Why study databases?

$ 200 000 000

$ 200 000 000 +

$ 13 000 000 / year

Both owned by Larry Ellison, CEO of Oracle

It pays to know databases !

Why go through all this?

- Database administrators are paid well
- Databases are everywhere (i.e. lots of job opportunities)
  - E.g. Google
  - at the doctor's office
  - payroll systems
  - on Wall Street
  - government (e.g. CIA)
  - scientific data
- Database research offers many exciting opportunities
  - Internet technologies
  - handling huge amounts of data
  - etc.

Databases in the wild

- Database assembles US warnings of Saddam threat – Reuters (1/23/2008)
  - can search by keywords
  - summarizes statistics
  - assembled from a number of sources
  - manual curation/entry
- Google
  - database of searches (google trends)
  - database of emails (gmail)
  - database of publications (google scholar)
  - ...
  - privacy issues
- Bio-medical databases
  - doctor's office, lab providers, hospitals, research institutes
  - insurance companies
  - who/how/when/how much information shared?

Data overload

- Commerce/e-commerce (Walmart > 500 TB of product data, 1 billion records added / day, also customer preferences, etc.)
- Library of congress (> 20TB)
- Scientific data
  - Sloan Digital Sky Survey (15 TB)
  - Biological Data (> 1 TB generated / day)
  - Climate data
- Surveillance data (e.g. sensor networks, traffic cameras)

Complex questions

- How do you get to Hershey Park from College Park given traffic, tolls, etc.
- Structure of terrorist networks: who will replace Osama Bin Laden if he is captured?
- Biological data: how do genes work together to create a living organism?

Efficiency

- Given a bank with millions of ATMs – how quickly can each transaction be made?
- Given a large biological dataset (e.g. 6TB Human Microbiome Data) – how quickly can you find all the genes and organisms?

Robustness and Concurrency

- What do you do if systems crash?
- How do you manage many (millions) of simultaneous queries (e.g. google)?
- How do you build a database on thousands of computers?
- How can you ensure privacy and security?

Database Management Systems

- Provide a means to address the questions we just raised
- Primarily deal with structured data (e.g. data that can be stored in spreadsheets)
- More advanced versions deal with graphs, plain text (e.g. searching/processing blogs), etc.
- We'll primarily focus on the former

Account		
bname	**acct_no**	**balance**
Downtown	A-101	500
Mianus	A-215	700
Perry	A-102	400
R.H	A-305	350

Customer		
cname	**cstreet**	**ccity**
Jones	Main	Harrison
Smith	North	Rye
Hayes	Main	Harrison
Curry	North	Rye
Lindsay	Park	Pittsfield

What we will cover…

- representing information
  - data modeling
- languages and systems for querying data
  - complex queries & query semantics
  - over massive data sets
- concurrency control for data manipulation
  - controlling concurrent access
  - ensuring transactional semantics
- reliable data storage
  - maintain data semantics even if you pull the plug