An Introduction to Data Mining

Graham.Williams@cmis.csiro.au

CSIRO Australia

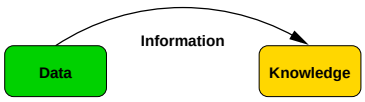
23 September 1999

Outline

- The Challenge
- The Process
- Privacy
- ACSys Data Mining

What Exactly is Being Mined?

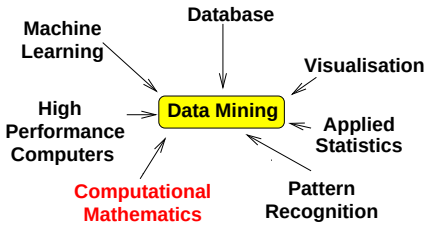
The non-trivial extraction of novel, implicit, and actionable knowledge from large datasets.



- Extremely large datasets (GB now and approaching TB)
- Discovery of the non-obvious
- Useful knowledge that can improve processes
- Can not be done manually

Technology to enable data exploration, data analysis, and data visualisation of very large databases at a high level of abstraction, without a specific hypothesis in mind.

And Where Has it Come From?



Typical Approaches to Data Mining

- **Predictive Modelling**
  tree induction, neural nets, regression
- **Database Segmentation**
  clustering, k-means, Kohonen maps,
- **Link Analysis**
  associations, sequential patterns, similar time sequences
- **Deviation Detection**
  visualisation, statistics, outlier detection

Data Mining is a Process

- A standard six step iterative process:
  1. Business Understanding (25%)
  2. Data Understanding (20%)
  3. Data Preparation (25%)
  4. Modelling (10%)
  5. Evaluation (20%)
  6. Deployment
- The “fun” has been in the modelling but the hard yakka is in the other steps
- Can Computational Mathematics help us here?

The CRISP-DM

- developed by NCR, Daimler-Benz, SPSS, OHRA
- **Cr**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
  http://www.crisp-dm.org
- Goal: Define and validate a **Data Mining Process Model**
  - applicable in diverse industry sectors
  - industry and tool neutral
  - large data mining projects executed faster, cheaper, more reliably and more manageably

Is Big Brother Watching

- Privacy is an important issue that **must** be addressed in most Data Mining exercises.
- Laws in many countries directly affect Data Mining and are required knowledge—penalties are often severe.
- The OECD Principles of Data Collection.

OECD Principles of Data Collection

- **Collection limitation**:
  Data should be obtained lawfully and fairly, while certain very sensitive data should not be held at all
- **Data quality**:
  Data should be relevant to the stated purposes, accurate, complete and up-to-date; proper precautions should be taken to ensure this accuracy
- **Purpose specification**:
  The purposes for which data will be used should be identifi ed, and the data should be destroyed if it no longer serves the given purpose

OECD Principles of Data Collection

- **Use limitation**:
  Use of data for purposes other than specifi ed is forbidden, except with the consent of the data subject or by authority of law
- **Security safeguards**:
  Agencies should establish procedures to guard against loss, corruption, destruction, or misuse of data
- **Openness**:
  It must be possible to acquire information about the collection, storage, and use of personal data

OECD Principles of Data Collection

- **Individual participation**:
  The data subject has a right to access and challenge the data related to him or her
- **Accountability**:
  A data controller should be accountable for complying with measures giving effect to all these principles

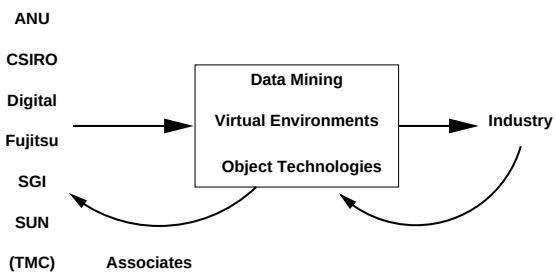
Amazon.com IS Watching?

Favourite Books by Area

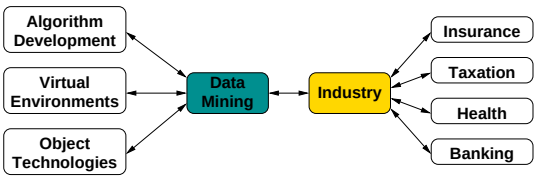
- San Jose
  - Bay Area Backroads
  - Eat, Drink, and Be Merry...
- New York City
  - For the Relief of Unbearable Urges
- Intel
  - Pentium Pro and Pentium II System Architecture
- U.S. Government
  - Mastering Microsoft Outlook 98

Making assessment by association ....

Advanced Computational Systems (ACSYS)



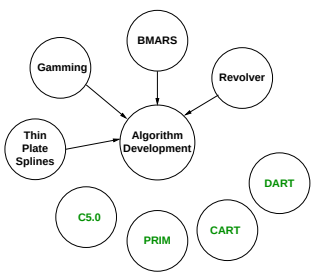
ACSys Data Mining



Research Projects

- Temporal Logics for Data Mining
- Intelligent Agents for Interestingness
- Parallel Algorithms
- Data Management
- Integrated Delivery

ACSys Parallel Algorithms



Application Projects

Mt Stromlo Observatory

NRMA Insurance Limited

Australian Taxation Office

Health Insurance Commission

Commercial Collaborators

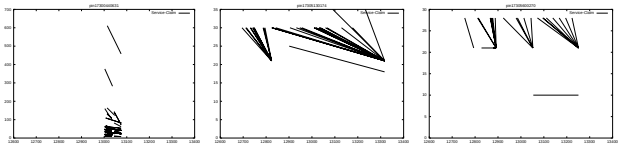
- **NRMA Insurance Ltd**
  Large database of motor vehicle insurance policy histories (millions of records with over 30 usable variables). Decision trees and regression splines form part of the toolkit. Interest in risk rating and premium setting.
- **Mt Stromlo Observatory**
  Nightly observations of the intensity of millions of stars over several years. Using Fourier, wavelet and auto regression techniques to extract features. Searching for MACHO events, to detect the missing dark matter of the universe, and performing clustering to find unusual stars.

Commercial Collaborators

- **Health Insurance Commission: Medicare**
  Terabytes of patient claims since the inception of Medicare. Using visualisation and hot spot analysis. Exploring public fraud, including doctor shoppers and other inappropriate practices. Moving on to Health Management.
- **Australian Taxation Office - ATO**
  Initial project hosted ATO Fraud Prevention and Control employee for 18 months. Assisted evaluation of data mining for fraud detection. Provided ATO with data mining tools developed in the program. Inspired development with VE Program of a profile visualisation tool. Led to a number of investigations by ATO auditors.

Example: Search for Interesting Patterns

A distinct group of behaviour identified as Medicare Claim Hoarders

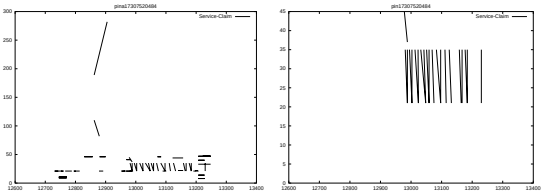


But there are millions of these plots

(plot visit to doctor versus dollar)

Inappropriate: Cash Only Regulars

Group of patients with very regular activity



All Medicare payments *versus* **Cash only payments**

(plot visit to doctor versus dollar)

Hunting for Statistical Outliers: Sofie

- Current collaborative work with NEC Japan
- Parametric Statistical Outlier Detection
  - Gaussian mixture model using Sequentially Discounting EM algorithm
- Non Parametric Statistical Outlier Detection
  - Kernel mixture model using Sequentially Discounting EM
- Application of these on real data feeding further development of algorithm

Visualisations with Virtual Environments

- Traditional mouse-keyboard-screen well explored
- VE dramatically increases workspace
- Immersive presentation with new interactive mechanisms
- Requires considerable compute power

Data Management

- Data comes in many forms from many locations
- Data regularly updated (versioning)
- Data Cleansing, Feature Extraction, Feature Derivation (versioning)
- Require tuned, fast access to data
- Algorithms want a consistent data model?

Data Management

- Is it possible to separate the algorithm from the data?
- Are there classes of *data interaction* that suit different algorithms?
- How do we best interact with heterogeneous sources of data?

Java Semantic Extension Framework

- Previous work in data management
  (Application Oriented DB, ODBC/JDBC)
- The “right” abstraction for Data Mining
- Research: Use of the Java SEF (and Persistent Java?) to provide an abstraction for accessing and manipulating data directly suited to the algorithms requirements. (Column-based access vs Row-based access)
- SEF: seamless access to data, wherever it may be?

Delivering the Integrated Application

- Use emerging standards to support delivery of tuned systems:
- Pluggable data mining:
  - Standard API to access data tuned for Data Mining
  - Standard model representations (PMML)
  - Suite of tools to process XML/PMML
  - XML Meta Descriptions of Analysis Tools
    Input requirements, Tuning Parameters, etc.

Delivering the Integrated Application

- Data Miner's Arcade
  - Java-based prototype environment
  - Provides Dataset API (move towards SEF)
  - Provides algorithm independent GUIs
  - GUIs tuned for algorithms (based on XML/XSL)
  - Scalable algorithms and scalable data management
  - Distributed data not (yet) a focus

Summary: The Data Mining Resource

- ACSys through ANU and CSIRO has developed a significant data mining facility
- Sun Enterprise: 10 processor; 5GB memory; $\frac{1}{2}$TB disk
- Large data sets from NRMA, ATO, HIC, Stromlo, for testing algorithms
- Datasets millions of rows, tens to hundreds of columns

http://www.cmis.csiro.au/alcd