

# COMP8400: Algorithms and Techniques for Data Mining

Data mining trends, social  
impacts, and course review

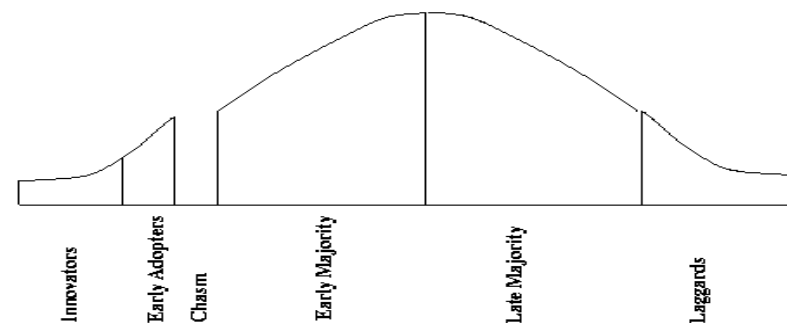
## Lecture outline

- Data mining – a hype?
  - Life cycle of technology adoption
  - Managers' business or everyone's?
- Social impacts: Threat to privacy and data security?
- Data mining trends
- Course review
  - Topics covered
  - Examination – details, questions, final course mark
  - Follow-up course, projects, etc.
  - Things to do now...

Is data mining a hype or will it be persistent?

- Data mining is a technology
- Technological life cycle
  - Innovators
  - Early adopters
  - Chasm
  - Early majority
  - Late majority
  - Laggards

Life cycle of technology adoption



- Data mining is at chasm!?
  - Existing data mining systems are too generic
  - Need business-specific data mining solutions and smooth integration of business logic with data mining functions

## Data mining: Managers' business or everyone's?

- Data mining will surely be an important tool for managers' decision making
  - Bill Gates: "Business @ the speed of thought"
- The amount of the available data is increasing, and data mining systems will be more affordable
- Multiple personal uses
  - Mine your family's medical history to identify genetically-related medical conditions
  - Mine the records of the companies you deal with
  - Mine data on stocks and company performance
  - Many more applications
- Invisible/ubiquitous data mining
  - Build data mining functions into many intelligent tools

## Trends in data mining (1)

- Application exploration
  - Development of application-specific data mining systems
  - Invisible data mining (mining as built-in function)
- Scalable data mining methods
  - Constraint-based mining: use of constraints to guide data mining systems in their search for interesting patterns
- Integration of data mining with database systems, data warehouse systems, and Web database systems
- Invisible/ubiquitous data mining
- Real-time data mining
- Graph and link mining

## Social impacts: Threat to privacy and data security?

- Is data mining a threat to privacy and data security?
  - "Big Brother", "Big Banker", and "Big Business" are carefully watching you
- Profiling information is collected every time
  - Credit card, debit card, supermarket loyalty card, or frequent flyer card, or apply for any of the above
  - You surf the Web, rent a video, fill out a contest entry form
  - You pay for prescription drugs, or present your Medicare number when visiting a doctor
- Collection of personal data may be beneficial for companies and consumers, but there is also potential for misuse
  - Medical records, employee evaluations, etc.

## Trends in data mining (2)

- Standardisation of data mining language
  - A standard will facilitate systematic development, improve interoperability, and promote the education and use of data mining systems in industry and society (similar to SQL for databases)
- Visual data mining
- Distributed data mining
- New methods for mining complex types of data
  - Multi-relational data mining (not just one data set or database table)
  - More research is required towards the integration of data mining methods with existing data analysis techniques for complex types of data
- Web mining
- Privacy protection and information security in data mining

## Course review (1)

- Topics covered
  - Data mining process and data issues in data mining, incl. data warehousing
  - Data pre-processing, integration and linkage
  - Association rule mining
  - Cluster analysis
  - Classification and prediction
  - Mining data streams and time series
  - Privacy-preserving data mining
  - Text and Web data mining
  - Data mining trends and social impacts

## Examination details

- The exam is worth 50% of the final course mark
  - It will be marked out of 50 (3 minutes 36 seconds per mark!)
- Exam is 15 minutes reading time, then 3 hours working
  - **Wednesday 17<sup>th</sup> June, 14:15-17:30, Sports Hall** – *Make sure you check this!* (take warm clothes – often pretty cold in there)
  - No calculator needed (and **not permitted**)
  - Similar format to previous exams (which are on COMP8400 Web site)
- To pass COMP8400, you need a total mark of at least 50 out of 100
  - Sum of two assignment marks, paper presentation mark, and exam mark
  - If your final mark is between 45 and 49 (inclusive) you are eligible for a supplementary examination
  - Make sure you have read (and understand) the **assessment scheme** on the COMP8400 Web site

## Course review (2)

- Four labs (using the **Rattle** data mining tool)
  - 1) Introduction to **Rattle** and data exploration
  - 2) Association rule mining
  - 3) Decision trees
  - 4) Support vector machines
- Four tutorials (paper readings and discussions)
  - 1) Data cleaning, integration and data quality
  - 2) Association rules and rule measures
  - 3) Comparing classifiers and classifier technology
  - 4) Privacy-preserving data mining, and names in text mining
- Two assignments and one paper presentation

## Examination questions

- Main focus is on understanding and describing concepts and techniques
- No question on formulas, implementation details, etc.
- Questions will cover:
  - All lectures (main part of examination)
  - Text book material (as far as covered in lectures)
  - All labs and *first* paper in each tutorial
  - Material from assignments
- Previous exams are on COMP8400 Web page
  - See link to *Old Exams*
  - Work through it, and post your answers or questions onto *Forum*

## Examination question weighting

- Total marks out of 50 (same as 2008)
  - Data mining process and data issues in data mining, including data warehousing (7 marks)
  - Data pre-processing, integration and linkage (7 marks)
  - Association rule mining (5 marks)
  - Cluster analysis (5 marks)
  - Classification and prediction (7 marks)
  - Mining data streams and time series (4 marks)
  - Privacy-preserving data mining (4 marks)
  - Text and Web data mining (8 marks)
  - End-to-end data mining, data mining trends and social impacts (3 marks)

## What's next (follow up course)

- MATH3346 (*to be confirmed*)
  - 2<sup>nd</sup> semester
  - Level: 3<sup>rd</sup> year honours
  - Main emphasis on statistical and mathematical foundations of data mining algorithms
  - Labs using mainly R
  - Lecturers from mathematics and statistics
  - Course coordinator (TBC): john.maindonald@anu.edu.au
- Data mining student projects
  - For example MComp honours and implementation projects
  - Including data linkage projects, extension to Rattle, etc.
  - Please contact Peter Christen

## What now... things to do

- Marks for 2<sup>nd</sup> paper presentation will be released before exam (aimed mid next week)
- Contact hours in coming weeks (before exam) will be advertised on COMP8400 Web site and *Forum* soon
- For questions, preferably use the *Forum*!

Good luck to everybody!