CPSC 368: Databases in Data Science

Jessica Wong (jhmwong@cs.ubc.ca)

Piazza: Sign up by using the Piazza link in the Canvas sidebar
Course website: https://www.students.cs.ubc.ca/~cs-368/

Basic Housekeeping

- The Head of the Computer Science department has determined that this course will provide an **in-person and not a hybrid, online, or multi-access mode of instruction**.

COVID-19 Safety

- We expect and prefer that students wear a non-medical mask during our class meetings, for your own protection, and the safety and comfort of everyone else in the class.

- If you have not yet had a chance to get vaccinated against Covid-19, vaccines are available to you, free: http://www.vch.ca/covid-19/covid-19-vaccine#clinics

What is CPSC 368?

- How do we use a database to find the answers that we need?

- **Not** focused on database design (take CPSC 304 for that!)
- You **cannot** take CPSC 404 after taking this course. You need to take CPSC 304.
- This course is credit excluded with CPSC 304 (you cannot count both of these courses for credit towards graduation).

Learning Goals

- Define the term *database* and explain the purpose of having a database.
- Explain the high-level objectives of a *database management system* (DBMS), and explain how a DBMS relates to a database. List benefits that result from the usage of a DBMS.

Associated Textbook Chapter: Chapter 1

Why are databases *interesting*?

- They're useful for jobs
- DBMS encompasses most of CS
  - OS, languages, theory, AI, multimedia, logic,…
- Datasets are increasing in diversity and volume.
  - Digital libraries,
  - Interactive video
  - Human Genome project...
  - Amount doubles every *18* months (since 1990's)
  - For more fun; try combining them!
  - We put the data in big data
  - *Everyone* has data!

Real Life Example

"….provides each individual in Canada with a secure and private lifetime record of their key health history and care within the health system. The record is available electronically to authorized health care providers and the individual anywhere, anytime in support of high quality care."

Canada Health Infoway 2004

What is a database?

- A **database** is an organized collection of related data, usually stored on disk. It is typically:
  - Important data
  - Shared
  - Secured
  - Well-designed (minimal redundancy)
  - Variable size
- A DB typically models some real-world ***enterprise***
  - Entities (e.g., students, courses)
  - Relationships (e.g., Ting got 95% in CPSC 221)

What is a database management system?

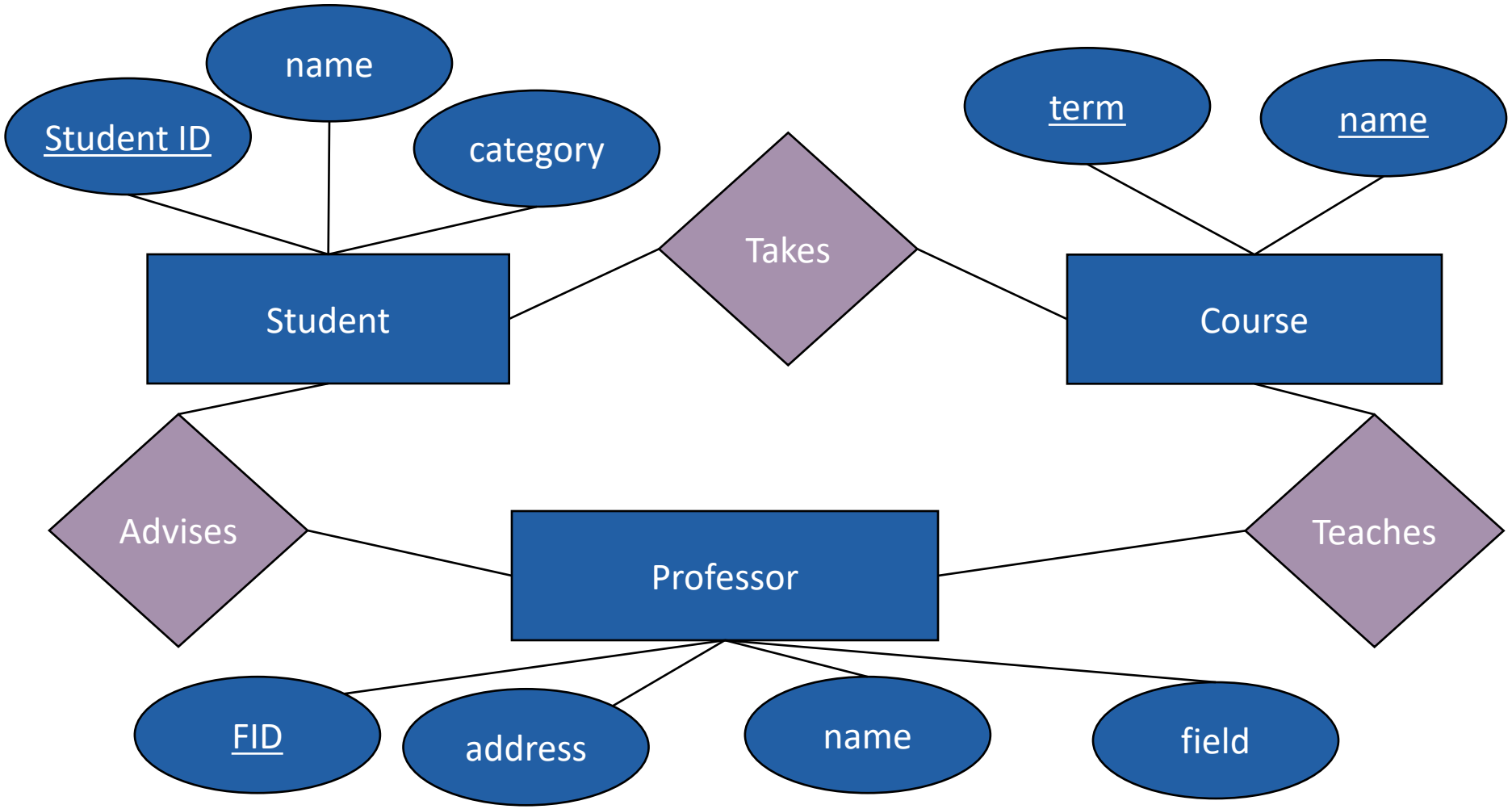
- A ***Database Management System (DBMS)*** is a bunch of software designed to store and manage databases. It is used to:
  - Define, modify, and query a database
  - Control access
  - Permit concurrent access
  - Maintain integrity
  - Provide loading, backup, and recovery
- Some common DBMSs are Microsoft SQL Server, MySQL, MariaDB, and Oracle
- In this class, we will use Oracle and MongoDB as our DBMS

Great! What's left for us to do?

In this course we'll ….

1. Conceptually model the concepts
2. *Logically* model concepts in a database
3. Query our data using a relational database and SQL
4. Look at how queries are processed
5. Talk about large scale data storage and processing by examining data warehousing
6. Discuss some data mining techniques
7. Examine issues surrounding the collection and mining of data
8. Examine non-relational databases and why they are useful

1. Conceptual modelling



2. Logical modelling

- Data model : a collection of tools for describing
  - data, data relationships, semantics, constraints
- We'll use the Relational Model

Students Table:

Student	Course	Term
Ying	CPSC 368	Winter 2, 2021
Andrew	CPSC 221	Summer 2, 2019

Separates logical and physical views of the data.

3. Query data using a relational database and SQL

- SQL is a way for you to grab data that you need from a database

- Example: Find all the students who have taken CPSC 304 in Winter Term 2, 2017.

```
SELECT  E.name
FROM Enroll E
WHERE E.course=“CPSC 304” and E.term=“Winter Term 2, 2017”
```

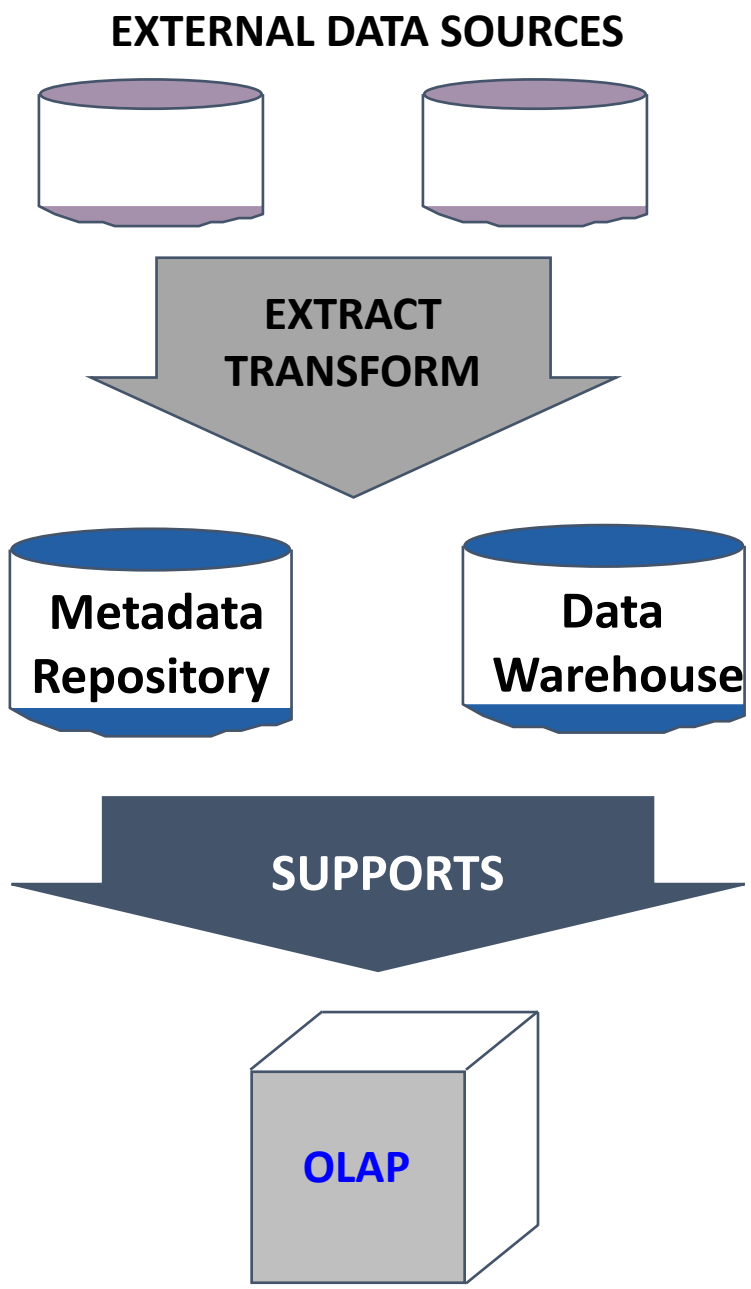
4. Looking at how queries are processed

- How can you improve the speed of a query?
- How does a database recover after a crash? (if time permits)

5. Data warehousing

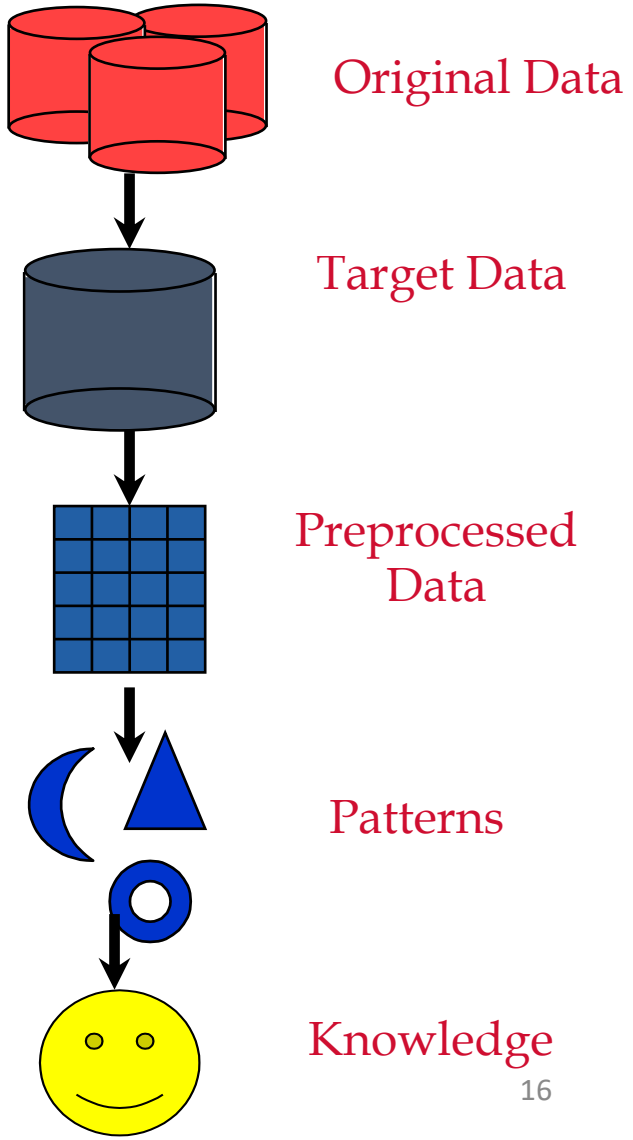
- Increasingly, organizations are analyzing current and historical data to identify useful patterns and support business strategies.

- The emphasis is on analysis of complex queries on very large datasets created by integrating data from across all parts of an enterprise.



6. Knowledge discovery and data mining

- *Knowledge discovery and data mining* is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.
- The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing



7.Data collection and privacy

- Just because you can do something, **should** you?

- What are some things you should consider before collecting data?

8. Non-relational databases

- Non-relational databases store data very differently from a relational database (it doesn't use a table model)
- Many different companies use them (e.g., Google)
- We'll examine a non-relational database by using MongoDB

Planned Assessments

- Assignments (7) : 32%
  - Assignments are not weighted equally
- Clickers: 3%
- In-Class Exercises: 3%
- Midterm: 22%
- Final: 40%

Minimum Passing Criteria

- Achieve an overall grade of 50% or better
- Pass the weighted average of the in-term and final exams **and** score at least 45% on the final exam

Platforms Used in this Class

- Canvas
- Jupyter/Syzygy
- Piazza
- Gradescope (more on this later in the term)

Canvas

- Holds all of our course content
- Your grades will be stored here
- Contains links to the other websites we use in this class (e.g., Piazza)
- **Important**: the tentative class schedule

Jupyter/Syzygy

- A few of the assignments will be programming based
- The programming assignments must be completed in Python
- You do not have to use Syzygy if you do not want to but we will not be responsible for installing/fixing any issues you encounter if you install your own copy of Jupyter

Clickers

- To make class more interactive and more tuned to how students are doing, we will use iClicker Cloud (it is free for UBC students at least this term): https://lthub.ubc.ca/guides/iclicker-cloud-student-guide/

- Clickers are graded for participation, not for correctness

- To allow for days when you are sick or have technology issues, you only need to score 80% of the total value of clickers to obtain full marks for this section of the course

In-Class Exercises

- To help you practice the concepts, we will have “in-class” exercises that you will complete
- These exercises will be due the day after class at 10PM
- Graded based on effort and completeness (does not necessarily have to be correct). We will take the best 90% of days to allow for the odd missed class.
- We will release the solution to the exercise after the deadline

Assignments

- Assignments are a mix of programming based and theoretical based questions. They are meant to give you practice with different aspects of handling data.
  - You will be allowed to complete the programming based assignments in pairs
  - 3 programming and 3 theory assignments
    - Four theory assignments if we include the syllabus quiz

- Due on Fridays (see the tentative schedule for specific dates)

- Typically released two weeks before the due date

Exams

- The in-term exam will be **in-class** on Thursday March 9, 2023
- It will be a paper based exam. No calculator, closed book, and closed neighbour.
- The final exam covers more topics than the midterms

Tutorials

- Attendance at tutorials is completely optional
- Tutorial material can be found on Canvas
- It’s a great way to practice class material
- Tutorials will start with a small lecture by your TAs. Depending on the tutorial, it can then become a mini-office hour or more of a lab

Slides

- A pre-class set of slides for each topic will be posted on Canvas prior to that topic's introduction in class
  - We'll work on clicker questions and in-class exercises together in class
  - The pre-class slides are subject to change
- After each class, I'll post the slides with the answer to the clicker questions so you can refer back to them if need be

The Collaboration/Cheating Policy

In general, you can collaborate as much as you want with whomever you want on turned-in work, with three restrictions:

- You must acknowledge everyone you collaborated with
- You may not take **any** record away from the collaboration.
- You cannot show anyone your answers (but you can discuss general strategies for how to approach a problem).
- You must spend at least an hour after collaborating and before working on your submission doing mindless activities

The exceptions are:

- Collaboration with the instructor and TAs is excluded from the above rules.
- You may *not* collaborate on the midterms, the final, unless explicitly stated.
- Follow the spirit of the rule and use common sense

Miscellaneous

- Most of the course material will be posted on Canvas but we will be using Piazza for our discussion board
- Readings are at the bottom of the title slide for a new set of slides.
- You are responsible for the material in the readings even if not explicitly covered in class.
  - Note that pre-lecture slides are subject to change and are only posted to help you prepare for the lecture.
  - Slides will be posted on Canvas
- **Piazza is the fastest way to get help!**

Communication

- If you have questions of a general nature, please post on Piazza.
- If you have specific concerns, then please email me or talk to me after class!
- Office hours will be posted on Canvas
- **PLEASE don't message us using Canvas.**

  The Canvas messaging system is not a communication avenue we use and messages sent here may not be seen/responded to until much later.

Learning Goals Revisited

- Define the term *database* and explain the purpose of having a database.
- Explain the high-level objectives of a *database management system* (DBMS), and explain how a DBMS relates to a database. List benefits that result from the usage of a DBMS.

TODOs

- Join Piazza
- Ensure that you have registered in a tutorial
- Get iClicker Cloud set up for next class (https://lthub.ubc.ca/guides/iclicker-cloud-student-guide/)
- Read through the syllabus and then do the syllabus quiz

Ask Me Things!

- For example:
  - What are your concerns?
  - Is anything unclear?
  - Is there something you want to talk about?
- Other topics welcome!
- **Disclaimer**: Registration (including prerequisite checking) is controlled solely by the CS department and I have no information about it whatsoever. I am also not allowed to sign any course add/drop or conflict forms.