CS 4449 Data Science Capstone

Claudio Delrieux
claudio.delrieux@du.edu

Midterm grading

You all did a great job! We had 12 A, 6 A- and 2 B.

Feedback and eventual resubmission is allowed upon request (I'll need some time, though).

Term project presentation

Please those of you who still didn't, drop me a line regarding the term project you choose.

Also notice that the original links to the data of the projects in the async are broken, so contact me for alternative links.

If its not any of the suggested ones and you plan to work with something different, we have to agree wrt the scope and depth.

Term project presentation

If you prefer you can team-up in two person groups.

We have to define ASAP a schedule so we can allot the presentation time during weeks 9 and 10.

We had ~12 mins per midterm presentation, and the term projects are more complex and elaborate. This will be very tight in two classes.

Term project presentation

Expected structure for the slides and the report:

1. Short pitch.
2. Why you considered it, what challenges you faced.
3. **Time breakdown**
4. ABCD.
5. The main presentation with your work.
6. Conclusions

Term project presentation

Expected deliverables:

1. Slide deck.
2. Report (~10 pages).
3. NB file or link to repository (GitHub, Colab).
4. Datasets (if not using the provided).
5. Ancillary material if any.

Term project presentation

Be sure to address all aspects as required in the syllabus. In the suggested term projects there are specific questions to be answered, supported by evidence derived from the data analysis.

If you will work with a different project, please state clearly what are the questions to be addressed by your analysis.

Term project presentation

During your presentation, devote the first minute to:

1. A description of the general scenario and the particular case or problem.
2. Why did you considered it, how was your approach, and the challenges you found.
3. A time breakdown for each of the stages (project assessment, data preparation and analysis, model development, experiments, wrap up) .

Term project presentation

Questions please?

Data Visualization and Visual Analytics

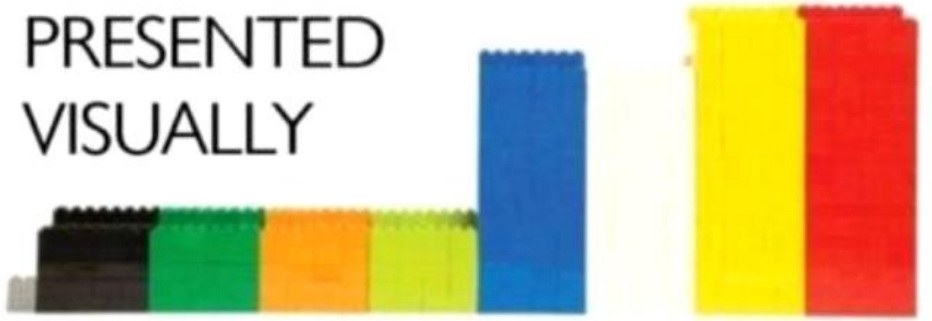
Visual Storytelling

DATA

SORTED

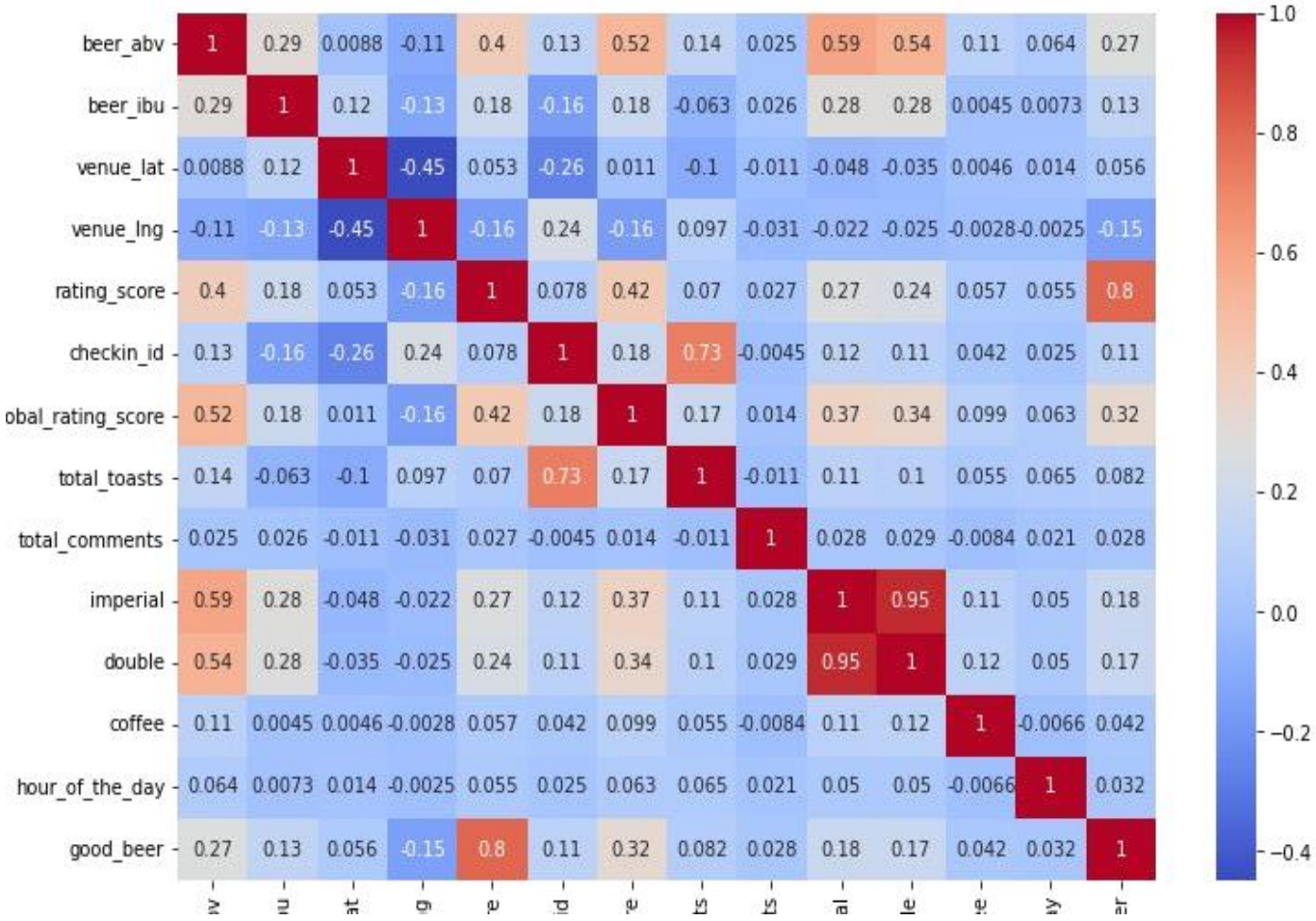
ARRANGED

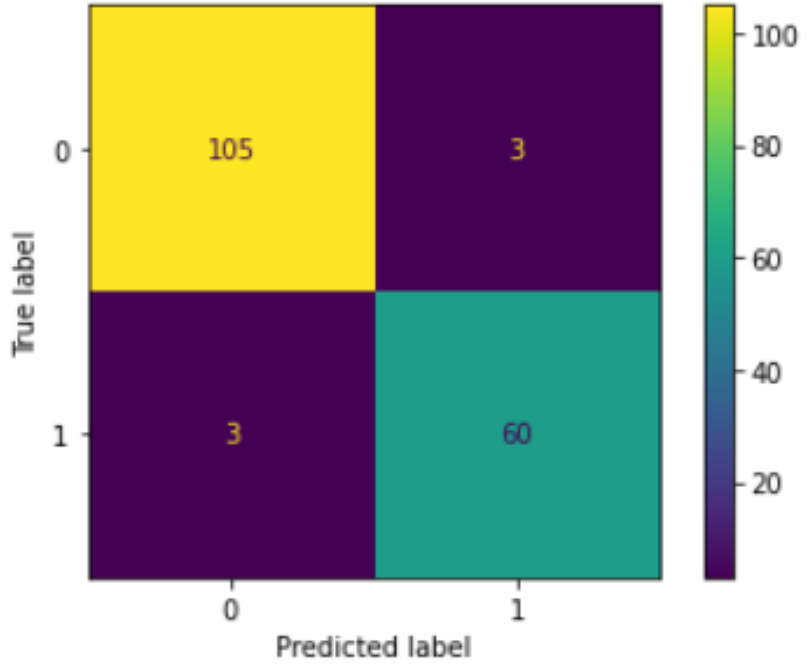
PRESENTED VISUALLY



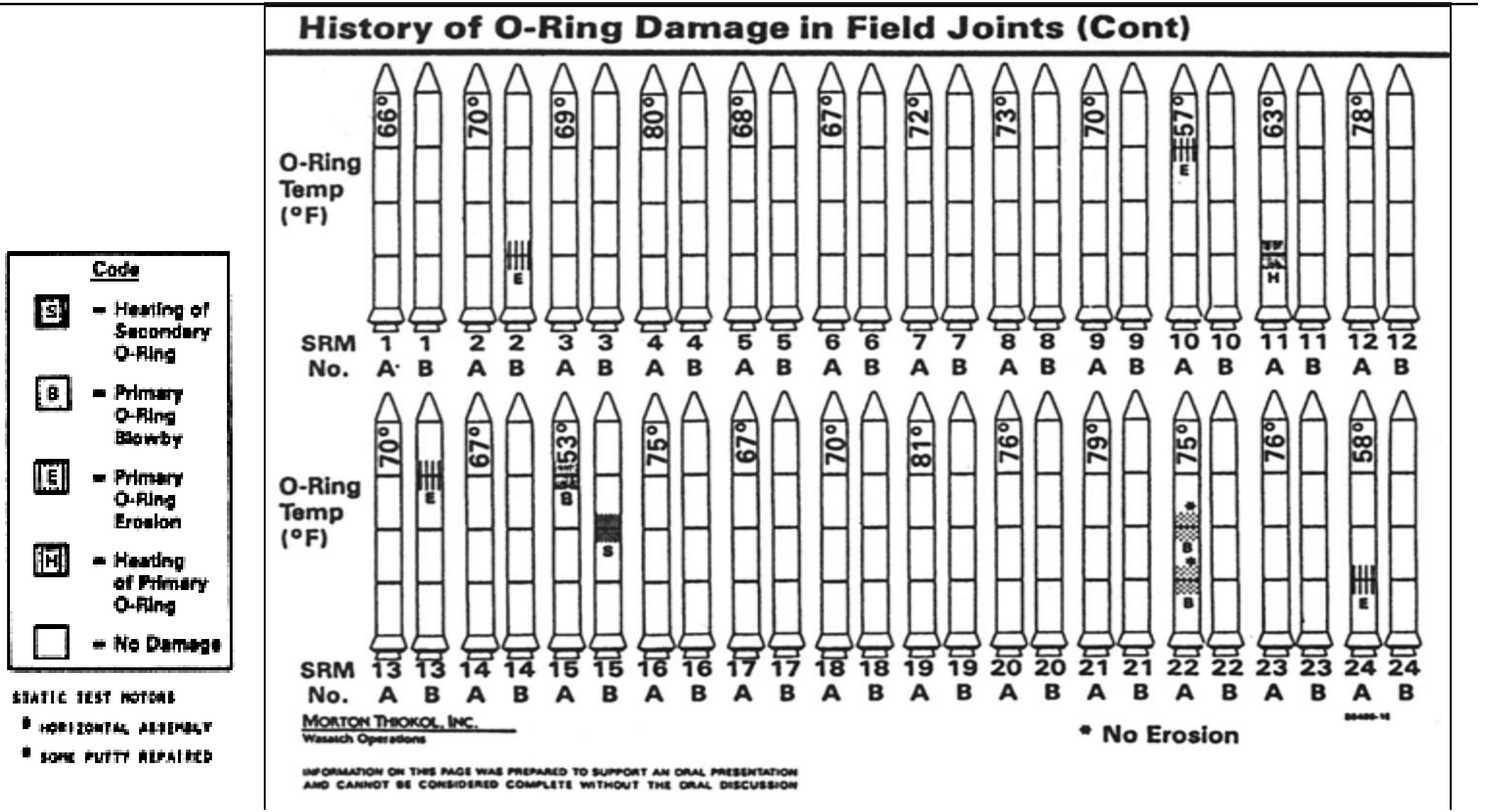
EXPLAINED WITH A STORY

How many ways to display a corr. matrix?



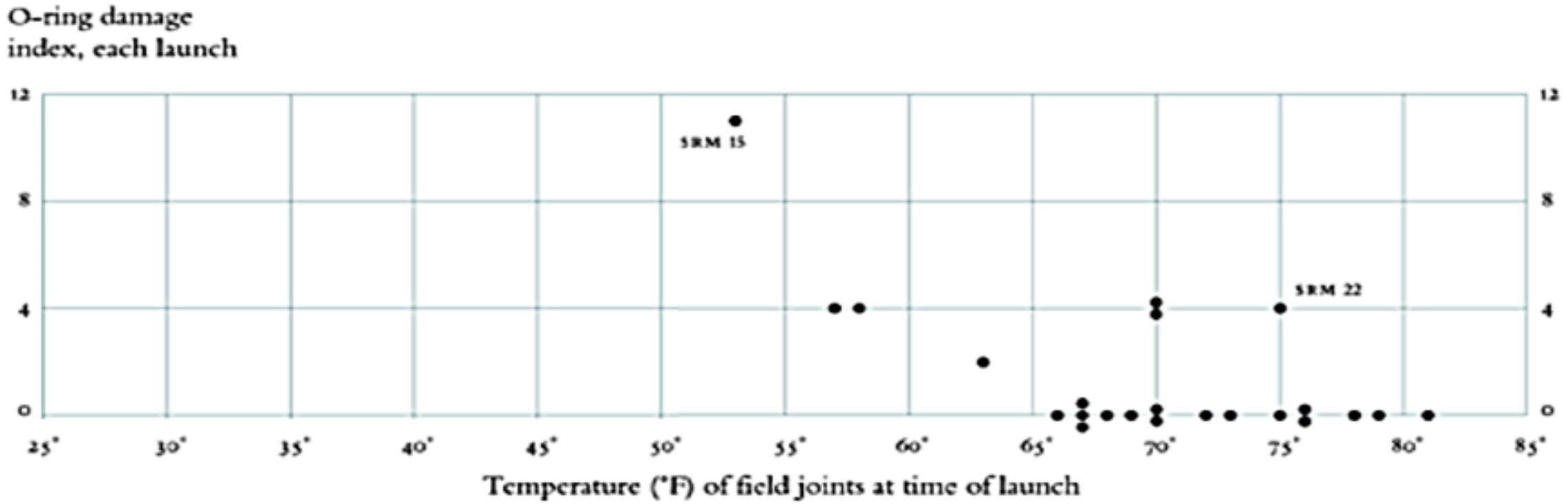


An example: the Challenger



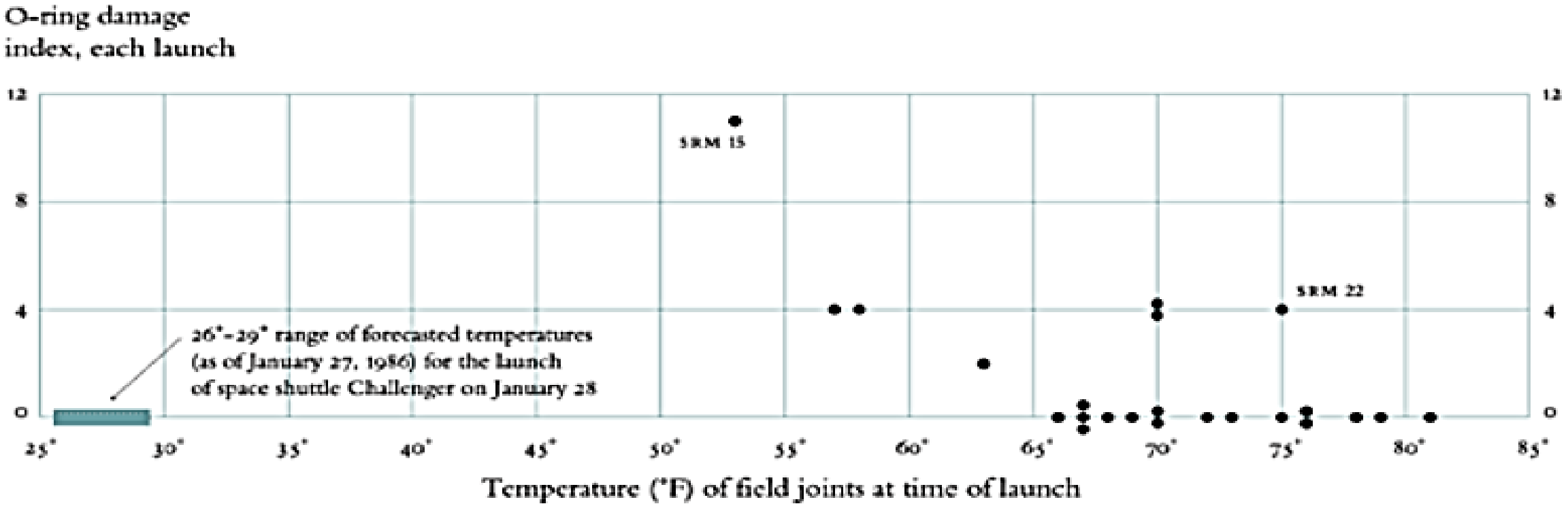
An example: the Challenger

O-ring damage vs. launch-time temperature



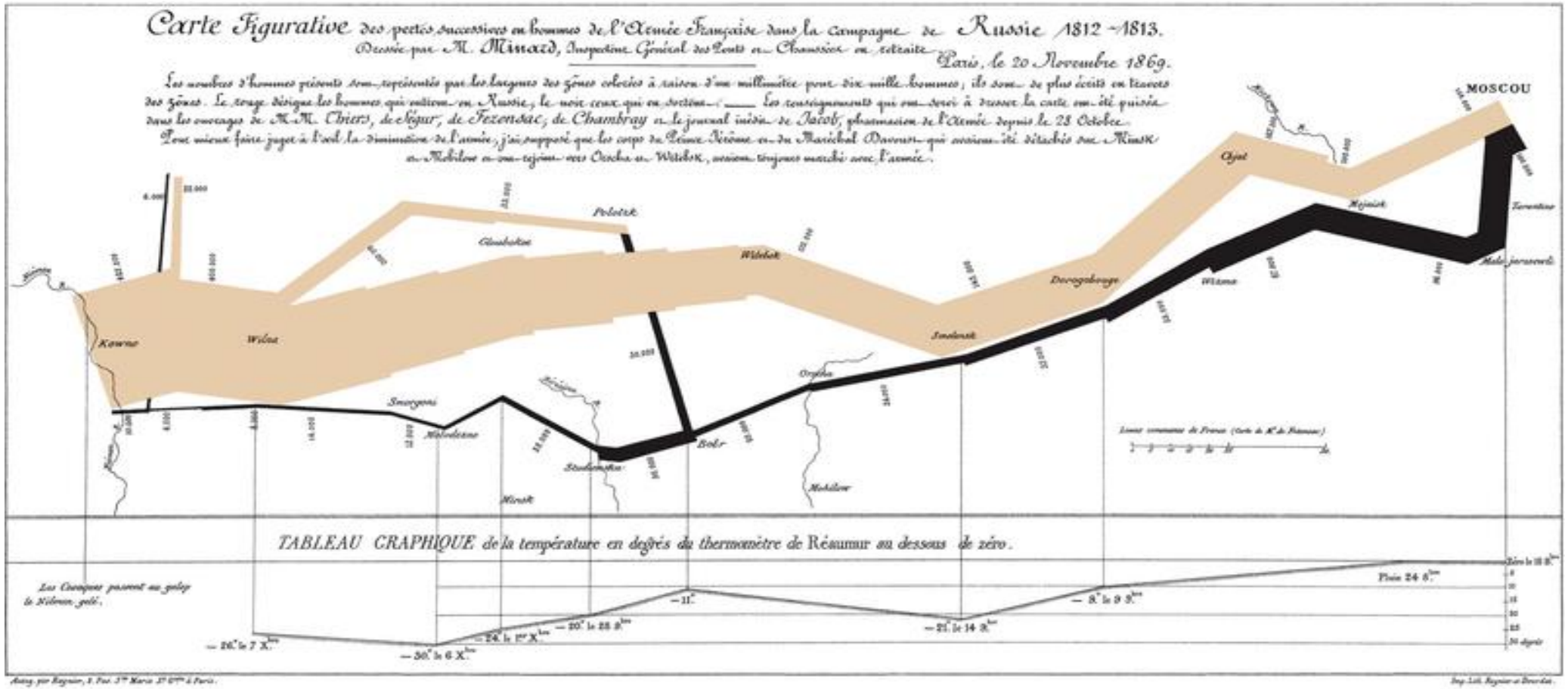
An example: the Challenger

Would you recommend launch with this forecast?



Data visualization

Lots of incredible historical examples. See E. Tufte's books!



A few definitions

According to Robert Spence, visualization is essentialy a mind activity that starts with a graphic representation and ends with a mental model in the observer.

Joan Costa states that a visual communicator translates "abstract data" into visually understandable messages, rendering complex data and phenomena plain to see.

Lev Manovich claims that the main issue in visualization is a *mapping* between certain data properties and visual attributes (positions, sizes, colors, etc.). The key aspect, then, is how data is visually encoded.

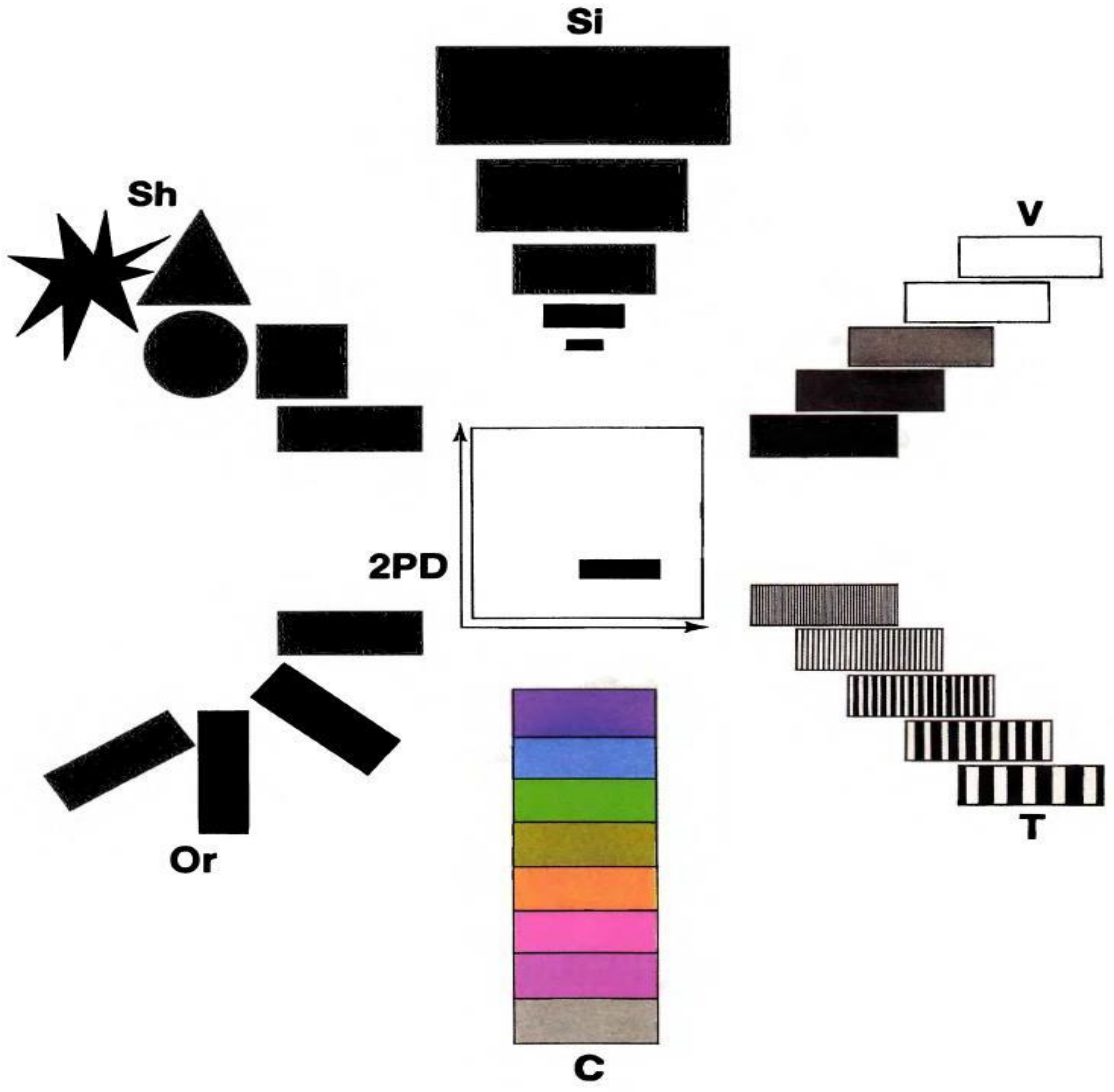
Stuart Card rephrases "the purpose of visualization is insight, not pictures".

A few definitions

Following Jacques Bértin, the visual attributes are eight:

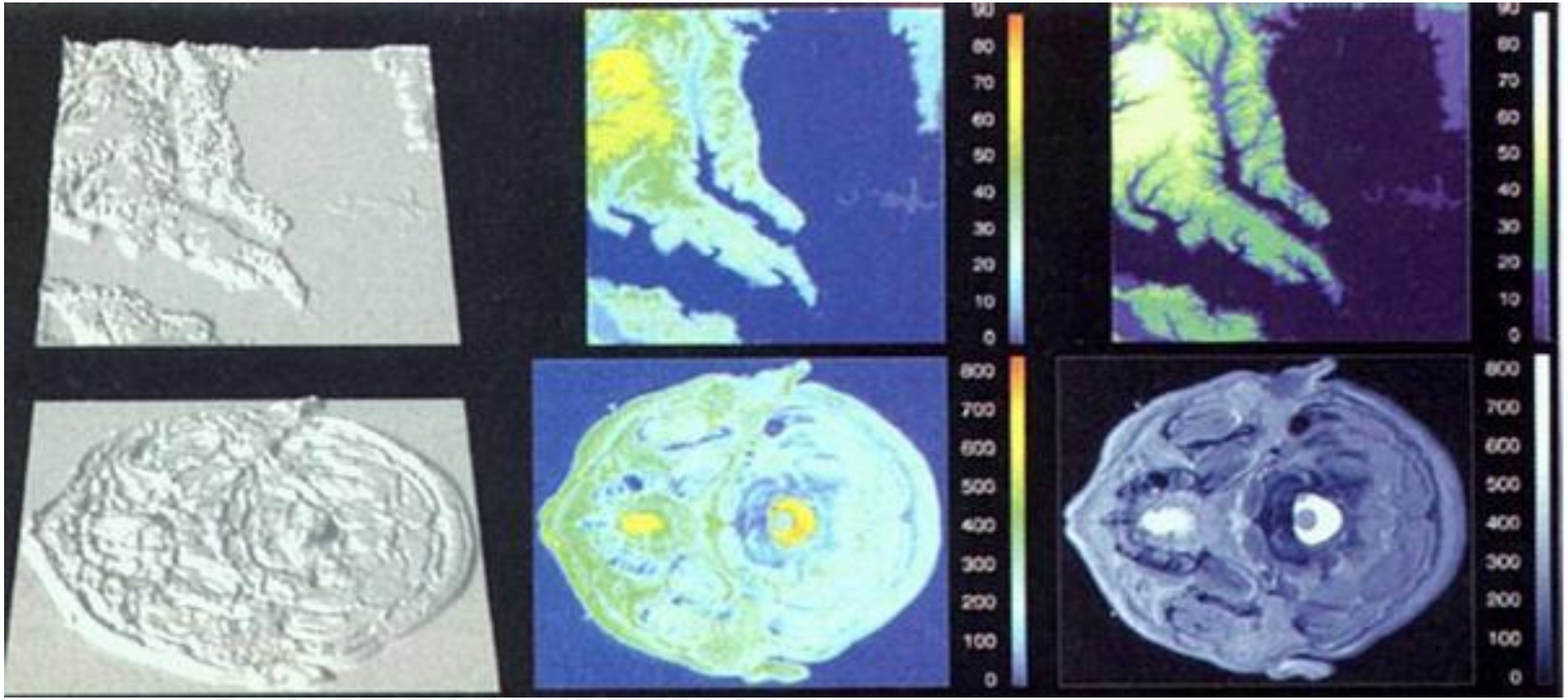
- 2D position
- Color
- Value
- Texture
- Shape
- Size
- Orientation

We will learn about their effectiveness in visualizing different data properties.



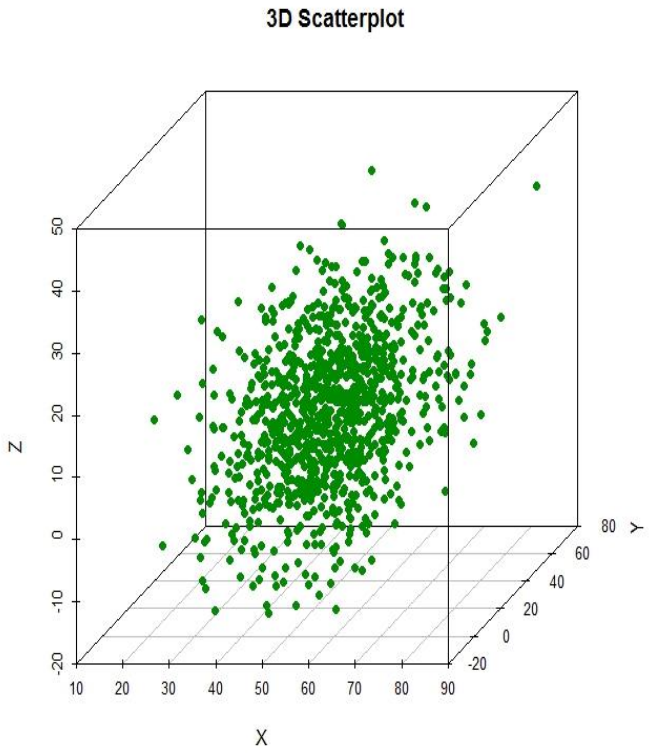
Visual Mapping

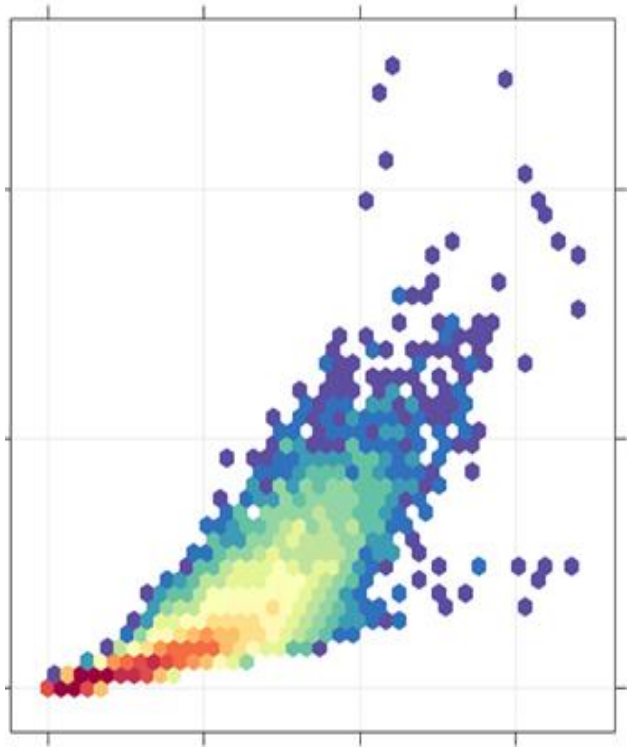
Several ways to map info to visual features.

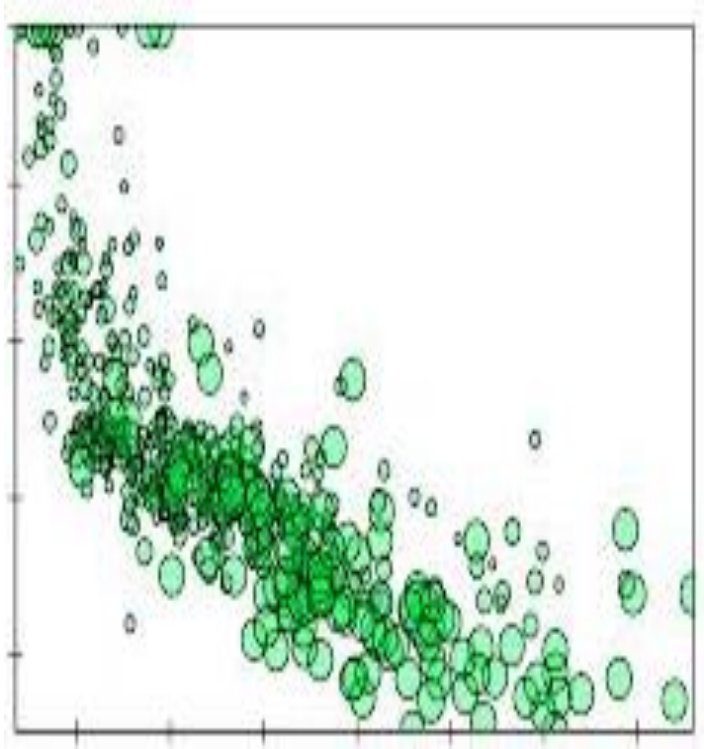


Visual Mapping

We have another “curse of dimensionality”.

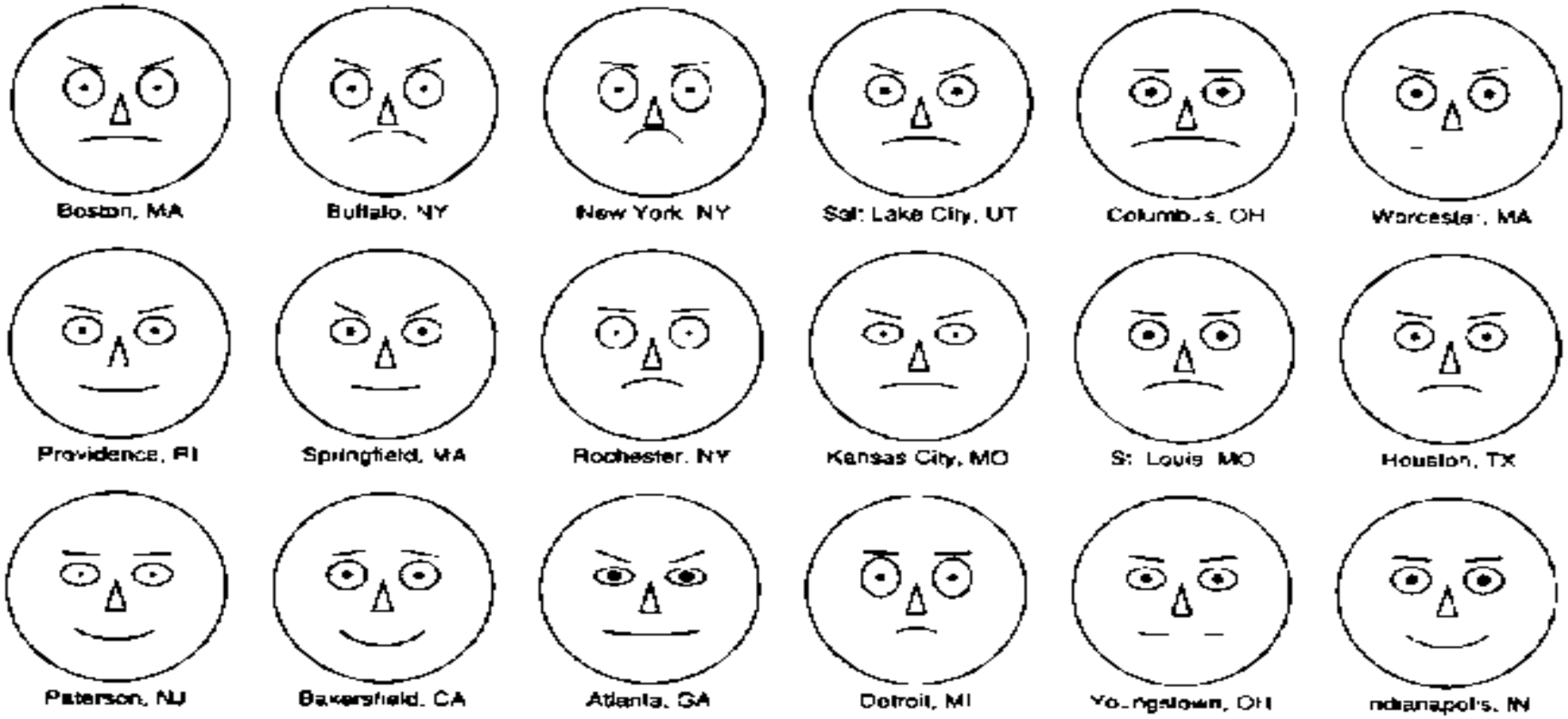






Visual Mapping

Chernoff plots



Grammar of Graphics

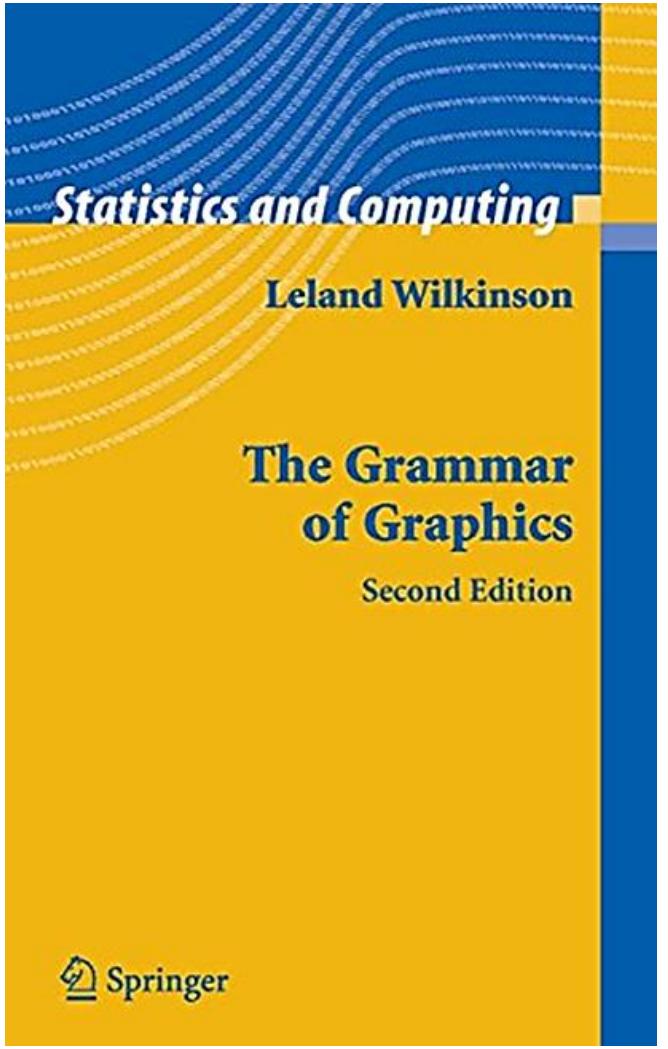
The basic idea behind the Grammar of Graphics is that a data visualization can be thought of as a **mapping** between data and visual elements.

This mapping can be broken down into layers of visual elements, where each layer represents a different aspect of the data.

Grammar of Graphics

This Grammar consists of three main components:

1. A set of visual elements or geometric objects, such as points, lines, and shapes, that can be used to represent data.



Grammar of Graphics

2. A set of aesthetic mappings, which specify how the data should be mapped to the visual elements. Aesthetics include things like color, shape, size, and position.
3. A set of scales, which define the mapping between data and aesthetics. Scales can be used to transform data from one domain to another, such as converting numbers to colors.

Grammar of Graphics

In addition to these three components, the Grammar of Graphics also includes a set of coordinate systems that define how the data should be mapped to the visual display.

Coordinate systems include things like Cartesian coordinates, polar coordinates, and geographic coordinates.

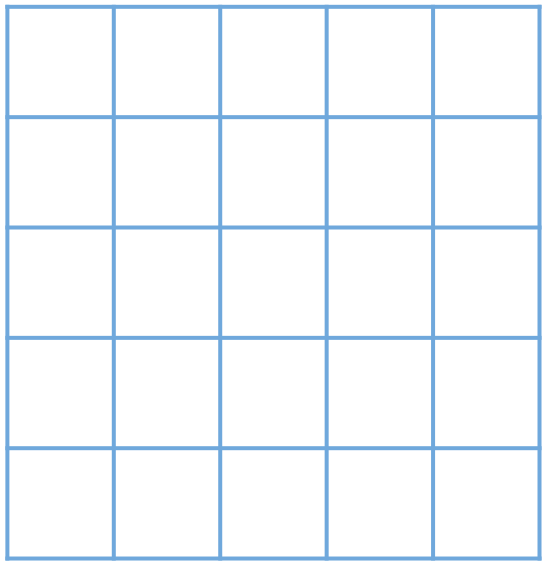
Grammar of Graphics

By breaking down a visualization into these independent components, the Grammar of Graphics allows for greater flexibility and modularity in creating complex visualizations.

It also provides a common language for communicating about and comparing different visualizations.

Grammar of Graphics

Dataset(s) Mapped (straight or using stat ansalyses) into **visible attributes** pertaining to **geometric objects**



- **2D or 3D coordinates**
- **Color (hue, brightness)**
- **Texture**
- **Shape, size, orientation**

pertaining to

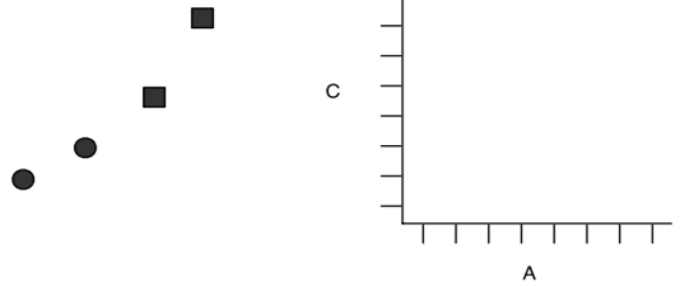
- **Scatterpoints**
- **Histograms**
- **Line, bar, pie charts**
- **Many other**

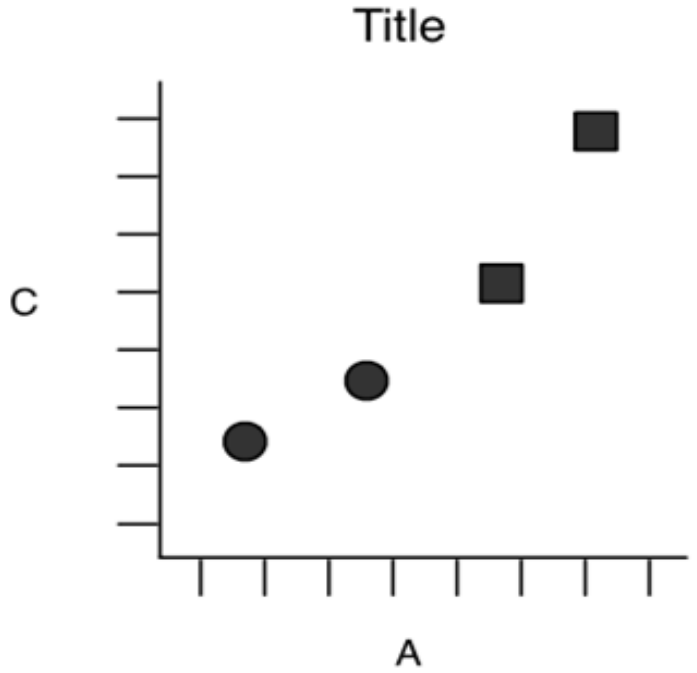
Grammar of Graphics

A	B	C	D
2	3	4	a
1	2	1	a
4	5	15	b
9	10	80	b

x	y	Shape
2	4	a
1	1	a
4	15	b
9	80	b

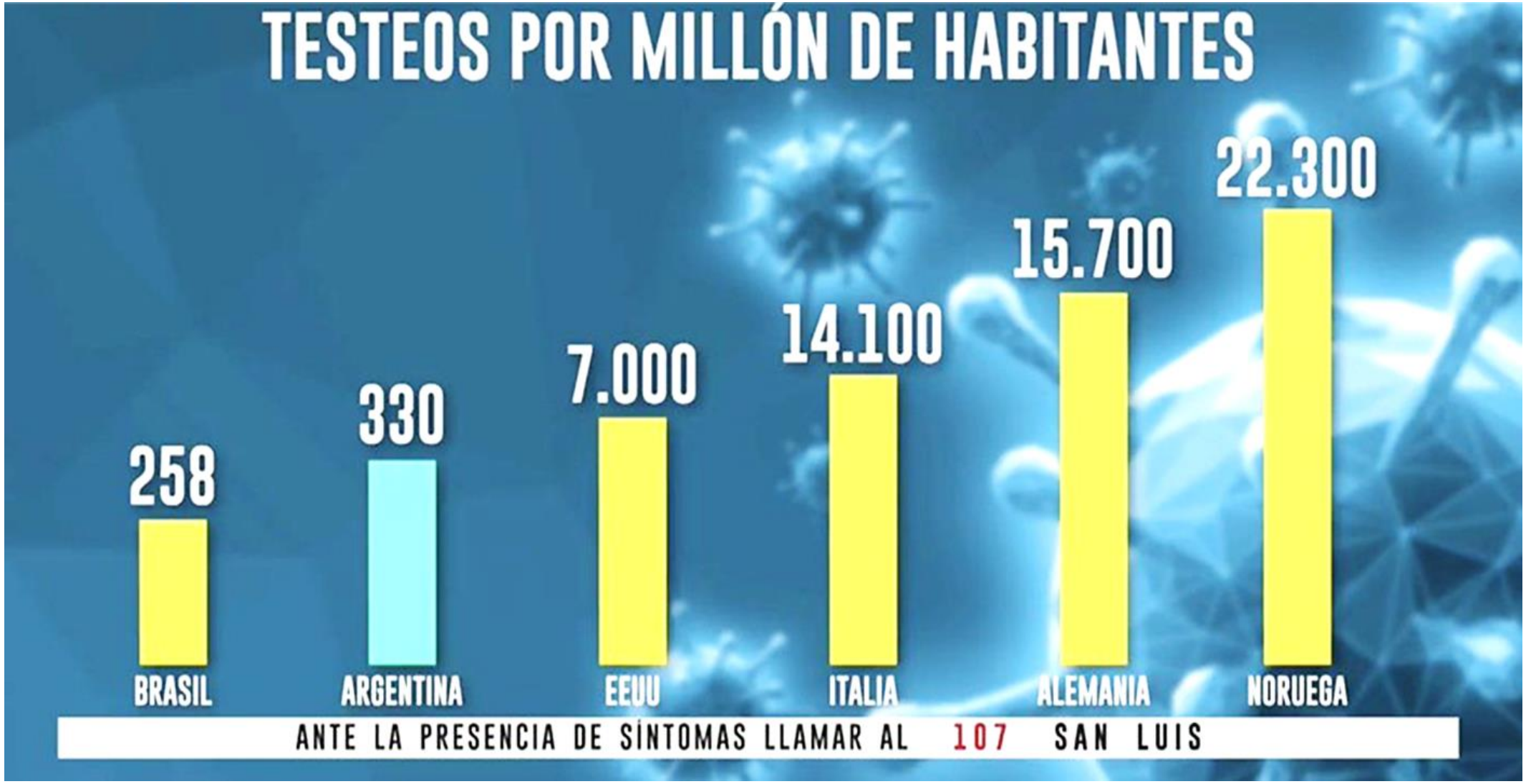
x	y	Shape
25	11	circle
0	0	circle
75	53	square
200	300	square



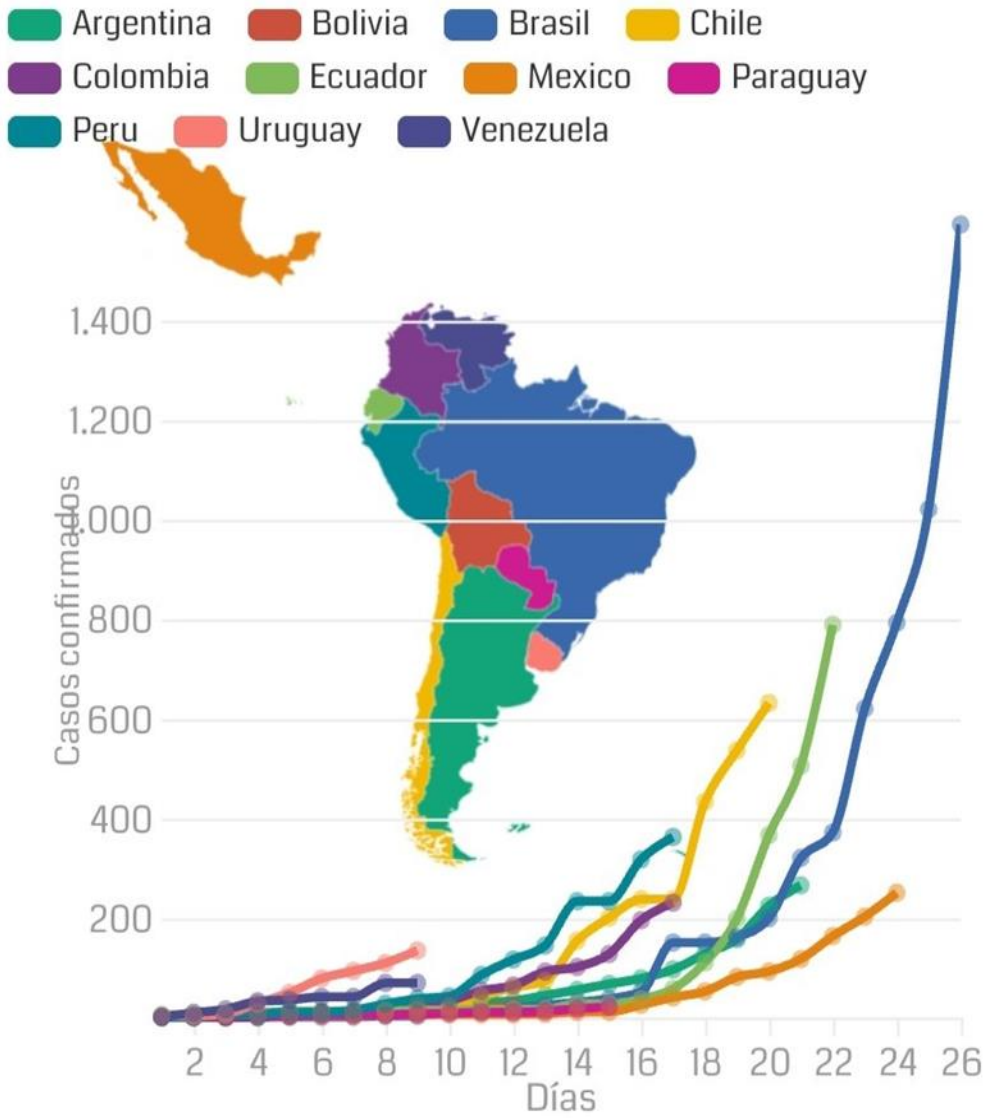


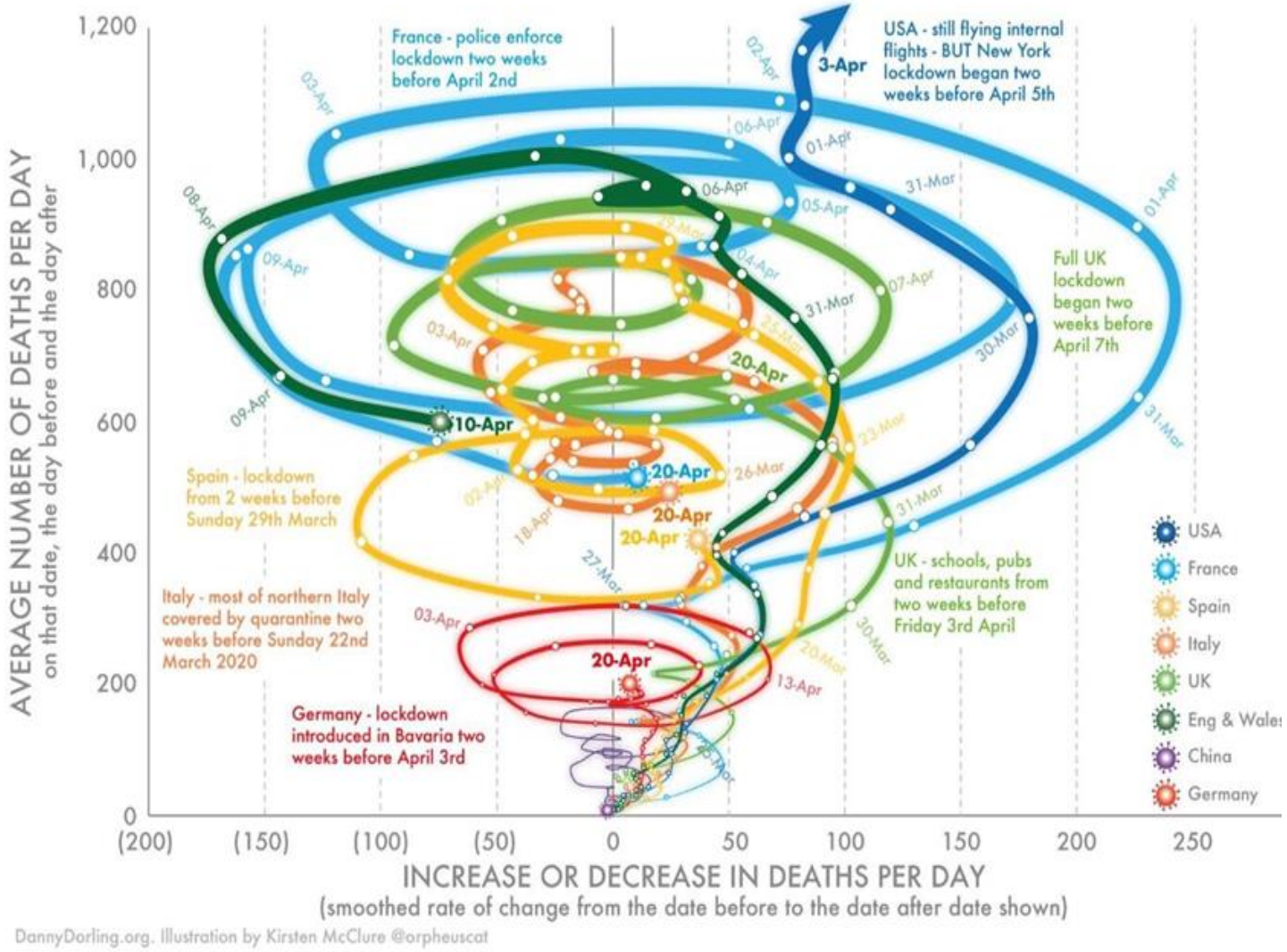
$$\text{floor}\left(\frac{x - \min(x)}{\text{range}(x)} * \text{screen width}\right).$$

How many ways to design a bad visualization?

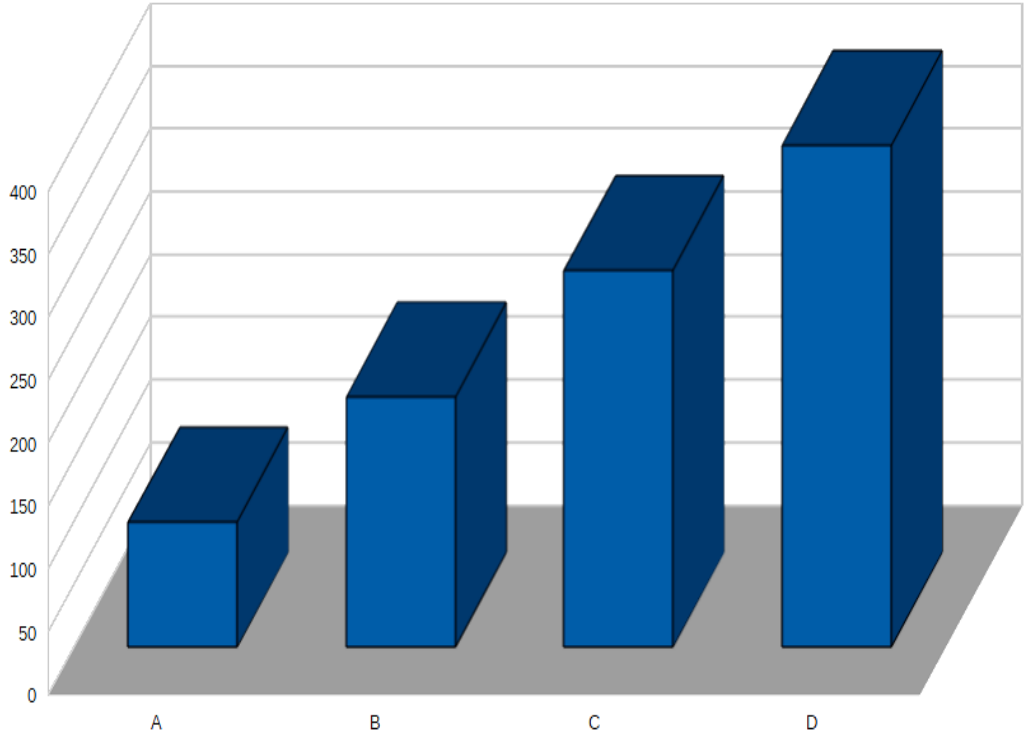


How many ways to design a bad visualization?



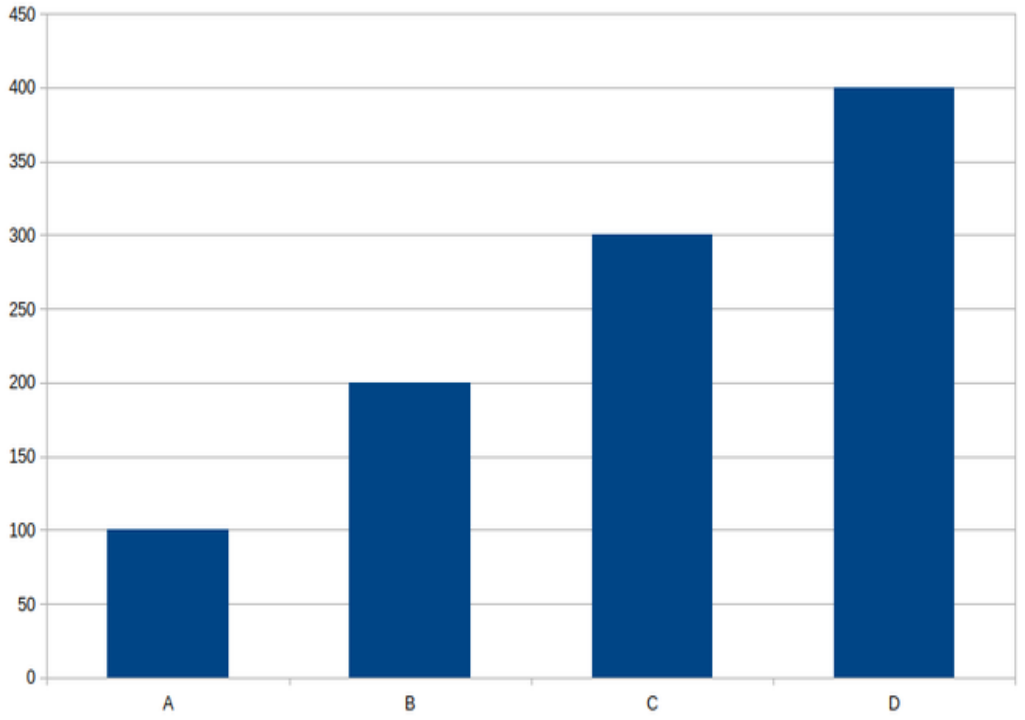


How many ways to design a bad visualization?



150	3 %
300	**84,8 %**
250	8,1 %
200	4,2 %

335 votos · Resultados finales



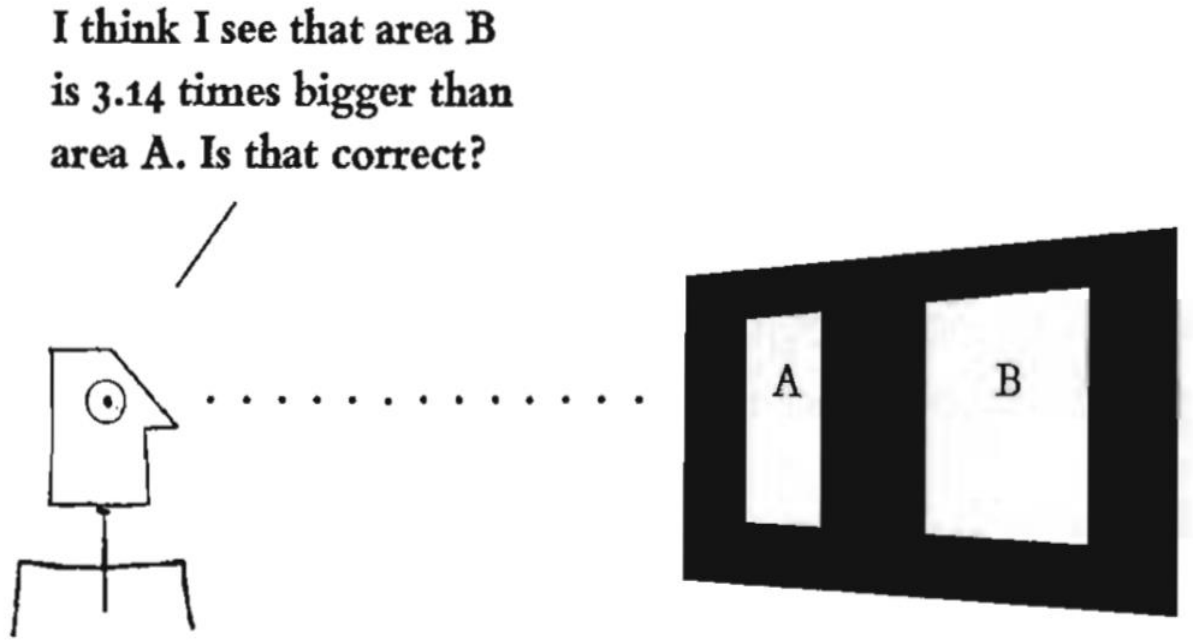
300	**96,9 %**
200	1,5 %
50	0,8 %
100	0,8 %

130 votos · Resultados finales

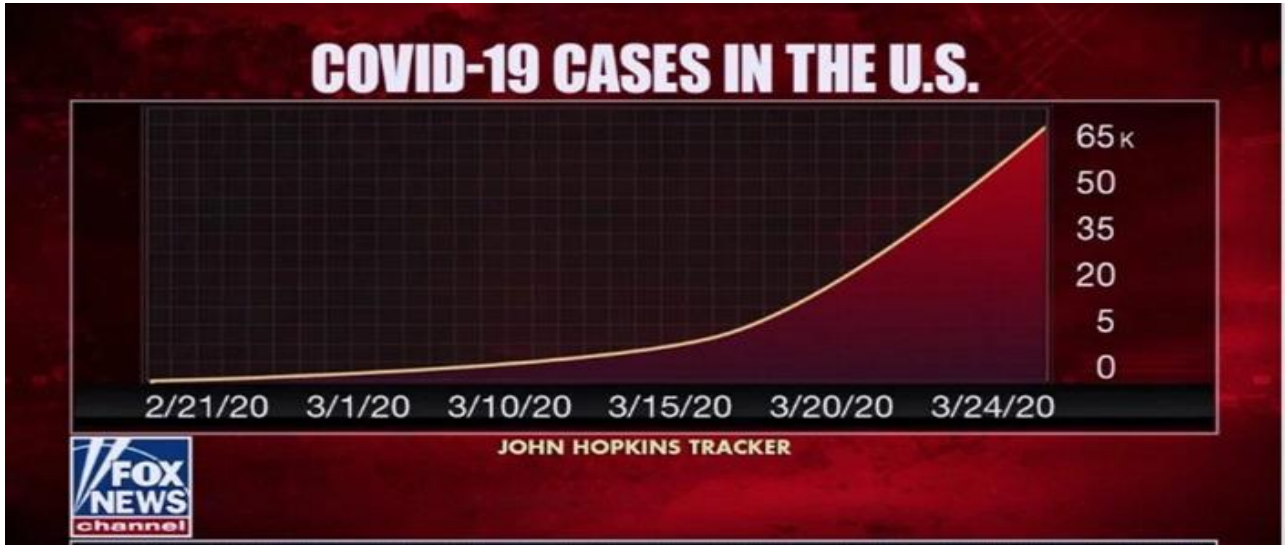
Integrity Principles

What constitutes **distortion** in a visualization?

According to Tufte, a visualization is distorted if the visual information presented is not faithful representation of the actual data.



Integrity Principles



Integrity Principles

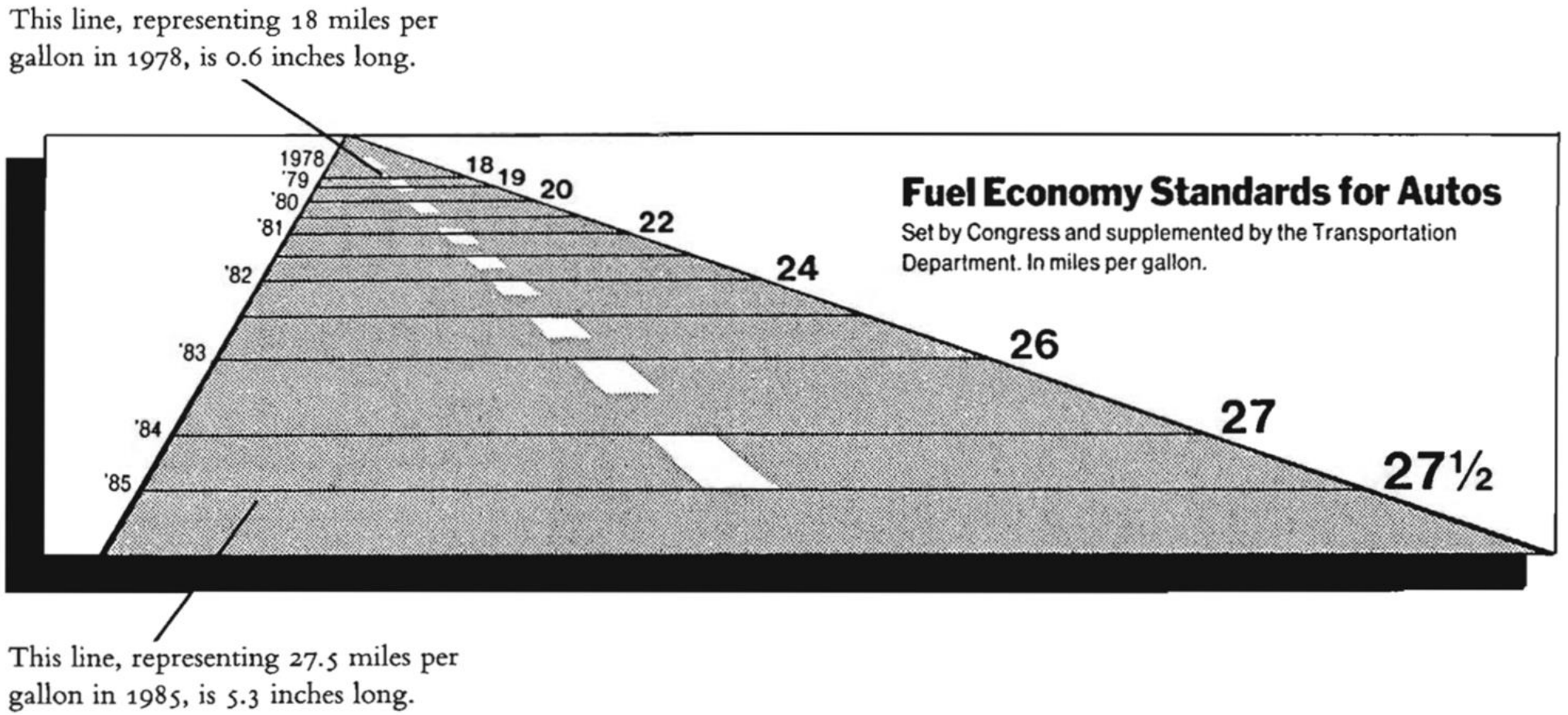
At least, qraphic representation of numeric quantities should be proportional to their actual graphic representation.

Also, there should be labels, ticks and scales to help diminish ambiguieties or errors.

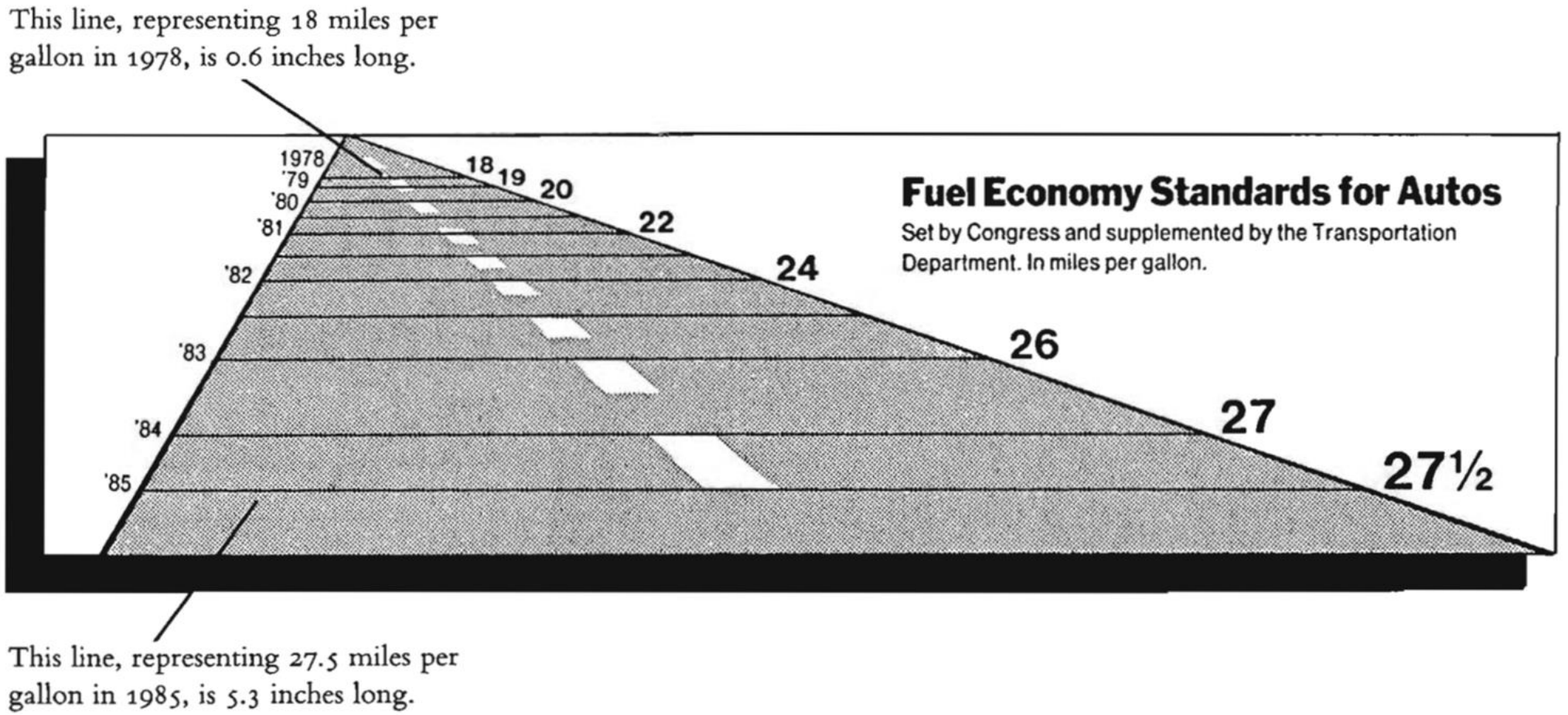
$$\text{Lie factor} = \frac{\text{quantities/proportions in the view}}{\text{quantities/proportions in data}}$$

A LF over 1.1 (say) is overestimating quantities/proportions, and the other way around.

Lie factor estimation

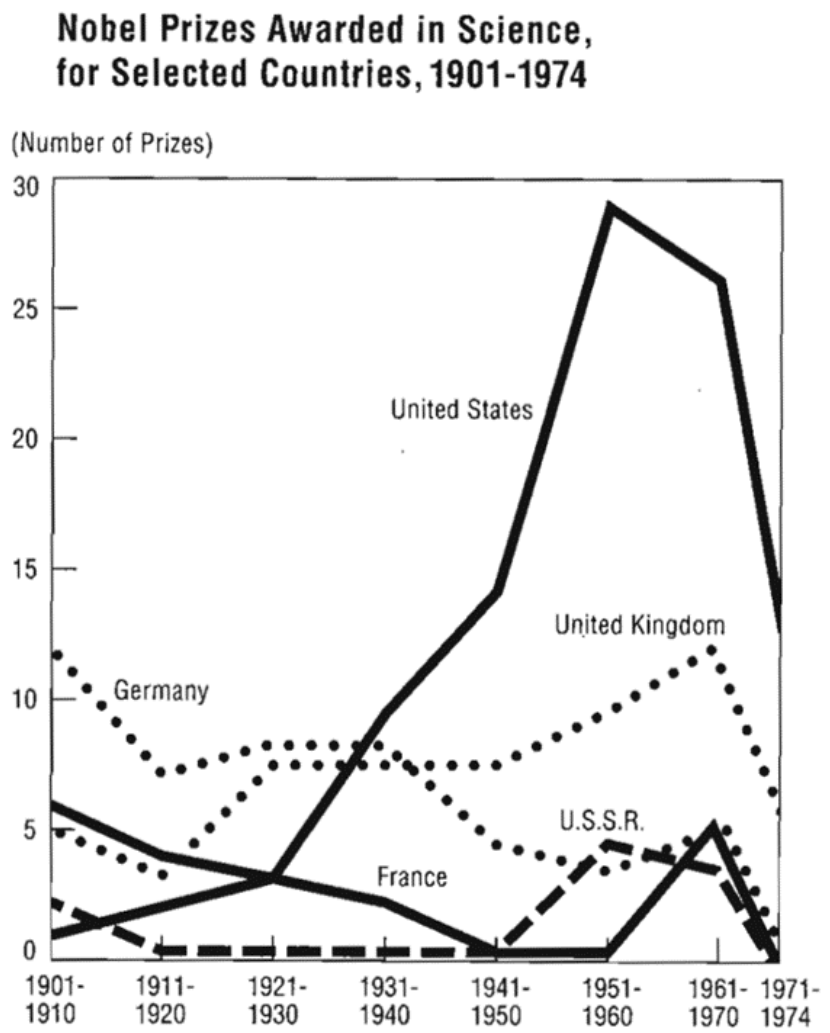


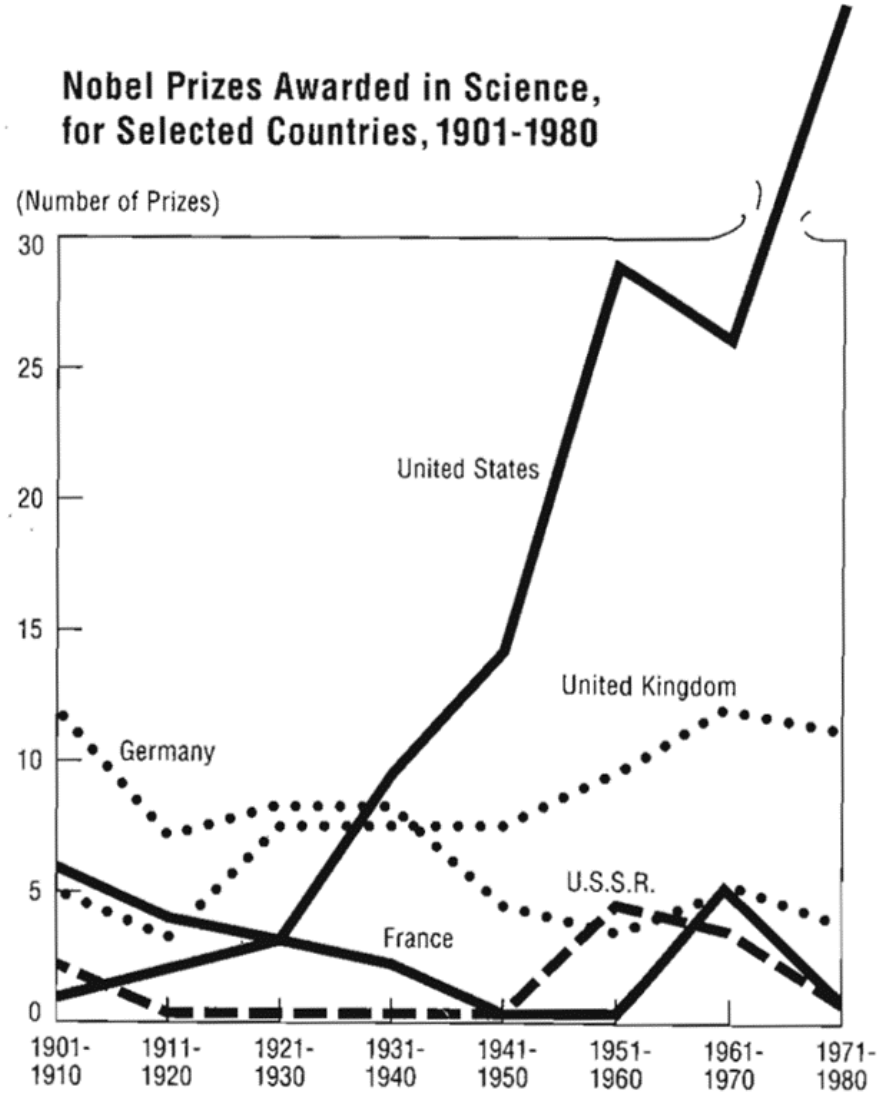
Lie factor estimation



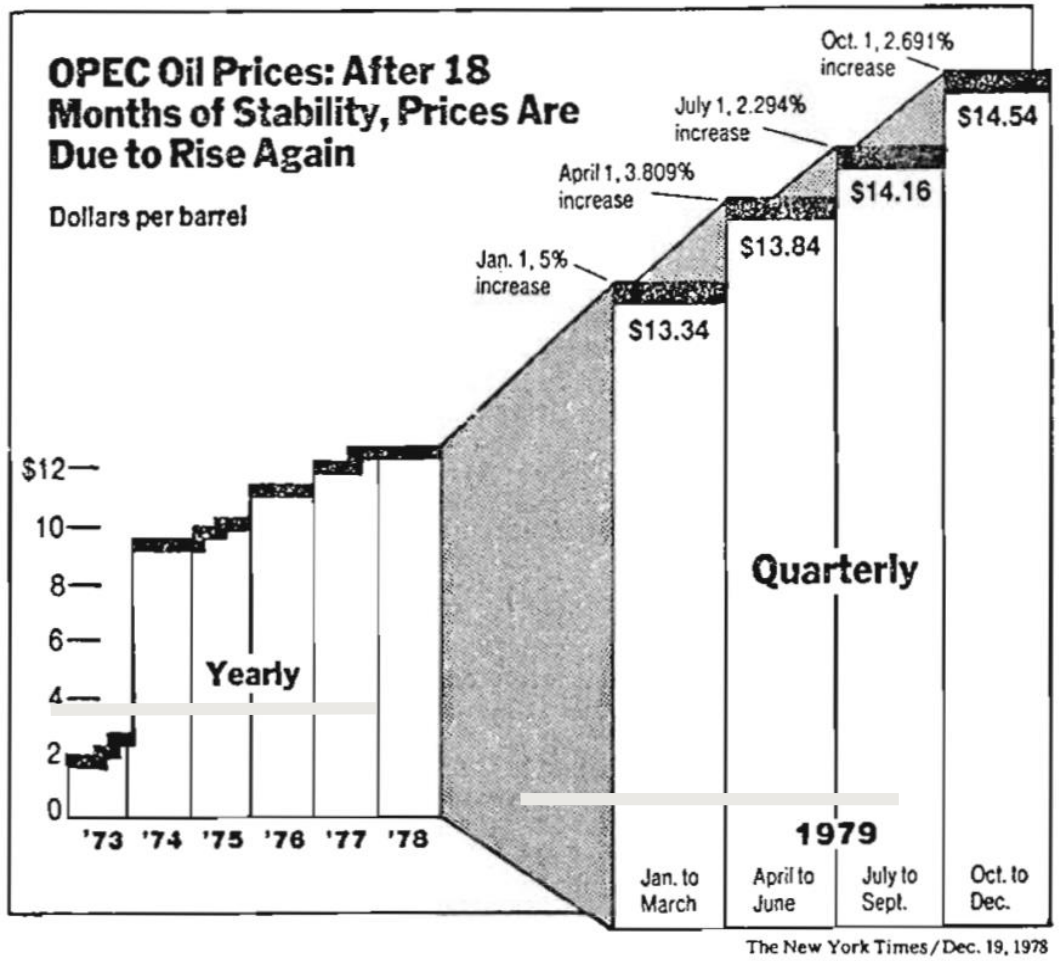
$$\frac{27.5 - 18.0}{18.0} \times 100 = 53\% \qquad \frac{5.3 - 0.6}{0.3} \times 100 = 783\% \qquad \frac{783}{53} = 14.8$$

Other ways to lie





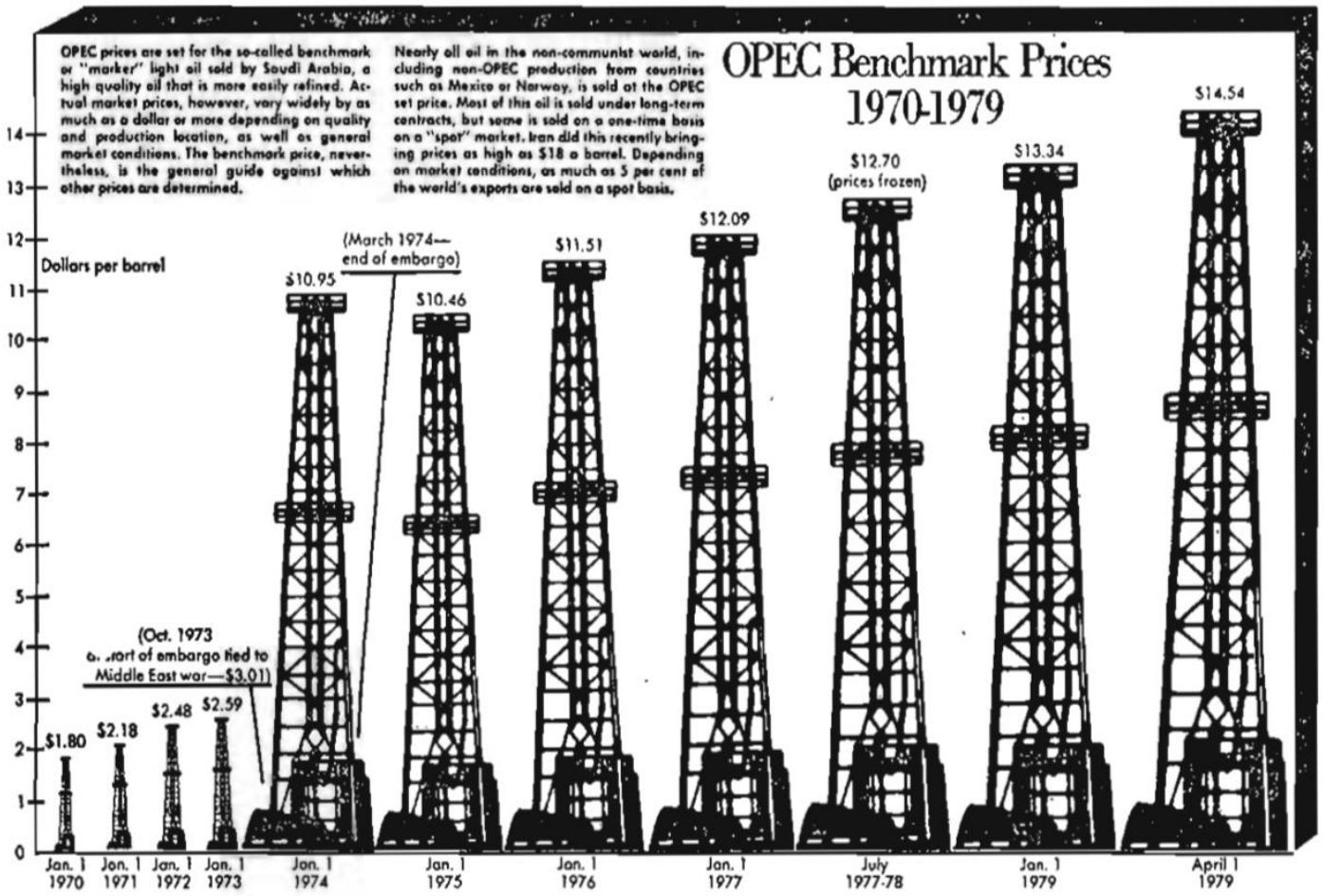
Other ways to lie

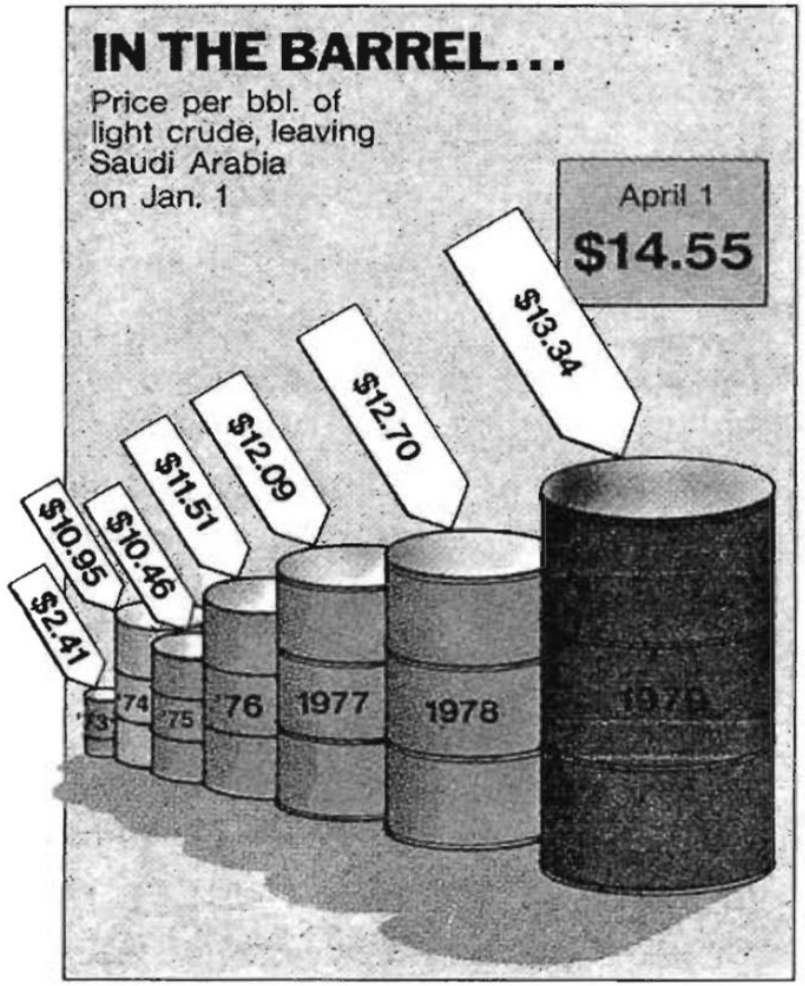


During this time	one vertical inch equals
1973–1978	$8.00
January–March 1979	$4.73
April–June 1979	$4.37
July–September 1979	$4.16
October–December 1979	$3.92

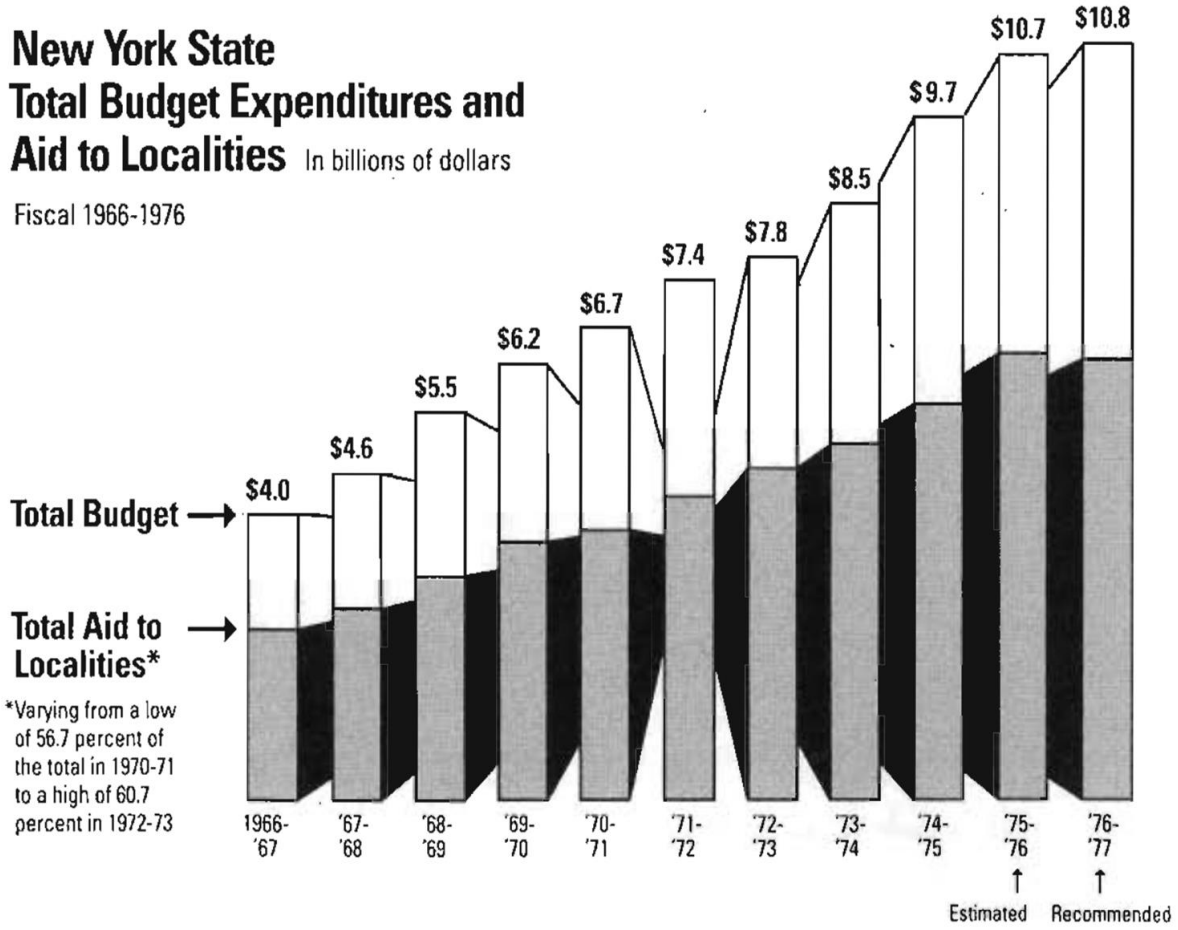
During this time	one horizontal inch equals
1973–1978	3.8 years
1979	0.57 years

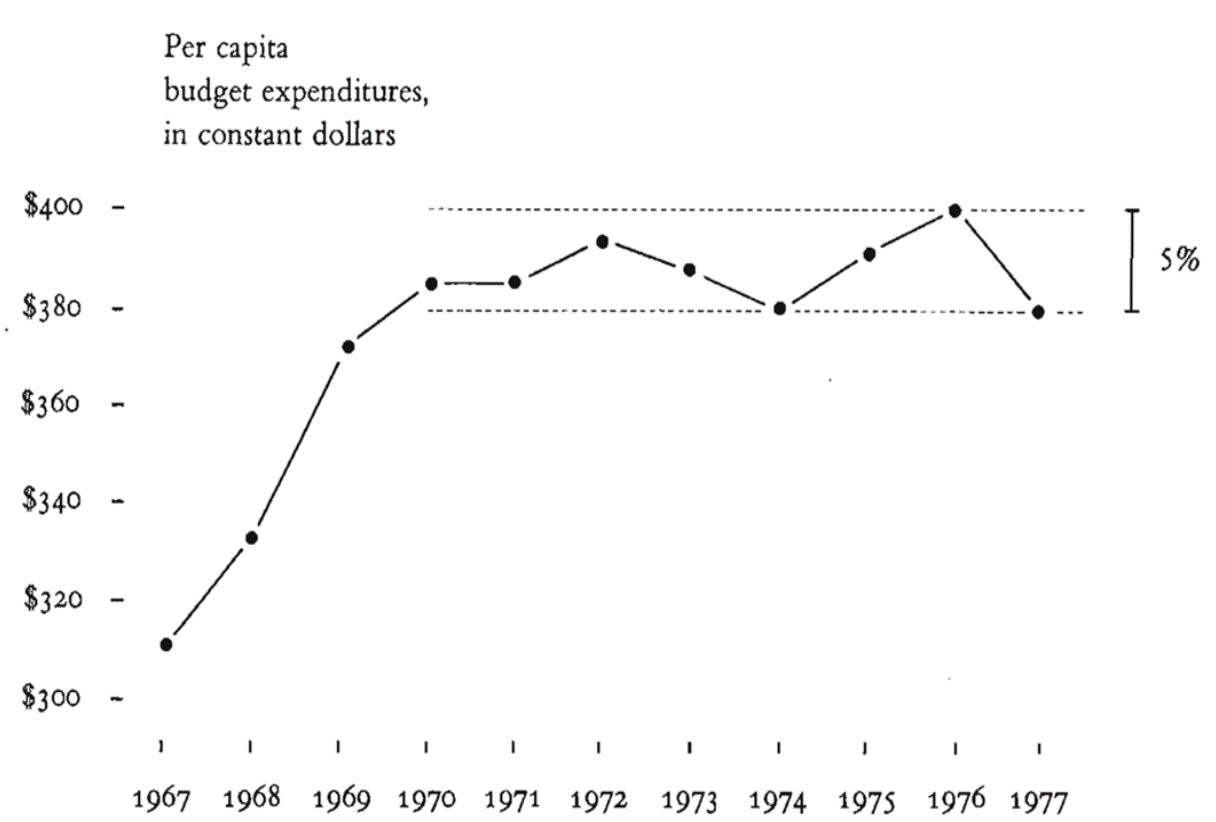
Other ways to lie



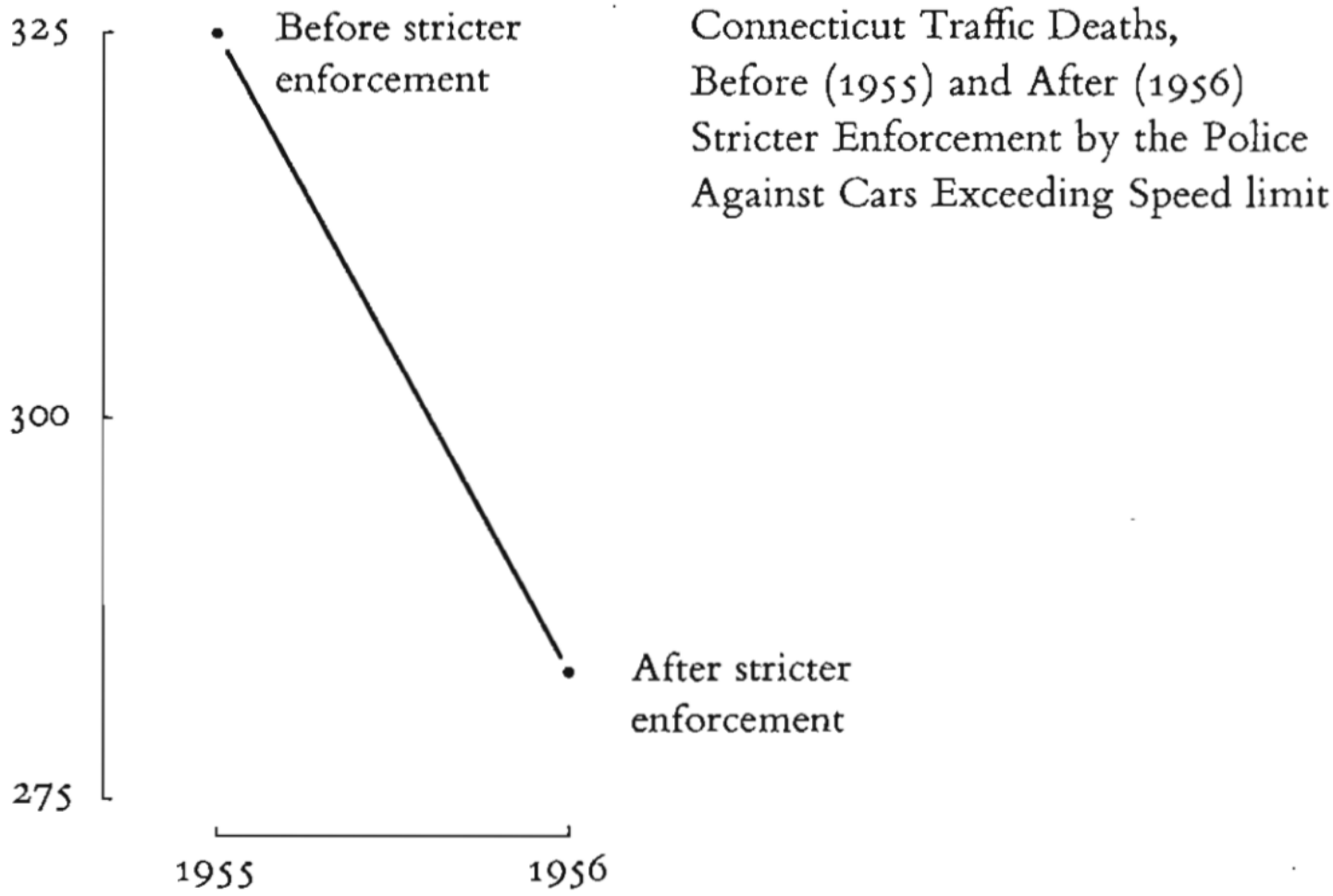


Adjusting to context

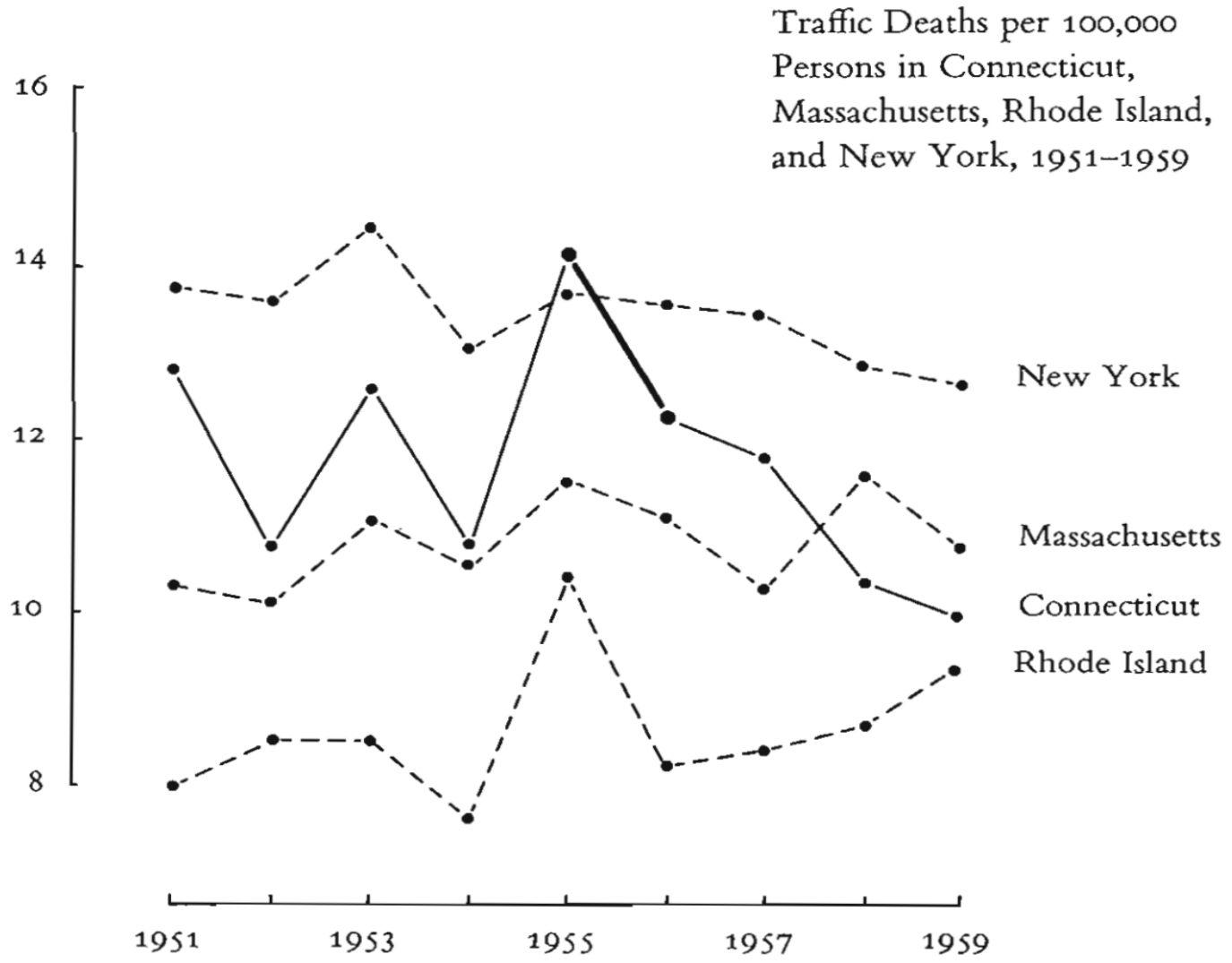




Adjusting to context

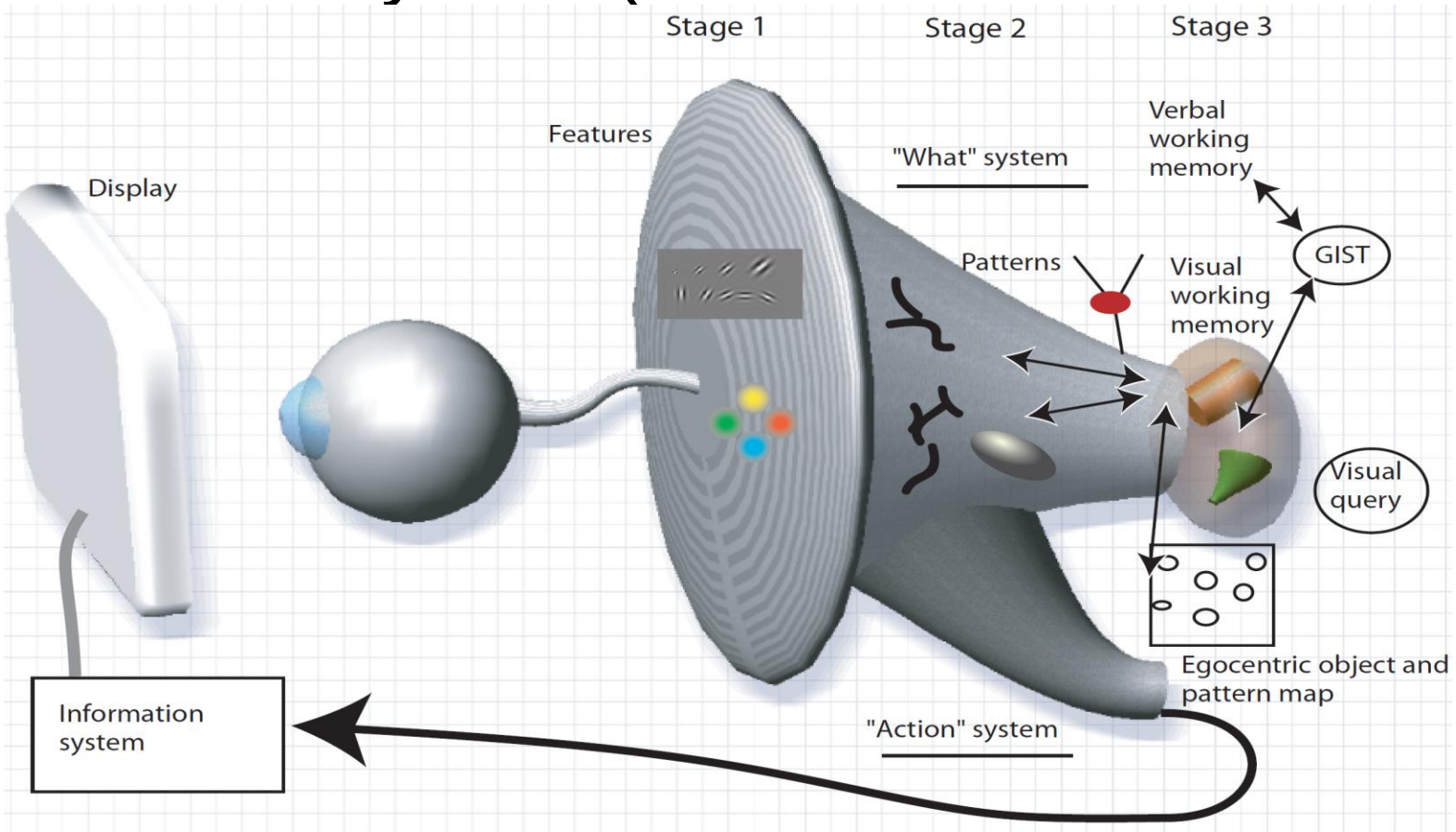


Adjusting to context



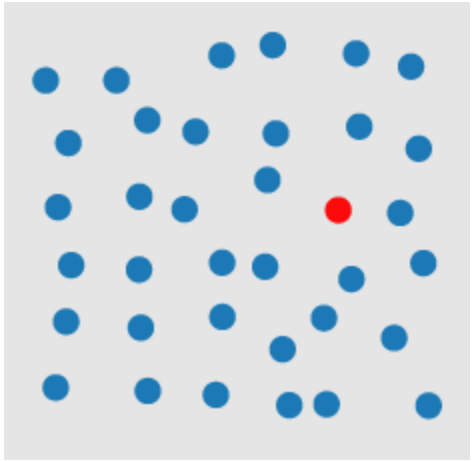
Perception in Visualization

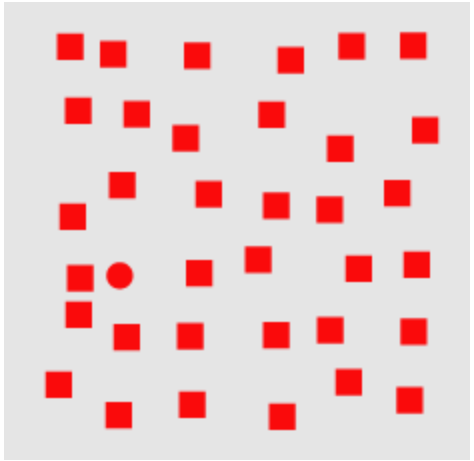
In visualization we intend to take advantage of the cognitive power of the human visual system (which we have to understand).



Preatentive features

You cannot help noticing them.

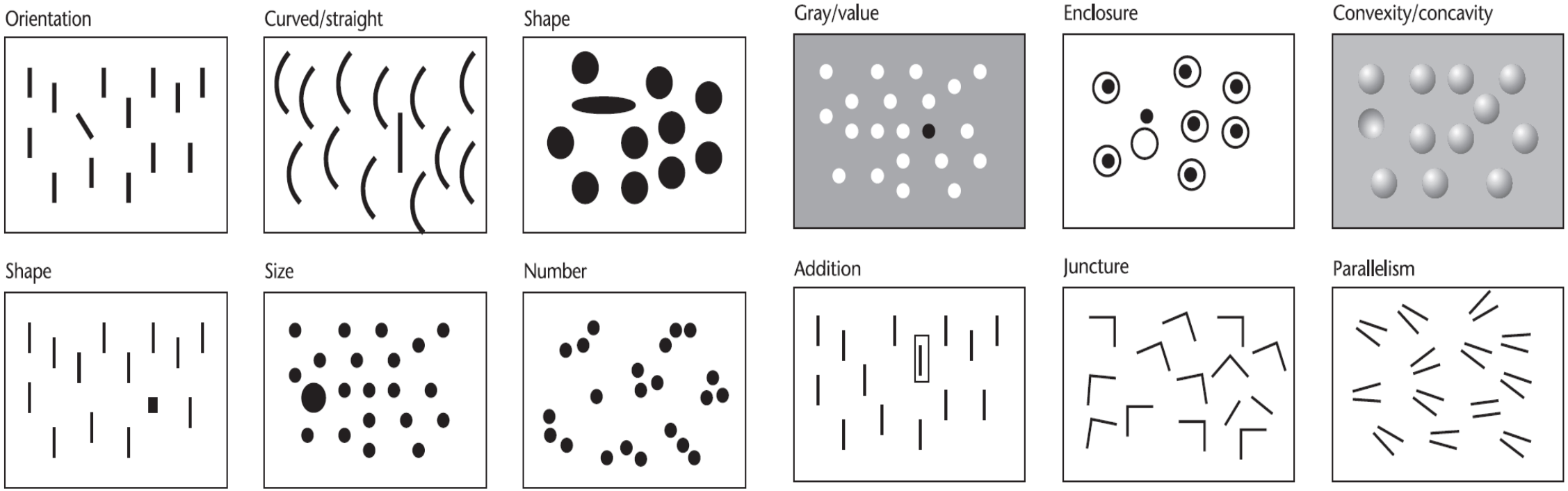




1281768756138976546984506985604982826762
9809858458224509856458945098450980943585
9091030209905959595772564675050678904567
8845789809821677654876364908560912949686
0985845822450985645894509845098094358598

Preatentive features

Several preatentive features, but cannot be used together.

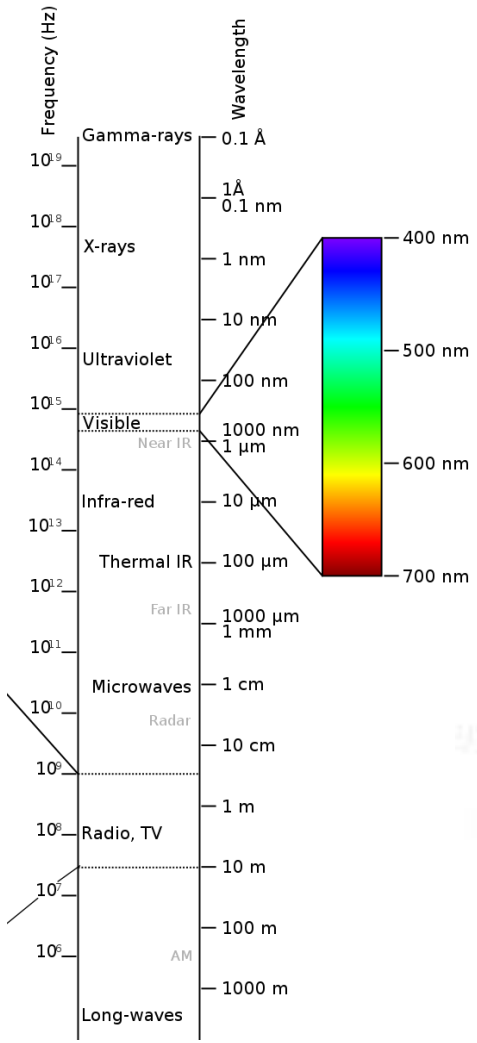


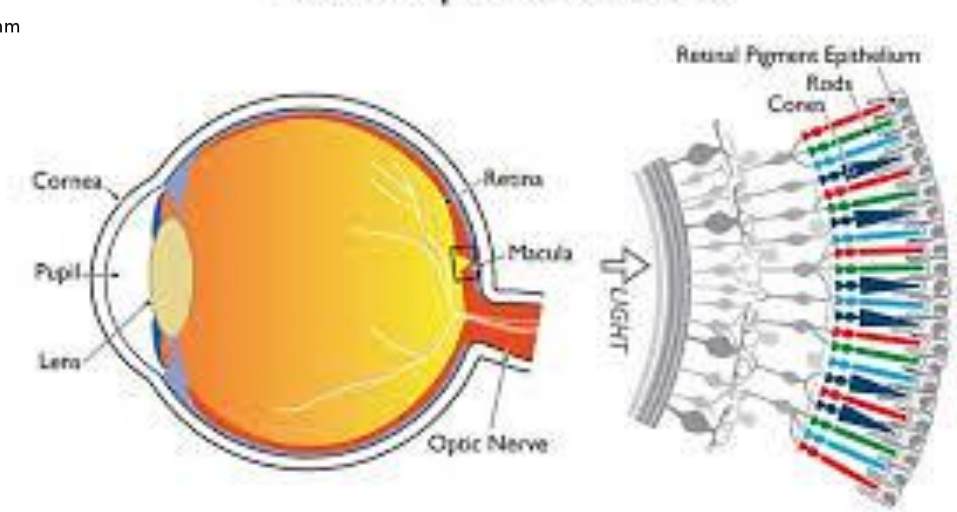
Color and color maps

Color is a perceptual feature (does not exist “outside there”).

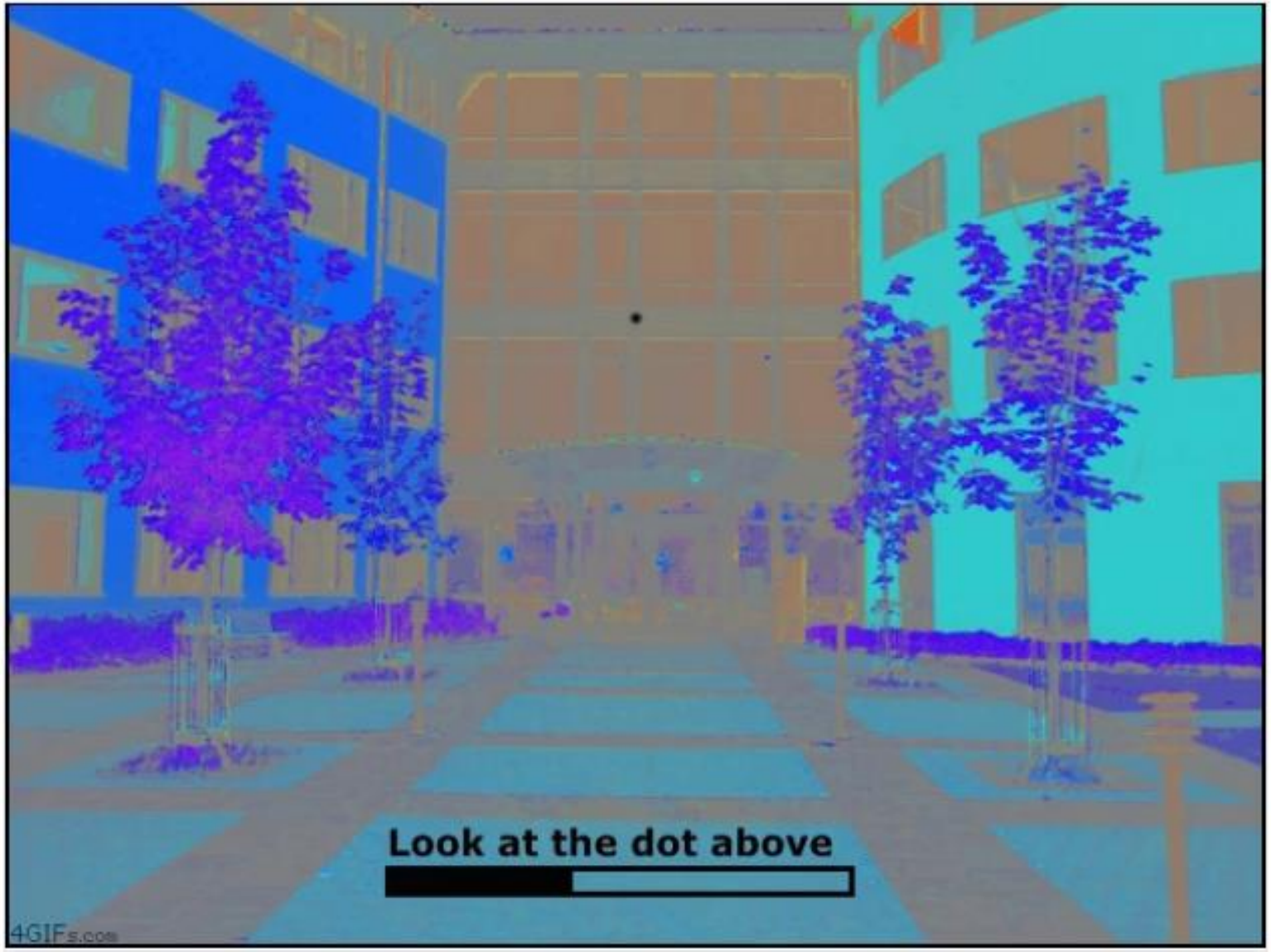
Depends strongly on human physiology.

Three receptors, hence our color space is 3D.





The brain tells a lot to the eyes!



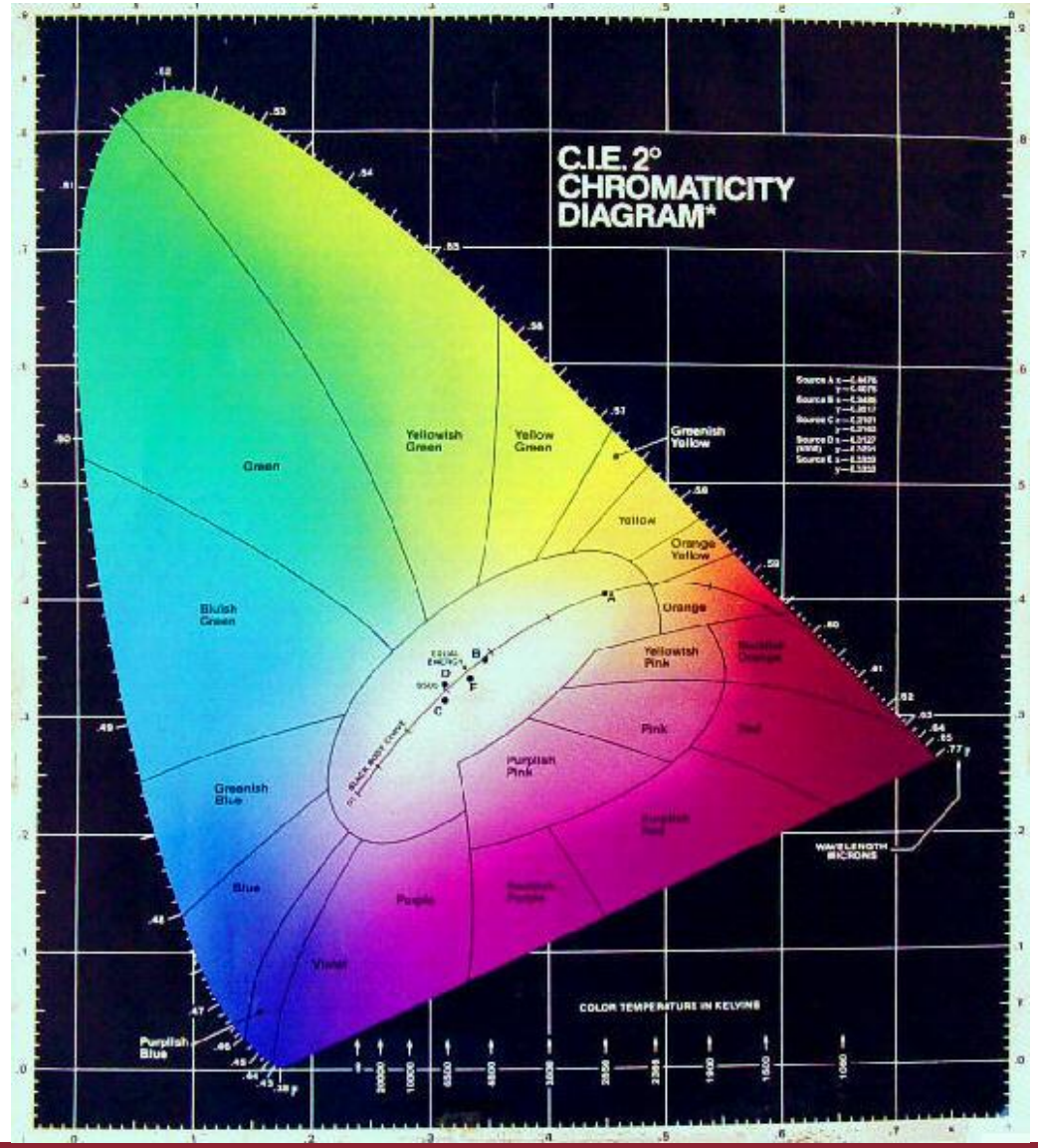
The brain tells a lot to the eyes!

Color and color maps

CIE "chromaticity" diagram organizes color stimuli regarding the relative activation of the three retina cones given a spectral distribution.

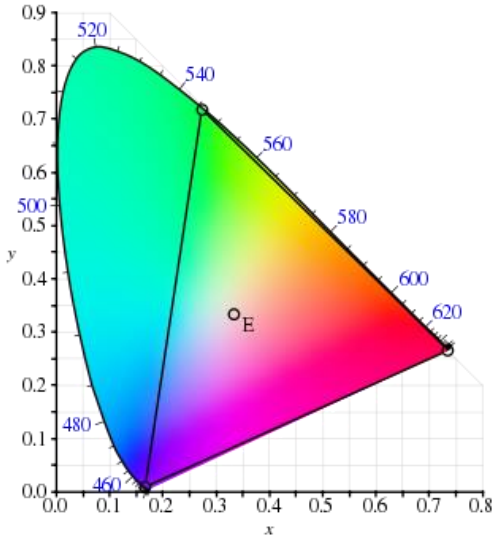
Chromaticity = (Hue, Saturation)

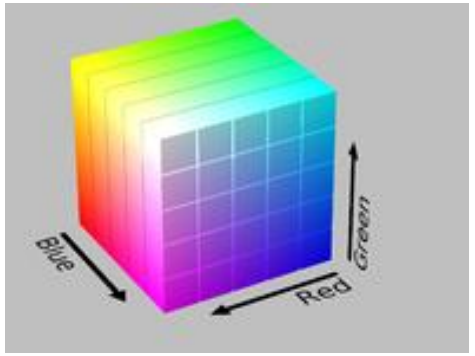
Stimuli = (Hue, Sat., Luminance)

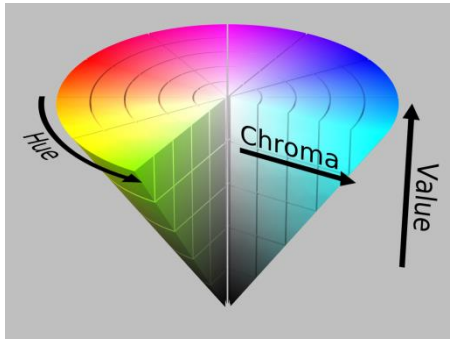


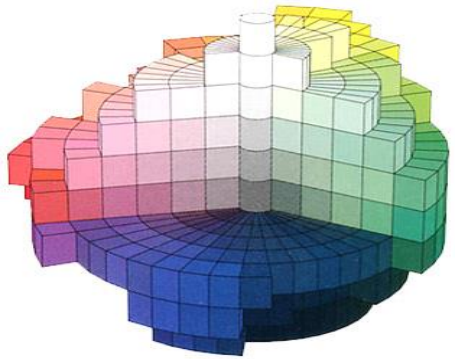
Color and color maps

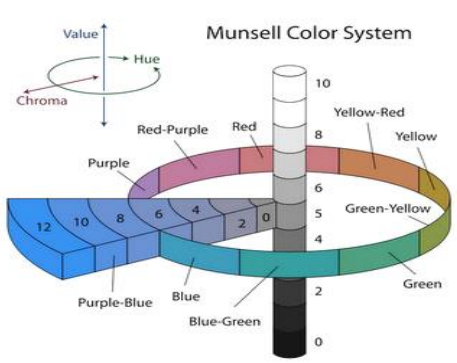
However, receptors don’t work independently, color perception is more complex and thus RGB space is not always adequate.

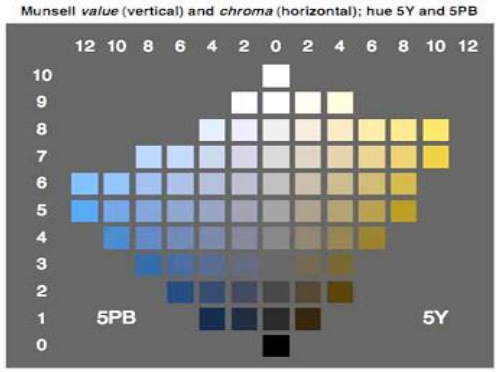




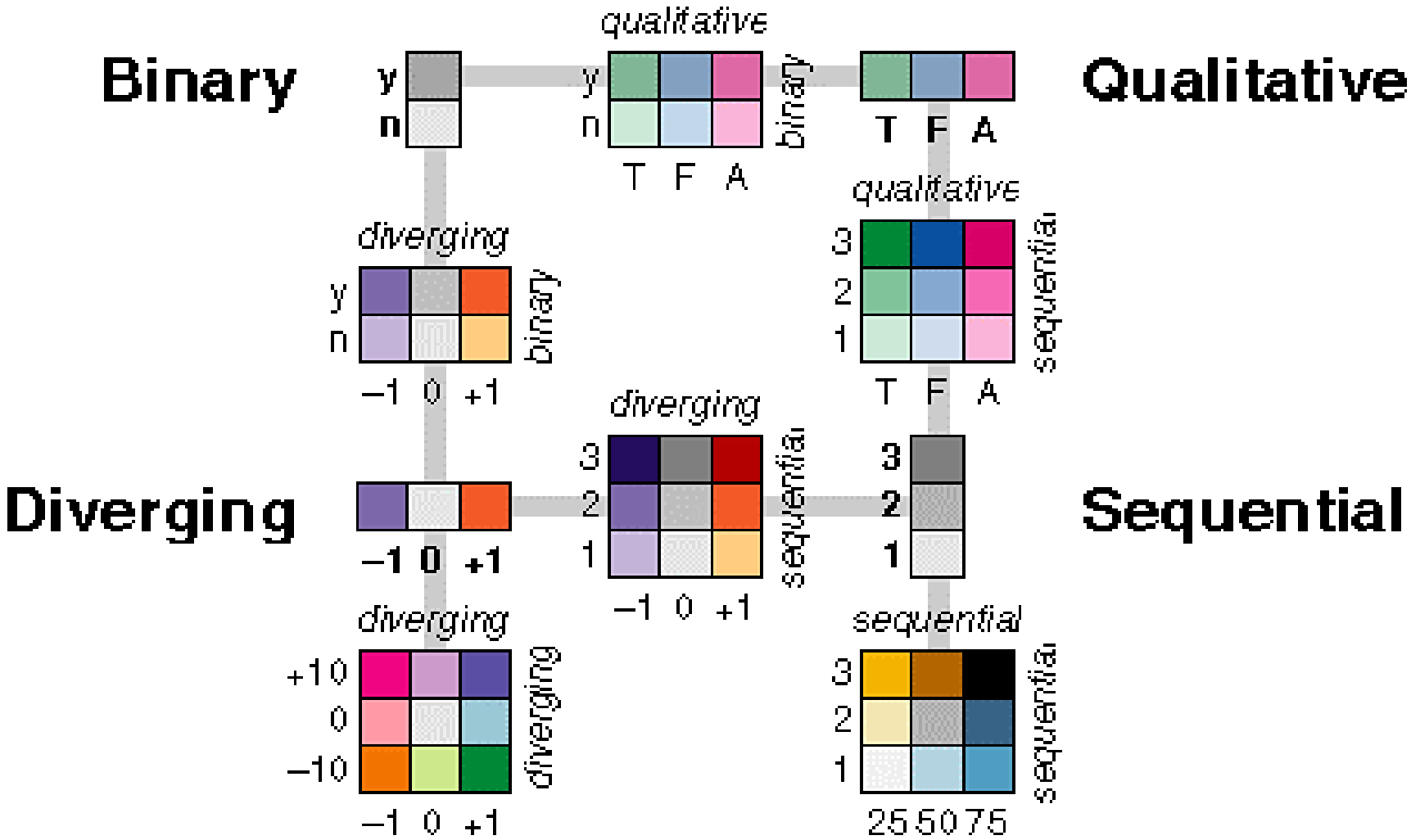






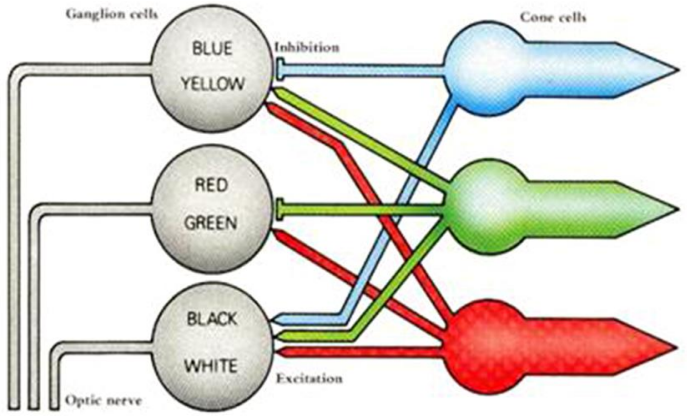


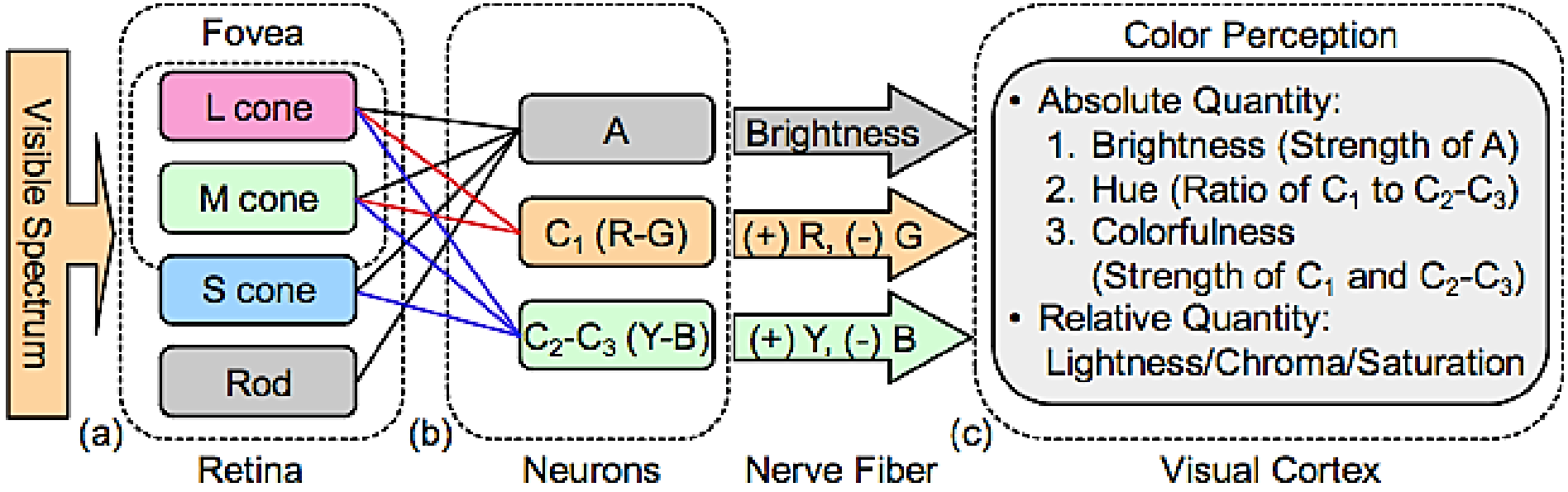
Color maps: Cynthia Brewer



Color maps: human perception

The "color" information of the receptors is processed in the retina in "opponent channels".

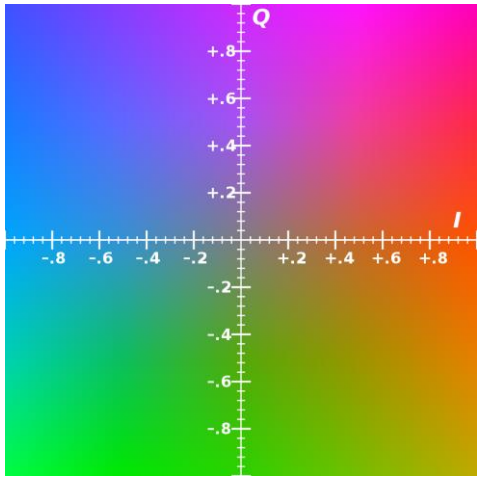


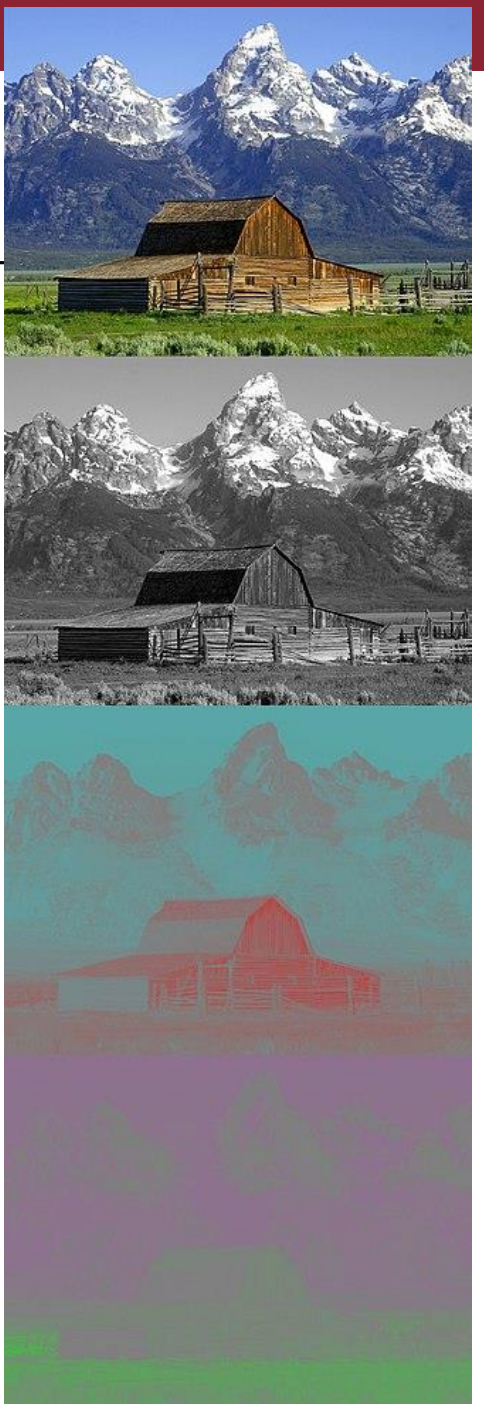


Color maps: human perception

This representation is best captured in the YIQ color space. Each opponent color has different importance in the attentional/perceptual field: Y gets 80% importance, I 15% and Q only 5%.

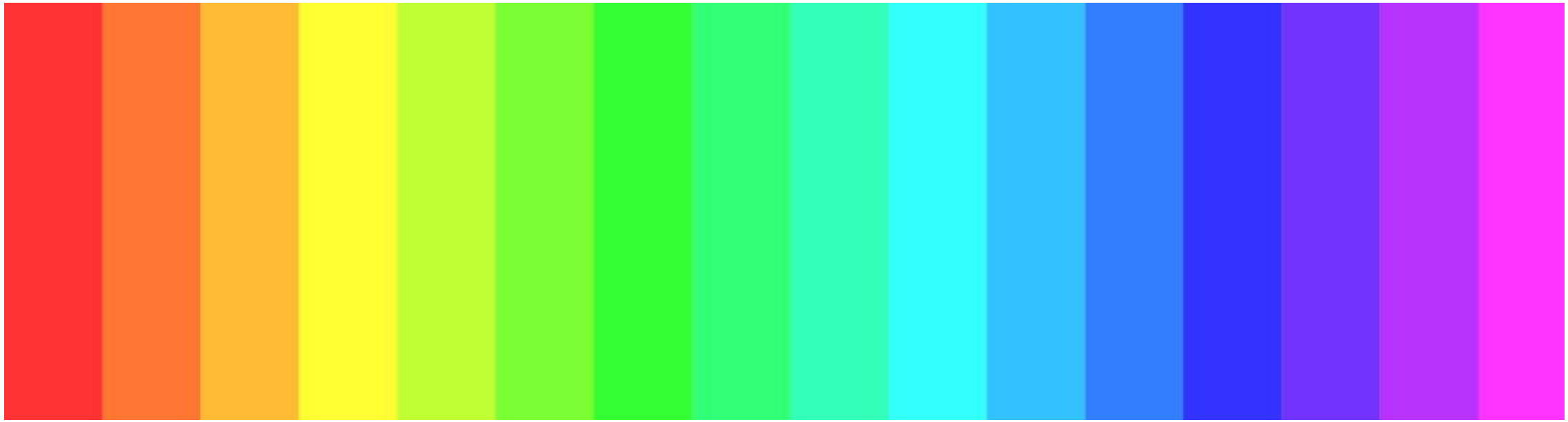
$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.311135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$





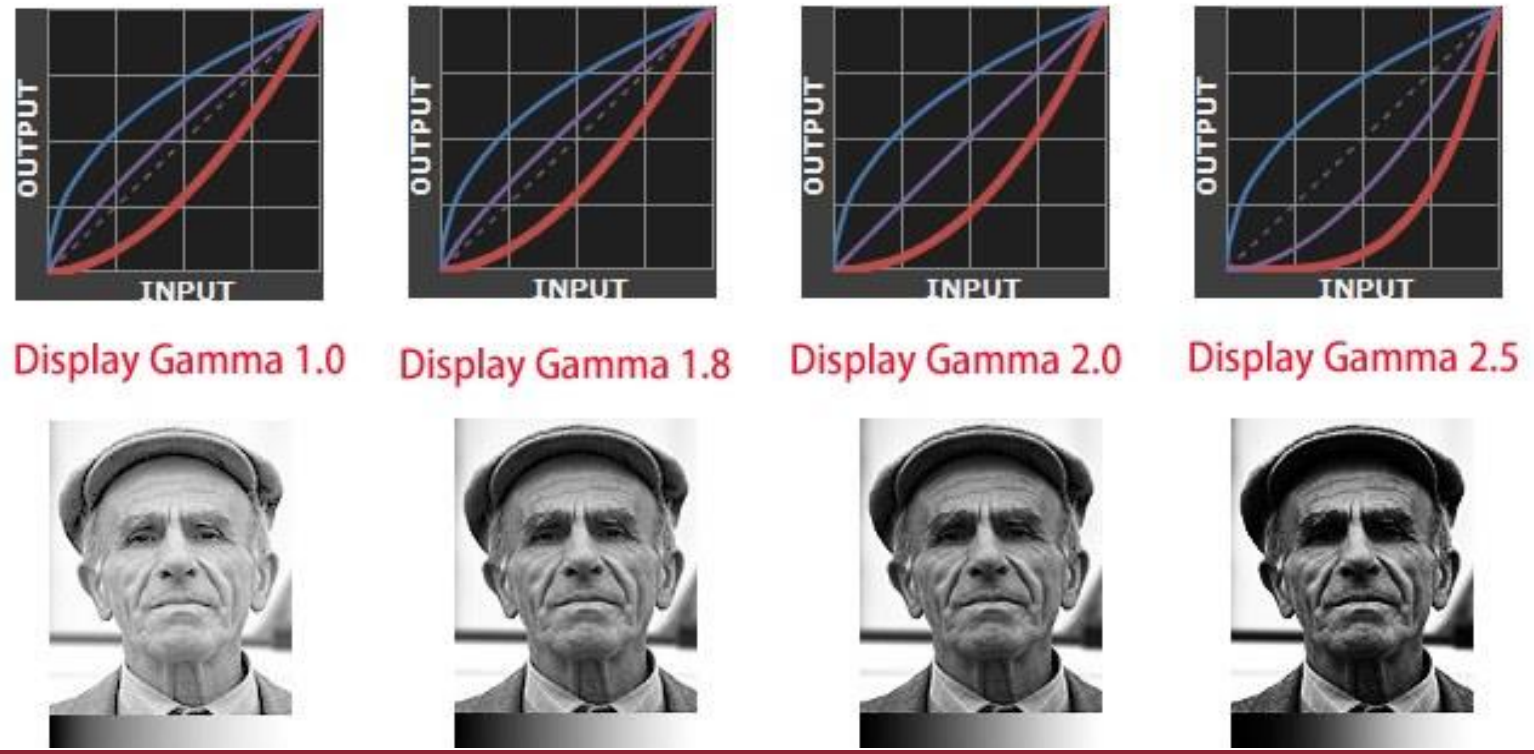
Color maps: human perception

As a consequence, human hue perception is not uniform. Also the incidence of the three RGB primaries in the perceived luminance is uneven.



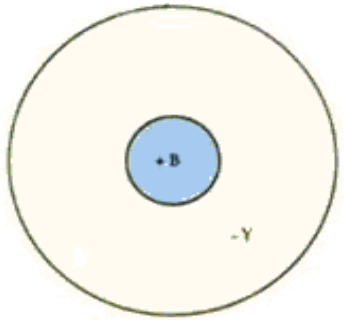
Color maps: human perception

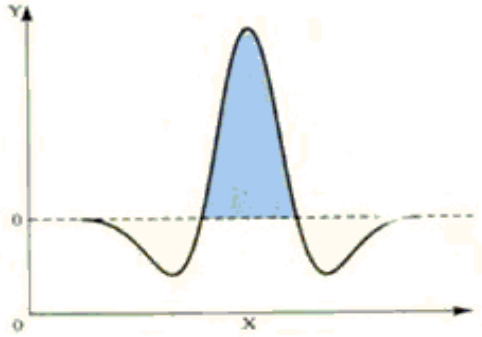
Another factor to take into account is the fact that information in the neural circuits is represented using logarithms. In image processing this is taken into account with the "gamma correction" that exponentiates the encoded brightness in the video signal.

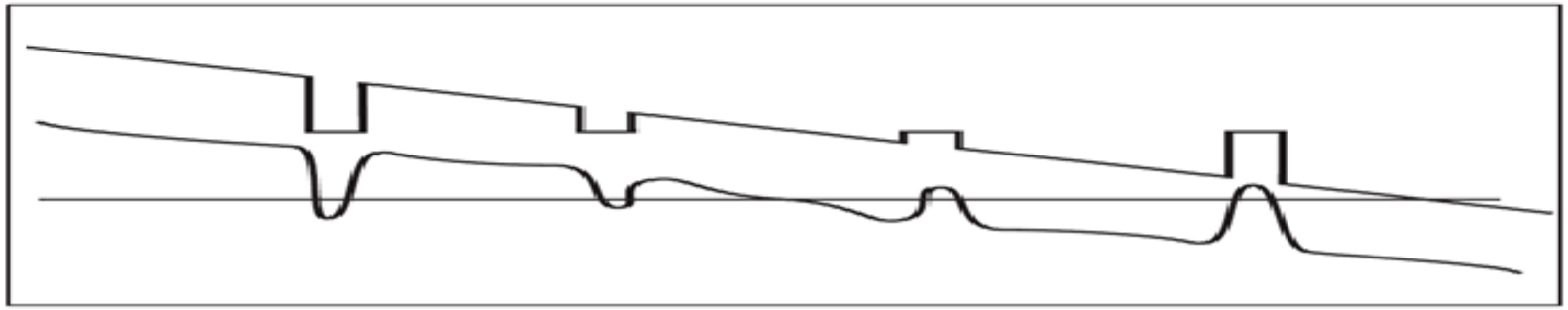


Color maps: human perception

This would entail a "fine detail" perceptive loss, thus the visual system enhances tiny local changes applying a "simultaneous contrast" process.

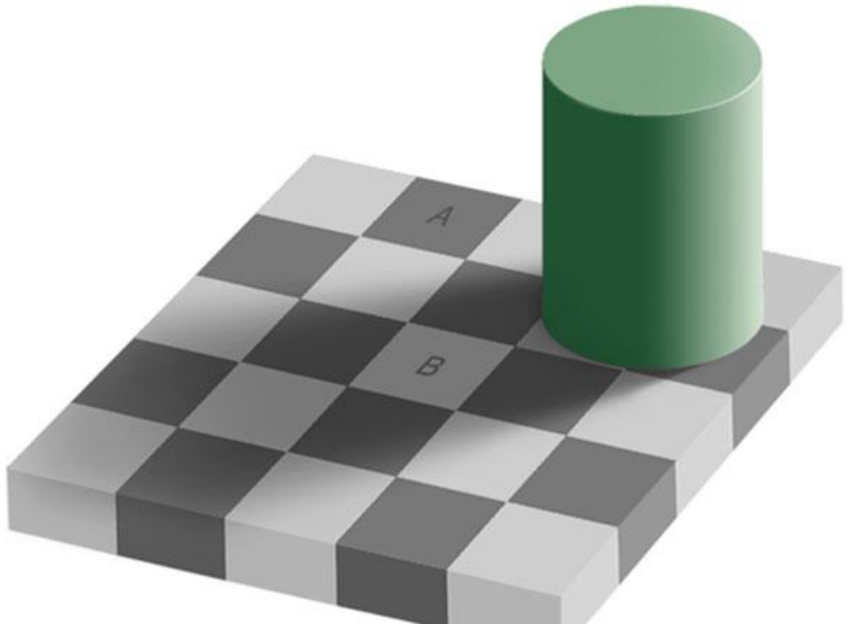


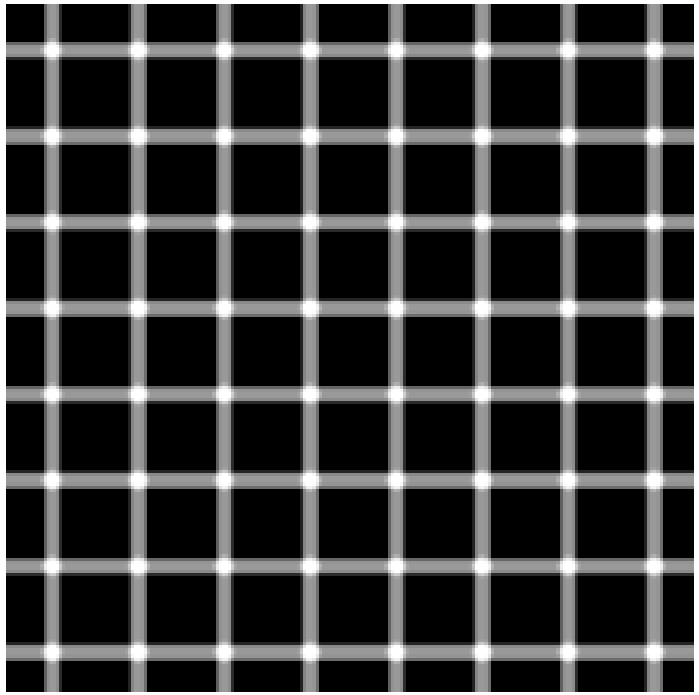


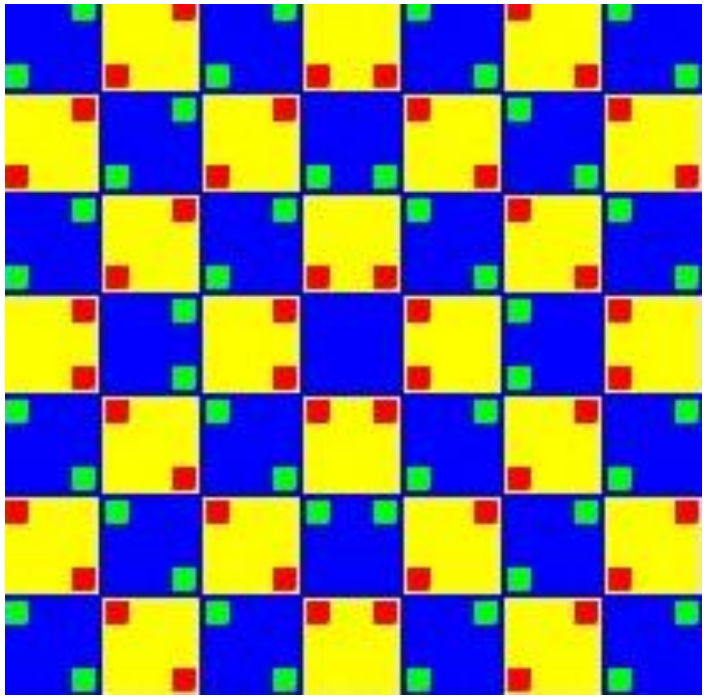


Color maps: human perception

This fact raises several perceptual illusions and paradoxes

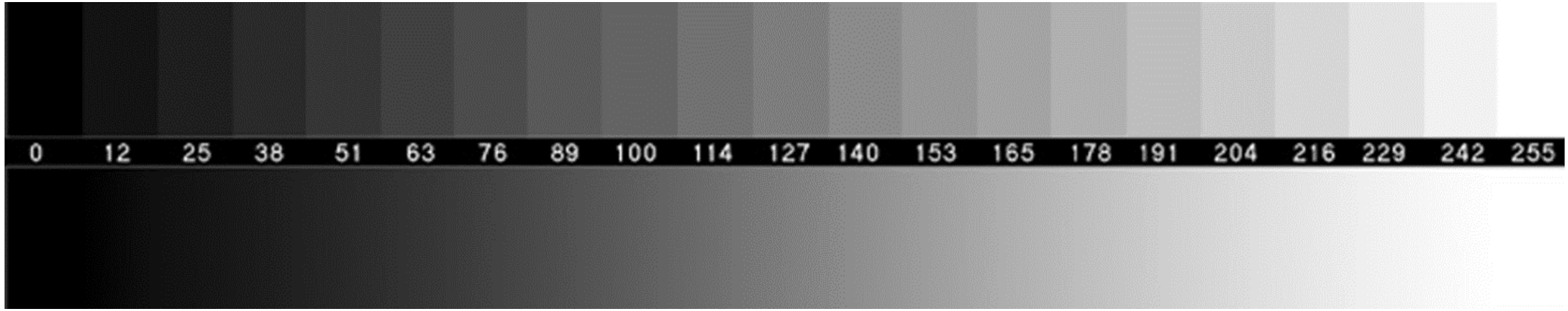






Color maps: human perception

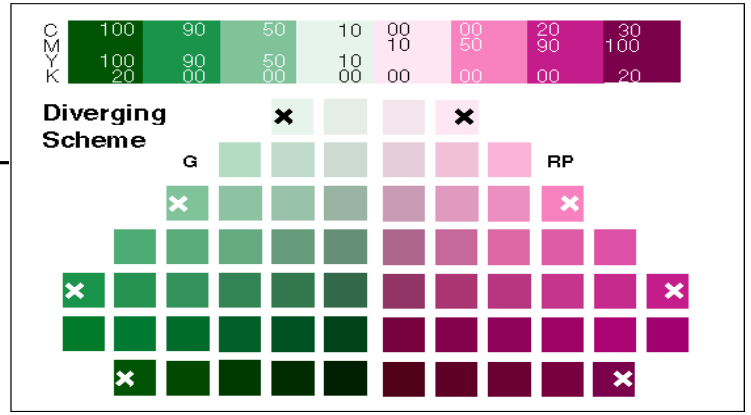
As a result, the use of "grayscale" or similar sequential color maps can be misleading on reading individual values. In the discontinuous map below, pay also attention to the "Mach band" effect:

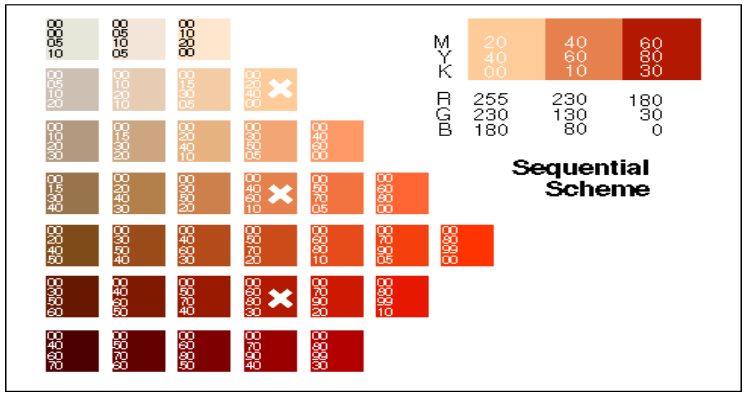


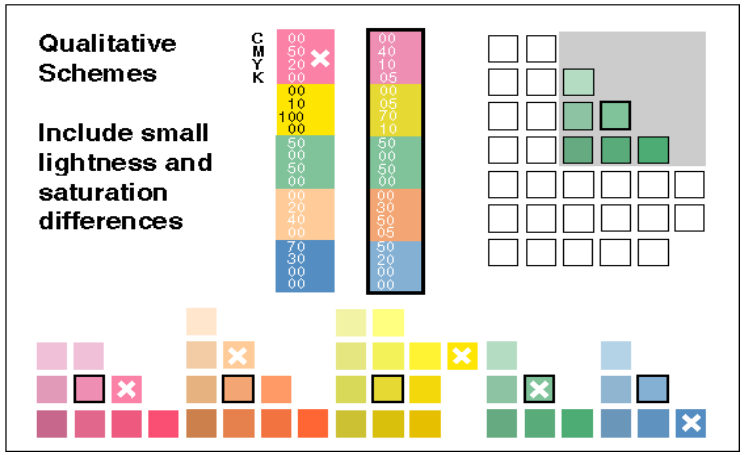
Color map design/use

Back to Cintia Brewer, now some principles can be better understood:

- In diverging schemes, use opposite hues and a neutral color for the neutral value.
- Use sequential schemes for ordinal values (f.e., education level).
- For nominal values, use hues well apart in chrominance (and no more than 6~8 thereof).

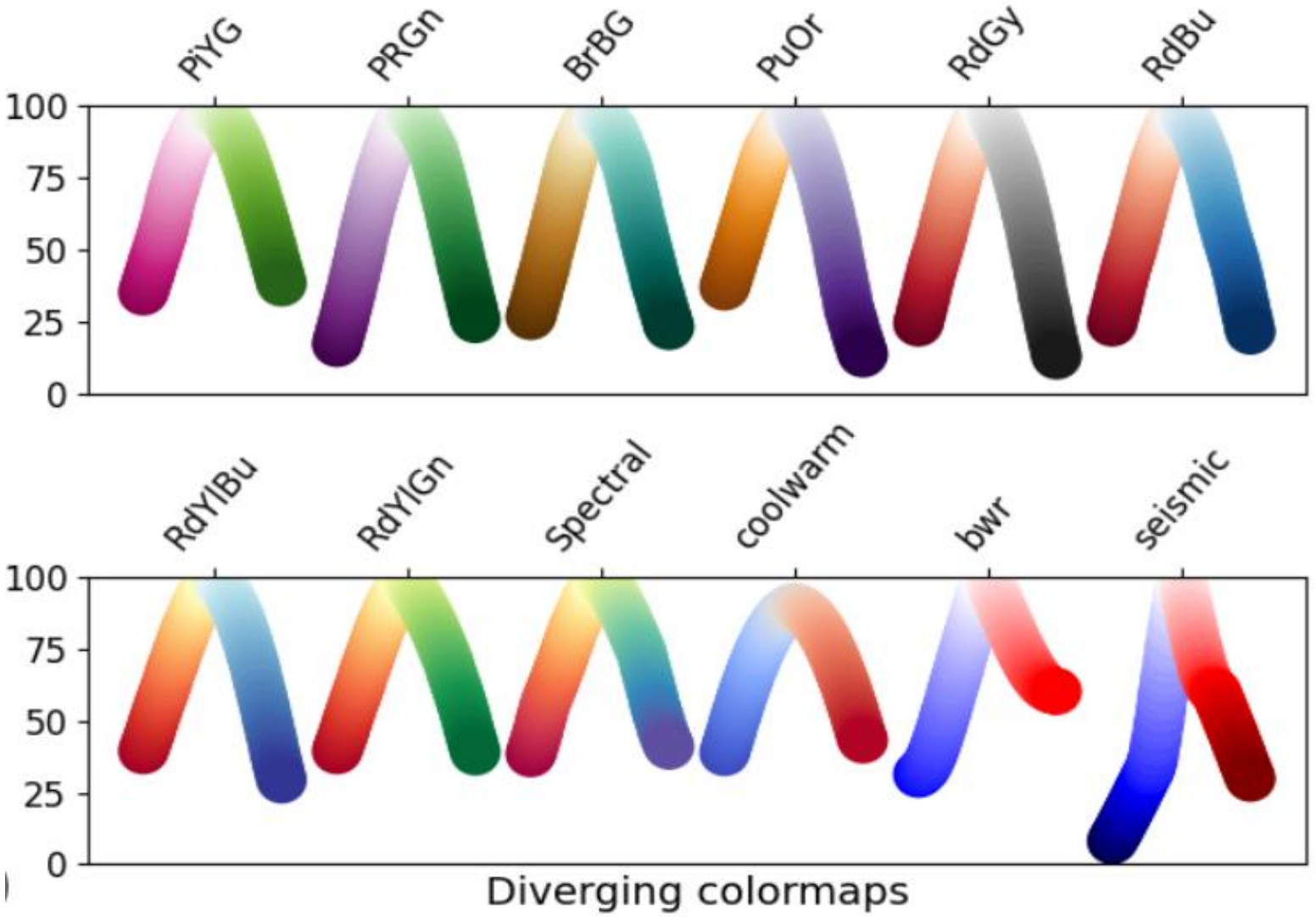






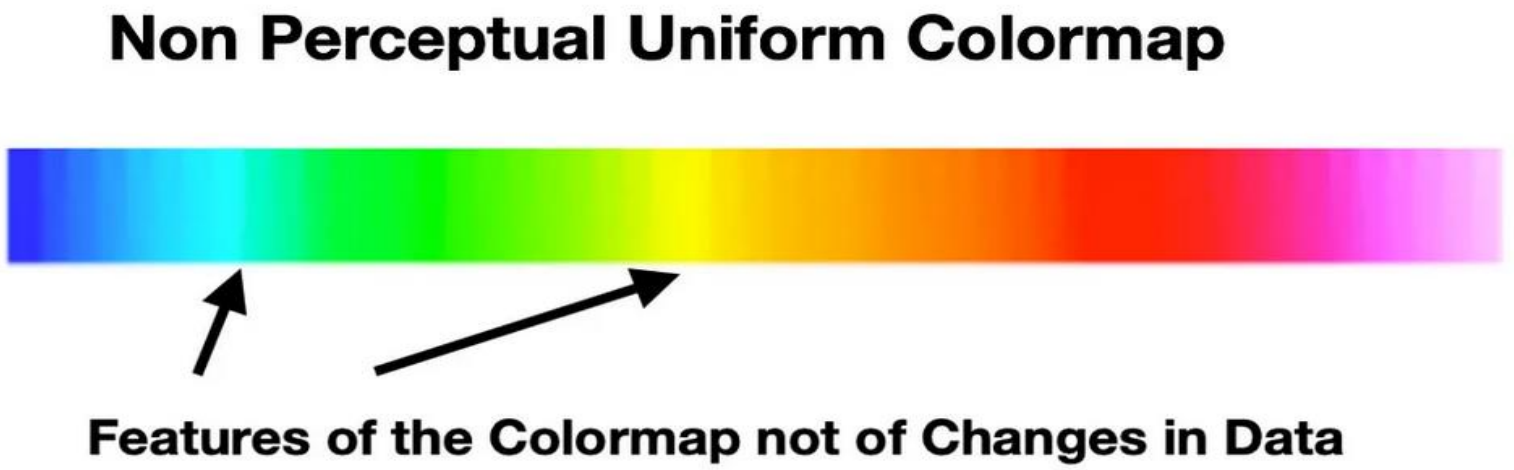
Color map design/use

These are the relative luminances of the most commonly MatPlotLib color maps.



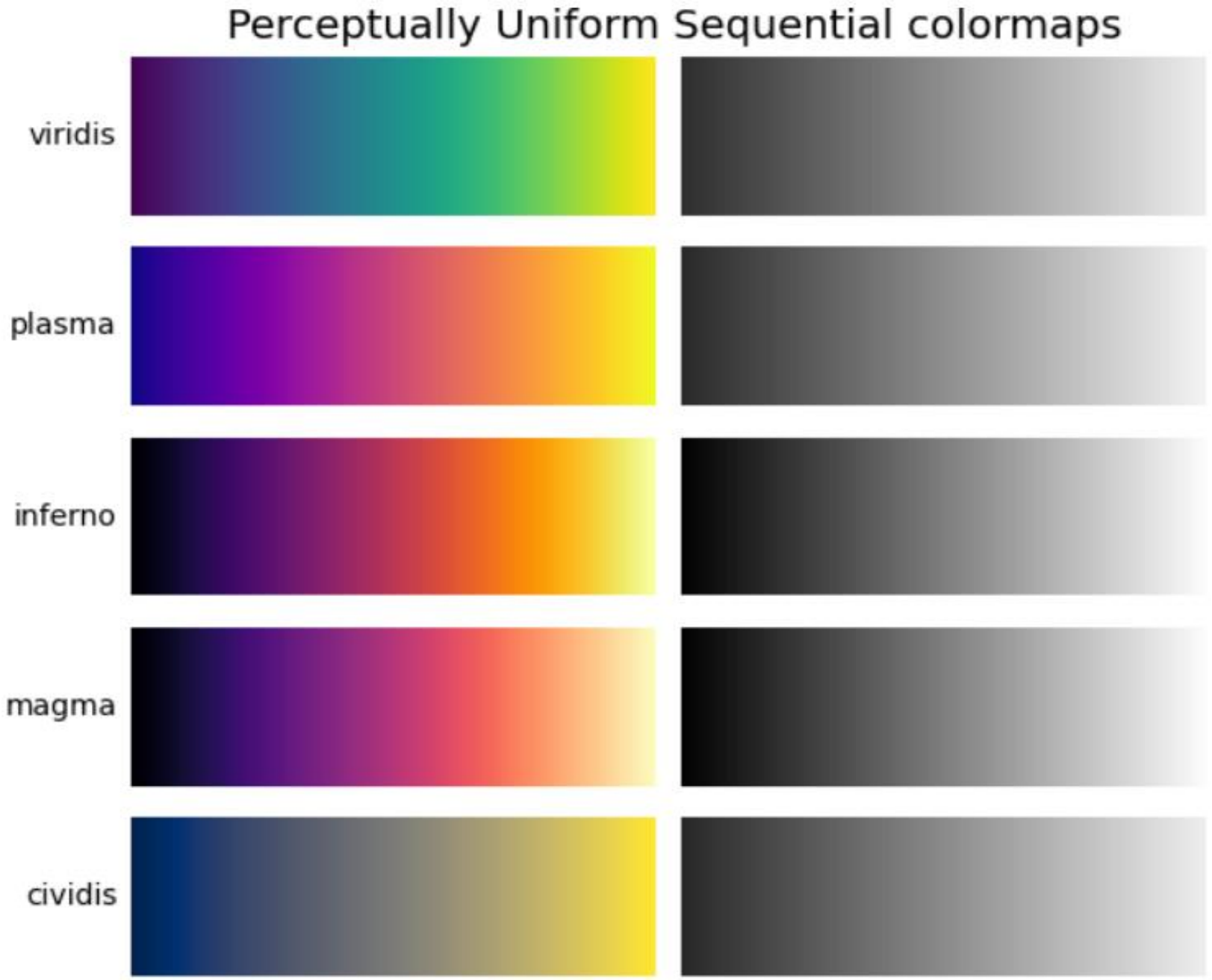
Color map design/use

What about continuous (quantitative) data?



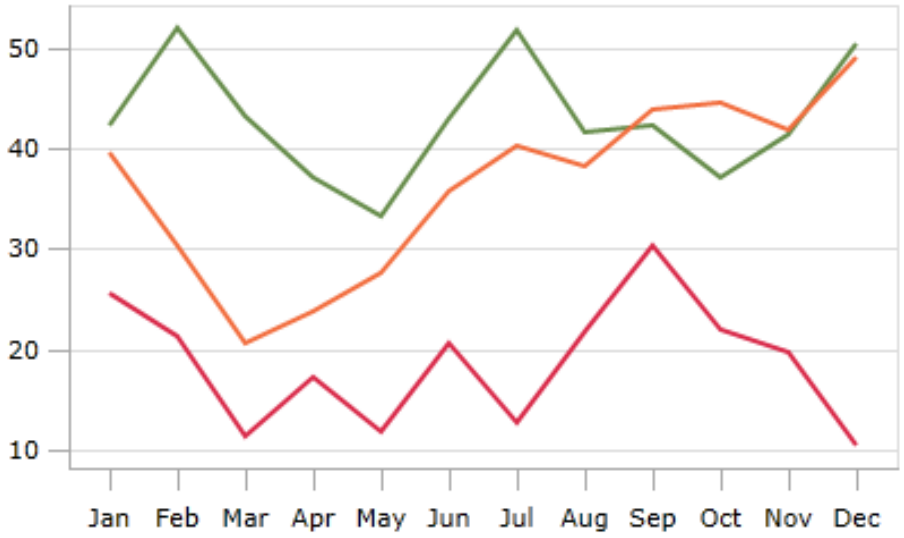
Color maps

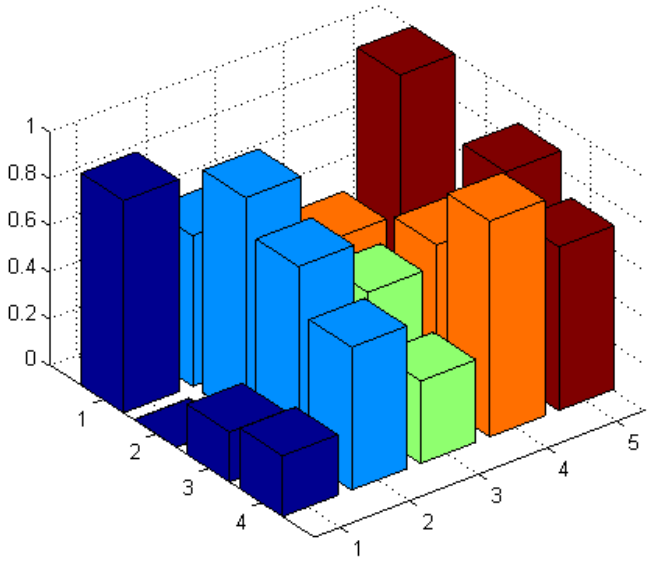
There are perceptually uniform sequential colormaps (they combine hue together with increasing brightness).

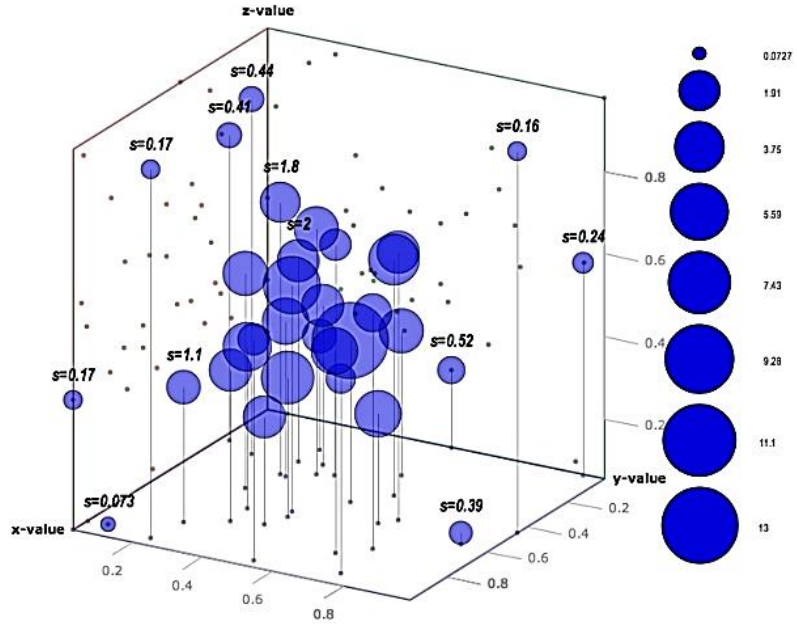


How many views are there?

We are mostly acquainted with several view for “rectangular” dataframes with one, two or more independent variables.

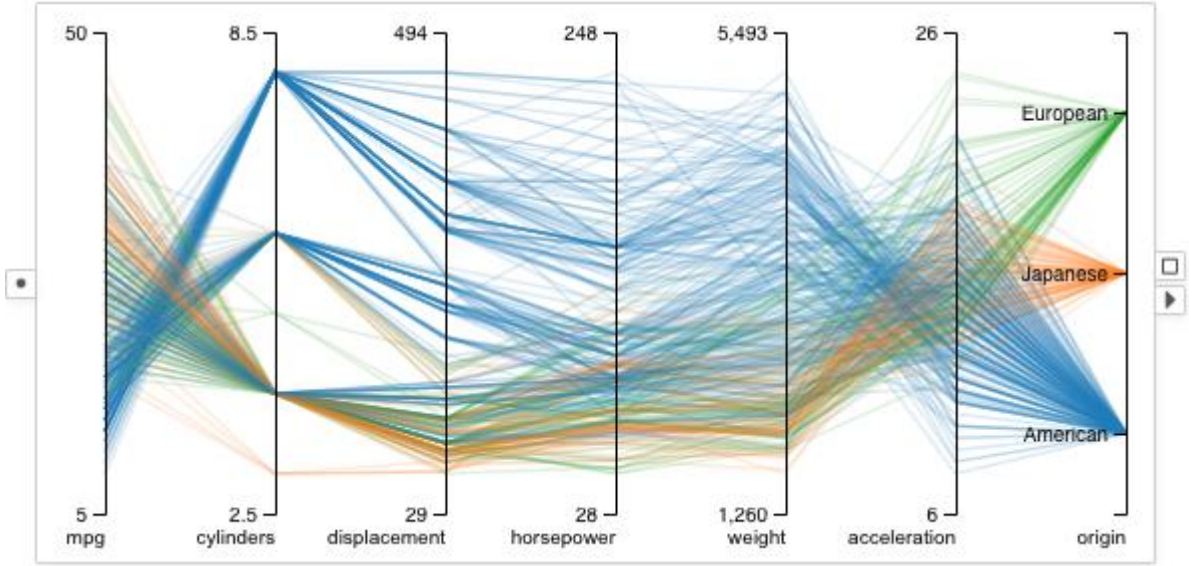


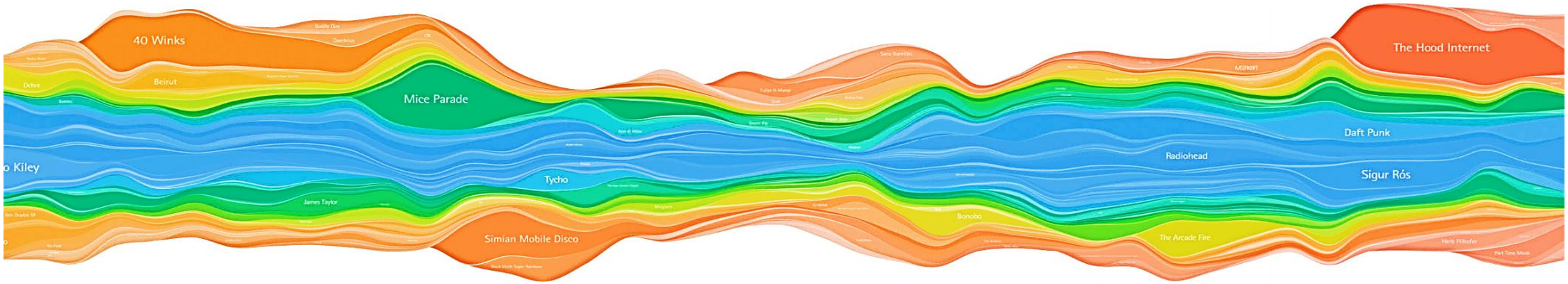




How many views are there?

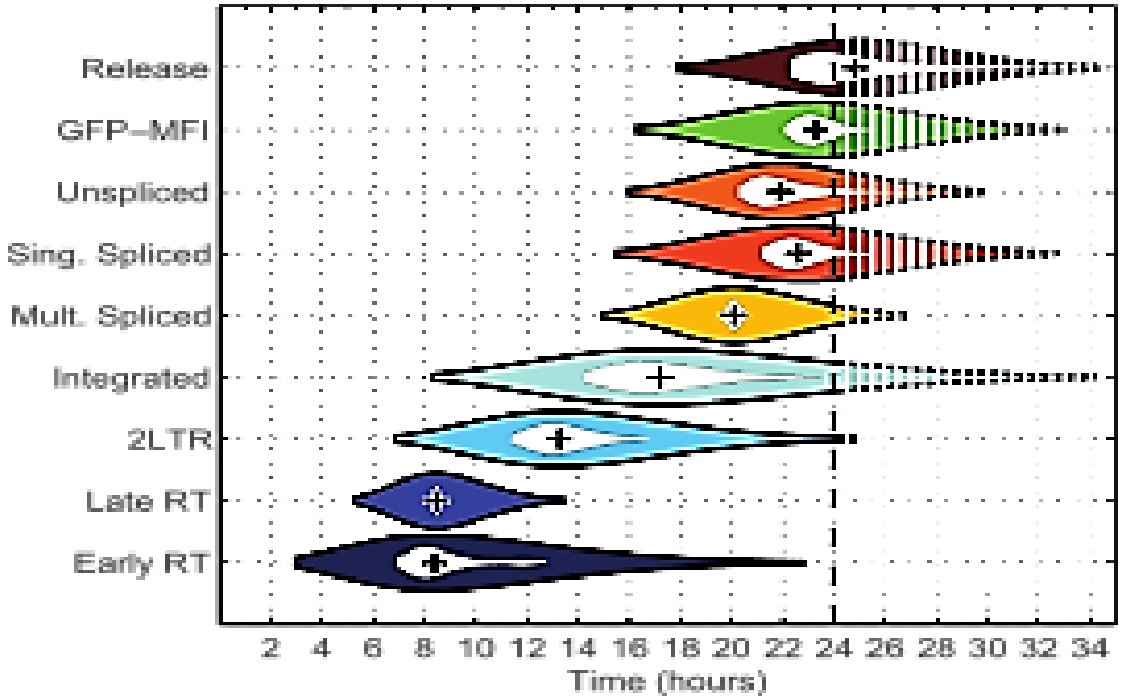
When the amount of variables is large, other kind of views are required.

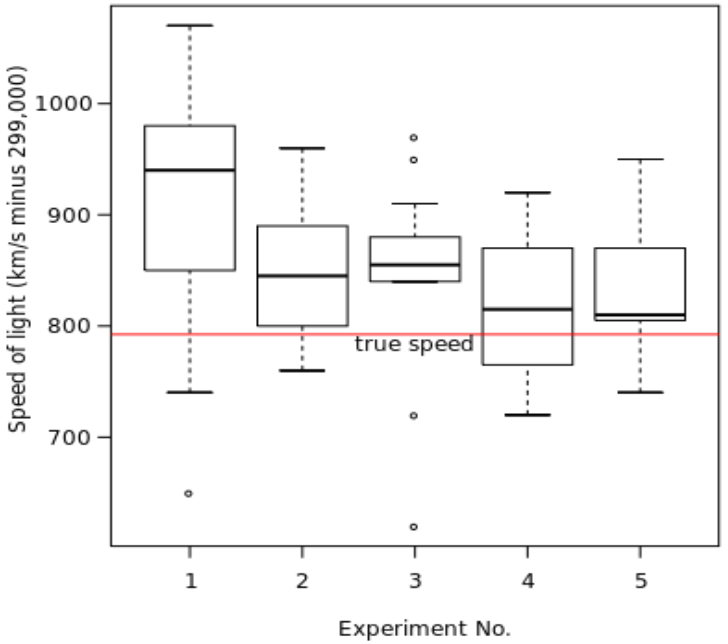




How many views are there?

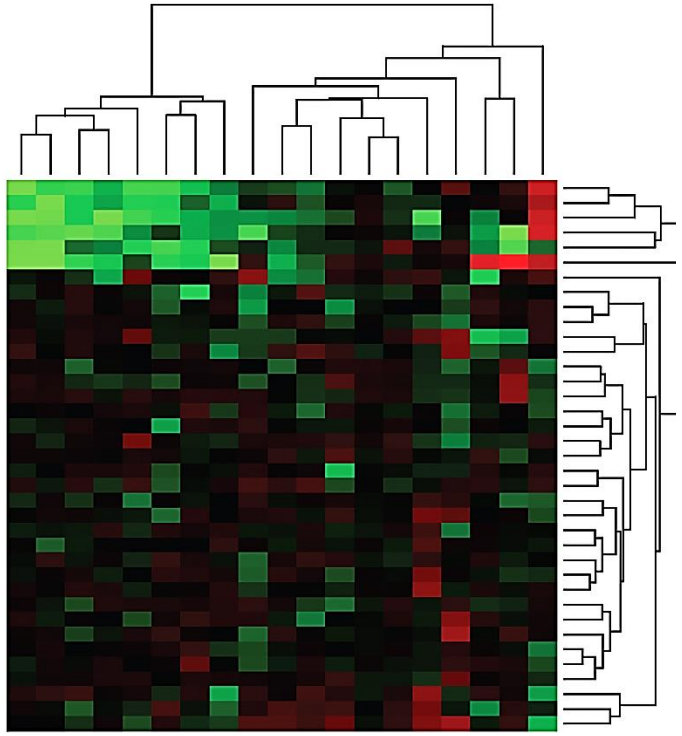
Population-wise (parametric or non-parametric) have specific views:

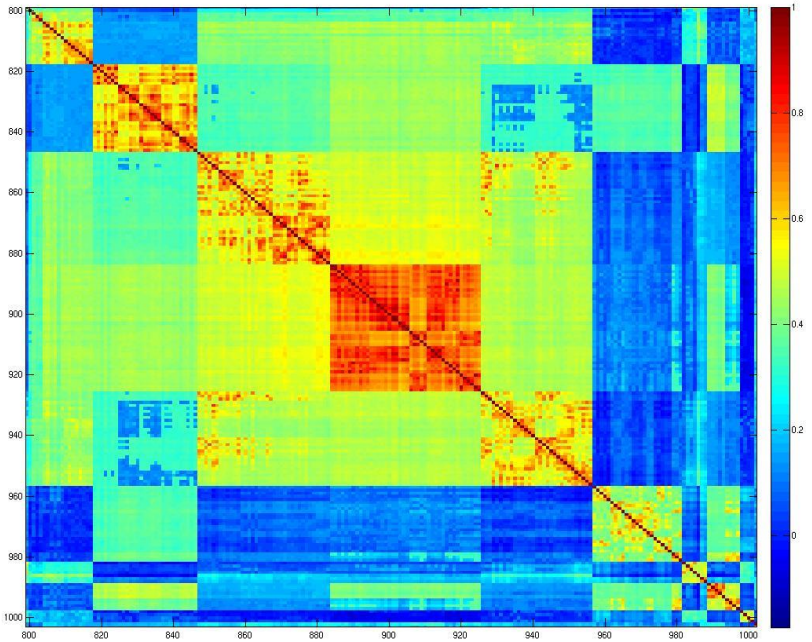


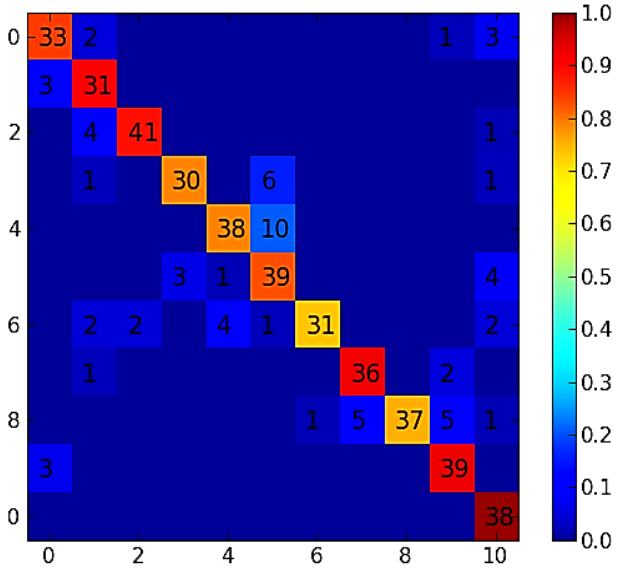


How many views are there?

Matrices (covariance, confusion, etc) typicaly require to associate position with color map.

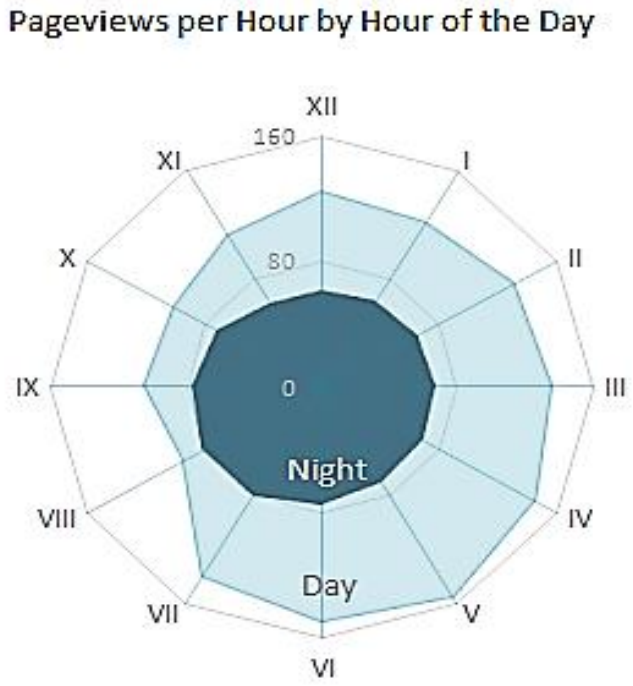


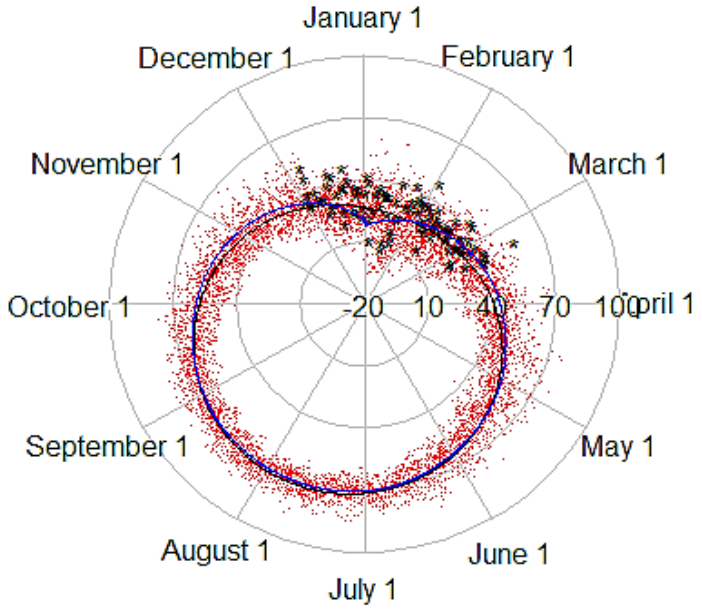


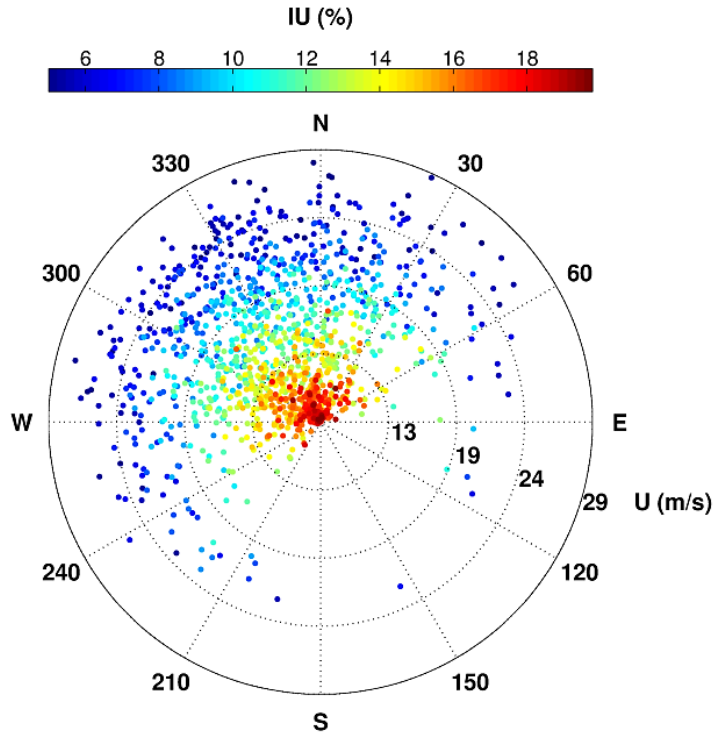


How many views are there?

An independent periodic variable (time, direction) is best visualized with radial plots.

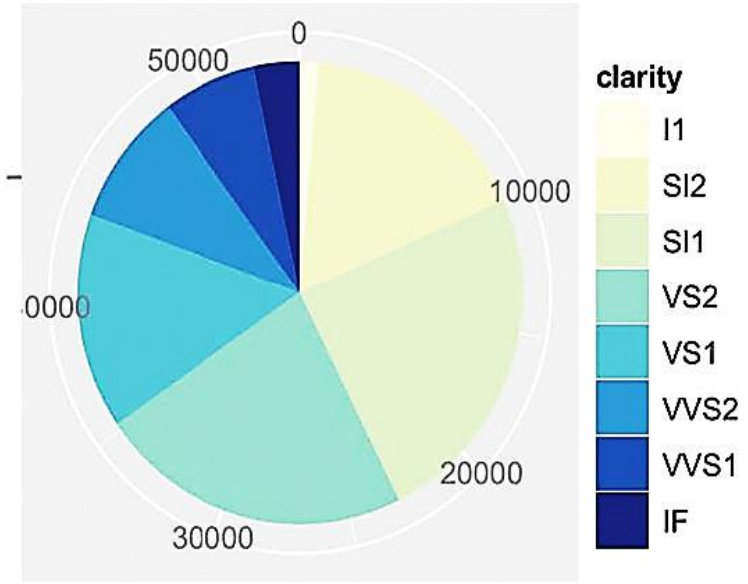


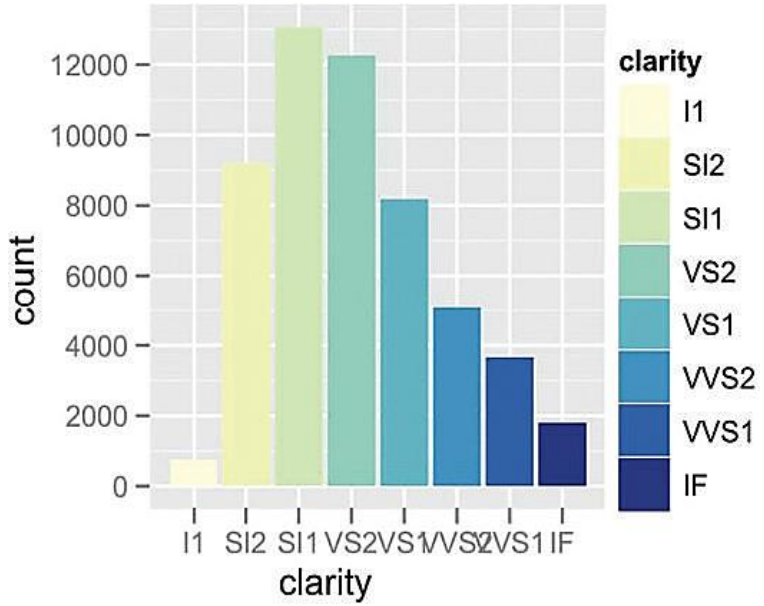


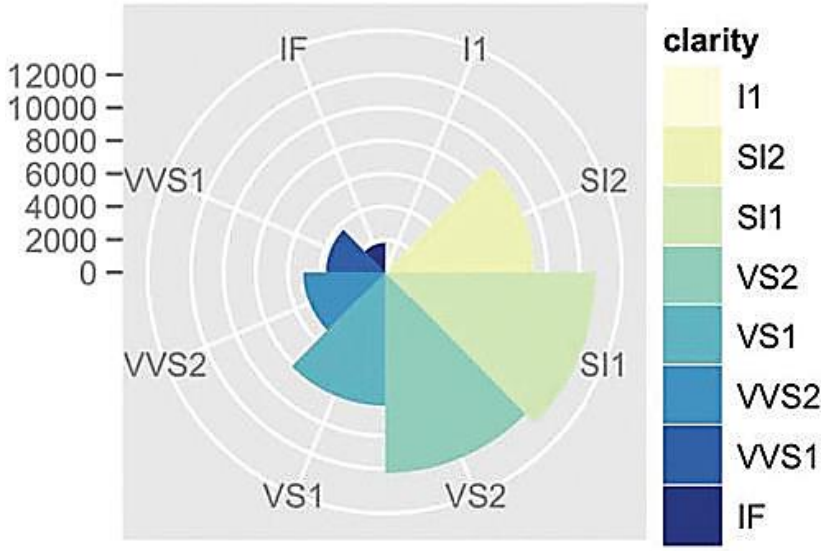


How many views are there?

Pie charts are a sort of radial plot. Is their use justified?

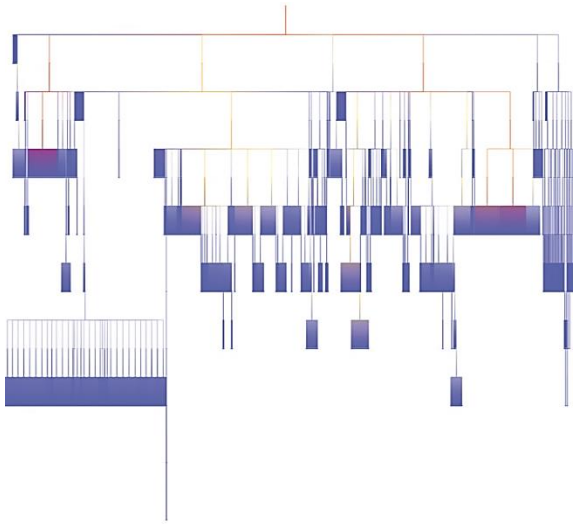


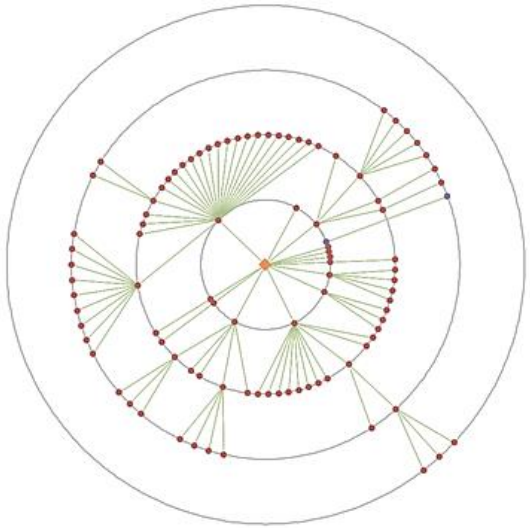


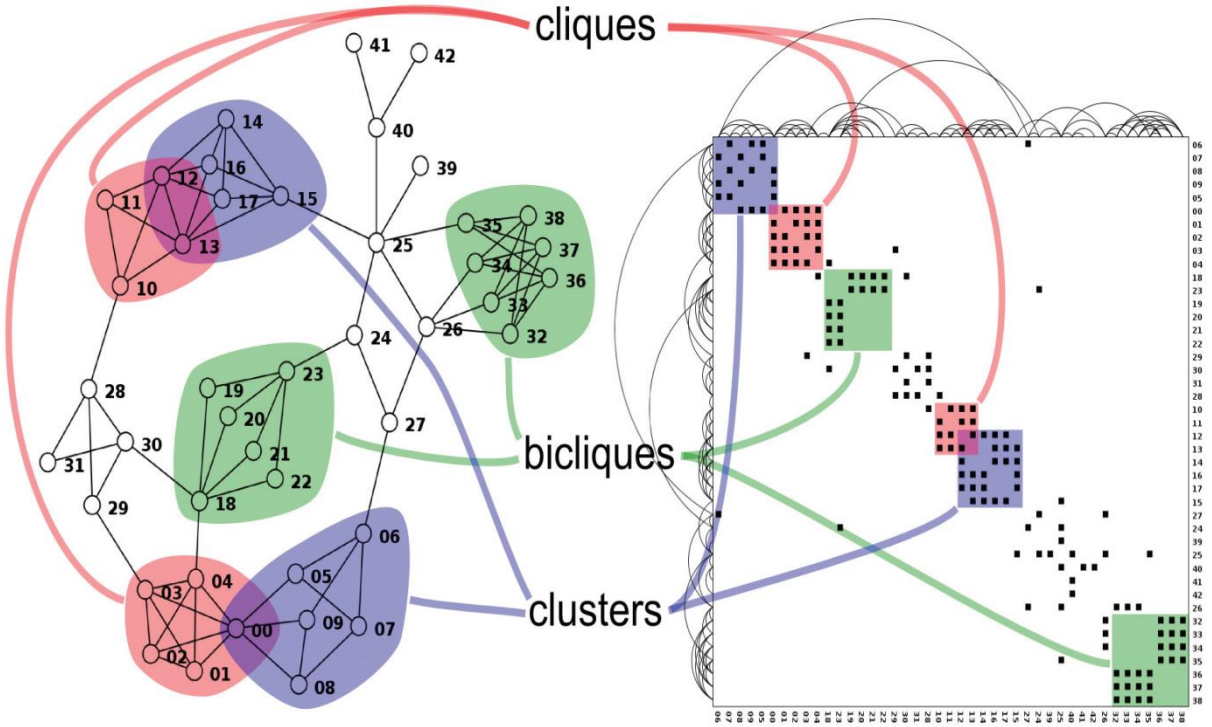


How many views are there?

Trees, graphs, etc. require special techniques.



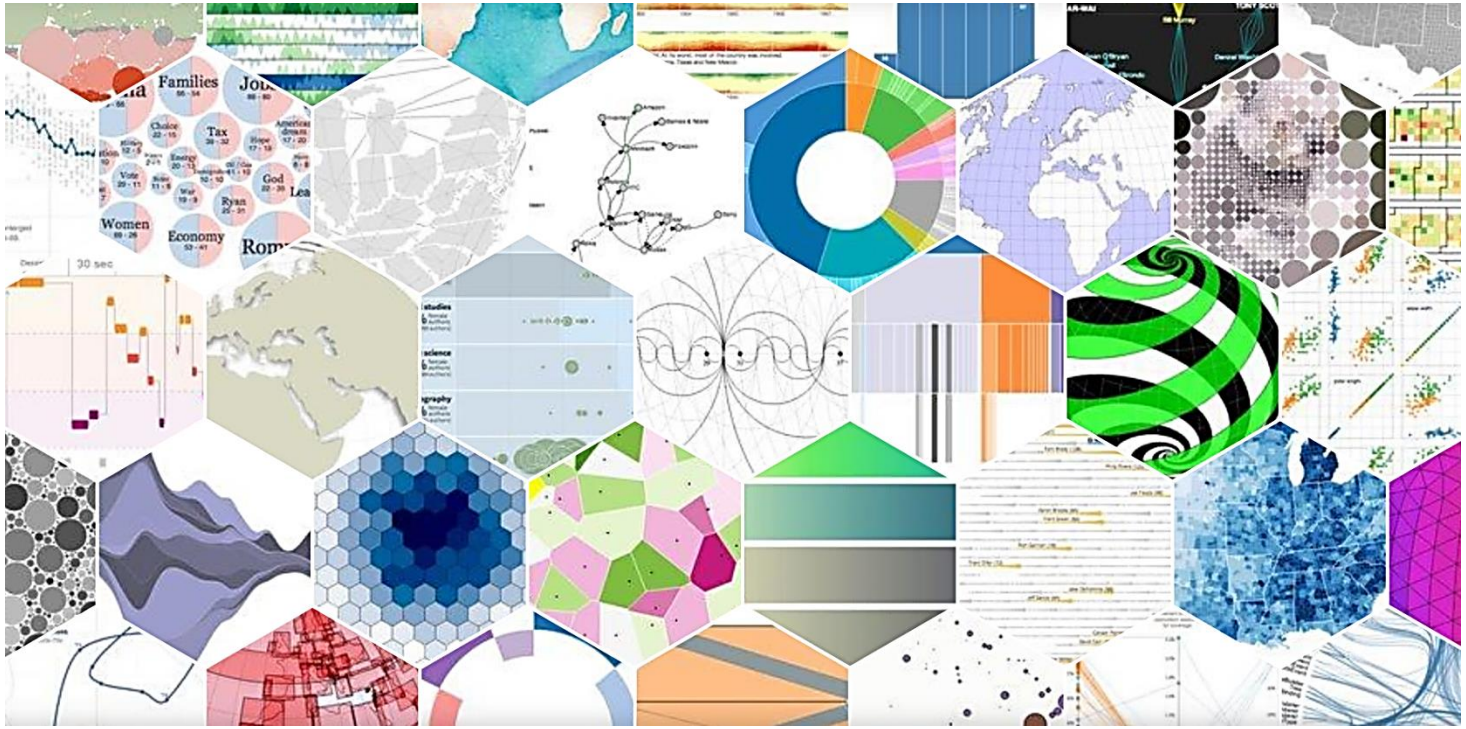




How many views are there?

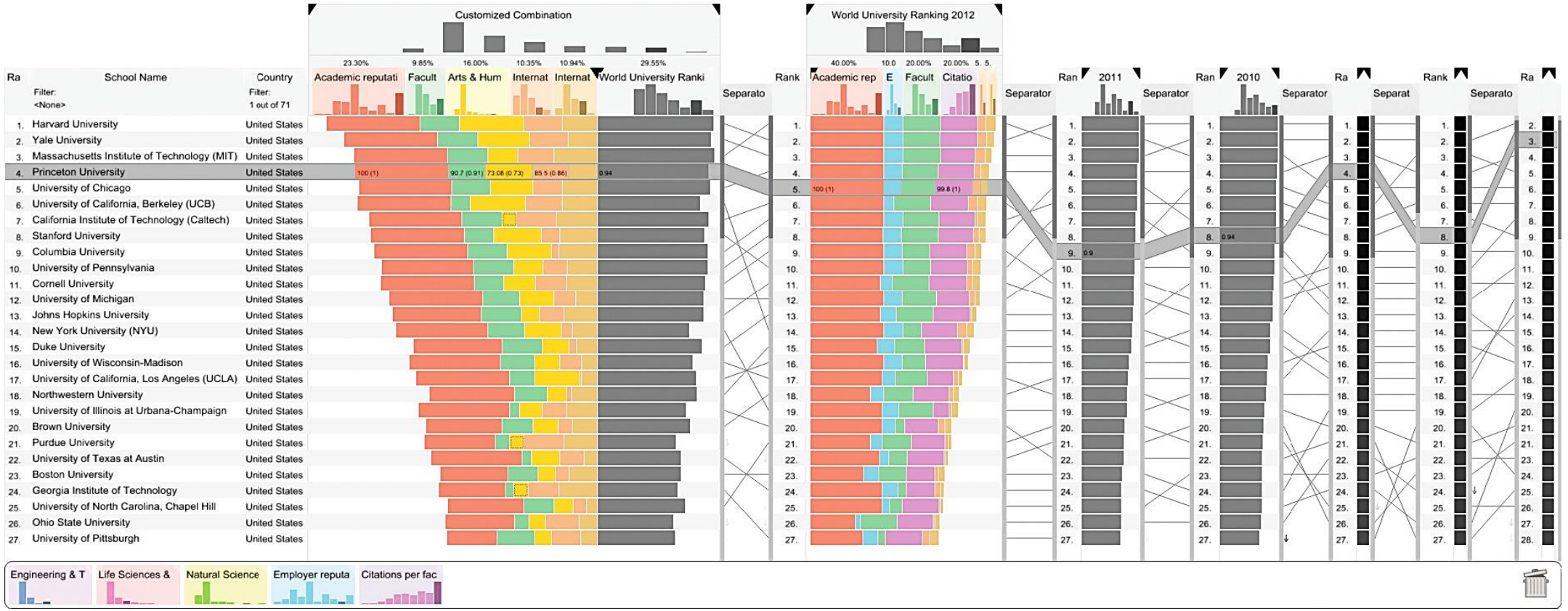
https://datavizcatalogue.com

https://d3js.org/



Acctionable views

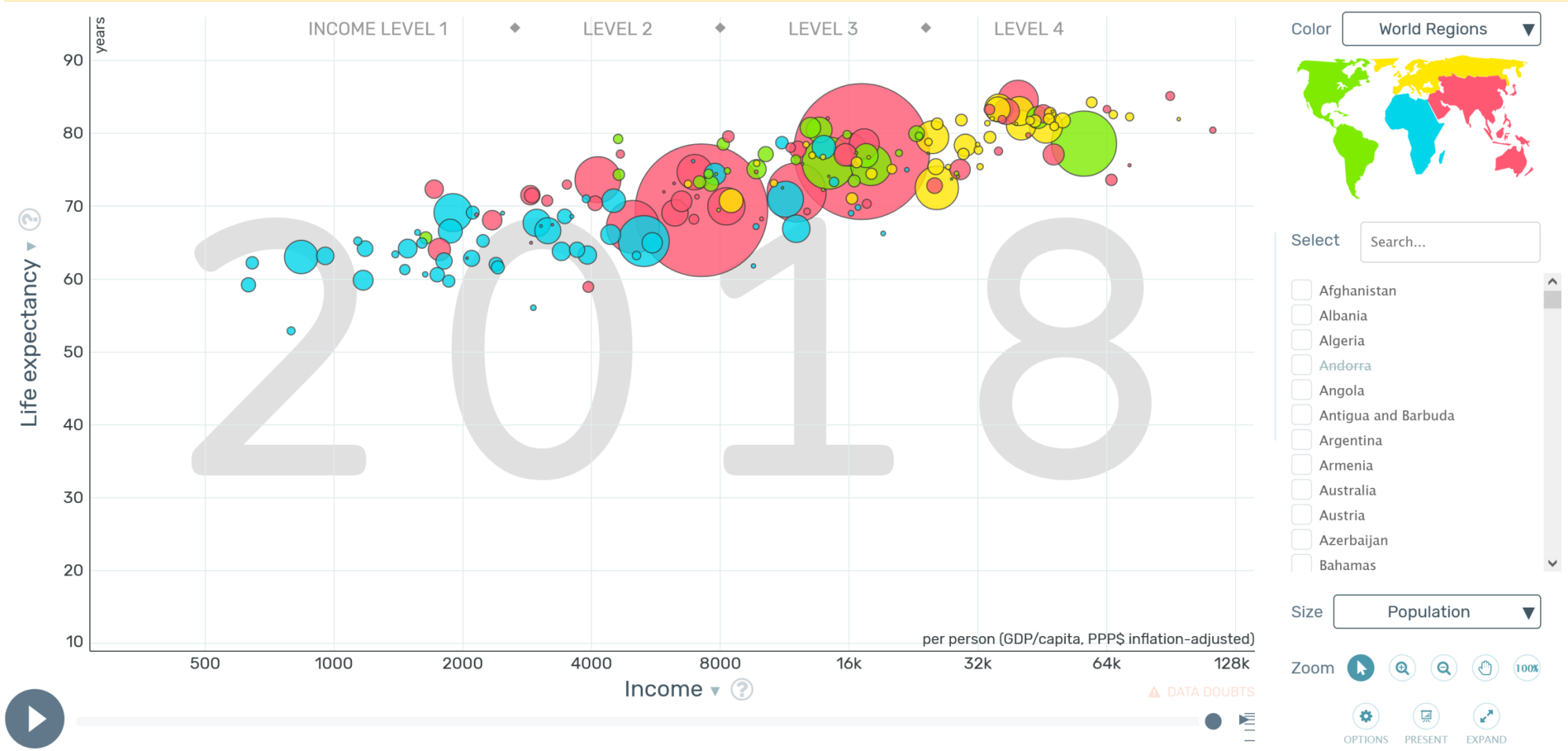
Actionable views allow users to interact and drill-down data.



Juxtaposed coordinated views

These views allow richer exploration and hypothesis formation.

Animated Views



What else?

Narrative: https://projects.fivethirtyeight.com/2016-election-forecast/

Visualization playground: https://playground.tensorflow.org

Tools: https://altair-viz.github.io/
https://holoviews.org/