FINDING INSIGHTS IN DATA

> **The visual representation and presentation of data facilitates understanding.**
>
> *Data Visualization: A Handbook of Data Driven Design* 2nd Edition
> Andy Kirk

selecting a chart to display the features of your data you think are most important

enhancing the chart using annotation, color, composition, and interactivity

comprehending the chart what has the consumer learnt, is the chart trustworthy
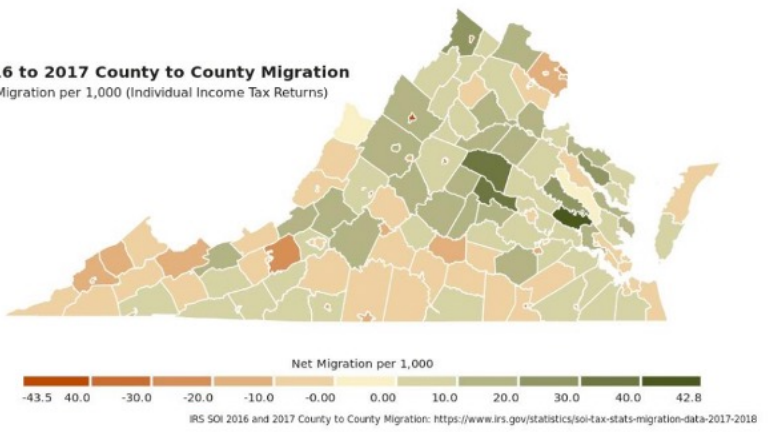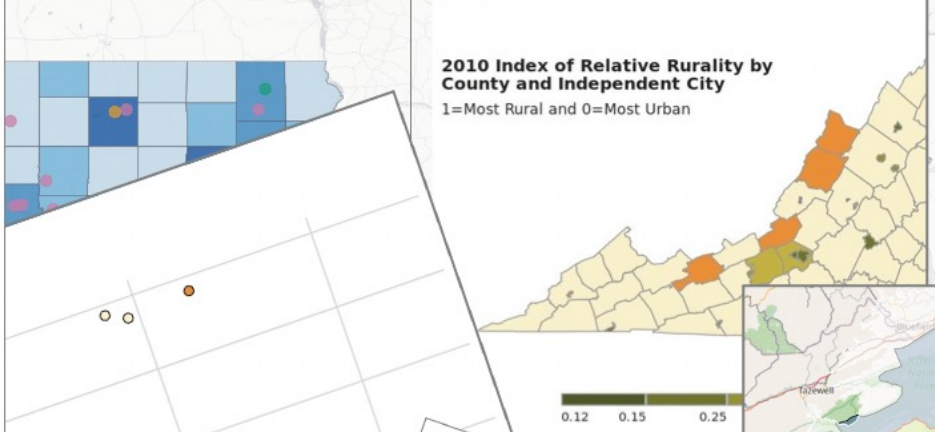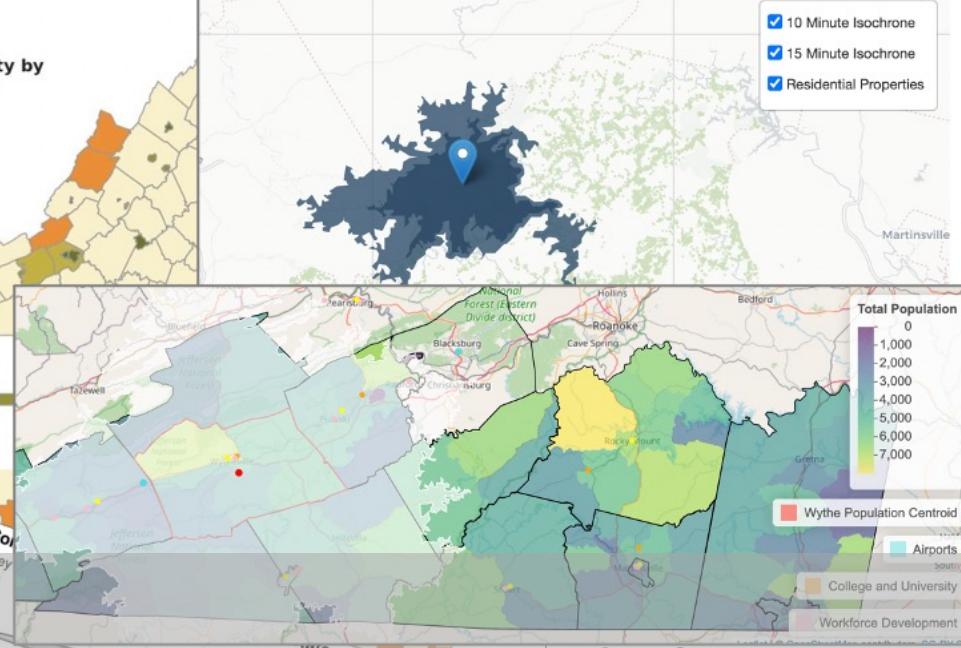
TAXONOMY OF A VISUALIZATION
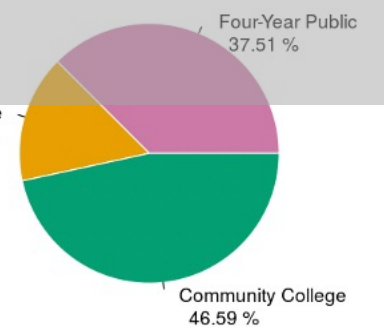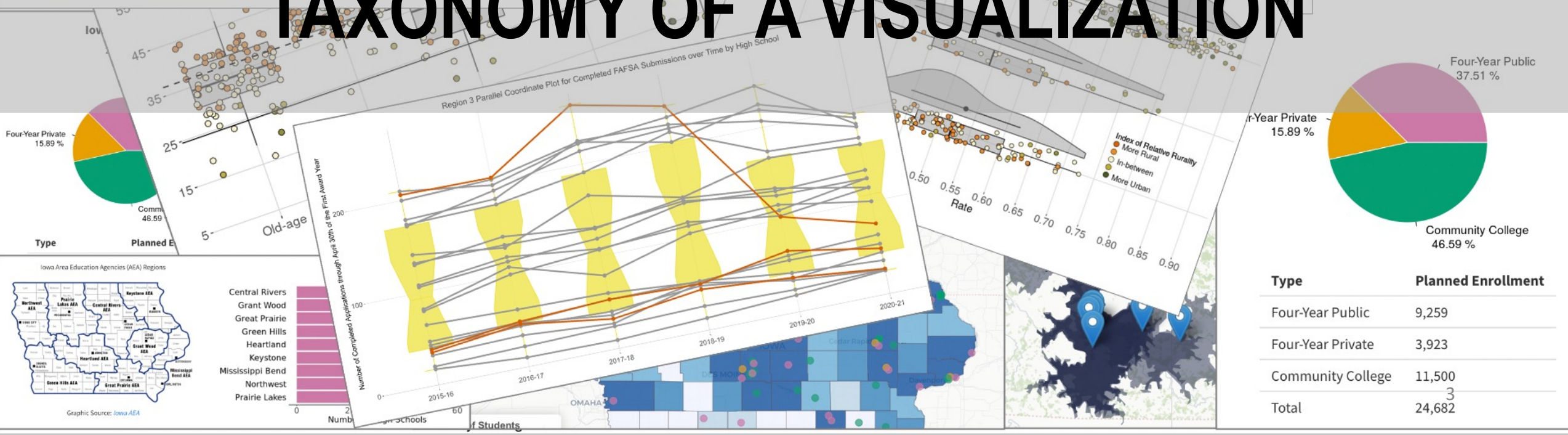
Visualization = Scaffolding + Content

Scaffolding consist of…
- Title
- Sub-title
- Axis Titles
- Axis Labels
- Legends
- Measurement
- Measurement Unit
- Scale
- Data Sources

How Charts Lie

Getting Smarter about Visual Information

Alberto Cairo
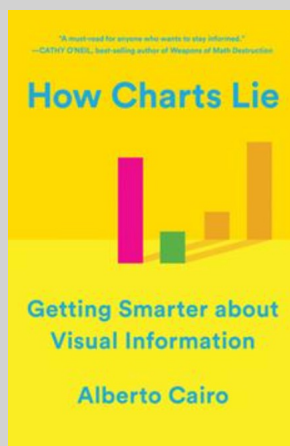
Adapted from

Science and engineering research space in academic institution, change over 2-year periods: FYs 1988–2017
The biennial survey cycle ran on even years from 1988 to 1998 and on odd years from 1999 to 2017.

Net assignable square feet in millions
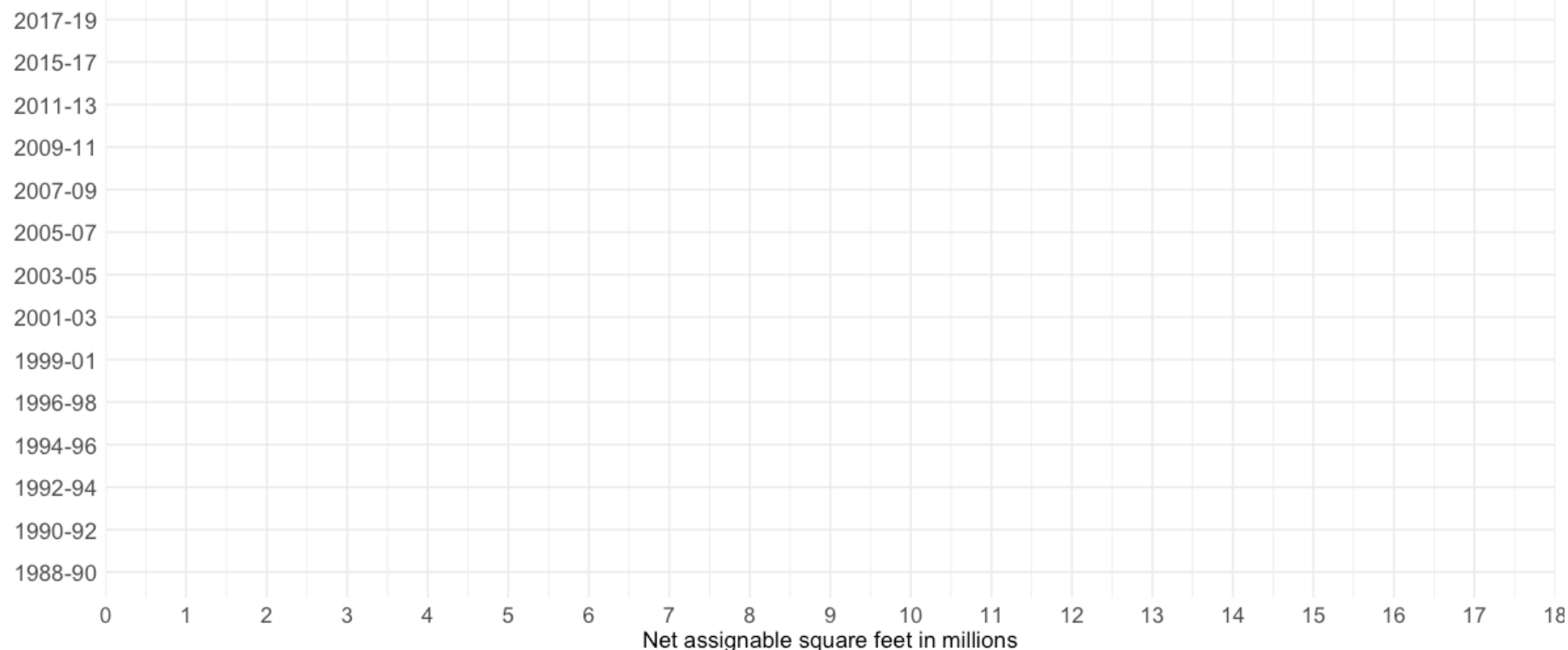
SOURCE: National Science Foundation, National Center for Science and Engineering Statistics, Survey of Science and Engineering Research Facilities

The **content** consists of…

- How the data are or represented or encoded
- An annotation layer that can be used to clarify a point of interest
- In this case the data are represented by the length of a bar and annotation is used to provide the length of the bar

Complete Visualization

A **bar chart** is a visualization of two variables, one with an ordinal level of measurement, year, and the other with a ratio level of measurement, square feet.

Science and engineering research space in academic institution, change over 2-year periods: FYs 1988–2017
The biennial survey cycle ran on even years from 1988 to 1998 and on odd years from 1999 to 2017.

SOURCE: National Science Foundation, National Center for Science and Engineering Statistics, Survey of Science and Engineering Research Facilities

Net assignable square feet in millions

Complete Visualization

A **bar chart** is a visualization of two variables, one with an ordinal level of measurement, year, and the other with a ratio level of measurement, square feet.

Science and engineering research space in academic institution, change over 2-year periods: FYs 1988–2017
The biennial survey cycle ran on even years from 1988 to 1998 and on odd years from 1999 to 2017.

Net assignable square feet in millions

SOURCE: National Science Foundation, National Center for Science and Engineering Statistics, Survey of Science and Engineering Research Facilities

Pew: minimal scaffold (legend, axes titles & labels)

Branding consists of…
- Title: black bolded
- Sub-title: grey italic different font than the title
- Axis Titles: grey same font as the title or minimal
- Axis Labels: grey same font as the title: minimum and maximum
- Legends: annotation
- Measurement: identified in sub-title
- Data Sources: identified at the bottom

Most who can do their job from home say they are currently doing so all or most of the time

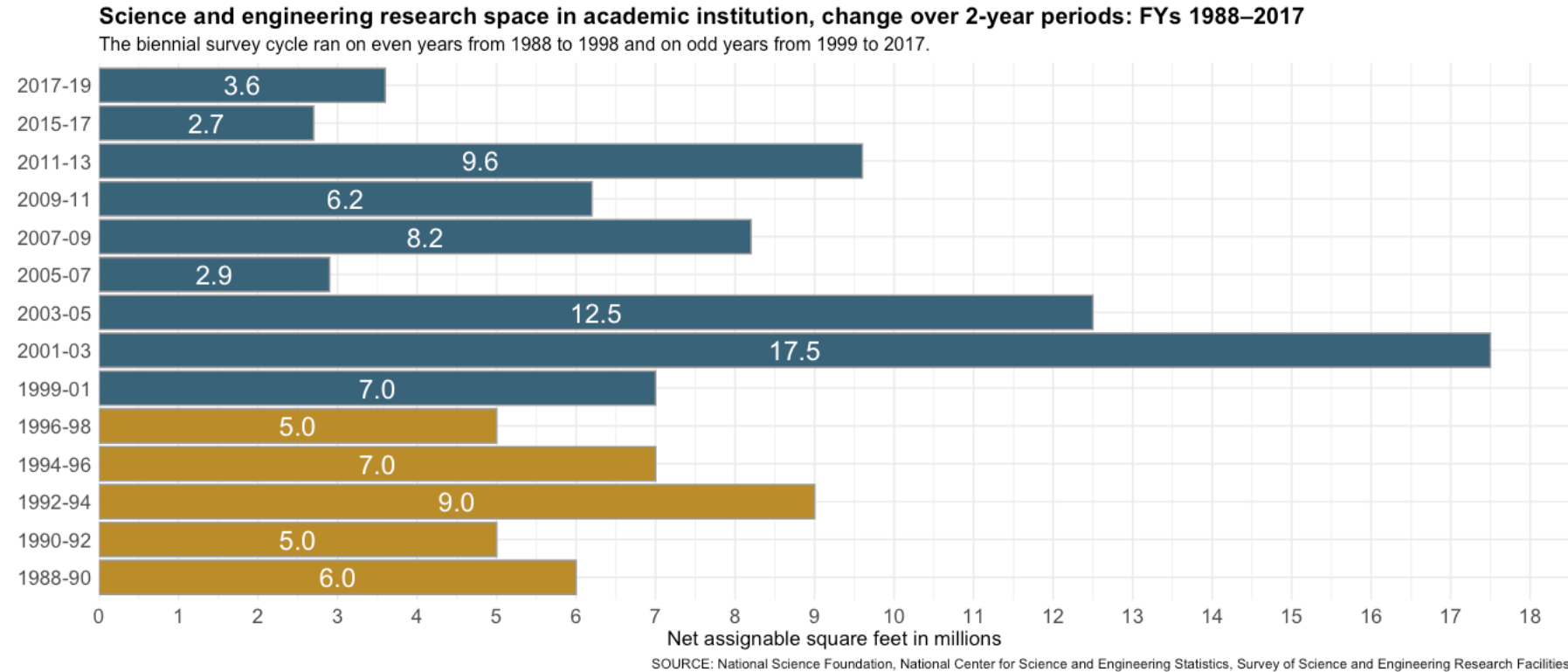*Among employed adults who say that, for the most part, the responsibilities of their job can be done from home, % saying they are currently working from home …*

	All of the time	Most of the time	Net
All who can do their work from home	55	16	71
Lower income	46	20	66
Middle income	52	15	67
Upper income	65	14	79

Note: Figures may not add to subtotals due to rounding. Family income tiers are based on adjusted 2019 earnings.
Source: Survey of U.S. adults conducted Oct. 13-19, 2020.
"How the Coronavirus Outbreak Has – and Hasn't – Changed the Way Americans Work"

PEW RESEARCH CENTER

While most Americans say schools should provide computers to at least some students during outbreak, parties divided on if this should be done for all

% of U.S. adults who say that K-12 schools have a responsibility to provide laptops or tablet computers to ___ in order to help them complete schoolwork at home during the coronavirus outbreak
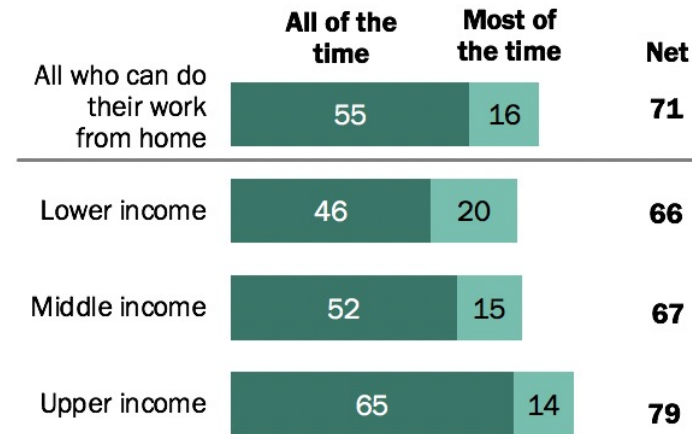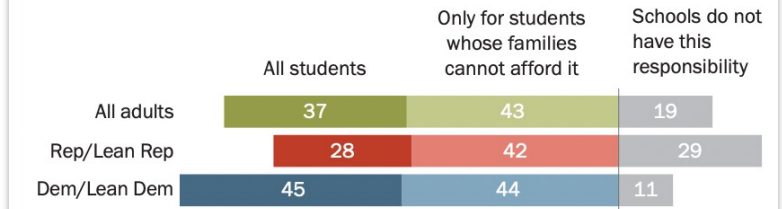
	All students	Only for students whose families cannot afford it	Schools do not have this responsibility
All adults	37	43	19
Rep/Lean Rep	28	42	29
Dem/Lean Dem	45	44	11

Note: Those who did not give an answer are not shown.
Source: Survey of U.S. adults conducted April 7-12, 2020.
"53% of Americans Say the Internet Has Been Essential During the COVID-19 Outbreak"

PEW RESEARCH CENTER

Restrictions on women's dress have risen around the world in the past five years

Number of countries where women experienced …

60 COUNTRIES — 61
41

Government restrictions
Countries that regulate women's religious clothing by law or policy

53 — 56

Social hostilities
Countries where women were harassed for their religious or secular dress

2013 — 2018   2013 — 2018
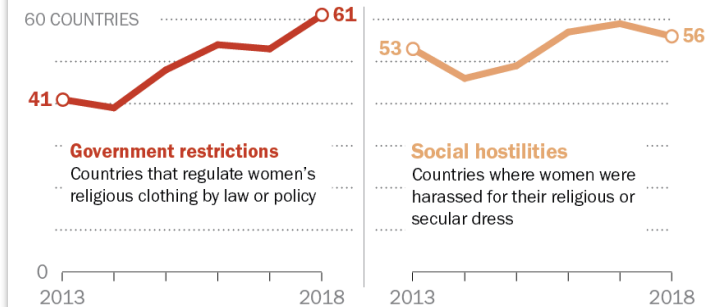
Note: Government restrictions include both rules that prohibit religious garb and rules that require it. Social hostilities data for each year also takes into account incidents from the two previous years to capture ongoing hostilities.
Source: Pew Research Center analysis of external data.
"In 2018, Government Restrictions on Religion Reach Highest Level Globally in More Than a Decade"

PEW RESEARCH CENTER

FiveThirtyEight: grey background & no legend

Branding consists of…

- Title: black bolded
- Sub-title: grey italic different font than the title
- Axis Titles: same color as title
- Axis Labels: grey same font as the title
- Legends: annotation
- Measurement: identified in sub-title
- Data Sources: identified at the bottom

Vulnerable populations tend to have the high-risk jobs
Share of workers among high and low likelihoods of being able to work from home and high and low physical proximity to others at work

FiveThirtyEight
SOURCE: MONGEY, PILOSSOPH AND WEINBERG
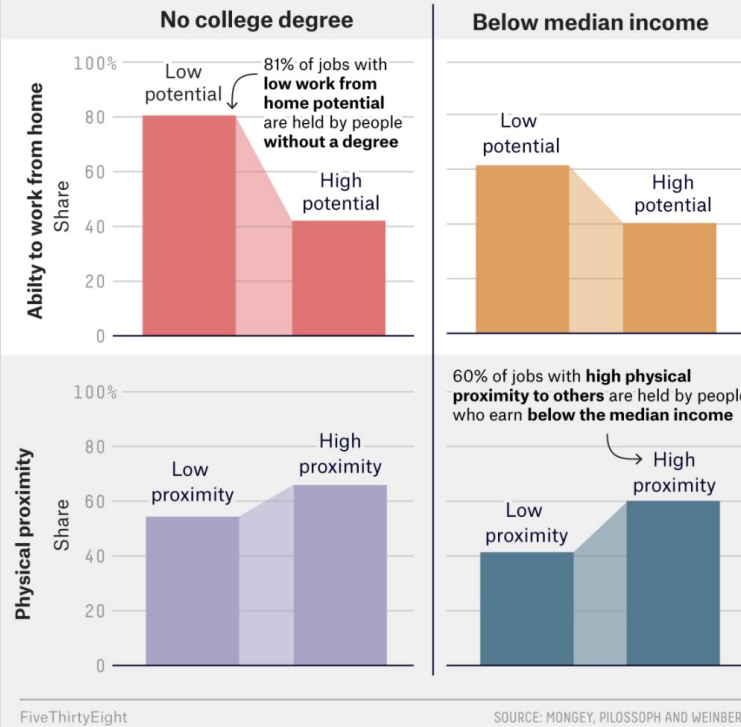
The Black unemployment rate remains very high
Unemployment rate by race, Jan. 2007 through Aug. 2020

FiveThirtyEight
SOURCE: BUREAU OF LABOR STATISTICS
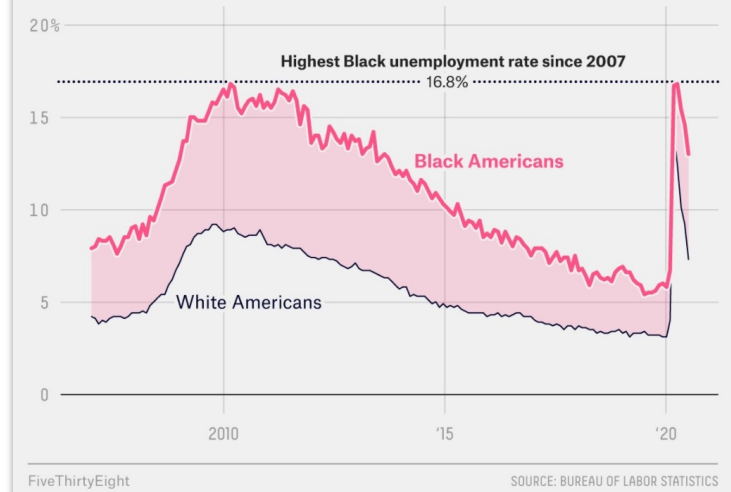
Black coaches aren't being replaced by Black coaches
NBA head coaches who left a team, broken down by race and by the race of the coaches replacing them at the start of the next season

FiveThirtyEight
SOURCE: BASKETBALL-REFERENCE.COM, THE INSTITUTE FOR DIVERSITY AND ETHICS IN SPORT
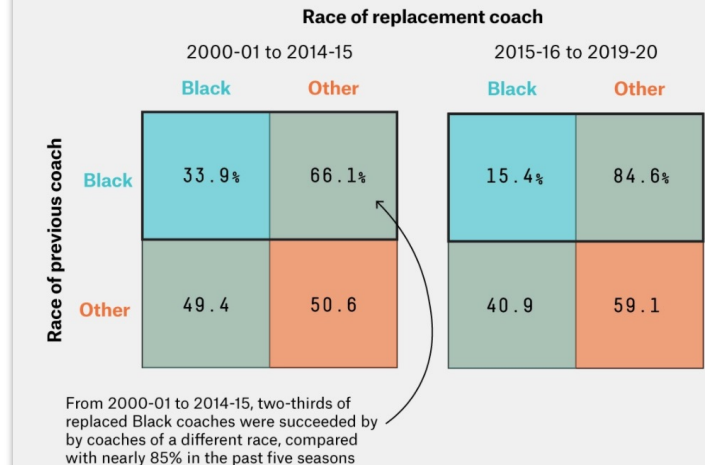
Tufte: *adding additional information to the axes*

Branding consists of…
packing a lot of information either using the axes
- (2) labels - the five-number summary
- (3) tick marks – marginal histograms

No grid lines

(2)

(3)
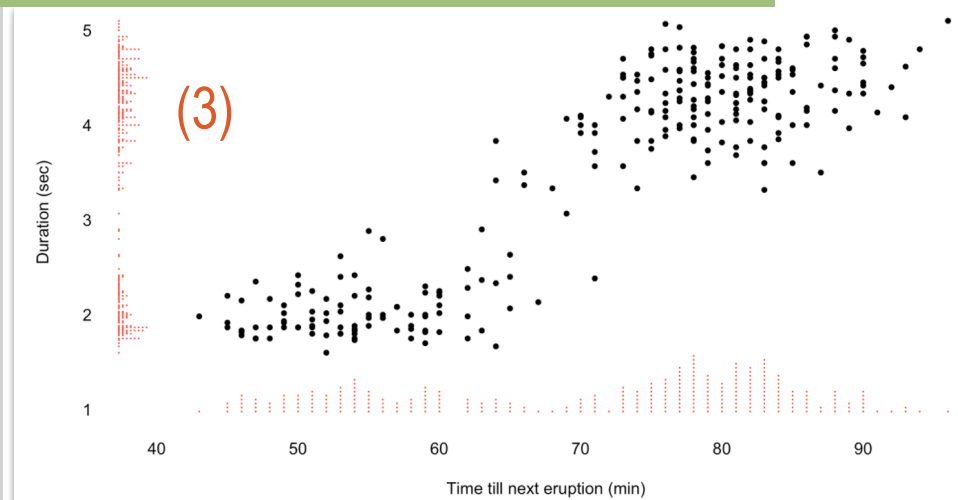
Dot Plots
(1) Dot-Dash, (2) Dot-Range frame,
(3) Dot-Marginal histogram

(1)

Sparklines (4)
Blue Dot = Maximum
Red Dot = Minimum
Grey Rectangle = area between the
1st and 3rd quartiles

Economist: axis labels on the right

Branding consists of…

- Title: black bolded
- Sub-title: grey italic different font than the title
- Axis Titles: none or minimal
- Axis Labels: grey same font as the title: minimum and maximum
- Legends: none
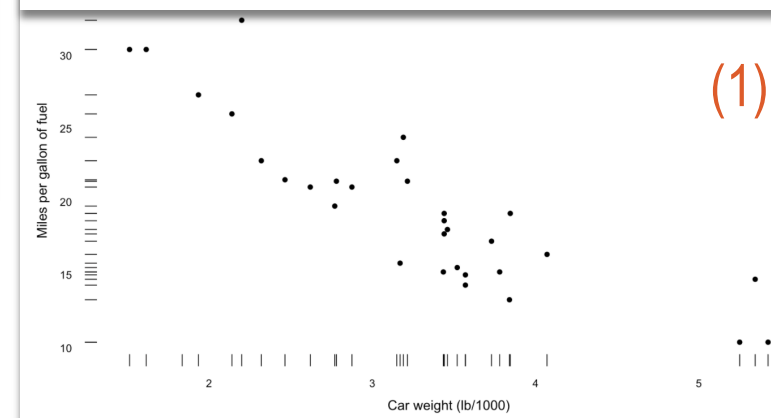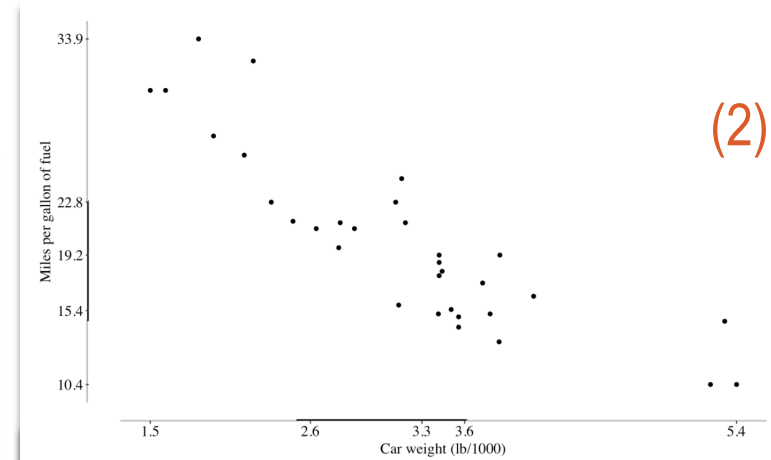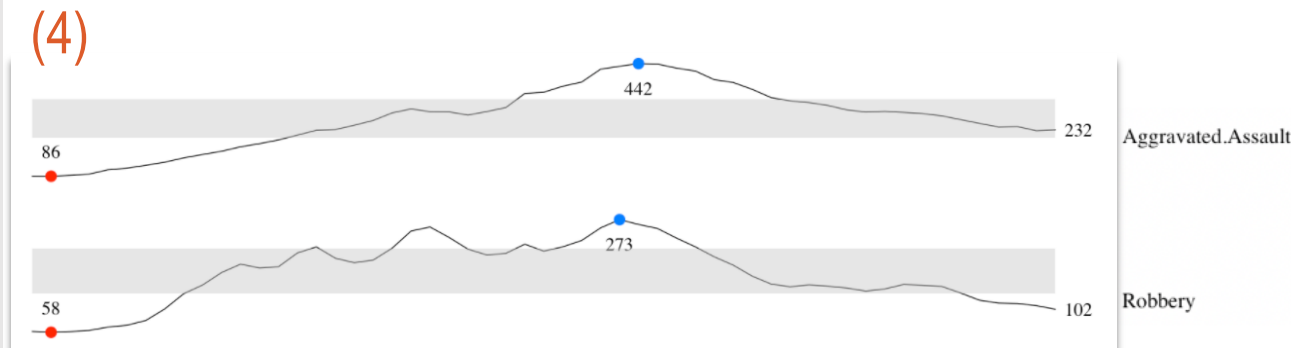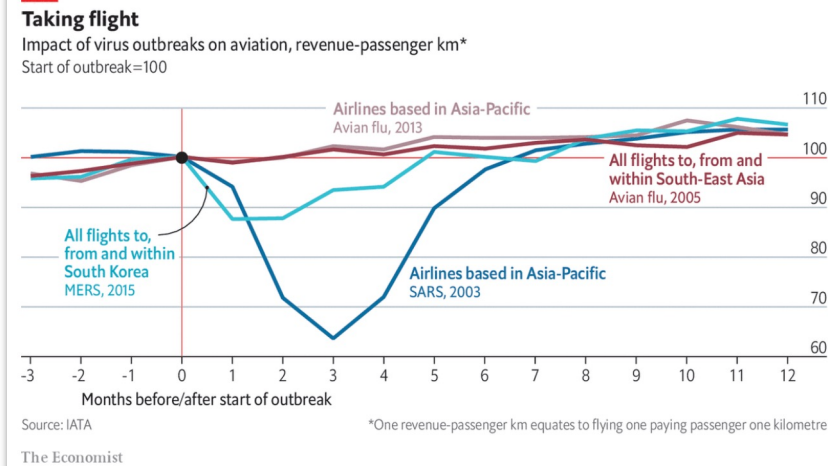- Measurement: identified in sub-title
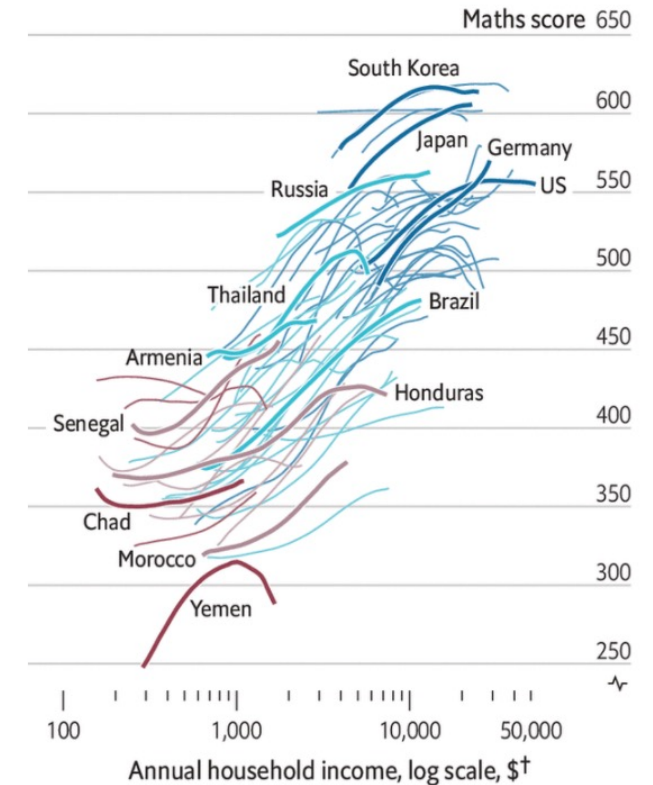- Data Sources: identified at the bottom

Taking flight
Impact of virus outbreaks on aviation, revenue-passenger km*
Start of outbreak=100

Airlines based in Asia-Pacific
Avian flu, 2013

All flights to, from and within South-East Asia
Avian flu, 2005

All flights to, from and within South Korea
MERS, 2015

Airlines based in Asia-Pacific
SARS, 2003

Months before/after start of outbreak

Source: IATA   *One revenue-passenger km equates to flying one paying passenger one kilometre

The Economist

Viral content
Subject matter of our 24 most popular articles*, 2020

- Covid-19
- US election
- Democracy index
- Cost of living index

Source: *The Economist*   *Daily charts and Graphic Detail print pieces

The Economist

It all adds up
Household income and maths test scores*
Between the 10th and 90th percentile of income in each country

Country income group
— High — Upper middle — Lower middle — Low

Maths score 650

South Korea
Japan Germany
Russia US
Thailand Brazil
Armenia
Honduras
Senegal
Chad
Morocco
Yemen
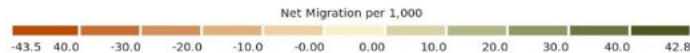
Annual household income, log scale, $†

*Estimated TIMSS score equivalent, out of 1,000
†In 2005 dollars, at purchasing-power parity (PPP)

Sources: "A Rosetta Stone for Human Capital", by D. Patel and J. Sandefur; World Bank; *The Economist*

The Economist
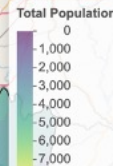
HOW VISUALIZATIONS DECEIVE
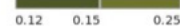## (or how to be a discriminating consumer)
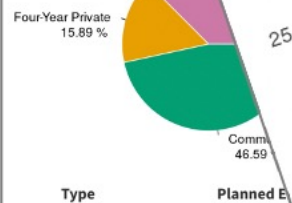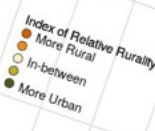
…by Being Poorly Designed

A tweet stated…

"Good news: America's high school graduation rate has increased to an all-time high.

The tweet included this visualization.
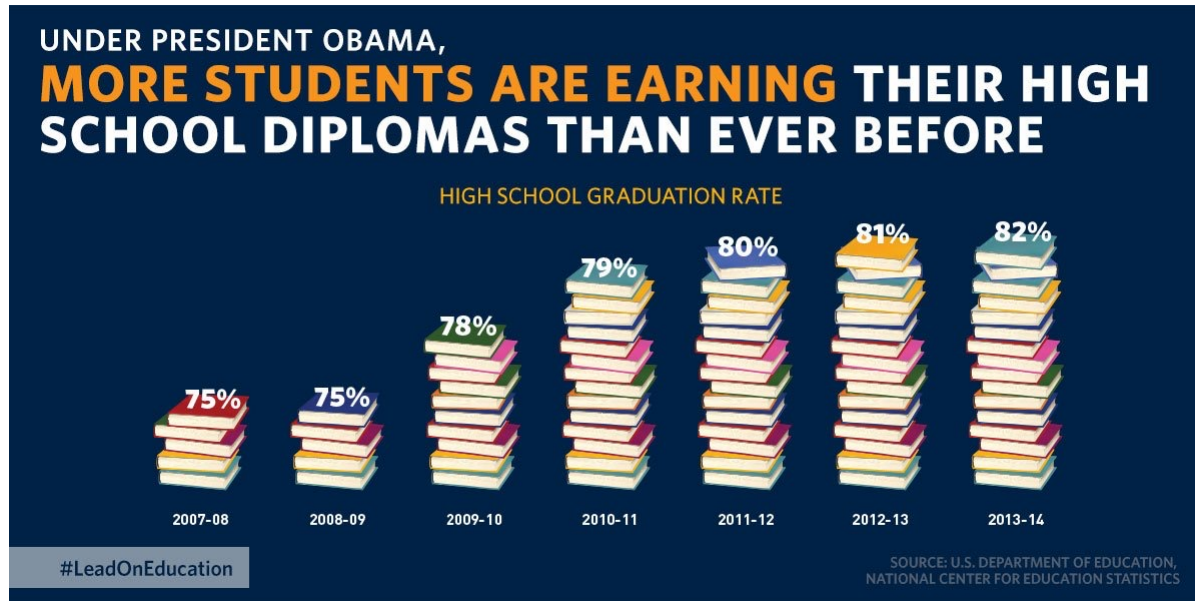
The visual encoding uses height to represent the high school graduation rate. In the absence of a vertical axis the number of books serves as a proxy for the percentage of graduates BUT the percentage of graduates per book is not consistent over time:

- in 2007-08 one book = 15.000%; and
- in 2013-14 one book = 5.125%.

…by Being Poorly Designed
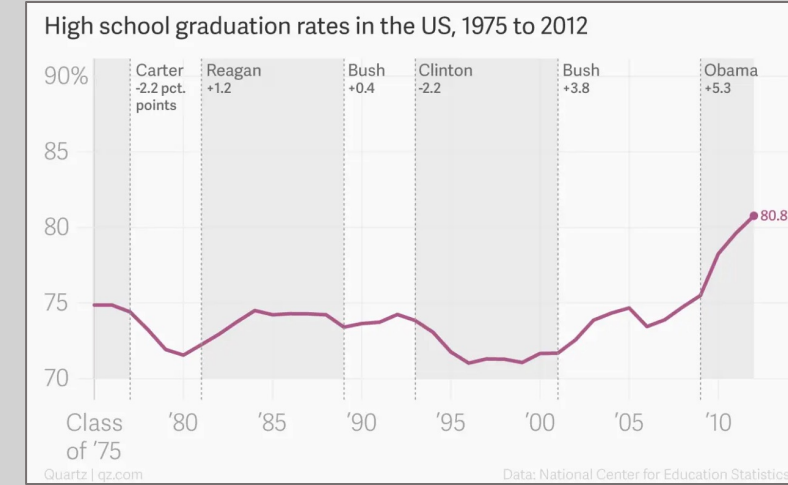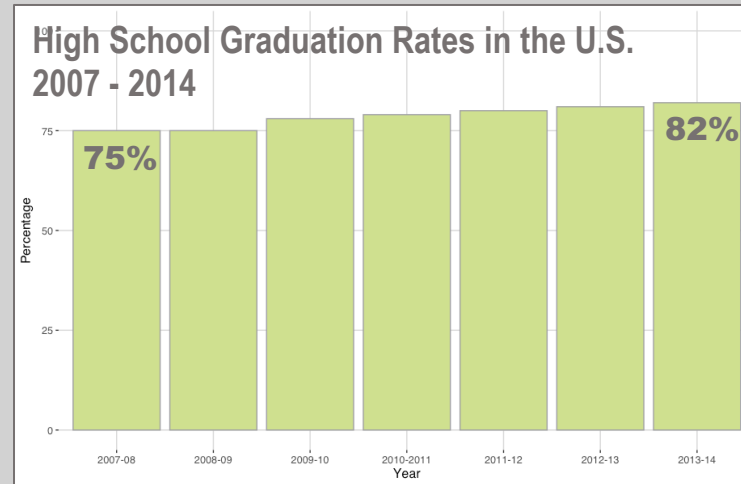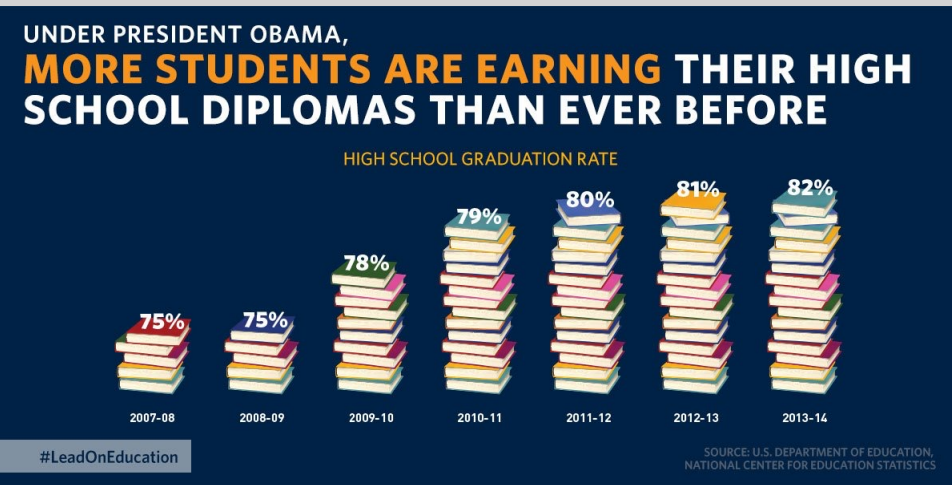
Take-aways:

- Focusing on the scaffolding – in this case on the measurement scale – will help identify distortions.
- When the method of encoding is height or length start at the baseline, in this case zero.
- Although the replot using a bar chart is less dramatic, you can call attention the percentage point increase of 7% by using annotation.
- When you want to show subtle changes over time us a line chart.
- In some case you didn't need a visualization – this might be one of them.

14

…by Displaying Insufficient Information

The headline of the pie chart stated:
> Look closely at this chart of federal spending... Somewhere within the tiny orange sliver at the bottom is the food stamp program….

A social media meme states 57% of federal spending goes to the military and just 1% goes to food and agriculture, including food stamps.

Education
6%

Government
6%

Veterans' Benefits
6%

Housing & Community
5%

Health
5%

International Affairs
3%

Energy & Environment
3%

Science
3%

Transportation
3%

Labor
2%

Food & Agriculture
1%

Military
57%

Adapted from an article on QUARTZ https://qz.com/580859/the-most-misleading-charts-of-2015-fixed/
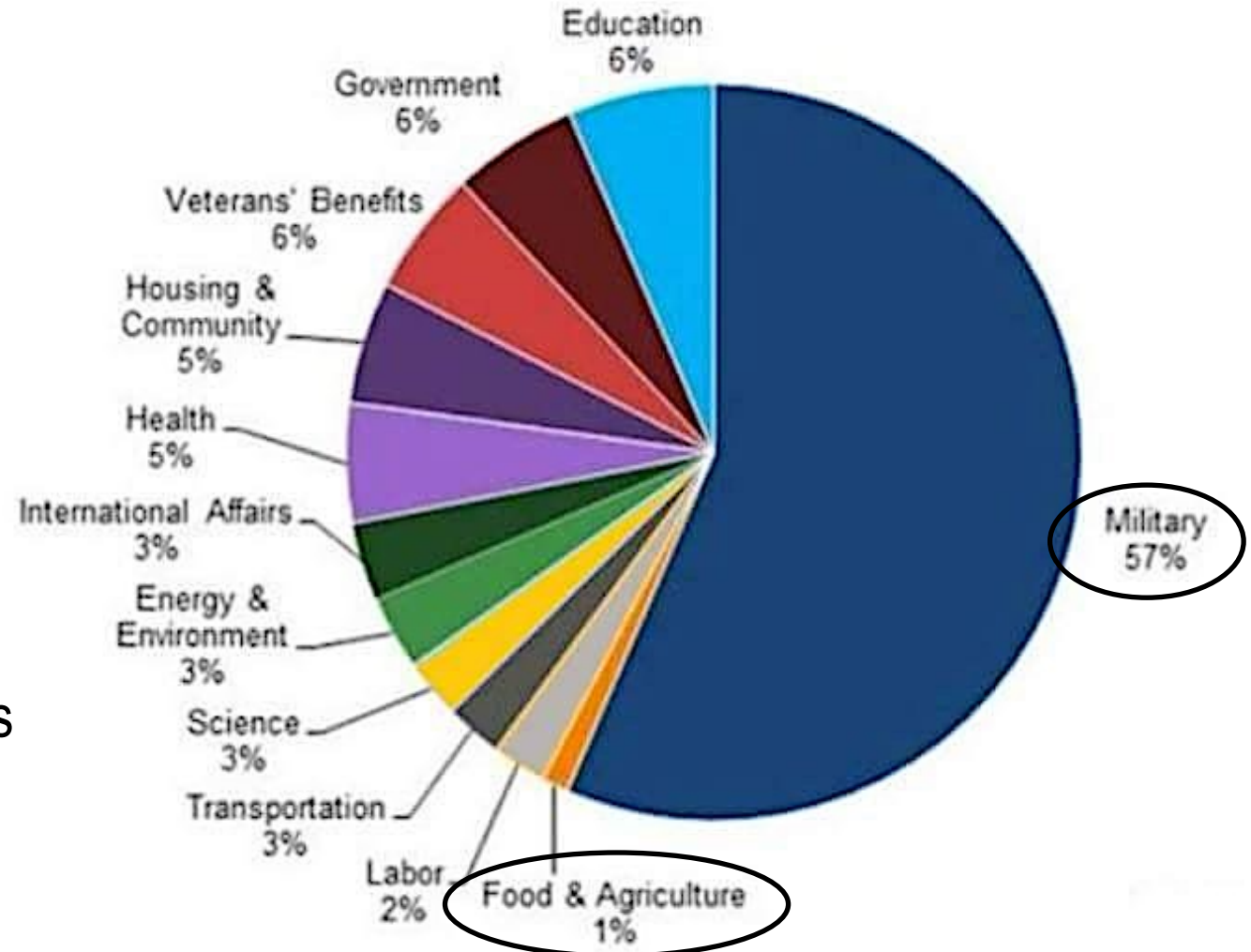
…by Displaying Insufficient Information

The visual encoding uses area to display federal budget allocations. It fails to inform the consumer it is not the entire federal budget.

The pie chart ONLY shows the discretionary spending and leaves out mandatory spending such as Medicare, Medicaid, and Social Security, which accounts for 60% of all government spending.

The pie chart (page 23) implies food stamps are part of discretionary spending when in fact they are part of a mandatory program.

Federal Mandatory and Discretionary Spending

- Health 28%
- Social Security 25.3%
- Defense/Homeland Security 16.2%
- Remainder 9.1%
- Veterans 4%
- Transportation 4%
- Food & Agriculture 4%
- Education 3%
- Int'l Affairs 2%
- Housing 1%
- Energy 1%
- Science 1%
- Labor 1%

16
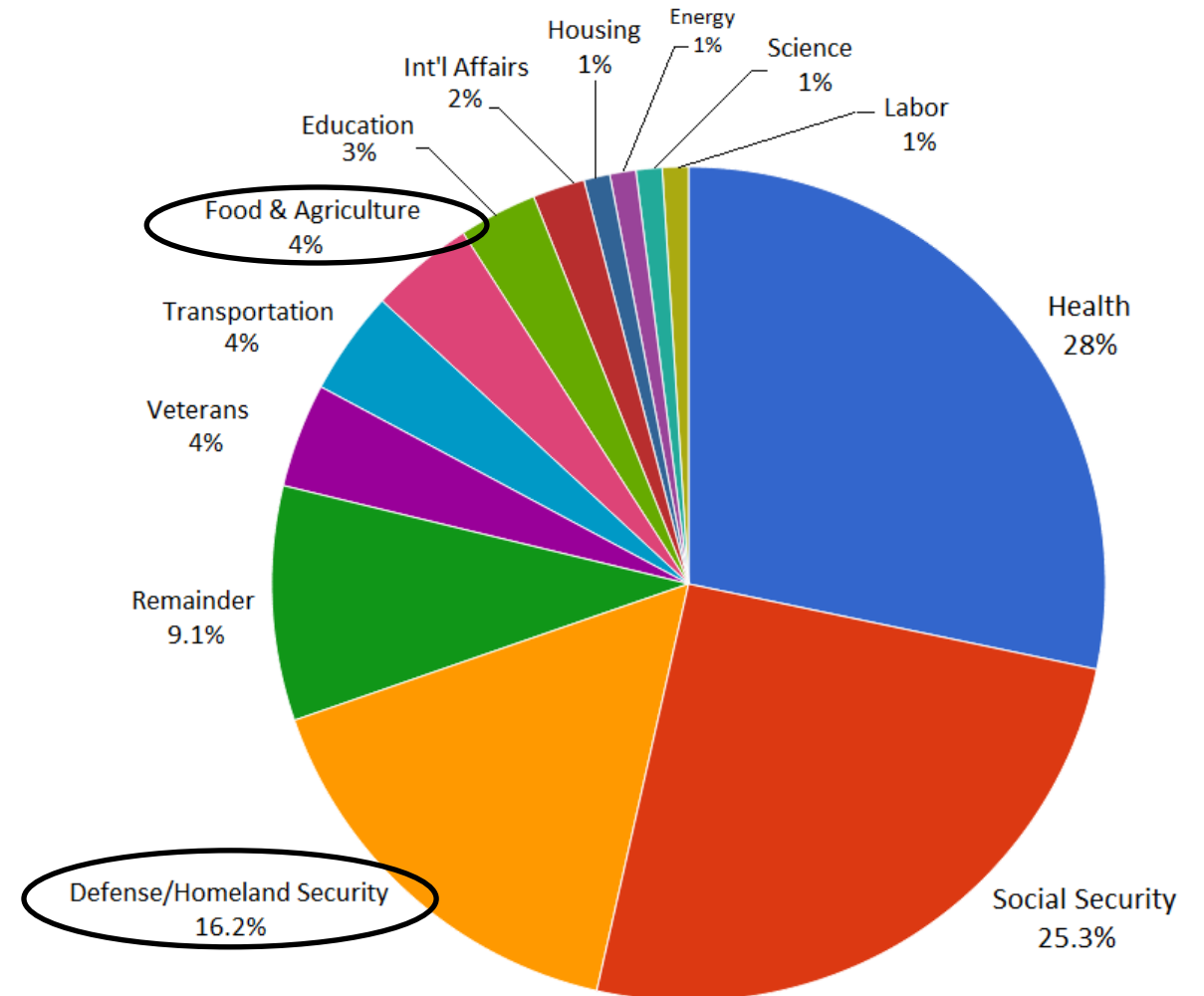
…by Displaying Insufficient Information

Federal Mandatory and Discretionary Spending

The pie chart on slide 24 reconstructed as a bar chart.

Health and Human Services	28%
Social Security	25
Defense and homeland security	16
Remainder*	9
Veterans	4
Transportation	4
Food and agriculture	4
Education	3
State Department and foreign aid	2
Housing and Urban Development	1
Energy and EPA	1
Science and NASA	1
Labor	1

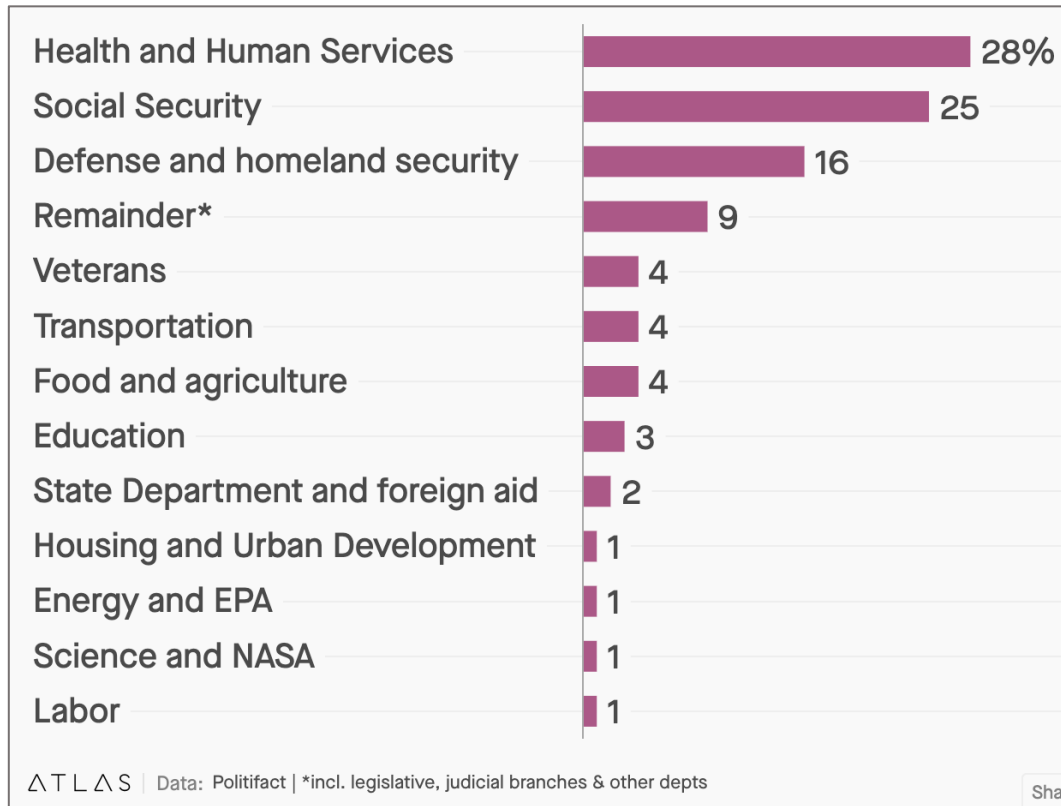ATLAS | Data: Politifact | *incl. legislative, judicial branches & other depts

Shar

Take-aways:

- If a reference to the the data source is provided check it out.
- If the data source is not provided (as in this case) ask people who have subject matter expertise that you trust.
- Pie charts are easy to interpret when the area of a slice is 25% or greater. When this is not the case a different visualization is advisable, for example the bar chart on the left.
- The Fact-checking Day website (https://factcheckingday.com) is dedicated to teaching information literacy and may help you determine the trustworthiness of a visualization.

…by Displaying Point Measurements in Areal Units

This is a typical U.S. election map. The map displays the election results aggregated by **county** - counties where candidate 3 received the most votes are red, counties where candidate 2 received the most votes are grey.

How the **citizens** voted or not voted is displayed in the bar chart.

How Registered Voters Responded in a National Election
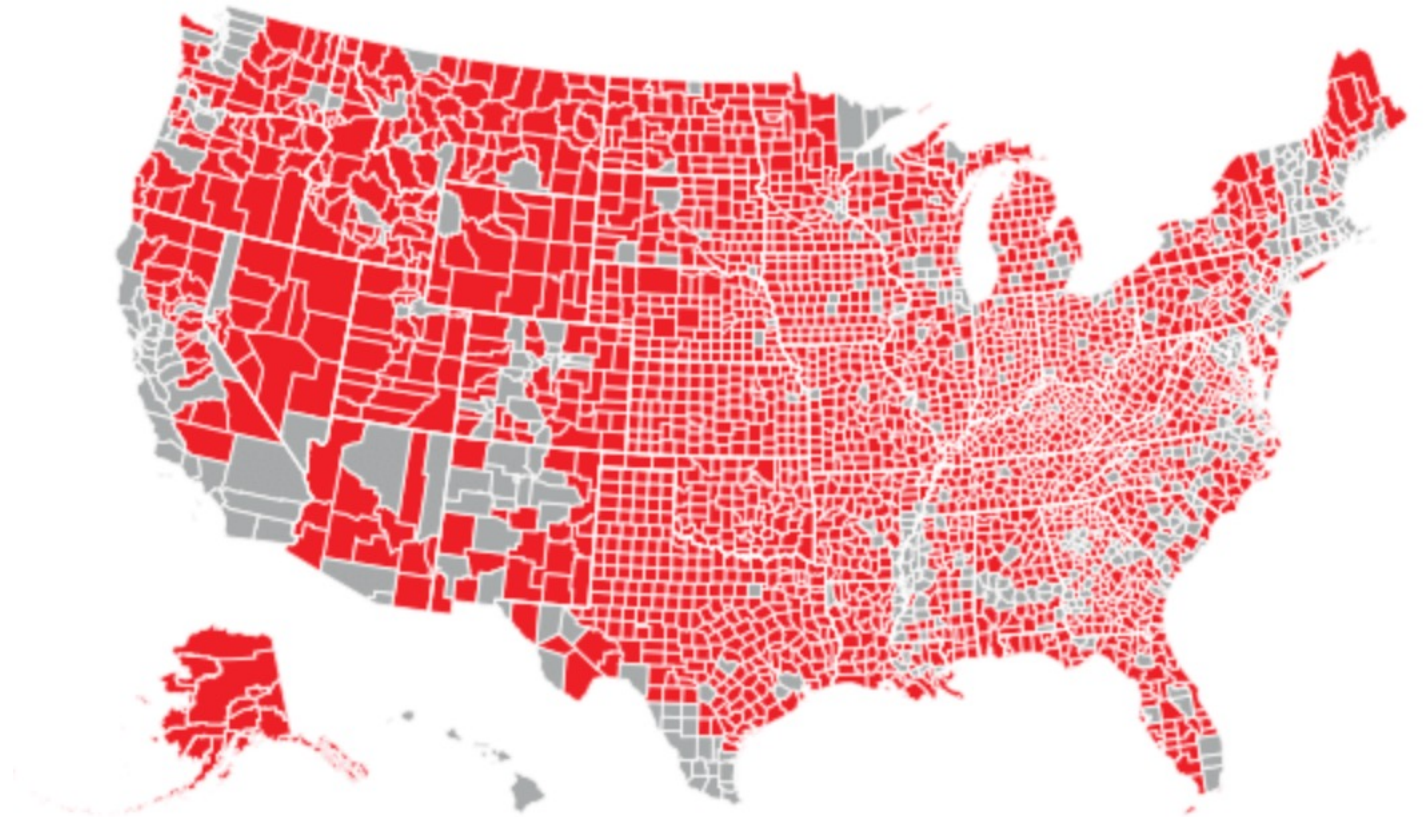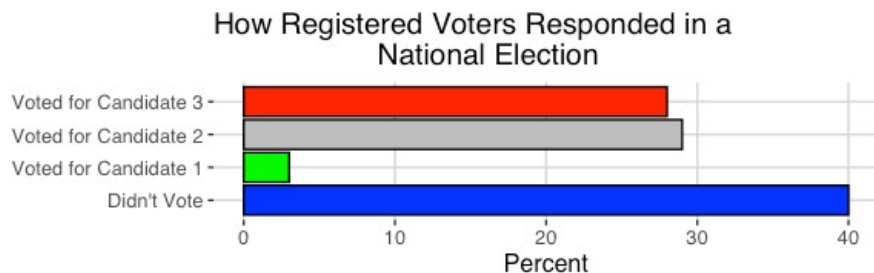
…by Displaying Point Measurements in Areal Units

The map fails to consider the population distribution. It fails to account for the fact that counties can be large in size but are sparsely populated.

http://www-personal.umich.edu/~mejn/election/2016/
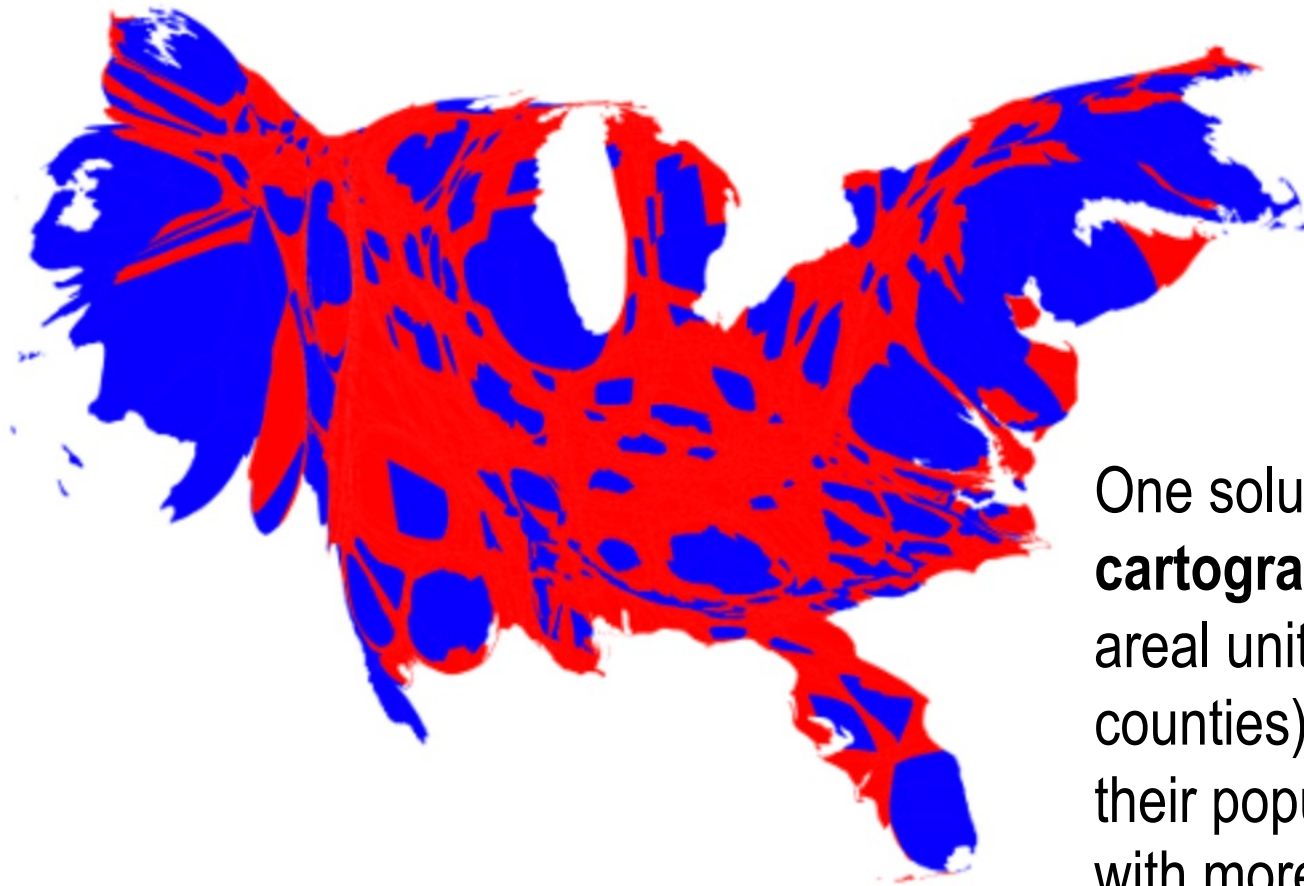
One solution to this is a **cartogram**, a map where the areal units (in this case, counties) are sized based on their population, so counties with more people appear much bigger.

…by Displaying Point Measurements in Areal Units

Here is another alternative (to the map on slide 29) by cartographer Kenneth Field where each dot (positioned approximately not exactly) represents a voter. Notice the vast areas of no dots.
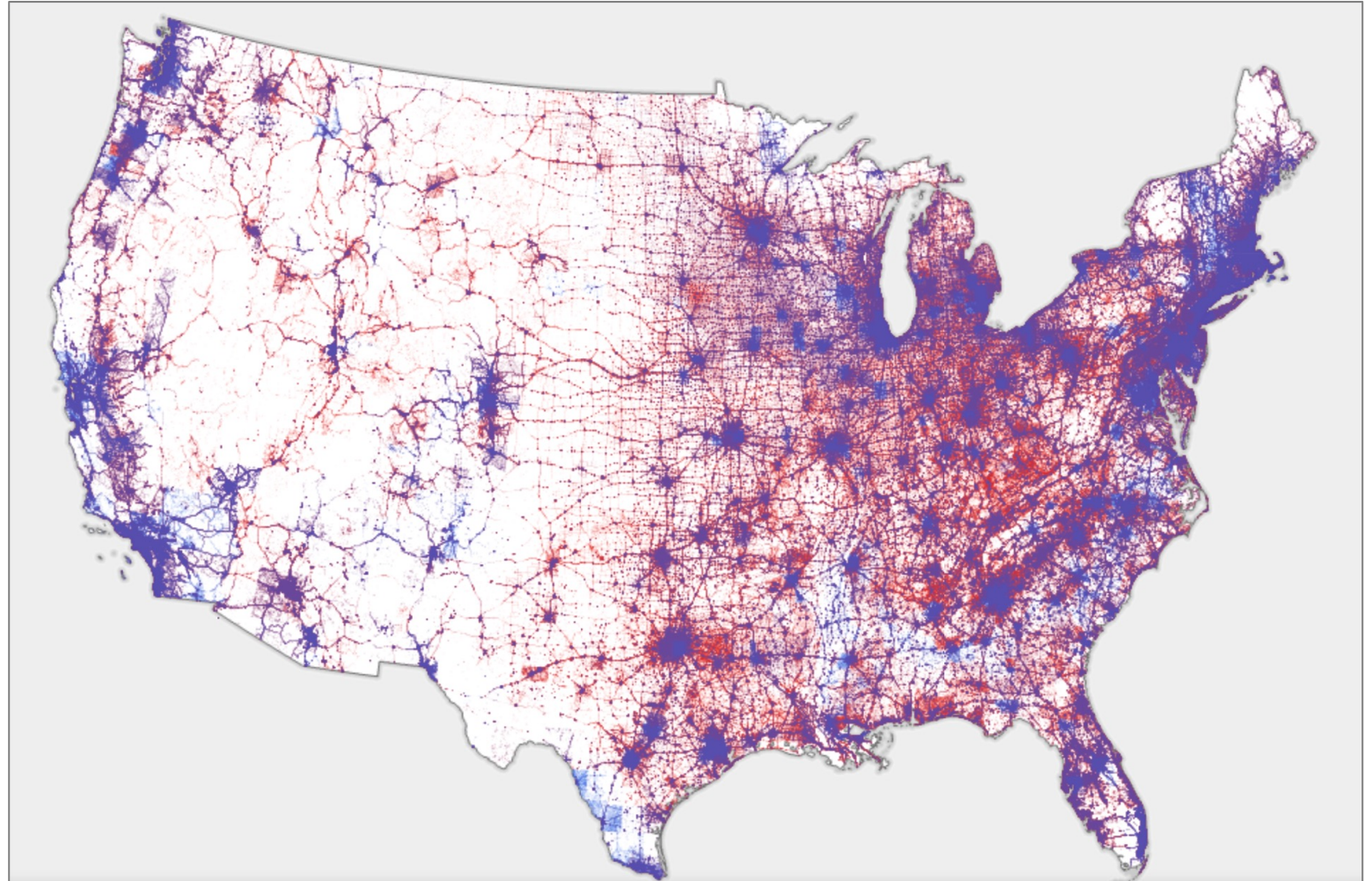
# …by Displaying Point Measurements in Areal Units

Take-aways:

- Be aware – of the "*modifiable areal unit problem*" with maps – when point based measures are aggregated by an areal unit (for example, county or state). The size of the areal unit can bias the interpretation if the visualization does not consider the total number possible point measures per areal unit.

- One-way points measures can be normalized to the areal unit is to divide them by the total number of possible point measures in the areal unit. For example, the two maps on the right the top displays the total number of COVID-19 cases/state, the bottom the total number of COVID-19 cases/million.

11/12/2020 Number of Active COVID-19 Cases


11/12/2020 Number of Active COVID-19 Cases/Million

https://coronavirus.1point3acres.com/#stat
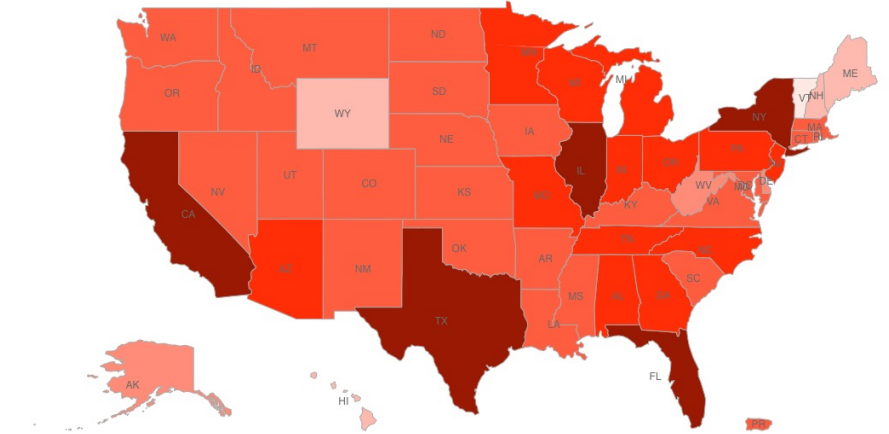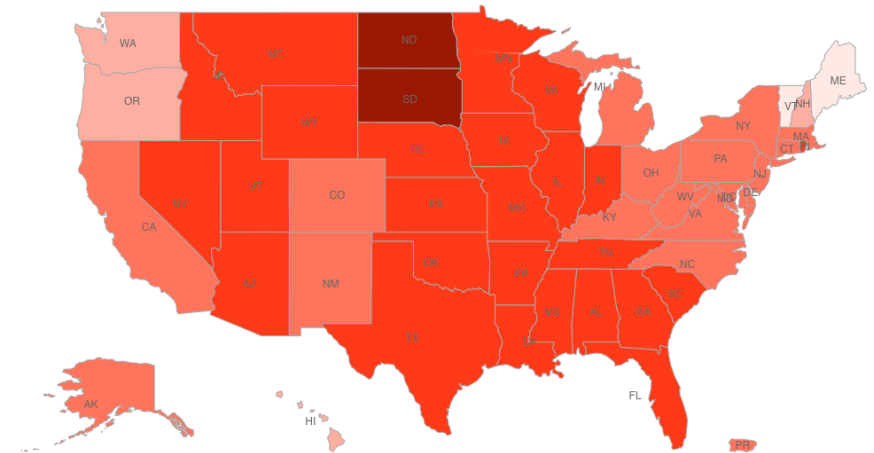
…by Concealing or Confusing Uncertainty

Visualizing the uncertainty in data is often ignored. Even when uncertainty is visualized it is often misinterpreted.

An exception to the visualization of uncertainty is ubiquitous during the hurricane season - the cone of uncertainty.

But what is not commonly know is how to interpret the cone…

- approximately one out of three times the hurricane will be outside the cone.
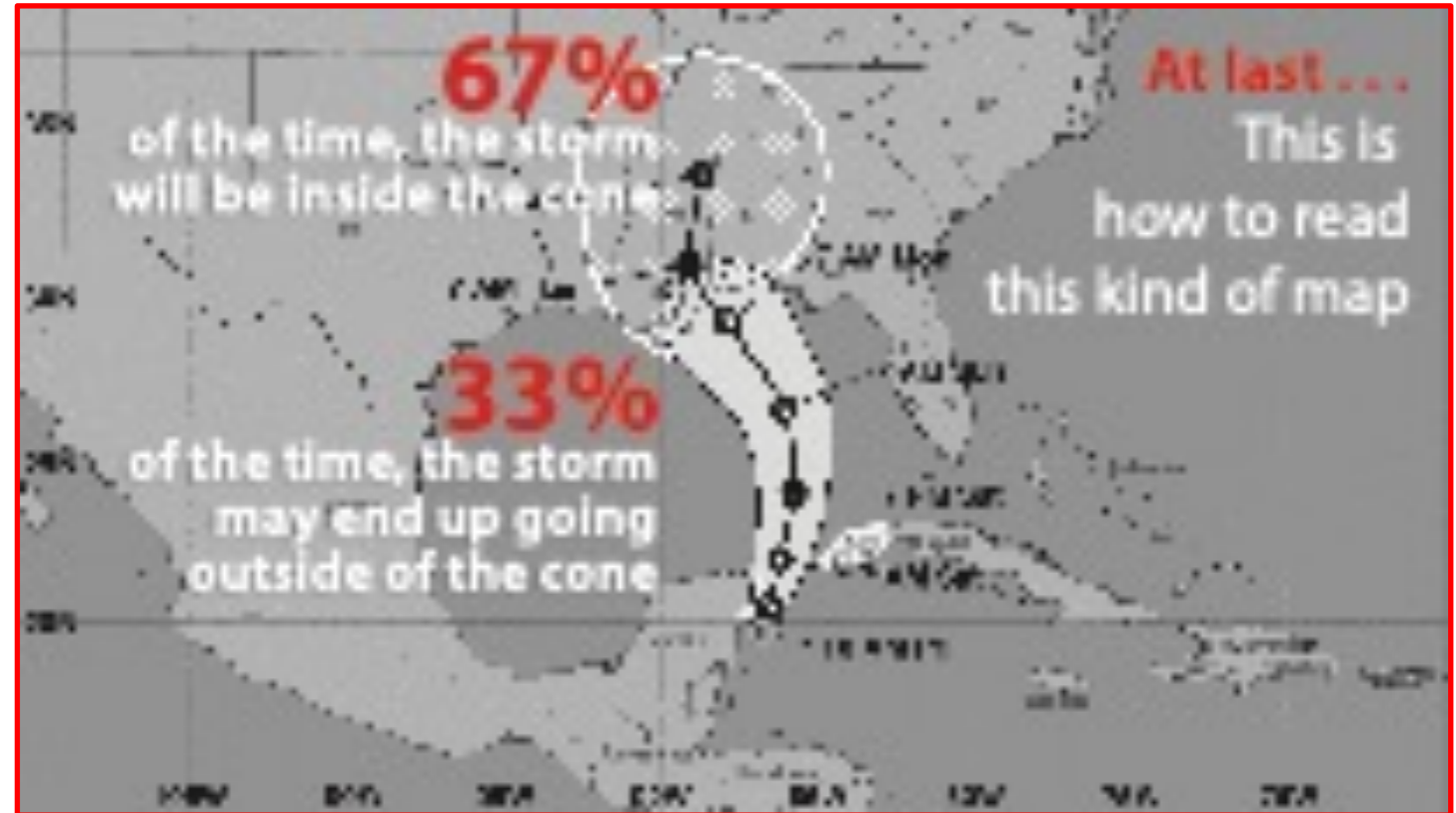

How Charts Lie (2019) by Alberto Cairo https://www.dropbox.com/sh/d1kb0jdrhkb43j9/AADTBfRvAh-mxmSxBRNZpLJja/5.CHAPTER5?dl=0&preview=PDF17.PercentagesRight.pdf&subfolder_nav_tracking=1

…by Concealing or Confusing Uncertainty

More commonly we are dealing with federal survey data like the American Community Survey (ACS) whose estimates come with a margin of error.

ACS estimates are based on a sample of the population this creates the uncertainty. Each time a random sample is taken the estimate will vary – this creates uncertainty in the estimate.

POPULATION

SAMPLE$_1$

SAMPLE$_2$

SAMPLE$_3$

Estimate$_1$

Estimate$_2$
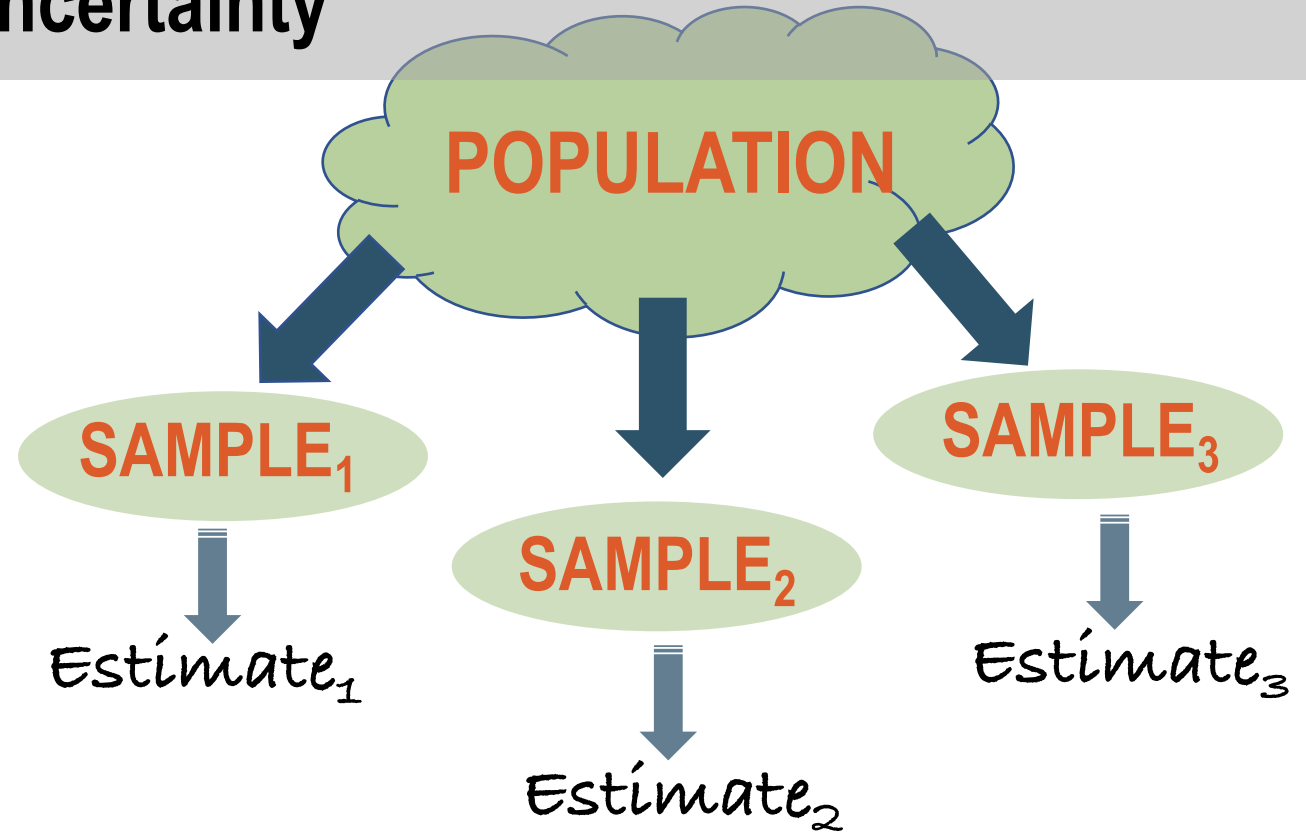
Estimate$_3$

Each time you take a sample and calculate the estimate – the estimate will be slightly different. The variation between the estimates is referred to as **sampling error**.

…by Concealing or Confusing Uncertainty

The **Census Household Pulse Survey** is collecting data to measure household experiences during the coronavirus pandemic. The bar chart below displays the percentage of adults in households where there was food scarcity in the last seven days along with a measure of variability.
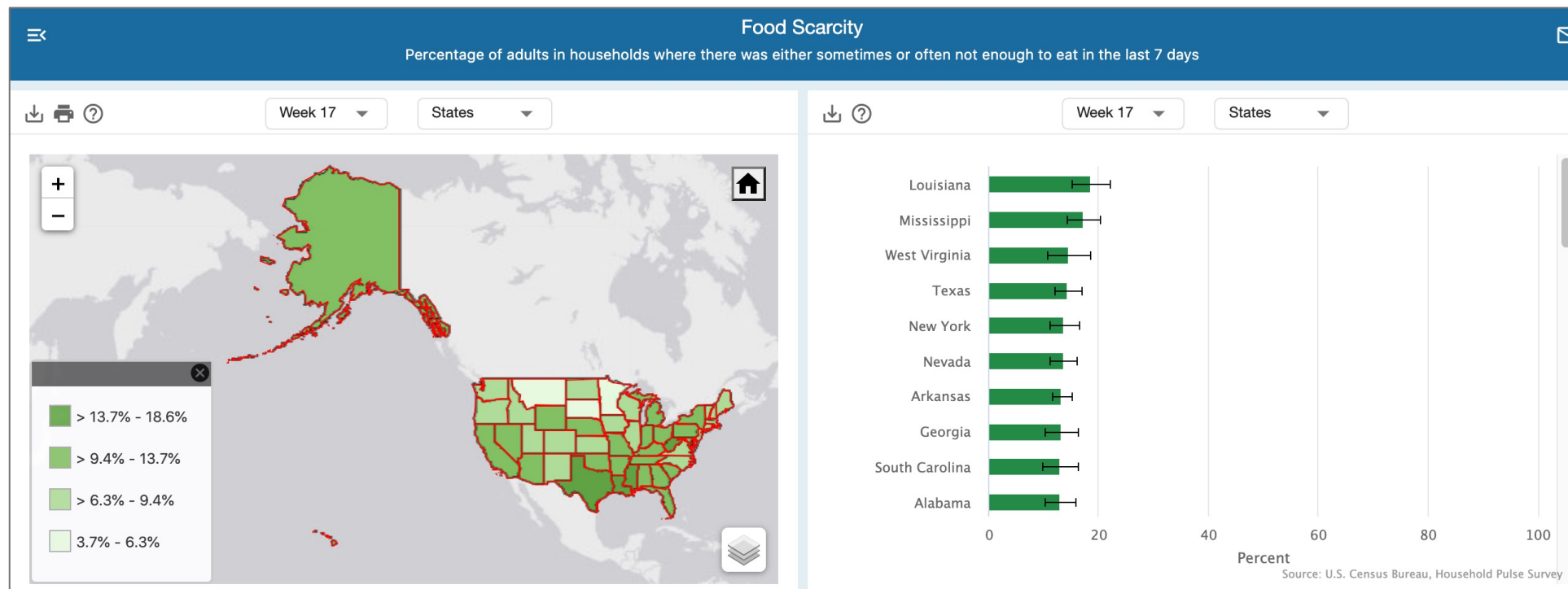
CENSUS Household Pulse Survey: https://www.census.gov/data-tools/demo/hhp/#/

…by Concealing or Confusing Uncertainty

Week	Area	CENSUS Household Pulse Survey: https://www.census.gov/data-tools/demo/hhp/#/	Total Individual Population age 18+	Measure Universe	Total Food Scarcity Number	Margin of Error +/-	Food Scarcity Percent Percent	Percent Margin of Error +/-
17	United States		249,170,916	218,462,392	23,883,801	1,078,539	10.9	0.5
17	Alabama		3,717,378	3,150,313	407,943	85,786	12.9	2.8
17	Alaska		524,925	484,809	62,482	11,626	12.9	2.4
17	Arizona		5,597,268	5,002,263	645,740	148,946	12.9	2.9
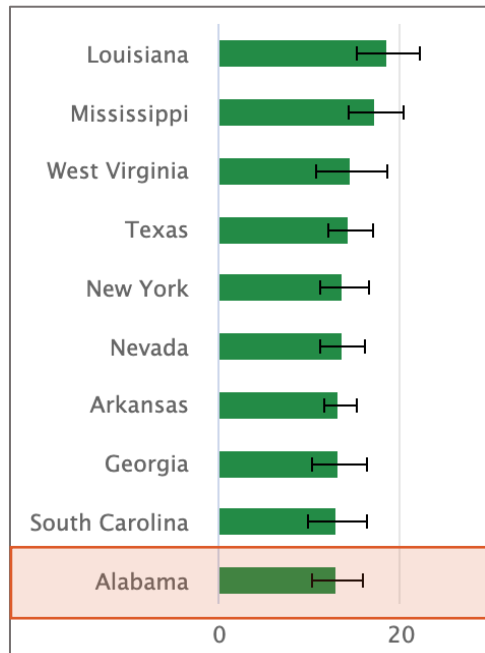
On the left is a bar chart that displays the percentage of households reporting food scarcity household with the margin of error.

The percent margin of error is interpreted as: if we were to sample AL households again the estimate would lie between 10.1 and 15.7, 9 times out of 10.

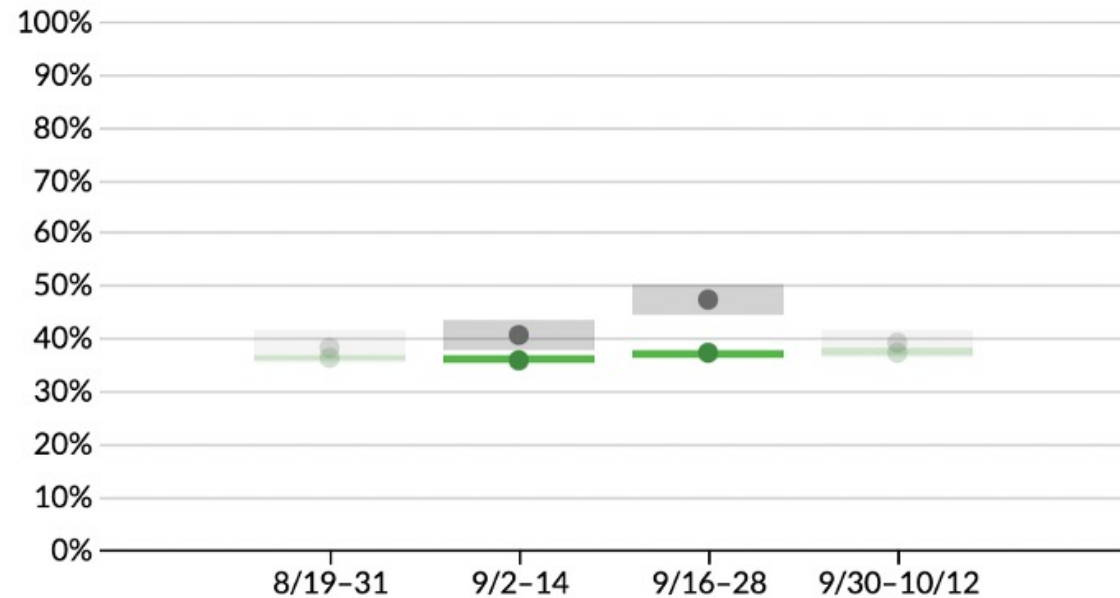$$12.9 \pm 2.8 = [10.1, 15.7]$$

90% Confidence Interval

…by Concealing or Confusing Uncertainty

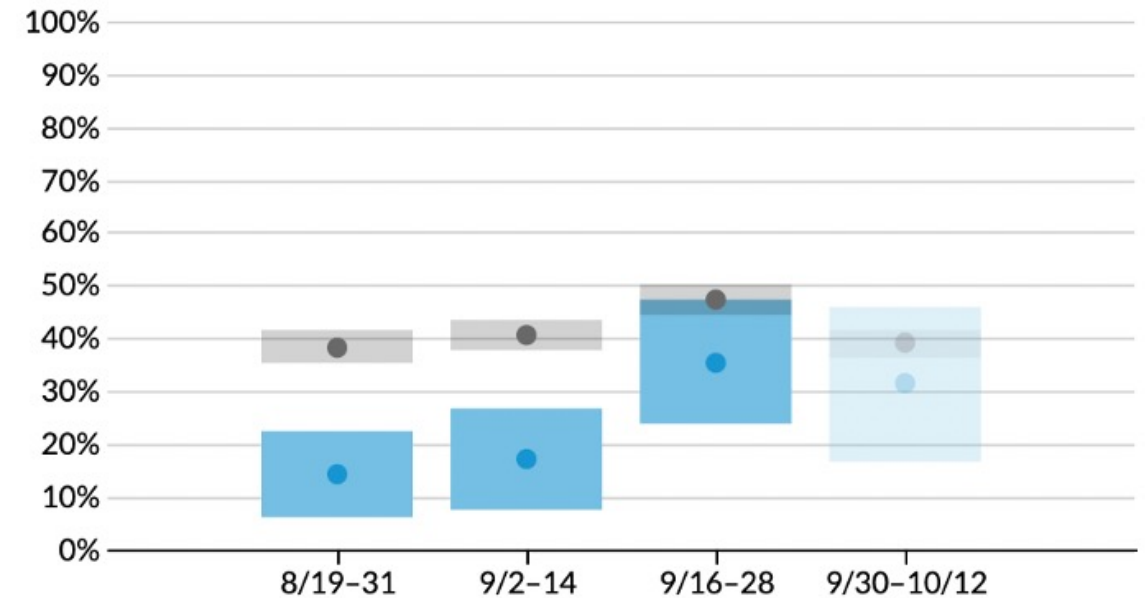**Share of adults that experienced symptoms of depression or anxiety disorders in the past week**


OREGON

…by Concealing or Confusing Uncertainty

Take-aways:

- You are more likely to see uncertainty reported in data tables than in a visualization. Even if you don't see uncertainty represented in a visualization don't assume it the data point is a precise representation
- One of the most challenging aspects of data visualization is the visualization of uncertainty. For example, think about how you might display uncertainty on a map.
- Nearly every data set we work with has some uncertainty, and whether and how we choose to represent this uncertainty can make a major difference in how accurately we perceive the meaning of the data.

Computer and Mathematical Scientists as a Percent of All Occupations by State

Area of the orange rectangle displays the SE

Area of the orange square displays the SE
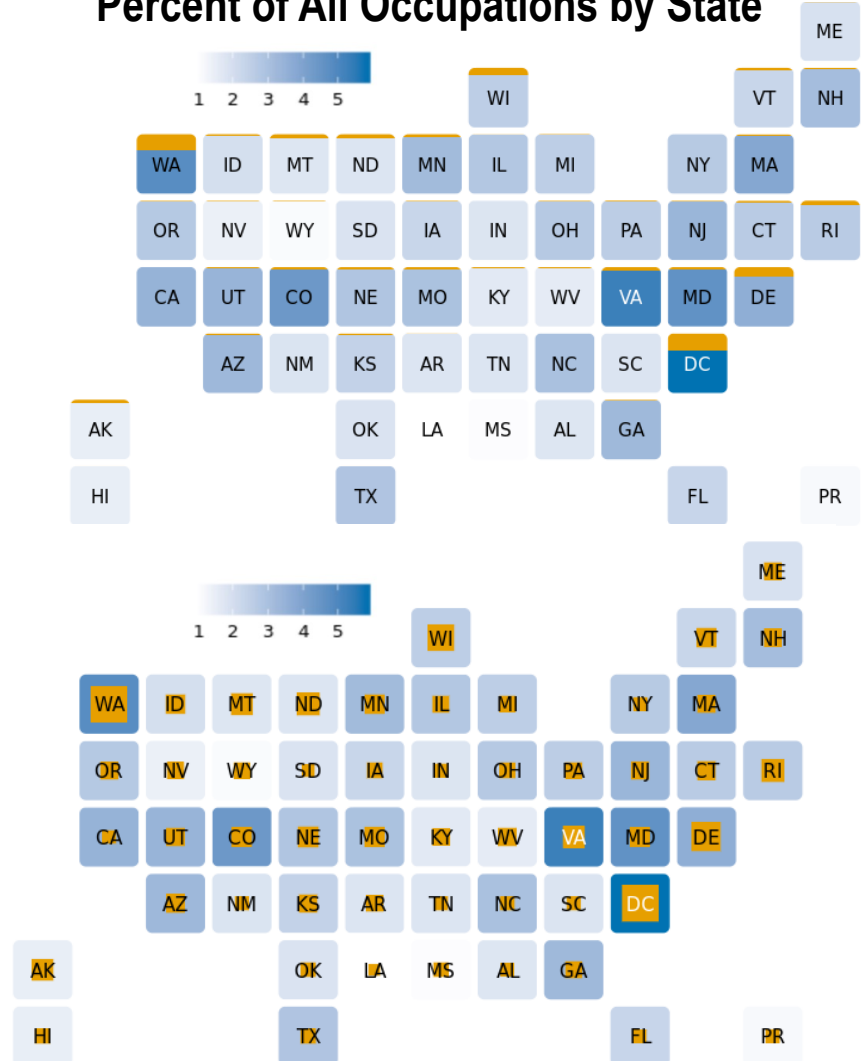
QUESTIONS?

WE ARE HERE TO HELP!

https://datascienceforthepublicgood.com/economic-mobility

CONTACT:

VICKI LANCASTER

Social & Decision Analytics Division
Biocomplexity Institute & Initiative | University of Virginia
Suite 2910 | 1100 Wilson Blvd | Arlington, VA   22209
val7zv@virginia.edu