Database Management Systems

Fall 2017

"Knowledge is of two kinds: we know a subject ourselves, or we know where we can find information upon it."

-- Samuel Johnson (1709-1784)

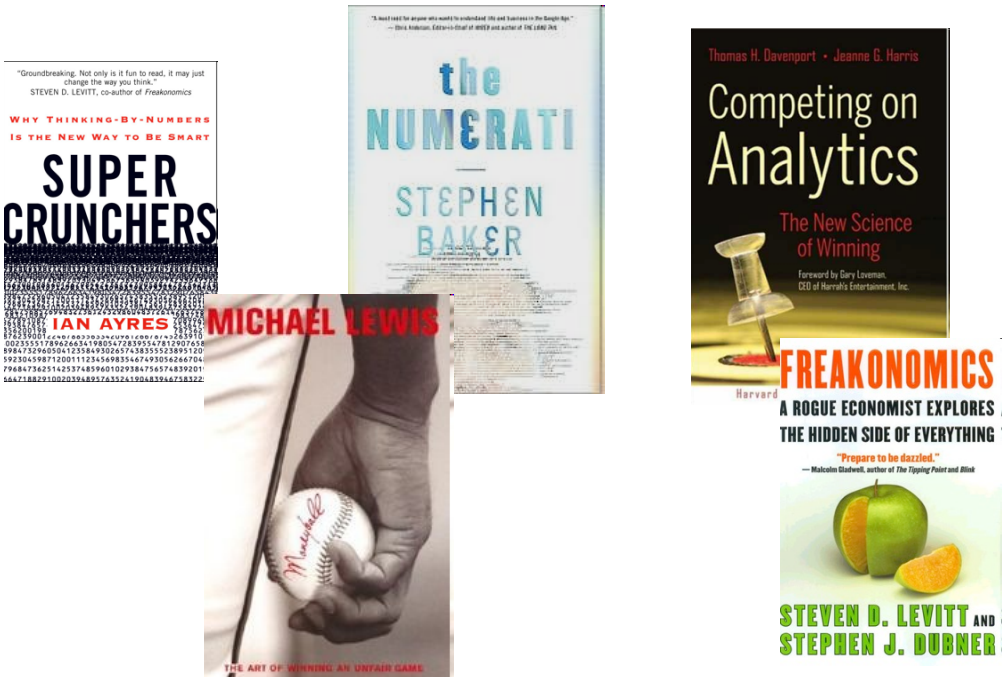
Queries for Today

- **Why?**
- Who?
- What?
- How?
- For instance?

Databases – Why Study Them?

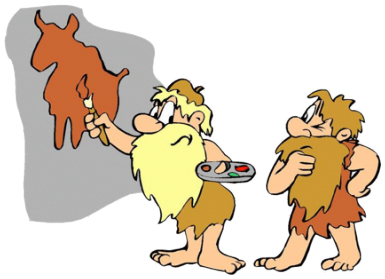
The quest for knowledge used to begin with grand theories.
Now it begins with massive amounts of data.
Welcome to the Petabyte Age.

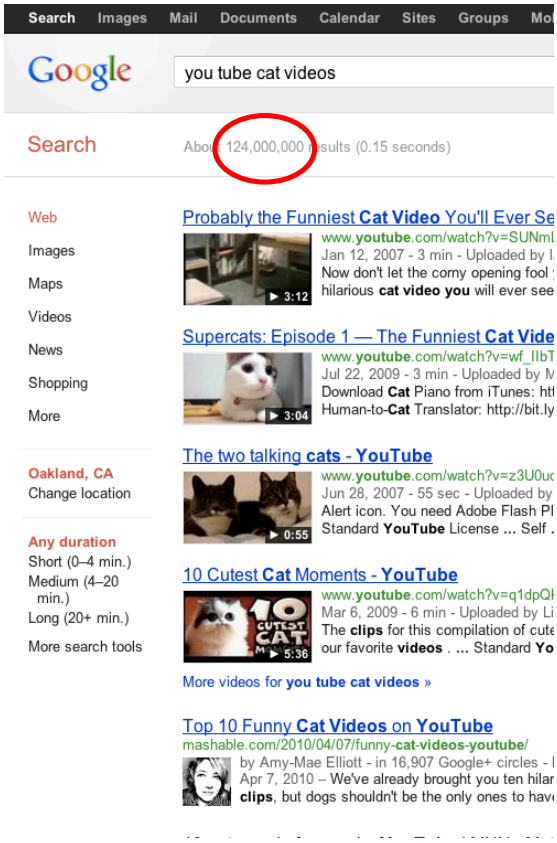
Databases – Why Study Them?

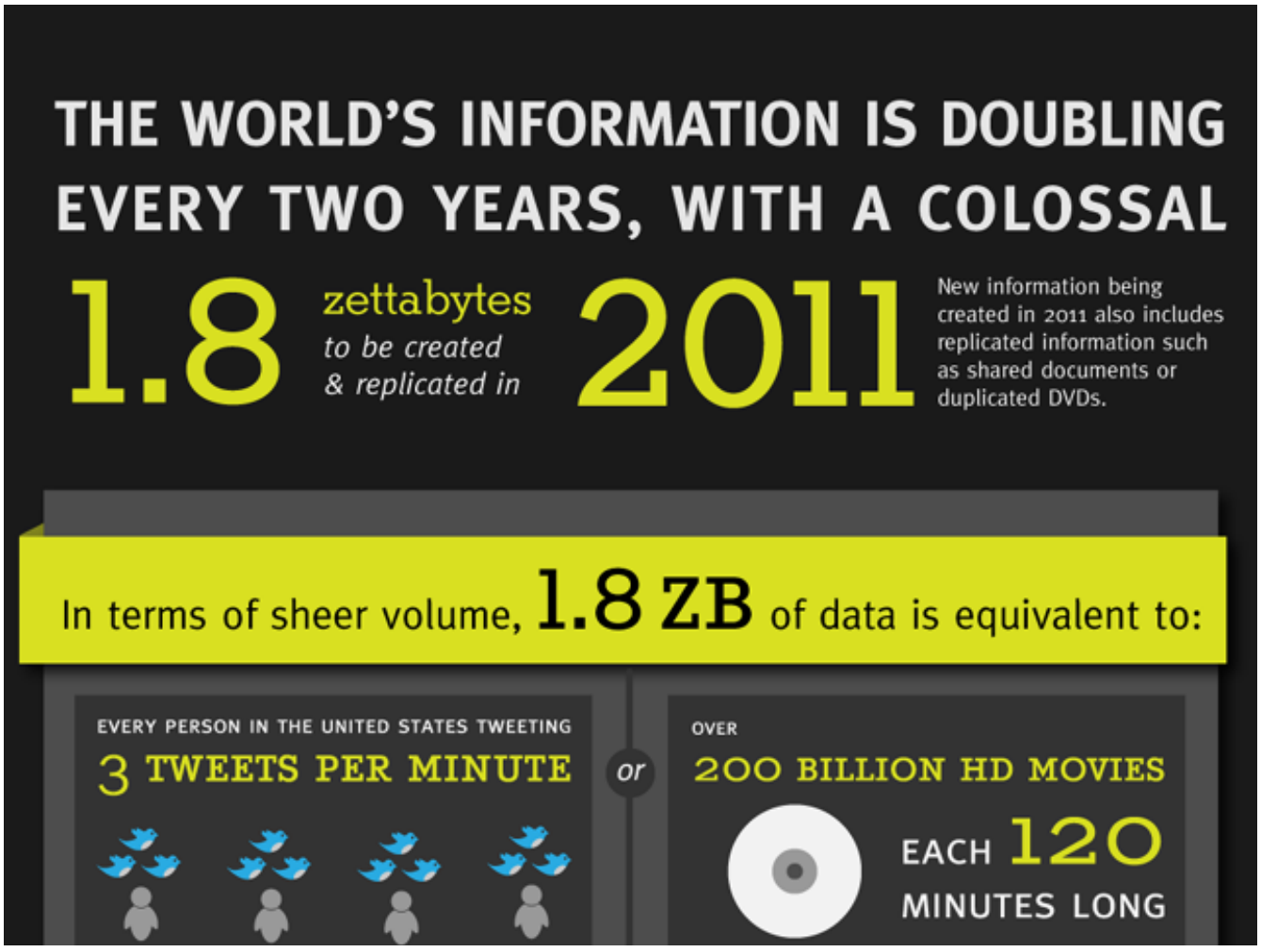


The “Big Data” Buzz – Why?

“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.” Eric Schmidt, Google (and others)







Storing 1.8 ZB of information would take:

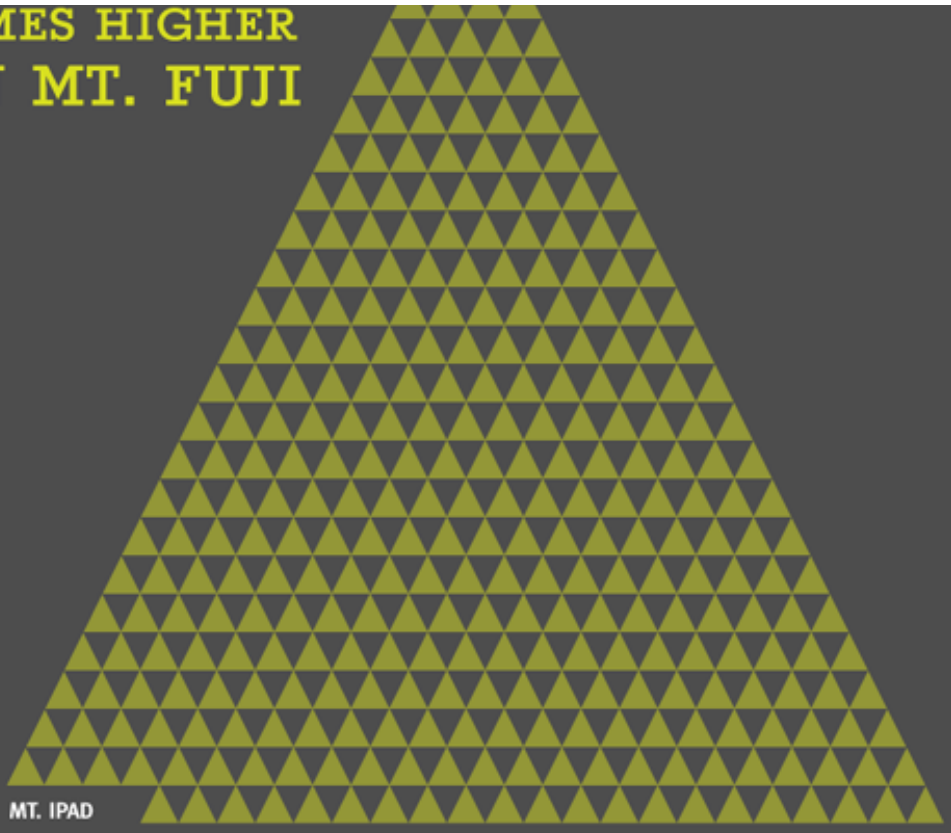
57.5 BILLION
32GB APPLE IPADS

1 = 10 billion

WITH THAT MANY IPADS WE COULD BUILD A MOUNTAIN OF IPADS THAT IS

25-TIMES HIGHER THAN MT. FUJI

25-TIMES HIGHER
THAN MT. FUJI

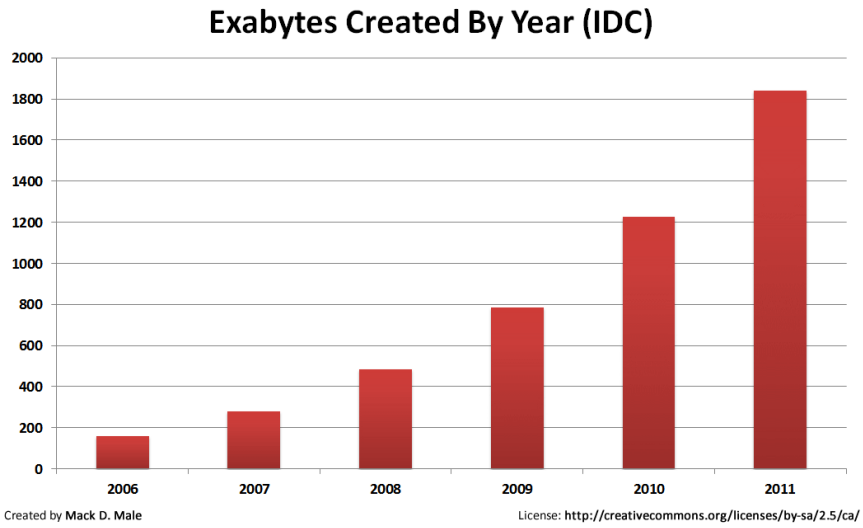


MT. FUJI ▲
3,776 m

MT. IPAD
94,400 m

SOURCES: *Extracting Value from Chaos* IDC Digital Universe study (sponsored by EMC Corporation)

The "Big Data" Buzz – Why?

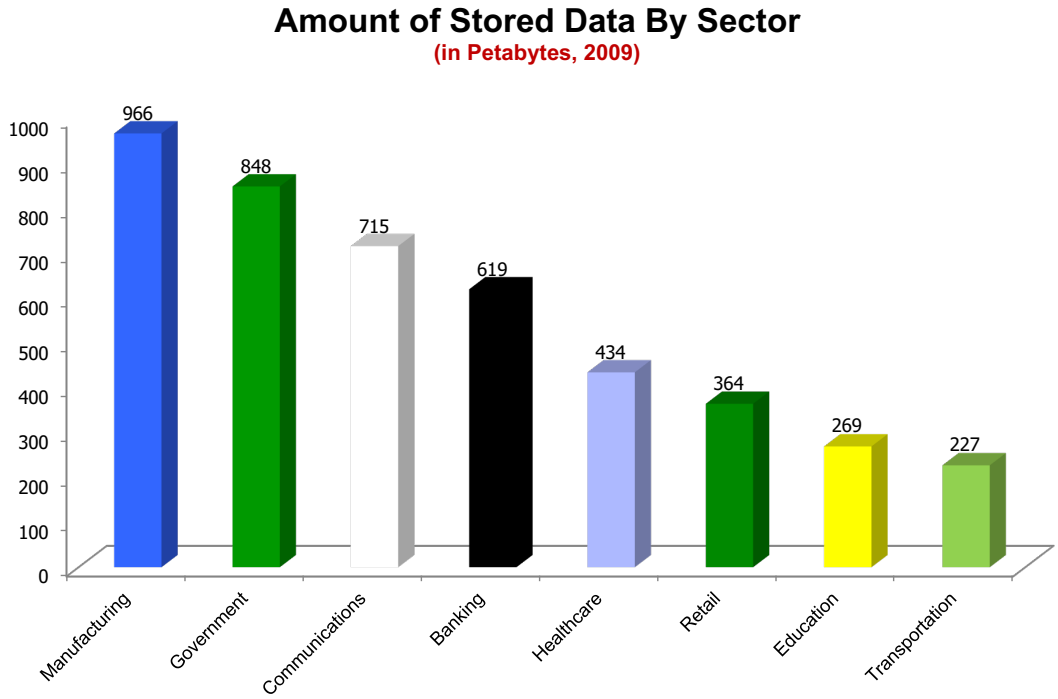


SI decimal prefixes		Binary usage
Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

"The sexy job in the next 10 years will be statisticians." "Data Scientists"?

Hal Varian
Prof. Emeritus UC Berkeley
Chief Economist, Google

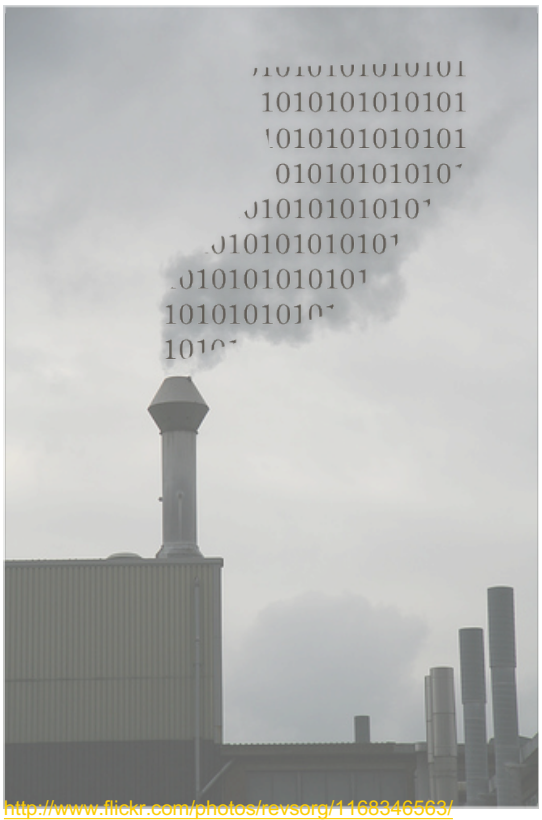
Some Numbers by Industry



Sources:
"Big Data: The Next Frontier for Innovation, Competition and Productivity."
US Bureau of Labor Statistics | McKinsley Global Institute Analysis

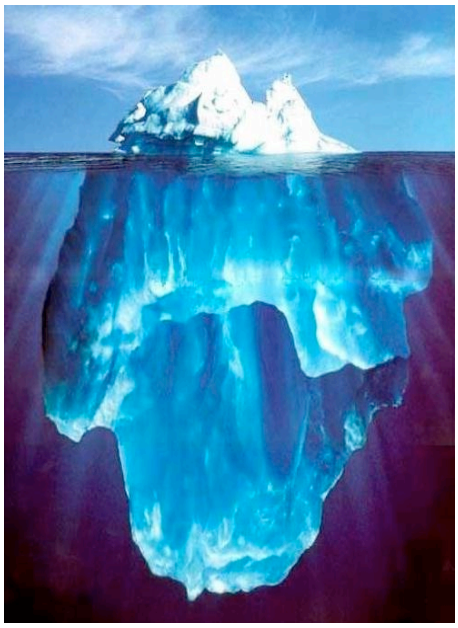
Industrial Revolution of Data!

- **UPC**
- **RFID**
- **GPS**
- **Sensornets**
- **Software Logs**
- **Microphones**
- **Cameras**
- **…**

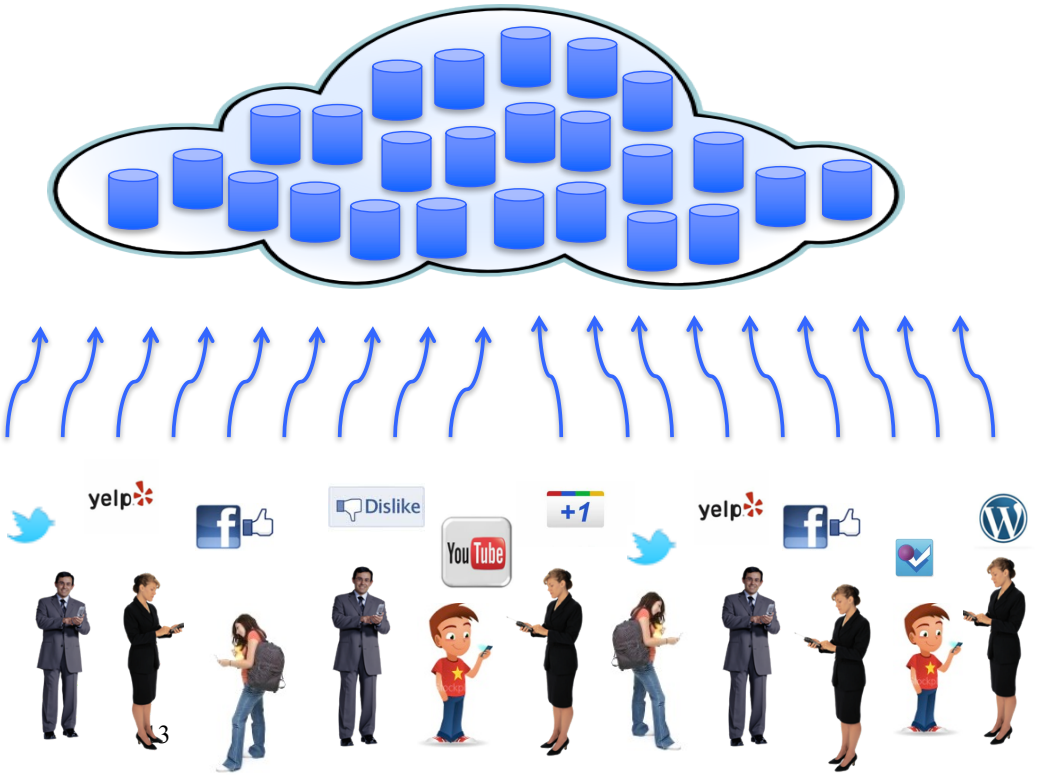


It's All Happening On-line

- **Every:**
  - Click
  - Ad impression
  - Wall post, friending, …
  - Billing event
  - Fast Forward, pause,…
  - Server request
  - Transaction
  - Network message
  - Fault
  - …
- **Generates Streams of Data that can be Analyzed**



User Generated Content

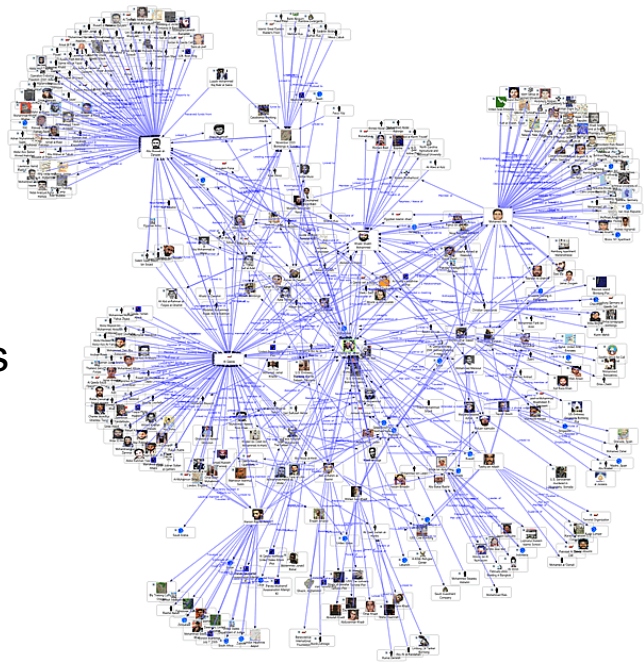


Credit: Mike Carey, UCI

Graph Data

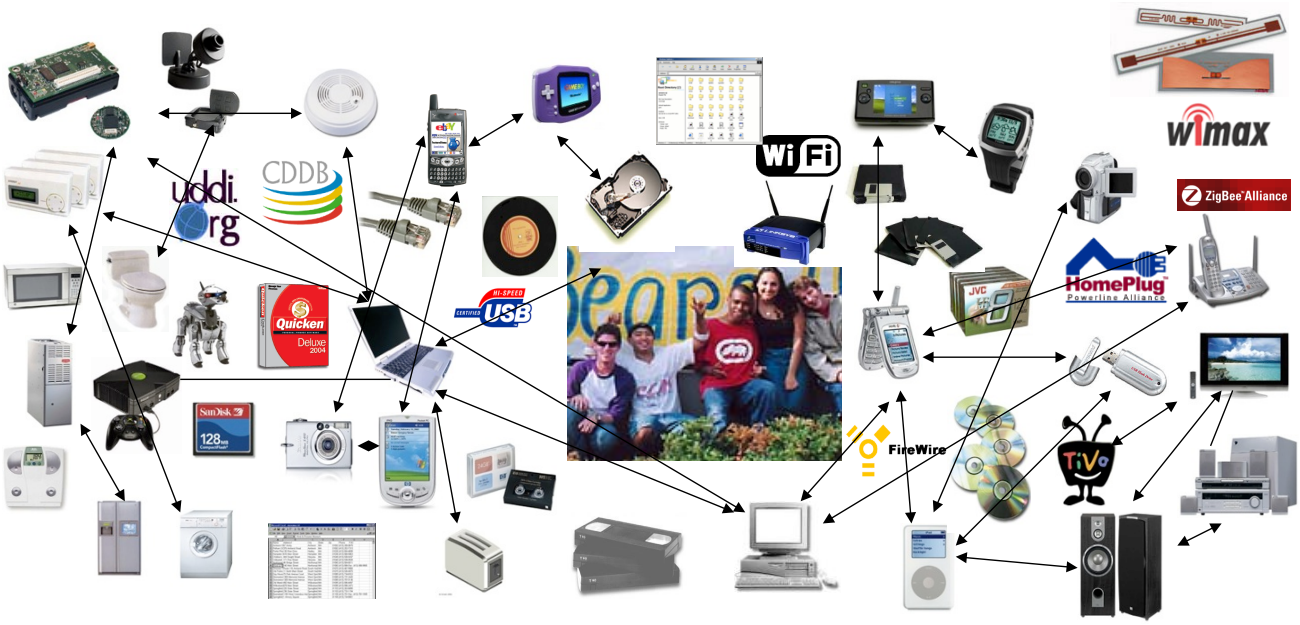
Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- …

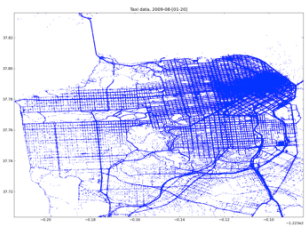


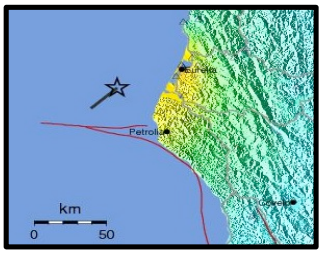
Some of these graphs can get quite large (e.g., Facebook's user graph)

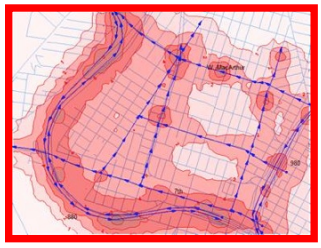
M2M - Internet of things



Fusion: e.g., NextGen Maps







Crowdsourcing + physical modeling + sensing + data assimilation

to produce:

From Alex Bayen, UCB

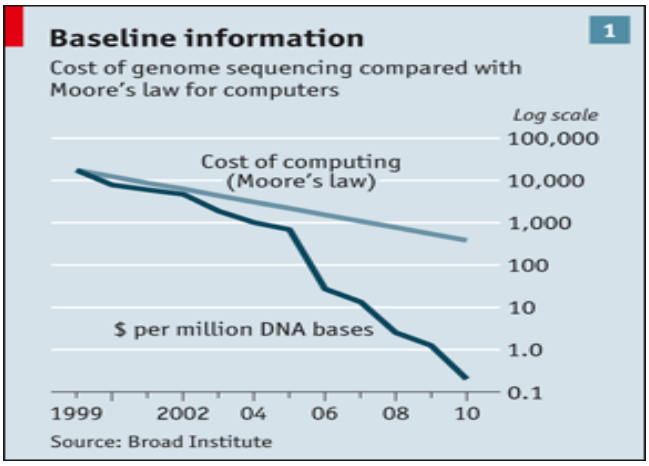
What can you do with the data

- **Reporting**
  - Post Hoc
  - Real time
- **Monitoring (fine-grained)**
- **Exploration**
- **Finding Patterns**
- **Root Cause Analysis**
- **Closed-loop Control**
- **Model construction**
- **Prediction**
- **…**

Big Data, Societal-Scale App?

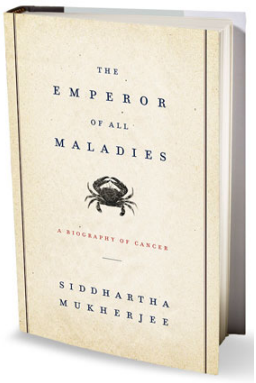
- **Cancer Tumor Genomics**
- **Vision: Personalized Therapy**
  - "…10 years from now, each cancer patient is going to want to get a genomic analysis of their cancer and will expect customized therapy based on that information."

Director, The Cancer Genome Atlas (TCGA), *Time Magazine,* 6/13/11



Opportunity or Obligation?

- Provocative Hypothesis: Given fast growing genomic databases, could CS now be a huge help in war on cancer?
- If a *chance* that we could help millions of cancer patients live longer and better lives, as moral people, aren't we obligated to try?



David Patterson, "Computer Scientists May Have What It Takes to Help Cure Cancer," *New York Times*, 12/5/2011

• UCSF cancer researchers + UCSC cancer genetic database + AMP Lab

The Cancer Genome Atlas: 5 PB = 20 cancers x 1000 genomes

So, In Summary…Why?

Data will be at the center of the major issues and events of your life.

As a computer professional, you'd better be on top of how to manage, use, and make sense of it.

Queries for Today

- Why?
- Who?
- **What?**
- How?
- For instance?

Who?

- **Haidar M. Harmanani**
  - Professor of Computer Science
  - Office: 810 Block A
  - Office Hours: TTh: 8:00-9:30 and 3:00-4:30
  - Email: haidar@lau.edu.lb
- **Course Website:**
  - http://vlsi.byblos.lau.edu.lb/classes/csc375/csc375.html
  - http://harmanani.github.io/csc375.html

What: Current Market

- **Relational DBMSs anchor the software industry**
  - Elephants: Oracle, IBM, Microsoft, Teradata, HP, EMC, …
  - Open source: MySQL, PostgreSQL
- **Obviously, Search**
  - Google & Bing
- **Open Source "NoSQL"**
  - Hadoop MapReduce
  - Key-value stores: Cassandra, Riak, Voldemort, Mongo, …
- **Cloud services**
  - Amazon, Google AppEngine, MS Azure, Heroku, …
- **Increasing use of custom code**

What will we learn?

- Design patterns for dealing with Big Data
- When, why and how to structure your data
- How MySQL and Oracle and (a bit of) Google work
- SQL ... and noSQL
- Managing concurrency
- Fault tolerance and Recovery
- Scaling out: parallelism and replication

- Audacity and Reverence.

What: Summing up

- Data is at the center of many things.
- For instance: computer science.

What: Summing up

You might think that we'll learn to apply computer science to Big Data.

The techniques we'll learn for Big Data are the key to scalable computer science.

Don’t forget Hal Varian’s prediction…

Google's Chief Economist

- **These professions barely have names:**
  - Cloud programmer
  - Data scientist
  - Scalable systems architect
  - Data-driven thinker
- **This will be a large fraction of the computing workforce.**

“ By 2018, the US could face a shortage of up to 190,000 workers with analytical skills” McKinsey Global Institute