



# 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

Return to the 2010 Census Homepage

www.census.gov/2010census/

United States Census 2010  
IT'S IN OUR HANDS

2010 Census Home Press & Media Partners Students & Teachers Census.gov

ABOUT DATA CONNECT MULTIMEDIA

A Look at Your Community

View 2010 Census statistics for local areas down to the block level. Statistics include population counts, age, sex, race, ethnicity and household information.

See More

< 1 2 3 4 >

Map showing Congressional District boundaries for Massachusetts. A callout box highlights "AL - Congressional District #4" with the following data:

Race	Total Population = 164,481
White	135,911
African American	35,571
Asian	2,200
Native Hawaiian/Pacific Islander	348
American Indian/Alaska Native	1,000
Two or more races	9,254

Population Finder

Select a state to begin

Select a state

Interactive Map

Use the Interactive Population Map to explore 2010 Census statistics.

Census Briefs and Reports

www.census.gov/2010census/

2010 Census: District of Columbia Profile

Population by Sex and Age

Total Population: 601,723

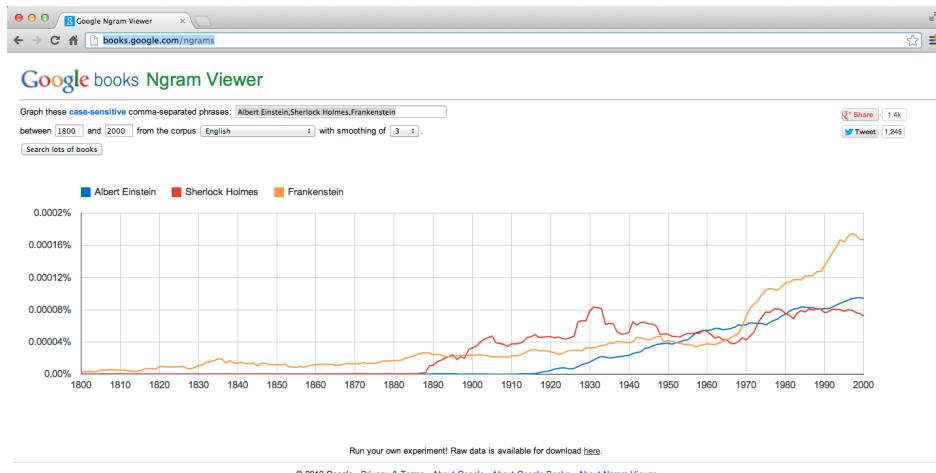
2010 Census: State Population Profile Maps

View detailed population and housing data from the 2010 Census for each state. Each map includes a pie chart showing population by race, a map showing population density, and a bar chart illustrating housing occupancy rates.

See More

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

# Descriptive analysis



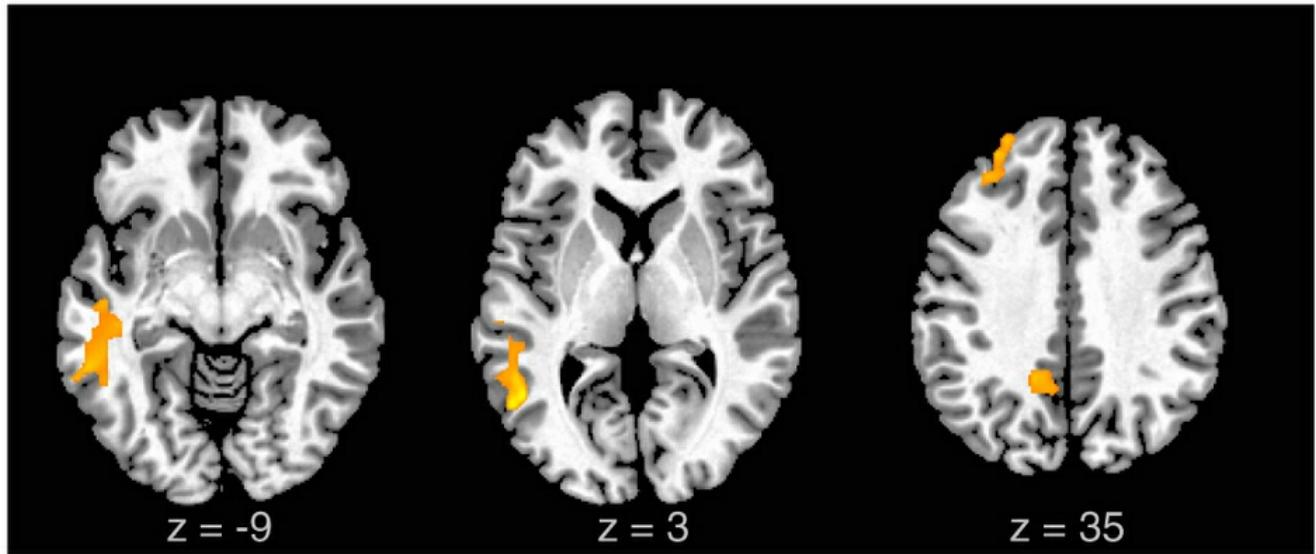
<http://books.google.com/ngrams>

# 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

The Sloan Digital Sky Survey (SDSS) is one of the most ambitious and influential surveys in the history of astronomy. Over eight years of operations (SDSS-I, 2000-2005; SDSS-II, 2005-2008), it obtained deep, multi-color images covering more than a quarter of the sky and created 3-dimensional maps containing more than 930,000 galaxies and more than 120,000 quasars.

SDSS data have been released to the scientific community and the general public in annual increments, with the final public data release from SDSS-II occurring in October 2008. That release, [Data Release 7](#), is available through this website.

Meanwhile, SDSS is continuing with the [Third Sloan Digital Sky Survey \(SDSS-III\)](#), a program of four new surveys using SDSS facilities. SDSS-III began observations in July 2008 and released [Data Release 8](#) in January 2011 and [Data Release 9](#) in August 2012. SDSS-III will continue operating and releasing data through 2014.

[Data Release 9](#) contains the first release of BOSS spectroscopy to the public as well as several significant updates to the cumulative SDSS archive.

[Data Release 8](#) contains all images from the SDSS telescope - the largest color image of the sky ever made. It also includes measurements for nearly 500 million stars and galaxies, and spectra of nearly two million. All the images, measurements, and spectra are available free online. You can [browse through sky images](#), look up data for individual objects, or search for objects anywhere in the sky based on any criteria.

The SDSS used a dedicated 2.5-meter telescope at Apache Point Observatory, New Mexico, equipped with two powerful special-purpose instruments. The 120-megapixel camera imaged 1.5 square degrees of sky at a time, about eight times the area of the full moon. A pair of spectrographs fed by optical fibers measured spectra of (and hence distances to) more than 600 galaxies and quasars in a single observation. A custom-designed set of software pipelines kept pace with the enormous data flow from the telescope. The two key technologies that enabled the SDSS, optical fibers and the digital imaging detectors known as CCDs, were the discoveries awarded the [2009 Nobel Prize in Physics](#).

During its first phase of operations, 2000-2005, the SDSS imaged more than 8,000 square degrees of the sky in five optical bandpasses, and it obtained spectra of galaxies and quasars selected from 5,700 square degrees of that imaging. It also obtained repeated imaging (roughly 30 scans) of a 300 square degree stripe in the southern Galactic cap.

With new financial support and an expanded collaboration including 25 institutions around the globe, SDSS-II carried out three distinct surveys:

- [The Sloan Legacy Survey](#) completed the original SDSS imaging and spectroscopic goals. The final dataset includes 230 million celestial objects detected in 8,400 square degrees of imaging and spectra of 930,000 galaxies, 120,000 quasars, and 225,000 stars.
- [SFD4DE](#) (the Sloan Extension for Galactic Understanding and Exploration) mapped the structure and history of the Milky Way galaxy with new imaging of

<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.<sup>a</sup>; Pope, C. Arden III<sup>b</sup>; Dockery, Douglas W.<sup>c</sup>; Wang, Yun<sup>a</sup>; Ezzati, Majid<sup>d</sup>; Dominici, Francesca<sup>a</sup>

FREE

SDC

[Article Outline](#)

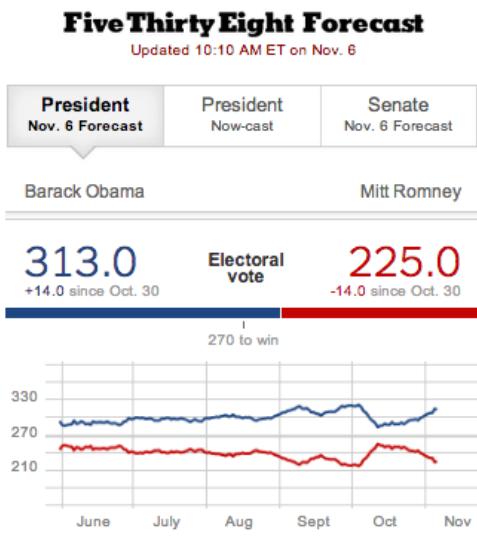
[Correia et al. \(2013\) Epidemiology](#)

# 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

The screenshot shows a web browser displaying an article from Forbes. The title of the article is "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did". The article is authored by Kashmir Hill, a Forbes Staff member. The text discusses how retailers like Target use consumer data to predict pregnancy. On the right side of the article, there is a large red Target logo with the word "TARGET" below it. To the right of the logo, the Coviden logo is visible. The page includes social sharing buttons for various platforms like Facebook, Twitter, and LinkedIn.

<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

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 Nieuworp, 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., Jaap 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 Print Download Citation Supplementary Material E-Mail Save Article Alert Reprints Permissions Share/Bookmark

**TOPICS**  
Gastroenterology > Bacterial Infections >

**MORE IN**  
Research >

**TRENDS**  
Most Viewed (Last Week)

**ORIGINAL ARTICLE**  
Duodenal Infusion of Donor Feces for

SUBSCRIBE OR RENEW TODAY >

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



## 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 AASHTO 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).

PAVEMENT AND MATERIALS

[http://www.fhwa.dot.gov/resourcecenter/teams/pavement/pave\\_3pdg.pdf](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

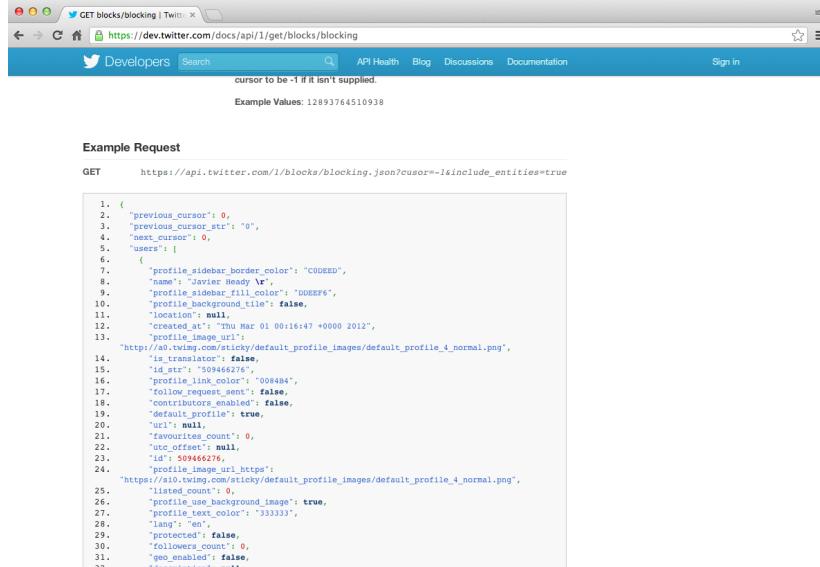
**Quantitative:** Height, weight, blood pressure

# What do data look like?

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGACGGGTTCACCGAGAATGCCGAGACGGATCTGTATGCCGTCTGCTGCGTACAAGACAGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa^b_aa^aa^YaX)aZ^azM^z]Yra]YSG[[ZREQLHESDHNDDHNMEEDDMPPENITKFLFEEDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTCACACTCGCAGTATGGGTCGGCAGCACAGGCAGCGGTAGCCTGCCTTGGCTGGCCTTCGAAAA
+HWI-EAS121:4:100:1783:1611#0/1
a``^\\_`_``^a``^a_`_]a_\`]\`a____`_``^]X)_]XTV_\`])NX_XVX])_TTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTATGTTTGAATATGCTTATCTAACGGTTATTTAGATGTTGGTCTTATTCTAACGGTCATATATTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbabaabbbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``at]a__V\])_`_`^a`_]a_abbaV_
@HWI-EAS121:4:100:1783:1394#0/1
CCCTCTTATTGGTCTGGTGATCCCCATATCTCCGGTTGTCGTTAACCGATCATGCCATTACTCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^][aabbb][`a_abbb`^`bbbbbaaabaaab_Vza_`_bab_X`[a\HV_[_]_`[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGACGGGTTCACCGAGAATGCCGAGACGGATCTGTATGCCGTCTGCTTGAaaaaaaaaaACA
+HWI-EAS121:4:100:1783:207#0/1
abba^Xa``^aa]ba_bba[a_0_a`aa^aa^a]^V]X_a^YS\R_\H[_]\ZTDUZZUSOPX]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAAATTCAAGGACAATGTAATGGCTGCACAAAAAAATACATCTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbaabbbba`bb`ab_0_bab_Q_bbabaa_a
```

[http://brianknaus.com/software/srtoolbox/s\\_4\\_1\\_sequence80.txt](http://brianknaus.com/software/srtoolbox/s_4_1_sequence80.txt)

# What do data look like?



The screenshot shows a browser window with the URL [https://dev.twitter.com/docs/api/1/get\(blocks/blocking](https://dev.twitter.com/docs/api/1/get(blocks/blocking)). The page title is "GET blocks/blocking | Twitter". The main content area displays an "Example Request" for the API endpoint. The request is a GET request to [https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include\\_entities=true](https://api.twitter.com/1/blocks/blocking.json?cursor=-1&include_entities=true). The response body contains a JSON object with 32 properties, representing a user profile. Some properties include:

```
1. {
2.   "previous_cursor": 0,
3.   "previous_cursor_str": "0",
4.   "next_cursor": 0,
5.   "users": [
6.     {
7.       "profile_sidebar_border_color": "CODEED",
8.       "name": "Savier Heady 🌻",
9.       "profile_sidebar_fill_color": "DDEEF6",
10.      "profile_use_background_image": false,
11.      "location": null,
12.      "created_at": "Thu Mar 01 00:16:47 +0000 2012",
13.      "profile_image_url": "http://ai.twimg.com/avatar/default_profile_images/default_profile_4_normal.png",
14.      "follow_request_sent": false,
15.      "id": 509466276,
16.      "profile_link_color": "008484",
17.      "follow_request_sent": false,
18.      "contributors_enabled": false,
19.      "default_profile": true,
20.      "url": null,
21.      "favourites_count": 0,
22.      "utc_offset": null,
23.      "id": 509466276,
24.      "profile_banner_url": "https://ai.twimg.com/sticky/default_profile_images/default_profile_4_normal.png",
25.      "listed_count": 0,
26.      "profile_use_background_image": true,
27.      "profile_text_color": "333333",
28.      "lang": "en",
29.      "protected": false,
30.      "followers_count": 0,
31.      "geo_enabled": false,
32.      "description": null,
```

[https://dev.twitter.com/docs/api/1/get\(blocks/blocking](https://dev.twitter.com/docs/api/1/get(blocks/blocking)

# What do data look like?

ALLERGIES		MEDICATION HISTORY
Last Updated: 01 Dec 2011 @ 0851		Last Updated: 11 Apr 2011 @ 1737
Allergy Name:	TRIMETHOPRIM	Medication: AMLODIPIINE BESYLATE 10MG TAB
Location:	DAYT29	Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TABLET FOR GRAPEFRUIT JUICE--
Date Entered:	09 Mar 2011	Status: Active
Action:		Refills Remaining: 3
Allergy Type:	DRUG	Last Filled On: 28 Aug 2010
A Drug Class:	ANTI-INFECTIVES, OTHER	Initially Ordered On: 13 Aug 2010
Observed/Historical:	HISTORICAL	Quantity: 45
Comments:	The reaction to this allergy was MILD (NO SQUELAE)	Days Supply: 90
Allergy Name:	TRAMADOL	Pharmacy: DAYTON
Location:	DAYT29	Prescription Number: 2718953
Date Entered:	09 Mar 2011	
Action:	URINARY RETENTION	Medication: IBUPROFEN 600MG TAB
Allergy Type:	DRUG	Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY WITH FOOD
A Drug Class:	NON-OPIOID ANALGESICS	Status: Active
Observed/Historical:	HISTORICAL	Refills Remaining: 3
Comments:	gradually worsening difficulty emptying bladder	Last Filled On: 28 Aug 2010
		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](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?

The screenshot shows a SoundCloud page for the set "DarwinTunes" by user "uncoolbob". The main content features a large waveform visualization of the audio snapshot. Below the waveform are standard SoundCloud interaction buttons: Like, Repost, Share, and a link to the full playlist. To the right, the user's stats are displayed: 481 plays, 94 favorites, and 24 likes. A "Follow" button is also present. On the left side of the main content area, there is a decorative illustration of a tree with red flowers. Below the illustration, a text box encourages users to follow "uncoolbob" and others on SoundCloud, with links to "Sign up for SoundCloud" and "Sign in". Further down, a descriptive text explains the project: "Audio snapshots from DarwinTunes.org at various intervals during evolution. As of 2012 we are at 3500 generations and still going strong... you can take part in this interactive experiment here: darwintunes.org/evolve-music". It also states that the playlist contains 43 tracks with a total duration of 1.27.08. At the bottom of the main content, there is a link to the full playlist: "uncoolbob DarwinTunes - evolution of music commentary". The right sidebar contains a summary of the user's activity: 24 likes, 2 reposts, and links to "View all".

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

# What do data look like?

The screenshot shows the Data.gov homepage. At the top, there's a navigation bar with links for HOME, ABOUT, DATA, METRICS, OPEN GOVERNMENT, BLOGS, and COMMUNITIES. Below the navigation is a large map of a coastal area with red and orange shaded regions. Overlaid on the map is the text "SANDY DAMAGE ESTIMATES BY BLOCK GROUP". To the right of the map, there's a sidebar titled "Latest Datasets" which lists various datasets such as "Mississippi River Centerline - Headwa...", "1997 Red River of the North Flood Bou...", etc. Below the map, there are three sections: "DATA AND TOOLS" (with a screenshot of a dashboard), "COMMUNITIES" (with a screenshot of a network graph), and "OPEN GOVERNMENT" (with a screenshot of the American flag).

<http://www.data.gov/>

# What do data look like? Rarely

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	<b><i>id</i></b>	<b><i>problem_id</i></b>	<b><i>subject_id</i></b>	<b><i>start</i></b>	<b><i>stop</i></b>	<b><i>time_left</i></b>										
2	1	498	17	1307119989	1307120016	2369	A									
3	2	150	15	1307119991	1307120009	2376	D									
4	3	313	16	1307119994	1307120009	2374	C									
5	4	12	13	1307119995	1307120019	2368	B									
6	5	273	14	1307119996	1307120028	2357	A									
7	6	101	19	1307119998	1307120021	2364	B									
8	7	103	18	1307119999	1307120048	2337	B									
9	8	162	12	1307120001	1307120042	2342	C									
10	9	70	15	1307120001	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	1307120022	1307120152	2272	B									
15	14	232	14	1307120020	1307120158	2277	C									
16	15	344	15	1307120041	1307120117	2268	B									
17	16	160	17	1307120076	1307120249	2136	D									
18	17	516	16	1307120094	1307120159	2226	B									
19	18	472	12	1307120100	1307120170	2211	A									
20	19	43	15	1307120122	1307120140	2245	C									
21	20	353	13	1307120140	1307120199	2186	C									
22	21	218	15	1307120152	1307120272	2113	E									
23	22	69	16	1307120161	1307120188	2197	D									
24	23	562	16	1307120160	1307120193	2081	D									
25	24	121	19	1307120253	1307120294	2091	E									
26	25	297	15	1307120277	1307120342	2043	B									
27	26	495	13	1307120284	1307120353	2032	E									
28	27	94	14	1307120286	1307120343	2042	E									
29	28	22	18	1307120290	1307120365	2024	C									
30	29	64	19	1307120310	1307120385	2000	B									
31	30	502	16	1307120320	1307120336	2049	B									
32	31	44	16	1307120339	1307120352	2033	A									
33	32	315	14	1307120340	1307120392	2023	B									
34	33	308	15	1307120342	1307120353	2032	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	1307120380	1307120415	1970	E									
39	38	257	14	1307120401	1307120427	1950	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

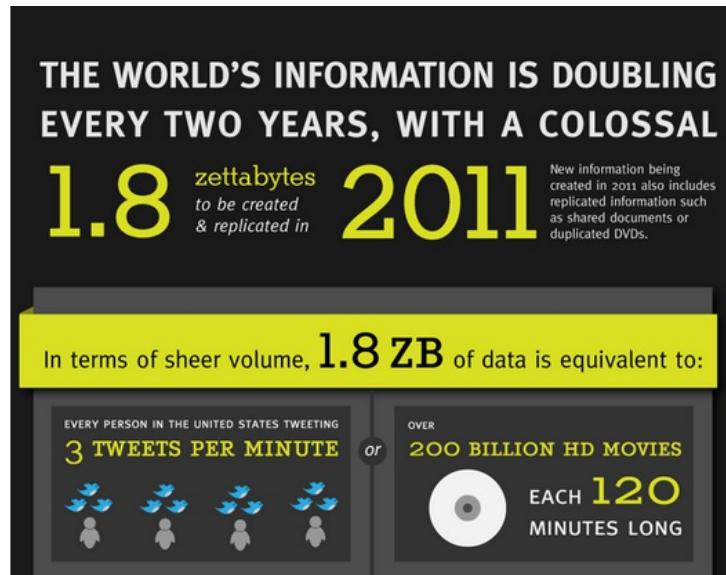
- The most important thing in data science is the question
- The second most important is the data
- Often the data will limit or enable the questions
- But having data can't save you if you don't have a question



# What about big data?

Jeffrey Leek  
Johns Hopkins Bloomberg School of Public Health

# How much is there?

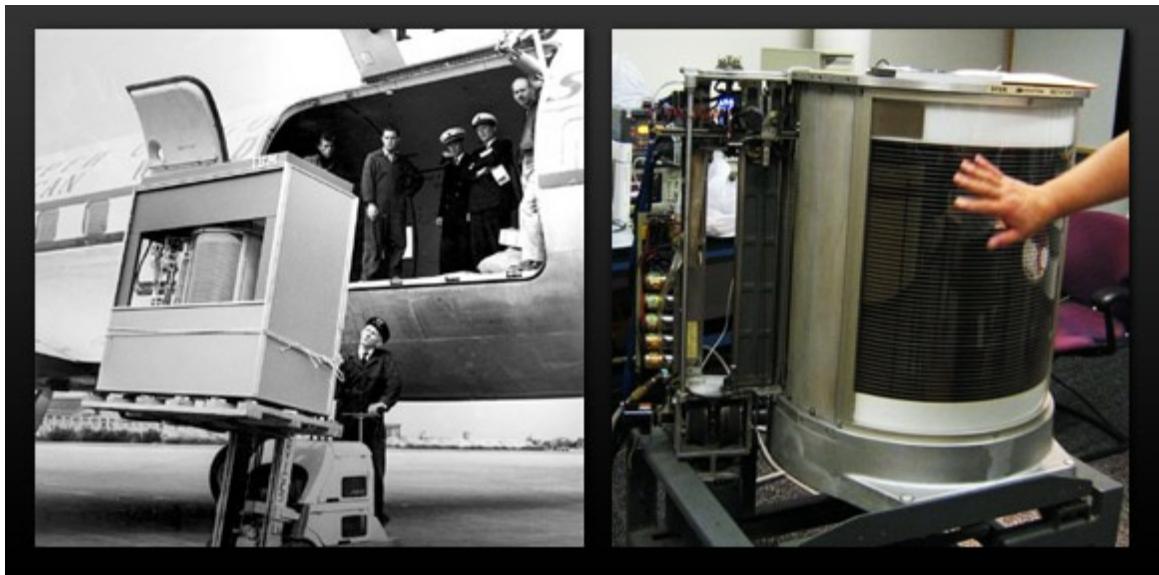


<http://mashable.com/2011/06/28/data-infographic/>

# 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, rather than the actions and characteristics of individuals. The dataset contains summary properties of 30 billion conversations among 240 million people. From the data, we construct a communication graph with 180 million nodes and 1.3 billion undirected edges, creating the largest 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 claim that people are separated by ``six degrees of separation'' and find that the average path length among Messenger users is 6.6. We also find that people tend to communicate more with each other when they have similar age, language, and location, and that 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

The screenshot shows a web browser window with the following details:

- Title Bar:** "Don't use Hadoop - your d" (partially visible)
- Address Bar:** "www.chrisstucchio.com/blog/2013/hadoop\_hatred.html"
- Content Area:**
  - Header:** "Chris Stucchio" (orange text), "Home" (red), "Blog" (red), "Code" (red), "Work" (red)
  - Section Title:** "Don't use Hadoop - your data isn't that big" (orange text)
  - Text:** "Posted: Mon, 16 Sep 2013"
  - Tags:** "big data", "buzzwords", "hadoop"
  - Social Sharing:** "Follow @stucchio", "Tweet", "2,169", "submit", "Like", "Share", "1,055 people like this. Sign Up to see what your friends like.", "g+", "+537 Recommend this on Google", "RSS feed icon".
  - Text:** "So, how much experience do you have with Big Data and Hadoop?" they asked me. I told them that I use Hadoop all the time, but rarely for jobs larger than a few TB. I'm basically a big data neophyte - I know the concepts, I've written code, but never at scale.
  - Text:** The next question they asked me. "Could you use Hadoop to do a simple group by and sum?" Of course I could, and I just told them I needed to see an example of the file format.

[http://www.chrisstucchio.com/blog/2013/hadoop\\_hatred.html](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<sup>1,2</sup>, Holly K Dressman<sup>1,3</sup>, Andrea Bild<sup>1,3</sup>, Richard F Riedel<sup>1,2</sup>, Gina Chan<sup>4</sup>, Robyn Sayer<sup>4</sup>, Janiel Cragun<sup>4</sup>, Hope Cottrill<sup>4</sup>, Michael J Kelley<sup>2</sup>, Rebecca Petersen<sup>5</sup>, David Harpole<sup>5</sup>, Jeffrey Marks<sup>5</sup>, Andrew Berchuck<sup>1,6</sup>, Geoffrey S Ginsburg<sup>1,2</sup>, Phillip Febbo<sup>1,2,3</sup>, Johnathan Lancaster<sup>4</sup> & Joseph R Nevins<sup>1,2,3</sup>

**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†

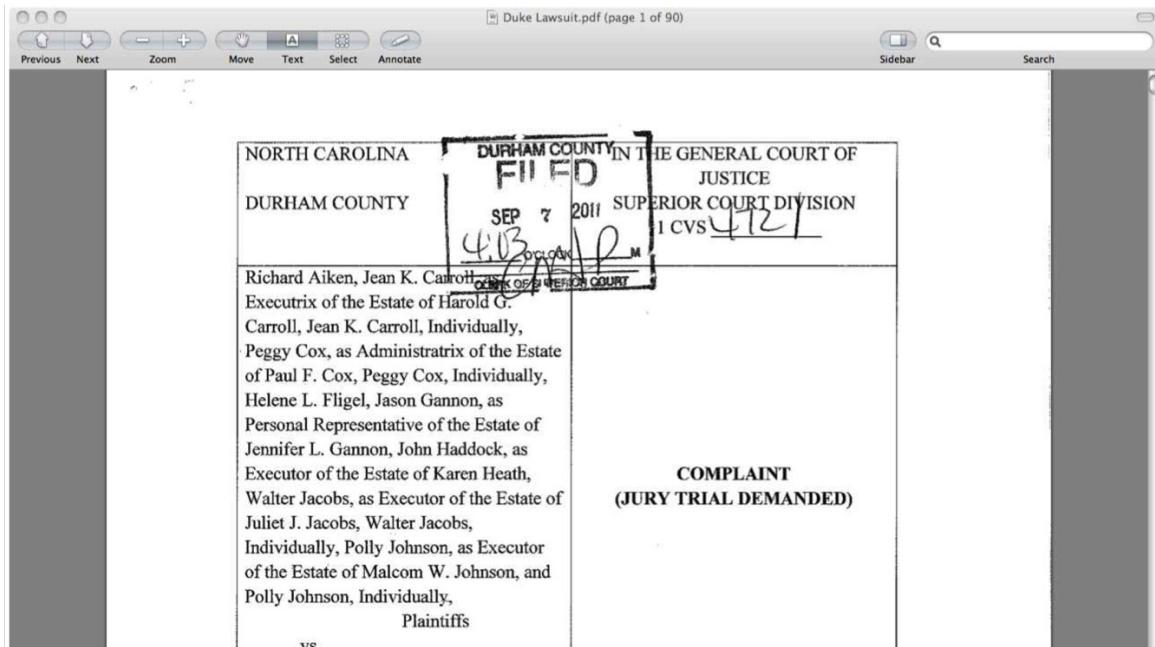
*U.T. M.D. Anderson Cancer Center*

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

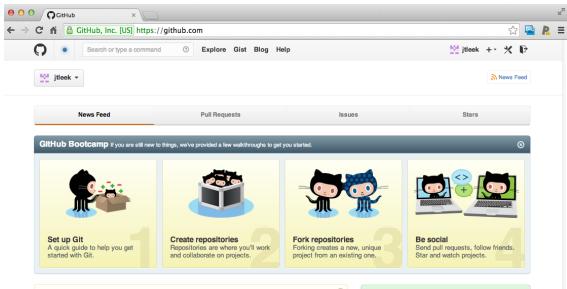
Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation 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

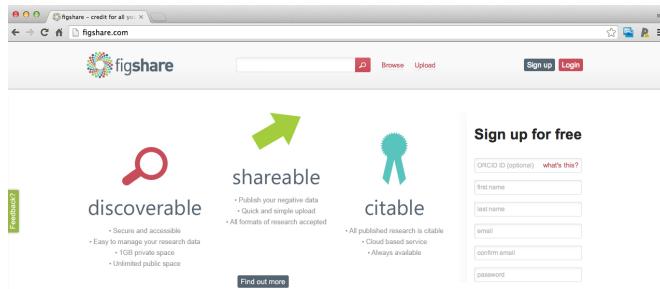
In this study, we have used (insert statistical method here) to compile unique DNA methylation signatures from normal human heart, lung, and kidney using the Illumina Infinium 27 K methylation arrays and compared those to gene expression by RNA sequencing. We have identified unique signatures of global 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?

The Leek group guide to data sharing — Edit

A screenshot of a GitHub repository page for 'datasharing'. The repository has 25 commits, 1 branch, 0 releases, and 8 contributors. The master branch is selected. A merge pull request #9 from nikai3d/patch-1 is shown. The README.md file contains the text 'How to share data with a statistician'. Below the repository details, there is a section titled 'How to share data with a statistician' with a list of target audiences.

25 commits 1 branch 0 releases 8 contributors

branch: master / [datasharing](#) / [+](#)

Merge pull request #9 from nikai3d/patch-1 ...

jtleek authored 6 days ago latest commit e53857faa4 [edit](#)

[README.md](#) fix typo 6 days ago

[README.md](#)

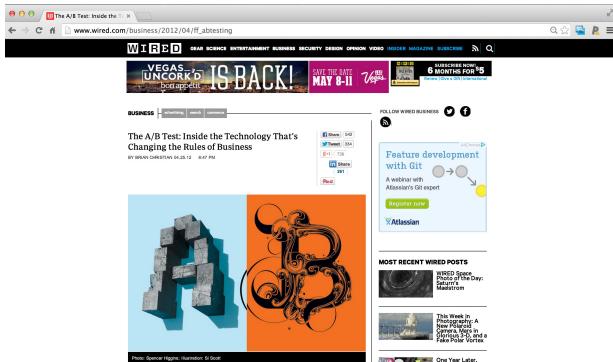
## How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician. The target audiences I have in mind are:

- Scientific collaborators who need statisticians to analyze data for them
- Students or postdocs in scientific disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean data sets

<https://github.com/jtleek/datasharing>

# Formulate your question in advance



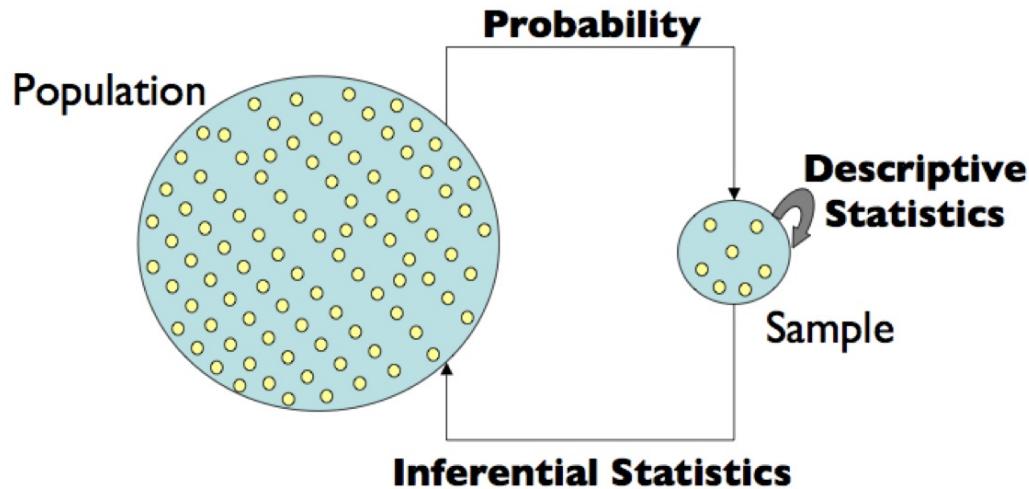
**Question:** Does changing the text on your website improve donations?

**Experiment:**

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

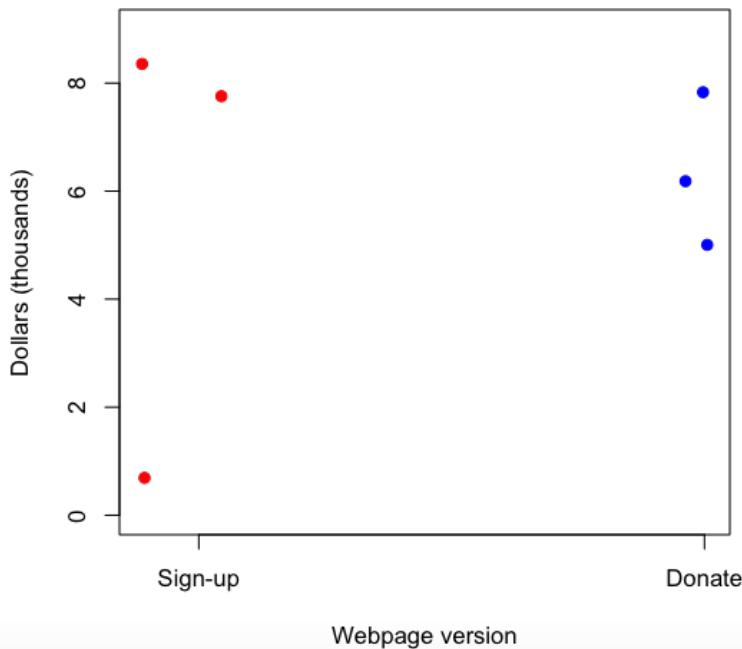
[http://www.wired.com/business/2012/04/ff\\_abtesting](http://www.wired.com/business/2012/04/ff_abtesting)

# Statistical inference

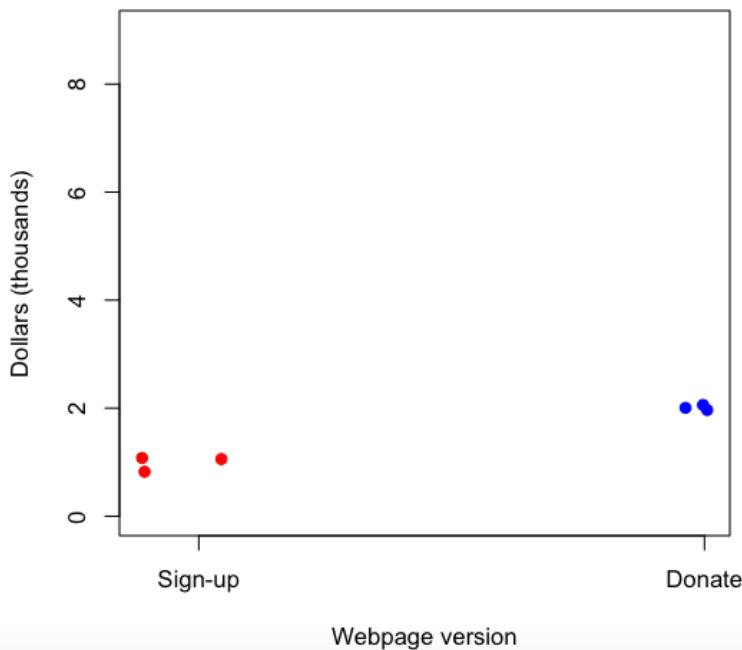


<http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf>

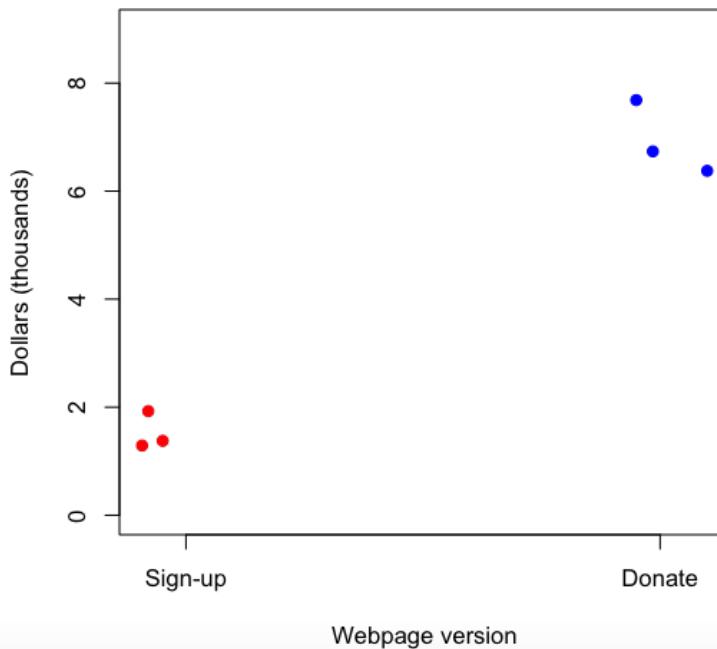
# Variability - Scenario 1



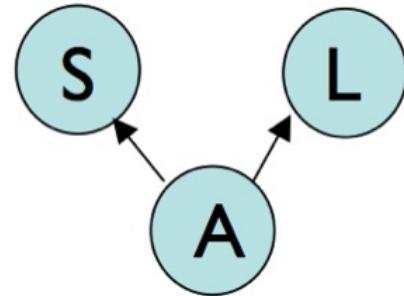
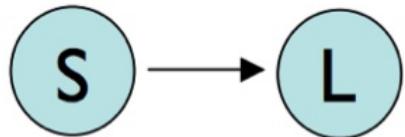
# Variability - Scenario 2



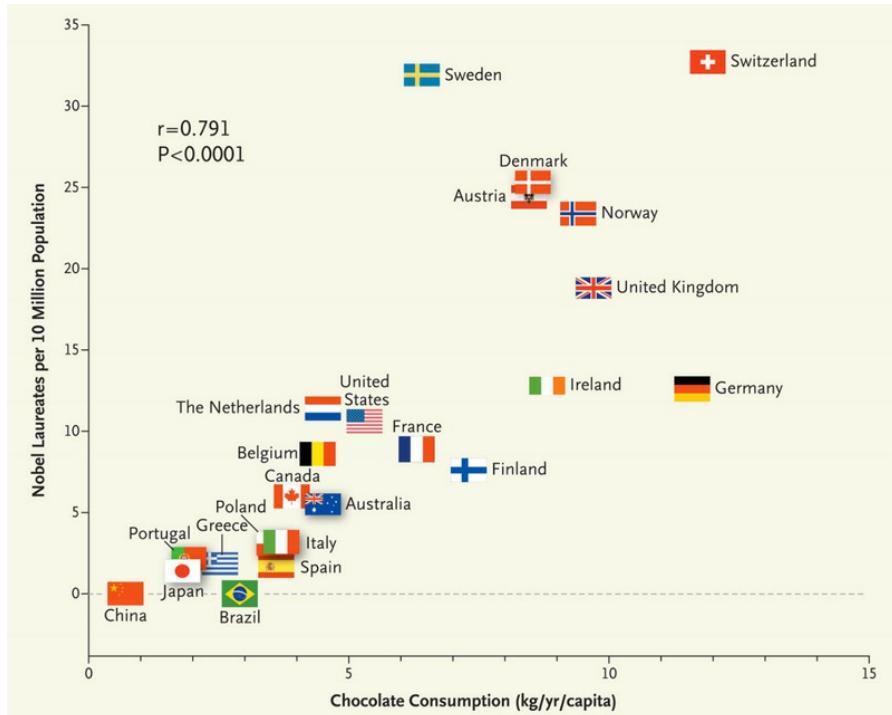
# Variability - Scenario 3



# Confounding



# Correlation is not causation\*



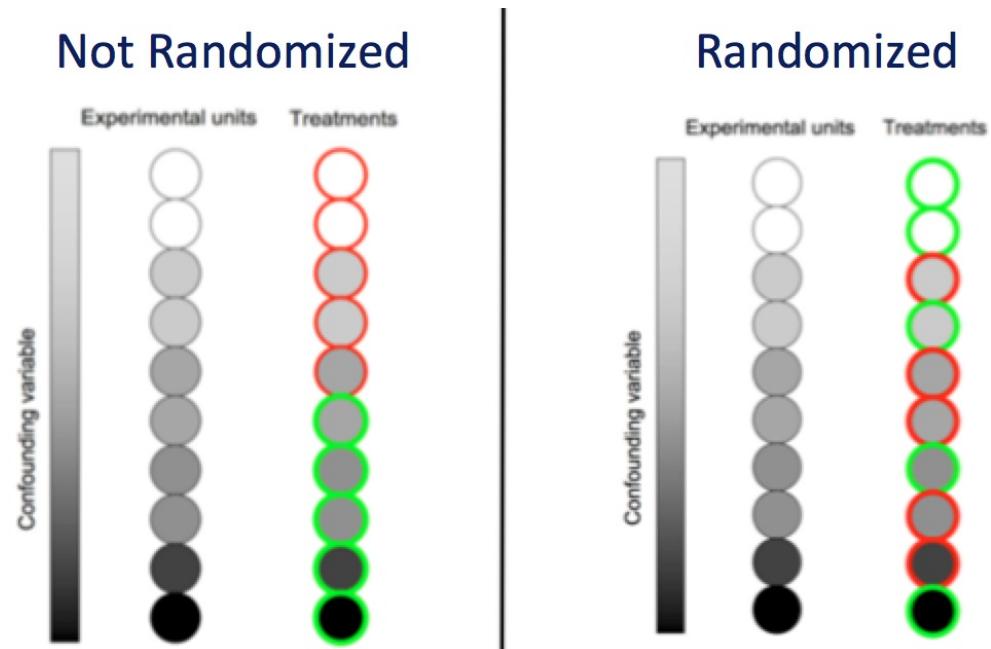
<http://www.nejm.org/doi/full/10.1056/NEJMoa1211064>

Sometimes called spurious correlation\*

# Randomization and blocking

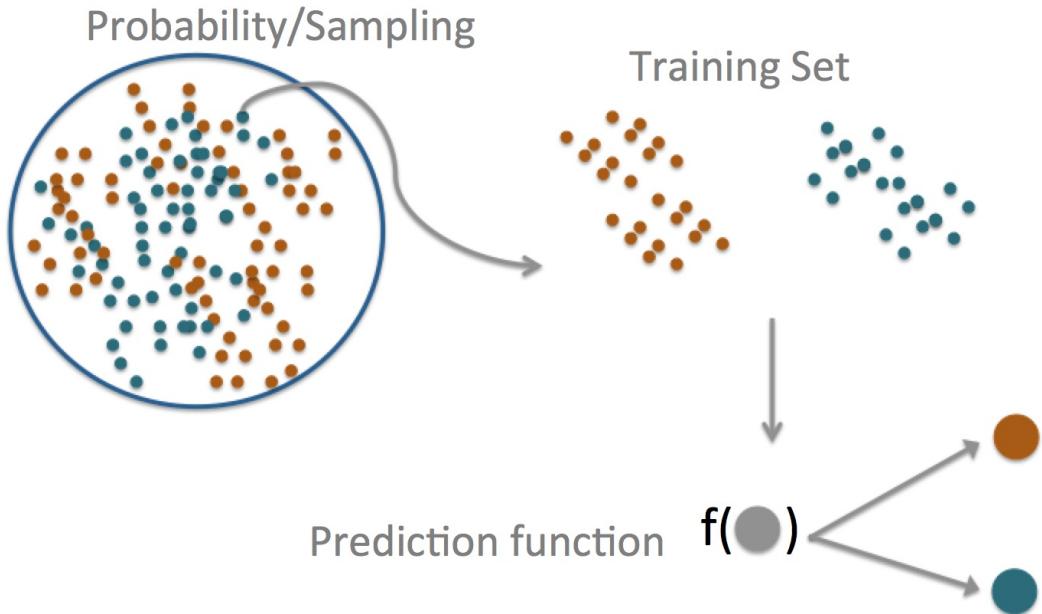
- If you can (and want to) fix a variable
  - Website always says Obama 2014 on it
- If you don't fix a variable, stratify it
  - If you are testing sign up phrases and have two website colors, use both phrases equally on both.
- If you can't fix a variable, randomize it

# Why does randomization help?

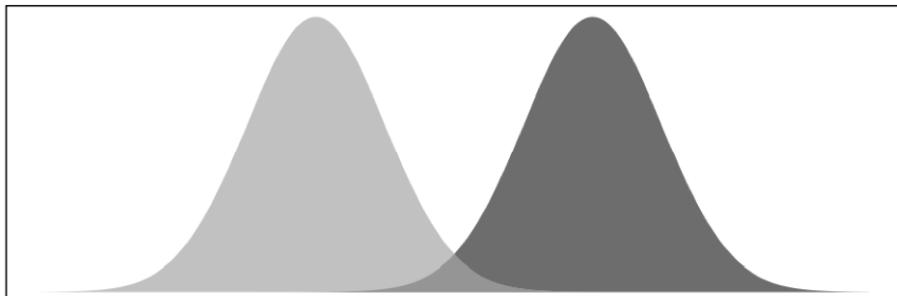
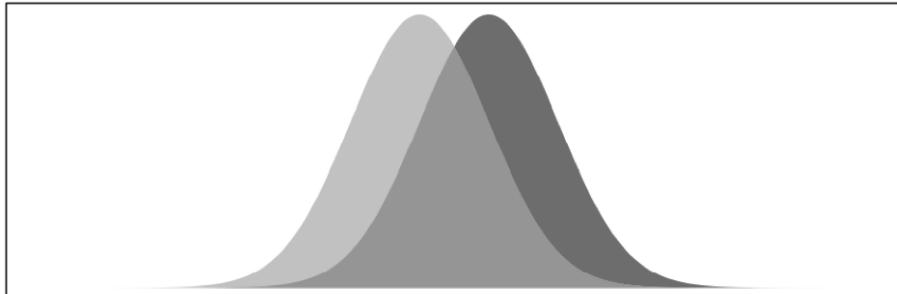


<http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdf>

# Prediction



# Prediction versus inference



# Prediction key quantities

		DISEASE	
		+	-
TEST	+	TP	FP
	-	FN	TN

Sensitivity

→  $\Pr(\text{positive test} \mid \text{disease})$

Specificity

→  $\Pr(\text{negative test} \mid \text{no disease})$

Positive Predictive Value

→  $\Pr(\text{disease} \mid \text{positive test})$

Negative Predictive Value

→  $\Pr(\text{no disease} \mid \text{negative test})$

Accuracy

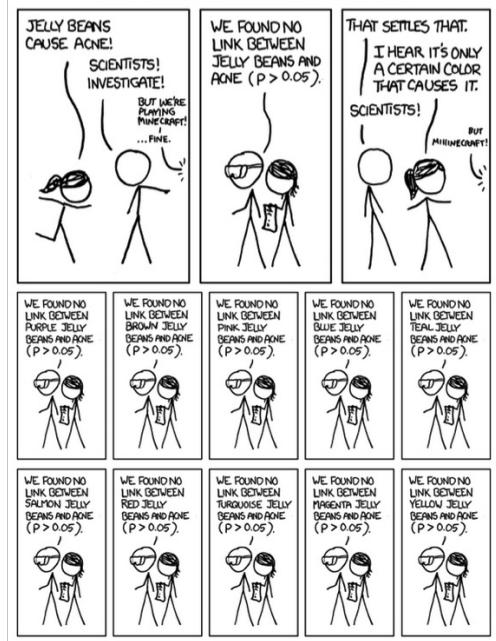
→  $\Pr(\text{correct outcome})$

# Beware data dredging



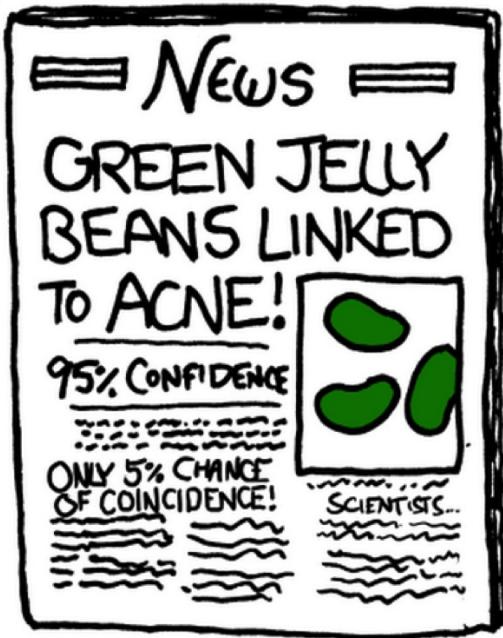
<http://xkcd.com/882/>

# Beware data dredging



<http://xkcd.com/882/>

# Beware data dredging



<http://xkcd.com/882/>

# Summary

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