

W4111
Introduction to Databases
Fall 2015

Computer Science Department
Columbia University

CS4111-Introduction to Databases,EugeneWu

1

Data

CS4111-Introduction to Databases,EugeneWu

2

Data
is for serious business

CS4111-Introduction to Databases,EugeneWu

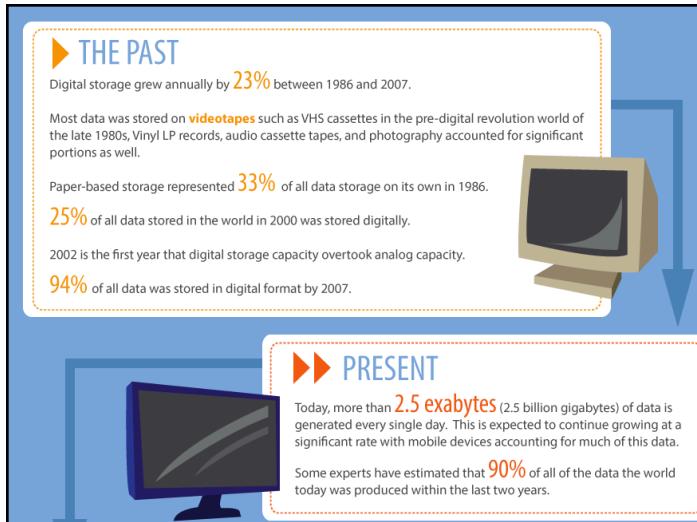
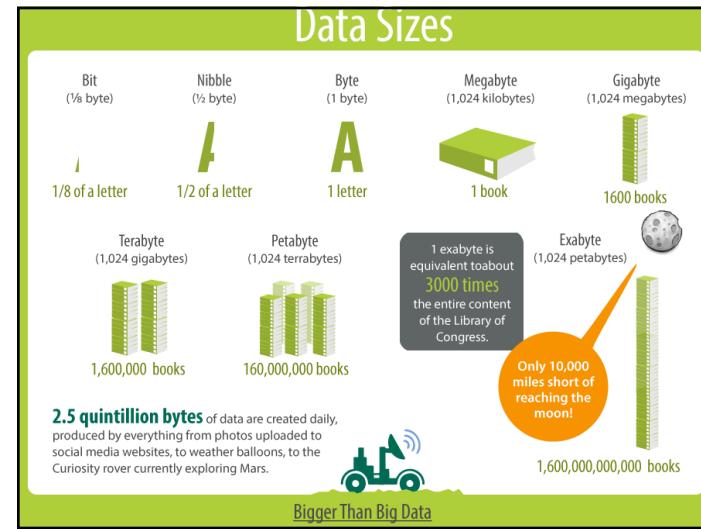
3

Data
is at the center of most things.

CS4111-Introduction to Databases,EugeneWu

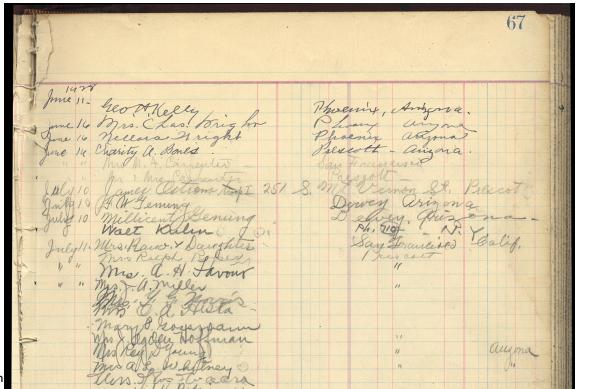
4

Data
is at the center of everything



How did we get here?

Data was Manual



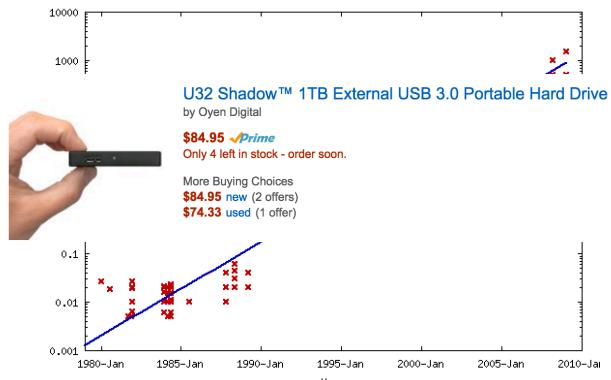
Data was Expensive



CS4111-Introduction to Databases, Eugene Wu

10

Data is Cheap



CS4111-Introduction to Databases, Eugene Wu

11

Data is Automated

Physical devices



CS4111-Introduction to Databases, Eugene Wu

12

Data is Automated

Physical devices
Software logs

CS4111-Introduction to Databases, Eugene Wu

13

Data is Ubiquitous

Physical devices
Software logs
Phones



CS4111-Introduction to Databases, Eugene Wu

14

Data is Ubiquitous

Physical devices
Software logs
Phones
GPS/Cars



CS4111-Introduction to Databases, Eugene Wu

15

Data is Everywhere

Physical devices
Software logs
Phones
GPS/Cars
Internet of Things



CS4111-Introduction to Databases, Eugene Wu

16

All this data, what are we doing with it?

CS4111-Introduction to Databases, Eugene Wu

17

What are we doing with data?

Health

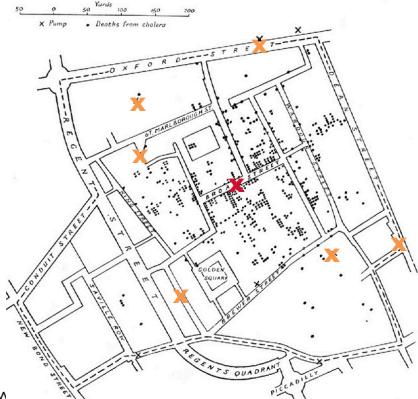


CS4111-Introduction to Databases, Eugene Wu

18

What are we doing with data?

Health



CS4111-Introduction to Databases, Eugene Wu

What are we doing with data?

Health

Investigative Journalism

PRO PUBLICA | Patient Safety

Surgeon Scorecard

by Sasi Wei, Olga Pierce and Marshall Allen, ProPublica, Updated July 15, 2015

Guided by experts, ProPublica calculated death and complication rates for surgeons performing one of eight elective procedures in Medicare, carefully adjusting for differences in patient health, age and hospital quality. Use this database to know more about a surgeon before your operation.

CS4111-Introduction to Databases, Eugene Wu

20

What are we doing with data?

Health
Investigative Journalism
Recommendations

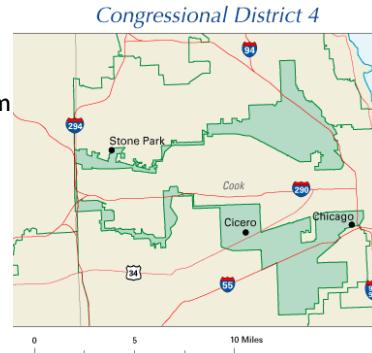


CS4111-Introduction to Databases, Eugene Wu

22

What are we doing with data?

Health
Investigative Journalism
Recommendations
Politics



CS4111-Introduction to Databases, Eugene Wu

24

What are we doing with data?

Health
Investigative Journalism

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



Kashmire Hill, FORBES STAFF

Welcome to The Not-So Private Parts where technology & privacy collide

FOLLOW ON FORBES (2079)

Opinions expressed by Forbes Contributors are their own.

CS4111-Introduction to Databases, Eugene Wu

23

What are we doing with data?

Health
Investigative Journalism
Recommendations
Politics

A thumbnail from TIME magazine's website featuring an article about the 2012 election and data crunchers who helped Obama win.

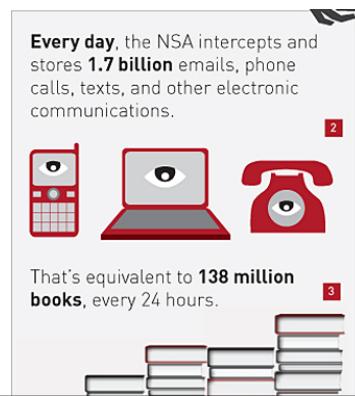
A thumbnail from the Mayday.us website, which is associated with the political action committee. It features the text "WE'RE HERE TO FIGHT" and "Join the movement".

CS4111-Introduction to Databases, Eugene Wu

25

What are we doing with data?

Health
Investigative Journalism
Recommendations
Politics
Surveillance

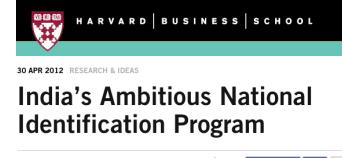


CS4111-Introduction to Databases, Eugene Wu

28

What are we doing with data?

Health
Investigative Journalism
Recommendations
Politics
Surveillance
Identity



30 APR 2012 RESEARCH & IDEAS

India's Ambitious National Identification Program

Comments 39 Email Print Download Share Recommend Share 92

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."

CS4111-Introduction to Databases, Eugene Wu

27

What data?

CS4111-Introduction to Databases, Eugene Wu

28

What data?

Fake data



CS4111-Introduction to Databases, Eugene Wu

29

What data?

Fake data
Biased data

CS4111 -Introduction to Databases, Eugene Wu

30

What data?

Fake data
Biased data
Incorrect data

CS4111 -Introduction to Databases, Eugene Wu

31

What data?

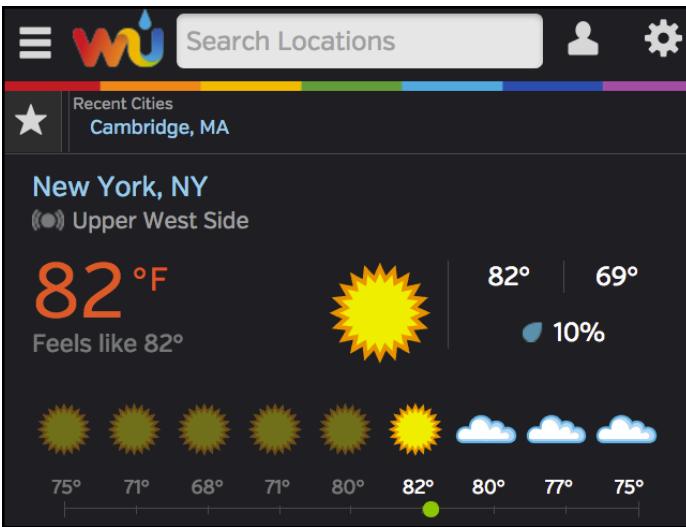
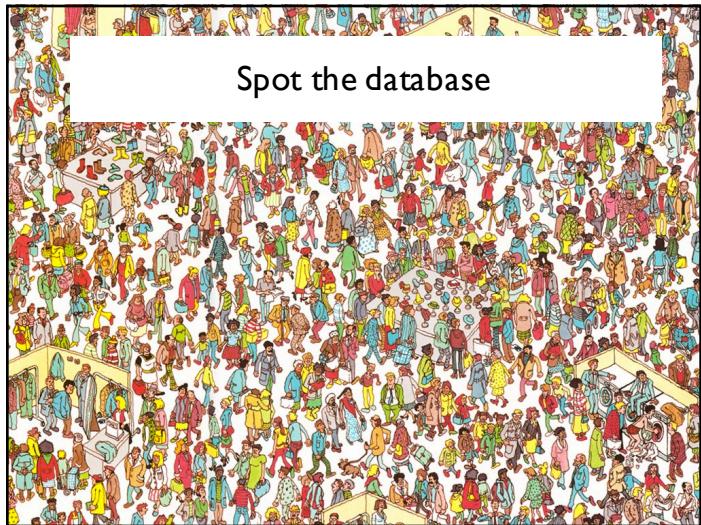
Fake data
Biased data
Incorrect data
Mixed data



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

CS4111 -Introduction to Databases, Eugene Wu

33

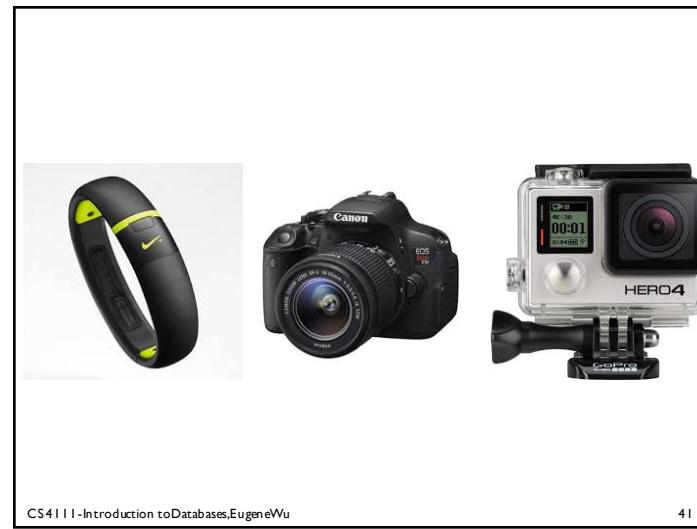


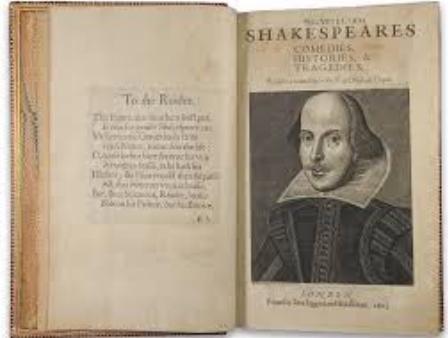


The image shows a screenshot of the Twitter mobile interface. At the top, there are three notification icons: Home, Notifications, and Messages. The Notifications icon is highlighted with a black box. The main feed shows several tweets, including one from "the treeship" (@thetreeship) and another from "IOC MEDIA" (@iocmedia). On the right side, there is a large image of a sunset over a body of water. At the bottom left is the slide navigation indicator "CS4111-Introduction toDatabases,EugeneWu" and at the bottom right is the page number "39".

```

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-292194659-127.0.1.1-50010-13247633001
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
10
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-292194659-127.0.1.1-50010-1324763300176
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
10
2012-01-04 00:01:23,411 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
, cliID: DFSClient_-603743753, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300
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_-2054881890, offset: 0, srvID: DS-292194659-127.0.1.1-50010-1324763300
2012-01-04 00:01:23,494 INFO org.apache.hadoop.hdfs.server.datanode.DataNode: Receiving
10
2012-01-04 00:01:23,498 INFO org.apache.hadoop.hdfs.server.datanode.DataNode.clienttrace
, cliID: DFSClient_-2054881890, offset: 0, srvID: DS-292194659-127.0.1.1-50010-132476330
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
  
```





CS4111-Introduction to Databases, Eugene Wu

42

What is a Database?

Structured data

CS4111-Introduction to Databases, Eugene Wu

45

What is a Database?

Lots of
Structured data

CS4111-Introduction to Databases, Eugene Wu

46

Database Management System (DBMS)

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

CS4111-Introduction to Databases, Eugene Wu

47

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.

CS4111-Introduction to Databases, Eugene Wu

48

Is a script a DBMS?

Java/Python/etc Script

Data stored in variables (RAM)

Very fast access

CS4111-Introduction to Databases, Eugene Wu

49

Is Excel a DBMS?

Microsoft office security

Visually access/modify/compute over data cells

Click save to store persistently

CS4111-Introduction to Databases, Eugene Wu

50

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

CS4111-Introduction to Databases, Eugene Wu

51

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. $\diagdown(\cup)\diagup$

CS4111 -Introduction toDatabases,EugeneWu

52

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. $\diagdown(\cup)\diagup$

CS4111 -Introduction toDatabases,EugeneWu

53

Is the file system a DBMS?

The screenshot shows a Microsoft Word interface. The ribbon is visible at the top, with the 'Insert' tab highlighted. A red box is drawn around the 'Recovered File 2' button in the ribbon. Below the ribbon, the slide content area contains the text 'COMS W4111 Introduction to Databases'.

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, deity
Lots of users
Ad-hoc data access
Data formats csv? tsv? custom format?

CS4111 -Introduction toDatabases,EugeneWu

55

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.

CS4111-Introduction to Databases, Eugene Wu

56

Database Management System (DBMS)

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

CS4111-Introduction to Databases, Eugene Wu

57

DBMSes in the Wild

Classic Relational

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

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...

CS4111-Introduction to Databases, Eugene Wu

58

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 st order logic

Scalable Computer Science

CS4111-Introduction to Databases, Eugene Wu

59

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

CS4111-Intro1

60

2 Key Concepts

Data Independence

Declarative Languages

Serve to insulate application programmers
from the system implementation

CS4111-Introduction to Databases, Eugene Wu

61

Data Independence

External Schema
Describe how users see data

External Schema

Conceptual Schema
Describes logical structure

Conceptual Schema

Physical Schema
Describes files and indexes

Physical Schema

“Data”

CS4111-Introduction to Databases, Eugene Wu

62

Example App: Guuber

Users(`uid int`, `name str`, `age int`)

Drivers(`did int`, `name str`)

Rides(`uid int`, `did int`, `distance float`, `drive_time float`)



CS4111-Introduction to Databases, Eugene Wu

63

Data Independence

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

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

What is the number of adults?

CS4111-Introduction to Databases, Eugene Wu

64

Data Independence

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

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

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

CS4111-Introduction to Databases, Eugene Wu

65

Data Independence

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

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

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

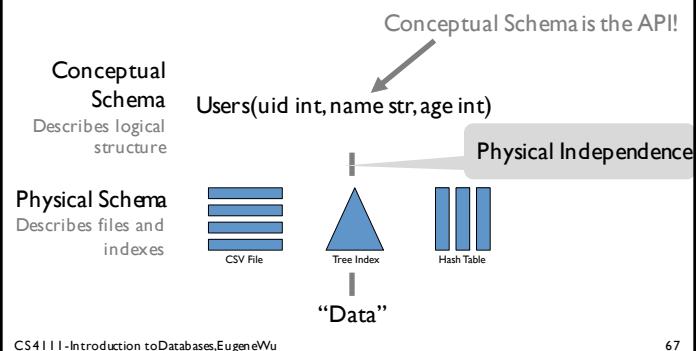
CS4111-Introduction to Databases, Eugene Wu

66

Data Independence

Conceptual Schema
Describes logical structure

Physical Schema
Describes files and indexes



67

Data Independence

Users(uid int, name str, age int)
Drivers(did int, name str)
Rides(uid int, did int, distance float, drive_time float)

“Welcome back Mr. Wu”

CS4111-Introduction to Databases, Eugene Wu

68

Data Independence

Users(uid int, **fname str, lname str**, age int)
Drivers(did int, name str)
Rides(uid int, did int, distance float, drive_time float)

“Welcome back Mr. Wu”

CS4111-Introduction to Databases, Eugene Wu

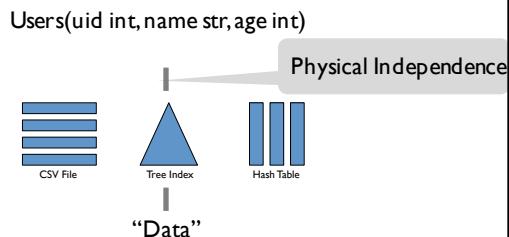
69

Data Independence

Conceptual Schema
Describes logical structure

Users(uid int, name str, age int)

Physical Schema
Describes files and indexes



CS4111-Introduction to Databases, Eugene Wu

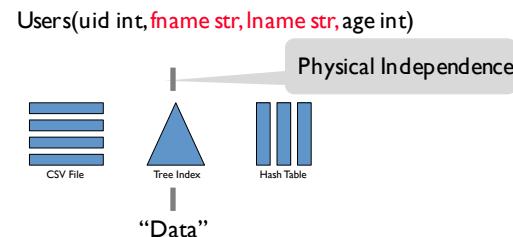
70

Data Independence

Conceptual Schema
Describes logical structure

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

Physical Schema
Describes files and indexes



CS4111-Introduction to Databases, Eugene Wu

71

Data Independence

External Schema

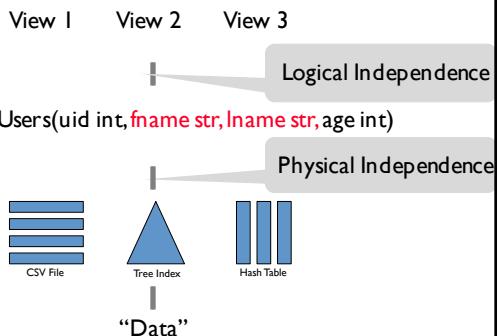
Describe how users see data

Conceptual Schema

Describes logical structure

Physical Schema

Describes files and indexes



CS4111 - Introduction to Databases, Eugene Wu

72

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

CS4111 - Introduction to Databases, Eugene Wu

73

Declarative

What you want, not how to do it.

"Make me a sandwich"

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

Buy from pb&j store
Make BLT
½ Tuna
Veggie

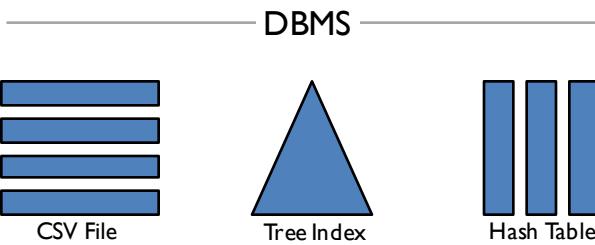
What if on 1st shelf?
Out of wheat bread?
No counter space?

CS4111 - Introduction to Databases, Eugene Wu

74

Declarative

"I want all highly rated fast drivers"



CS4111 - Introduction to Databases, Eugene Wu

75

Declarative

“I want all highly rated fast drivers”

DBMS



Declarative

“I want all highly rated fast drivers”

DBMS



Classic Components in Databases

Concurrency Control
Transactions
Atomicity
Recovery and Logs

Transaction: Execution of a DB Program

Def: **atomic** sequence of DBMS actions

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

Transaction: Execution of a DB Program

Def: **atomic** sequence of DBMS actions

Each fully executed transaction must leave DB in **consistent state** if DB is consistent before transaction

- Users specify simple **integrity constraints** on data, and DBMS enforces the constraints.
- DBMS does not understand semantics of its data e.g., doesn't know how bank interest is computed
- User's responsibility to ensure transaction (run alone) preserves consistency

CS4111-Introduction to Databases, Eugene Wu

80

Concurrency Control

Concurrently running multiple user programs needed for good performance

Disk accesses are frequent & slow. Keep CPU working on several user programs while waiting.

Concurrency can cause inconsistencies

- e.g., check cleared while account balance being computed.
- Really hard to program against

DBMS ensures such problems don't arise

- programmers can pretend to use a single-user system.

CS4111-Introduction to Databases, Eugene Wu

81

Scheduling Concurrent Transactions

Transactions T_1, \dots, T_n are run concurrently

Equivalent to a **serial** ordering (as if no concurrency)

Locks: T_i requests and waits for lock before read/write.
e.g., T_i locks the database, updates, then releases
e.g., T_i locks the table, updates, then releases
e.g., T_i locks rows, updates, then releases

Will talk about how this works later in course.

CS4111-Introduction to Databases, Eugene Wu

82

Atomicity

Def: Xact fully completes, or never happened even after failures e.g., crashes

Record all actions Xact did during execution in a log

1. **Write ahead logging:** before making any change, ensure the change is safely recorded in log
2. After failure, read log and undo any incomplete Xacts

CS4111-Introduction to Databases, Eugene Wu

83

The Log

A log record contains enough info to undo actions:

Transaction id

T_i writes an object: old and new values

Log record *must* be safely stored before the changed data

T_i commits/aborts: store commit/abort action

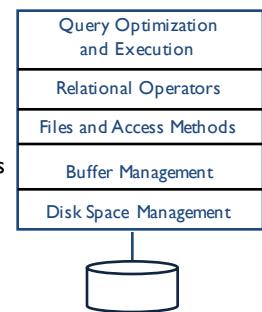
All logging, recovery and concurrency control activities hidden away from user.

Classic Structure of a DBMS

Typical layered architecture

DBMS, not OS, manages memory and disk

Doesn't show concurrency control & recovery components



Database Courses at Columbia

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

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

CS4111-Introduction to Databases,EugeneWu

88

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

CS4111-Introduction to Databases,EugeneWu

89

Administrivia

CS4111-Introduction to Databases,EugeneWu

90

Next Up

Set up your environment on the cloud

HW0 is out.
Due by Monday 9/14 10AM sharp.

Must be completed on time to stay in this class

CS4111-Introduction to Databases,EugeneWu

91

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.

CS4111-Introduction to Databases, Eugene Wu

92

Your Instructor: **Eugene Wu**

Contact

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

ewu@cs.columbia.edu

7LW1A Schapiro CEPSR

1-212-939-7088

Office hours

Weds 4-5PM

By appointment by email

CS4111-Introduction to Databases, Eugene Wu

93

Class Resources

Class web page

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

Discussion board

piazza (linked from courseworks/website)

Announcements from class staff:

CourseWorks → @columbia.edu email

CS4111-Introduction to Databases, Eugene Wu

94

Your TAs

Anthony Dubis

Mengqing Wang

Jincheng Li

Sania Arif

All TA office hours in CS TA Room (see class web page)

TA office hours will be posted on class web page

CS4111-Introduction to Databases, Eugene Wu

95

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.

Class Information: Lectures

Mondays and Wednesdays
2:30 - 4 p.m.
833 Mudd

(here)

Grading Information

Midterm: 15% (current est: 10/28)
Final : 40% (last day of class, 12/9)
HW: 15% (4 HWs equally weighed)
Project I: 15%
Project 2: 5%

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

Exam Dates

Midterm: 10/28, in class
Final: 12/9 last day of class, in class

If you cannot make the final exam,
do not take this course

Homework

Homeworks usually due at 10AM of due date.

No extensions or exceptions.

Three grace late days for hws throughout the semester.

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

Check full details on web site.

CS4111-Introduction to Databases, Eugene Wu

100

Projects (more details soon)

Two projects.

Teams of two

Run on Microsoft Azure cloud infrastructure

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

Language is Python; w/ a "minimal programming" option.

Project 1

Model and build your own database web application

Explore "traditional" relational database features.

Project 2

TBD

CS4111-Introduction to Databases, Eugene Wu

101

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.

CS4111-Introduction to Databases, Eugene Wu

102

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

CS4111-Introduction to Databases, Eugene Wu

103

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 rightaway if you have any questions
or are falling behind.

CS4111-Introduction to Databases, Eugene Wu

104

Textbook

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

Available from

Bookculture bookstore 536 W. 112th St.

Online retailers

Upperclass-persons

On reserve in Engineering Library

CS4111-Introduction to Databases, Eugene Wu

105

Contests and Rewards

“A+ letter-and-lunch” reward

Students who earn an A+ will get

Personal letter of congratulations from the instructor.

Lunch at Faculty House with instructor and A+ students

Top candidates for cs4111 TA positions in the future.

Project I contest

Four best projects chosen as contest winners.

Winners get:

Option to discuss and demo your project in class.

10% boost in your Project I grade.

CS4111-Introduction to Databases, Eugene Wu

106

On-going Feedback

Please provide feedback throughout the course.

- What is useful or confusing in lecture
- Thoughts about software stack
- Thoughts about assignments

Email me, come to office hours, talk to staff or:

CS4111-Introduction to Databases, Eugene Wu

107

On-going Feedback

Use form on website

Feedback form

Please share your comments and suggestions for the course!

* Required

Feedback *

Share what worked or what was confusing/difficult

Improvements

What change would you suggest to improve things?

CS4111-ir

108

Slides borrow material from

Prof. Gravano

Prof. Hellerstein & Franklin@Cal

Prof. Madden & Stonebraker@MIT

(and by transitivity Raghu Ramakrishnan and Johannes Gehrke)

CS4111-Introduction to Databases, Eugene Wu

109

Useful info

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

MW 2:30-4P, Seeley W. Mudd 833

ewu@cs.columbia.edu

DO HOMEWORK 0!

CS4111-Introduction to Databases, Eugene Wu

110