

Welcome to

Comp 115: Databases

<http://www.cs.tufts.edu/comp/115/>

Instructor: *Manos Athanassoulis*
email: *manos@cs.tufts.edu*

Today

big data

data-driven world

databases & database systems



when you see this, I want you to
speak up!
[and you can always interrupt me]

no smartphones



no laptop



Big Data

marketing term ...

but ...

science / government / business / personal data

exponentially growing data collections

So, it is all good!

How big is “Big”?



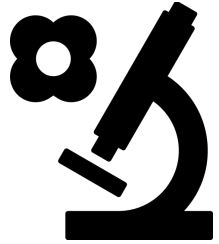
Every day, we create 2.5 exabytes* of data — 90% of the data in the world today has been created in the last two years alone.

[Understanding Big Data, IBM]

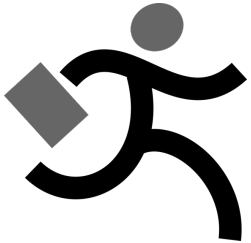
*exabyte = 10^9 GB



Using Big Data



experimental physics (IceCube, CERN)
biology
neuroscience



data mining business datasets
machine learning for corporate and consumer



data analysis for fighting crime

... are only some examples

Data-Driven World

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.The Google logo, with the word "Google" in its characteristic multi-colored font.

Big Data V's

Volume

Velocity

Variety

Veracity

Information is transforming traditional business.

[“Data, data everywhere”, Economist]

Data-Driven World

Reporting

Discovery

Logging

Exploration

Transactions

Data-to-Insight

Business Analysis

Automated Decisions

*Behind all these: use &
manage data*

Comp 115

we live in a ***data-driven*** world

Comp115 is about the ***basics*** for
storing, using, and managing data

your lecturer (that's me!)

Manos Athanassoulis

name in greek: Μάνος Αθανασούλης

grew up in Greece

enjoys playing basketball and the sea

BSc and MSc @ University of Athens, Greece

PhD @ EPFL, Switzerland

Research Intern @ IBM Research Watson, NY

Postdoc @ Harvard University

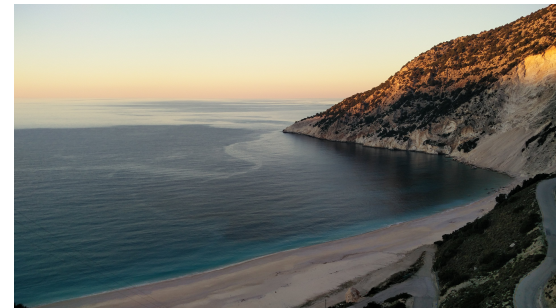
some awards:

SNSF Postdoc Mobility Fellowship

IBM PhD Fellowship



photo for VISA / conferences



Myrtos, Kefalonia, Greece

<http://manos.athanassoulis.net>

Office: Halligan Hall 228B

Office Hours: M/W after class

your awesome TAs



Elif



Sam



Deanna



Taus

your awesome head TA

Sam Lasser
grad Student in PL



ta115@cs.tufts.edu

Data

to make data usable and manageable we
organize them in collections

Databases

a large, integrated, *structured* collection of data

intended to model some real-world enterprise

Examples: a university, a company, social media

University: students, professors, courses

what is missing?

-- how to connect these?

-- enrollment, teaching



What about a company? What about social media?

Database Systems

a.k.a. database management systems (DBMS)

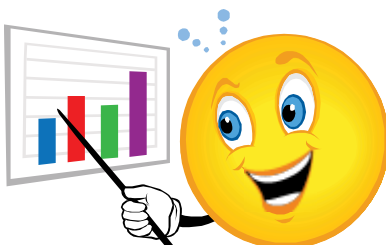
a.k.a. data systems



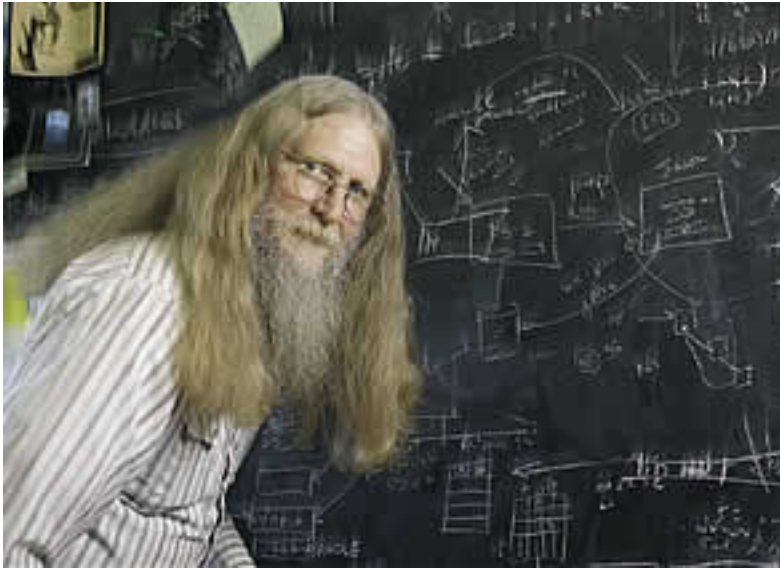
Sophisticated
pieces of software...



... which store, manage,
organize, and facilitate
access to my databases ...



... so I can do things (and ask questions) that are
otherwise hard or impossible



*“relational databases
are the foundation of
western civilization”*

Bruce Lindsay, IBM Research

ACM SIGMOD Edgar F. Codd Innovations award 2012

Ok but what really IS a database system?

Is the WWW a DBMS?



Is a File System a DBMS?



Is Facebook a DBMS?



Is the WWW a DBMS?

Not really!

Fairly sophisticated search available

web crawler *indexes* pages for fast search

.. but

data is unstructured and untyped

no well-defined “correct answer”

cannot update the data

freshness? consistency? fault tolerance?

web sites **use** a *DBMS* to provide these functions

e.g., amazon.com (Oracle), facebook.com (MySQL and others)

“Search” vs. Query

What if you wanted to find out which actors donated to Barack Obama’s presidential campaign 8 years ago?

Try “actors donated to obama” in your favorite search engine.

The screenshot shows a Google search interface. The search bar contains the text "actors donated to obama". Below the search bar, the word "Search" is highlighted in red. To the right of "Search", it says "About 424,000,000 results (0.20 seconds)". On the left side, there is a vertical menu with options: "Everything", "Images", "Maps", "Videos", "News", "Shopping", and "More". Below this menu, there are links for "All results", "Related searches", and "More search tools". The main content area displays several search results. The second result, "Actors Called to Play 'Young People' at Obama Town Hall. - HUMA...", is circled in red. This result includes the URL "www.humanevents.com/article.php?id=39415" and a snippet: "14 Oct 2010 – Obama can't afford an unscripted moment at youth town hall, so casting call goes out to pack the audience." Other visible results include "Dead actor Roy Scheider donates to Barack Obama campaign ..." and "Actor Kal Penn joining Obama administration in Valerie Jarrett's ...".

Google - actors donated to obama

Search About 424,000,000 results (0.20 seconds)

Everything
Images
Maps
Videos
News
Shopping
More

All results
Related searches
More search tools

[Dead actor Roy Scheider donates to Barack Obama campaign ...](#)
latimesblogs.latimes.com/washington/2008/04/dead-actor-roy.html
23 Apr 2008 – Dead voters regularly vote in Barack **Obama's** hometown, but dead donors is something new

[Actors Called to Play 'Young People' at Obama Town Hall. - HUMA...](#)
www.humanevents.com/article.php?id=39415
14 Oct 2010 – **Obama** can't afford an unscripted moment at youth town hall, so casting call goes out to pack the audience.

[Actor Kal Penn joining Obama administration in Valerie Jarrett's ...](#)
blogs.suntimes.com/sweet/.../actor_kal_penn_joining_obama_a.html
7 Apr 2009 – **Actor** Kal Penn worked as a floor whip during the 2008 Democratic National Convention. ... around the country for the **Obama** presidential campaign, is going to leave...
... At first i was furious they **gave** his character the axe.

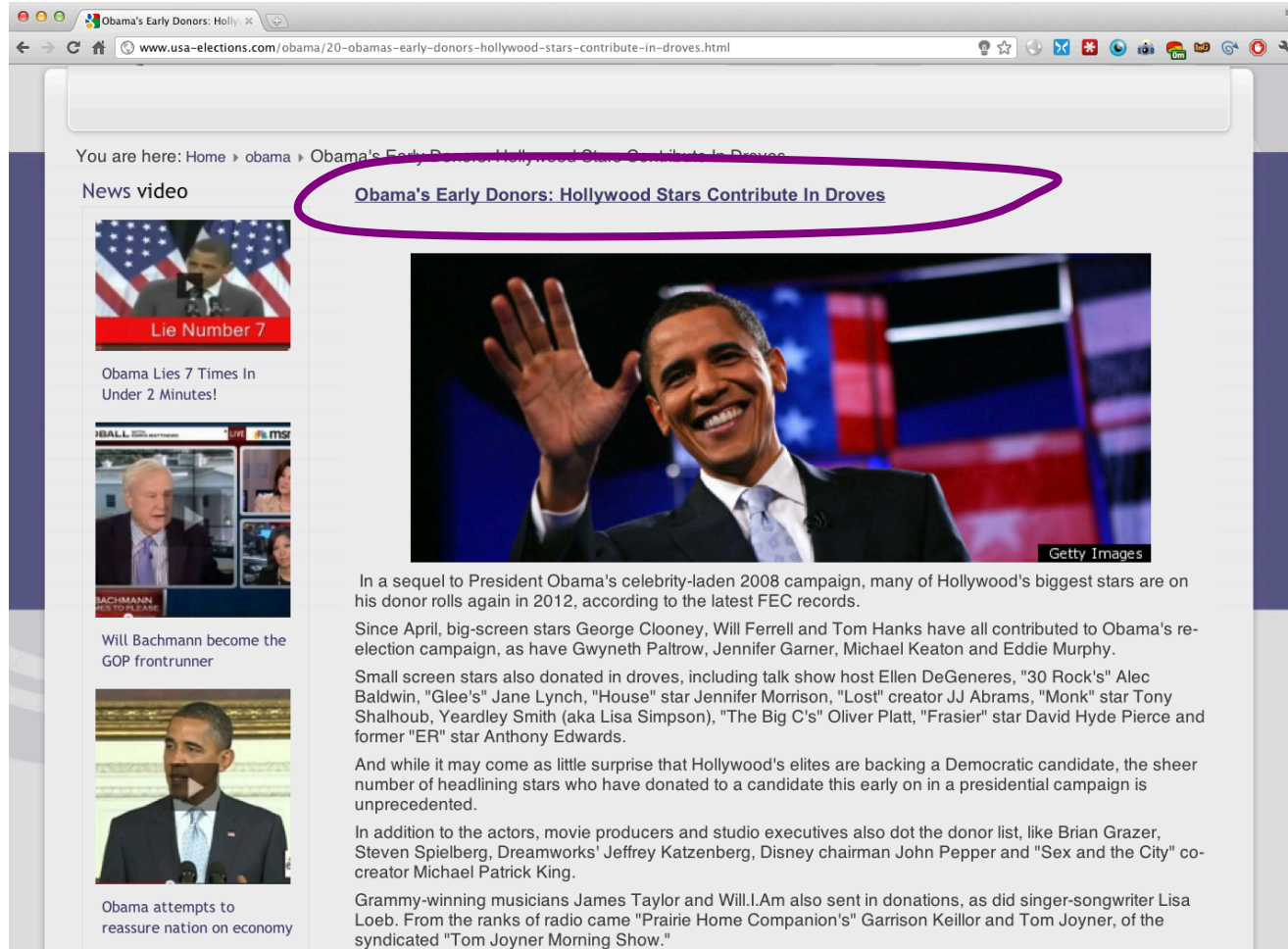
[Obama's Early Donors: Hollywood Stars Contribute In Droves](#)
www.usa-elections.com/obama/20-obamas-early-donors-hollywood-s...
Obama's Early Donors: Hollywood **Stars** Contribute In Droves ... Small screen stars also **donated** in droves, including talk show host Ellen DeGeneres, "30 ...

[Don't call it a Hollywood fundraiser | Campaign 2012 | Washington ...](#)
campaign2012.washingtonexaminer.com/.../dont-call-it-hollywood-...
26 Oct 2011 – President Barack **Obama** looks out the window of his limousine after ... "not a fundraiser," and attendees "were not asked to **donate** to the **Obama** ... **Actor** Kal Penn (Kalpen Modi) formerly worked for **Obama** in the White House ...

“Search” vs. Query

“Search” can return only what’s been “stored”

E.g., best match at Google:



The screenshot shows a web browser window with the address bar displaying www.usa-elections.com/obama/20-obamas-early-donors-hollywood-stars-contribute-in-droves.html. The page content includes a breadcrumb trail: "You are here: Home > obama > Obama's Early Donors: Hollywood Stars Contribute In Droves". Below this, a purple oval highlights the article title "Obama's Early Donors: Hollywood Stars Contribute In Droves". The left sidebar features three video thumbnails: "Lie Number 7" (Obama), "Will Bachmann become the GOP frontrunner" (Will Bachmann), and "Obama attempts to reassure nation on economy" (Obama). The main article area has a large photo of Barack Obama waving, credited to "Getty Images". The text of the article discusses President Obama's celebrity donors in 2012, mentioning George Clooney, Will Ferrell, Tom Hanks, Gwyneth Paltrow, Jennifer Garner, Michael Keaton, and Eddie Murphy, as well as smaller screen stars like Ellen DeGeneres, Alec Baldwin, Jane Lynch, Jennifer Morrison, JJ Abrams, Tony Shalhoub, Yeadley Smith, Oliver Platt, David Hyde Pierce, and Anthony Edwards. It also mentions movie producers and studio executives like Brian Grazer, Steven Spielberg, Jeffrey Katzenberg, John Pepper, and Michael Patrick King, and musicians like James Taylor, Will.I.Am, and Lisa Loeb.

Obama's Early Donors: Hollywood Stars Contribute In Droves

In a sequel to President Obama's celebrity-laden 2008 campaign, many of Hollywood's biggest stars are on his donor rolls again in 2012, according to the latest FEC records.

Since April, big-screen stars George Clooney, Will Ferrell and Tom Hanks have all contributed to Obama's re-election campaign, as have Gwyneth Paltrow, Jennifer Garner, Michael Keaton and Eddie Murphy.

Small screen stars also donated in droves, including talk show host Ellen DeGeneres, "30 Rock's" Alec Baldwin, "Glee's" Jane Lynch, "House" star Jennifer Morrison, "Lost" creator JJ Abrams, "Monk" star Tony Shalhoub, Yeadley Smith (aka Lisa Simpson), "The Big C's" Oliver Platt, "Frasier" star David Hyde Pierce and former "ER" star Anthony Edwards.

And while it may come as little surprise that Hollywood's elites are backing a Democratic candidate, the sheer number of headlining stars who have donated to a candidate this early on in a presidential campaign is unprecedented.

In addition to the actors, movie producers and studio executives also dot the donor list, like Brian Grazer, Steven Spielberg, Dreamworks' Jeffrey Katzenberg, Disney chairman John Pepper and "Sex and the City" co-creator Michael Patrick King.

Grammy-winning musicians James Taylor and Will.I.Am also sent in donations, as did singer-songwriter Lisa Loeb. From the ranks of radio came "Prairie Home Companion's" Garrison Keillor and Tom Joyner, of the syndicated "Tom Joyner Morning Show."

A “Database Query” Approach

where can we find
data for “all actors”?



where can we find
data for “all donations”?



A “Database Query” Approach

www.imdb.com/search/name?gender=male&sort=alpha,asc&start=16684

IMDb Find Movies, TV shows, Celebrities and more... All


Males

Sorted by Name Ascending

16,684-16,733 of 1,865,455 names.


Sort by: [STARMeter](#) | [A-Z](#) | [Height](#) | [Birth Date](#) | [Deaths](#)

16684.




[Adam Sandler](#)
Producer, [Grown Ups](#)
Adam Sandler was born on September 9, 1966, in Queens, New York. He took his first stand-up comedy gig at 17, when he spontaneously took the stage at a local comedy club. He nurtured his talent and became a successful comedian, actor, and producer.

16685.




[Adam Sandler](#)
Producer, [Episode #38.2](#)

16686.



[Adam Sandoval](#)
Actor, [Unspeakable](#)

16687.



[Adam Sandroni](#)
Actor, [Joey's Girl](#)

OpenSecrets.org Center for Responsive Politics

Search... GO

(e.g. Donors, Politicians, Corporations and more)

[Politicians & Elections](#) [Influence & Lobbying](#) [News & Analysis](#) [Resources](#) [Take Action](#) [About Us](#) [Donate!](#)

Home » [Politicians & Elections](#) » [Presidential](#) » **Presidential Donor Lookup Results**

In Politicians & Elections:

Presidential Donor Lookup Results

Your search has generated too many results. Only the top 1000 records are being displayed. If you would like to refine your search, return to the [form page](#).

Search Criteria:

Donor name: (all)

Cycle selected: 2008

[Start another search](#)

☐ Sort by Name

☐ Sort by Date (Descending)

☒ Sort by Amount

[Sort](#)

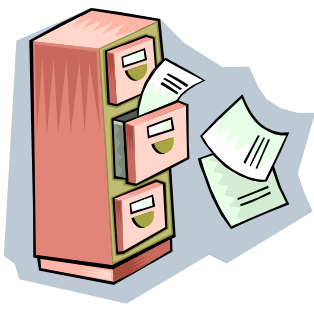
[Save/Share:](#) [BOOKMARK](#) [Print](#) [E-mail](#)

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) ... [21](#) [Next](#)

Candidate	Contributor	Employer	Date	Amount
Obama, Barack	Budinger, William Aspen, CO 81611	Not employed	7/31/08	\$30,800
Obama, Barack	BOSLER, JAMES FORT WORTH, TX 76126	NOT EMPLOYED/RETIRED	8/28/08	\$28,500
Obama, Barack	HIGDON, JOE WASHINGTON, DC 20008	NOT EMPLOYED/RETIRED	8/28/08	\$28,500
Obama, Barack	MYERS, DEBRA RANCHO PALOS VERDE, CA 90275	SELF EMPLOYED/PHYSICIAN	8/31/08	\$28,500
Obama, Barack	MYERS, WOODROW DR JR INDIANAPOLIS, IN 46204	MYERS VENTURES LLC/MGR DIRECTOR	8/31/08	\$28,500

“IMDB Actors” JOIN “OpenSecrets”

Contributor	Employer	Date	Amount
ROCK, CHRIS MR NEW YORK,NY 10019	ACTOR	4/20/07	\$9,200
DOUGLAS, MICHAEL UNIVERSAL CITY,CA 91608	ACTOR/ PRODUCER	3/30/07	\$4,600
DOUGLAS, MICHAEL UNIVERSAL CITY,CA 91608	ACTOR/ PRODUCER	3/30/07	\$2,300
ROCK, CHRIS MR NEW YORK,NY 10019	ACTOR	4/20/07	\$2,300
CARIDES, GEORGIA NEW YORK,NY 10017	ACTOR	5/18/07	\$1,000
CARTER COVINGTON, CLAUDIA CHARLOTTE,NC 28207	ACTORS THEATRE PART TIME/ACTOR/NEW	5/20/08	\$1,000
FOX, RICK ENCINO,CA 91316	ACTOR/PRODUCER	6/16/08	\$1,000
HILDRETH, THOMAS W LOS ANGELES,CA 90068	ACTOR	9/29/08	\$1,000
RENNER, CARL BEVERLY HILLS,CA 90210	ACTOR/BESSONE@ROADRUNNER.COM	8/28/08	\$1,000
SIMMONS, HENRY WEST HOLLYWOOD,CA 90046	ACTOR	6/4/07	\$1,000



Is a File System a DBMS?

Not really!

Thought Experiment 1:

- You and your project partner are editing the same file.
- You both save it at the same time.
- Whose changes survive?



A) Yours **B) Partner's** **C) Both** **D) Neither** **E) ???**

Thought Experiment 2:

- You're updating a file.
- The power goes out.
- Which of your changes survive?



A) All **B) None** **C) All Since last save** **D) ???**

Is Facebook a DBMS?

Is the data structured & typed?



Does it offer well-defined queries?

Not really!

Does it offer properties like “durability” and “consistency”?

Facebook is a data-driven company that uses several database systems (>10) for different use-cases (internal or external).

Why take this class?

computation to information

corporate, personal (web), science (big data)

database systems everywhere

data-driven world, data companies

DBMS: much of CS as a practical discipline

languages, theory, OS, logic, architecture, HW

Comp 115 in a nutshell

model

data representation model

query

query languages – ad hoc queries

access (concurrently multiple reads/writes)
ensure *transactional* semantics

store (reliably)
maintain *consistency/semantics* in *failures*

A “free taste” of the class

data modeling

query languages

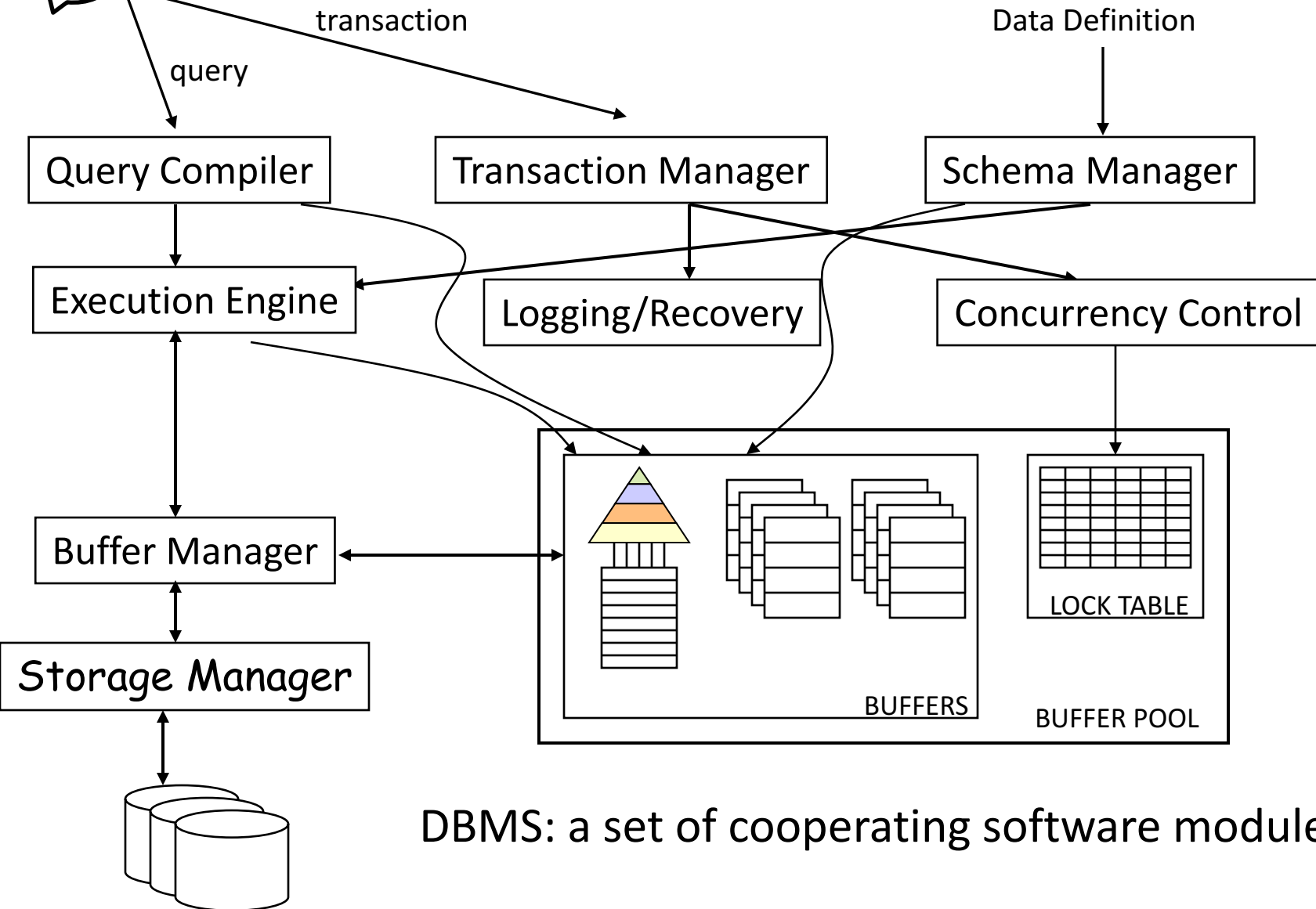
concurrent, fault-tolerant data management

DBMS architecture

Coming in next class

Discussion on *database systems designs*

Components of a "classic" DBMS



Describing Data: Data Models

data model : a collection of concepts describing data

relational model is the most widely used model today
key concepts

relation : basically a table with rows and columns

schema : describes the columns (or fields) of each table

Schema of “University” Database

Students

***sid**: string, **name**: string, **login**: string, **age**: integer, **gpa**: real*

Courses

***cid**: string, **cname**: string, **credits**: integer*

Enrolled

***sid**: string, **cid**: string, **grade**: string*



Levels of Abstraction

what the users *see*

External Schema 1

External Schema 2

what is the *data model*

Conceptual Schema

how the data is *physically* stored
e.g., files, indexes

Physical Schema

Schemata of “University” Database

Conceptual Schema

Students

***sid**: string, **name**: string, **login**: string, **age**: integer, **gpa**: real*

Courses

***cid**: string, **cname**: string, **credits**: integer*

Enrolled

***sid**: string, **cid**: string, **grade**: string*

Physical Schema

relations stored in heap files

indexes for sid/cid

Schemata of “University” Database

External Schema

a “view” of data that can be derived from the existing data

example: Course Info

Course_Info (***cid***: string, ***enrollment***:integer)

Data Independence

Abstraction offers “application independence”

Logical data independence

Protection from changes in *logical* structure of data

Physical data independence

Protection from changes in *physical* structure of data

Q: Why is this particularly important for DBMS?

Applications can treat DBMS as
black boxes!



Queries

”Bring me all students with gpa more than 3.0”

“SELECT * FROM Students WHERE gpa>3.0”

SQL – a powerful declarative query language

treats DBMS as a black box

What if we have multiples accesses?

Concurrency Control

multiple users/apps

Challenges



how frequent access to slow medium

how to keep CPU busy

how to avoid *short jobs* waiting behind *long ones*

e.g., ATM withdrawal while summing all *balances*

interleaving actions of *different* programs

Concurrency Control

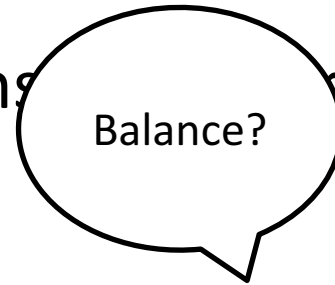
Problems with *interleaving* actions of programs



Bill



Move 100 from
savings to checking



Alice

Bad interleaving:

Savings -= 100

Print balances

Checking += 100

Printout is missing 100\$!

Concurrency Control

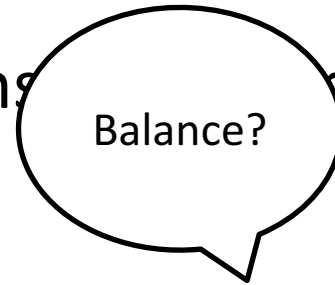
Problems with *interleaving* actions of programs



Bill



Move 100 from
savings to checking



Alice

What is a correct interleaving?

Savings -= 100

Checking += 100

Print balances



How to achieve this interleaving?



Scheduling Transactions

Transactions: atomic sequences of **R**eads & **W**rites

$$T_{\text{Bill}} = \{R1_{\text{Savings}}, R1_{\text{Checking}}, W1_{\text{Savings}}, W1_{\text{Checking}}\}$$
$$T_{\text{Alice}} = \{R2_{\text{Savings}}, R2_{\text{Checking}}\}$$

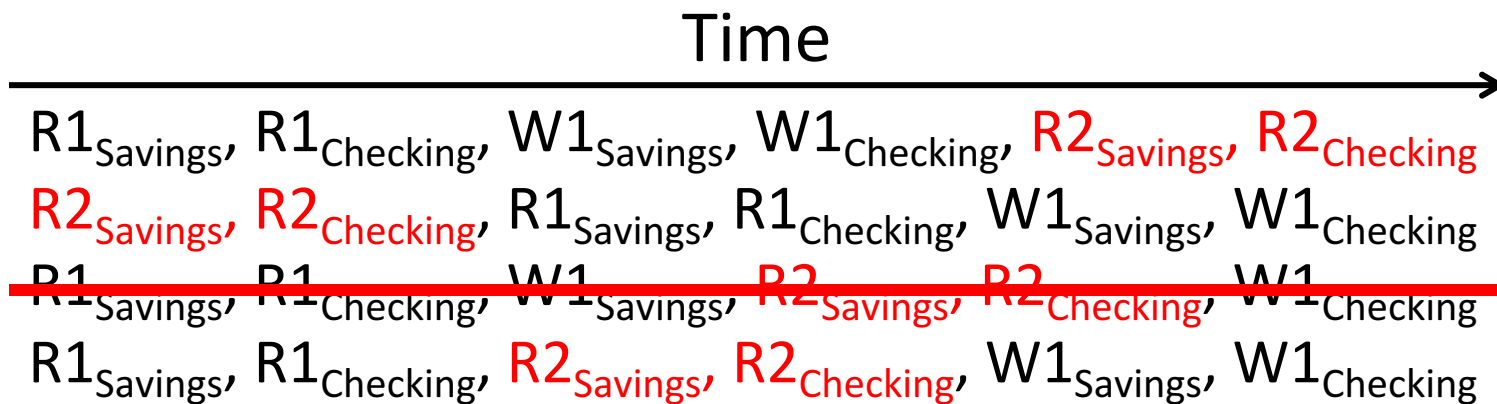
How to avoid previous problems?



Scheduling Transactions

All interleaved executions equivalent to a serial

All actions of a transaction executed as a whole



How to achieve one of these?



Locking



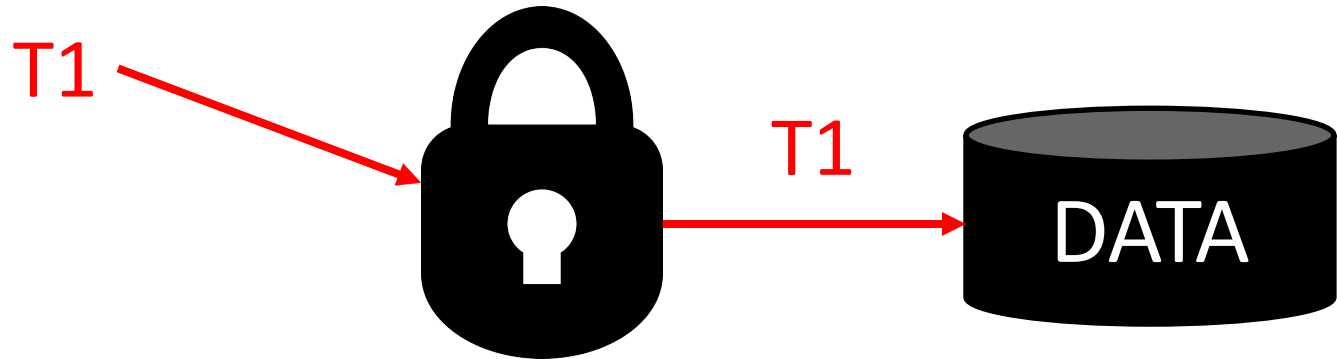
before an object is accessed a lock is requested

Locking



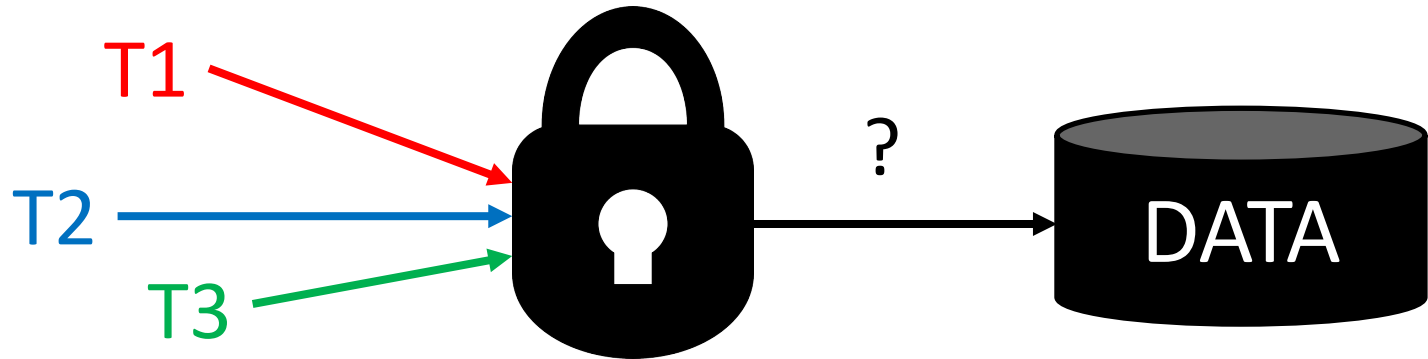
before an object is accessed a lock is requested

Locking



before an object is accessed a lock is requested

Locking



locks are held until the end of the transaction

*[this is only one way to do this, called
“strict two-phase locking”]*

Locking

$T_1 = \{R1_{Savings}, R1_{Checking}, W1_{Savings}, W1_{Checking}\}$

$T_2 = \{R2_{Savings}, R2_{Checking}\}$

Both should lock *Savings* and *Checking*

What happens:

if T1 locks Savings & Checking ?

T2 has to wait

if T1 locks Savings & T2 locks Checking ?

we have a deadlock



How to solve deadlocks?

we need a mechanism to undo

also when a transaction is incomplete
e.g., due to a crash



what can be an undo mechanism?



log every action before it is applied!

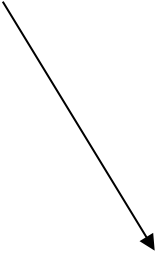
Transactional Semantics

Transaction: one execution of a user program

multiple executions → multiple transactions

Every transaction:

Logging → ***Atomic***
Consistent
Isolated
Durable

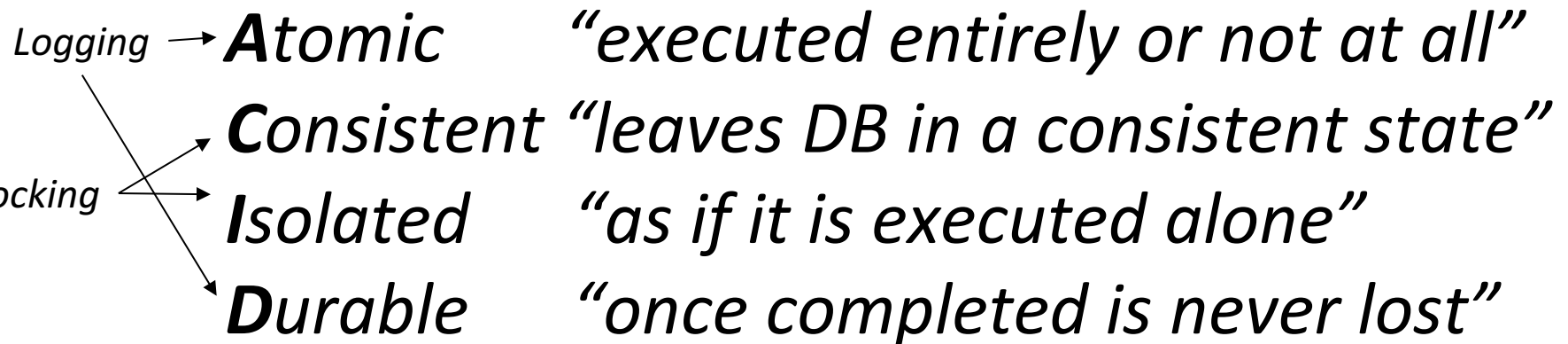


Transactional Semantics

Transaction: one execution of a user program

multiple executions → multiple transactions

Every transaction:



Who else needs transactions?



lots of data

lots of users

frequent updates

background game analytics

Scaling games to epic proportions,

by W. White, A. Demers, C. Koch, J. Gehrke and R. Rajagopalan

ACM SIGMOD International Conference on Management of Data, 2007

Only “classic” DBMS?

No, there is much more!

NoSQL & Key-Value Stores: No transactions, focus on queries

Graph Stores

Querying raw data without loading/integrating costs

Database queries in large datacenters

New hardware and storage devices

... many exciting open problems!

<http://www.cs.tufts.edu/comp/115/>

Next time in ...

Comp 115: Databases

Database Systems Architectures

Class administrativa

Class project administrativa

<http://www.cs.tufts.edu/comp/115/>

Additional Accommodations

If you require additional accommodations please contact the Student Accessibility Services office at Accessibility@tufts.edu or 617-627-4539 to make an appointment with an SAS representative to determine which are the appropriate accommodations for your case.

Please be aware that accommodations cannot be enacted retroactively, making timeliness a critical aspect for their provision.

More details about accessibility services in the syllabus:

<http://www.cs.tufts.edu/comp/115/syllabus.html>