

W4111  
Introduction to Databases  
Fall 2016  
Eugene Wu

Computer Science Department  
Columbia University

CS 4111-Introduction to Databases

1

Data

CS 4111-Introduction to Databases

2

Data  
is for serious business

CS 4111-Introduction to Databases

3

IT Executive

#### Deliver the Real-Time Enterprise

With growing data volumes and aggressive service level expectations, maximize the potential of your IT organization while delivering the real-time enterprise.



More for IT Executives >

→ Are you unlocking the full potential of your database infrastructure?

→ Are you unleashing the full potential of your database professionals?

→ Does your data seamlessly meet availability, security, and compliance requirements?

→ Does your database enable competitive business operations and analytics?

CS 4111-Introduction to Databases

4

Data  
is at the center of most things.

CS 4111-Introduction to Databases

5

Data  
is at the center of *everything*

CS 4111-Introduction to Databases

6



Betting worth billions. Elite players. Violent threats. Covert messages with Sicilian gamblers. And suspicious matches at Wimbledon. Leaked files expose match-fixing evidence that tennis authorities have kept secret for years.

CS 4111-Introduction to Databases

7



CS 4111-Introduction to Databases

8

**QUIZ**

Can You Guess The TV Show Based On The Wedding Dress?

This should be easy, considering we were all basically at their weddings.

Ali Velez · A half hour ago · 11 responses

"A K E I"

A BUZZFEED NEWS / BBC INVESTIGATION

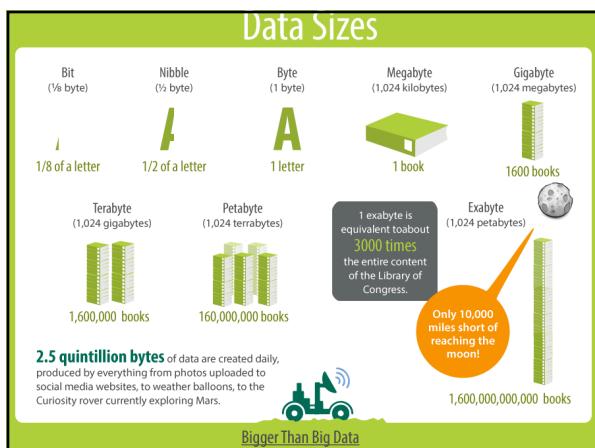
CS 4111-Introduction to Databases

9

The investigation into men's tennis by BuzzFeed News and the BBC is based on a cache of leaked documents from inside the sport – the Fixing Files – as well as an original analysis of the betting activity on 26,000 matches and interviews across three continents with gambling and match-fixing experts, tennis officials, and players.

CS 4111-Introduction to Databases

10



Produced by everything from photos uploaded to social media websites, to weather balloons, to the Curiosity rover currently exploring Mars.

How did we get here?

CS 4111-Introduction to Databases

12

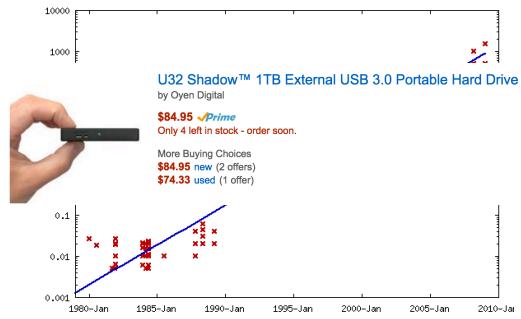
## Data was Expensive



CS 4111-Introduction to Databases

13

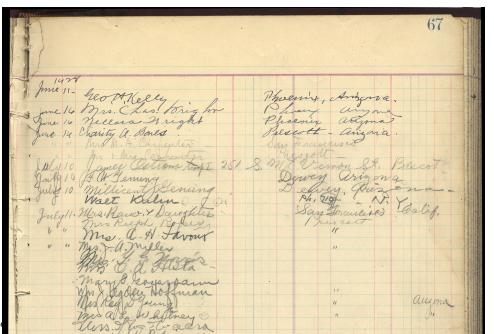
## Data is Cheap



CS 4111-Introduction to Databases

14

## Data was Manual



CS 4111-Intr

15

## Data is Automated

### Physical devices



CS 4111-Introduction to Databases

16

## Data is Automated

Physical devices  
Software logs



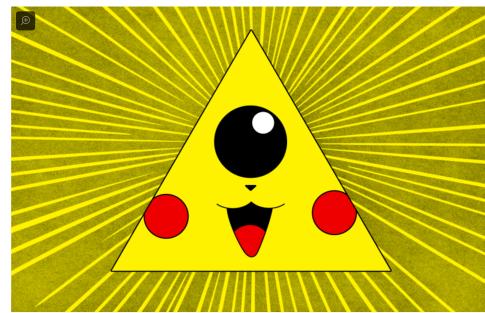
CS 4111-Introduction to Databases

17

## Pokémon Go Is a Government Surveillance Psyop Conspiracy

Ashley Feinberg  
7/17/16 3:00pm - Filed to: WAKE UP MAREEPLE

13M 260 71 0



CS 4 Illustration by Carter

18

## Data is *Ubiquitous*

Physical devices  
Software logs  
Phones  
GPS/Cars



CS 4111-Introduction to Databases

19

## Data is *Everywhere*

Physical devices  
Software logs  
Phones  
GPS/Cars  
Internet of Things



Leeo, a night light that listens for your smoke detector, then calls your smartphone.

CS 4111-Introduction to Databases

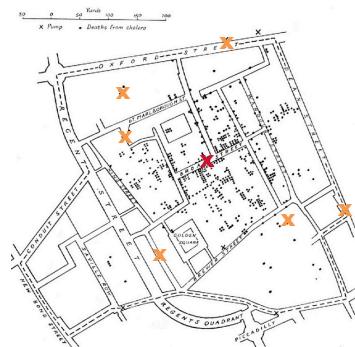
All this data, what are we doing with it?

CS 4111-Introduction to Databases

21

## What are we doing with data?

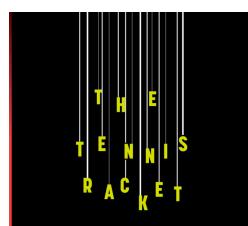
Health



CS 4111-Introduction to Databases

## What are we doing with data?

Health  
Investigative Journalism



CS 4111-Introduction to Databases

23

## What are we doing with data?

Health  
Investigative Journalism  
Recommendations



CS 4111-Introduction to Databases

24

## What are we doing with data?

Health



Forbes Tech

2 FREE Issues of Forbes

FEB 16, 2012 @ 11:02 AM 2,814,982 VIEWS

### How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill, FORBES STAFF  
Welcome to the Not-So Private Parts where technology & privacy collide

FOLLOW ON FORBES (2079)

Twitter Facebook RSS

Opinions expressed by Forbes Contributors are their own.

CS 4111-Introduction to Databases

READ MORE

25

## What are we doing with data?

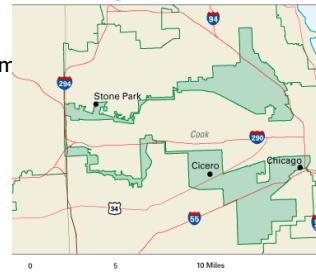
Health

Investigative Journalism

Recommendations

Politics

Congressional District 4



CS 4111-Introduction to Databases

26

## What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics



CS 4111-Introduction to Databases

27



DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

### The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 186,253 DATASETS

CS 4111-Introduction to Databases

28

## What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Surveillance

Every day, the NSA intercepts and stores 1.7 billion emails, phone calls, texts, and other electronic communications.



That's equivalent to 138 million books, every 24 hours.



CS 4111-Introduction to Databases

## What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Surveillance

Identity



HARVARD | BUSINESS | SCHOOL

30 APR 2012 RESEARCH & IDEAS

India's Ambitious National Identification Program

Comments (30) Email Print Download Share Recommend Share (62)

The Unique Identification Authority of India has been charged with implementing a nationwide program to register and assign a unique 12-digit ID to every Indian resident—some 1.2 billion people—by 2020. In a new case, Professor Tarun Khanna and HBS India Research Center Executive Director Anjali Raina discuss the complexities of this massive data management project.

"YOU ARE BASICALLY DENIED ALMOST EVERYTHING IF YOU CAN'T PROVE WHO YOU ARE."

CS 4111-Introduction to Databases

30

## What data?

CS 4111-Introduction to Databases

31

## What data?

Fake data



32

The screenshot shows a news article from The Verge. The headline reads "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day". The author is James Vincent, and the date is March 24, 2016. The article discusses how Microsoft's AI chatbot, Tay, learned discriminatory language from Twitter users. The page includes a sidebar with news items like "Apple is selling Microsoft Office 365 as an accessory for the iPad Pro" and "Siri and Alexa aren't speaking my language".

CS 4111-Introduction to Databases

33

## What data?

Fake data

Biased data

Incorrect data

CS 4111-Introduction to Databases

34

## What data?

Fake data

Biased data

Incorrect data

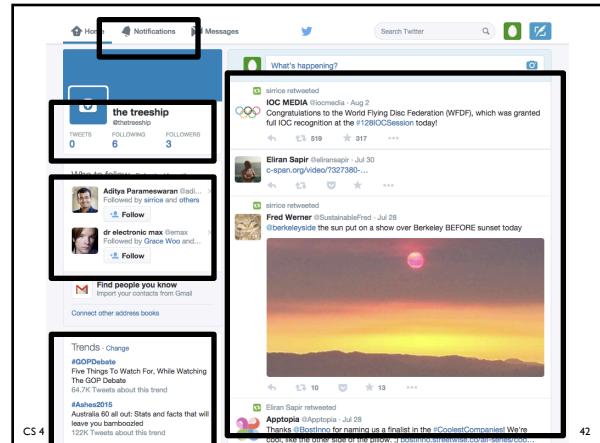
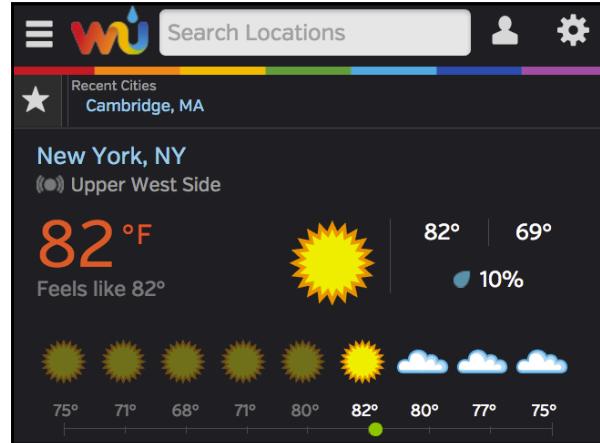
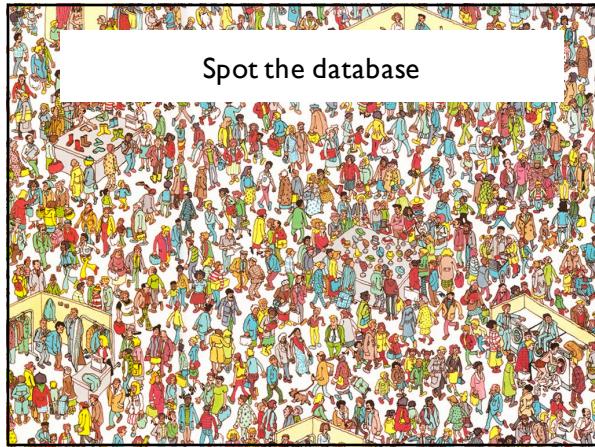
Mixed data

The screenshot shows the homepage of the Sunlight Foundation. The header features the foundation's logo and navigation links for BLOG, TOOLS, API, POLICY, ISSUES, PRESS, ABOUT, and CONTACT. There are also buttons for DONATE and JOIN. Below the header, a banner states "Making government & politics more accountable & transparent.". On the right side, there is a sidebar with sections for Reservation, SIDES, About, Menu, and Reviews. The menu lists "Fruit Plate" and "Patatas Bravas, Spicy-Tangy Sauce and Rosemary Aioli".

CS 4111-Introduction to Databases

36

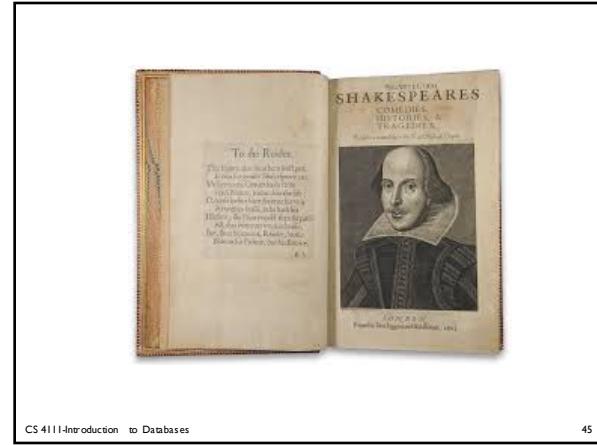
Data will be crucial to  
how we live  
as individuals and as a society



```

2012-01-04 00:01:23,180 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
010
2012-01-04 00:01:23,184 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-132476330017
2012-01-04 00:01:23,185 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResp
2012-01-04 00:01:23,291 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
010
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-132476330017
2012-01-04 00:01:23,293 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResp
2012-01-04 00:01:23,324 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
010
2012-01-04 00:01:23,326 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-132476330017
2012-01-04 00:01:23,327 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResp
2012-01-04 00:01:23,409 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
010
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-13247633008
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResp
2012-01-04 00:01:23,433 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
cliID: DFSClient_-2054881899, offset: 0, srvID: DS-292194659-127.0.1.1-50010-13247633008
2012-01-04 00:01:23,494 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
010
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
, cliID: DFSClient_-2054881899, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763308
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: PacketResp
2012-01-04 00:01:23,523 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
010

```



## What is a Database?

Structured data

CS 4111-Introduction to Databases

48

## What is a Database?

Lots of  
Structured data

CS 4111-Introduction to Databases

49

## Database Management System (DBMS)

A system to **store, manage** and **access** databases

CS 4111-Introduction to Databases

50

## Database Management System (DBMS)

System to **safely** and **reliably** store **lots of persistent** structured data and is **convenient** for **multiple users** to **efficiently** access and modify.

CS 4111-Introduction to Databases

51

## Is a program a DBMS?

Java/Python/Javascript etc  
Data stored in variables (RAM)  
Very fast access

CS 4111-Introduction to Databases

52

## Is a program a DBMS?

Java/Python/Javascript etc  
Data stored in variables (RAM)  
Very fast access

What about crashes? Restarts?

CS 4111-Introduction to Databases

53

## Is Excel a DBMS?

Visually access/modify/compute over data cells  
Click save to store persistently

CS 4111-Introduction to Databases

54

## Is Excel a DBMS?

Visually access/modify/compute over data cells  
Click save to store persistently

What about sharing? Huge tables? Multiple tables?

CS 4111-Introduction to Databases

55

## Is the file system a DBMS?

Manages files that are persistently stored on disk  
Open/read/seek/write access to files  
Access via file names  
Access control via permissions

CS 4111-Introduction to Databases

56

## Is the file system a DBMS?

You and a friend edit the same text file  
Save at the same time  
What happens?

1. Your changes survive
2. Friend's changes survive
3. Both changes survive
4. No changes survive
5.  $\backslash\backslash(\backslash)\backslash\backslash$

CS 4111-Introduction to Databases

57

## Is the file system a DBMS?

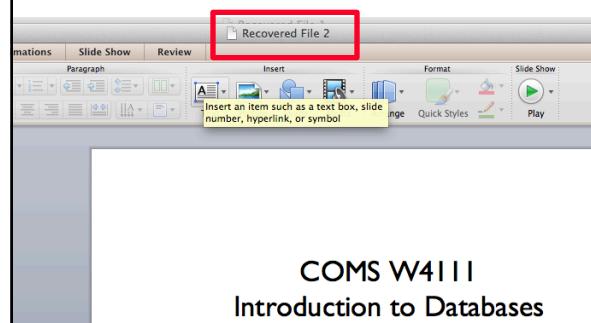
You edit a text file

Computer crashes

What happens?

1. All changes survive
2. No changes survive
3. Changes from last save survive
4.  $\backslash(\cup)\backslash$

## Is the file system a DBMS?



## Want Guarantees from DBMS

You want to write a hot new app on a DBMS.  
What do you *not* want to worry about?

**Failures** disk, machine, human, corruption  
**Lots of users**  
**Ad-hoc data access**  
**Data formats** csv? tsv? custom format?

## Database Management System (DBMS)

System to **safely** and **reliably** store **lots** of **persistent** structured data that is **convenient** for **multiple** users to **efficiently** access and modify.

## Database Management System (DBMS)

**Safe** Consistent and correct data after failures  
**Reliable** 99.99%+ Uptime  
**Lots** >>RAM (terabytes)  
**Persistent** Lives longer than one program/process  
**Convenient** Physical Independence, Declarative.  
**Multiple Users** Concurrent access. Access control.  
**Efficient** Fast: 100k+ queries / sec

## DBMSes in the Wild

### Classic Relational

\$\$: Oracle, IBM, Microsoft, Teradata, EMC, etc  
Free: MySQL, PostgreSQL, SQLite

### New Relational

In-Memory, Column-store, Streaming

### Non-traditional

Search (Google, Bing, Lucene), Scientific, Geographic

### NoSQL

Big Data: Hadoop, Spark, etc

Key-value: Mongo, BerkeleyDB, Cassandra, etc

### DBMS-as-a-Service

Microsoft Azure, Amazon Redshift/RDS, etc...

## Encompasses most of CS

OS	DBMS directly manages hardware
Languages	SQL is a domain specific language
Theory	Algorithms, models, NP-complete
AI/ML	Knowledge Discovery
Logic	Relational Algebra = 1 <sup>st</sup> order logic

CS 4111-Introduction to Databases

64

## Good time to learn!

Cloud programmer  
Data science  
Data engineer  
Machine learning engineer



### DATA Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil  
FROM THE OCTOBER 2012 ISSUE

65

## 2 Key Concepts

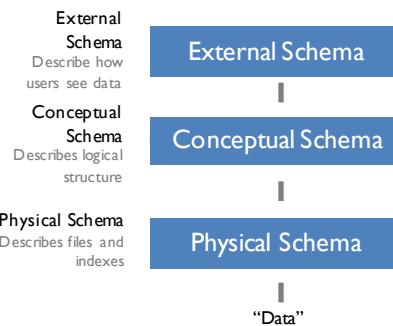
Data Independence  
Declarative Languages

Serve to insulate application programmers  
from the system implementation

CS 4111-Introduction to Databases

66

## Data Independence



CS 4111-Introduction to Databases

67

## Data Independence

UID	Name	Age
0	Eugene Wu	17
1	Luis Gravano	20
2	Ken Ross	30

0,Eugene Wu,17  
1,Luis Gravano,20  
2,Ken Ross,30  
CSV File

What is the number of adults?

CS 4111-Introduction to Databases

69

## Data Independence

UID	Name	Age
0	Eugene Wu	17
1	Luis Gravano	20
2	Ken Ross	30

0,Eugene Wu,17  
1,Luis Gravano,20  
2,Ken Ross,30  
CSV File

```
n = 0
for line in csv_file:
    attributes = line.split(",")
    if attributes[2] >= 18:
        n += 1
```

CS 4111-Introduction to Databases

70

## Data Independence

UID	Name	Age
0	Eugene Wu	17
1	Luis Gravano	20
2	Ken Ross	30

0,1,2  
Eugene Wu,Luis ...  
17,20,30  
CSV File

```
n = 0
for line in csv_file:
    attributes = line.split(",")
    if attributes[2] == 18:
        n += 1
```

CS 4111-Introduction to Databases

71

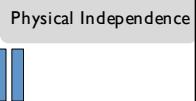
## Data Independence

Conceptual Schema  
Describes logical structure

Physical Schema  
Describes files and indexes

Users(uid int, name str, age int)

Conceptual Schema is the API!



CS 4111-Introduction to Databases

72

## Data Independence

Users(uid int, name str, age int)

“Welcome back Eugene Wu”

CS 4111-Introduction to Databases

73

## Data Independence

Users(uid int, fname str, lname str, age int)

“Welcome back Mr. Wu”

CS 4111-Introduction to Databases

74

## Data Independence

Conceptual Schema  
Describes logical structure

Users(uid int, name str, age int)

Physical Schema  
Describes files and indexes



Physical Independence

CS 4111-Introduction to Databases

75

## Data Independence

Conceptual Schema  
Describes logical structure

Users(uid int, fname str, lname str, age int)

Physical Schema  
Describes files and indexes

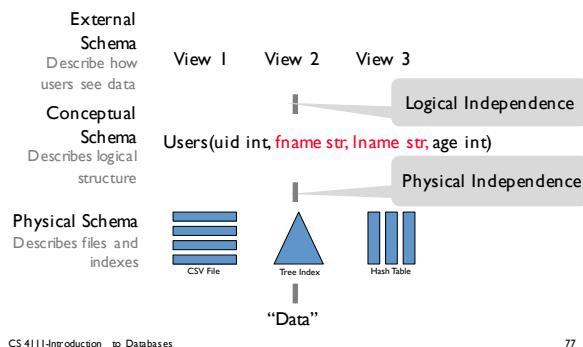


Physical Independence

CS 4111-Introduction to Databases

76

## Data Independence



## Data Independence

### Physical Independence

Protection from changes in physical structure of data

### Logical Independence

Protection from changes in logical structure of data

One of most important properties of a DBMS

CS 4111-Introduction to Databases

78

## Declarative

### What you want, not how to do it.

"Get me a sandwich"

"Take two slices of wheat bread out of the 2<sup>nd</sup> shelf, put them next to each other..."

Go to store and buy BLT  
Make PB&J  
Get Falafel delivered

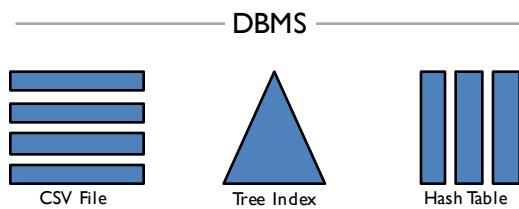
What if on 1<sup>st</sup> shelf?  
Out of wheat bread?  
No counter space?

CS 4111-Introduction to Databases

79

## Declarative

"I want W4111 instructors for Spring 2016"



CS 4111-Introduction to Databases

80

## Declarative

"I want W4111 instructors for Spring 2016"

DBMS



CS 4111-Introduction to Databases

81

## Classic Components in Databases

Durability (Recovery and Logs)

Transactions

Concurrency Control

Atomicity

CS 4111-Introduction to Databases

83

## Durability (Write-Ahead Log)

After crash: Want data to be accessible  
Idea: Write modifications to a *log*, then apply to data structures  
Each record allows redo/undo entire action

CS 4111-Introduction to Databases

84

## Transaction: Execution of a DB Program

Def: *atomic* (indivisible) sequence of DBMS actions

```
Begin;
<read Evan's account>
<deduct from Evan's account>
<increase Eugene's account>
Commit; (or Abort;)
```

CS 4111-Introduction to Databases

85

## Transaction: Execution of a DB Program

Def: *atomic* (indivisible) sequence of DBMS actions  
Each fully executed transaction must leave DB in *consistent state* if DB is consistent before transaction

CS 4111-Introduction to Databases

86

## Concurrency Control

Concurrently running multiple user programs needed for good performance  
Disks and networks are slow. Keep CPU working  
Use multiple CPUs  
**Concurrency can cause inconsistencies**

- e.g., check cleared while account balance being computed.
- Really hard to ensure correctness

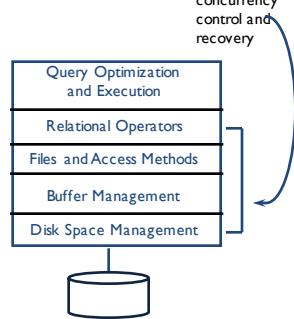
**DBMS hides concurrency:** write “single user” program

CS 4111-Introduction to Databases

87

## Classic Structure of a DBMS

Typical layered architecture  
DBMS, not OS, manages memory and disk  
Doesn't show concurrency control & recovery components



CS 4111-Introduction to Databases

90

## Database Courses at Columbia

CS 4111-Introduction to Databases

91

## **COMS W4111 - Intro to Databases**

Prerequisites: CS3137 or CS3134; fluency in Python

Intro to DBMSes

Data Models Entity-relation, Relational, ...

Relational Algebra

SQL

Applications + SQL cursors, APIs, embedded ...

Normalization

Peek at DBMS internals:

Storage and indexing

Query optimization

Transaction Processing

CS 4111-Introduction to Databases

92

## **COMS W4112-Database Sys. Impl.**

Storage Methods and Indexing

Query Processing and Optimization for INF Relations,  
including external sorting

Materialized Views and Use in Query Optimization

Query Processing and Optimization for ORDBMSs

Transaction Processing and Recovery

Parallel & Distributed DBMSes: Query Proc. and  
Optimization

Parallel and Distributed Databases: Transaction Processing

Performance Considerations Beyond I/Os

CS 4111-Introduction to Databases

93

## **COMS E6111-Advanced Databases**

Prerequisites: CS4111; fluency in Java or Python

Information Retrieval

Web Search

Distributed Information Retrieval and Web Search

Data Mining

Data Warehousing OLAP, Decision Support

Information Extraction

Scalable Visualization and Interaction

Supporting data analysis

Exploration, explanation and exhibition techniques

CS 4111-Introduction to Databases

94

## **Administrivia**

CS 4111-Introduction to Databases

95

## **Next Up**

Set up environment and perform data analysis

HW0 is out.

Due by Monday 10AM sharp

CS 4111-Introduction to Databases

96

## **Your Instructor: Eugene Wu**

B.S. @U.C. Berkeley

Ph.D. @MIT

PostDoc @U.C. Berkeley

Assistant Professor since Fall 2015

Databases, visualization, data analysis  
data cleaning, crowdsourcing

CS 4111-Introduction to Databases

97

## Your Instructor: **Eugene Wu**

### Contact

eugenewu.net  
ewu@cs.columbia.edu  
421 Mudd

### Office hours

Weds after class  
By appointment

## Class Resources

### Class web page

<http://www.cs.columbia.edu/~coms4111>

### Discussion board

piazza (linked from courseworks/website)

### Announcements from class staff:

Piazza

## Your TAs

Varun Jagdish Shetty  
Mengqi Wang  
Aayush Mudgal  
Qi Wang

All TA office hours in CS TA Room (see class web page)  
TA office hours will be posted on class web page

## Class Information: Prerequisites

COMS W3134 - *Data Structures in Java* or  
COMS W3137 - *Data Structures and Algorithms*  
(equivalent courses taken elsewhere are acceptable as well)

### Fluency in **Python**

You need permission from the instructor if you don't have the prerequisites.

## Grading Information

Midterm: 15%  
Midterm 2: 40% (last day of class)  
HW: 15% (4 HWs equally weighted)  
Project 1: 15%  
Project 2: 5%  
Scribe notes: up to or greater than 5% extra credit

Median grade: B+ or slightly higher.  
Alternative or make-up exams will not be given.  
All homework assignments are equally weighted.  
Project 1 has higher weight than Project 2.

## Exam Dates

Midterm 1: in class

Midterm 2: last day of class, in class

If you cannot make the midterms,  
do not take this course

## Homework

Homeworks usually due at 10AM of due date (before class)  
**No extensions or exceptions.**

**Three grace late days** for hw's throughout the semester.

After using all grace days, **25% grade deduction** per late day.

**Check full details on web site.**

CS 4111-Introduction to Databases

105

## Projects (more details soon)

### Two projects.

Teams of two

Get CS account if your team doesn't have a computer

Language is Python

#### Project 1

Model and build your own database web application  
Explore "traditional" relational database features.

#### Project 2

TBD

CS 4111-Introduction to Databases

106

## Projects (cont.)

**No extensions or exceptions** for project submission.

**3 grace late days total** for project.

After using all grace days, **25% grade deduction** per late day.

**Check full details on web site.**

CS 4111-Introduction to Databases

107

## Scribe Notes

### W4111 Scribe Notes

The goal of these scribe notes is to eventually create a document that can replace and surpass the expensive textbook. These notes are meant to supplement the lecture slides, which do not include detailed information nor full examples, and address the issue that the same questions are repeatedly asked on Piazza.

<https://github.com/w4111/scribenotes/wiki>

CS 4111-Introduction to Databases

108

## Collaboration Policy

Read Syllabus on course site for allowed conduct

**CS Dept academic honesty policies**  
<http://www.cs.columbia.edu/education/honesty>

We will not tolerate *any* cheating

CS 4111-Introduction to Databases

109

## Collaboration Policy

Discussing lectures and course material strongly encouraged

Homework and exams are *individual*. No exceptions  
Any libraries or code however minor must be disclosed.

Projects are done in teams; no collaboration between teams.

Contact the instructor right away if you have any questions  
or are falling behind.

CS 4111-Introduction to Databases

110

## Textbook

Raghu Ramakrishnan, Johannes Gehrke: *Database Management Systems*, 3<sup>rd</sup> edition, McGraw-Hill, 2002

Available from

*Bookculture* bookstore 536 W. 112th St.

Online retailers

Upperclass-persons

CS 4111-Introduction to Databases

111

Slides borrow material from  
Prof. Gravano

Prof. Hellerstein & Franklin@Cal

Prof. Madden & Stonebraker@MIT

(and by transitivity Raghu Ramakrishnan and Johannes Gehrke)

CS 4111-Introduction to Databases

115

## Useful info

<http://www.cs.columbia.edu/~coms4111>

**DO HOMEWORK 0!**

CS 4111-Introduction to Databases

116