Finding and Telling Stories with R

Andrew Flowers
andrew.w.flowers@gmail.com
@andrewflowers

About me

- 2014 to 2016: Quantitative editor/writer at FiveThirtyEight

- 2017: Freelance data scientist and journalist

andrewflowers.github.io

Sorry, the topic of this talk has been changed.


MAKE AMERICA
GREAT AGAIN

6 Types of Data Stories and How to Tell Them

Andrew Flowers
andrew.w.flowers@gmail.com
@andrewflowers

Motivation

Principles of good data journalism are

principles of good data storytelling.

So, what is this "data journalism" you speak of?

One answer: Empirical social science **on deadline**.

What do we mean by that?

- Story leads, data follows.
- Use rigorous but interpretable methods.
- Be accurate. Be fast. Be transparent.

FiveThirtyEight

TheUpshot

NICAR

theguardian

PRICEONOMICS

Why do data journalists use R?

Preaching to the choir

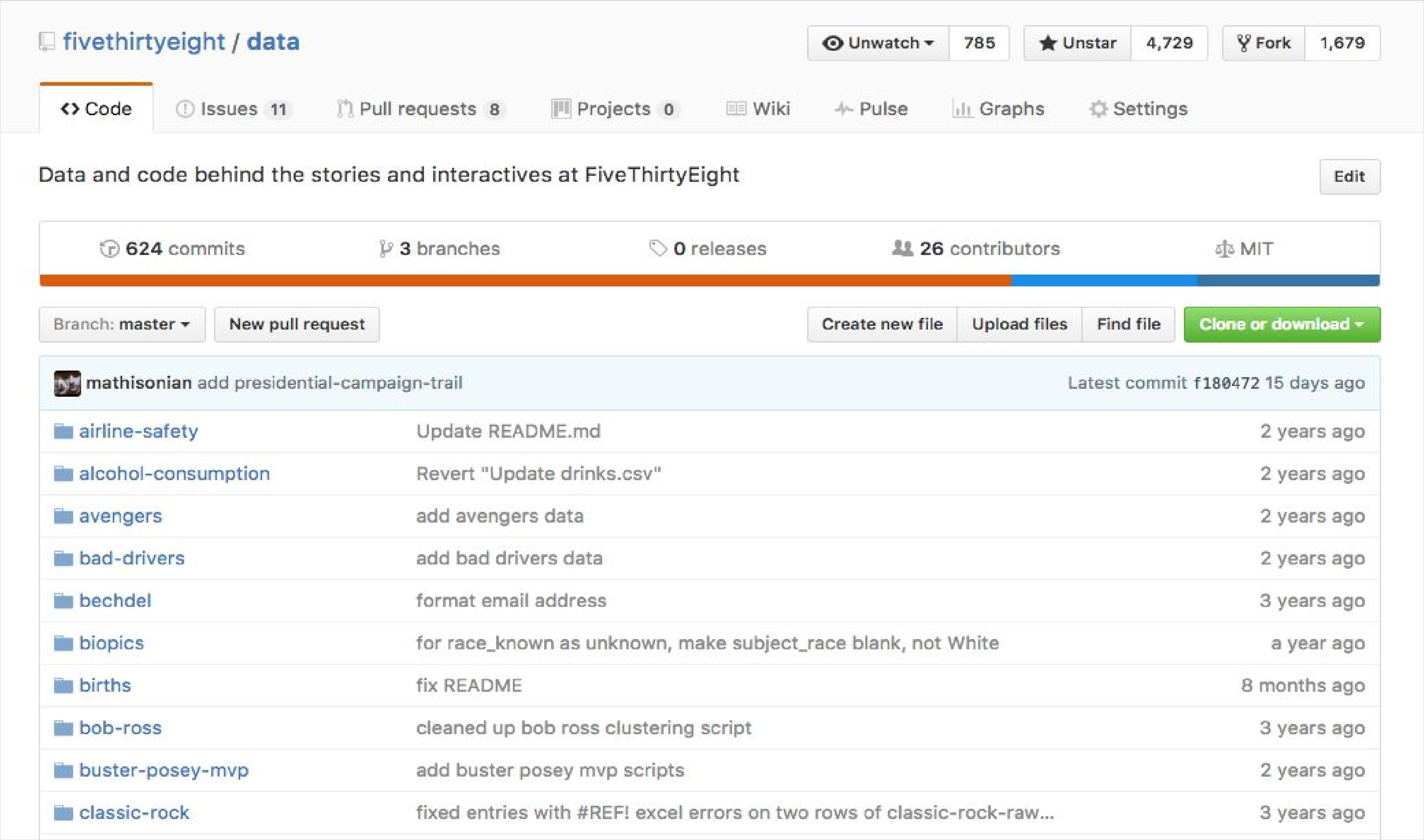
Five big reasons

1. Open source (transparency, GitHub)
2. ggplot2 (custom theme, weird charts)
3. Data wrangling (speed, handles messy data)
4. Collaboration (RStudio/Git/GitHub integration)
5. Interactives (Shiny prototypes, data processing)

Bottom line: R is the best tool for rapid, shareable data analysis

That said...some staffers still use Stata and Excel

Open source journalism: github.com/fivethirtyeight/data



BIG NEWS: fivethirtyeight R package now on CRAN

fivethirtyeight: Data and Code Behind the Stories and Interactives at 'FiveThirtyEight'

Data and code behind the stories and interactives at 'FiveThirtyEight' <https://github.com/fivethirtyeight/data>.

Version:	0.1.0
Depends:	R (≥ 3.2.4)
Suggests:	knitr, rmarkdown, dplyr, readr, ggplot2, broom, magrittr, scales, stringr, ggthemes
Published:	2017-01-09
Author:	Albert Y. Kim [cre], Chester Ismay [aut], Jennifer Chunn [aut]

Created by:

- Albert Y. Kim - Middlebury College
- Chester Ismay - Reed College/Pacific University
- Jennifer Chunn - Seattle University

~~The~~ Six Types of Data Stories

What motivates a data story?
What makes it *worth telling*?

1. Novelty
2. Outlier
3. Archetype
4. Trend
5. Debunking
6. Forecast

1. Novelty

“What kind of stories should we be suspicious of? … It's the stories that you like the most, that you find the most rewarding, the most inspiring. The stories that don't focus on opportunity cost, or the complex, unintended consequences of human action, because that very often does not make for a good story. So often a story is of triumph, of struggle; there are opposing forces, which are either evil or ignorant; there is a person on a quest, someone making a voyage, and a stranger coming to town. And those are your categories, but don't let them make you too happy.”

-- Tyler Cowen, Be suspicious of stories

Example: Uber vs. Taxis in NYC

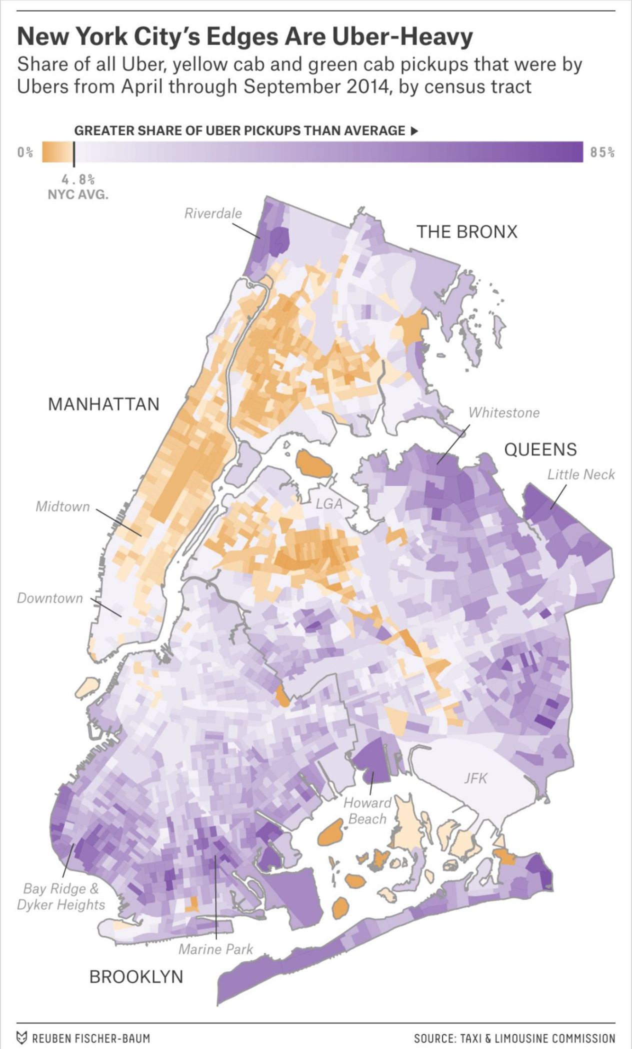
Data obtained via FOIA request:
https://github.com/fivethirtyeight/uber-tlc-foil-response

Uber TLC FOIL Response

This directory contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. Trip-level data on 10 other for-hire vehicle (FHV) companies, as well as aggregated data for 329 FHV companies, is also included. All the files are as they were received on August 3, Sept. 15 and Sept. 22, 2015.

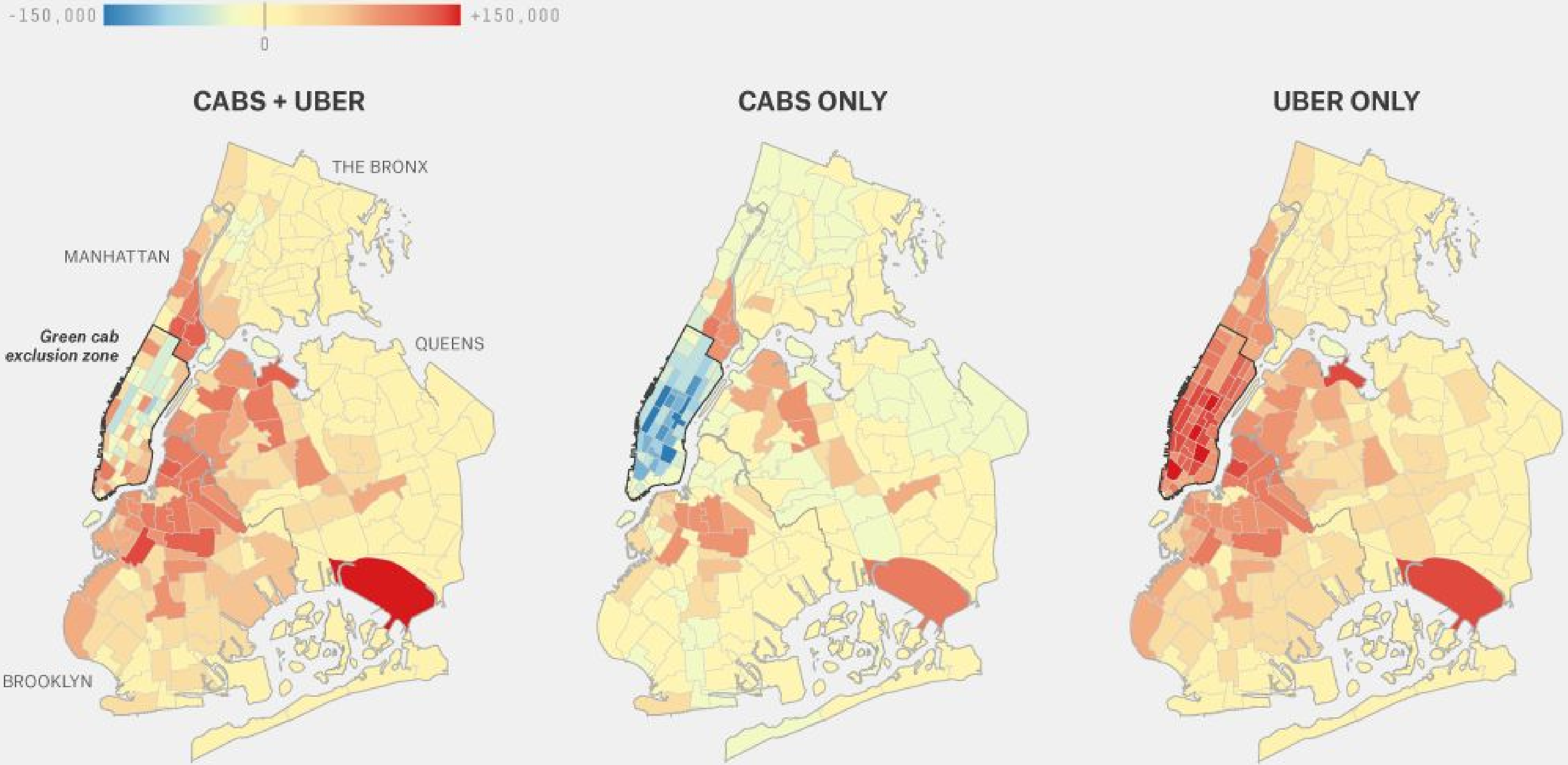
FiveThirtyEight obtained the data from the NYC Taxi & Limousine Commission (TLC) by submitting a Freedom of Information Law request on July 20, 2015. The TLC has sent us the data in batches as it continues to review trip data Uber and other HFV companies have submitted to it. The TLC's correspondence with FiveThirtyEight is included in the files `TLC_letter.pdf`, `TLC_letter2.pdf` and `TLC_letter3.pdf`. TLC records requests can be made here.

This data was used for four FiveThirtyEight stories: Uber Is Serving New York's Outer Boroughs More Than Taxis Are, Public Transit Should Be Uber's New Best Friend, Uber Is Taking Millions Of Manhattan Rides Away From Taxis, and Is Uber Making NYC Rush-Hour Traffic Worse?.



Are Ubers Supplementing Or Replacing Cabs?

Change in number of Uber and taxi pickups by taxi zone, April-June 2014 versus April-June 2015



ELLA KOEZE

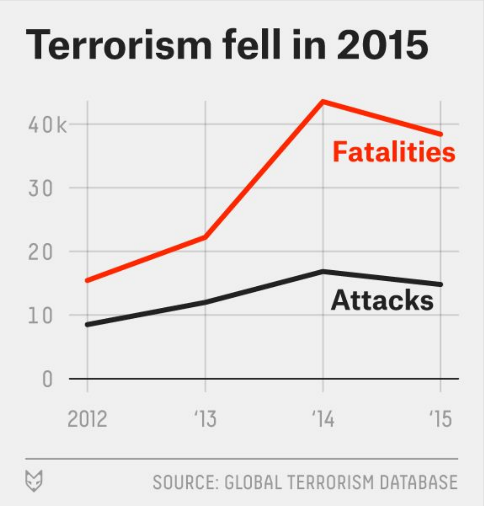
SOURCE: TAXI & LIMOUSINE COMMISSION

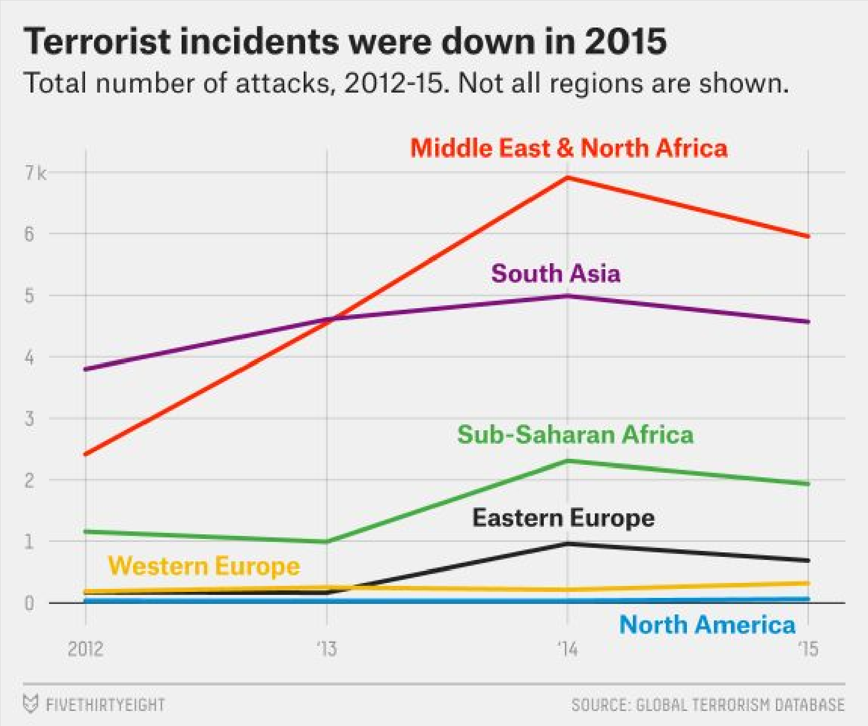
Danger: **Triviality**

Tactic: **Simple summaries**

Ask yourself: *Is the data meaningful to others?*

Example: Global terrorism trends





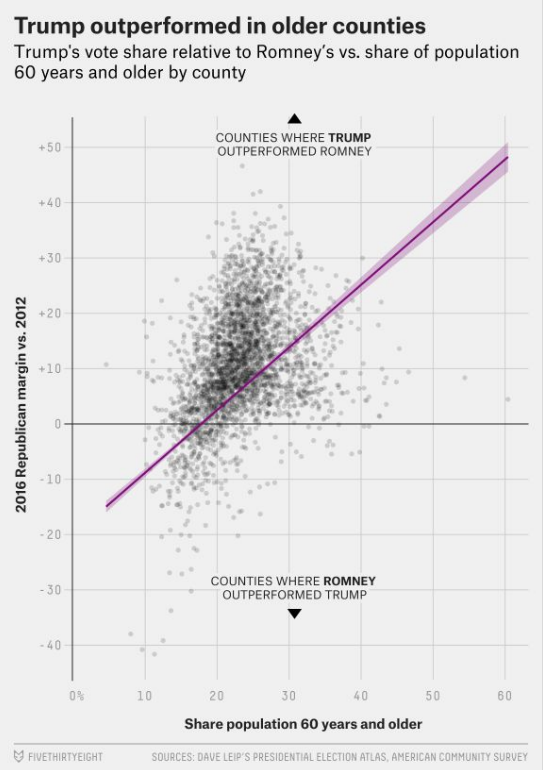
Deadliest terrorist groups in 2015

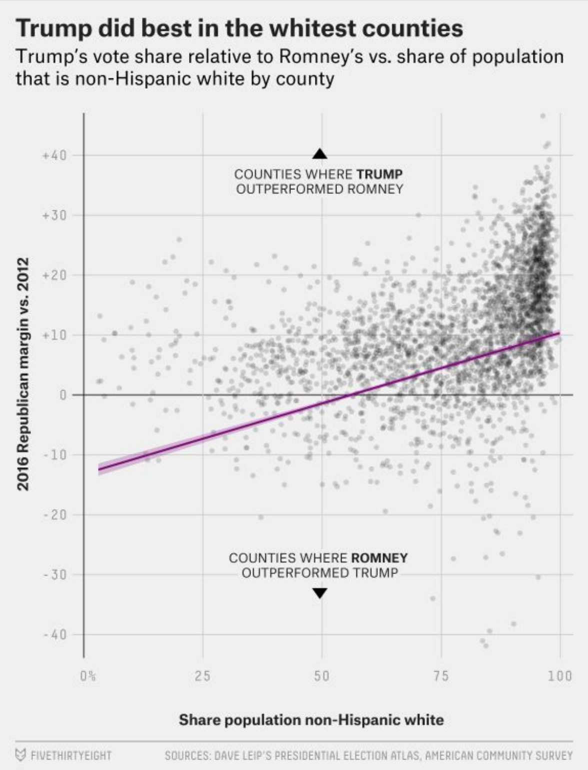
RANK	GROUP/AFFILIATION	FATALITIES
1	Islamic State	8420
2	Boko Haram	6299
3	Taliban	5215
4	Al-Shabaab	1586
5	Huthi Extremists	1306
6	Al-Nusrah Front	924
7	Sinai Province of the Islamic State	604
8	Donetsk People's Republic	597
9	Fulani Militants	572
10	Tehrik-i-Taliban Pakistan (TTP)	368

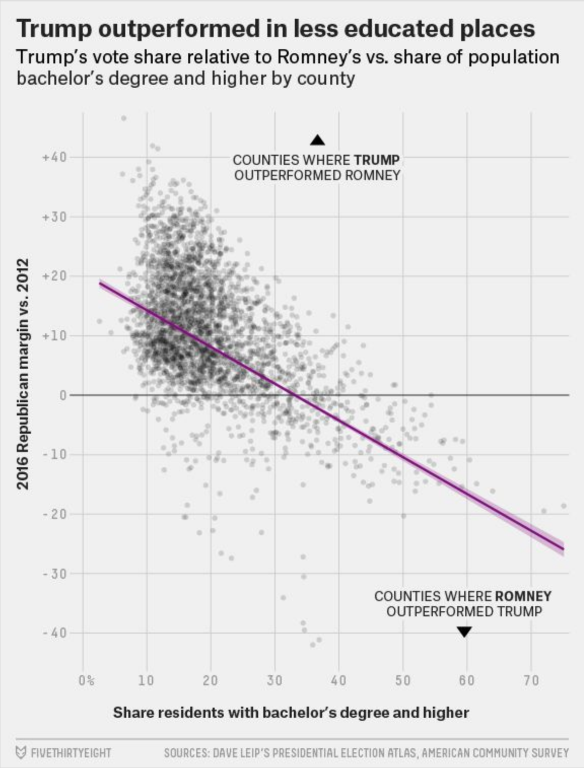
Excludes attacks where group is unknown.

SOURCE: GLOBAL TERRORISM DATABASE

Example:
Counties where Trump outperformed Romney



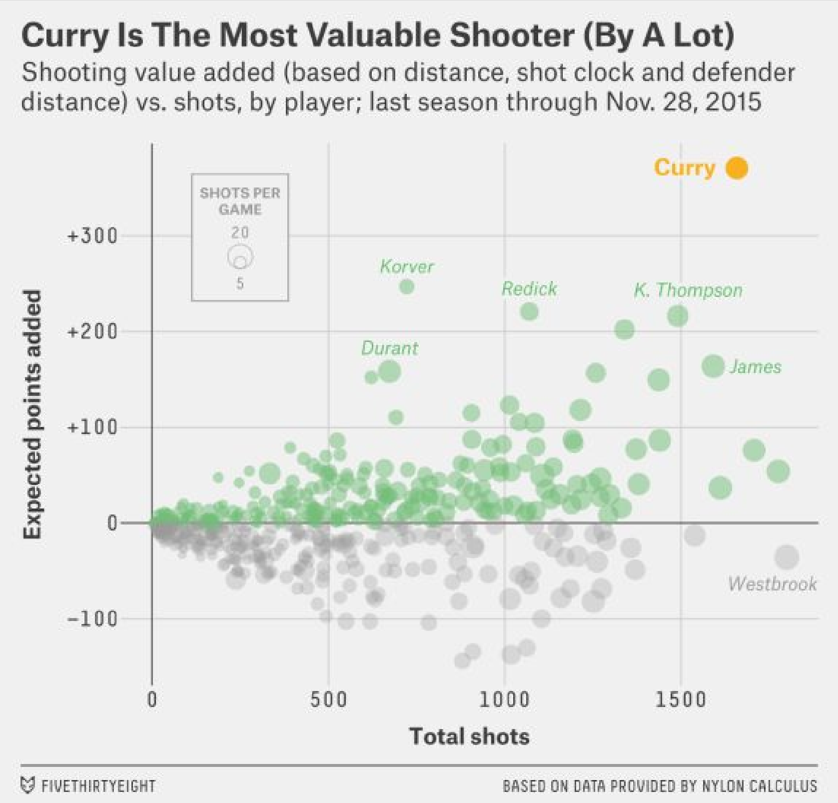


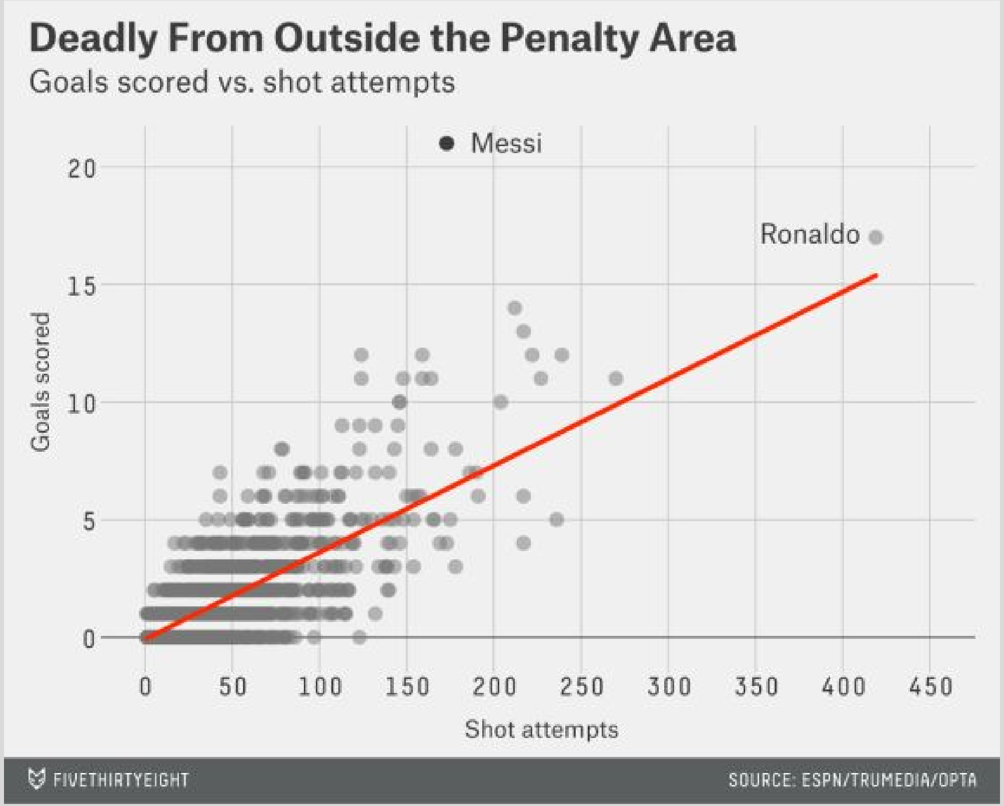



DREW ANGERER/GETTY IMAGES

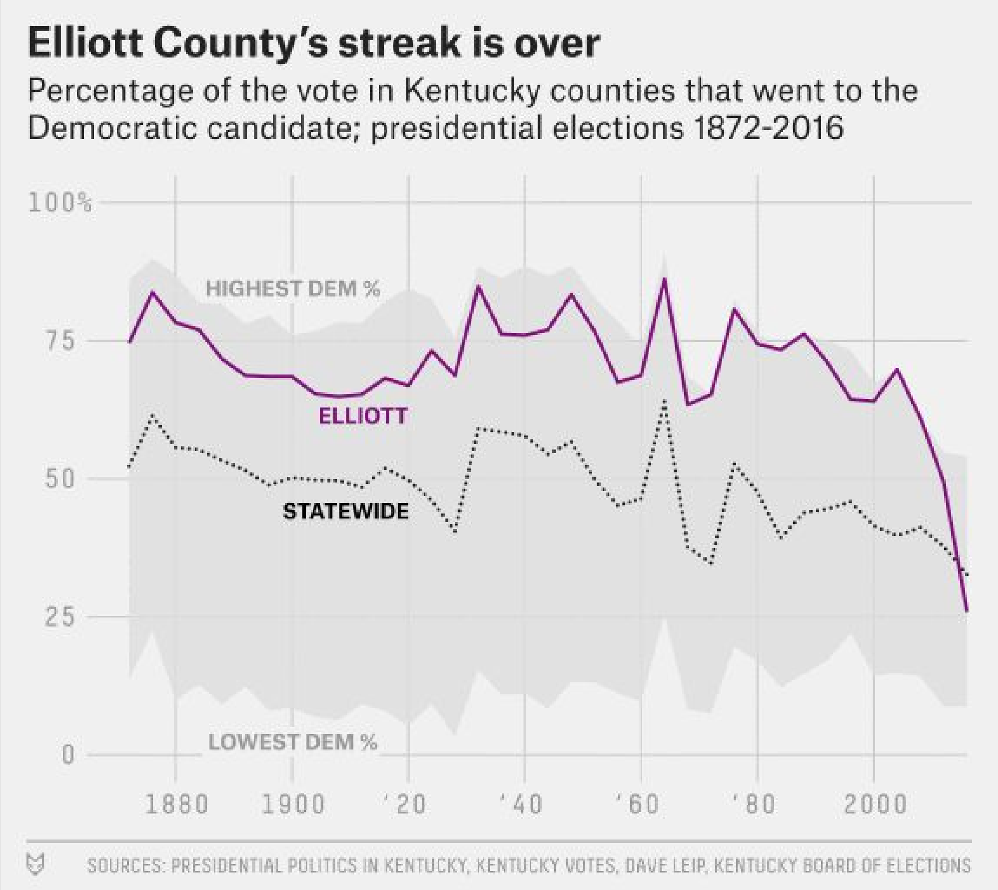
2. Outlier

Example: Stephen Curry and Lionel Messi





Example: Elliott County, Kentucky



Example: Lucrative college majors

Students walk across the campus of UCLA in Los Angeles. KEVORK DJANSEZIAN / GETTY IMAGES

SEP 12, 2014 AT 7:37 AM

The Economic Guide To Picking A College Major

By Ben Casselman

RECOMMENDED

Senate 2016: The Democrats Strike Back

Why Pennsylvania Could Decide The 2016 Election

MAJOR	# OF MAJORS	EARNINGS (x1,000) MED.	25TH	75TH	% WORKING IN JOBS PART-TIME	NON-COLLEGE	LOW-PAYING
Petroleum Eng.	2,339	$110	$95	$125	13	19	10
Mining & Mineral Eng.	756	$75	$55	$90	23	42	8
Metallurgical Eng.	856	$73	$50	$105	19	28	0
Naval Architecture & Marine Eng.	1,258	$70	$43	$80	12	16	0
Chemical Eng.	32,260	$65	$50	$75	18	20	4
Nuclear Eng.	2,573	$65	$50	$102	11	37	13

Branch: **master** ▾ **data** / **college-majors** / **college-majors-rscript.R** Find file Copy path

BenCasselman Create college-majors-rscript.R 613c858 on Sep 11, 2014

1 contributor

166 lines (148 sloc) | 9.17 KB Raw Blame History

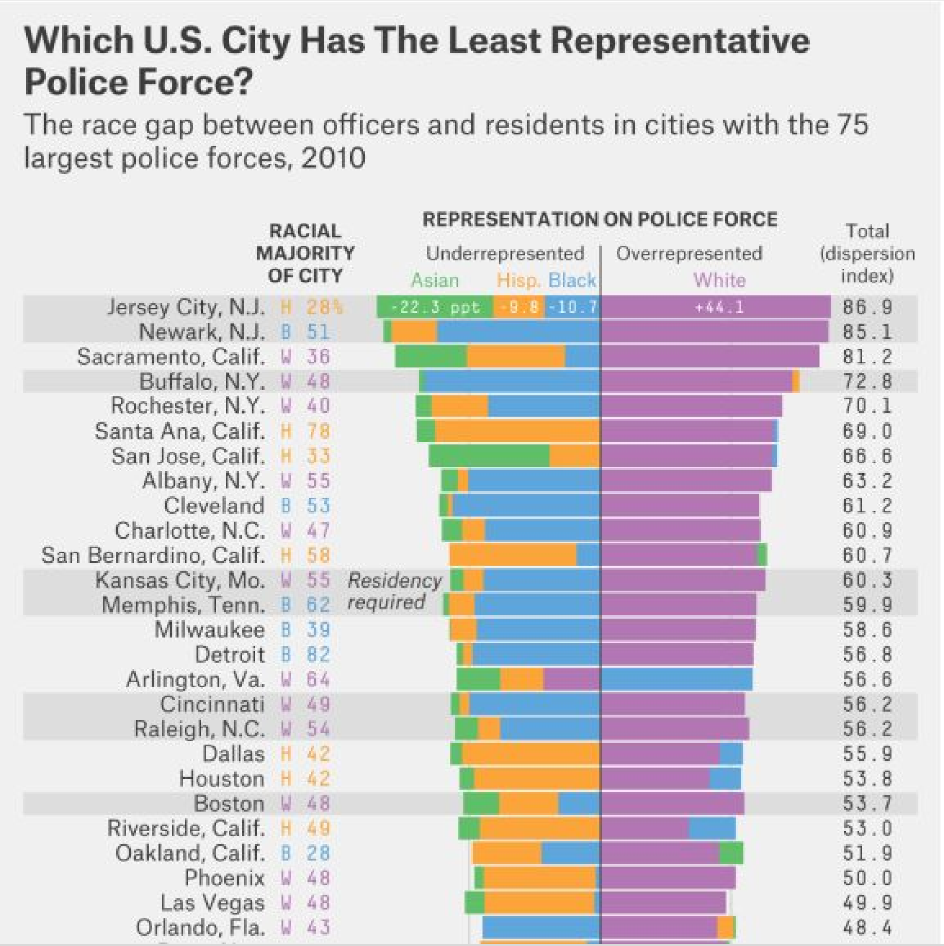
```r
###############################################################
#
# COLLEGE MAJORS AND EARNINGS
# This is the code used to generate data for FiveThirtyEight's
# story on earnings by college major.
# Analysis is based off the 2010-2012 American Community Survey
# microdata.
# Download data here: http://www.census.gov/acs/www/data_documentation/pums_data/
# Documentation here: http://www.census.gov/acs/www/data_documentation/pums_documentation/

# First download data and select records for which college major
# (variable FOD1P) is present. Save into data frame as MAJORS1012.
# Also download github_majorslist.csv

require(dplyr)

load("MAJORS1012")
MajorsList <- read.csv("github_majorslist.csv",header=TRUE,stringsAsFactors=FALSE)

working <- merge(MAJORS1012,MajorsList,by="FOD1P") # Check merge properly
MAJORS1012  <- working
rm(working)
```

Danger: **Spurious result**

Tactic: **Characters**

Ask yourself: *Is this really so different?*

Example: Police residency requirements

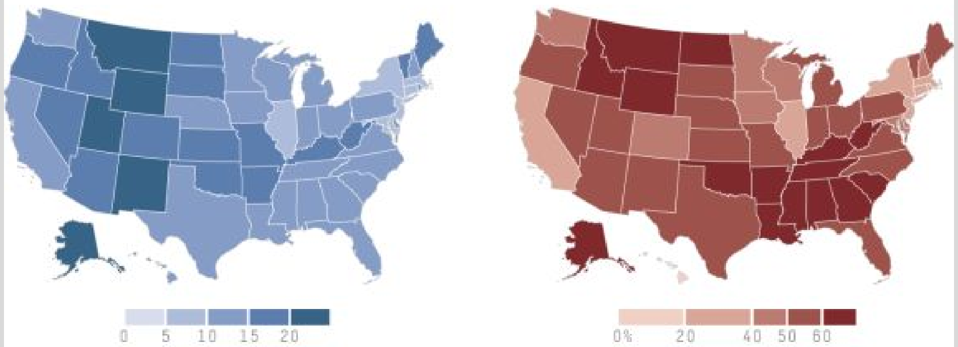


Example: Wyoming suicides

The West has high rates of gun suicide

Suicides per 100,000*

Share of suicides by firearm



*Age-adjusted

FIVETHIRTYEIGHT

SOURCE: CENTERS FOR DISEASE CONTROL AND PREVENTION

Suicides per 100,000 **by age group**

AGE	SUICIDE RATE
5-14 years old	0.9
15-24	11.2
25-34	14.8
35-44	16.4
45-54	19.9
55-64	18.3
65-74	14.9
75-84	17.1
85+	18.6

SOURCE: CENTERS FOR DISEASE CONTROL AND PREVENTION

Suicides per 100,000 by **race and ethnicity**

RACE	SUICIDE RATE
White	16.0
American Indian or Alaska Native	11.1
Asian or Pacific Islander	6.0
Hispanic or Latino	5.9
Black or African-American	5.5

Age adjusted

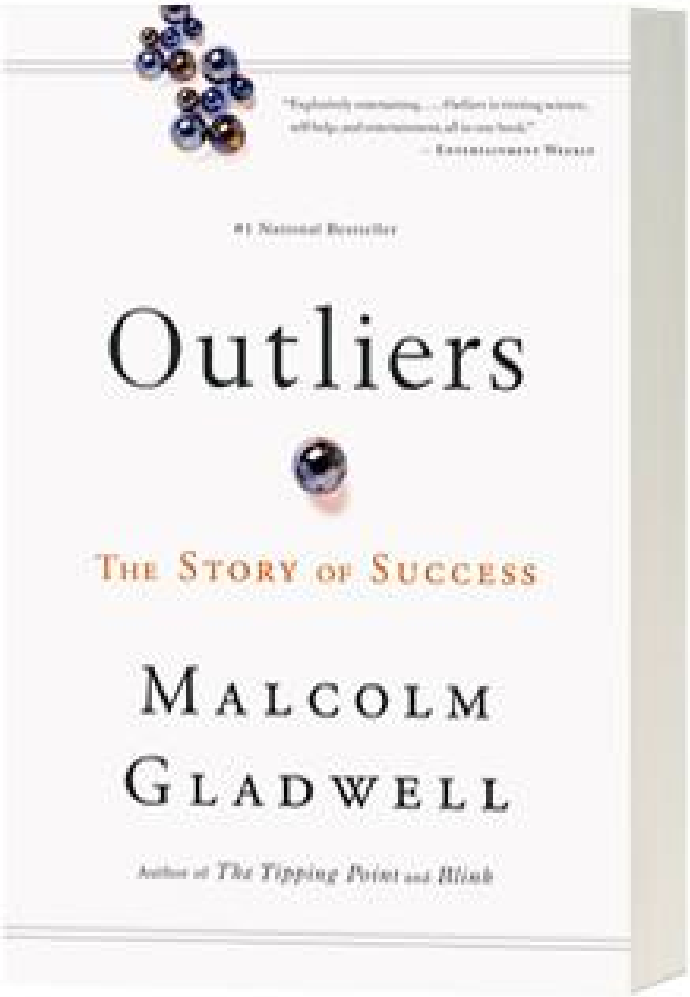
SOURCE: CENTERS FOR DISEASE CONTROL AND PREVENTION

Suicides per 100,000 **by gender**

AGE	SUICIDE RATE
Male	20.4
Female	5.5

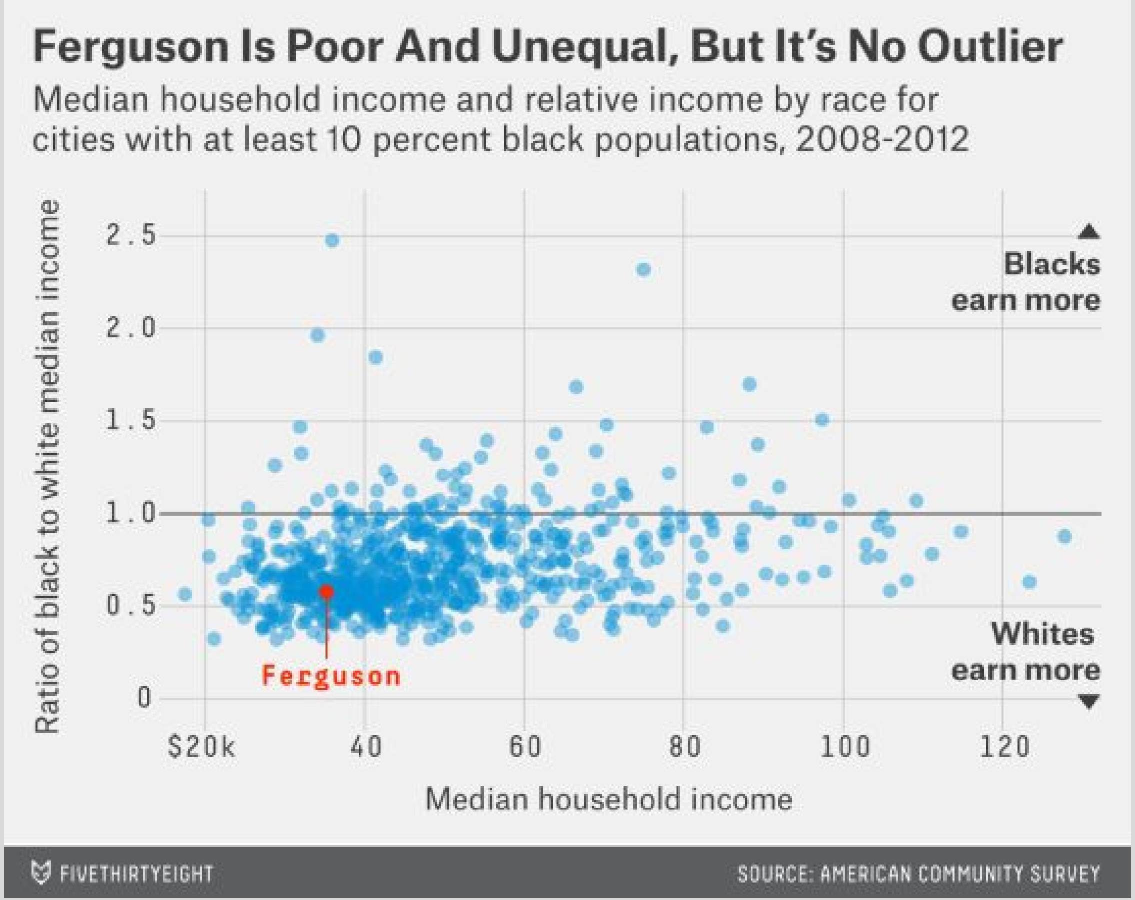
Age adjusted

SOURCE: CENTERS FOR DISEASE CONTROL AND PREVENTION


#1 National Bestseller
Outliers
The Story of Success
Malcolm
Gladwell
Author of The Tipping Point and Blink

3. Archetype

Example: Ferguson, Missouri



Danger: **Oversimplification**

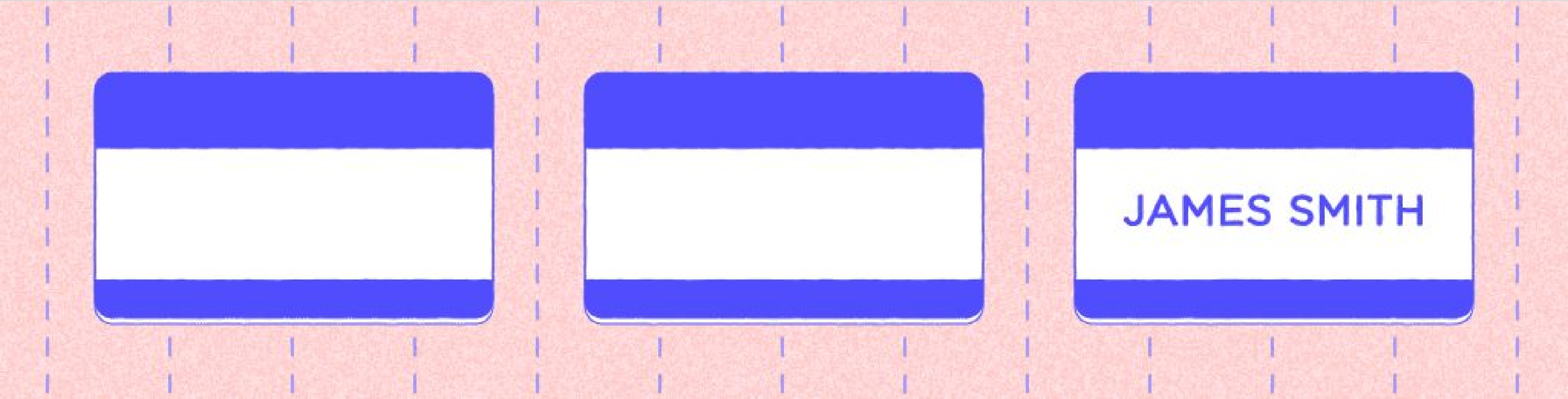
Tactic: **Modeling**

Ask yourself: *What variables am I leaving out?*

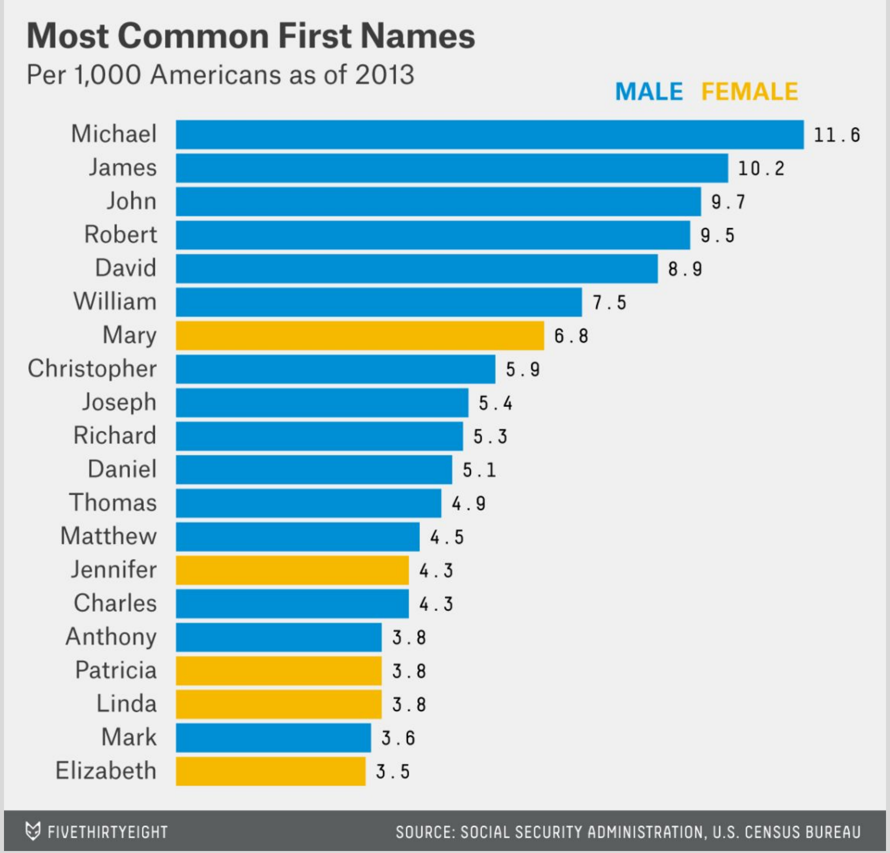
“Everything should be made as simple as possible, but not simpler.”

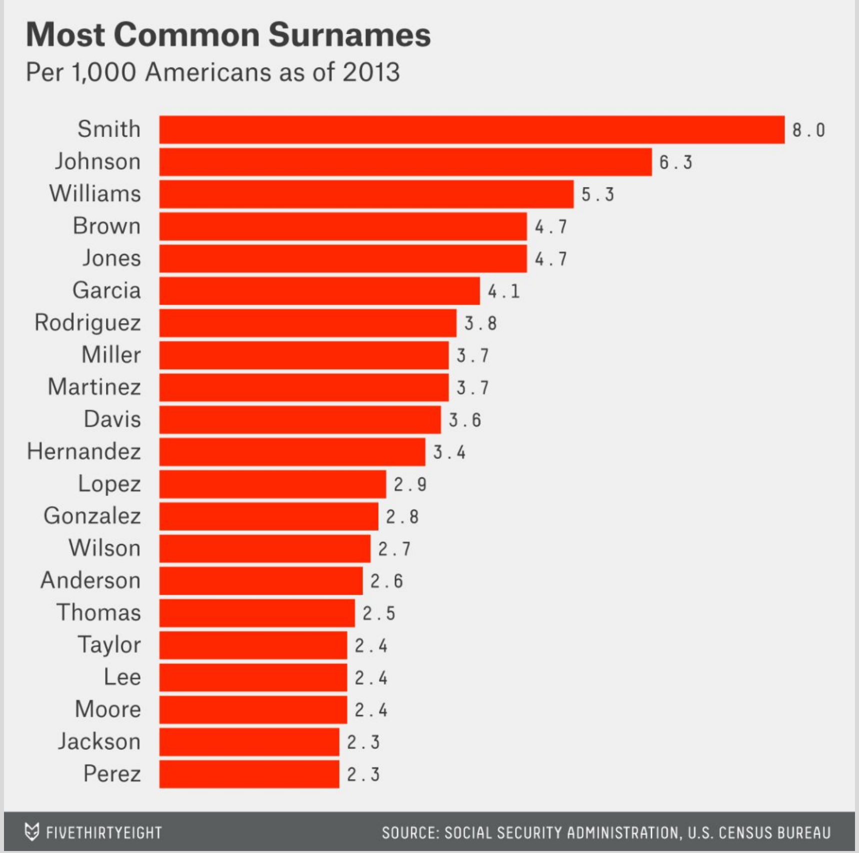
-- Albert Einstein *(maybe?)*

Example: What’s the most common name in America?



It’s Michael Smith, right?

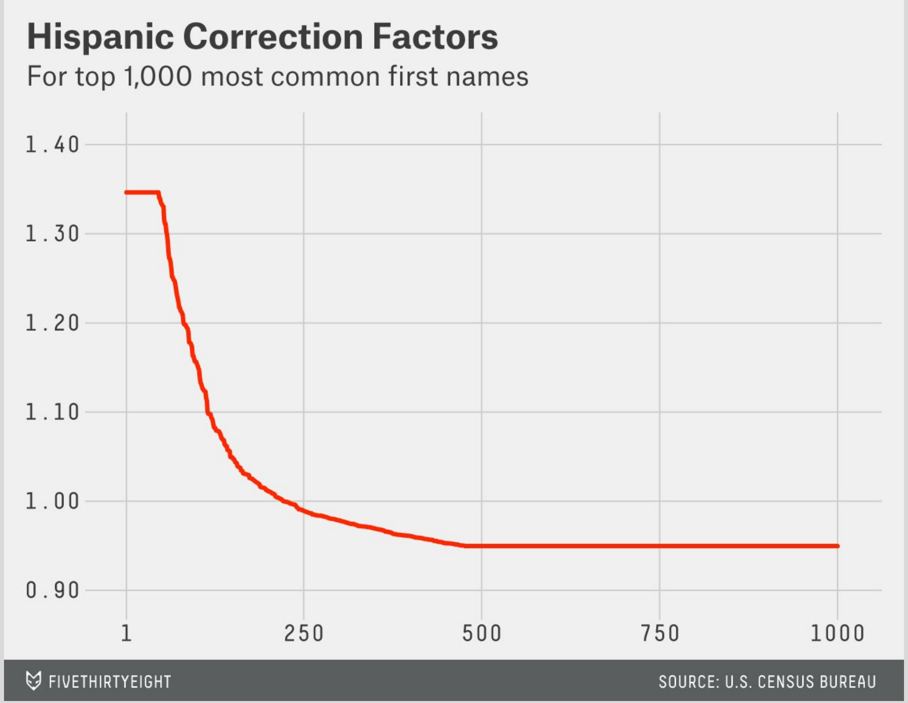




But names aren't consistently correlated

	Miller	Anderson	Martin	Smith	Thompson	Wilson	Moore	White	Taylor	Davis	Johnson	Brown	Jones	Thomas	Williams	Jackson	Lee	Garcia	Martinez	Rodriguez
John	8.4	6.9	13.4	-23.5	3.6	2.5	3.6	1.7	0.4	-3.8	-35.2	-34.3	-26.5	5.1	-11.3	-27.7	-8.3	-66.4	-63.6	-69.5
Michael	18.1	2.4	13.0	5.6	6.7	2.6	15.7	-1.7	0.6	2.5	-4.8	4.2	-2.4	2.9	-8.1	-16.8	-9.8	-54.7	-54.4	-63.2
James	20.3	24.0	35.9	21.1	33.6	40.9	43.0	41.0	37.4	32.8	9.6	25.4	8.9	24.8	20.7	19.8	12.2	-84.4	-83.9	-87.4
Robert	29.6	27.0	16.8	8.7	18.0	22.0	17.2	18.1	39.2	7.2	5.9	14.2	12.6	4.8	-1.3	-1.8	12.1	-57.4	-55.1	-61.0
David	27.9	28.1*	15.8	10.8	11.0	12.6	8.8	16.2	-1.5	-44.7	3.5	10.3	5.0	0.9	-7.3	-15.4	10.1	-30.5	-28.5	-31.7
Mary	21.5	18.2	26.2	14.7	18.5	18.9	22.8	22.5	16.0	14.8	11.7	13.1	12.2	18.9	13.1	8.7	-16.4	-42.4	-41.3	-54.0
William	22.5	7.0	33.3	16.6	29.3	21.5	31.9	38.2	29.2	23.1	4.3	22.3	15.4	19.8	-59.1	9.6	-11.4	-82.3	-80.2	-74.2
Richard	24.1	22.9	12.6	6.2	8.4	1.7	7.3	12.2	5.0	5.2	-2.0	3.9	-4.6	-2.3	-17.0	-21.6	-6.6	-43.2	-43.0	-47.7
Thomas	12.9	3.0	22.1	-2.6	-3.5	9.7	32.2	13.2	4.3	-5.9	-15.0	1.8	-3.5	-81.1	-11.5	-11.0	-8.9	-77.8	-75.6	-84.2
Jennifer	24.9	17.3	22.5	16.7	15.2	14.8	11.5	14.8	14.9	8.7	10.3	6.5	24.0	-1.6	-7.4	-6.3	22.5	-32.9	-32.2	-33.2
Patricia	20.9	10.7	25.3	17.1	16.7	6.7	19.0	17.2	19.4	11.8	7.1	15.2	11.2	15.9	10.6	8.8	-22.9	-16.0	-16.9	-23.3
Joseph	-8.9	-34.8	0.1	-27.6	-24.8	-26.1	-20.4	-10.6	-23.9	-26.7	-33.9	-23.9	-39.0	-7.5	-26.4	-26.0	-26.4	-45.8	-40.3	-52.8
Linda	34.0	25.8	29.6	27.5	24.8	24.4	29.3	26.1	28.6	24.4	22.9	19.3	21.1	17.4	15.3	16.1	6.5	-49.2	-50.0	-56.5
Maria	-77.6	-77.3	-52.1	-78.8	-76.4	-77.4	-79.3	-77.7	-80.3	-79.7	-78.3	-78.5	-79.4	-75.6	-79.4	-80.9	-77.5	663.9	614.1	639.8
Charles	36.3	25.7	29.3	38.4	38.1	45.7	49.0	43.4	41.2	43.6	24.8	39.9	30.5	37.0	33.3	29.0	4.6	-83.7	-84.5	-88.5
Barbara	35.5	25.3	25.1	24.0	20.6	25.8	24.6	25.2	24.7	24.2	20.3	26.7	16.2	16.6	14.8	18.5	-18.7	-66.9	-67.5	-67.9
Mark	42.1	45.9	-38.5	4.1	32.4	22.0	-23.1	-16.8	11.0	7.3	14.7	-14.0	-16.7	2.5	-7.3	-23.4	-33.6	-67.0	-63.0	-71.3
Daniel	19.9	-19.1	10.3	-5.2	-6.4	-13.9	-8.5	-5.0	21.3	-22.2	-17.8	-15.5	-25.1	-25.3	-28.7	-32.3	-1.3	35.1	33.2	24.0
Susan	32.1	28.4	15.8	3.9	5.4	3.5	3.5	-0.9	3.0	-5.8	-6.5	-4.6	-17.1	-1.5	-24.2	-31.9	3.7	-68.2	-68.0	-71.7
Elizabeth	13.3	7.4	13.5	1.7	9.1	7.9	8.4	4.0	1.7	-2.4	-3.7	-3.4	-5.5	-3.3	-13.9	-18.2	-20.2	16.3	23.4	22.5

It’s James Smith, but Maria Garcia is rising



Branch: master ▾ data / most-common-name / **most-common-name.R**

Find file | Copy path

andrewflowers most common name data and script — bef7964 on Nov 20, 2014

1 contributor

218 lines (157 sloc) | 10.5 KB

Raw | Blame | History

```r
#### Story: "Dear Mona, What’s The Most Common Name In America?"
#### Url: http://fivethirtyeight.com/features/whats-the-most-common-name-in-america/
#### Authors: Mona Chalabi (Mona.Chalabi@fivethirtyeight.com) and Andrew Flowers (andrew.flowers@fivethirtyeight.com)

require(babynames)
require(dplyr)
require(reshape2)
require(zoo)
require(datasets)

# Census population parameters
pop2000 <- 276059000 # year 2000 population
pop2013 <- 316128839 # year 2013 population

hispPopShare <- .171 # Hispanic share of overall population
foreignPopShare <- .127 # Foreign-born share of overall population

# Census growth rates by racial categories, 2000-2013
whiteGrowth  <- 1.01155164
blackGrowth  <- 1.138799977
aianGrowth  <- 1.110695106
asianGrowth  <- 1.553975166
twoRaceGrowth  <-  1.817182595
hispGrowth <- 1.531490233

```

4. Trend

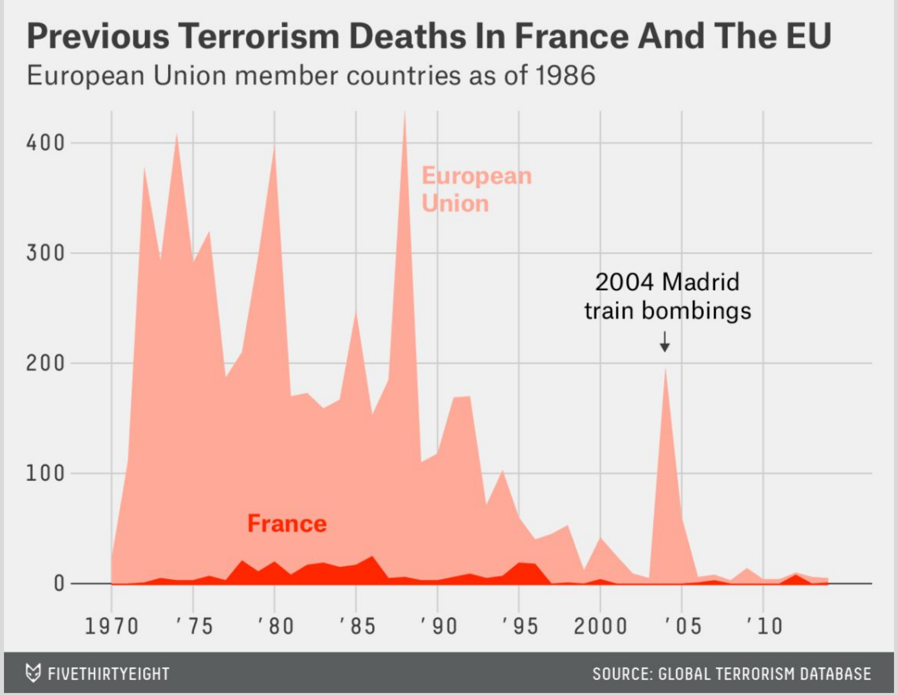
Example: Contextualize breaking news #1

NOV 14, 2015 AT 6:26 PM

The Rise Of Terrorism Inspired By Religion In France

By Carl Bialik

Filed under Terrorism



```
##### France analysis #####

# Look at incidents in France
france <- rawData %>% filter(country_txt=="France")
dim(france)

# Incidents by year -- 1993 is missing from this data
table(france$iyear, useNA="ifany")
france %>% group_by(iyear) %>%
  summarize(incidents=n()) %>%
  arrange(desc(iyear))

# Fatalities by yeaer
fraFatByYear <- france %>% group_by(iyear) %>%
  summarize(fatalities=sum(nkill, na.rm=T)) %>%
  arrange(desc(fatalities))

# Add in 1993 data -- 5 fatalities in France
fraFatByYear <- rbind(fraFatByYear,
                      data.frame(iyear=1993,
                                 fatalities=stats1993[match("France", stats1993$Country),]$`Number Killed`))

# Analysis: France had 274 fatalities from terrorism incidents between 1972 and 2014.
fraFatByYear
sum(fraFatByYear$fatalities, na.rm=T)
```

Danger: **Variance**

Tactic: **Be conservative**

Ask yourself: *Is this signal or noise?*

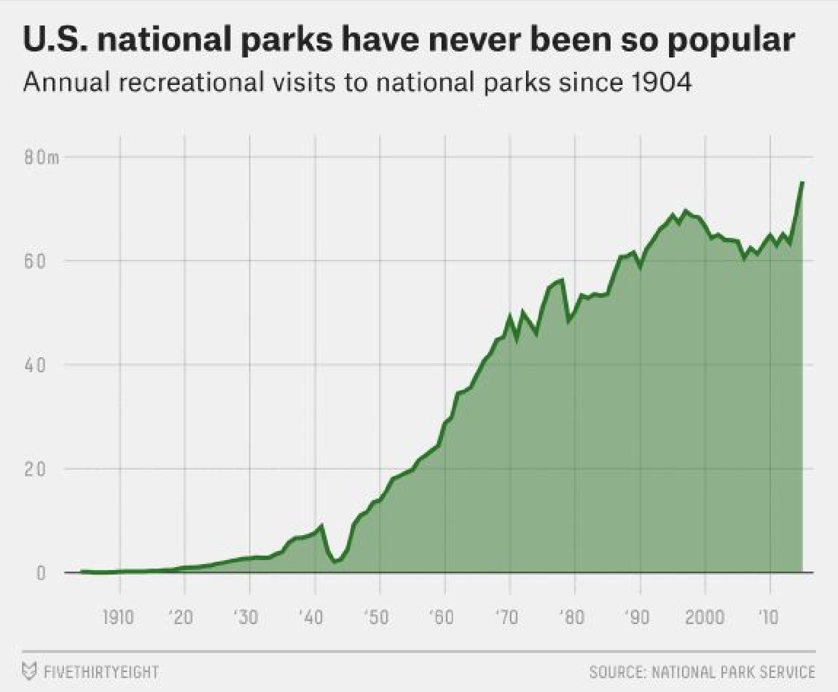
Example: Contextualize breaking news #2

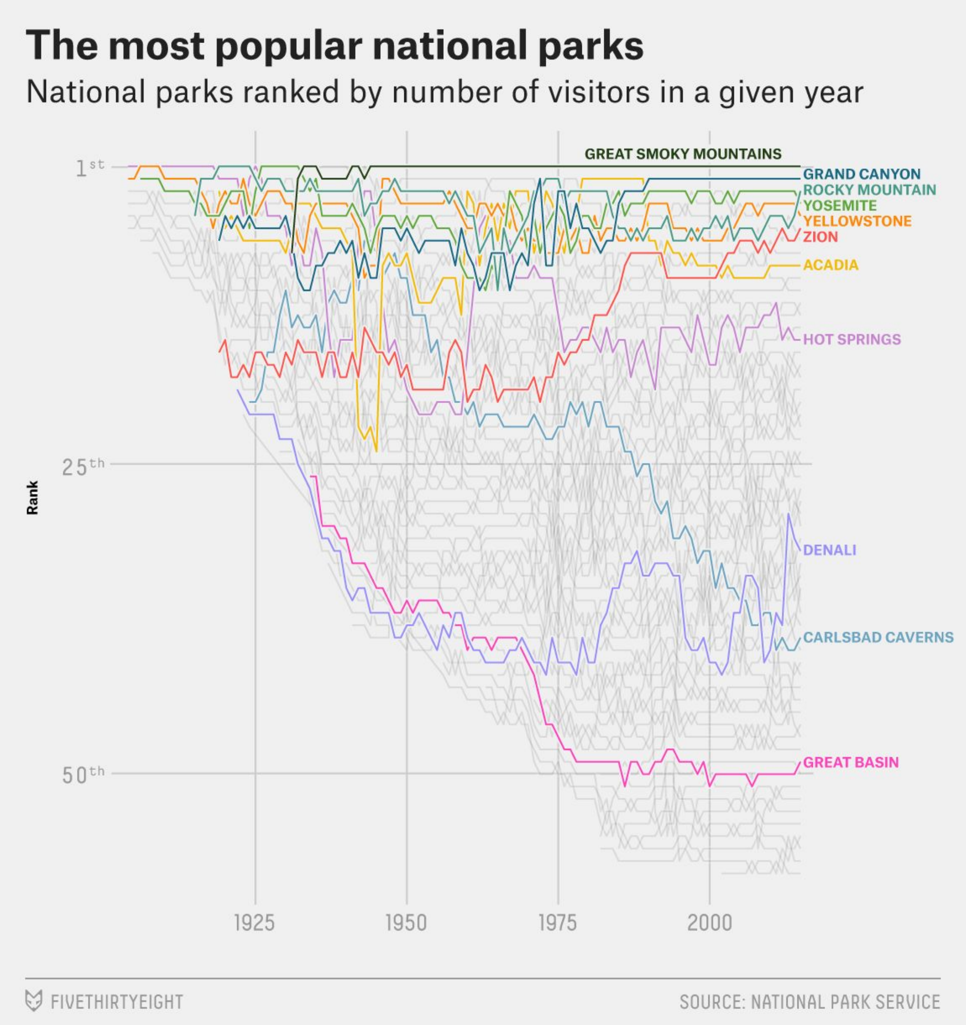
JUL 8, 2016 AT 4:29 PM

The Dallas Shooting Was Among The Deadliest For Police In U.S. History


```r
# Dallas shooting scraping

library(dplyr)
library(rvest)
library(readr)
library(tidyr)
library(lubridate)
library(stringr)
library(ggplot2)

yearly_url <- 'https://www.odmp.org/search/year/'

years <- seq(1791, 2016)

all_data <- data.frame()

# Scrape data

for (year in years){

  new_url <- paste0(yearly_url, year)

  selector_yearly <- 'p , p a'

  raw_data <- read_html(new_url) %>%
    html_nodes(selector_yearly) %>%
    html_text() %>%
    as.data.frame()

  names(raw_data) <- c("incident")

  clean_data <- raw_data %>%
    separate(incident, sep = '\n\t\t\t\t\t\t\t\t\t',
             into = c("person", "dept","eow", "cause")) %>%
    filter(!is.na(dept))

  all_data <- rbind(all_data, clean_data)
```

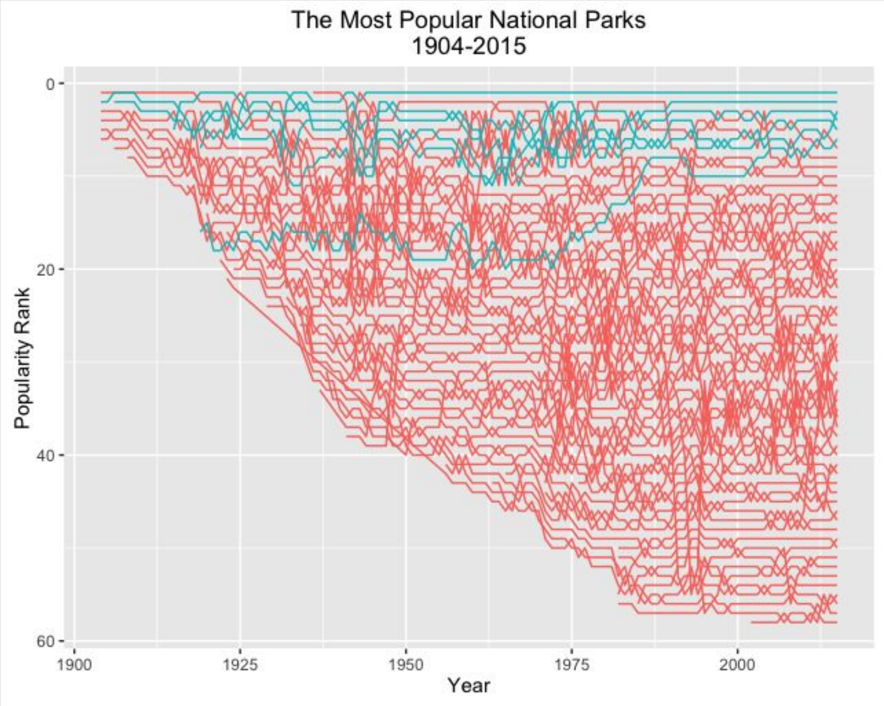
Example:
National parks visitation since 1904

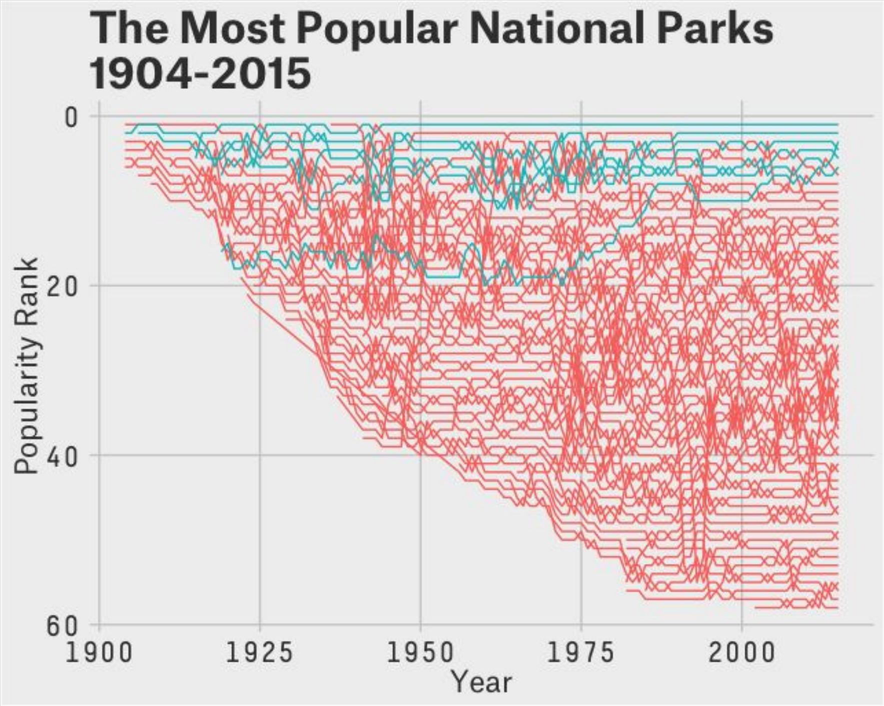




Custom ggplot theme

theme538 PRIVATE





5. Debunking

Example: Bechdel test

A Walmart employee puts Lionsgate's "The Hunger Games: Catching Fire" Blu-ray Combo Pack and DVD on the rack prior to the midnight release at Walmart on March 6, 2014 in Orange, California. JEROD HARRIS / GETTY IMAGES

APR 1, 2014 AT 1:52 PM

The Dollar-And-Cents Case Against Hollywood's Exclusion of Women

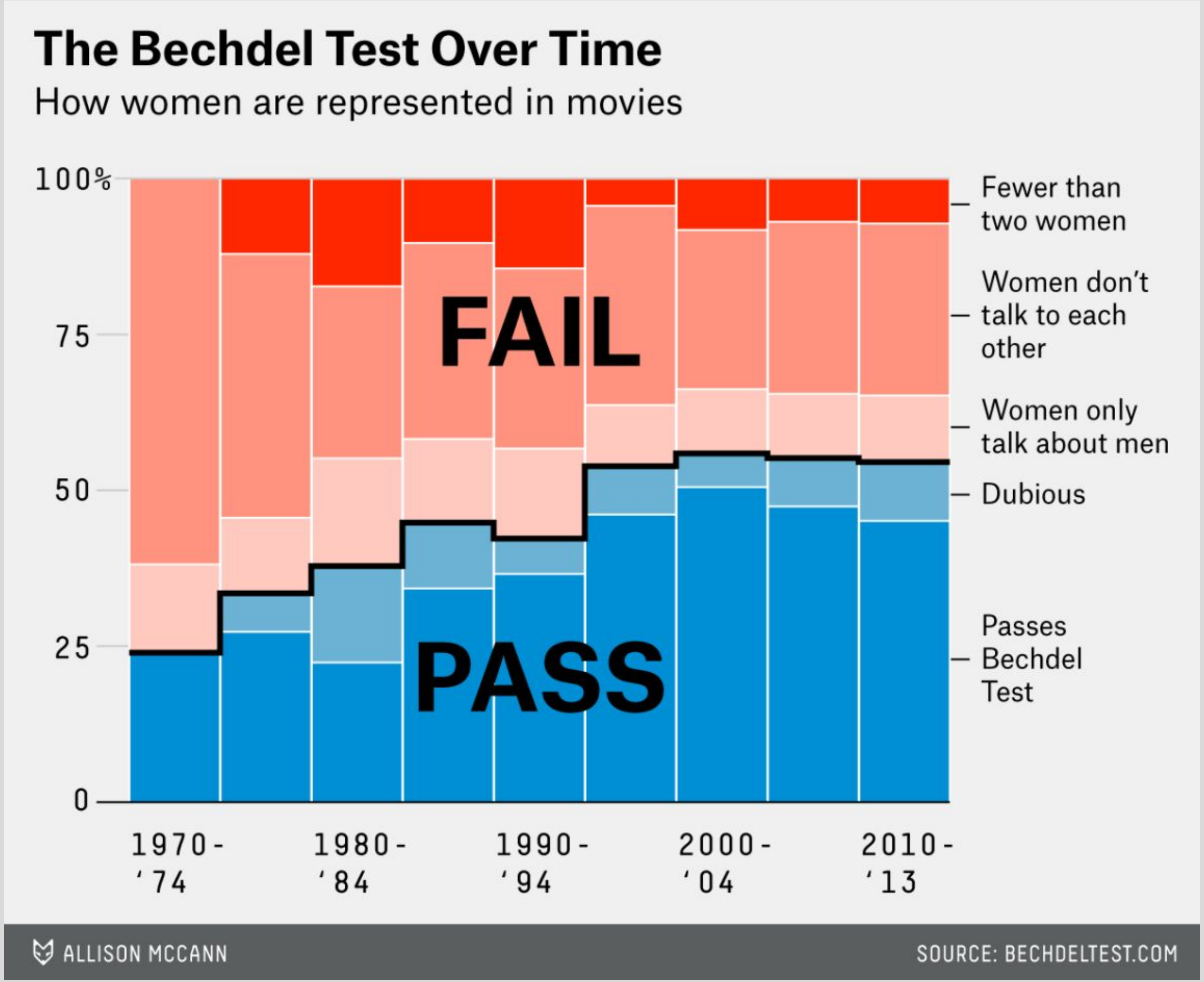
By Walt Hickey

Filed under Movies

RECOMMENDED

Trump's Scorning Of Data May Not Hurt Him, But It'll Hurt The GOP

Senate 2016: The Democrats Strike Back


The Bechdel Test Over Time
How women are represented in movies
100%
75
50
25
0
FAIL
PASS
Fewer than two women
Women don't talk to each other
Women only talk about men
Dubious
Passes Bechdel Test
1970-'74
1980-'84
1990-'94
2000-'04
2010-'13
ALLISON MCCANN
SOURCE: BECHDELTEST.COM

Branch: **master** ▾ data / bechdel / **analyze-bechdel.R**

Find file | Copy path

ascheink format email address — 3e71708 on Apr 8, 2014

2 contributors

79 lines (57 sloc) | 2.7 KB

Raw | Blame | History

```r
# Calculates summary statistics and conducts basic regression analysis to determine
# whether movies which pass the Bechdel test do better or worse at the box office,
# using data from www.bechdeltest.com and www.the-numbers.com

# By Andrew Flowers <andrew.flowers@fivethirtyeight.com>
# See also http://fivethirtyeight.com/features/the-dollar-and-cents-case-against-hollywoods-exclusion-of-women/

# Install and load required packages
# install.packages(c("gdata", "cwhmisc"))
library(gdata)
library(cwhmisc)

# Load data
rawData<-read.csv("movies.csv", na.strings="#N/A")
```

Danger: **Confirmation bias**

Tactic: **Showcase failures**

Ask yourself: *How much do I want to debunk this?*

Example: P-hacking in nutrition studies

JAN 6, 2016 AT 6:00 AM

You Can't Trust What You Read About Nutrition

We found a link between cabbage and innie bellybuttons, but that doesn't mean it's real.

By Christie Aschwanden

Filed under Nutrition

Nutrition Studies

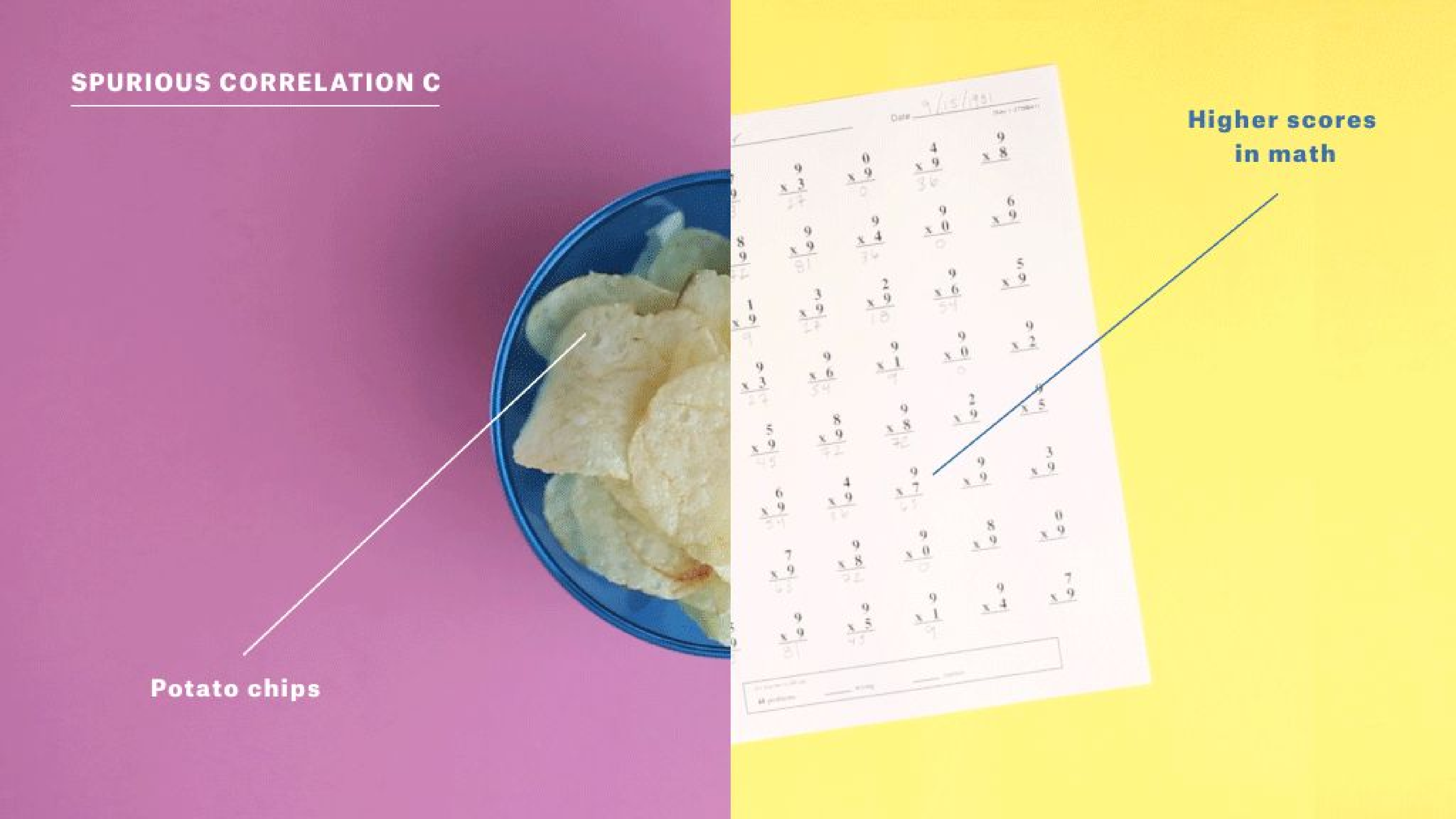
This directory contains data and code behind the story You Can't Trust What You Read About Nutrition.

Many studies of diet and nutrition include multiple variables with vast amounts of data, making it easy to p-hack your way to sexy (and false) results. We learned this firsthand when we invited readers to take a survey about their eating habits known as the food frequency questionnaire and answer a few other questions about themselves. We ended up with 54 complete responses and looked for associations much as researchers look for links between foods and dreaded diseases. It was easy to find them.

Warning: This is evil (statistical) work. Do not go to the dark side. Do not try this at home.


SPURIOUS CORRELATION A
Cabbage
Innies


SPURIOUS CORRELATION B
Dog ownership
Egg rolls


SPURIOUS CORRELATION C
Higher scores
in math
Potato chips

6. Forecast

Example: Underdog beats the FiveThirtyEight forecast

Biggest college football championship upsets, 1975-2016

SEASON	CHAMPION	BOWL RESULT	BOWL OPPONENT	PREGAME ELO WIN PROBABILITY
2002	Ohio State	31-24, Fiesta Bowl	Miami (FL)	22.2%
1983	Miami (FL)	31-30, Orange Bowl	Nebraska	25.9
2006	Florida	41-14, BCS Championship	Ohio State	28.6
1992	Alabama	34-13, Sugar Bowl	Miami (FL)	30.5
2016	**Clemson**	**35-31, CFP Championship**	**Alabama**	**32.9**
1981	Clemson	22-15, Orange Bowl	Nebraska	34.0
2000	Oklahoma	13-2, Orange Bowl	Florida State	35.8
2005	Texas	41-38, Rose Bowl	USC	39.8
2014	Ohio State	42-20, CFP Championship	Oregon	40.1
2010	Auburn	22-19, BCS Championship	Oregon	41.9

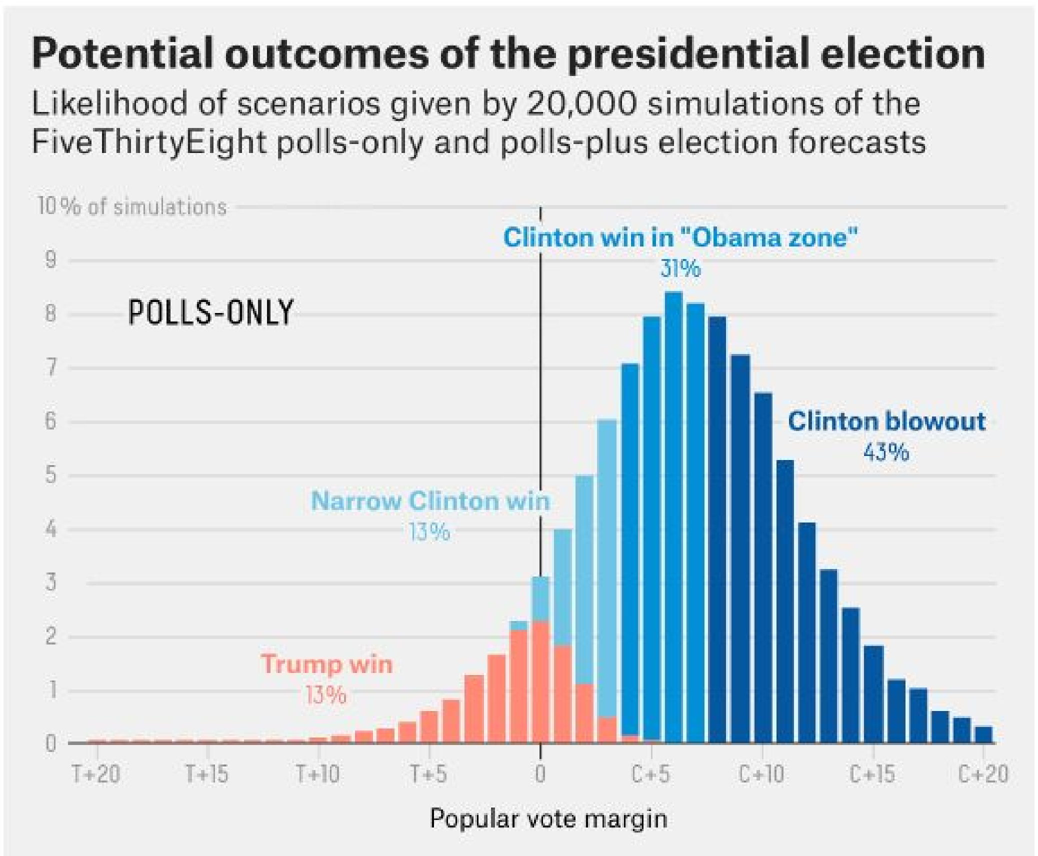
SOURCE: COLLEGE FOOTBALL AT SPORTS-REFERENCE.COM

Danger: **Overfitting**

Tactic: **Simulations and scenarios**

Ask yourself: *Am I properly conveying the uncertainty in my model?*

Example: Election outcomes



Note: this simulation was run on Oct. 21, 2016.

What makes a ***data*** story worth telling?

But where is the danger?

1. Novelty
2. Outlier
3. Archetype
4. Trend
5. Debunking
6. Forecast

A. Triviality
B. Spurious result
C. Oversimplification
D. Variance
E. Confirmation bias
F. Overfitting