INFS4205/7205 Advanced Techniques for High Dimensional Data

Course Review

Semester 1, 2021

University of Queensland

+ INFS 4205/7205

- Learning objectives
  - The foundation and the latest in large-scale data management and processing: **complex data types** and **similarity-based queries**
    - How to represent and manage complex data to enable efficient processing of advanced queries
    - Data type: 1D →2D → 3D → 2D+Time → HD
      - Relations, Operations, Indexes, Queries
        - Hash, Hash + Tree, Tree
        - Space Centric, Object Centric
        - Filter and Refine
  - Applicable to a wide range of **applications**, including spatial, multimedia and other types of data

What We have Covered

1. Spatial data types and spatial databases
2. Spatial indexing mechanisms
3. Spatial algorithms and query processing
4. Spatiotemporal data management
5. High-dimensional indexing and search
6. Multimedia databases
7. Route Planning in Road Network

+ Spatial Databases

- Key topics
  - Spatial data types and modelling
  - Spatial relationships, operations and queries
  - Spatial indexing and query processing
  - SDBMS architectures
  - Advanced queries (kNN and skyline)
- Goals
  - Understand how spatial data is different from the relational data
  - Understand how these differences affect those relational techniques we learned before
  - Understand what spatial DBMS is

+ Spatial Data Types

- Spatial data
- Spatial data types
  - RDBSM data types: numbers, strings and dates (BLOBs?)
  - SDBMS data types: points, lines and polygons
    - Usages are very similar compared with relational data types
    - Indexing and processing are quite different
- Spatial data types can help us to understand how other data types should be managed and processed
  - Such as multimedia data
  - Feature extraction and embedding can transform many type of data into vector data

+ Spatial Data

- Spatial Data Relations
  - Topological, Direction, Metric
- Spatial Data Operations
  - Basic Spatial Operations
    - Selection, Projection, Amalgamation
    - Containment, Region, Intersection, Overlay, Fusion, Windowing, Clipping, Centre, Boundary
    - Area, Perimeter, Distance
  - Other Spatial Operations
    - Nearest Neighbours, Similarity Search, Skyline Queries
  - Complex Spatial Queries
    - Multiple predicates/operations, spatial sub-queries

+ Spatial Queries

- Normal use of (spatial) data types, operations and constants

(1) Understand SQL
(2) Define non-SQL components

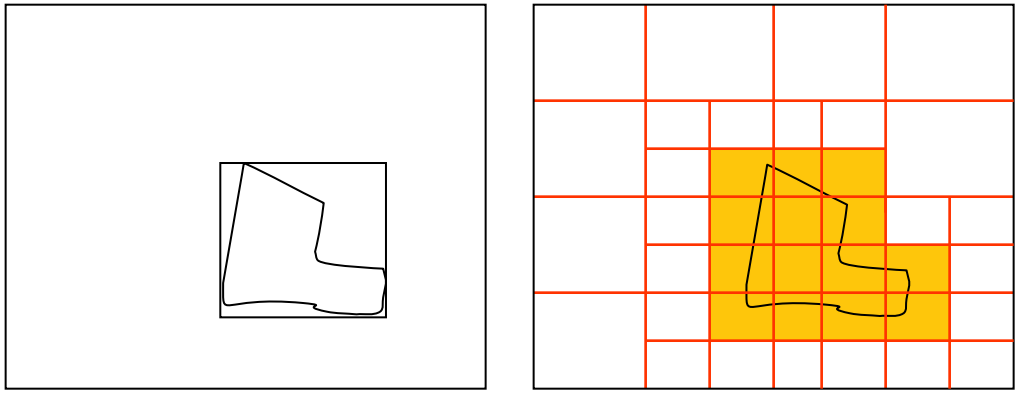
```
SELECT  river.name, road.name,
          intersection(river.route, road.route)
FROM    river, road
WHERE   river.route intersect road.route;

SELECT  river.name,
          length(intersection(river.route, Queensland))
FROM    river
WHERE   river.route intersect Queensland;
```

Spatial Indexing

- Purpose:
  - Efficiency in processing spatial selection, join and other spatial operations
- Two strategies to organize space and objects
  - Map spatial objects into 1D space and use a standard index structure (B-tree)
  - Dedicated external data structures
- Basic ideas
  - Approximation
    - Bounding box, Grids
  - Hierarchical Data Organization

Object Approximation

- A fundamental idea of spatial indexing is the use of approximation
- Continuous Approximation
  - Object centric
  - Example:
    - Use of MBRs (Minimum Bounding Rectangles)
    - R-Tree
- Grid Approximation
  - Space centric
    - Faster mapping
    - Uniform / Non-uniform
    - High-D?
  - Example:
    - Quadtree



Data Access Methods

- One dimensional
  - Hashing and B-Trees
- Line Data
  - Segment Tree, Interval Tree
- Point data
  - Hashing: GRID and EXCELL
  - Hierarchical
    - Quadtree: point and region quadtrees
    - kd-Tree
    - Z-values and B-tree
- Polygon data
  - Transformation: End point mapping and z-values
  - Overlapping: R-tree and R*-tree
  - Clipping: R^+-tree

(1) Understand how an indexing structure indexes data (and how they are different form other methods)

(2) Understand how an index can support query processing (for point and range queries, and for join and other advanced queries)

+ Region Quad-tree

- The order of data insertion is not important
- Data distribution is still important
  - Region quad-tree could be unbalanced

- Main differences between point and region quad-trees?
  - Region quad-tree: regular decomposition
    - No need to store x & y coordinates in internal nodes
    - Hard to be balanced
  - Point quad-tree: irregular decomposition (governed by input)
    - Some data can be stored at internal nodes
    - Can be balanced by sorting
    - How to insert / delete?
- They both work in high dimensional spaces

+ Transformation: Using Z-Ordering

- (0,0), (0,1), (1,0), (1,1)

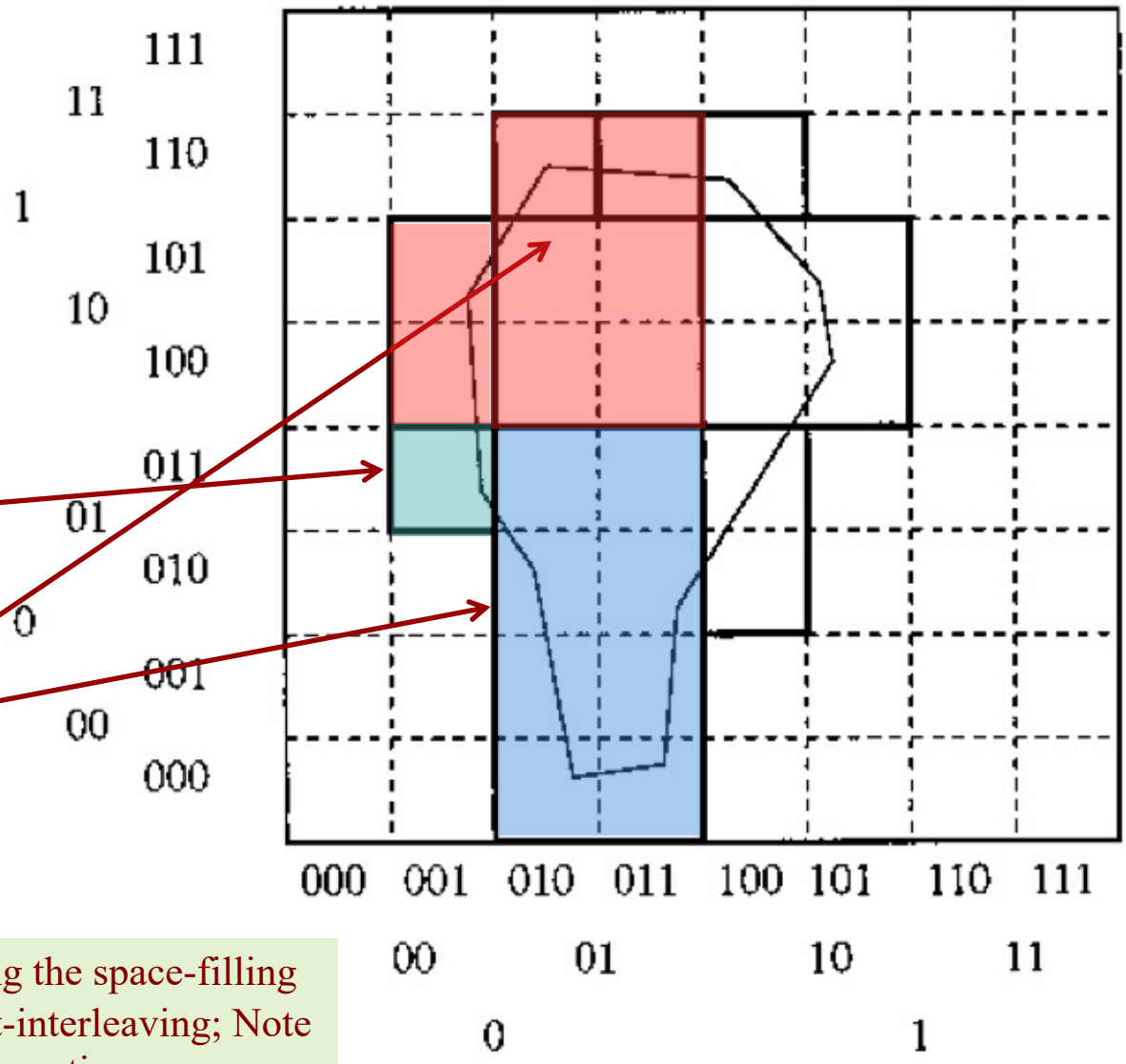
- **Further decompose**

(0,0) -> (00, 01)

-> (001,011)

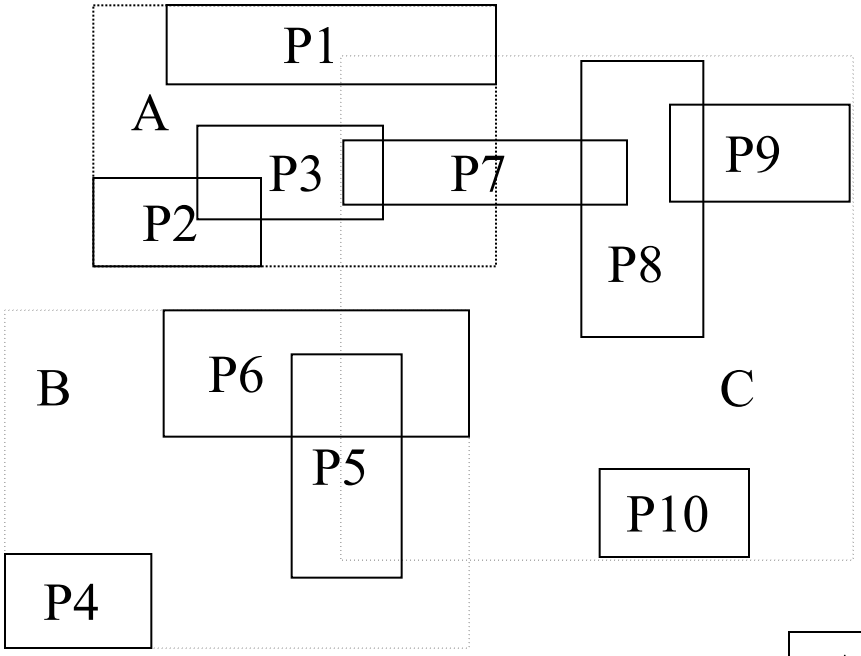
(01,00), (01,01)

(0,1) …



Multiple ways of mapping: following the space-filling curse, recursive decondition, and bit-interleaving; Note about base-5 number, use of 1 for the entire space

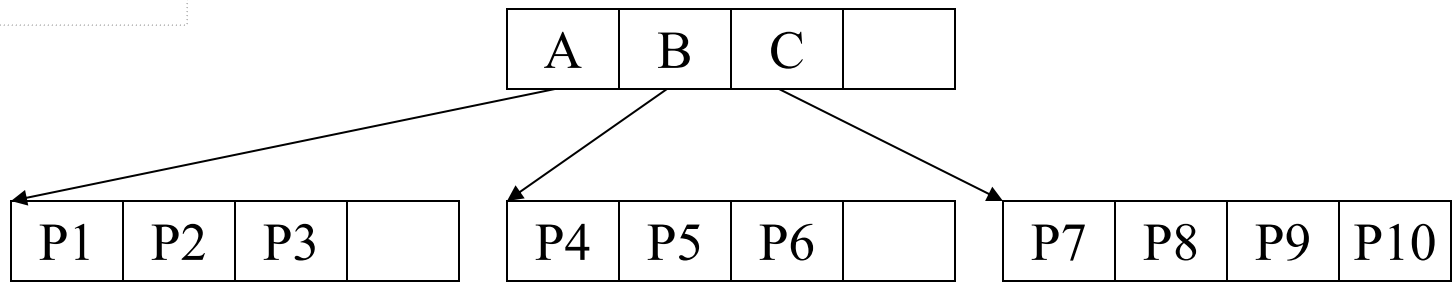
+ R-Tree



Must allow **overlapping** cells

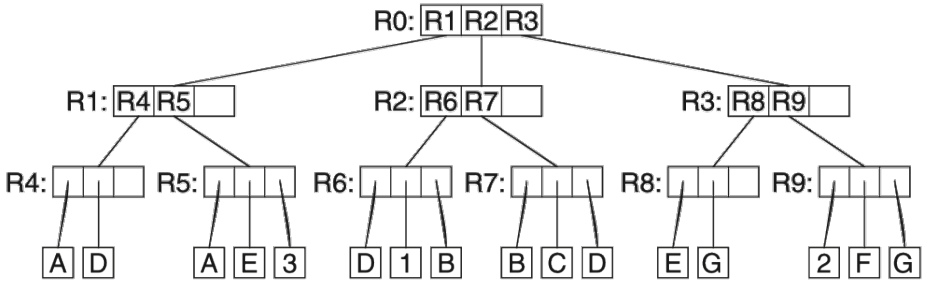
multiple cells need to be examined to **search** for an object
(e.g., find object P7)

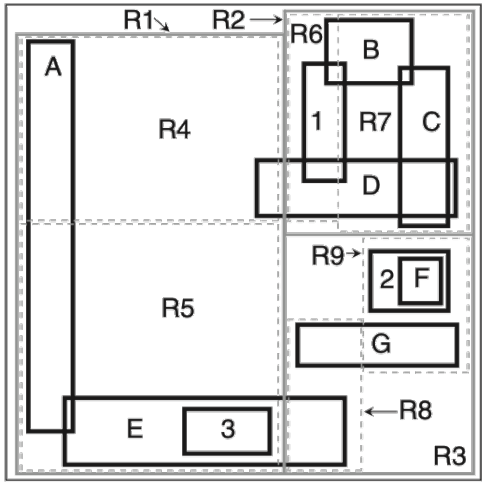
Insert only once
where to insert? How to split?
Multiple methods including R* tree



+ R+-tree

- How R+-tree differs from R-tree?
- Relative advantages and disadvantages?
- Search and update strategies?
- Suitability?





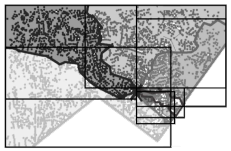
- Note: again, R-tree and R+-tree work for high dimensional spaces

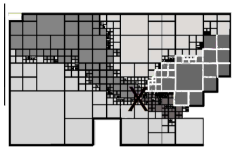
+ Query Execution Plan

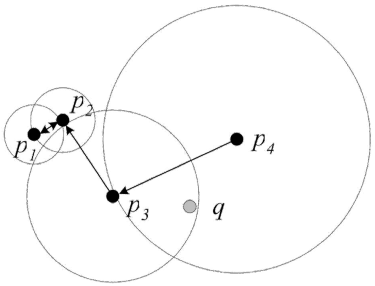
- Define the schema and indexes first
- Describe the procedure of using indexes step-by-step
- Follow the filter-and-refine approach
- Be able to explain the purpose of each step and savings made
- Many simple examples in the lecture notes
  - Polygon Intersection
  - More complex example related to kNN and Skyline query processing, and in research papers

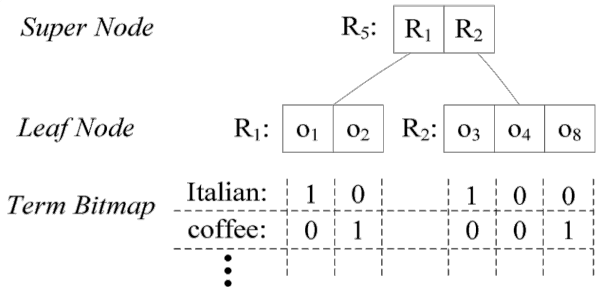
+ kNN and Skyline Query Processing

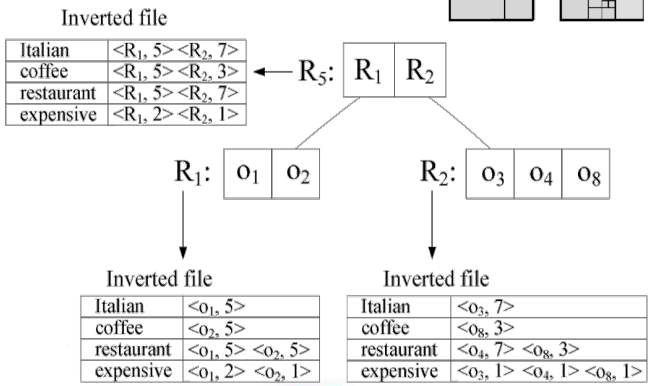
- Goals
  - Understand how spatial indexing structures can be used to support complex queries
  - Enhance the understanding of spatial databases from in-depth knowledge of advanced spatial processing
  - Provide a brief view of the frontier of spatial database research

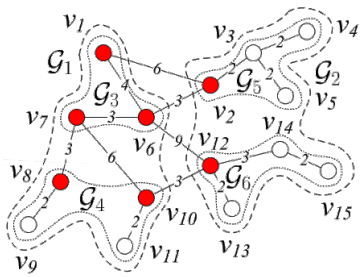


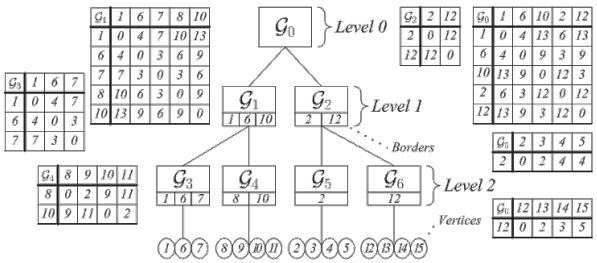






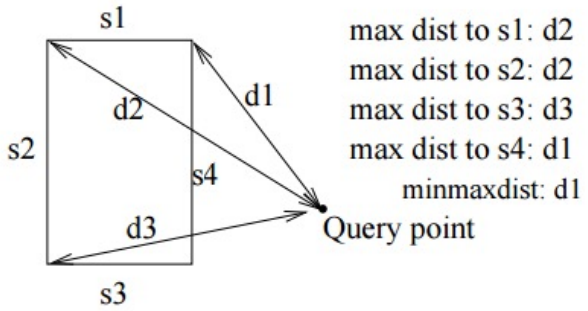






Pruning Strategies

- MinDist(p, R) ≤ NN(p) ≤ MinMaxDist(p, R)



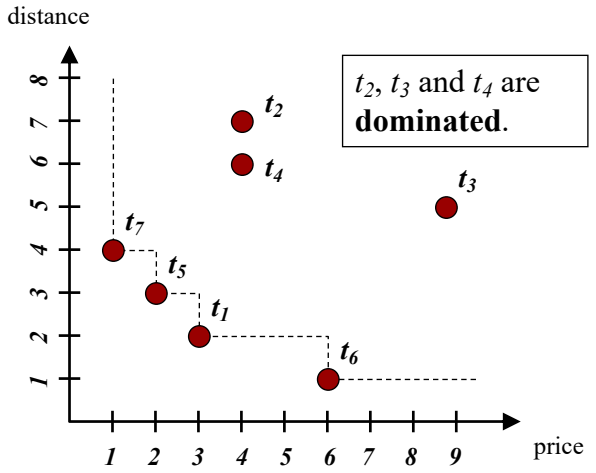
- Strategy:
  - **(downward pruning)** An MBR R is discarded if there exists another R' such that MINDIST(p, R)>MINMAXDIST(p, R')
  - **(downward pruning)** An object o is discarded if there exists an R such that $d(p, o)$ > MINMAXDIST(p, R)
  - **(upward pruning)** An MBR R is discarded if a point q is found such that MINDIST(p, R)> $d(p, q)$

kNN Query Processing

- What is kNN query?
- Naïve approach?
  - And problems?
- How this can be improved by using an R-tree?

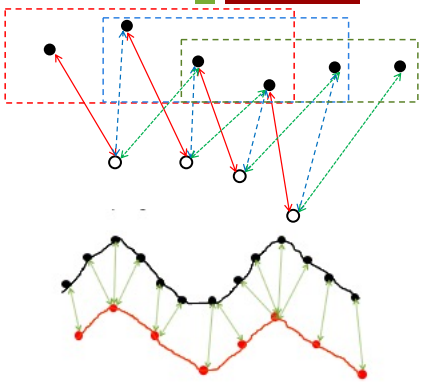
Skyline Query Processing

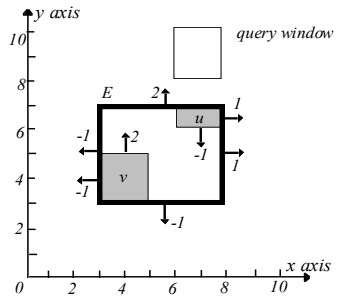
- What is the skyline query?
  - How to define a skyline query?
- Why people are interested in that?
- Basic Algorithms
  - Naïve approach
  - NN based approach
  - The state-of-the-art: BBS

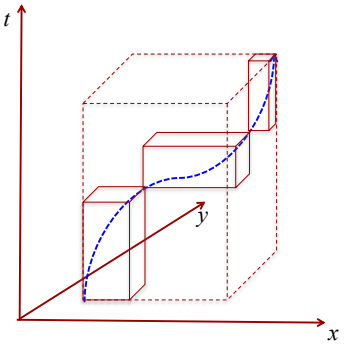


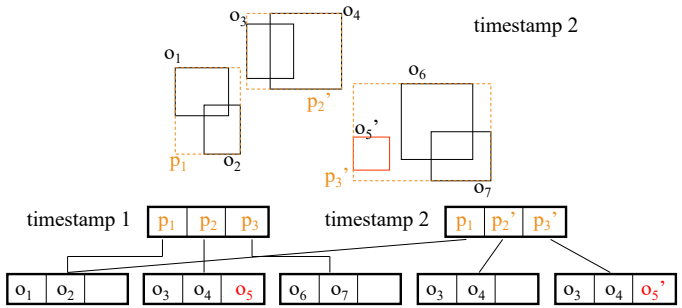
+ Spatiotemporal Databases

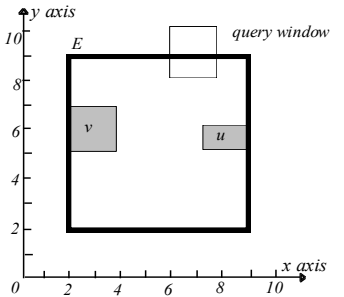
- Definitions: moving objects and trajectories
- Spatiotemporal queries
  - Spatiotemporal point and window queries
  - Trajectory similarity queries – and similarity measures
    - Lock-Step Windows, LCSS, EDR, DTW,…
- Spatiotemporal indexing methods
  - Simple ways of indexing spatiotemporal and issues
    - 3D R-Tree
    - HR-Tree
    - TPR-Tree





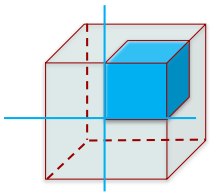


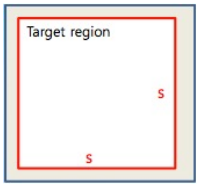


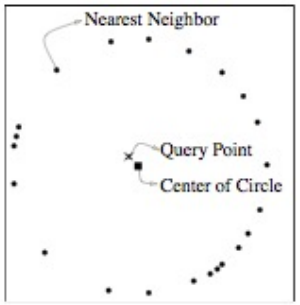


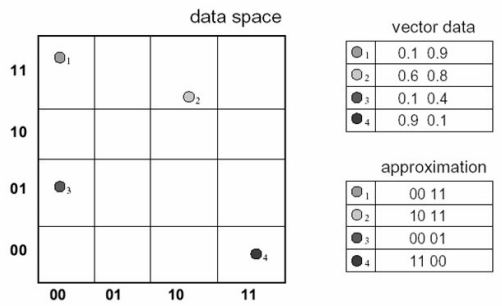
+ High Dimensional Search

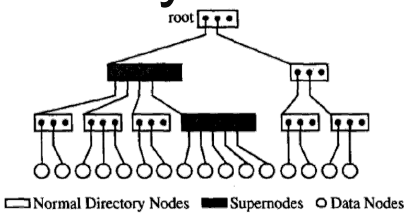
- In theory and in practice…
- Dimensionality Curse
  - Many interesting observations
    - Numbers of partitions, Central hollow / High overlapping, Nearest Neighbor is not near…
  - The performance degrades rapidly as dimensionality increases, and eventually underperforms linear scan
  - However, linear scan needs to search the whole data file - affected volume is 100%
- Example Indexes: Why do they still exist?
  - R-Tree -> X-Tree
  - Z-Order -> Pyramid
  - Grid -> VA-File
  - Cluster -> iDistance

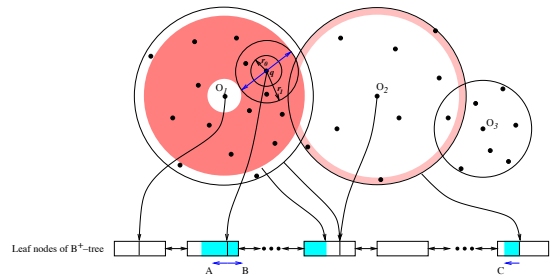






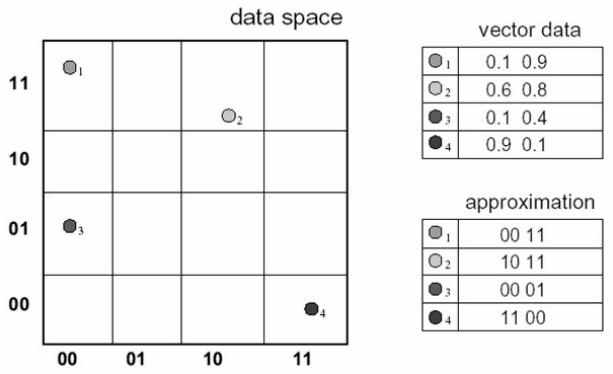






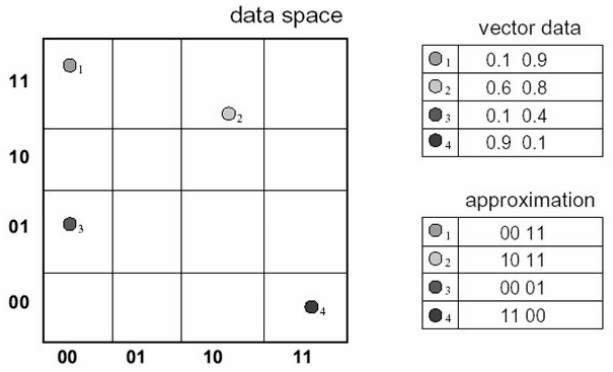
+ Vector Approximation (VA) File

- In high-D spaces, all tree-based indexing structures examine large fraction of leave nodes
  - Since the space of nodes are heavily overlapped
- Better to scan the whole data set and avoid performing seeks altogether
- Natural question: how to speed-up linear scan?
  - Approximation to compress vector data
    - Reduces the amount of data to be read during search



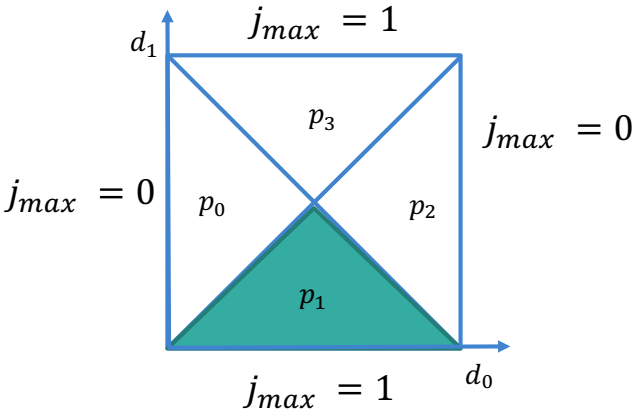
+ Vector Approximation (VA) File

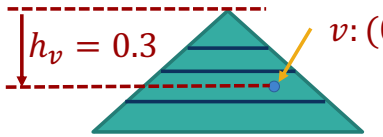
- In high-D spaces, all tree-based indexing structures examine large fraction of leave nodes
  - Since the space of nodes are heavily overlapped
- Better to scan the whole data set and avoid performing seeks altogether
- Natural question: how to speed-up linear scan?
  - Approximation to compress vector data
    - Reduces the amount of data to be read during search



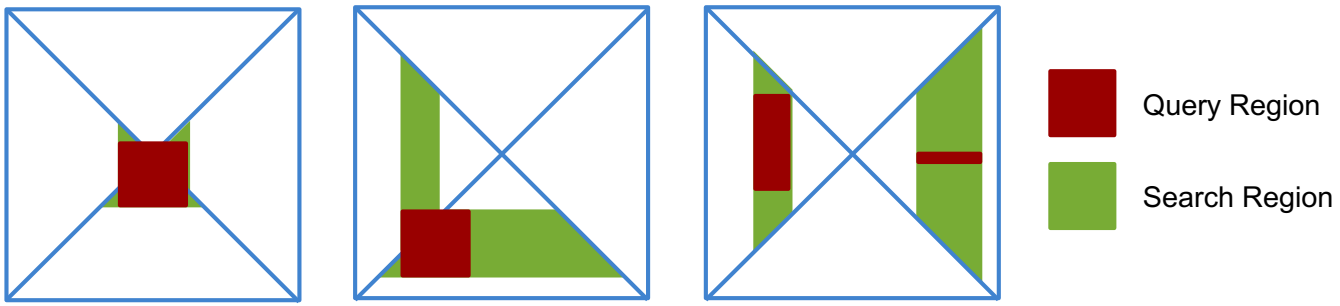
+ Pyramid Technique

- Divide the data space such that the resulting partitions are shaped like peels of an onion
  - Index by B-Tree
  - Pyramid Value = Pyramid Numbering + Point Height



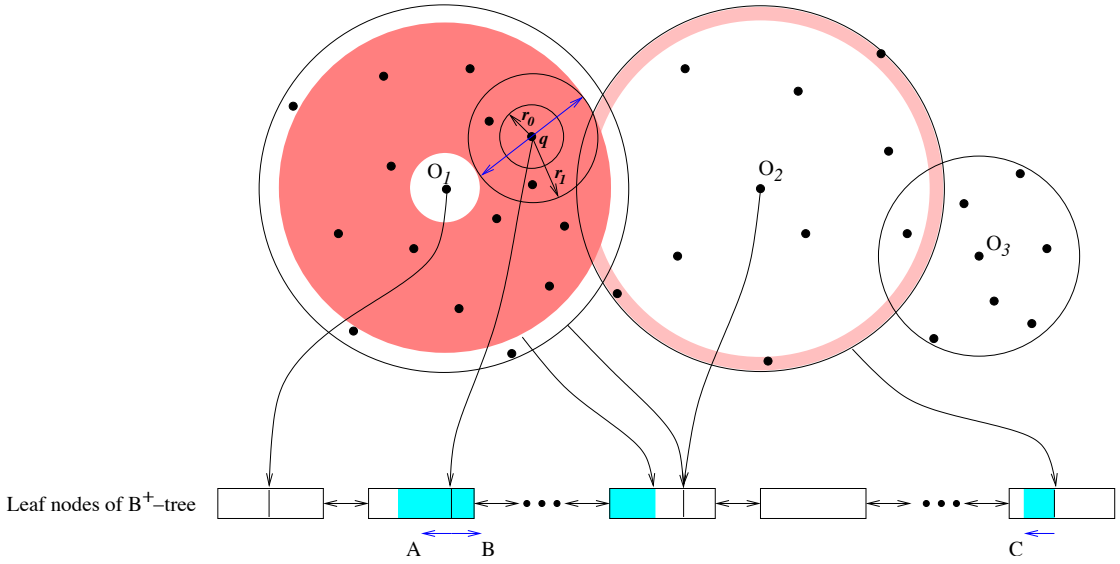


$pv_v = 1 + 0.3 = 1.3$



+ iDistance

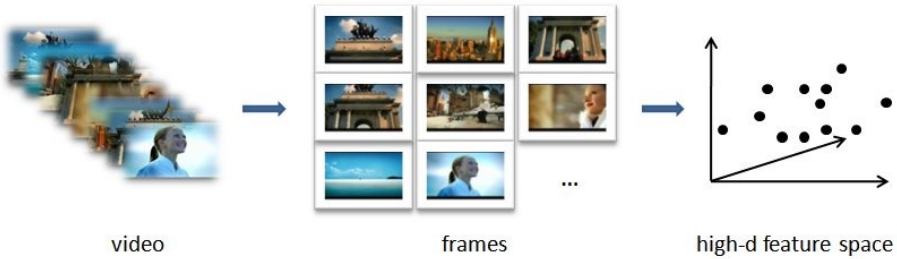
- iDistance is simple and efficient for clustered data
  - It is a distance and partition based index
  - It can be used for approximate search
  - The index can be integrated to existing systems easily
    - iDistance -> B-Tree

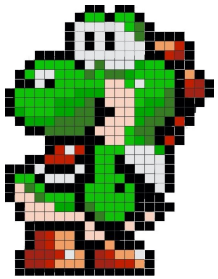


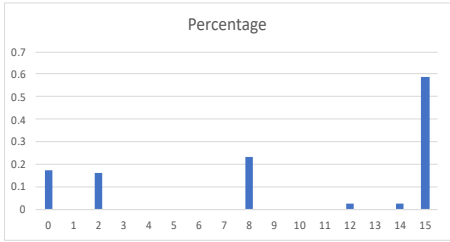
$$[dist(O_i, q) - querydist(q),$$
$$\min(dist\_max_i, dist(O_i, q) + querydist(q)]$$

+ MMDB

- Data representations
  - Image: basic format, abstraction and features (colour, texture and shape)
  - Video: structure

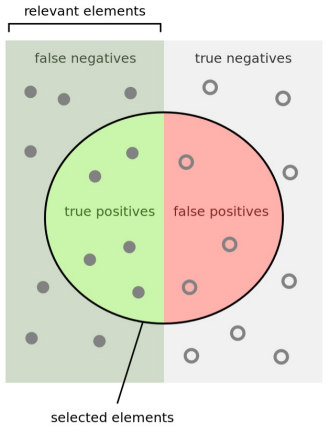






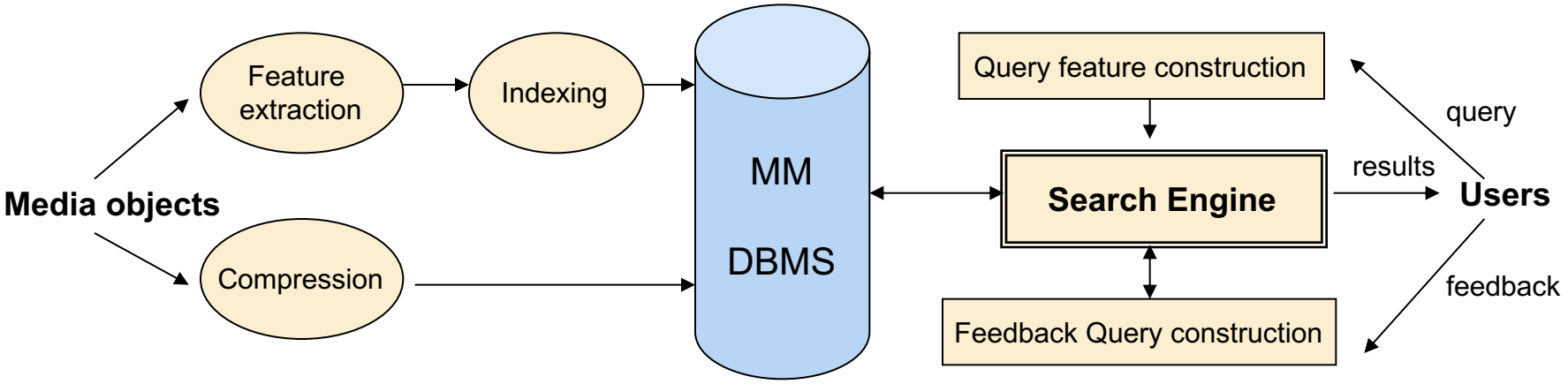
- Text-based vs Content-based search

- Similarity measures
  - P-norm, ViTri, mean distance and DTW
  - Query results evaluation: Precision and Recall



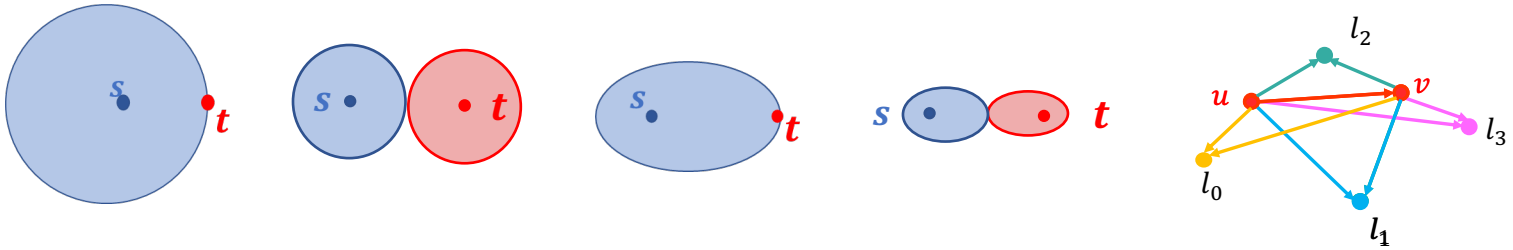
+ MMDB

- General knowledge:
  - MMDB architecture
  - The process of multimedia search
  - Scalability issues, and how do we we deal with it?

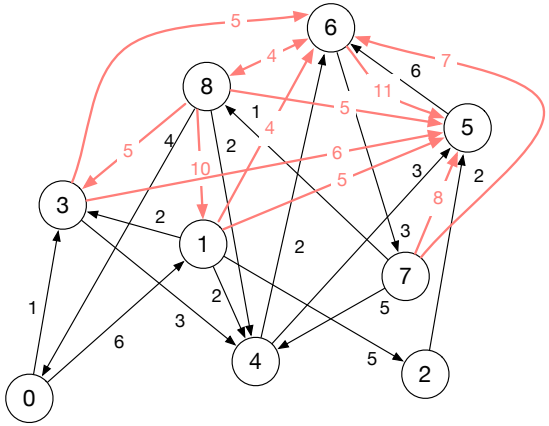


+ Route Planning

- Index Free
  - Dijkstra's, Bi-Dijkstra's, A*, Bi-A*, Landmark



- Index Based
  - Contraction Hierarchy, 2 Hop Labeling



Node	Out-Label	In-Label
0	(4,4) (1,6) (3,1)	(4,10) (5,14) (6,8)
1	(4,2)	(4,16)
2	(4,14) (1,22) (5,2)	(4,21) (1,5)
3	(4,3)	(4,11) (1,2) (5,15) (6,9)
4		
5	(4,12) (1,20)	(4,3)
6	(4,6) (1,14)	(4,2) (5,6)
7	(4,3) (1,11) (3,6) (0,5)	(4,5) (5,9) (6,3)
8	(4,2) (1,10) (3,5) (0,4)	(4,6) (5,10) (6,4) (7,1)

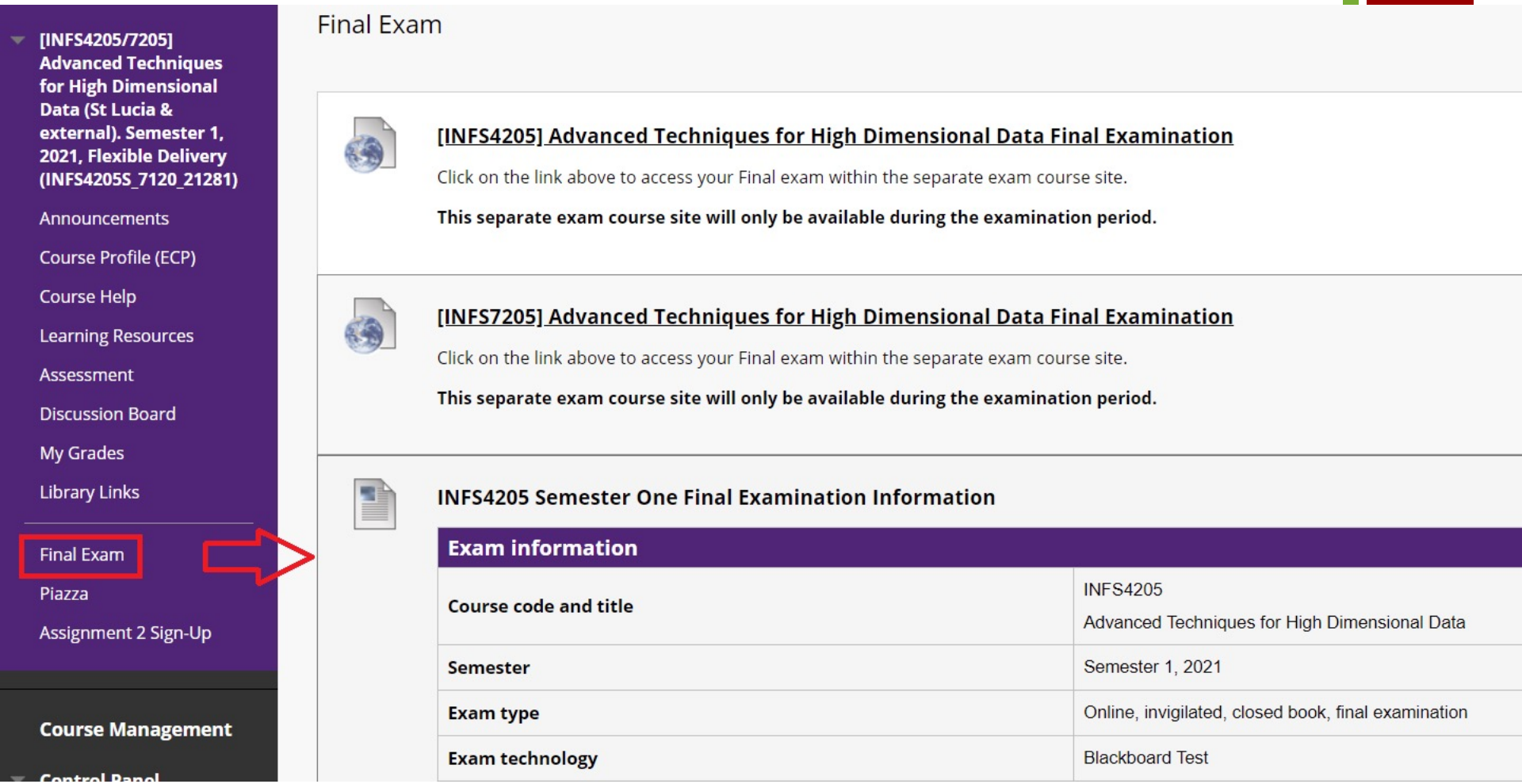
+ Trends and Open Issues

- Social media
  - With text, rich link structures, location and multimedia information
- Trajectory data
  - Moving objects, spatiotemporal data…
- Semantics and tagging
- Integration and value

Prepare for the Exam

- Exam scope and format
  - Blackboard Test with ProctorU
  - Entire course, excluding Week 12 Lecture
  - Close-book, 120 + 10 minutes, 50 marks
  - “Double Pass”
    - At least 50 marks in total and at least 25 marks in the final exam, to pass this course
- Consultations with the tutors and Q&A
  - Tutorial sessions in week 13

+ Final Exam


[INFS4205/7205] Advanced Techniques for High Dimensional Data (St Lucia & external). Semester 1, 2021, Flexible Delivery (INFS4205S_7120_21281)
Announcements
Course Profile (ECP)
Course Help
Learning Resources
Assessment
Discussion Board
My Grades
Library Links
Final Exam
Piazza
Assignment 2 Sign-Up
Course Management
Final Exam
[INFS4205] Advanced Techniques for High Dimensional Data Final Examination
Click on the link above to access your Final exam within the separate exam course site.
This separate exam course site will only be available during the examination period.
[INFS7205] Advanced Techniques for High Dimensional Data Final Examination
Click on the link above to access your Final exam within the separate exam course site.
This separate exam course site will only be available during the examination period.
INFS4205 Semester One Final Examination Information
Exam information
Course code and title
INFS4205
Advanced Techniques for High Dimensional Data
Semester
Semester 1, 2021
Exam type
Online, invigilated, closed book, final examination
Exam technology
Blackboard Test

+ Exam Information

Exam information	
Course code and title	INFS4205 Advanced Techniques for High Dimensional Data
Semester	Semester 1, 2021
Exam type	Online, invigilated, closed book, final examination
Exam technology	Blackboard Test
Exam date and time	Refer to your personal exam timetable for the scheduled date and time of this exam. The examination duration from the time your proctor starts your examination will be: 2 hours 10 minutes including 10 minutes reading time..
Exam window	At the time selected by you within the scheduled exam window, you will be required to connect to ProctorU. After an on-boarding process, your proctor will start the exam. You will then have the duration of the exam listed to complete your exam. **Please note**: It may take up to 30 minutes to connect you with a proctor, however your exam timer does not start until your proctor starts your exam. After an exam is started, you will have the duration of the exam to complete and submit your response. **Please note:** You will not be able to access the examination after this time.
Permitted materials	The recommended materials are listed below. Any calculator permitted - unrestricted.
Recommended materials	Ensure the following materials are available during the exam: Water bottle; unrestricted number of blank working paper
Instructions	You need to answer all of the questions in the Blackboard Test.

+ Exam Information

Who to contact	Given the nature of this examination, responding to student queries and/or relaying corrections to exam content during the exam may not be feasible. At the end of the exam there will be a free text box field. Please use this to specify any assumptions you have made in completing the exam and which questions those assumptions relate to. You may also include queries you may have made with respect to a particular question, should you have been able to 'raise your hand' in an examination room. If you experience any technical difficulties when connected to an invigilator, talk to your online invigilator via the webcam or chat functions. If the technical trouble cannot be resolved, you should ask for an email (or transcript of the chat) documenting any technical advice provided to support your request for a deferred exam. If your invigilator advises you to contact UQ or you experience any technical difficulties when not connected to an invigilator, contact the Library AskUs service for advice as soon as practicable: **Chat:** support.my.uq.edu.au/app/chat/chat_launch_lib **Phone:** +61 7 3506 2615 **Email:** examsupport@library.uq.edu.au You should also ask for an email documenting the advice provided so you can provide this to the course coordinator immediately at: h.yin1@uq.edu.au
Important exam condition information	Academic integrity is a core value of the UQ community and as such high academic integrity expectations apply to all examinations, whether undertaken face-to-face or online. This means: • You are not permitted to access any online or hard copy resources during this closed book exam, and hence you cannot cut-and-paste material other than your own work as answers. • You are not permitted to consult any other person – whether directly, online, or through any other means – about any aspect of this examination during the period that it is available. • If it is found that you have given or sought outside assistance with this examination, then that will be deemed to be cheating. Undertaking this online examination deems your commitment to UQ's academic integrity pledge as summarised in the following declaration: *"I certify that I have completed this examination in an honest, fair and trustworthy manner, that my submitted answers are entirely my own work, and that I have neither given nor received any unauthorised assistance on this examination".*

+ Exam Information

- **Format**: Multiple-choice questions (MCQ) and short-answer questions. There are 5 MCQ questions, and each question is assigned 2 marks. There are 6 groups of short-answer questions with the total marks 40. Each group contains 2-4 subquestions.
- Repeated submissions **will not be allowed**. Once you submit your answers, you cannot go back to the exam.
- Please click **Save** once you finish each question in case of any unforeseen technical problem.
- This test will **save and submit automatically** when the time expires.
- This test **can be saved and resumed** at any point until time has expired. The timer will continue to run if you leave the test.

Thanks, and All the Best!