Types of Data Analysis Questions

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Types of Data Analysis Questions

In approximate order of difficulty

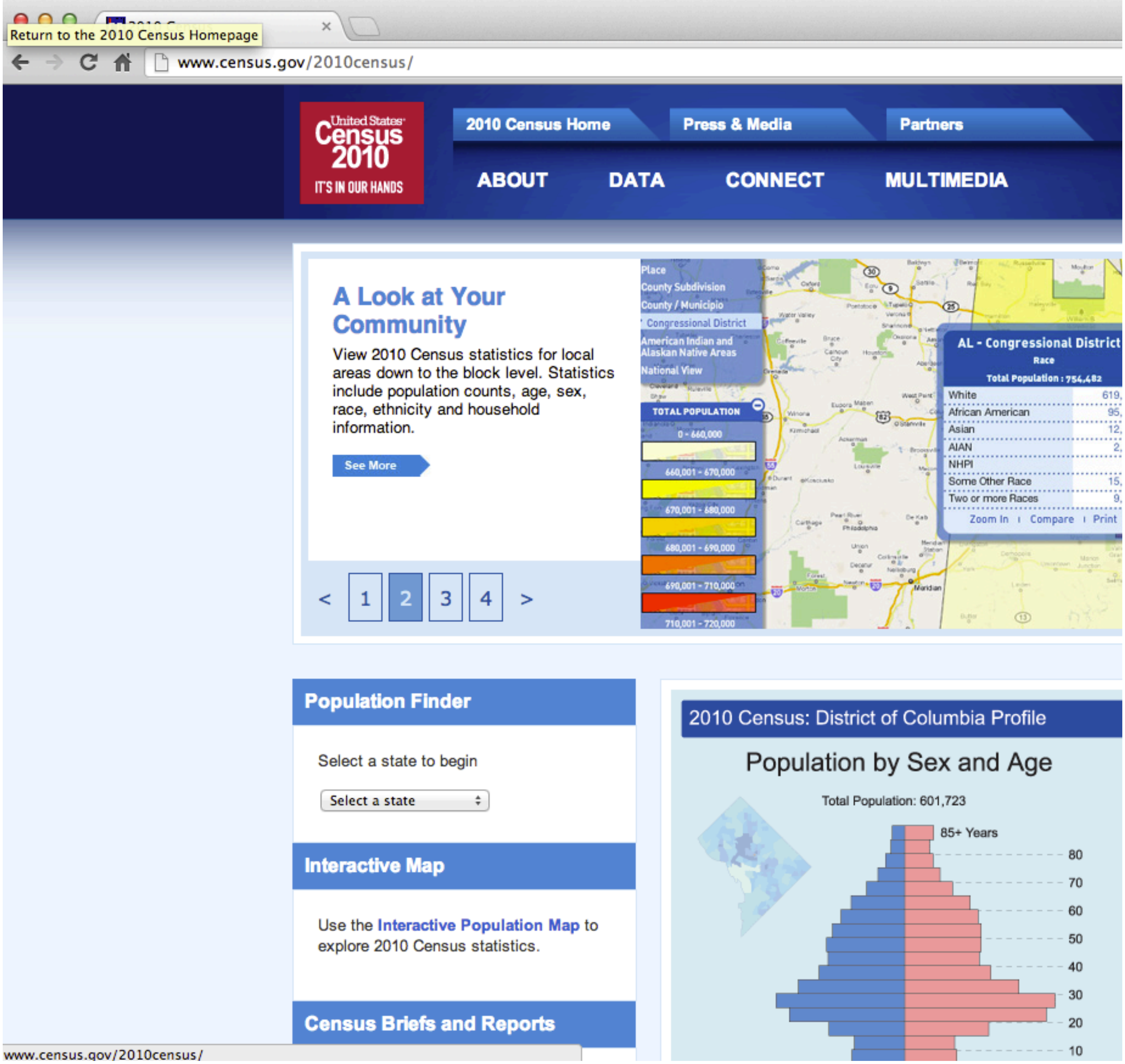
- Descriptive
- Exploratory
- Inferential
- Predictive
- Causal
- Mechanistic

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



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.

AL - Congressional District

Race

Total Population : 754,482

White	619,
African American	95,
Asian	12,
AIAN	2,
NHPI	
Some Other Race	15,
Two or more Races	9,

Population Finder

Select a state to begin

Interactive Map

Use the Interactive Population Map to explore 2010 Census statistics.

Census Briefs and Reports

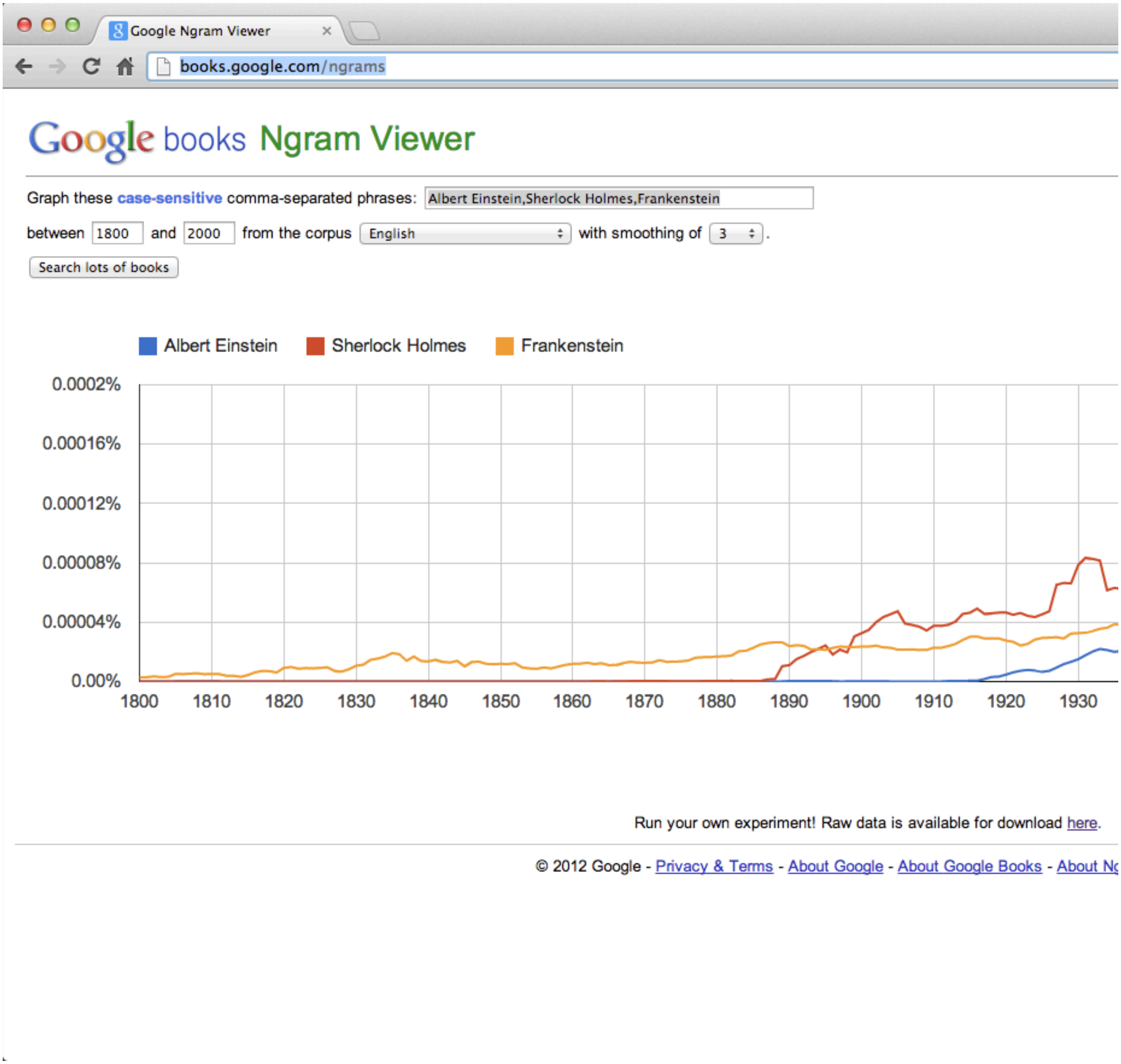
2010 Census: District of Columbia Profile

Population by Sex and Age

Total Population: 601,723

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

Descriptive analysis



Graph these case-sensitive comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein

between 1800 and 2000 from the corpus English with smoothing of 3.

Run your own experiment! Raw data is available for download here.

© 2012 Google - Privacy & Terms - About Google - About Google Books - About Ng

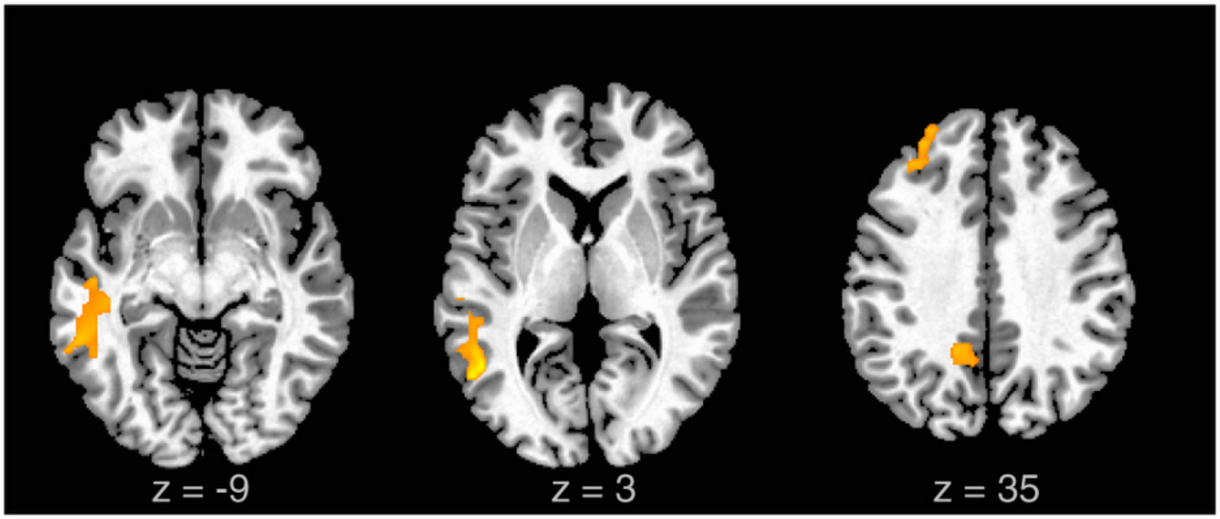
http://books.google.com/ngrams

About exploratory analysis

Goal: Find relationships you didn't know about

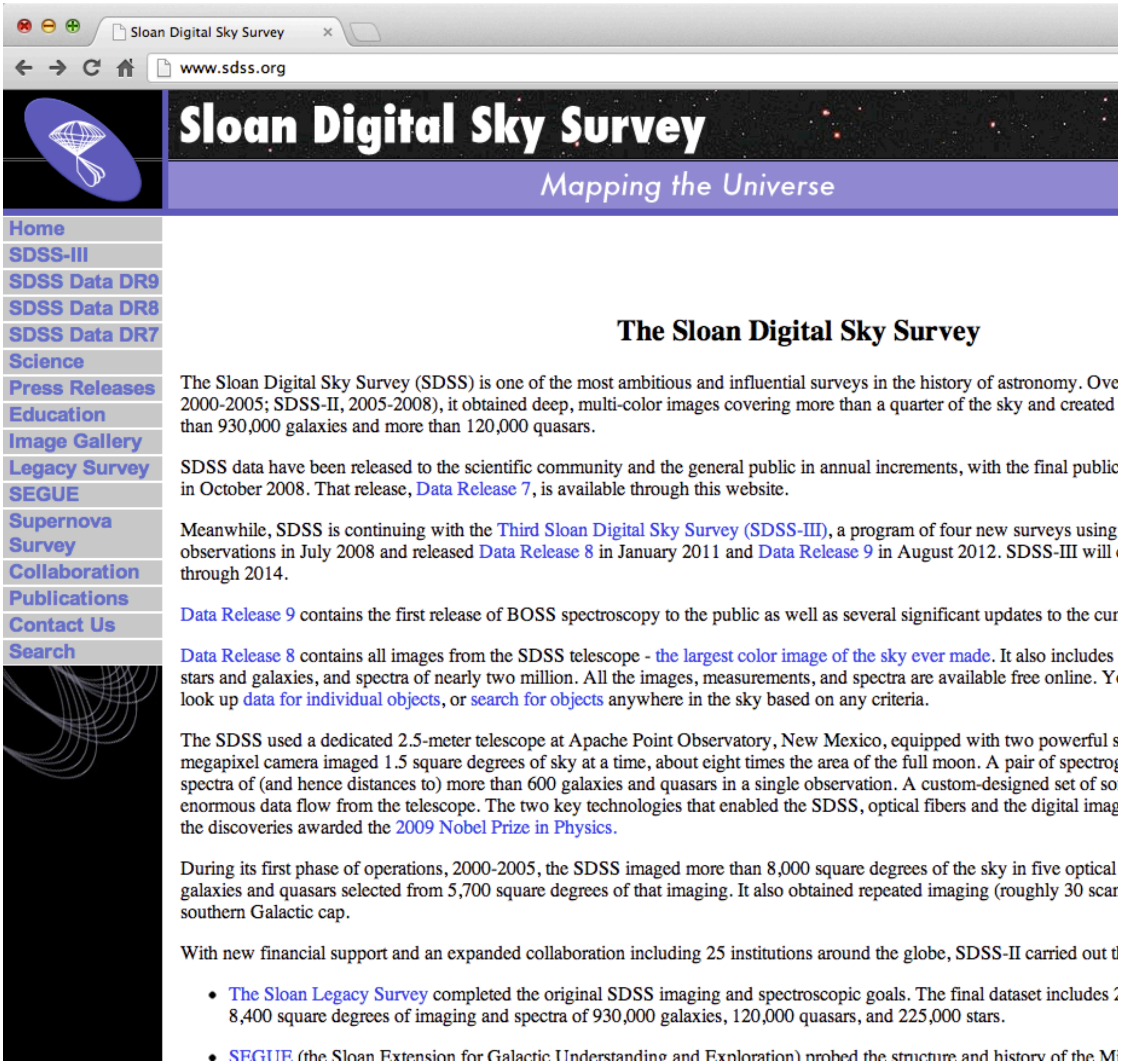
- Exploratory models are good for discovering new connections
- They are also useful for defining future studies
- Exploratory analyses are usually not the final say
- Exploratory analyses alone should not be used for generalizing/predicting
- Correlation does not imply causation

Exploratory analysis



Liu et al. (2012) Scientific Reports

Exploratory analysis



The Sloan Digital Sky Survey

The Sloan Digital Sky Survey (SDSS) is one of the most ambitious and influential surveys in the history of astronomy. Ove 2000-2005; SDSS-II, 2005-2008), it obtained deep, multi-color images covering more than a quarter of the sky and created 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 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 observations in July 2008 and released Data Release 8 in January 2011 and Data Release 9 in August 2012. SDSS-III will through 2014.

Data Release 9 contains the first release of BOSS spectroscopy to the public as well as several significant updates to the cur

Data Release 8 contains all images from the SDSS telescope - the largest color image of the sky ever made. It also includes stars and galaxies, and spectra of nearly two million. All the images, measurements, and spectra are available free online. Y 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 s megapixel camera imaged 1.5 square degrees of sky at a time, about eight times the area of the full moon. A pair of spectrog spectra of (and hence distances to) more than 600 galaxies and quasars in a single observation. A custom-designed set of so enormous data flow from the telescope. The two key technologies that enabled the SDSS, optical fibers and the digital imag 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 galaxies and quasars selected from 5,700 square degrees of that imaging. It also obtained repeated imaging (roughly 30 scar southern Galactic cap.

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

- The Sloan Legacy Survey completed the original SDSS imaging and spectroscopic goals. The final dataset includes 2 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 Mi

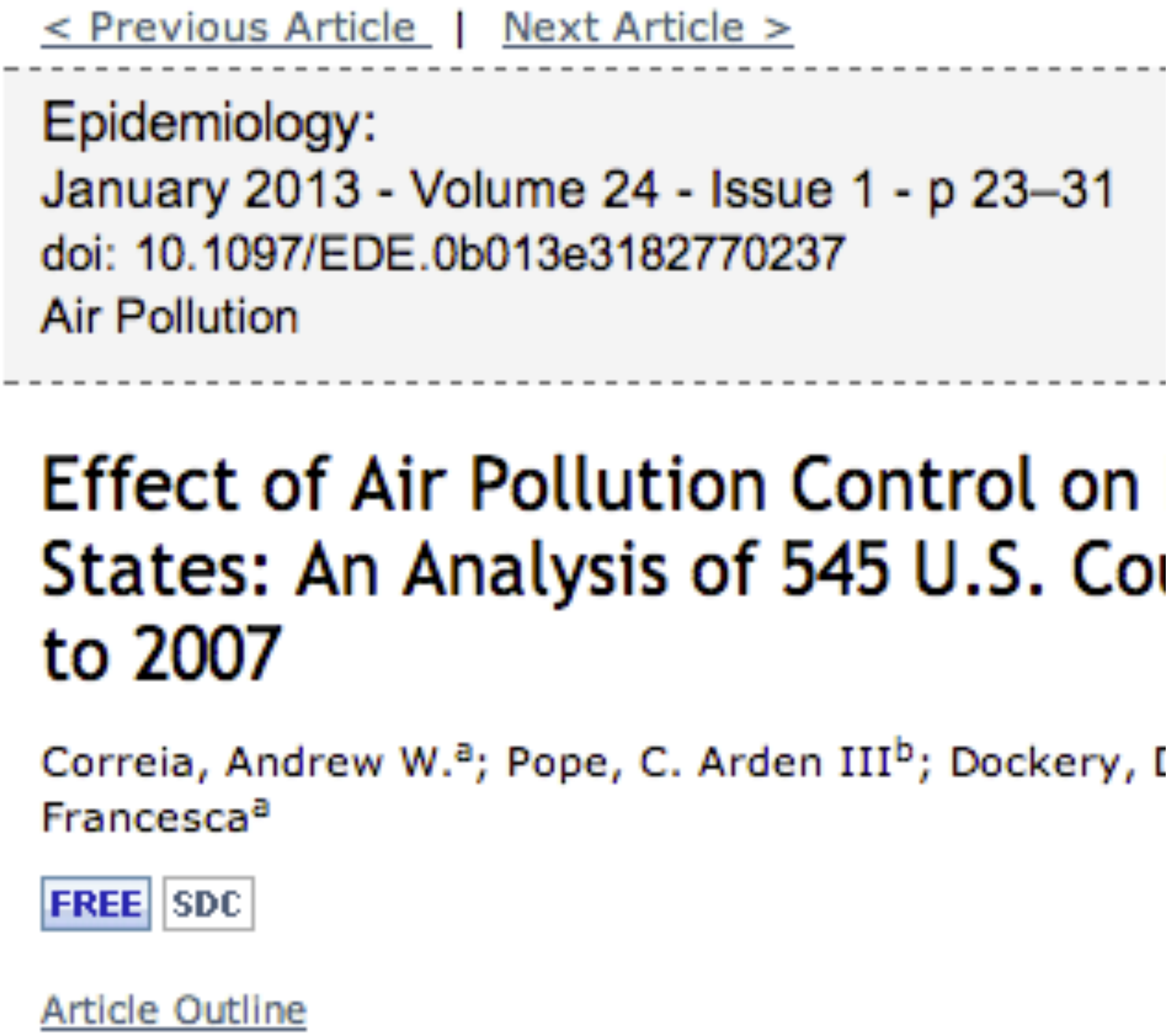
http://www.sdss.org/

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
States: An Analysis of 545 U.S. Co
to 2007

Correia, Andrew W.[a]; Pope, C. Arden III[b]; Dockery,
Francesca[a]

FREE SDC

Article Outline

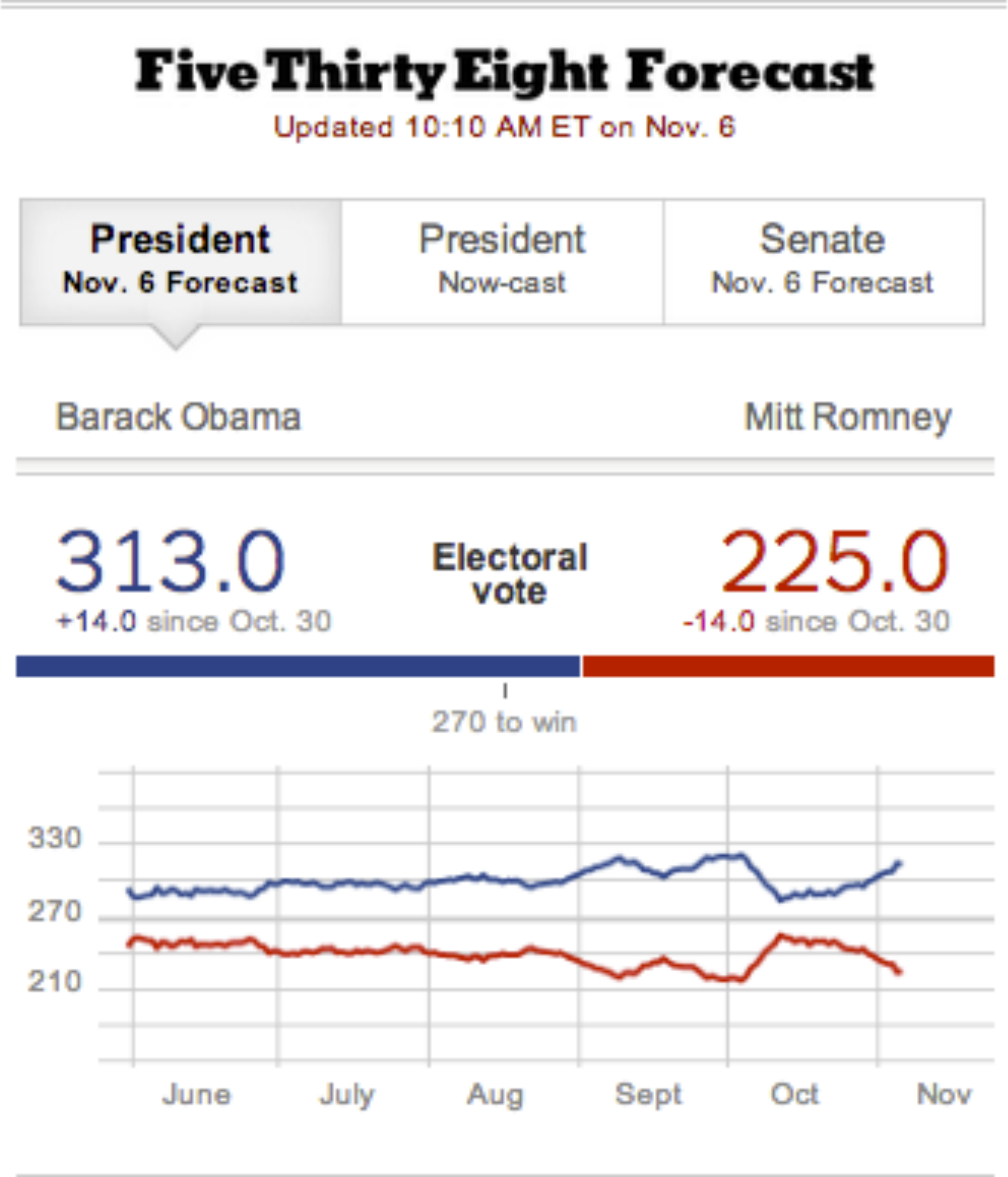
Correia et al. (2013) Epidemiology

About predictive analysis

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

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

Predictive analysis



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

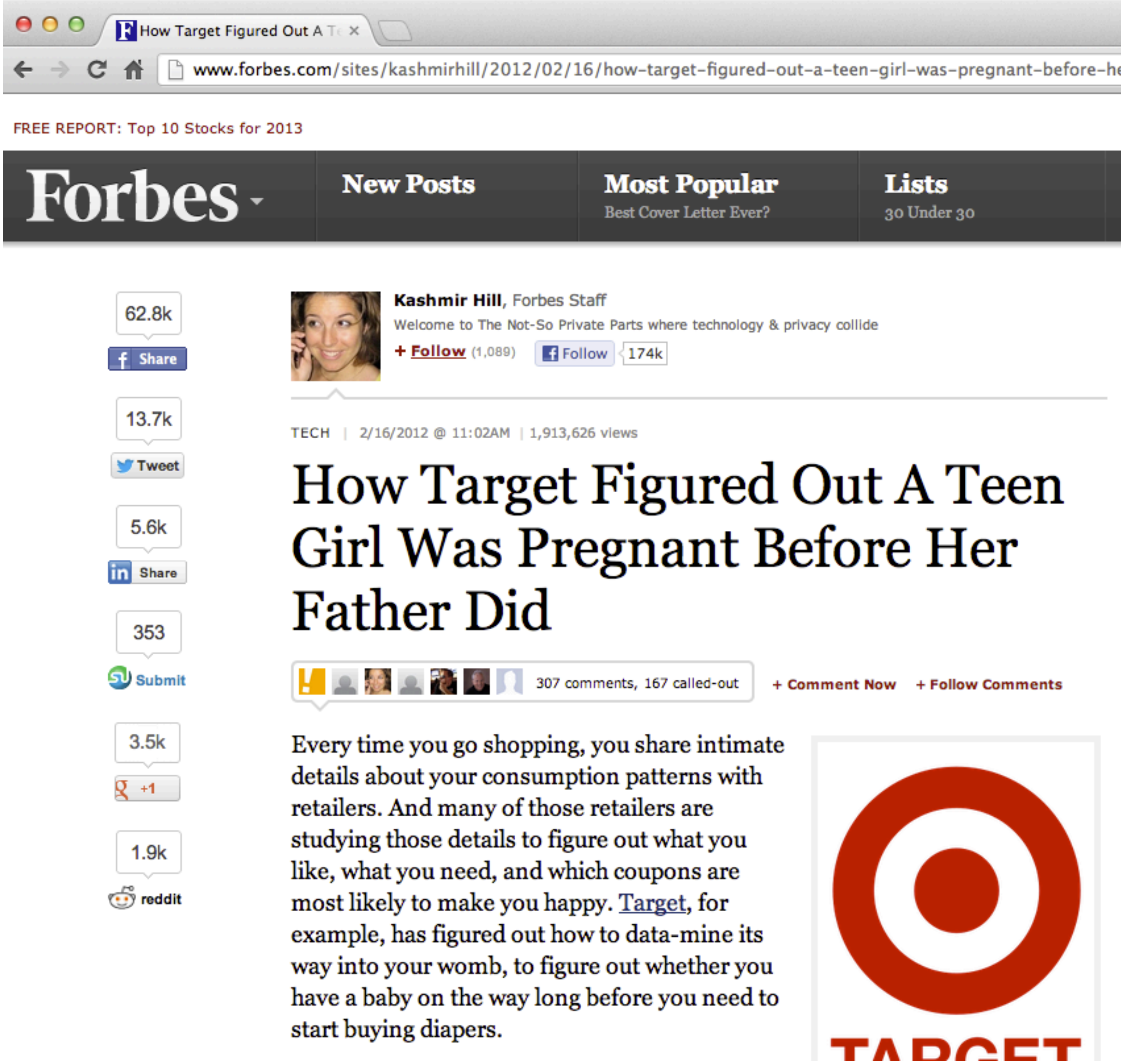
Electoral vote

225.0
-14.0 since Oct. 30

270 to win

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

Predictive analysis


How Target Figured Out A Te
www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-he
FREE REPORT: Top 10 Stocks for 2013
Forbes
New Posts
Most Popular
Best Cover Letter Ever?
Lists
30 Under 30
62.8k
Share
13.7k
Tweet
5.6k
Share
353
Submit
3.5k
+1
1.9k
reddit
+ Follow (1,089)
Follow
174k
307 comments, 167 called-out
+ Comment Now
+ Follow Comments
TARGET

Kashmir Hill, Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

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

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

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.

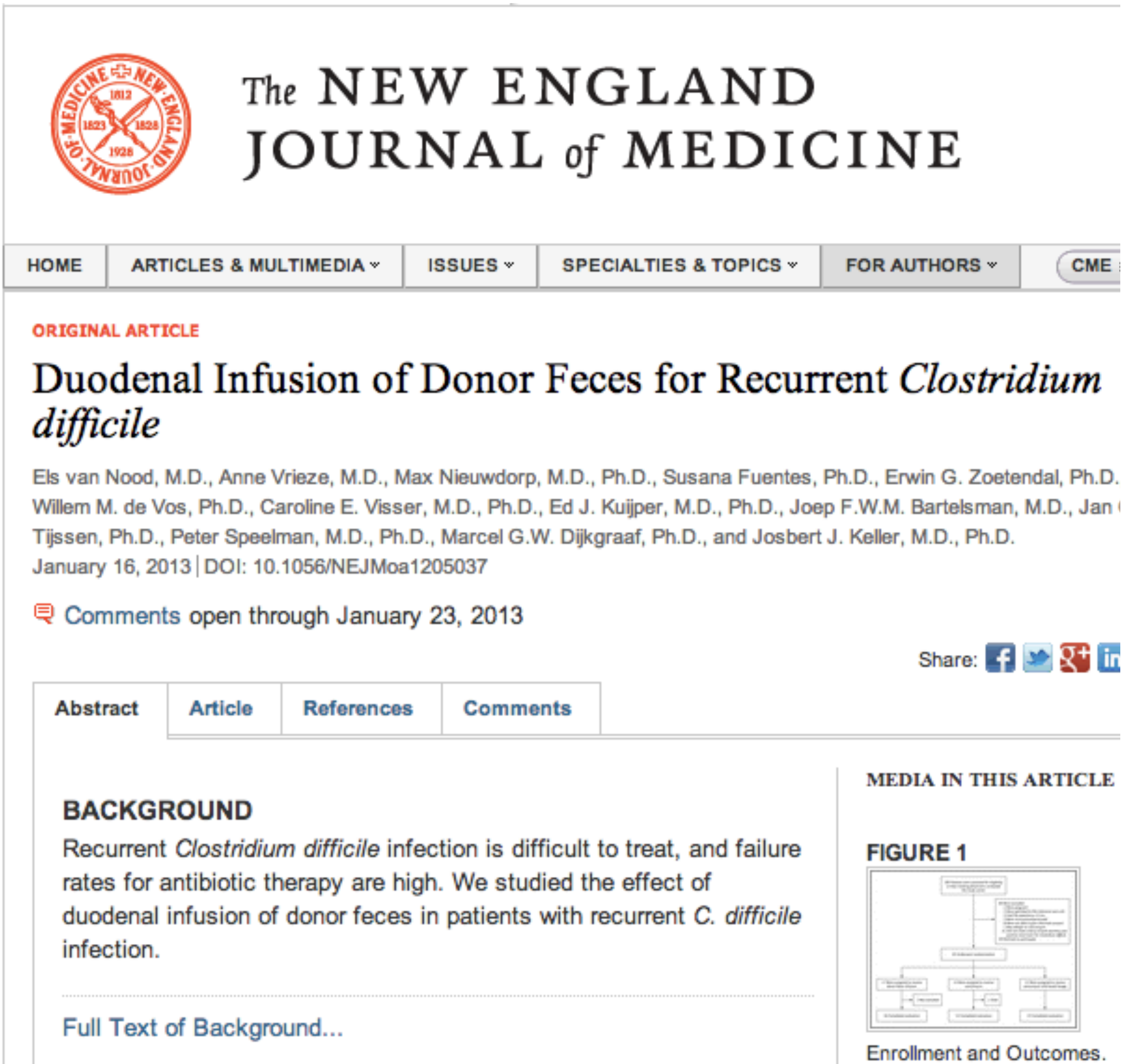
http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

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



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 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

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...

Enrollment and Outcomes.

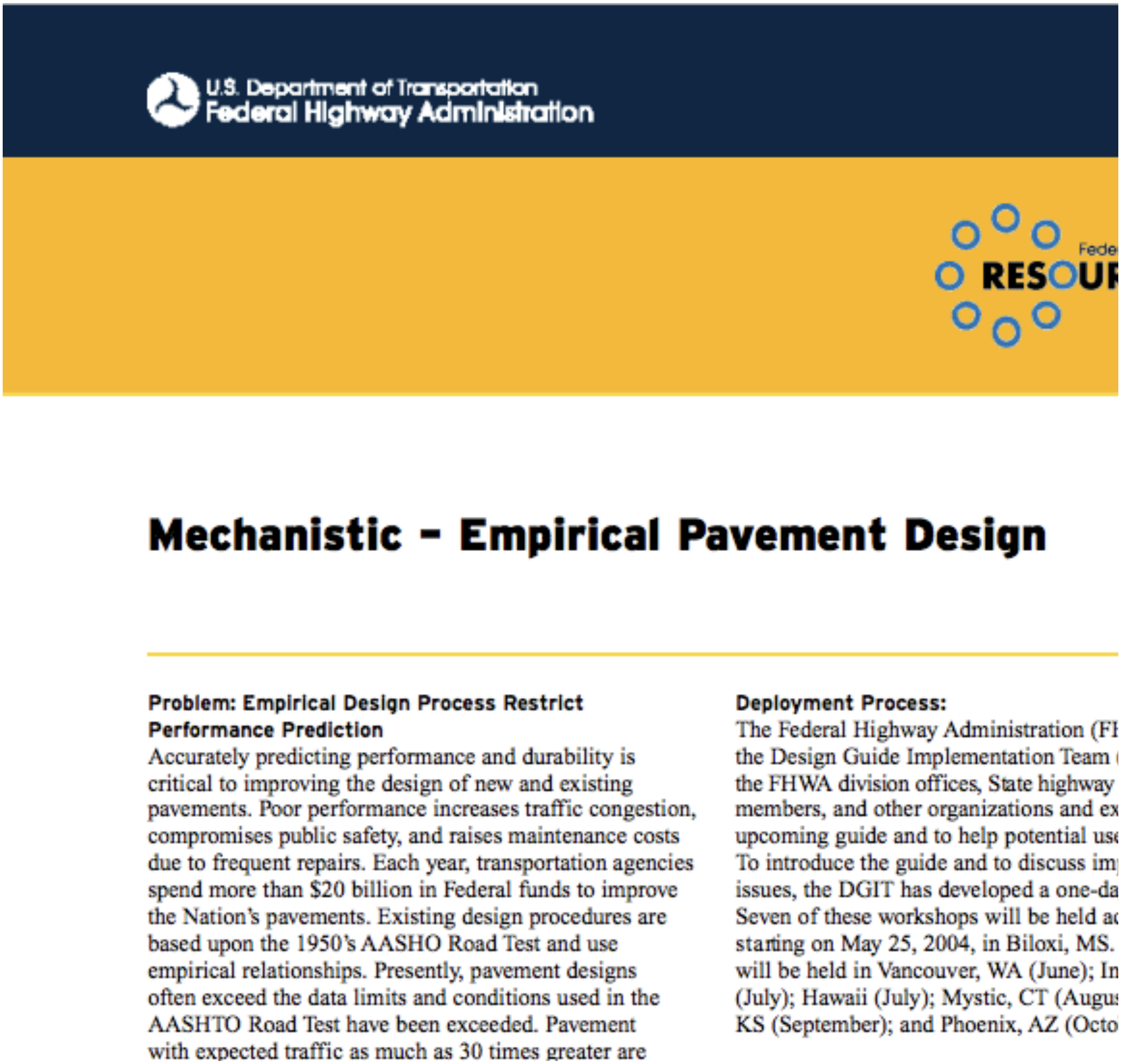
van Nood et al. (2013) NEJM

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

Fede
RESOUR

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 (FI
the Design Guide Implementation Team (
the FHWA division offices, State highway
members, and other organizations and ex
upcoming guide and to help potential use
To introduce the guide and to discuss im
issues, the DGIT has developed a one-da
Seven of these workshops will be held ac
starting on May 25, 2004, in Biloxi, MS.
will be held in Vancouver, WA (June); In
(July); Hawaii (July); Mystic, CT (Augus
KS (September); and Phoenix, AZ (Octo

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