CS4225/CS5425 Big Data Systems for Data Science

Course Overview

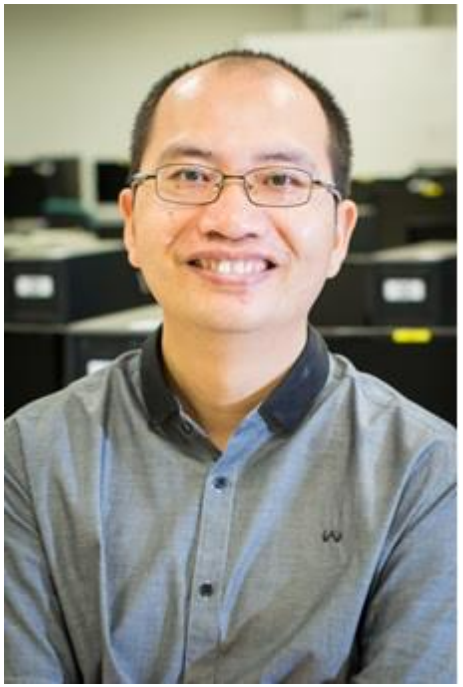
Bingsheng He
School of Computing
National University of Singapore
hebs@comp.nus.edu.sg

School of Computing

Learning Objectives

- Not a lot, just to give you an idea of who I am, what I do, and most of all what this course is about.
- Why?
  - It’s important for you to have a road-map that you can refer to when you can’t connect the dots together. ☺

About Bingsheng



- Bingsheng HE
- Office: COM3 #02-12
- Phone: 6516-7998
- Email: hebs@comp.nus.edu.sg
- Consultation:
  - By appointment through email
- My research interests: big data systems, cloud computing, parallel and distributed computing
- Industrial experiences and consultation

About Ai Xin

- Ai XIN
- Office: COM3 #B1-24
- Phone: 660 16657
- Email: aixin@comp.nus.edu.sg
- Consultation:
  - By appointment through email
- My research interests: Machine Learning and Deep Learning, Data Analytics and Big Data
- Industrial experiences: BHP Marketing Asia

Teaching Assistants

- Responsibility
  - Tutorials
  - Assist you in matters pertaining to the coding assignments
- We are fortunate to have the following great TAs.
  - CHEN XIHAO
  - CHEN JIQING
  - Edward Lim Jun Heng
  - QUEK ZHI HENG
  - Kaushik Kumar
  - Mark (Man Chun) Wong
  - Chew Cheng Yap
  - Johan Kok Zhi Kang
  - Lau Chee Loong, Desmond
- Each assignment will indicate the TAs in charge of them, so you can contact the relevant TAs (or lecturers) for assistance

Course Components

- Lectures: 13 weeks, 2 hours/week
  - Bingsheng for the first half (Weeks 1-6)
  - Ai Xin for the 2nd half (Weeks 7-13)
- Tutorials: 6 tutorials
  - Two sessions for introducing coding assignments
- Assignments: 2 coding assignments
- Assessment for both CS4225 and CS5425:
  - Two coding assignments (25% each)
  - Two tests (25% each)
- **All materials are available in the course website on Canvas.**

Schedule

Week		Lecture Topic	Tutorial	Assessment
1	Mon, 9 Jan ~ Fri, 13 Jan	Course Overview and Introduction		
2	Mon, 16 Jan ~ Fri, 20 Jan	Principles of Big Data Systems		
3	Mon, 23 Jan ~ Fri, 27 Jan	MapReduce/Hadoop- Intro		
4	Mon, 30 Jan ~ Fri, 3 Feb	Performance analysis of Big Data Systems: MapReduce	Tut:Assgn1	Assgn1 released
5	Mon, 6 Feb ~ Fri, 10 Feb	MapReduce - Relational Databases		
6	Mon, 13 Feb ~ Fri, 17 Feb	MapReduce - Data Mining	Tut:Hadoop	
	Sat, 18 Feb ~ Sun, 26 Feb	Recess Week		
7	Mon, 27 Feb ~ Sat, 4 Mar	NoSQL Overview		Assgn 1 due
8	Mon, 6 Mar ~ Fri, 10 Mar	Apache Spark I	Tut: Assgn2	Assgn 2 released
9	Mon, 13 Mar ~ Fri, 17 Mar	Apache Spark II	Tut:Spark	Midterm
10	Mon, 20 Mar ~ Fri, 24 Mar	Large Graph Processing	Tut:Graph processing	
11	Mon, 27 Mar ~ Fri, 31 Mar	Stream Processing	Tut:Stream processing	
12	Mon, 3 Apr ~ Fri, 7 Apr	No Lecture (NUS Well-Being Day)		Assgn2 due
13	Mon, 10 Apr ~ Fri, 14 Apr	Test		Week13 test

- Tutorial starts from Week 4
- **Pay attention to the time management**

Assessment

- 2 assignments (25% each)
- Midterm (25%) + Week 13 test (25%) – held in-person. **Please make sure you can come in-person for the tests.**
  - Midterm: 14-15:30pm March 18, Saturday.
    - For both Grp L1 & L2.
    - Venue: UTOWN AUDITORIUM 1/2.
  - Week 13 test: Apr 13 Thu
    - **Grp L1**: 630pm-8pm; **Grp L2**: 5pm-630pm
- (No marks for attendance)
- (Note: all in-lecture poll / quizzes are ungraded)

Lecture

- In-person lecture
- Two groups:
  - Grp L1: Thu 18:30-20:30 ~~LT15~~ I3-AUD
  - Grp L2: Thu 14:00-16:00 LT15
- Bingsheng for the first half, Ai Xin for the 2nd half
- Video recording will be given by CSIT
- Format
  - Session 1 about 45 minutes
  - 10 minutes for some short quiz (ungraded) and a break
  - Session 2 about 45 minutes

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)
- Only content discussed in class can appear on tests. Content marked “optional”, or only appearing in the powerpoint notes, will not appear on tests.

Tutorials

- Starts from Week 4
- Tutorials are not counted for final grade
- 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)
  - Sufficient materials will be given to help you do the required set up
  - Assignment 1 is in Java
  - Assignment 2 is flexible – you can use Java, Scala or Python
- Submission on Canvas
- Deadline
  - Assignment 1: Feb 27, 2023, Mon 11:59pm.
  - Assignment 2: Apr 3, 2023, Mon 11:59pm.
- Assignment documents will be available on Canvas.

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.

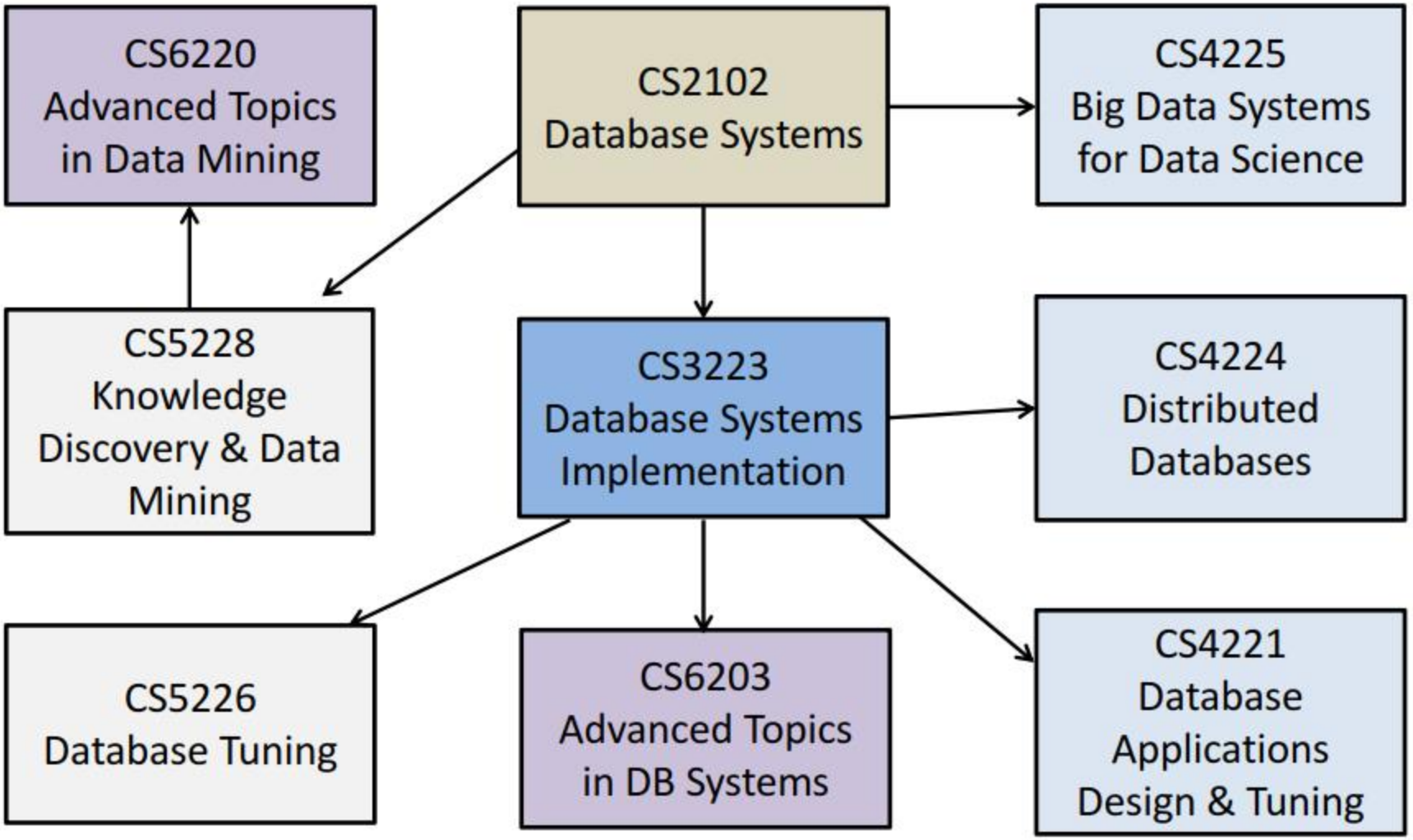
Coding Assignments (cont’)

- You'll need an SoC account for assignments. If you don't have one, please create your account at https://mysoc.nus.edu.sg/~newacct , then enable cluster service at: https://mysoc.nus.edu.sg/~myacct

Test

- Midterm (25%) + Week13 test (25%)
  - Midterm: 14-15:30pm March 18, Saturday.
    - For both Grp L1 & L2.
    - Venue: UTOWN AUDITORIUM 1/2.
  - Week 13 test: Apr 13 Thu
    - **Grp L1**: 630pm-8pm; **Grp L2**: 5pm-630pm
  - Held in person; open book + notes, but no electronics usage
- Focus is on understanding and application, not facts / memorization
- 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



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.

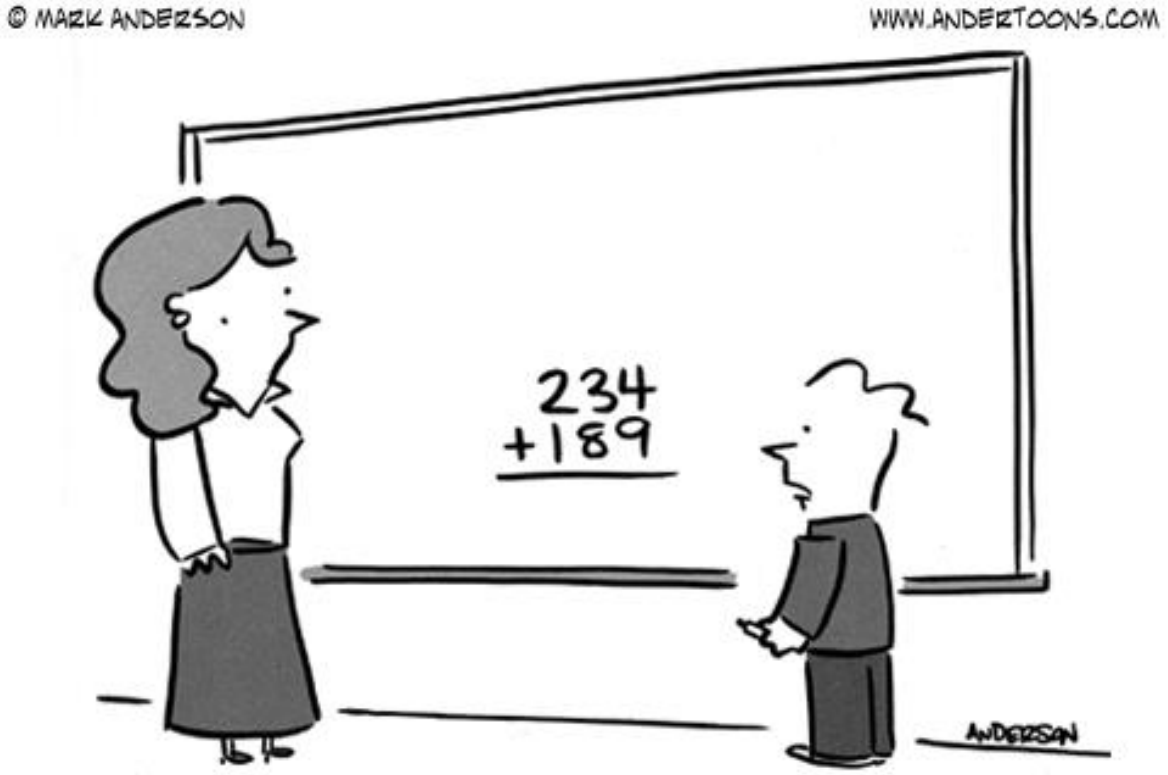
Course Policies

- 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 Canvas.
  - Files: Lecture notes, assignments, lab exercises
  - Forum:
    - We have an especially big class this semester. So, 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.
- Feedback and comments are always open.

Questions?



"Does this count as big data?"