Types of Data Science Questions

Jeffrey Leek
Johns Hopkins Bloomberg School of Public Health

Types of Data Science Questions

In approximate order of difficulty

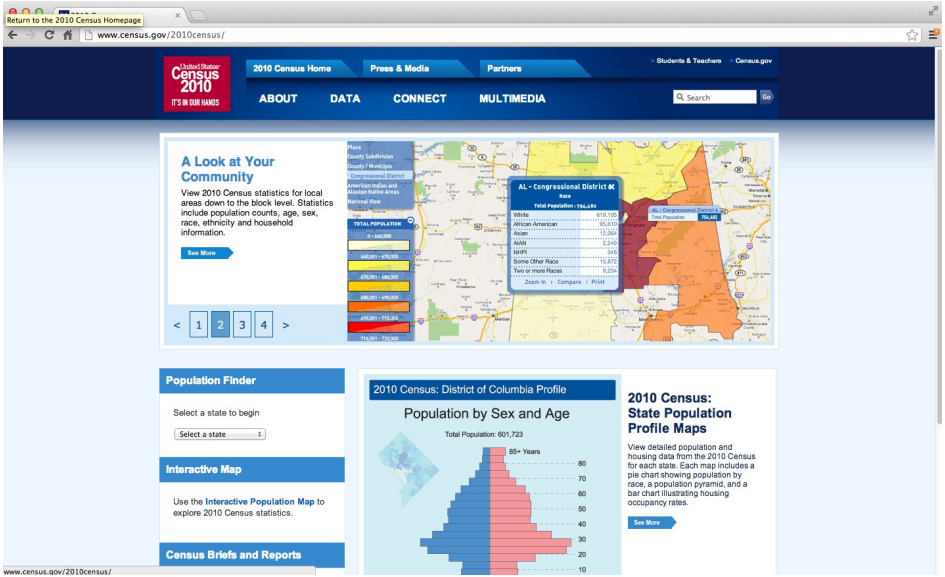
- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

About descriptive analyses

Goal: Describe a set of data

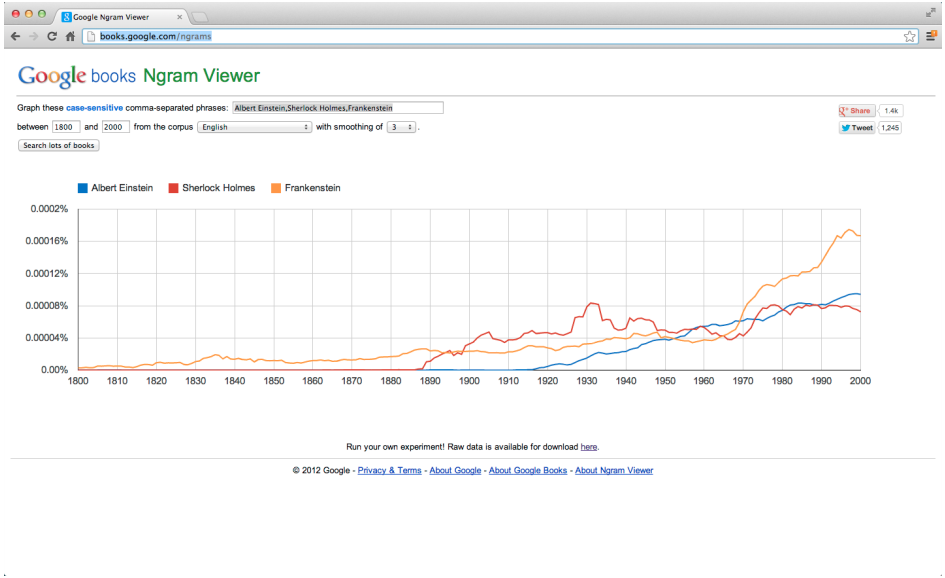
- The first kind of data analysis performed
- Commonly applied to census data
- The description and interpretation are different steps
- Descriptions can usually not be generalized without additional statistical modeling
  so just describing, not interpreting meaning from description

Descriptive analysis



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

Descriptive analysis



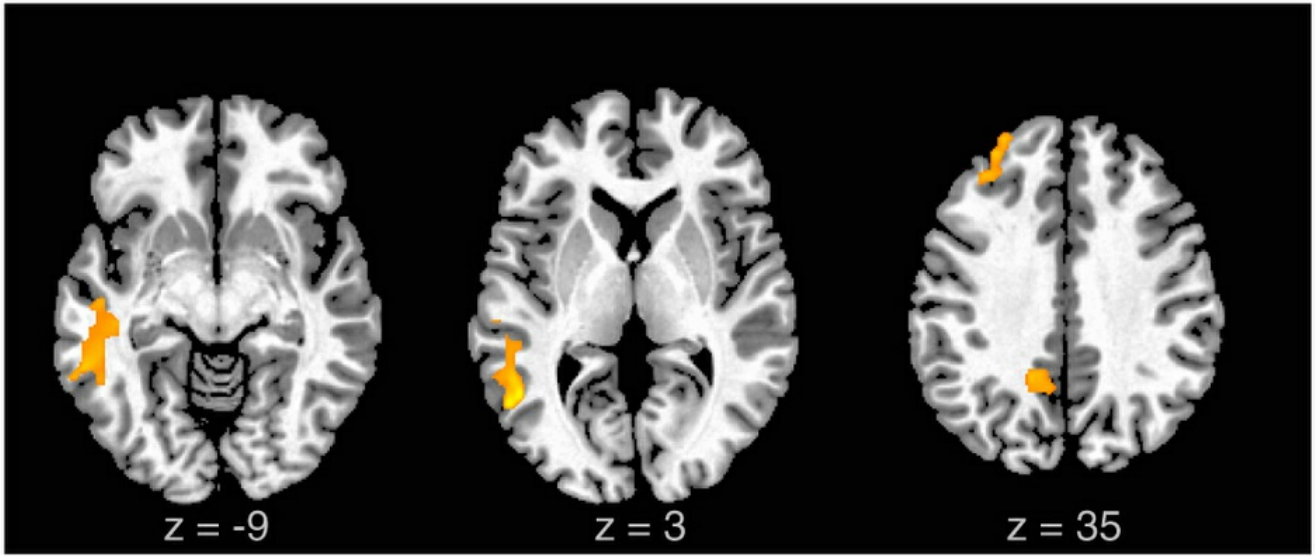
http://books.google.com/ngrams

About exploratory analysis

Goal: Find relationships you didn't know about

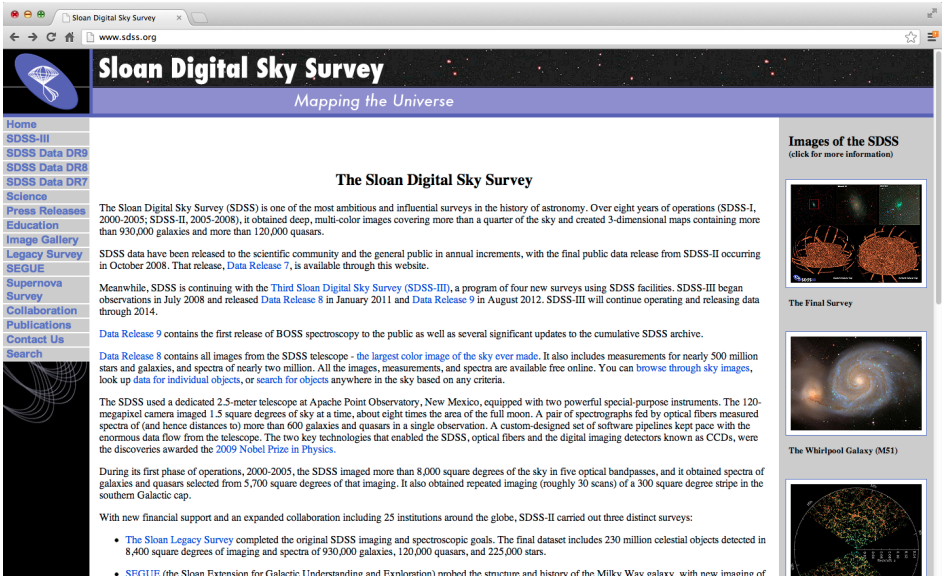
- Exploratory models are good for discovering new connections
- They are also useful for defining future studies
- Exploratory analyses are usually not the final say
- Exploratory analyses alone should not be used for generalizing/predicting just finding relationships btw variables
- Correlation does 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 part for a whole

- 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

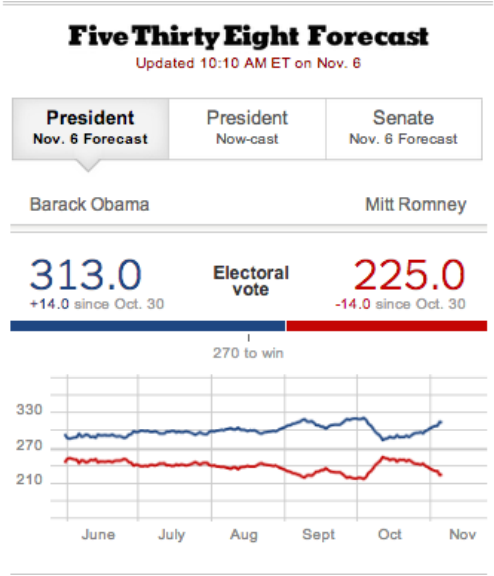
Correia et al. (2013) Epidemiology

About predictive analysis

Goal: To use the data on some objects to predict values for another object what does one say about the other

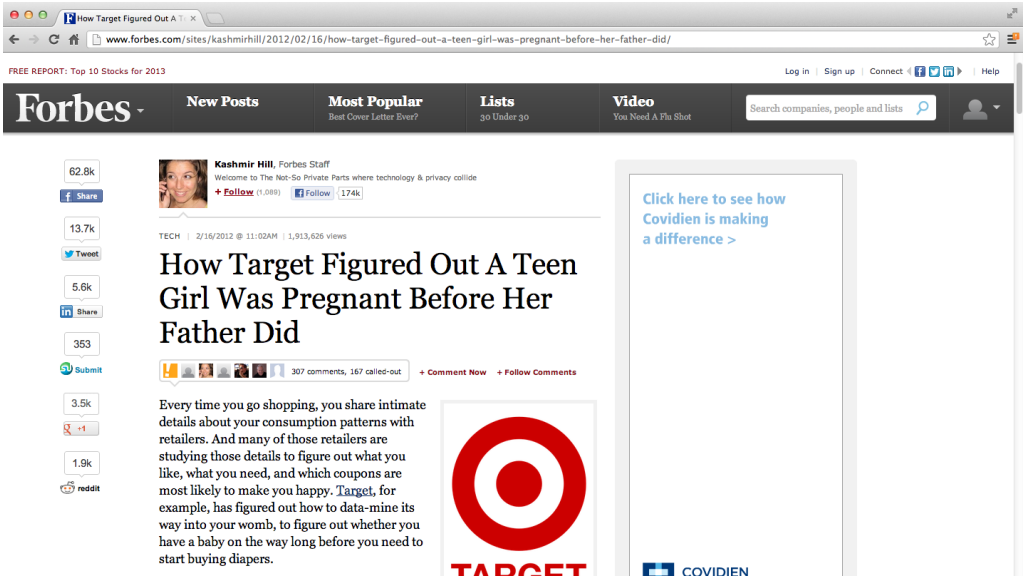
- If X predicts Y it does not mean that X causes Y
- Accurate prediction depends heavily on measuring the right variables
- Although there are better and worse prediction models, more data and a simple model works really well
- Prediction is very hard, especially about the future references

Predictive analysis



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

Predictive analysis



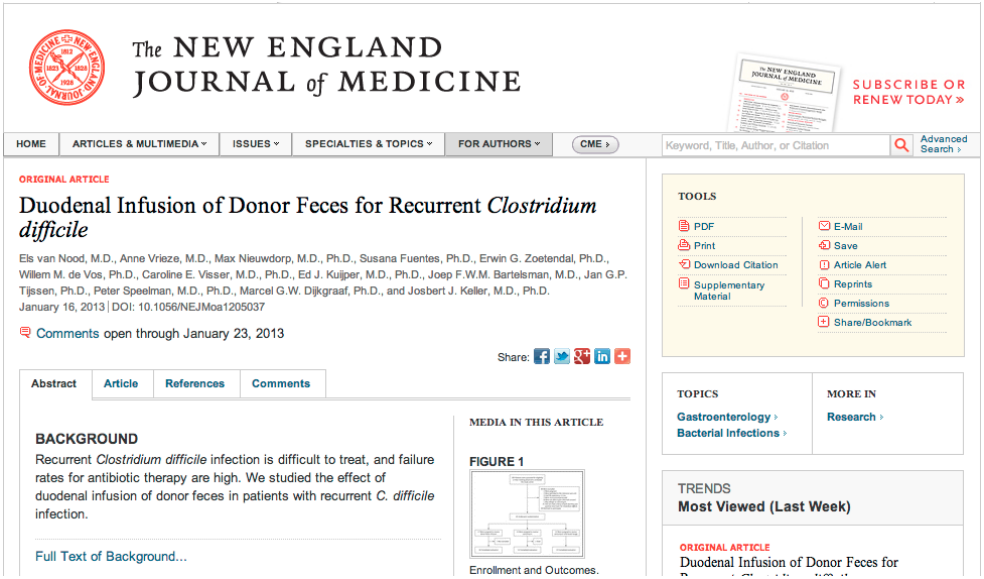
http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

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