Data Analytics and Business Intelligence (8696/8697)

Introducing Data Mining

Graham.Williams@togaware.com

Chief Data Scientist
Australian Taxation Office

Adjunct Professor, Australian National University
Adjunct Professor, University of Canberra
Fellow, Institute of Analytics Professionals of Australia

Graham.Williams@togaware.com
http://datamining.togaware.com

Overview

Overview

A Little History

- Database Research — where to now?
- Machine Learning — algorithms for symbolic learning.
- Statistics — long heritage and role of sampling.
- Data Mining is "one of the most important core technologies being developed today, in the same league as medical research, energy research, and environmental research." (David Hand, 2006)
- Data mining is a core technology underlying new developments in medicine, biotechnology, finance, environmental research.

Data Mining in Practise

- Data Mining research - CSIRO 1995
- Data Mining in practise - Health Insurance Commission 1995
- Data Mining projects with:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - ...

Data Mining in Practise

- Data Mining research - CSIRO 1995
- Data Mining in practise - Health Insurance Commission 1995
- Data Mining projects with:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - ...

Data Mining in Practise

- Data Mining research - CSIRO 1995
- Data Mining in practise - Health Insurance Commission 1995
- Data Mining projects with:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - ...

Data Mining in Practise

- Data Mining research - CSIRO 1995
- Data Mining in practise - Health Insurance Commission 1995
- Data Mining projects with:
  - Esanda Finance
  - NRMA
  - Mount Stromlo
  - Health Insurance Commission
  - Commonwealth Bank
  - Department of Health
  - Australian Taxation Office
  - Australian Customs Service
  - Department of Veteran Affairs
  - ...

Data Mining in Use Today

- Amazon — What else should I buy
- eBay — How to support sellers
- Facebook and LinkedIn — Who should I be firends with?
- Google — Who, what, when, how — GoolgeNow
- Finance — Bank products and fraud
- Health care — Best treatment
- Insurance — Fraud
- Government — Better Services and Compliance

Overview

Data is Fundamental

Sherlock Holmes:

> *"It is a capital mistake to theorize before one has data. Insensibly, one begins to twist facts to suit theories, instead of theories to suit facts."*

A Scandal in Bohemia (1891)
Arthur Conan Doyle

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) $\rightarrow$ low
  - Low resale value $\rightarrow$ high
  - Low resale + own a luxury car $\rightarrow$ low
  - Low resale + own a luxury car expired insurance $\rightarrow$ no
  - Angry call after claim demanding action $\rightarrow$ high
  - Customer calls after the 20th of the month $\rightarrow$ low
- But so many combinations!

Digital Detectives

The Economist, 21 September 2006

- *Combine dozens of clues to spot suspicious patterns in mountains of transactions.*
- Car Insurance: the Monday morning fraudsters — separating the genuine from the fraud
  - Claimant nearly injured (driver side impact) → low
  - Low resale value → high
  - Low resale + own a luxury car → low
  - Low resale + own a luxury car expired insurance → no
  - Angry call after claim demanding action → high
  - Customer calls after the 20th of the month → low
- But so many combinations!

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Credit Card Fraud

- Association for Payment Clearing Services
  - Buy petrol and pay by card using a machine → hmmm...
  - Soon after buy a diamond ring → very high
  - Purchase sports shoes → marginally high
  - Purchase two → higher
  - Teenage sizes → high
  - City has vibrant black market → highest
- Professional fraudsters are innovative and dynamic — we need to monitor and score individual cards and transactions.
- Block cards with sudden spikes in purchases.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Digital Footprints

We leave behind us, every day, a growing digital footprint.

- Store Purchase - loyalty cards and credit cards
- Building Access
- Computer Login
- eToll Records
- Mobile Phone
- Cameras with sophisticated image recognition

We need due diligence in collection and analysis of data — privacy protocols.

Taxation Records - Historic Perspective

- **Data Collection**
  5500 years ago Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- **Data Analysis**
  Since then people have sought ways to add value to this otherwise statically recorded information — either for good or bad!

Source http://www.crystalinks.com/cuneiformtablets.html

Taxation Records - Historic Perspective

- **Data Collection**
  5500 years ago Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- **Data Analysis**
  Since then people have sought ways to add value to this otherwise statically recorded information — either for good or bad!

Source http://www.crystalinks.com/cuneiformtablets.html

Taxation Records - Historic Perspective

- **Data Collection**
  5500 years ago Sumerian (Iraq) and Elam (Iran) peoples marked their tax records onto dried mud tablets.
- **Data Analysis**
  Since then people have sought ways to add value to this otherwise statically recorded information — either for good or bad!

Source `http://www.crystalinks.com/cuneiformtablets.html`

Overview

Machine Learning as Model Building

Data Mining is a data driven approach to understanding the world.

Deploys machine learning and statistical modelling approaches to build models from large data collections.

We are aiming to build models of the real world, as an architect would build models to get a better understanding of how things will look.

Machine Learning as Model Building

Data Mining is a data driven approach to understanding the world.

Deploys machine learning and statistical modelling approaches to build models from large data collections.

We are aiming to build models of the real world, as an architect would build models to get a better understanding of how things will look.

Machine Learning as Model Building

Data Mining is a data driven approach to understanding the world.

Deploys machine learning and statistical modelling approaches to build models from large data collections.

We are aiming to build models of the real world, as an architect would build models to get a better understanding of how things will look.

Data Mining as Model Building

- Build models of the world (regression, decision trees, neural networks, association rules, fuzzy systems, random forests, support vector machines, boosted stumps, . . .) from data that represent snippets of information about the world.
- Use these models to understand and discover patterns of interest that may provide knowledge deployable in improving business processes.

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.

Knowledge from data any way you can.

- 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.

Data Mining as Model Building

- Build models of the world (regression, decision trees, neural networks, association rules, fuzzy systems, random forests, support vector machines, boosted stumps, . . .) from data that represent snippets of information about the world.
- Use these models to understand and discover patterns of interest that may provide knowledge deployable in improving business processes.

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.

Knowledge from data any way you can.

- 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.

Data Mining as Model Building

- Build models of the world (regression, decision trees, neural networks, association rules, fuzzy systems, random forests, support vector machines, boosted stumps, . . .) from data that represent snippets of information about the world.
- Use these models to understand and discover patterns of interest that may provide knowledge deployable in improving business processes.

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.

Knowledge from data any way you can.

- 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.

Data Mining as Model Building

- Build models of the world (regression, decision trees, neural networks, association rules, fuzzy systems, random forests, support vector machines, boosted stumps, . . .) from data that represent snippets of information about the world.
- Use these models to understand and discover patterns of interest that may provide knowledge deployable in improving business processes.

- The non-trivial extraction of novel, implicit, and actionable knowledge from large databases and in a *timely* manner.

Knowledge from data any way you can.

- 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.

Data Mining Adds Value to Data

- Competitive business environment — knowledge is power
- Improved knowledge for better decision support
- Volumes of data unexplored — wealth of information
- Data sources accessible — massive data warehouses
- Available data mining tools and computational power

Overview

Business Problems

- Customer Profiling
- Customer Segmentation
- Direct Marketing
- Customer Retention
- Basket Analysis
- Fraud Detection
- Compliance
- Adverse Drug Reactions
- ...

Customer Relationship Management

- All-of-business view of customer
- E-commerce providing significant data for mining
- Amazon.com — a data mining company!
- ATO (and other Government Departments)—CRM through data mining

Fraud

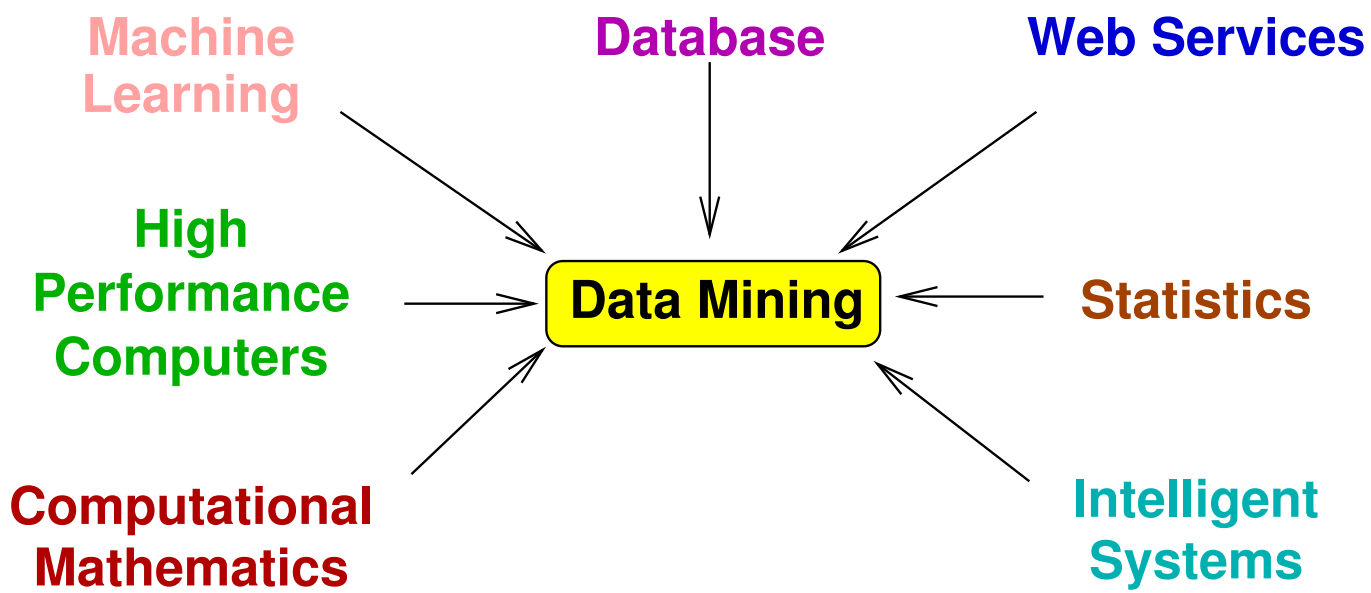
- Disguised amongst the mass of data
- A small percentage of all transactions
- A small percentage of a very large budget
  0.1% of $1billion = $1million
- NRMA: Fraud costs $70 per policy
  Dob in a fraud: Canberra Times 16 Aug 2003
- ATO and serious non-compliance

Characteristics

- Extremely large databases
- Discovery of the non-obvious
- No specific hypothesis — c.f. statistical hypothesis testing
- Useful knowledge to improve processes
- Too large to be done manually

Overview

Technologies



A Framework for Data Mining

A Formal Framework with Three Elements:

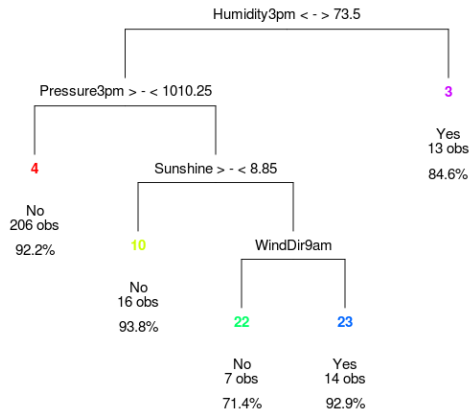
- Knowledge Representation - Language and Sentences
- Measure of Goodness
- Heuristic Search - near infinite search space

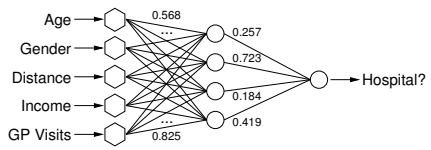
Machine Learning

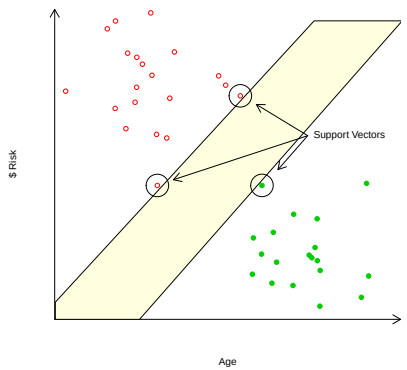
Representing Knowledge Languages:

- Decision Trees
- Classification Rules
- Neural Networks
- Support Vector Machine
- K Nearest Neighbours
- Cluster Centroids
- Association Rules

Heuristic search for the best sentences (models) in the language by some measure (criteria of best model).







Statistics

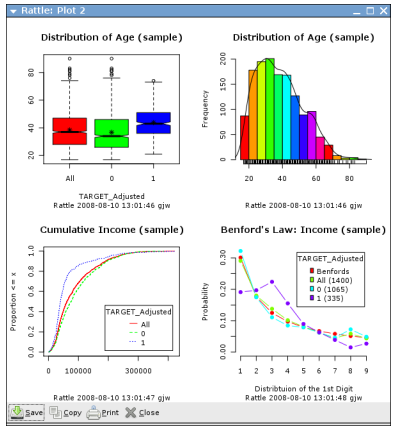
Supporting sampling and uncertainty and modelling:

- Data, Counting, Probabilities, Hypothesis testing
- Exploratory data analysis
- Predictive models:
  CART, MARS, PRIM
- Statistical thinking and
  avoiding data dredging
- Computational requirements
  $\Rightarrow$ sampling
  Data mining attempts to avoid sampling
- Data mining: hypothesis generation c.f.
  hypothesis testing

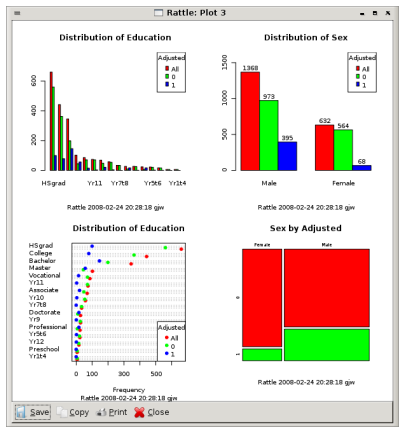
Be aware of the Bonferroni correction: testing 2 hypotheses, then use a p-value threshold of 0.025.

Visualisation

- Data visualisation is key to understanding our data, its distribution, and models



- Exploratory data analysis—has long history in Statistics, now significantly enhanced by computer science



- Visualisation in Data Mining
  - Understanding the data
  - Visualising the process
  - Visualising the results of mining

See Minard's visualisation of Napolean's march on Moscow.

High Performance Computing

- Most algorithms are computationally expensive
- Very large datasets versus slow algorithms
  - Parallel computers
  - Multiprocessor computers
- Moving back to 64bit processors
  - 32bit is limited to 4GB of data
  - 64bit is limited to 16 EB (exabytes)

(16,000 PB = 16,000,000 TB = 16,000,000,000 GB)

Software Engineering

- SE Practices apply to DM practises
  Model development process
  Repeatability and Evidence

- Developing APIs for data and models SOAP
- Delivery of data mining through web services

- Interoperability through PMML (XML)
  Predictive Modelling Markup Language

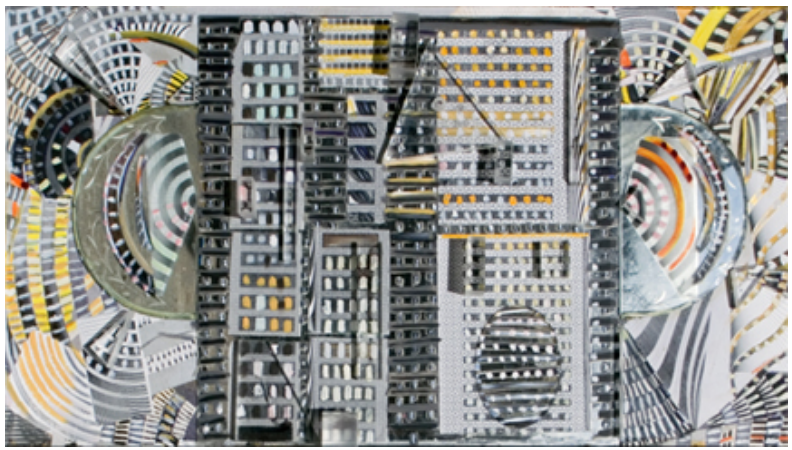
Overview

Techniques

Data mining can be descriptive or predictive.

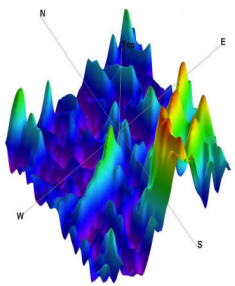
Descriptive: Identify human-interpretable patterns.

- Visualisation
- Concept Description
- Cluster Analysis
- Association Analysis
- Sequential Patterns
- Anomaly Detection



Predictive: Predict unknown or future outcomes.

- Regression—predict a continuous quantity
- Classification—predict class of entities



Concept Description

- **Characterise** *and* **discriminate** *between different concepts (classes or groups) of entities (customers)*
- Example: Amazon.com interested in the characteristics of people who tend to purchase many books so that they can provide a better service for them.
- Characterisation identifies general characteristics or features of a target class of data.
- Discrimination compares features of different classes for those that differentiate.

Cluster Analysis

- *Identify groups within data that maximise intraclass similarity and minimises interclass similarity.*
- Example: Cluster all customers based on personal characteristics and products they've purchased.
- Building models from **unlabelled** data: *unsupervised learning*.

Association Analysis

- *Identify* **attribute-value** *conditions that occur frequently together in a given set of data.*
- Example:

  $Age \in [20, 30] \cap Income \in [\$20K, \$30K] \rightarrow MP3player$

  Support=2%; Confidence=60%

Anomaly Detection

Also referred to as Outlier Analysis

- *Identify entities that do not comply with the general behaviour or model of data.* Exceptions to the rule!
- Example: Odd patterns can be easily hidden amongst 10 million transactions, but may be indicative of fraud.
- . . . but not likely — find those who live at the edge

Classification

- *Find a* **model** *which describes and distinguishes data classes or concepts for the purpose of using the model to predict the class of previously unseen entities.*
- Example: Build a model to predict whether customers are likely to purchase a download of a particular MP3 music file.
- Building models from **labelled** data: *supervised learning*.

Overview

Challenges

1. Scalability
2. Dimensionality
3. Complex and Heterogeneous Data
4. Data Quality
5. Data Matching and Linking
6. Privacy
7. Streaming: Anytime Data Mining
   give me the answer now

Lecture Summary

- Data Mining is about Analysing Data;
- Technology that is all pervasive today;
- Draws on many disciplines;
- Key disciplines of Statistics and Machine Learning.

This document, sourced from DataMiningL.Rnw revision 436, was processed by KnitR version 1.6 of 2014-05-24 and took 1.1 seconds to process. It was generated by gjw on nyx running Ubuntu 14.04 LTS with Intel(R) Xeon(R) CPU W3520 @ 2.67GHz having 4 cores and 12.3GB of RAM. It completed the processing 2014-06-21 20:22:53.