MONASH University

Information Technology

FIT5186 Intelligent Systems

Lecture 8

Data Mining and Knowledge Discovery

Learning Objectives

- Understand
  - the purposes, algorithms and methodologies of data mining
  - the directed and undirected knowledge discovery approaches
  - the statistical approaches to data mining and knowledge discovery
  - the principles of decision trees for knowledge discovery
  - the applications of data mining methodologies and their advantages and limitations via case studies

Lecture Outline

- What is data mining?
  - Methodology; Holistic Approach
- Data Mining Purposes and Algorithms
- Knowledge Discovery Approach
  - Directed and Undirected Knowledge Discovery
  - Typically, we use a combination of both directed and undirected knowledge discovery (a.k.a. supervised and unsupervised learning in NN).
- Statistical Approaches
  - Correlation, linear regression, moving average, multiple regression, factor analysis
- Decision Trees
- Case studies

Data Mining

- “Data mining is the efficient discovery of valuable, nonobvious information from a large collection of data” J. Bigus, 1996

- Information is often buried in the data.
  - Consumers may be switching brands and inventory will soon be out of balance (even though profits may be uniform).
  - A customer may have a series of unusual transactions (stolen credit card?)
  - Customers are leaving ... Why? What do they have in common?

Data Mining (continued)

- Business data is a valuable commodity.
  - Represents the current state of business
    - Profit is a lagged indicator of performance.
  - Can be used in the decision making process.
- The idea of extracting information from data collections is certainly not new.
  - Statistical techniques can be used to discover relationships and make predictions (using historic data for profit or stock forecasts).
  - Database tools (like SQL and Crystal Reports) facilitate complex queries to extract information.

Data Mining (continued)

- Traditional techniques are not always helpful for data mining.
  - Statistical techniques make many assumptions about the data.
  - Statistical results need careful interpretation.
    - e.g. a strong correlation factor does not necessarily indicate a causal relationship.
  - Query tools only help if you know what you are looking for.
- With data mining, we are trying to discover **new** (**hidden**) facts.

Data Mining (continued)

- Remember, data mining is all about ***efficiently*** discovering ***nonobvious*** relationships in large data sets.
  - We are not talking about complex queries or statistical tests to verify suspicions.
  - We are talking about the automated discovery of **new** relationships among the raw data.
  - The operation is **efficient** if the value of this new information exceeds the cost of retrieving it.

Data Mining (continued)

- Modern times have changed the scene:
  - Computerisation of business transactions and operations has led to a flood of business data.
  - Storage/processing technologies allow gigabytes and terabytes of data to be examined.
  - Advanced algorithms (such as neural networks) now exist for automated knowledge discovery.
- Combining these factors, there is a solution for businesses drowning in data that can result in increased knowledge and insight into the business (past, present and future).

Some Applications of Data Mining

- Marketing
  - Market segmentation using customer demographics and purchase patterns
    - Advertising and product development can then be undertaken with specific target groups in mind
  - Database marketing using market information with an existing database.
    - It helps to identify customers who are likely to switch to a competitor (target promotions).

Some Applications of Data Mining (continued)

- Retail
  - Market basket analysis
    - Mining point of sale transactions to discover associations between products.
    - Used to determine promotions and shop layout.
    - e.g. Magnum Opus Software (http://www.rulequest.com/) for pattern discovery (i.e. finding new and unanticipated patterns from data).

Some Applications of Data Mining (continued)

- Retail
  - Temporal spending patterns
    - Mining to detect the temporal relationship between products.
    - e.g. within 6 months most customers who purchase a printer need to buy a replacement cartridge (so encourage them to buy a spare as initial purchase, otherwise they might buy elsewhere).

Some Applications of Data Mining (continued)

- Finance
  - Fraudulent credit card transactions
    - Mine the data to discover unusual or suspicious transactions.
  - Credit risk assessment and bankruptcy prediction
    - Use data of existing customers to learn to classify new customers as good or bad credit risks.
  - Interest and exchange rate prediction.
  - Trading strategies.
  - These financial applications have been extremely successful.

Some Applications of Data Mining (continued)

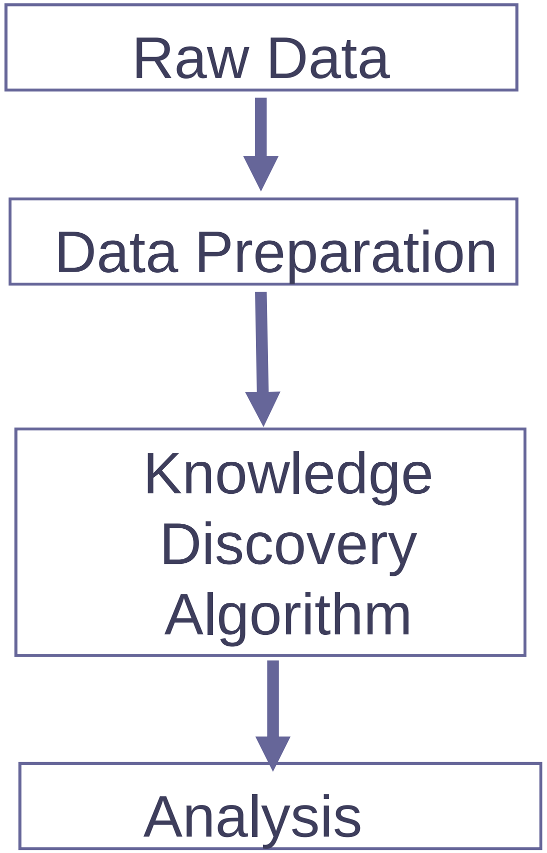
- Insurance
  - Claim modelling
  - Customer retention analysis
  - Fraud detection
- Business Intelligence or Business Analytics; Web mining
  - Market oriented names for data mining (knowledge extraction).
- And many more applications previously discussed regarding neural networks.
- Data mining is one of the most successful applications of neural networks.

Data Mining Methodology

- As a business process, data mining is characterised by 4 main steps:
  1. Identify the problem;
  2. Analyse the data;
  3. Take action;
  4. Measure the outcome.
- Steps 1 and 3 concern mainly with business issues, and Step 2 is the focus of data mining.
  – Link to the problem.
  – Link to insights that are actionable.

Analysing the Data (Stage 2)

Assuming the business problem and goals are identified, the data analysis process consists of 4 sub activities.



Raw Data

Data Preparation

Knowledge Discovery Algorithm

Analysis

Analysing the Data (Stage 2) (continued)

- Surveys show that anywhere from 50% to 80% of the resources spent on a data mining operation involve data preparation.
  - – Data preparation may involve one or all the following activities
    - Cleansing the data to remove errors and missing data;
    - Selection of the data relevant to the goals;
    - Pre-processing the data (e.g. combining variables or using summarised forms (averages);
    - Selection of data representation (data type) e.g. binary, continuous, or 1-out-of-N encoding;
    - Choosing the quantity of data to be used;
    - Consideration of noise in the data.
  - – Essential to eventual success.

Knowledge Discovery Algorithms

- A knowledge discovery algorithm extracts (discovers) meaningful (new) information (relevant to the business goals) from data.

- The type of knowledge discovery algorithm depends on the *purpose* of the data mining.

Purposes of Data Mining

- **Classification** – known classes, so it is a matter of learning how to class objects.
- **Clustering** – classes are not known prior to clustering.
- **Estimation** – deals with continuous value outcomes, e.g. estimating a family's income from other characteristics.
- **Associations** (affinity grouping) – determining what objects go together often (typically by counting how often it happens).
- **Forecasting (prediction**) – similar to classification and estimation, but the emphasis is on future events – it is not possible to check the results by confirming the classification or estimate by calling the family, for example.
- **Description and visualisation** – a picture can be worth a thousand words for data understanding.

Data Mining Purposes and Algorithms

Purpose	Algorithm	Application Examples
Classification	MFNN; Decision Trees; Discriminant Analysis	Loan Applicant Evaluation; Target Marketing
Prediction/ Forecasting	MFNN; Regression; Box-Jenkins (time series) ARMA (autoregressive moving average)	Sales Forecasting; Interest Rate (Exchange Rate or Stock Price) Prediction
Clustering	SOFM; K-Means	Market Segmentation; Location Analysis
Estimations	MFNN; Regression; Curve Fitting	Income Estimation
Associations	SOFM; Statistics; Rough Set Theory	Market Basket Analysis

Data Mining Result Analysis

- The information extracted from knowledge discovery algorithms need to be examined to determine how useful and meaning it is.
- Applying business knowledge to determine the information is actionable. If yes, take action and measure the outcome, and the process is complete.
- If the information does not result in actionable decisions, the knowledge discovery stage needs to be re-examined, probably using a different algorithm, changing the data mining purpose or revising the business goals.
- As a business process, data mining is an iterative process which may require several iterations through its steps before the desired outcome is achieved.

NN for Knowledge Discovery

- Neural networks are just one (very versatile) knowledge discovery technique.
  - Suitable for many of the purposes of data mining, especially classification, prediction and clustering.
  - A useful *modelling* technique for all of the data mining purposes considered.
  - Very good at fitting highly complex nonlinear relationships.
- There are other steps that should be followed for complete knowledge discovery.

Modelling in Knowledge Discovery

- Parametric modelling is the development of specific mathematical equations to characterise the relationship between variables (e.g. speed and gravity).
  - It requires a lot of knowledge about the problem.
  - This degree of understanding is frequently unavailable in real-world problems.
- Models that rely heavily on data, rather than domain-specific human expertise are called non-parametric of data-driven models. Multiple examples are provided from data sets so that the general patterns can be "learnt" by the non-parametric algorithm.
- The continuum is from *no data/much expert knowledge* (**parametric**) to *much data/little knowledge* (**non-parametric**). Many intermediate (hybrid) states exist between these two extremes.

Hypothesis Testing

- Often you will have feelings about relationships amongst the data.
  - You will use analytical methods to verify these feelings (e.g. hypotheses).
  - Data mining can also inform you of additional relationships you might not have realised existed.
    - e.g. males under 25 have more car accidents than any other age group.

Descriptive Statistics

- Before knowledge discovery, some descriptive statistics should be used to help become familiar with the data.
- What are the averages for each field (variable) across the entire data set?
  - What are the correlations between the fields?
  - What are the breakdowns of each field by significant characteristics
    - e.g. number of accidents per age group per geographic area for males and females.
    - Pivot tables in Excel can do this easily.

Descriptive Statistics (continued)

- Profiling the data in this way is useful for several reasons:
  - It helps you become familiar with the variables, their possible values, their meaning, etc.
  - It gives you an opportunity to formulate useful hypotheses.
  - It is likely to raise warning flags about inconsistencies in the data or definition problems which could make the results of data mining meaningless.

Summarisation

- What is the right level of detail for the data?
  - It depends on the analysis required
  - e.g. do we need to know times down to the minute, or is the hour or day sufficient?
  - Summarisation helps (“Do not delete – summarise”):
    - Too much detail may create files that are too big to handle.
    - There may be too few examples at the finest level of detail.

Knowledge Discovery Approach

- The approach (algorithm) can be directed or undirected.
- **Directed Knowledge Discovery** involves trying to explain the relationships between a particular field (variable) in terms of the others.
  - e.g. credit worthiness in terms of income, age, etc.
  - We know the answers for the example (training) data.
- In **Undirected Knowledge Discovery**, we ask the (computer) model (algorithm) to identify patterns in the data that may be significant without guidance.

Knowledge Discovery Approach (continued)

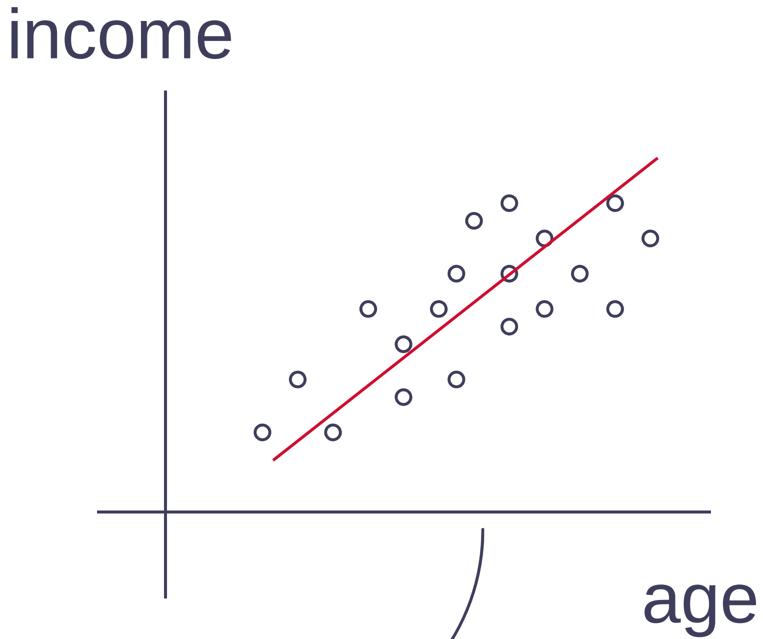
- Clustering is the most common outcome of undirected knowledge discovery.
  - Clustering takes all of the data and segments it into groups which have a lot in common.
  - We can then investigate these clusters and see what they have in common.
  - It helps target business decisions towards significant clusters of customers that may have been unnoticed in large data sets.
- Typically, we use a combination of both *directed and undirected knowledge discovery* (a.k.a. *supervised and unsupervised learning* in NN).

Knowledge Discovery Approach (continued)

- So knowledge discovery may involve:
  - Initial descriptive statistics;
  - Hypothesis testing (with SQL and statistics);
  - Directed knowledge discovery;
    - Using neural networks (MFNN) or decision trees, etc.
  - Undirected knowledge discovery;
    - Using clustering techniques such as SOFM or K-means.
  - Integration of this knowledge with existing business knowledge;
  - Possible generation of new hypotheses and models from the new knowledge.

Statistical Approaches

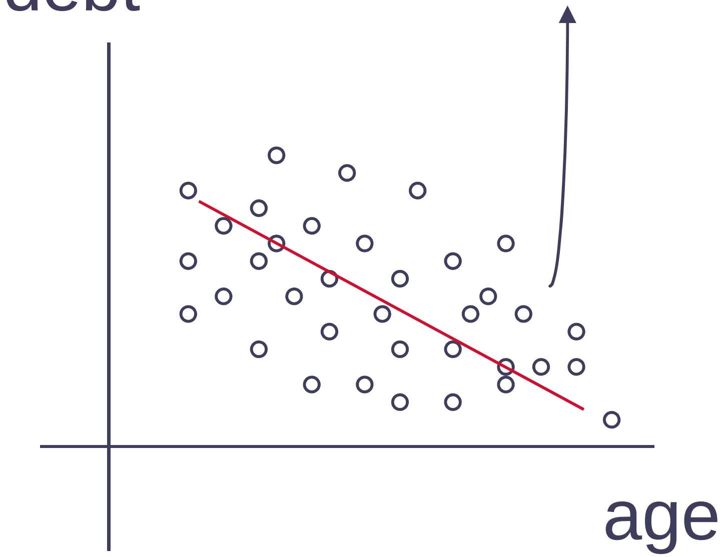
- Correlation
  - Measures the strength of the *linear* relationship between two data variables:
    - +1 means strong positive linear relationship.
    - -1 means strong negative linear relationship.
    - 0 means no linear relationship.
  - Useful for finding associations amongst data.
  - Attempts to fit a straight line through data to establish a linear relationship.



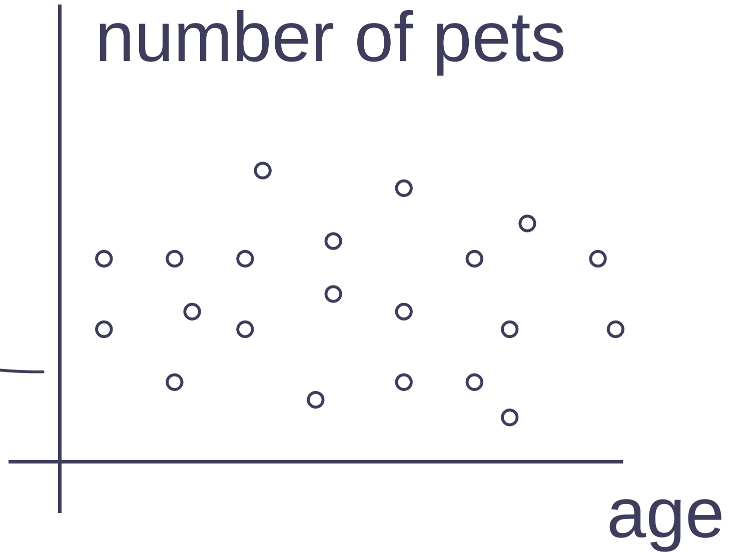
Strong positive correlation
(say *r* = 0.95)

Weak negative correlation

(say *r* = -0.4)



No correlation
(r = 0)



Statistical Approaches (continued)

- Example: use realty.xls to find correlations between house characteristics and price.
  - In Excel: use the CORREL() function.
- Does correlation imply a causal relationship?
  - No. Only that the variables *appear* to be related. (Correlation does not imply causation)
  - *Structural equation modeling* (SEM) is typically used to model causal relationships among (latent) variables.
  - As an extension of the general linear model (GLM) of which multiple regression is a part, SEM may be used as a more powerful alternative to multiple regression, path analysis, factor analysis, time series analysis, and analysis of covariance.
- Causes of misleading correlations:
  - Extreme values (outliers); non-linear relationships; luck.

Statistical Approaches (continued)

- Linear Regression
  - Determines the equation of the straight line which best fits the data.
  - Can be used for prediction and forecasting.
  - Example: use anz.xls to find linear regression (predicting today's closing price using today's opening price): Data, Data Analysis, Regression
  - OK, provided that the relationship is roughly linear.

Statistical Approaches (continued)

- Moving Averages
  - Considers the average of previous values rather than just one variable.
  - Used for time series prediction and forecasting.
  - Example: use moving average to improve prediction for anz.xls.
- Multiple Regression
  - Allows several input variables to be used to predict or forecast.

Statistical Approaches (continued)

- Factor Analysis
  - Also known as Principal Component Analysis.
  - It is used to uncover the latent structure (dimensions) of a set of variables.
  - As a preprocessing technique, it reduces a large number of variables to a smaller number of factors for modeling purposes, where the large number of variables precludes modeling all the measures individually.
    - e.g. combining income, mortgage payment and number of cars into a *factor* called (labelled) “affluence”*.*
  - It can be used to select a subset of variables from a larger set, based on which original variables have the highest correlations with the principal component factors.
  - It can also be used to identify clusters of cases and/or outliers.

Statistical Approaches: Limitations

- Not suited to many types of problems
  - Generally OK if the relationships are linear.
    - In real business situations, the relationships commonly are not linear.
  - Tend to make assumptions about the statistical distribution of the data that cannot be guaranteed for real data.
- Need to be interpreted with caution.
  - Correlation does not imply causality.
- Different statistical techniques are required for each data mining purpose.

Neural Nets: Data Mining Engines

- Clearly, neural networks are a very versatile tool for data mining applications.
- They have several further advantages over more traditional techniques:
  - Can be used as a “black box” approach;
  - Inherently non-linear approach;
  - Results in models which can be used for decision support;
  - Can be used for both directed and undirected knowledge discovery.

More Knowledge Discovery Techniques

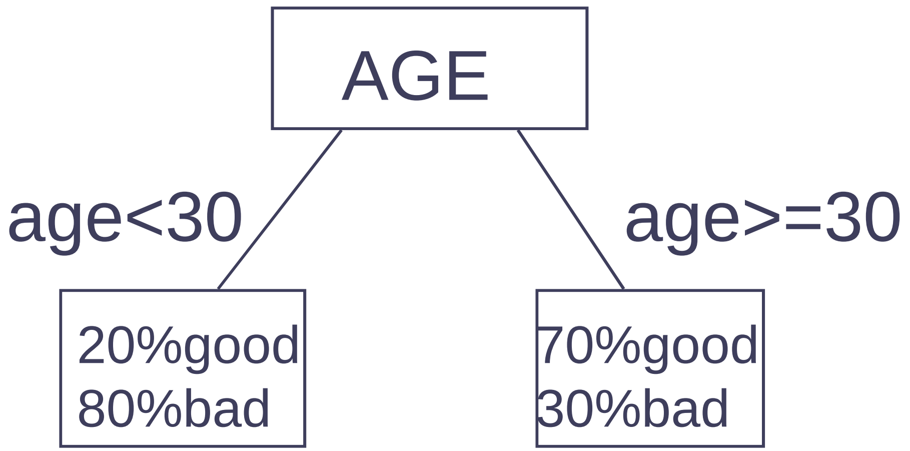
- Apart from neural networks and regression analysis, many commercial data mining packages provide:
  - Decision trees for directed knowledge discovery;
  - Genetic algorithms for directed knowledge discovery;
  - Link analysis for undirected knowledge discovery;
  - Data visualisation tools for undirected knowledge discovery;
  - And many more ...

Decision Trees

- Unlike neural networks, decision trees are used to discover rules (expressed in English) that separate the data into groups.
  - Need a target variable.
- At each step of the algorithm, a variable is chosen which is thought to most distinguish the known groups in the data.
  - Statistical tests performed to determine which variable is chosen.

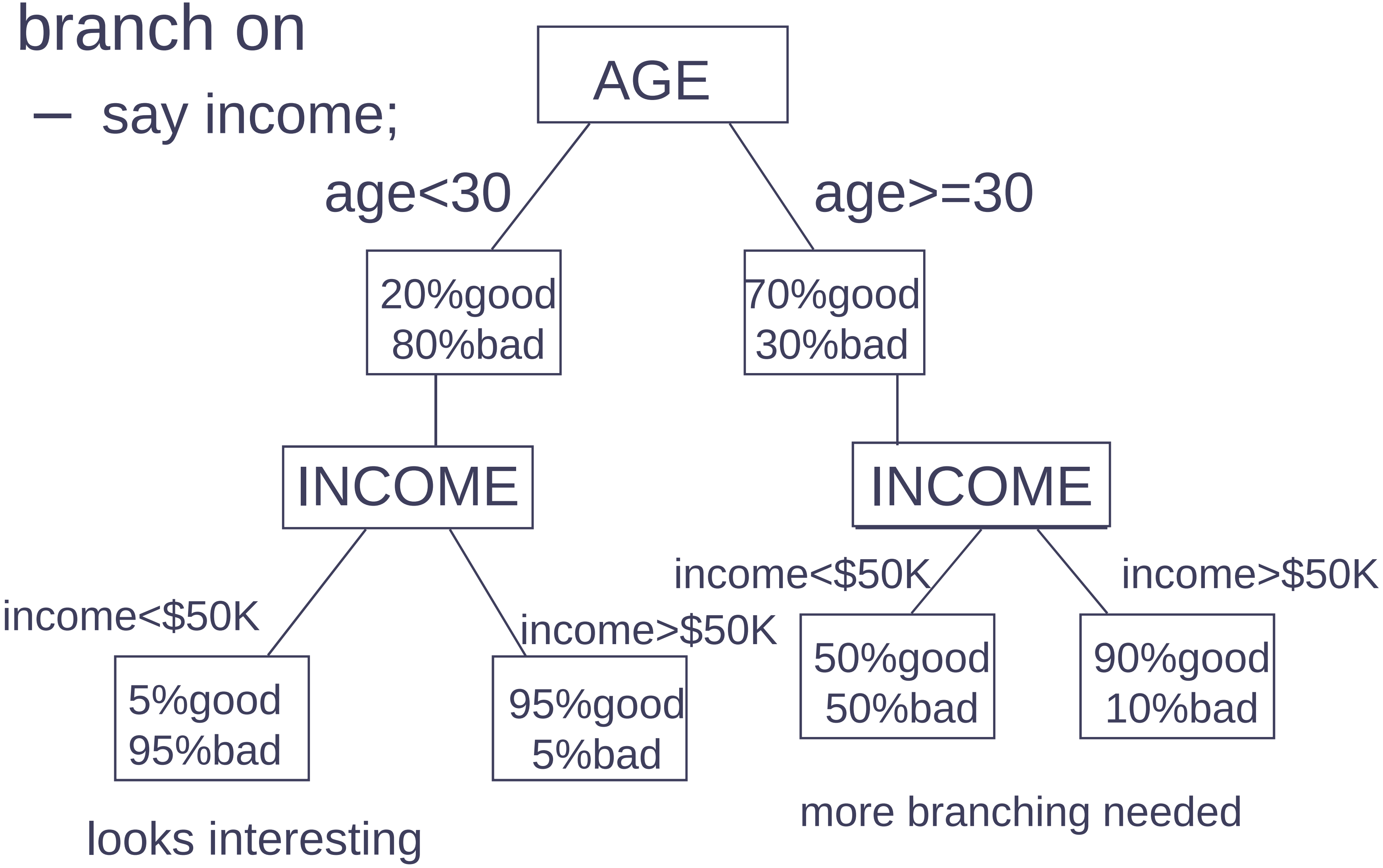
Decision Trees (continued)

- This variable becomes a tree node, and several branches are created.
  - e.g. suppose in the credit example, age is found to help distinguish good and bad credit risks (credit risk is target variable).
  - A branch might be created for people under 30, and another branch for people 30 or over.
  - Calculate statistics for each branch.



Decision Trees (continued)

- Another significant variable is then chosen to branch on
  - say income;



Decision Trees (continued)

- Procedure continues until no more statistically significant variables to branch on.
- If an interesting “pure” final node is found (i.e. almost completely one homogenous group), then the English rules defining the group can be found by reading the tree.
  - e.g. If (income < $50K) AND (age < 30) THEN probability of being bad credit risk = 0.95.
  - Of course we need to pay attention to the sample size in each node (1 person always gives a “pure” node!).

Directed Knowledge Discovery

- Identify sources of pre-classified data
- Prepare data for analysis
- Select a KD technique (e.g. neural networks - MFNN)
- Divide data into training and test sets
- Use training data to build a model
- Tune the model by applying test data
- Take action based on result
- Measure the effect of the actions taken
- Restart the data mining process to take advantage of new data generated by the actions taken ...

Undirected Knowledge Discovery

- Identify data source
- Prepare data for analysis
- Select an undirected KD algorithm (e.g. K-means or self-organising neural networks)
- Use the technique to discover hidden structures in the data (clusters)
- Inspect the clusters for patterns
- Identify potential targets for directed knowledge discovery
- Generate new hypotheses to test ...

The Holistic Approach to Data Mining Methodology

- Conduct descriptive statistical analysis
- Generate hypotheses
- Define data needed to test hypotheses
- Get data and prepare it for analysis
- Test hypotheses (typically with statistical tests, directed and undirected knowledge discovery algorithms)
- Evaluate results of queries and models
- Take action or generate new hypotheses
- Measure the effect of the action
- Restart the data mining process ...

A Case Study

- A bank wants to use data mining to improve its marketing of a loan product.
  - This case study illustrates how the two styles of data mining, knowledge discovery (directed and undirected) and hypothesis testing can be used together.
  - It generalises to many other significant business areas where data mining can help.

A Directed Knowledge Discovery Example

A Directed Knowledge Discovery (KD) Example

- The bank used directed KD (a neural network model) to learn to recognise likely prospects.
  - It used a training set of current home equity line holders.
  - It developed a neural network (MFNN) model for propensity to use a home equity line.
  - Each customer was given a score.
  - Scores allowed ranking of new customers according to their propensity, so only top 11% were mailed (in order to reduce costs).

A Undirected Knowledge Discovery Example

A Undirected Knowledge Discovery Example

- The customer database was segmented using an undirected clustering technique.
  - 12 clusters were produced.
  - One cluster was proved very interesting
    - 39% of the customers in this cluster had both business and personal accounts with the bank.
    - This cluster accounted for 27% of the 11% of customers who received high scores for propensity.
    - High correlation then between business accounts and home equity loans.

A New Hypothesis ...

- The bank hypothesised that a significant number of their home loan customers use their home equity to start a new business.
  - Hypothesis tested using statistics across database, and found valid.
  - New marketing campaign aimed at these same types of customers.

Another Case Study

- A bank was concerned about the rate of attrition of its cheque accounts.
  - They had already used a neural network model for predicting customer attrition.
  - This model did not provide enough warning to allow corrective action.
  - They needed to know what types of customers were they losing and why?

Another Undirected Knowledge Discovery Example

Another Undirected Knowledge Discovery Example

- The data can be segmented into similar clusters, then we can look for patterns within the clusters.
- We could segment according to fixed categories such as account type, branch, postcode, demographic profile.
- But it is sometimes better to let the data segment itself into clusters.

Another Undirected Knowledge Discovery Example

- All data is fed into a clustering algorithm
  - We do not tell the algorithm which customers are attriters or non-attriters.
- Once the data is clustered, we look inside each cluster.
  - Some clusters will have a high rate of attrition (higher than the average for the entire data set).
  - What else do the customers in such clusters have in common?
- One cluster had high attrition, but on close inspection most of the customers in this cluster had died!

  (Nothing to be done here).

Another Undirected Knowledge Discovery Example

- Another cluster had high attrition with many of the customers sharing the following characteristics:
  - Several accounts; mostly phone banking;
  - Call after hours; wait when they call.
- These customers are cheap to service, but they are leaving. What can be done?
- Business decisions are needed now.
  - Hire more call centre staff?
  - Provide these customers with another phone number (which is prioritised)?
  - Upgrade the call centre to adjust queue when they ring (just these customers)?

Another Undirected Knowledge Discovery Example

- The bank chose the second option.
  - These customers were notified of a special number.
  - Other customers were not, so the queues were kept short.
  - Other new customers fitting the description of the cluster will also be given the new number.
- The other clusters need to be analysed too of course.
- Directed KD can also be used within clusters to predict which customers are going to attrite, and the data is now refined.

Some Data Mining Packages

- IBM - Intelligent Miner
- SAS - SAS Enterprise Miner
- GhostMiner http://www.fqs.pl/business_intelligence/ghostminer
- Purple Insight (Silicon Graphics) - MineSet
- SPSS – Clementine (IBM)
- WEKA (free from http://www.cs.waikato.ac.nz/~ml/weka/)

- Most of these products use neural networks and other techniques (e.g. statistics, decision trees, fuzzy logic, genetic algorithms, etc.).

A Case Study – Car Insurance

- Based on a consulting work for AAMI (Australian Associated Motor Insurers, Ltd.).
- AAMI had invested in a data warehouse (SAS system).
- Data mining was the next obvious step, but was it worth the (large) expense?
- This case study investigates the potential of data mining in the insurance industry.

Business Understanding

- Insurance is a highly competitive industry.
- AAMI expected that data mining could give them a competitive advantage.
- Two operational objectives:

1. Market growth

- Growth in the number of customers achieved through marketing, mergers, and take-overs.
- But growth does not necessarily mean higher profitability (low premiums/high claims).

Business Understanding (continued)

2. Profitability

- Profitability based on: pricing of premiums and cost of claims.
- Pricing of premiums affects both customer acquisition (growth) and retention.
- If we could predict the average cost of claims, we could examine the impact on profitability.

Purposes of Data Mining (for this case study)

- To demonstrate the potential of data mining as a methodology for achieving market growth and profitability.
- If we can

  1) identify the effect of premium price on customer acquisition and retention, and
  2) predict the average claim costs and frequency of claims.

- Then we can predict the relationships between price, growth, and profitability.
  - Price can be set low enough to maximise customer acquisition and retention, but high enough to cover the cost of claims, so maximising growth and profit.

Data Mining Strategy

- Two separate problems:

1. **Growth related**

   To determine how the premium price affects acquisition and retention of customers

2. **Profit related**

   To identify which customers are likely to cost more in claims

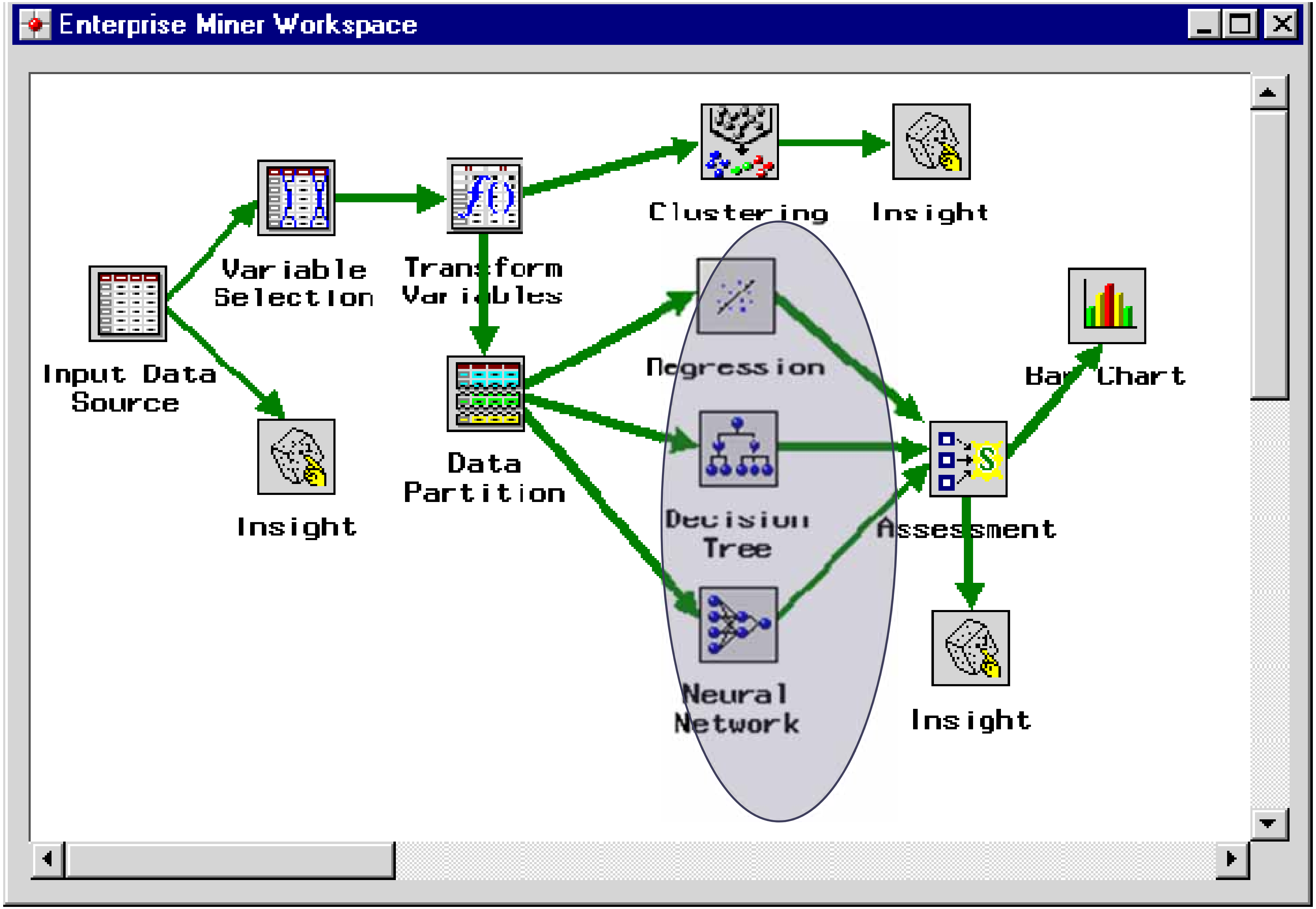
Problem 1 - Customer Retention

- First step of data mining done (identifying the problem), now for data analysis:
- Preliminary work needed to extract appropriate data from the data warehouse.
- Details on each policy holder:
  - Demographic (e.g. age group, postcode)
  - Policy information (e.g. premium, sum insured)
  - History (e.g. rating, years on rating, claim history)
  - Differences in premium and sum insured between the current policy and the renewal policy.
  - Which policies renewed and which terminated.

Data Understanding

- Management suggest 3 factors:
  - Pricing of premium;
  - Customer service, and
  - Insured value of car.
- Descriptive Statistics
  - Final data set contained 20,914 Victorian policy holders whose policies were due for renewal in a particular month, of which 7.1% terminated.
  - Amount of premium, insured value of car were significant between those who renewed and those who terminated their policies.
  - Customer service was not significant.
  - Age, time as a customer, etc. were not significant.
- Directed Knowledge Discovery
  - Use the premium and insured value to classify customers as likely to renew or termination their policies.
  - The model will identify customers who will renew better than identifying customers who will terminate, because it has had fewer examples to learn from.

Data Modelling



Results Assessment

- The data is divided into a training set and a test set.
- Three modelling techniques, including regression analysis, decision trees and neural networks, were used and compared.
- Neural networks gave the best result for classifying policy holders as likely to renew or terminate.

Results Analysis

- Classification accuracy with a 0.5 decision threshold
  - The customer is classified as terminated if the likelihood or probability generated by the NN model exceeds 0.5.

Total: 20,914		Classified as renewed	Classified as terminated	Row accuracy
19,427	Actually renewed	19,407	20	99.9%
1,487	Actually terminated	1,112	375	25.2%
	Column accuracy	94.6%	95%	

- Classification accuracy with a 0.1 decision threshold
  - The customer is classified as terminated if the likelihood or probability generated by the NN model exceeds 0.1.

Total: 20,914		Classified as renewed	Classified as terminated	Row accuracy
19,427	Actually renewed	17,174	2,253	88.4%
1,487	Actually terminated	731	756	50.8%
	Column accuracy	95.9%	25.1%	

Problem 2 - Claims Analysis

- Aims:
  - Understand where the growth has come from
  - Understand the impact of growth on profitability
  - Generate a tool for predicting the average claim costs of groups of policy holders
  - Develop a strategy for increasing profitability based on this analysis
- Management was not able to give insights into claims (e.g. cannot predict claims), so Undirected Knowledge Discovery was used rather than a classification model.

Key Performance Indicators (KPIs)

- Important to know KPIs to ensure data mining is relevant (i.e. the business understanding step of data mining).
- KPI 1: Cost ratio (CR)
  - Sum of claim costs / sum of earned premiums
  - Lower value is better (i.e. more earning from premiums relative to claims from those policies).
- KPI 2: Frequency ratio (FR)
  - Number of claims / number of policies
  - Low is good.

Analysis of Growth

- Data understanding step of data mining:
  - Temporal (time) study data from 2006, 2007, 2008.
  - 86% growth between 2006 and 2008.
  - Exceptional growth (200% vs. 86%) in (1) young people, (2) certain high risk areas of NSW, and (3) cars over $40K.
  - CR, FR fairly uniform across age groups, gender, risk area, etc. (existing variables used by company analysis)
  - The above implied premiums OK, but is there other hidden information in the data?

Clustering

- Modelling stage of data mining:
  - Self organise data into clusters to reveal hidden information.
- Calculate the cost ratio (CR) and frequency ratio (FR) for the new clusters.
- Many clusters found with CR between 0.7 and 0.9 (normal, so not new information).
  - Some with CR > 1.3 (premiums too low)
  - A few with CR < 0.4 (maybe the relatively high premiums will affect retention (policy renewal) of these customers)

Pricing the Premiums

- The clusters with lower and higher CR than average were interesting (and new!).
- Take action' stage of data mining

  1: Predict average claim costs by cluster.

  2: Classify customer retention.

  3: Adjust premium prices to retain customers but still make a profit.

Conclusion

- Successful demonstration of potential of the data mining approach to provide solutions to an industry.
- AAMI purchased SAS Enterprise Miner, hired a full time data miner to continue investigation.
- Many more issues to investigate including real time operation of the system.

Another Case Study:
The Emergency Department (ED)

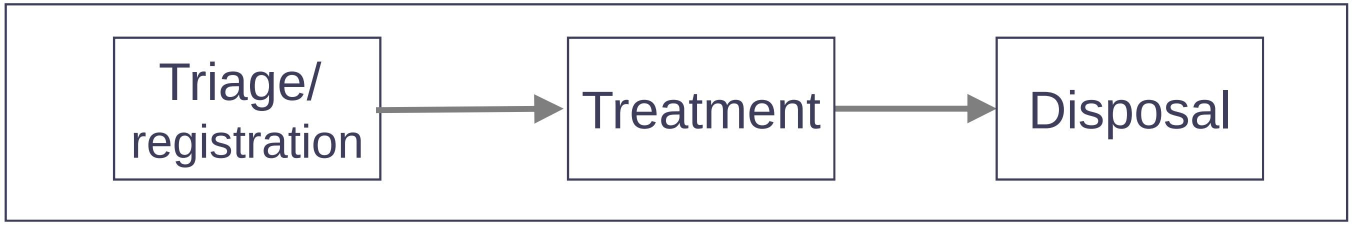
- Metropolitan Acute-care hospital
    - The main way patients get admitted to hospital.
    - High volumes of patients (every 8 minutes).
    - Much variety in presentations
        - Complete range of demographics, urgencies and severities at any given time
    - Shared resources (laboratories, x-ray, etc.)

ED Operations

- The government encourages “Save more lives but use less resources”
- Drive to efficiency “Do more with less”
- Cuts “resulted” in excessive patient wait and other politicised problems
- Many efforts to “fix” ED operations

ED Processes

- General ED process known



- Detailed formal models
  - Function view: what happens when (event process diagrams)
  - Organisation view: who does what
  - Data view: how patient data is used
- Existing workflow system

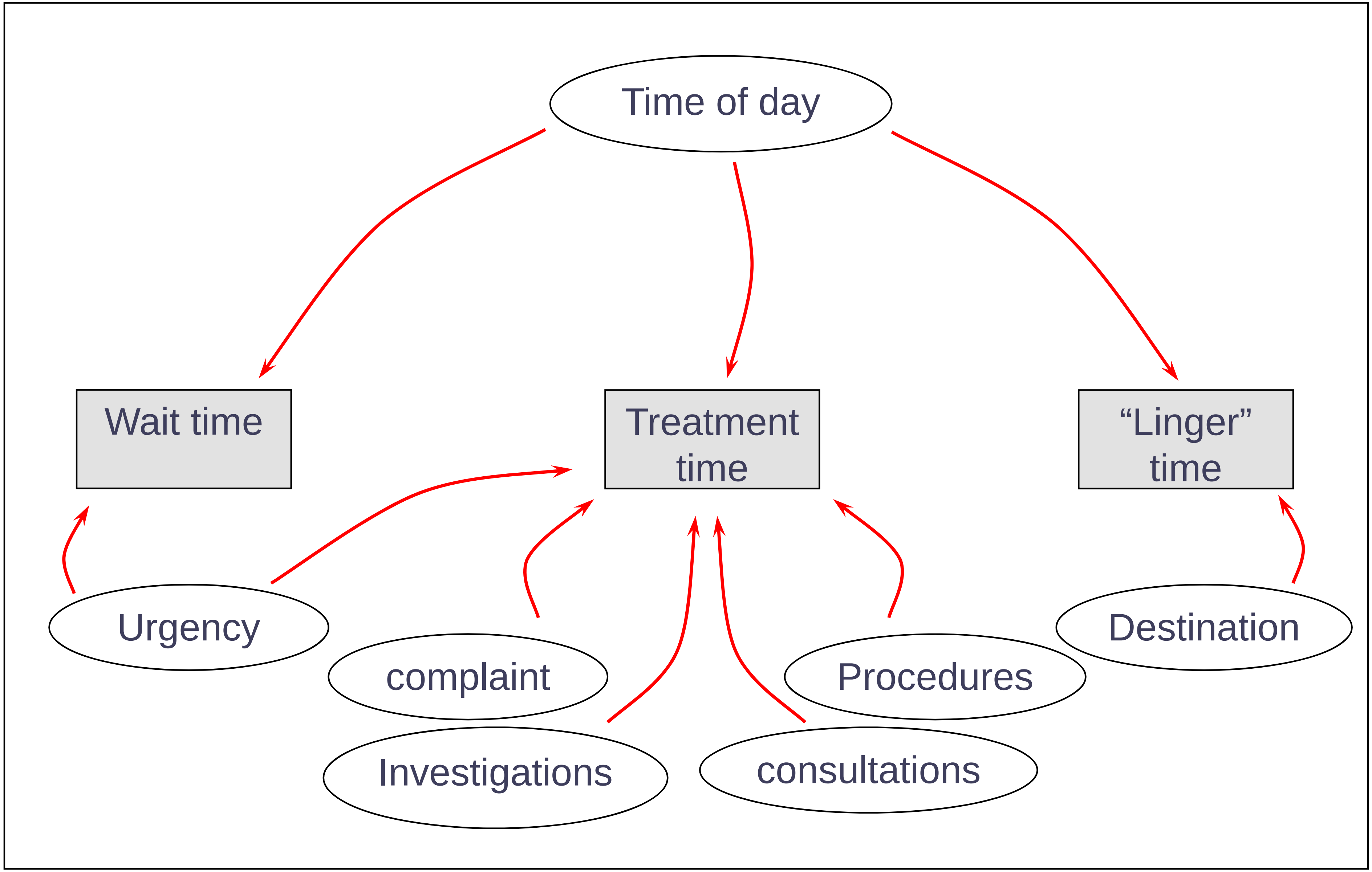
ED Casemix

- Clinicians group patients by “case”
  - Used for performance management and funding.
- ED groupings correlate ED **cost of treatment** to:
  - Urgency
  - Disposition (treated and discharged, admitted to hospital, etc.)
  - Age
- Using multivariate regression, but inaccurate.
- Can the data tell us more about what happens in the ED?

ED Data

- Much data captured
  - Demographics: Age, sex
  - Symptoms, diagnosis
  - Urgency
  - Key time points
  - Medical procedures performed
- Many errors in data (e.g. text age, not numerical) – cleanup and data preparation took months!

Statistical Analyses



Data Modelling

- Put usual medical data in a black box and shake it all up.
- Box = Descriptive statistics, regression, decision trees, supervised Neural networks, Self Organising Maps.

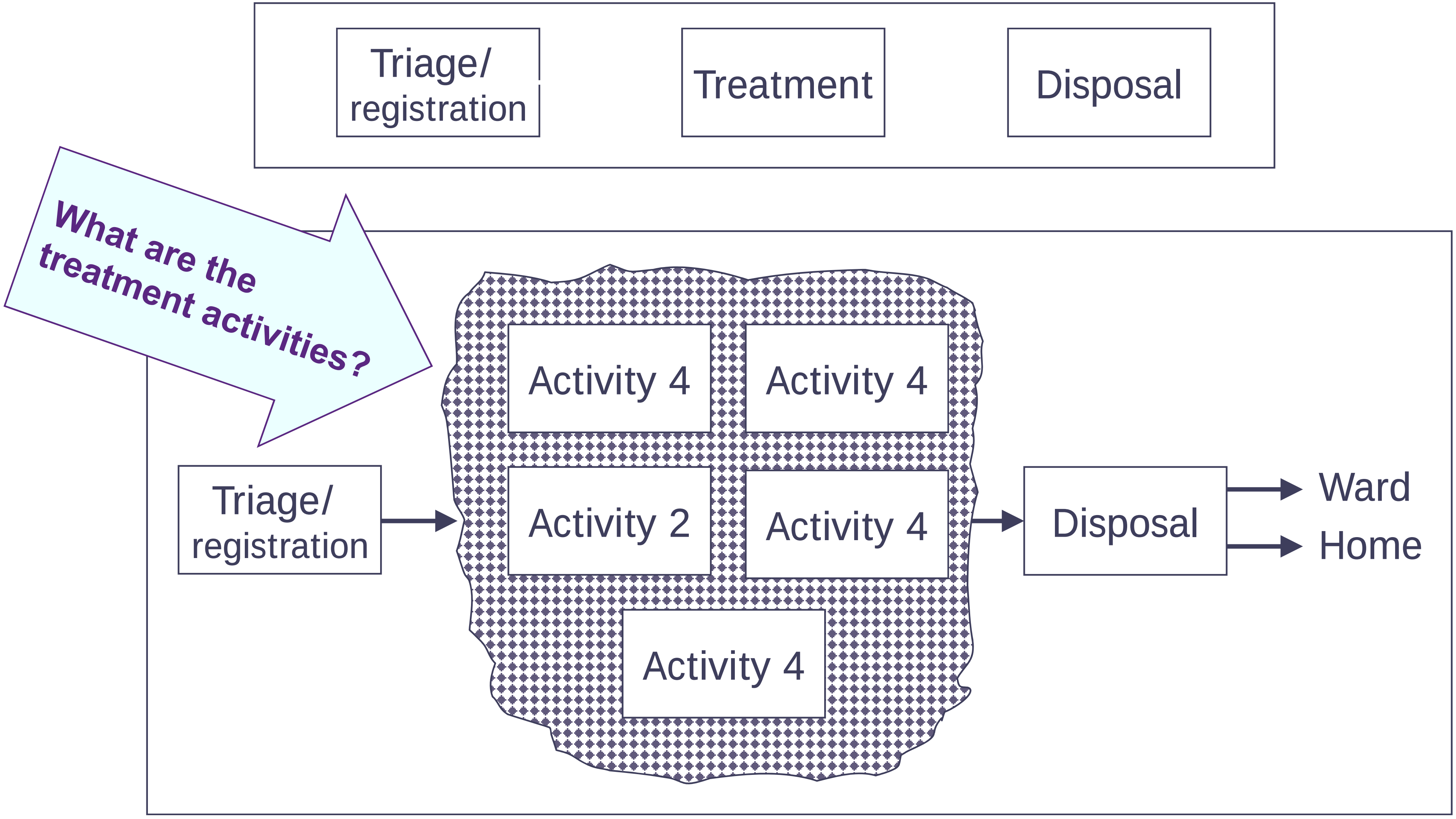
- Nothing useful fell out!

Data ≠ Answer

- Some logic required.
- Data carries medical procedures carried out on each patient.

- Can we find similar groups of procedures?

The Data Driven Idea



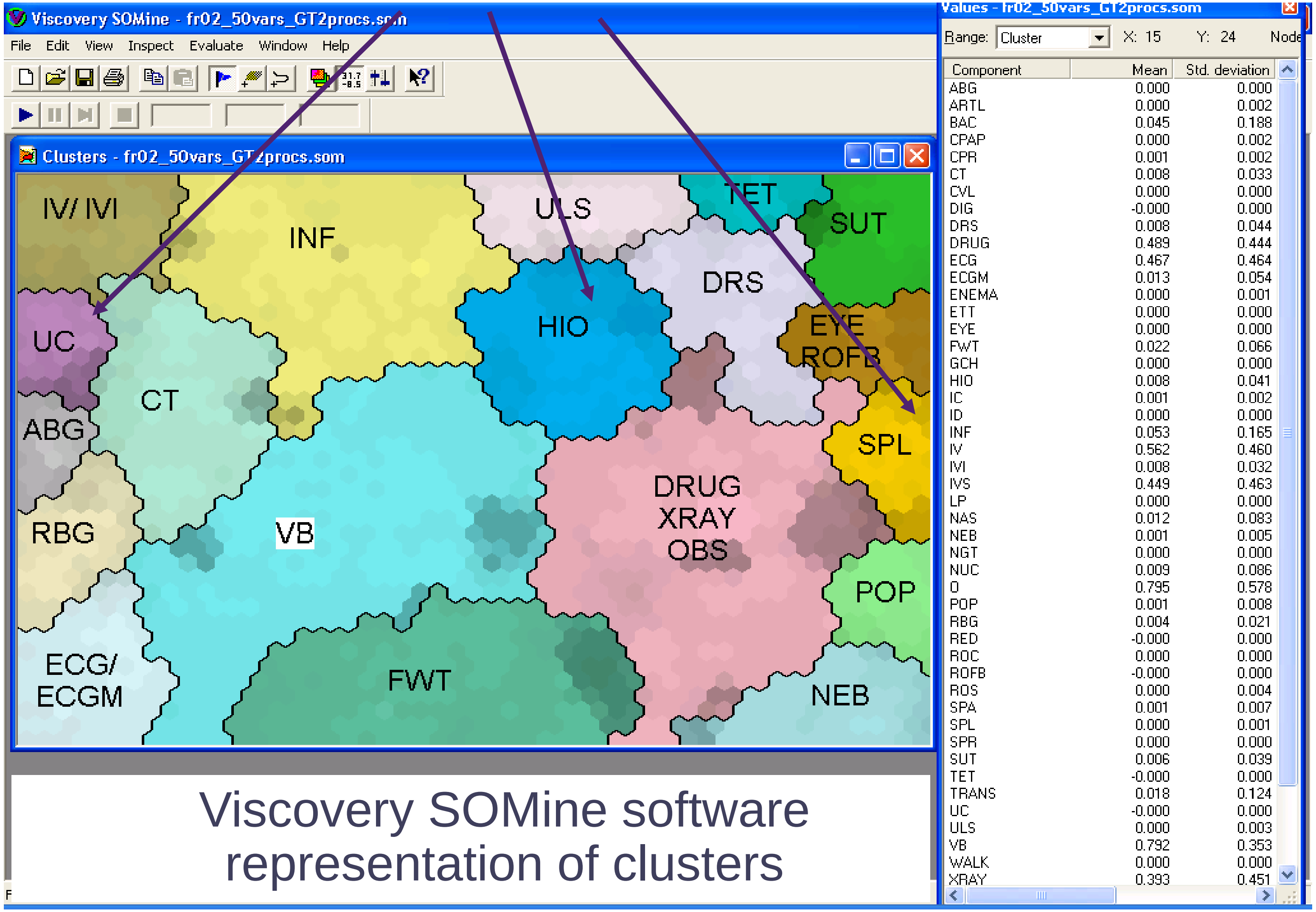
Working Hypothesis

- It is possible to find groups of patients who have similar sets of medical procedures? (e.g. undergo similar treatment, resource needs, etc.)

- Put medical procedures in SOM “box” and shake.

Self Organised Mapping

1. Clusters of patients who undergo similar procedures



Range: Cluster X: 15 Y: 24 Node

Component	Mean	Std. deviation
ABG	0.000	0.000
ARTL	0.000	0.002
BAC	0.045	0.188
CPAP	0.000	0.002
CPR	0.001	0.002
CT	0.008	0.033
CVL	0.000	0.000
DIG	-0.000	0.000
DRS	0.008	0.044
DRUG	0.489	0.444
ECG	0.467	0.464
ECGM	0.013	0.054
ENEMA	0.000	0.001
ETT	0.000	0.000
EYE	0.000	0.000
FWT	0.022	0.066
GCH	0.000	0.000
HIO	0.008	0.041
IC	0.001	0.002
ID	0.000	0.000
INF	0.053	0.165
IV	0.562	0.460
IVI	0.008	0.032
IVS	0.449	0.463
LP	0.000	0.000
NAS	0.012	0.083
NEB	0.001	0.005
NGT	0.000	0.000
NUC	0.009	0.086
O	0.795	0.578
POP	0.001	0.008
RBG	0.004	0.021
RED	-0.000	0.000
ROC	0.000	0.000
ROFB	-0.000	0.000
ROS	0.000	0.004
SPA	0.001	0.007
SPL	0.000	0.001
SPR	0.000	0.000
SUT	0.006	0.039
TET	-0.000	0.000
TRANS	0.018	0.124
UC	-0.000	0.000
ULS	0.000	0.003
VB	0.792	0.353
WALK	0.000	0.000
XRAY	0.393	0.451

Viscovery SOMine software representation of clusters

2. Detail of procedures in each cluster

Results

- Distinct clusters of “Treatment clusters” of patients emerge.
- Grouping seems medically sound.
- Confirmation of clusters by alignment with patient complaint and diagnosis (text mining).
- Confirmation of clusters in data from different hospitals.
- Confirmation using k-means.

Implications

- Data gave “as is” picture of emergency department activity.
  - Can we better forecast resource requirements?
  - Can we predict services requirements (laboratories, imaging, hospital admissions)?
- Design of predictive Decision Support Systems integrated into ED information systems.
- Improved costing (Activity Based Costing).

Data Visualisation

- Like good writing, good graphical displays of data communicate ideas with clarity, precision, and efficiency.
- Like poor writing, bad graphical displays distort or obscure the data, thus making it harder to understand or compare.

Data Visualisation

- Apart from clustering techniques that can be represented visually (e.g. with *Viscovery SOMine* for instance), it is helpful to have visual tools for inspecting the data.
  - This is part of the hypothesis formulation phase.
  - Also part of undirected knowledge discovery.
- There are many such products.
  - Really statistical and graphical tools more than data mining, although they are an important part of the data mining process.
  - GGobi: an open source visualization program for exploring high-dimensional data. http://www.ggobi.org/

A Final Comment

- Data mining has not been without criticism.
  - Some people say it is the same old thing with a trendy name.
  - It is more the methodology of data mining that is new, together with techniques for handling massive amounts of data.
- There are many myths about data mining.
  - Some reports of unsuccessful attempts to save dying businesses by investing in quick data mining solutions.

A Final Comment (continued)

- “The new technology cycle typically goes like this:

  Enthusiasm for an innovation leads to spectacular assertions. Ignorant of the technology’s true capabilities, users jump in without adequate preparation and training. Then, sobering reality sets in. Finally, frustrated and unhappy, users complain about the new technology and urge a return to ‘business as usual’.”

  Small (1997)

- Vendors who market products as “black box” solutions are contributing to this cycle.
- Nothing can replace a thorough understanding of what is inside the “black box” to get the best results.

Suggested Readings on Data Mining

- Berry, M. and Linoff, G. (2000). *Mastering Data Mining,* Wiley.
- Kennedy, R.L. et al. (1997). *Solving Data Mining Problems through Pattern Recognition*, Prentice Hall.
- Miller, T. (2005). Data and Text Mining - A Business Applications Approach, Prentice Hall.
- Shmueli, G., Bruce, P.C. and Patel, N.R. (2016). Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner, 3rd Edition, Wiley.

Week 8 Tutorial

- Use ART1 to learn to cluster 4 binary patterns according to similarity with
  - a) $\rho = 0.7$
  - b) $\rho = 0.3$
- Make sure you understand the algorithm (somewhat tedious).
- Make sure you understand what the vigilance factor is really doing.