Introduction to Big Data with Apache Spark

BerkeleyX

This Lecture

Course Goals

Brief History of Data Analysis

Big Data and Data Science – Why All the Excitement?

Where Big Data Comes From

Course Goals

1. Learn about Data Science
   - » Where Big Data Comes from
   - » Observation and Experimentation
   - » The Elements of Data Science
     - Data Acquisition
     - Data Preparation
     - Analysis
     - Data Presentation
     - Data Products

Course Goals

2. Learn how to perform Data Science
   - » Understanding Data Quality
   - » Cleaning and manipulating datasets
   - » Using and parsing data representations
   - » Using basic Machine Learning algorithms and libraries
   - » Writing big data applications
   - » Performing Exploratory Data Analysis

Course Goals

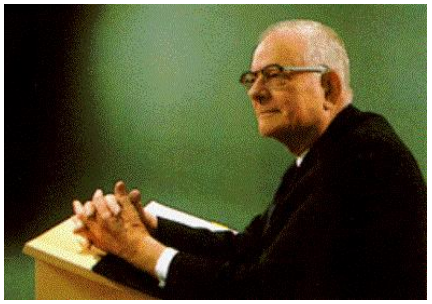
3. Learn to write Apache Spark programs
   - » History and development
   - » Conceptual model
   - » How the Spark cluster model works
   - » Spark essentials (transformations, actions, persistence, broadcast variables, accumulators, Key-Value pairs, pySpark API)
   - » Debugging Spark programs
   - » Using Spark `mllib` for Machine Learning

Brief Data Analysis History

- R. A. Fisher
  - » 1935: "The Design of Experiments"

"correlation does not imply causation"

- W. E. Demming
  - » 1939: "Quality Control"

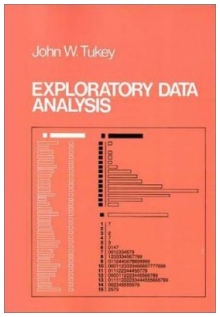


Images: http://culturacientifica.wikispaces.com/CONTRIBUCIONES+DE+SIR+RONALD+FISHER+A+LA+ESTADISTICA+GENETICA
http://es.wikipedia.org/wiki/William_Edwards_Deming

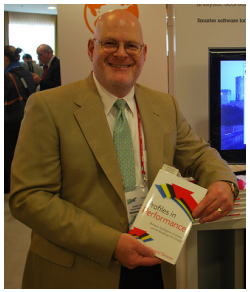
Brief Data Analysis History

- Peter Luhn
  - » 1958: "A Business Intelligence System"

- John W. Tukey
  - » 1977: "Exploratory Data Analysis



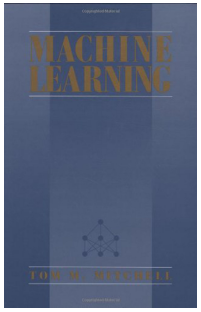
- Howard Dresner
  - » 1989: "Business Intelligence"



Images: http://www.businessintelligence.info/definiciones/business-intelligence-system-1958.html
http://www.betterworldbooks.com/exploratory-data-analysis-id-0201076160.aspx
https://www.flickr.com/photos/42266634@N02/4621418442

Brief Data Analysis History

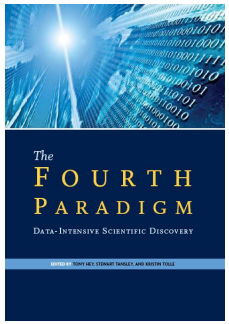
- Tom Mitchell
  - » 1997: "Machine Learning book"



- Google
  - » 1996: "Prototype Search Engine"

Google

- Data-Driven Science eBook
  - » 2007: "The Fourth Paradigm"



Images: http://www.amazon.com/Machine-Learning-Tom-M-Mitchell/dp/0070428077
http://www.google.com/about/company/history/
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Brief Data Analysis History

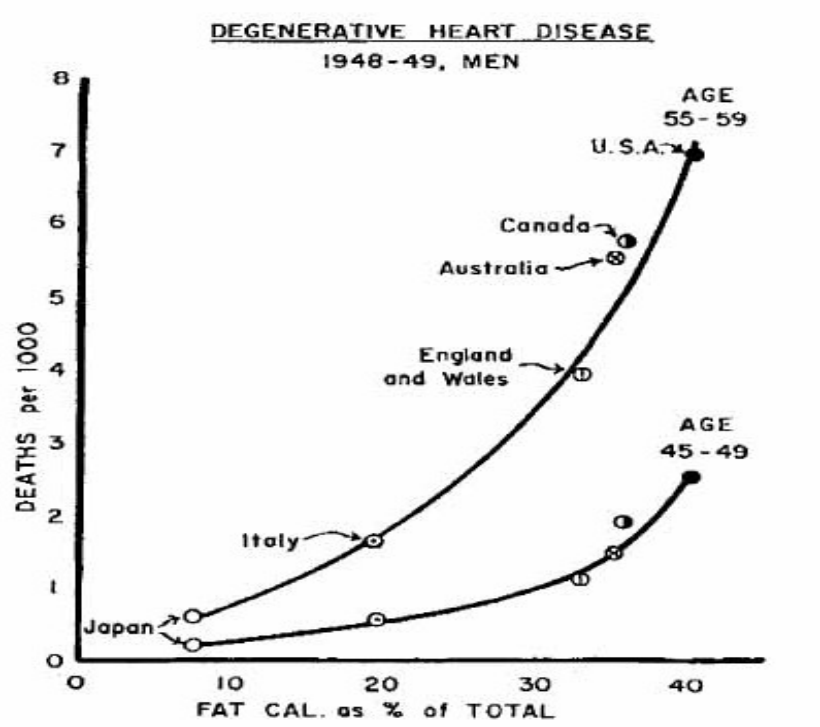
- Peter Norvig
  - » 2009: "The Unreasonable Effectiveness of Data"

- Exponential growth in data volume
  - » 2010: "The Data Deluge"

Images: http://en.wikipedia.org/wiki/Peter_Norvig
http://www.economist.com/node/15579717

Data Makes Everything Clearer (part I)?

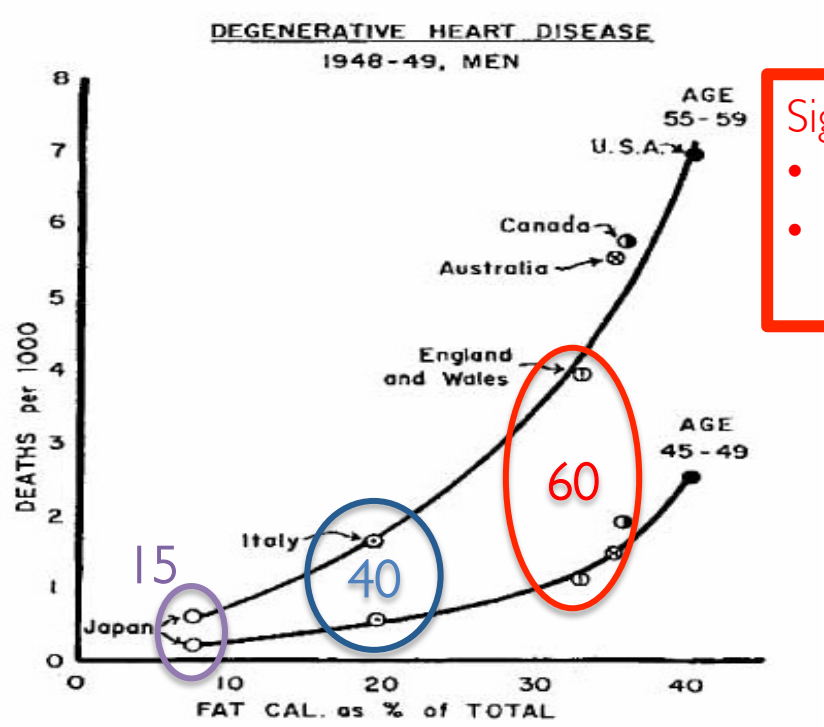
- Seven Countries Study (Ancel Keys)
  - » Started in 1958, followed13,000 subjects total for 5-40 years



http://en.wikipedia.org/wiki/Seven_Countries_Study

Data Makes Everything Clearer (part I)?

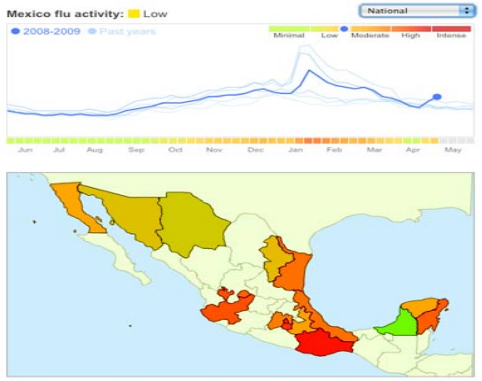
- Seven Countries Study (Ancel Keys)
  - » Started in 1958, followed 13,000 subjects total for 5-40 years



Significant controversy
- Only studied subset of 21 countries with data
- Failed to consider other factors (e.g., per capita annual sugar consumption in pounds)

http://en.wikipedia.org/wiki/Seven_Countries_Study

Big Data: Why all the Excitement?



- Nowcasting vs Forecasting

- Example – Google Flu Trends:
  - » February 2010 detected outbreak two weeks ahead of CDC data
  - » Initially 97% accurate but overestimated during 2011-13 including one interval in 2012-13 period where GFT was off by 2x
  - » New models are estimating which cities are most at risk for spread of the Ebola virus

https://www.google.org/flutrends/

Why All the Excitement?

elections2012

Live results | President | Senate | House | Governor | Choose your

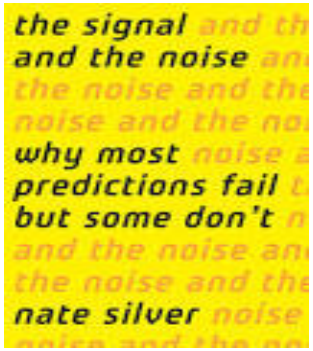
Numbers nerd Nate Silver's forecasts prove all right on election night

FiveThirtyEight blogger predicted the outcome in all 50 states, assuming Barack Obama's Florida victory is confirmed

Luke Harding
guardian.co.uk, Wednesday 7 November 2012 10.45 EST

http://www.theguardian.com/world/2012/nov/07/nate-silver-election-forecasts-right

USA 2012 Presidential Election



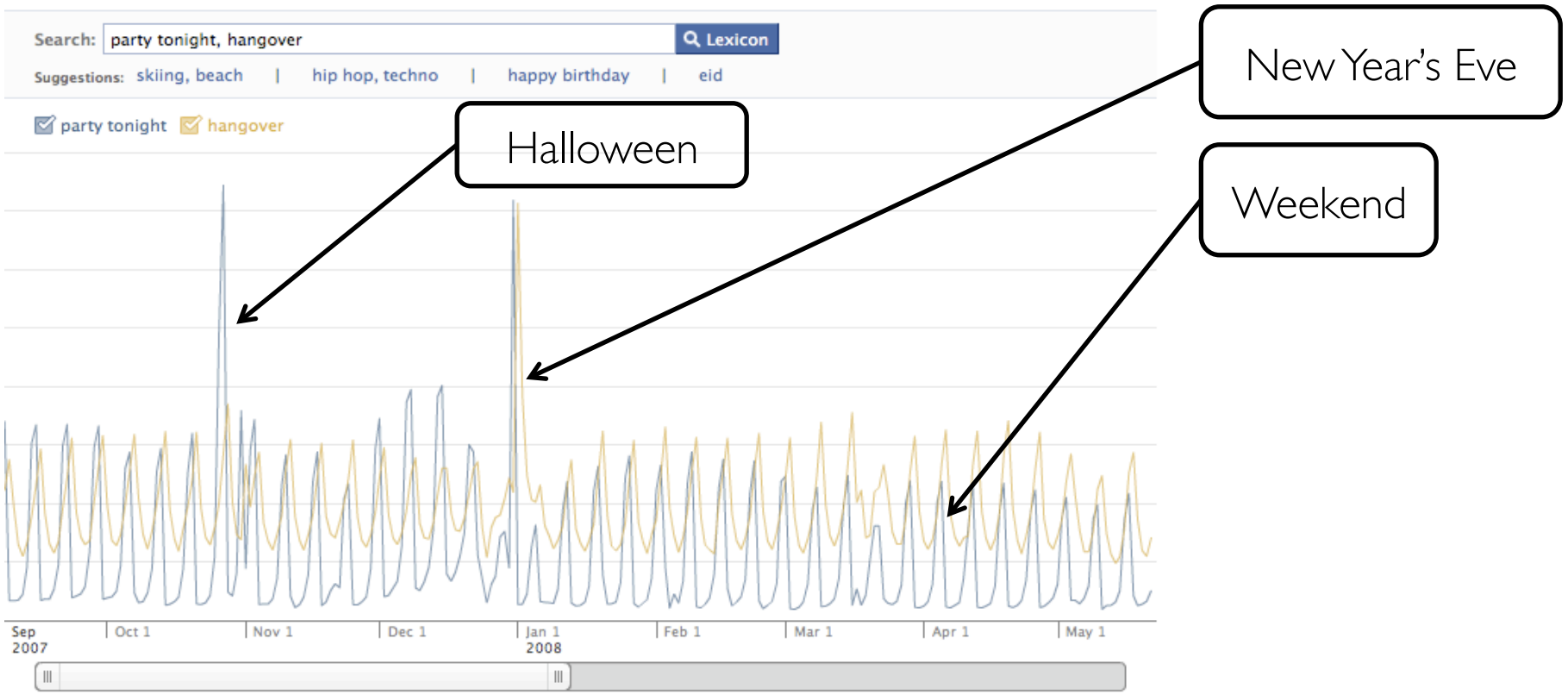
Big Data and Election 2012 (cont.)

…that was just one of several ways that Mr. Obama's campaign operations, some unnoticed by Mr. Romney's aides in Boston, helped save the president's candidacy. In Chicago, the campaign recruited a team of behavioral scientists to build an extraordinarily sophisticated database

…that allowed the Obama campaign not only to alter the very nature of the electorate, making it younger and less white, but also to create a portrait of shifting voter allegiances. The power of this operation stunned Mr. Romney's aides on election night, as they saw voters they never even knew existed turn out in places like Osceola County, Fla.

New York Times, Wed Nov 7, 2012

Example: Facebook Lexicon



Example: Facebook Lexicon

Data Makes Everything Clearer (part II)?

Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]
1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA
*** E-mail: Corresponding spechler@princeton.edu**

Abstract

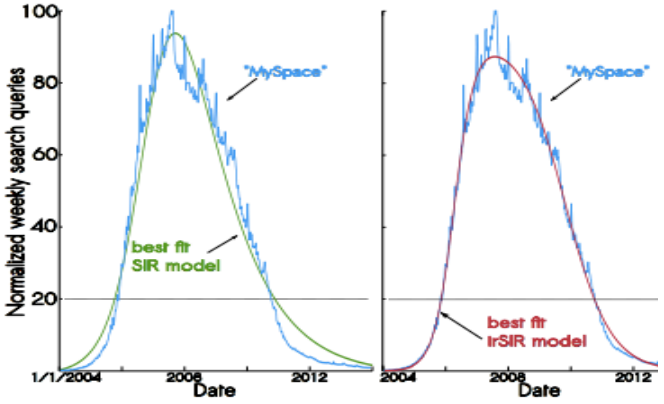
The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

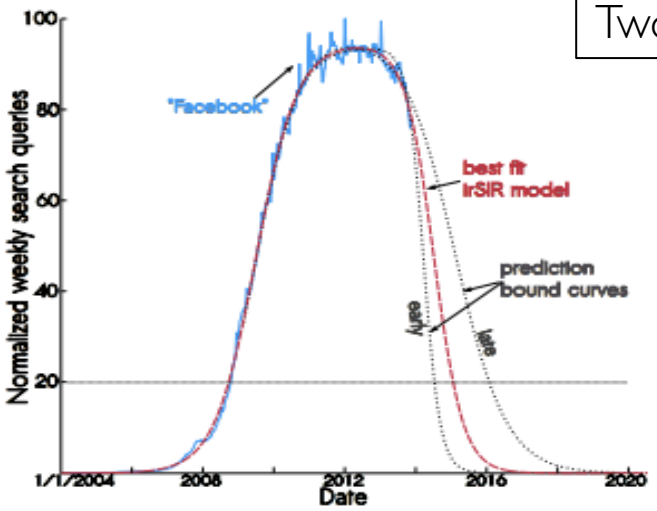
"Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years."

http://arxiv.org/abs/1401.4208

Data Makes Everything Clearer (part II)?

Google Trends searches for "MySpace"



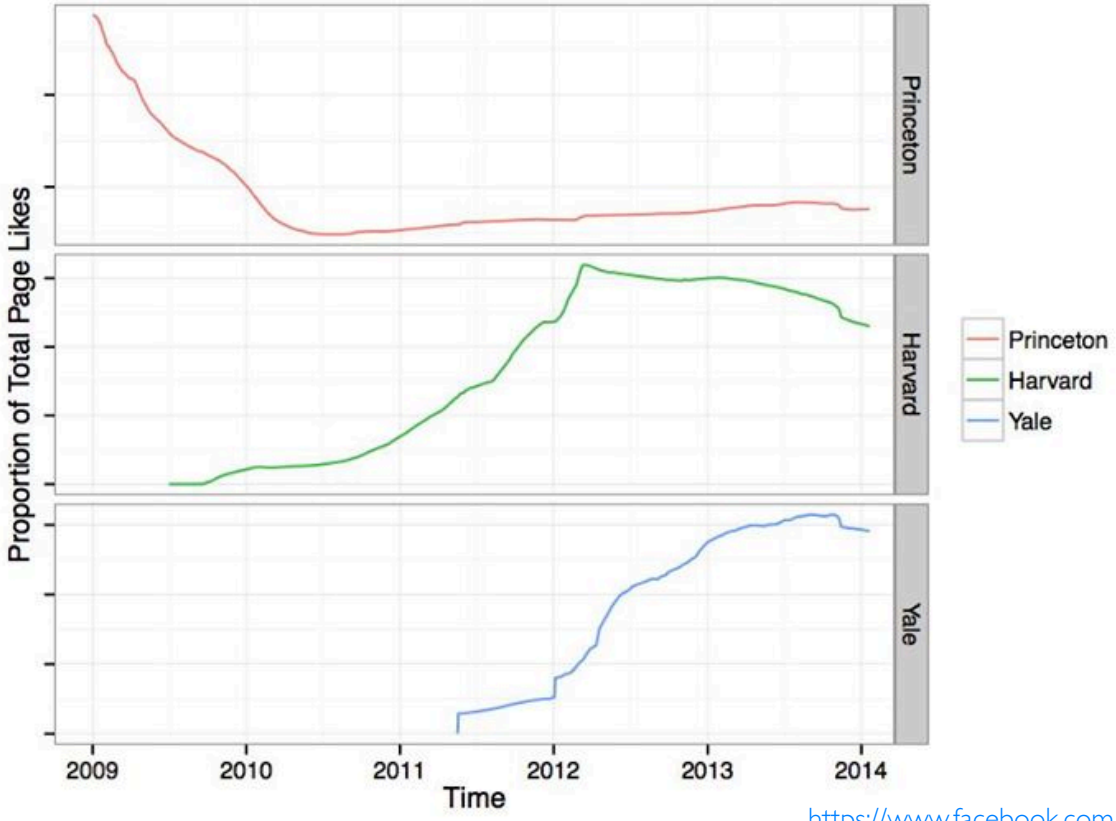


Two Figures from the paper

Searches for "Facebook"

http://arxiv.org/abs/1401.4208

Data Makes Everything Clearer (part II)?



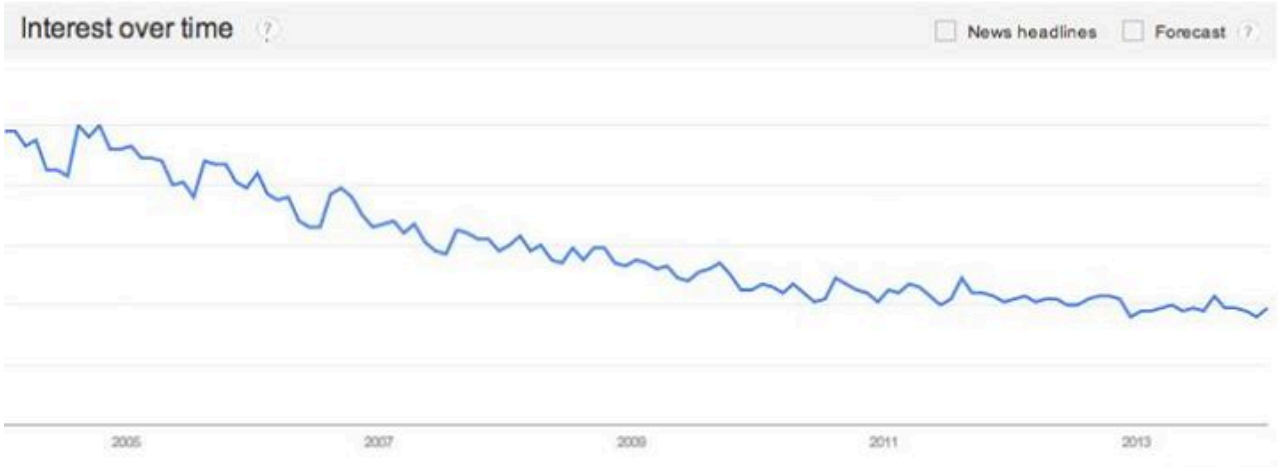
In keeping with the scientific principle "correlation equals causation," our research unequivocally demonstrated that Princeton may be in danger of disappearing entirely.

https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849

Data Makes Everything Clearer (part II)?

… and based on Princeton search trends:

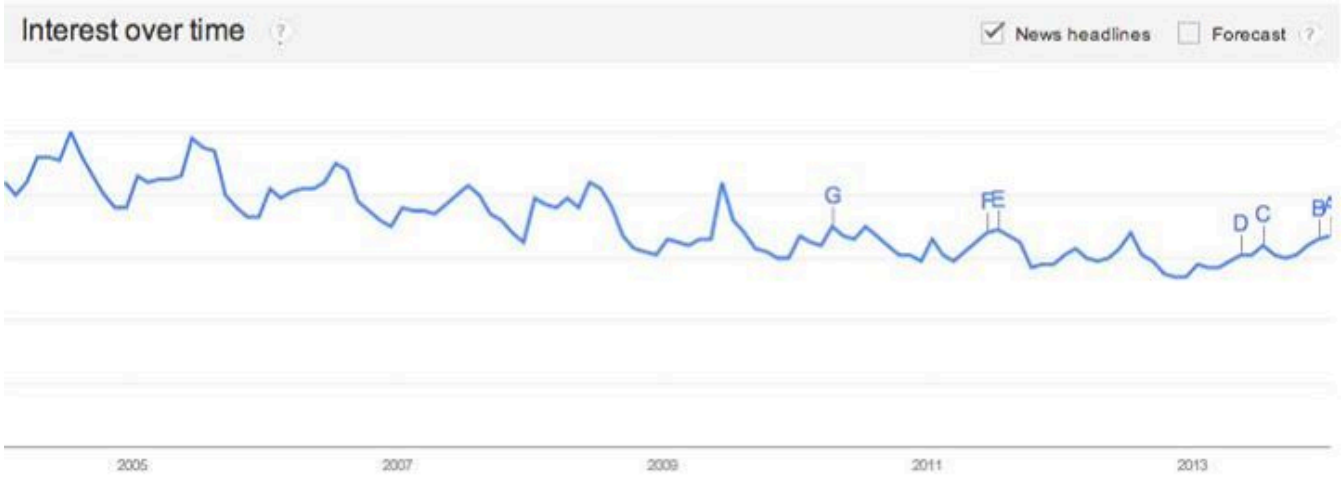
"This trend suggests that Princeton will have only half its current enrollment by 2018, and by 2021 it will have no students at all,…"



https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849

Data Makes Everything Clearer (part II)?

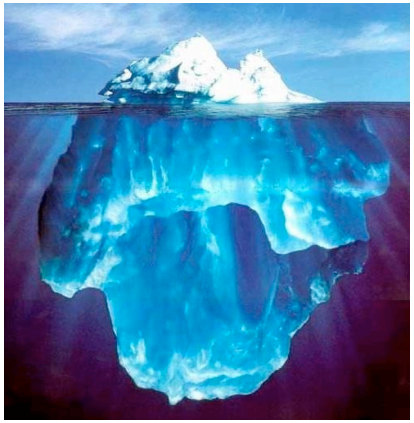
While we are concerned for Princeton University, we are even more concerned about the fate of the planet — Google Trends for "air" have also been declining steadily, and our projections show that by the year 2060 there will be no air left:



https://www.facebook.com/notes/mike-develin/debunking-princeton/10151947421191849

Where Does Big Data Come From?

- It's all happening online – could record every:
  - » Click
  - » Ad impression
  - » Billing event
  - » Fast Forward, pause,…
  - » Server request
  - » Transaction
  - » Network message
  - » Fault
  - » …

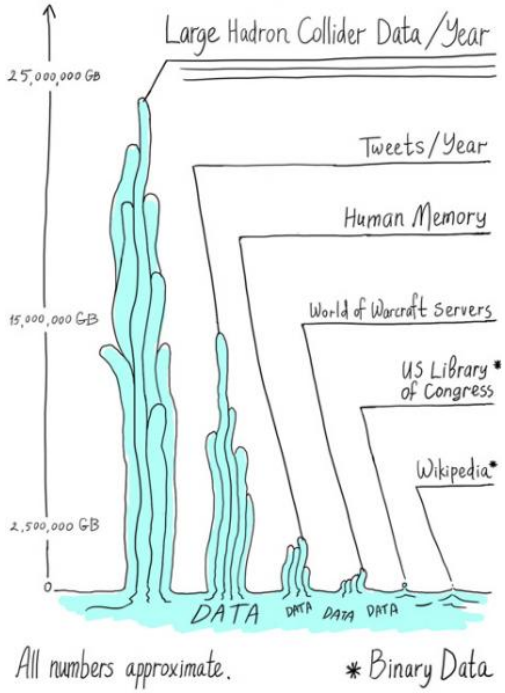


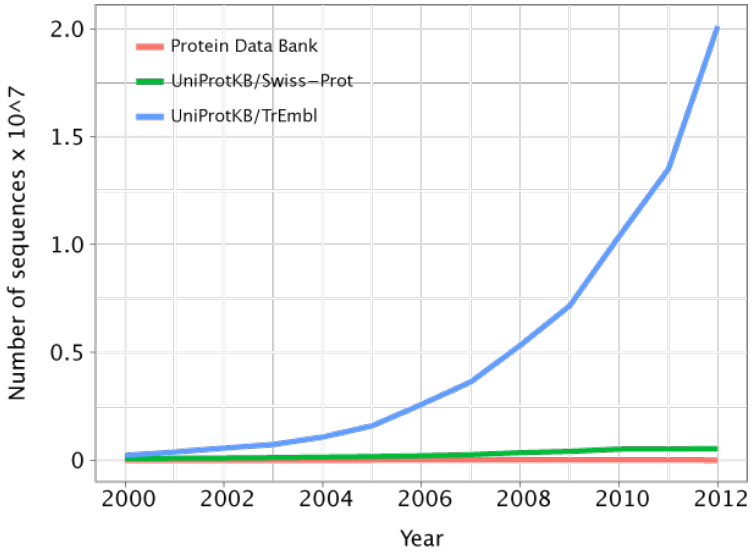
Where Does Big Data Come From?

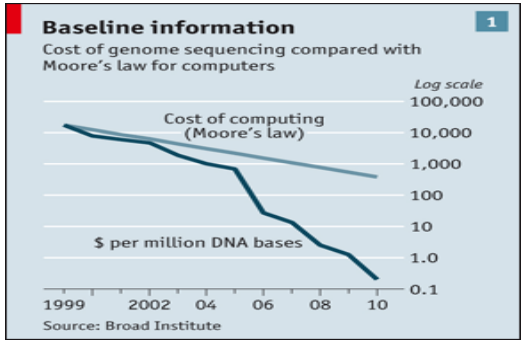
- User Generated Content (Web & Mobile)
  - » Facebook
  - » Instagram
  - » Yelp
  - » TripAdvisor
  - » Twitter
  - » YouTube
  - » …

Where Does Big Data Come From?

- Health and Scientific Computing

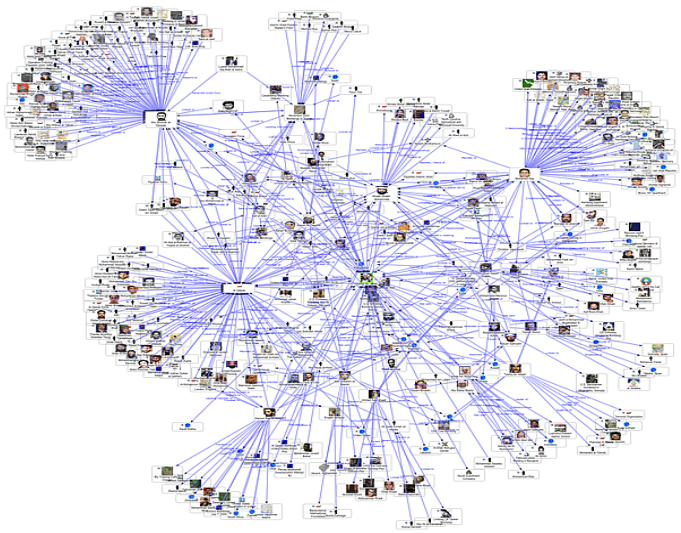






Images: http://www.economist.com/node/16349358
http://gorbi.irb.hr/en/method/growth-of-sequence-databases/
http://www.symmetrymagazine.org/article/august-2012/particle-physics-tames-big-data

Graph Data



Lots of interesting data has a graph structure:

- Social networks
- Telecommunication Networks
- Computer Networks
- Road networks
- Collaborations/Relationships
- …

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

Log Files – Apache Web Server Log

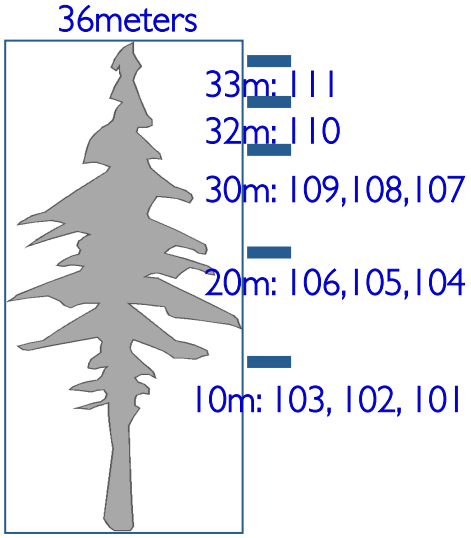
```
uplherc.upl.com - - [01/Aug/1995:00:00:07 -0400] "GET / HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/ksclogo-medium.gif
HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/MOSAIC-logosmall.gif
HTTP/1.0" 304 0
uplherc.upl.com - - [01/Aug/1995:00:00:08 -0400] "GET /images/USA-logosmall.gif HTTP/
1.0" 304 0
ix-esc-ca2-07.ix.netcom.com - - [01/Aug/1995:00:00:09 -0400] "GET /images/launch-
logo.gif HTTP/1.0" 200 1713
uplherc.upl.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/WORLD-logosmall.gif
HTTP/1.0" 304 0
slppp6.intermind.net - - [01/Aug/1995:00:00:10 -0400] "GET /history/skylab/
skylab.html HTTP/1.0" 200 1687
piweba4y.prodigy.com - - [01/Aug/1995:00:00:10 -0400] "GET /images/launchmedium.gif
HTTP/1.0" 200 11853
tampico.usc.edu - - [14/Aug/1995:22:57:13 -0400] "GET /welcome.html HTTP/1.0" 200 790
```

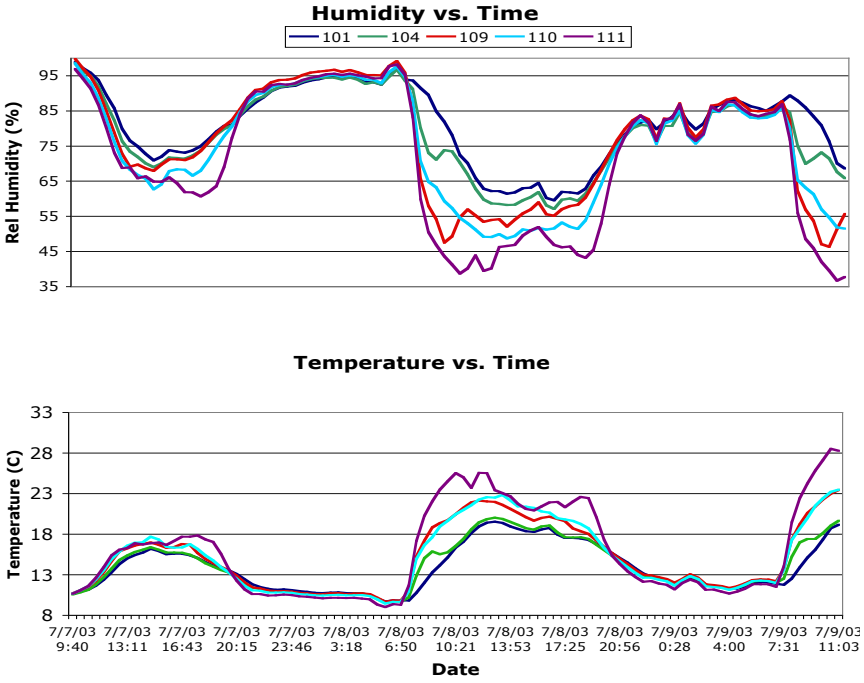
Machine Syslog File

```
dhcp-47-129:CS100_1> syslog -w 10
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting
read:]: unexpected field ID 23 with type 8.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMUser read:]:
unexpected field ID 17 with type 12.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -
[EDAMAuthenticationResult read:]: unexpected field ID 6 with type 11.
Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -
[EDAMAuthenticationResult read:]: unexpected field ID 7 with type 11.
Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting
read:]: unexpected field ID 19 with type 8.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMAccounting
read:]: unexpected field ID 23 with type 8.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMUser read:]:
unexpected field ID 17 with type 12.  Skipping.
Feb  3 15:18:11 dhcp-47-129 Evernote[1140] <Warning>: -[EDAMSyncState read:]:
unexpected field ID 5 with type 10.  Skipping.
Feb  3 15:18:49 dhcp-47-129 com.apple.mtmd[47] <Notice>: low priority

thinning needed for volume Macintosh HD (/) with 18.9 <= 20.0 pct free space
```

Internet of Things: Example Measurements





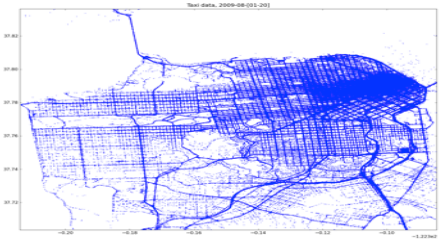
Redwood tree humidity and temperature at various heights

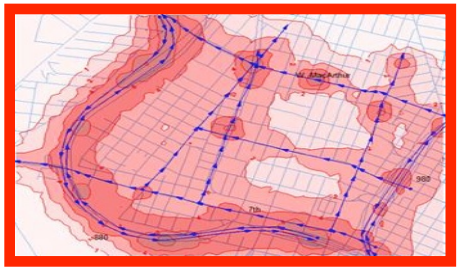
Internet of Things: RFID tags

- California FasTrak Electronic Toll Collection transponder
- Used to pay tolls
- Collected data also used for traffic reporting
  - » http://www.511.org/

http://en.wikipedia.org/wiki/FasTrak

What Can You do with Big Data?





Crowdsourcing + Physical modeling + Sensing + Data Assimilation

=

http://traffic.berkeley.edu