

# COMP8400: Algorithms and Techniques for Data Mining

## Introduction to course and data mining overview

### Very short introduction to data mining (1)

- Many government agencies, businesses, and research projects collect massive amounts of data
  - Ten largest decision support databases range from 17 to 100 Terabytes
  - Ten largest transaction-processing databases range from 6 to 23 Terabytes
  - Sizes have tripled between 2003 and end of 2005!
  - Source: [http://wintercorp.com/VLDB/2005\\_TopTen\\_Survey/TopTenProgram.html](http://wintercorp.com/VLDB/2005_TopTen_Survey/TopTenProgram.html)
  - Also: [http://www.businessintelligencelowdown.com/2007/02/top\\_10\\_largest\\_.html](http://www.businessintelligencelowdown.com/2007/02/top_10_largest_.html)
- Questions arise:
  - Is there any new, unexpected and potentially useful information contained in this data?
  - Can we use historical data to predict future outcomes (e.g. customer behaviour, predict fraudulent transactions, etc.)

### Lecture outline

- Very short introduction to data mining
- Course overview, syllabus and objectives
- Course modules
- Course coordinator contact details
- Course resources (Web site and text book)
- Lectures, tutorials and lab details
- Proposed assessment scheme
- What to do next

### Very short introduction to data mining (2)

- Data mining involves:
  - Database and data warehouse technologies
  - Machine learning and artificial intelligence
  - Statistics and numerical mathematics
  - Parallel and high-performance computing
  - Visualisation
- Data mining is applied in many areas, including:
  - Bioinformatics and health
  - Governments (statistics, census, taxation, social welfare)
  - Credit card and insurance companies
  - Terror, crime and fraud detection
  - Networking and telecommunications

## Very short introduction to data mining (3)

- Data mining techniques:
  - Data cleaning, pre-processing, and integration
  - Cluster analysis
  - Rule discovery (association rules)
  - Outlier detection
  - Predictive modelling
  - Classification
- Data mining applications:
  - Spatial and temporal data mining
  - Text and Web mining
  - Sequence mining (e.g. DNA, proteins)
  - Multimedia data mining (audio, images, video)

## Course syllabus

- This course introduces students to the concepts, algorithms and techniques of data mining. Topics include data warehousing, data pre-processing and integration, data mining process, algorithms and techniques, applications, social and security aspects related to data mining.
- The activities in the course will be some combination of lectures, tutorials and practical labs, reading of research papers, as well as smaller project works, as appropriate to the topic.

## Course overview

- 10 two-hour lectures
- 4 tutorials (based on selected paper readings)
- 4 practical labs (using *Rattle* data mining tool)
- 2 assignments
- Student presentation
- Final written examination

## Course objectives

- This course provides a practical focus on the technology and research in the data mining area. It focuses on the algorithms and techniques, and less on the mathematical and statistical foundations.
- Students will learn about the data quality issues involved in data mining, how different data mining techniques and algorithms work and how to apply them, the use of (open source) data mining tools, and the direction of research in data mining.

## Course modules

- Introduction (1 hour)
- The data mining process (1 hour)
- Data issues in data mining (incl. data warehouses) and data pre-processing (2 hours)
- Data integration and data linkage (2 hours)
- Mining frequent patterns and associations (2 hours)
- Cluster analysis (2 hours)
- Classification and prediction (4 hours)
- Mining time series and data streams (1 hour)
- Privacy-preserving data mining (1 hour)
- Web mining (1 hour)
- Text data mining (1 hour)
- End-to-end data mining (external lecturer, TBC) (1 hour)
- Data mining trends and social issues (1 hour)

## Course Web page

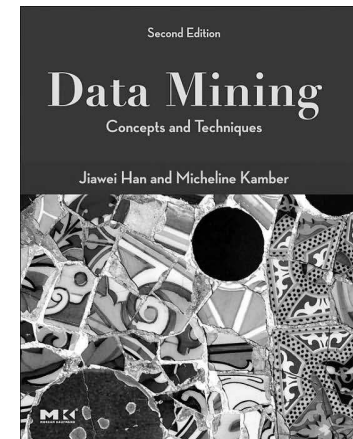
- URL: <http://cs.anu.edu.au/student/comp8400>
- News and forum
- Course schedule
- Lecture slides (normally 1-2 days before lecture); please print out (4-up) and bring along (for making your own notes)
- Lab exercises and tutorial papers
- Links to further material

## Course coordinator contact

- Course coordinator and lecturer:  
Dr Peter Christen  
Office: N330 (CSIT building, 3<sup>rd</sup> floor)  
Phone: 6125 5690
- Course e-mail: [comp8400@cs.anu.edu.au](mailto:comp8400@cs.anu.edu.au)
- Contact hours available on COMP8400 Web site (currently being finalised)
- Discussion forum (news, help, questions, etc.):  
<https://cs.anu.edu.au/streams/forum.php>  
(Data Mining)

## Text book

- *Data Mining – Concepts and Techniques* (2<sup>nd</sup> Edition)  
by Jiawei Han and Micheline Kamber, Morgan Kaufmann, 2006, ISBN 1-55860-901-6
- Around \$90 at Coop bookshop
- Book Web site with resources:  
<http://www-faculty.cs.uiuc.edu/~hanj/bk2/>



## Lectures

- One lecture slot per week:
  - Thursday 9-11, CSIT N101 (*to be confirmed*), possibly ENGN T
  - No lecture in first week after semester break (week 8), Peter is at PAKDD conference in Thailand
- Most lecture material is based on text book slides
- Please ask questions or provide comments any time!

## Practical lab sessions

- Four practical laboratory sessions (currently) scheduled in weeks 3, 5, 7, and 10.
- Using open source (free) data mining tool *Rattle* (<http://rattle.togaware.com>), developed in Canberra (by Graham Williams, ATO/Togaware)
- Possibly *Weka* (<http://www.cs.waikato.ac.nz/~ml/>) or *KNIME* (<http://www.knime.org>)
- Venue and time:
  - Thursday 14-16, CSIT building (no. 108), lab room N115/N116 (*to be confirmed*)
  - Time to be discussed and confirmed

## Tutorial sessions

- Four tutorial sessions (currently) scheduled in weeks 4, 6, 9, and 11
- Idea: Read research papers that provide a good overview of an area before tutorial, and then discuss in tutorials
  - Specific questions will be provided for each tutorial paper
- Venue and time:
  - Thursday 14-16, CSIT building (no. 108), lab room N115/N116 (*to be confirmed*)
  - Time to be discussed and confirmed

## Proposed assessment scheme

- Two assignments (worth 15% each)
  - Short essay, programming or data mining project
- Presentation (worth 20%)
  - Based on a research paper chosen by the student
  - 5-15 minutes presentation open to all interested in data mining
- Final written exam (worth 50%)
  - In normal ANU exam period, 3 hours
- Any changes to this assessment scheme will be announced on COMP8400 Web page and in lectures
  - Finalised by Friday 6th March 2009 (end of week 2)

## Final course mark and plagiarism

- To pass COMP8400 you must score at least 50 out of 100
  - Supplementary exam for students with 45 to 49 marks
- For assignments, no group work permitted!
  - We do take plagiarism very seriously!
  - If you use sources from the Internet make sure you make proper attribution (citations in quotes, proper references to scientific papers, etc.)
  - Always ask if you are in doubt!

## What now... things to do

- Discuss lab/tutorial times (so nobody has a clash...)  
(fill in lab/tutorial list with your preferences)
- Get a copy of the text book (Coop bookshop)
- Visit COMP8400 Web site, get familiar with it
- Inspect forum
- Any questions, problems, issues..?