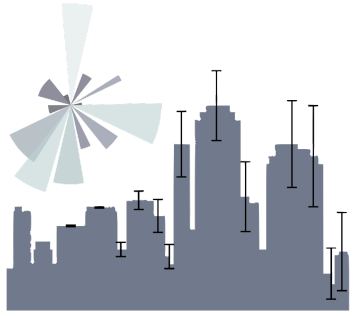
principles of Urban Science 9



visualizations

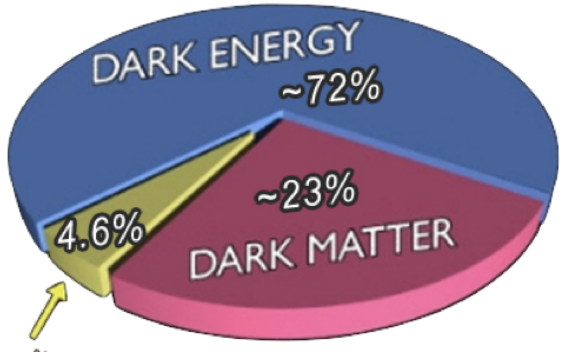
dr.federica bianco | *fbb.space* | *fedhere* | *fedhere*

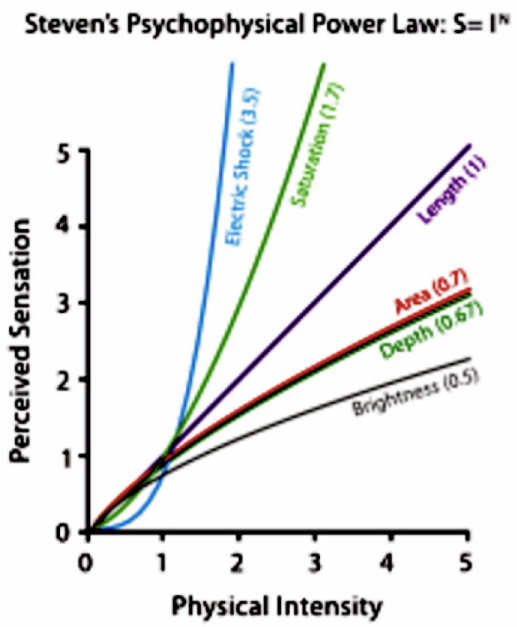
this slide deck: slides.com/federicabianco/pus2020_9

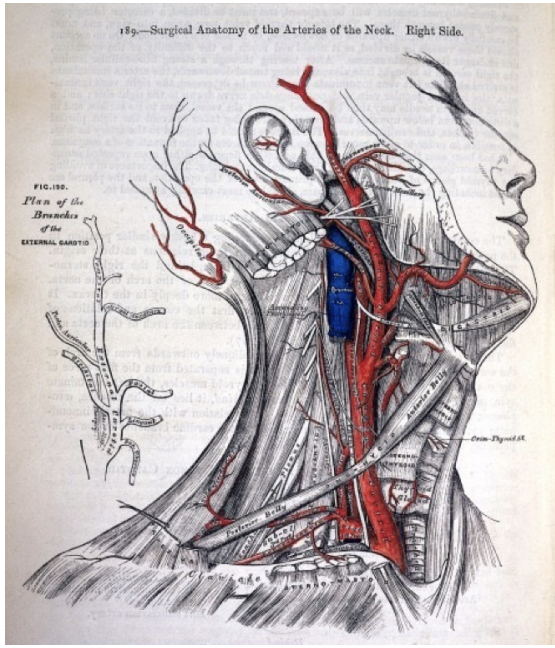
- Descriptive data viz
  - Lie with statistics
  - Tufte’s rules
- Exploratory data viz

  Jer Thorp

  - Psychophysics
  - Esthetics vs(??) functionality
    - color blindness
    - the third dimension
  - Interactivity








why?

In [3]:
```
import pylab as pl
from pylab import imread
%pylab inline
im = imread("secret.png")
im
```

executed in 32ms, finished 16:07:56 2019-10-19

```
Populating the interactive namespace from numpy and matplotlib
```

Out[3]:
```
array([[[0.3882353 , 0.49019608, 0.54901963, 1.        ],
        [0.48235294, 0.5803922 , 0.6509804 , 1.        ],
        [0.44705883, 0.54509807, 0.6156863 , 1.        ],
        ...,
        [0.80784315, 0.7254902 , 0.5568628 , 1.        ],
        [0.78039217, 0.68235296, 0.5254902 , 1.        ],
        [0.8745098 , 0.7607843 , 0.627451  , 1.        ]],

       [[0.4392157 , 0.5411765 , 0.6       , 1.        ],
        [0.3647059 , 0.46666667, 0.5254902 , 1.        ],
        [0.44313726, 0.54509807, 0.6       , 1.        ],
        ...,
        [0.78039217, 0.69803923, 0.5294118 , 1.        ],
        [0.8352941 , 0.7372549 , 0.5803922 , 1.        ],
        [0.83137256, 0.7176471 , 0.58431375, 1.        ]],

       [[0.28627452, 0.39215687, 0.4588235 , 1.        ],
        [0.3882353 , 0.49803922, 0.54901963, 1.        ],
        [0.49019608, 0.5117647 , 0.4745098 , 1.        ],
        ...,
        [0.8039216 , 0.7294118 , 0.5078434 , 1.        ],
        [0.7764706 , 0.68235296, 0.53333336, 1.        ],
        [0.8       , 0.69803923, 0.5686275 , 1.        ]],

       ...,
```

why?

computers understand data as numbers,

we (people) do not.

In [3]:
```
import pylab as pl
from pylab import imread
%pylab inline
im = imread("secret.png")
im
```

executed in 32ms, finished 16:07:56 2019-10-19

```
Populating the interactive namespace from numpy and matplotlib
```

Out[3]:
```
array([[[0.3882353 , 0.49019608, 0.54901963, 1.        ],
        [0.48235294, 0.5803922 , 0.6509804 , 1.        ],
        [0.44705883, 0.54509807, 0.6156863 , 1.        ],
        ...,
        [0.80784315, 0.7254902 , 0.5568628 , 1.        ],
        [0.78039217, 0.68235296, 0.5254902 , 1.        ],
        [0.8745098 , 0.7607843 , 0.627451  , 1.        ]],

       [[0.4392157 , 0.5411765 , 0.6       , 1.        ],
        [0.3647059 , 0.46666667, 0.5254902 , 1.        ],
        [0.44313726, 0.54509807, 0.6       , 1.        ],
        ...,
        [0.78039217, 0.69803923, 0.5294118 , 1.        ],
        [0.8352941 , 0.7372549 , 0.5803922 , 1.        ],
        [0.83137256, 0.7176471 , 0.58431375, 1.        ]],

       [[0.28627452, 0.39215687, 0.45882353, 1.        ],
        [0.3882353 , 0.49803922, 0.54901963, 1.        ],
        [0.29803923, 0.4117647 , 0.42745098, 1.        ],
        ...,
        [0.803[illegible]94118 , 0.5[illegible]434, 1.        ],
        [0.7764706 , 0.68235296, 0.[illegible]3333336, 1.        ],
        [0.8[illegible], 0.69803923, 0.568[illegible]375 , 1.        ]],

       ...,
```

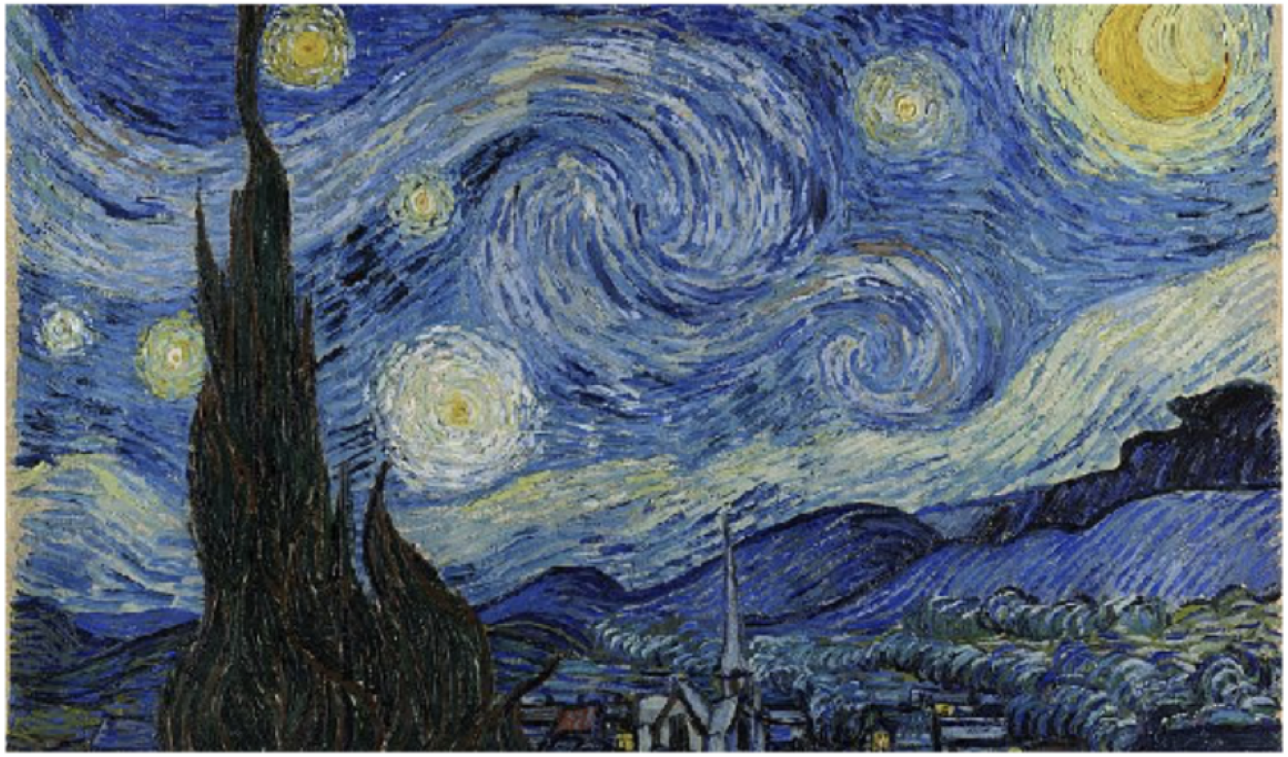
computers understand data as numbers,

we (people) do not.

what is this?

In [2]:
```
pl.figure(figsize(10,10))
pl.imshow(im)
pl.axis('off');
```

executed in 177ms, finished 16:14:21 2019-10-19



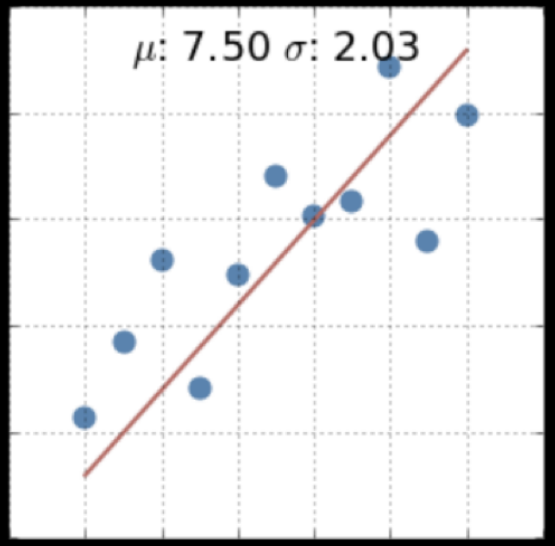
Van Gogh starry night

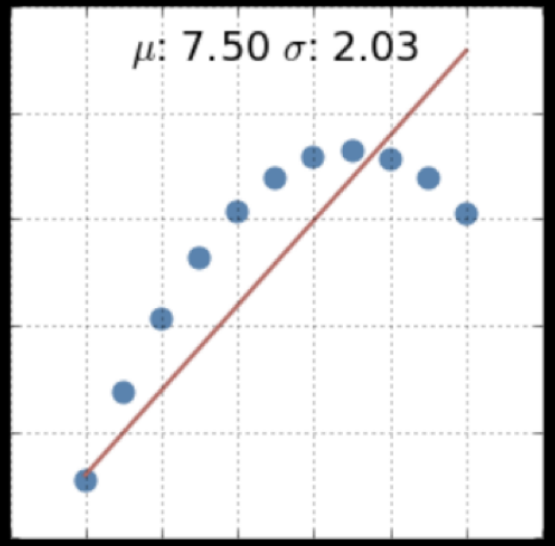
computers understand data as numbers,

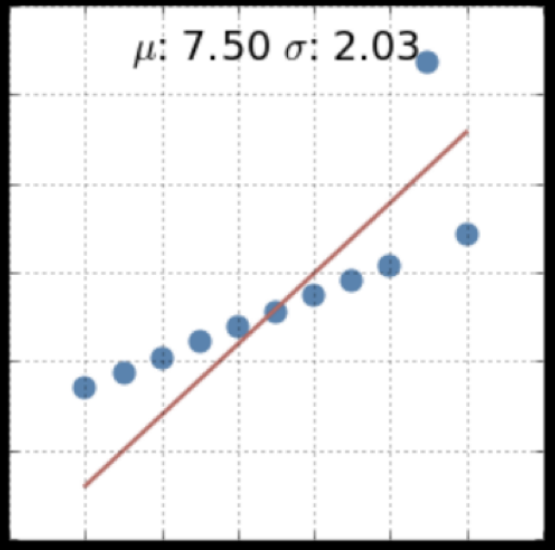
we (people) do not.

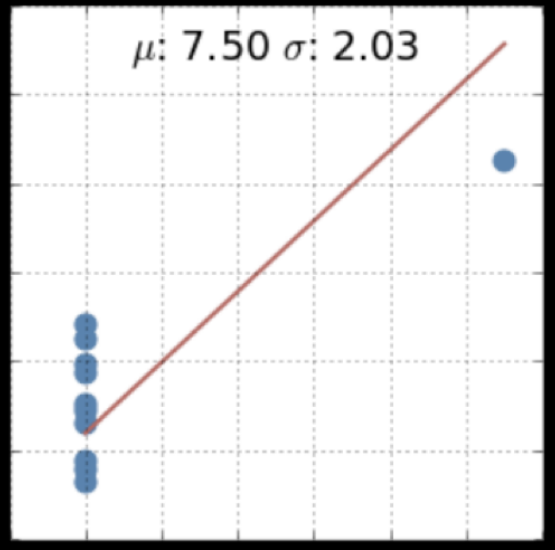
I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

what is this?









(Francis Anscombe, 1973) comprises four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

the moral of the story is: look at your data!

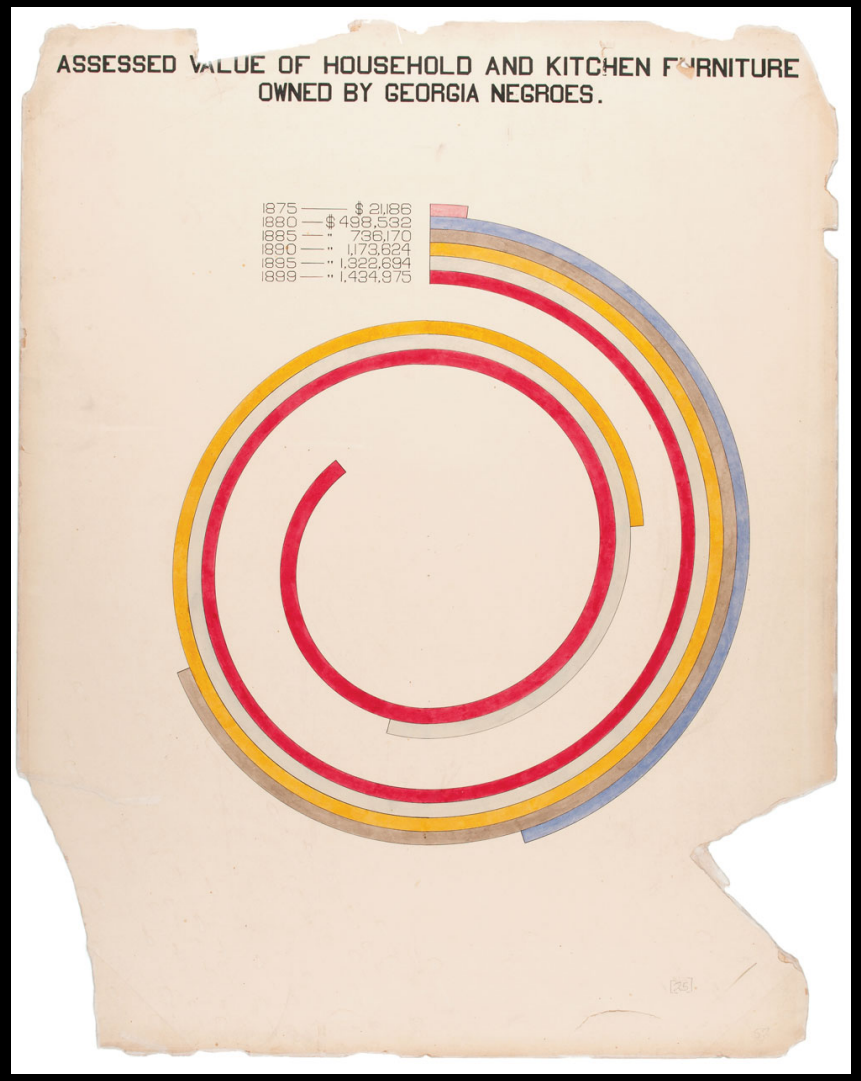
Anscombe's quartet?

https://github.com/fedhere/DSPS/blob/master/lab8/Anscombe's%20Quartet.ipynb

we visualize to
communicate
(Tufte)
and to
explore
(Thorp)

“Du Bois was aware that while unmoving prose and dry presentations of charts and graphs might catch attention from specialists, this approach would not garner notice beyond narrow circles of academics,” Aldon Morris writes in the essay “American Negro at Paris, 1900.” “Such social science was useless to the liberation of oppressed peoples. Breaking from tradition, Du Bois was among the first great American public intellectuals whose reach extended beyond the academy to the masses.”

https://hyperallergic.com/476334/how-w-e-b-du-bois-meticulously-visualized-20th-century-black-america/



while eyesight is the most developed sense for humanity in general consider perceptual differences to assure accessibility and equality! sonification, tactile data 3D printed, and accessible colors and visual properties

https://www.revealnews.org/article/watch-oklahomas-earthquake-explosion/

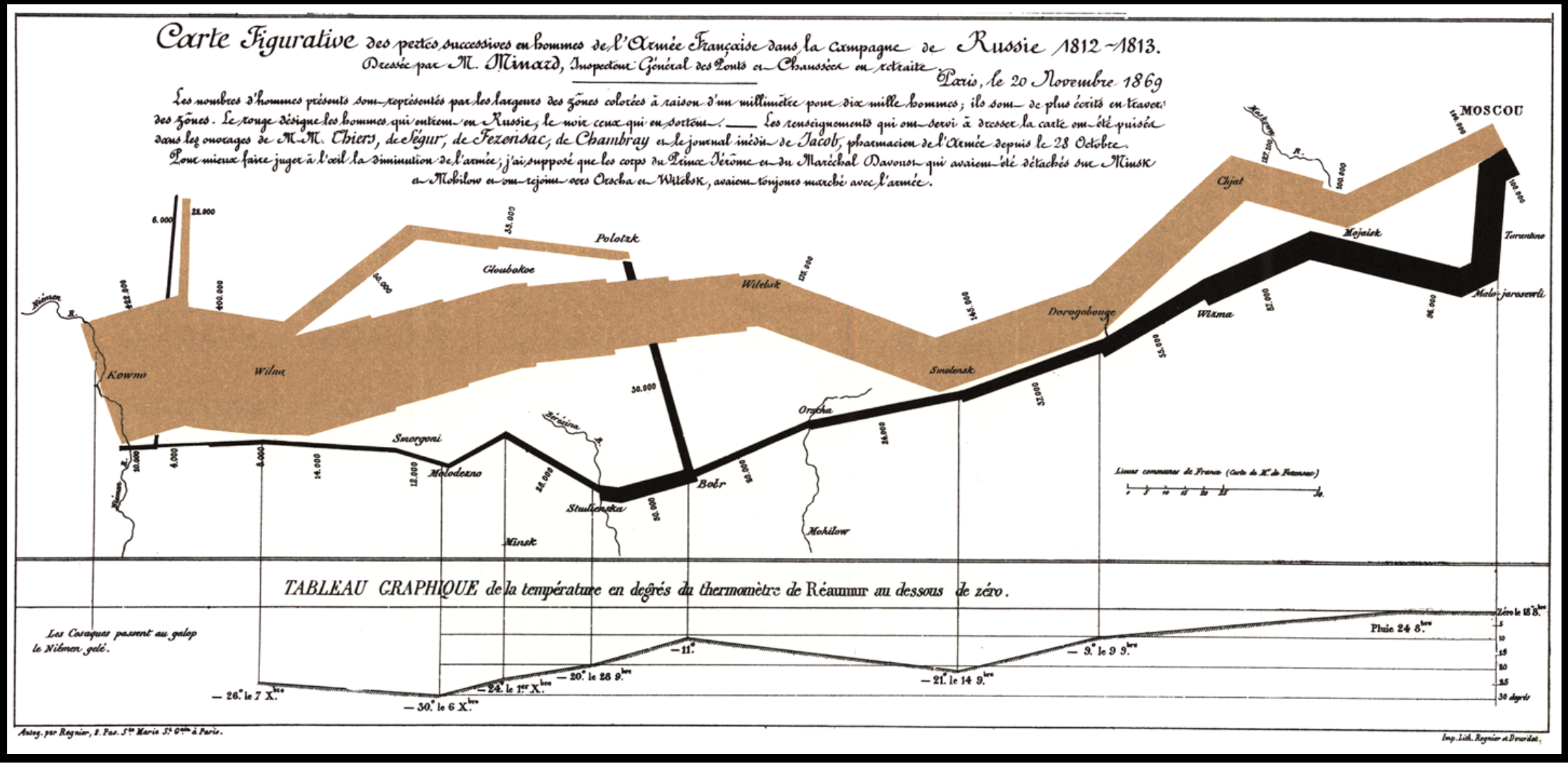

how?

a few historical plots and why they made history

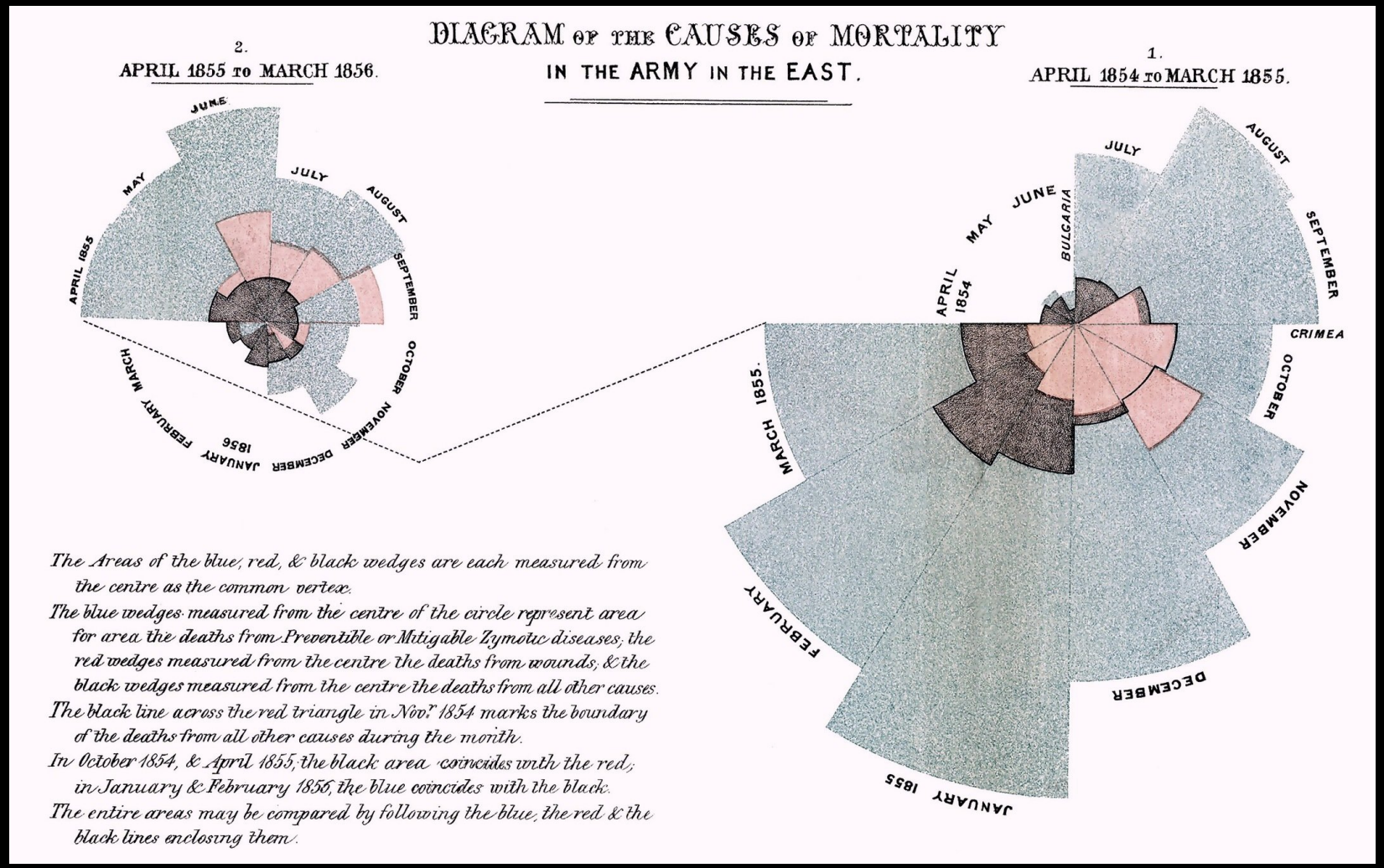
a few historical plots and why they made history

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812-1813.

The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October.In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jèrôme and under Marshal Davoust, who were sent to Minsk and Mobilow and who rejoined near Orscha and Witebsk, had always marched with the army.



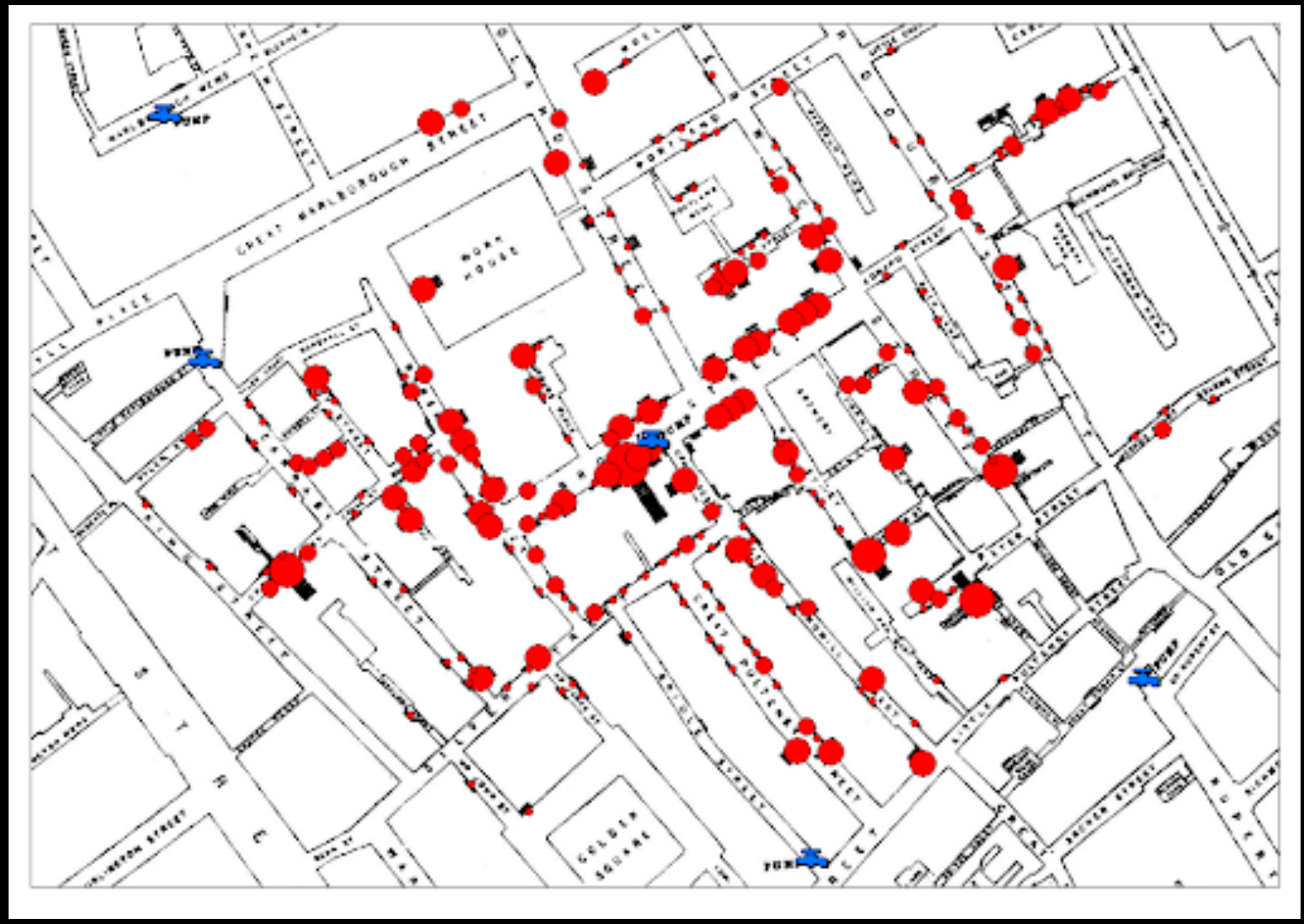
a few historical plots and why they made history



Florence Nightingale Coxcombs

http://timelyportfolio.github.io/rCharts_micropolar/nightingale/index.html

a few historical plots and why they made history

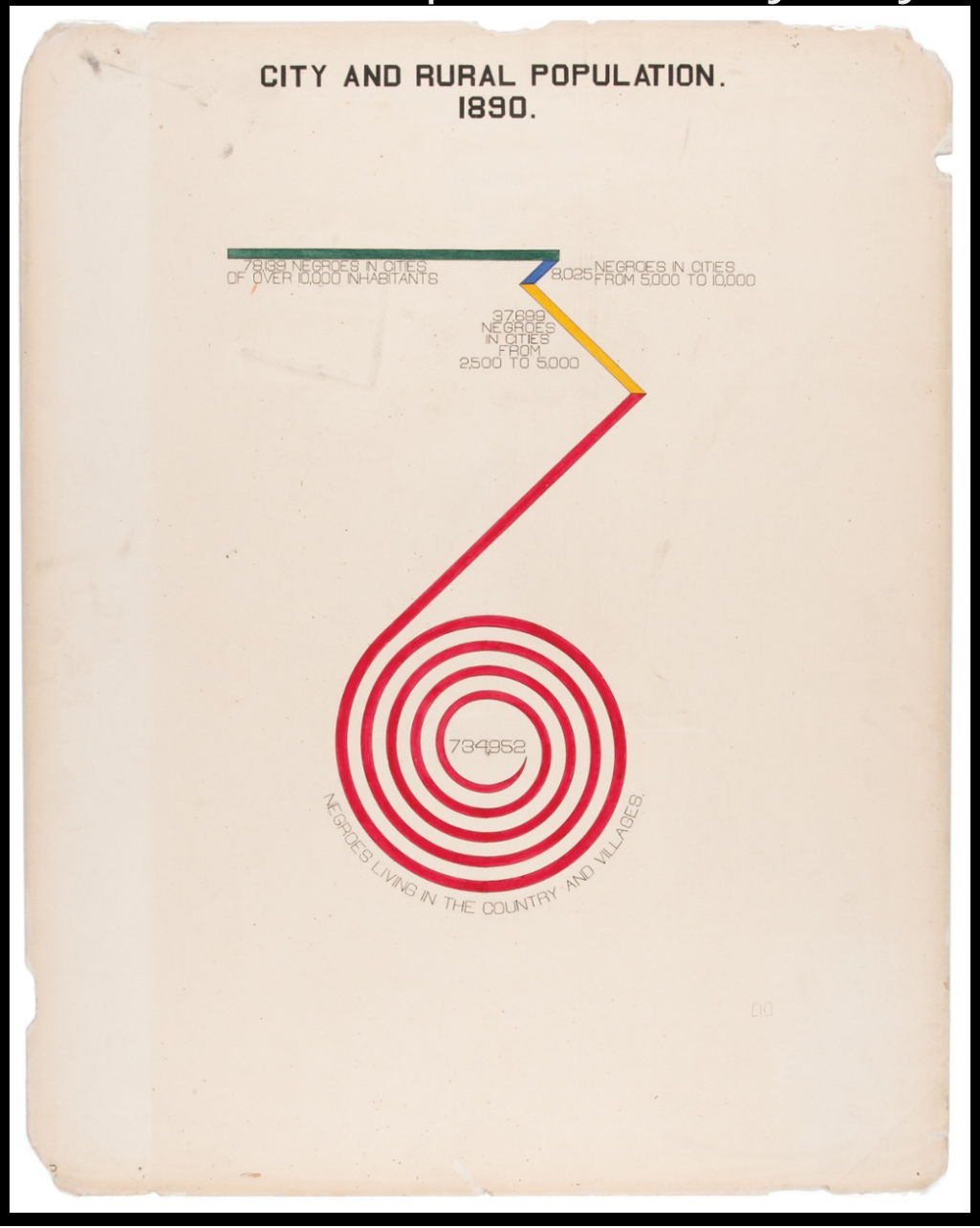


Dr. John Snow 1854

A Cholera Map shows clustering - generally referred to as the first example of data science

https://medium.com/public-health/john-snow-early-big-data-science-d62b4dacd71b

a few historical plots and why they made history



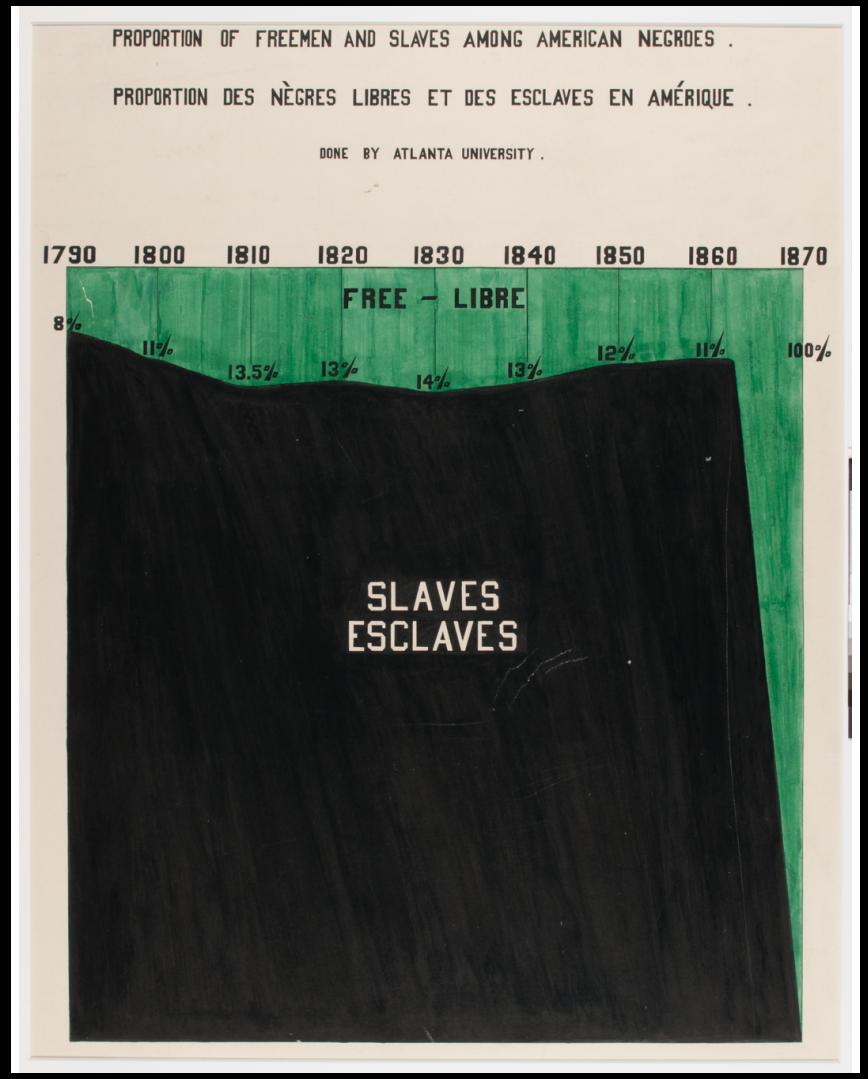
W.E.B. Du Bois

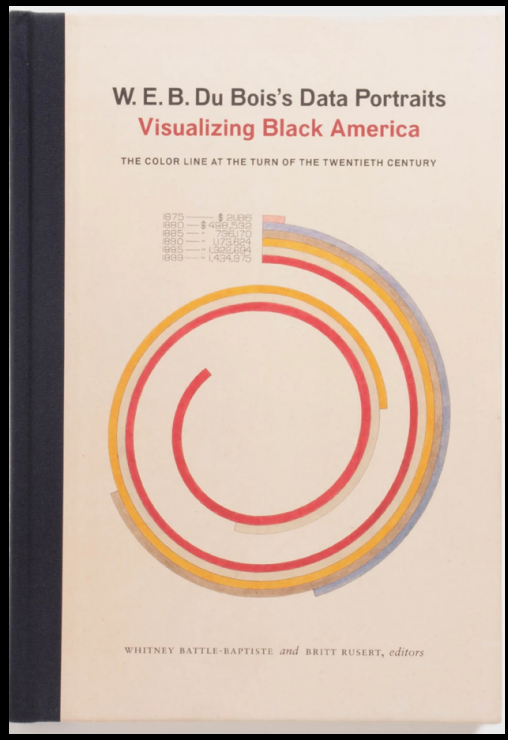
February 23, 1868 – August 27, 1963

American sociologist, socialist, historian, civil rights activist, Pan-Africanist, author, writer and editor

https://inspirehep.net/record/1082448/plots

a few historical plots and why they made history





W.E.B. Du Bois

Smithsonian Magazine

https://policyviz.com/podcast/episode-136-web-dubois-data-portraits/

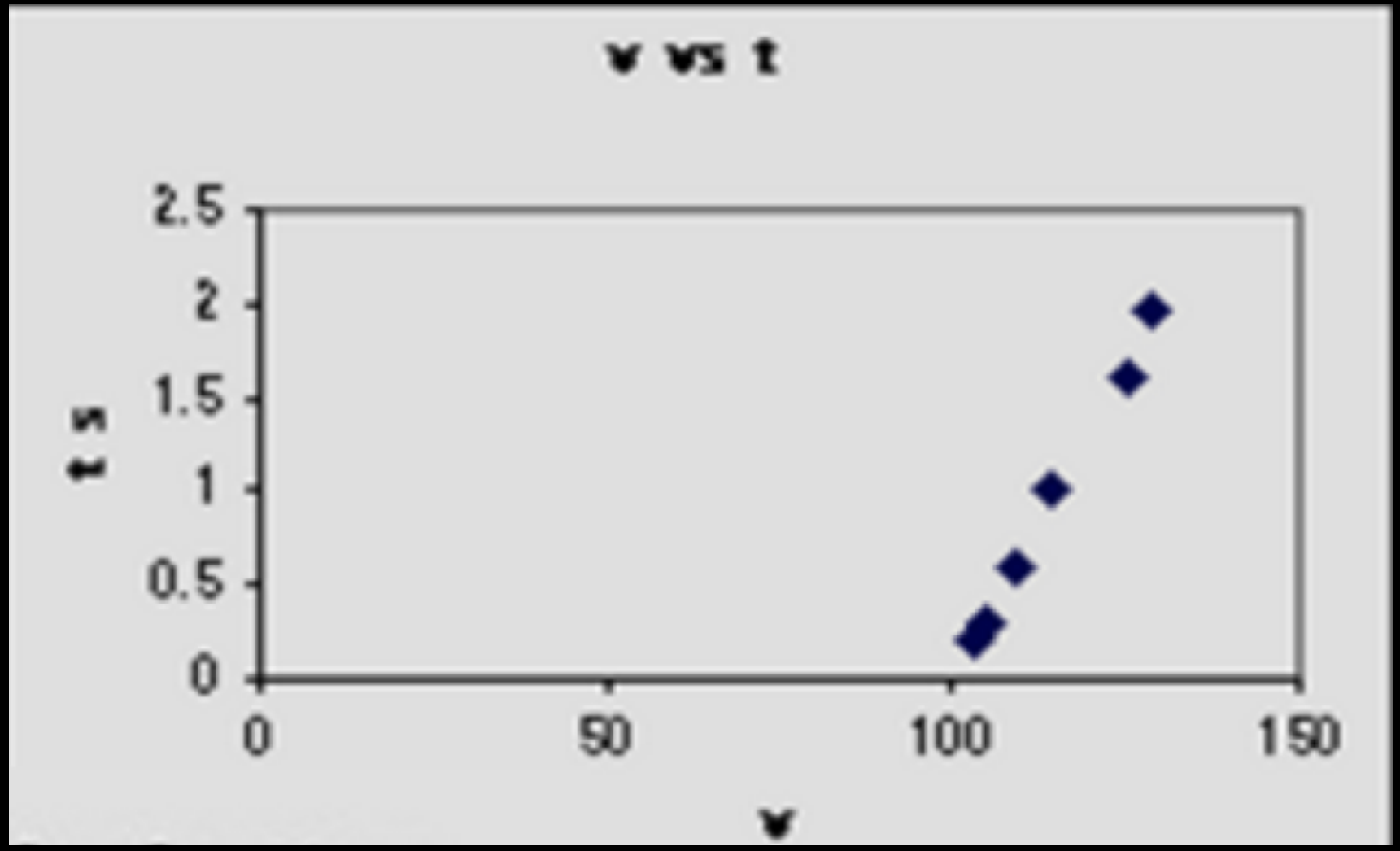
what makes a *bad* visualization?

Ambiguity | distortion | distraction.

Ambiguity | distortion | distraction.

round 1 :

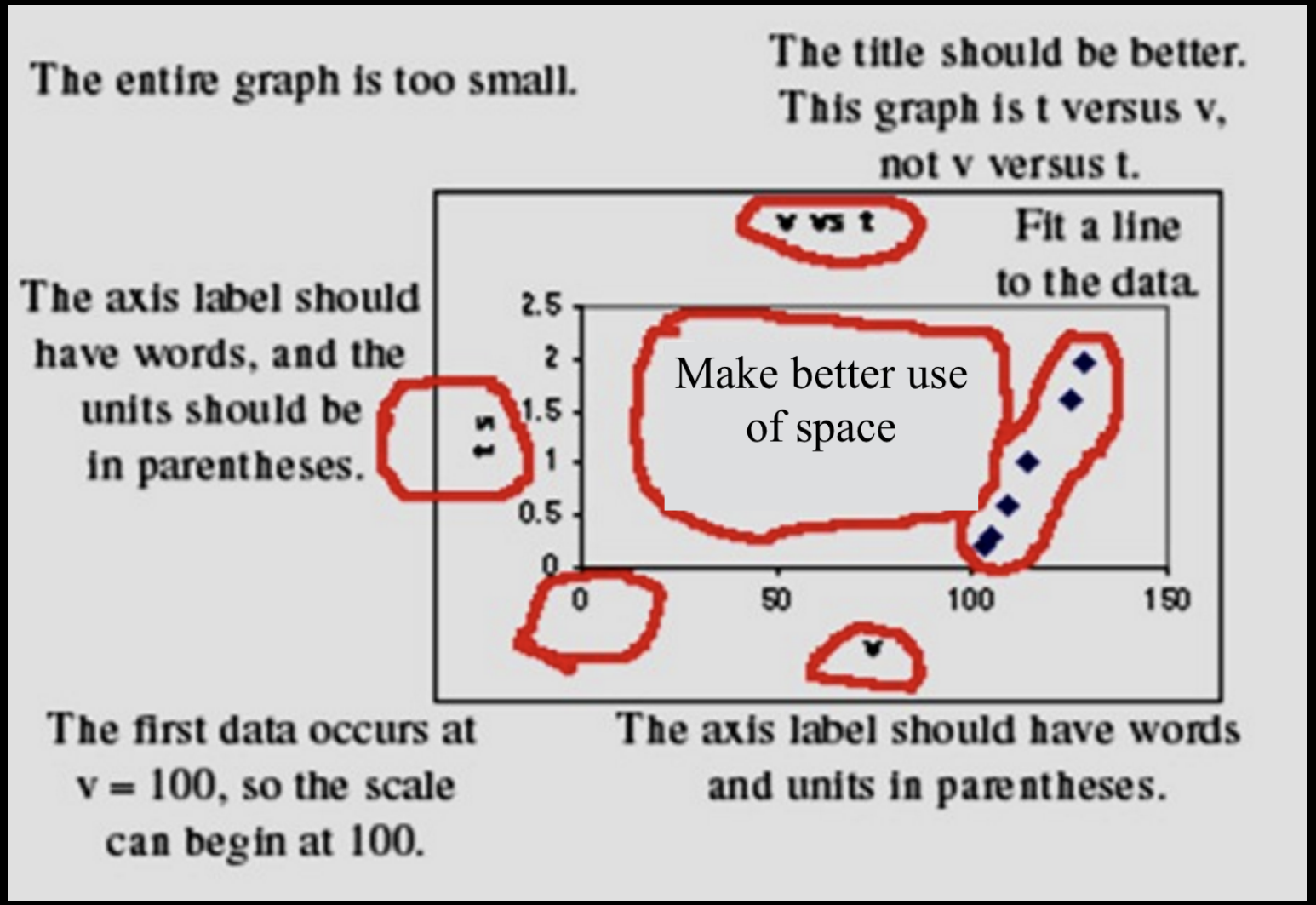
what is wrong with this plot????



6 wrong things with this plot...

Prof. Vern Lindberg

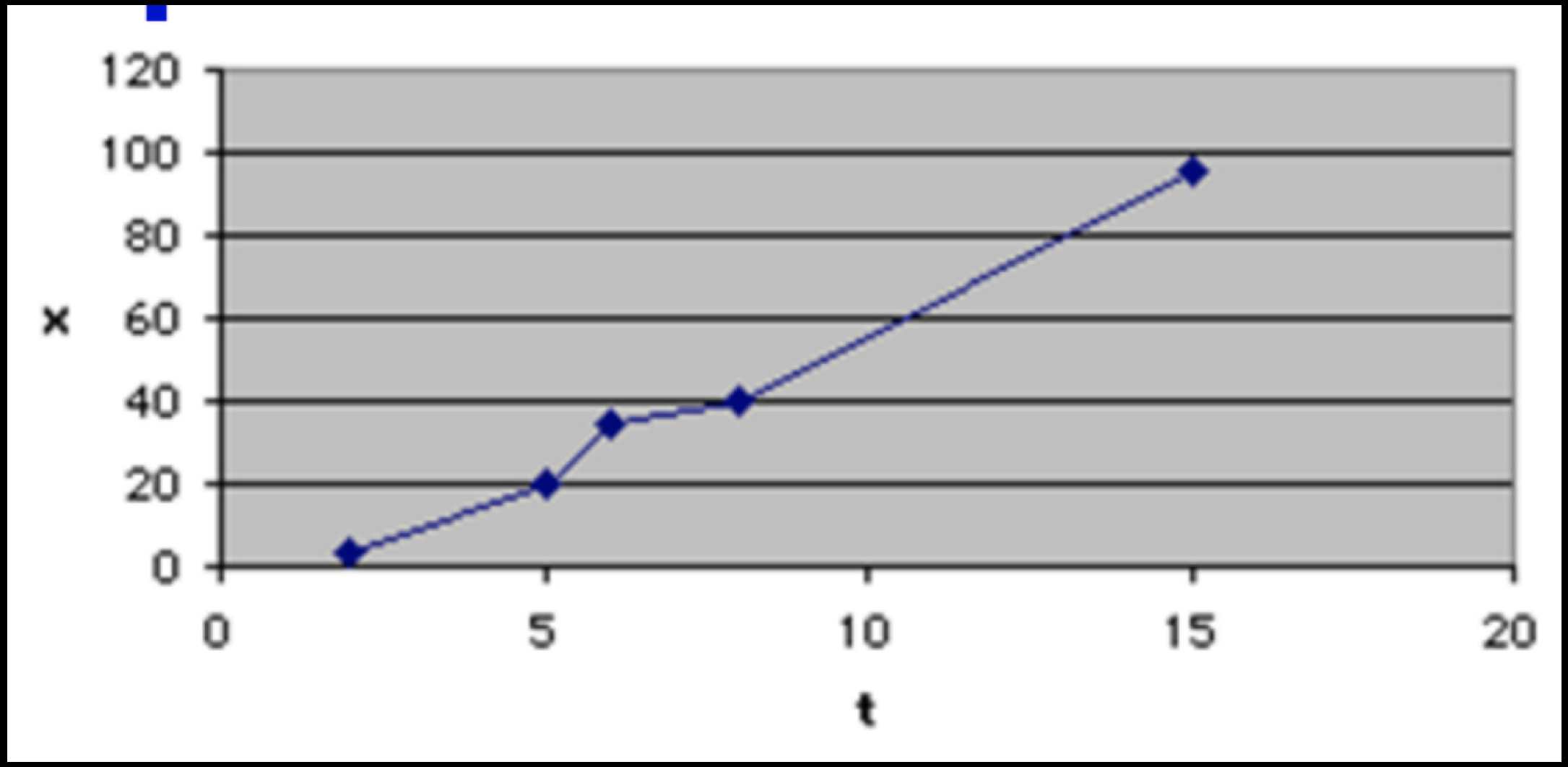
http://www.geol.lsu.edu/jlorenzo/geophysics/graphing/graphingpart3.html



6 wrong things with this plot...

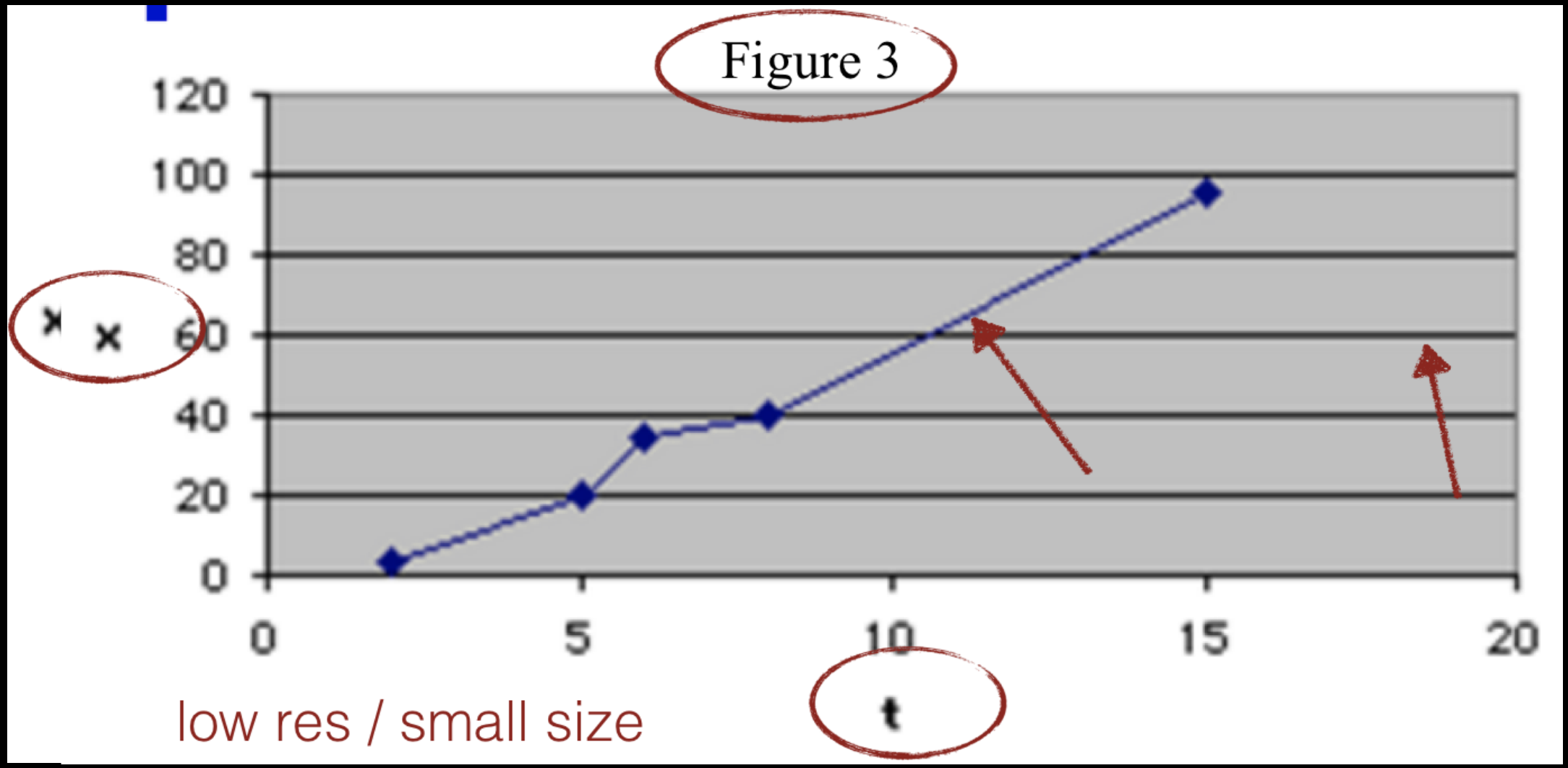
Prof. Vern Lindberg

http://www.geol.lsu.edu/jlorenzo/geophysics/graphing/graphingpart3.html



Prof. Vern Lindberg

http://www.geol.lsu.edu/jlorenzo/geophysics/graphing/graphingpart3.html



Prof. Vern Lindberg

http://www.geol.lsu.edu/jlorenzo/geophysics/graphing/graphingpart3.html

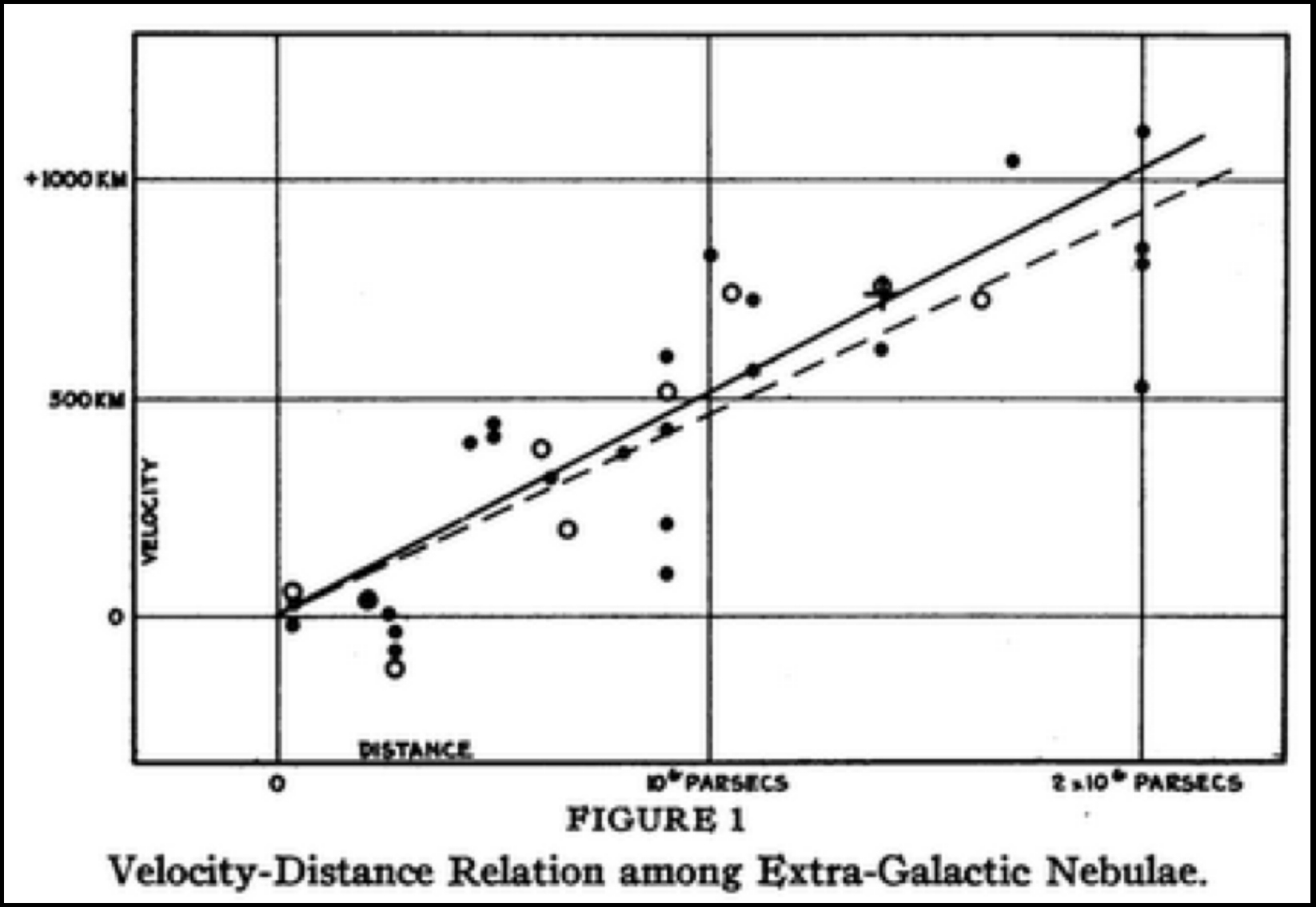


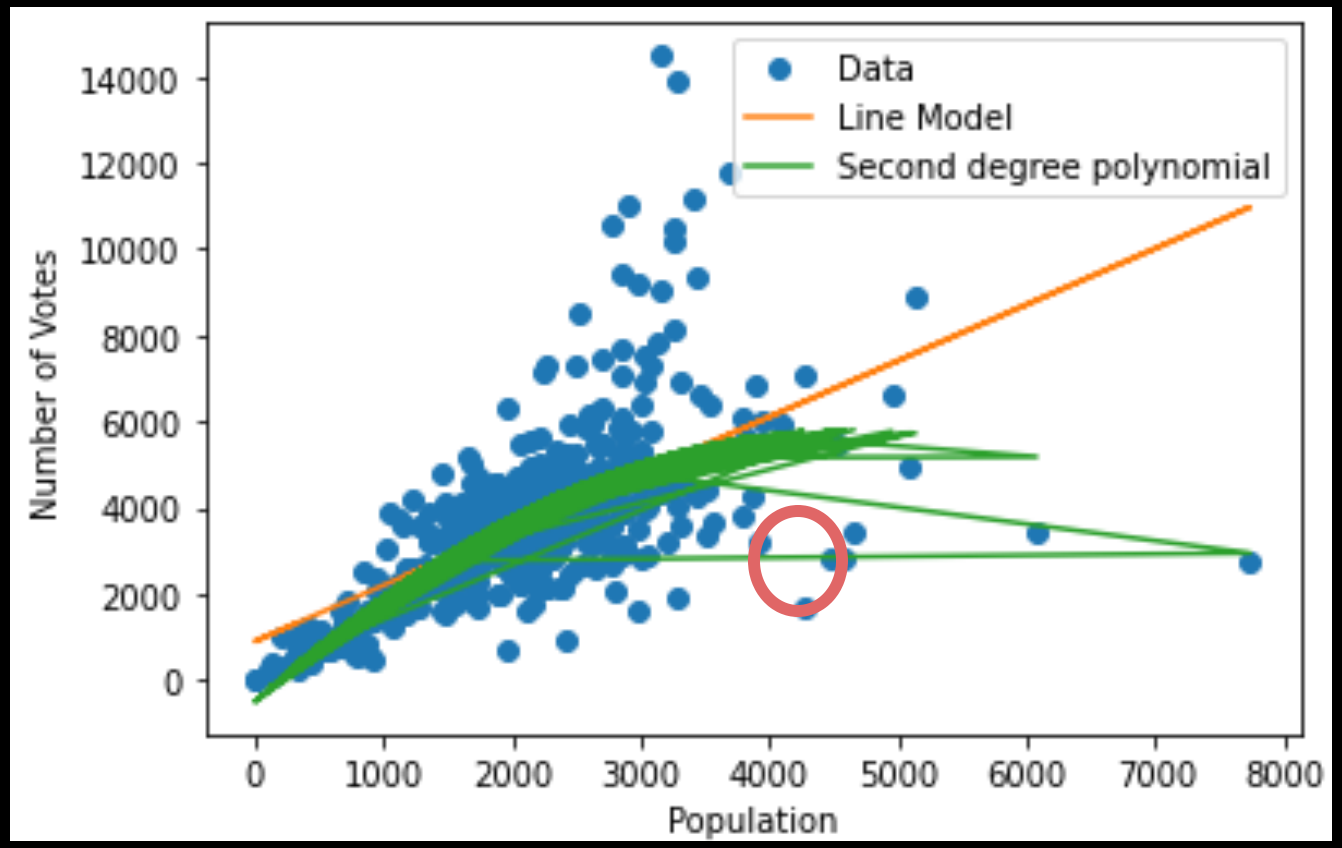
FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

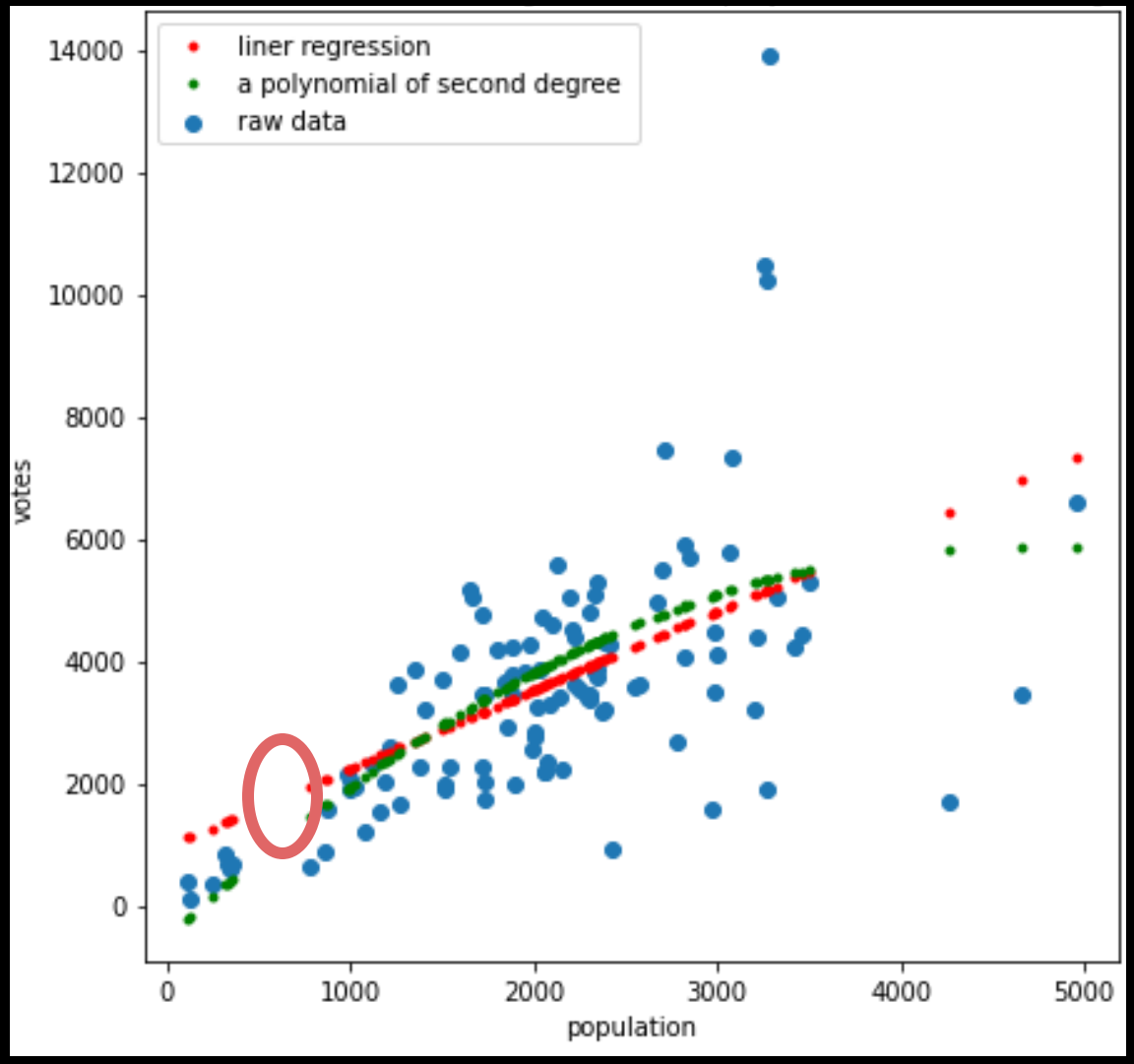
Edwin Hubble

January 17, 1929

Velocity-Distance Relation among Extra-Galactic Nebulae. Radial velocities, corrected for solar motion, are plotted against distances estimated from involved stars and mean luminosities of nebulae in a cluster. The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.



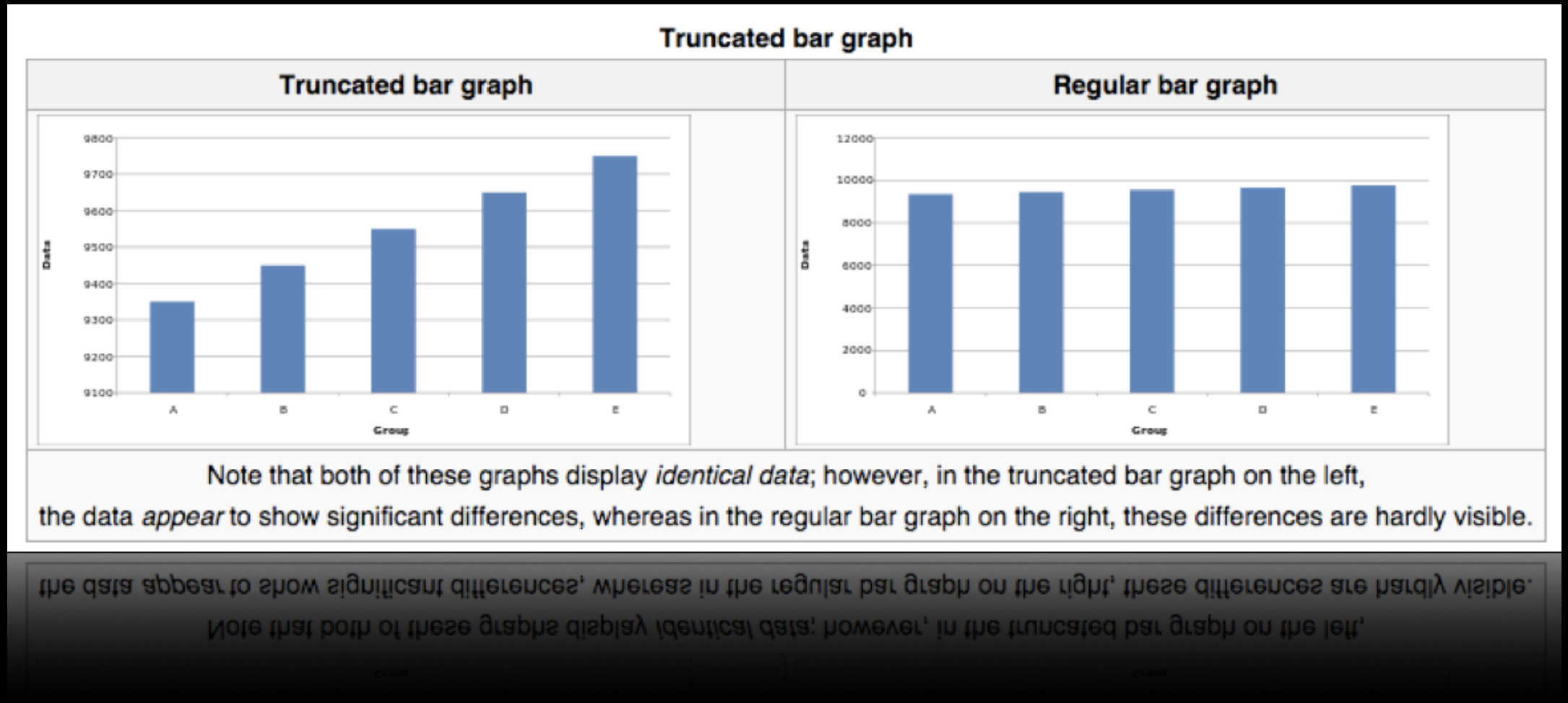
what is the model predicting?



Ambiguity | distortion | distraction.

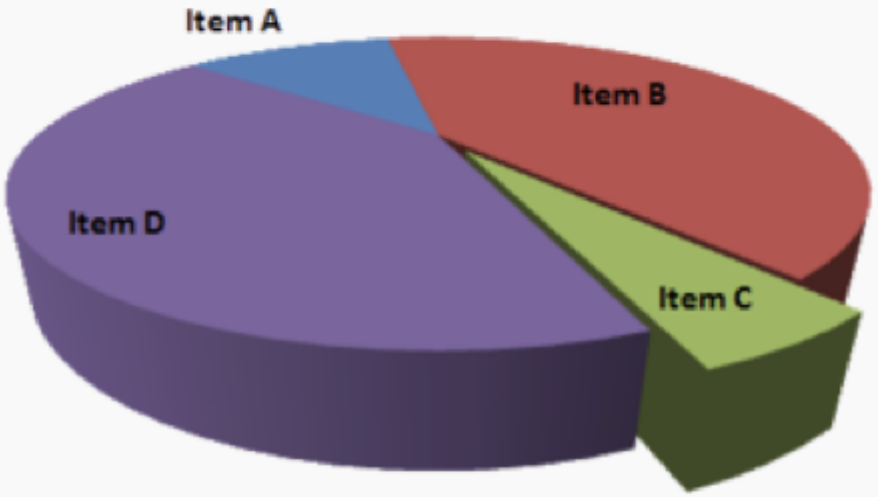
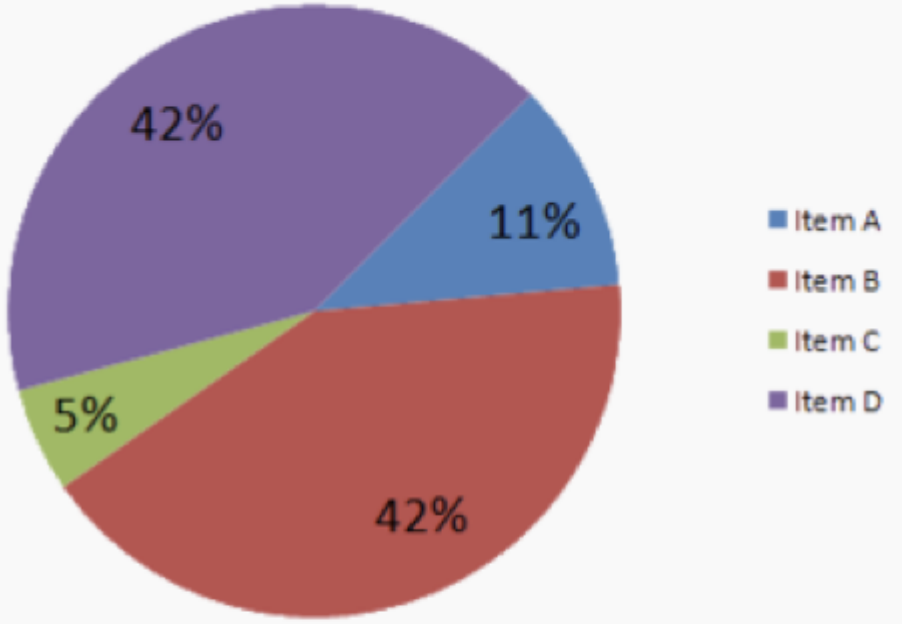
Ambiguity | distortion | distraction.

(=misleading)



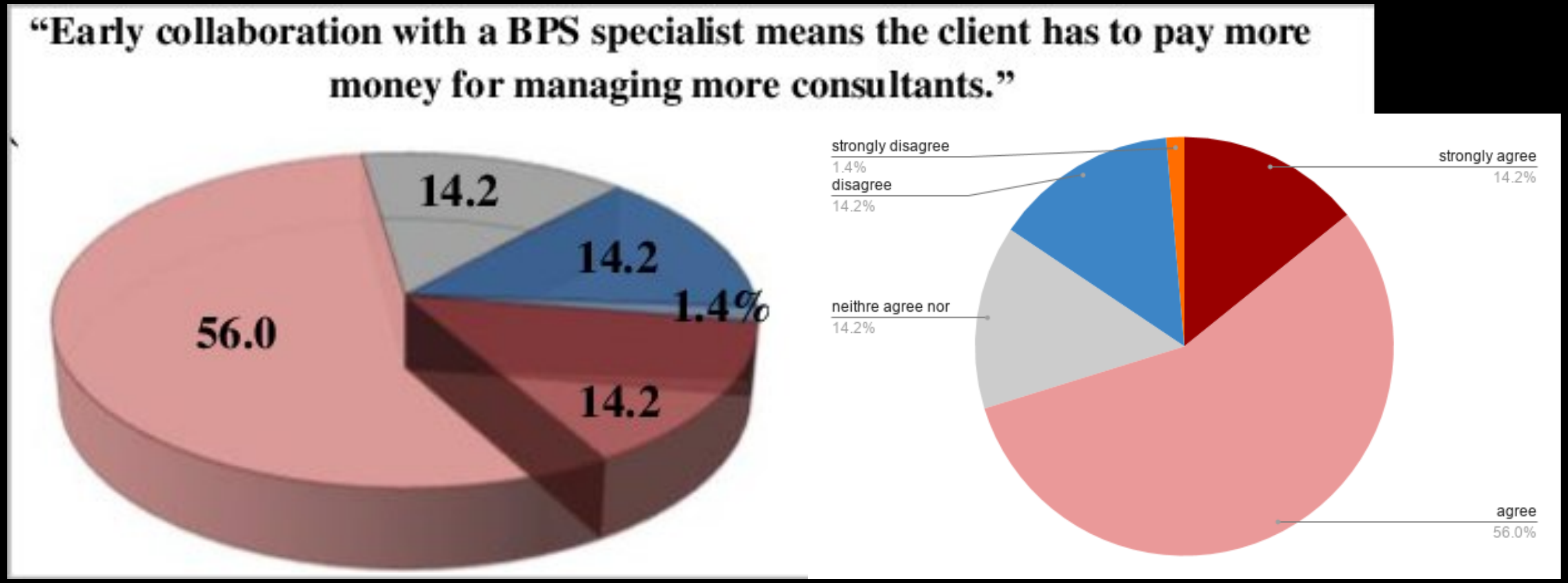
http://gizmodo.com/how-to-lie-with-data-visualization-1563576606

Comparison of pie charts

Misleading pie chart	Regular pie chart
	

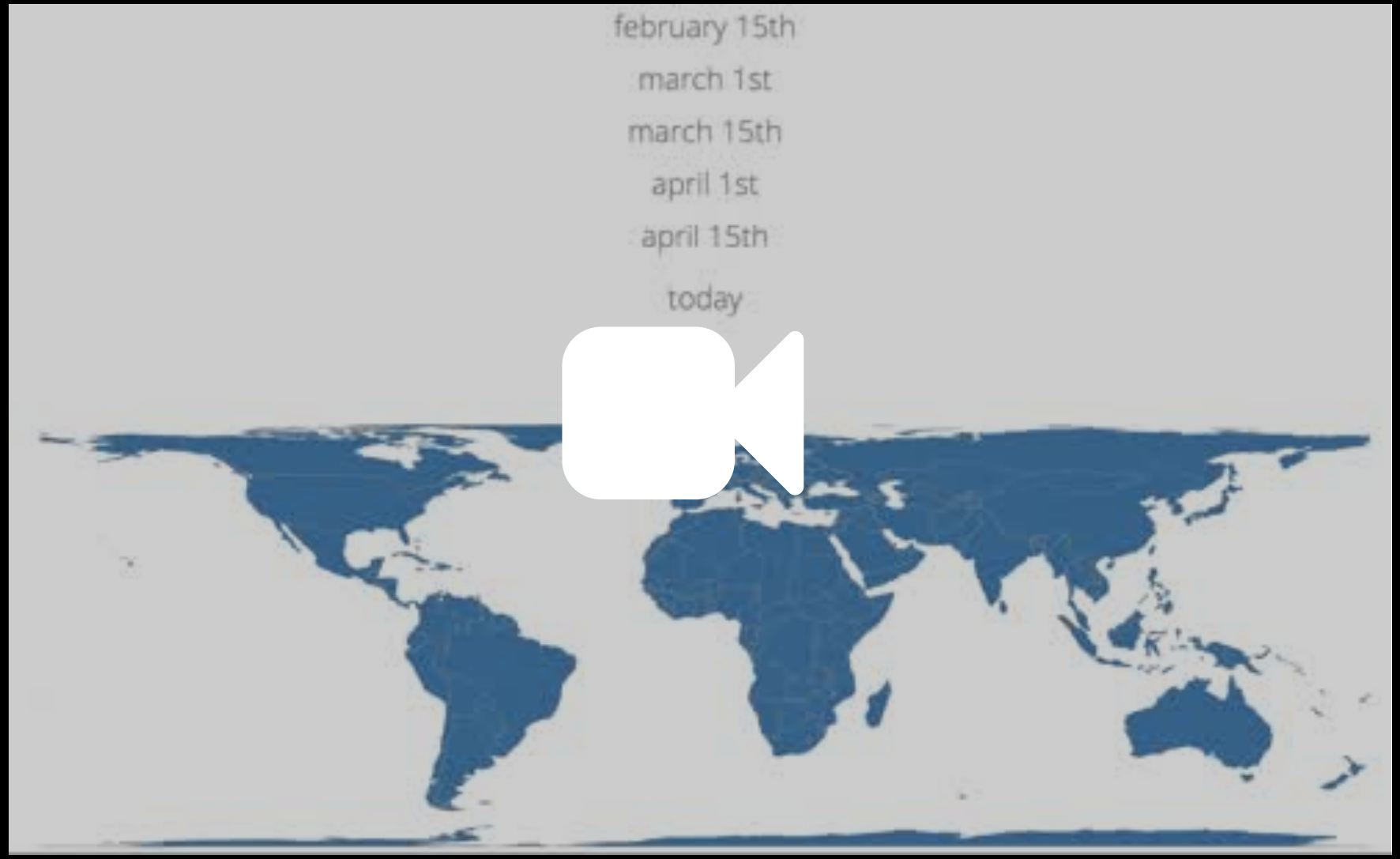
In the misleading pie chart, Item C appears to be at least as large as Item A, whereas in actuality, it is less than half as large.

https://en.wikipedia.org/wiki/Misleading_graph



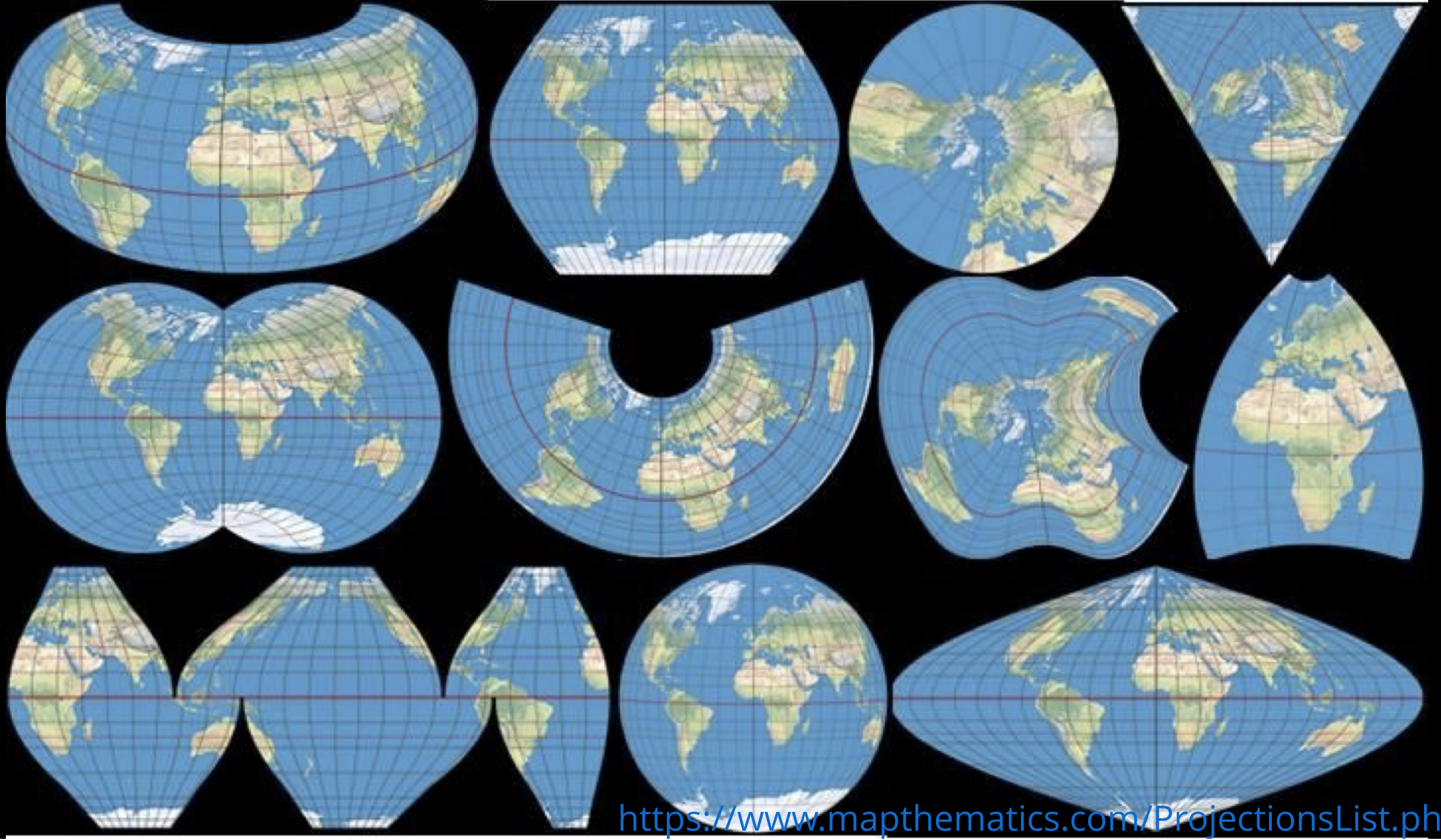
Exactly this plot is in the front page of the Plank collaboration website! Plank is am $800M mission to study the earliest Universe

http://planck.caltech.edu/epo/epo-planckScience5.html



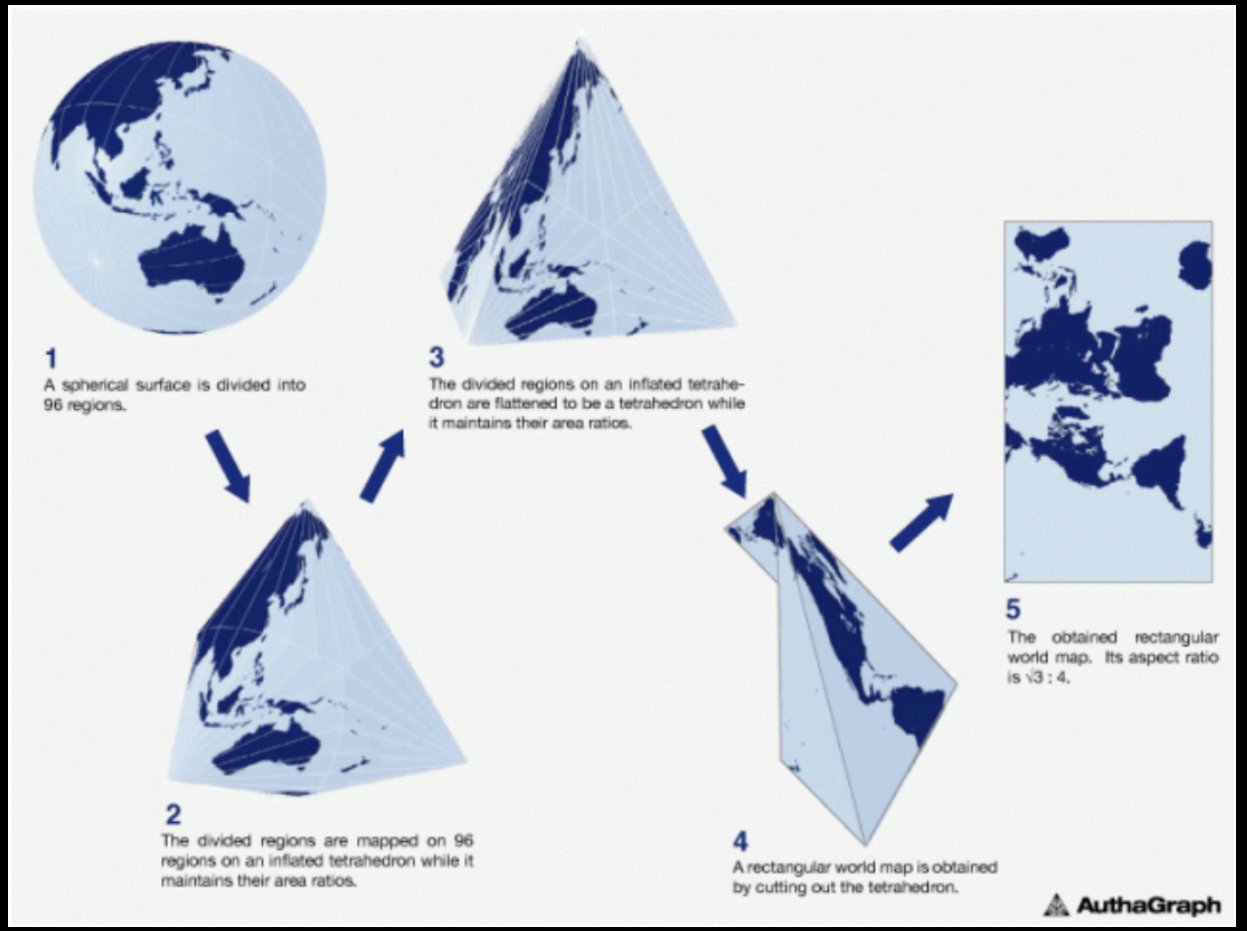
sometimes we use distortion

https://www.d3-graph-gallery.com/cartogram



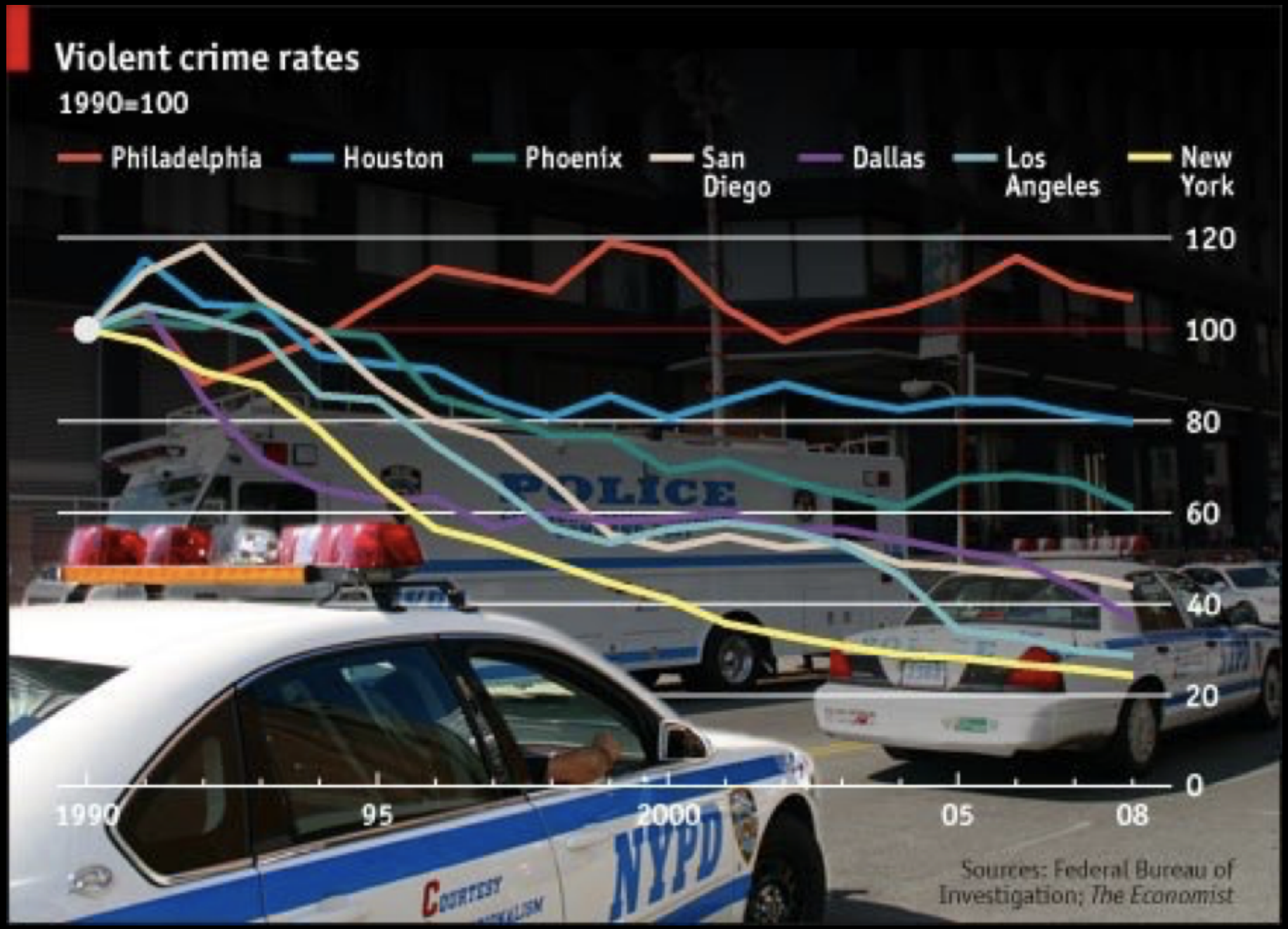
https://www.mapthematics.com/ProjectionsList.php

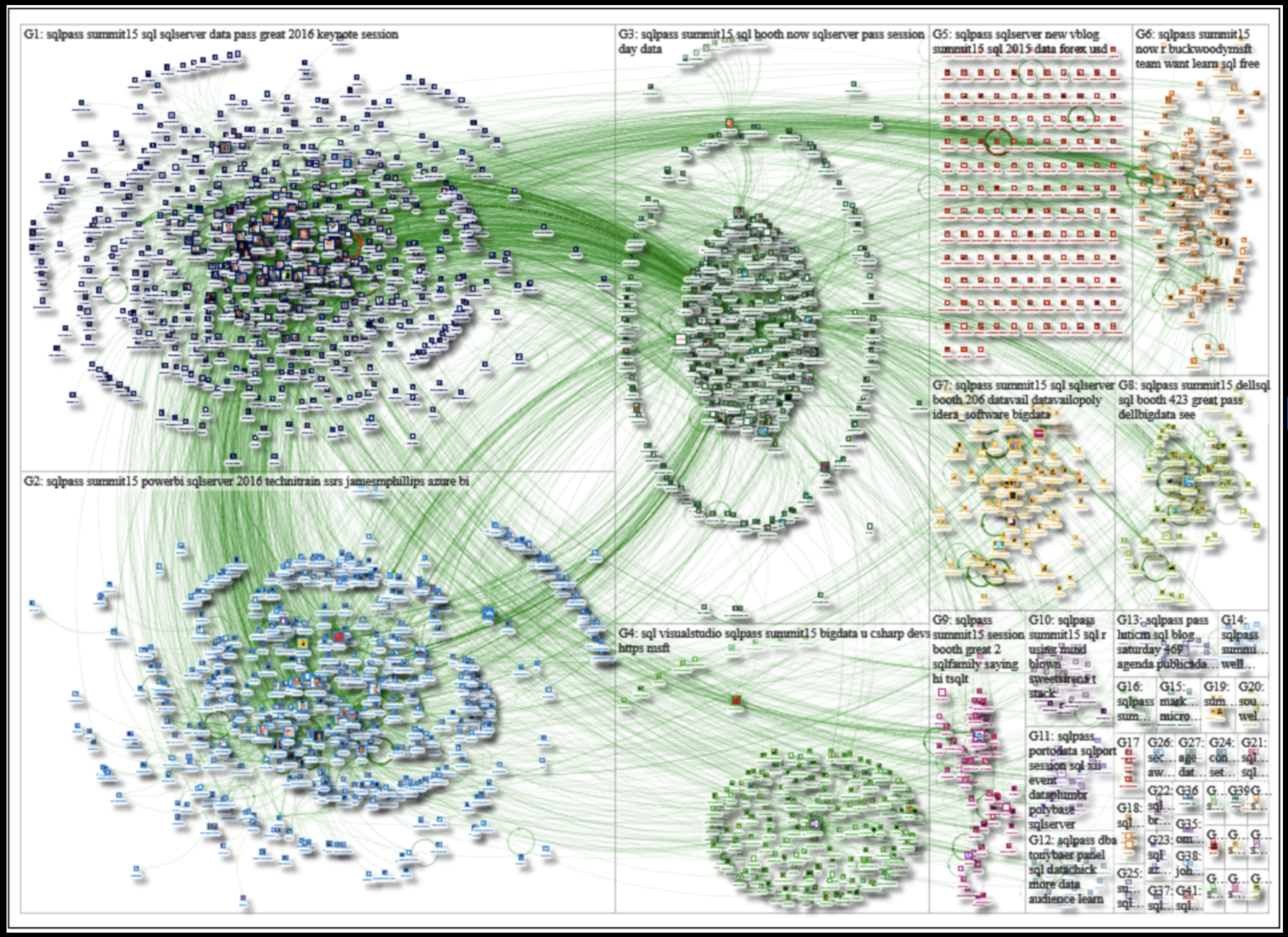
Hajime Narukawa, a Tokyo-based architect and artist, broke the globe up into 96 regions and folded it into a tetrahedron and then a pyramid before finally flattening it into a two-dimensional sheet. This won him the 2016 Japan's prestigious Good Design prize.



http://blogs.discovermagazine.com/d-brief/2016/11/03/most-accurate-world-map/#.XayzIpNKjOQ

Ambiguity | distortion | distraction.


Violent crime rates
1990=100
Philadelphia
Houston
Phoenix
San Diego
Dallas
Los Angeles
New York
120
100
80
60
40
20
0
1990
95
2000
05
08
Sources: Federal Bureau of Investigation; The Economist



Sometime the distraction is a consequence of the complexity of the data.

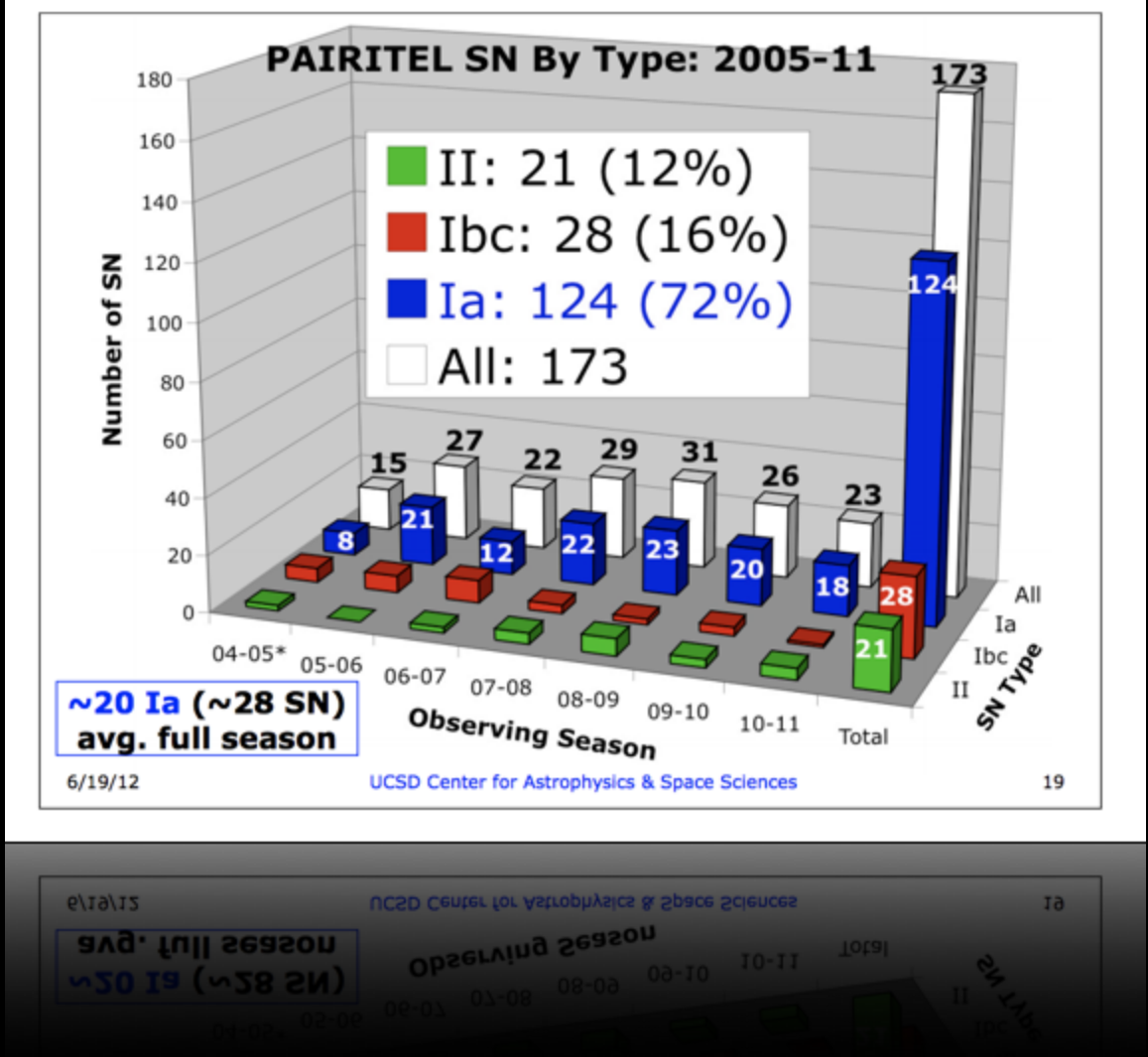
http://www.nodexlgraphgallery.org/Pages/InteractiveGraph.aspx?graphID=56967

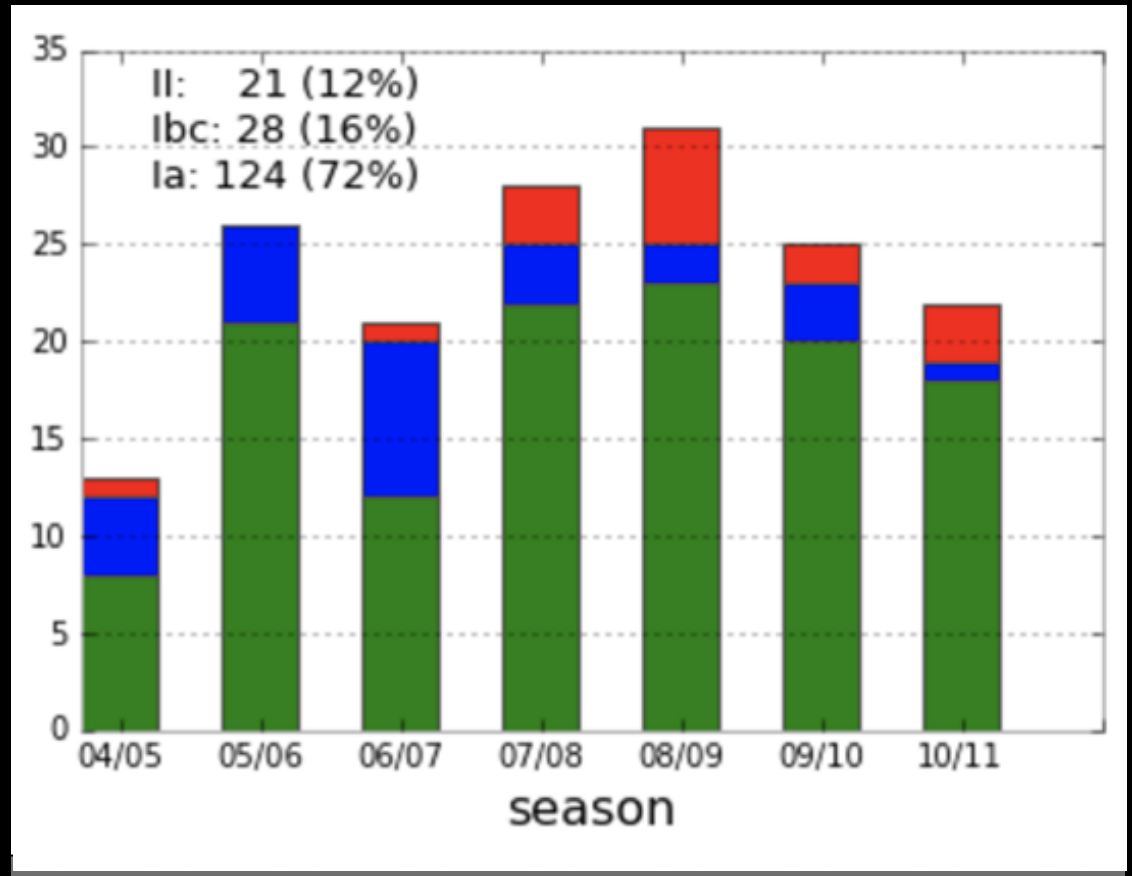
Sometime the distraction is a consequence of the complexity of the data.

https://i.imgur.com/RzYaLZg.gif

http://i.imgur.com/RzYaLZg.gif

Ambiguity, distortion, distraction.

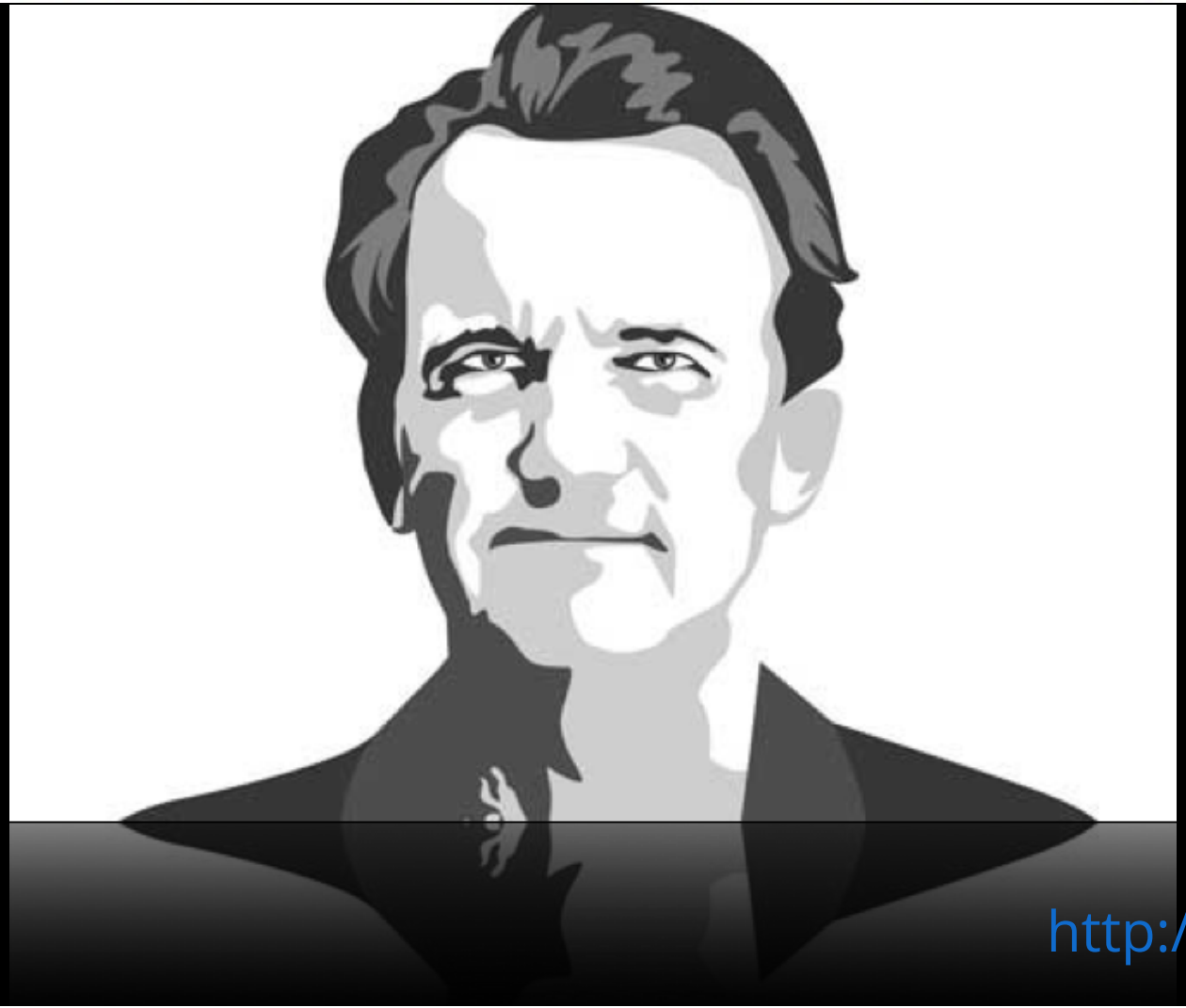




what makes a good visualization?

Tufte's rules

	1999.1.1	65 months	2004.4.28	low	high		2003.4.28	12 months	2004.4.28	low	high
Euro foreign exchange	$	1.1608		1.1907	.8252	1.2858	$	1.1025		1.1907	1.0783	1.2858
Euro foreign exchange	¥	121.32		130.17	89.30	140.31	¥	132.54		130.17	124.80	140.31
Euro foreign exchange	£	0.7111		0.6665	.5711	0.7235	£	0.6914		0.6665	0.6556	0.7235



Edward Tufte

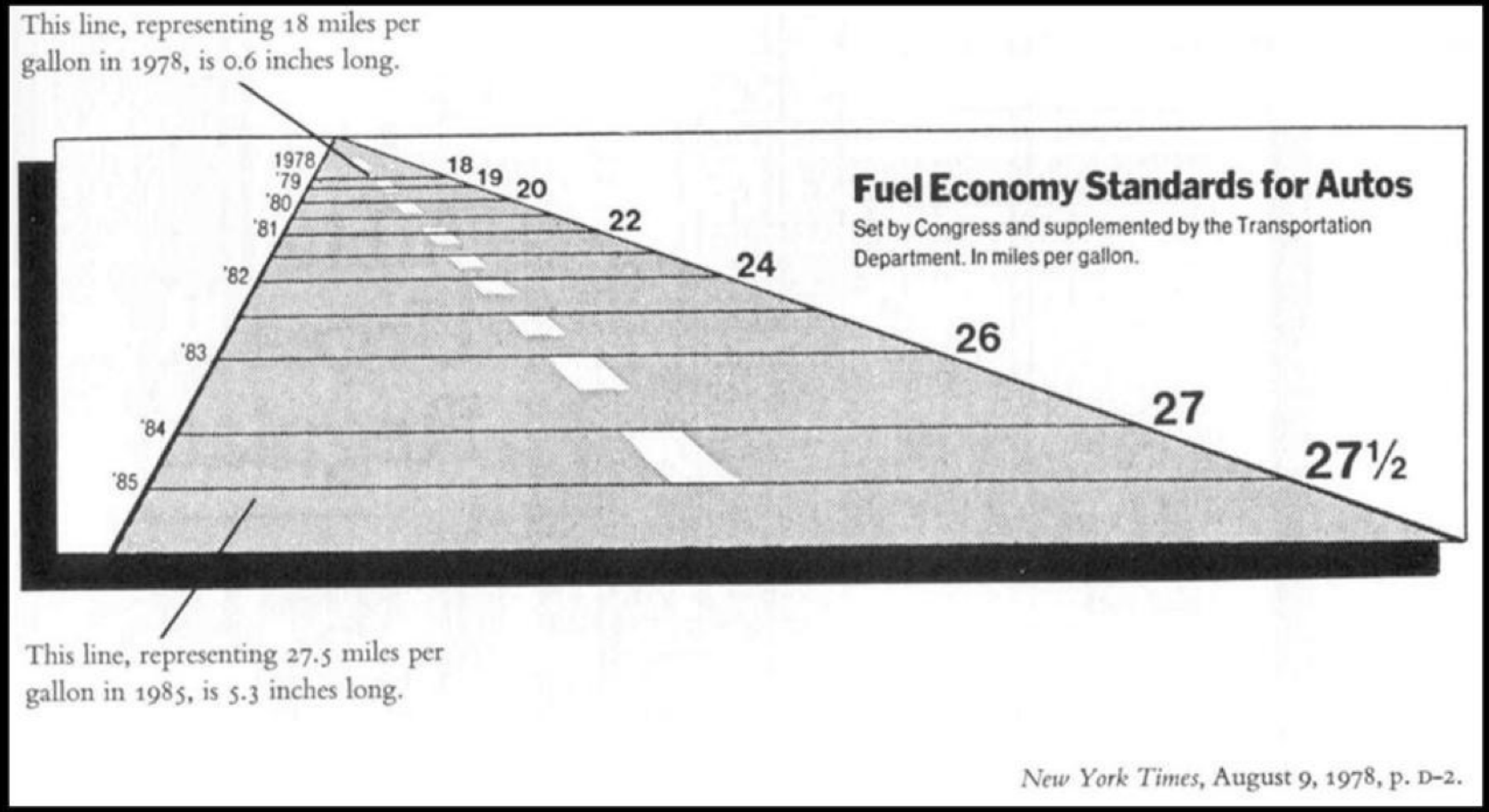
http://okhaos.com/tufte.pdf

Tufte’s rules:

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

Tufte's rules:

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

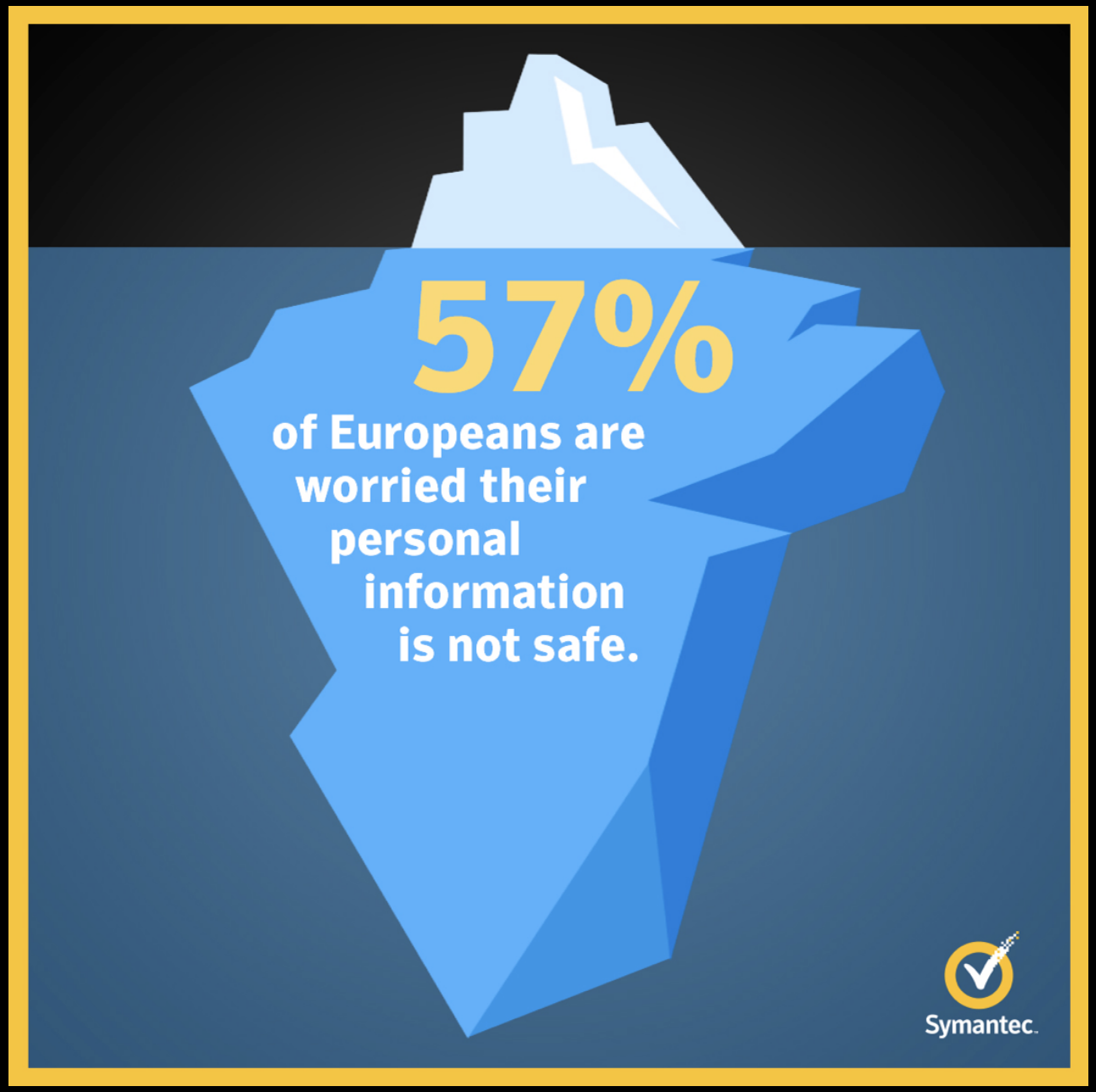


Tufte’s rules:

1. The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented ("lie factor")

effect size ~ 1

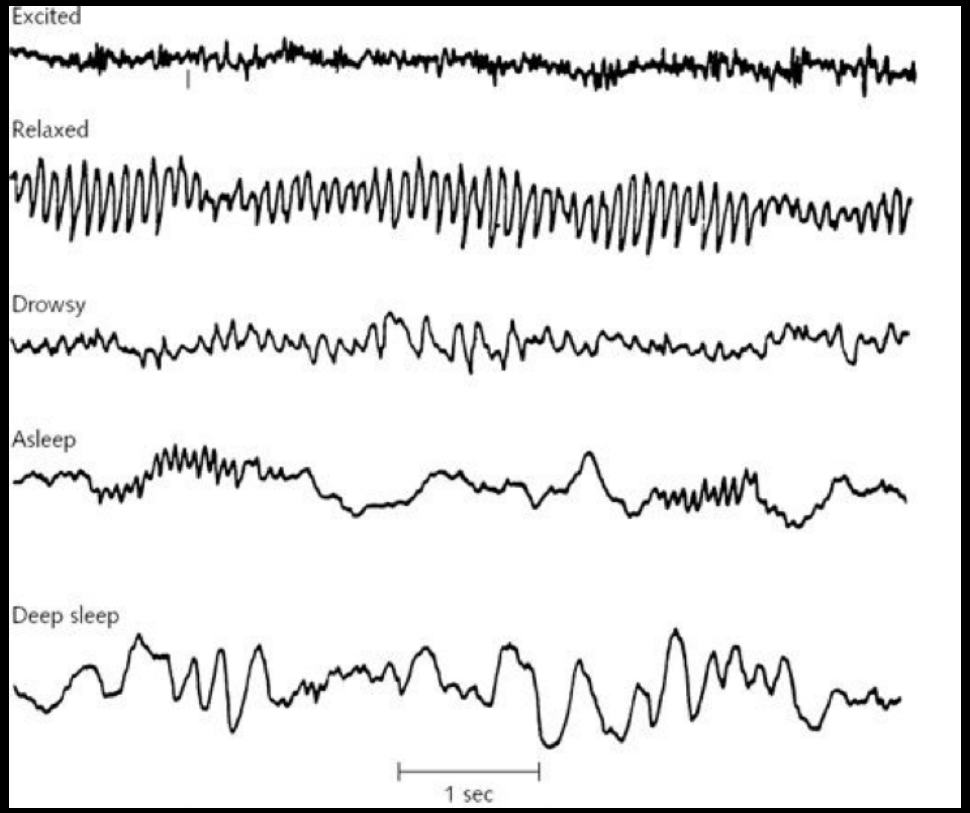
Tufte’s rules:



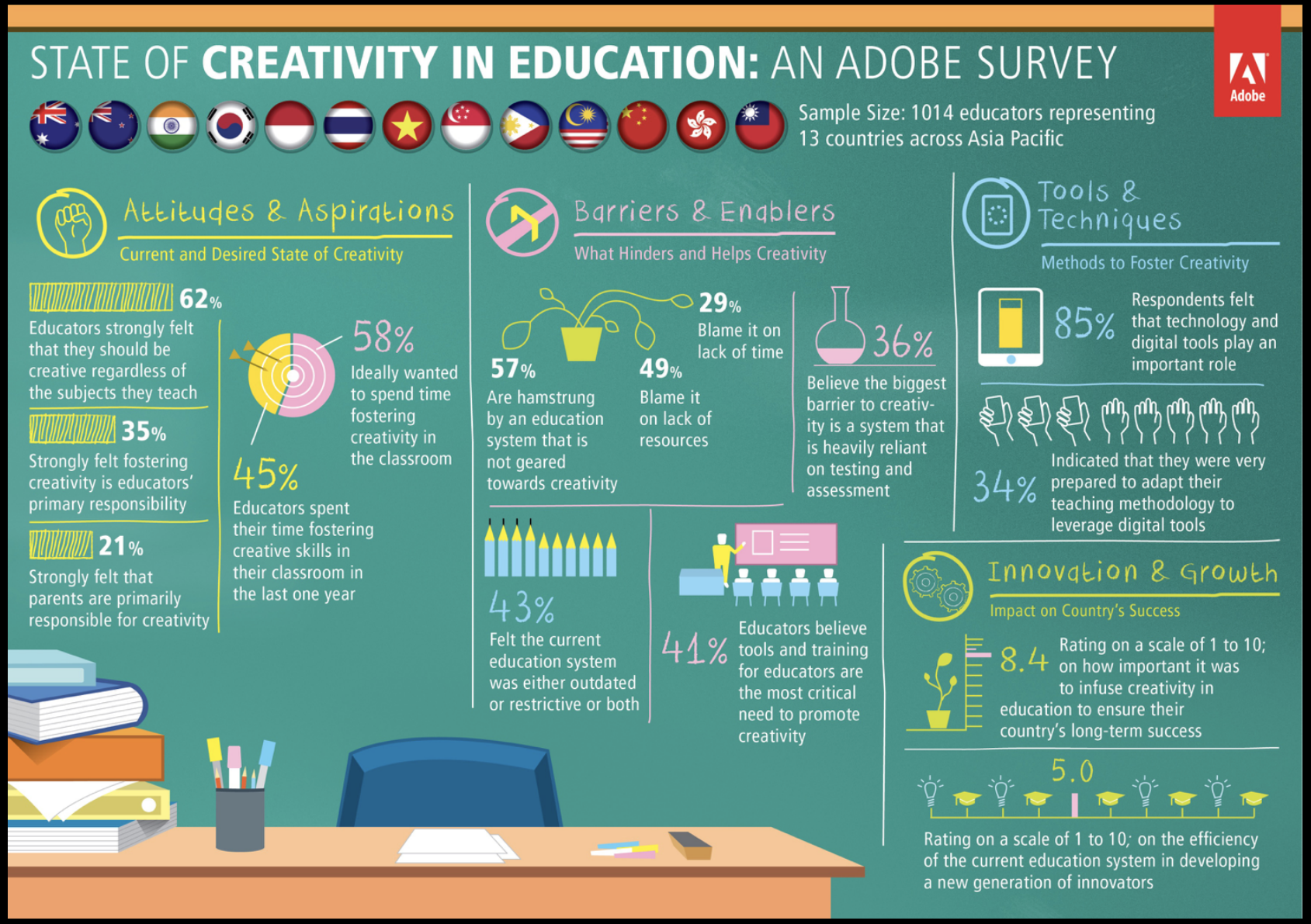
Keep lie factor ~1

Tufte’s rules:

$$\textbf{Data-ink ratio} = \frac{\text{amount of data}}{\text{amount of ink}}$$



Tufte’s rules:



Tufte’s rules:

1. The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented ("lie factor")

effect size ~ 1

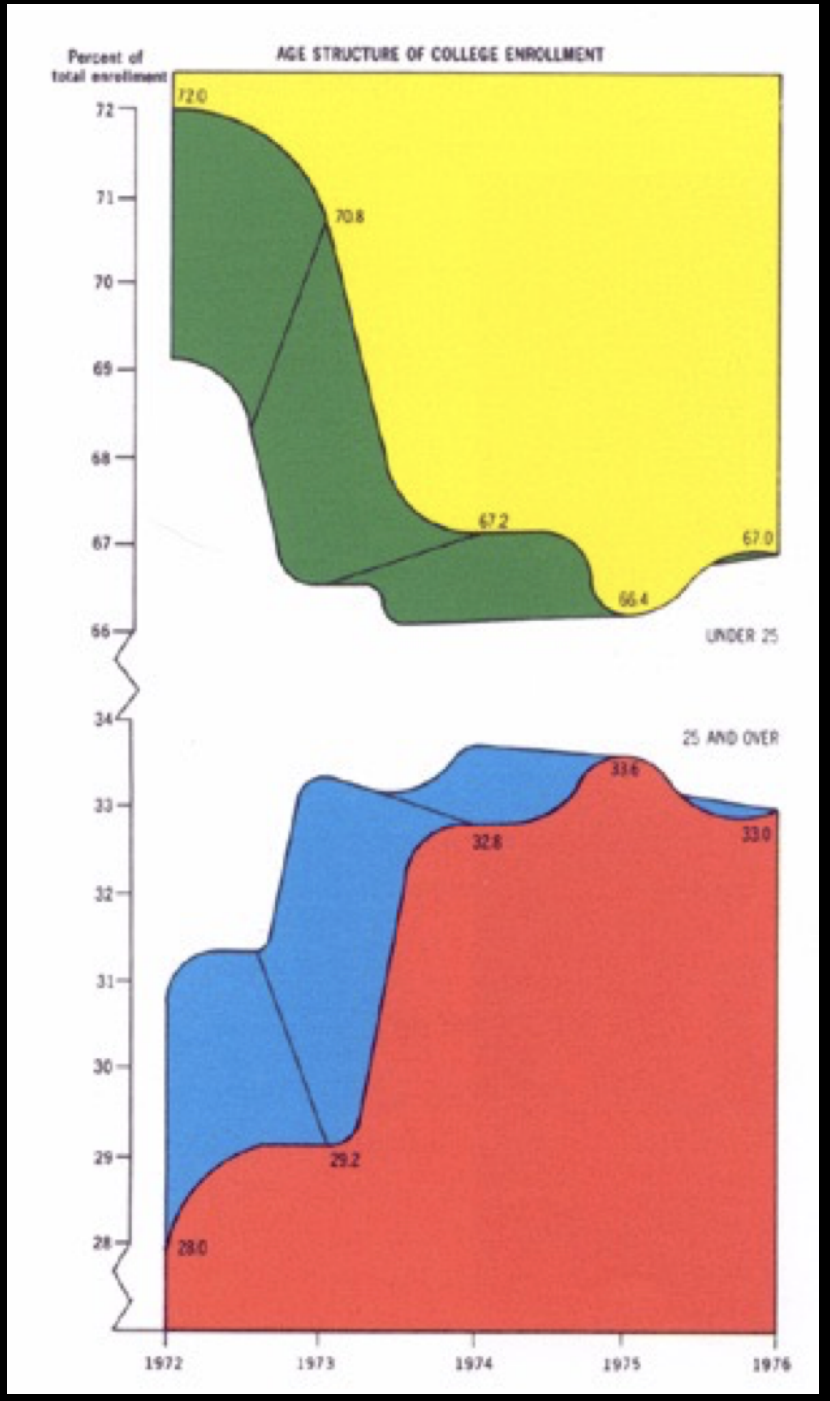
2. Clear, detailed and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graph itself. Label important events in the data

data/ink -> large

Tufte’s rules:

Chart Junk

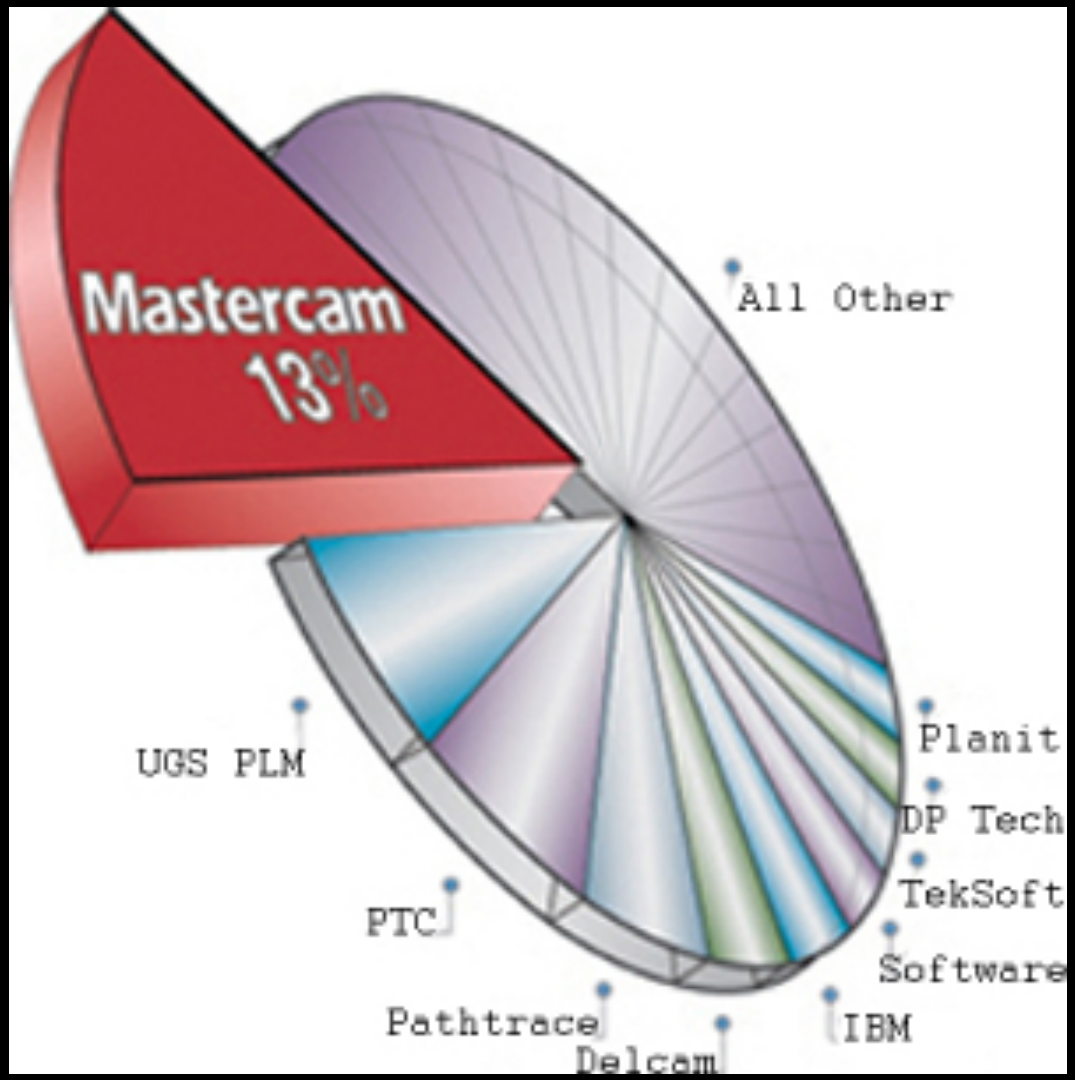
the excessive and unnecessary use of graphical effects



Tufte’s rules:

Chart Junk

the excessive and unnecessary use of graphical effects



Tufte’s rules:

Chart Junk

the excessive and unnecessary use of graphical effects

Tufte’s rules:

1. The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented ("lie factor")

effect size ~ 1

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

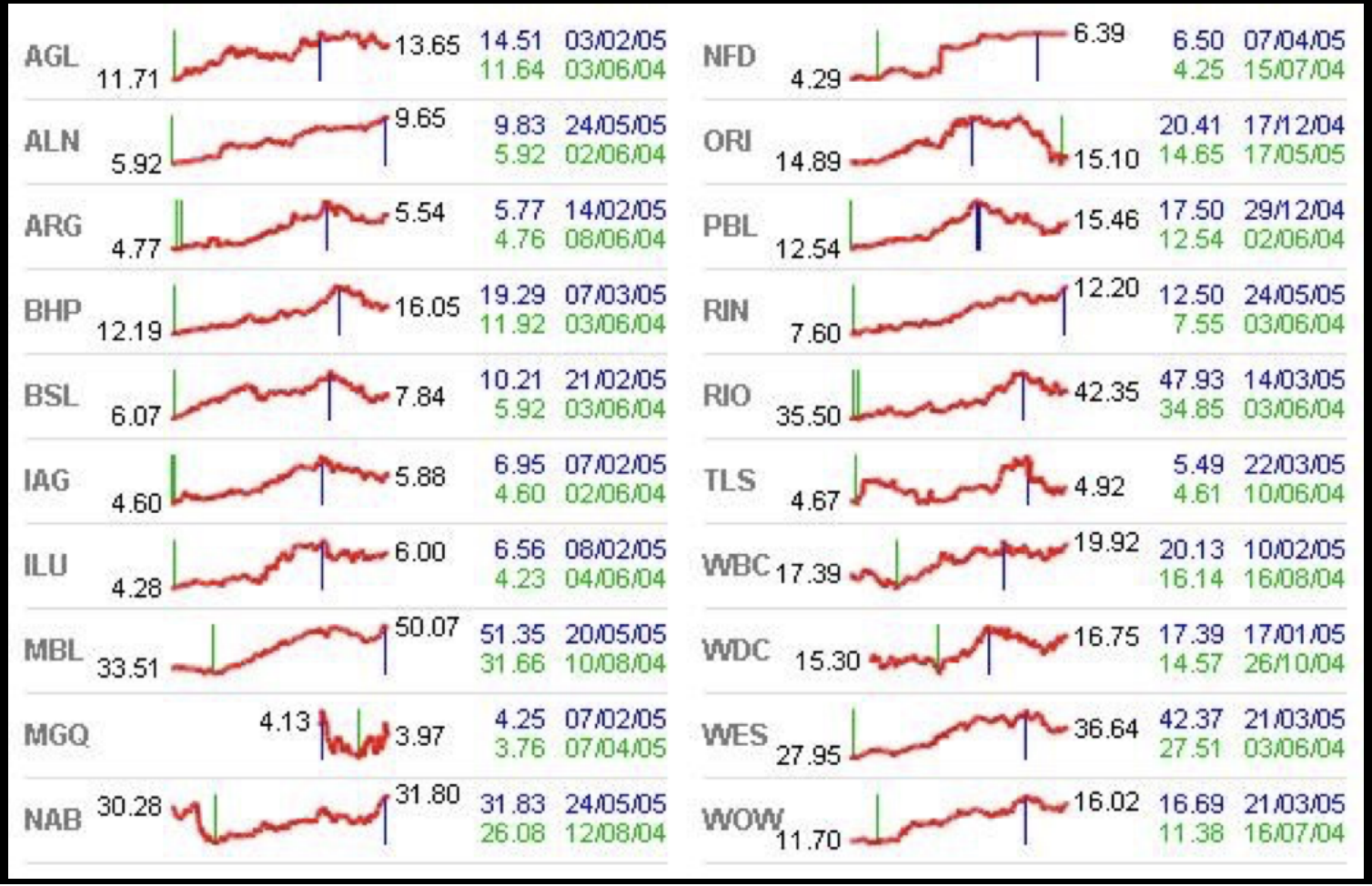
data/ink -> large

3. Show data variation, not design variation

no chart junk

Tufte’s rules:

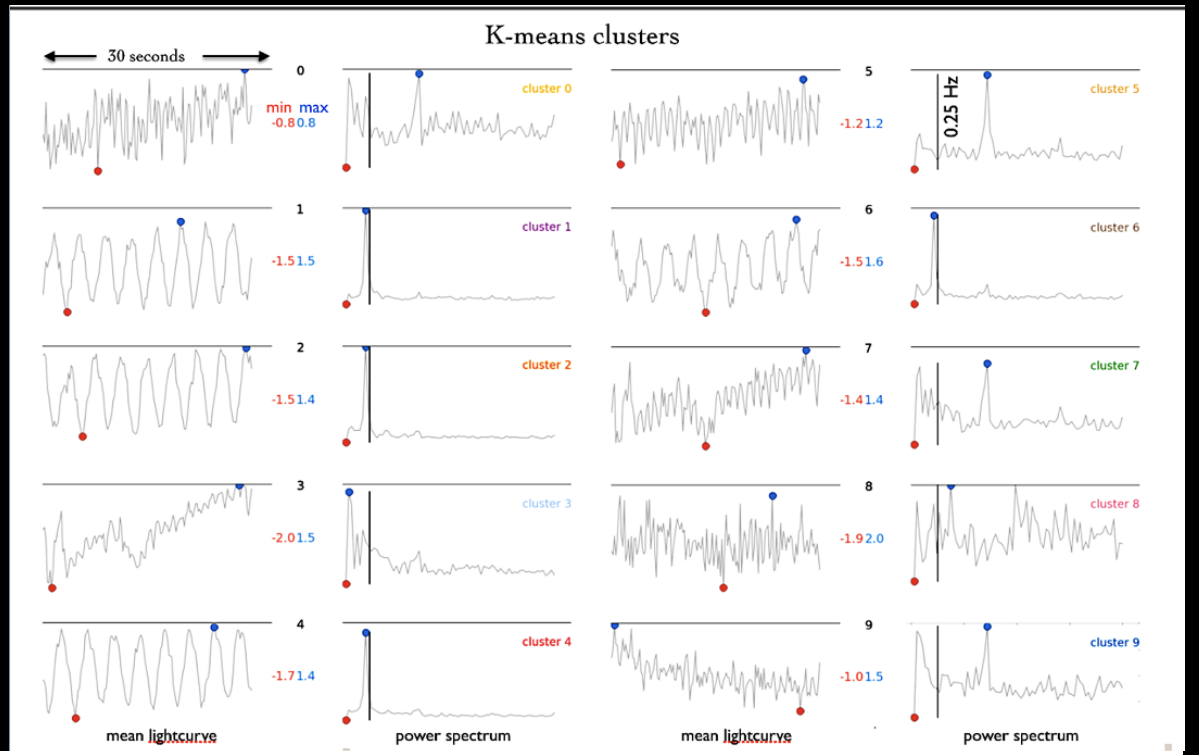
Small multiples encourage comparison



sparkline graph

Tufte’s rules:

Small multiples



sparklpy

DOI 10.5281/zenodo.35387 Code Health

module to create Tufte-style spark line plots for time series, including astronomical ones (in magnitude!)

sparkline graph

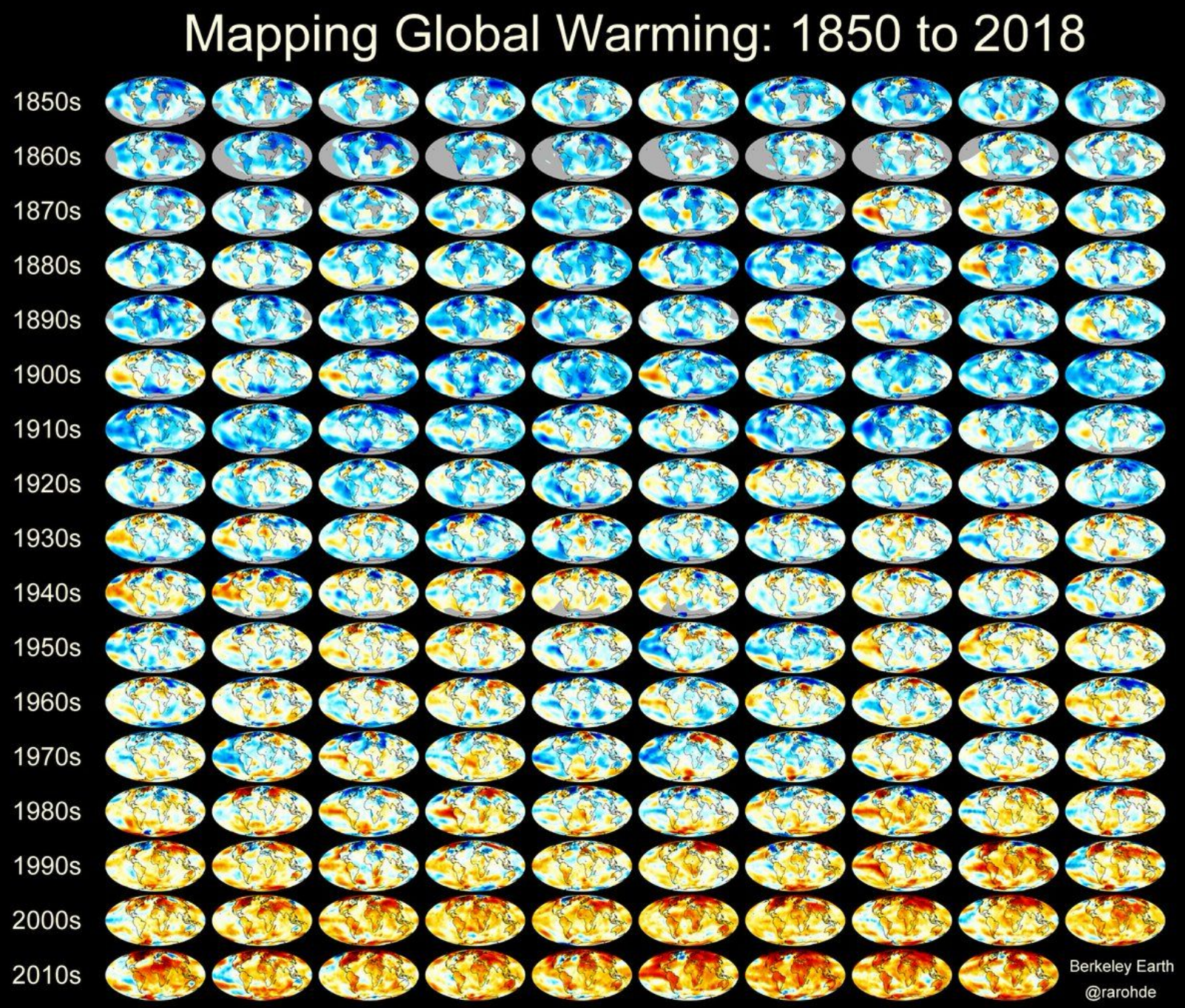
https://github.com/fedhere/sparklpy

Tufte's rules:

Small multiples

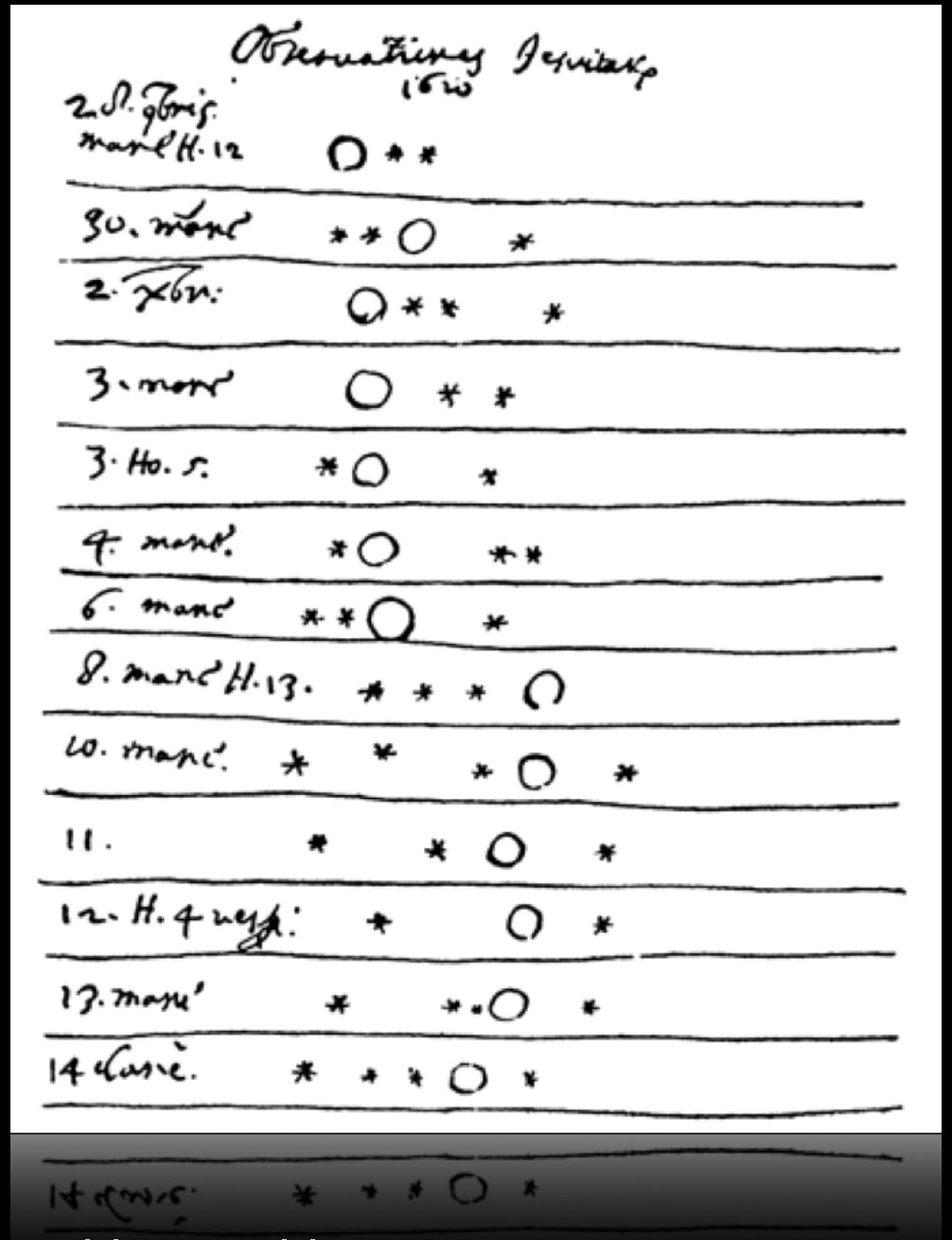
work really well with maps!

https://mahb.stanford.edu/whats-happening/167-tiny-maps-tell-major-story-climate-change/



Tufte’s rules:

Small multiples



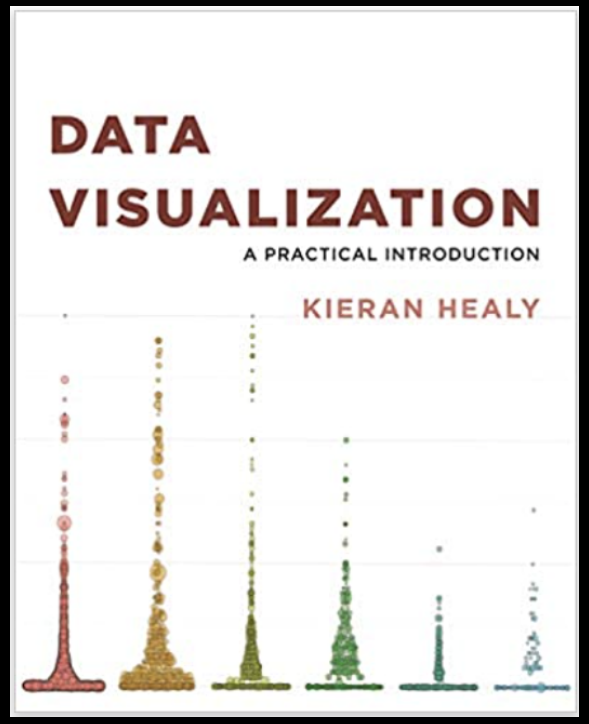
Galileo Galilei, Jupiter moons, 1610

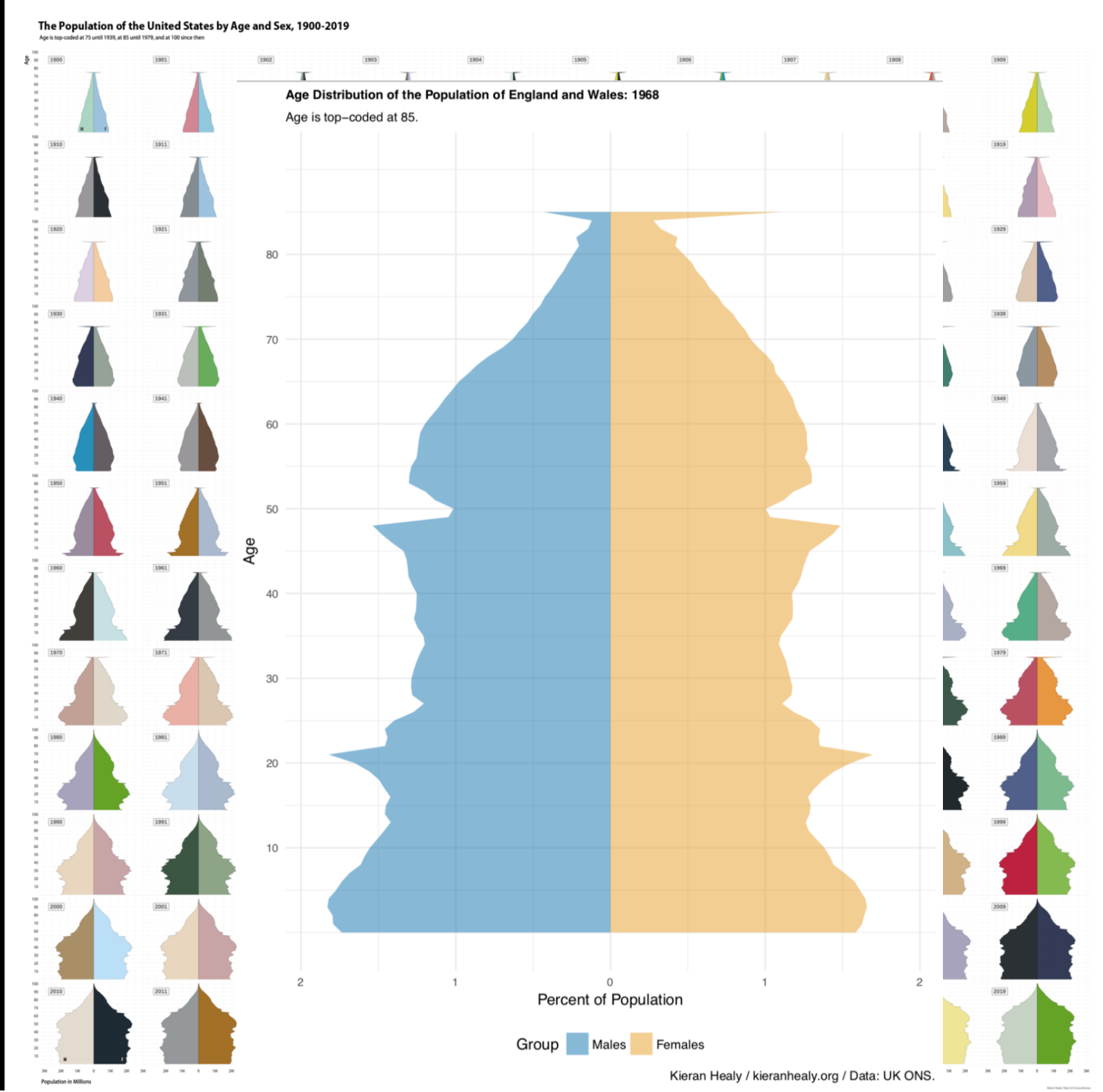
Tufte’s rules:

Small multiples

Keiran Healy

(Data Viz A Practical Intro)



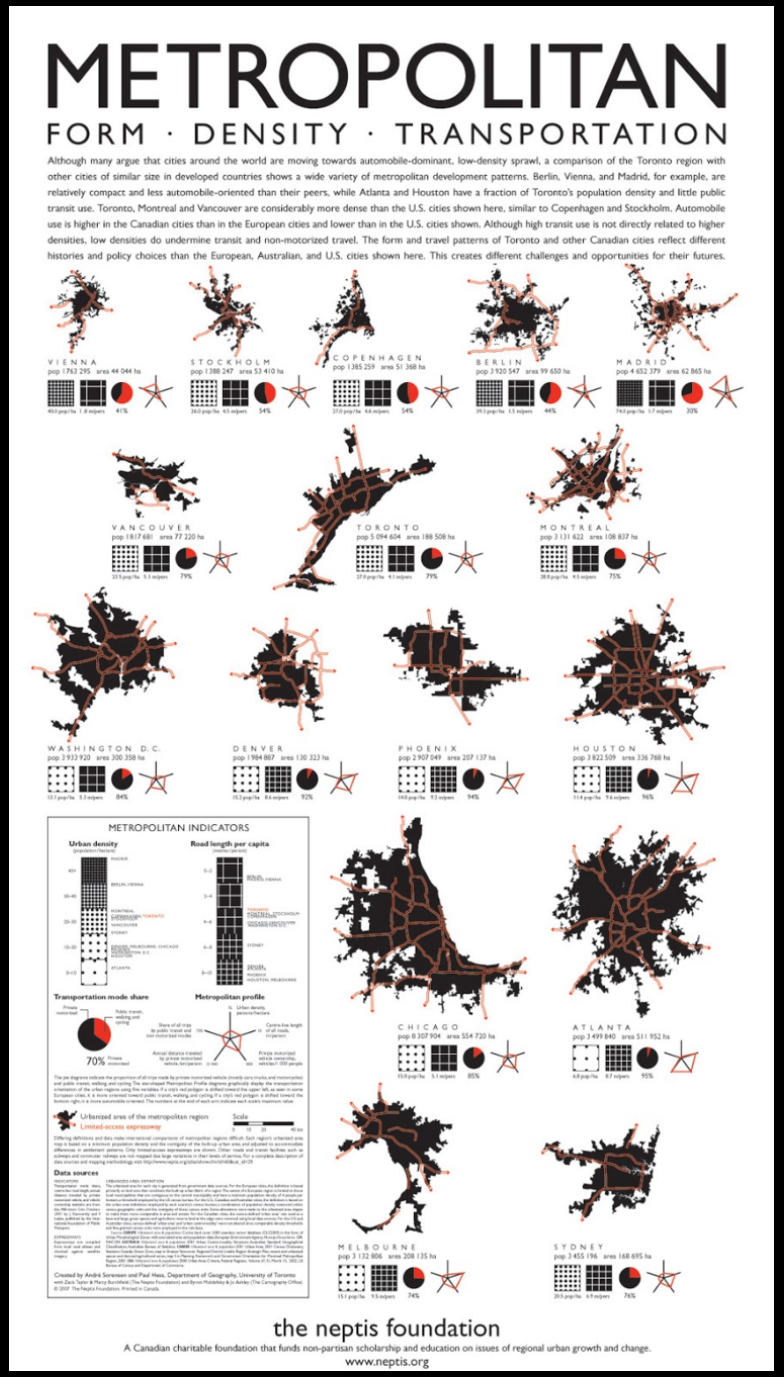


Tufte’s rules:

Small multiples

… missing the point.

https://vividmaps.com/comparing-metropolitan-form-density/



Tufte’s rules:

1. The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented ("lie factor")

effect size ~ 1

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

data/ink -> large

3. Show data variation, not design variation

no chart junk

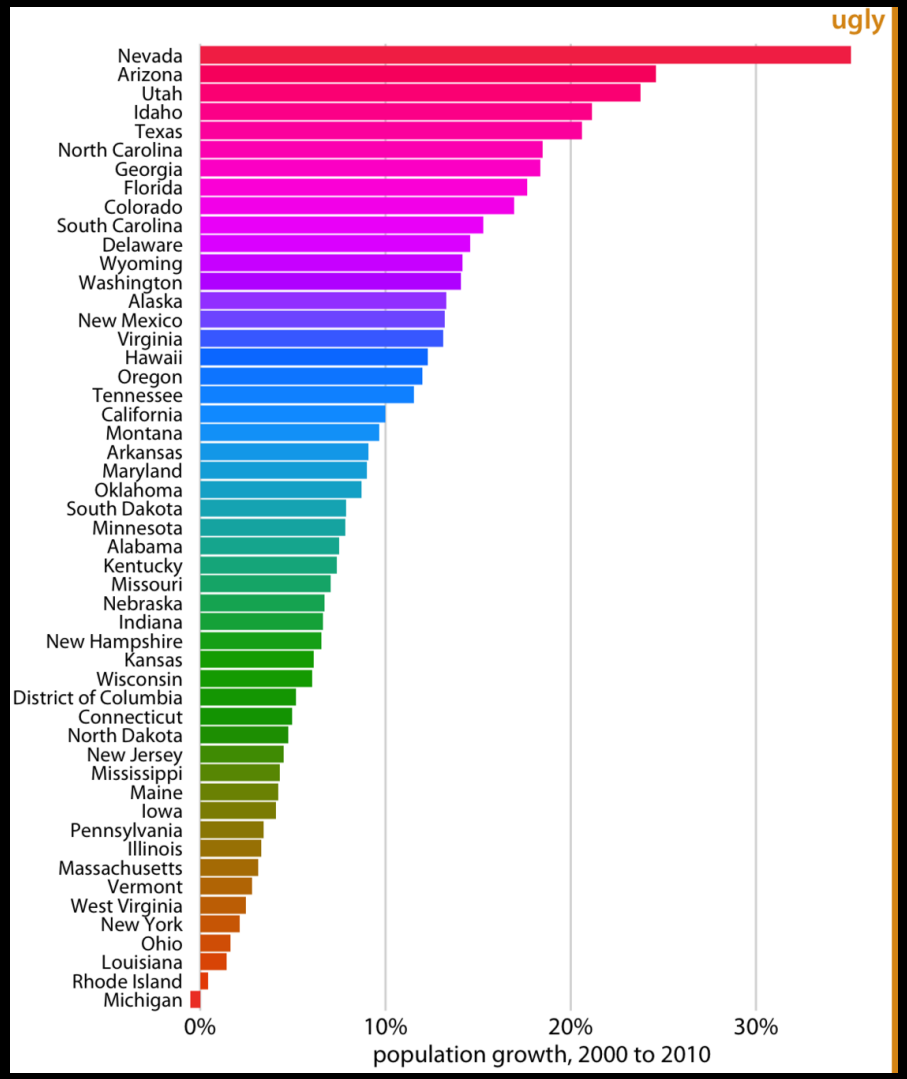
4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.

use small-multiples

Tufte's rules:

every feature should be associated with only 1 graphical element

(here color is redundant with length)



Tufte’s rules:

1. The representation of numbers, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented ("lie factor")

effect size ~ 1

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

data/ink -> large

3. Show data variation, not design variation

no chart junk

4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.

use small-multiples

5. The number of information carrying (variable) dimensions depicted should not exceed the number of dimensions in the data. Graphics must not quote data out of context.

avoid redundancy in communication

what makes a *good* visualization?

Jer Thorp

Using visualizations to understand the data, not to communicate a result

there definitely are historical precedents:

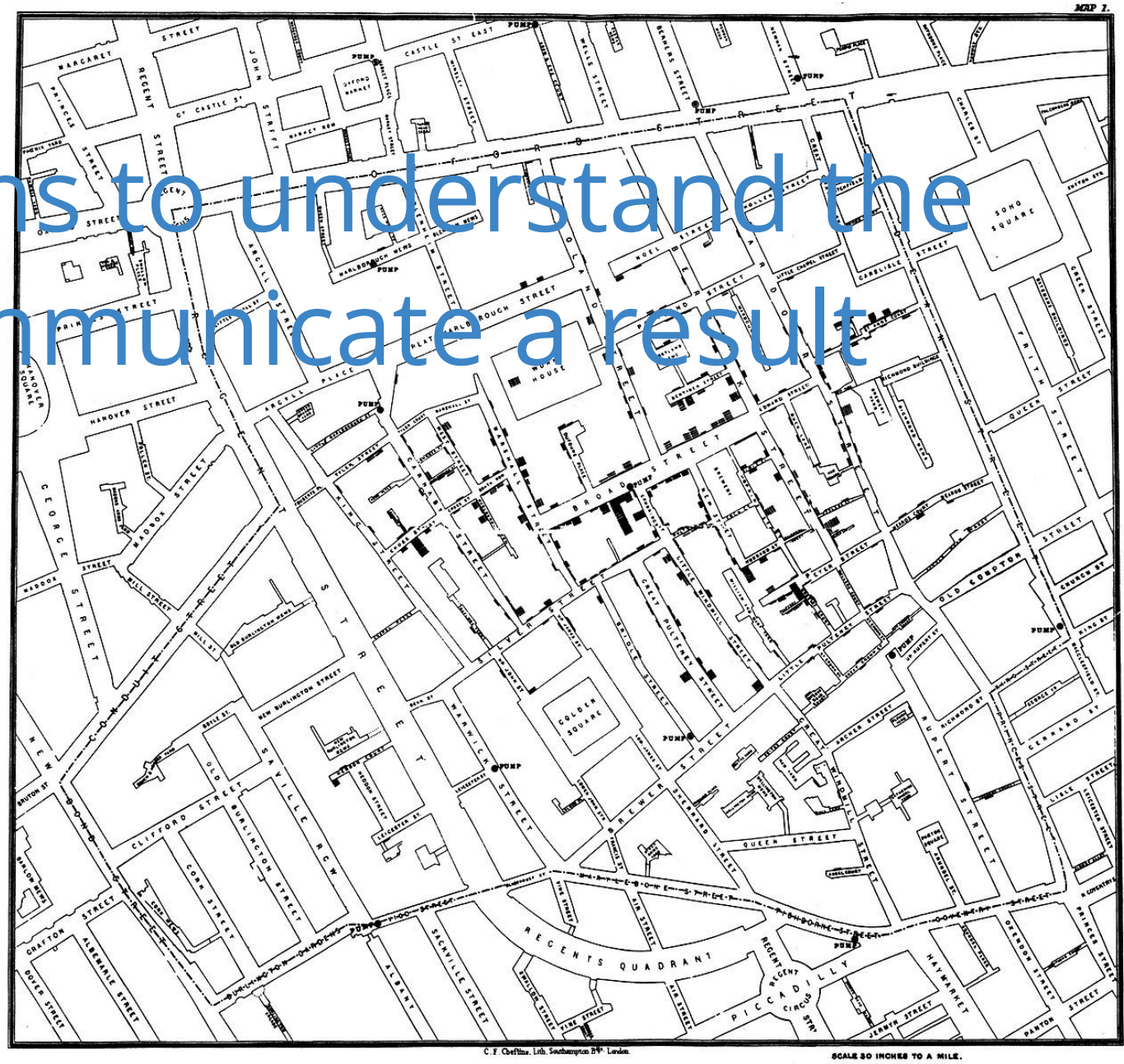
John Snow's map of cholera,

considered the first

"data science project"

uses "clustering" to drive causal inference

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak#cite_ref-FOOTNOTESnow1855[httpsarchiveorgstreamb28985266p38mode1up_38]_19-0



John Snow - Published by C.F. Cheffins, Lith, Southhampton Buildings, London, England, 1854

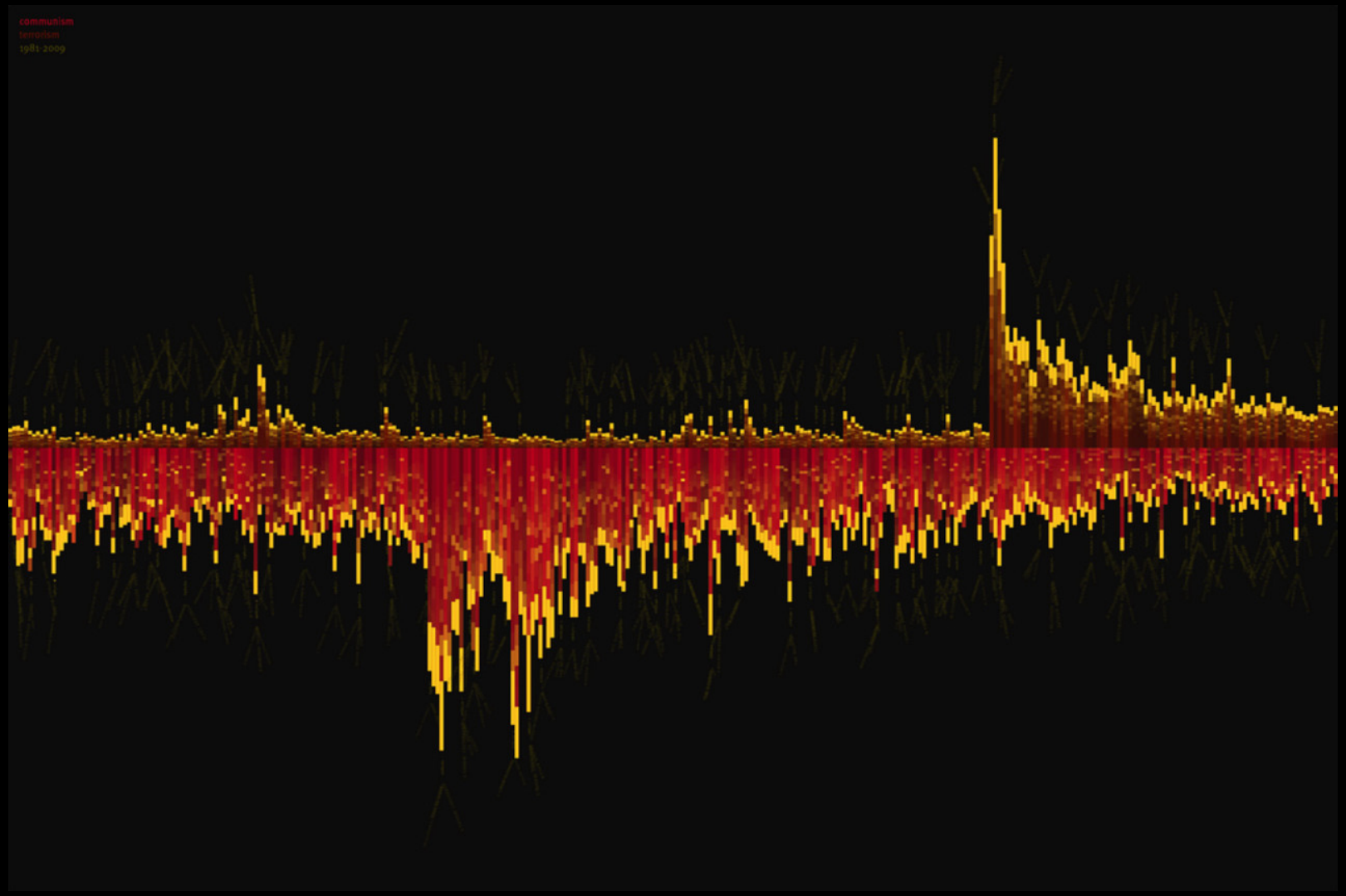
Using visualizations to understand the data, not to communicate a result

but only recently visualizations to aid science exploration became a well developed and active field of research

why the paradigm shift?

Jer Thorp

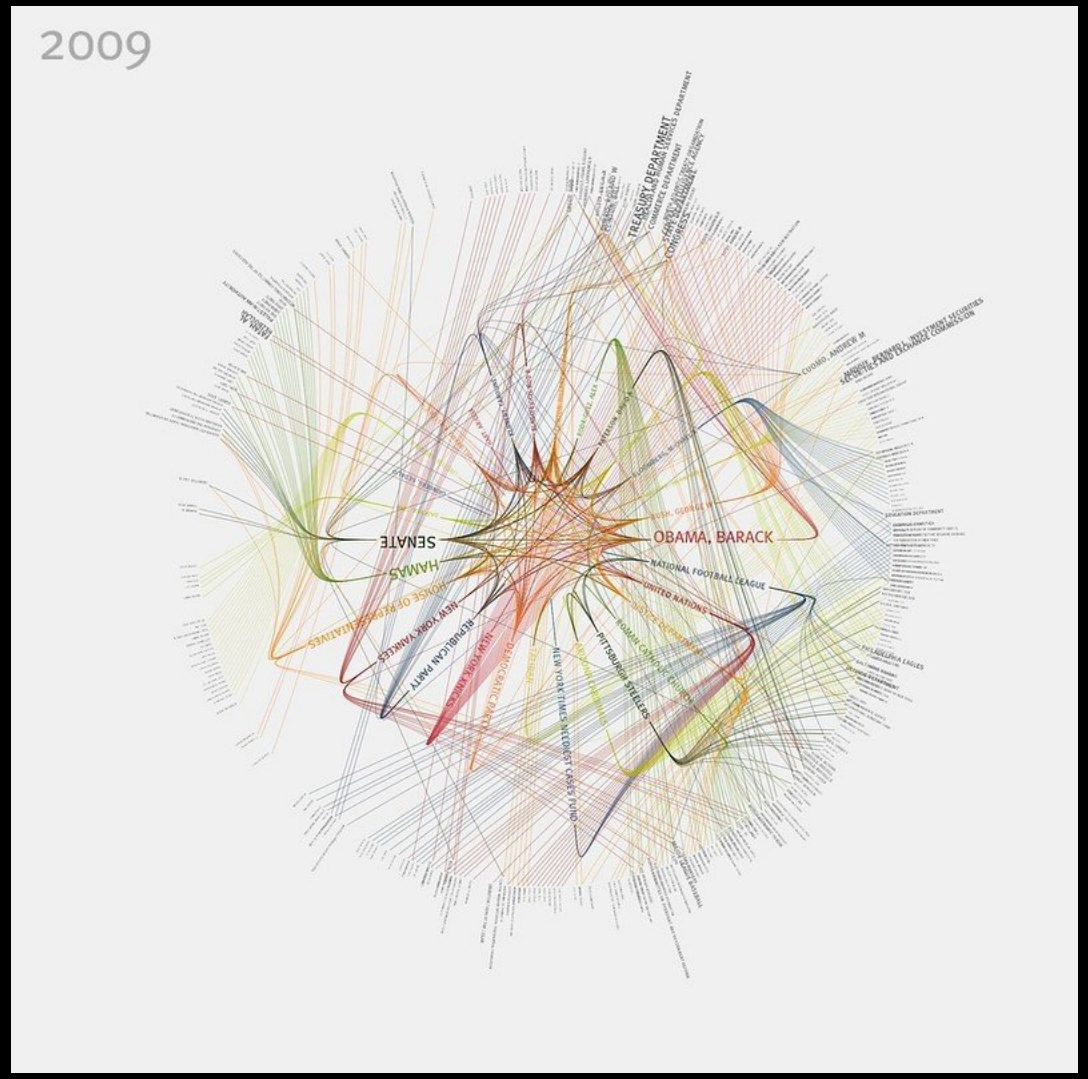
Big data:
increased data volume



One of Thorp's projects is a visualization of the number of times the terms "communism" (bottom) and "terrorism" (top) appeared in The New York Times, from 1981 until 2009. The spike for "terrorism" is the reflection of 9/11. As the word "terrorism" is used more and more, the use of the word "communism" decreases. (Image courtesy Jer Thorp; flickr.com/photos/blprnt/)

Big data:

increased data complexity



https://www.flickr.com/photos/blprnt/32912268016/in/album-72157614008027965/

These visualizations show the top organizations and personalities for every year from 1985 to 2001. Connections between these people & organizations are indicated by lines.

Data is from the newly-released NYTimes Article Search API: developer.nytimes.com

For more information, and source code to access the NYTimes API, visit my blog: blog.blprnt.com

https://player.vimeo.com/video/112183497?api=1

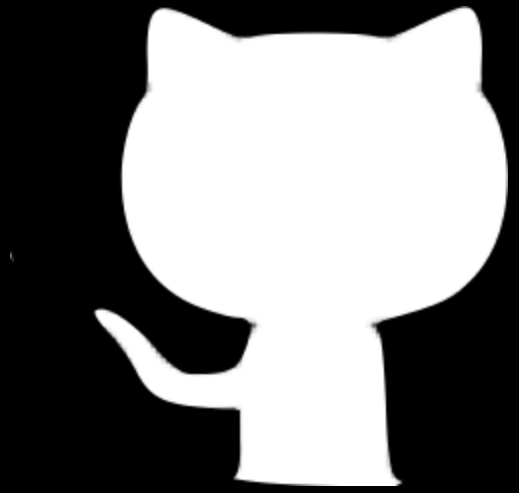
round 2 :

what is wrong with this plot????

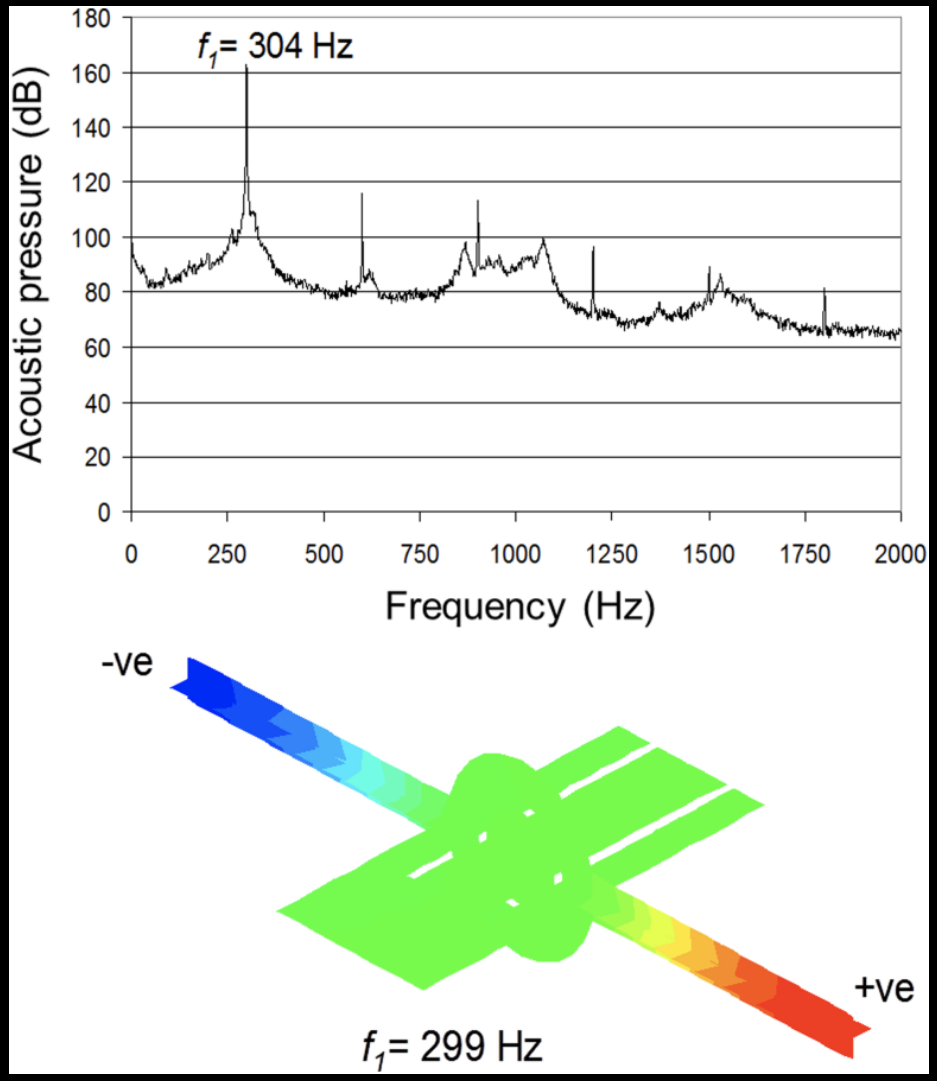
Tufte's rules version

in-class exercise

https://github.com/fedhere/PUS2020_FBianco/blob/master/classdemo/badplotgoodplot.ipynb



Tufte’s rules:

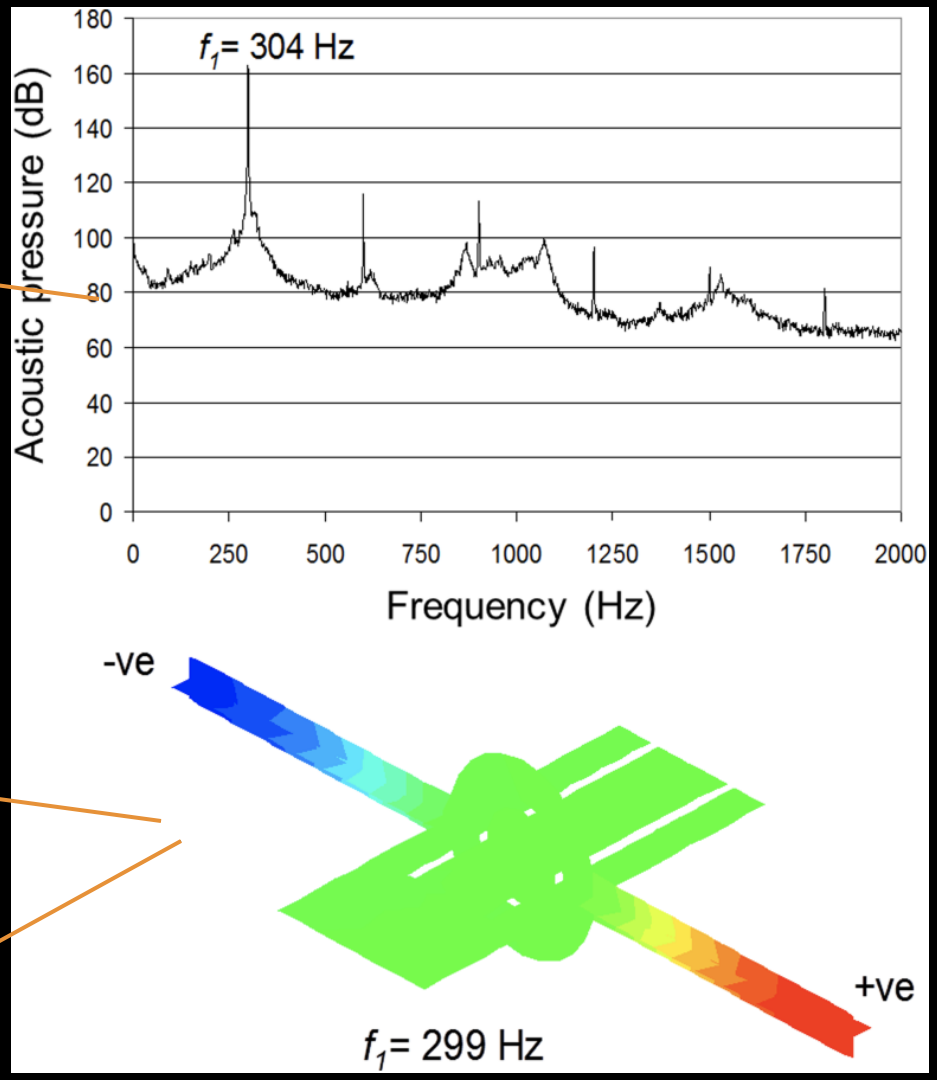


Tufte’s rules:

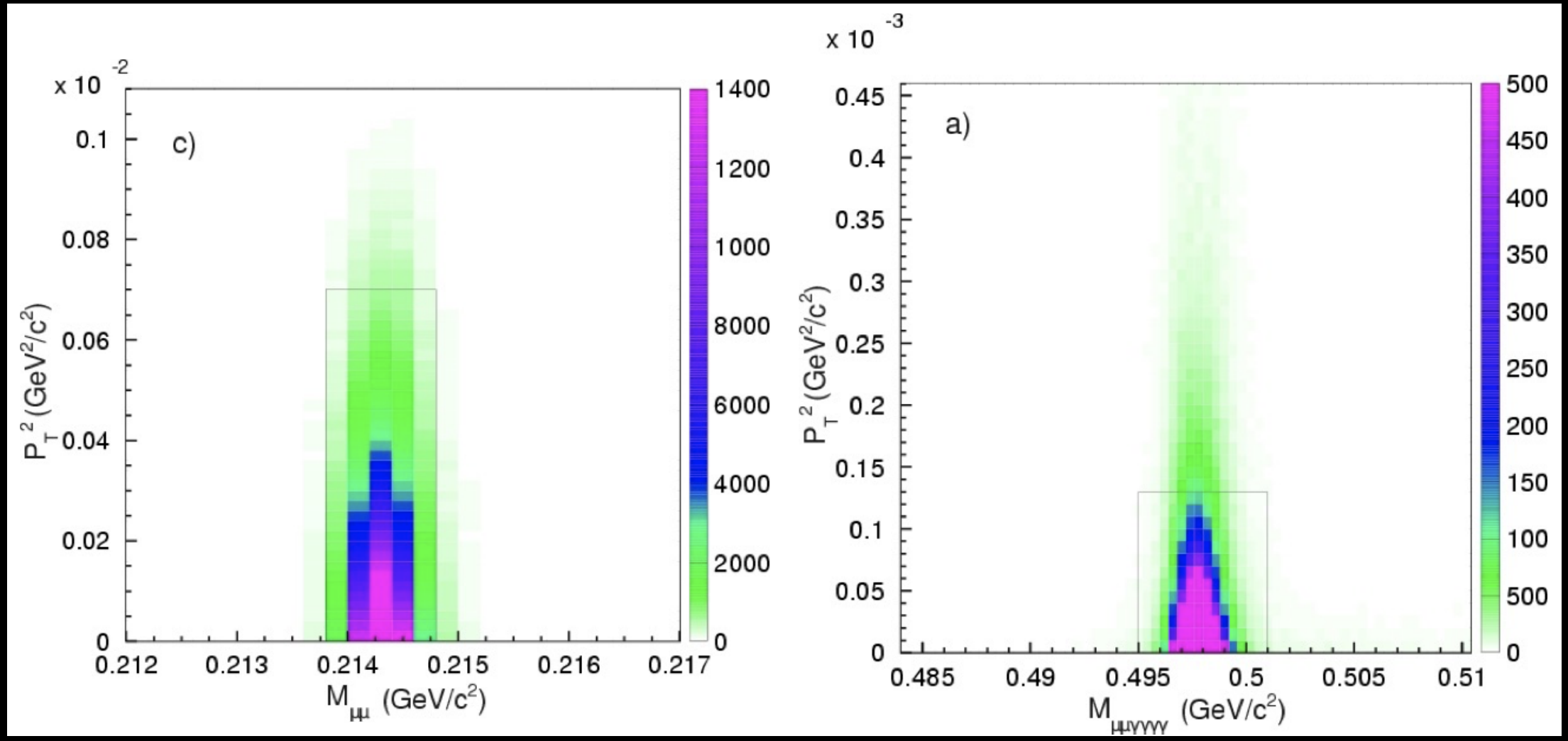
low data/ink ratio

chart junk

2 graphical elements for frequency
(color and position)

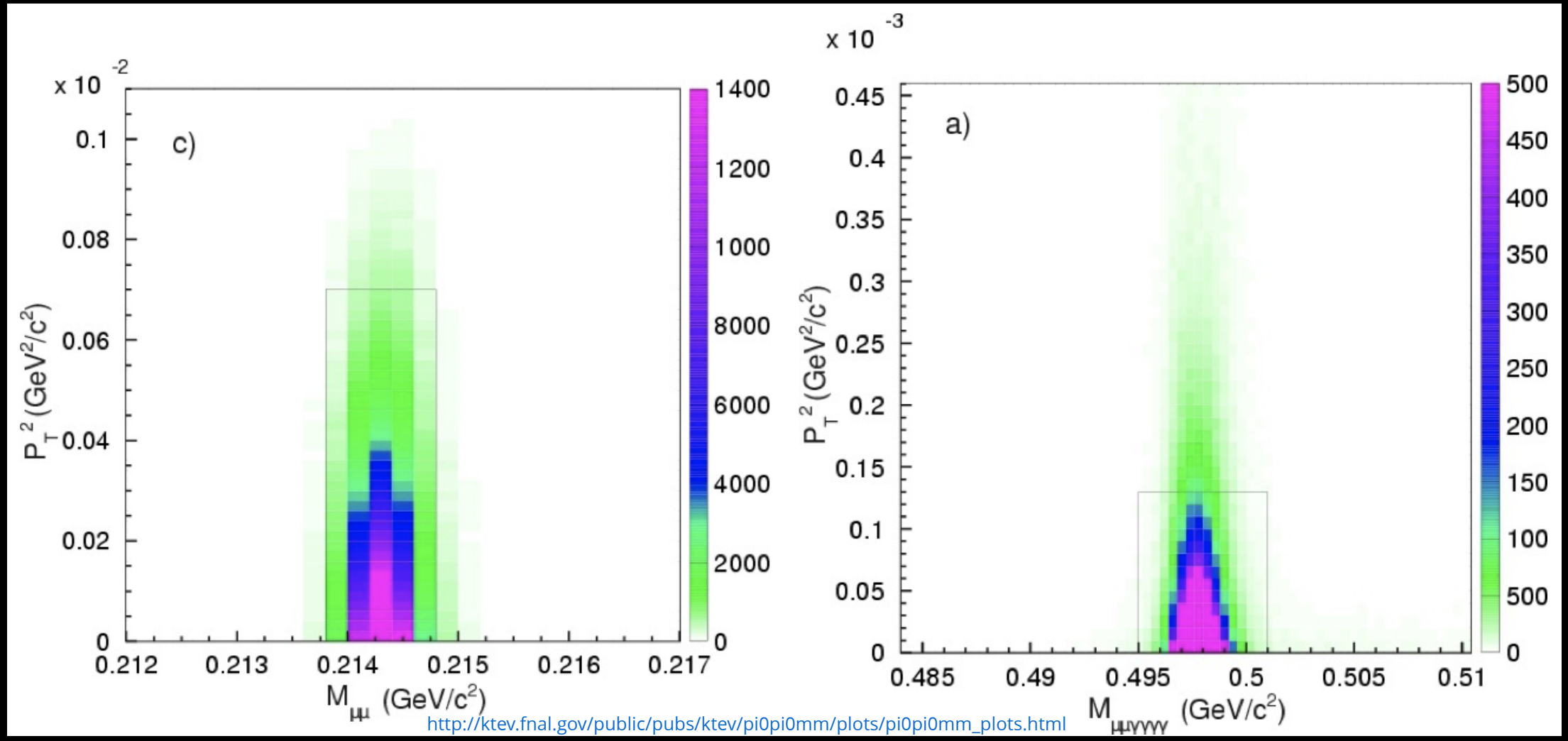


Tufte’s rules:



Tufte's rules:

low data/ink ratio
comparison but scale out of context
high effect-size due to the choice of color map (more on this later)



http://ktev.fnal.gov/public/pubs/ktev/pi0pi0mm/plots/pi0pi0mm_plots.html

Graphic Vocabulary

Graphic Vocabulary

What graphical elements are available and what elements are appropriate to convey certain information?

The ideal of all research is:

1. precise investigation of each individual phenomenon — in isolation,

2. the reciprocal effect of phenomena upon each other — in combinations,

3. general conclusions which are to be drawn from the above two divisions.

My objective in this book extends only to the first two parts. The material in this book does not suffice to cover the third part which, in any case, cannot be rushed.

The investigation should proceed in a meticulously exact and pedantically precise manner. Step by step, this "tedious" road must be traversed — not the smallest alteration in the nature, in the characteristics, in the effects

Point, Line, and Plane, Wassily Kandinsky, 1926

point line plane

position

size

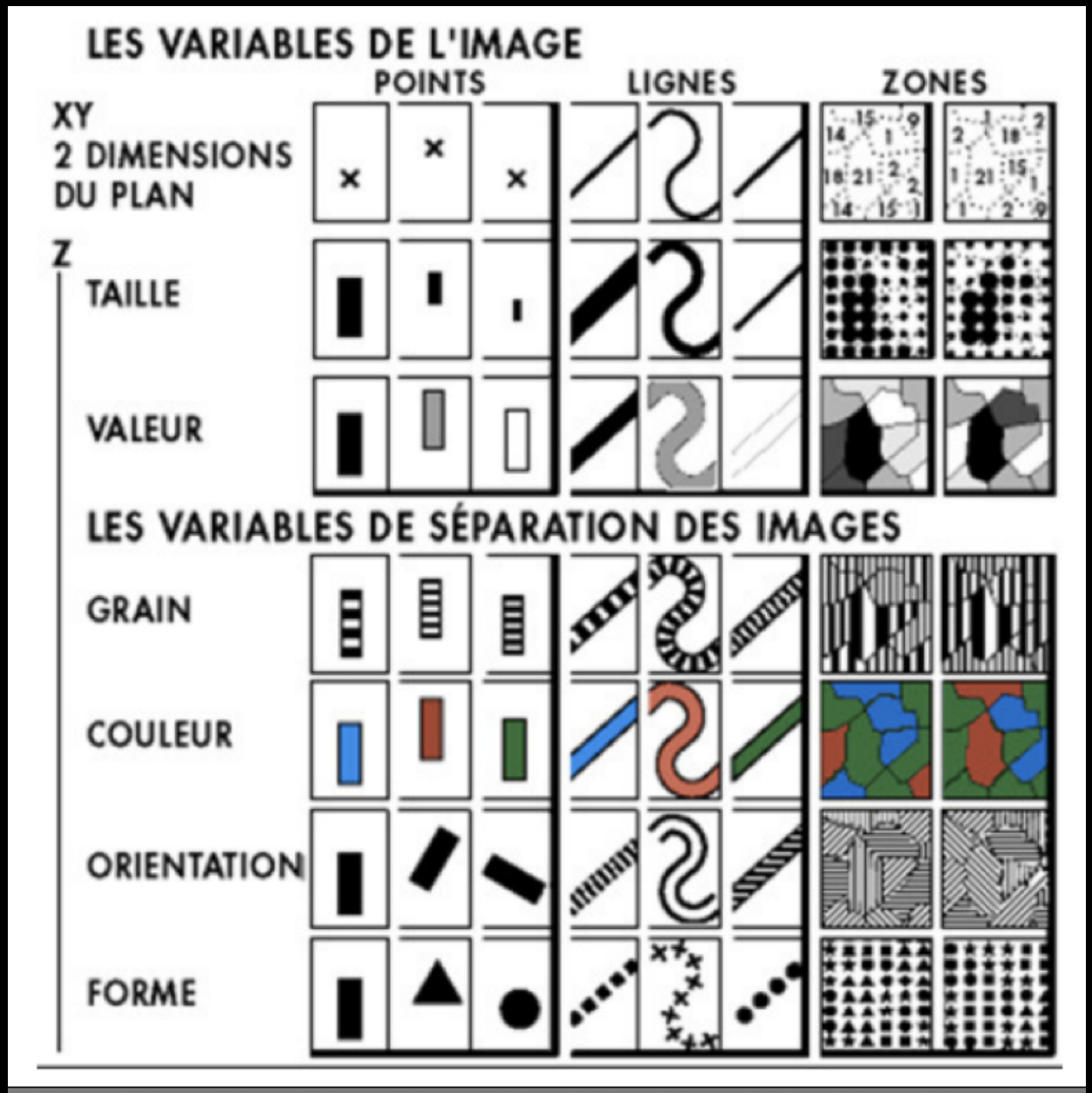
intensity

texture

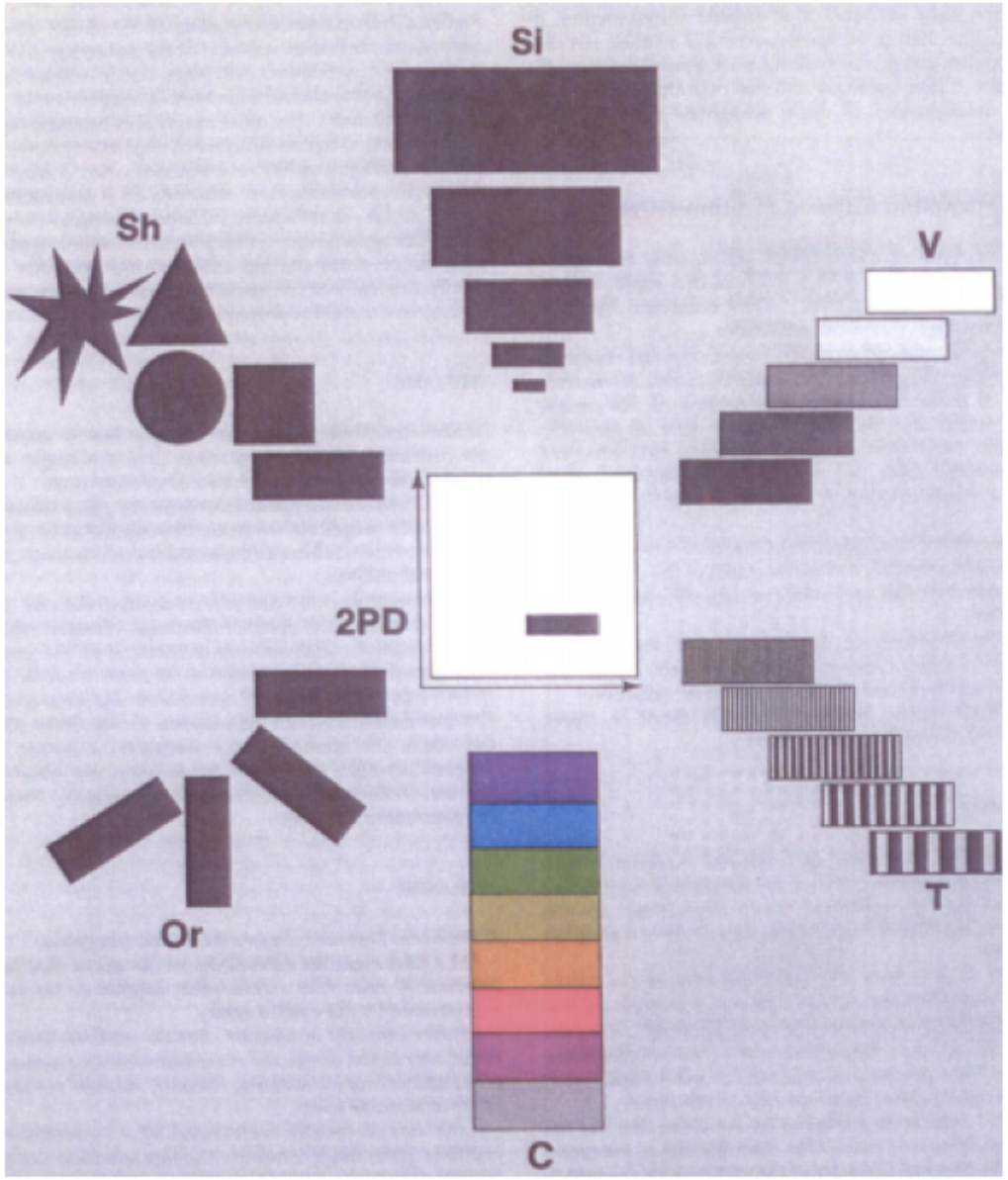
color

orientation

shape



Jacques Bertin: Semiology of Graphics, 1967 Gauthier-Villars, 1998 EHESS



- Size
- Value (Density)
- Texture
- Color
- Orientation
- Shape

- 3D
- Animation/Time

Bertin, Semiology of Graphics, 1967 Gauthier-Villars, 1983 UW Press

data types

graphical elements work differently on different data types

- **Continuous:** **distance to the closest supermarket**

Continuous data may be:

- **Continuous Ordinal:** Earthquakes (not-linear scale)
- **Interval:** F temperature - interval size preserved
- **Ratio**: Car speed - 0 is naturally defined

- **Discrete:** **any countable, e.g. number of people**

Discrete data may be:

- **Counts:** number of people with highschool degree
- **Ordinal:** survey response Good/Fair/Poor

- **Categorical:** **urban vs rural census tract**

Data may also be:

- **Censored:** age > 90
- **Missing:** “Prefer not to answer” (NA / NaN)

data types

graphical elements work differently on different data types

continuous

ordered

categorical