Lecture Notes

Flexible Delivery
&
External

INFS3200 Advanced Database Systems
Semester 1, 2021

Course Introduction

Professor Xue Li

The University Of Queensland
AUSTRALIA

+ The Students

- Semester 1&2, 2020

Course	Semester 1	Semester 2
INFS3200 (Undergraduates)	162	92 (FD45 + EX47)
INFS7907 (Postgraduates)	150	157 (FD82 + EX75)

- Semester 1&2, 2021

Course	Semester 1	Semester 2
INFS3200	254 (FD132 + EX122)	

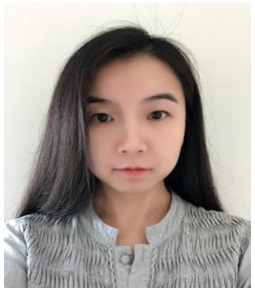
- Multiple Programs:
  - Business, Business Management, Commerce, Engineering, Science, CompSci, IT, Data Science, Exchange students, …

+ Weekly Plan

Week	Lecture	Tutorial	Practical	Assessment Due (40%)	Remarks
1	RDBMS review; Course Introduction				
2	Distributed Database Design	To. Tutor Introduction	Po. Tutor Guidance on Tool Installation		P1. out
3	Distributed Query Processing	T1. DDB Design	P1.		
4	Distributed Transaction Management	T2. DDB Query	P1.		
5	Data Warehouse Design	T3. DDB Transaction	P1.		
6	Data Warehouse Implementation	T4. DW Design	P2.	P1. Due (Online Submission 5%)	• Assignment Out • P2. Out
Semester Break (No Lecture, Tutorial, and Practical)					
7	Modern Data Management Platforms	T5. DW Implementation	P2.		
8	Data Integration	T6. SQL vs. NOSQL	P2.		
9	Data Quality Management	T7. **Pop Quiz with Tutor**	P3.	P2. Due (Online Submission 5%)	P3. Out
10	Data Security and Privacy	T8. Data Integration	P3.		
11	On the Sharing of Large Data sets	T9. Data Quality	P3.		
12	Applications of Blockchain Technology	T10. Security and Privacy	Tutor Consultation	P3. Due (Online Submission 5%)	
13	Course Review	T11. **Pop Quiz with tutor**		Assignment Due (Online Submission 25%)	
Revision Period (Week 14)					
Examination Period (Weeks 14-15) (60%)					

+ The Teaching Team

- Course Coordinator & Lecturer
  - Professor Dr Xue Li
  - Email: xue.li@uq.edu.au
  - Room: 78-652, ph: 3365 4044



- Tutors
  - Ms **Lin Yue**, l.yue@uq.edu.au
  - Ms **Kirti Khade**, k.khade@uq.net.au
  - Mr **Rischan Mafrur**, r.mafrur@uq.edu.au
  - Ms **Sivangi Mund**, s.mund@uq.net.au
  - Ms **Tejeshvini Ashre** t.ashre@uq.net.au

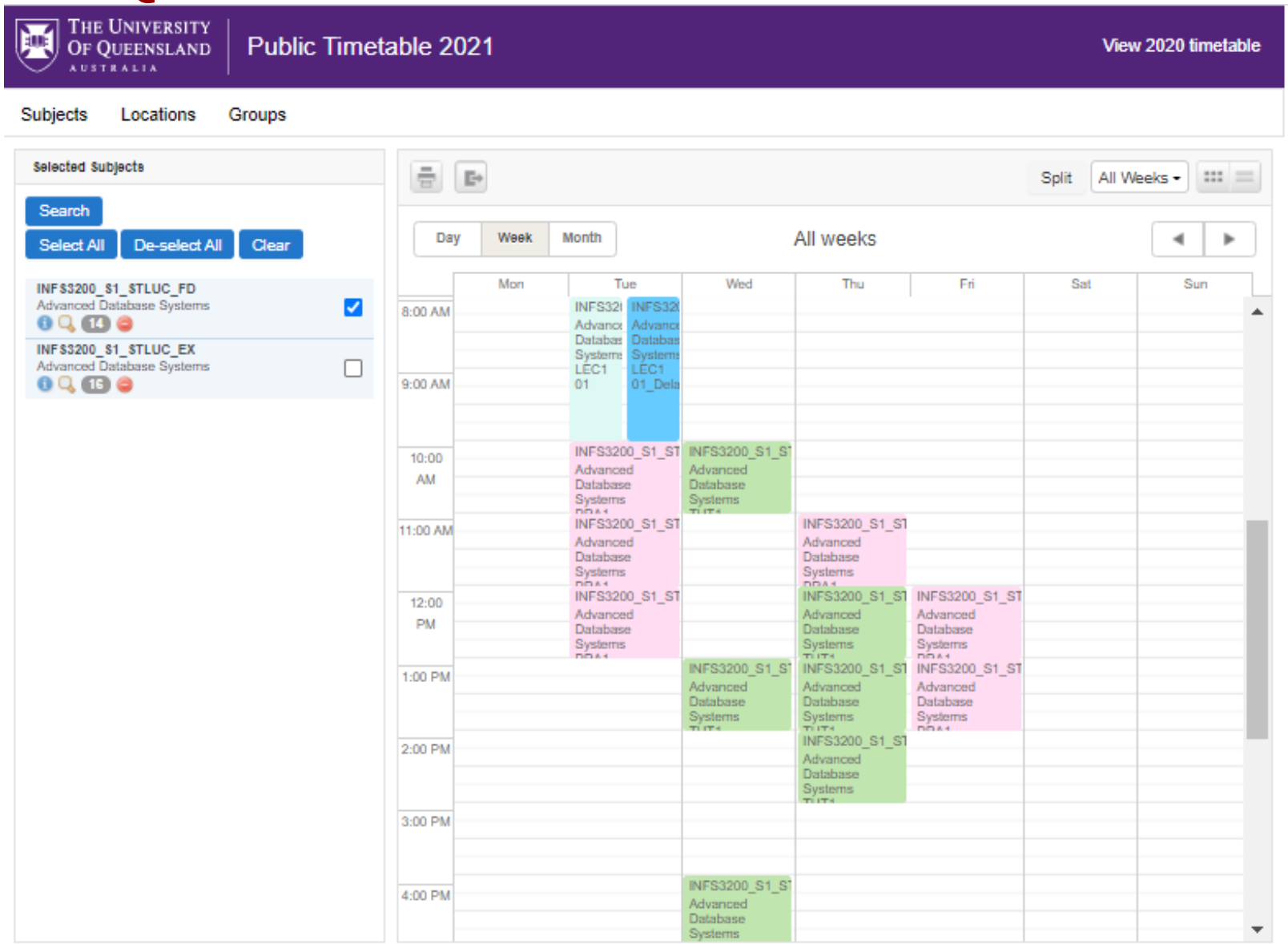
Course Website

- http://www.itee.uq.edu.au/~infs3200
- https://blackboard.elearning.uq.edu.au / or https://learn.uq.edu.au/

- ***Discussion Board*** on Blackboard – used in this class for Q&A
- You can find the following information on the web:
  - Latest news (Announcements)
  - Teaching materials
  - Lecture notes, tutorials, and practicals
  - Assessment information and lodgement
  - Links to additional information
  - Marking results
  - …

PLEASE CHECK THE WEBSITE / BLACKBORAD REGULARLY!

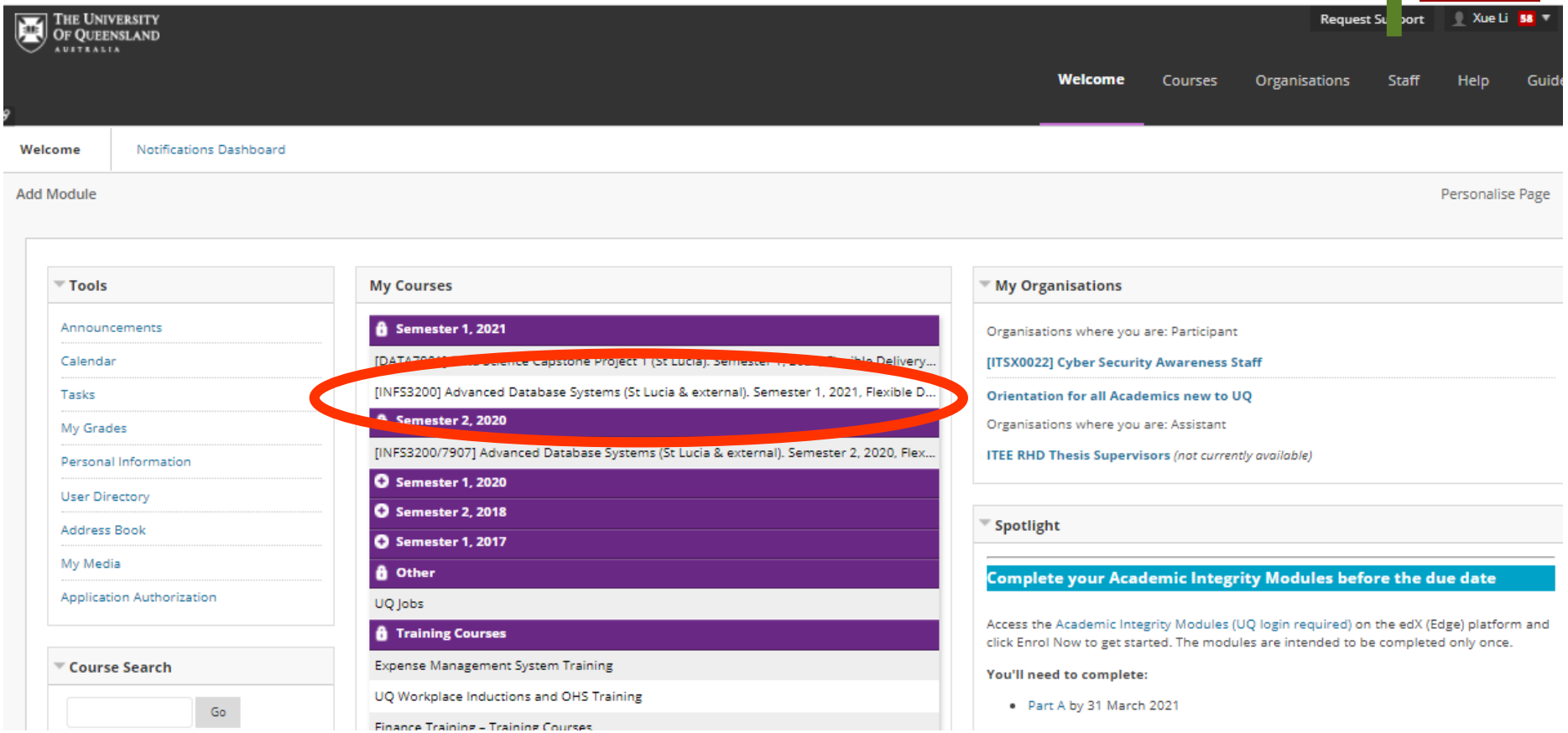
UQ Public Timetable Information



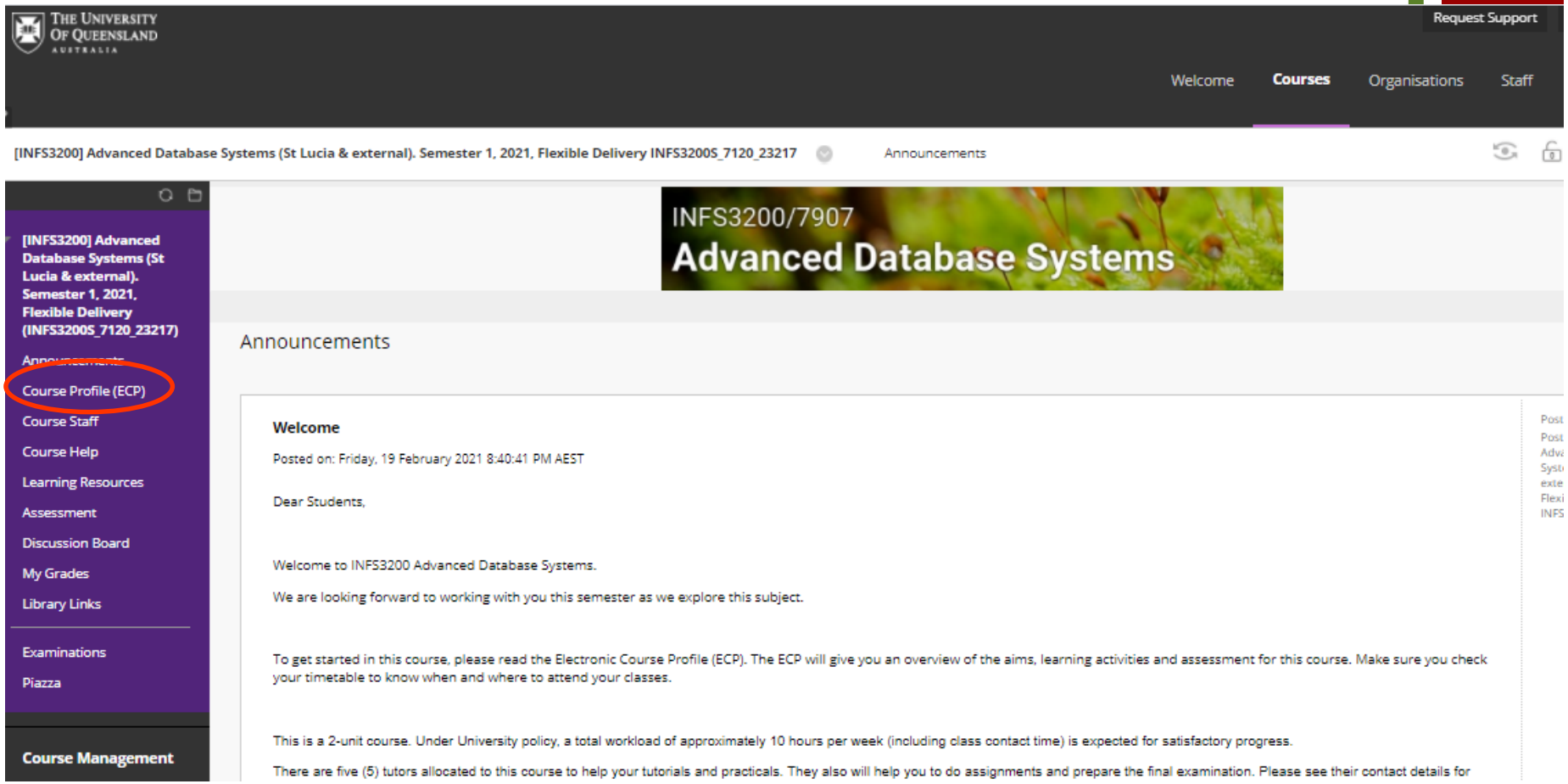
https://timetable.my.uq.edu.au/odd/timetable/#subjects

Blackboard – Course Website

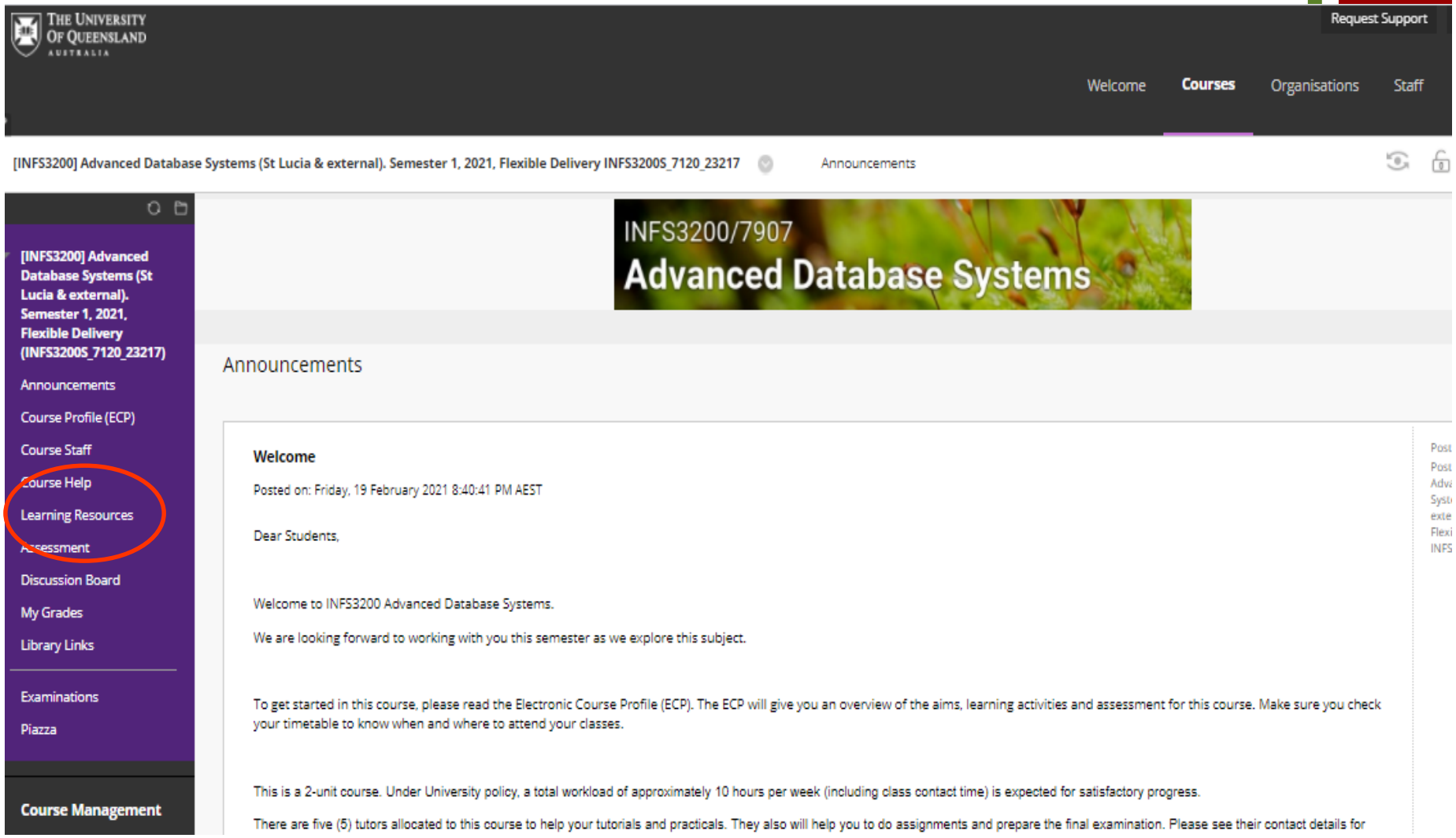
- https://learn.uq.edu.au/



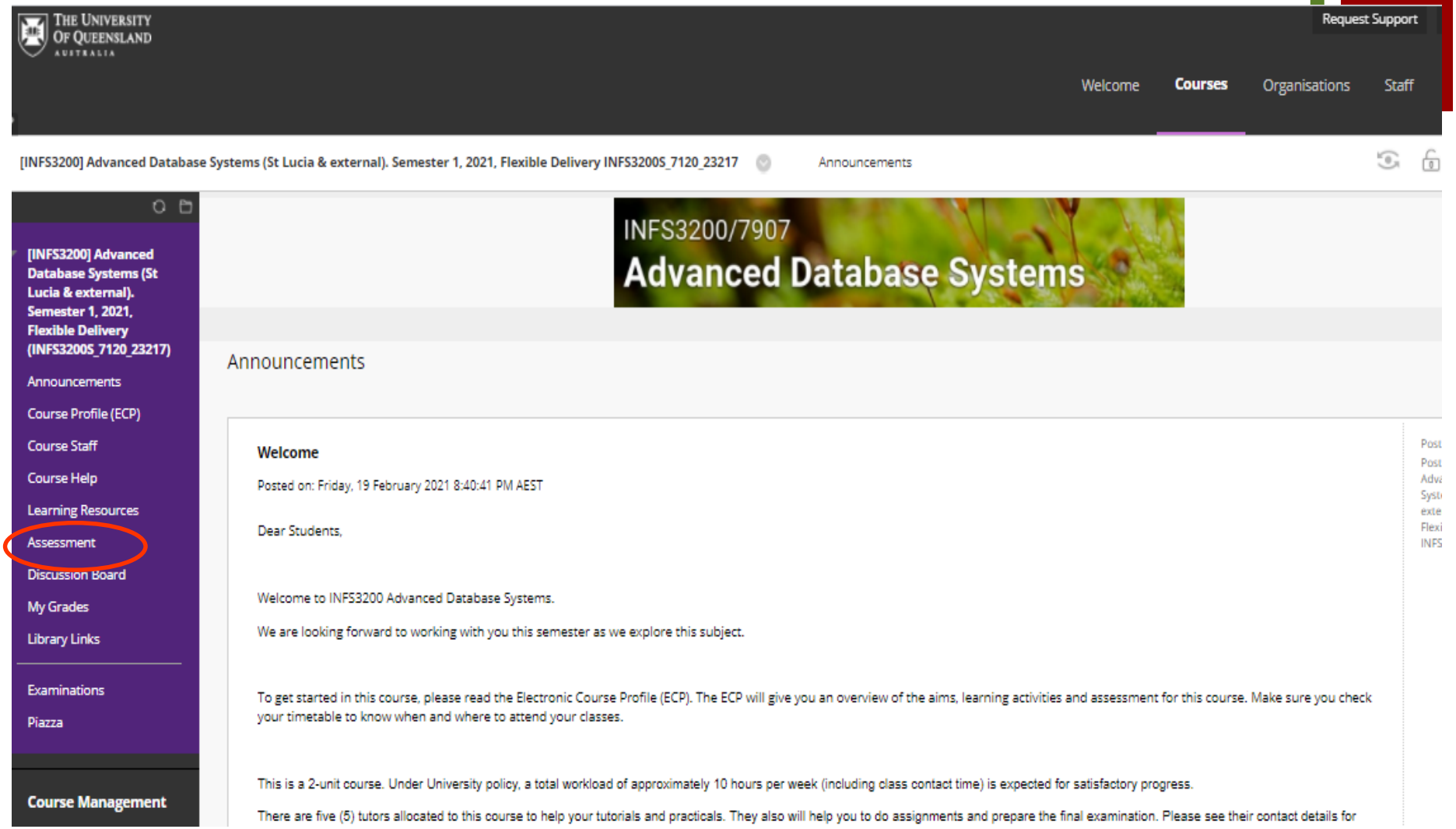
Blackboard – Course Website



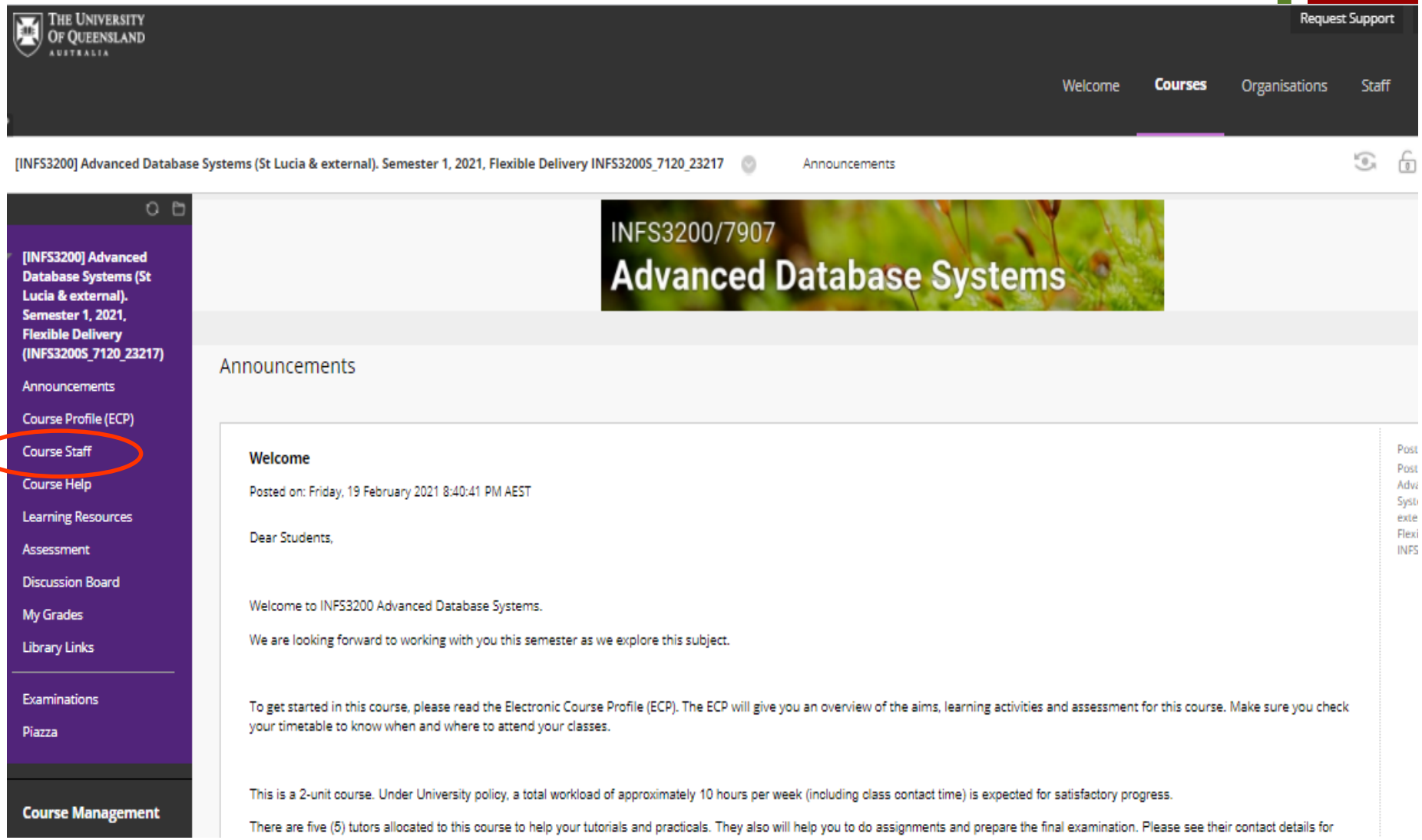
Blackboard – Course Website



+ Blackboard – Course Website

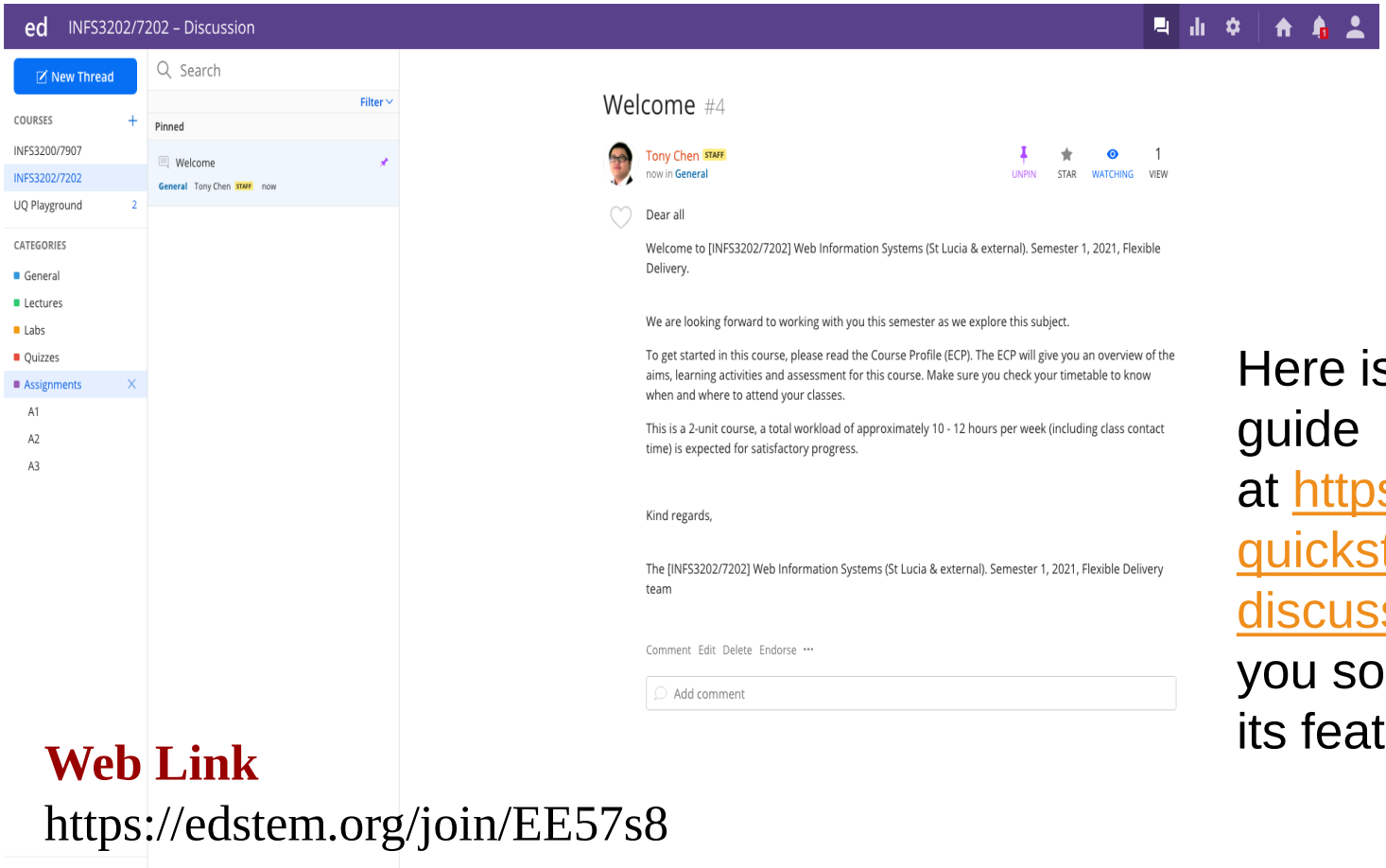


+ Blackboard – Course Website



ED-Discussion

"**Ed-Discussion**" is a online discussion platform to replace **Piazza**. This is an Australian product (data will be hosted in an Australian data center) and is used at many Australian universities and also Harvard, Stanford, Princeton, Yale & Berkeley amongst others. So, we will use it.



Web Link

https://edstem.org/join/EE57s8

Here is a quick start guide at https://edstem.org/quickstart/ed-discussion.pdf to give you some idea about its features.

Course Delivery Mode

- Due to the COVID-19, the course will be delivered in two modes:
- **Flexible mode:**
  - Online lectures (pre-recorded lecture videos in case of unpredictable situations, e.g. Internet outage);
  - Physical Tutorial and Practical sessions on campus (in-person teaching by tutors);
  - Online Tutorial and Practical sessions (pre-recorded videos).
- **External mode:**
  - Online lectures (pre-recorded lecture videos in case of unpredictable situations, e.g. Internet outage);
  - Online Tutorial and Practical sessions (pre-recorded videos).
- All modes will have teaching staff supervisions.
- Weekly consultation and discussion
  - Discussion Forum (Ed Discussion)

COVID-Safe Teaching

Help us to help you stay safe!

- Room capacities have been reduced to enable appropriate physical distancing

Please only attend the session you're signed up to

- Use hand sanitiser on your way in and out of the classroom - let us know if the supplies are low
- Try to stay at least 1.5m from your classmates and teaching staff
- Don't share pens, calculators, etc. with classmates
- Let a staff member know (email is ok) if you feel uncomfortable – there is an external version of this course available.

If you feel unwell, please stay at home and get tested – **class materials are all available online**

COVID-Safe Teaching – Close Range Teaching

Tutorials and practicals involve teaching staff helping individual students at close range, e.g. at a computer screen in computer labs (50-N301)

- Physical distancing (1.5m) is encouraged but may not be possible at all times
- **Use of a face mask is suggested** during close range teaching
  - Teaching staff will also wear a face mask
  - Teaching staff will not be able to help you at close range if you are not wearing a mask
- Use hand sanitiser on your way in and out of the classroom
- Use sanitising wipes to wipe down equipment / surfaces before and after use

If you feel unwell, please stay at home and get tested – class materials are all available online

Extension Policy

- Extensions to all other assessment items must be requested via my.UQ. You can find instructions on how to submit your request online.
- The application and supporting documentation (e.g. medical certificate) must be submitted to the ITEE **Coursework Studies office (78-425)** or by email to enquiries@itee.uq.edu.au. If submitted electronically, you must retain the original documentation for a minimum period of six months to provide as verification should you be requested to do so. Failure to produce the original documentation for verification may result in the approval of your extension being rescinded.
- An extension application granted on medical grounds will not be approved for any more than the number of calendar days the medical certificate indicates you were unfit for study. Students who are ill for more than 14 days should consider applying for withdrawal without academic penalty.
- Requests must be made at least 48 hours prior to the submission deadline, unless the medical or other circumstances are such that you could not reasonably be expected to have applied by then. Requests for extensions which are received on or after the due date may not be considered.

UQ INFS Courses

- INFS1200/7900 Introduction to Information Systems
- INFS2200/7903 Relational Database Systems
- INFS3200 Advanced Database Systems
- INFS3202/7202 Web Information Systems
- INFS3208/7208 Cloud Computing
- INFS4203/7203 Data Mining
- INFS4205/7205 Advanced Techniques for High Dimensional Data
- INFS7410 Information Retrieval and Web Search
- INFS7450 Social Media Analytics

These courses are offered by the **Data Science** Research Group in ITEE: one of the strongest database research groups in the world

+ INFS3200 Modules

- Module 1: Distributed Databases
  - 3 weeks
- Module 2: Data Warehousing
  - 2 weeks
- Mid-term break
- Module 3: Data Integration and Data Quality Control
  - 3 weeks
- Other: Data Security and Privacy
  - 1 week

Lectures + pracs + tutorials + exam

+ Readings

- Recommended readings:
  - Elmasri & Navathe (2011) ***Fundamentals of Database Systems***, 6th edition, Addison-Wesley
  - Selected papers and articles (will be made available on the course website)
  - Other books for some chapters (available in UQ library)

+ Lectures

- 2-Hour lecture session per week
- Lecture slides will be available online as PDF files
  - Likely to be updated before or after the lecture
- Recommended readings
  - Will be listed on the last slide of each module, and made available online when possible

+ Tutorial/Practical Sessions

- Tutorial and practical sessions starts from **1st Mar (Monday of Week 2) to 28/May (Friday of W13)**
- You must get permission to contact the tutors if you want to change session or catch up missing ones
- Tutorials will be focusing on the contents delivered in lectures
- Practicals will be focusing on using ORACLE products

+ Tutorials (Assessable in the Final Exam)

- 11 x 1-Hour weekly tutorials
  - Starting from week 2 Physical tutorial **(09-210, etc**)
  - See mySI-net for the schedule – *signup optional*
  - *NOTE: EX-P06, FD-P04, FD-T05, FD-T07 are cancelled and those students who have selected these sessions need to re-select them.*
- Tutorials will consist of questions and problem solving, similar to those in the assessments
  - Tutorial participation is strongly recommended
  - Solutions will be made available online
  - Tutorials might also cover additional knowledge (with explicit notes for those non-examinable content)
  - No marks given for attending tutorials
  - There are two weeks “Pop Quiz” (*free discussion questions without marks*) are arranged in this semester.

+ Practicals (Assessable 15%)

- 10 x 1-Hour practical sessions
  - Online Practicals starting from Week 2
  - Physical practical sessions are started from Week 3 **(50-N301)**
  - See mySI-net for the schedule – *signup necessary*
  - *NOTE: EX-P06, FD-P04, FD-T05, FD-T07 are cancelled and those students who have selected these sessions need to re-select them.*
- Practical exercises will allow hands-on experiences on the theoretical and conceptual topics discussed in lectures and tutorials
- Prac work is assessable (15%)
  - 3 pracs with 15 marks in total

+ The Assignment (Assessable 25%)

- An individual Assignment, Due in Week 13
- Detail to be Announced in Week 6
- The content covered in the Assignment will also be examable in the final Exam.
- Online submission of an Assignment Report document
- Help and Discussions will be given by tutors during Week 6 – 13 tutorials.

Final Exam (60%) & Final Grading

- **Closed-**book exam – 120 minutes (plus 10 minutes perusal)
- **Cover all materials** in this course.
- Mainly on **concepts and understanding.**
- The final grade (Grade 1-7, with score 1% - 100%) is determined by the total scores from the 3 Practicals (**15%** = 3 x 5%), the assignment (**25%**) and the final exam (**60%**).
- To pass the course, you must pass **the half of marks** in the Final Exam.
- *All Post-/under-graduate (Flexible Delivery/External) Students will take the same Final Exam questions.*

Final Score	Grade
85-100	**7**
75-84	**6**
65-74	**5**
50-64	**4**
45-49	3
20-44	2
0-19	1




To survive is to ask questions!

Consultation

- Problems with understanding Lectures, Tutorial problems and Pracs
  - The lecturer will make time for the online consultation available Immediately after weekly lectures.
  - Using ED-Discussion.
  - Tutors at the next available session time or appointment via email
  - Make sure you have read the ECP before asking questions.
  - Check the Noticeboard on the INFS3200 Course Website frequently **(at least once a week)**.

“Not Knowing is not an excuse”

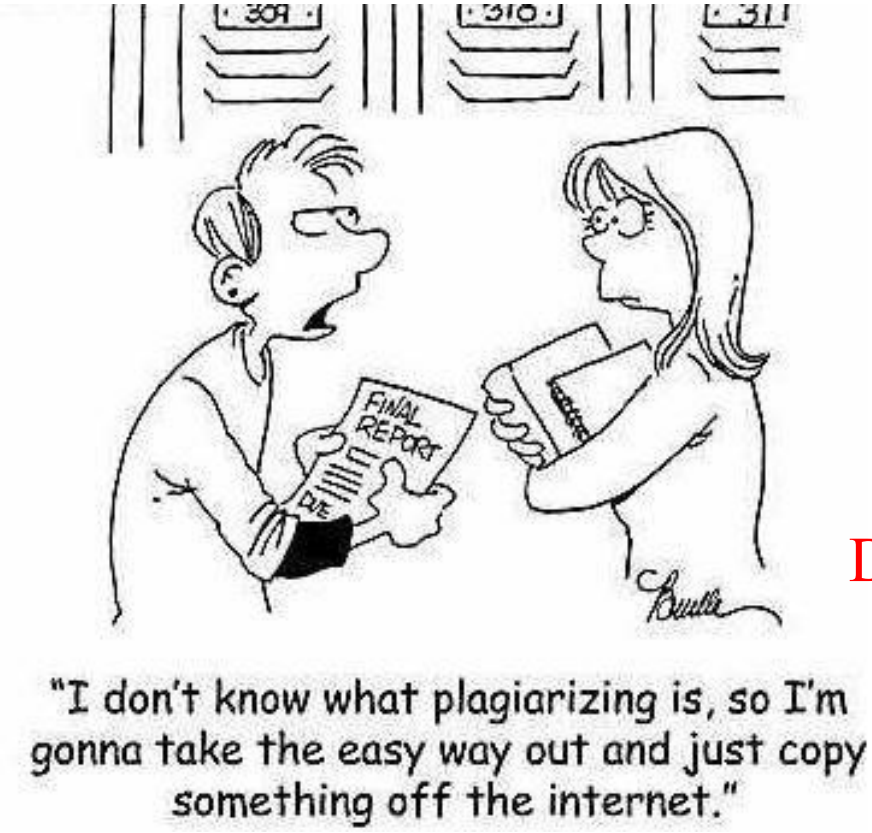


Image source: https://www.pinterest.com.au/wassef87/academic-dishonesty-and-integrity/

Ignorance is not a defense!

Get familiar with Academic Integrity at UQ

Don’t risk getting on the academic misconduct register

https://www.uq.edu.au/integrity/

+ UQ INFS Courses

- INFS1200/7900 Information Systems
- INFS2200/7903 Relational Database Systems
- INFS3200 Advanced Database Systems
- INFS3202/7202 Web Information Systems
- INFS3208/7208 Cloud Computing
- INFS4203/7203 Data Mining
- INFS4205/7205 Advanced Techniques for High Dimensional Data
- INFS7410 Information Retrieval and Web Search
- INFS7450 Social Media Analytics

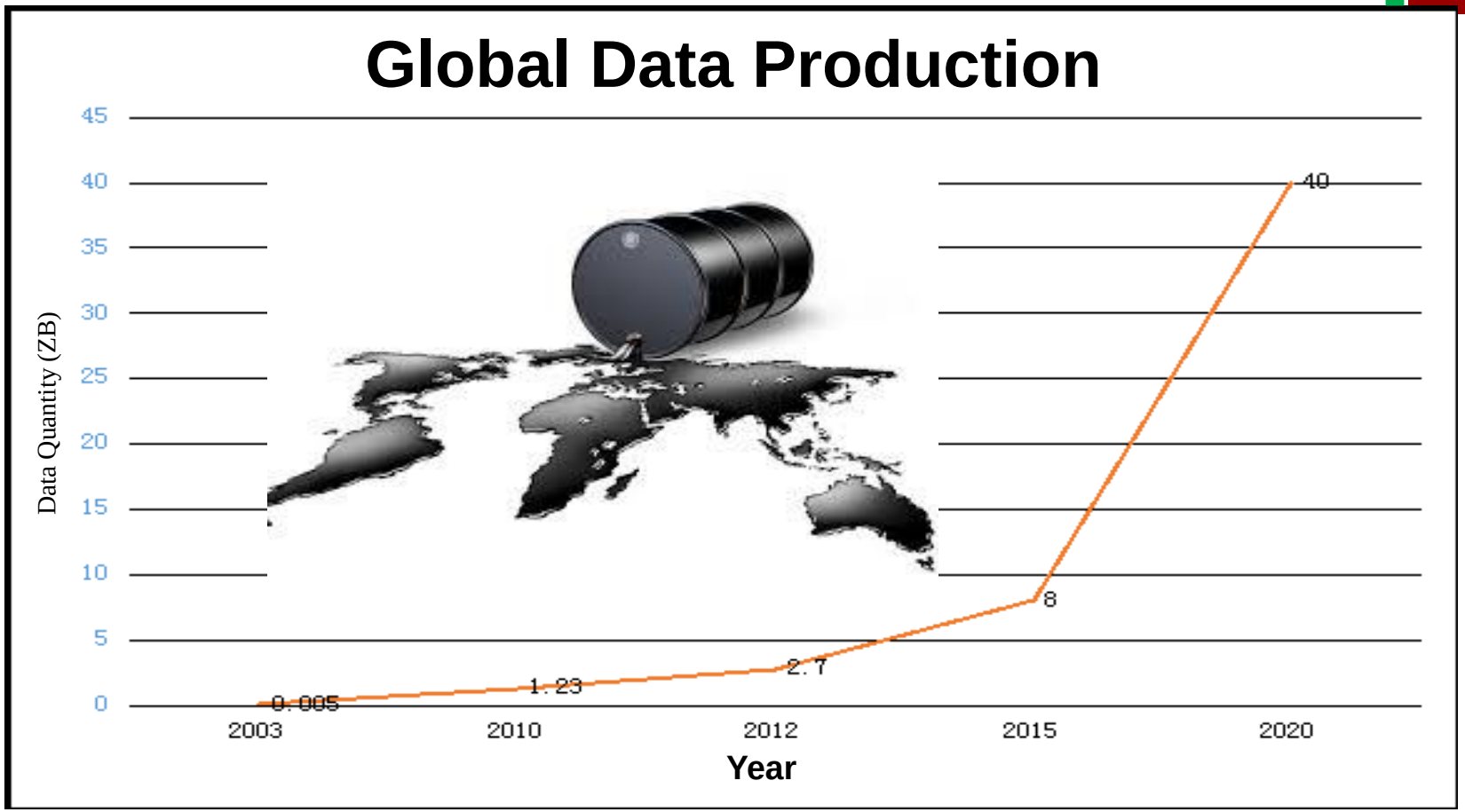
These courses are offered by the **Data Science** Research Group in ITEE: one of the strongest database research groups in the world

About Data

- What is a piece of Data?
  - **meaning**: it's the description and measurement about objects (such as patient records, observed temperature values, financial records, etc.
  - **context**: the time it is collected, the constraints imposed on it (data domain: ranges, conditions, cardinality, precision, etc), the format (e.g., inch vs, cm, ponds vs Kg, ets), etc.
  - **Structure**: dependency, 1:m, m:n, 1:1, associations, correlations, dimensionality, etc.
  - **Constraints**: monotonous changes, patterned changes, dynamics of changes, etc.
- What is the baggiest issue/problem with all data management systems?
  - Data redundancy (Update Anomaly)

“Data is the new oil of industry.”

- Dr Leif Hanlen, Technology Director, DATA61.



+ “Business is now all about Data”

“Agriculture is about data, health is about data, finance is about data, and sport is about data.”

- Prof Hugh Durrant-Whyte, Sydney University

+ INFS1200 vs INFS2200

INFS1200

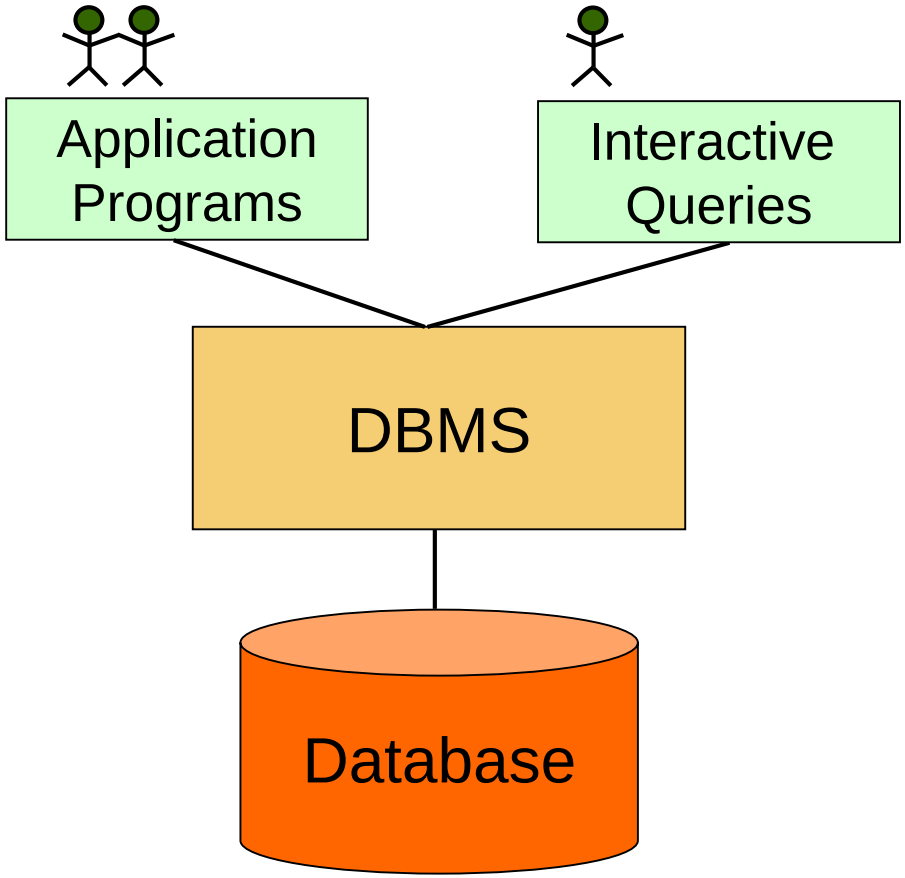
INFS2200

+ INFS2200 vs INFS3200

INFS2200

INFS3200

What is Database System?



+ RDBMS

- Key Terms in RDB
  - Tuple, Row, Record
  - Relation, Table, Relational Schema
  - Primary Key, Foreign Key
  - ER Modelling, FDs, Normalization
  - Meta Data, System Catalogue, Data Dictionary
  - SQL Query
  - Relational Algebra
- Query processing
  - Query optimisation and scheduling
  - Data indexing
- Transaction processing
  - Failures, schedules, recoverability, serialisability
- Concurrency control and recovery
  - Locking, time stamping, recovery techniques
- Security management
  - Access control and privileges

+ Example: Normal Forms

Customer

Customer ID	First Name	Surname	Telephone Number
123	Pooja	Singh	555-861-2025, 192-122-1111
456	San	Zhang	(555) 403-1659 Ext. 53; 182-929-2929
789			

1NF

Electric Toothbrush Models

Manufacturer	Model	Model Full Name	Manufacturer Country
Forte	X-Prime	Forte X-Prime	Italy
Forte	Ultraclean	Forte Ultraclean	Italy
Dent-o-Fresh	EZbrush	Dent-o-Fresh EZbrush	USA
Kobayashi	ST-60	Kobayashi ST-60	Japan
Hoch			
Hoch			

2NF

Tournament Winners

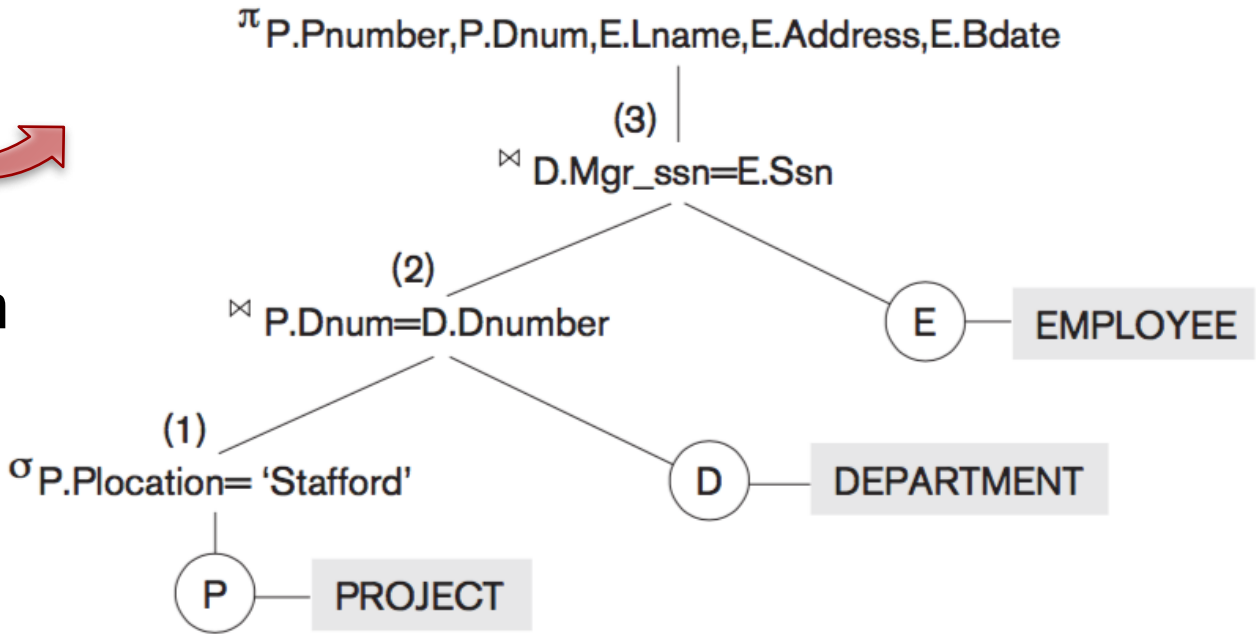
Tournament	Year	Winner	Winner Date of Birth
Indiana Invitational	1998	Al Fredrickson	21 July 1975
Cleveland Open	1999	Bob Albertson	28 September 1968
Des Moines Masters	1999	Al Fredrickson	21 July 1975
Indiana Invitational	1999	Chip Masterson	14 March 1977

3NF

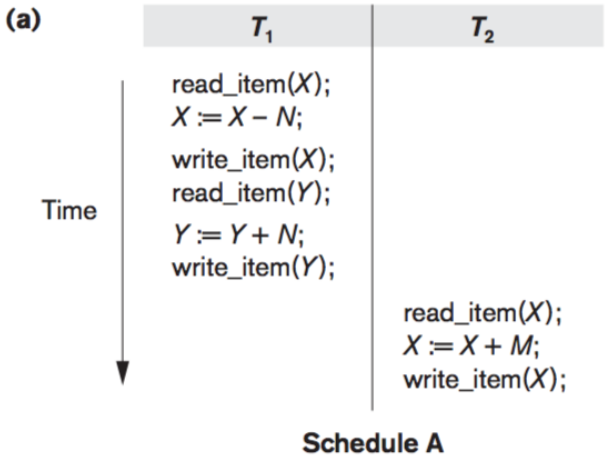
+ Example: Query Processing

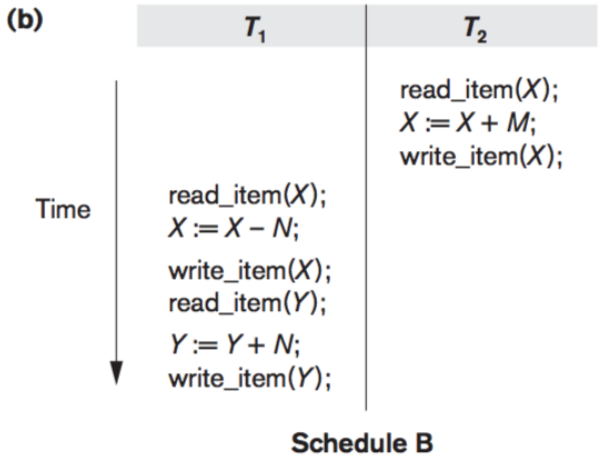

```
SELECT  P.Pnumber, P.Dnum, E.Lname, E.Address, E.Bdate
FROM    PROJECT AS P, DEPARTMENT AS D, EMPLOYEE AS E
WHERE   P.Dnum=D.Dnumber AND D.Mgr_ssn=E.Ssn AND
        P.Plocation= 'Stafford';
```

- Parsing
- Optimisation
- Execution



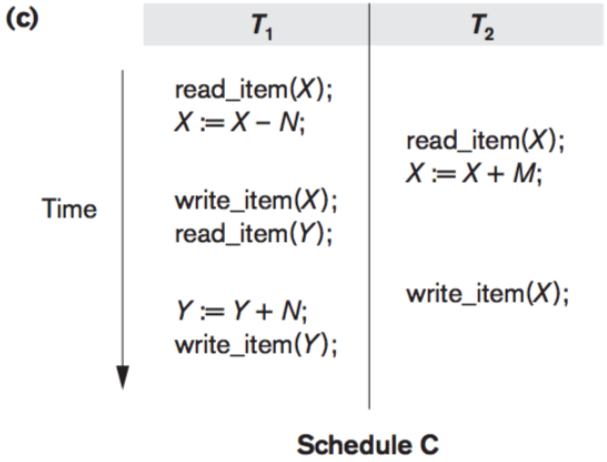
+ Example: Transaction Management

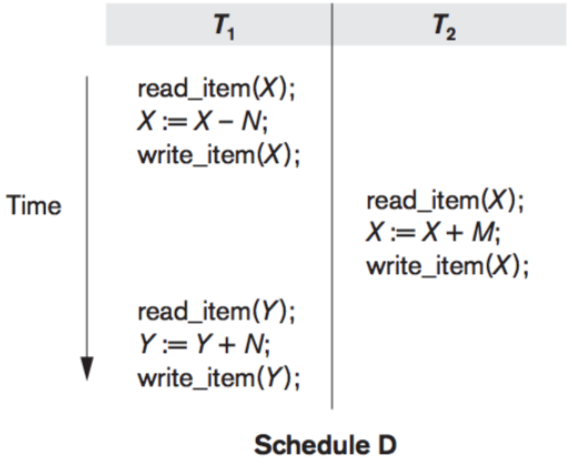




Serial

What is ACID?





Non-serial

+ Example: Concurrency Control

T_1	T_2

read_lock(Y);
read_item(Y);
unlock(Y);
write_lock(X);
read_item(X);
$X := X + Y$;
write_item(X);
unlock(X);
	read_lock(X);
	read_item(X);
	unlock(X);
	write_lock(Y);
	read_item(Y);
	$Y := X + Y$;
	write_item(Y);
	unlock(Y);

T_1'	T_2'

read_lock(Y);
read_item(Y);
write_lock(X);
unlock(Y)
read_item(X);
$X := X + Y$;
write_item(X);
unlock(X);
	read_lock(X);
	read_item(X);
	write_lock(Y);
	unlock(X)
	read_item(Y);
	$Y := X + Y$;
	write_item(Y);
	unlock(Y);

T_1	T_2

read_lock(Y);
read_item(Y);
unlock(Y);	

	read_lock(X);
	read_item(X);
	unlock(X);
	write_lock(Y);
	read_item(Y);
	$Y := X + Y$;
	write_item(Y);
	unlock(Y);

write_lock(X);
read_item(X);
$X := X + Y$;
write_item(X);
unlock(X);	

What is 2-Phase Locking?

+ Example: Recovery

T_1
read_item(A)
read_item(D)
write_item(D)

T_2
read_item(B)
write_item(B)
read_item(D)
write_item(D)

T_3
read_item(A)
write_item(A)
read_item(C)
write_item(C)

T_4
read_item(B)
write_item(B)
read_item(A)
write_item(A)

[start_transaction,T_1]
[write_item, T_1, D, 20]
[commit, T_1]
[checkpoint]
[start_transaction, T_4]
[write_item, T_4, B, 15]
[write_item, T_4, A, 20]
[commit, T_4]
[start_transaction, T_2]
[write_item, T_2, B, 12]
[start_transaction, T_3]
[write_item, T_3, A, 30]
[write_item,T_2, D, 25] ← System crash

What are undo and redo?

What is checkpoint?

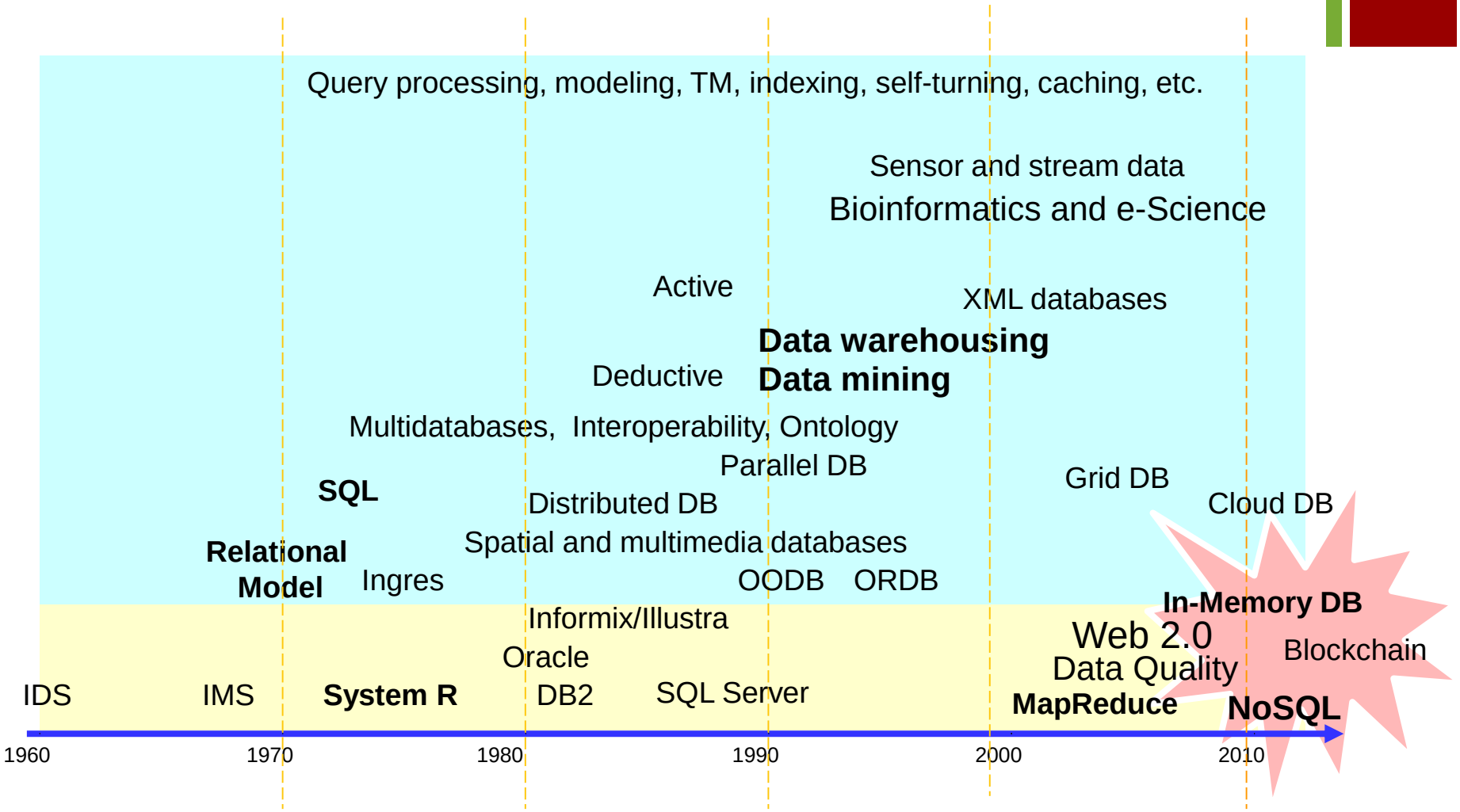
+ (Once) Advantages of RDBMS

- Separation of data from applications
- Push-down common functions (general-purpose systems!)
- Separation of physical structures and logical structures
- Relational model and theory
- Non-procedural query language
- Concurrency control and recovery
- High performance query processing

+ Limitations of RDBMS

- (*to be filled in classroom*)

+ A Brief History of DBMS



+ Definitions of “Big Data”

- Definition 1: the data volume is too large for the current technology to process
- Definition 2: from causality to association

+ Characteristics of “Big Data”

+ Another V: Value

- Big Data Challenges
  - Volume
    - How do we process a large amount of data?
  - Variety
    - How to handle a large number of types of data?
  - Velocity
    - How to deal with fast changing data and applications?
  - Veracity
    - How accurate and applicable is the data?
  - Value
    - How to derive value from data to support decision making?

+ Flu Prediction

- "Detecting Influenza Epidemics using Search Engine Query Data", *Nature* 2009
- 50M common search terms
- 2003-2008 CDC seasonal flu spreading trends
  - Training data: 2003-2006; Test data: 2007–2008
- 450M combinations ➔ 45 search terms with strong association
- Validation： successful predication of 2009 bird flu outbreak
- Deployed in 29 countries by WHO

+ Obama's 2012 Election

- "*The first big data election*"
- Started to collect data since 2008 (using cookies): 80 different types of information
- Data integration: personal, polls, donations, volunteers, consumers, social media…
- Using 66000 machines to process every night
- Personalised donation invitations, precisely targeted ads, state-based strategies for swing voters
- Precise outcome prediction from states to states

+ Election prediction from DS group

Queensland Professor claims algorithm will accurately predict US election results

NOVEMBER 4, 2016 8:25PM

Donald Trump

Victoria Craw news.com.au

A QUEENSLAND Univer
Hillary Clinton in the US
media data.

RIGHT NOW IN TECHNOLOGY

1184 READERS Netflix: Marvel's Iron Fist, Love S2 S2 on Stan | The Real ...

792 Sydney, Perth, Melbourne, Brisbane,

HOME NATIONAL INTERNATIONAL POLITICS BUSINESS SPORTS ENTERTAINMENT TECH OPINI

HOME » INTERNATIONAL / US PRESIDENTIAL ELECTION 2016

Big data expert in Queensland Uty predicts US election result

NARADA DESK | Updated On : November 5, 2016 3:52 pm

Professor Li has reportedly predicted the 2013 Australian federal election and the 2015 Queensland election with 95 per cent accuracy

FINANCIAL REVIEW

NEWS BUSINESS MARKETS STREET TALK REAL ESTATE OPINION TECHNOLOGY

Dec 23 2015 at 12:15 AM | Updated Dec 23 2015 at 9:40 AM

Australian computer scientist makes an algorithm that predicts elections

Professor Xue Li tested the search engine on the 2013 federal election, and it predicted the number of seats with 95 per cent accuracy. **Louie Douvis**

https://shorthand.uq.edu.au/changemakers/issue2/how-big-data-is-drawing-a-global-picture/

+ What to Cover in This Course

Distributed DB

- Design
- Query Processing
- Transaction Management

Data Preprocessing

- Integration
- Quality Management

Fundamental Issues

- Data Privacy
- Data Security

Data Warehouse

- Design
- Implementation

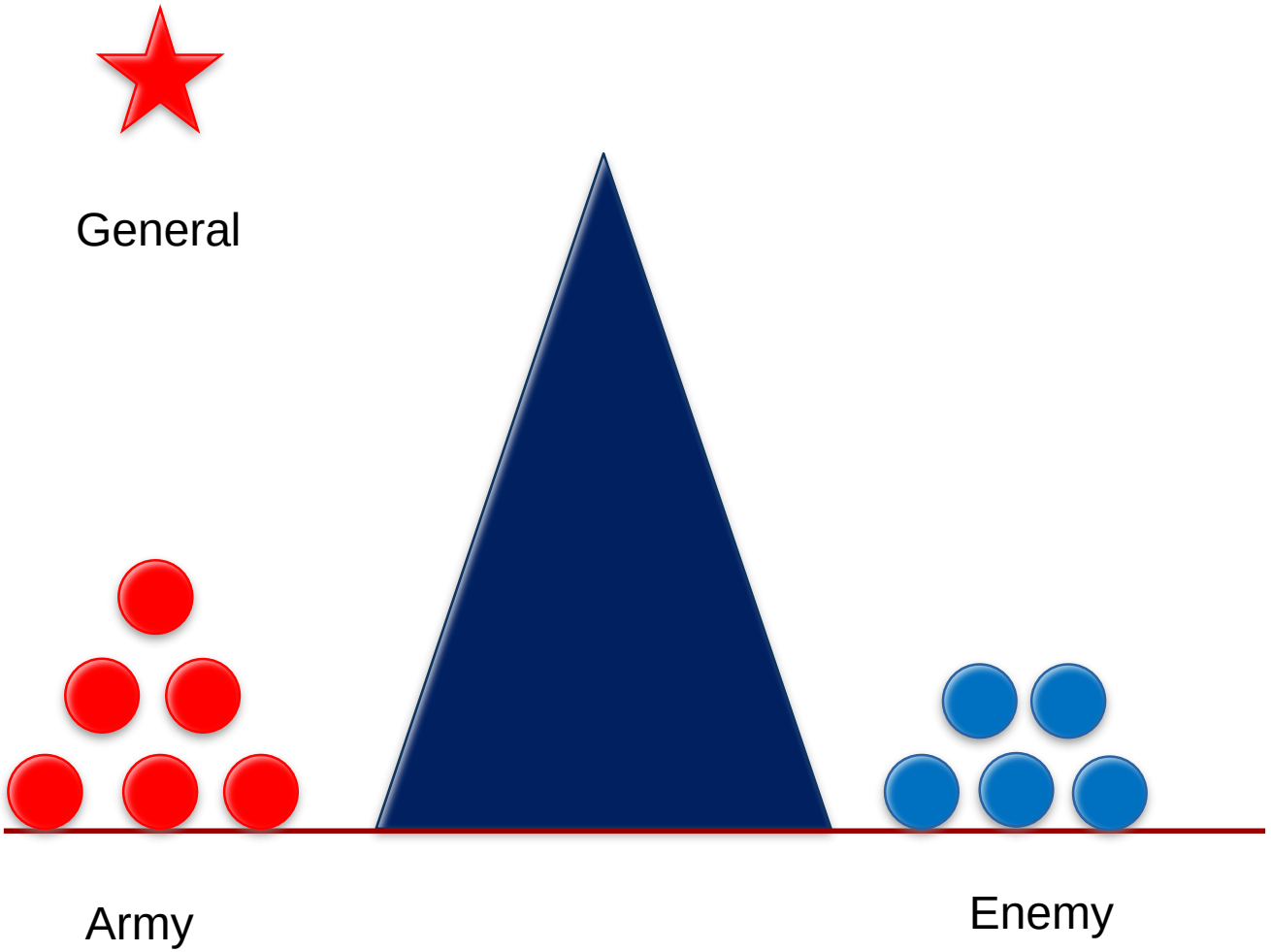
Some Real Systems

- Modern Computing Platforms
- Blockchain

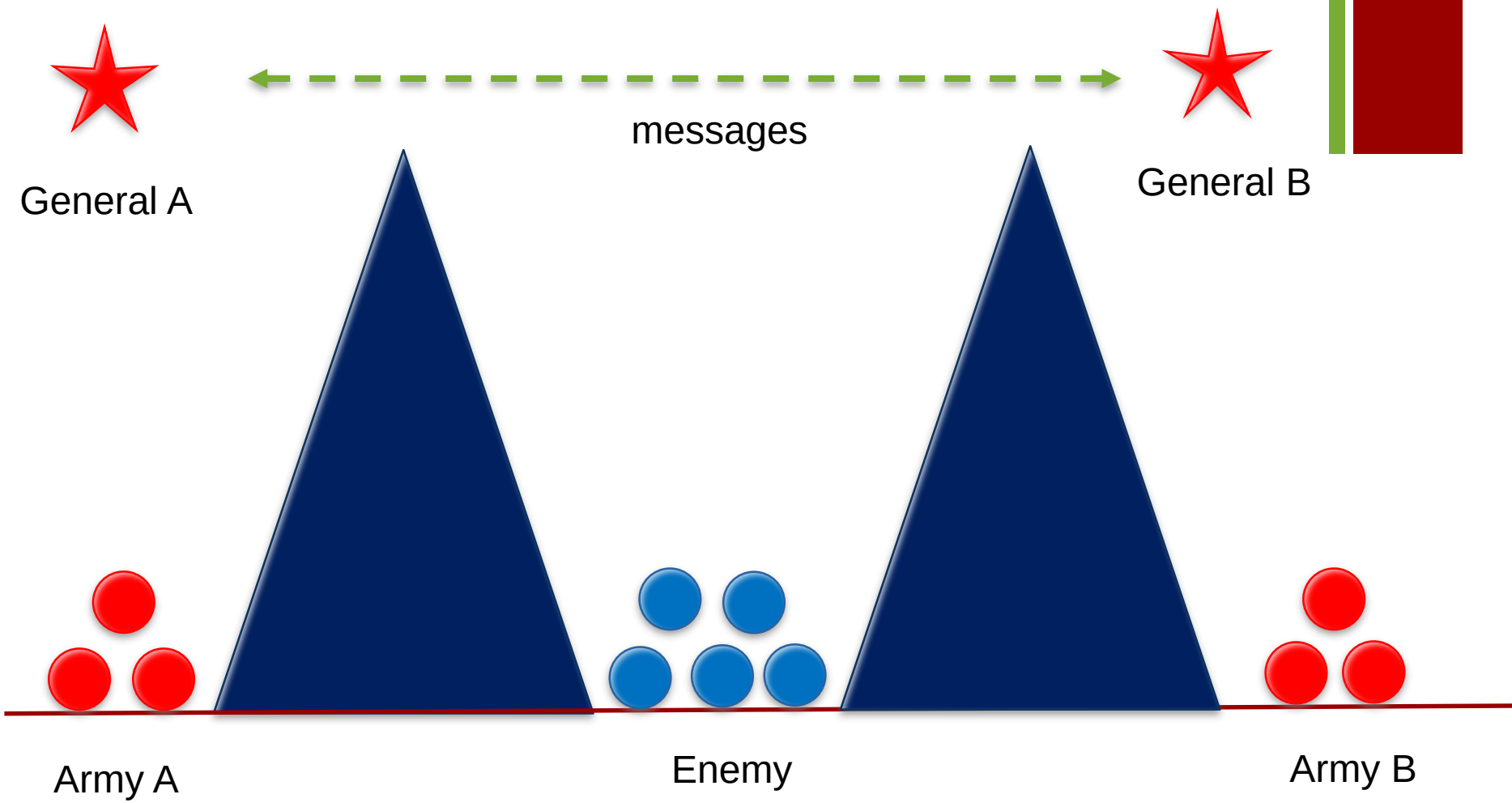
+ A Taste of DDB…

- Multiple processors
  - Opportunities for parallelism
  - Opportunities for reliability
  - Challenges with synchronization
- Heterogeneous and autonomous components

+ One General's Problem



+ Two Generals' Problem



+ A Taste of Data Warehousing

- Consider a supermarket database

Day	Product	Store	Value
9/2/2017	Milk	Toowong	3412
10/2/2017	Milk	Toowong	2918
9/2/2017	Bread	Toowong	2918
10/2/2017	Bread	Toowong	3445
9/2/2017	Milk	Kenmore	5440
10/2/2017	Milk	Kenmore	4992
9/2/2017	Bread	Kenmore	2918
10/2/2017	Bread	Kenmore	3067

Sales

Store	District	Region
Toowong	West	Brisbane
Kenmore	West	Brisbane

Stores

Product	Kind	Type	Class
Milk	Dairy	Perishable	Food
Bread	Bakery	Perishable	Food

Products

+ Normalisation and Redundancy

- Two key features of RDBMS
  - Normalised database design
  - Removal of redundancy
- Data analytics queries
  - How much milk sold at Toowong shop on 9/2/2017?
  - How much dairy products sold in West Brisbane in the 1st quarter of 2017?
  - Dimensions
    - Product: *milk*, dairy, perishable…
    - Store: *Toowong*, West, Brisbane…
    - Day: *9/2/2-17*, week, month, quarter, year…

+ A Taste of Data Integration, Quality, and Privacy

	gender	GPA	age	postcode	program
0	F	6.24	28	4106	Bachelor Business Management/Arts
1	F	2.65	20	4067	Bachelor Business Management/Science
2	M	5.48	30	4065	Bachelor Engineering /Arts
3	M	4.87	32	4066	Bachelor Business Management/Arts
4	M	3.68	22	4104	Bachelor Science - St Lucia
5	F	3.49	48	4104	Bachelor Information Technology/Arts
6	M	2.56	29	4104	Bachelor Mathematics/Science
7	M	5.06	41	4105	Bachelor Business Management/Science
8	F	5.84	41	4105	Bachelor Arts/Social Science
9	M	5.01	26	4064	Bachelor Arts/International Hotel and Tourism ...

Anonymous student data

Sporting club membership

	first_name	last_name	gender	age	postcode	is_student	member_id
1	Rose	Hayes	F	20	4067	True	25646
2	John	Wheeler	M	41	4105	True	69447
3	Mildred	Hicks	F	41	4105	True	98351
4	Ronald	Fisher	M	26	4064	True	90919
5	Karen	Mitchell	F	26	4103	True	77890
6	Jack	Patterson	M	57	4102	True	39891
7	Donald	Matthews	M	21	4067	True	38655
8	Thomas	Montgomery	M	19	4101	True	85481
9	Jonathan	Harris	M	47	4000	True	12279

+ Summary

- This is an advanced course for data management
  - Based on your RDBMS knowledge
  - Distributed Databases: distribution and replication
  - Data warehousing: support data analytics queries
  - Database security and privacy issues
- Course website
  - Accessible through UQ Blackboard system
  - Students should check this site at least once a week
- Help us to make this course better!
  - Your feedback and suggestions are very important to us! Feel free to talk to us.
- Next: Distributed Database Design