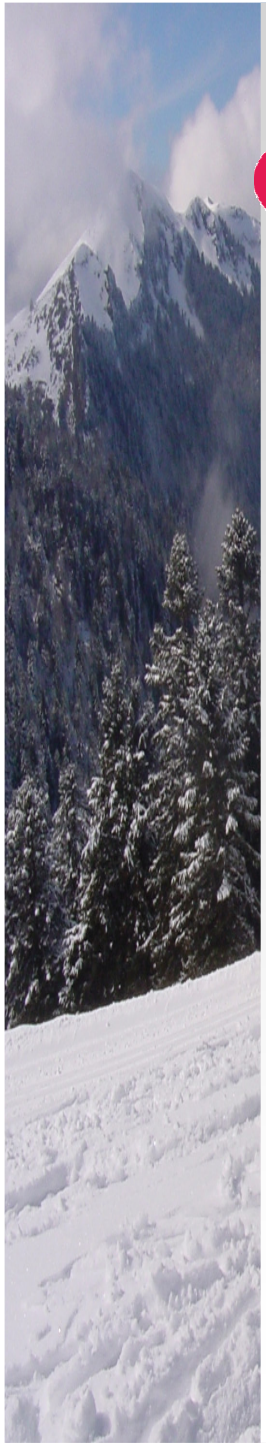
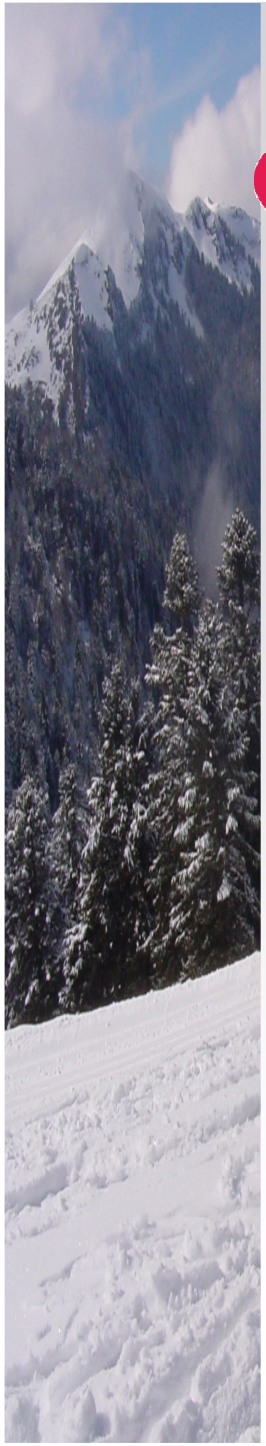
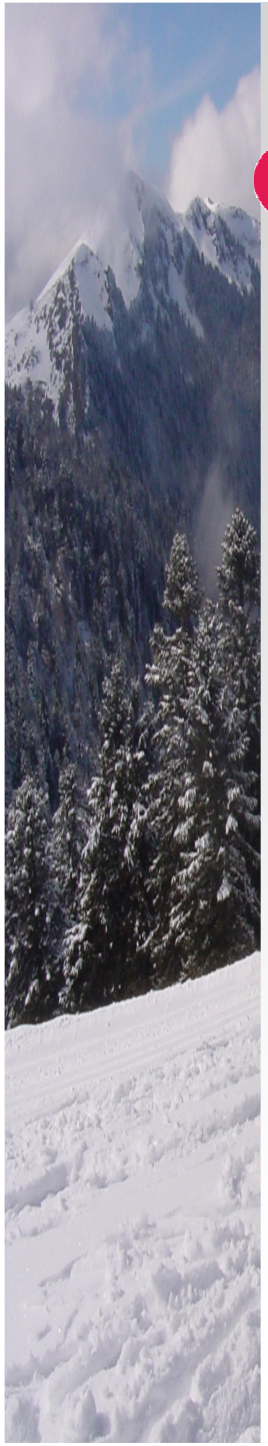


Successful Data Mining in Practice: Where Do We Start?



Outline

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



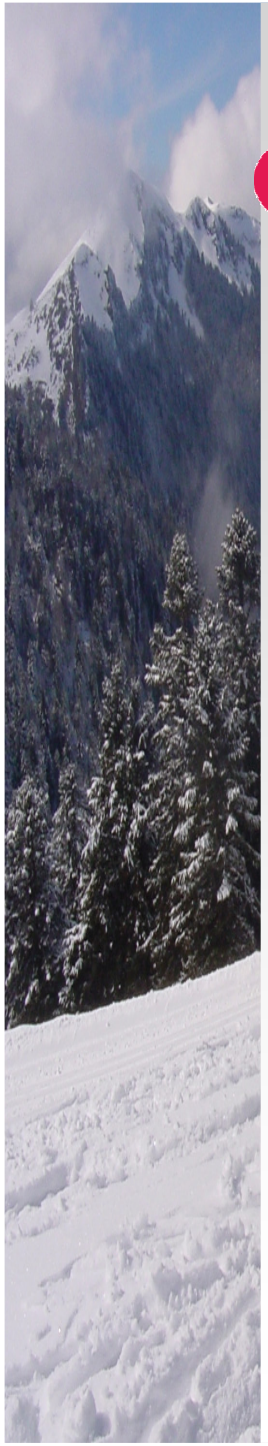
Data Mining Is…

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

"the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" D.Hand

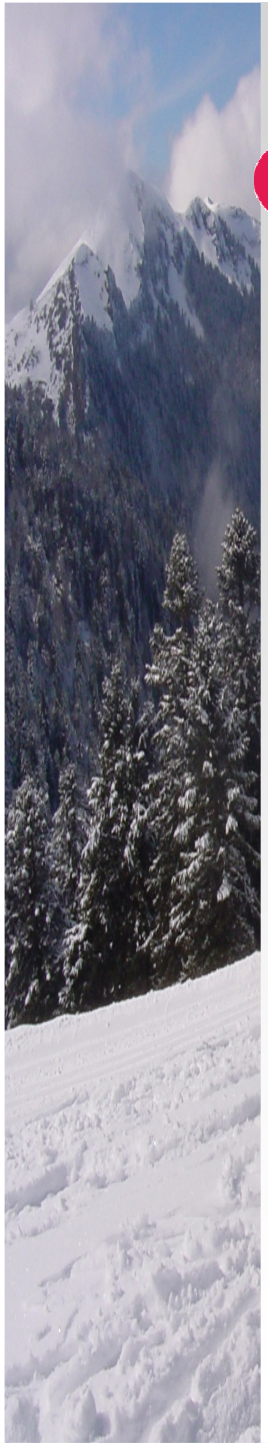
"a knowledge discovery process of extracting previously unknown, actionable information from very large data bases"--- Zornes

" 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



Paralyzed Veterans of America

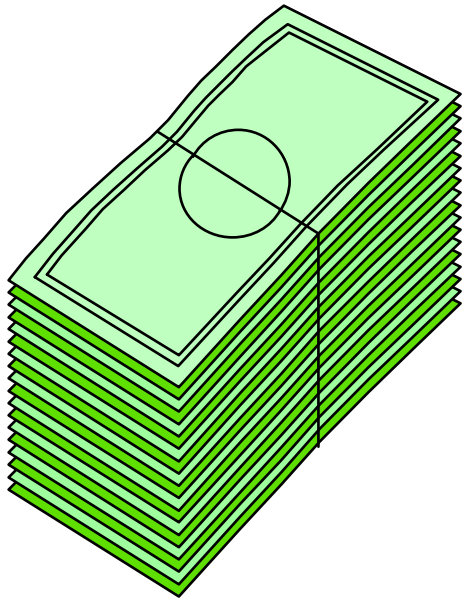
- KDD 1998 cup
- Mailing list of 3.5 million potential donors
- Lapsed donors
  - Made their last donation to PVA 13 to 24 months prior to June 1997
  - 200,000 (training and test sets)
- Who should get the current mailing?
- Cost effective strategy?

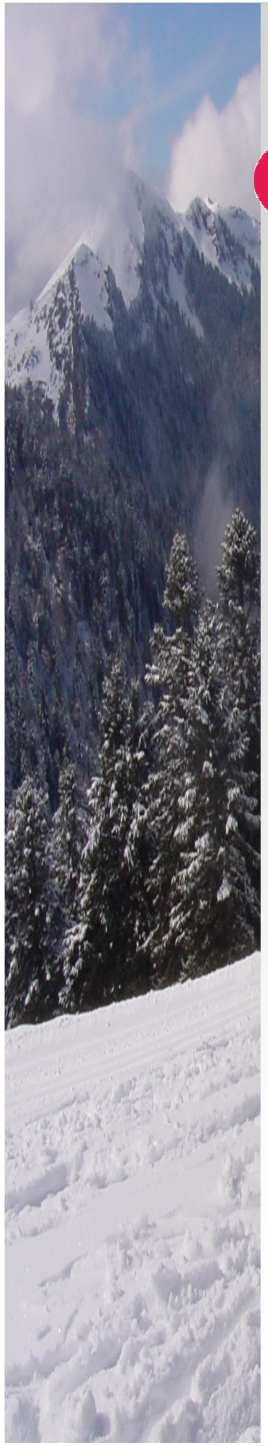


Results for PVA Data Set

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

- Using data mining techniques, this was increased 41.37%



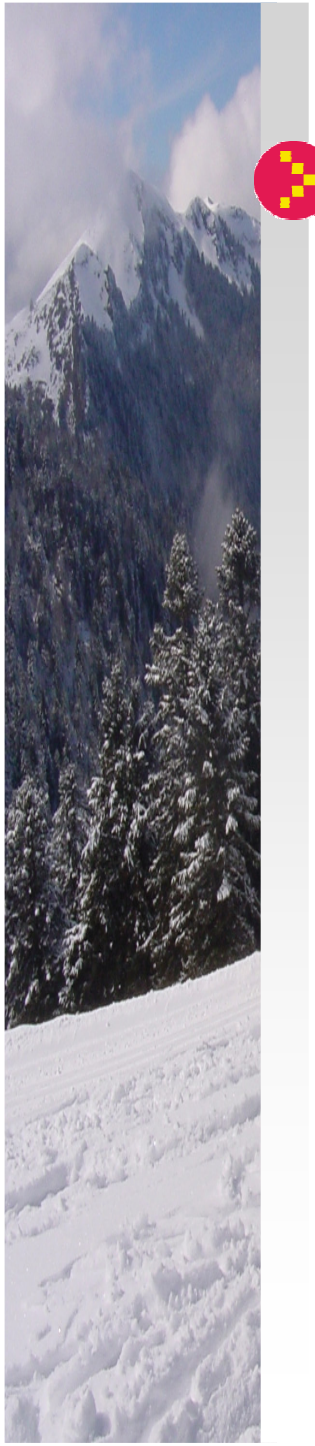


KDD CUP 98 Results

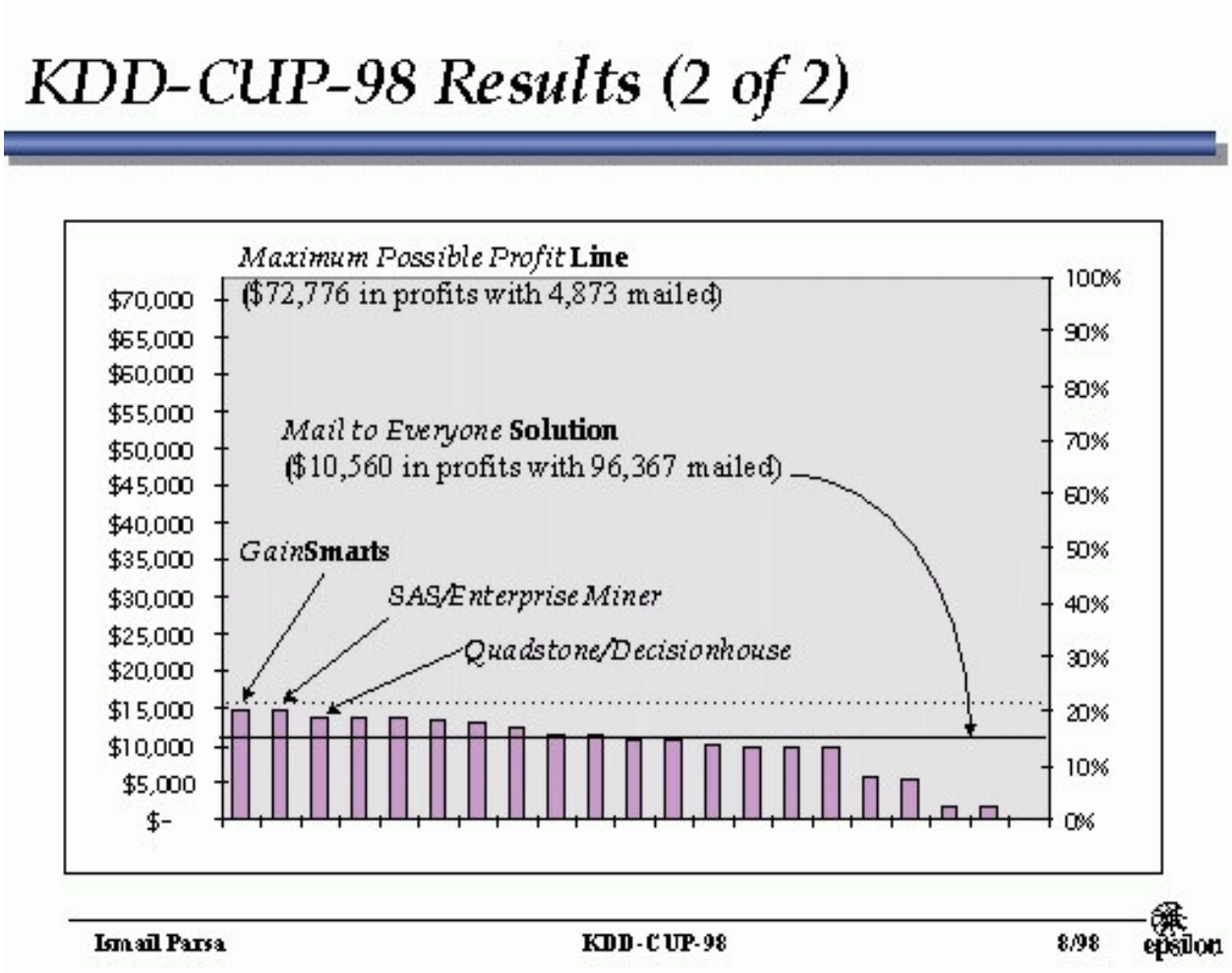
KDD-CUP-98 Results (1 of 2)

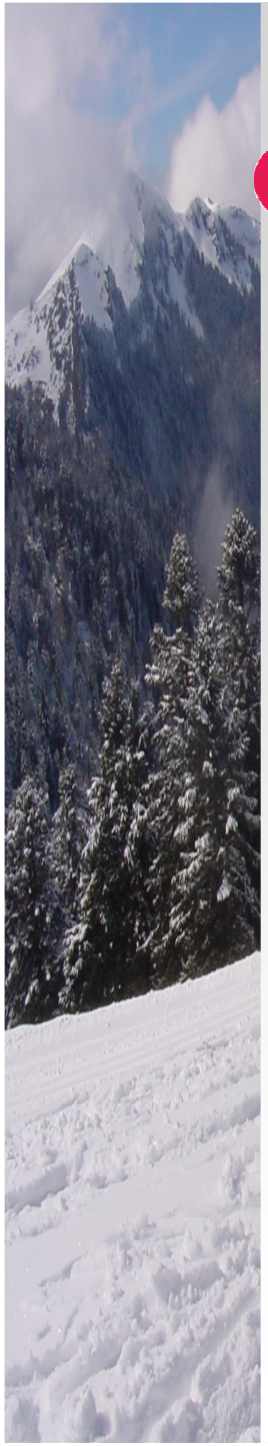
Participants	Sum of Actual Profits	Number Mailed	Average Profits
GainSmarts	$ 14,712.24	56,330	0.26
SAS/Enterprise Miner	$ 14,662.43	55,838	0.26
Quadstone/Decisionhouse	$ 13,954.47	57,836	0.24
# 4	$ 13,824.77	55,650	0.25
# 5	$ 13,794.24	51,906	0.27
# 6	$ 13,598.05	55,830	0.24
# 7	$ 13,040.46	60,901	0.21
# 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



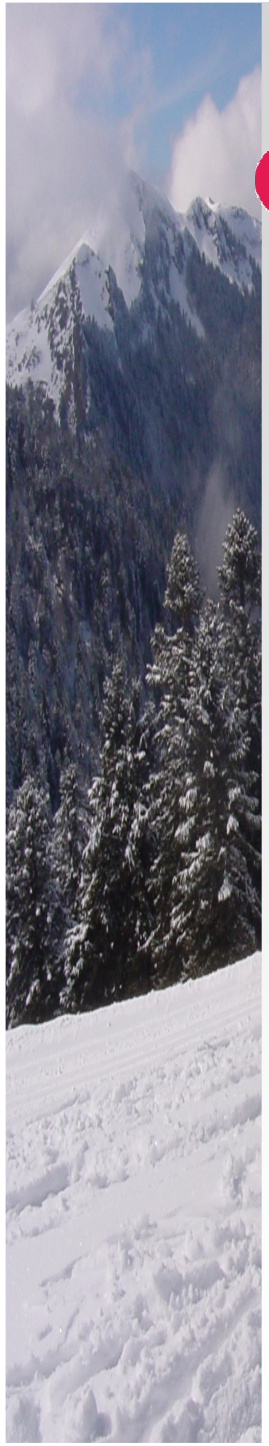
KDD CUP 98 Results 2





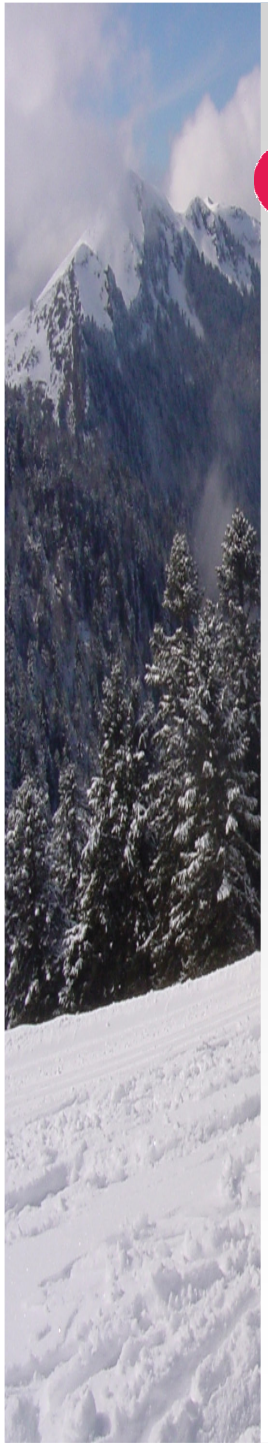
Reason for Data Mining

Data = $$



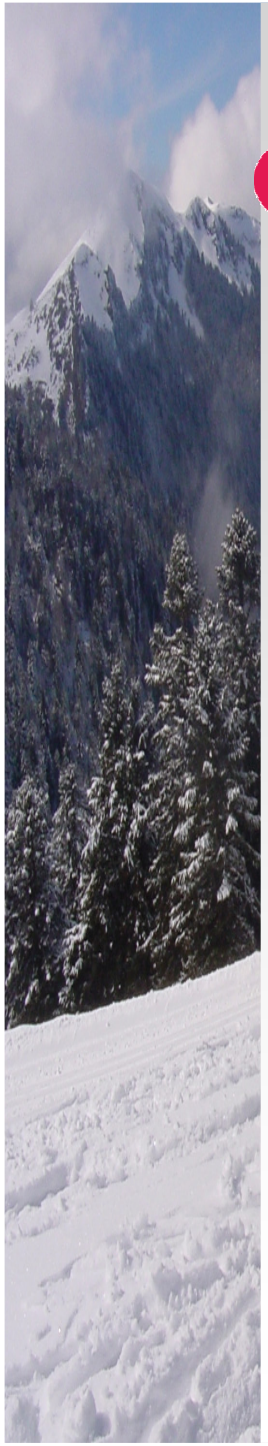
Case Study

- 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 environme



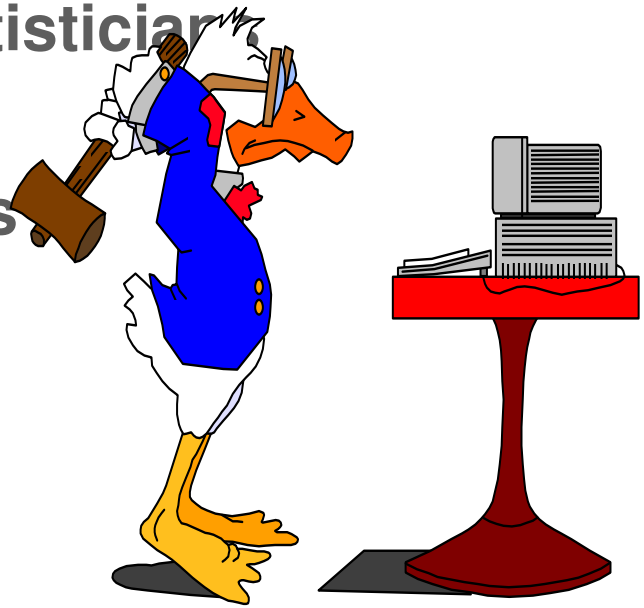
Case Study II -- Car Insurance

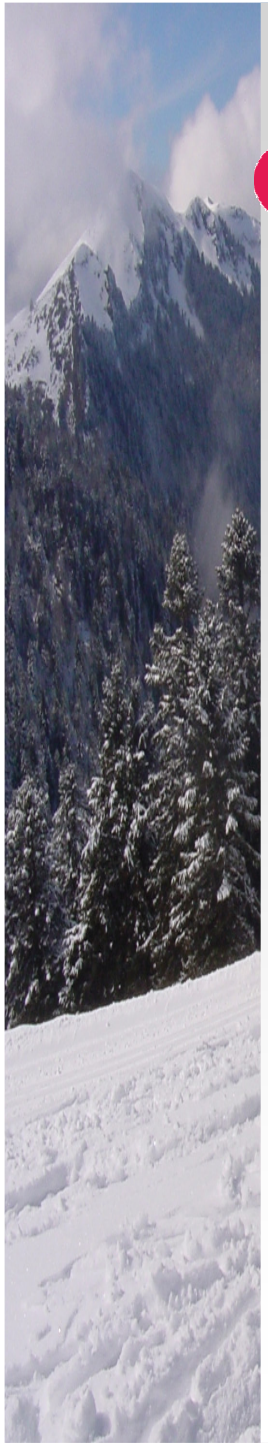
- 42800 mature policies
- 65 potential predictors
  - **Tree model found industry, vehicle age, numbers of vehicles, usage and location**



What's Different?

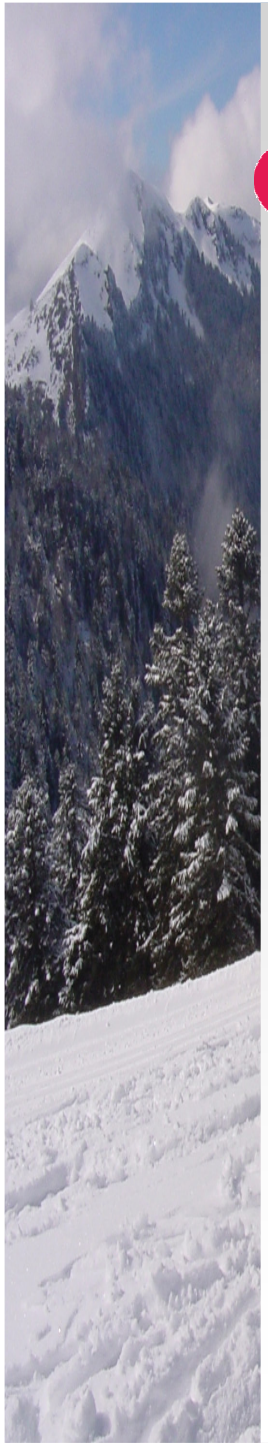
- Users
  - **Domain experts, not statisticians**
  - **Have too much data**
  - **Want *automatic* methods**
  - **Want useful information**
- Problem size
  - **Number of rows**
  - **Number of variables**





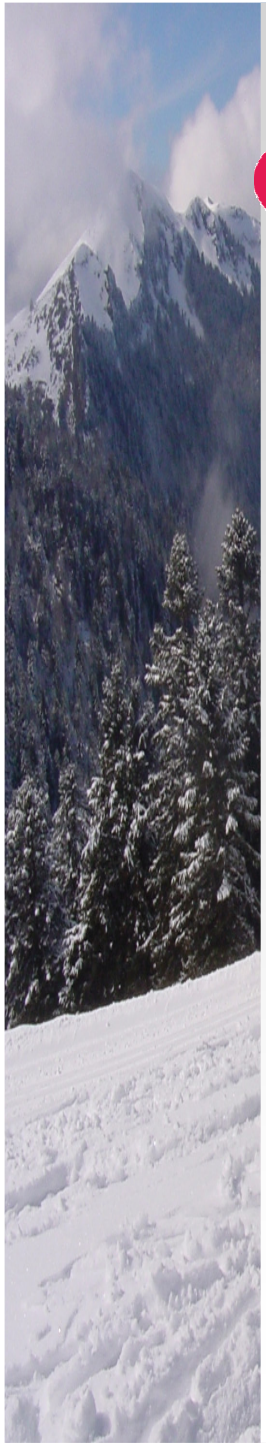
Data Mining Myths

- Data mining tools need no guidance.
- Data mining models explain behavior.
- Data mining requires no data analysis skill.
- Data mining eliminates the need to understand your business and your data
- Data mining tools are “different” from statistics.



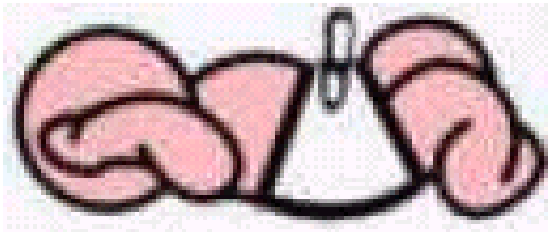
Associations

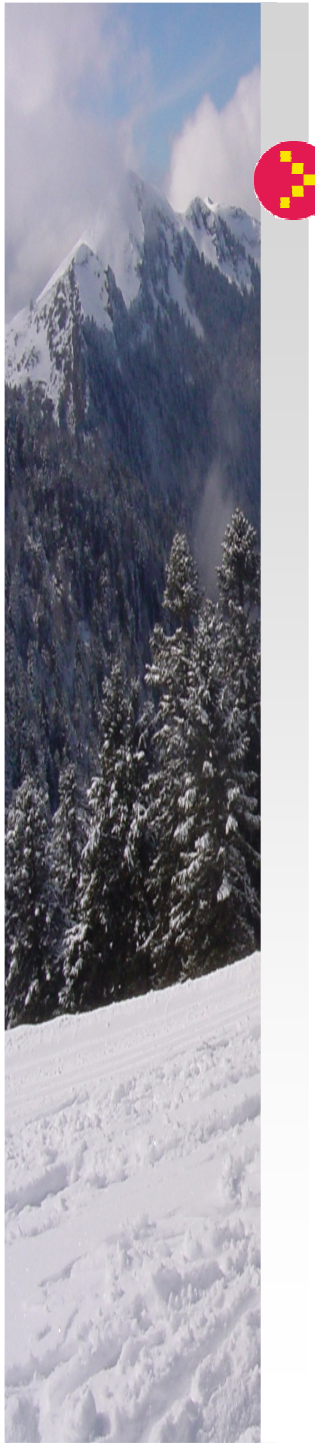
- If item A is part of an event, then x % of the time (the confidence factor) item B is part of the event.
  - If low fat cottage cheese and non-fat yogurt are bought, then 85% of the time skim milk is purchased.
  - If corn chips are purchased, 65% of the time cola is purchased, unless there is a promotion, in which case 85% of the time cola is purchased.
- Quiz: What grocery item's purchases is most highly associated with shampoo purchases?



Beer and Diapers

- Made up story?
- Unrepeatable -- happened once.
- Lessons learned?





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

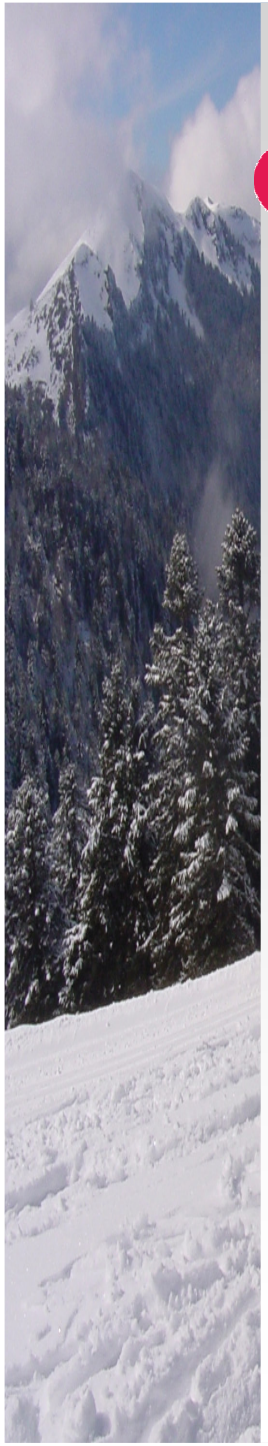
PowerPoint, what else?

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,…



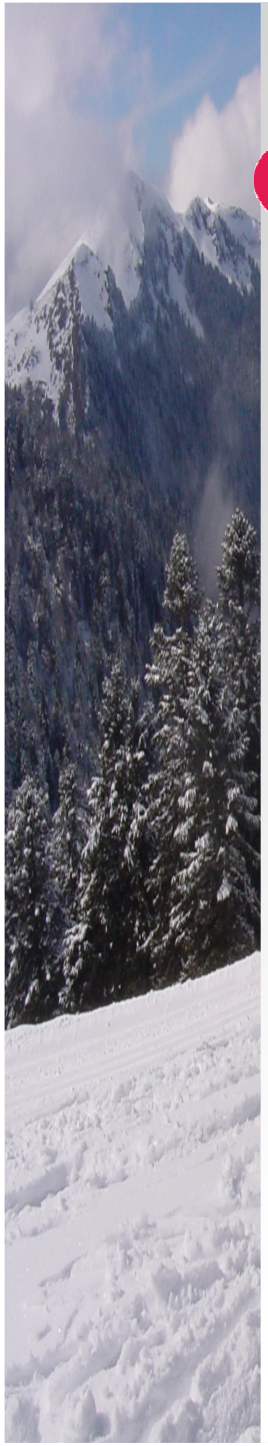
Data Mining vs. Statistics

- Flexible models
- Automatic
- Prediction often most important
- Computation matters

- Variable selection and overfitting are problems

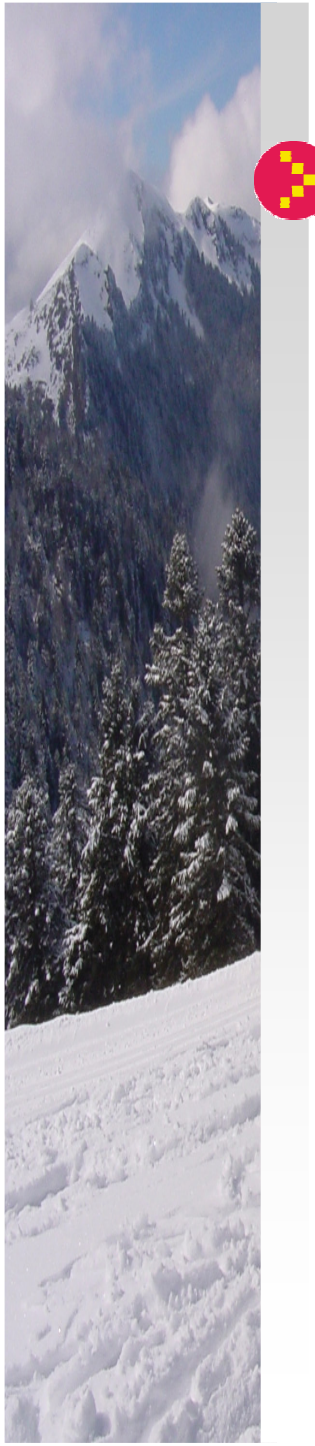
- Particular model and error structure
- Understanding, confidence intervals
- Computation not critical

- Variable selection and model selection are still problems

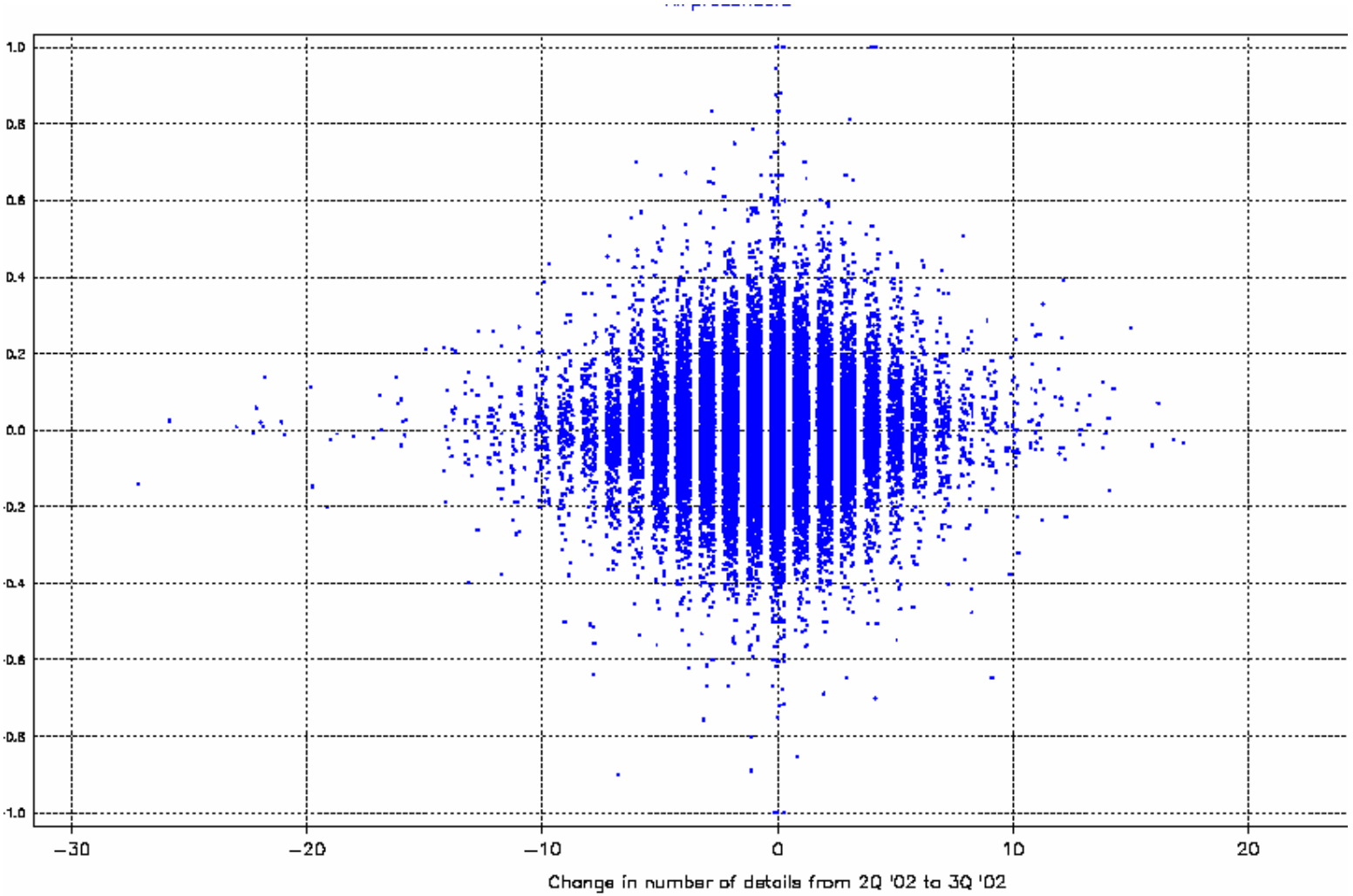


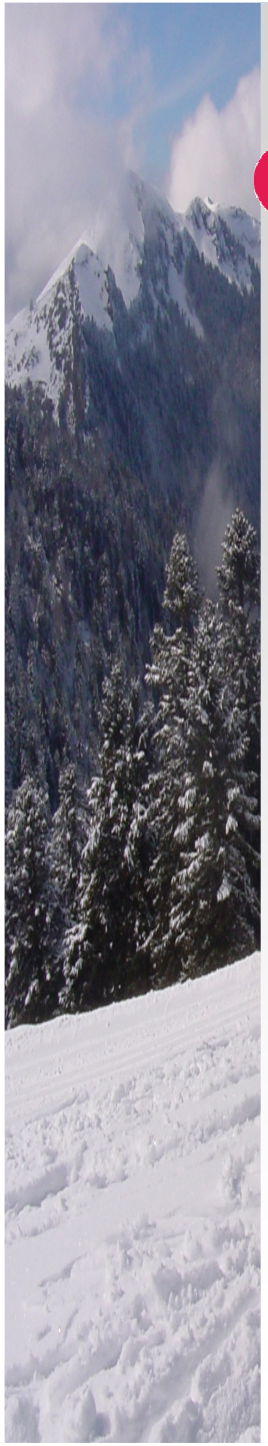
What's the Same?

- 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



Hidden trends?

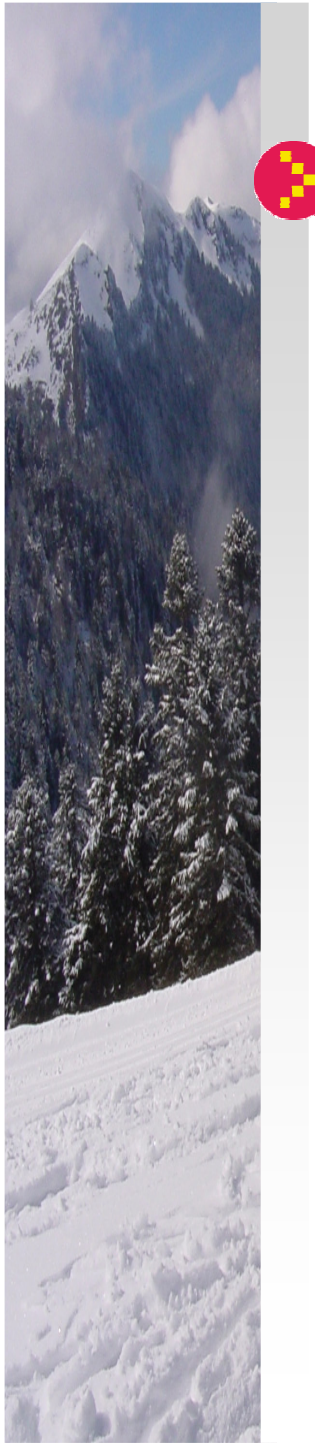




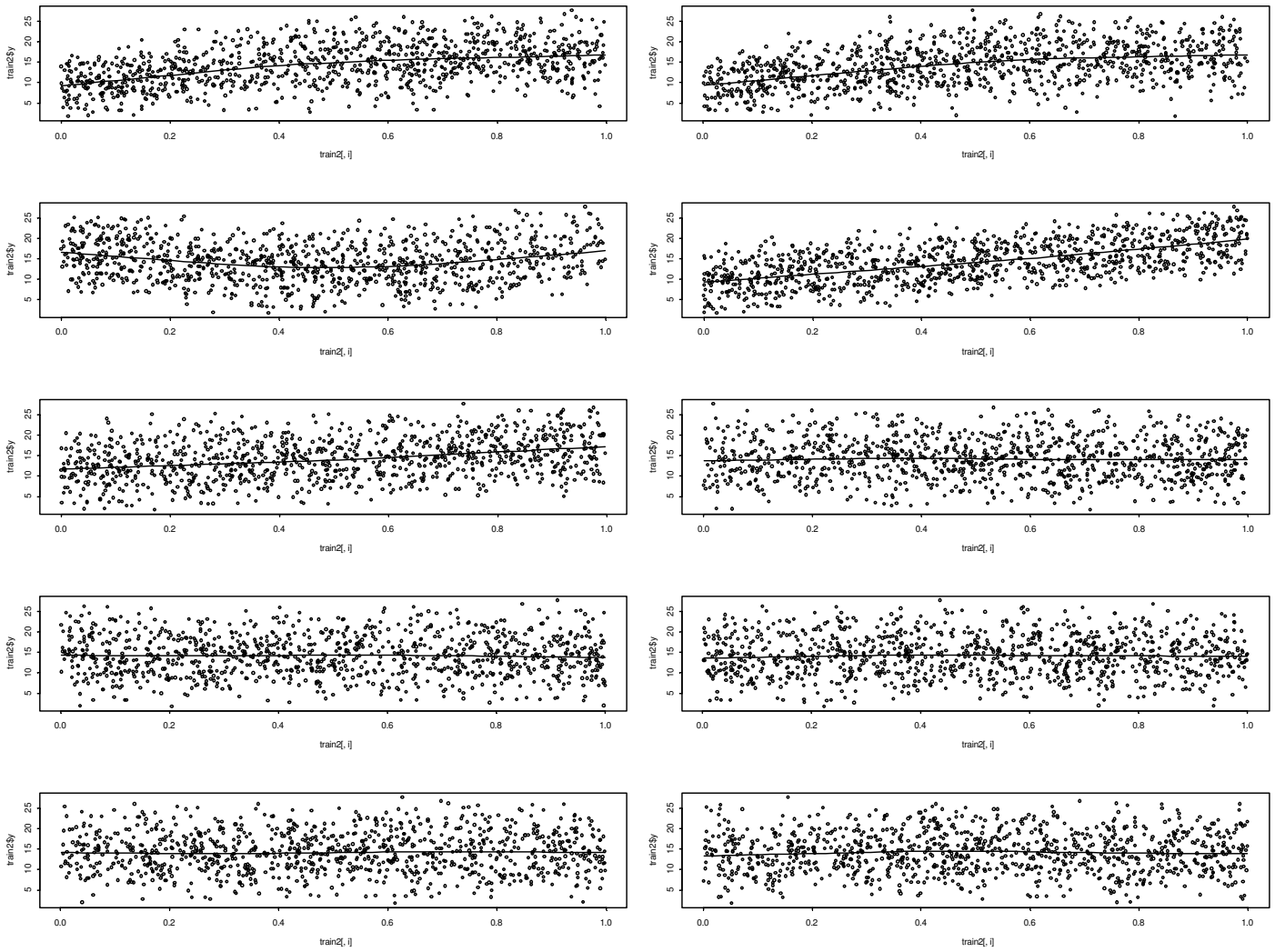
Data Preparation

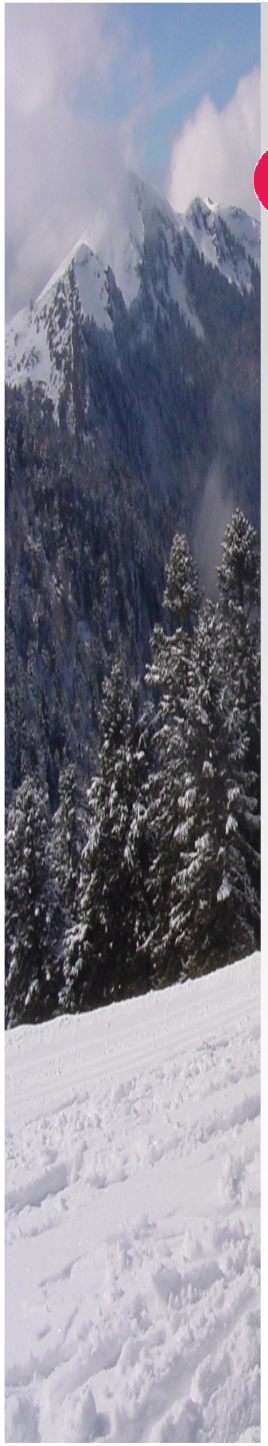
- Build data mining database
- Explore data
- Prepare data for modeling

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



“Toy” Problem

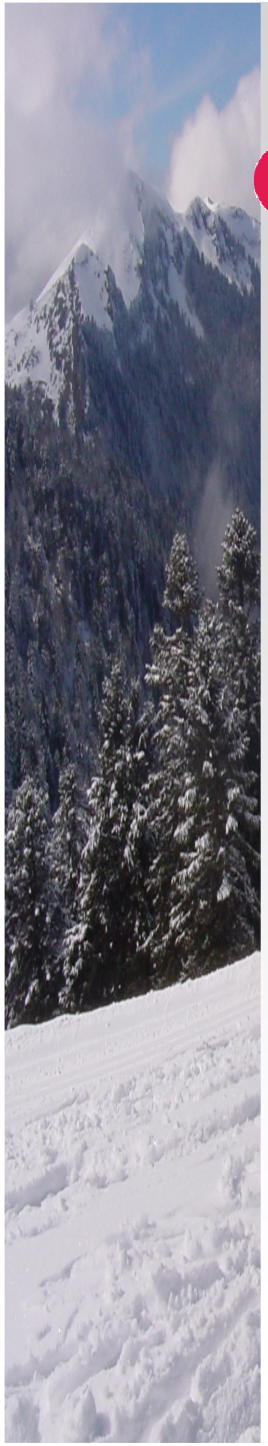




Linear Regression

Term	*Estimate*	*Std Error*	*t Ratio*	*Prob>|t|*
Intercept	0.806	0.427	1.890	0.059
x1	7.269	0.273	26.590	<.0001
x2	7.289	0.281	25.940	<.0001
x3	-0.719	0.287	-2.500	0.012
x4	9.769	0.273	35.810	<.0001
x5	4.834	0.275	17.590	<.0001
x6	-0.456	0.280	-1.630	0.104
x7	0.123	0.270	0.460	0.647
x8	-0.349	0.276	-1.270	0.206
x9	-0.578	0.285	-2.030	0.043
x10	0.080	0.280	0.280	0.777

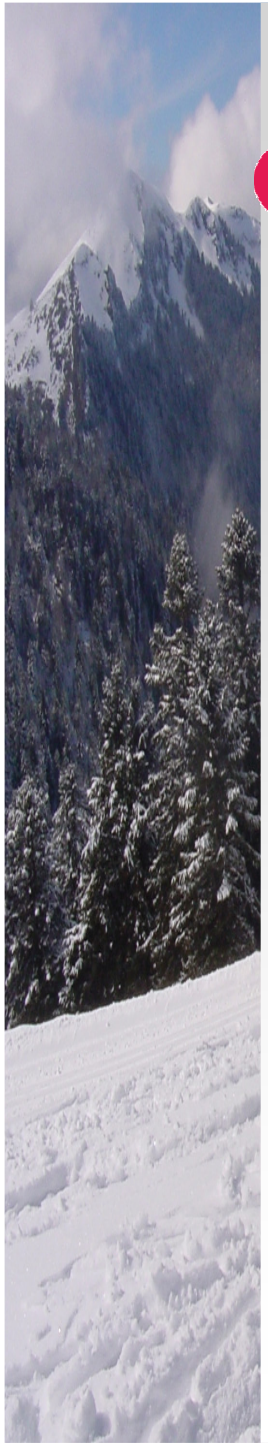
R-squared: 76.1% Train 73.3% Test



Stepwise Regression

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		0.561	0.328	1.710	0.087
x1		7.252	0.273	26.550	<.0001
x2		7.311	0.280	26.110	<.0001
x3		-0.767	0.286	-2.690	0.007
x4		9.747	0.272	35.790	<.0001
x5		4.799	0.274	17.510	<.0001
x9		-0.609	0.284	-2.140	0.032

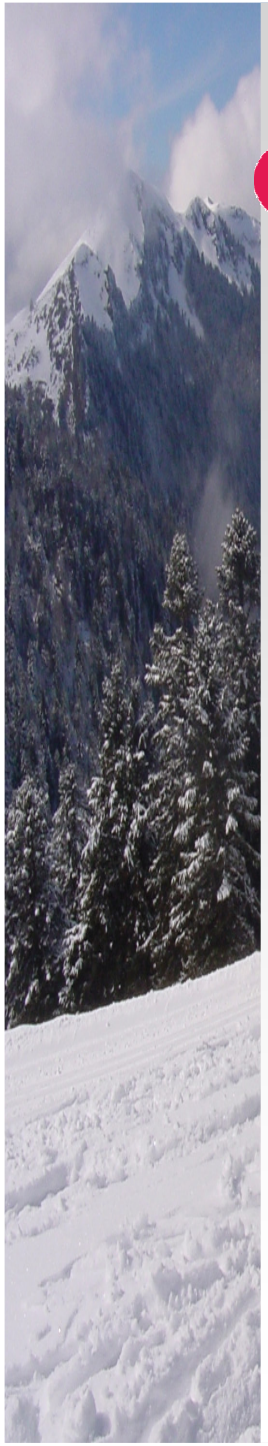
R-squared 76.0% on Train 73.4% Test



Stepwise 2ND Order Model

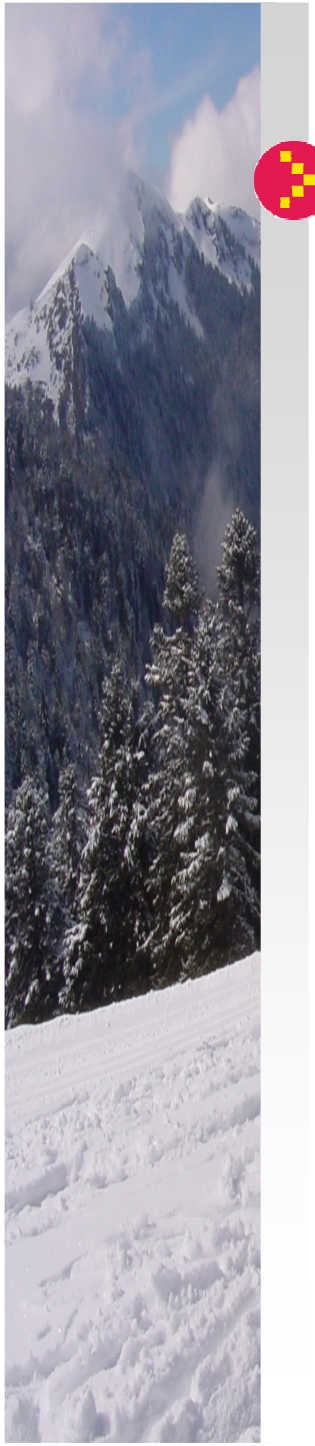
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.000	0.000	.	.
x1	7.204	0.169	42.510	<.0001
(x1-0.49573)*(x1-0.49573)	-12.137	0.682	-17.790	<.0001
x2	7.313	0.173	42.380	<.0001
(x2-0.48895)*(x2-0.48895)	-11.289	0.688	-16.410	<.0001
x3	-1.010	0.179	-5.660	<.0001
(x3-0.46706)*(x3-0.46706)	20.658	0.703	29.390	<.0001
x4	10.169	0.172	59.070	0.000
x5	5.135	0.168	30.610	<.0001
(x5-0.49425)*(x5-0.49425)	1.714	0.694	2.470	0.014
x7	0.244	0.165	1.480	0.140
x8	0.079	0.171	0.460	0.646
(x1-0.49573)*(x2-0.48895)	2.370	0.639	3.710	0.000
(x2-0.48895)*(x4-0.49038)	-0.322	0.626	-0.510	0.607
(x3-0.46706)*(x7-0.4962)	1.273	0.626	2.030	0.042
(x4-0.49038)*(x8-0.4975)	-1.015	0.603	-1.680	0.092
(x7-0.4962)*(x8-0.4975)	-1.283	0.601	-2.130	0.033

R-squared 90.0% Train 88.5% Test

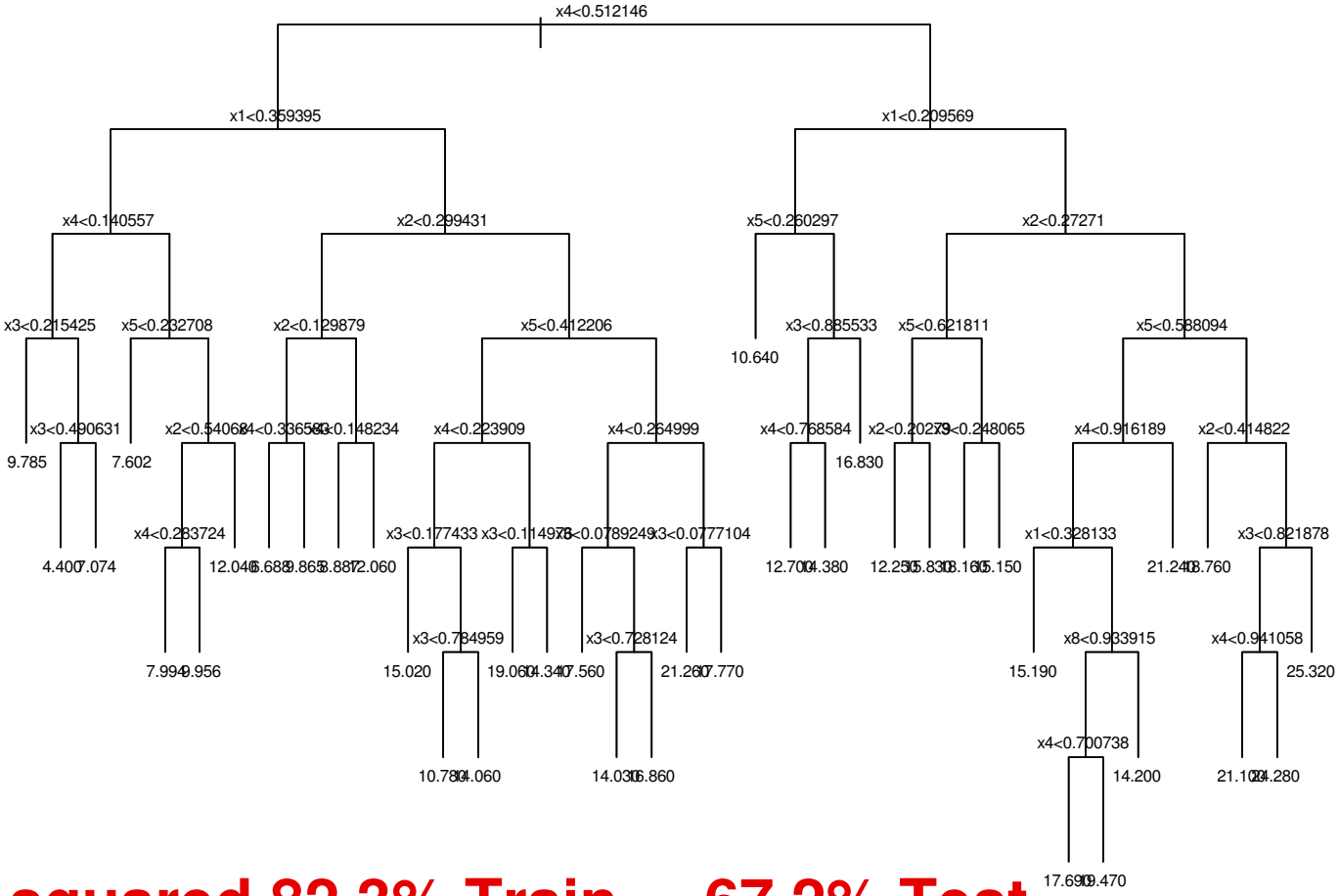


Next Steps

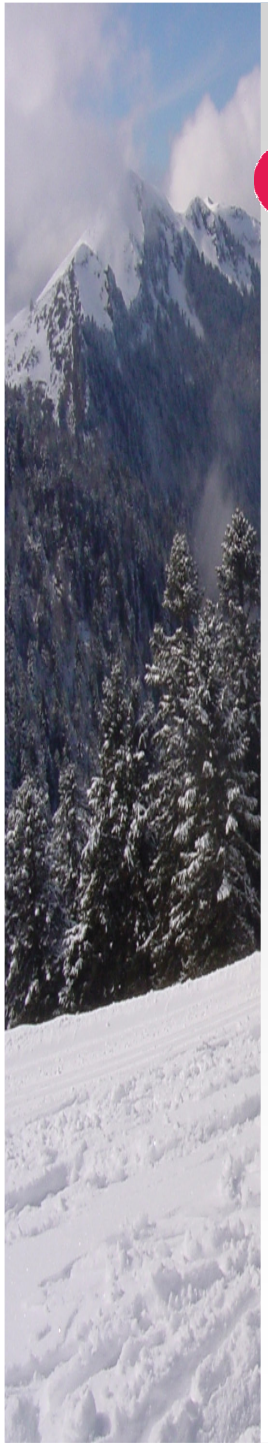
- Higher order terms?
- When to stop?
- Transformations?
- Too simple: underfitting – bias
- Too complex: inconsistent predictions, overfitting – high variance
- Selecting models is Occam's razor
  - Keep goals of interpretation vs. Prediction in mind



Tree Model



R –squared 82.3% Train 67.2% Test



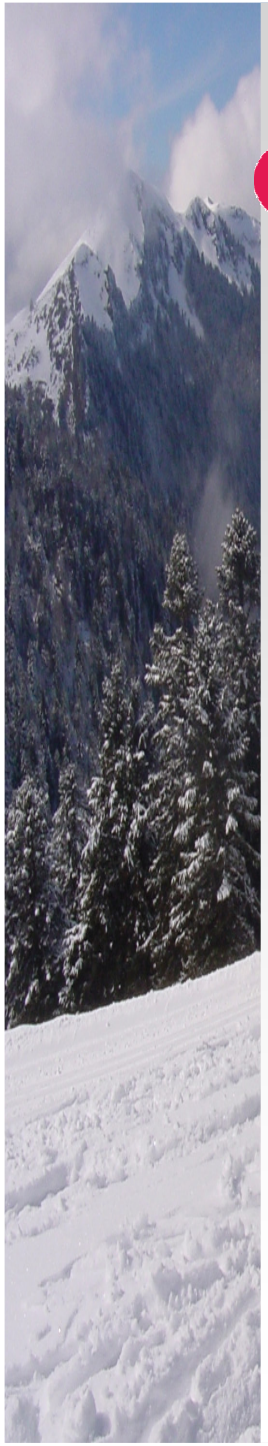
Feature Creation

- New predictor based on original predictors
- Often linear:

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

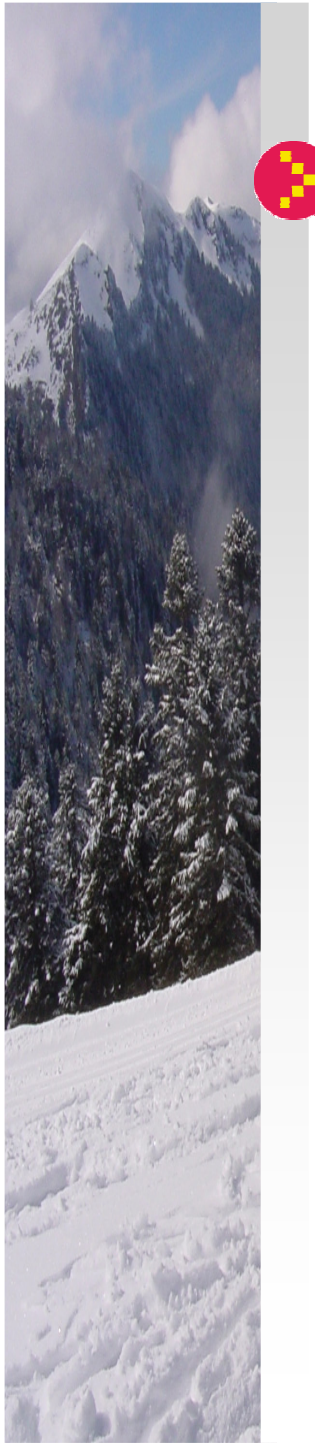
  - Principal components
  - Factor analysis
  - Multidimensional scaling

- But also simple transformations and ratios

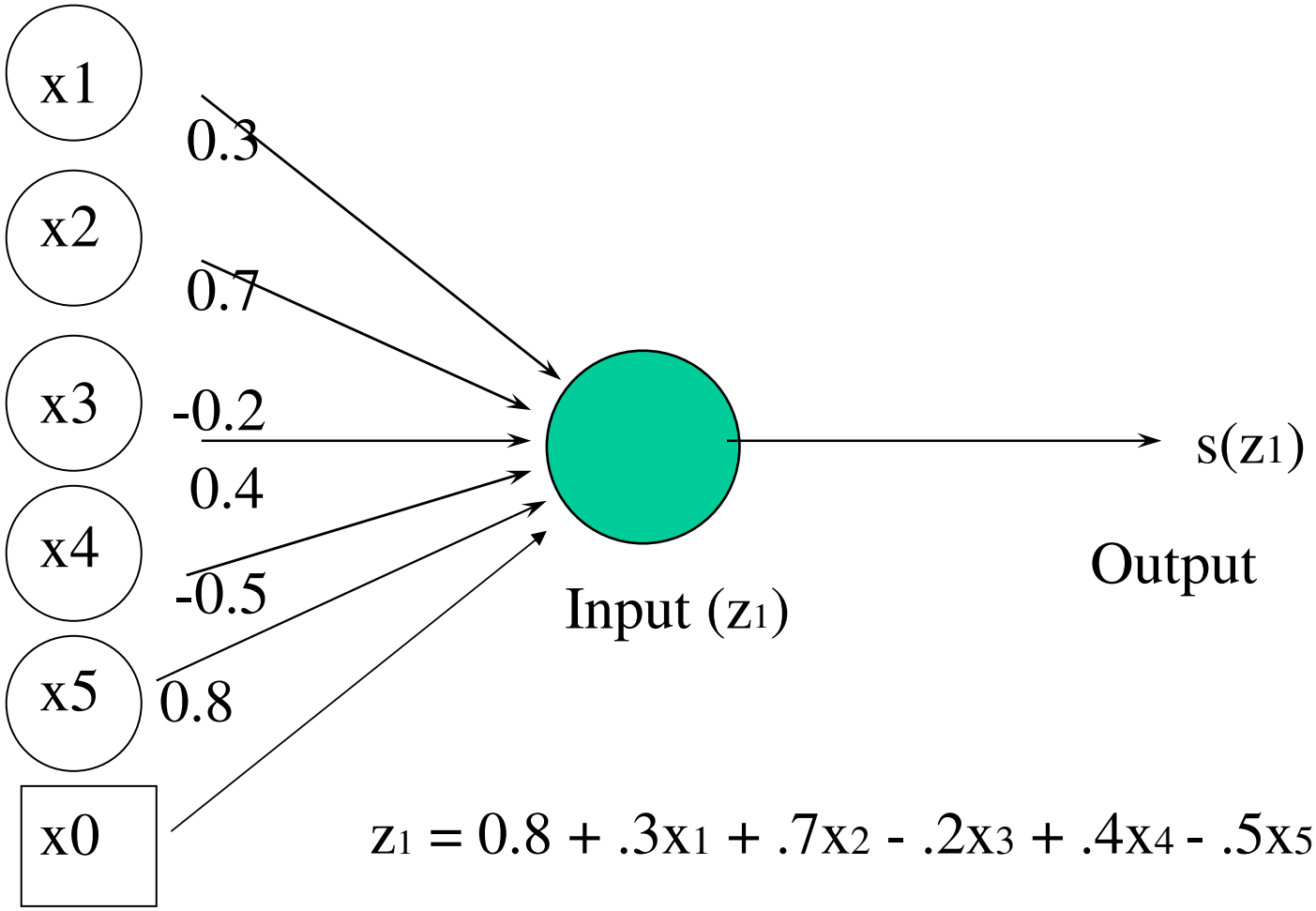


How to Use New Features

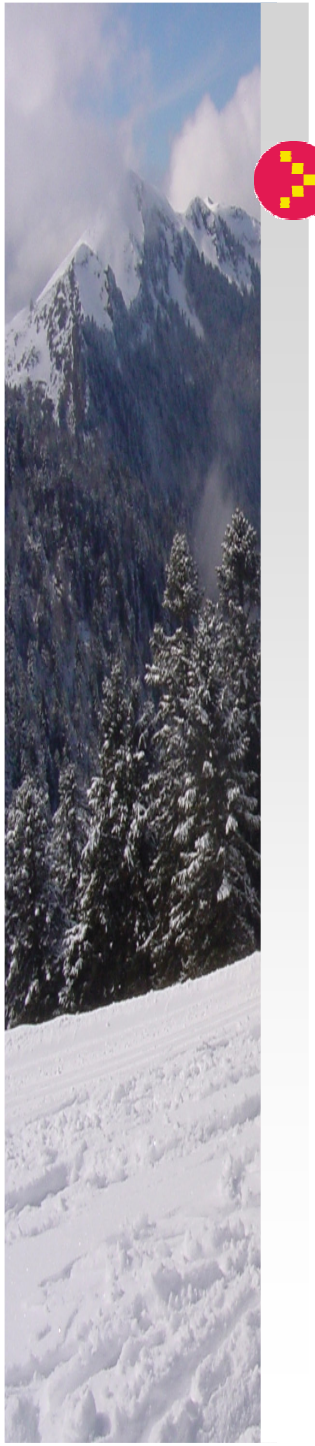
- **In original regression**
  - **Find new features, use as predictors**
- **Find them automatically**
  - Projection pursuit regression
    - ✓ Z's are linear, f's are arbitrary
  - Neural network
    - ✓ Z's are linear, f's are sigmoidal
      - The z's are the hidden nodes
      - The f's are the activation functions
      - The b's are the weights



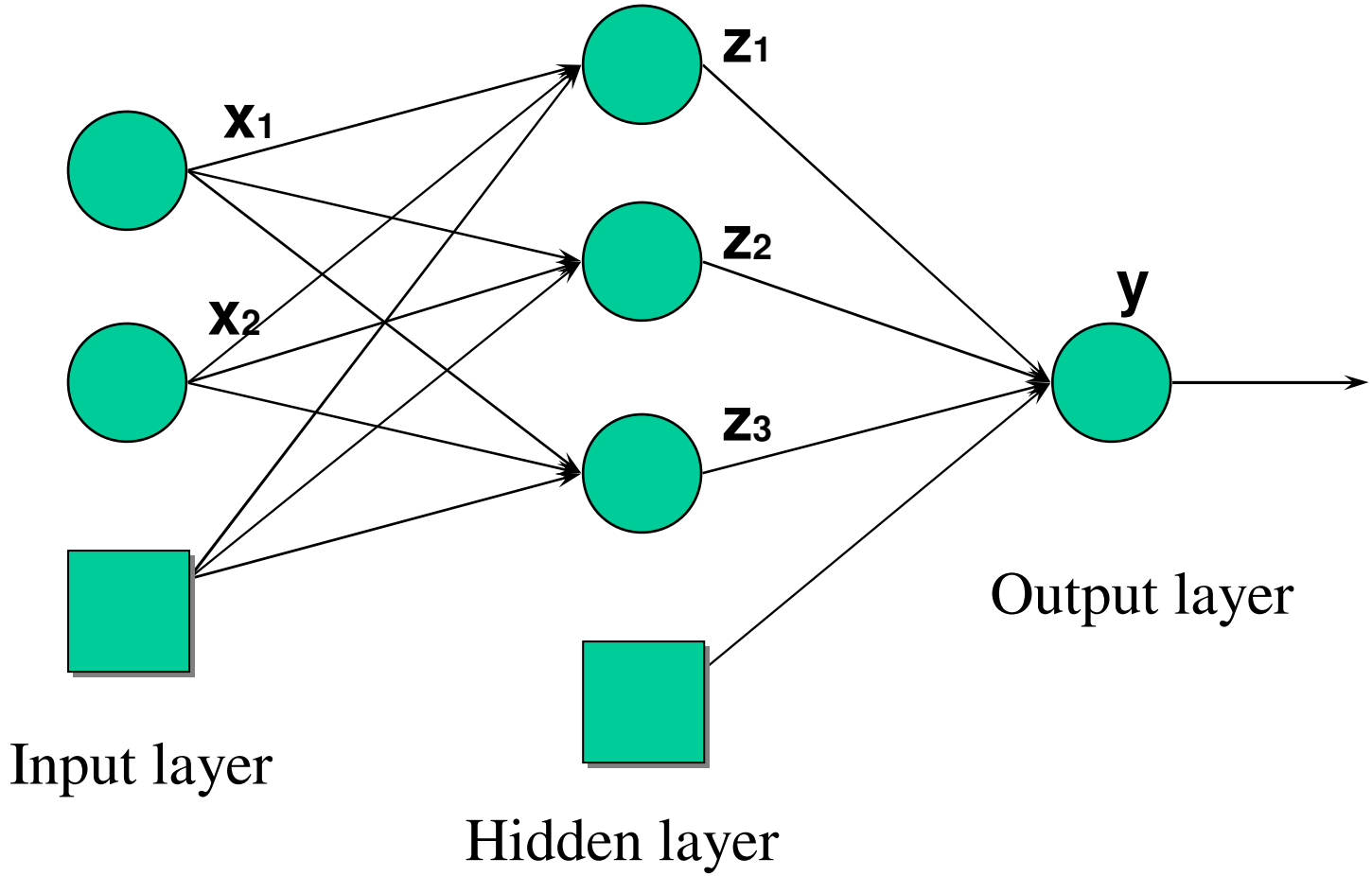
A Single Neuron



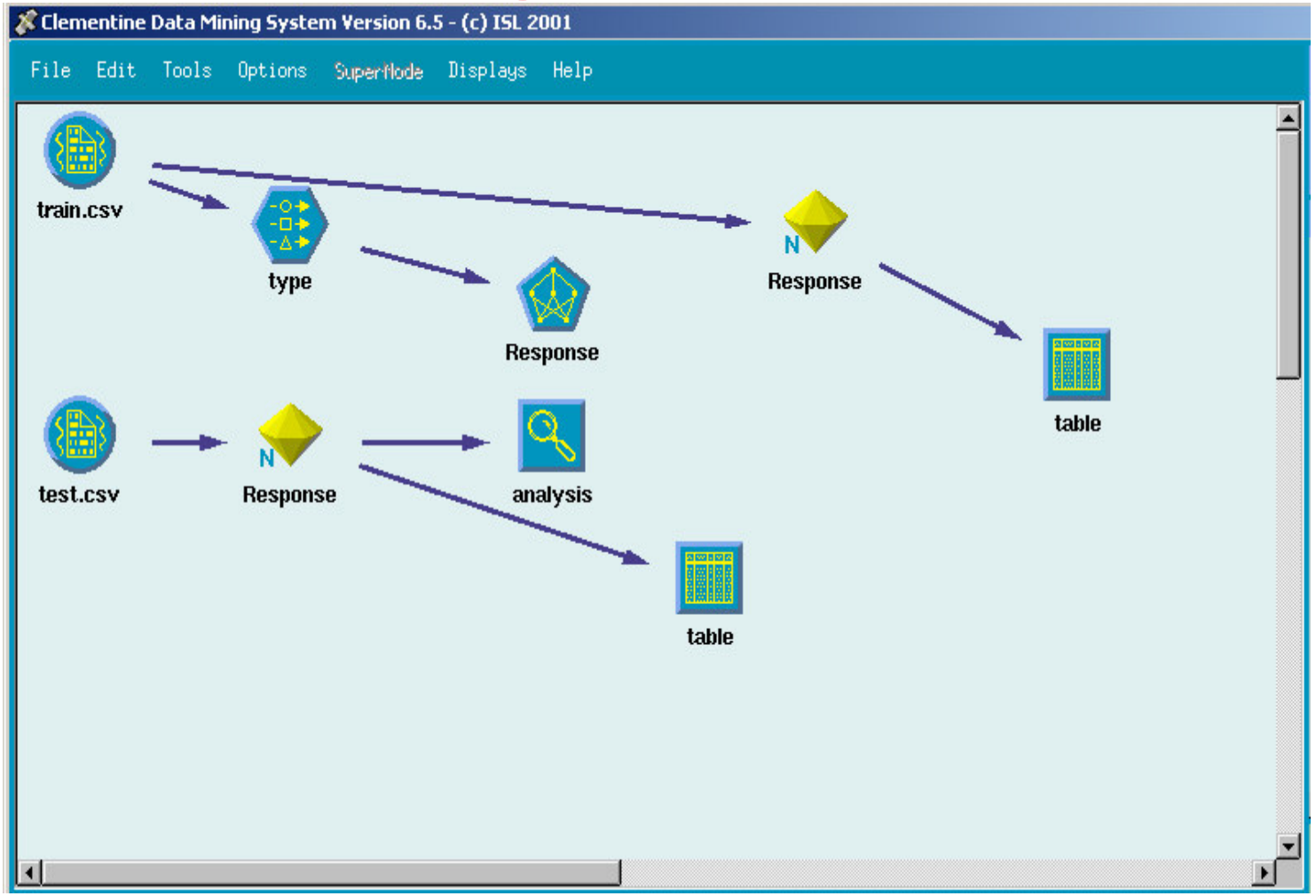
$$z_1 = 0.8 + .3x_1 + .7x_2 - .2x_3 + .4x_4 - .5x_5$$

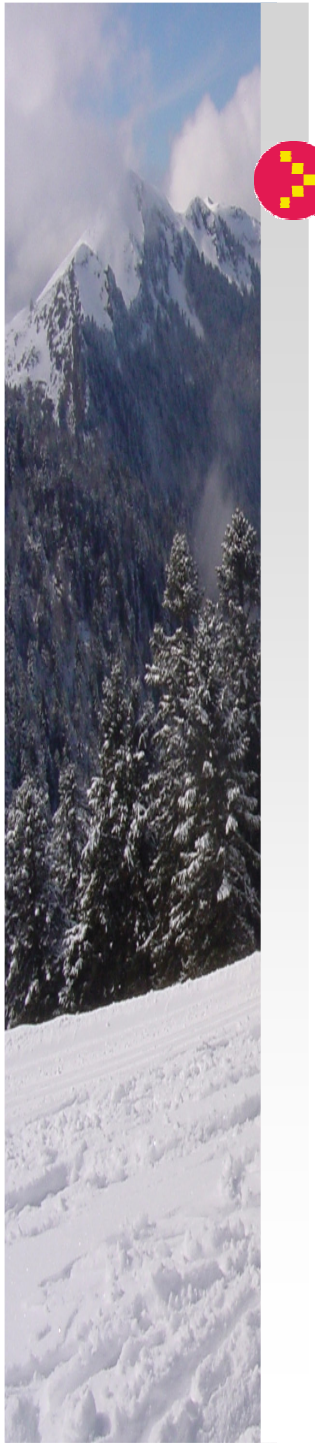


Layered Architecture

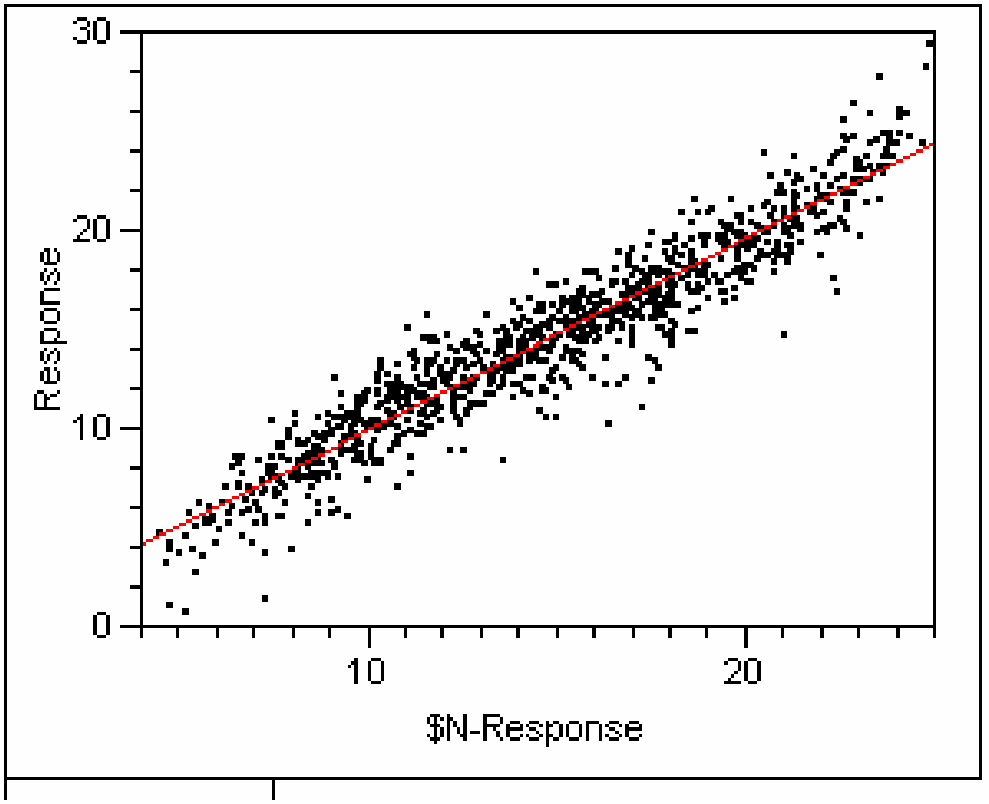


Running a Neural Net

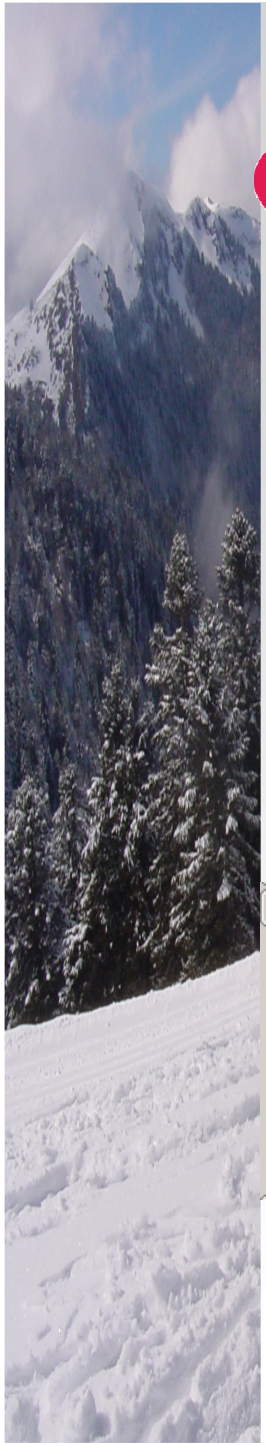




Predictions for Example

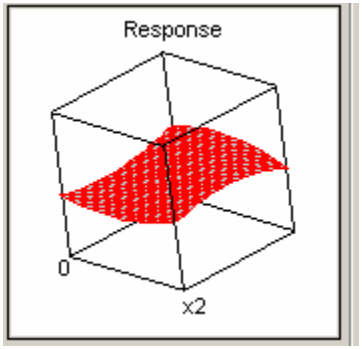


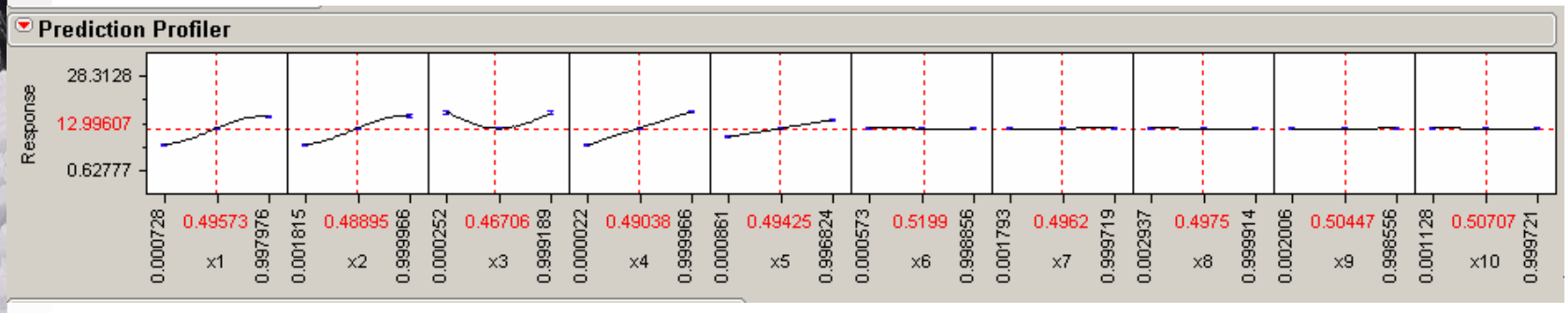
R squared 92.7% Train 90.6% Test

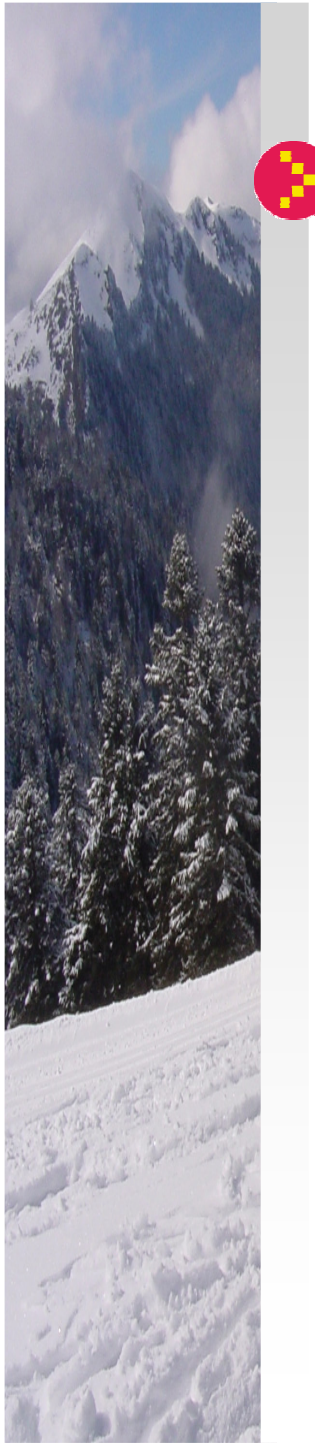


What Does This Get Us?

- Enormous flexibility
- Ability to fit anything
  - **Including noise**
- Interpretation?

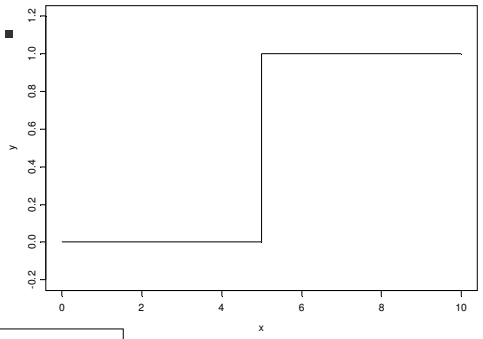


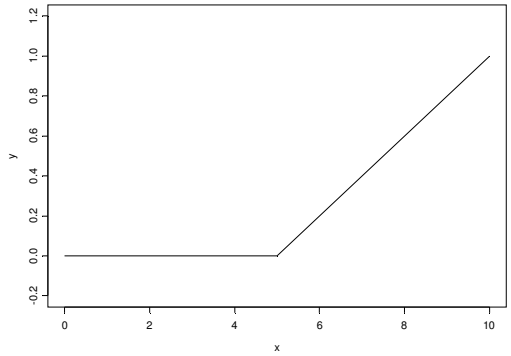


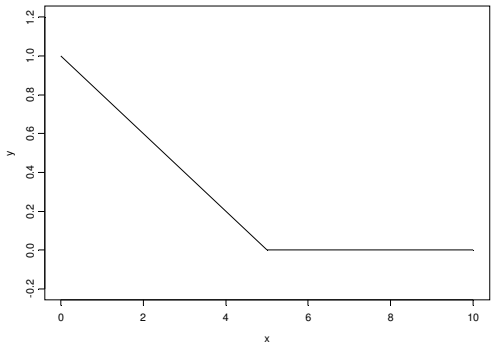


MARS

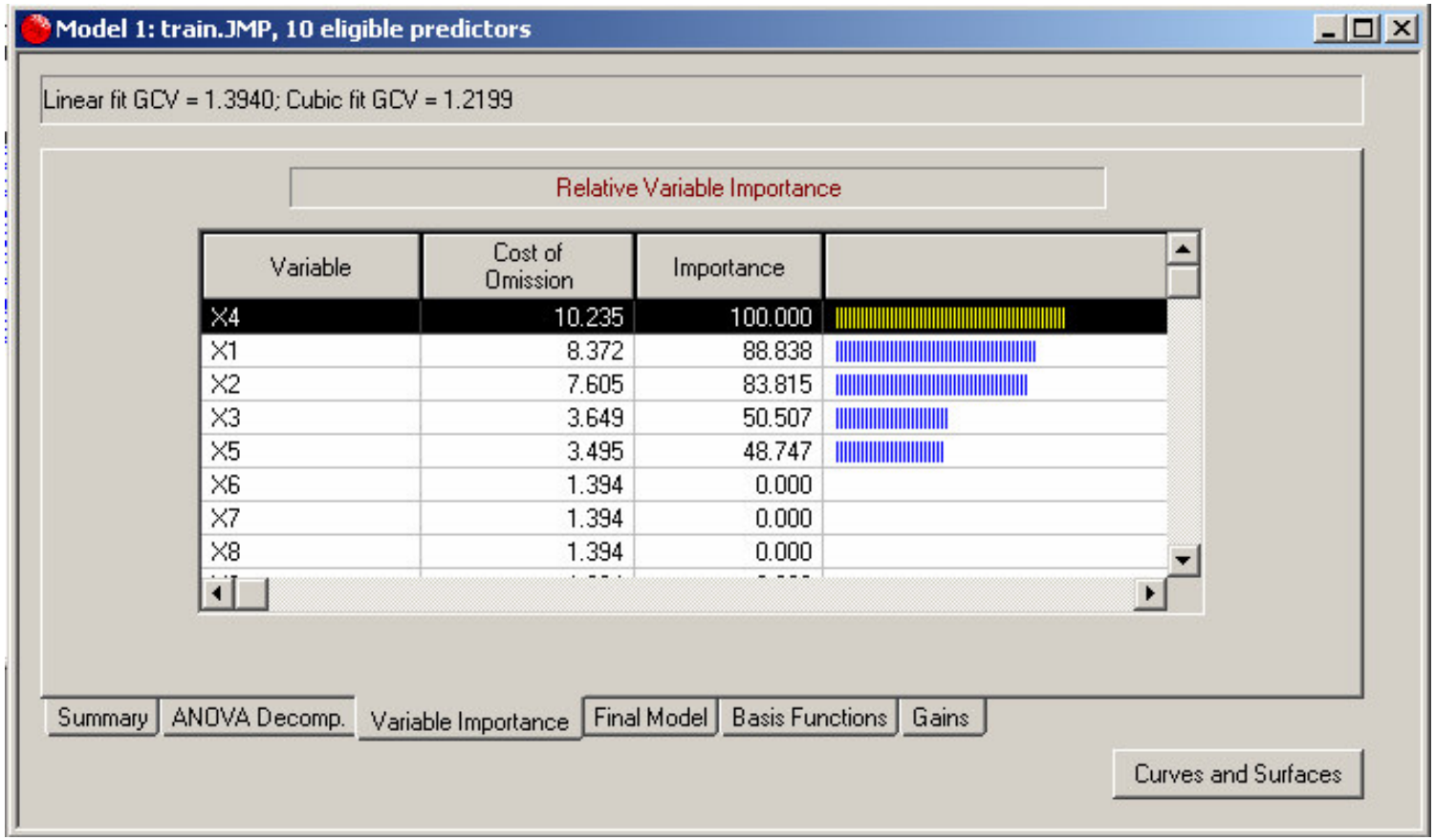
- **Multivariate Adaptive Regression Splines**
- **What do they do?**
  - Replace each step function in a tree model by a pair of linear functions.



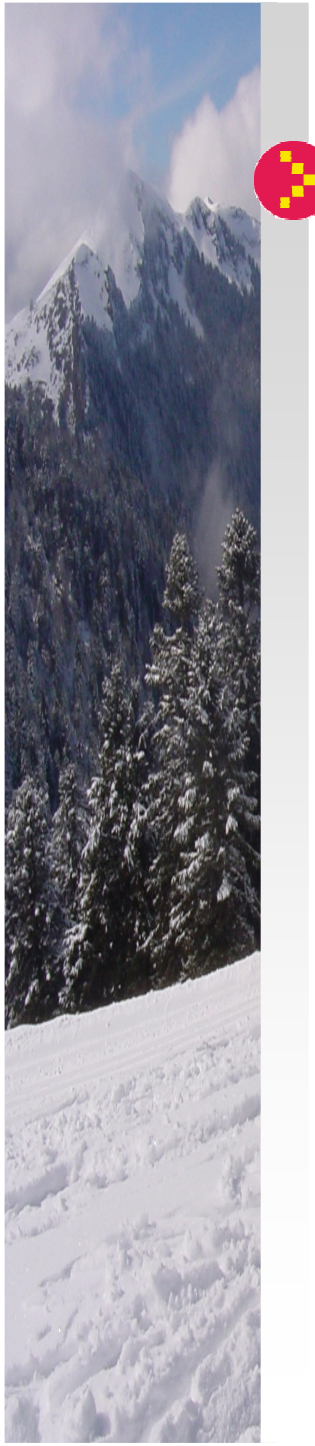




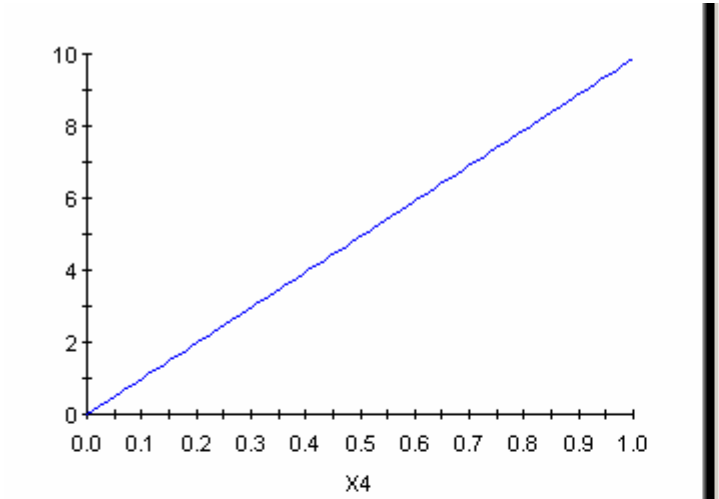
MARS Variable Importance

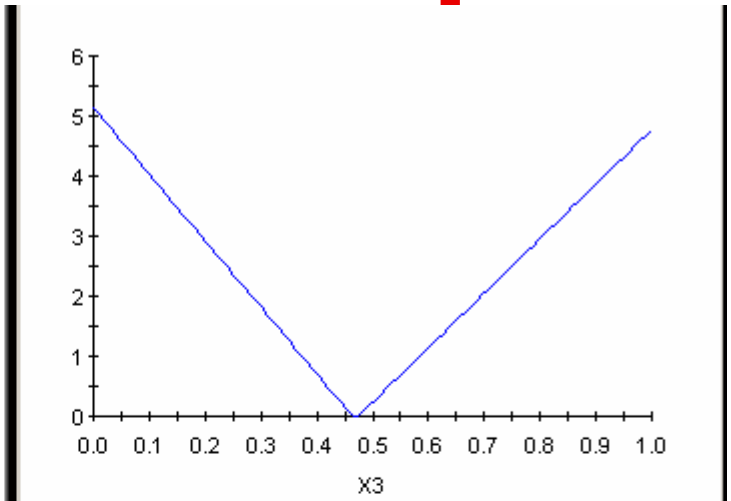


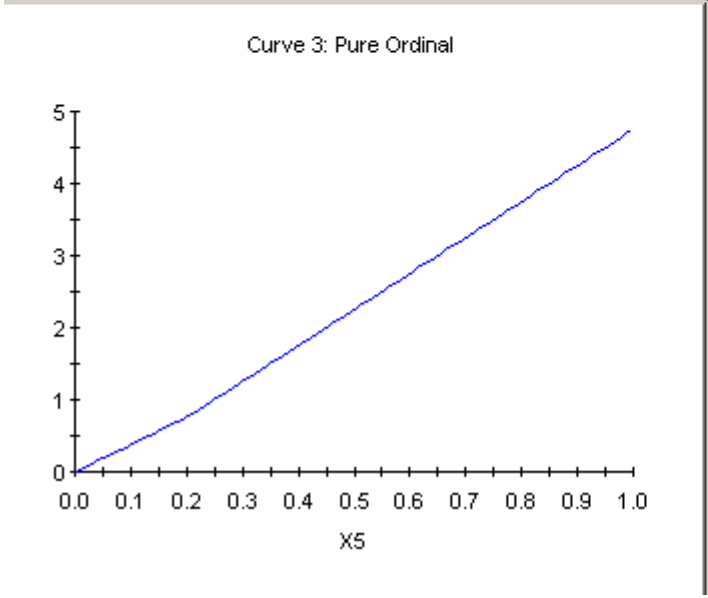
R-squared 95.0% Train 94.3% Test

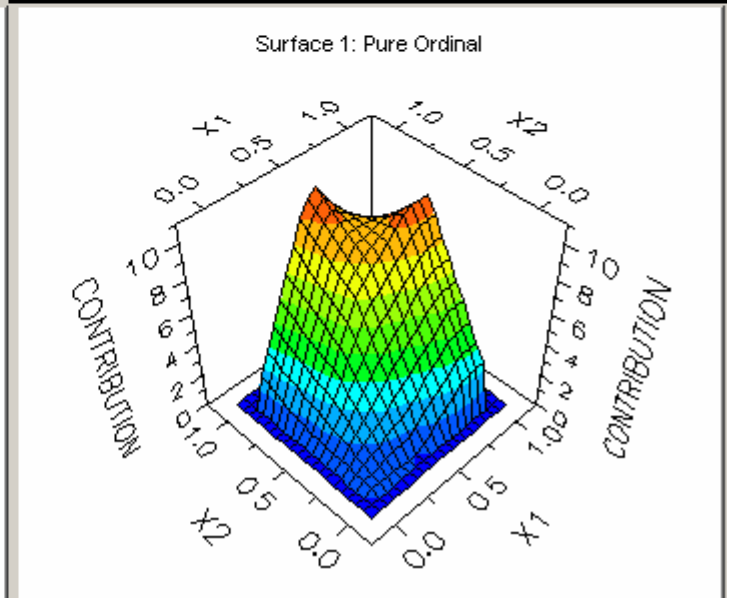


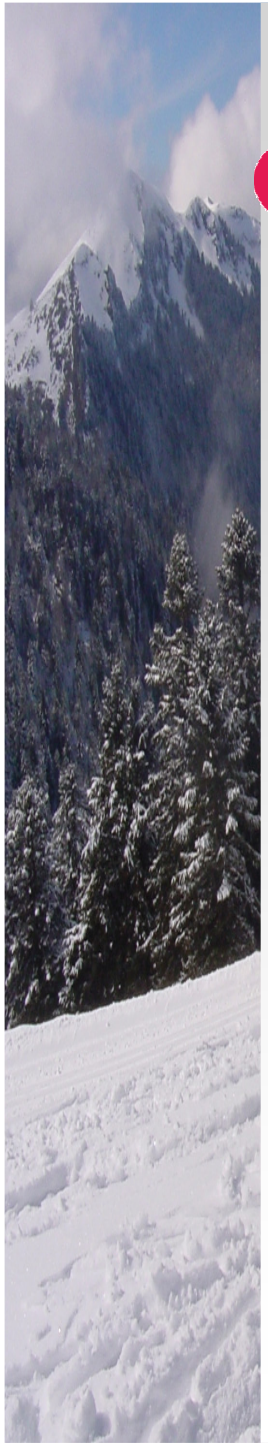
MARS Function Output





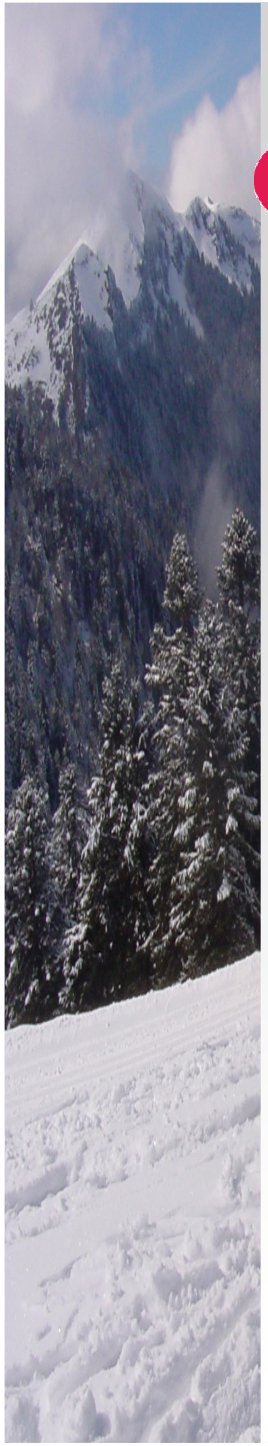






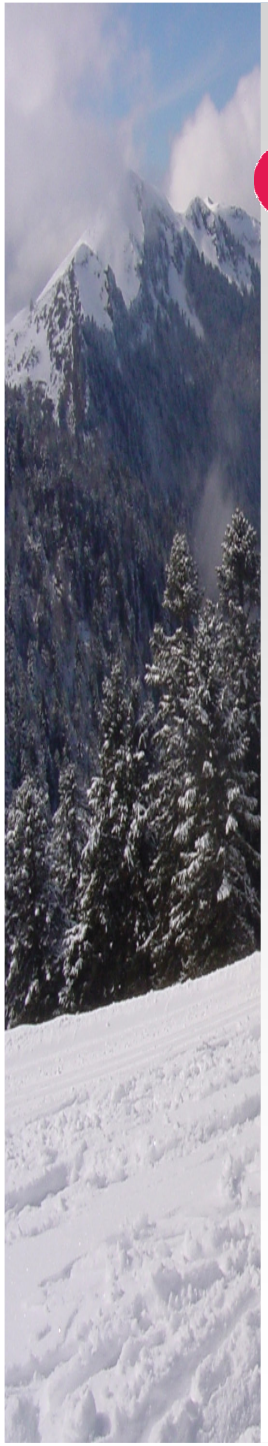
How Do We Really Start?

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



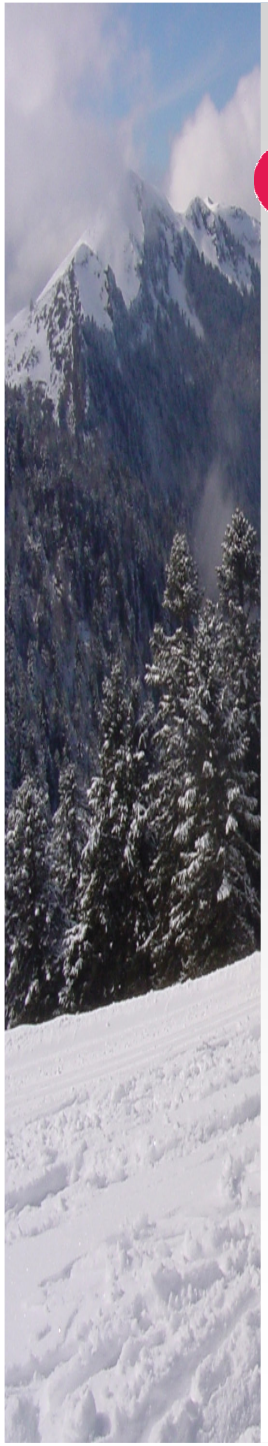
Where to Start

- 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)



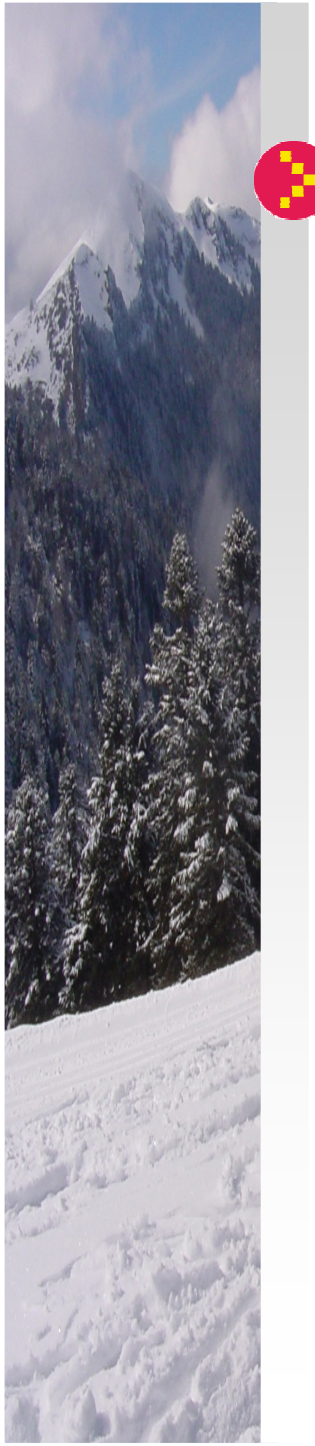
Exploratory Models

- **EDM**
  - **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



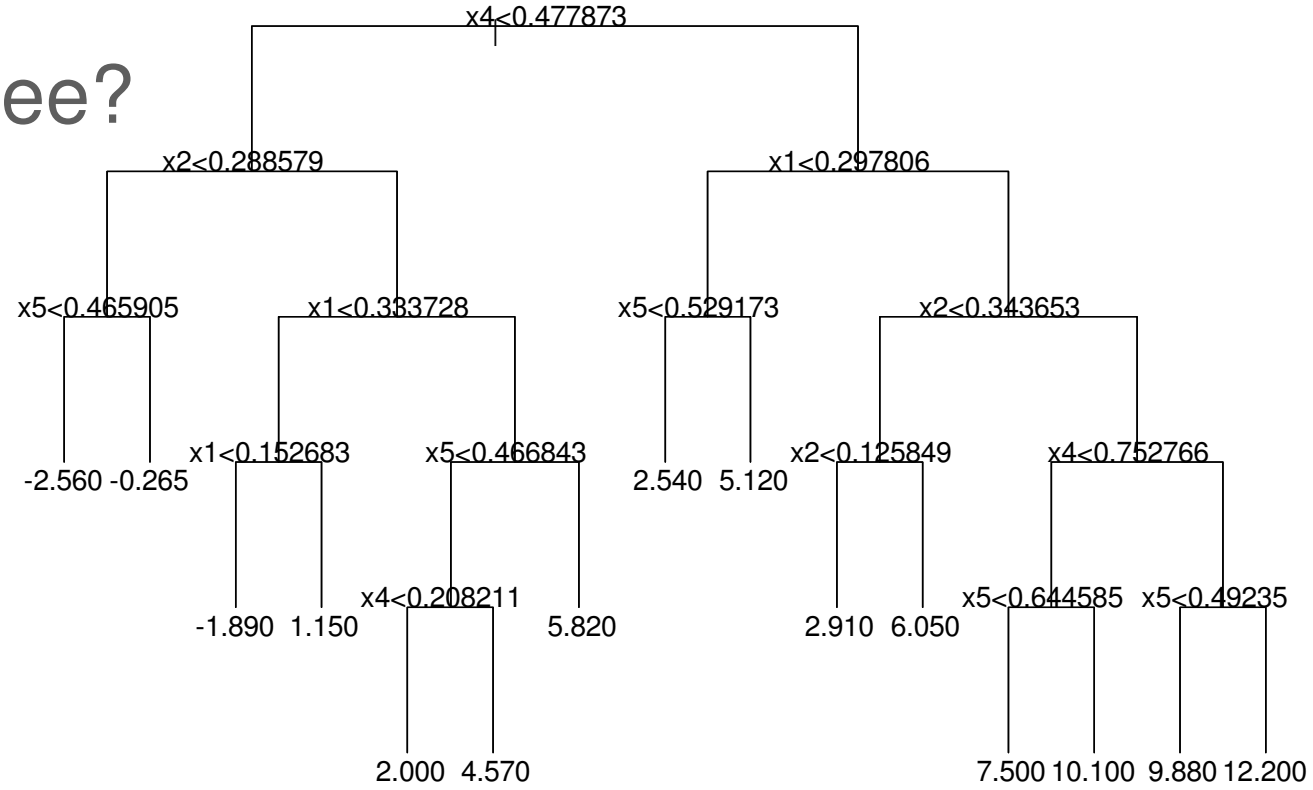
More Realistic

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

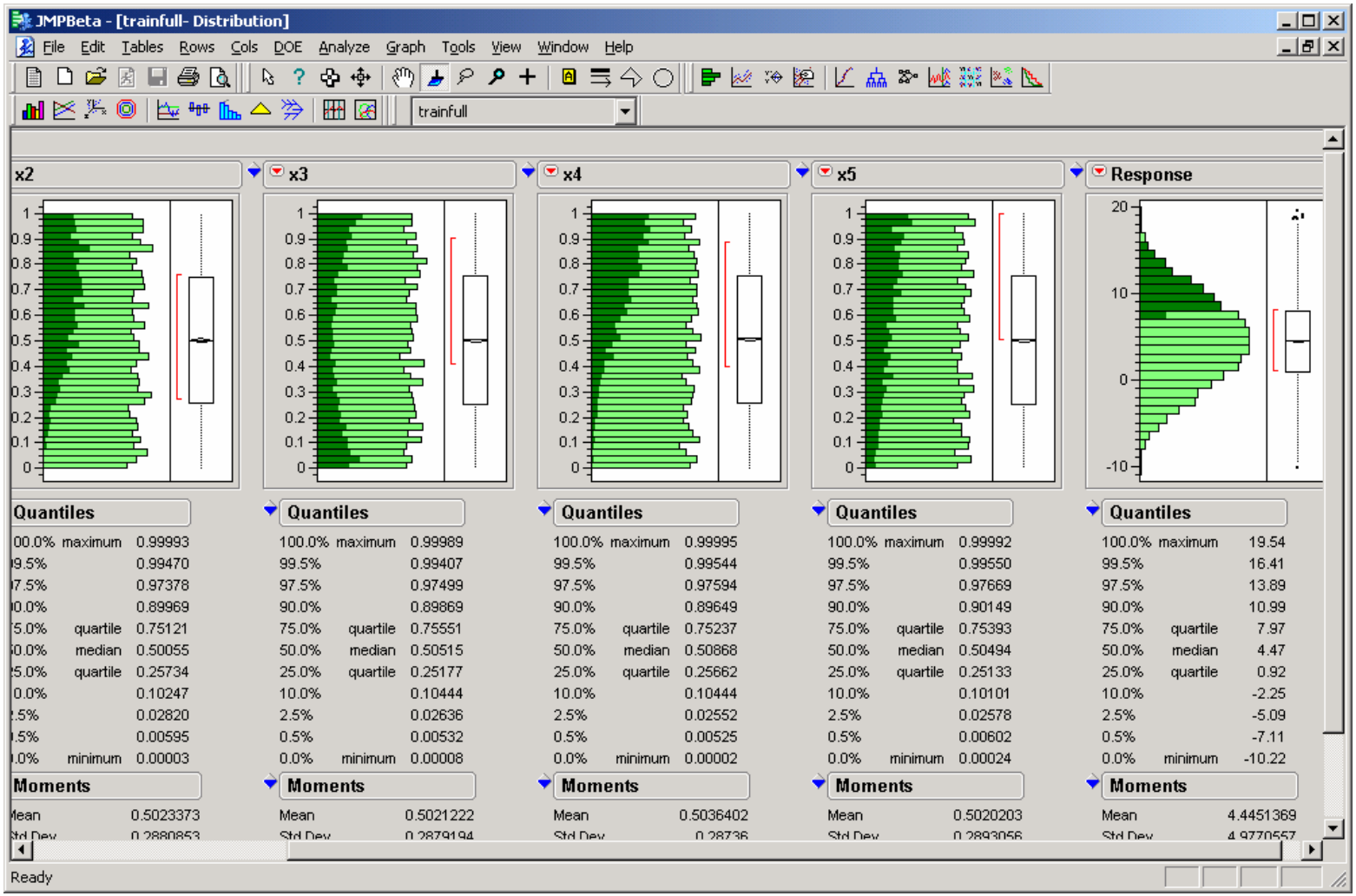


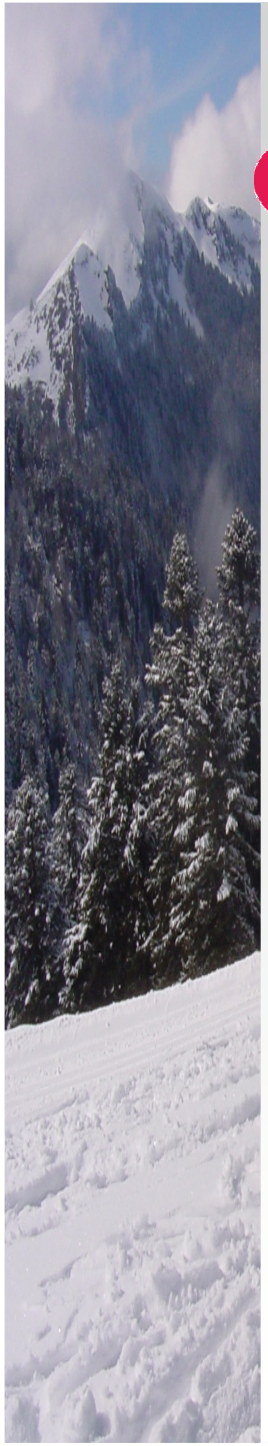
Start With a Simple Model

- Tree?



Brushing

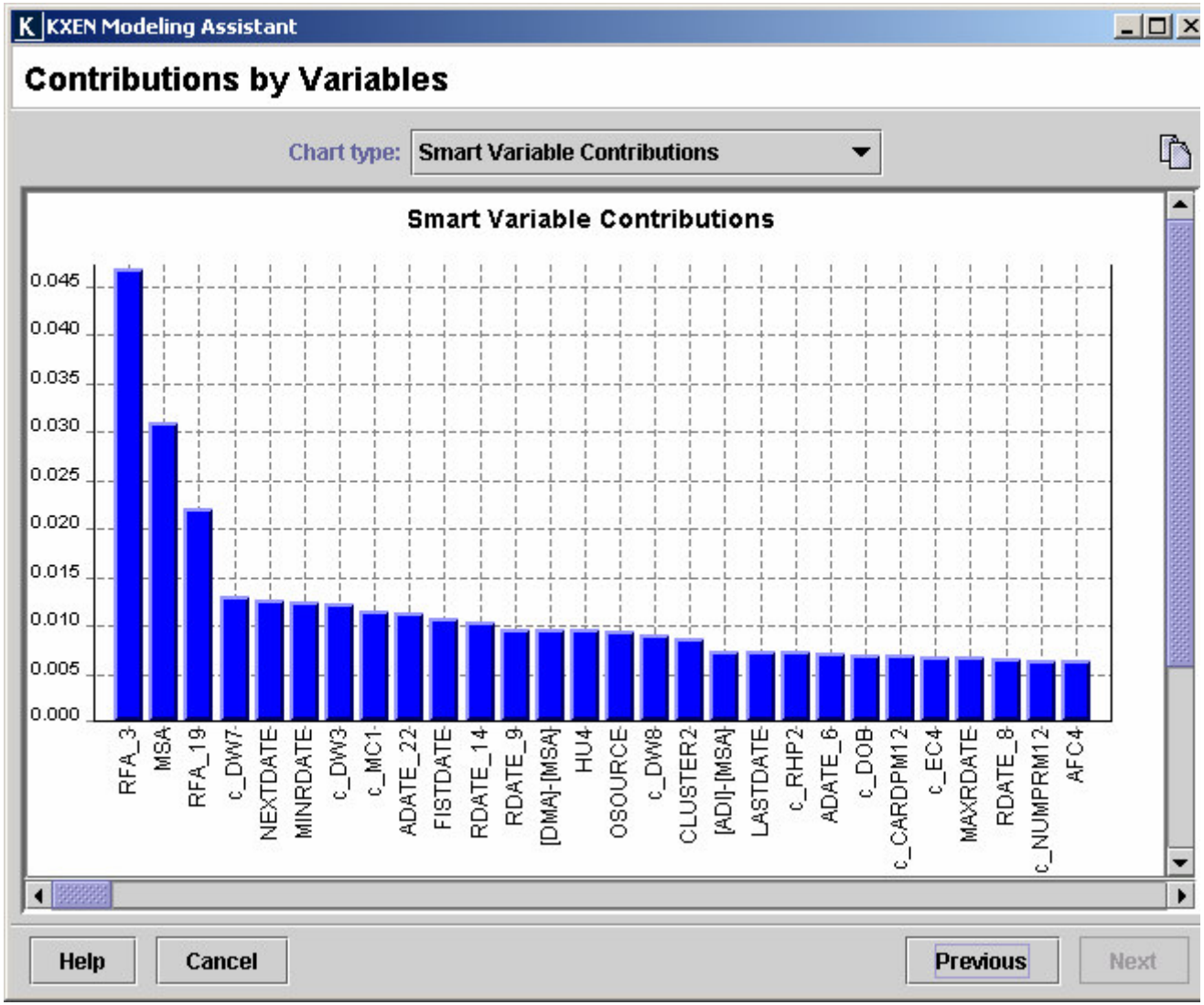

JMPBeta - [trainfull- Distribution]
File Edit Tables Rows Cols DOE Analyze Graph Tools View Window Help
trainfull
x2
x3
x4
x5
Response
Quantiles
Moments
Ready

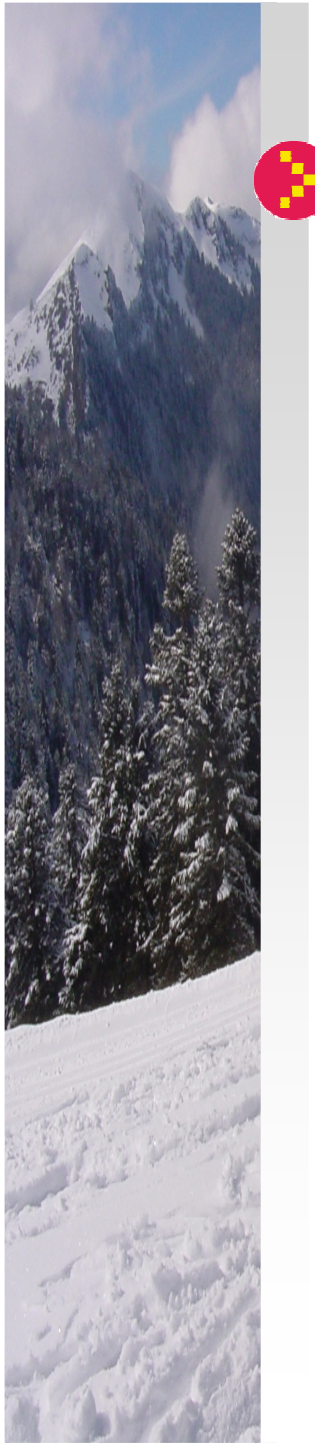


Paralyzed Veterans

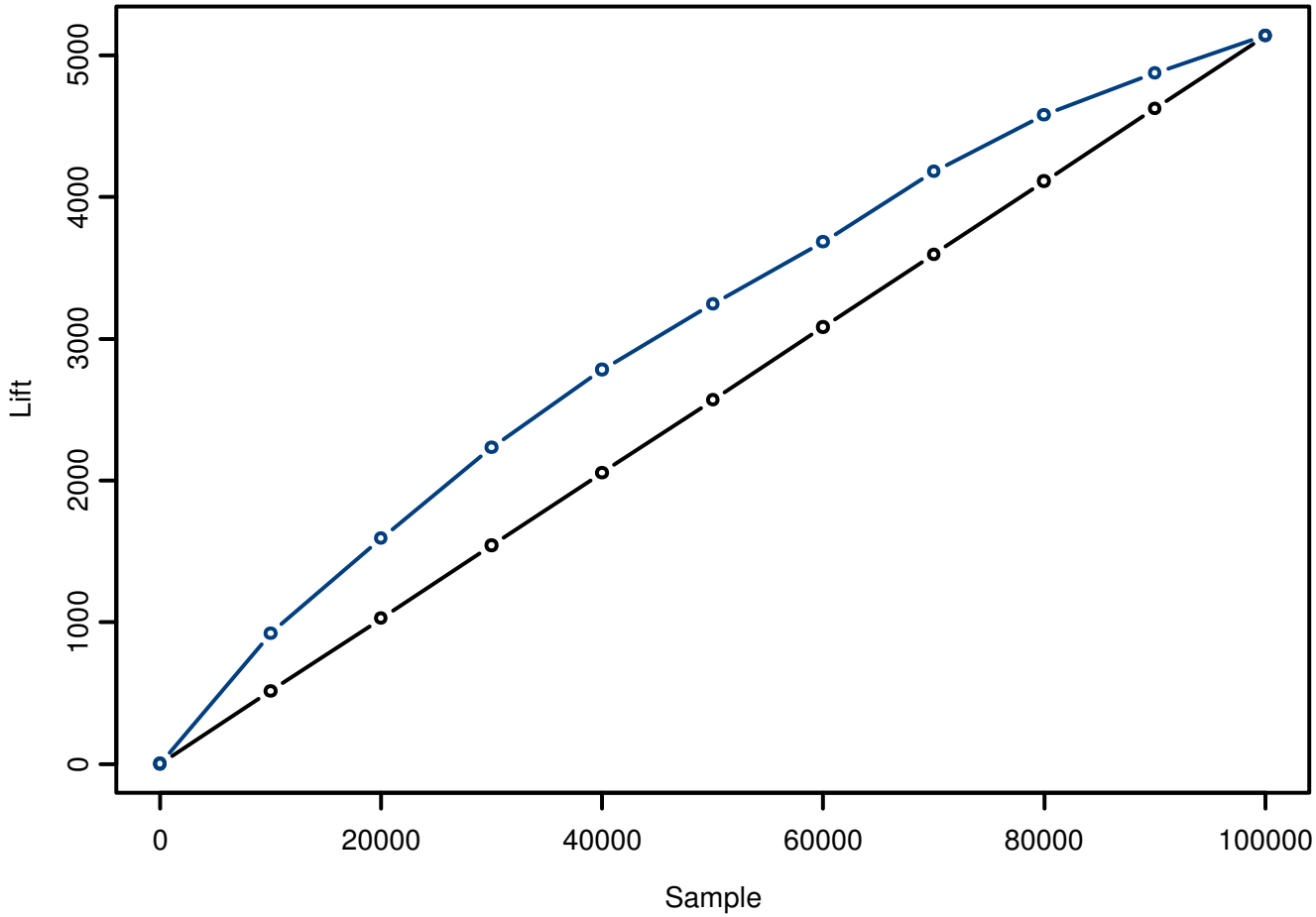
- Back to reality
- Categorical predictors
- Missing data
- Almost 500 potential predictors

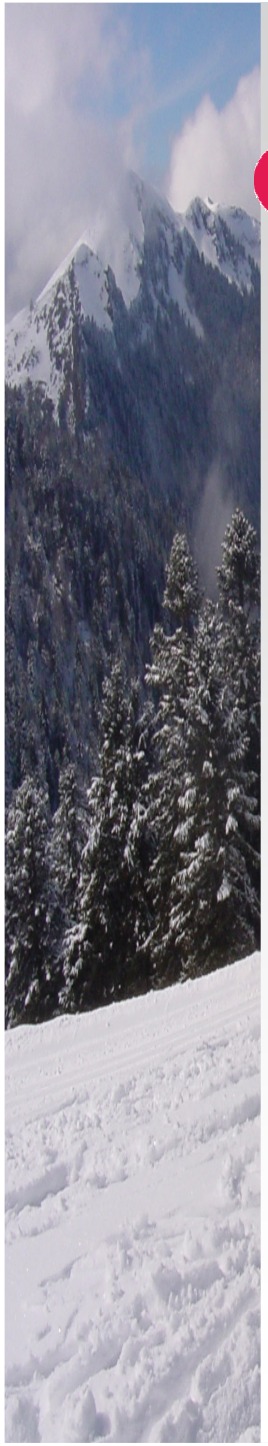
Variable Importance





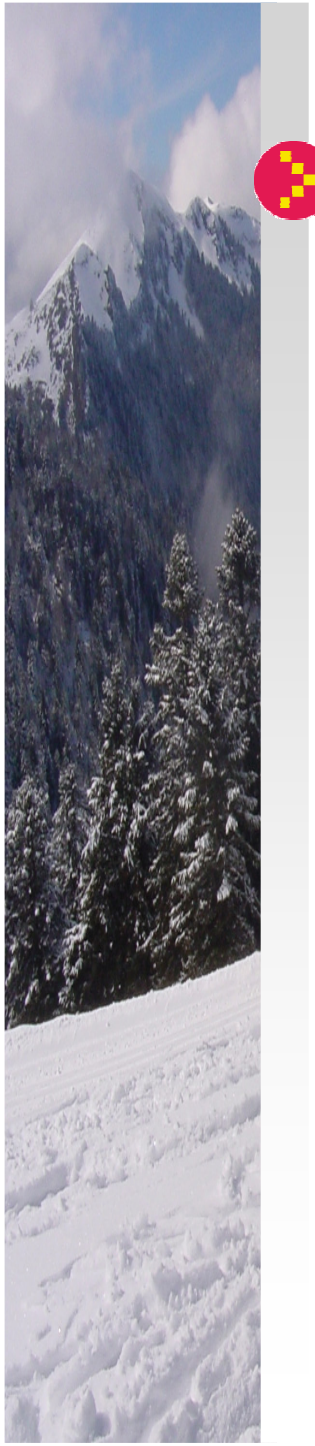
Lift Curve


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

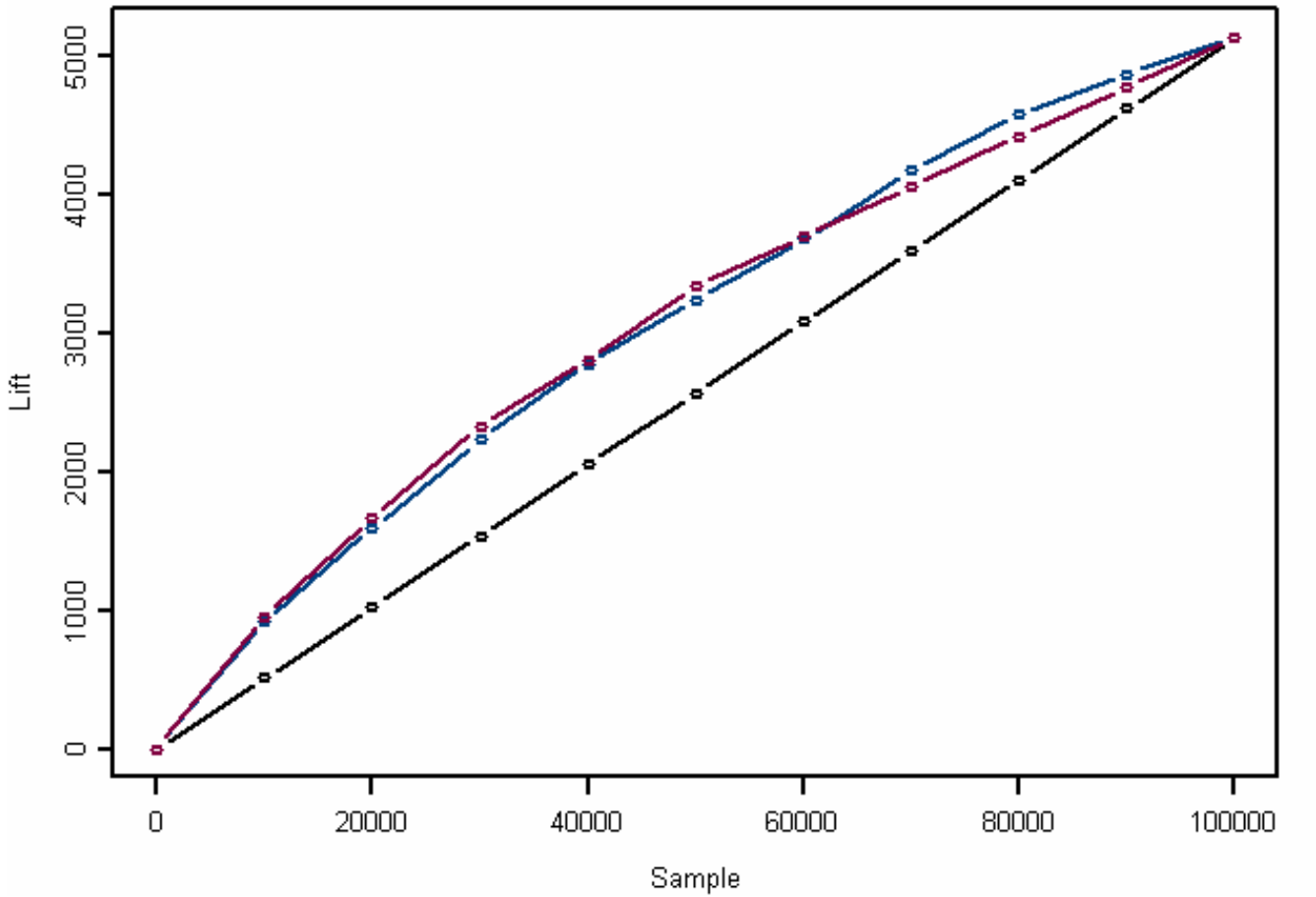


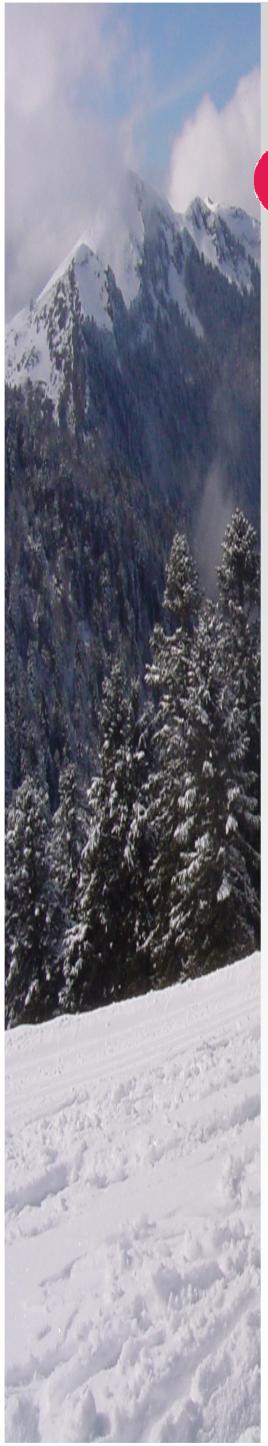
Tree Model

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

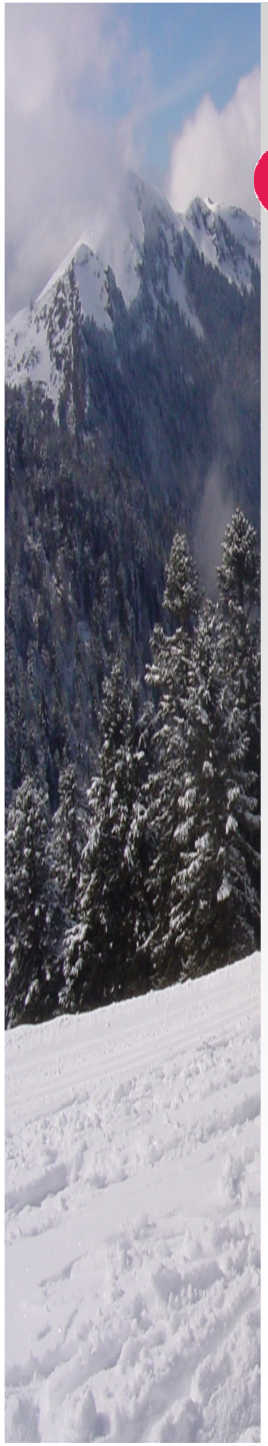


Is This the Answer?

- **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**

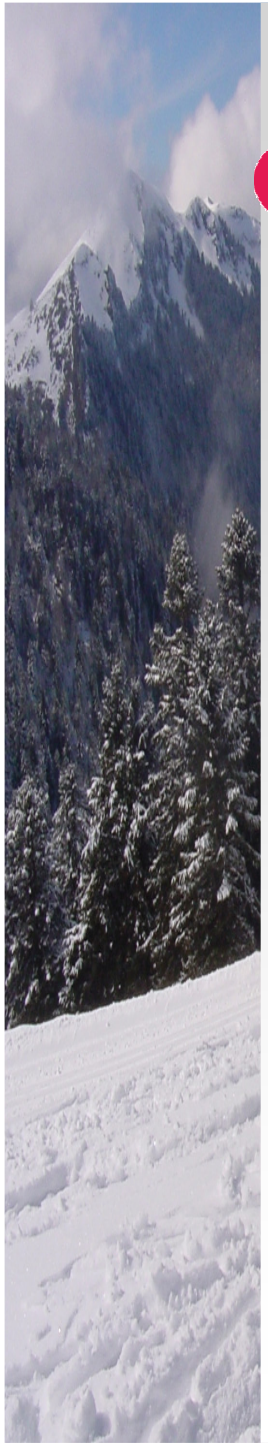
What Did We Learn?

- 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**



Recap

- Problem formulation
- Data preparation
  - **Data definitions**
  - **Data cleaning**
  - **Feature creation, transformations**
- EDM – exploratory modeling
  - **Reduce dimensions**
- Graphics
- Second phase modeling
- Testing, validation, implementation



Take Home Messages

- Data preparation is most of the work
- Dealing with missing values
- What to do first?
  - **Use an exploratory model**
- Which algorithm to use?
  - **All– this is the fun part, but beware of overfitting**
  - **Each tells you something**
- Results
  - **Keep goals in mind**
  - **Test models in real situations**