Data Science

Gaurav Sood

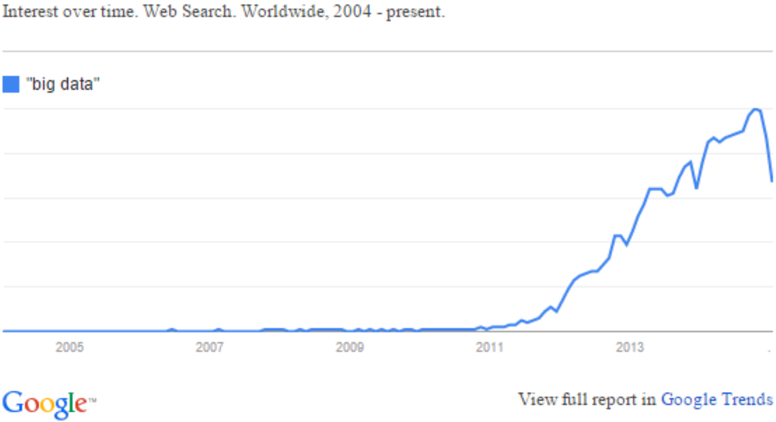
Spring 2015

Big Data

Big Data

Lots of hype recently.

Big Data



Interest over time. Web Search. Worldwide, 2004 - present.
"big data"
2005
2007
2009
2011
2013
Google™
View full report in Google Trends

Big Data

But where's the cheese?

Some examples

Fishing Out Fishy Figures


Previous | Next | Latest Americas view

All latest updates

Statistics in Argentina

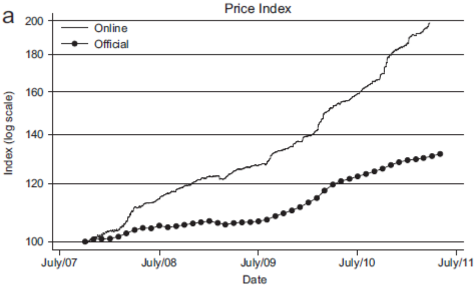
Fishy figures

Sep 17th 2014, 19:53 BY H.C. | BUENOS AIRES

Timekeeper Like 107 Tweet 9

"The CPINu's August inflation figure of 1.3% is less than half the 2.65% of the CPI Congreso, a compilation of private estimates gathered by opposition members of Congress." (Economist)

Fishing Out Fishy Figures



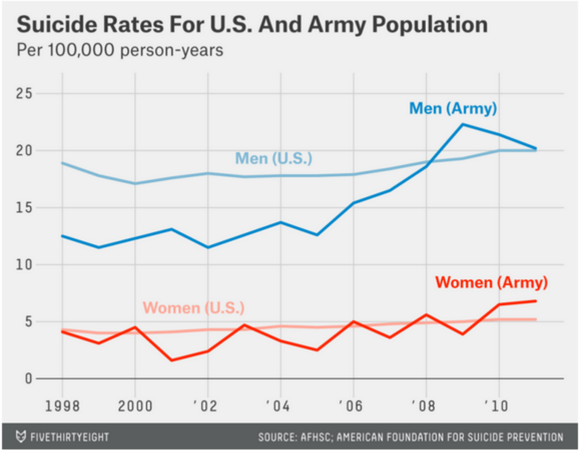
Source: Online vs Official Price Indexes: Measuring Argentina's Inflation By Alberto Cavallo

Suicide Prevention in the Army

Suicide Prevention in the Army

"In 2012, more soldiers committed suicide than died while fighting in Afghanistan: 349 suicides compared to 295 combat deaths."

Suicide Prevention in the Army



Suicide Prevention in the Army

"Research has repeatedly shown that doctors are not accurate in predicting who is at risk of suicide."

Suicide Prevention in the Army

"The soldiers with the highest 5 percent of risk scores committed over half of all suicides in the period covered — at an extraordinary rate of about 3,824 suicides per 100,000 person-years."

538 Article
STARRS paper

Reducing Crime

Minority Report

Reducing Crime

Predictive, ‘CompStat’, ‘HotSpot’ Policing

Reducing Crime

PredPol: Predictive Policing

LAPD, Atlanta PD

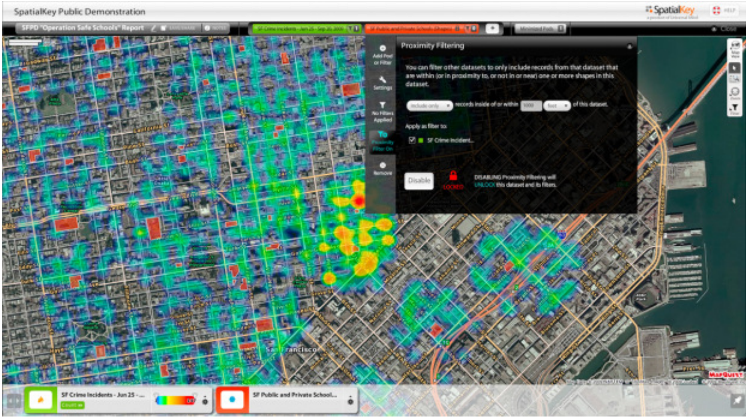
Based off earthquake prediction algorithm

Reducing Crime

"During a four-month trial in Kent, 8.5% of all street crime occurred within PredPol's pink boxes, with plenty more next door to them; predictions from police analysts scored only 5%. An earlier trial in Los Angeles saw the machine score 6% compared with human analysts' 3%."

Economist

Reducing Crime



Web Search

Web Search

What's New
Check Email
YAHOO!
Personalize
Help
Yahoo! Pager
instant messaging
Yahoo! Pager
now works with chat
Yahoo! Mail
free email for life
Search
advanced search
Yahoo! Auctions - 1000's of items to bid on - Pokemon, Beanie Babies, video games, Furbys...
Shopping - Yellow Pages - People Search - Maps - Travel Agent - Classifieds - Personals - Games - Chat
Email - Calendar - Pager - My Yahoo! - Today's News - Sports - Weather - TV - Stock Quotes - more...
Arts & Humanities
Literature, Photography...
News & Media
Full Coverage, Newspapers, TV...
Business & Economy
Companies, Finance, Jobs...
Recreation & Sports
Sports, Travel, Autos, Outdoors...
Computers & Internet
Internet, WWW, Software, Games...
Reference
Libraries, Dictionaries, Quotations...
Education
College and University, K-12...
Regional
Countries, Regions, US States...
Entertainment
Cool Links, Movies, Humor, Music...
Science
Biology, Astronomy, Engineering...
Government
Military, Politics, Law, Taxes...
Social Science
Archaeology, Economics, Languages...
Health
Medicine, Diseases, Drugs, Fitness...
Society & Culture
People, Environment, Religion...
In the News
· NATO - Serbia war
· Giant bacterium discovered
· Lakers release Rodman
more...
Marketplace
· Charity Auctions - for the Kosovo relief effort
· Find a new job!
more...
Inside Yahoo!
· Y! Movies - showtimes, reviews
· Y! Clubs - create your own
· Y! Visa - instant credit while you wait
more...

Web Search

– Human Curation, Ad-hoc automation

Web Search

- Human Curation, Ad-hoc automation
- Google crawls over 20 billion URLs a day (Sullivan 2012).

Web Search

- Human Curation, Ad-hoc automation
- Google crawls over 20 billion URLs a day (Sullivan 2012).
- Google answers 100 billion search queries a month (Sullivan 2012).

Web Search

- Human Curation, Ad-hoc automation
- Google crawls over 20 billion URLs a day (Sullivan 2012).
- Google answers 100 billion search queries a month (Sullivan 2012).
- "... a typical search returns results in less than 0.2 seconds" (Google)

Web Search

- Human Curation, Ad-hoc automation
- Google crawls over 20 billion URLs a day (Sullivan 2012).
- Google answers 100 billion search queries a month (Sullivan 2012).
- “. . . a typical search returns results in less than 0.2 seconds” (Google)
- Page Rank

Side-effects of Drugs

Side-effects of Drugs

"Adverse drug events cause substantial morbidity and mortality and are often discovered after a drug comes to market."

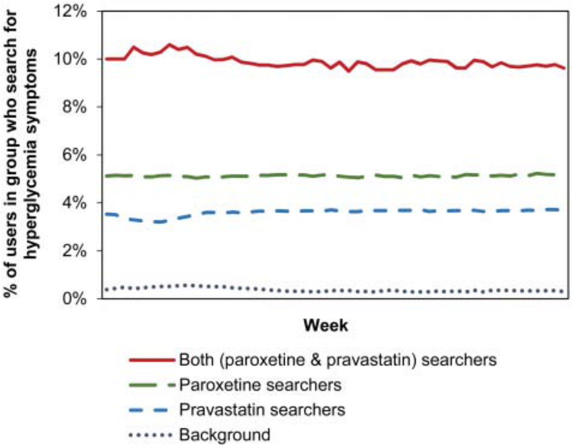
Side-effects of Drugs

FDA collects this information from “physicians, pharmacists, patients, and drug companies” but these reports “are incomplete and biased”

Side-effects of Drugs

"paroxetine and pravastatin, whose interaction was reported to cause hyperglycemia after the time period of the online logs used in the analysis"

Side-effects of Drugs



Web-scale Pharmacovigilance: Listening to Signals from the Crowd. By White et al.

Flu Season

How many got the sniffles?

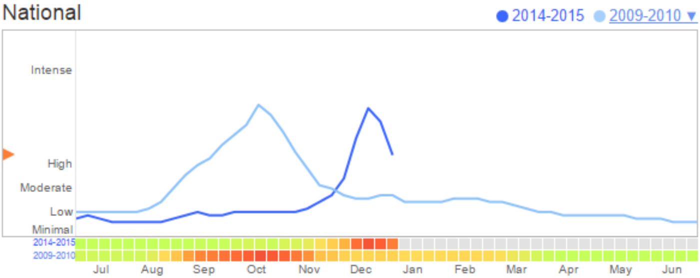
Flu Season

How many got the sniffles in the past month?

Flu Season

Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. [Learn more »]

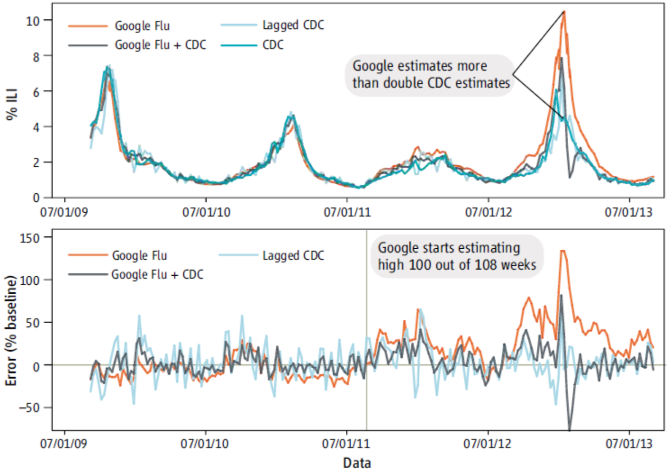


Google Flu Trends

Flu Season

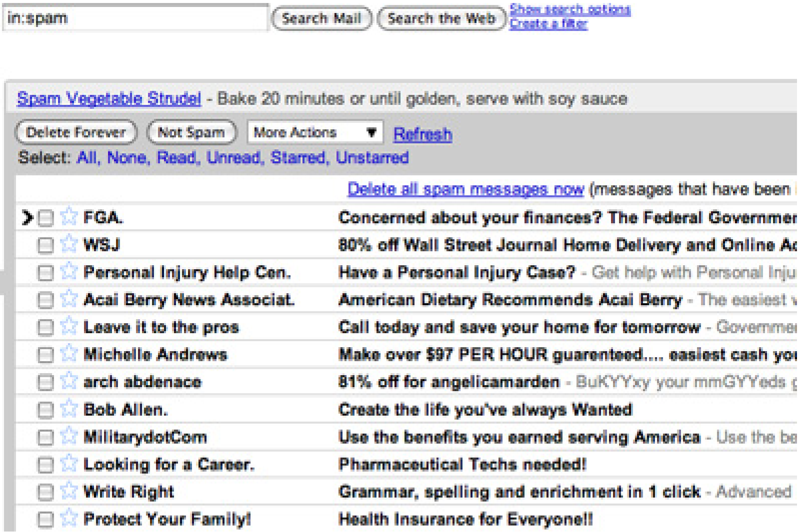
Google Flu is sick.

Flu Season



The Parable of Google Flu: Traps in Big Data Analysis. By Lazer et al.

Spam or Ham


in:spam
Search Mail
Search the Web
Show search options
Create a filter
Spam Vegetable Strudel - Bake 20 minutes or until golden, serve with soy sauce
Delete Forever
Not Spam
More Actions
Refresh
Select: All, None, Read, Unread, Starred, Unstarred
Delete all spam messages now (messages that have been
FGA.
Concerned about your finances? The Federal Governmer
WSJ
80% off Wall Street Journal Home Delivery and Online Ac
Personal Injury Help Cen.
Have a Personal Injury Case? - Get help with Personal Inju
Acai Berry News Associat.
American Dietary Recommends Acai Berry - The easiest
Leave it to the pros
Call today and save your home for tomorrow - Governme
Michelle Andrews
Make over $97 PER HOUR guarenteed.... easiest cash yo
arch abdenace
81% off for angelicamarden - BuKYYxy your mmGYYeds
Bob Allen.
Create the life you've always Wanted
MilitarydotCom
Use the benefits you earned serving America - Use the be
Looking for a Career.
Pharmaceutical Techs needed!
Write Right
Grammar, spelling and enrichment in 1 click - Advanced
Protect Your Family!
Health Insurance for Everyone!!

Spam or Ham

- “According to Commtouch’s Internet Threats Trend Report for the first quarter of 2013, an average of 97.4 billion spam e-mails and 973 million malware e-mails were sent worldwide each day in Q1 2013 (h/t Softpedia).”

Spam or Ham

- "According to Commtouch's Internet Threats Trend Report for the first quarter of 2013, an average of 97.4 billion spam e-mails and 973 million malware e-mails were sent worldwide each day in Q1 2013 (h/t Softpedia)."
- Spam Filter

Spam or Ham

- "According to Commtouch's Internet Threats Trend Report for the first quarter of 2013, an average of 97.4 billion spam e-mails and 973 million malware e-mails were sent worldwide each day in Q1 2013 (h/t Softpedia)."
- Spam Filter
- $logit[p(spam)] = \alpha + f'\beta$ where f is frequencies.

Vote


A 61-million-person experiment in social influence and political mobilization

Robert M. Bond[1], Christopher J. Fariss[1], Jason J. Jones[2], Adam D. I. Kramer[3], Cameron Marlow[3], Jaime E. Settle[1] & James H. Fowler[1,4]

Human behaviour is thought to spread through face-to-face social networks, but it is difficult to identify social influence effects in observational studies[9–13], and it is unknown whether online social with all users of at least 18 years of age in the United States who accessed the Facebook website on 2 November 2010, the day of the US congressional elections. Users were randomly assigned to a 'social

Vote

A 61-million-person experiment in social influence and political mobilization

Robert M. Bond[1], Christopher J. Fariss[1], Jason J. Jones[2], Adam D. I. Kramer[3], Cameron Marlow[3], Jaime E. Settle[1] & James H. Fowler[1,4]

Human behaviour is thought to spread through face-to-face social networks, but it is difficult to identify social influence effects in observational studies[9–13], and it is unknown whether online social with all users of at least 18 years of age in the United States who accessed the Facebook website on 2 November 2010, the day of the US congressional elections. Users were randomly assigned to a 'social

.39% direct effect, and .01 to .1% indirect effect.

Vote for Obama

– Obama 2012 Campaign

Vote for Obama

- Obama 2012 Campaign
- Highly customized messaging: Soccer Moms, NASCAR dads ...

Vote for Obama

– Obama 2012 Campaign
– Highly customized messaging: Soccer Moms, NASCAR dads . . .
– Used SQL database Vertica for ‘speed-of-thought’ analyses.

What do we mean by big data?

What do we mean by big data?

– Big in rows (size n)
  Big in columns (dimensions p)

What do we mean by big data?

- Big in rows (size n)
  Big in columns (dimensions p)
- Hard to extract value from

What do we mean by big data?

- Big in rows (size n)
  Big in columns (dimensions p)
- Hard to extract value from
- ‘Big data’ is high **volume**, high **velocity** and high **variety** information assets that demand cost-effective, innovative forms of information processing[.]

  Gartner, Inc.’s “3Vs” definition.

Sources of (Big) Data

– Data as the by-product of other activities

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase
  - Moving your body (Fitbit)

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase
  - Moving your body (Fitbit)
  - Moving yourself without moving your body (Snapshot)

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase
  - Moving your body (Fitbit)
  - Moving yourself without moving your body (Snapshot)
  - Data were always being generated. They just weren't being captured.

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase
  - Moving your body (Fitbit)
  - Moving yourself without moving your body (Snapshot)
  - Data were always being generated. They just weren't being captured.
    - Cheaper, smaller sensors help.

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase
  - Moving your body (Fitbit)
  - Moving yourself without moving your body (Snapshot)
  - Data were always being generated. They just weren't being captured.
    - Cheaper, smaller sensors help.
    - So does cheaper storage. (1950 $\sim$ \$10,000/MB. 2015 $\lll$ \$.0001/MB)

Sources of (Big) Data

- Data as the by-product of other activities
  - Click trail, clicks before a purchase
  - Moving your body (Fitbit)
  - Moving yourself without moving your body (Snapshot)
  - Data were always being generated. They just weren't being captured.
    - Cheaper, smaller sensors help.
    - So does cheaper storage. (1950 $\sim$ \$10,000/MB. 2015 $\lll$ \$.0001/MB)
- Data as the primary goal of activities

  Telescopes, Genetic sequencers, 61 million person experiments . . .

How big are the data?

How big are the data?

- Web

  20 billion webpages, each 20 KB = 400 TB

  Say all stored on a single disk.

How big are the data?

– Web

20 billion webpages, each 20 KB = 400 TB

Say all stored on a single disk.

Read speed $\sim$ 50MB/sec.

How big are the data?

- Web

  20 billion webpages, each 20 KB = 400 TB

  Say all stored on a single disk.

  Read speed $\sim$ 50MB/sec.

  92 days to read from disk to memory.

How big are the data?

– Web

20 billion webpages, each 20 KB = 400 TB

Say all stored on a single disk.

Read speed $\sim$ 50MB/sec.

92 days to read from disk to memory.

How big are the data?

- Web
  20 billion webpages, each 20 KB = 400 TB
  Say all stored on a single disk.
  Read speed $\sim$ 50MB/sec.
  92 days to read from disk to memory.

- Astronomy
  Apache Point Telescope $\sim$ 200 GB/night.
  Large Synoptic Survey Telescope: 3 billion pixel camera
  $\sim$ 30TB/night. In 10 years $\sim$ 60 PB

How big are the data?

- Web

  20 billion webpages, each 20 KB = 400 TB

  Say all stored on a single disk.

  Read speed $\sim$ 50MB/sec.

  92 days to read from disk to memory.

- Astronomy

  Apache Point Telescope $\sim$ 200 GB/night.

  Large Synoptic Survey Telescope: 3 billion pixel camera

  $\sim$ 30TB/night. In 10 years $\sim$ 60 PB

- Life Sciences

  High Throughput sequencer $\sim$ 1 TB/day

How big are the data?

- Web
  20 billion webpages, each 20 KB = 400 TB
  Say all stored on a single disk.
  Read speed $\sim$ 50MB/sec.
  92 days to read from disk to memory.

- Astronomy
  Apache Point Telescope $\sim$ 200 GB/night.
  Large Synoptic Survey Telescope: 3 billion pixel camera
  $\sim$ 30TB/night. In 10 years $\sim$ 60 PB

- Life Sciences
  High Throughput sequencer $\sim$ 1 TB/day

- CIA
  REDACTED

Implications for Statistics, Computation.

Implications for Statistics

Implications for Statistics

– Little data, Big data

Implications for Statistics

– Little data, Big data

  Sampling still matters

Implications for Statistics

- Little data, Big data
  Sampling still matters
- Everything is significant (The Starry Night)

Implications for Statistics

– Little data, Big data
  Sampling still matters

– Everything is significant (The Starry Night)

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}} \tag{1}$$

  Benjamini and Hochberg (1995) (FDR), cost of false discovery, Familywise error rate (Bonferroni)

Implications for Statistics

- Little data, Big data
  Sampling still matters

- Everything is significant (The Starry Night)

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}} \tag{1}$$

  Benjamini and Hochberg (1995) (FDR), cost of false discovery, Familywise error rate (Bonferroni)

- Inverting a matrix

Implications for Statistics

– Little data, Big data
  Sampling still matters

– Everything is significant (The Starry Night)

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}} \qquad (1)$$

  Benjamini and Hochberg (1995) (FDR), cost of false discovery, Familywise error rate (Bonferroni)

– Inverting a matrix
  (Stochastic) Gradient Descent (Ascent), BFGS, ...

Implications for Statistics

- Little data, Big data
  Sampling still matters

- Everything is significant (The Starry Night)

$$\text{False discovery Proportion} = \frac{\text{\# of FP}}{\text{\# of Sig. Results}} \tag{1}$$

  Benjamini and Hochberg (1995) (FDR), cost of false discovery, Familywise error rate (Bonferroni)

- Inverting a matrix
  (Stochastic) Gradient Descent (Ascent), BFGS, ...

- Causal inference

Implications for Statistics

- Little data, Big data
  Sampling still matters

- Everything is significant (The Starry Night)

$$\text{False discovery Proportion} = \frac{\text{\# of FP}}{\text{\# of Sig. Results}} \tag{1}$$

  Benjamini and Hochberg (1995) (FDR), cost of false discovery, Familywise error rate (Bonferroni)

- Inverting a matrix
  (Stochastic) Gradient Descent (Ascent), BFGS, ...

- Causal inference
  Large p may help

Implications for Statistics

- Little data, Big data

  Sampling still matters

- Everything is significant (The Starry Night)

$$\text{False discovery Proportion} = \frac{\#\text{ of FP}}{\#\text{ of Sig. Results}} \tag{1}$$

  Benjamini and Hochberg (1995) (FDR), cost of false discovery, Familywise error rate (Bonferroni)

- Inverting a matrix

  (Stochastic) Gradient Descent (Ascent), BFGS, ...

- Causal inference

  Large p may help

  Passive observation as things change arbitrarily may help

Implications for Computation

- Conventional Understanding of what is computationally tractable: Polynomial time algorithm (N^k)

Implications for Computation

- Conventional Understanding of what is computationally tractable: Polynomial time algorithm (N^k)
- Now it is $(N^k)/m$, where m is the number of computers.

Implications for Computation

- Conventional Understanding of what is computationally tractable: Polynomial time algorithm (N^k)
- Now it is $(N^k)/m$, where m is the number of computers.
- For really big data: N*log(N)

  Traversing a binary tree, sort and search N log(N)

  Streaming application

MapReduce and PageRank

20 billion webpages, each 20 KB = 400 TB
Say all data stored on a single disk. Read speed $\sim$ 50MB/sec.
92 days to read from disk to memory.

Solution, Problem

– Parallelize, if 1000 computers, then just ~ 1 hour

Solution, Problem

- Parallelize, if 1000 computers, then just ~ 1 hour
- But ...

Solution, Problem

- Parallelize, if 1000 computers, then just ~ 1 hour
- But ...
  - Nodes can fail.
    Say single node fails once every 1000 days.
    But 1000 failures per day if 1 million servers.

Solution, Problem

- Parallelize, if 1000 computers, then just ~ 1 hour
- But . . .
  - Nodes can fail.
    Say single node fails once every 1000 days.
    But 1000 failures per day if 1 million servers.
  - Network bandwidth $\sim$ 1GBps.
    If you have 10 TB of day – takes 1 day.

Solution, Problem

- Parallelize, if 1000 computers, then just ~ 1 hour
- But ...
  - Nodes can fail.
    Say single node fails once every 1000 days.
    But 1000 failures per day if 1 million servers.
  - Network bandwidth $\sim$ 1GBps.
    If you have 10 TB of day – takes 1 day.
  - Distributed programming can be very very hard.

Solution: MapReduce

– Store data redundantly

Solution: MapReduce

- Store data redundantly
  - Distributed File Systems, e.g. GFS, HDFS

Solution: MapReduce

- Store data redundantly
  - Distributed File Systems, e.g. GFS, HDFS
  - Typical usage pattern: Data rarely updated. Read often. Updated through appends.

Solution: MapReduce

- Store data redundantly
  - Distributed File Systems, e.g. GFS, HDFS
  - Typical usage pattern: Data rarely updated. Read often. Updated through appends.
  - Implementation:
    Data kept in chunks, machines called 'chunk servers'
    Chunks replicated. Typically 3x. One in a completely separate rack.
    Master node (GFS)/Name Node (HDFS) tracks metadata

Solution: MapReduce

- Store data redundantly
  - Distributed File Systems, e.g. GFS, HDFS
  - Typical usage pattern: Data rarely updated. Read often. Updated through appends.
  - Implementation:
    Data kept in chunks, machines called ‘chunk servers’
    Chunks replicated. Typically 3x. One in a completely separate rack.
    Master node (GFS)/Name Node (HDFS) tracks metadata
- Minimize data movement

Solution: MapReduce

- Store data redundantly
  - Distributed File Systems, e.g. GFS, HDFS
  - Typical usage pattern: Data rarely updated. Read often. Updated through appends.
  - Implementation:
    Data kept in chunks, machines called 'chunk servers'
    Chunks replicated. Typically 3x. One in a completely separate rack.
    Master node (GFS)/Name Node (HDFS) tracks metadata
- Minimize data movement
- Simple programming model

More About MapReduce

- Name comes from 2004 paper MapReduce: Simplified Data Processing on Large Clusters by Dean and Ghemawat.

More About MapReduce

– Name comes from 2004 paper MapReduce: Simplified Data Processing on Large Clusters by Dean and Ghemawat.

– Implementation - Hadoop (via Yahoo)/Apache

MapReduce Example

- Count each distinct word in a huge document, e.g. urls

MapReduce Example

- Count each distinct word in a huge document, e.g. urls
  - Map function produces key, value pairs
    Word frequency of every word in each document

MapReduce Example

- Count each distinct word in a huge document, e.g. urls
  - Map function produces key, value pairs
    Word frequency of every word in each document
  - Send different words to different computers

MapReduce Example

- Count each distinct word in a huge document, e.g. urls
  - Map function produces key, value pairs
    Word frequency of every word in each document
  - Send different words to different computers
  - Combine

```
map(String input_key, String input_value):
    // input_key: document name
    // input_value: document contents
    for each word w in input_value:
      EmitIntermediate(w, ``1'');


  reduce(String output_key, Iterator intermediate_values):
    // output_key: a word
    // output_values: a list of counts
    int result = 0;
    for each v in intermediate_values:
      result += ParseInt(v);
    Emit(AsString(result));
```

PageRank

– The PageRank Citation Ranking: Bringing Order to the Web By Page et al.

PageRank

- The PageRank Citation Ranking: Bringing Order to the Web By Page et al.
- Among the top 10 data mining algorithms

PageRank

- The PageRank Citation Ranking: Bringing Order to the Web By Page et al.
- Among the top 10 data mining algorithms
- Search

PageRank

- The PageRank Citation Ranking: Bringing Order to the Web By Page et al.
- Among the top 10 data mining algorithms
- Search
  - Searching for similarity

PageRank

- The PageRank Citation Ranking: Bringing Order to the Web By Page et al.
- Among the top 10 data mining algorithms
- Search
  - Searching for similarity
  - Works well when you look for a document in your computer. Any small trusted corpora.

PageRank

- The PageRank Citation Ranking: Bringing Order to the Web By Page et al.
- Among the top 10 data mining algorithms
- Search
  - Searching for similarity
  - Works well when you look for a document in your computer. Any small trusted corpora.
  - Web - lots of matches.

PageRank

- The PageRank Citation Ranking: Bringing Order to the Web By Page et al.
- Among the top 10 data mining algorithms
- Search
  - Searching for similarity
  - Works well when you look for a document in your computer. Any small trusted corpora.
  - Web - lots of matches.
  - Web - lots of false positives. Some of it malware peddling sites.

PageRank

– How to order conditional on similarity?

PageRank

- How to order conditional on similarity?
  - One can rank by popularity if you have complete web traffic information.

PageRank

- How to order conditional on similarity?
  - One can rank by popularity if you have complete web traffic information.
  - But those data are hard to get.

PageRank

- How to order conditional on similarity?
  - One can rank by popularity if you have complete web traffic information.
  - But those data are hard to get.
  - Or you can do ad hoc automation and human curation.

PageRank

- How to order conditional on similarity?
  - One can rank by popularity if you have complete web traffic information.
  - But those data are hard to get.
  - Or you can do ad hoc automation and human curation.
  - But costly to implement. And don't scale.

PageRank

- How to order conditional on similarity?
  - One can rank by popularity if you have complete web traffic information.
  - But those data are hard to get.
  - Or you can do ad hoc automation and human curation.
  - But costly to implement. And don't scale.

- Innovation: see Internet as a graph

  Nodes = webpages, Edges = hyperlinks

  Ranks based on linking patterns alone.

PageRank

- How to order conditional on similarity?
  - One can rank by popularity if you have complete web traffic information.
  - But those data are hard to get.
  - Or you can do ad hoc automation and human curation.
  - But costly to implement. And don't scale.
- Innovation: see Internet as a graph
  Nodes = webpages, Edges = hyperlinks
  Ranks based on linking patterns alone.
- Currency is in-links.

PageRank

– Each in-link is a vote.

PageRank

- Each in-link is a vote.
- But each vote is not equal.

PageRank

- Each in-link is a vote.
- But each vote is not equal.
- In-links from more important pages count for more

PageRank

- Each in-link is a vote.
- But each vote is not equal.
- In-links from more important pages count for more
- Value of each vote:

PageRank

- Each in-link is a vote.
- But each vote is not equal.
- In-links from more important pages count for more
- Value of each vote:
  - Proportional to importance of source page

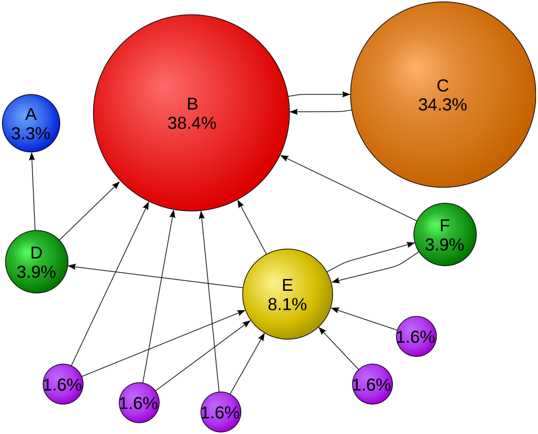
PageRank

- Each in-link is a vote.
- But each vote is not equal.
- In-links from more important pages count for more
- Value of each vote:
  - Proportional to importance of source page
  - Inversely proportional to number of outgoing links on the source page

PageRank

- Each in-link is a vote.
- But each vote is not equal.
- In-links from more important pages count for more
- Value of each vote:
  - Proportional to importance of source page
  - Inversely proportional to number of outgoing links on the source page
  - Say page i has importance r_i, and n outgoing links, each vote $= r_i/n$

Page Rank Example



Source: Wikipedia

PageRank

Say page i gets links from pages j ($n = 5$, r_j) and k ($n = 2$, r_k)

PageRank

Say page i gets links from pages j ($n = 5$, r_j) and k ($n = 2$, r_k)

$$r_i = \frac{r_j}{5} + \frac{r_k}{2}$$

PageRank

Say page i gets links from pages j ($n = 5$, r_j) and k ($n = 2$, r_k)

$$r_i = \frac{r_j}{5} + \frac{r_k}{2}$$

Page j will have its own outlinks, and each will have a value r_j

PageRank

Say page i gets links from pages j ($n = 5$, r_j) and k ($n = 2$, r_k)

$$r_i = \frac{r_j}{5} + \frac{r_k}{2}$$

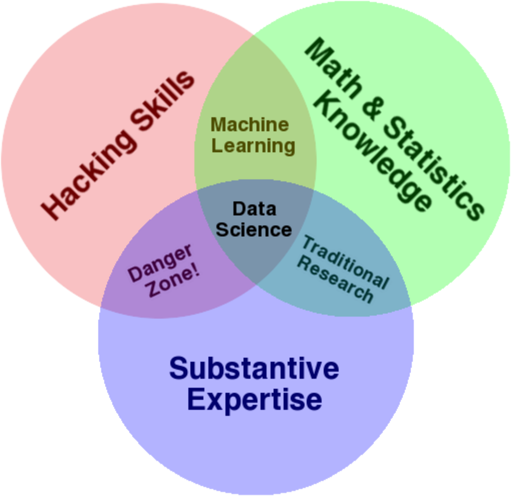
Page j will have its own outlinks, and each will have a value r_j

$$r_i = \sum \frac{r_j}{d_j}$$

over j where j tracks all the pages pointing to i.

Class

Data Science



Data Science Venn Diagram. By Drew Conoway

Course Outline

- Get your own (big) data

  Scrape it, clean it

  Basics of webscraping

  Basics of regular expressions

Course Outline

- Get your own (big) data
  Scrape it, clean it
  Basics of webscraping
  Basics of regular expressions

- Manage your (big) data
  Store it, organize it, use it
  Relational Databases, SQL (SQLite)
  Other databases

Course Outline

- Get your own (big) data
  Scrape it, clean it
  Basics of webscraping
  Basics of regular expressions

- Manage your (big) data
  Store it, organize it, use it
  Relational Databases, SQL (SQLite)
  Other databases

- Analyze your (big) data
  Cross-validation
  (Not) Maximally Likely
  Numerical Optimization

Prerequisites

– Basic but important

Prerequisites

- Basic but important
- Statistics:
  - Probability theory, some combinatorics
  - Linear Regression and other standard estimation techniques

Prerequisites

- Basic but important
- Statistics:
  - Probability theory, some combinatorics
  - Linear Regression and other standard estimation techniques
- Computation:
  - Have written a loop
  - Have written a function

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.
  - `R Markdown` For Documenting.

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.
  - `R Markdown` For Documenting.

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.
  - `R Markdown` For Documenting.
- Python

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.
  - `R Markdown` For Documenting.
- Python
  - Academic Version `Enthought Python Distribution`

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.
  - `R Markdown` For Documenting.
- Python
  - Academic Version `Enthought Python Distribution`
  - Eclipse, PyDev, Aptana Studio etc.

Software and Programming

- Open Source
  License fees add up if you are running software on 1000's of machines
- R
  - `RStudio` IDE for R, Makes it easier to code.
  - `R Markdown` For Documenting.
- Python
  - Academic Version `Enthought Python Distribution`
  - Eclipse, PyDev, Aptana Studio etc.
- SQLite