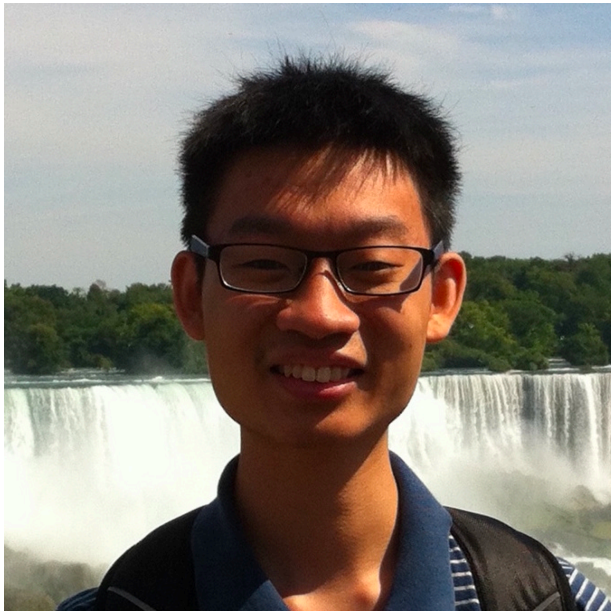
CS4225/CS5425 Big Data Systems for Data Science

Course Overview

Bryan Hooi
School of Computing
National University of Singapore
bhooi@comp.nus.edu.sg

About Bryan



- Office: COM2 #03-15
- Email: bhooi@comp.nus.edu.sg
- Office Hours:
  - Fridays 3 – 4pm or by appointment
- My research interests: graphs, robust machine learning, anomaly detection

Teaching Assistant

- Responsibility
  - Tutorials
  - Assist you in matters pertaining to the coding assignments
- We are fortunate to have the following great TAs.
  - Nicholas Lim, e0045287@u.nus.edu
  - Wang Yiwei, e0409763@u.nus.edu
  - Li Shen, e0474115@u.nus.edu

Assessment

- 2 assignments (50%)
- Week 13 test (50%) – held during lecture hours
- (No marks for attendance – it is fine to rely on video lectures)
- (Note: all in-lecture Zoom poll quizzes are ungraded)

Schedule

Week	Date	Topics	Tutorial	Due Dates
1	12 Aug	Overview and Introduction		
2	29 Aug	MapReduce - Introduction		
3	26 Aug	MapReduce and Relational Databases		
4	2 Sep	MapReduce and Data Mining	Tutorial: Hadoop	Assignment 1 released
5	9 Sep	NoSQL Overview 1		
6	16 Sep	NoSQL Overview 2	Tutorial: NoSQL	
Recess				
7	30 Sep	Apache Spark 1		Assignment 1 due, Assignment 2 released (3 Oct)
8	7 Oct	Apache Spark 2	Tutorial: Spark	
9	14 Oct	Large Graph Processing 1		
10	21 Oct	Large Graph Processing 2	Tutorial: Large Graph Processing	
11	28 Oct	Stream Processing		Assignment 2 due (31 Oct)
12	4 Nov	Deepavali – No Class		
13	11 Nov	**Test**		

Lecture

- Zoom (login with your NUS account)
  - Go to LumiNUS > Conferencing
  - Recorded lectures will be available on Conferencing > Previous
- Format:
  - We will divide each lecture into sessions of video of ~35 minutes, followed by discussion and Q&A.
    - You can type your question into **Chat** at any time during the video broadcast.
  - In-class zoom quizzes will be held during the breaks
  - Example:
    - 6:30-7:05pm, Part 1 video; 7:05-7:15pm, discuss and/or quiz
    - 7:15-7:50pm, Part 2 video; 7:50-8:05pm, discuss and/or quiz.

Lectures

- Reference textbooks
  - Jimmy Lin and Chris Dyer. 2010. Data-Intensive Text Processing with Mapreduce. Morgan and Claypool Publishers. https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf
  - Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. Mining of Massive Datasets (3rd ed.). Cambridge University Press. http://www.mmds.org/
- Study materials
  - Related chapters in the reference textbooks +
  - The related technical articles (for the state of the art)

Tutorials

- Starts from Week 4
- All tutorial questions will be available on the course website before the tutorial
- Recommended to attempt questions before tutorial
- Some questions are samples for tests

Coding Assignments

- Two coding assignments on Hadoop and Spark (50% total)
  - Analytics tasks
  - Sufficient materials are given on each analytics task.
- Submission to LumiNUS
  - Requirements for submission can be found in lab manuals
- Deadline
  - Assignment 1: Oct 3, 2021, Sun 11:59pm.
  - Assignment 2: Oct 31, 2021, Sun 11:59pm.
- Lab manuals and other supplement documents will be available in LumiNUS.

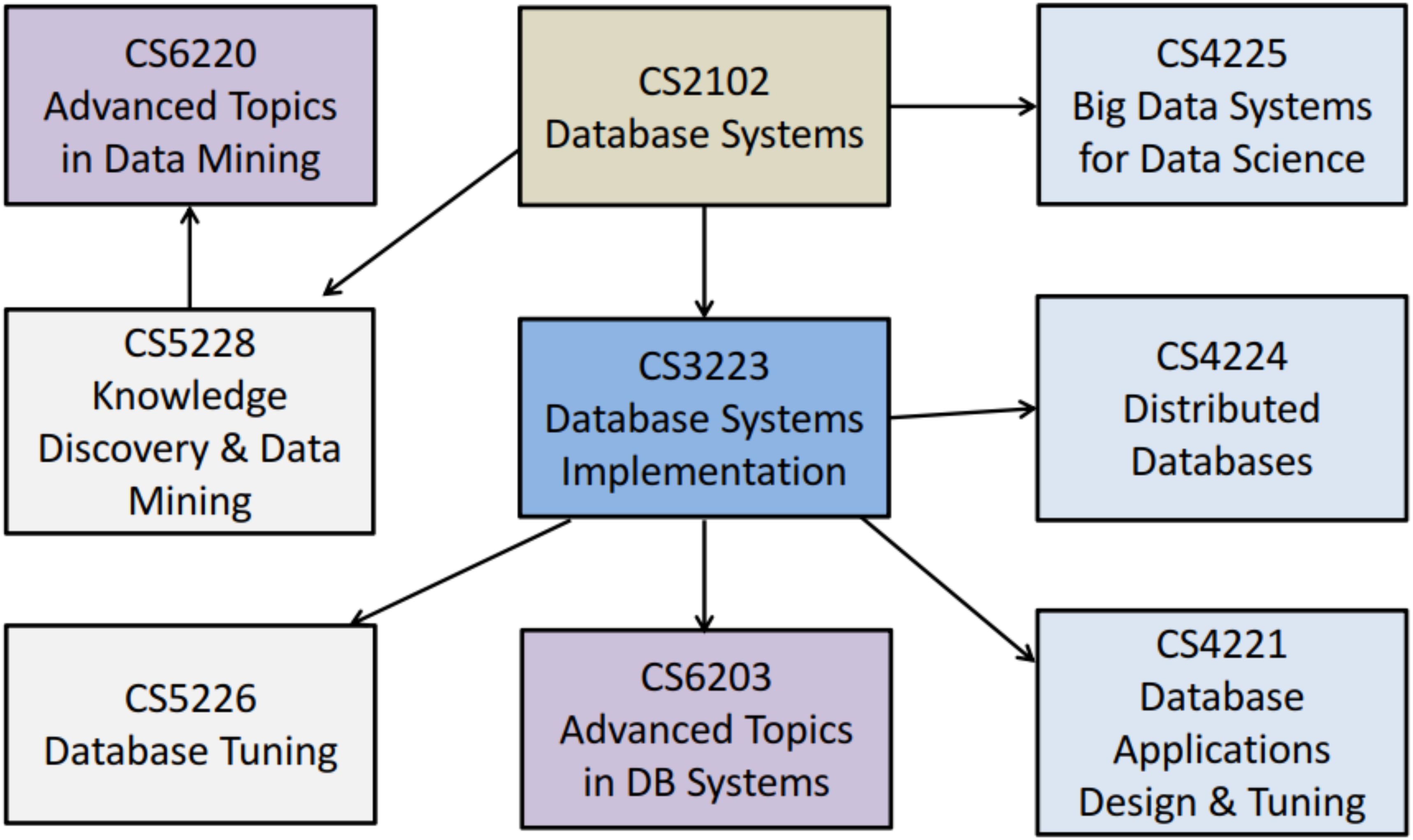
Coding Assignments (cont')

- **Individual** assignments
- Hadoop/Spark Resources
  - on your local machine
  - on computing cluster
- My expectations
  - Self-learning is important.
    - This course does *not* teach programming.
    - You're expected to pick up Hadoop/Spark with the provided materials and other online materials.

Test

- Test (50% in the final mark)
  - Date: Nov 11, 2021 (in the normal lecture hours)
  - **Open book & internet; on Zoom**
- Example questions
  - Integrative: Require you to combine knowledge from different chapters of the textbook
  - "Application": Require you to apply your knowledge of fundamental concepts to reasonably practical scenarios.
  - "Why not": Example, Tommy proposed a solution A to solve problem B in the lecture. Tell me what is the problem with solution A and how to overcome this problem
- Examples will be given during tutorial sessions

Database Courses @ SoC



CS6220
Advanced Topics
in Data Mining

CS2102
Database Systems

CS4225
Big Data Systems
for Data Science

CS5228
Knowledge
Discovery & Data
Mining

CS3223
Database Systems
Implementation

CS4224
Distributed
Databases

CS5226
Database Tuning

CS6203
Advanced Topics
in DB Systems

CS4221
Database
Applications
Design & Tuning

Relationships with Other Course

- This course has some overlaps with the following course
  - CS5344: Big Data Analytics Technology
- If you have already taken/or taking the above course, you should not take this course.

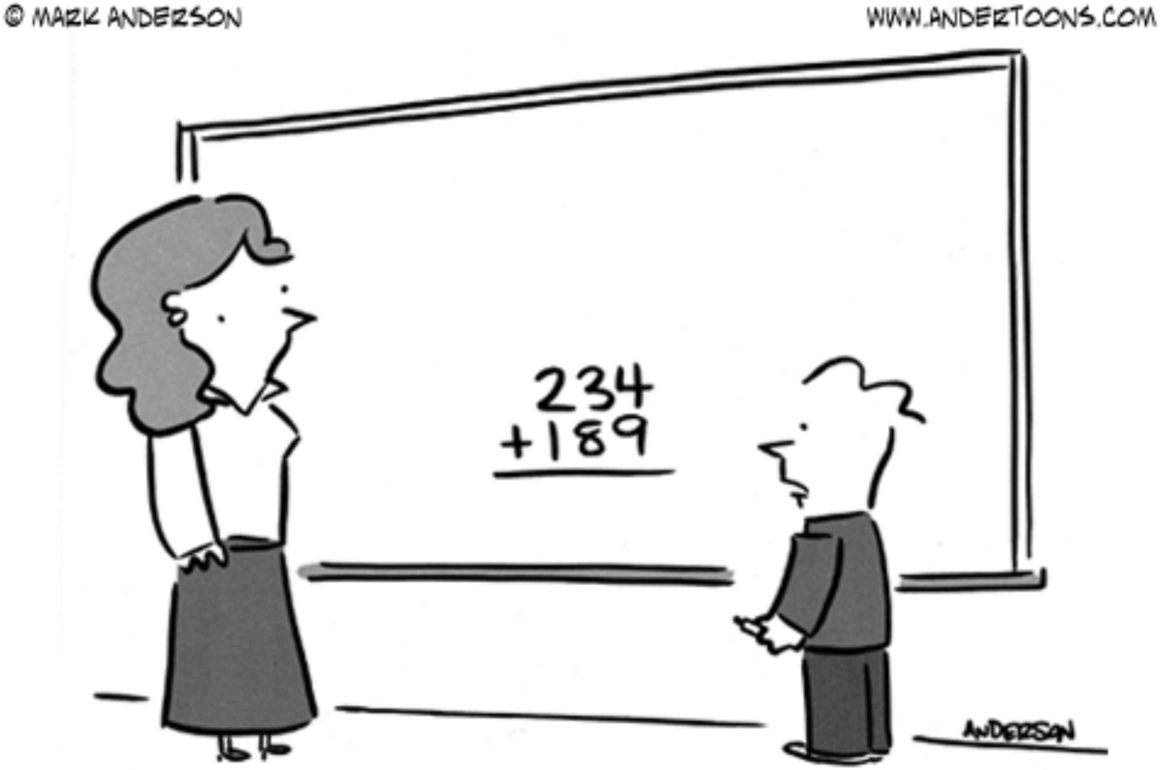
Course Policies

- Zero-tolerance for plagiarism
- Plagiarism resources
  - http://www.cdtl.nus.edu.sg/ug/resources/plagiarism.htm
- Plagiarism prevention
  - http://cit.nus.edu.sg/plagiarism-prevention/

Take-away

- All materials are available at course site in LumiNUS.
  - Workbin (Files): Lecture notes, assignments, lab exercises
  - Forum: Ask course-related technical questions in the forum.
    - If you have questions of general interest, it is recommended to ask them on the forum as your question may help other students as well.
    - But if you prefer asking over email, that is totally fine as well.
    - We will maintain a frequently asked questions list from previous and current semesters in the forum as well.
- Feedback and comments are always open.

Questions?



"Does this count as big data?"