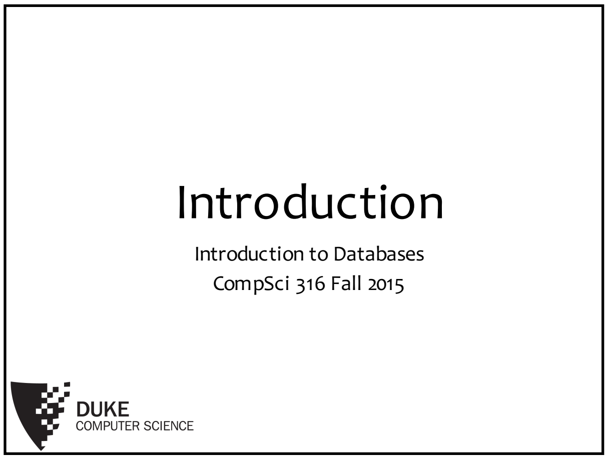
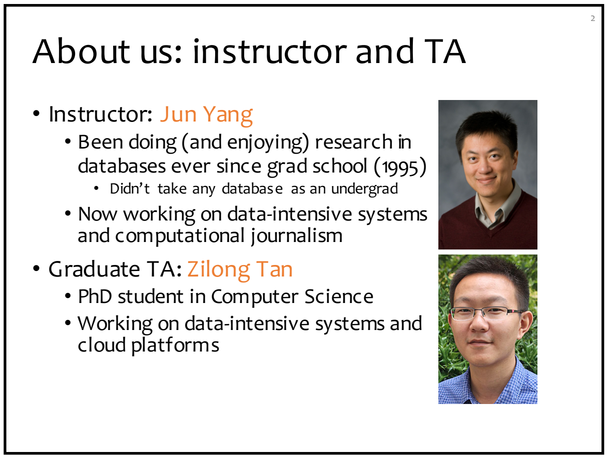


Introduction

Introduction to Databases
CompSci 316 Fall 2015

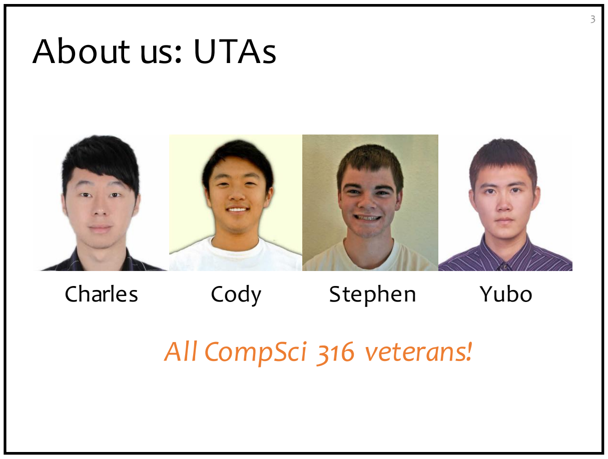
DUKE
COMPUTER SCIENCE



2

About us: instructor and TA

- Instructor: Jun Yang
  - Been doing (and enjoying) research in databases ever since grad school (1995)
    - Didn't take any database as an undergrad
  - Now working on data-intensive systems and computational journalism
- Graduate TA: Zilong Tan
  - PhD student in Computer Science
  - Working on data-intensive systems and cloud platforms



3

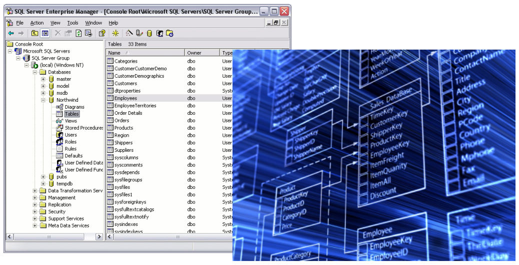
About us: UTAs

Charles Cody Stephen Yubo

All CompSci 316 veterans!

What comes to your mind…

…when you think about "databases"?



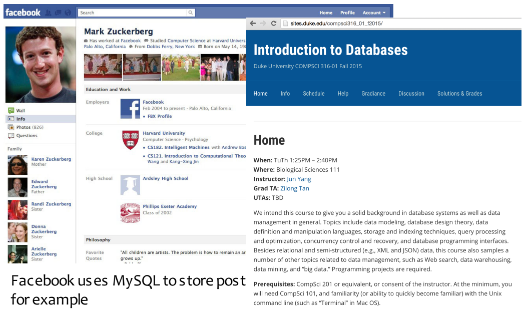
SQL Server Enterprise Manager - [Console Root\Microsoft SQL Servers\SQL Server Group…

Tables 33 Items

Name	Owner	Type
Categories	dbo	User
CustomerCustomerDemo	dbo	User
CustomerDemographics	dbo	User
Customers	dbo	User
dtproperties	dbo	Syst
Employees	dbo	User
EmployeeTerritories	dbo	User
Order Details	dbo	User
Orders	dbo	User
Products	dbo	User
Region	dbo	User
Shippers	dbo	User
Suppliers	dbo	User
syscolumns	dbo	Syst
syscomments	dbo	Syst
sysdepends	dbo	Syst
sysfilegroups	dbo	Syst
sysfiles	dbo	Syst
sysfiles1	dbo	Syst
sysforeignkeys	dbo	Syst
sysfulltextcatalogs	dbo	Syst
sysfulltextnotify	dbo	Syst
sysindexes	dbo	Syst
sysindexkeys	dbo	Syst

http://www.quackit.com/pix/database/tutorial/dbms_sql_server.gif
http://webstoresltd.com/wp-content/uploads/2013/06/database-design.jpg

But these use databases too…



Mark Zuckerberg

Has worked at Facebook · Studied Computer Science at Harvard Universi… Palo Alto, California · From Dobbs Ferry, New York · Born on May 14, 19…

Education and Work

Employers: Facebook — Feb 2004 to present · Palo Alto, California — FBX Profile

College: Harvard University — Computer Science · Psychology — CS182. Intelligent Machines with Andrew Bos… — CS121. Introduction to Computational Theo… Wang and Kang-Xing Jin

High School: Ardsley High School — Phillips Exeter Academy, Class of 2002

Family: Karen Zuckerberg, Mother; Edward Zuckerberg, Father; Randi Zuckerberg, Sister; Donna Zuckerberg, Sister; Arielle Zuckerberg, Sister

Philosophy

Favorite Quotes: "All children are artists. The problem is how to remain an ar… grows up."

sites.duke.edu/compsci316_01_f2015/

Introduction to Databases

Duke University COMPSCI 316-01 Fall 2015

Home

When: TuTh 1:25PM – 2:40PM
Where: Biological Sciences 111
Instructor: Jun Yang
Grad TA: Zilong Tan
UTAs: TBD

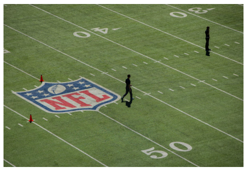
We intend this course to give you a solid background in database systems as well as data management in general. Topics include data modeling, database design theory, data definition and manipulation languages, storage and indexing techniques, query processing and optimization, concurrency control and recovery, and database programming interfaces. Besides relational and semi-structured (e.g., XML and JSON) data, this course also samples a number of other topics related to data management, such as Web search, data warehousing, data mining, and "big data." Programming projects are required.

Prerequisites: CompSci 201 or equivalent, or consent of the instructor. At the minimum, you will need CompSci 101, and familiarity (or ability to quickly become familiar) with the Unix command line (such as "Terminal" in Mac OS).

Facebook uses MySQL to store post for example

WordPress uses MySQL to manage components of a website (pages, links, menus, etc.)

And these…



In a data-driven N.F.L., the pings (12 per sec. per player) may soon outstrip the X's and O's

ASHLEY MADISON®
Life is short. Have an affair.®
Please Select
See Your Matches »
Over 32,430,000 anonymous members!

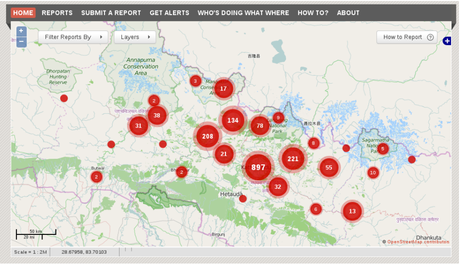
…P.S. To Ashley Madison's Development Team: You should be embarrased [sic] for your train wreck of a database (and obviously security), not sanitizing your phone numbers to your database is completely amateur, it's as if the entire site was made by Comp Sci 1XX students.

— Creators of CheckAshleyMadison.com

http://www.nytimes.com/2015/08/23/sports/football/in-a-data-driven-nfl-the-pings-may-soon-outstrip-the-xs-and-os.html
http://www.washingtonpost.com/news/the-intersect/wp/2015/08/19/how-to-see-if-you-or-your-spouse-appear-in-the-ashley-madison-leak/

7

And this…

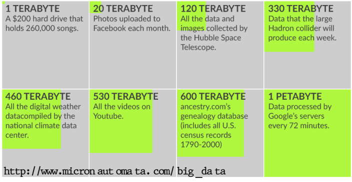


http://schoo lofdata.o rg/2015/0 6/15/how -open-map -data-is -helping- save-liv es-in-nep al/

8

Challenges

- Moore's Law:
  Processing power doubles every 18 months
- But amount of data doubles every 9 months
  - Disk sales (# of bits) doubles every 9 months
  - Parkinson's Law:
    Data expands to fill the space available for storage



1 TERABYTE A $200 hard drive that holds 260,000 songs.	**20 TERABYTE** Photos uploaded to Facebook each month.	**120 TERABYTE** All the data and images collected by the Hubble Space Telescope.	**330 TERABYTE** Data that the large Hadron collider will produce each week.
460 TERABYTE All the digital weather datacompiled by the national climate data center.	**530 TERABYTE** All the videos on Youtube.	**600 TERABYTE** ancestry.com's genealogy database (includes all U.S. census records 1790-2000)	**1 PETABYTE** Data processed by Google's servers every 72 minutes.

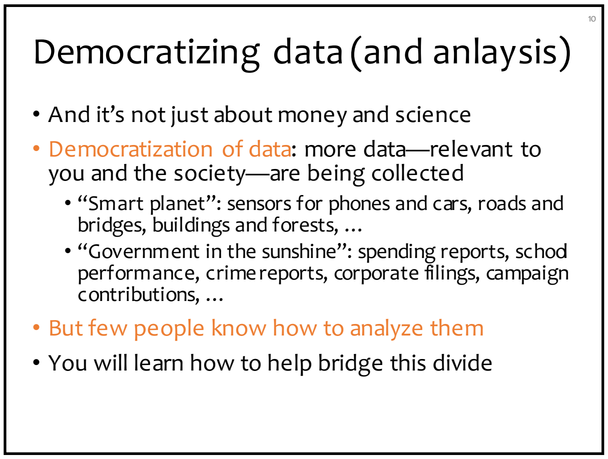
http://www.micron aut oma ta. com/ big_da ta

9

Moore's Law reversed

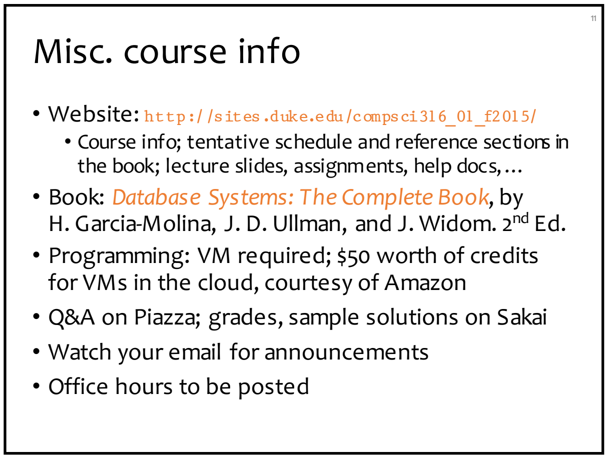
Time to process all data
_______ every 18 months!

- Does your attention span _____ every 18 months?
  - No, so we need smarter data management techniques



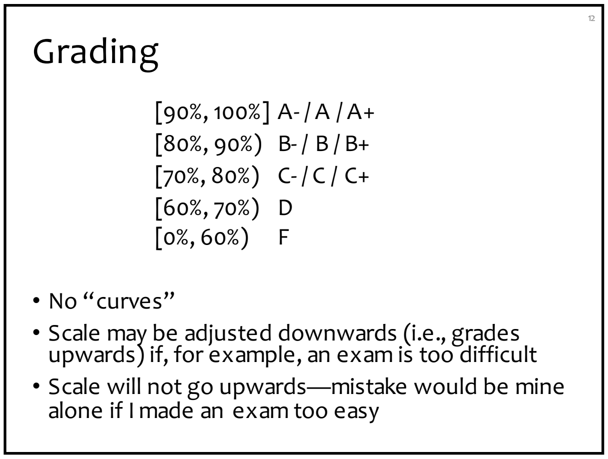
Democratizing data (and anlaysis)

- And it's not just about money and science
- Democratization of data: more data—relevant to you and the society—are being collected
  - "Smart planet": sensors for phones and cars, roads and bridges, buildings and forests, …
  - "Government in the sunshine": spending reports, schod performance, crime reports, corporate filings, campaign contributions, …
- But few people know how to analyze them
- You will learn how to help bridge this divide



Misc. course info

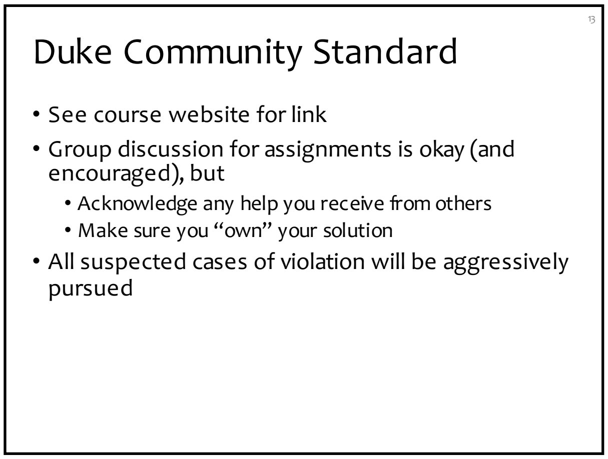
- Website: `http://sites.duke.edu/compsci316_01_f2015/`
  - Course info; tentative schedule and reference sections in the book; lecture slides, assignments, help docs, …
- Book: *Database Systems: The Complete Book*, by H. Garcia-Molina, J. D. Ullman, and J. Widom. 2nd Ed.
- Programming: VM required; $50 worth of credits for VMs in the cloud, courtesy of Amazon
- Q&A on Piazza; grades, sample solutions on Sakai
- Watch your email for announcements
- Office hours to be posted



Grading

[90%, 100%]	A- / A / A+
[80%, 90%)	B- / B / B+
[70%, 80%)	C- / C / C+
[60%, 70%)	D
[0%, 60%)	F

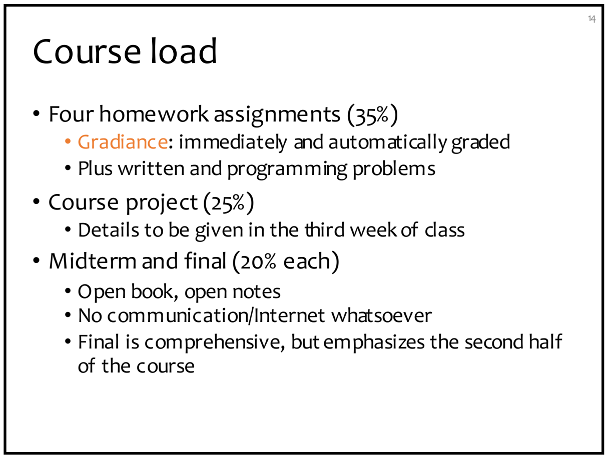
- No "curves"
- Scale may be adjusted downwards (i.e., grades upwards) if, for example, an exam is too difficult
- Scale will not go upwards—mistake would be mine alone if I made an exam too easy



13

Duke Community Standard

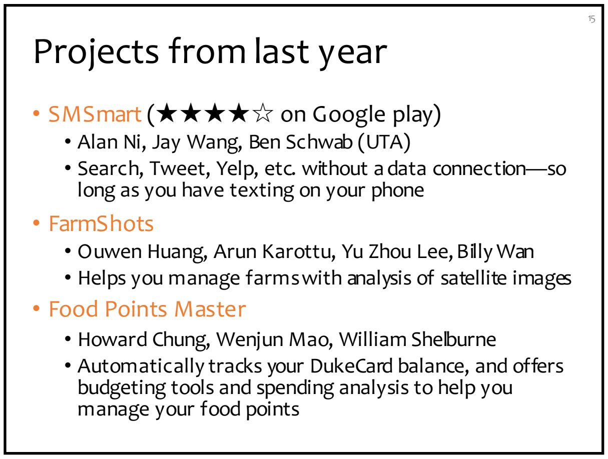
- See course website for link
- Group discussion for assignments is okay (and encouraged), but
  - Acknowledge any help you receive from others
  - Make sure you “own” your solution
- All suspected cases of violation will be aggressively pursued



14

Course load

- Four homework assignments (35%)
  - Gradiance: immediately and automatically graded
  - Plus written and programming problems
- Course project (25%)
  - Details to be given in the third week of class
- Midterm and final (20% each)
  - Open book, open notes
  - No communication/Internet whatsoever
  - Final is comprehensive, but emphasizes the second half of the course



15

Projects from last year

- SMSmart (★★★★☆ on Google play)
  - Alan Ni, Jay Wang, Ben Schwab (UTA)
  - Search, Tweet, Yelp, etc. without a data connection—so long as you have texting on your phone
- FarmShots
  - Ouwen Huang, Arun Karottu, Yu Zhou Lee, Billy Wan
  - Helps you manage farms with analysis of satellite images
- Food Points Master
  - Howard Chung, Wenjun Mao, William Shelburne
  - Automatically tracks your DukeCard balance, and offers budgeting tools and spending analysis to help you manage your food points

Projects from earlier years

- Expose.js: natural language querying
  - E.g.: "*find beers served by bar with name Satisfaction*"
  - Ben Schwab, James Hong, Jesse Hu, 2013
- Pickup Coordinator: an iPhone app that lets you coordinate carpool/pickups with others
  - Adam Cue, Kevin Esoda, Kate Yang, 2012
- Mobile Pay: quick way to make a transaction between two people on their phones
  - Michael Deng, Kevin Gao, Derek Zhou, 2012
- FriendsTracker app: where are my friends?
  - Anthony Lin, Jimmy Mu, Austin Benesh, Nic Dinkins, 2011

More past examples

- ePrint iPhone app
  - Ben Getson and Lucas Best, 2009
- Making iTunes social
  - Nick Patrick, 2006; Peter Williams and Nikhil Arun, 2009
- Duke Schedulator: ditch ACES—plan visually!
  - Alex Beutel, 2008
- SensorDB: manage/analyze sensor data from forest
  - Ashley DeMass, Jonathan Jou, Jonathan Odom, 2007
- Facebook^{+}
  - Tyler Brock and Beth Trushkowsky, 2005
- Web-based K-ville tenting management
  - Zach Marshall, 2005

Your turn to be creative

http://www.yummymummyclub.ca/sites/default/files/styles/large/public/field/image/teaching_kids_creative_skills.jpg

19

So, what is a database system?

From Oxford Dictionary:

- Database: an organized body of related information
- Database system, DataBase Management System (DBMS): a software system that facilitates the creation and maintenance and use of an electronic database

20

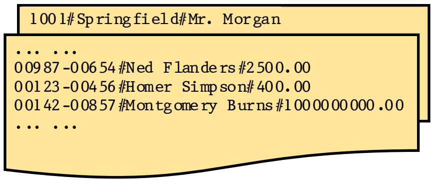
What do you want from a DBMS?

- Keep data around (persistent)
- Answer questions (queries) about data
- Update data

- Example: a traditional banking application
  - Data: Each account belongs to a branch, has a number, an owner, a balance, …; each branch has a location, a manager, …
  - Persistency: Balance can't disappear after a power outage
  - Query: What's the balance in Homer Simpson's account? What's the difference in average balance between Springfield and Capitol City accounts?
  - Modification: Homer withdraws $100; charge accounts with lower than $500 balance a $5 fee

21

Sounds simple!



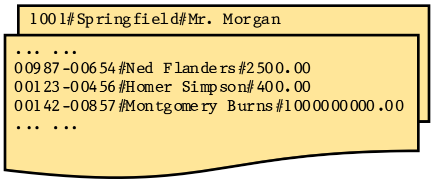
```
1001#Springfield#Mr. Morgan
```

```
... ...
00987-00654#Ned Flanders#2500.00
00123-00456#Homer Simpson#400.00
00142-00857#Montgomery Burns#100000000.00
... ...
```

- Text files
- Accounts/branches separated by newlines
- Fields separated by #'s

22

Query by programming



```
1001#Springfield#Mr. Morgan
... ...
00987-00654#Ned Flanders#2500.00
00123-00456#Homer Simpson#400.00
00142-00857#Montgomery Burns#100000000.00
... ...
```

- What's the balance in Homer Simpson's account?
- A simple script
  - Scan through the accounts file
  - Look for the line containing "Homer Simpson"
  - Print out the balance

23

Query processing tricks

- Tens of thousands of accounts are not Homer's

☞And the list goes on...

What happens when the query changes to: *What's the balance in account 00142-00857?*

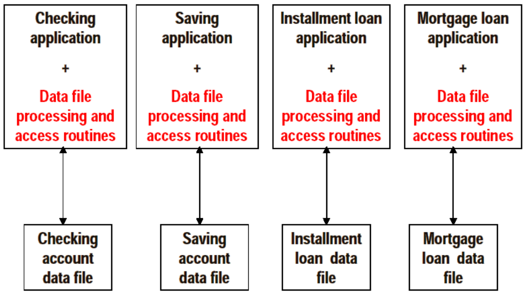
24

Observations

- There are many techniques—not only in storage and query processing, but also in concurrency control, recovery, etc.
- These techniques get used over and over again in different applications
- Different techniques may work better in different usage scenarios

25

The birth of DBMS – 1



Checking application + Data file processing and access routines

Saving application + Data file processing and access routines

Installment loan application + Data file processing and access routines

Mortgage loan application + Data file processing and access routines

Checking account data file

Saving account data file

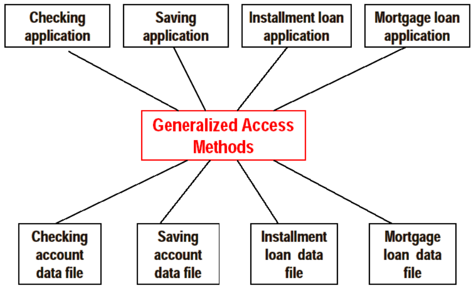
Installment loan data file

Mortgage loan data file

From Hans-J. Schek's *VLDB 2000* slides

26

The birth of DBMS – 2



Checking application

Saving application

Installment loan application

Mortgage loan application

Generalized Access Methods

Checking account data file

Saving account data file

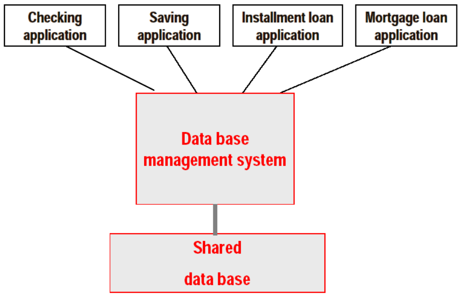
Installment loan data file

Mortgage loan data file

From Hans-J. Schek's *VLDB 2000* slides

27

The birth of DBMS – 3



Checking application

Saving application

Installment loan application

Mortgage loan application

Data base management system

Shared data base

From Hans-J. Schek's *VLDB 2000* slides

28

Early efforts

- "Factoring out" data management functionalities from applications and standardizing these functionalities is an important first step
  - CODASYL standard (circa 1960's)

  ☞Bachman got a Turing award for this in 1973
- But getting the abstraction right (the API between applications and the DBMS) is still tricky

29

CODASYL

- Query: Who have accounts with 0 balance managed by a branch in Springfield?
- Pseudo-code of a CODASYL application:

```
Use index on account(balance) to get accounts with 0 balance;
For each account record:
  Get the branch id of this account;
  Use index on branch(id) to get the branch record;
  If the branch record's location field reads "Springfield":
    Output the owner field of the account record.
```

- Programmer controls "navigation": accounts → branches
  - How about branches → accounts?

30

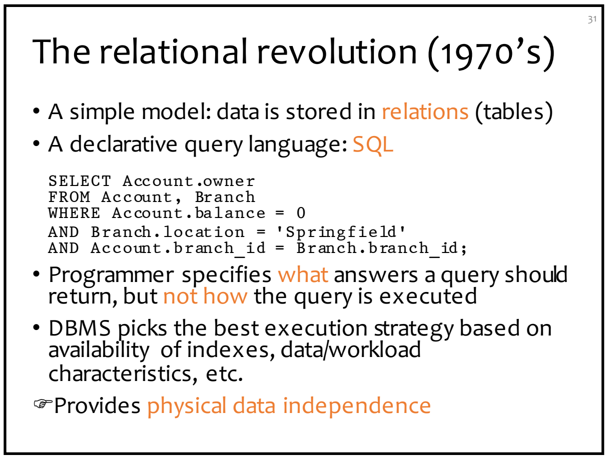
What's wrong?

- The best navigation strategy & the best way of organizing the data depend on data/workload characteristics

With the CODASYL approach

- To write correct code, programmers need to know how data is organized physically (e.g., which indexes exist)
- To write efficient code, programmers also need to worry about data/workload characteristics

☞Can't cope with changes in data/workload characteristics



31

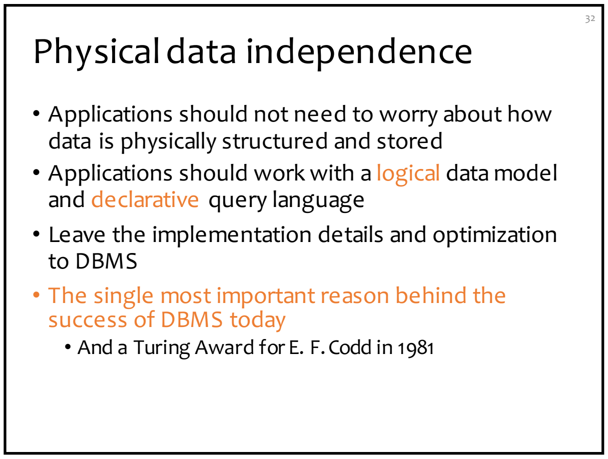
The relational revolution (1970's)

- A simple model: data is stored in relations (tables)
- A declarative query language: SQL

```
SELECT Account.owner
FROM Account, Branch
WHERE Account.balance = 0
AND Branch.location = 'Springfield'
AND Account.branch_id = Branch.branch_id;
```

- Programmer specifies what answers a query should return, but not how the query is executed
- DBMS picks the best execution strategy based on availability of indexes, data/workload characteristics, etc.

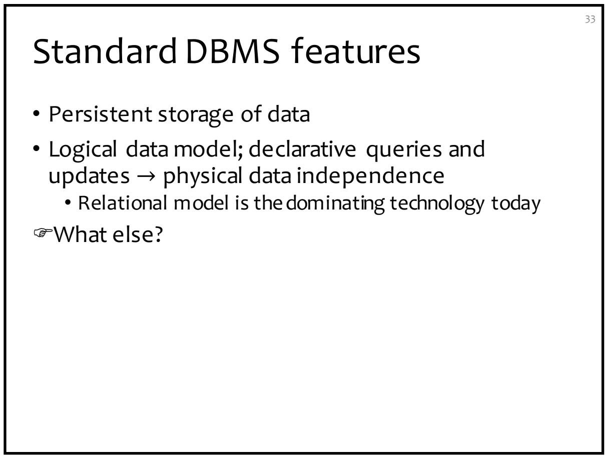
☞Provides physical data independence



32

Physical data independence

- Applications should not need to worry about how data is physically structured and stored
- Applications should work with a logical data model and declarative query language
- Leave the implementation details and optimization to DBMS
- The single most important reason behind the success of DBMS today
  - And a Turing Award for E. F. Codd in 1981

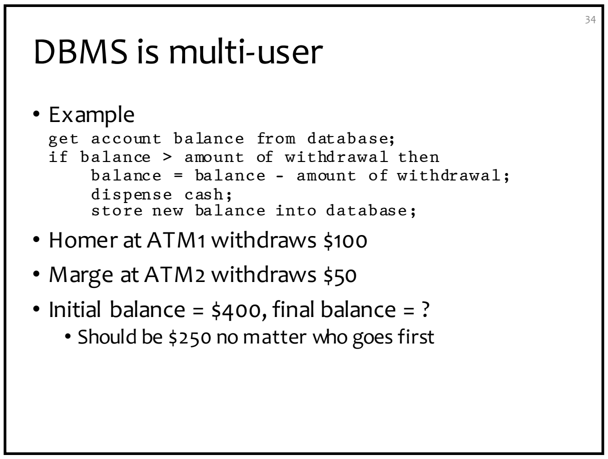


33

Standard DBMS features

- Persistent storage of data
- Logical data model; declarative queries and updates → physical data independence
  - Relational model is the dominating technology today

☞What else?



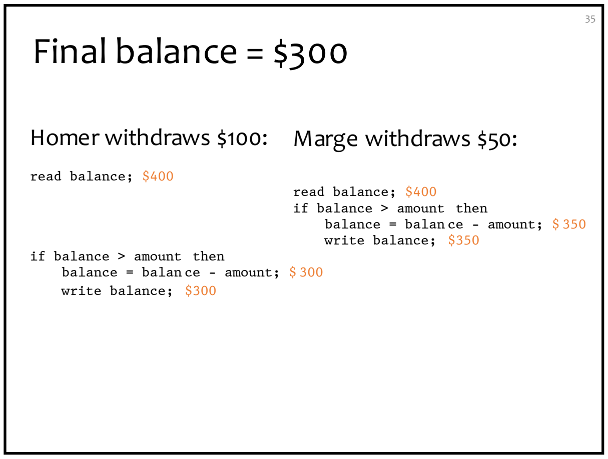
34

DBMS is multi-user

- Example

```
get account balance from database;
if balance > amount of withdrawal then
     balance = balance - amount of withdrawal;
     dispense cash;
     store new balance into database;
```

- Homer at ATM1 withdraws $100
- Marge at ATM2 withdraws $50
- Initial balance = $400, final balance = ?
    - Should be $250 no matter who goes first

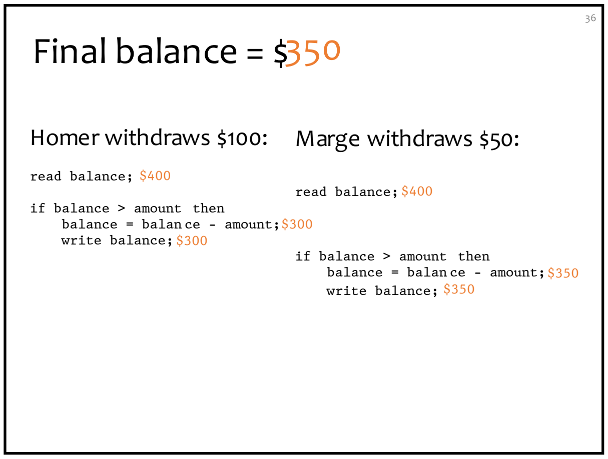


35

Final balance = $300

Homer withdraws $100: Marge withdraws $50:

```
read balance; $400
                                  read balance; $400
                                  if balance > amount then
                                      balance = balance - amount; $350
                                      write balance; $350
if balance > amount then
    balance = balance - amount; $300
    write balance; $300
```



36

Final balance = $350

Homer withdraws $100: Marge withdraws $50:

```
read balance; $400
                                  read balance; $400
if balance > amount then
    balance = balance - amount;$300
    write balance;$300
                                  if balance > amount then
                                      balance = balance - amount;$350
                                      write balance; $350
```

37

Concurrency control in DBMS

- Similar to concurrent programming problems?
  - But data not main-memory variables
- Similar to file system concurrent access?
  - Lock the whole table before access
    - Approach taken by MySQL in the old days
    - Still used by SQLite (as of Version 3)
  - But want to control at much finer granularity
    - Or else one withdrawal would lock up all accounts!

38

Recovery in DBMS

- Example: balance transfer

```
decrement the balance of account X by $100;
increment the balance of account Y by $100;
```

- Scenario 1: Power goes out after the first instruction
- Scenario 2: DBMS buffers and updates data in memory (for efficiency); before they are written back to disk, power goes out
- How can DBMS deal with these failures?

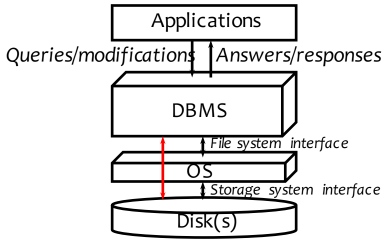
39

Standard DBMS features: summary

- Persistent storage of data
- Logical data model; declarative queries and updates → physical data independence
- Multi-user concurrent access
- Safety from system failures
- Performance, performance, performance
  - Massive amounts of data (terabytes~petabytes)
  - High throughput (thousands~millions transactions/hour)
  - High availability (≥ 99.999% uptime)

40

DBMS architecture today



- Much of the OS may be bypassed for performance and safety
- We will be filling in many details of the DBMS box throughout the semester

41

AYBABTU?

"Us" = relational databases

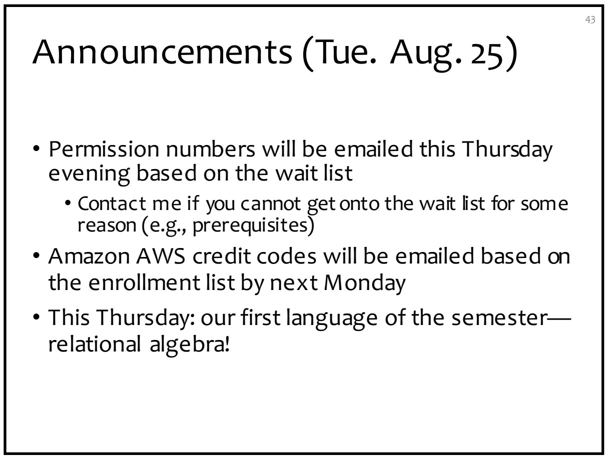
- Most data are not in them!
  - Personal data, web, scientific data, system data, …
- Text and semi-structured data management
  - XML, JSON, …
- "NoSQL" and "NewSQL" movement
  - MongoDB, Cassandra, BigTable, HBase, Spanner, HANA…
- This course will look beyond relational databases

Use of AYBABTU inspired by Garcia-Molina
Image: http://upload.wikimedia.org/wikipedia/en/0/03/Aybabtu.png

42

Course components

- Relational databases
  - Relational algebra, database design, SQL, app programming
- XML
  - Data model and query languages, app programming, interplay between XML and relational databases
- Database internals
  - Storage, indexing, query processing and optimization, concurrency control and recovery
- Advanced topics (TBD)
  - Data warehousing and data mining, Web search and indexing, parallel data processing/MapReduce, etc.



43

Announcements (Tue. Aug. 25)

- Permission numbers will be emailed this Thursday evening based on the wait list
  - Contact me if you cannot get onto the wait list for some reason (e.g., prerequisites)
- Amazon AWS credit codes will be emailed based on the enrollment list by next Monday
- This Thursday: our first language of the semester—relational algebra!