Make good (visualization) choices

Laura Garrison, ICTP, 15. Nov 2023

Let’s warm up with some *bad* visualizations

VOTELINE

SATURDAY'S RESULT

Can Julia Gillard win the next federal election?

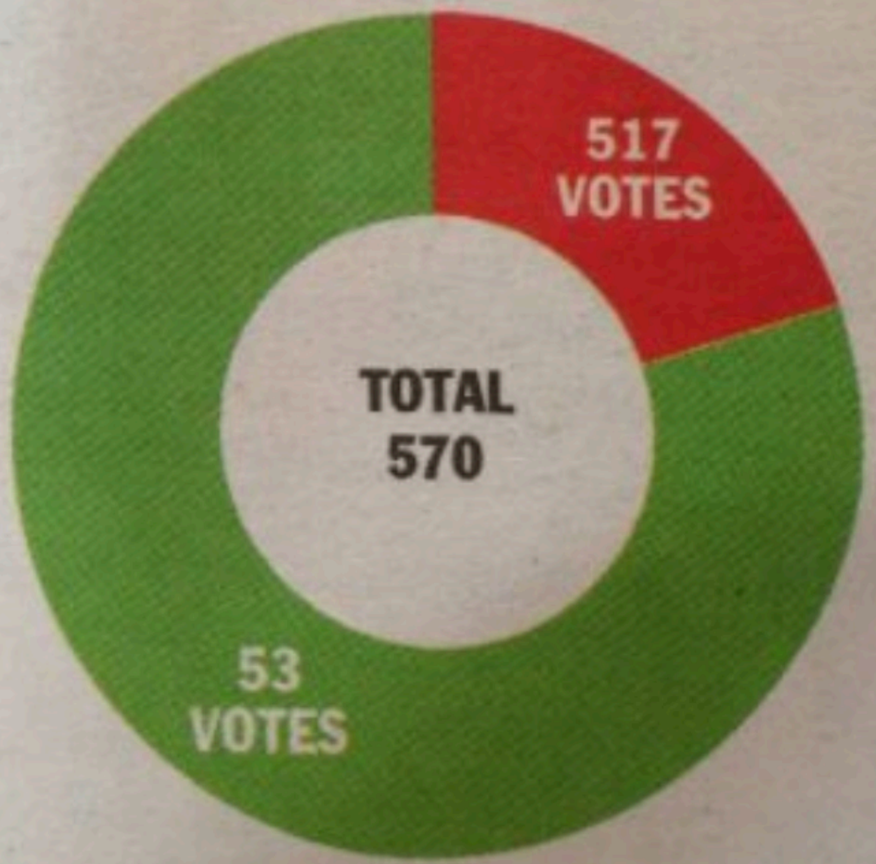
10% YES | 90% NO

TODAY'S QUESTION

Do you like the plan to remove a car lane of the Princes Bridge for sole use by cyclists?

YES	1900 956 434
NO	1900 956 435

Calls cost 38.5c including GST. You can also have your say at heraldsun.com.au or facebook.com/heraldsun



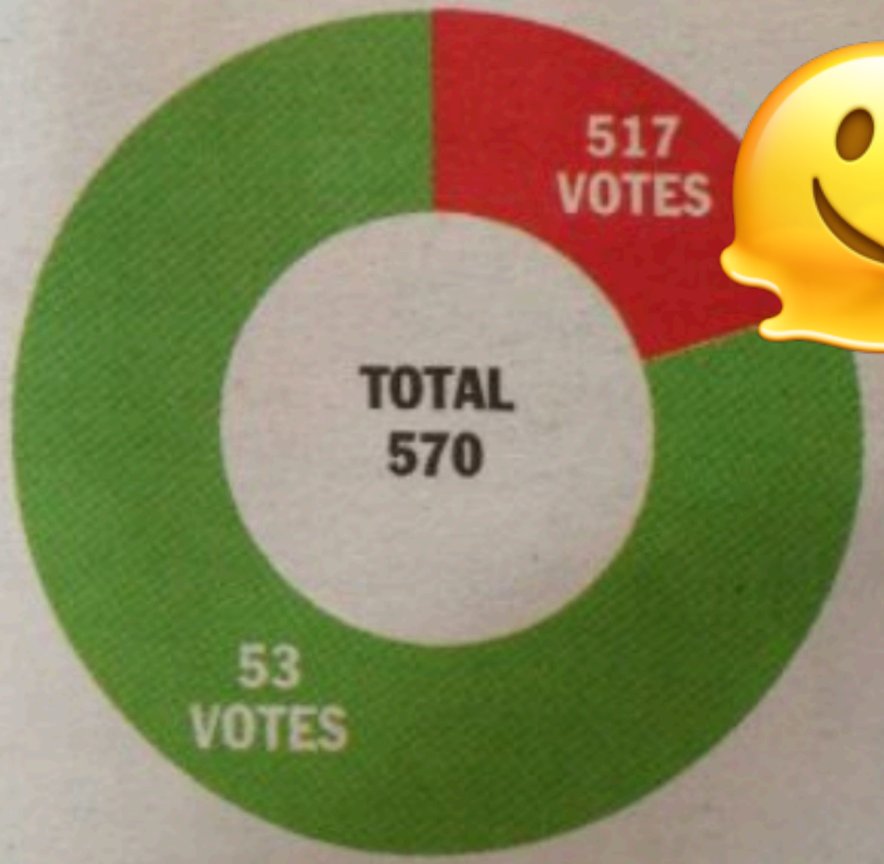
https://viz.wtf/

VOTELINE

SATURDAY'S RESULT

Can Julia Gillard win the next federal election?

10% YES | 90% NO



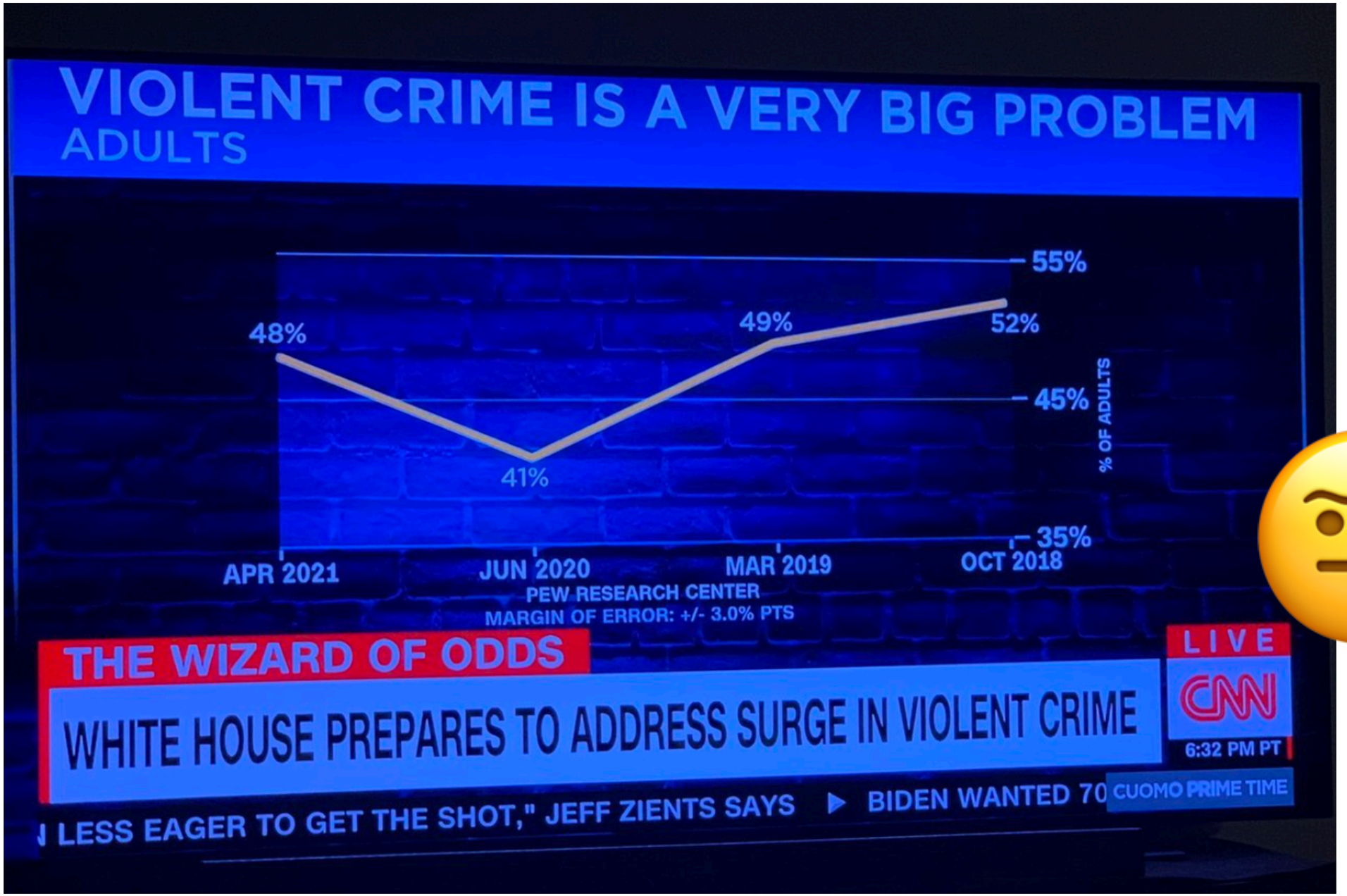
TODAY'S QUESTION

Do you like the plan to remove a car lane of the Princes Bridge for sole use by cyclists?

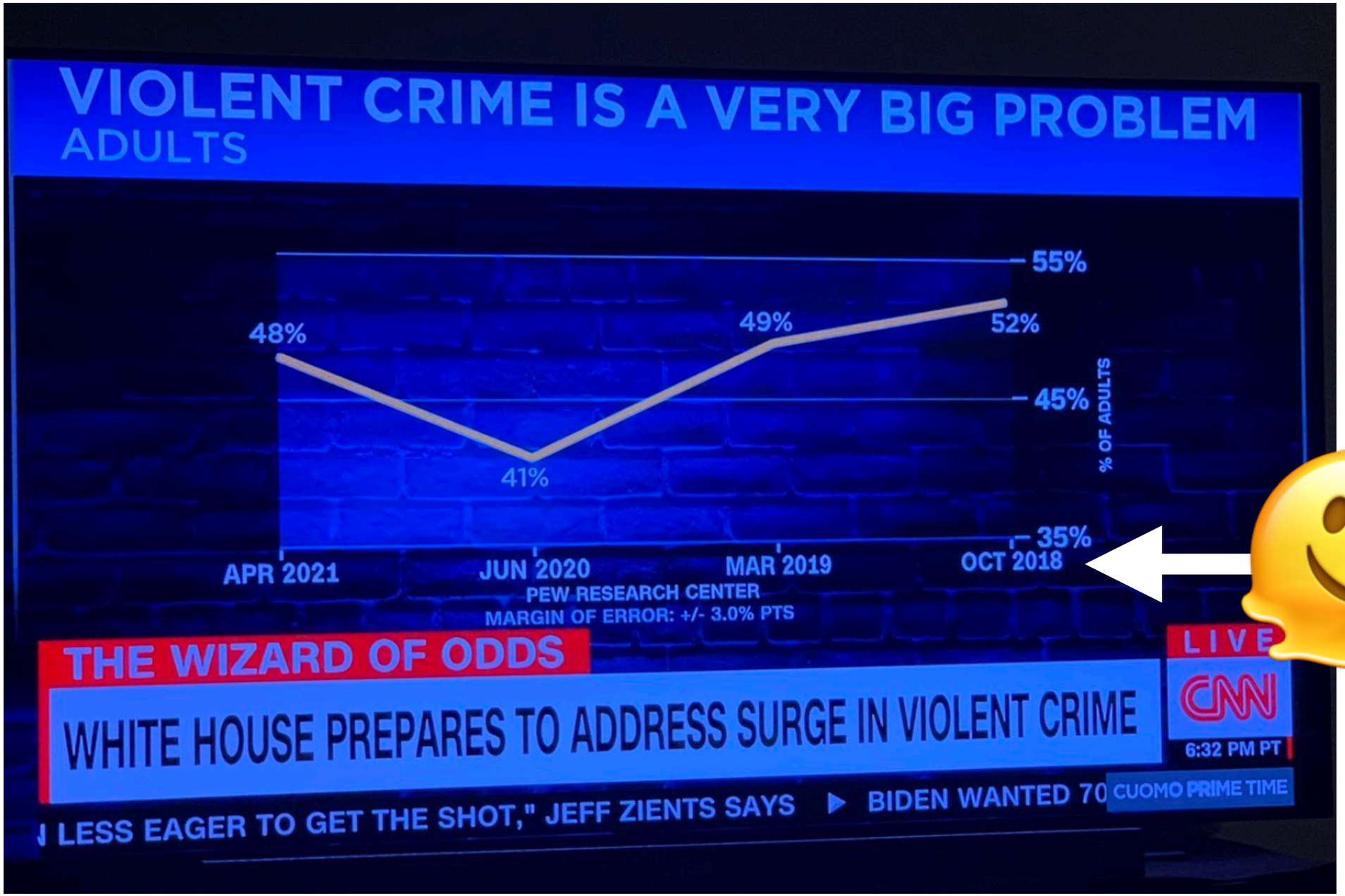
YES 1900 956 434
NO 1900 956 435

Calls cost 38.5c including GST. You can also have your say at heraldsun.com.au or facebook.com/heraldsun

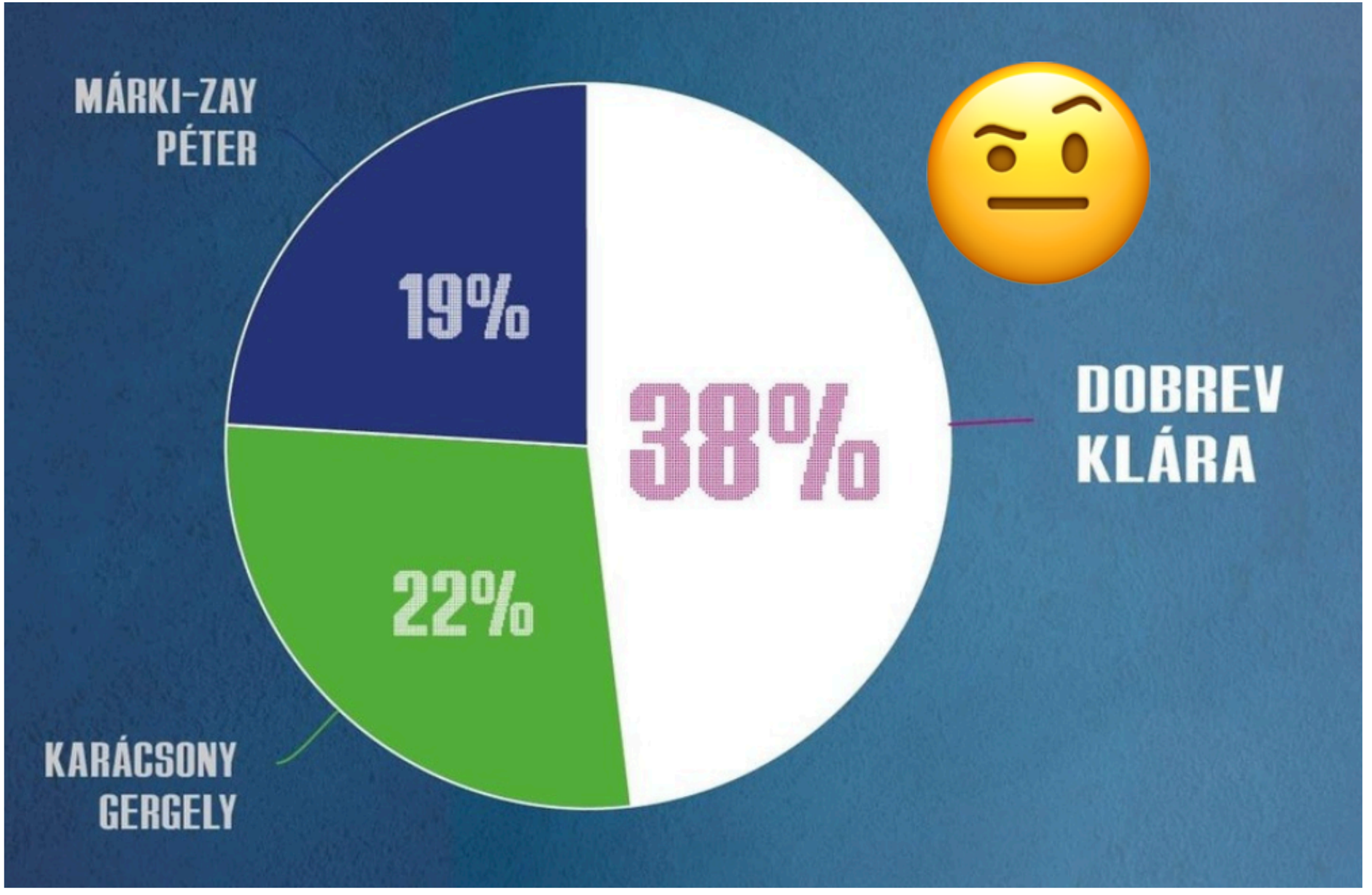
https://viz.wtf/

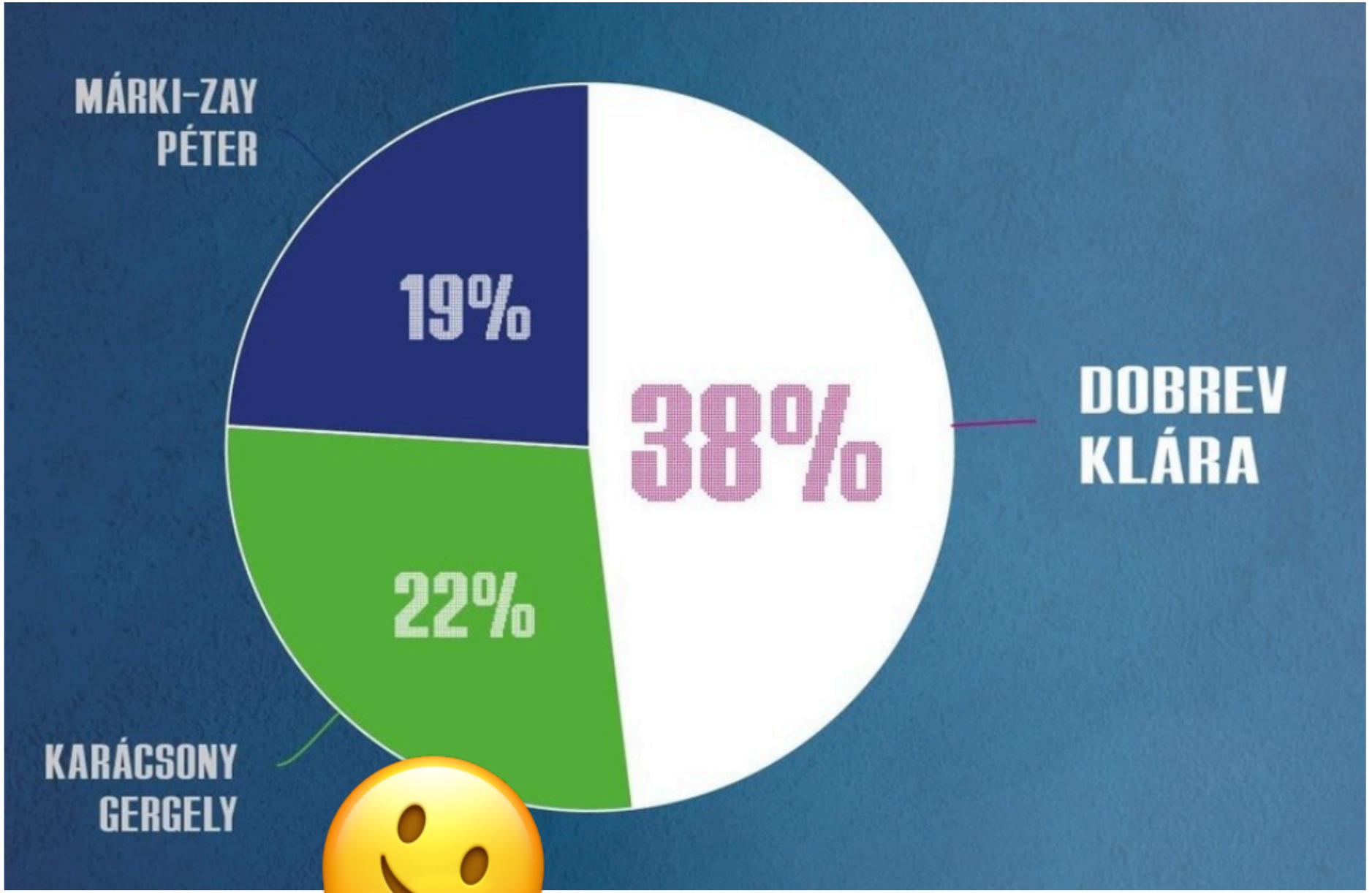


https://viz.wtf/

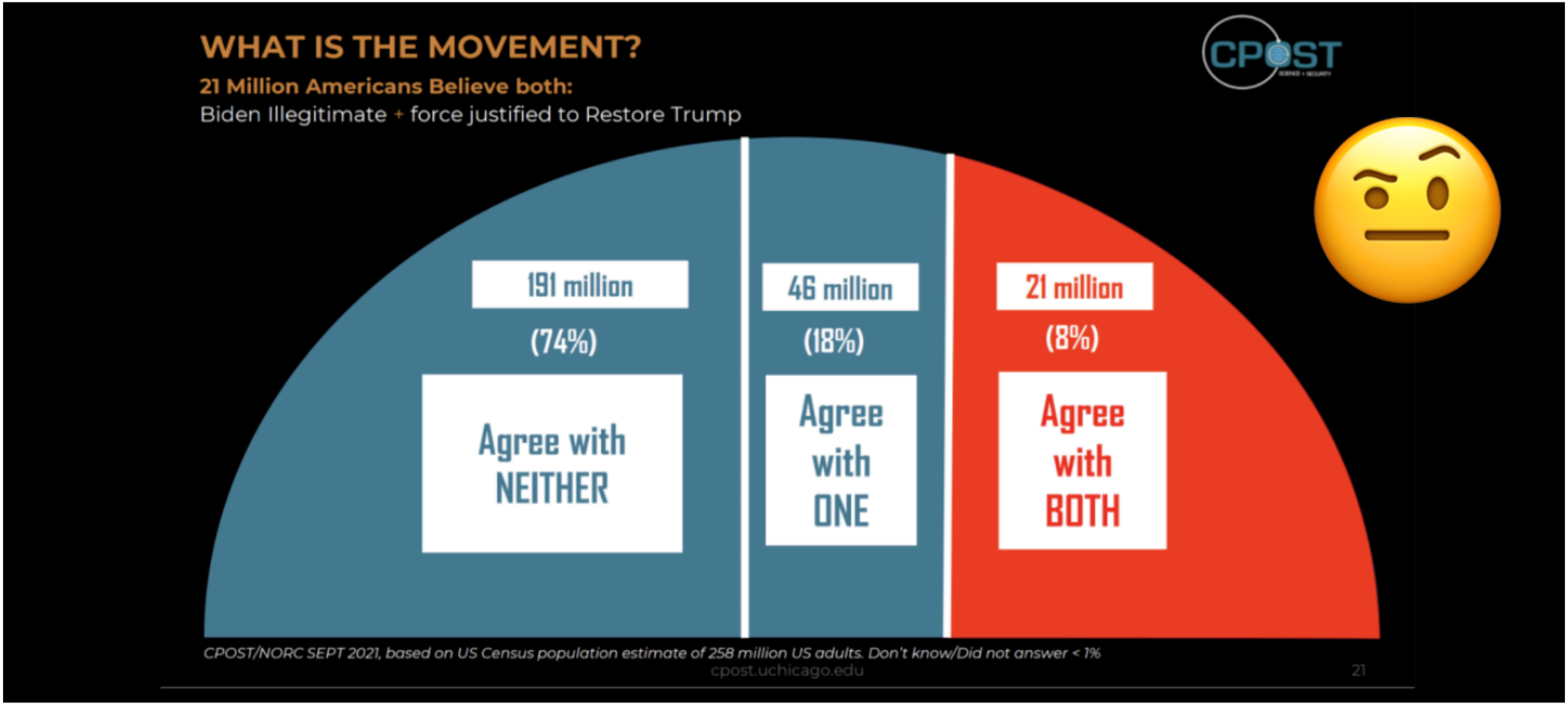


https://viz.wtf/

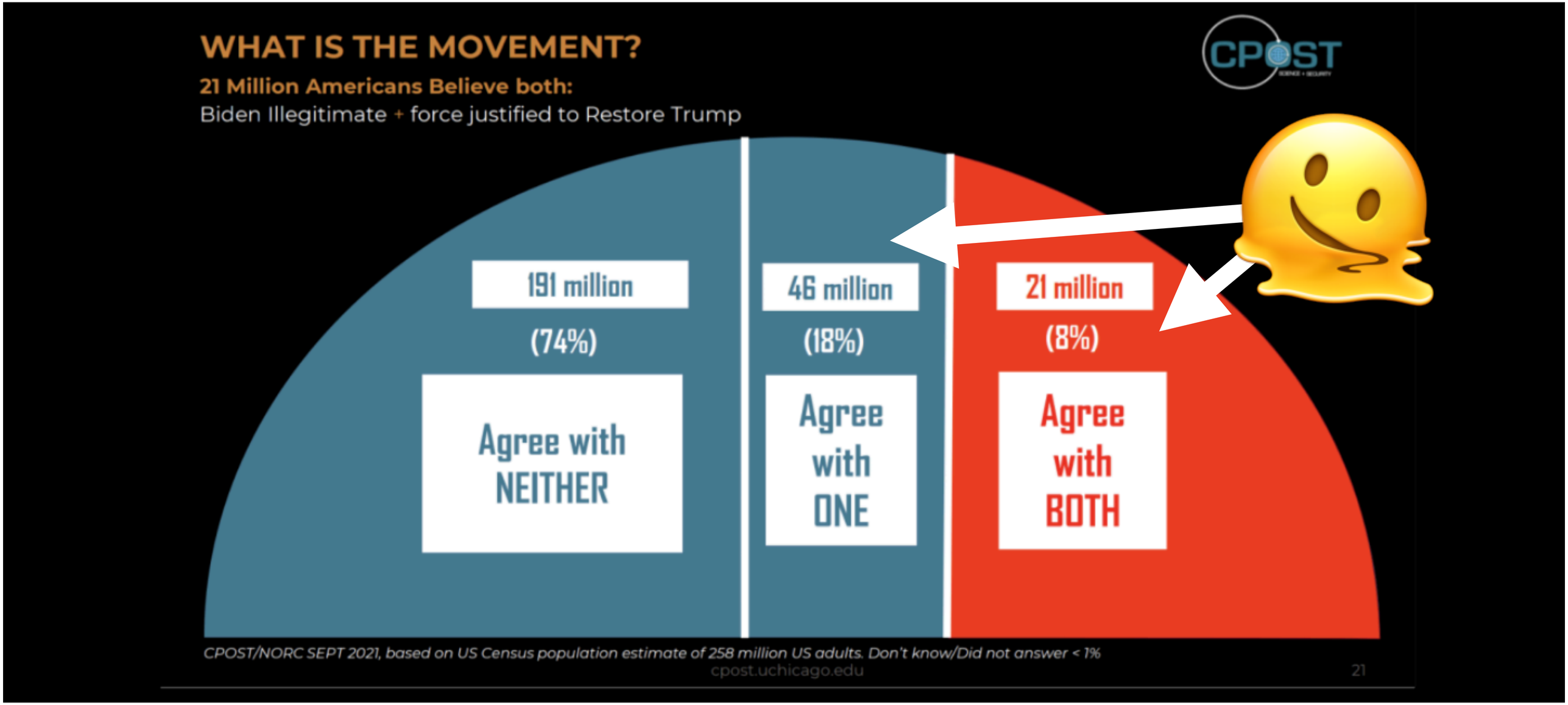

MÁRKI-ZAY PÉTER
19%
38%
DOBREV KLÁRA
22%
KARÁCSONY GERGELY



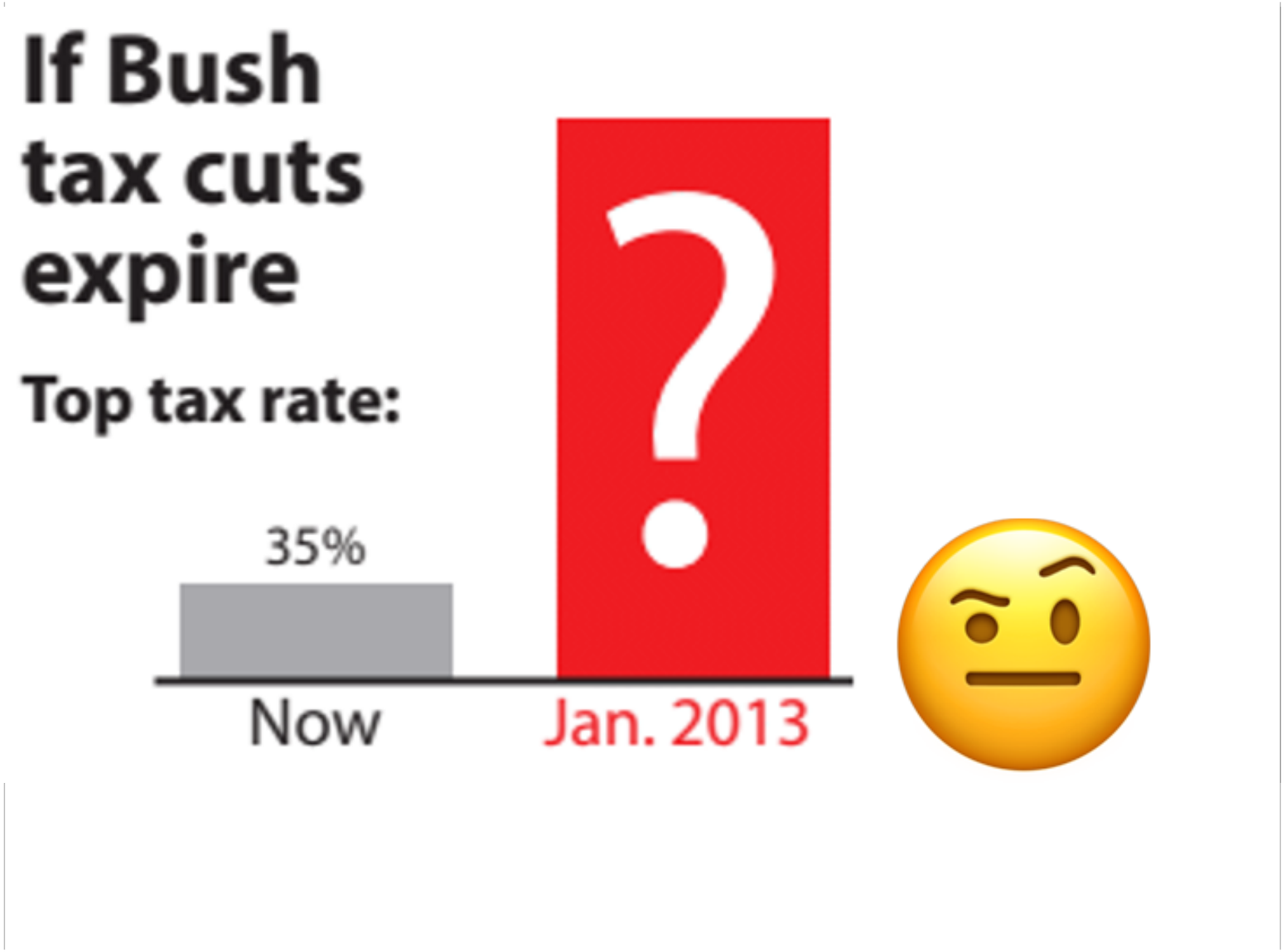
38 + 19 + 22 = ...



https://viz.wtf/

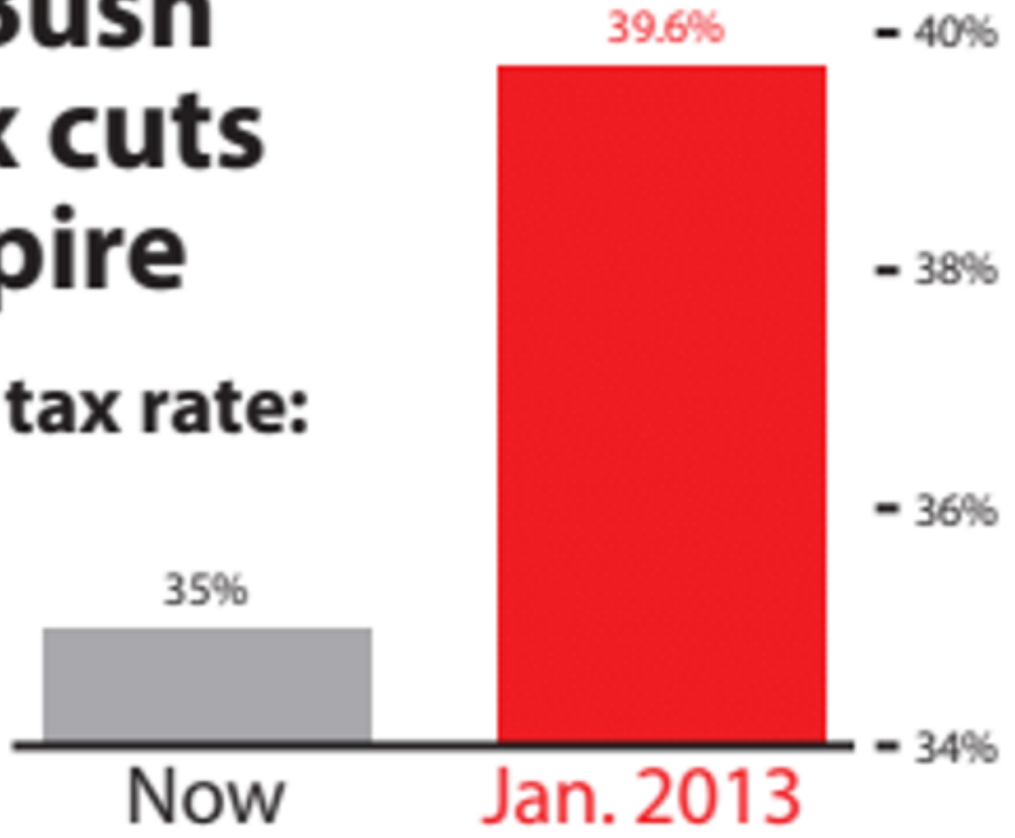


https://viz.wtf/ - Area encoding?


If Bush tax cuts expire
Top tax rate:
35%
?
Now
Jan. 2013

If Bush tax cuts expire

Top tax rate:



What can we do to make a *good* visualization?

(Some) tips for making a good visualization

1. Define your **goals**
2. **Show** the data (go beyond summary statistics)
3. Be **honest** with your visuals
4. Respect **common associations**
5. Design a **hierarchy of information**
6. Avoid taxing working **memory**
7. Tell a **story**
8. Reflect on **uncertainty and unknowns**

Define your goals.

In 1-3 sentences, can you say what you want to visualize and why?

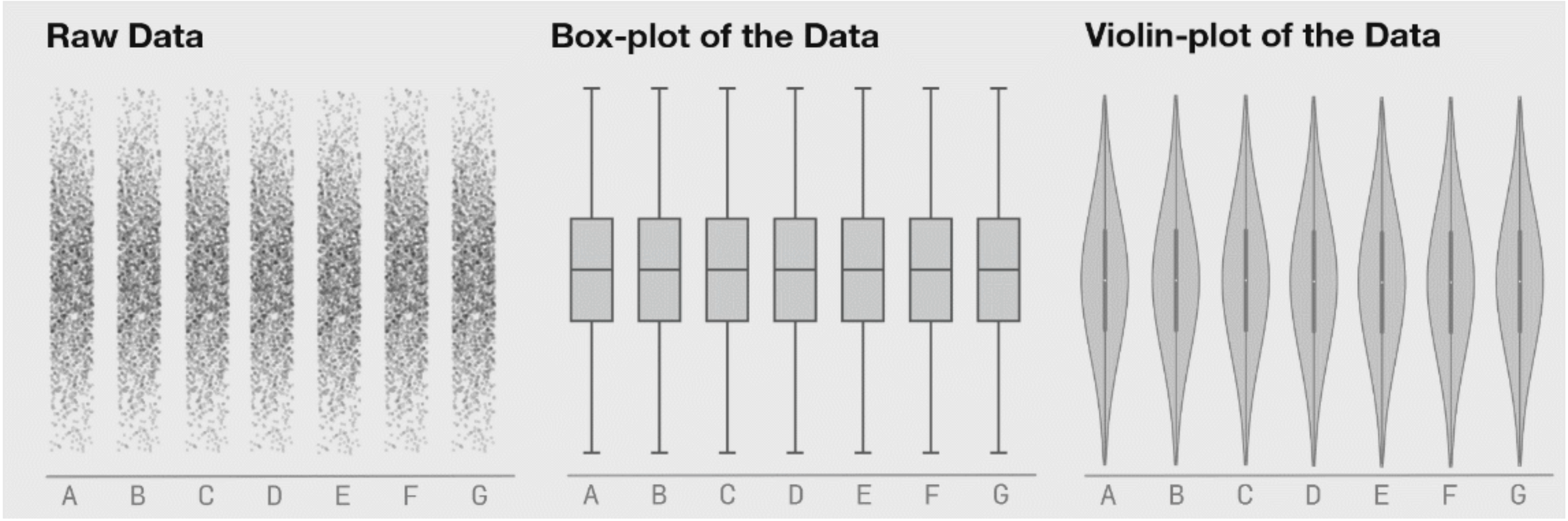
Define your goals.

I want to see what the titanic dataset looks like.

I want to compare the difference in survival between passengers in 1st, 2nd, and 3rd class to understand how social class played a role in whether passengers lived or died on the titanic.

Show the data

Go beyond summary statistics

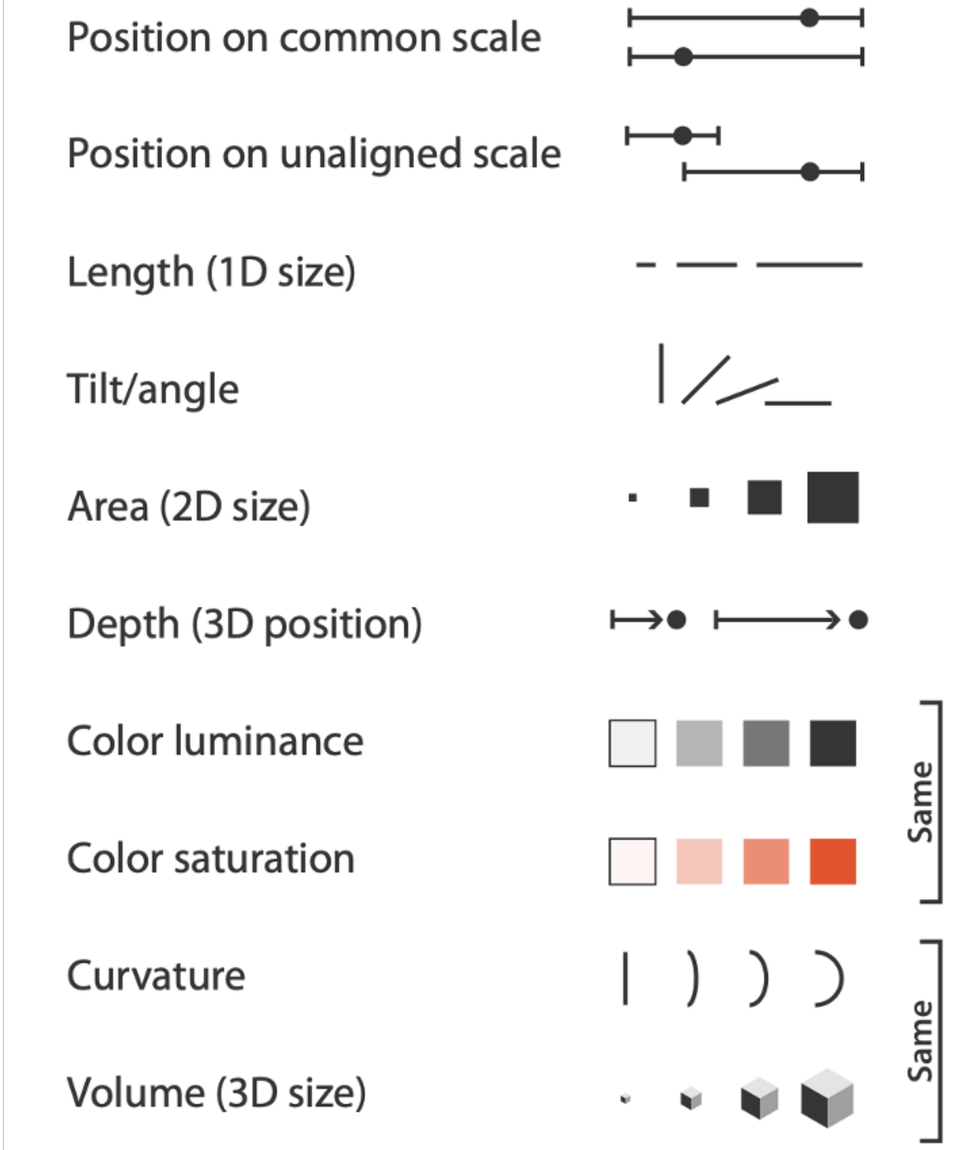


Matejka, J., & Fitzmaurice, G. (2017). Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In Proceedings of the 2017 CHI conference on human factors in computing systems (pp. 1290-1294).

Be honest with your visuals.

Is your visualization true to the underlying data?
Are you *hiding* something?

Magnitude Channels: Ordered Attributes

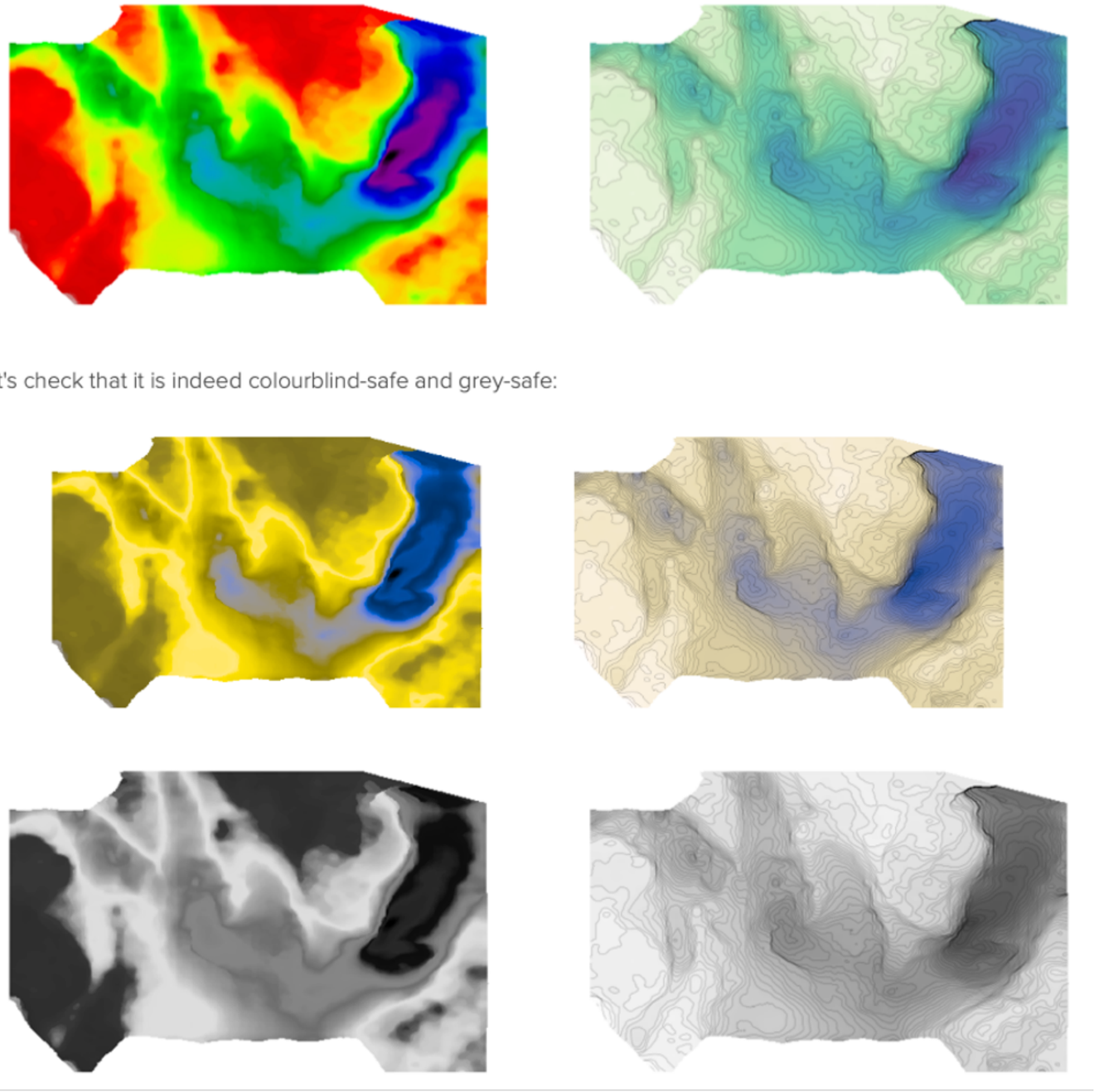


Identity Channels: Categorical Attributes

Friendly reminder about color…

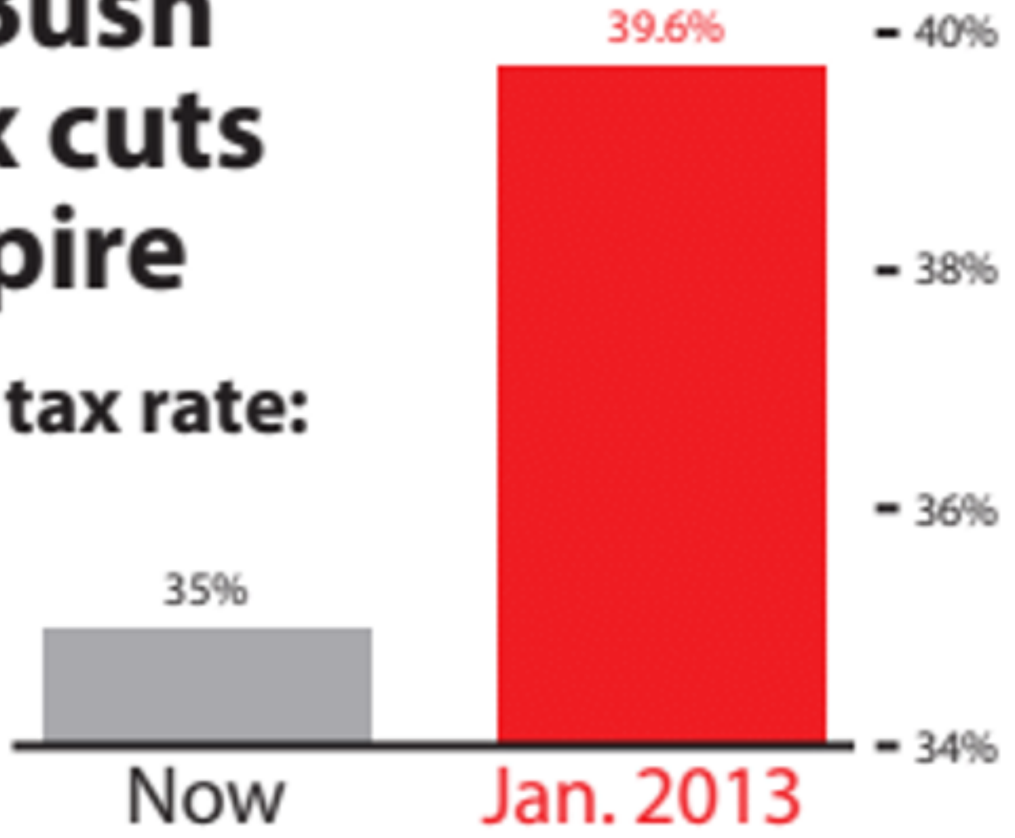
- **For the love of god, do not use jet (rainbow)**
- **I will haunt your dreams if you use jet**
- **Remember accessibility issues! (Red/green)**

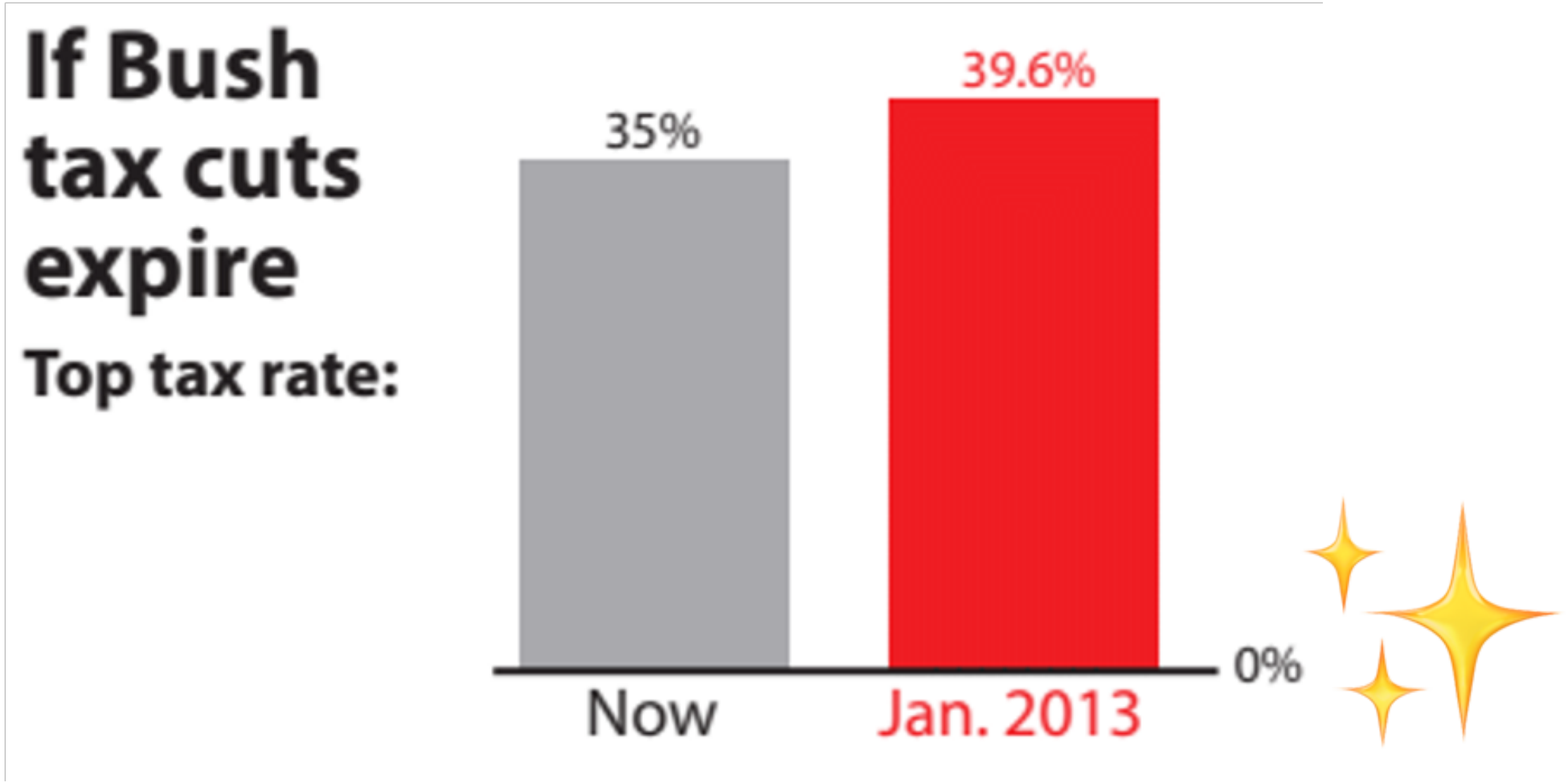
For more on color to honestly depict data, see separate color presentation :)



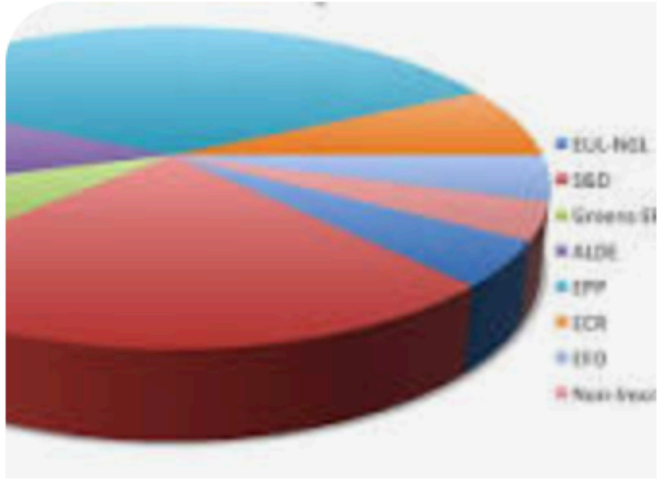
If Bush tax cuts expire

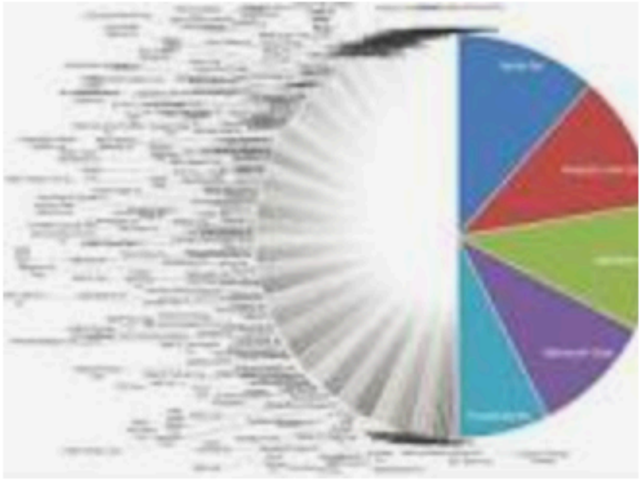
Top tax rate:

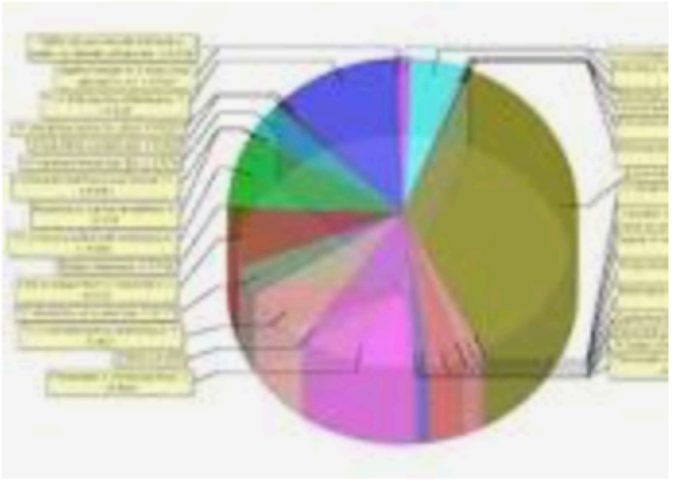


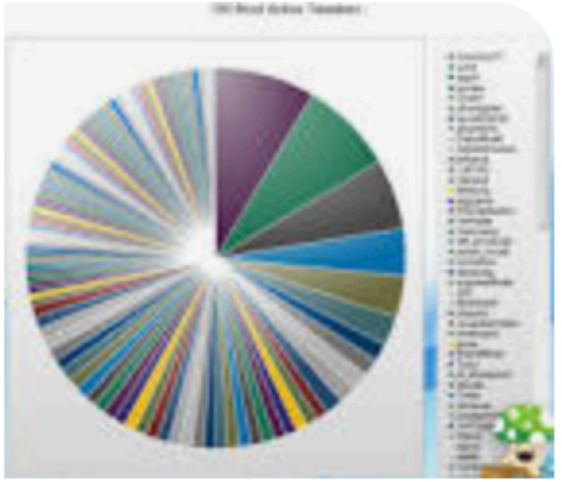


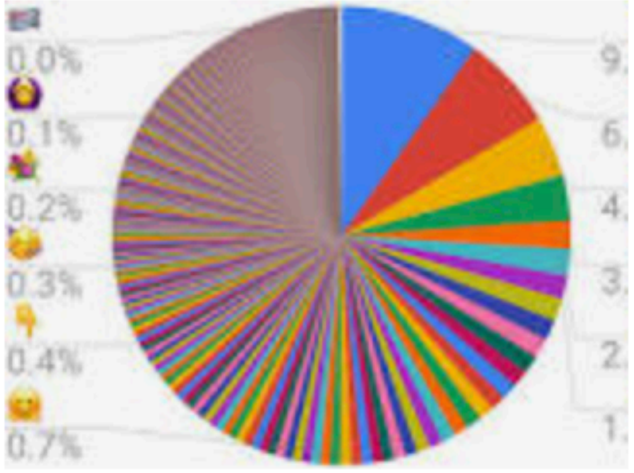
Principle of “proportional ink” (Tufte)

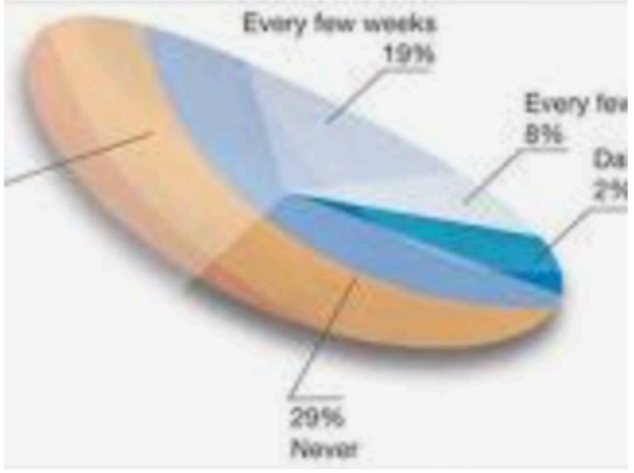

EUL-NGL
S&D
Greens-EF
ALDE
EPP
ECR
EFD

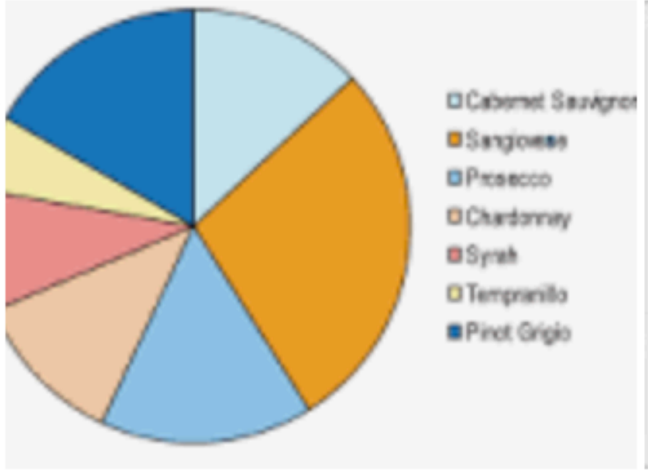


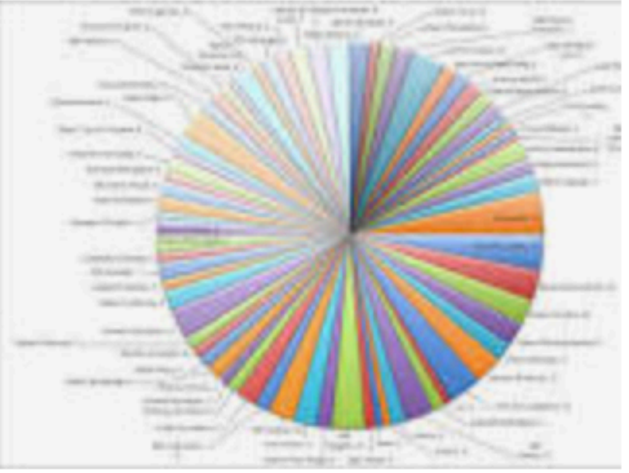


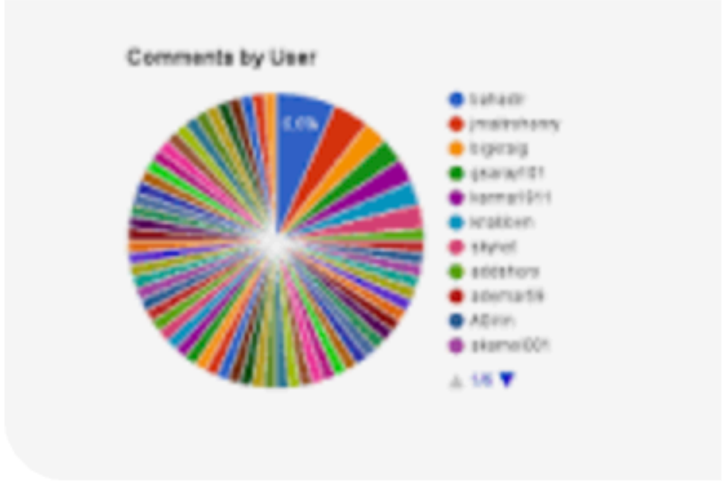


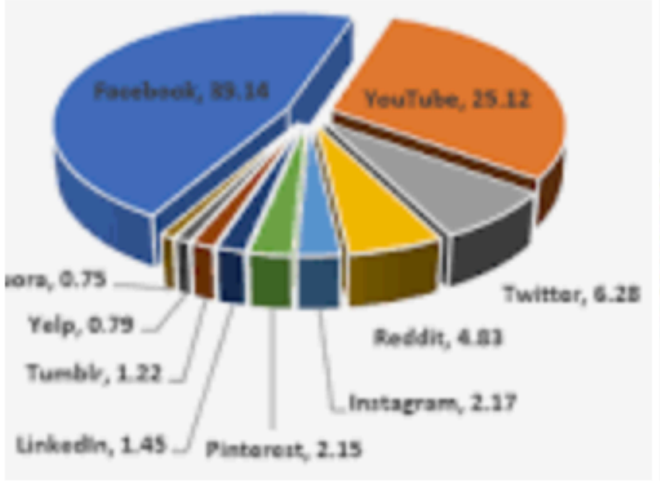

0.0%
0.1%
0.2%
0.3%
0.4%
0.7%

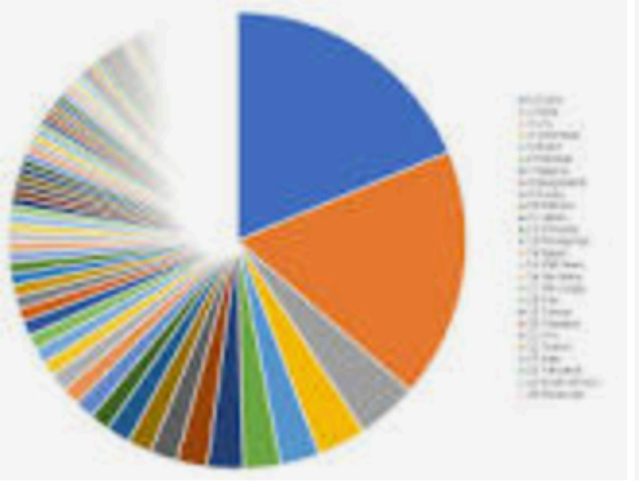

Every few weeks
19%
Every few
8%
Da
2%
29%
Never


Cabernet Sauvignon
Sangiovese
Prosecco
Chardonnay
Syrah
Tempranillo
Pinot Grigio

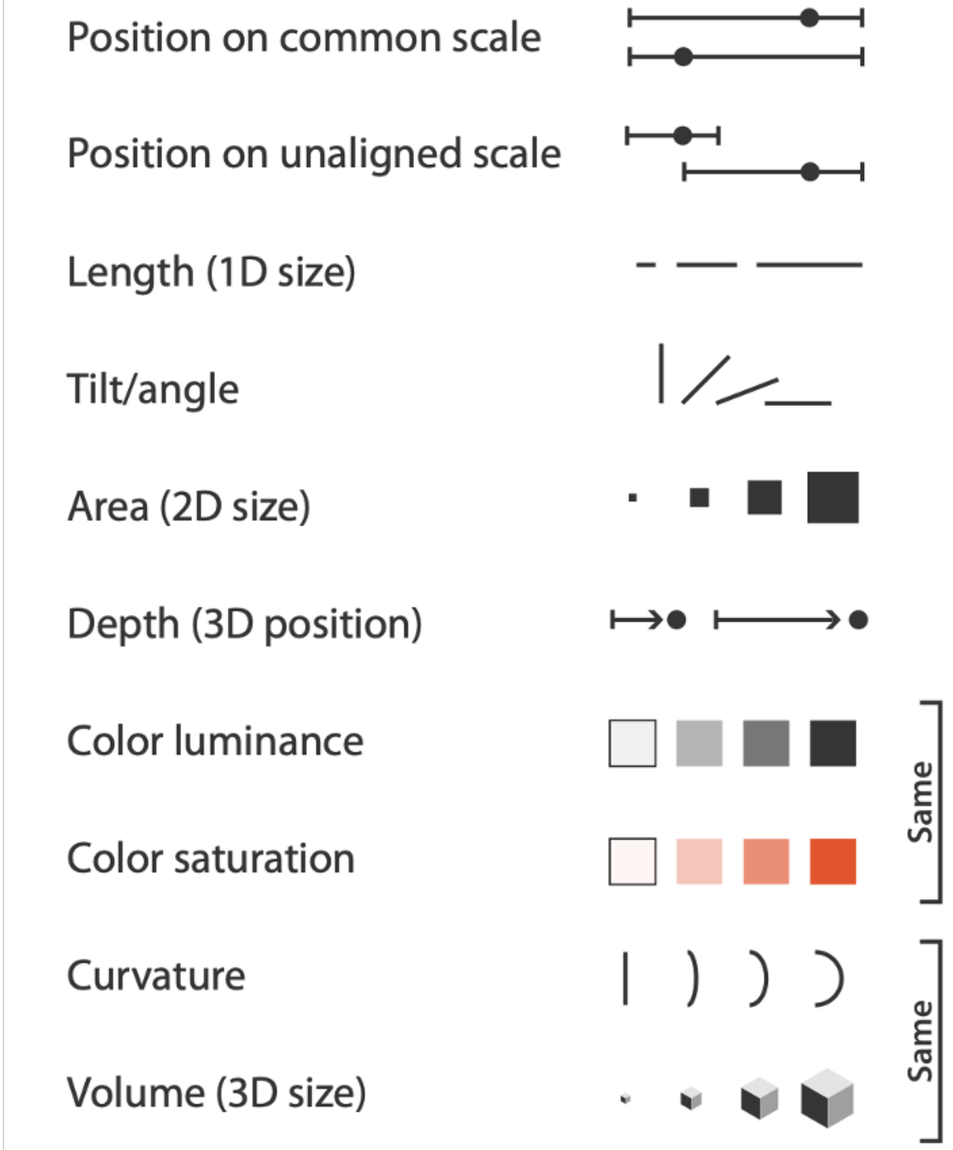



Comments by User


Facebook, 39.14
YouTube, 25.12
Twitter, 6.28
Reddit, 4.83
Instagram, 2.17
Pinterest, 2.15
LinkedIn, 1.45
Tumblr, 1.22
Yelp, 0.79



Magnitude Channels: Ordered Attributes

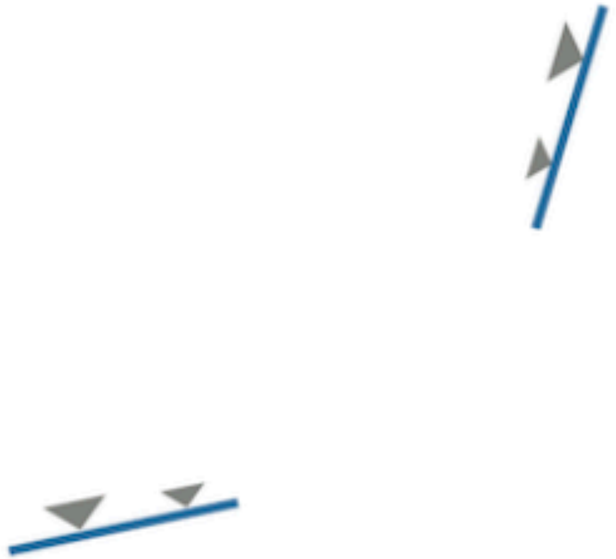


Identity Channels: Categorical Attributes

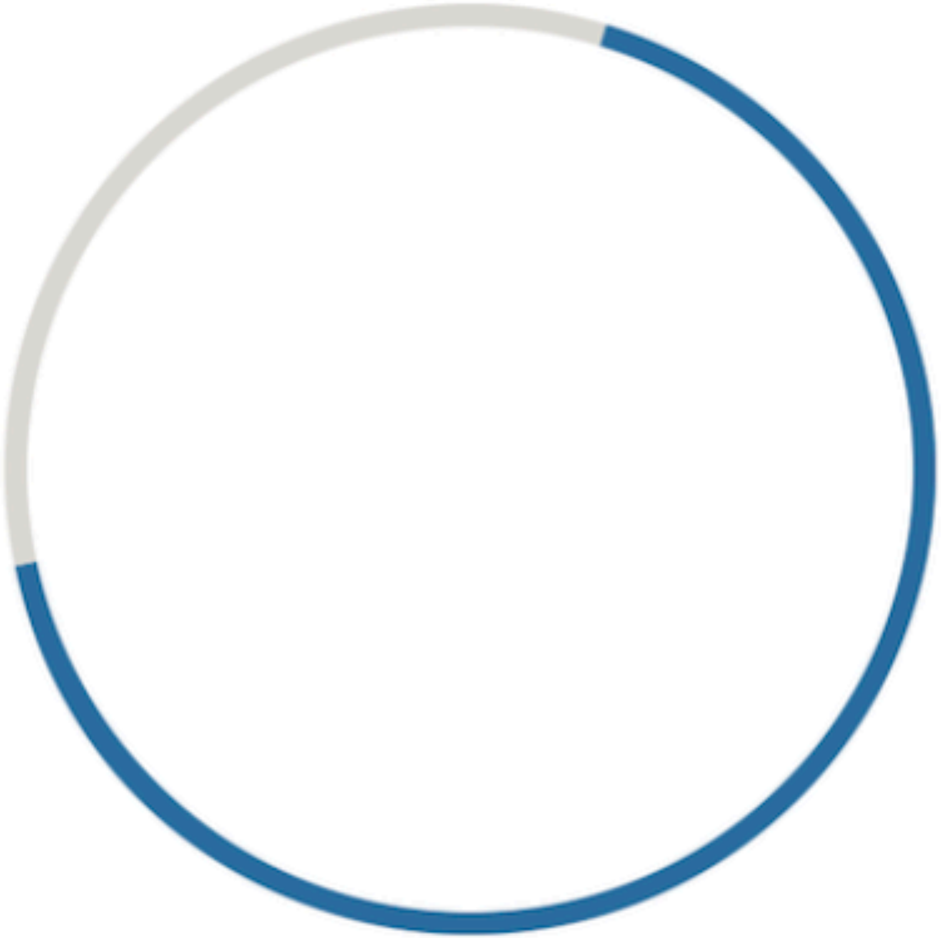
What is the value of the grey segment?

What is the value of the grey segment?

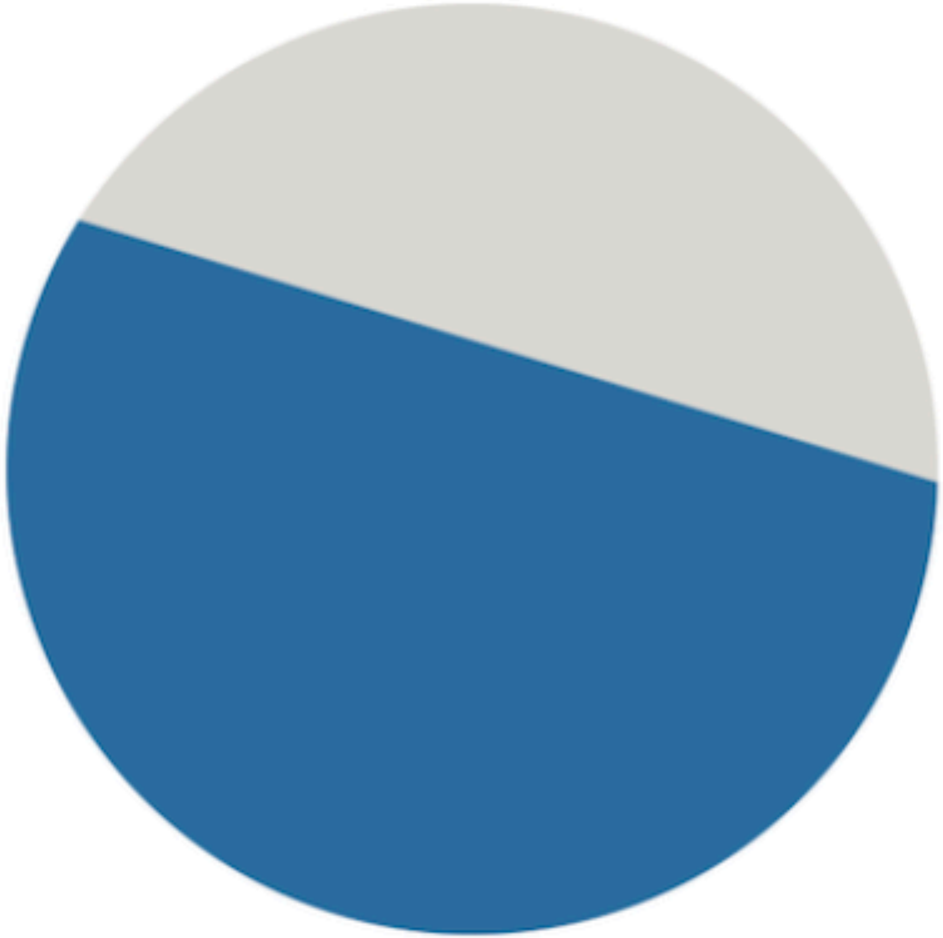
What is the value of the grey segment?



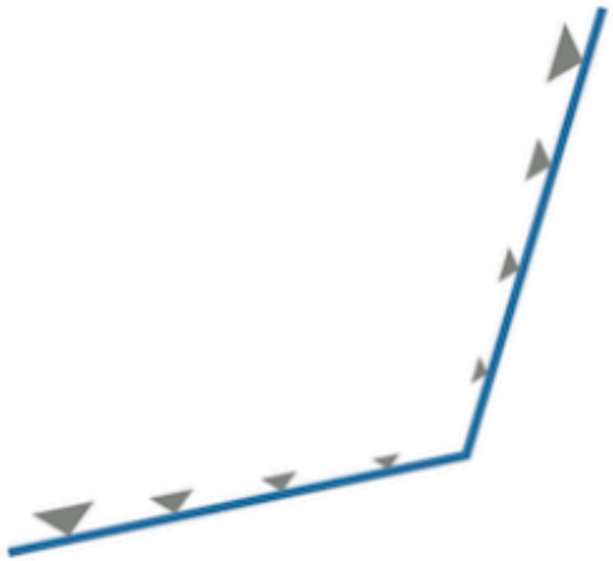
What is the value of the grey segment?

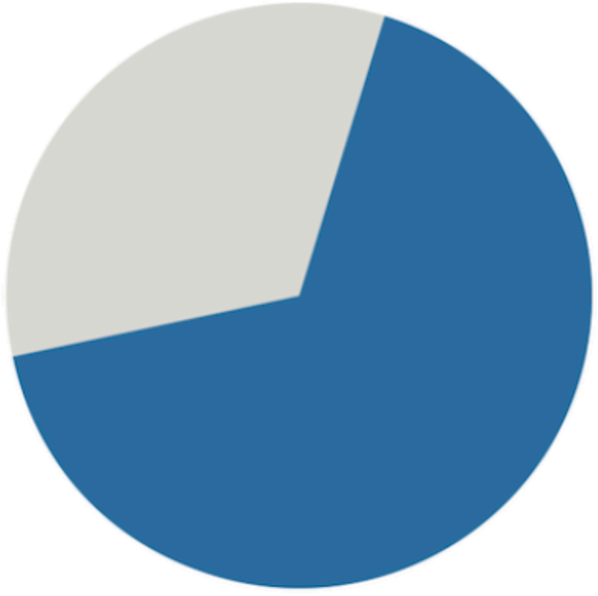


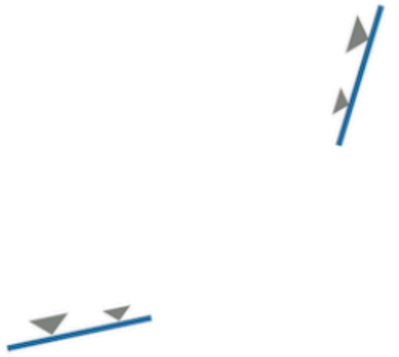
What is the value of the grey segment?

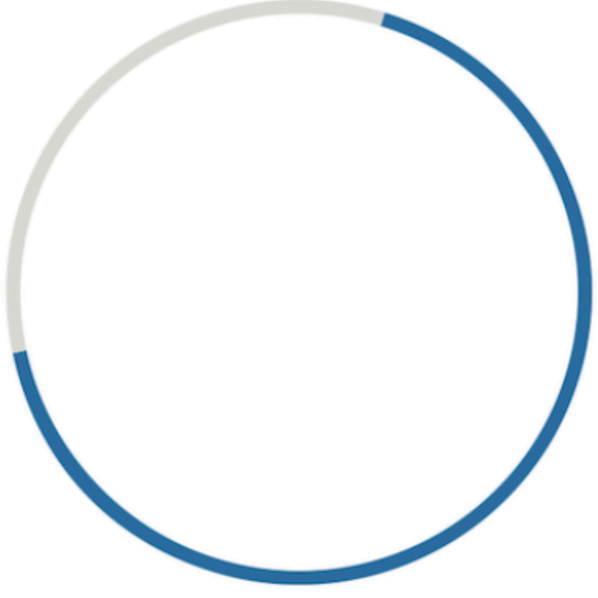


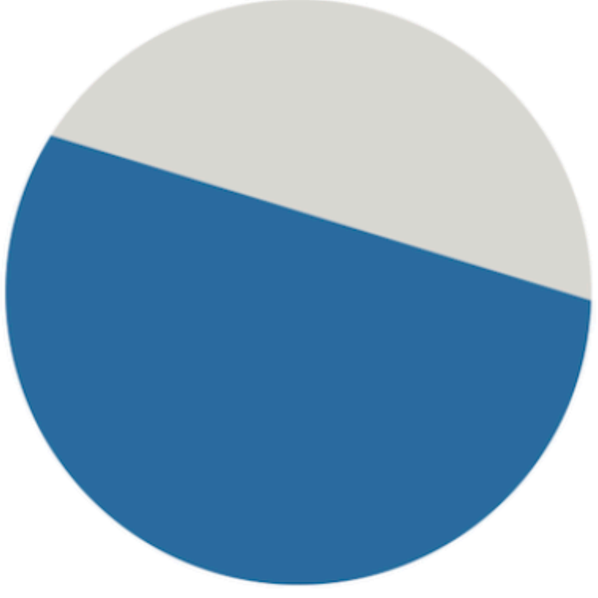
What is the value of the grey side of the segment?

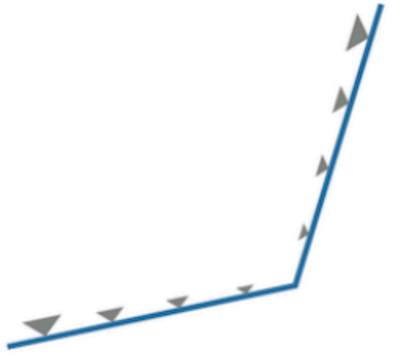






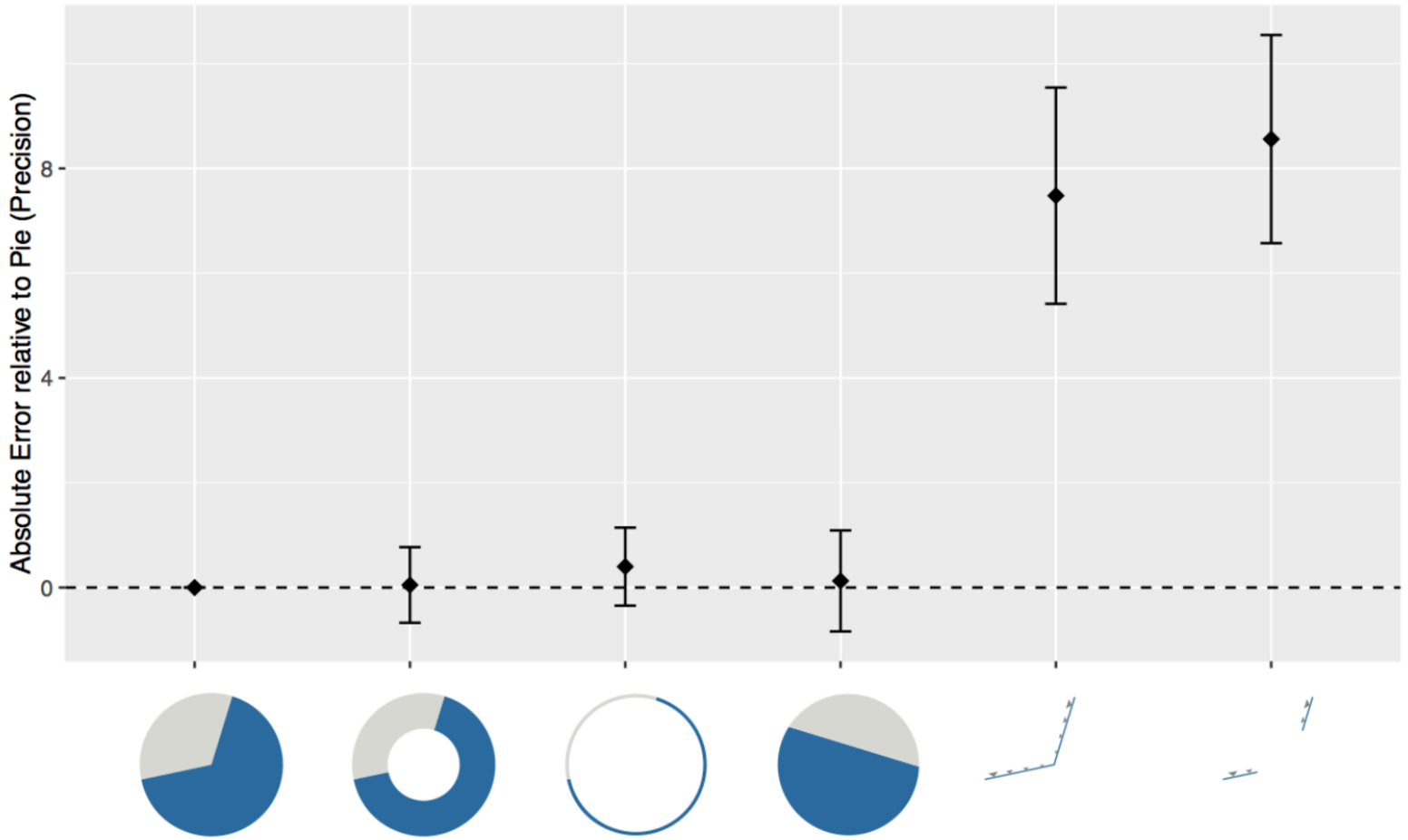






Pie charts

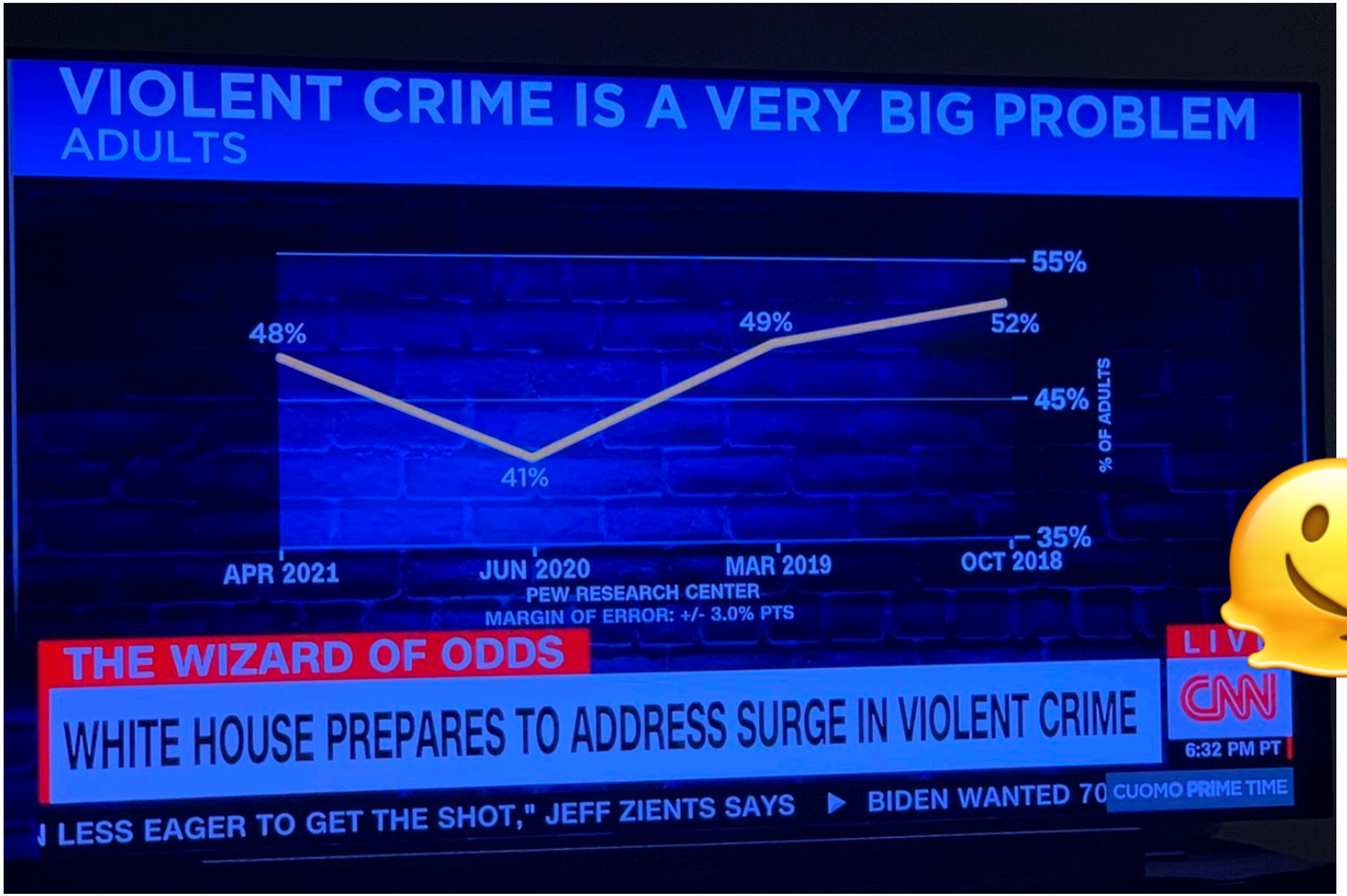
- **Pie charts aren't always bad!**
  - Part-to-whole
  - < 5 categories
- **Reading channel**
  - ~~Angle~~
  - Area
  - Arc length



https://eagereyes.org/blog/2016/an-illustrated-tour-of-the-pie-chart-study-results

Respect common associations

Follow directional and semantic conventions to reduce how hard your audience needs to think!



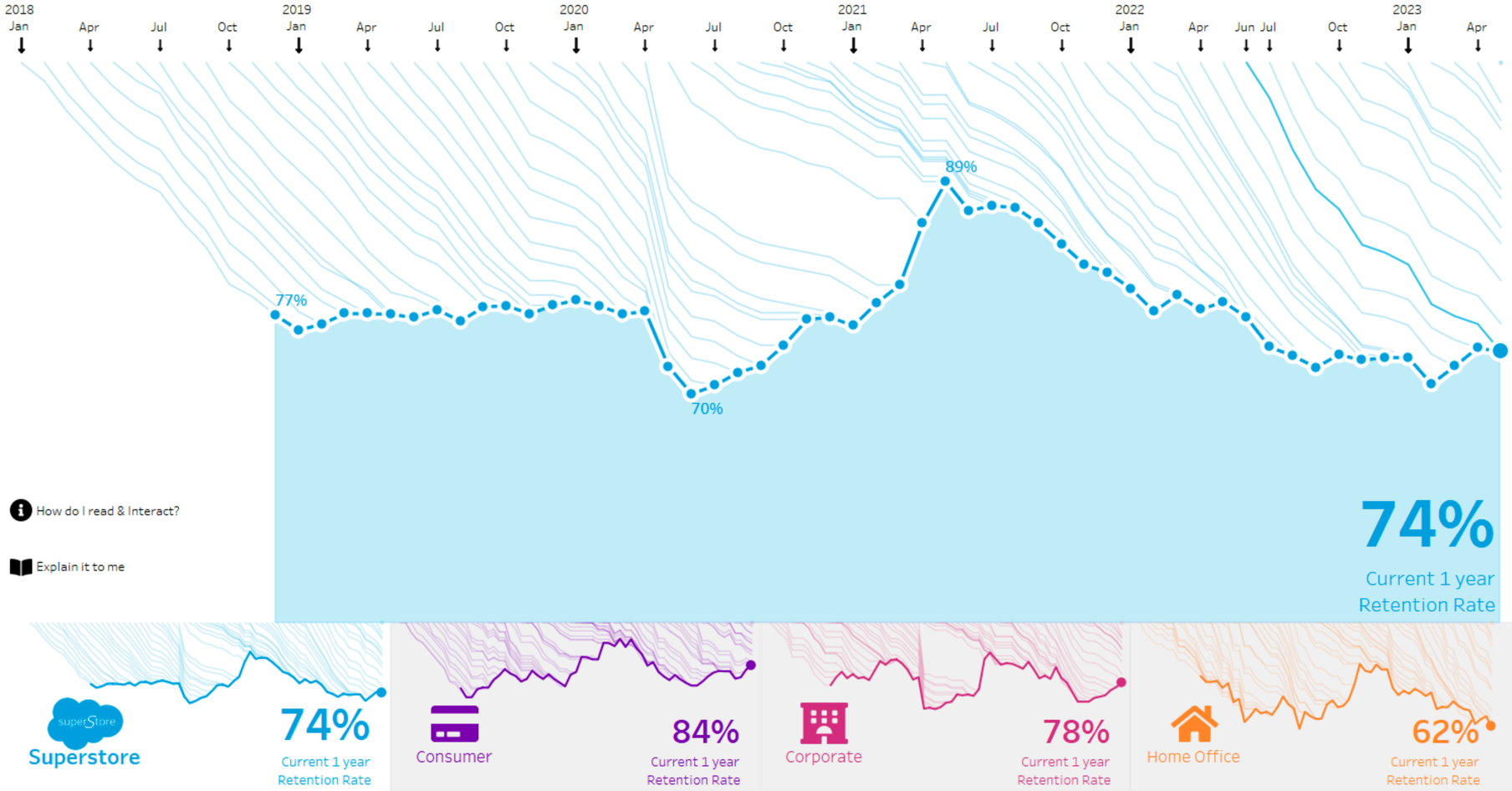
https://viz.wtf/

Employee Retention

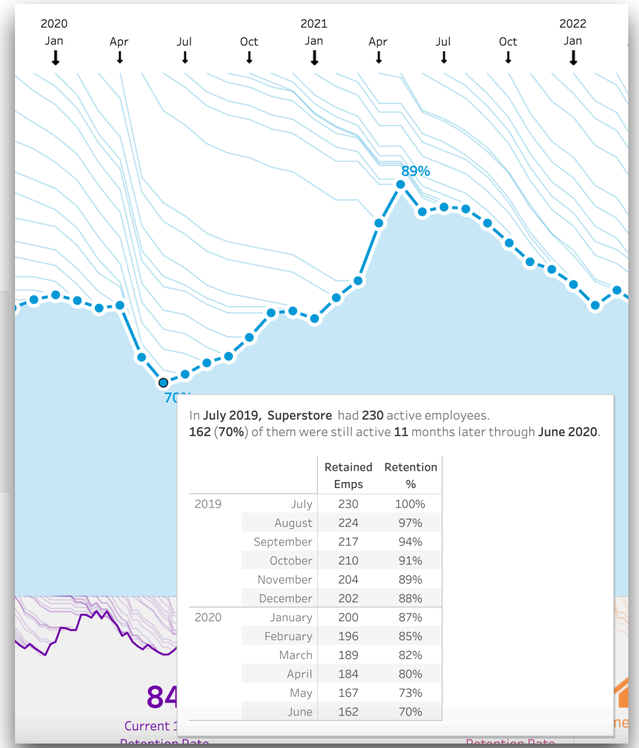
Created by: Jack Hineman
@jackhineman

Explore trends in Employee Retention at **Superstore** or drill into the three segments (**Consumer**, **Corporate**, **Home Office**)

Lightning Bolts: represent the % of Retained Employees from a Starting Point. | *Mountain Range: shows the rolling retention rate over the # of selected months.*

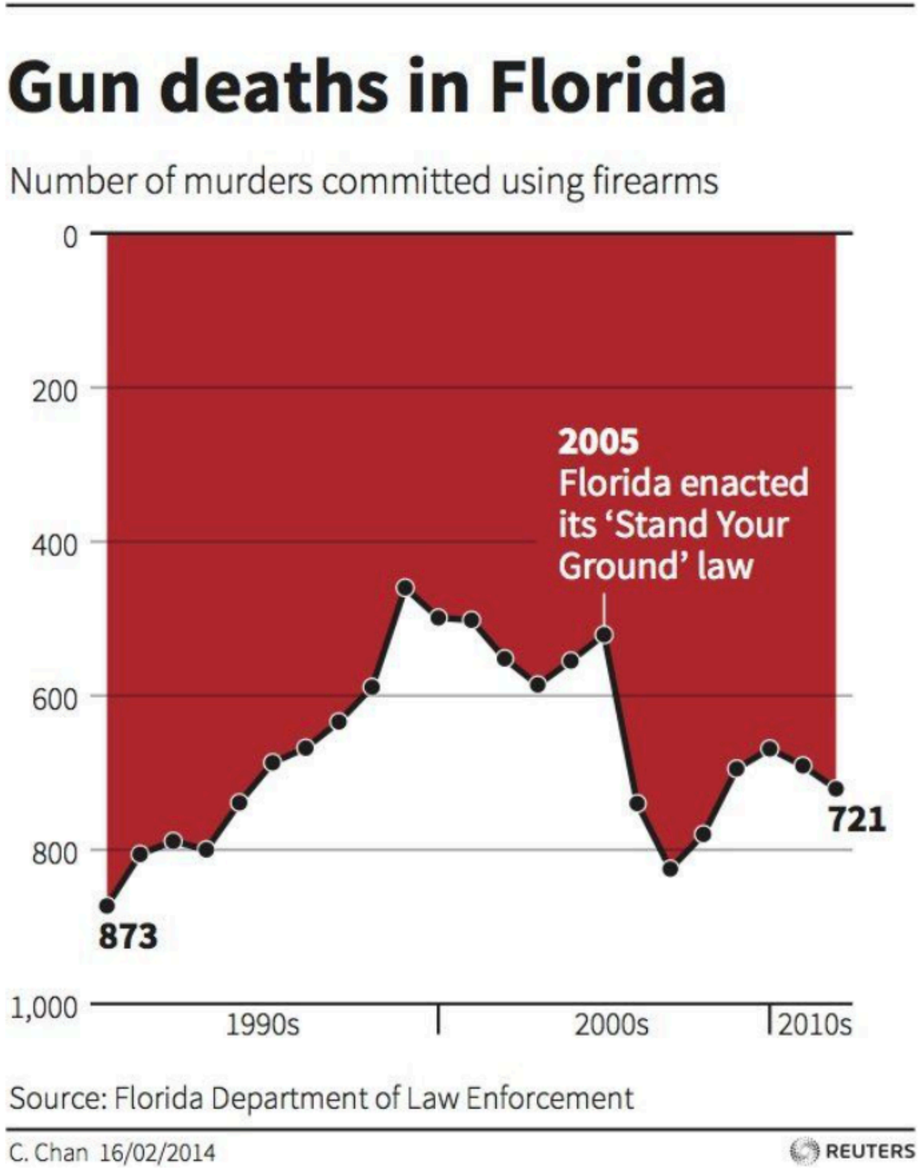


Info on hover:

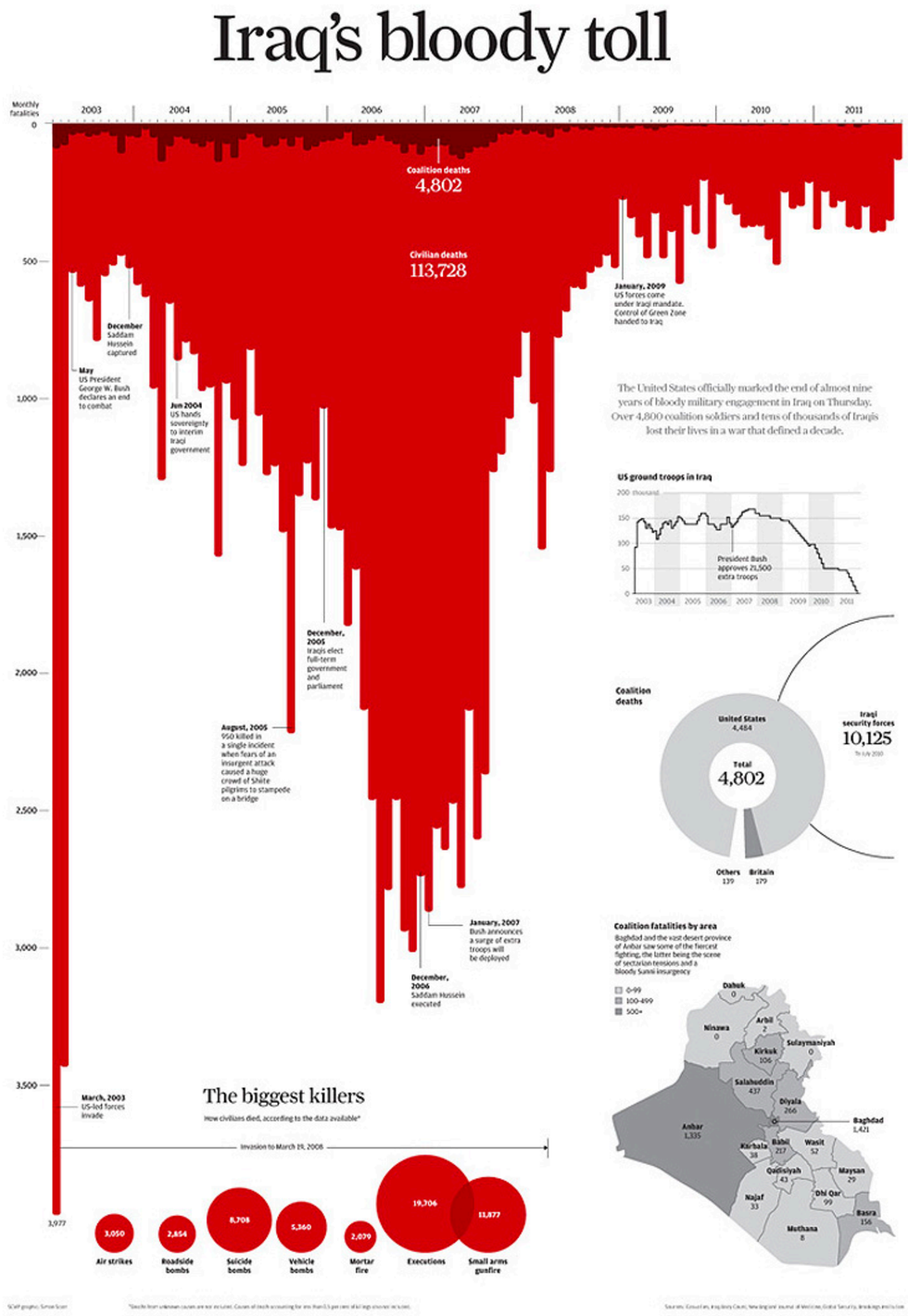


		Retained Emps	Retention %
2019	July	230	100%
	August	224	97%
	September	217	94%
	October	210	91%
	November	204	89%
	December	202	88%
2020	January	200	87%
	February	196	85%
	March	189	82%
	April	184	80%
	May	167	73%
	June	162	70%

- Down -> up == more/increasing
- Left -> right == more/increase/a progression



twitter.com/GraphCrimes

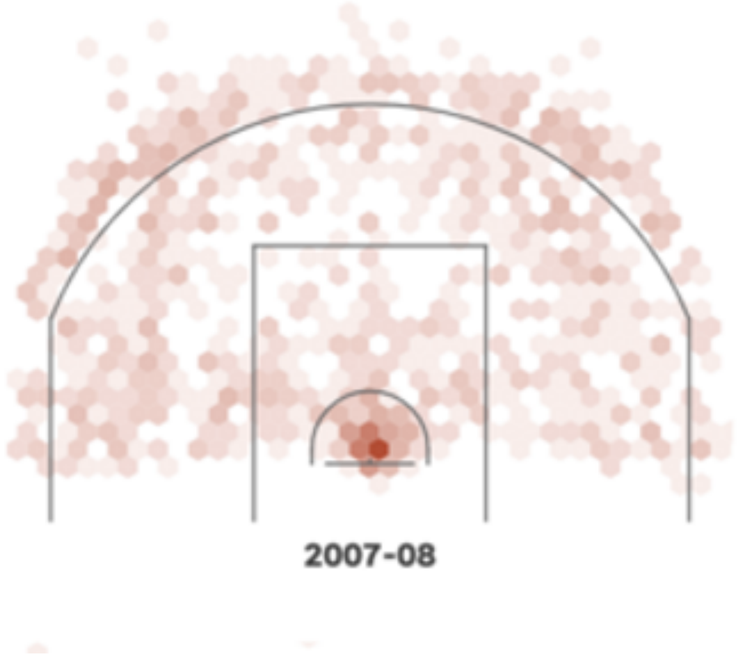


It can be okay to break convention, if done thoughtfully!

The semantics work here: We read the shape, not the inverse shape as in gun deaths chart.

LeBron James has captured the scoring title. We visualized every shot. (USA Today)

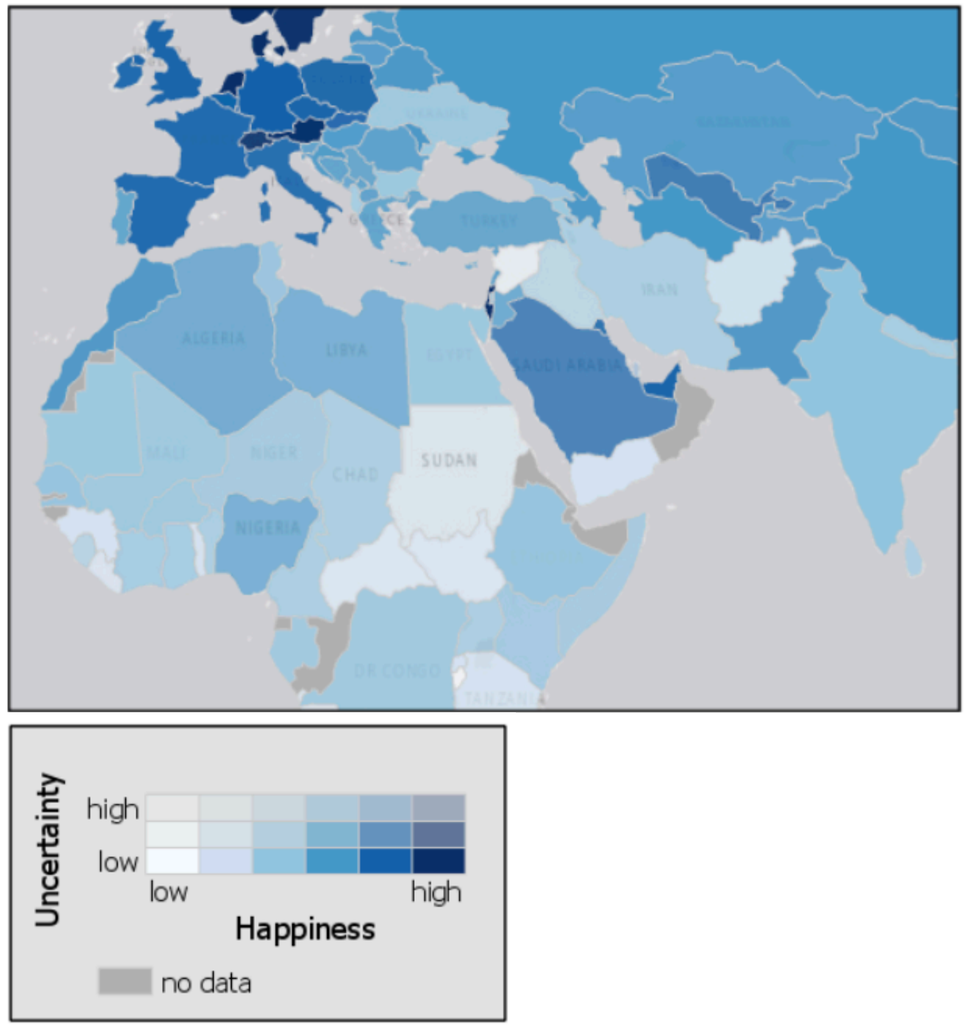
Opacity == Frequency



Cary Anderson, Penn State University.

Data Source: The World Happiness Report (Helliwell et al. 2018), Natural Earth.

Opacity == Uncertainty

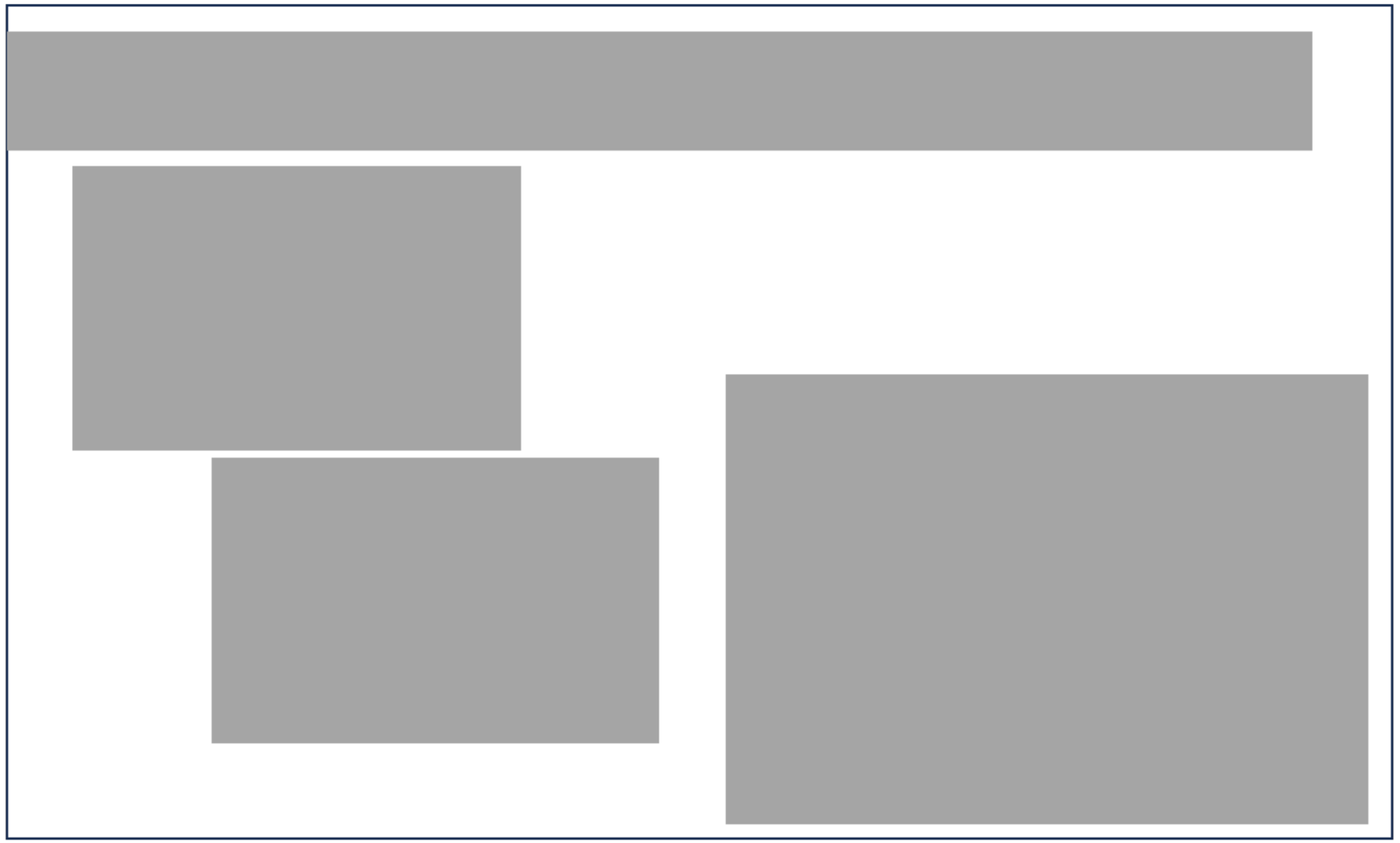


Opacity/tint *can* mean different things.

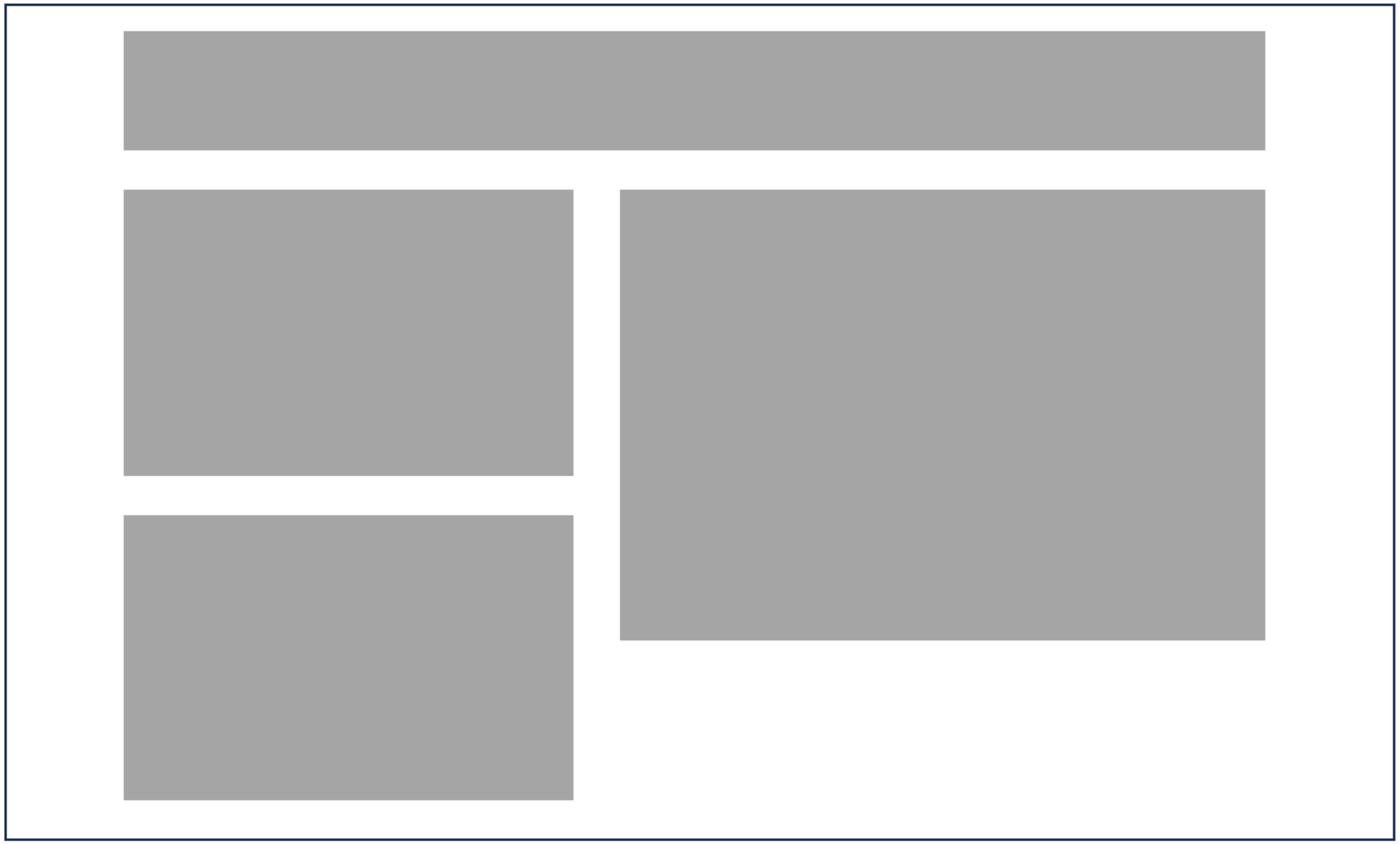
Design a hierarchy of information

Guide the eye through the visualization

Align elements on a grid



Align elements on a grid



Group for comparison

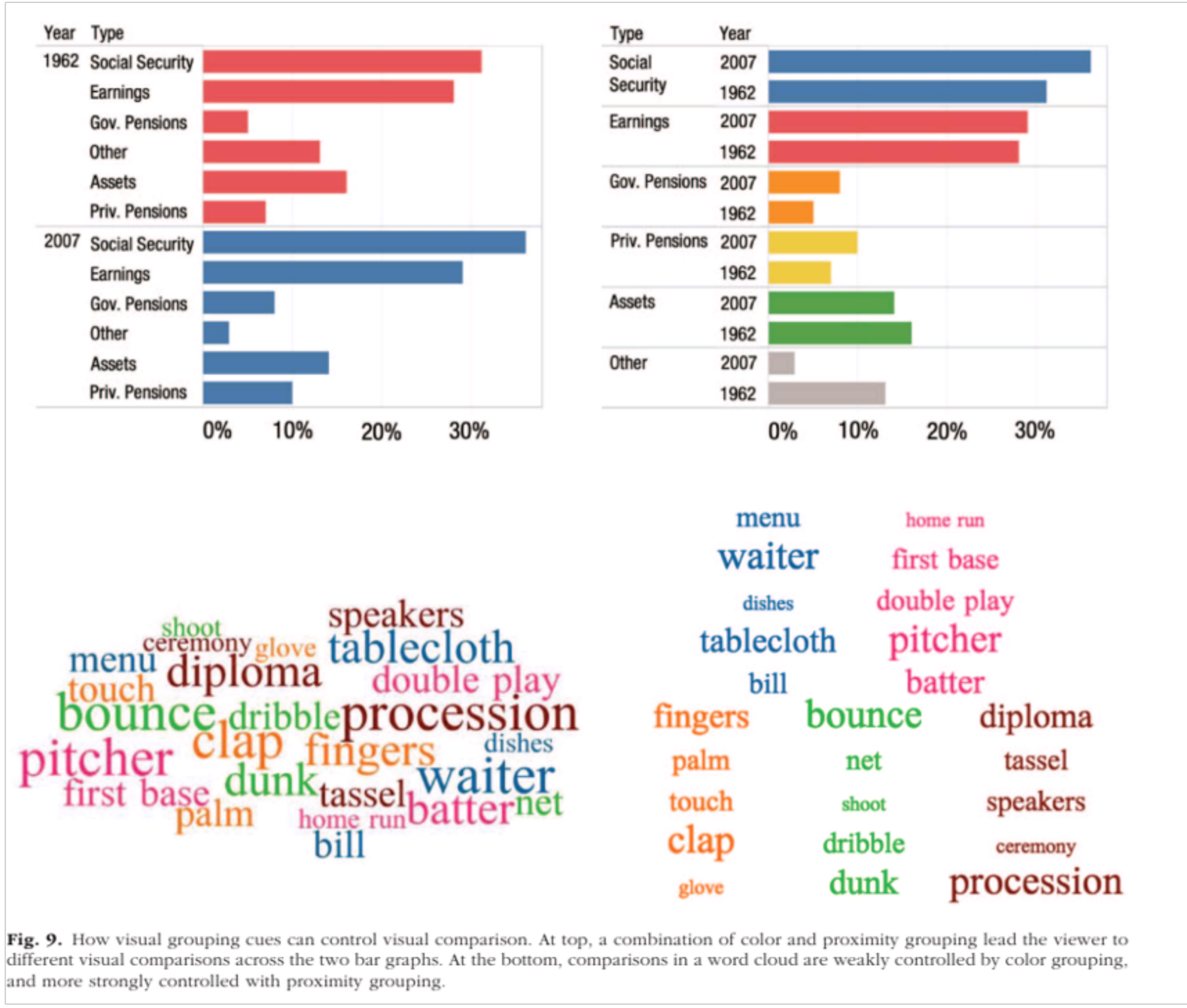


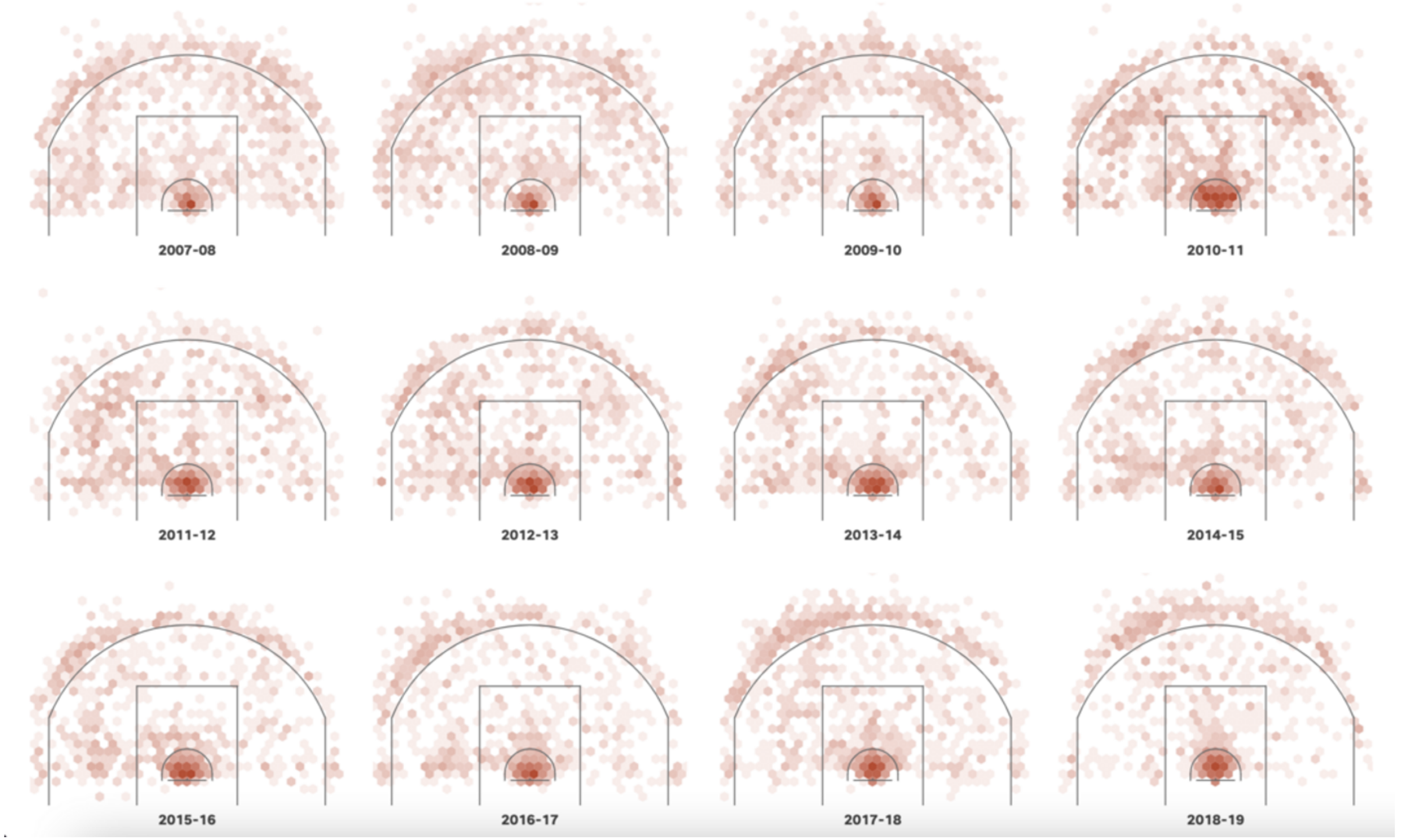
Fig. 9. How visual grouping cues can control visual comparison. At top, a combination of color and proximity grouping lead the viewer to different visual comparisons across the two bar graphs. At the bottom, comparisons in a word cloud are weakly controlled by color grouping, and more strongly controlled with proximity grouping.

Franconeri et al. 2021: The science of visual data communication: What works

Visually subset

LeBron James has captured the scoring title. We visualized every shot.

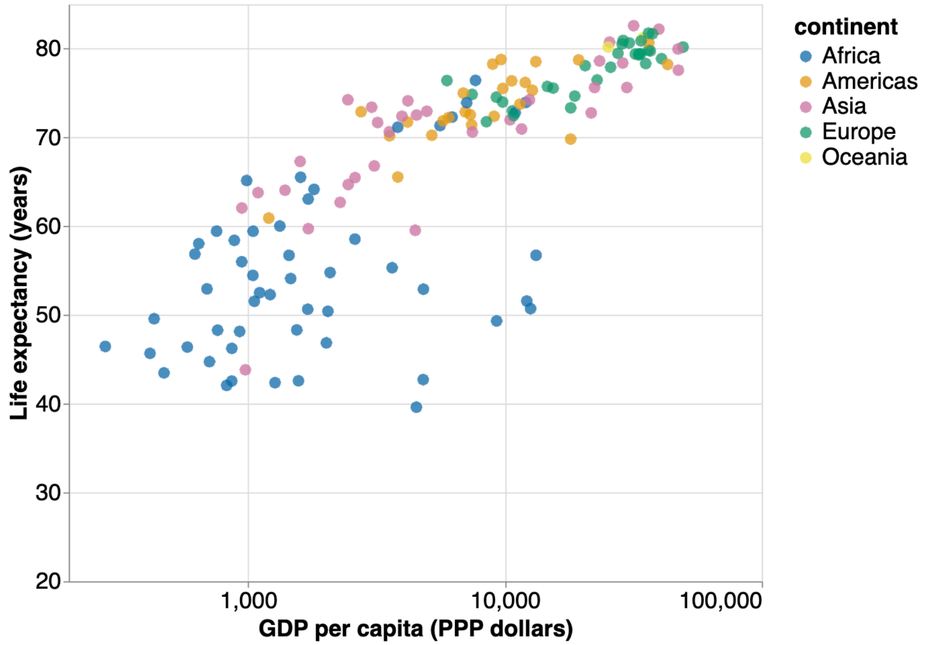
by USA Today



link

Text is an equal partner

- **Basic anatomy of a chart**
  - Title
  - Axis labels
  - Caption
- **Consider different weights, and/or golden ratio or similar to create consistent text hierarchy (**typescale.com**)**



Life expectancy as function of the gross domestic product
Gross domestic product (GDP) per capita measures the value of everything produced in a country during a year, divided by the number of people.
The unit is in purchasing power parities (PPP dollars), fixed to 2017 prices.
Data is adjusted for inflation and differences in the cost of living between countries.

Image source: Radovan Bast

Especially relevant in article figures, posters!

Semantic Snapping for Guided Multi-View Visualization Design

Yngve S. Kristiansen, Laura Garrison, and Stefan Bruckner, *Member, IEEE Computer Society*

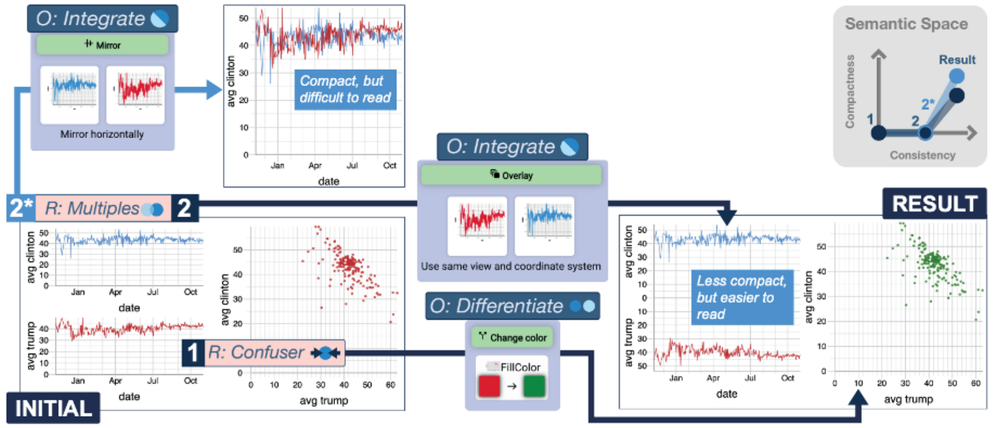


Fig. 1. Semantic snapping allows the user to perform iterative operations to improve the compactness and consistency of a multi-view visualization. Underlying algebraic rules called *relations* define the available *operations* for each iteration. In this example showing the 2016 US Election poll percentages and pollsters, from our initial composition we (1) identify a *confuser* relation between the bottom left and rightmost views showing the same color (red). We *differentiate* these views by selecting green as the fill color for the scatter plot. We next identify a *multiples* relation in the two left views. We resolve this through one of two *integration* operations. (2*) *Overlay* produces an unsatisfactory result, so we revert and (2) perform a *mirroring* operation to arrive at our resulting composition. The semantic map to the right illustrates our path through semantic space.

Abstract—Visual information displays are typically composed of multiple visualizations that are used to facilitate an understanding of the underlying data. A common example are dashboards, which are frequently used in domains such as finance, process monitoring and business intelligence. However, users may not be aware of existing guidelines and lack expert design knowledge when composing

Interactive Hierarchical Data Exploration through Dimensional Bundling

This research is supported by the University of Bergen and the Trond Mohn Foundation in Bergen (BFS13SS), Visualizing Data Science for Large Scale Hypothesis Management in Imaging Biomarker Discovery (MDI), and by the Federal State of Saxony-Anhalt, Germany (FKZ: I 88).

Laura Garrison[1,2], Juliane Müller[3], Stefanie Schreiber[3,4], Steffen Oeltze-Jafra[3,4], Helwig Hauser[1,2], & Stefan Bruckner[1,2]

[1]Dept. of Informatics, Univ. of Bergen, Bergen, Norway
[2]Mohn Medical Imaging & Visualization Centre, Bergen, Norway
[3]Center for Behavioral Brain Sciences, Magdeburg, Germany
[4]Dept. of Neurology, Otto-von-Guericke Univ., Magdeburg, Germany

MOTIVATION

The identification of interesting patterns and relationships is essential to exploratory data analysis. This becomes increasingly difficult in high dimensional datasets, and even more difficult when these data contain mixed numerical and categorical data types. To the first high-dimensionality problem, while dimensionality reduction techniques are frequently applied to reduce the analysis space, these approaches may unintentionally bury key dimensions within a larger grouping and obfuscate meaningful patterns. To the second problem of mixed data types within a dataset, previous approaches have either treated these different data types separately, or used one type to filter the other. This presents a missed opportunity to identify relationships between these different data types.

CONCEPTUAL FLOW

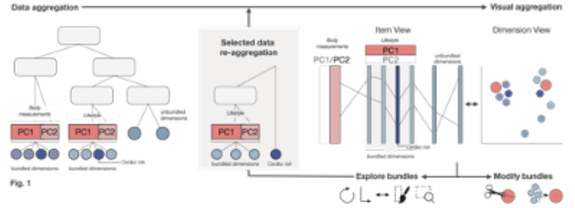


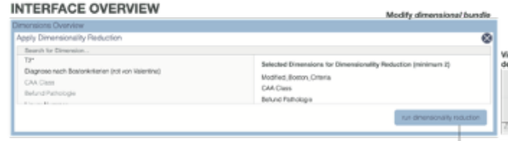
Fig. 1

Our Integrated Dual Analysis + DimLift conceptual workflow (Fig. 1) and interface (Fig. 2) combines statistical analysis with user interactions.

INTEGRATED DUAL ANALYSIS

The first phase of this project [1] extends the Dual Analysis framework [2], a technology that enables the simultaneous joint exploration of both items and dimensions, to allow for joint analysis of numerical and categorical data. This joint analysis is made possible through our computation of common statistical measures for all dimensions.

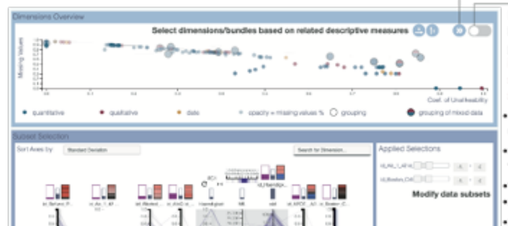
INTERFACE OVERVIEW



DIMLIFT

Algorithm 1: *Dimensional bundle creation for two or more dimensions*

```
initialize pool = all dimensions in dataset
do
    mark all dims in pool as possibly contributing
    initialize new bundle
    perform FAMD on pool
    for all dimensions in pool
        if PC1 loading ≥ contribution threshold
            move dimension from pool to new bundle
        else
            mark dimension as non-contributing
while pool contains dimensions marked as
    non-contributing
for all bundles
    perform FAMD on bundle
    store PC1 and PC2 for bundle
```



Dimensions and bundles are plotted against interchangable statstical measures:

- coefficient of unalikeability
- standard deviation/stdDev
- variance/VA
- modality
- percentage missingness

The second phase proposes a novel approach to dimensionality reduction via the creation of *dimensional bundles* [3]. Generated through an iterative dimensionality reduction or user-driven approach, dimensional bundles are expressive groups of dimensions that contribute similarly to the variance of a dataset.

CASE STUDY

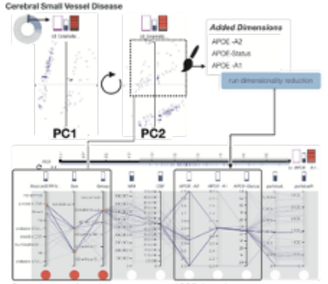


Fig. 3

Fig. 2

Radar plot on hover demonstrates descriptive measure values for a given dimension or bundle. Overall deviation plots normalized aggregation of deviations from subset selection.

We exemplify the power of our integrated dual analysis and dimensional bundling techniques for data exploration and hypothesis generation in an expert case study on clinical cohort data (Fig. 3). Interactive exploration and reconstruction methods via a layered parallel coordinates plot allow users to lift interesting and subtle relationships to the surface, e.g., Boston STRIVE criteria alongside diagnostic group and key test/imaging dimensions. Users may modify bundles as they form new hypotheses, e.g., addition of APOE to an existing bundle.

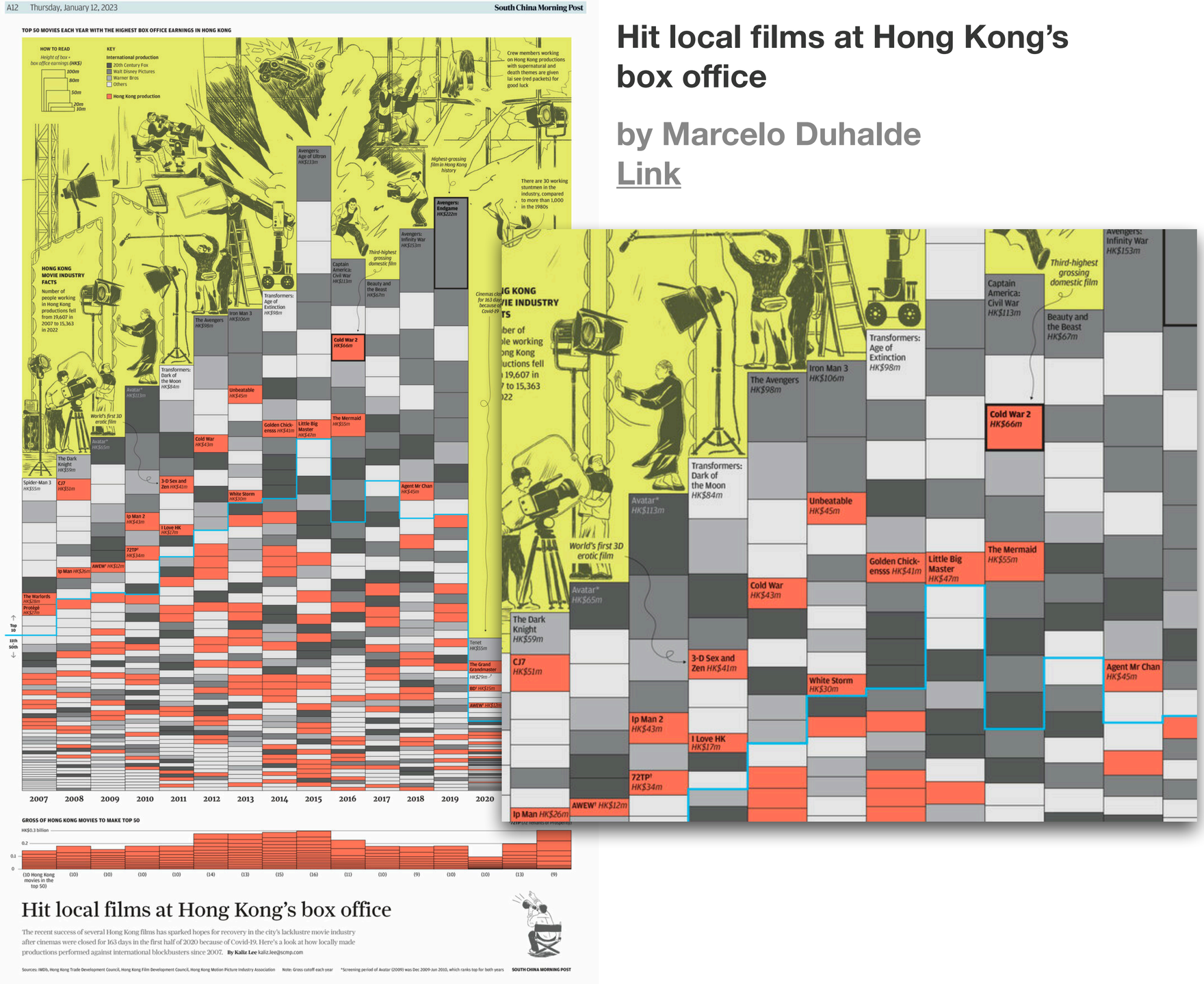
Explore and identify interesting patterns in high dimensional mixed-type datasets.

REFERENCES [1] J. Müller, L. Garrison, S. Schreiber, S. Bruckner, H. Hauser, and S. Oeltze-Jaffra, "Integrated Dual Analysis of Quantitative and Qualitative High-Dimensional Data," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 6, pp. 2653-2665, 1 June 2021, [2] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data," IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, pp. 2591–2599, 2011, [3] L. Garrison, J. Müller, S. Schreiber, S. Oeltze-Jafra, H. Hauser and S. Bruckner, "DimLift: Interactive Hierarchical Data Exploration Through Dimensional Bundling," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 6, pp. 2908-2922, 1 June 2021.

Design with grey

Use color strategically to highlight elements of interest in your data!

J. Schwabish. *Better data visualizations: A guide for scholars, researchers, and wonks* (Columbia University Press, 2021).



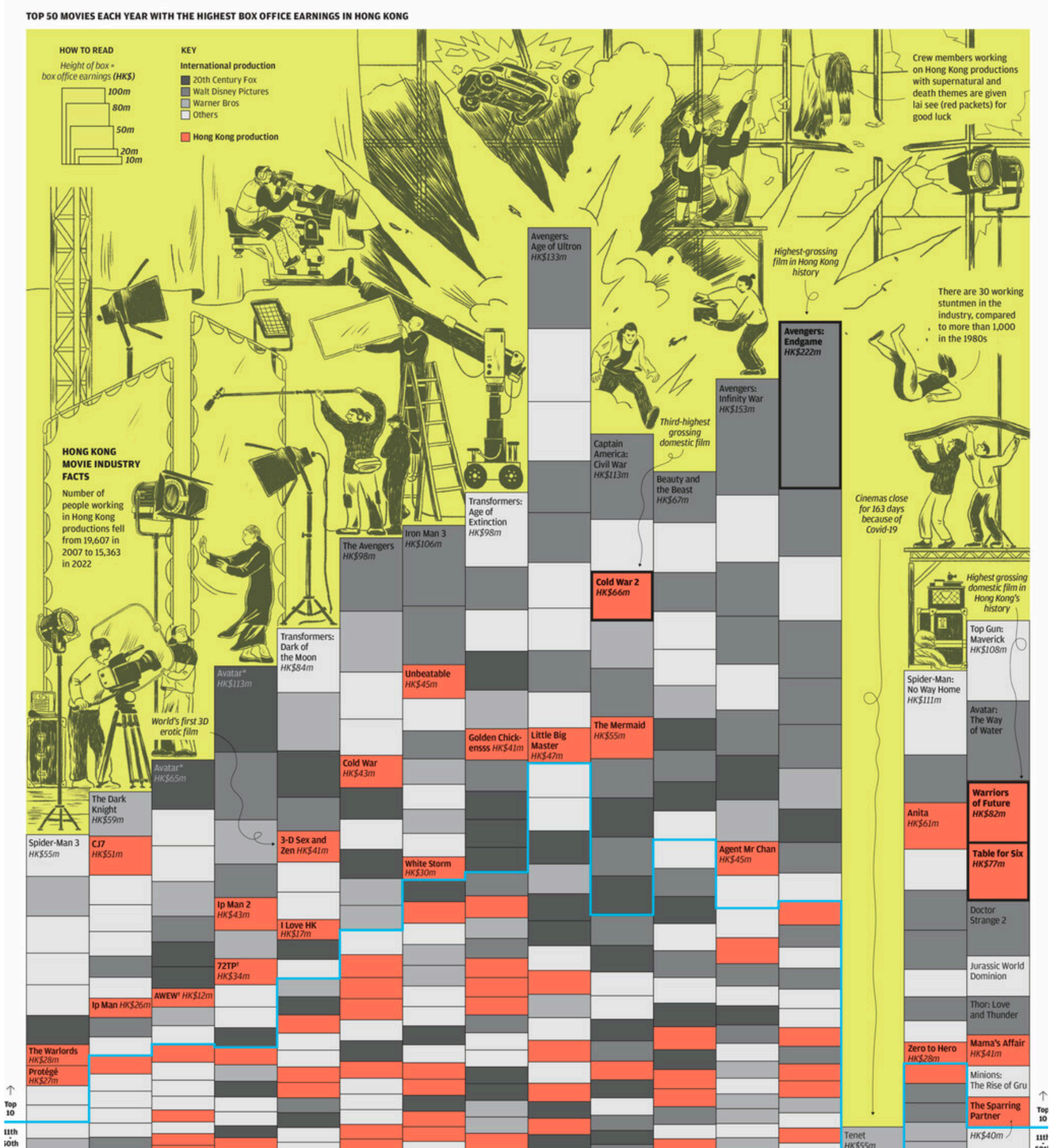
Hit local films at Hong Kong's box office

by Marcelo Duhalde

Link

Tell a story

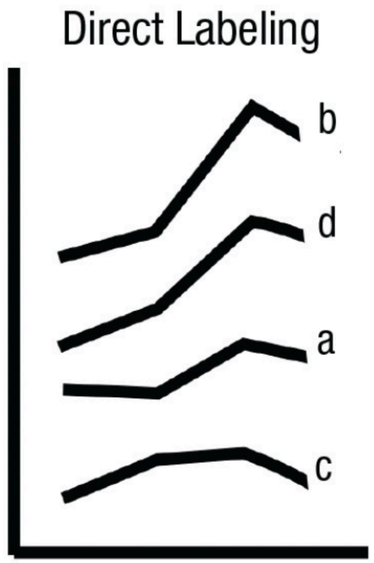
- **Explain trends**
- **Highlight key messages**
- **Add explainers**

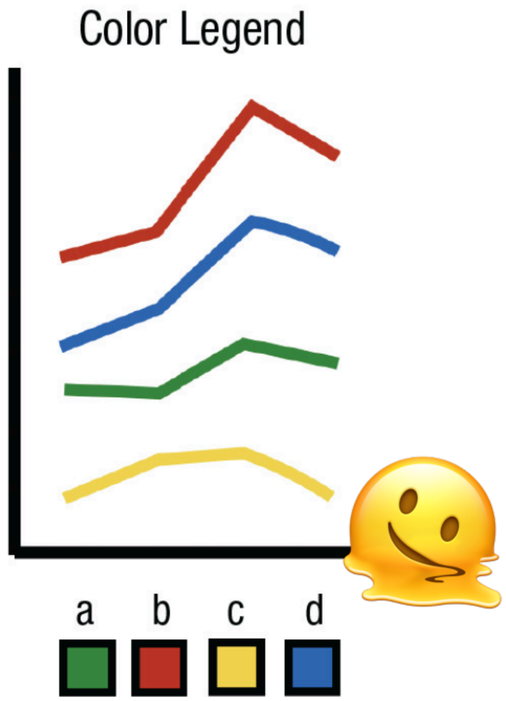


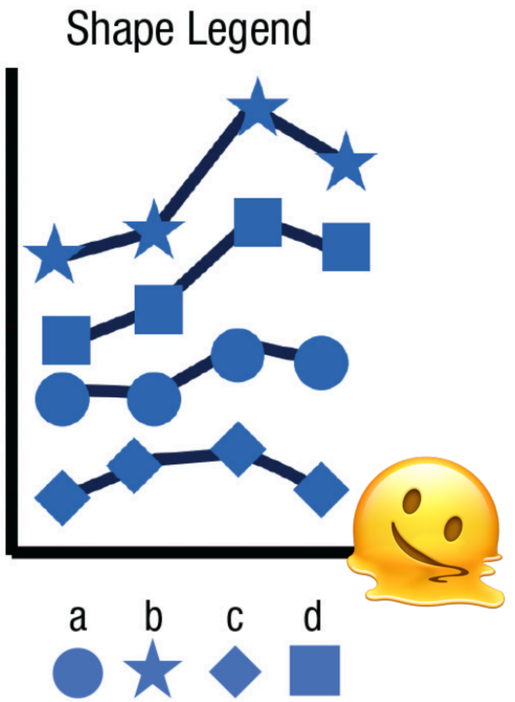
J. Christiansen, Building Science Graphics. CRC Press 2023

Avoid taxing working memory

- **Declutter to minimize distractions** (unless elements help to tell story)
- **Integrate legends where possible and where it makes sense!**
- Be careful with animation







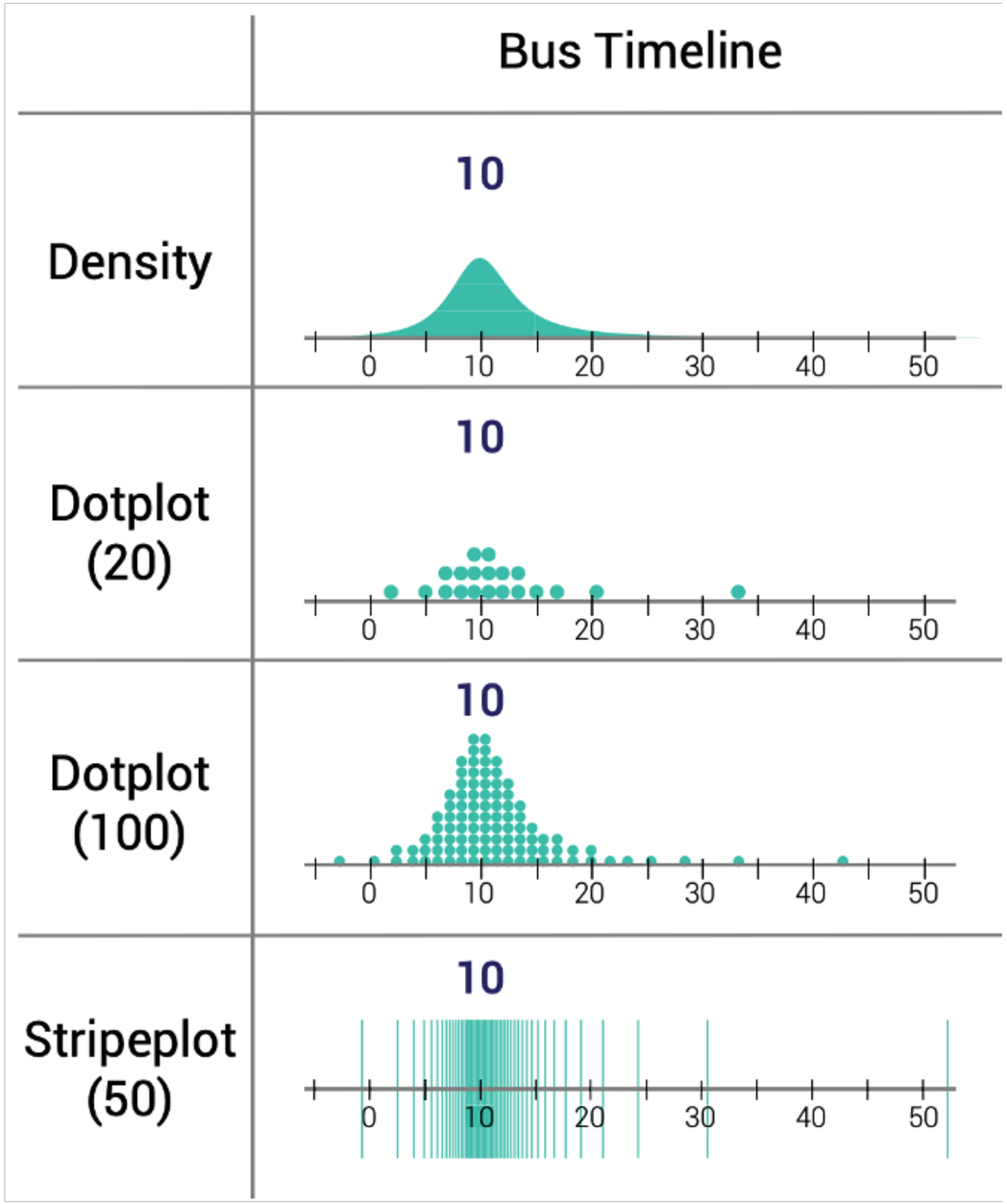
The Science of Visual Data Communication: What Works, Franconeri et al. 2021

Reflect on uncertainties, unknowns

Data biases? Missingness?

Showing uncertainty

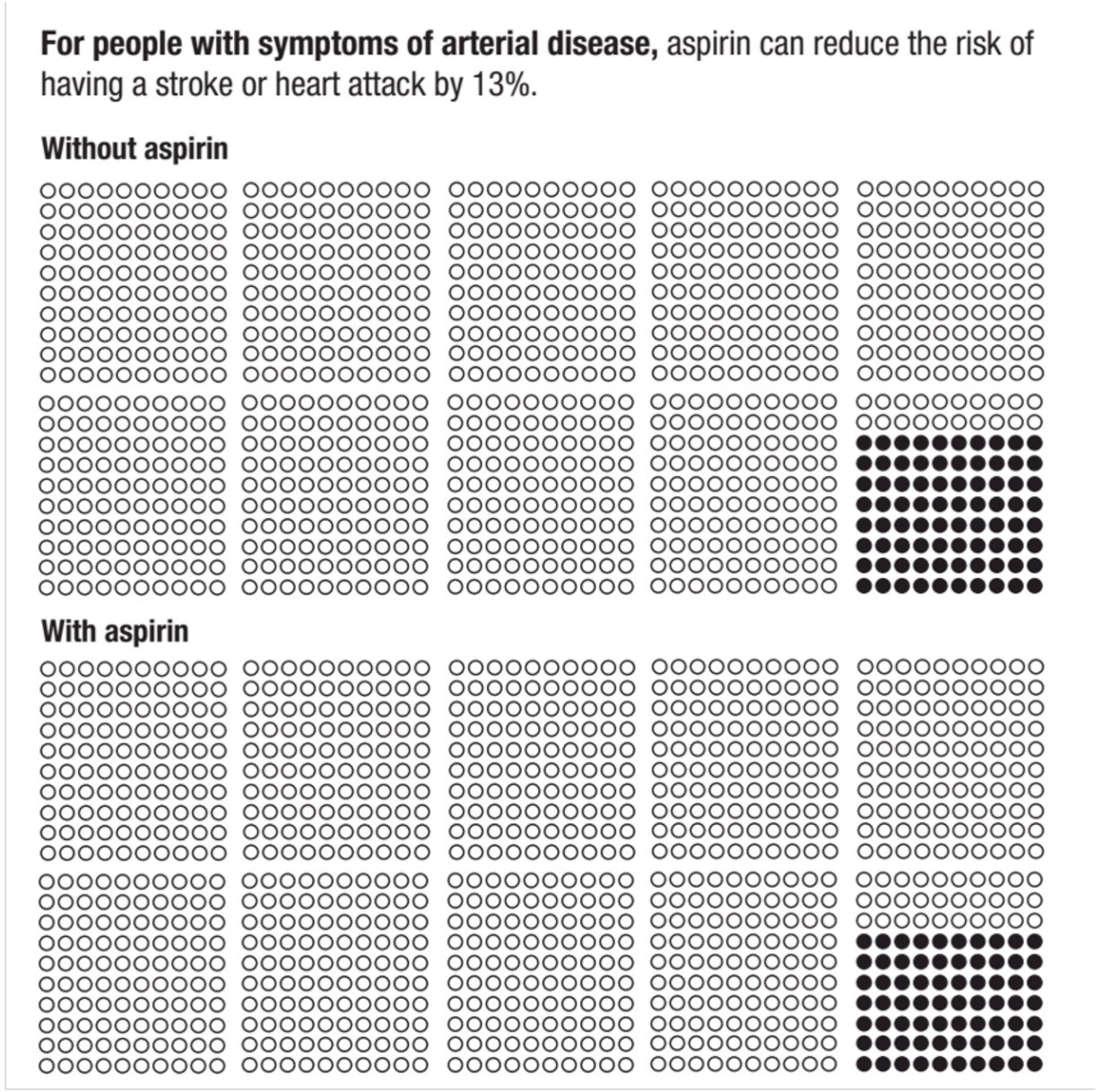
- **Error bars are often misinterpreted**
- **Consider a probability distribution instead of summary stats** (back again to showing data)
- **Different ways to show distribution…**



M. Kay et al. "When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems." *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016.

Showing uncertainty

- **Absolute vs relative risk**
- **Consider showing probabilities as frequencies, not percentages**
- **Icon arrays with same denominator**



The Science of Visual Data Communication: What Works, Franconeri et al. 2021

(Some) tips for making a good visualization

1. Define your goals
2. Show the data (go beyond summary statistics)
3. Be honest with your visuals
4. Respect common associations
5. Design a hierarchy of information
6. Avoid taxing working memory
7. Tell a story
8. Reflect on uncertainty and unknowns

TL;DR

1. Know **why** you are making a visualization
2. Know **who** you are visualizing for
3. Tell a **story**

Want more?

Further reading

- Christiansen, J., 2022. Building Science Graphics: An Illustrated Guide to Communicating Science Through Diagrams and Visualizations. CRC Press.
- Franconeri, S.L., Padilla, L.M., Shah, P., Zacks, J.M. and Hullman, J., 2021. The science of visual data communication: What works. *Psychological Science in the public interest*, *22*(3), pp.110-161.
- Gleicher M. Considerations for visualizing comparison. IEEE transactions on visualization and computer graphics. 2017 Aug 29;24(1):413-23.
- Nature Points of View: Data visualization article collection (https://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html)
- Rogowitz, B.E., Treinish, L.A. and Bryson, S., 1996. How not to lie with visualization. *Computers in physics*, *10*(3), pp.268-273.
- Schwabish J. Better data visualizations: A guide for scholars, researchers, and wonks. Columbia University Press; 2021 Dec 31