SCIENTIFIC VISUALIZATION

Nicolas P. Rougier – Nicolas.Rougier@inria.fr
Inria - University of Bordeaux

November 14, 2021

ON THE IMPORTANCE OF VISION

... 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

What is common to these data sets?

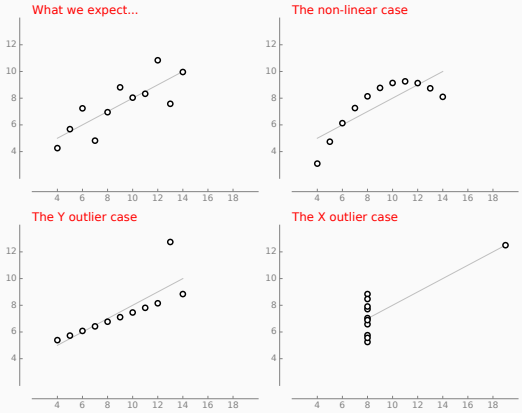
Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.12
Linear regression	y=3.00+0.500*x
R squared	0.666
p value	0.0021

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

ANSCOMBE'S QUARTET

What is common to these data sets?

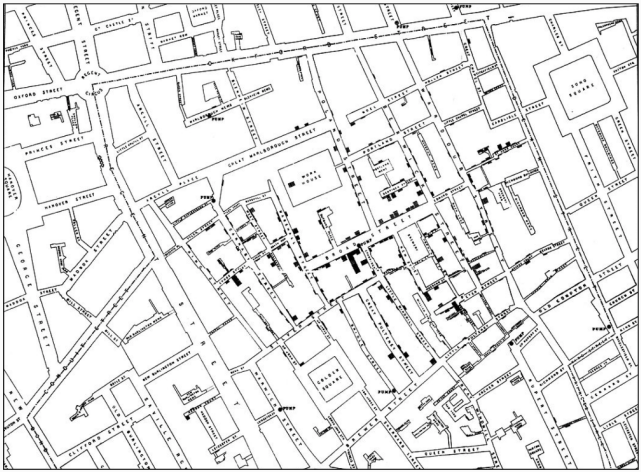
Mean of x	9
Sample variance of x	11
Mean of y	7.5
Sample variance of y	4.12
Linear regression	y=3.00+0.500*x
R squared	0.666
p value	0.0021



A computer should make both calculations and graphs – *Francis Anscombe (1918-2001)*

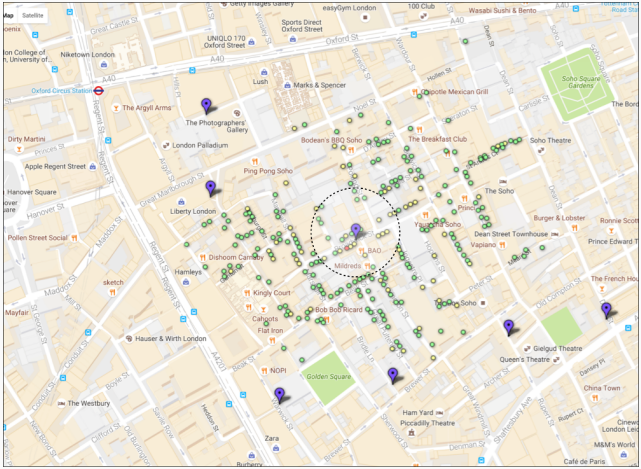
CHOLERA EPIDEMIC, LONDON, 1854

John Snow (1813-1858) is considered one of the fathers of modern epidemiology, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854.



CHOLERA EPIDEMIC, LONDON, 1854

John Snow (1813-1858) is considered one of the fathers of modern epidemiology, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854.

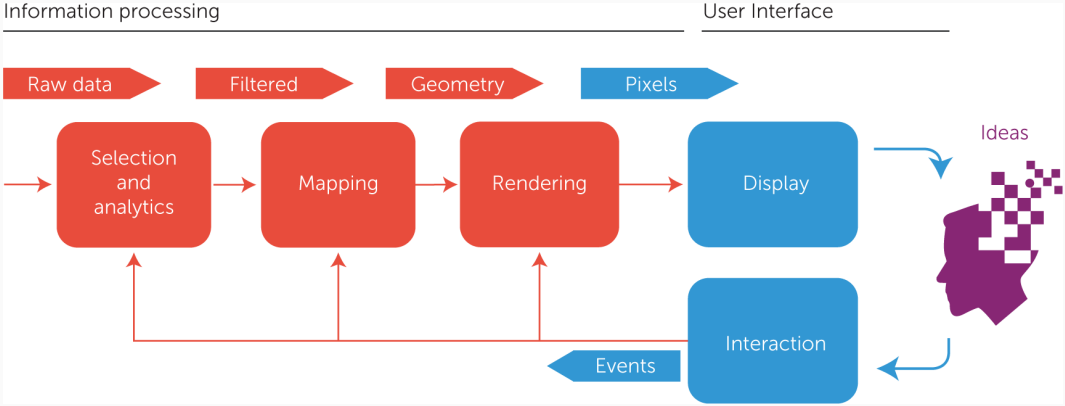


WHAT IS DATA VISUALISATION?

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

THE VISUALIZATION PIPELINE



From Scalable Real-Time Visualization Using the Cloud, Holliman & Watson, 2015.

QUANTITATIVE VS QUALITATIVE DATA

Quantitative (values or observations that can be measured)

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

Qualitative (values or observations that can be sorted into groups or categories)

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

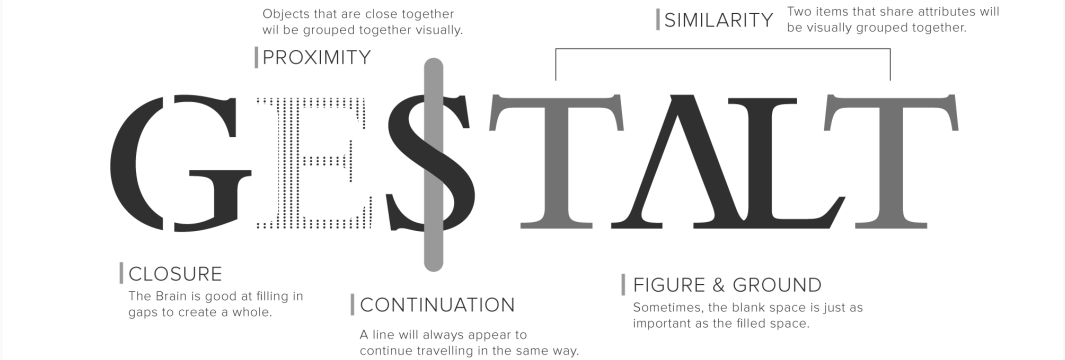
GRAPHICAL ELEMENTS

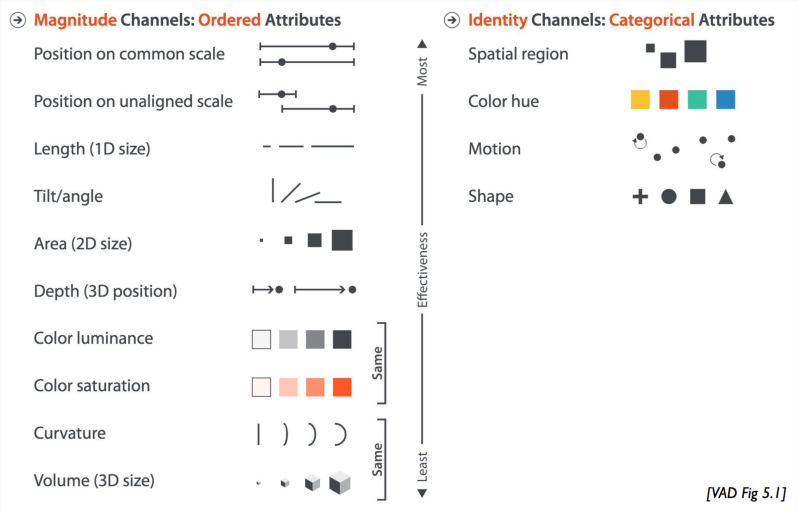
A scientific figure can be fully described by a set of graphic primitives with different attributes:

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

Questions is thus how to organize and link them to the underlying data.

PRINCIPLES OF VISUAL PERCEPTION



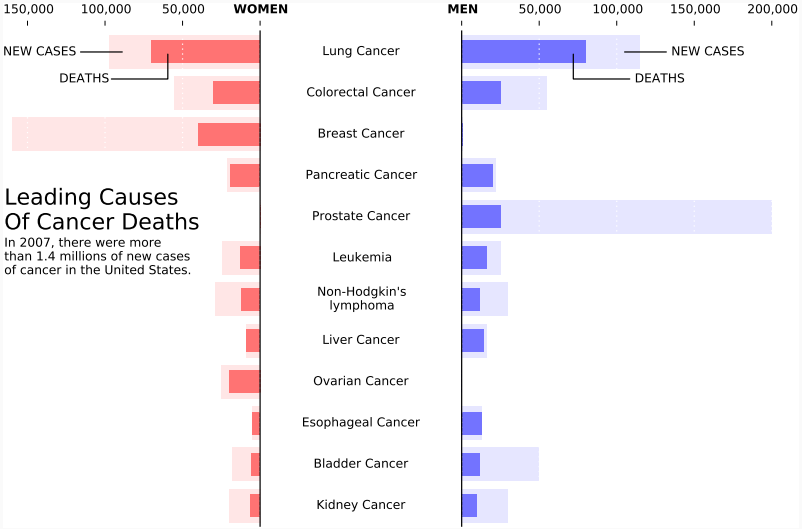

Magnitude Channels: Ordered Attributes
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Most
Effectiveness
Least
Identity Channels: Categorical Attributes
Spatial region
Color hue
Motion
Shape
[VAD Fig 5.1]

DATA VISUALIZATION CATALOGUE (S. RECEBBA)

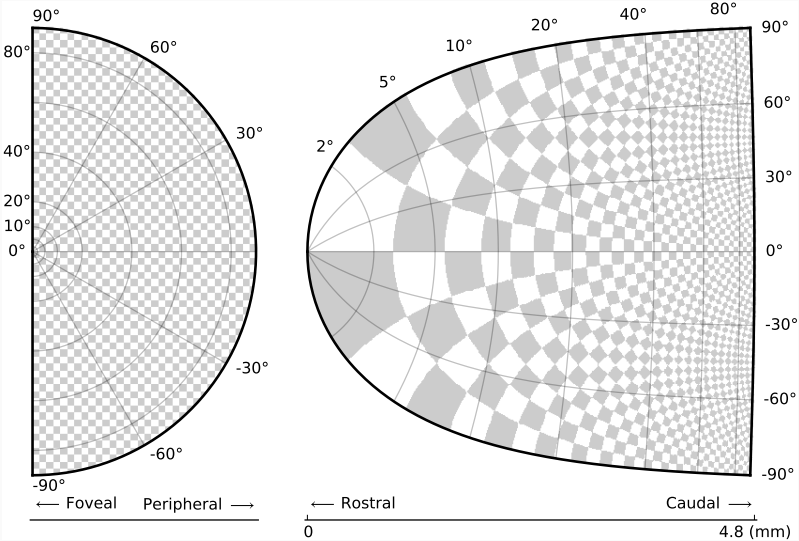
datavizcatalogue.com

Ten simple rules for better figures

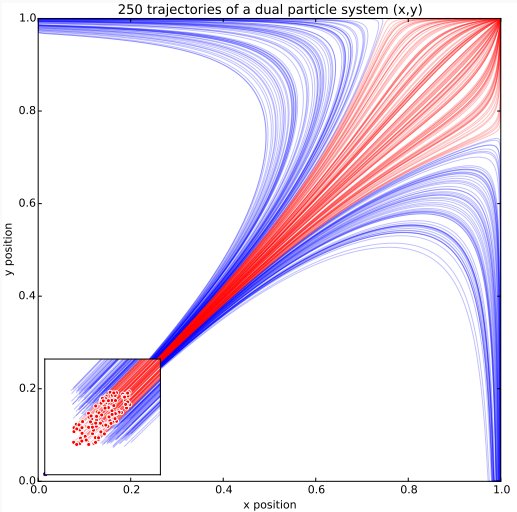
RULE 1: KNOW YOUR AUDIENCE

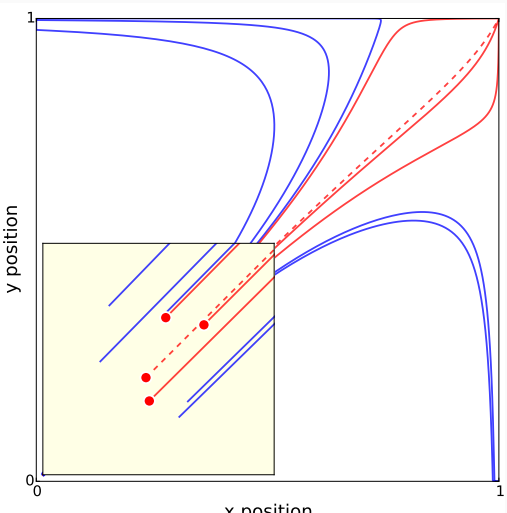


RULE 2: IDENTIFY YOUR MESSAGE

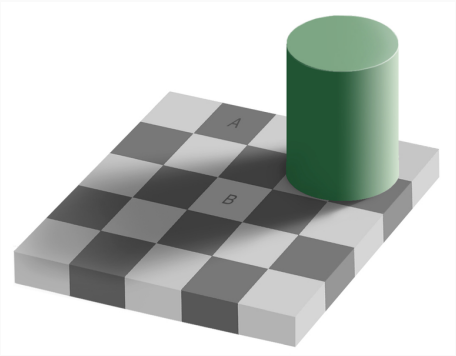


RULE 3: ADAPT THE FIGURE

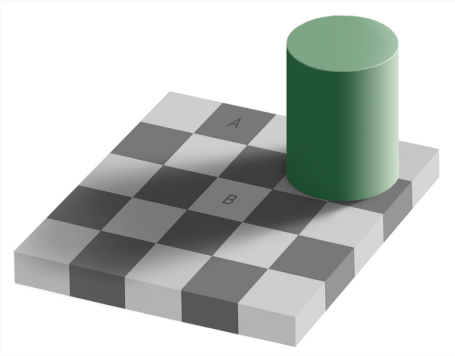




RULE 4: CAPTIONS ARE NOT OPTIONAL

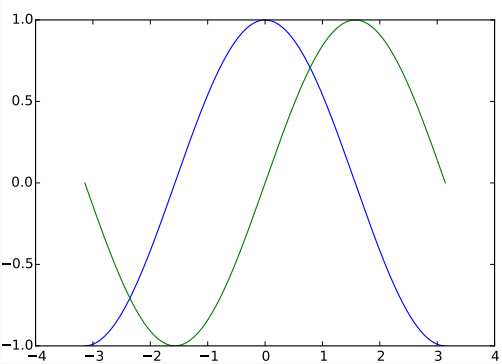


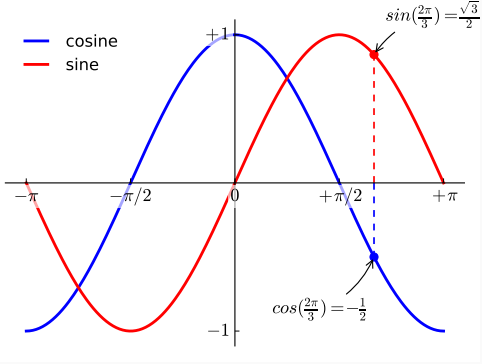
Optical illusion



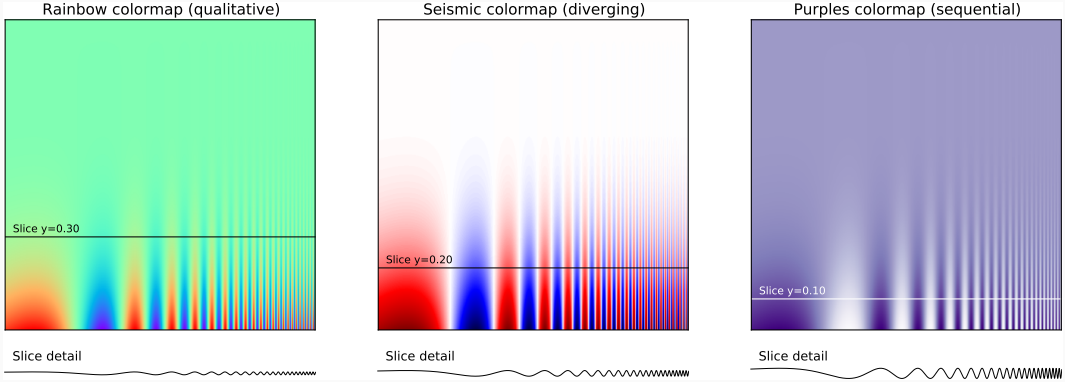
The A and B patches are actually the same color even though we perceived them at being different color.

RULE 5: DO NOT TRUST THE DEFAULTS

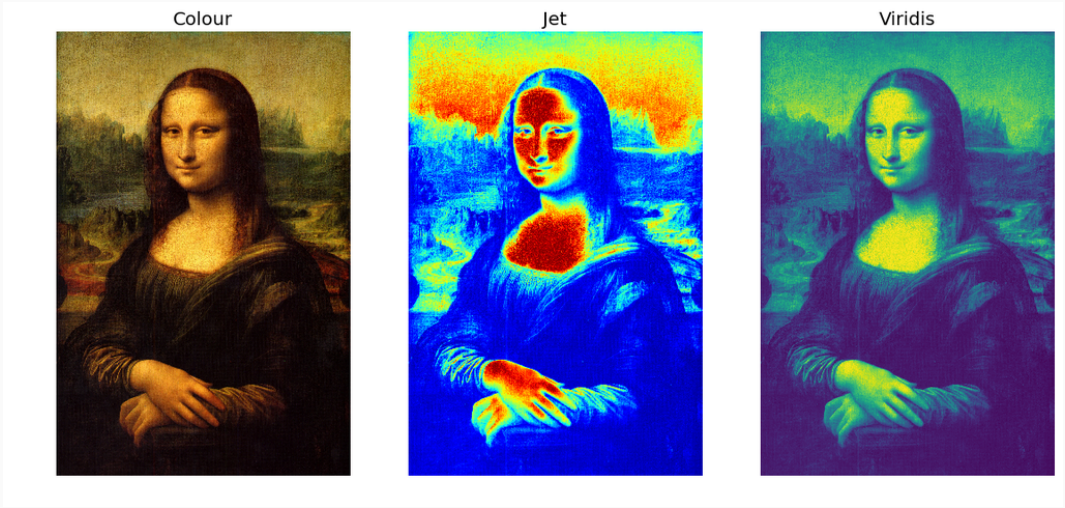




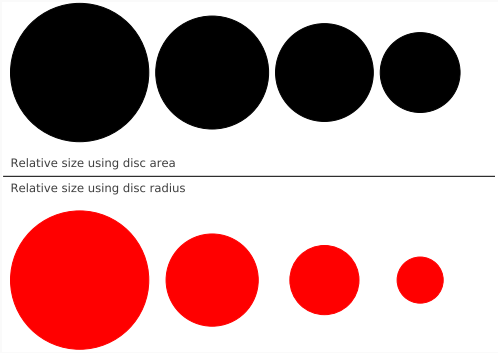
RULE 6: USE COLOR EFFECTIVELY



RULE 6 BIS: ABOVE ALL, NO JET. EVER.

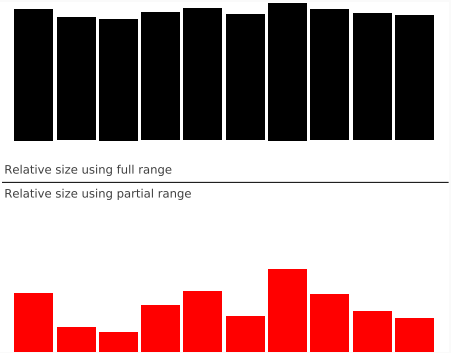


RULE 7: DO NOT MISLEAD THE READER



Relative size using disc area

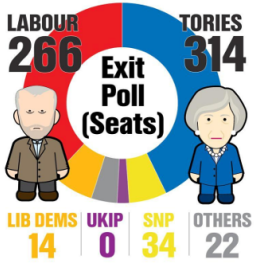
Relative size using disc radius

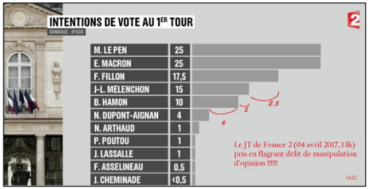


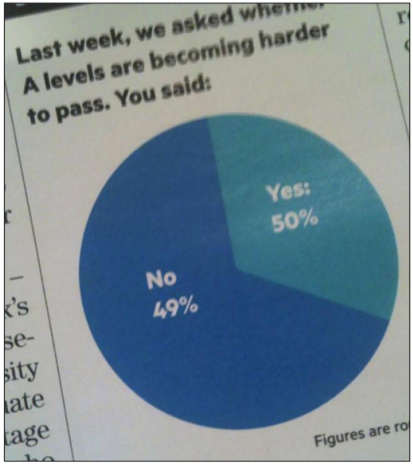
Relative size using full range

Relative size using partial range

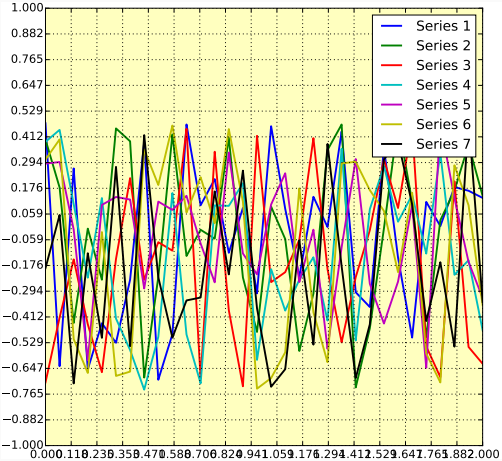
RULE 7: DO NOT MISLEAD THE READER. REALLY.

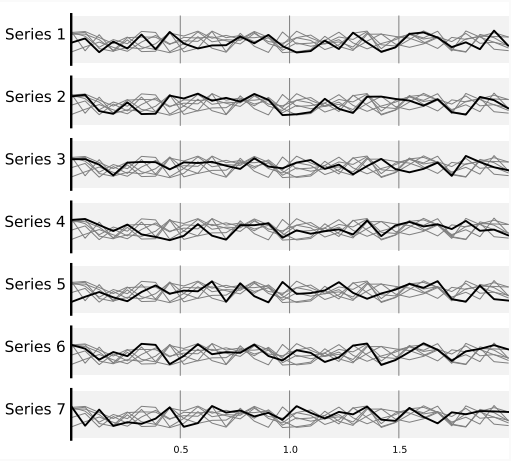




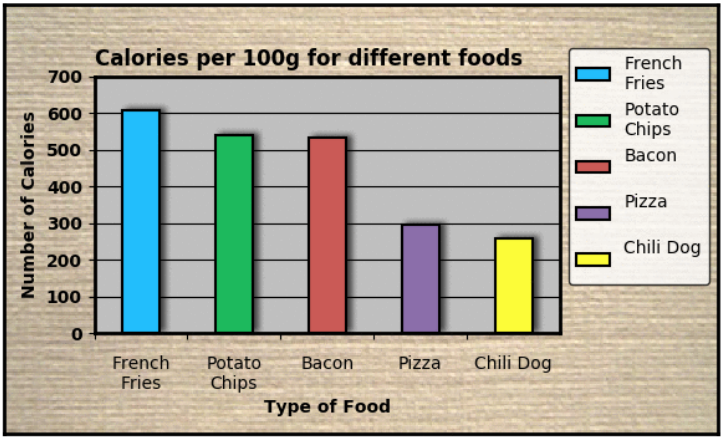


RULE 8: AVOID CHARTJUNK





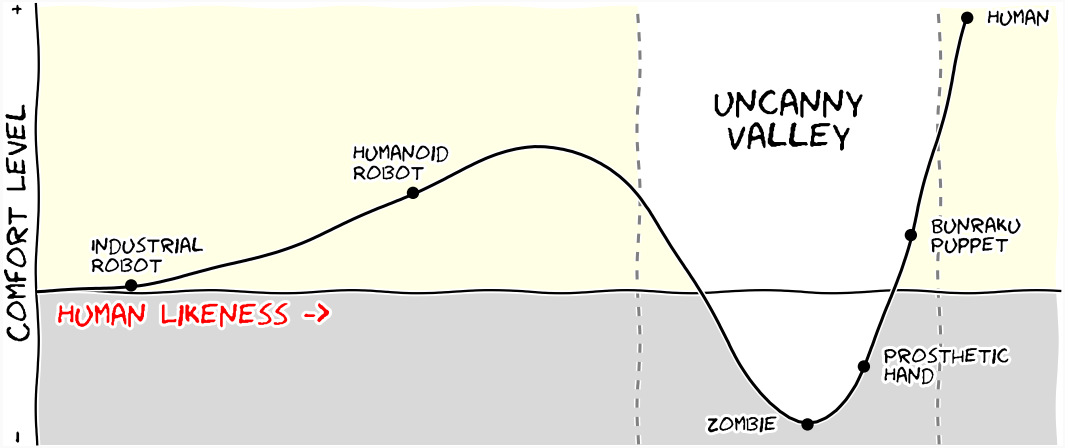
Less is More



A remake of www.darkhorseanalytics.com

Made with matplotlib 2.0

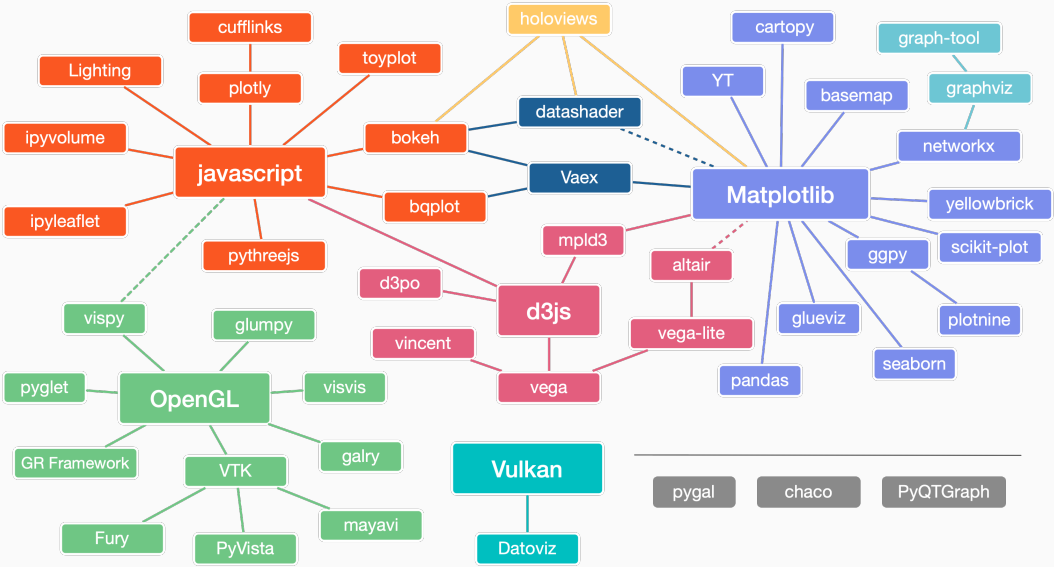
RULE 9: MESSAGE TRUMPS BEAUTY



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



A NOTE ABOUT FORMATS

Standard data formats

CSV – Comma-Separated Values

JSON – JavaScript Object Notation

Standard vector formats

PDF – Portable Document Format

SVG – Scalable Vector Graphics

Standard bitmap formats

PNG – Portable Network Graphics (lossless)

JPG – Joint Photographic Experts Group (lossy)

ANALYZING 1.1 BILLION NYC TAXI AND UBER TRIPS, WITH A VENGEANCE

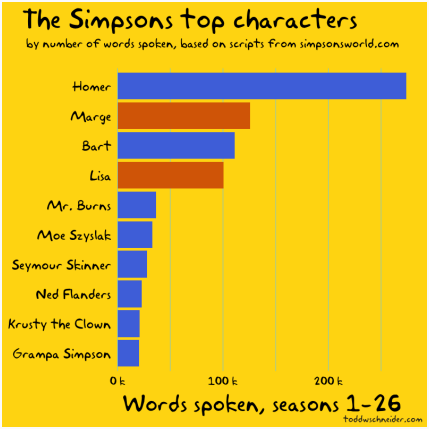
The New York City Taxi & Limousine Commission has released a staggeringly detailed historical dataset covering over 1.1 billion individual taxi trips in the city from January 2009 through June 2015. Taken as a whole, the detailed trip-level data is more than just a vast list of taxi pickup and drop off coordinates: it's a story of New York.

Todd W. Schneider (`toddwschneider.com`)

THE SIMPSONS BY THE DATA

Analysis of 27 seasons of Simpsons data reveals the show's most significant side characters, a pattern of patriarchy, declining TV ratings, and more

Todd W. Schneider (`toddwschneider.com`)



CONCLUSION

Schedule

- 15/11/2021 : Introduction + study
- 06/12/2021 : Dataviz catalogue + project
- 13/12/2021 : Layout and projection + project
- 10/01/2021 : Advanced concepts + project
- 24/01/2021 : Project

Project

The goal of the project is to analyse wedding data (from INSEE) and to produce a one page PDF comparing data from 2018 and 2019. Your PDF must have at least 3 figures (with caption) and contains text to introduce your analysis.

REFERENCES

Books

- Scientific Visualization: Pyhpn + Matplotlib, N. Rougier, 2021
- Fundamentals of Data Visualization, C. Wilke, 2018
- Visualization Analysis and Design ($), T. Munzner, 2014.
- Trees, maps, and theorems ($), J.-L. Doumont, 2009.
- The Visual Display of Quantitative Information ($), E.R. Tufte, 1983.

Other resources

- A Tour through the Visualization Zoo, J. Heer, M. Bostock, and V. Ogievetsky, 2010.
- The most misleading charts of 2015, fixed, K. Collins, 2015.
- Data is beautiful / reddit.
- From data to viz