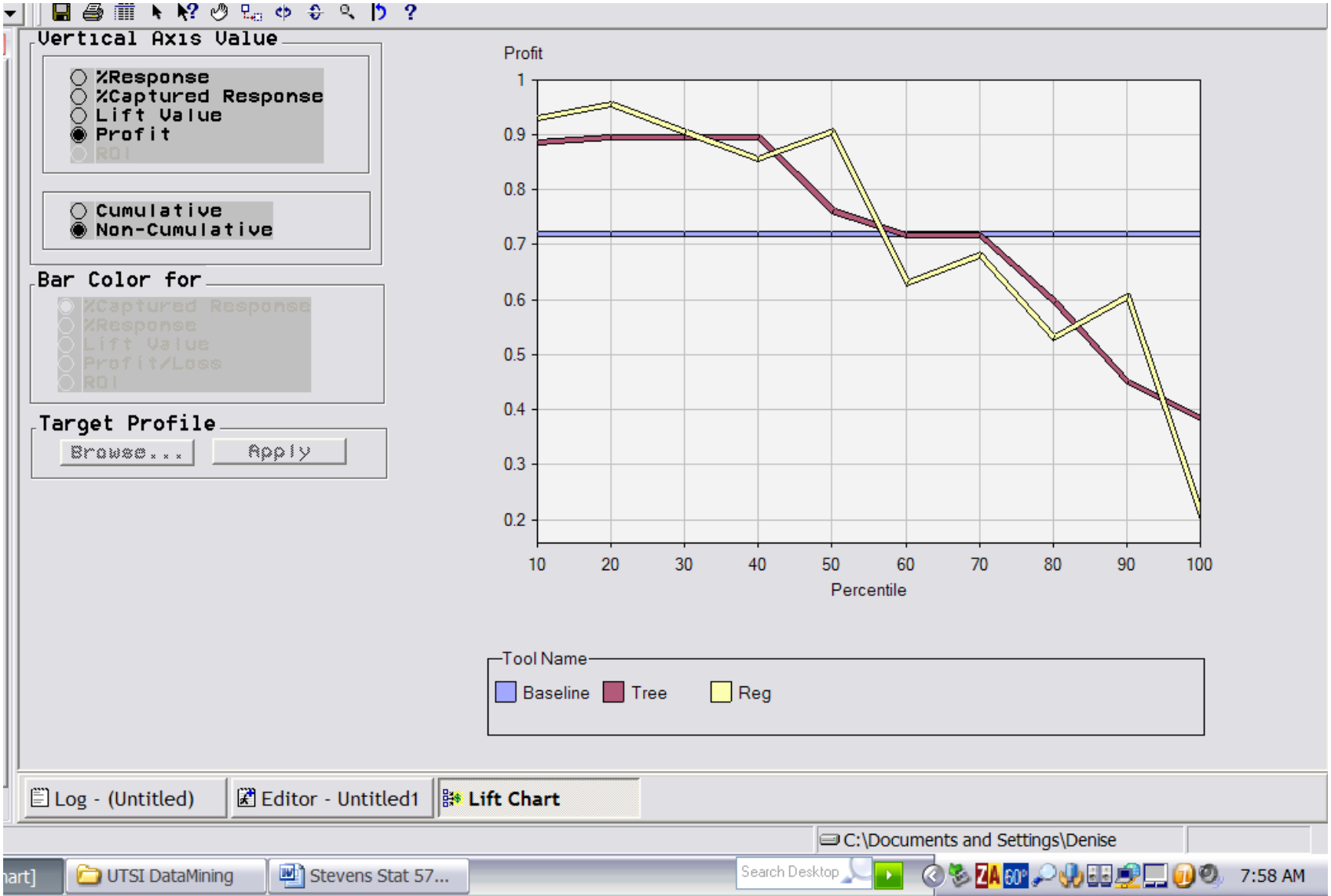
Small Number of Useful Variables

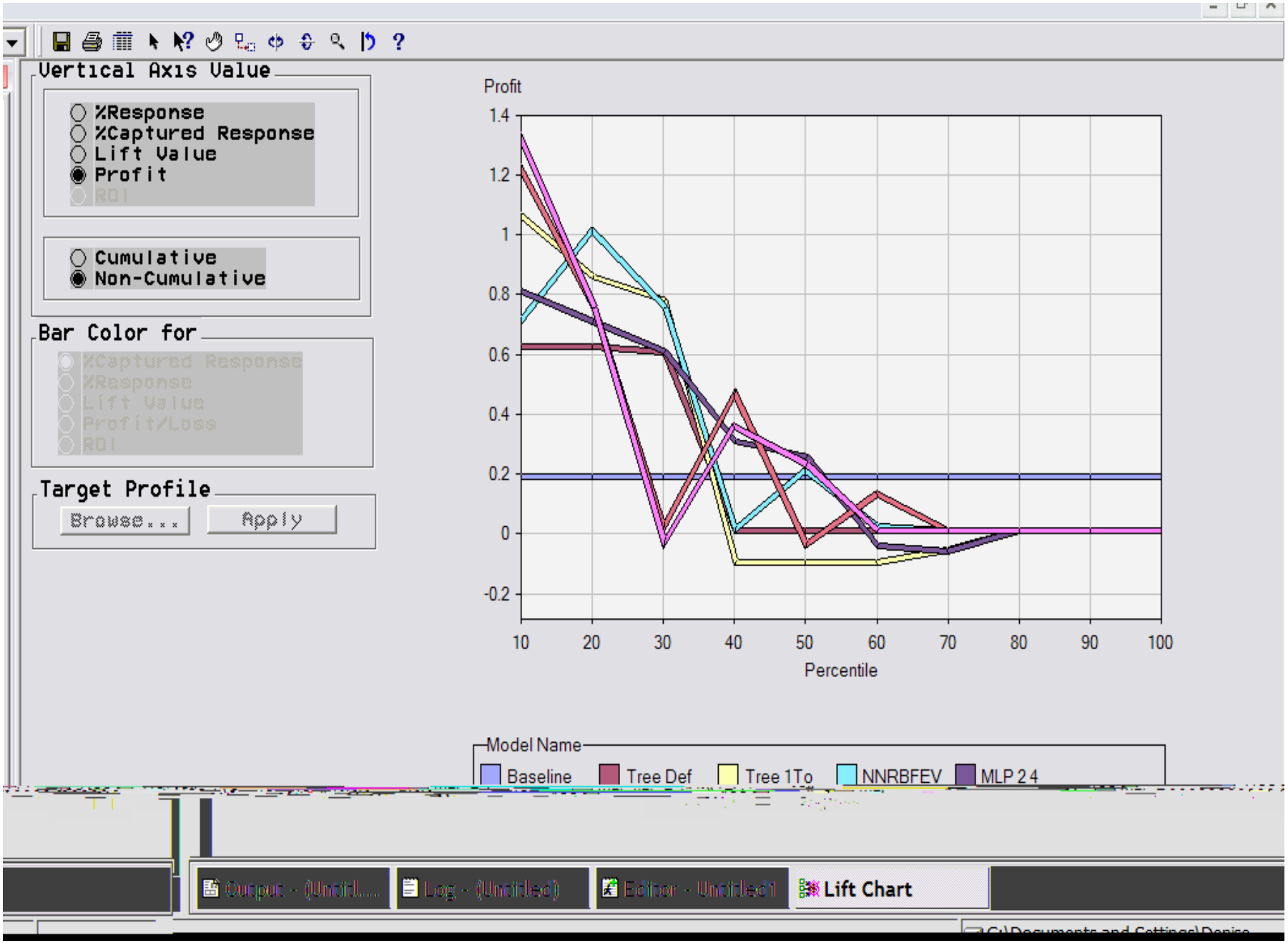


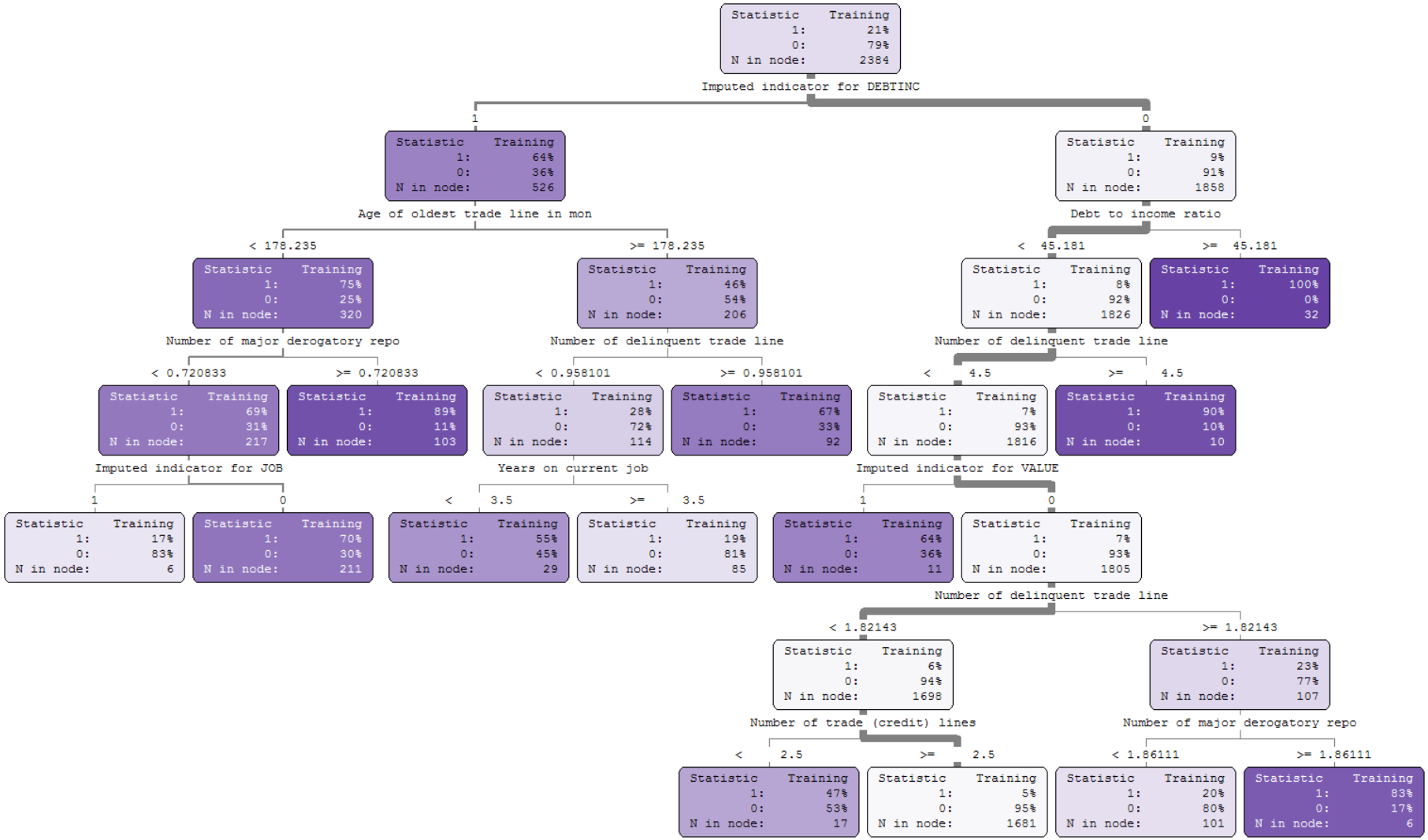
The Normal Curve



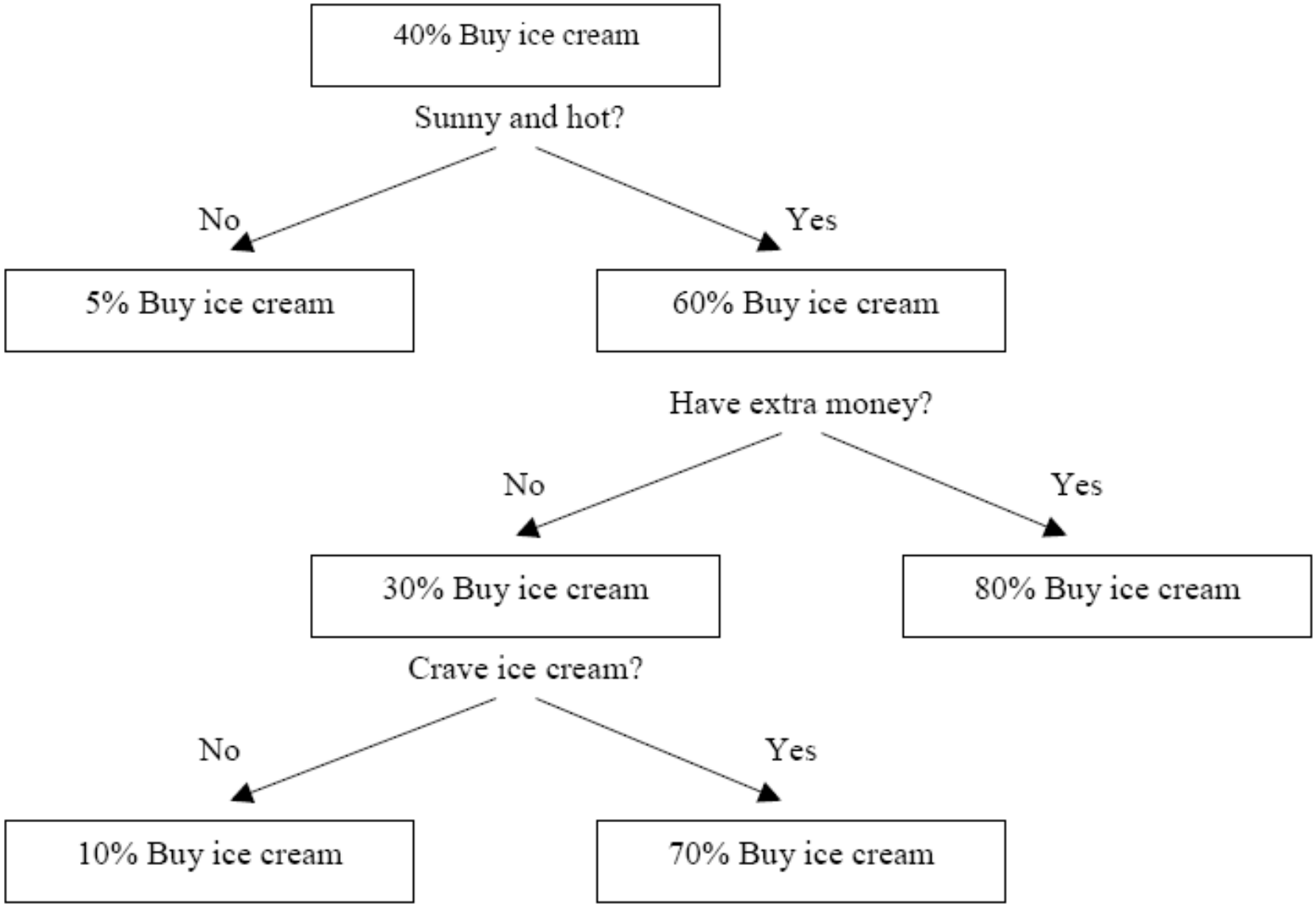
Comparing Methods and Profit vs Marketing Cost



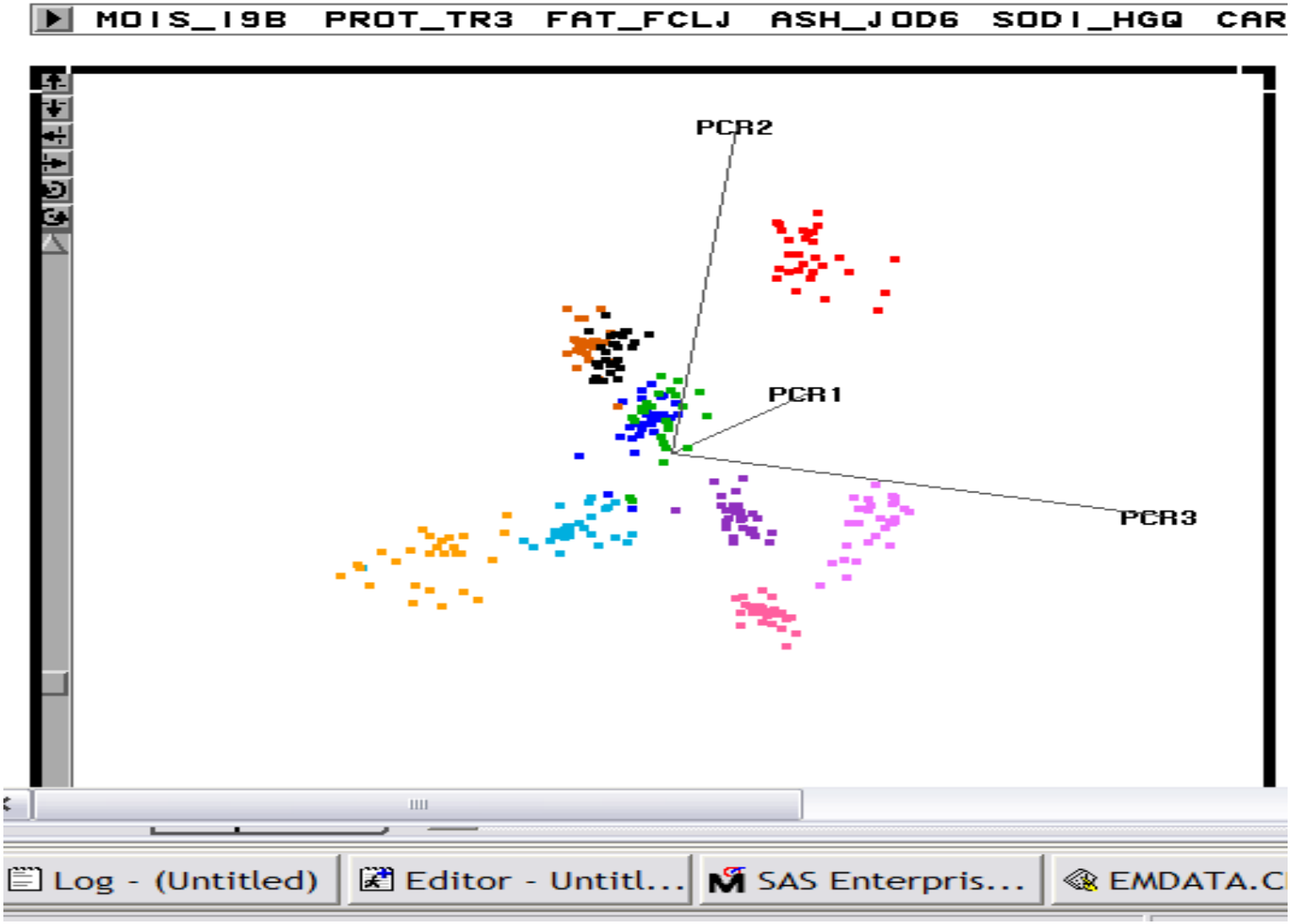

Vertical Axis Value
%Response
%Captured Response
Lift Value
Profit
ROI
Cumulative
Non-Cumulative
Bar Color for
%Captured Response
%Response
Lift Value
Profit/Loss
ROI
Target Profile
Browse...
Apply
Profit
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
10
20
30
40
50
60
70
80
90
100
Percentile
Model Name
Baseline
Tree Def
Tree 1To
NNRBFEV
MLP 2 4
Output - (Untitl...
Log - (Untitled)
Editor - Untitled1
Lift Chart

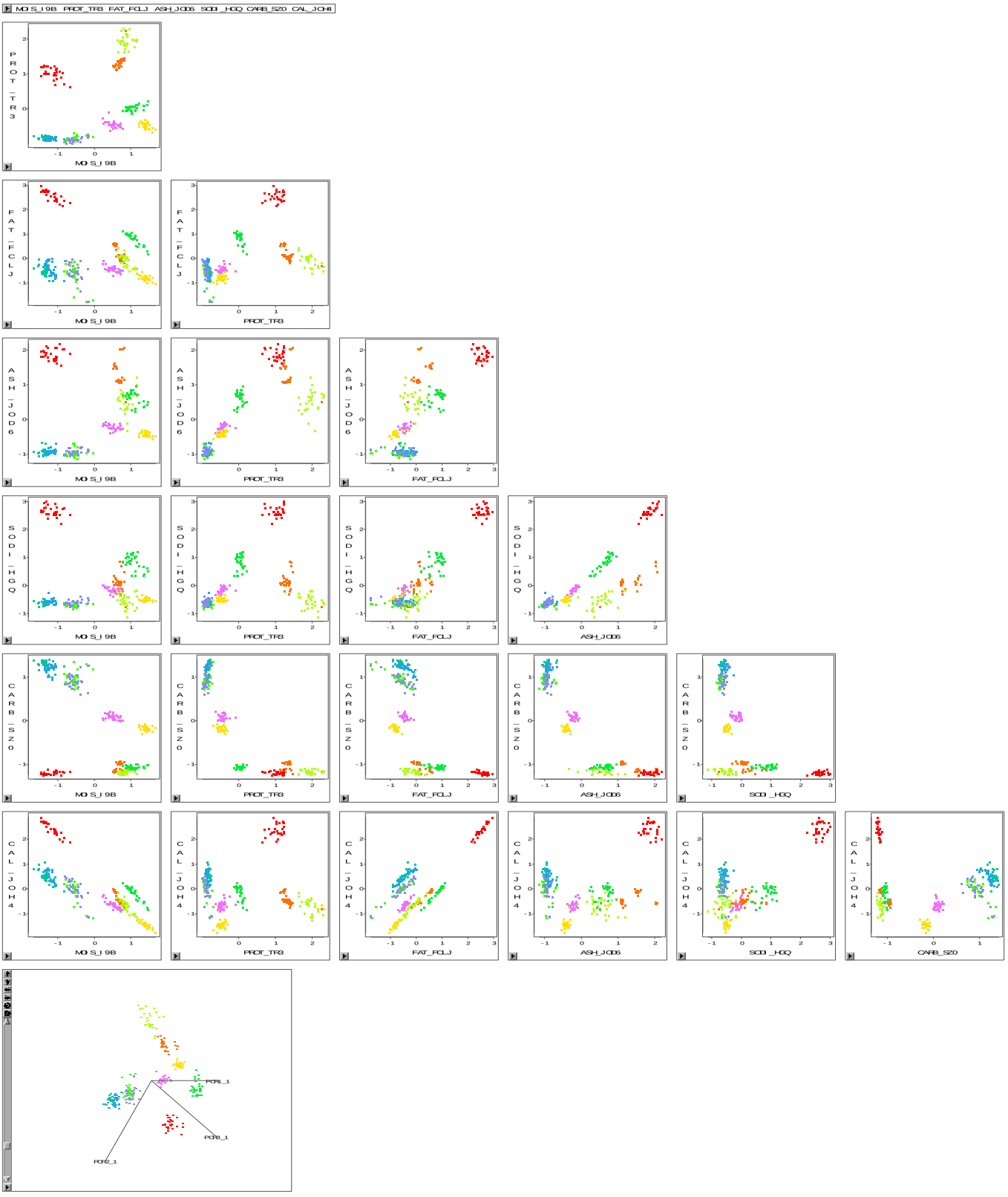

Statistic Training
1: 21%
0: 79%
N in node: 2384
Imputed indicator for DEBTINC
1
Statistic Training
1: 64%
0: 36%
N in node: 526
Age of oldest trade line in mon
< 178.235
Statistic Training
1: 75%
0: 25%
N in node: 320
Number of major derogatory repo
< 0.720833
Statistic Training
1: 69%
0: 31%
N in node: 217
Imputed indicator for JOB
1
Statistic Training
1: 17%
0: 83%
N in node: 6
0
Statistic Training
1: 70%
0: 30%
N in node: 211
>= 0.720833
Statistic Training
1: 89%
0: 11%
N in node: 103
>= 178.235
Statistic Training
1: 46%
0: 54%
N in node: 206
Number of delinquent trade line
< 0.958101
Statistic Training
1: 28%
0: 72%
N in node: 114
Years on current job
< 3.5
Statistic Training
1: 55%
0: 45%
N in node: 29
>= 3.5
Statistic Training
1: 19%
0: 81%
N in node: 85
>= 0.958101
Statistic Training
1: 67%
0: 33%
N in node: 92
0
Statistic Training
1: 9%
0: 91%
N in node: 1858
Debt to income ratio
< 45.181
Statistic Training
1: 8%
0: 92%
N in node: 1826
Number of delinquent trade line
< 4.5
Statistic Training
1: 7%
0: 93%
N in node: 1816
Imputed indicator for VALUE
1
Statistic Training
1: 64%
0: 36%
N in node: 11
0
Statistic Training
1: 7%
0: 93%
N in node: 1805
Number of delinquent trade line
< 1.82143
Statistic Training
1: 6%
0: 94%
N in node: 1698
Number of trade (credit) lines
< 2.5
Statistic Training
1: 47%
0: 53%
N in node: 17
>= 2.5
Statistic Training
1: 5%
0: 95%
N in node: 1681
>= 1.82143
Statistic Training
1: 23%
0: 77%
N in node: 107
Number of major derogatory repo
< 1.86111
Statistic Training
1: 20%
0: 80%
N in node: 101
>= 1.86111
Statistic Training
1: 83%
0: 17%
N in node: 6
>= 4.5
Statistic Training
1: 90%
0: 10%
N in node: 10
>= 45.181
Statistic Training
1: 100%
0: 0%
N in node: 32

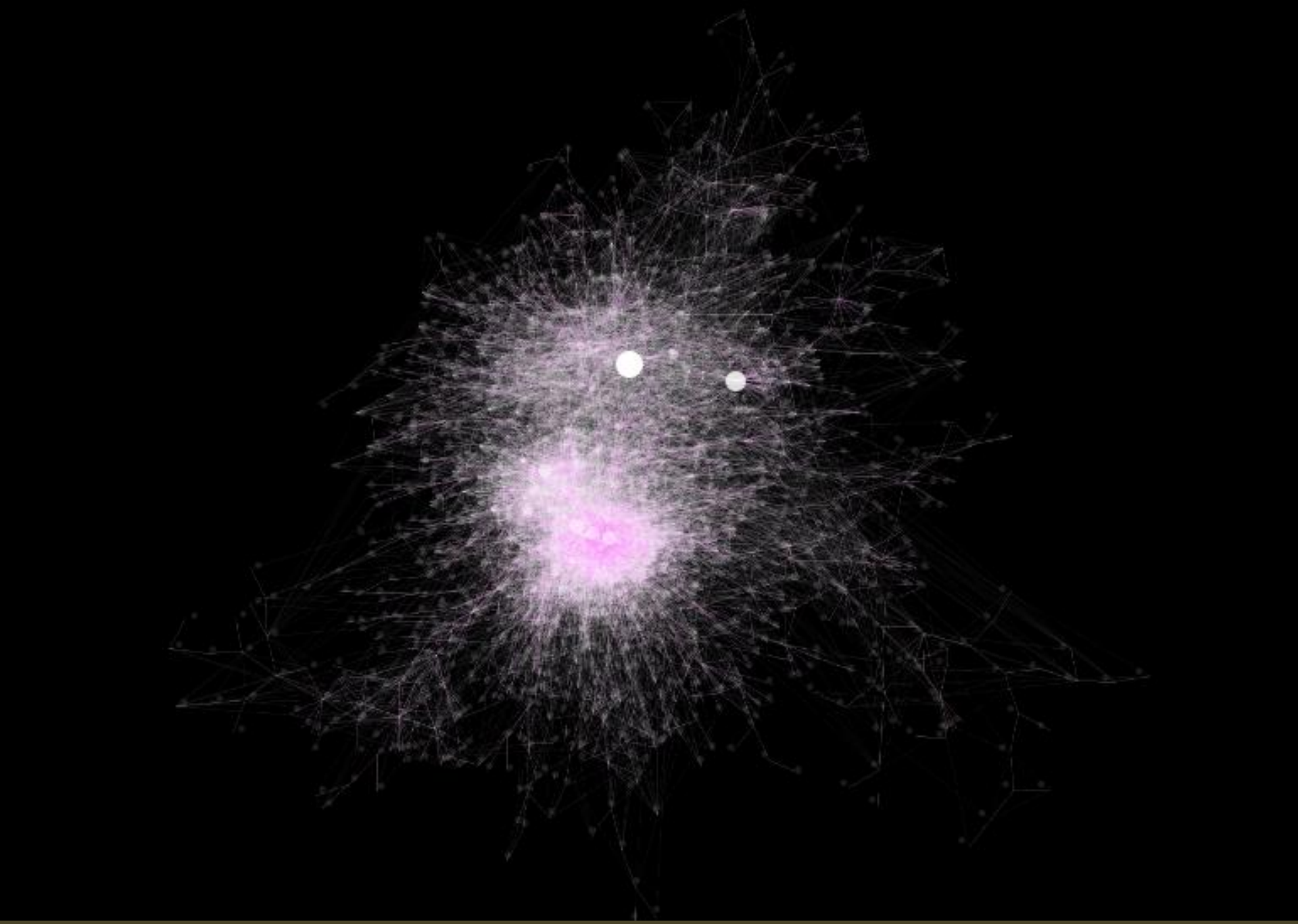
Decision Trees for Predictive Modeling
Padraic G. Neville SAS Institute Inc. 4 August 1999



Clustering As in Different Brands

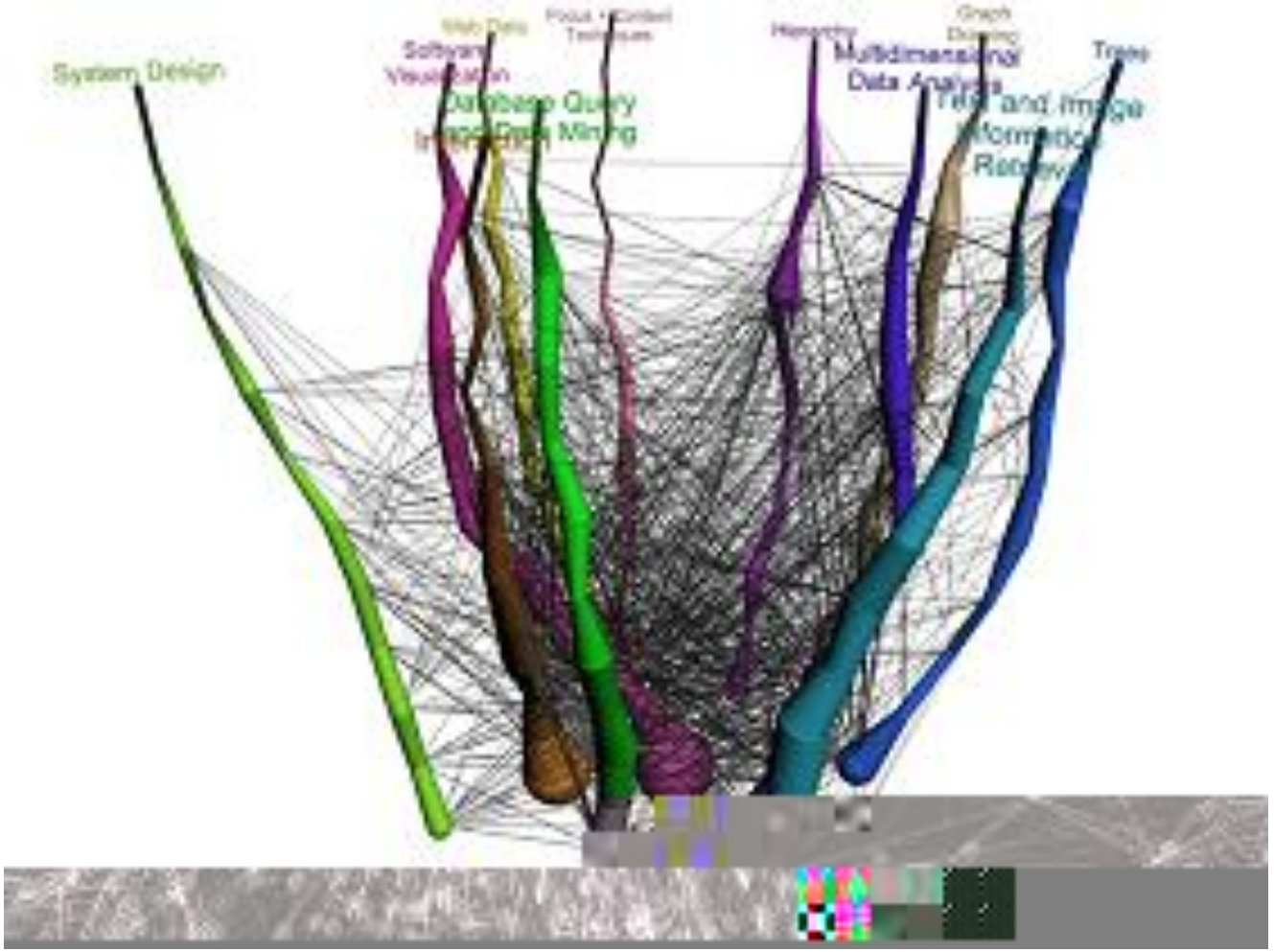



MOS_J9B
PROT_TR3
FAT_FCLJ
ASH_JOD6
SOD_HGQ
CARB_SZ0
CAL_JOH4
PCR6_1
PCR8_1
PCR2_1

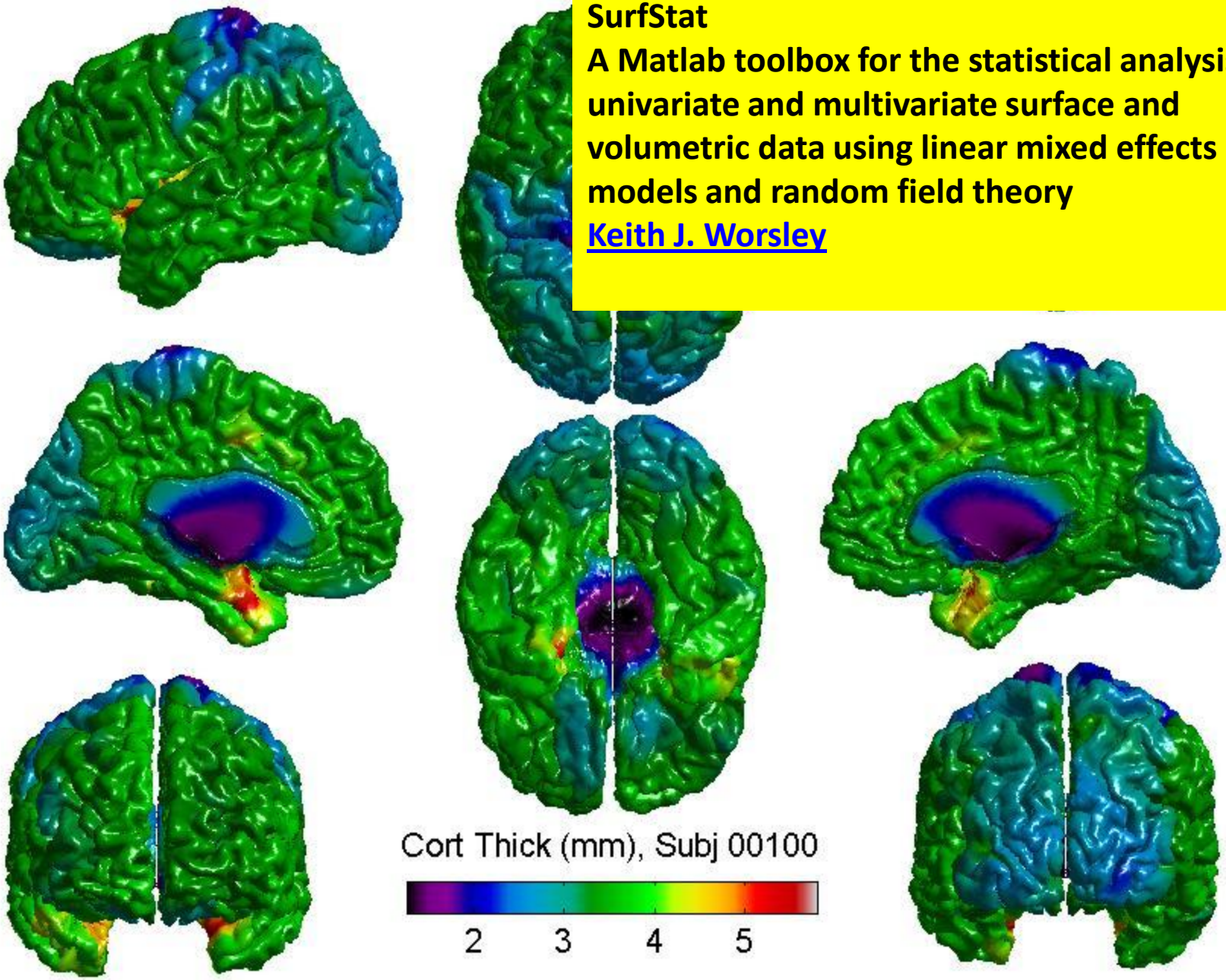


Data Mining Art found at http://datamining.typepad.com/data_mining/dataviz/page/2/

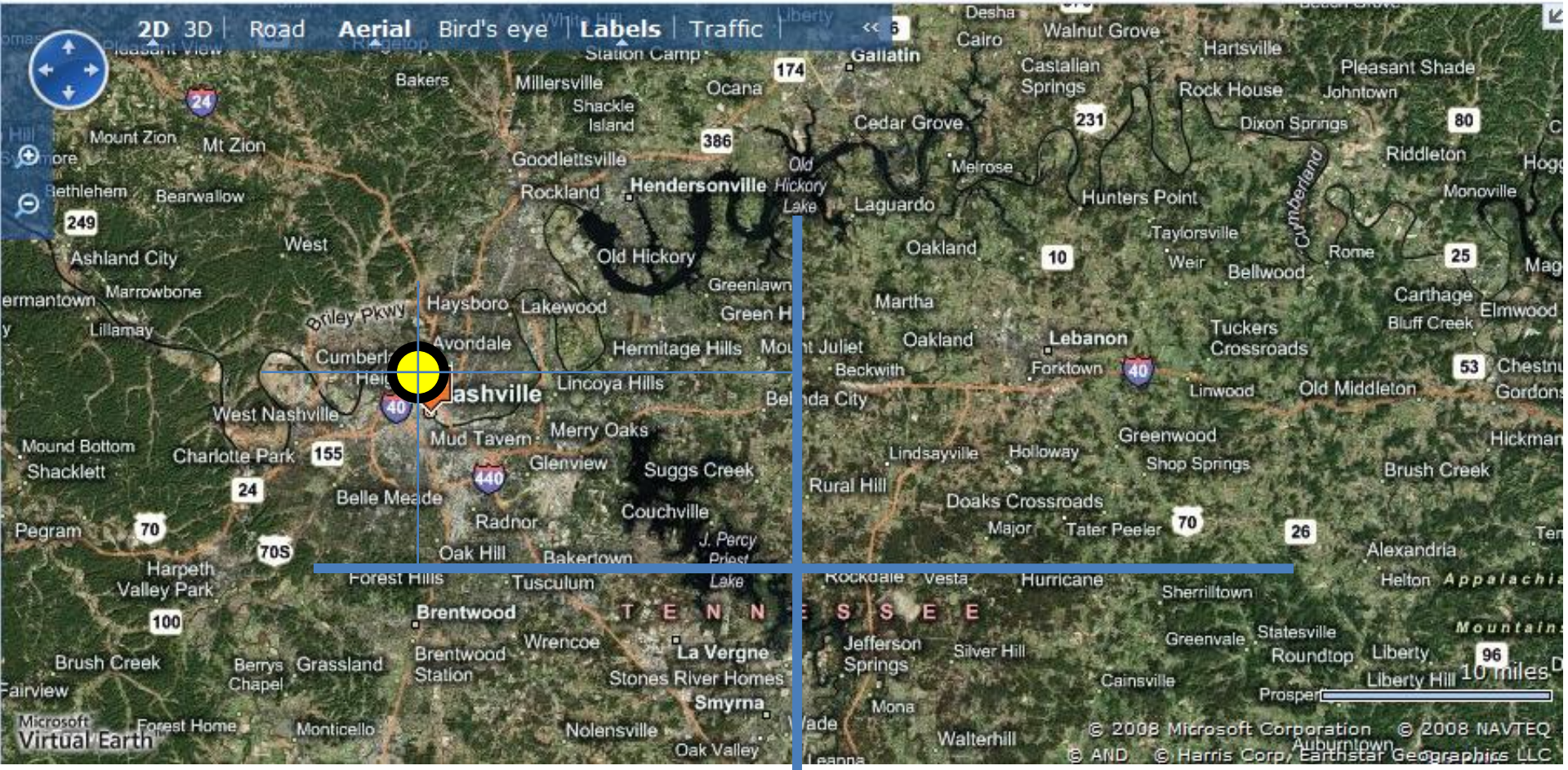
Data Mining Art found at http://datamining.typepad.com/data_mining/dataviz/page/2/


System Design
Web Data
Focus + Context Techniques
Hierarchy
Graph Drawing
Trees
Multidimensional Data Analysis
Text and Image
Database Query
Data Mining

SurfStat
A Matlab toolbox for the statistical analysis of univariate and multivariate surface and volumetric data using linear mixed effects models and random field theory
Keith J. Worsley



Latitude 36.19N and *Longitude* -86.78W



Nashville, TN, USA

Genealogical Tree
On You Tube

http://www.youtube.com/watch?v=CnniJR5Ah7g