Successful Data Mining in Practice: Where do we Start?

Richard D. De Veaux

Department of
Mathematics and Statistics
Williams College
Williamstown MA, 01267
deveaux@williams.edu

http://www.williams.edu/Mathematics/rdeveaux

Outline

- What is it?
- Why is it different?
- Types of models
- How to start
- Where do we go next?
- Challenges

Reason for Data Mining

Data = $$

Data Mining Is…

"the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." --- Fayyad

"finding interesting structure (patterns, statistical models, relationships) in data bases".--- Fayyad, Chaduri and Bradley

"a knowledge discovery process of extracting previously unknown, actionable information from very large data bases"--- Zornes

" a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions."
---Edelstein

What is Data Mining?

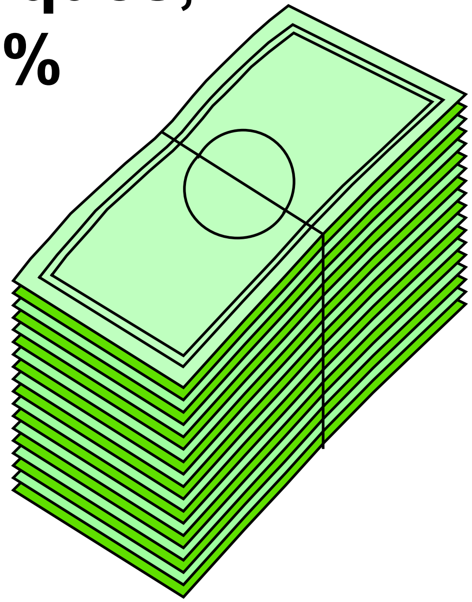
Paralyzed Veterans of America

- **KDD 1998 cup**
- **Mailing list of 3.5 million potential donors**
- **Lapsed donors**
  - Made their last donation to PVA 13 to 24 months prior to June 1997
  - 200,000 (training and test sets)
- **Who should get the current mailing?**
- **Cost effective strategy?**

Results for PVA Data Set

- If entire list (100,000 donors) are mailed, net donation is $10,500

- Using data mining techniques, this was increased 41.37%



KDD CUP 98 Results

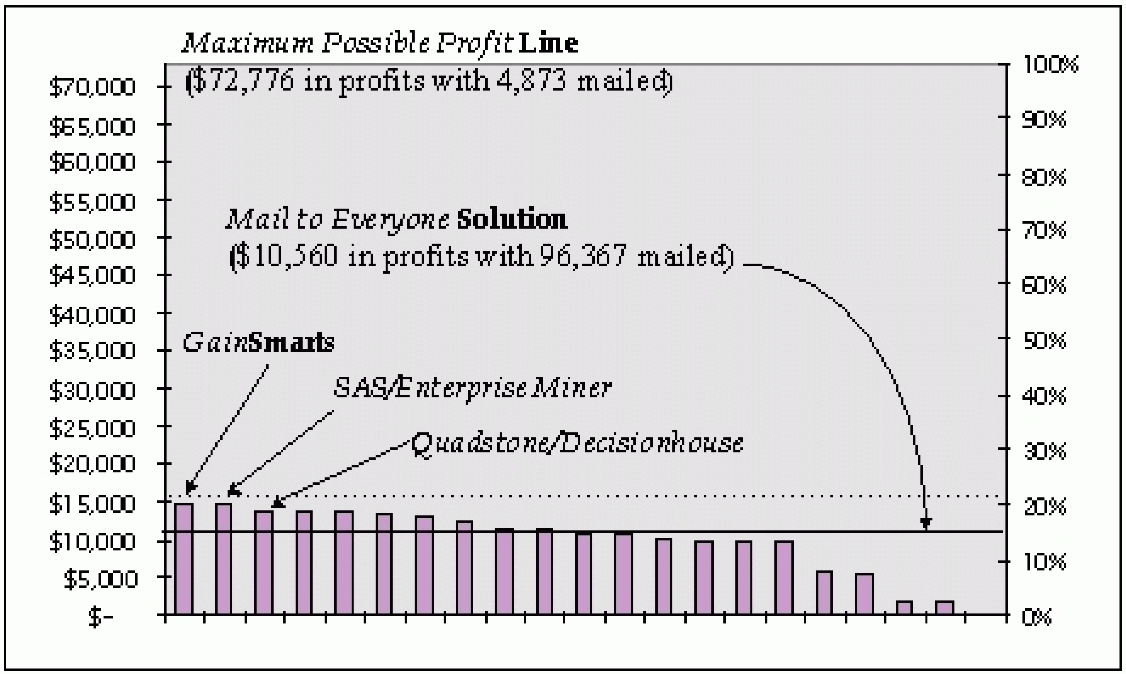
KDD-CUP-98 Results (1 of 2)

Participants	Sum of Actual Profits		Number Mailed	Average Profits
GainSmarts	$	14,712.24	56,330	0.26
SAS/Enterprise Miner	$	14,662.43	55,838	0.26
Quadstone/Decisionhouse	$	13,954.47	57,836	0.24
# 4	$	13,824.77	55,650	0.25
# 5	$	13,794.24	51,906	0.27
# 6	$	13,598.05	55,830	0.24
# 7	$	13,040.46	60,901	0.21
# 8	$	12,298.23	48,304	0.25
# 9	$	11,422.77	56,144	0.20
# 10	$	11,276.46	90,976	0.12
# 11	$	10,719.88	62,432	0.17
# 12	$	10,706.34	65,286	0.16
# 13	$	10,112.08	64,044	0.16
# 14	$	10,048.72	76,994	0.13
# 15	$	9,740.72	54,195	0.18
# 16	$	9,463.77	79,294	0.12
# 17	$	5,682.91	51,477	0.11
# 18	$	5,483.67	30,539	0.18
# 19	$	1,924.69	50,475	0.04
# 20	$	1,706.17	42,270	0.04
# 21	$	(53.68)	1,551	-0.03

Ismail Parsa KDD-CUP-98 8/98 epsilon

KDD CUP 98 Results 2

KDD-CUP-98 Results (2 of 2)



Ismail Parsa KDD-CUP-98 8/98 epsilon

Why Is This Hard?

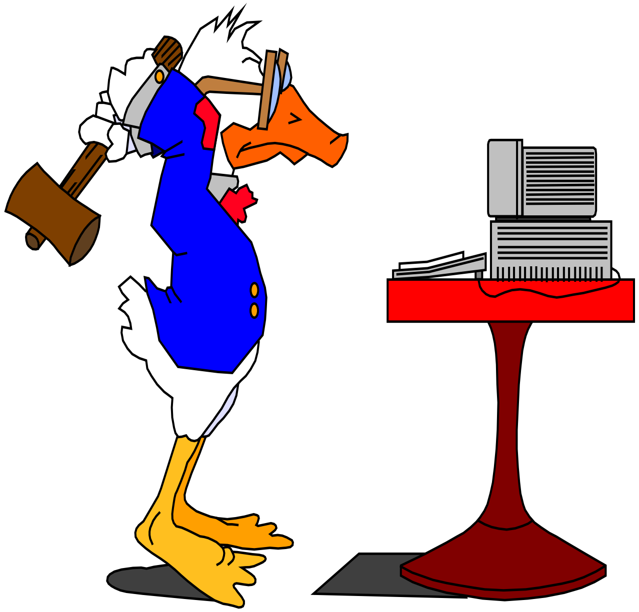
- Size of Data Set
- Signal/Noise ratio
- Example #1 – PVA on

Why Is It Taking Off Now?

- **Because we can**
  - Computer power
  - The price of digital storage is near zero
- **Data warehouses already built**
  - Companies want return on data investment

What's Different?

- **Users**
  - Domain experts, not statisticians
  - Have too much data
  - Want *automatic* methods
  - Want useful information
- **Problem size**
  - Number of rows
  - Number of variables



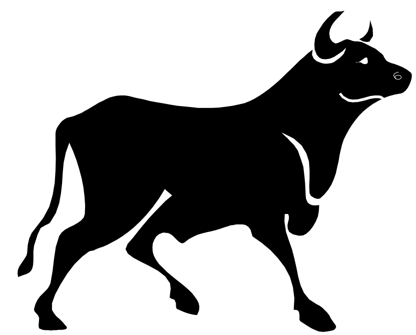
Data Mining Data Sets

- **Massive amounts of data**
- **UPS**
  - 16TB -- library of congress
  - Mostly tracking
- **Low signal to noise**
  - Many irrelevant variables
  - Subtle relationships
  - Variation

Financial Applications

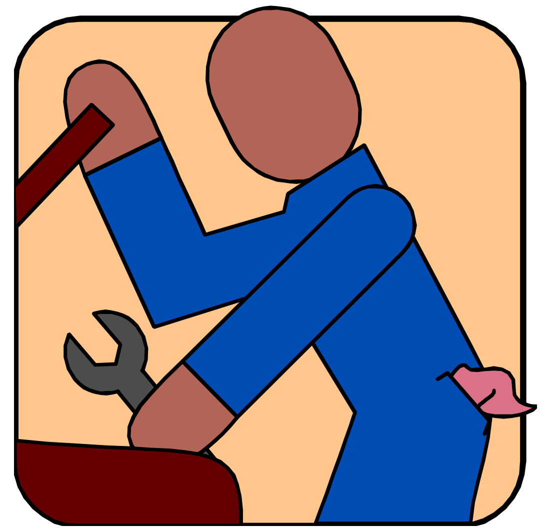
- **Credit assessment**
  - ➢ Is this loan application a good credit risk?
  - ➢ Who is likely to declare bankruptcy?

- **Financial performance**
  - ➢ What should be a portfolio product mix



Manufacturing Applications

- **Product reliability and quality control**
- **Process control**
  - What can I do to improve batch yields?
- **Warranty analysis**
  - Product problems
  - Fraud
  - Service assessment



Medical Applications

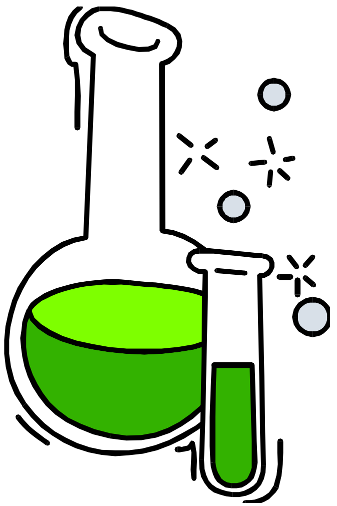
- **Medical procedure effectiveness**
  - Who are good candidates for surgery?
- **Physician effectiveness**
  - Which tests are ineffective?
  - Which physicians are likely to over-prescribe treatments?
  - What combinations of tests are most effective?

E-commerce

- Automatic web page design
- Recommendations for new purchases
- Cross selling

Pharmaceutical Applications

- **Clinical trial databases**
- **Combine clinical trial results with extensive medical/demographic data base to explore:**
  - Prediction of adverse experiences
  - Who is likely to be non-compliant or drop out?
  - What are alternative (I.E., Non-approved) uses supported by the data?



Example: Screening Plates

- **Biological assay**
  - Samples are tested for potency
  - 8 x 12 arrays of samples
  - Reference compounds included
- **Questions:**
  - Correct for drift
  - Recognize clogged dispensing tips

Pharmaceutical Applications

- **High throughput screening**
  - Predict actions in assays
  - Predict results in animals or humans
- **Rational drug design**
  - Relating chemical structure with chemical properties
  - Inverse regression to predict chemical properties from desired structure
- **DNA snips**

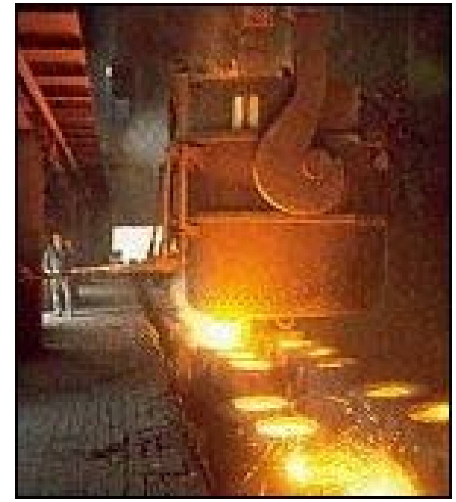
Pharmaceutical Applications

- **Genomics**
  - ➢ Associate genes with diseases
  - ➢ Find relationships between genotype and drug response (e.g., dosage requirements, adverse effects)
  - ➢ Find individuals most susceptible to placebo effect

Fraud Detection

- **Identify false:**
  - Medical insurance claims
  - Accident insurance claims
- **Which stock trades are based on insider information?**
- **Whose cell phone number has been stolen?**
- **Which credit card transactions are from stolen cards?**

Case Study I



- **Ingot cracking**
  - 953 30,000 lb. Ingots
  - 20% cracking rate
  - $30,000 per recast
  - 90 potential explanatory variables
    - ✓Water composition (reduced)
    - ✓Metal composition
    - ✓Process variables
    - ✓Other environmental variables

Case Study II – Car Insurance

- **42800 mature policies**
- **65 potential predictors**
  - Tree model found industry, vehicle age, numbers of vehicles, usage and location

Data Mining and OLAP

- **On-line analytical processing (OLAP): users deductively analyze data to verify hypothesis**
  - Descriptive, not predictive
- **Data mining: software uses data to inductively find patterns**
  - Predictive or descriptive
- **Synergy**
  - OLAP helps users understand data before mining
  - OLAP helps users evaluate significance and value of patterns

Data Mining vs. Statistics

Large amount of data:

1,000,000 rows, 3000 columns

1,000 rows, 30 columns

Data Collection

Happenstance Data

Designed Surveys, Experiments

Sample?

Why bother? We have big, parallel computers

You bet! We even get error estimates.

Reasonable Price for Sofware

$1,000,000 a year

$599 with coupon from Amstat News

Presentation Medium

PowerPoint, what else?

Overhead foils, of course!

Nice Place for a Meeting

Aspen in January, Maui in February,…

Indianapolis in August, Dallas in August, Baltimore in August, Atlanta in August,…

Data Mining Vs. Statistics

- Flexible models
- Prediction often most important
- Computation matters
- Variable selection and overfitting are problems

- Particular model and error structure
- Understanding, confidence intervals
- Computation not critical
- Variable selection and model selection are still problems

What's the Same?

- **George Box**
  - ➢All models are wrong, but some are useful
  - ➢Statisticians, like artists, have the bad habit of falling in love with their models
- **The model is no better than the data**
- **Twyman's law**
  - ➢If it looks interesting, it's probably wrong
- **De Veaux's corollary**
  - ➢If it's not wrong, it's probably obvious

Knowledge Discovery Process

- Define business problem
- Build data mining database
- Explore data
- Prepare data for modeling
- Build model
- Evaluate model
- Deploy model and results

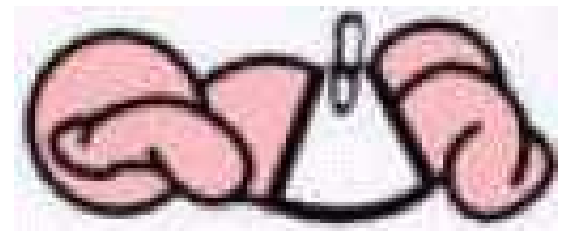
Note: This process model borrows from CRISP-DM: CRoss Industry Standard Process for Data Mining

Data Mining Myths

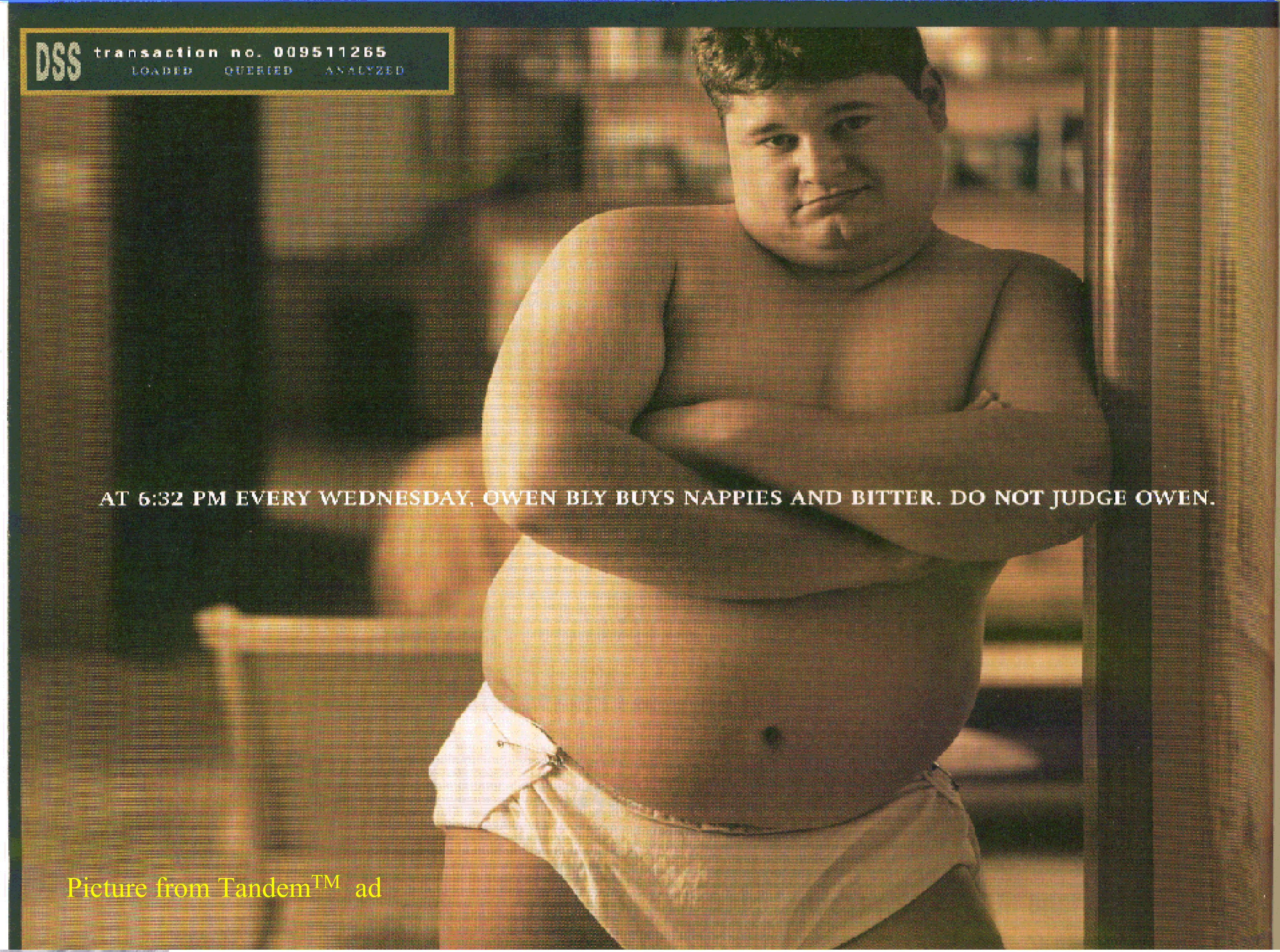
- Find answers to unasked questions
- Continuously monitor your data base for interesting patterns
- Eliminate the need to understand your business
- Eliminate the need to collect good data
- Eliminate the need to have good data analysis skills

Beer and Diapers

- Made up story?
- Unrepeatable --
  Happened once.



- Lessons learned?
- Imagine being able to see nobody coming down the road, and at such a distance
- De Veaux's theory of evolution



Picture from Tandem™ ad

Successful Data Mining

- **The keys to success:**
  - Formulating the problem
  - Using the right data
  - Flexibility in modeling
  - Acting on results
- **Success depends more on the way you mine the data rather than the specific tool**

Types of Models

- **Descriptions**
- **Classification (categorical or discrete values)**
- **Regression (continuous values)**
  - Time series (continuous values)
- **Clustering**
- **Association**

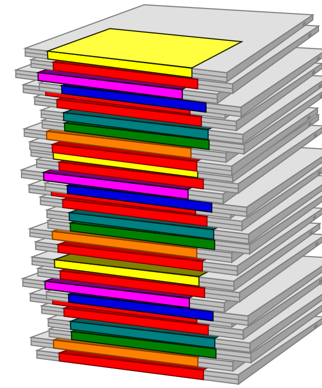
Data Preparation

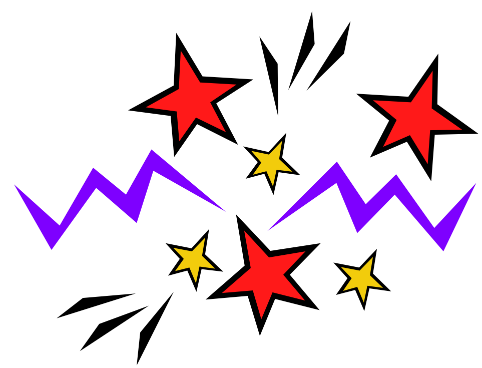
- Build data mining database
- Explore data
- Prepare data for modeling

60% to 95% of the time is spent preparing the data

Data Challenges

- **Data definitions**
  - Types of variables
- **Data consolidation**
  - Combine data from different sources
  - NASA mars lander
- **Data heterogeneity**
  - Homonyms
  - Synonyms
- **Data quality**





Data Quality

Missing Values

- **Random missing values**
  - ➢ Delete row?
    - ✓Paralyzed Veterans
  - ➢ Substitute value
    - ✓Imputation
    - ✓Multiple Imputation
- **Systematic missing data**
  - ➢ Now what?

Missing Values -- Systematic

- Ann Landers: 90% of parents said they wouldn't do it again!!
- Wharton Ph.D. Student questionnaire on survey attitudes
- Bowdoin college applicants have mean SAT verbal score above 750

The Depression Study

- **Designed to study antidepressant efficacy**
  - Measured via Hamilton Rating Scale
- **Side effects**
  - Sexual dysfunction
  - Misc safety and tolerability issues
- **Late '97 and early '98.**
- **692 patients**
- **Two antidepressants + placebo**

The Data

- **Background info**
  - Age
  - Sex
- **Each received either**
  - Placebo
  - Anti depressant 1
  - Anti depressant 2
- **Dosages**
- **At time points 7 and 14 days we also have:**
  - Depression scores
  - Sexual dysfunction indicators
  - Response indicators

Example #2

- Depression Study data
- Examine data for missing values

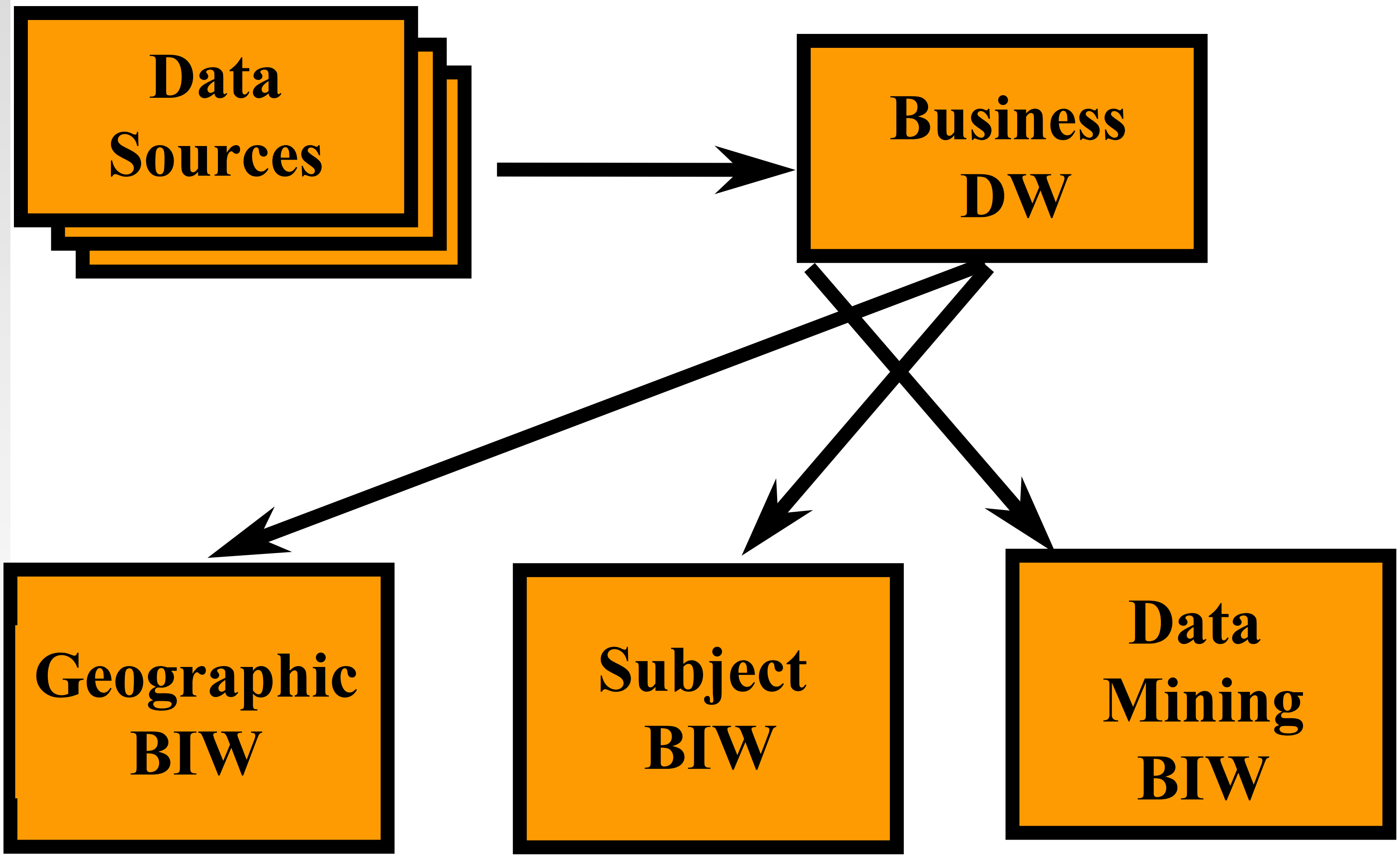
Build Data Mining Database

- Collect data
- Describe data
- Select data
- Build metadata
- Integrate data
- Clean data
- Load the data mining database
- Maintain the data mining database

Data Warehouse Architecture

- **Reference: *Data Warehouse from Architecture to Implementation* by Barry Devlin, Addison Wesley, 1997**
- **Three tier data architecture**
  - Source data
  - Business data warehouse (BDW): the reconciled data that serves as a system of record
  - Business information warehouse (BIW): the data warehouse you use

Data Mining BIW



Metadata

- **The data survey describes the data set contents and characteristics**
  - Table name
  - Description
  - Primary key/foreign key relationships
  - Collection information: how, where, conditions
  - Timeframe: daily, weekly, monthly
  - Cosynchronus: every Monday or Tuesday

Relational Data Bases

- Data are stored in tables

Items

ItemID	ItemName	price
C56621	top hat	34.95
T35691	cane	4.99
RS5292	red shoes	22.95

Shoppers

Person ID	person name	ZIPCODE	item bought
135366	Lyle	19103	T35691
135366	Lyle	19103	C56621
259835	dick	01267	RS5292

RDBMS Characteristics

- **Advantages**
  - All major DBMSs are relational
  - Flexible data structure
  - Standard language
  - Many applications can directly access RDBMSs
- **Disadvantages**
  - May be slow for data mining
  - Physical storage required
  - Database administration overhead

Data Selection

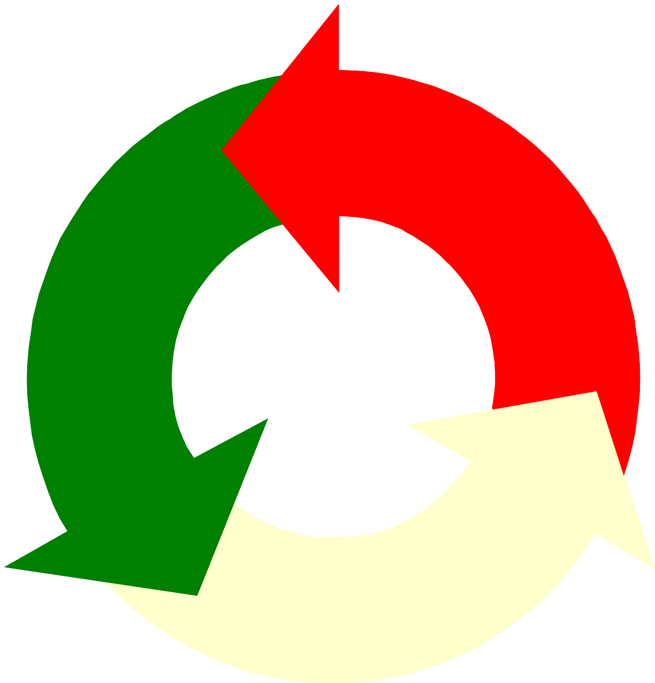
- **Compute time is determined by the number of cases (rows), the number of variables (columns), and the number of distinct values for categorical variables**
  - Reducing the number of variables
  - Sampling rows
- **Extraneous column can result in overfitting your data**
  - Employee ID is predictor of credit risk

Sampling Is Ubiquitous

- The database itself is almost certainly a sample of some population
- Most model building techniques require separating the data into training and testing samples

Model Building

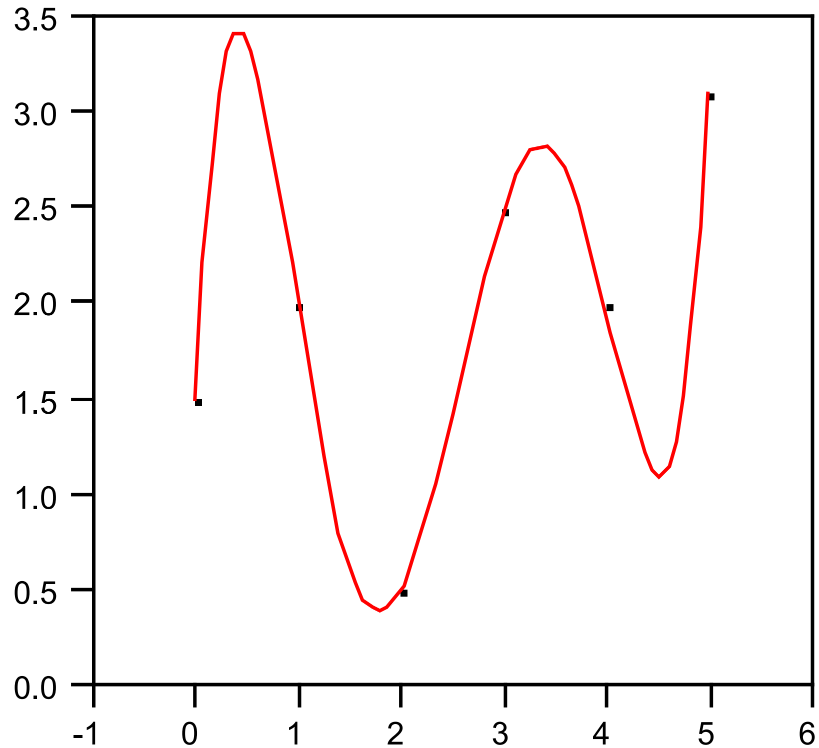
- **Model building**
  - Train
  - Test
- **Evaluate**



Overfitting in Regression

Classical overfitting:

- Fit 6th order polynomial to 6 data points



Overfitting

- **Fitting non-explanatory variables to data**
- **Overfitting is the result of**
  - ➢ Including too many predictor variables
  - ➢ Lack of regularizing the model
    - ✓ Neural net run too long
    - ✓ Decision tree too deep

Avoiding Overfitting

- **Avoiding overfitting is a balancing act**
  - Fit fewer variables rather than more
  - Have a reason for including a variable (other than it is in the database)
  - Regularize (don’t overtrain)
  - Know your field.

> **All models should be as simple as possible but no simpler than necessary**
>
> **Albert Einstein**

Evaluate the Model

- **Accuracy**
  - Error rate
  - Proportion of explained variation
- **Significance**
  - Statistical
  - Reasonableness
  - Sensitivity
  - Compute value of decisions
    - ✓ The "so what" test

Simple Validation

- *Method :* split data into a training data set and a testing data set. A third data set for validation may also be used
- *Advantages:* easy to use and understand. Good estimate of prediction error for reasonably large data sets
- *Disadvantages*: lose up to 20%-30% of data from model building

Train vs. Test Data Sets

Train

Age	Income	Job Yrs	OK
41	29,000	8	Y
32	54,000	5	Y
26	29,000	2	N

Test

Age	Income	Job Yrs	OK	Model
39	29,000	4	Y	N
29	54,000	5	Y	Y

N-fold Cross Validation

- **If you don't have a large amount of data, build a model using all the available data.**
  - What is the error rate for the model?
- **Divide the data into N equal sized groups and build a model on the data with one group left out.**

1	2	3	4	5	6	7	8	9	10

N-fold Cross Validation

- The missing group is predicted and a prediction error rate is calculated
- This is repeated for each group in turn and the average over all N repeats is used as the model error rate
- *Advantages:* good for small data sets. Uses all data to calculate prediction error rate
- *Disadvantages:* lots of computing

Regularization

- A model can be built to closely fit the training set but not the real data.
- Symptom: the errors in the training set are reduced, but increased in the test or validation sets.
- Regularization minimizes the residual sum of squares adjusted for model complexity.
- Accomplished by using a smaller decision tree or by pruning it. In neural nets, avoiding over-training.

Example #3

- **Depression Study data**
- **Fit a tree to DRP using all the variables**
  - Continue until the model won't let you fit any more
- **Predict on the test set**

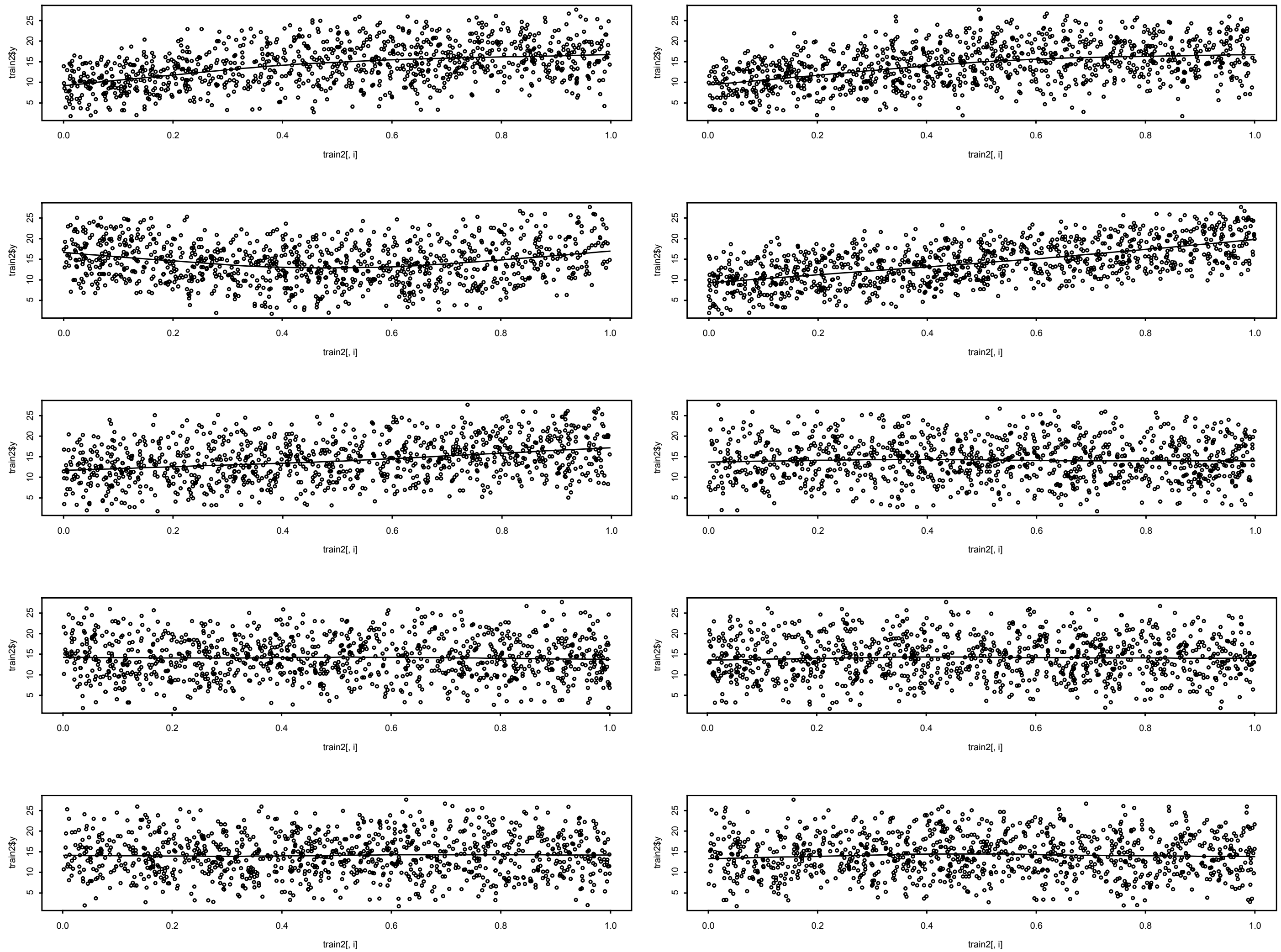
Opaque Data Mining Tools

- **Visualization**
- **Regression**
  - Logistic regression
- **Decision trees**
- **Clustering methods**

Black Box Data Mining Tools

- Neural networks
- K nearest neighbor
- K-means
- Support vector machines
- Genetic algorithms (not a modeling tool)

“Toy” Problem



Linear Regression

Term	*Estimate*	*Std Error*	*t Ratio*	*Prob>|t|*
Intercept	-0.900	0.482	-1.860	0.063
x1	4.658	0.292	15.950	<.0001
x2	4.685	0.294	15.920	<.0001
x3	-0.040	0.291	-0.140	0.892
x4	9.806	0.298	32.940	<.0001
x5	5.361	0.281	19.090	<.0001
x6	0.369	0.284	1.300	0.194
x7	0.001	0.291	0.000	0.998
x8	-0.110	0.295	-0.370	0.714
x9	0.467	0.301	1.550	0.122
x10	-0.200	0.289	-0.710	0.479

R-squared: 73.5% Train 69.4% Test

Stepwise Regression

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		-0.625	0.309	-2.019	0.0439
x1		4.619	0.289	15.998	<.0001
x2		4.665	0.292	15.984	<.0001
x4		9.824	0.296	33.176	<.0001
x5		5.366	0.28	19.145	<.0001

R-squared 73.3% on Train 69.8% Test

Stepwise 2^{ND} Order Model

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.074	0.248	-8.356	0.000
x1	4.352	0.182	23.881	0.000
x2	4.726	0.183	25.786	0.000
x3	-0.503	0.182	-2.769	0.006
(x3-0.48517)*(x3-0.48517)	20.450	0.687	29.755	0.000
x4	9.989	0.186	53.674	0.000
x5	5.185	0.176	29.528	0.000
x9	0.391	0.188	2.084	0.038
(x9-0.51161)*(x9-0.51161)	-0.783	0.743	-1.053	0.293
(x1-0.51811)*(x2-0.48354)	8.815	0.634	13.910	0.000
(x1-0.51811)*(x3-0.48517)	-1.187	0.648	-1.831	0.067
(x1-0.51811)*(x4-0.49647)	0.925	0.653	1.416	0.157
(x2-0.48354)*(x3-0.48517)	-0.626	0.634	-0.988	0.324

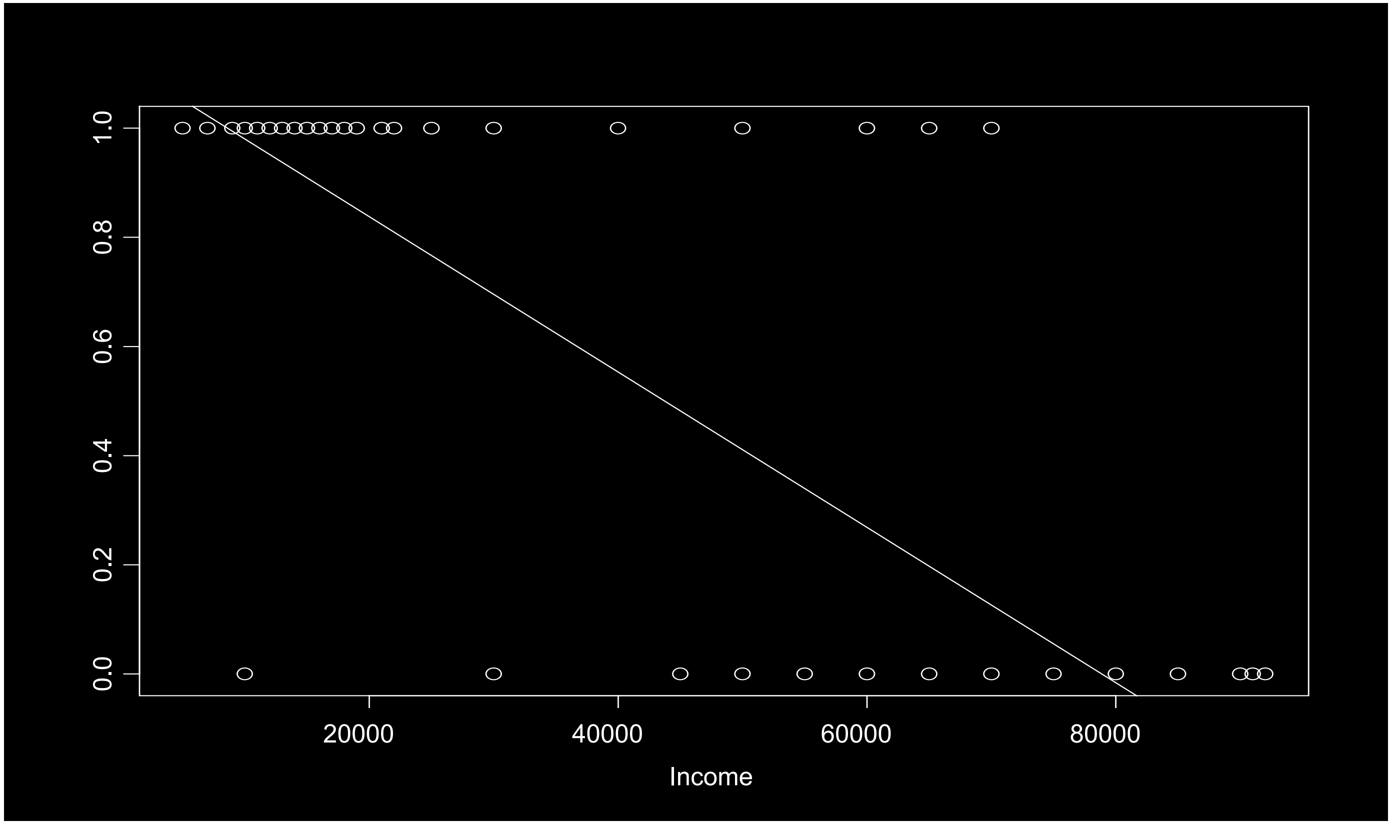
R-squared 89.7% Train 88.9% Test

Next Steps

- **Higher order terms?**
- **When to stop?**
- **Transformations?**
- **Too simple: underfitting – bias**
- **Too complex: inconsistent predictions, overfitting – high variance**
- **Selecting models is Occam's razor**
  - Keep goals of interpretation vs. prediction in mind

Logistic Regression

What happens if we use linear regression on 1-0 (yes/no) data?



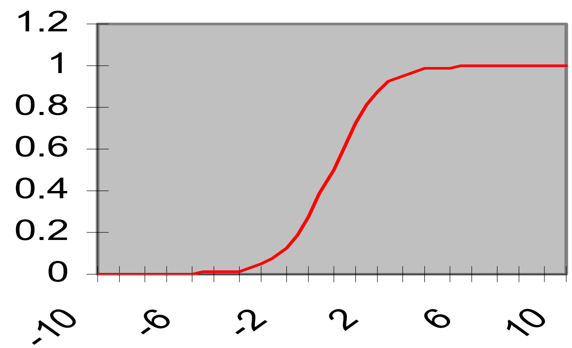
Example #4

- Depression Study data
- Fit a linear regression to DRP using HAMA 14

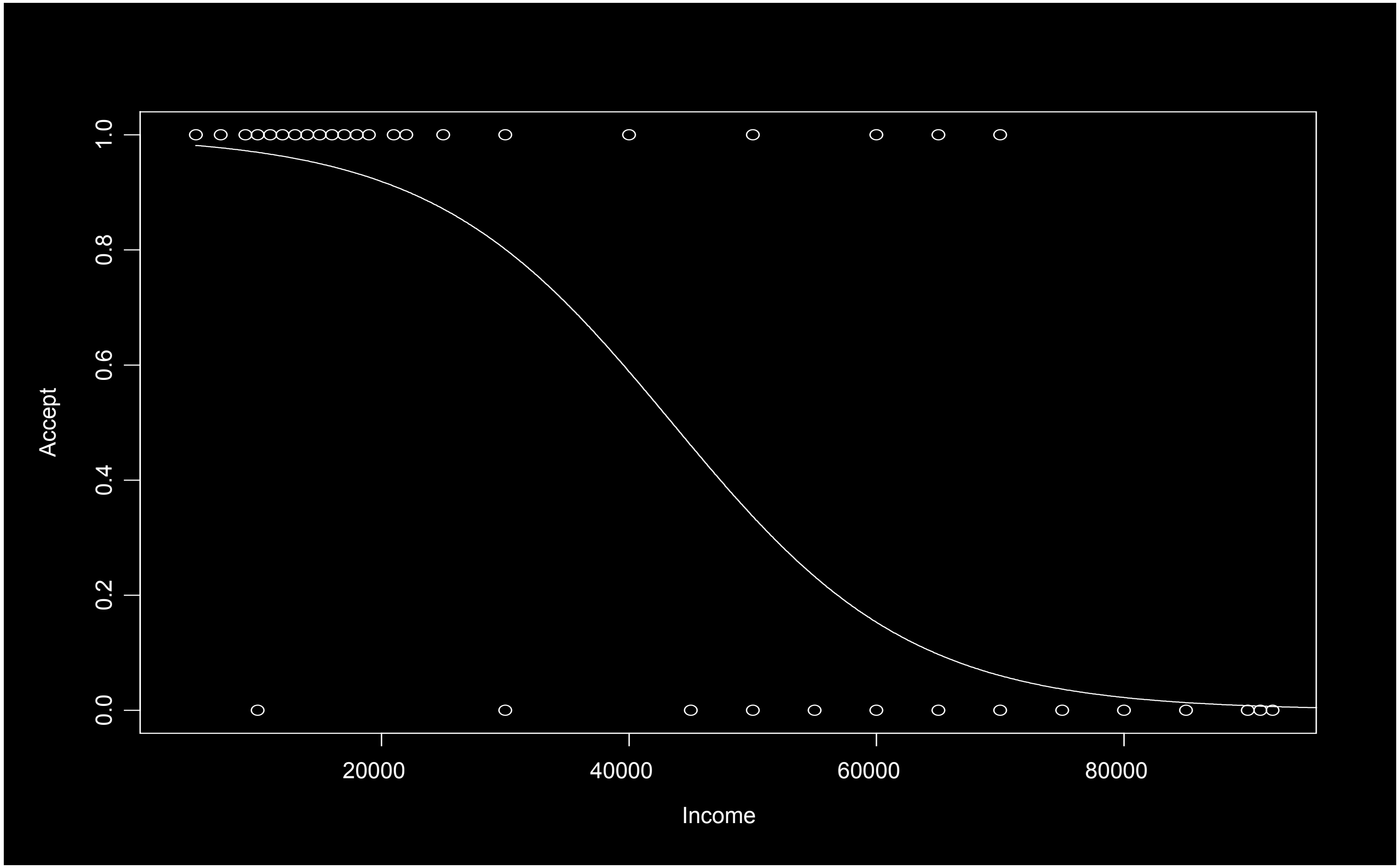
Logistic Regression II

- Points on the line can be interpreted as probability, but don't stay within [0,1]
- Use a sigmoidal function instead of linear function to fit the data

$$f(I) = \frac{1}{1+e^{-I}}$$



Logistic Regression III



Example #5

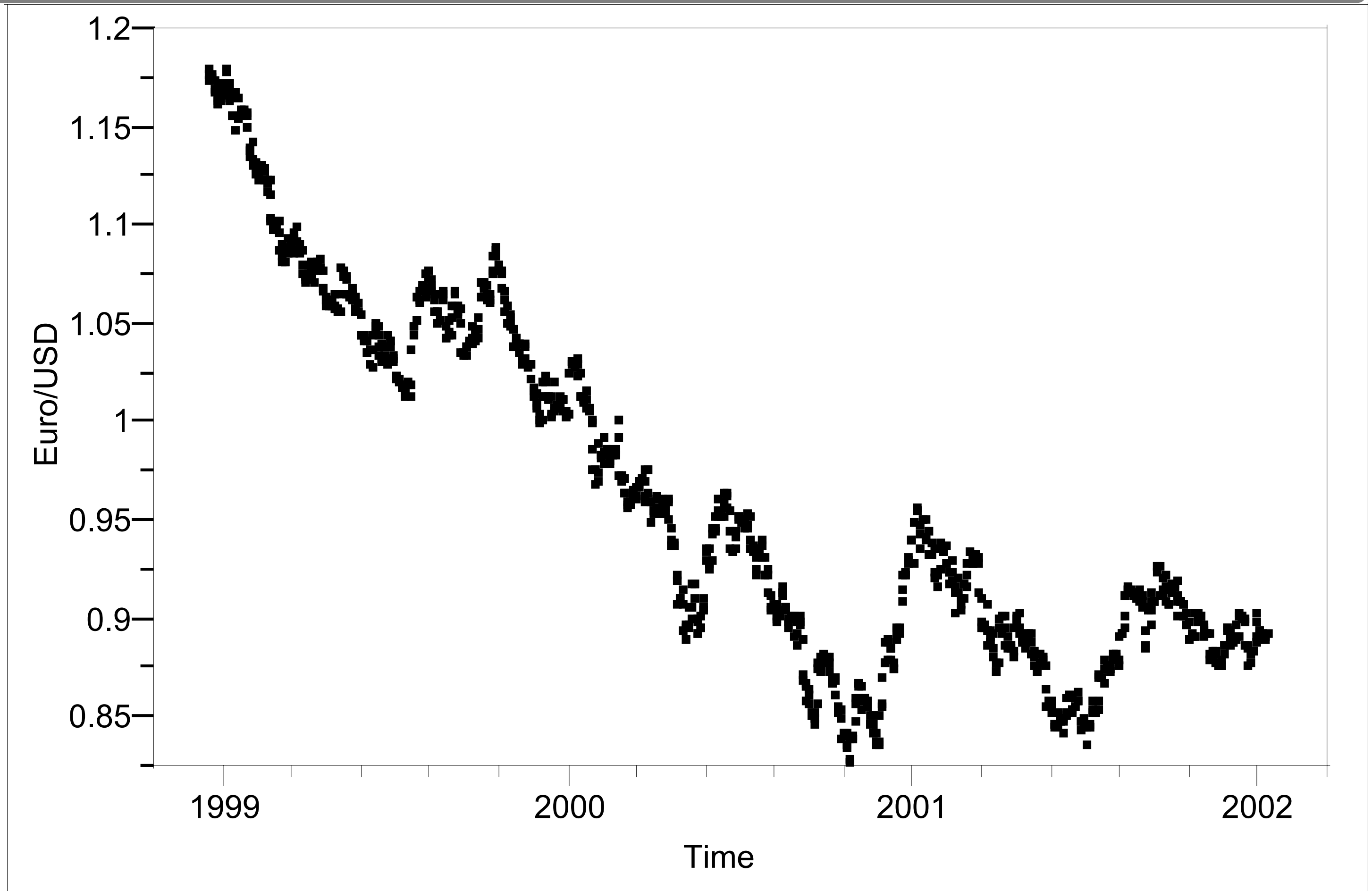
- Depression Study data
- Fit a logistic regression to DRP using HAMA 14

Regression - Summary

- Often works well
- Easy to use
- Theory gives prediction and confidence intervals
- Key is variable selection with interactions and transformations
- Use logistic regression for binary data

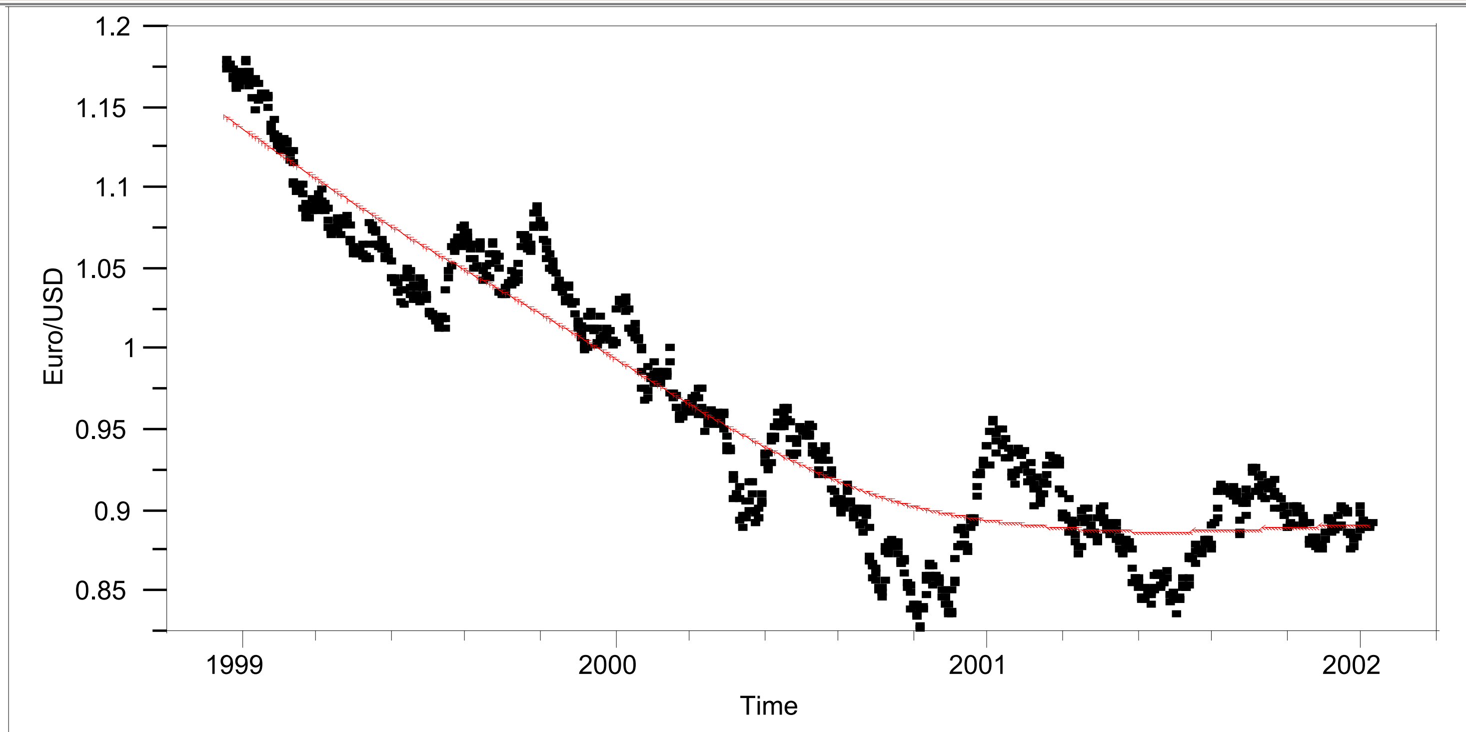
Smoothing – What's the Trend?

Bivariate Fit of Euro/USD By Time



Scatterplot Smoother

Bivariate Fit of Euro/USD By Time



Smoothing Spline Fit, lambda=10.44403

Smoothing Spline Fit, lambda=10.44403

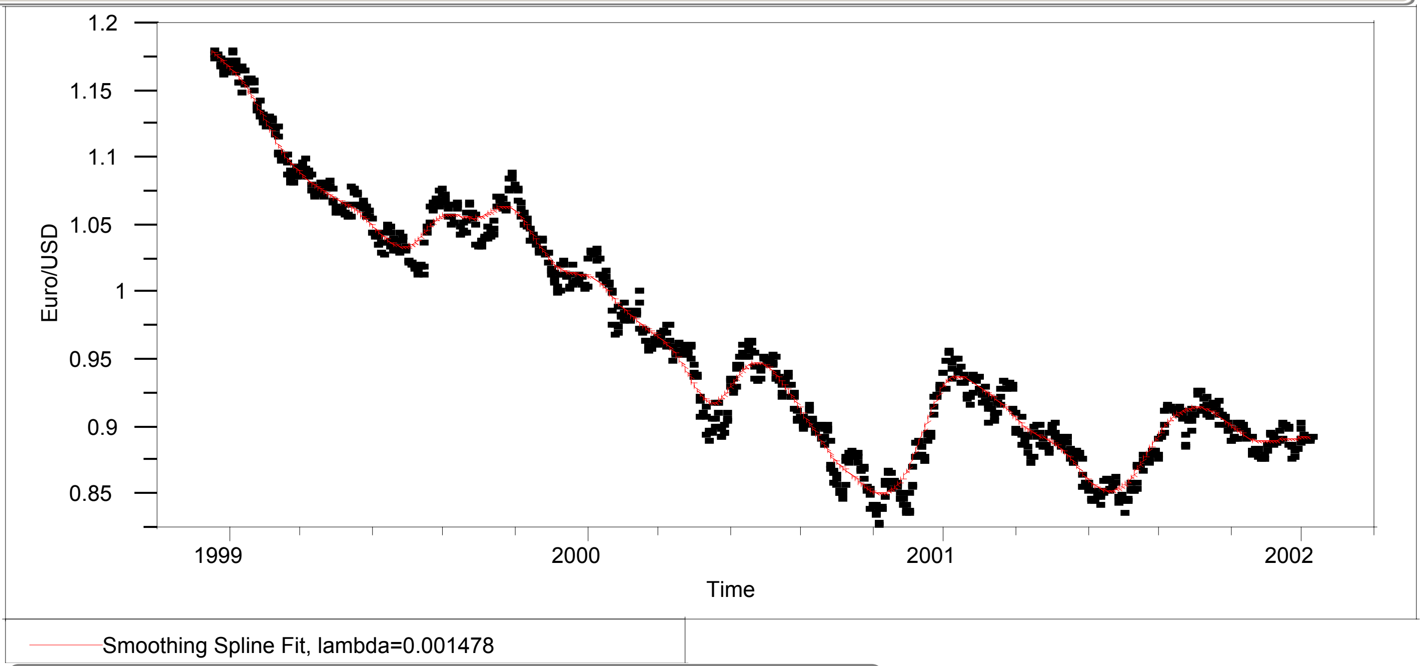
R-Square	0.90663
Sum of Squares Error	0.806737

Change Lambda:

Less Smoothing

Usually these smoothers have choices on how much smoothing

Bivariate Fit of Euro/USD By Time



Smoothing Spline Fit, lambda=0.001478

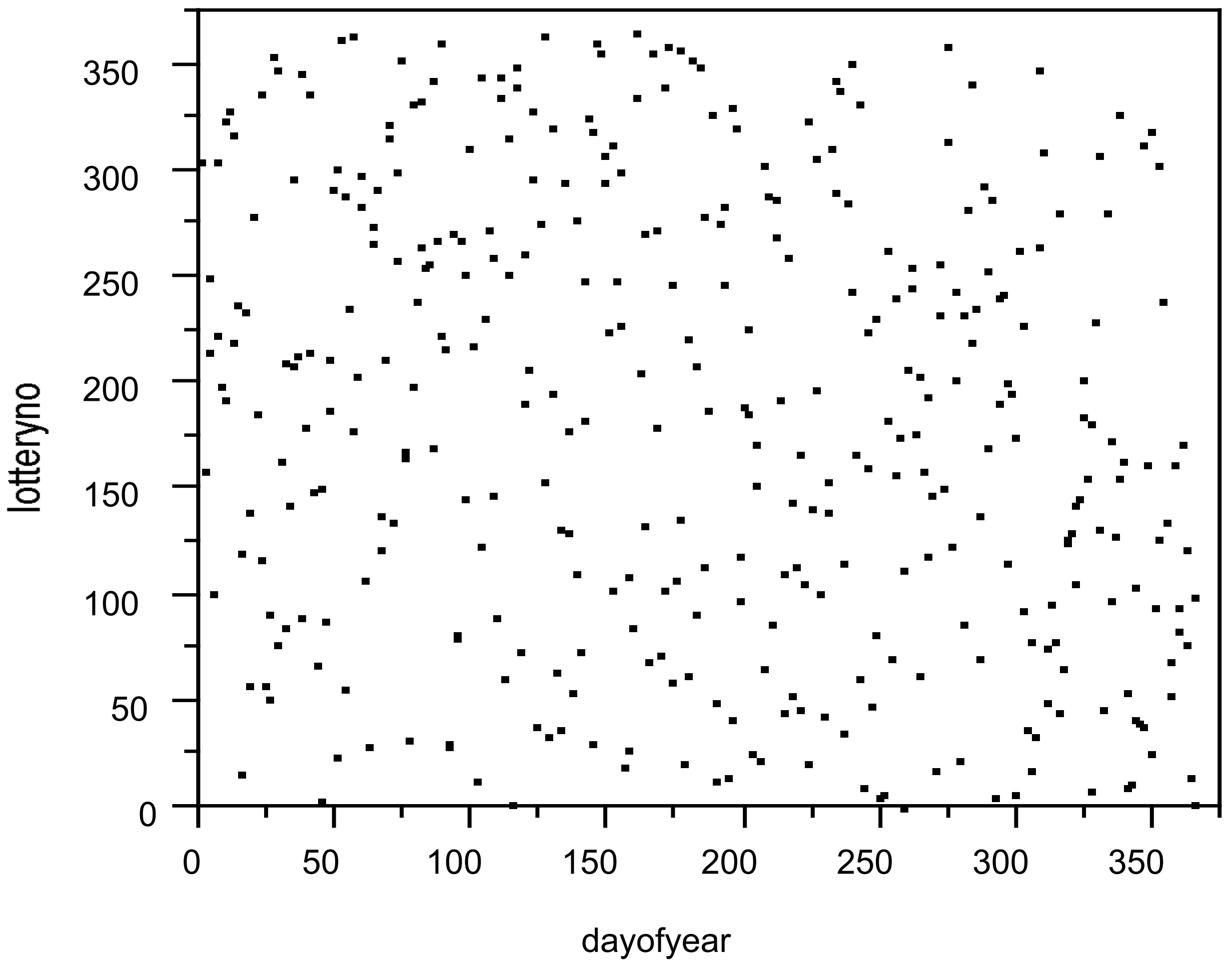
R-Square	0.986559
Sum of Squares Error	0.116135

Change Lambda:

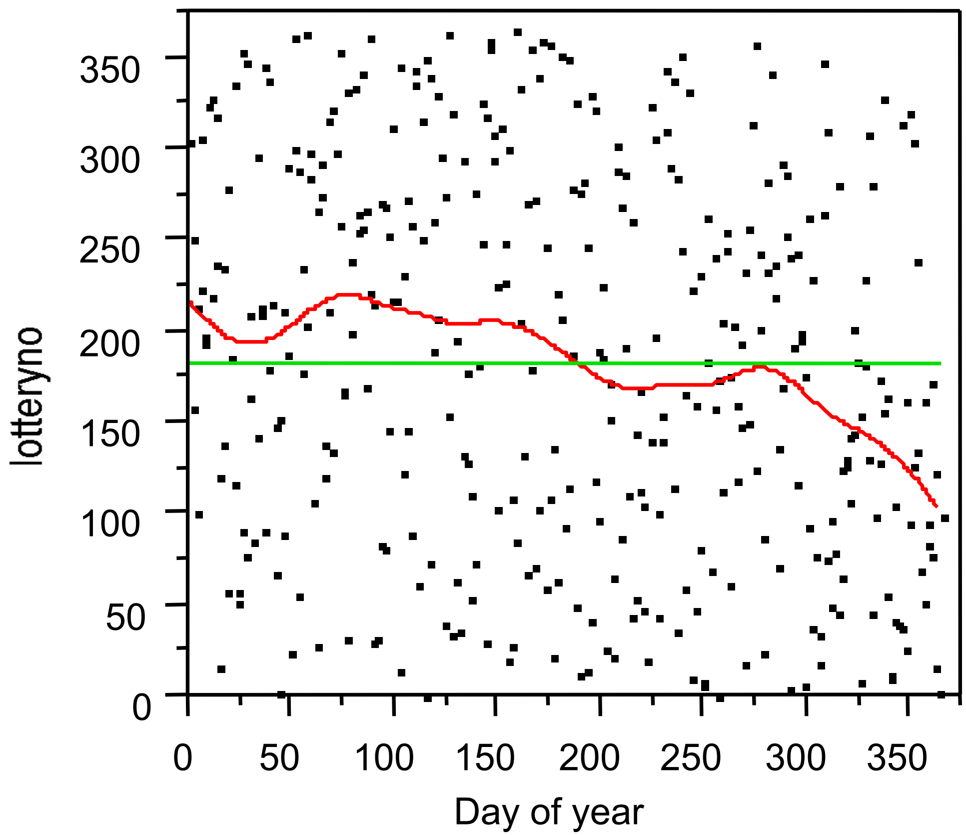
Example #6

- Fit a linear regression to Euro Rate over time
- Fit a smoothing spline

Draft Lottery 1970



Draft Data Smoothed



More Dimensions

- **Why not smooth using 10 predictors?**
  - Curse of dimensionality
  - With 10 predictors, if we use 10% of each as a neighborhood, how many points do we need to get 100 points in cube?
  - Conversely, to get 10% of the points, what percentage do we need to take of each predictor?
  - Need new approach

Additive Model

- **Cant get**

$$\hat{y} = f\,(x_1,\ldots,x_p)$$

- **So, simplify to:**

$$\hat{y} = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

- **Each of the f_i are easy to find**
  - Scatterplot smoothers

Create New Features

- **Instead of original x's use linear combinations**

$$z_i = \alpha + b_1 x_1 + \ldots + b_p x_p$$

  - Principal components
  - Factor analysis
  - Multidimensional scaling

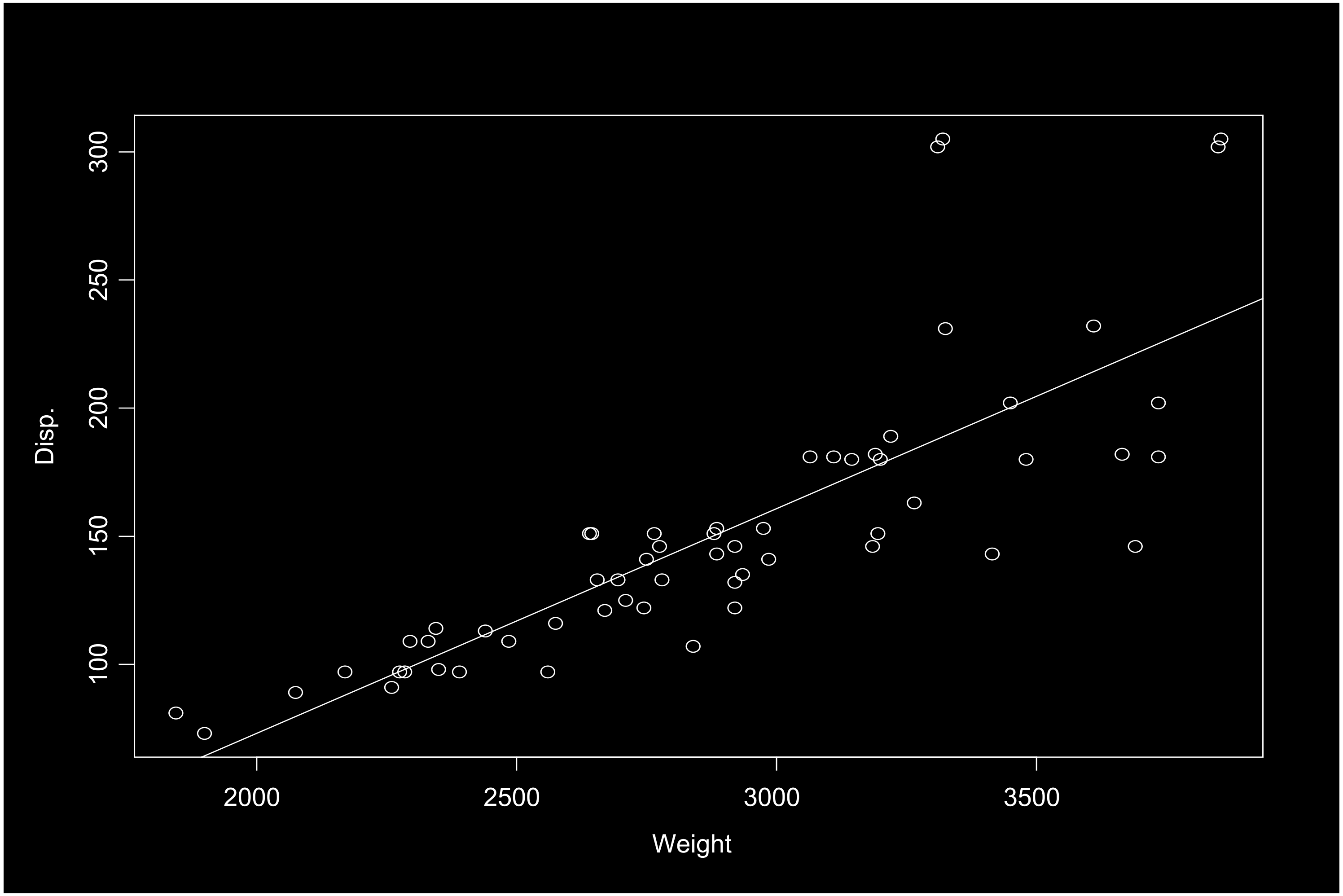
How To Find Features

- **If you have a response variable, the question may change.**
- **What are interesting directions in the predictors?**
  - High variance directions in X - PCA
  - High covariance with Y -- PLS
  - High correlation with Y -- OLS
  - Directions whose smooth is correlated with *y* - PPR

Principal Components

- **First direction has maximum variance**
- **Second direction has maximum variance of all directions perpendicular to first**
- **Repeat until there are as many directions as original variables**
- **Reduce to a smaller number**
  - Multiple approaches to eliminate directions

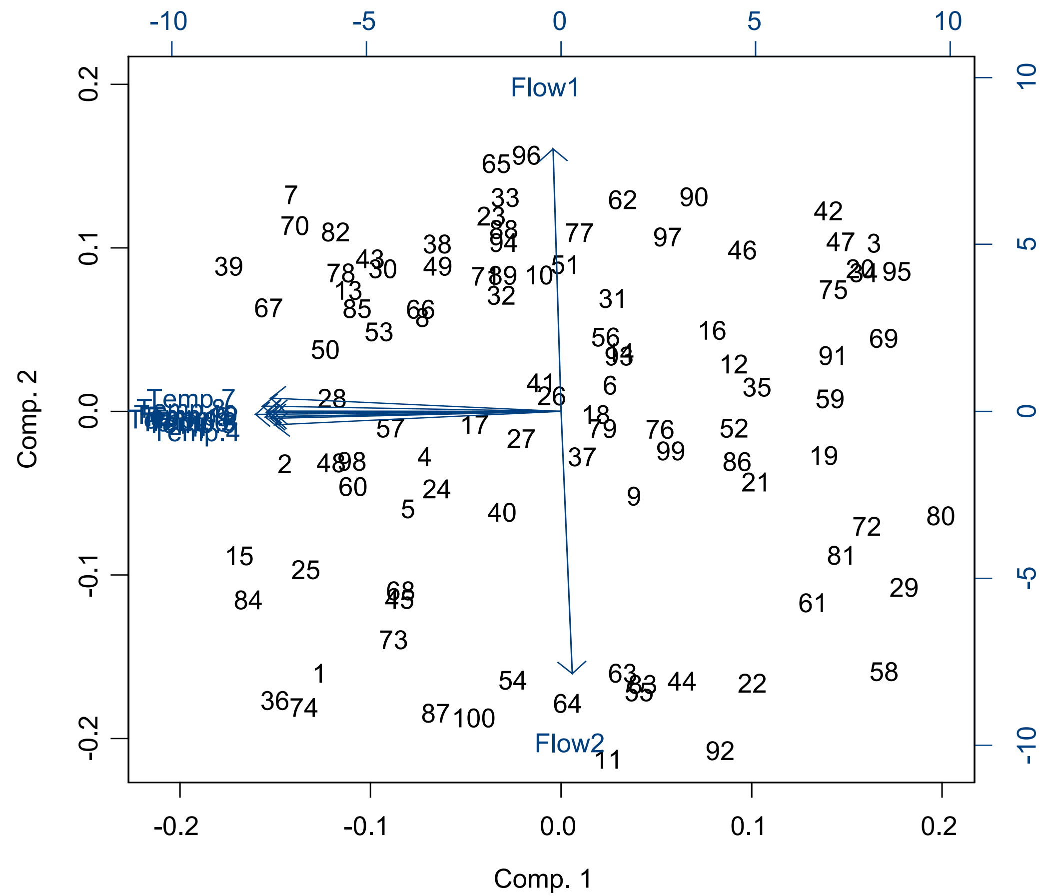
First Principal Component



When Does This Work Well?

- **When you have a group of highly correlated predictor variables**
  - Census information
  - History of past giving
  - 10 temperature sensors

Biplot



Advantages of Projection

- Interpretation
- Dimension reduction
- Able to have more predictors than observations

Disadvantages

- Lack of interpretation
- Linear

Going Non-linear

- The features are all linear:

$$\hat{y} = b_0 + b_1 z_1 + \ldots + b_k z_k$$

- But you could also use them in an *additive* model:

$$\hat{y} = f_1(z_1) + f_2(z_2) + \ldots + f_p(z_p)$$

Examples

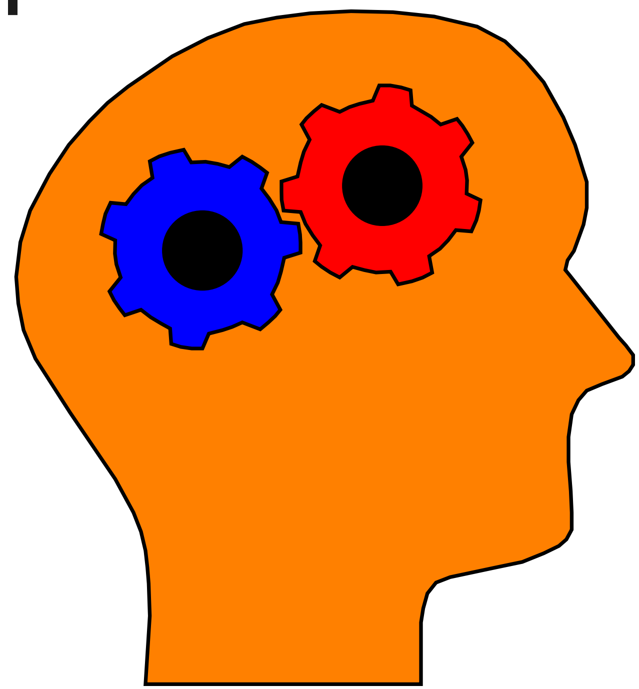
- **If the f's are arbitrary, we have projection pursuit regression**
- **If the f's are sigmoidal we have a neural network**

$$\hat{y} = \alpha + b_1 s_1(z_1) + b_2 s_2(z_2) + \ldots + b_p s_p(z_p)$$

- The z's are the hidden nodes
- The s's are the activation functions
- The b's are the weights

Neural Nets

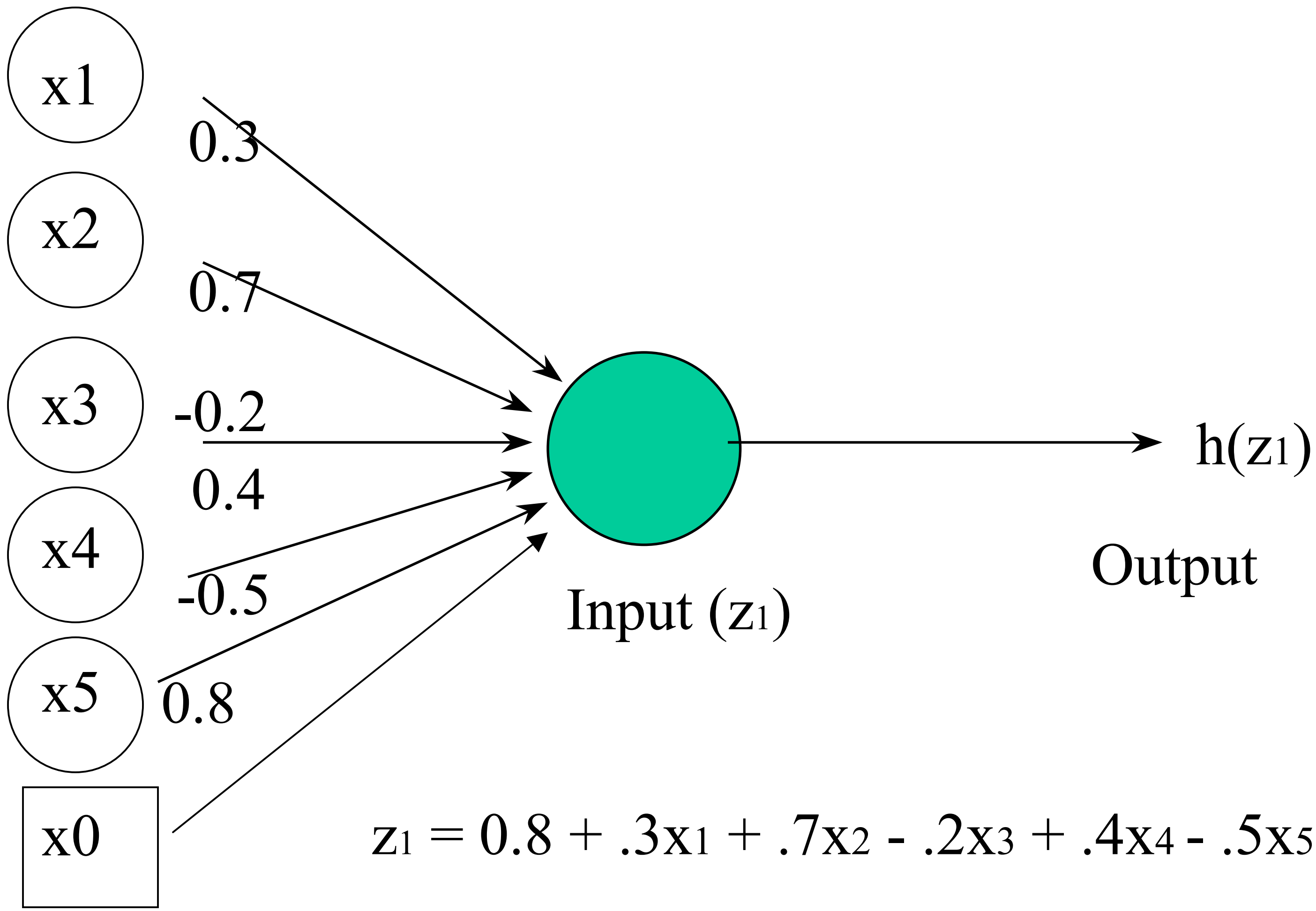
- **Don’t resemble the brain**
  - Are a statistical model
  - Closest relative is projection pursuit regression



History

- **Biology**
  - Neurode (McCulloch & Pitts, 1943)
  - Theory of learning (Hebb, 1949)
- **Computer science**
  - Perceptron (Rosenblatt, 1958)
    Adaline (Widrow, 1960)
  - *Perceptrons* (Minsky & Papert, 1969)
  - Neural nets (Rummelhart, others 1986)

A Single Neuron



$$z_1 = 0.8 + .3x_1 + .7x_2 - .2x_3 + .4x_4 - .5x_5$$

Single Node

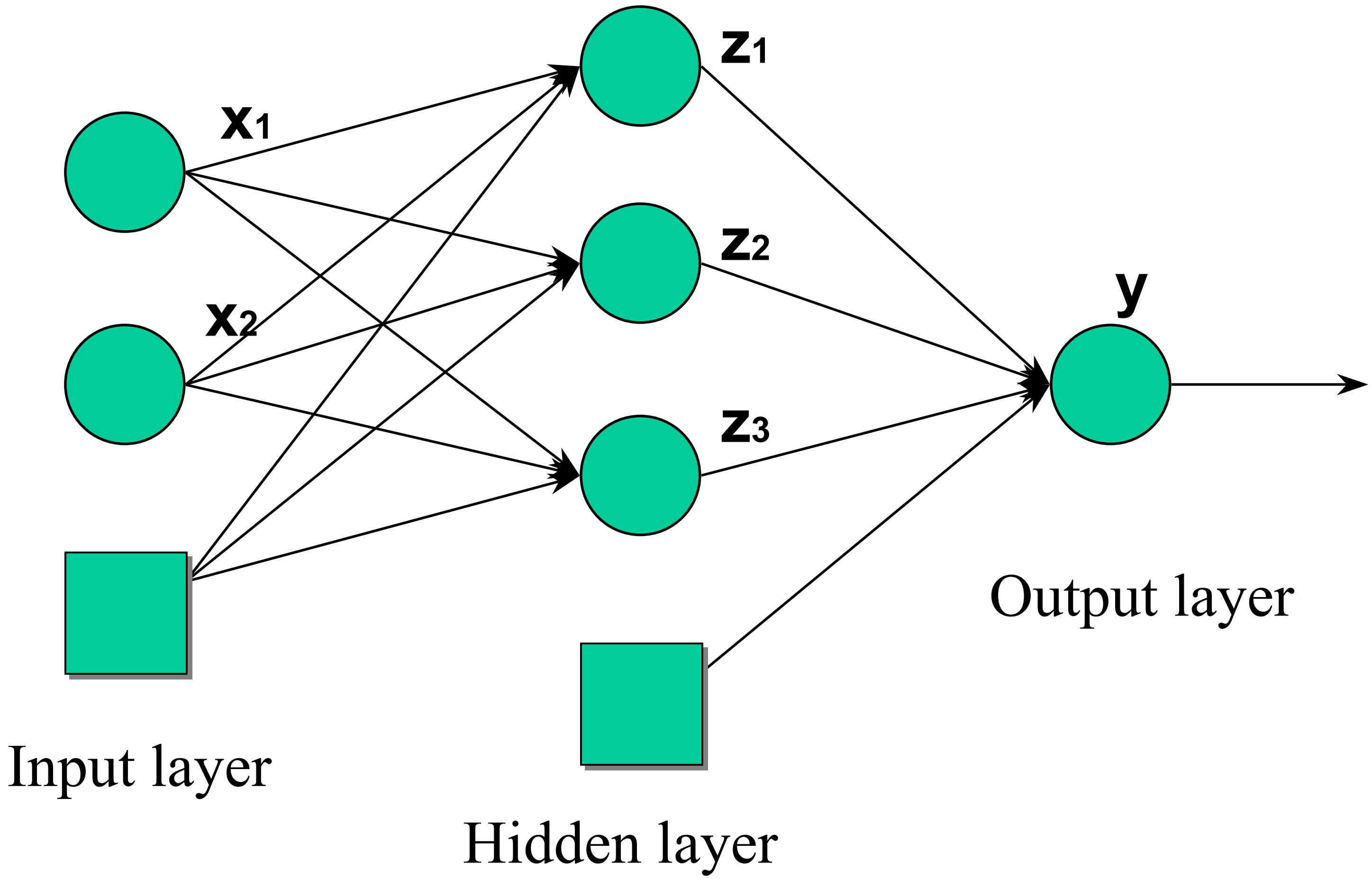
Input to outer layer from “hidden node”:

$$I = z_l = \sum_j w_{1jk} x_j + \theta_l$$

Output:

$$\hat{y}_k = h(z_{kl})$$

Layered Architecture



Neural Networks

Create lots of features – hidden nodes

$$z_l = \sum_j w_{1jk} x_j + \theta_l$$

Use them in an additive model:

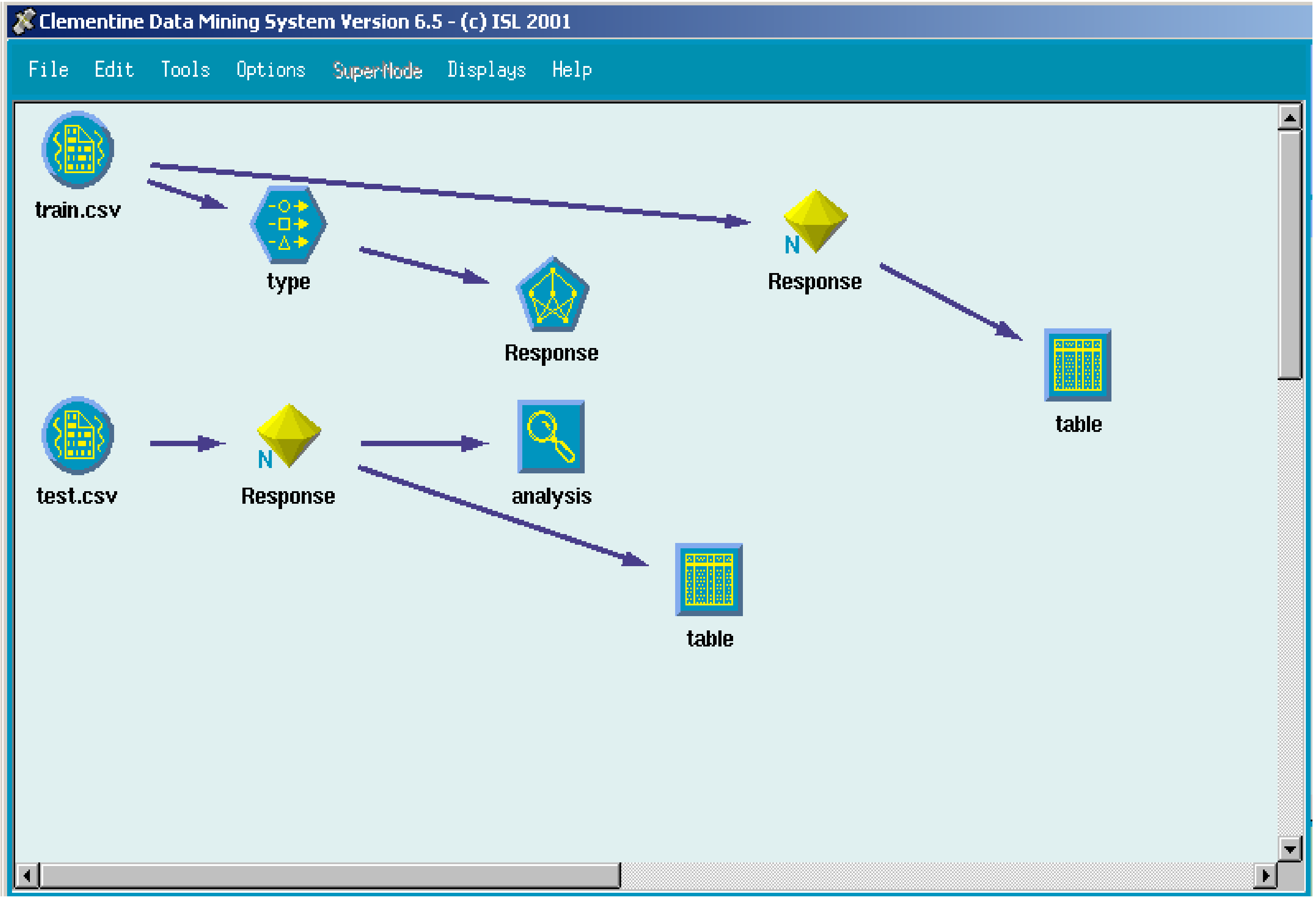
$$\hat{y}_k = w_{21}\, h(z_1) + w_{22}\, h(z_2) + \ldots + \theta_j$$

Put It Together

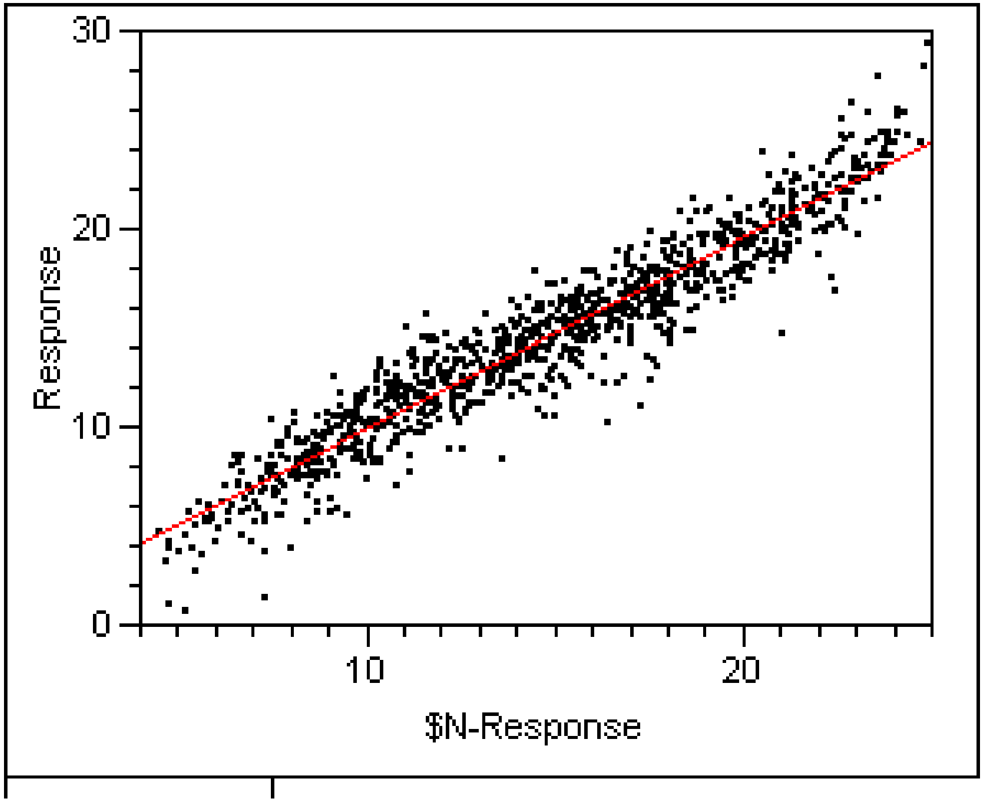
$$\hat{y}_k = \tilde{h}\Big(\sum_l w_{2kl} \;\; h\Big(\sum_j w_{1jk} x_j + \theta_l\Big) + \theta_j\Big)$$

The resulting model is just a flexible non-linear regression of the response on a set of predictor variables.

Running a Neural Net



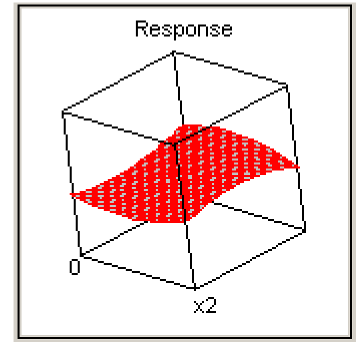
Predictions for Example

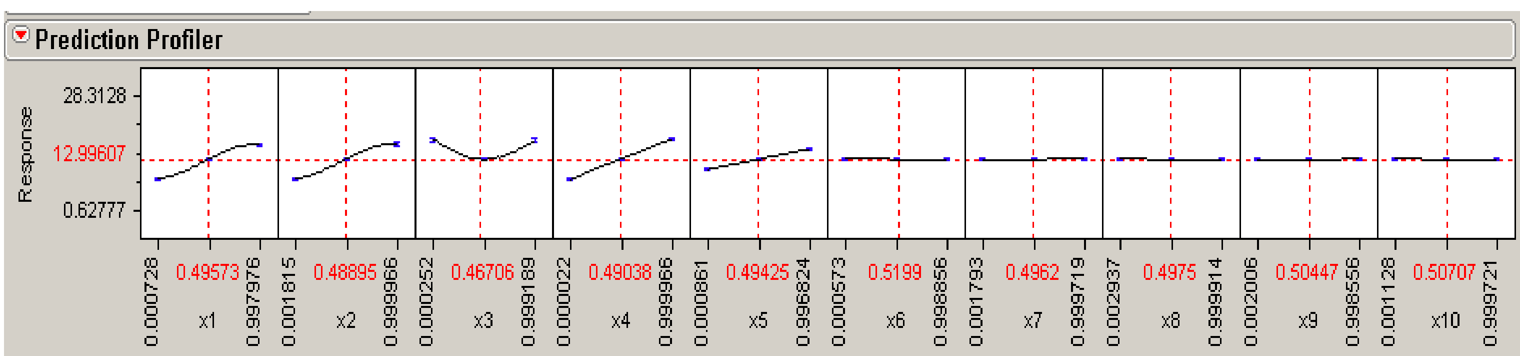


R^2 89.5% Train 87.7% Test

What Does This Get Us?

- **Enormous flexibility**
- **Ability to fit anything**
  - Including noise
  - Not just the elephant – the whole herd!
- **Interpretation?**





Example #7

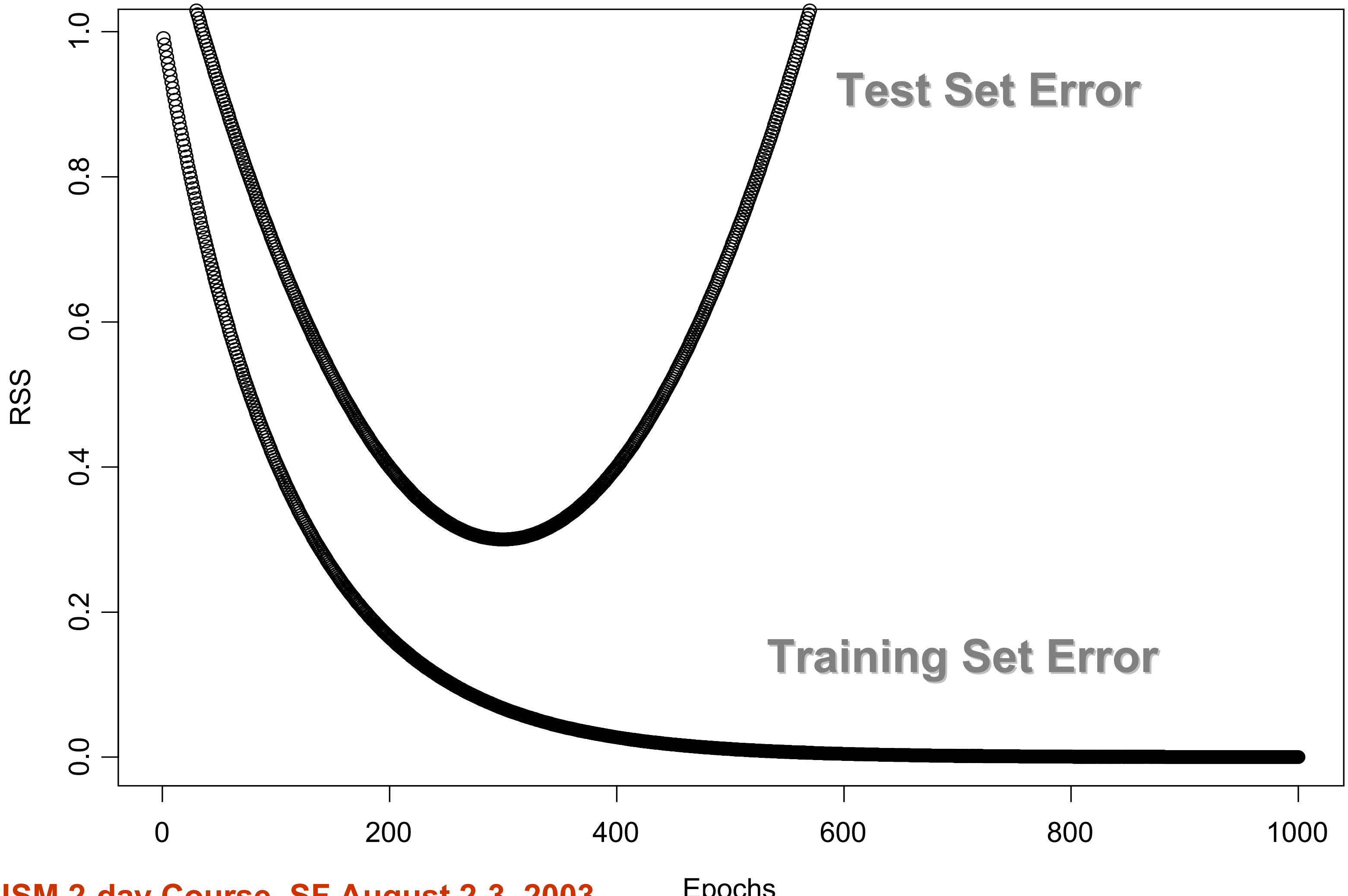
- Fit a neural net to the “toy problem” data.
- Look at the profiler.
- How does it differ from the full regression model?

Running a Training Session

- **Initialize weights**
  - Set range
  - Random initialization
  - With weights from previous training session
- **An *epoch* is one time through every row in data set**
  - Can be in random order or fixed order

Training the Neural Net

Error as a function of training



Stopping Rules

- **Error (RMS) threshold**
- **Limit -- early stopping rule**
  - Time
  - Epochs
- **Error rate of change threshold**
  - E.G. No change in RMS error in 100 epochs
- **Minimize error + complexity -- weight decay**
  - De Veaux et al *Technometrics 1998*

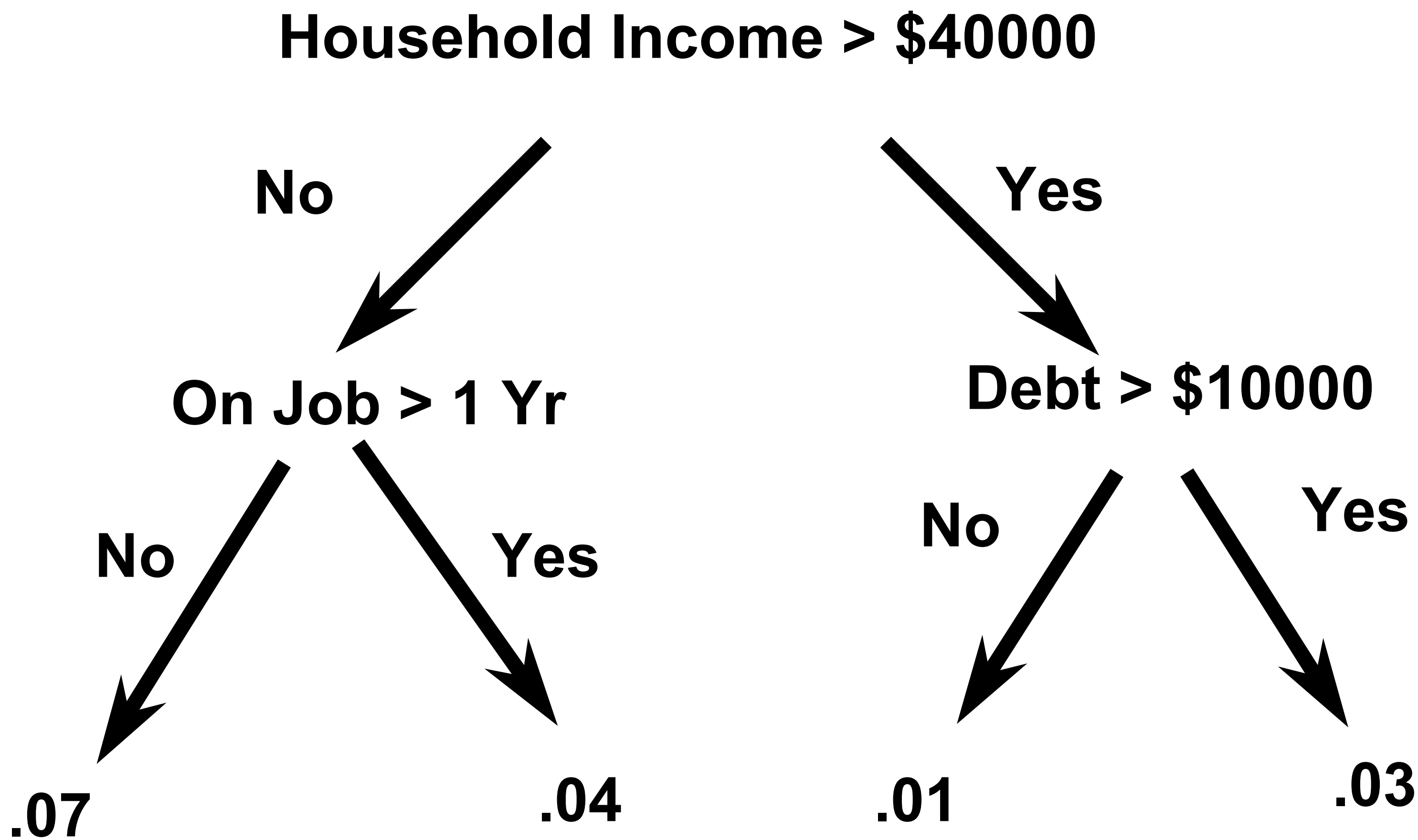
Neural Net Pro

- **Advantages**
  - ➢ Handles continuous or discrete values
  - ➢ Complex interactions
  - ➢ In general, highly accurate for fitting due to flexibility of model
  - ➢ Can incorporate known relationships
    - ✓So called grey box models
    - ✓See De Veaux et al, *Environmetrics* 1999

Neural Net Con

- **Disadvantages**
  - Model is not descriptive (black box)
  - Difficult, complex architectures
  - Slow model building
  - Categorical data explosion
  - Sensitive to input variable selection

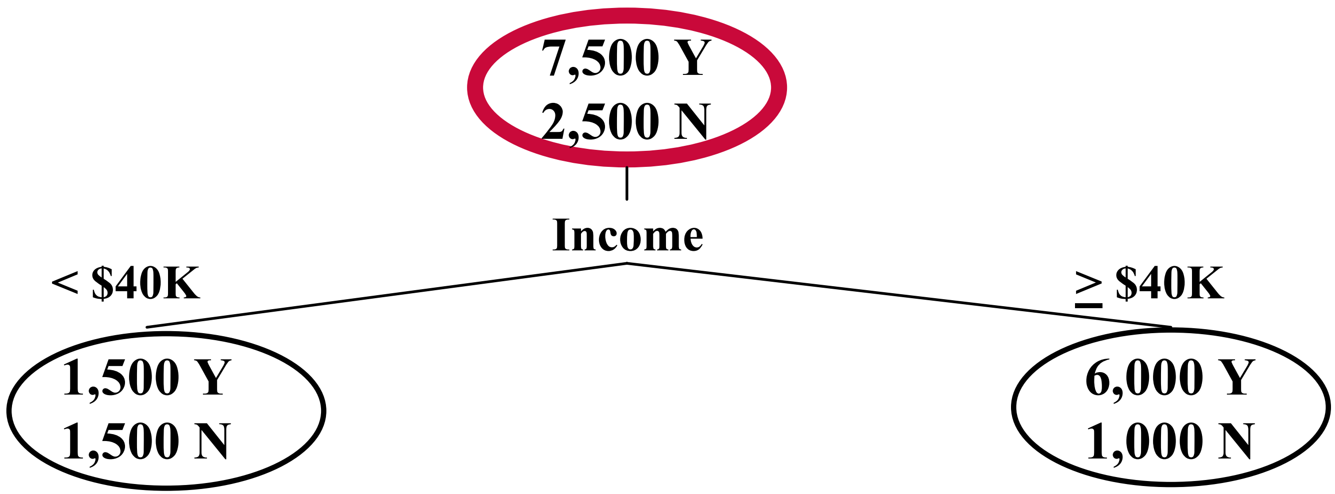
Decision Trees



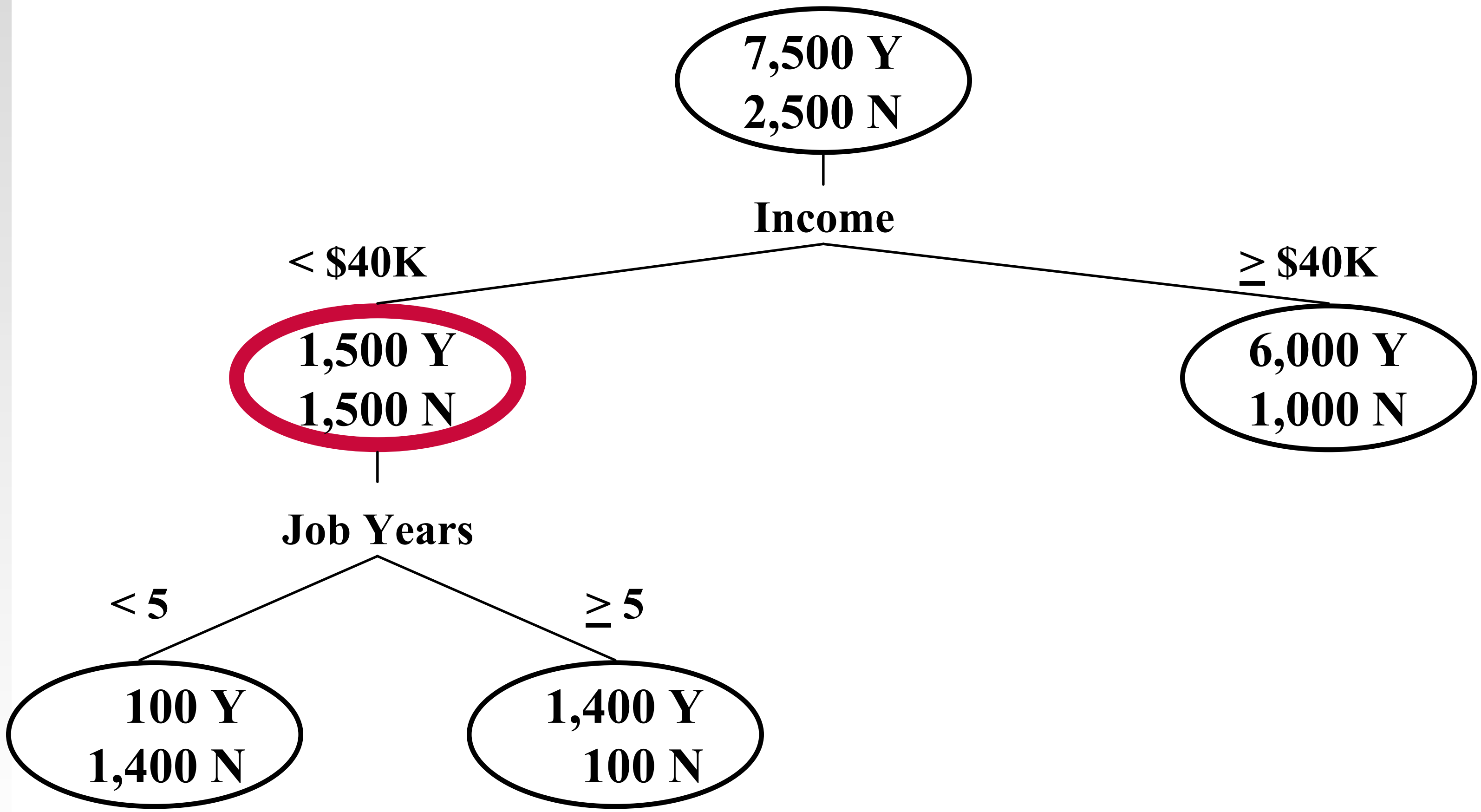
Determining Credit Risk

- **11,000 cases of loan history**
  - 10,000 cases in training set
    - 7,500 good risks
    - 2,500 bad risks
  - 1,000 cases in test set
- **Data available**
  - predictor variables
    - Income: continuous
    - Years at job: continuous
    - Debt: categorical (High, Low)
  - response variable
    - Good risk: categorical (Yes, No)

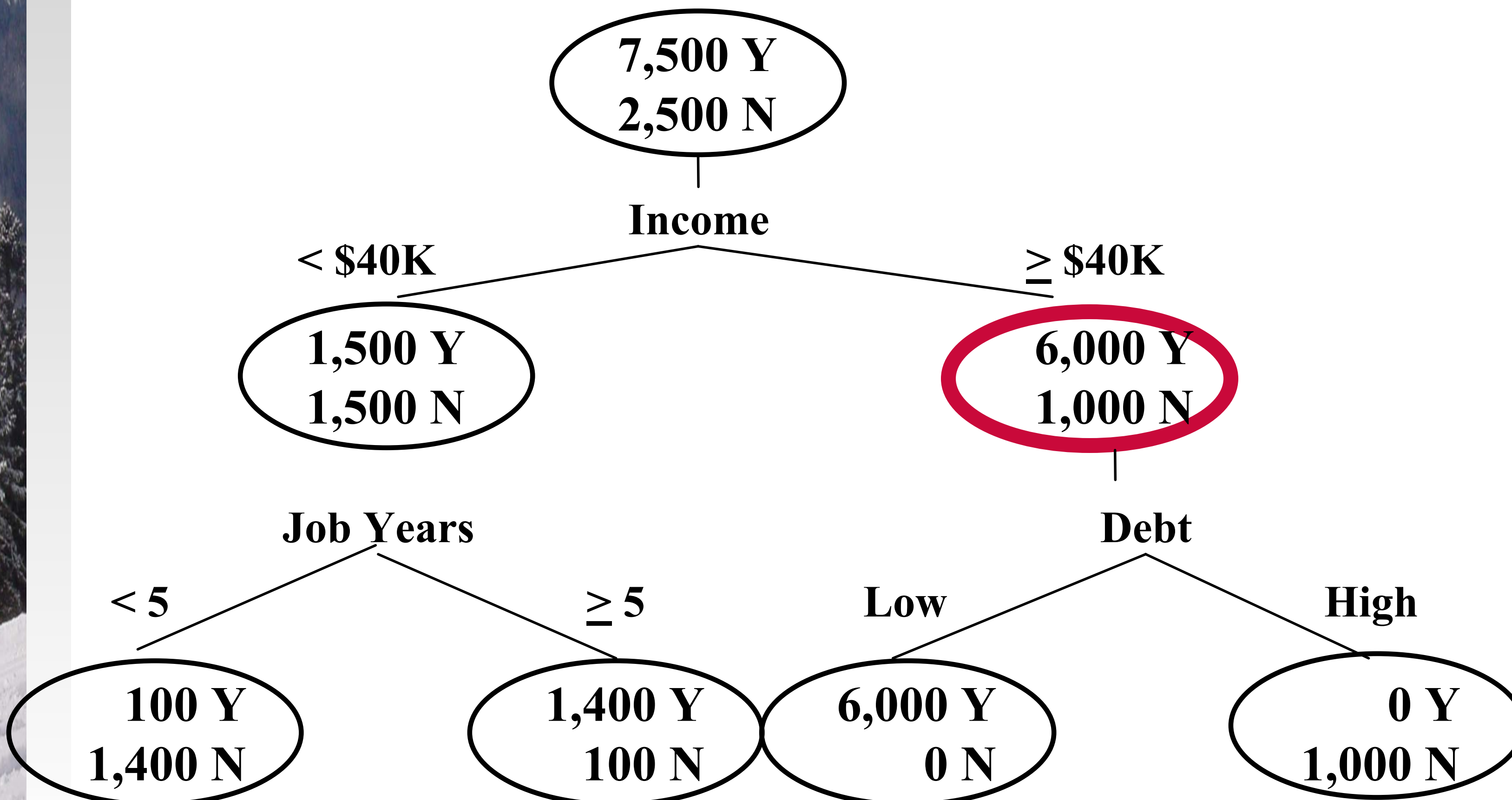
Find the First Split



Find an Unsplit Node and Split It



Find an Unsplit Node and Split It



Class Assignment

- The tree is applied to new data to classify it
- A case or instance will be assigned to the largest (or *modal*) class in the leaf to which it goes
- Example:

  1,400 Y
  100 N

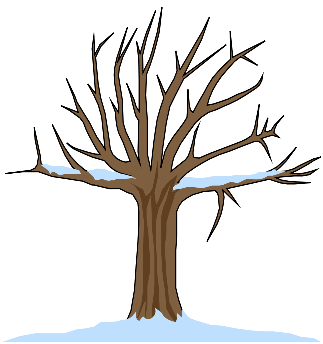
- All cases arriving at this node would be given a value of “yes”

Tree Algorithms

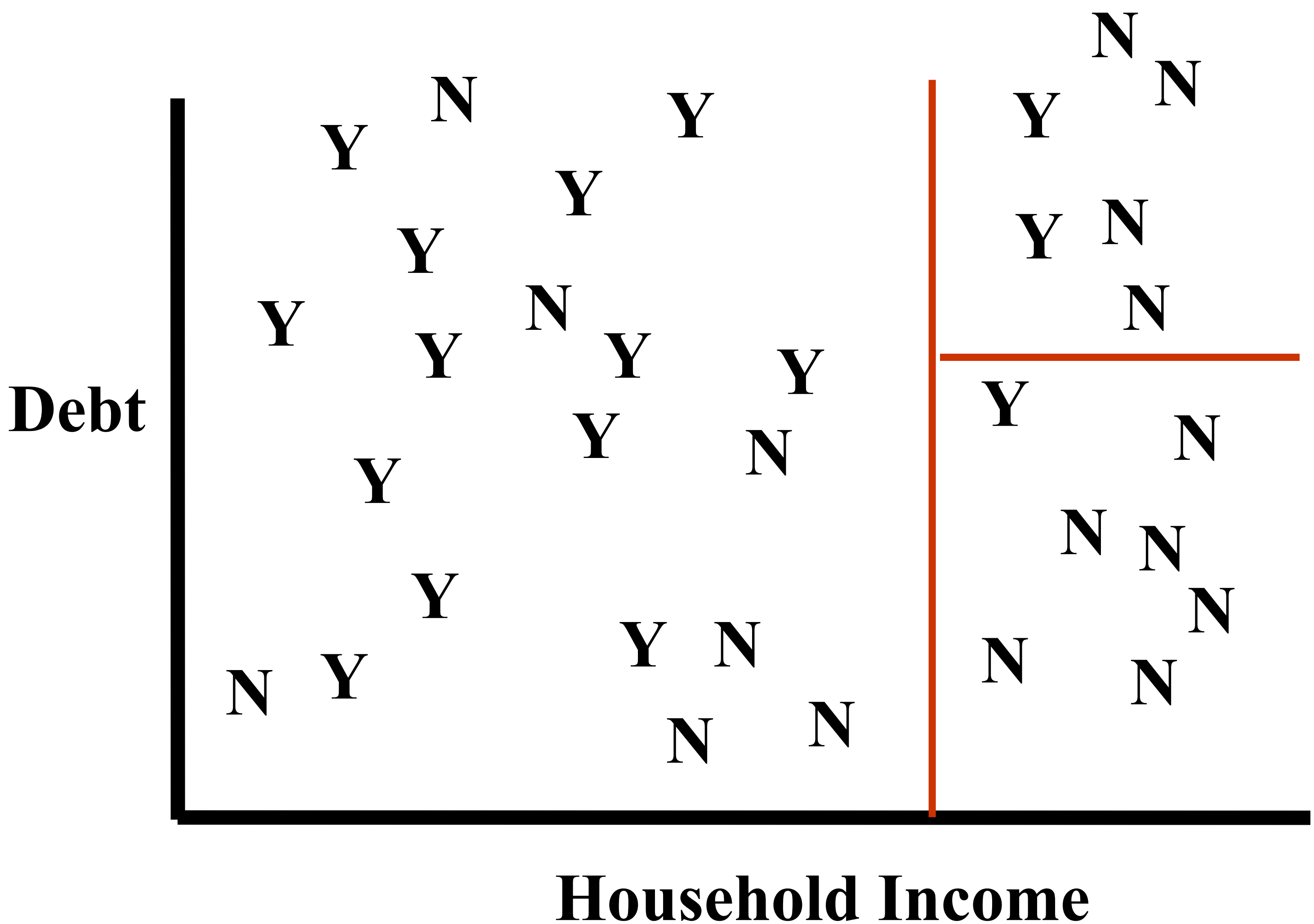
- CART (Breiman, Friedman, Olshen, stone)
- C4.5, C5.0, cubist (Quinlan)
- CHAID
- Slip (IBM)
- Quest (SPSS)

Decision Trees

- Find split in predictor variable that best splits data into heterogeneous groups
- Build the tree inductively basing future splits on past choices (greedy algorithm)
- Classification trees (categorical response)
- Regression tree (continuous response)
- Size of tree often determined by cross-validation



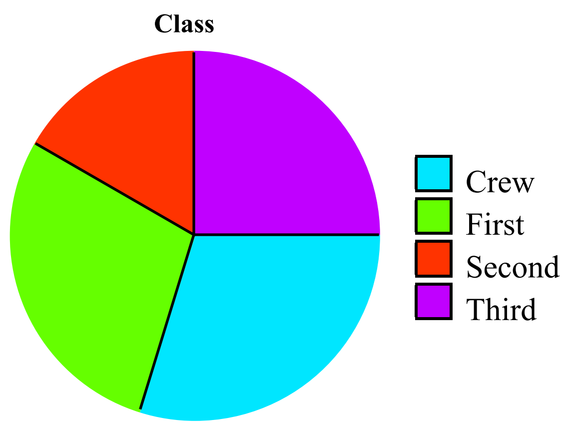
Geometry of Decision Trees



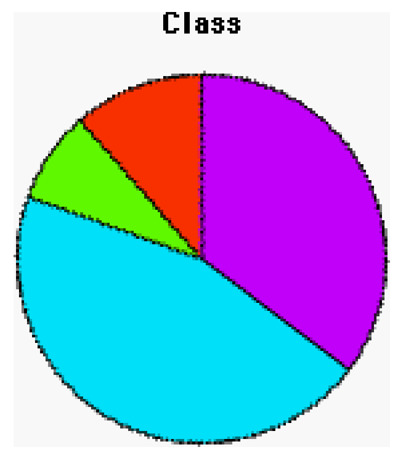
Two Way Tables -- Titanic

		Ticket Class				
		Crew	First	Second	Third	Total
	Lived	212	202	118	178	**710**
Survival	Died	673	123	167	528	**1491**
	Total	**885**	**325**	**285**	**706**	**2201**

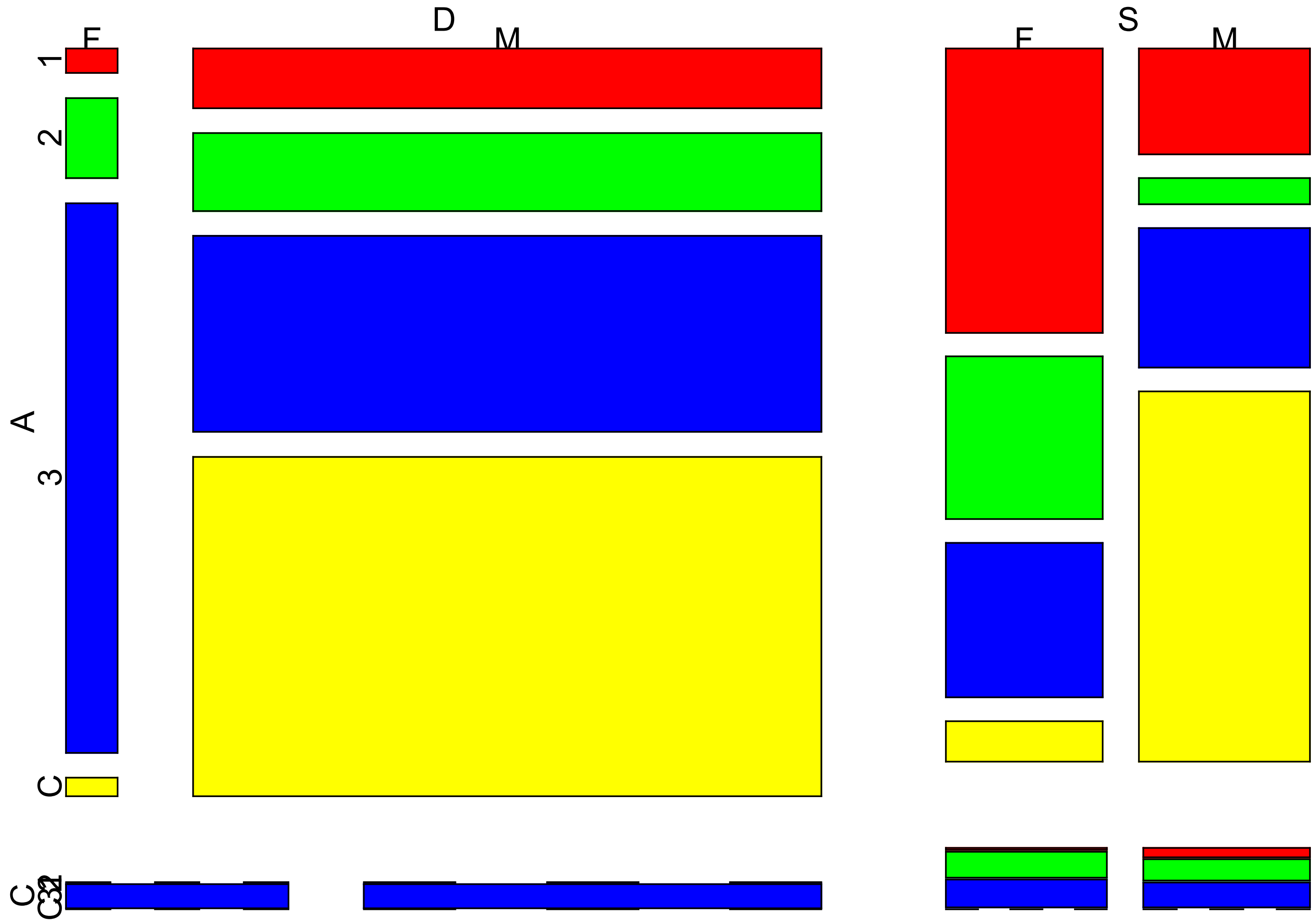
Survivors



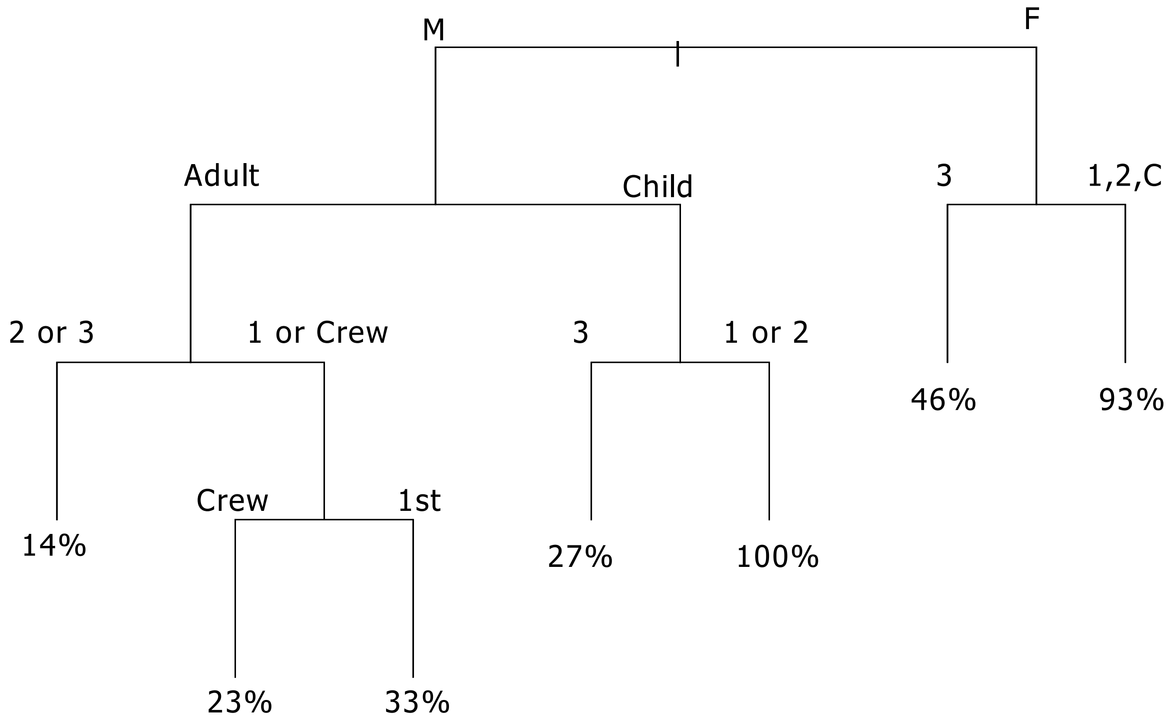
Non-Survivors



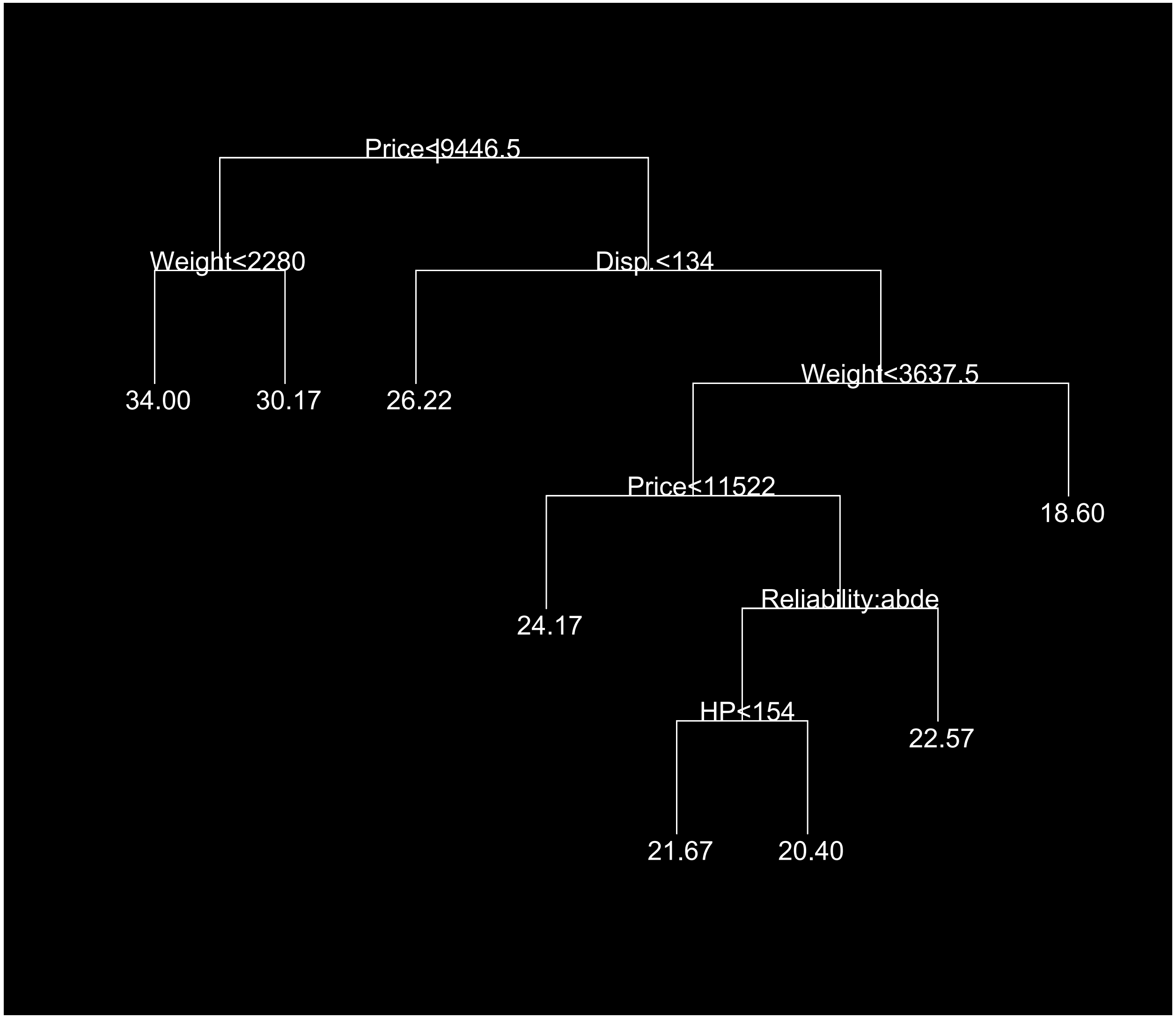
Mosaic Plot



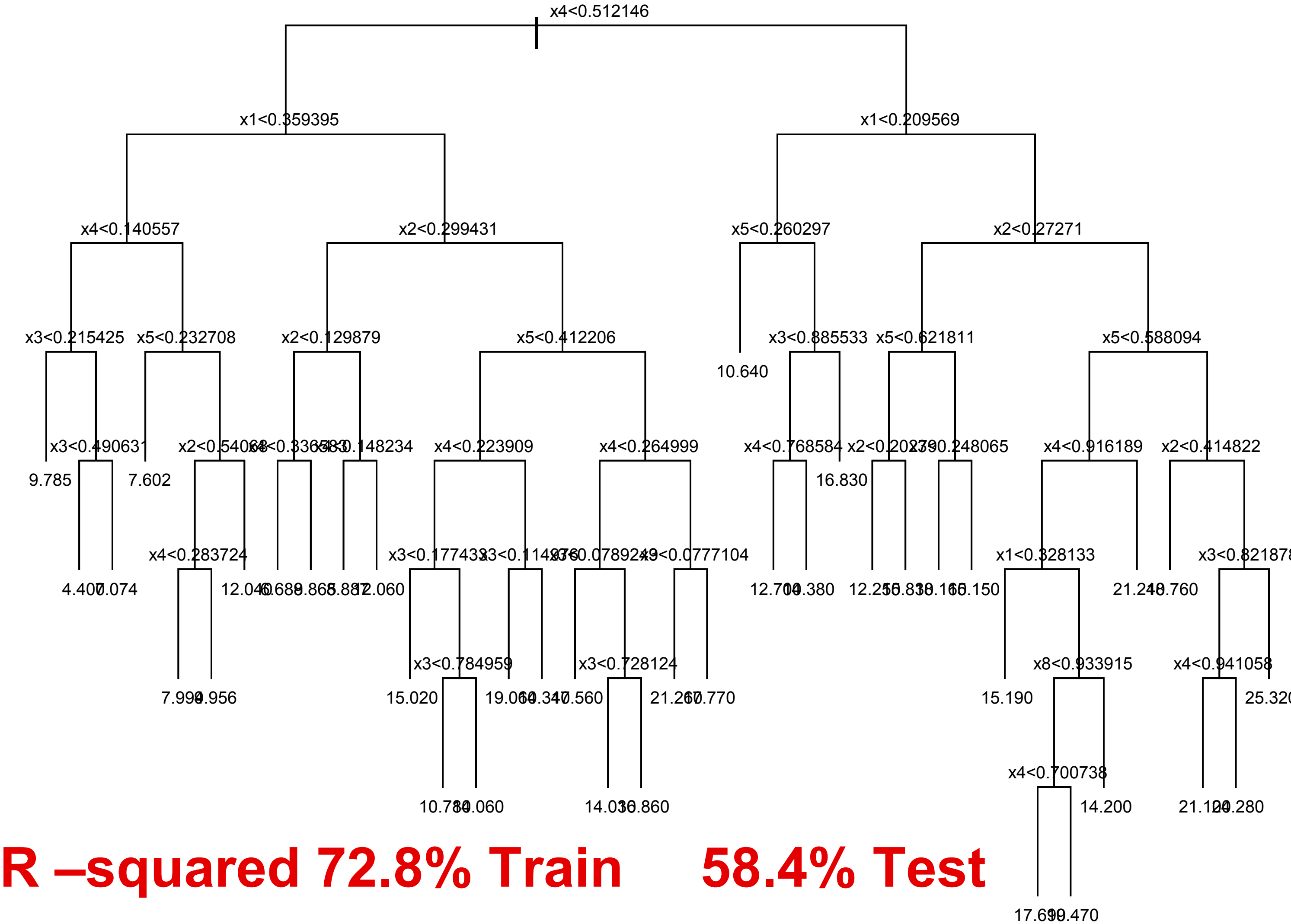
Tree Diagram



Regression Tree



Tree Model



R –squared 72.8% Train 58.4% Test

Example #8

- **Fit a tree to the “toy problem data”**
- **Fit a tree to the Depression study data**
  - Fit various strategies for missing values

Tree Advantages

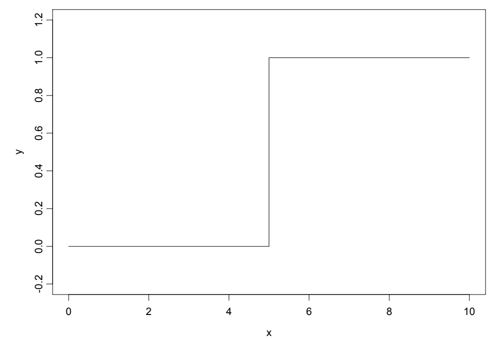
- **Model explains its reasoning -- builds rules**
- **Build model quickly**
- **Handles non-numeric data**
- **No problems with missing data**
    - Missing data as a new value
    - Surrogate splits
- **Works fine with many dimensions**

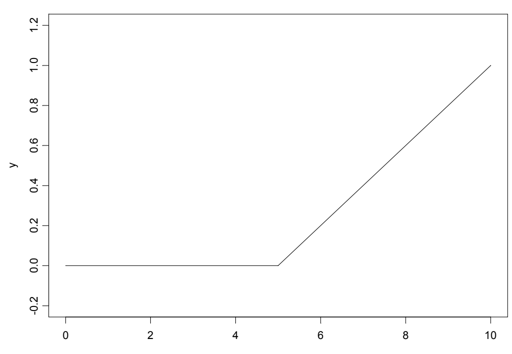
What's Wrong With Trees?

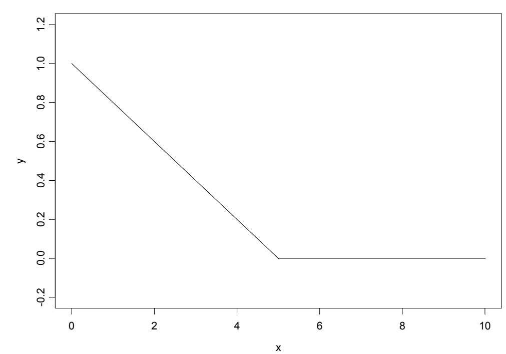
- Output are step functions – big errors near boundaries
- Greedy algorithms for splitting – small changes change model
- Uses less data after every split
- Model has high order interactions -- all splits are dependent on previous splits
- Often non-interpretable

MARS

- **Multivariate Adaptive Regression Splines**
- **What do they do?**
  - ➢ Replace each step function in a tree model by a pair of linear functions.







How Does It Work?

- Replace each step function by a pair of linear basis functions.
- New basis functions may or may not be dependent on previous splits.
- Replace linear functions with cubics after backward deletions.

Algorithm Details

- Fit the response y with a constant (i.e. Find its mean)
- Pick the variable and knot location which give the best fit in terms of residual sum of squares error.
- Repeat this process on every other variable. Limit typically on number of basis functions allowed.

Details II

- Model has too many basis functions. Perform backward elimination of individual terms that do not improve the fit enough to justify the increased complexity.
- Fit the resulting model with a smooth function to avoid discontinuities.

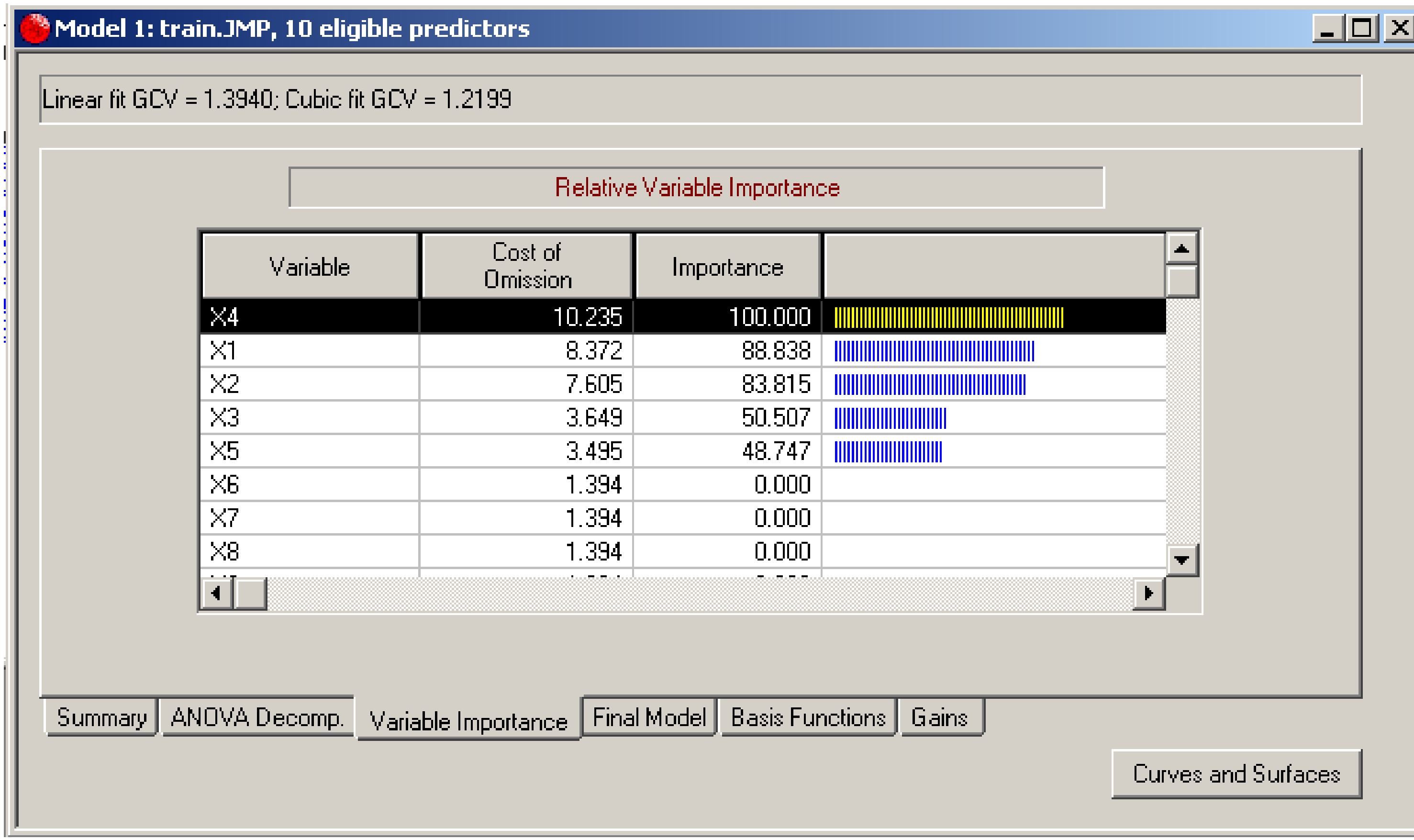
MARS Output

MARS modeling, version 3.5 (6/16/91)

forward stepwise knot placement:

basfn(s)	gcv	#indbsfns	#efprms	var	knot	parent
0	25.67	0.0	1.0			
1	17.36	1.0	7.0	4.	0.9308E-02	0.
3 2	12.26	3.0	14.0	1.	0.7059	0.
5 4	7.794	5.0	21.0	2.	0.6765	0.
7 6	6.698	7.0	28.0	3.	0.6465	1.
9 8	5.701	9.0	35.0	5.	0.3413	0.
11 10	5.324	11.0	42.0	1.	0.3754	4.
13 12	5.052	13.0	49.0	3.	0.3103	5.
15 14	5.869	15.0	56.0	4.	0.3269	2.
17 16	6.998	17.0	63.0	1.	0.5097	5.
19 18	8.761	19.0	70.0	3.	0.4290	0.
21 20	11.59	21.0	77.0	3.	0.8270	3.
23 22	20.83	23.0	84.0	3.	0.5001	2.
25 24	58.24	25.0	91.0	10.	0.2250	9.
26	461.7	26.0	97.0	10.	0.4740E-02	8.

MARS Variable Importance



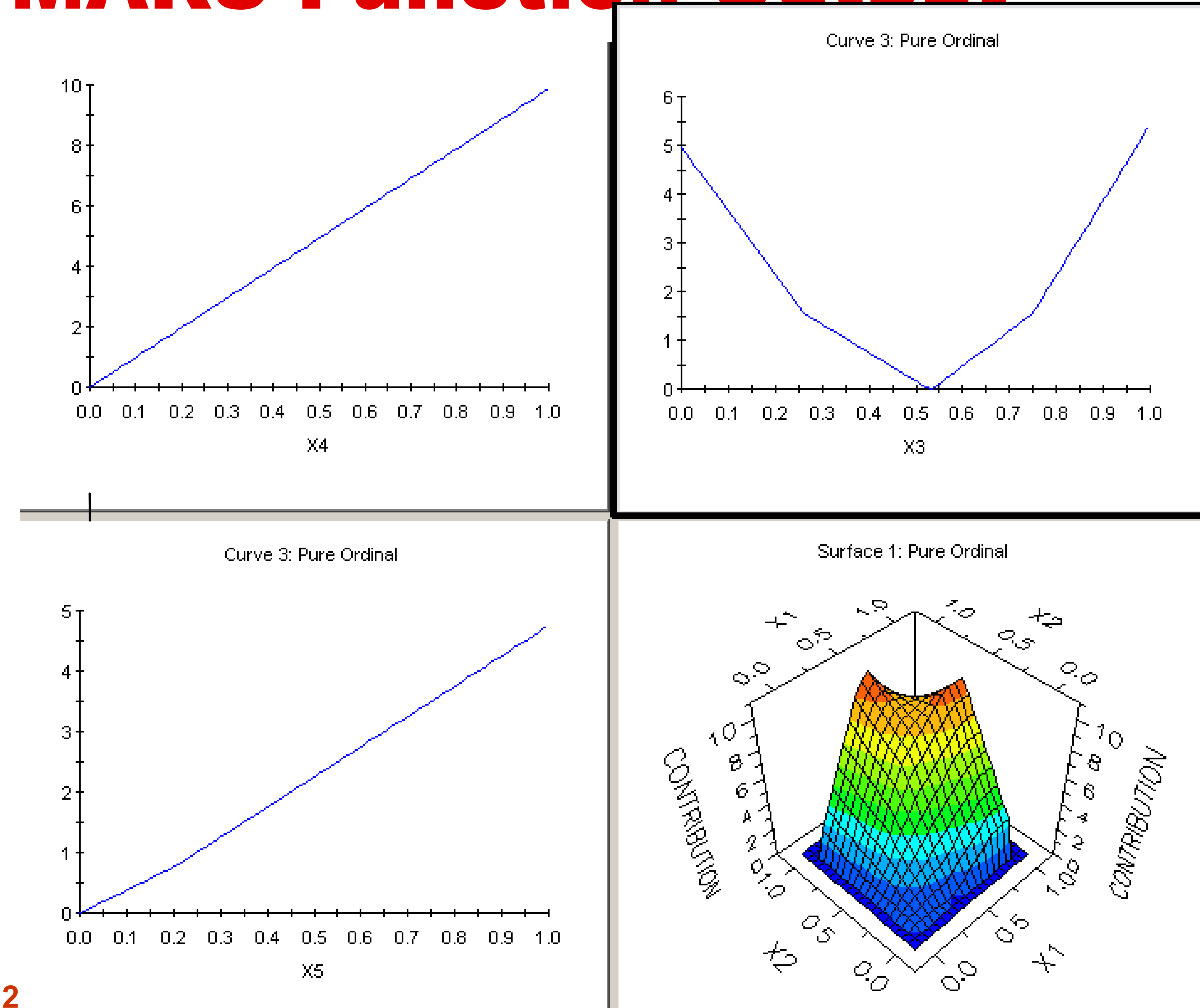
Model 1: train.JMP, 10 eligible predictors

Linear fit GCV = 1.3940; Cubic fit GCV = 1.2199

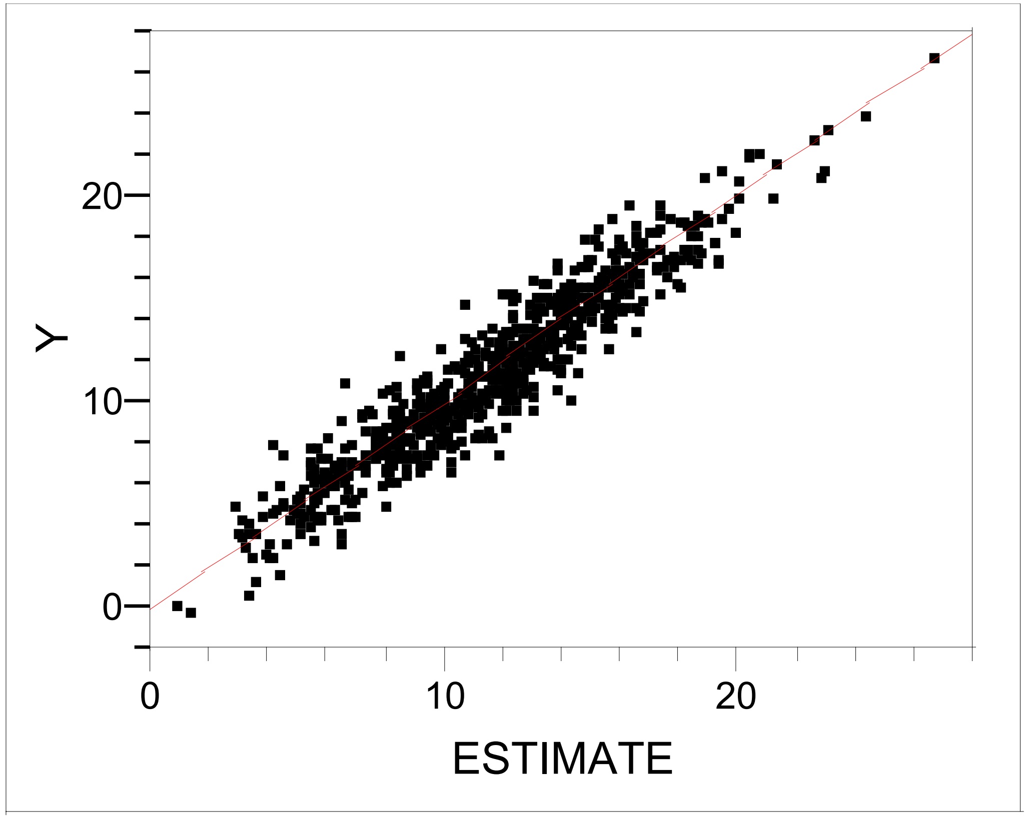
Relative Variable Importance

Variable	Cost of Omission	Importance	
X4	10.235	100.000	
X1	8.372	88.838	
X2	7.605	83.815	
X3	3.649	50.507	
X5	3.495	48.747	
X6	1.394	0.000	
X7	1.394	0.000	
X8	1.394	0.000	

MARS Function Output



Predictions for Example



R^2 = 89.6% Training Set 89.0% Test Set

Example #9

- Fit Mars to the “toy problem data”
- Compare to other models

Summary of MARS Features

- Produces smooth surface as a function of many predictor variables
- Automatically selects subset of variables
- Automatically selects complexity of model
- Tends to give low order interaction models preference
- Amount of smoothing and complexity may be tuned by user

K-Nearest Neighbors(KNN)

- **To predict *y* for an *x*:**
  - Find the *k* most similar *x*'s
  - Average their *y*'s
- **Find *k* by cross validation**
- **No training (estimation) required**
- **Works embarrassingly well**
  - Friedman, KDDM 1996

Collaborative Filtering

- Goal: predict what movies people will like
- Data: list of movies each person has watched

Lyle	Andre, Starwars
Ellen	Andre, Starwars, Hiver
Fred	Starwars, Batman
Dean	Starwars, Batman, Rambo
Jason	Emilie Poulin, Chocolat

Data Base

- Data can be represented as a sparse matrix

	Starwars	Batman	Rambo	Andre	Destin d'Emilie	Chocolat
Lyle	y			y		
Ellen	y			y	y	
Fred	y	y				
Dean	y	y	y			
Jason	y				y	y
Karen	?	?	?	y	?	?

- Karen likes Andre. What else might she like?
- CDNow doubled e-mail responses

Clustering

- **Turn the problem around**
- **Instead of predicting something about a variable, use the variables to group the observations**
  - K-means
  - Hierarchical clustering

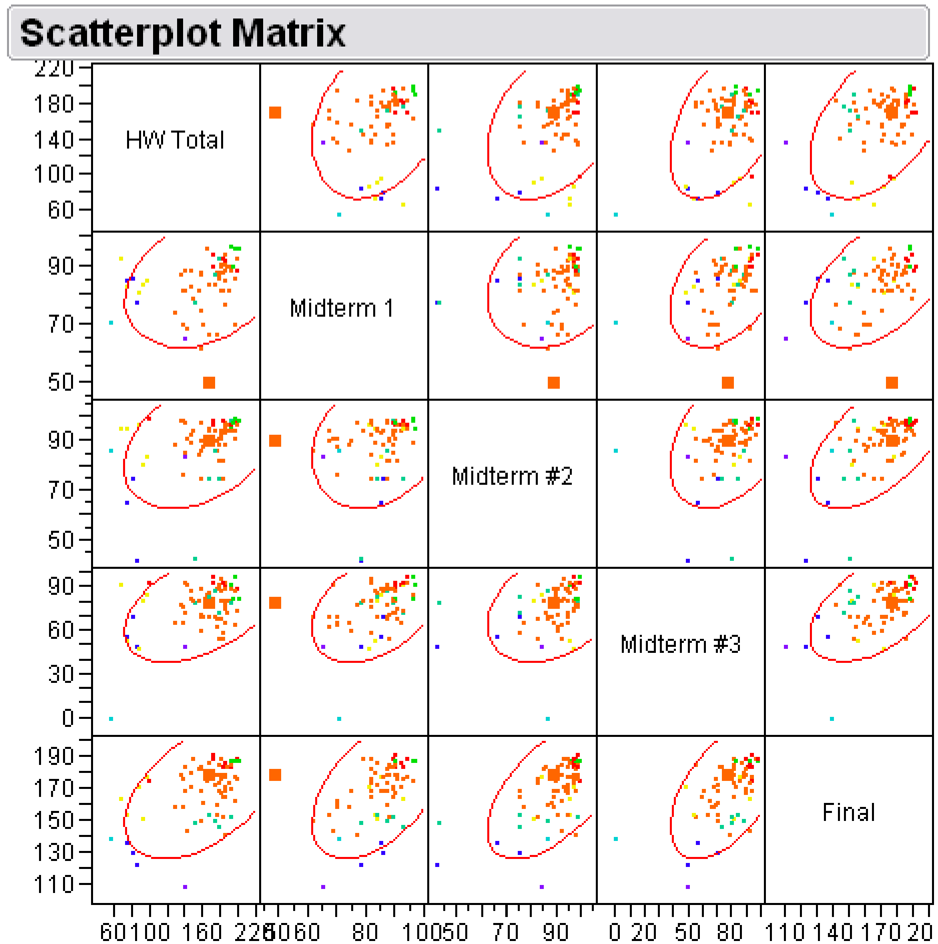
K-Means

- Rather than find the K nearest neighbors, find K clusters
- Problem is now to group observations into clusters rather than predict
- Not a predictive model, but a segmentation model

Example

- **Final Grades**
  - Homework
  - 3 Midterms
  - Final
- **Principal Components**
  - First is weighted average
  - Second is difference between 1 and 3rd midterms and 2nd

Scatterplot Matrix



Principal Components

Principal Components: on Correlations

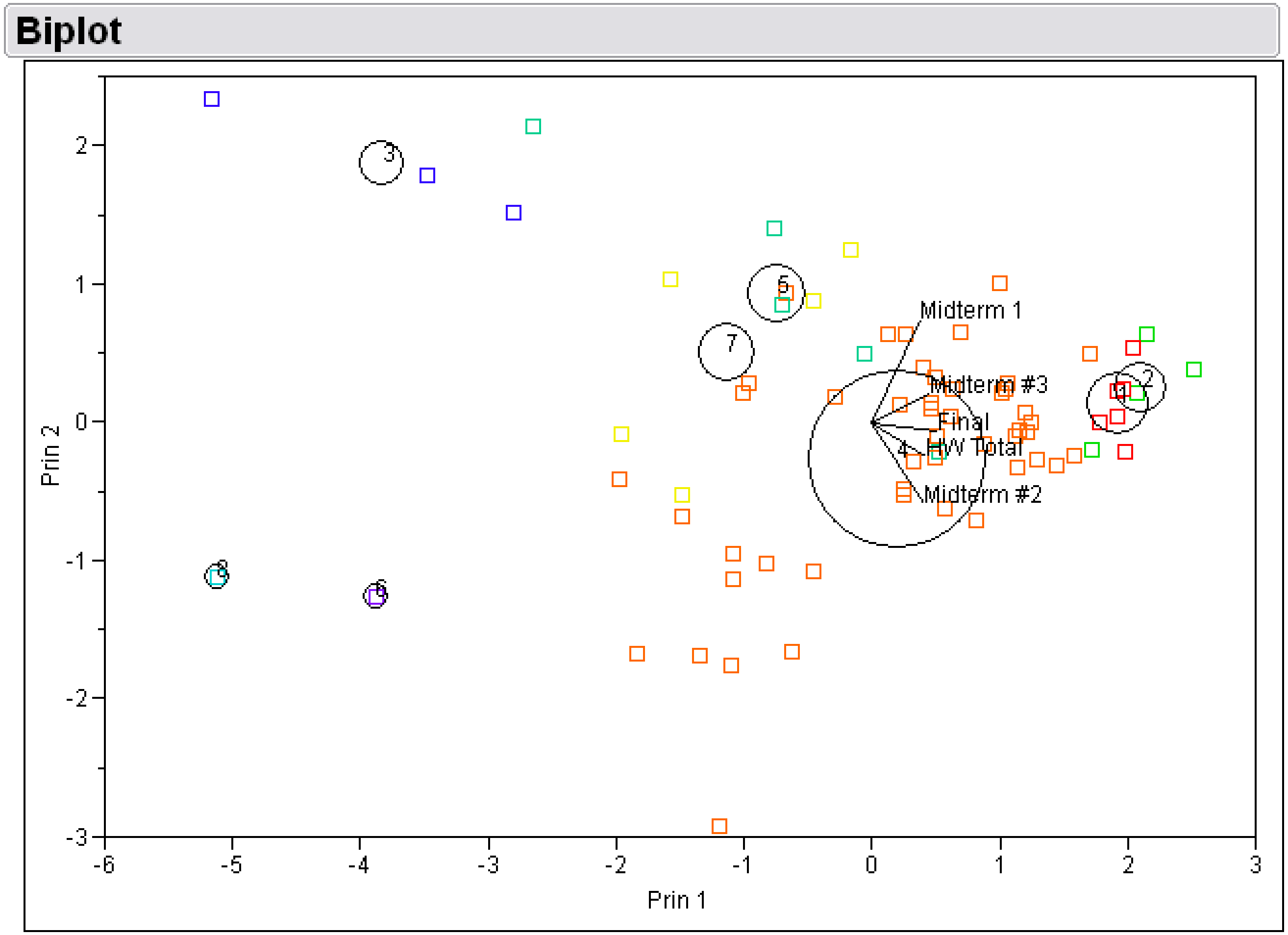
Eigenvalue	2.7060	0.8074	0.6725	0.4823	0.3317
Percent	54.1200	16.1484	13.4499	9.6469	6.6349
Cum Percent	54.1200	70.2683	83.7182	93.3651	100.0000
Eigenvectors					
HW Total	0.43295	-0.24849	-0.65677	0.55784	0.09094
Midterm 1	0.38549	0.74425	0.31141	0.35434	0.27378
Midterm #2	0.41892	-0.58041	0.51664	-0.05152	0.46697
Midterm #3	0.46495	0.20845	-0.38284	-0.74856	0.18296
Final	0.52181	-0.06342	0.24122	-0.01627	-0.81562

Cluster Means

Cluster Means

Cluster	HW Total	Midterm 1	Midterm #2	Midterm #3	Final
1	183.833333	91.3333333	97.8333333	93.3333333	188
2	195.75	94.75	98.25	89	188
3	81	83	61	59	130.333333
4	169.234043	82.7446809	91.2978723	77.8085106	172
5	172.2	86.2	75.8	81	151.2
6	139	65	84	50	110
7	84.4	85.2	90.8	72.4	164.4
8	56	71	87	0	139

Biplot



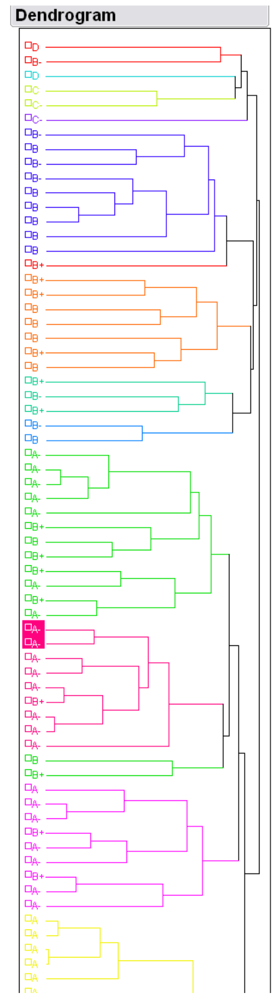
Eigenvalues

2.7059981 0.8074182 0.6724939 0.4823449 0.3317448

Hierarchical Clustering

- **Define distance between two observations**
- **Find closest observations and form a group**
  - Add on to this to form hierarchy

Grade Example



Example

- **Data on fifty states**
- **Find Clusters**
- **Examine Hierarchical Cluster**
  - Do clusters make sense?
  - What did we learn?

Genetic Algorithms

- **Genetic algorithms are a search procedure**
- **Part of the optimization toolbox**
- **Typically used on LARGE problems**
  - Molecular chemistry - rational drug design
  - Survival of the fittest
- **Can replace any optimization procedure**
  - May be very slow on moderate problems
  - May not find optimal point

Support Vector Machines

- **Mainly a classifier although can be adapter for regression**
- **Black box**
  - Uses a linear combination of transformed features in very high dimensions to separate points
  - Transformations (kernels) problem dependent
- **Based on Vapnik's theory**
  - See Friedman, Hastie and Tibshirani for more

Bagging and Boosting

- **Bagging (Bootstrap Aggregation)**
  - Bootstrap a data set repeatedly
  - Take many versions of same model (e.g. tree)
  - Form a committee of models
  - Take majority rule of predictions
- **Boosting**
  - Create repeated samples of weighted data
  - Weights based on misclassification
  - Combine by majority rule, or linear combination of predictions

MART

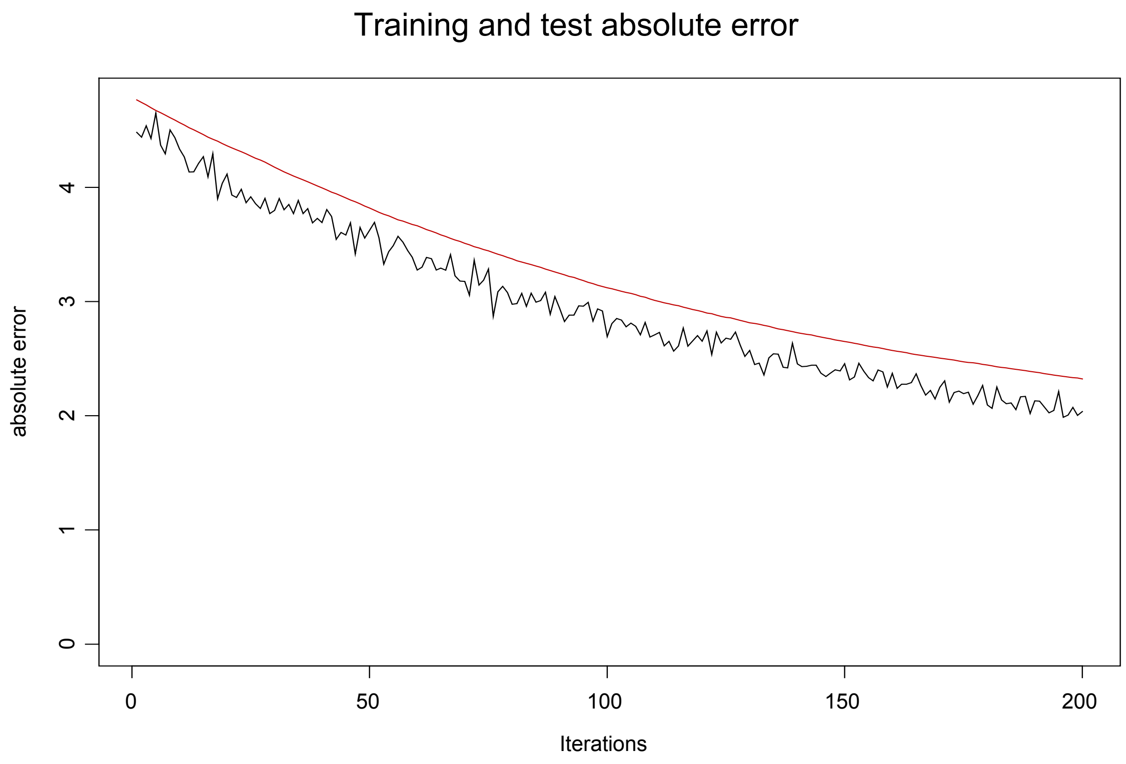
- **Boosting Version 1**
  - Use logistic regression.
  - Weight observations by misclassification
    - ✓ Upweight your mistakes
  - Repeat on reweighted data
  - Take majority vote

- **Boosting Version 2**
  - – use CART with 4-8 nodes
  - Use new tree on residuals
  - Repeat many, many times
  - Take predictions to be the sum of all these trees

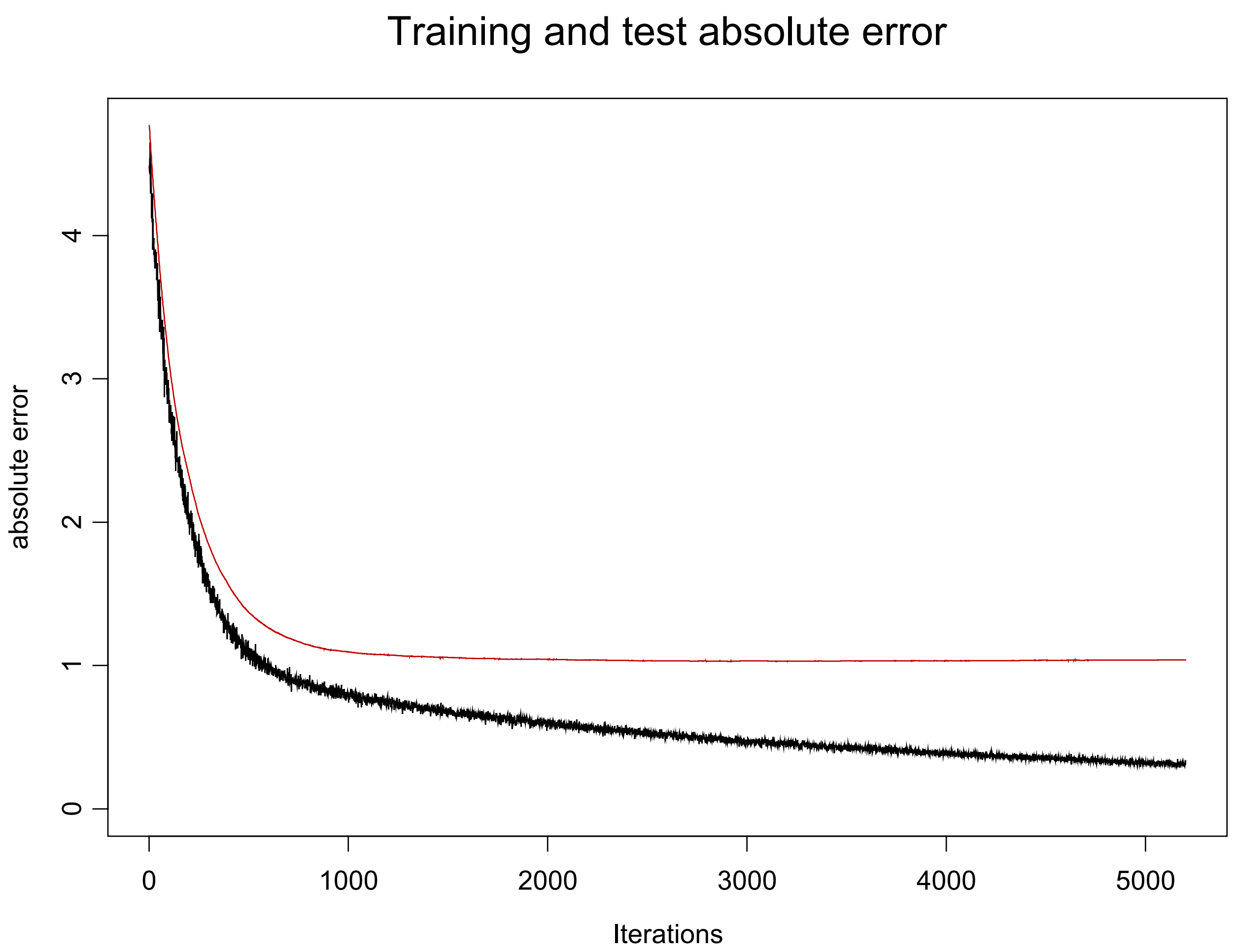
Upshot of MART

- Robust – because of loss function and because we use trees
- Low interaction order because we use small trees (adjustable)
- Reuses all the data after each tree

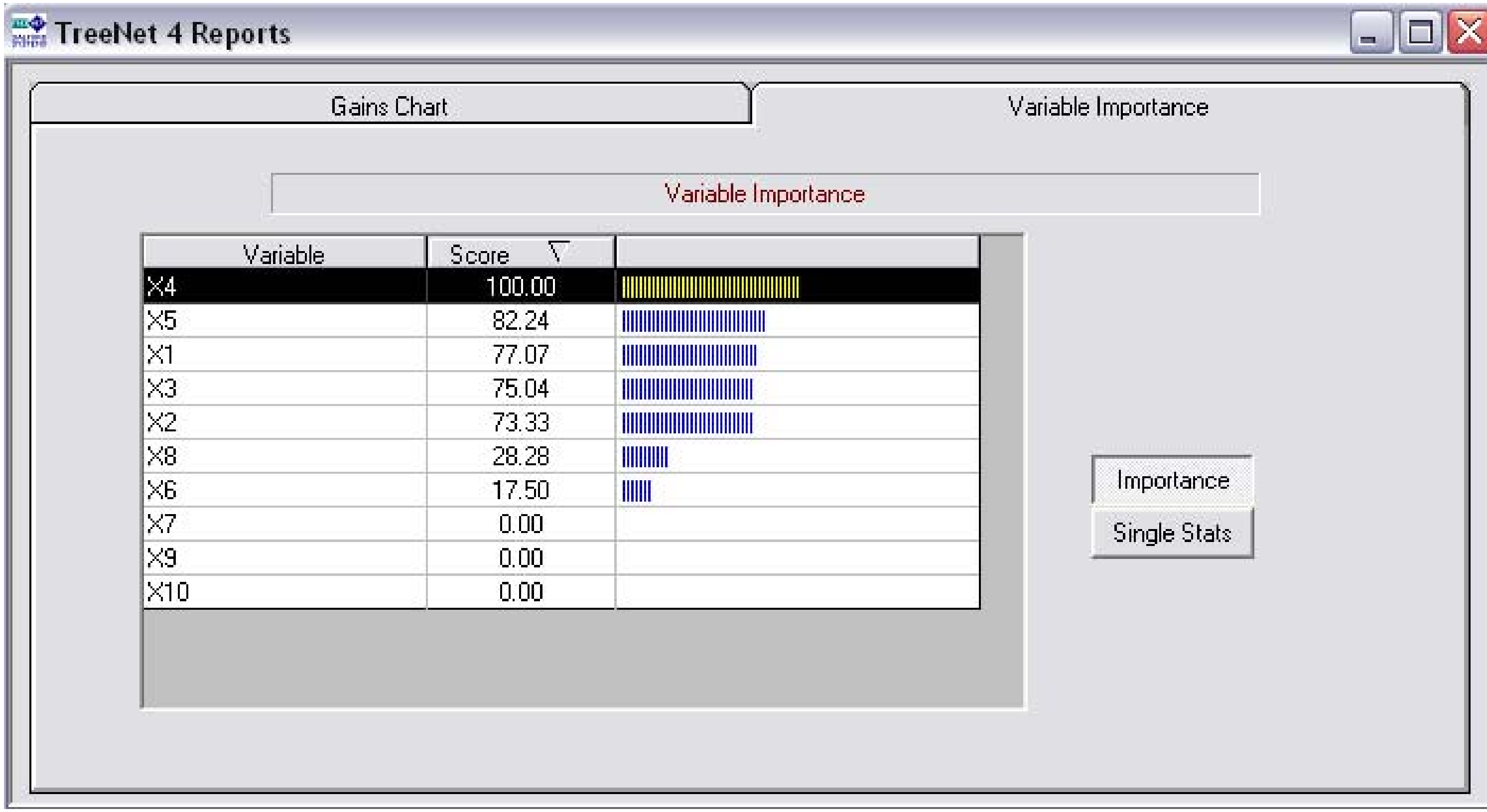
MART in action



More MART

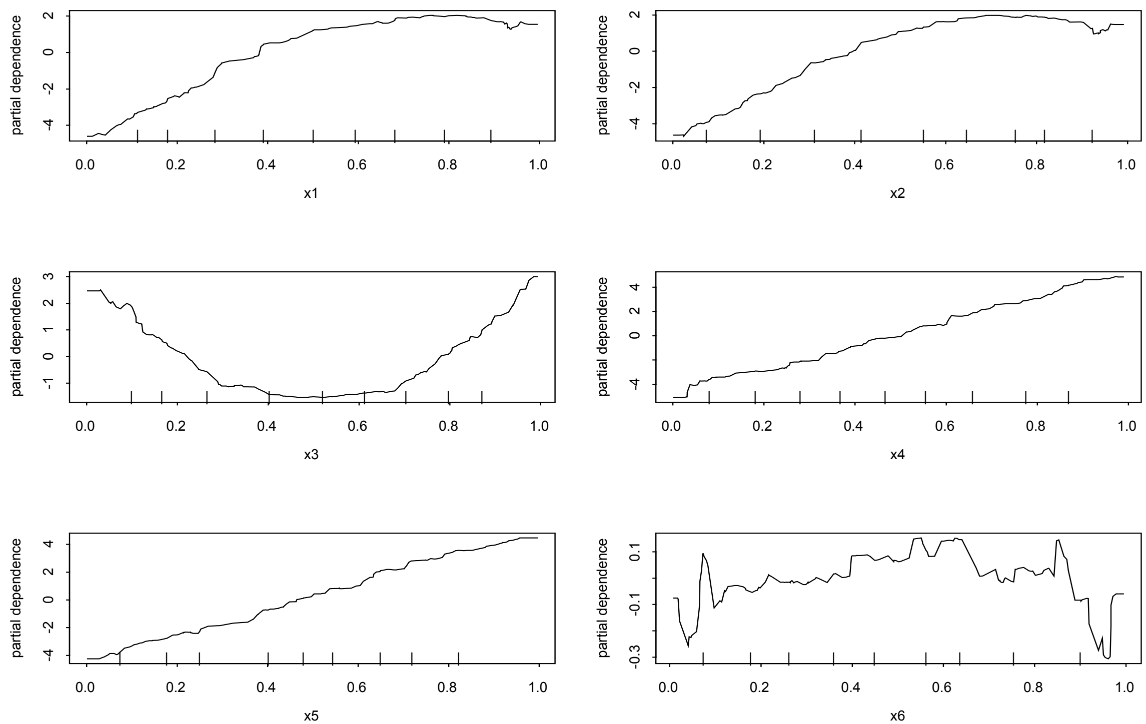

Training and test absolute error
absolute error
0
1
2
3
4
0
1000
2000
3000
4000
5000
Iterations

MART summary

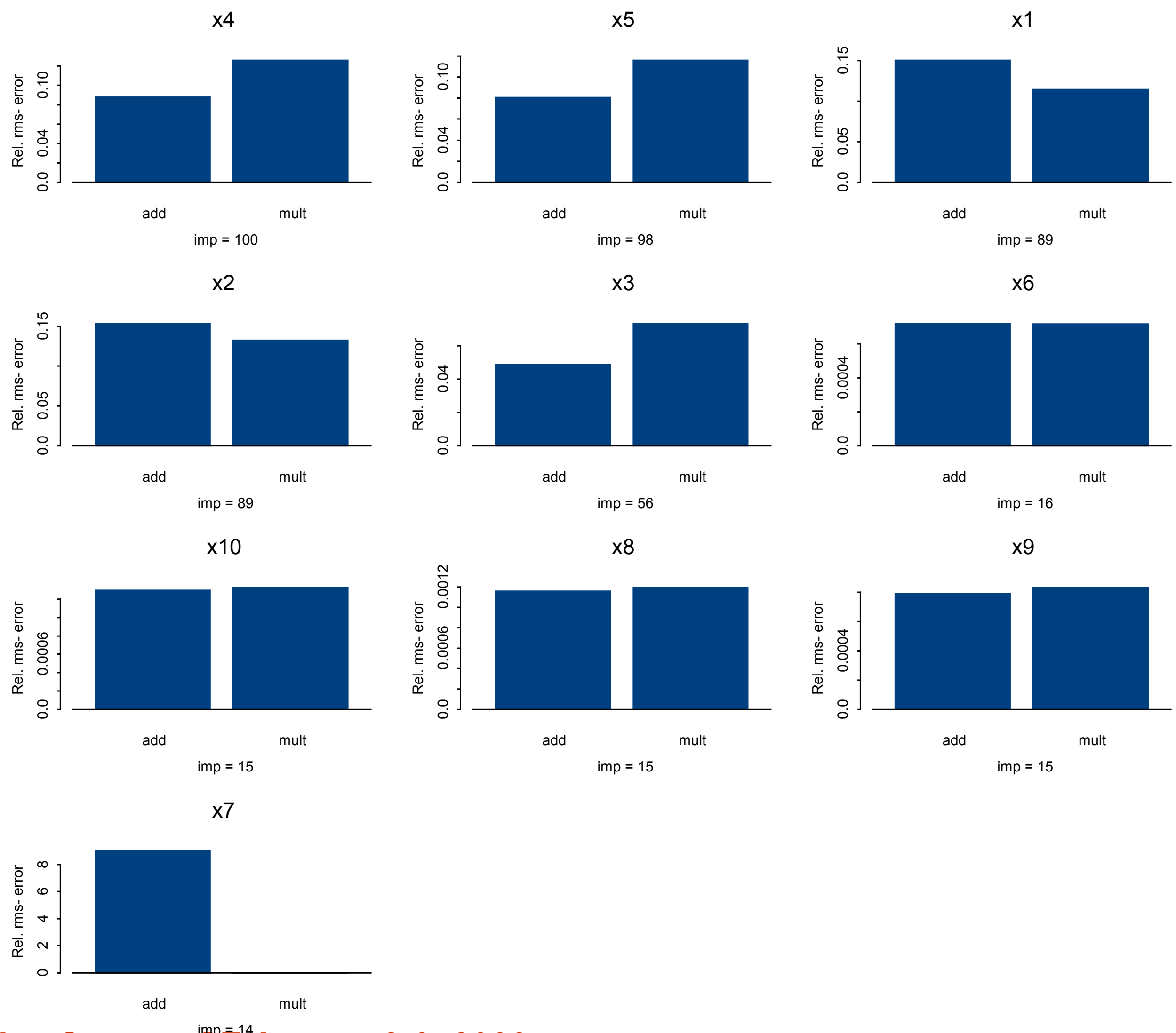


Variable	Score	
X4	100.00	
X5	82.24	
X1	77.07	
X3	75.04	
X2	73.33	
X8	28.28	
X6	17.50	
X7	0.00	
X9	0.00	
X10	0.00	

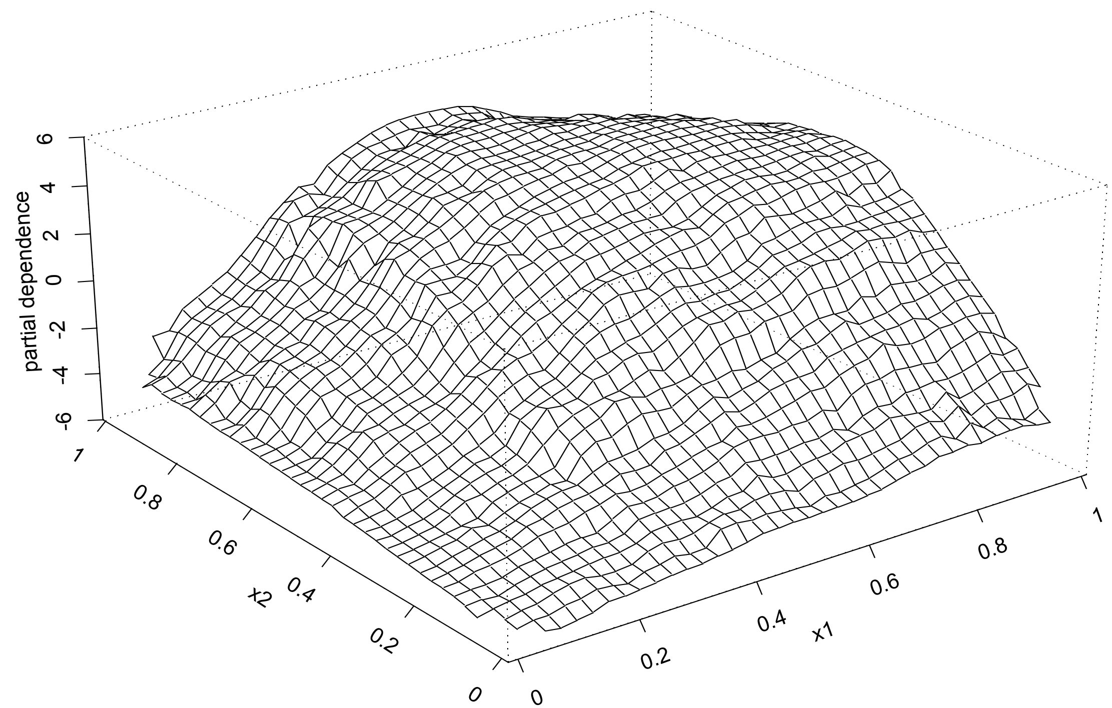
Single variable plots



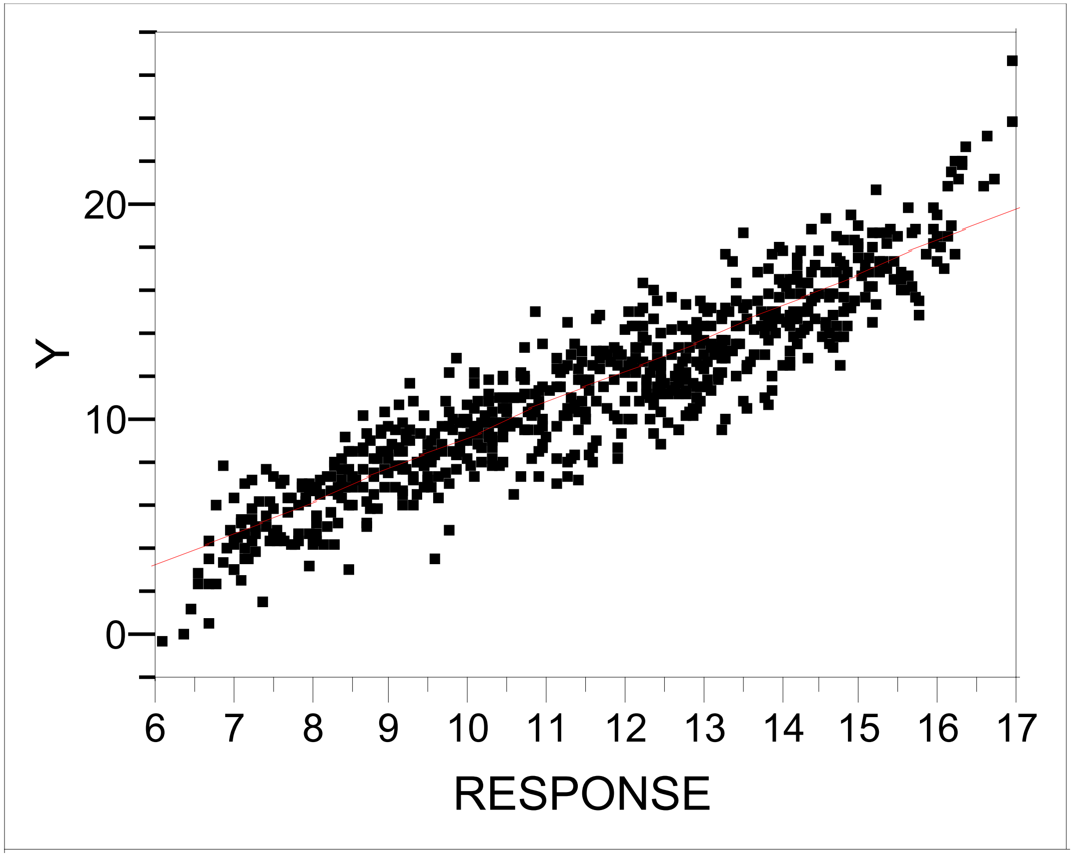
Interaction order?



Pairplots



MART Results



R squared 84.2% Train 78.4% Test

How Do We Really Start?

- **Life is not so kind**
  - Categorical variables
  - Missing data
  - 500 variables, not 10
- **481 variables – where to start?**

Where to Start

- **Three rules of data analysis**
  - Draw a picture
  - Draw a picture
  - Draw a picture
- **Ok, but how?**
  - There are 90 histogram/bar charts and 4005 scatterplots to look at (or at least 90 if you look only at y vs. X)

Exploratory Data Models

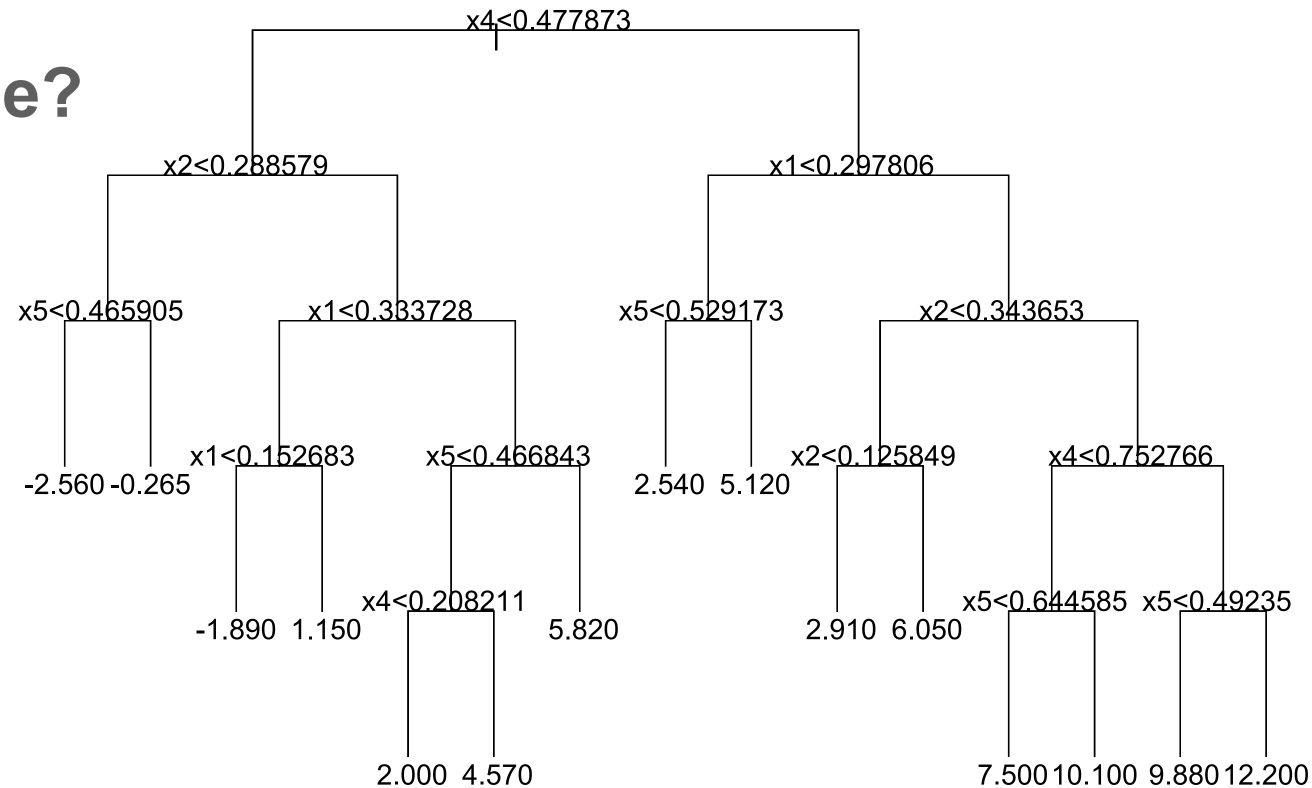
- **Use a tree to find a smaller subset of variables to investigate**
- **Explore this set graphically**
  - Start the modeling process over
- **Build model**
  - Compare model on small subset with full predictive model

More Realistic

- **200 predictors**
- **10,000 rows**
- **Why is this still easy?**
  - ➢ No missing values
  - ➢ All continuous predictors

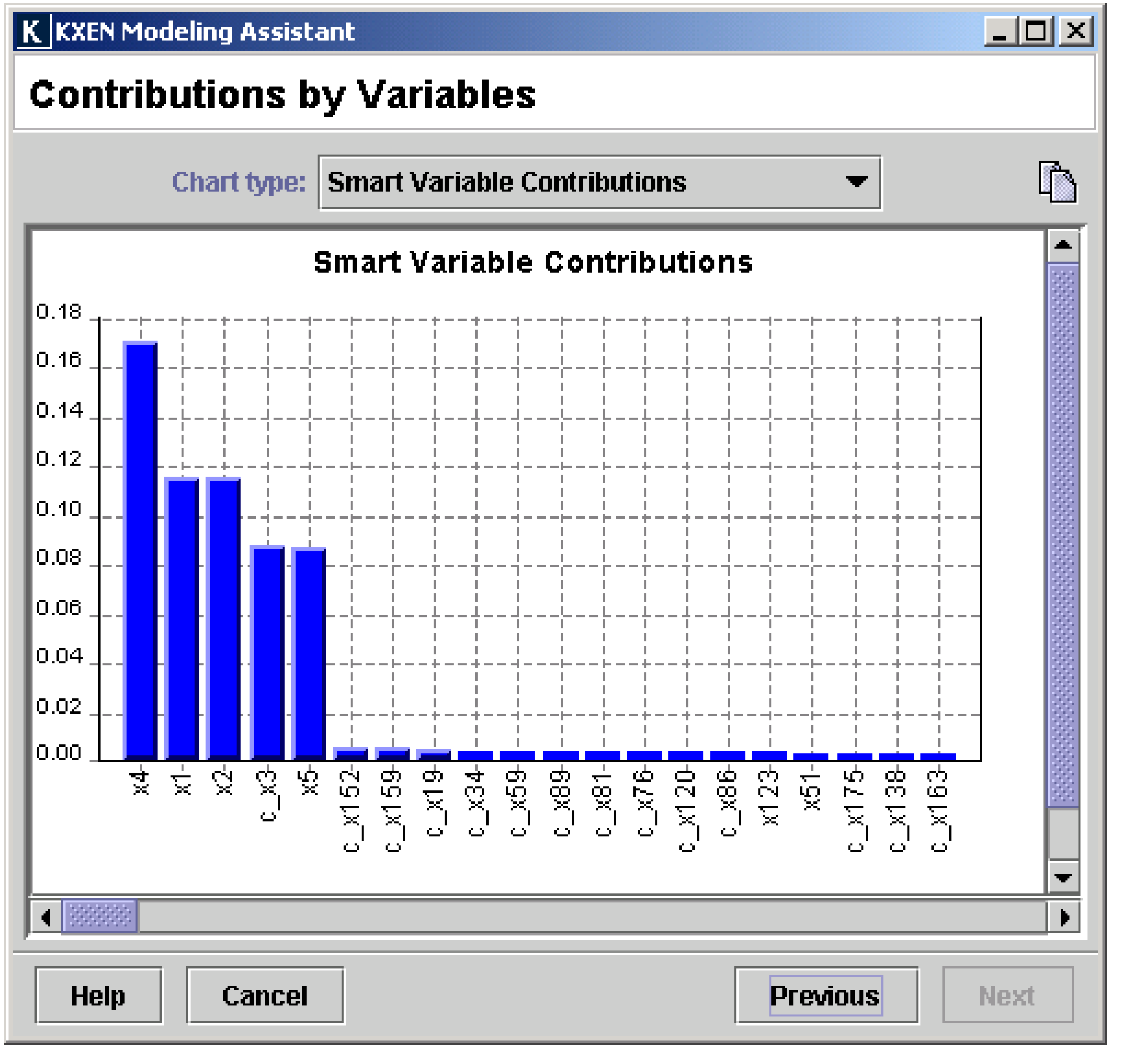
Start With a Simple Model

- Tree?

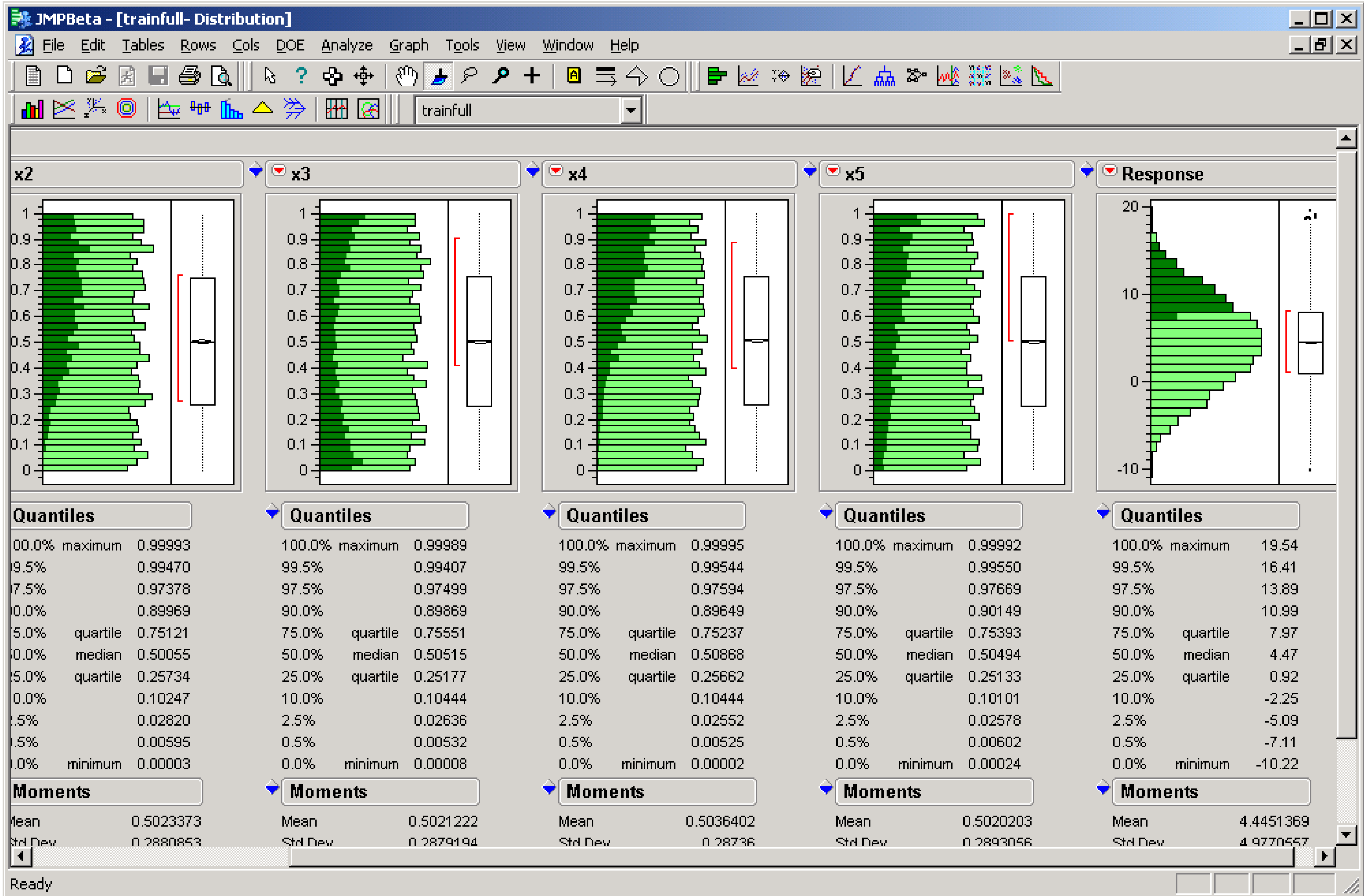


Automatic Models

- KXEN

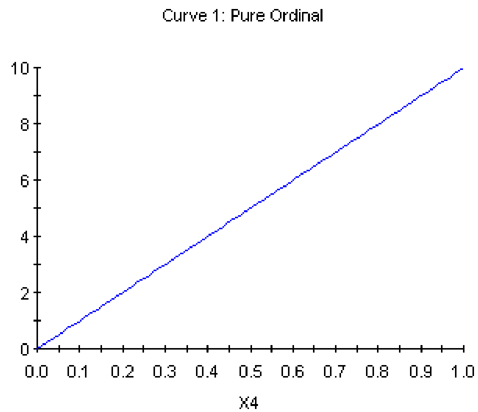


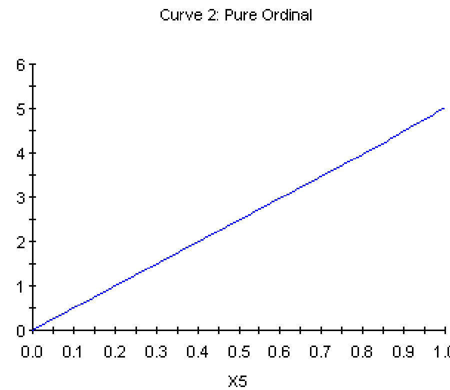
Brushing

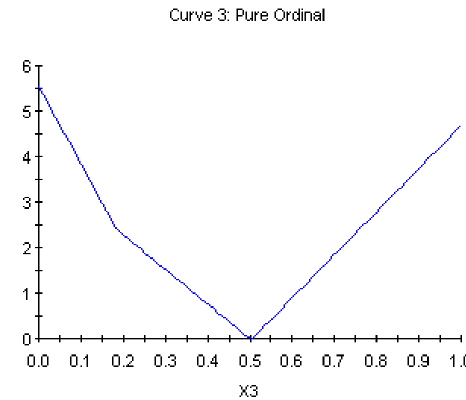


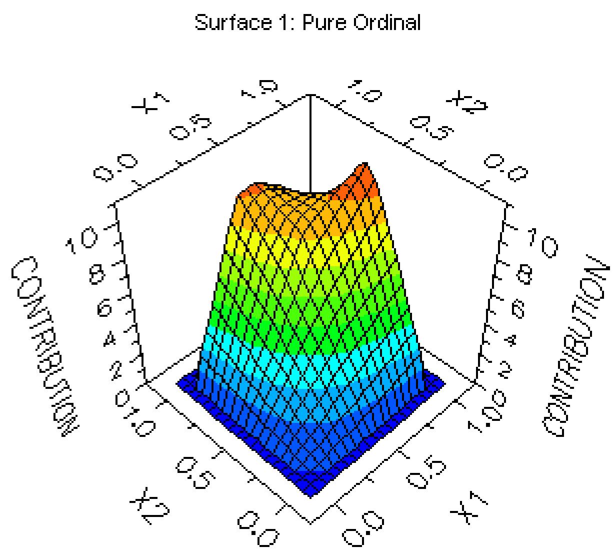
JMPBeta - [trainfull- Distribution]
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
trainfull
x2
x3
x4
x5
Response
Quantiles
00.0% maximum 0.99993
9.5% 0.99470
7.5% 0.97378
0.0% 0.89969
5.0% quartile 0.75121
0.0% median 0.50055
5.0% quartile 0.25734
0.0% 0.10247
.5% 0.02820
.5% 0.00595
.0% minimum 0.00003
Moments
lean 0.5023373
Quantiles
100.0% maximum 0.99989
99.5% 0.99407
97.5% 0.97499
90.0% 0.89869
75.0% quartile 0.75551
50.0% median 0.50515
25.0% quartile 0.25177
10.0% 0.10444
2.5% 0.02636
0.5% 0.00532
0.0% minimum 0.00008
Moments
Mean 0.5021222
Quantiles
100.0% maximum 0.99995
99.5% 0.99544
97.5% 0.97594
90.0% 0.89649
75.0% quartile 0.75237
50.0% median 0.50868
25.0% quartile 0.25662
10.0% 0.10444
2.5% 0.02552
0.5% 0.00525
0.0% minimum 0.00002
Moments
Mean 0.5036402
Quantiles
100.0% maximum 0.99992
99.5% 0.99550
97.5% 0.97669
90.0% 0.90149
75.0% quartile 0.75393
50.0% median 0.50494
25.0% quartile 0.25133
10.0% 0.10101
2.5% 0.02578
0.5% 0.00602
0.0% minimum 0.00024
Moments
Mean 0.5020203
Quantiles
100.0% maximum 19.54
99.5% 16.41
97.5% 13.89
90.0% 10.99
75.0% quartile 7.97
50.0% median 4.47
25.0% quartile 0.92
10.0% -2.25
2.5% -5.09
0.5% -7.11
0.0% minimum -10.22
Moments
Mean 4.4451369
Ready

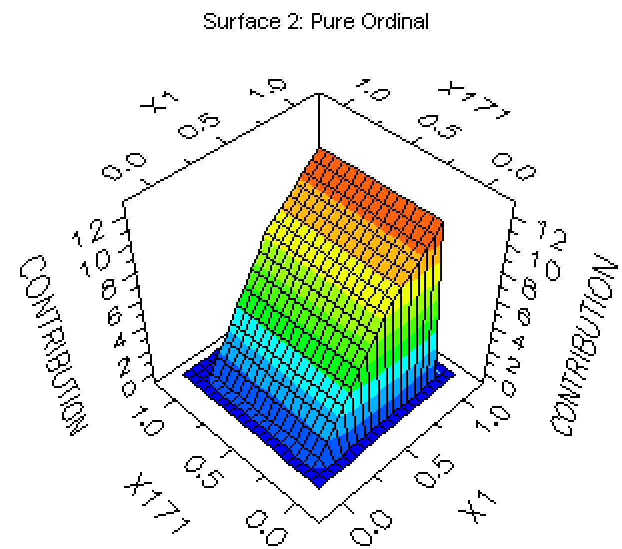
MARS Output


Curve 1: Pure Ordinal
0 2 4 6 8 10
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
X4


Curve 2: Pure Ordinal
0 1 2 3 4 5 6
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
X5


Curve 3: Pure Ordinal
0 1 2 3 4 5 6
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
X3


Surface 1: Pure Ordinal
X1
X2
CONTRIBUTION
CONTRIBUTION


Surface 2: Pure Ordinal
X1
X171
CONTRIBUTION
CONTRIBUTION

Variable Importance

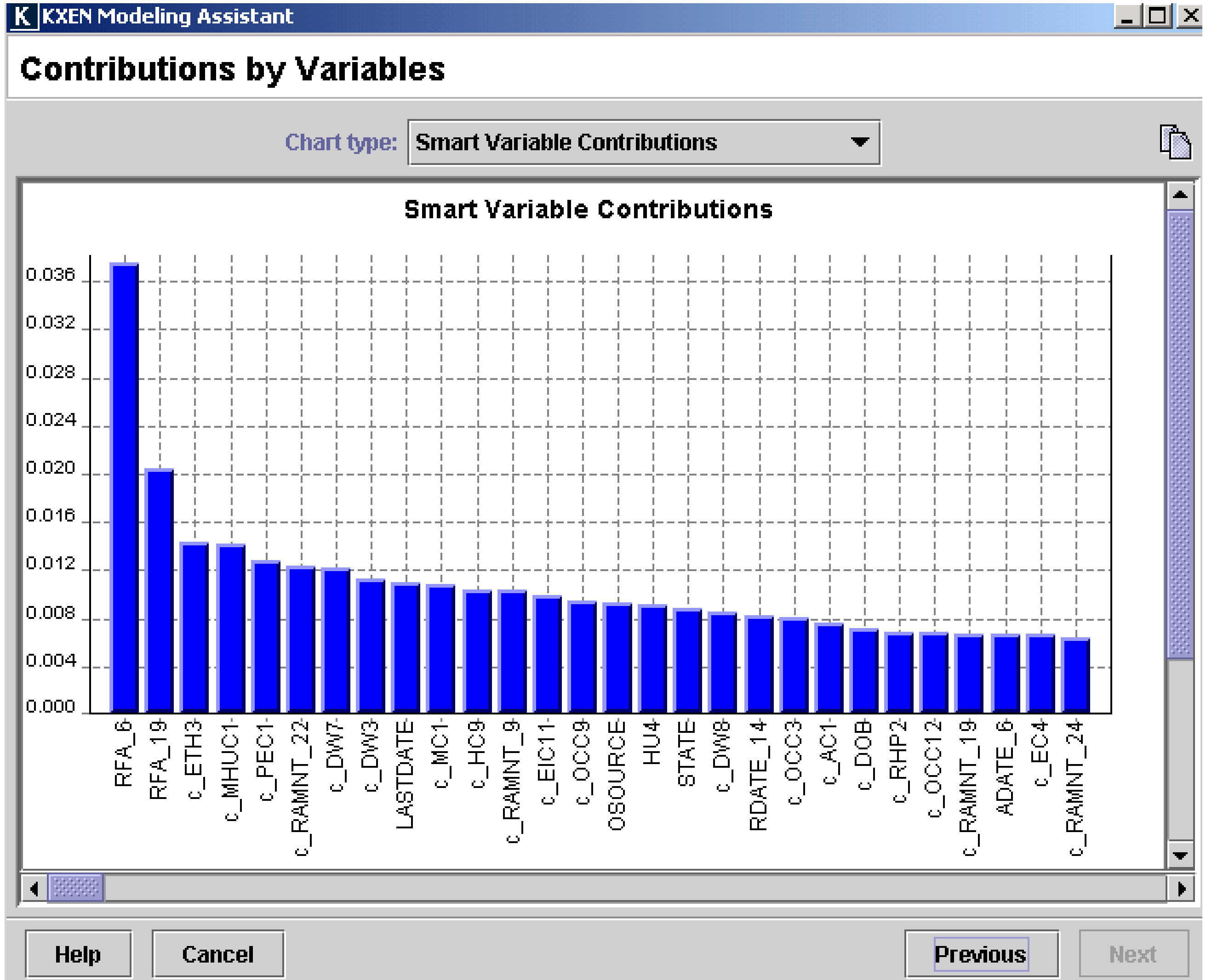
Relative Variable Importance

Variable	Cost of Omission	Importance
X4	9.479	100.000
X1	7.863	89.726
X2	7.585	87.839
X3	3.294	50.381
X5	3.292	50.347
X171	1.191	0.856
X6	1.191	0.000
X7	1.191	0.000

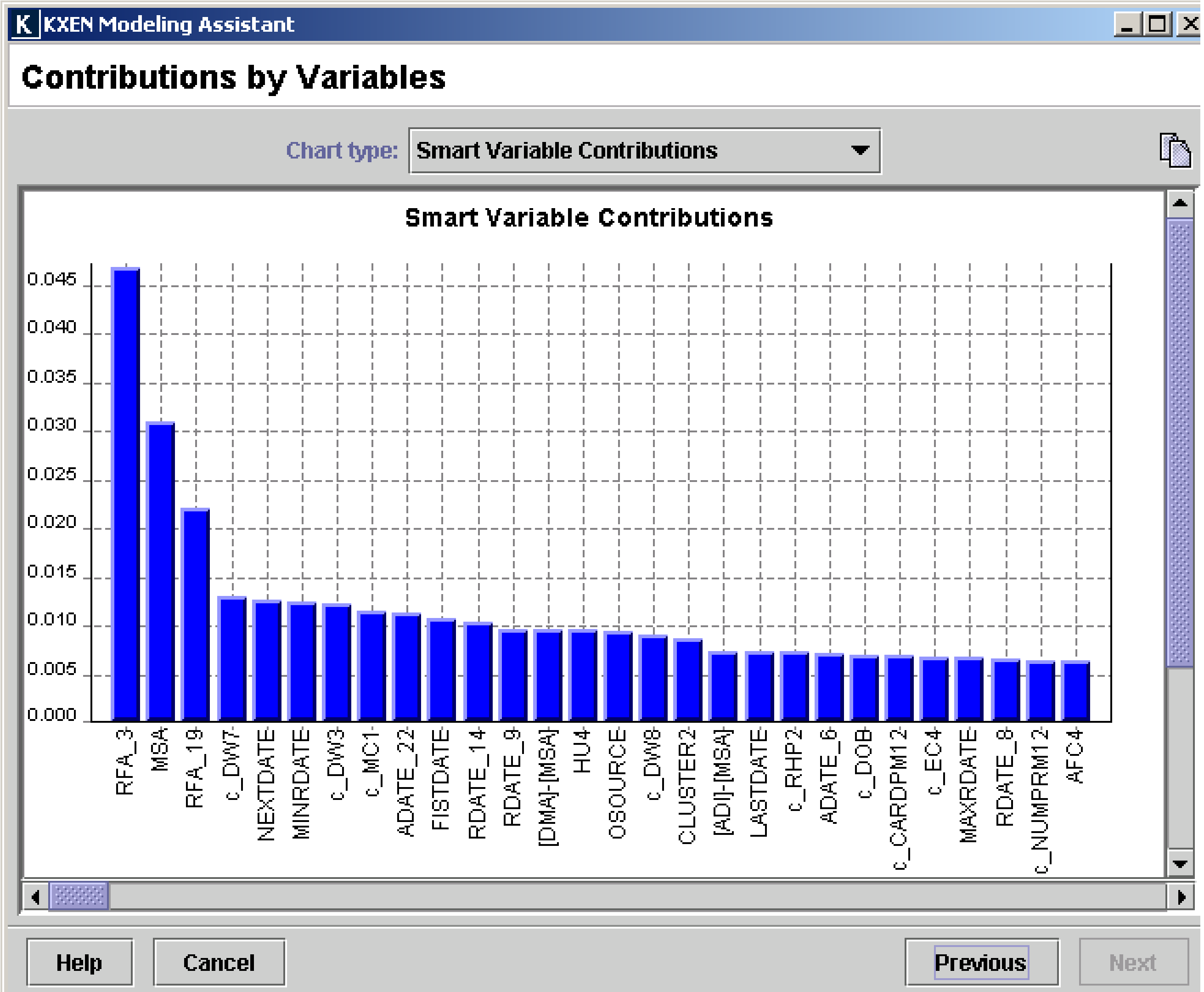
Back to Real Problem

- Missing values
- Many predictors
- Coding issues

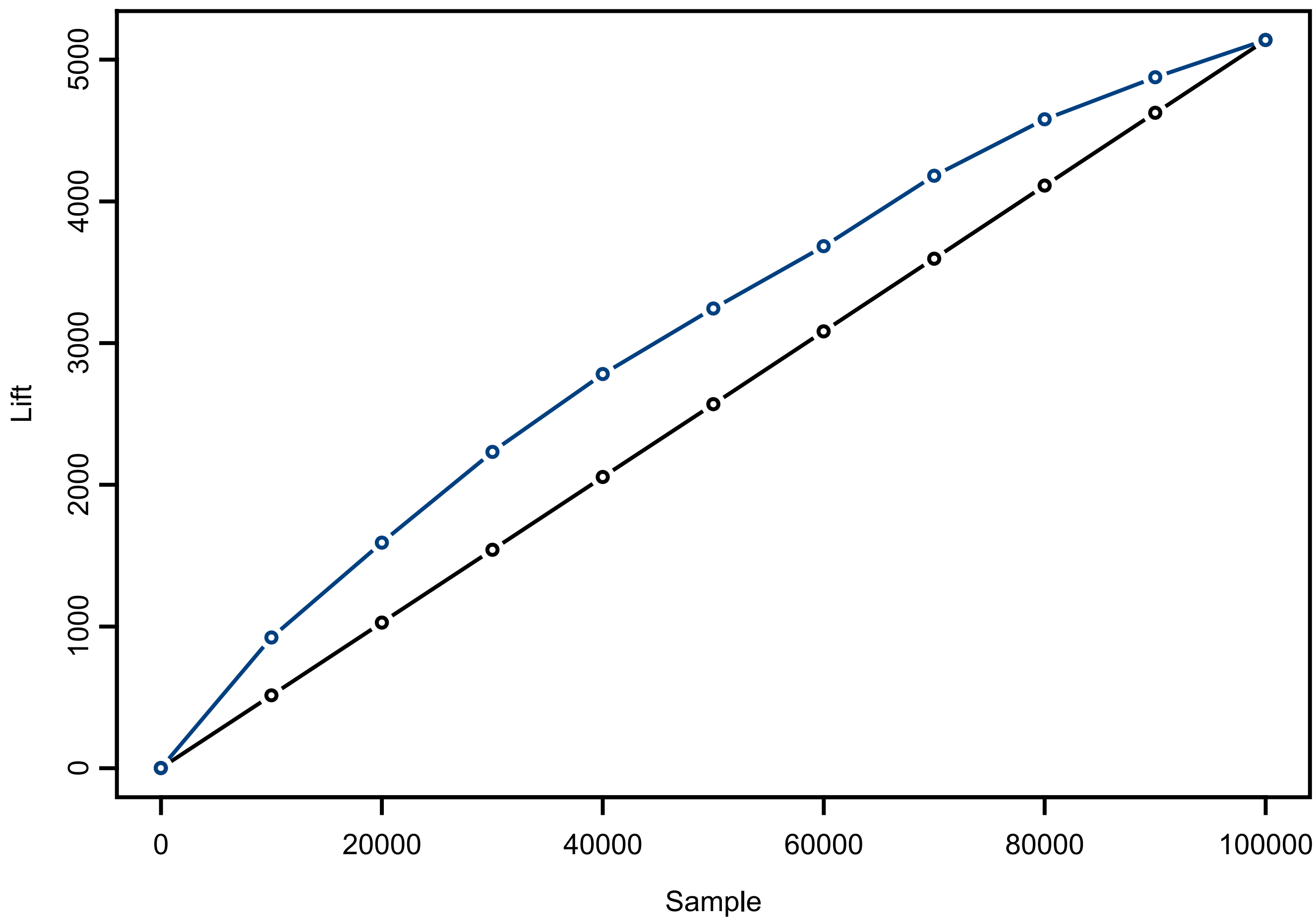
KXEN Variable Importance



With Correct Codings

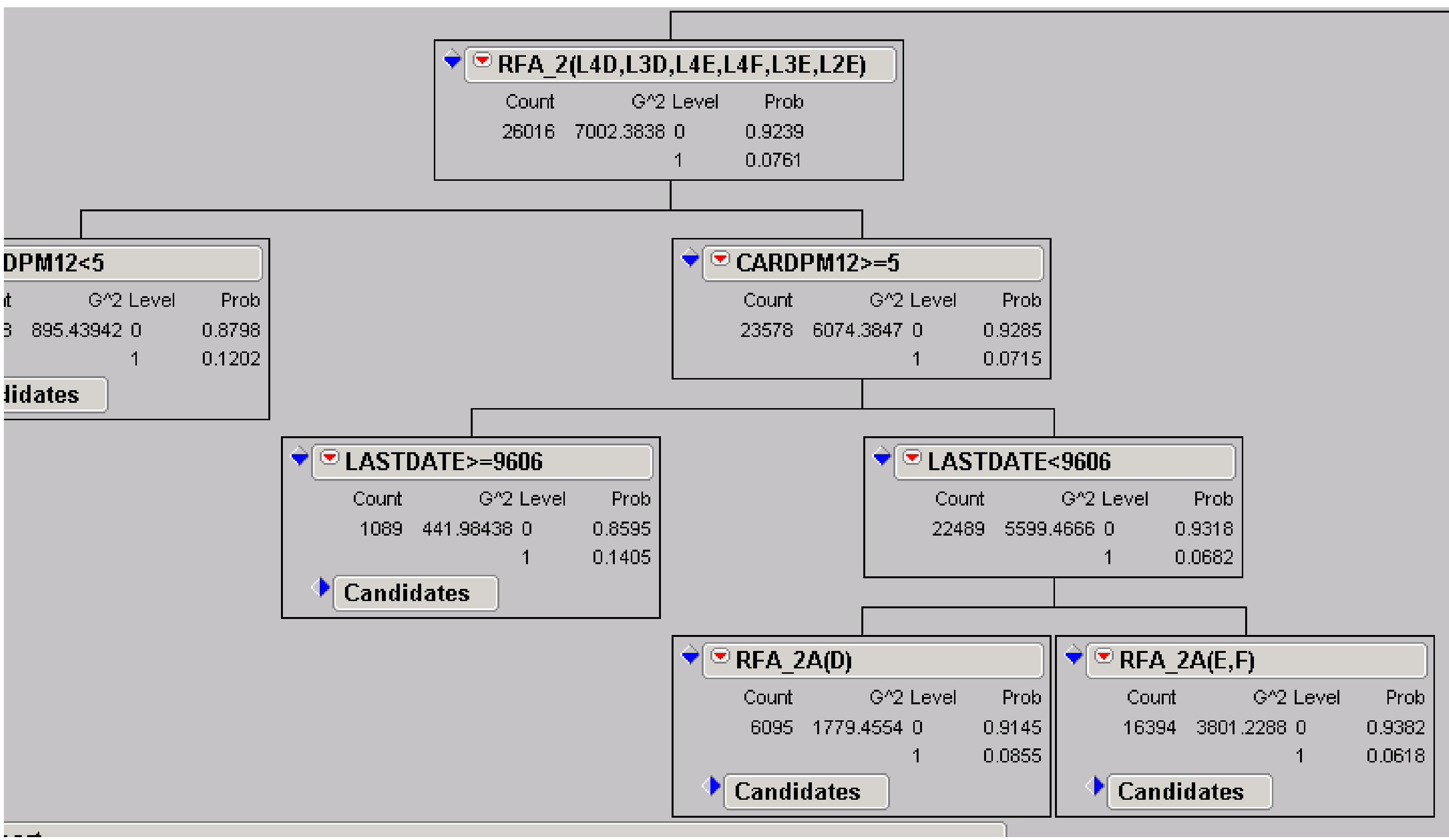


Lift Curve



Lift
0
1000
2000
3000
4000
5000
0
20000
40000
60000
80000
100000
Sample

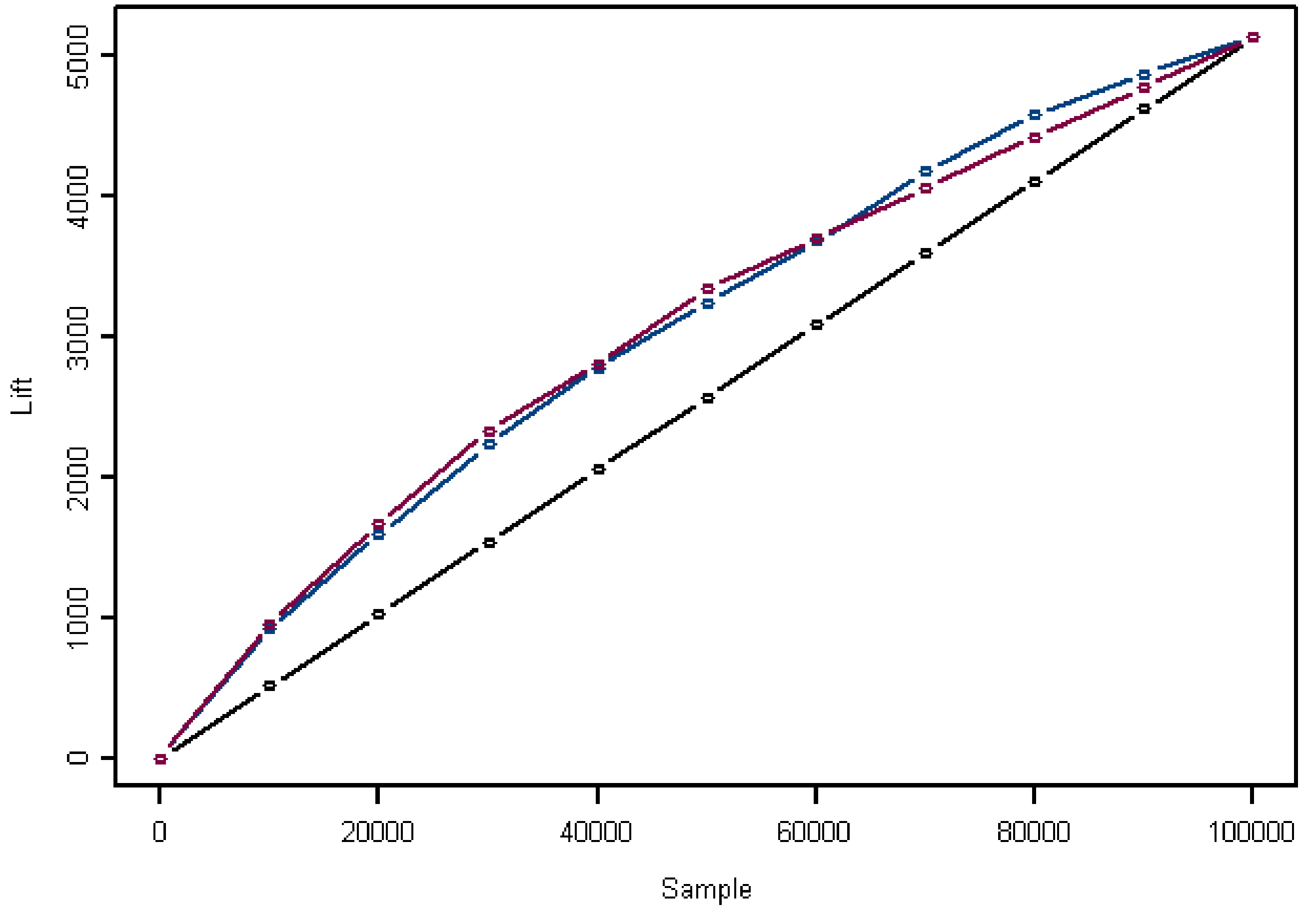
Exploratory Model



Tree Model

- **Tree model on 40 key variables as indentified by KXEN**
  - Very similar performance to KXEN model
  - More coarse
  - Based only on
    - ✓ RFA_2
    - ✓ Lastdate
    - ✓ Nextdate
    - ✓ Lastgift
    - ✓ Cardprom

Tree vs. KXEN



Lift
Sample
0
1000
2000
3000
4000
5000
0
20000
40000
60000
80000
100000

Is This the Answer?

- Actual question is to predict profit
  - ➢ Two stage model
    - ✓Predict response (yes/no)
    - ✓Then predict amount for responders
  - ➢ Use amounts as weights
    - ✓Predict amount directly
    - ✓Predict yes/no directly using amount as weight
- Start these models building on what we learned from simple models

What Did We Learn?

- **Toy problem**
  - Functional form of model
- **PVA data**
  - Useful predictor – increased sales 40%
- **Insurance**
  - Identified top 5% of possibilities of losses
- **Ingots**
  - Gave clues as to where to look
  - Experimental design followed

Interpretation or Prediction? Which Is Better?

- None of the models represents reality
- All are models and therefore wrong
- Answer to which is better is completely situation dependent

Why Interpretable?

- Depends on goal of project
- For routine applications goal may be predictive
- For breakthrough understanding, black box not enough

Spatial Analysis

- **Warranty data showing problem with ink jet printer**
- **Black box model shows that zip code is most important predictor**
  - Predictions very good
  - What do we learn?
  - Where do we go from here?

Zip Code?

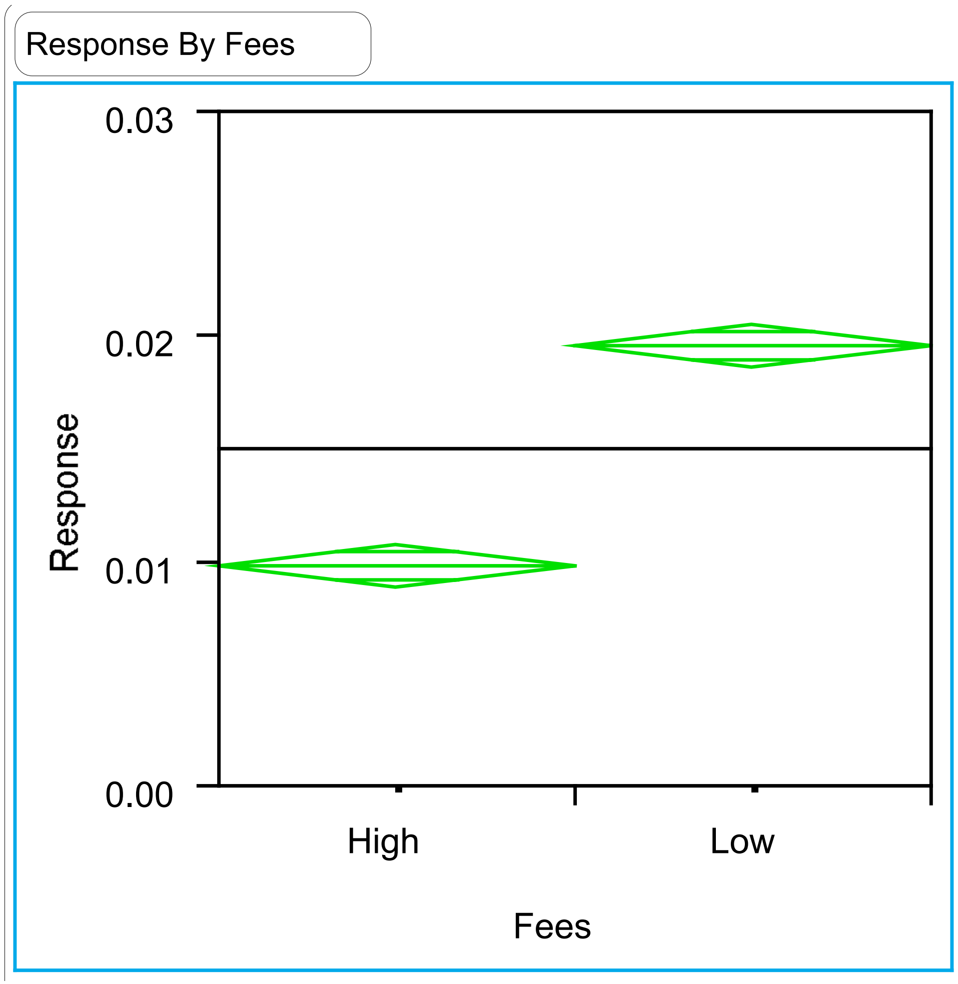
Data Mining – DOE Synergy

- Data Mining is exploratory
- Efforts can go on simultaneously
- Learning cycle oscillates naturally between the two

One at a Time Strategy

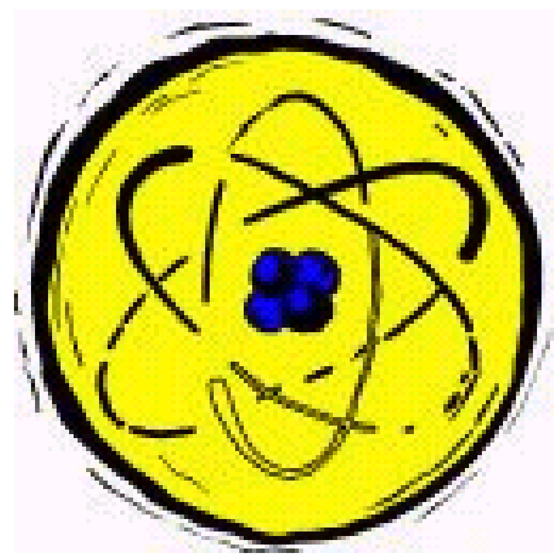
- Fixed Price
- Sent out 50,000 at Low Fee -- 50,000 at High Fee
- Estimated difference

Low Fees Gives 1% Lift



Chemical Plant

- Current product within spec 30% of the time
- 12,000 lbs/hour of product
- 30 years worth of data
- 6000 input variables
- Find model to optimize production



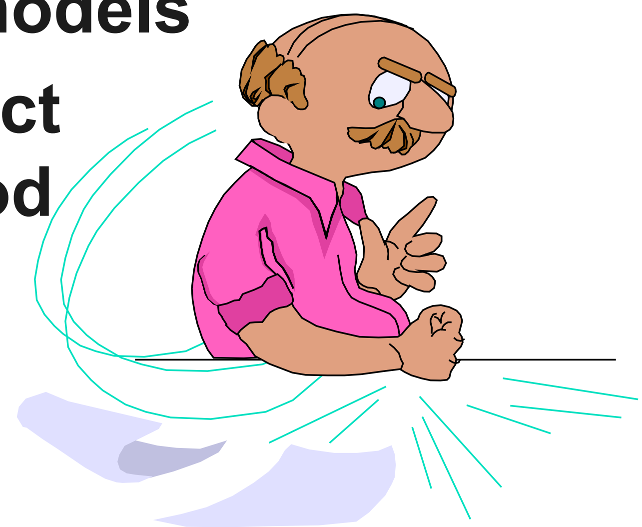
The Good News

- **We used 2 plants, 2 scenarios each**
  - 2 “good” runs, 2 “bad” runs each to maximize difference

- **Each of four models fit very well - R^2 over 80%**

The Bad News

- All four models were different
- No variables were the same
- The one variable known to be important (Methanol injection rate) didn't appear in models
- Models unable to predict outside their time period

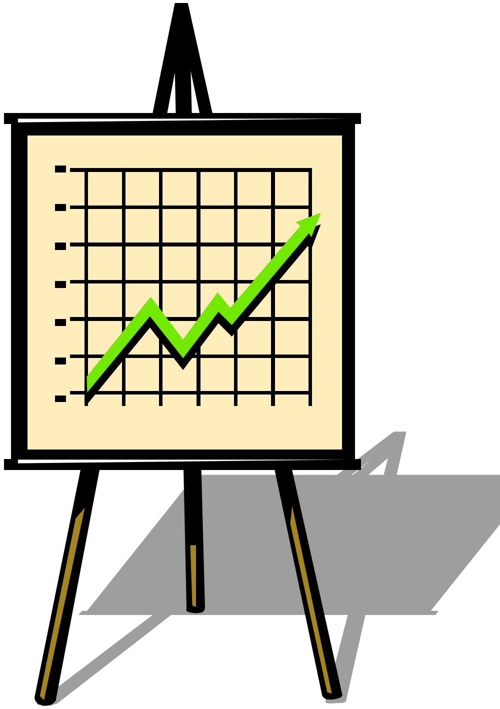


What really happened

- 6 months of incremental experimental design

- Increased specification percentage from 30% to 45%

- Profit increased $12,000,000/year



Challenges for data mining

- Not algorithms
- Overfitting
- Finding an interpretable model that fits reasonably well

Recap

- **Problem formulation**
- **Data preparation**
  - ➢Data definitions
  - ➢Data cleaning
  - ➢Feature creation, transformations
- **EDM – exploratory modeling**
  - ➢Reduce dimensions

Recap II

- Graphics
- Second phase modeling
- Testing, validation, implementation

Which Method(s) to Use?

- No method is best
- Which methods work best when?

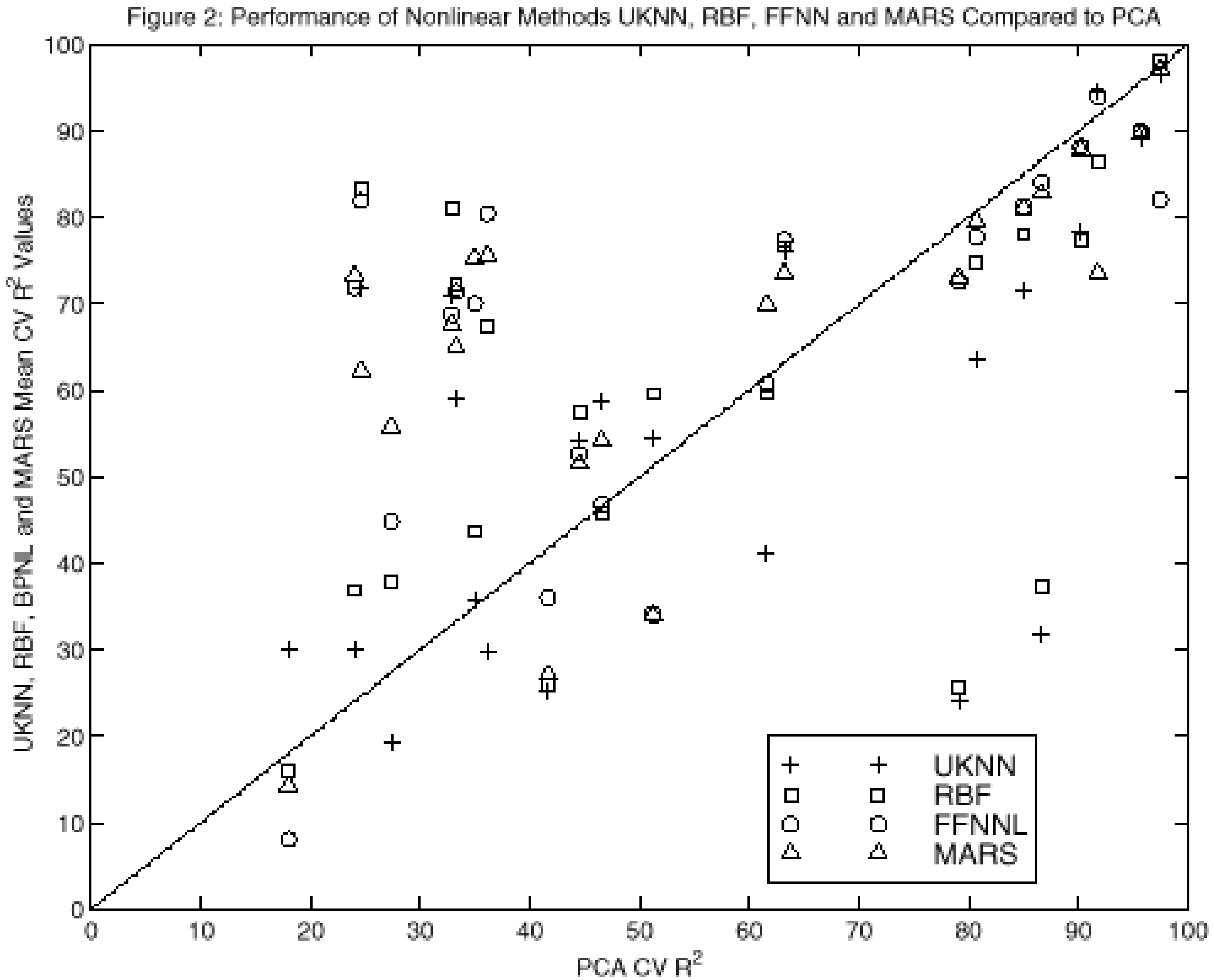
Competition Details

- **Ten real data sets from chemistry and chemical engineering literature**
  - No missing data
  - No replication of predictor points
- **Methodology**
  - Cross validated accuracy on predicted values (AVG RSS over CV samples)
  - Cross validation used to select parameters
    - ✓ Size of network, # of components for PCR

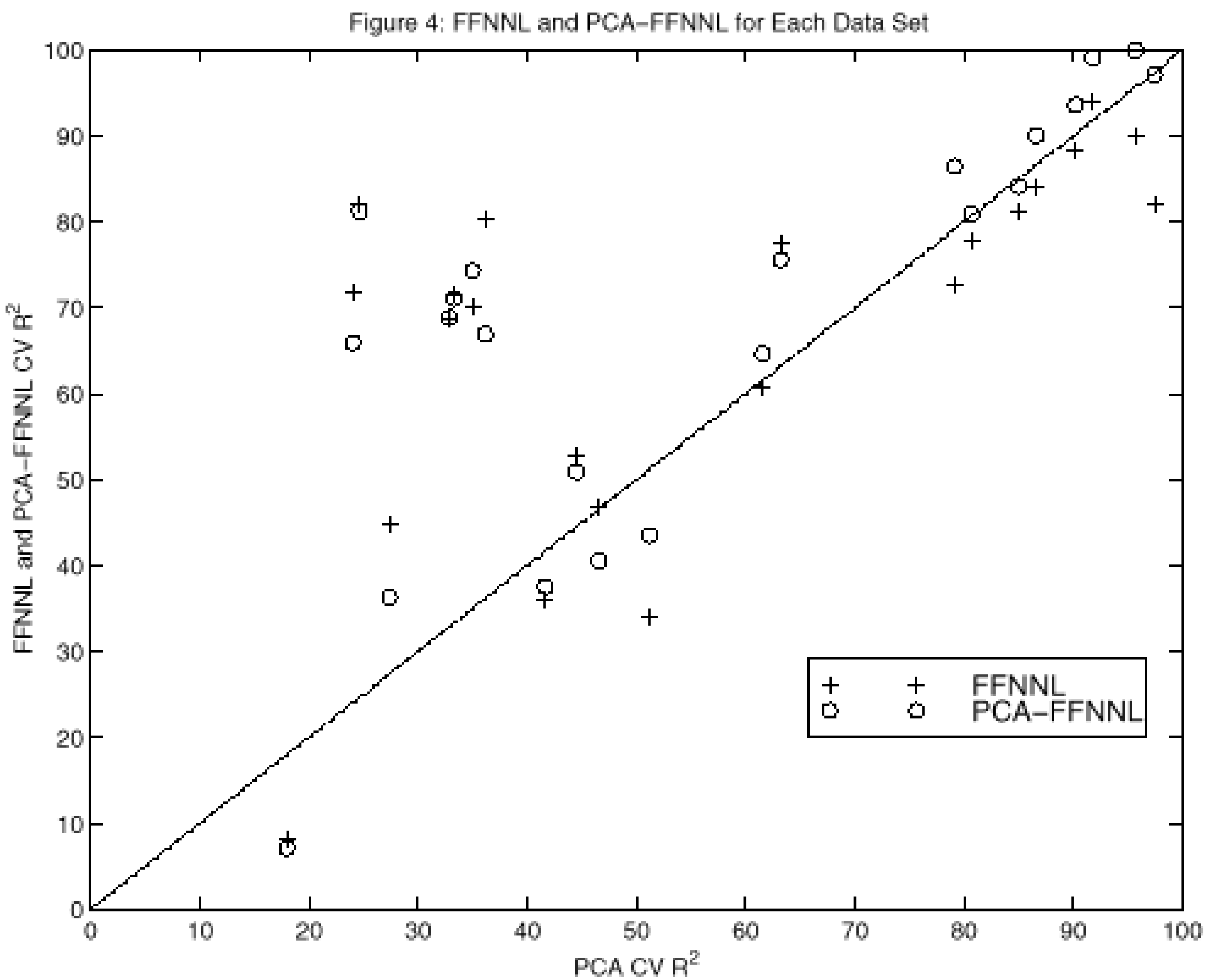
The Data Sets

Name	Description	n	p	r	VIF max
Gambino	chemical analysis	37	4	2	2.12
Benzyl	Chemical analysis	68	4	1	3.52
CIE	Wine testing for color	58	3	1	17.74
Periodic	Periodic table analysis	54	10	3	25.69
Venezia	Water quality analysis	156	15	1	147.62
Wine	Wine quality analysis	38	17	3	24.95
Polymer	Polymerization process	61	10	4	∞
3M	NIR for adhesive tape	34	219	2	∞
NIR	NIR for soybeans	60	175	3	∞
Runpa	NIR for composite material	45	466	2	∞

Nonlinear methods vs. best linear



Hybrid methods



Summary of Results

- **Many data sets one encounters in Chemometrics are inherently linear**
  - Linear methods work well for these!
  - Hence the historical success and popularity of such methods as PLS
- **When data sets are linear, CV R2 > .70, non-linear methods perform *worse* than linear methods**
- **But when CV R2<.40, non-linear methods may perform much better**

Summary Continued

- Hybrid methods -- those starting with linear (*e.g.* PCR) and then using a nonlinear method on the residuals always does well

Recommendations

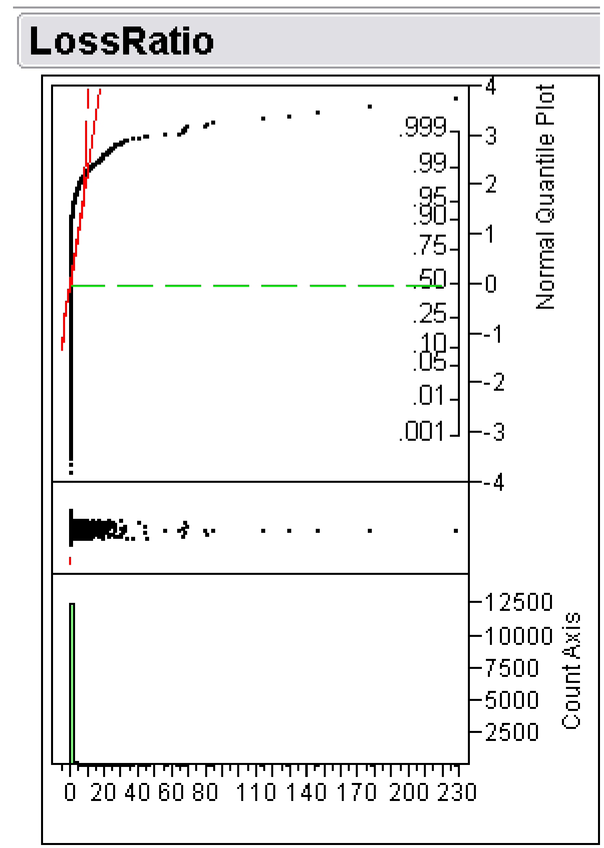
- **Start linear**
- **Assess linearity -- CV R2?**
- **Consider a nonlinear method**
  - Black box -- RBF NN or FF nn
  - Opaque -- MARS, KNN?
- **Consider the nonlinear method on the residuals from the linear method**
- **Cross validate!**

Case Study III
Church Insurance

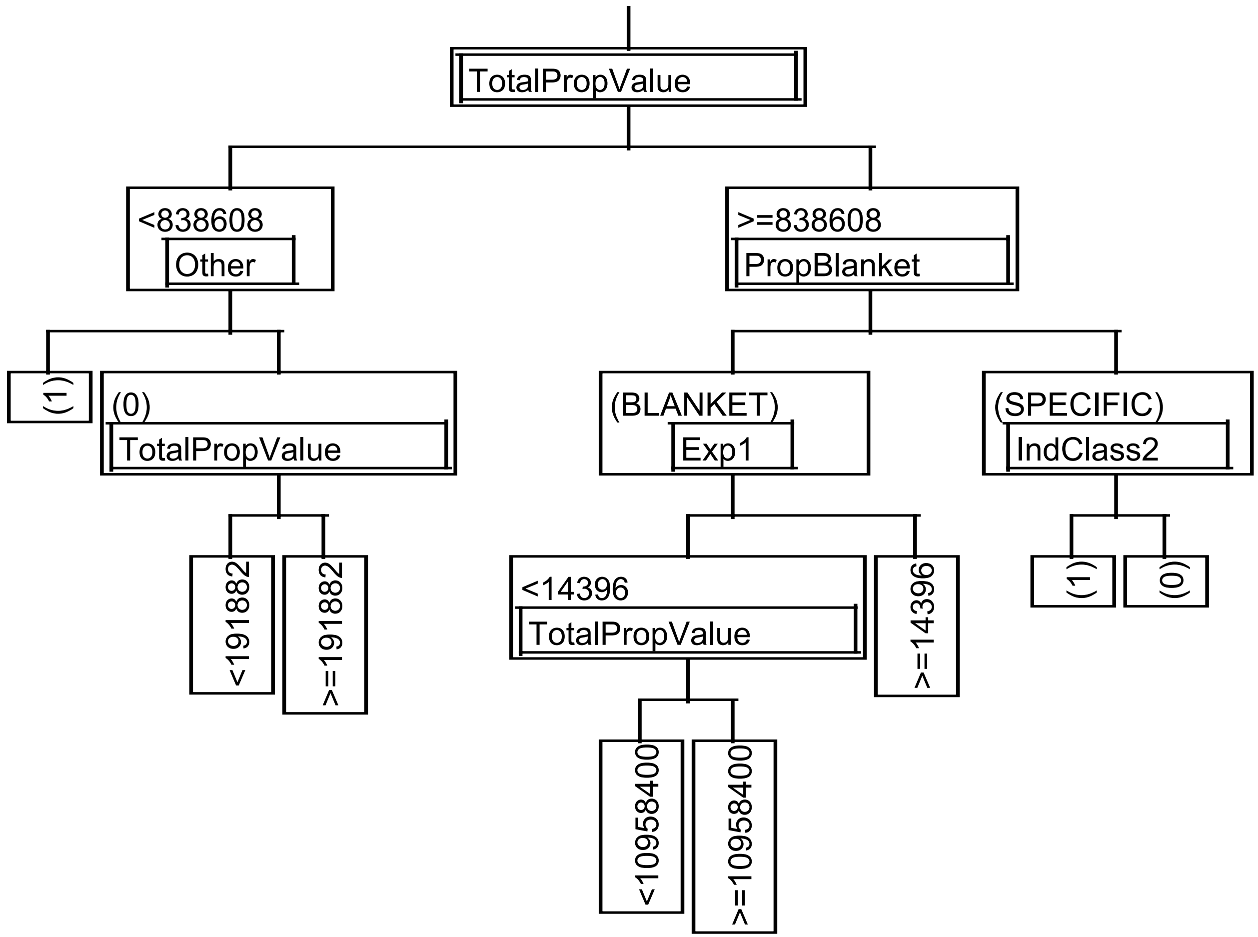
- **Loss Ratio for church policy**
  - Some Predictors
    - ✓ Net Premium
    - ✓ Property Value
    - ✓ Coastal
    - ✓ Inner100 (a.k.a., highly-urban)
    - ✓ High property value Neighborhood
    - ✓ Indclass1 (Church/House of worship)
    - ✓ Indclass2 (Sexual Misconduct – Church)
    - ✓ Indclass3 (Add'l Sex. Misc. Covg Purchased)
    - ✓ Indclass4 (Not-for-profit daycare centers)
    - ✓ Indclass5 (Dwellings – One family (Lessor's risk))
    - ✓ Indclass6 (Bldg or Premises – Office – Not for profit)
    - ✓ Indclass7 (Corporal Punishment – each faculty member)
    - ✓ Indclass8 (Vacant land- not for profit)
    - ✓ Indclass9 (Private, not for profit, elementary, Kindergarten and Jr. High Schools)
    - ✓ Indclass10 (Stores – no food or drink – not for profit)
    - ✓ Indclass11 (Bldg or Premises – Bank or office – mercantile or manufacturing – Maintained by insured (lessor's risk) – not for profit)
    - ✓ Indclass12 (Sexual misconduct – diocese)

Churches – First Steps

- **Select Test and Training sets**
- **Look at data**
  - ➢ Transform Loss Ratio?
  - ➢ Categorize Loss Ratio?
  - ➢ Outliers
- **Tree**



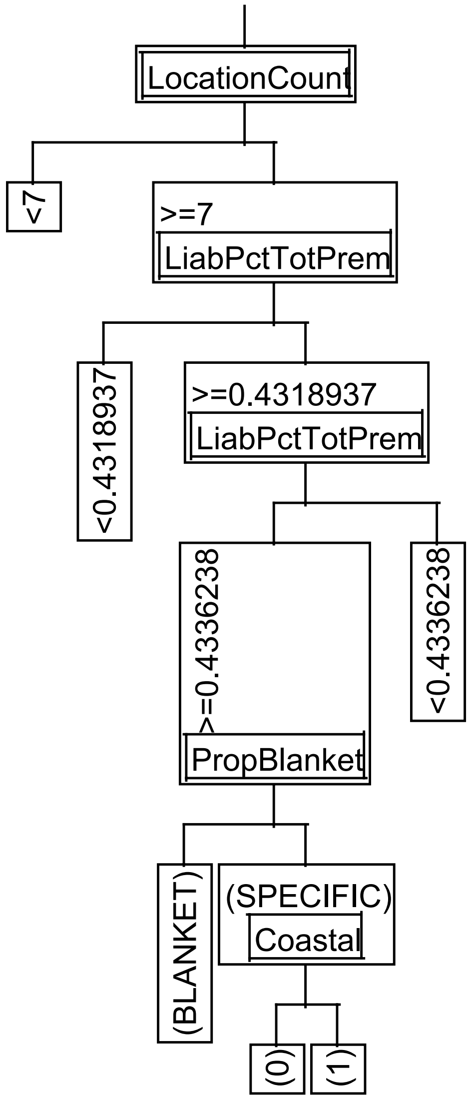
First Tree


TotalPropValue
<838608
Other
>=838608
PropBlanket
(1)
(0)
TotalPropValue
(BLANKET)
Exp1
(SPECIFIC)
IndClass2
<191882
>=191882
<14396
TotalPropValue
>=14396
(1)
(0)
<10958400
>=10958400

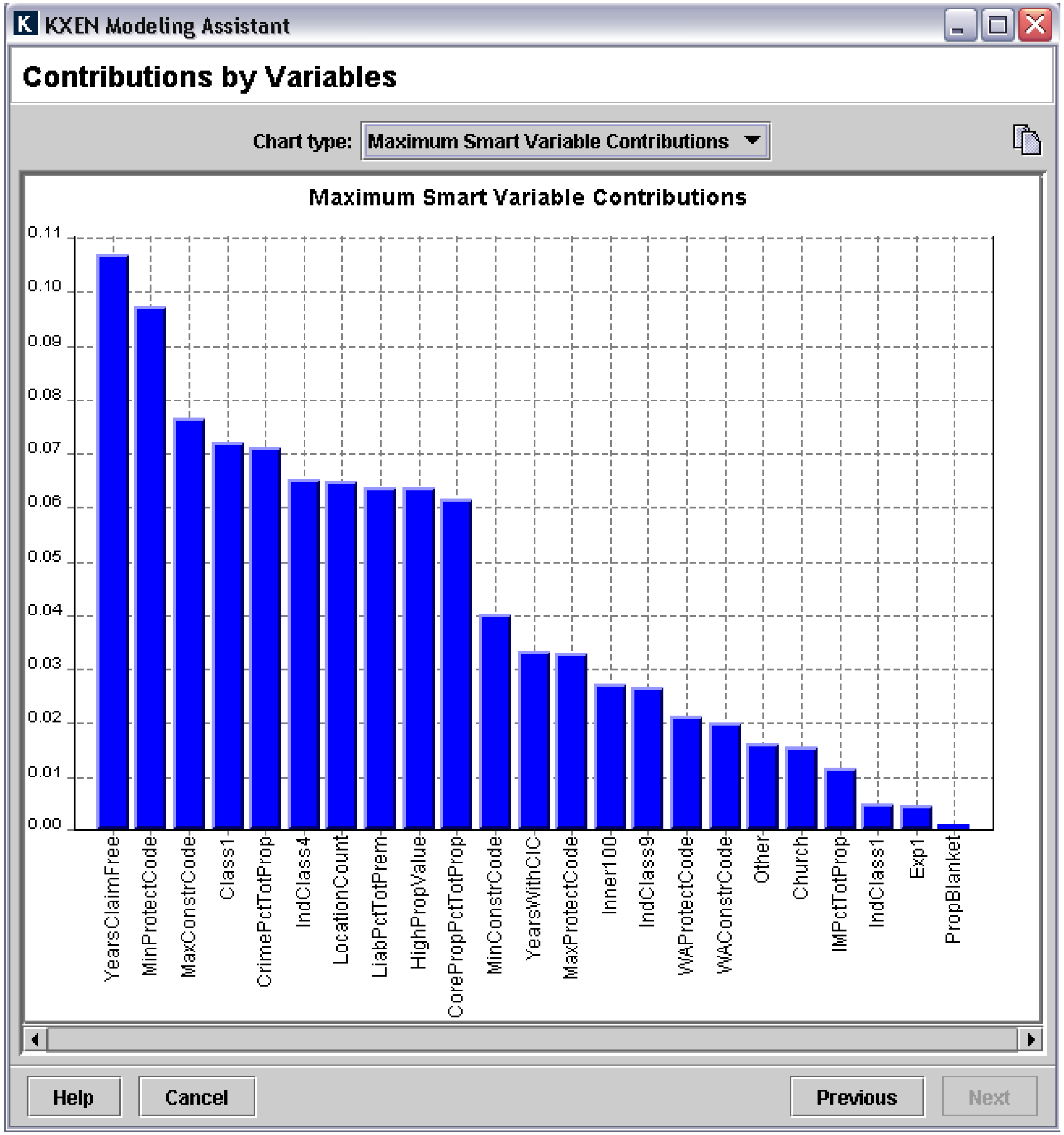
Unusable Predictors

- Size of policy not of use in determining likely high losses
- Decided to eliminate all policy size predictors

Next Tree

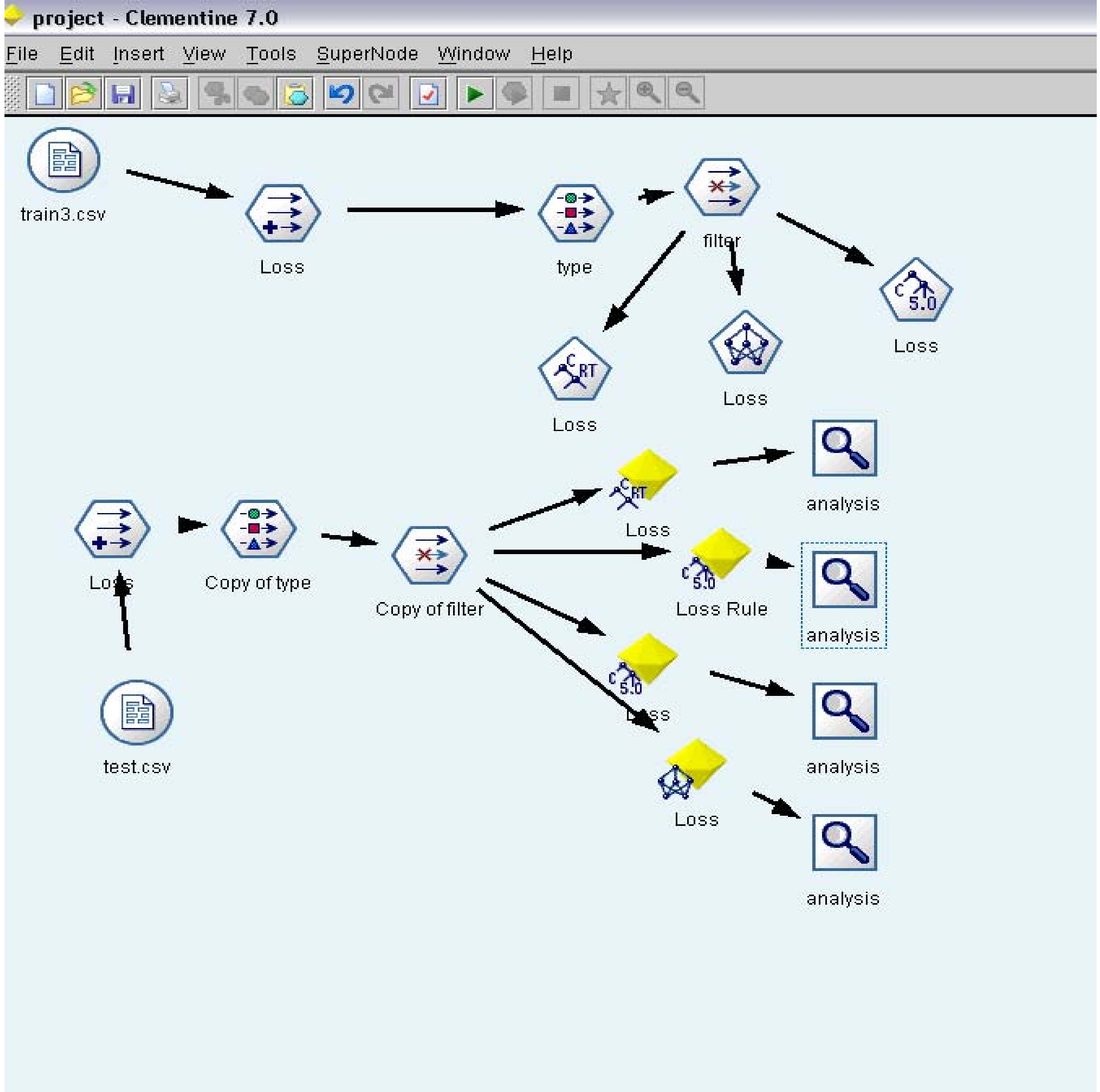


Where to go from here?


KXEN Modeling Assistant
Contributions by Variables
Chart type: Maximum Smart Variable Contributions
Maximum Smart Variable Contributions
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
YearsClaimFree
MinProtectCode
MaxConstrCode
Class1
CrimePctTotProp
IndClass4
LocationCount
LiabPctTotPrem
HighPropValue
CorePropPctTotProp
MinConstrCode
YearsWithCIC
MaxProtectCode
Inner100
IndClass9
WAProtectCode
WAConstrCode
Other
Church
IMPctTotProp
IndClass1
Exp1
PropBlanket
Help
Cancel
Previous
Next

Churches – Next Steps

- **Investigated**
  - Sources of missing
  - Interactions
  - Nonlinearities
- **Response**
  - Loss Ratio
  - Log LR
  - Categories
  - 0-1
  - Direct Profit
  - Two Stage – Loss and Severity


project - Clementine 7.0
File Edit Insert View Tools SuperNode Window Help
train3.csv
Loss
type
filter
Loss
Loss
Loss
analysis
Loss
Copy of type
Copy of filter
Loss
Loss Rule
analysis
analysis
Loss
test.csv
Loss
analysis

Working Model

- **Outliers influenced any assessment of expected loss ratio severely**
- **Eliminated top 15 outliers from test and train**
- **Randomly assigned train and test multiple times**
  - Two stage model
    - ✓ Positive loss (Tree)
    - ✓ Severity on positive cases (Tree)
- **Consistently identified top few percentiles of high losses**
- **Estimated savings in low millions of dollars/year**

Opportunities

- **Predictive model can tell us**
  - Who
  - What factors
- **Sensitivity analysis can help us even with black box models**
- **Causality?**
  - Experimental Design

Data Mining Tools

- **Software for specific methods**
  - Neural nets or trees or regression or association rules
- **General tool packages**
- **Vertical package solutions**

General Data Mining Tools

- **Examples**
  - SAS: Enterprise Miner
  - SPSS: Clementine
- **Characteristics**
  - Neural nets, decision trees, nearest neighbors, etc
  - Nice GUI
  - Assume data is clean and in a nice tabular form

State of the Market

- **Good products are available**
  - ➢ Strong model building
  - ➢ Fair deployment
  - ➢ Poor data preparation (except KXEN)
- **Products differ in size of data sets they handle**
- **Performance often depends on undocumented feature selection**

Next Steps

- **Time to start getting experience**
  - Develop a strategy
  - Set up a research group
  - Select a pilot project
    - ✓ Control scope
    - ✓ Minimize data problems
    - ✓ Real value to solution but not mission critical
- **Communication**
  - Statisticians as partners
  - Statisticians as consultants and teachers
- **Enormous opportunity**

Take Home Messages I

- **You have more data than you think**
  - ➢Save it and use it
  - ➢Let non-statisticians use it
- **Data preparation is most of the work**
- **Dealing with missing values**

Take Home Messages II

- **What to do first?**
  - Use a (tree)
- **Which algorithm to use?**
  - All– this is the fun part, but beware of overfitting
- **Results**
  - Keep goals in mind
  - Test models in real situations

For More Information

- **Two Crows**
  - http//www.twocrows.com
- **KDNuggets**
  - http://www.kdnuggets.com

M. Berry and G. Linoff, **Data Mining Techniques**,John Wiley, 1997

J. Friedman, T. Hastie and R. Tibshirani, **The Elements of Statistical Learning,** Springer-Verlag, 2001

U. Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, **Advances in Knowledge Discovery and Data Mining**, MIT press, 1996

Dorian Pyle, **Data Preparation for Data Mining**, Morgan Kaufmann, 1999

C. Westphal and T. Blaxton, **Data Mining Solutions**, John Wiley, 1998

Vasant Dhar and Roger Stein, **Seven methods for transforming corporate data into business intelligence**, Prentice Hall 1997

David J. Hand, H. Mannila, P. Smyth , **Principles of Data Mining** , MIT Press, 2001