

Wattle ► My courses ► COMP3420

Course information

course public website is here


Evaluations (SELT)

You have no outstanding surveys to complete at this time.

[SELT information and privacy](#)

Settings

▼ Course administration

 Turn editing on Edit settings

► Users

 Filters Grades Outcomes Backup Download Import

► Question bank

 Legacy course files

► Switch role to...

► My profile settings

COMP

Do Complete SELT !
Current staff and future students
value your feedback :)

Lexing coordinates, and Lexing and Peter each lecture part of the class, as specified in the course schedule table (link below).

You can contact the lecturers through email comp3420@cs.anu.edu.au. Further details about Peter and Lexing's availability are given below.

 [Course Schedule for COMP3420, Semester 1 2013](#) [Course Information Sheet](#) [2nd textbook "Mining of Massive Datasets"](#) [Further material \(text book and reading material, old examinations, external links\)](#) [ANU timetable information for S1-2013](#) [Course information at Study@ANU](#) [News forum](#) [Discussion forum](#) [Peter's contact details \(including his weekly timetable\)](#)

Lexing's contact hours are:

Wednesdays 4-5pm, and by appointment Thursdays 4-5pm

Lecture notes for lectures 01-13 can be obtained from either the links below, or as http://cecs.anu.edu.au/~xlx/comp3420/lecXX_2013.pdf where "XX" is the lecture number.

Accessing lecture notes (and other materials hosted on CECS server) from the CS lab: use Coral CDN build by the PlanetLab academic community by adding "nyud.net" to the desired URL, i.e., http://cecs.anu.edu.au.nyud.net/~xlx/comp3420/lecXX_2013.pdf

We will start the Wed and Thu lectures at 2:05pm, to allow you to come from the preceding class

DLD

22nd May 2013 at 2:00 pm

15th May 2013 at 2:00 pm

9th May 2013 at 2:00 pm

8th May 2013 at 2:00 pm

2nd May 2013 at 2:00 pm

1st May 2013 at 2:00 pm

24th Apr 2013 at 2:00 pm

17th Apr 2013 at 2:00 pm

28th Mar 2013 at 2:00 pm

27th Mar 2013 at 2:00 pm

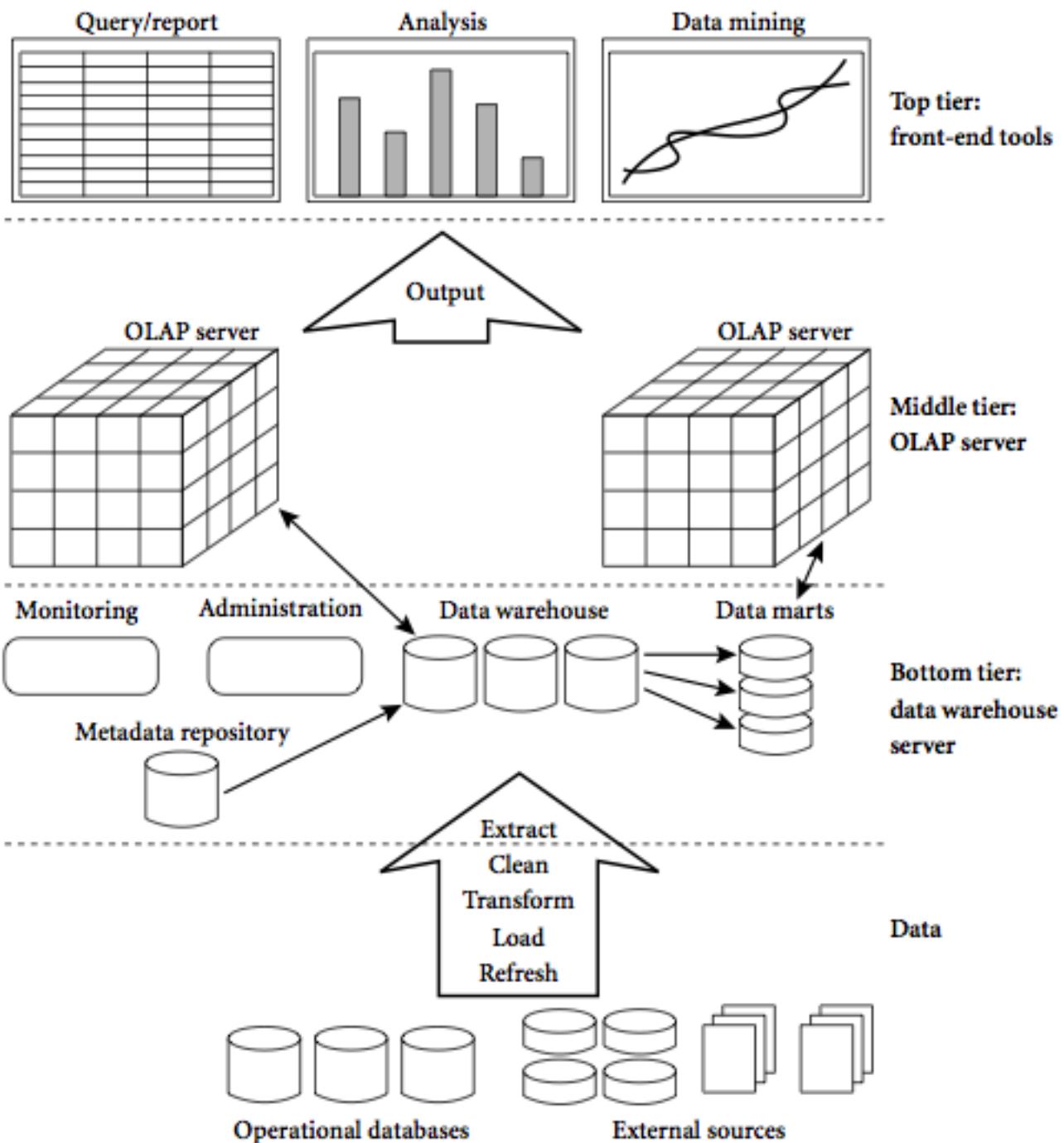
[Older >>](#)

Course Review, Exam Matters, Q&A

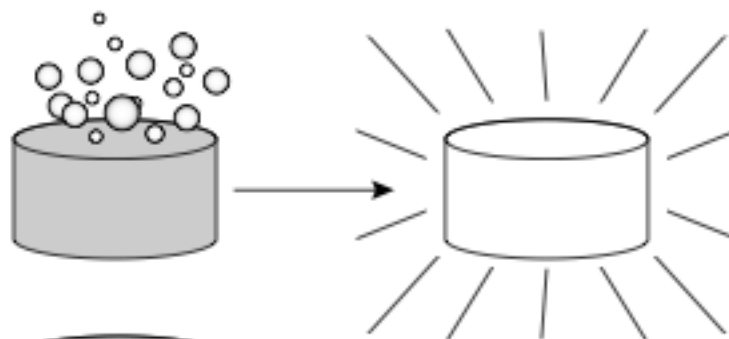
COMP3420

Main course topics

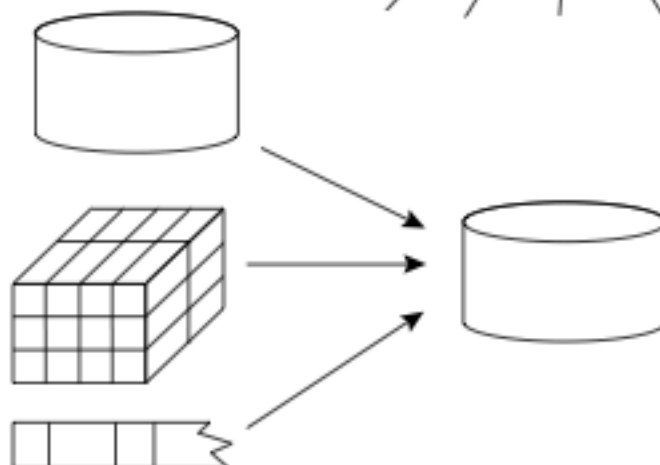
- What is data mining
- Knowing you data
- Data pre-processing
- OLAP operations
- Data cube
computation
- “Social” network
representation and
description
- Association rule mining
- Cluster analysis
- Classification
- Outlier detection
- Privacy-preserving
data mining



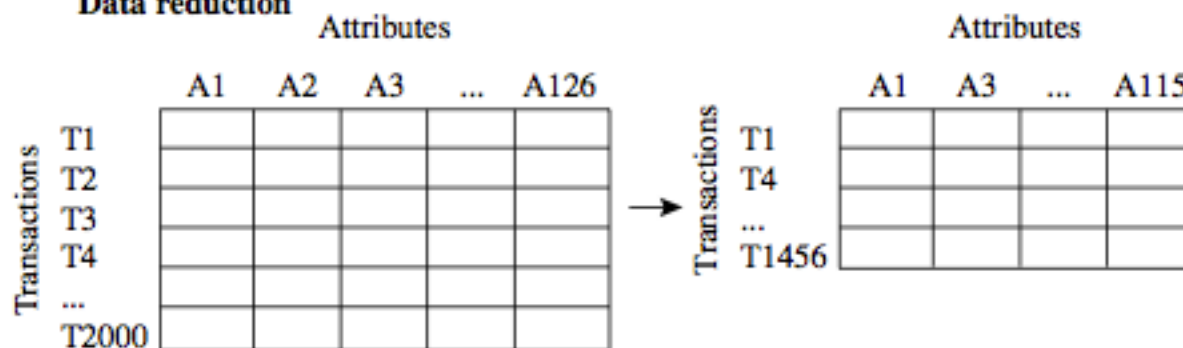
Data cleaning



Data integration

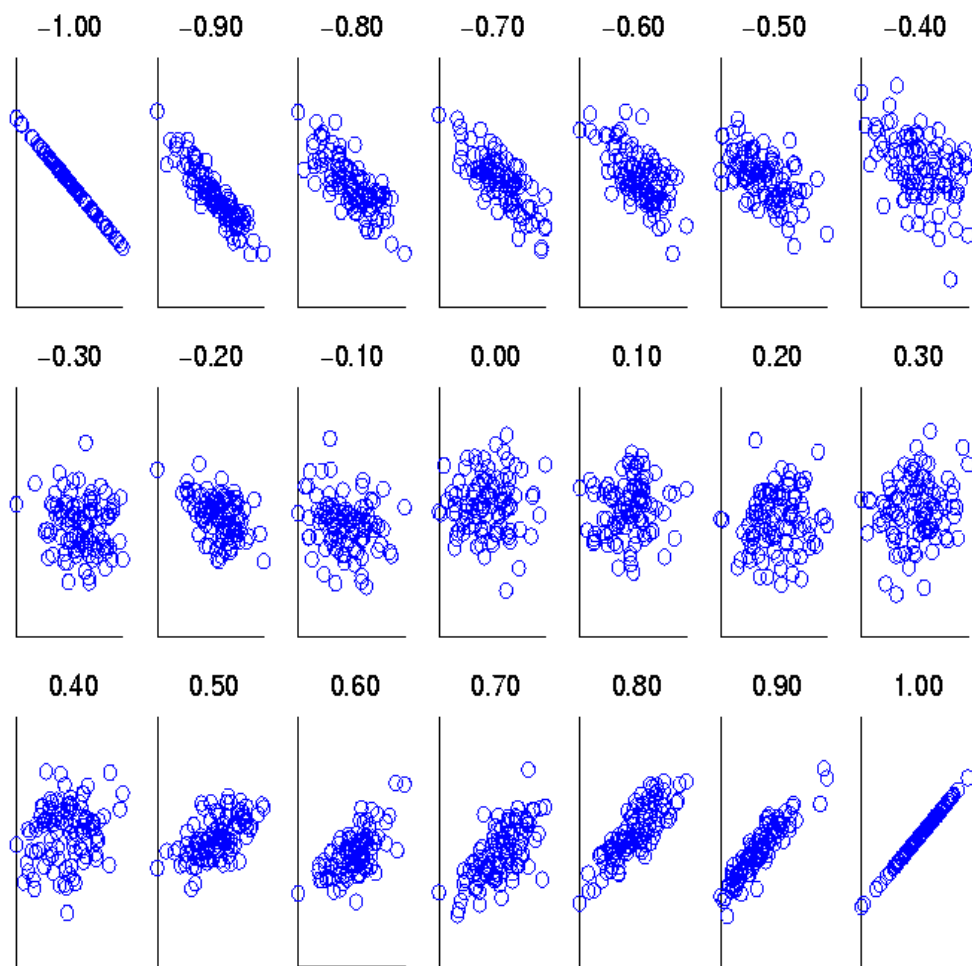


Data reduction



Data transformation

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$



What is correlation, what does it imply / not imply?

How to compute this? What can we use it for (data normalization, data reduction)

What is the discrete version of correlation?

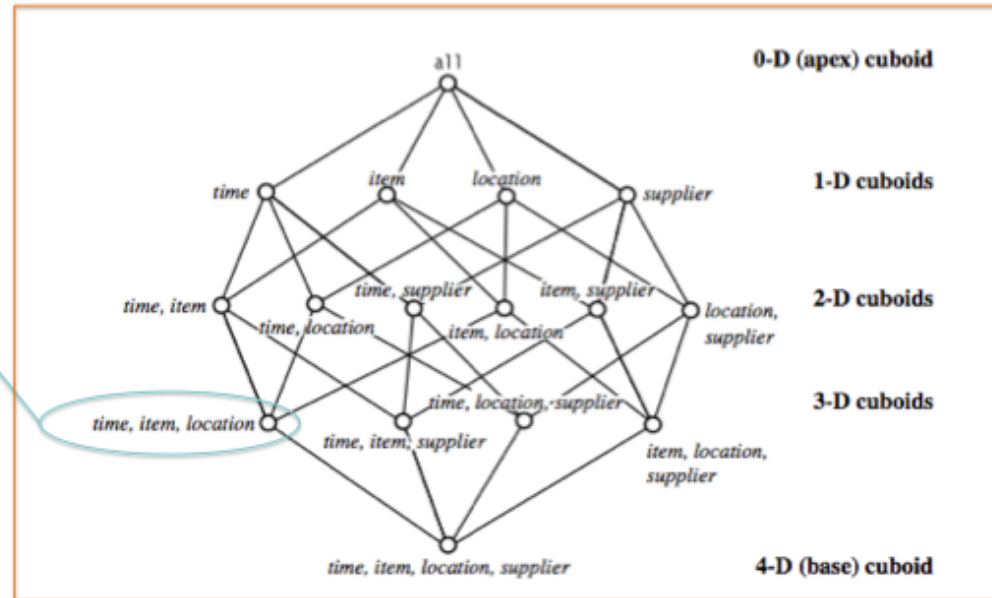
Data Cell, Cube, Cuboid, Lattice

A 3-dimensional cuboid,
OR A Data Cube ⁽²⁾

		location (cities)							
		Chicago	New York	Toronto	Vancouver				
time (quarters)	Q1	605	825	14	400	682	925	698	
	Q2	680	952	31	512	728	1002	789	
	Q3	812	1023	30	501	784	984	870	
	Q4	927	1038	38	580				
		computer	home entertainment	phone	security	item (types)			

A cell in the cuboid:
<Q1, computer, Chicago>

A data cube ⁽¹⁾
OR a lattice of cuboids



Note:

Data Cube has two different definitions*:

(1) A data cube is a lattice of cuboids. (textbook Page 158 and 113)

(2) A data cube (or OLAP cube) is a high dimensional array that contain summarized data

(textbook example 1.2 on page 13, or http://www.computerworld.com/s/article/91640/Data_Cubes)

Cuboid: a multi-dimensional array that contains measures along a number of dimensions.

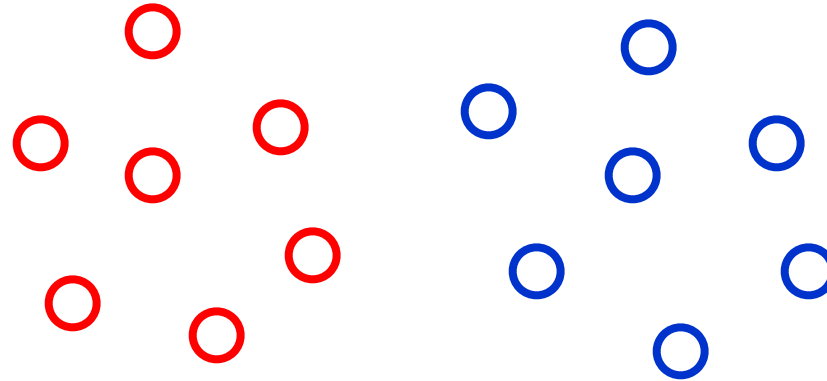
Lattice: (a drawing of) the collection of cuboids and their relationships, as shown in the example above.

Cube Cell: One value in the multi-dimensional array of a cuboid that contains a measure .

*-- wikipedia writeup explicitly acknowledge the two different views, see http://en.wikipedia.org/wiki/OLAP_cube .

How to construct data cubes? What does it mean? Who cares?

How many (dimensions, cells, levels, cuboids, lattice nodes, descendants, ancestors) ?? ...
efficiency concerns (iceberg, holistic)



How to represent a network? Matrix, edge list, adjacency list

What are the macro-level components of a network?

-- connected components (strong vs weak), density of edges

How to describe a node's position in the network?

-- centrality measures (degree, betweenness, closeness, ...)

-- what do they mean?

How to find “groups” or “communities” in a network?

Where can this be used?

Main course topics

- What is data mining
- Knowing you data
- Data pre-processing
- OLAP operations
- Data cube
computation
- “Social” network
representation and
description
- Association rule mining
- Cluster analysis
- Classification
- Outlier detection
- Privacy-preserving
data mining

Course review (contd)

- Six tutorials/labs (mix theoretical / practical tasks)
 - 1) Data warehouse and pre-processing overview
 - 2) Data cleaning, integration and pre-processing
 - 3) OLAP and data cube computation
 - 4) Introduction to **Rattle** and data exploration
 - 5) Association rule mining and clustering in **Rattle**
 - 6) Decision trees and support vector machines in **Rattle**
- Two Quizzes
- Two assignments
 - 1) Data warehousing and data pre-processing
 - 2) Data mining

Final examination

- Weight of examination is 55% of course mark
(split into roughly 50% on Data warehouse topics and 50% on data mining topics)
- Written examination of 3 hours duration
- Permitted material is one A4-sheet with notes on both sides (handwritten, not typed), plus a dictionary WITH WRITTEN APPROVAL ONLY
- Please provide short, clear answers rather than long descriptions
- Make sure that you write clearly – if we cannot read your answer you might lose marks!
- Exam will be on **Thursday 2 June, ~9-12:30**

More about Data Mining:

S2 Courses

COMP4650/COMP6490 Document Analysis

Do a Kaggle Contest to sharpen your skills!

Research projects with CS and/or NICTA

Example exam questions (1)

- What is data warehouse? How is it different from a transaction database?
- What are the three-layers of a data warehouse?
- Where should we get metadata about a DW?
- How do we spot “dirty” data? What are the methods to clean them?
- What are different methods for finding correlation?
- What are key OLAP operations? Which ones do you use to find X?
- How many cuboids are there in a data cube? How many cells are there in a data cube?

Example exam questions (2)

- What is *support*, as used in association rules mining, measuring?
- What are two main bottlenecks in Apriori algorithm?
- Give three examples where cluster analysis can be used.
- What does a dendrogram illustrate?
- What is the difference between classification and prediction?
- Describe a weakness of the neural network classifier.
- Describe two challenges of privacy-preserving data mining.