

COS 597A:

Principles of
Database and Information Systems

Professor Andrea LaPaugh

What makes a *database system*?

- **Large integrated** collection of data
- Uniform access/modification mechanisms
 - Rich **query** language
- **Model** of data organization
 - Levels of abstraction

Database systems ubiquitous
Behind many Web pages

What DB systems provide?

- Uniform **interface***
 - Uniform **models** of data*
 - Data **integrity**
 - Data **security**
 - Data **reliability**
 - **Concurrency**
 - **Efficiency**
- *like **abstract data types**
but **large**: disk vs memory

Is **overhead**

Some Current Database Models

- **Entity relationship** model
 - External “information” view
 - conceptual
- **Relational** model
 - Foundation of organization and access
- **XML** model
 - Semi-structured versus fully structured
 - Large amounts info within one element
 - Databases meet Web

Relational Model

Dominant DB model

- Formal underpinnings
- SQL most widely used DB language
 - ‘Q’ is for query

Historical staying power

Introduced 1970 by Edgar Codd

What his motivations?

How do they compare to modern concerns?

Flat model

vs older hierarchical and newer XML tree models

Levels of Abstraction

1. **Logical** (e.g. relational) model
 2. **Data organization**
 - indexing
 3. **Physical** model
 - File organization
 - File storage
- Determines access and manipulation methods

Database Algorithms

- Data **entry**
 - Index use
- Query **evaluation**
 - requests for data satisfying specified constraints
 - Efficiency
- Achieve **concurrency**
- Achieve **robustness**

Performance issues?

- **Efficiency of algorithms**
- **Large amounts** data
 - disk I/O!
- **Distributed** across **network**
 - Where is data?
 - Where should data be?

General Information Systems

- Semi-structured data
 - XML
- Unstructured data
 - No predefined structure **useful** to query/management system
 - Information retrieval systems
- properties share with database systems
 - Large **collection** of information
 - Desire uniform **access mechanisms**

Access mechanisms

A way to get at **specific parts** of the information.

A **query** is a request for data or information satisfying **specified constraints**

“all students taking Italian”
“information on small villages in Italy”

- What questions do you want to ask?
- Range of expectations
 - Query for information **know is** (or isn't) **there**
 - Query for info **will know when see it**
 - Predictability of results?
 - o “**Surprise me**” – *Data Mining*

How do you answer questions?

- **Models** of data/information
- **Correctness**

- In database systems, models of data and correct search well-defined
- In information retrieval, these #1 issues

Our syllabus Part 1: Models and Queries

- Structured Database models
 - The entity-relationship model
 - constraints
 - The relational model
 - Algebra, calculus and SQL
- Semi-structured and unstructured data
 - XML and the tree model
 - bridging database systems and IR systems
 - Information Retrieval

Our syllabus Part 2: Storing, Retrieving and Maintaining

- File Organization
- Indexing Methods
 - B+ Trees
 - Dynamic hashing
- Relational Query Evaluation
 - Optimization
- Transactions
- Crash recovery
 - logging

Our syllabus Part 3: Current Research

- advances in fundamentals and applications
- trend in research:

traditional databases



unstructured information retrieval systems

Graduate Focus

- Emphasize **fundamental models and methods**
 - expressiveness of languages
 - relationships through constraints
 - effectiveness and efficiency
- De-emphasize how use standard DB systems
 - still opportunity to do so

Graduate Focus

- **Explore interaction** with “other” research areas
 - research techniques applied to database/info systems
 - example: advanced data structures
 - example: caching in information systems
 - database/info system concepts applied to research
 - example: how integrate heterogeneous data sets in genomics
 - example: how structure data for network monitoring

Course logistics- overview

Web page has all: READ!!

<http://www.cs.princeton.edu/courses/archive/fall11/cos597A/>

- Texts
 - **Required: *Database Management Systems*** by Ramakrishnan and Gehrke, 3rd Edition, McGraw-Hill, 2003
 - reserved books in library
 - online readings
- 2 take-home tests (20% each)
- 5 problem sets (15%)
- Project (35%) – your choosing with approval
- Class Participation and oral presentation (10%)
- My office hours Mondays 4:30-5:30 or by appointment