DS4200: Information Presentation and Visualization

Introduction

Xiaoyi Yang
Khoury College of Computer Sciences
Northeastern University

Xiaoyi Yang, PhD

- BS in Statistics and Math at University of Wisconsin-Madison
- Master and PhD in Statistics at Carnegie Mellon University
- Tenure-track Assistant Professor at Creighton Univeristy
- Thesis: *Learning social networks from text data using covariate information*
- Teaching: *Information presentation and visualization*
  Foundations of Data Science
- New faculty, still learning and happy to communicate!

TAs

Online TA (Piazza and online office hour)
- Soumyae Tyagi: tyagi.so@northeastern.edu
- Yujin Park: park.yuj@northeastern.edu

Quiz TA (Quiz grading and onsite office hour)
- Jing Cheng: cheng.jing2@northeastern.edu
- Mia Khan: khan.mia@northeastern.edu

Homework TA (Homework grading and onsite office hour)
- Eduardo Puerta: puerta.e@northeastern.edu
- Wo Wei (Willy) Lin: lin.wo@northeastern.edu

TA office hours

Office hours will start from next week (Sep 11th)
Tentative schedule will be posted on Canvas announcement and Piazza

Onsite office hour: Snell Library 047 for Thursday 6-8 PM and Snell Library 009 for Friday 1-5 PM and 6-8 PM.

Online office hour: Wednesday and Thursday 7-10 PM, Saturday 2-4 PM, Friday 9-11 AM
zoom link: https://northeastern.zoom.us/j/98219149610

Any changes of the TA office hour will be posted on Piazza.

What will be in this course

- Discuss the relations between human, computer and data
- Learn the options and guidelines to design a good visualization
- Require programming in Python, JavaScript, HTML, and CSS.
- Learn some basic use of ArcGIS and Tableau. Discuss the pros and cons for AI use in the data visualization.
- Requires extensive writing.
- By the end of this course, you should be able to
  - Analysis a given data and make appropriate visualizations
  - Comment on given visualizations
  - Design your own website

Syllabus

Outline

- What is data visualization
- Good and bad data visualization examples
- Design rules of thumb
- Tufte's "Graphical Integrity" principles
- A fuzzy gray area of interpretation and opinion on integrity (e.g., role of pictograms)
- Reading for this week: chapter 1 and 6

What is visualization?

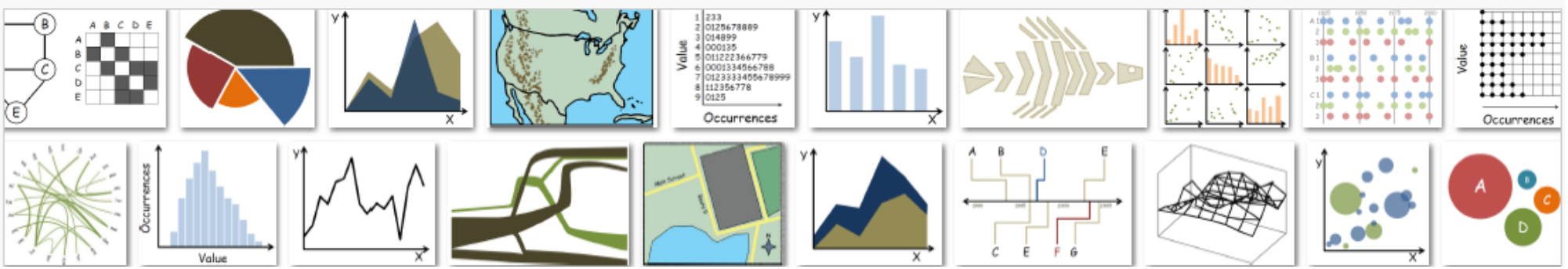
What is visualization?

visualization:the visual representation of data to reinforce human cognition

What is visualization?

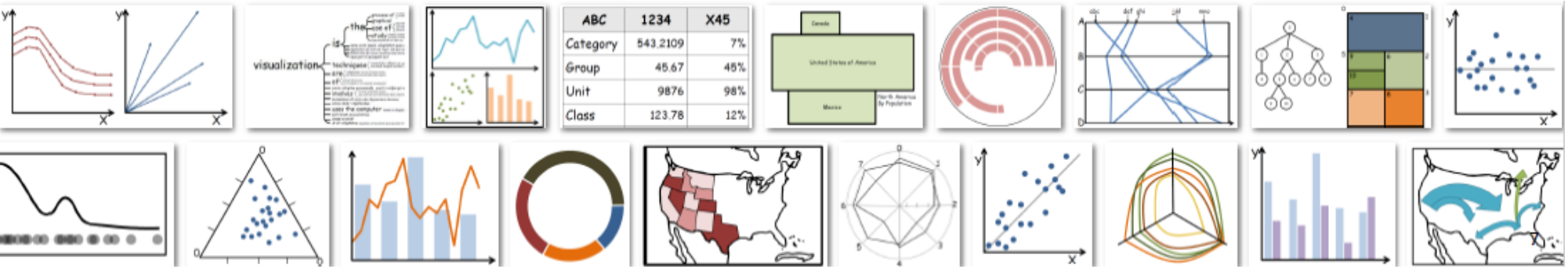
visualization:the (static or interactive) visual representation of (abstract or spatial) data to reinforce human cognition

Visualization

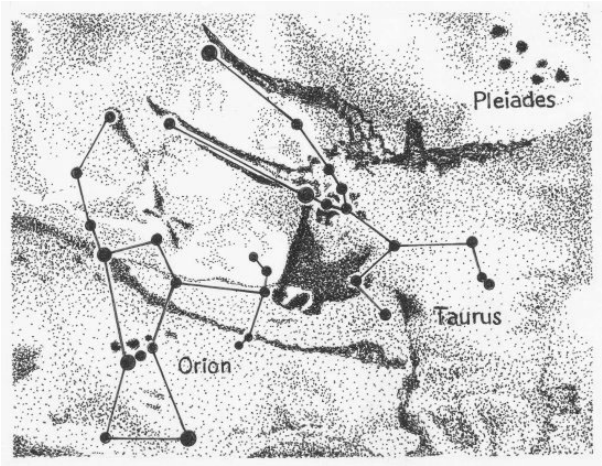


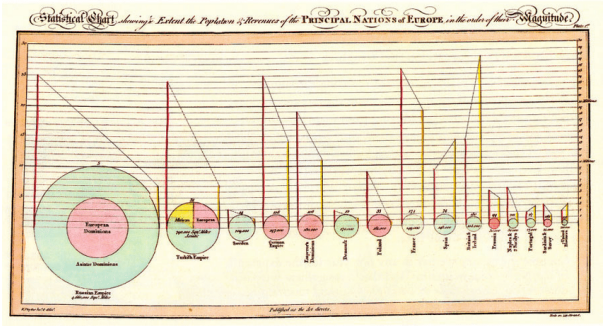
What is visualization?

visualization:the (static or interactive) visual representation of (abstract or spatial) data to reinforce human cognition



History of Visualization

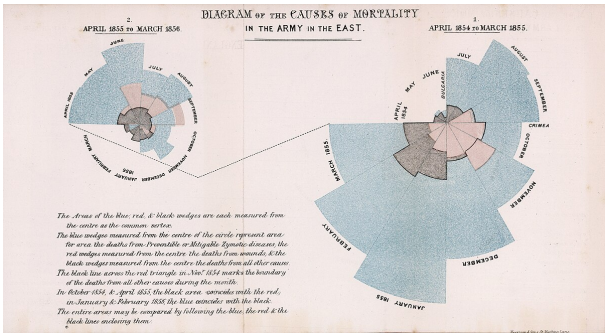




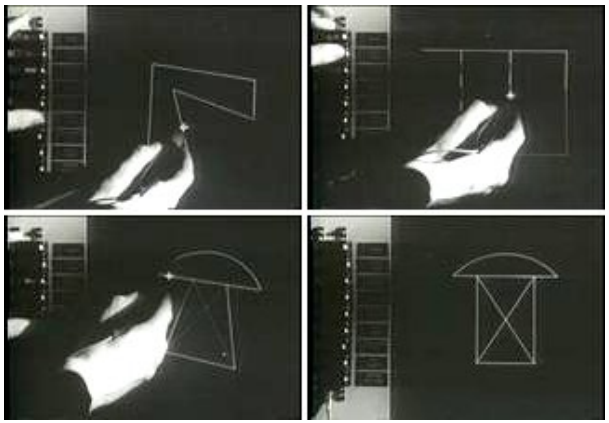
phonograms/pictographs Cartography Playfair

https://www.youtube.com/watch?v=PIjw6UGmthc

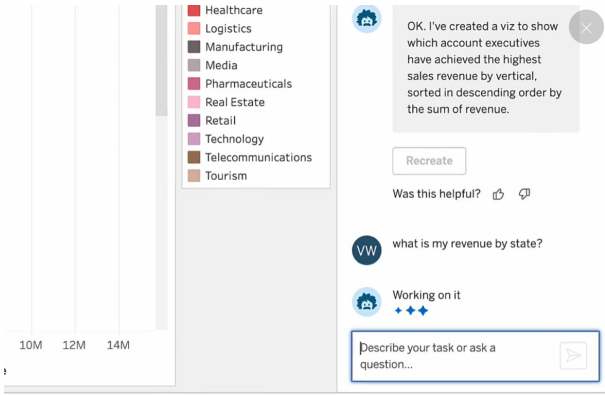
History of Visualization



coxcomb diagram



Sketchpad



AI-assisted

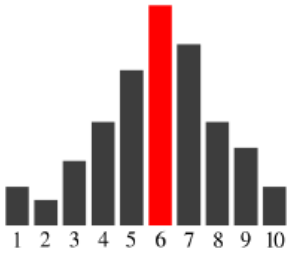
Visualization related terms

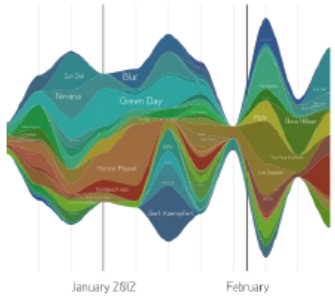
Any terms that may related to visualization?

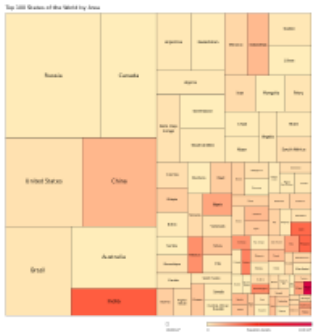
Visualization related terms

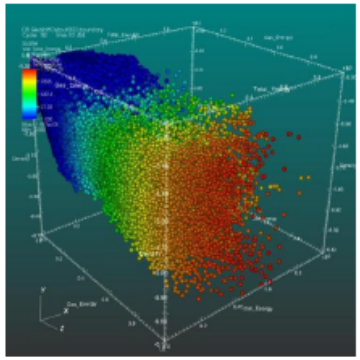
Any terms that may related to visualization?

- Computer graphics
- Computer vision
- Human computer interaction
- Design
- Art
- Statistics
- Psychology









Motivation

Ok, but why do we need visualization?

In class Sketching: "Three numbers"

3,8,36

- Break out into groups of 3 students.
- Together Sketch as many possible visualizations as you can of these three numbers.
- As a class we will discuss some of the designs and themes.
- Save your sketch! You need to upload it to the weekly quiz!

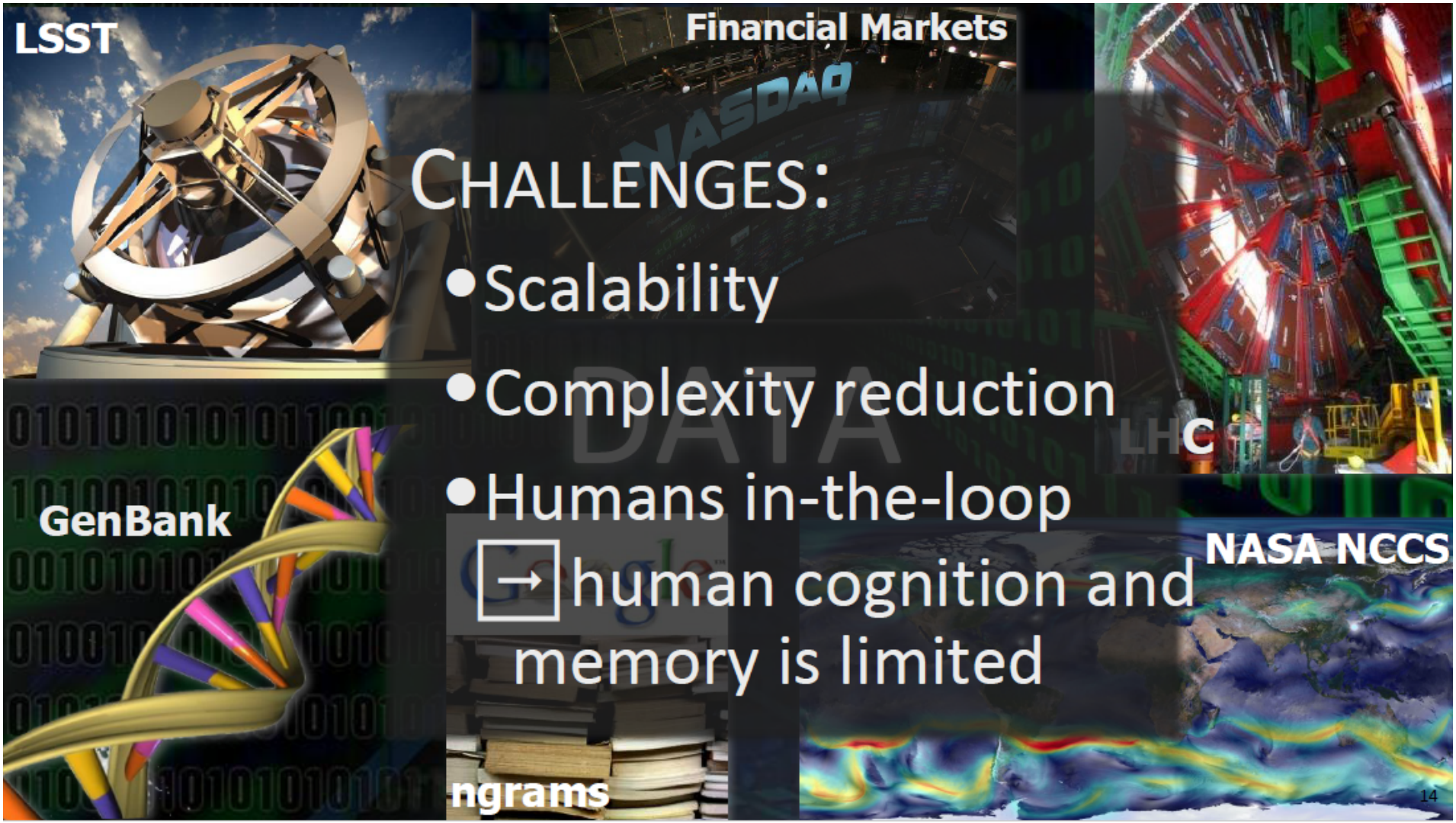
Another example: random letters

H, U, S, Q, H, D, N, X, N, A, P,
O, E, I, F, S, U, T, U, W, B, I, K,
P, J, Y, T, M, N, S, I, P, S, T

Another example: random letters

H, U, S, Q, H, D, N, X, N, A, P,
O, E, I, F, S, U, T, U, W, B, I, K,
P, J, Y, T, M, N, S, I, P, S, T

Motivation


LSST
Financial Markets
NASDAQ
CHALLENGES:
• Scalability
• Complexity reduction
• Humans in-the-loop
→ human cognition and memory is limited
GenBank
ngrams
NASA NCCS
14

Numerical example

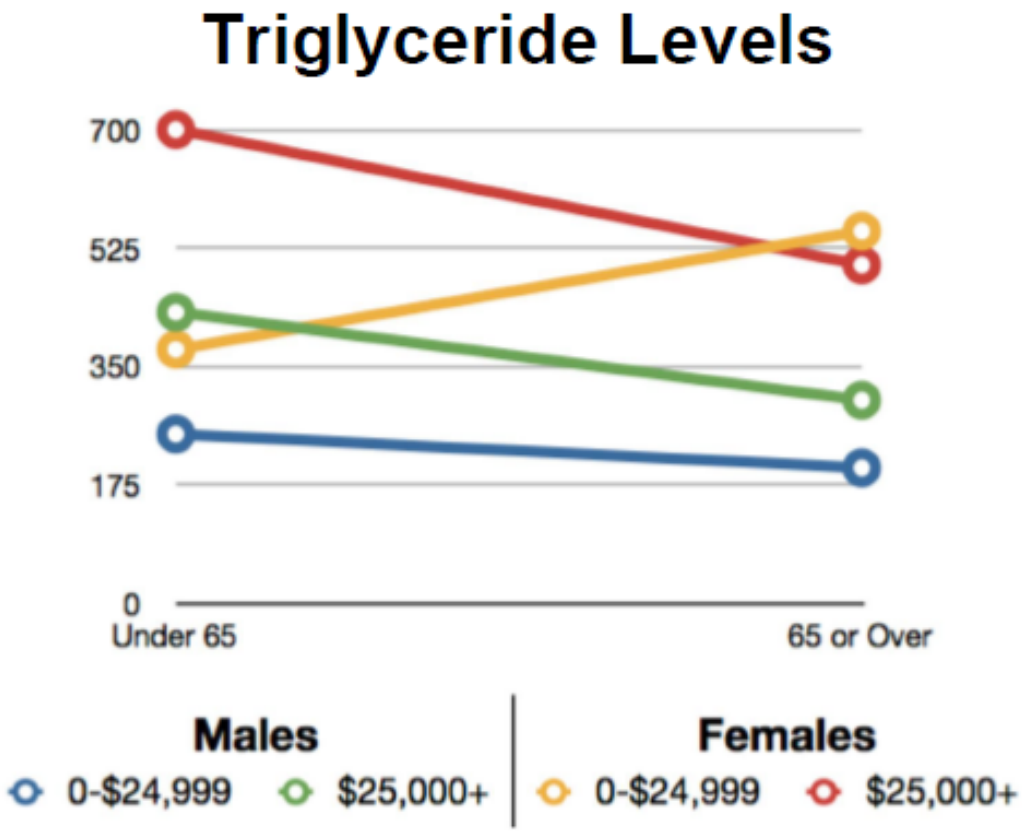
Which gender and income level shows a different effect of age on triglyceride levels?

Triglyceride Levels

	Males		Females	
Income Group	Under 65	65 or Over	Under 65	65 or Over
0-$24,999	250	200	375	550
$25,000+	430	300	700	500

Numerical example

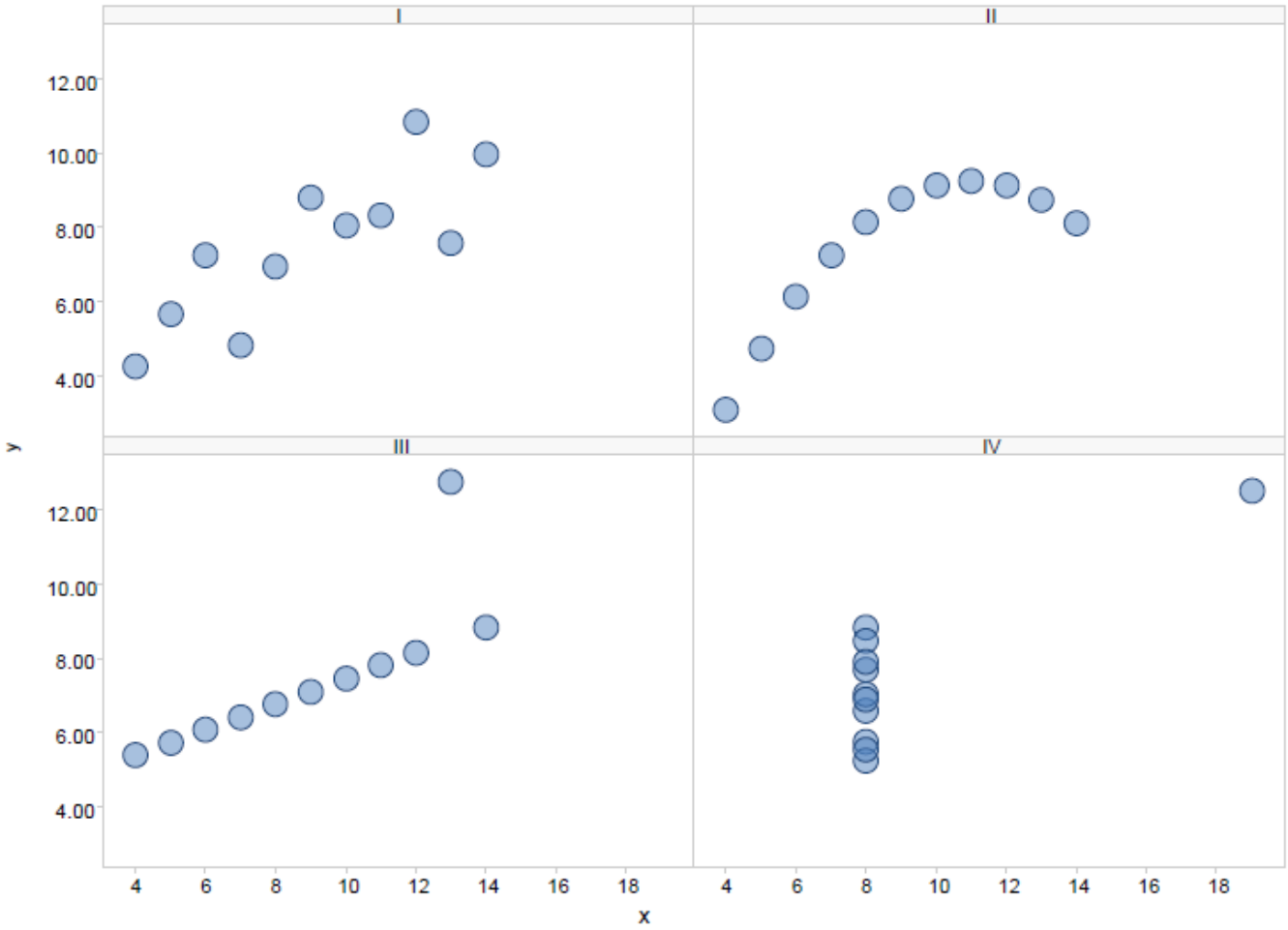
Which gender and income level shows a different effect of age on triglyceride levels?



Numerical example

I		II		III		IV	
x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

Numerical example

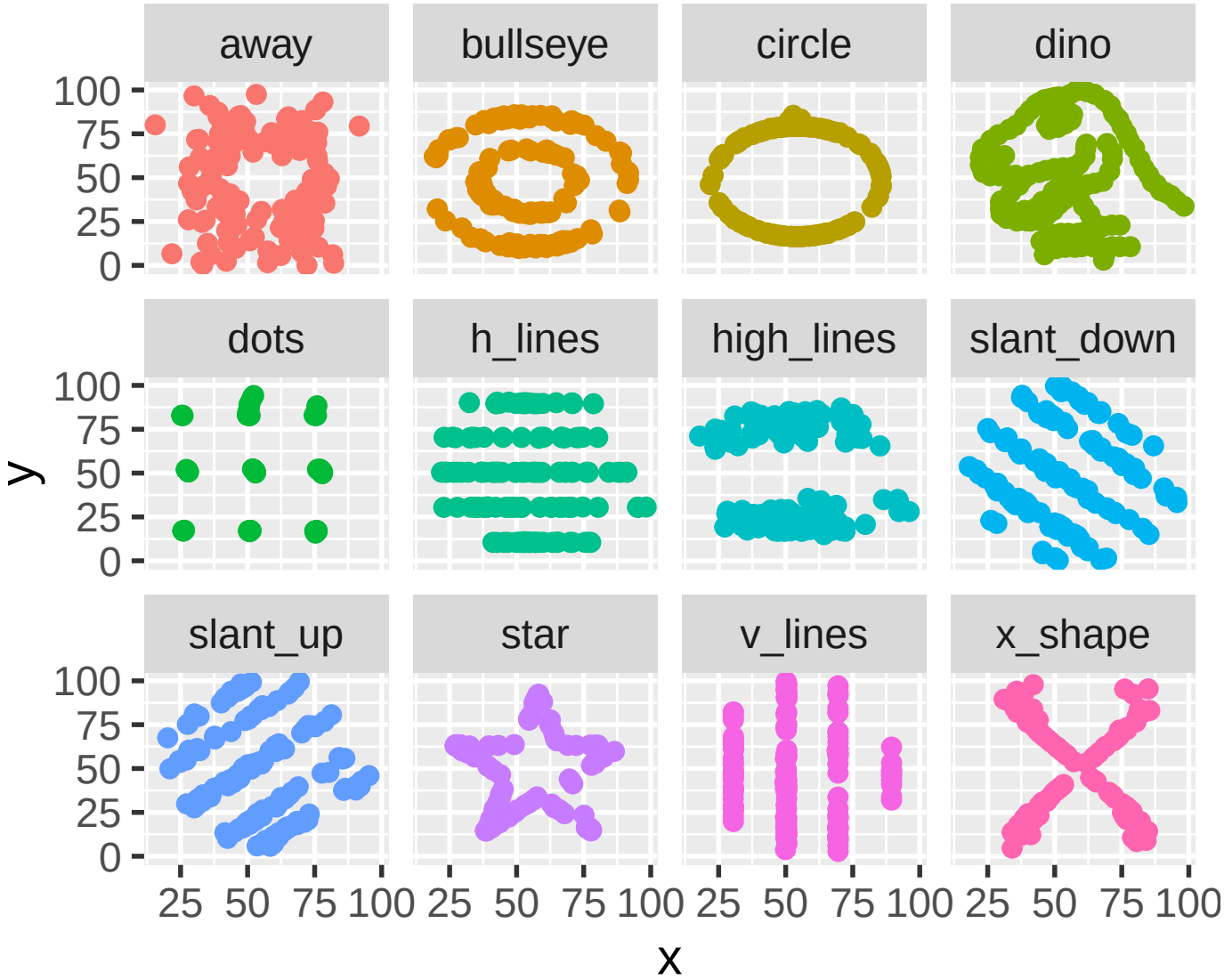


Numerical example

If I tell you the correlation between two variables is 0....

Numerical example

If I tell you the correlation between two variables is 0....



There are three types of lies: lies, damned lies, and statistics

- Mark Twain (maybe)

No catalogue of techniques can convey a willingness to look for what can be seen, whether or not anticipated. Yet this is at the heart of exploratory data analysis. ... the picture examining eye is the best finder we have of the wholly unanticipated.

- Tukey, 1980

Why visualize your data?

- Help cognition
- Expand memory
- Generate hypotheses
- Answer questions
- Make decisions
- Find patterns
- Record
- Clarify
- Communicate
- Inspire

Why visualize your data?

- RECORD information
- ANALYZE data to support reasoning
- CONFIRM hypotheses
- COMMUNICATE ideas to others

Be careful: "change blindness"

Change blindness: a perceptual phenomenon in which individuals fail to detect significant changes in their visual environment, even when those changes are quite noticeable.

https://www.youtube.com/watch?v=FWSxSQsspiQ

https://www.youtube.com/watch?v=1nL5ulsWMYc

Good example

https://www.youtube.com/watch?v=hVimVzgtD6w

Good example

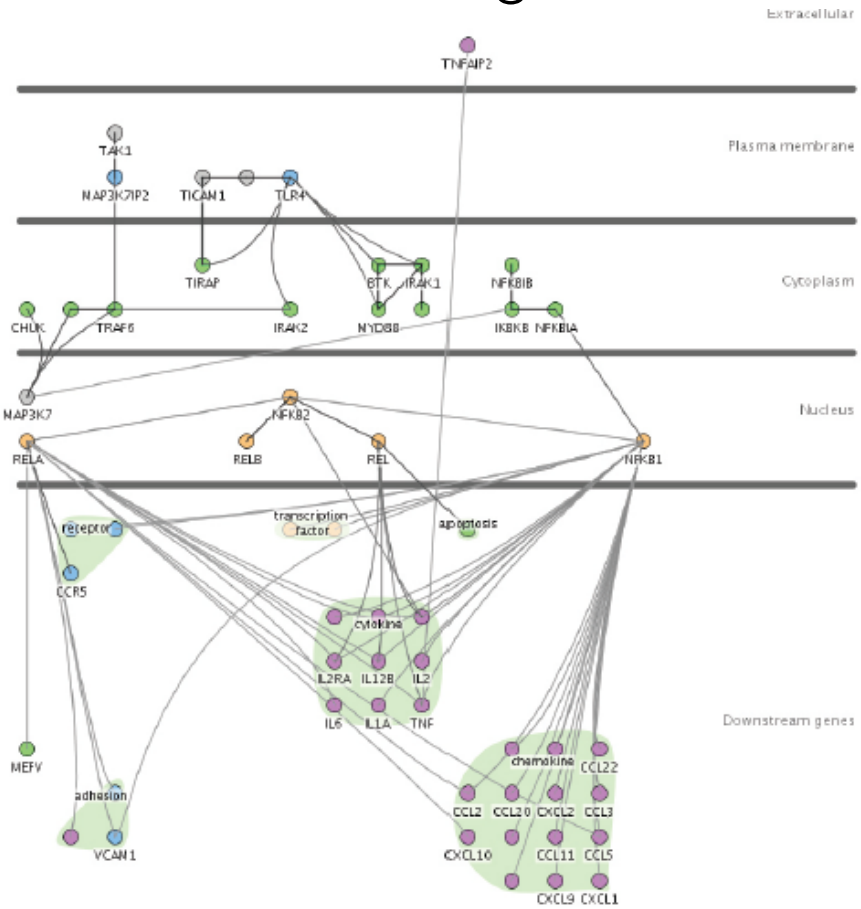
https://www.youtube.com/watch?v=hVimVzgtD6w

Generate hypotheses, Answer questions, Find patterns, Record, Clarify, Communicate, Inspire

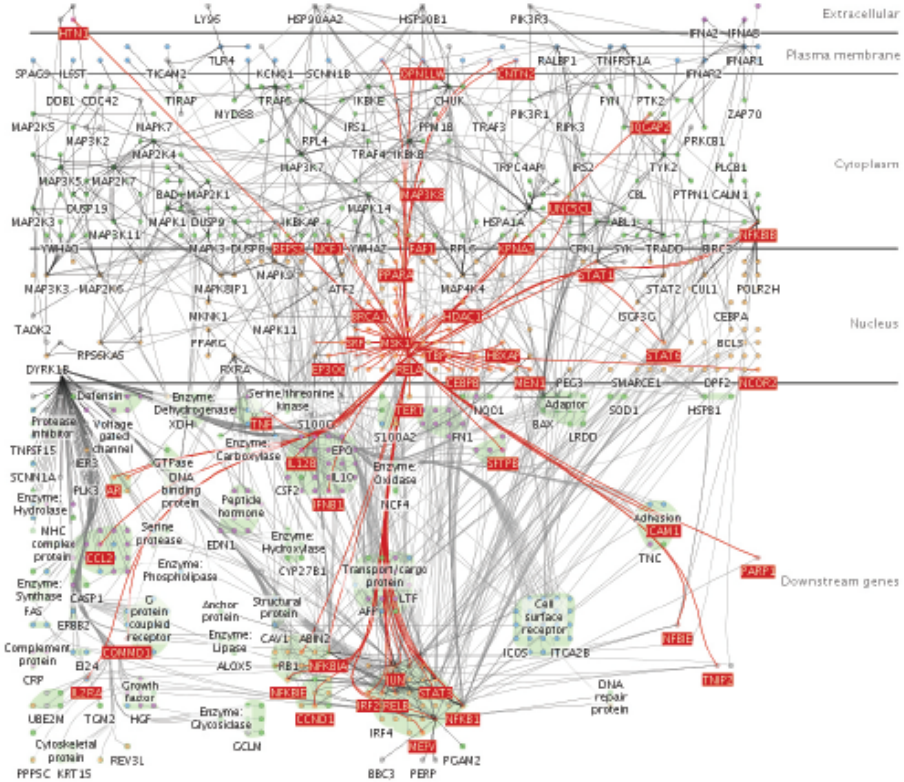
https://www.gapminder.org/tools/

Bad example

The Cerebral vis tool captures the style of hand-drawn diagrams in biology textbooks with vertical layers that correspond to places within a cell where interactions between genes occur.



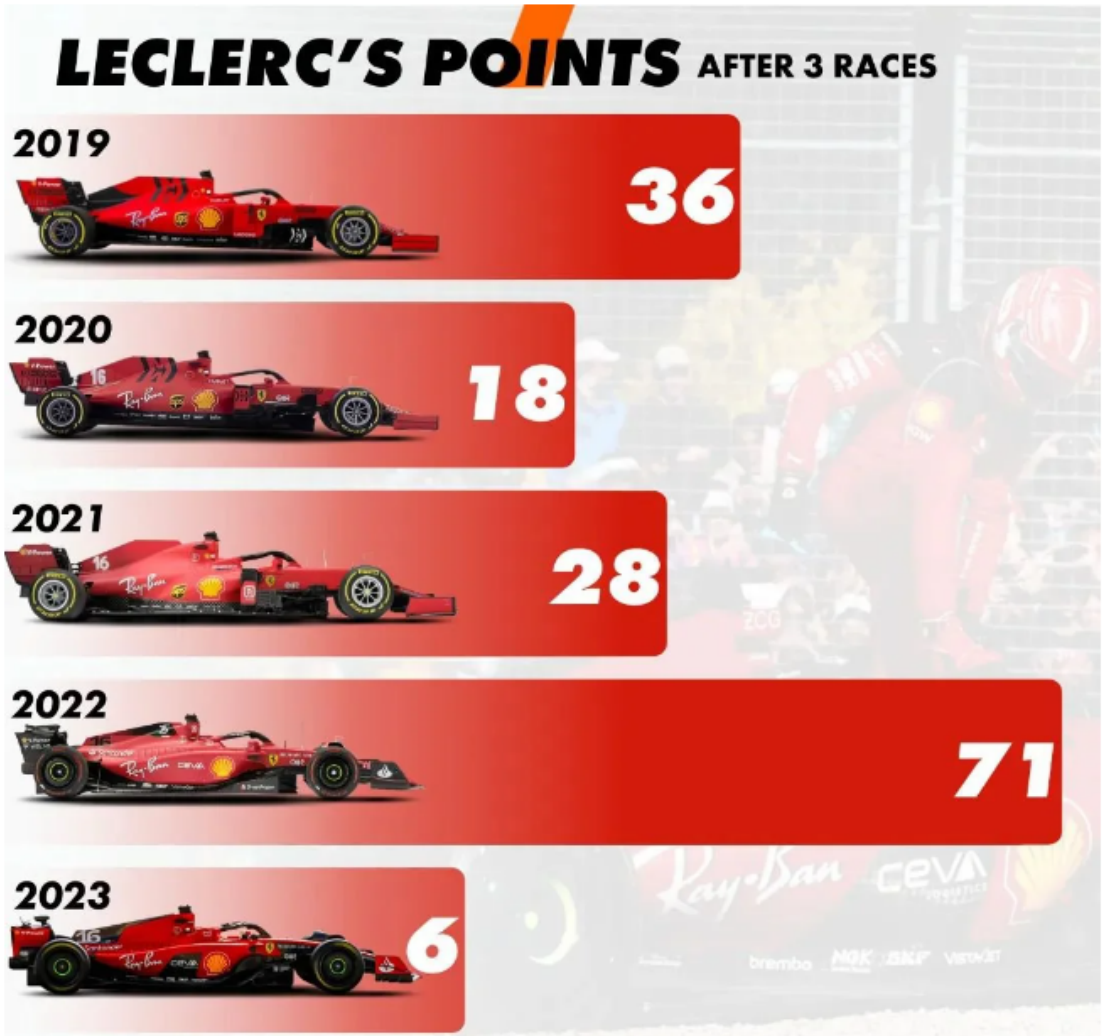
(a)



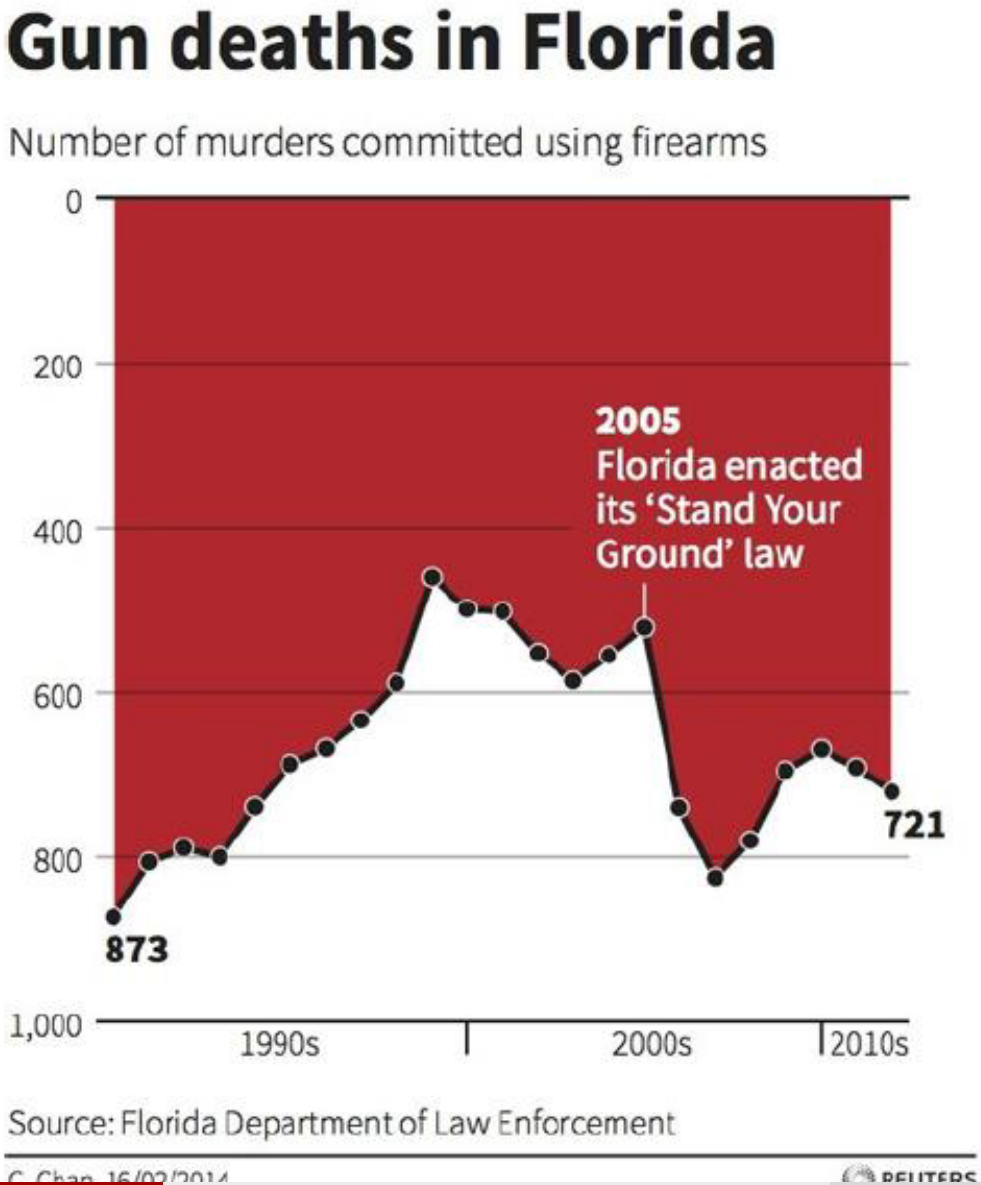
(b)

Bad example

The following image shows the points obtained by Charles Leclerc after three races in different years.

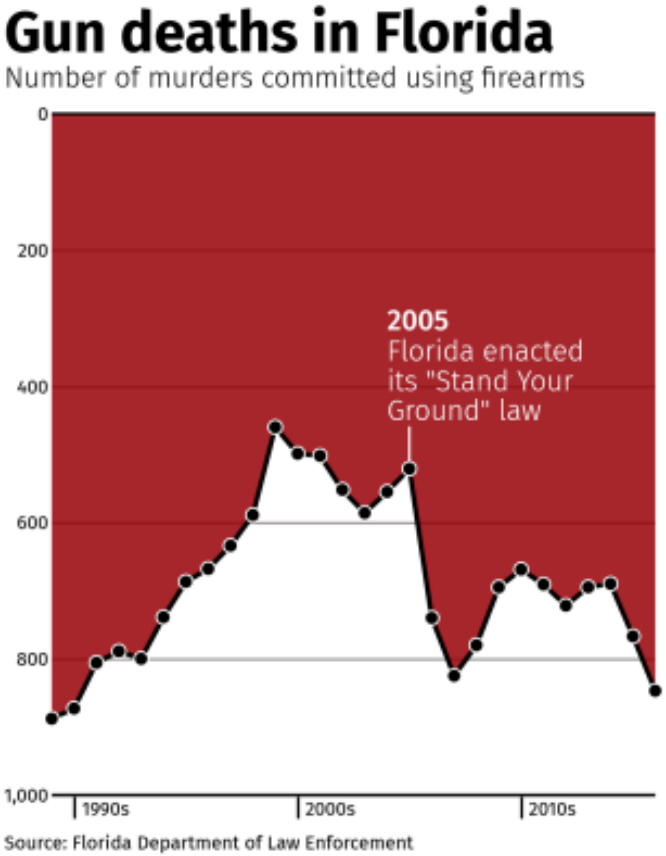


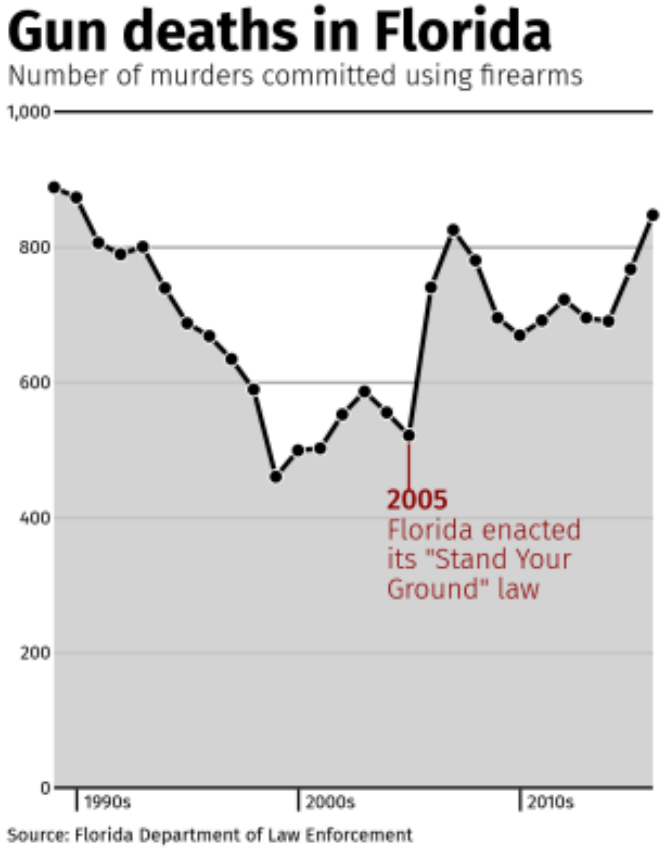
Bad example


Gun deaths in Florida
Number of murders committed using firearms
0
200
400
600
800
1,000
2005
Florida enacted its 'Stand Your Ground' law
873
721
1990s
2000s
2010s
Source: Florida Department of Law Enforcement

Design Rules of Thumb

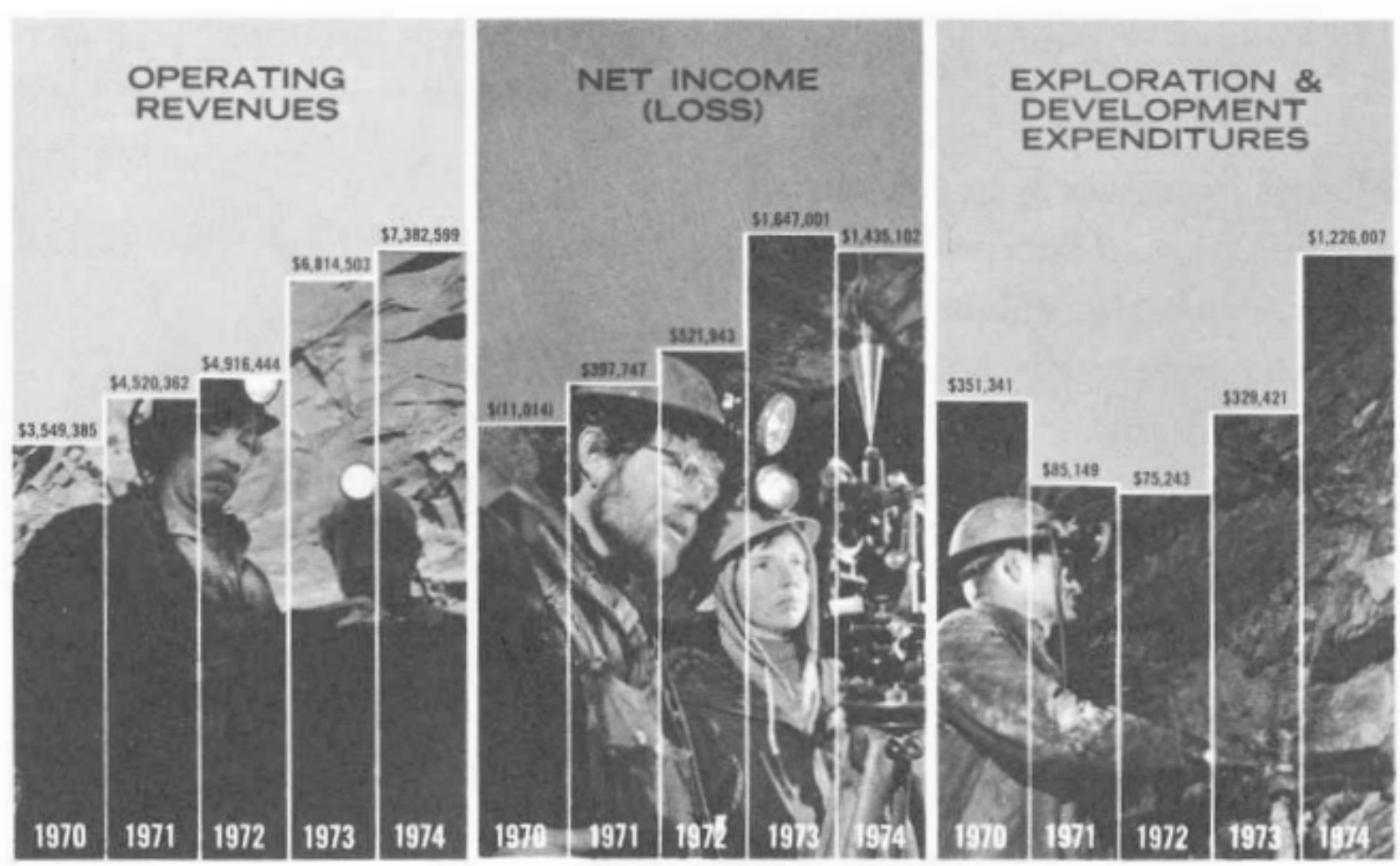
1. Function first, form next





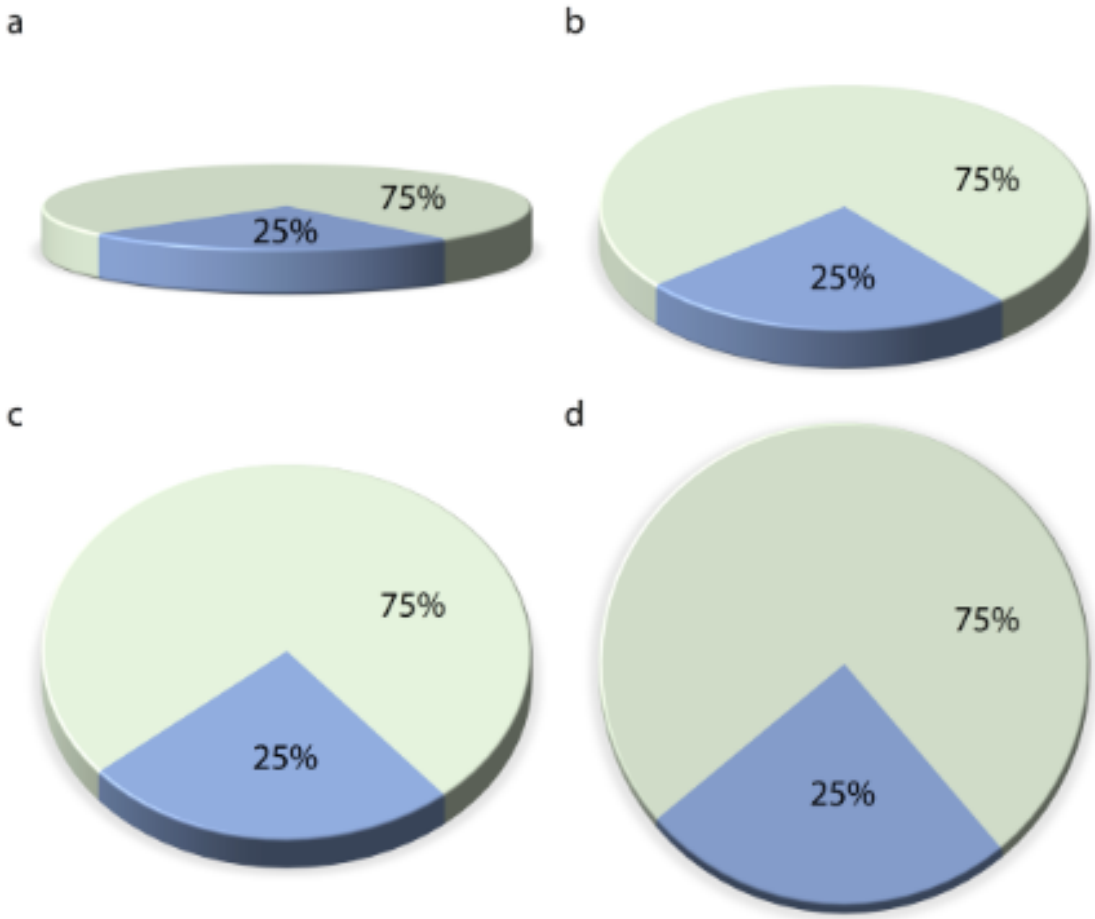
Design Rules of Thumb

1. Function first, form next



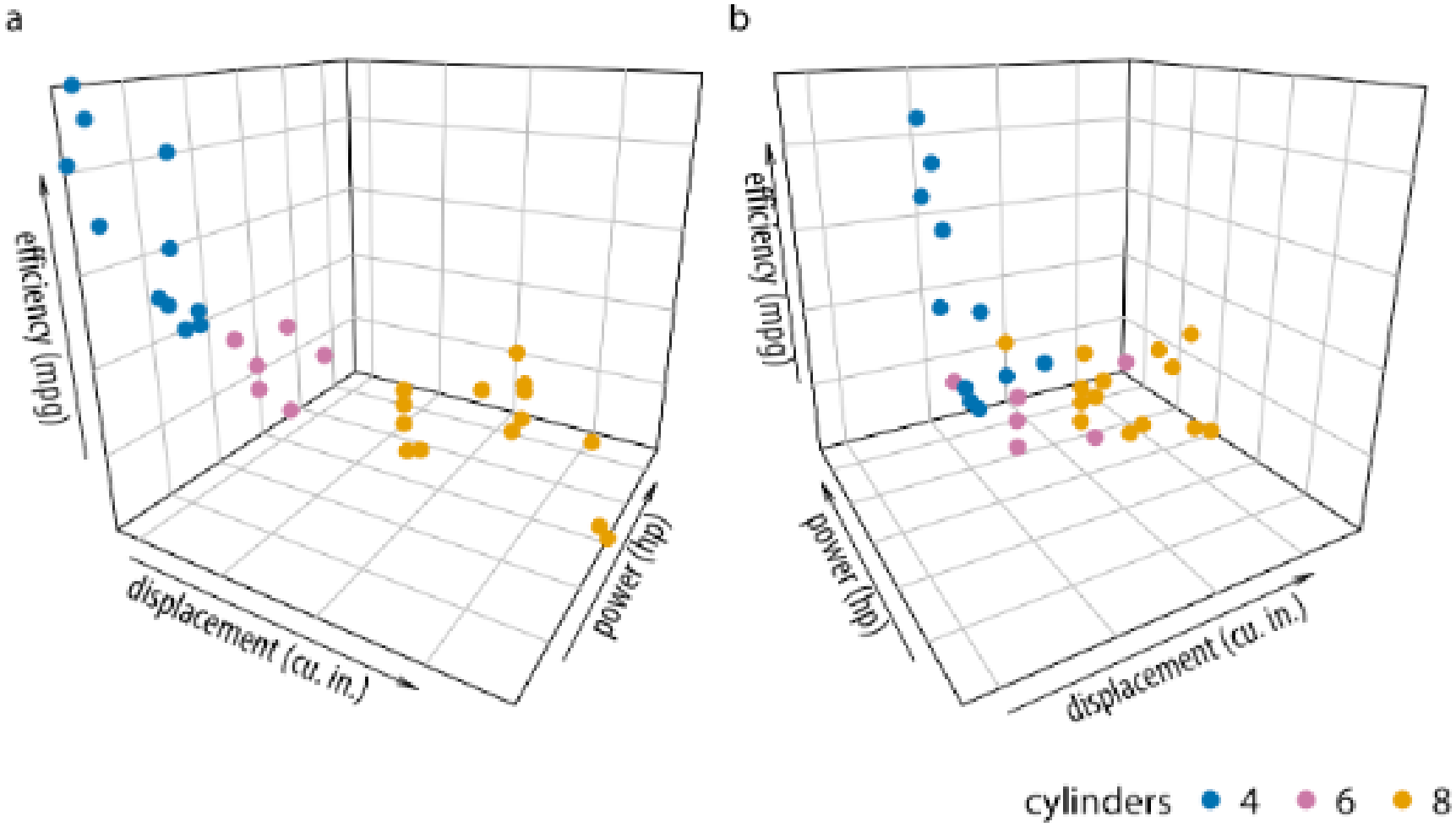
Design Rules of Thumb

2. No unjustified 3D



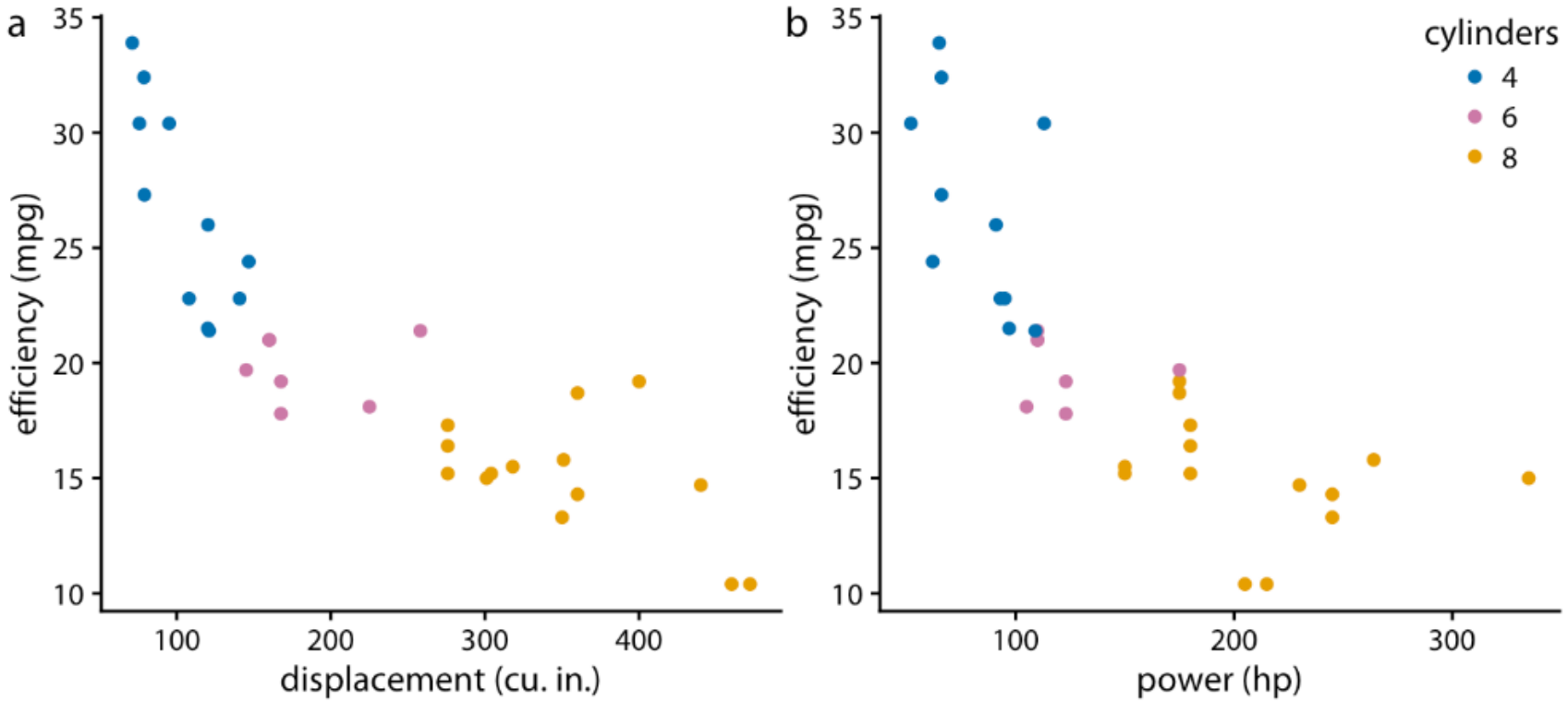
Design Rules of Thumb

2. No unjustified 3D



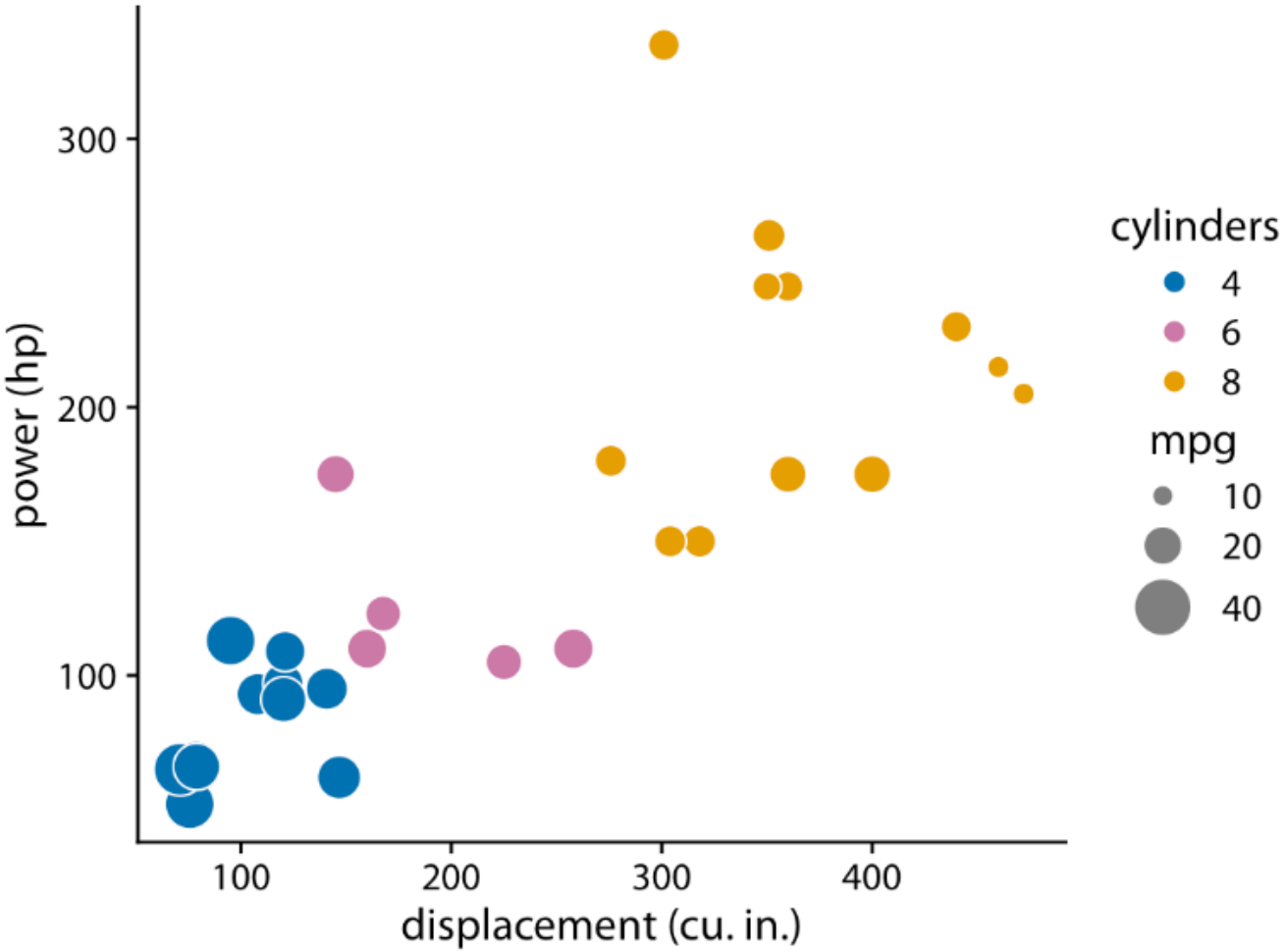
Design Rules of Thumb

2. No unjustified 3D



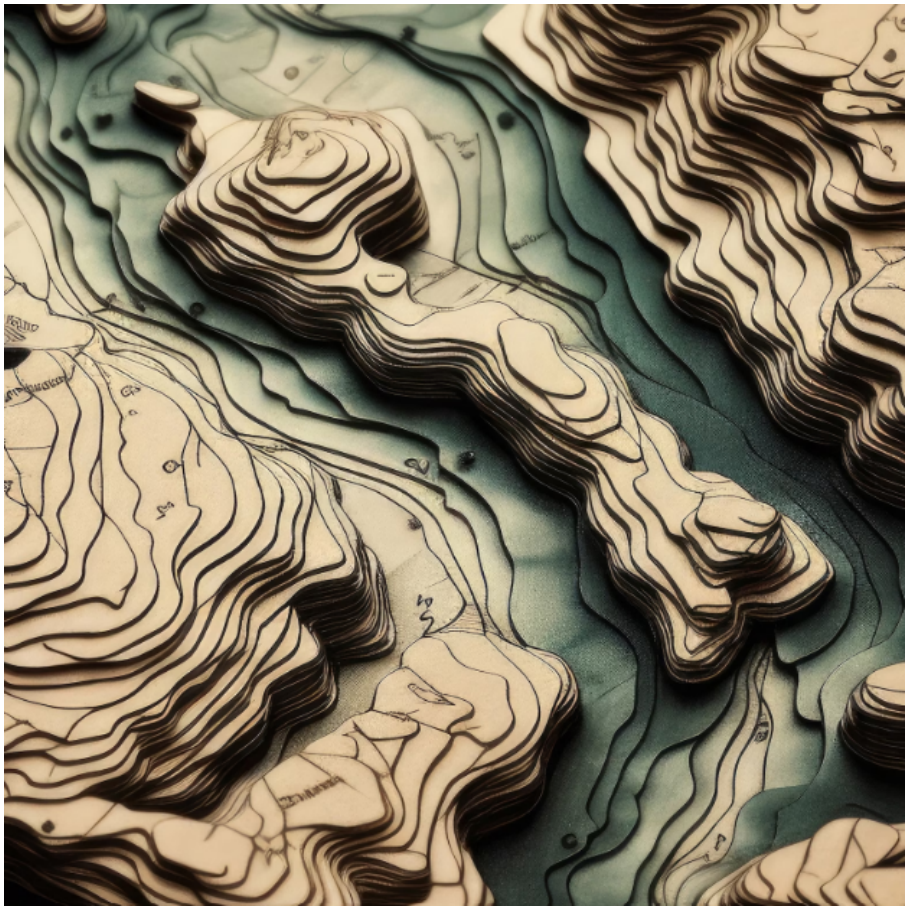
Design Rules of Thumb

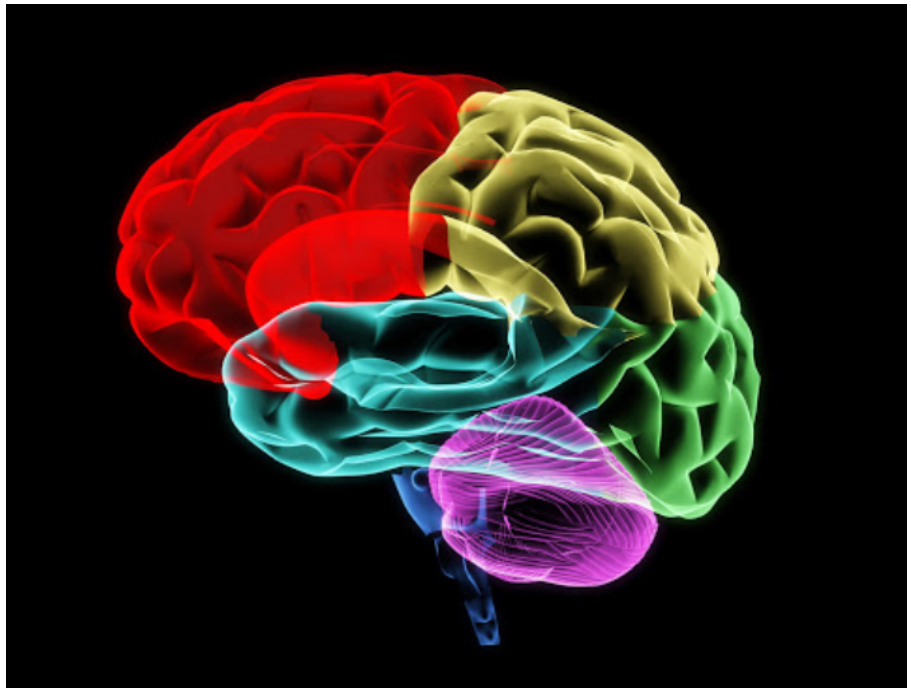
2. No unjustified 3D



Design Rules of Thumb

2*. No unjustified 3D





Design Rules of Thumb

3. No unjustified 2D

Task: what is the color for "Delta"?

Design Rules of Thumb

3. No unjustified 2D

Task: what is the color for "Delta"?

Node	Color
Alpha	White
Beta	Maroon
Delta	Purple
Epsilon	Teal
Eta	Mustard Yellow
Gamma	Green
Theta	Orange
Zeta	Pink

Design Rules of Thumb

3. No unjustified 2D

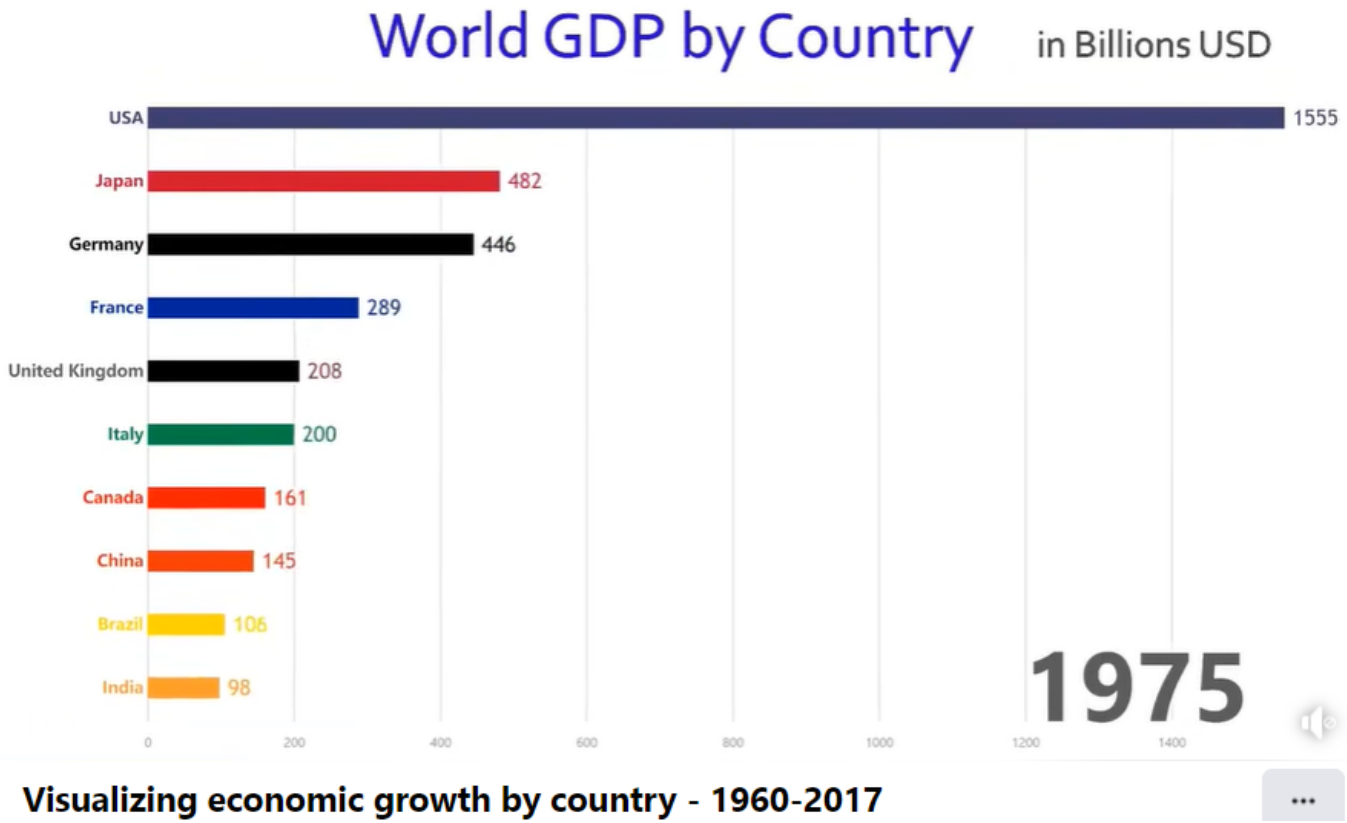
Task: what is the color for "Delta"?

Node	Color
Alpha	White
Beta	Maroon
Delta	Purple
Epsilon	Teal
Eta	Mustard Yellow
Gamma	Green
Theta	Orange
Zeta	Pink

If the task doesn't need a 2D visualization, then don't use one.

Design Rules of Thumb

4. Eyes beat memory



https://www.facebook.com/visualcapitalist/videos/visualizing-economic-growth-by-country-1960-2017/2169324383292546/

Design Rules of Thumb

- Function first, form next
- No unjustified 3D
- No unjustified 2D
- Eyes beat memory

Edward Tufte

Tufte (age 73) during his one-day course in Dallas, May 21, 2015

Edward Rolf Tufte is an American statistician and professor emeritus of political science, statistics, and computer science at Yale University. He is noted for his writings on information design and as a pioneer in the field of data visualization.

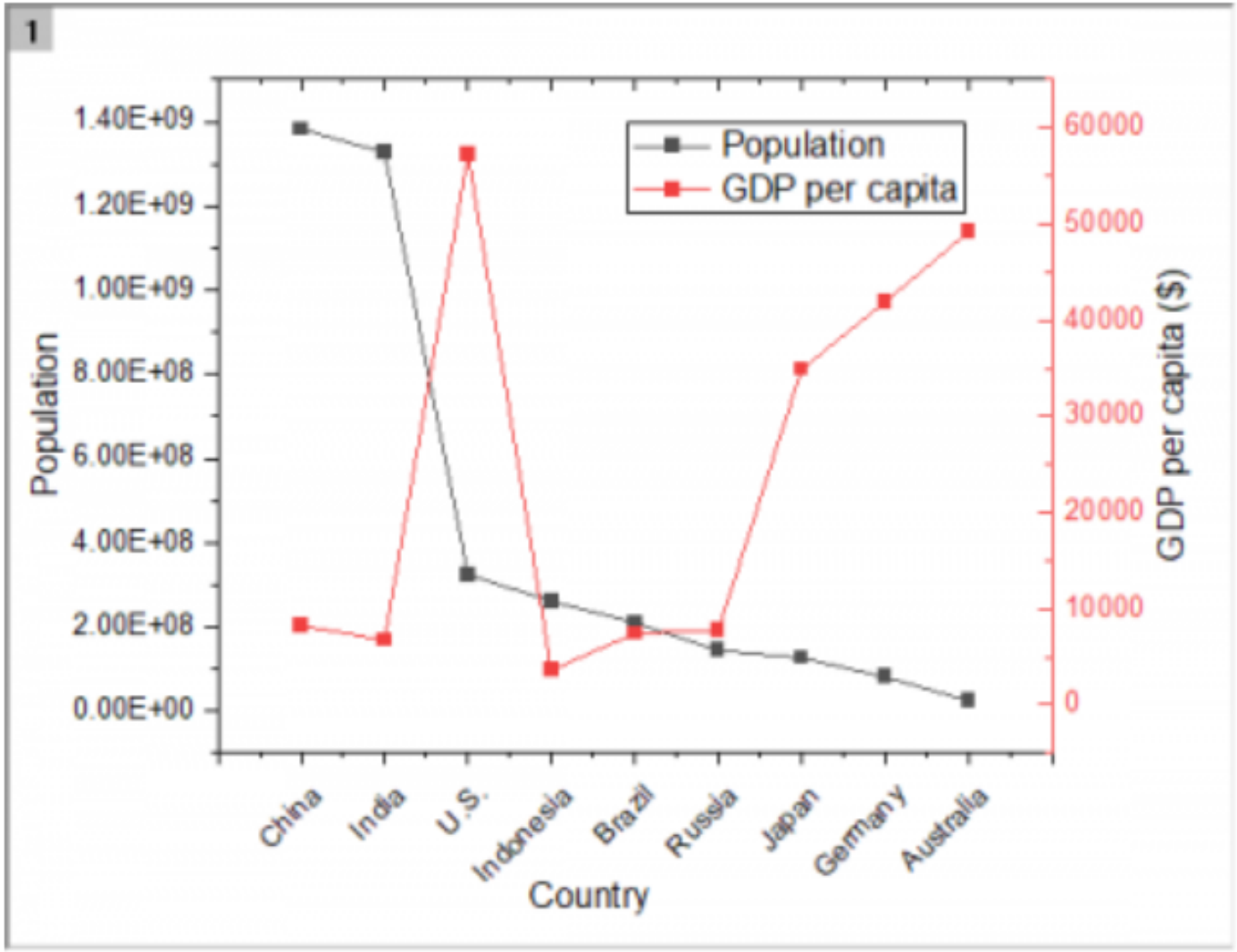
Graphical Integrity 1

"Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data. "

Tufte, "Visual Display of Quantitative Information"

(Axes and axis labels, titles, annotations, legends, etc.)

Double axis


1
Population
GDP per capita
1.40E+09
1.20E+09
1.00E+09
8.00E+08
6.00E+08
4.00E+08
2.00E+08
0.00E+00
60000
50000
40000
30000
20000
10000
0
GDP per capita ($)
China
India
U.S.
Indonesia
Brazil
Russia
Japan
Germany
Australia
Country

Two titles

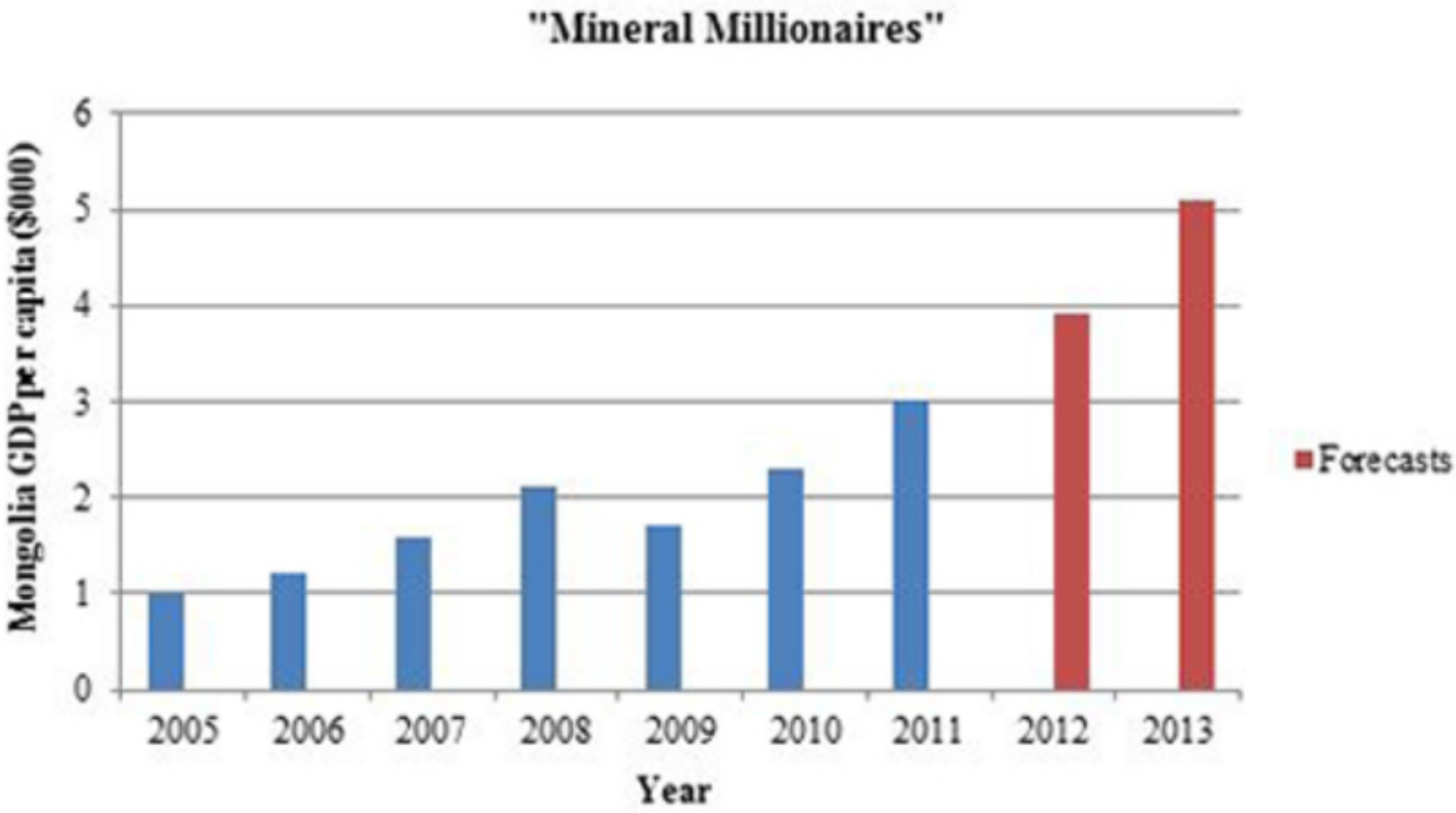
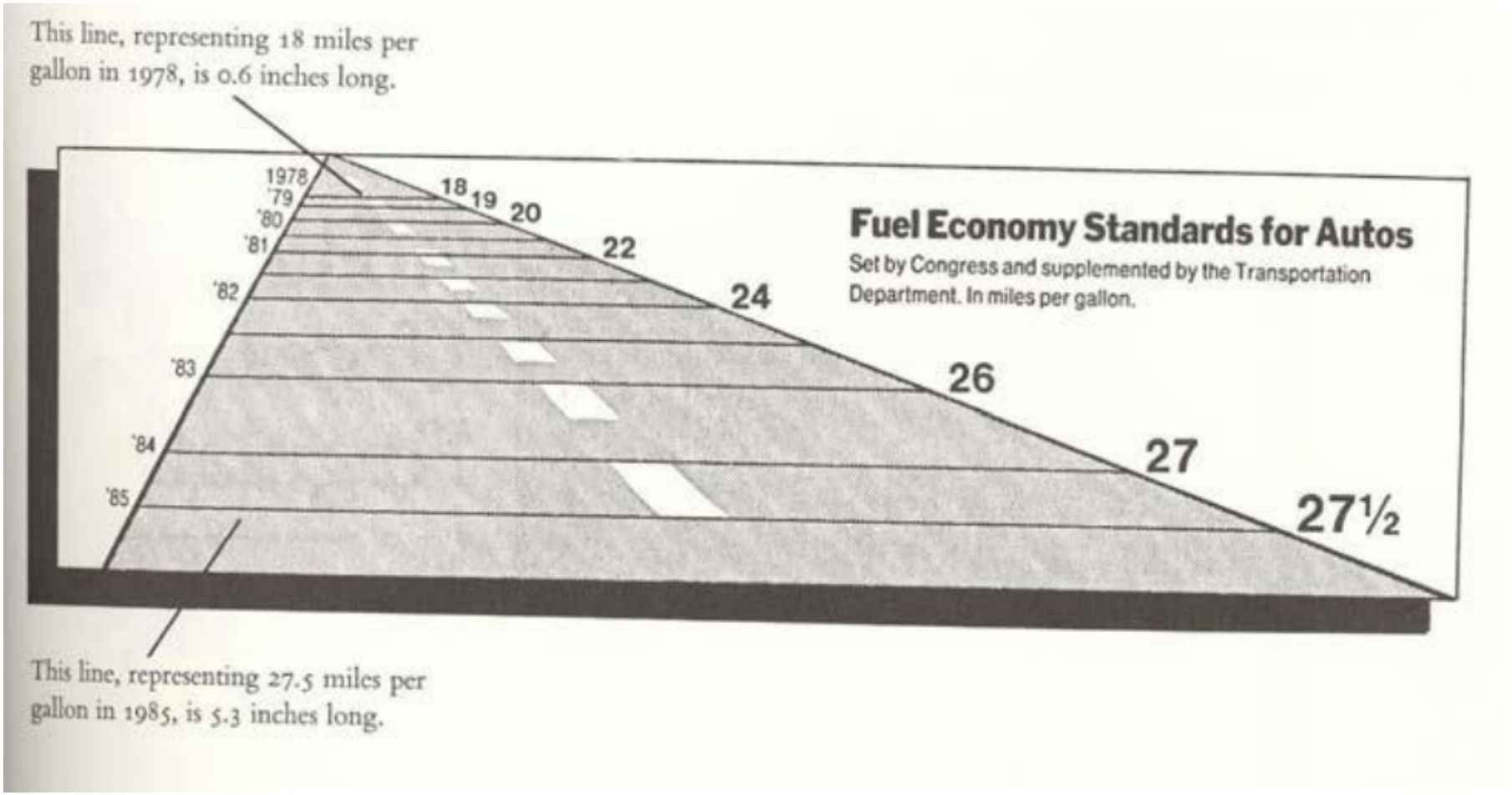


Figure 1.7. Expectations of large GDP per capita growth from a Financial Times article.

Graphical Integrity 2

"The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured."

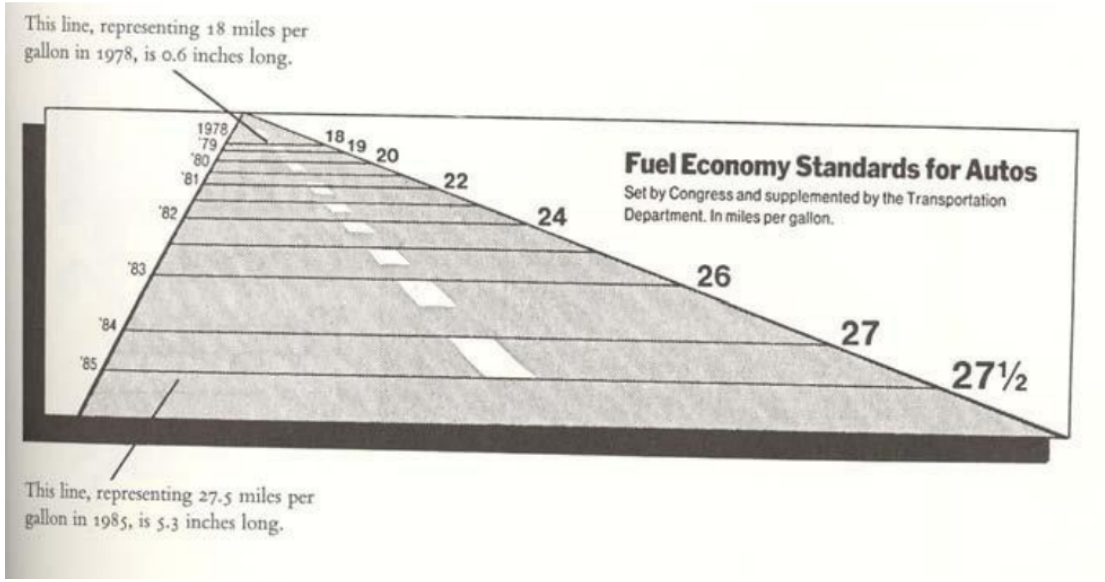
Lie factor


This line, representing 18 miles per gallon in 1978, is 0.6 inches long.
1978
'79
'80
'81
'82
'83
'84
'85
18
19
20
22
24
26
27
27½
Fuel Economy Standards for Autos
Set by Congress and supplemented by the Transportation Department. In miles per gallon.
This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Lie factor

$$\text{Lie factor} = \frac{\text{Size of effect in graphic}}{\text{Size of effect in data}}$$

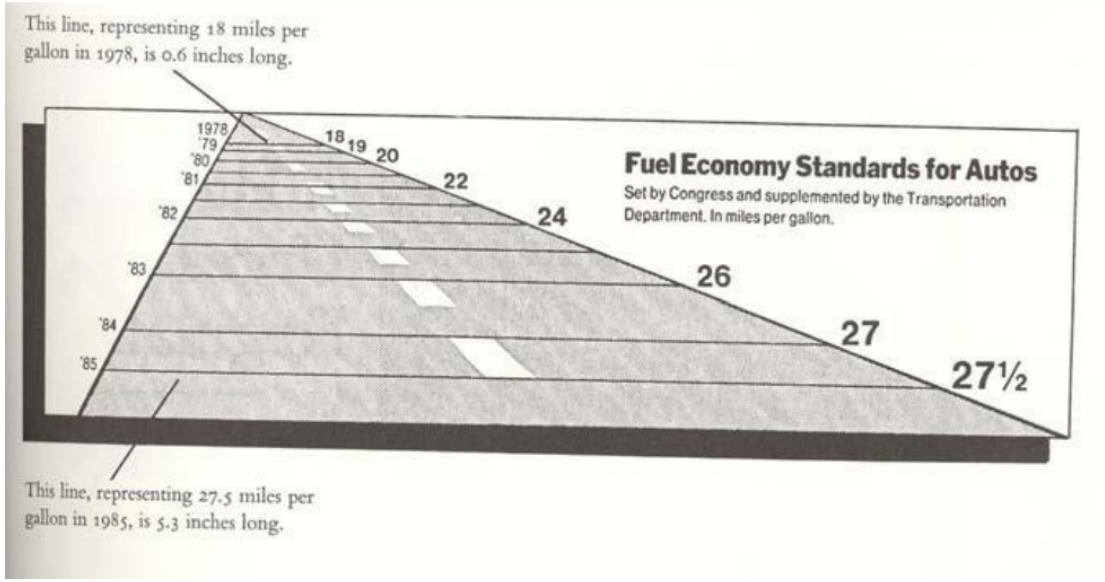
- Lie factor ≥ 1, overstating
- Lie factor ≈ 1, accurate
- Lie factor ≤ 1, understating



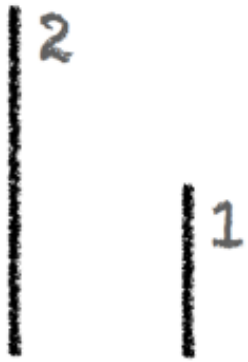
Lie factor

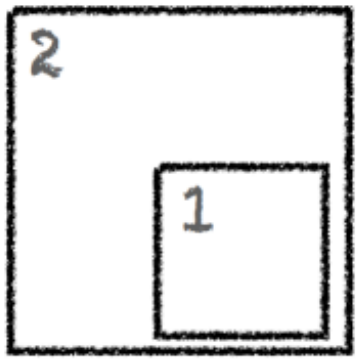
$$\text{Lie factor} = \frac{\text{Size of effect in graphic}}{\text{Size of effect in data}}$$

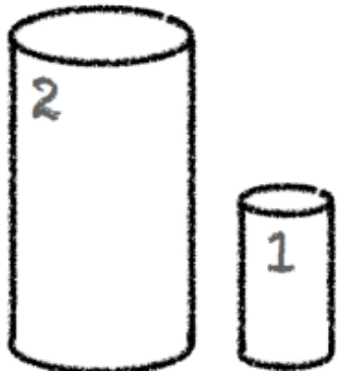
- image $= \frac{5.3-0.6}{0.6} = 7.83$
- data $= \frac{27-18}{18} = 0.53$
- lie factor $= \frac{7.83}{0.53} = 14.8$
- lie factor ≥ 1, overstating



Lie factor: calculate for yourself







Graphical Integrity 3

"The number of information-carrying (attribute) dimensions depicted should not exceed the number of dimensions in the data."

Chart Junk

- Excessive Gridlines
- Gratuitous 3D Effects
- Overly Decorative Fonts and Colors
- Unnecessary Illustrations
- Excessive Labels and Annotations
- Unrelated Graphics

Chart Junk: How to change?

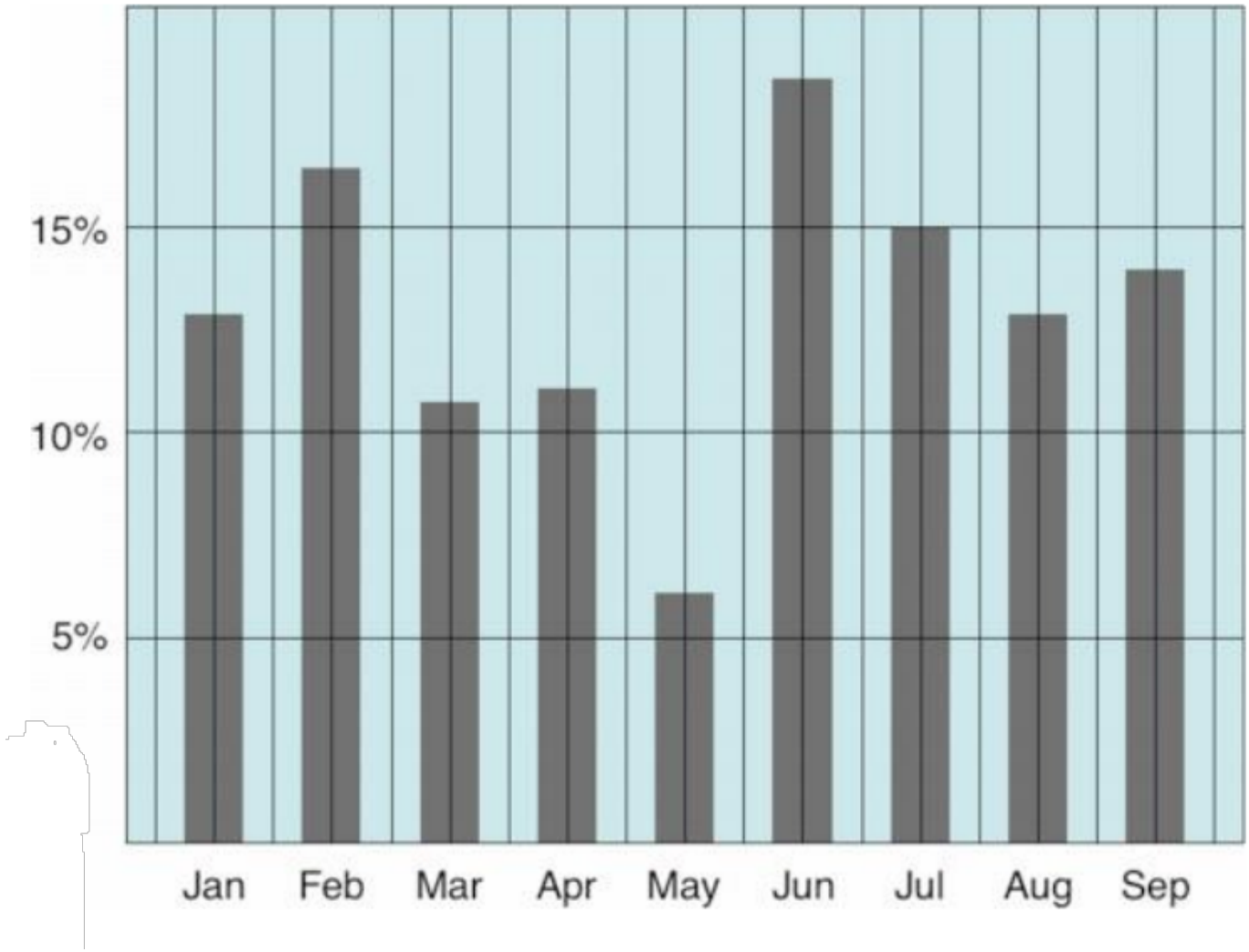
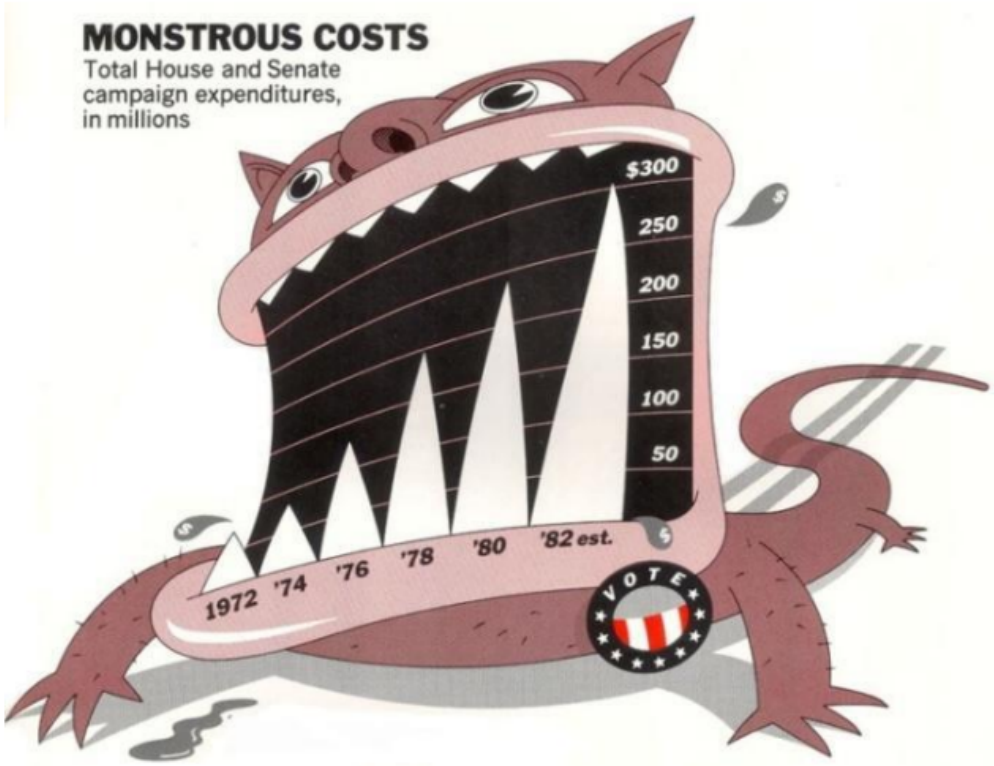


Chart Junk


MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
1972
'74
'76
'78
'80
'82 est.
VOTE

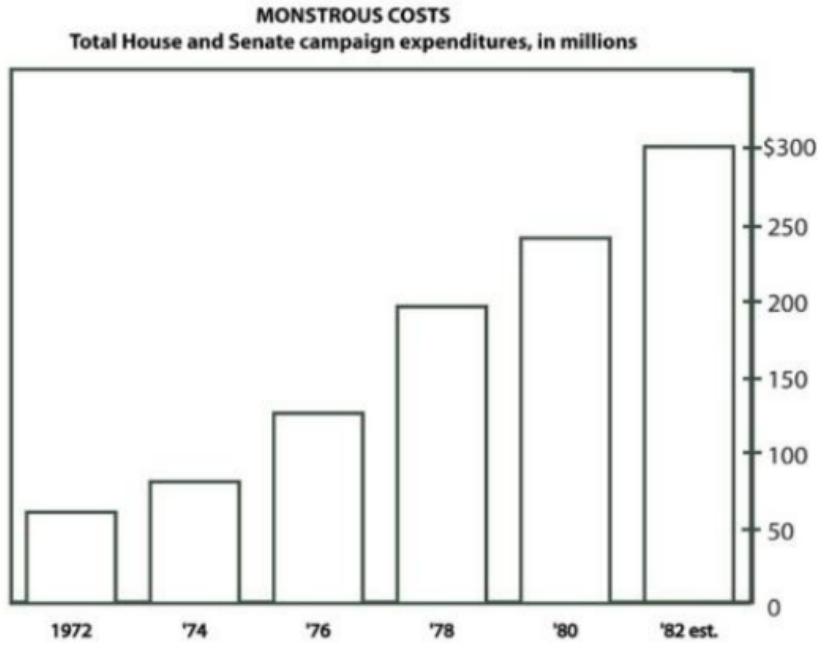
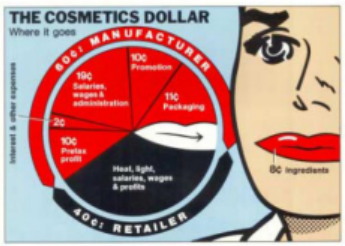
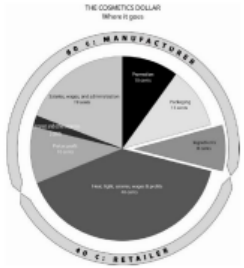

MONSTROUS COSTS
Total House and Senate campaign expenditures, in millions
$300
250
200
150
100
50
0
1972
'74
'76
'78
'80
'82 est.

Chart Junk Debate

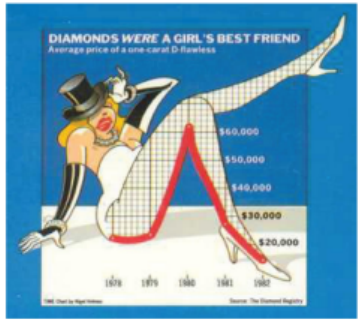
Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts





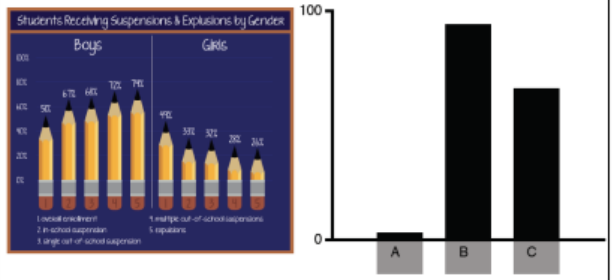
Bateman, et al. (2010)

Benefitting InfoVis with Visual Difficulties



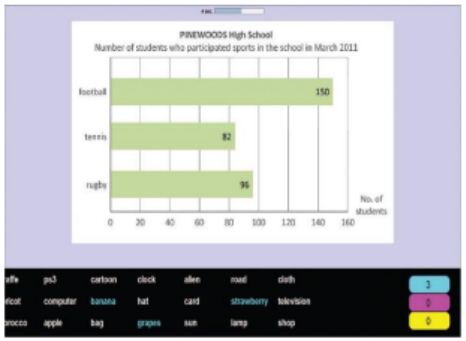
Hullman, et al. (2011)

An Evaluation of the Impact of Visual Embellishments in Bar Charts



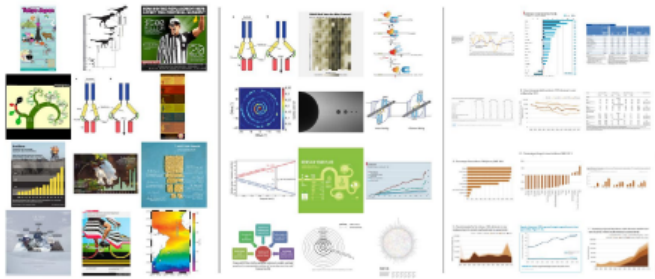
Skau, et al. (2015)

An Empirical Study on Using Visual Embellishments in Visualization



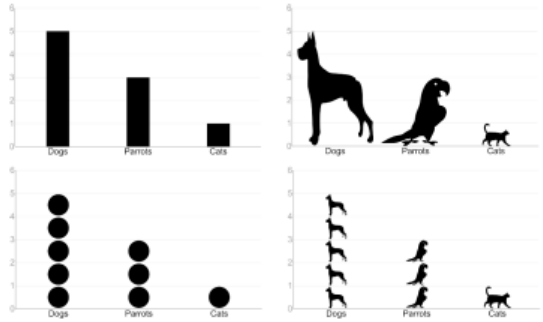
Borgo, et al. (2012)

What makes a visualization memorable?



Borkin, et al. (2013)
Borkin, et al. (2015)

ISOTYPE Visualization – Working Memory, Performance, and Engagement with Pictographs



Haroz, et al. (2015)

Chart Junk Debate

- Chart junk can persuade, help with memorability, engage
- Chart junk can bias, limit data-ink ratio, clutter, lower trust

Take-away: it depends on your audience, task, and context...