Types of Data Science Questions

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Types of Data Science Questions

In approximate order of difficulty

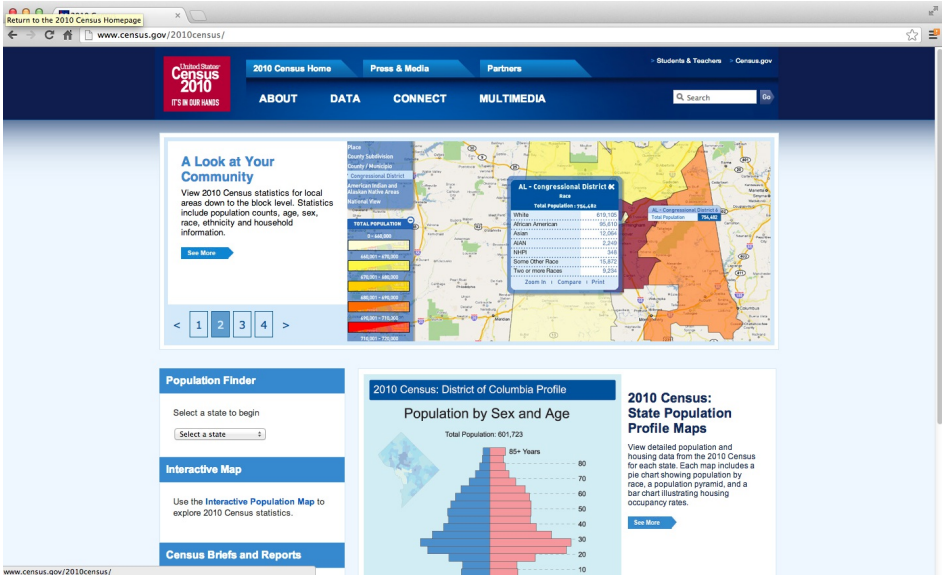
- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

About descriptive analyses

Goal: Describe a set of data

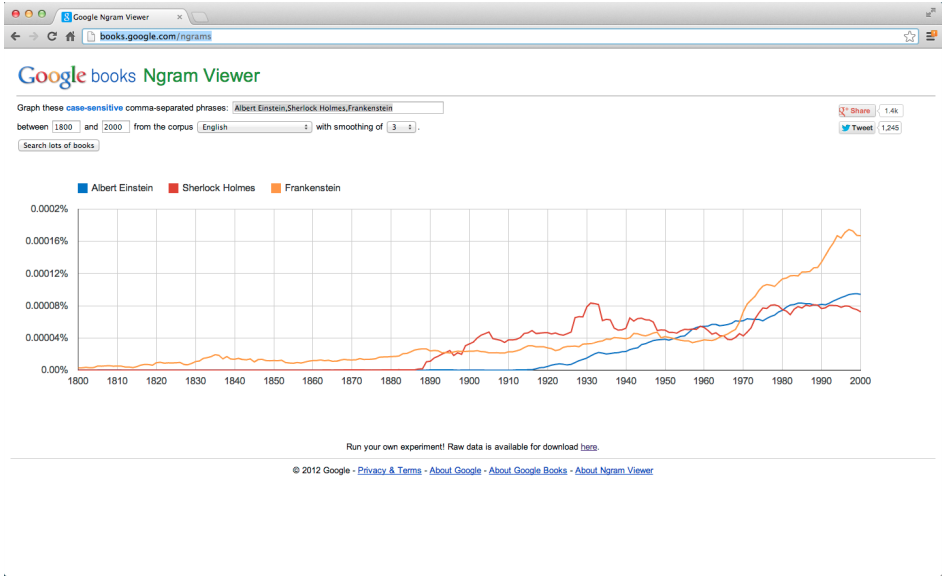
- The first kind of data analysis performed
- Commonly applied to census data
- The description and interpretation are different steps
- Descriptions can usually not be generalized without additional statistical modeling

Descriptive analysis



http://www.census.gov/2010census/

Descriptive analysis



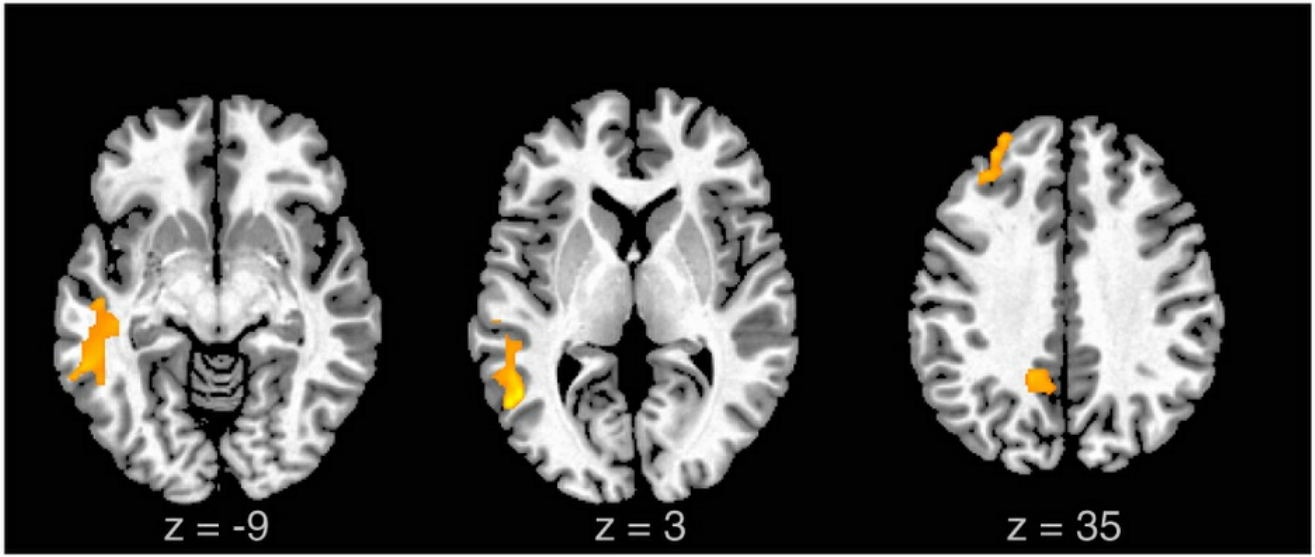
http://books.google.com/ngrams

About exploratory analysis

Goal: Find relationships you didn't know about

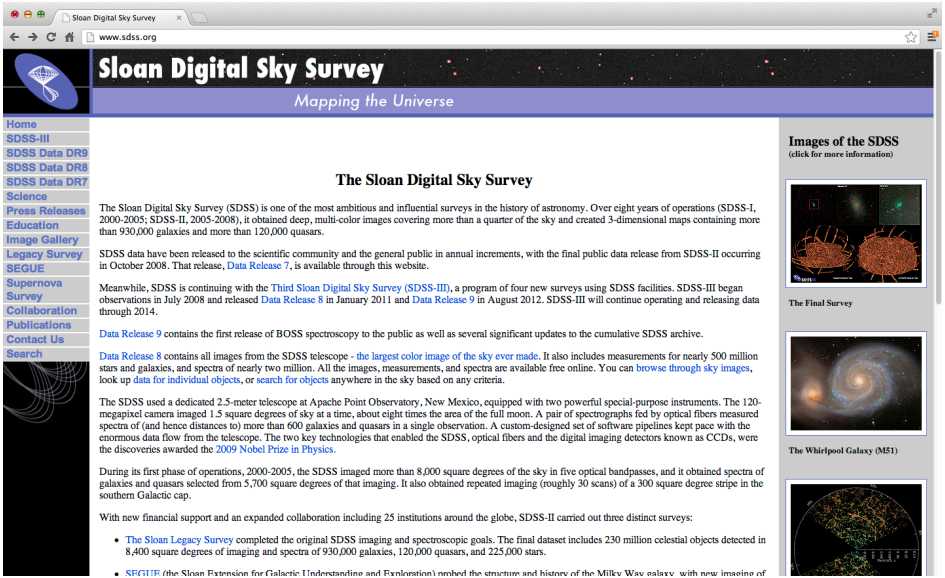
- Exploratory models are good for discovering new connections
- They are also useful for defining future studies
- Exploratory analyses are usually not the final say
- Exploratory analyses alone should not be used for generalizing/predicting
- Correlation does not imply causation

Exploratory analysis



Liu et al. (2012) Scientific Reports

Exploratory analysis



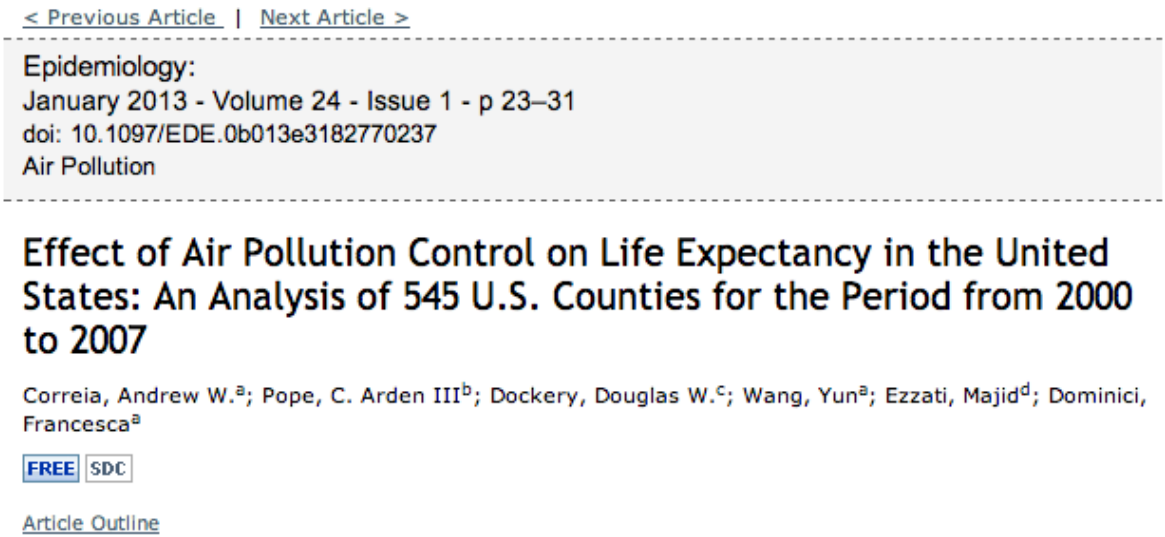
http://www.sdss.org/

About inferential analysis

Goal: Use a relatively small sample of data to say something about a bigger population

- Inference is commonly the goal of statistical models
- Inference involves estimating both the quantity you care about and your uncertainty about your estimate
- Inference depends heavily on both the population and the sampling scheme

Inferential analysis



< Previous Article | Next Article >

Epidemiology:
January 2013 - Volume 24 - Issue 1 - p 23–31
doi: 10.1097/EDE.0b013e3182770237
Air Pollution

Effect of Air Pollution Control on Life Expectancy in the United States: An Analysis of 545 U.S. Counties for the Period from 2000 to 2007

Correia, Andrew W.[a]; Pope, C. Arden III[b]; Dockery, Douglas W.[c]; Wang, Yun[a]; Ezzati, Majid[d]; Dominici, Francesca[a]

FREE SDC

Article Outline

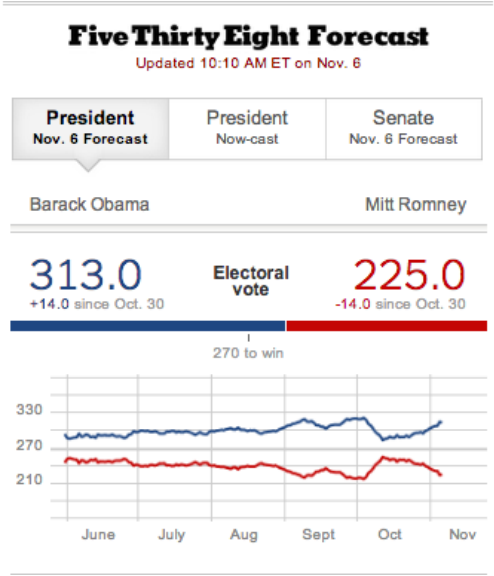
Correia et al. (2013) Epidemiology

About predictive analysis

Goal: To use the data on some objects to predict values for another object

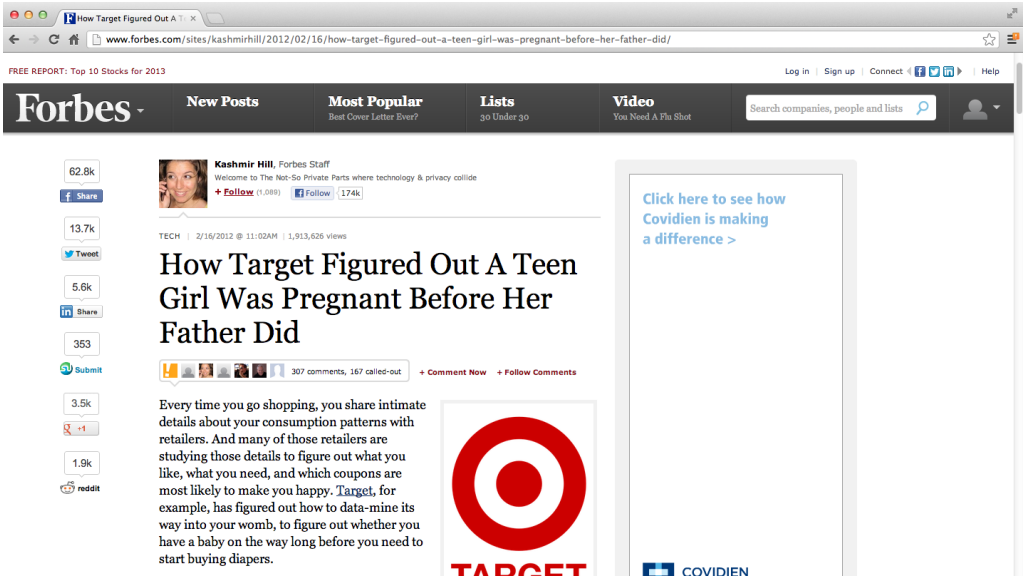
- If X predicts Y it does not mean that X causes Y
- Accurate prediction depends heavily on measuring the right variables
- Although there are better and worse prediction models, more data and a simple model works really well
- Prediction is very hard, especially about the future references

Predictive analysis



http://fivethirtyeight.blogs.nytimes.com/

Predictive analysis



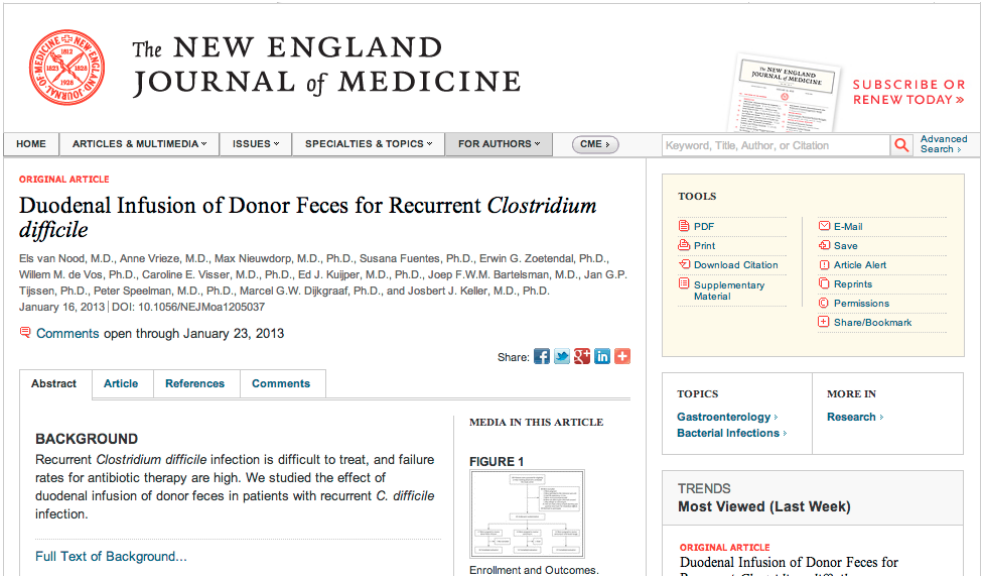
http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

About causal analysis

Goal: To find out what happens to one variable when you make another variable change.

- Usually randomized studies are required to identify causation
- There are approaches to inferring causation in non-randomized studies, but they are complicated and sensitive to assumptions
- Causal relationships are usually identified as average effects, but may not apply to every individual
- Causal models are usually the "gold standard" for data analysis

Causal analysis



The NEW ENGLAND JOURNAL of MEDICINE

SUBSCRIBE OR RENEW TODAY »

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS | CME | Keyword, Title, Author, or Citation | Advanced Search

ORIGINAL ARTICLE

Duodenal Infusion of Donor Feces for Recurrent *Clostridium difficile*

Els van Nood, M.D., Anne Vrieze, M.D., Max Nieuwdorp, M.D., Ph.D., Susana Fuentes, Ph.D., Erwin G. Zoetendal, Ph.D., Willem M. de Vos, Ph.D., Caroline E. Visser, M.D., Ph.D., Ed J. Kuijper, M.D., Ph.D., Joep F.W.M. Bartelsman, M.D., Jan G.P. Tijssen, Ph.D., Peter Speelman, M.D., Ph.D., Marcel G.W. Dijkgraaf, Ph.D., and Josbert J. Keller, M.D., Ph.D.
January 16, 2013 | DOI: 10.1056/NEJMoa1205037

Comments open through January 23, 2013

Share:

Abstract | Article | References | Comments

BACKGROUND

Recurrent *Clostridium difficile* infection is difficult to treat, and failure rates for antibiotic therapy are high. We studied the effect of duodenal infusion of donor feces in patients with recurrent *C. difficile* infection.

Full Text of Background...

MEDIA IN THIS ARTICLE

FIGURE 1

Enrollment and Outcomes.

TOOLS

PDF | E-Mail | Print | Save | Download Citation | Article Alert | Supplementary Material | Reprints | Permissions | Share/Bookmark

TOPICS

Gastroenterology | Bacterial Infections

MORE IN

Research

TRENDS

Most Viewed (Last Week)

ORIGINAL ARTICLE

Duodenal Infusion of Donor Feces for

van Nood et al. (2013) NEJM

About mechanistic analysis

Goal: Understand the exact changes in variables that lead to changes in other variables for individual objects.

- Incredibly hard to infer, except in simple situations
- Usually modeled by a deterministic set of equations (physical/engineering science)
- Generally the random component of the data is measurement error
- If the equations are known but the parameters are not, they may be inferred with data analysis

Mechanistic analysis

U.S. Department of Transportation
Federal Highway Administration

Federal Highway Administration
RESOURCE CENTER

PAVEMENT AND MATERIALS

Mechanistic – Empirical Pavement Design

Problem: Empirical Design Process Restrict Performance Prediction

Accurately predicting performance and durability is critical to improving the design of new and existing pavements. Poor performance increases traffic congestion, compromises public safety, and raises maintenance costs due to frequent repairs. Each year, transportation agencies spend more than $20 billion in Federal funds to improve the Nation's pavements. Existing design procedures are based upon the 1950's AASHO Road Test and use empirical relationships. Presently, pavement designs often exceed the data limits and conditions used in the AASHTO Road Test have been exceeded. Pavement with expected traffic as much as 30 times greater are

Deployment Process:

The Federal Highway Administration (FHWA) organized the Design Guide Implementation Team (DGIT) to inform the FHWA division offices, State highway agencies, industry members, and other organizations and experts about the upcoming guide and to help potential users prepare for it. To introduce the guide and to discuss implementation issues, the DGIT has developed a one-day workshop. Seven of these workshops will be held across the Nation, starting on May 25, 2004, in Biloxi, MS. Other workshops will be held in Vancouver, WA (June); Indianapolis, IN (July); Hawaii (July); Mystic, CT (August); Kansas City, KS (September); and Phoenix, AZ (October).

http://www.fhwa.dot.gov/resourcecenter/teams/pavement/pave_3pdg.pdf

What is data?

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Definition of data

> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

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

Definition of data

> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

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

Set of items: Sometimes called the population; the set of objects you are interested in

Definition of data

> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

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

Variables: A measurement or characteristic of an item.

Definition of data

> “Data are values of qualitative or quantitative variables, belonging to a set of items.”

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

Qualitative: Country of origin, sex, treatment

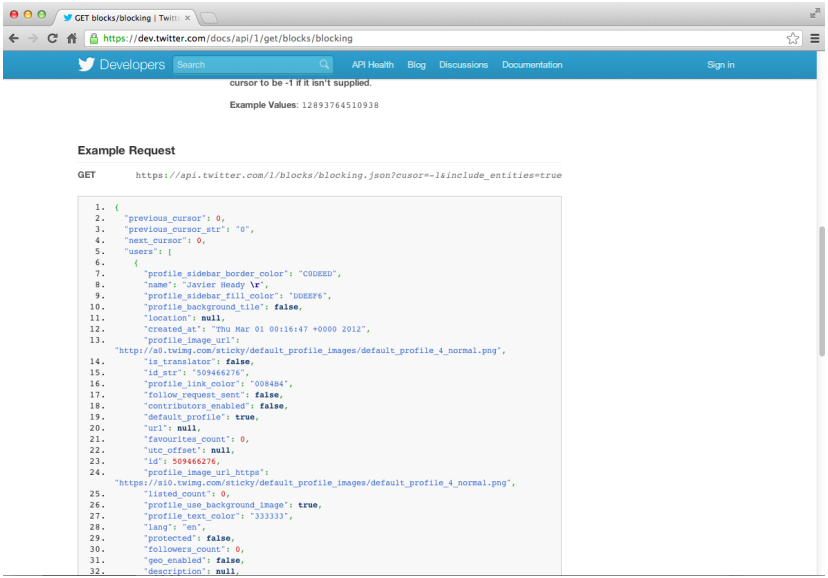
Quantitative: Height, weight, blood pressure

What do data look like?

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```^\__`_```^a``a`^a_^__]a_]\]`a_______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaababbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```

http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt
```

What do data look like?



https://dev.twitter.com/docs/api/1/get/blocks/blocking

What do data look like?

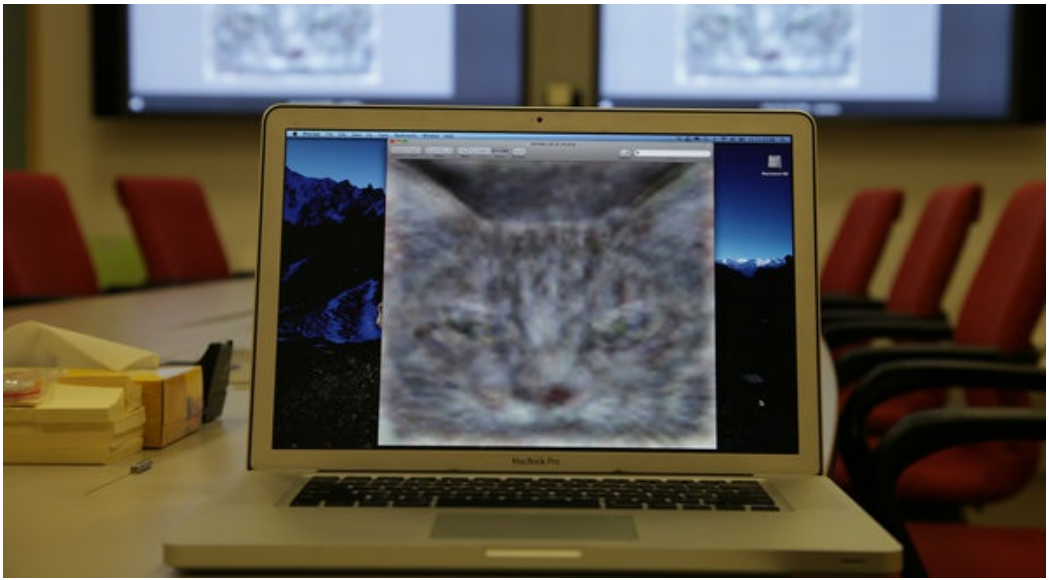
```
-------------------------   ALLERGIES   -------------------------      -------------------------   MEDICATION HISTORY   ----------

ast Updated: 01 Dec 2011 @ 0851                                       Last Updated: 11 Apr 2011 @ 1737

                                                                      Medication: AMLODIPINE BESYLATE 10MG TAB
llergy Name:        TRIMETHOPRIM                                      Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR
ocation:            DAYT29                                            GRAPEFRUIT JUICE--
ate Entered:        09 Mar 2011                                       Status: Active
eaction:                                                              Refills Remaining: 3
llergy Type:        DRUG                                              Last Filled On: 20 Aug 2010
A Drug Class:       ANTI-INFECTIVES,OTHER                             Initially Ordered On: 13 Aug 2010
bserved/Historical: HISTORICAL                                        Quantity: 45
omments:            The reaction to this allergy was MILD (NO SQUELAE) Days Supply: 90
                                                                      Pharmacy: DAYTON
llergy Name:        TRAMADOL                                          Prescription Number: 2718953
ocation:            DAYT29
ate Entered:        09 Mar 2011                                       Medication: IBUPROFEN 600MG TAB
eaction:            URINARY RETENTION                                 Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOO
llergy Type:        DRUG                                              Status: Active
A Drug Class:       NON-OPIOID ANALGESICS                             Refills Remaining: 3
bserved/Historical: HISTORICAL                                        Last Filled On: 20 Aug 2010
omments:            gradually worsening difficulty emptying bladder   Initially Ordered On: 01 Jul 2010
```

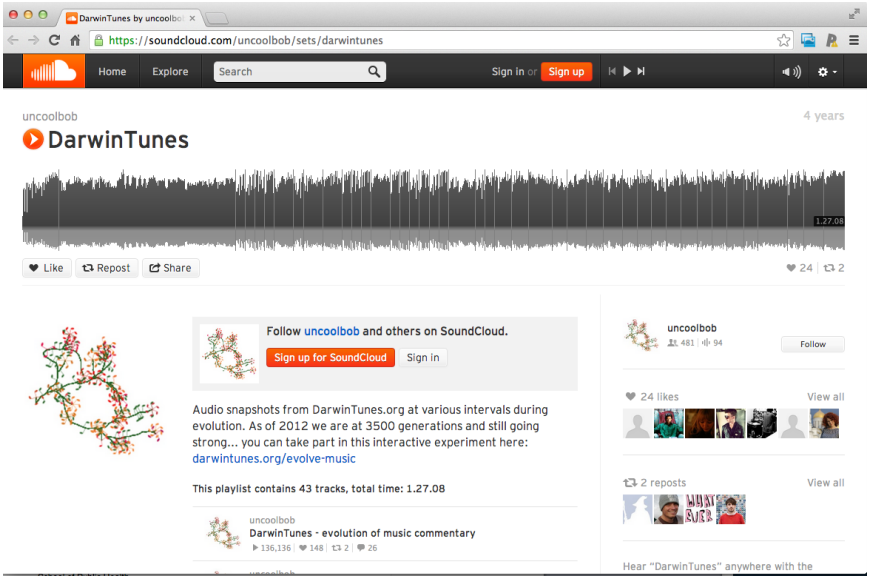
http://blue-button.github.com/challenge/

What do data look like?



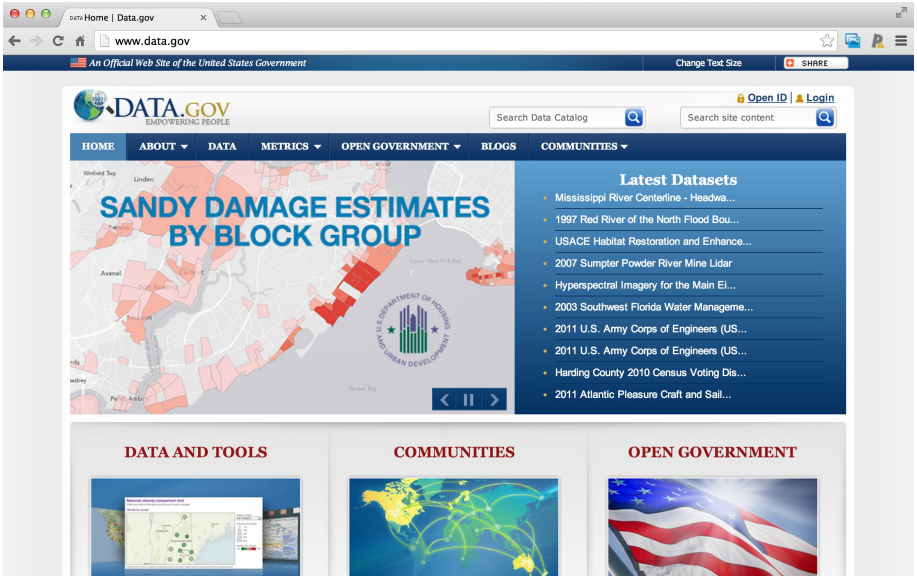
http://www.nytimes.com/2012/06/26/technology/in-a-big-network-of-computers-evidence-of-machine-learning.html?pagewanted=all&_r=0

What do data look like?



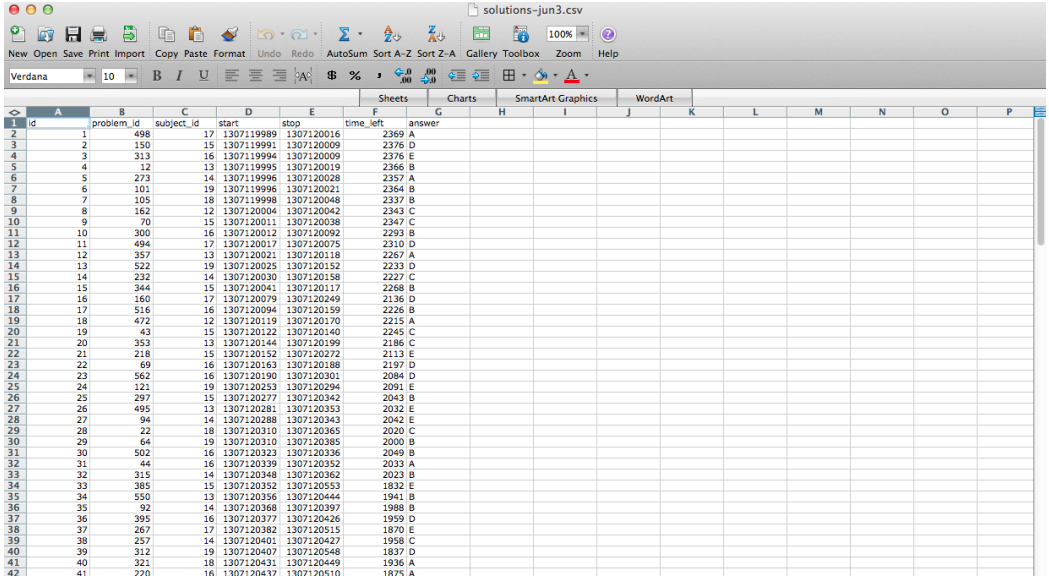
http://www.pnas.org/content/109/30/12081.full https://soundcloud.com/uncoolbob/sets/darwintunes

What do data look like?



http://www.data.gov/

What do data look like? Rarely



	A	B	C	D	E	F	G
1	id	problem_id	subject_id	start	stop	time_left	answer
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	2364	B
8	7	105	18	1307119998	1307120048	2337	B
9	8	162	12	1307120004	1307120042	2343	C
10	9	70	15	1307120011	1307120038	2347	C
11	10	300	16	1307120012	1307120092	2293	B
12	11	494	17	1307120017	1307120075	2310	D
13	12	357	13	1307120021	1307120118	2267	A
14	13	522	19	1307120025	1307120152	2233	D
15	14	232	14	1307120030	1307120158	2227	C
16	15	344	15	1307120041	1307120117	2268	B
17	16	160	17	1307120079	1307120249	2136	D
18	17	516	16	1307120094	1307120159	2226	B
19	18	472	12	1307120119	1307120170	2215	A
20	19	43	15	1307120122	1307120140	2245	C
21	20	353	13	1307120144	1307120199	2186	C
22	21	218	15	1307120152	1307120272	2113	E
23	22	69	16	1307120163	1307120188	2197	D
24	23	562	16	1307120190	1307120301	2084	D
25	24	121	19	1307120253	1307120294	2091	E
26	25	297	15	1307120277	1307120342	2043	B
27	26	495	13	1307120281	1307120353	2032	E
28	27	94	14	1307120288	1307120343	2042	E
29	28	22	18	1307120310	1307120365	2020	C
30	29	64	19	1307120310	1307120385	2000	B
31	30	502	16	1307120323	1307120336	2049	B
32	31	44	16	1307120339	1307120352	2033	A
33	32	315	14	1307120348	1307120362	2023	B
34	33	385	15	1307120352	1307120553	1832	E
35	34	550	13	1307120356	1307120444	1941	B
36	35	92	14	1307120368	1307120397	1988	B
37	36	395	16	1307120377	1307120426	1959	D
38	37	267	17	1307120382	1307120515	1870	E
39	38	257	14	1307120401	1307120427	1958	C
40	39	312	19	1307120407	1307120548	1837	D
41	40	321	18	1307120431	1307120449	1936	A
42	41	220	16	1307120437	1307120510	1875	A

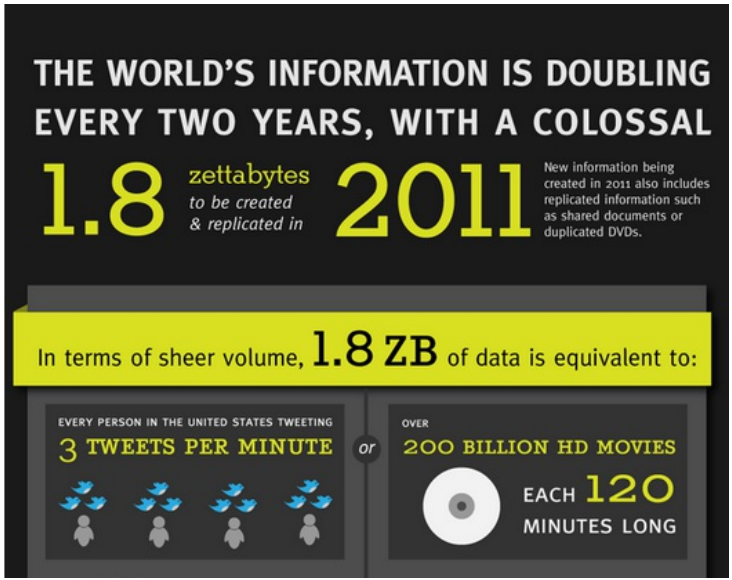
The data is the second most important thing

- The most important thing in data science is the question
- The second most important is the data
- Often the data will limit or enable the questions
- But having data can't save you if you don't have a question

What about big data?

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

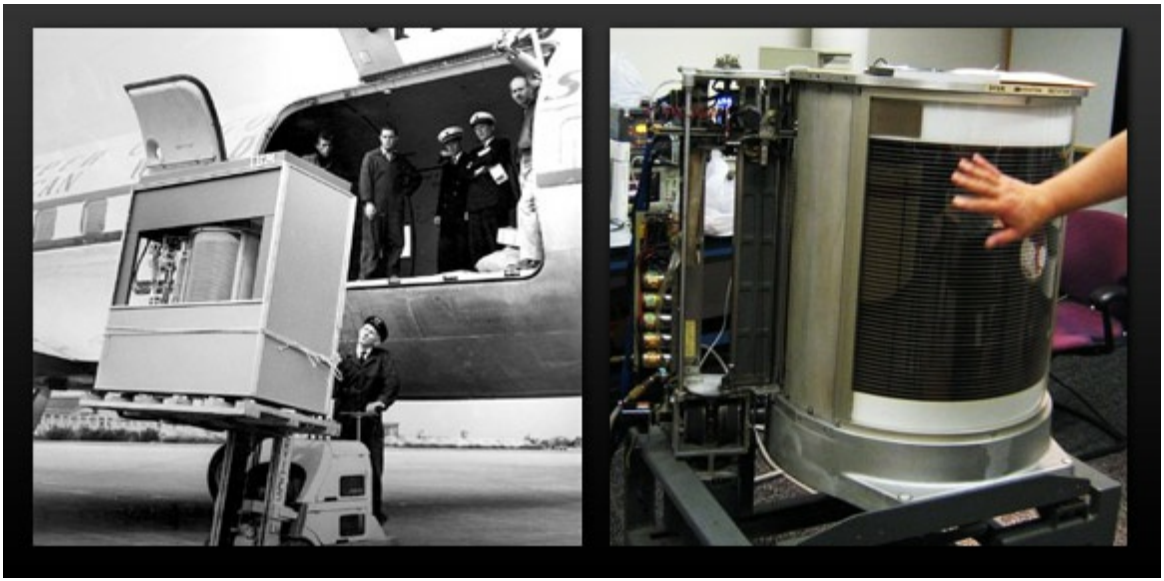
How much is there?



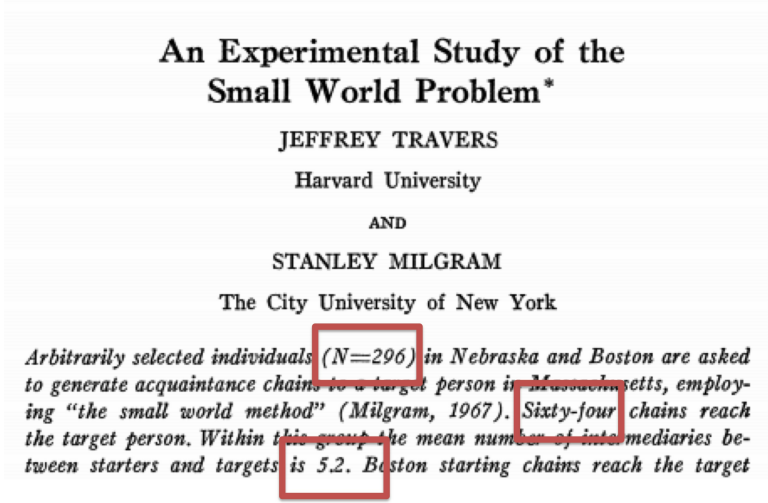
http://mashable.com/2011/06/28/data-infographic/

So what about big data?

Depends on your perspective



Why big data now?



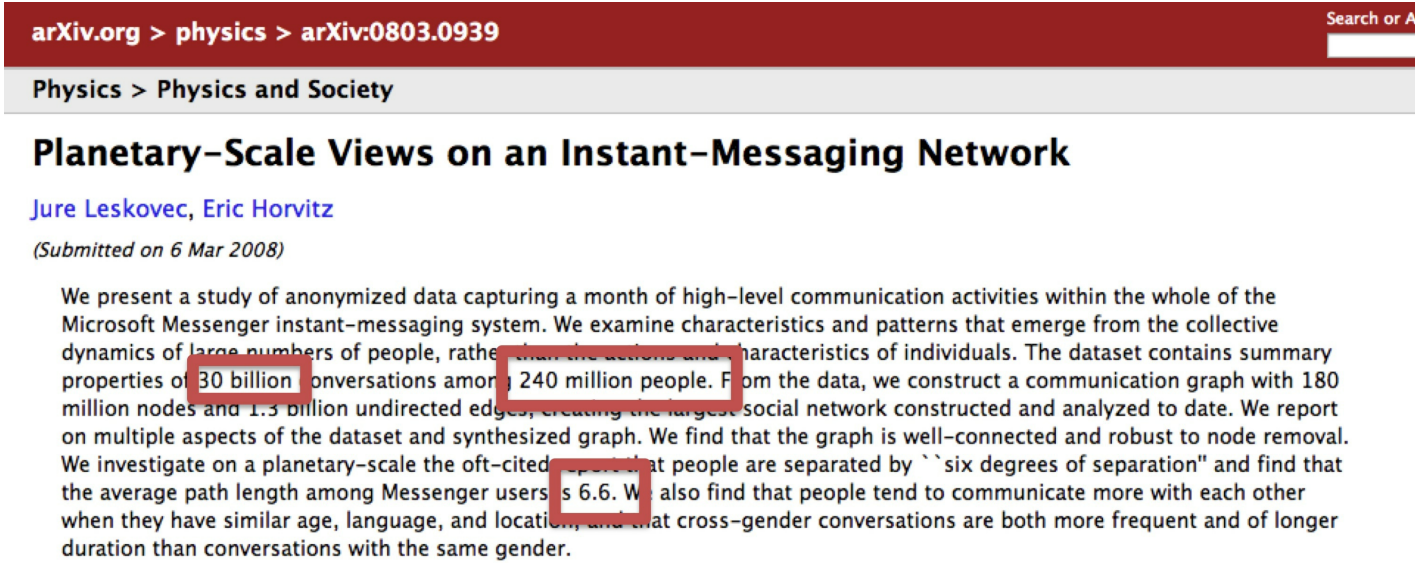
An Experimental Study of the Small World Problem*

JEFFREY TRAVERS
Harvard University
AND
STANLEY MILGRAM
The City University of New York

Arbitrarily selected individuals (N=296) in Nebraska and Boston are asked to generate acquaintance chains to a target person in Massachusetts, employing "the small world method" (Milgram, 1967). Sixty-four chains reach the target person. Within this group the mean number of intermediaries between starters and targets is 5.2. Boston starting chains reach the target

Travers and Milgram (1969) Sociometry

Why big data now?



arXiv.org > physics > arXiv:0803.0939

Search or A

Physics > Physics and Society

Planetary-Scale Views on an Instant-Messaging Network

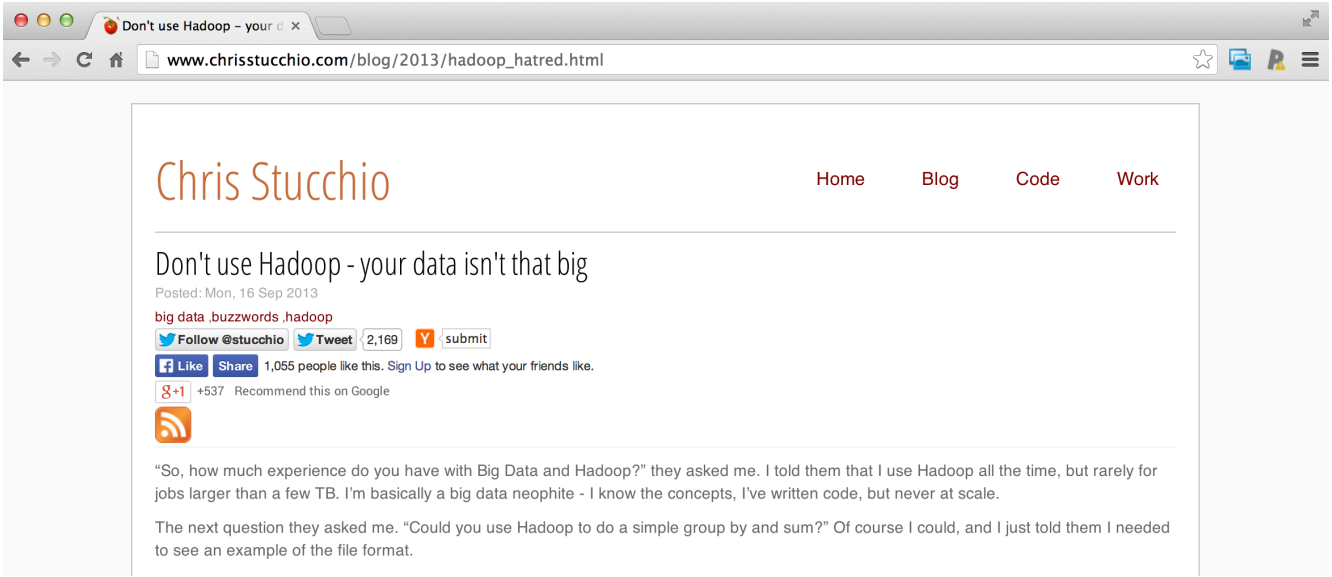
Jure Leskovec, Eric Horvitz

(Submitted on 6 Mar 2008)

We present a study of anonymized data capturing a month of high-level communication activities within the whole of the Microsoft Messenger instant-messaging system. We examine characteristics and patterns that emerge from the collective dynamics of large numbers of people, rathe haracteristics of individuals. The dataset contains summary properties of 30 billion onversations amon 240 million people. F om the data, we construct a communication graph with 180 million nodes and 1.3 billion undirected edg social network constructed and analyzed to date. We report on multiple aspects of the dataset and synthesized graph. We find that the graph is well-connected and robust to node removal. We investigate on a planetary-scale the oft-cited t people are separated by ``six degrees of separation'' and find that the average path length among Messenger users s 6.6. W also find that people tend to communicate more with each other when they have similar age, language, and locati t cross-gender conversations are both more frequent and of longer duration than conversations with the same gender.

Leskovec and Horvitz WWW '08

Big or small - you need the right data



http://www.chrisstucchio.com/blog/2013/hadoop_hatred.html

Big or small - you need the right data

> “The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data...”

Tukey

Big or small - you need the right data

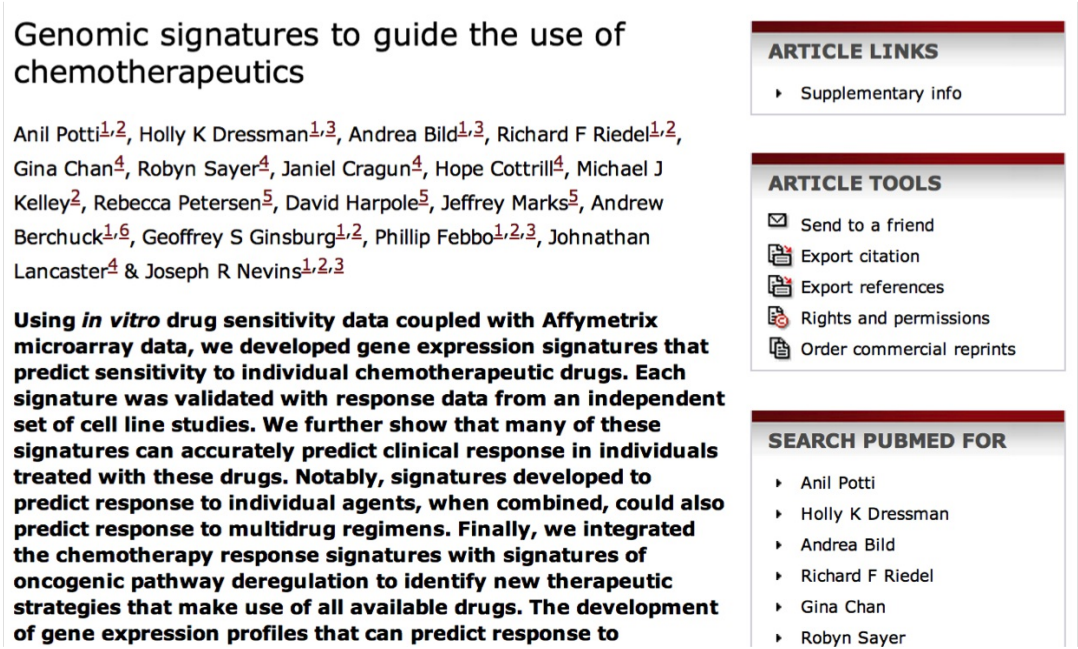
> “...no matter how big the data are.”

Leek

Experimental design

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Why you should care - an exciting result!



Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1,2,3], Johnathan Lancaster[4] & Joseph R Nevins[1,2,3]

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to

ARTICLE LINKS

- Supplementary info

ARTICLE TOOLS

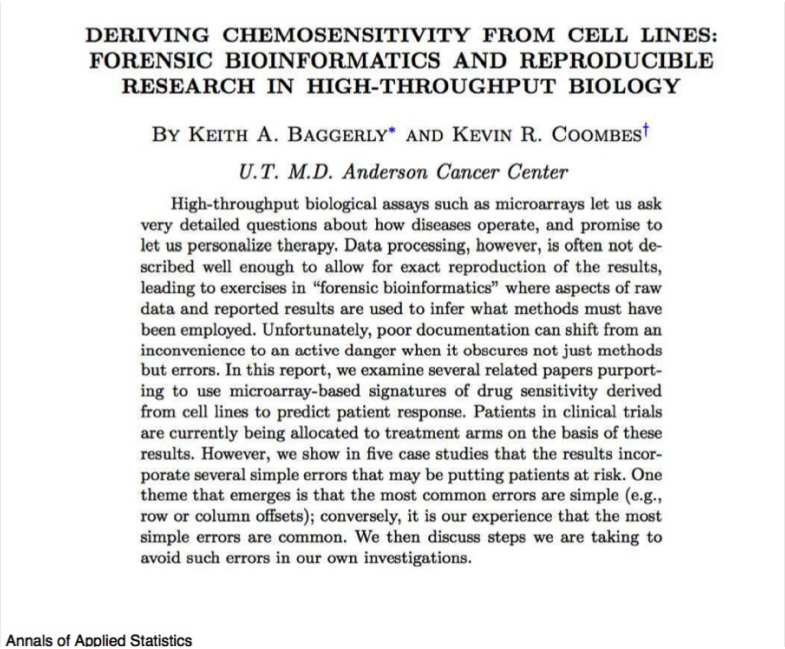
- Send to a friend
- Export citation
- Export references
- Rights and permissions
- Order commercial reprints

SEARCH PUBMED FOR

- Anil Potti
- Holly K Dressman
- Andrea Bild
- Richard F Riedel
- Gina Chan
- Robyn Sayer

http://www.nature.com/nm/journal/v12/n11/full/nm1491.html

Why you should care - uh oh!



DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly* and Kevin R. Coombes†

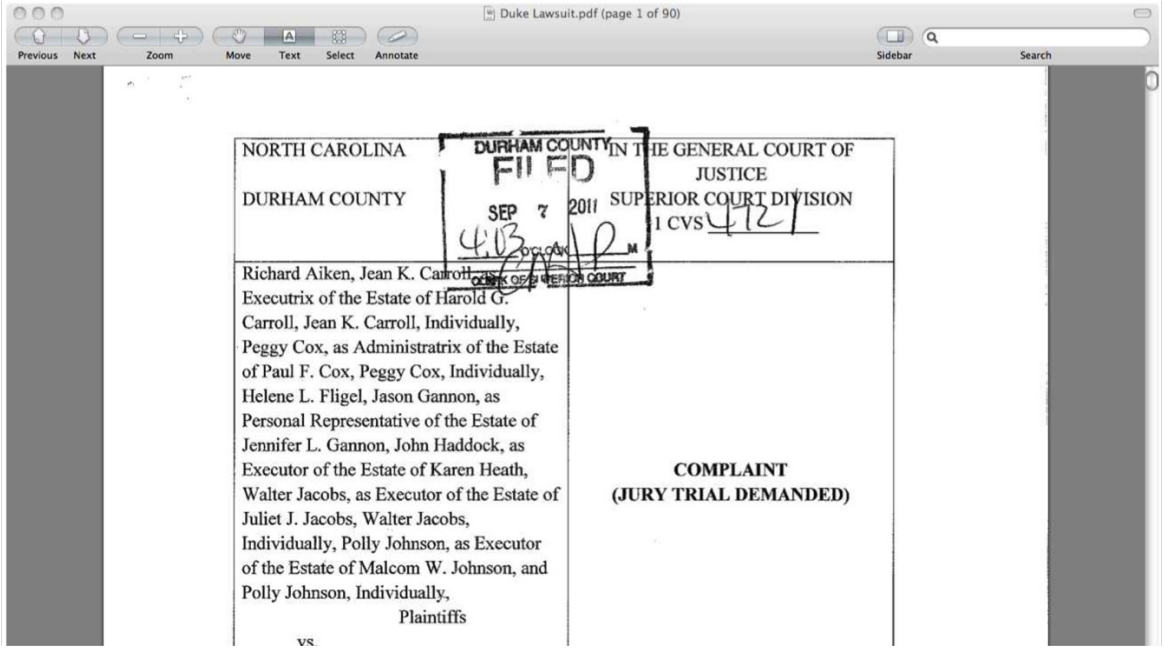
U.T. M.D. Anderson Cancer Center

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results, leading to exercises in "forensic bioinformatics" where aspects of raw data and reported results are used to infer what methods must have been employed. Unfortunately, poor documentation can shift from an inconvenience to an active danger when it obscures not just methods but errors. In this report, we examine several related papers purporting to use microarray-based signatures of drug sensitivity derived from cell lines to predict patient response. Patients in clinical trials are currently being allocated to treatment arms on the basis of these results. However, we show in five case studies that the results incorporate several simple errors that may be putting patients at risk. One theme that emerges is that the most common errors are simple (e.g., row or column offsets); conversely, it is our experience that the most simple errors are common. We then discuss steps we are taking to avoid such errors in our own investigations.

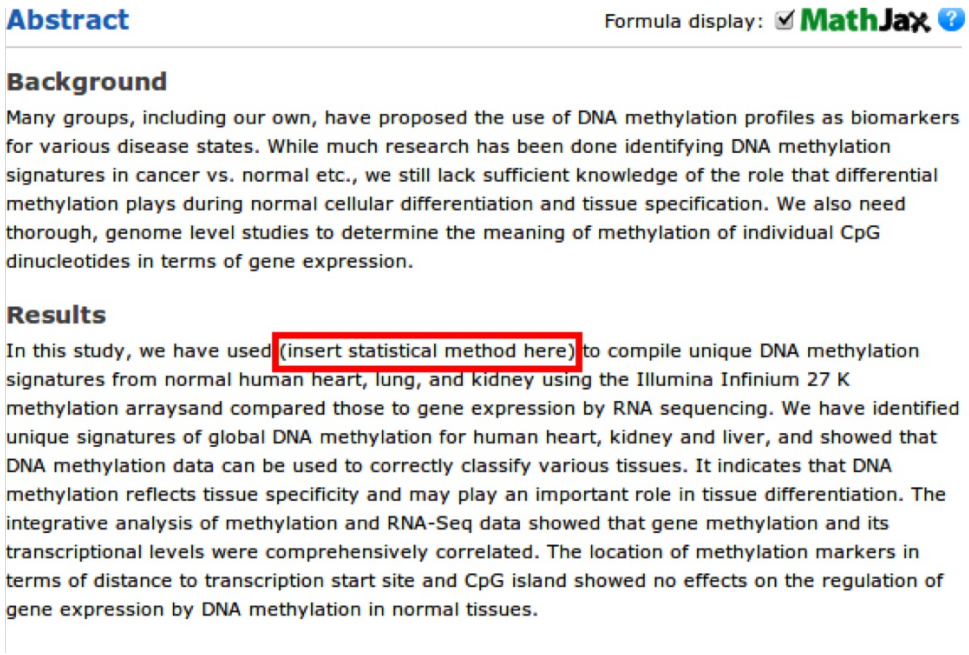
Annals of Applied Statistics

http://arxiv.org/pdf/1010.1092.pdf

Why you should care - serious trouble



Know and care about the analysis plan!



Abstract

Formula display: MathJax

Background

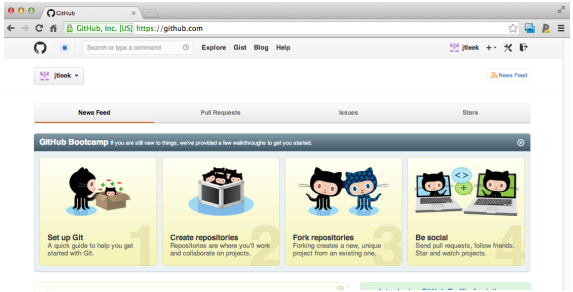
Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation plays during normal cellular differentiation and tissue specification. We also need thorough, genome level studies to determine the meaning of methylation of individual CpG dinucleotides in terms of gene expression.

Results

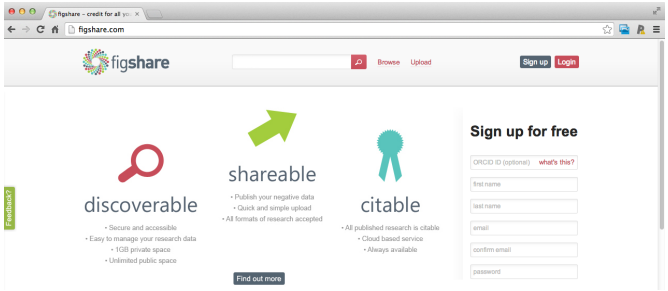
In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arraysand compared those to gene expression by RNA sequencing. We have identified unique signatures of global DNA methylation for human heart, kidney and liver, and showed that DNA methylation data can be used to correctly classify various tissues. It indicates that DNA methylation reflects tissue specificity and may play an important role in tissue differentiation. The integrative analysis of methylation and RNA-Seq data showed that gene methylation and its transcriptional levels were comprehensively correlated. The location of methylation markers in terms of distance to transcription start site and CpG island showed no effects on the regulation of gene expression by DNA methylation in normal tissues.

http://nsaunders.wordpress.com/2012/07/23/we-really-dont-care-what-statistical-method-you-used/

Have a plan for data and code sharing

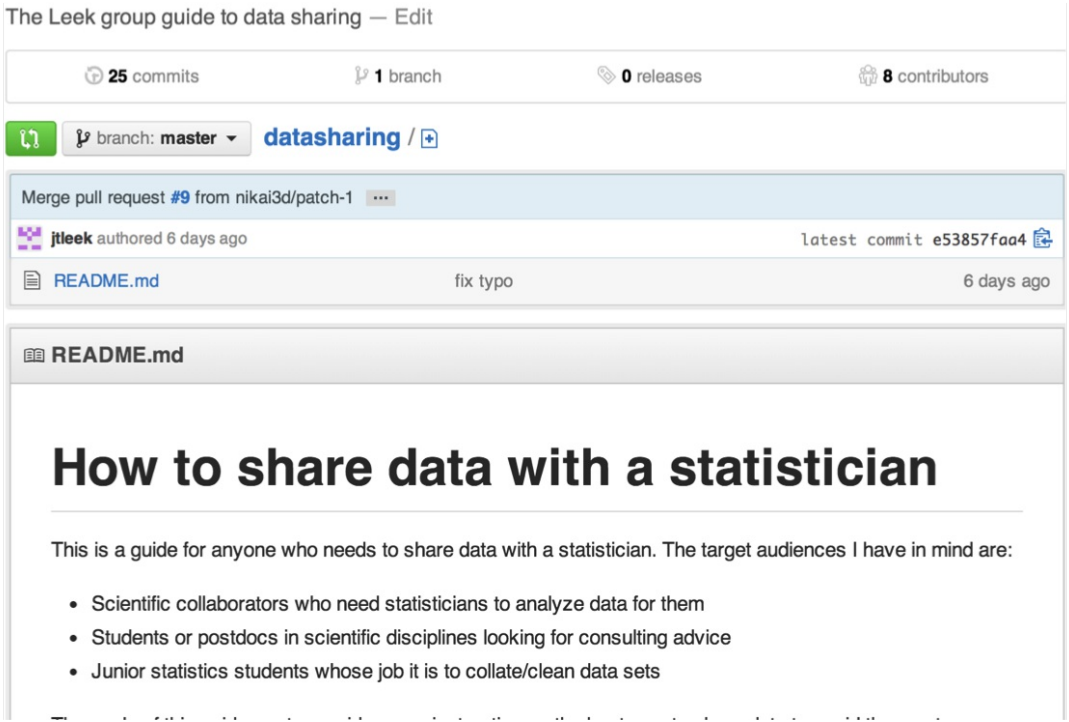


https://github.com/



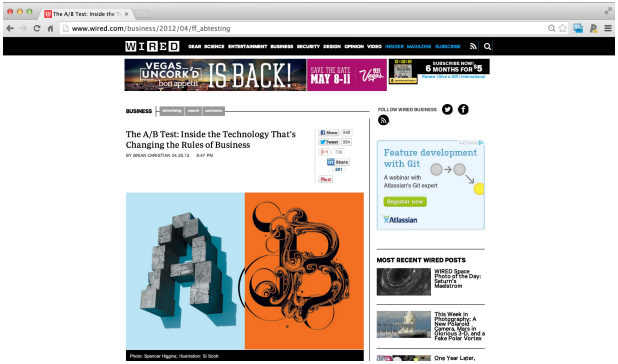
http://figshare.com/

May I recommend?



https://github.com/jtleek/datasharing

Formulate your question in advance



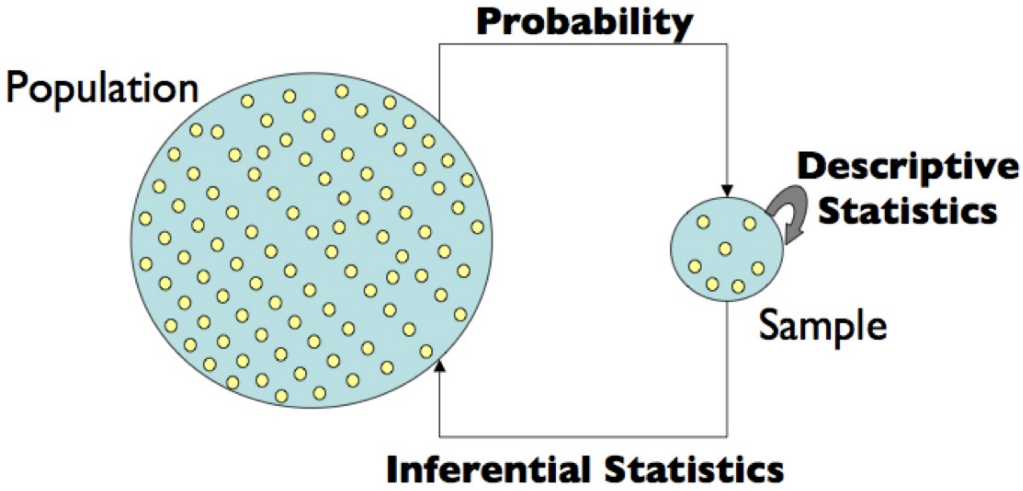
Question: Does changing the text on your website improve donations?

Experiment:

1. Randomly show visitors one version or the other
2. Measure how much they donate
3. Determine which is better

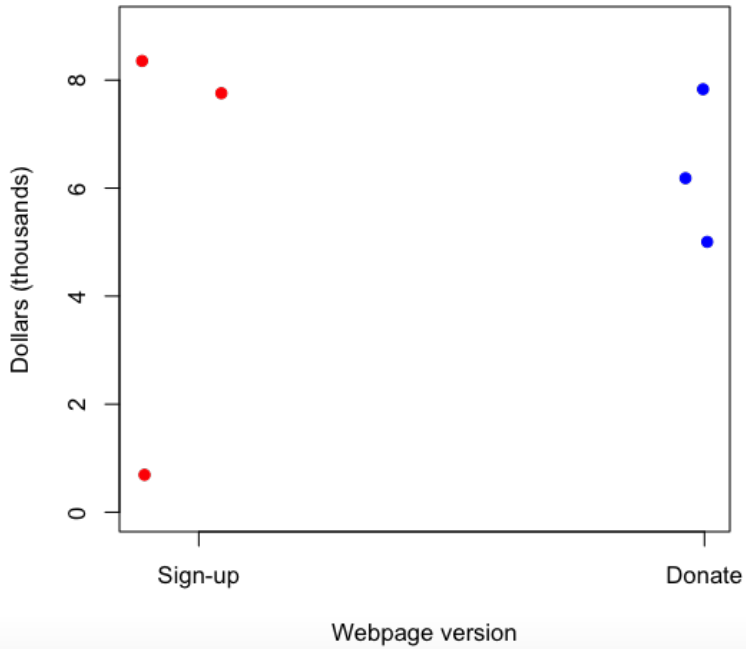
http://www.wired.com/business/2012/04/ff_abtesting

Statistical inference

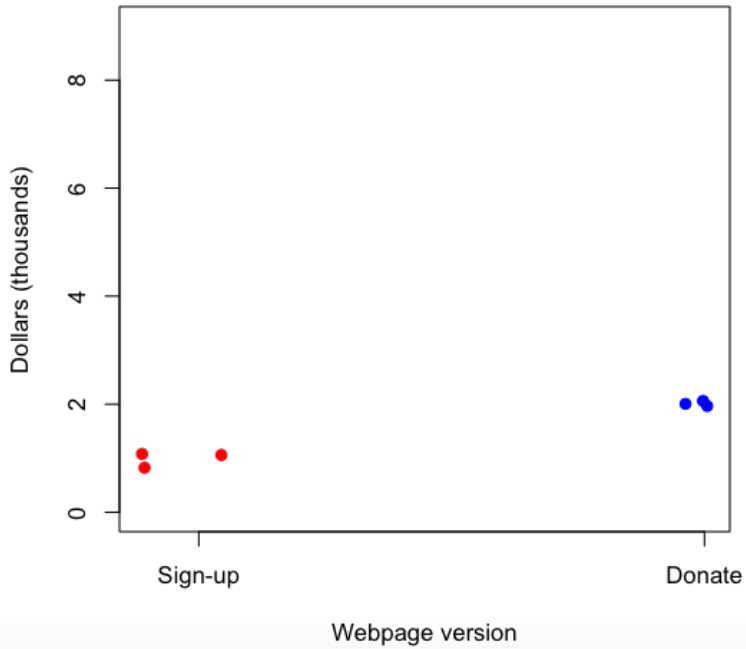


http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf

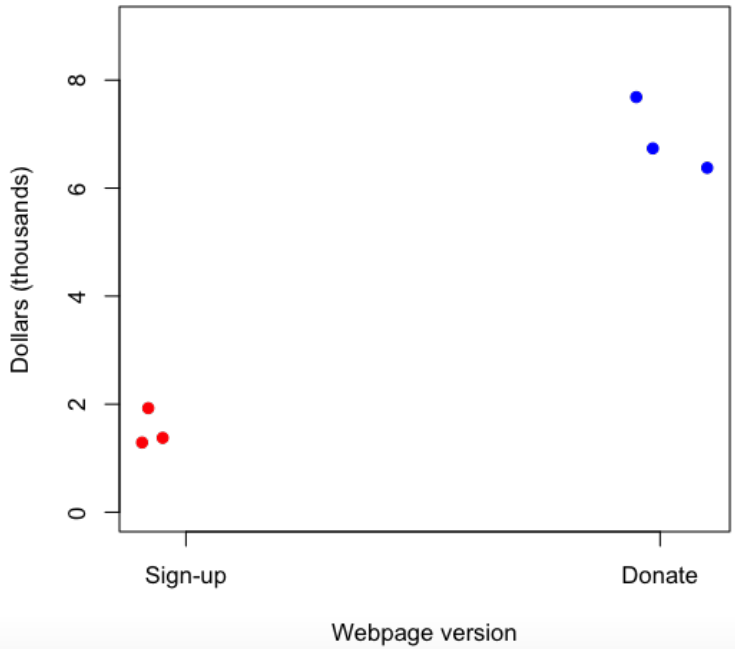
Variability - Scenario 1



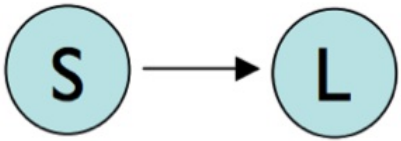
Variability - Scenario 2

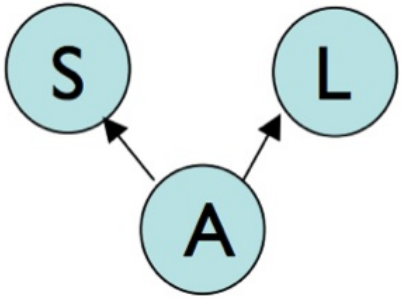


Variability - Scenario 3

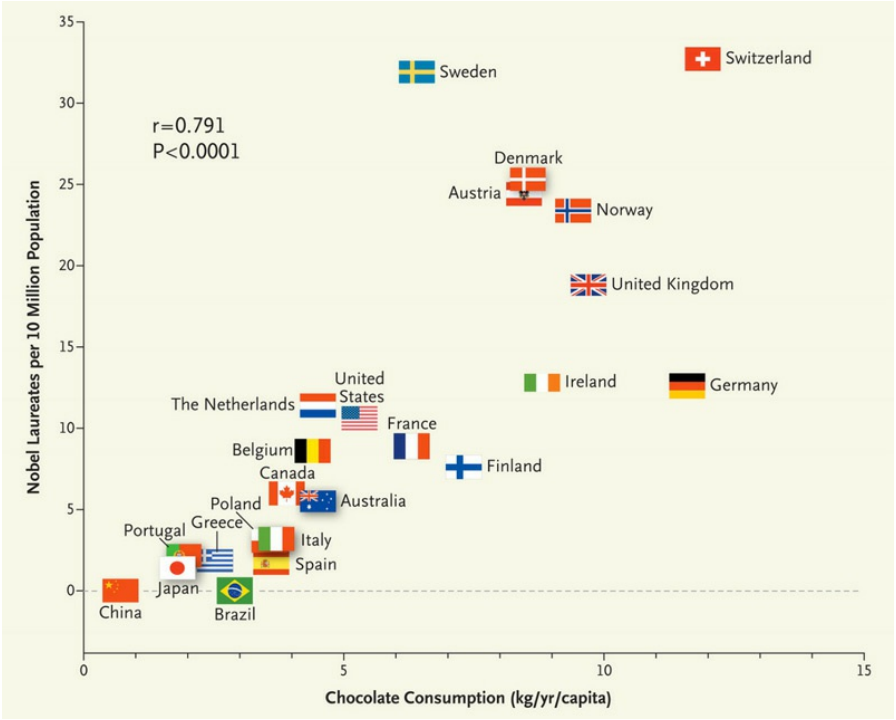


Confounding





Correlation is not causation*



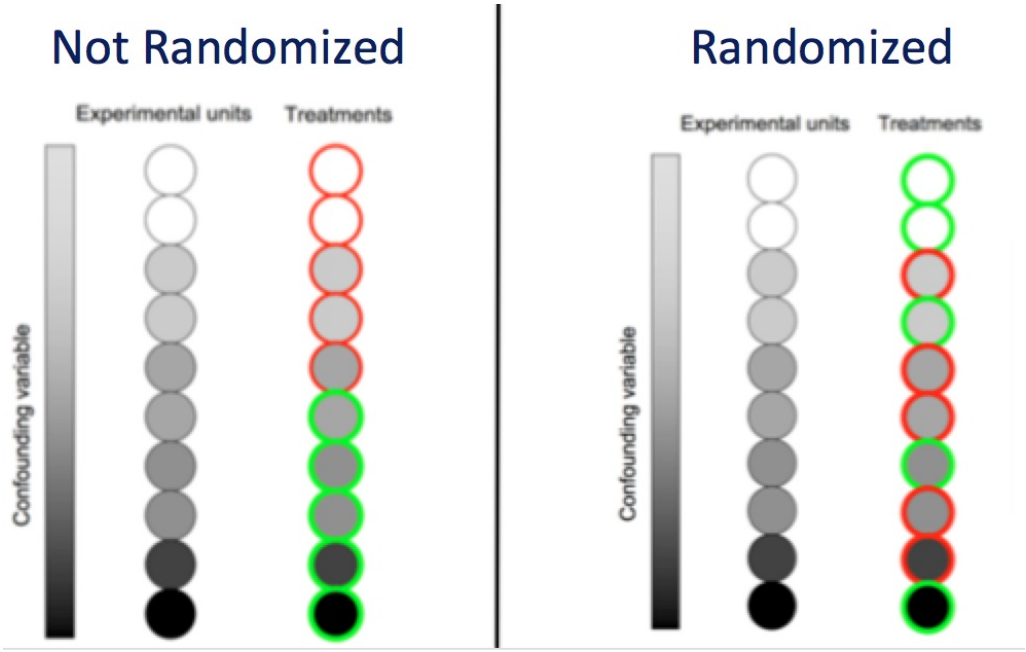
http://www.nejm.org/doi/full/10.1056/NEJMon1211064

*Sometimes called spurious correlation**

Randomization and blocking

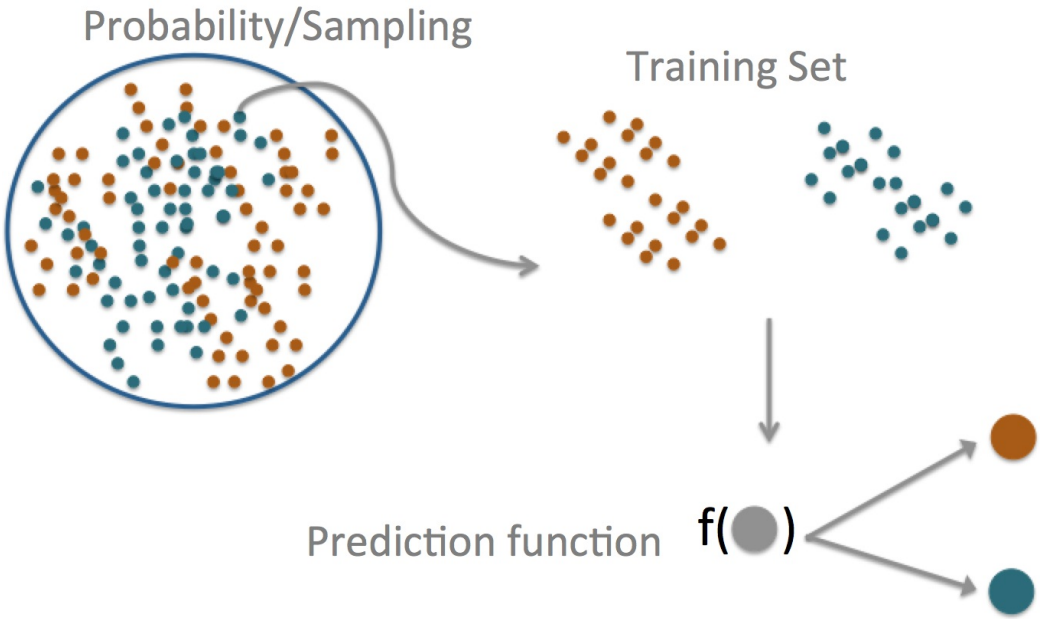
- If you can (and want to) fix a variable
  - Website always says Obama 2014 on it
- If you don't fix a variable, stratify it
  - If you are testing sign up phrases and have two website colors, use both phrases equally on both.
- If you can't fix a variable, randomize it

Why does randomization help?

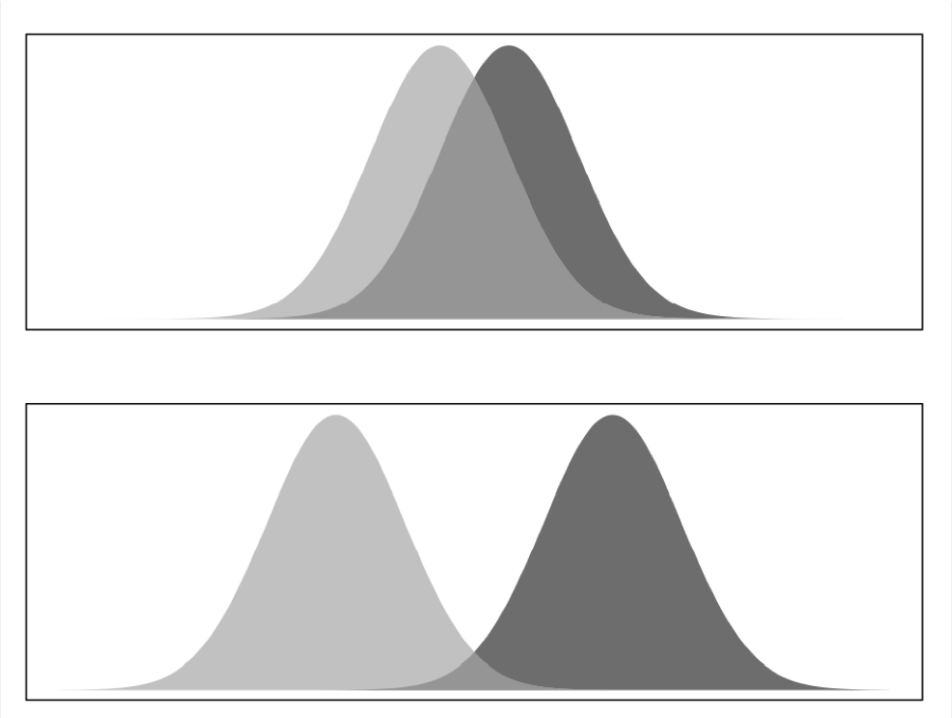


http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdf

Prediction

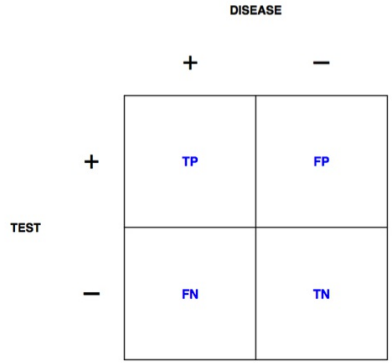


Prediction versus inference



http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/

Prediction key quantities



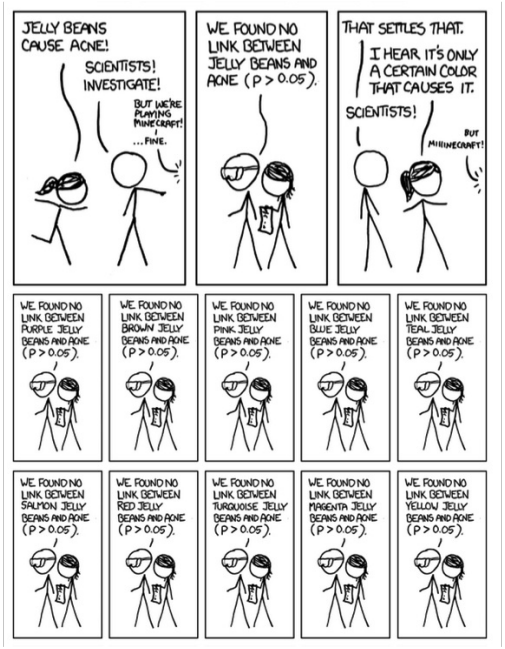
Sensitivity	→	Pr (positive test | disease)
Specificity	→	Pr (negative test | no disease)
Positive Predictive Value	→	Pr (disease | positive test)
Negative Predictive Value	→	Pr (no disease | negative test)
Accuracy	→	Pr (correct outcome)

http://www.biostat.jhsph.edu/~iruczins/teaching/140.615/

Beware data dredging

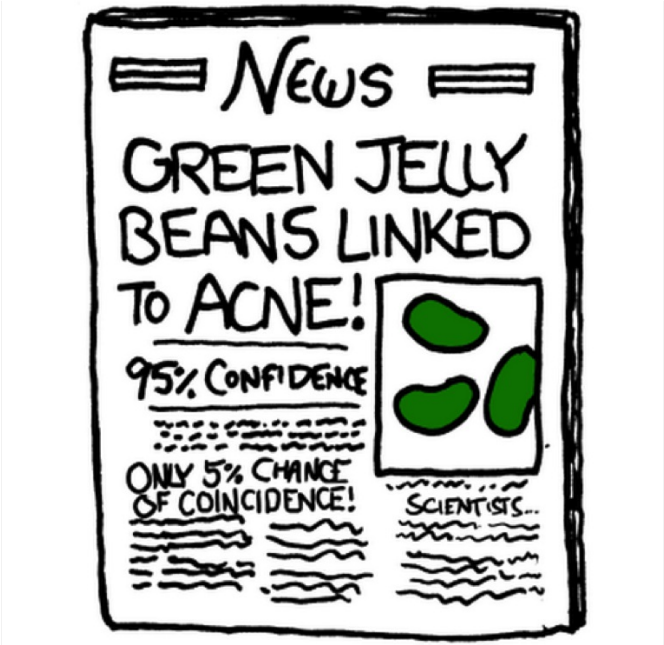
http://xkcd.com/882/

Beware data dredging



http://xkcd.com/882/

Beware data dredging



http://xkcd.com/882/

Summary

- Good experiments
  - Have replication
  - Meaure variability
  - Generalize to the problem you care about
  - Are transparent
- Prediction is not inference
  - Both can be important
- Beware data dredging