



# 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

The screenshot shows the 2010 Census website interface. At the top, there is a navigation bar with links for '2010 Census Home', 'Press & Media', and 'Partners'. Below this is a secondary navigation bar with 'ABOUT', 'DATA', 'CONNECT', and 'MULTIMEDIA'. A search bar is located on the right side of this bar.

The main content area features a section titled 'A Look at Your Community' with a sub-header 'View 2010 Census statistics for local areas down to the block level. Statistics include population counts, age, sex, race, ethnicity and household information.' Below this is a 'See More' button.

To the right of this section is a map of Alabama with a pop-up window for 'AL - Congressional District 11'. The pop-up window displays a table of demographic data:

Total Population: 125,444	
White	63,700
African American	30,810
Asian	10,264
ALN	2,243
Hispanic	945
Some Other Race	18,672
Two or more Races	8,204

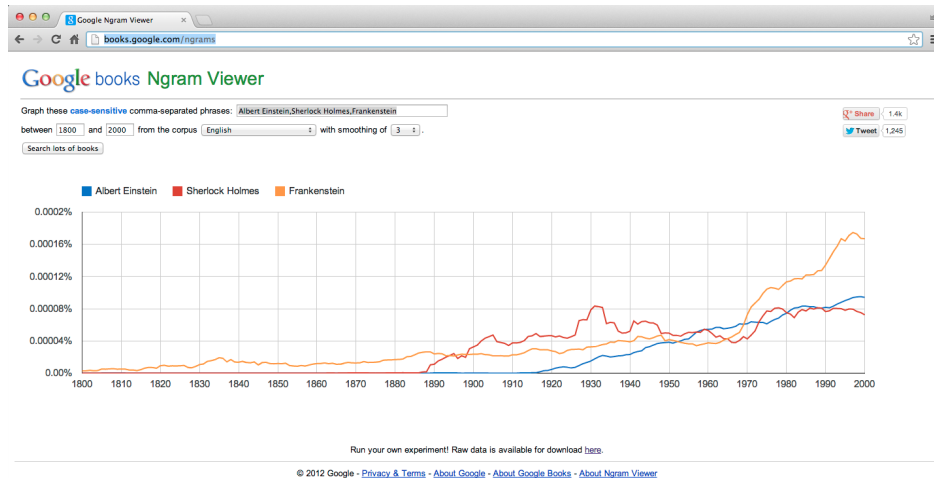
Below the map is a 'Population Finder' section with a dropdown menu to 'Select a state to begin' and another to 'Select a state'. Below that is an 'Interactive Map' section with a link to 'Use the Interactive Population Map to explore 2010 Census statistics.' At the bottom left is a 'Census Briefs and Reports' section.

On the right side of the main content area, there is a section titled '2010 Census: District of Columbia Profile' with a sub-header 'Population by Sex and Age'. Below this is a population pyramid for the District of Columbia, showing a total population of 601,723. The pyramid is divided into two halves: males on the left and females on the right. The age groups are listed on the right side of the pyramid, ranging from 0-4 to 85+ years.

Below the population pyramid is a section titled '2010 Census: State Population Profile Maps' with a sub-header 'View detailed population and housing data from the 2010 Census for each state. Each map includes a pie chart showing population by race, a population pyramid, and a bar chart illustrating housing occupancy rates.' Below this is a 'See More' button.

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

# Descriptive analysis



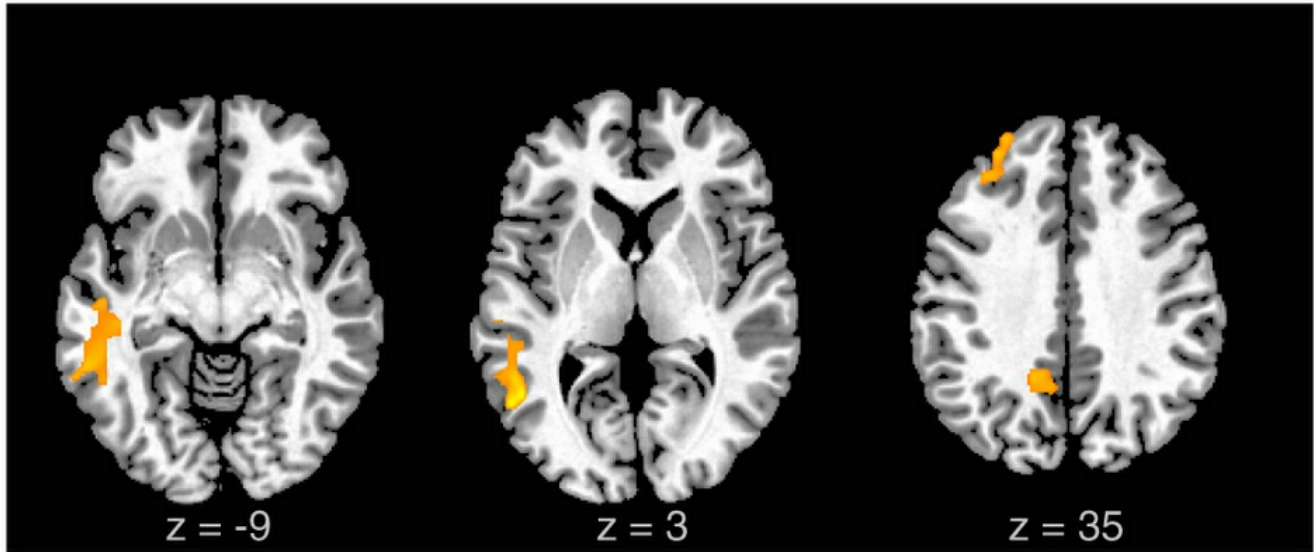
<http://books.google.com/ngrams>

# About exploratory analysis

**Goal:** Find relationships you didn't know about

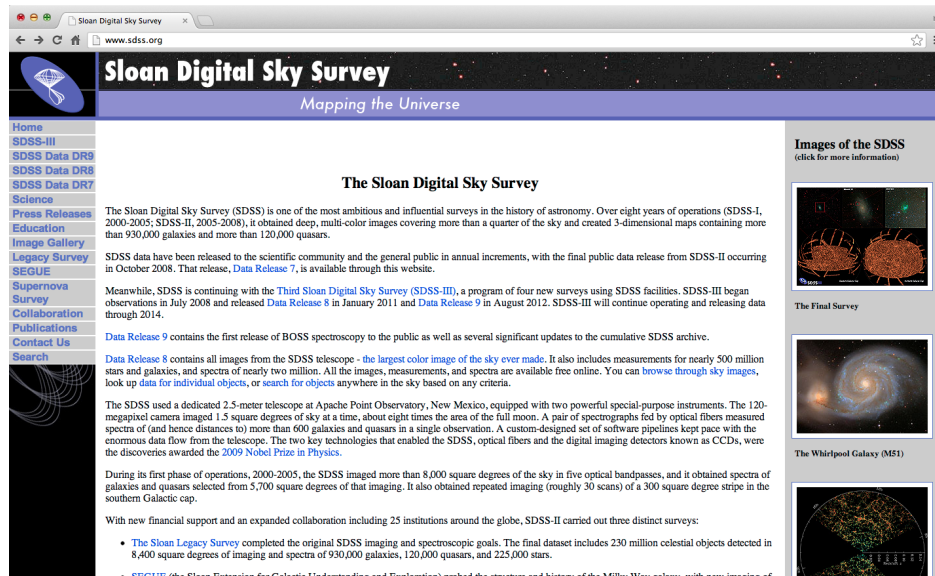
- Exploratory models are good for discovering new connections
- They are also useful for defining future studies
- Exploratory analyses are usually not the final say
- Exploratory analyses alone should not be used for generalizing/predicting
- [Correlation does not imply causation](#)

# Exploratory analysis



[Liu et al. \(2012\) Scientific Reports](#)

# Exploratory analysis



The screenshot shows the Sloan Digital Sky Survey website. The browser address bar displays "www.sdss.org". The page header features the Sloan Digital Sky Survey logo and the tagline "Mapping the Universe". A navigation menu on the left includes links for Home, SDSS-III, SDSS Data DR9, SDSS Data DR8, SDSS Data DR7, Science, Press Releases, Education, Image Gallery, Legacy Survey SEGUE, Supernova Survey, Collaboration, Publications, Contact Us, and Search. The main content area is titled "The Sloan Digital Sky Survey" and contains several paragraphs of text describing the survey's history, data releases, and scientific goals. On the right side, there is a section titled "Images of the SDSS" with three sub-sections: "The Final Survey" showing a star field, "The Whirlpool Galaxy (M51)" showing a spiral galaxy, and another image of a star field.

## The Sloan Digital Sky Survey

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 3,700 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.
- [SEGUE](#) (the Sloan Extension for Galactic Understanding and Exploration) probed 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

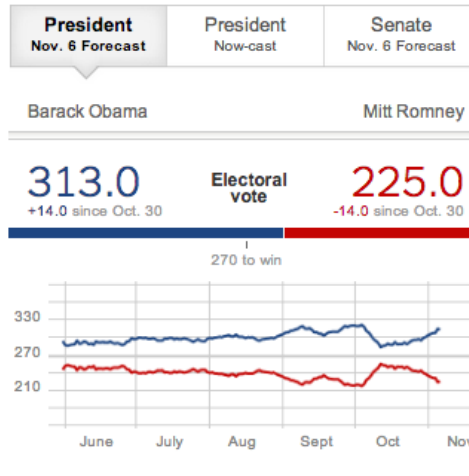
**Goal:** To use the data on some objects to predict values for another object

- 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

## Five Thirty Eight Forecast

Updated 10:10 AM ET on Nov. 6



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

# Predictive analysis

The screenshot shows a web browser displaying a Forbes article. The browser's address bar shows the URL: [www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/](http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/). The Forbes navigation bar includes sections for 'New Posts', 'Most Popular', 'Lists', and 'Video'. The article is by Kashmir Hill, a Forbes Staff member, and is dated February 16, 2012. The article title is 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did'. The main text begins with 'Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.' A large red Target logo is prominently displayed. To the right of the article is a sidebar with a 'Click here to see how Covidien is making a difference >' link and a Covidien logo.

62.8k  
Share  
13.7k  
Tweet  
5.6k  
Share  
353  
Submit  
3.5k  
+1  
1.9k  
reddit

**Kashmir Hill**, Forbes Staff  
Welcome to The Not-So Private Parts where technology & privacy collide  
[+ Follow](#) (1,089) [Follow](#) (174k)

TECH | 2/16/2012 @ 11:02AM | 1,913,626 views

## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

307 comments, 167 called-out [+ Comment Now](#) [+ Follow Comments](#)

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

**TARGET**

[Click here to see how Covidien is making a difference >](#)  
**COVIDIEN**

<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 screenshot shows the NEJM website interface. At the top left is the NEJM logo. The main header reads "The NEW ENGLAND JOURNAL of MEDICINE". A navigation bar includes links for HOME, ARTICLES & MULTIMEDIA, ISSUES, SPECIALTIES & TOPICS, FOR AUTHORS, and CME. A search bar is on the right. The article title is "Duodenal Infusion of Donor Feces for Recurrent *Clostridium difficile*". Below the title is the author list: "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. Barteldsman, 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." The publication date is "January 16, 2013" and the DOI is "10.1056/NEJMoa1205037". There is a "Comments" section with the text "Comments open through January 23, 2013". A "Share:" section includes icons for Facebook, Twitter, LinkedIn, and a plus sign. Below the title are tabs for "Abstract", "Article", "References", and "Comments". The "Background" section starts with "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." A "Full Text of Background..." link is present. To the right, under "MEDIA IN THIS ARTICLE", is "FIGURE 1" with a flowchart titled "Enrollment and Outcomes." The right sidebar contains a "TOOLS" section with links for PDF, Print, Download Citation, Supplementary Material, E-Mail, Save, Article Alert, Reprints, Permissions, and Share/Bookmark. Below that is a "TOPICS" section with "Gastroenterology" and "Bacterial Infections", and a "MORE IN" section with "Research". At the bottom of the sidebar is a "TRENDS" section titled "Most Viewed (Last Week)" with a link to the current article.

[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).

[http://www.fhwa.dot.gov/resourcecenter/teams/pavement/pave\\_3pdg.pdf](http://www.fhwa.dot.gov/resourcecenter/teams/pavement/pave_3pdg.pdf)