Data and Big Data

Welcome to the Age of Information

Notes for CSC 100 - The Beauty and Joy of Computing
The University of North Carolina at Greensboro

Reminders

Big thing for this week:

Project Proposal Presentations: This Friday

Homework 3

- Should have completed online fractal tutorial
- Definitely should be playing around with drawing in BYOB
- HW 3 due: Wednesday, November 6

Data...

What is data? Is it the same as information?

> "You can have data without information, but you cannot have information without data." - Daniel Keys Moran

Data is being collected, generated, and stored far, far faster than ever before. How much?

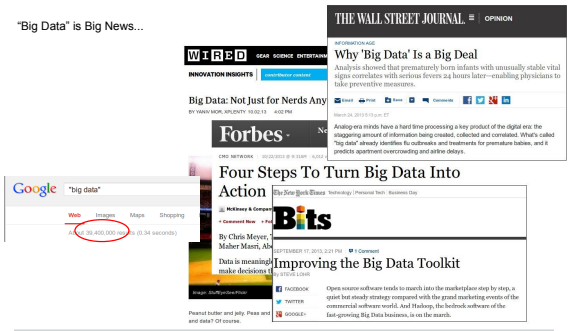
> "In 2012, every day **2.5 quintillion bytes of data** (1 followed by 18 zeros) are created, with 90% of the world's data created in the last two years alone." *http://marciaconner.com/blog/data-on-big-data/*

The result is a flood of data…

… or "Big Data"

Everyone is talking about “Big Data”

“Big Data” is Big News...



Measuring Data

The basic unit of information is a bit

… or a byte (usually 8 bits)

“usually”?!? Really? Really!!
Almost always 8 bits, but not always
When it’s important that we refer to 8 bits, the term used is “octet”

What is a kilobyte (kB)?
Memory (RAM) sizes must be a power of 2, so 1 kB was traditionally 2^{10}=1024 bytes
Different from SI units version of “kilo” (1000, as in kilometer, kilogram, …)
But it’s close!

So traditionally, 1 MB = 2^{20} bytes = 1,048,576 bytes ; 1 GB = 2^{30} B = 1,073,741,824 B
So off by over 7% for 1GB

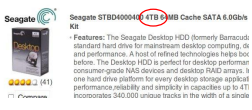
Hard drive manufacturers revolted!
Wanted to advertise a 229 B drive as 537 MB rather than 512 MB - back to SI units!

Now: RAM typically in power-of-two units (some suggest KiB/MiB/GiB for this), and persistent storage in SI units. What about flash drives? If it’s important, get clarification!

How much is...

1kB?
- Paragraph of text

1MB?
- 4 megapixel JPEG-compressed image



1GB?
- 30 minutes of DVD-quality SD TV
- 3.5 minutes at Blu-ray HD rate

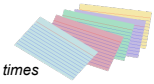
1TB?
- 2,000 hours of audio (uncompressed)
- 17,000 hours as MP3s (255,000 4-minute songs)

1PB?
- DNA of the entire population of the US - three times over!
- Two months data from the planned Large Synoptic Survey Telescope

See also http://www.jamesshuggins.com/h/tek1/how-big.htm

A Flood of Data

- Human genome: Just over 3 billion base pairs
  - *Typing in 12pt on 8.5x11 paper fits 2880 characters*
  - *So the human genome would be over a million pages (printed two-sided, an 86 foot high stack of paper)*
- Facebook (source: http://thesocialskinny.com/100-social-media-statistics-for-2012/)
  - *Around a billion users*
  - *Around 420 million status updates per day*
  - *On index cards, would be a stack 53 miles high!*
  - *... or end-to-end would stretch around the world 1.3 times*



- Large Synoptic Survey Telescope
  - *16 terabytes (16,000,000,000,000 bytes) will be captured per day*
  - *No human being will ever see most of this data*
- *Walmart customer transaction database*
  - *Estimated to be approximately 2.5 petabytes*

So much data available

Some publicly-available big datasets

Some examples of available data:

- data.gov
  - Over 91,000 datasets on Oct 30, 2013
  - Census data, USGS Topo maps, house price indexes, NOAA Geophysical Data Center, …
- Amazon web services public data sets (http://aws.amazon.com/datasets)
  - Web crawl of over 5 billion web pages
  - "1000 Genomes Project"
  - Japan Census Data
  - Google Books Ngrams
- UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html)
  - All sorts of data for machine learning experiments
- NCBI (http://www.ncbi.nlm.nih.gov/genome)
  - Genome information including sequences of many organisms
  - From the U.S. National Library of Medicine

Making sense out of data - viewing

Visualization

How can we view data? A couple of on-line examples:

"Many Eyes" hosted by IBM:

http://www-958.ibm.com/software/data/cognos/manyeyes/

D3: Data-Drive Documents

Technology allows interactive data presentation

Example: http://benschmidt.org/Degrees/

D3 toolkit: http://d3js.org/

Other viz tools:

http://selection.datavisualization.ch/

Making sense out of data - processing

Data mining: Finding patterns

What do you do with lots of data - find patterns!

> An old urban legend: *A supermarket analyzed purchasing data and found a correlation between purchases of diapers and beer that no one knew about. They put the two closer together in the store and….. profit!*
>
> The real story: *Almost true… It was Osco Drug stores, not a supermarket. And while they found the correlation (between 5:00 and 7:00pm) they didn't actually change anything as a result - that was just a "what if…" comment that became legend.*

There are certainly real stories that are even more astounding:

> http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
>
> From the story: "*As Pole's computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a "pregnancy prediction" score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.*"
>
> See also: Video at the end of this lecture.

Data Analytics

A valuable and growth-area skill

Job postings mentioning "Big Data":

Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows the percentage of jobs we find that contain your search terms.

Valuable contests to demonstrate skills and develop techniques:

> http://www.kaggle.com/

Summary

Main take-aways:

- More data than ever before being collected and used
- Must be able to manage the data
- Making sense out of the data is a very valuable skill
  - Analysis, mining, and visualization are all parts of this

What you should have gotten from this lecture:

- A sense for data sizes
- An idea of what data is out there: available and private
- Some ideas and pointers for how data is used

And finally… a video

Relevant information:

- *NY Times Magazine* story on data mining
  http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
- Forbes story "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did":
  http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

The entertaining take:

http://www.colbertnation.com/the-colbert-report-videos/408981/february-22-2012/the-word---surrender-to-a-buyer-power