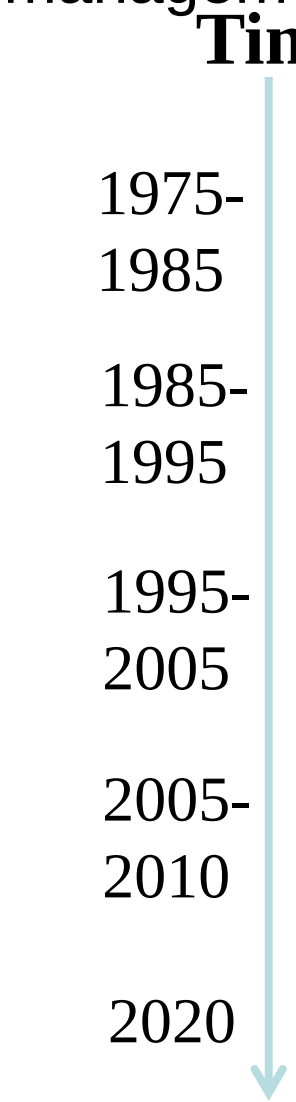
CPS 216: Advanced Database Systems (Data-intensive Computing Systems)

Shivnath Babu

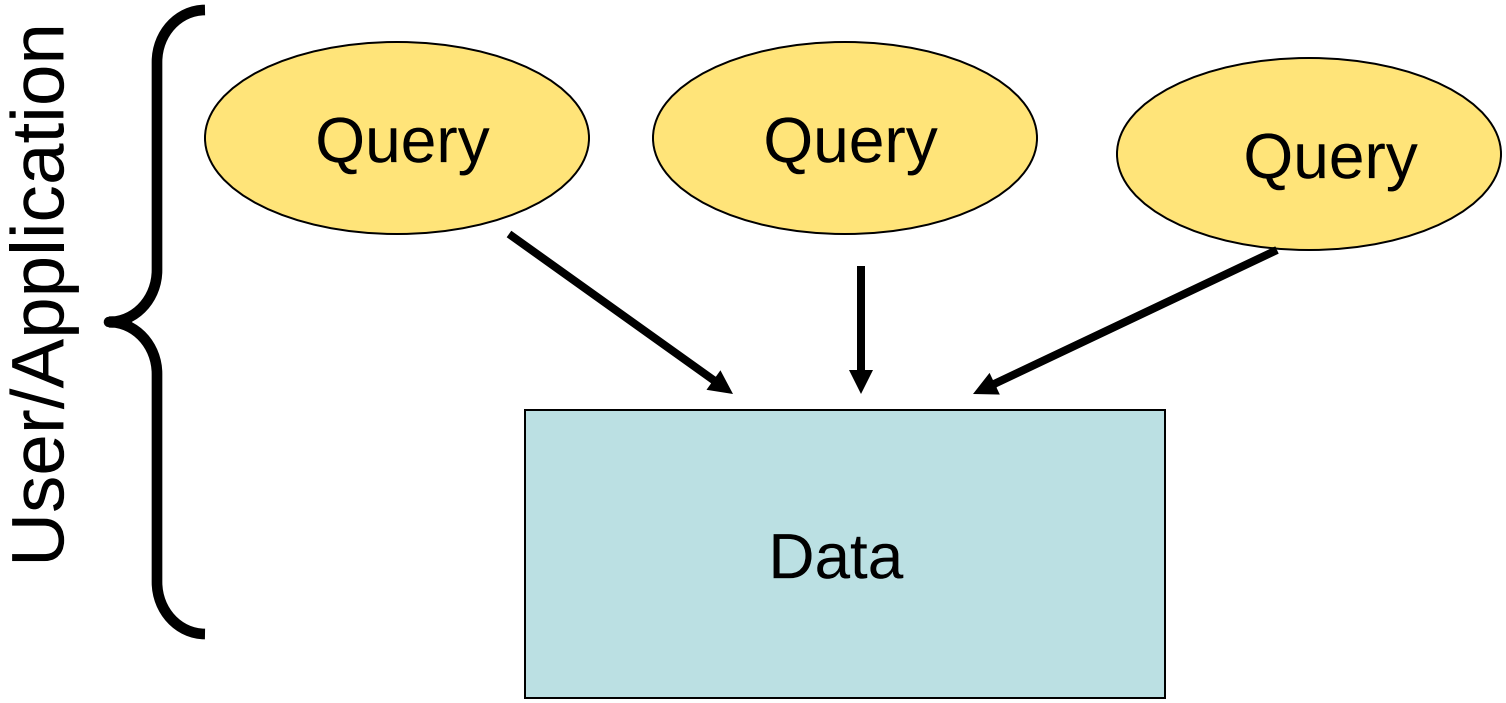
A Brief History

Relational database management systems



Let us first see what a relational database system is

Data Management



DataBase Management System (DBMS)

Example: At a Company

Query 1: Is there an employee named “Nemo”?
Query 2: What is “Nemo’s” salary?
Query 3: How many departments are there in the company?
Query 4: What is the name of “Nemo’s” department?
Query 5: How many employees are there in the “Accounts” department?

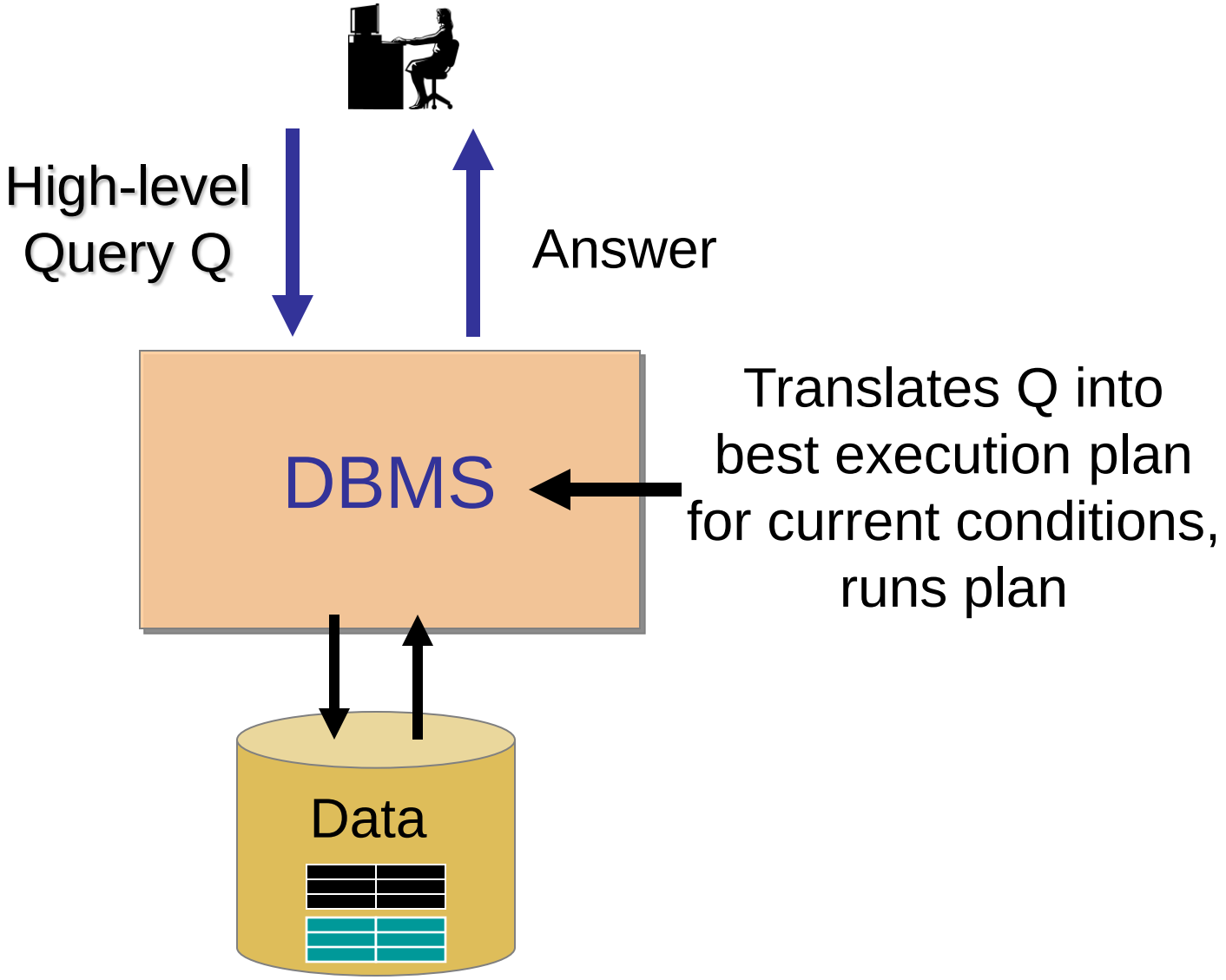
Employee

ID	Name	DeptID	Salary	…
10	Nemo	12	**120K**	…
20	Dory	156	**79K**	…
40	Gill	89	**76K**	…
52	Ray	34	**85K**	…
…	…	…	…	…

Department

ID	Name	…
12	IT	…
34	Accounts	…
89	HR	…
156	Marketing	…
…	…	…

DataBase Management System (DBMS)



Example: Store that Sells Cars

Owners of Honda Accords who are <= 23 years old

Make	Model	OwnerID	ID	Name	Age
Honda	Accord	12	12	Nemo	**22**
Honda	Accord	156	156	Dory	**21**

Join (Cars.OwnerID = Owners.ID)

Filter (Make = Honda and Model = Accord)

Filter (Age <= 23)

Cars

Make	Model	OwnerID
Honda	Accord	12
Toyota	Camry	34
Mini	Cooper	89
Honda	Accord	156
…	…	…

Owners

ID	Name	Age
12	Nemo	**22**
34	Ray	**42**
89	Gill	**36**
156	Dory	**21**
…	…	**…**

DataBase Management System (DBMS)

High-level Query Q

Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

Data

A Brief History

Relational database management systems

Time

1975-1985

1985-1995

1995-2005

2005-2010

2020

Assumptions and requirements changed over time

Semi-structured and unstructured data (Web)

Hardware developments

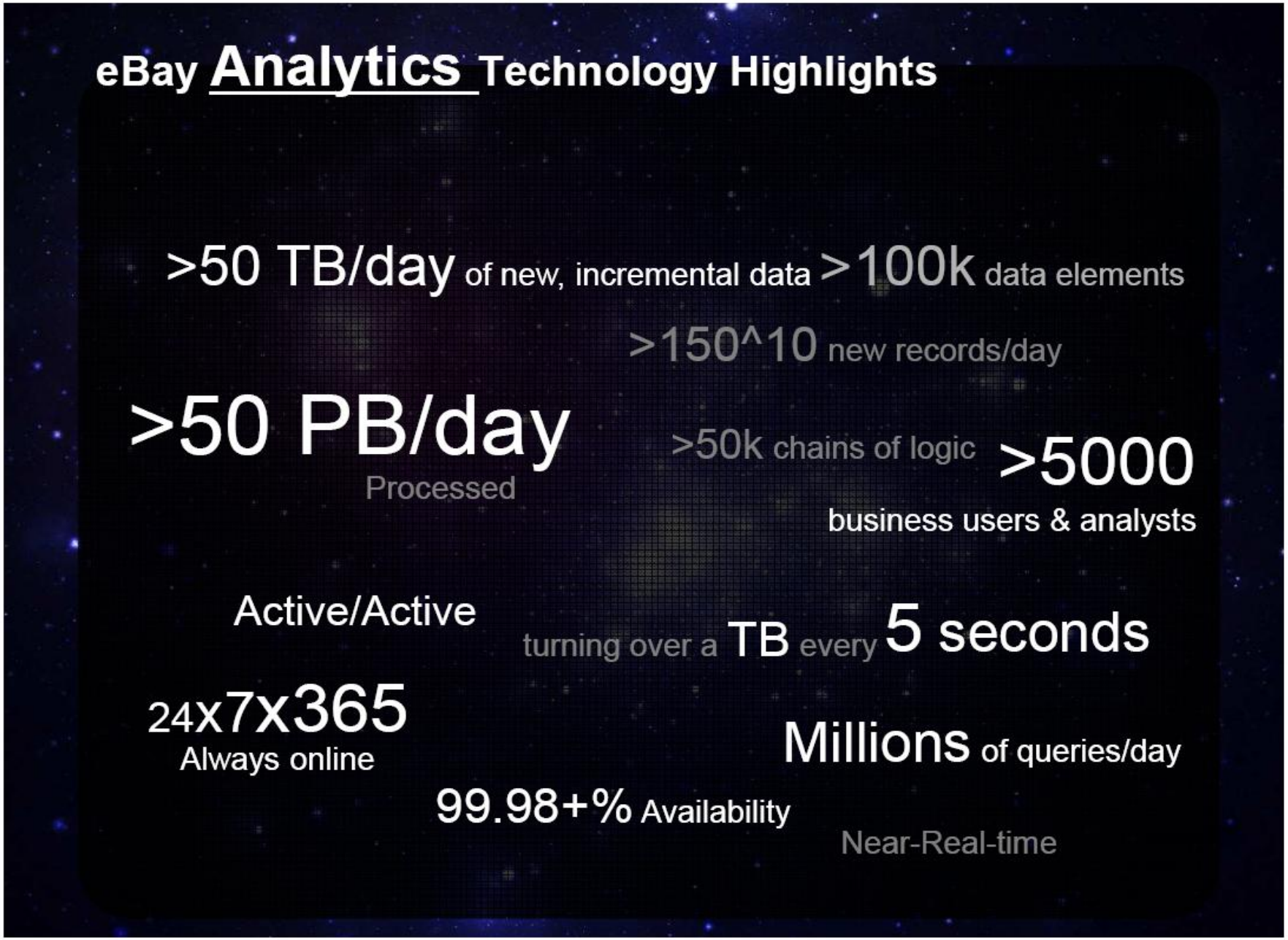
Developments in system software

Changes in data sizes

Big Data: How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- Facebook has 36 PB of user data + 80-90 TB/day (6/2010)
- CERN's LHC: 15 PB a year (any day now)
- LSST: 6-10 PB a year (~2015)

From http://www.umiacs.umd.edu/~jimmylin/



From: http://www.cs.duke.edu/smdb10/

NEW REALITIES

The quest for knowledge used to begin with grand theories.

Now it begins with massive amounts of data.

Welcome to the Petabyte Age.

From: http://db.cs.berkeley.edu/jmh/

THE NEW PRACTITIONERS

"Looking for a career where your services will be in high demand?

… Provide a scarce, complementary service to something that is getting ubiquitous and cheap.

the sexy job in the next ten years will be statisticians

So what's ubiquitous and cheap?
Data.

And what is complementary to data?
Analysis.

Hal Varian, UC Berkeley, Chief Economist @ Google

From: http://db.cs.berkeley.edu/jmh/

THE NEW PRACTITIONERS

- Aggressively Datavorous
- Statistically savvy
- Diverse in training, tools

From: http://db.cs.berkeley.edu/jmh/

FOX AUDIENCE NETWORK

- Greenplum parallel DB
  - 42 Sun X4500s ("Thumper") *each* with:
    - 48 500GB drives
    - 16GB RAM
    - 2 dual-core Opterons
- Big and growing
  - 200 TB data (mirrored)
  - Fact table of 1.5 trillion rows
  - Growing 5TB per day
    - 4-7 Billion rows per day

- Also extensive use of R and Hadoop

Yahoo! runs a 4000 node Hadoop cluster (probably the largest). Overall, there are 38,000 nodes running Hadoop at Yahoo!

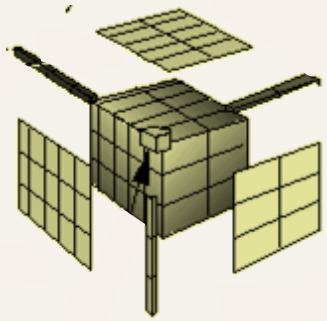
From: http://db.cs.berkeley.edu/jmh/

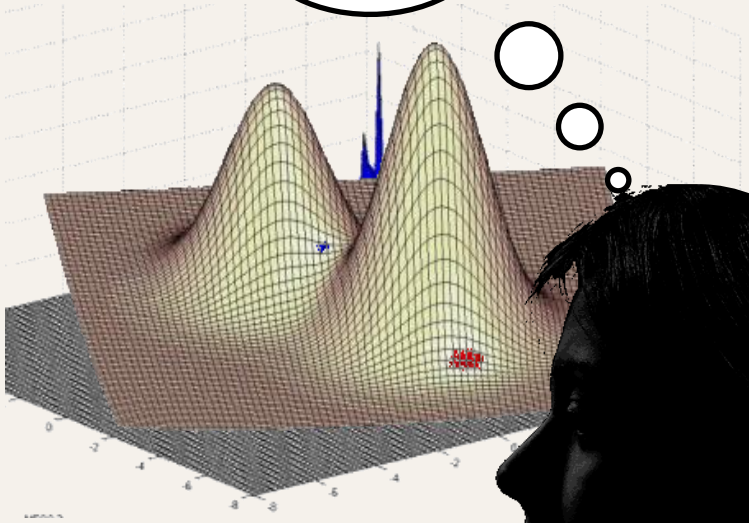
As reported by FAN, Feb, 2009

A SCENARIO FROM FAN

How many female WWF fans under the age of 30 visited the Toyota community over the last 4 days and saw a Class A ad?

How are these people similar to those that visited Nissan?





Open-ended question about statistical *densities (distributions)*

From: http://db.cs.berkeley.edu/jmh/

MULTILINGUAL DEVELOPMENT

- SQL or MapReduce
- Sequential code in a variety of languages
  - Perl
  - Python
  - Java
  - R
- Mix and Match!

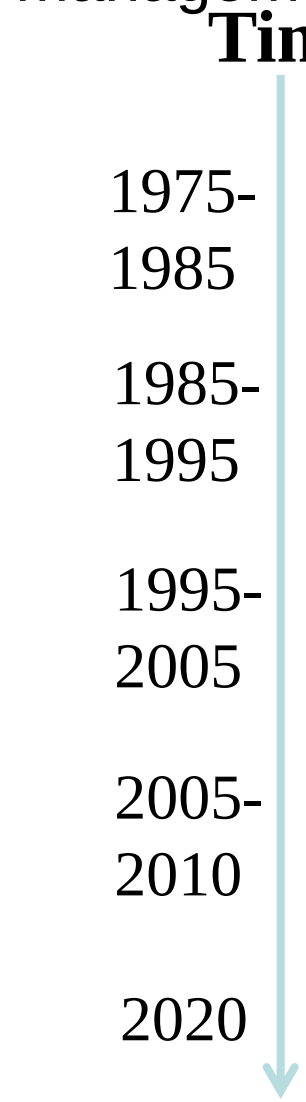
From: http://db.cs.berkeley.edu/jmh/

The Next Gen = Cloud Computing

From: http://outsideinnovation.blogs.com/pseybold/2009/03/-sun-will-shine-in-blue-cloud.html

Teaching/Learning Methodology

Relational database management systems



Assumptions and requirements changed over time

Semi-structured and unstructured data (Web)

Hardware developments

Developments in system software

Changes in data sizes

Course Outline

- Principles of query processing ***(30%)***
  - Indexes
  - Query execution plans and operators
  - Query optimization
- Data storage ***(10%)***
  - Databases Vs. filesystems (Google/Hadoop Distributed FileSystem)
  - Flash memory and Solid State Drives
- Scalable data processing ***(35%)***
  - Parallel query plans and operators
  - Systems based on MapReduce
  - Scalable key-value stores
- Concurrency control and recovery ***(15%)***
  - Consistency models for data (ACID, BASE, Serializability)
  - Write-ahead logging
- Information retrieval and Data mining ***(10%)***
  - Web search (Google PageRank, inverted indexes)
  - Association rules and clustering

Course Logistics

- Web: http://www.cs.duke.edu/courses/fall10/cps216
- TA: Gang Luo
- References:
  - *Hadoop: The Definitive Guide*, by Tom White
  - *Database Systems: The Complete Book*, by H. Garcia-Molina, J. D. Ullman, and J. Widom
- Grading:
  - Project 35% (Hopefully, on Amazon Cloud!)
  - Homework Assignments 15%
  - Midterm 25%
  - Final 25%