COMP5339: Data Engineering

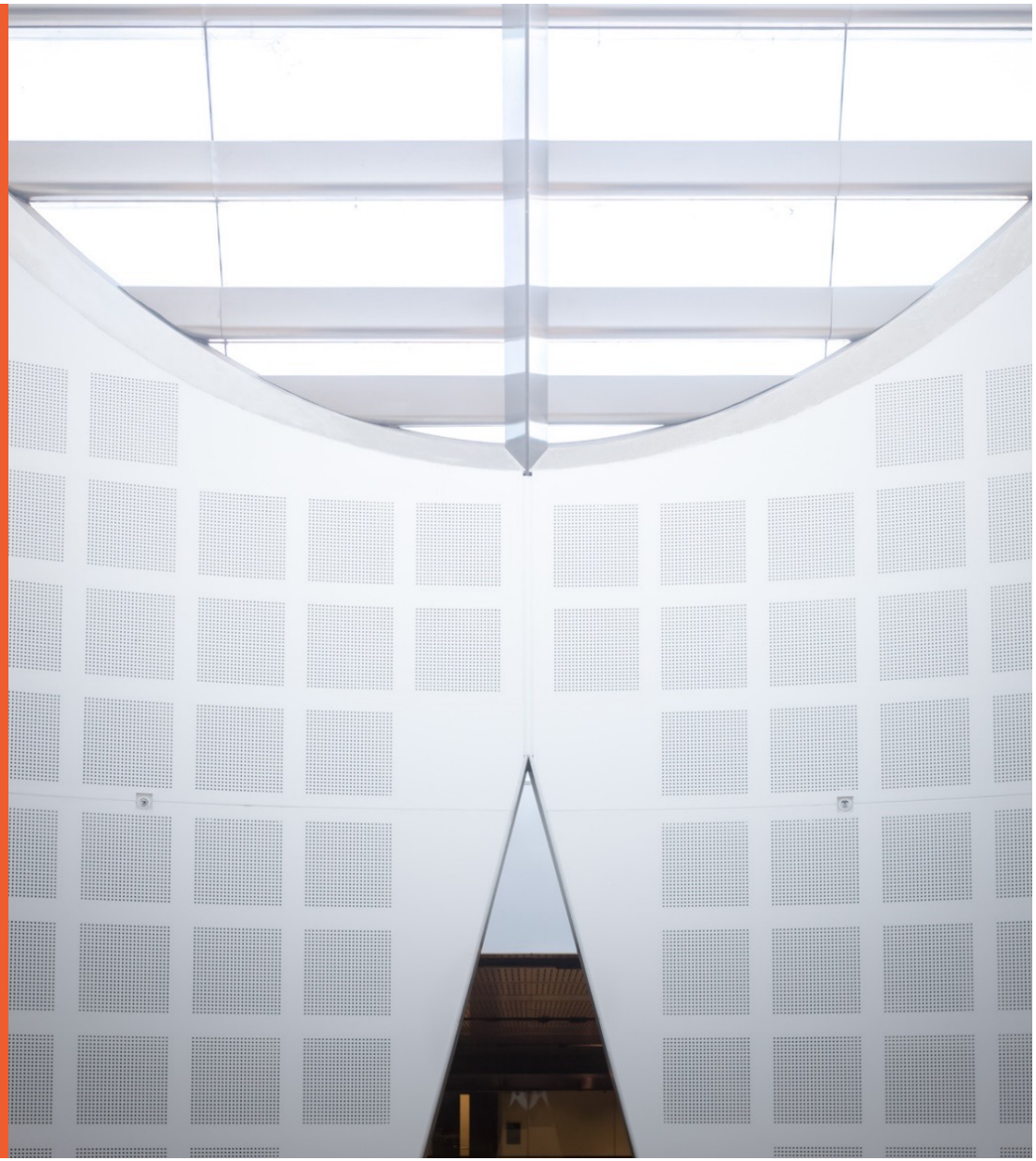
Week 13: Review

Presented by

A/Prof Uwe Roehm

School of Computer Science

THE UNIVERSITY OF SYDNEY



Outline

- General Summary
- General UoS Review
- UoS Evaluation
- Examination Tips

Overview of Lectures

Week	Topic
Week 1	Introduction to Data Engineering and Data Pipelines
Week 2	Data Acquisition and Data Cleaning
Week 3	Databases and Data Analysis in SQL ; Data Warehousing
Week 4	Web Scraping and Web APIs
Week 5	Semistructured Data and NoSQL
Week 6	Spatial-Data Engineering
Week 7	Time-Series Data
Week 8	Processing Unstructured Data
Semester Break	
Week 9	Stream Data Processing
Week 10	Scalable Data Engineering
Week 11	Data Ops and Scaling ML Pipelines
Week 12	Data Privacy and Security; Technology Choices
Week 13	Unit of Study Review

Assessment Package

Component	COMP5339
Tutorial Quiz 1 (Wk 7)	5%
Tutorial Quiz 2 (Wk 12)	5%
Assignment 1 (Wk 8)	15%
Assignment 2 (Wk 13)	25%
Final Exam	50%

- All progressive marks will be consolidated in https://canvas.sydney.edu.au
  - Report any errors/omissions within 10 days
- A pass requires at least:
  - **a) >= 40% in exam marks;** and
  - **b) >= 50% overall mark.**
- All submitted work must be your own:
  - https://www.sydney.edu.au/students/academic-integrity/breaches.html

Progressive Assessment during Semester

- Quiz 1: results available in Canvas & Gradescope
- Quiz 2 (last week) currently marked
- Assignment 1: Results and marking rubrics available in Canvas
- Assignment 2
  - **Due this** ~~Friday, 7 November,~~ **Sunday 9 November 23:59 23:59**
  - Extended for 2 days; FAQ post in Ed discussion forum
  - Note: While AI writing tools are allowed, you must properly **acknowledge any AI assistance in your submissions.** This includes citing the specific AI tools used and describing how they contributed to the work.
    - also keep a record/log of how you used those tools

https://www.sydney.edu.au/students/academic-integrity/artificial-intelligence.html

Final Examination

Objective

Assess understanding of unit material; understanding of data engineering lifecycle including principles, data models, technologies, and undercurrents (security, data architecture, DataOps, etc).

Content

- Questions about all lecture and tutorial material
- Examples and more details
  see next few slides

Format

- **In-person, written exam** (on campus)
  - scheduled: **Fri 21 November, 1pm (AEST)**
  - 2 hours 10 min duration
  - Various exam rooms -> check timetable
- restricted open-book
  1. allowed: 1 page own notes
  2. allowed: bilingual dictionary
- 50% of final mark

SIT policy: You must get 40% on the exam and 50% overall to pass COMP5339

More exam details in Part 3…

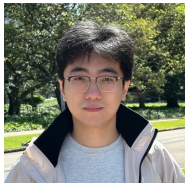
Shout-out to the Teaching Team

- Lecturer: Uwe Roehm
- TAs: **Rohan**, Vinit and Kiran
- Tutor Team:


Chen


Haoyu


Hengzhi


Jichao

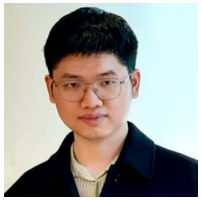

Kiran


Rohan


Sirui


Tianyi


Vinit


Xiaocheng


Yiran

How to make your USS feedback count

Your Unit of Study Survey (USS) feedback is **confidential**.

It's a way to share what you enjoyed and found most useful in your learning, and to provide constructive feedback. It's also a way to 'pay it forward' for the students coming behind you, so that their **learning experience** in this class is as good, or even better, than your own.

When you complete your USS survey (https://student-surveys.sydney.edu.au), please:

Be specific.
Which class tasks, assessments or other activities helped you to learn? *Why* were they helpful? Which one(s) *didn't* help you to learn? *Why* didn't they work for you?

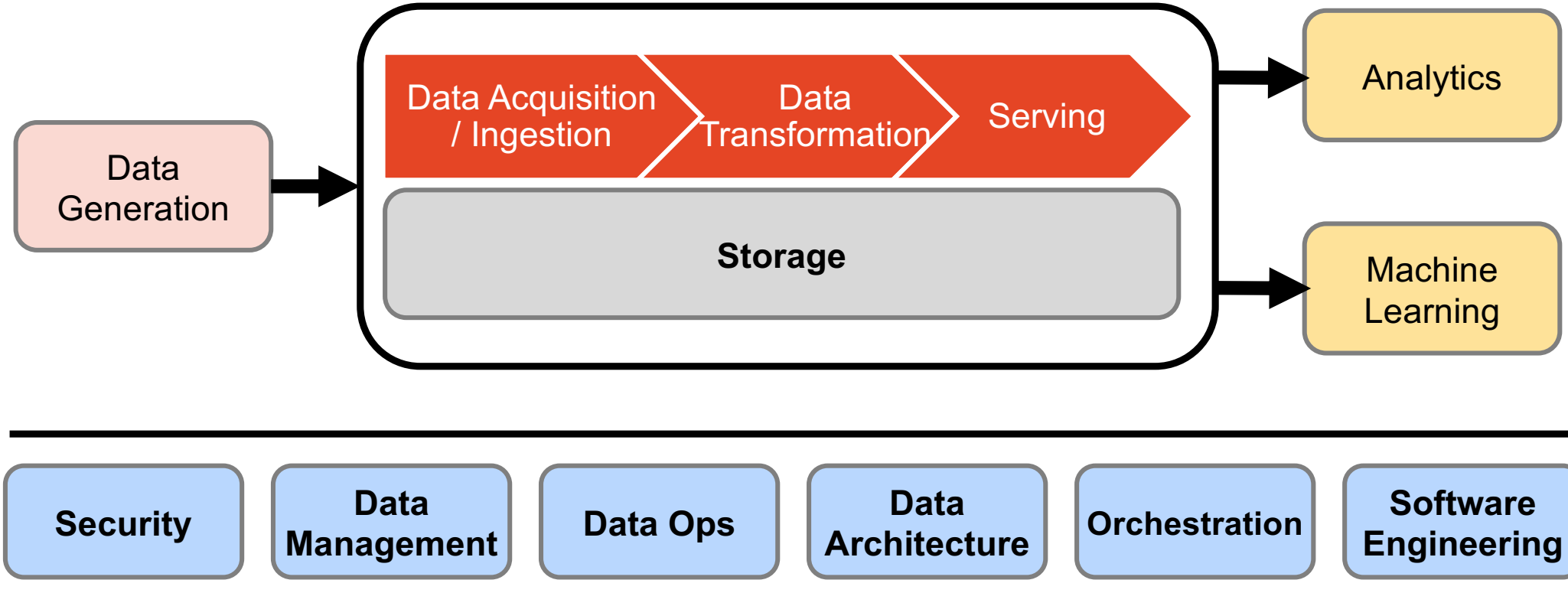
Be constructive.
What practical changes can you suggest to class tasks, assessments or other activities, to help the next class learn better?

Be relevant.
Imagine you are the teacher. What sort of feedback would you find most useful to help make your teaching more effective?

Content Review

THE UNIVERSITY OF SYDNEY

Data Engineering Lifecycle



[J. Reis and M. Housley: Fundamentals of Data Engineering, 2022]

Review: Data Acquisition and Data Cleaning

- Important Aspects:
  - **Data Sources:** file-based, databases, web scraping, APIs, Publish/Subscribe
  - **Push / Pull / Poll access patterns**
  - **ETL vs ELT**
  - **Data Cleaning** issues

- Example questions:
  - Identify data quality issues in some given example use case, and suggest fixes.
  - Differences between ETL and ELT? When use which?
  - Given a use case scenario, which data acquisition patterns would you suggest to use to meet the scenario's requirements?

Review: Databases and Data Warehouses

- Important Aspects:
  - **Databases:** relational data model, SQL query language, DBMS as source or sink
  - **Column vs Row Stores;** Schema-first vs Schema-late
  - **OLAP, Star Schema**
  - **Data Warehouses vs Data Lakes, ETL vs ELT**

- Example questions:
  - Differences between OLTP and OLAP? Give a primary use case for each.

    [or in general, comparing any other two related techniques in DE]
  - Given a use case scenario, which database solution would you suggest as source/sink to use to meet the scenario's requirements?

Review: Web Scraping and Web APIs

- Important Aspects:
  - **Web Scraping** process (Reconnaissance, retrieval, extraction, cleaning, …)
  - HTML, DTD and CSS selectors / content extraction
  - **Web API** programming; REST vs SOAP/XML web services
  - Ethics, authentication API keys, robots.txt

- Example questions:
  - Explain the different parts of the web scraping process.
  - Given some HTML document, which sub-part selected by a CSS selector?
  - When would you use web scraping and when web APIs at data ingestion?

Review: Semistructured Data and NoSQL

- Important Aspects:
  - **Semistructured Data: JSON vs XML**
  - Examples for content selection: XPath, CSS Selectors
  - JSON/XML support in relational databases (eg. JSON/JSONB data types)
  - NoSQL Databases:
    - Document databases such as MongoDB
    - Graph databases such as Neo4J
- Example questions:
  - Differences between JSON and XML? When use which?
  - Given some XML document, which sub-part selected by a XPath expression?
  - When would you use a NoSQL document databases, when a relational DB?

Review: Spatial and Temporal Data Engineering

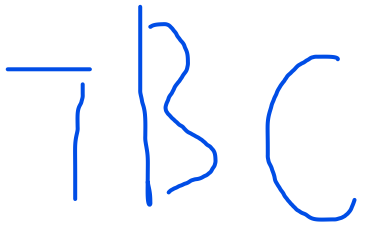
- Important Aspects:
  - **OGC Spatial Data Model**, coordinate systems, topological operations
  - SDBMS vs GIS; tool support: **PostGIS**, **GeoPandas**, GeoJSON and KML
  - **Temporal databases**, temporal data types, **kinds of time**, NOW handling
  - Timeseries data representations

- Example questions:
  - Differences between SDBMS and RDBMS? When use which?
  - Role of coordinate system?
  - Differences between valid and transaction time?

week8

Review: Unstructured Data ; Machine Learning Pipelines

- Important Aspects:
  - **Unstructured Data:** Text, Images, Video, …
  - **Feature Extraction:** for text bag-of-words TF/IDF or word embeddings (eg. BERT) for images white vs black box approaches; meta-data
  - **Scaling ML Pipelines**: ML-to-Data or Data-to-ML?
    - FeatureStore, MADlib MADlib为什么适合



- Example questions:
  - Differences between TF/IDF and BERT for feature extraction?
  - What is Apache MADlib?
  - Provide an example of feature extraction for unstructured data.

Review: Stream Data Processing

- Important Aspects:
  - **Data Stream Processing**
  - notions of time (event time etc);
  - **window processing;** types of windows; watermarking
  - **examples: Kafka and Apache Flink**

- Example questions:
  - Differences between stream processing systems and database systems? When would you use which?
  - Role of Pub/Sub broker such as Kafka?
  - Given a real-time data analysis scenario, what data architecture would you suggest? Would a publish/subscribe system such as Kafka help?

Review: Scalable Data Engineering

- Important Aspects:
  - **Scale-up vs Scale-out, CAP Theorem**
  - Principles of Good Data Architectures, such as Availability, Security, Elasticity, Scalability (Throughput, Latency)
  - **Data Replication, Data Partitioning / Sharding**
  - **Platforms:** MapRecuce, Hive, Spark, Flink

- Possible exam questions:
  - What is the meaning of the CAP theorem?
  - Differences between Apache Spark and Apache Flink? When use which?
  - Role of data replication? Which problem does it address?

Review: DataOps, Security, Privacy

- Important Aspects:
  - **DataOps Lifecycle**
  - **Personally Identifiable Information (PII) + Sensitive Information**
  - Data Minimalism and Principle of Least Privilege
  - **Encryption At-rest and In-Transit,** Logging, Monitoring, Network Access
  - Data Backup and Recovery

- Possible exam questions:
  - Explain four key functions of the DataOps Lifecycle.
  - What is the difference between encryption at-rest and in-transit?
  - Your data pipeline is processing some sensitive information; which key security and data privacy measures do you need to implement?

EXAM ARRANGEMENTS

THE UNIVERSITY OF SYDNEY

Exam Schedule

Friday, 21 November, 13:00 – 15:10 (early afternoon)

Venue: multiple rooms (please check your exam timetable)

- On-campus, in-person written exam
- Restricted open book exam: handwritten notes, printed notes
  - One A4 sheet of paper of own notes (hand-written or typed; double-sided)
  - Bilingual dictionary allowed too
- Two-hour, supervised in-person examination
  - Read through the in-person exam site for advice and preparation tips.
    - cf. https://www.sydney.edu.au/students/exams/in-person.html
  - You will need to bring one form of valid photo identification to your exam
    - **Don't forget to bring your University student card!**

Exam Overview

- There will be five to six main topics covered.
  - Organised in 4 sections…
- Combination of different types of questions.
  - **Short & Long answer questions, some multiple choice questions too**
  - Typically with increasing level of difficulty.
  - Follows the style of the tutorial Quiz 1 and Quiz 2
    but note: more emphasise on text questions than in the tutorial quizzes…
- The exam will have a total of 50 marks
- You need to get a **at least 20/50 in the final exam** to pass

Sample Exam Questions

- available on Canvas -> Modules

- Notes: The example exam only gives examples of exam-style short-answer questions. The actual exam will be longer and also have a section of multiple-choice questions, similar to Quiz 1 and Quiz2.

Exam Techniques

- You will get an exam script with exam questions which you can answer in the order of your liking
  - make sure that you answer **all** questions

- Questions will be a mix of multiple choice, design and short answer questions
  - The latter have to be answered in the answer boxes provided in the exam script
  - Whenever there is a tip on how to format your answer, please follow this
    - E.g. to label answer parts to correspond to different question parts

Exam Techniques (cont'd)

- In the short-answer questions, check for "Justify your choice", "Briefly explain", "Describe", "Why?", "Discuss" or "Give an example" parts.
  - Such questions test your *understanding* of an area.
    - A simple yes/no is not enough!
  - Please answer BRIEFLY
    (one or two sentences are typically OK, but NOT a whole page)
    - E.g. if you have to compare two techniques, a good approach is to first define the techniques in one sentence each, before actually comparing them in more detail (and don't forget that last part ;)
  - Say it in your own words, don't just copy from the textbook.
  - Please write complete, English sentences!
    - You want the marker to understand what you wanted to say…

Special Consideration

- If you are unwell for the exam, please apply for special consideration.
  - If your special consideration application is approved, you will be able to take a replacement exam later.

- If you are approved for a replacement exam, please ensure you attend it.
  - Special consideration for the replacement exam may result in an oral examination (viva).

Exam Advice

- Plan how you will allocate time (wisely)
  - Use “reading time”
- Answer everything (get the “easy marks”)
  - If you are uncertain about a question during the exam, answer to the best of your ability
- Write clearly
- If you need more space, use blank pages at the end of the exam booklet but leave a forwarding pointer in the provided space (where the marker will be looking)

Exam Advice

- Find the room location before the exam day!
- Bring spare pens (either black or blue)
- Bring water, and have clothing in layers
- Have your student ID and put it on the desk

FINAL ACTIVITIES

Final Activities

- **Assignment 2 due on Sunday 9 November, 23:59**
  - Deadline extended to include weekend before study week
- Tutorials this week with assignment help

- Review your marks on Canvas/Gradescope
- Complete USS survey: https://student-surveys.sydney.edu.au/students/
- Review lecture slides, and quizzes

- Good luck!!! You are going to do great!