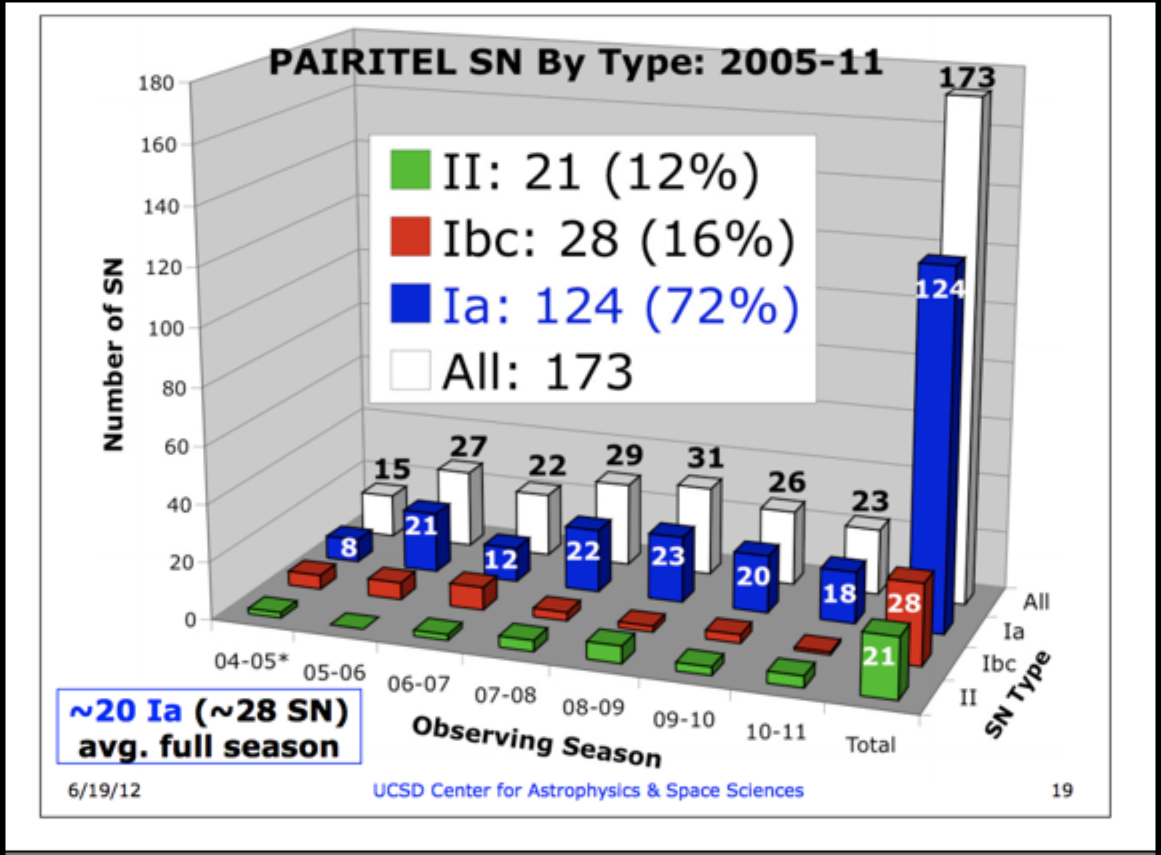
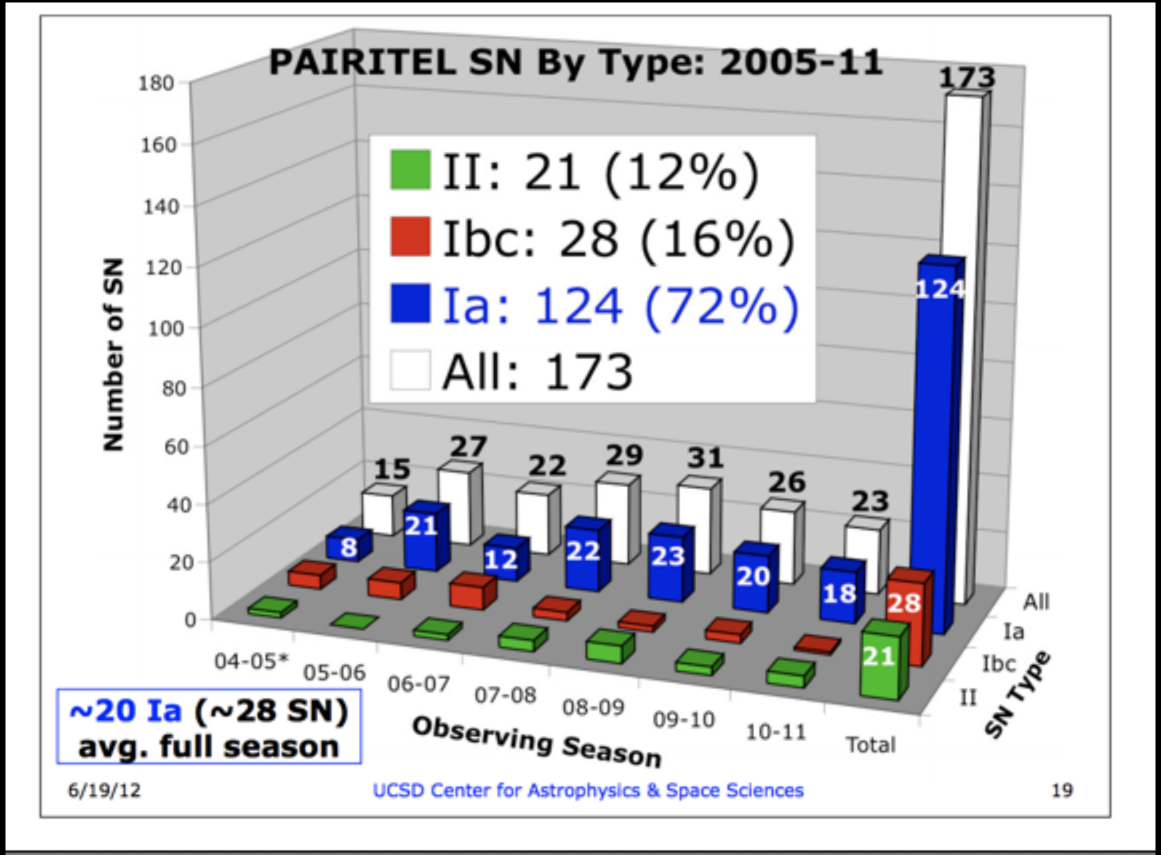
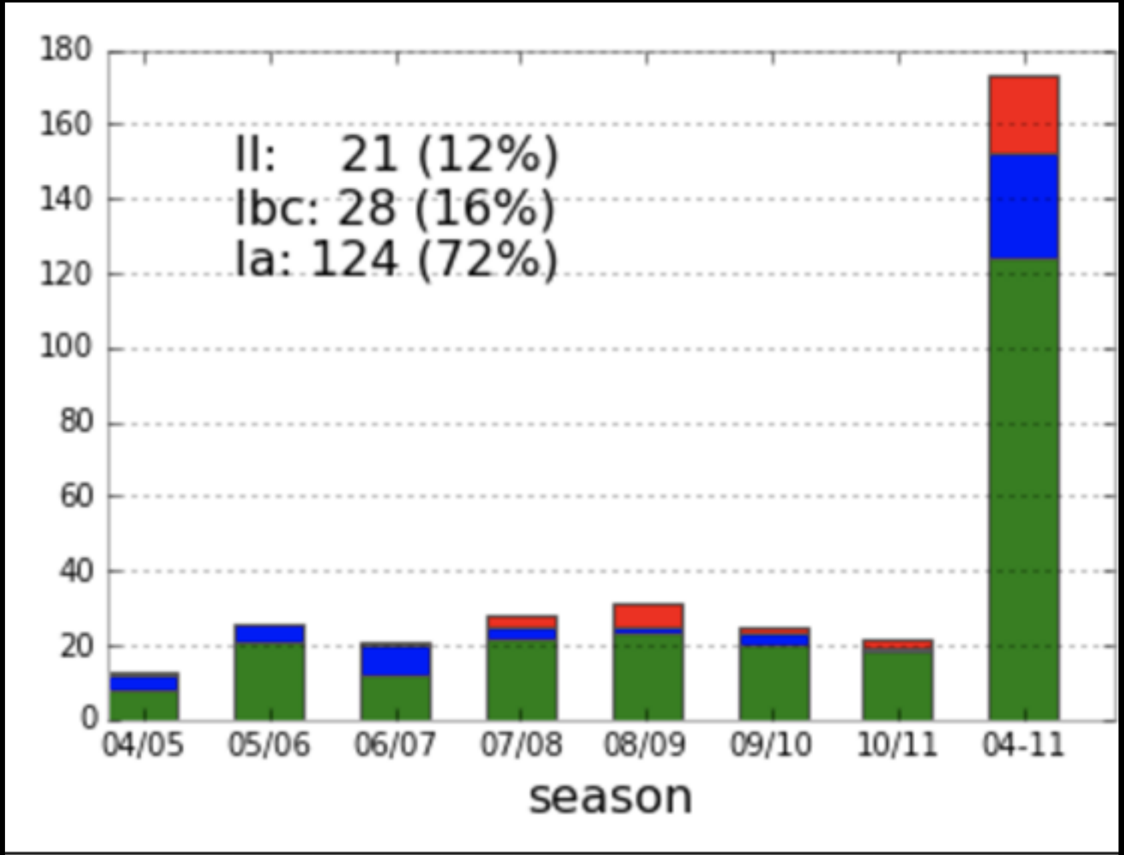
how would you improve these plots?

A true story of a plot created for in clusion in a paper of mine

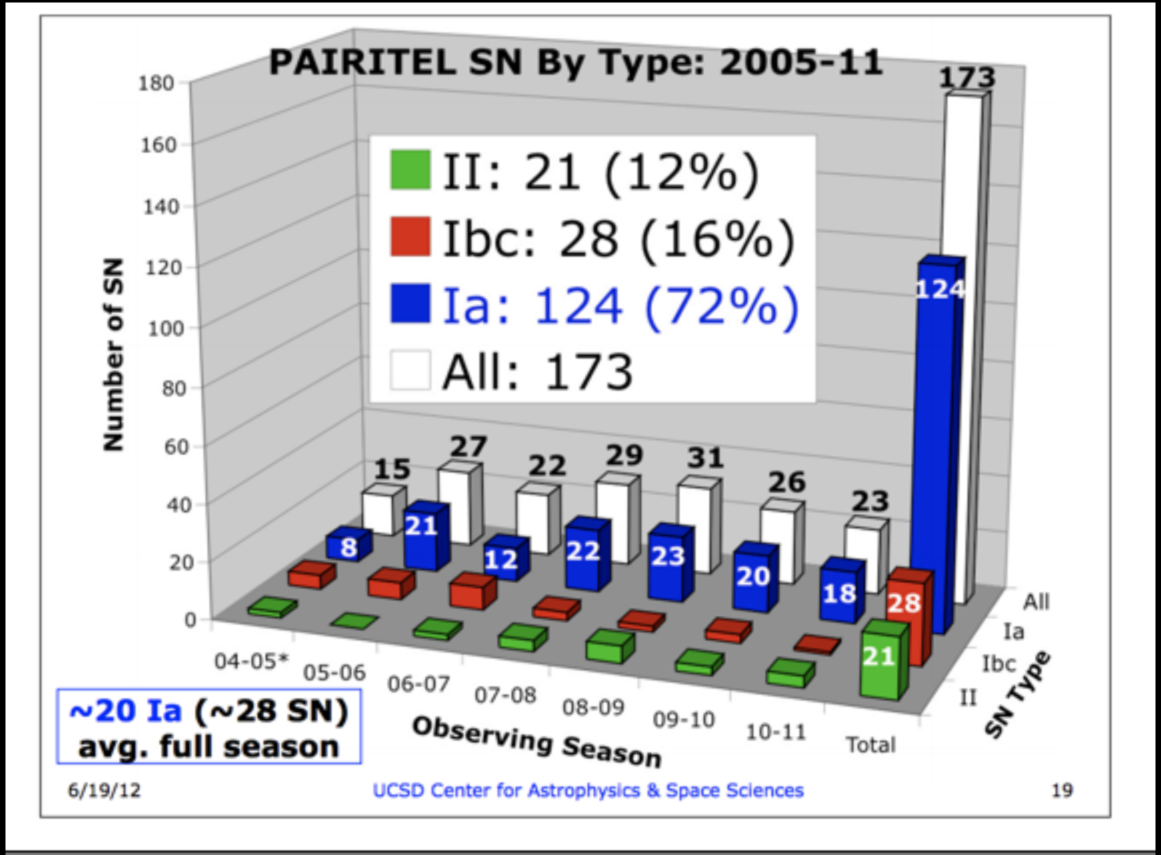


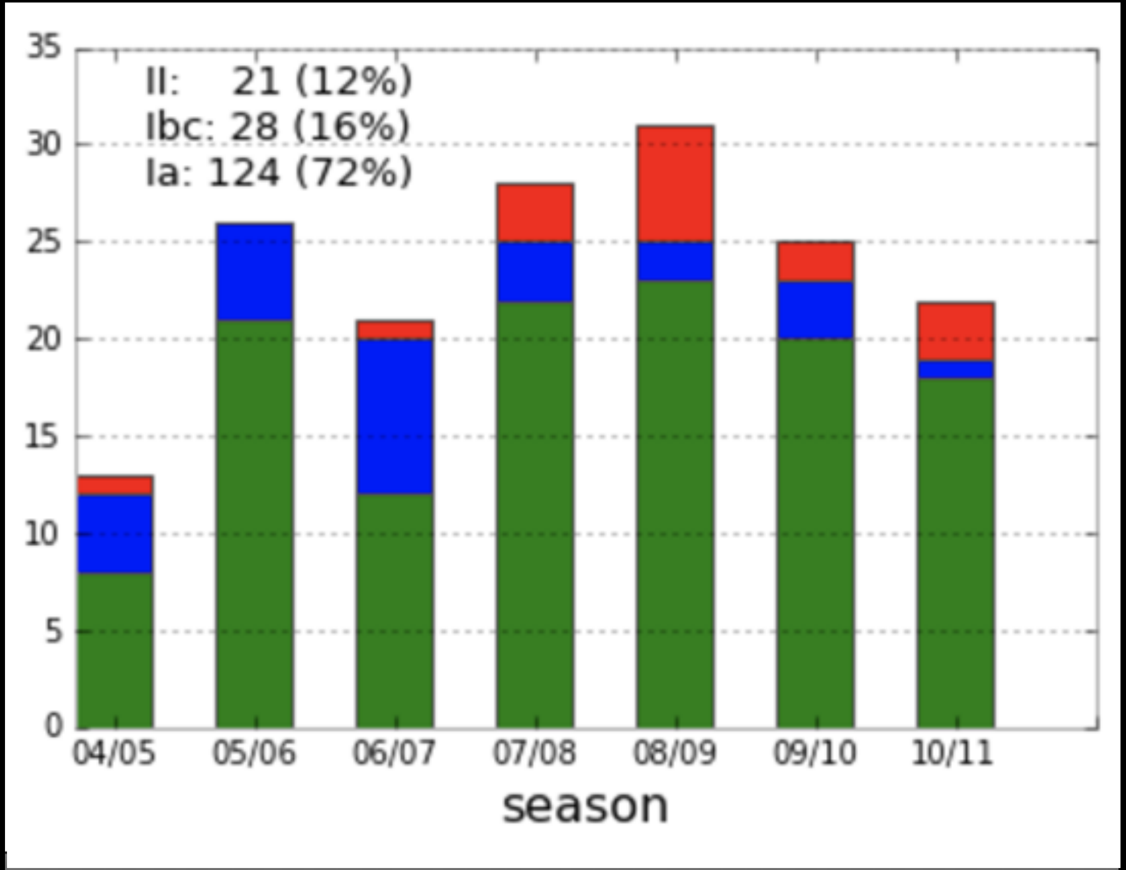
A true story of a plot created for in clusion in a paper of mine



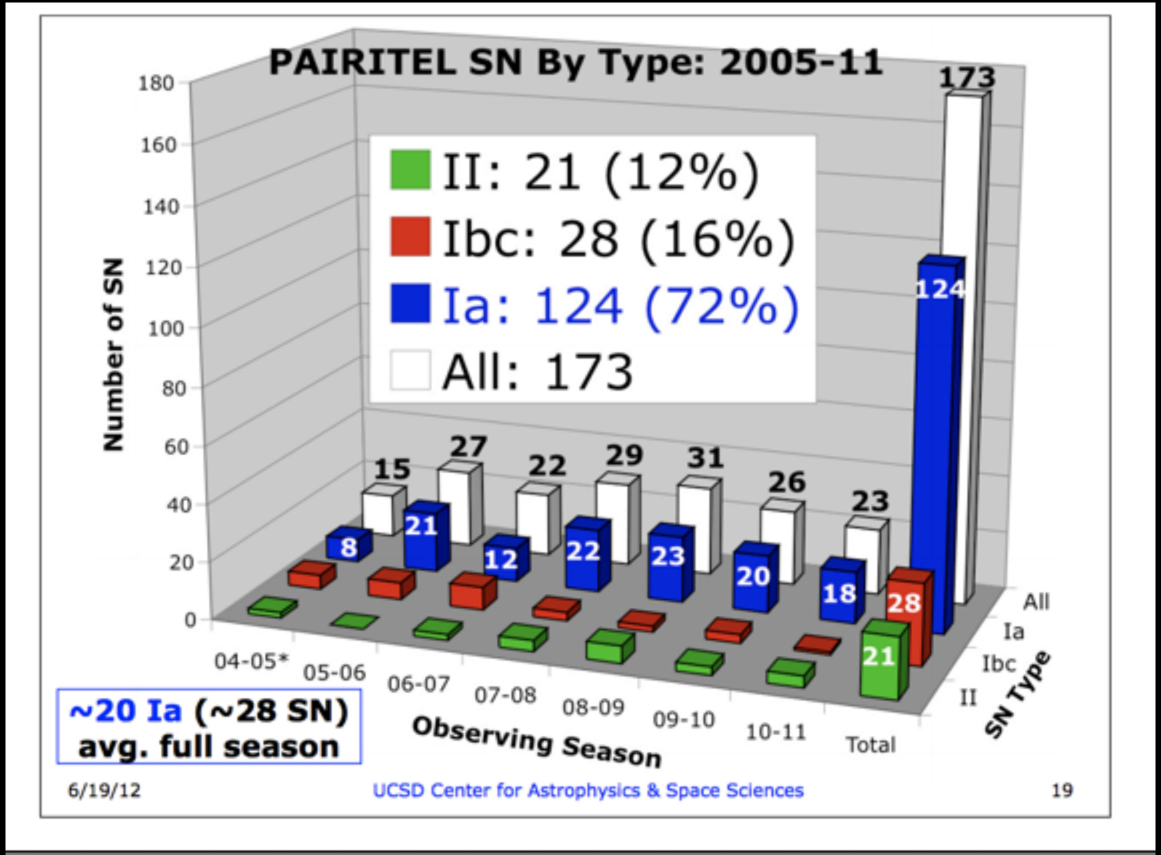


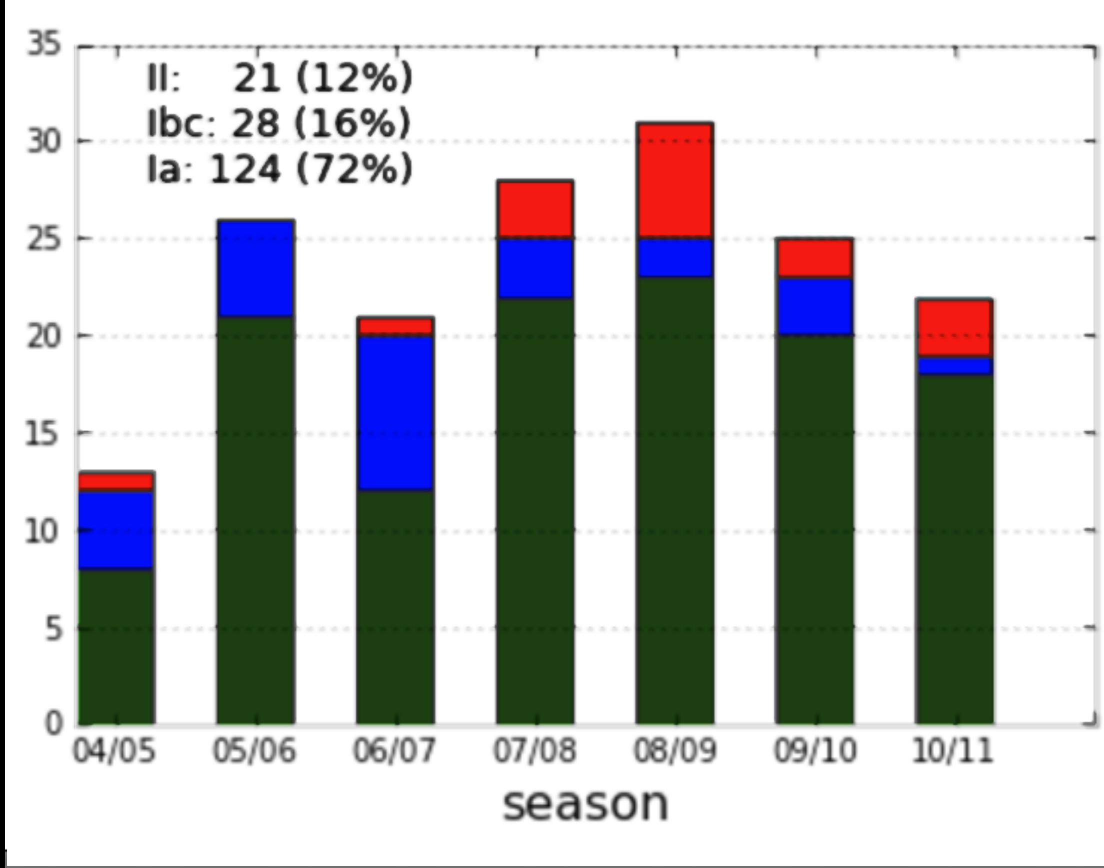
A true story of a plot created for in clusion in a paper of mine



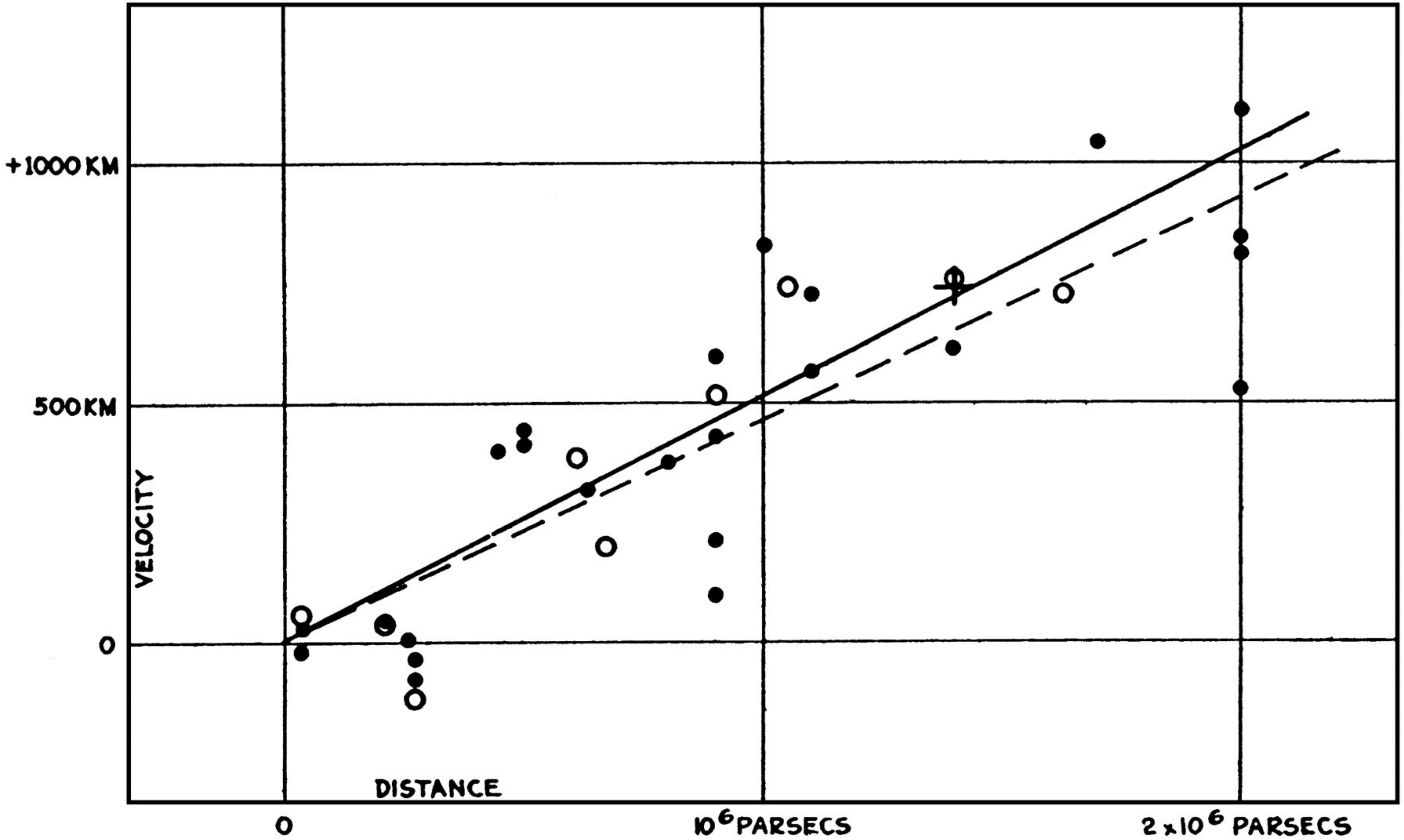


A true story of a plot created for in clusion in a paper of mine





how would you improve these plots?

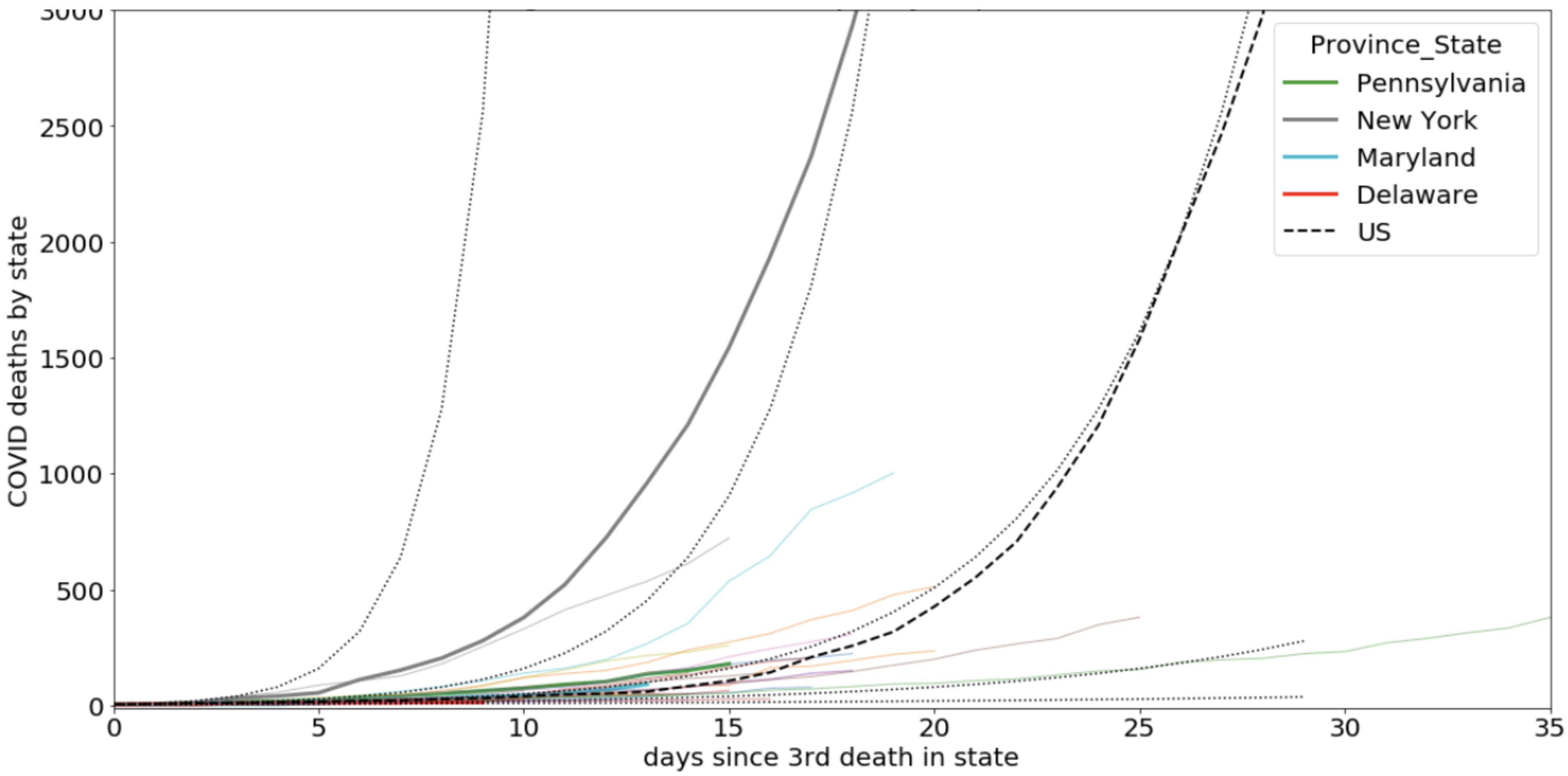


https://www.pnas.org/doi/10.1073/pnas.253679910

Plot A

E. Hubble 1929

how would you improve these plots?

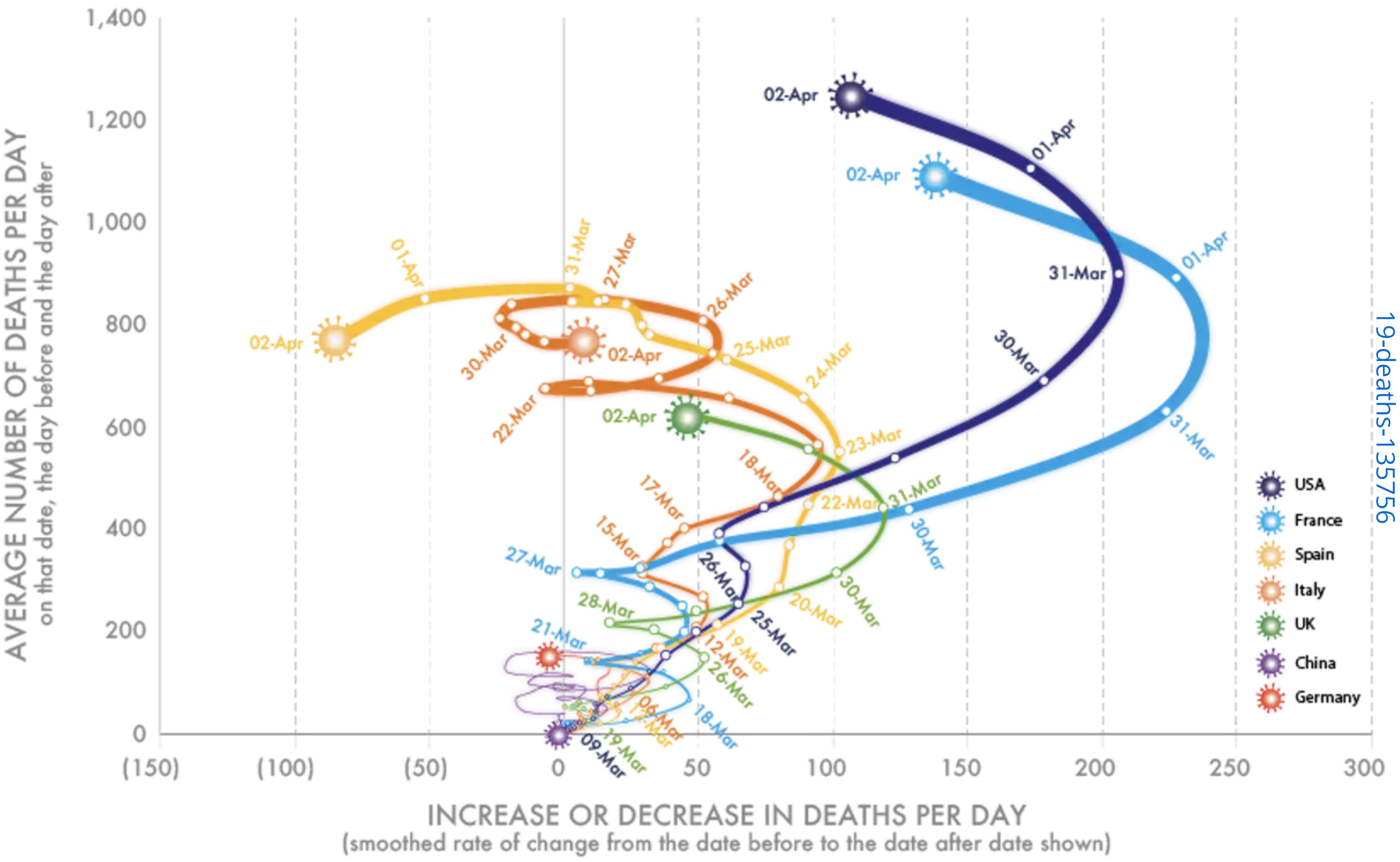


Plot B

2020

how would you improve these plots?

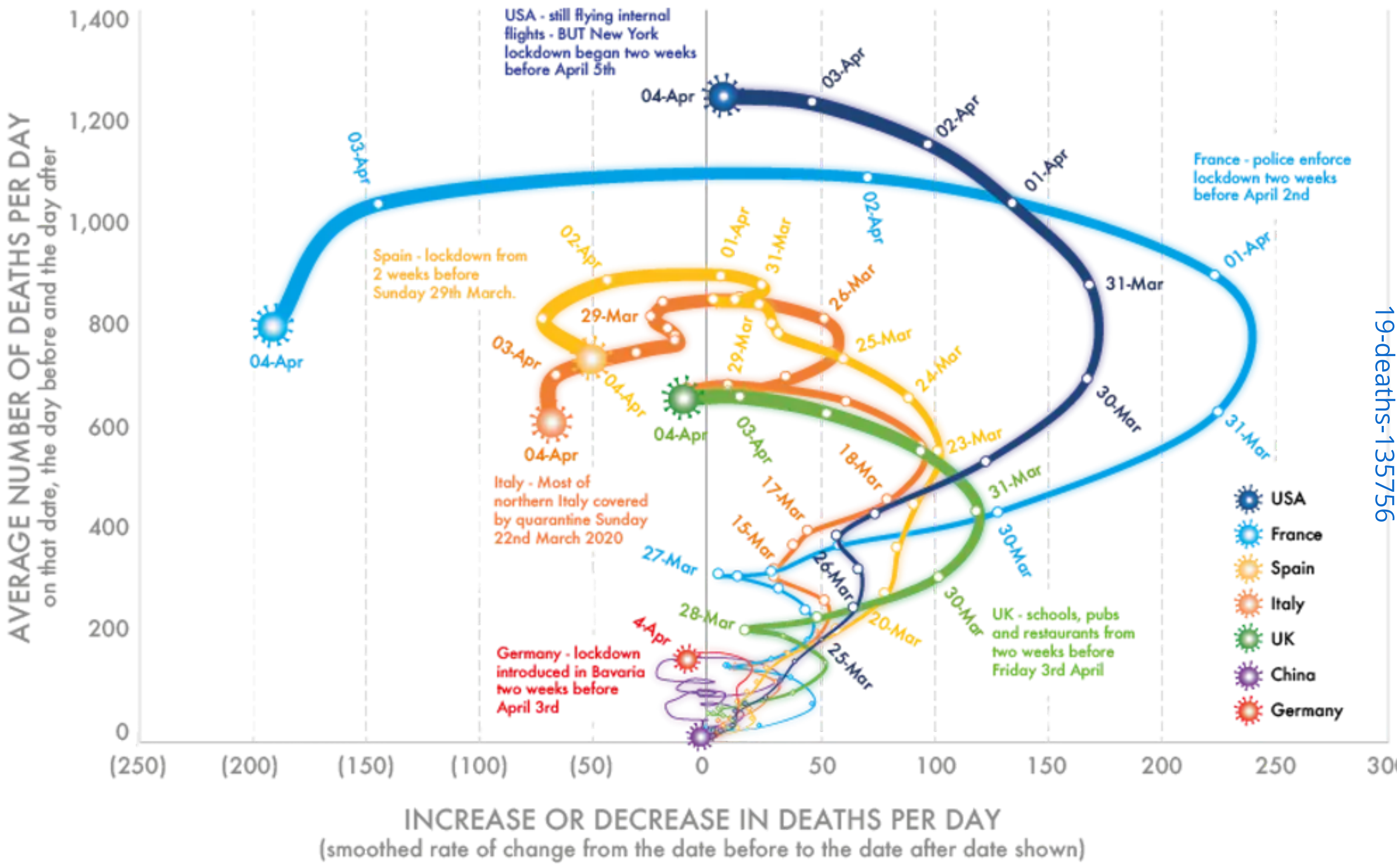
Plot C



https://theconversation.com/three-graphs-that-show-a-global-slowdown-in-covid-19-deaths-135756

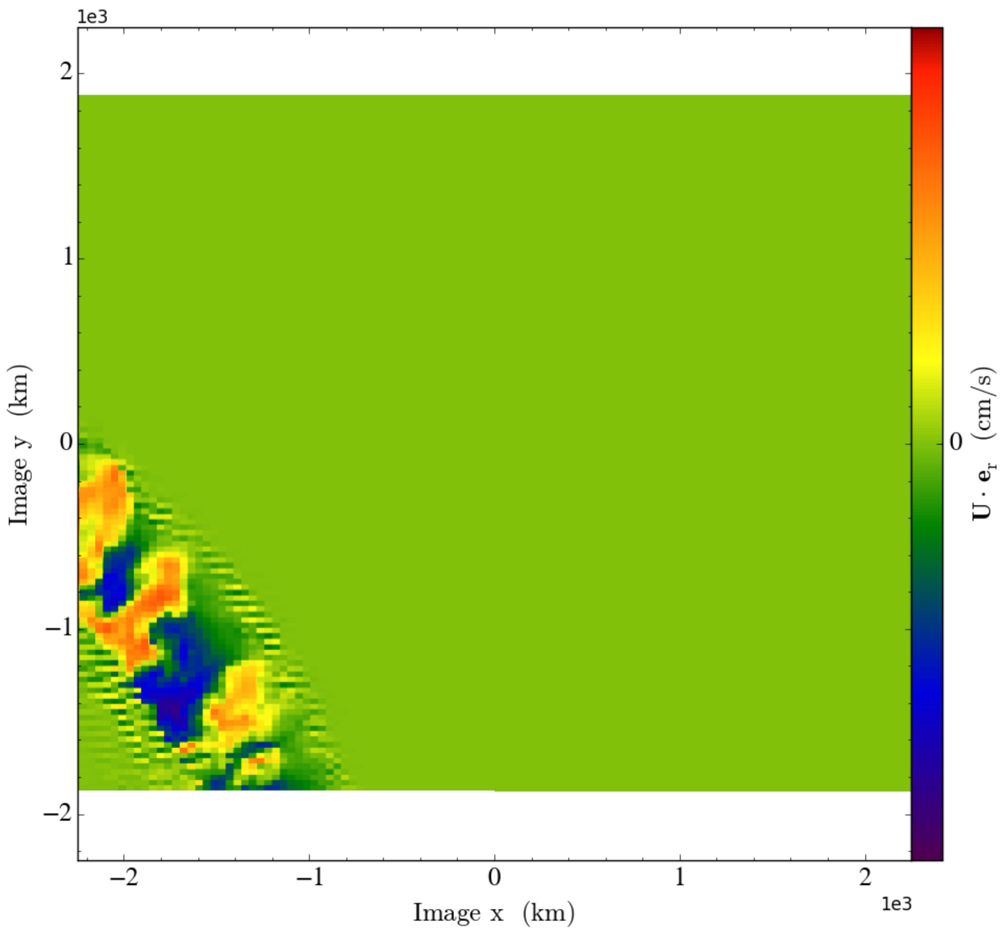
how would you improve these plots?

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



https://theconversation.com/three-graphs-that-show-a-global-slowdown-in-covid-19-deaths-135756

how would you improve these plots?

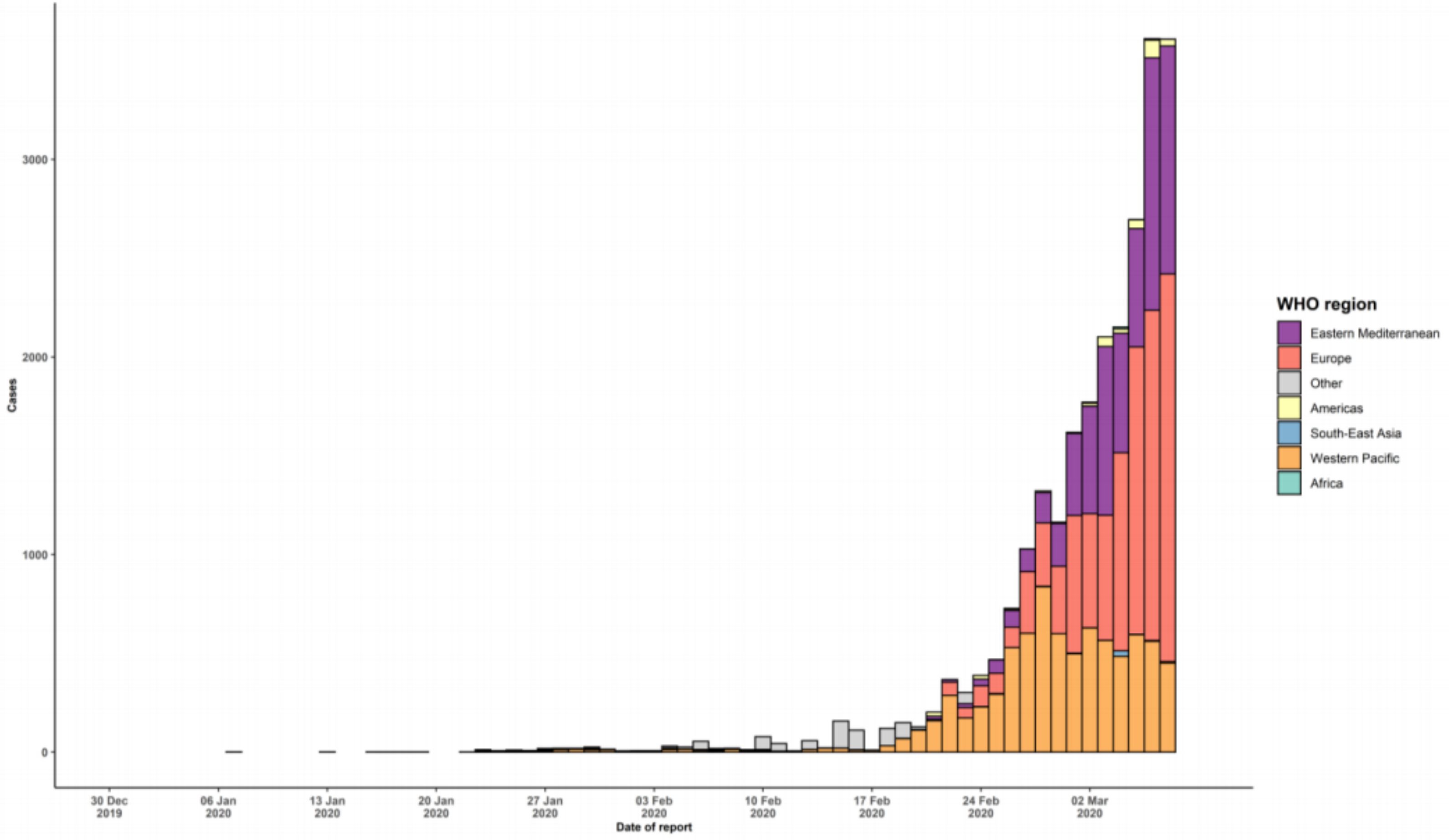


Plot D

(no author, no need to shame, but this was published in a peer reviewed journal)

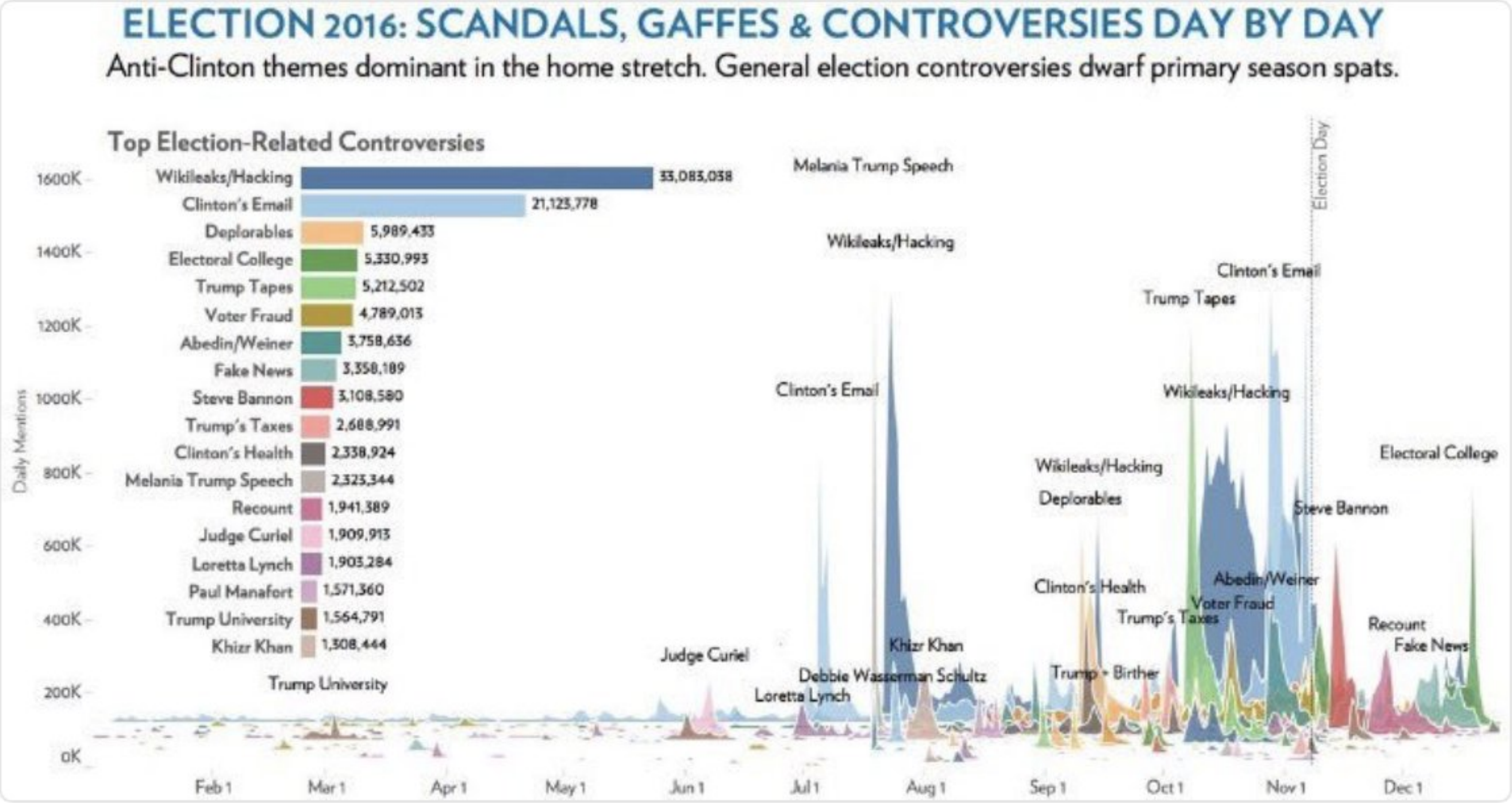
how would you improve these plots?

Figure 2. Epidemic curve of confirmed COVID-19 cases reported outside of China (*n*=24,727), by date of report and WHO region through 08 March 2020



Plot E

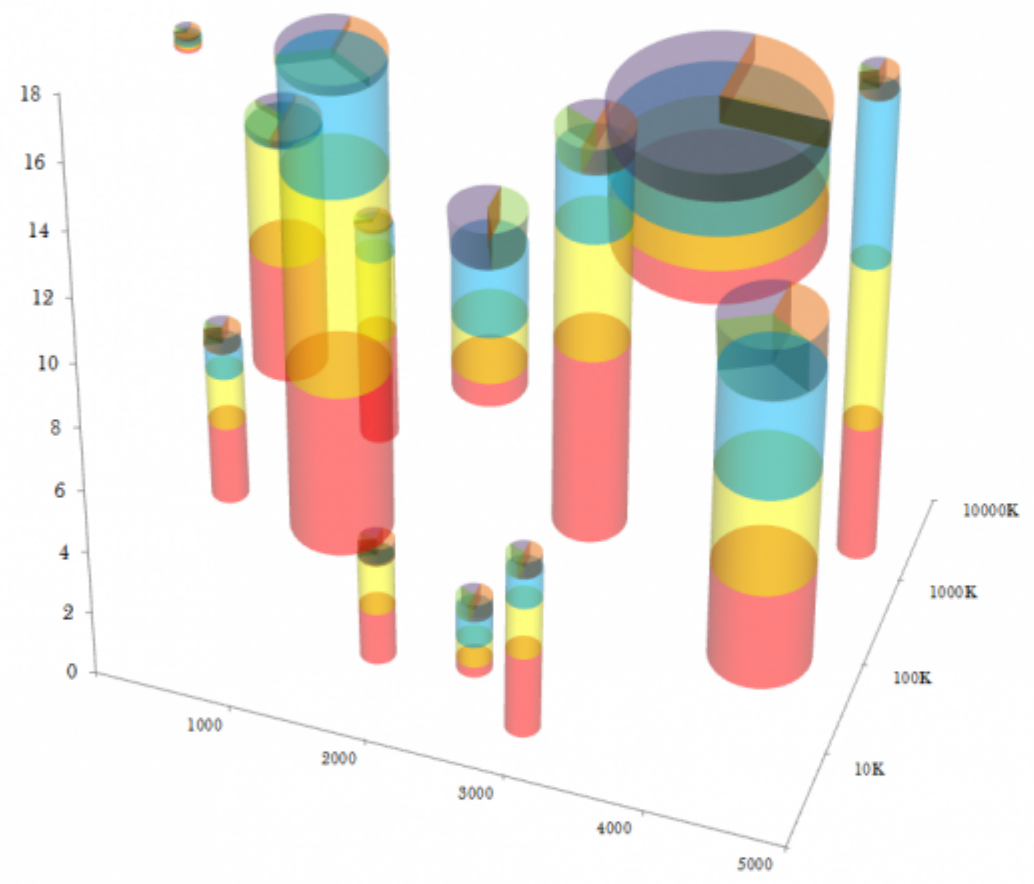
how would you improve these plots?



Plot F

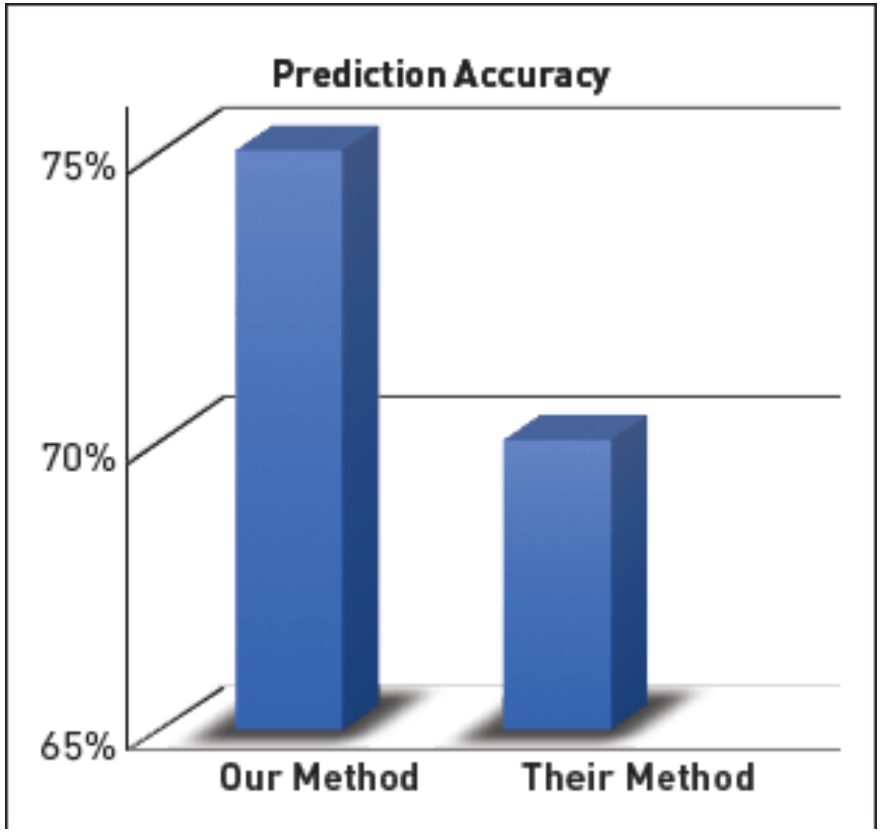
how would you improve these plots?

Plot G

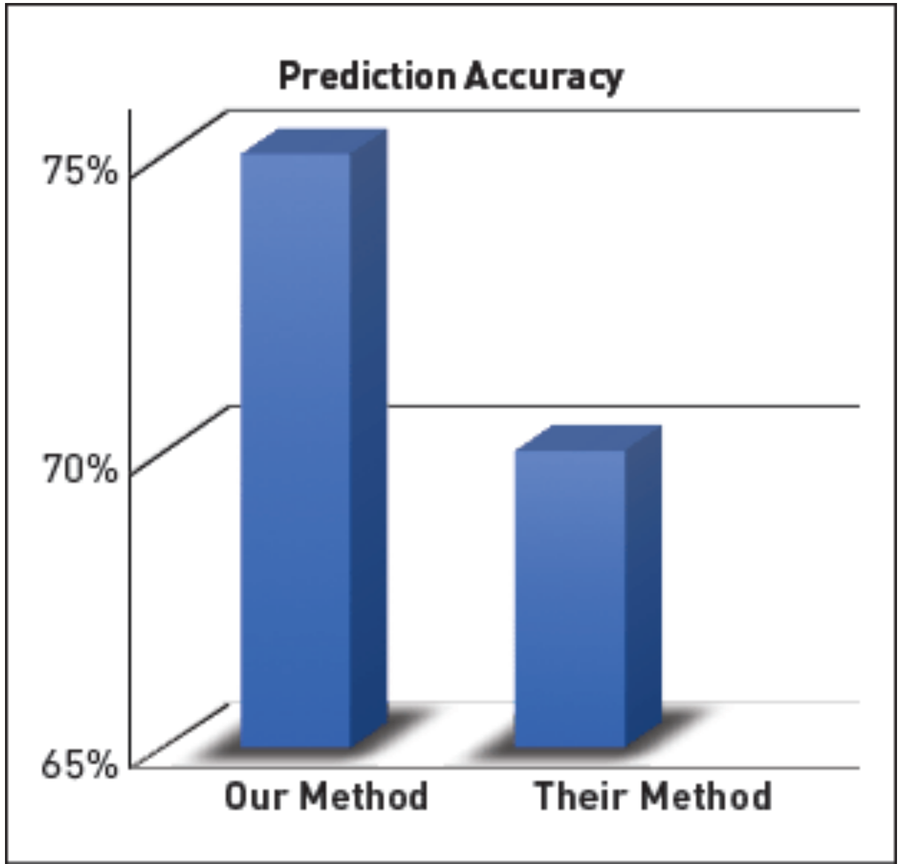


how would you improve these plots?

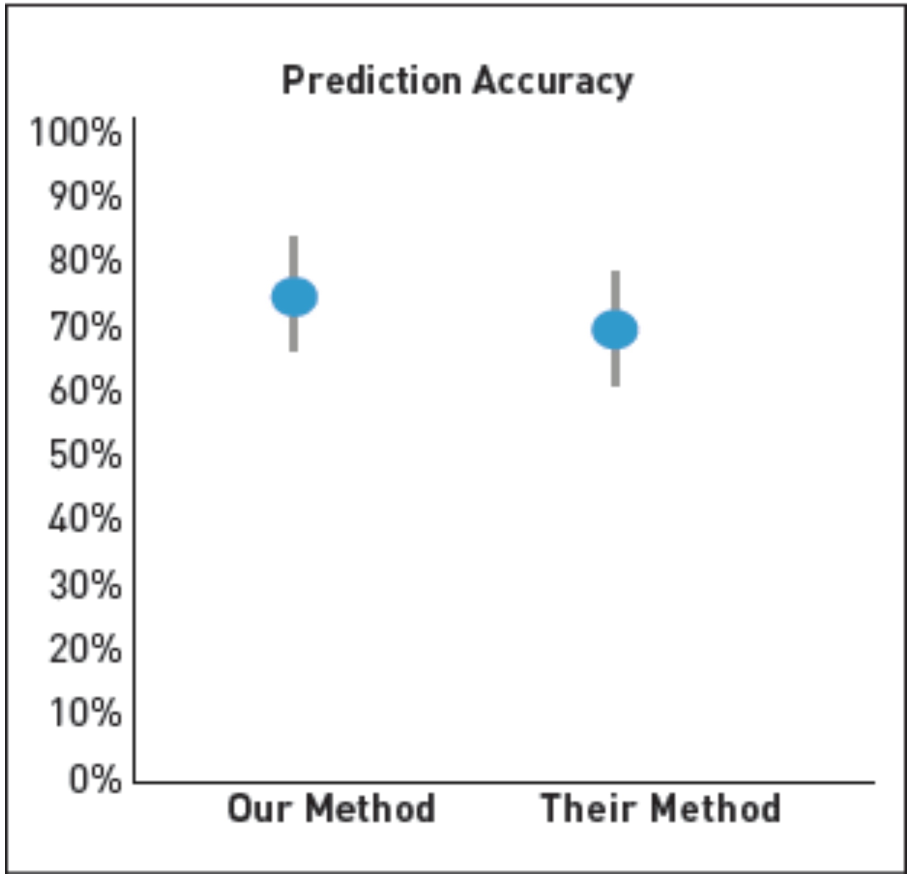
Plot H



how would you improve these plots?



Misleading



Improved

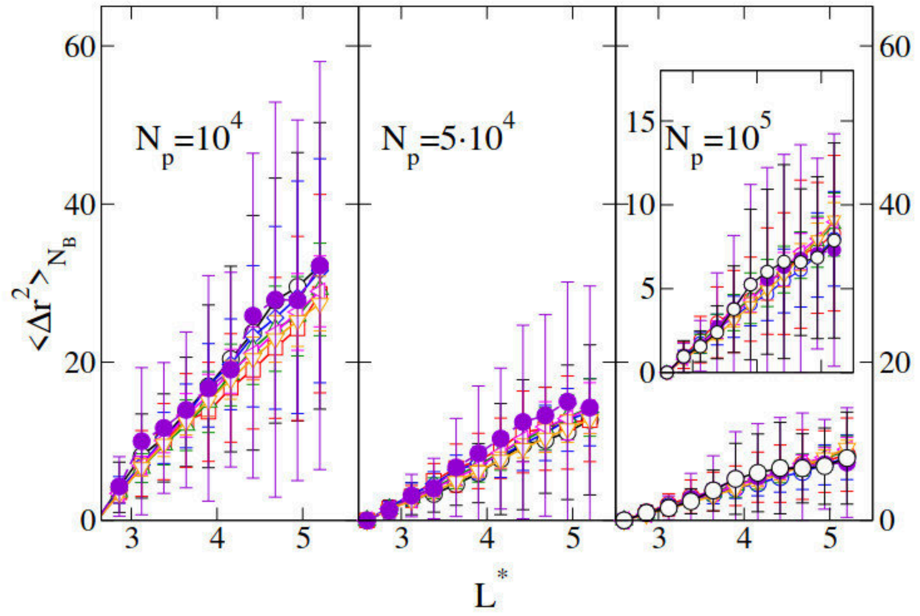
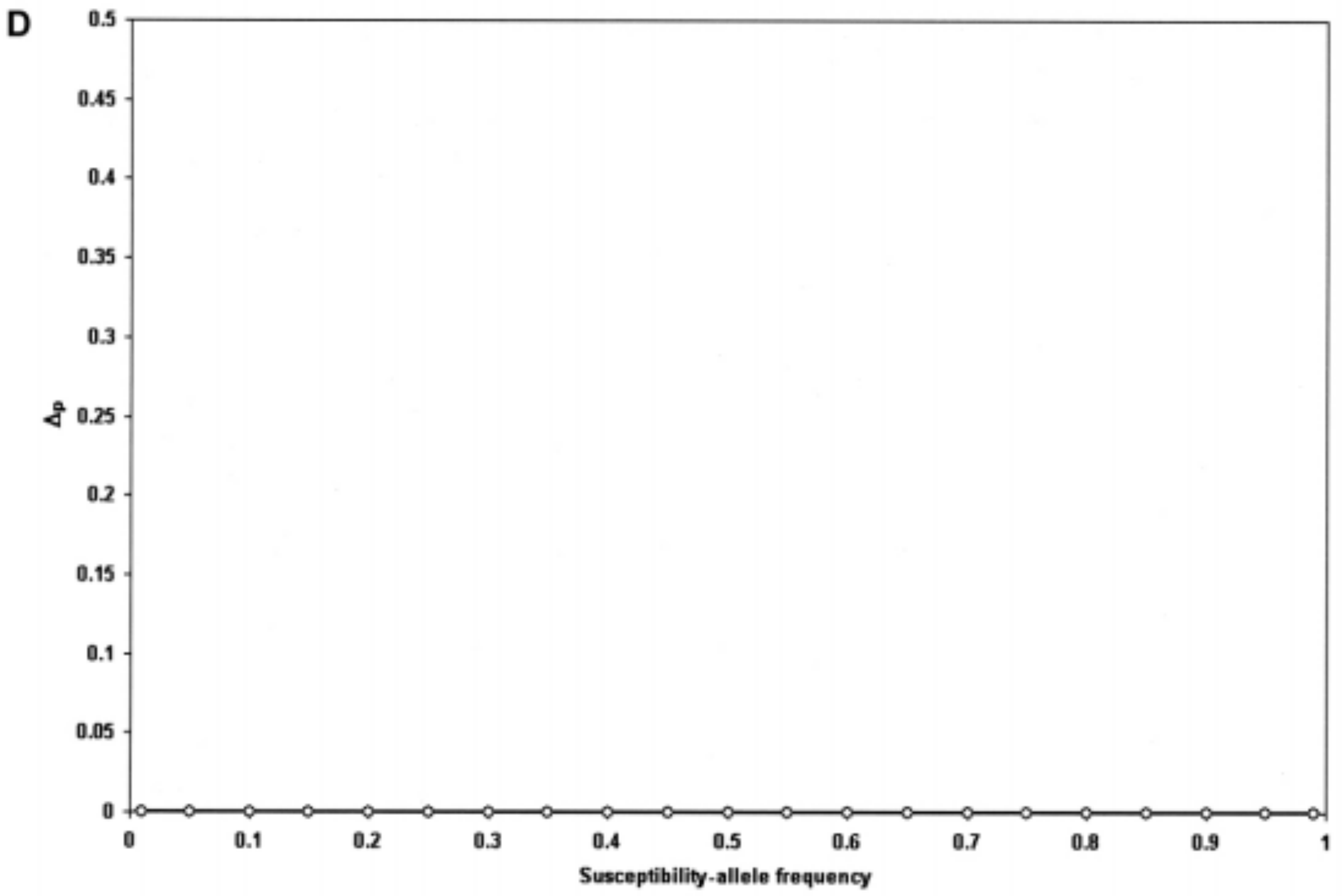
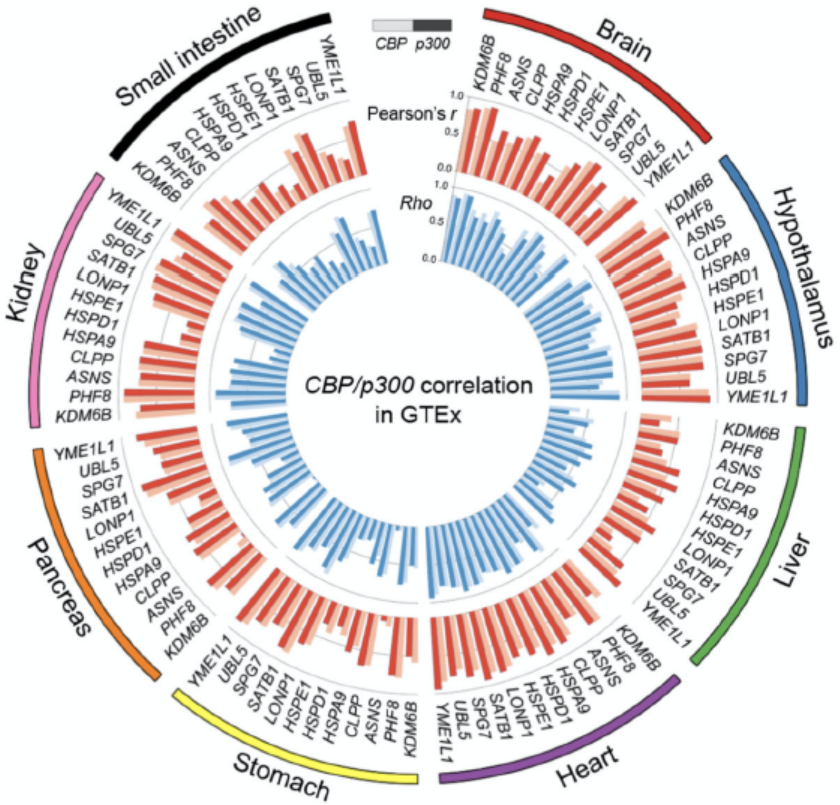


FIG. 6: MSDs as a function of bacteria aspect ratio L^* for colonies containing 1 to 64 bacteria and $N_p = 10^4$ (left), $N_p = 5 \cdot 10^4$ (middle) and $N_p = 10^5$ (right) polymer particles with diameter $\sigma_p = 0.1\sigma$. Symbols refer to colonies containing 1 (●), 2 (○), 4 (□), 8 (◇), 16 (△), 32 (◁) and 64 (▽) bacteria. The inset in the right frame magnifies the panel where it is included. Solid lines are guides for the eye and error bars represent the standard deviation of the mean.

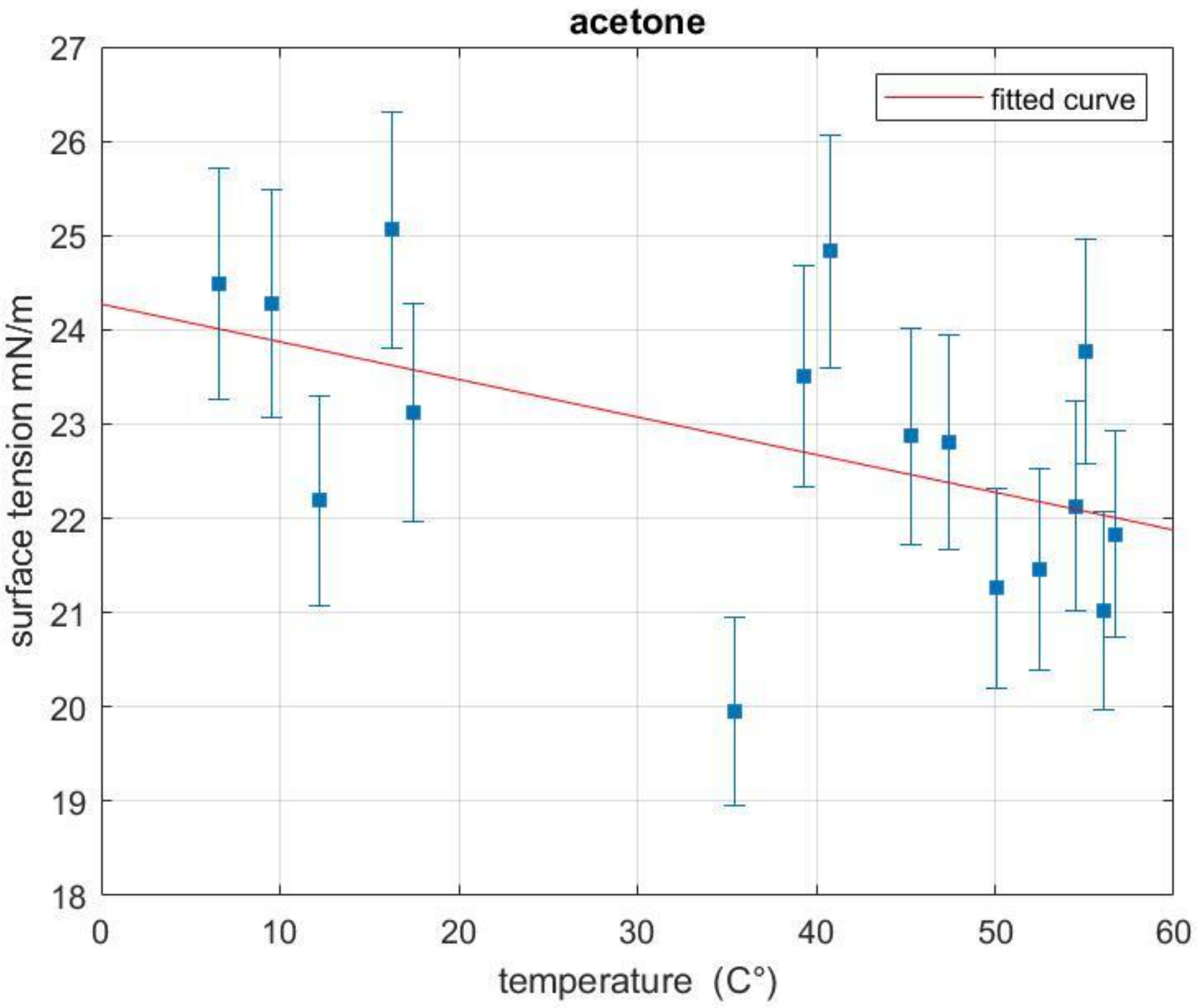


D
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Δp
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Susceptibility-allele frequency

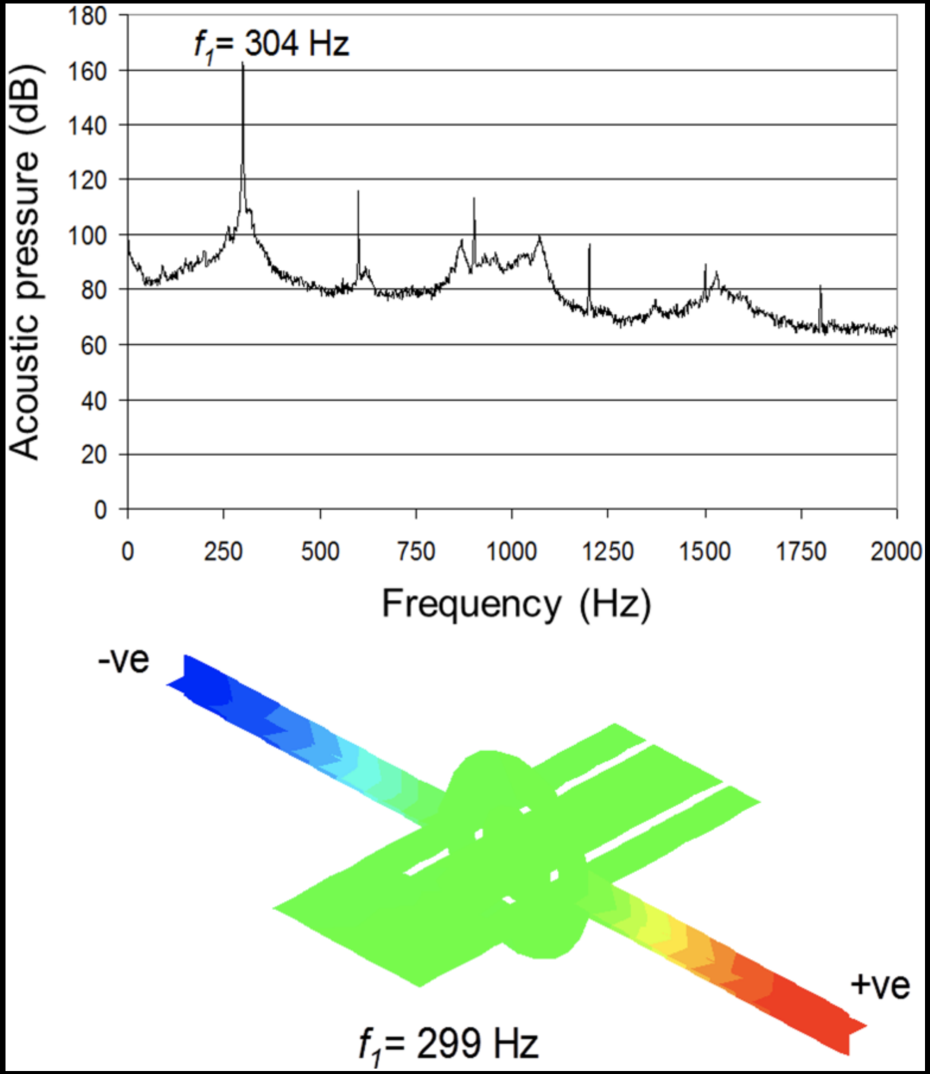
d



CBP
p300
Pearson's r
Rho
1.0
0.5
0.0
CBP/p300 correlation in GTEx
Brain
Hypothalamus
Liver
Heart
Stomach
Pancreas
Kidney
Small intestine
KDM6B
PHF8
ASNS
CLPP
HSPA9
HSPD1
HSPE1
LONP1
SATB1
SPG7
UBL5
YME1L1


acetone
fitted curve
surface tension mN/m
temperature (C°)
27
26
25
24
23
22
21
20
19
18
0
10
20
30
40
50
60

Tufte’s rules:



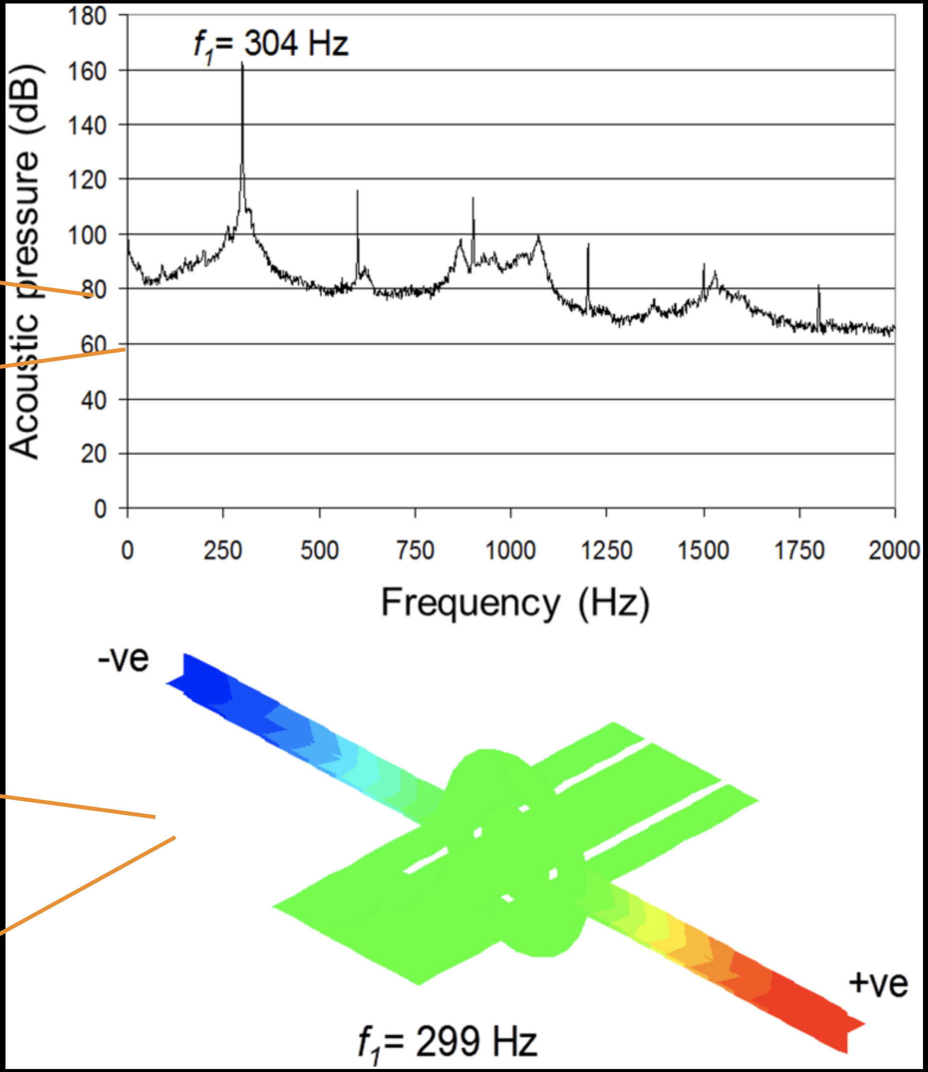
Tufte’s rules:

low data/ink ratio

no comparison

chart junk

2 graphical elements for frequency
(color and position)

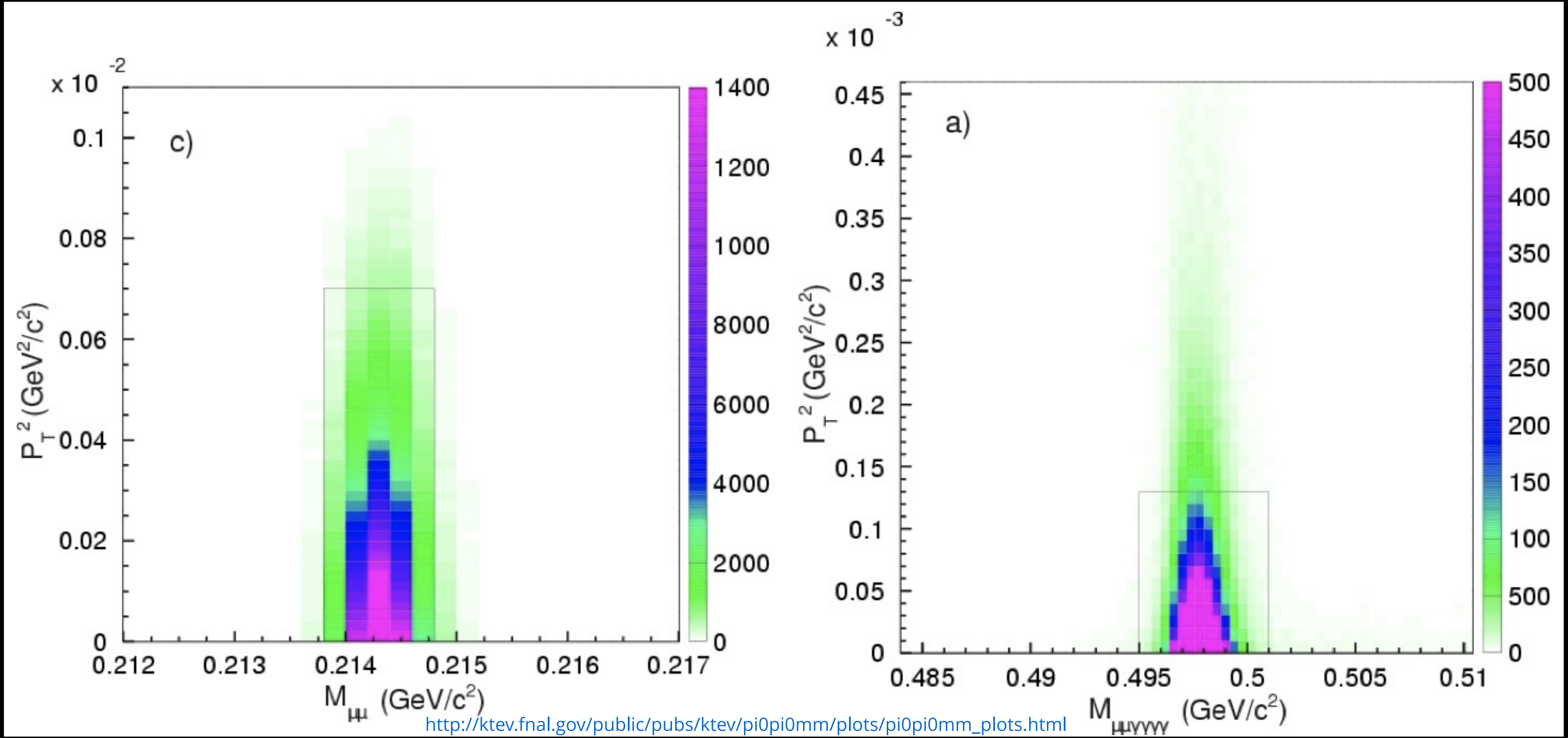


Tufte’s rules:

low data/ink ratio

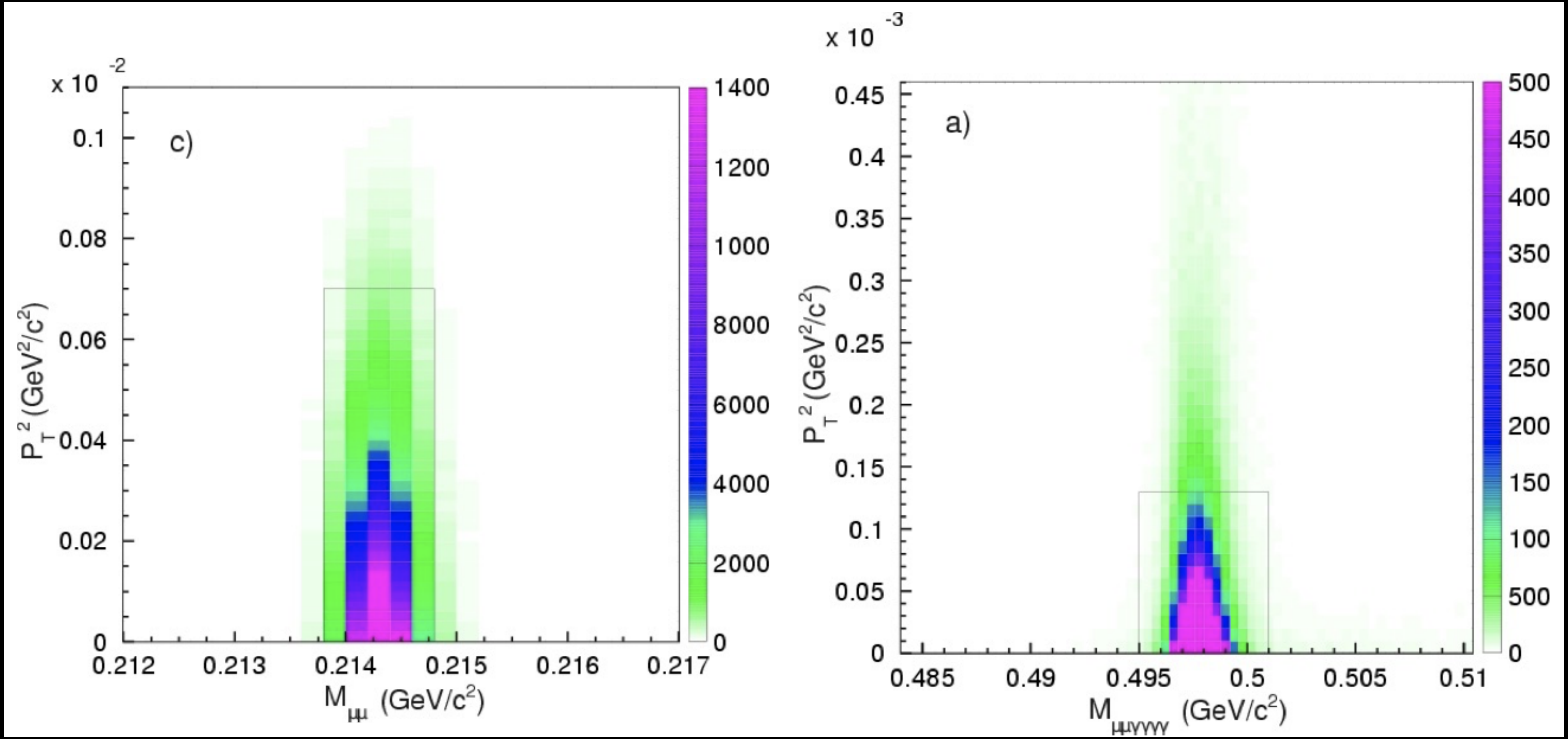
comparison but scale out of context

high effect-size due to the choice of color map (more on this later)



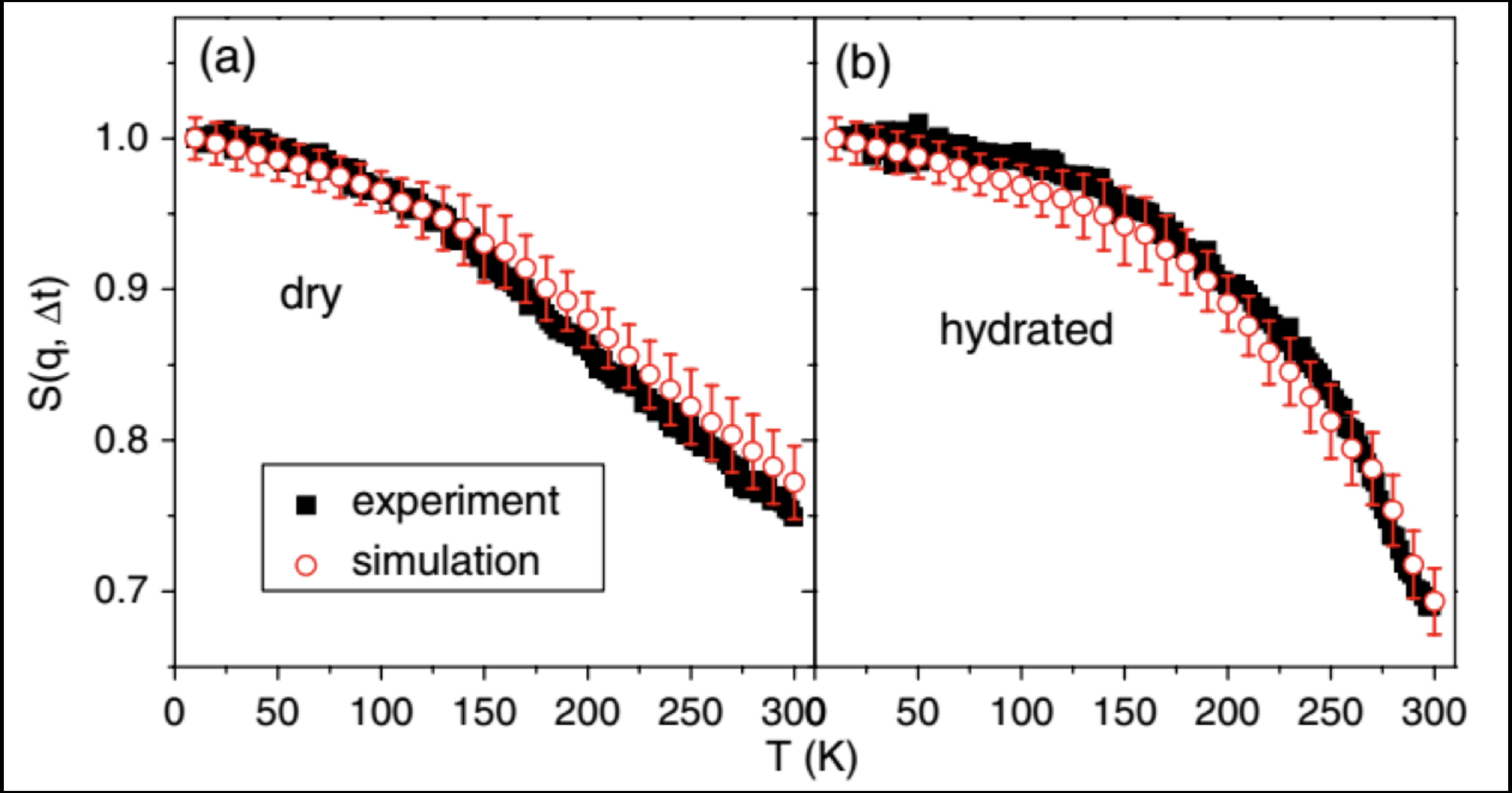
http://ktev.fnal.gov/public/pubs/ktev/pi0pi0mm/plots/pi0pi0mm_plots.html

Tufte's rules:

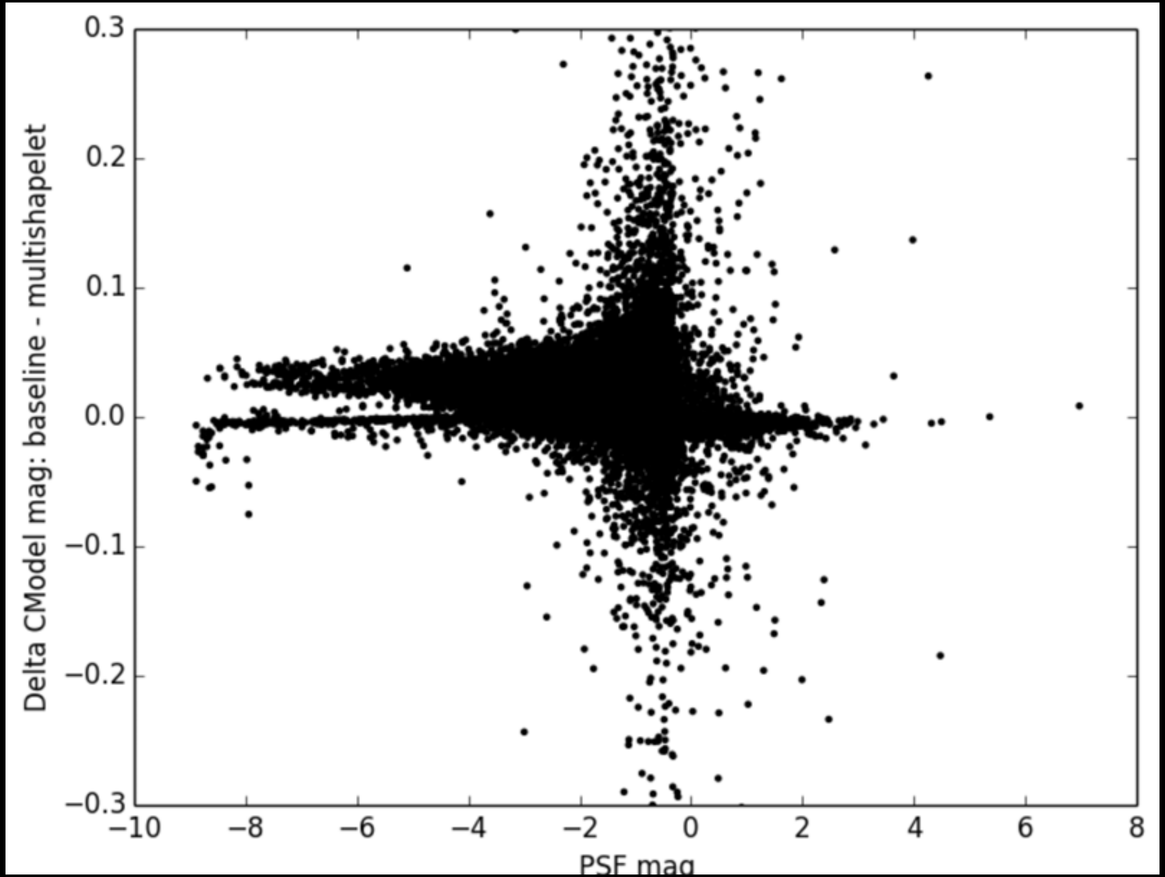


Tufte’s rules:

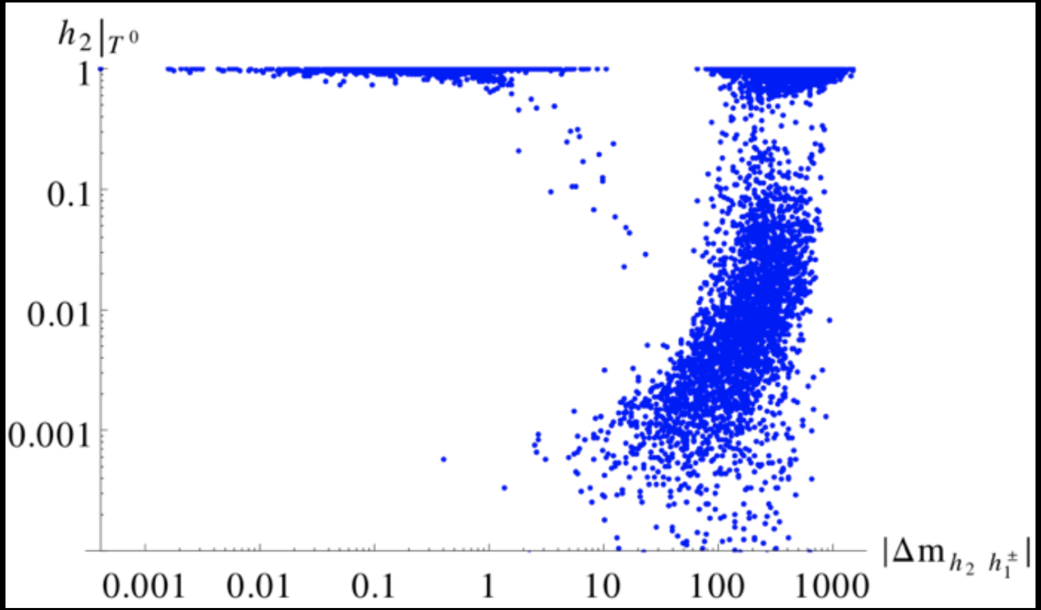
http://ins.sjtu.edu.cn/people/lhong/english/research.html

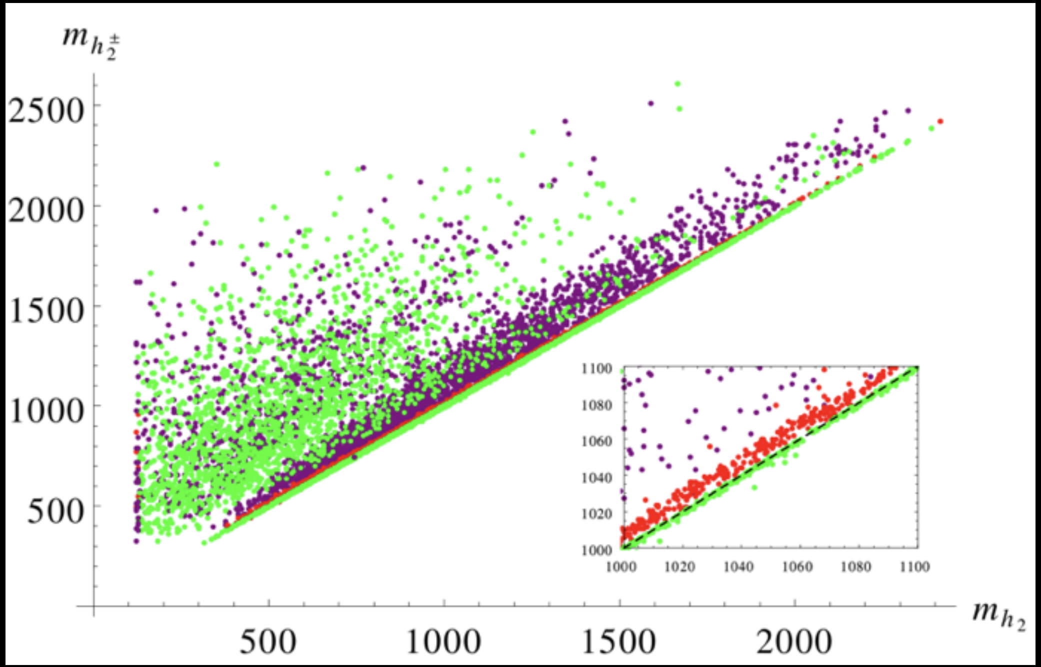


Tufte’s rules:

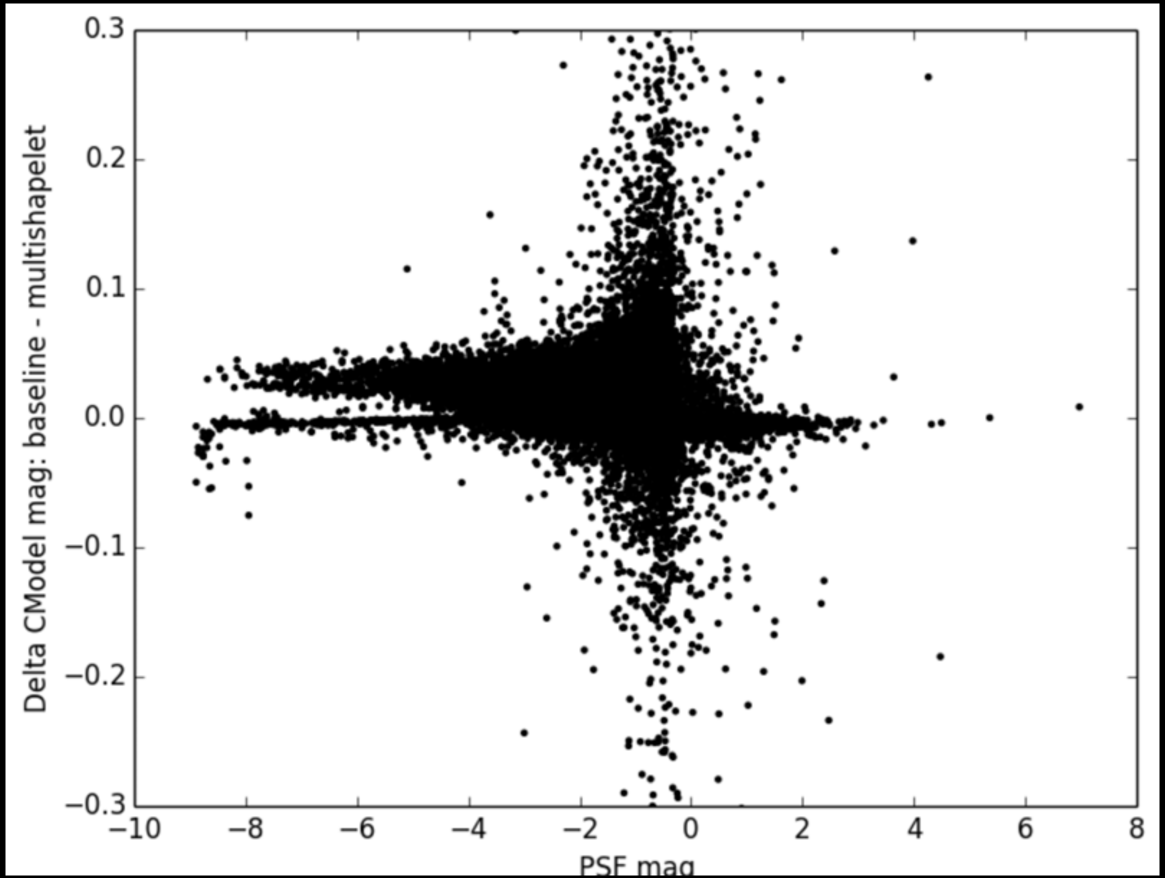


http://inspirehep.net/record/1411674/plots

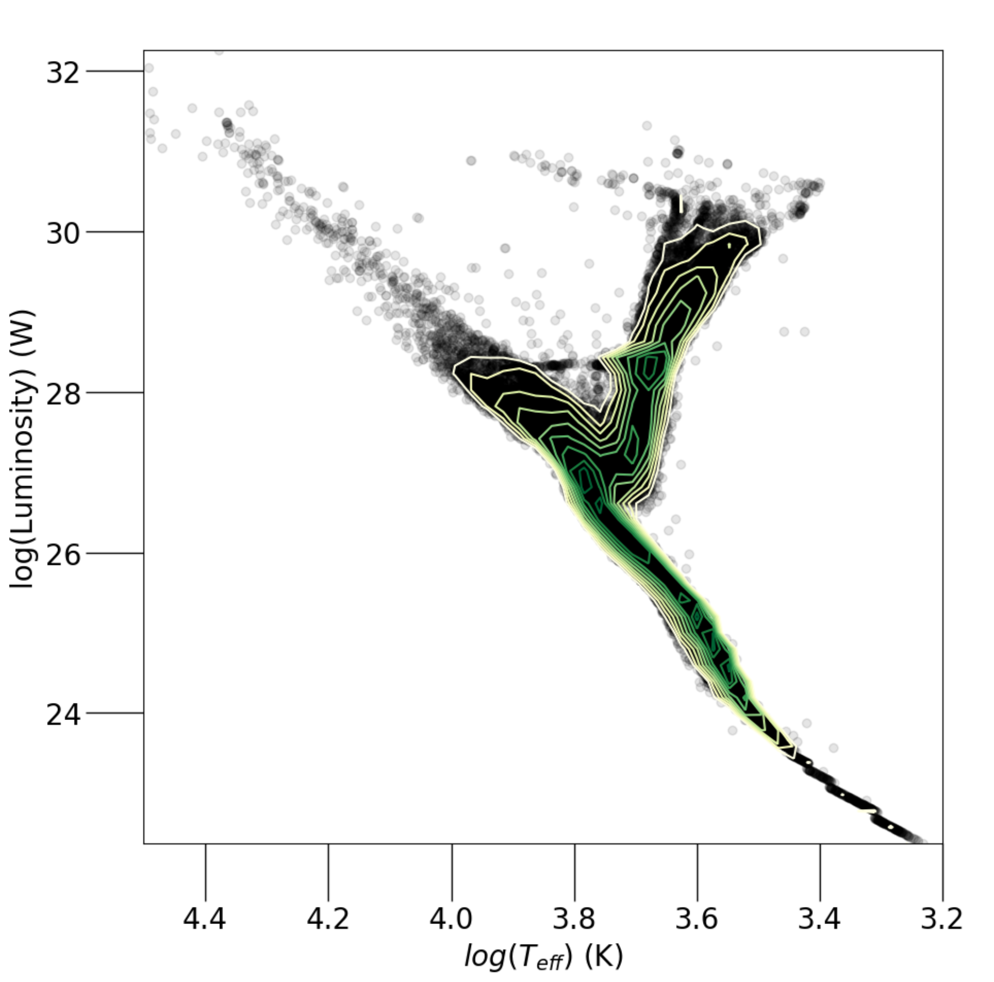




Tufte's rules:



http://inspirehep.net/record/1411674/plots



a better alternative

https://github.com/fedhere/DSPS/blob/master/lab8/badPlotgoodPlot.ipynb

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

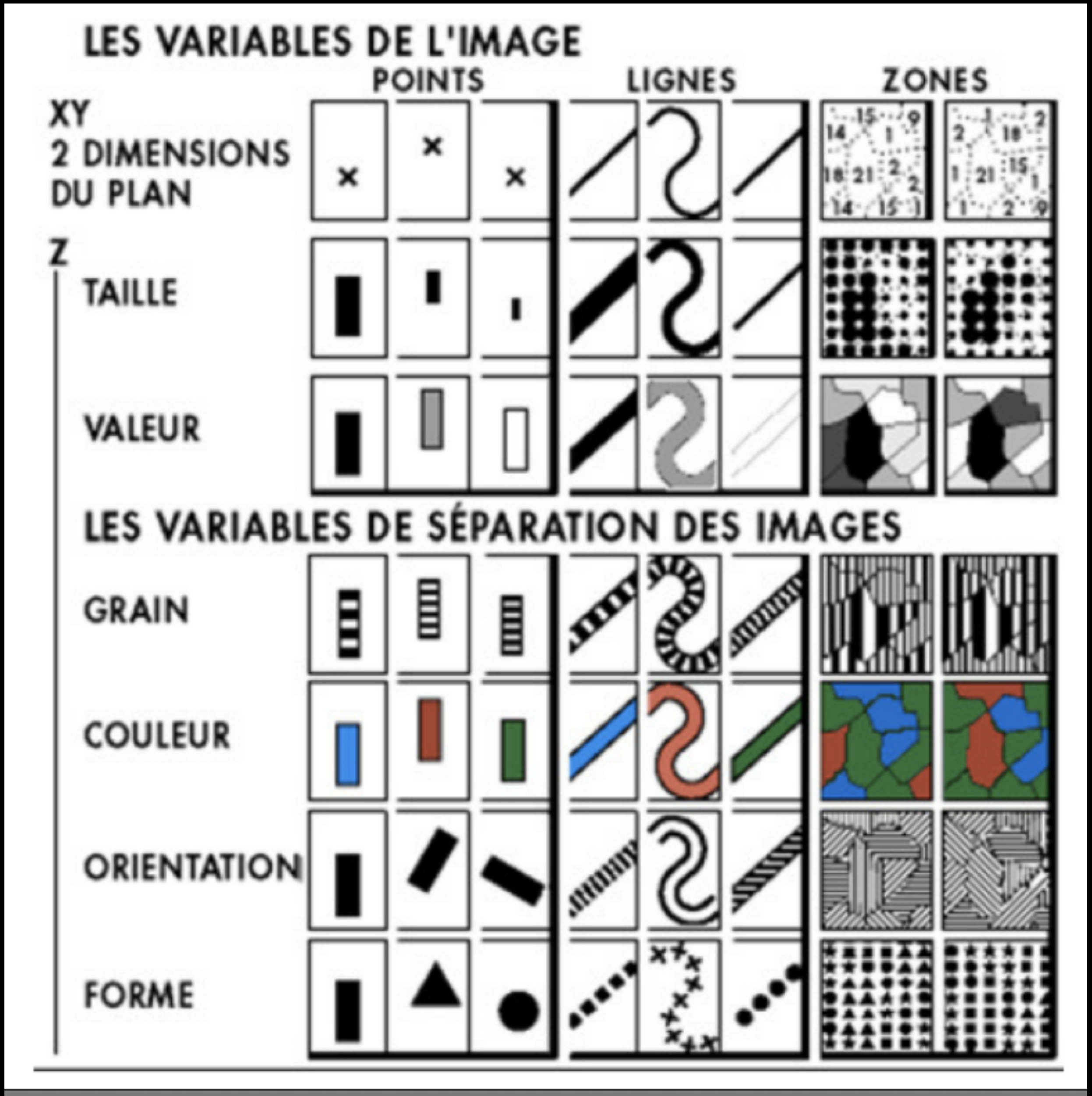
intensity

texture

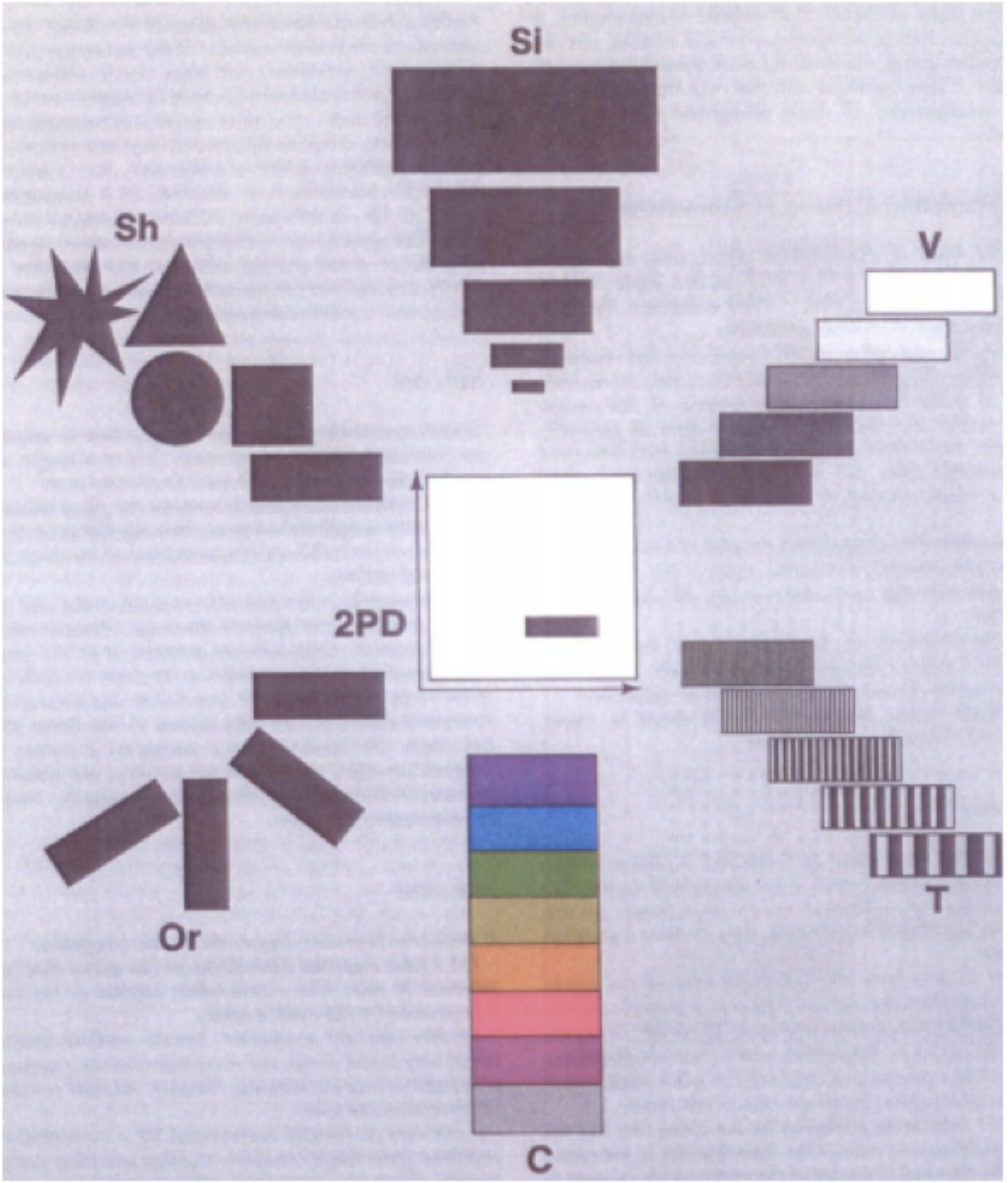
color

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 star (can take any value)**

Continuous data may be:

- **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**

Discrete data may be:

- **Counts:** number of bacteria at time *t* in section *A*
- **Ordinal:** survey response Good/Fair/Poor

- **Categorical: fermion - bosons: any object by class**
  -

Data may also be:

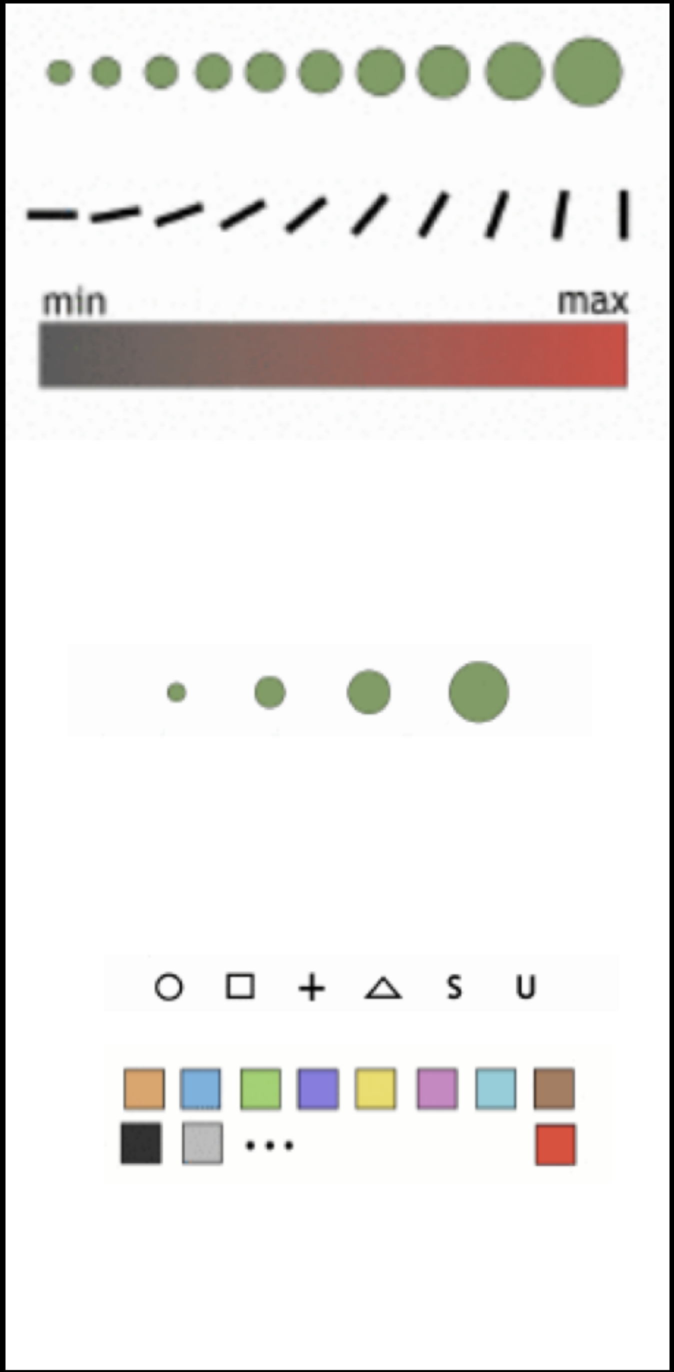
- **Censored:** star mass >30 Msun
- **Missing:** “Prefer not to answer” (NA / NaN)

data types

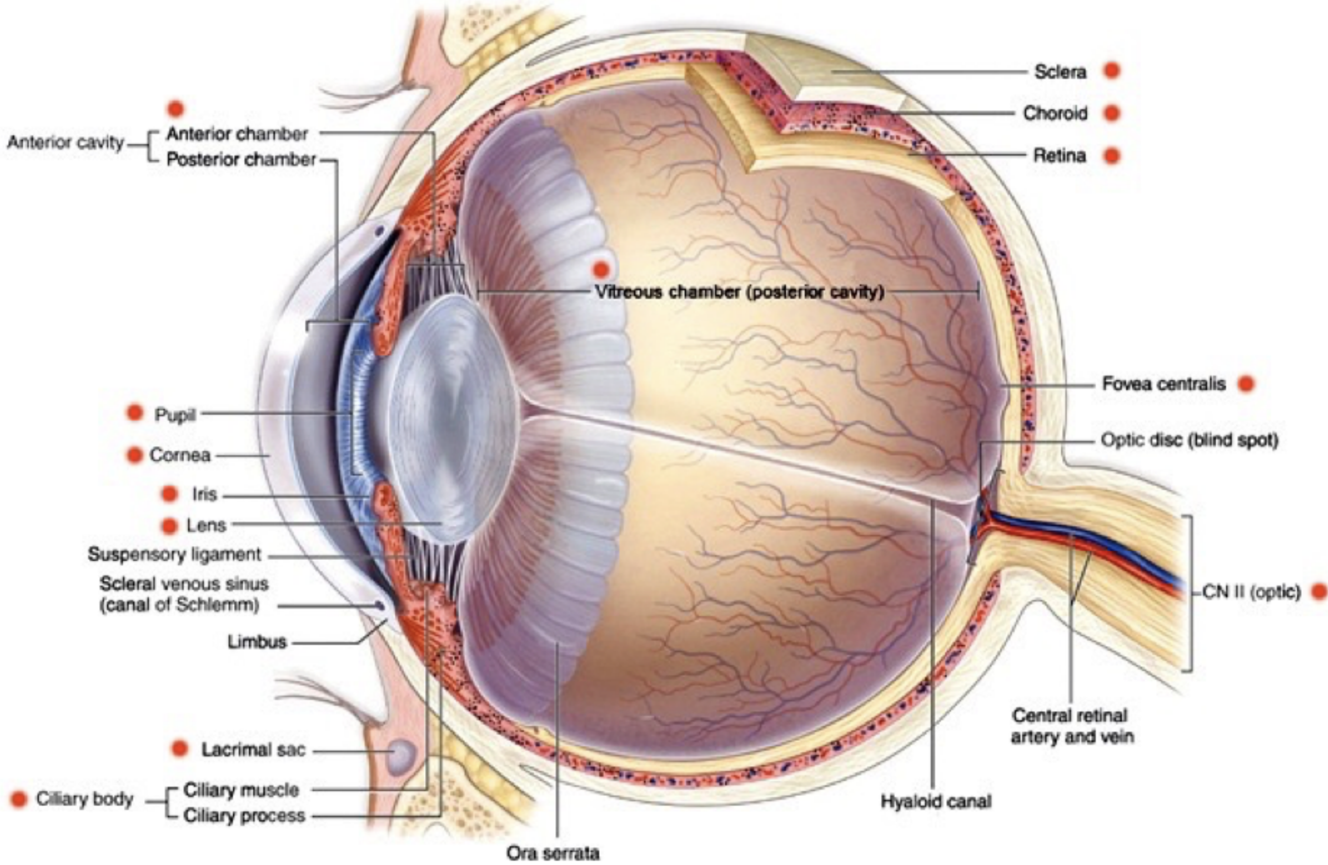
continuous

ordered

categorical



graphical elements work differently on different data types



psychophysics

The study of human perception

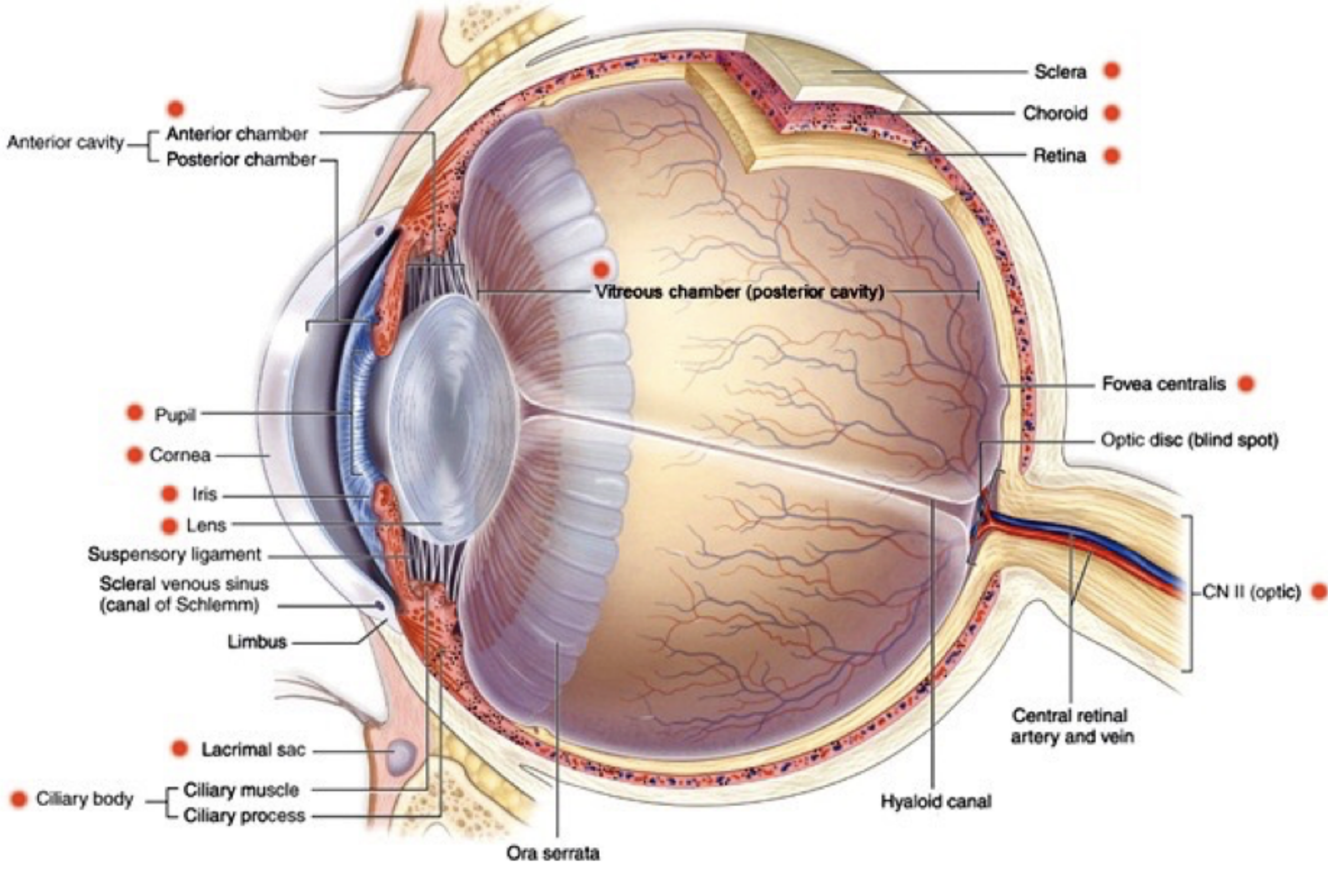
why do we visualize? cause our brain is best suited to receive visual sitmuly (compared to tactile or auditorial for example)

alternatives:

sonification
https://en.wikipedia.org/wiki/Sonification

tactile maps

https://www.nasa.gov/audience/foreducators/a-feel-for-astronomy.html



psychophysics

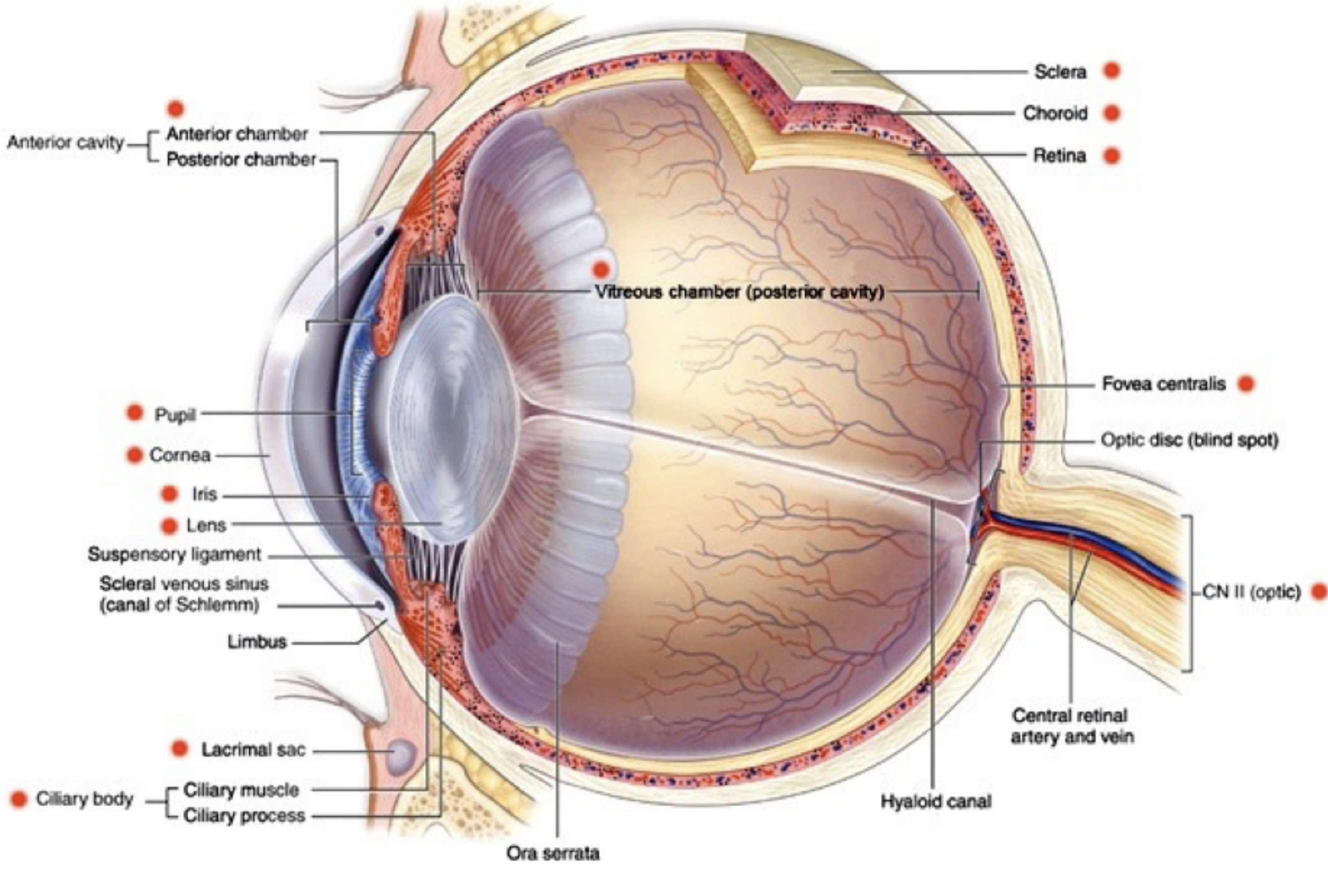
why do we visualize? cause our brain is best suited to receive visual sitmuly (compared to tactile or auditorial for example)

alternatives:

sonification
https://en.wikipedia.org/wiki/Sonification

tactile maps

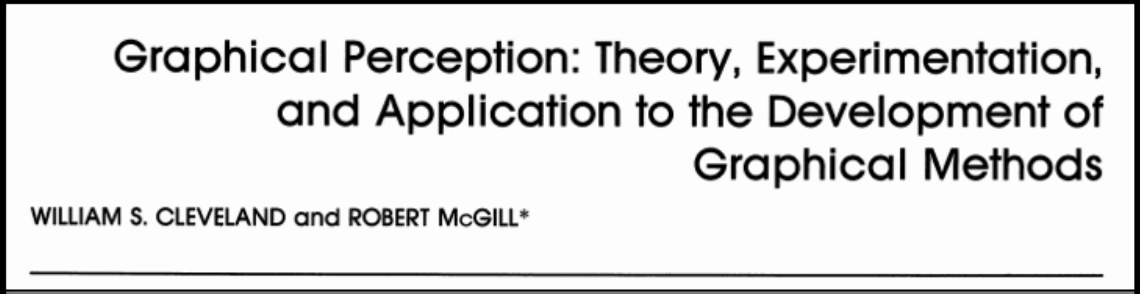
https://www.nasa.gov/audience/foreducators/a-feel-for-astronomy.html



psychophysics

exploring these alternative is important as a social justice issue to make science and data science accessible to vision impaired people but also: can the different ways in which we process information can give new insight?

e.g.: we process visual stimuli but sound stimuli holistically (we hear a chord, not the notes in it)


Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

psychophysics

The study of human perception
here limited to visual stimuli

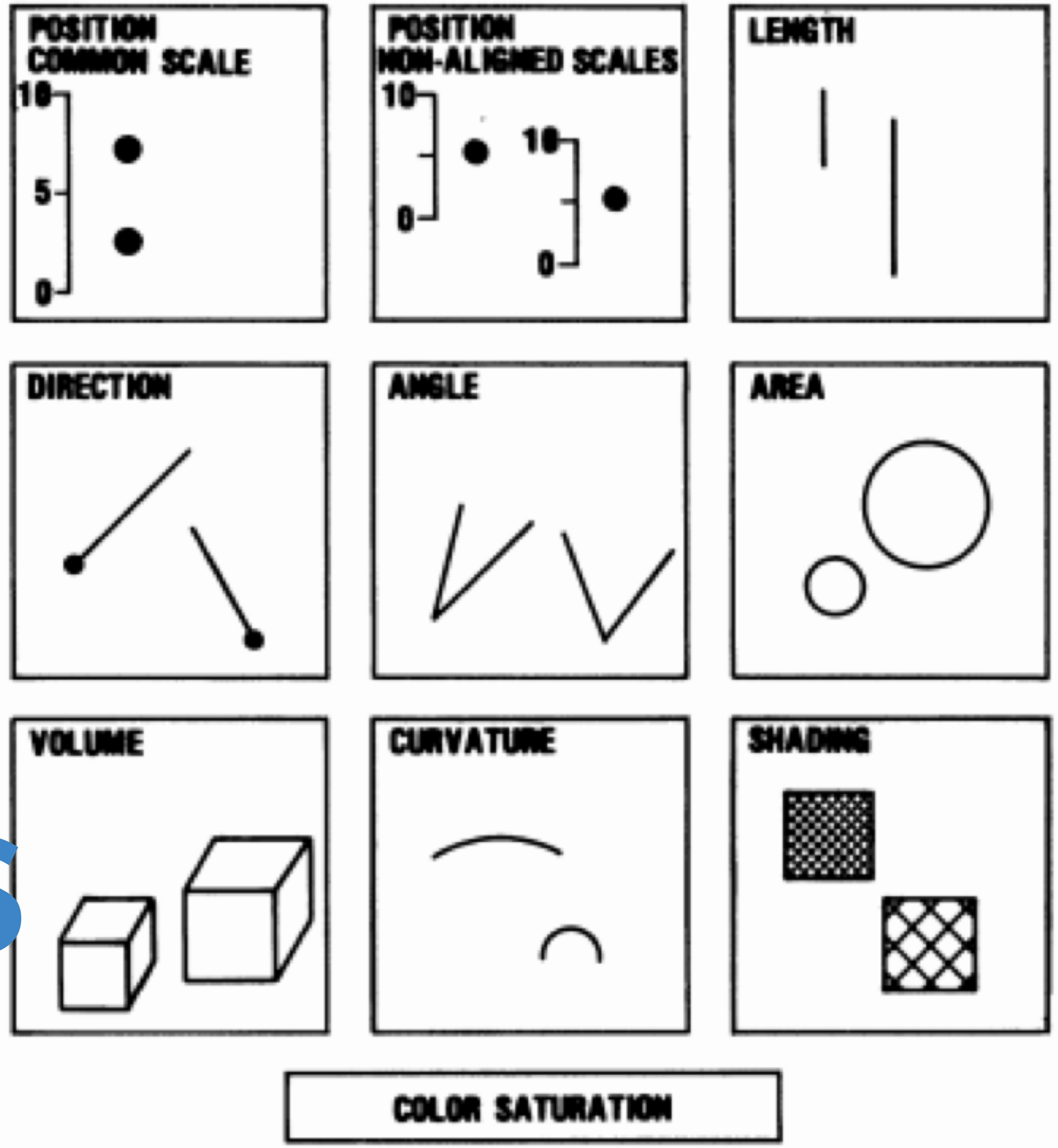


Figure 1. Elementary perceptual tasks.

Stevens 1975

Psychophysical power law

The apparent magnitude of all sensory channels follows a power law based on the stimulus intensity

$$S = I^n$$

S sensation, I intensity

Stevens 1975

response to length: $I = S$

when shown something 4x as long we perceive it as being 4x as long

response to brightness: $I = \sqrt{S}$

when shown something 4x as bright we perceive it as being 2x as bright

response to saturation: $I = S^{1.7}$

when shown something 4x as saturated we perceive it as being 11x as saturated

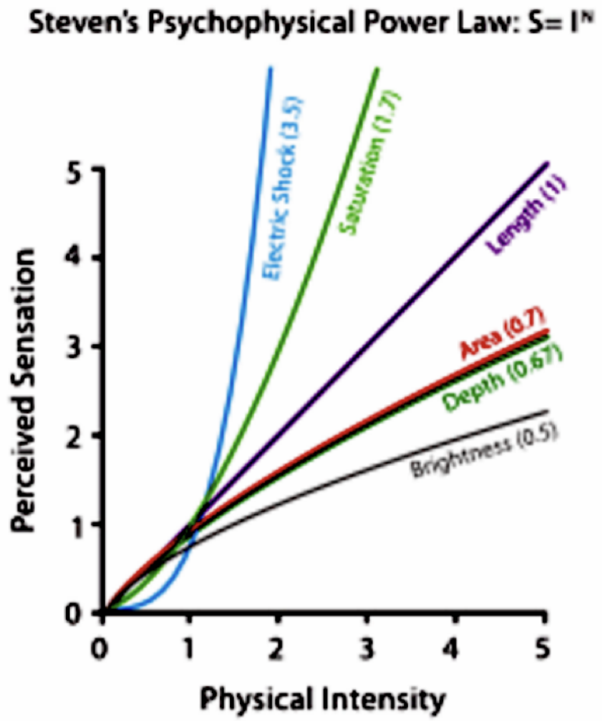


Figure 5.7. Stevens showed that the apparent magnitude of all sensory channels follows a power law $S = I^n$, where some sensations are perceptually magnified compared with their objective intensity (when $n > 1$) and some compressed (when $n < 1$). Length perception is completely accurate, whereas area is compressed and saturation is magnified. Data from Stevens [Stevens 75, p. 15].

Stevens 1975

response to electroshock:$I = S^{3.5}$

when given an electroshock 4x as strong

we perceive it as 128x as strong

(personally, I do not know of any electroshock based visualizations)

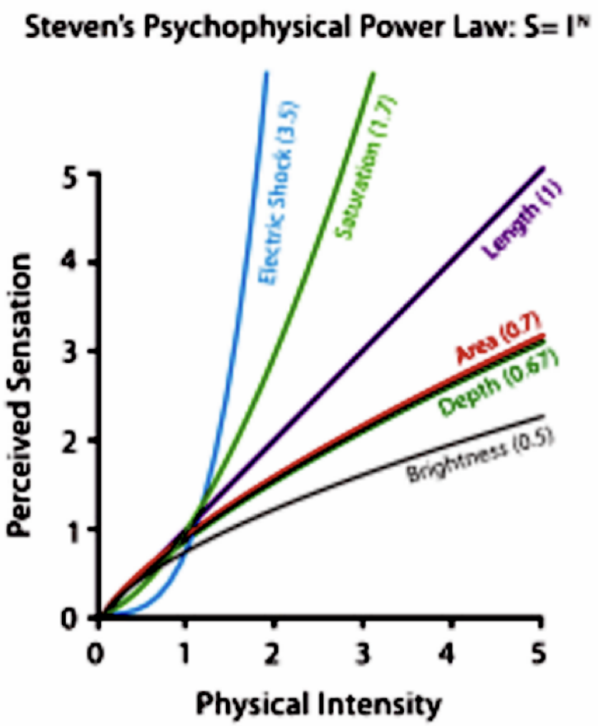
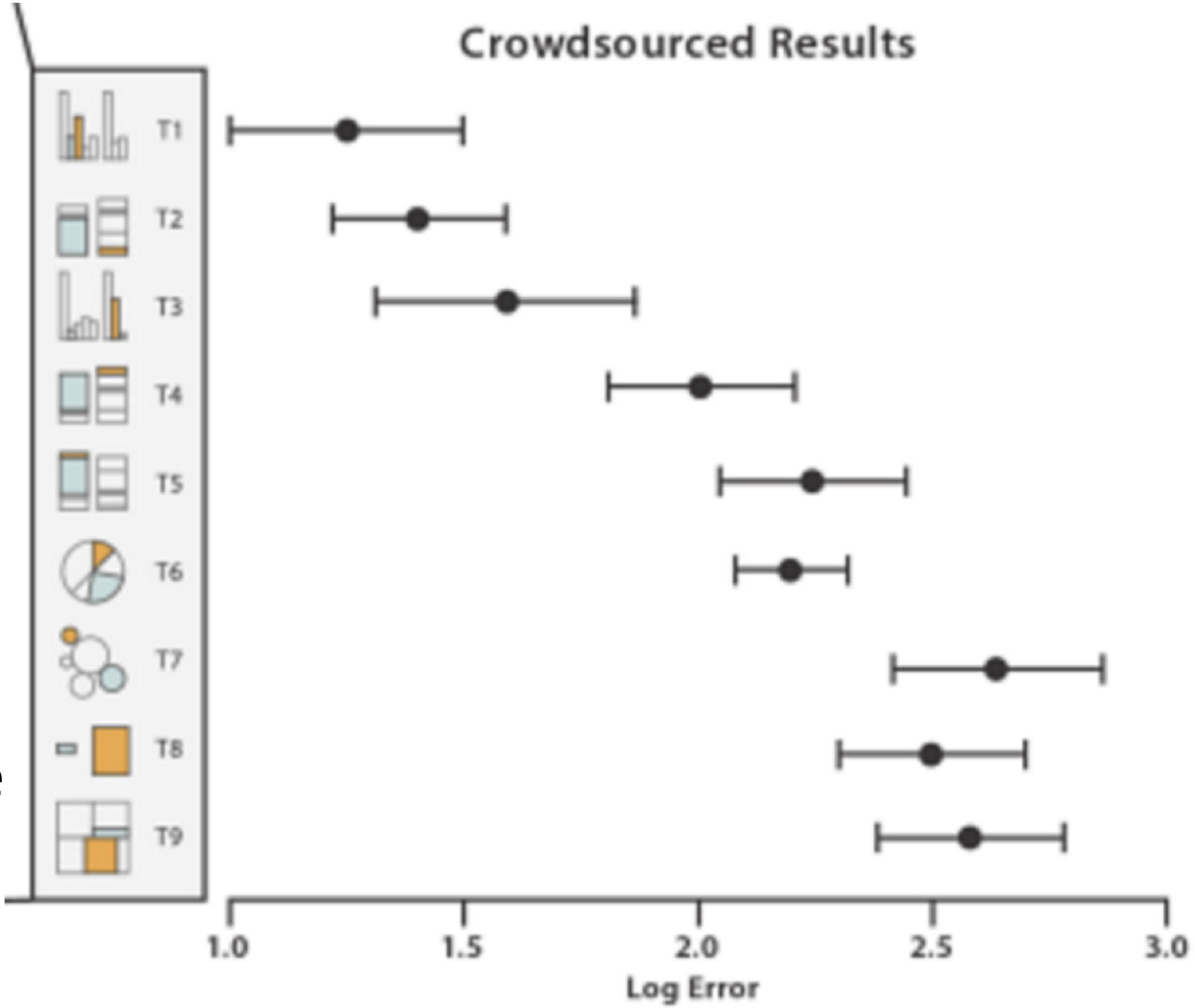
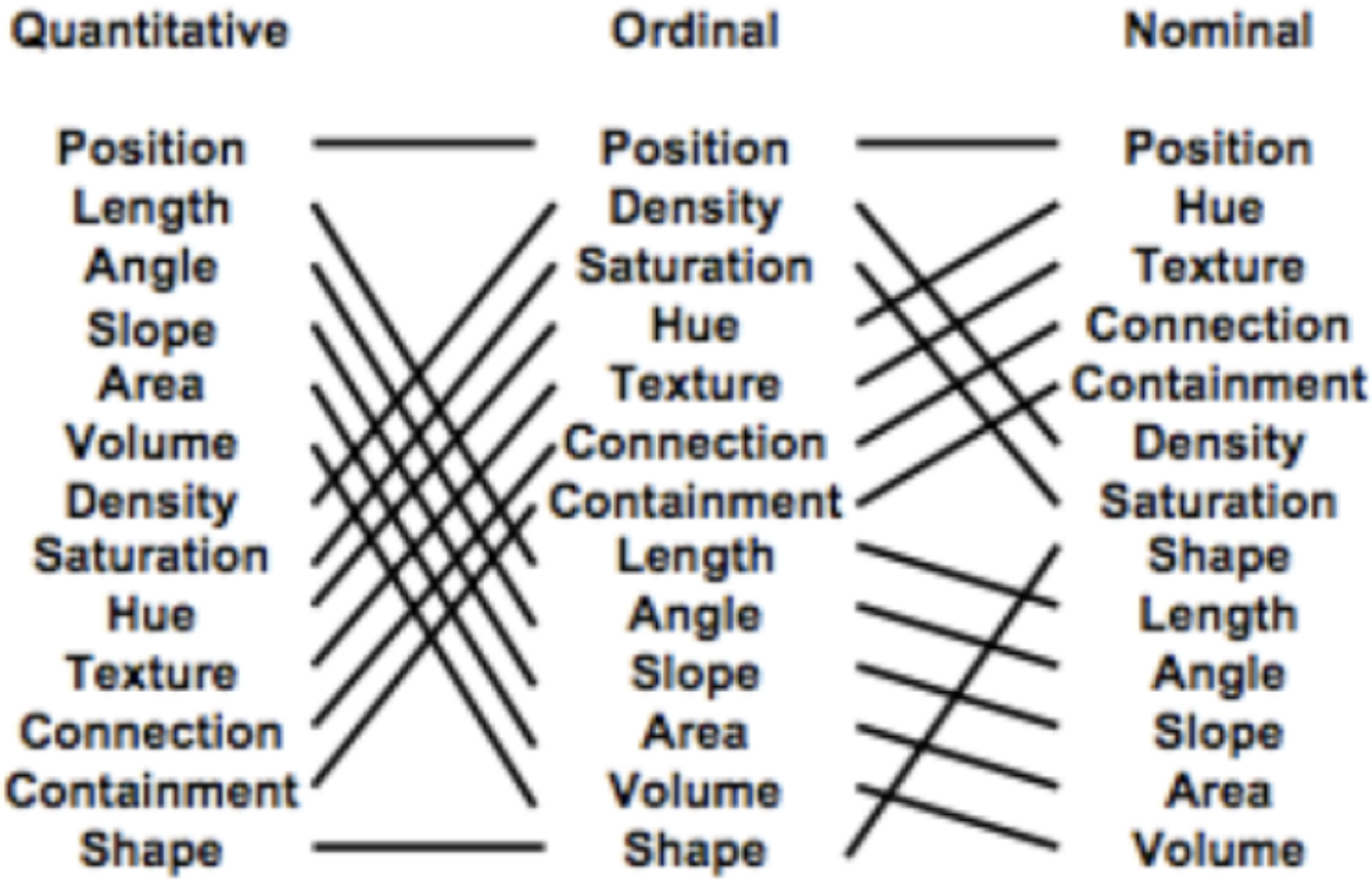


Figure 5.7. Stevens showed that the apparent magnitude of all sensory channels follows a power law $S = I^n$, where some sensations are perceptually magnified compared with their objective intensity (when $n > 1$) and some compressed (when $n < 1$). Length perception is completely accurate, whereas area is compressed and saturation is magnified. Data from Stevens [Stevens 75, p. 15].

Heer and Bostock 2010

modern version gets uncertainties to these quantities by crowdsourcing the tests





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

Weber law

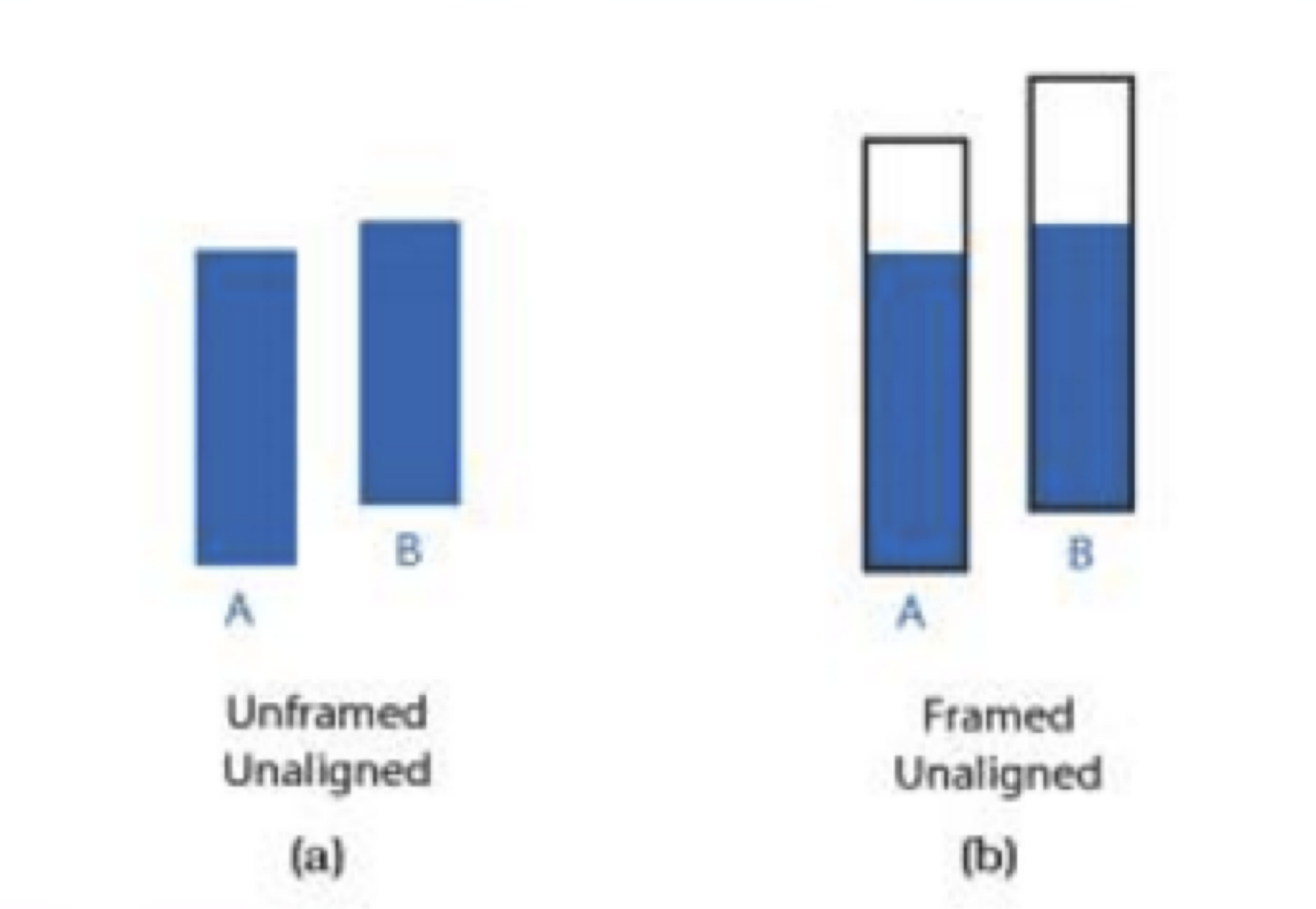
We judge based on relative differences

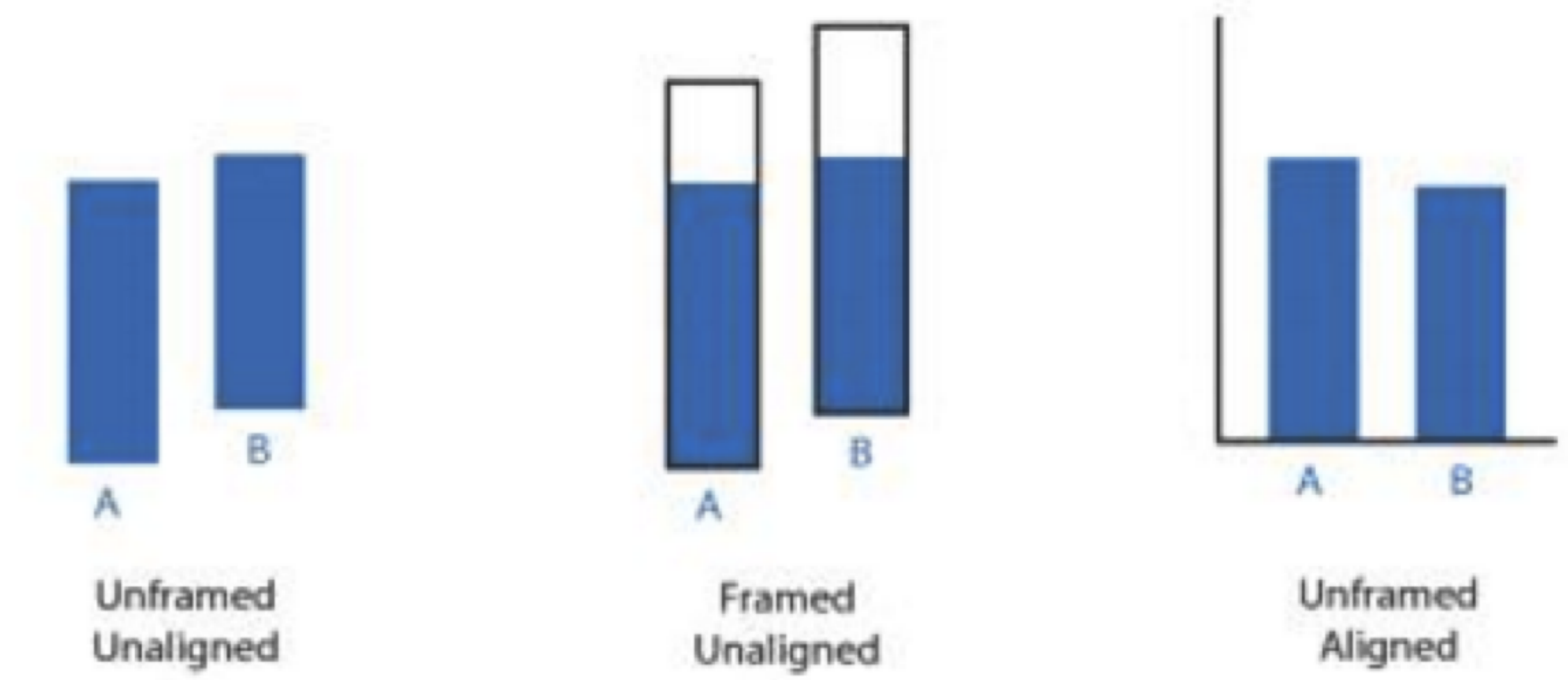
The detectable difference in stimulus intensity is a fixed percentage of the object magnitude

$$\delta I / I = K$$

I intensity, K constant

(a)


A
B
Unframed
Unaligned
(a)
A
B
Framed
Unaligned
(b)

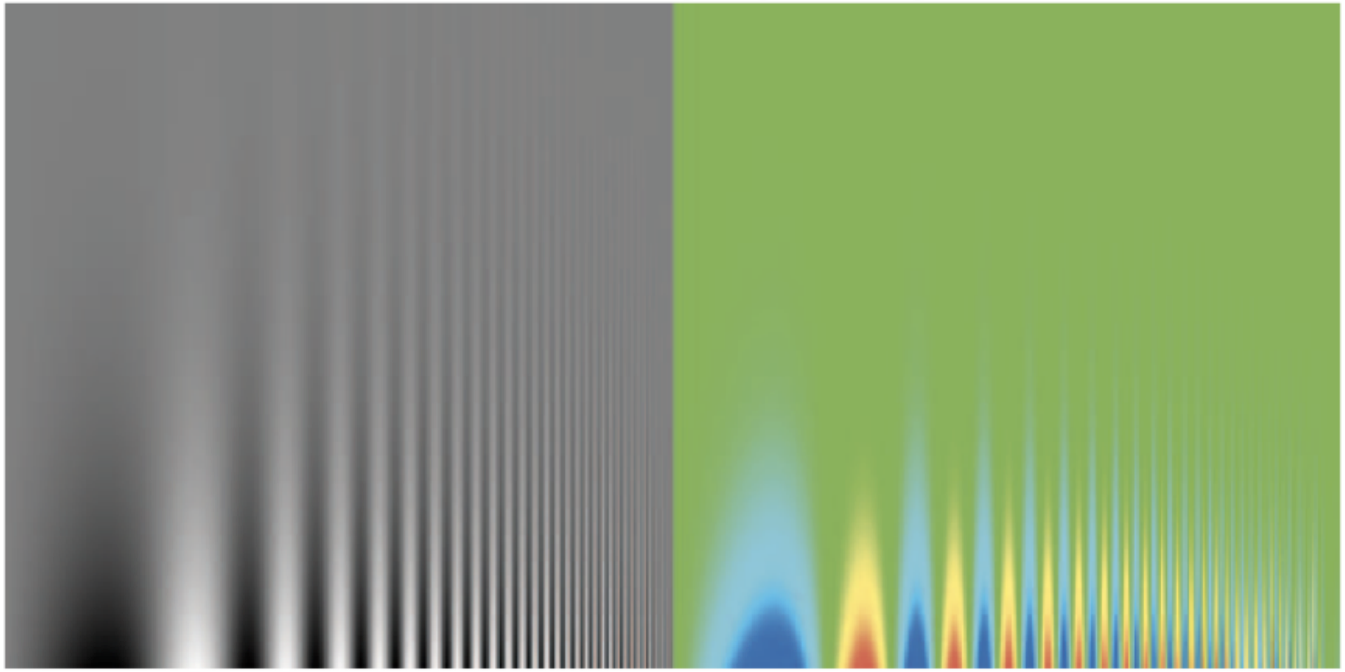

A
B
Unframed
Unaligned
A
B
Framed
Unaligned
A
B
Unframed
Aligned
(a) (b) (c)

Color theory

(and good practice)

Good and Bad color choices

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)

http://blog.visual.ly/rainbow-color-scales/

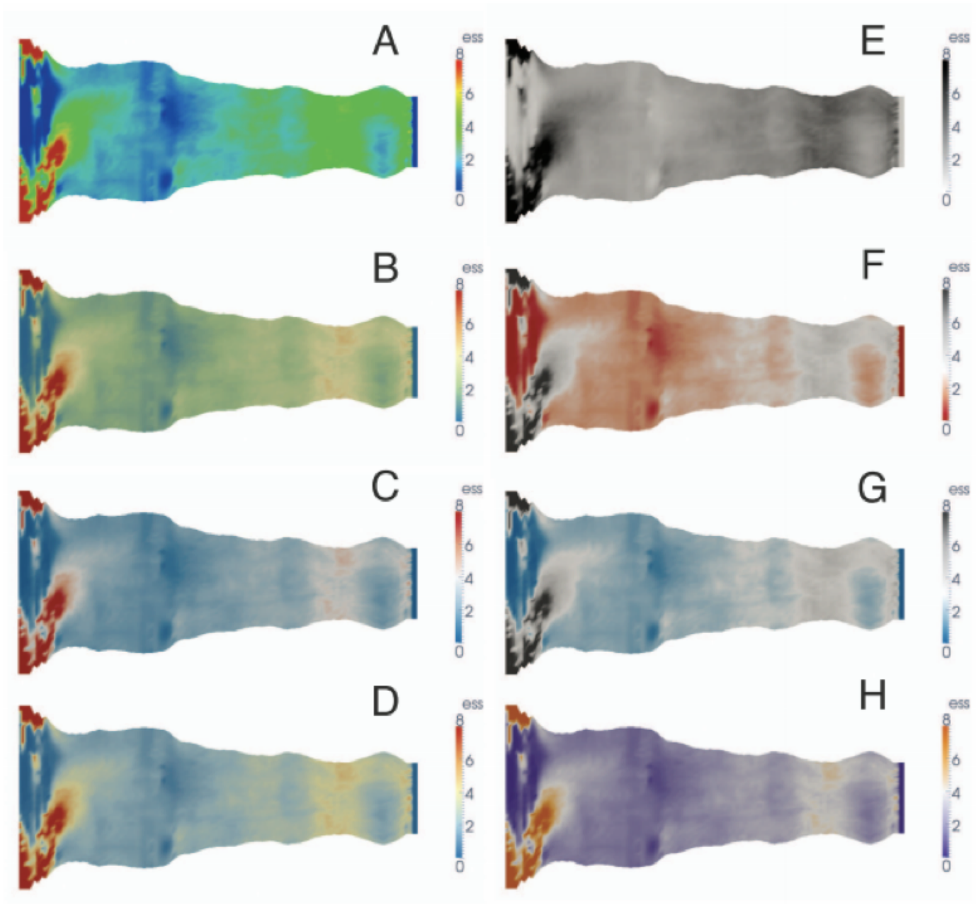


Fig. 4. Color schemes presented during the qualitative user study. The rainbow scheme (A) was preferred by most since it is what they are accustomed to viewing. The next most popular scheme was the red-black diverging scale (F). The grayscale image (E) was unanimously disliked since participants assume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

very real consequences of bad color choices

Borkin et al. 2011

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

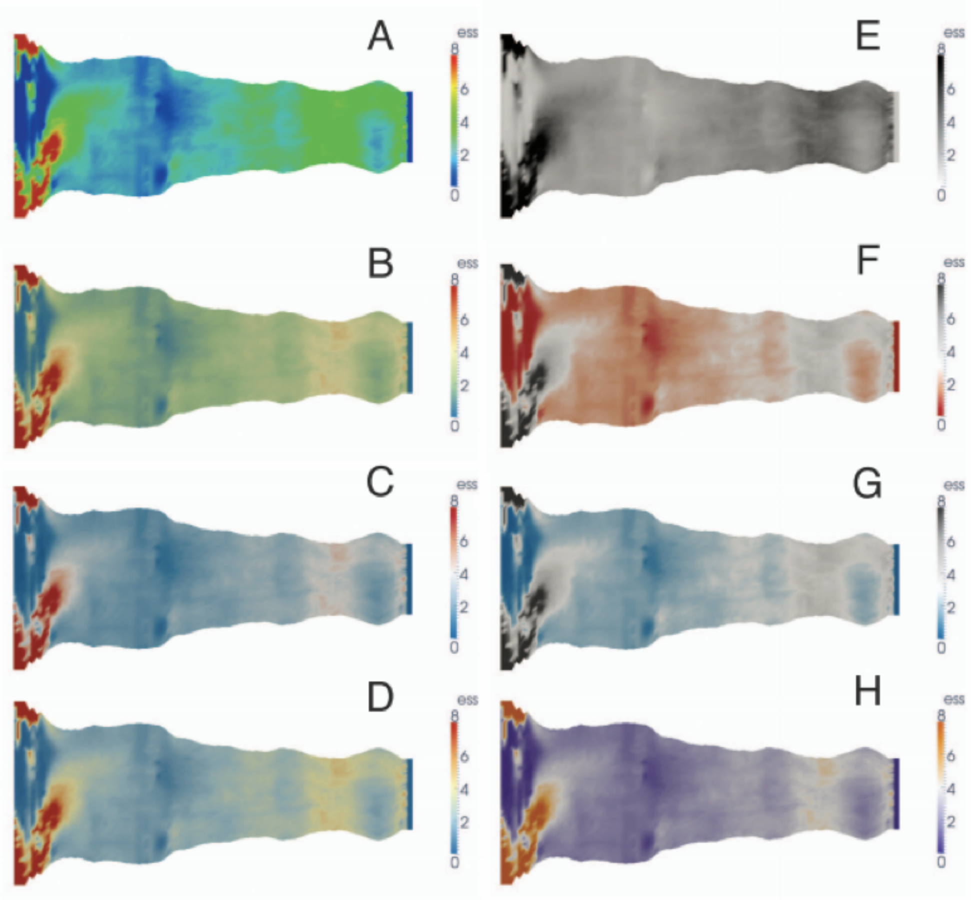


Fig. 4. Color schemes presented during the qualitative user study. The rainbow scheme (A) was preferred by most since it is what they are accustomed to viewing. The next most popular scheme was the red-black diverging scale (F). The grayscale image (E) was unanimously disliked since participants assume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

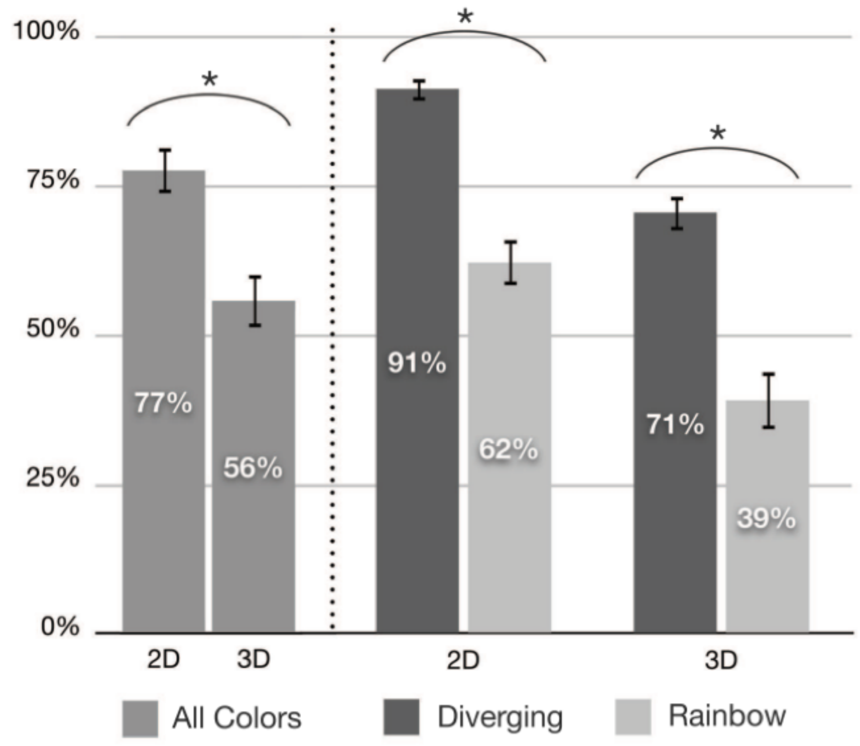


Fig. 7. **Average percent of low ESS regions identified** broken down by 2D and 3D representation, and color. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. Participants were more accurate in 2D and when using the diverging color map.

Borkin et al. 2011

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

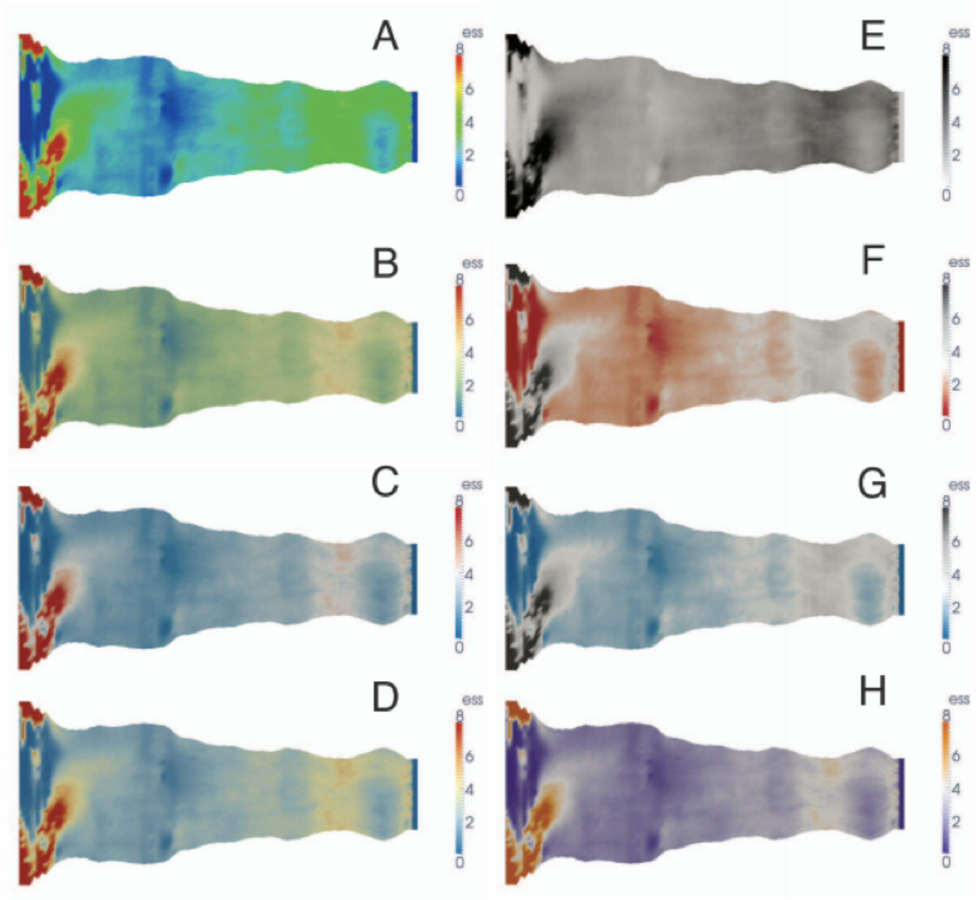
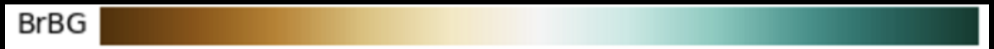


Fig. 4. Color schemes presented during the qualitative user study. The rainbow scheme (A) was preferred by most since it is what they are accustomed to viewing. The next most popular scheme was the red-black diverging scale (F). The grayscale image (E) was unanimously disliked since participants assume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

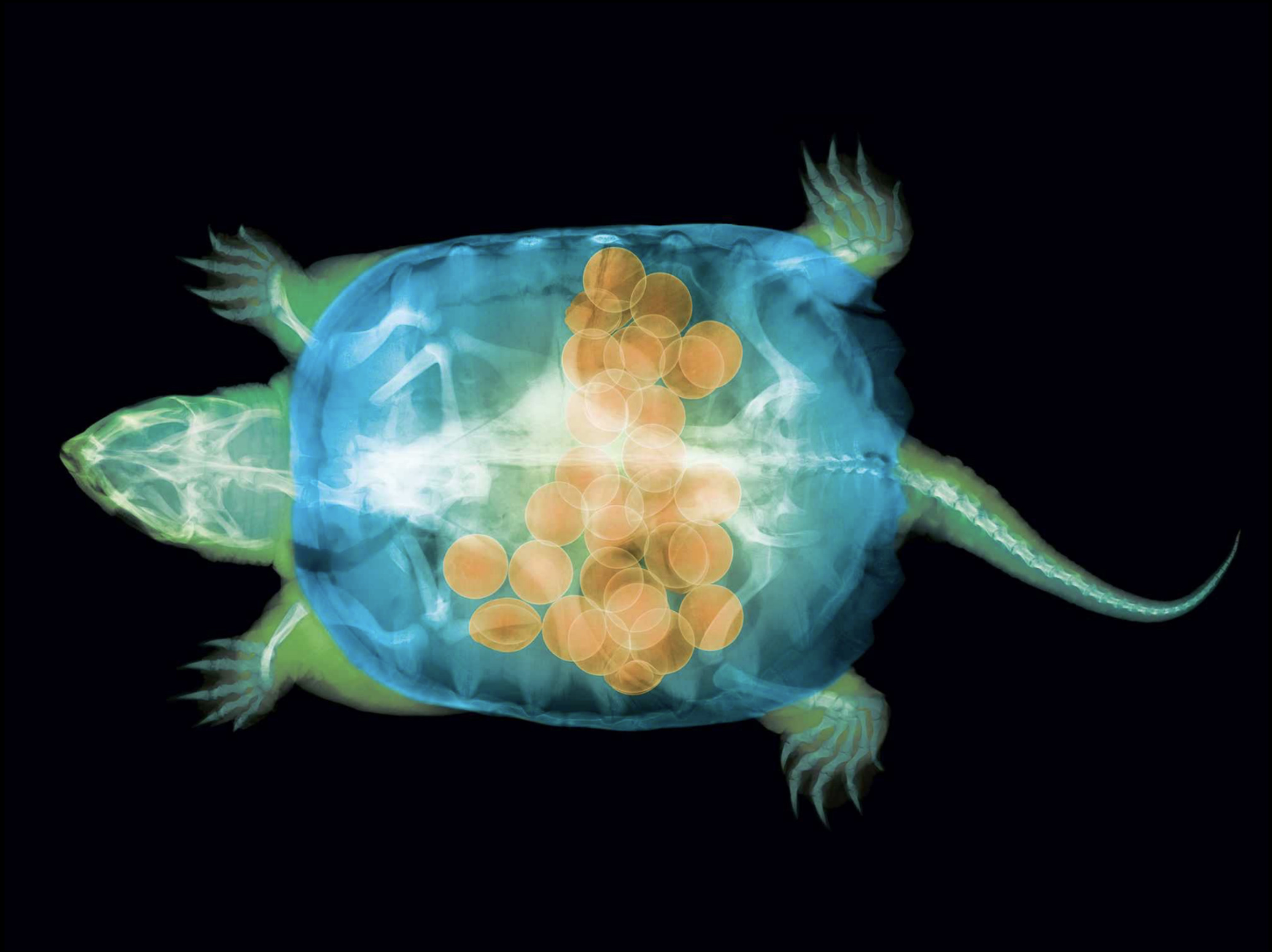
1) Never use Rainbow

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



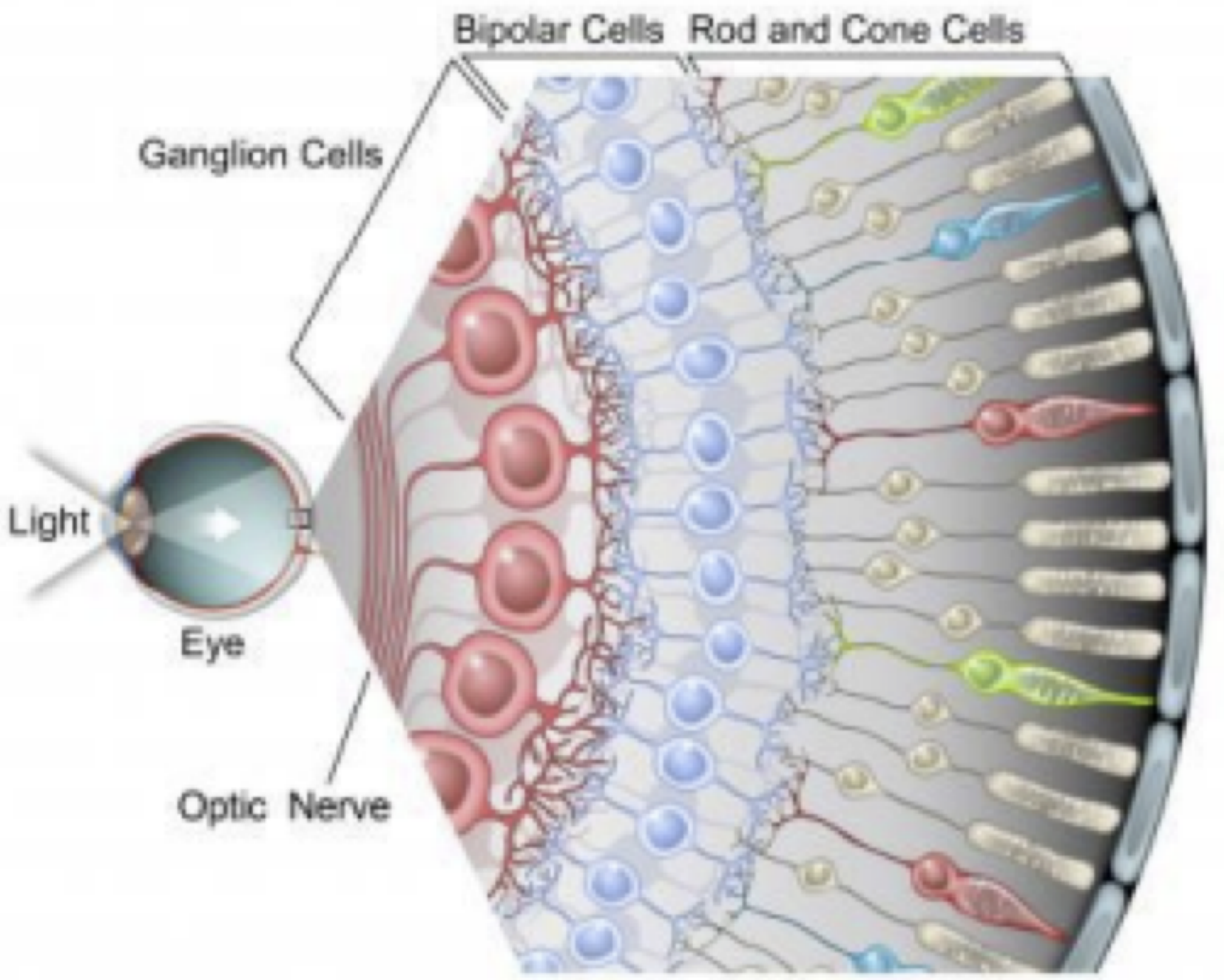
3) Choose a *perceptually uniform* color map for continuous data that does not have a focal point (a special point inside the range)

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



http://www.popsci.com/2015-vizzies-science-visualizations-video-images?image=0

Eye Physiology and color perception deficiencies



color blindness

Color blindness (color vision deficiency, CVD) affects approximately

1 in 12 men (8%) and 1 in 200 women

in the world.

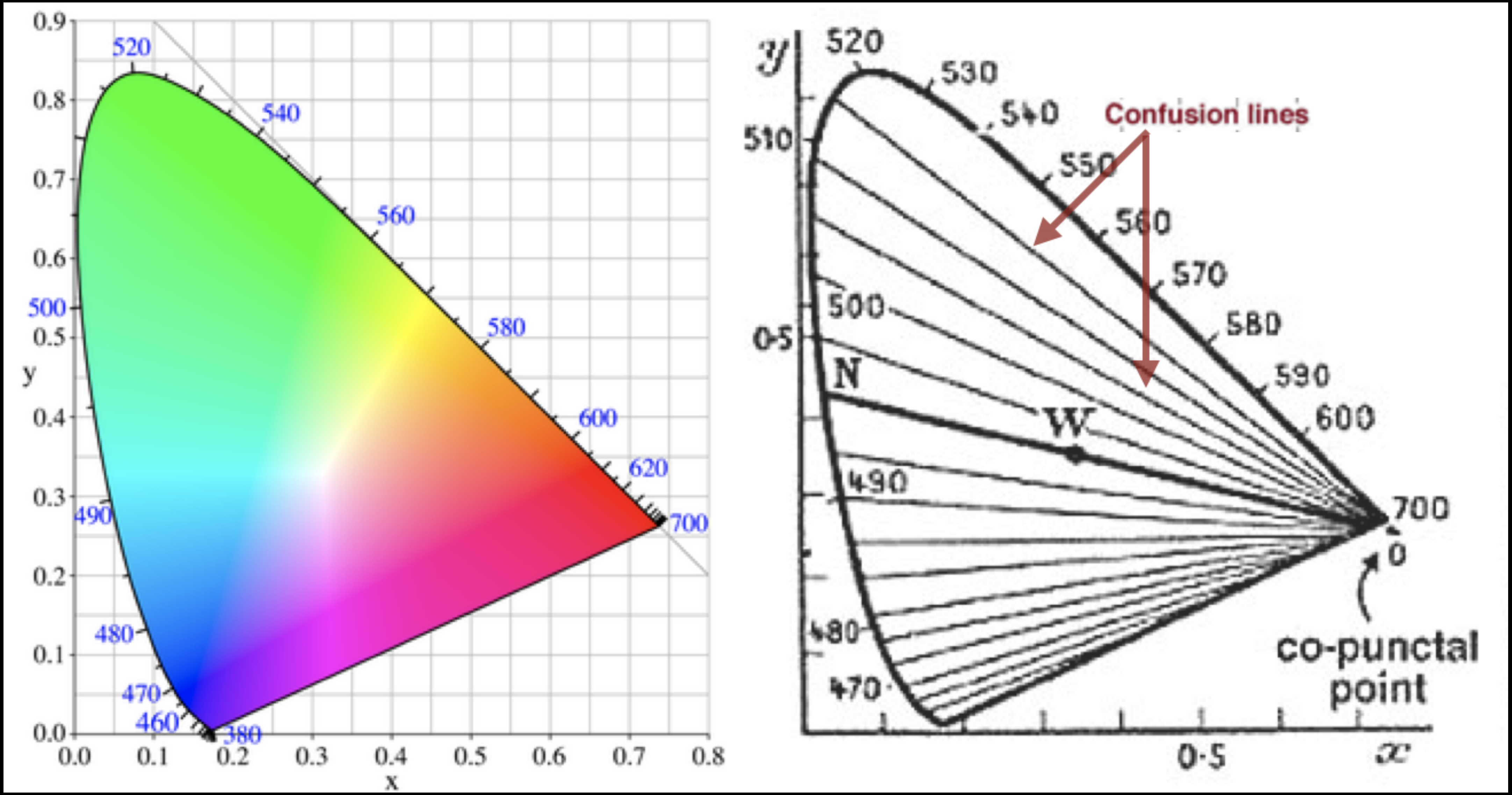
Worldwide, there are approximately 300 million people with colour blindness, almost the same number of people as the entire population of the USA!

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

color blindness

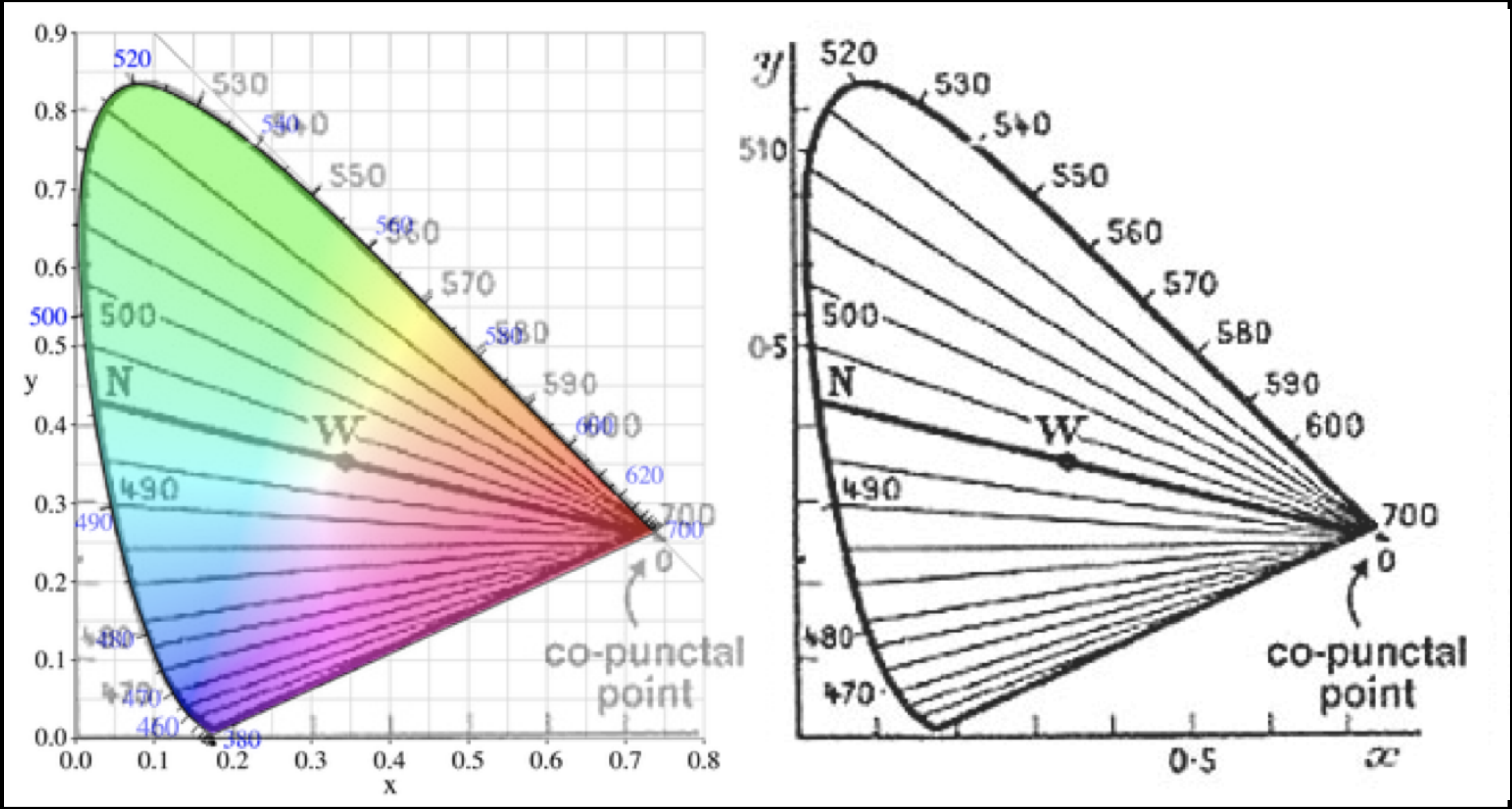
http://www.colourblindawareness.org/colour-blindness/

color blindness



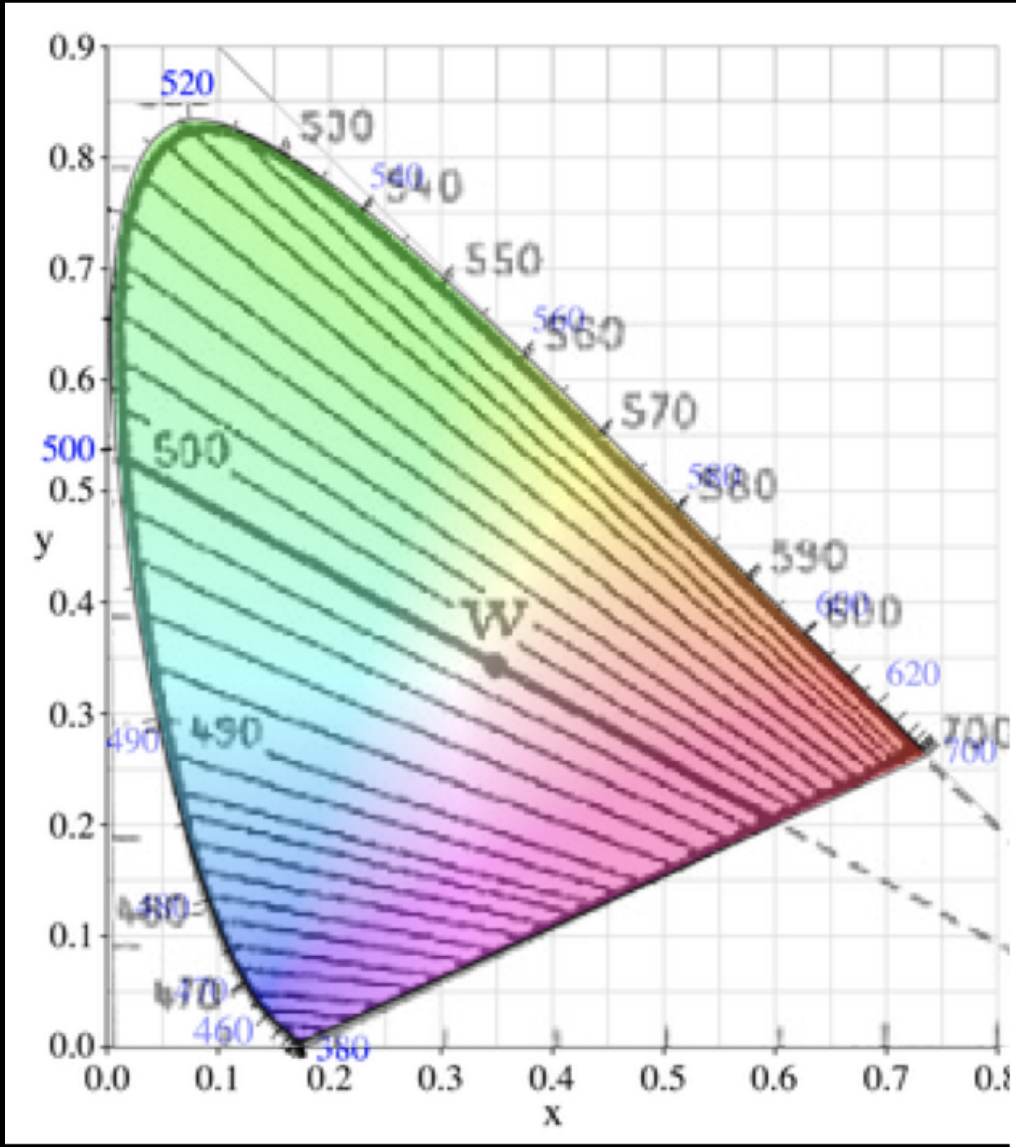
Protanopia

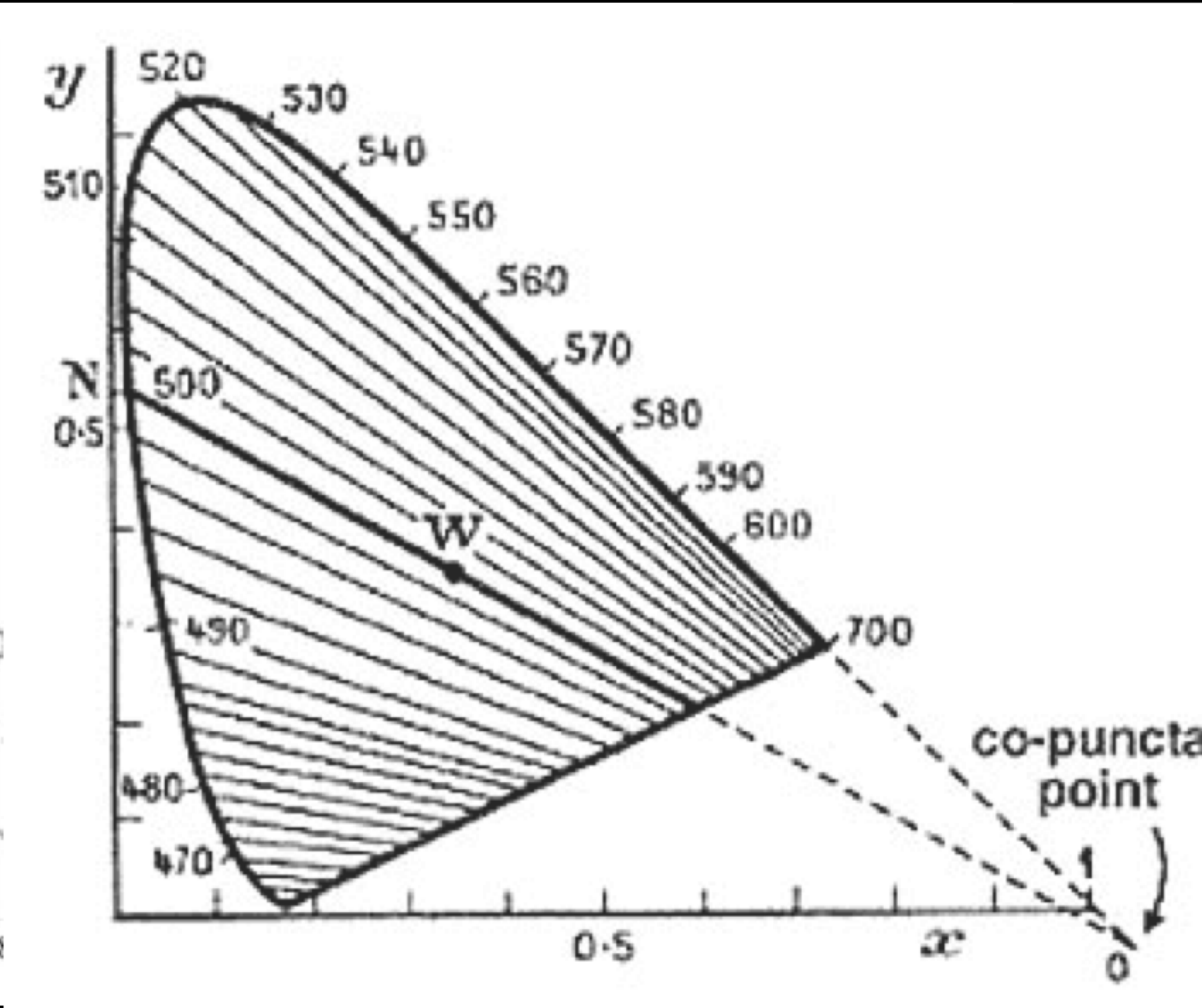
color blindness



Protanopia (red-blind)

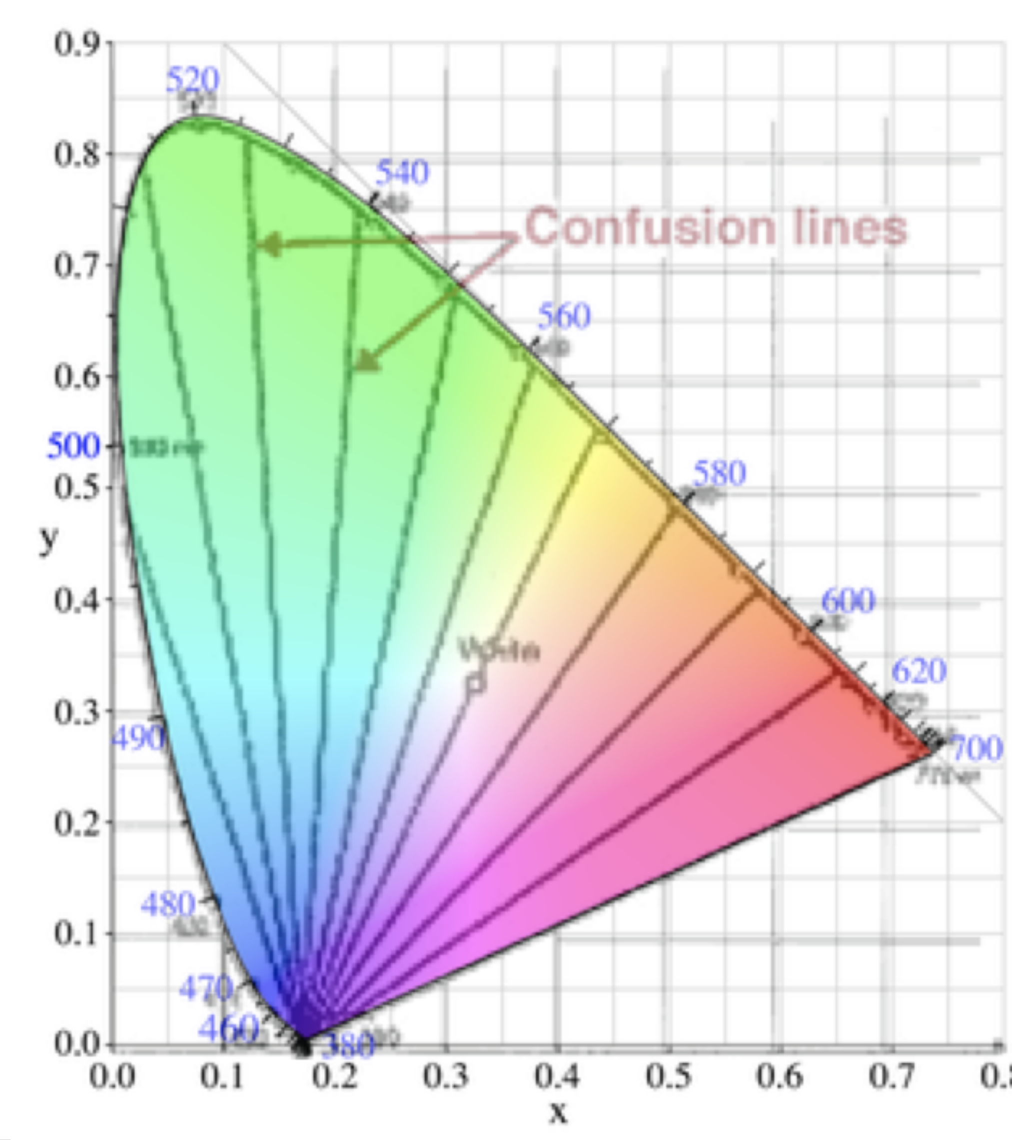
color blindness

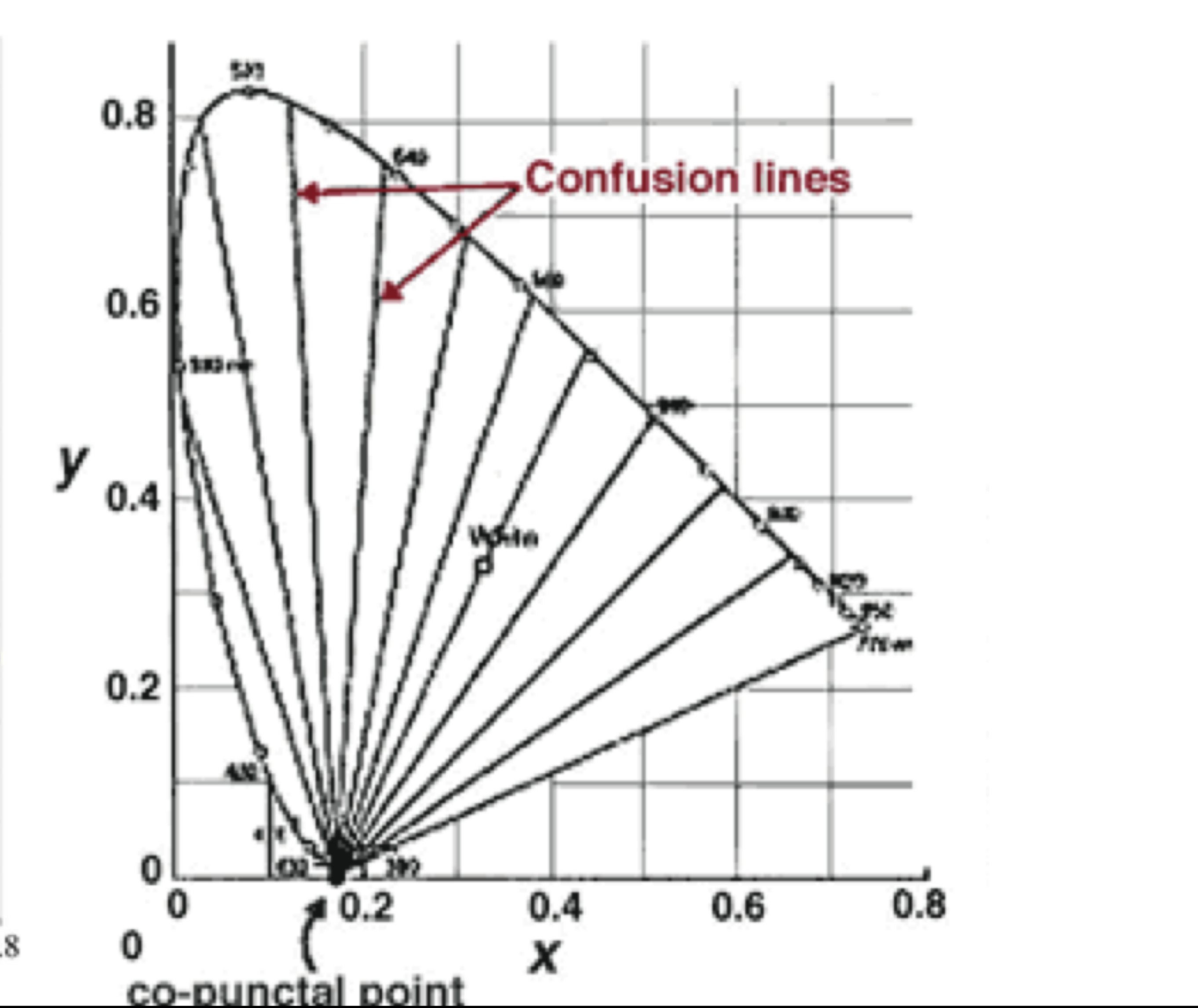




Protanopia (green-blind)

color blindness





Tritanopia (blue-blind)

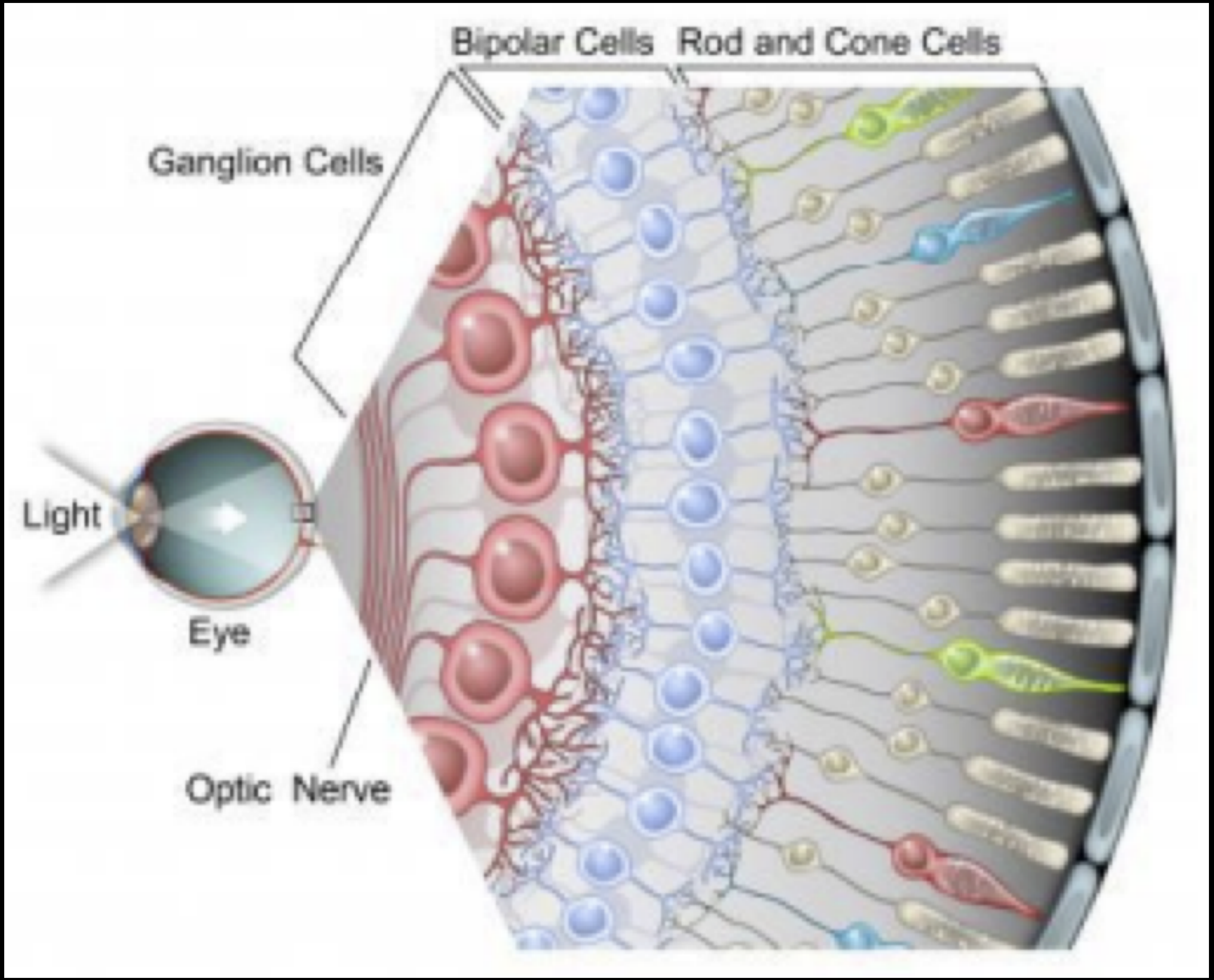
small differences can still be percieved as colors are also associated to brightness

Rods | **Cones**

Brightness | **Color**

R G B

brightness: 31% 59% 10%



http://colororacle.org/

Kelly 1965 designed a list of 22 maximally contrasting colors for colorblind compliance (the “Kelly colors”):

https://medium.com/@rjurney/kellys-22-colours-of-maximum-contrast-58edb70c90d1

TABLE I—COLORS OF MAXIMUM CONTRAST

Color Serial or selection number	General color name	ISCC-NBS centroid number	ISCC-NBS color-name (abbreviation)	Munsell renotation of ISCC-NBS Centroid Color
1	white	263	white	2.5PB 9.5/0.2
2	black	267	black	N 0.8/
3	yellow	82	v.Y	3.3Y 8.0/14.3
4	purple	218	s.P	6.5P 4.3/9.2
5	orange	48	v.O	4.1YR 6.5/15.0
6	light blue	180	v.l.B	2.7PB 7.9/6.0
7	red	11	v.R	5.0R 3.9/15.4
8	buff	90	gy.Y	4.4Y 7.2/3.8
9	gray	265	med.Gy	3.3GY 5.4/0.1
10	green	139	v.G	3.2G 4.9/11.1
11	purplish pink	247	s.pPk	5.6RP 6.8/9.0
12	blue	178	s.B	2.9PB 4.1/10.4
13	yellowish pink	26	s.yPk	8.4R 7.0/9.5
14	violet	207	s.V	0.2P 3.7/10.1
15	orange yellow	66	v.OY	8.6YR 7.3/15.2
16	purplish red	255	s.pR	7.3RP 4.4/11.4
17	greenish yellow	97	v.gY	9.1Y 8.2/12.0
18	reddish brown	40	s.rBr	0.3YR 3.1/9.9
19	yellow green	115	v.YG	5.4GY 6.8/11.2
20	yellowish brown	75	deep yBr	8.8YR 3.1/5.0
21	reddish orange	34	v.rO	9.8R 5.4/14.5
22	olive green	126	d.OlG	8.0GY 2.2/3.6

"#023fa5", "#7d87b9", "#bec1d4", "#d6bcc0", "#bb7784", "#8e063b", "#4a6fe3", "#8595e1", "#b5bbe3", "#e6afb9", "#e07b91", "#d33f6a", "#11c638", "#8dd593", "#c6dec7", "#ead3c6", "#f0b98d", "#ef9708", "#0fcfc0", "#9cded6", "#d5eae7", "#f3e1eb", "#f6c4e1", "#f79cd4"

visualizations for data exploration?

Jer Thorp

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

there definitely are historical precedents:

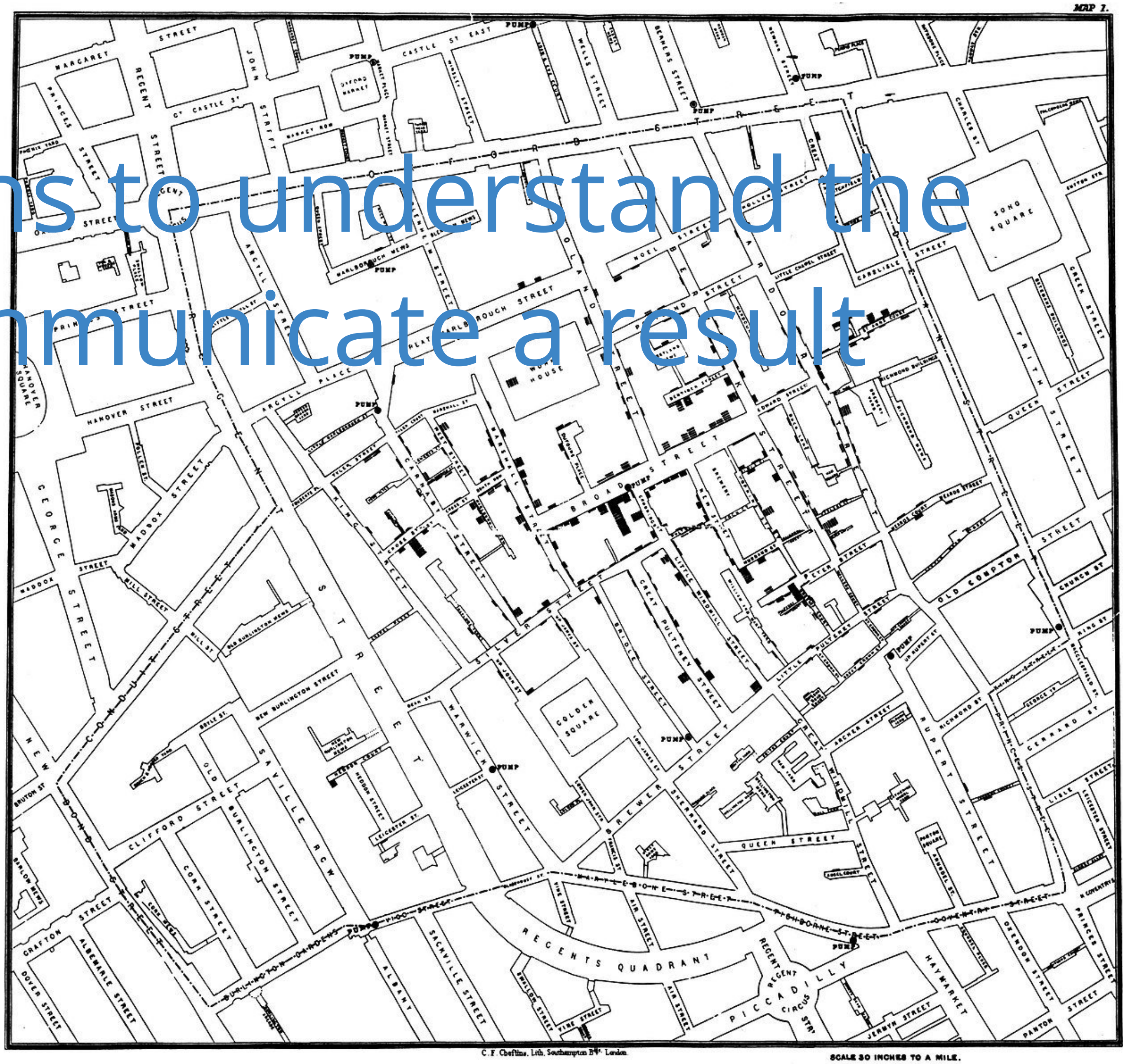
John Snow's map of cholera,

considered the first

"data science project"

uses "clustering" to drive causal inference

https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_
break#cite_ref-
FOOTNOTESnow1855[httpsarchiveorgstreamb28985266p
38mode1up_38]_19-0



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

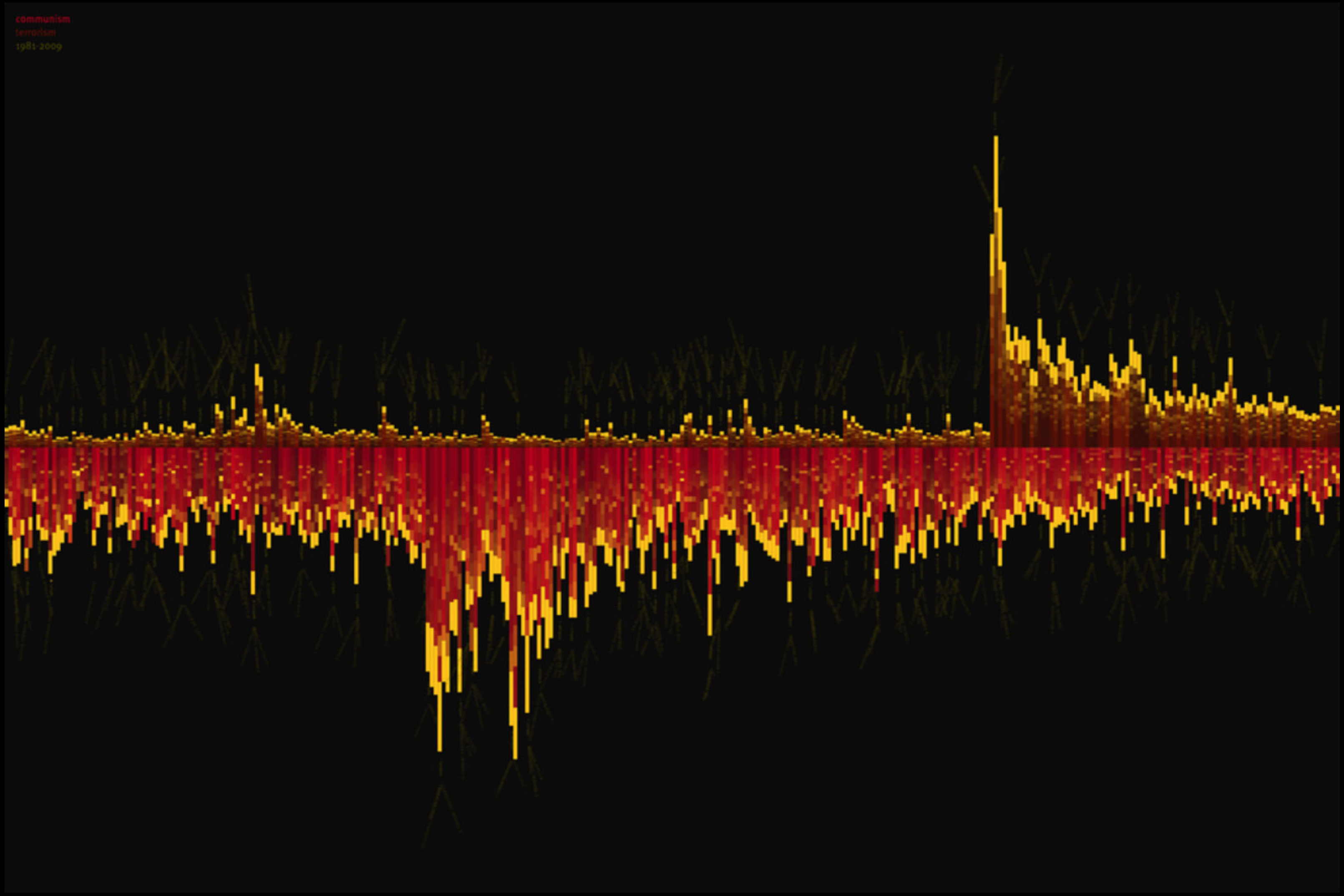
Using visualizations to understand the data, not to communicate a result

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

why the paradigm shift?

Jer Thorp

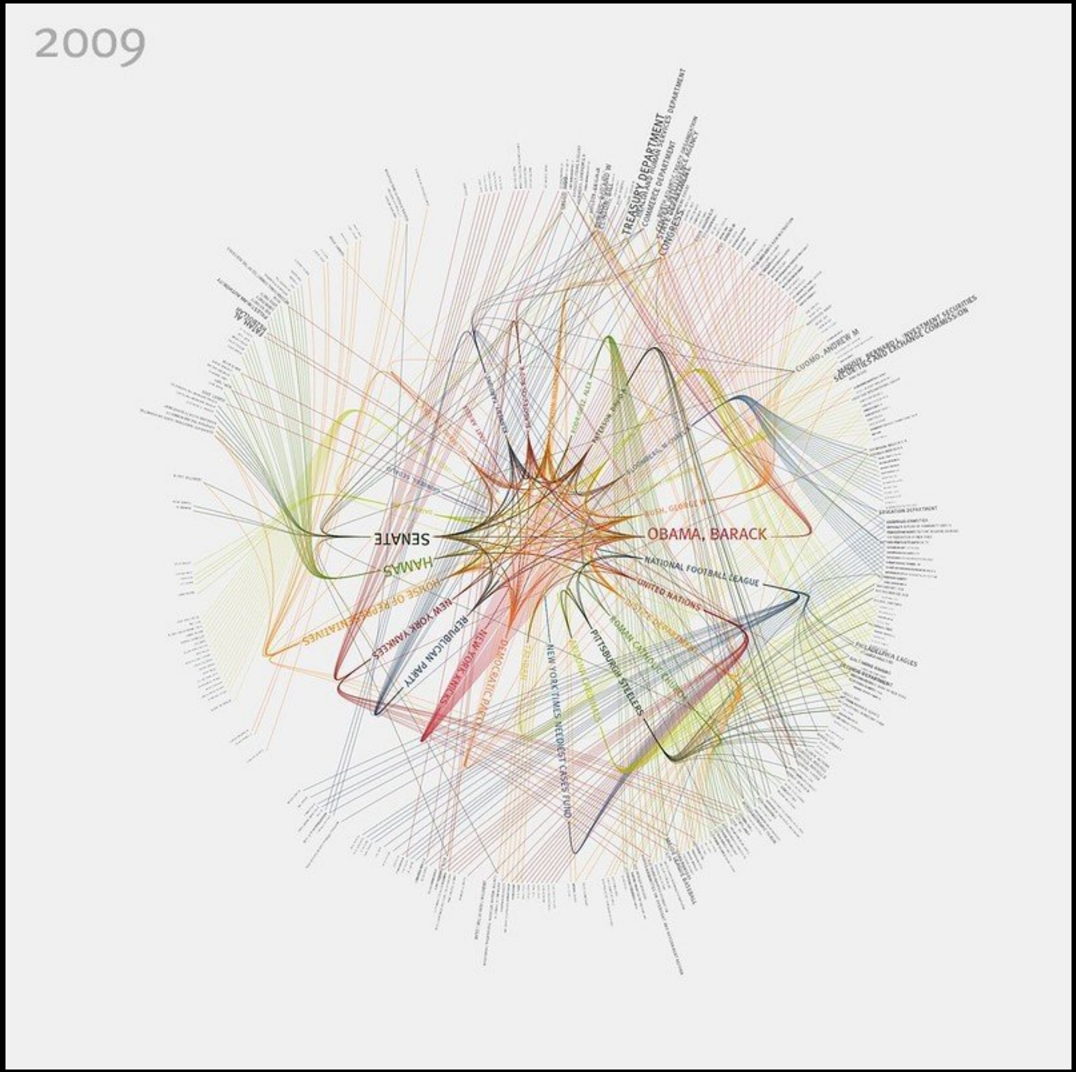
Big data:
increased data volume



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/3291268016/in/album-72157614008027965/

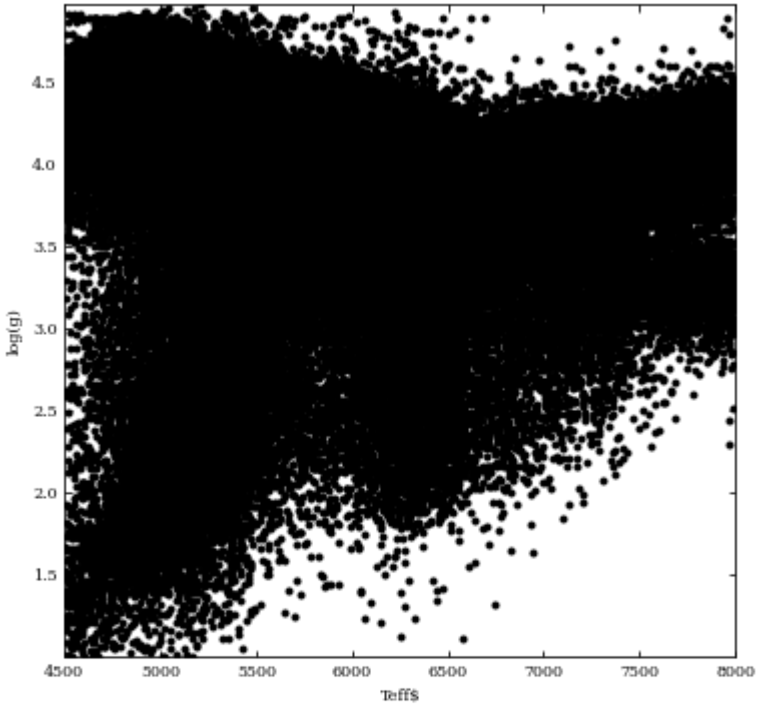
These visualizations show the top organizations and personalities for every year from 1985 to 2001. Connections between these people & organizations are indicated by lines.

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

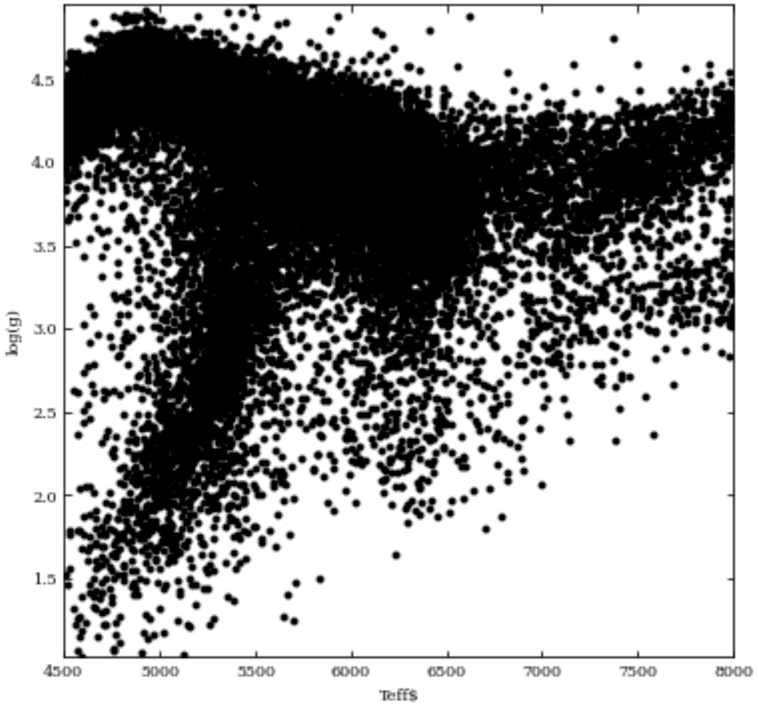
A common problem: too many points

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



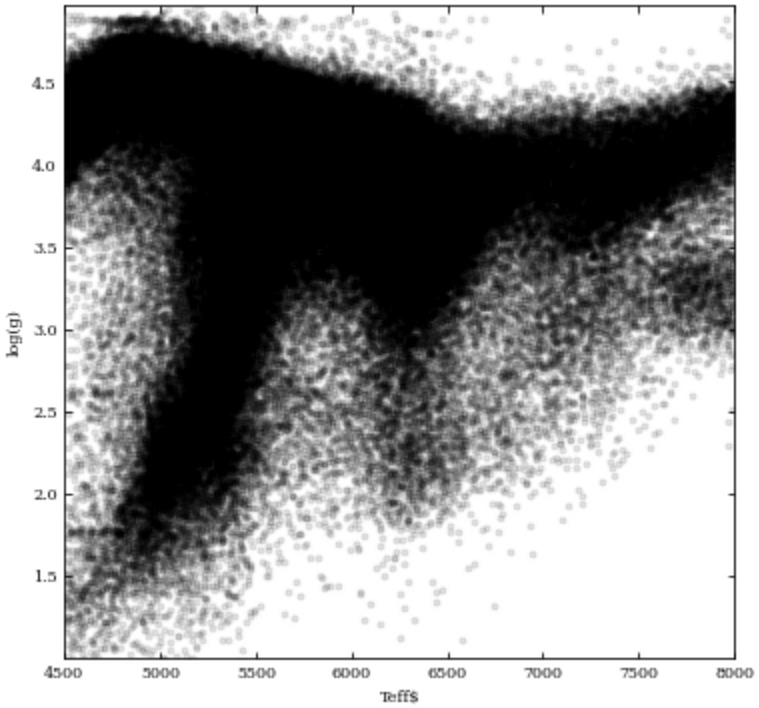
solution: subsample

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



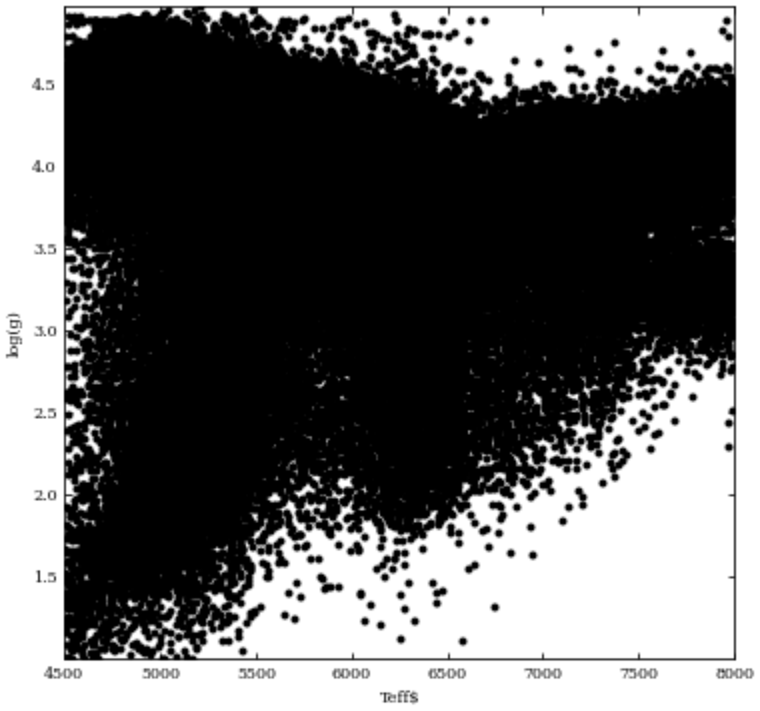
solution: alpha

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



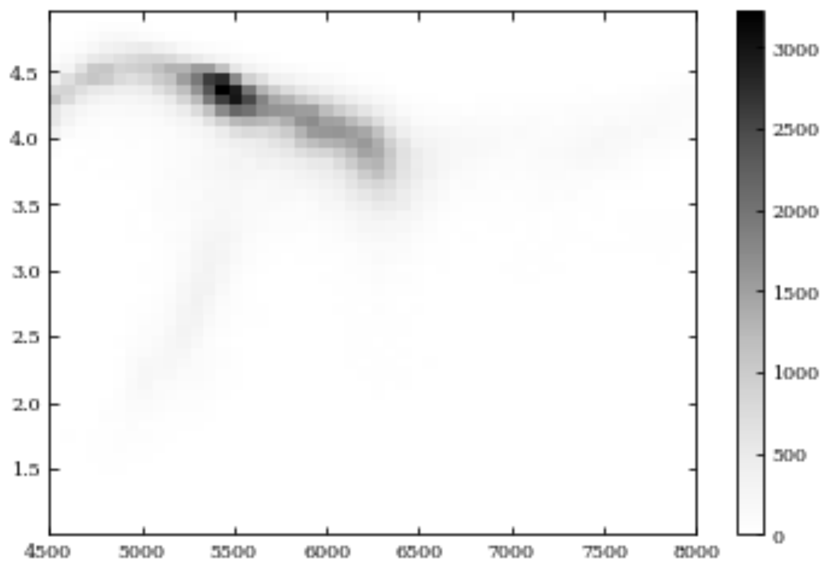
A common problem: too many points

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



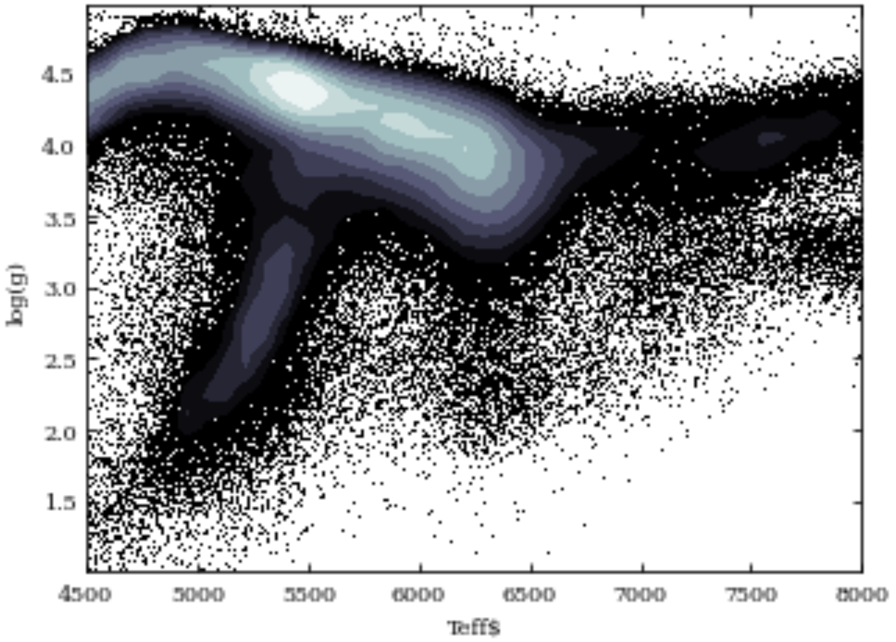
solution: density histograms

```
1 plt.hist2d(Teff, logg, bins=(50, 50), c
```



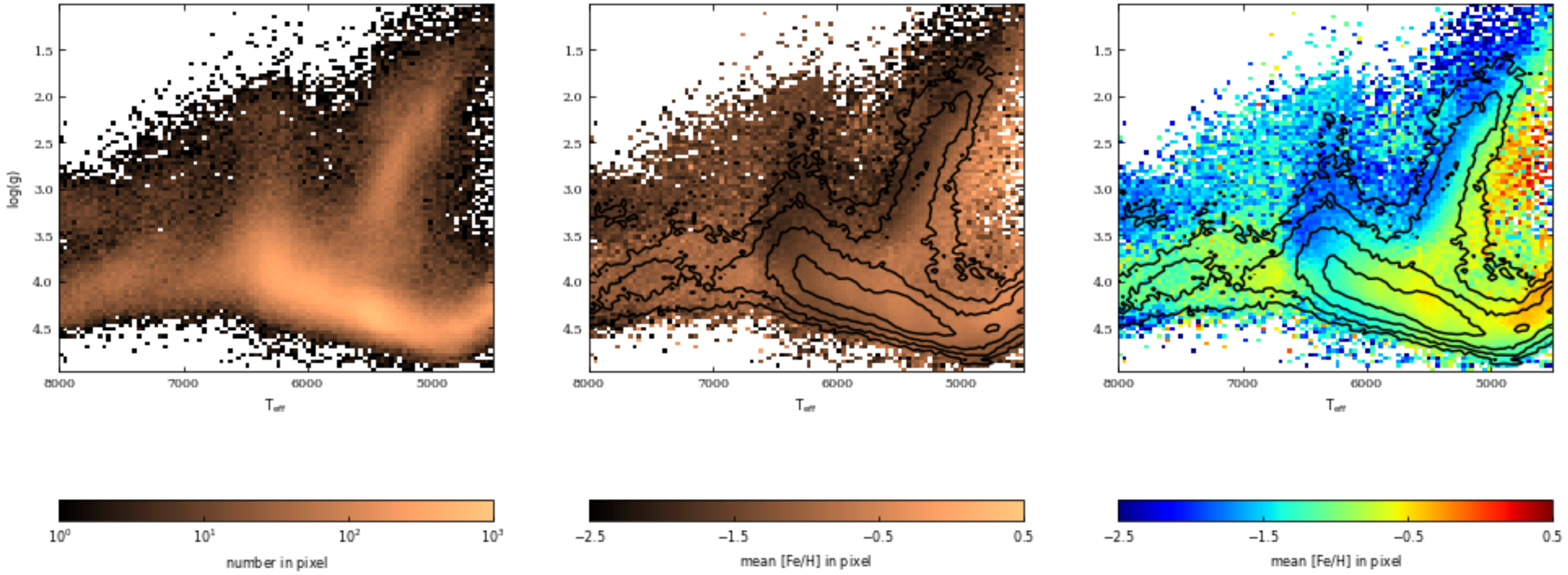
solution: scatter contours

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



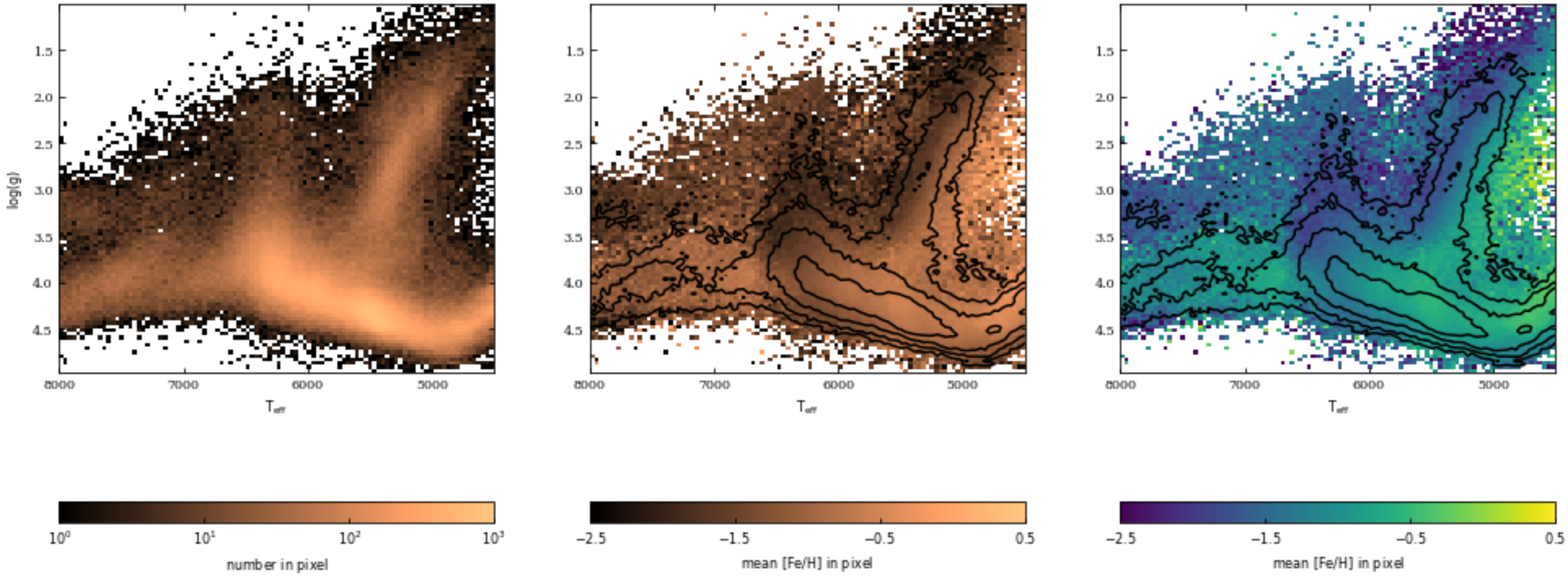
A common problem: too many points

Bad Color Choice!



A common problem: too many points

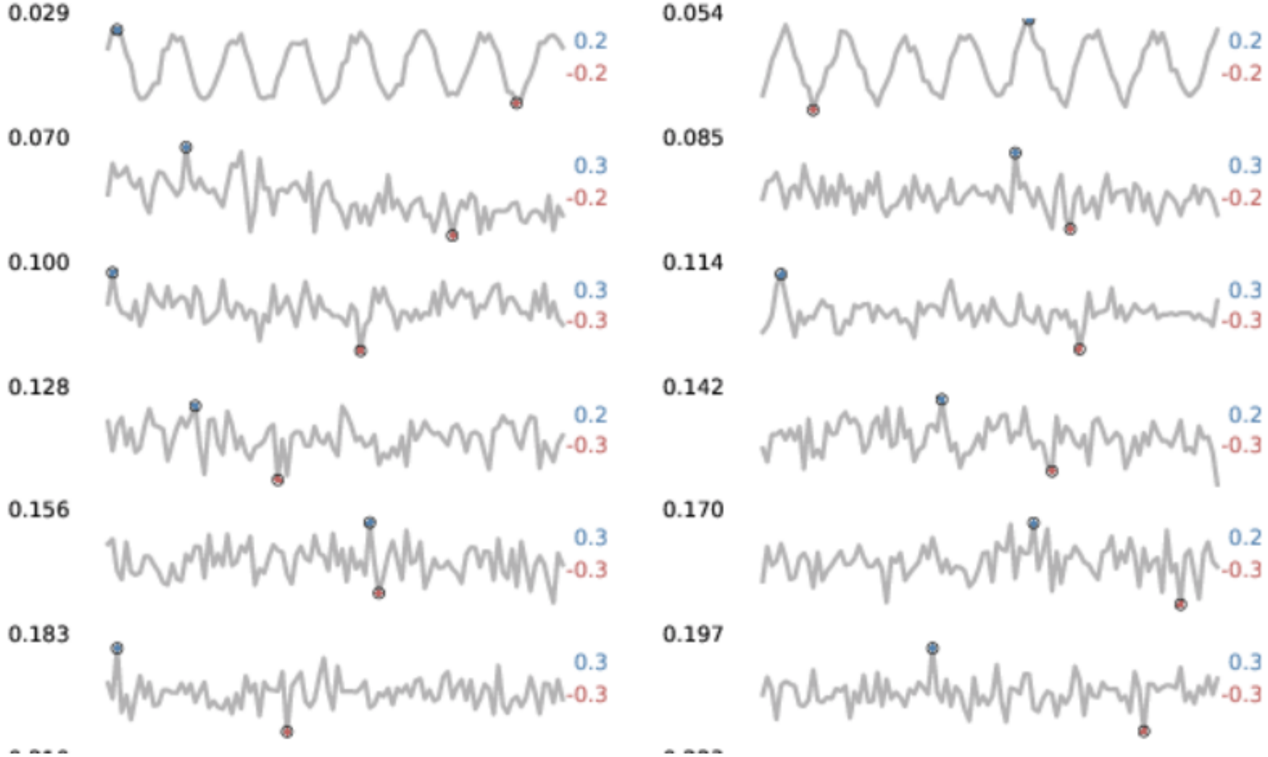
http://www.astroml.org/book_figures/chapter1/fig_SSPP_metallicity.html

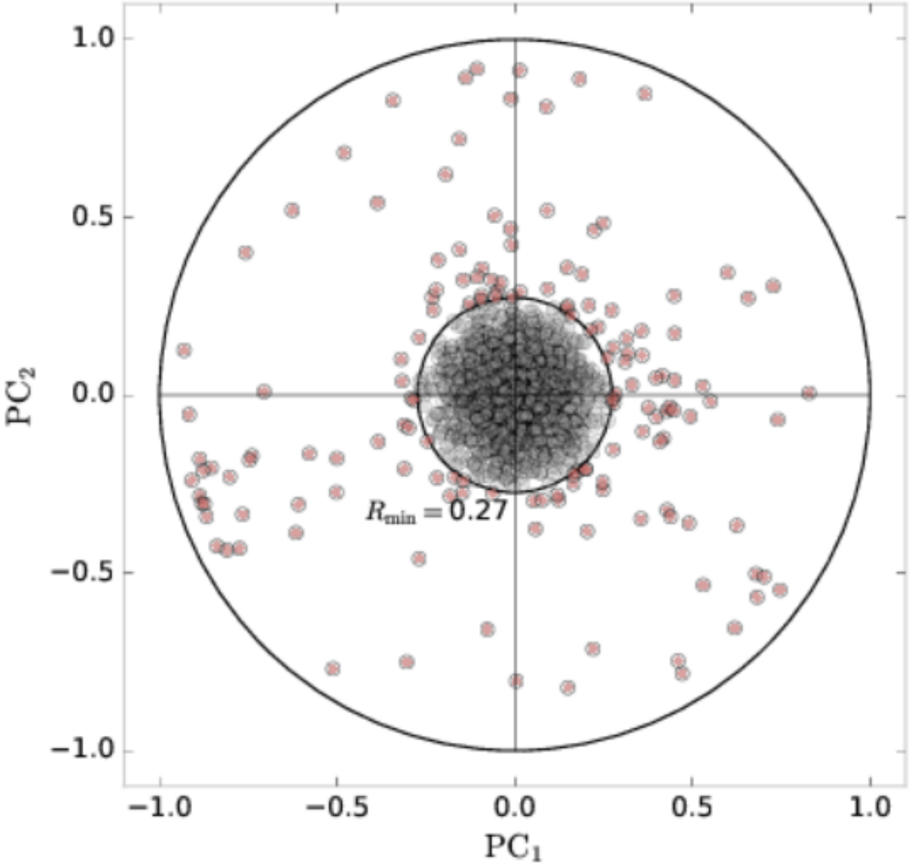


A common problem: too many dimensions

Solution: dimensionality reduction to 2 dimension

e.g. Principal Component Analysis



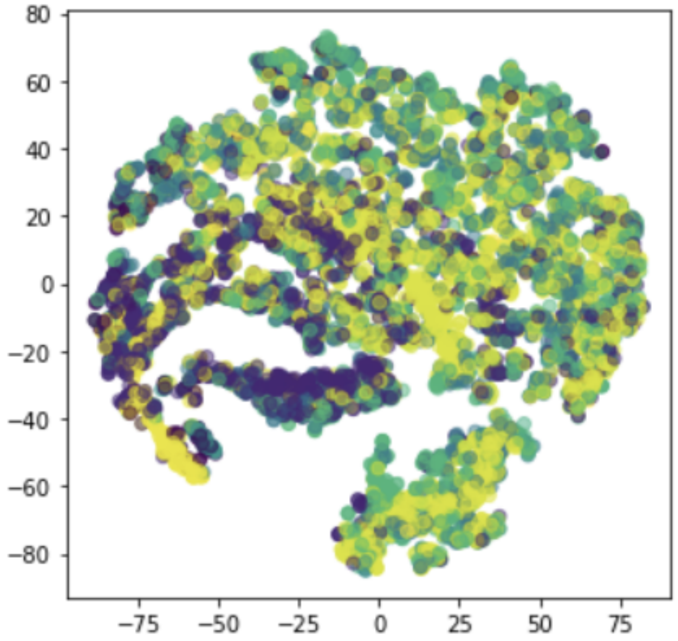


Bianco et al 2015 https://arxiv.org/abs/1611.04633

A common problem: too many dimensions

Solution: dimensionality reduction to 2 dimension

e.g. t-SNE



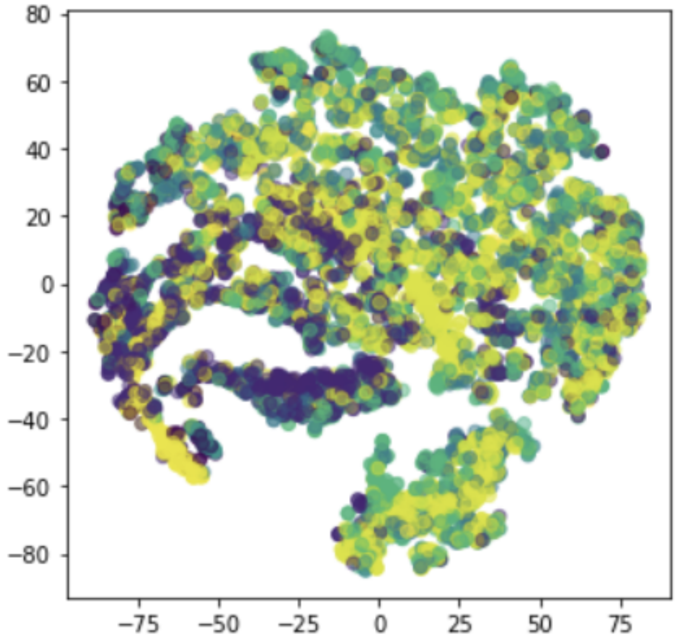
https://github.com/fedhere/MLTSA_FBianco/tree/master/HW5

Figure: the 2D projection of the 79-dimensional feature space we created by extracting features from the time series. The separation of the objects in the t-distribution based stocastic neighbourhood embedding (t-SNE) is promising: t-SNE is an embedding (a transformation to a different coordinate space) which is designed to preserve Eucledianpairwise distances existing in the higher dimensional parameter space. How many clusters can you see in this embedding? A note: since the t_SNE produces a projection of the feature space on an ideal set of coordinates, this is the only cxase in which you are allowed not to label your axes!

A common problem: too many dimensions

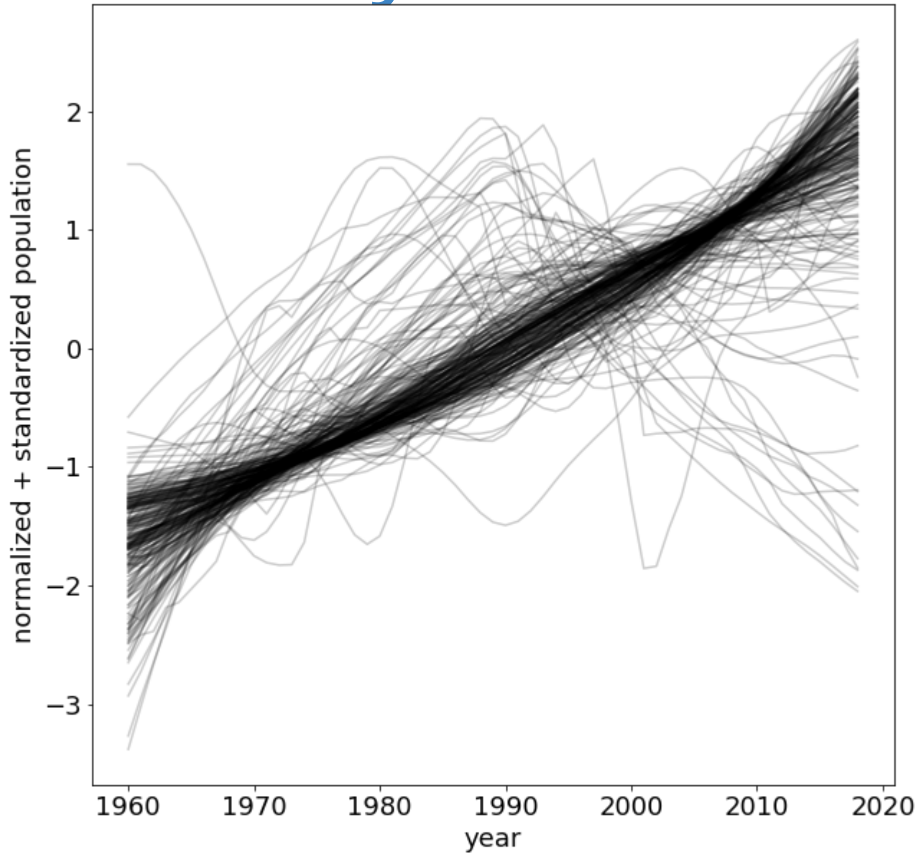
Solution: dimensionality reduction to 2 dimension

e.g. t-SNE



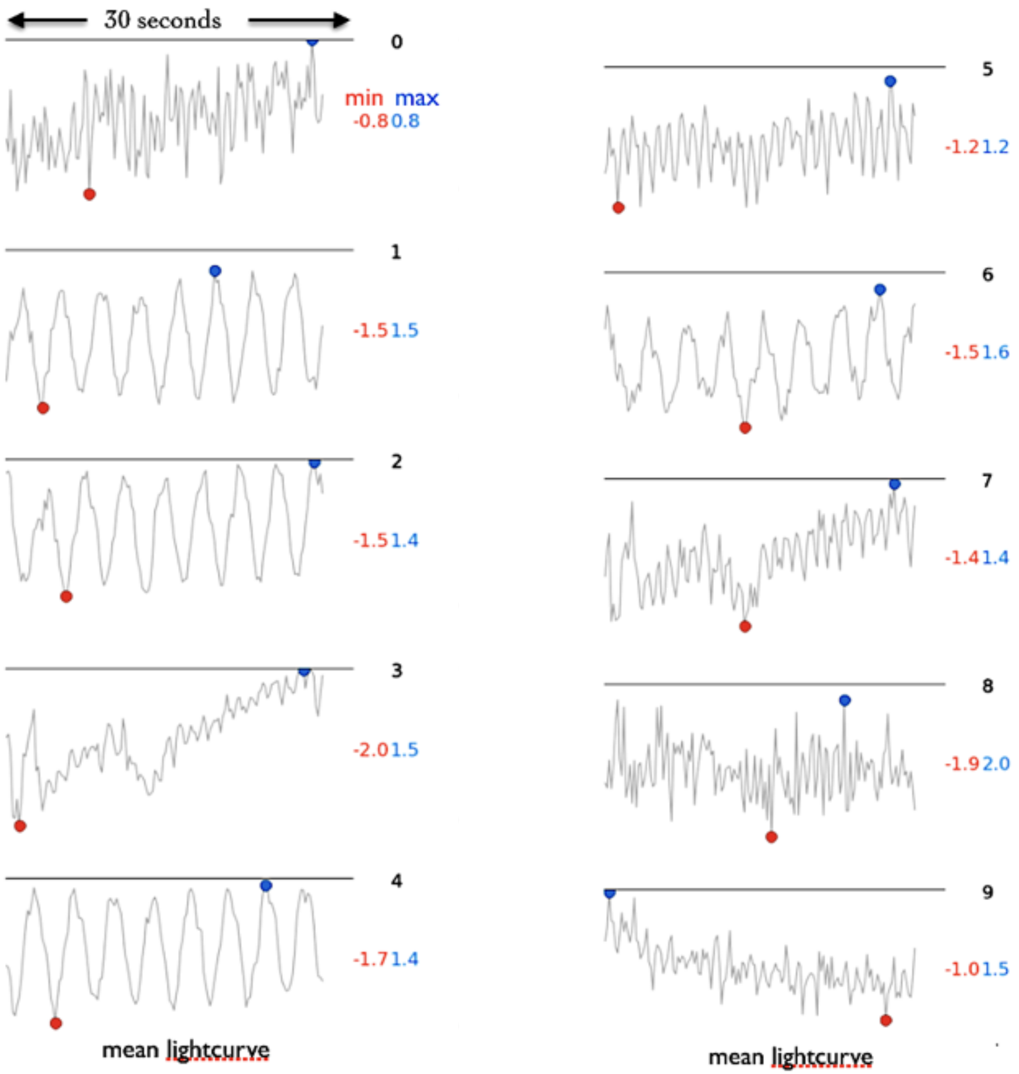
The t-distributed Stochastic Neighbor Embedding (SNE) method, (Maaten & Hinton, 2008), improving upon SNE (Hinton & Roweis 2003). SNE works by *embedding multidimensional Euclidean distances with conditional probabilities*. The similarity between x_i and x_j is the conditional probability Px_j |x_i that x_i will choose xj as a neighbor under the normal distribution. Do the same in the full dimensional and lower dimensional space, SNE then attempts to minimize the Kullback-Leibler (KL) divergence between the two probability. However, SNE is computationally very expensive; *t-SNE attempts to resolve this issue by looking at a "symmetric" SNE and redefines the lower dimensional distribution using a Student t-distribution.*

too many time series



		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

too many time series

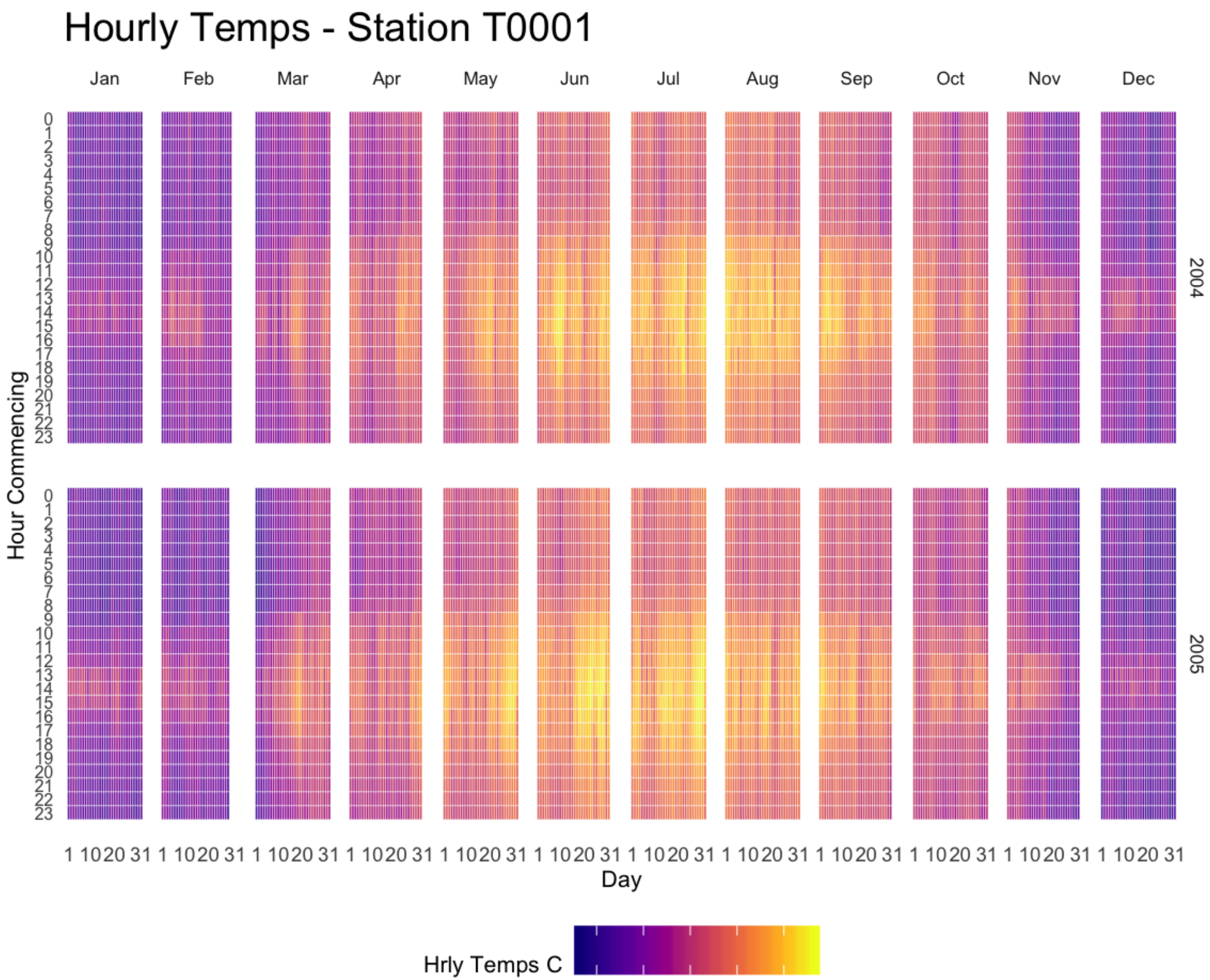


Tufte's small multiples and spakrlines

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

enable comparison by giving the data center stage

too many time series

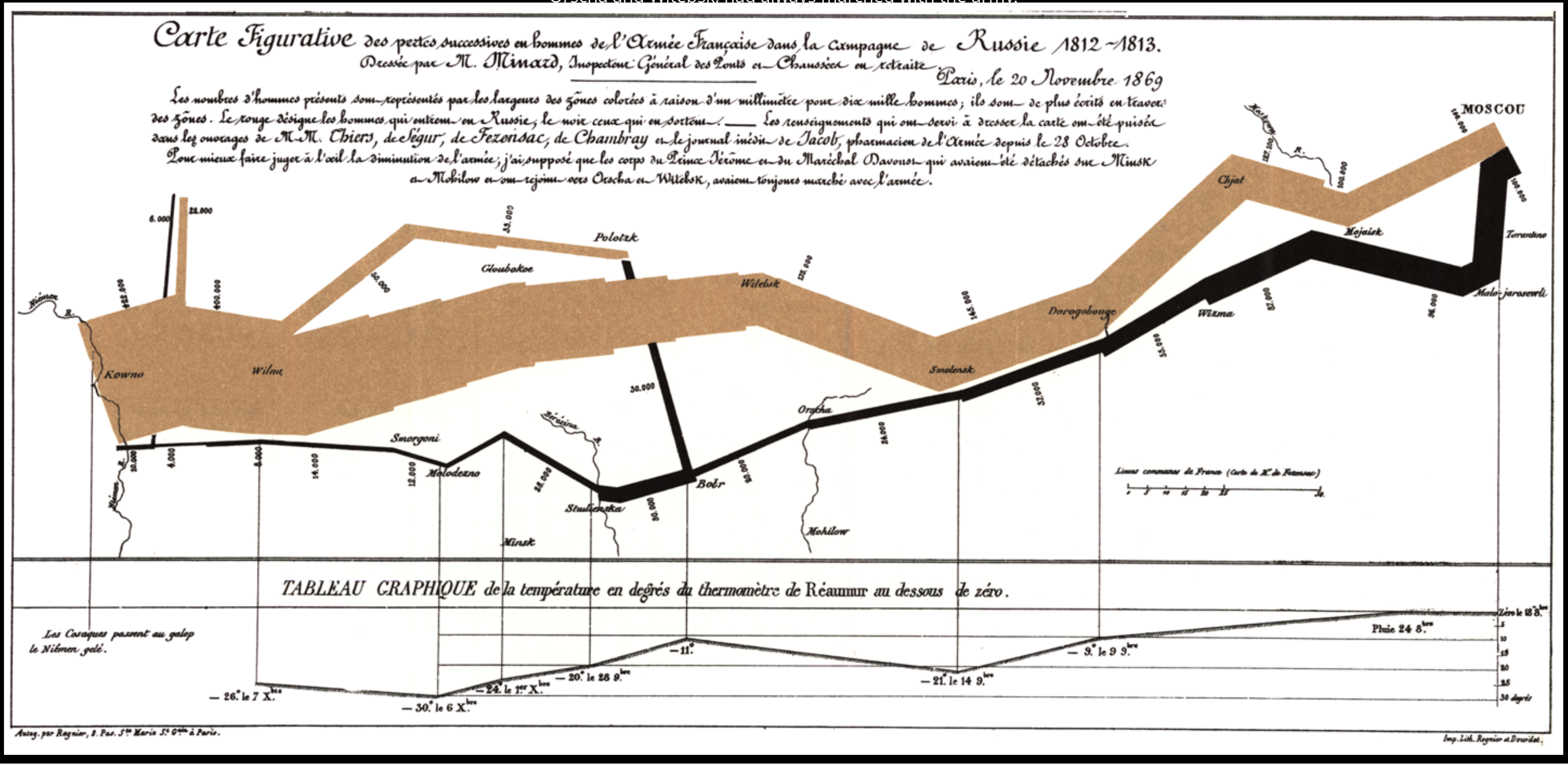


Time series heatmaps

Minard's russian campaign : so why is this plot so good?

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

The numbers of soldiers present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand soldiers; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The information 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.



so there is a thing called "the rule of 7": you cannot put more than 7 pieces of information in your plot because that his the maximum number of things a person can remember. Well, that 7 comes from a test where people are told several words and asked to repeat them back. On average people remember 7... +/-4 ...

The number of information elements that are shown in a plot depends on how effectively you can show them. This plot contains (at least) the following features:
space (distance, however approximate), time, size of the army, rate of lives lost (highly covariant with size of the army), purpose (going on the attack toward Moskow or retreating, indicated by the color), topography (changes of direction, rivers), temperature, the last 2 are conveying a causal connection by showing the lives lost (decrease in width of the army size) in conjunction with critical temperatures and rivers)

Tufte's rules lie factor=1, data/ink ratio high, no chart junk, #graphical elements<#features

Munzner's rules functional use of color, no unjustified 3D, eyes over memory (granted.... they did not have animations back then)

Be thoughtful and make sure your visualizations are (in this order):

honest

clear

convincing

beautiful

key concepts

Identify the purpose of your visualization:

visualize to communicate results

visualize to understand data and guide analysis

key concepts

Edwaed tufte (anything)

Wassily Kandinsky, *Point, Line, and Plane,* 1926

Tamara Munzner

Visualization Analysis & Design, 2014

(link to a talk slide-deck about her book: http://www.cs.ubc.ca/~tmm/talks/minicourse14/vad15london.pdf)

color maps http://www.kennethmoreland.com/color-maps/

Kelly colors https://medium.com/@rjurney/kellys-22-colours-of-maximum-contrast-58edb70c90d1

resources

7 Great Visualizations from History

https://web.archive.org/web/20171114145335/http://data-informed.com/7-great-visualizations-history/

7 classical vis papers

https://web.archive.org/web/20190226213626/http://fellinlovewithdata.com/guides/7-classic-foundational-vis-papers

Six Lessons from the Bauhaus: Masters of the Persuasive Graphic

http://blog.visual.ly/six-lessons-from-the-bauhaus-masters-of-the-persuasive-graphic/

Using preattemptive processing elements

https://pdfs.semanticscholar.org/0456/bc9cdf02c3a446e252cf2e6b83145e17749a.pdf

resources

Any of these papers:

7 classical vis papers

https://web.archive.org/web/20190226213626/http://fellinlovewithdata.com/guides/7-classic-foundational-vis-papers

reading