Master of Technology in Knowledge Engineering

Data Mining Methodology and Methods

Introduction to Data Mining

Dr. Zhu Fangming
Institute of Systems Science,
National University of Singapore.
E-mail: isszfm@nus.edu.sg

Agenda for Day 1

- Data Mining – What and Why?
- Applications Overview
- Identifying and planning data mining projects
- Data Mining Planning Workshop
- Data Mining Tools Overview/Demo

The Big Data Rainstorm!

- **Twitter** - 500m tweets/day (Jan 2017)
- **Facebook** – 2 billion active users (Jan 2018)
- 269 billion **emails** were sent and received in 2017
- 373 million **blogs** in 2017
- **Google** - >130 trillion pages in its index (2016)
- **Amazon**'s S3 cloud service had 2 trillion objects in 2013, with approximately 1,100,000 requests/second.
- **LinkedIn** has over 500m users (mid-2017)
- **Wal-Mart** handles 1M customer transactions/hour
- **USA** – 104 billion card transactions in 2015

Where is the deluge coming from?

- The Internet
  - Web searches, Website logs (pageviews, ad clicks…)
  - E-Commerce transactions
  - Emails, Blogs, Tweets, Social Media

- Growth of devices & tracking
  - Cellphones, Barcoding ,Tagging technologies (RFID, GPS,..) for logging goods, vehicles, people, operational events, …
  - E.g. Singapore Smart Nation pilot ~ 1000 sensors in Jurong to monitor traffic, street lights, waste bins…..

- Science & Research
  - The Sloan Digital Sky Survey amassed more in its first few weeks (in 2000) than all data collected in the history of astronomy ~ 200 GB/night
- Growth of deployed Enterprise Applications which collect huge amounts of enterprise data

ERP & CRM

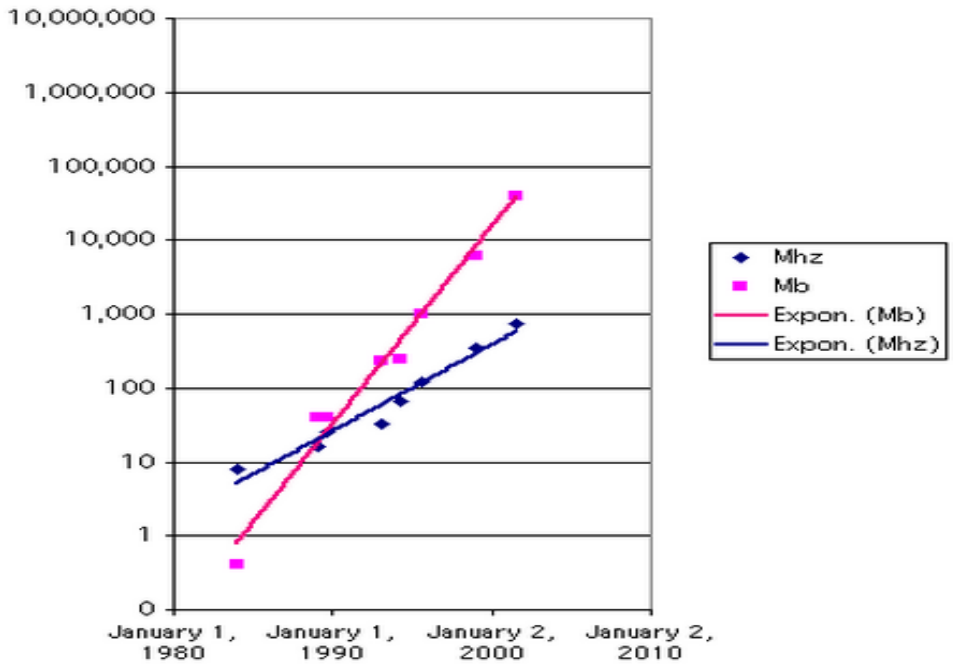
- Enterprise Resource Planning (ERP) Systems
  - Consolidate all Enterprise data into a common database
  - Before ERP, each department kept its data separate with limited scope & volume, very hard to access and integrate and see the company-wide picture!

Source : http://en.wikipedia.org/wiki/Enterprise_resource_planning

Customer Relationship Management (CRM) Systems help track, store and integrate customer data across all touch points

Enablers of the Deluge

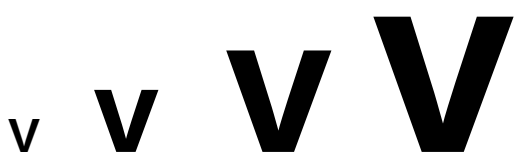
- Moore's Law
  - Processing power doubles every two years
  - "the number of transistors on integrated circuits doubles approximately every two years"
- Kryder's Law
  - Magnetic disk storage density is increasing at faster pace much faster than Moore's Law ~ doubling every 18months



These are only observations – but they are still holding true!

Characteristics of the deluge

- The 4 V's of Big Data
  - Volume => very large amount of data
  - Velocity => its arriving fast, changing often (often real-time)
  - Variety => the data is in many different formats
  - Veracity => there is often uncertainty about the data



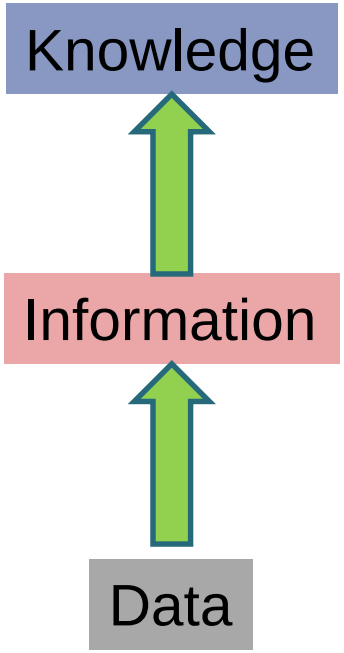
The Data Challenge

- Data is increasingly not just business numbers…
  - 350 million photos uploaded to Facebook every day* (Aug 2013)
  - 100 hours of video are uploaded to YouTube every minute** (and 6 billion hours watched each month!)
- Data is increasingly unstructured…
  - Structured data ~ databases, XML etc.
  - Unstructured ~ text, documents, images, video, audio

* http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9
** http://www.youtube.com/yt/press/statistics.html

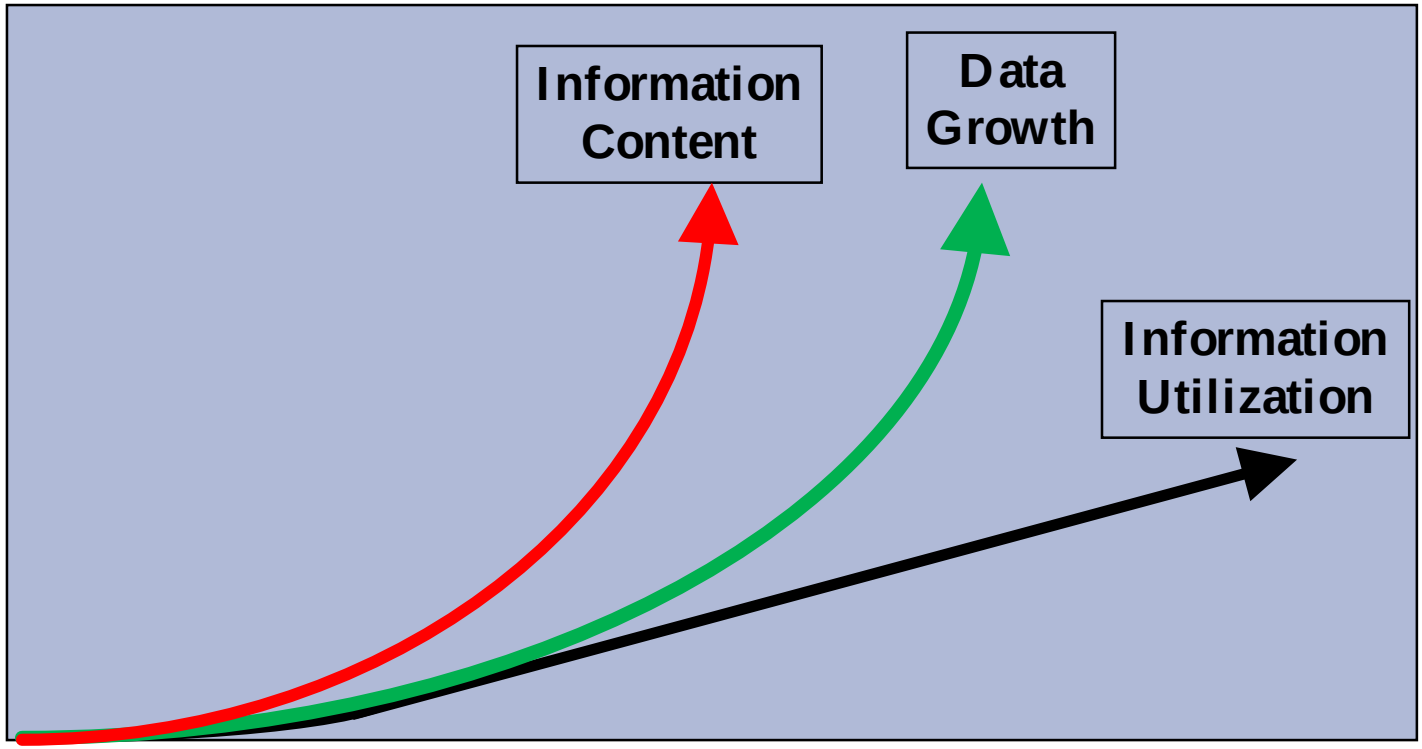
The Data Challenge

- We are not short of data, but short of information
- Data can hold critical decision making information
  - Research by IBM- only 1% of collected data is ever analysed
- Knowledge is buried in the data... we need to dig it out!



The Data Challenge

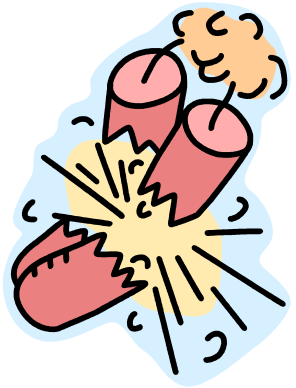
- As the number of records & fields grows, the potential number of relationships in a database grows at an exponential rate
- Utilisation is currently growing at a near linear rate



The Data Challenge

- Consider making discoveries about buying patterns (high or low propensity to buy) given 2 factors:
  - Age (young, old), Sex (male, female)
- There are 8 possible relationships that must be explored, e.g.
  - If young & male then buy is high
  - If old & male then buy is high
  - If young & female then buy is high
  - If old & female then buy is high

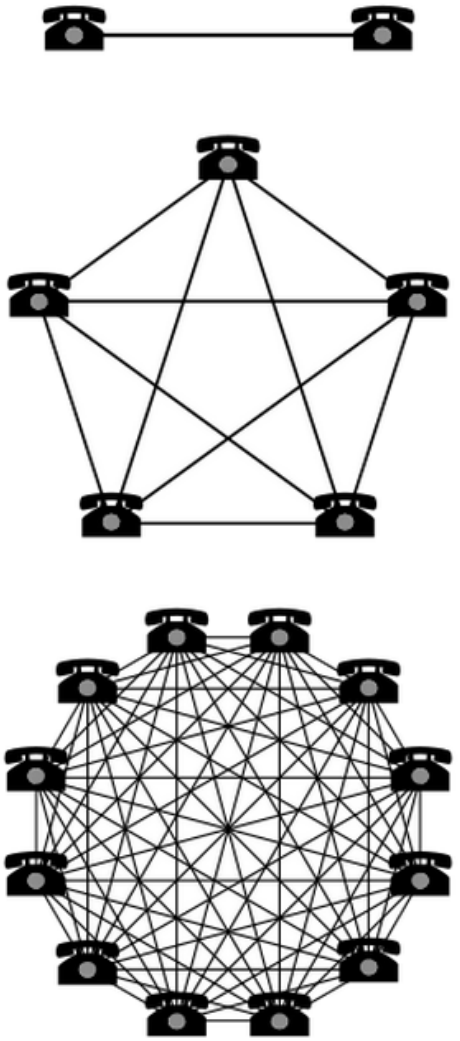
  (plus same again with "buy is low")



- How many relationships are possible given 4 boolean factors?
  - Age(Y/O), Sex(M/F), Marital Status (T/F), Family history (T/F)

The Data Challenge

- Network Data Mining suffers even more!
- Metcalf's Law
  - The value of a telecommunications network is proportional to the square of the number of connected users of the system (n^2)

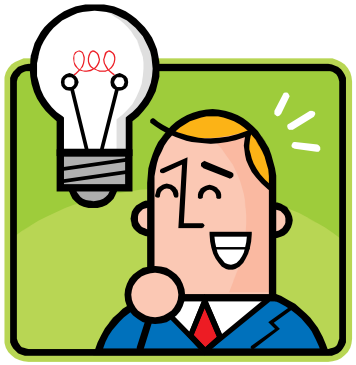


Enter Data Mining

- Some definitions of Data Mining:
  - An approach and a set of techniques to allow information and **knowledge** to be **extracted** from data
  - The practice of searching through large amounts of computerized data to **find useful patterns or trends**
  - The process of analyzing data from different perspectives and **summarizing** it into useful information

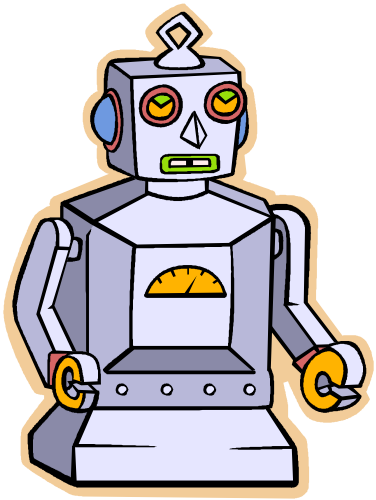
Related Jargon

- Knowledge Discovery in Databases (KDD)
  - The process of discovering useful knowledge from a collection of data. The knowledge can be in many formats
  - Knowledge Extraction is a term also used
- Knowledge Acquisition
  - Acquiring knowledge from a human expert for an Expert System
- Insights Creation
  - The utilization of data, information, and knowledge to produce valuable, previously unknown useful patterns to enable effective decision-making



What is Machine Learning?

- The focus is not on deriving knowledge or insights or aiding decision making but on enabling machines to learn to do tasks by themselves
  - *"The science of getting computers to act without being explicitly programmed" (Coursera/Stanford)*
  - *"A branch of artificial intelligence, concerns the construction and study of systems that can learn from data" (Wikipedia)*



- In the past decade, ML has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome

Data Analytics versus Data Mining?

- Data Analytics
  - The discovery and communication of *meaningful* patterns in data (all domains: business, science, engineering, etc…). Analytics often favors data visualization to communicate insight. (wikipedia)
- Business Analytics
  - Using data to gain insights and drive business planning, make ***business decisions.*** Extracting *useful* knowledge from data to solve ***business problems***

Analytics is much more goal driven…

"*Analytics usually comes with hypotheses testing. The analyst has something in mind and is looking to answer a question and has a hypothesis about that question. Data mining is more the act of discovery that lacks a hypothesis. Data mining looks, in some cases, for patterns, often through vast amounts of data, and data mining is looking for patterns and relationships that were not anticipated.*"

Business Analytics is Very Broad

Business Intelligence

Data mining is a core tool for Advanced Business Analytics

Data and Knowledge driven optimisation to determine how to do things better, cut costs etc.

Business Analytics: Focus on Actionable Outcomes

- Data Mining discoveries with actionable outcomes….
  - 72% of customers who bought baby diapers also bought beer on Thursday nights
    ➔ *position diaper & beer next to the other and have paired discounts*
  - 89% of BBA students purchased PCs within 6 months of receiving loans
    ➔ *cross-sell computer loans to these students*
  - CableTV subscribers who often watch Nickelodeon (infer-subscriber has a family with young children)
    ➔ *mail invitations for Disney-on-ice premiere*

Data Mining Examples: Target

- Examined historical buying data & demographics for all women who had signed Target baby registries in the past
  - Identified ~ 25 products that, when analyzed together, allowed each shopper to be assigned a "pregnancy prediction" score.
  - Estimated due date to within a small window, so Target could send coupons timed to very specific stages of pregnancy
  - Example Patterns found
    - buying larger quantities of unscented lotion were bought around the 3month period
    - sometime in the first 20 weeks, they loaded up on supplements like calcium, magnesium and zinc
    - buying lots of scent-free soap, extra-big bags of cotton balls, hand sanitizers and washcloths together signals they could be getting close to their delivery date

Examples: Parkson & Customer Analytics

Operates 31 retail stores in Malaysia (also in China & Vietnam)

"To maximize its market reach and revenue, the retail chain Parkson sought to better understand its customers' needs and shopping habits. It therefore uses data mining for **predictive analysis** of customer trends and behaviour, to enhance its marketing campaigns. As a result, the number of shoppers responding to Parkson's mailers increased by 40% and spending per shopper among those who responded to the mailers was also higher."

Source: Ronnie Tay, CEO IDA, March 2012

Examples: Hertz Car Rental

- Use feedback from customers to know if vehicles are run well, and whether the fleet is clean and appealing.
- Must keep track of 10,000's of customer touch points a day, divided over 8,300 locations in 146 countries
- Was done by manually processing paper surveys that took weeks to analyse. Whenever action was needed, it usually was too late and a customer was lost.

- Introduced *"Voice of the Customer"* analytics system that automatically captures customer experiences in real-time, transforming the information into useable intelligence
  - Used IBM Content Analytics System
  - Automatically categorizes comments received via email and online social media with descriptive terms such as "Vehicle Cleanliness," "Staff Courtesy," and "Mechanical Issues."
  - Hertz now gets a full picture of its customers and whether add-on features such as satellite radio enhanced the user experience

Major Application Areas

- **Marketing & Customer Analytics/CRM**
  - Acquire New Customers:
    - E.g. Which residents in a postal district should receive a discount coupon in the mail for new store location?
  - Increase Sales:
    - What is the next best product for this customer? (up-selling)
    - What other products is this customer likely to purchase? (cross-selling)
    - Product placement: should I place diapers next to beer?
    - Pricing tolerance: What is the highest price the market will bear?
  - Retain Customers /Increase Loyalty:
    - Who are my most profitable customers?
    - Which customers are most at risk of defection (churn)?

Major Application Areas

- **Prediction & Forecasting**
  - Financial Forecasting
    - What weekly revenue increase can be expected after the Mother's day sale?
    - What will the stock market do in the next few months?
  - Demand Forecasting
    - What will sales be next week/month?
    - How many hotel rooms, airline seats, hospital beds will be occupied
  - Insurance Rate Setting:
    - How likely is it that this individual will have a claim?
  - Fraud Detection:
    - How can I identify a fraudulent purchase or fraudulent insurance claim?

Major Application Areas

- **Optimisation**
  - Supply/Warehouse Optimization:
    - E.g. How much 60inch LED HDTV inventory should I hold? (Too many = costly; Too little = lost revenue)
  - Staffing Optimization:
    - E.g. What are the best times and best days to have technical experts on the showroom floor?
  - Network Optimization
    - E.g. Optimize supply chains and logistics networks ~ how many trucks & drivers needed?, find shortest routes etc,

Applications across Sectors

Data Mining can be applied in all sectors - anywhere that has data:

- ✓ **Retail** targeted marketing, customer profiling
- ✓ **Finance** credit risk analysis, investment analysis, forecasting, CRM
- ✓ **Insurance** fraud detection, risk analysis, developing pricing models, CRM
- ✓ **Telecoms** fraud detection, customer retention (churn) & CRM
- ✓ **Healthcare** best practices, customised care, ins. fraud detection
- ✓ **Biotech** drug discovery, cell biology research, medical discoveries
- ✓ **Manufg.** equipment failure analysis, inventory modeling
- ✓ **Gov.** law enforcement, tax cheats, anti-terror
- ✓ **Web** e-commerce, search engines, bots, …

What's happening in Singapore?

- NTUC Fairprice, NTUC Link, NTUC Unity Healthcare and CK Tang is using **Customer Analytics** to gain a deeper understanding of what their customers really want and, to help assess the effectiveness of their marketing campaigns and understand their customers' purchase pattern and trends.

- Best Denki is using **Business Analytics** for Inventory Optimisation to improve inventory visibility, and inventory management, to enable greater operational efficiency and reduce costs.

- YG Marketing, distributor of fashion apparel (represents brands such as Callaway, Van Heusen, Arnold Palmer, Daniel Hechter and Pierre Cardin in Singapore) is using **Business Analytics** to enhance its business-to-business supply chain to manage the distribution, replenishment and sales at various retail stores.

Source: IDA Business Analytics Seminar & Exhibition , 31st May 2012

What's happening in Singapore?

- People's Assocation uses **Business Analytics** as a way to consolidate and access a wealth of information to increase the efficiency of its internal operations and to improve its service delivery to Singapore citizens. They can now "extract the data and slice it and dice it in whichever form we want to see it – by ethnic profile, by age profile, by neighborhood profile – any way that is useful to us" – Tan Boon Huat, CEO

- CITIBANK Singapore uses **Advanced Analytics** to stay ahead of the competition. Advanced analytics allowed CITIBANK to "unearth customer behaviour and information that shows the potential for the bank to build a deeper relationship that were not apparent before analytics" – Eric Sandosham, Director Decision Management

Source: www.sas.com

What's happening in Singapore?

- Marina Bay Sands uses **Business Analytics** for its Patron Value Optimization & Hotel Revenue Optimization. Analytics is used to make sense of the huge data generated from the convention centre, hotel, shopping, casino, etc. By coordinating pricing strategies with guest satisfaction efforts, they are able to keep profits up. Based on a customer's past stays, casino floor activity and other purchases, SAS creates a forecast of what the customer will likely spend on the next visit.

- DHL Singapore uses **Business Analytics** to monitor and analyze how its drivers were driving the trucks - from acceleration and braking to idling. The data was collected and analysed to see how driving behaviour affected fuel usage leading to a 12% improvement in fuel efficiency. DHL also uses analytics to optimize route planning and vehicle usage, and to improve carbon efficiency through carbon counting and the development of a carbon dashboard.

Source: www.ida.gov.sg

Some points to bear in mind…

- Data Mining and Data Analytics are not a panacea for your business pain – it could be that there is no useful hidden pattern or value in your data!
- Success requires a good understanding of the business domain as well as the mining techniques & tools
- Data is NEVER 100% analytics-ready:
  - Missing, erroneous data should be expected!!
  - Data preparation & cleaning is difficult & time consuming; often 60% of the analytics effort is spent here