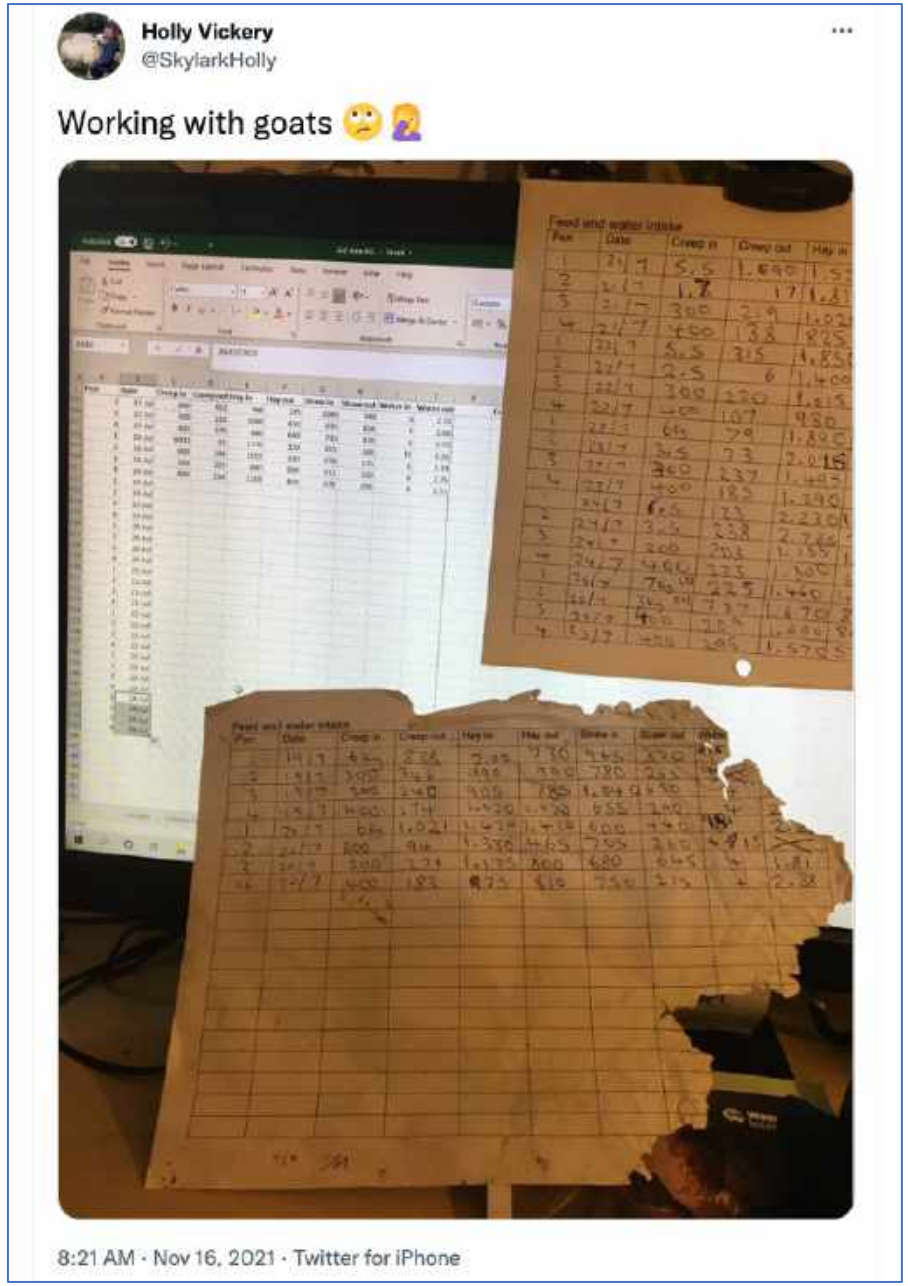
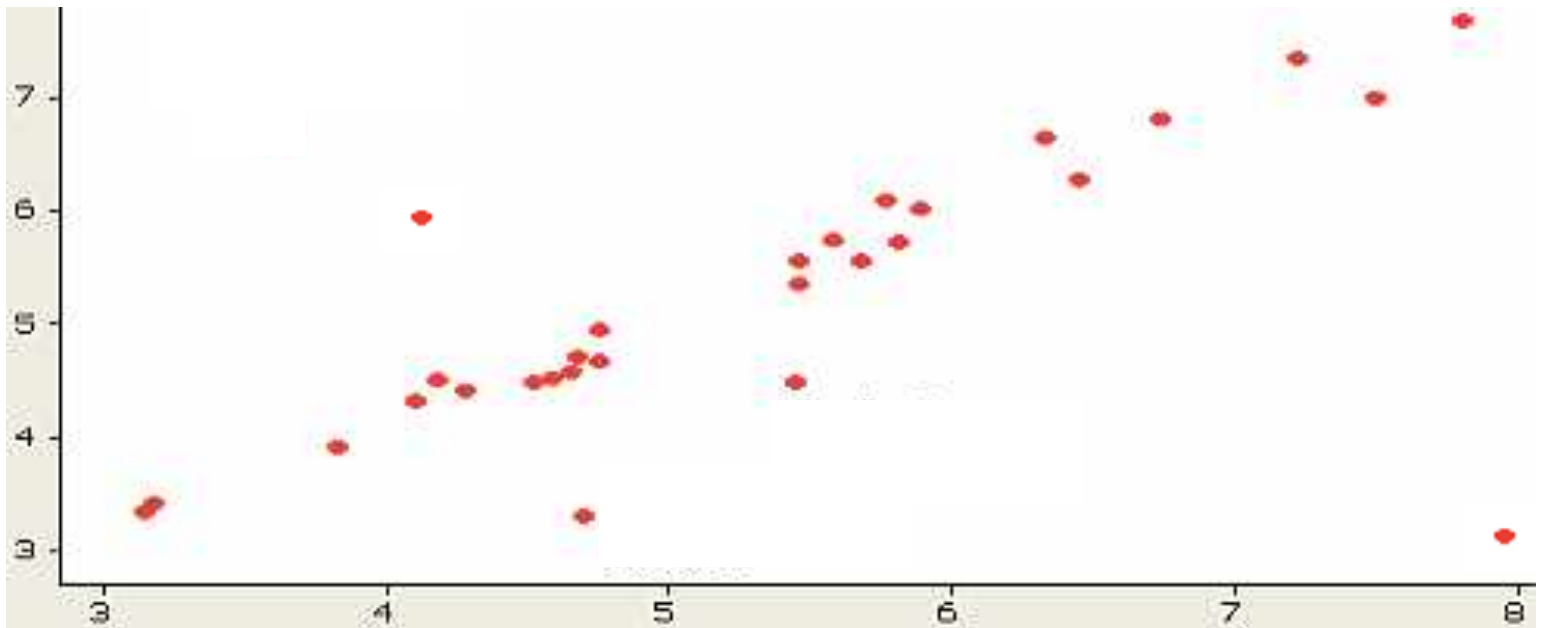

Holly Vickery
@SkylarkHolly
Working with goats 😕🤦‍♀️
8:21 AM · Nov 16, 2021 · Twitter for iPhone

Why do we visualize data?

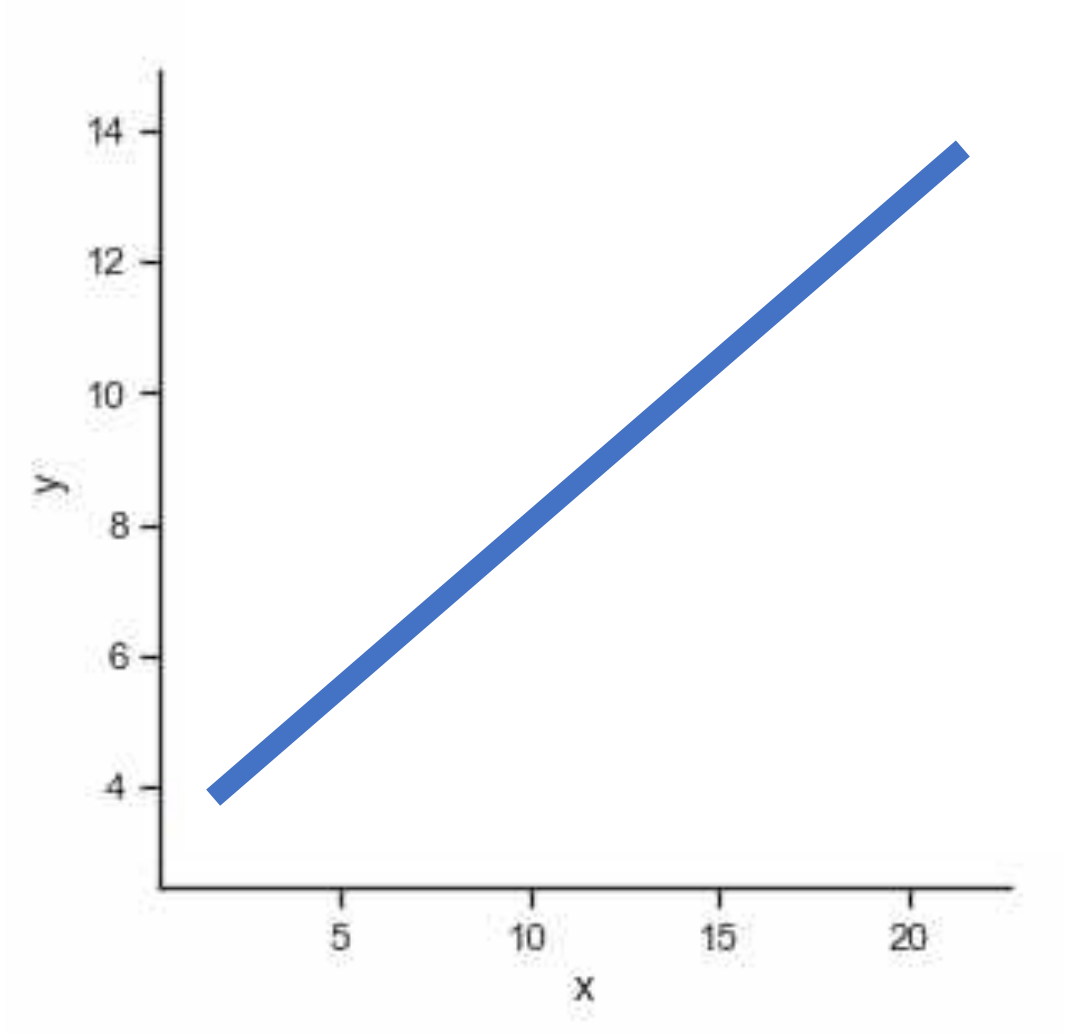
1) To check our dataset for errors and better understand our results.

2) To better communicate those results.

1) To check our dataset for errors…


7
6
5
4
3
3
4
5
6
7
8

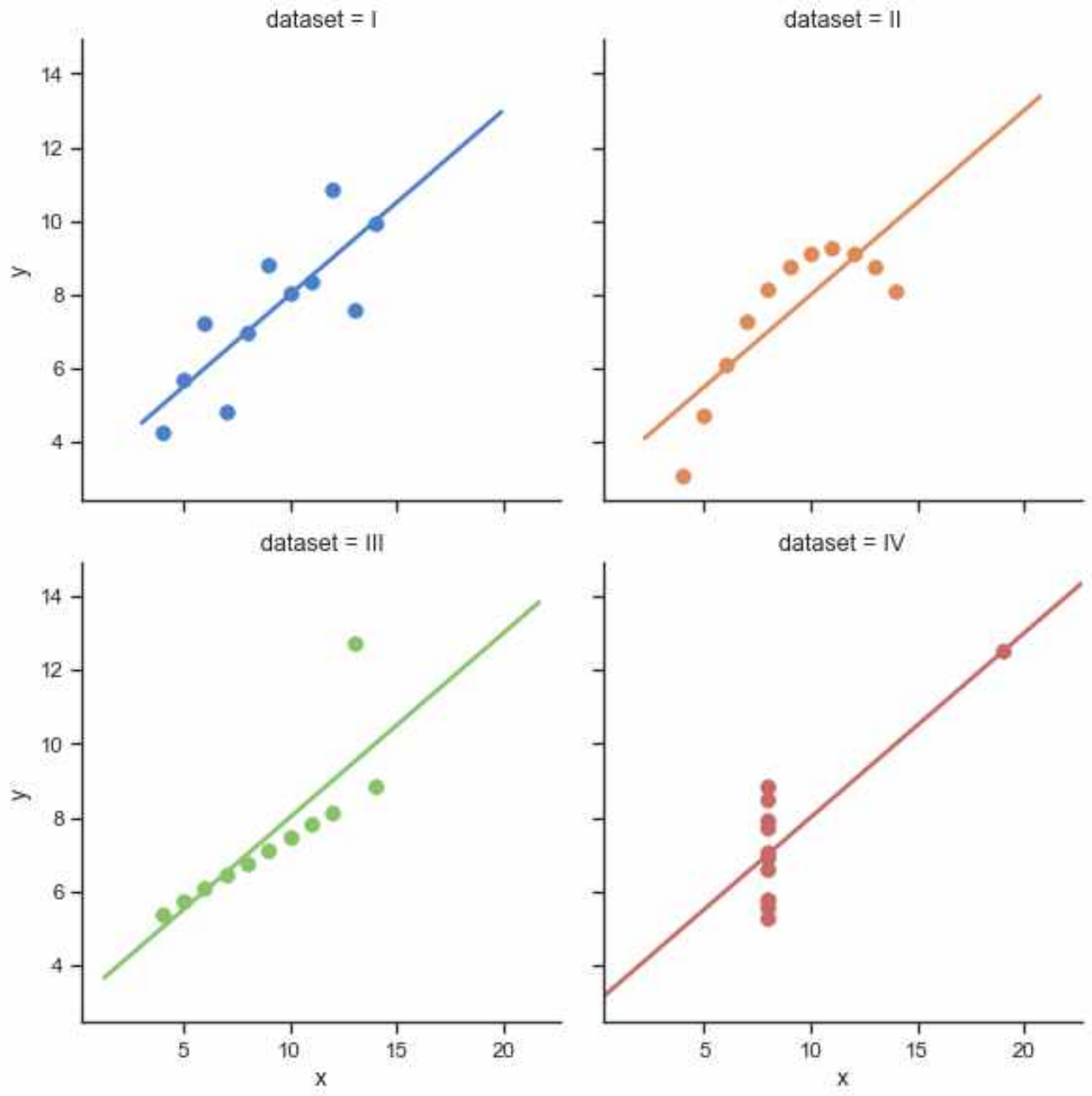
…and better understand our results.



Anscombe's Quartet

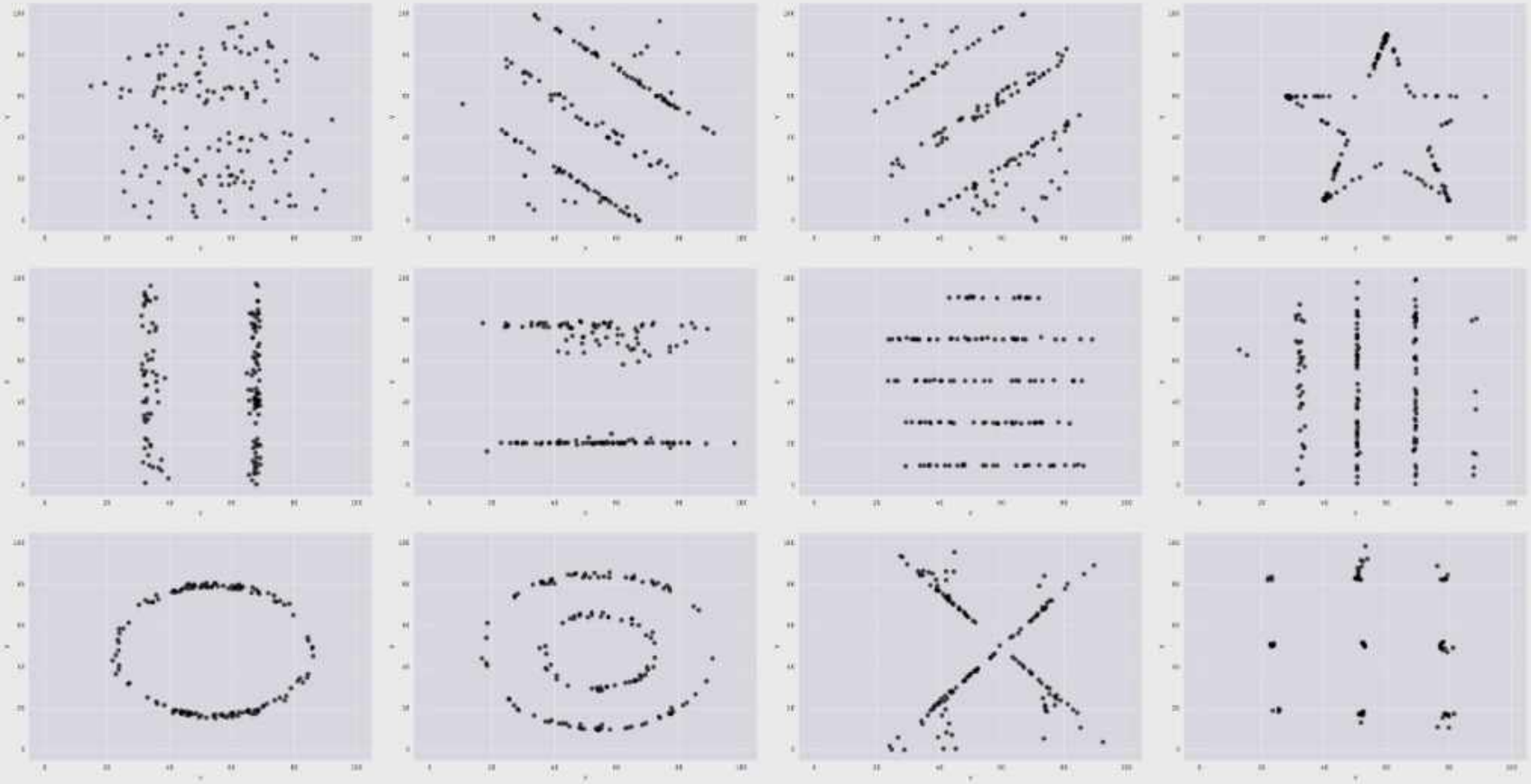
All four of these datasets have:

Mean = 9
Variance = 10
$r^2 = 0.81$



Alberto Cairo's Datasaurus

X Mean	54.26
Y Mean	47.83
X SD	16.76
Y SD	26.93
Corr.	-0.06

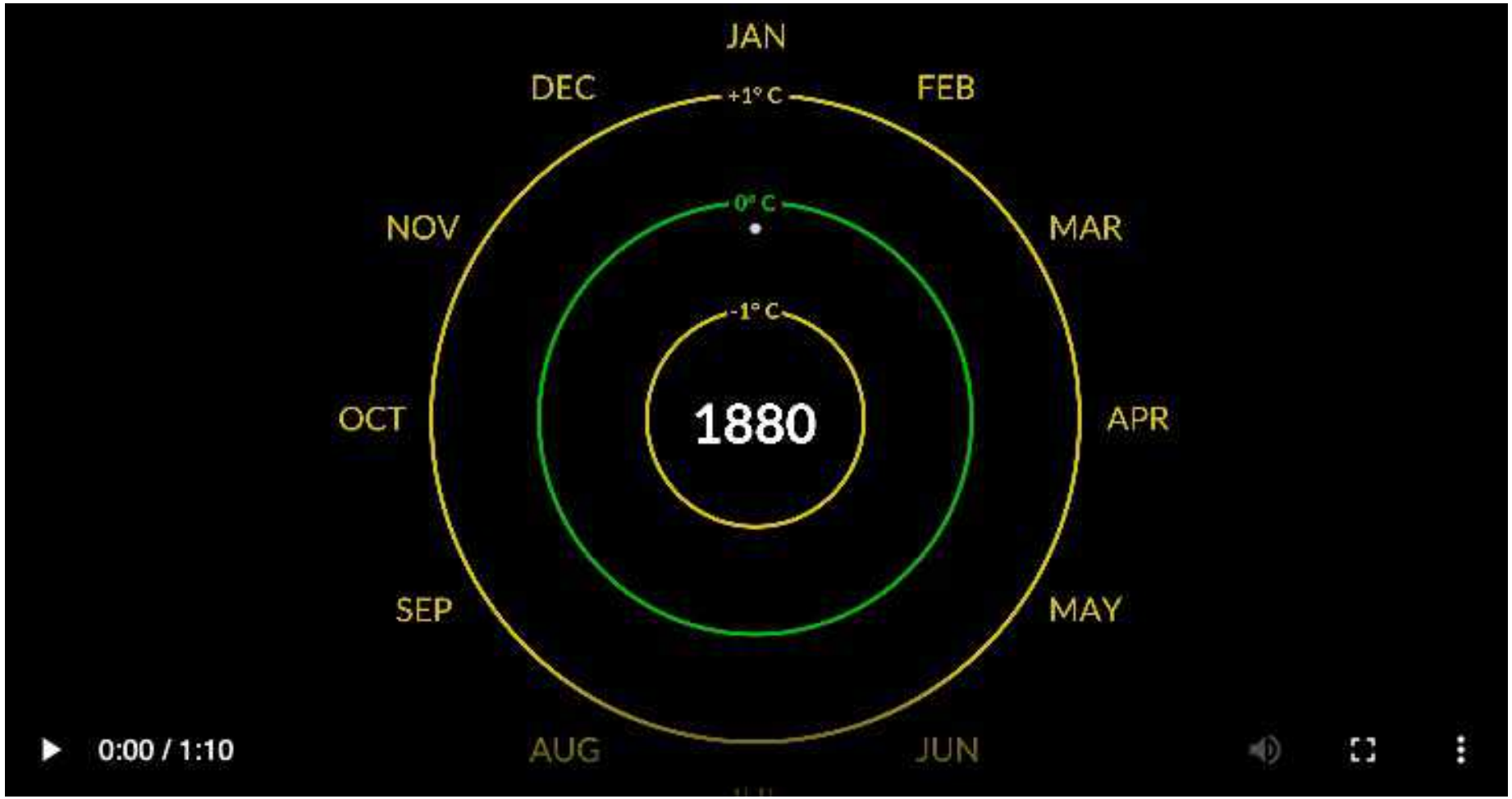


Always Plot Your Data

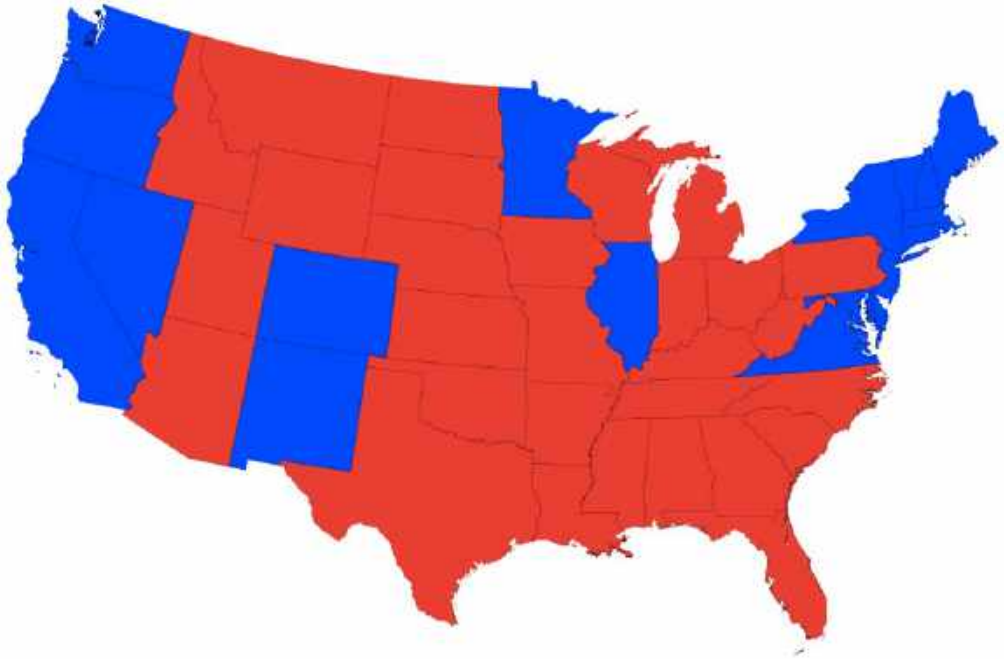
2) To better communicate those results.

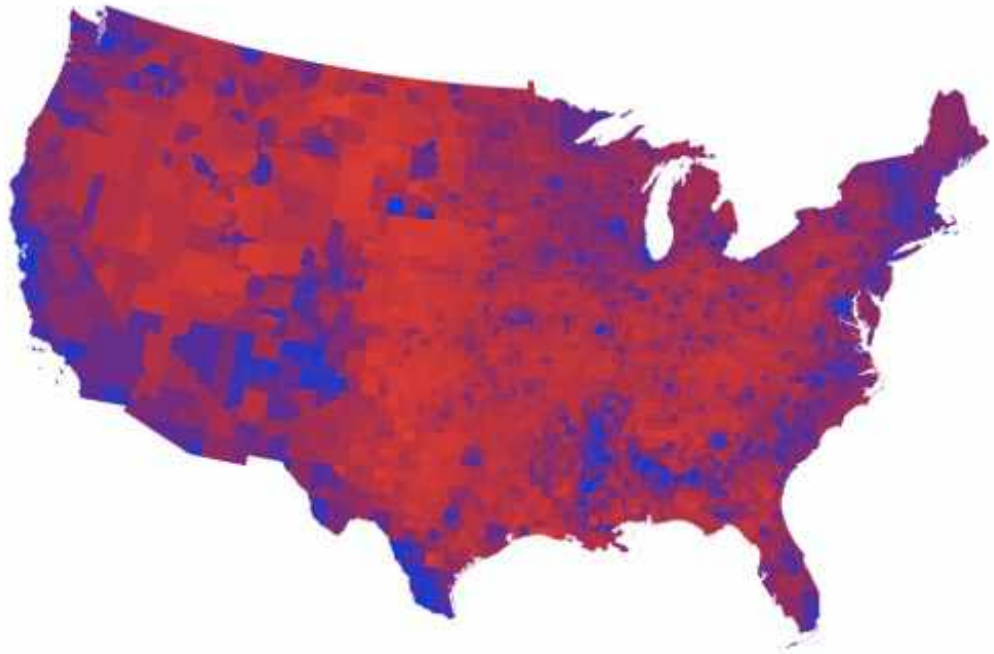
Trite but true:

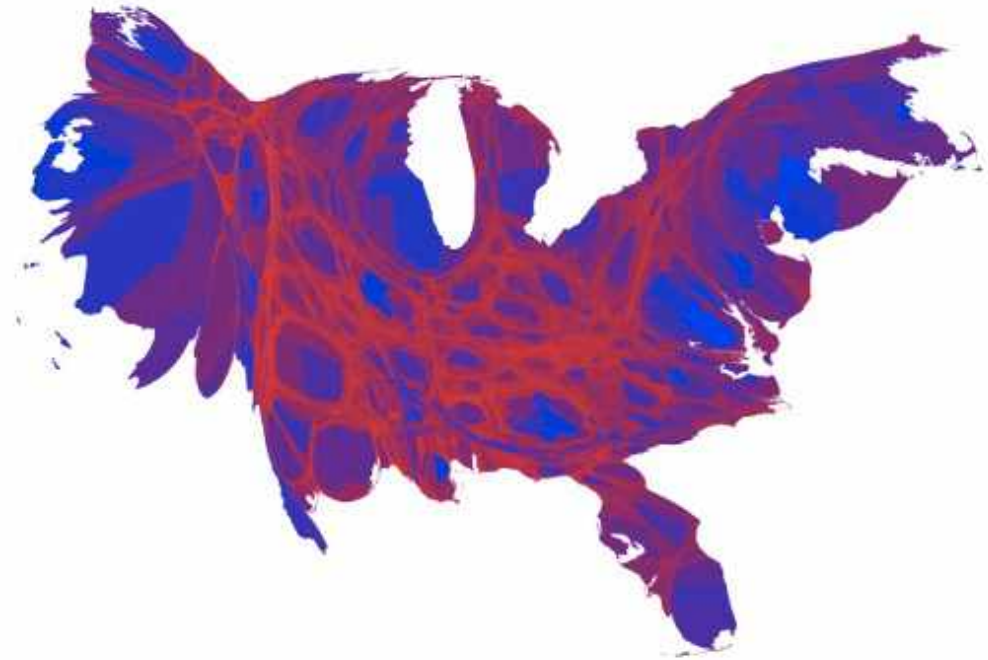
A picture is worth a thousand words

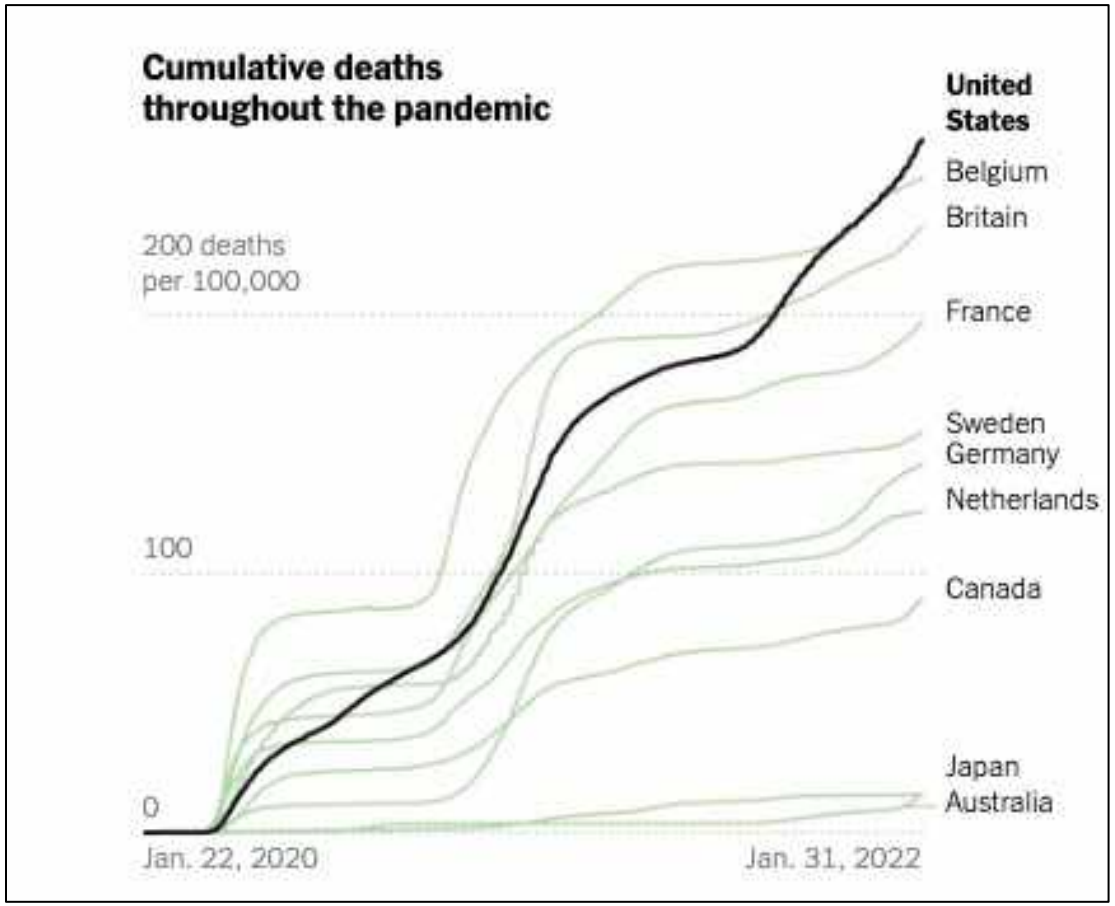


[Nasa Climate Spiral]

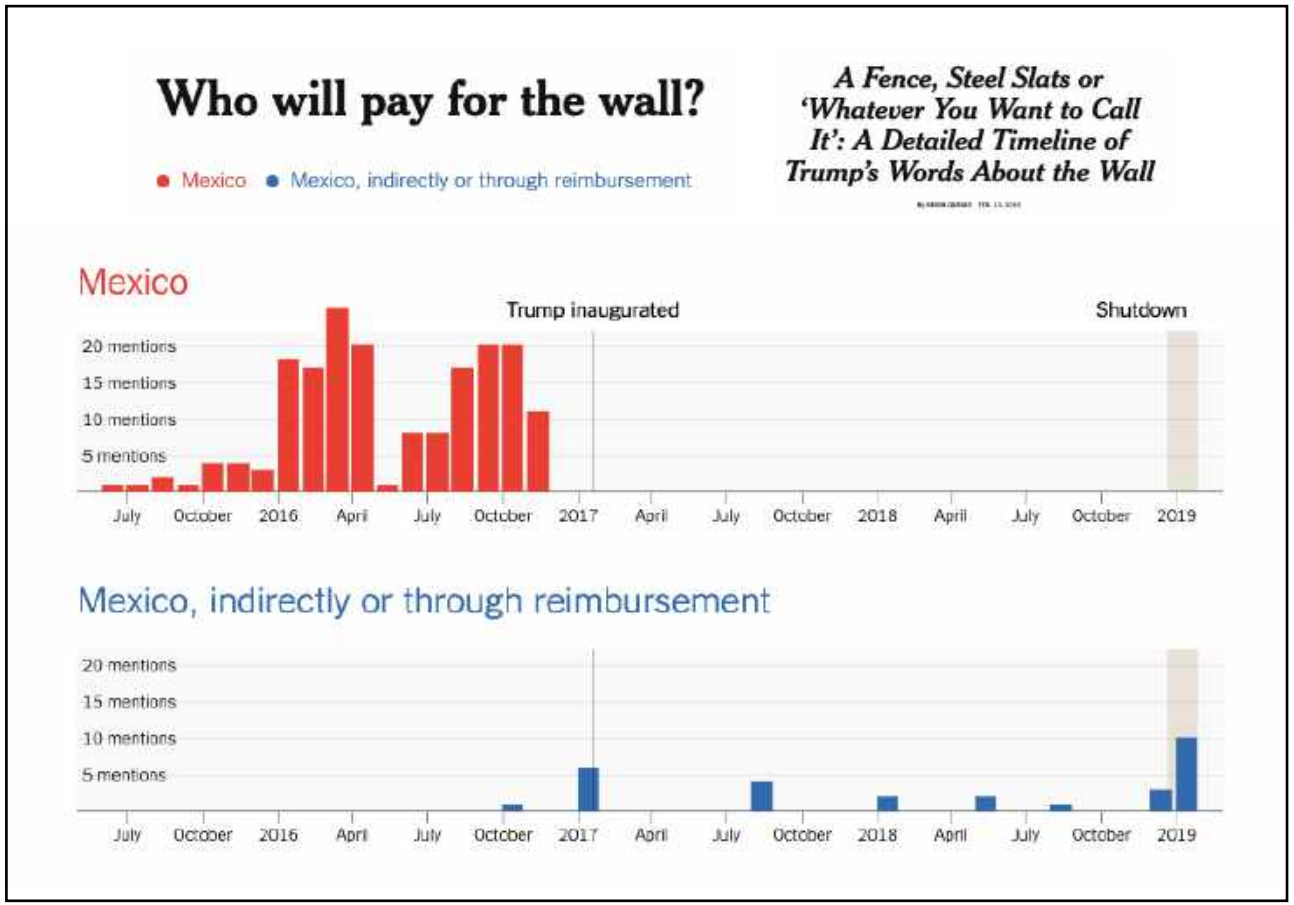






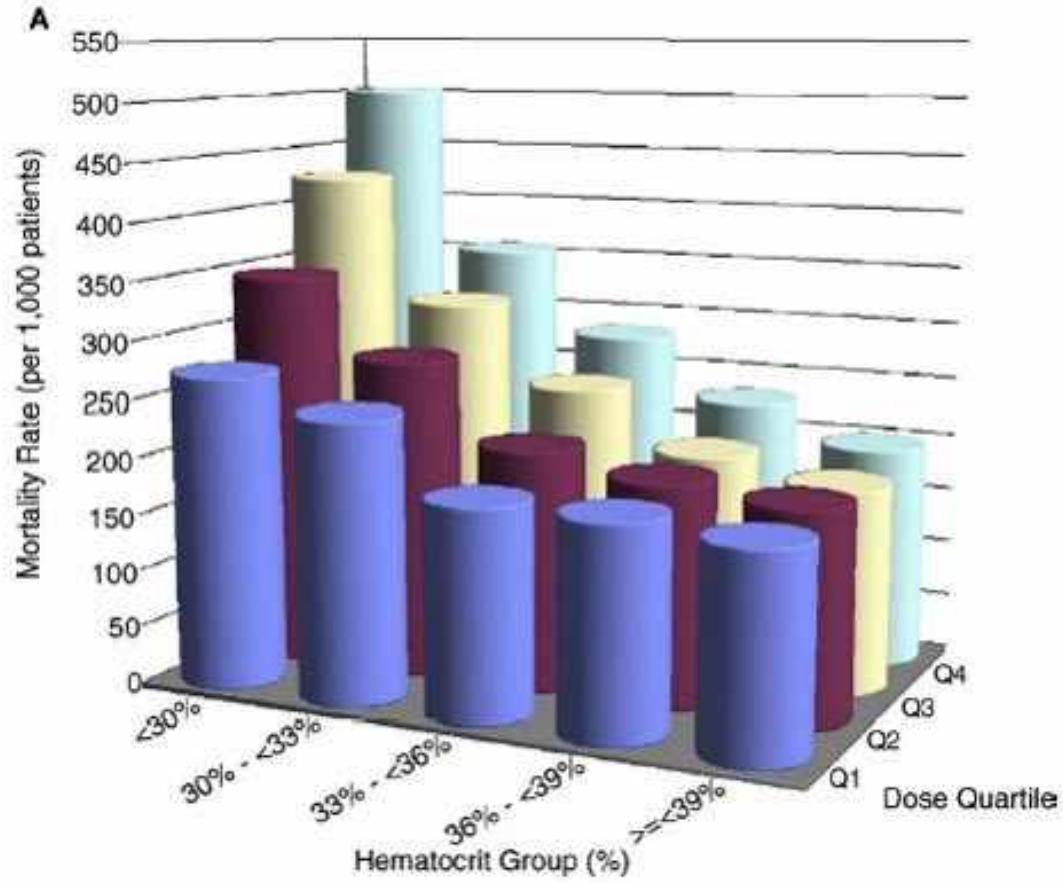


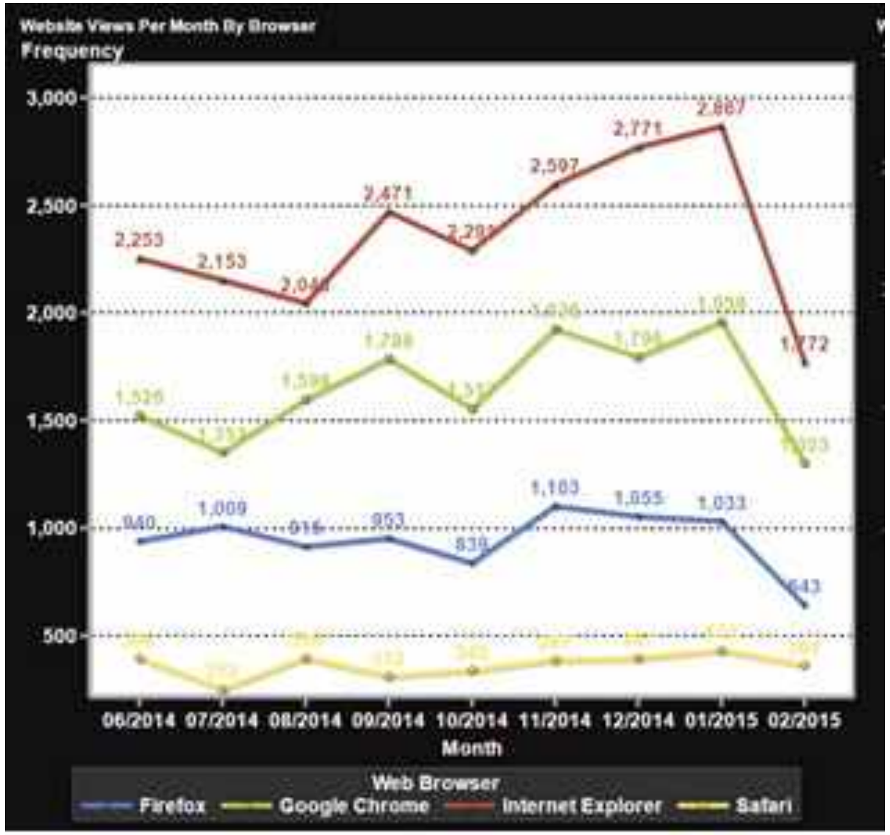
NY Times Covid Dashboard

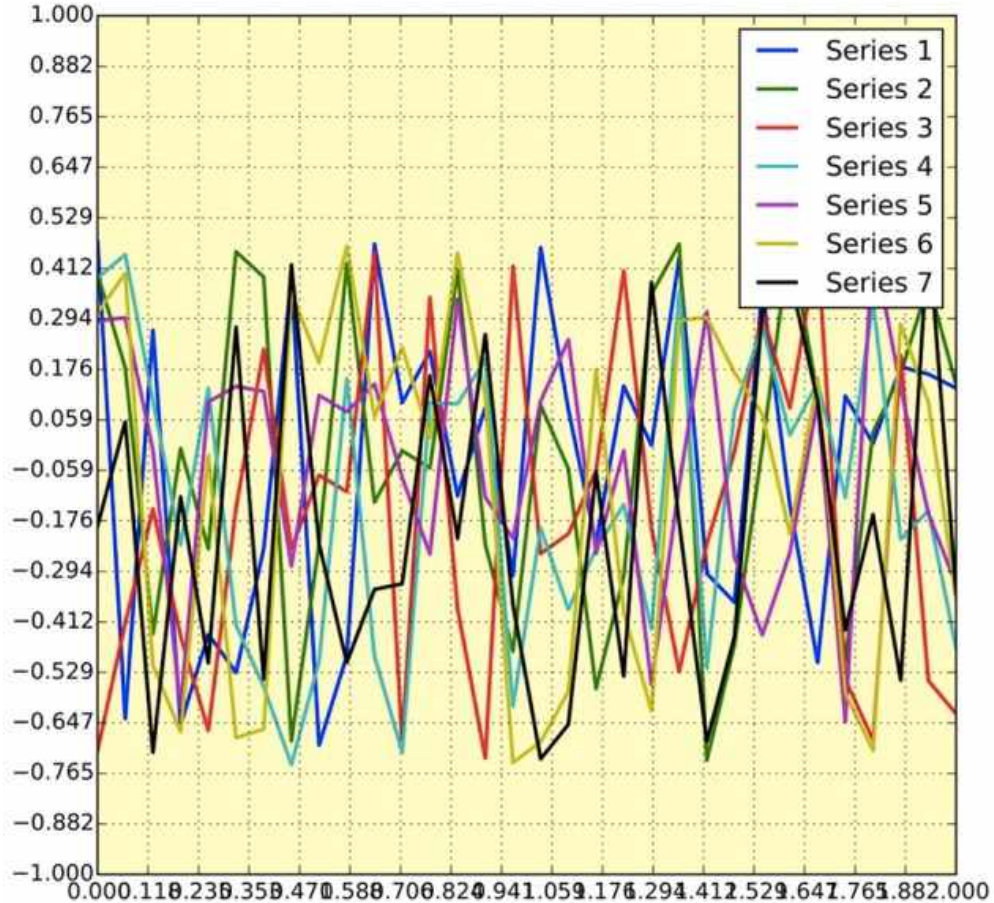


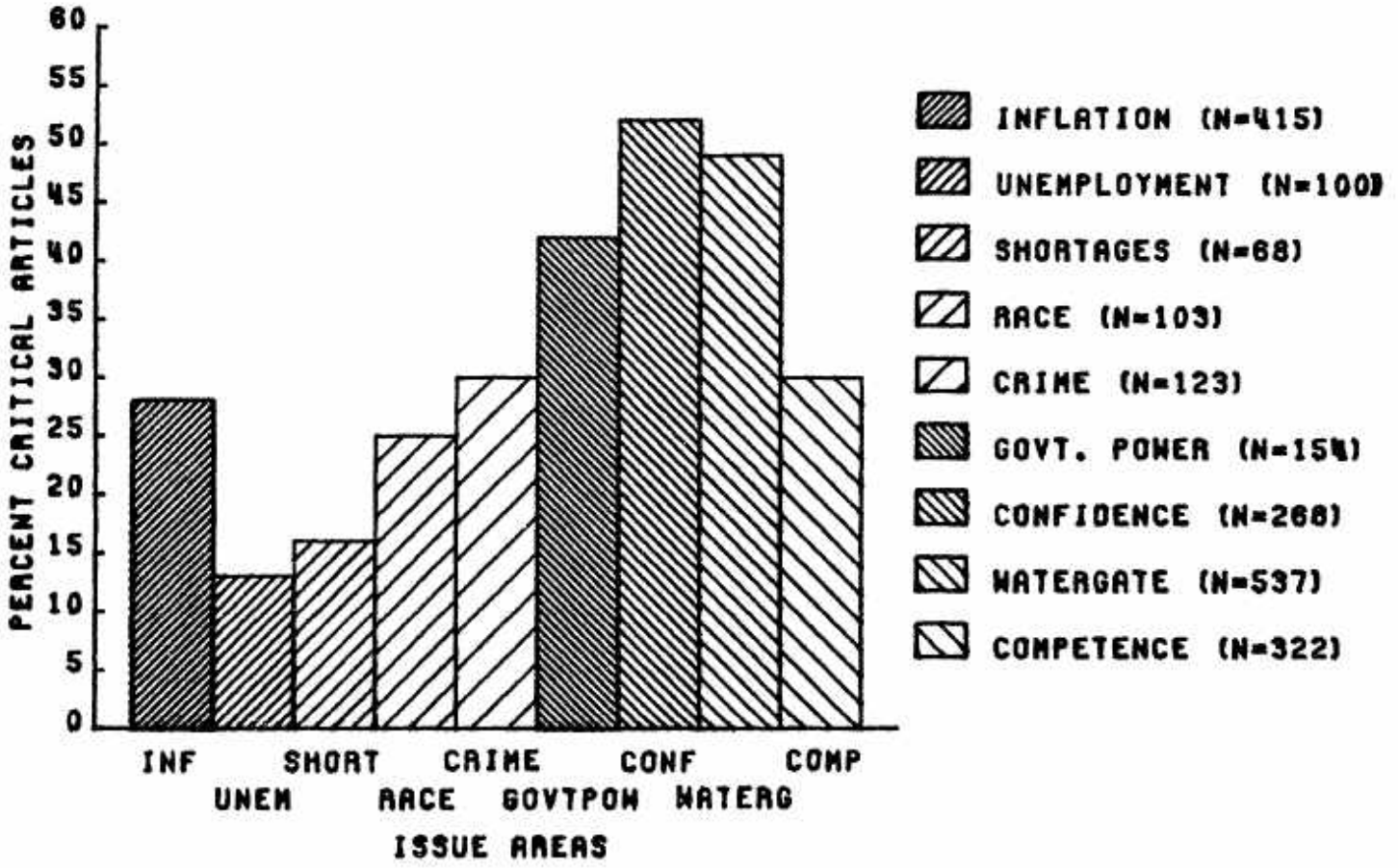
NY Times Election Coverage

We need to talk.


A
Mortality Rate (per 1,000 patients)
550
500
450
400
350
300
250
200
150
100
50
0
<30%
30% - <33%
33% - <36%
36% - <39%
>=<39%
Hematocrit Group (%)
Q4
Q3
Q2
Q1
Dose Quartile


Website Views Per Month By Browser
Frequency
3,000
2,500
2,000
1,500
1,000
500
06/2014
07/2014
08/2014
09/2014
10/2014
11/2014
12/2014
01/2015
02/2015
Month
Web Browser
Firefox
Google Chrome
Internet Explorer
Safari


1.000
0.882
0.765
0.647
0.529
0.412
0.294
0.176
0.059
−0.059
−0.176
−0.294
−0.412
−0.529
−0.647
−0.765
−0.882
−1.000
0.000
1.882
2.000
Series 1
Series 2
Series 3
Series 4
Series 5
Series 6
Series 7



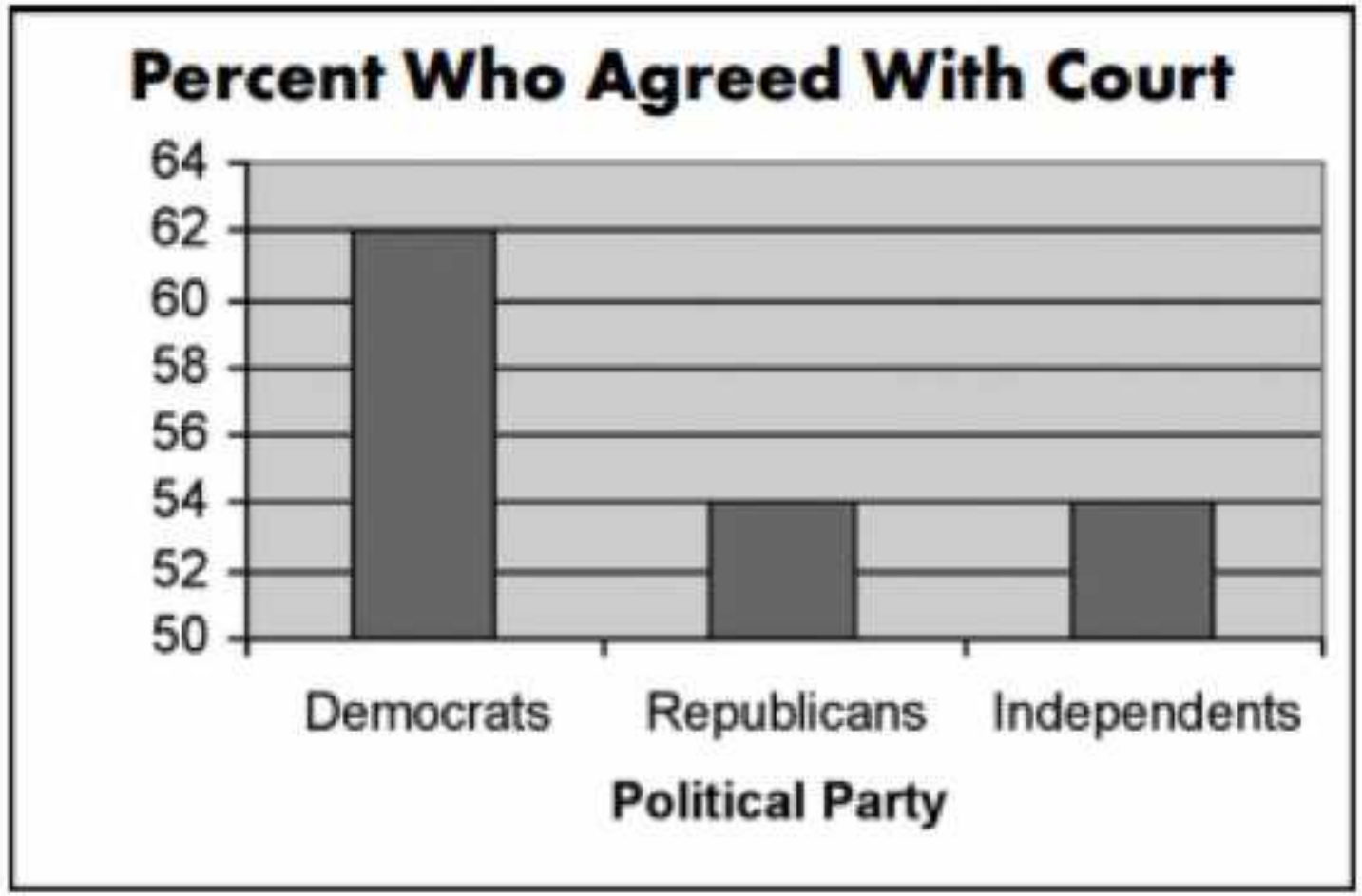
Source: Center for Political Studies Media Content Analysis Study, 1974; available through the University of Michigan, ICPSR. Not to be cited without full bibliographical reference to the present article.

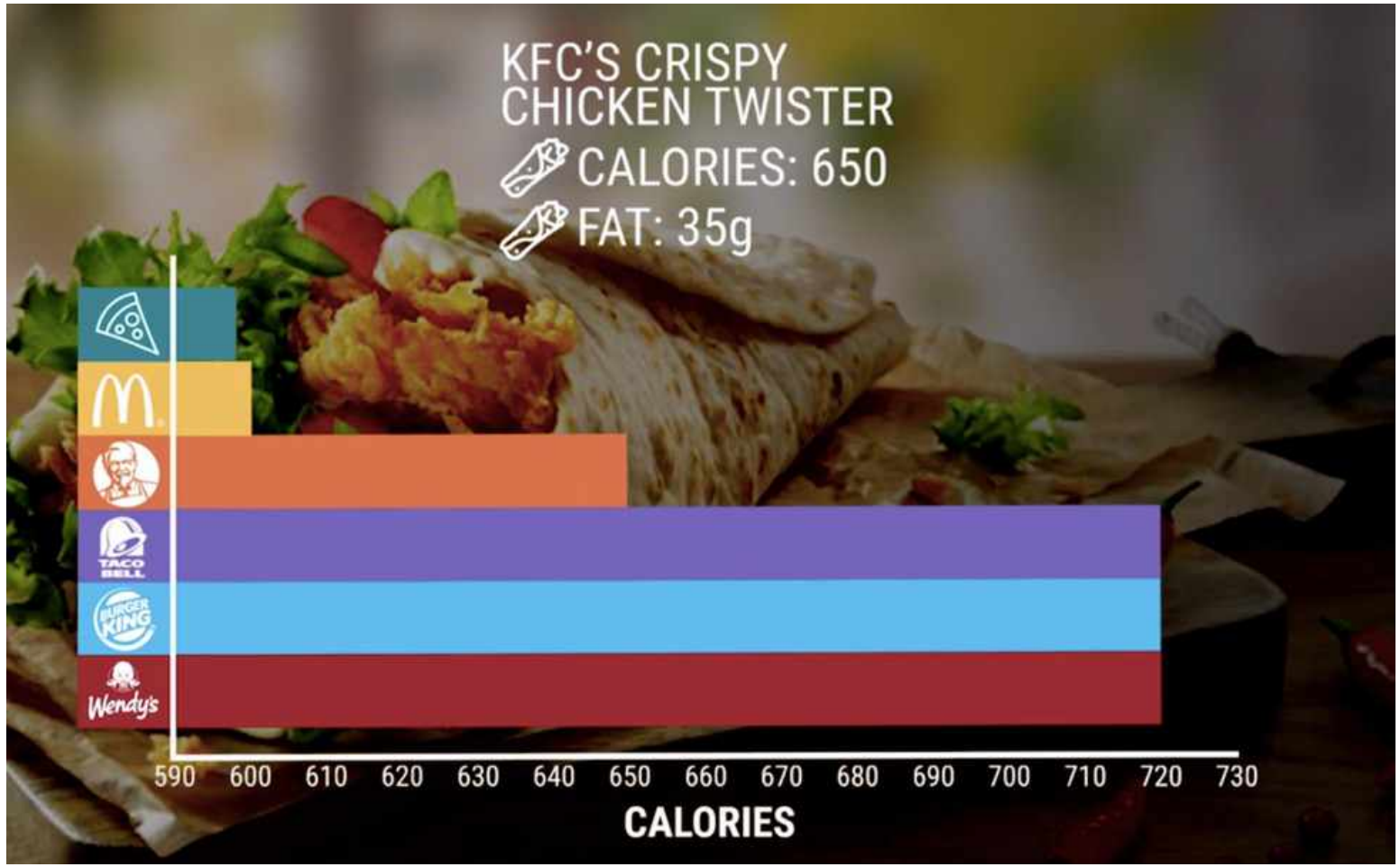
Tufte calls this “Chart Junk”

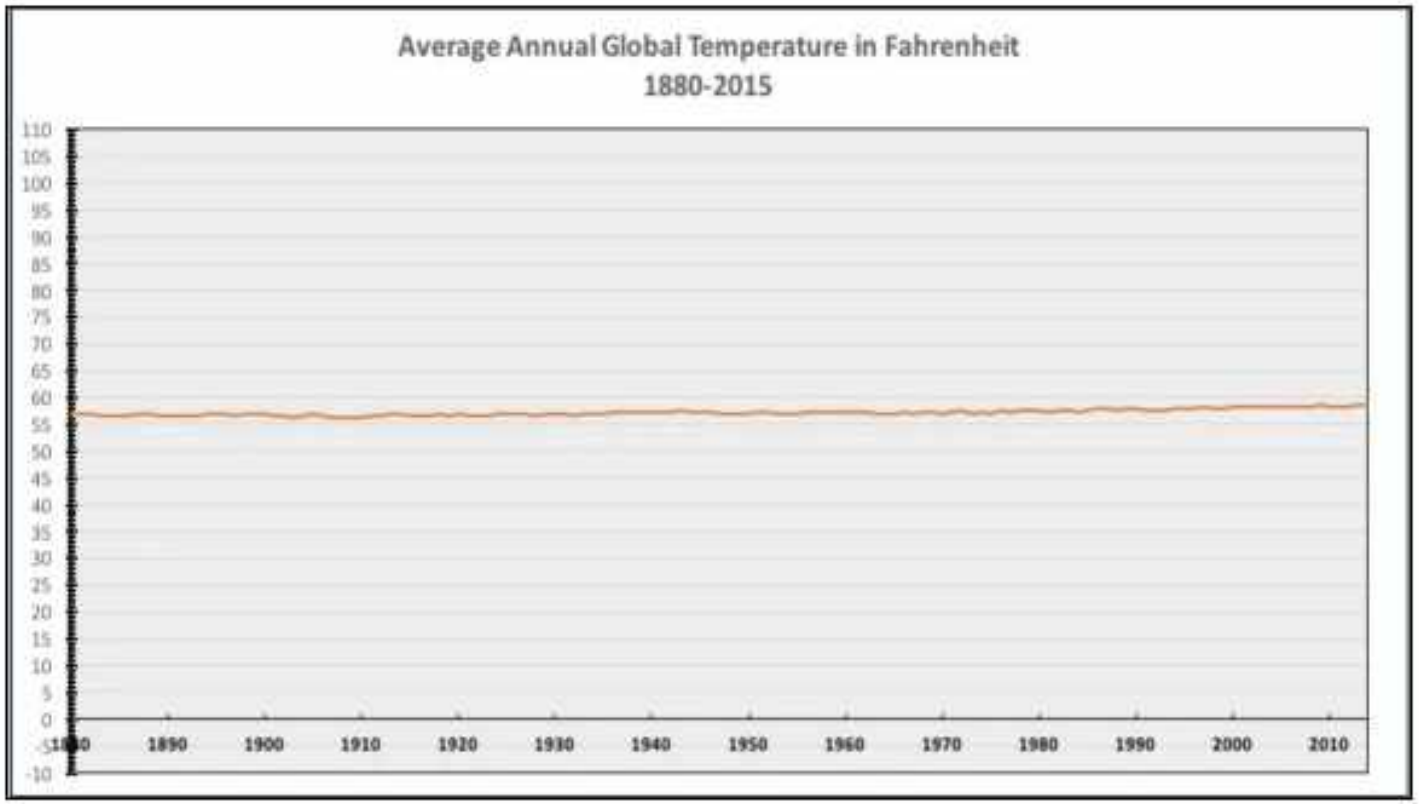
Be aware of how your visualization choices influence your ability to validate, interpret, and communicate your results.

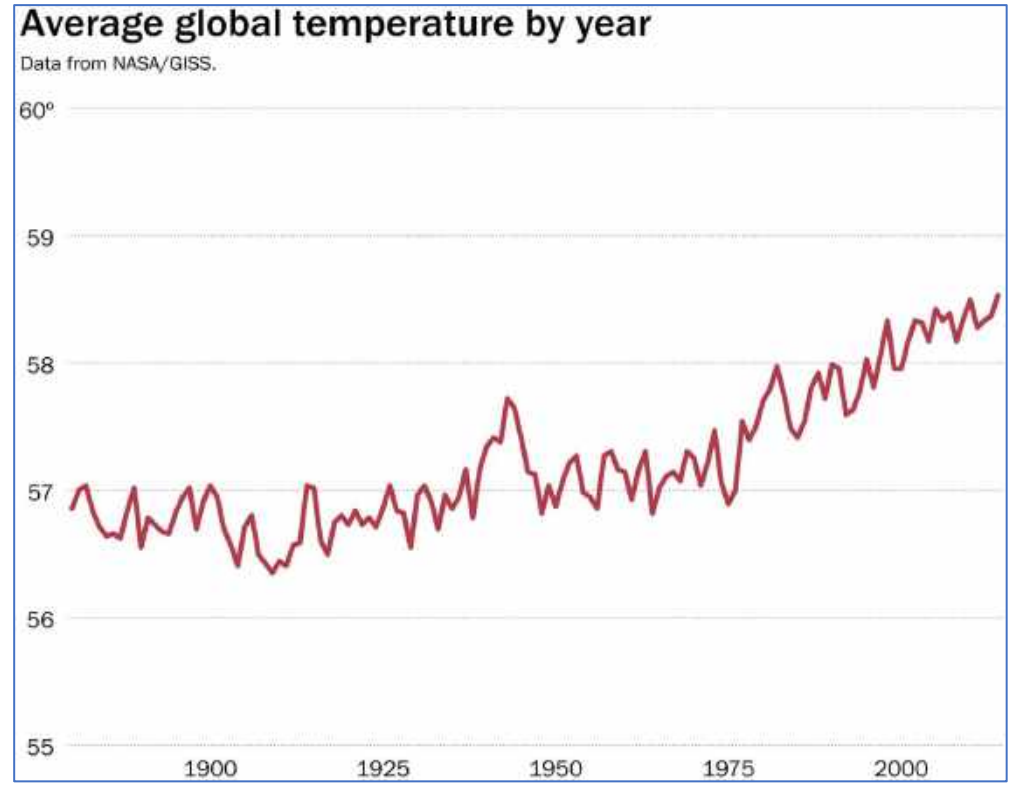
Be aware of how decisions (or mistakes) made by others can lead you to draw incorrect conclusions.

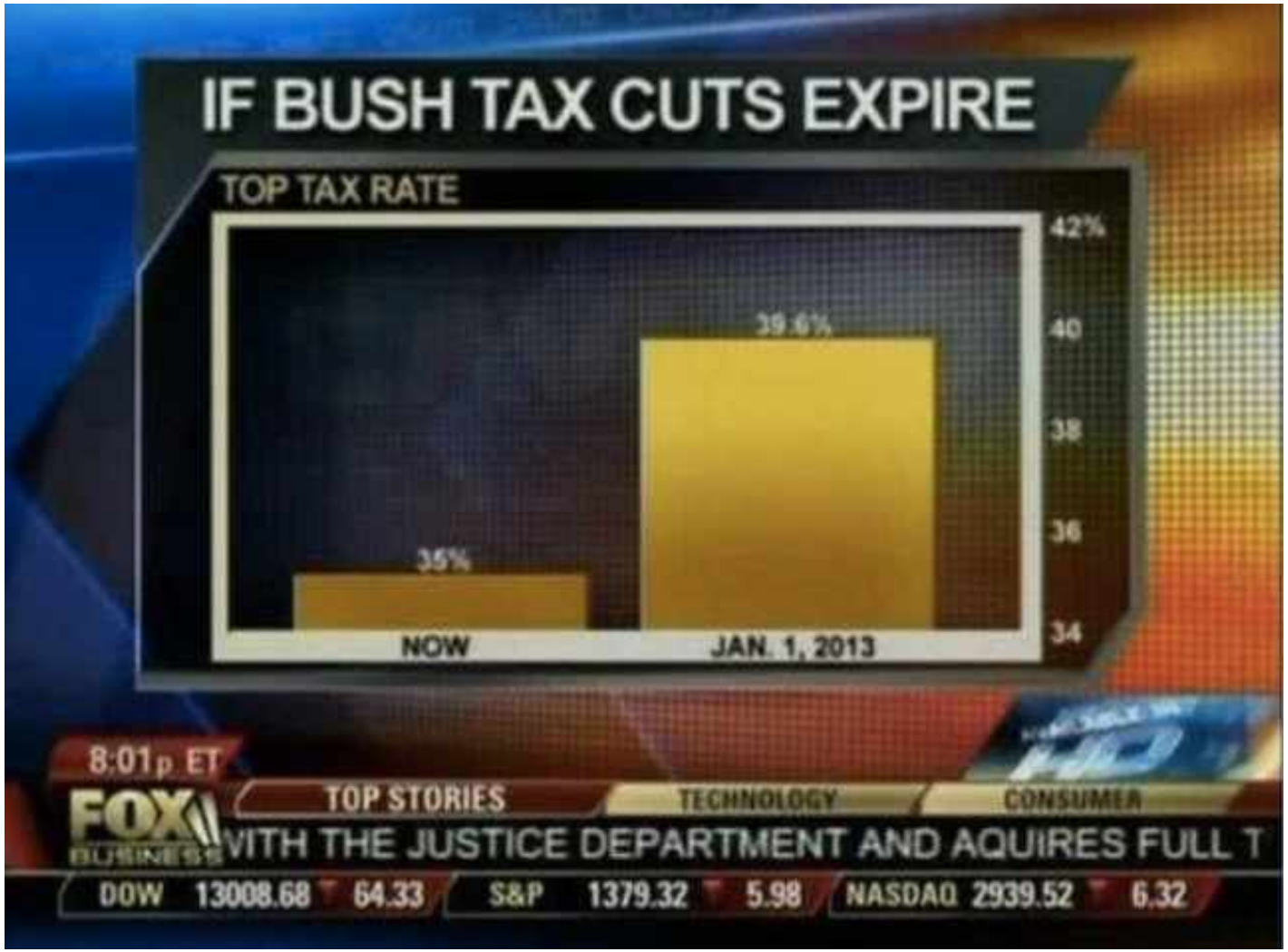
Be aware.

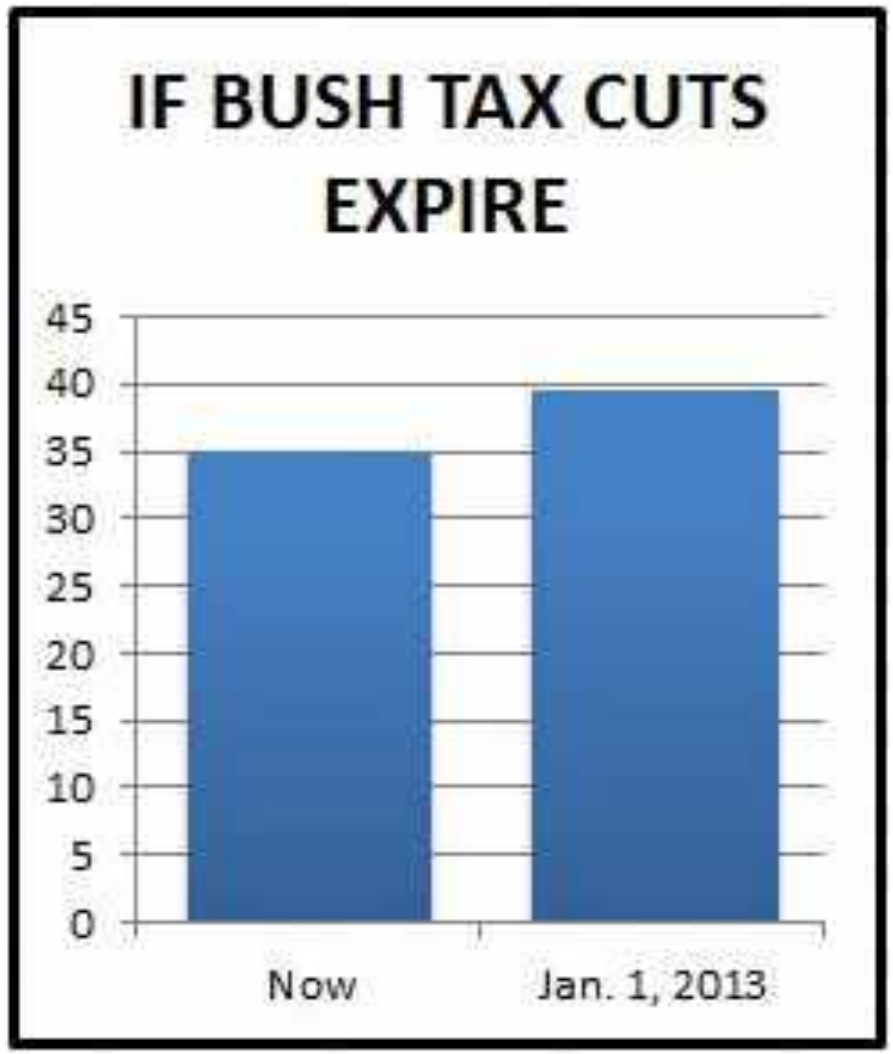

Percent Who Agreed With Court
64
62
60
58
56
54
52
50
Democrats
Republicans
Independents
Political Party

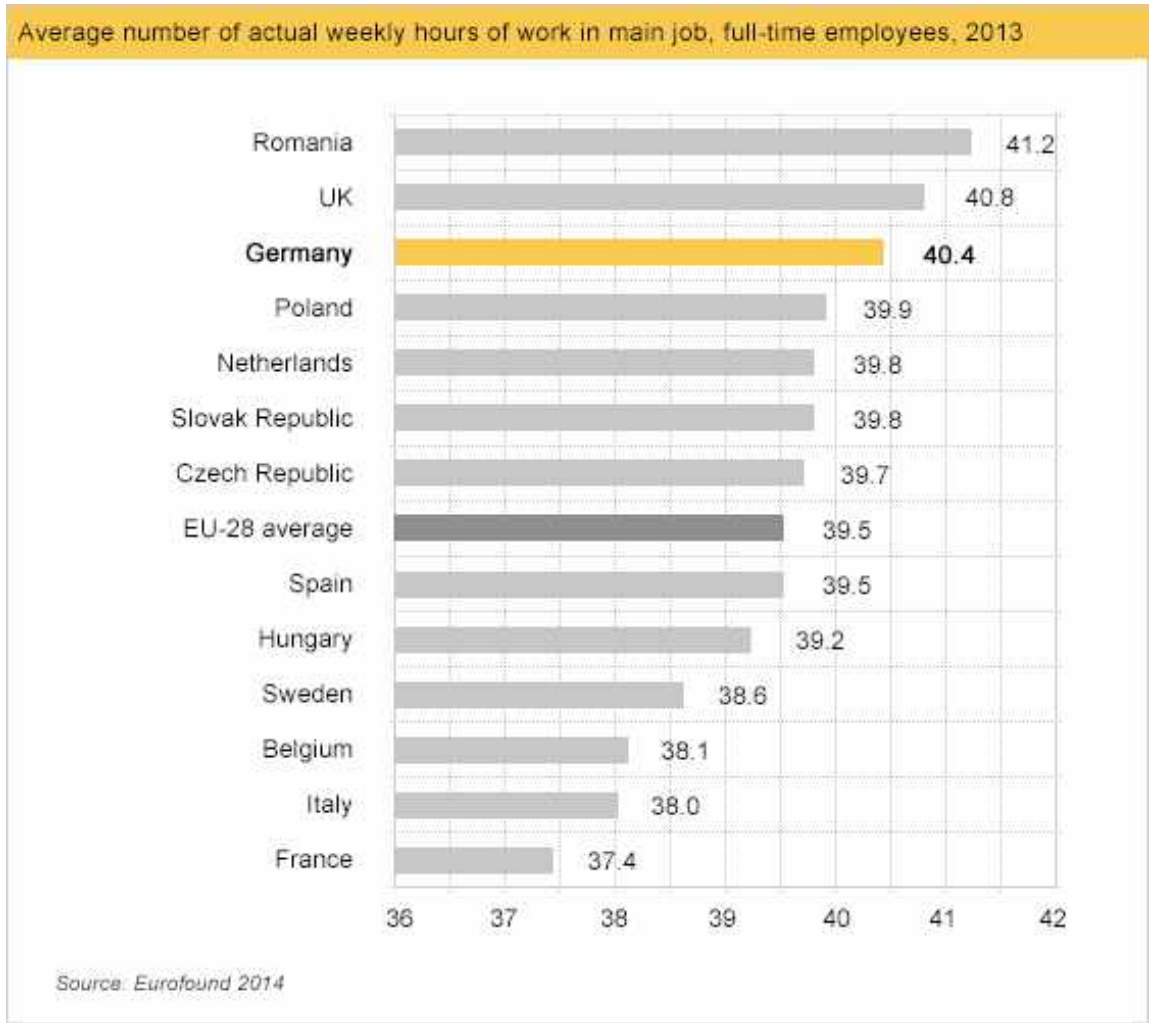

KFC'S CRISPY
CHICKEN TWISTER
CALORIES: 650
FAT: 35g
TACO BELL
BURGER KING
Wendy's
590 600 610 620 630 640 650 660 670 680 690 700 710 720 730
CALORIES

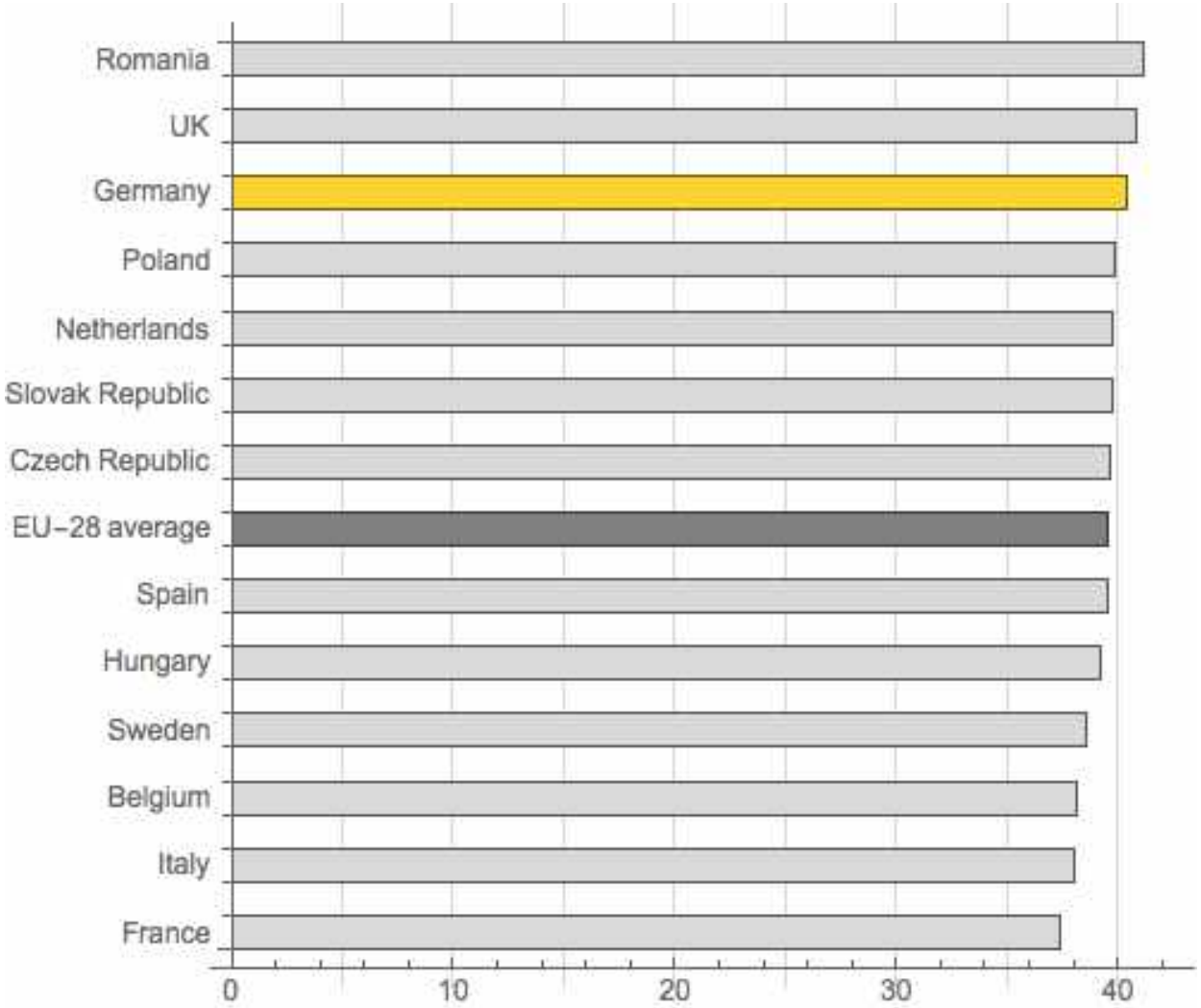

Average Annual Global Temperature in Fahrenheit
1880-2015
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
-5
-10
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010


Average global temperature by year
Data from NASA/GISS.
60°
59
58
57
56
55
1900
1925
1950
1975
2000

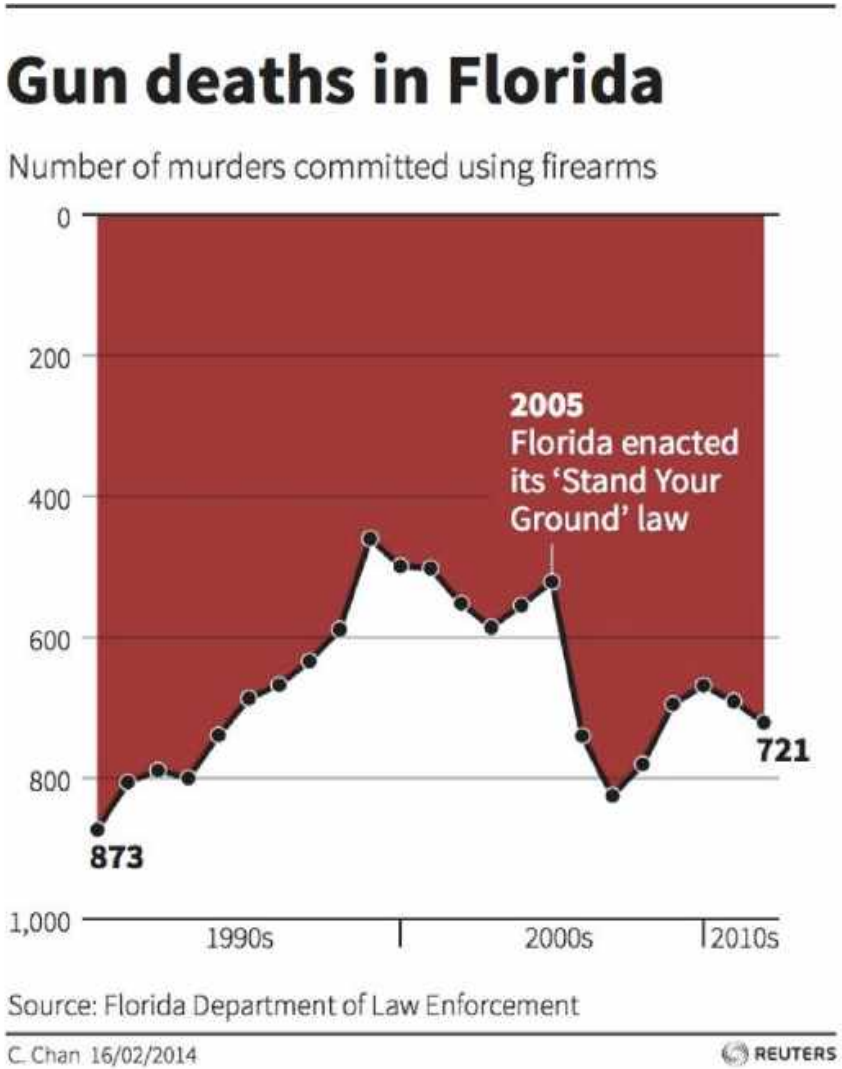

IF BUSH TAX CUTS EXPIRE
TOP TAX RATE
42%
40
38
36
34
39.6%
35%
NOW
JAN. 1, 2013
8:01p ET
FOX BUSINESS
TOP STORIES
TECHNOLOGY
CONSUMER
VITH THE JUSTICE DEPARTMENT AND AQUIRES FULL T
DOW 13008.68 64.33
S&P 1379.32 5.98
NASDAQ 2939.52 6.32

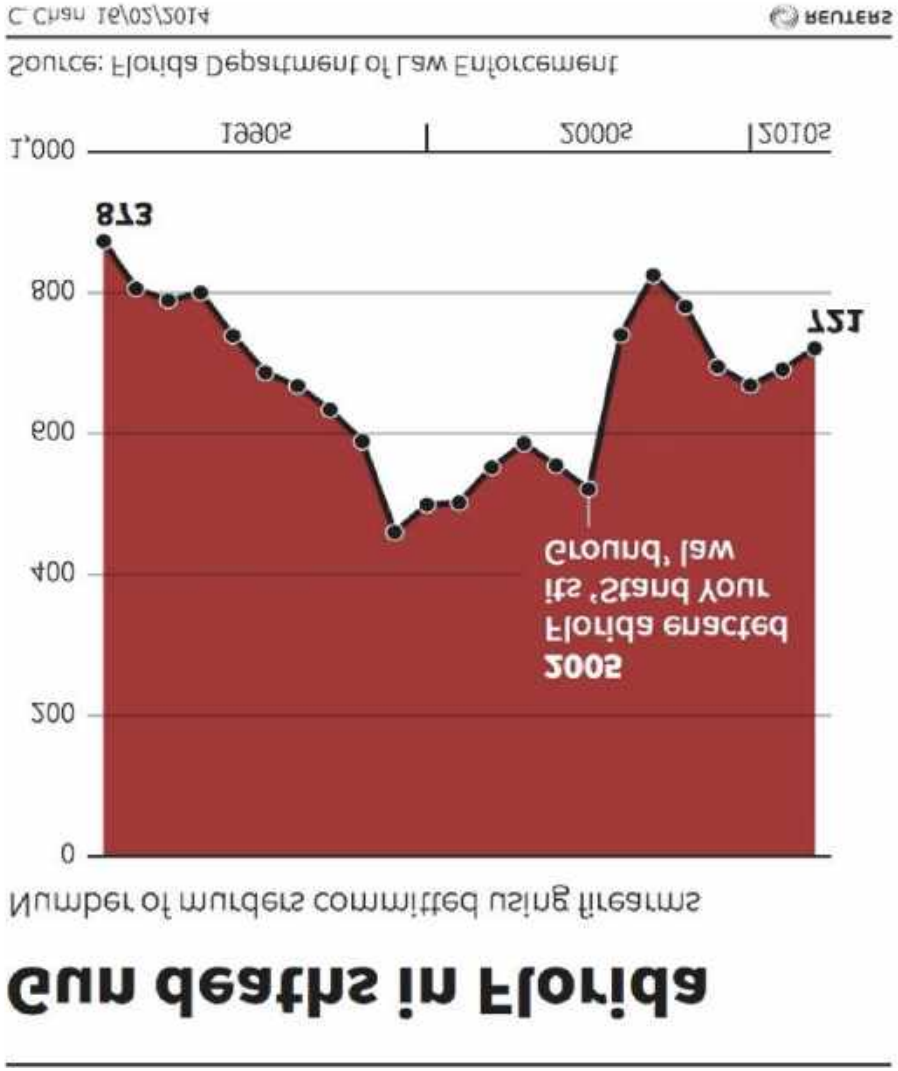

IF BUSH TAX CUTS EXPIRE
45
40
35
30
25
20
15
10
5
0
Now
Jan. 1, 2013



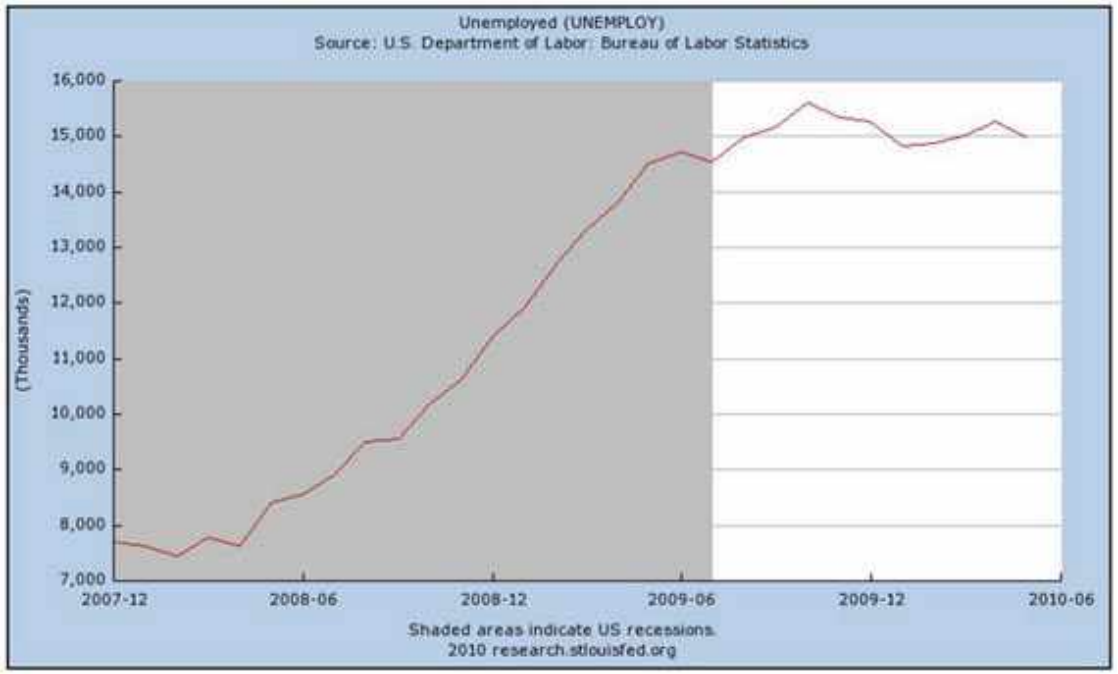


Be aware of the values chosen for axes.

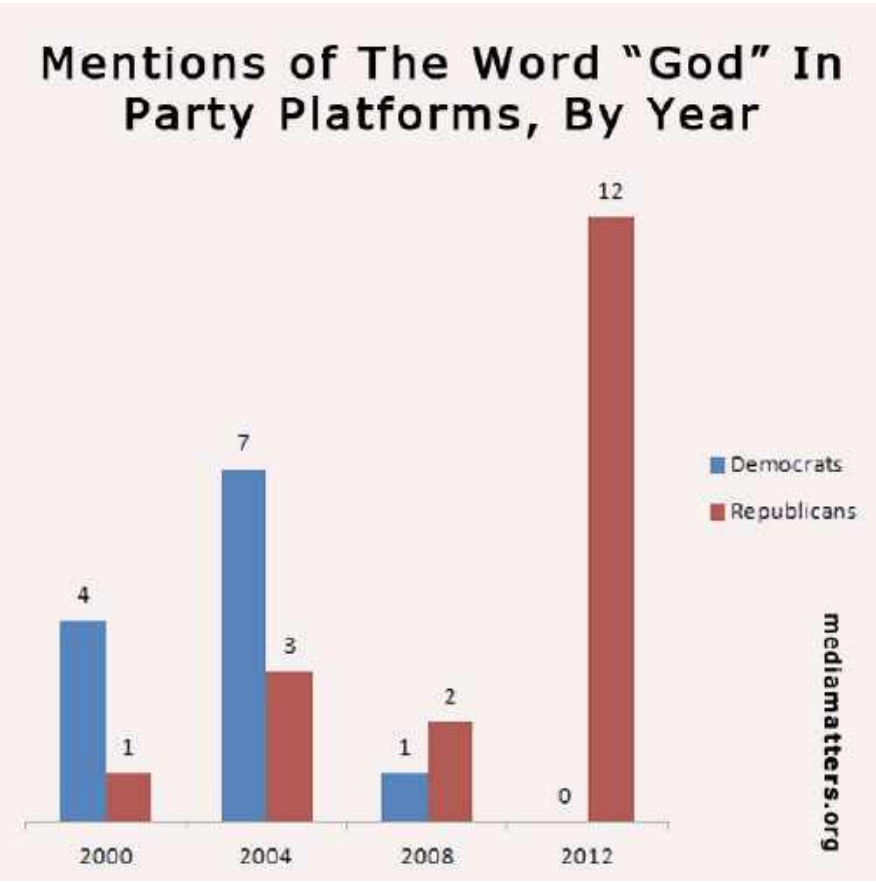




Be aware of how shading can alter interpretation

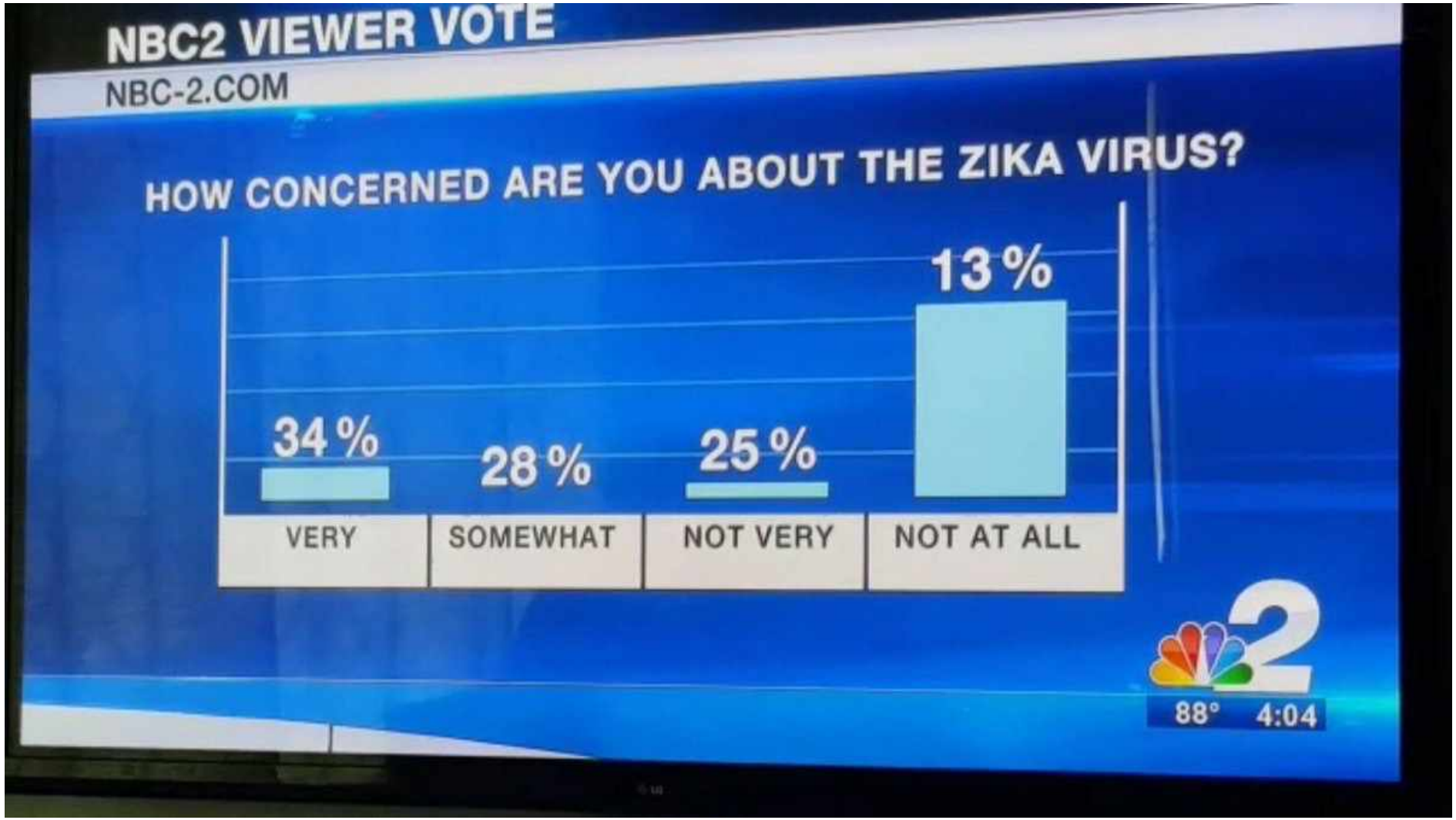


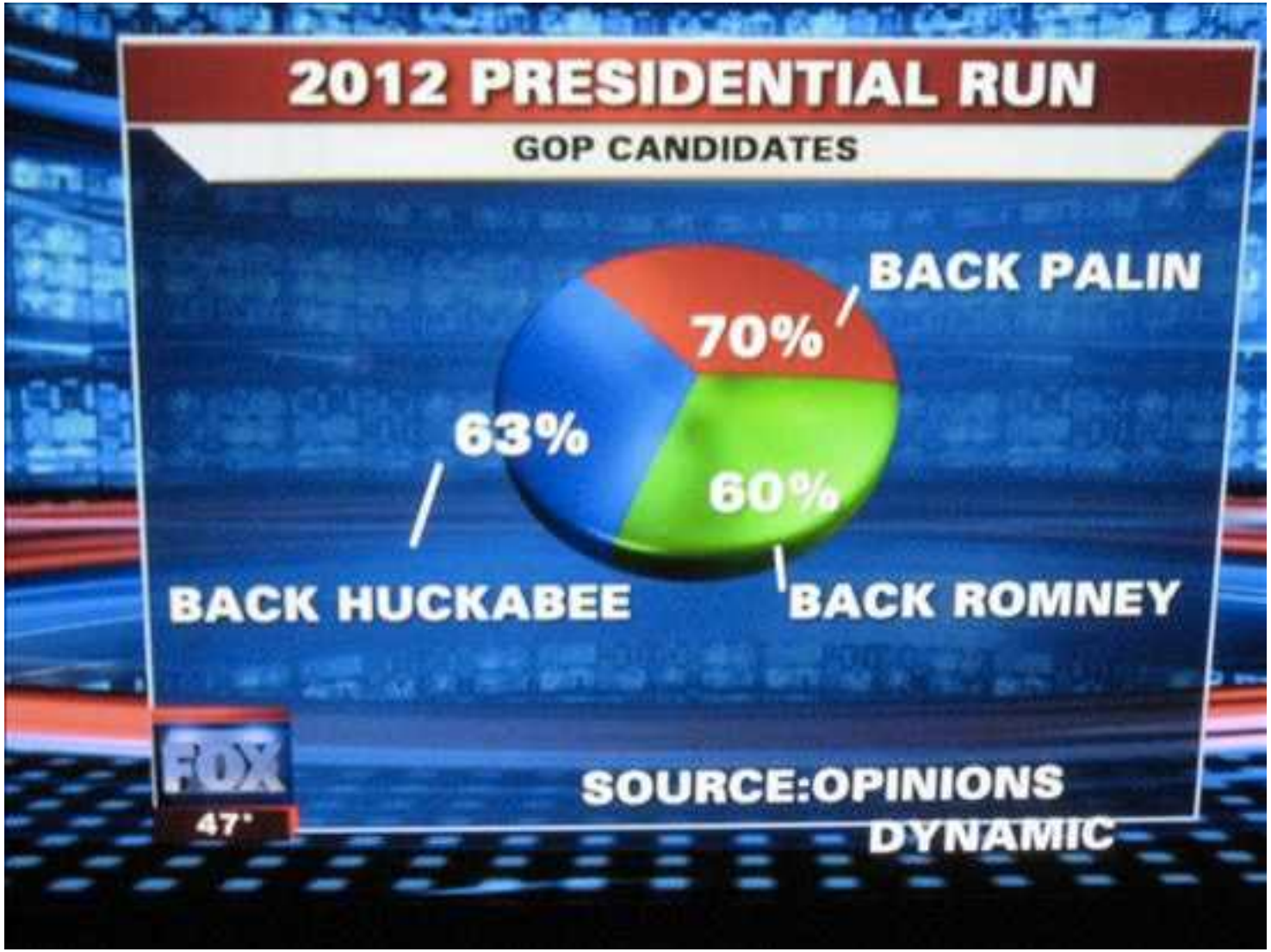
Be aware of truncating datasets



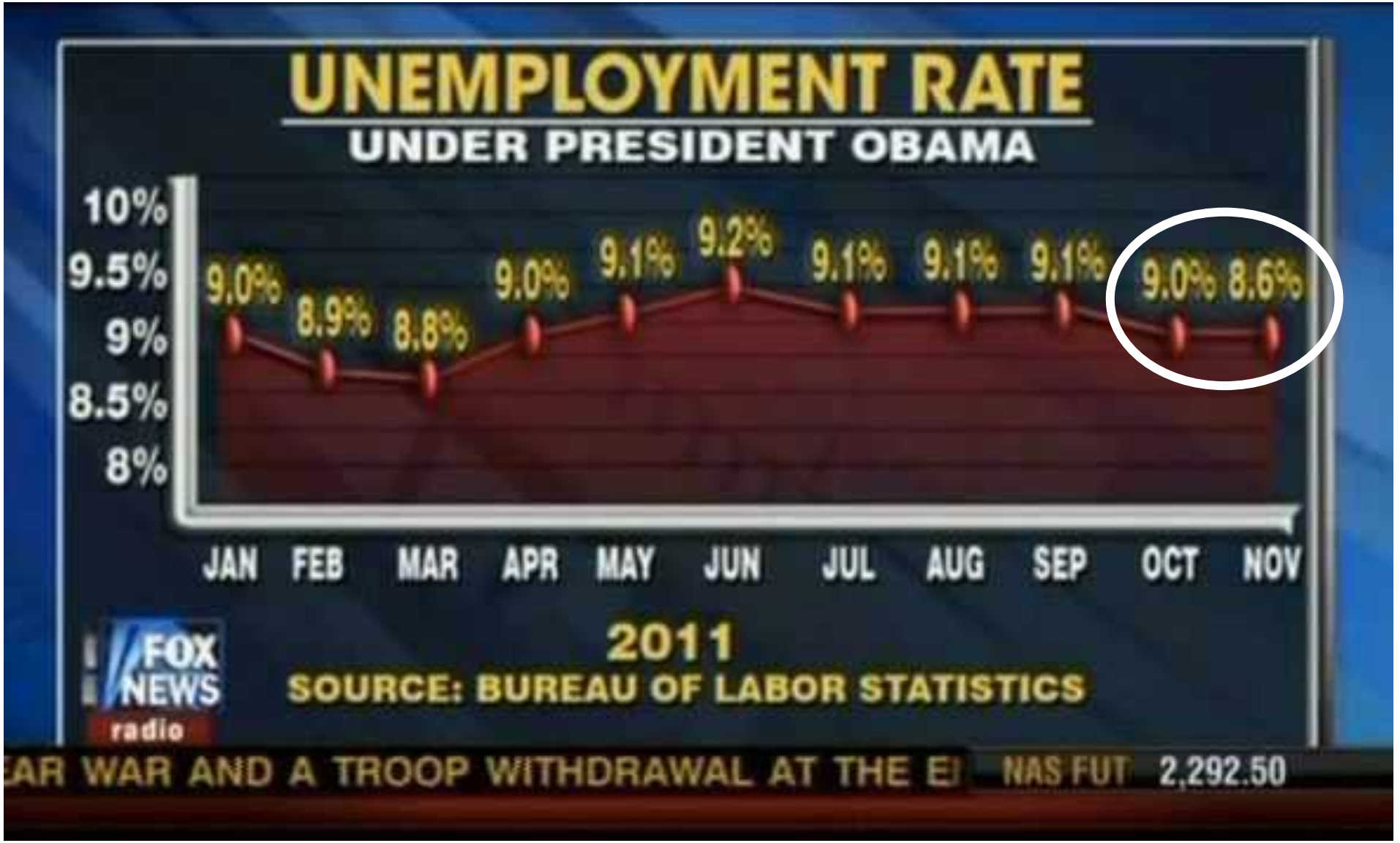
Be aware of the order in which values are presented

Be aware of Legends and Captions.

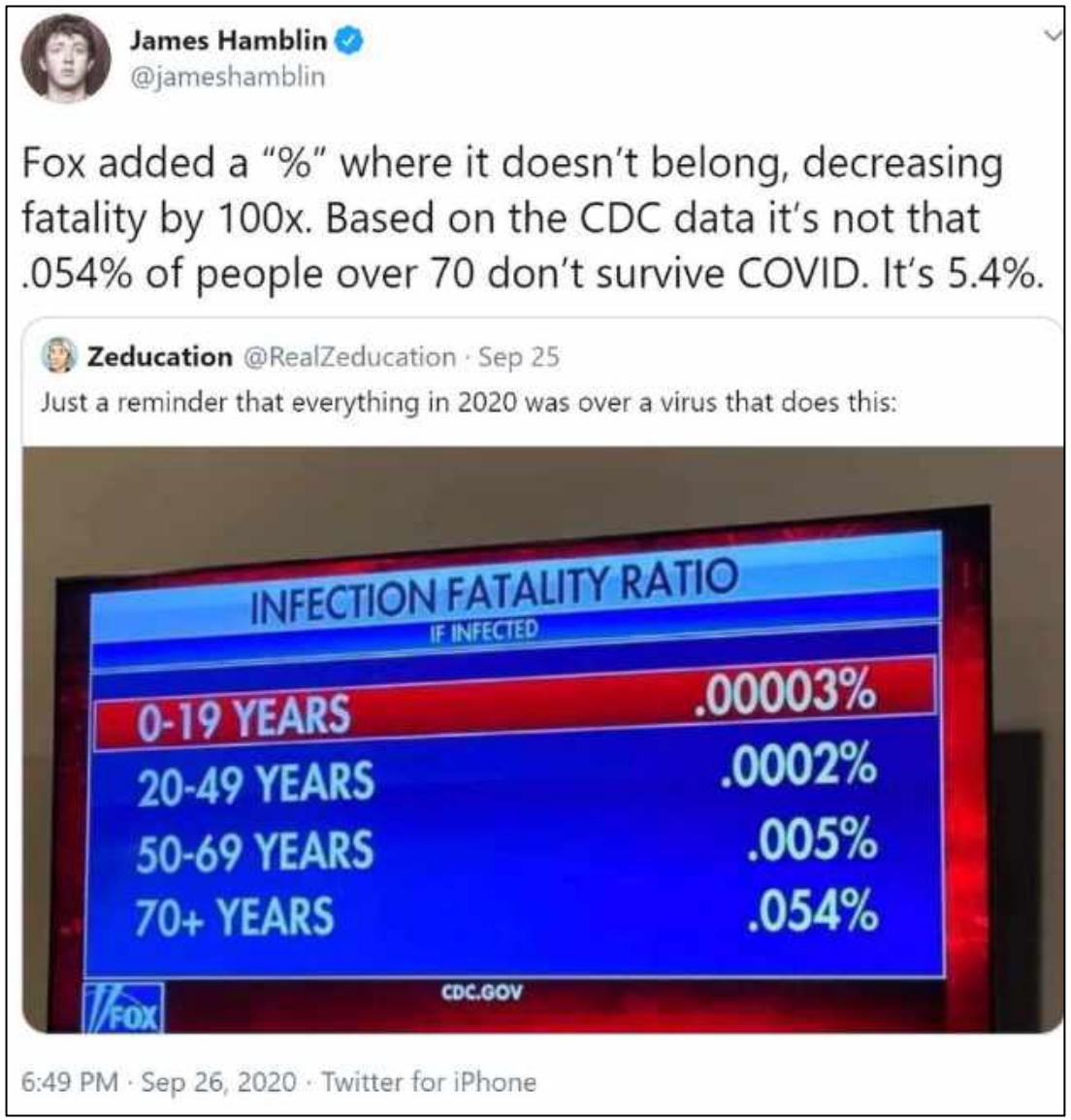

NBC2 VIEWER VOTE
NBC-2.COM
HOW CONCERNED ARE YOU ABOUT THE ZIKA VIRUS?
34 %
28 %
25 %
13 %
VERY
SOMEWHAT
NOT VERY
NOT AT ALL
2
88°
4:04



Be aware of what you are telling the software to plot.



Be aware of people with nefarious intent.

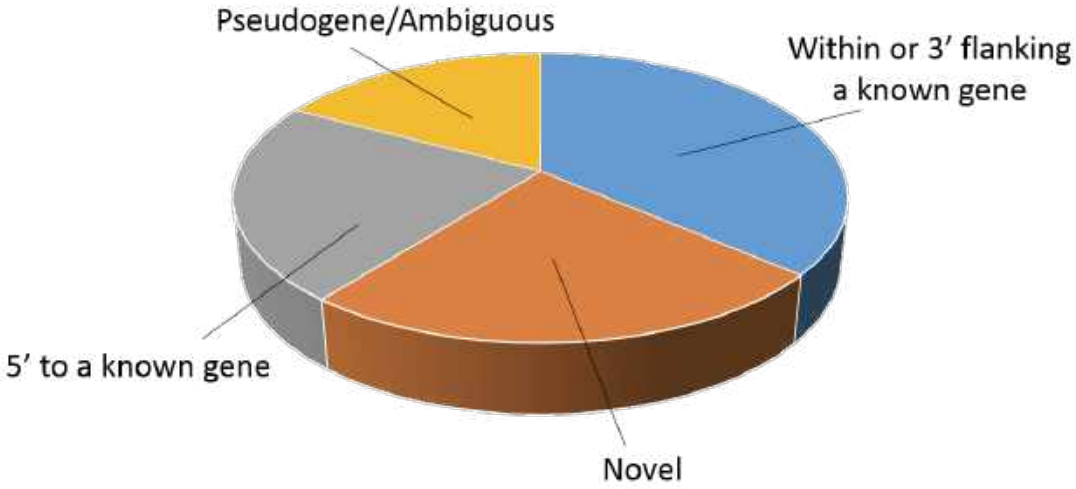


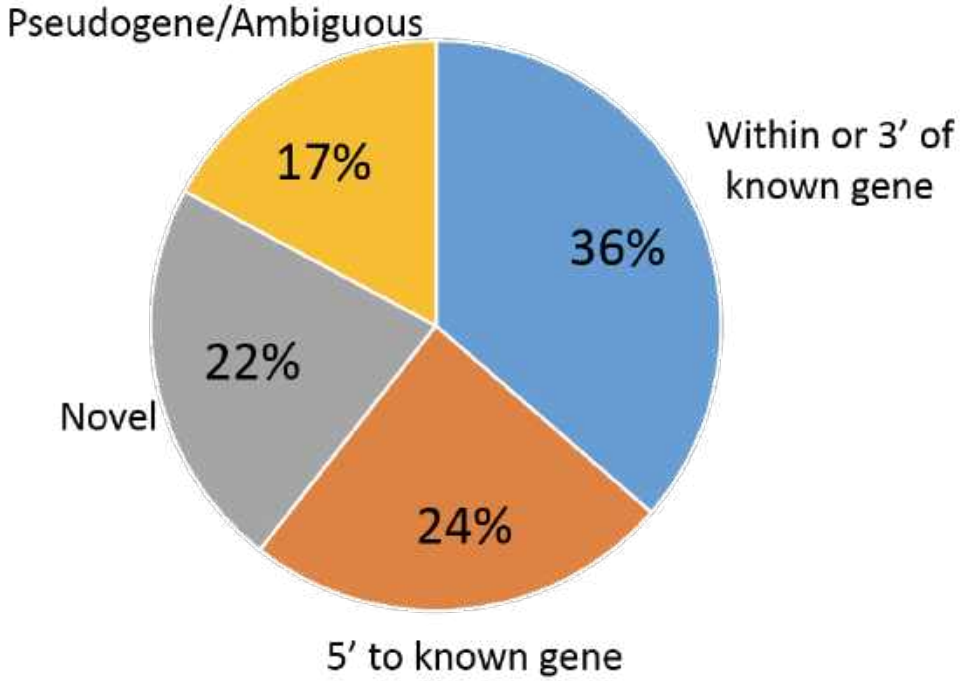
Be aware of the difference between proportions and percentages.

Be aware of whatever this is.

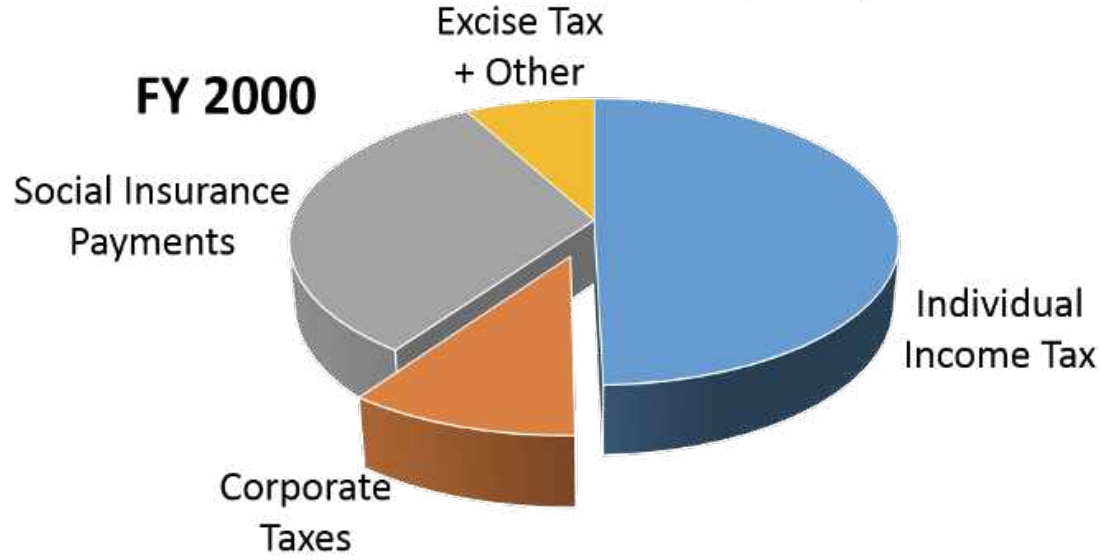
Be aware
of pie charts.

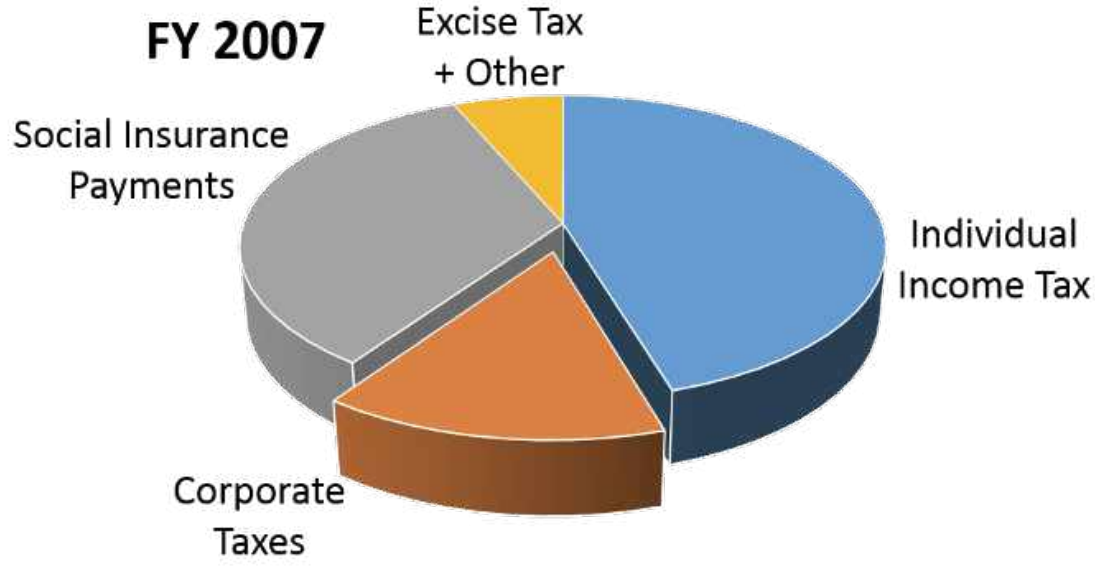
Distribution of All TFBS Regions

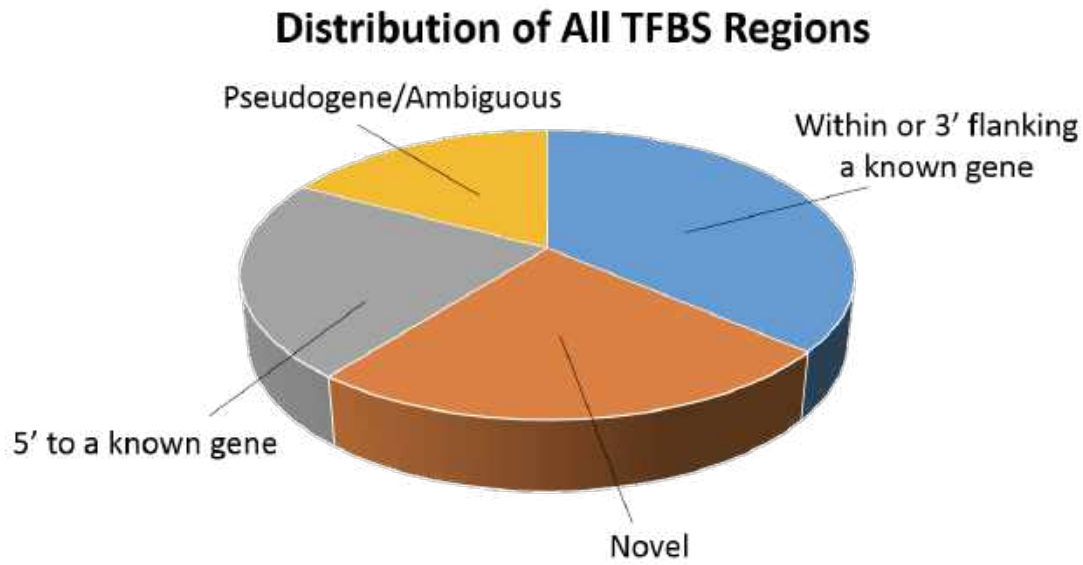




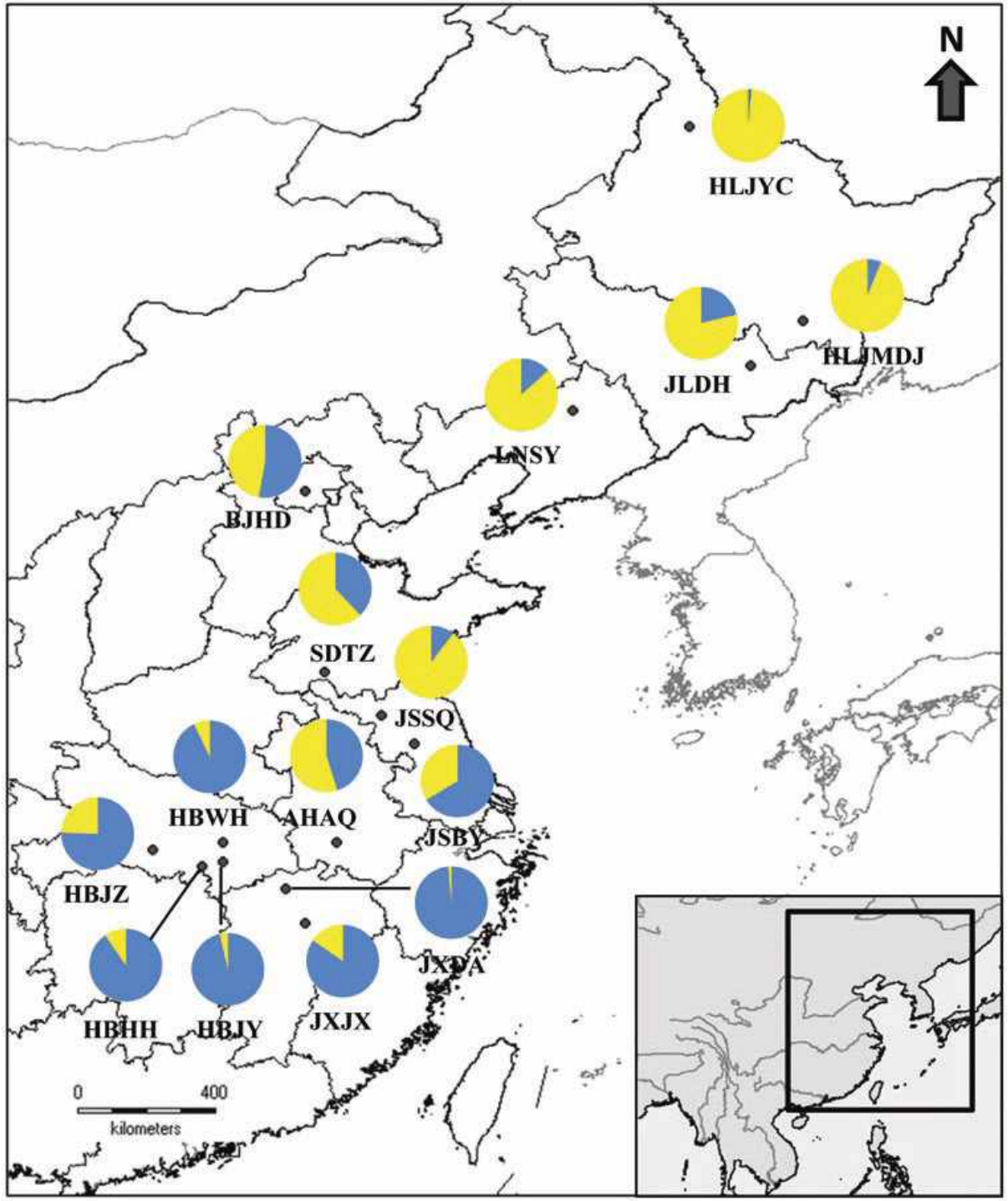
Federal Government Receipts by Source







Within or 3’ of known gene	36%
5' to known gene	24%
Novel	22%
Pseudogene/ambiguous	17%



Geographic locations of the sampling sites and corresponding genetic assignments (pie charts) by STRUCTURE for the 15 Zizania latifolia populations in China. Downloaded from https://academic.oup.com/aobpla/article-abstract/10/6/ply072/5225188 by guest on 25 December 2018

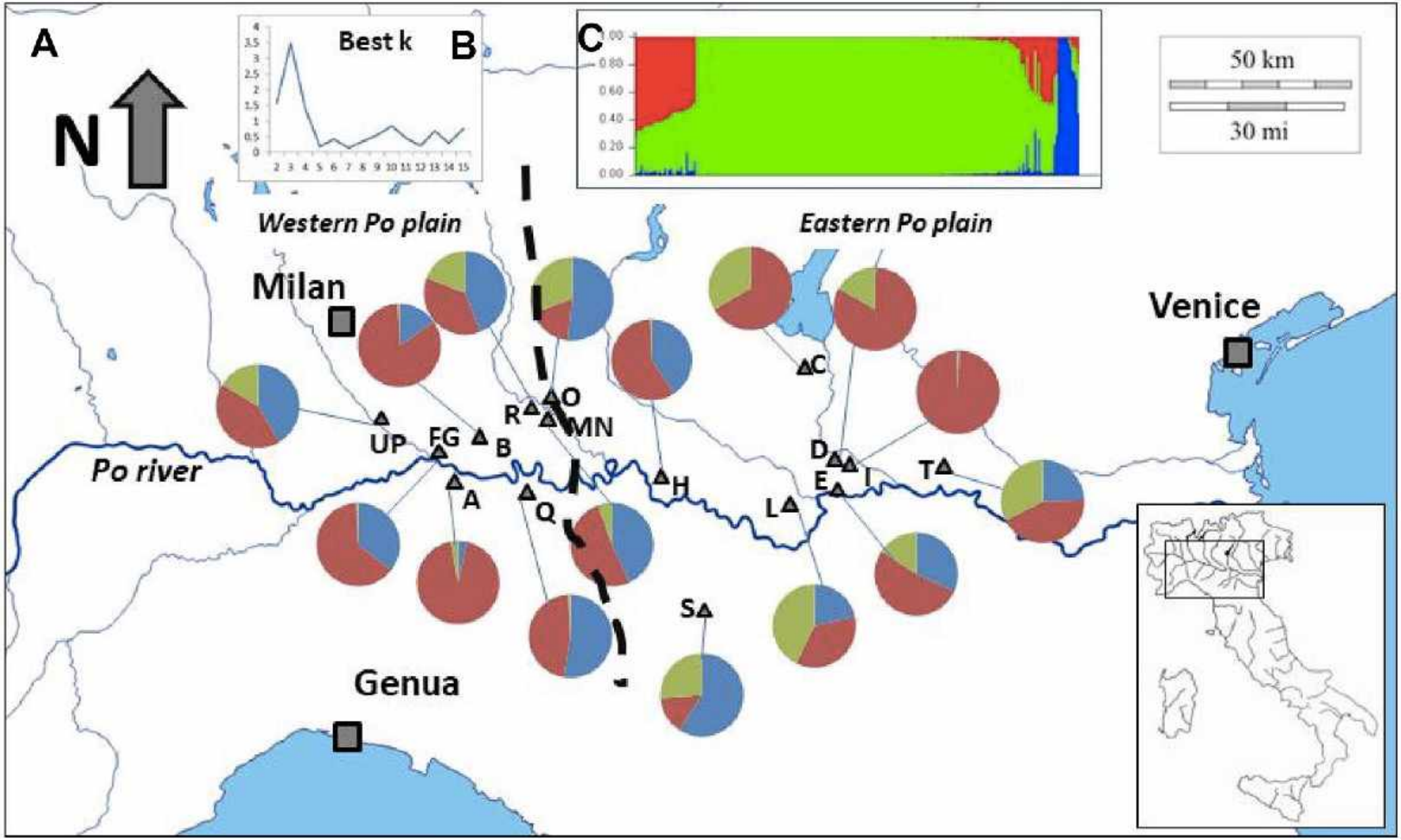
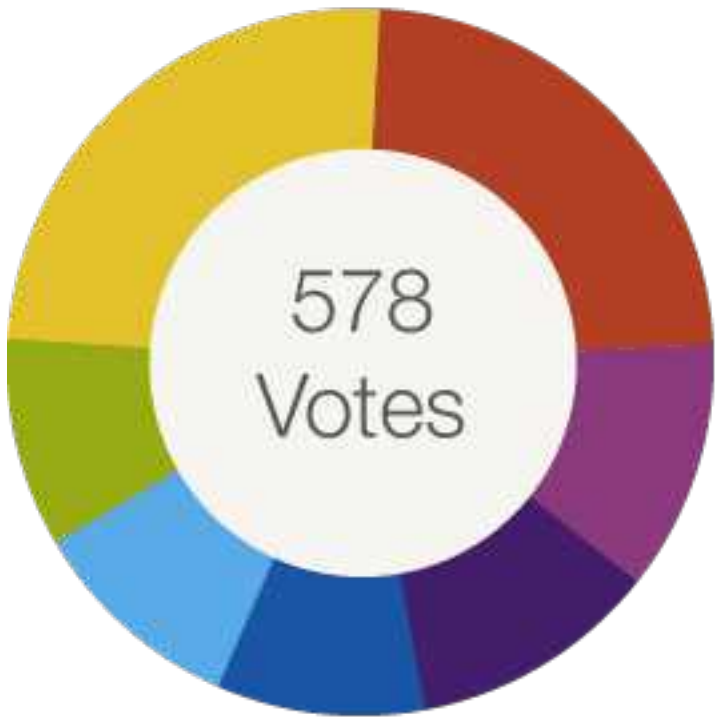
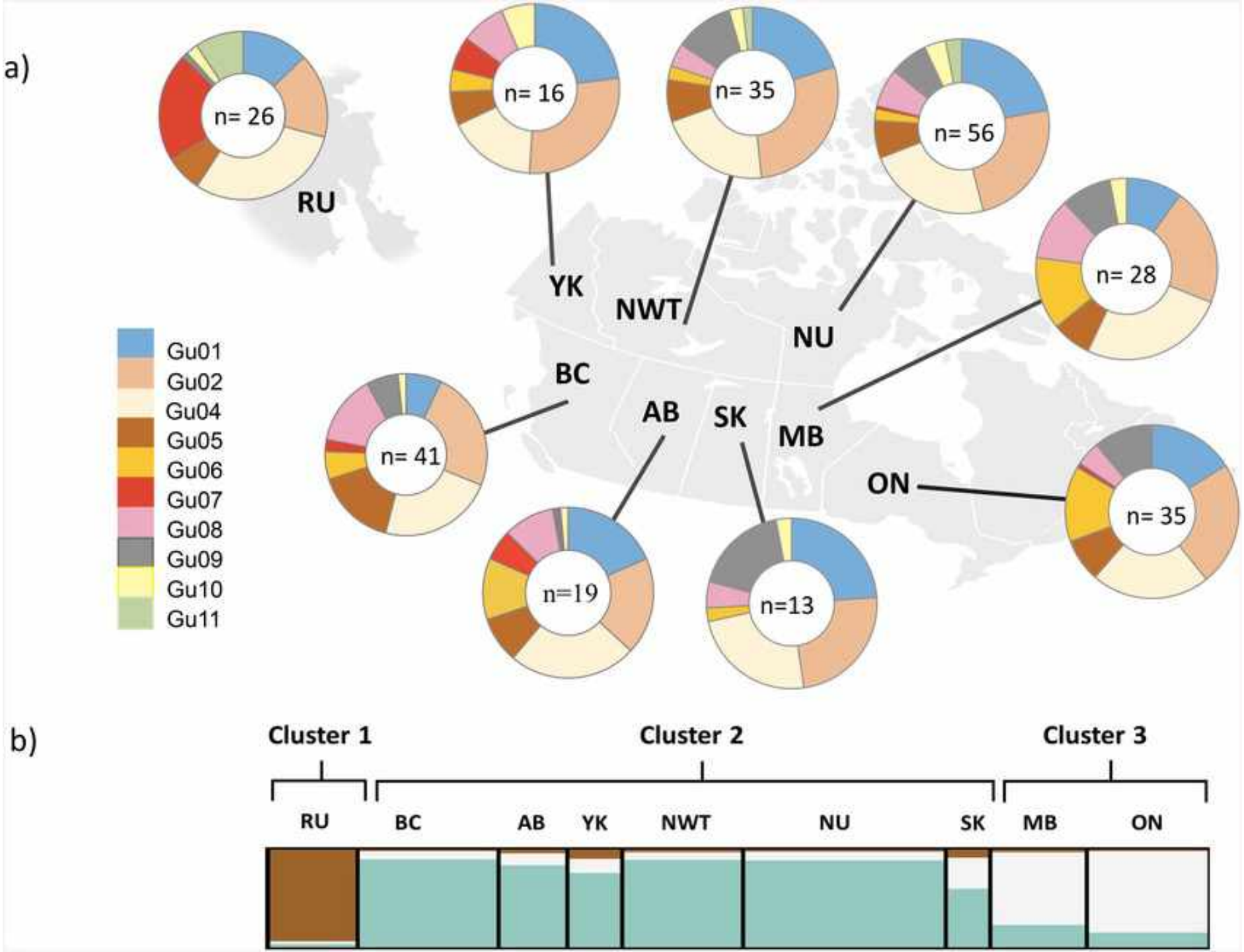


Fig. 2 A Spatial genetic structure and population clusters of L. aestivum inferred by Bayesian clustering implemented in STRU CTU RE. At each location, pie charts in the map indicate mean proportion of membership of individuals for K = 3 genetic groups; B results of the ΔK calculation (see "Materials and methods" for details); C in the bar diagram different colours (q values) represent the proportion of ancestry in each of the K populations.



Ecuador	30%
Brasil	10%
Argentina	10%
Peru	9%
Colombia	8%
Bolivia	8%
Others	25%



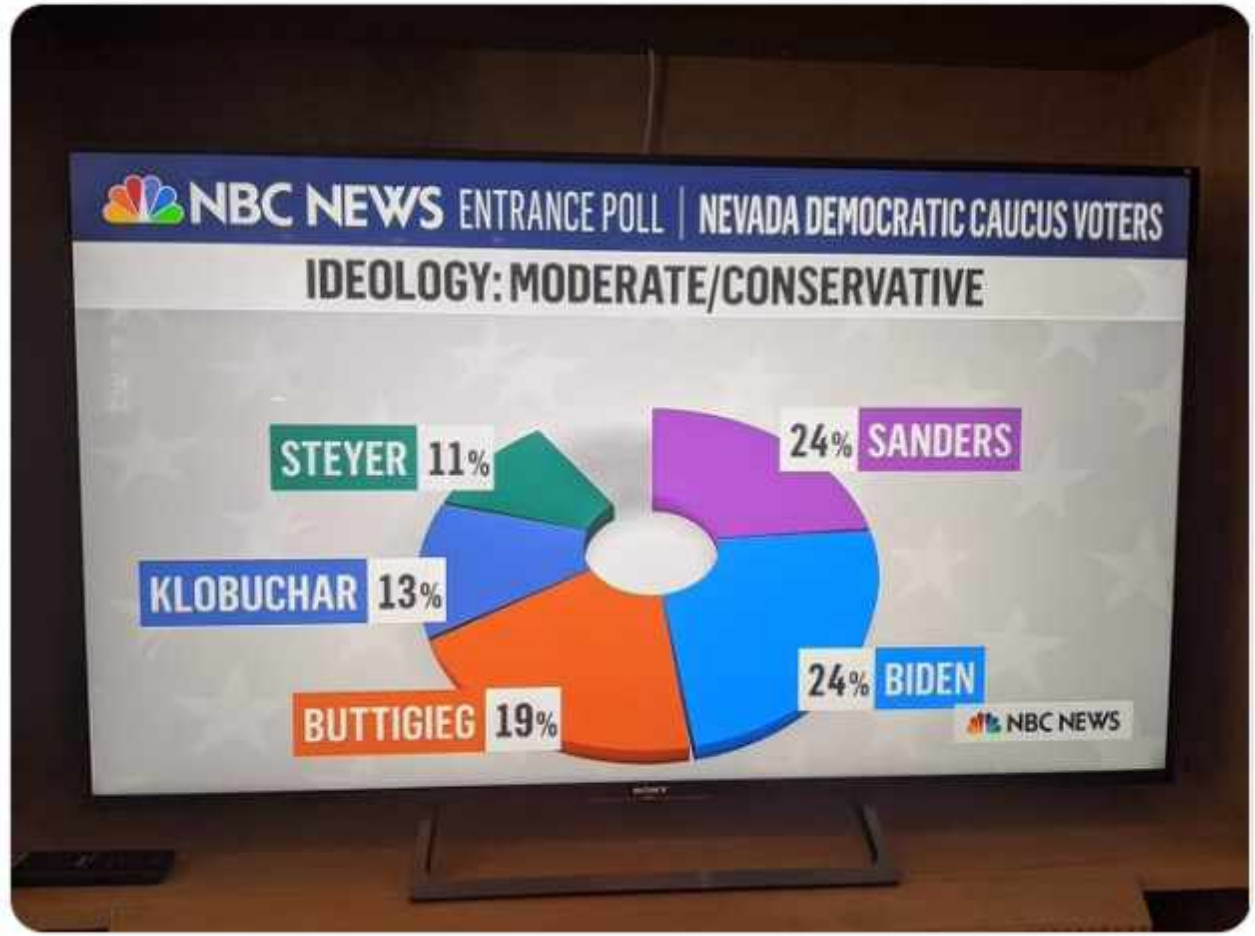
Patterns of microsatellite and MHC genetic variation within nine sampled regions. (a) Relative frequency distribution of ten MHC alleles per sampled region. Each color of the pie chart represents an MHC allele, while its size is proportional to the frequency of that allele within a location. Numbers within pie charts denote sample size. (b) STRUCTURE barplot of population membership scores for inferred k = 3 genetic clusters for 11 microsatellites.

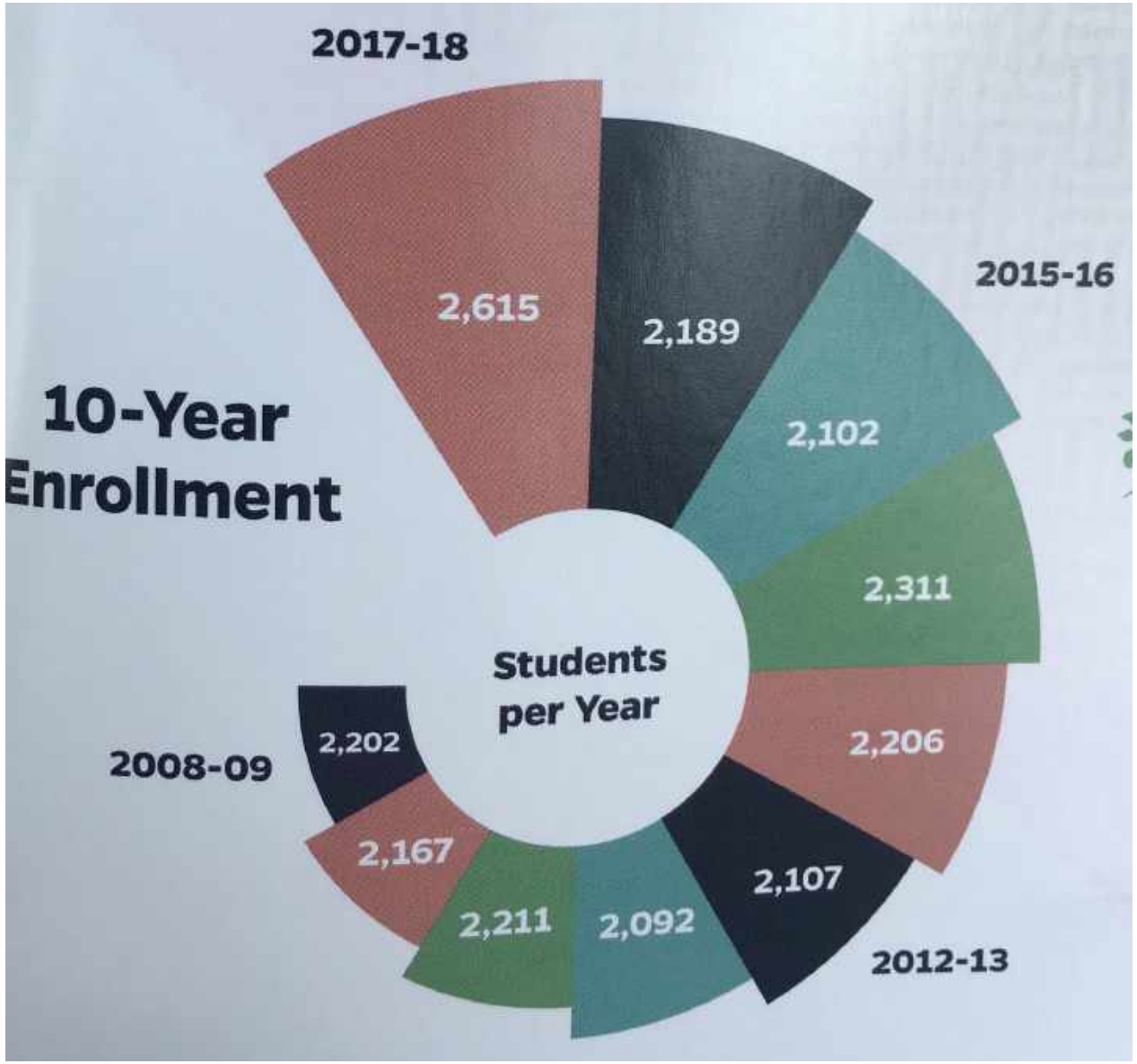
Dr Sardonicus 😎🦮☕

@scottmcmorrow

Replying to @ParkerMolloy

Look how perspective distortion was used to make Biden and Buttigieg look larger than Sanders and the rest of the field.





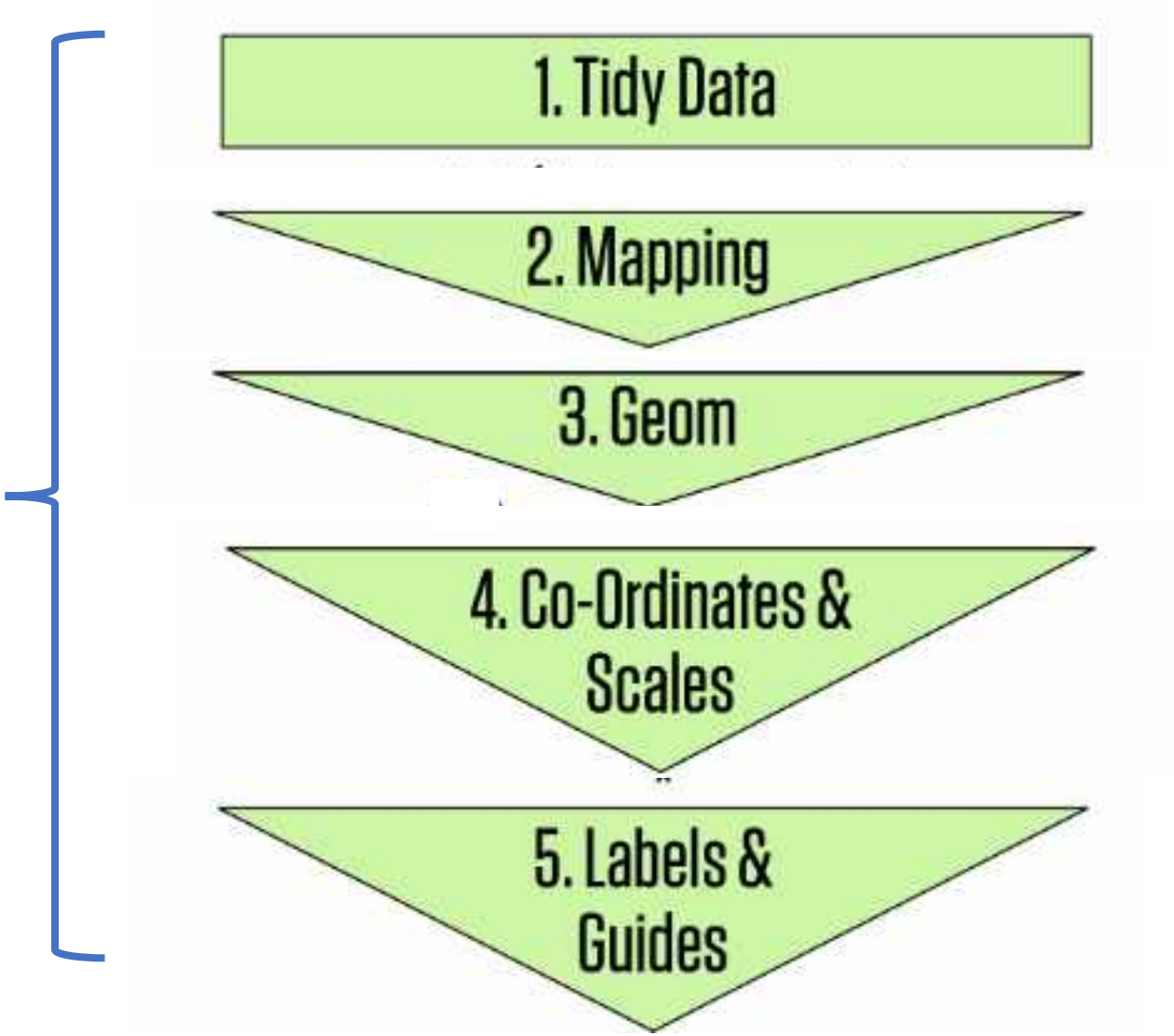
(Be aware of pre-made infographics leading you astray)

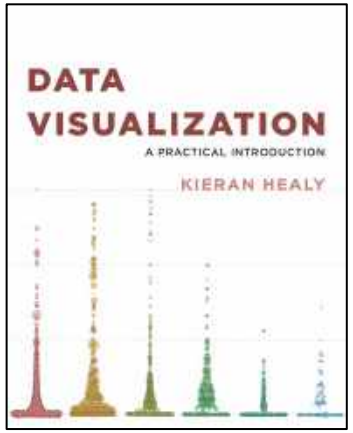
Be aware of making plots accessible.

ggplot2

"Grammar of Graphics"

(Leland Wilkenson – Object Oriented Design)





Images from Kieran Healy's *Data Visualization: A practical Introduction*
https://socviz.co

1. Tidy Data

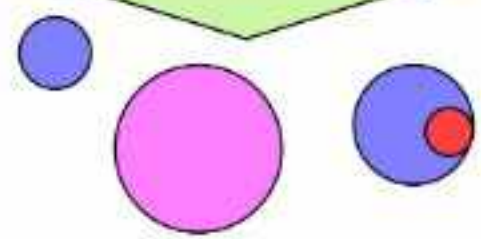
```
p ← ggplot(data = gapminder, ...
```

gdp	lifexp	pop	continent
340	65	31	Euro
227	51	200	Amer
909	81	80	Euro
126	40	20	Asia

2. Mapping

```
p ← ggplot(data = gapminder,
           mapping = aes(x = gdp,
           y = lifexp, size = pop,
           color = continent))
```

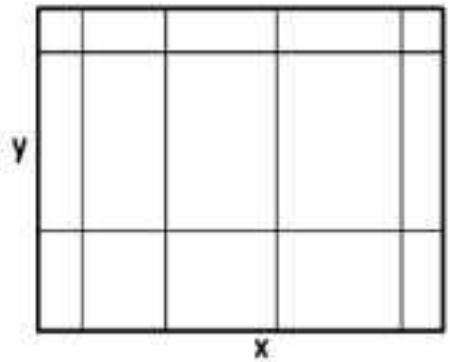
3. Geom



```
p + geom_point()
```

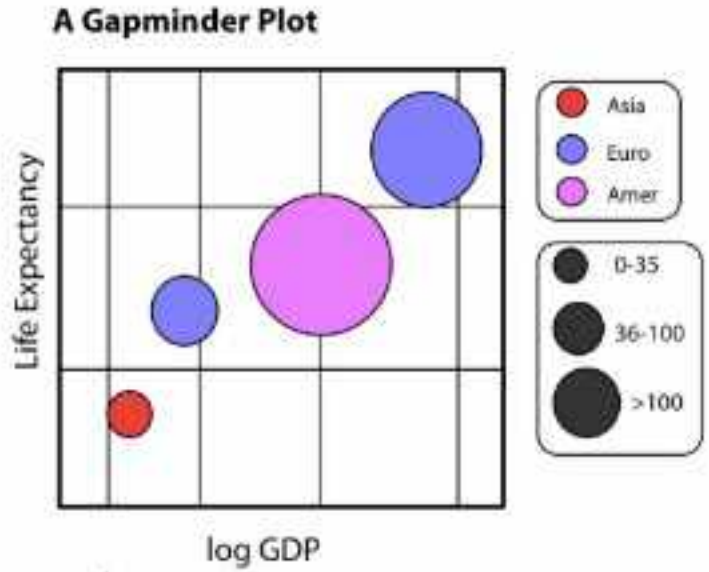
4. Co-Ordinates & Scales

```
p + coord_cartesian() +
    scale_x_log10()
```



5. Labels & Guides

```
p + labs(x = "log GDP",
         y = "Life Expectancy",
         title = "A Gapminder Plot")
```



Does it seem like a lot?
Fortunately, ggplot2 starts us off with some decent defaults