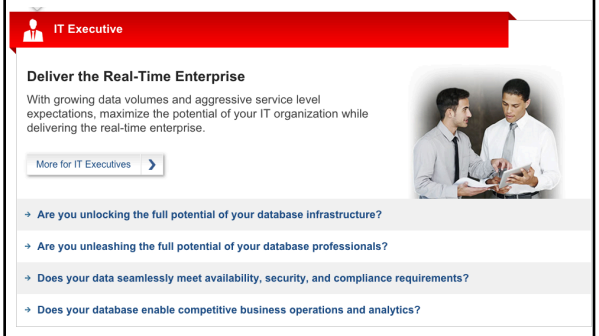
W4111
Introduction to Databases
Fall 2016
Eugene Wu

Computer Science Department
Columbia University

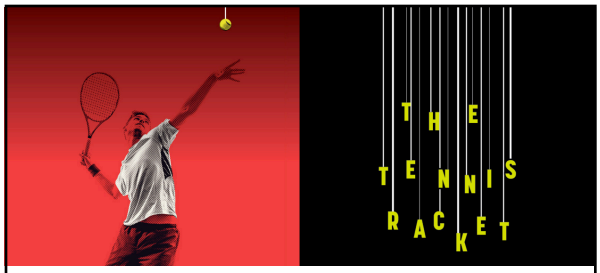
Data

Data
is for serious business

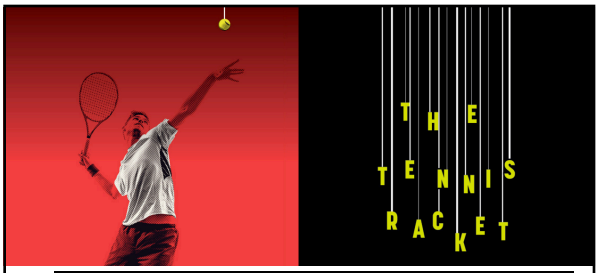


Data
is at the center of most things.

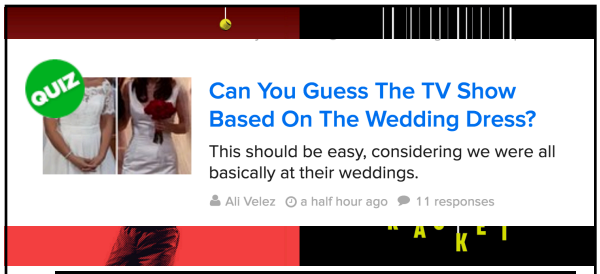
Data
is at the center of *everything*



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.



A BUZZFEED NEWS / BBC INVESTIGATION



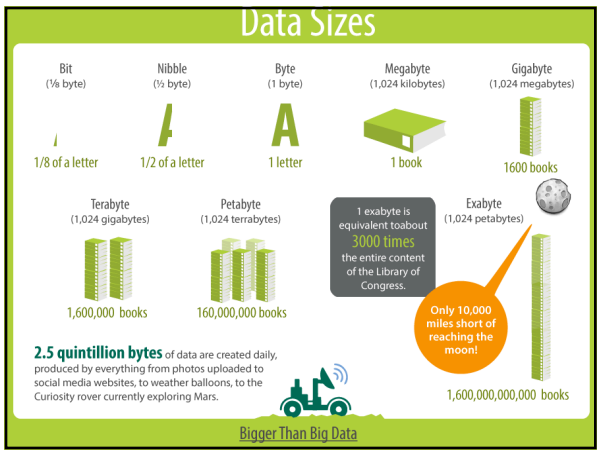
Can You Guess The TV Show Based On The Wedding Dress?

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

Ali Velez a half hour ago 11 responses

A BUZZFEED NEWS / BBC INVESTIGATION

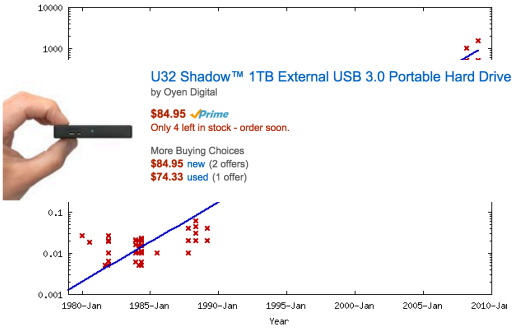
The investigation into men's tennis by BuzzFeed News and the BBC is based on a cache of leaked documents from inside the sport – the Fixing Files – as well as an original analysis of the betting activity on 26,000 matches and interviews across three continents with gambling and match-fixing experts,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



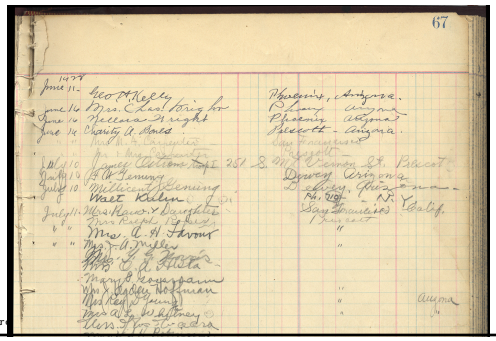
How did we get here?

Data was *Expensive*

Data is *Cheap*



Data was *Manual*

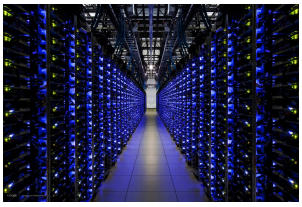


Data is *Automated*

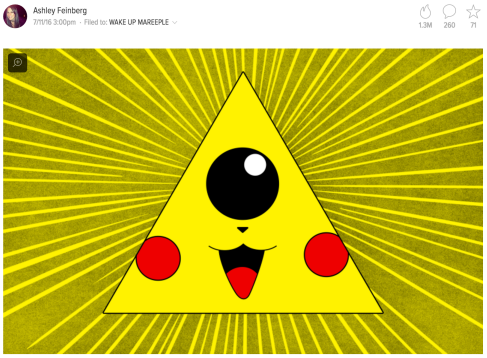
Physical devices

Data is *Automated*

Physical devices
Software logs



Pokémon Go Is a Government Surveillance Psyop Conspiracy



Data is *Ubiquitous*

Physical devices
Software logs
Phones
GPS/Cars

Data is *Everywhere*

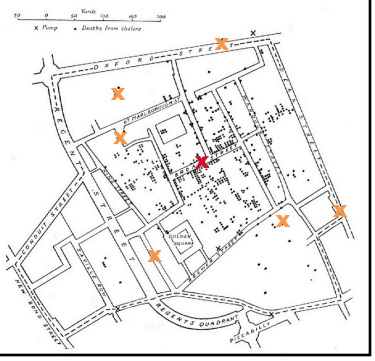
Physical devices
Software logs
Phones
GPS/Cars
Internet of *Things*

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

All this data, what are we doing with it?

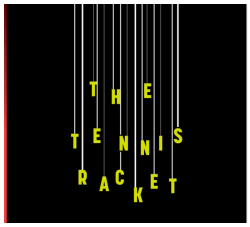
What are we doing with data?

Health



What are we doing with data?

Health
Investigative Journalism



What are we doing with data?

Health
Investigative Journalism
Recommendations

TARGET
NETFLIX
okcupid

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)

Opinions expressed by Forbes Contributors are their own.

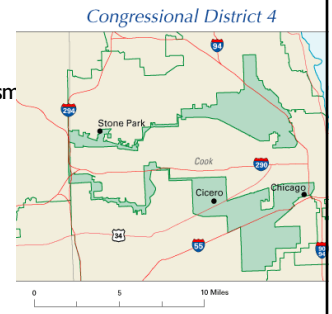
What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics



What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Surveillance

What are we doing with data?

Health

Investigative Journalism

Recommendations

Politics

Surveillance

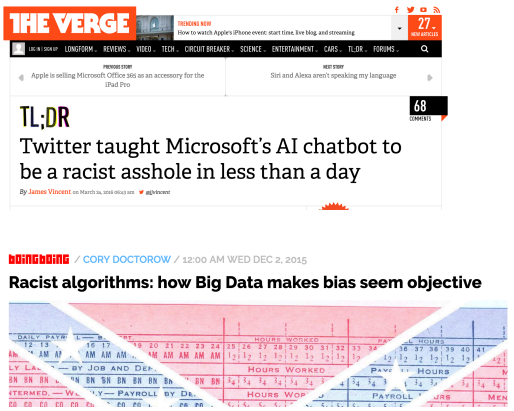
Identity

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

What data?

What data?

Fake data

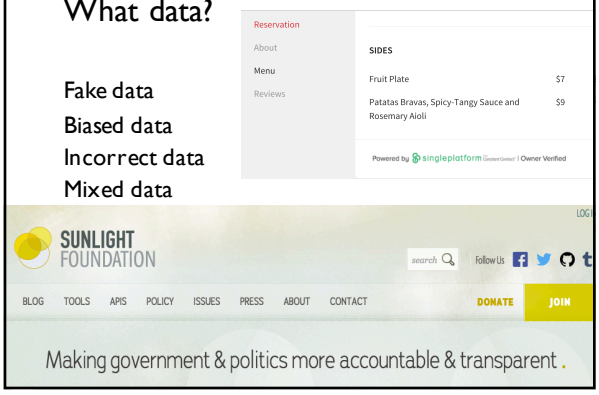


What data?

Fake data
Biased data
Incorrect data

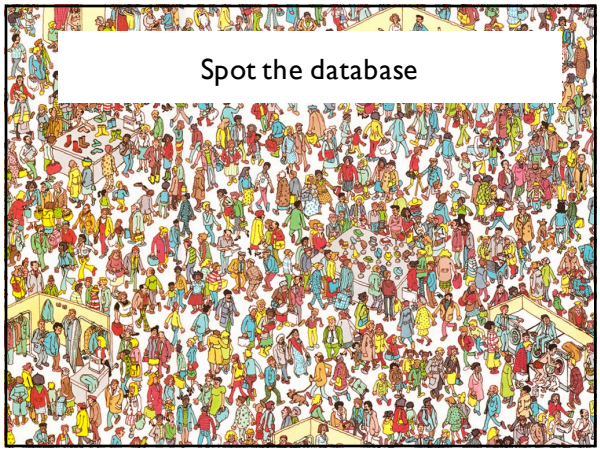
What data?

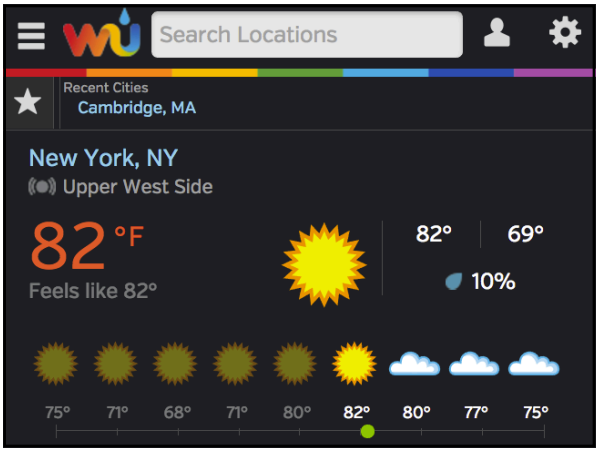
Fake data
Biased data
Incorrect data
Mixed data

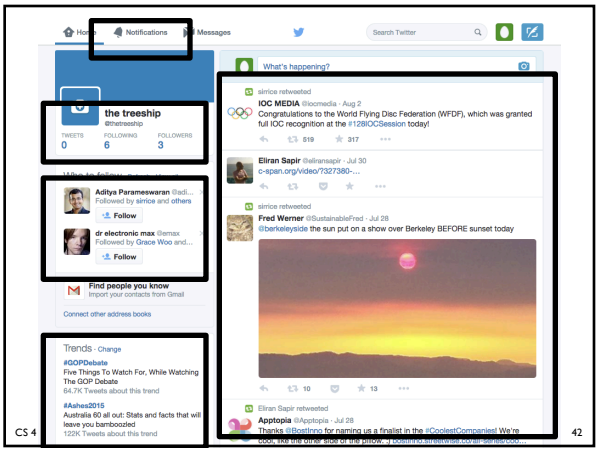


Data will be crucial to
how we live
as individuals and as a society

Spot the database







```
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
liID: 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
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
```

What is a Database?

Structured data

What is a Database?

Lots of

Structured data

Database Management System (DBMS)

A system to store, manage and access databases

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.

Is a program a DBMS?

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

Is a program a DBMS?

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

What about crashes? Restarts?

Is Excel a DBMS?

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

Is Excel a DBMS?

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

What about sharing? Huge tables? Multiple tables?

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

Is the file system a DBMS?

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

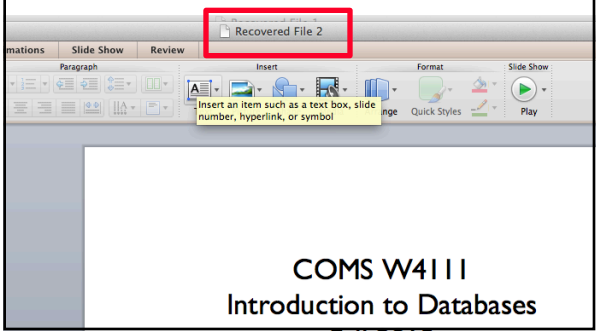
1. Your changes survive
2. Friend's changes survive
3. Both changes survive
4. No changes survive
5. ¯_(ツ)_/¯

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. ¯_(ツ)_/¯

Is the file system a DBMS?



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
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 = 1st order logic

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

2 Key Concepts

Data Independence

Declarative Languages

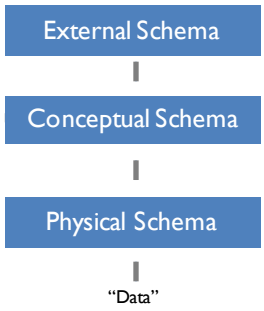
Serve to insulate application programmers from the system implementation

Data Independence

External Schema
Describe how users see data

Conceptual Schema
Describes logical structure

Physical Schema
Describes files and indexes



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?

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
```

Data Independence

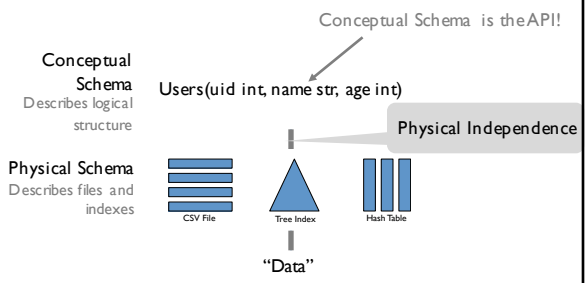
UID	Name	Age
0	Eugene Wu	17
1	Luis Gravano	20
2	Ken Ross	30

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
```


Data Independence



Data Independence

Users(uid int, name str, age int)

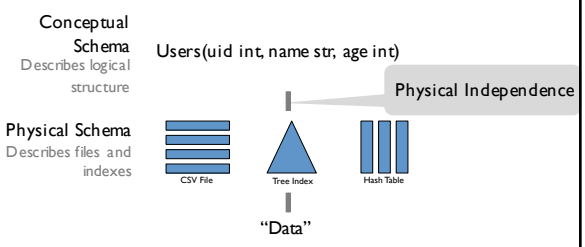
“Welcome back Eugene Wu”

Data Independence

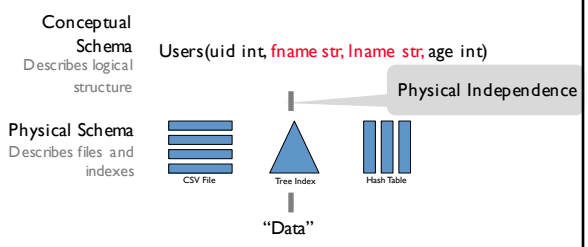
Users(uid int, fname str, lname str, age int)

“Welcome back Mr. Wu”

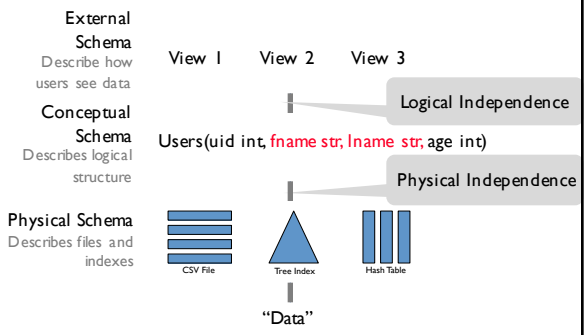
Data Independence



Data Independence



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

Declarative

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

"Get me a sandwich"

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

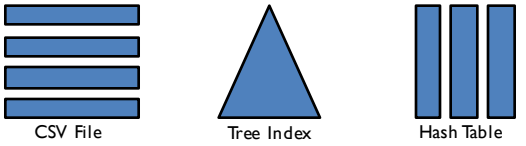
"Take two slices of wheat bread out of the 2nd shelf, put them next to each other…"

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

Declarative

"I want W4111 instructors for Spring 2016"

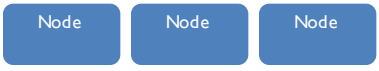
DBMS



Declarative

"I want W4111 instructors for Spring 2016"

DBMS



Classic Components in Databases

Durability (Recovery and Logs)
Transactions
Concurrency Control
Atomicity

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

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;)
```


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

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

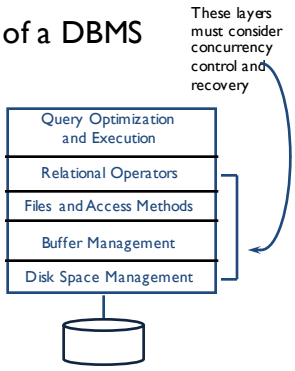
- e.g., check cleared while account balance being computed.
- *Really* hard to ensure correctness

DBMS hides concurrency: write "single user" program

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 1NF 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

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

Administrivia

Next Up

Set up environment and perform data analysis

HW0 is out.
Due by Monday 10AM sharp

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

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 weighed)
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 hws throughout the semester.

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

Check full details on web site.

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

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.

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

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

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.

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*

Slides borrow material from
Prof. Gravano
Prof. Hellerstein & Franklin@Cal
Prof. Madden & Stonebraker@MIT

(and by transitivity Raghu Ramakrishnan and Johannes Gehrke)

Useful info

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

DO HOMEWORK 0!