CS 4400 Database Systems

Meeting 1: Introduction
Brandon Myers
University of Iowa
One of my research projects

```sql
select count(*), mass from Particles
where x > 0 and x < 10
group by mass/10
```

Why is 4400 important?

Data management is critical in business and science.
eCommerce business

Digital media content

Warehouse operations

Website

Users

Stock

Orders

Analysts predicting demand for Harry Potter

“Users who viewed this item also liked...”

An internet full of shoppers
Astronomy

Application Layer - framework-based pipelines process raw data to products

Middleware API

Application Framework

Data Acquisition Infrastructure

Alert Production

Data Release Production

Calibration Products Production

Observatory Control System

Image Process

Detection

Association

Moving Object

Alert Processing

Moving Object

Deep Detection

Photometric Calib

Astrometric Calib

Image Coadd

Classification

Science Data QA

Illumination Correction

Pupil Ghost Images

Crosstalk Matrix

Flats & Bias

Fringe Images

User Tools

Query, Data Quality Analysis, Monitoring

Data Products

Eng/Facility Data Archive

Image Archive

Source Catalog

Object Catalog

Orbit Catalog

Alert Archive

Calibration Data Products

database

https://commons.wikimedia.org/wiki/File:Large_Synoptic_Survey_Telescope_profile_render_2013.png

https://www.lsst.org/about/dm/pipelines
Administrivia

• **Course website:**
  • [http://homepage.cs.uiowa.edu/~bdmyers/cs4400_fa16/](http://homepage.cs.uiowa.edu/~bdmyers/cs4400_fa16/)
  • Go read the syllabus; I’ll talk about parts of it but not all the details
  • Ask questions if anything is unclear
Administrivia

• Instructor: Brandon Myers
  room: 201K MLH

• Debug your brain (DYB): 4-5pm Thursdays
  • Show up even if you don’t think you have any questions!
  • Come work with peers, we’ll go through practice problems, etc
  • Dedicated to 4400

• Free-style office hour: 11am-12 Wednesdays
  • Also can show up 4-5pm Tuesday, but that time is first dedicated to CS2630 students DYB

• By appointment

• E-mail: bdmyers@uiowa.edu
...continued

• Teaching assistant: maybe

• How to ask questions:
  • It is a personal or private question: in person or by “email” (try to use ICON’s inbox rather than regular email)
  • It is a non-personal question that others will benefit from: on the ICON Discussions board, in class, or in DYB/OHs
    • If you email me a non-personal question then I’ll tell you to post it on Discussions
ICON Discussions

• Use the ICON discussions board: post replies, post your own topics, as long as it is related to 4400
  • Teach each other. Your peers might answer the question first
  • Of course, academic integrity policy applies. Intellectual arguments are great. Posting answers or critical "ahas" is not okay. If in doubt, ask the staff
Textbook

Late policy

• 4 slip days
  • Meant to replace other excuses for not turning in assignments on time
  • No fractional slip days
  • Use them carefully: you don’t want to fall behind!

• Besides that, late work is 50% of earned points up to one day late and 0% afterwards

• Tell the course staff before the deadline if possible about longer term issues
Be successful in CS4400

• There are ~8 homeworks, many involving programming
• Come to class; reading slides is a poor substitute
• Active learning, peer teaching, and other activities to replace the lack of labs/discussions
  • This is a small class, which makes it easier to have group activities!
• Attend Debug Your Brain and/or office hours
• Help your classmates in class on ICON
• Midterm in class (mid October, exact date TBA)
• Keep on top of announcements in ICON/website
Peer instruction

• Think, answer, discuss...
• Participation counts, *not* right answers

• Some class meetings involve use of your computer
  (need at least 1 per pair of students)
  • If bringing a laptop presents a hardship, email me
Databases and Database management systems (DBMS)

• Examples of databases

• Examples of DBMSs
An example: online music streaming service

• What data must it contain?

• What capabilities are needed?
Summary of data management requirements

1. Able to describe real-world entities in terms of stored data
2. Persistently store large datasets
3. Efficiently query & update
4. Change structure (e.g., add attributes)
5. Concurrency control: enable simultaneous updates
6. Crash recovery
7. Security and integrity, provenance

DBMS provides these so that users can focus on application logic
Data structures and databases

• In CS2310 (or equivalent) was all about data structures
• What is the difference between a database and a data structure?
Data warehouses to data lakes

- Conventionally, businesses would have:
  1. Business operations supported by: a DBMS for transactions (e.g., sales, supply chain orders)
  2. Business intelligence supported by: a DBMS for storing a structured and indexed archive of recent and historical data (think library) called a data warehouse. Employees analyzed the data to inform decisions.

- Today, companies like Microsoft refer to data lakes, replacing the carefully maintained databases of a data warehouse with enormous quantities of raw data
- When the data needs to be analyzed, it is transformed with parallel processing systems
- in 4400 we’ll explore XML (semi-structured data), parallel processing, and non-relational systems (“NoSQL”)
People and databases

1. App developer: writes programs that update and query the data in the DB
2. DB designer: models the data by choosing tables and their attributes
3. DB admin (“DBA”): operates the database, diagnoses performance problems
4. Data analyst: data mining (inferring useful information), data integration (combining disparate data)
5. DBMS implementer: builds the DBMS

In 4400 we’ll try to give you some experience in all of these roles, although 4 and 5 are huge topics that demand their own courses
Rough schedule of CS4400

• ~2.5 weeks: How do we query *relational databases*\(^1\)?
• ~2.5 weeks: How does a relational database management system work?
• ~1 week: What is the mathematical logic behind relational databases queries?
• ~1 week: Semi-structured (non-table) data, specifically XML
• ~2.5 weeks: How do you design a good database?
• ~1.5 weeks: How do DBMSs support updates (transactions) from many concurrent users?
• ~4 weeks: Parallel processing and beyond relational databases

1. Databases where data is stored as tables with rows and columns
What to do now

• HW 1 is out later in the week
  • To get a head start, install and test sqlite (on the Resources page) of the website (not ICON)
• Read the syllabus online
• Check ICON and reply to the discussion question
Attribution

• Some slides inspired or quoted from UW CSE 344
  • https://courses.cs.washington.edu/courses/cse344/