

#DATAVIZ

A 2h00 crash-course on scientific visualization

Nicolas P. Rougier

Advanced Scientific Programming in Python, Camerino, Italy, 2018

What is scientific visualisation ?

“Visualisation is a method of computing. It transforms the symbolic into the geometric, enabling researchers to observe their simulations and computations. Visualisation offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights.”

Visualisation in Scientific Computing, NSF report, 1987

“For example, about 50 percent of the cerebral cortex of primates is devoted exclusively to visual processing, and the estimated territory for humans is nearly comparable.”

The MIT Encyclopedia of the Cognitive Sciences

Anscombe's quartet, 1973

The purpose of computing is insight, not numbers
Richard Hamming, 1962

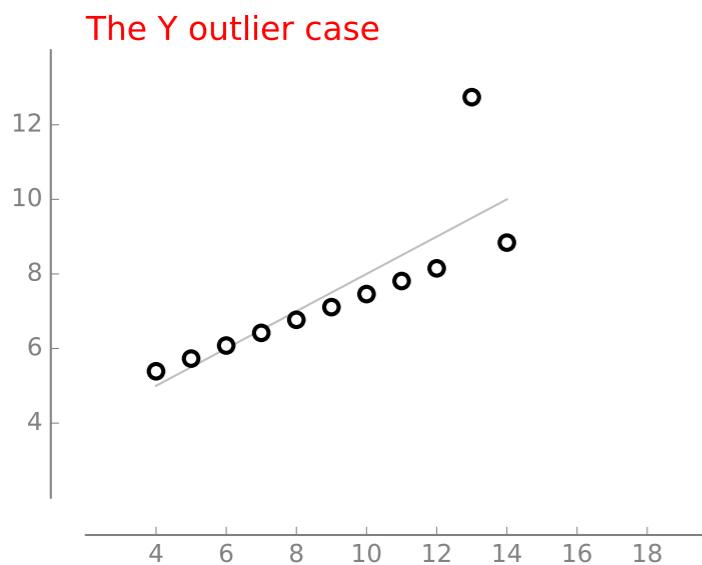
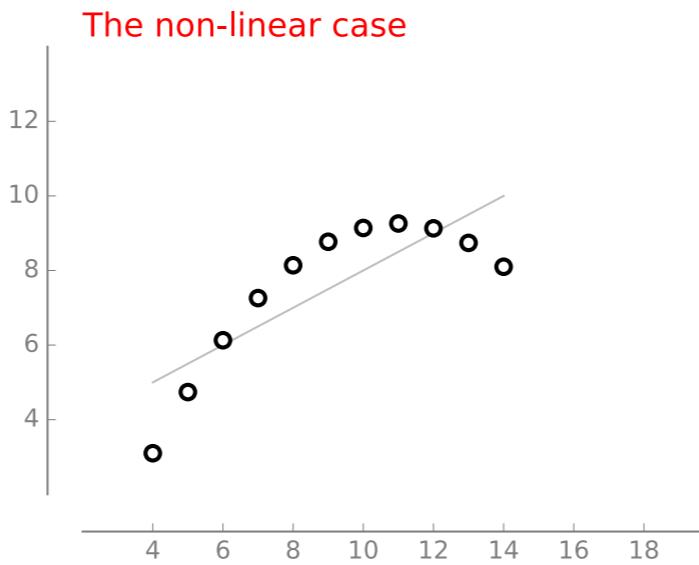
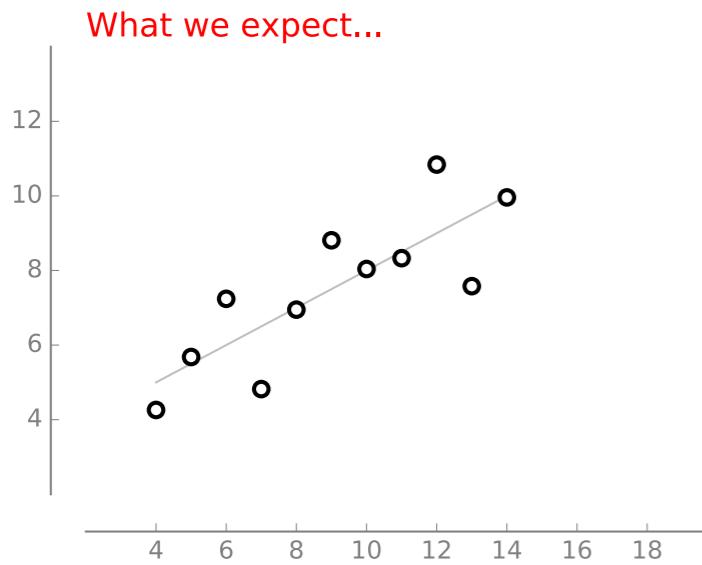
| I | | II | | III | | IV | |
|------|-------|------|------|------|-------|------|-------|
| x | y | x | y | x | y | x | y |
| 10.0 | 8.04 | 10.0 | 9.14 | 10.0 | 7.46 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.0 | 8.14 | 8.0 | 6.77 | 8.0 | 5.76 |
| 13.0 | 7.58 | 13.0 | 8.74 | 13.0 | 12.74 | 8.0 | 7.71 |
| 9.0 | 8.81 | 9.0 | 8.77 | 9.0 | 7.11 | 8.0 | 8.84 |
| 11.0 | 8.33 | 11.0 | 9.26 | 11.0 | 7.81 | 8.0 | 8.47 |
| 14.0 | 9.96 | 14.0 | 8.10 | 14.0 | 8.84 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.0 | 6.13 | 6.0 | 6.08 | 8.0 | 5.25 |
| 4.0 | 4.26 | 4.0 | 3.10 | 4.0 | 5.39 | 19.0 | 12.50 |
| 12.0 | 10.84 | 12.0 | 9.13 | 12.0 | 8.15 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.0 | 7.26 | 7.0 | 6.42 | 8.0 | 7.91 |
| 5.0 | 5.68 | 5.0 | 4.74 | 5.0 | 5.73 | 8.0 | 6.89 |

What is common to these data sets?

Mean of x: 9
Sample variance of x: 11
Mean of y: 7.50
Sample variance of y: 4.12
Linear regression: $y=3.00+0.500*x$
R squared: 0.666
p value 0.0021

But having a closer look at the data...

Anscombe's quartet, 1973



*“A computer
should make
both calculations
and graphs”*

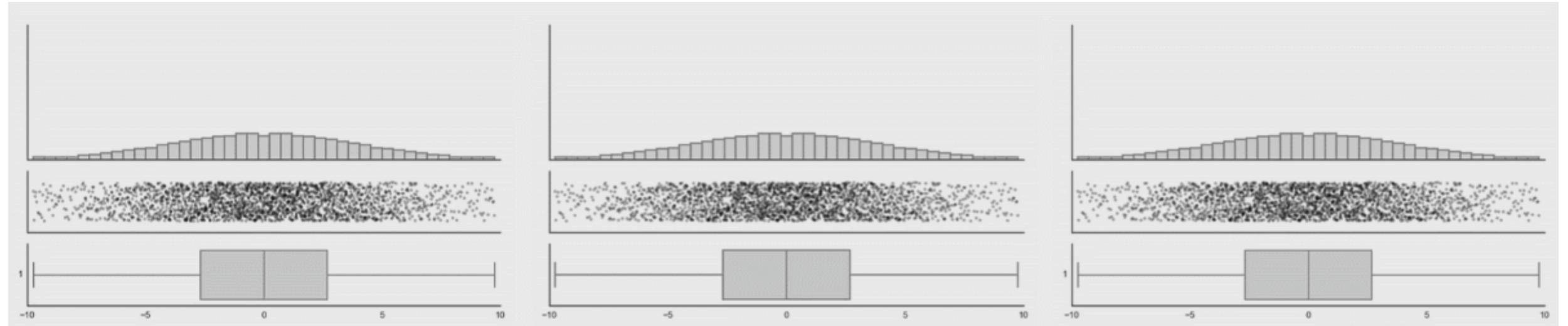
Francis Anscombe (1918-2001)

What is common to these data sets?

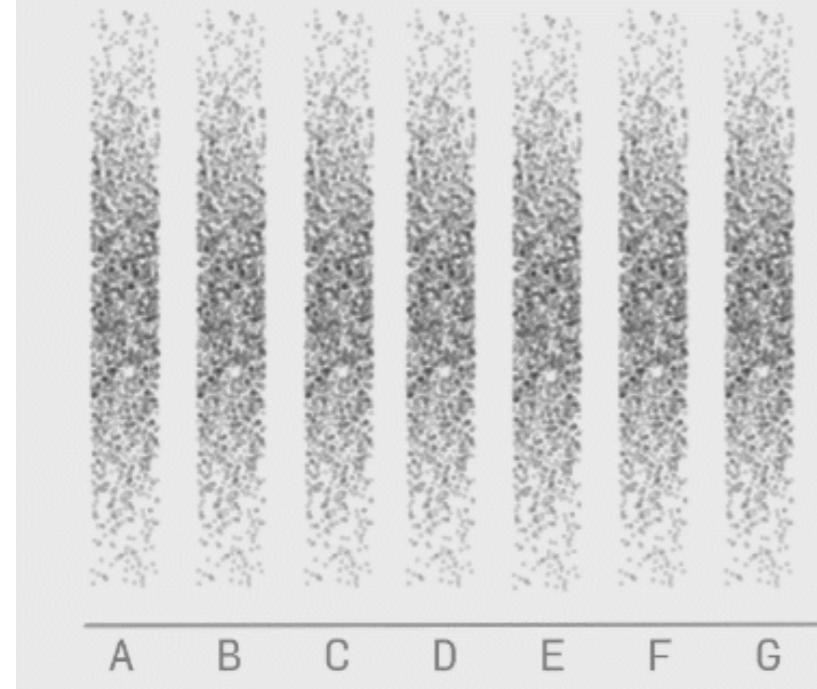
| | |
|-----------------------|------------------|
| Mean of x: | 9 |
| Sample variance of x: | 11 |
| Mean of y: | 7.50 |
| Sample variance of y: | 4.12 |
| Linear regression: | $y=3.00+0.500*x$ |
| R squared: | 0.666 |
| p value | 0.0021 |

But having a closer look at the data...

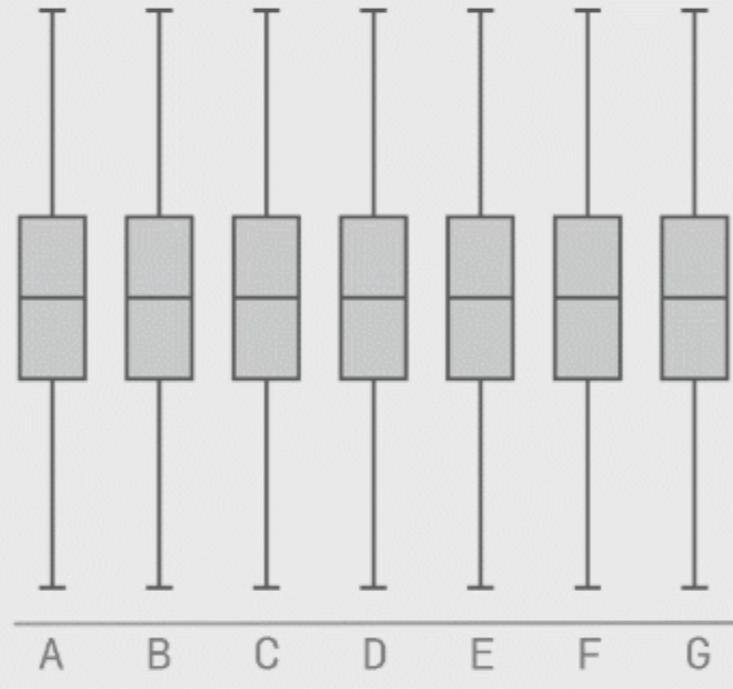
Autodesk research, 2017



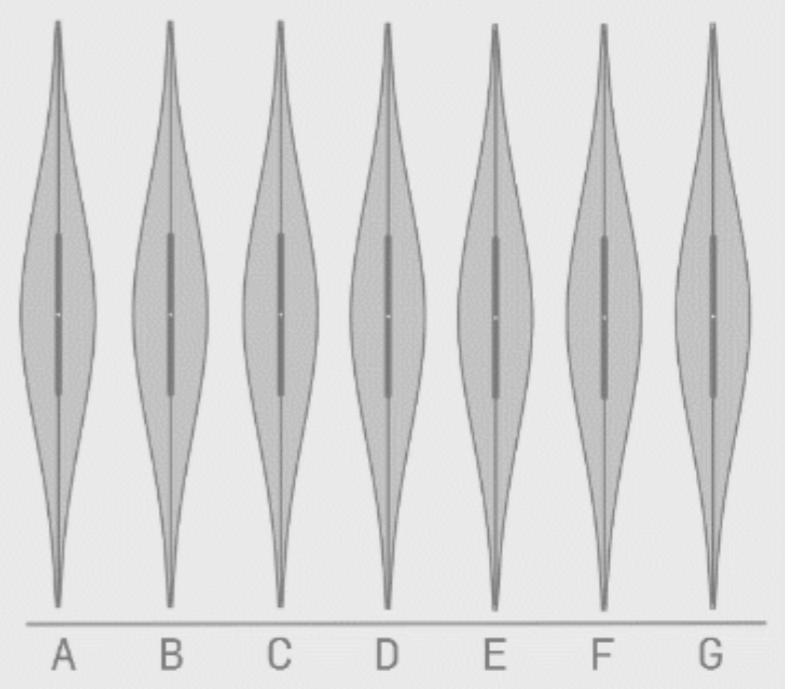
Raw Data



Box-plot of the Data



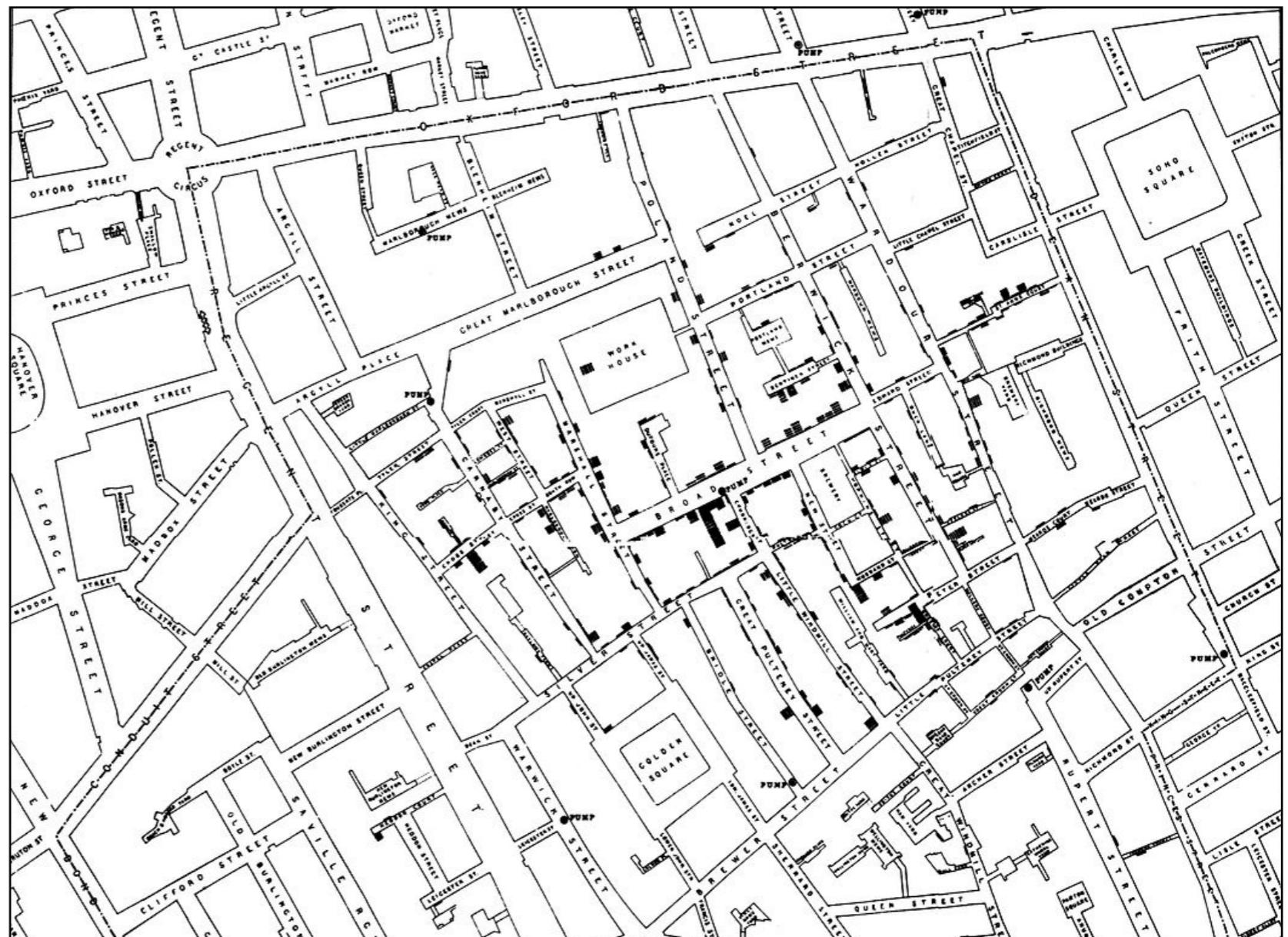
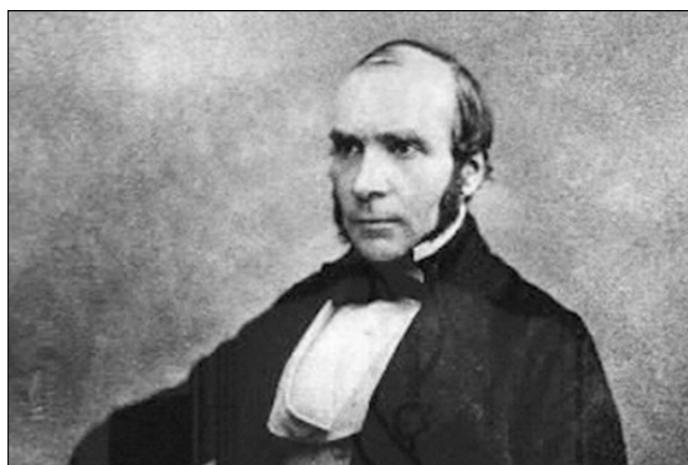
Violin-plot of the Data



Cholera epidemic, London, 1854

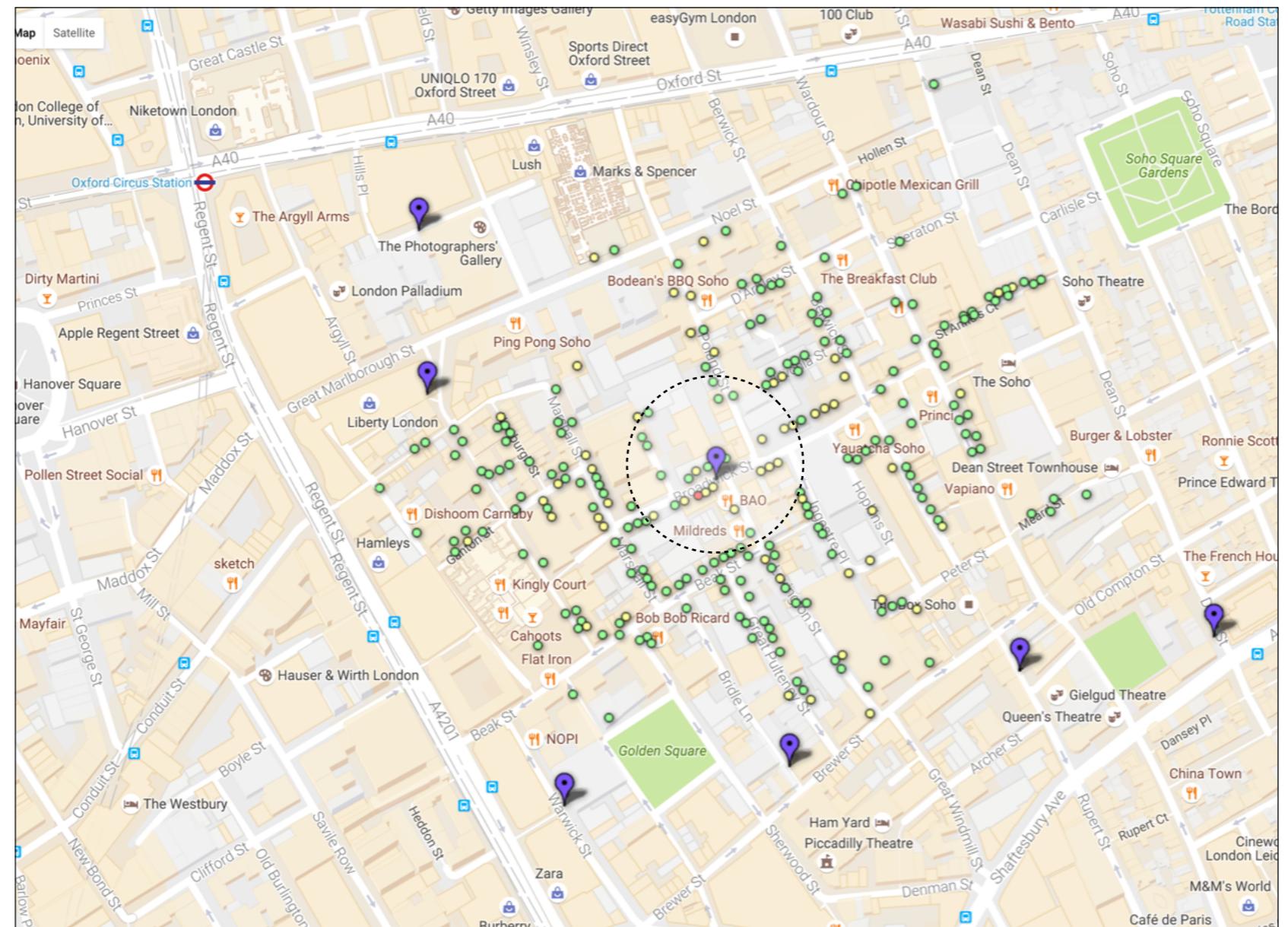
The most terrible outbreak of cholera which ever occurred in this kingdom (J. Snow)

John Snow (1813-1858)



Cholera epidemic, London, 1854

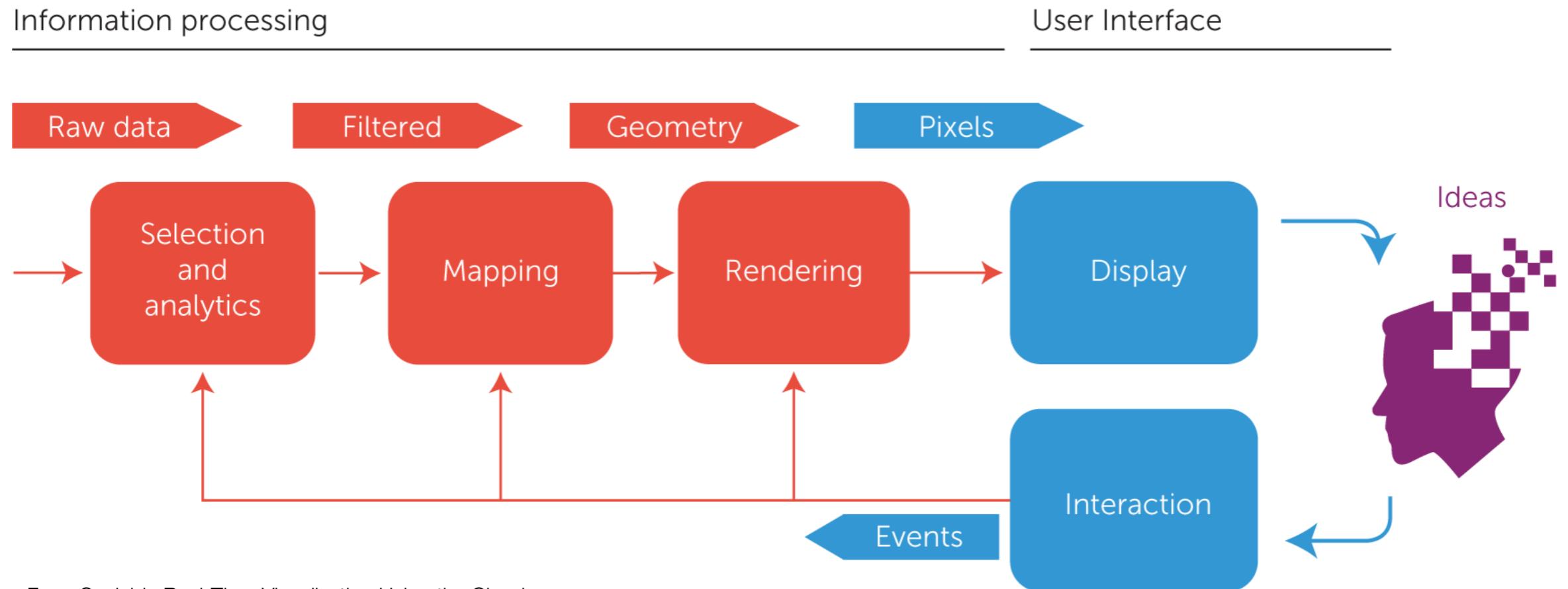
The most terrible outbreak of cholera which ever occurred in this kingdom (J. Snow)



Jon Snow (no connection at all)



The visualisation pipeline



From Scalable Real-Time Visualization Using the Cloud
Issue No.06 - Nov.-Dec. (2015 vol.2)
Nick Holliman , Newcastle University, UK
Paul Watson , Newcastle University, UK
DOI Bookmark: <http://doi.ieeecomputersociety.org/10.1109/MCC.2015.131>

Data type

Quantitative vs. Categorical Data: A Difference Worth Knowing (S.Few)

Quantitative → values or observations that can be measured

- Continuous (e.g. *temperature*)
- Discrete (e.g. *number of inhabitants*)

Categorical → values or observations that can be sorted into groups or categories

- Nominal (e.g. *nationality*)
- Ordinal (e.g. *months*)
- Interval (e.g. *age groups*)

Graphical elements

A scientific figure can be fully described by a set of graphic primitives with different attributes:

- Points, markers, lines, areas, ...
- Position, color, shape, size, orientation, curvature, ...
- Helpers, text, axis, ticks, ...
- Interaction, animation, ...

But who want to describe each individual elements? Describing a figure in terms of such graphic primitives would be a very tedious and complex task.

This is exactly where visualization libraries are useful because they will automatize most of the work (more or less depending on the library).

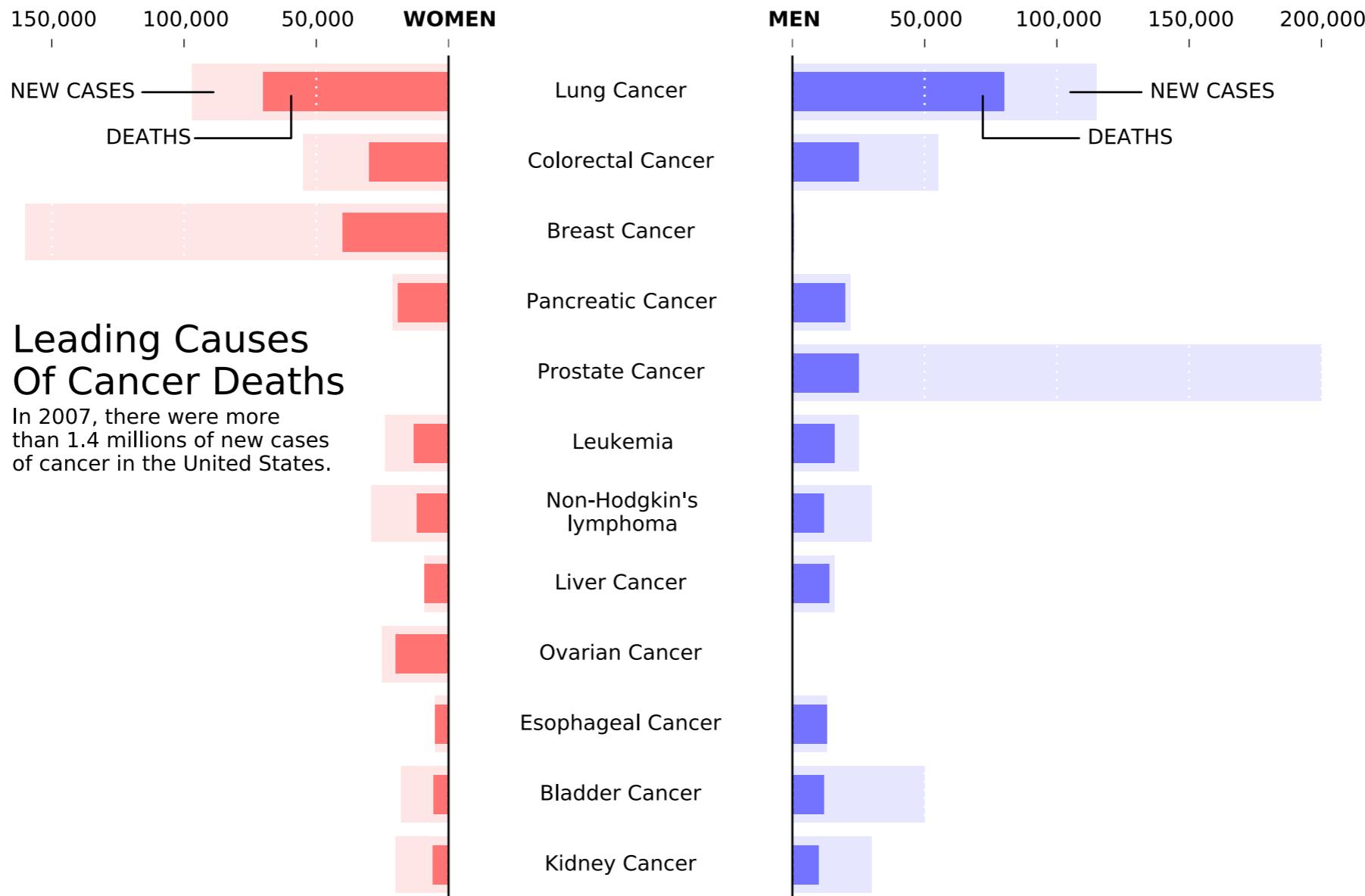
Visualization type

Data Visualisation catalogue by S. Rebecca

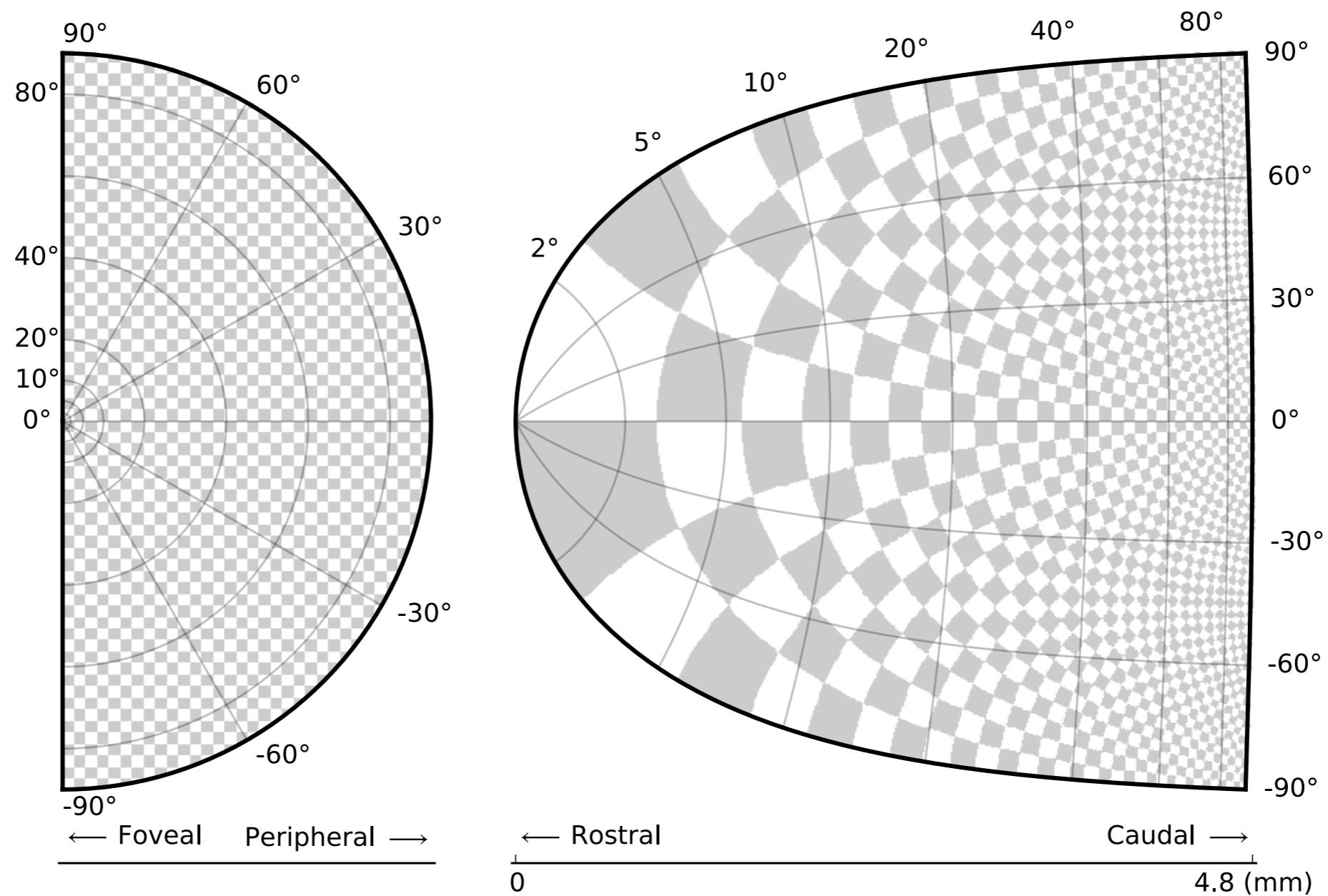


10 Simple Rules for Better Figures

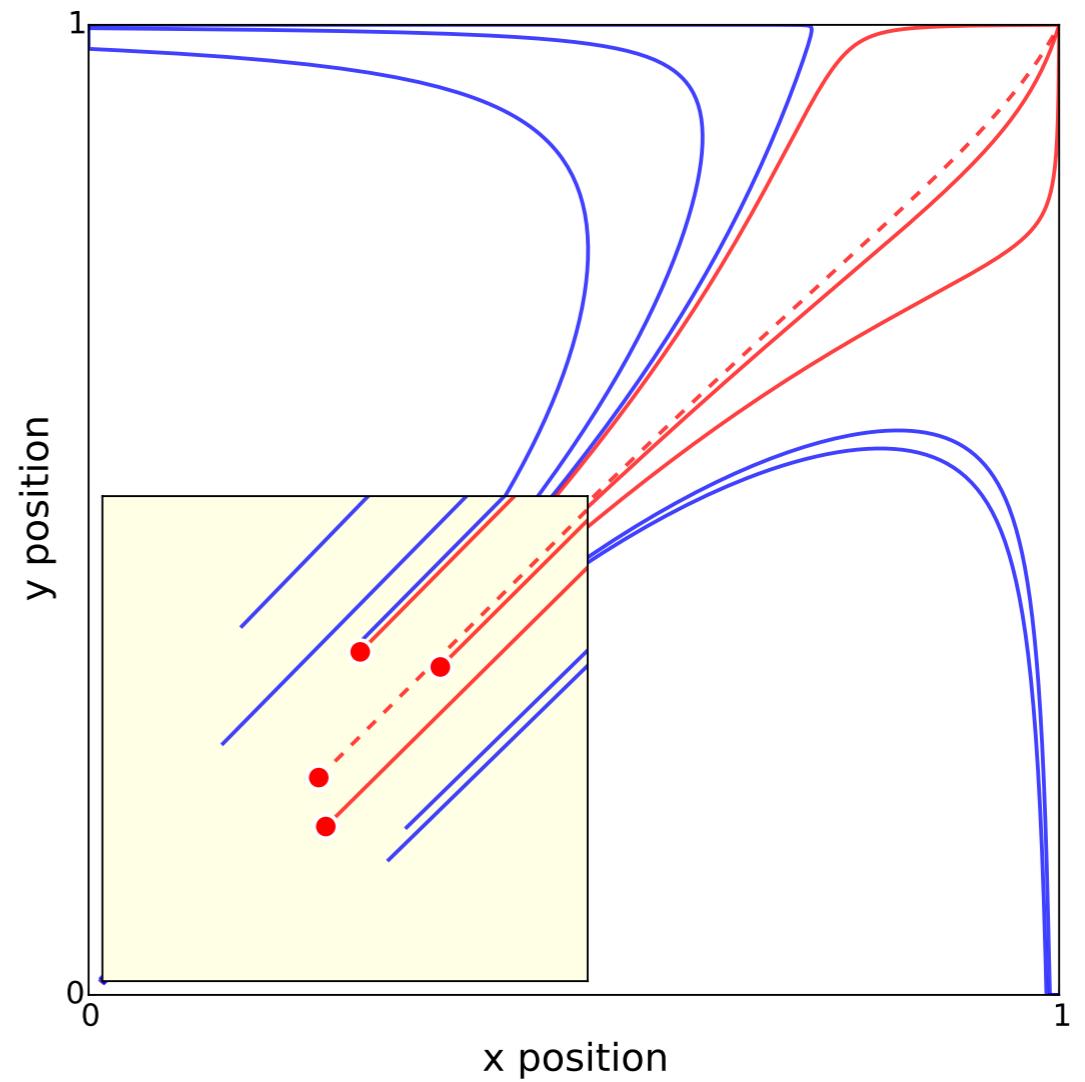
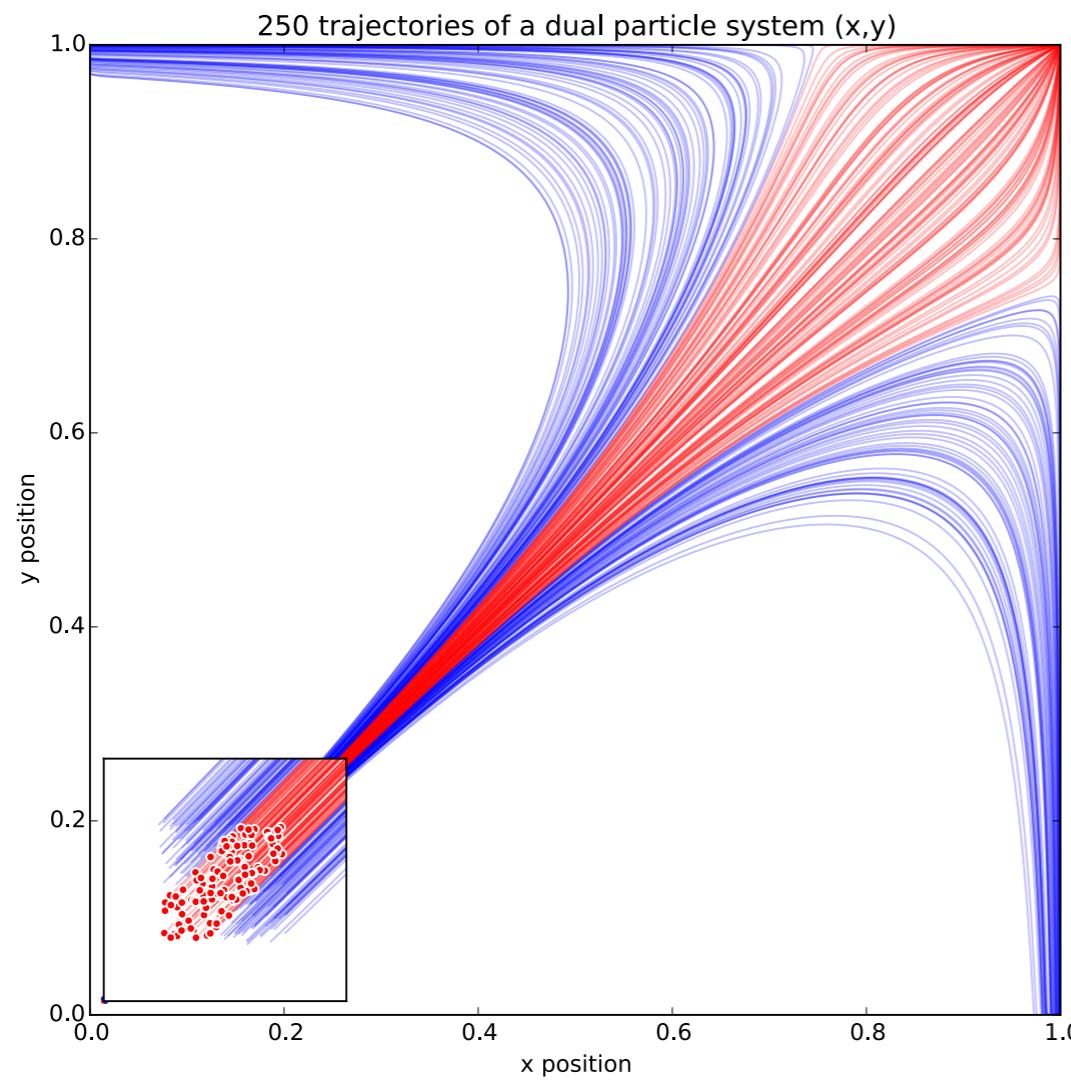
Rule 1: Know your audience



Rule 2: Identify your message



Rule 3: Adapt the figure



Rule 4: Captions are not optional

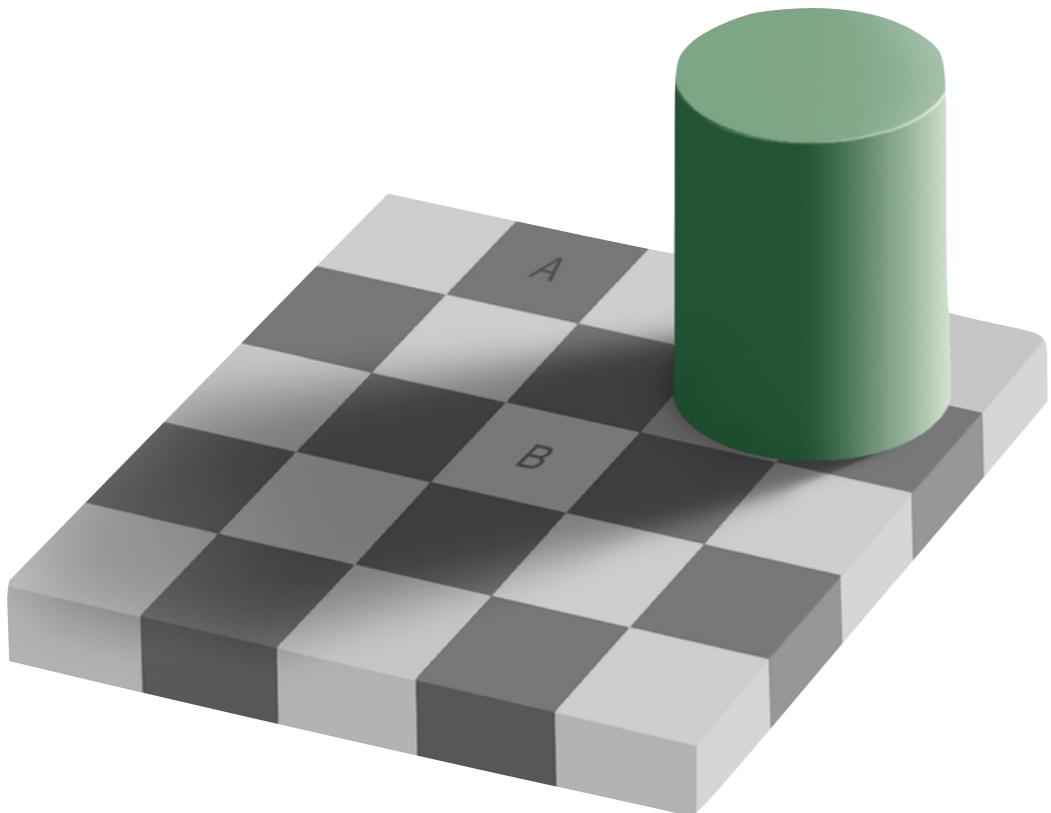


Figure 1. Optical illusion

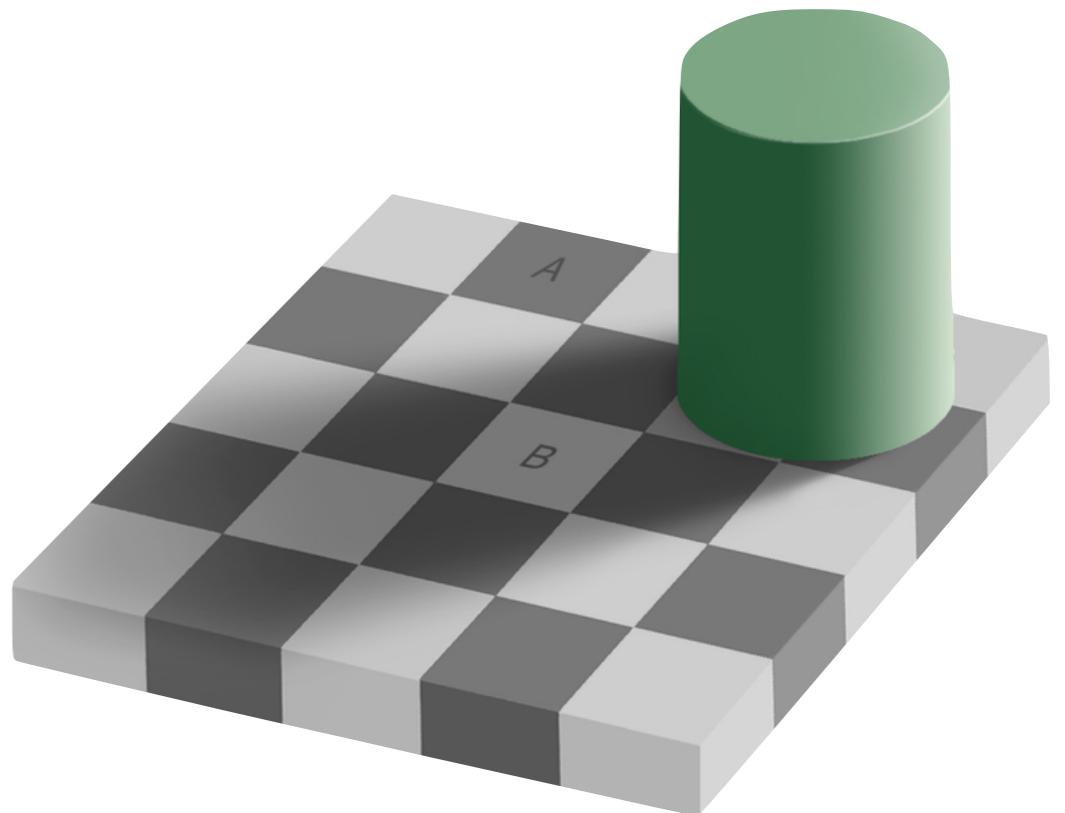
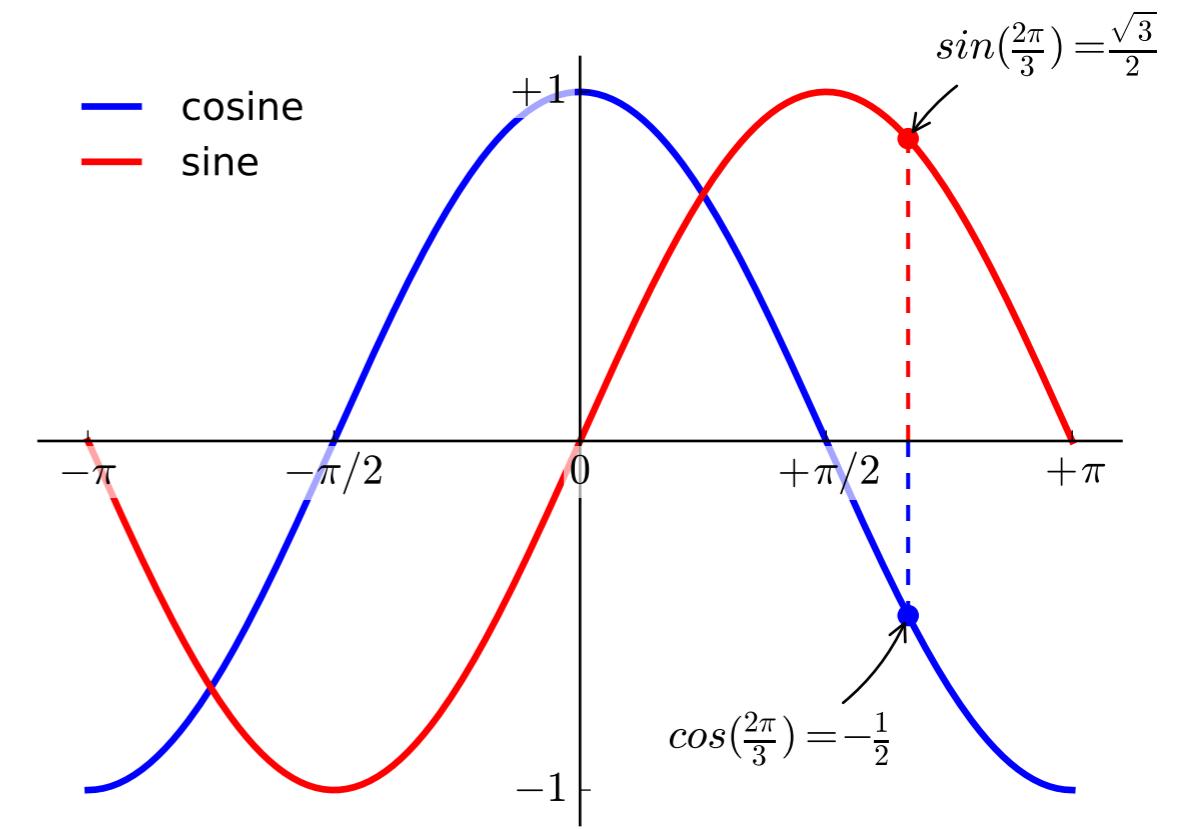
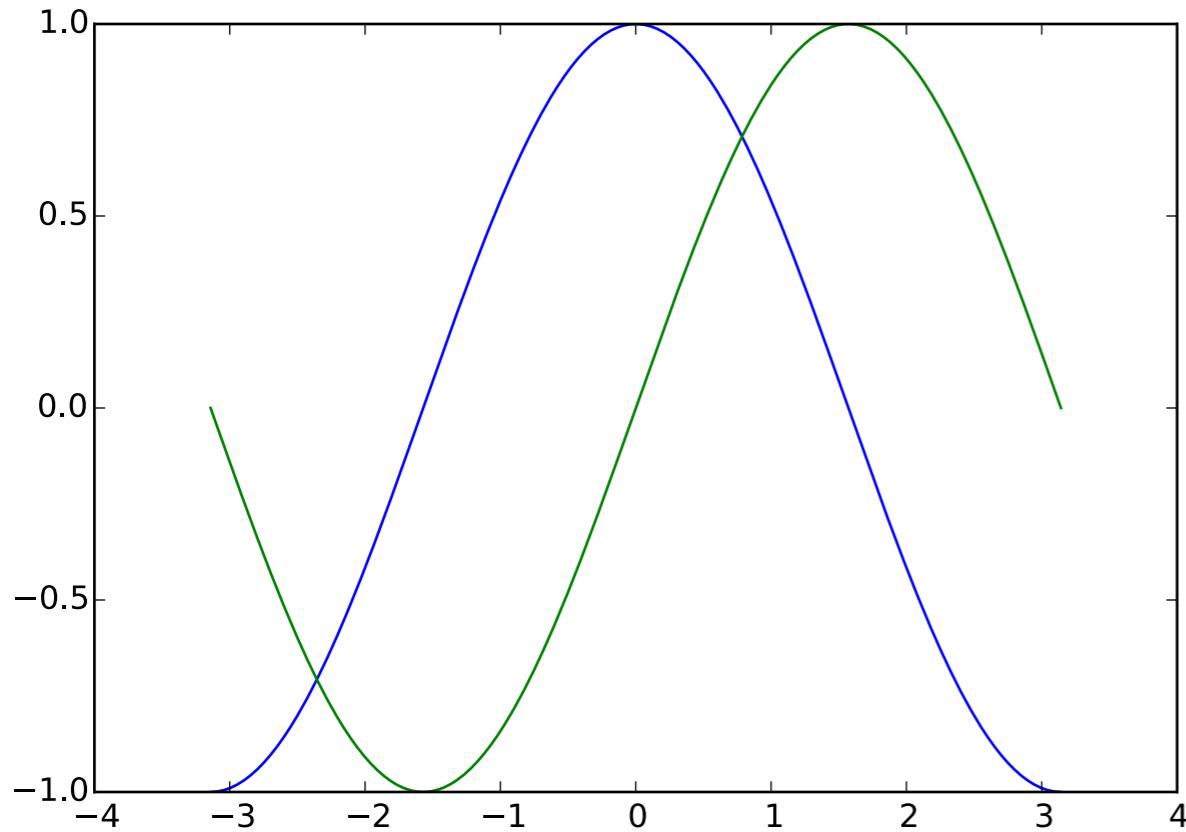
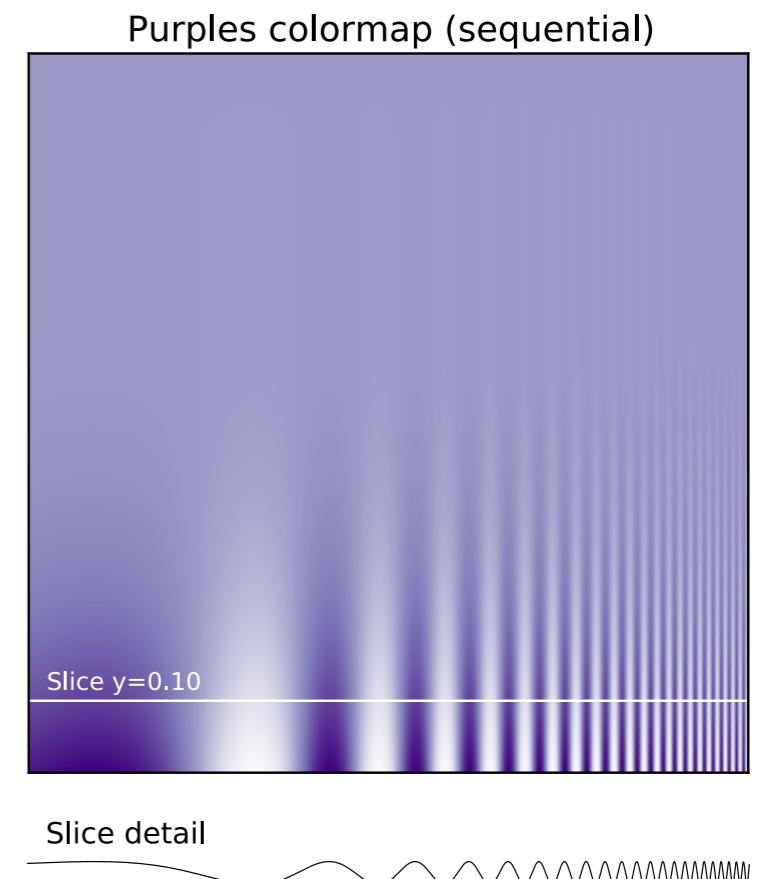
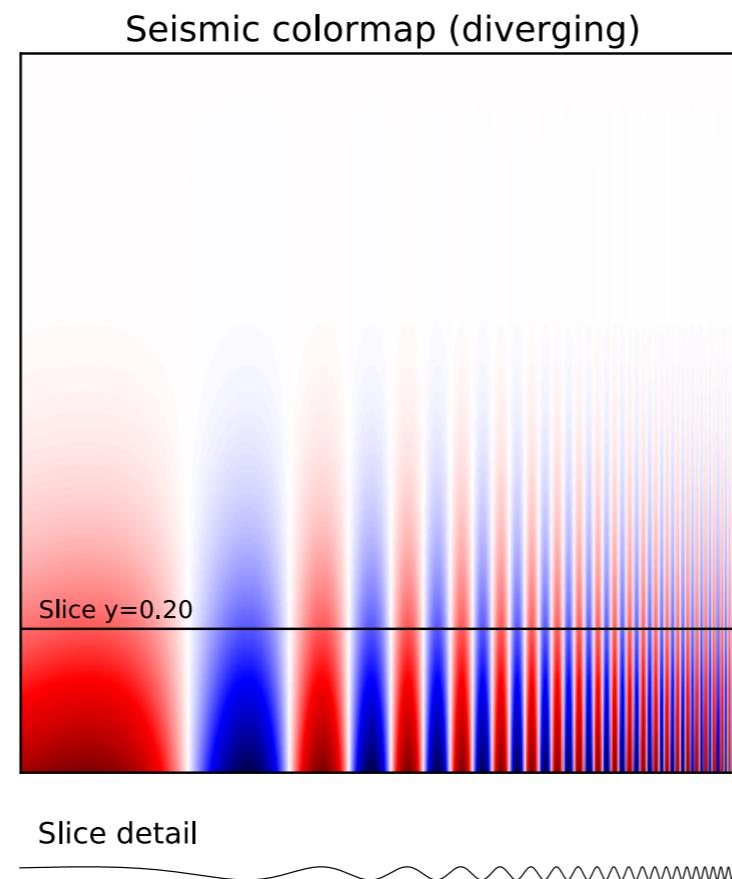
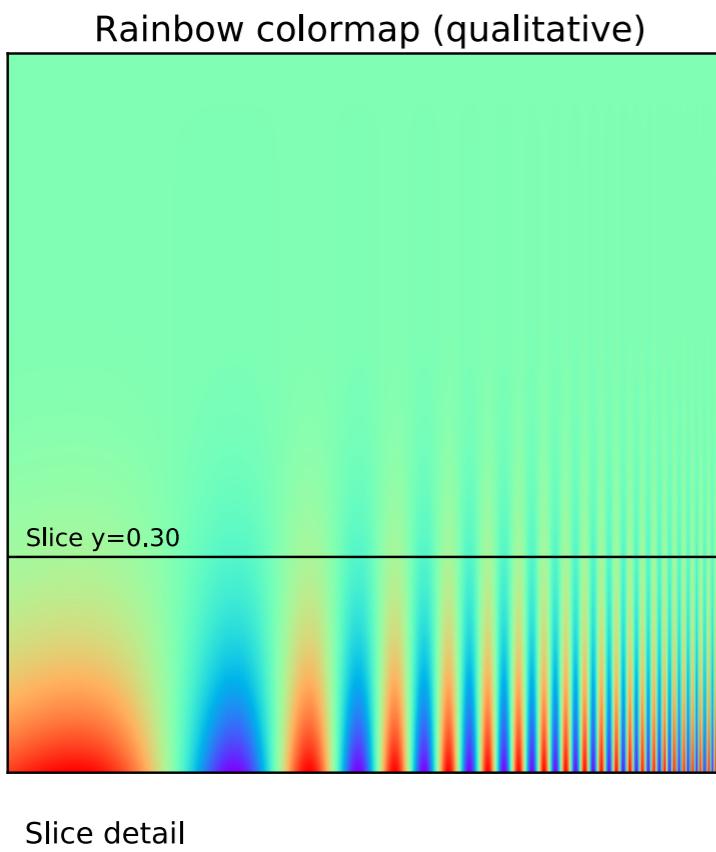


Figure 1. The A and B patches are actually the same color even though we perceived them at being different color.

Rule 5: Do not trust the defaults



Rule 6: Use color effectively

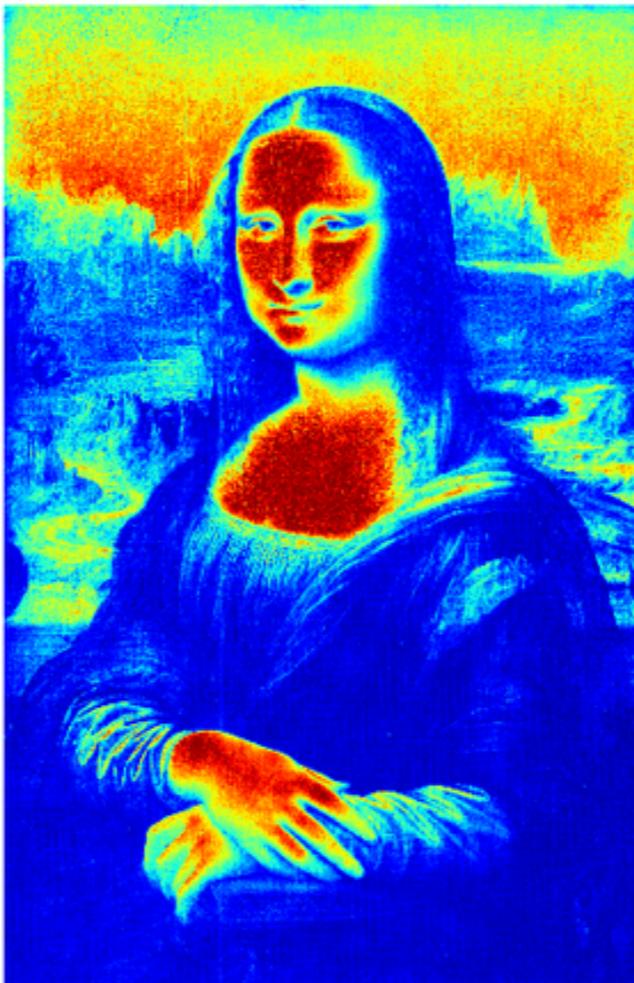


Rule 6 bis: Above all, do no harm (no jet, ever !)

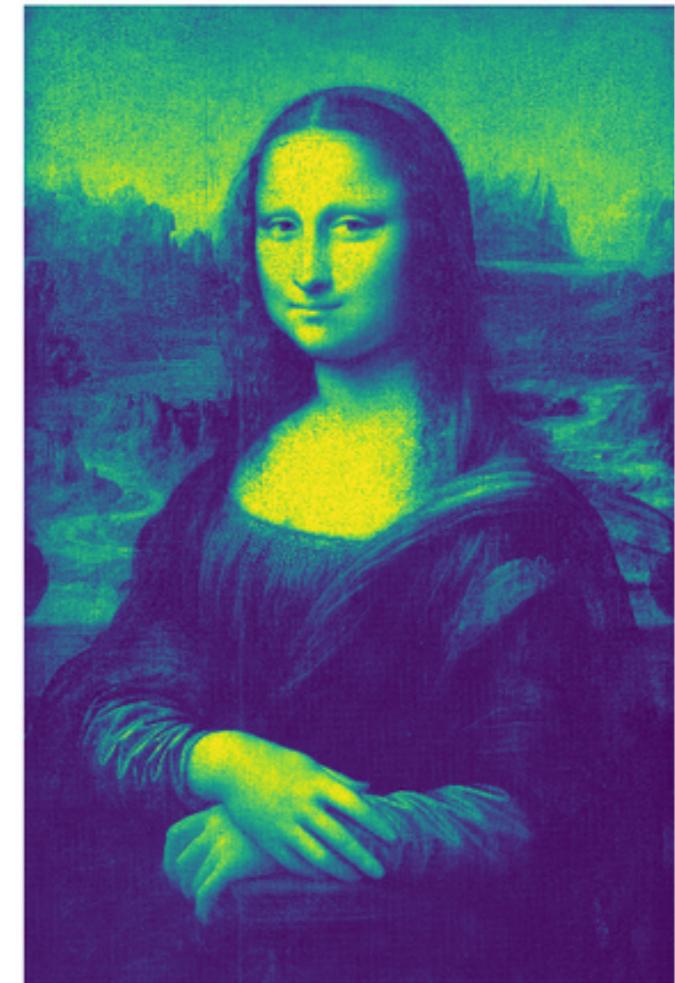
Colour



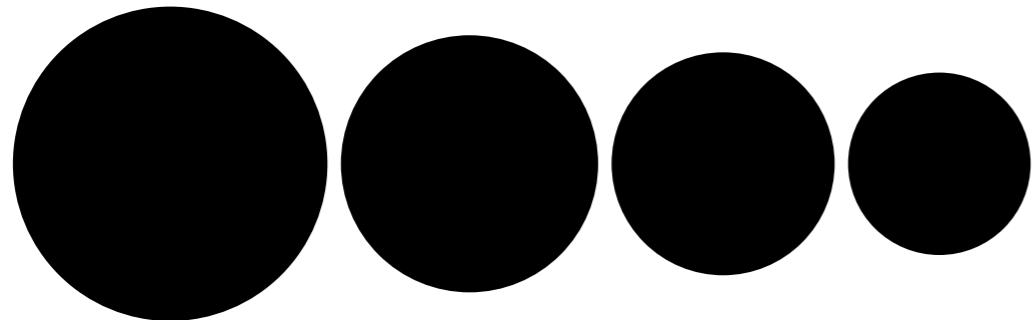
Jet



Viridis

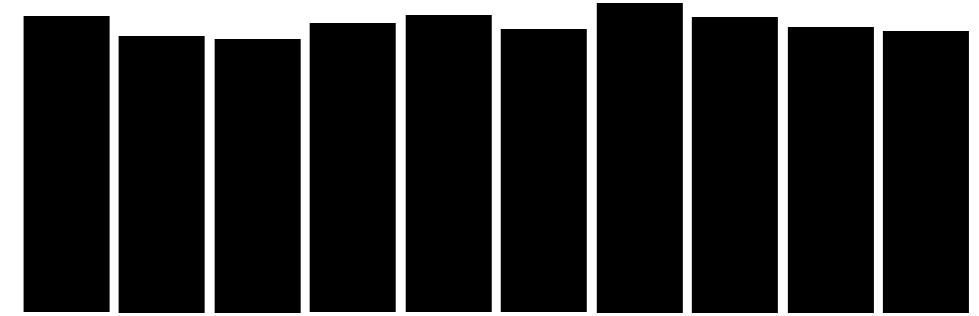


Rule 7: Do not mislead the reader



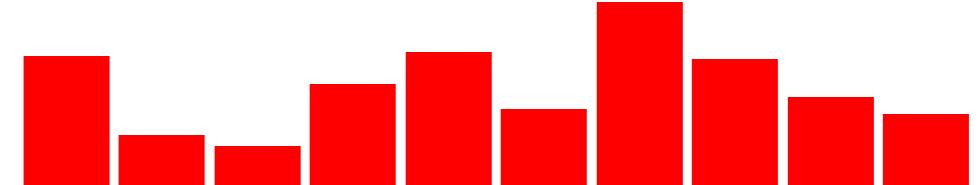
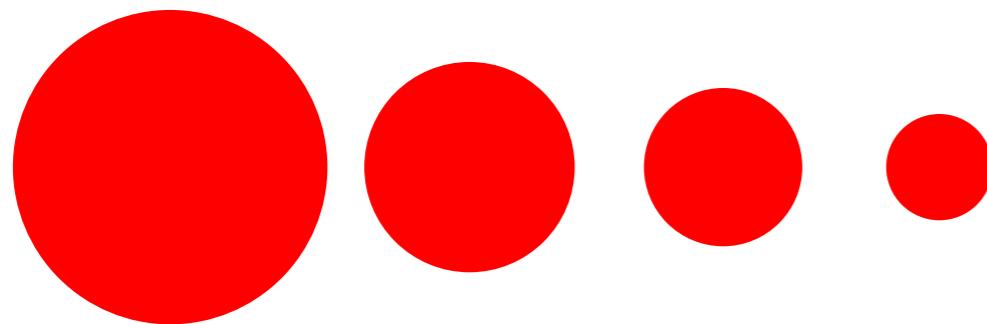
Relative size using disc area

Relative size using disc radius

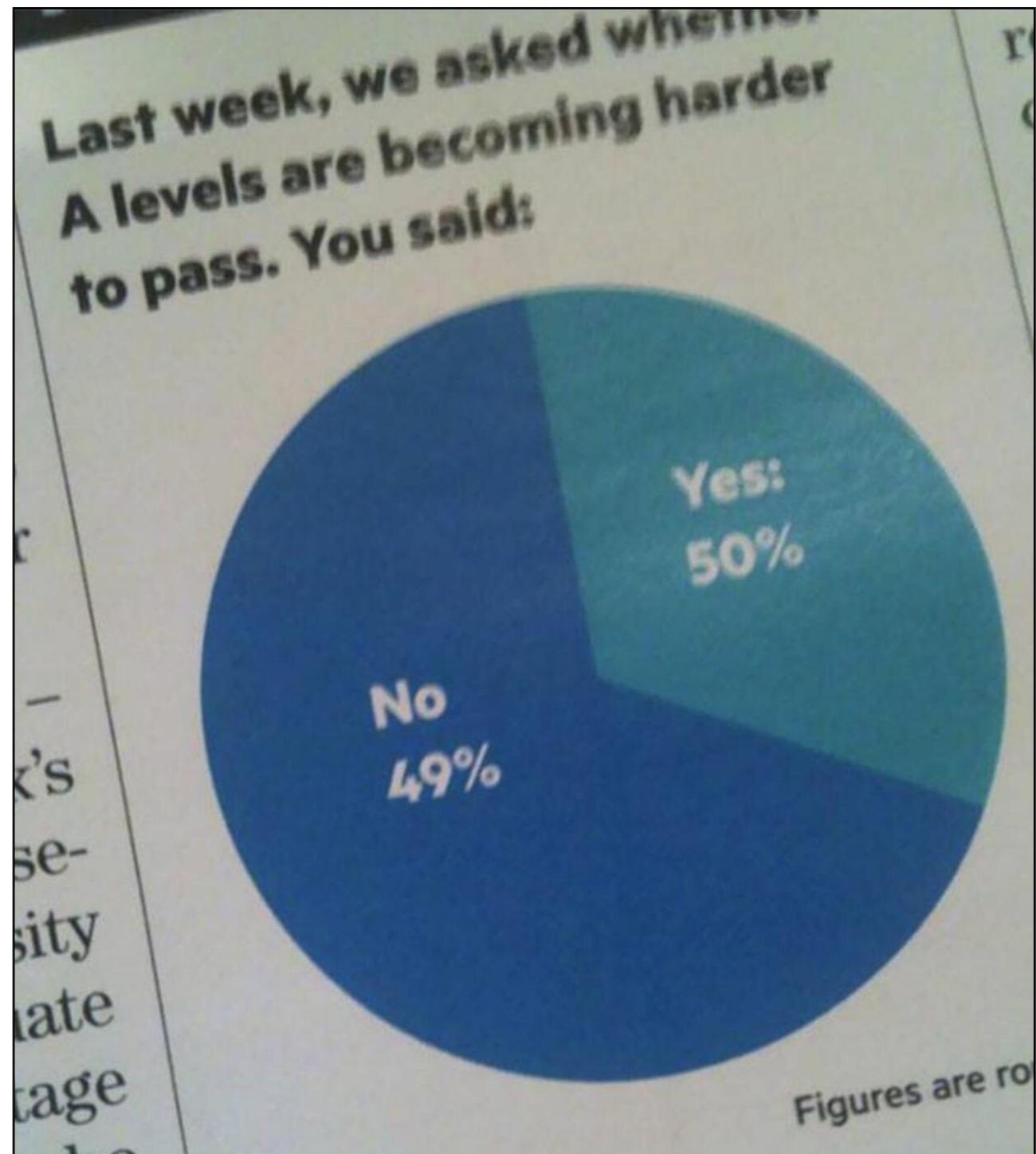
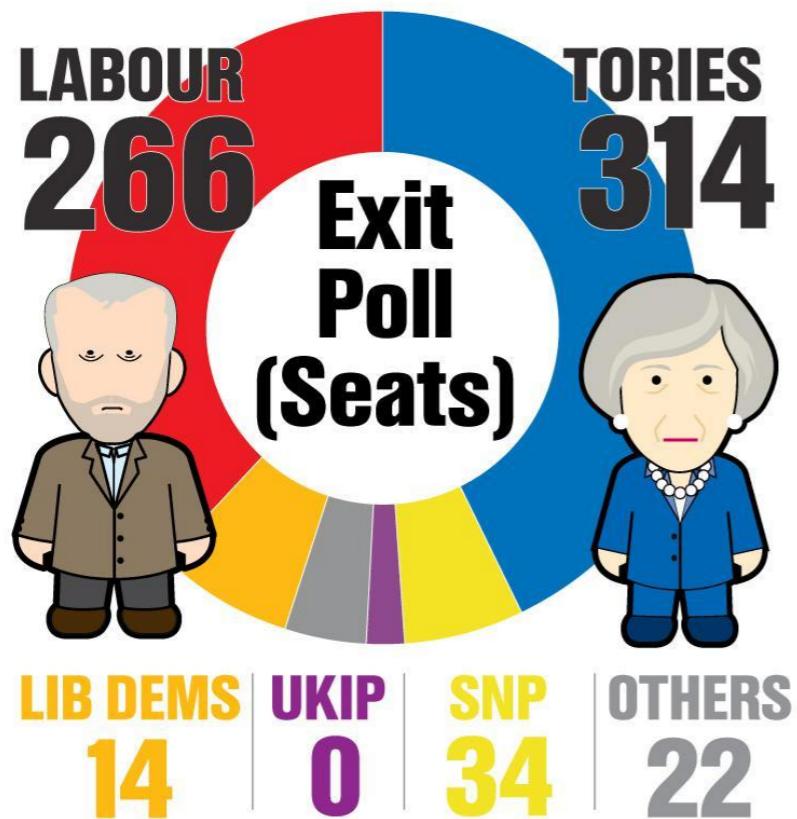


Relative size using full range

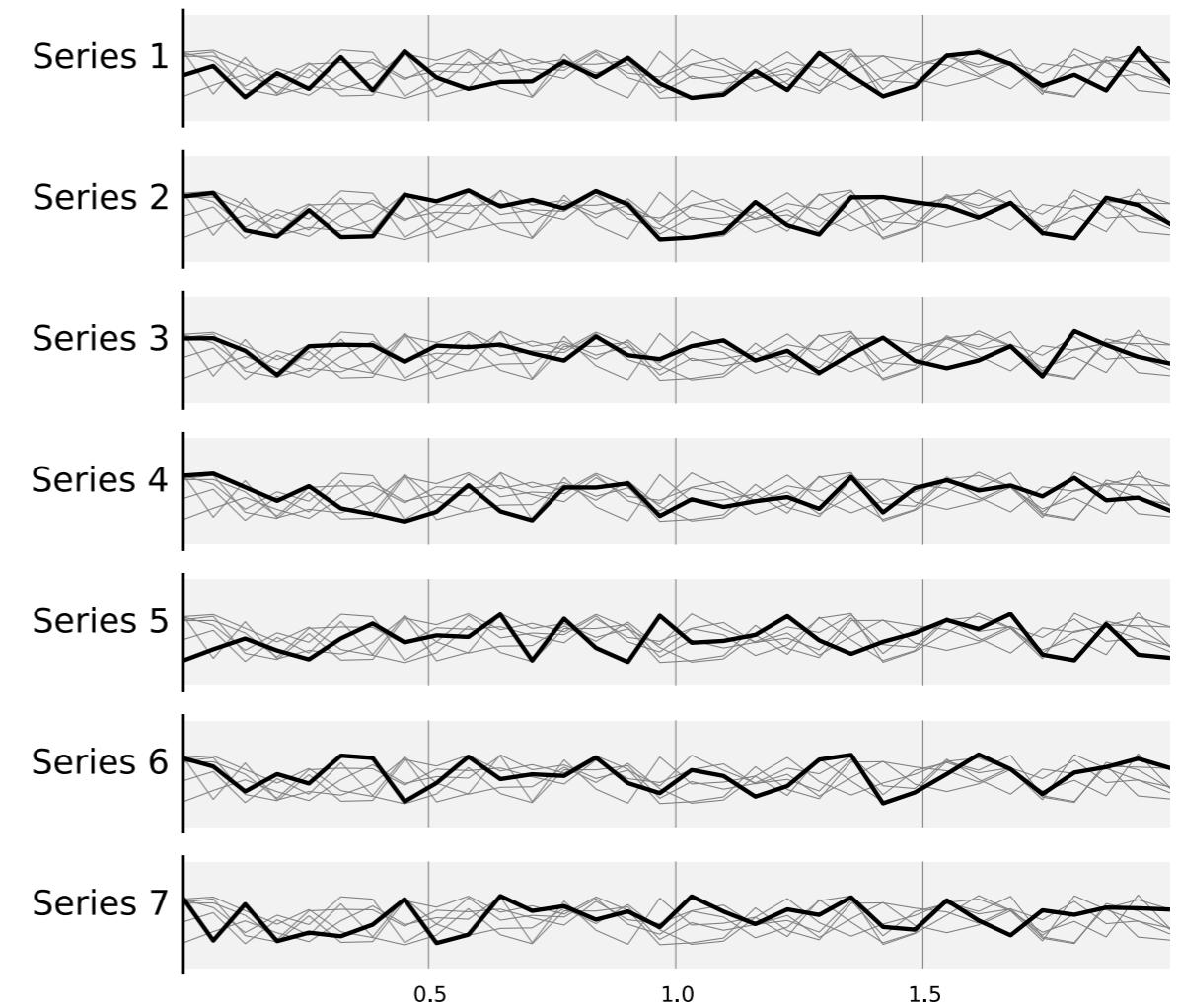
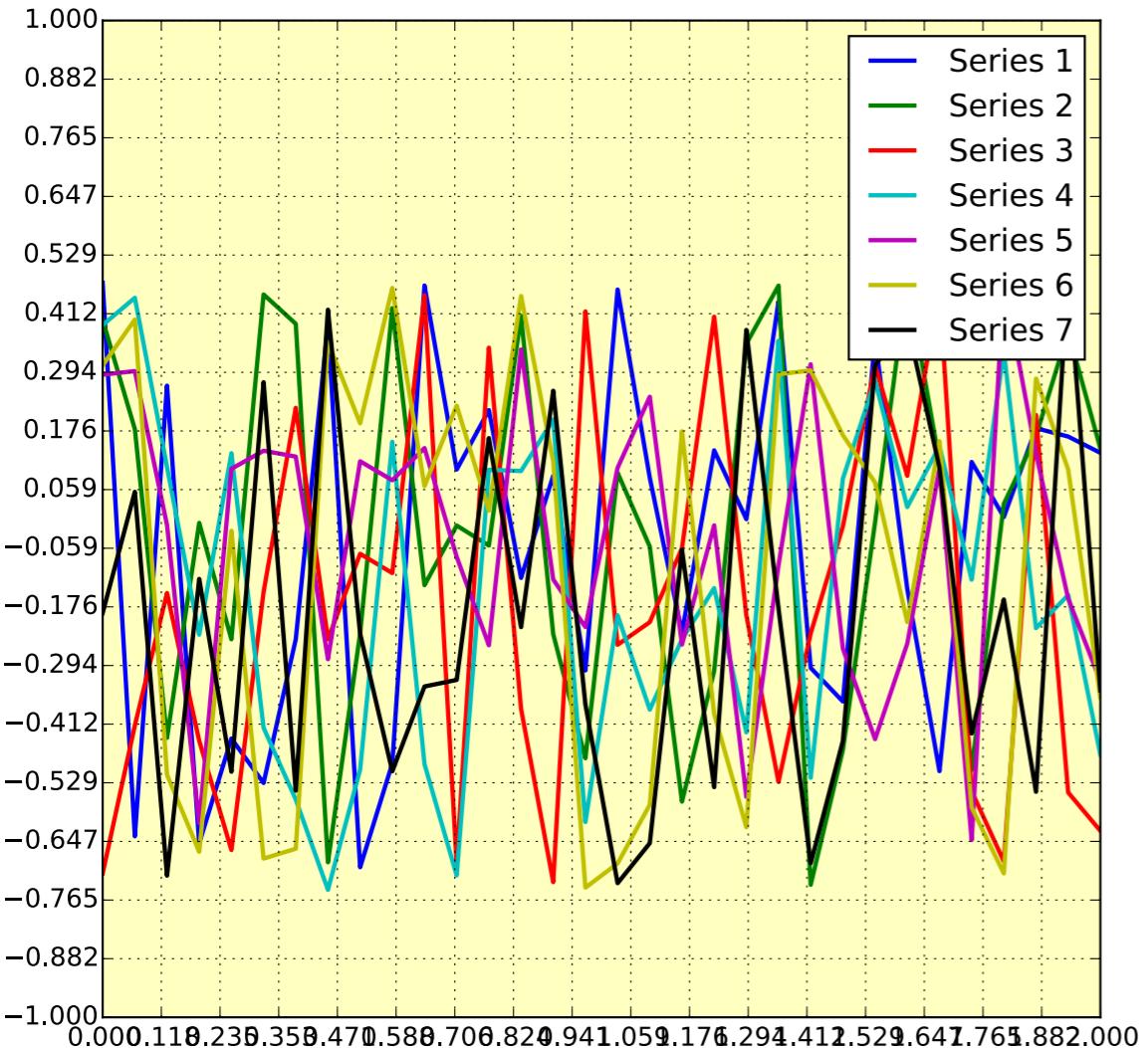
Relative size using partial range



Rule 7: Do not mislead the reader (bis)

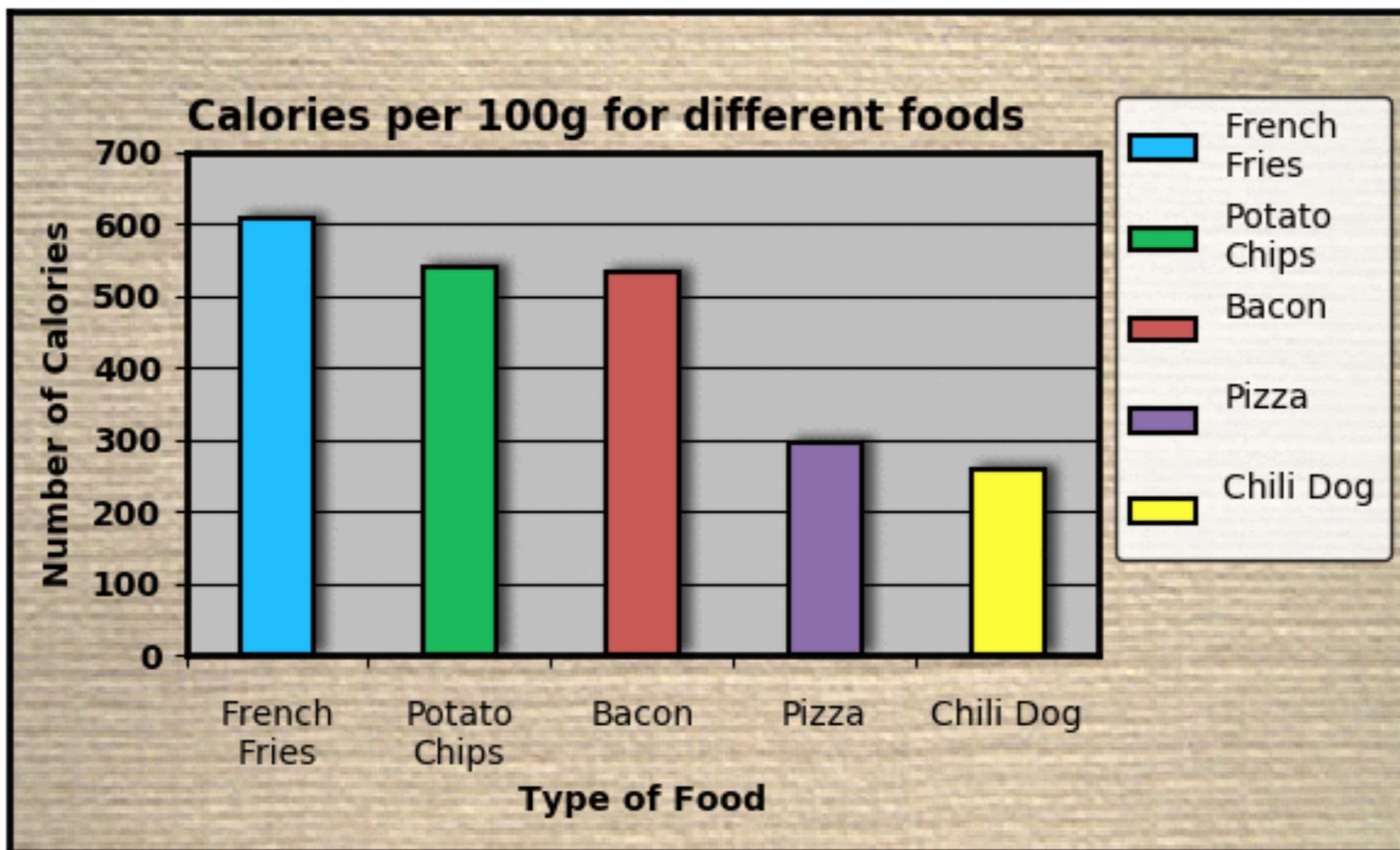


Rule 8: Avoid “Chartjunk”



Rule 8 bis: Less is more

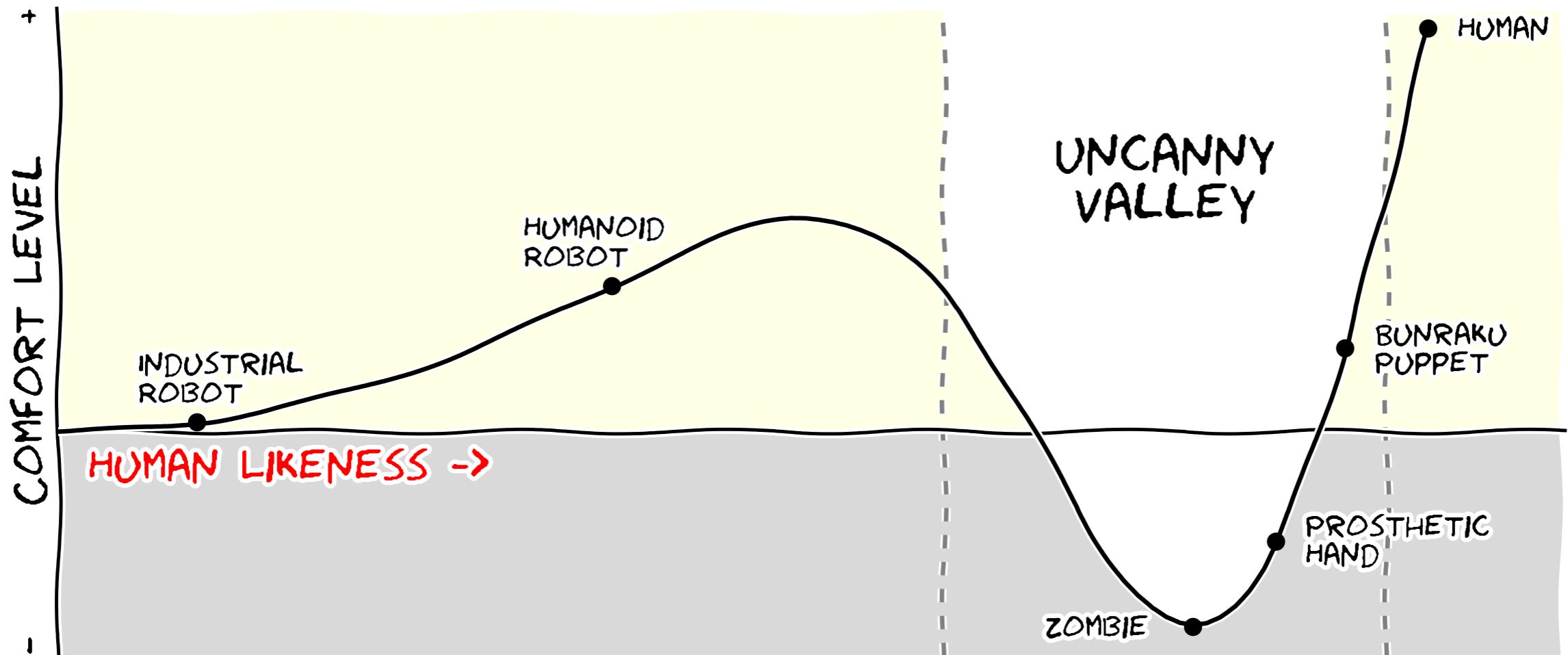
Less is More



A remake of www.darkhorseanalytics.com

Made with matplotlib 2.0

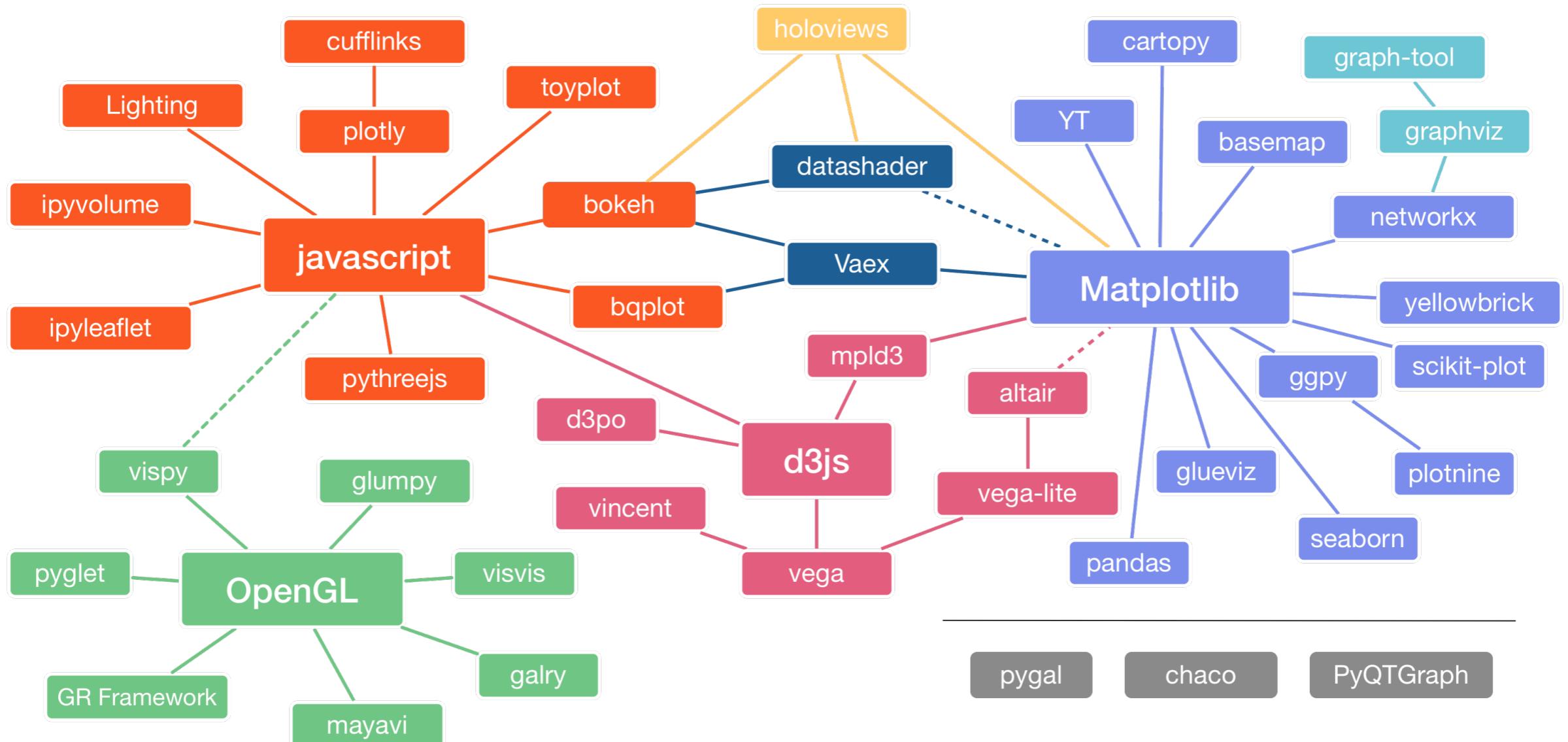
Rule 9: Message trumps beauty



Rule 10: Get the right tool

- PDFCrop (remove white borders)
<http://pdfcrop.sourceforge.net>
- GraphViz (easy graph)
<http://www.graphviz.org>
- ImageMagick (scripted image processing)
<http://www.imagemagick.org/script/index.php>
- Gimp (bitmap image manipulation)
<https://www.gimp.org>
- Inkscape (vector image manipulation)
<https://www.inkscape.org>
- Tikz (scripted vector art)
<http://www.texample.net/tikz/examples/all/>
- And many, many, many others ...

Rule 10: Get the right tool



Enough, theory, let's practice !

github.com/rougier/ASPP-2018-dataviz