

Data Mining in the Real World: Five lessons learned in the pit

Richard D. De Veaux
Professor of Statistics
Department of Mathematics and Statistics
Williams College
Williamstown MA

deveaux@williams.edu



Data Mining Is...

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

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

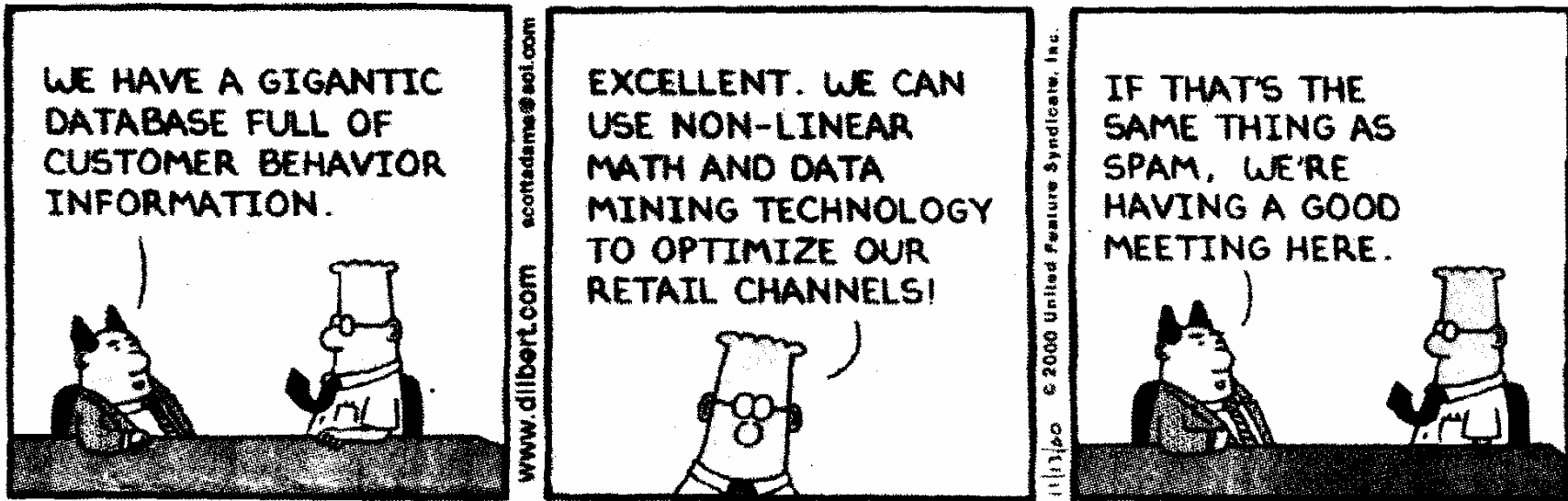
“a knowledge discovery process of extracting previously unknown, actionable information from very large data bases” --- Zornes

“ 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

“Statistics in the role of Detective” – John Sall, JMP

What is Data Mining?

DILBERT By SCOTT ADAMS



Lesson 1: Learn to make friends – you'll need them

- **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?**



What's "Hard"? --Example

The screenshot shows the JMP software interface with a data table and a summary of fit for an Oneway Anova. The data table has the following columns: ODATEDW, OSOURCE, TCODE, STATE, ZIP, MAILCODE, PVASTATE, DOB, NOEXCH, RECINHSE, RECP3, RECPGVG, RECSWEEP, and MC. The summary of fit for the Oneway Anova shows the following statistics:

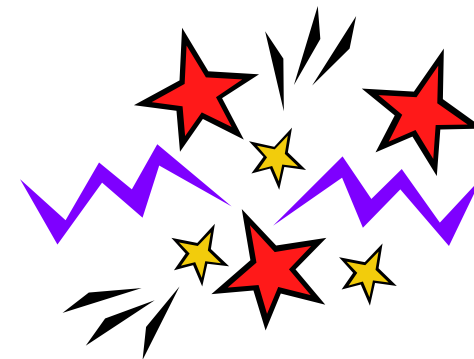
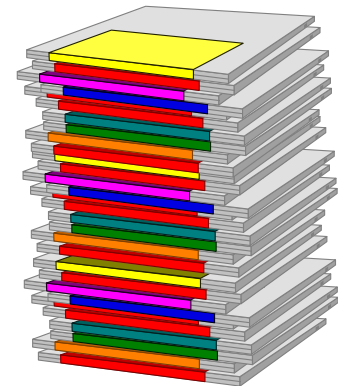
Statistic	Value
Std Err Mean	3.1122959
upper 95% Mean	60.513171
lower 95% Mean	48.313039
N	94649

Metadata

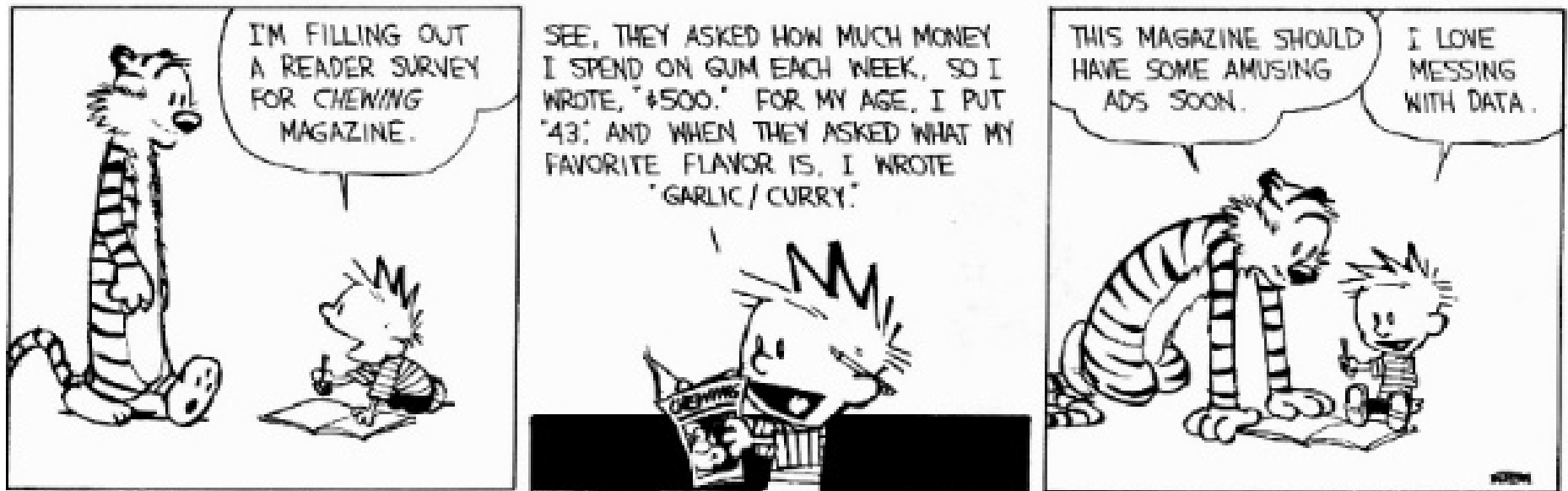
- **The data dictionary describes the data set contents and characteristics**
 - Table name
 - Description
 - Primary key/foreign key relationships
 - Collection information: how, where, conditions
 - Timeframe: daily, weekly, monthly
 - Cosynchronous: every Monday or Tuesday
 - Who owns this information?

Data Challenges

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



Data Quality



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

Data Preparation

- **Build data mining database**
 - Combining sources
 - Synchronizing sources
- **Explore data**
- **Prepare data for modeling**

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



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
Quads tone/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

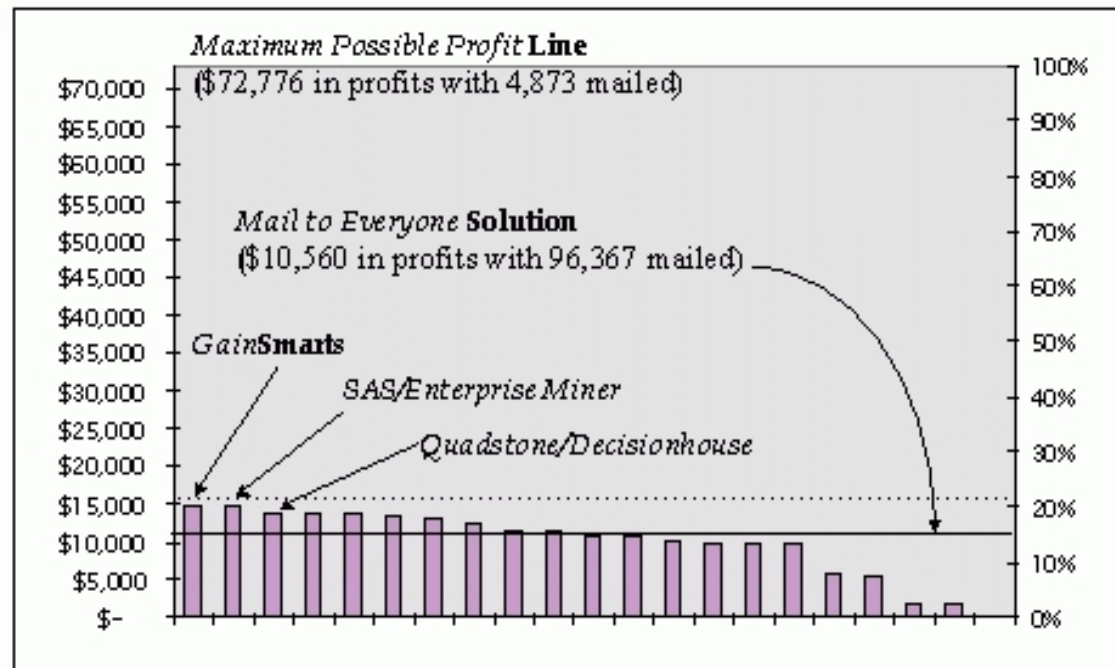
KDD-CUP-98

8/98



KDD CUP 98 Results 2

KDD-CUP-98 Results (2 of 2)



Ismail Parsa

KDD-CUP-98

8/98



Students in Data Mining Class

KDD-CUP-98 Results (1 of 2)

Participants	Sum of Actual Profits	Number Mailed	Average Profits
Student #1 \$15,024	\$ 14,712.24	56,330	0.26
Student #2 \$14,695	\$ 14,662.43	55,838	0.26
Student #3 \$14,345	\$ 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



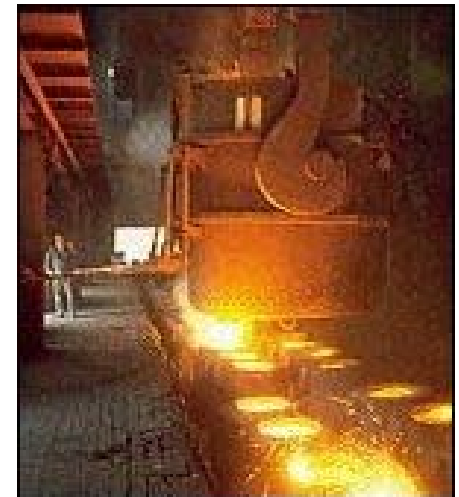
Lesson 2: Twyman's Law

- **“If it looks interesting, it’s probably wrong”**

- **De Veaux’ Corrolary to Twyman’s Law**
 - “If it isn’t wrong, it’s probably obvious

Ingots cracking

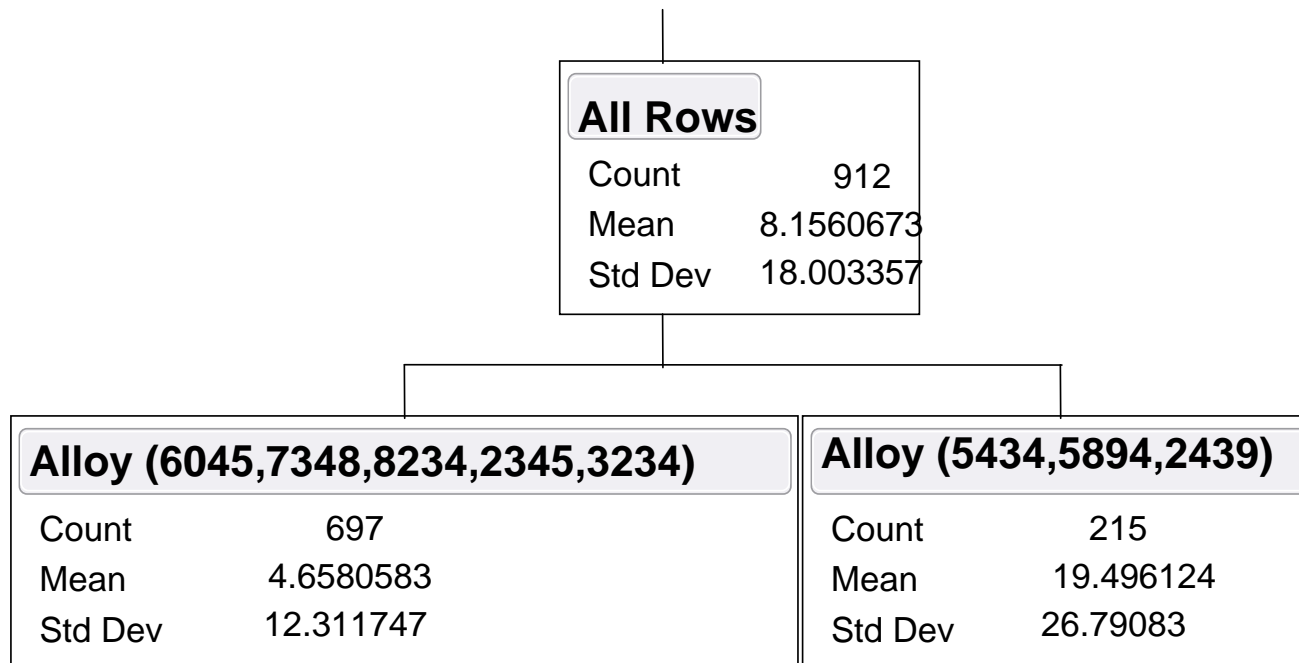
- **953 30,000 lb. Ingots**
 - 20% cracking rate
 - \$30,000 per recast
 - 90 potential explanatory variables
 - Water composition (reduced)
 - Metal composition
 - Process variables
 - Other environmental variables



Data Processing

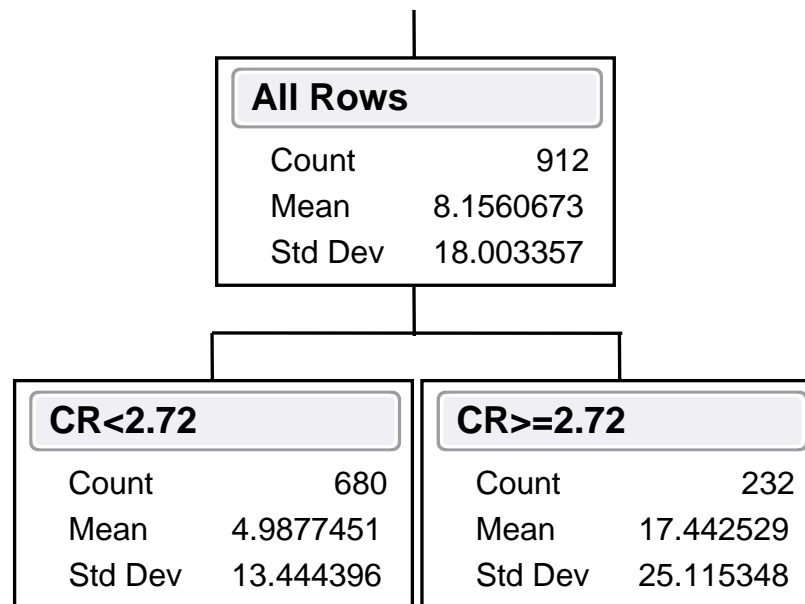
- **Five months to consolidate process data**
- **Three months to analyze and reduce dimension of water data**
- **Eight months after starting projects, statisticians received flat file:**
 - 960 ingots (rows)
 - 149 variables

First Tree



We know that – some alloys are hard to make. That’s why we gave you the data in the first place.

Second Tree



What do you think is *in* those alloys?

One More Time

- **Third Tree- Looks like Manganese matters**

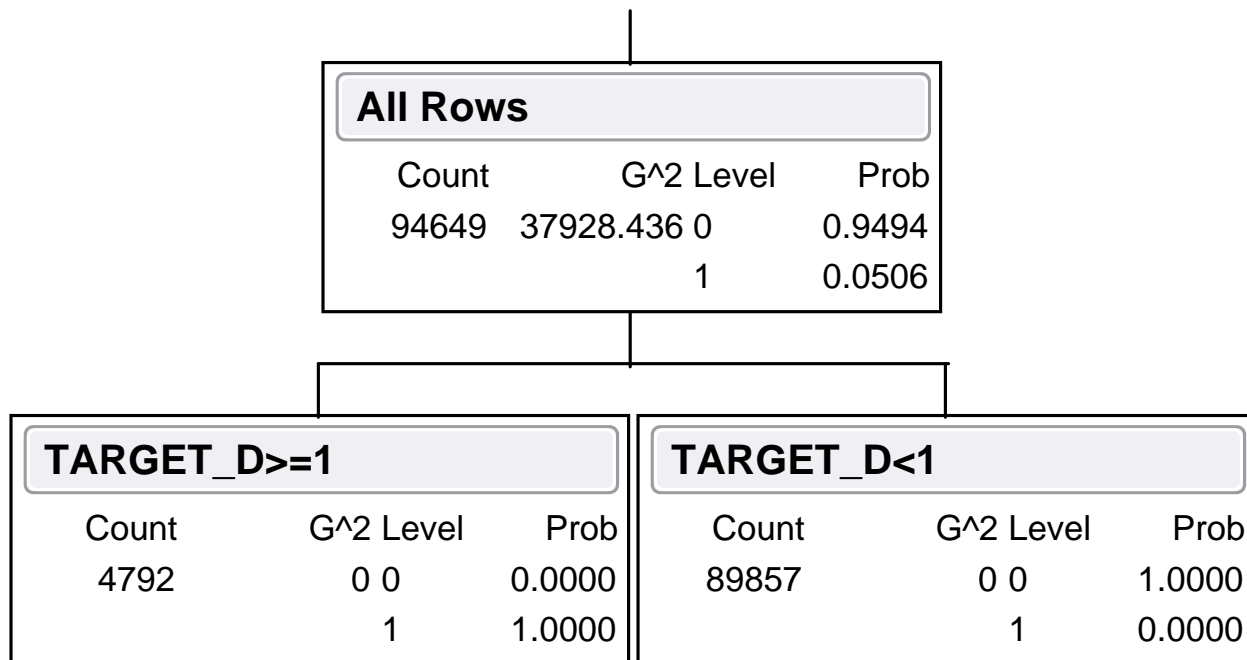
- OH!
- Did that solve it?
 - Experimental design
 - Enabled us to *focus* on important variables



I mean “Hmm.. That’s funny.”

-Issac Asimov

PVA Data: Herb's Tree



Lesson 3: Know When to Fold 'em

■ Liability for churches

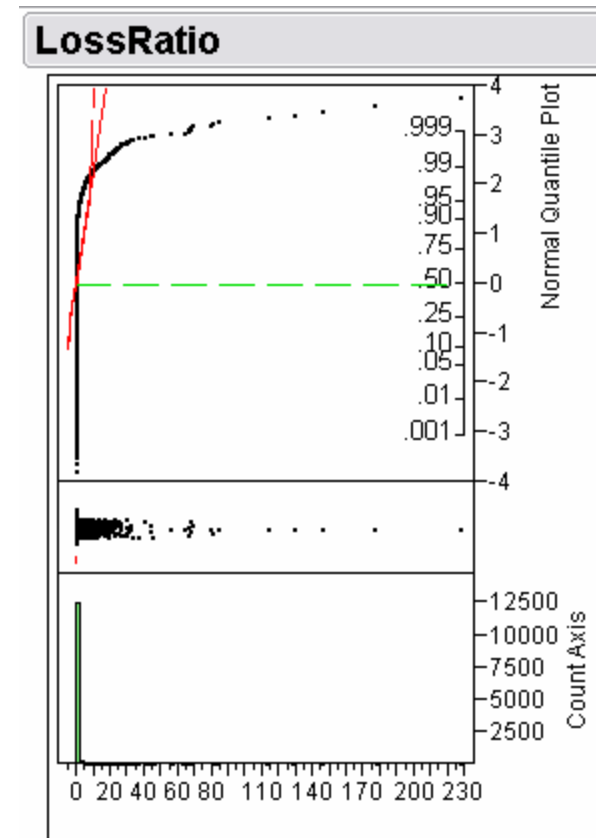
■ Some Predictors

- 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)

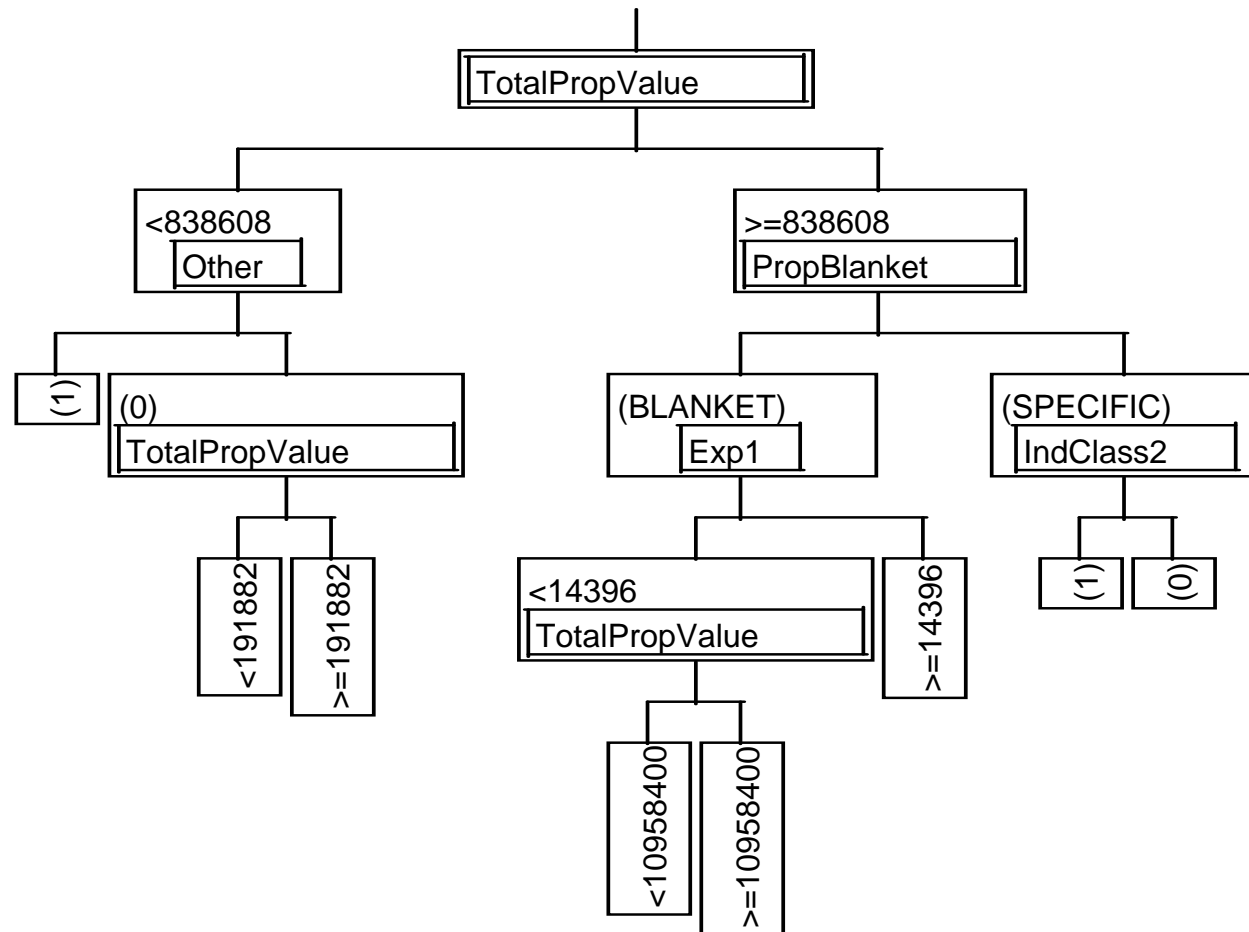


Churches – First Steps

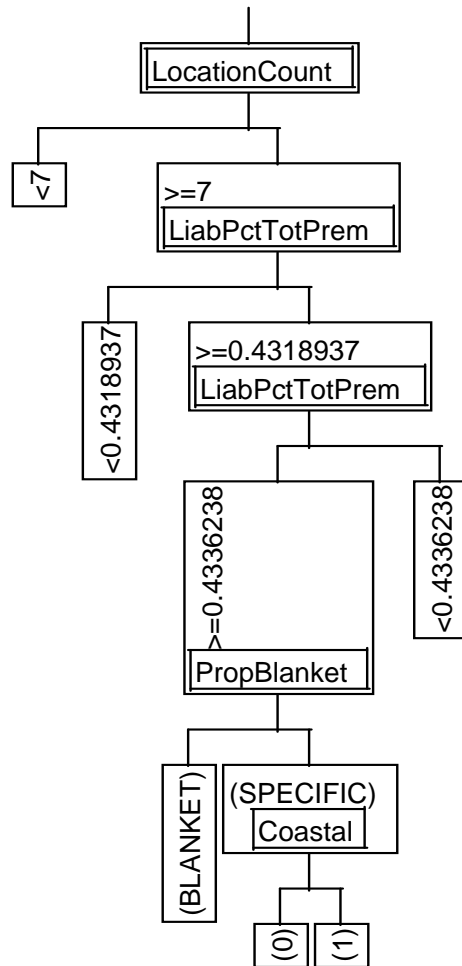
- **Select Test and Training sets**
- **Look at data**
 - Transform Loss Ratio?
 - Categorize Loss Ratio?
 - Outliers
- **Tree**



First Tree



Next Tree

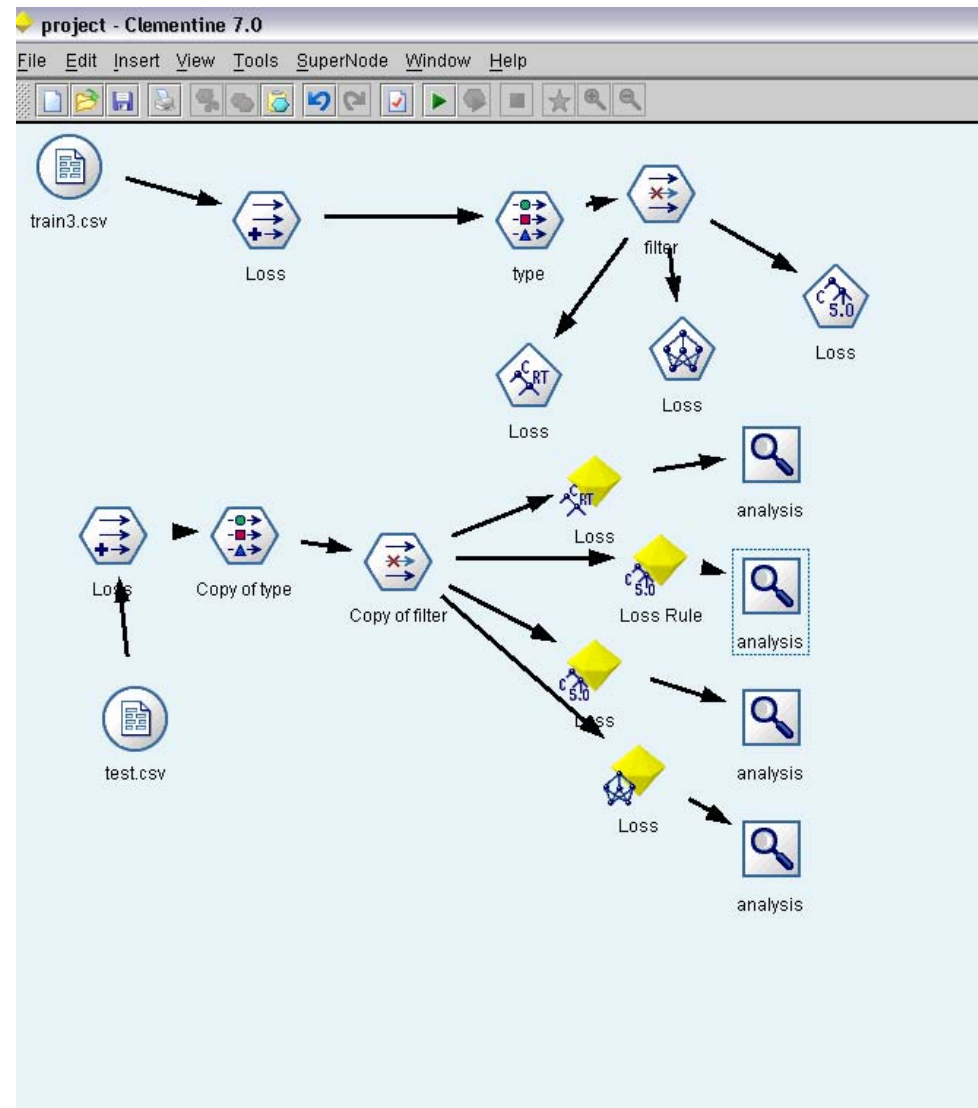


Where to go from here?

Churches – Next Steps

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

Clementine



Model Selection

- **After trying:**
 - Multiple regression
 - Trees
 - Neural networks

- **None of the models had a cross validated R^2 greater than 1%**

- **What does that tell you?**

- **Automatic data processing**
 - Missing values treated as category
 - Each variables broken into quartiles and appropriate number of degrees of freedom chosen
 - Categorical variables at k levels generate m dummy variables. Typically $m \ll k$.
 - Summary of model fit : KI (model quality) and KR (model reliability)
- **For Church data KI = -0.012 KR = 0.034**

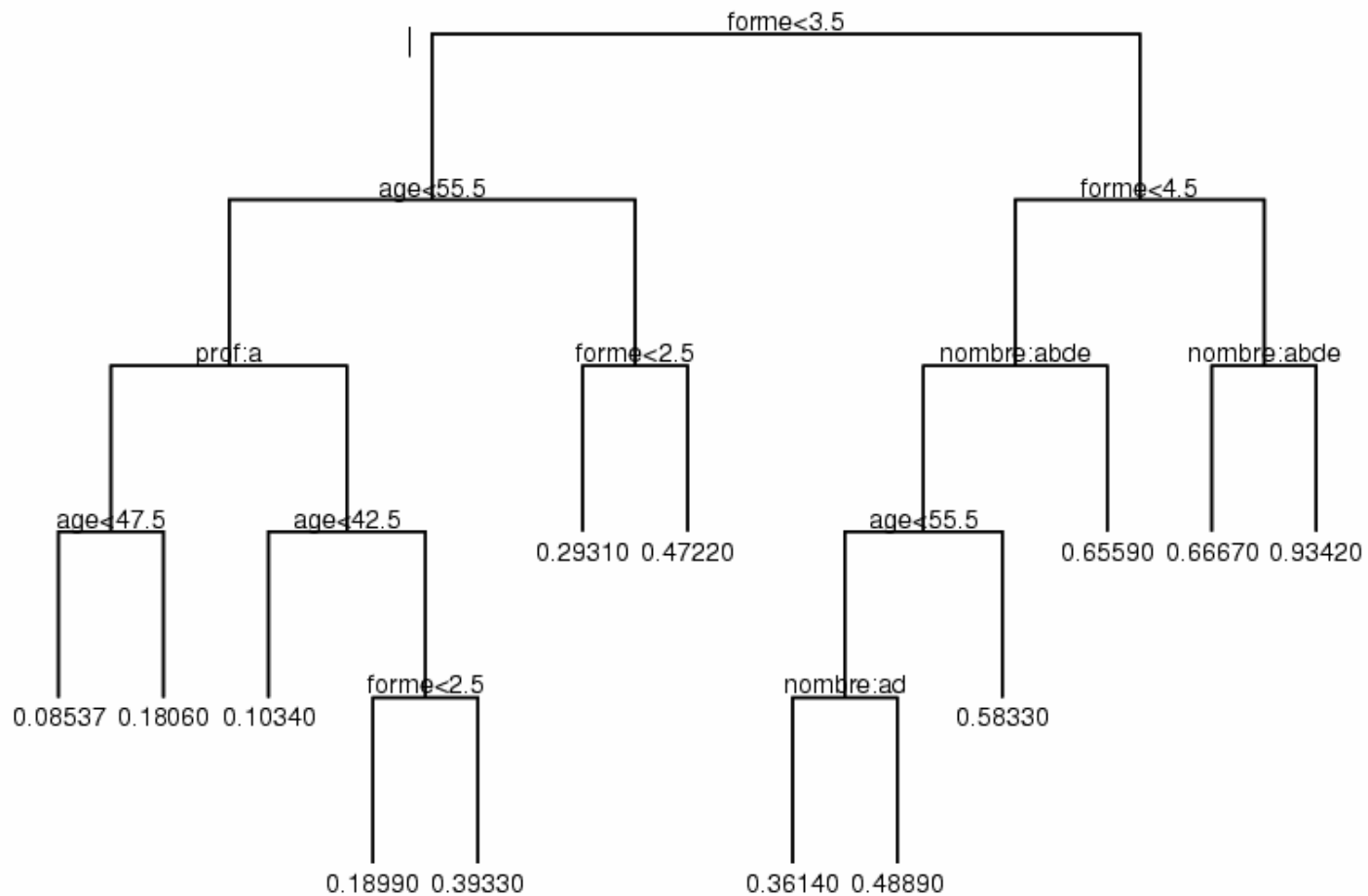
Lesson 4: Know When to Hold 'em

- **Breast cancer data from mammograms**
 - Error rates by trained radiologists are near 25% for both false positives and false negatives
- **Early detection of breast cancer is crucial**
- **Cumulative type I error over a decade is near 100% leading to needless biopsies**
- **Newer equipment (MRI) is increasingly used for screening but is prohibitively expensive for the developing world**

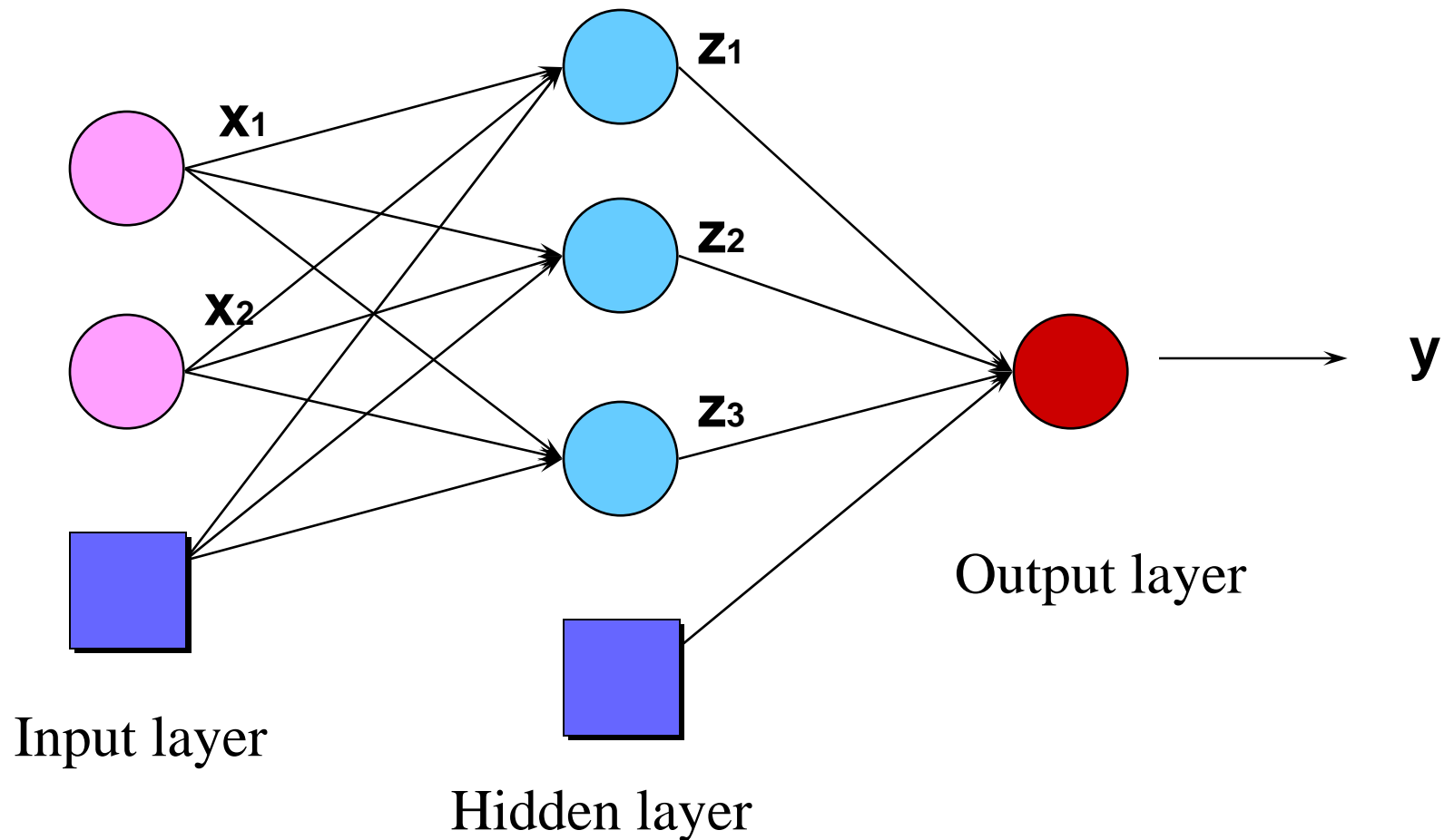
The Data

- **1618 mammograms showing clustered microcalcifications**
 - Biostatistics Dept Institut Curie
- **Variables**
 - Response: Malignant or not
 - Predictors: Age, Tissue Type (light/dense) Size (mm), Number of microcalc, Number of suspicious clusters, Shape of microcalc (1-5), Polyshape?(y/n), Shape of cluster (1,2,3), Retro (cluster near nipple?), Deep? (y/n)

Tree model



Neural Network

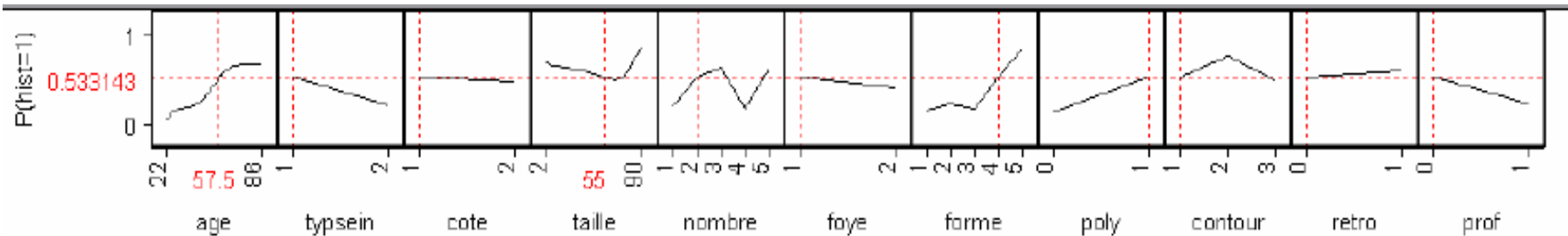


Put It Together

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

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

Sensitivity Profiler



	False Positive Rate	False Negative Rate
Simple Tree	32.2%	33.7%
Neural Network	25.5%	31.7%
Radiologists	22.4%	30.8%

Bagged Trees

- **B(ootstrap) Agg(regat)ed Trees**
- **Sample with replacement from training data**
 - Fit a “small” tree with a subset of predictors
 - Predict response
 - Repeat 1000 times
 - Average the predictions over the 1000 trees
 - Random Forests
 - Also selects random subset of predictors

Boosted Trees

- **Fit a small tree**
 - Downweight the data that are correctly predicted
 - Refit a small tree with weighted data
 - Repeat
 - Average the trees with weights proportional to % correct
 - Implementation in TreeNet – Salford Systems
- **Avoids overfitting**

Results

- **Split data into train and test (62.5% - 37.5%)**
- **Repeat random splits 1000 times**
 - For each iteration, count false positives and false negatives on the 600 test set cases

	False Positive Rate	False Negative Rate
Simple Tree	32.2%	33.7%
Neural Network	25.5%	31.7%
Boosted Trees	24.9%	32.5%
Bagged Trees	19.3%	28.8%
Radiologists	22.4%	30.8%

Lesson 5: Machines are Smart – You are Smarter

- **Why do modelers like interpretability?**
- **Black boxes are not interpretable, but there may be important information**

Case Study – Warranty Data

- **A new backpack inkjet printer is showing higher than expected warranty claims**

- What are the important variables?
- What's going on?



- **A neural networks shows that Zip code is the most important predictor**



Zip Code?



Data Mining – DOE Synergy

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

Take Home Messages

- **You have more data than you think**
 - Learn to make friends so you can use it
 - Listen to others so that analysis makes sense – Twyman's Law
- **Data preparation is most of the work**
 - Make friends
 - Modeling is the fun part
- **Keep abreast of technological developments**
 - Automatic modeling techniques
 - Web site
- **Don't worry about machines replacing you**
 - There's plenty of work left

Thank You!