CSC 343
Introduction to Databases

Nosayba El-Sayed (based on slides from Diane Horton)
Fall 2015

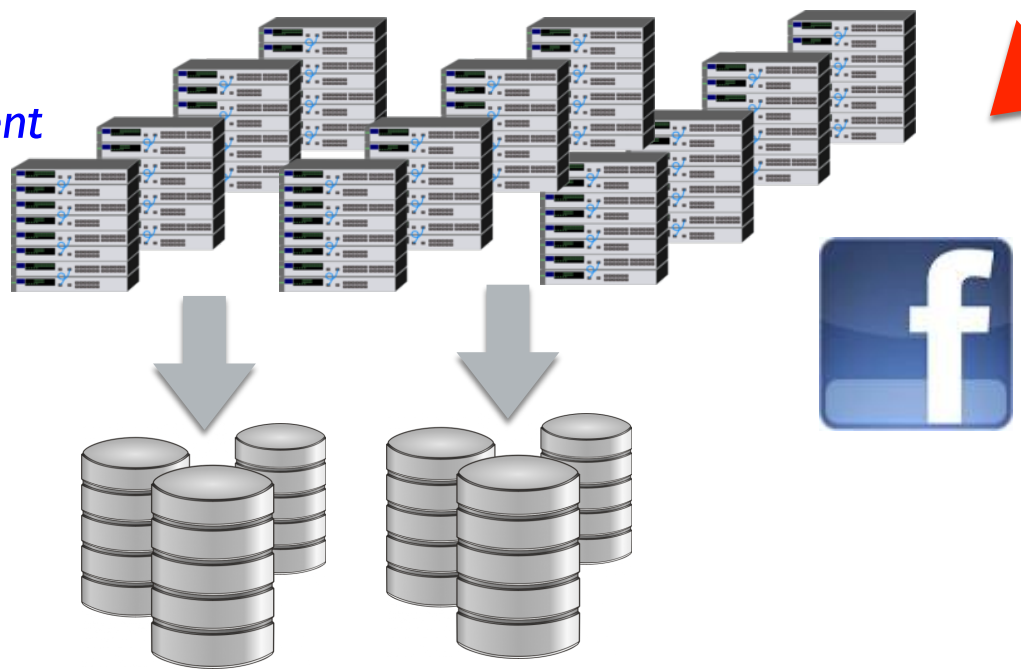
http://www.cdf.toronto.edu/~csc343h/fall

UNIVERSITY OF TORONTO

DCS50

Why study databases?

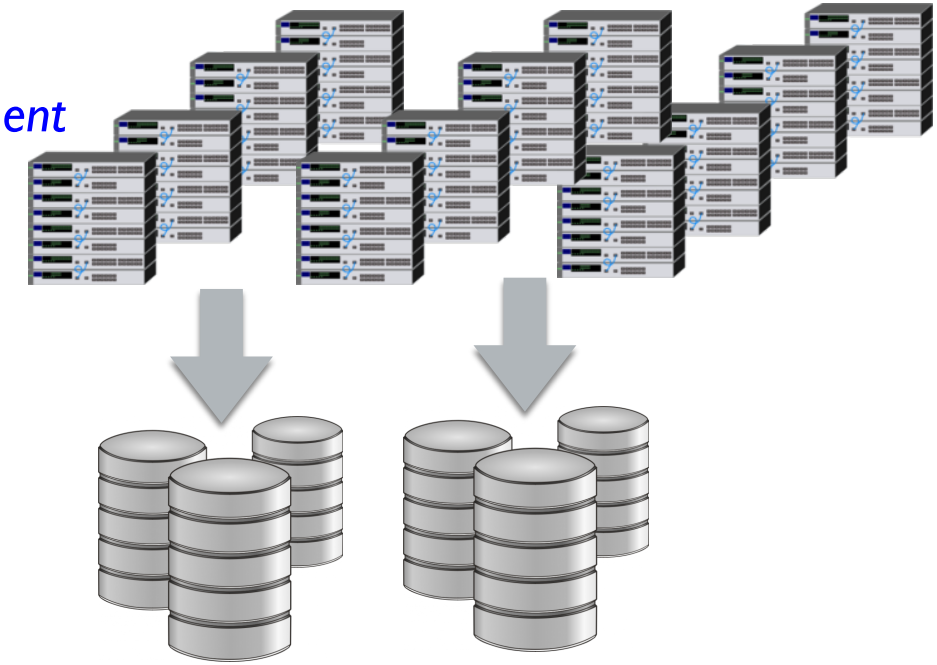
Database Management System (DBMS)



Storage

Why study databases?

Database Management System (DBMS)

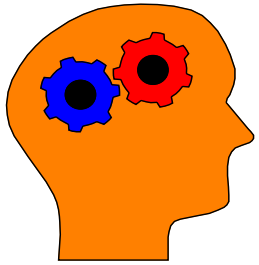


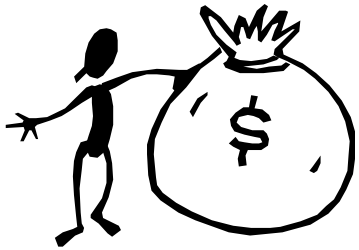
Storage

Why study databases?

- Interesting concepts and techniques.
- Spans computer science, including OS, languages, theory, AI, multimedia, logic.
- Databases have become increasingly important
  - shift from a focus on computation to information
  - data increases in volume and diversity.

- Jobs: In demand and well paid.
- Research: Many open problems.





Our first hour or so..

- Some key concepts
- Examples to motivate the course
- Admin info

Databases and DBMSs

- Databases are <u>everywhere</u>, often behind the scenes.
- DBMS (Database Management System):
  "A powerful tool for creating and managing large amounts of data efficiently and allowing it to persist over long periods of time, safely."
  [Ullman and Widom, FCDB]
- Database:
  a collection of data managed by a DBMS.

Data models

- Every DBMS is based on some data model: a notation for describing data, including
  - the structure of the data
  - constraints on the content of the data
  - operations on the data

- Some specific data models:
  - network & hierarchical data models — of historic interest
  - relational data model
  - semistructured data model

The relational data model

- Main concept is a "relation."
  Based on the concept of relations in *math*.
- Can think of as tables of rows and columns.

Teams

Name	Home Field	Coach
Rangers	Runnymede CI	Tarvo Sinervo
Ducks	Humber Public	Maeve Mahar
Choppers	High Park	Tom Cole

Games

Home team	Away team	Home goals	Away goals
Rangers	Ducks	3	0
Ducks	Choppers	1	1
Rangers	Choppers	4	2
Choppers	Ducks	0	5

Example ...

- A dataset scraped from Twitter

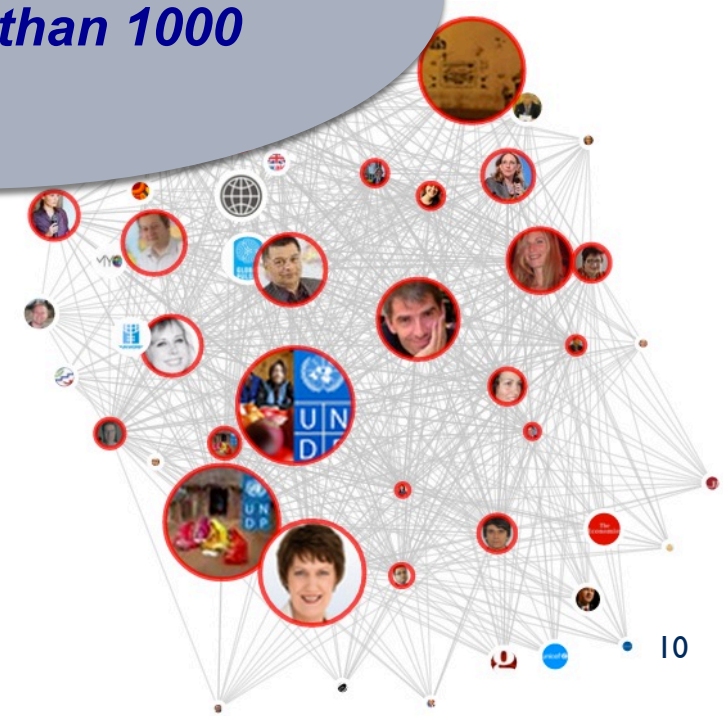
twitter-scrape.txt

```
iamwill
will.i.am

http://dipdive.com/member/iamwill
i.am.will...and all the other i.am's are jockin' my style
ENDBIO
kanyewest
RonConwayFacts
Oprah
dickc
BarackObama
END
yorchopolis
Jorge Aranda
Toronto
http://catenary.wordpress.com/
I'm a PhD student at the University of Toronto
vegetarian, and a boardgame collector.
ENDBIO
zuzelvp
LilaFontes
jhariono
swcarpentry
mike_conley
mfeathers
irvingreid
cjmuise
adinscannell
jonpipitone
algore
END
mattcohler
Matt Cohler
```

Who has the most followers?

Who is in Toronto, mentions DJ in their bio, and has more than 1000 followers?



Example ...

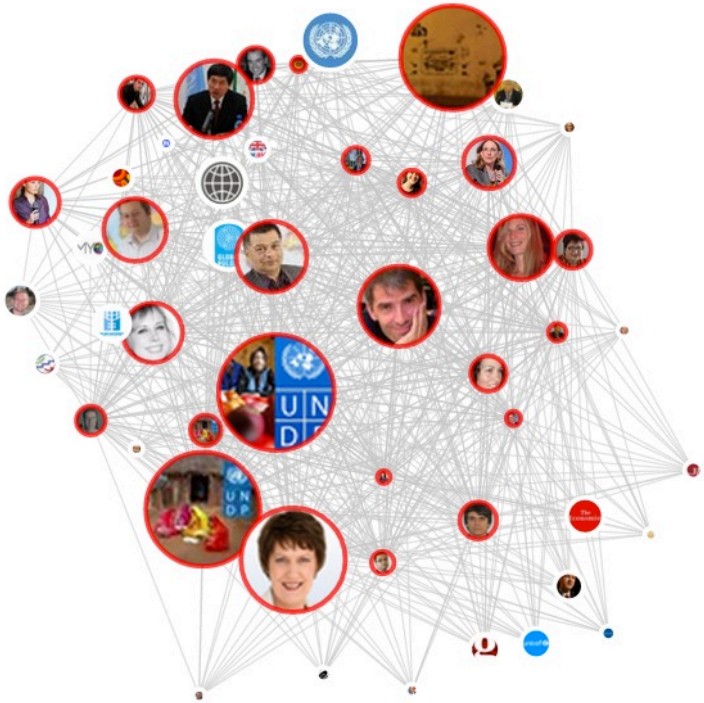
- A dataset scraped from Twitter
- Defining a schema that expresses its structure

twitter-schema.sql (~/Desktop/CSC343/mylectures...

```sql
DROP SCHEMA IF EXISTS Twitter CASCADE;
CREATE SCHEMA Twitter;
SET SEARCH_PATH TO Twitter;

CREATE TABLE Profile (
        ID VARCHAR(50),
        name VARCHAR(50),
        location VARCHAR(50),
        url VARCHAR(150),
        bio VARCHAR(500),
        PRIMARY KEY (ID)
);

CREATE TABLE Follows (
        a VARCHAR(50),
        b VARCHAR(50),
        PRIMARY KEY(a, b),
        FOREIGN KEY (a) REFERENCES Profile(ID)
);
~
~
~
~
~
~
~
~
```

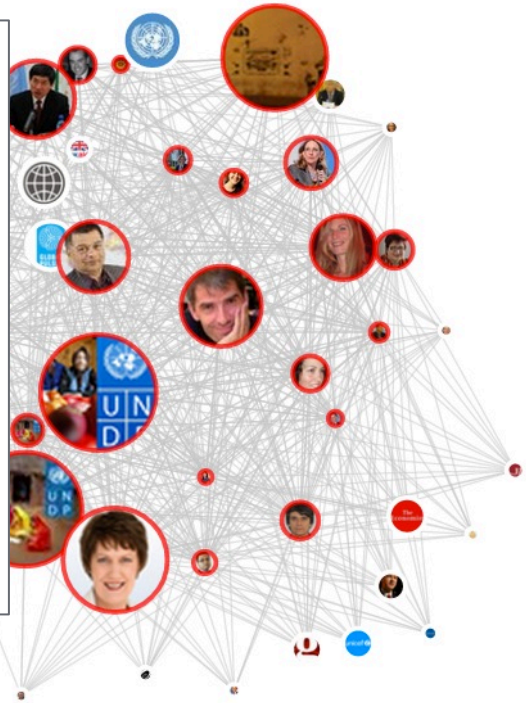


Example ...

- A dataset scraped from Twitter
- Defining a schema that expresses its structure
- Creating an instance that contains the data
- Writing some queries on the data…

```
>> select id, name, location
from profile
where location = 'Toronto'
and bio like '%DJ%';

       id          |        name        | location
-------------------+--------------------+----------
 yorchopolis       | Jorge Aranda       | Toronto
 lance_            | lance underscore   | Toronto
 zuzelvp           | Zuzel Vera         | Toronto
 karenreid         | karenreid          | Toronto
 torontoist        | Torontoist         | Toronto
 dianelynnhorton   | dianelynnhorton    | Toronto
```

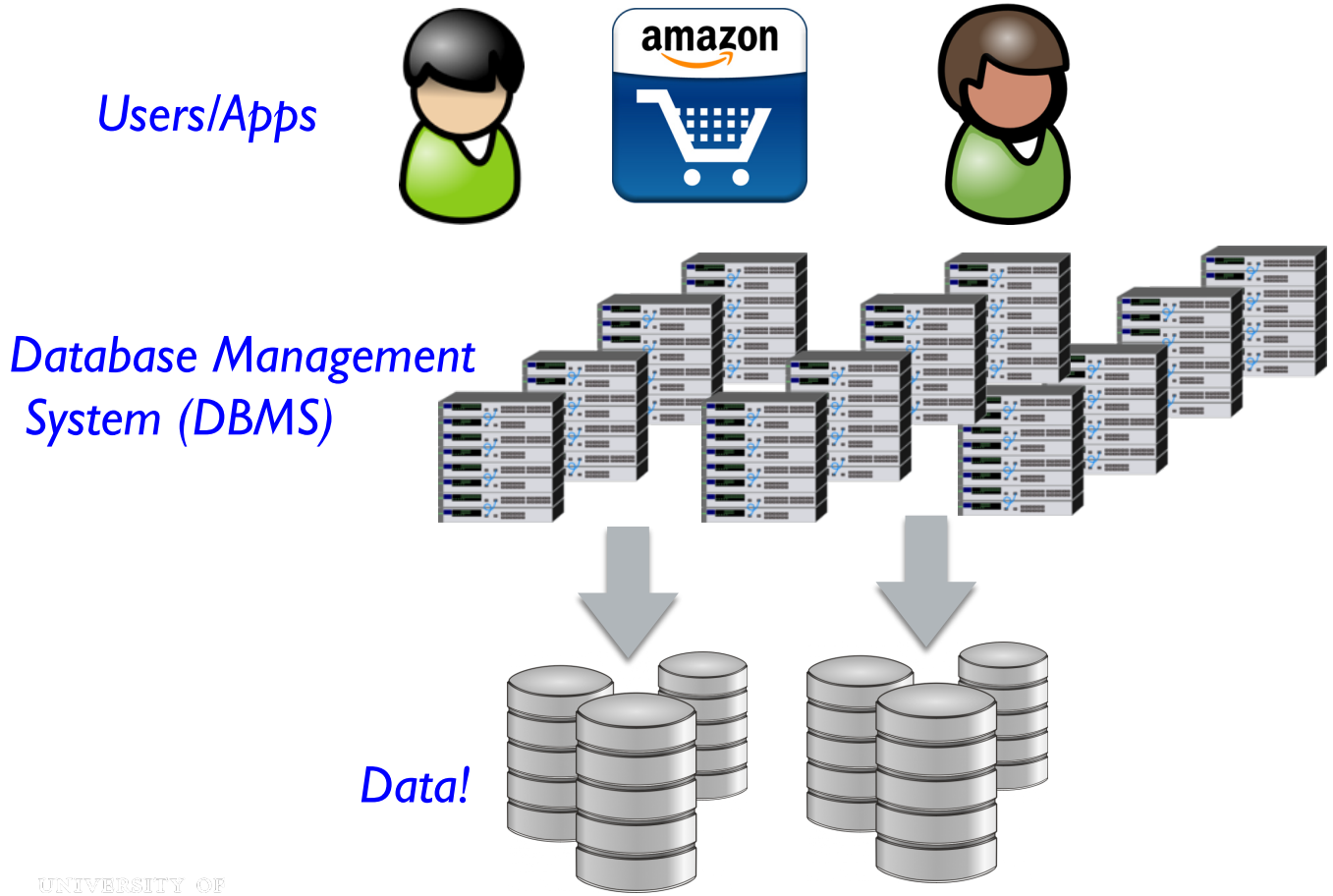


What a DBMS provides..

- Ability to specify the logical structure of the data
  - explicitly
  - and have it enforced
- Ability to query or modify the data.
- Good performance under heavy loads (huge data, many queries).
- Durability of the data.
- Concurrent access by multiple users/processes.

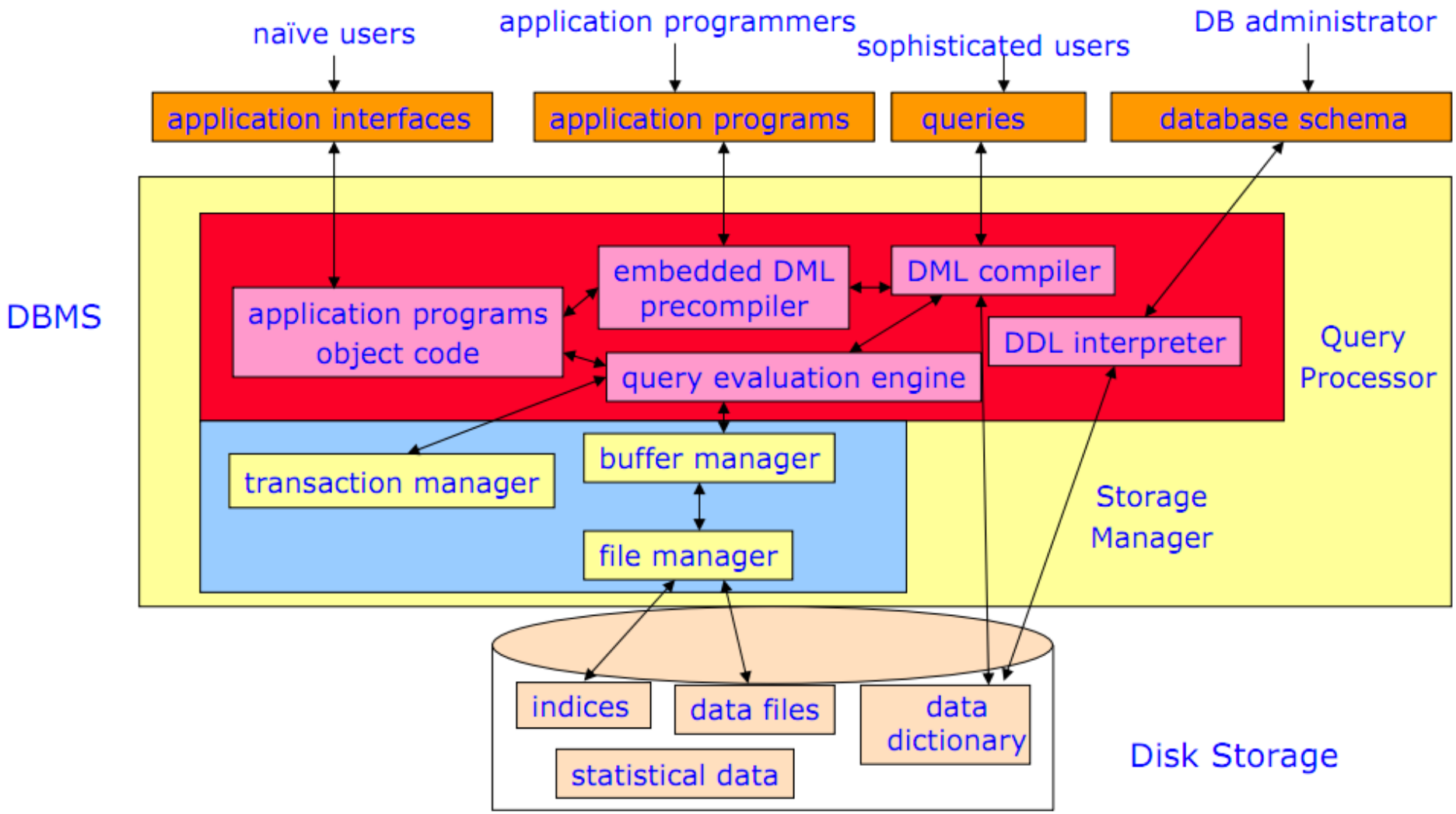
Overall architecture of a DBMS

- The DBMS sits between the data and the users or between the data and an application program



Overall architecture of a DBMS

- No like, seriously..?



Overall architecture of a DBMS

- The DBMS sits between the data and the users or between the data and an application program
- Within the DBMS are layers of software for:
  - parsing "queries"
  - implementing the fundamental operations
  - optimizing queries
  - maintaining indices on the data
  - accessing the files that store the data and indices
  - management of buffers
  - management of disk space

A “semi-structured” example ...

- An xml dataset scraped from imdb.com
- No schema required, no instance made
- We can immediately write queries on the data
- A much *looser* approach

*imdb-subset.xml

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<movies>
<movie>
   <title>Club Sandwich</title>
   <imdb_key>0150155</imdb_key>
   <year>1931</year>
   <rating votes="0">0</rating>
   <genre>Animation</genre>
   <genre>Short</genre>
   <keyword>al-falfa</keyword>
   <credits>
      <director>Frank Moser</director>
      <author>Paul Terry (I)</author>
   </credits>
   <country>USA</country>
   <languages>
      <language>English</language>
   </languages>
</movie>
```

A “semi-structured” example …

```
<books search-terms=“database+design”>
        <book>
                <title>Database Design for Mere Mortals </title>
                <author>Michael J. Hernandez</author>
                <date>13/03/2003 </date>
        </book>
        <book id=“B2” >
                <title>Beginning Database Design</title>
                <subtitle>From Novice to Professional</subtitle>
                <author>Clare Churcher</author>
        </book>
</books>
```

What this course is about

- csc443 is about implementation of the DBMS itself
- csc343 is about using DBMSs:
  - defining schemas and instances
  - writing queries
  - connecting to code written in a general-purpose language (e.g. Java!)
  - rigorous underlying principles

CSC343 - Administrative Info!

UNIVERSITY OF TORONTO

DCS50

Admin Stuff..

Important: Read the course syllabus

- Contact:
  - website and Piazza: required reading
  - your questions: to Piazza please
  - *personal* matters: email or visit me in O.H.
- Office hours:
  - Tuesdays 3-5pm
  - Room: BA 3219

Prerequisites

- For A&S students, the prerequisites are: (1)CSC165H1/CSC240H1/(MAT135H1, MAT136H1)/ MAT135Y1/MAT137Y1/MAT157Y1; (2)CSC207H1
- Prerequisite for Engineering students only: ECE345H1/ CSC190H1/CSC192H1

- Email me immediately if you don't have the prerequisites (nosayba@cs.toronto.edu).
  Include your unofficial ROSI transcript.

- Engineering students, contact me if you need permission.

Active lectures (kind of..)

- Goal: get your gears turning in class!
- Activities like:
  - team problem solving, reviewing other students' solutions, and short quizzes.
- Weekly "lecture prep activities" will get you ready.
  - exercises, reading, watching videos
- All three hours will be here, with me.
  - Relax: some weeks will have tutorials delivered by TAs ;-)

Benefits of active learning

- Exercise your knowledge and skills in class, with support.
- We'll know where the difficulties are.
- Get more from when I'm lecturing.

What it requires

- Doing the lecture prep.
- Being active in class, including working with others and looking at each other's solutions to problems.
- A positive, encouraging environment.

Course Marking Scheme

Work	Weight	Comment
3 assignments	30%	10% each
weekly lecture prep	7%	due Sundays 11pm
weekly in-class exercise	3%	due in lectures
midterm	15%	Oct 27
Final exam	45%	You must get >= 40% in exam mark to pass the course

Recommended Resources

- Ullman and Widom,
  "A First Course in Database Systems", third edition.
- Jennifer Widom's online mini-courses from Stanford.

Assignment Policies

- You may work with a partner on assignments.
- Can be from any section on StGeorge campus.
- Can change partners between assignments.
- You may <u>not</u> dissolve a partnership without permission.
- Assignments must be submitted via <u>MarkUs</u>.
- Your code must run on our lab computers ("cdf").
- Late policy:
  - You have **6 grace tokens** that can be used for **2-hour** extension each.
  - No submission allowed after all tokens are exhausted

Your To-do list

- Anyone new to the cdf labs:
  - Find out your account on our cdf machines. See the course website for details.
  - Try logging in.
- Read the course syllabus.
- Bookmark the course website.
- Do the class prep due Sunday night.