

VISUALIZING DATA

Veracity, Accuracy, Accountability



Alberto Cairo

Dialogues in Research Ethics
University of Miami, 2020

We are living through a Golden Age of visualization



We are living through a Golden Age of visualization



Graphic by the Electronic Visualization Laboratory (EVL) at the University of Illinois at Chicago (UIC)

The most-viewed piece *ever* published by The Washington Post online

Q

Sections

The Washington Post

Democracy Dies in Darkness

Alberto Cairo To...

f

t

e

i

n

p

t

2.6k

Health

Why outbreaks like coronavirus spread exponentially, and how to “flatten the curve”

By Harry Stevens March 14, 2020

PLEASE NOTE

The Washington Post is providing this story for free so that all readers have access to this important information about the coronavirus. For more free stories, [sign up for our daily Coronavirus Updates newsletter](#).

<https://www.washingtonpost.com/graphics/2020/world/corona-simulator/>

Common misconceptions when talking about visualization:

1. “A picture is worth a thousand words”
2. “Visualization is intuitive”
3. “The data should speak for itself”
4. “Show, don’t tell!”

Visualizations can't be designed based just on our personal preferences—although these *are* important.

Visualization is a bit like writing: beyond some conventions and constraints regarding symbols, visual grammar, perception, and cognition, visualization **can't be based on “rules” that are set in stone.**

Instead, when designing visualizations, we need to be guided by **reasoned, justifiable choices.**

“Facts give us **reasons** [...] when they count in favor of our having some belief or desire, or acting in some way.”

Derek Parfit, *On What Matters*

Reasoning about visualization.

Key questions:

1. Why to visualize?
2. What to visualize?
3. Who to visualize for?
4. How much to visualize?
5. How to visualize it?
6. What style to use?

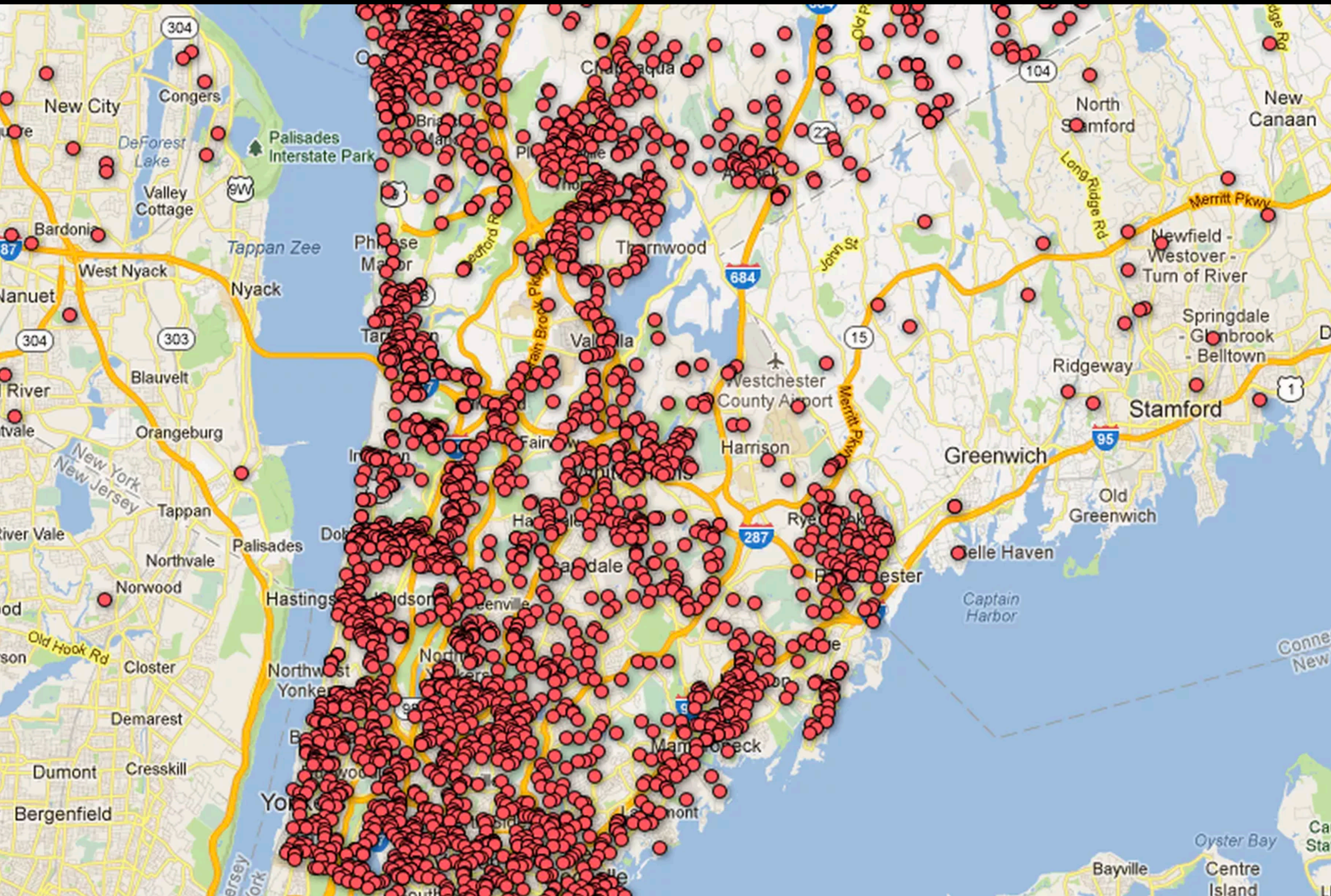
Note: this is just a tentative list of questions



I. Why should my visualization exist?

Do the potential benefits of designing my visualization outweigh the possible harm it might cause?

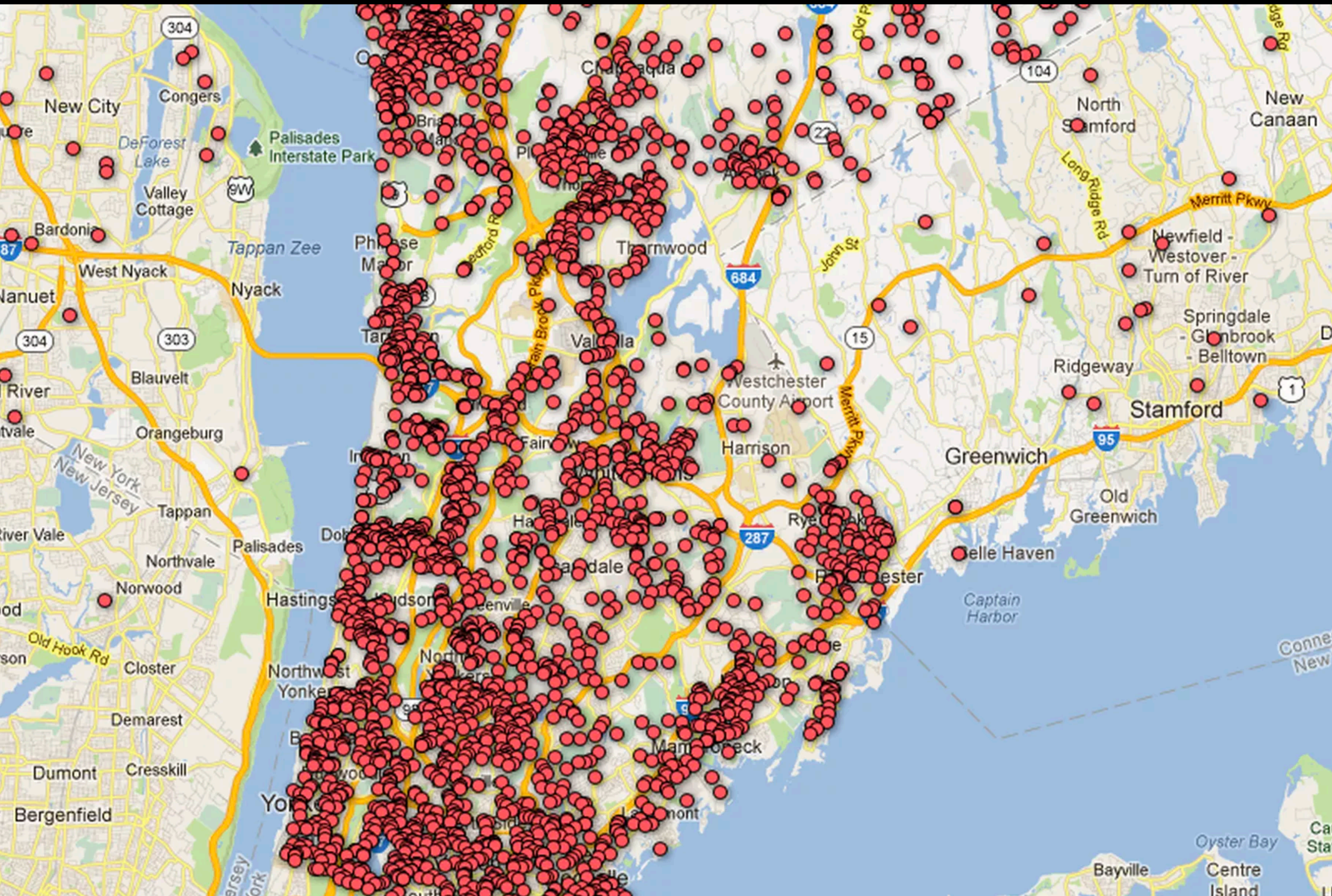
Published Dec. 23, 2012 (the Sandy Hook Elementary School shooting was on Dec. 14)



"Where are the gun permits in your neighborhood?" That's the question posed by **The Journal News**, a New York newspaper that published a Google map on Sunday that shows the names and addresses of pistol or revolver permits in Westchester and Rockland counties."

<https://www.theverge.com/2012/12/25/3802960/new-york-newspaper-posts-map-with-names-addresses-of-gun-owners>

Published Dec. 23, 2012 (the Sandy Hook Elementary School shooting was on Dec. 14)



“We felt sharing information about gun permits in our area was important in the aftermath of the Newtown shootings.”

**Janet Hasson,
president and
publisher of the
Journal News
Media Group**

Published Dec. 23, 2012 (the Sandy Hook Elementary School shooting was on Dec. 14)

WHY?

Why should this data be made public?

Why should it be made public through a map?

Why should it be *this type* of map?

Even if we decided that this data is worth publishing, wouldn't a different map be better?

What are the potential consequences of my decisions?

Are the benefits worth the risk of harm?



2. What to visualize?

Do I understand my data, its limitations, uncertainty, or glitches?
What or who is being measured (*o not being measured,*) and why?

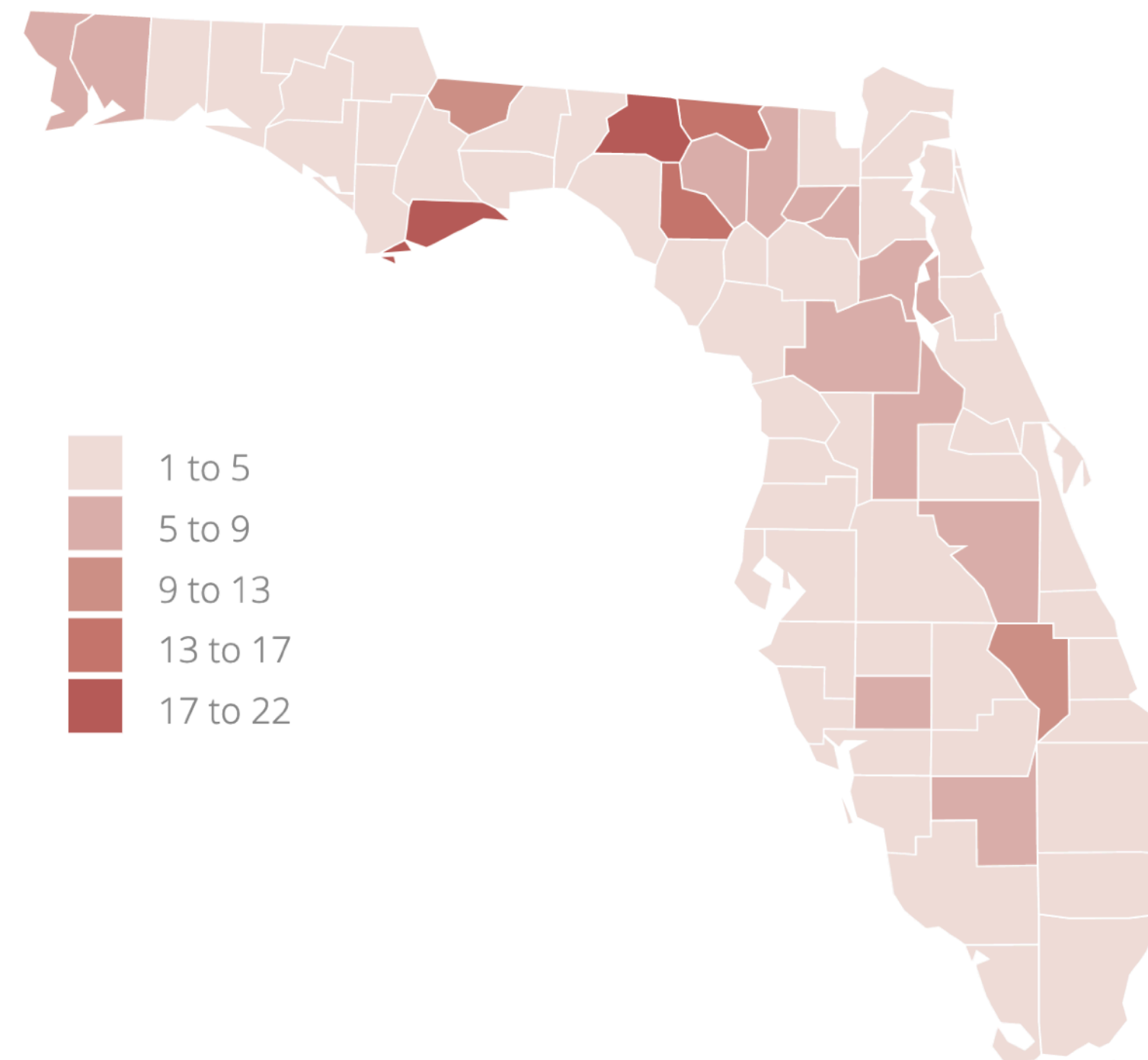
AT SCHOOL

WITHOUT A ROOF

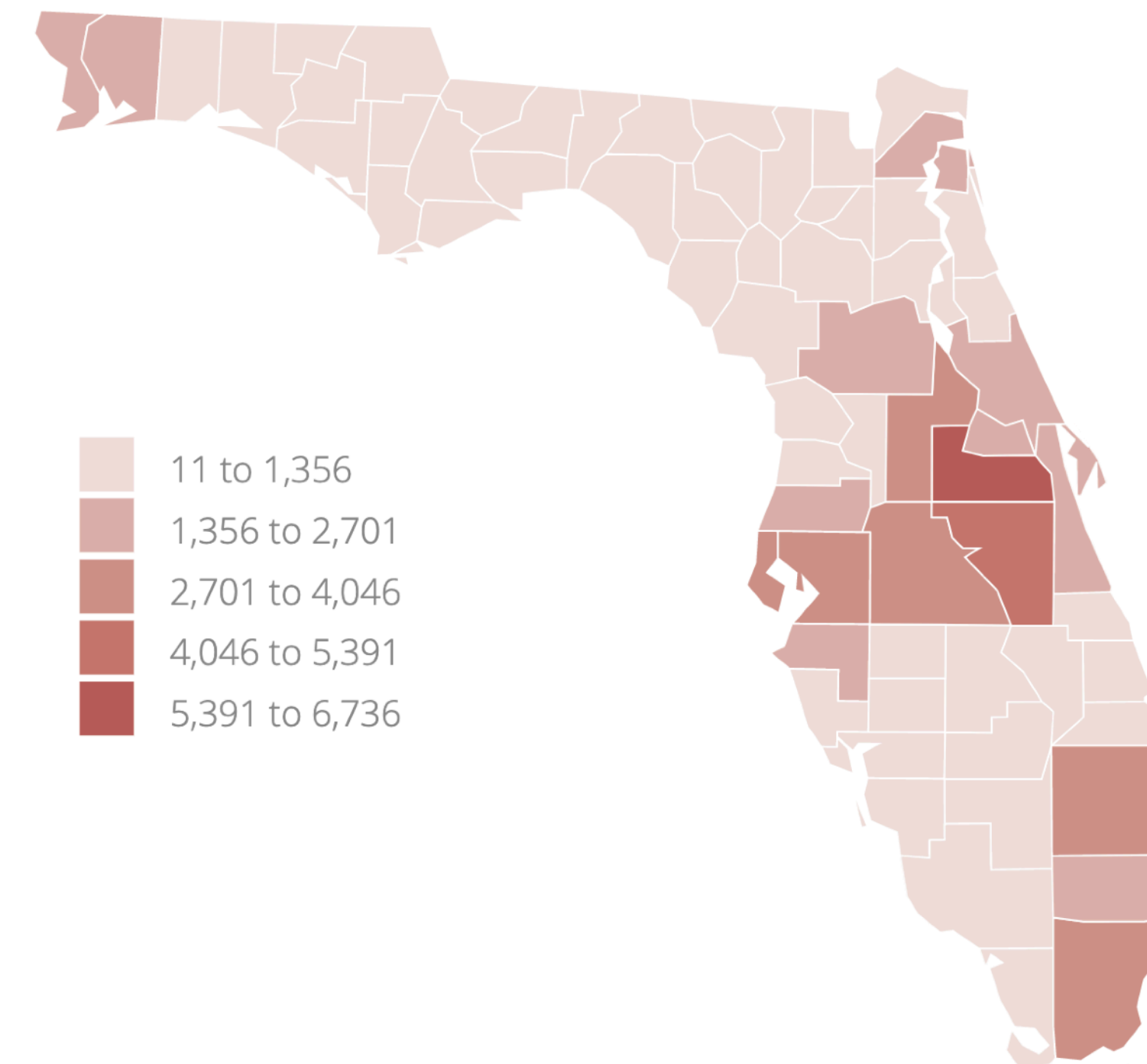
[http://
www.lmelgar.me/
without-a-roof/](http://www.lmelgar.me/without-a-roof/)

In Florida more than 71,000 students are homeless. During the last decade, this population rocketed as a result of the recession and how hard it has become for the poorest families to find affordable housing.

Percentage Total



Percentage Total



Kansas is the nation's porn capital, according to Pornhub

Blue states watch more porn. But what's the matter with Kansas?

According to [Pornhub Insights](#), Kansas leads the nation in porn pageviews per capita at roughly 194. They don't specify what interval this is over (monthly, weekly, etc), but the state-by-state comparison is nonetheless interesting.

Plotting Obama vote share in 2012 versus porn consumption, it looks like blue states consume more porn per capita than red ones. Aside from Kansas - a clear outlier - and Georgia, the remaining top ten per-capita porn consumers are all blue. Similarly, New Mexico and Maine are the only blue states in the bottom ten per-capita porn consumers.

<https://wonkviz.tumblr.com/post/82488570278/kansas-is-the-nations-porn-capital-according-to>

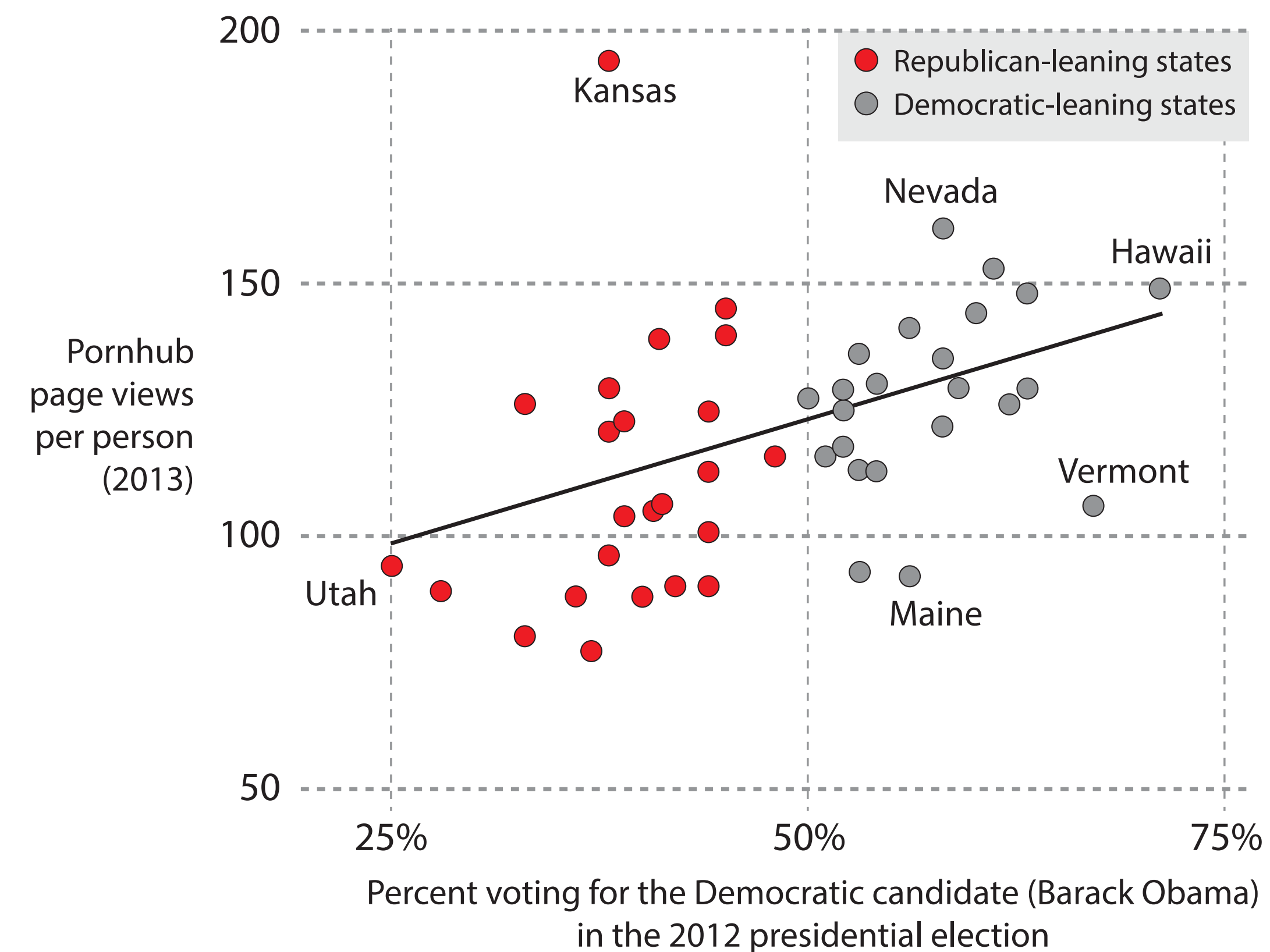
Kansas is the nation's porn capital, according to Pornhub

Blue states watch more porn. But what's the matter with Kansas?

According to [Pornhub Insights](#), Kansas leads the nation in porn pageviews per capita at roughly 194. They don't specify what interval this is over (monthly, weekly, etc), but the state-by-state comparison is nonetheless interesting.

Plotting Obama vote share in 2012 versus porn consumption, it looks like blue states consume more porn per capita than red ones. Aside from Kansas - a clear outlier - and Georgia, the remaining top ten per-capita porn consumers are all blue. Similarly, New Mexico and Maine are the only blue states in the bottom ten per-capita porn consumers.

<https://wonkviz.tumblr.com/post/82488570278/kansas-is-the-nations-porn-capital-according-to>



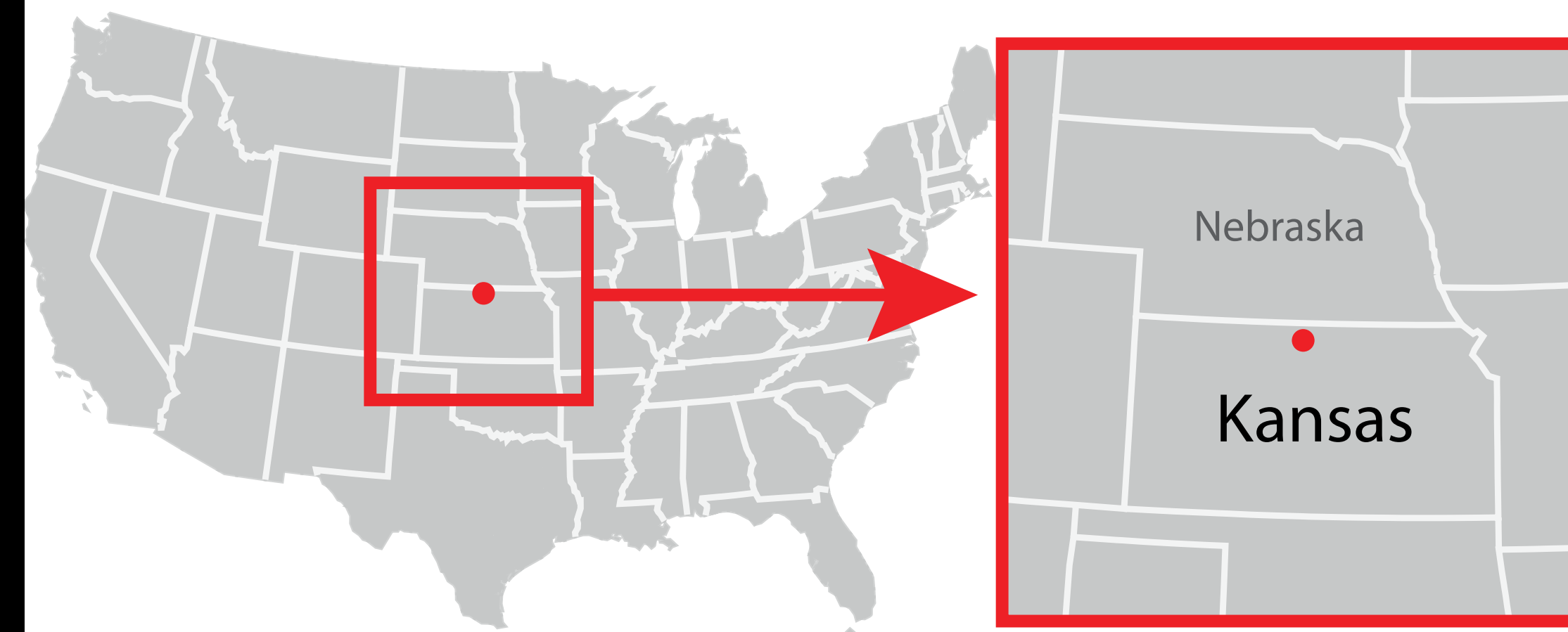
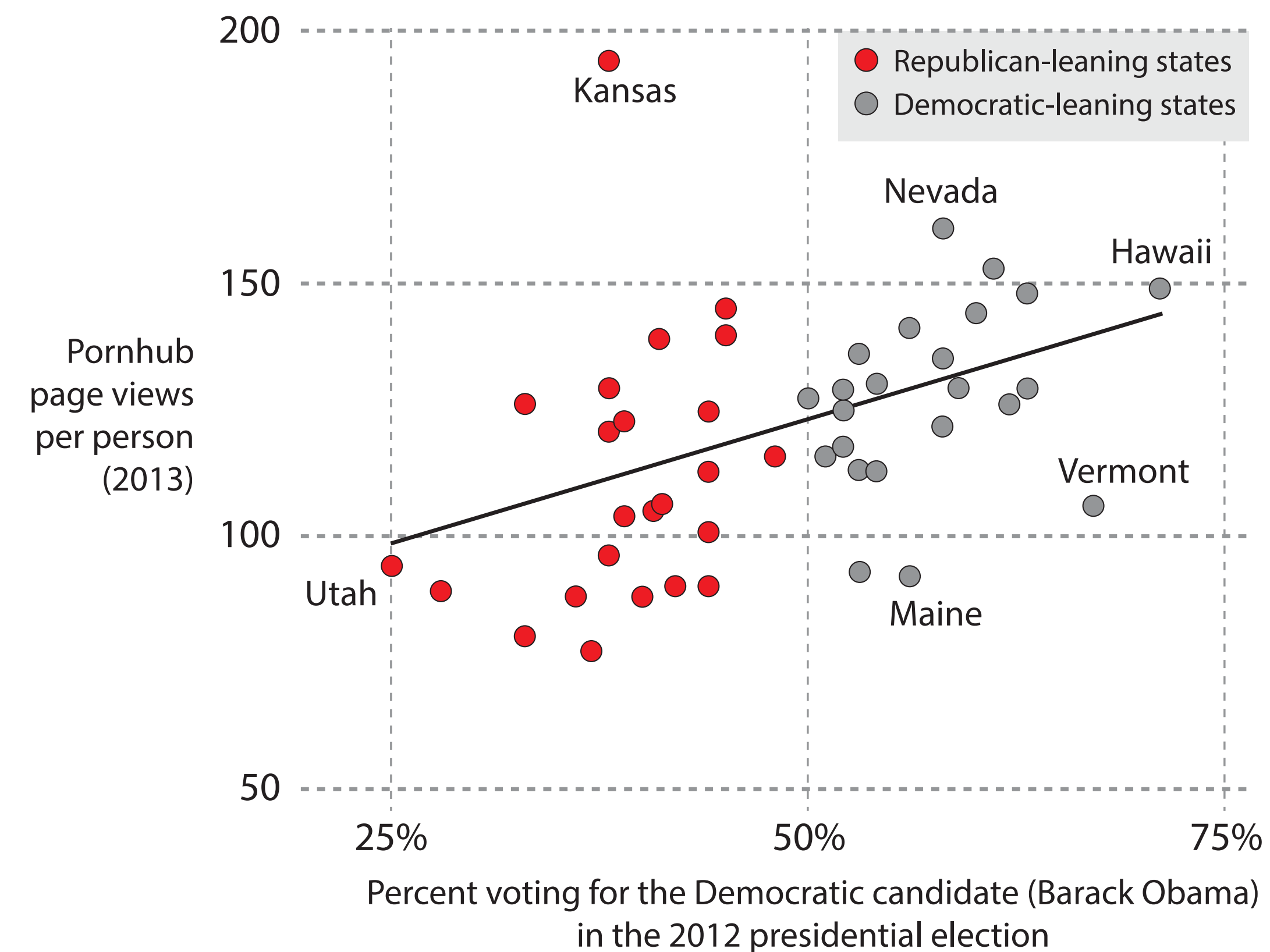
Kansas is the nation's porn capital, according to Pornhub

Blue states watch more porn. But what's the matter with Kansas?

According to [Pornhub Insights](#), Kansas leads the nation in porn pageviews per capita at roughly 194. They don't specify what interval this is over (monthly, weekly, etc), but the state-by-state comparison is nonetheless interesting.

Plotting Obama vote share in 2012 versus porn consumption, it looks like blue states consume more porn per capita than red ones. Aside from Kansas - a clear outlier - and Georgia, the remaining top ten per-capita porn consumers are all blue. Similarly, New Mexico and Maine are the only blue states in the bottom ten per-capita porn consumers.

<https://wonkviz.tumblr.com/post/82488570278/kansas-is-the-nations-porn-capital-according-to>



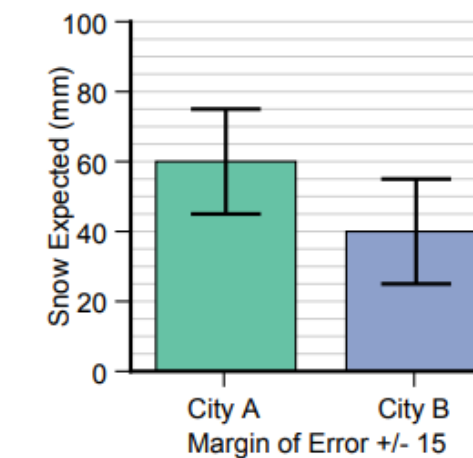
• Geographic center of the contiguous United States

Disclosing limitations and uncertainty

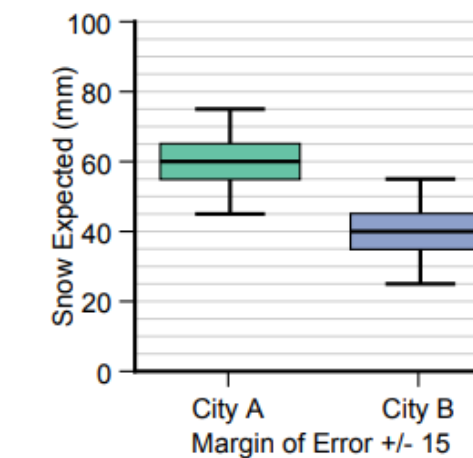
Uncertainty and graphicacy
How should statisticians, journalists, and designers reveal uncertainty in graphics for public consumption?

Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error

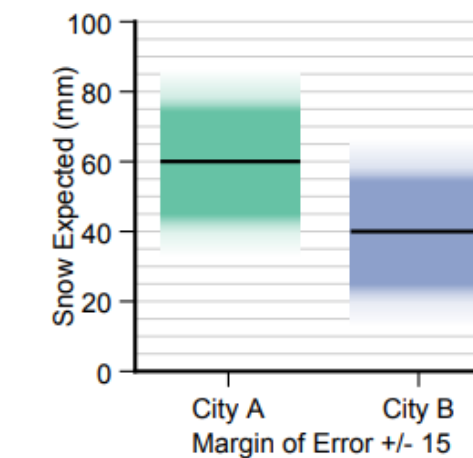
Michael Correll *Student Member, IEEE*, and Michael Gleicher *Member, IEEE*



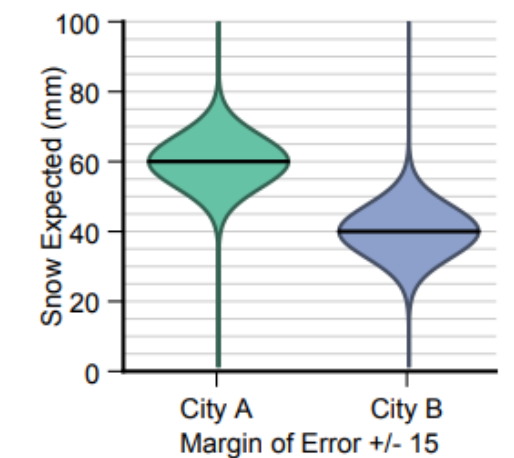
(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.



(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.



(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.



(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

<https://ec.europa.eu/eurostat/cros/powerfromstatistics/OR/PfS-OutlookReport-Cairo.pdf>

<https://graphics.cs.wisc.edu/Papers/2014/CGI4/Preprint.pdf>

Collection of papers about visualizing uncertainty:

<https://www.dropbox.com/sh/jk4ginxyai6ylqu/AABvqdyTlhJtyFN9nKNHyX9Ba?dl=0>



3. Who to visualize for?

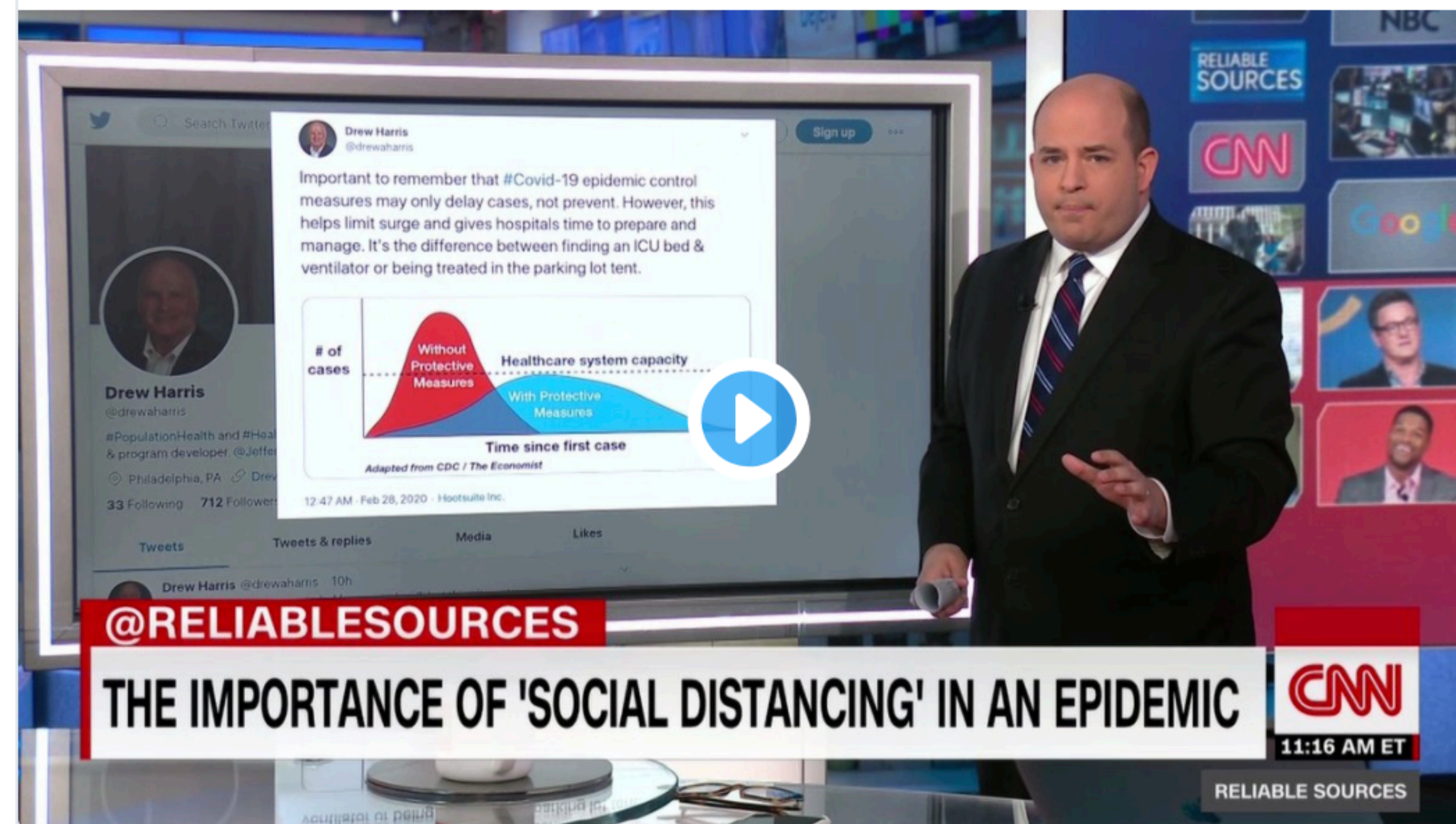
Have I thought about how my intended audience will access my graphic? Will they understand it? Can I explain it better?



Good journalism isn't just showing charts. It's also about explaining them: twitter.com/brianstelter/s...

Brian Stelter  @brianstelter

This infographic is worth a thousand words – showing why "social distancing" and other protective measures helps to slow an outbreak. Hat tips to CDC, @theeconomist, @drewharris, and @CT_Bergstrom

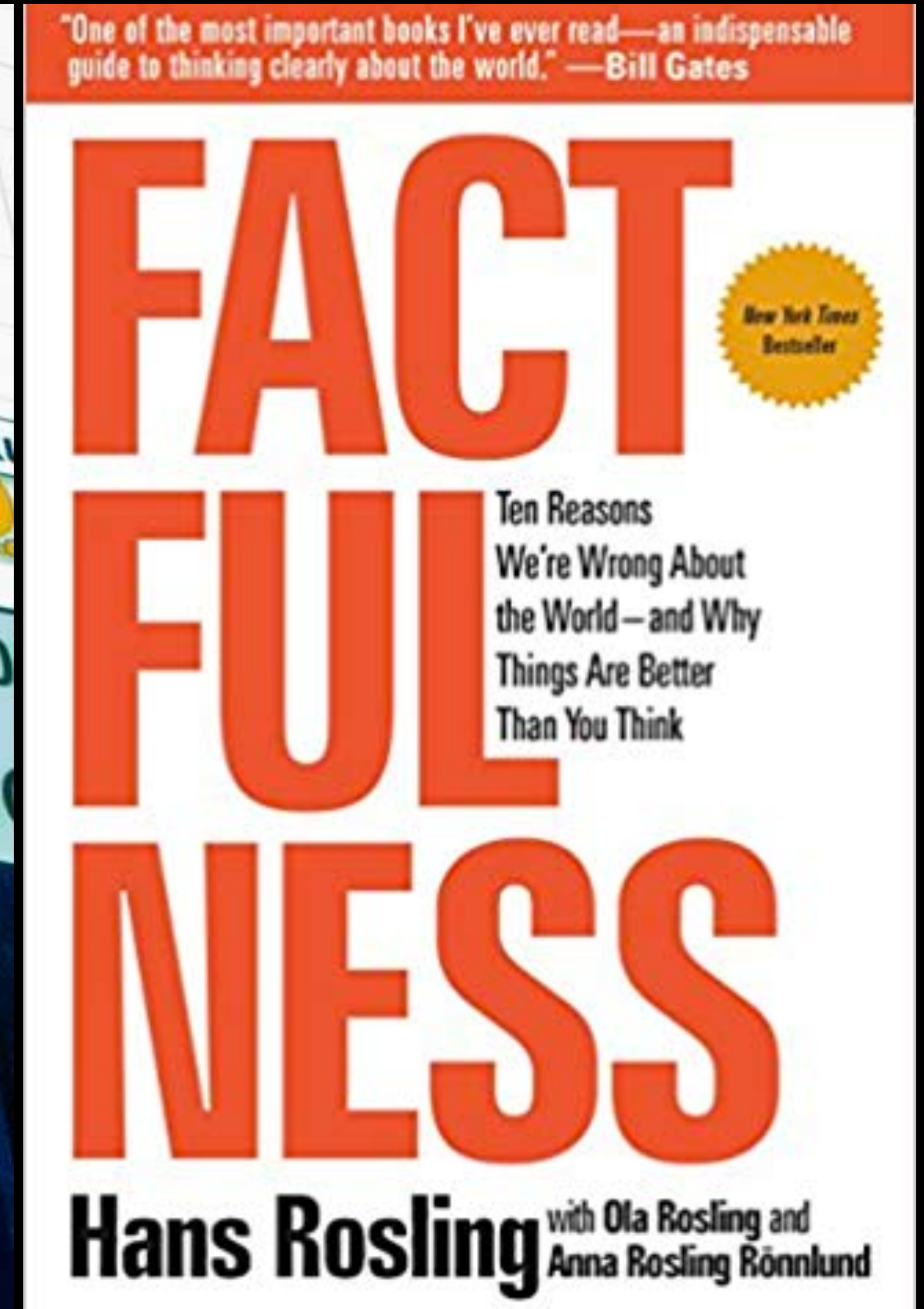
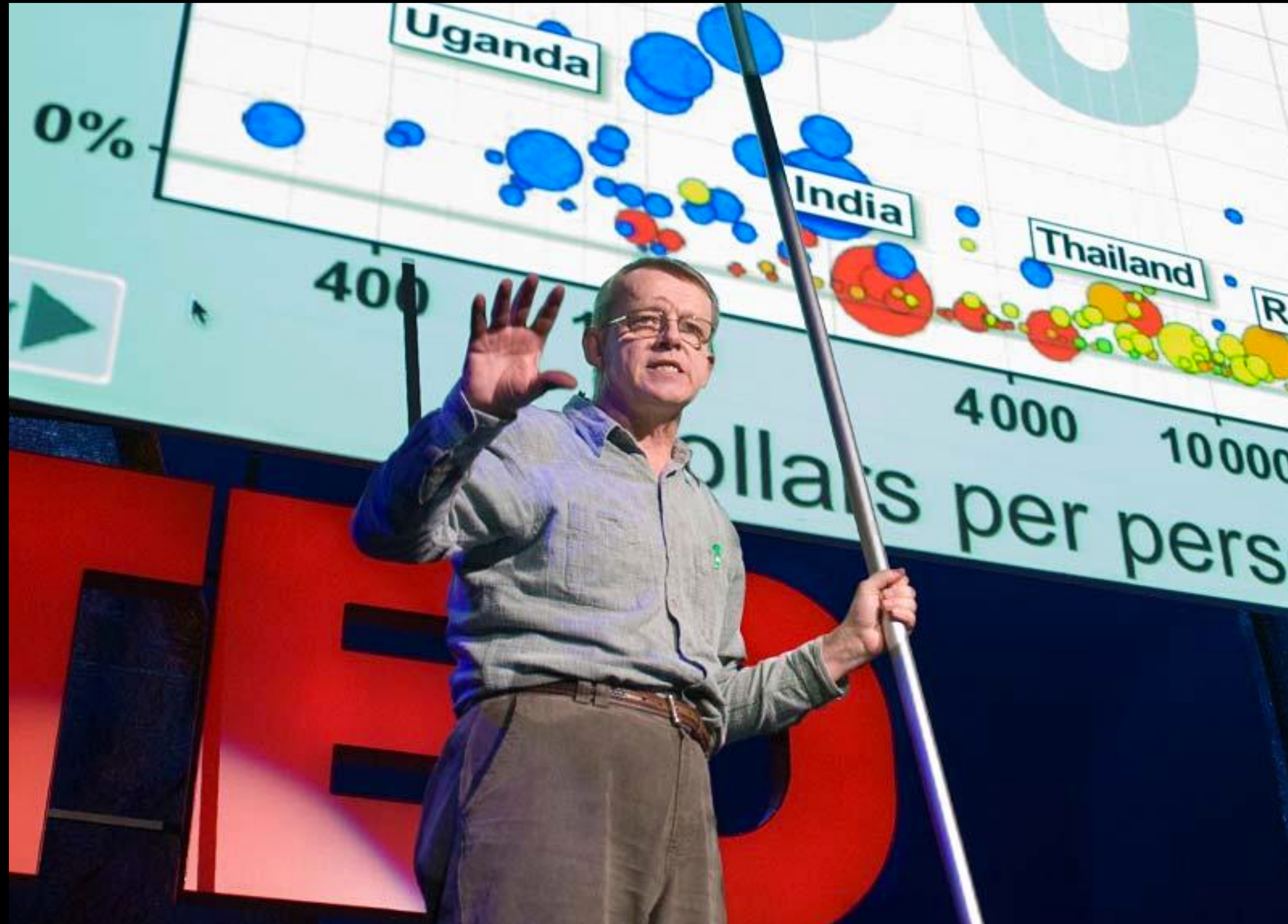


♡ 227 5:58 PM - Mar 8, 2020



<https://twitter.com/AlbertoCairo/status/1236773377865658370>

Show **AND** tell



Hans Rosling, www.gapminder.org

Show **AND** tell

BBC FOUR

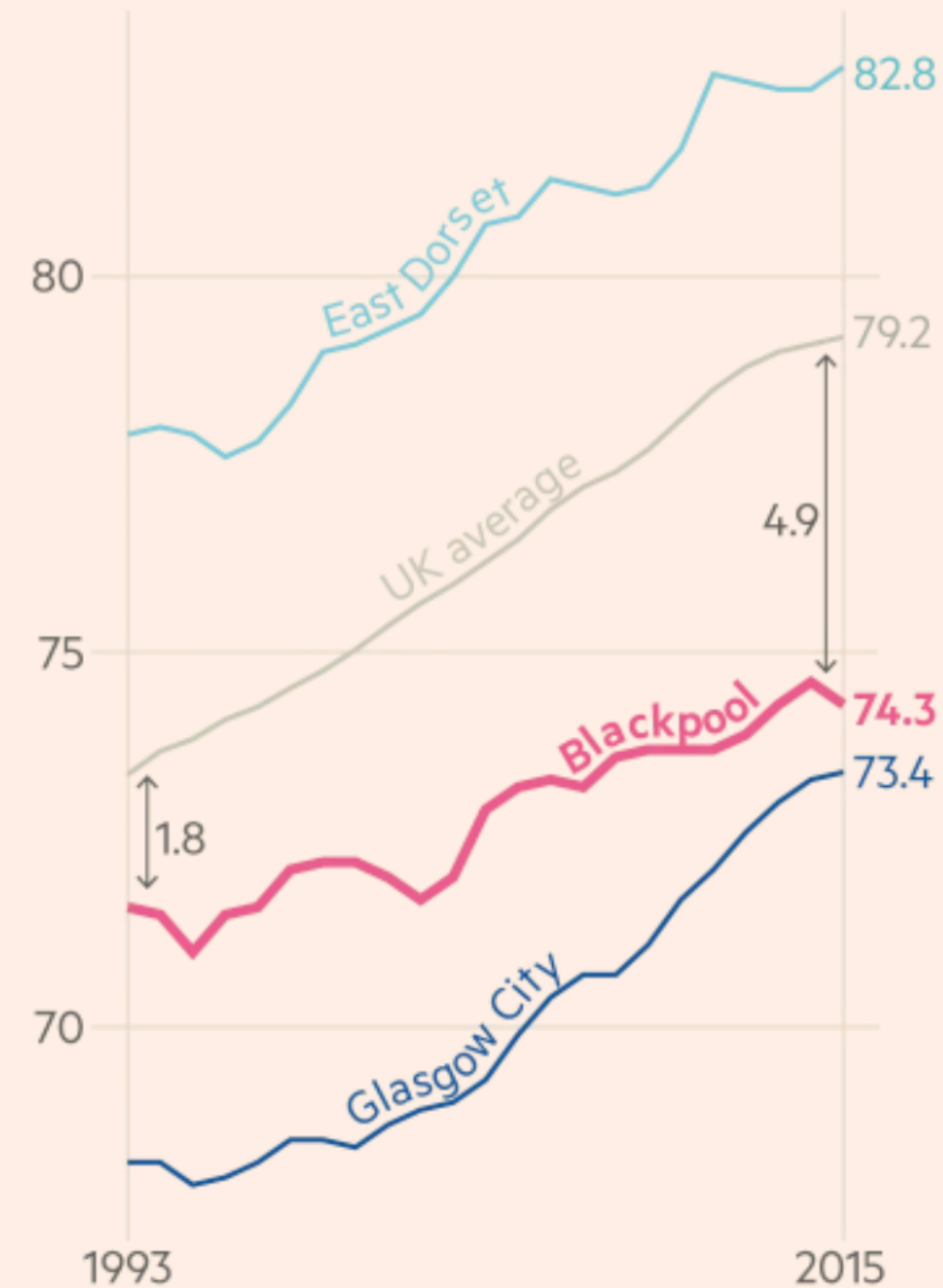


Hans Rosling, *The Joy of Stats*

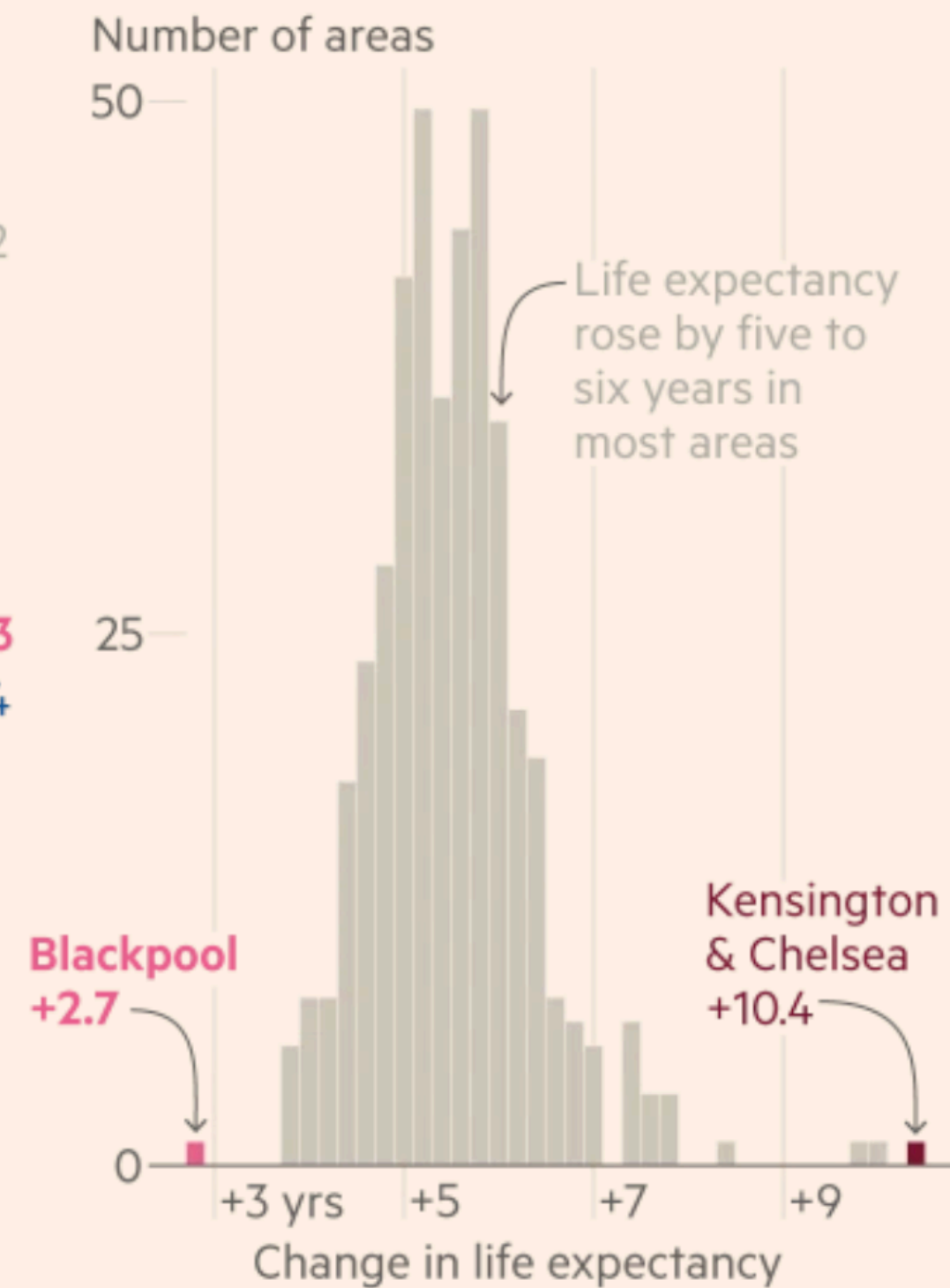
Alberto Cairo • University of Miami • www.thefunctionalart.com • Twitter: @albertocairo

Boys born in **Blackpool** can expect to live just 74 years — the second lowest in the UK, and up by just 2.7 years since 1993

Male life expectancy at birth in selected local authorities, 1993-2015



Distribution of change in male life expectancy at birth from 1993 to 2015, all UK local authorities



Source: ONS

Graphic by John Burn-Murdoch / @jburnmurdoch

© FT

“I and my colleagues here at the FT, we really do think one of the most valuable things we can do as data visualization practitioners is add this expert annotation layer.”

John Burn-Murdoch

Financial Times

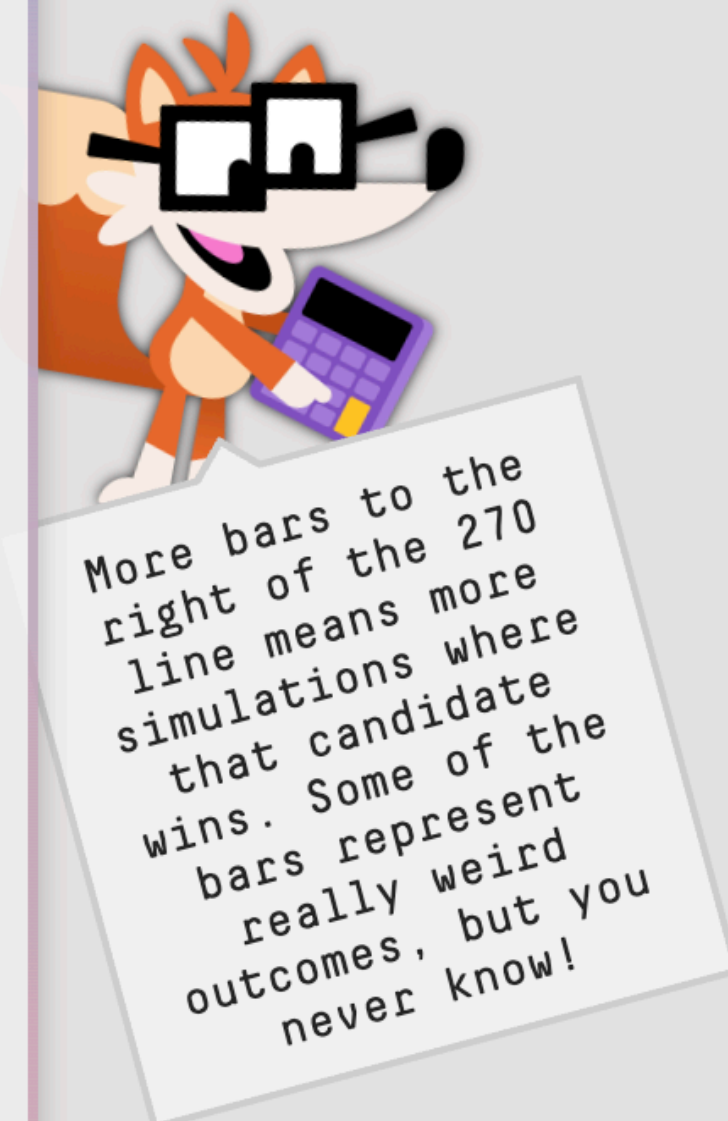
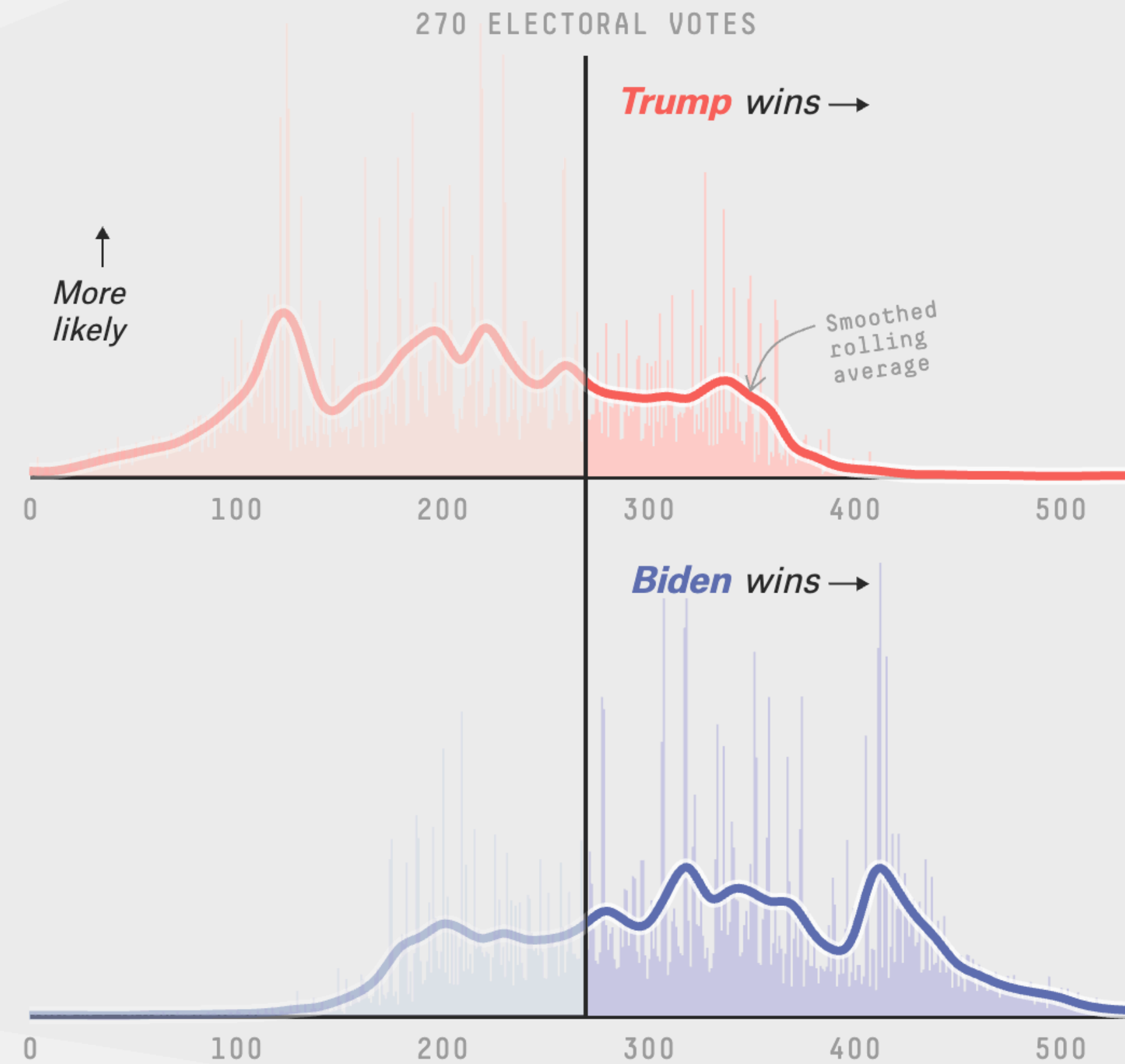
<https://policyviz.com/podcast/episode-155-john-burn-murdoch/>

“Design secrets behind the FT’s best charts of the year”

<https://www.ft.com/content/4743ce96-e4bf-11e7-97e2-916d4fbac0da>

Every outcome in our simulations

All possible Electoral College outcomes for each candidate, with higher bars showing outcomes that appeared more often in our 40,000 simulations



<https://projects.fivethirtyeight.com/2020-election-forecast/>

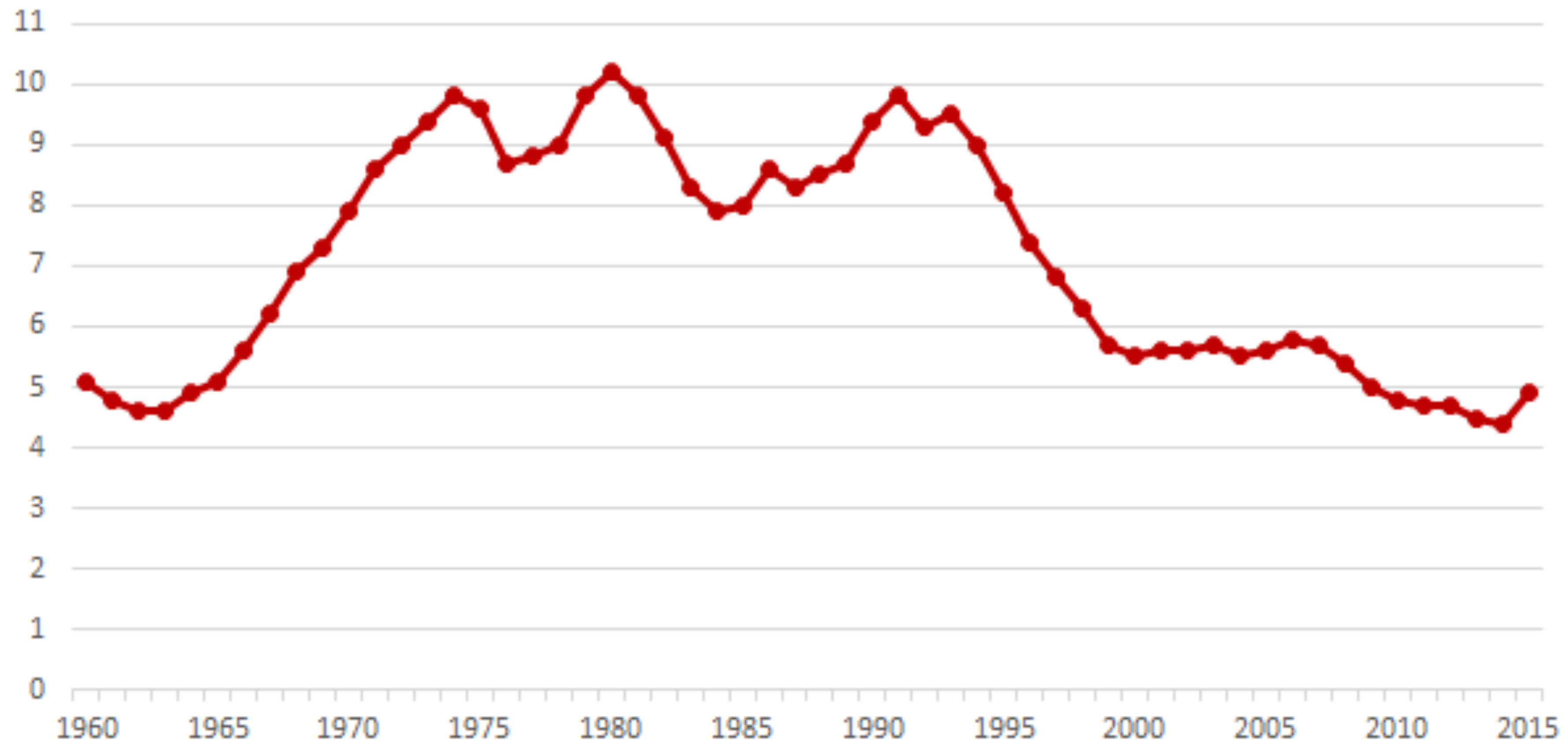


4. How much to visualize?

Am I showing too little?

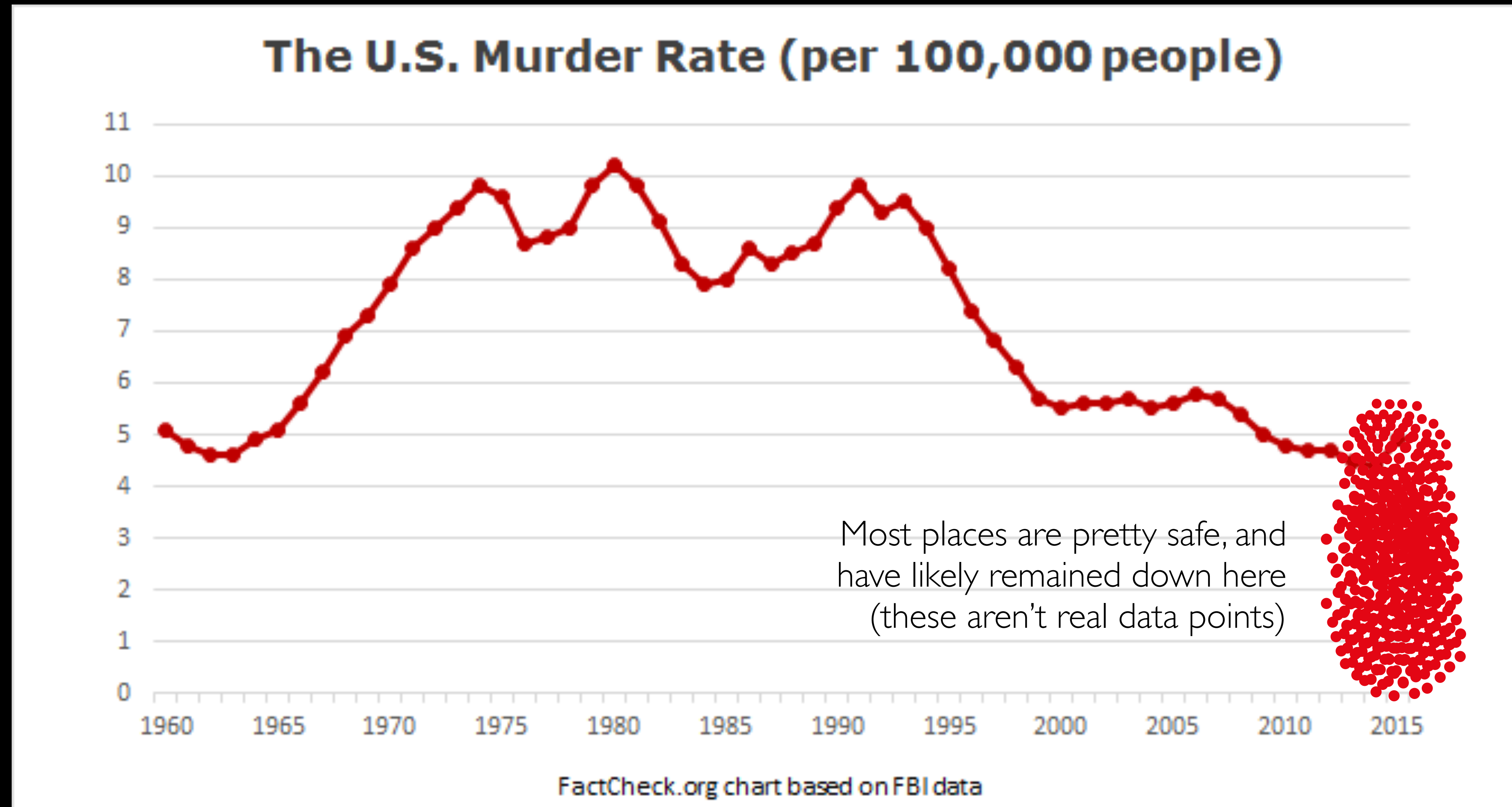
Am I showing too much?

The U.S. Murder Rate (per 100,000 people)

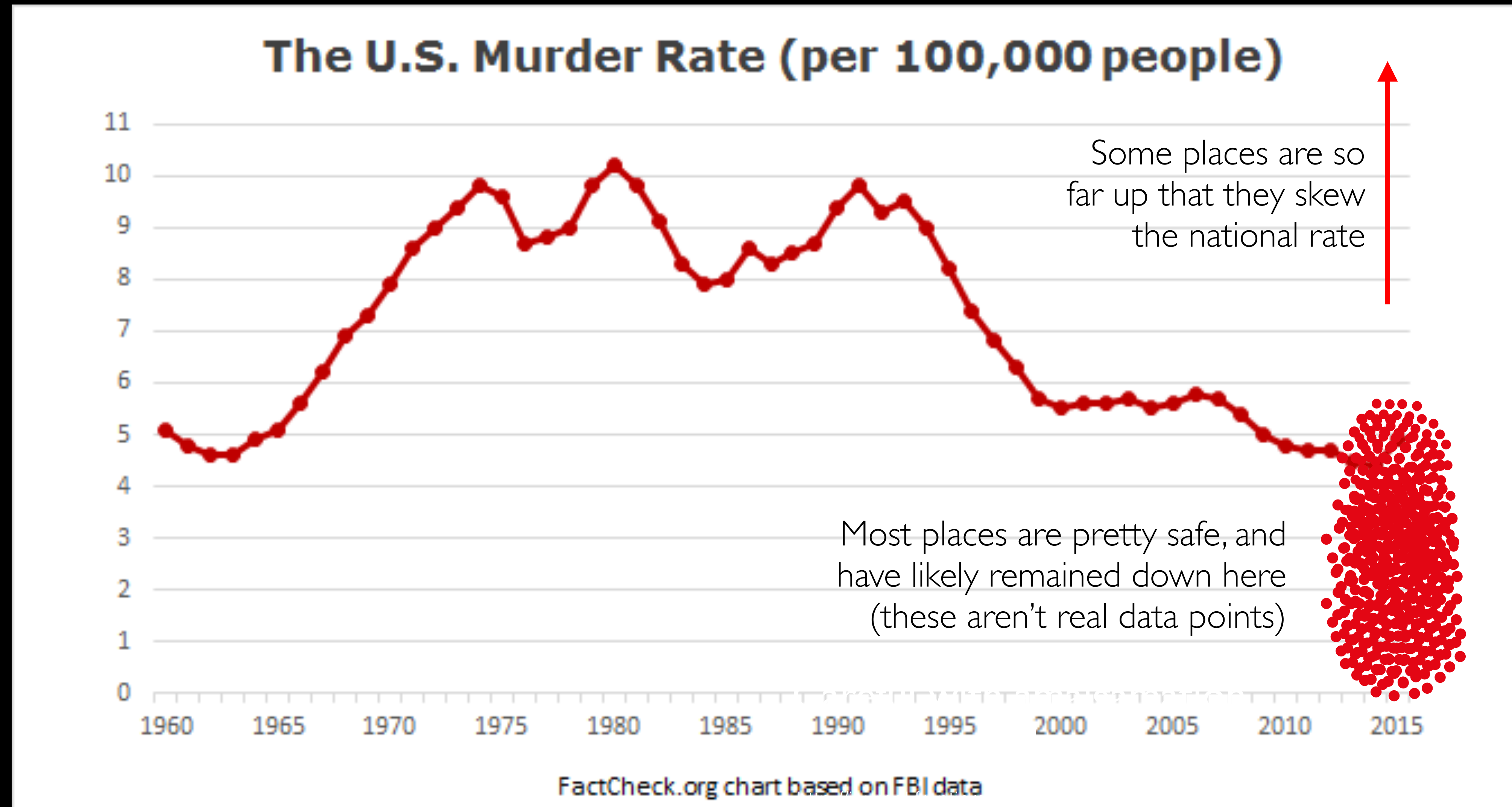


FactCheck.org chart based on FBI data

The danger of aggregating data too much,
and presenting just averages and other statistical summaries



The danger of aggregating data too much,
and presenting just averages and other statistical summaries





5. How to visualize it?

What types of charts or maps should I use?

What is the best way to organize the visualization?

Figure 2 - Main nationalities of arriving migrants – 2016

Greece

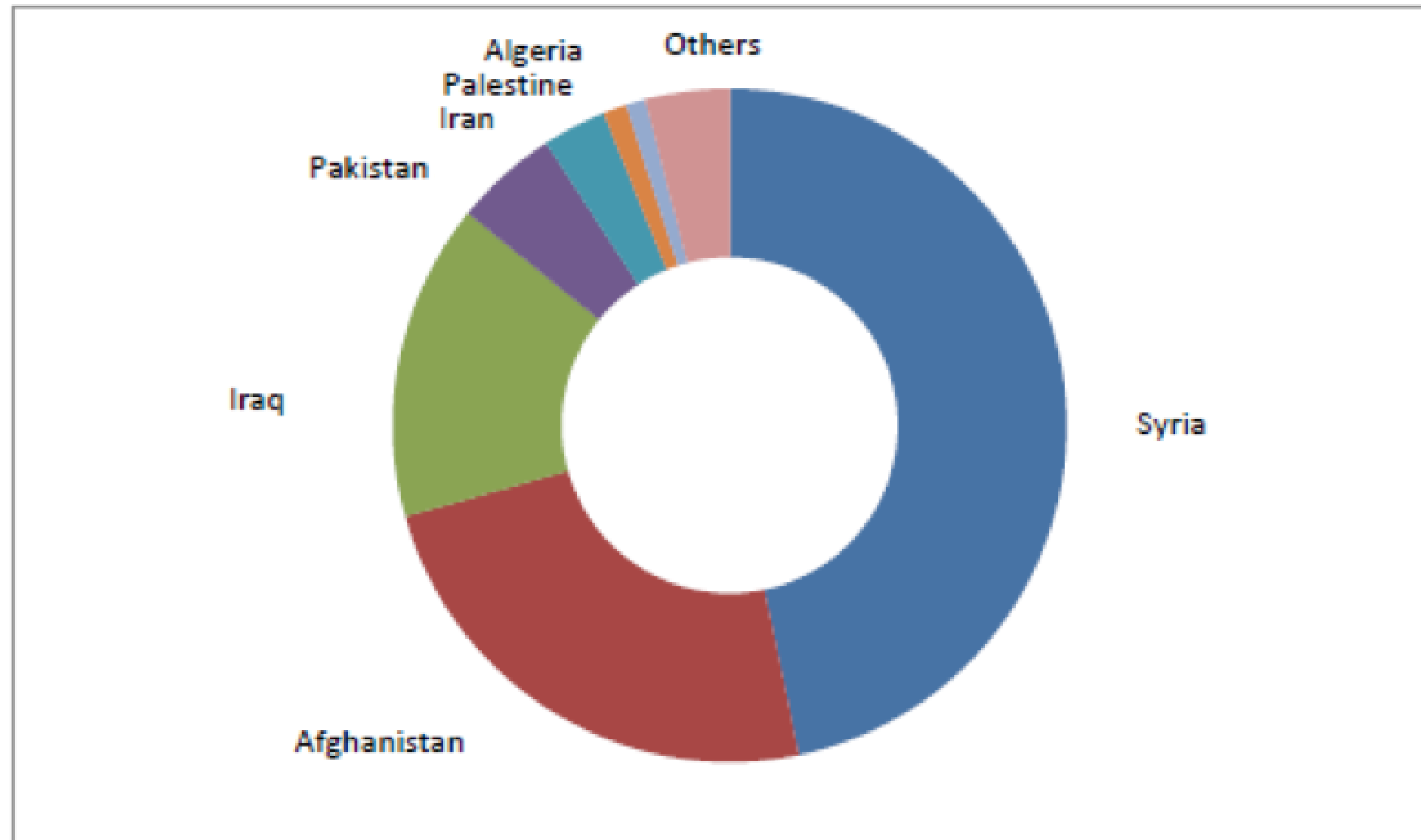
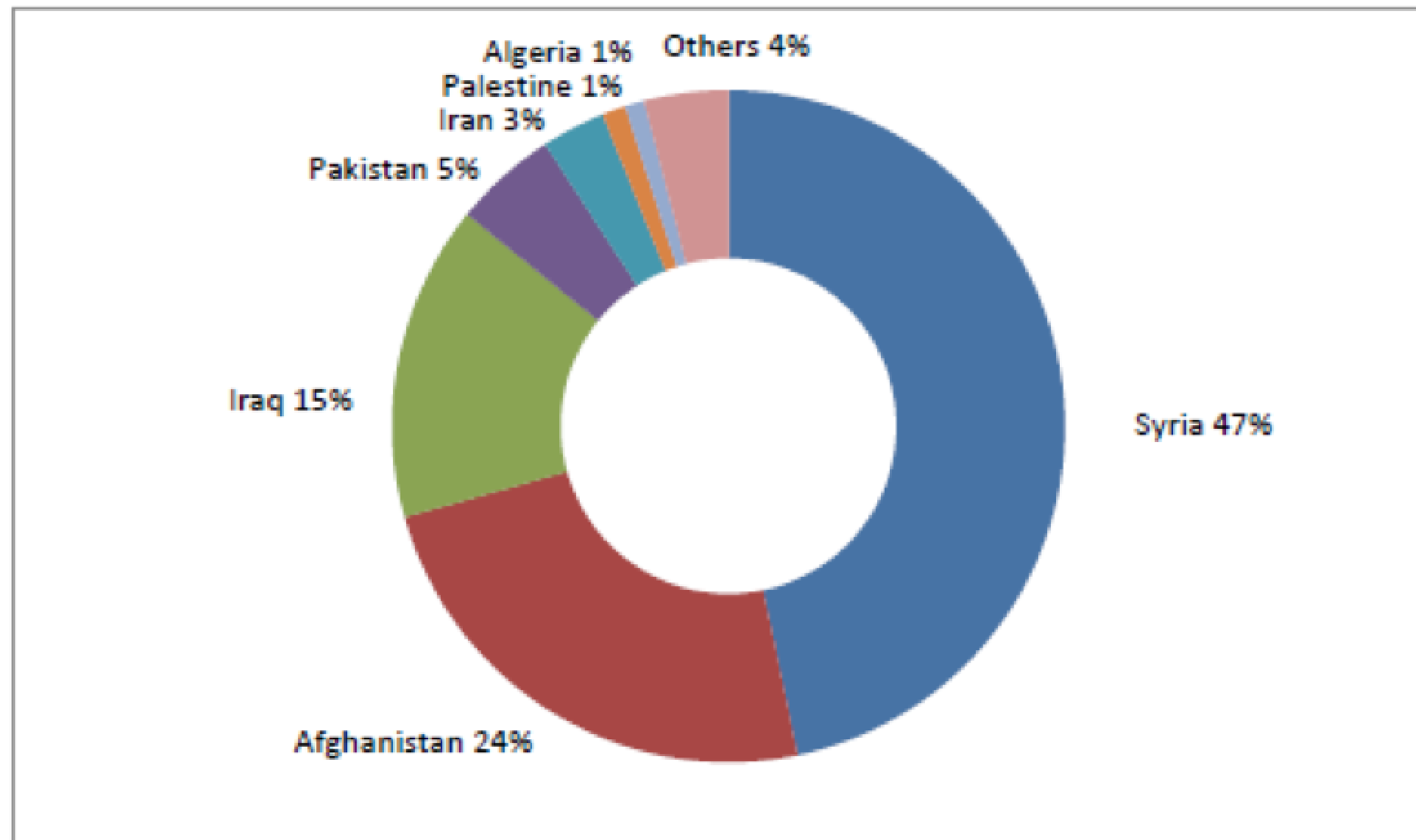
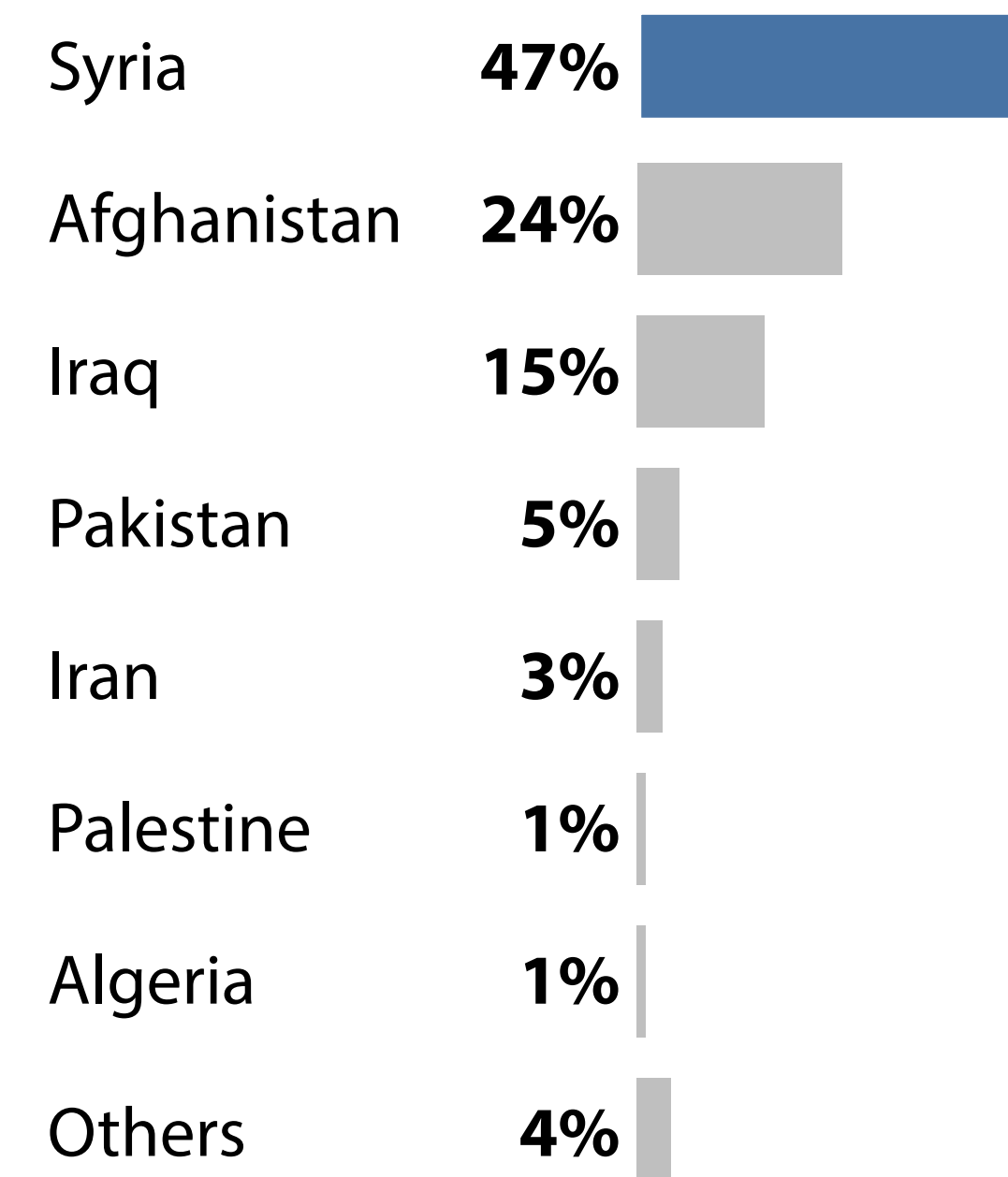
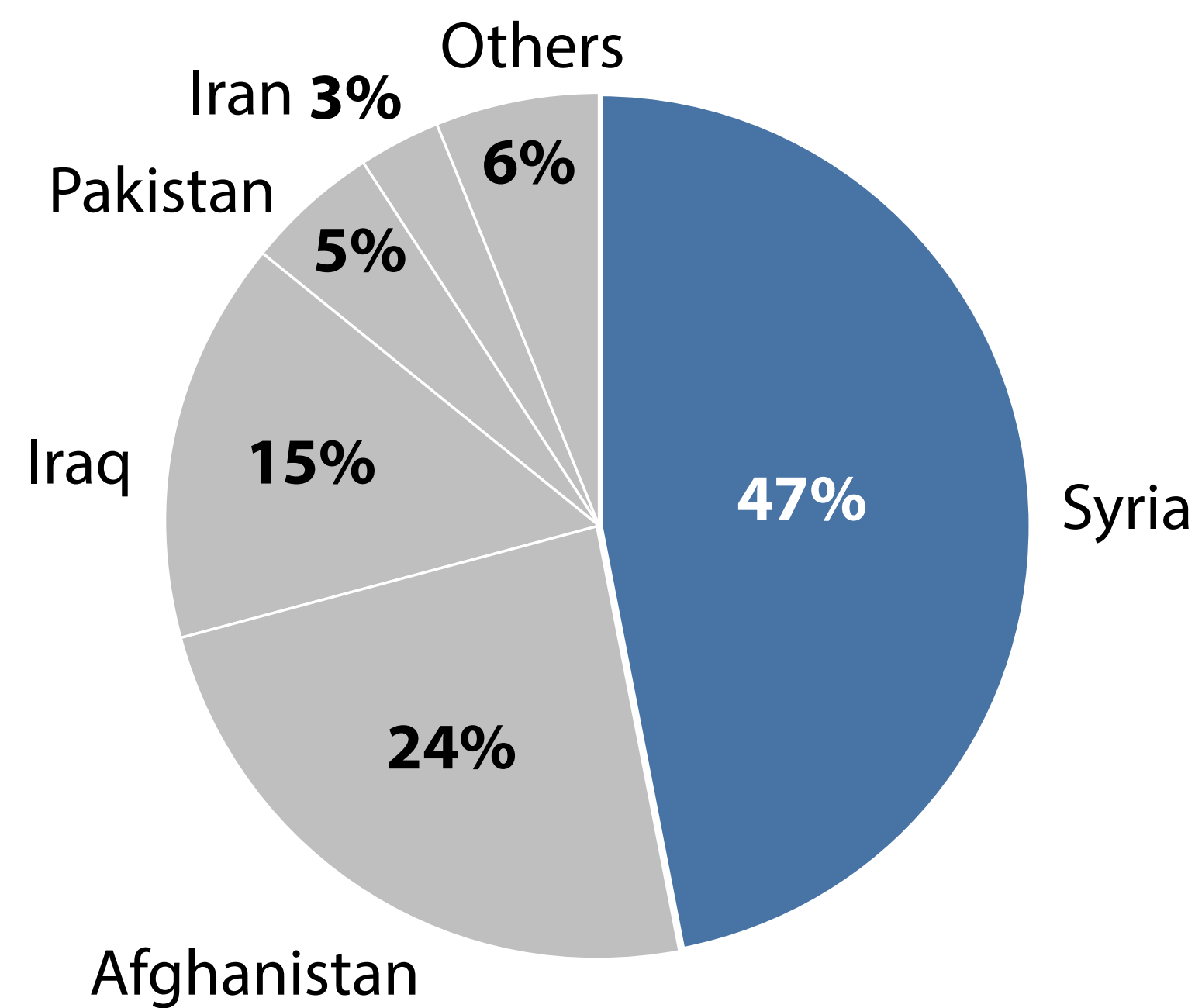
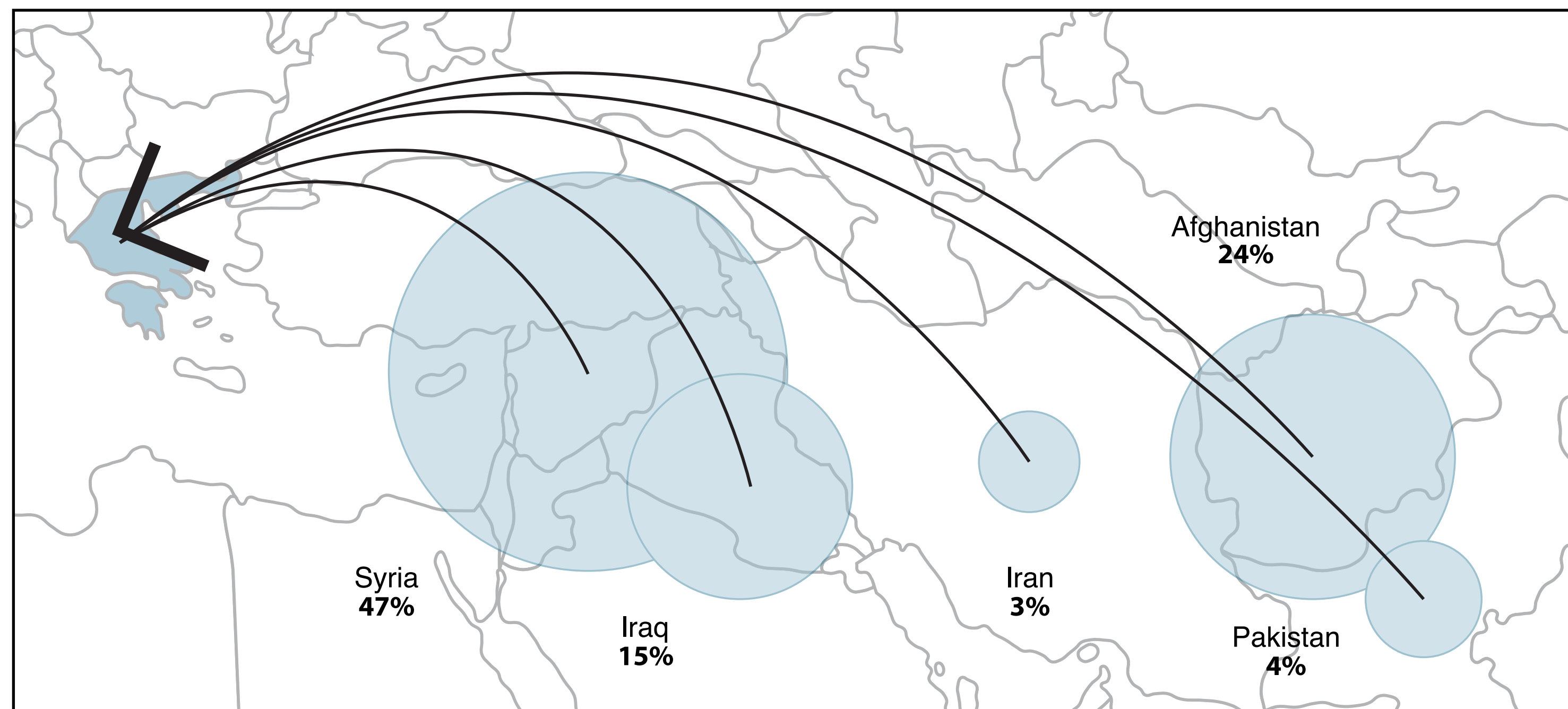


Figure 2 - Main nationalities of arriving migrants – 2016

Greece





The Data Visualisation Catalogue

About • Suggest • Shop • Resources

Search by Function

View by List



Deviation

Emphasize variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/negative/neutral).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar
A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar
Perfect for presenting survey results which involve sentiment (eg. disagree/neutral/agree).

Spine chart
Splits a single value into 2 contrasting components (eg. Male/Female).

Surplus/deficit filled line
The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that unless you tell them otherwise, many readers will assume the relationship you show them to be causal (ie. one causes the other).

Example FT uses
Inflation & unemployment, income & life expectancy

Scatterplot
The standard way to show the relationship between two continuous variables, each of which has its own axis.

Line + Column
A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot
Usually used to show how the relationship between 2 variables has changed over time.

Bubble
Like a scatterplot, but adds additional detail by storing the circles according to a third variable.

XY heatmap
A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar
Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column
See above.

Ordered proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot
Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope
Perfect for showing how ranks have changed over time or vary between categories.

Lollipop chart
Lollipops draw more attention to the data value than standard bar/columns and can also show rank and value effectively.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equity in the data.

Example FT uses
Income distribution, population, (age)sex distribution

Histogram
The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot
Summarise multiple distributions by showing the median, quartiles and range of the data.

Violin plot
Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple averages).

Population pyramid
A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Dot strip plot
Good for showing individual values in a distribution, can be a problem when there are many dots have the same value.

Dot plot
A simple way of showing the change or range (min/max) of data across multiple categories.

Barcode plot
Like dot strip plots, good for displaying all the data in a table; they work best when highlighting individual values.

Cumulative curve
A good way of showing how unequal a distribution is; y axis is always cumulative frequency, x axis is always a measure.

Change over Time

Give emphasis to changing trends. These can be short (or 'spike') movements or extended series (trending upwards or downwards). Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series

Line
The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column
Columns work well for showing the size and proportion of data at the same time – as long as the data are not too complicated.

Line + column
A good way of showing the relationship over time between an amount (columns) and a rate (line).

Stock price
Usually focused on day-to-day activity, these charts show the opening/closing and high/low points of each day.

Slope
Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart
Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Fan chart (projections)
Use to show the uncertainty in future projections – usually this grows the further forward to projection.

Connected scatterplot
A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap
A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline
Great when date and duration are key elements of the story in the data.

Circle timeline
Good for showing discrete values of varying size across multiple categories (eg. earthquakes by continents).

Seismogram
Another alternative to the circle timeline for showing series where there are big variations in the data.

Part-to-whole

Show how a single entity can be broken down into its constituent elements, if the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column
A simple way of showing the part-to-whole relationships but can be difficult to read with more than a few components.

Proportional stacked bar
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie
A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut
Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg. total).

Treemap
Use for hierarchical data; the size and proportion of data at the same time – as long as the data are not too complicated.

Voronoi
A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Sunburst
Another way of visualising hierarchical part-to-whole relationships – usually only with whole numbers (do not slice off an arm to represent a decimal).

Arc
A hemicycle, often used for visualising political results in parliaments.

Gridplot
Good for showing % information, they work best when used on whole numbers and work well in multiple layout form.

Venn
Generally only used for schematic representation.

Waterfall
Can be useful for showing part-to-whole relationships where some of the components are negative.

Magnitude

Show size comparisons. These can be relative (size being able to be larger/smaller) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation

Column
The standard way to compare the size of things. Must always start at 0 on the axis.

Bar
See above. Good when the data are not time series and labels have long category names.

Paired column
As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar
See above.

Proportional stacked bar
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)
Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop chart
Lollipop charts draw more attention to the data value than standard bar/column – does not HAVE to start at zero (but preferable).

Radar chart
A space-efficient way of showing value of multiple variables – but make sure they are organised in a way that makes sense to reader.

Parallel coordinates
An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Spatial

Used only when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Locator maps, population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (categorical)
The standard approach for putting data on a map – should always be rates rather than values and use a sensible base geography.

Proportional symbol (count/magnitude)
Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map
For showing unambiguous movement across a map.

Contour map
For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values.

Equalized cartogram
Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal values.

Scaled cartogram (value)
Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density
Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map
Grid-based data values mapped with an intensity colour scale. As choropleth map – but not trapped to an administrative unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information relationship graphs.

Sankey
Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall
Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord
A complex but powerful diagram which can illustrate 2-way flows (and net flows) in a matrix.

Network
Used for showing the strength and inter-connectedness of relationships of varying types.

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Jan Both; Li Fei; Graham Parish; Billy Ehrenberg; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribeca

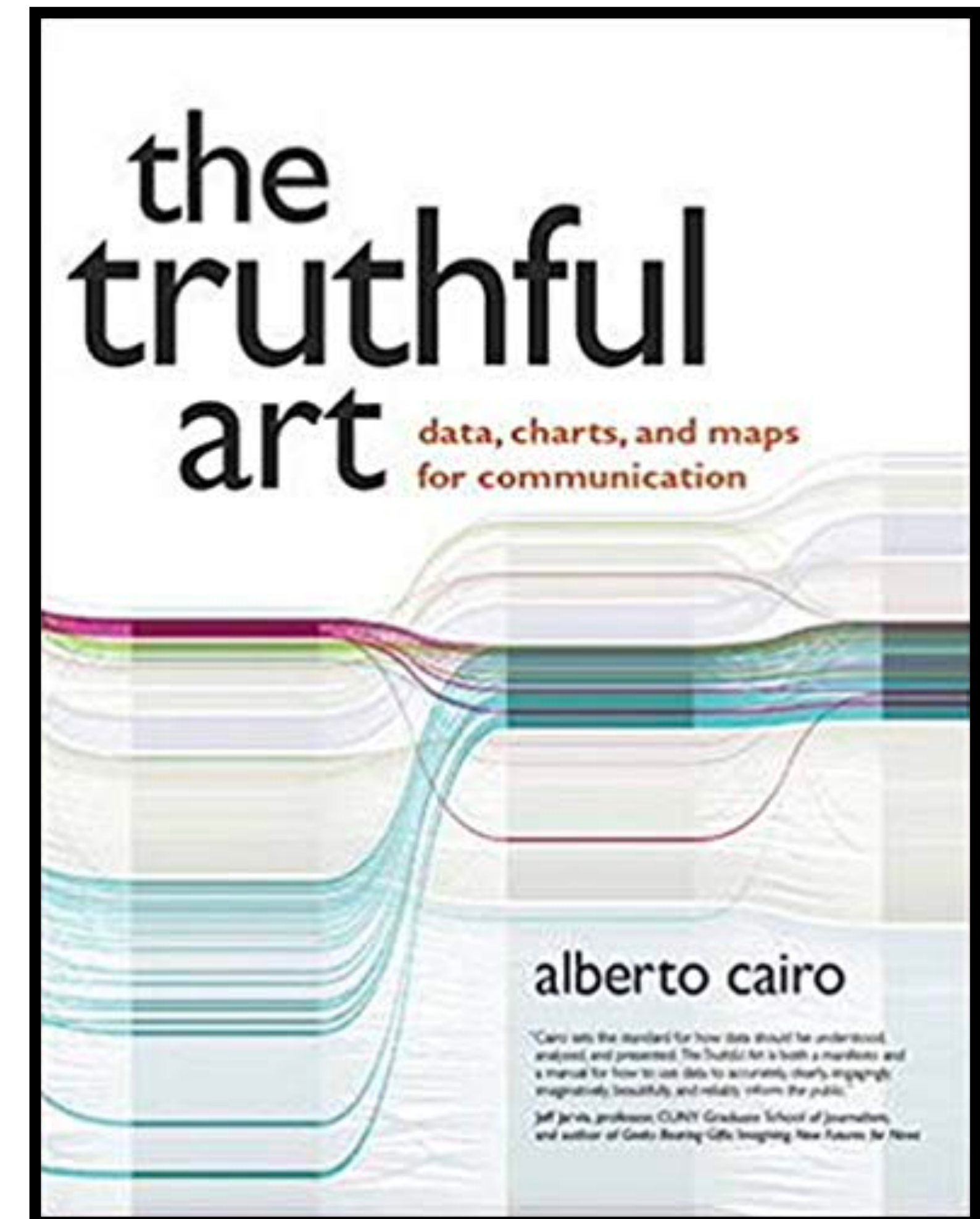
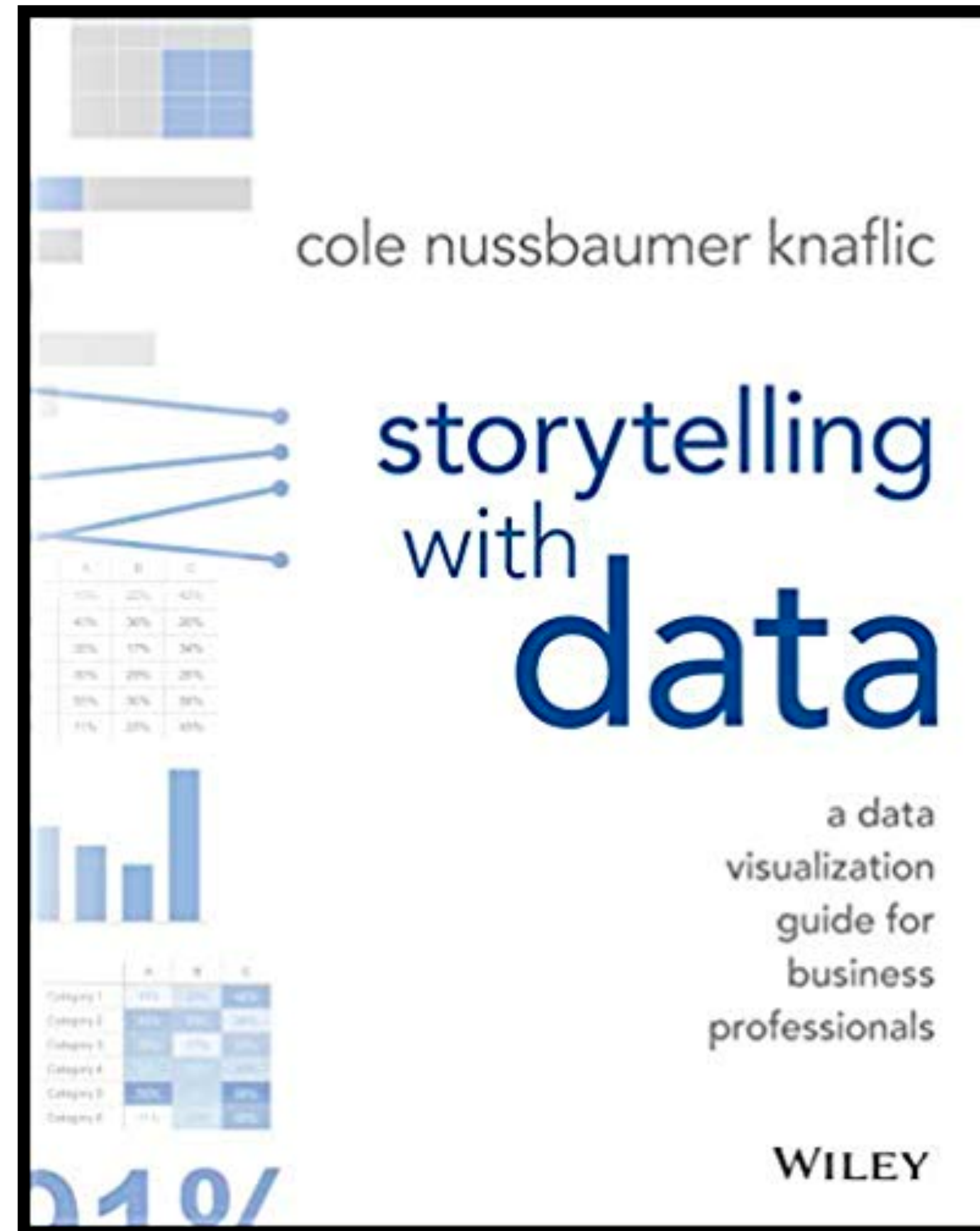
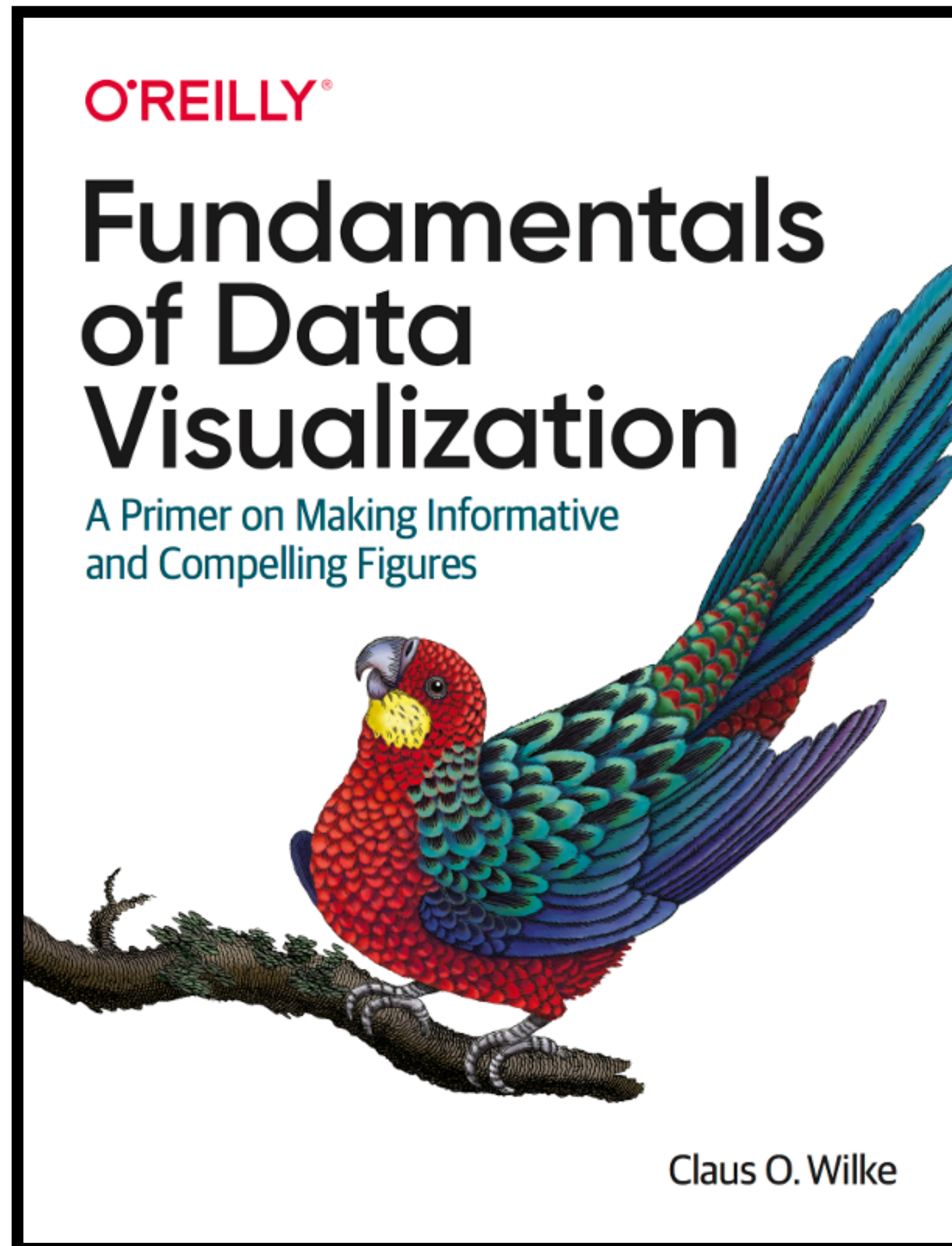
ft.com/vocabulary



<http://www.datavizcatalogue.com/>

<https://github.com/ft-interactive/chart-doctor/blob/master/visual-vocabulary/Visual-vocabulary.pdf>

Books to make design choices



Draft available online:
<https://serialmentor.com/dataviz/>



6. What visual style to use?

Not all visualizations need to be minimalist.

Not all visualizations need to be flashy and innovative, either.

Standard visualizations

Appropriate for graphics we use all the time

Total Deficits and Surpluses

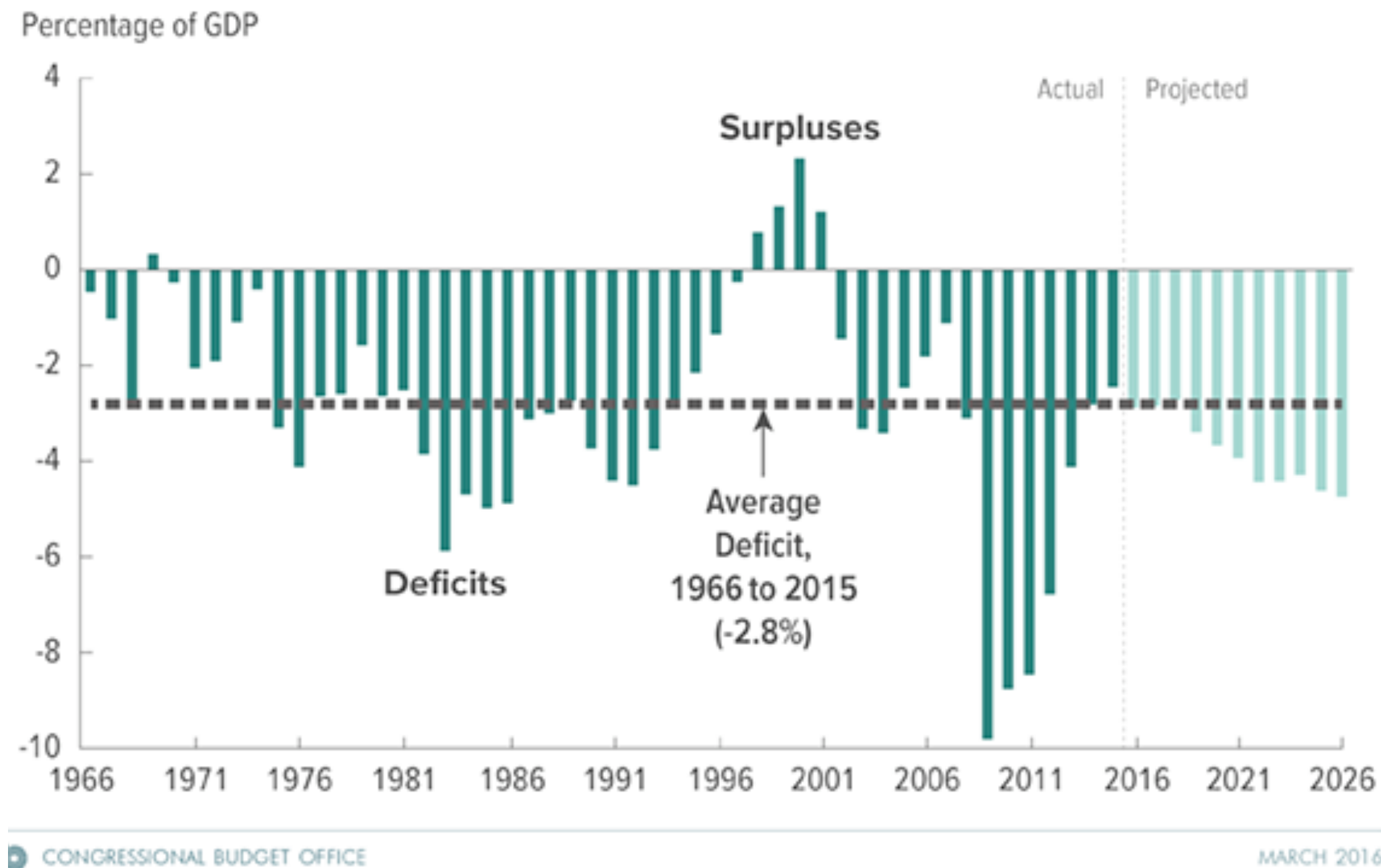
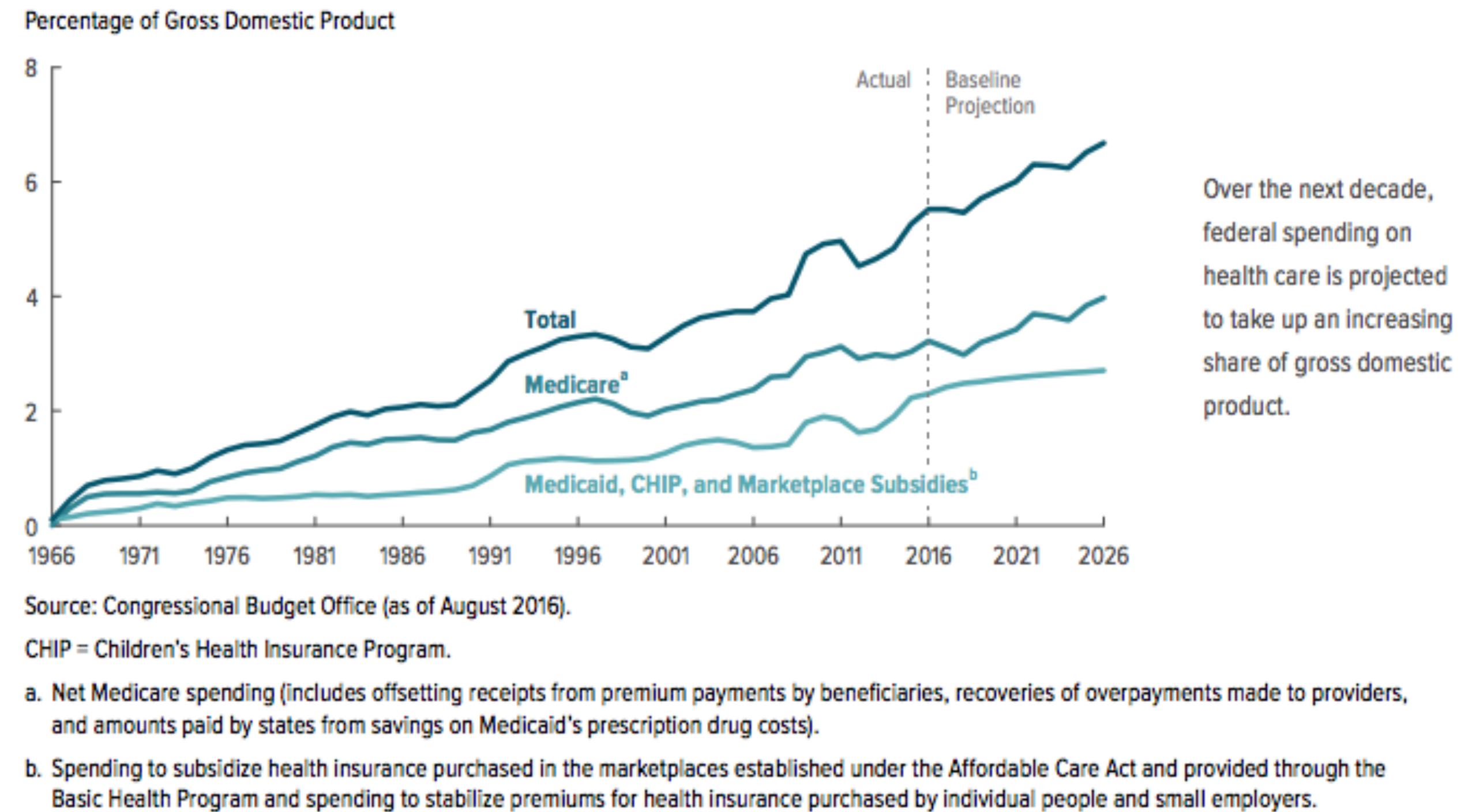


Figure 5-1.

Federal Spending on the Major Health Care Programs, by Category

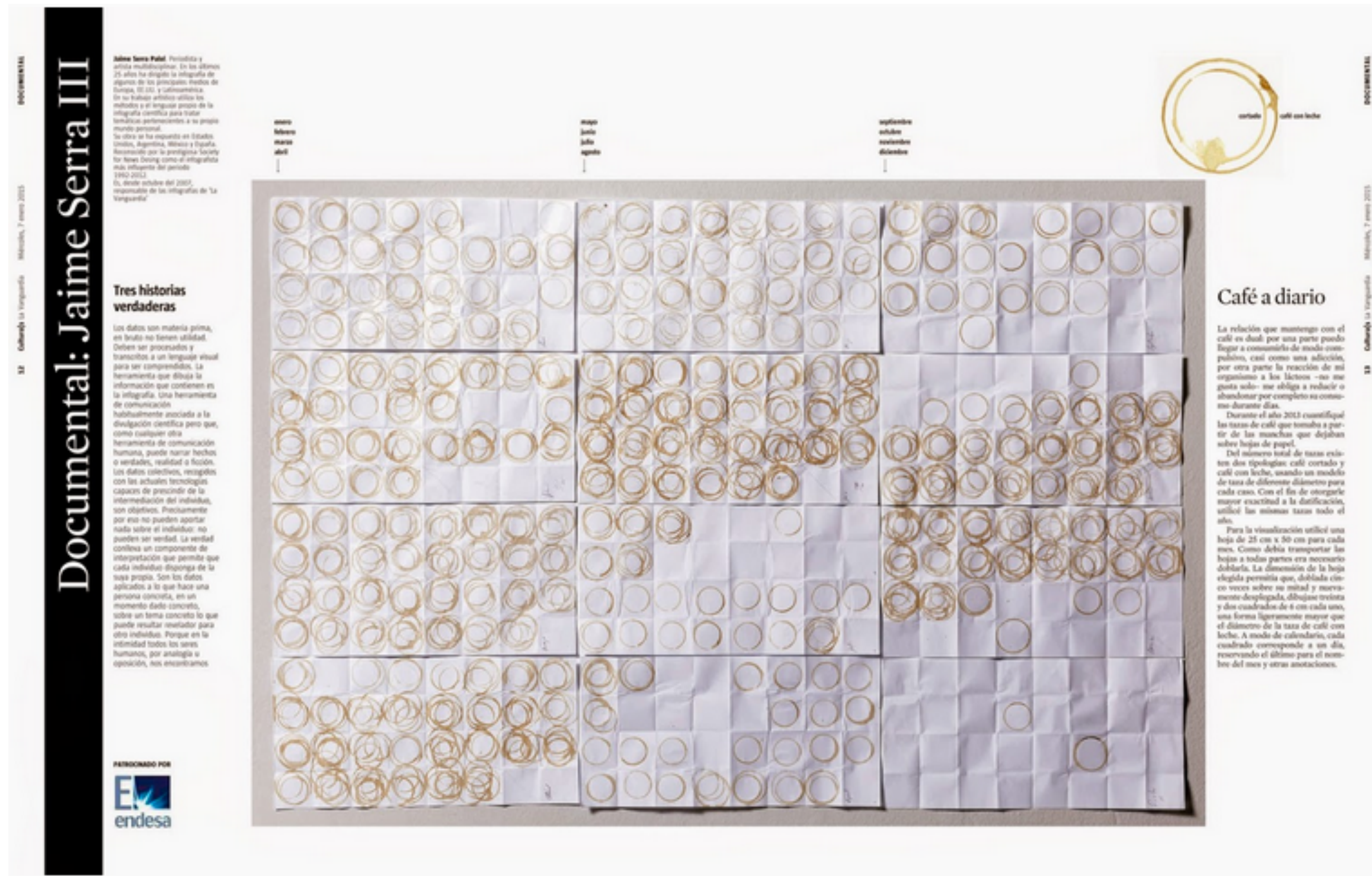




<https://jaimeserra-archivos.blogspot.com/>

Fully customized visualizations:

Appropriate for one-time use when we want to provoke curiosity, surprise
—or simply a smile



<http://visualoop.com/28792/portfolio-of-the-week-jaime-serra>

Making data feel “warmer”

Q

Sections

The Washington Post
Democracy Dies in Darkness

Alberto Cairo Touri...

AT THE EPICENTER

What if all covid-19 deaths in the United States had happened in your neighborhood?

Find out what would happen if your neighborhood was the epicenter of the coronavirus pandemic in the United States.

Updated Sept. 24 at 11:43 a.m.
Data updated on Sept. 29, 2020

In partnership with

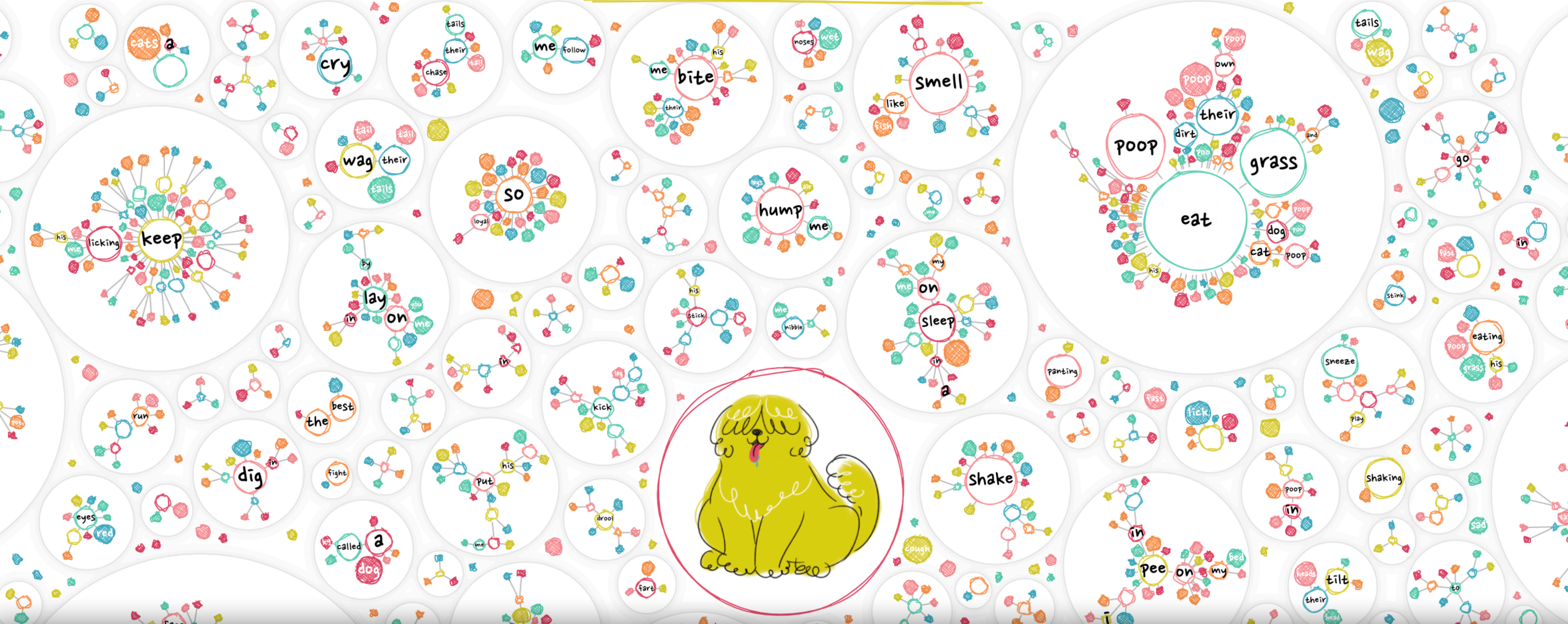
Lupa | Google News Initiative

Enter your address in the USA

USE MY LOCATION

<https://www.washingtonpost.com/graphics/2020/national/coronavirus-deaths-neighborhood/>

The purpose of visualization isn't visualization per se. The purpose of visualization is to help people **make sense of the world** through a combination of visuals and words.



The End.

www.thefunctionalart.com , www.albertocairo.com , alberto.cairo@gmail.com