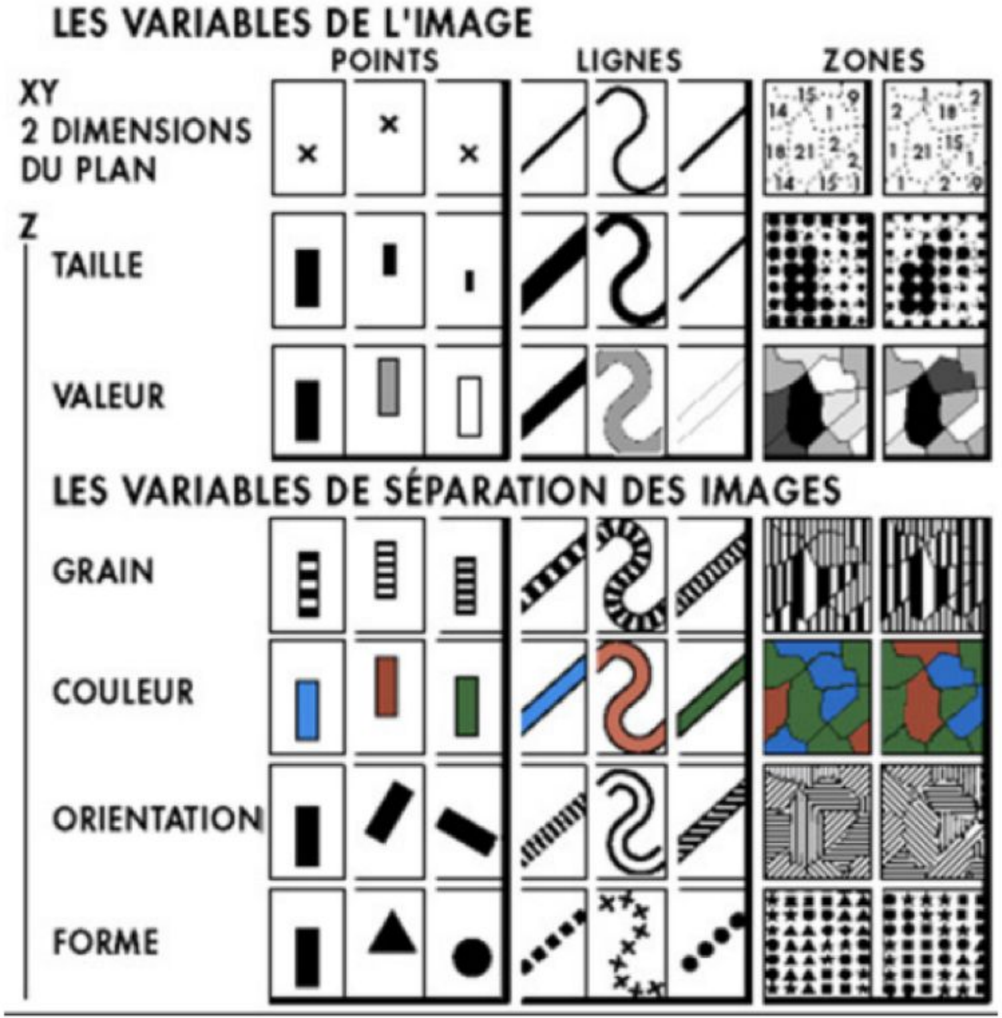
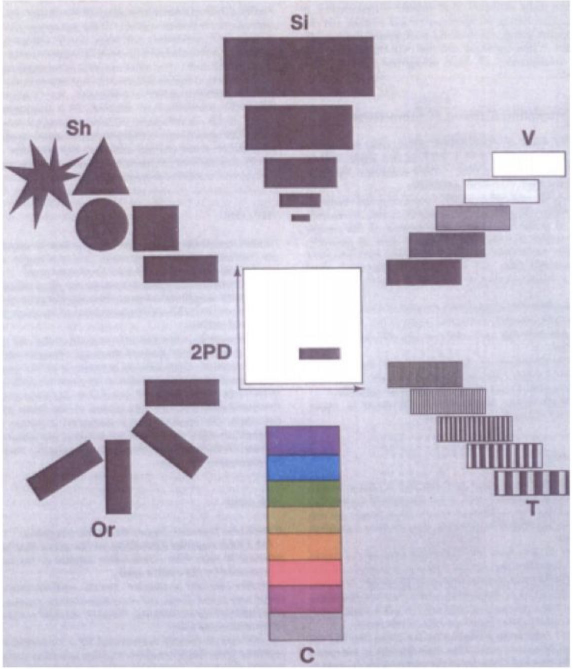
Graphic vocabulary



Graphical features



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

- 3D
- Animation/Time

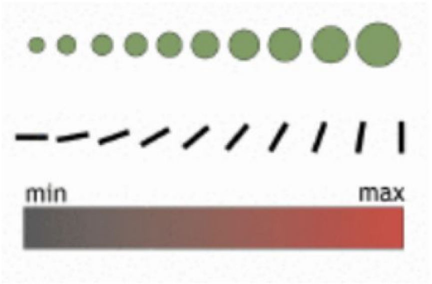
Bertin, Semiology of Graphics, 1967 Gauthier-Villars, 1983 UW Press

Data types

- **Continuous: distance to the closest star (can take any value)**
  - **Continuous Ordinal:** Earthquakes (notlinear scale)
  - **Interval:** F temperature - interval size preserved
  - **Ratio**: Car speed - 0 is naturally defined
- **Discrete: any countable, e.g. number of brain synapses**
  - **Counts:** number of bacteria at time *t* in section *A*
  - **Ordinal:** survey response Good/Fair/Poor
- **Categorical:** type of medication taken
- **Censored:** individual dead/alive
- **Missing:** “Prefer not to answer” (NA / NaN)

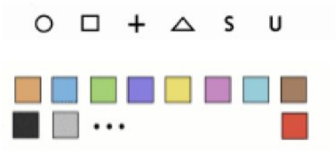
Data types

Continuous



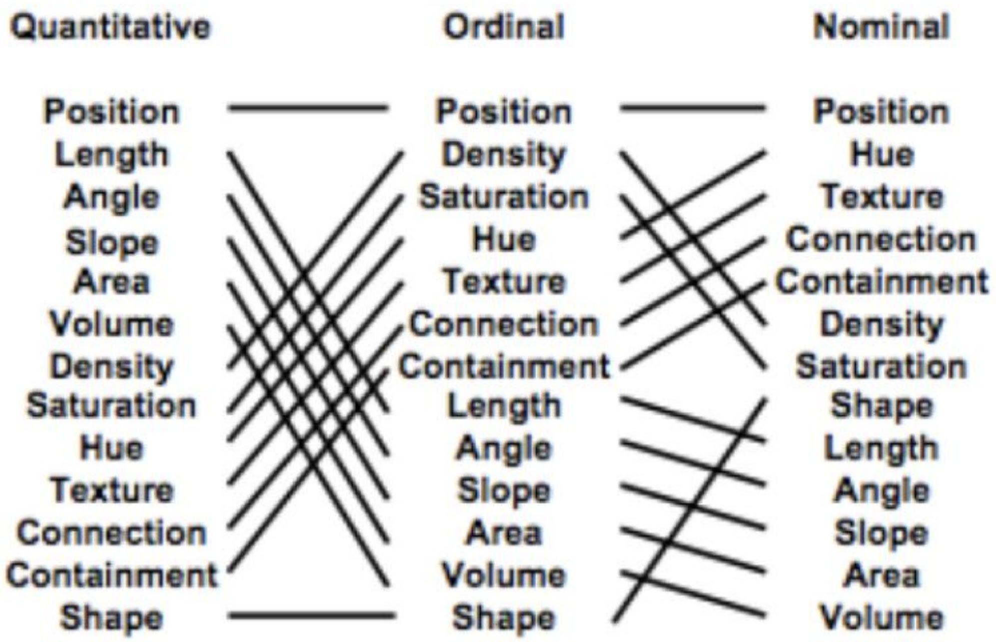
Ordered

Categorical



graphical elements work differently on different data types

Data types



[Mackinlay, Automating the Design of Graphical Presentations of Relational Information, ACM TOG 5:2, 1986]

Colors

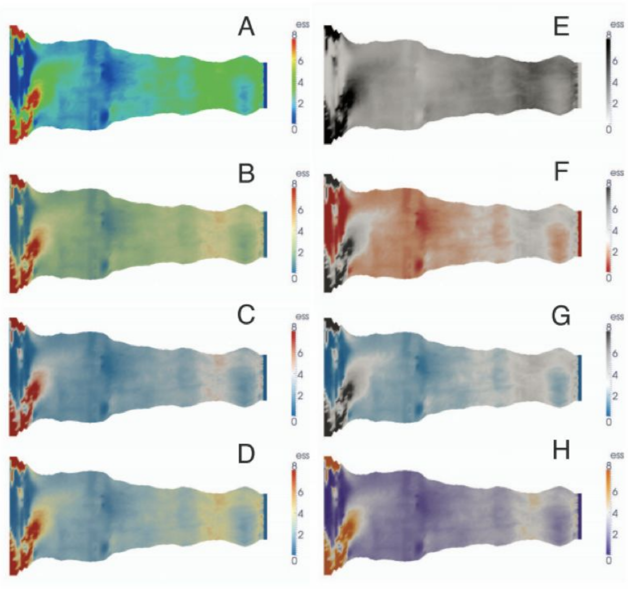


Fig. 4. Color schemes presented during the qualitative user study. The rainbow scheme (A) was preferred by most since it is what they are accustomed to viewing. The next most popular scheme was the red-black diverging scale (F). The grayscale image (E) was unanimously disliked since participants assume black-and-white images to be raw radiological data, while color indicates that the data has been processed or simulated.

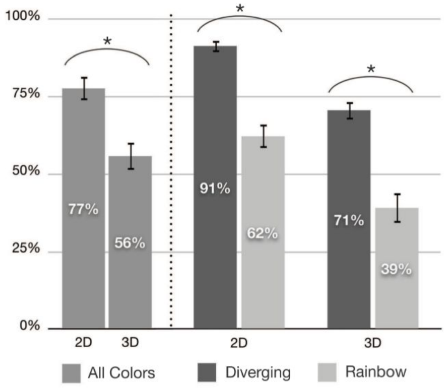


Fig. 7. **Average percent of low ESS regions identified** broken down by 2D and 3D representation, and color. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. Participants were more accurate in 2D and when using the diverging color map.

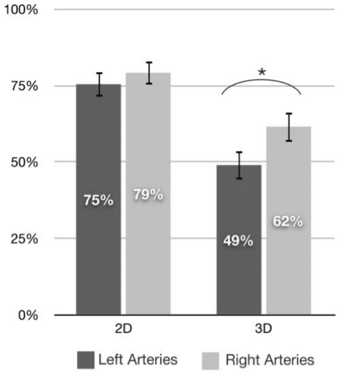


Fig. 8. **Average percent of low ESS regions identified** broken down by 2D and 3D representation, and left and right artery systems. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. In 3D, users were less accurate identifying regions in the most complex data sets (i.e., left artery systems). Whereas in 2D, performance was the same regardless of task complexity.

Evaluation of artery visualization for heart disease diagnostic - identification of endothelial shear stress (ESS)

http://www.eecs.harvard.edu/~kgajos/papers/2011/borkin11-infoviz.pdf

Colors

1) Never use rainbow

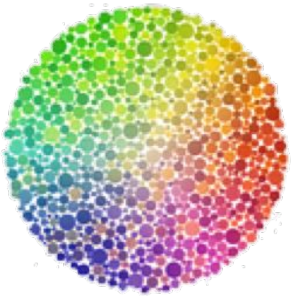
2) Use *diverging* color maps for data for which the center value is “special” (e.g., 0 with data ranging from positive to negative. In a diverging colormap the center of the range is white or black

3) Choose a *perceptually uniform* color map for continuous data that does not have a focal point (a special point inside the range)

4) Choose a sequential color map if your data range represents a progression (reflects some property of the data)

Color blindness

use the http://colororacle.org/ app to test your plots for color-blindness



http://www.colourblindawareness.org/colour-blindness/

Several tools to pick discrete colors/color palettes

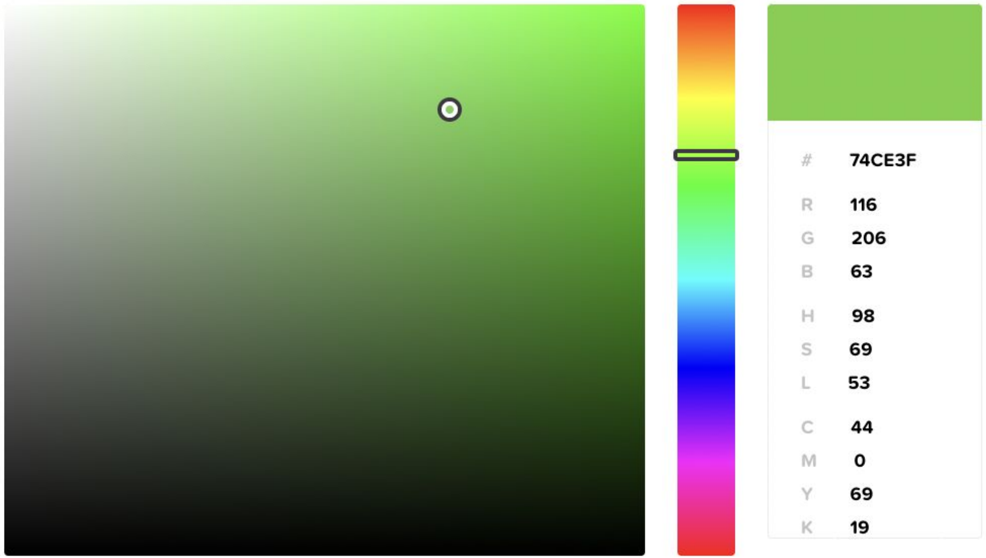
https://colorbrewer2.org/

https://www.color-hex.com/color-palettes/

https://htmlcolorcodes.com/

Finding colors: named colors, RGB and HEX

Finding colors: named colors, RGB and HEX



More examples of coded colors available

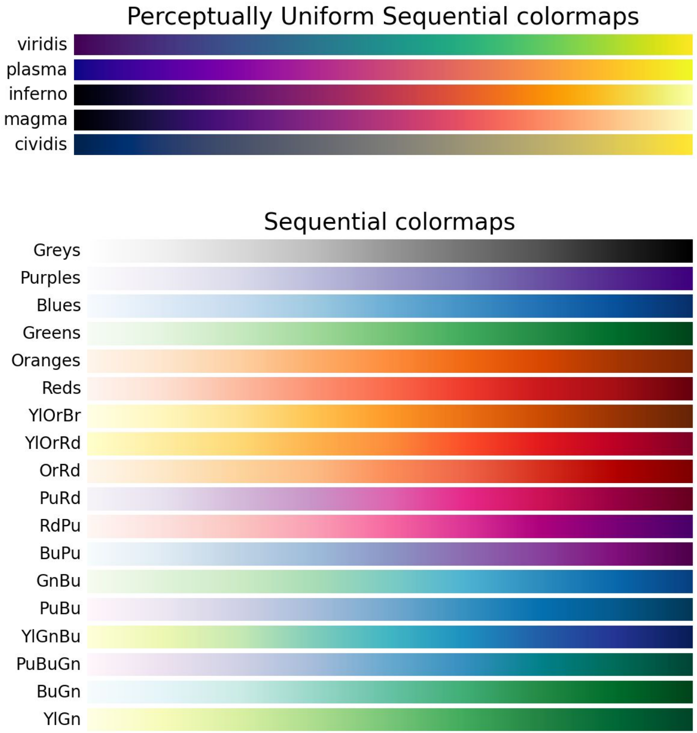
X11/CSS4	xkcd	
#00FFFF	#13EAC9	aqua
#7FFFD4	#04D8B2	aquamarine
#F0FFFF	#069AF3	azure
#F5F5DC	#E6DAA6	beige
#000000	**#000000**	**black**
#0000FF	#0343DF	blue
#A52A2A	#653700	brown
#7FFF00	#C1F80A	chartreuse
#D2691E	#3D1C02	chocolate
#FF7F50	#FC5A50	coral
#DC143C	#8C000F	crimson
#00FFFF	**#00FFFF**	**cyan**
#00008B	#030764	darkblue
#006400	#054907	darkgreen
#FF00FF	#ED0DD9	fuchsia
#FFD700	#DBB40C	gold
#DAA520	#FAC205	goldenrod

X11/CSS4	xkcd	
#008000	#15B01A	green
#808080	#929591	grey
#4B0082	#380282	indigo
#FFFFF0	#FFFFCB	ivory
#F0E68C	#AAA662	khaki
#E6E6FA	#C79FEF	lavender
#ADD8E6	#7BC8F6	lightblue
#90EE90	#76FF7B	lightgreen
#00FF00	#AAFF32	lime
#FF00FF	#C20078	magenta
#800000	#650021	maroon
#000080	#01153E	navy
#808000	#6E750E	olive
#FFA500	#F97306	orange
#FF4500	#FE420F	orangered
#DA70D6	#C875C4	orchid
#FFC0CB	#FF81C0	pink

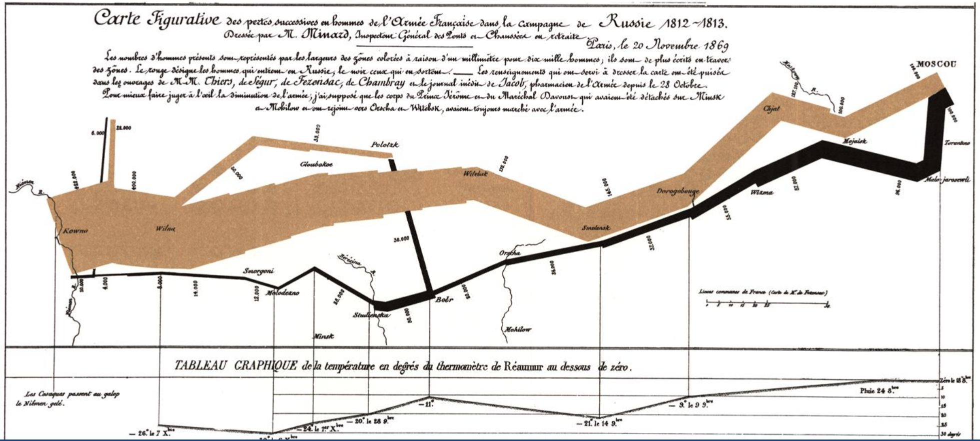
X11/CSS4	xkcd	
#DDA0DD	#580F41	plum
#800080	#7E1E9C	purple
#FF0000	#E50000	red
#FA8072	#FF796C	salmon
#A0522D	#A9561E	sienna
#C0C0C0	#C5C9C7	silver
#D2B48C	#D1B26F	tan
#008080	#029386	teal
#FF6347	#EF4026	tomato
#40E0D0	#06C2AC	turquoise
#EE82EE	#9A0EEA	violet
#F5DEB3	#FBDD7E	wheat
#FFFFFF	**#FFFFFF**	**white**
#FFFF00	#FFFF14	yellow
#9ACD32	#BBF90F	yellowgreen

https://matplotlib.org/stable/tutorials/colors/colors.html

Color palettes

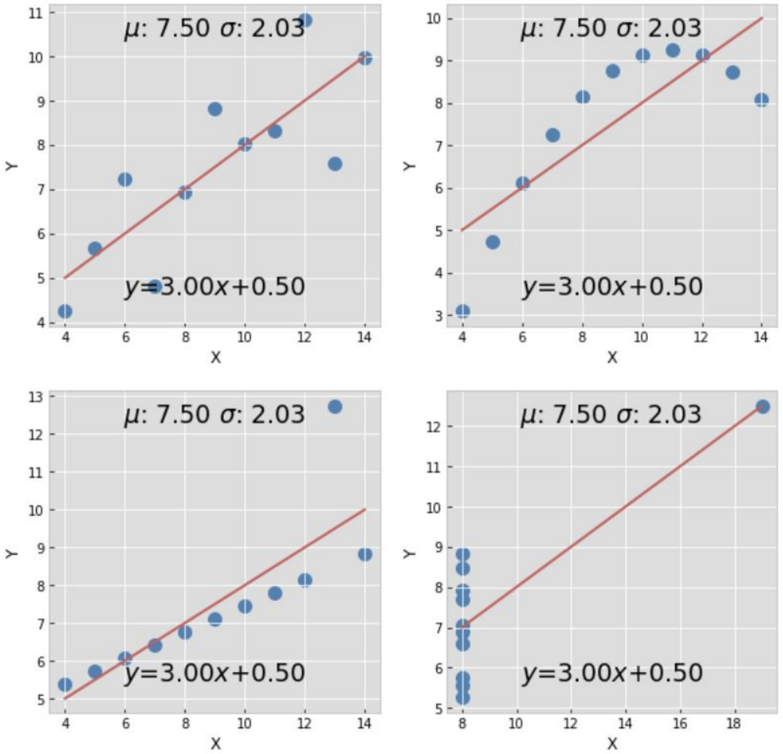


https://matplotlib.org/stable/tutorials/colors/colormaps.html



Data Visualization

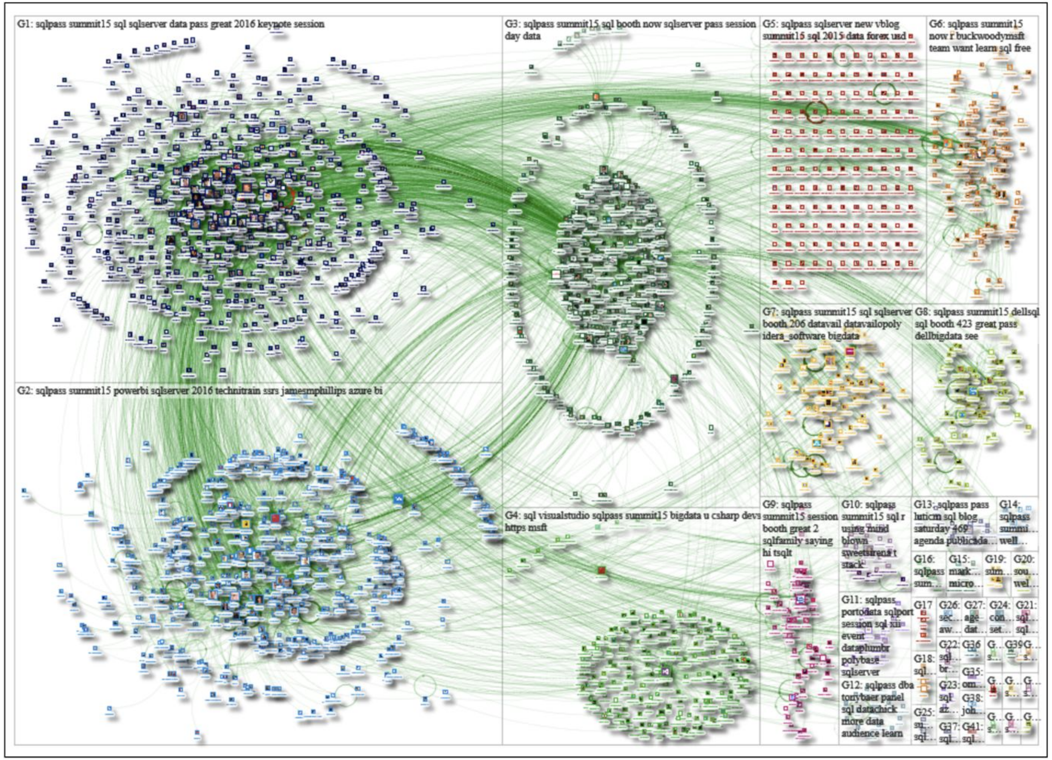
The importance of visualization



Anscombe's quartet

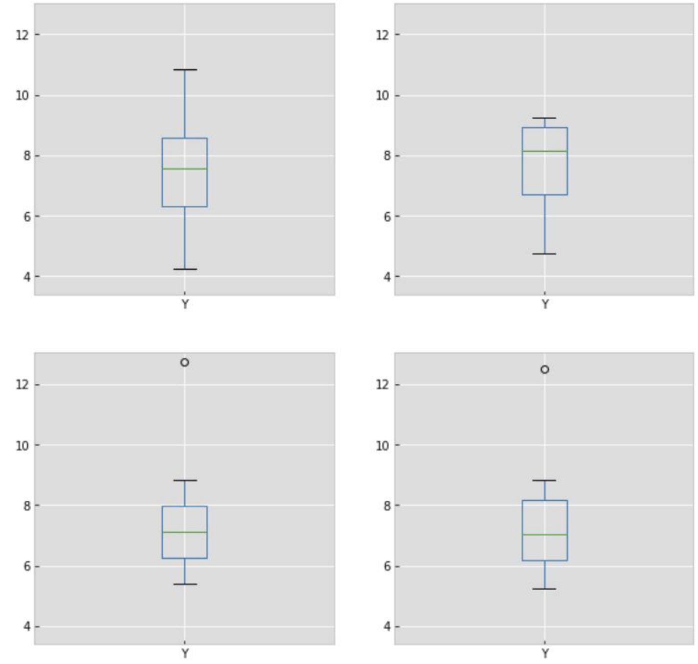
(Francis Anscombe, 1973) comprises four datasets that have **nearly identical simple descriptive statistics**, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.

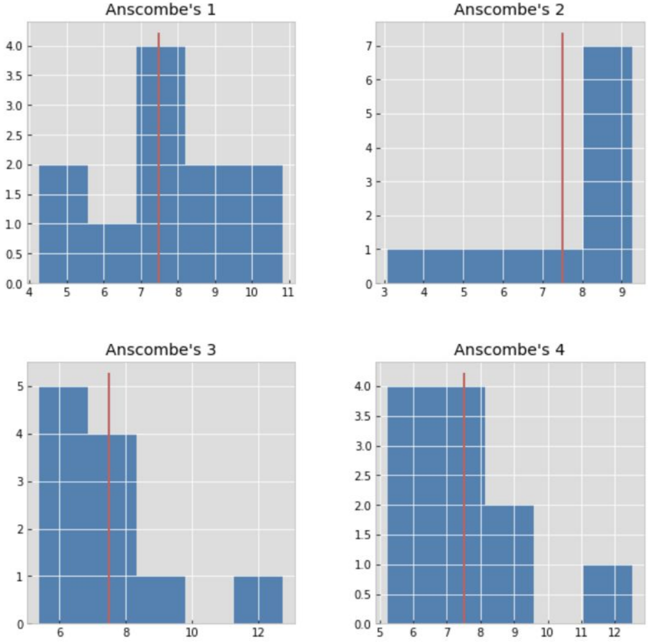
The problem with big data



the larger the data, and especially the higher the number of dimensions, the harder to design a visualization that is effective

Visualization of derived data products may be effective

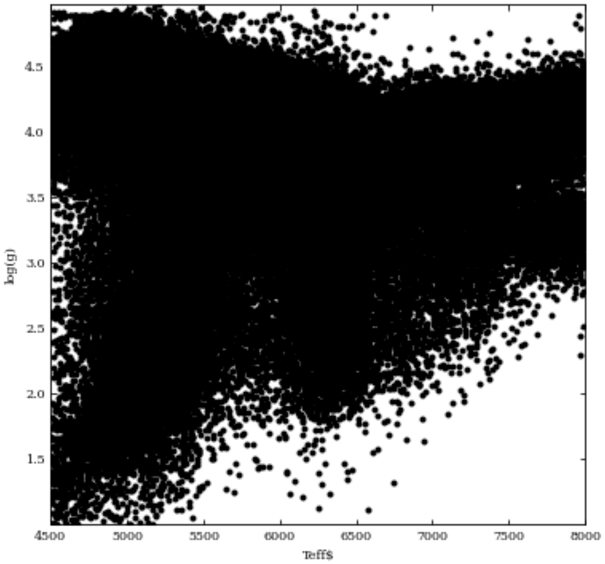




… the larger the data the harder it is to collectively understand the descriptive statistics

A common problem: too many points

```
plt.plot(Teff, logg, 'k.')
```

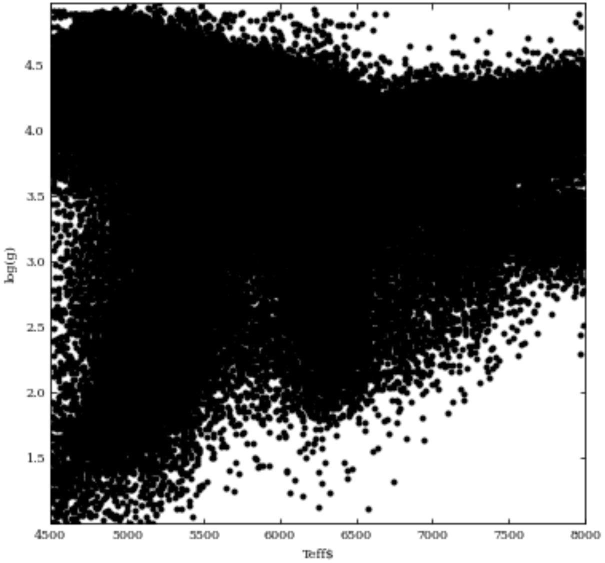


A common problem: too many points

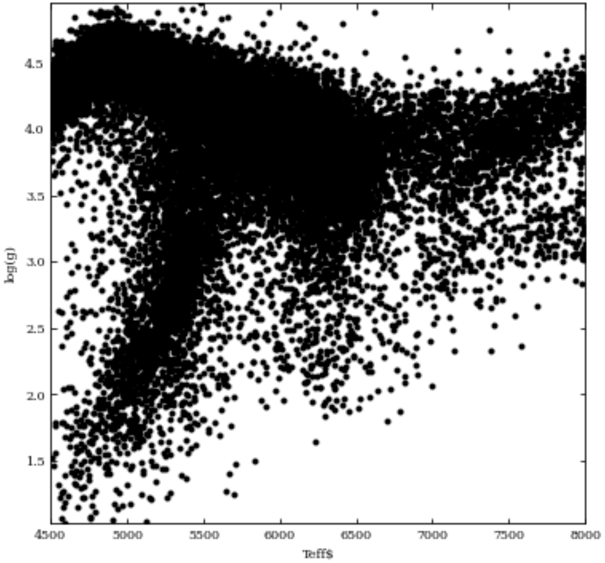
solution: subsample

solution: alpha

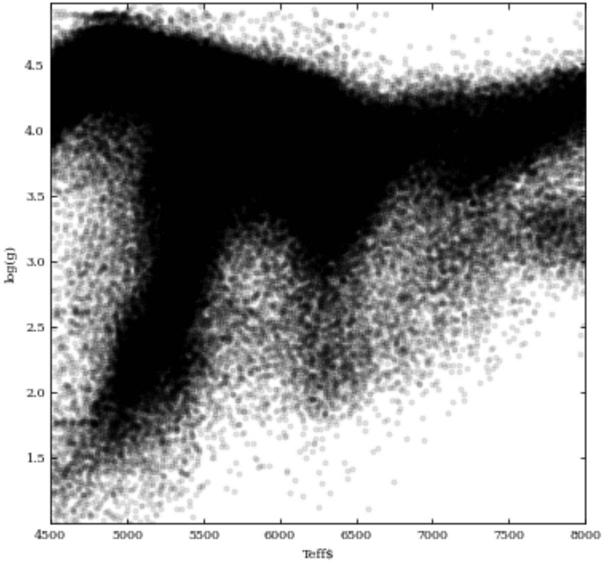
```
plt.plot(Teff, logg, 'k.')
```




```
plt.plot(Teff[::10], logg[::10], 'k.')
```

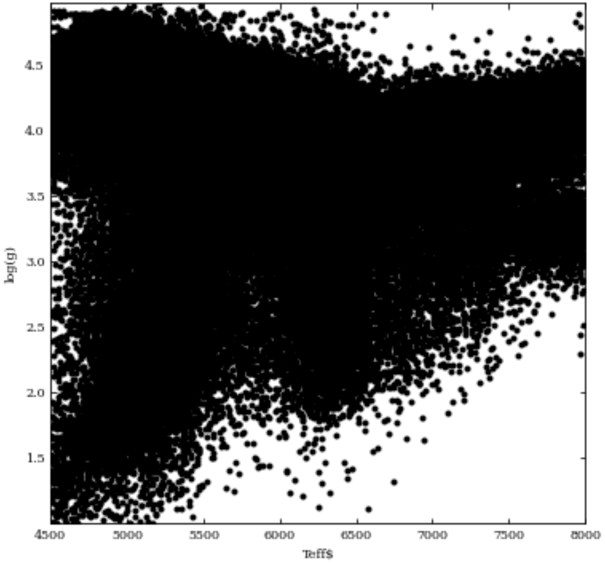



```
plt.plot(Teff, logg, 'k.', alpha=0.1)
```



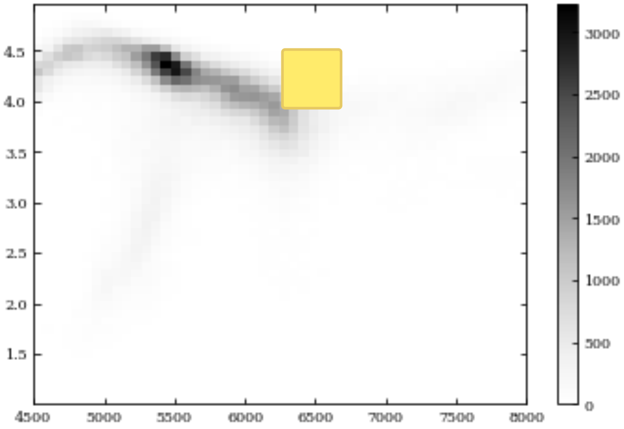
A common problem: too many points

```
plt.plot(Teff, logg, 'k.')
```



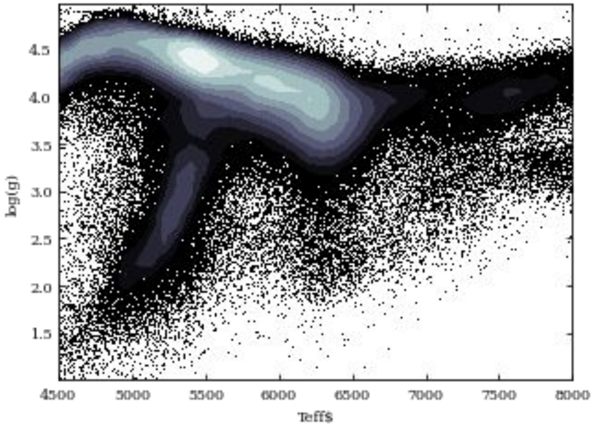
solution: density histograms

```
plt.hist2d(Teff, logg, bins=(50, 50), cmap=plt.cm.Greys)
```



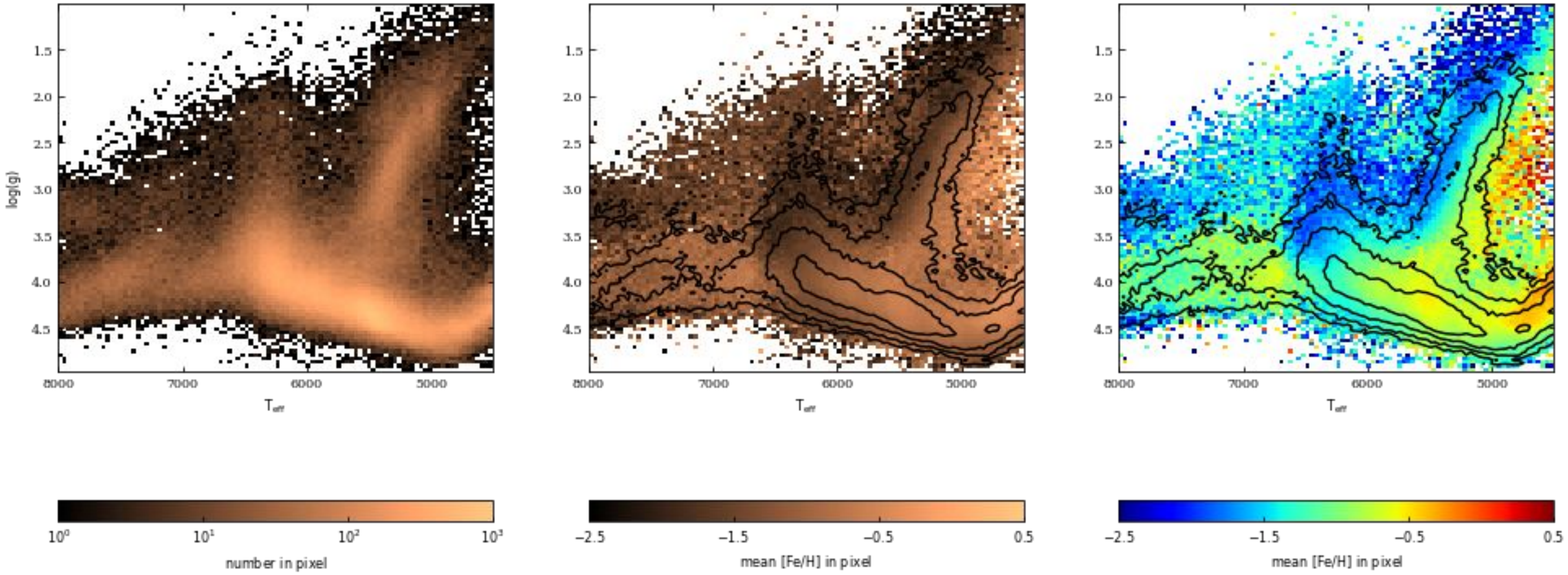
solution: scatter contours

```
plt.plot(Teff, logg, 'k.', alpha=0.1)
```



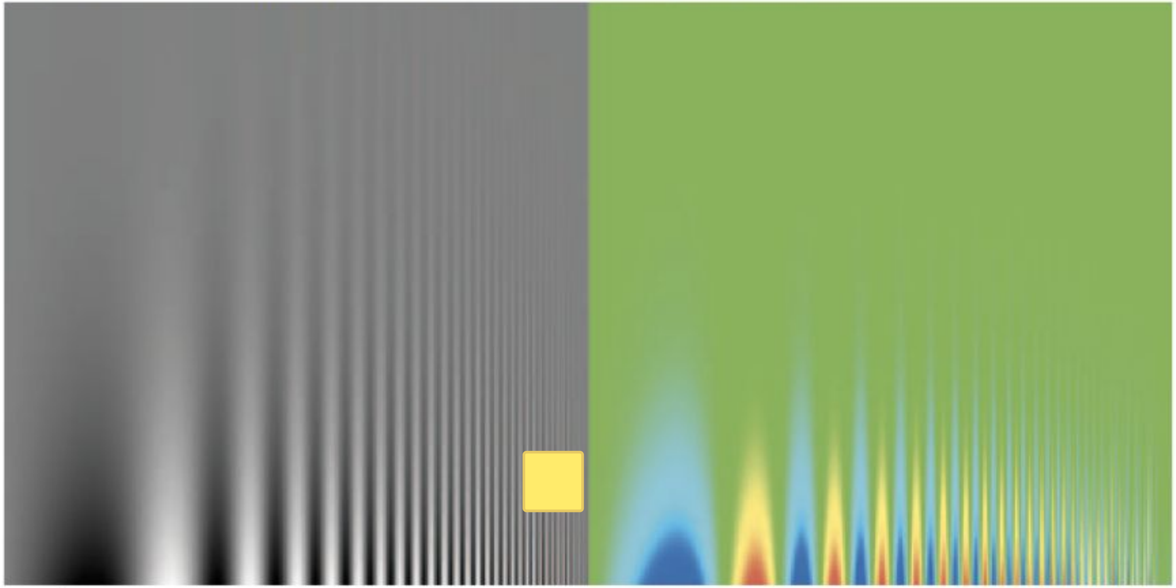
A common problem: too many points

Bad color choice!!!



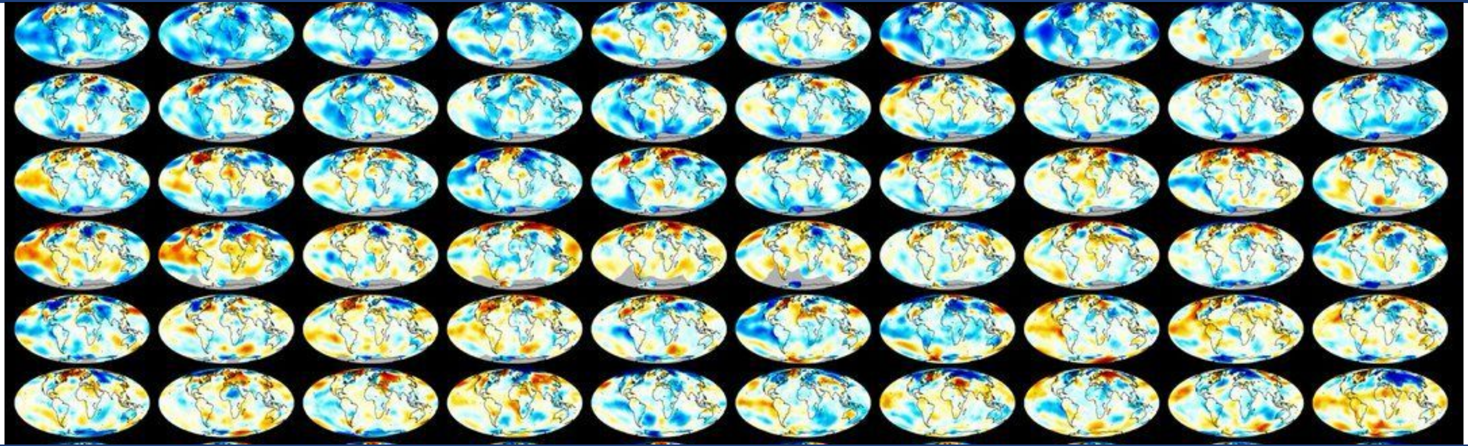
A common problem: too many points

5. **Detail is actually harder to see in a rainbow.**
The logic that it is easier to see detail in a range when you add colors seems to make sense, but in reality, more detail can be seen in a single hue image with a high brightness range.



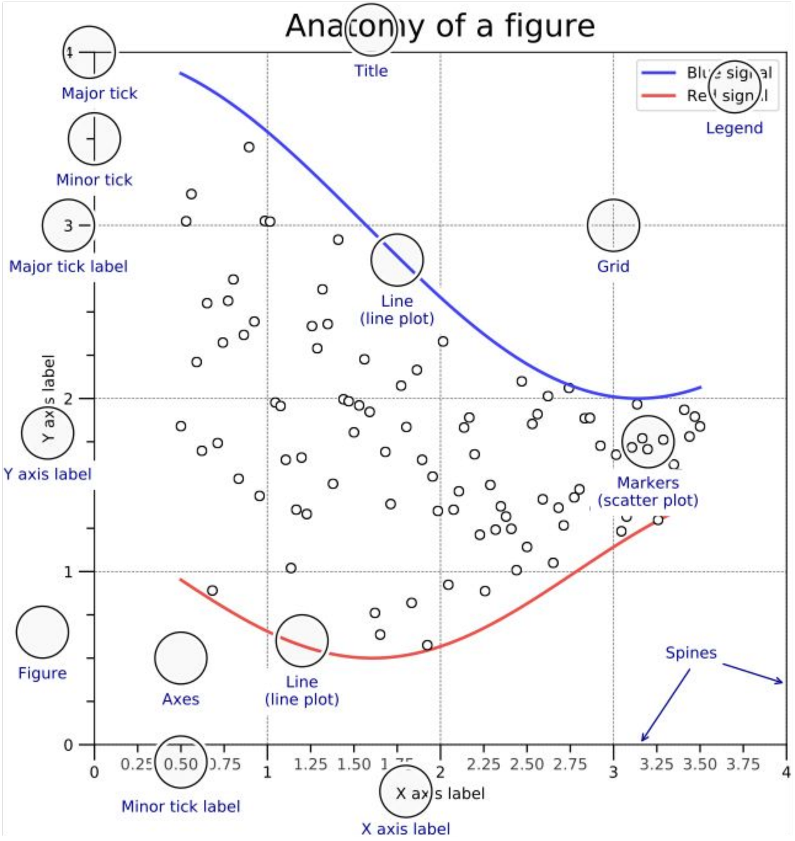
(source)

https://sciencenode.org/visualization/dear-nasa-no-more-rainbows-please.php



What makes a good plot

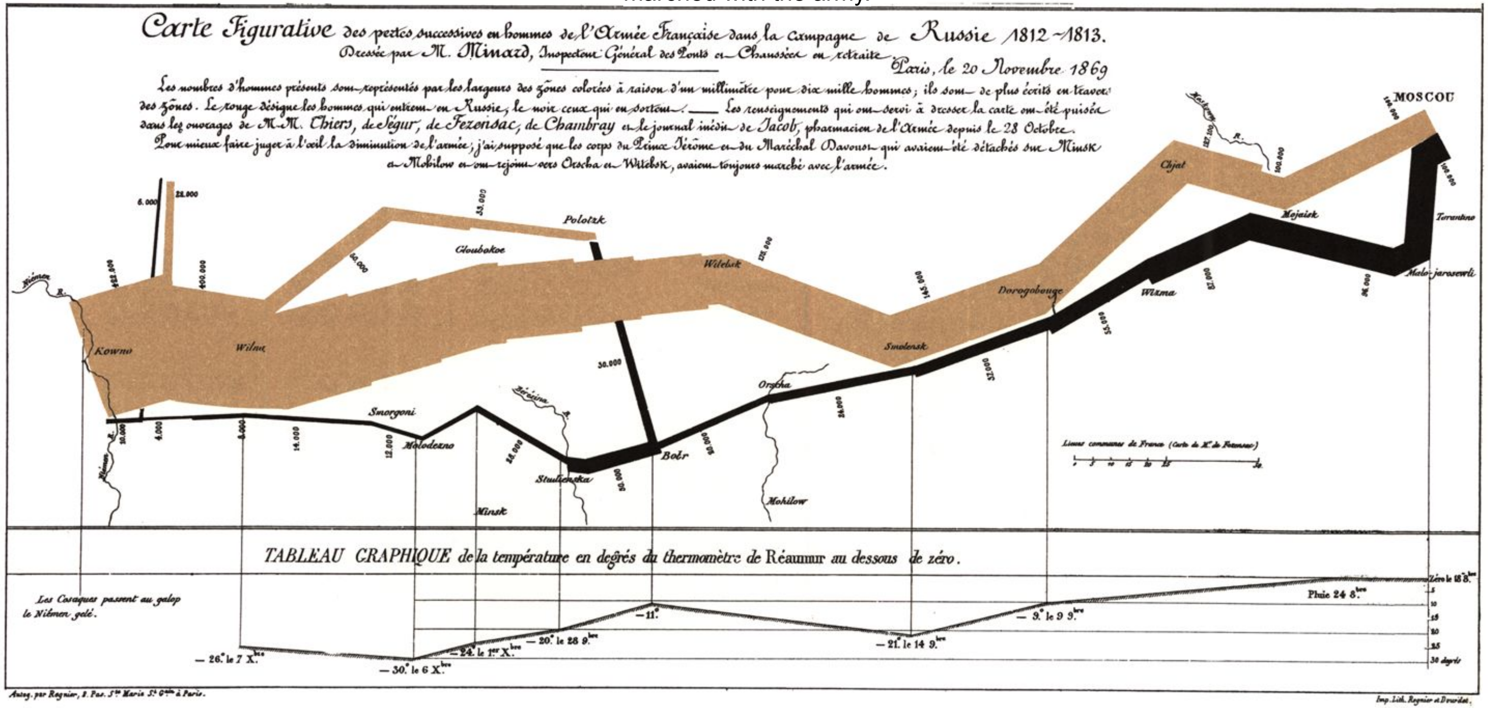
Anatomy of a figure



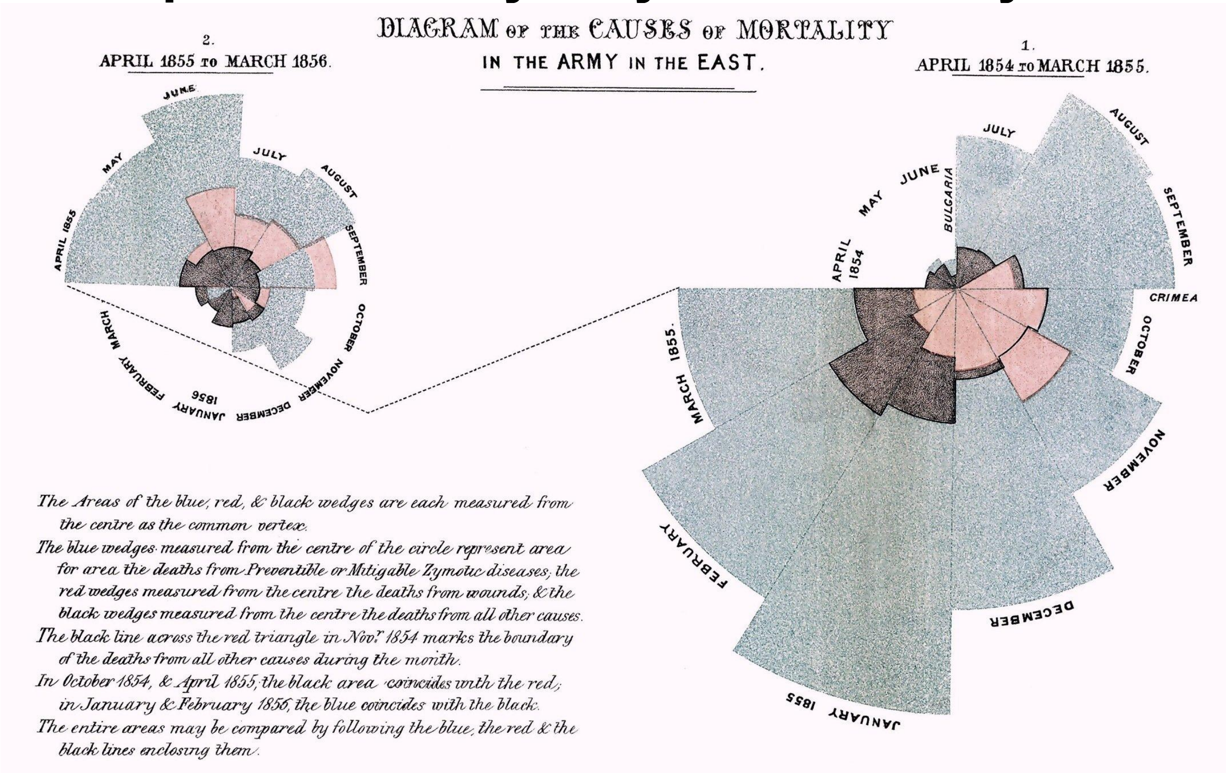
A few historical plots and why they made history

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.

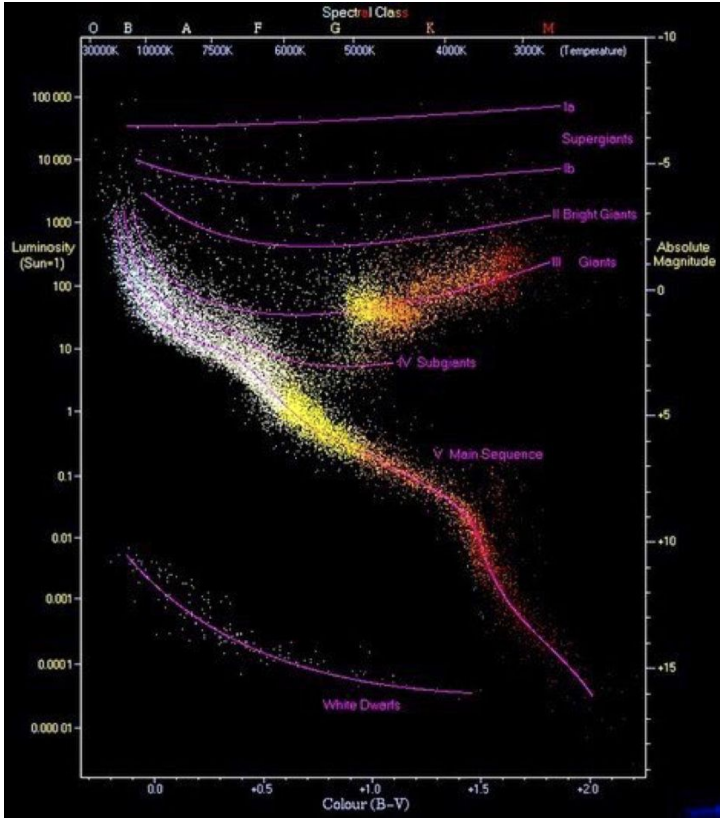


A few historical plots and why they made history



Florence Nightingale Coxcombs : http://timelyportfolio.github.io/rCharts_micropolar/nightingale/index.html

A few historical plots and why they made history



H-R diagram: the life of a star (~1910)

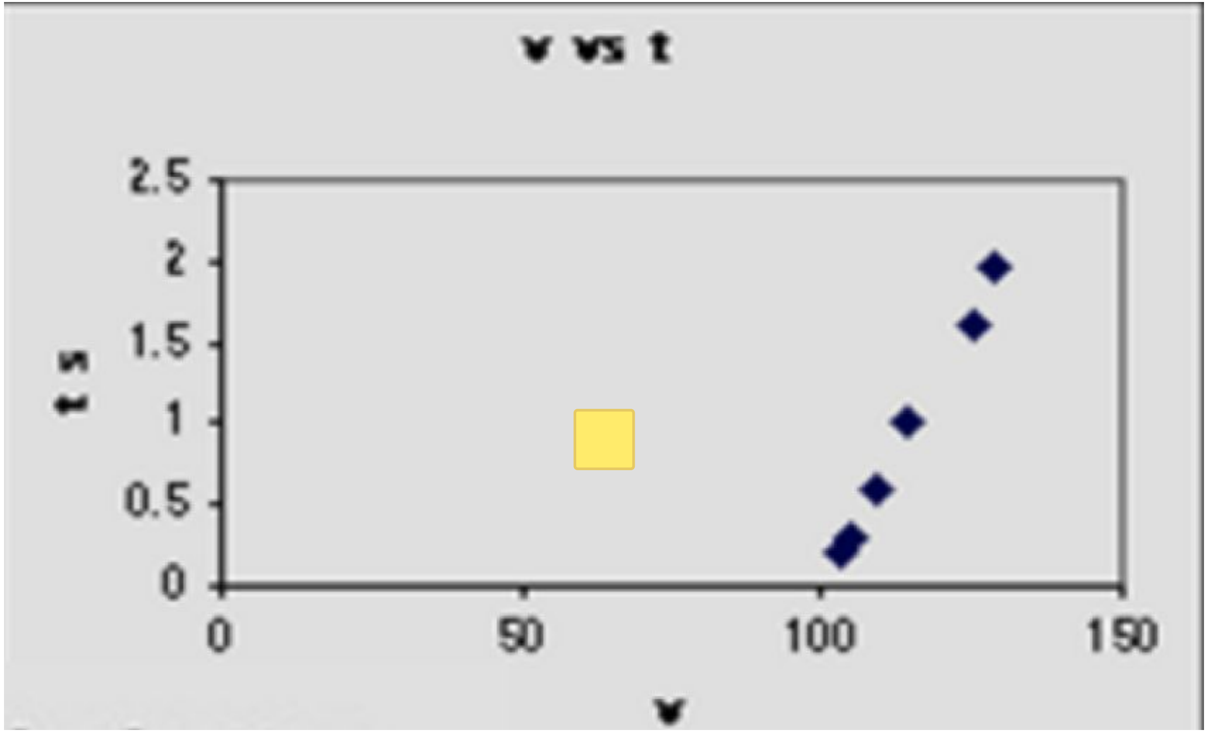
https://en.wikipedia.org/wiki/Hertzsprung%E2%80%93Russell_diagram

What makes a bad plot?

Ambiguity | distortion | distraction.

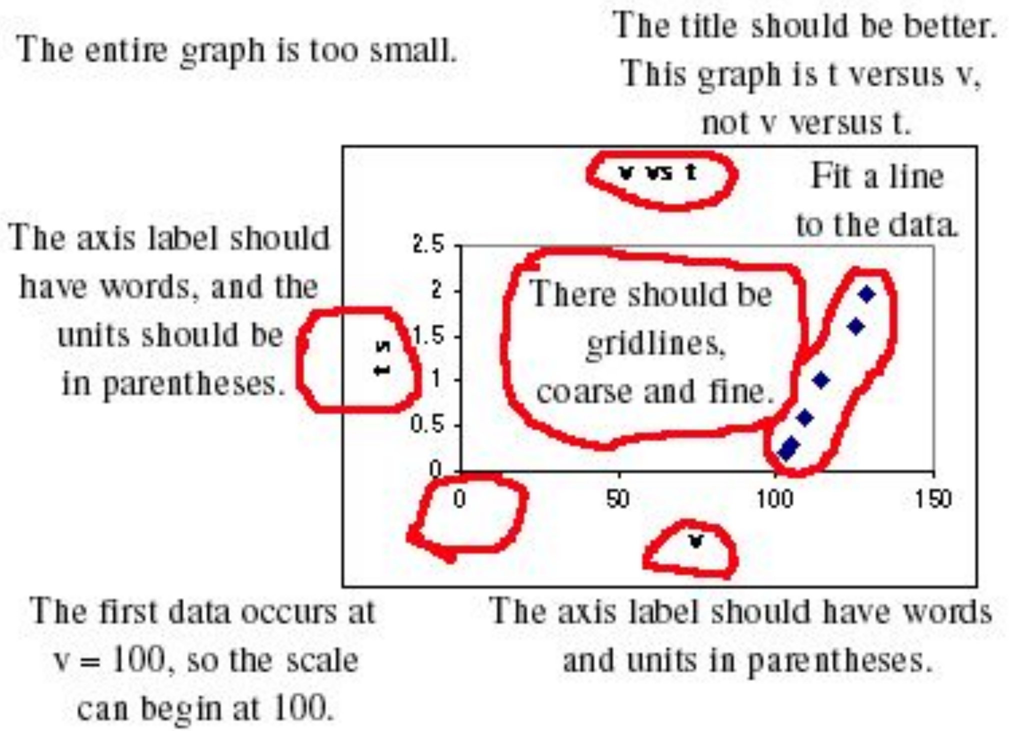
Ambiguity | distortion | distraction.

What is wrong with this plot?



http://www.geol.lsu.edu/jlorenzo/geophysics/graphing/badlinear1.html

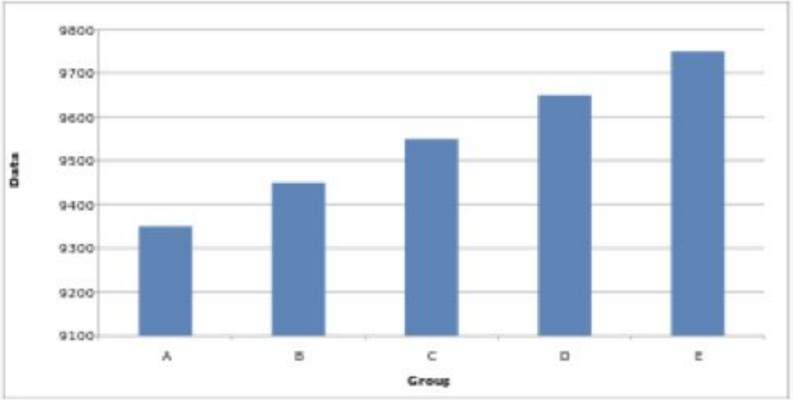
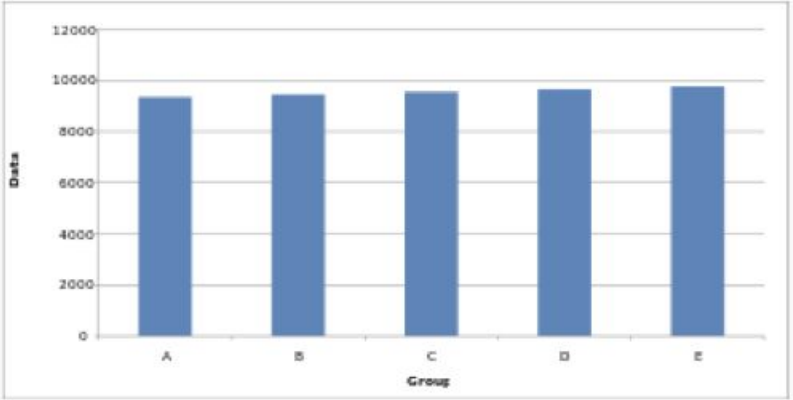
What is wrong with this plot?



Ambiguity | **distortion** | distraction.

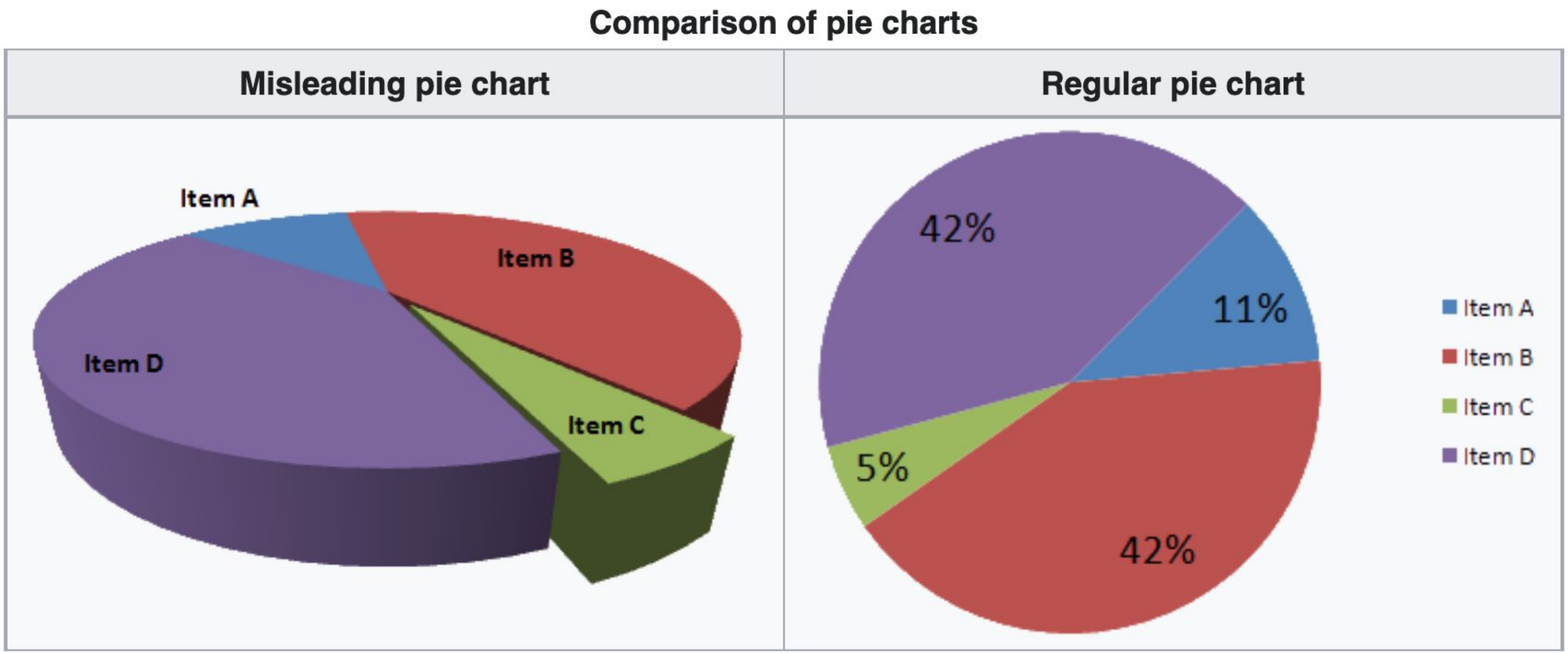
Distortion = misleading representation

Truncated bar graph

Truncated bar graph	Regular bar graph
	

Note that both of these graphs display *identical data*; however, in the truncated bar graph on the left, the data *appear* to show significant differences, whereas in the regular bar graph on the right, these differences are hardly visible.

Distortion = misleading representation

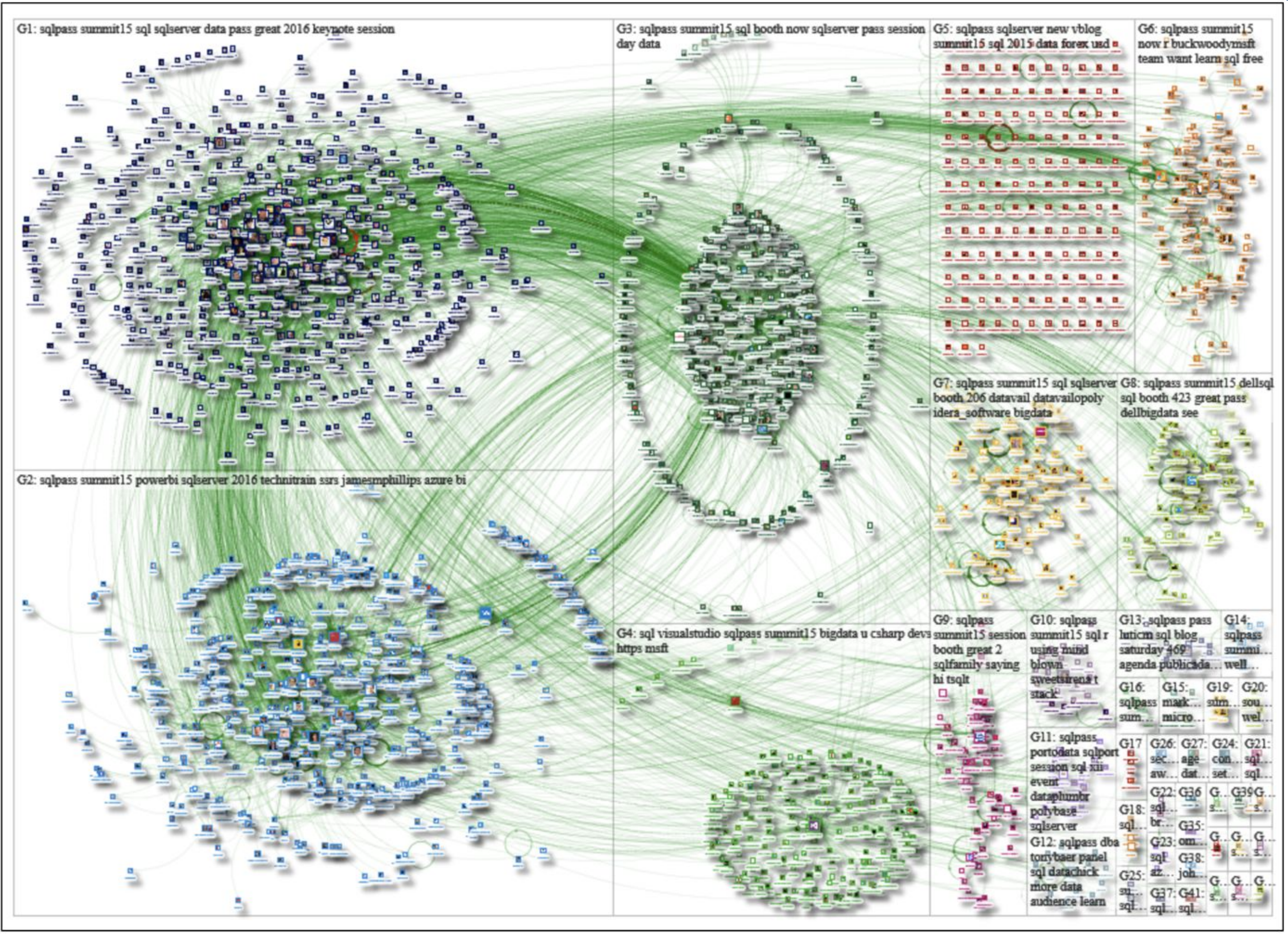


Item C appears to be at least as large as Item A in the misleading pie chart, whereas in actuality, it is less than half as large.

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

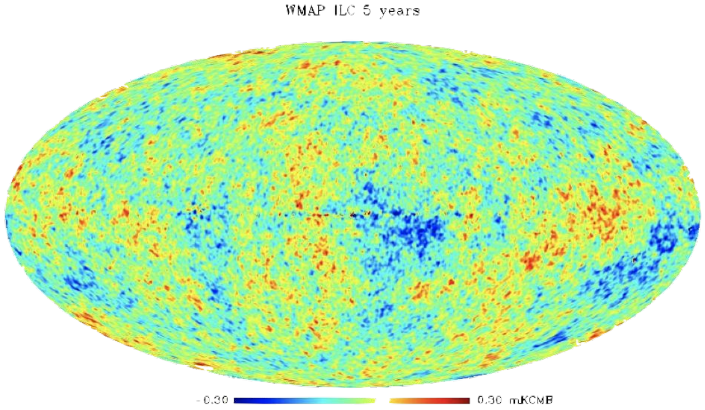
Ambiguity | distortion | **distraction.**

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
POLICE
NYPD



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

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

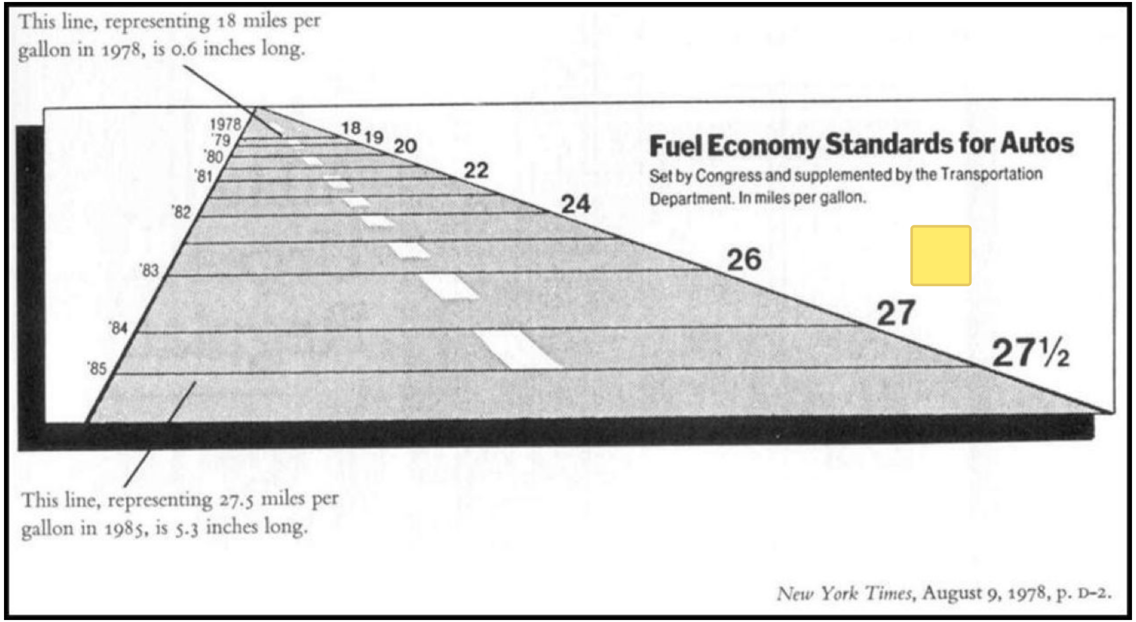


What makes a good visualization?

Tufte's rules

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

Data-ink ratio

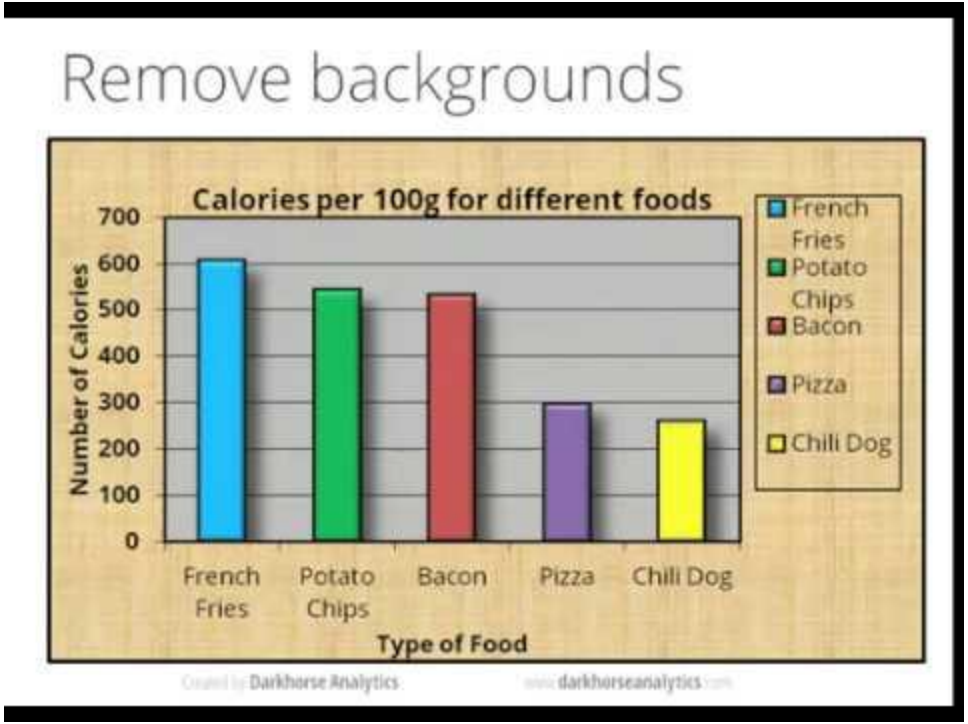
The data-ink ratio is the proportion of Ink that is used to present actual data compared to the total amount of ink (or pixels) used in the entire display. (Ratio of Data-Ink to non-Data-Ink).

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

$$= \text{proportion of a graphic's ink devoted to the non-redundant display of data-information}$$

$$= 1.0 - \text{proportion of a graphic that can be erased}$$

Data-ink ratio



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")

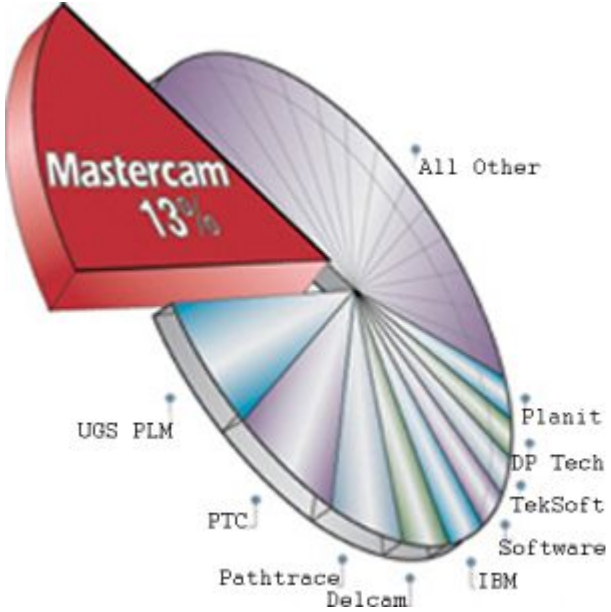
Effect size ~ 1

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

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

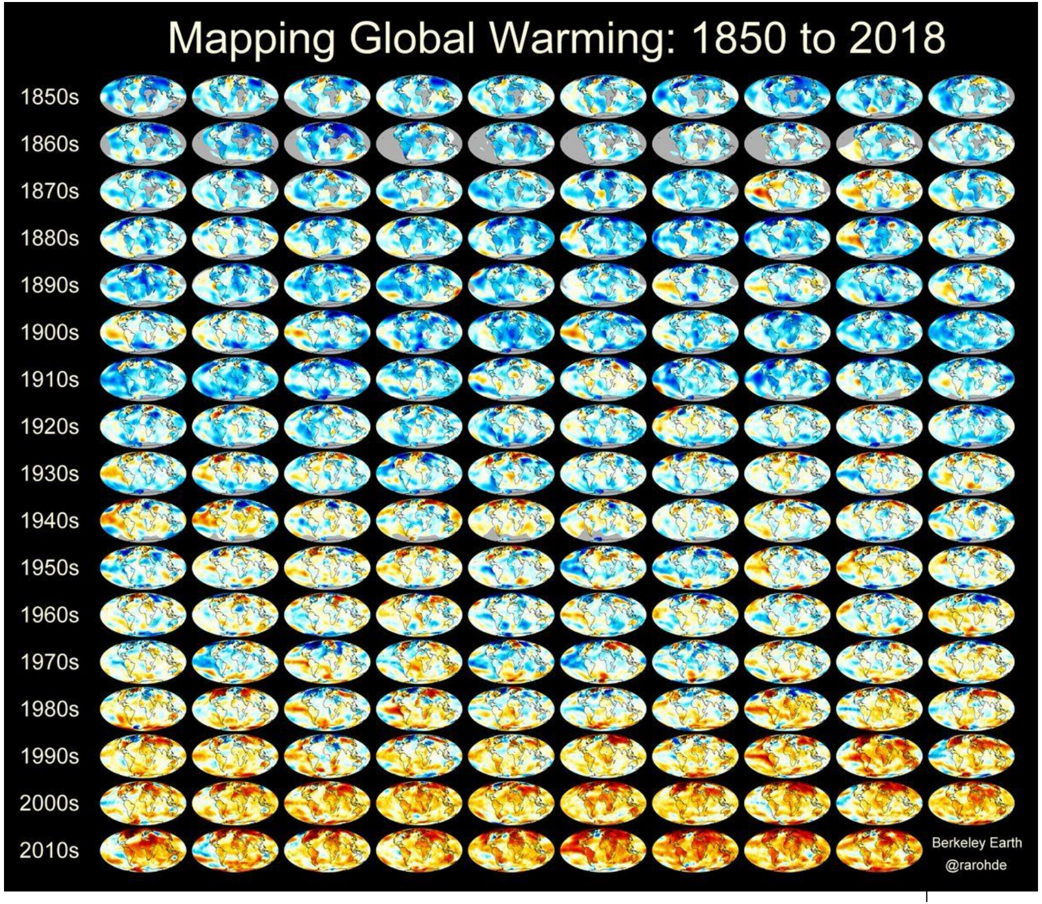
3) Show data variation, not design variation

no chart junk

ETH zürich

Small multiples

encourage comparison



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

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

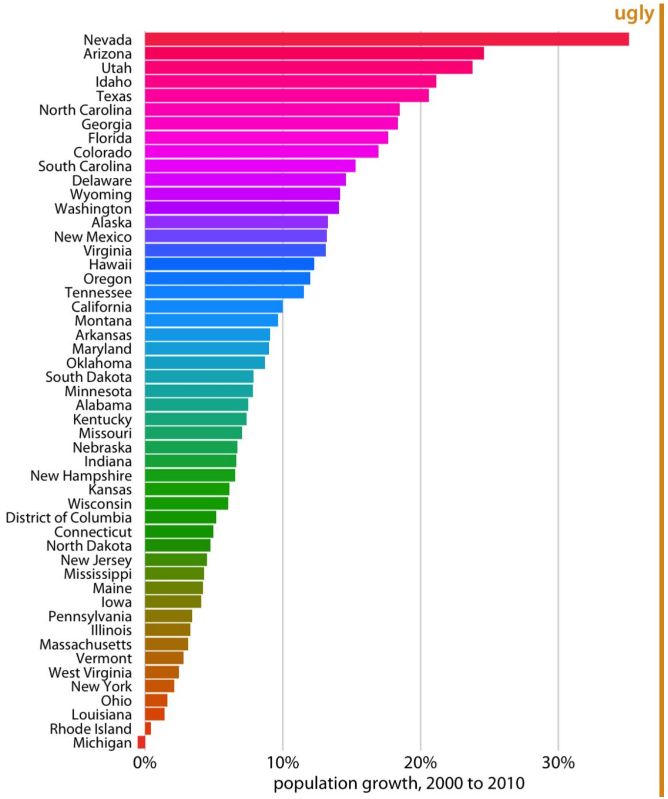
3) Show data variation, not design variation

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

Redundancy of information



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 redundant communication

Tufte's rules

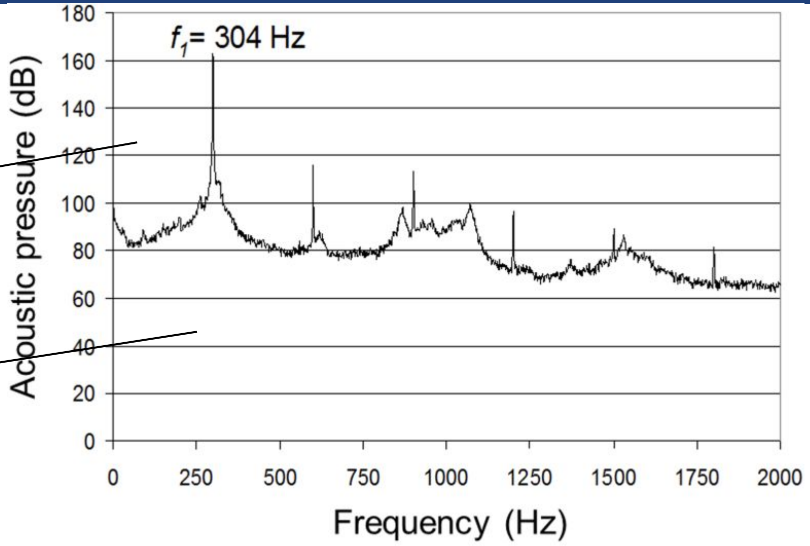
low data/ink ratio

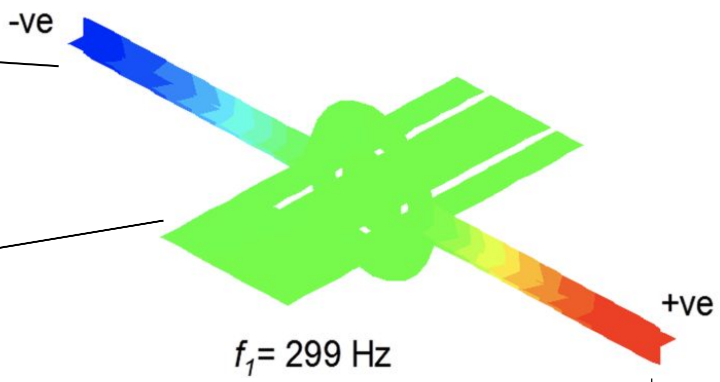
no comparison

chart junk

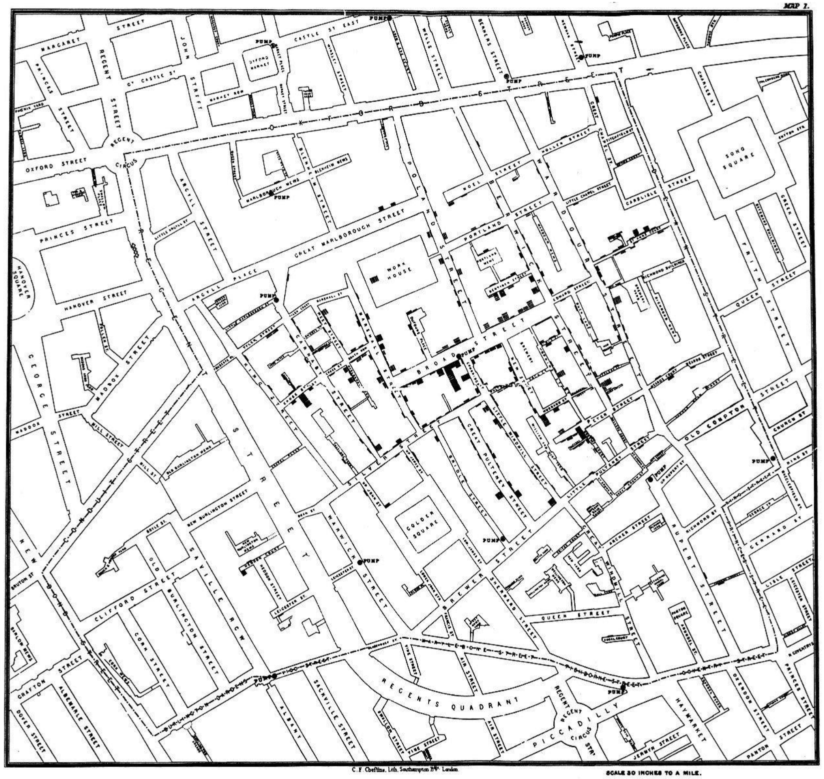
2 graphical elements for frequency

(color and position)





Use visualization to understand data, not to communicate a result



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[httpsarchiveorgstreamb28985266page38mode1up_38]_19-0

More rules

Function first, Form next

no unjustified beauty

Get it right in Black & White

no unjustified color

consider designing your plot in BW first

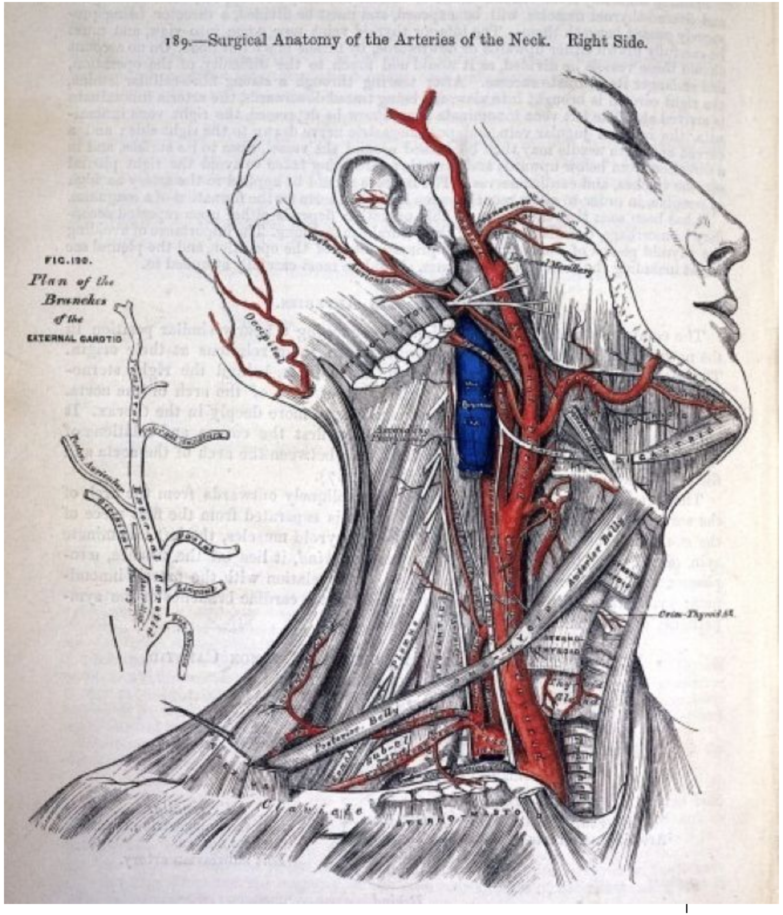
No Unjustified 3D

Eyes over Memory

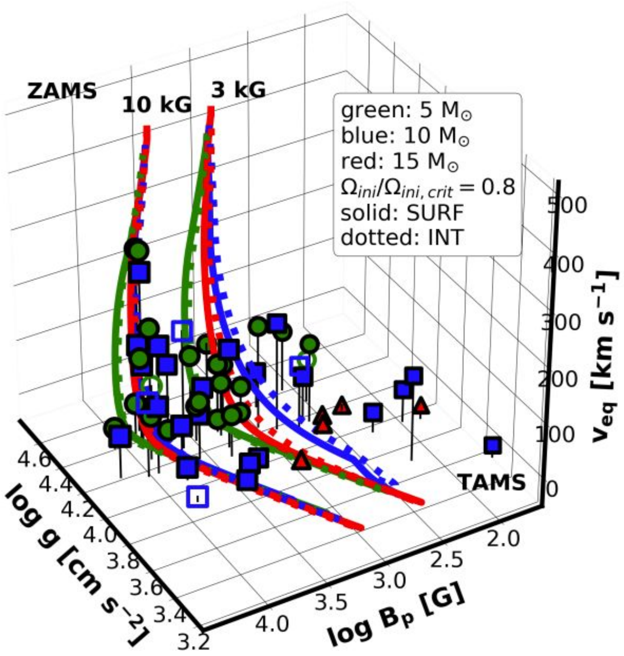
no unjustified animation

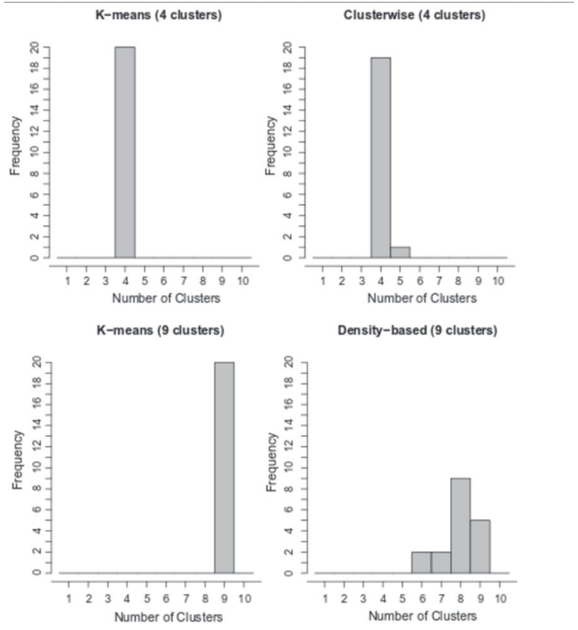
More examples

Functional use of colors (and distortion)



More examples





No unjustified 3D

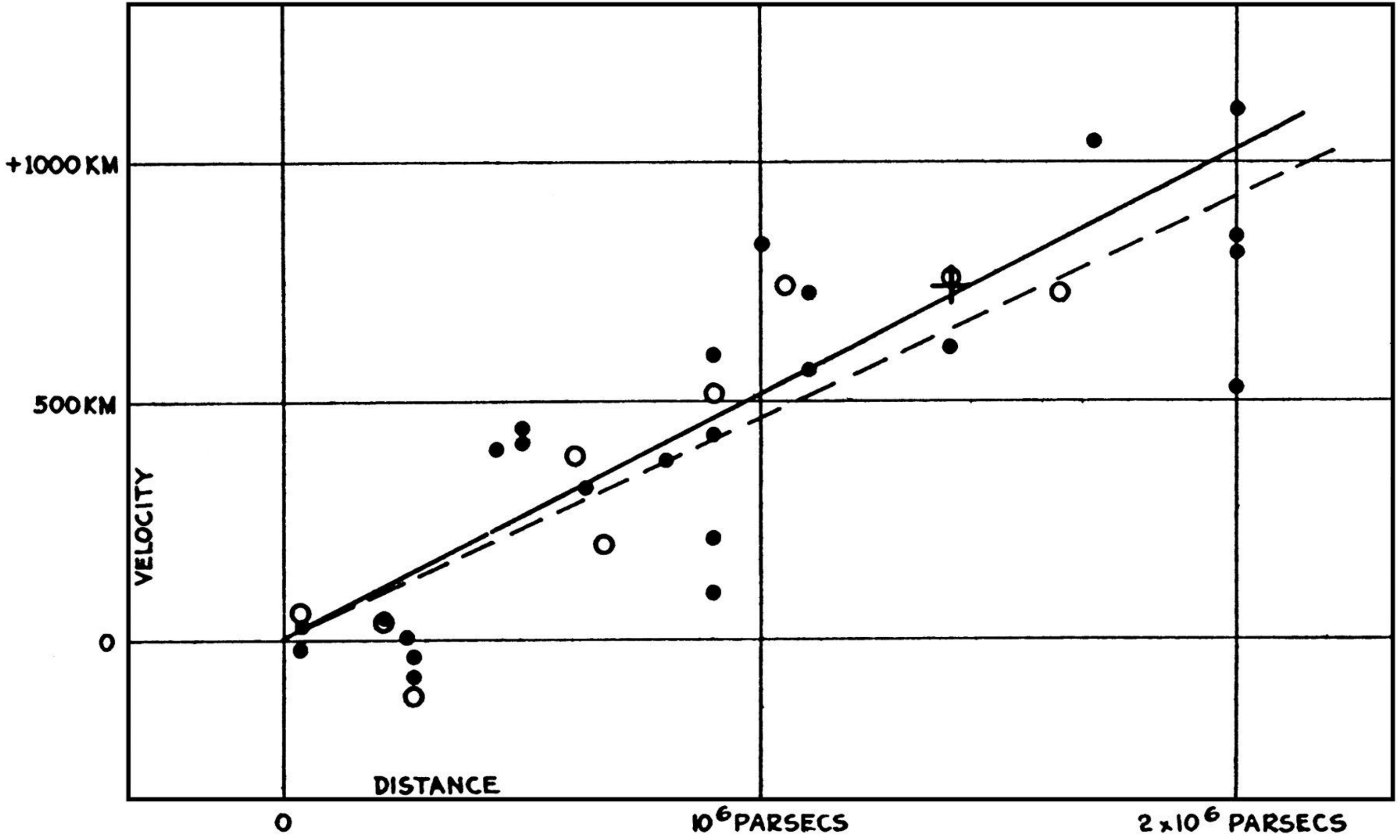
(or 2D either!)

obstruction

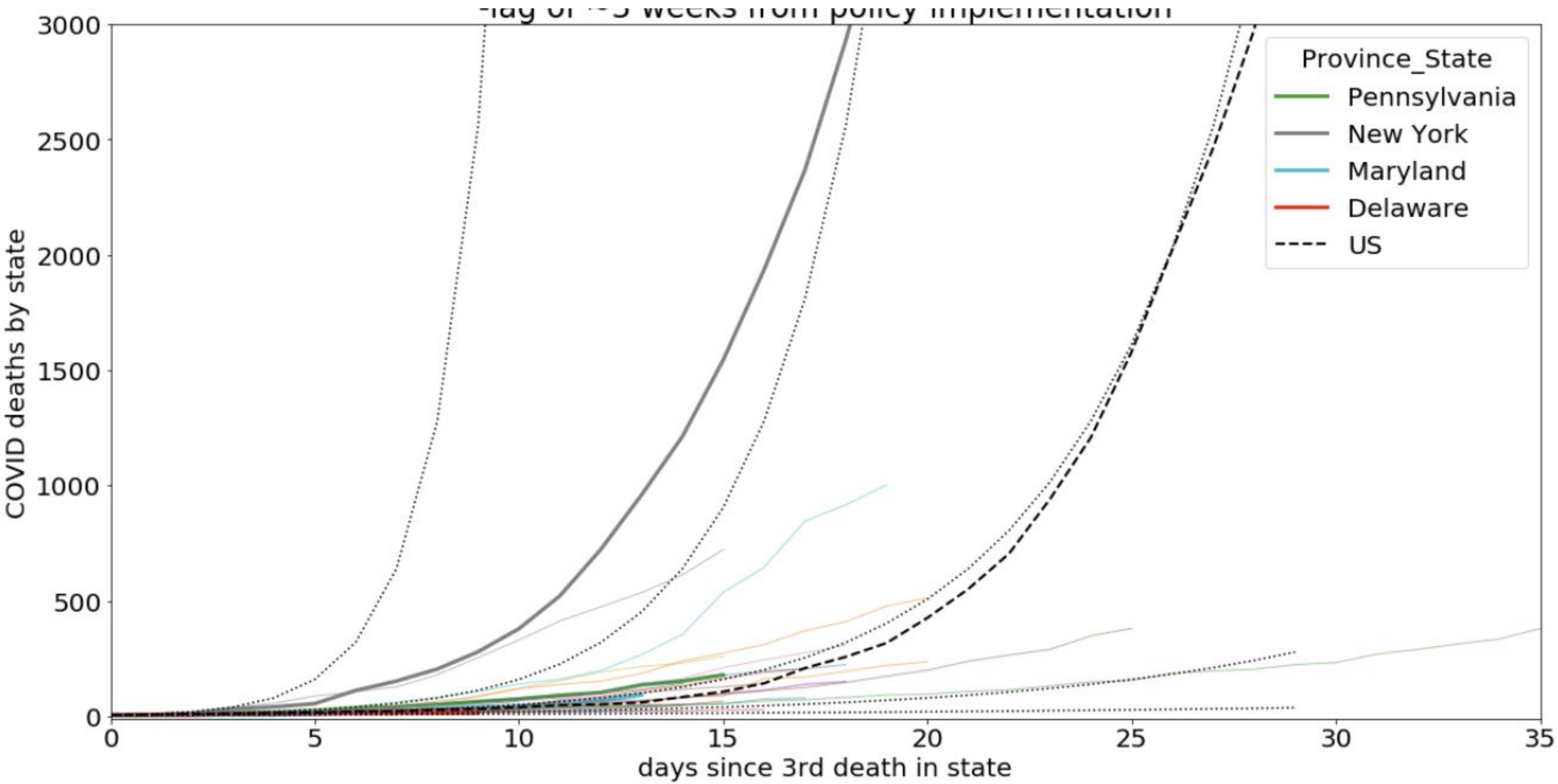
clutter

deformation

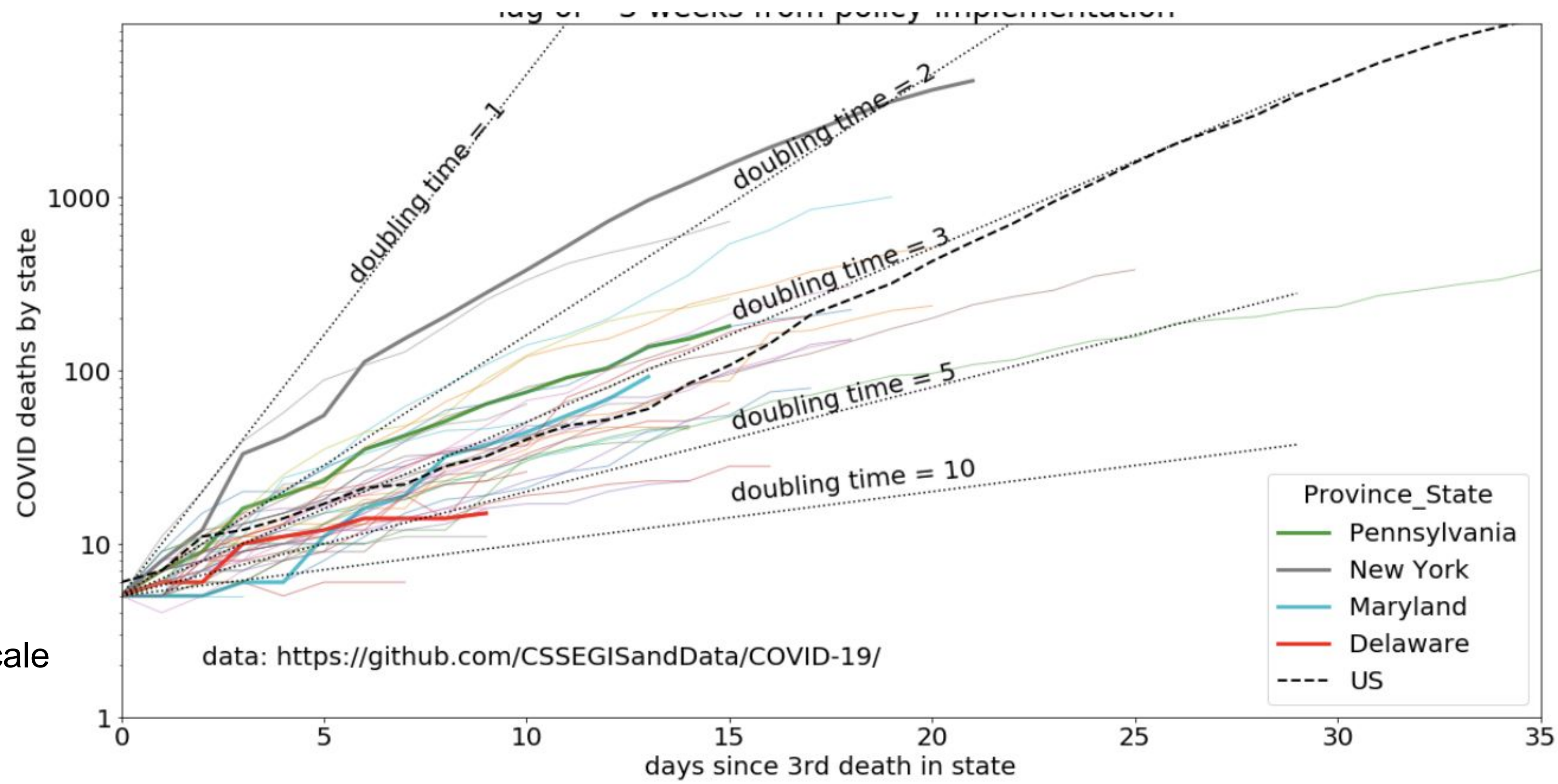
How would you improve these plots?



How would you improve these plots?

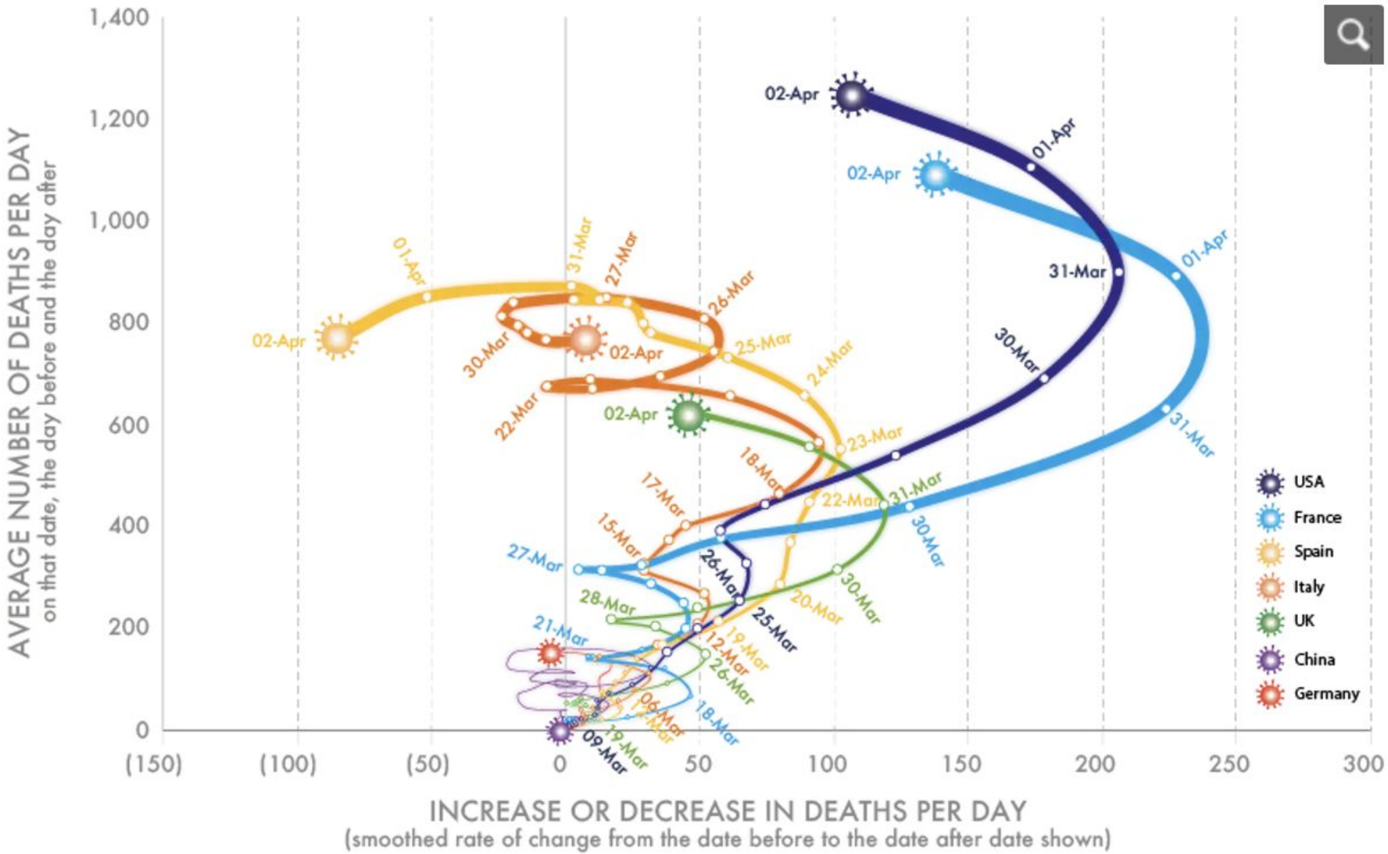


How would you improve these plots?



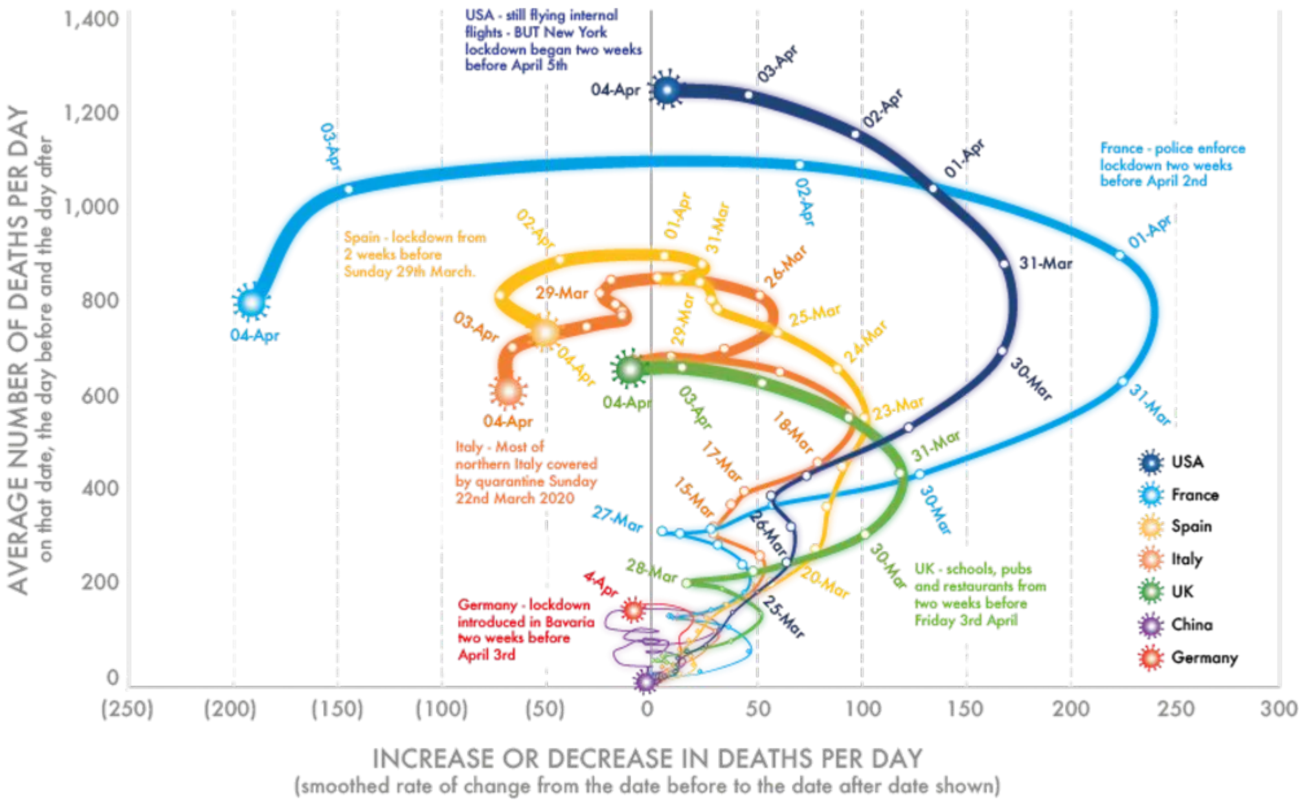
Log scale

How would you improve these plots?

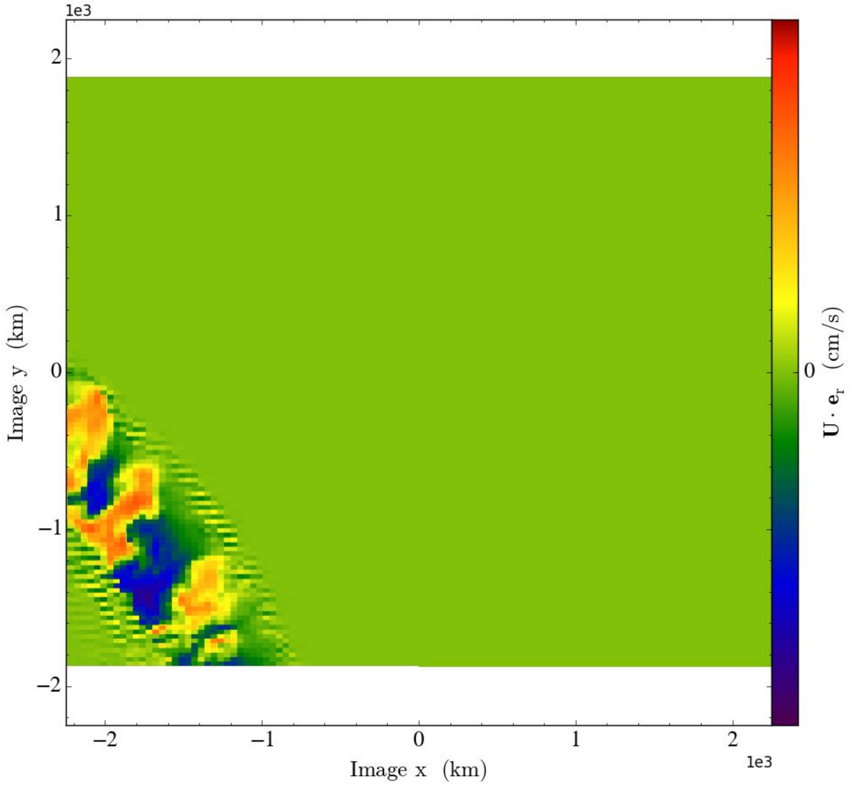


How would you improve these plots?

I would say this plot is at the limit of confusion (information saturation)

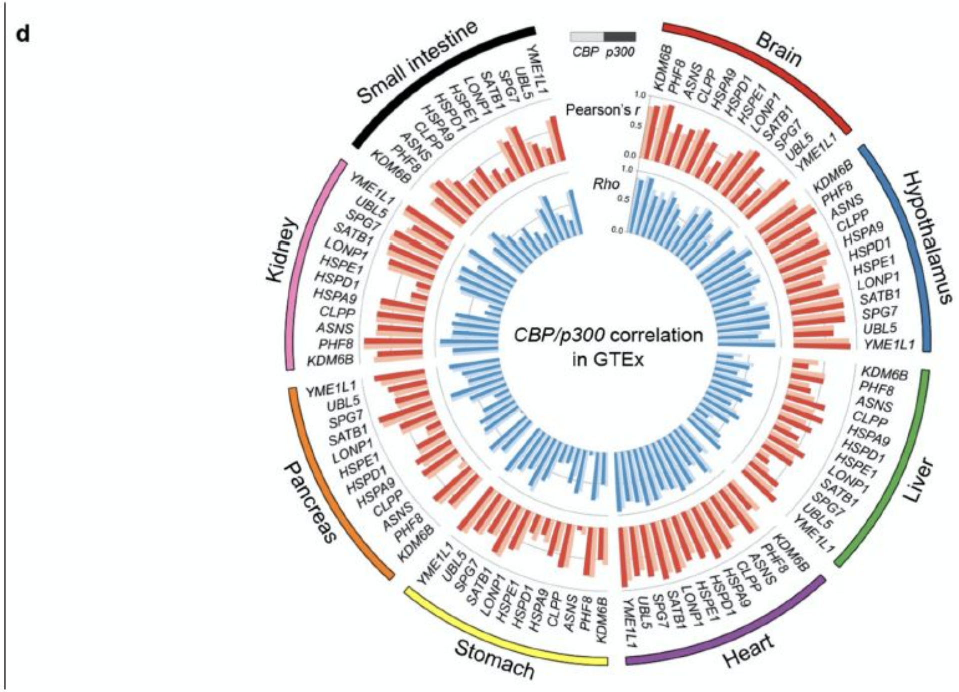


How would you improve these plots?



ETH zürich

How would you improve these plots?



Li et al. Nat Aging. Author manuscript; available in PMC 2021 March 13.

To conclude:

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.