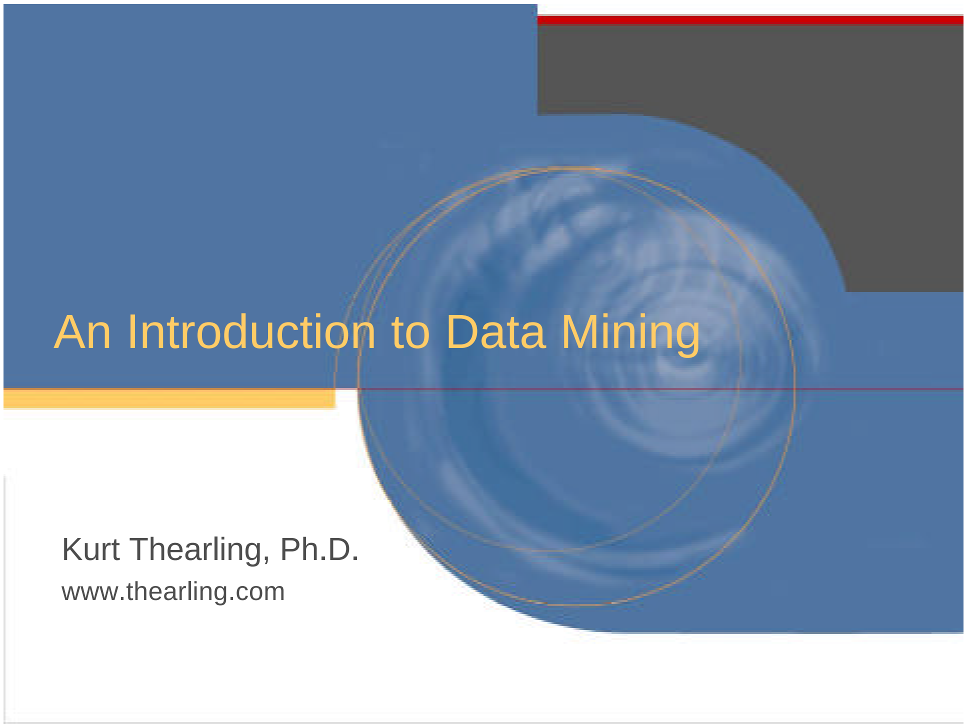
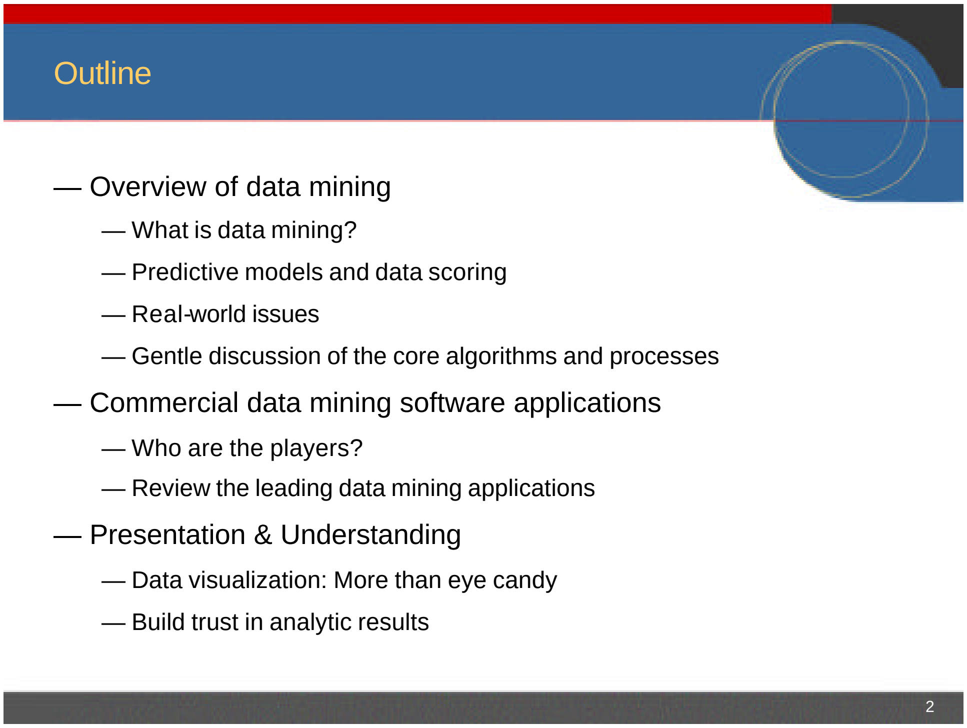


An Introduction to Data Mining

Kurt Thearling, Ph.D.
www.thearling.com

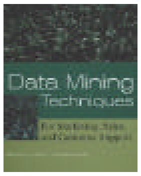


Outline

- Overview of data mining
  - What is data mining?
  - Predictive models and data scoring
  - Real-world issues
  - Gentle discussion of the core algorithms and processes
- Commercial data mining software applications
  - Who are the players?
  - Review the leading data mining applications
- Presentation & Understanding
  - Data visualization: More than eye candy
  - Build trust in analytic results

Resources

- Good overview book:
  - *Data Mining Techniques* by Michael Berry and Gordon Linoff



- Web:
  - My web site (recommended books, useful links, white papers, …)
    - > http://www.thearling.com
  - Knowledge Discovery Nuggets
    - > http://www.kdnuggets.com
- DataMine Mailing List
  - majordomo@quality.org
  - send message “subscribe datamine-l”

A Problem...

- You are a marketing manager for a brokerage company
  - Problem: Churn is too high
    - > Turnover (after six month introductory period ends) is 40%
  - Customers receive incentives (average cost: $160) when account is opened

  - Giving new incentives to everyone who might leave is very expensive (as well as wasteful)
  - Bringing back a customer after they leave is both difficult and costly

… A Solution

- One month before the end of the introductory period is over, predict which customers will leave
  - If you want to keep a customer that is predicted to churn, offer them something based on their predicted value
    - > The ones that are not predicted to churn need no attention
  - If you don't want to keep the customer, do nothing
- How can you predict future behavior?
  - Tarot Cards
  - Magic 8 Ball

The Big Picture

- Lots of hype & misinformation about data mining out there
- Data mining is part of a much larger process
  - 10% of 10% of 10% of 10%
  - Accuracy not always the most important measure of data mining
- The data itself is critical
- Algorithms aren't as important as some people think
- If you can't understand the patterns discovered with data mining, you are unlikely to act on them (or convince others to act)

Defining Data Mining

— The automated extraction of predictive information from (large) databases

— Two key words:

  - ✍ Automated
  - ✍ Predictive

— Implicit is a statistical methodology

— Data mining lets you be proactive

  - — Prospective rather than Retrospective

Goal of Data Mining

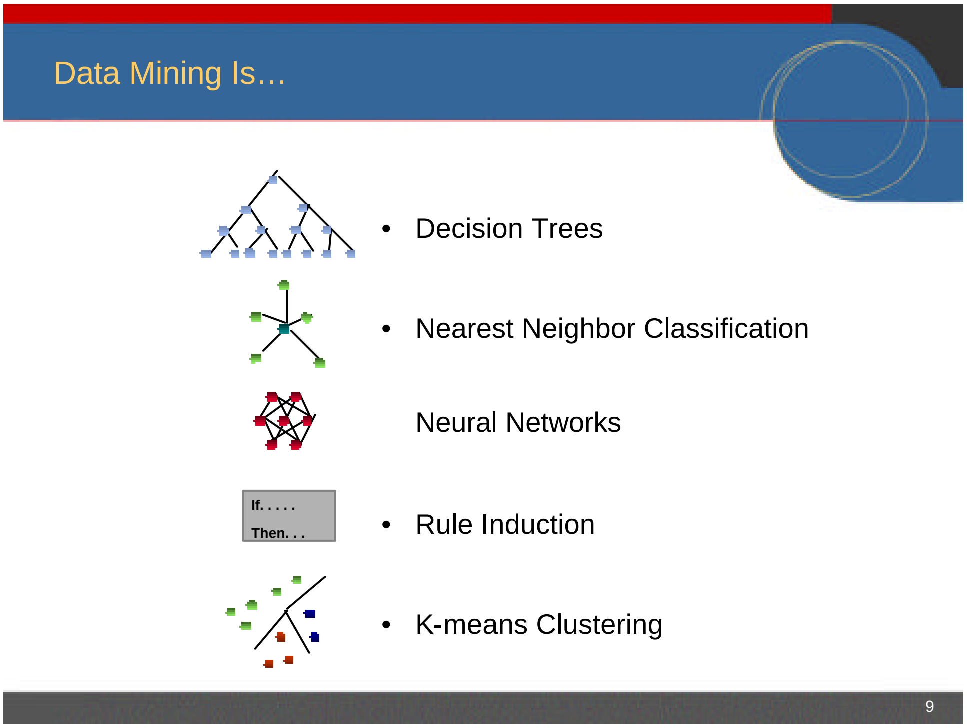
— Simplification and automation of the overall statistical process, from data source(s) to model application

— Changed over the years

  - — Replace statistician ✍ Better models, less grunge work

— 1 + 1 = 0

  - — Many different data mining algorithms / tools available
  - — Statistical expertise required to compare different techniques
  - — Build intelligence into the software

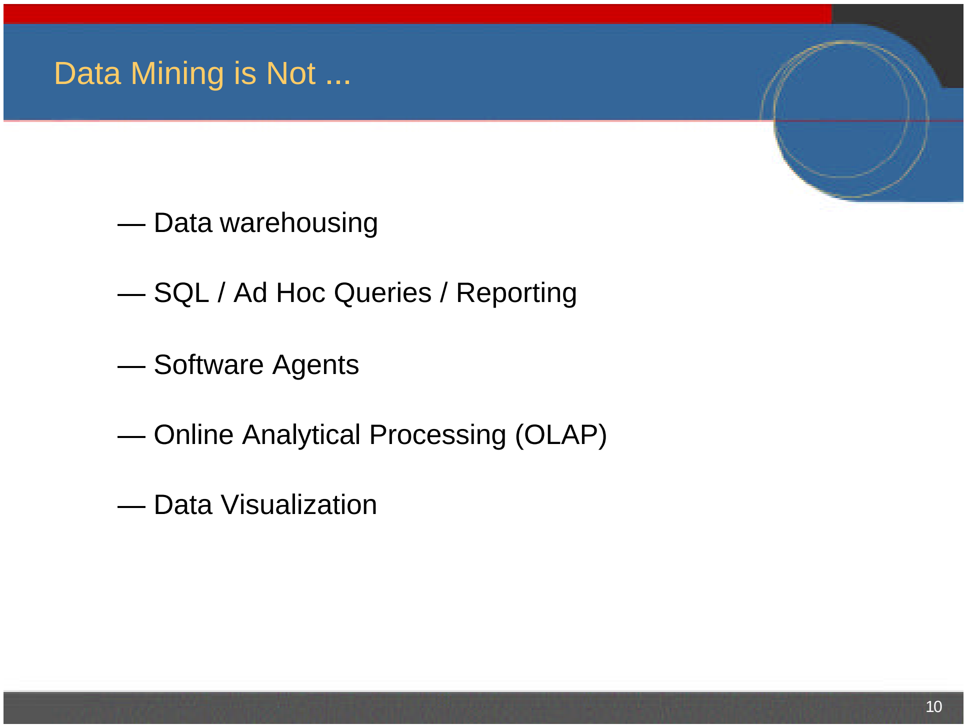


Data Mining Is…

- Decision Trees
- Nearest Neighbor Classification

Neural Networks

- Rule Induction
- K-means Clustering



Data Mining is Not ...

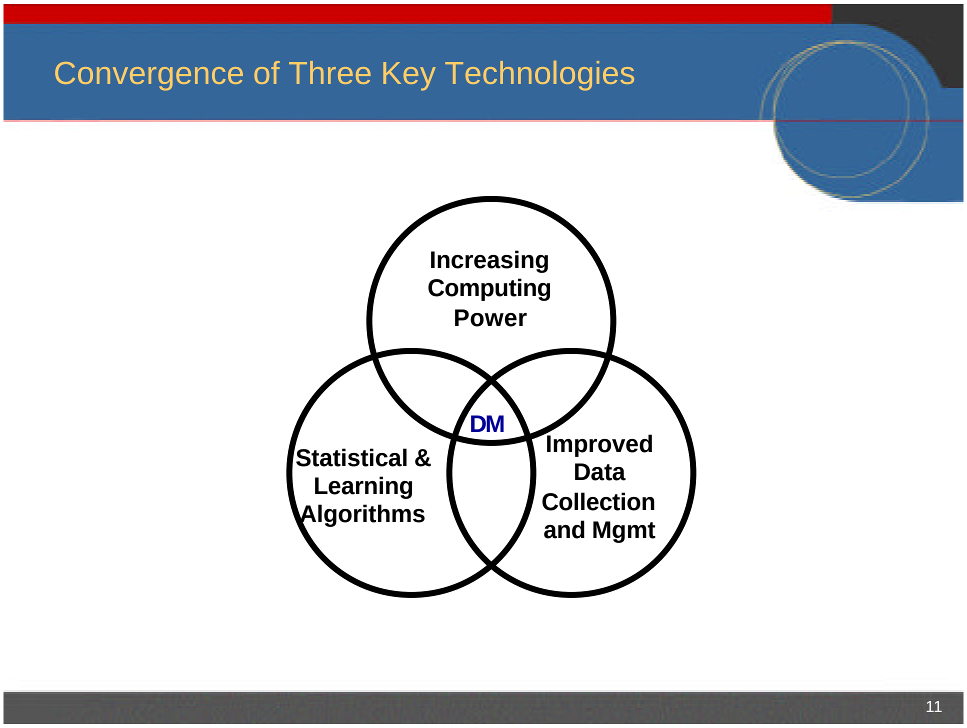
— Data warehousing

— SQL / Ad Hoc Queries / Reporting

— Software Agents

— Online Analytical Processing (OLAP)

— Data Visualization



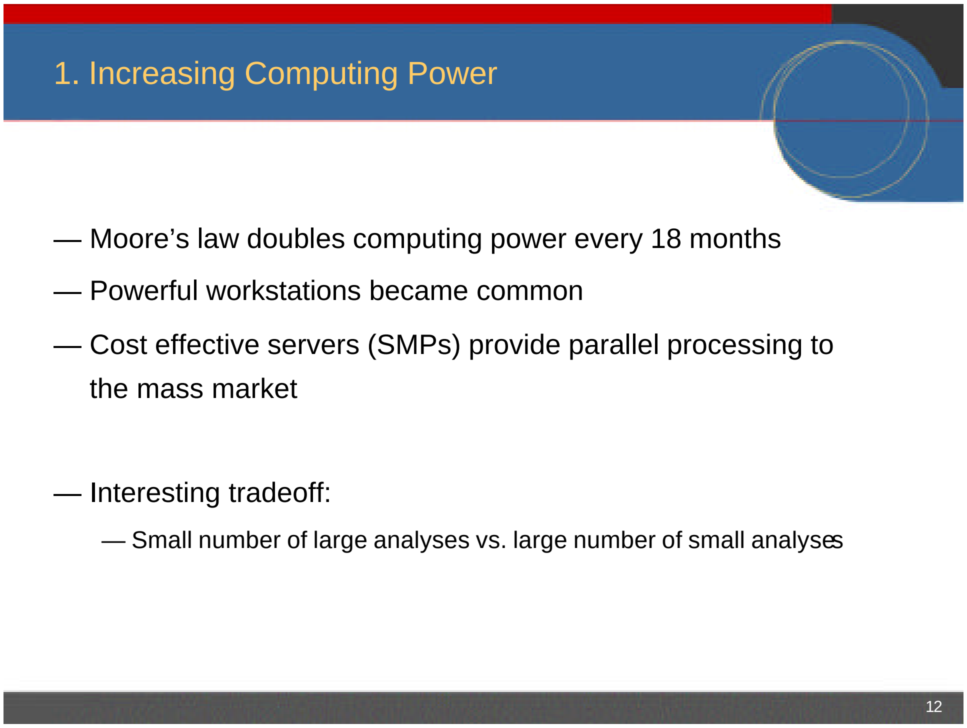
Convergence of Three Key Technologies

Increasing Computing Power

Statistical & Learning Algorithms

Improved Data Collection and Mgmt

DM



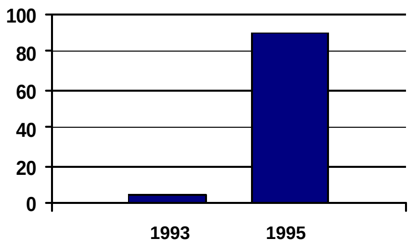
1. Increasing Computing Power

- Moore’s law doubles computing power every 18 months
- Powerful workstations became common
- Cost effective servers (SMPs) provide parallel processing to the mass market

- Interesting tradeoff:
  - Small number of large analyses vs. large number of small analyses

2. Improved Data Collection and Management

% CIOs Building Data Warehouses



- — Data Collection ✍ Access ✍ Navigation ✍ Mining
- — The more data the better (usually)

3. Statistical & Machine Learning Algorithms

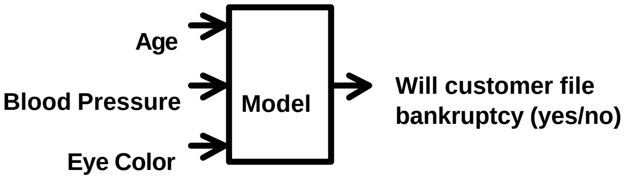
- — Techniques have often been waiting for computing technology to catch up
- — Statisticians already doing “manual data mining”
- — Good machine learning is just the intelligent application of statistical processes
- — A lot of data mining research focused on tweaking existing techniques to get small percentage gains

Common Uses of Data Mining

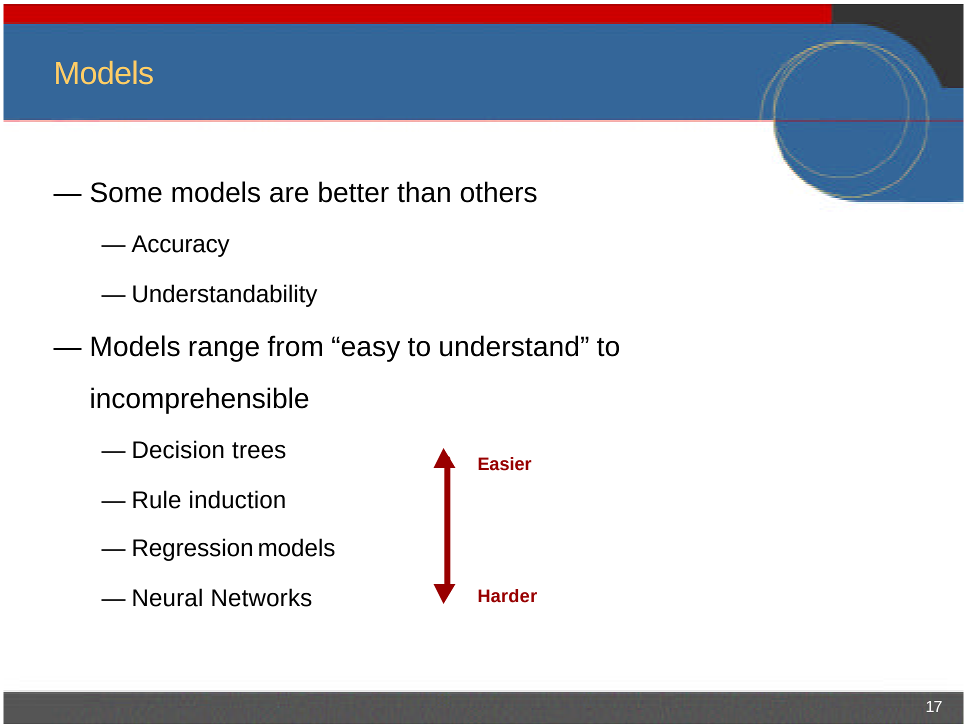
- Direct mail marketing
- Web site personalization
- Credit card fraud detection
  - Gas & jewelry
- Bioinformatics
- Text analysis
  - SAS lie detector
- Market basket analysis
  - Beer & baby diapers:

Definition: Predictive Model

- A “black box” that makes predictions about the future based on information from the past and present



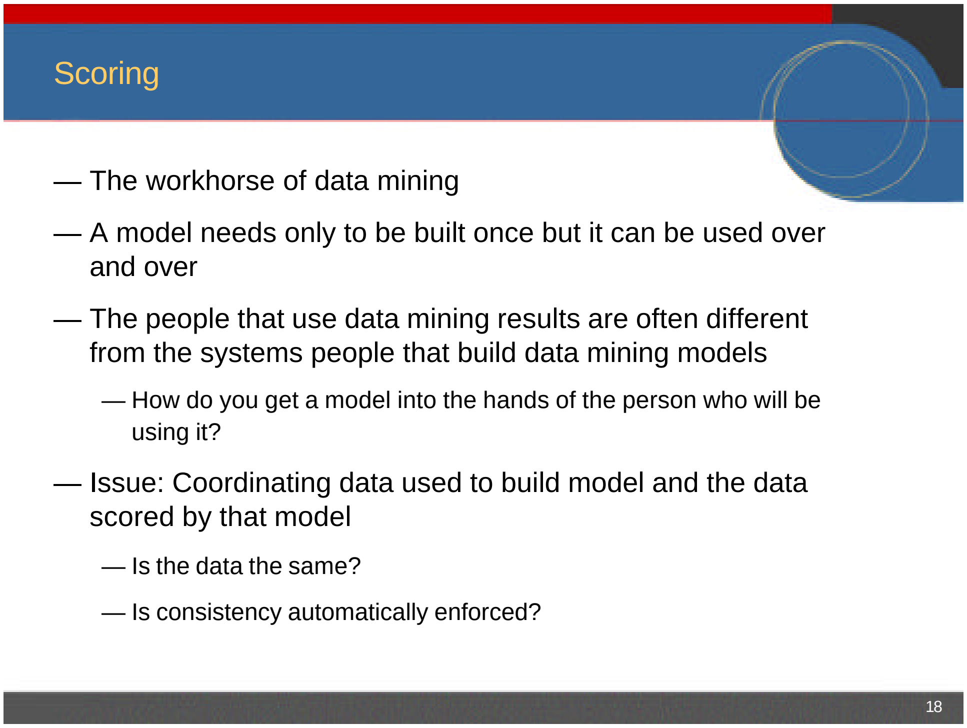
- Large number of inputs usually available



Models

- Some models are better than others
  - Accuracy
  - Understandability
- Models range from “easy to understand” to incomprehensible
  - Decision trees
  - Rule induction
  - Regression models
  - Neural Networks

Easier
Harder



Scoring

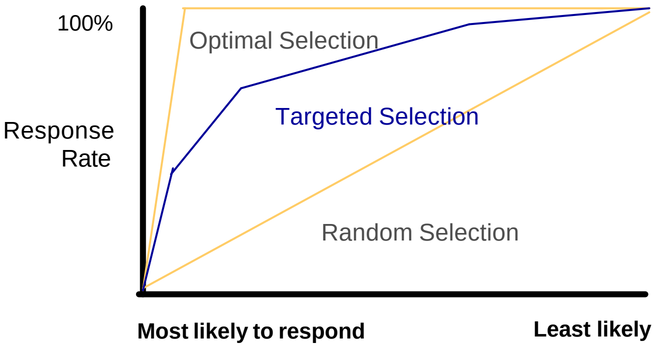
- The workhorse of data mining
- A model needs only to be built once but it can be used over and over
- The people that use data mining results are often different from the systems people that build data mining models
  - How do you get a model into the hands of the person who will be using it?
- Issue: Coordinating data used to build model and the data scored by that model
  - Is the data the same?
  - Is consistency automatically enforced?

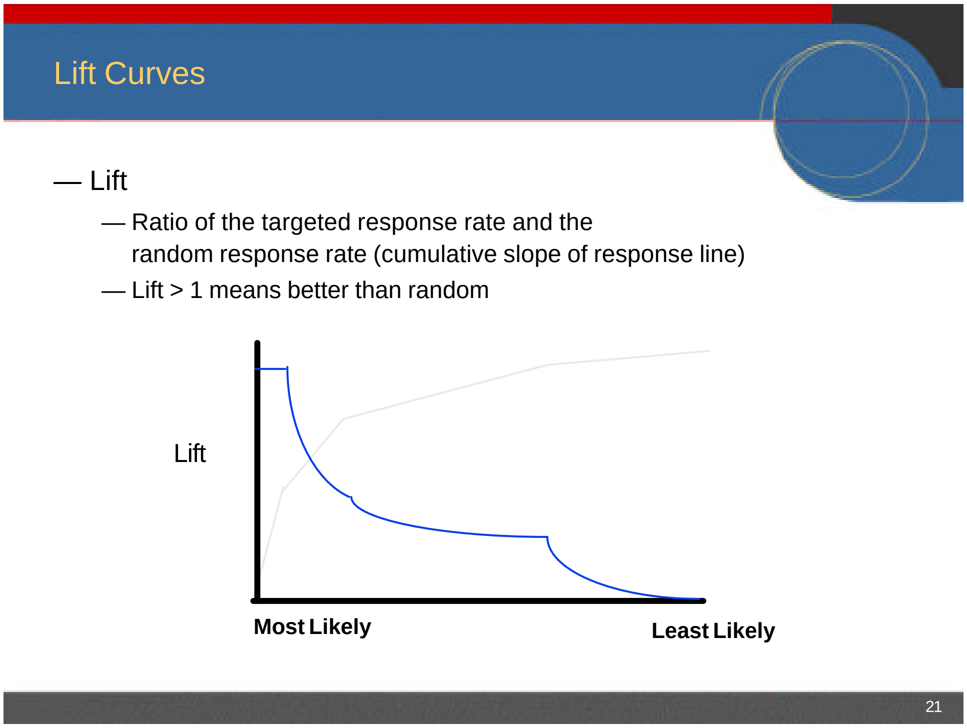
Two Ways to Use a Model

- — Qualitative
  - — Provide insight into the data you are working with
    - > If city = New York and 30 < age < 35 …
    - > Important age demographic was previously 20 to 25
    - > Change print campaign from Village Voice to New Yorker
  - — Requires interaction capabilities and good visualization
- — Quantitative
  - — Automated process
  - — Score new gene chip datasets with error model every night at midnight
  - — Bottom-line orientation

How Good is a Predictive Model?

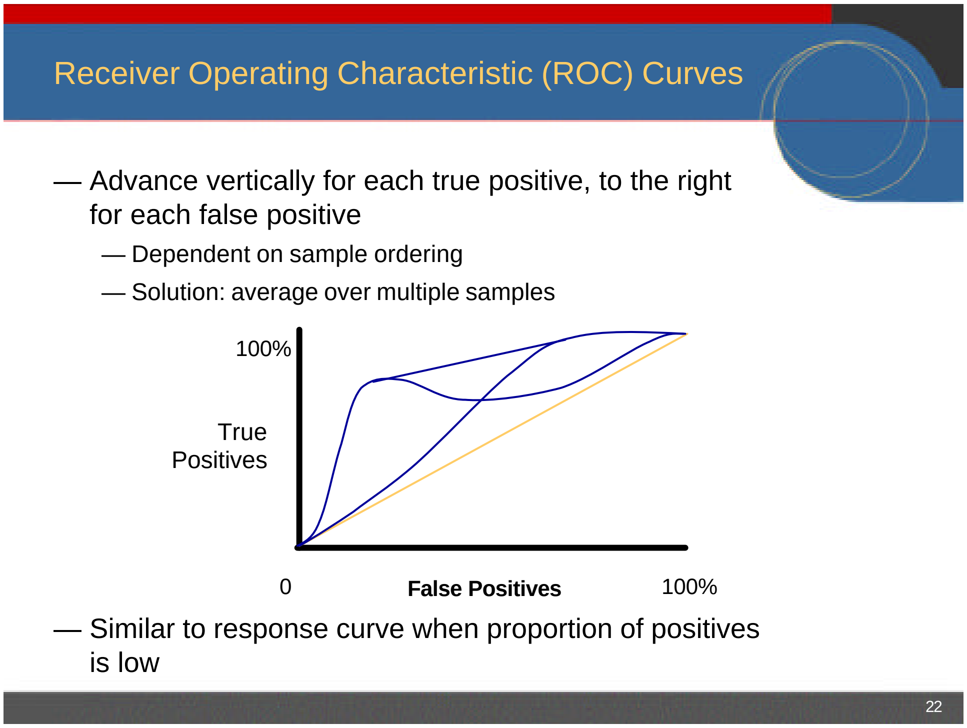
- — Response curves
  - — How does the response rate of a targeted selection compare to a random selection?





Lift Curves

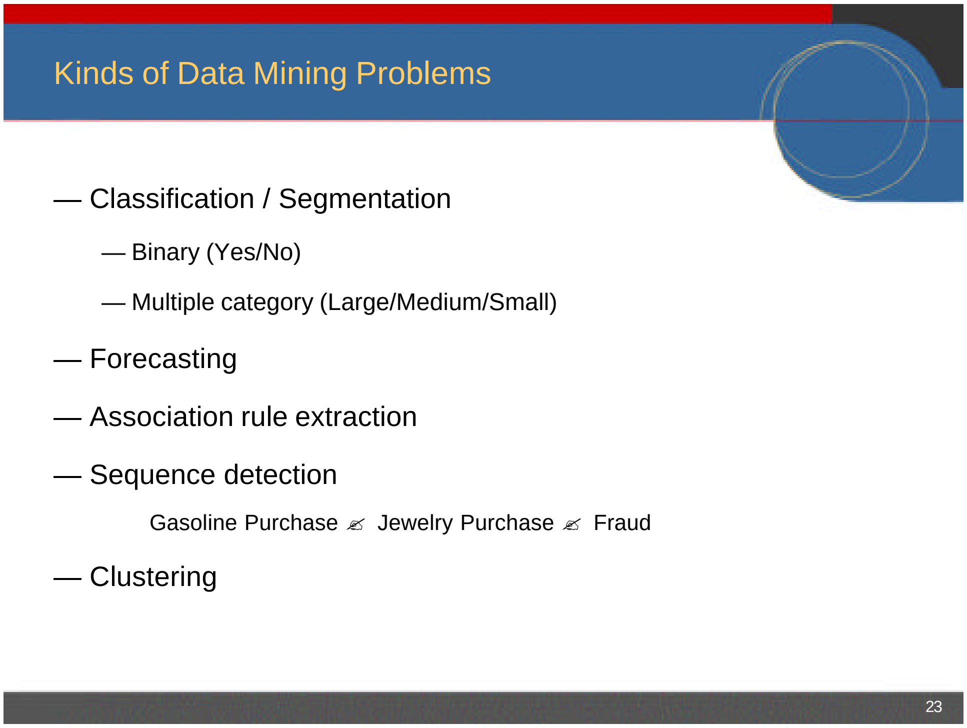
- — Lift
  - — Ratio of the targeted response rate and the random response rate (cumulative slope of response line)
  - — Lift > 1 means better than random



Receiver Operating Characteristic (ROC) Curves

- — Advance vertically for each true positive, to the right for each false positive
  - — Dependent on sample ordering
  - — Solution: average over multiple samples

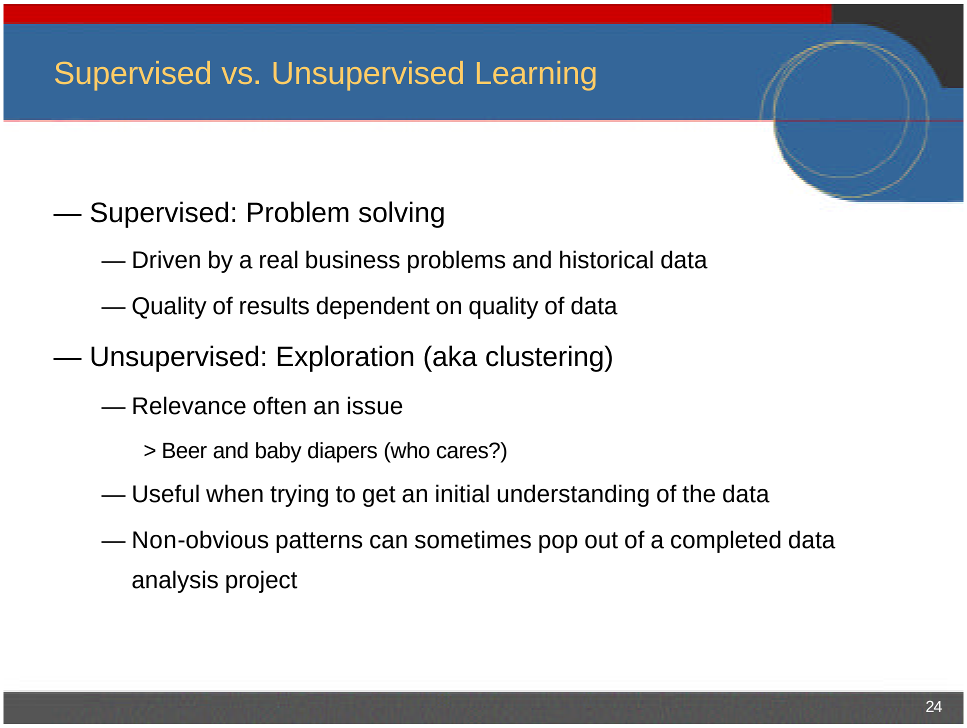
- — Similar to response curve when proportion of positives is low



Kinds of Data Mining Problems

- Classification / Segmentation
  - Binary (Yes/No)
  - Multiple category (Large/Medium/Small)
- Forecasting
- Association rule extraction
- Sequence detection

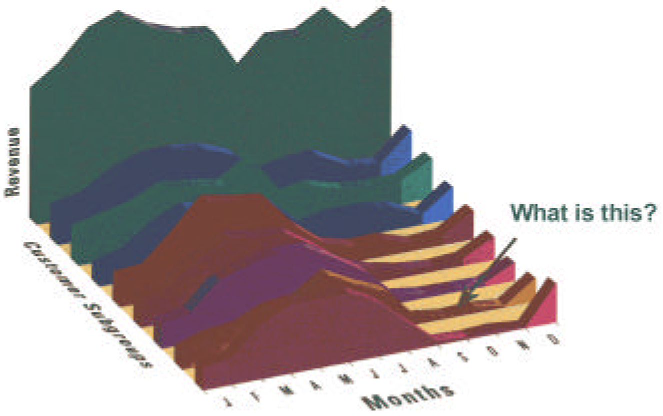
  Gasoline Purchase ✍ Jewelry Purchase ✍ Fraud
- Clustering



Supervised vs. Unsupervised Learning

- Supervised: Problem solving
  - Driven by a real business problems and historical data
  - Quality of results dependent on quality of data
- Unsupervised: Exploration (aka clustering)
  - Relevance often an issue
    - > Beer and baby diapers (who cares?)
  - Useful when trying to get an initial understanding of the data
  - Non-obvious patterns can sometimes pop out of a completed data analysis project

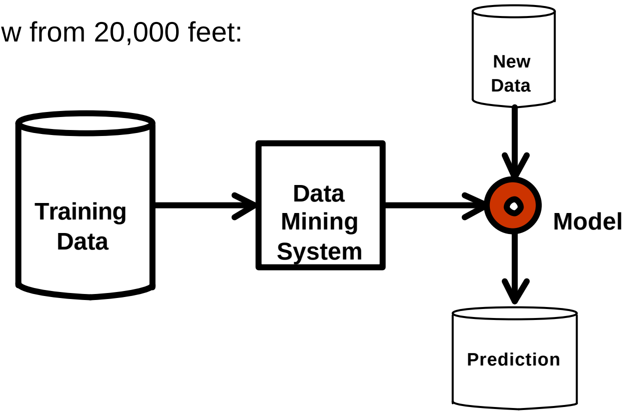
Sometimes the Data Tells You Something You Should Have Already Known

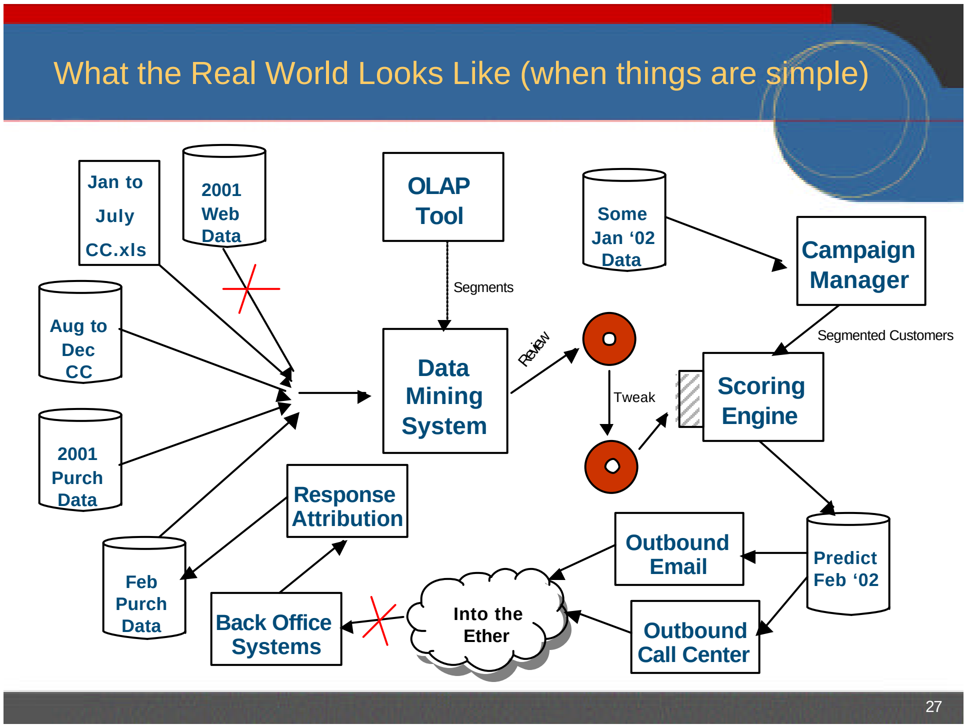


What is this?

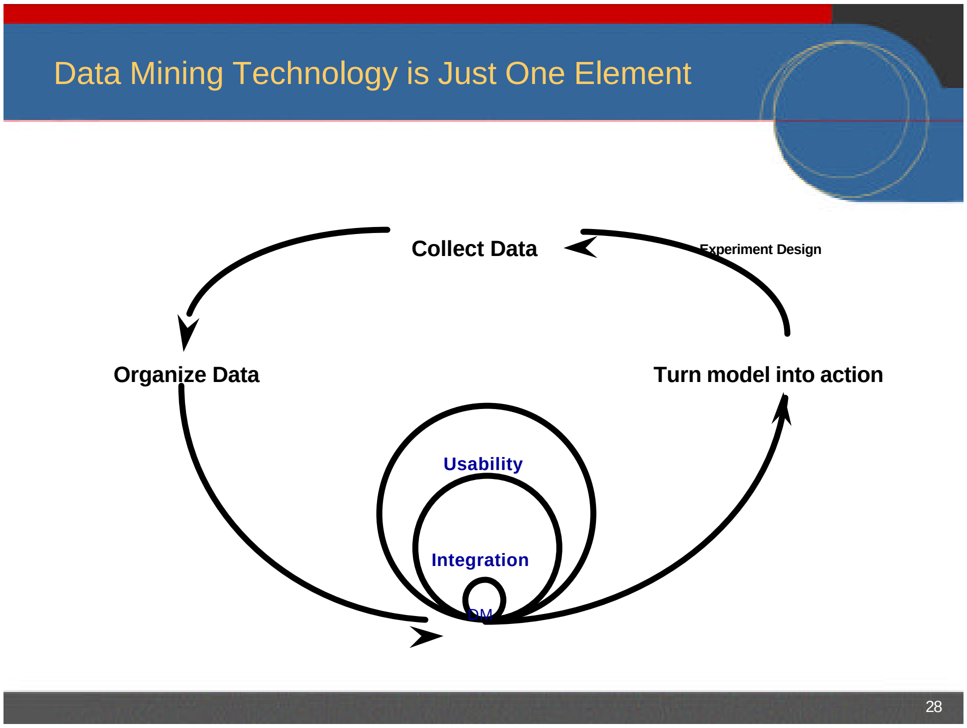
How are Predictive Models Built and Used?

— View from 20,000 feet:





What the Real World Looks Like (when things are simple)

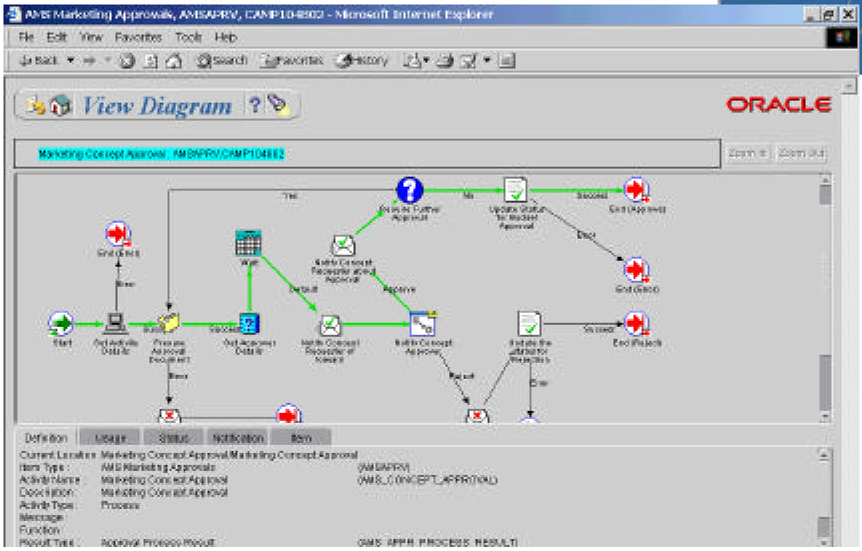


Data Mining Technology is Just One Element

Data Mining Fits into a Larger Process

- Easy in a ten person company, harder in a 50,000 person organization with offices around the world
- Run-of-the-mill office politics
  - Control of budget, personnel
  - Data ownership
  - Legal issues
- Application specific issues
  - Goals need to be identified
  - Data sources & segments need to be defined
- Workflow management is one option to deal with complexity
  - Compare this to newspaper publishing systems, or more recently, web content management
    - > Editorial & advertising process flow

Example: Workflow in Oracle 11i


AMS Marketing Approvals, AMSAPPRV, CAMP104802 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
View Diagram
ORACLE
Zoom In
Zoom Out
Start
Get Activity Details
Prepare Approval Document
Get Approver Details
Notify Concept Requestor of Timeout
Notify Concept Approver
Notify Concept Requestor about Approval
Wait
Is there Further Approval
Update Status for Marked Approval
End (Approved)
End (Error)
End (Rejected)
End (Error)
Update the status for Rejection
Success
Error
Default
Approve
Reject
Yes
No
Definition
Usage
Status
Notification
Item

Marketing Concept Approval : AMSAPPRV,CAMP104862

Current Location: Marketing Concept Approval/Marketing Concept Approval
Item Type : AMS Marketing Approvals (AMSAPPRV)
Activity Name : Marketing Concept Approval (AMS_CONCEPT_APPROVAL)
Description : Marketing Concept Approval
Activity Type : Process
Message :
Function :
Result Type : Approval Process Result (AMS APPR PROCESS RESULT)

What Caused this Complexity?

- Volume
  - Much more data
    - > More detailed data
    - > External data sources (e.g., GO Consortium, …)
  - Many more data segments
- Speed
  - Data flowing much faster (both in and out)
  - Errors can be easily introduced into the system
    - > “I thought a 1 represented patients who didn’t respond to treatment”
    - > “Are you sure it was table X23Jqqiud3843, not X23Jqguid3483?”
- Desire to include business inputs to the process
  - Financial constraints

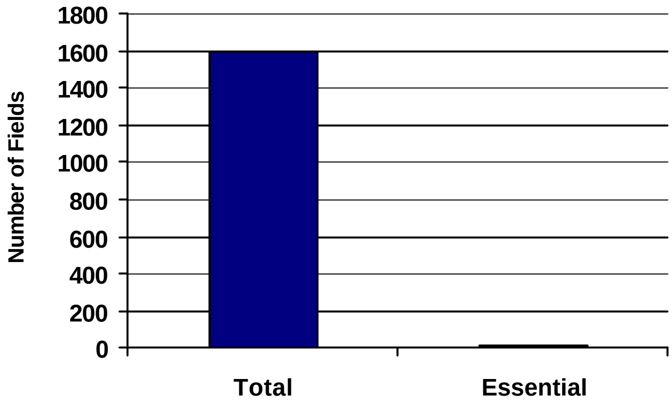
Legal and Ethical Issues

- Privacy Concerns
  - Becoming more important
  - Will impact the way that data can be used and analyzed
  - Ownership issues
  - European data laws will have implications on US
- Government regulation of particular industry segments
  - FDA rules on data integrity and traceability
- Often data included in a data warehouse cannot legally be used in decision making process
  - Race, Gender, Age
- Data contamination will be critical

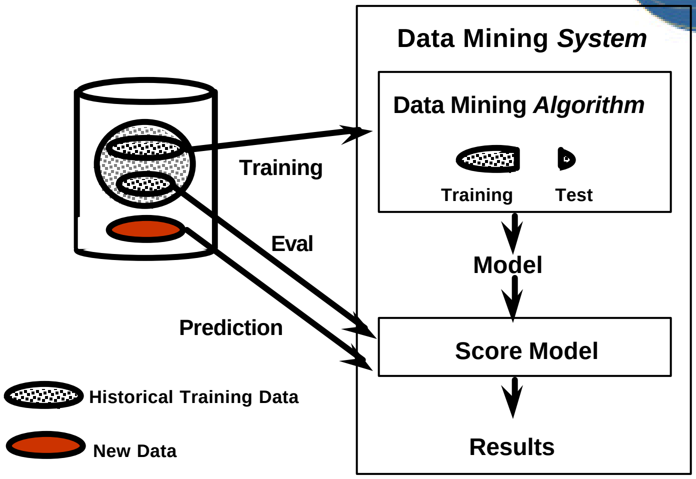
Data is the Foundation for Analytics

- If you don’t have good data, your analysis will suffer
  - Rich vs. Poor
  - Good vs. Bad (quality)
- Missing data
- Sampling
  - Random vs. stratified
- Data types
  - Binary vs. Categorical vs. Continuous
  - High cardinality categorical (e.g., zip codes)
- Transformations

Don’t Make Assumptions About the Data



The Data Mining Process



Data Mining *System*

Data Mining *Algorithm*

Training

Test

Model

Score Model

Results

Training

Eval

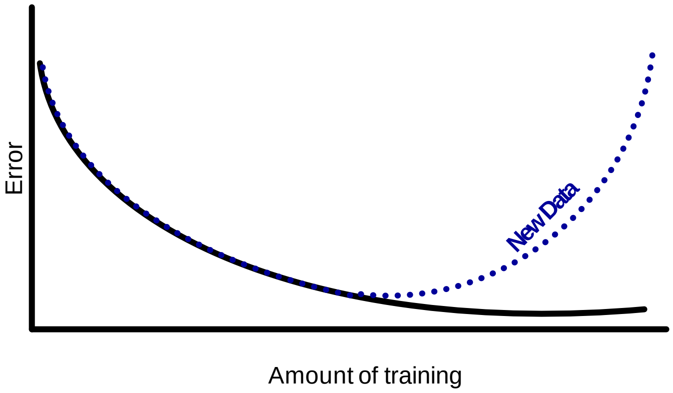
Prediction

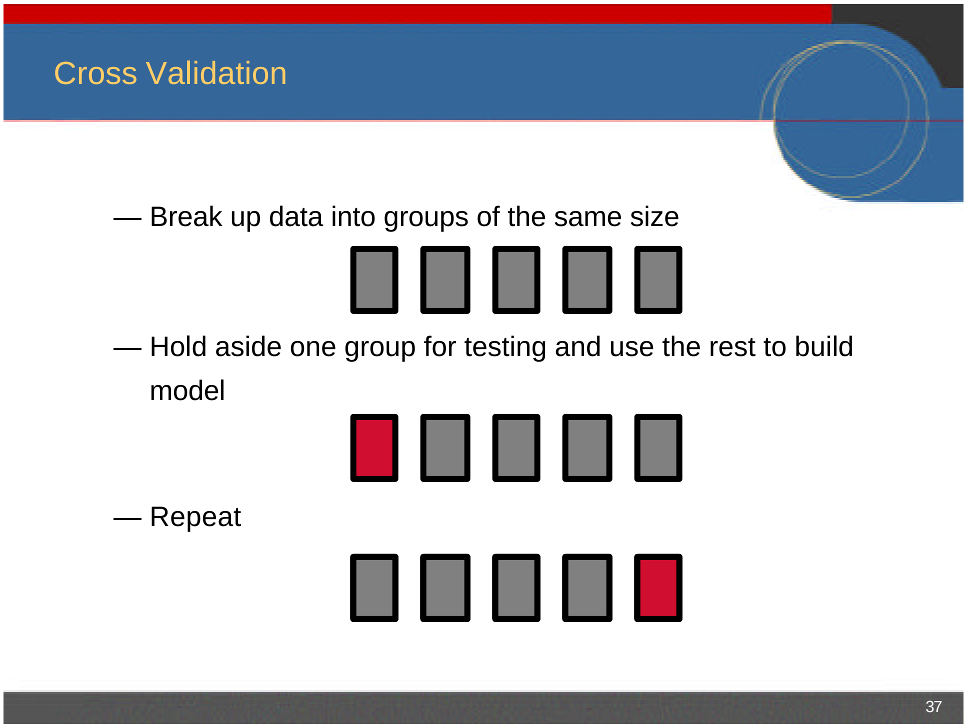
Historical Training Data

New Data

Generalization vs. Overfitting

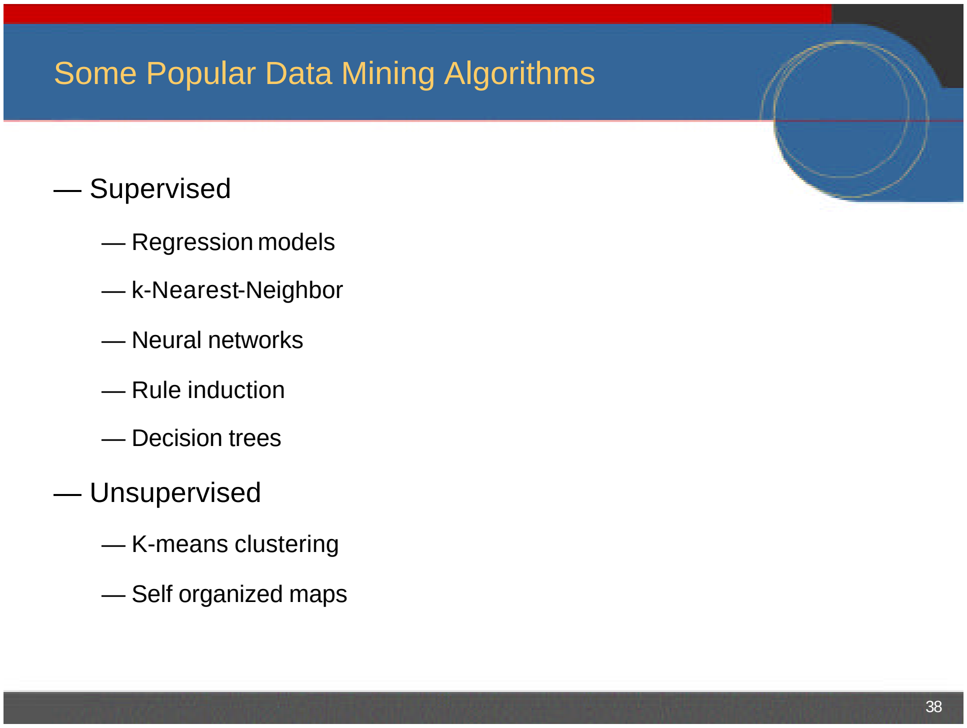
— Need to avoid overfitting (memorizing) the training data





Cross Validation

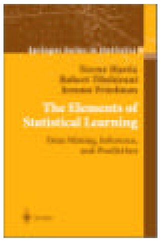
- Break up data into groups of the same size
- Hold aside one group for testing and use the rest to build model
- Repeat



Some Popular Data Mining Algorithms

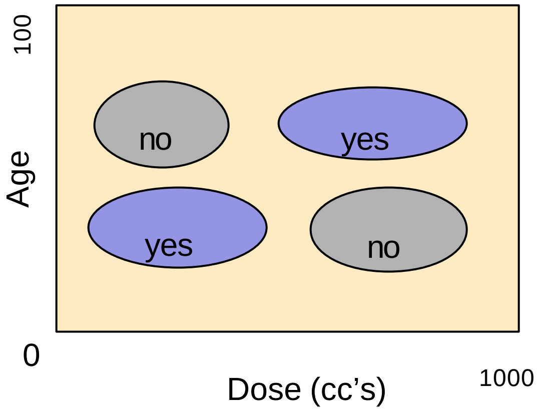
- Supervised
  - Regression models
  - k-Nearest-Neighbor
  - Neural networks
  - Rule induction
  - Decision trees
- Unsupervised
  - K-means clustering
  - Self organized maps

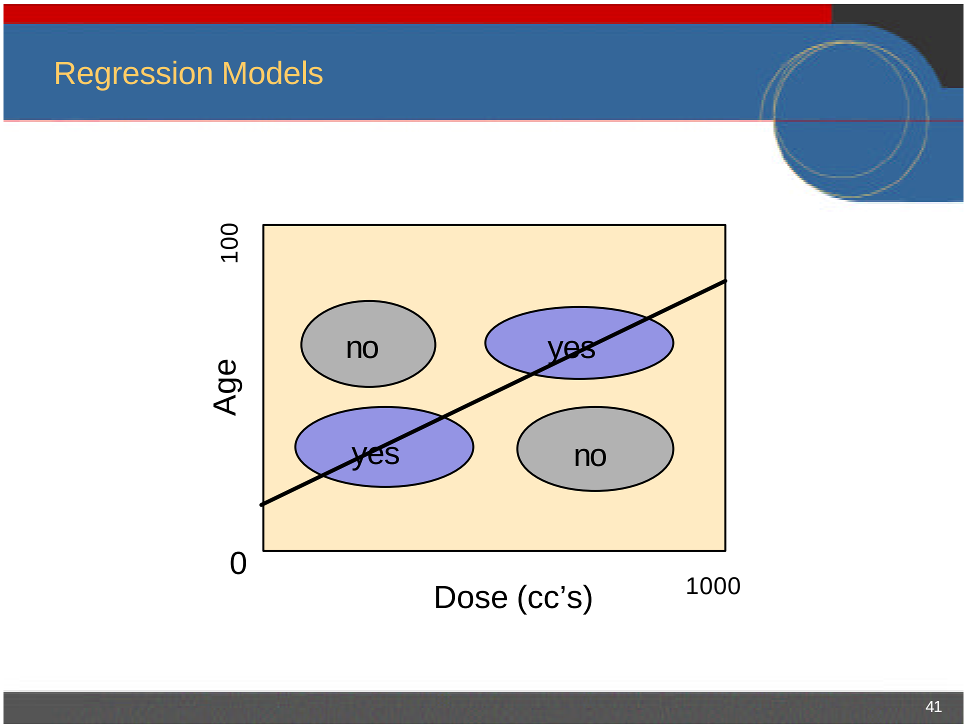
Two Good Data Mining Algorithm Books



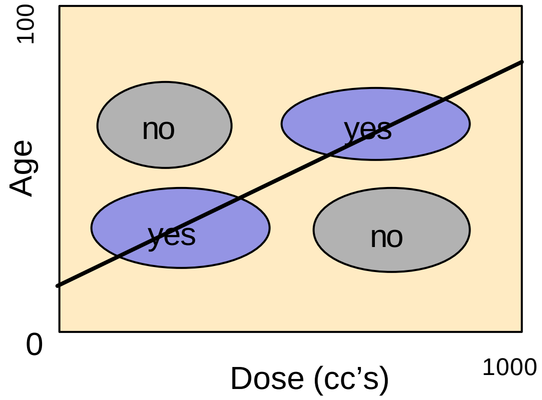
- *Intelligent Data Analysis: An Introduction* by Berthold and Hand
  - More algorithmic
- *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* by Hastie, Tibshirani, and Friedman
  - More statistical

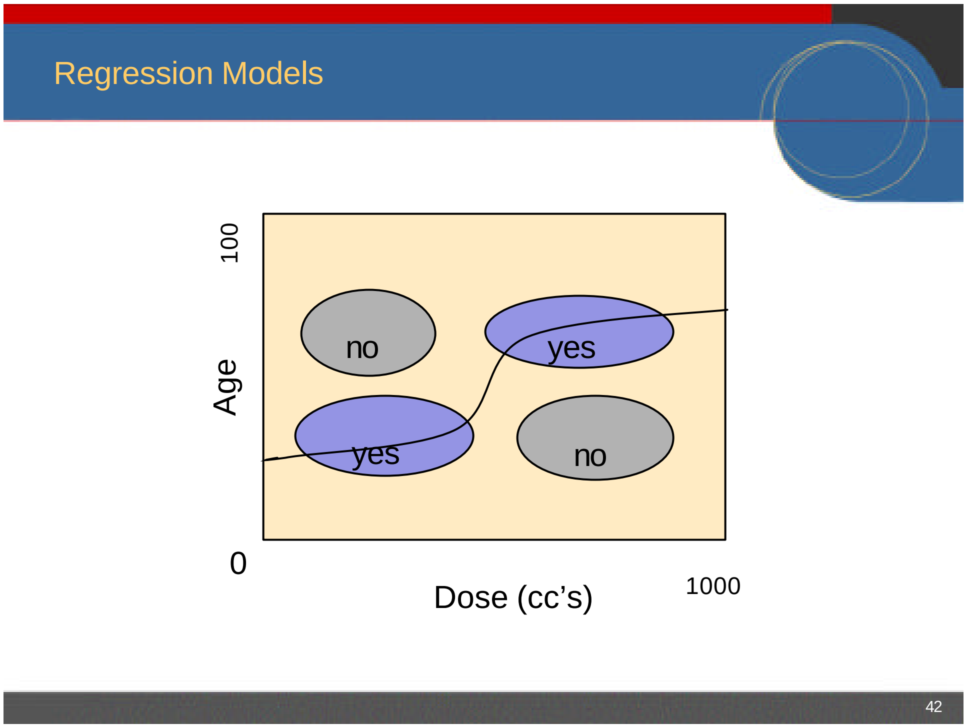
A Very Simple Problem Set



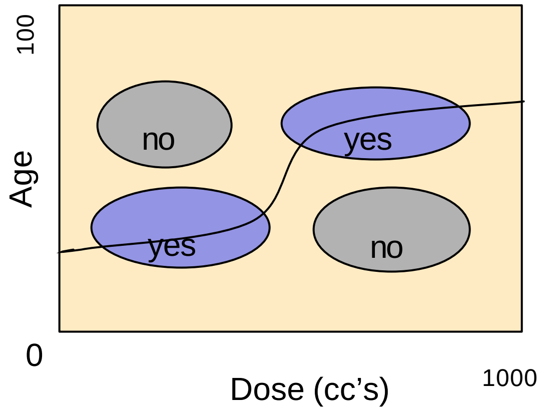


Regression Models





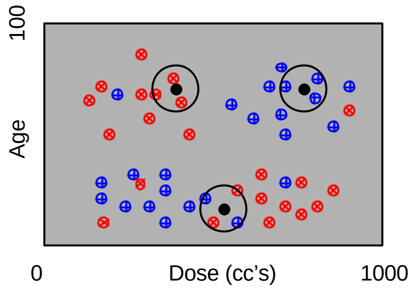
Regression Models



k-Nearest-Neighbor (kNN) Models

— Use entire training database as the model

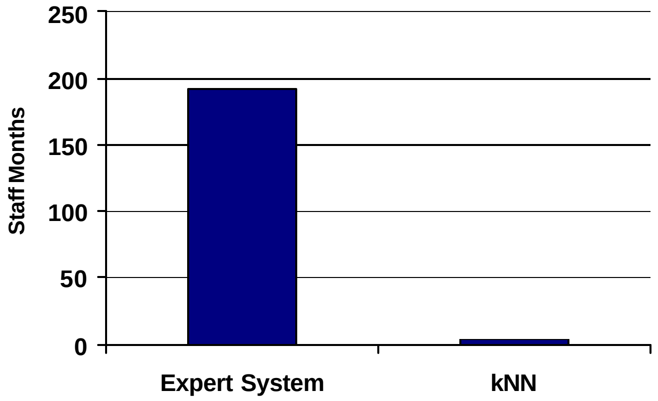
— Find nearest data point and do the same thing as you did for that record



— Very easy to implement. More difficult to use in production.

— Disadvantage: Huge Models

Time Savings with kNN



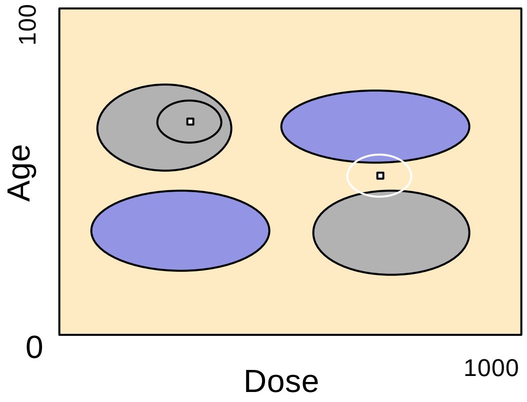
Developing a Nearest Neighbor Model

- Model generation:
  - What does “near” mean computationally?
  - Need to scale variables for effect
  - How is voting handled?
  - Confidence Function
- Conditional probabilities used to calculate weights
- Optimization of this process can be mechanized

Example of a Nearest Neighbor Model

- Weights:
  - Age: 1.0
  - Dose: 0.2
- Distance = $\sqrt{?\,\text{Age}^{2} + ???????\,\text{Dose}^{2}}$
- Voting: 3 out of 5 Nearest Neighbors (k = 5)
- Confidence = 1.0 - D(v) / D(v’)

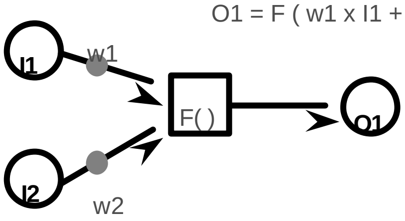
Example: Nearest Neighbor



(Feed Forward) Neural Networks

- Very loosely based on biology
- Inputs transformed via a network of simple processors
- Processor combines (weighted) inputs and produces an output value

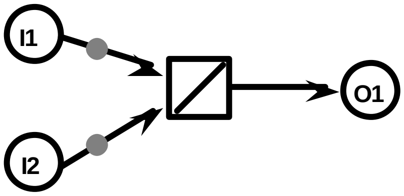
O1 = F (w1 x I1 + w2 x I2)



- Obvious questions: What transformation function do you use and how are the weights determined?

Processor Functionality Defines Network

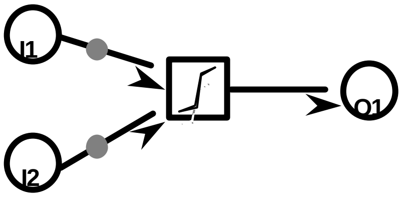
— Linear combination of inputs:



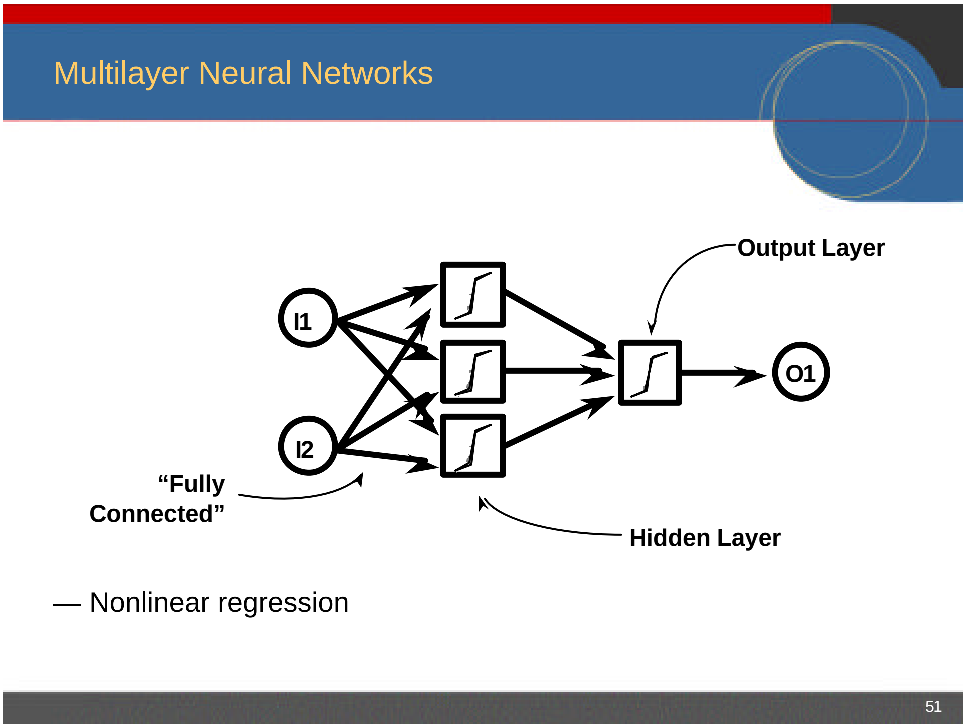
— Simple linear regression

Processor Functionality Defines Network (cont.)

— Logistic function of a linear combination of inputs

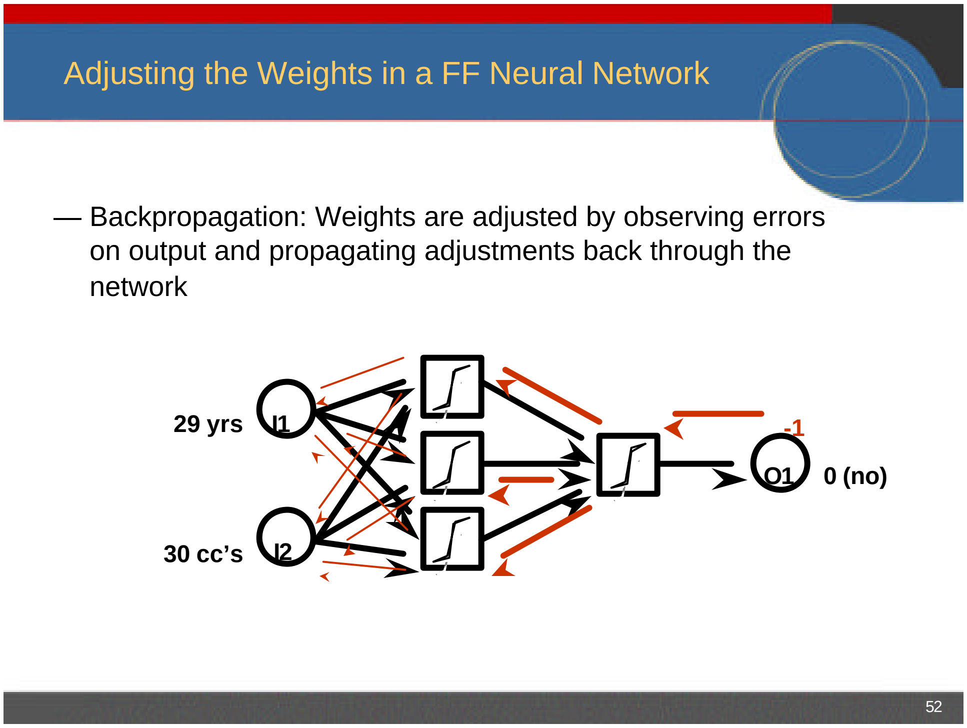


— Logistic regression
— Classic “perceptron”



Multilayer Neural Networks

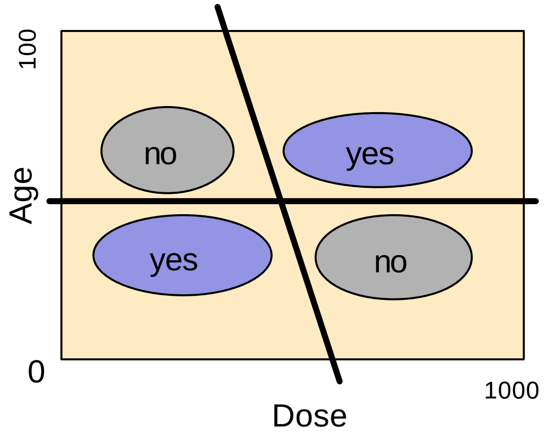
— Nonlinear regression



Adjusting the Weights in a FF Neural Network

— Backpropagation: Weights are adjusted by observing errors on output and propagating adjustments back through the network

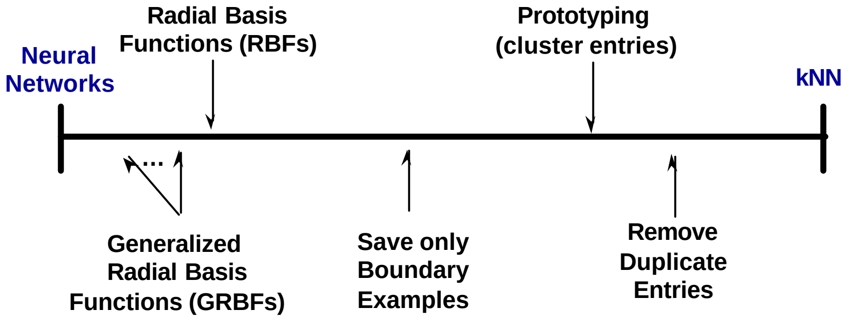
Neural Network Example



Neural Network Issues

- — Key problem: Difficult to understand
  - — The neural network model is difficult to understand
  - — Relationship between weights and variables is complicated
    - > Graphical interaction with input variables (sliders)
  - — No intuitive understanding of results
- — Training time
  - — Error decreases as a power of the training size
- — Significant pre-processing of data often required
- — Good FAQ: ftp.sas.com/pub/neural/FAQ.html

Comparing kNN and Neural Networks



Neural Networks

Radial Basis Functions (RBFs)

Prototyping (cluster entries)

kNN

…

Generalized Radial Basis Functions (GRBFs)

Save only Boundary Examples

Remove Duplicate Entries

Rule Induction

If Car = Ford and Age = 30…40
Then Defaults = Yes — **Weight = 3.7**

If Age = 25…35 and Prior_purchase = No
Then Defaults = No — **Weight = 1.2**

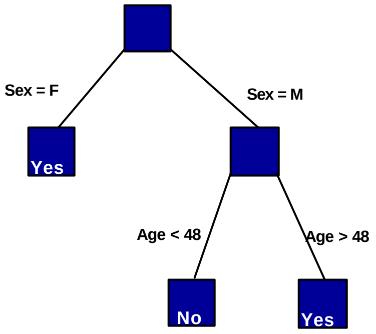
- Not necessarily exclusive (overlap)
- Start by considering single item rules
  - If A then B
    - A = Missed Payment, B = Defaults on Credit Card
  - Is observed probability of A & B combination greater than expected (assuming independence)?
    - If It is, rule describes a predictable pattern

Rule Induction (cont.)

- Look at all possible variable combinations
  - Compute probabilities of combinations
  - Expensive!
  - Look only at rules that predict relevant behavior
  - Limit calculations to those with sufficient support
- Move onto larger combinations of variables
  - n^3, n^4, n^5, ...
  - Support decreases dramatically, limiting calculations

Decision Trees

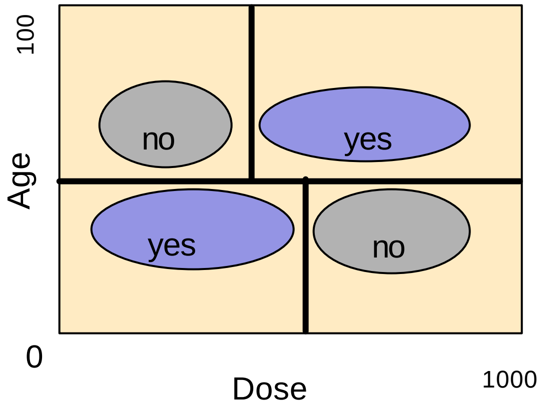
- A series of nested if/then rules.



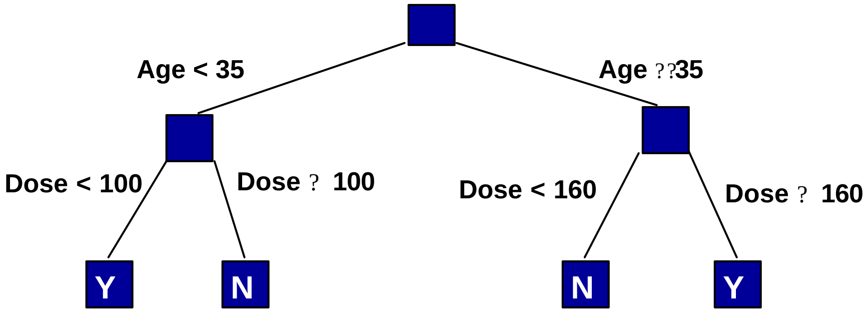
Types of Decision Trees

- CHAID: Chi-Square Automatic Interaction Detection
  - Kass (1980)
  - n-way splits
  - Categorical Variables
- CART: Classification and Regression Trees
  - Breimam, Friedman, Olshen, and Stone (1984)
  - Binary splits
  - Continuous Variables
- C4.5
  - Quinlan (1993)
  - Also used for rule induction

Decision Tree Model



One Benefit of Decision Trees: Understandability



Supervised Algorithm Summary

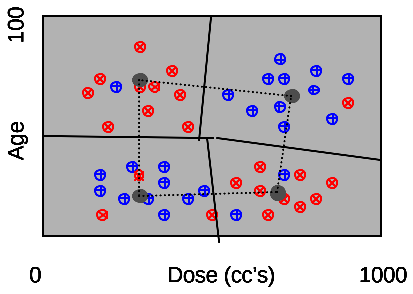
- kNN
  - Quick and easy
  - Models tend to be very large
- Neural Networks
  - Difficult to interpret
  - Can require significant amounts of time to train
- Rule Induction
  - Understandable
  - Need to limit calculations
- Decision Trees
  - Understandable
  - Relatively fast
  - Easy to translate into SQL queries

Other Supervised Data Mining Techniques

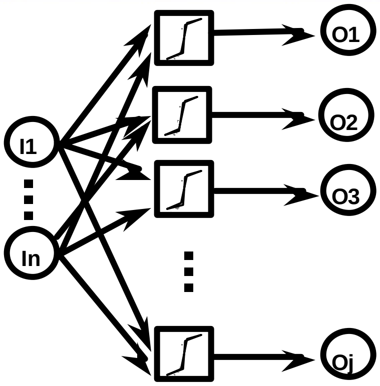
- — Support vector machines
- — Bayesian networks
  - — Naïve Bayes
- — Genetic algorithms
  - — More of a search technique than a data mining algorithm
- — Many more...

K-Means Clustering

- — User starts by specifying the number of clusters (K)
- — K datapoints are randomly selected
- — Repeat until no change:
  - — Hyperplanes separating K points are generated
  - — K Centroids of each cluster are computed



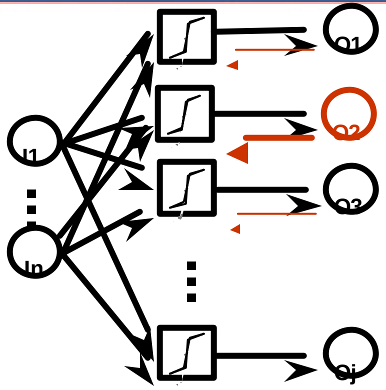
Self Organized Maps (SOM)



— Like a feed-forward neural network except that there is one output for every hidden layer node

— Outputs are typically laid out as a two dimensional grid (initial applications were in computer vision)

Self Organized Maps (SOM)



— Inputs are applied and the “winning” output node is identified

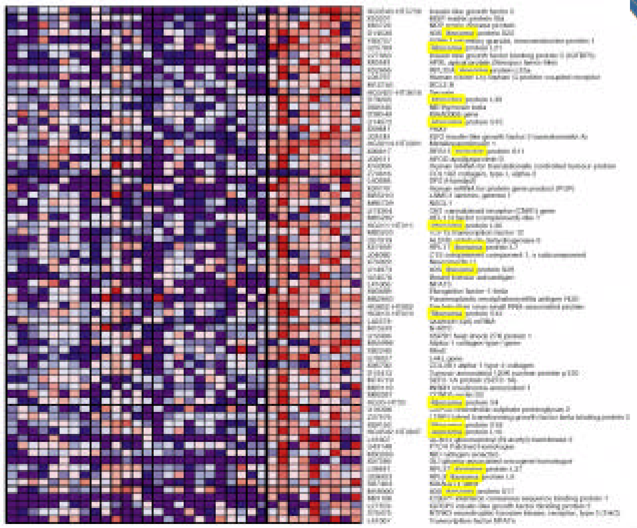
— Weights of winning node adjusted, along with weights of neighbors (based on “neighborliness” parameter)

— SOM usually identifies fewer clusters than output nodes

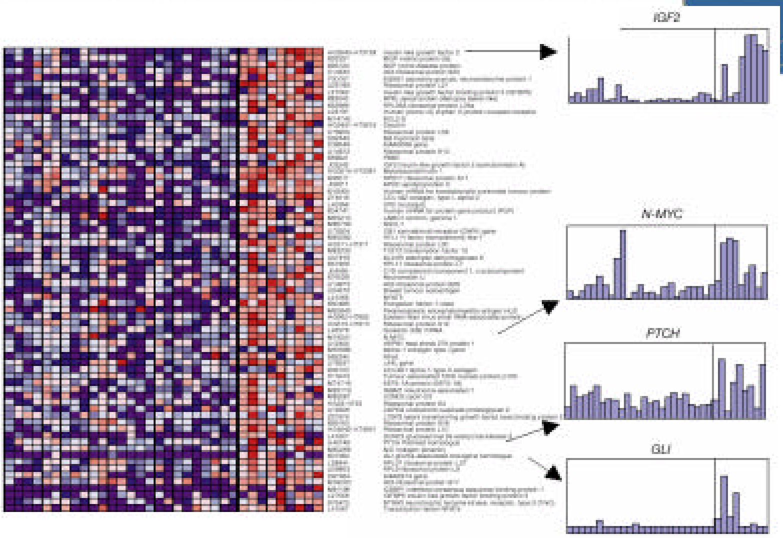
Text Mining

- — Unstructured data (free-form text) is a challenge for data mining techniques
- — Usual solution is to impose structure on the data and then process using standard techniques
  - — Simple heuristics (e.g., unusual words)
  - — Domain expertise
  - — Linguistic analysis
- — Example: Cymfony BrandManager
  - — Identify documents ✍ extract theme ✍ cluster
- — Presentation is critical

Text Can Be Combined with Structured Data



Text Can Be Combined with Structured Data

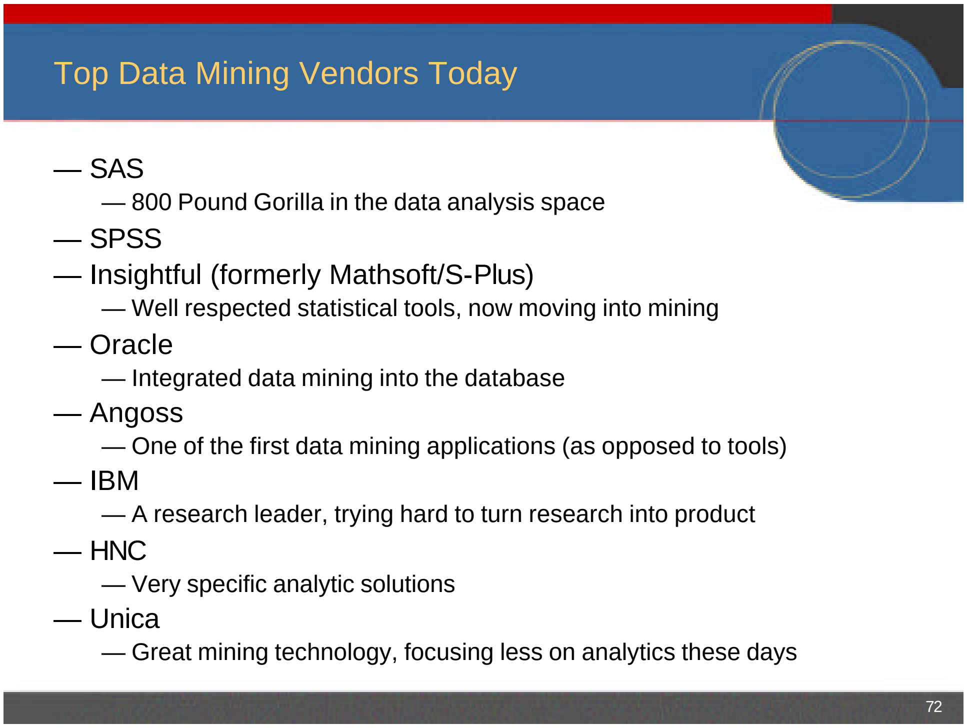


Commercial Data Mining Software

- It has come a long way in the past seven or eight years
- According to IDC, data mining market size of $540M in 2002, $1.5B in 2005
  - Depends on what you call “data mining”
- Less of a focus towards applications as initially thought
  - Instead, tool vendors slowly expanding capabilities
- Standardization
  - XML
    - > CWM, PMML, GEML, Clinical Trial Data Model, …
  - Web services?
- Integration
  - Between applications
  - Between database & application

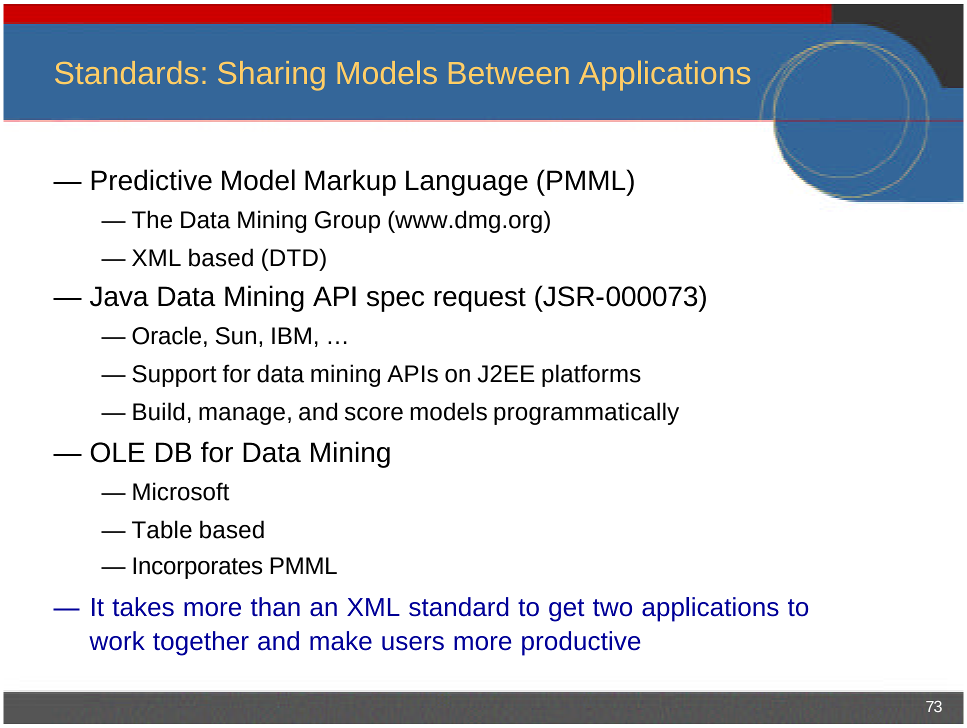
What is Currently Happening in the Marketplace?

- Consolidation
  - Analytic companies rounding out existing product lines
    - > SPSS buys ISL, NetGenesis
  - Analytic companies expanding beyond their niche
    - > SAS buys Intrinsic
  - Enterprise software vendors buying analytic software companies
    - > Oracle buys Thinking Machines
    - > NCR buys Ceres
- Niche players are having a difficult time
- A lot of consulting
- Limited amount of outsourcing
  - Digimine



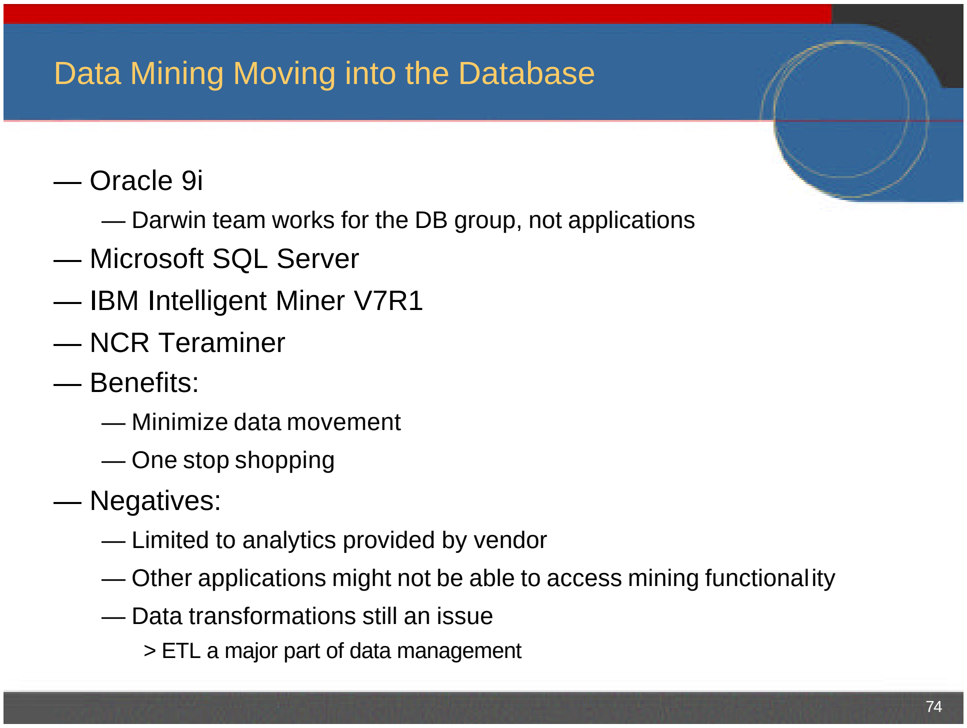
Top Data Mining Vendors Today

- SAS
  - 800 Pound Gorilla in the data analysis space
- SPSS
- Insightful (formerly Mathsoft/S-Plus)
  - Well respected statistical tools, now moving into mining
- Oracle
  - Integrated data mining into the database
- Angoss
  - One of the first data mining applications (as opposed to tools)
- IBM
  - A research leader, trying hard to turn research into product
- HNC
  - Very specific analytic solutions
- Unica
  - Great mining technology, focusing less on analytics these days



Standards: Sharing Models Between Applications

- Predictive Model Markup Language (PMML)
  - The Data Mining Group (www.dmg.org)
  - XML based (DTD)
- Java Data Mining API spec request (JSR-000073)
  - Oracle, Sun, IBM, …
  - Support for data mining APIs on J2EE platforms
  - Build, manage, and score models programmatically
- OLE DB for Data Mining
  - Microsoft
  - Table based
  - Incorporates PMML
- It takes more than an XML standard to get two applications to work together and make users more productive

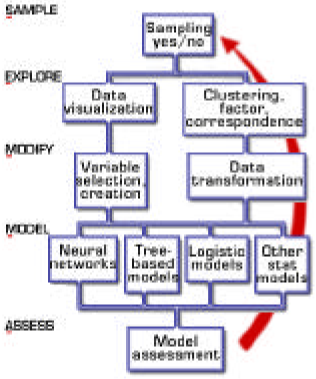


Data Mining Moving into the Database

- Oracle 9i
  - Darwin team works for the DB group, not applications
- Microsoft SQL Server
- IBM Intelligent Miner V7R1
- NCR Teraminer
- Benefits:
  - Minimize data movement
  - One stop shopping
- Negatives:
  - Limited to analytics provided by vendor
  - Other applications might not be able to access mining functionality
  - Data transformations still an issue
    - > ETL a major part of data management

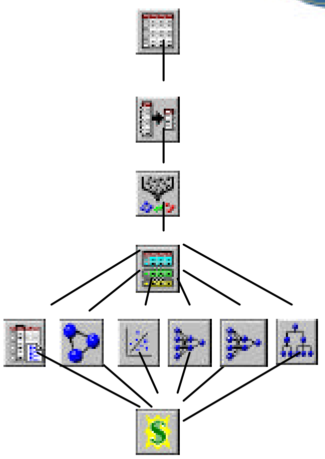
SAS Enterprise Miner

- Market Leader for analytical software
  - Large market share (70% of statistical software market)
    - >30,000 customers
    - >25 years of experience
- GUI support for the SEMMA process
  - Workflow management
- Full suite of data mining techniques

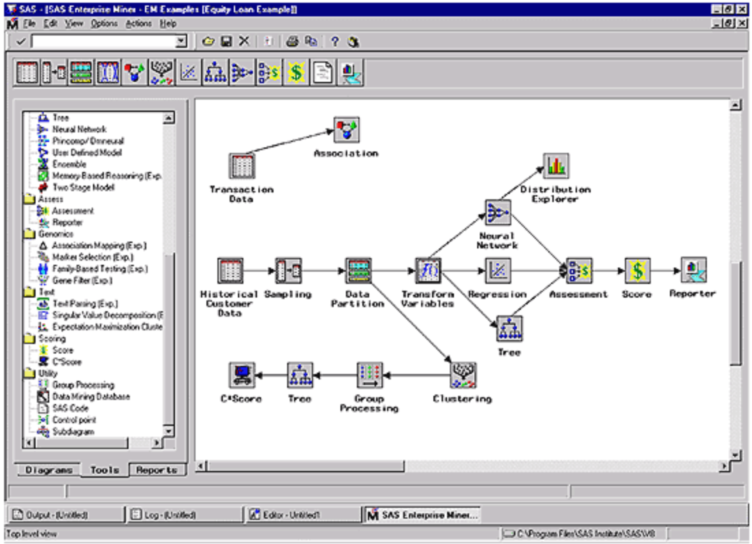


Enterprise Miner Capabilities

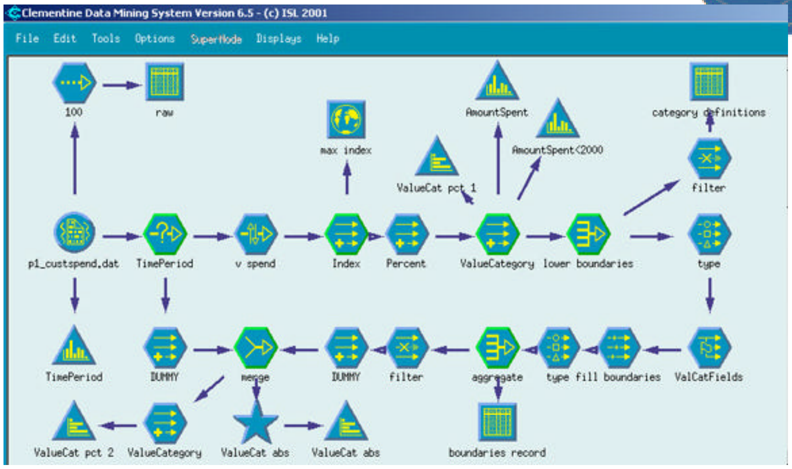
- Regression Models
- K Nearest Neighbor
- Neural Networks
- Decision Trees
- Self Organized Maps
- Text Mining
- Sampling
- Outlier Filtering
- Assessment



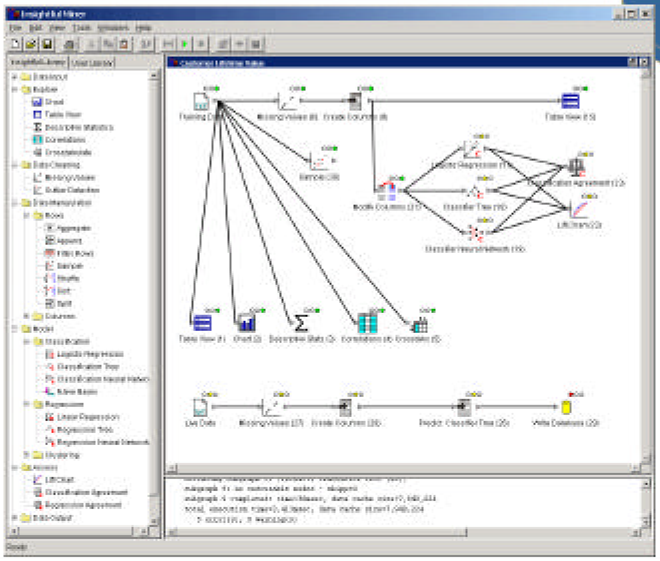
Enterprise Miner User Interface



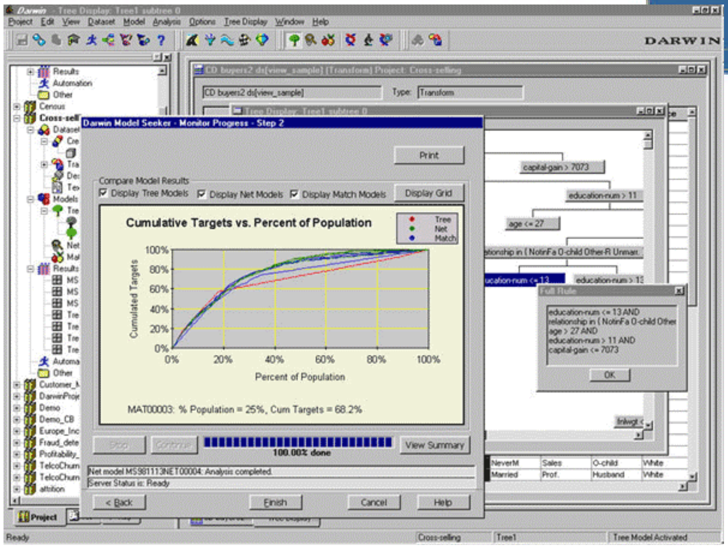
SPSS Clementine



Insightful Miner



Oracle Darwin



CD buyers2 ds[view_sample] Type: Transform

☑ Display Tree Models ☑ Display Net Models ☑ Display Match Models

MAT00003: % Population = 25%, Cum Targets = 68.2%

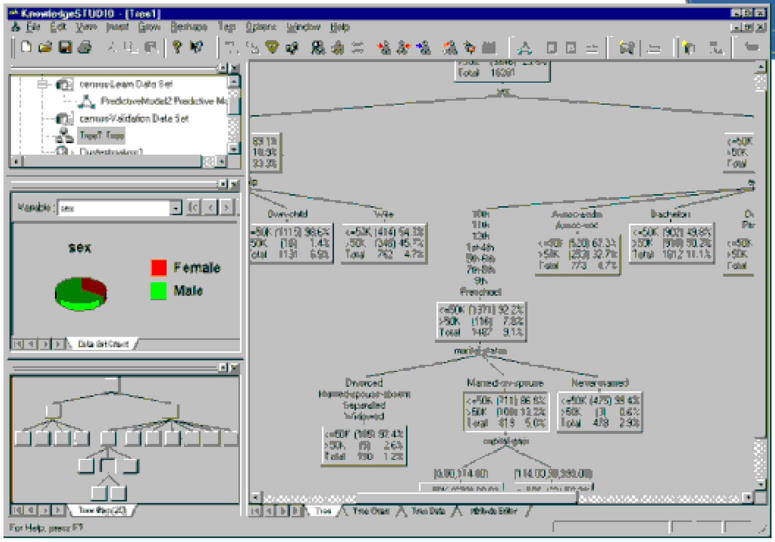
Net model MS981113NET00004: Analysis completed.

Server Status is: Ready

education-num <= 13 AND
relationship in (NotinFa 0-child Other
age > 27 AND
education-num > 11 AND
capital-gain <= 7073

NeverM	Sales	O-child	White
Married	Prof.	Husband	White

Angoss KnowledgeSTUDIO

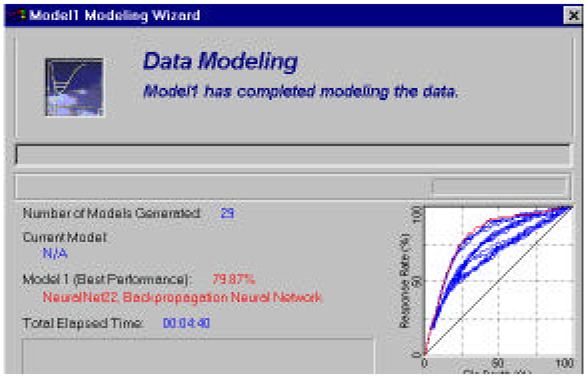


Usability and Understandability

- — Results of the data mining process are often difficult to understand
- — Graphically interact with data and results
  - — Let user ask questions (poke and prod)
  - — Let user move through the data
  - — Reveal the data at several levels of detail, from a broad overview to the fine structure
- — Build trust in the results

User Needs to Trust the Results

— Many models – which one is best?



Model1 Modeling Wizard

Data Modeling

Model1 has completed modeling the data.

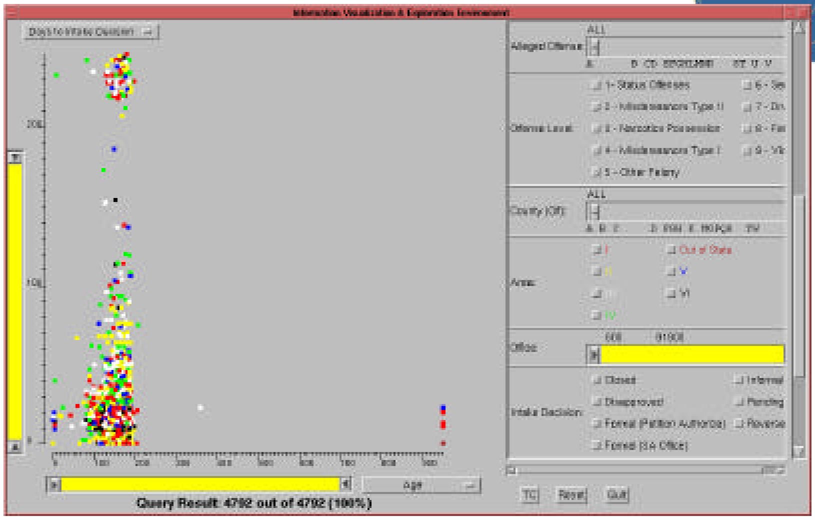
Number of Models Generated: 29

Current Model:
N/A

Model 1 (Best Performance): 79.87%
NeuralNet22, Backpropagation Neural Network

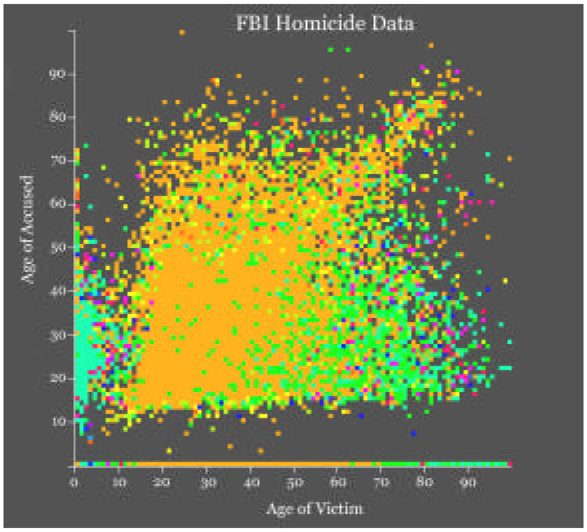
Total Elapsed Time: 00:04:40

Visualization Can Help Identify Data Problems



Query Result: 4792 out of 4792 (100%)

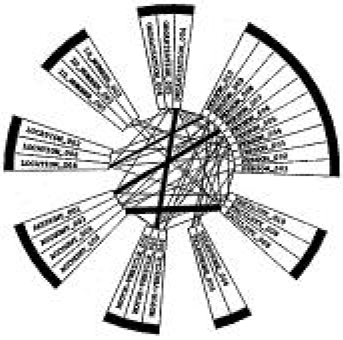
Visualization Can Provide Insight



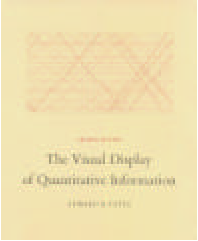
Visualization can Show Relationships

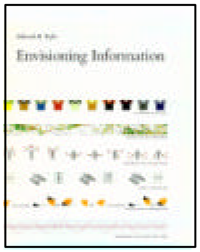
— NetMap

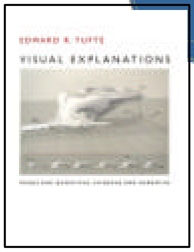
- — Correlations between items represented by links
- — Width of link indicated correlation weight
- — Originally used to fight organized crime



The Books of Edward Tufte



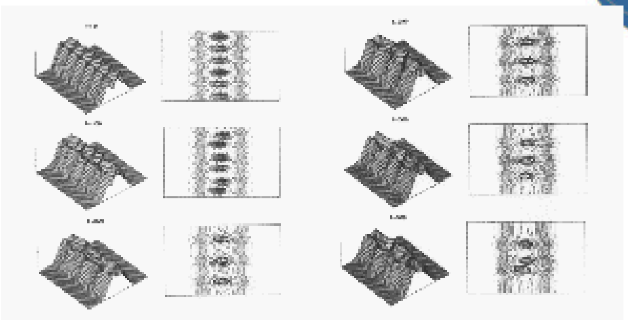




- — *The Visual Display of Quantitative Information* (1983)
- — *Envisioning Information* (1993)
- — *Visual Explanations* (1997)

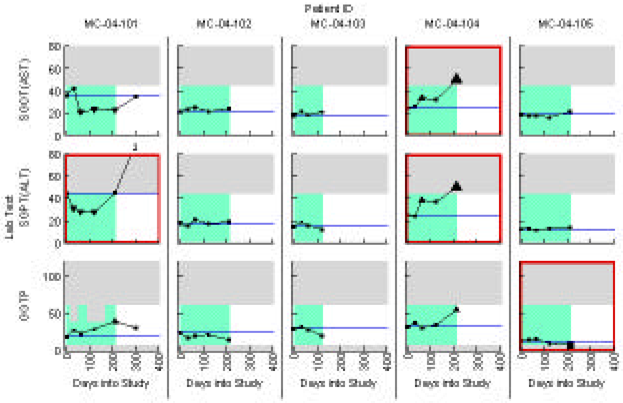
- — Basic idea: How do you accurately present information to a viewer so that they understand what you are trying to say?

Small Multiples

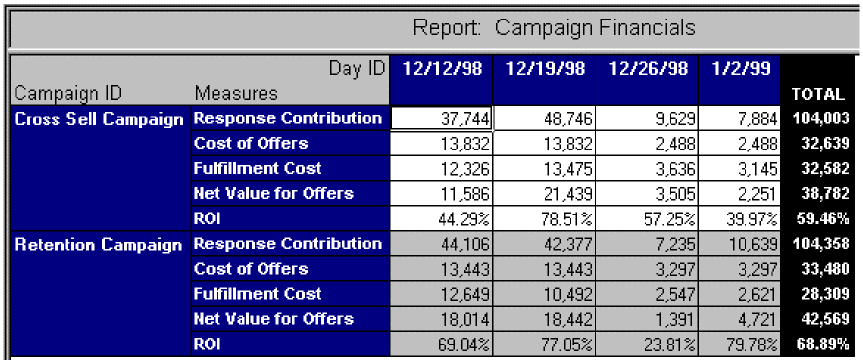


- — Coherently present a large amount of information in a small space
- — Encourage the eye to make comparisons

PPD Informatics: CrossGraphs



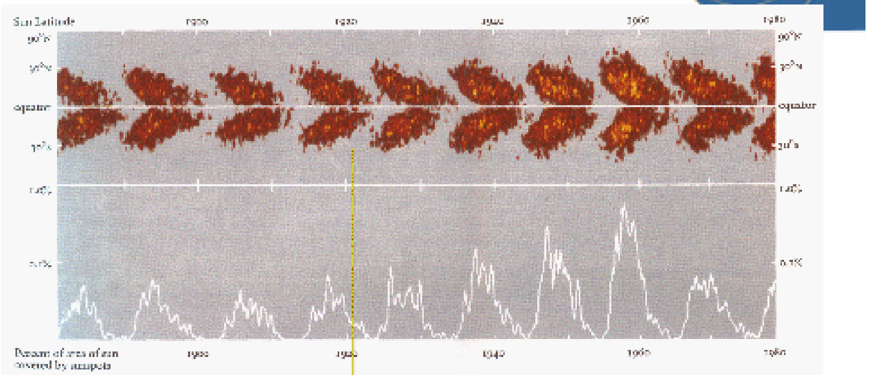
OLAP Analysis



Report: Campaign Financials

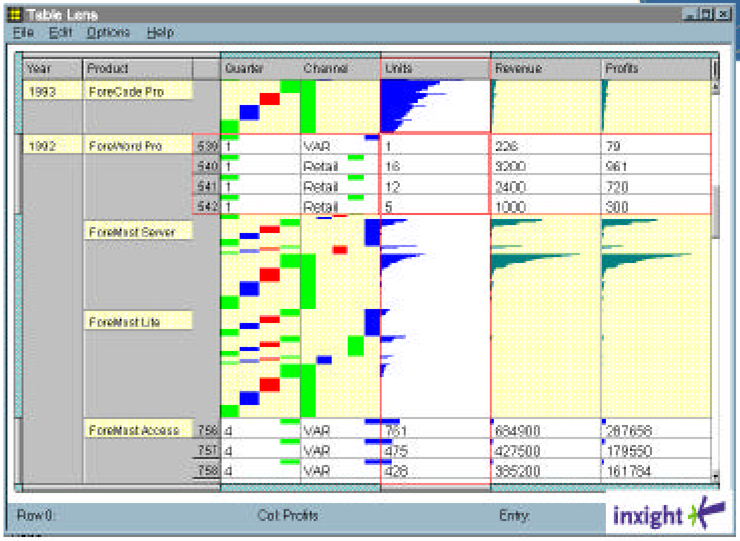
Campaign ID	Measures / Day ID	12/12/98	12/19/98	12/26/98	1/2/99	TOTAL
Cross Sell Campaign	Response Contribution	37,744	48,746	9,629	7,884	104,003
	Cost of Offers	13,832	13,832	2,488	2,488	32,639
	Fulfillment Cost	12,326	13,475	3,636	3,145	32,582
	Net Value for Offers	11,586	21,439	3,505	2,251	38,782
	ROI	44.29%	78.51%	57.25%	39.97%	59.46%
Retention Campaign	Response Contribution	44,106	42,377	7,235	10,639	104,358
	Cost of Offers	13,443	13,443	3,297	3,297	33,480
	Fulfillment Cost	12,649	10,492	2,547	2,621	28,309
	Net Value for Offers	18,014	18,442	1,391	4,721	42,569
	ROI	69.04%	77.05%	23.81%	79.78%	68.89%

Micro/Macro

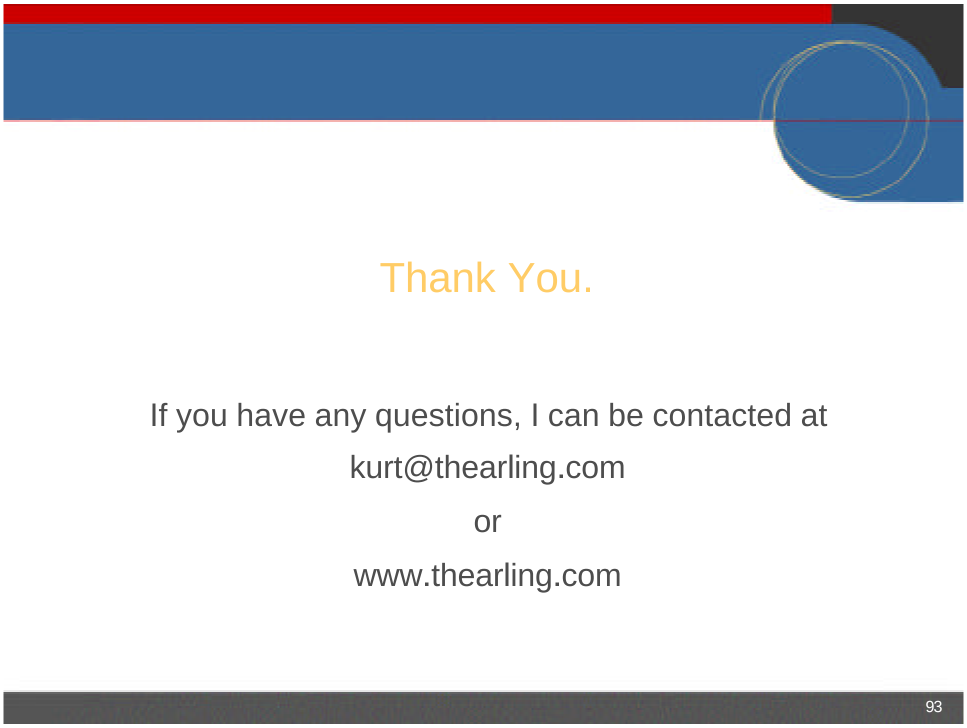


— Show multiple scales simultaneously

Inxight: Table Lens



Year	Product		Quarter	Channel	Units	Revenue	Profits
1993	ForeCode Pro						
1992	ForeWord Pro	539	1	VAR	1	226	79
		540	1	Retail	16	3200	961
		541	1	Retail	12	2400	720
		542	1	Retail	5	1000	300
	ForeMost Server						
	ForeMost Lite						
	ForeMost Access	756	4	VAR	761	684900	287658
		757	4	VAR	475	427500	179550
		758	4	VAR	428	385200	161784



Thank You.

If you have any questions, I can be contacted at
kurt@thearling.com
or
www.thearling.com