MY472 – Week 9: Exploratory Data Analysis

Pablo Barberá & Akitaka Matsuo

MY 472: Data for Data Scientists

December 4, 2018

Course website: lse-my472.github.io

Course outline

1. Introduction to data
2. The shape of data
3. Cloud computing
4. Basics of HTML and CSS
5. Using data from the internet
6. (Reading week)
7. Working with APIs
8. Creating and managing databases
9. Interacting with online databases
10. Exploratory data analysis
11. Parallel computing

Seminar schedule

9 Online databases
- 4th marked assignment (in groups)
- Deadline: December 14th

10 Exploratory data analysis

11 Course wrap-up
- 5th marked assignment (individual)
- Deadline: December 21st

Take-home exam due January 18

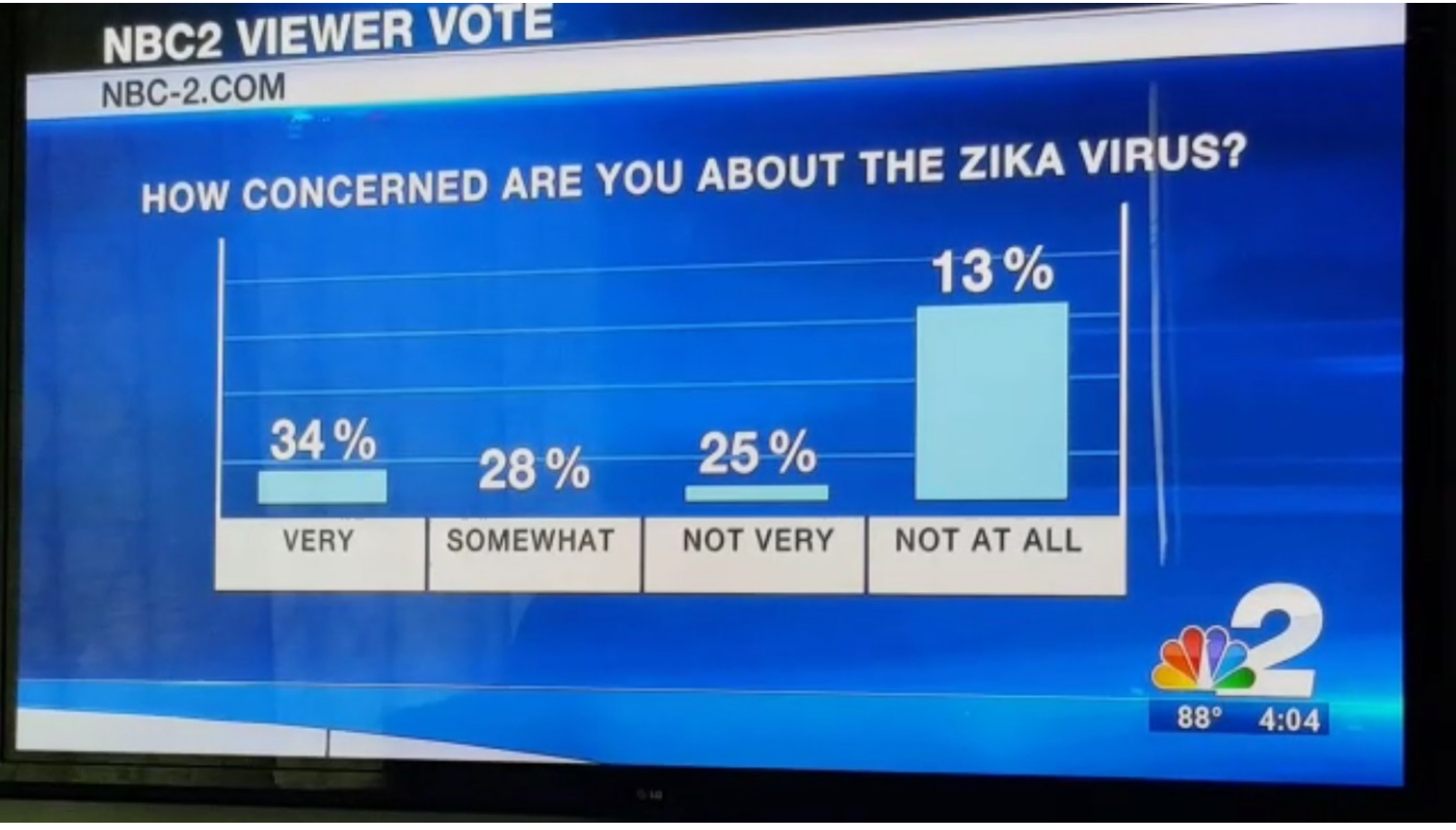
Plan for today

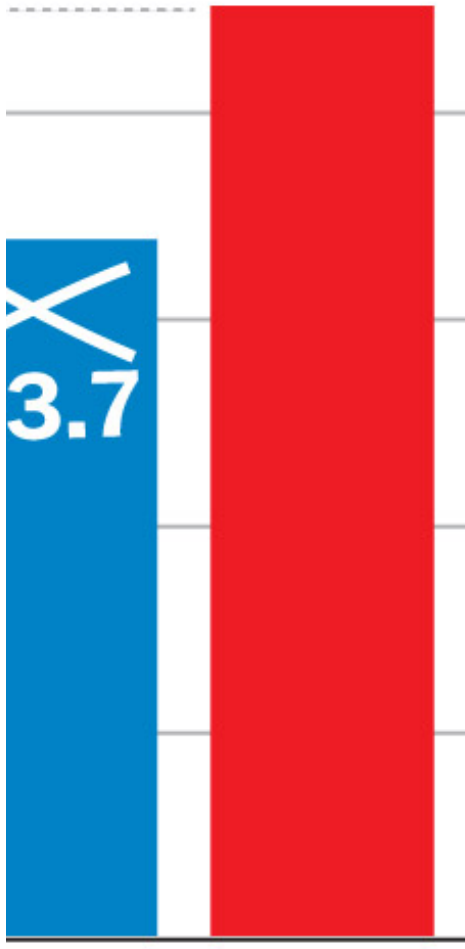
- Data visualization
  - How (not) to lie with graphs
  - Principles of data visualization
  - `ggplot2`
- Teaching evaluations

Data visualization: why?

Often the most effective way to describe, explore, and summarize data...is to visualize the data

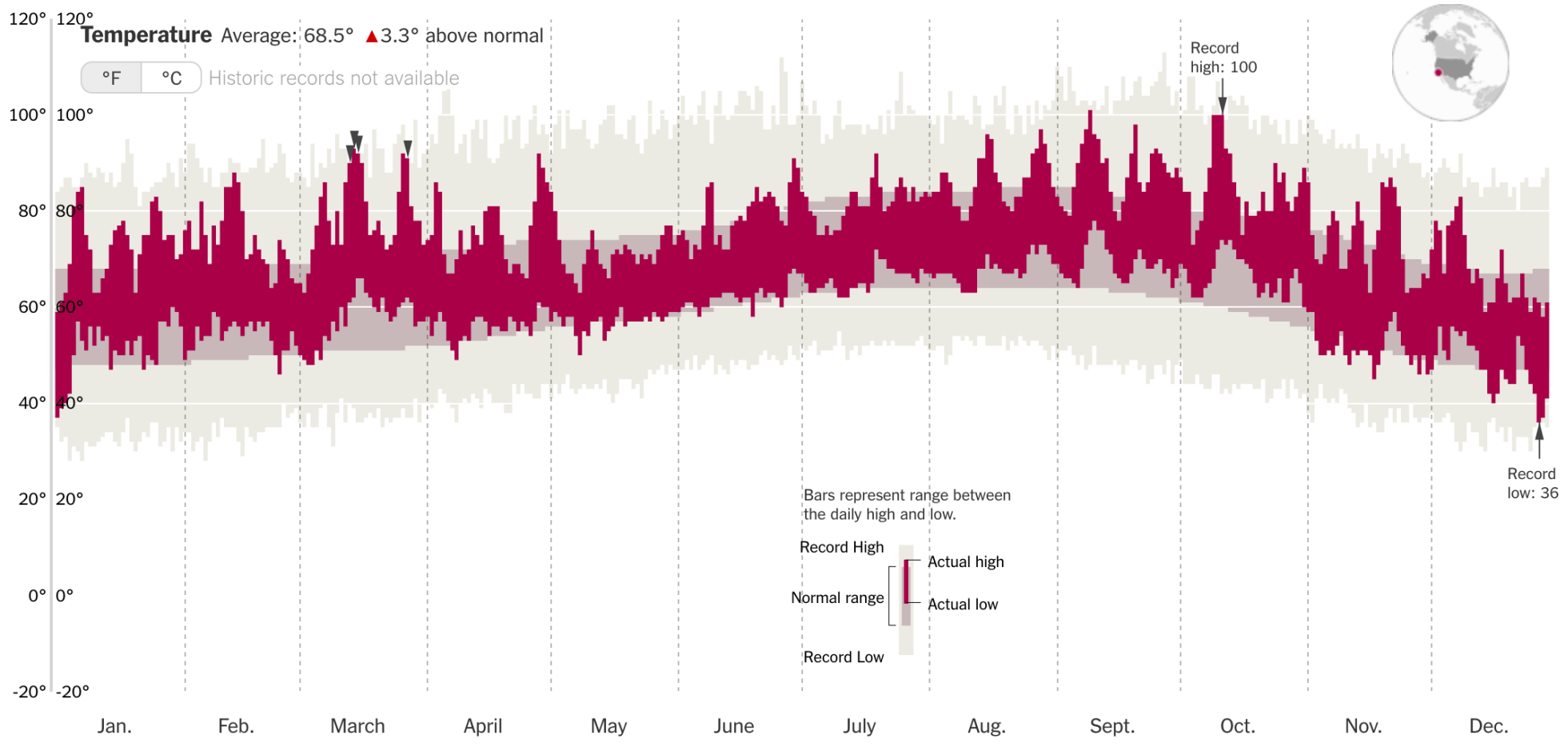
`see 01-anscombe.Rmd`

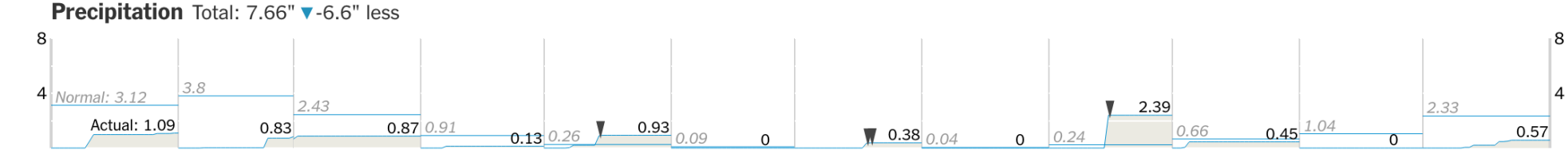

NBC2 VIEWER VOTE
NBC-2.COM
HOW CONCERNED ARE YOU ABOUT THE ZIKA VIRUS?
34 %
28 %
25 %
13 %
VERY
SOMEWHAT
NOT VERY
NOT AT ALL
2
88°
4:04



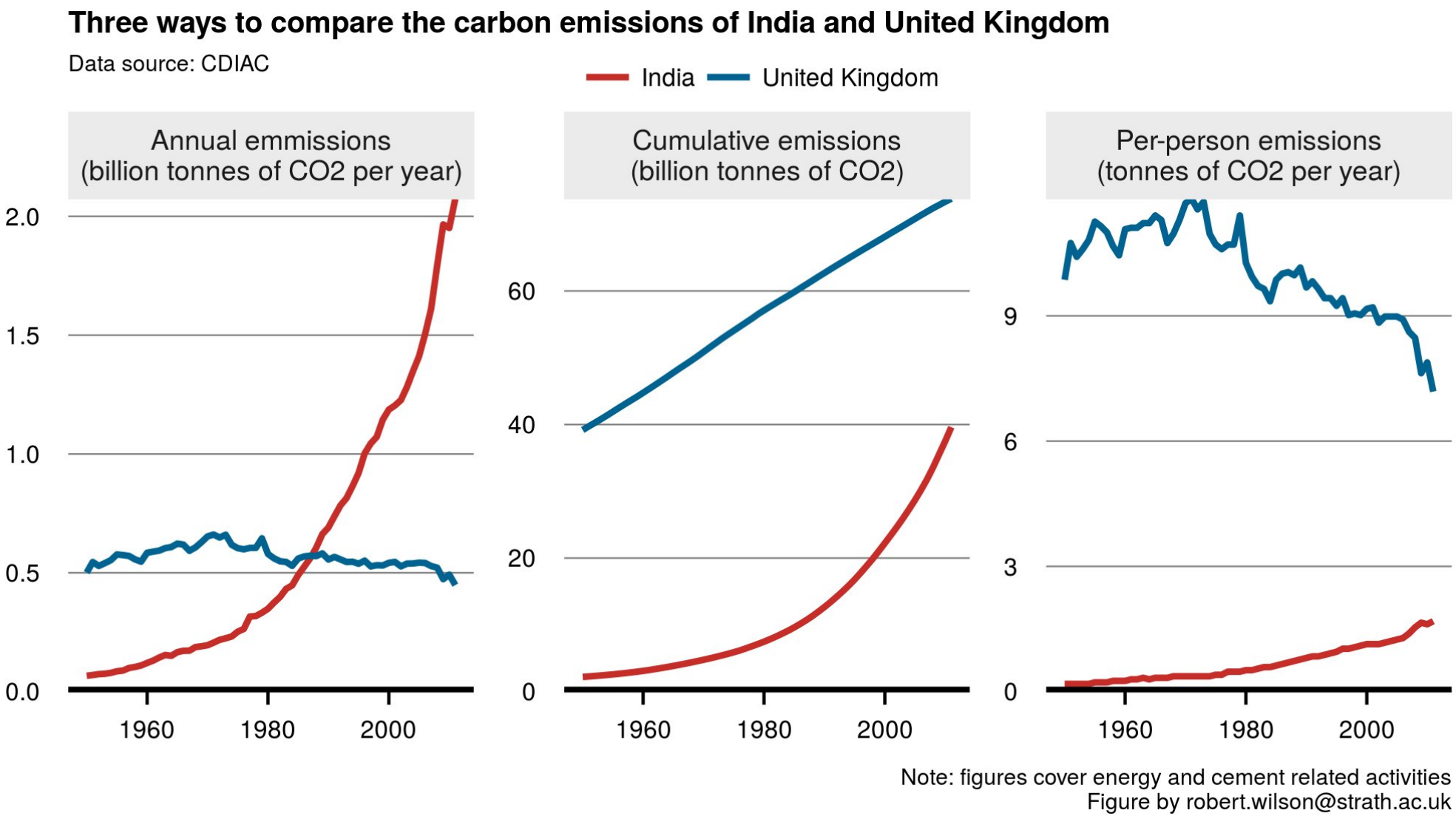
Source: Washington Post

Los Angeles, Calif.

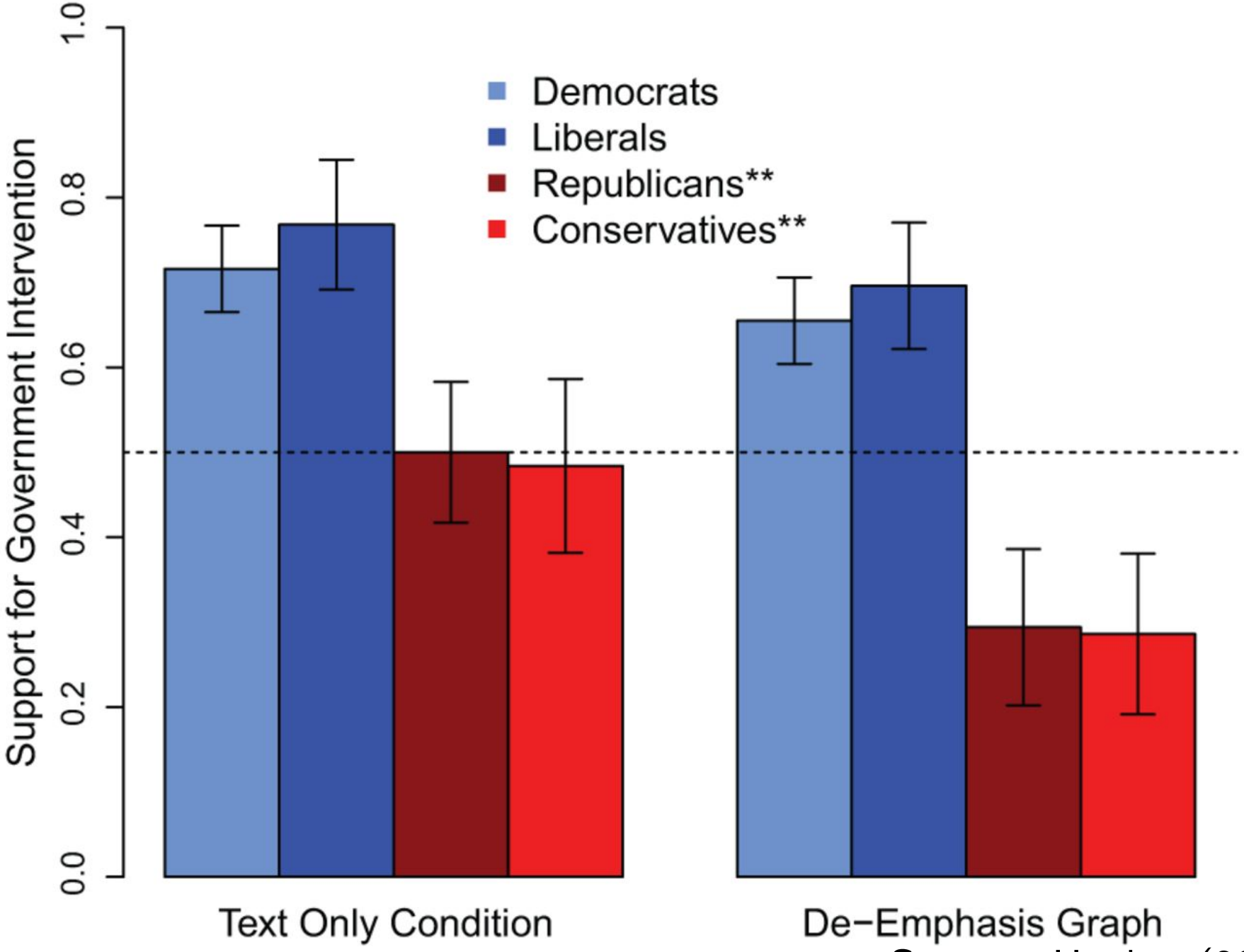




Cumulative monthly precipitation, in inches, compared with normal. Precipitation totals are rainfall plus the liquid equivalent of any frozen precipitation.



Source: New York Times



Source: Hughes (2015)

Data visualization

General principles (Tufte)

- Show the data
- Avoid distorting what the data have to say
- Allow viewer to compare
- Serve a clear purpose: description, exploration, tabulation or decoration
- Be closely integrated with the statistical and verbal descriptions of the dataset
- **Graphics reveal data**: e.g. Anscombe Quartet

Data visualization

Specific guidelines

- Maximize data-to-ink ratio
- Avoid misleading decisions:
  - Y axis starts at 0
  - Comparison of areas is hard
  - Use comparable units
  - Erase chart junk
- Use text to inform and contextualize. Add annotations
- Appropriate use of scales (x/y axes, color, size, shape...)
- Use small multiples to facilitate comparisons
- Always cite your sources

Data visualization with ggplot2

Why `ggplot2`?

- Based on "Grammar of Graphics" (Wilkinson, 2005)
  - $\rightarrow$ powerful, consistent, modular.
- Compact, parsimonious code
- Sensible defaults for quick exploratory plots
- But also easy to customize, extend
- Excellent online resources (and easy to google)

What is the grammar of graphics?

The grammar of graphics.

A statistical graph is a mapping from data to aesthetic attributes (color, shape, size) of geometric objects (points, lines, bars). The plot may also contain statistical transformations of the data and is drawn on a specific coordinate system. Faceting can be used to generate the same plot for different subsets of the data. It is the combination of these independent components that make up a graphic.

Hadley Wickham, *ggplot2, page 3*

What is the grammar of graphics?

Components of a graph:

- **data** What you want to visualize, including variables (columns) to be mapped to aesthetic attributes.
- **geom** Geometric objects that are drawn to represent the data: bars, lines, points, etc.
- **stats** Statistical transformations of the data, such as binning or averaging.
- **scales** Map values in the data space to values in an aesthetic space (color, shape, size...)
- **coord** Coordinate system; provides axes and gridlines to make it possible to read the graph.
- **facets** Breaking up the data into subsets, to be displayed independently on a grid

ggplot2

see 02_ggplot2_basics

see 03_scales_axes_legends