Introduction to data analysis

doc. Ing. Oleksii Yehorchenkov, DrSc.

Types of Data Science Questions

The data analysis question

Define the data analytic question first

Data can be used to answer many questions, but not all of them.

> *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.*
>
> *John Tukey*

Before performing a data analysis the key is to define the type of question being asked. Some questions are easier to answer with data and some are harder. This is a broad categorization of the types of data analysis questions, ranked by how easy it is to answer the question with data.

The data analysis question type flow chart

In approximate order of difficulty:

Descriptive

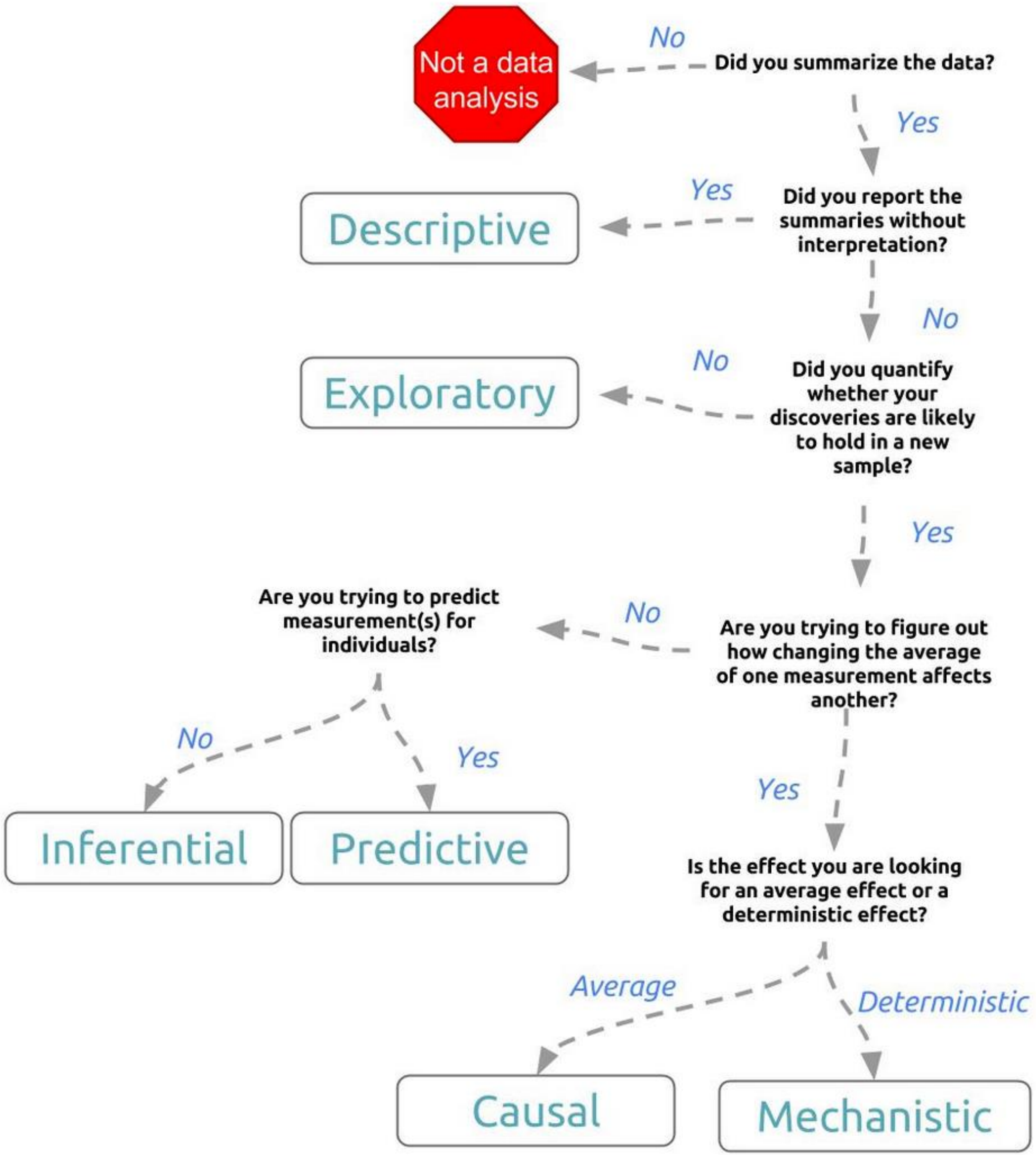
Exploratory

Inferential

Predictive

Causal

Mechanistic



Descriptive

A descriptive data analysis seeks to summarize the measurements in a single data set without further interpretation.

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 (*узагальнені*) without additional statistical modelling

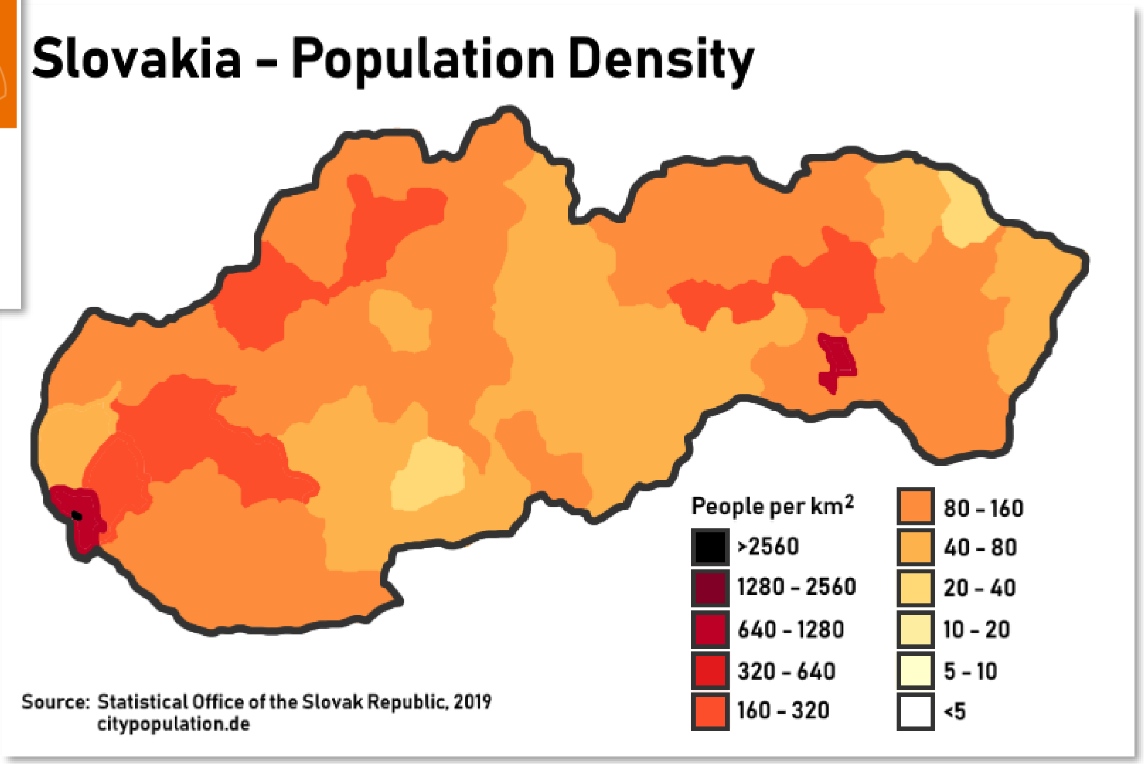
Descriptive

An example is the United States Census. The Census collects data on the residence type, location, age, sex, and race of all people in the United States at a fixed time. The Census is descriptive because the goal is to summarize the measurements in this fixed data set into population counts and describe how many people live in different parts of the United States. The interpretation and use of these counts is left to Congress and the public, but is not part of the data analysis.

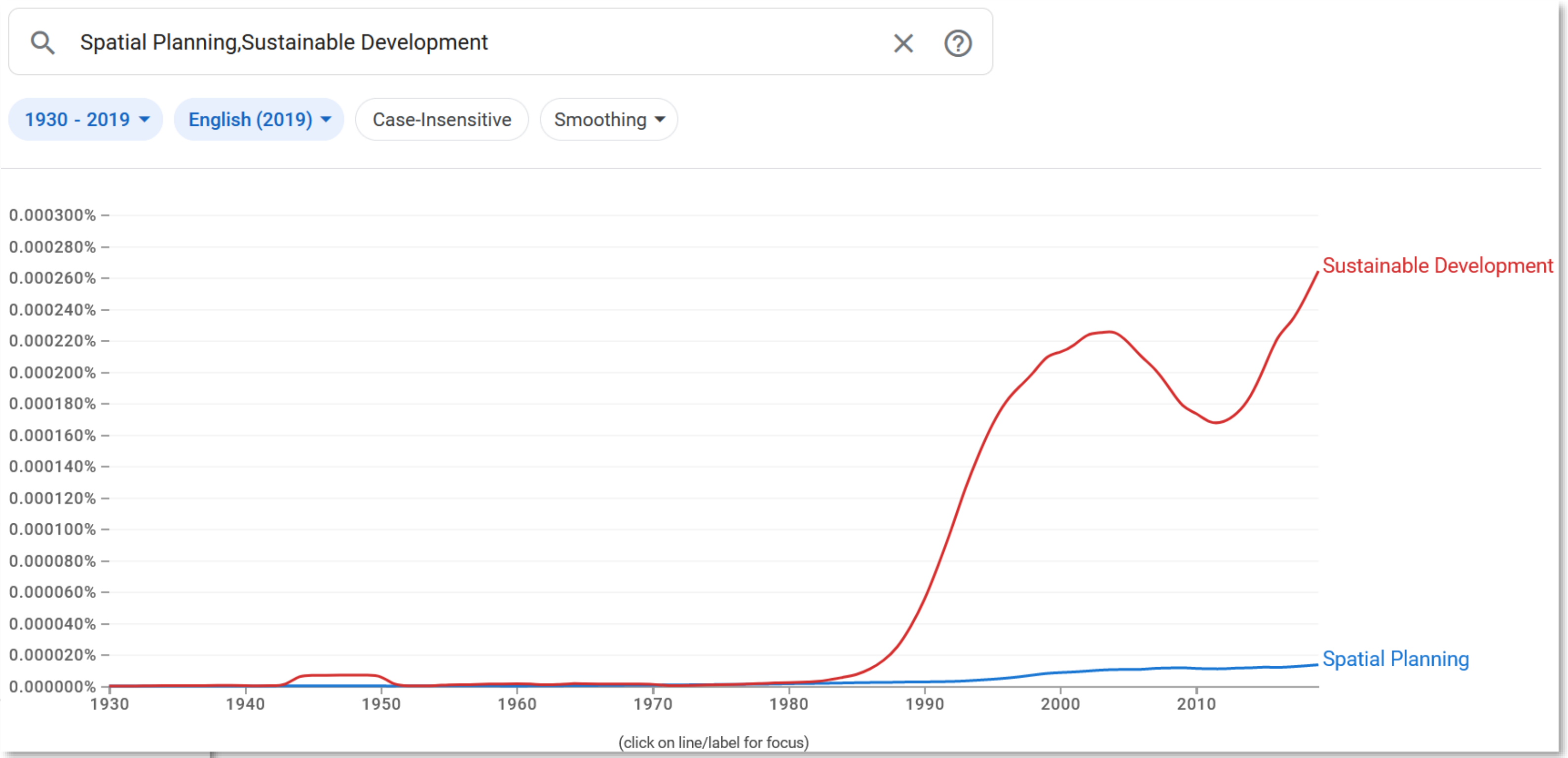
Descriptive

https://www.scitanie.sk/en

https://en.wikipedia.org/wiki/Slovakia

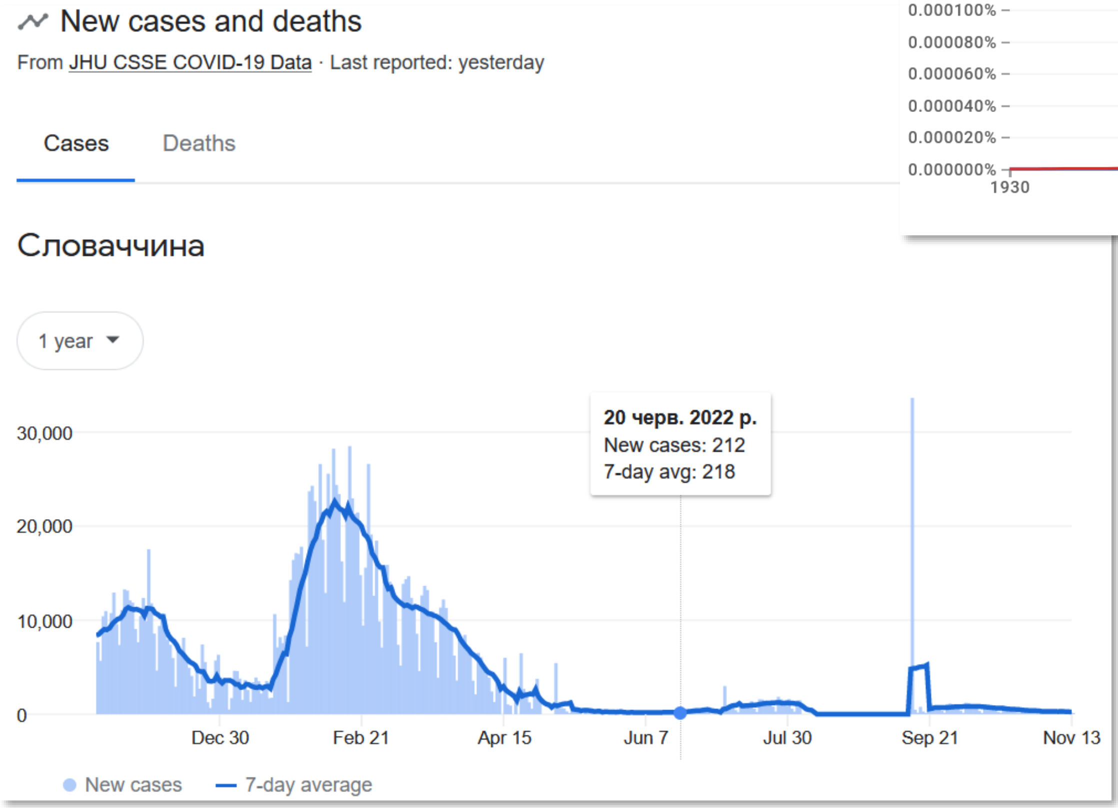


Descriptive



https://books.google.com/ngrams

covid 19 statistics google



Exploratory

An exploratory data analysis builds on a descriptive analysis by searching for discoveries, trends, correlations, or relationships between the measurements of multiple variables to generate ideas or hypotheses.

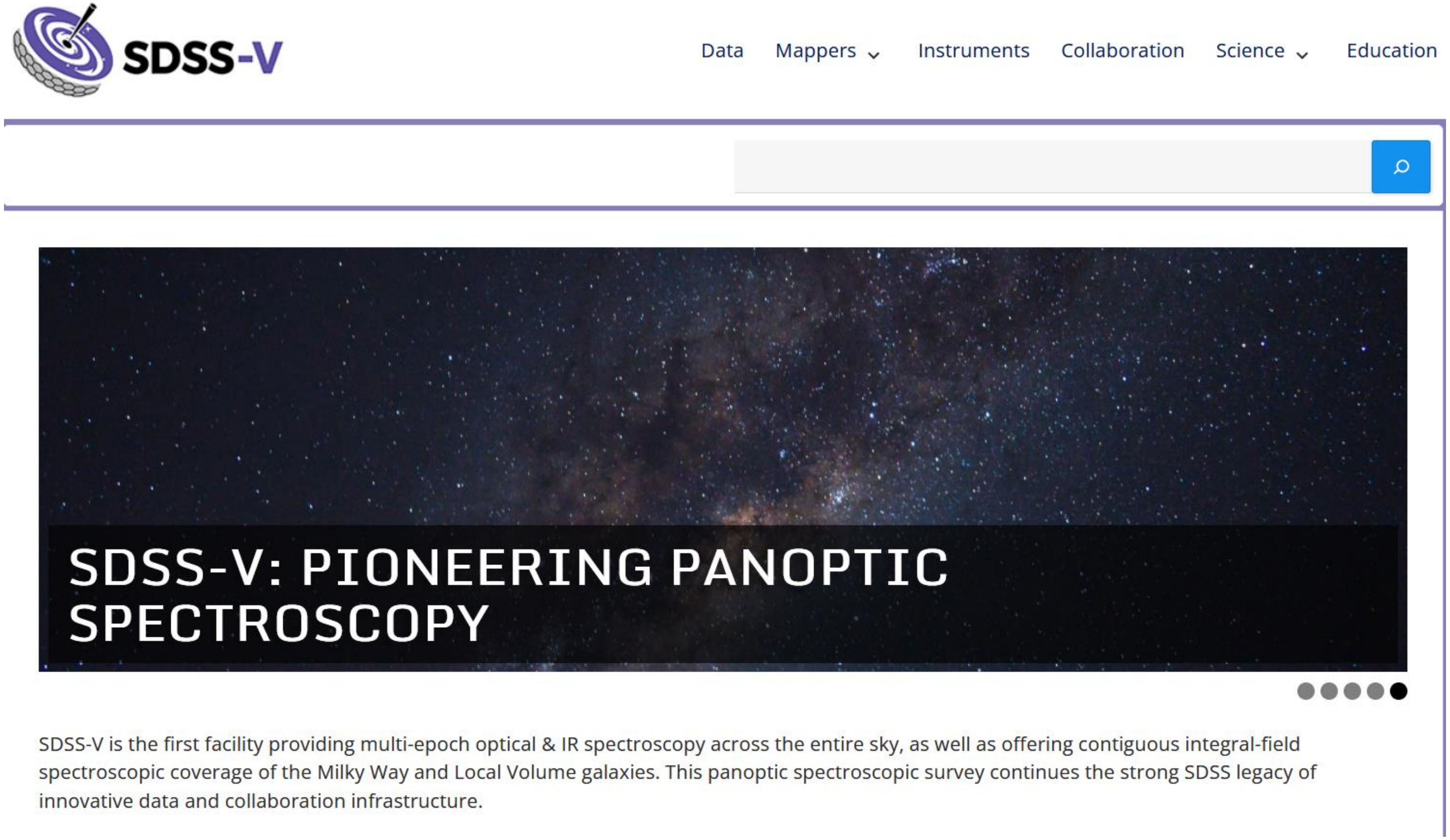
Goal: Find relationships you didn’t know about

- 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
- **Correlations does not imply causation**

Exploratory

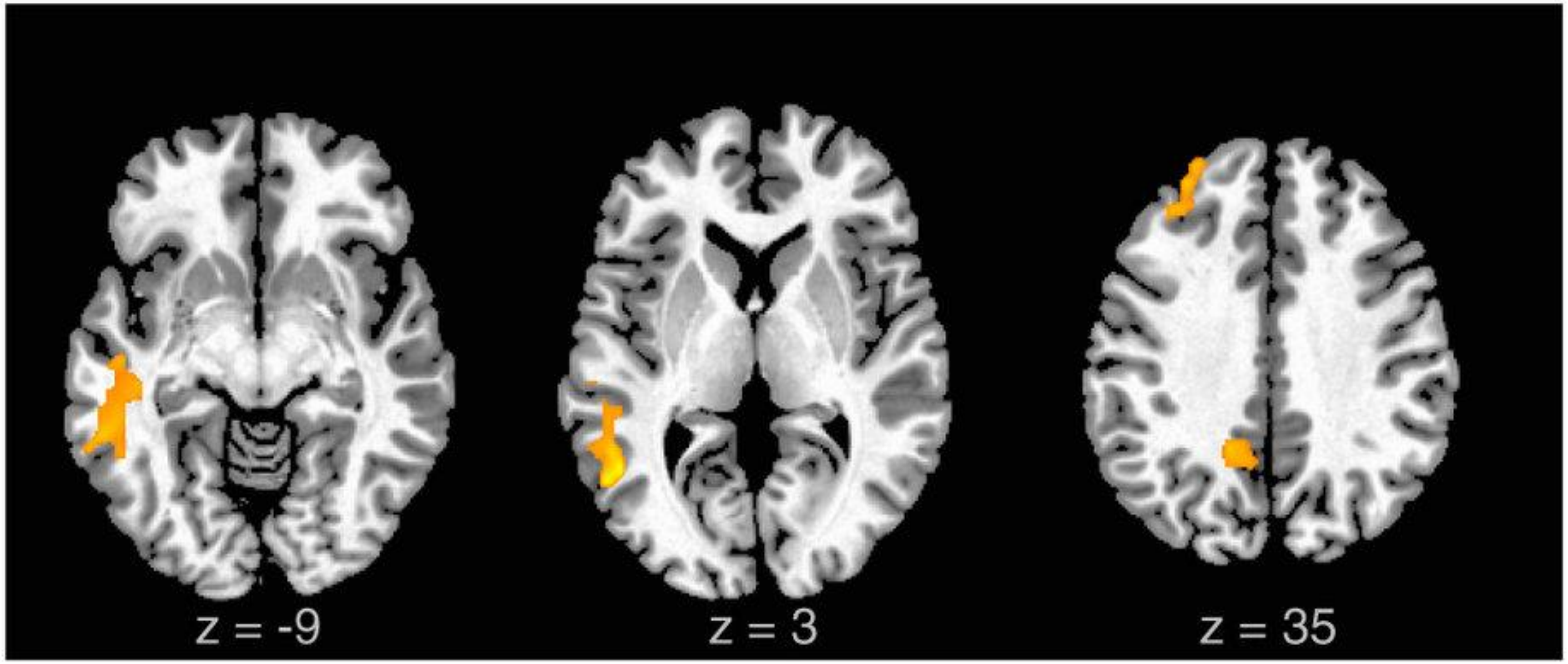
An example is the discovery of a four-planet solar system by amateur astronomers using public astronomical data from the Kepler telescope. The data was made available through the **planethunters.org** website, that asked amateur astronomers to look for a characteristic pattern of light indicating potential planets. An exploratory analysis like this one seeks to make discoveries, but rarely can confirm those discoveries. In the case of the amateur astronomers, follow-up studies and additional data were needed to confirm the existence of the four-planet system.

Exploratory



http://www.sdss.org/

Exploratory



http://www.nature.com/srep/2012/121115/srep00834/full/srep00834.html

Inferential

An inferential data analysis goes beyond an exploratory analysis by quantifying whether an observed pattern will likely hold beyond the data set in hand. Inferential data analyses are the most common statistical analysis in the formal scientific literature.

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

An example is a study of whether air pollution correlates with life expectancy at the state level in the United States. The goal is to identify the strength of the relationship in both the specific data set and to determine whether that relationship will hold in future data. In non-randomized experiments, it is usually only possible to observe whether a relationship between two measurements exists. It is often impossible to determine how or why the relationship exists – it could be due to unmeasured data, relationships, or incomplete modeling.

Inferential

< Previous Abstract | Next Abstract >

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]

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

SDC

Abstract | Author Information

Background: In recent years (2000–2007), ambient levels of fine particulate matter ($PM_{2.5}$) have continued to decline as a result of interventions, but the decline has been at a slower rate than previous years (1980–2000). Whether these more recent and slower declines of $PM_{2.5}$ levels continue to improve life expectancy and whether they benefit all populations equally is unknown.

Methods: We assembled a data set for 545 U.S. counties consisting of yearly county-specific average $PM_{2.5}$, yearly county-specific life expectancy, and several potentially confounding variables measuring socioeconomic status, smoking prevalence, and demographic characteristics for the years 2000 and 2007. We used regression models to estimate the association between reductions in $PM_{2.5}$ and changes in life expectancy for the period from 2000 to 2007.

Results: A decrease of 10 $\mu g/m^3$ in the concentration of $PM_{2.5}$ was associated with an increase in mean life expectancy of 0.35 years (SD = 0.16 years, $P = 0.033$). This association was stronger in more urban and densely populated counties.

Conclusions: Reductions in $PM_{2.5}$ were associated with improvements in life expectancy for the period from 2000 to 2007. Air pollution control in the last decade has continued to have a positive impact on public health.

View Full Text
Article as PDF (663 KB)
Article as EPUB
Print this Article
Add to My Favorites
Export to Citation Manager
Alert Me When Cited
Request Permissions

Related Links

- Articles in PubMed by Andrew W. Correia
- This article in PubMed
- Articles in Google Scholar by Andrew W. Correia
- Other articles in this journal by Andrew W. Correia

Readers Of this Article Also Read

- Estimating the Generation Interval of Influenza A (H1N1) in a Range of Social Se...
- Childhood and Adolescent Exposures and the Risk of Endometriosis
- Early-term Birth (37–38 Weeks) and Mortality in Young Adulthood

http://journals.lww.com/epidem/Abstract/2013/01000/Effect_of_Air_Pollution_Control_on_Life_Expectancy.4.aspx

Predictive

While an inferential data analysis quantifies the relationships among measurements at population-scale, a predictive data analysis uses a subset of measurements (the features) to predict another measurement (the outcome) on a single person or unit.

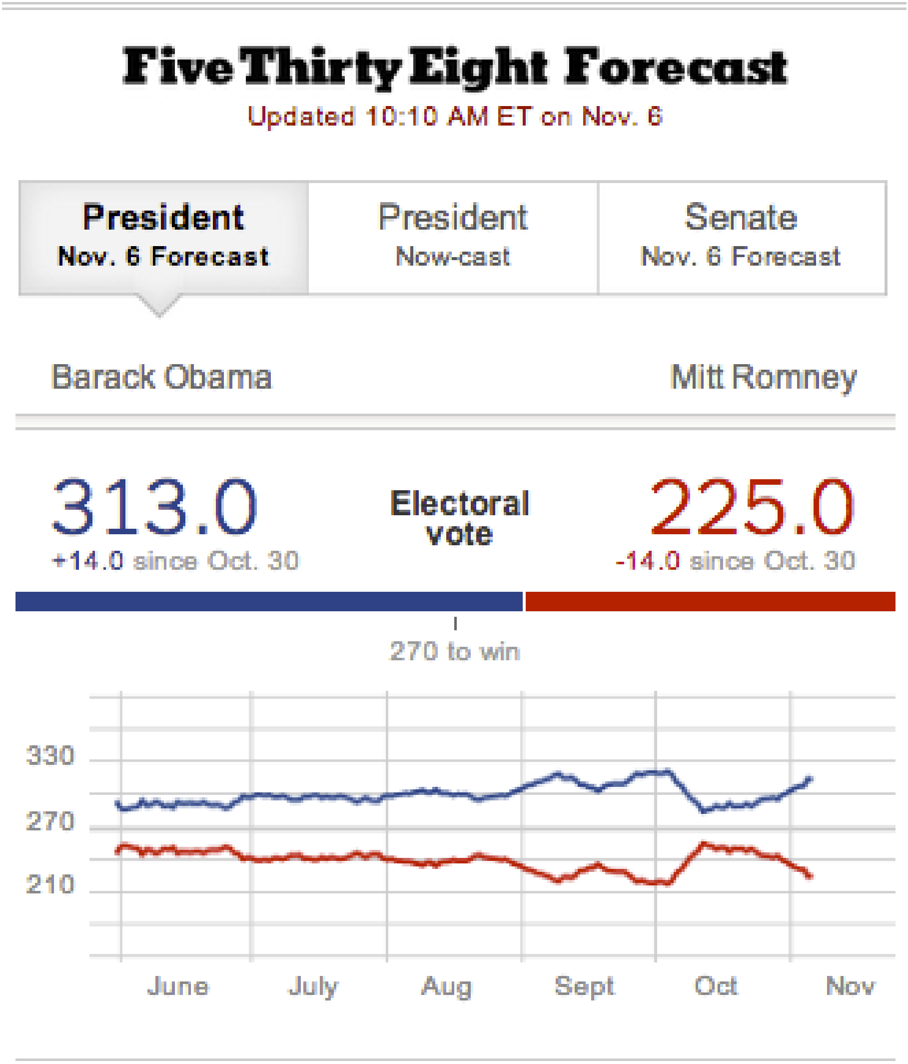
Goal: To use the data on some objects to predict values for another object

- 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

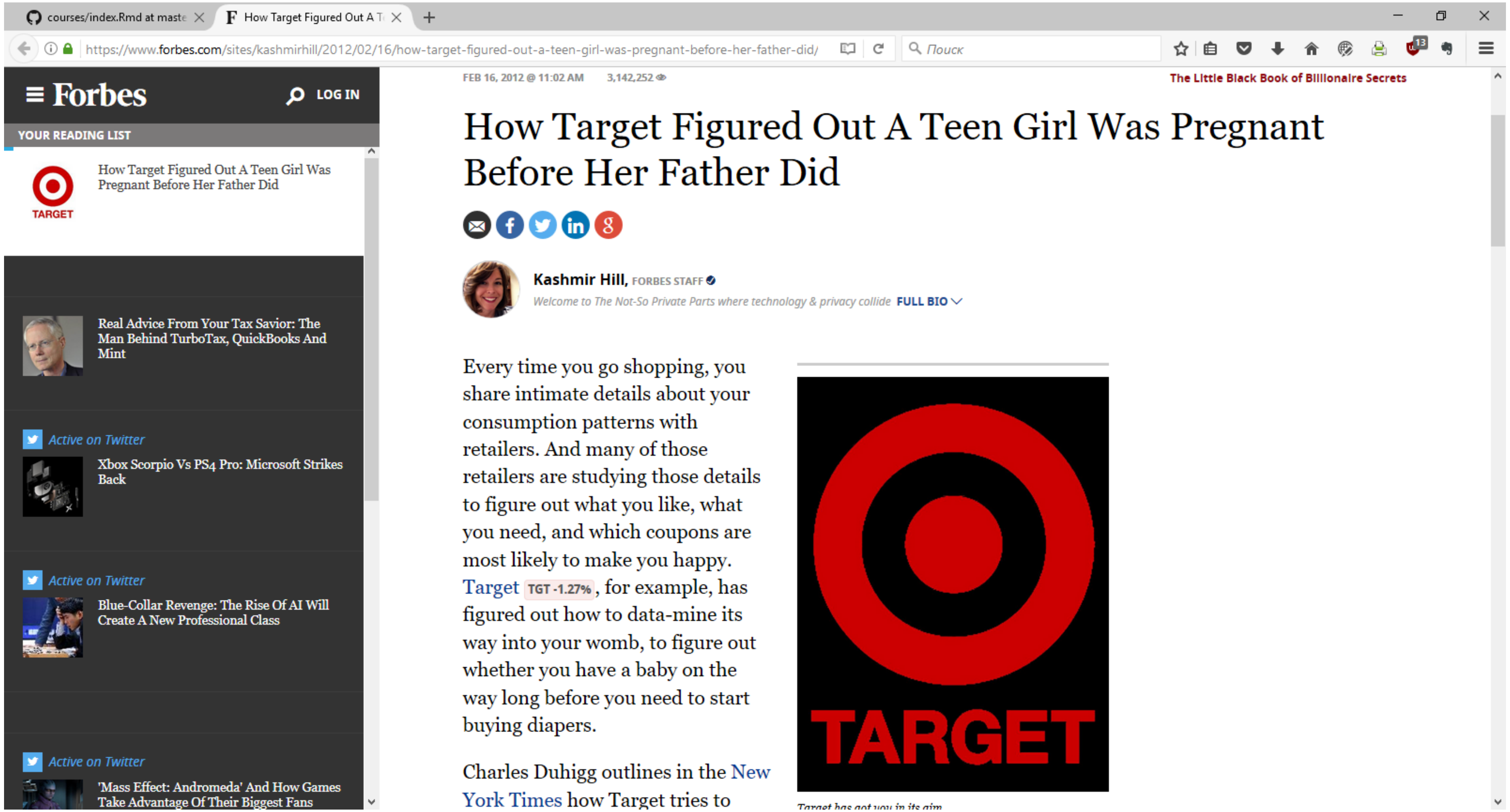
An example is when organizations like FiveThirtyEight.com use polling data to predict how people will vote on election day. In some cases, the set of measurements used to predict the outcome will be intuitive. There is an obvious reason why polling data may be useful for predicting voting behavior. But predictive data analyses only show that you can predict one measurement from another, they don’t necessarily explain why that choice of prediction works.

Predictive



FiveThirtyEight Forecast
Updated 10:10 AM ET on Nov. 6
President
Nov. 6 Forecast
President
Now-cast
Senate
Nov. 6 Forecast
Barack Obama
Mitt Romney
313.0
+14.0 since Oct. 30
Electoral vote
225.0
-14.0 since Oct. 30
270 to win
330
270
210
June
July
Aug
Sept
Oct
Nov

Predictive



https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#2e4133326668

Causal

A causal data analysis seeks to find out what happens to one measurement if you make another measurement change.

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

Causal analysis is plausible reasoning applied to diagnosing observed effect(s), for example, diagnosing cause of biological impairment in a stream. Sir Bradford Hill basically defined the application of causal analysis when he enumerated the elements of causality for associating cigarette smoking with lung cancer.

Mechanistic

Causal data analyses seek to identify average effects between often noisy variables. For example, decades of data show a clear causal relationship between smoking and cancer. If you smoke, it is a sure thing that your risk of cancer will increase. But it is not a sure thing that you will get cancer. The causal effect is real, but it is an effect on your average risk. A mechanistic data analysis seeks to demonstrate that changing one measurement always and exclusively leads to a specific, deterministic behavior in another. The goal is to not only understand that there is an effect, but how that effect operates.

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 (механістичний)

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 being designed using empirical procedures based upon the AASHO Road Test.

Solution: The Mechanistic Empirical Design Procedure

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).

The FHWA plans to develop additional State and regional workshops, training courses, and other educational resources over the next few years, as needed. As State agencies begin to implement the guide, DGIT will arrange

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

Interpolated Values

Temp

Time	68.0	68.5	69.0	69.5	70.0	70.5	71.0	71.5	72.0	72.5	73.0	73.5	74.0	74.5	75.0
0.025	2504.08	2638.15	2707.32	2750.09	2784.91	2851.19	2911.62	2940.67	2961.40	2983.17	3000.06	3006.32	3041.01	3125.78	3026.85
0.05	2507.26	2635.76	2704.79	2746.66	2779.96	2846.35	2907.00	2934.98	2955.07	2976.69	2993.64	2999.35	3034.49	3126.43	3036.68
0.075	2510.83	2633.45	2702.58	2743.62	2775.40	2841.84	2902.75	2929.64	2949.08	2970.51	2987.50	2992.60	3027.98	3126.97	3046.32
0.1	2513.93	2631.34	2700.70	2740.99	2771.27	2837.66	2898.88	2924.66	2943.43	2964.66	2981.67	2986.08	3021.49	3127.39	3055.77
0.125	2515.14	2629.60	2699.17	2738.77	2767.61	2833.83	2895.40	2920.07	2938.14	2959.14	2976.16	2979.83	3015.06	3127.71	3065.02
0.15	2514.31	2628.58	2698.02	2736.99	2764.49	2830.38	2892.31	2915.87	2933.23	2953.97	2970.99	2973.86	3008.70	3127.95	3074.08
0.175	2511.84	2628.88	2697.25	2735.66	2762.00	2827.31	2889.59	2912.08	2928.72	2949.17	2966.17	2968.21	3002.47	3128.11	3082.93
0.2	2508.10	2629.91	2696.87	2734.79	2760.22	2824.68	2887.26	2908.72	2924.62	2944.75	2961.71	2962.89	2996.39	3128.21	3091.57
0.225	2503.37	2631.32	2696.88	2734.37	2759.24	2822.57	2885.29	2905.80	2920.96	2940.73	2957.65	2957.93	2990.50	3128.25	3099.99
0.25	2497.84	2632.93	2697.28	2734.42	2759.10	2821.05	2883.68	2903.34	2917.76	2937.13	2953.97	2953.36	2984.86	3128.24	3108.19
0.275	2491.66	2634.64	2698.05	2734.91	2759.76	2820.23	2882.43	2901.33	2915.02	2933.97	2950.71	2949.20	2979.52	3128.18	3116.14
0.3	2484.92	2636.35	2699.18	2735.85	2761.12	2820.16	2881.55	2899.79	2912.78	2931.26	2947.88	2945.48	2974.53	3128.07	3123.83
0.325	2477.71	2638.00	2700.64	2737.22	2763.09	2820.81	2881.06	2898.72	2911.04	2929.03	2945.47	2942.21	2969.96	3127.90	3131.26
0.35	2470.07	2639.54	2702.41	2739.01	2765.59	2822.11	2880.97	2898.13	2909.82	2927.29	2943.52	2939.43	2965.89	3127.66	3138.38
0.375	2462.06	2640.93	2704.45	2741.19	2768.54	2823.98	2881.29	2898.00	2909.13	2926.05	2942.01	2937.16	2962.39	3127.30	3145.19
0.4	2453.70	2642.15	2706.75	2743.75	2771.89	2826.33	2882.03	2898.34	2908.97	2925.33	2940.96	2935.42	2959.55	3126.79	3151.66
0.425	2445.03	2643.15	2709.26	2746.67	2775.62	2829.13	2883.20	2899.16	2909.34	2925.14	2940.37	2934.25	2957.45	3126.07	3157.75
0.45	2436.07	2643.94	2711.97	2749.92	2779.68	2832.32	2884.78	2900.44	2910.23	2925.48	2940.24	2933.67	2956.16	3125.09	3163.42
0.475	2426.82	2644.48	2714.84	2753.48	2784.06	2835.88	2886.78	2902.19	2911.63	2926.34	2940.57	2933.71	2955.74	3123.85	3168.63
0.5	2417.31	2644.77	2717.84	2757.32	2788.73	2839.78	2889.19	2904.40	2913.52	2927.71	2941.36	2934.34	2956.22	3122.46	3173.31
0.525	2407.54	2644.80	2720.95	2761.44	2793.67	2844.01	2891.99	2907.04	2915.89	2929.57	2942.61	2935.55	2957.60	3121.27	3177.39
0.55	2397.51	2644.56	2724.14	2765.79	2798.87	2848.55	2895.19	2910.11	2918.72	2931.90	2944.30	2937.30	2959.85	3120.88	3180.74
0.575	2387.24	2644.05	2727.39	2770.37	2804.31	2853.38	2898.77	2913.60	2921.99	2934.68	2946.43	2939.57	2962.89	3121.69	3183.21
0.6	2376.71	2643.25	2730.67	2775.14	2809.97	2858.49	2902.71	2917.48	2925.67	2937.89	2948.99	2942.35	2966.66	3123.41	3184.53

What is data?

Definition of Data

Data is a collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted.

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

Population: the set of objects you are interested in.

Variables: A measurement or characteristic of an item.

- **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_babbabaabbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
@HWI-EAS121:4:100:1783:1837#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATATCGTATGCCGTCTTCTGCTTTAATAAAAAAAAA
```

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

What do data look like?

XML

```xml
<?xml version="1.0"?>
<catalog>
   <book id="bk101">
      <author>Gambardella, Matthew</author>
      <title>XML Developer's Guide</title>
      <genre>Computer</genre>
      <price>44.95</price>
      <publish_date>2000-10-01</publish_date>
      <description>An in-depth look at creating applications
      with XML.</description>
   </book>
   <book id="bk102">
      <author>Ralls, Kim</author>
      <title>Midnight Rain</title>
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2000-12-16</publish_date>
      <description>A former architect battles corporate zombies,
      an evil sorceress, and her own childhood to become queen
      of the world.</description>
   </book>
   <book id="bk103">
      <author>Corets, Eva</author>
      <title>Maeve Ascendant</title>
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2000-11-17</publish_date>
      <description>After the collapse of a nanotechnology
      society in England, the young survivors lay the
      foundation for a new society.</description>
   </book>
   <book id="bk104">
      <author>Corets, Eva</author>
      <title>Oberon's Legacy</title>
      <genre>Fantasy</genre>
      <price>5.95</price>
      <publish_date>2001-03-10</publish_date>
      <description>In post-apocalypse England, the mysterious
      agent known only as Oberon helps to create a new life
      for the inhabitants of London. Sequel to Maeve
      Ascendant.</description>
   </book>
```

https://msdn.microsoft.com/en-us/library/ms762271(v=vs.85).aspx

What do data look like?

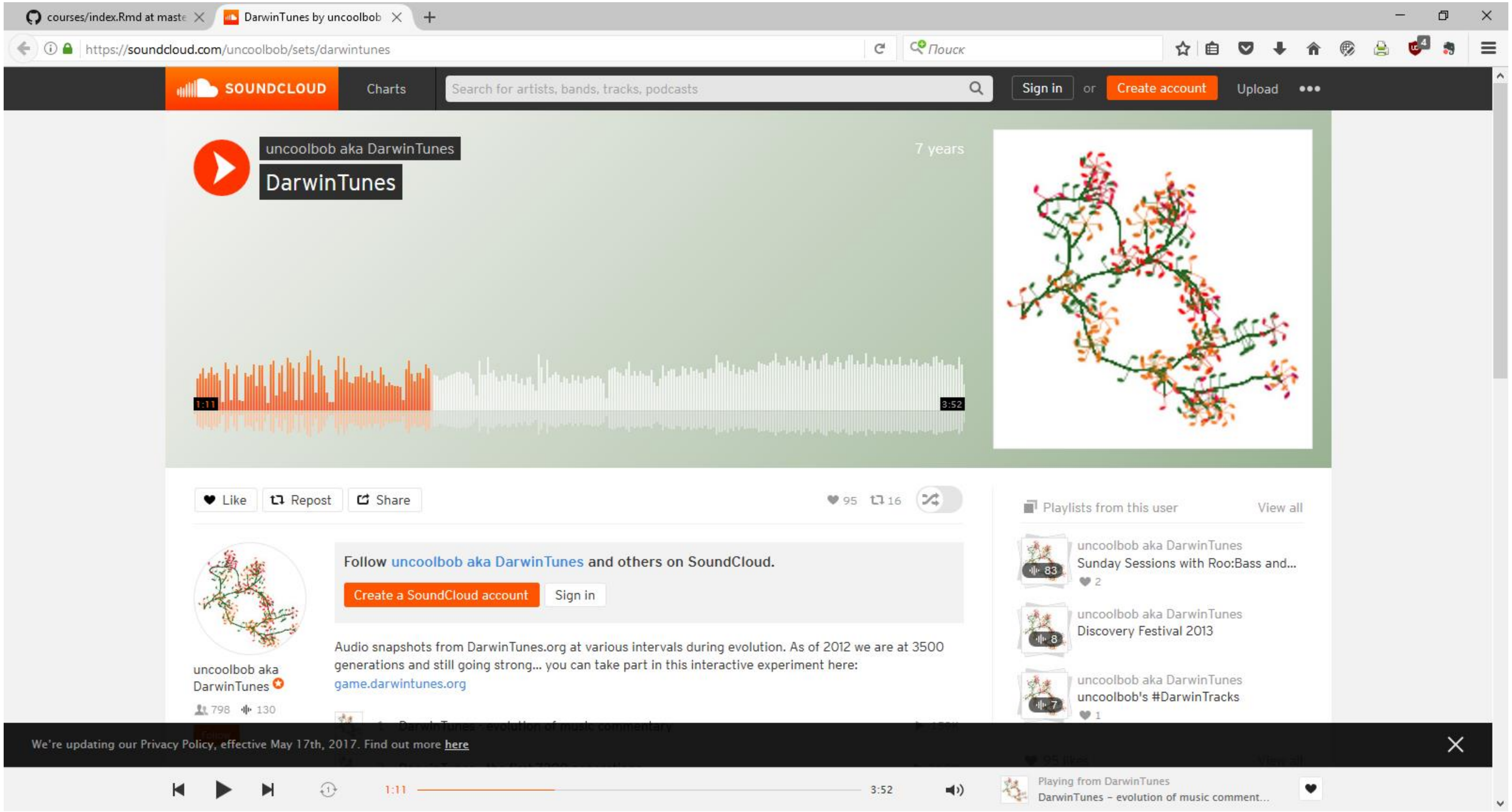
```
# Demographics
First Name: Ellen
Last Name: Ross
Gender: Female
Martial Status: Married
Religious Affiliation: Christian
Ethnicity: Asian
Language Spoken: English
Address: 17 Daws Road, Portland, OR 97006
Telephone: 415-555-1229
Birthday: March 7, 1960


# Guardian
Role: Sister
First Name: Martha
Last Name: Shan
Address: 1357 Amber Drive, Beaverton, OR 97006
Telephone: 816-276-6909


# Provider
Name of Provider: Ashby Medical Center
Address: 1002 Healthcare Dr, Portland, OR 97266
Telephone: 415-555-1200


# Allergies
Allergy Name: Penicillin
Reaction: Hives
Severity: Moderate to severe

Allergy Name: Codeine
Reaction: Shortness of Breath
Severity: Moderate

Allergy Name: Bee Stings
Reaction: Anaphylactic Shock
Severity: Severe


# Immunizations
Date: May 2001
Immunization Name: Influenza virus vaccine, IM
Type: Intramuscular injection
Dose Quantity (value / unit): 50 / mcg
Education/Instructions: Possible flu-like symptoms for three days

Date: April 2000
Immunization Name: Tetanus and diphtheria toxoids, IM
Type: Intramuscular injection
Dose Quantity (value / unit) 50 / mcg
```

http://healthdesignchallenge.com/files/sample-challenge-record.txt

What do data look like?

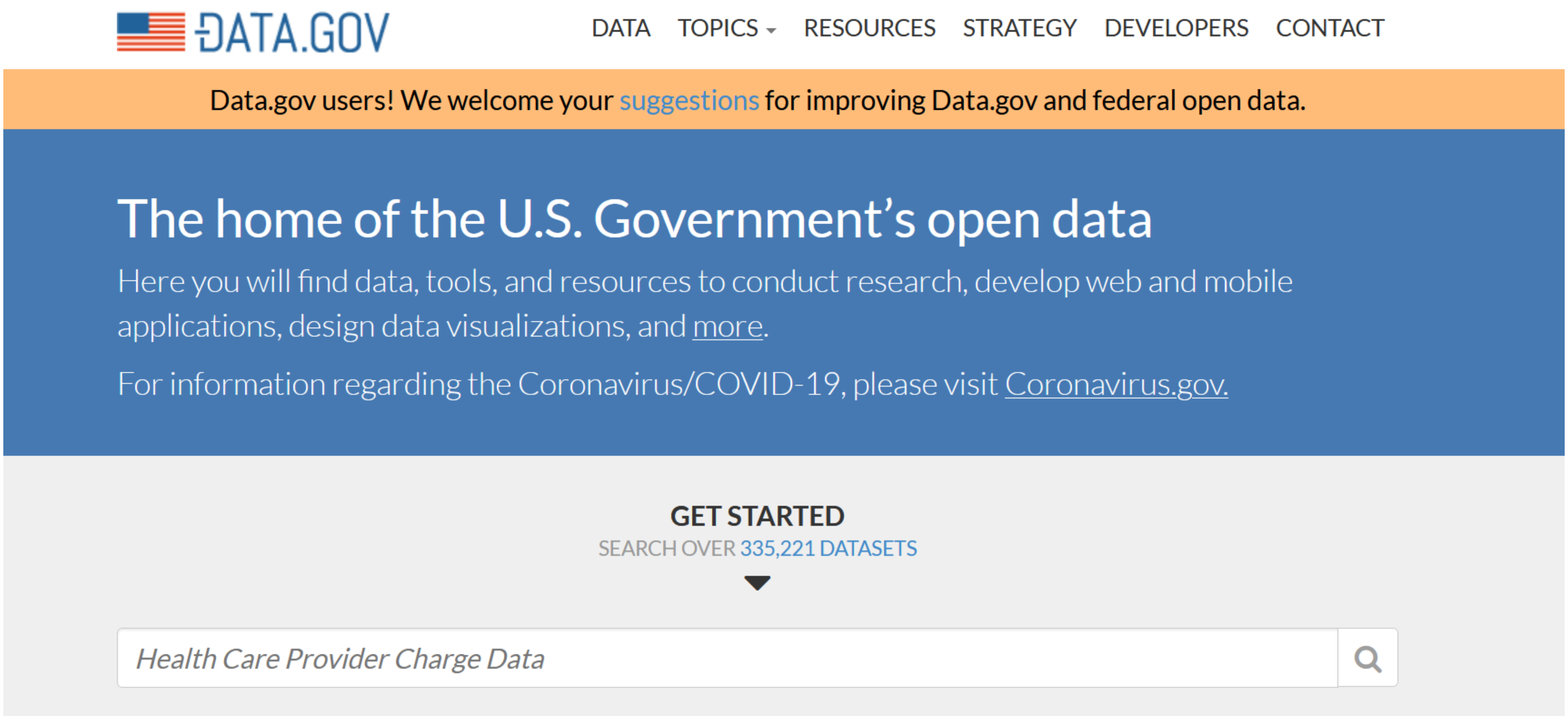
https://www.dreamstime.com/stock-photo-collage-different-cute-cats-many-faces-image46520920

What do data look like?



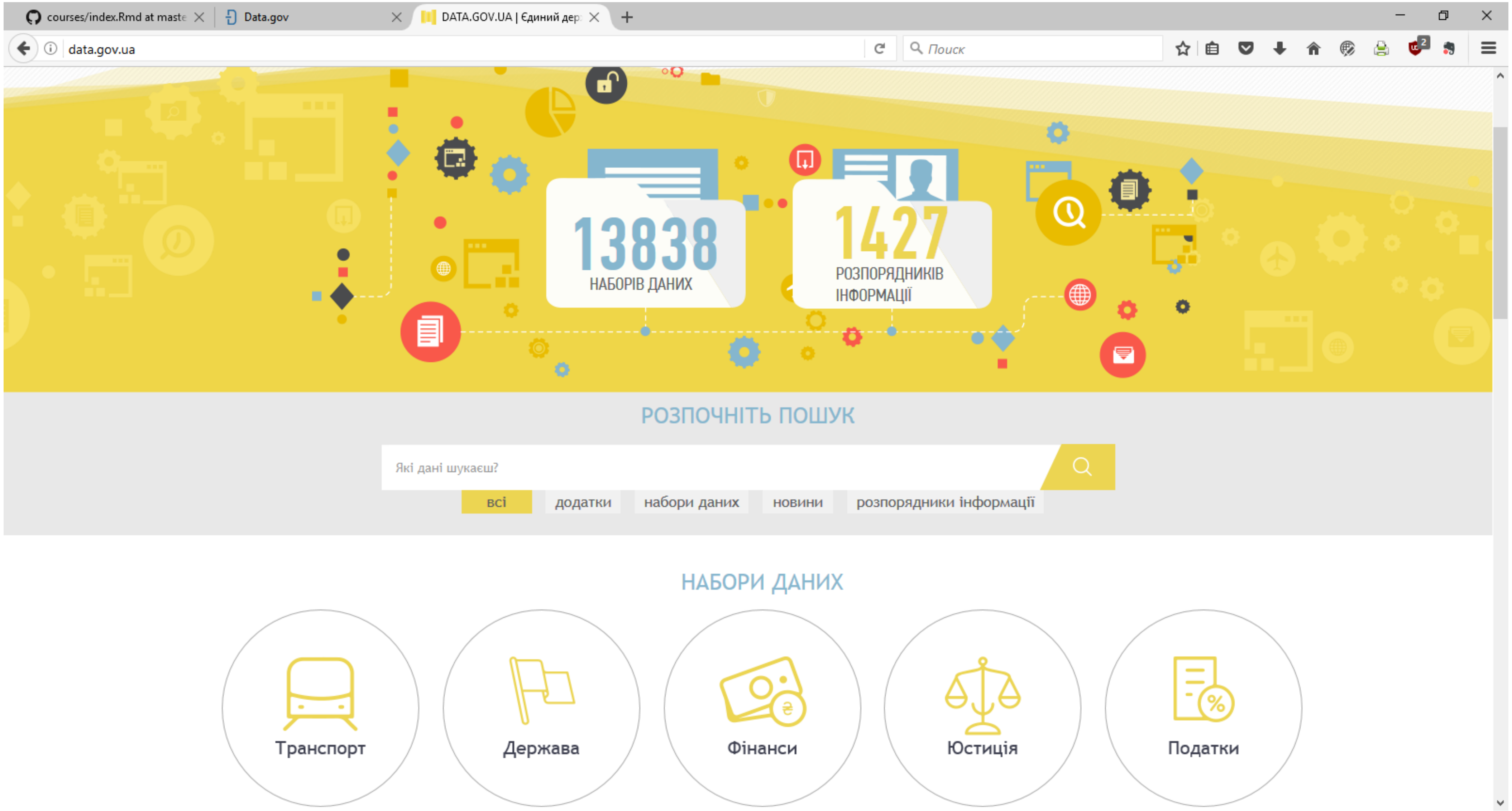
https://soundcloud.com/uncoolbob/sets/darwintunes

What do data look like?



http://data.gov/

What do data look like?



http://data.gov.ua/

What do data look like? Rarely

openaccess/MetObjects.csv at master · metmuseum/openaccess · GitHub - Mozilla Firefox (Не отвечает)

https://media.githubusercontent.com/media/metmuseum/openaccess/master/MetObjects.csv

Поиск


```
Object Number,Is Highlight,Is Public Domain,Object ID,Department,Object Name,Title,Culture,Period,Dynasty,Reign,Portfolio,Artist Role,Artist Prefix,Artist Display Name,Artist Display Bio,Artist Suffix,Artist Alpha Sort,Artist Nationality,Artist Begin Date,Artist End Date,Object Date,Object Begin Date,Object End Date,Medium,Dimensions,Credit Line,Geography Type,City,State,County,Country,Region,Subregion,Locale,Locus,Excavation,River,Classification,Rights and Reproduction,Link Resource,Metadata Date,Repository
1979.486.1,False,False,1,American Decorative Arts,Coin,One-dollar Liberty Head Coin,,,,,,Maker,,James Barton Longacre,"American, Delaware County, Pennsylvania 1794-1869 Philadelphia, Pennsylvania",,"Longacre, James Barton",American,"1794      ","1869      ",1853,1853,1853,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1979",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/1,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
1980.264.5,False,False,2,American Decorative Arts,Coin,Ten-dollar Liberty Head Coin,,,,,,Maker,,Christian Gobrecht,1785-1844,,"Gobrecht, Christian",,"1785      ","1844      ",1901,1901,1901,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1980",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/2,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.9,False,False,3,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/3,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.10,False,False,4,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/4,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.11,False,False,5,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/5,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.12,False,False,6,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/6,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.13,False,False,7,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/7,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.14,False,False,8,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/8,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
67.265.15,False,False,9,American Decorative Arts,Coin,Two-and-a-Half Dollar Coin,,,,,,,,,,,,,,,1909-27,1909,1927,Gold,Diam. 11/16 in. (1.7 cm),"Gift of C. Ruxton Love, Jr., 1967",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/9,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
1979.486.3,False,False,10,American Decorative Arts,Coin,Two-and-a-half-dollar Indian Head Coin,,,,,,Maker,,Bela Lyon Pratt,1867-1917,,"Pratt, Bela Lyon",,"1867      ","1917      ",1912,1912,1912,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1979",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/10,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
1979.486.2,False,False,11,American Decorative Arts,Coin,Two-and-a-half-dollar Liberty Head Coin,,,,,,Maker,,Christian Gobrecht,1785-1844,,"Gobrecht, Christian",,"1785      ","1844      ",1907,1907,1907,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1979",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/11,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
1979.486.7,False,False,12,American Decorative Arts,Coin,Twenty-dollar Liberty Head Coin,,,,,,Maker,,James Barton Longacre,"American, Delaware County, Pennsylvania 1794-1869 Philadelphia, Pennsylvania",,"Longacre, James Barton",American,"1794      ","1869      ",1876,1876,1876,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1979",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/12,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
1979.486.4,False,False,13,American Decorative Arts,Coin,Five-dollar Indian Head Coin,,,,,,Maker,,Bela Lyon Pratt,1867-1917,,"Pratt, Bela Lyon",,"1867      ","1917      ",1910,1910,1910,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1979",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/13,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
1979.486.5,False,False,14,American Decorative Arts,Coin,Five-dollar Liberty Head Coin,,,,,,Maker,,Christian Gobrecht,1785-1844,,"Gobrecht, Christian",,"1785      ","1844      ",1907,1907,1907,Gold,Dimensions unavailable,"Gift of Heinz L. Stoppelmann, 1979",,,,,,,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/14,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.49,False,False,15,American Paintings and Sculpture,Coin,"Coin, 1/2 Real",,,,,,,,,,,,,,,1665-1700,1665,1700,Silver,Diam. 1/2 in. (1.3 cm),"Gift of Mrs. Russell Sage, 1916",Made in,,,,Mexico,,,,,,,Silver,,http://www.metmuseum.org/art/collection/search/15,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.27,False,False,16,American Paintings and Sculpture,Peso,"Coin, 1/4 Peso",,,,,,Artist,,Mexican Artist,,,"Mexican, Artist",,,,1800-1900,1800,1900,Bronze or copper,Diam. 1 1/8 in. (2.9 cm),"Gift of Mrs. Russell Sage, 1916",Made in,,,,Mexico,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/16,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.28,False,False,17,American Paintings and Sculpture,Peso,"Coin, 1/4 Peso",,,,,,Artist,,Mexican Artist,,,"Mexican, Artist",,,,1867,1867,1867,Bronze or copper,Diam. 1 1/8 in. (2.9 cm),"Gift of Mrs. Russell Sage, 1916",Made in,,,,Mexico,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/17,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.29,False,False,18,American Paintings and Sculpture,Peso,"Coin, 1/4 Peso",,,,,,Artist,,Mexican Artist,,,"Mexican, Artist",,,,1860,1860,1860,Bronze or copper,Diam. 1 1/8 in. (2.9 cm),"Gift of Mrs. Russell Sage, 1916",Made in,,,,Mexico,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/18,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.30,False,False,19,American Paintings and Sculpture,Peso,"Coin, 1/4 Peso",,,,,,,,,,,,,,,1859,1859,1859,Bronze or copper,Diam. 1 1/8 in. (2.9 cm),"Gift of Mrs. Russell Sage, 1916",Made in,,,,Mexico,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/19,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.31,False,False,20,American Paintings and Sculpture,Peso,"Coin, 1/4 Peso",,,,,,,,,,,,,,,1860,1860,1860,Bronze or copper,Diam. 1 1/8 in. (2.9 cm),"Gift of Mrs. Russell Sage, 1916",Made in,,,,Mexico,,,,,,,Metal,,http://www.metmuseum.org/art/collection/search/20,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
16.74.32,False,False,21,American Paintings and Sculpture,Peso,"Coin, 1/4 Peso",,,,,,,,,,,,,,,1859,1859,1859,Bronze or copper,Diam. 1 1/8 in. (2.9 cm),"Gift of Mrs. Russell Sage, 1916",Made
w.metmuseum.org/art/collection/search/21,4/3/2017 8:00:08 AM,"Metropolitan Museum of Art, New York, NY"
```


Передача данных с media.githubusercontent.com...

https://github.com/metmuseum/openaccess, https://www.kaggle.com/metmuseum/the-metropolitan-museum-of-art-open-access

What do data look like? Rarely

MetObjects - Excel

Вход

Файл Главная Вставка Разметка страницы Формулы Данные Рецензирование Вид Разработчик Надстройки Команда Что вы хотите сделать? Поделиться

A1 Object Number

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Object Num	Is Highligh	Is Public D	Object ID	Department	Object Name	Title	Culture	Period	Dynasty	Reign	Portfolio	Artist Role	Artist Pref	Artist Disp	Artist Disp	Artist Suff	Artist Alpl	Artist Nati	Artist Beg	Artist End	Object Da	Object I
2	1979.486.1	False	False	1	American Deci	Coin	One-dollar Liberty Head Coin						Maker		James Bar	American, Delaware		Longacre,	American	1794	1869	1853	18
3	1980.264.5	False	False	2	American Deci	Coin	Ten-dollar Liberty Head Coin						Maker		Christian (	1785–1844		Gobrecht, Christian		1785	1844	1901	19
4	67.265.9	False	False	3	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
5	67.265.10	False	False	4	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
6	67.265.11	False	False	5	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
7	67.265.12	False	False	6	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
8	67.265.13	False	False	7	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
9	67.265.14	False	False	8	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
10	67.265.15	False	False	9	American Deci	Coin	Two-and-a-Half Dollar Coin															1909–27	19
11	1979.486.3	False	False	10	American Deci	Coin	Two-and-a-half-dollar Indian Head Coin						Maker		Bela Lyon	1867–1917		Pratt, Bela Lyon		1867	1917	1912	19
12	1979.486.2	False	False	11	American Deci	Coin	Two-and-a-half-dollar Liberty Head Coin						Maker		Christian (	1785–1844		Gobrecht, Christian		1785	1844	1907	19
13	1979.486.7	False	False	12	American Deci	Coin	Twenty-dollar Liberty Head Coin						Maker		James Bar	American, Delaware		Longacre,	American	1794	1869	1876	18
14	1979.486.4	False	False	13	American Deci	Coin	Five-dollar Indian Head Coin						Maker		Bela Lyon	1867–1917		Pratt, Bela Lyon		1867	1917	1910	19
15	1979.486.5	False	False	14	American Deci	Coin	Five-dollar Liberty Head Coin						Maker		Christian (	1785–1844		Gobrecht, Christian		1785	1844	1907	19
16	16.74.49	False	False	15	American Pair	Coin	Coin, 1/2 Real															1665–170(	16
17	16.74.27	False	False	16	American Pair	Peso	Coin, 1/4 Peso						Artist		Mexican Artist			Mexican, Artist				1800–190(	18
18	16.74.28	False	False	17	American Pair	Peso	Coin, 1/4 Peso						Artist		Mexican Artist			Mexican, Artist				1867	18
19	16.74.29	False	False	18	American Pair	Peso	Coin, 1/4 Peso						Artist		Mexican Artist			Mexican, Artist				1860	18
20	16.74.30	False	False	19	American Pair	Peso	Coin, 1/4 Peso															1859	18
21	16.74.31	False	False	20	American Pair	Peso	Coin, 1/4 Peso															1860	18
22	16.74.32	False	False	21	American Pair	Peso	Coin, 1/4 Peso															1859	18
23	16.74.43	False	False	22	American Pair	Coin	Coin, 1/4 Real															1881	18
24	16.74.44	False	False	23	American Pair	Coin	Coin, 1/4 Real															1878	18
25	16.74.33	False	False	24	American Pair	Centavos	Coin, 10 Centavos															1860–70	18
26	16.74.34	False	False	25	American Pair	Centavos	Coin, 10 Centavos															1860–70	18
27	16.74.35	False	False	26	American Pair	Centavos	Coin, 10 Centavos															1860–70	18
28	16.74.36	False	False	27	American Pair	Centavos	Coin, 10 Centavos															1860–70	18
29	16.74.38	False	False	28	American Pair	Centavos	Coin, 10 Centavos															1860–70	18
30	16.74.39	False	False	29	American Pair	Centavos	Coin, 10 Centavos															1860–70	18
31	16.74.37	False	False	30	American Pair	Centavos	Coin, 10 Centavos															1885	18
32	16.74.40	False	False	31	American Pair	Centavos	Coin, 10 Centavos															1885	18
33	09.09.2015	False	False	32	American Pair	Pesos	Coin, 20 Pesos						Artist		Mexican Artist			Mexican, Artist				1866	18
34	64.62	False	True	33	American Deci	Bust	Bust of Abraham	American					Maker		James Gill	American	1861–ca. 1	Gillinder &	American	1861	1930	1876	18

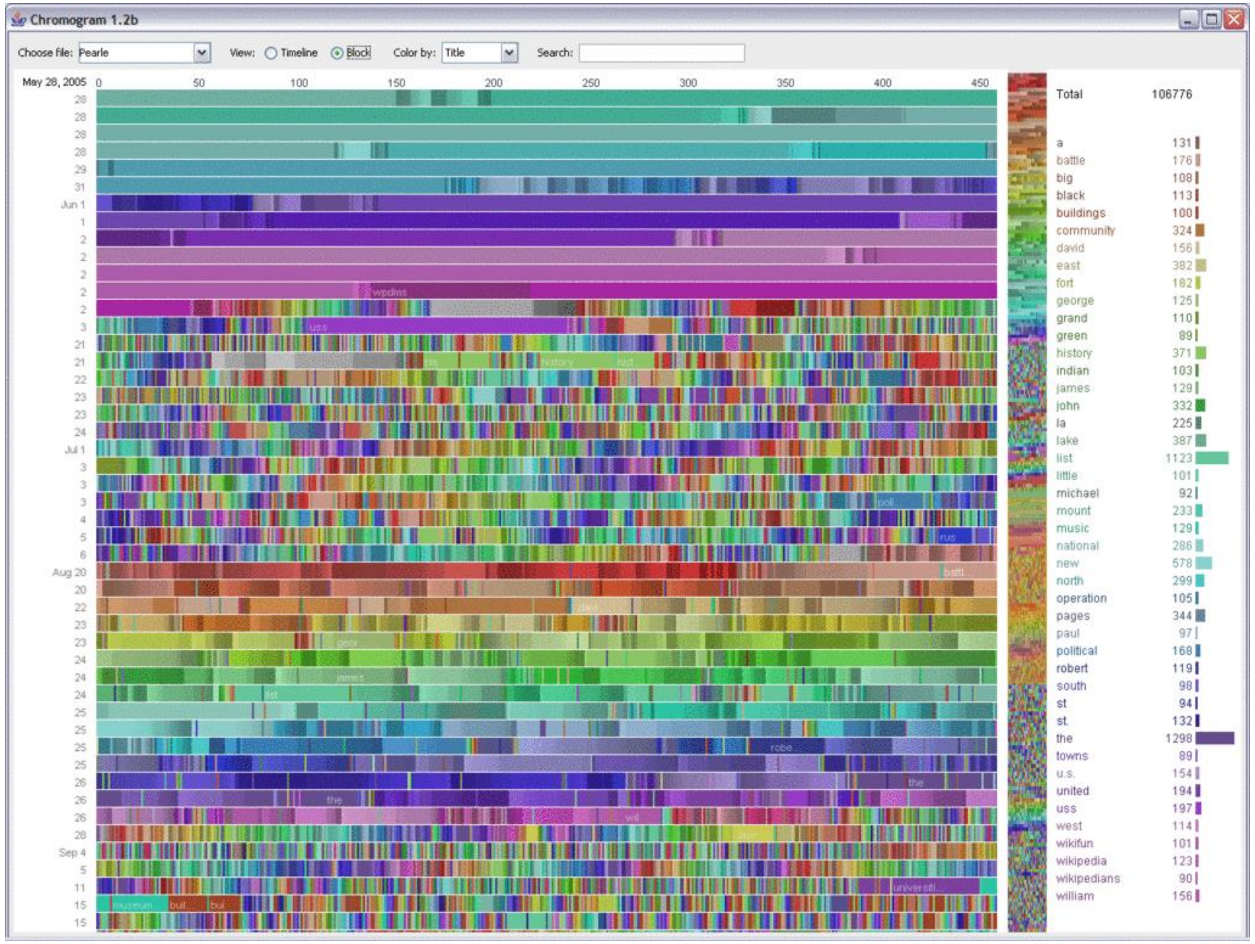
MetObjects

Готово 100 %

https://github.com/metmuseum/openaccess, https://www.kaggle.com/metmuseum/the-metropolitan-museum-of-art-open-access

The data is the second most important thing

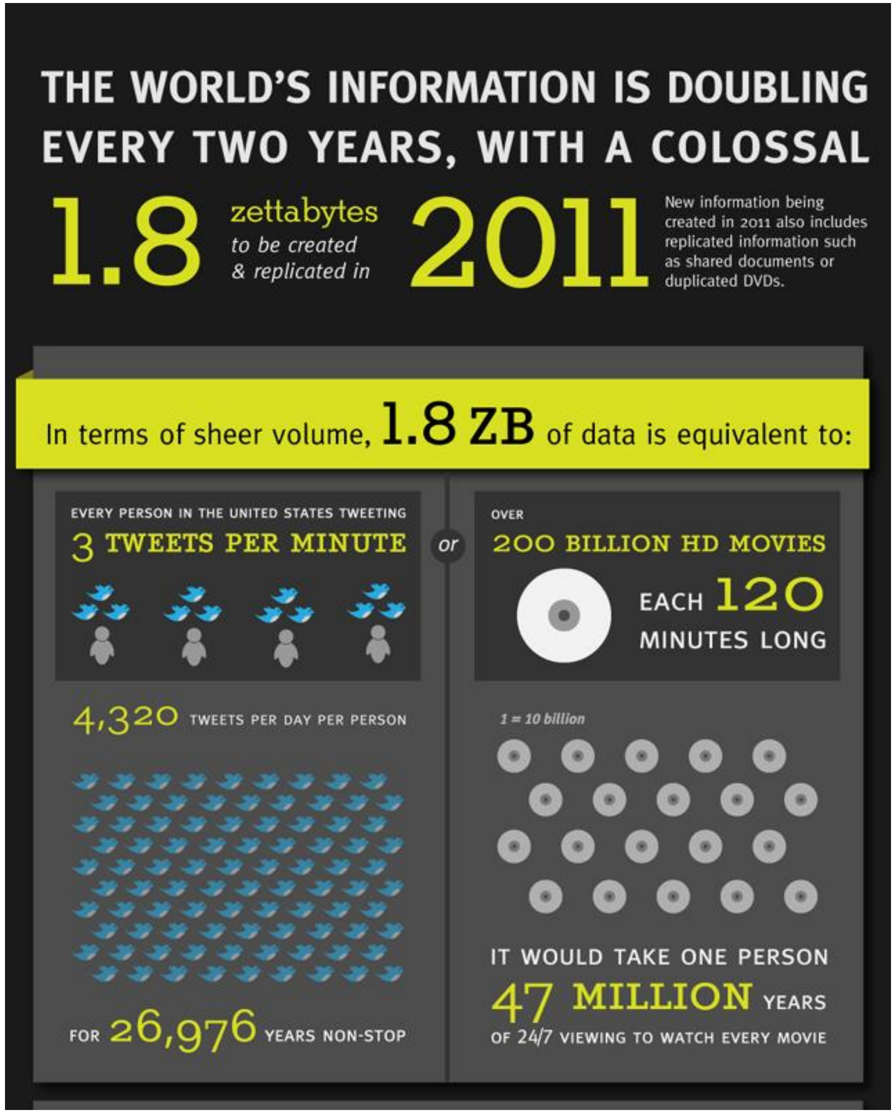
- The most important thing in data science is the question
- The second most important is the data
- Often the data will limit or enable the questions
- **But having data can't save you if you don't have a question**

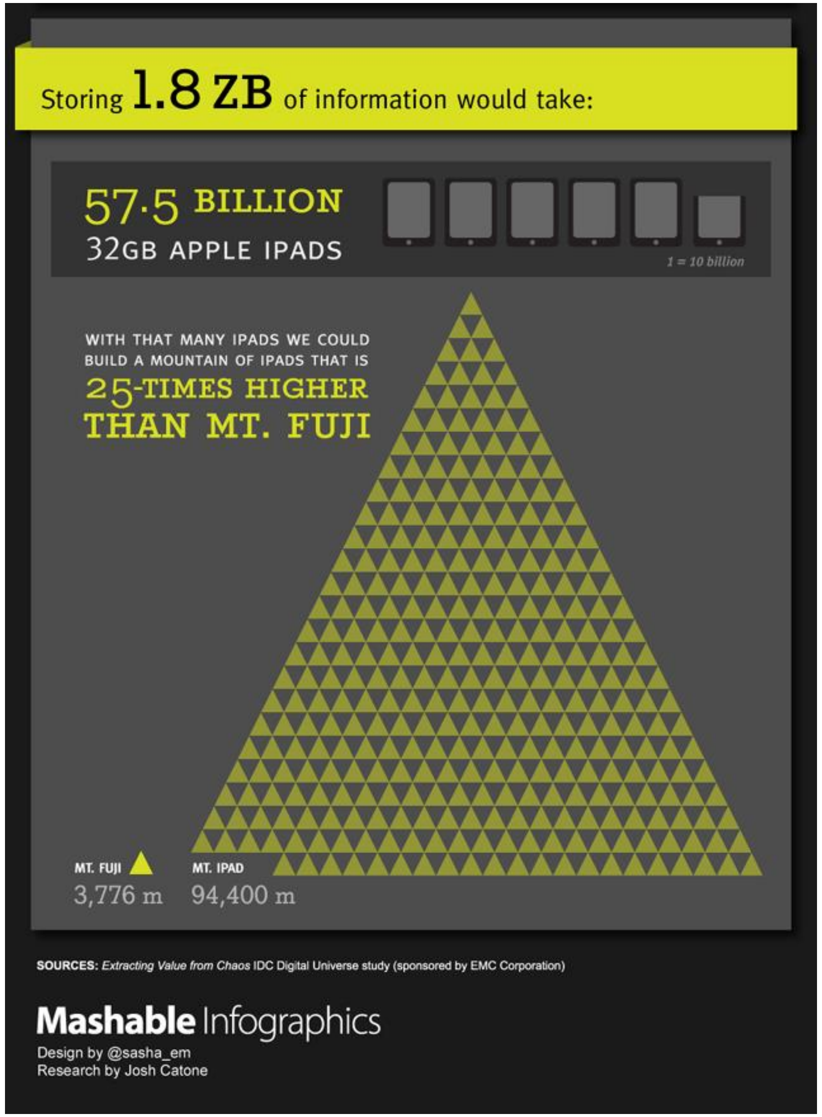


What about Big Data?

Auther: Fernanda B. Viégas - User activity on Wikipedia, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=10090013

How much is there?

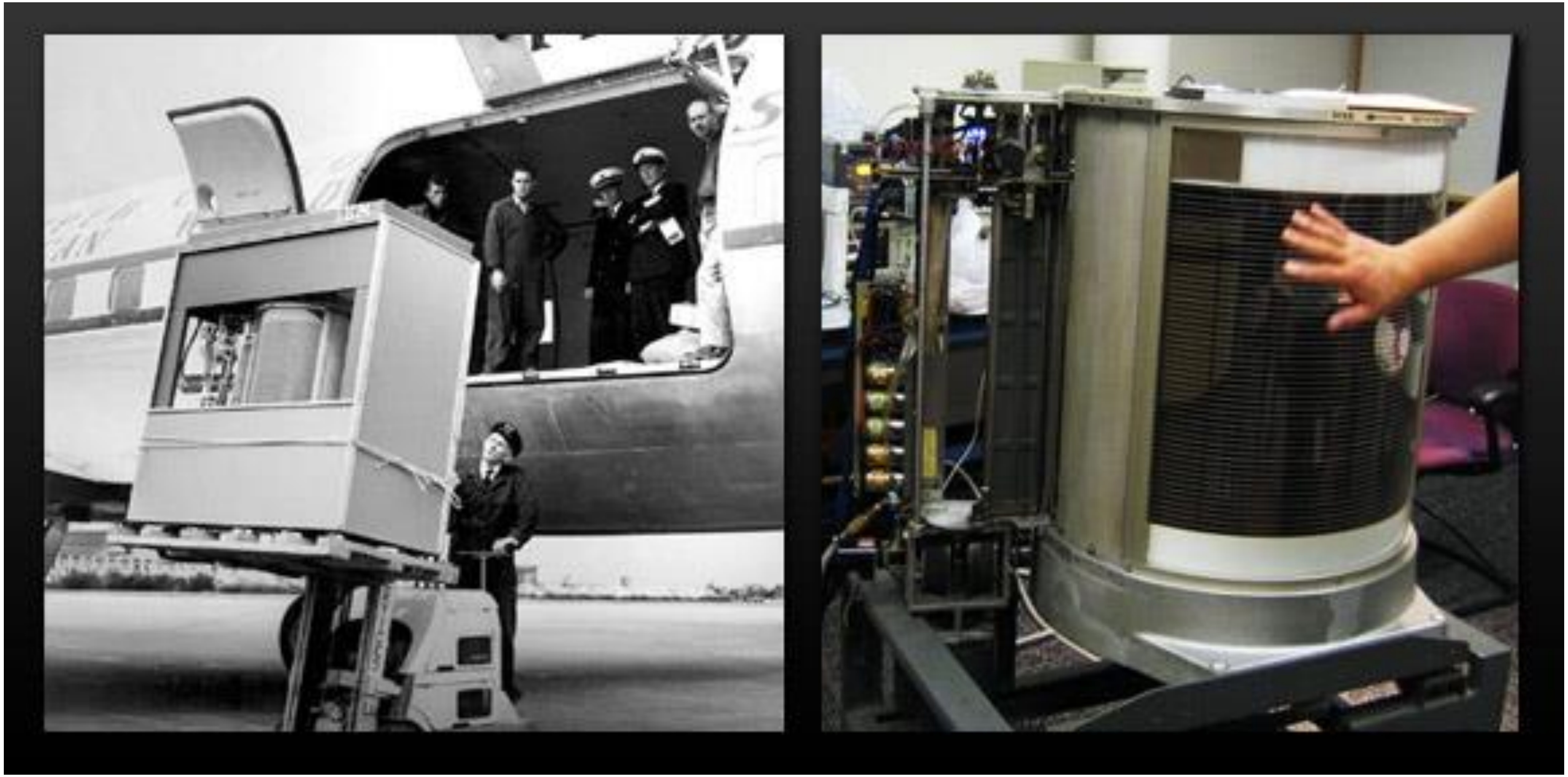




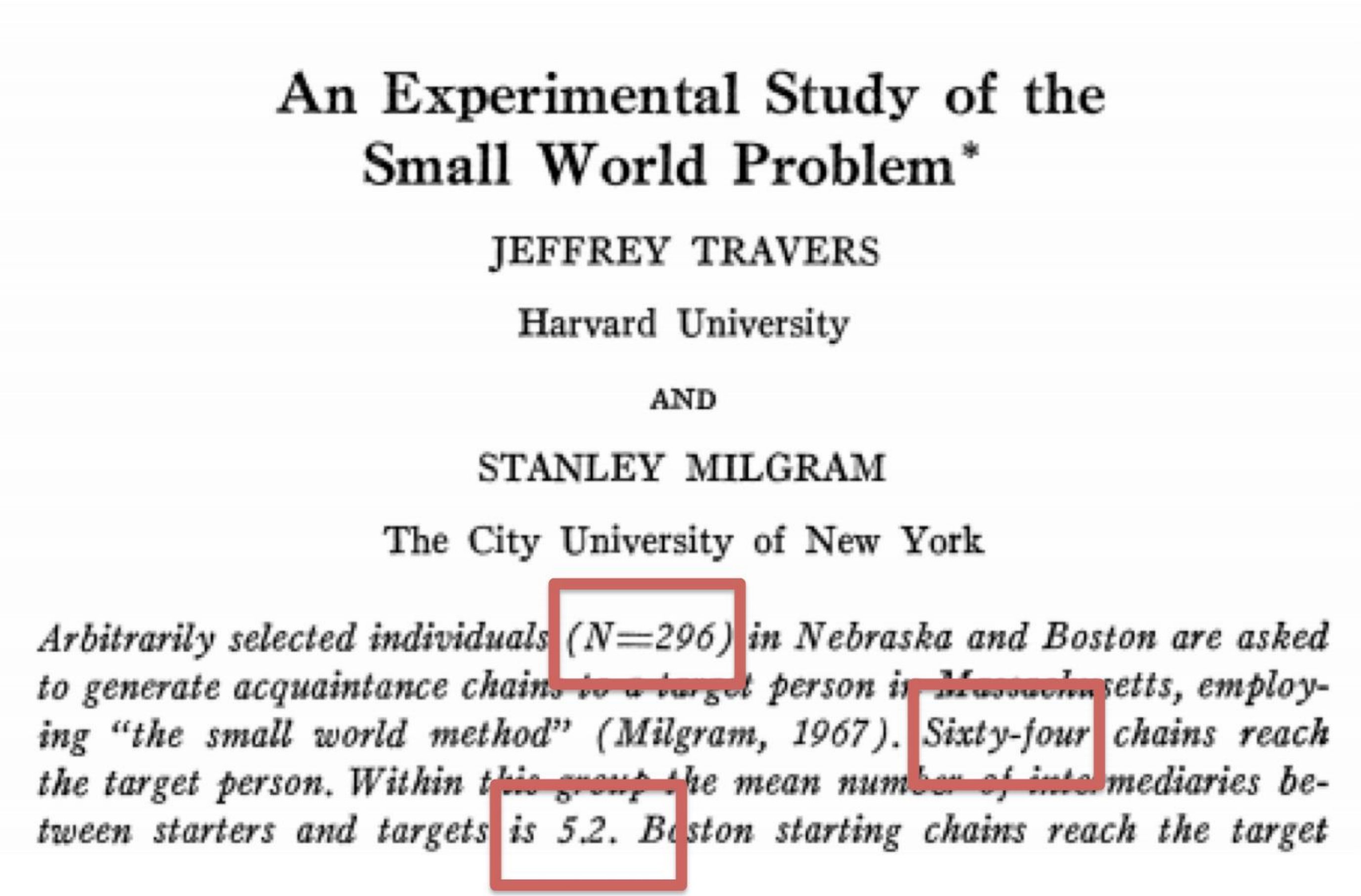
http://mashable.com/2011/06/28/data-infographic/

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

http://www.jstor.org/discover/10.2307/2786545?uid=3739704&uid=2&uid=4&uid=3739256&sid=21101674727517

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

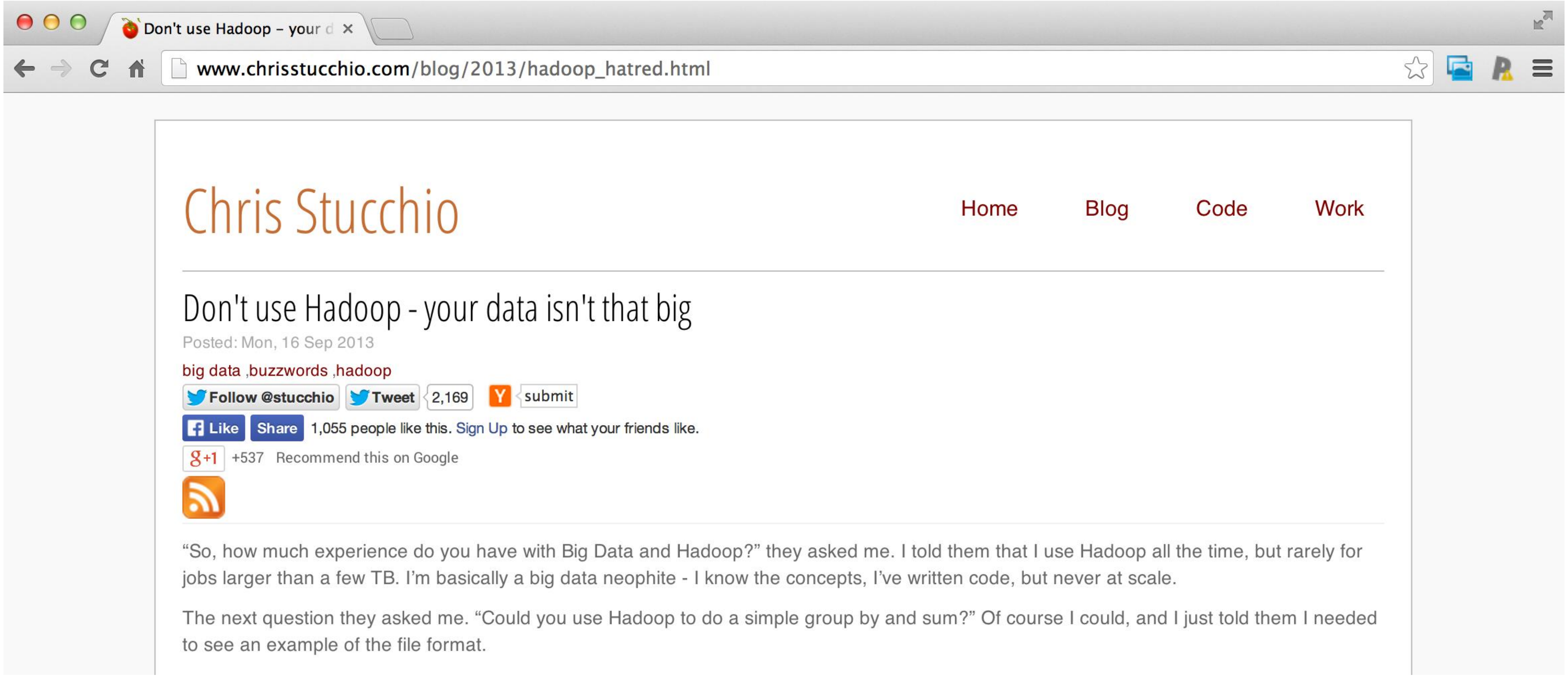
Jure Leskovec, Eric Horvitz

(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 report 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.

http://arxiv.org/abs/0803.0939

Big or small - you need the right data



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

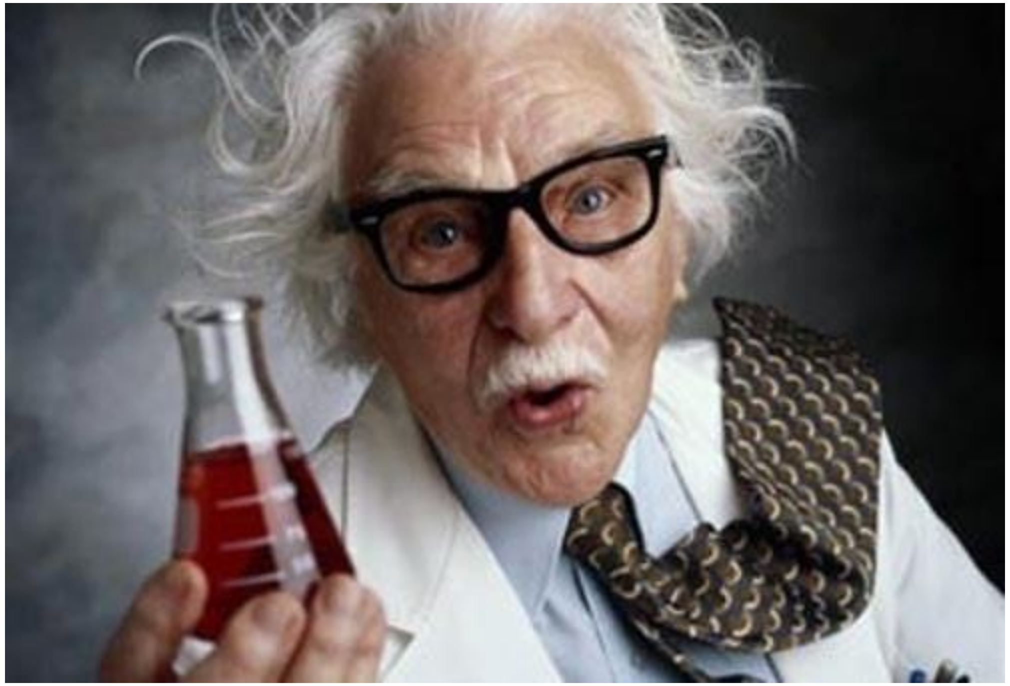
Big or small - you need the right data

“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...”

John Tukey

“...no matter how big the data are.”

Jeff Leek



Experimental Design

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

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†

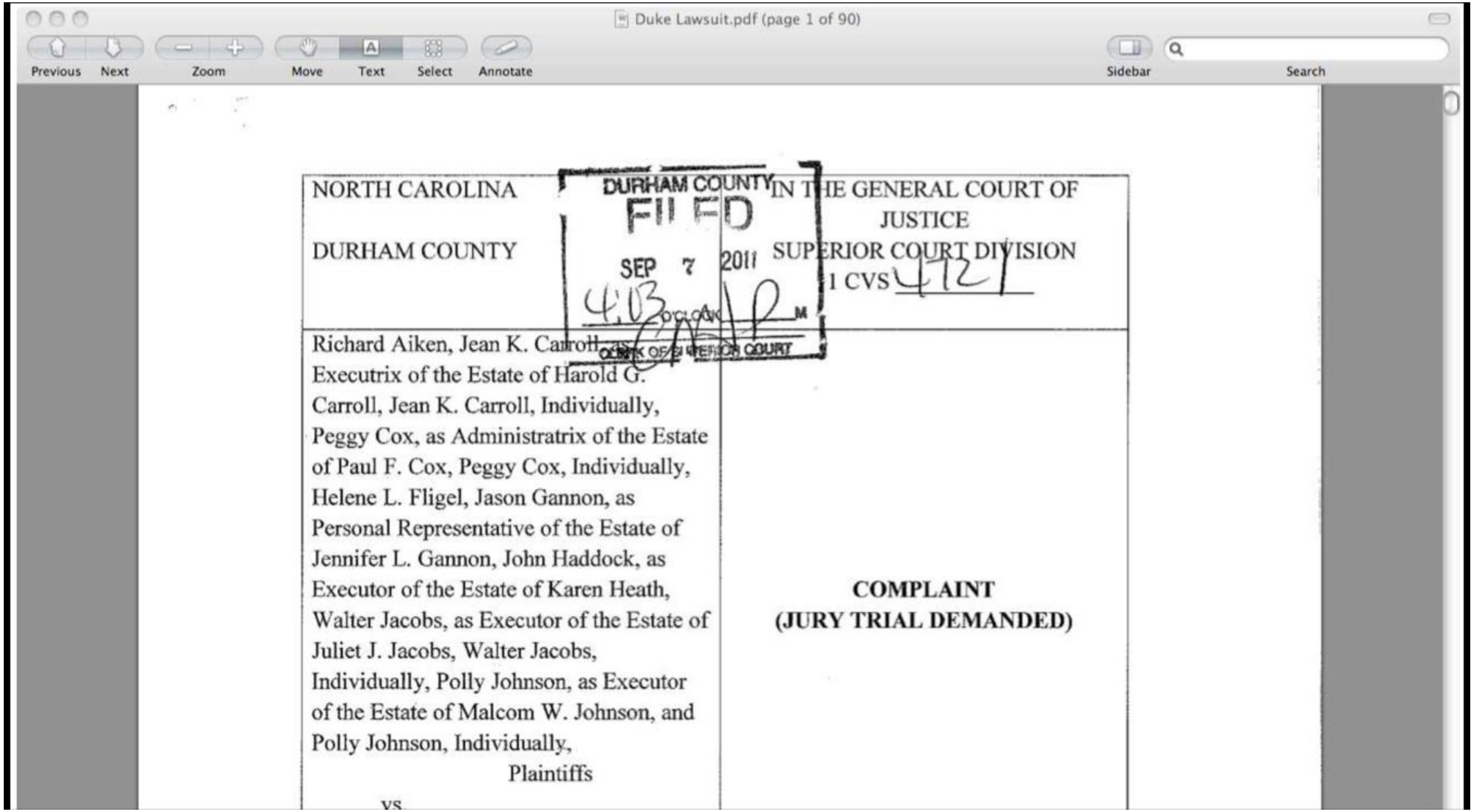
U.T. M.D. Anderson Cancer Center

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



NORTH CAROLINA DURHAM COUNTY	IN THE GENERAL COURT OF JUSTICE SUPERIOR COURT DIVISION 1 CVS 4727
Richard Aiken, Jean K. Carroll, as Executrix of the Estate of Harold G. Carroll, Jean K. Carroll, Individually, Peggy Cox, as Administratrix of the Estate of Paul F. Cox, Peggy Cox, Individually, Helene L. Fligel, Jason Gannon, as Personal Representative of the Estate of Jennifer L. Gannon, John Haddock, as Executor of the Estate of Karen Heath, Walter Jacobs, as Executor of the Estate of Juliet J. Jacobs, Walter Jacobs, Individually, Polly Johnson, as Executor of the Estate of Malcom W. Johnson, and Polly Johnson, Individually, Plaintiffs vs.	**COMPLAINT** **(JURY TRIAL DEMANDED)**

DURHAM COUNTY
FILED
SEP 7 2011
4:03 O'CLOCK P M
CLERK OF SUPERIOR COURT

Know and care about the analysis plan!

Abstract Formula display: ☑ MathJax

Background

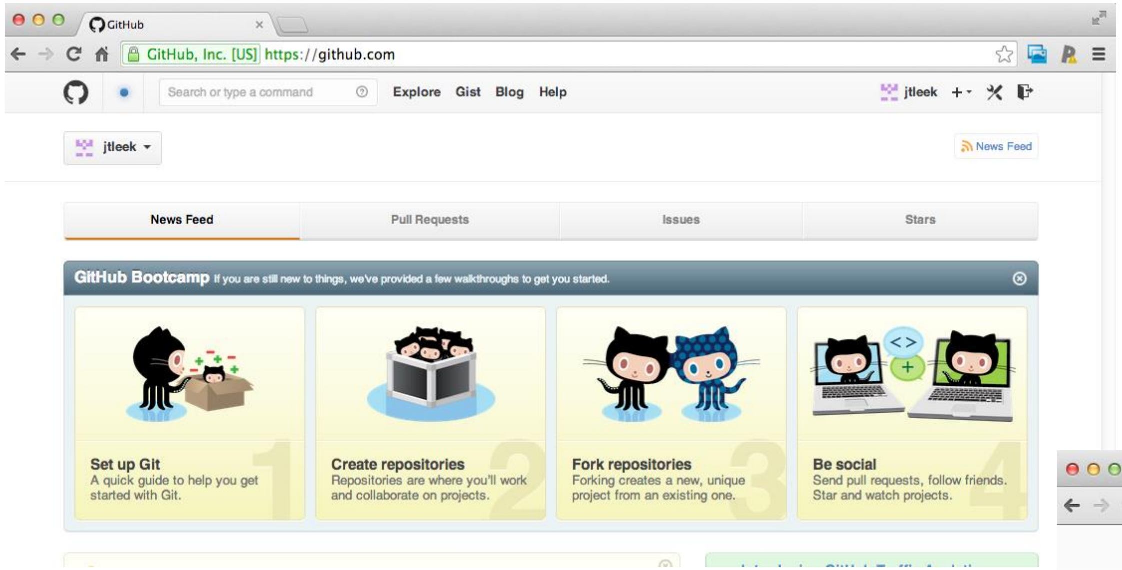
Many groups, including our own, have proposed the use of DNA methylation profiles as biomarkers for various disease states. While much research has been done identifying DNA methylation signatures in cancer vs. normal etc., we still lack sufficient knowledge of the role that differential methylation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

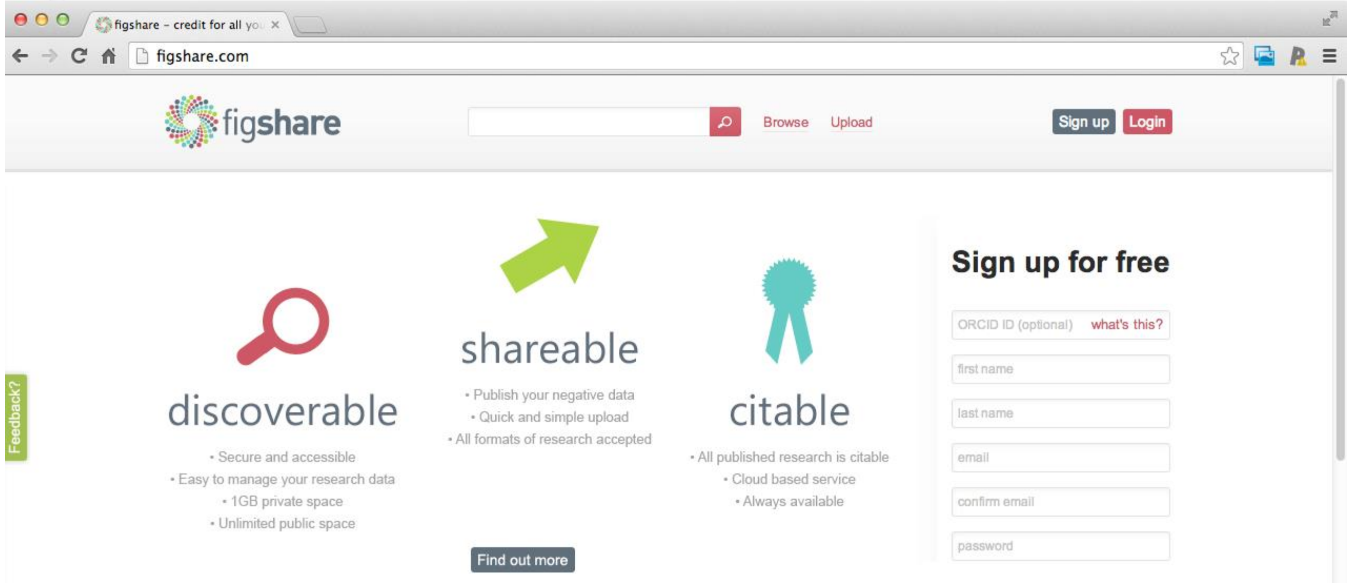
http://nsaunders.wordpress.com/2012/07/23/we-really-dont-care-what-statistical-method-you-used/

Have a plan for data and code sharing

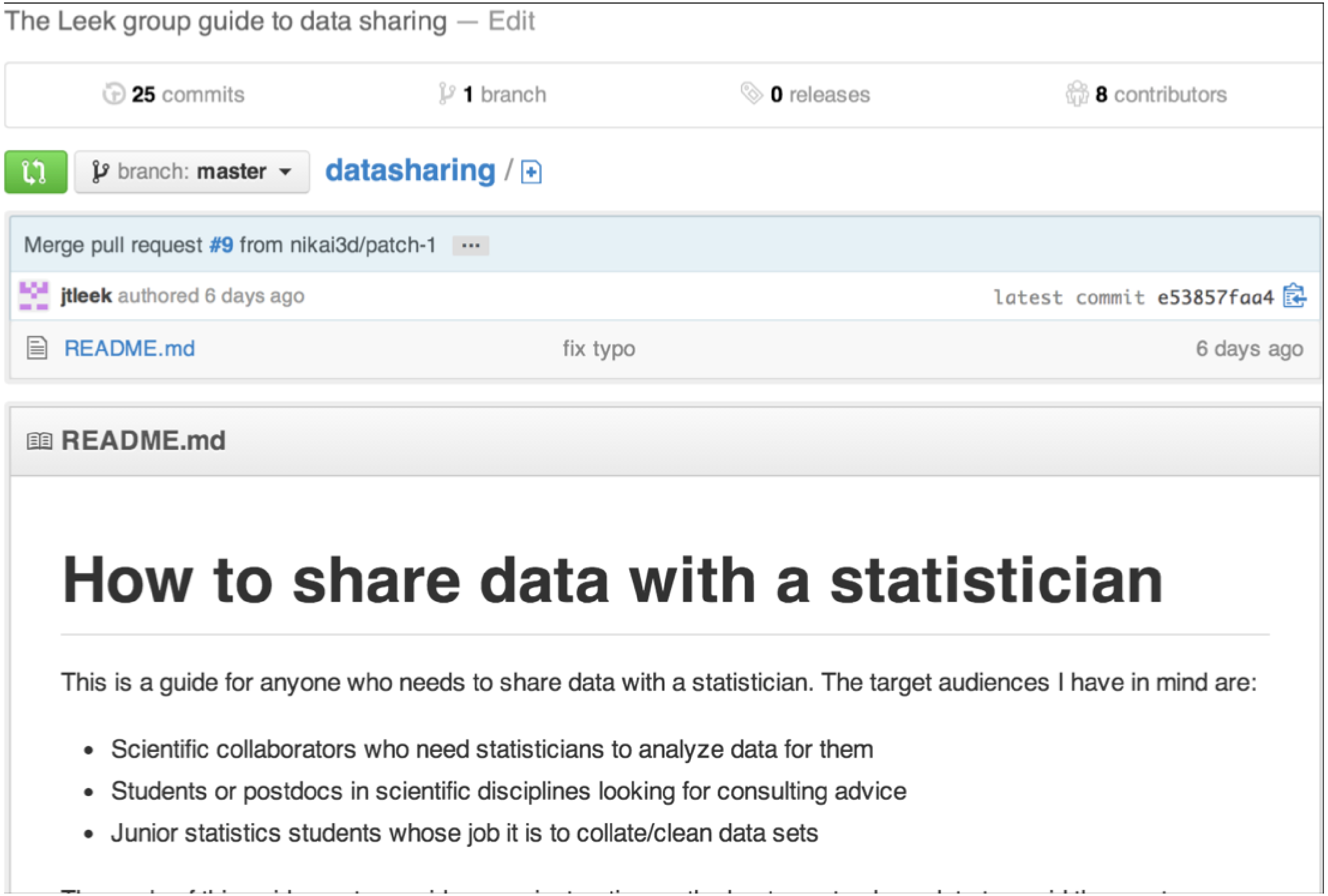


https://github.com/

http://figshare.com/

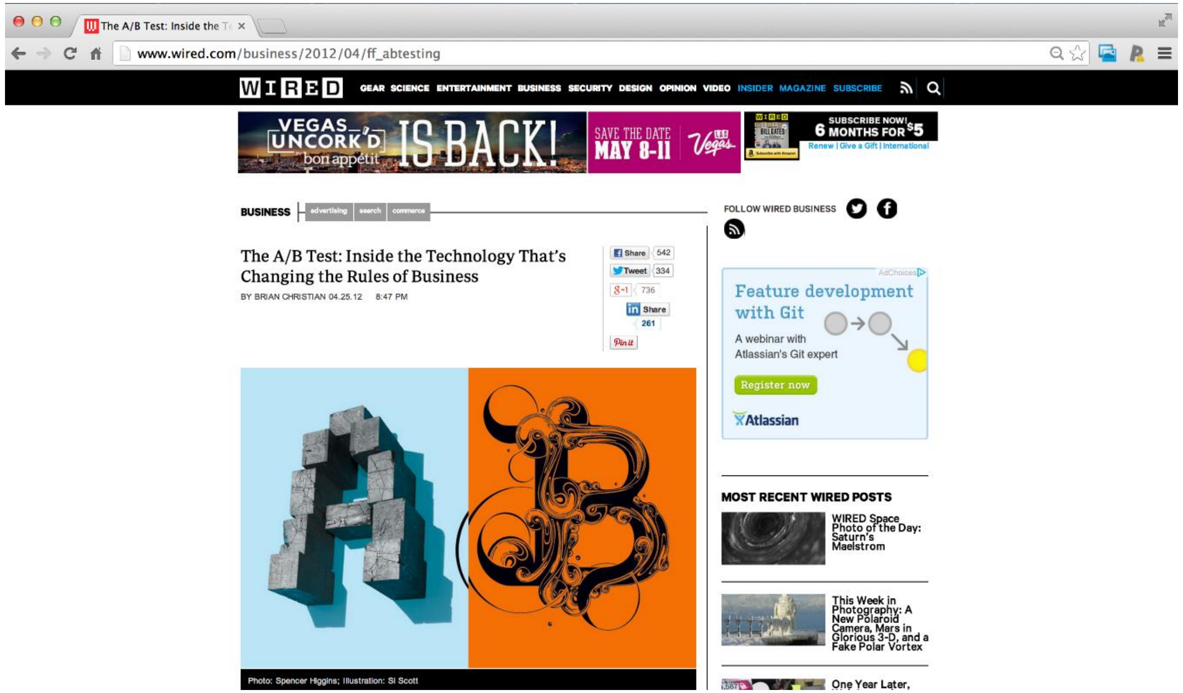


May I recommend?



https://github.com/jtleek/datasharing

Formulate your question in advance



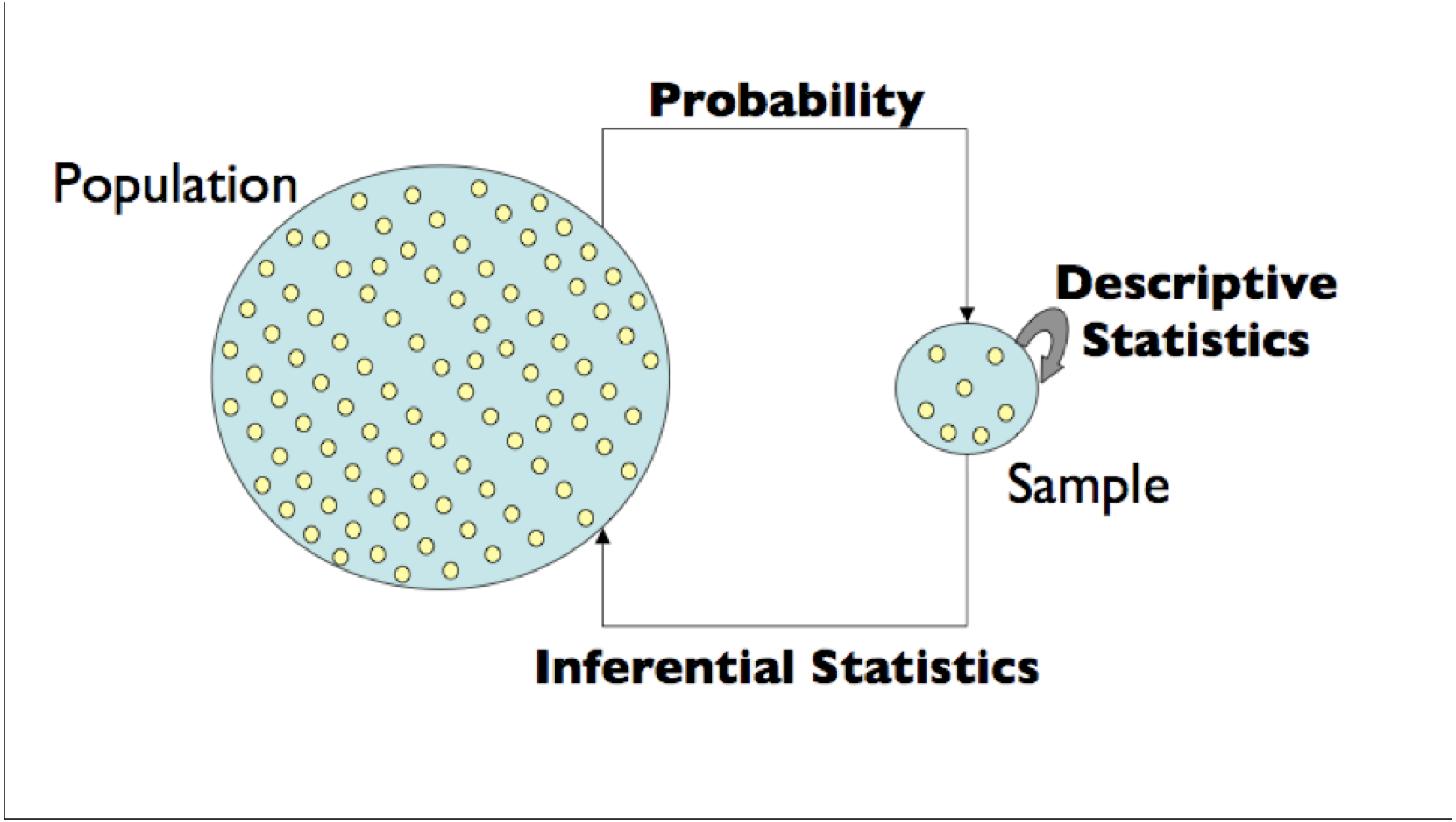
Question: Does changing the text on your website improve donations?

Experiment:

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

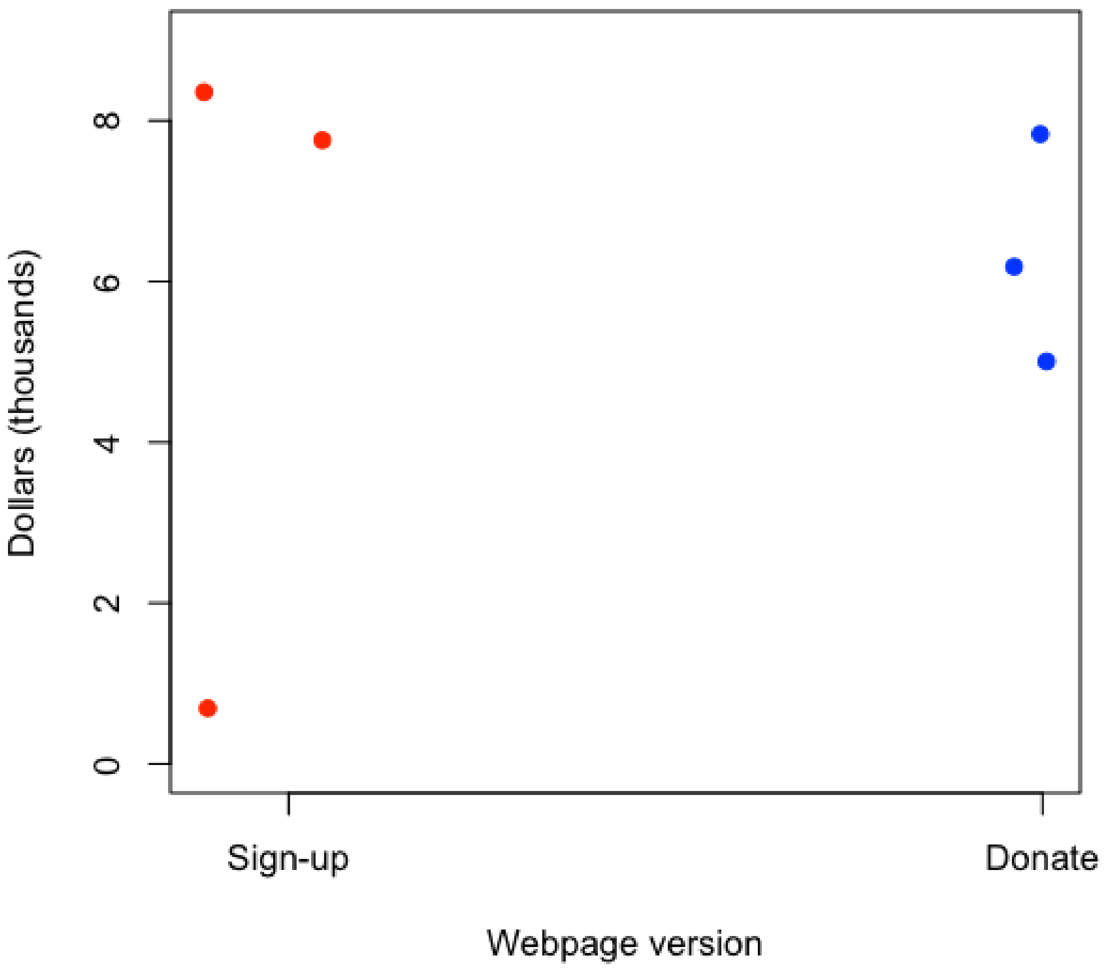
http://www.wired.com/business/2012/04/ff_abtesting

Statistical inference

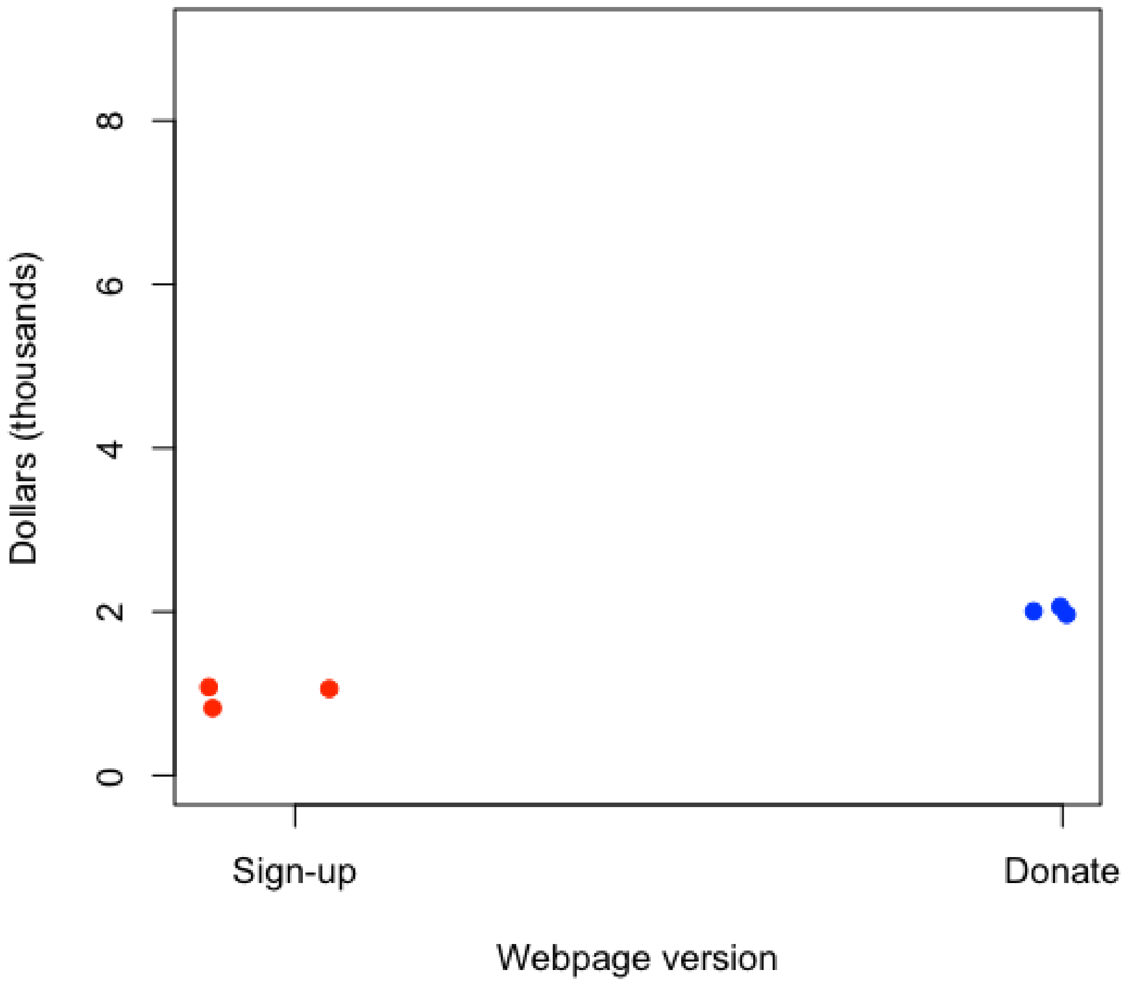


http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture2.pdf

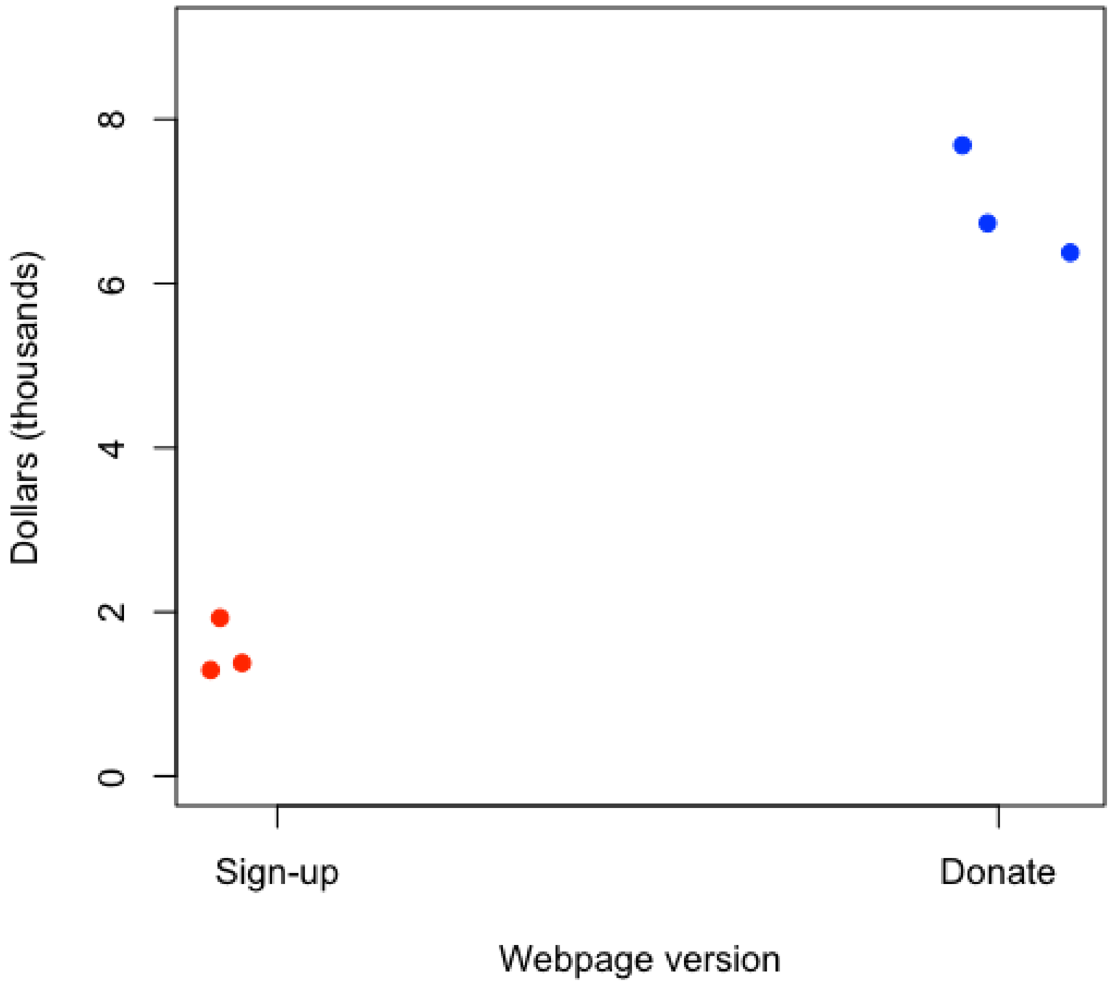
Variability - Scenario 1



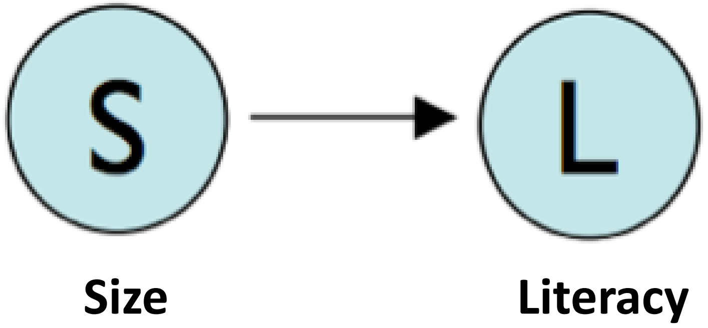
Variability (варіативність) - Scenario 2

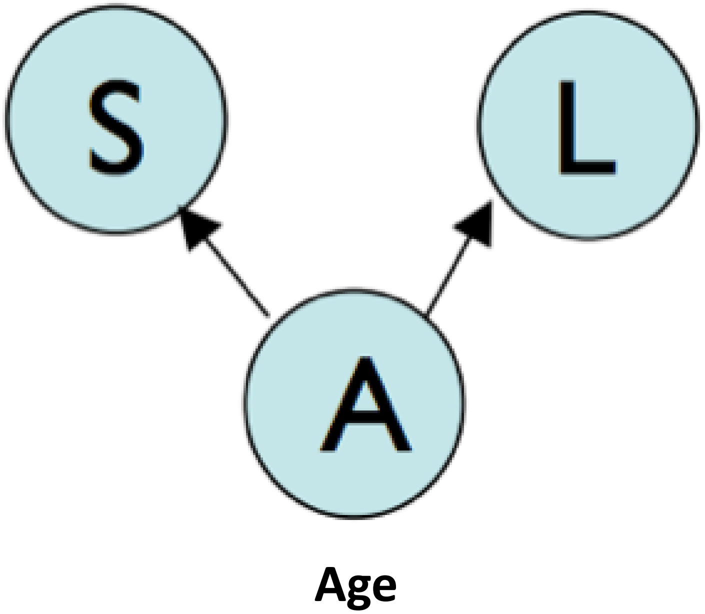


Variability (варіативність) - Scenario 3

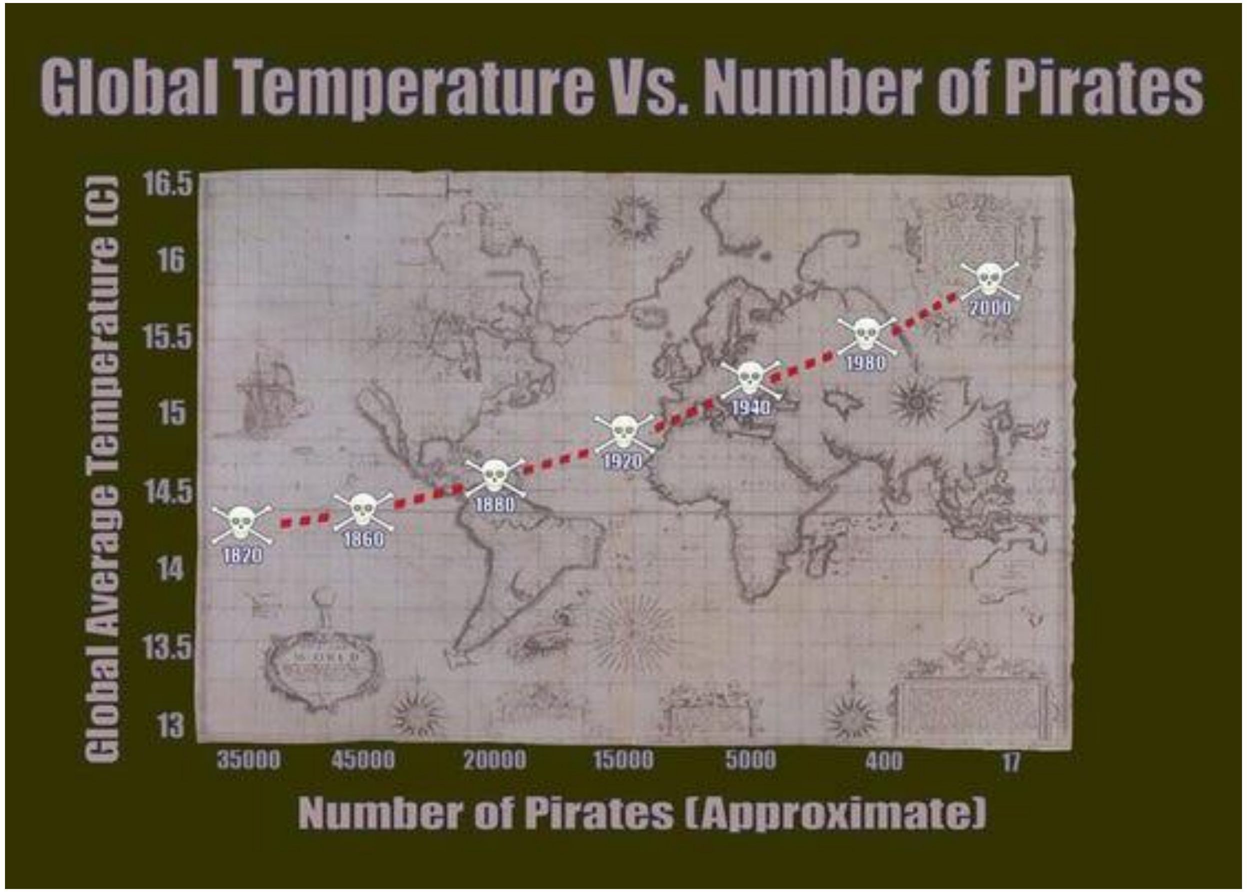


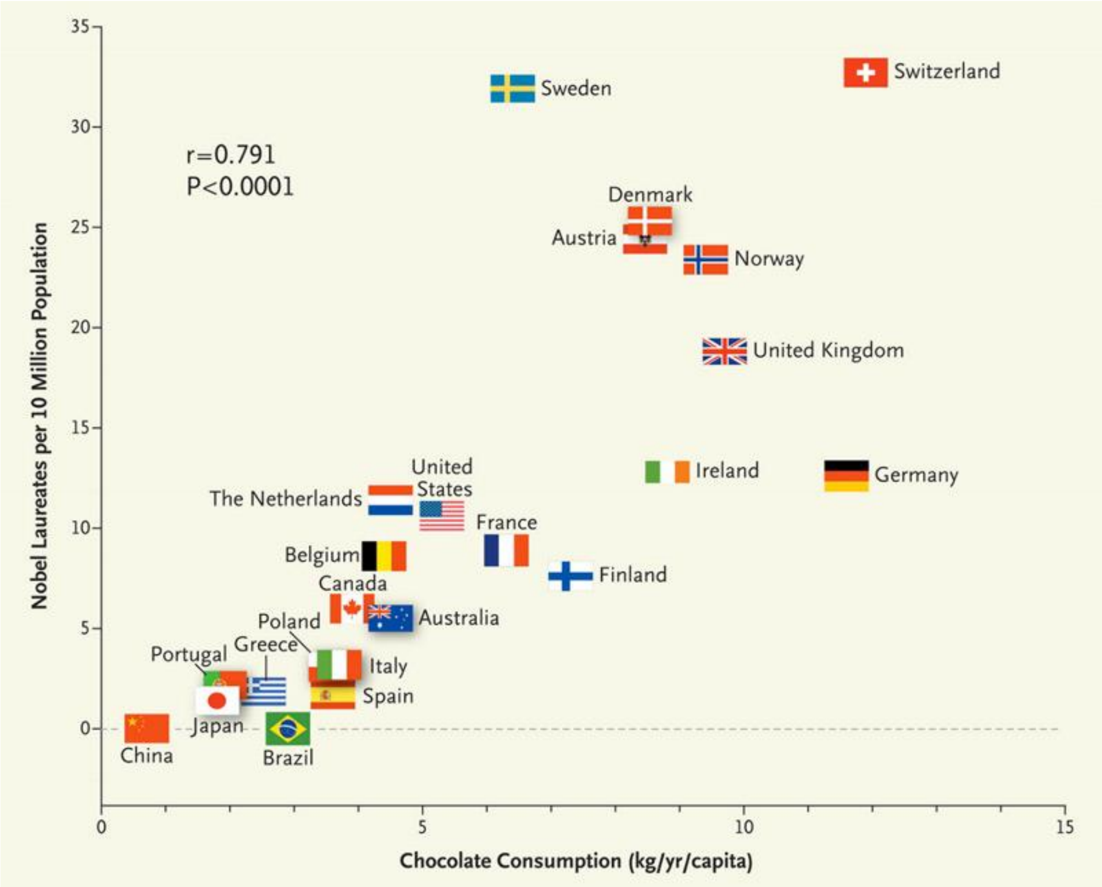
Confounding (спотворювання)





Correlation is not causation*





https://www.forbes.com/sites/erikaandersen/2012/03/23/true-fact-the-lack-of-pirates-is-causing-global-warming/?sh=de2ac883a679

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

*Sometimes called spurious correlation (хибна кореляція)**

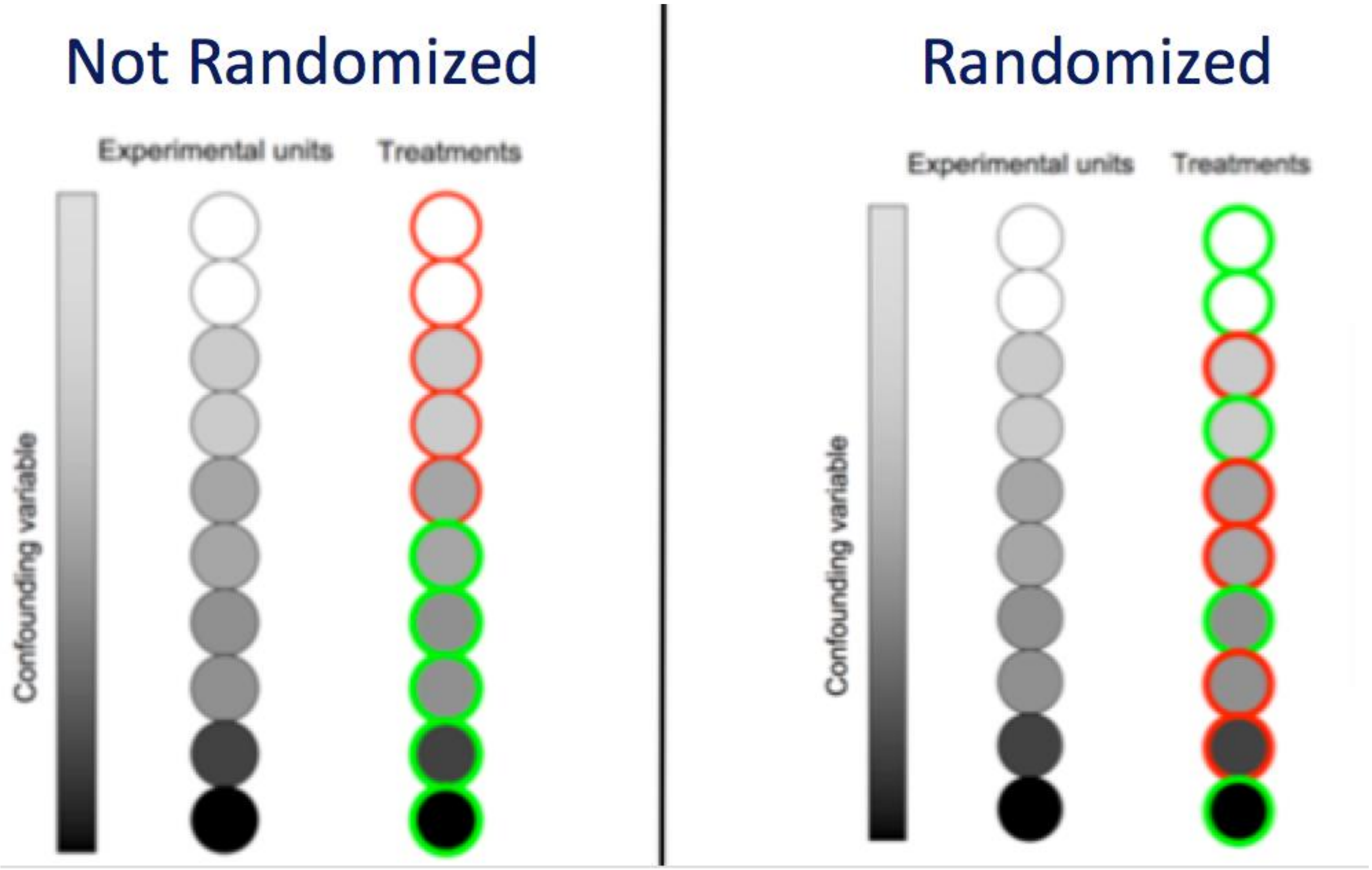
Randomization and blocking

- If you can (and want to) fix a variable

    Website always says Obama 2012 on it

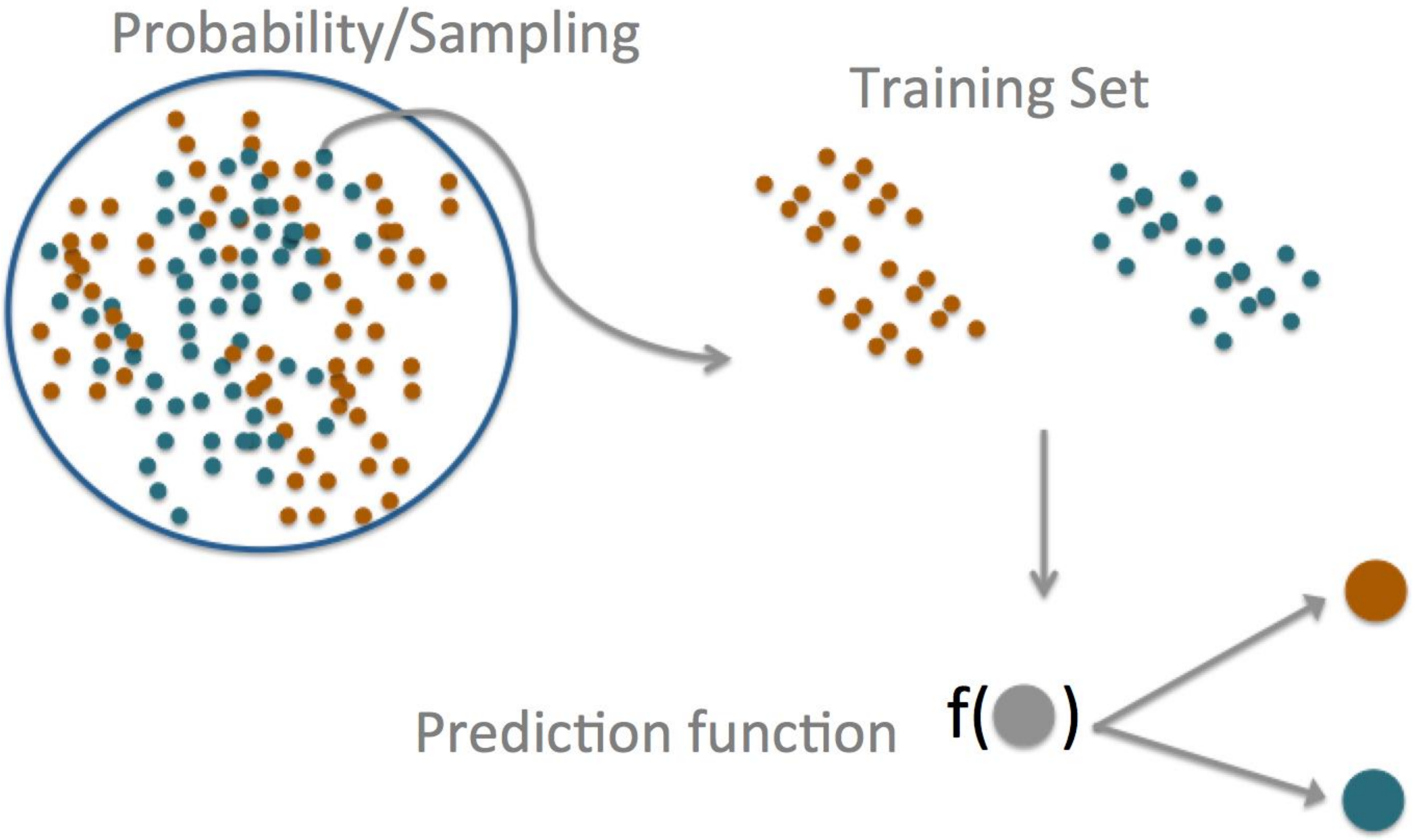
- If you don't fix a variable, stratify it
- *If you are testing sign up phrases and have two website colors, use both phrases equally on both.*
- If you can't fix a variable, randomize it

Why does randomization help?

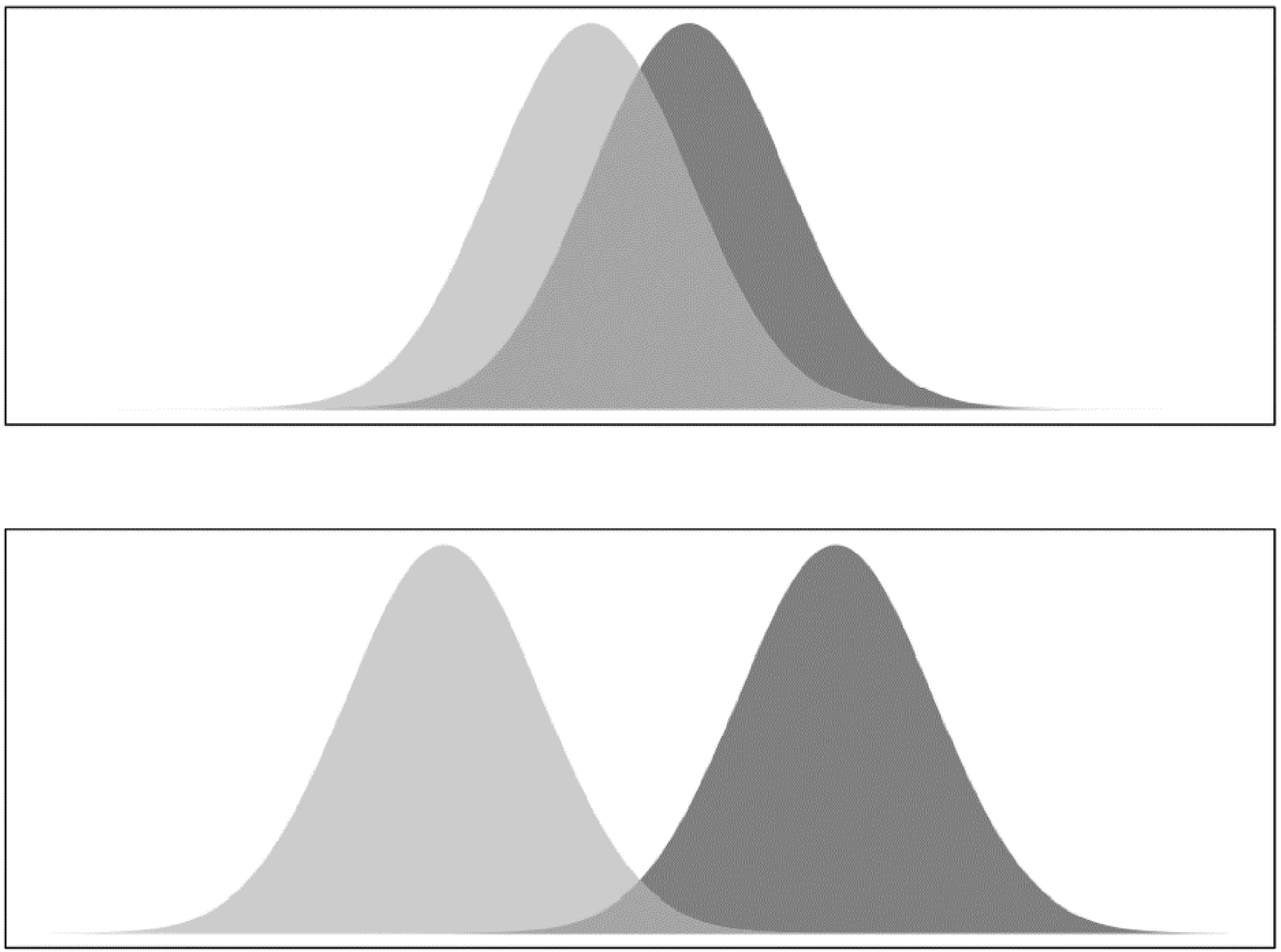


http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture1.pdf

Prediction



Prediction versus inference



http://www.biostat.jhsph.edu/%7Eiruczins/teaching/140.615/

Prediction key quantities

TEST \ DISEASE	+	−
+	TP	FP
−	FN	TN

TP – True Positive
FP – False Positive
FN – False Negative
TN – True Nevative

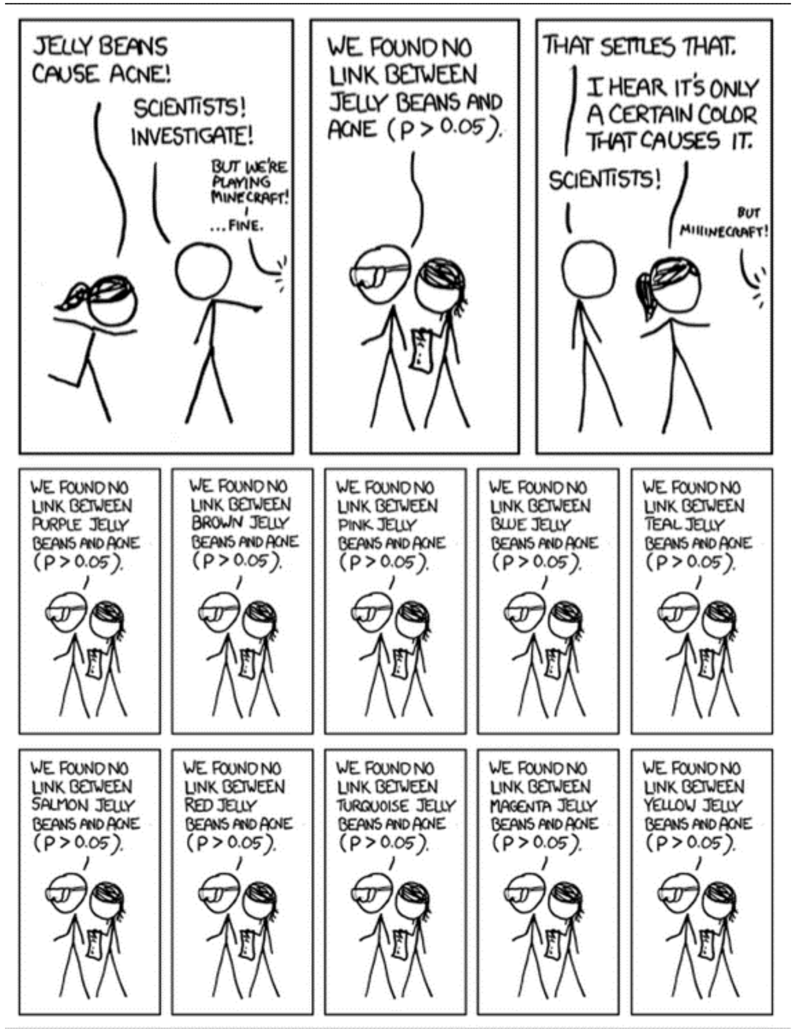
Sensitivity → Pr (positive test | disease)
Specificity → Pr (negative test | no disease)
Positive Predictive Value → Pr (disease | positive test)
Negative Predictive Value → Pr (no disease | negative test)
Accuracy → Pr (correct outcome)

http://www.biostat.jhsph.edu/%7Eiruczins/teaching/140.615/

Beware data dredging

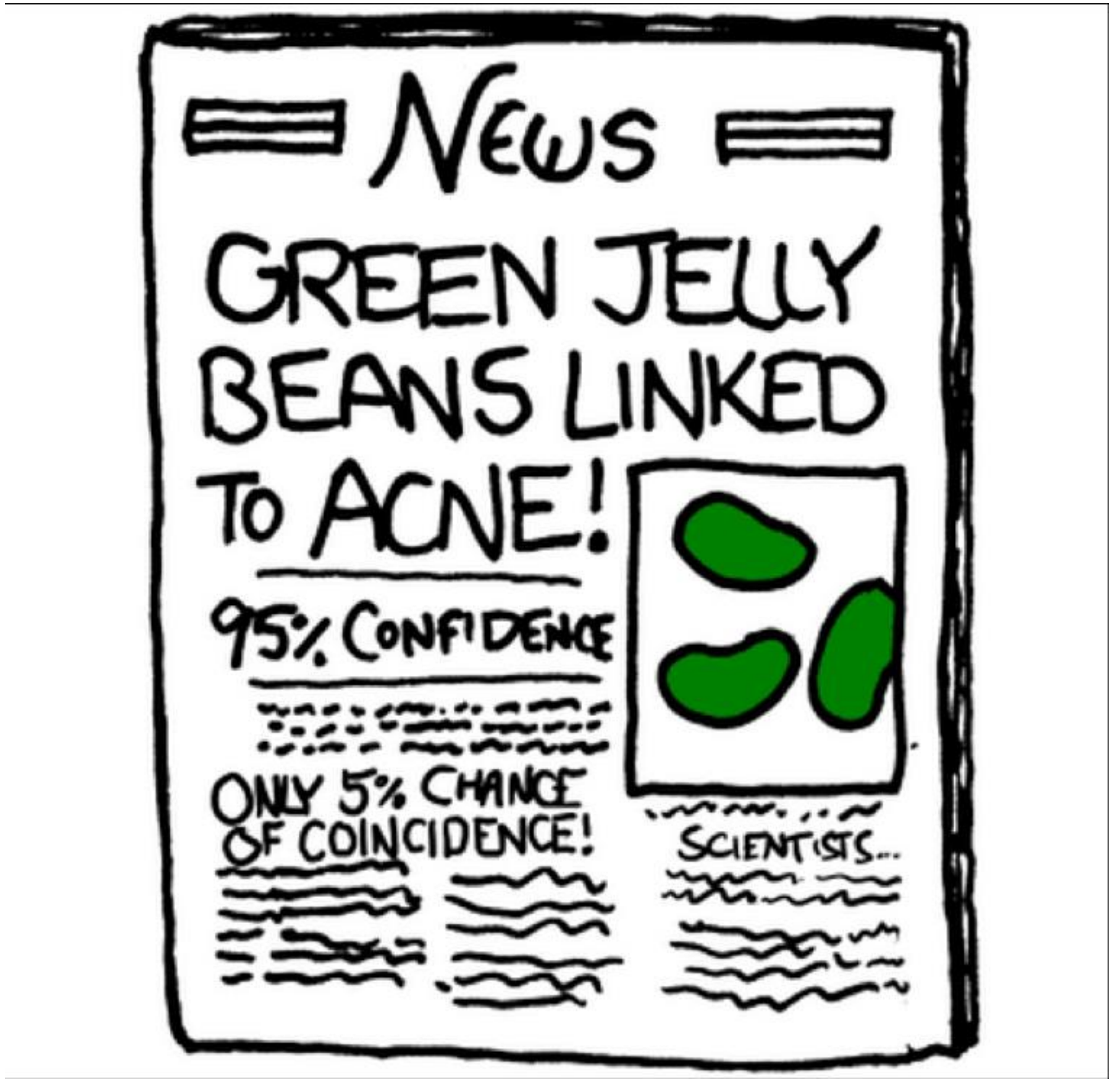
http://xkcd.com/882/

Beware data dredging



http://xkcd.com/882/

Beware data dredging



http://xkcd.com/882/

Summary

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

References

1. Course materials for the Data Science Specialization: https://www.coursera.org/specialization/jhudatascience/1 https://github.com/DataScienceSpecialization/courses
2. The Elements of Data Analytic Style. A guide for people who want to analyze data. Jeff Leek. This book is for sale at http://leanpub.com/datastyle
3. https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/