# CSE 444: Database Internals

## Lecture 1

## Introduction

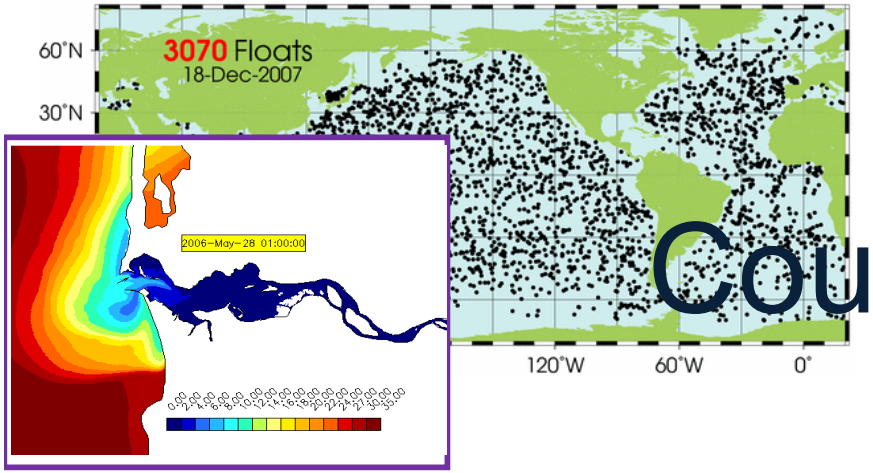
Course overload form: **http://tinyurl.com/hz9sxzd**

Code word is on the board (do not share the code word)

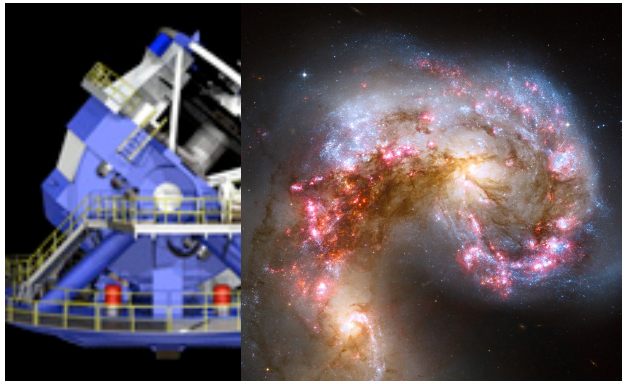
# Course Staff

- Instructor: Magdalena (Magda) Balazinska
  - magda@cs.washington.edu OH: Tuesday 4:30-5:20pm

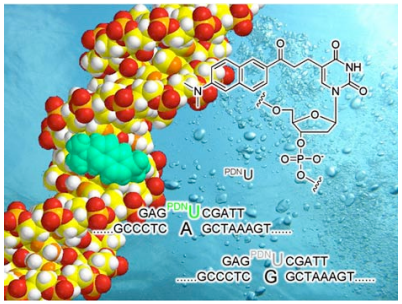
- TAs:
  - Jennifer Ortiz
  - Connor Moore
  - Chadi Moussi
  - Email addresses and office hour times and locations are on the course website
    - Every day one of us has office hours or there are sections




# Course Goals



- The world is drowning in data!
- Need computer scientists to help manage this data
  - Help domain scientists achieve new discoveries
  - Help companies provide better services (e.g. Facebook)
  - Help governments become more efficient
- This class: **principles of building data mgmt systems**
  - Learn how classical DBMSs are built
  - Learn key principles and techniques
  - Get hands-on experience building a (parallel) DBMS





# Course Format

- Lectures MWF, 12:30pm-1:20pm
- Sections: Th 9:30-10:20, 10:30-11:20
- Homeworks
  - 5 Labs + 6 Homeworks
- Quizzes:
  - 4 short quizzes in class

# Communication (part 1)

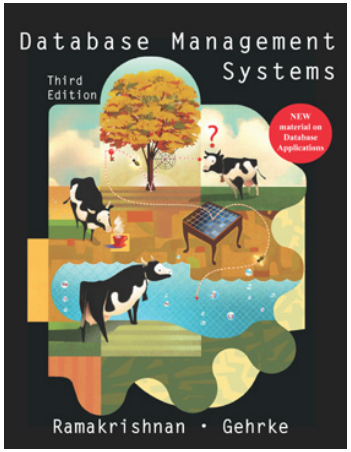
- Web page: http://www.cs.washington.edu/444
  - Lectures/Sections will be available there
  - Homeworks/Labs will be available there

- Mailing list
  - Announcements, group discussions
  - Your @uw.edu address is already subscribed

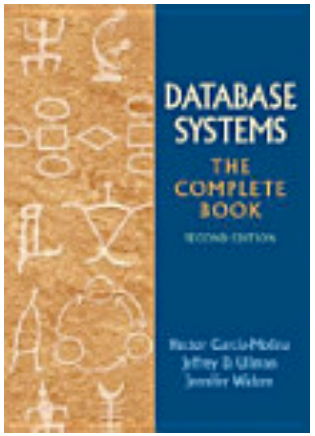
# Communication (part 2)

Message Board:

- Ask questions about the course, labs, homeworks
- Do not post any fragments of your code
- Do **not** send questions by email unless
  - You need to discuss a personal matter
  - You want to setup an appointment
  - A question has not been answered on the board




# Textbooks




## Recommended textbook (pick one)

- Database Management Systems. **Third Ed**. Ramakrishnan and Gehrke. McGraw-Hill.

- *Database Systems: The Complete Book*, Hector Garcia-Molina, Jeffrey Ullman, and Jennifer Widom. **Second edition**.

See course website for recommended chapters

# Other Readings

- See Website
- There is a section on reading assignments for 544M only

# Grading CSE444

- Labs:  40%
- Final project report 10%
- Six written assignments: 30%
- Four lab quizzes 20%

# Grading CSE 544M

- Same as CSE 444 plus
- Another 10% for the 4 paper reviews
- Then re-normalize to add up to 100%

- Graded separately from CSE 444

Acks: SimpleDB lab series originally developed by Prof. Sam Madden at MIT. We work with them on improving/extending.

# Six Labs

- Lab 1: Build a DBMS that can scan a relation on disk
  - **RELEASED! Part 1 of this lab is due on Friday!**
- Lab 2: Build a DBMS that can run simple SQL queries and also supports data updates

- Lab 3: Add a lock manager (transactions)
- Lab 4: Add a write-ahead log (transactions)

- Lab 5: Add a query optimizer
- Lab 6: Make your DBMS parallel (not this quarter)

Warning: I **will** run cheating-detecting software!

# About the Labs

Logistics:

- To be done **INDIVIDUALLY!**
- Each lab will take a **significant** amount of time
- Labs build on each other

Purpose

- Hands-on experience building a DBMS
- Deepen your understanding significantly
- We will build a *classical* DBMS
- In class we will discuss some *new-types* of DBMSs

# Six Homeworks

- Homework 1 has been released! Due next week
- Written assignments
- Help review material learned in class
- Prepare you for the labs
  - One homework before each corresponding lab
- Go beyond what we implement in labs

- To be done **INDIVIDUALLY**

# Quizzes

- One quiz in class for each of labs 1-4
- Test lab understanding
  - **No notes. No code. Answer from memory**
  - Likely only one or two open-ended questions
  - Example: “Explain how data is stored in SimpleDB”
  - Grades:
    - 10: Strength! Exceptional understanding and explanations
    - 9: You got it!
    - 6-8: Developing knowledge – some gaps
    - 0: Did not show up or wrote nothing

# Late Days

- Total of **4 late-days**
- Use in 24-hour chunks on hws or labs
- **At most 2 late-days per assignment**

- **No late-days can be applied to the final project due during finals week**

# Outline (this lecture and next)

- Review of DBMS goals and features
- Review of relational model
- Review of SQL

# Review: DBMS

- What is a database? Give examples
  - 
  - 
- What is a database management system? Give examples
  - 
  - 

# Review: DBMS

- What is a database? Give examples
  - A collection of related files
  - E.g. payroll, accounting, products
- What is a database management system? Give examples
  - A big C program written by someone else that manages the database; PostgreSQL, Oracle, …
  - In 444 you are that "someone else", implementing SimpleDB

# Review: Data Model

- What is a data model?
  - –
- What is the relational data model?

# Review: Data Model

- What is a data model?
  - A mathematical formalism for data
- What is the relational data model?
  - Data is stored in tables (aka relations)
  - Data is queried via relational queries
  - Queries are *set-at-a-time*

# Review: Transactions

- What is a transaction?
  - –
- What properties do transactions have?

# Review: Transactions

- What is a transaction?
  - A set of instructions that must be executed all or nothing
- What properties do transactions have?
  - ACID
  - Better: Serialization, recovery

# Review: Data Independence

The application should not be affected by changes of the physical storage of data

- Indexes
- Physical organization on disk
- Physical plans for accessing the data
- Parallelism: multicore, distributed

# Some Key Data Management Concepts

- Data models: Relational, XML, graph data (RDF)
- Schema v.s. Data
- Declarative query languages
  - Say what you want not how to get it
- Data independence
  - Physical: Can change how data is stored on disk without maintenance to applications
- Query compiler and optimizer
- Transactions: isolation and atomicity

# Course Content

**Focus: how to build a classical relational DBMS**

- Review of the relational model (lecture 1 and 2)
- DBMS architecture and deployments (lecture 3)
- Data storage, indexing, and buffer mgmt (lectures 4-6)
- Query evaluation (lectures 7-8)
- Query optimization (lectures 9-12)
- Transactions (lectures 13-20)
- Parallel query processing (lectures 21-24)
- Replication and distribution (lectures 25-26)
- Database as a service and NoSQL (lectures 27)

# Relational Model...

- Let's start our review of the relational model…
- We will continue next lecture