Introduction to Data Visualization

Halil Bisgin, PhD

What Is Data Visualization?

- Data visualization is the graphical representation of information and data.
- It provides a pictorial representation of the data to see trends, outliers, and patterns.

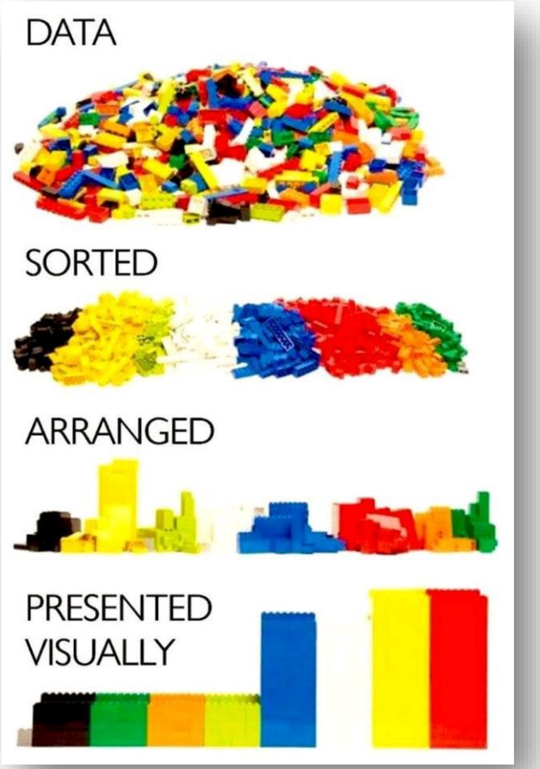
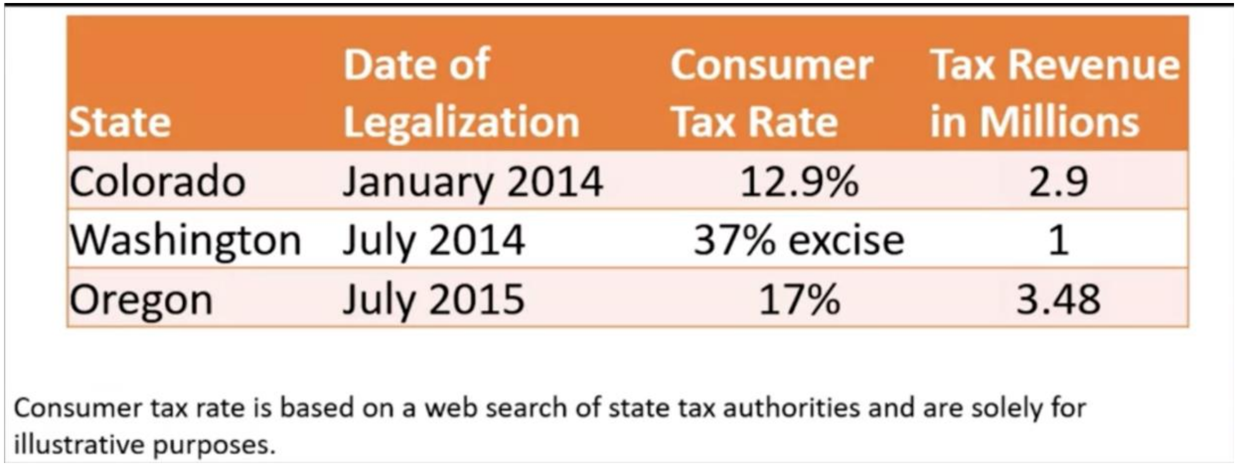
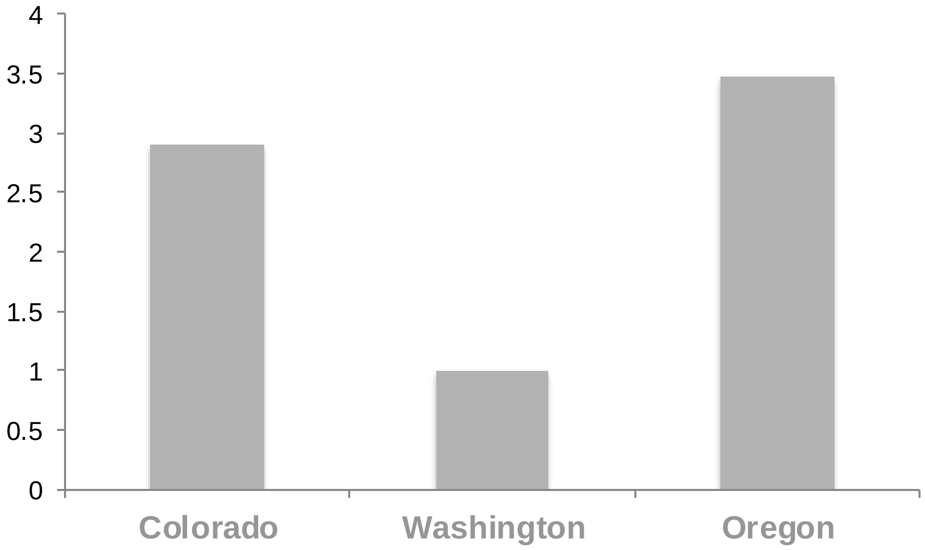


Table vs. Figure



State	Date of Legalization	Consumer Tax Rate	Tax Revenue in Millions
Colorado	January 2014	12.9%	2.9
Washington	July 2014	37% excise	1
Oregon	July 2015	17%	3.48

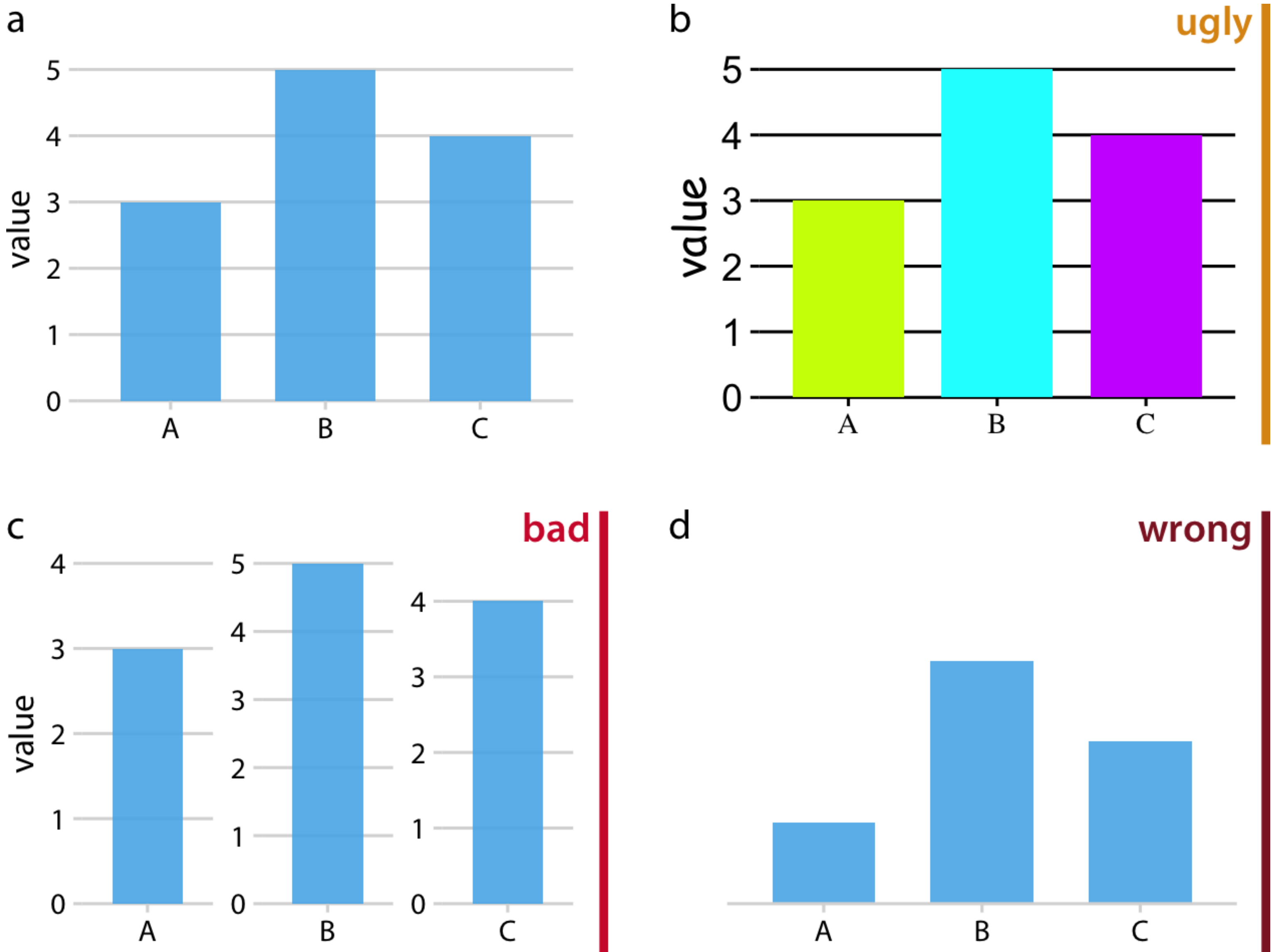
Consumer tax rate is based on a web search of state tax authorities and are solely for illustrative purposes.



Ugly, Bad, and Wrong

- Ugly:
  - *A figure that has aesthetic problems but otherwise is clear and informative*
- Bad:
  - *A figure that has problems related to perception; it may be unclear, confusing, overly complicated, or deceiving*
- Wrong:
  - *A figure that has problems related to mathematics; it is objectively incorrect*
- Among the good ones, there may always be better ones

Ugly, Bad, and Wrong

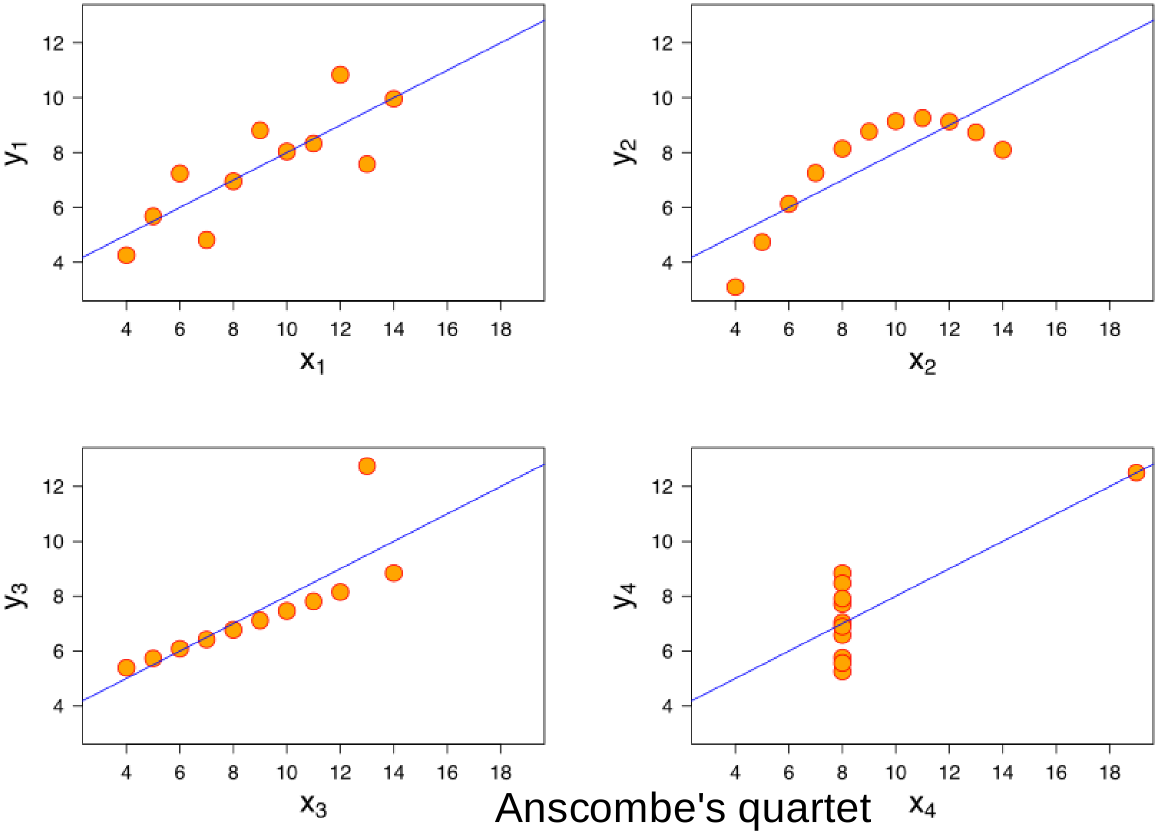


How can we make better figures?

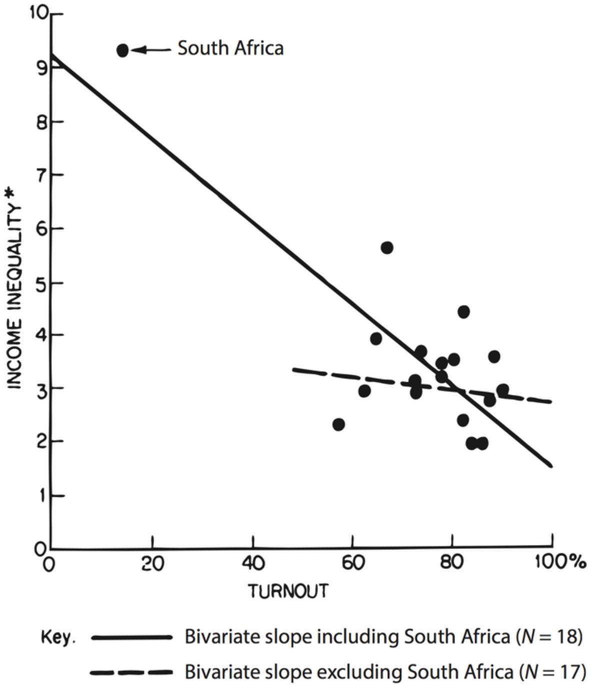
- It cannot be boiled down to a list of simple rules.
- Intended audience is also important regardless of how it looks.
  - *A scientific journal reader vs. general public*
- There is only so much that your software can do to keep you on the right track. Rest is up to you: doing right thing, being honest with your data and audience.
- However, there are still good visualization methods and principles we should use and stick to.

Why look at data?

- Statistical summary can be the same, but …



Anscombe's quartet

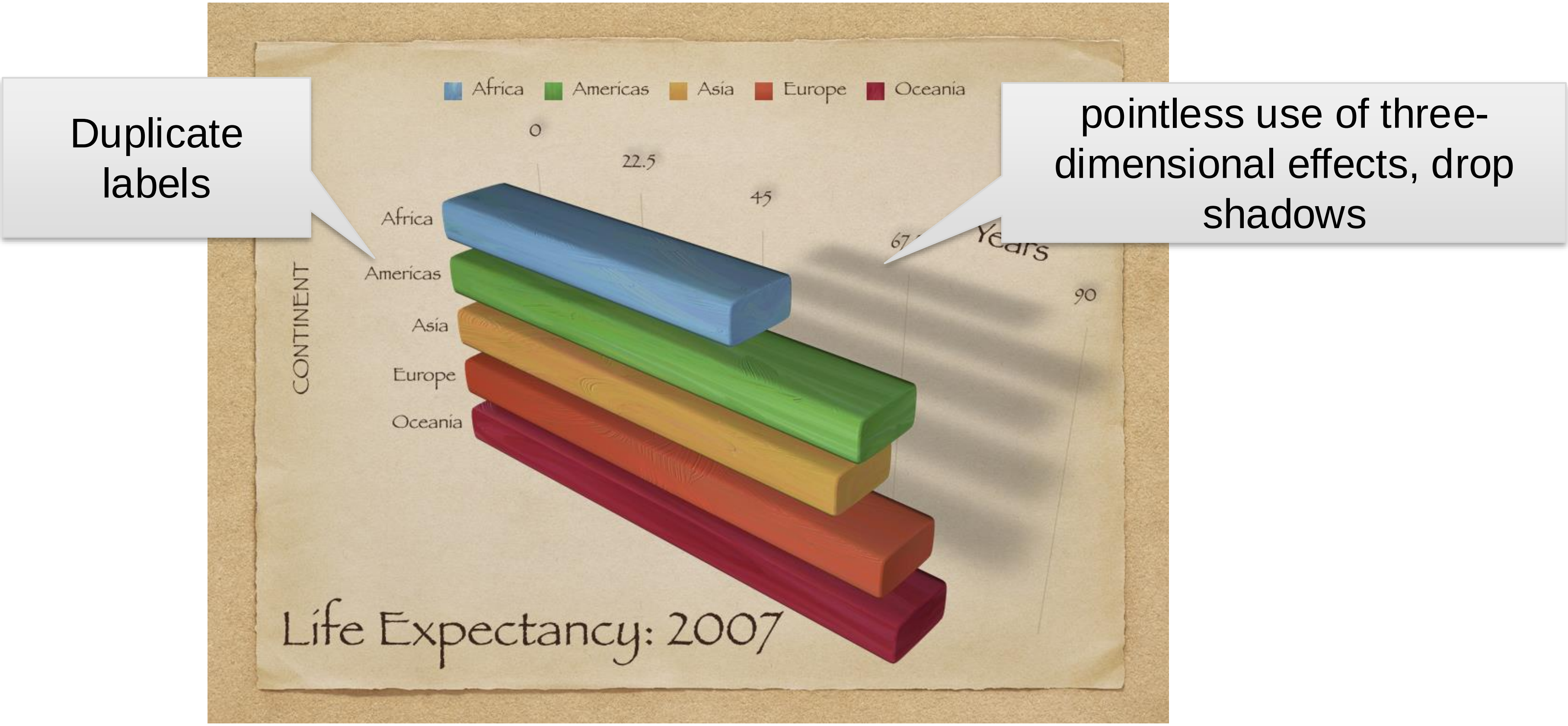


What makes bad figures bad?

- Parade of horribles
  - *Negative examples motivate good behavior*
- Negative examples often combine several kinds of badness that are better kept separate
- Our problems tend to come in three varieties
  - *Aesthetic: tacky, tasteless*
  - *Substantive: the way data represented*
  - *Perceptual: confusing or misleading because of how people perceive*

Bad taste

- Modest amount of info, but too much going on



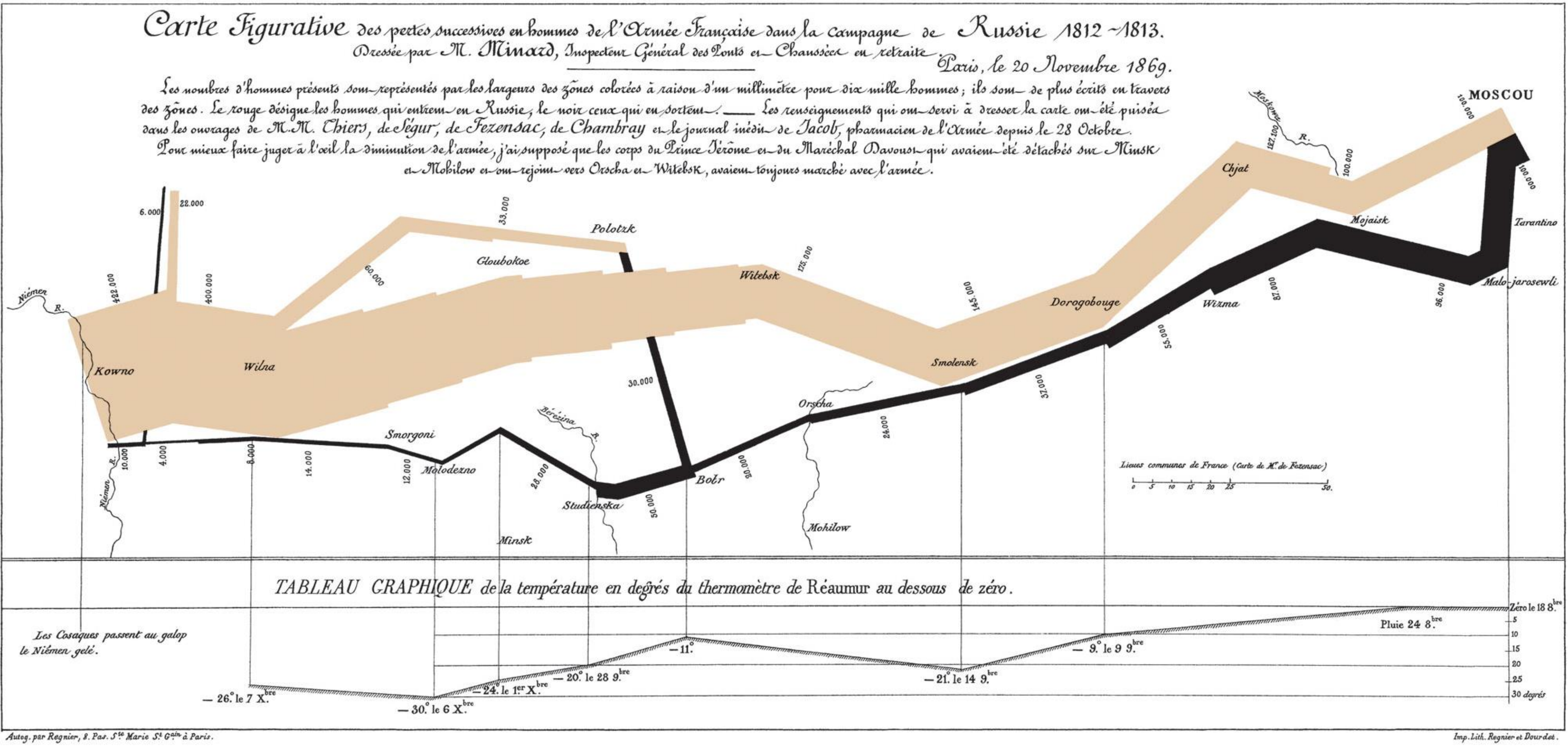
Duplicate labels

pointless use of three-dimensional effects, drop shadows

What is graphical excellence?

- The well-designed presentation of interesting data—a matter of substance, of statistics, and of design …
- Consists of complex ideas communicated with clarity, precision, and efficiency. …
- Gives the greatest number of ideas in the shortest time with the least ink in the smallest space …
- Nearly always multivariate …
- Requires telling the truth about the data. (Tufte,1983).

"may well be the best statistical graphic"



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.

Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.

Imp. Lith. Regnier et Dourdet.

Napoleon's march (retreat) on (from) Moscow by Charles Joseph Minard
(Paris, November 20, 1869)

Tufte’s comments on Minard’s graphic

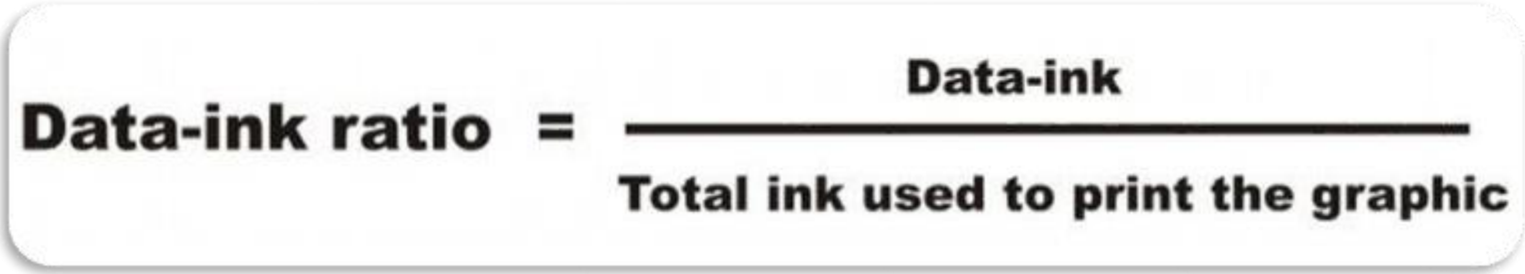
- Tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time.
- Six variables are plotted:
    - *the size of the army,*
    - *its location on a two-dimensional surface,*
    - *direction of the army’s movement, and*
    - *temperature on various dates during the retreat from Moscow*

Back to graphical excellence

- Tufte acknowledges that a *tour de force* such as Minard’s “*can be described and admired, but there are no compositional principles on how to create that one wonderful graphic in a million*” .
- The best one can do for “*more routine, workaday designs*” is to suggest some guidelines such as
  - *“have a properly chosen format and design,”*
  - *“use words, numbers, and drawing together,”*
  - *“display an accessible complexity of detail”, and*
  - *“avoid content-free decoration, including chartjunk”*

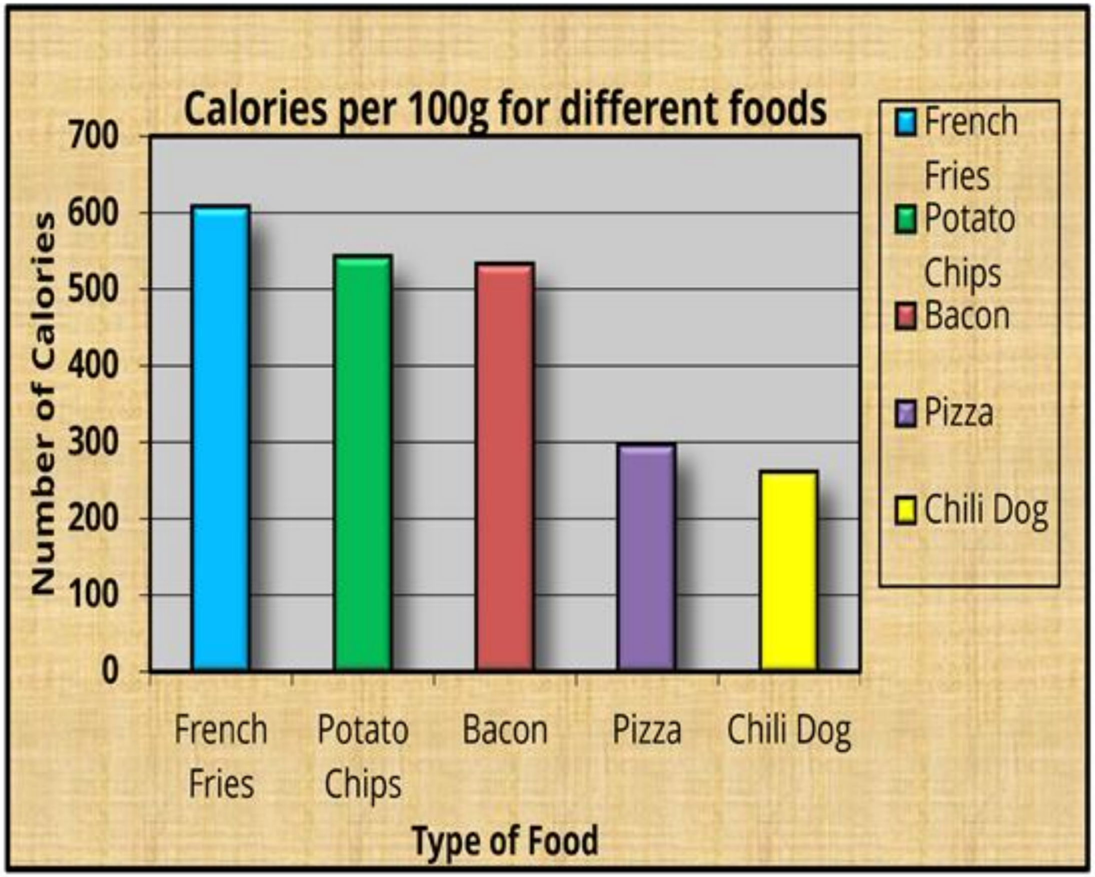
Graphical Heuristics

- Many experts criticize the inclusion of visual *embellishment* in charts and graphs.
- They claim that addition of chart junk, decorations and other kinds of non-essential imagery, can make interpretation more difficult and can distract readers
- They advocate plain and simple charts that maximize the proportion of data-ink (data-ink ratio).

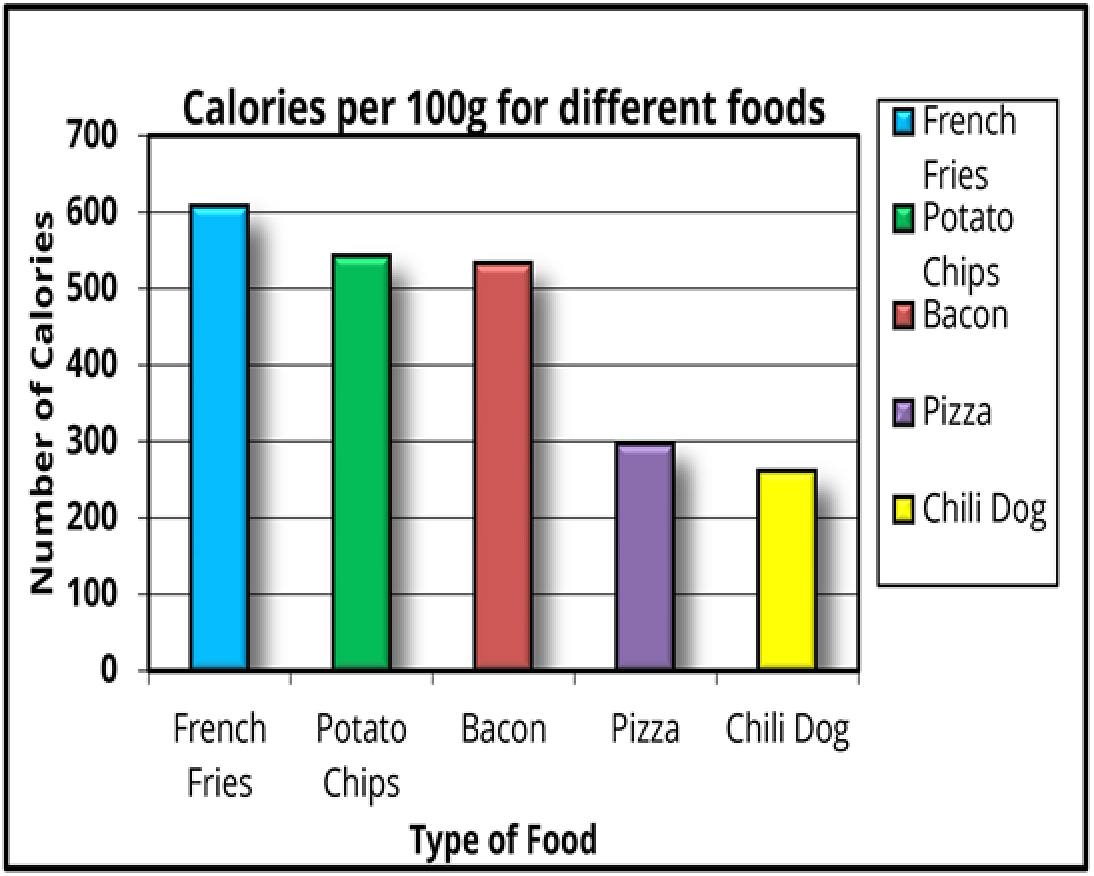


$$\textbf{Data-ink ratio} = \frac{\textbf{Data-ink}}{\textbf{Total ink used to print the graphic}}$$

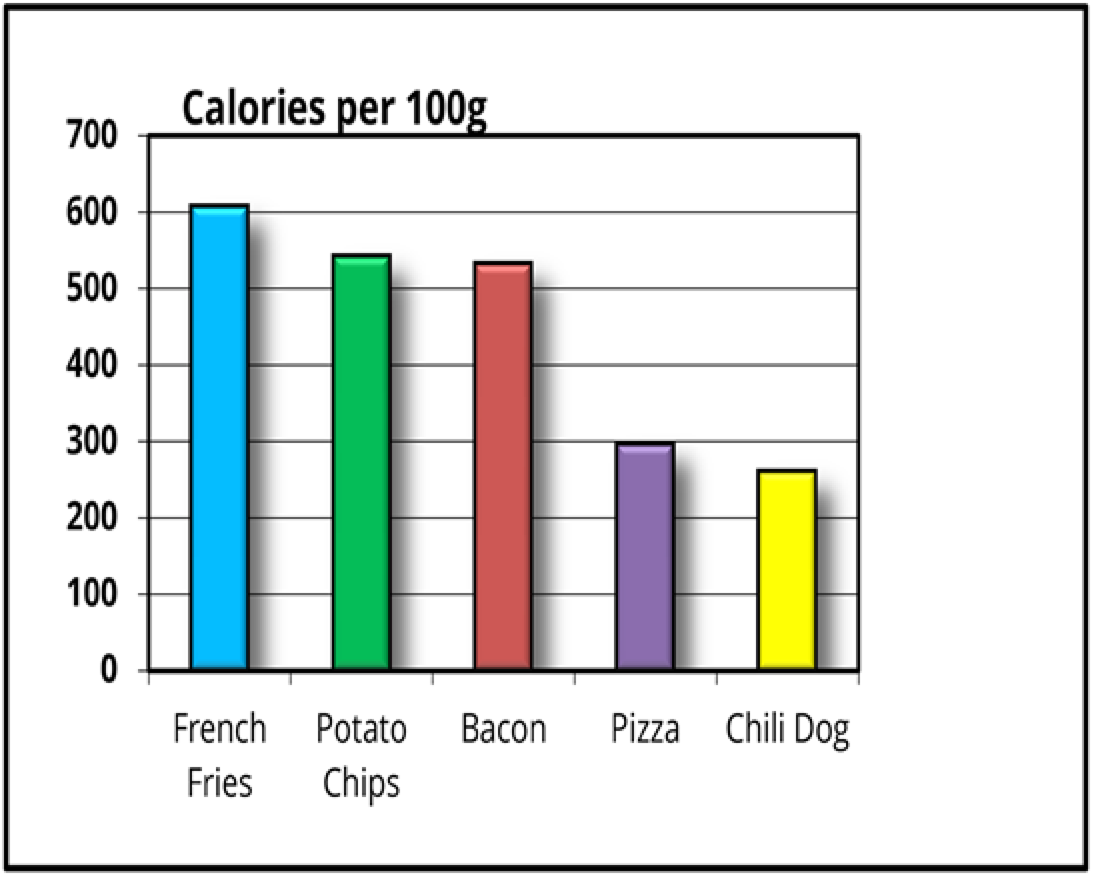
Remove backgrounds



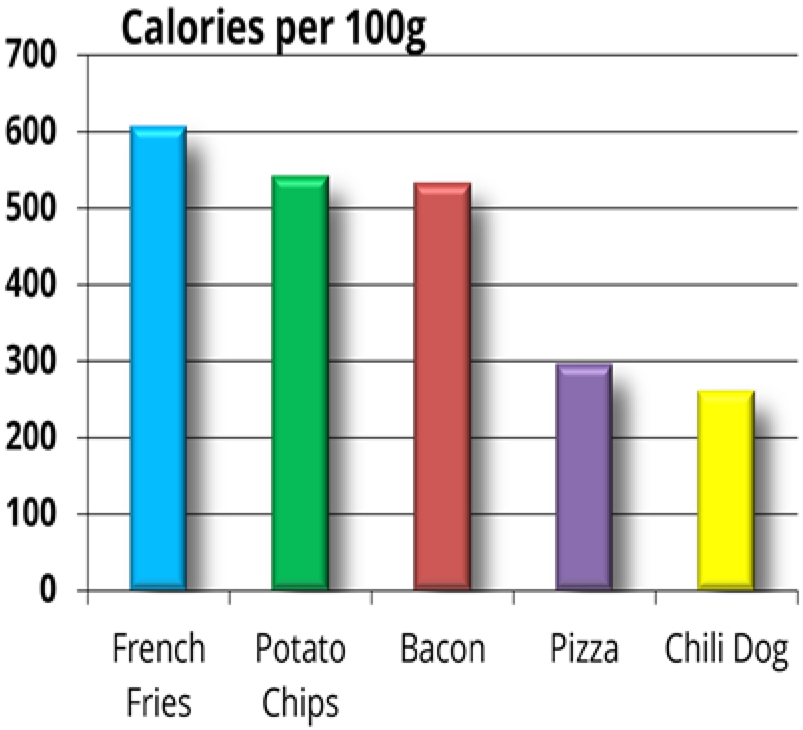
Remove redundant labels



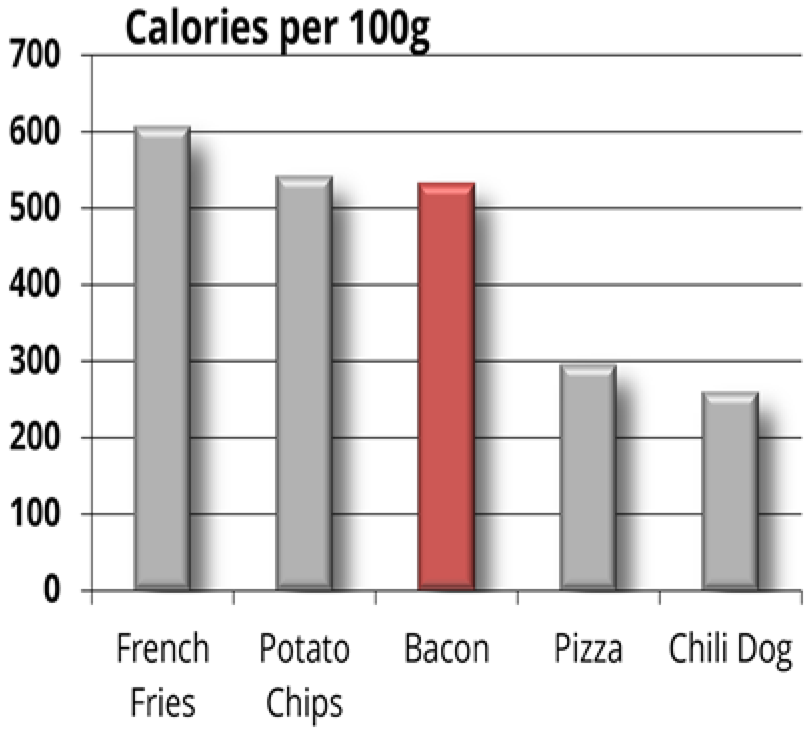
Remove borders



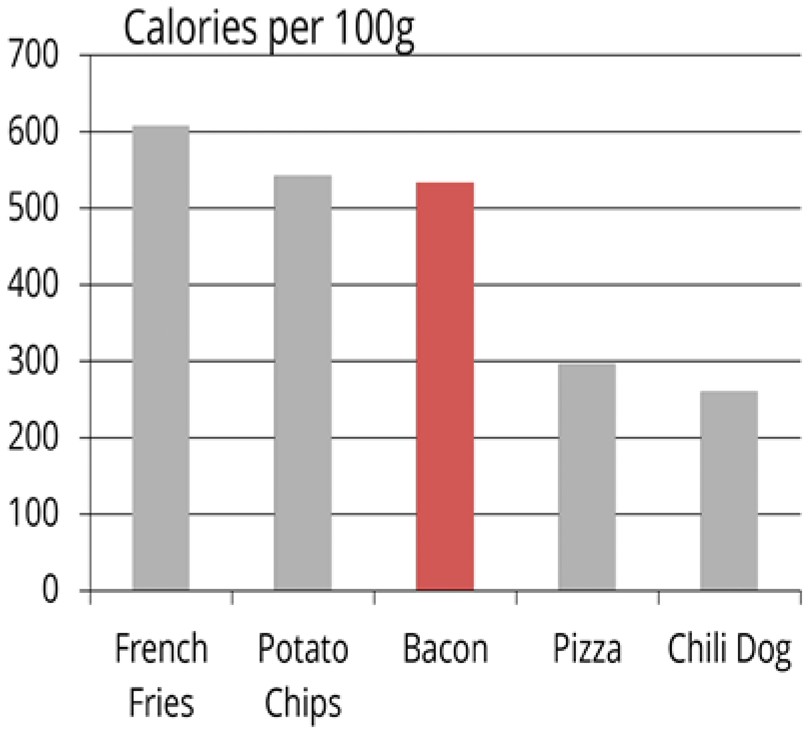
Reduce colors



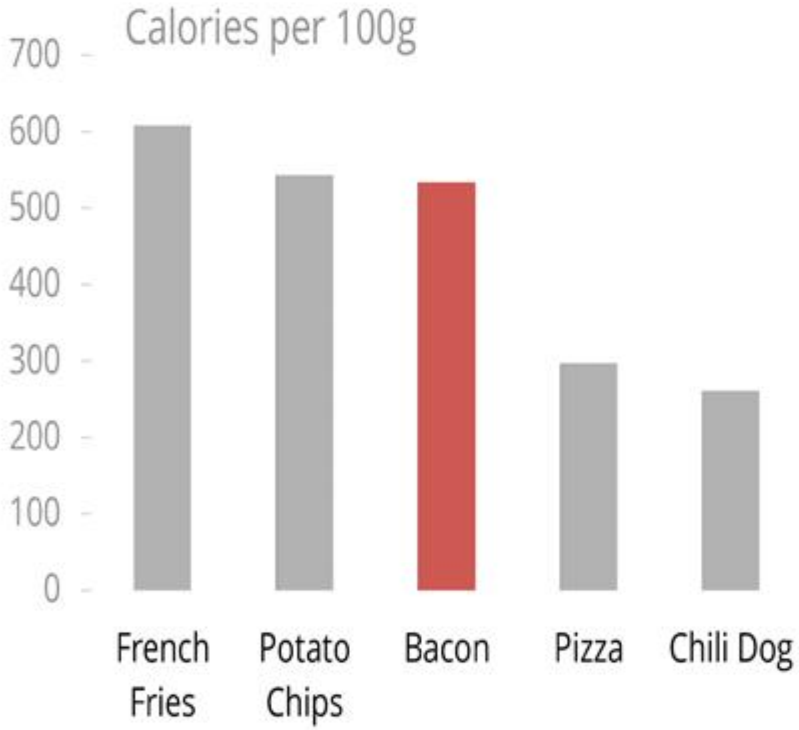
Remove special effects



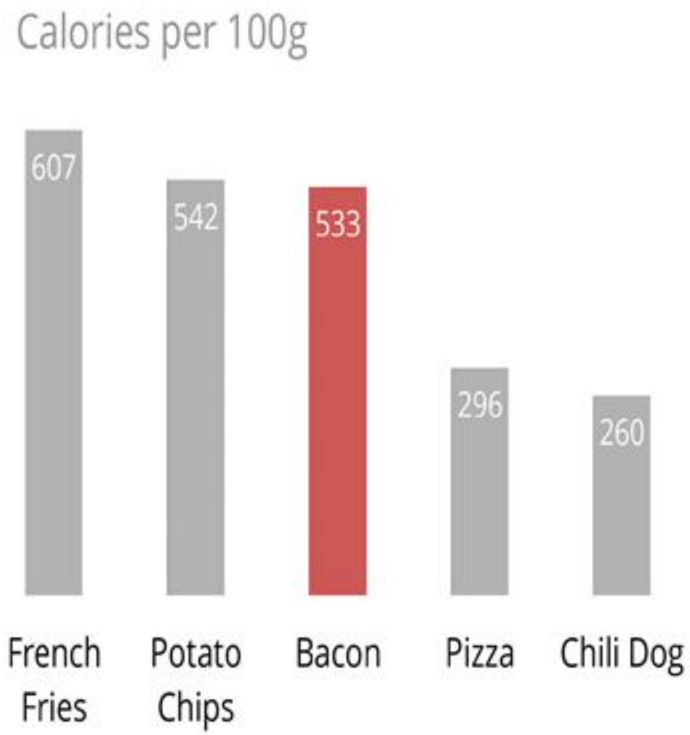
Lighten labels

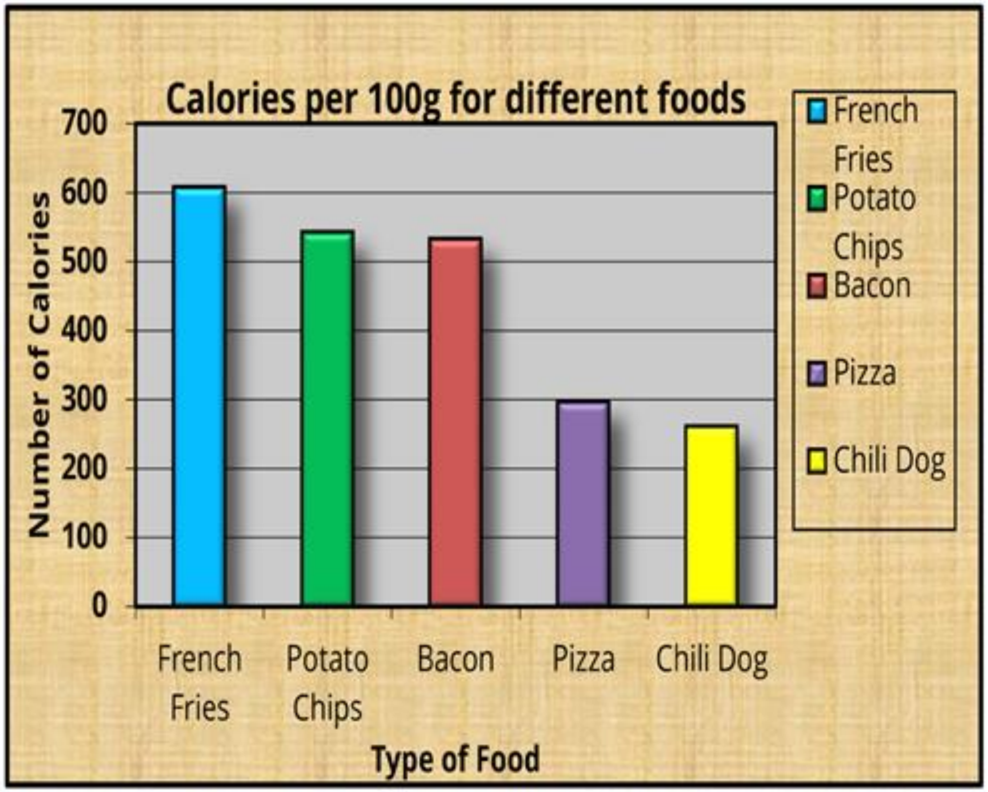


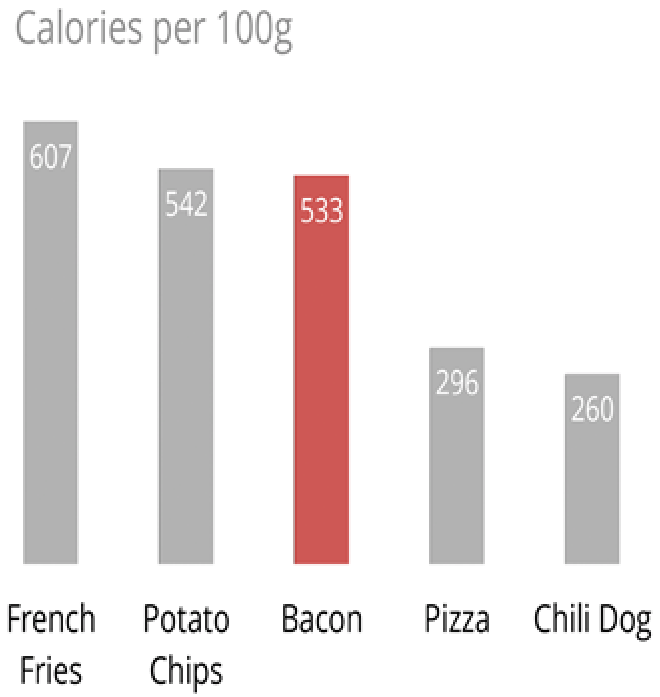
Direct label



Direct label




Calories per 100g for different foods
Number of Calories
700
600
500
400
300
200
100
0
French Fries
Potato Chips
Bacon
Pizza
Chili Dog
Type of Food
French Fries
Potato Chips
Bacon
Pizza
Chili Dog

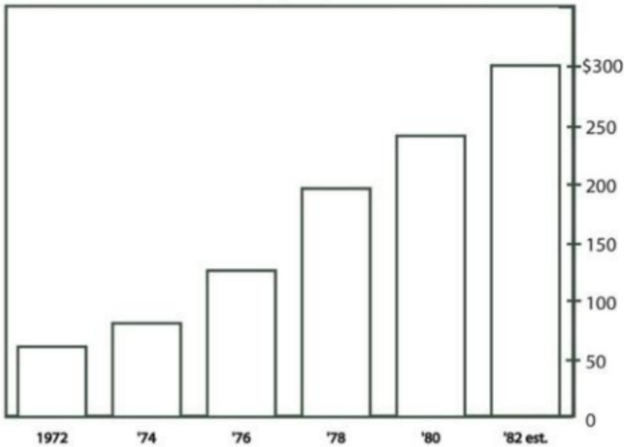

Calories per 100g
607
542
533
296
260
French Fries
Potato Chips
Bacon
Pizza
Chili Dog

How to maximize data-to-ink ratio?

- It is not hard to jettison tasteless junk!
- We can often
  - *clean up the typeface,*
  - *remove extraneous colors and backgrounds,*
  - *simplify, mute, or delete gridlines, superfluous axis marks, or needless keys and legends.*
  - *Remove excessive shading or patterning of chart features*
- Direct labeling of data is another great way to reduce this form of chartjunk

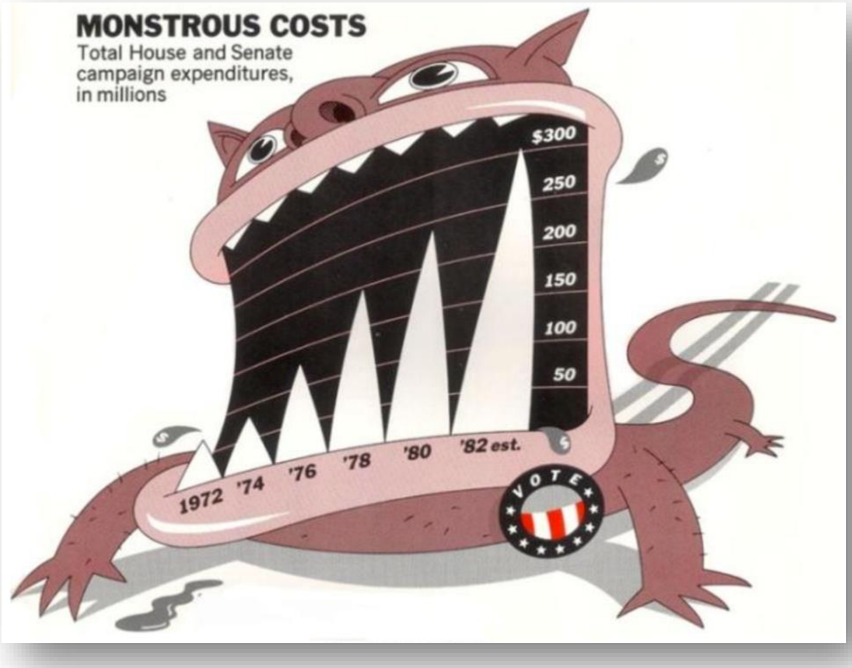
In practice, however, there are exceptions

- To have junk-free doesn't guarantee effectiveness
- There is evidence that highly embellished charts like Nigel Holmes's "Monstrous Costs" are often more easily recalled than their plainer alternatives.



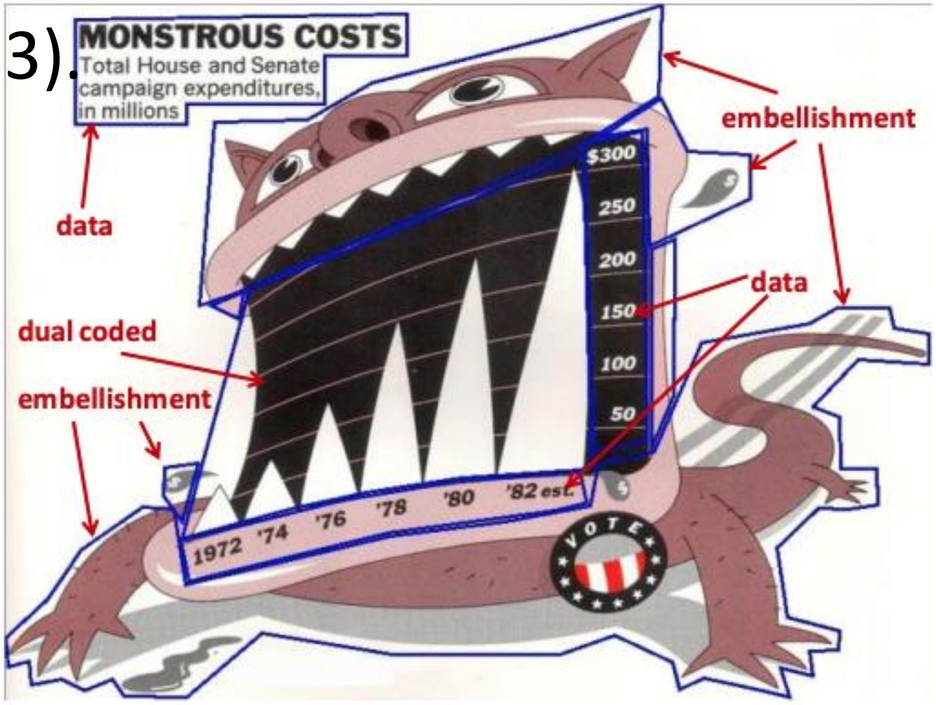
V.S

.



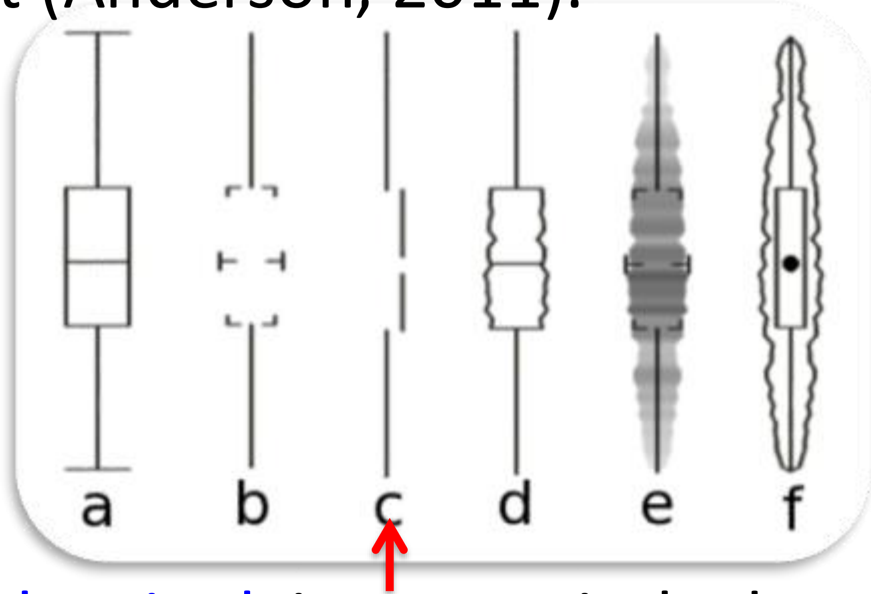
More on *Monstrous Costs*

- Not that easy to interpret, but easy to remember and enjoyable.
- Visually unique, "Infographic" style graphs are more memorable than more standard statistical visualizations (Borkin, 2013).
-



Data-to-ink and chartjunk

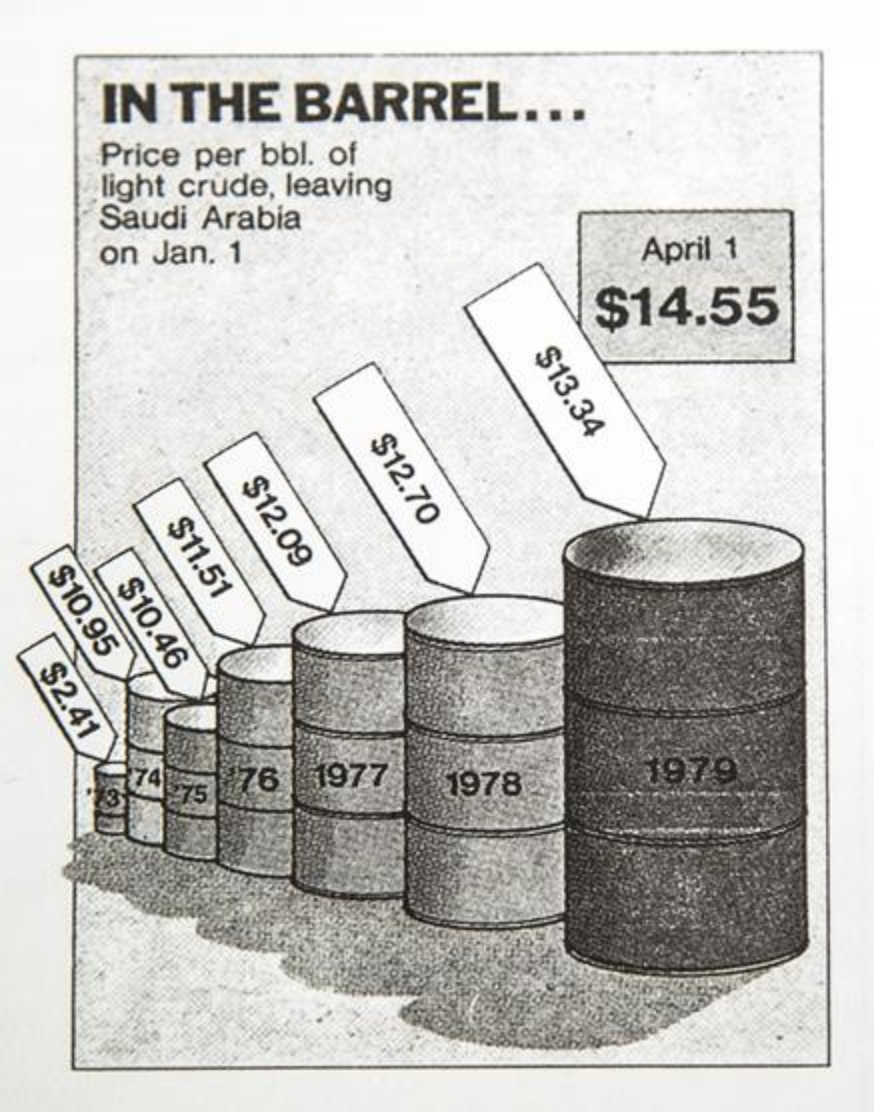
- Of the four kinds of boxplot, the minimalist version (C) from Tufte proved to be the most cognitively difficult (Anderson, 2011).



- While chartjunk is not entirely devoid of merit, bear in mind that ease of recall is only one virtue amongst many for graphics.

Lie factor

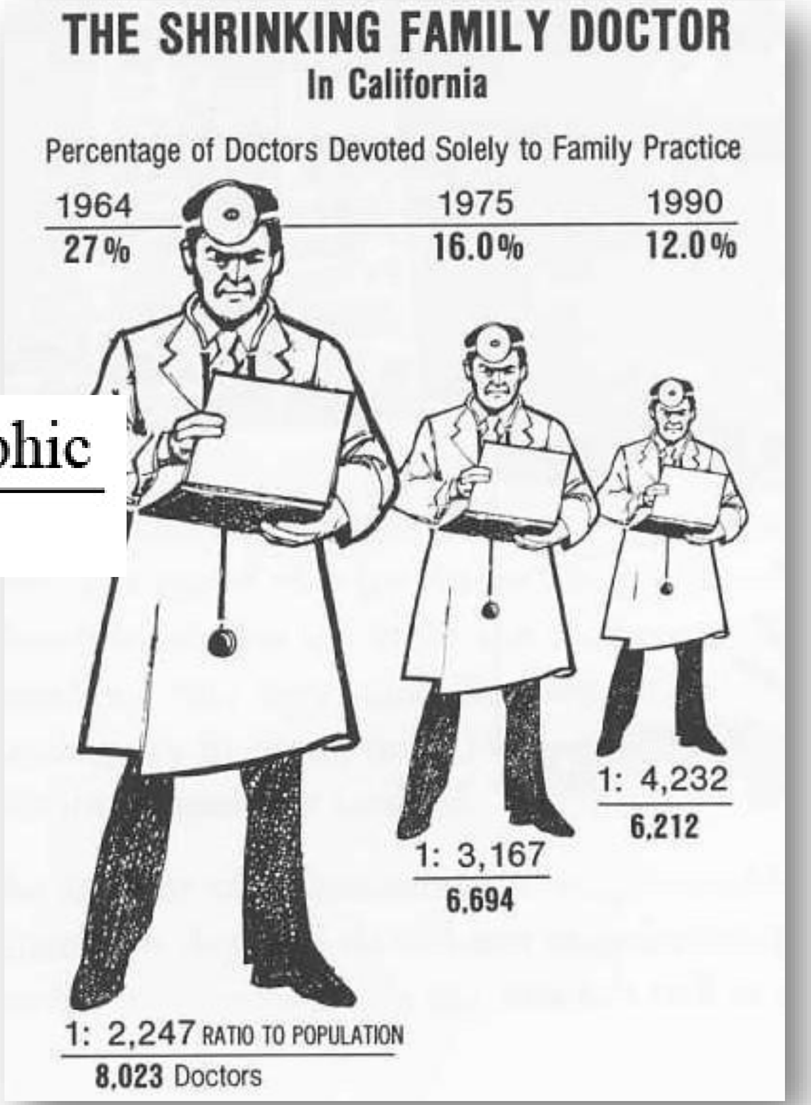
- *The "Lie Factor" is a value to describe the relation between the size of effect shown in a graphic and the size of effect shown in the data.*
- *"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented."-- Tufte*
- *If you consider volume, the lie factor is 9.4 times the stated prices*



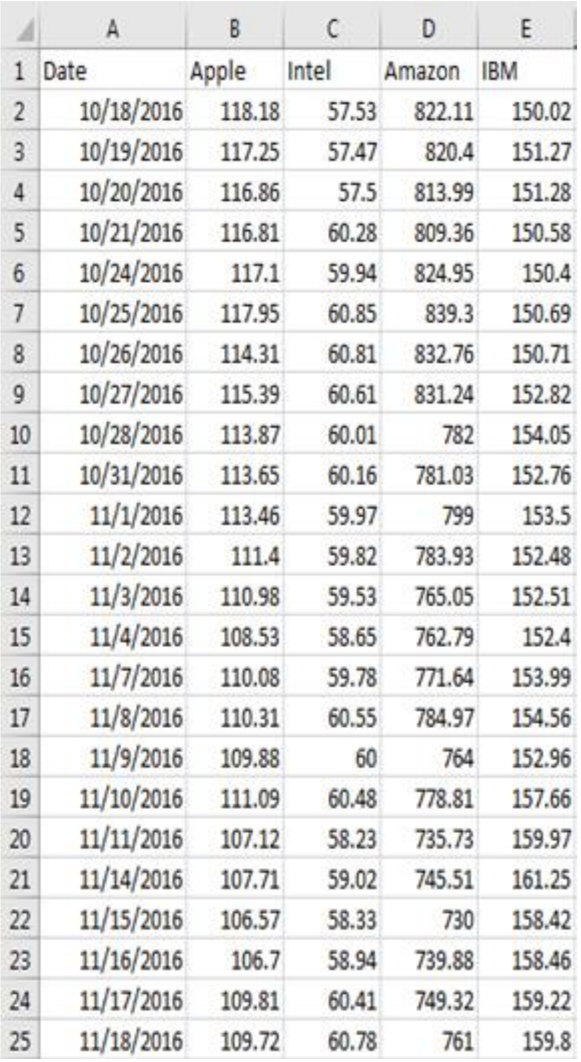
Lie factor

- Lie factor 2.8

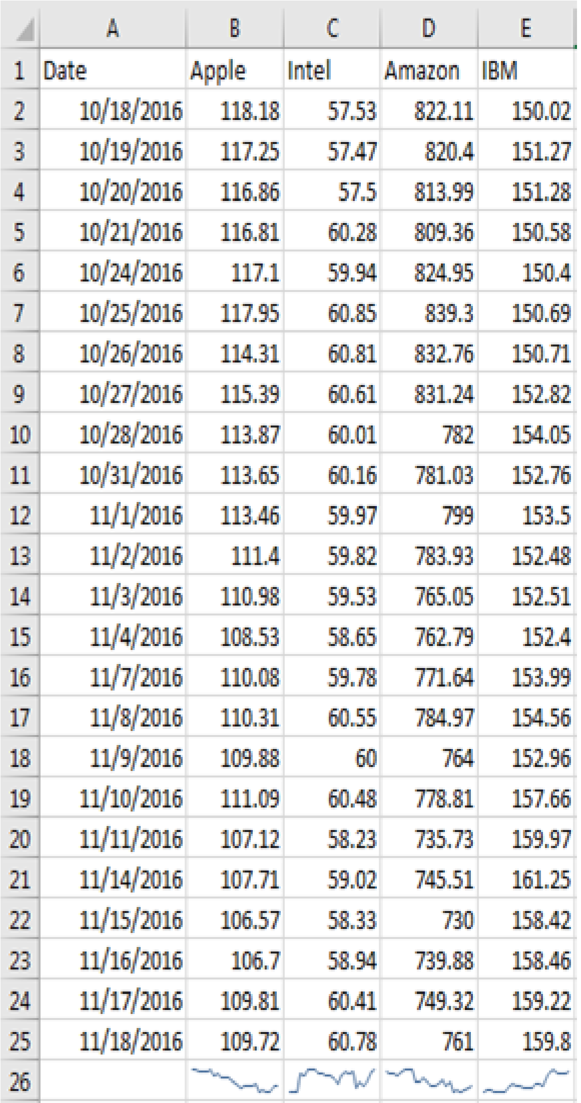
$$\textit{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$



Spark Lines

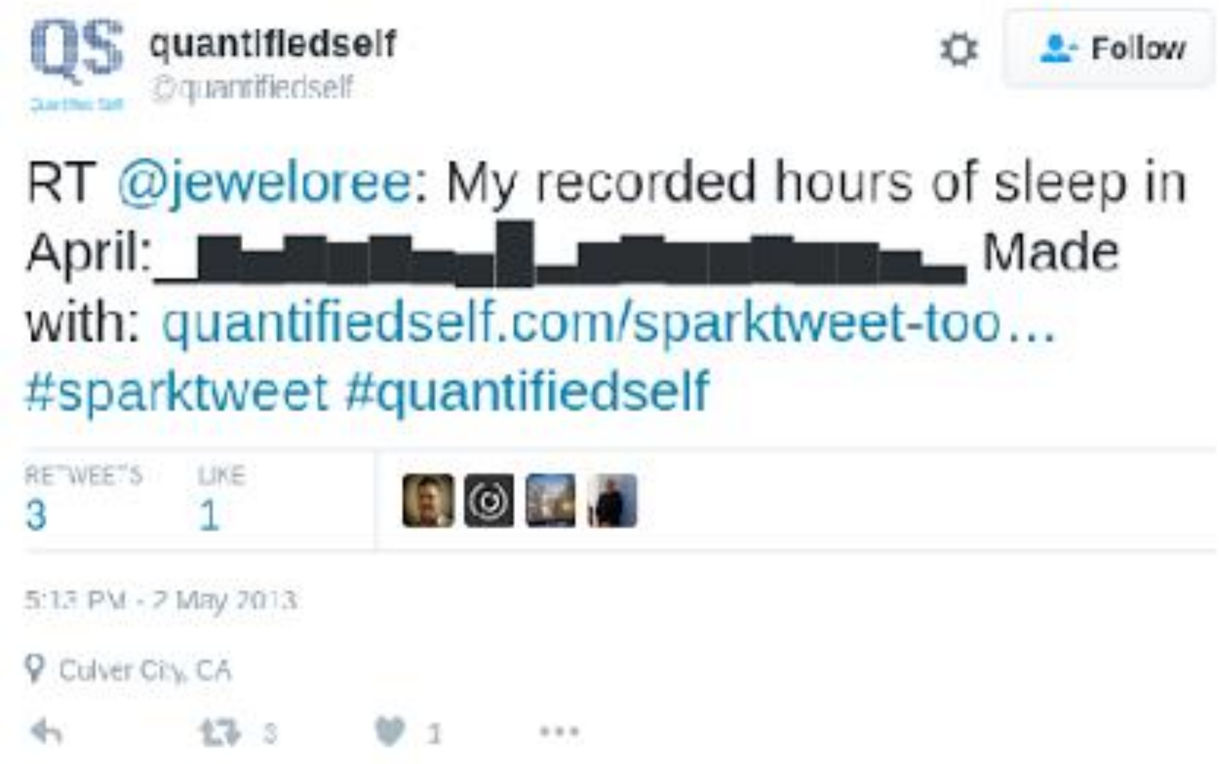


	A	B	C	D	E
1	Date	Apple	Intel	Amazon	IBM
2	10/18/2016	118.18	57.53	822.11	150.02
3	10/19/2016	117.25	57.47	820.4	151.27
4	10/20/2016	116.86	57.5	813.99	151.28
5	10/21/2016	116.81	60.28	809.36	150.58
6	10/24/2016	117.1	59.94	824.95	150.4
7	10/25/2016	117.95	60.85	839.3	150.69
8	10/26/2016	114.31	60.81	832.76	150.71
9	10/27/2016	115.39	60.61	831.24	152.82
10	10/28/2016	113.87	60.01	782	154.05
11	10/31/2016	113.65	60.16	781.03	152.76
12	11/1/2016	113.46	59.97	799	153.5
13	11/2/2016	111.4	59.82	783.93	152.48
14	11/3/2016	110.98	59.53	765.05	152.51
15	11/4/2016	108.53	58.65	762.79	152.4
16	11/7/2016	110.08	59.78	771.64	153.99
17	11/8/2016	110.31	60.55	784.97	154.56
18	11/9/2016	109.88	60	764	152.96
19	11/10/2016	111.09	60.48	778.81	157.66
20	11/11/2016	107.12	58.23	735.73	159.97
21	11/14/2016	107.71	59.02	745.51	161.25
22	11/15/2016	106.57	58.33	730	158.42
23	11/16/2016	106.7	58.94	739.88	158.46
24	11/17/2016	109.81	60.41	749.32	159.22
25	11/18/2016	109.72	60.78	761	159.8



	A	B	C	D	E
1	Date	Apple	Intel	Amazon	IBM
2	10/18/2016	118.18	57.53	822.11	150.02
3	10/19/2016	117.25	57.47	820.4	151.27
4	10/20/2016	116.86	57.5	813.99	151.28
5	10/21/2016	116.81	60.28	809.36	150.58
6	10/24/2016	117.1	59.94	824.95	150.4
7	10/25/2016	117.95	60.85	839.3	150.69
8	10/26/2016	114.31	60.81	832.76	150.71
9	10/27/2016	115.39	60.61	831.24	152.82
10	10/28/2016	113.87	60.01	782	154.05
11	10/31/2016	113.65	60.16	781.03	152.76
12	11/1/2016	113.46	59.97	799	153.5
13	11/2/2016	111.4	59.82	783.93	152.48
14	11/3/2016	110.98	59.53	765.05	152.51
15	11/4/2016	108.53	58.65	762.79	152.4
16	11/7/2016	110.08	59.78	771.64	153.99
17	11/8/2016	110.31	60.55	784.97	154.56
18	11/9/2016	109.88	60	764	152.96
19	11/10/2016	111.09	60.48	778.81	157.66
20	11/11/2016	107.12	58.23	735.73	159.97
21	11/14/2016	107.71	59.02	745.51	161.25
22	11/15/2016	106.57	58.33	730	158.42
23	11/16/2016	106.7	58.94	739.88	158.46
24	11/17/2016	109.81	60.41	749.32	159.22
25	11/18/2016	109.72	60.78	761	159.8
26					

Spark Lines



quantifiedself
@quantifiedself

Follow

RT @jeweloree: My recorded hours of sleep in April: ▁▆▅▅▆▄▂█▁▅▆▅▅▆▅▅▃ Made with: quantifiedself.com/sparktweet-too… #sparktweet #quantifiedself

RETWEETS 3 LIKE 1

5:13 PM - 2 May 2013

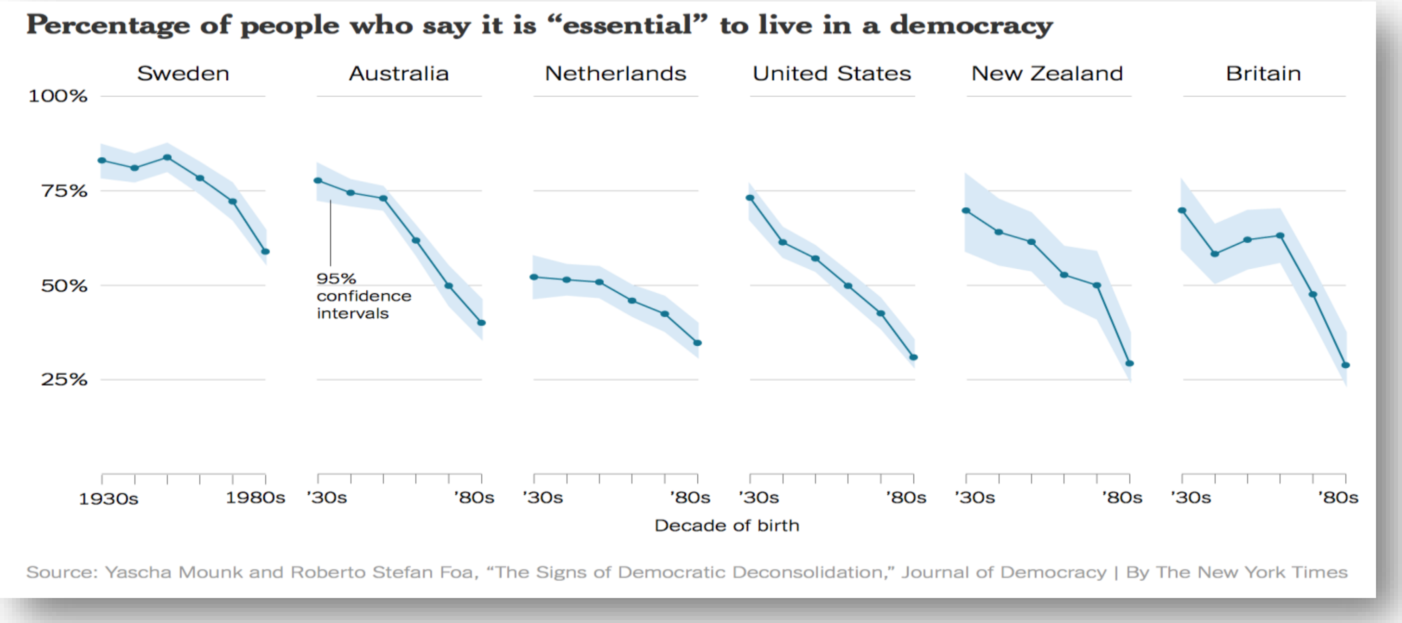
Culver City, CA

Bad data

- You are much more likely to make a good-looking, well-designed figure that misleads people because you have used it to display some bad data.
- Well-designed figures with little junk in their component parts are not by themselves a defense against cherry-picking your data.
- Indeed, it is even possible that, in a world where people are on guard against junky infographics, the “halo effect” accompanying a well-produced figure might make it easier to mislead some audiences.
- Good aesthetics does not make it much harder for you to mislead yourself as you look at your data.

A crisis of faith in democracy?*

- The graph reads as though people were asked to say whether they thought it was essential to live in a democracy.
- The results plotted show the percentage of respondents who said “Yes”, presumably in contrast to those who said “No”.

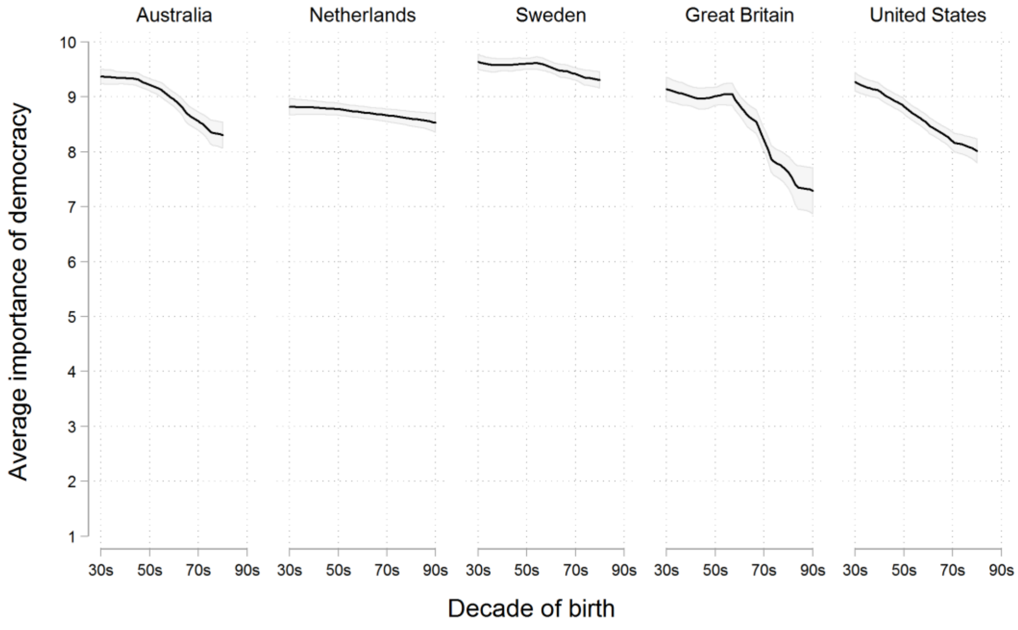


Source: Yascha Mounk and Roberto Stefan Foa, “The Signs of Democratic Deconsolidation,” Journal of Democracy | By The New York Times

*New York Times, 2016

Perhaps the crisis has been overblown*

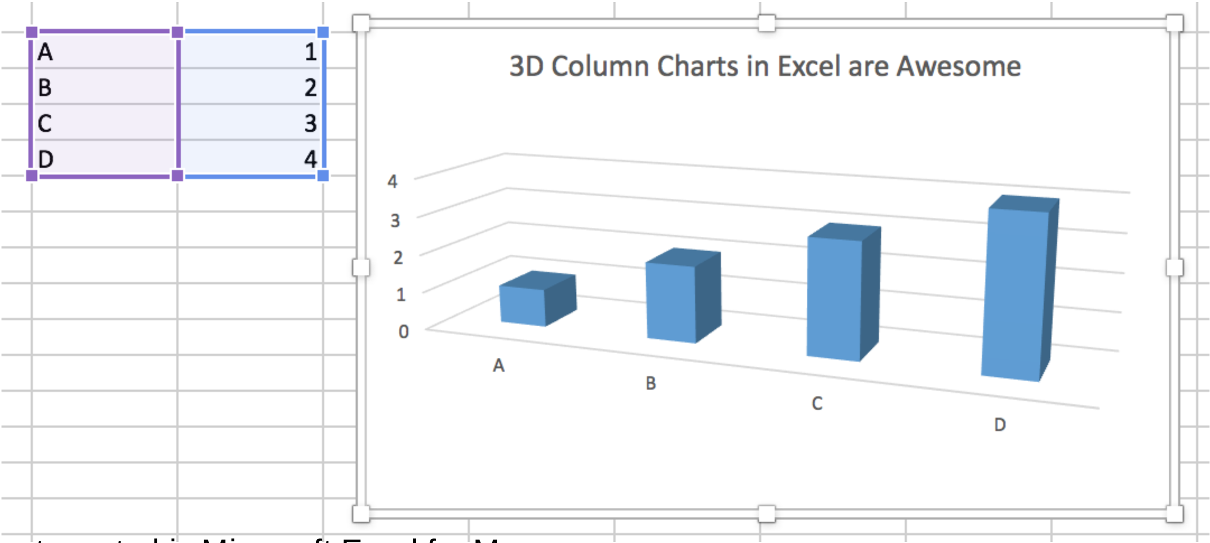
- In fact the survey question asked respondents to rate the importance of living in a democracy on a ten point scale, with 1 being “Not at all Important” and 10 being “Absolutely Important”.
- Figure redrawn with average values.



Graph by Erik Voeten, based on WVS 5

Bad perception

- Visualizations encode numbers in lines, shapes, and colors.
- Our interpretation of these encodings is partly conditional on how we perceive geometric shapes and relationships generally.

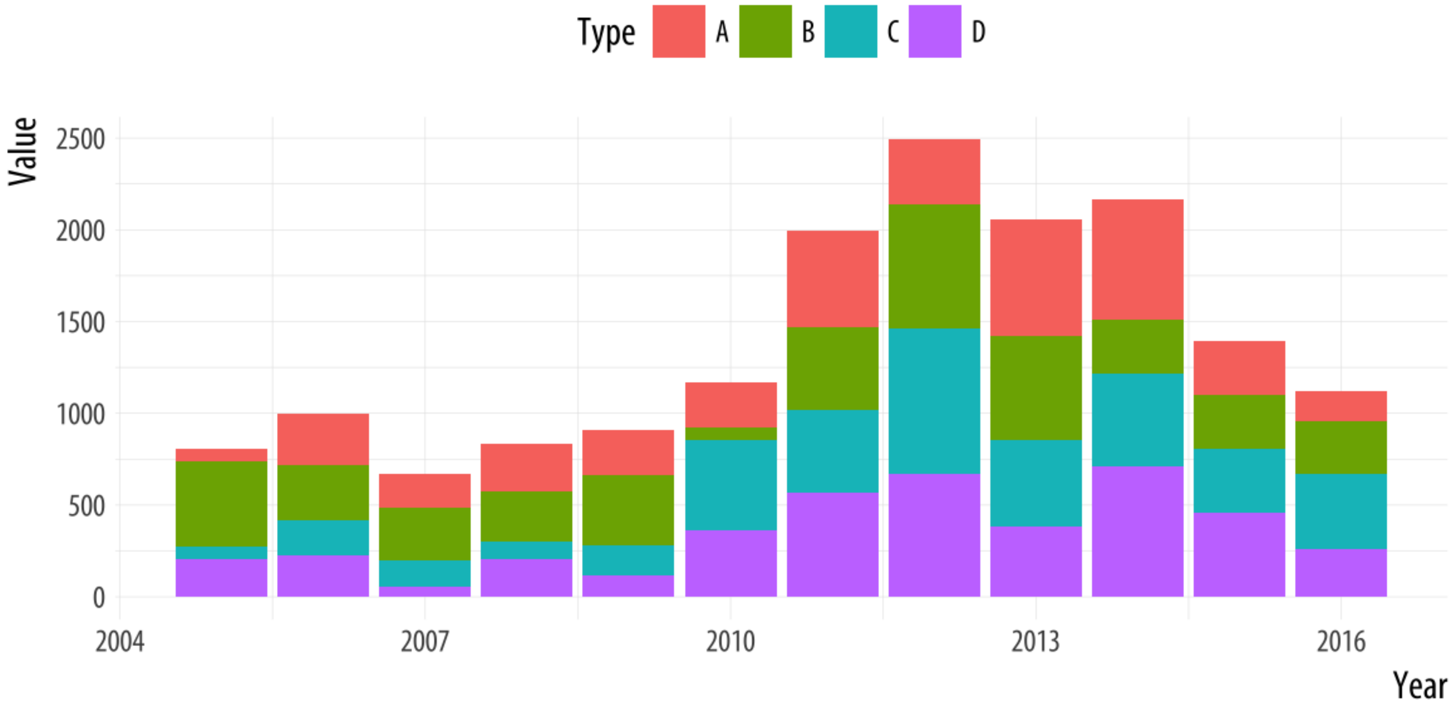


A	1
B	2
C	3
D	4

A 3-D Column Chart created in Microsoft Excel for Mac.
Although it may seem hard to believe, the values shown in the bars are 1, 2, 3, and 4.

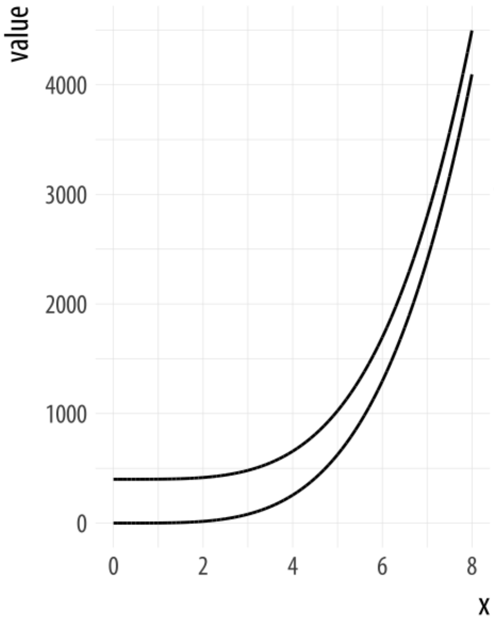
Relative comparisons need a stable baseline

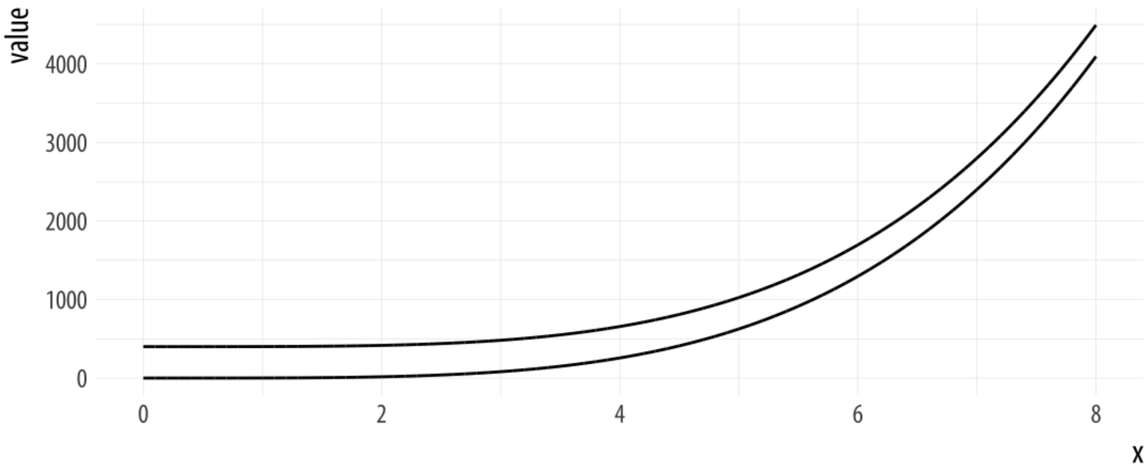
- The overall trend is readily interpretable.
- Easy to follow the category closest to the x-axis baseline (D).
- Other categories are not so easily grasped.



Different data?

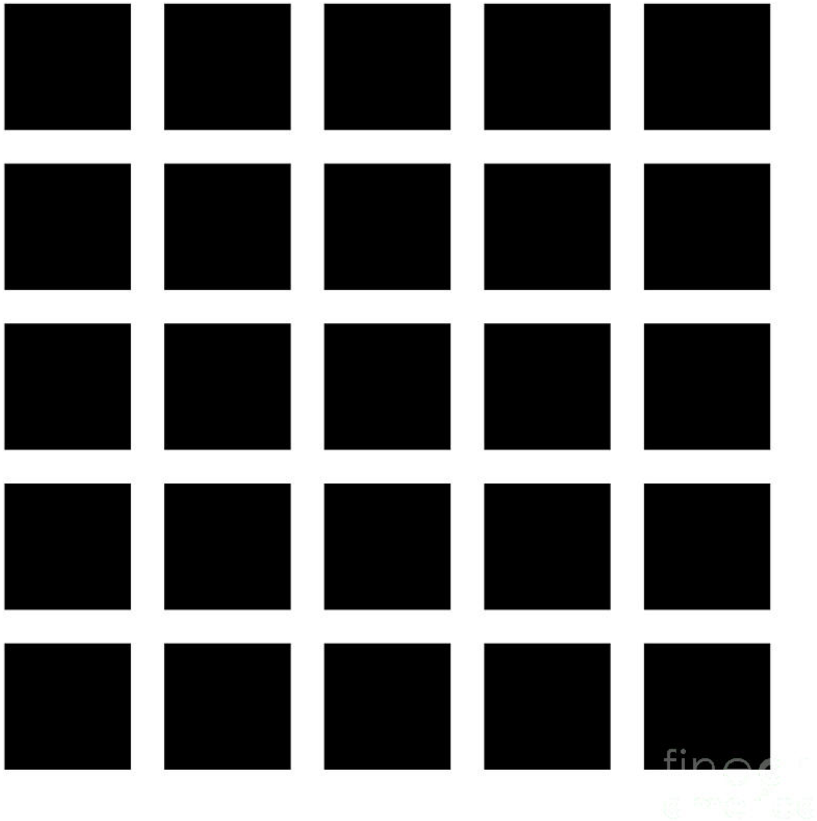
- Aspect ratio!





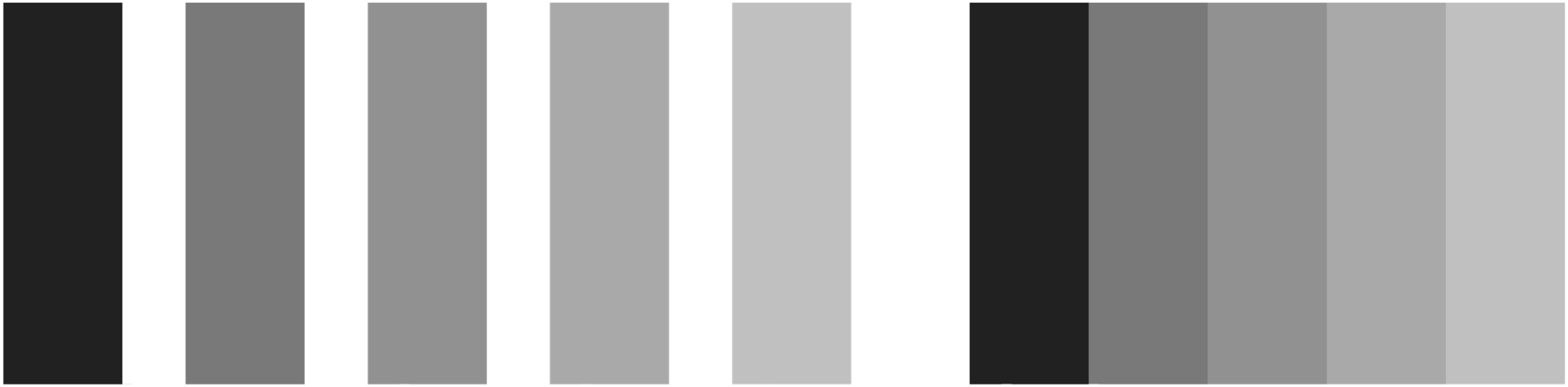
Perception and data visualization-1

- Hermann Grid Effect
- Ghostly blobs seem to appear at the intersections



Perception and data visualization-2

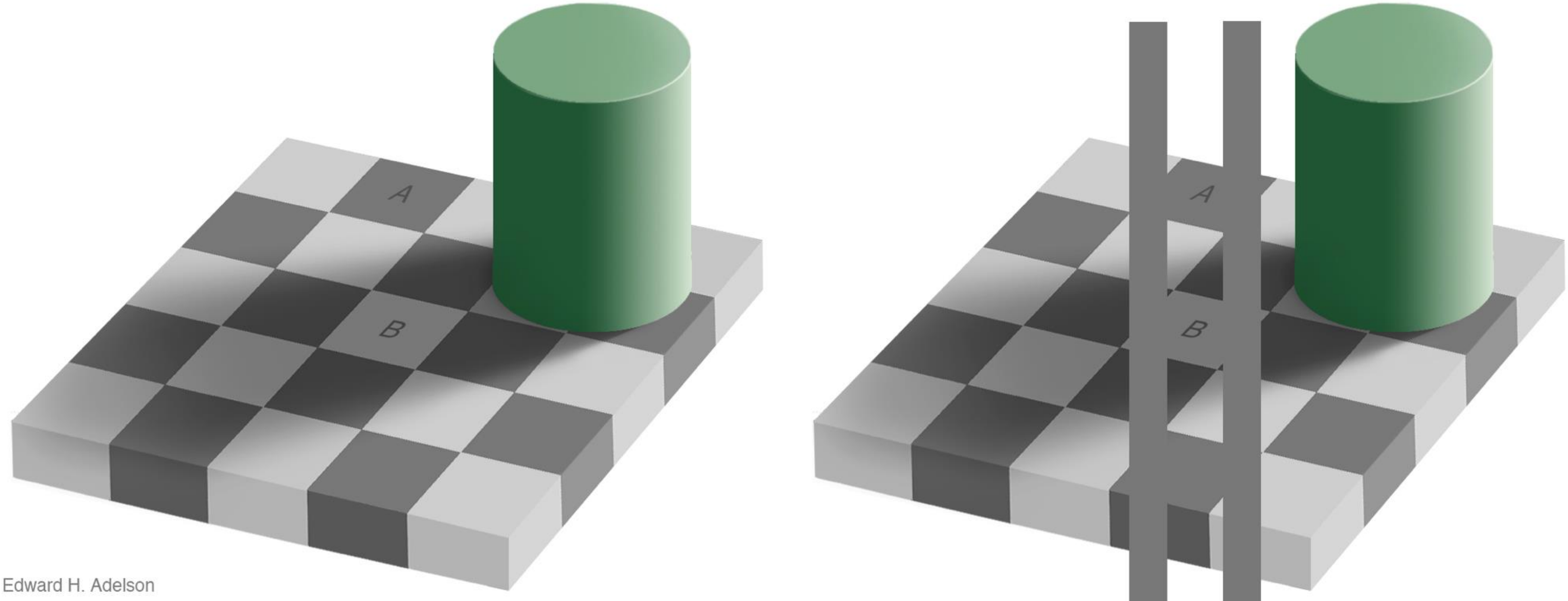
- Mach bands
- Our visual system is trying to construct a representation of what it is looking at based more on relative differences in the luminance (or brightness) of the bars



Perception and data visualization-3

A and B are the same color?

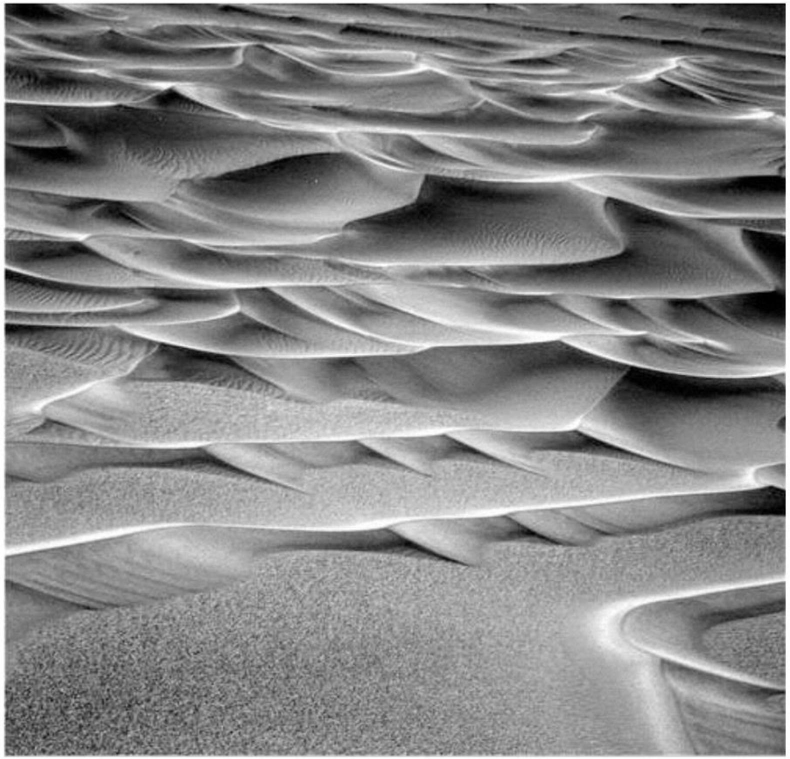
- Our visual system is attracted to edges, and we assess contrast and brightness in terms of relative rather than absolute values. To figure out the shade of the squares on the floor, we compare it to the nearby squares, and we also discount the shadows cast by other objects.

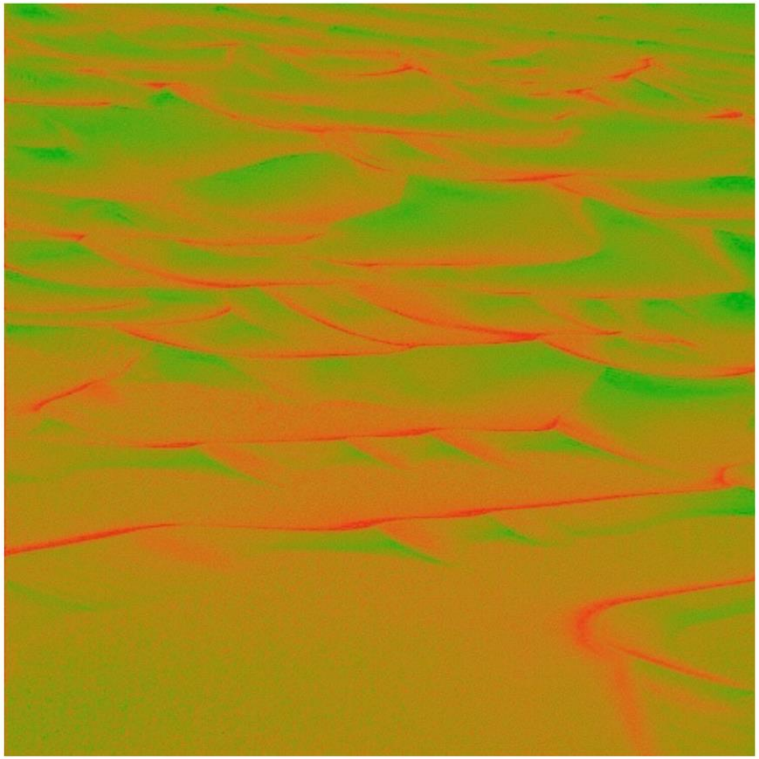


Edward H. Adelson

Perception and data visualization-4

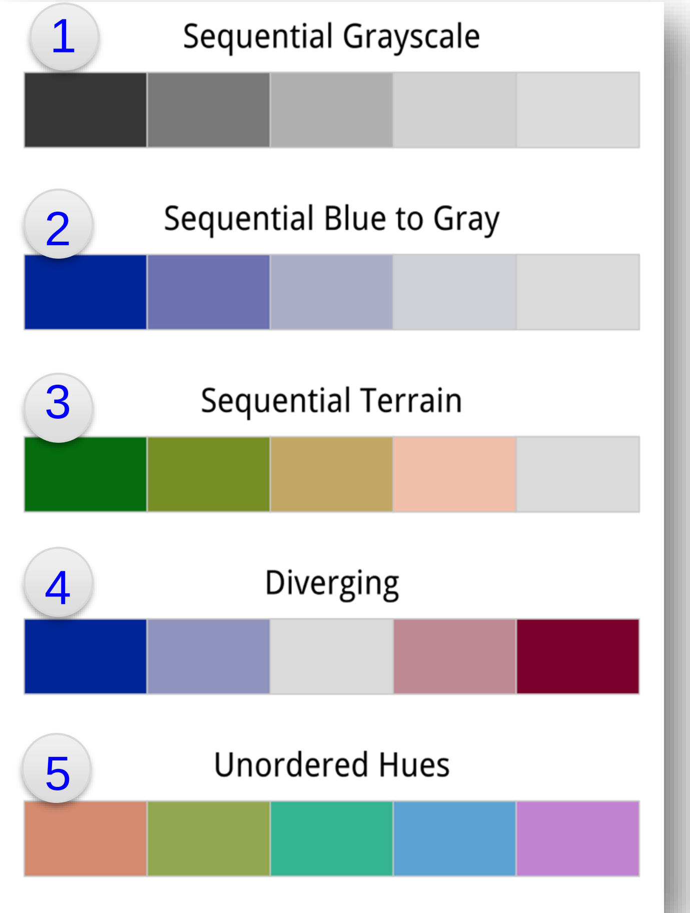
- Our ability to see edge contrasts is stronger for monochrome images than for color.
- In the grayscale version, the dunes and ridges are much more easily visible.





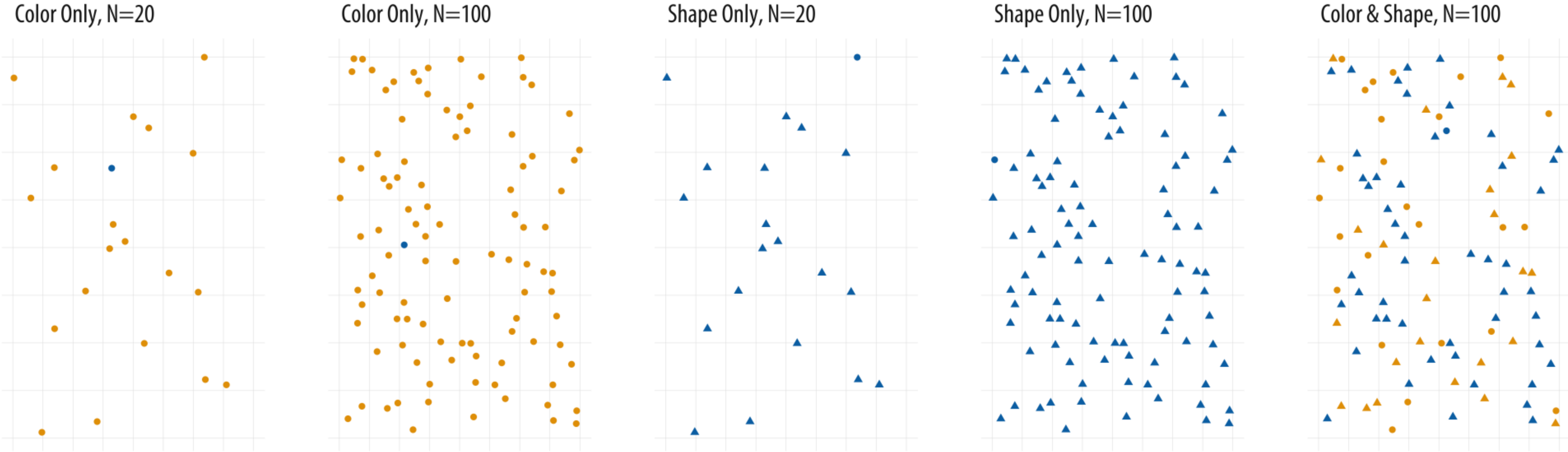
Perception and data visualization-4

- How to represent or encode data then?
- Need colors that are not just numerically but also perceptually uniform.
- Five sample palettes from R library
  1. *Varies only in luminance, or brightness*
  2. *Varies in both luminance and chrominance*
  3. *Varies in luminance, chrominance, and hue.*
  4. *Diverging, with a neutral midpoint.*
  5. *Balanced hues, suitable for unordered categories.*
- Good News: palettes are readily available.



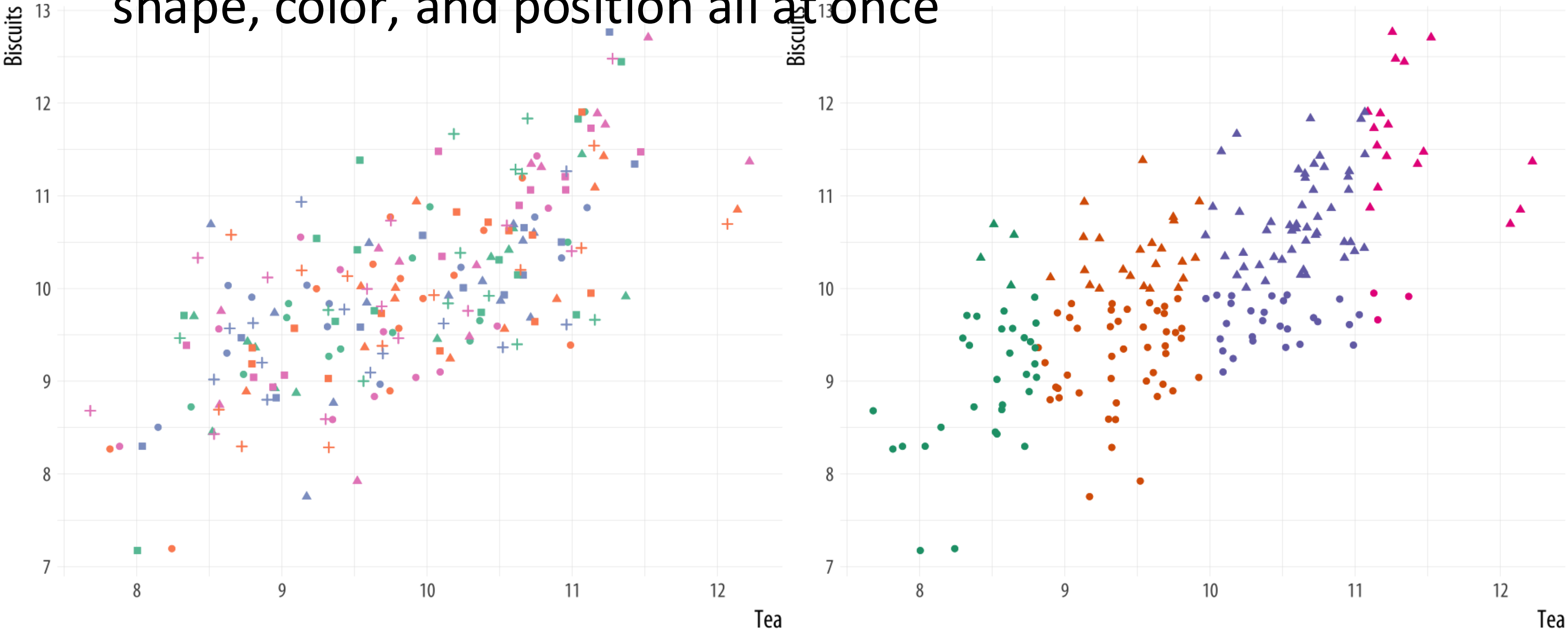
Preattentive pop-out

- Some objects in our visual field are easier to see than others.
- From our point of view it happens before or almost before the conscious act of looking at or for something
- Searching for the blue circle becomes progressively harder
- Think of shape and color as two distinct channels that can be used to encode information visually.



Multiple channels can be overtaxing

- Adding multiple channels to a graph is likely to overtax the capacity of the viewer very quickly.
- Even if our software allows us to, we should think carefully before representing different variables and their values by shape, color, and position all at once

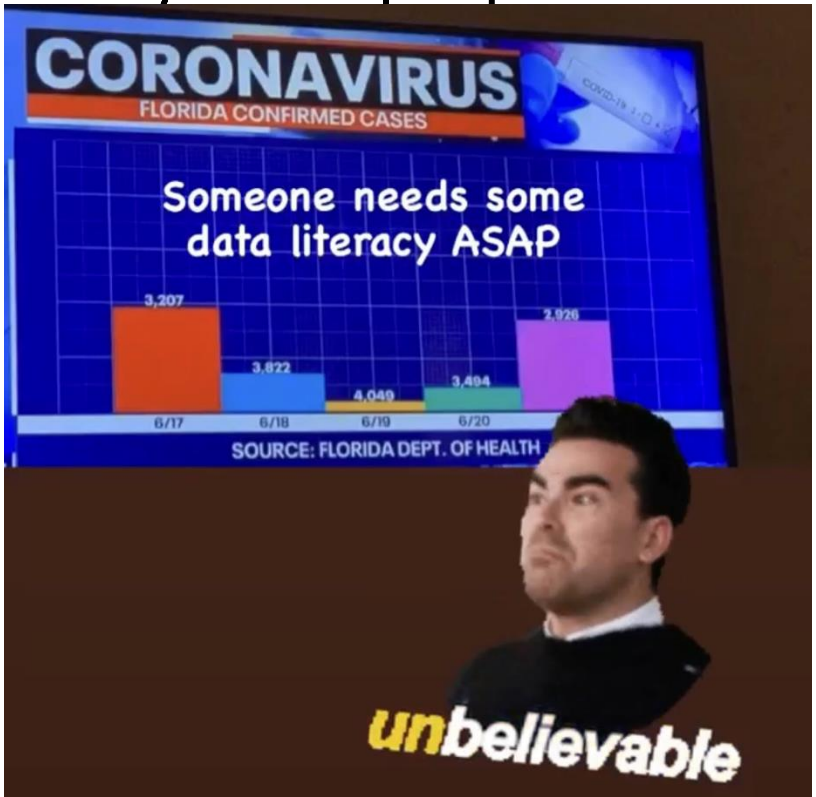


Lie factor-perception-ethics

- Can visualizations lie?

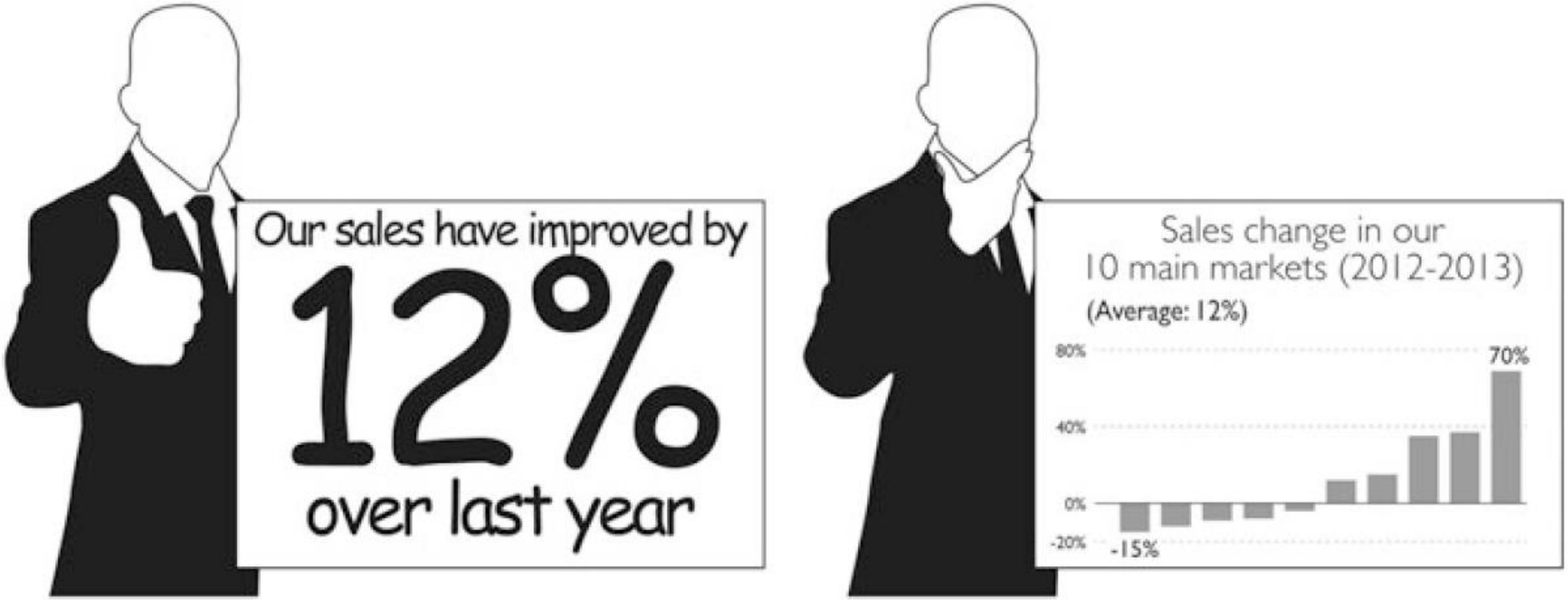
—Yes (Lie factor)

- Either mistakenly or on purpose.

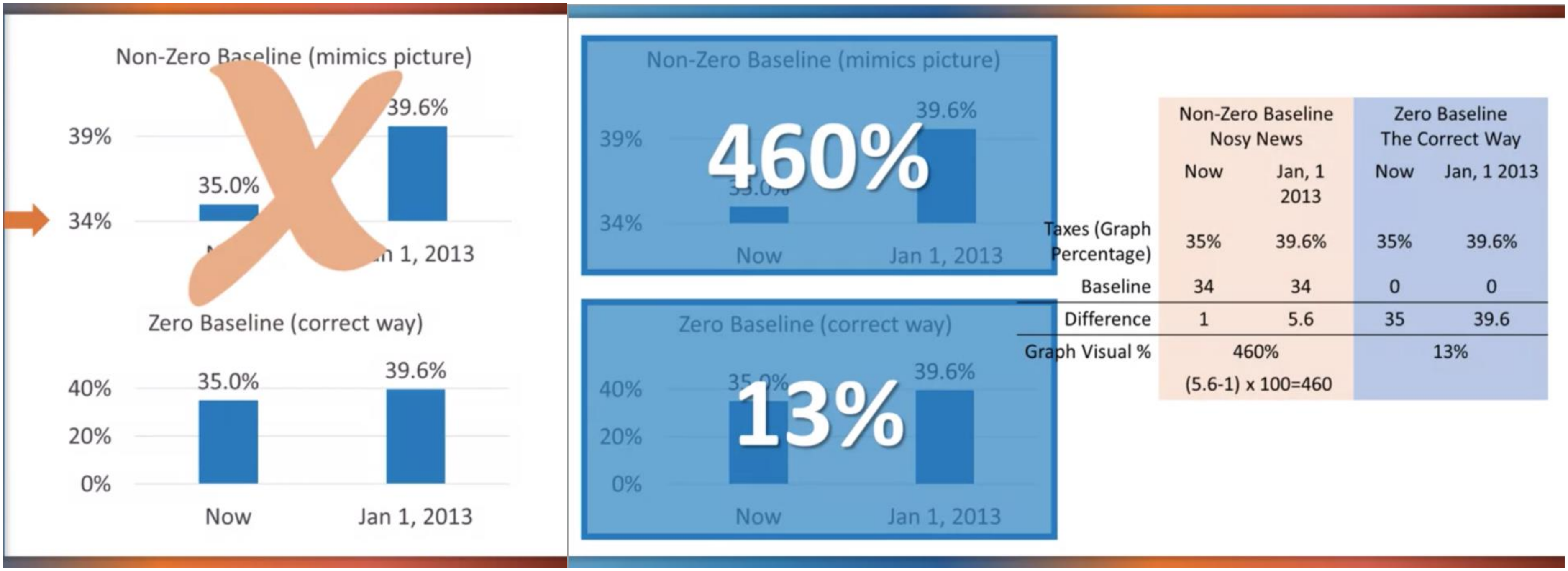


Lie factor-perception-ethics

- Information hiding and deception: not ethical, but plenty of examples, unfortunately.



Ethical issues in visualization

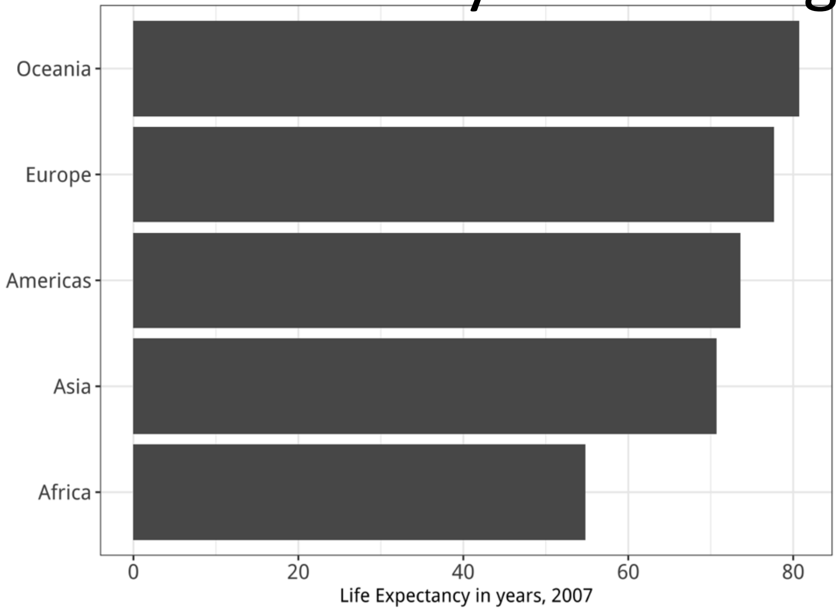


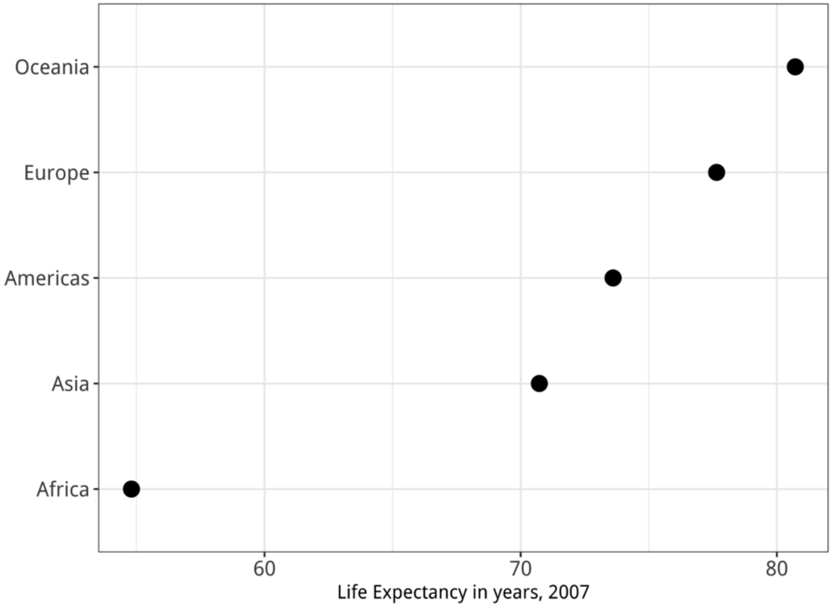
	Non-Zero Baseline Nosy News		Zero Baseline The Correct Way	
	Now	Jan, 1 2013	Now	Jan, 1 2013
Taxes (Graph Percentage)	35%	39.6%	35%	39.6%
Baseline	34	34	0	0
Difference	1	5.6	35	39.6
Graph Visual %	460%		13%	
	(5.6-1) x 100=460			

- Start your baseline at zero.
- Do not confuse your audience.

Ethical issues in visualization

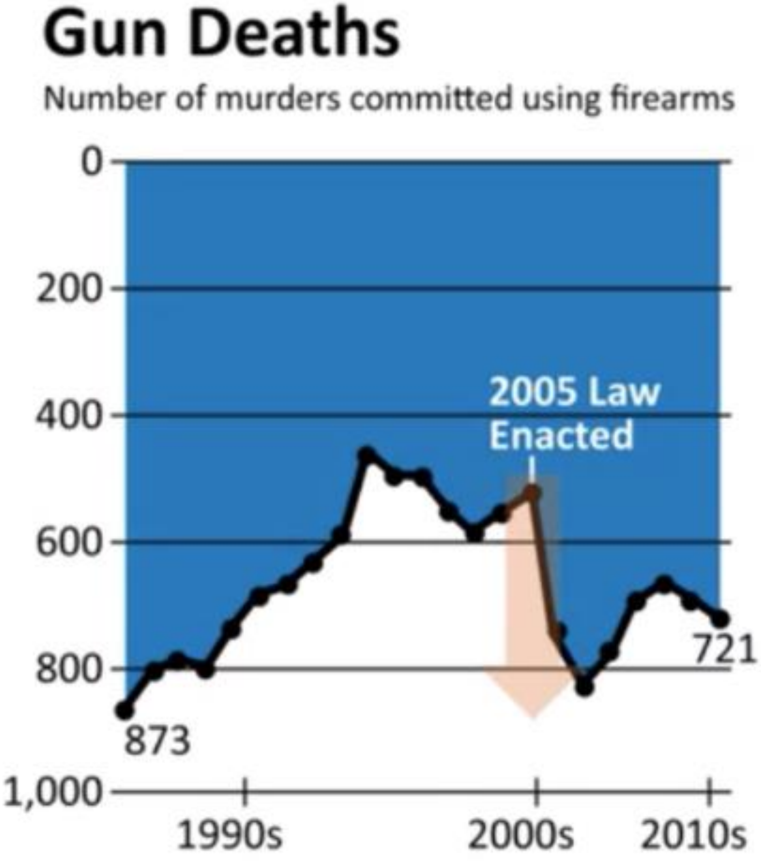
- Problems of honesty and good judgment
- Being honest with your data is a bigger problem than can be solved by rules of thumb about making graphs
- It would be a mistake to think that a dot plot deliberately misleading due to a different baseline.





Ethical issues in visualization

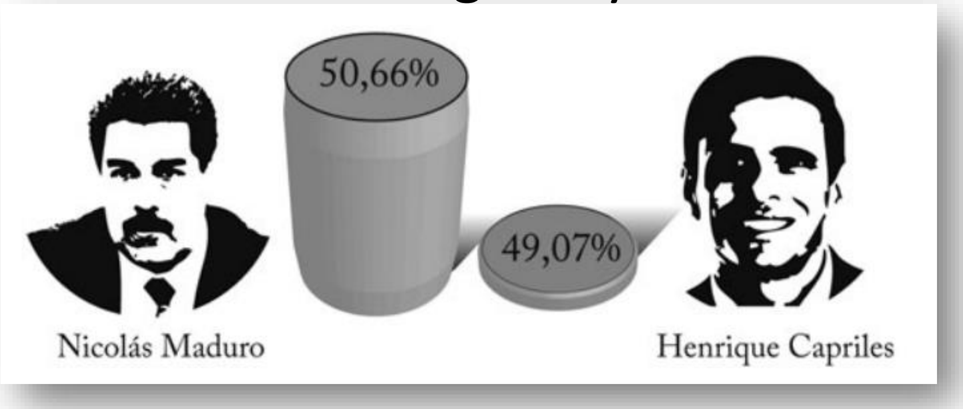
- Did you assume that there was a sharp decrease in gun death after 2005?
  - *A. Yes*
  - *B. No*



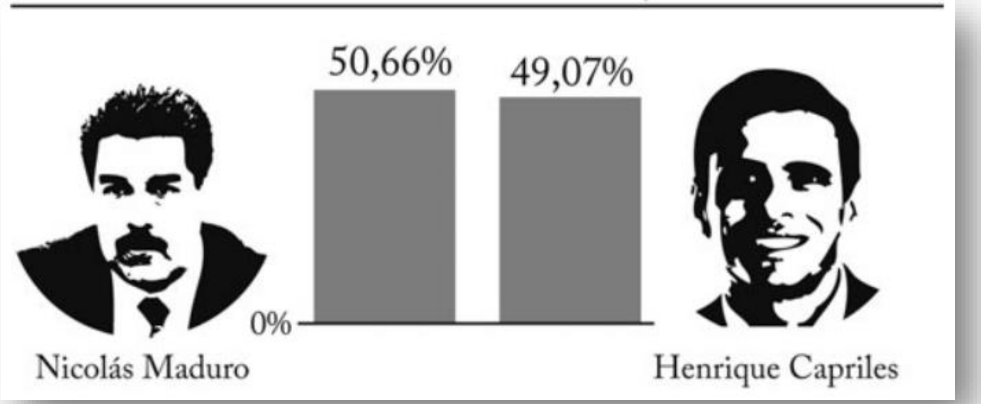
Ethical issues in visualization

Presidential election results in Venezuela in 2013.

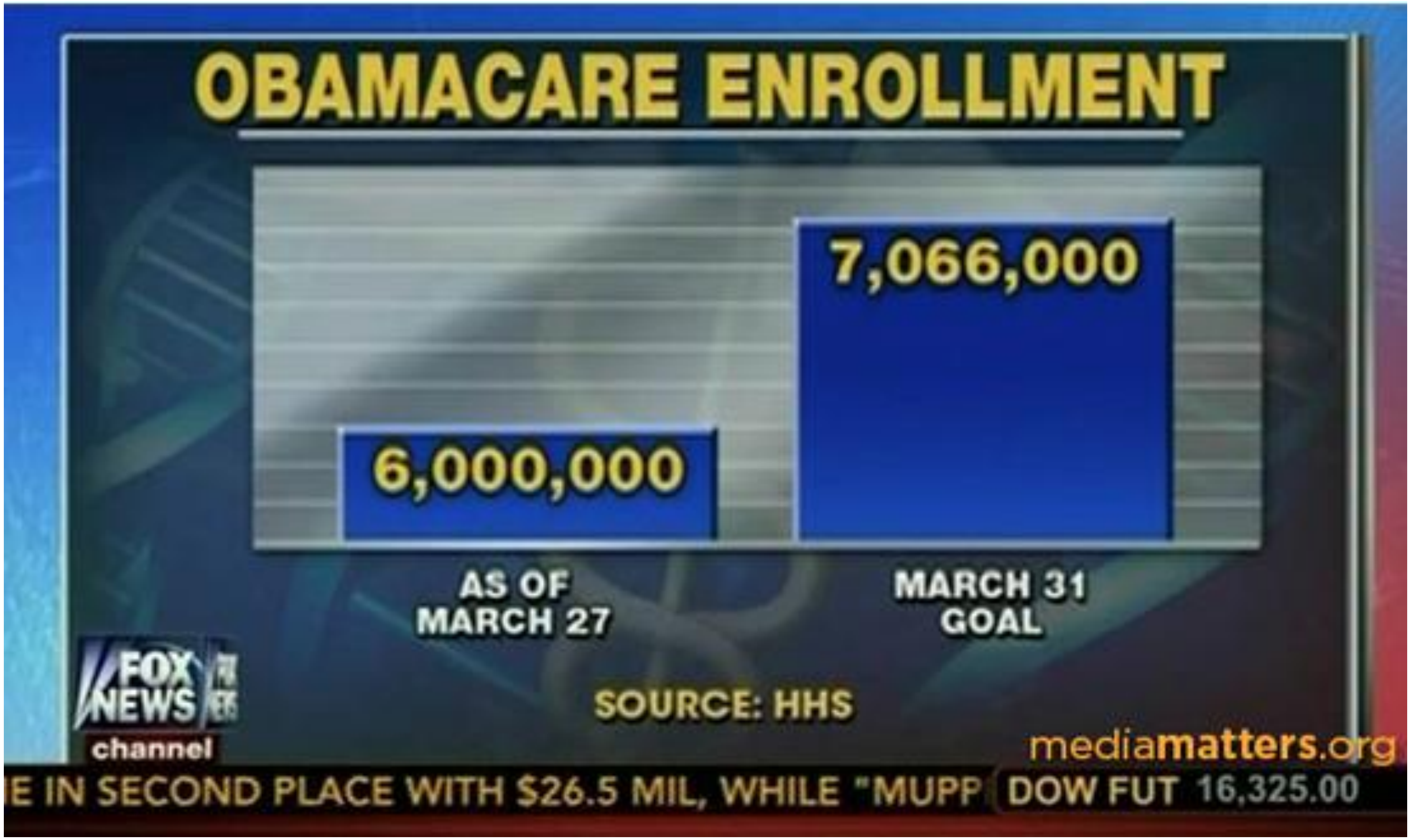
- The truncated Y-axis which greatly distorts the difference.



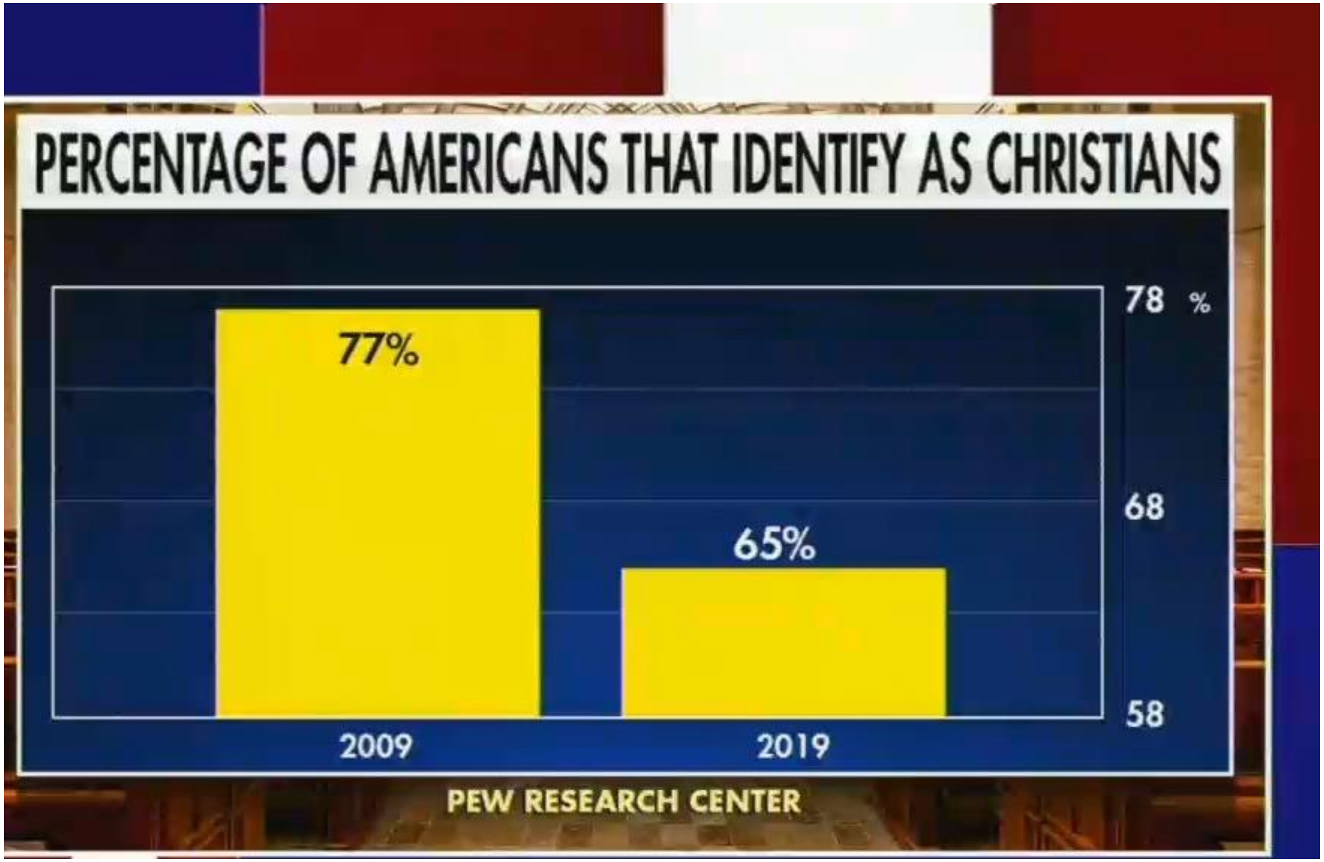
- 0-baseline has been added, and the 3D effect has been removed



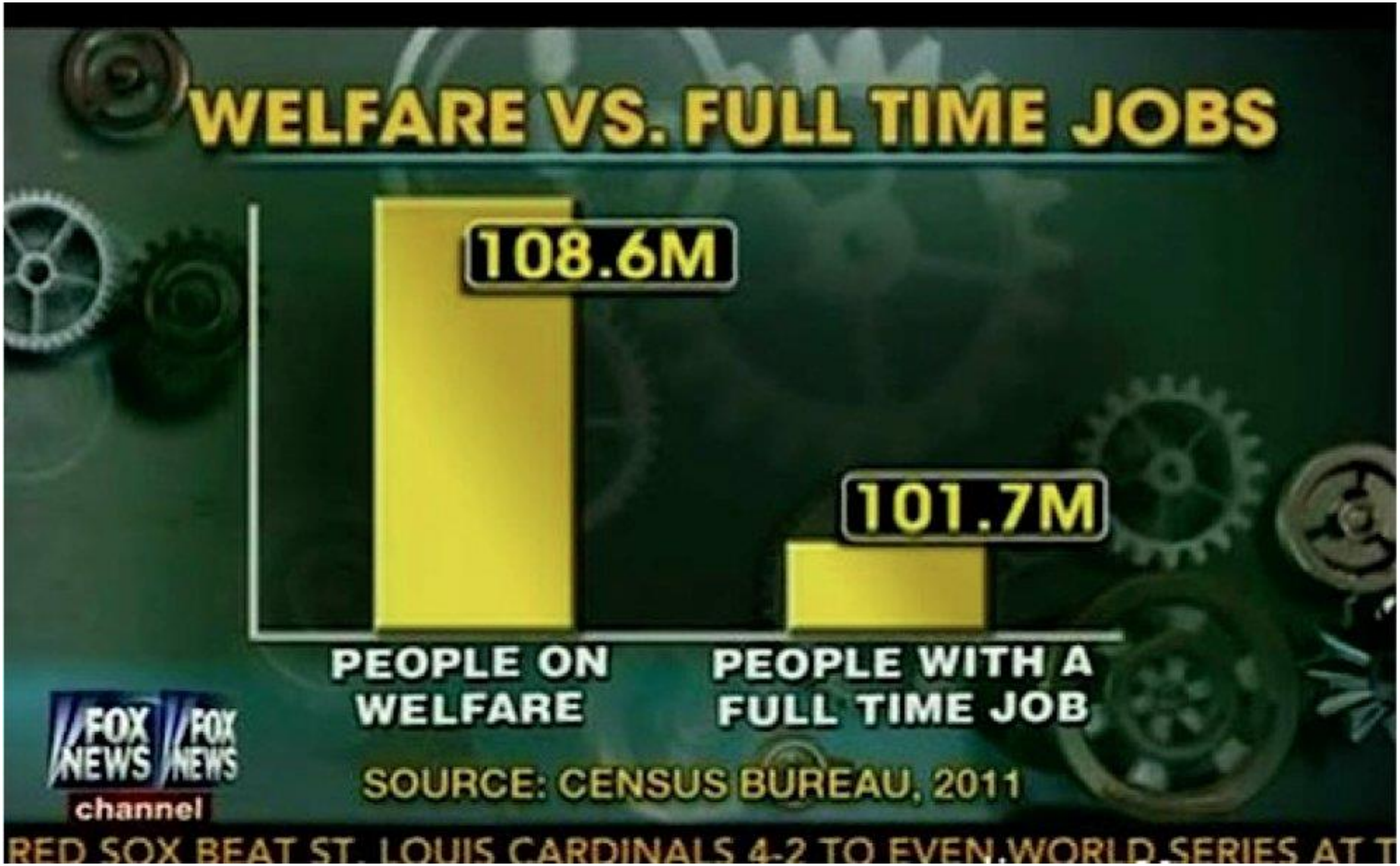
Ethical issues in visualization



Ethical issues in visualization



Ethical issues in visualization


WELFARE VS. FULL TIME JOBS
108.6M
101.7M
PEOPLE ON WELFARE
PEOPLE WITH A FULL TIME JOB
SOURCE: CENSUS BUREAU, 2011
FOX NEWS FOX NEWS channel
RED SOX BEAT ST. LOUIS CARDINALS 4-2 TO EVEN WORLD SERIES AT T