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

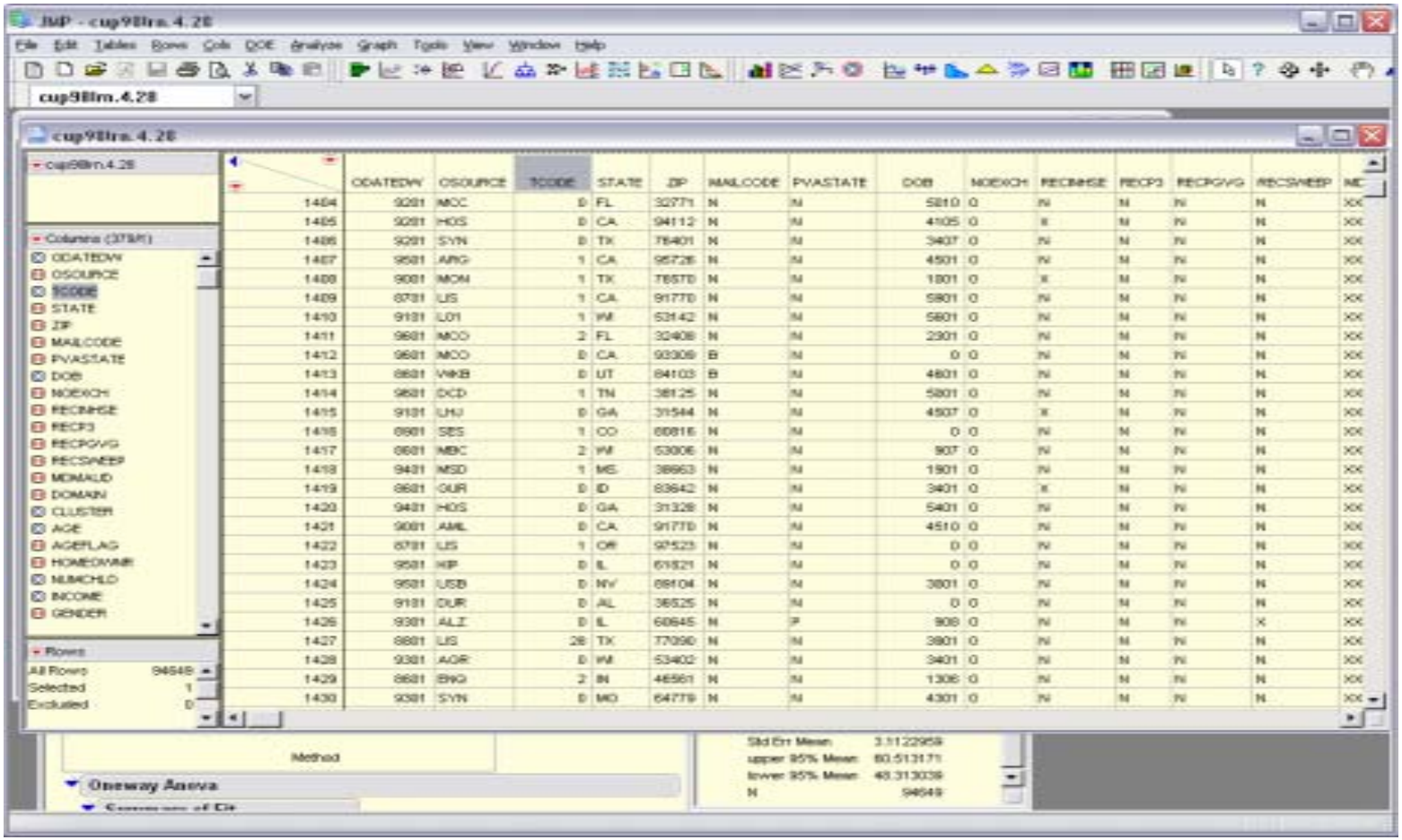

**Richard D. De Veaux**

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


# 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?**

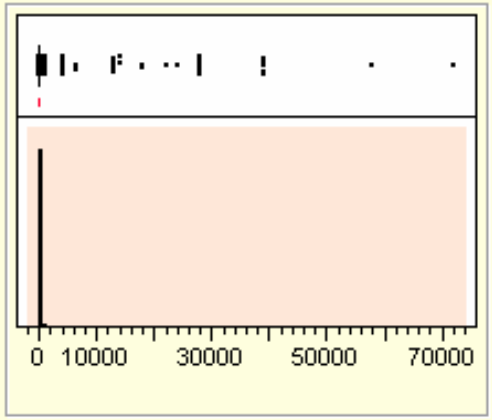
# What's "Hard"? --Example



# T-Code

Distributions

TCODE




Quantiles

| | | |
|---|---|---|
| 100.0% | maximum | 72002 |
| 99.5% | | 1002 |
| 97.5% | | 28 |
| 90.0% | | 28 |
| 75.0% | quartile | 2 |
| 50.0% | median | 1 |
| 25.0% | quartile | 0 |
| 10.0% | | 0 |
| 2.5% | | 0 |
| 0.5% | | 0 |
| 0.0% | minimum | 0 |

Moments

| | |
|---|---|
| Mean | 54.413105 |
| Std Dev | 957.50024 |
| Std Err Mean | 3.1122959 |
| upper 95% Mean | 60.513171 |
| lower 95% Mean | 48.313039 |
| N | 94649 |

# What does it mean?

| T-Code | Title | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | _ | 16 | DEAN | 48 | CORPORAL | 109 | LIC. | | | | |
| 1 | MR. | 17 | JUDGE | 50 | ELDER | 111 | SA. | | | | |
| 1001 | MESSRS. | 17002 | JUDGE & MRS. | 56 | MAYOR | 114 | DA. | | | | |
| 1002 | MR. & MRS. | 18 | MAJOR | 59002 | LIEUTENANT & MRS. | 116 | SR. | | | | |
| 2 | MRS. | 18002 | MAJOR & MRS. | 62 | LORD | 117 | SRA. | | | | |
| 2002 | MESDAMES | 19 | SENATOR | 63 | CARDINAL | 118 | SRTA. | | | | |
| 3 | MISS | 20 | GOVERNOR | 64 | FRIEND | 120 | YOUR MAJESTY | | | | |
| 3003 | MISSES | 21002 | SERGEANT & MRS. | 65 | FRIENDS | 122 | HIS HIGHNESS | | | | |
| 4 | DR. | 22002 | COLNEL & MRS. | 68 | ARCHDEACON | 123 | HER HIGHNESS | | | | |
| 4002 | DR. & MRS. | 24 | LIEUTENANT | 69 | CANON | 124 | COUNT | | | | |
| 4004 | DOCTORS | 26 | MONSIGNOR | 70 | BISHOP | 125 | LADY | | | | |
| 5 | MADAME | 27 | REVEREND | 72002 | REVEREND & MRS. | 126 | PRINCE | | | | |
| 6 | SERGEANT | 28 | MS. | 73 | PASTOR | 127 | PRINCESS | | | | |
| 9 | RABBI | 28028 | MSS. | 75 | ARCHBISHOP | 128 | CHIEF | | | | |
| 10 | PROFESSOR | 29 | BISHOP | 85 | SPECIALIST | 129 | BARON | | | | |
| 10002 | PROFESSOR & MRS. | 31 | AMBASSADOR | 87 | PRIVATE | 130 | SHEIK | | | | |
| 10010 | PROFESSORS | 31002 | AMBASSADOR & MRS | 89 | SEAMAN | 131 | PRINCE AND PRINCESS | | | | |
| 11 | ADMIRAL | 33 | CANTOR | 90 | AIRMAN | 132 | YOUR IMPERIAL MAJEST | | | | |
| 11002 | ADMIRAL & MRS. | 36 | BROTHER | 91 | JUSTICE | 135 | M. ET MME. | | | | |
| 12 | GENERAL | 37 | SIR | 92 | MR. JUSTICE | 210 | PROF. | | | | |
| 12002 | GENERAL & MRS. | 38 | COMMODORE | 100 | M. | | | | | | |
| 13 | COLONEL | 40 | FATHER | 103 | MLLE. | | | | | | |
| 13002 | COLONEL & MRS. | 42 | SISTER | 104 | CHANCELLOR | | | | | | |
| 14 | CAPTAIN | 43 | PRESIDENT | 106 | REPRESENTATIVE | | | | | | |
| 14002 | CAPTAIN & MRS. | 44 | MASTER | 107 | SECRETARY | | | | | | |
| 15 | COMMANDER | 46 | MOTHER | 108 | LT. GOVERNOR | | | | | | |
| 15002 | COMMANDER & MRS. | 47 | CHAPLAIN | | | | | | | | |

# 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

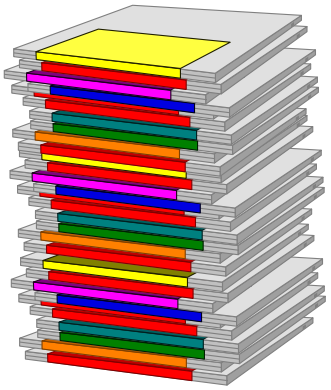
| ItemID | ItemName | price |
|---|---|---|
| C56621 | top hat | 34.95 |
| T35691 | cane | 4.99 |
| RS5292 | red shoes | 22.95 |

Shoppers

| Person ID | person name | ZIPCODE | item bought |
|---|---|---|---|
| 135366 | Lyle | 19103 | T35691 |
| 135366 | Lyle | 19103 | C56621 |
| 259835 | Dick | 01267 | RS5292 |

# Data Challenges

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



# Data Preparation

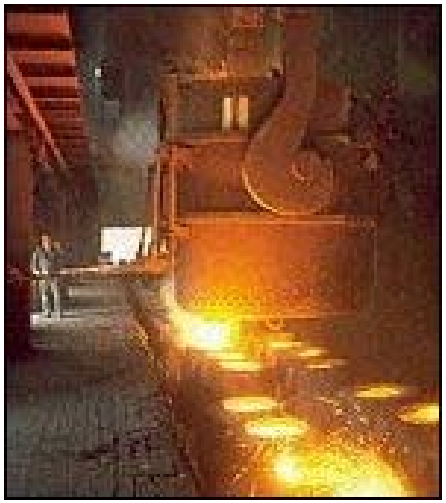
- **Build data mining database**
  - Combining sources
  - Synchronizing sources
- **Explore data**
- **Prepare data for modeling**

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

# 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

# Ingot cracking

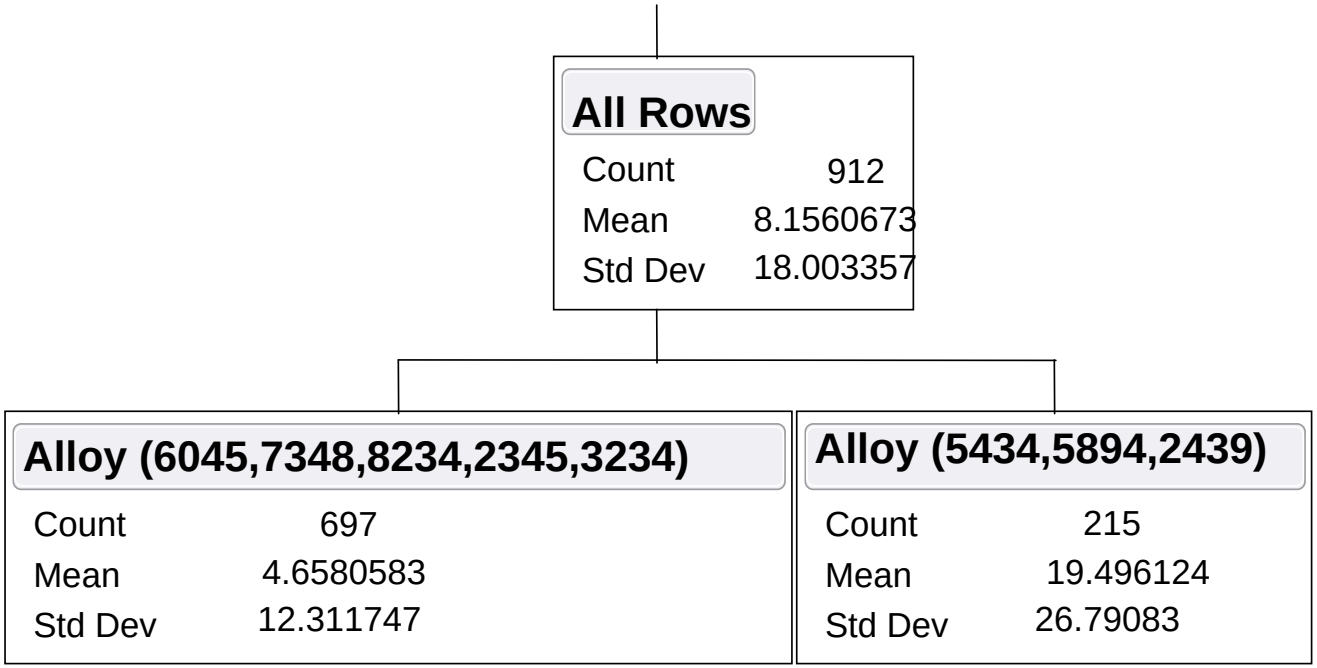


- **953 30,000 lb. Ingots**
  - 20% cracking rate
  - $30,000 per recast
  - 90 potential explanatory variables
    - ✓Water composition (reduced)
    - ✓Metal composition
    - ✓Process variables
    - ✓Other environmental variables

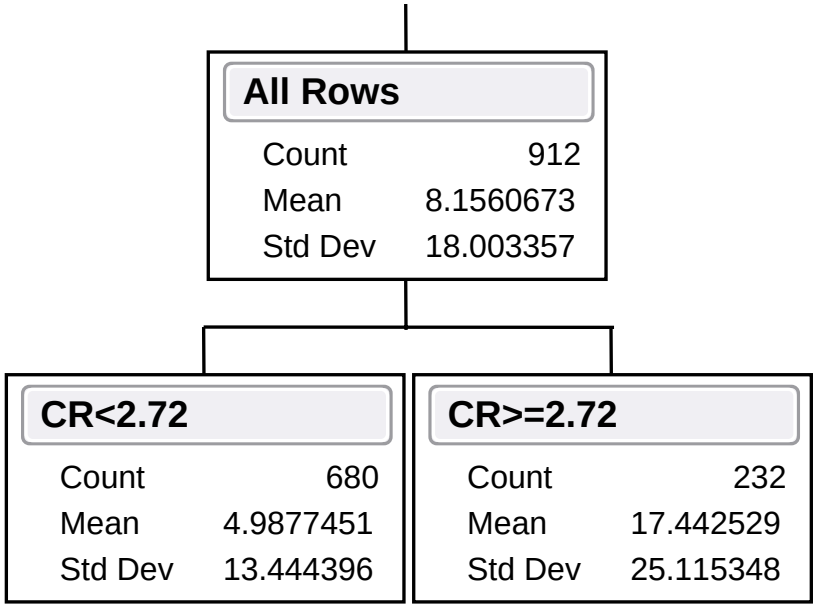
# Data Processing

- **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

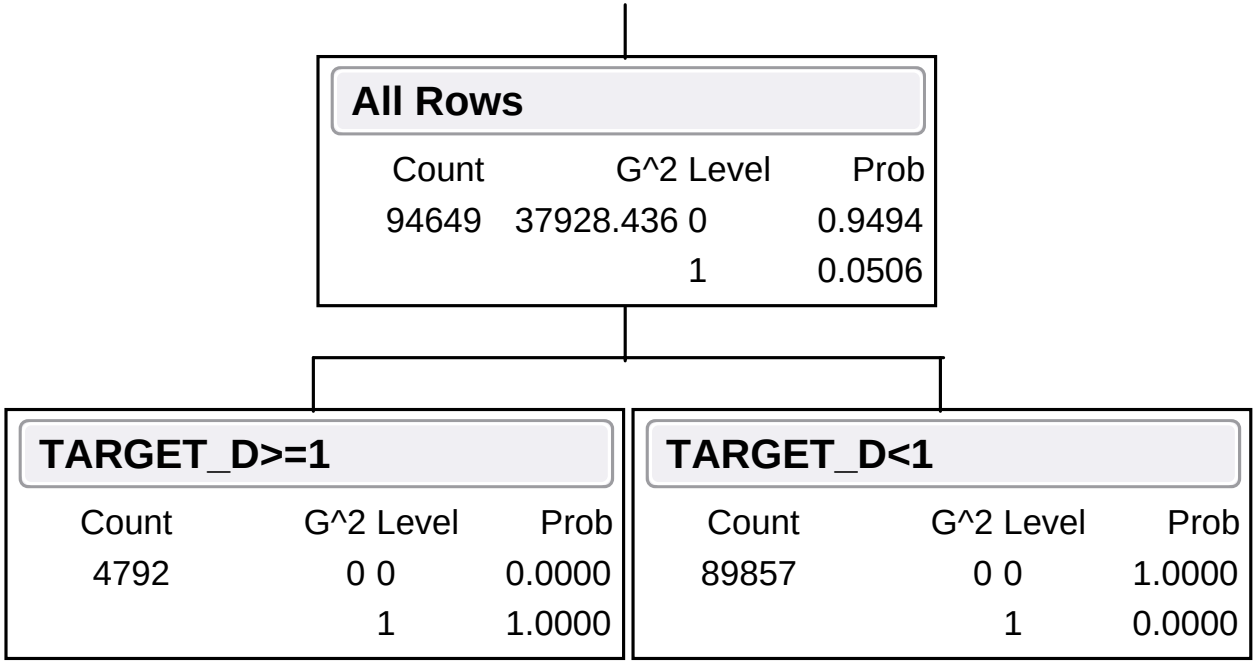
# First Tree


All Rows
Count 912
Mean 8.1560673
Std Dev 18.003357
Alloy (6045,7348,8234,2345,3234)
Count 697
Mean 4.6580583
Std Dev 12.311747
Alloy (5434,5894,2439)
Count 215
Mean 19.496124
Std Dev 26.79083

# Second Tree


All Rows
Count 912
Mean 8.1560673
Std Dev 18.003357
CR<2.72
Count 680
Mean 4.9877451
Std Dev 13.444396
CR>=2.72
Count 232
Mean 17.442529
Std Dev 25.115348

# PVA Data: Herb's Tree



# What did we learn?

- **Data mining gave clues for generating hypotheses**
- **Followed up with DOE**
- **DOE led to substantial process improvement**

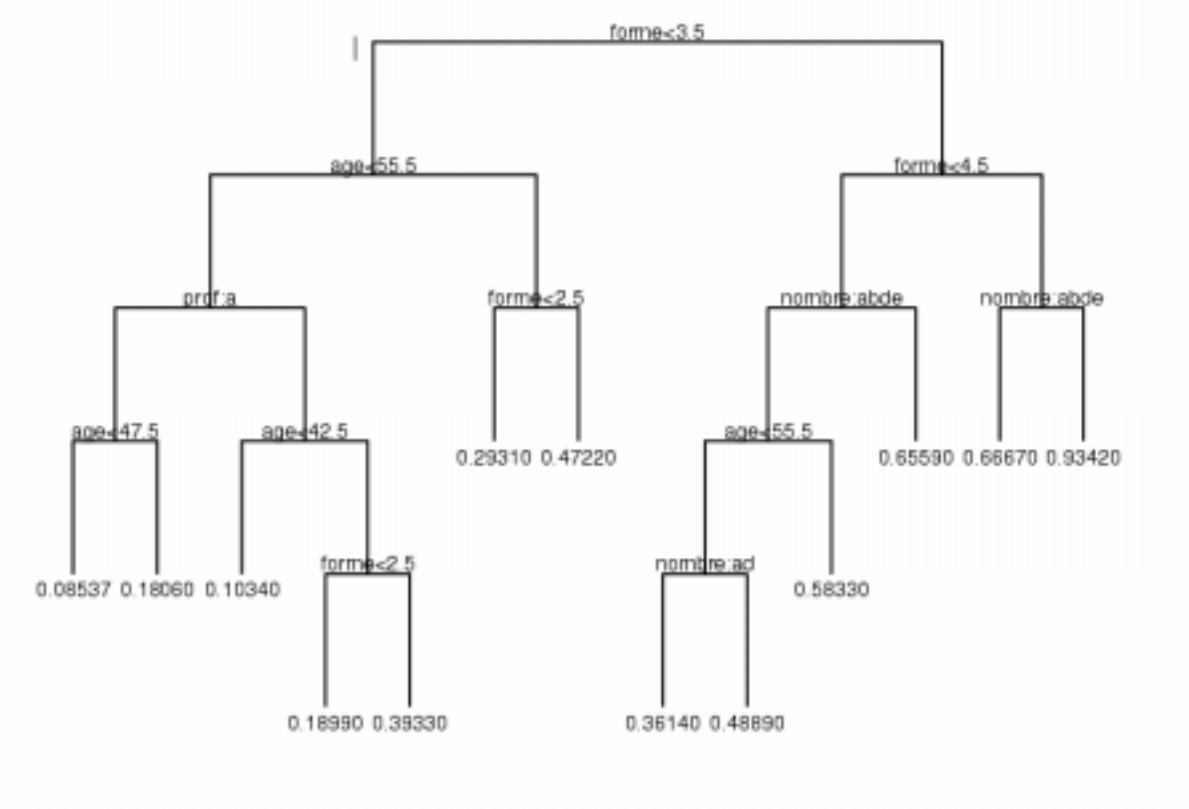
# Lesson 3: Know When to Hold 'em

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

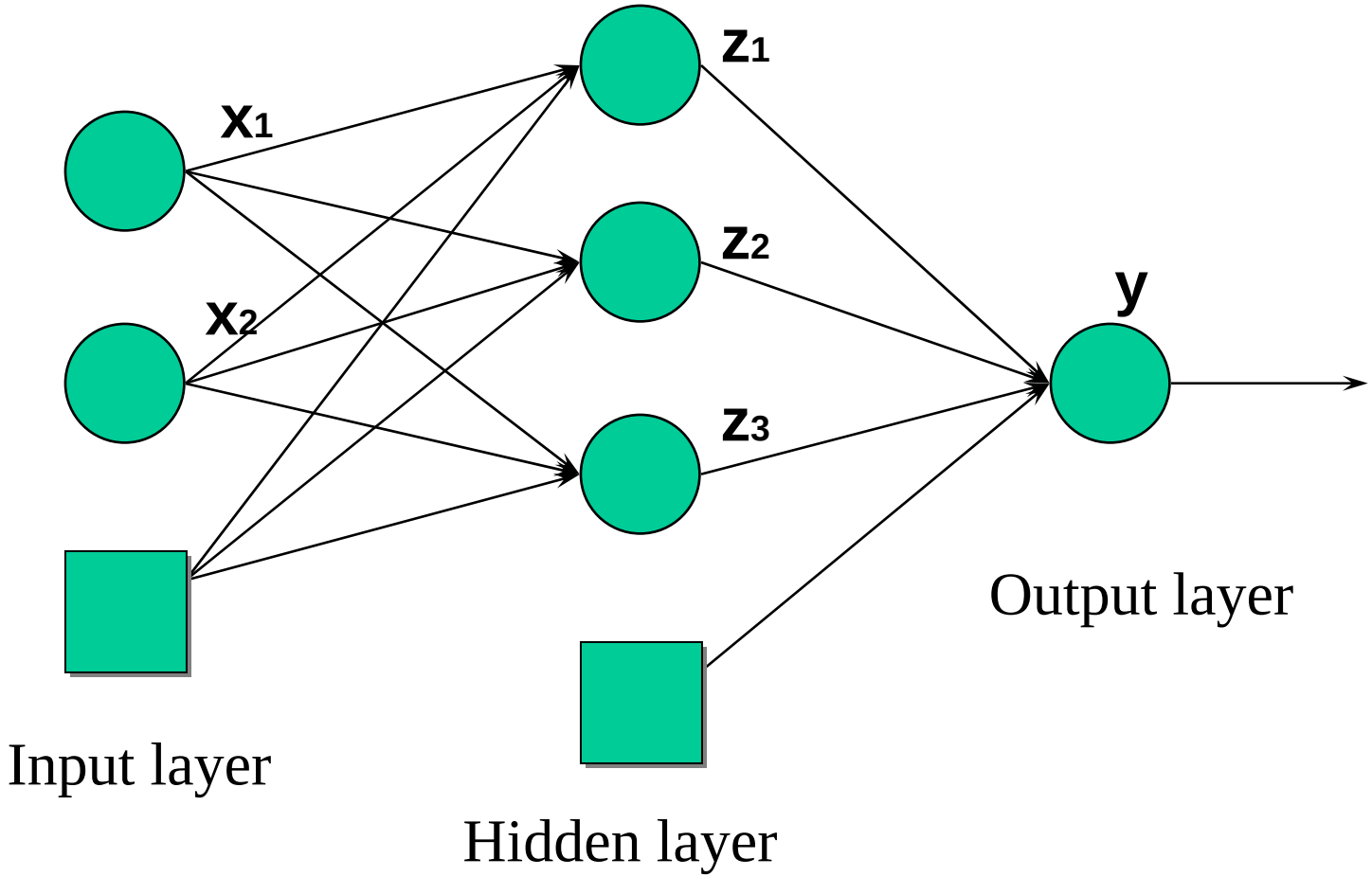
# The Data

- **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


forme<3.5
age<55.5
forme<4.5
prof:a
forme<2.5
nombre:abde
nombre:abde
age<47.5
age<42.5
0.29310 0.47220
age<55.5
0.65590 0.66670 0.93420
0.08537 0.18060 0.10340
forme<2.5
nombre:ad
0.58330
0.18990 0.39330
0.36140 0.48890

# Neural Network

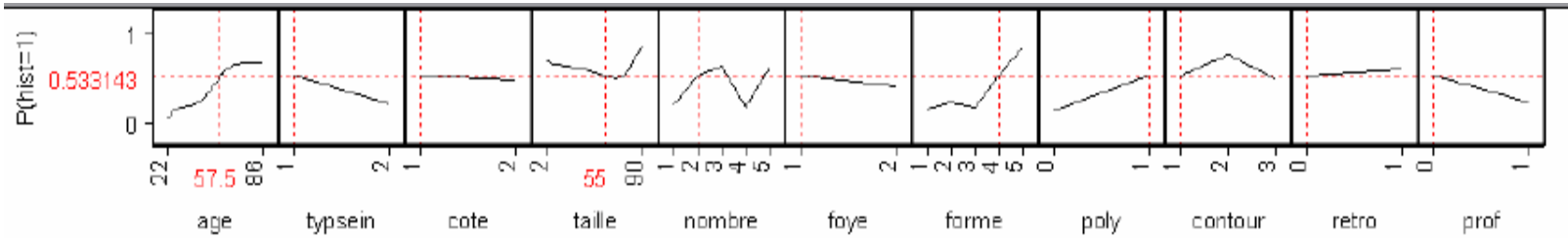


# Put It Together

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

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

# Neural Network



# 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

# 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
- **Avoids overfitting**

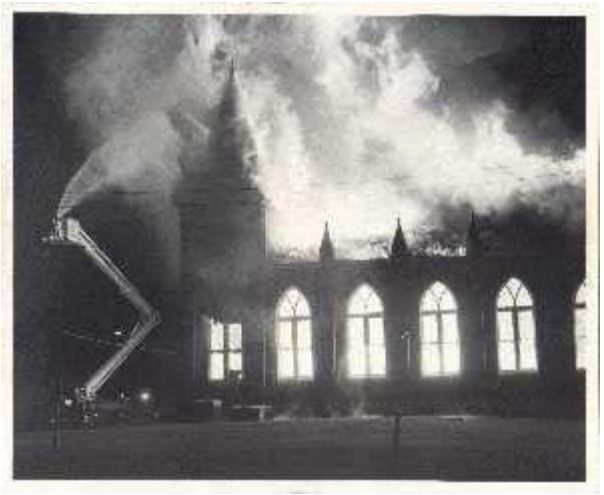
# Results

- **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 Positives | False Negatives |
|---|---|---|
| **Simple Tree** | 32.20% | 33.70% |
| **Neural Network** | 25.50% | 31.70% |
| **Boosted Trees** | 24.90% | 32.50% |
| **Bagged Trees** | 19.30% | 28.80% |
| **Radiologists** | **22.40%** | **35.80%** |

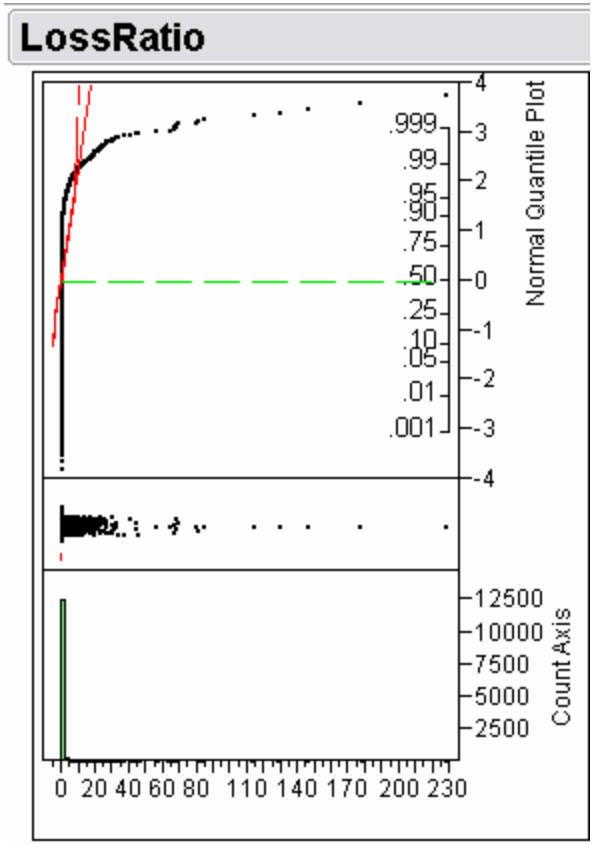
# Lesson 4: Know When to Fold 'em

- **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)

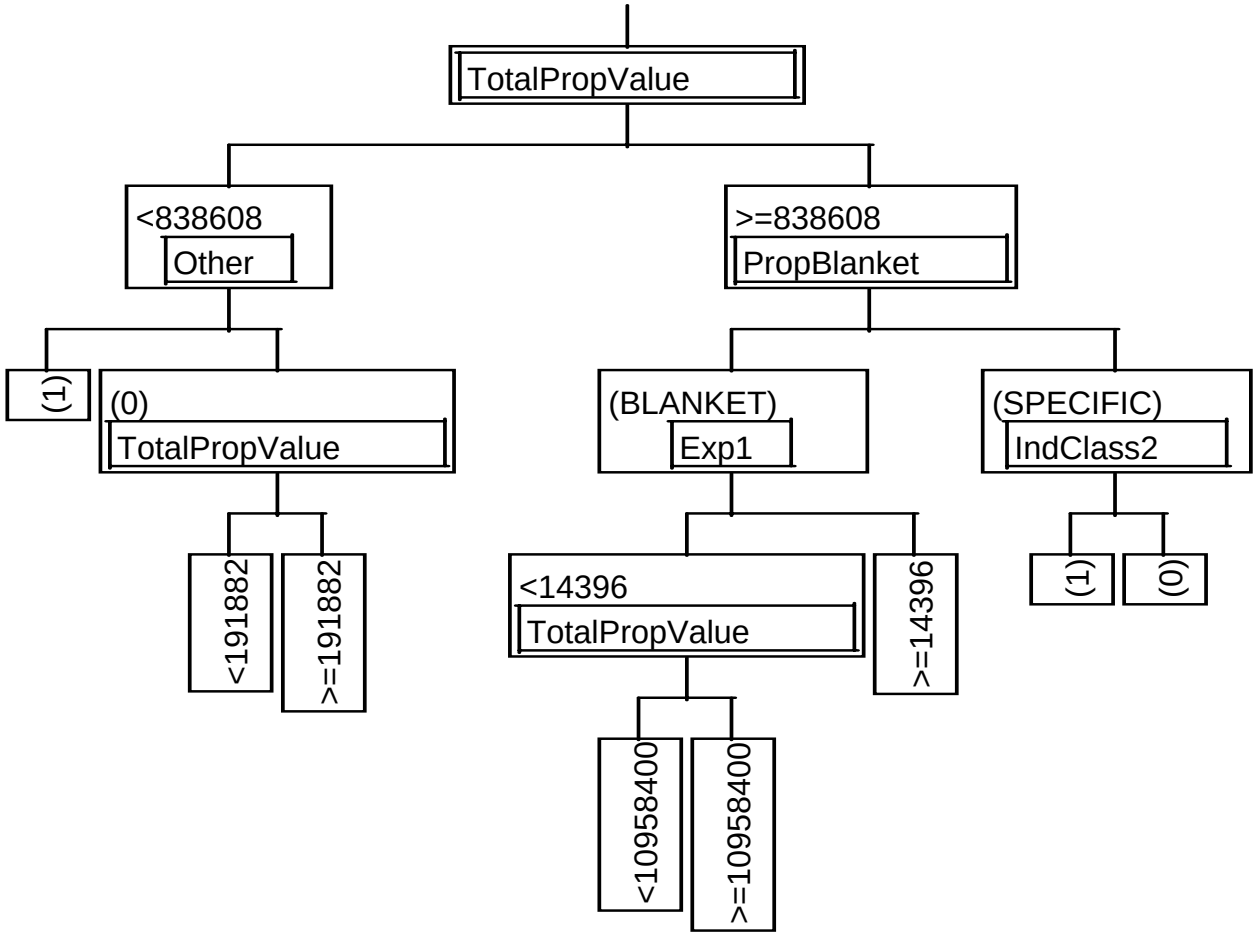


# Churches – First Steps

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



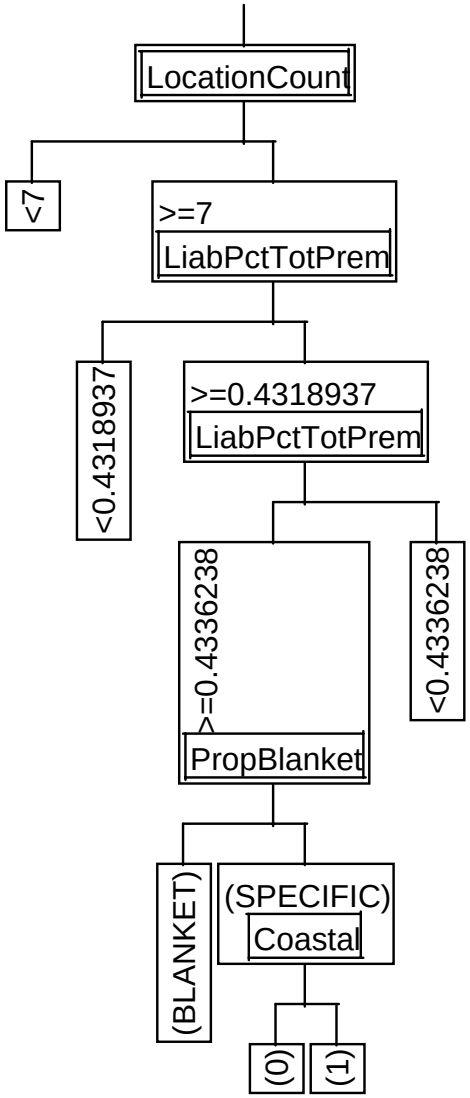
# First Tree


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**

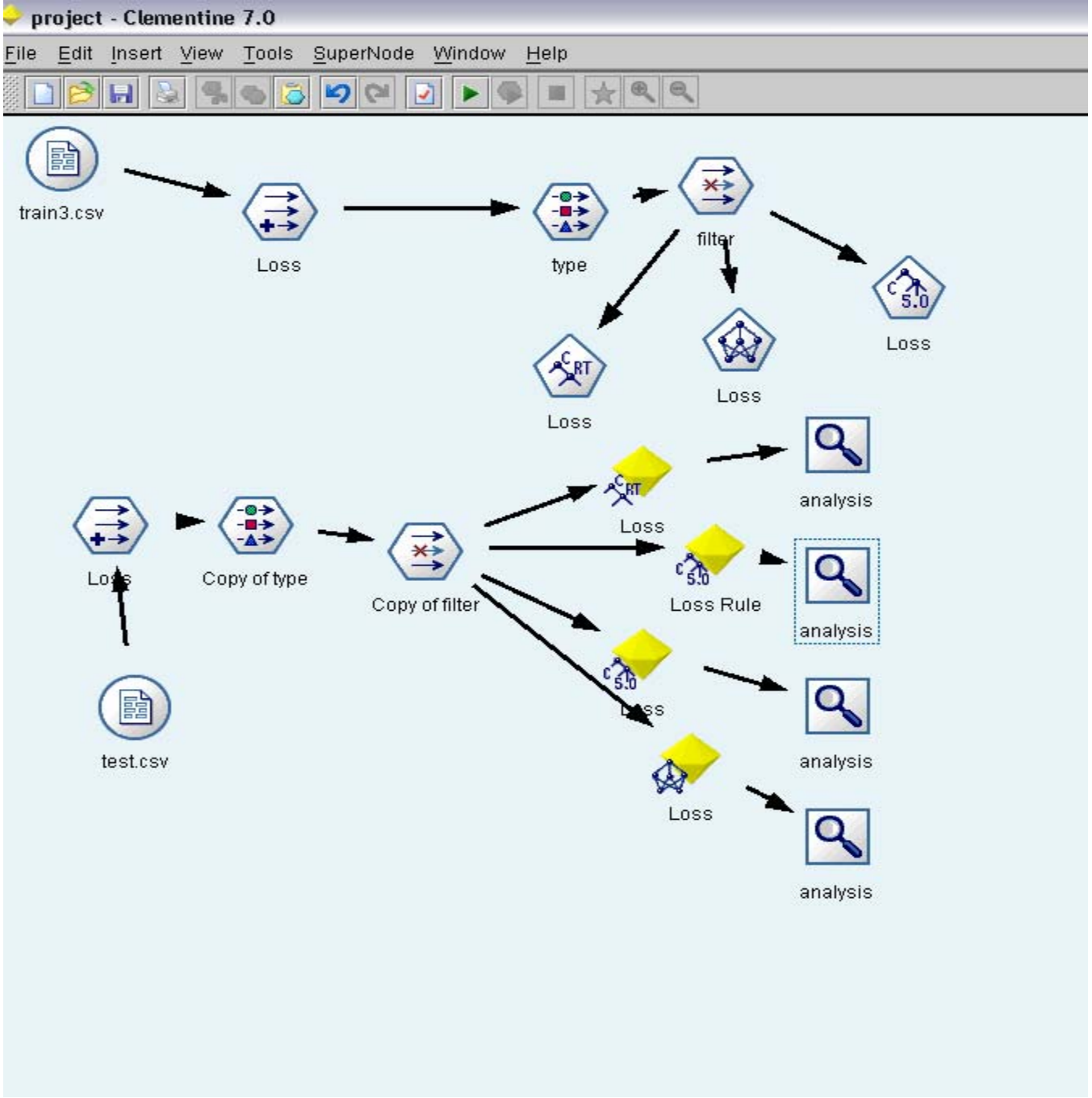
# Next Tree




Where to go from here?

# 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
Loss
Copy of type
Copy of filter
Loss
Loss Rule
analysis
analysis
analysis
analysis
test.csv
Loss

# KXEN

- **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 << k.
  - Summary of model fit : KI and KR
- **For Church data KI = -0.012  KR = 0.034**

# Lesson 5: Machines are Smart – You are Smarter

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

# 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**

# Take Home Messages

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