Introduction to Data Visualization

Morine Amutorine
Benjamin Akera
Elaine Nsoesie

Instructor introductions …

Contacts:

Morine - morine.amutorine@one.un.org / Twitter: @M_moryn

Elaine - onelaine@bu.edu / Twitter: @ensoesie

Ben - akeraben@gmail.com

Resources

Github Repo
https://github.com/ensoesie/DSA_Visualization

Google Trends
https://trends.google.com

Twitter
https://developer.twitter.com

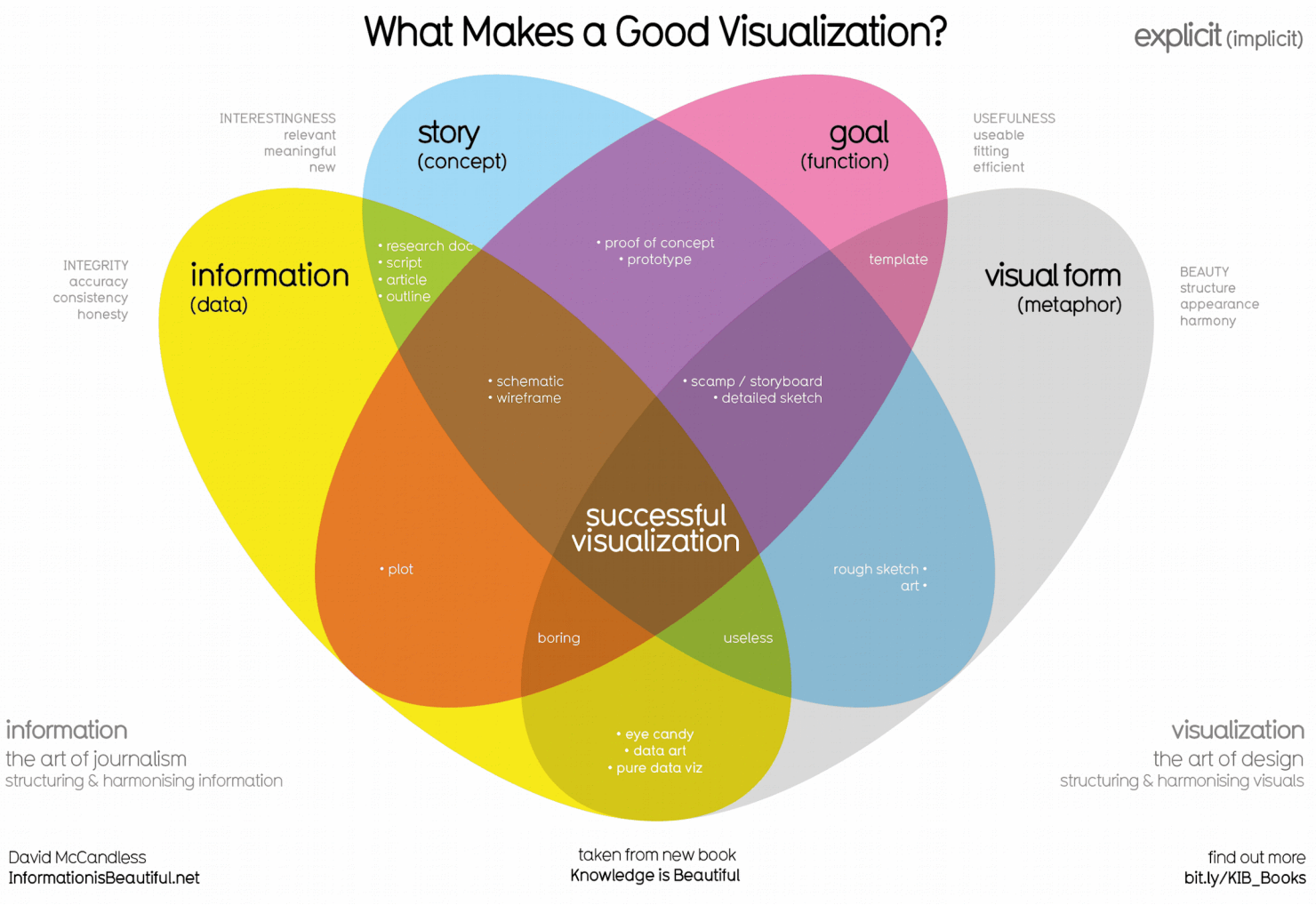
Why visualize data?

A picture is worth a thousand words

It is easier to remember pictures than text

Useful for understanding data

Can summarize large amounts of complex data



Visualization in Data Science can be used to:

- Explore data
- Analyze data
- Communicate findings
- Quickly draw attention to key messages

How to use visualizations to communicate effectively?

1 Decide on what your visualization should convey

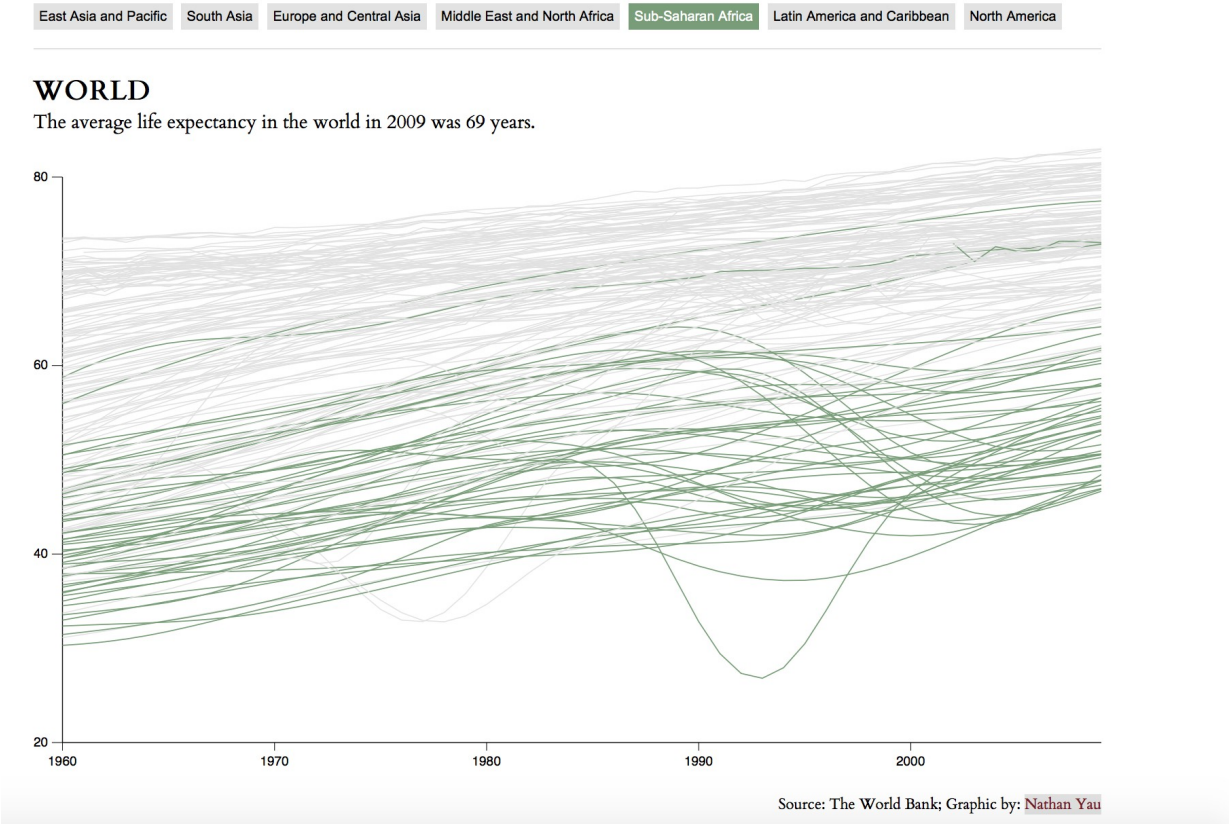
FOCUS ON THE DATA

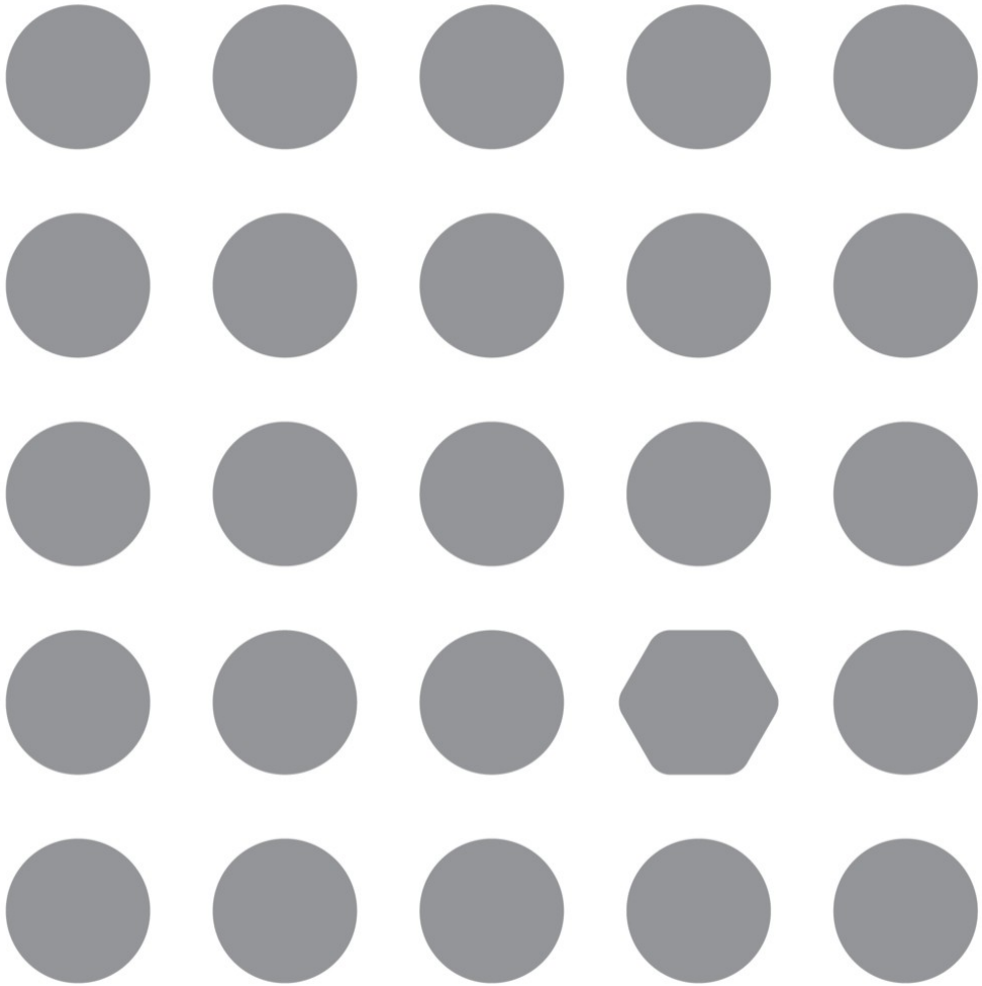
The style and structure of your visualization will depend on its purpose

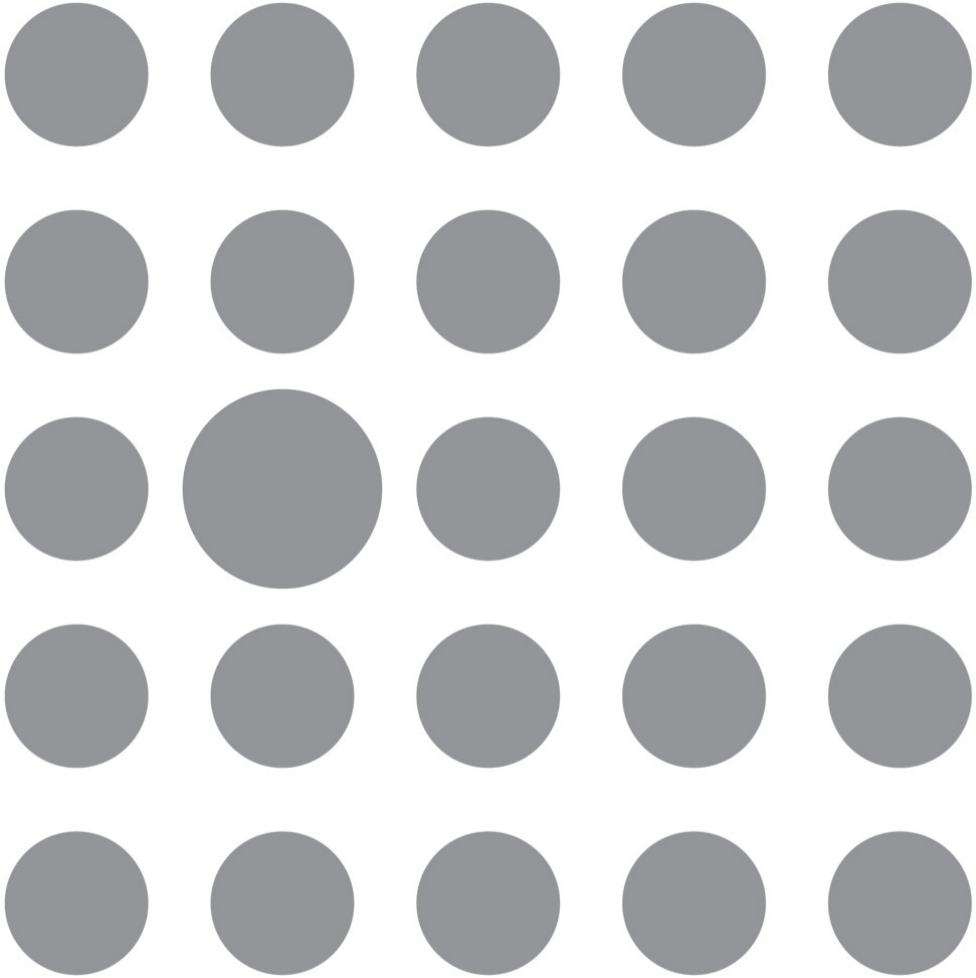
Design for a specific audience

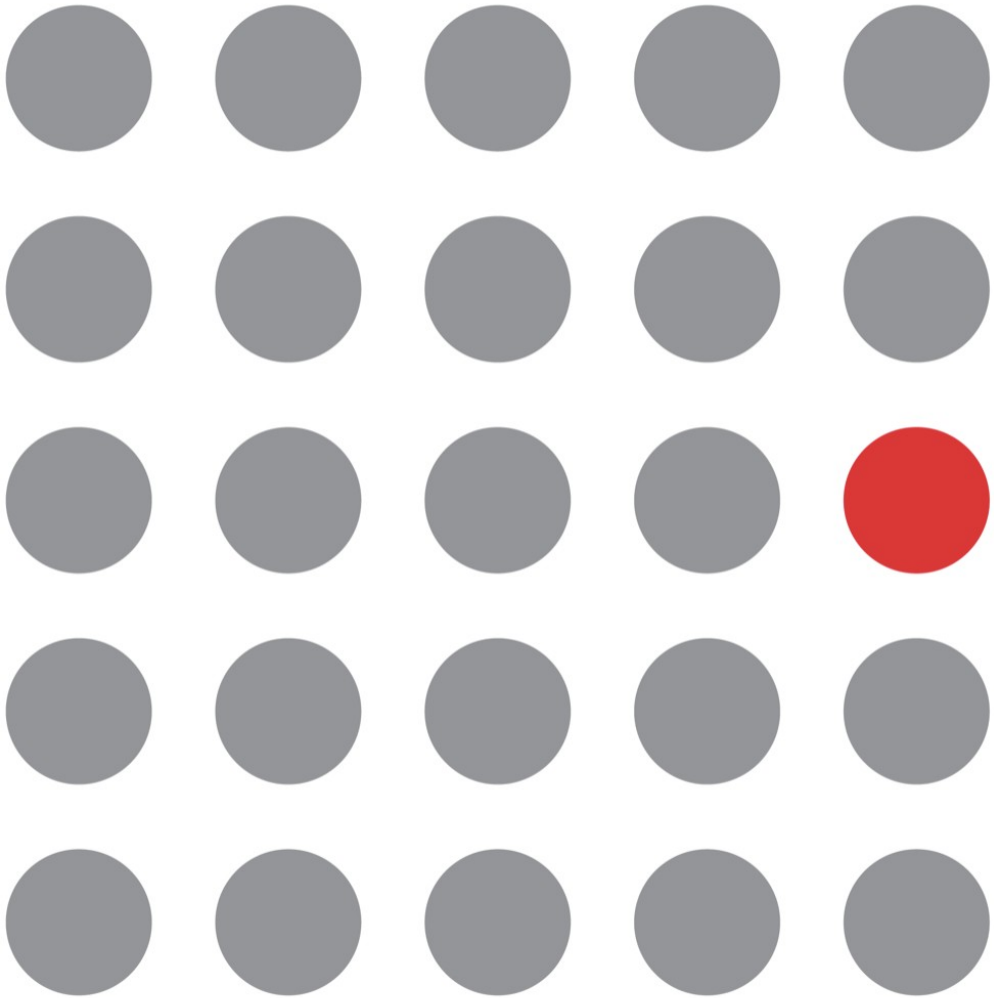
Tell a good story with a clear message

2 Use color and size to highlight and suppress information





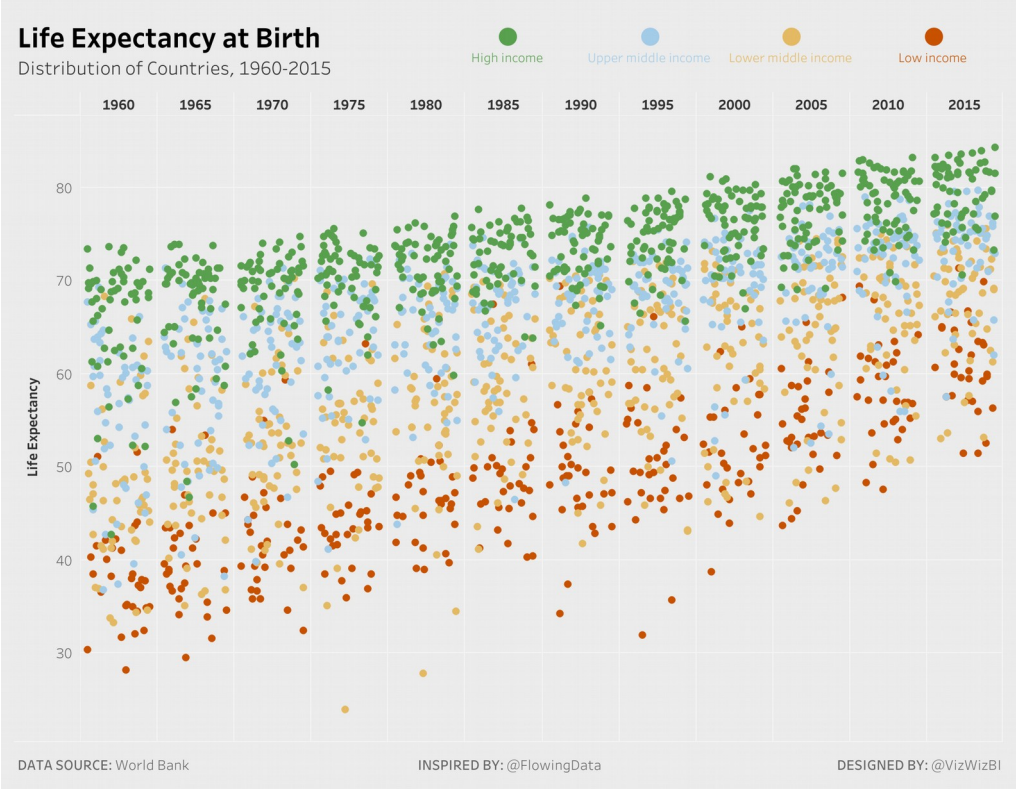




https://venngage.com/blog/how-to-pick-colors/

3 Use length and position to express quantitative information. Use color for categorical information

Scatter plots and **bar charts** allow for **more accurate comparison of information over time** compared to pie charts

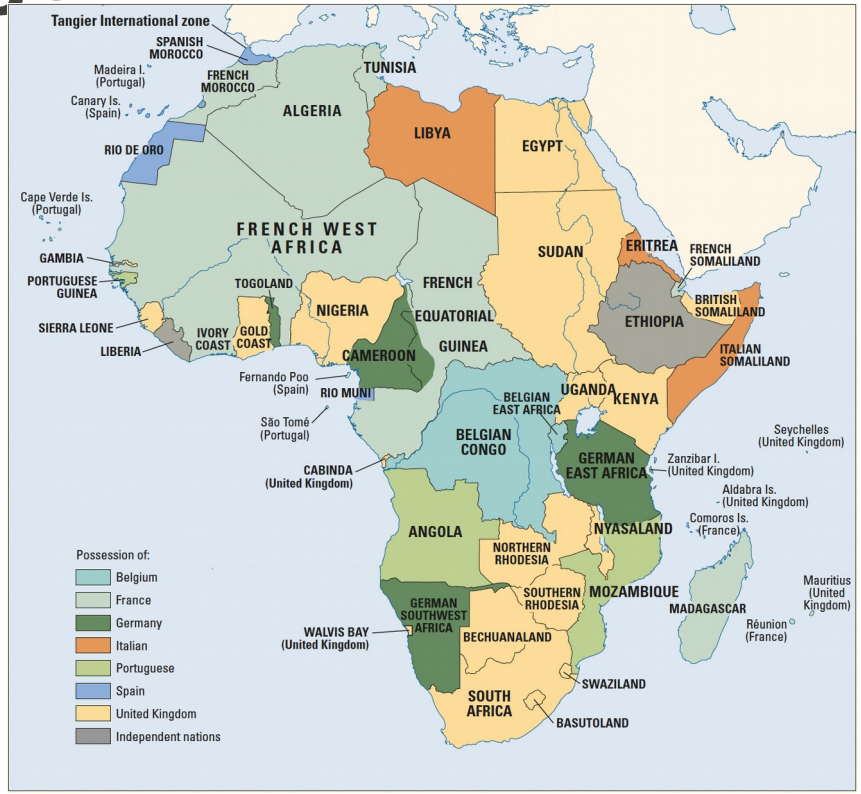


4 Think carefully about color selection and usage

Use color to create groupings

Add a **single color** to a black and white image

Use black and white to add contrast to an image with a single color gradient



https://africaindata.org/#/3

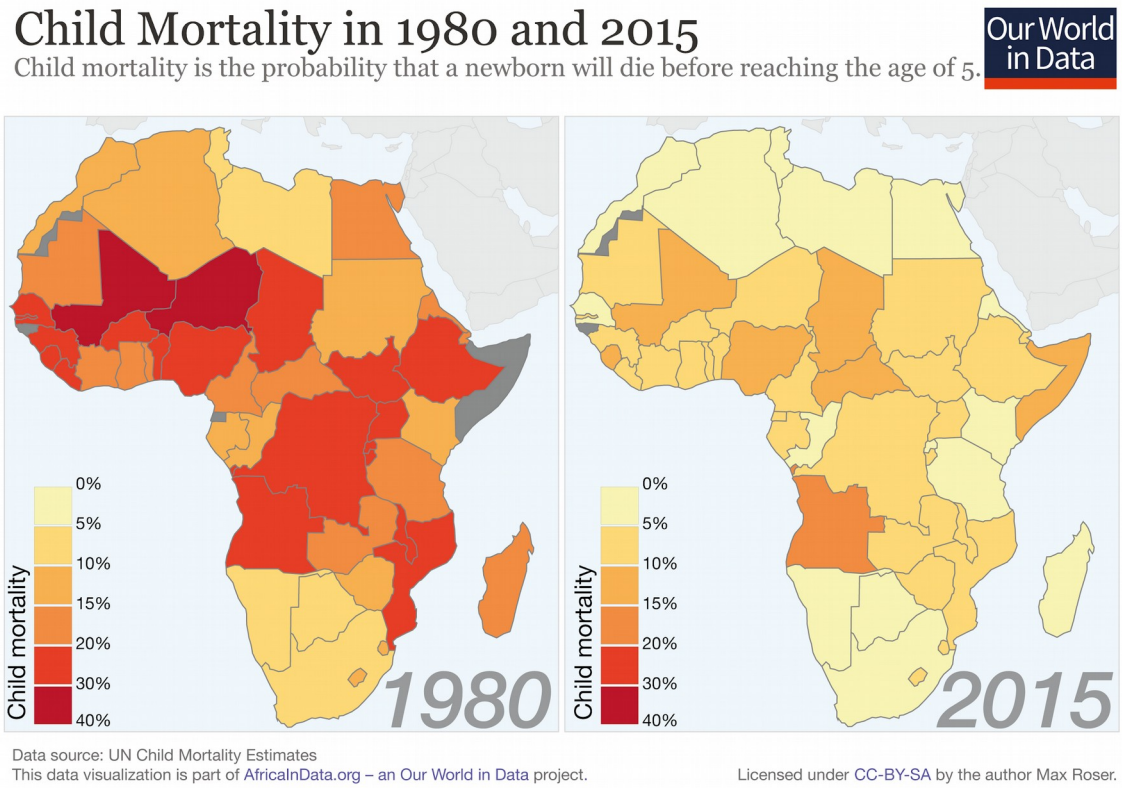
4 Think carefully about color selection and usage

Some colors have pre-established meanings

Consider those with color blindness

Red	Green	Blue
Stop	Moving	Water
Dangerous	Money	Cool
Hot	Plants	Safe

5 Use all available space and proper scales

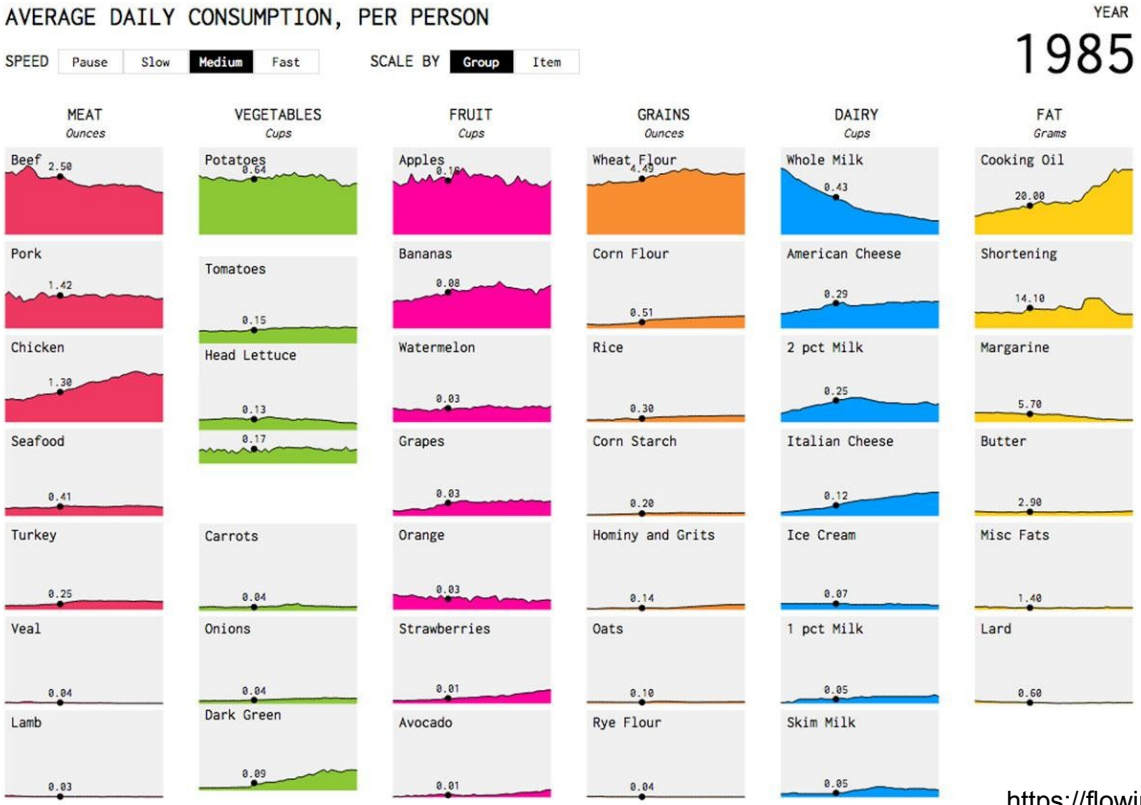


Scale does not always have to include zero

Optimize the ratio between plot objects to capture accurate relationships

Transform data to a different scale e.g. use log scale to show percentage change over time

6 Use text and labels to improve interpretation



Use meaningful titles

Label axis, as needed

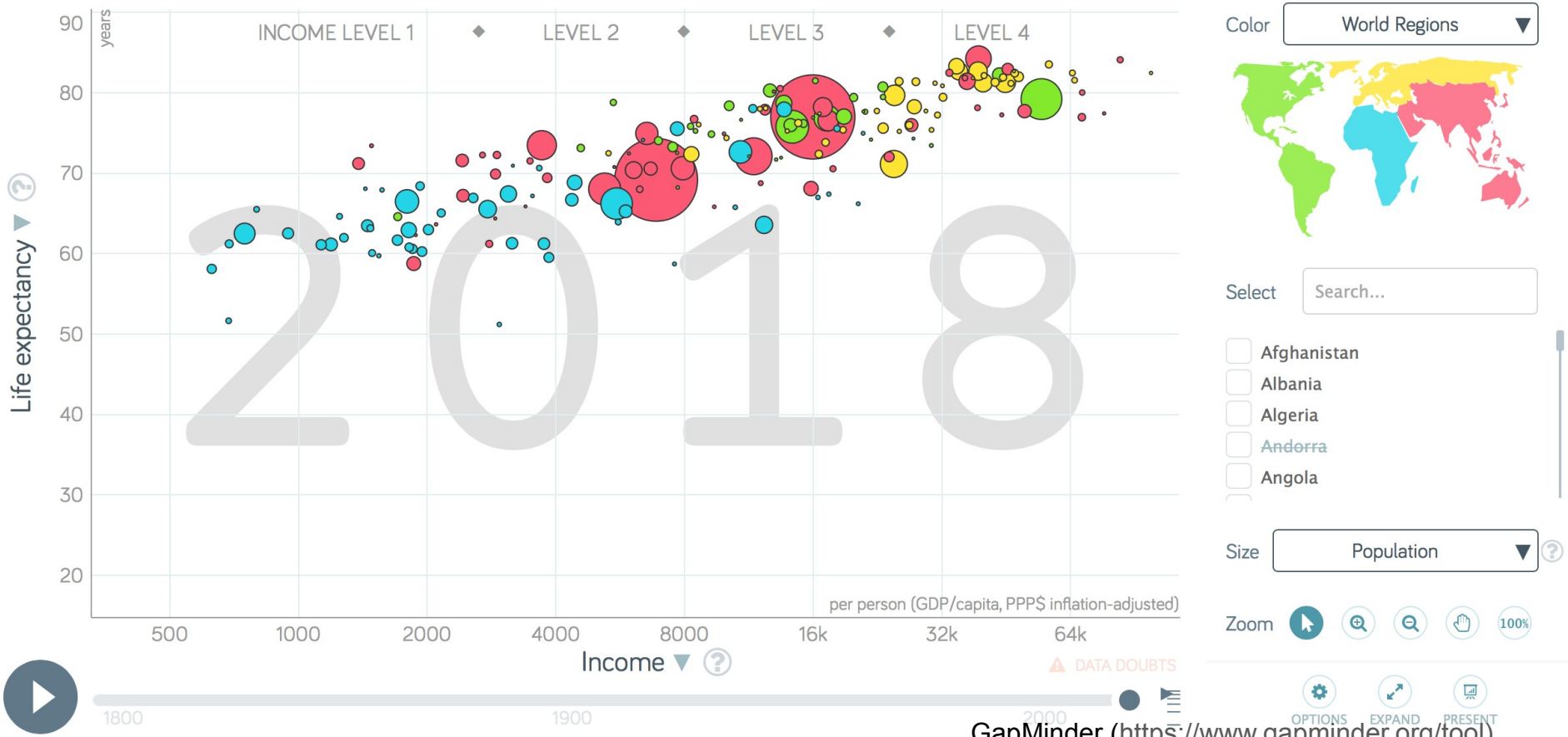
Add texts directly to the image - do not always rely on legends

Lines should not obstruct points

Use colors (e.g. light grey) and weight that lessen focus on tick marks and grids

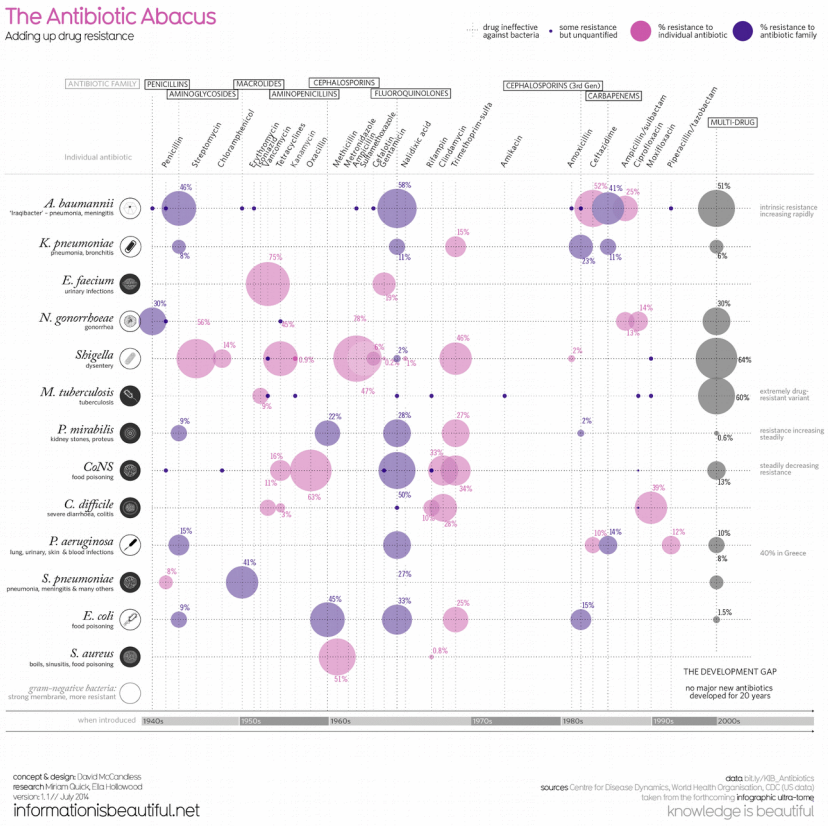
https://flowingdata.com/2016/05/17/the-changing-american-diet/

7 Balance complexity and clarity



GapMinder (https://www.gapminder.org/tool)

7 Balance complexity and clarity (infographics)



Templates and examples available online

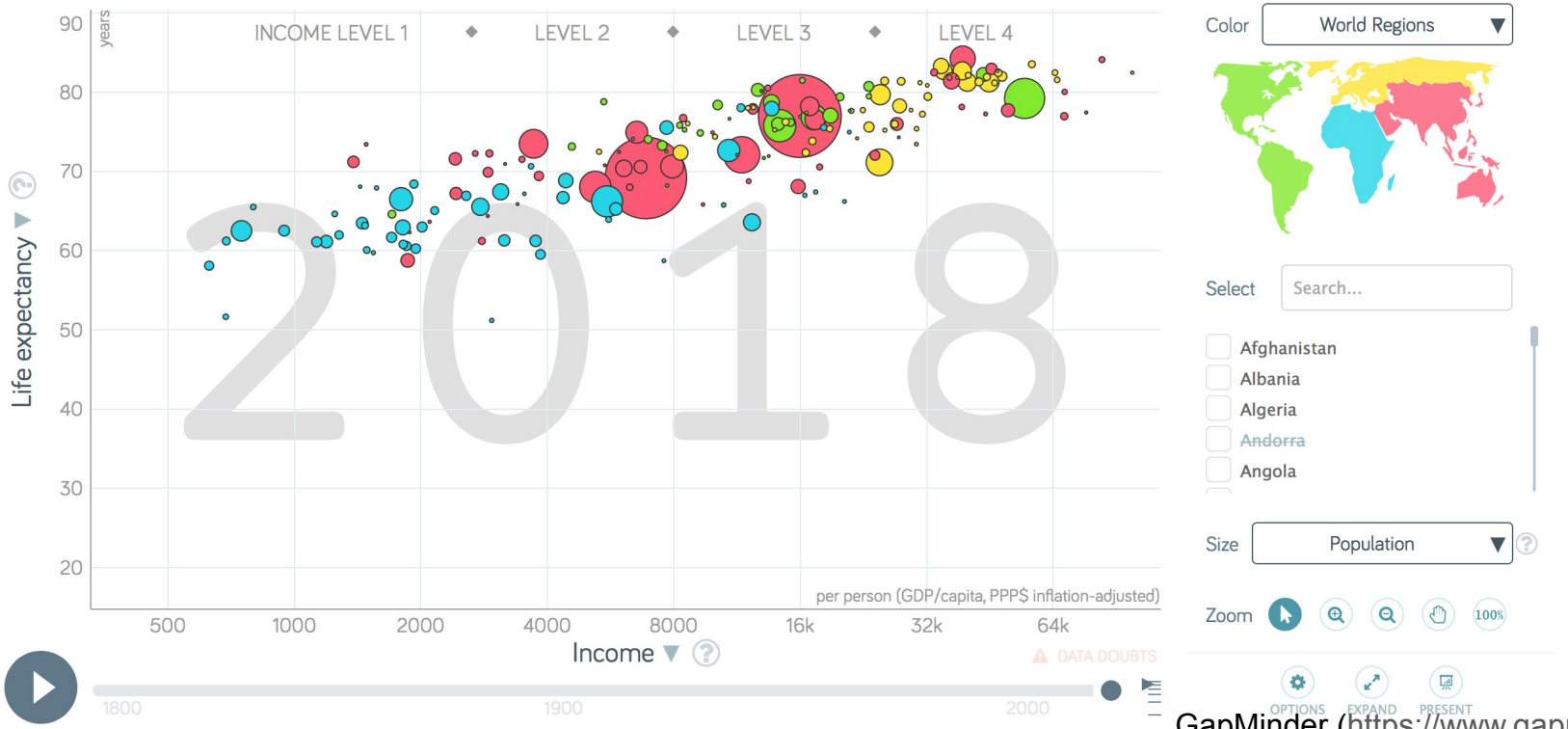
Can combine visualizations from python with manual editing

Examples

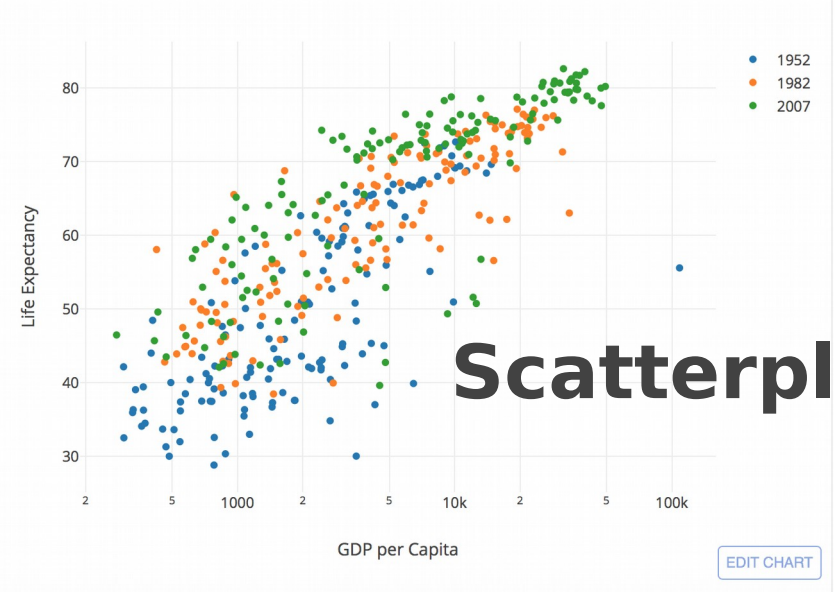
Bubbles

When to use?

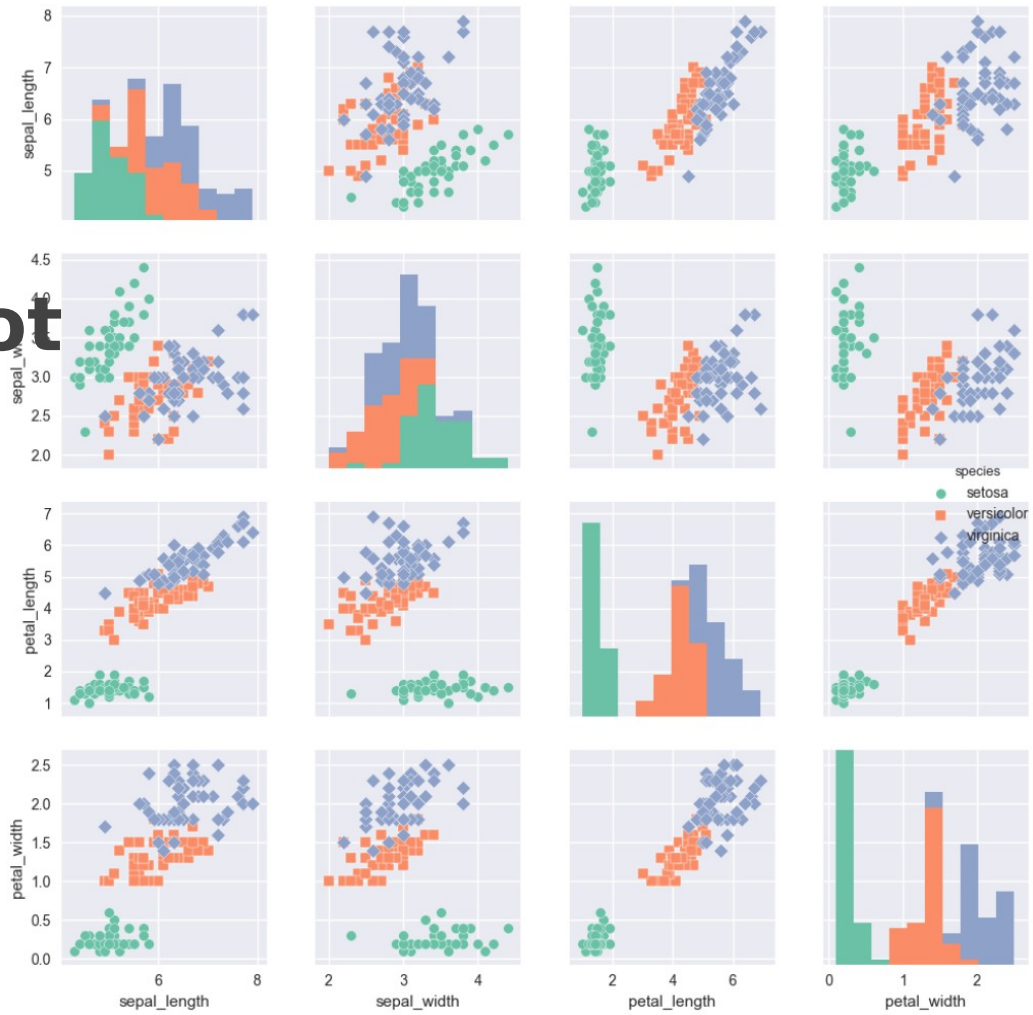
Visualize correlation/association



GapMinder (https://www.gapminder.org/tool)



Scatterplot



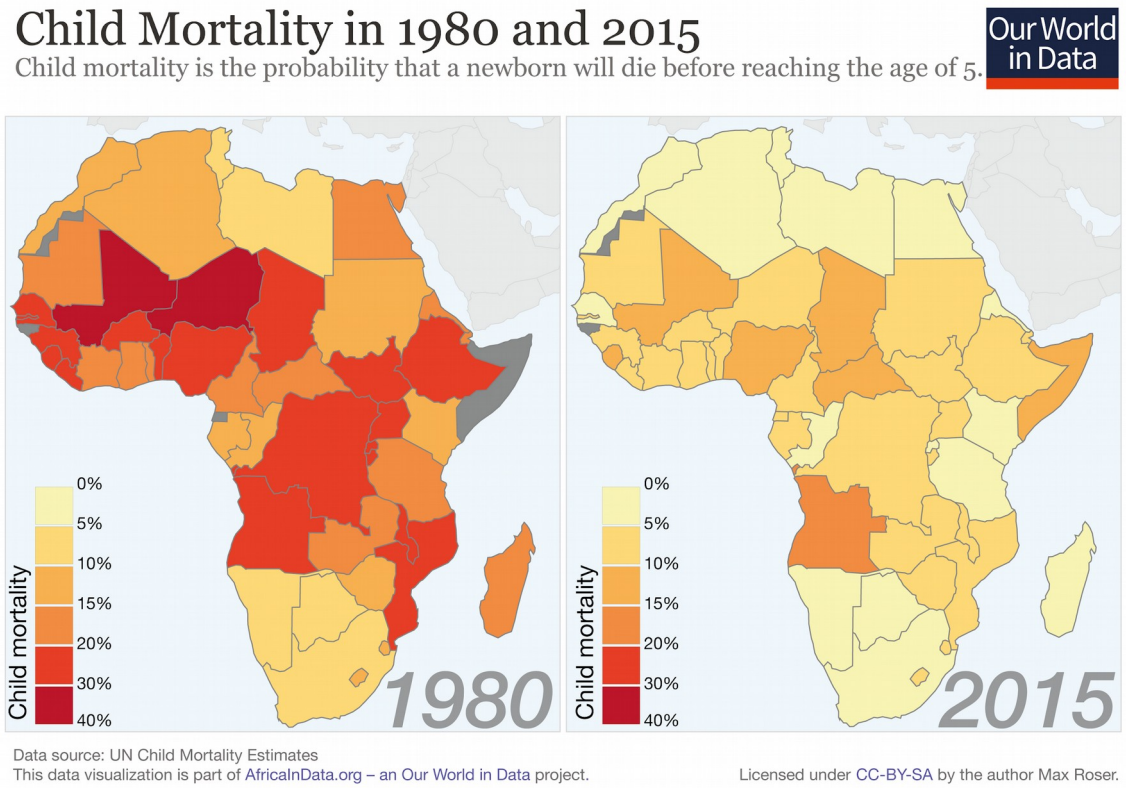
- Connected scatter
  Correlogram
  Heatmap

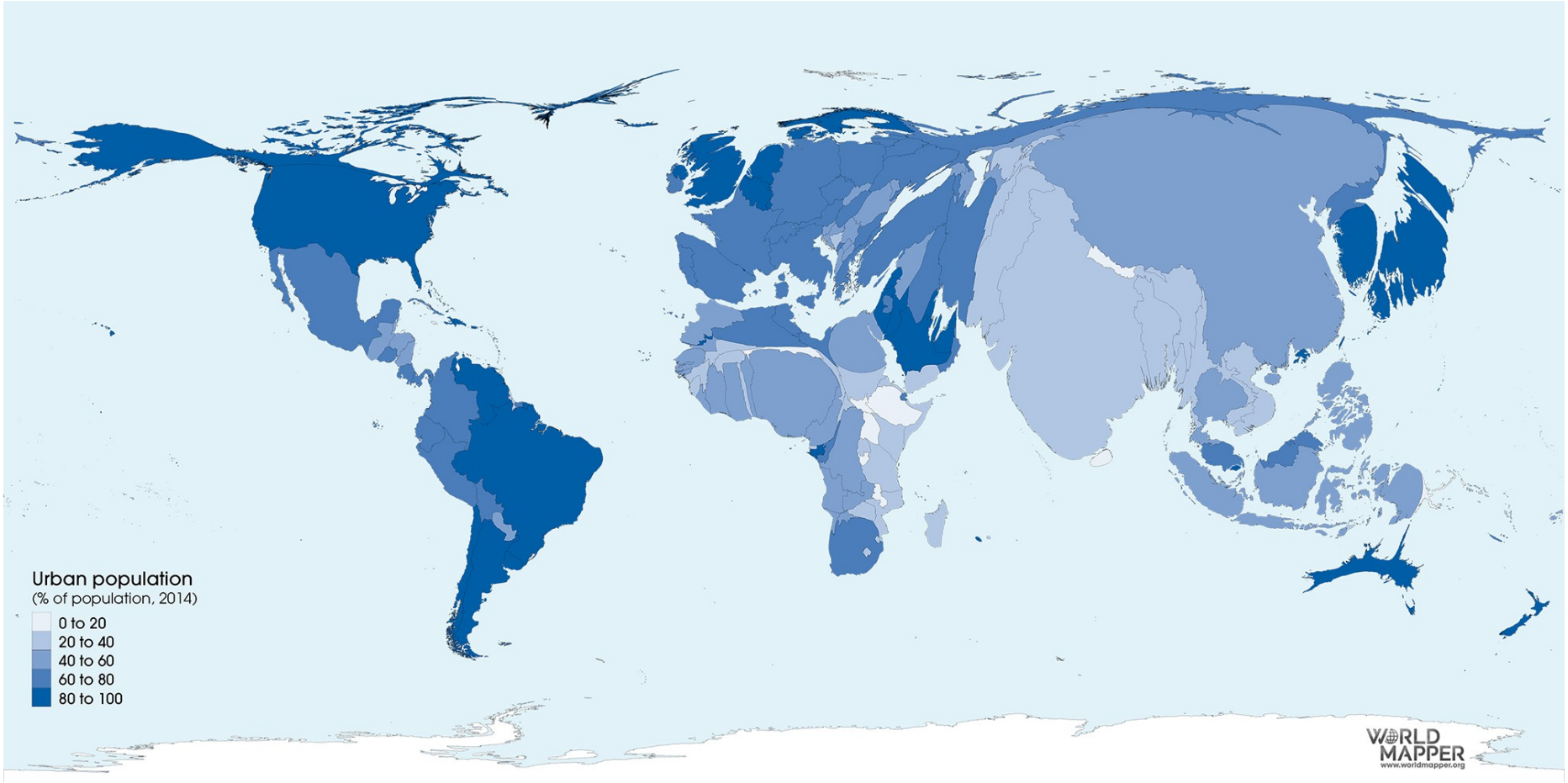
https://python-graph-gallery.com

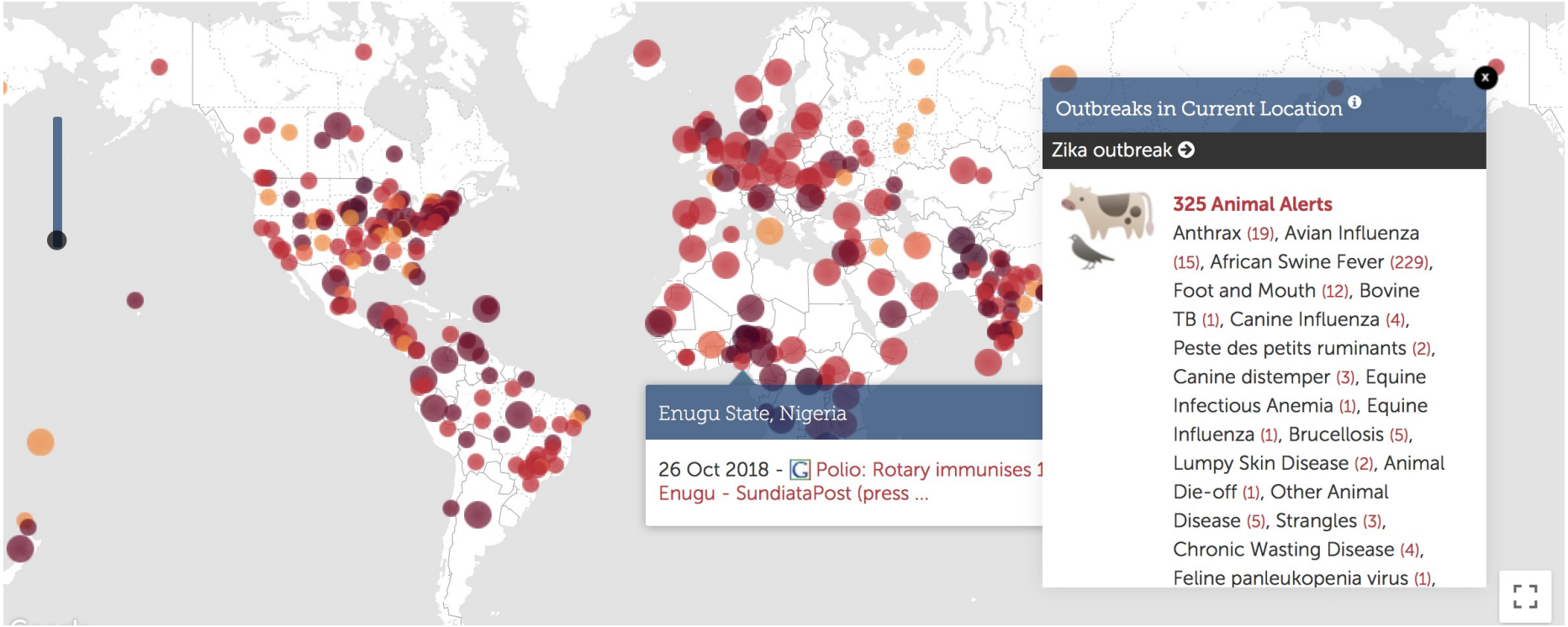
Maps

When to use?

Useful for spatial visualizations




Urban population
(% of population, 2014)
0 to 20
20 to 40
40 to 60
60 to 80
80 to 100
WORLD MAPPER
www.worldmapper.org



- Maps with bubbles
- Maps with pins

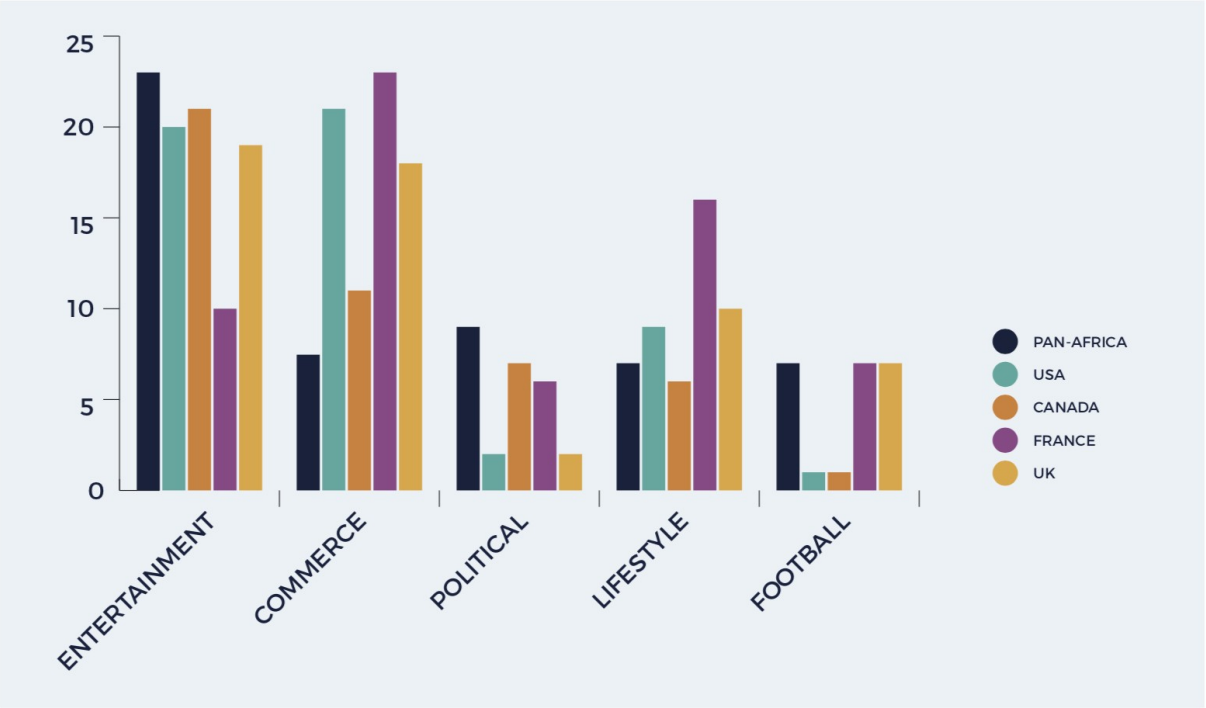
healthmap.org

When to use?

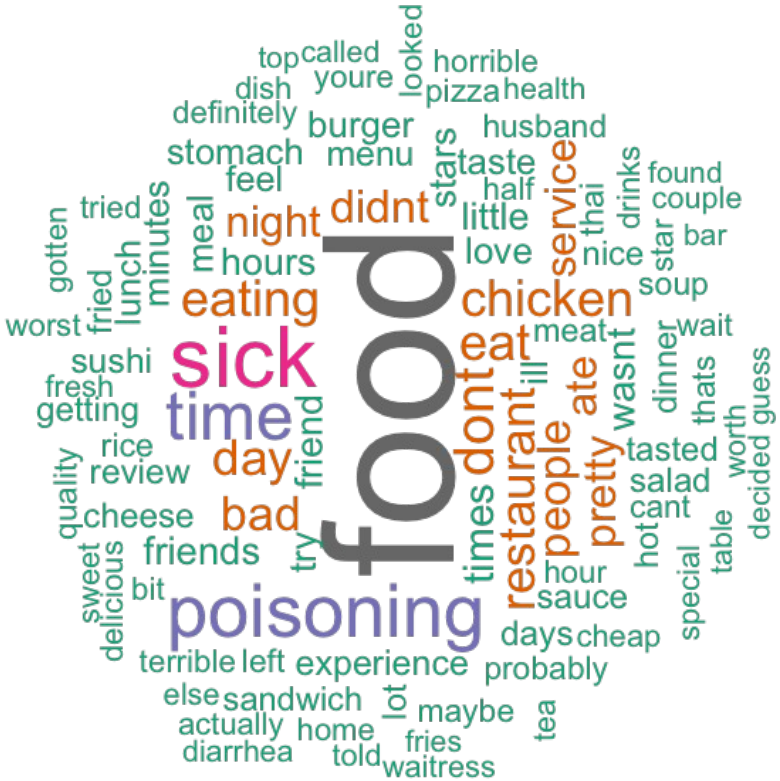
Useful for rankings

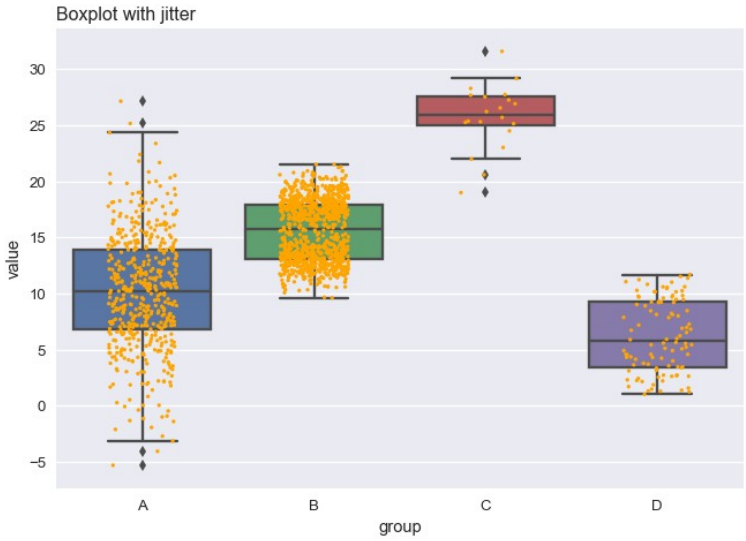
Bar plots

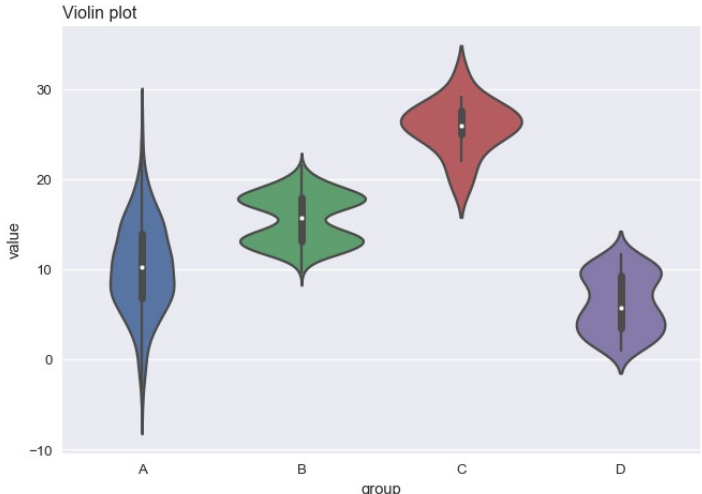
Top five themes of hashtags around the world



How Africa Tweets. https://portland-communications.com/publications/how-africa-tweets-2018/







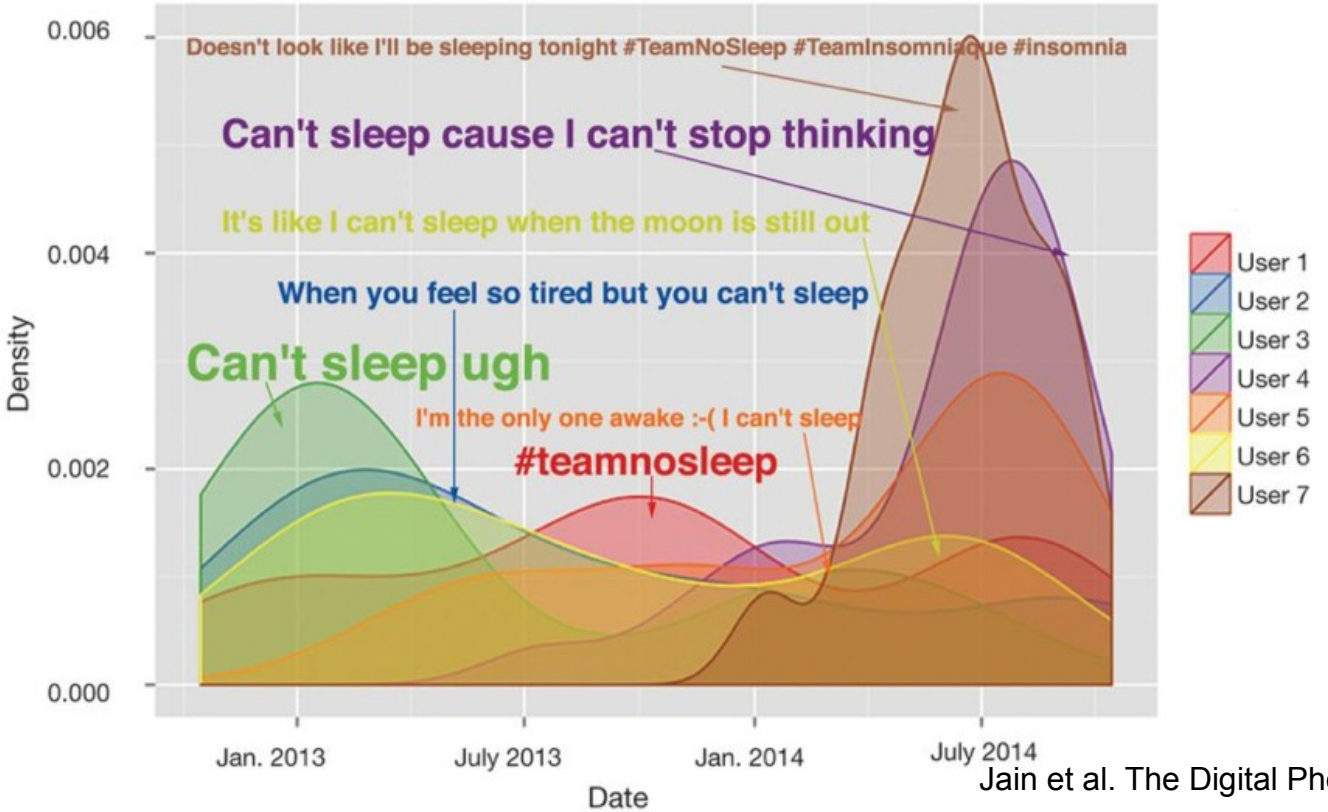
- Box plot
  Lollipop plot
  Word cloud

https://python-graph-gallery.com

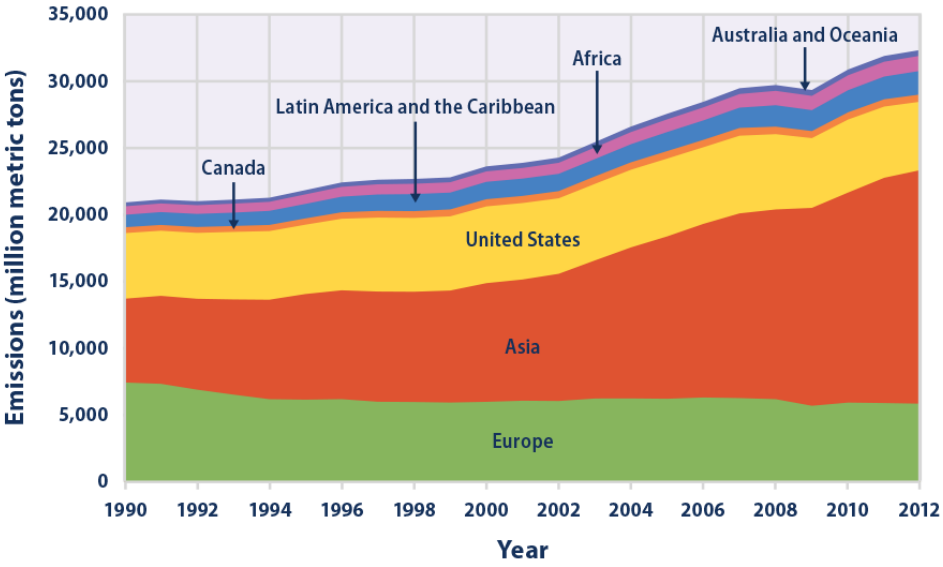
When to use?

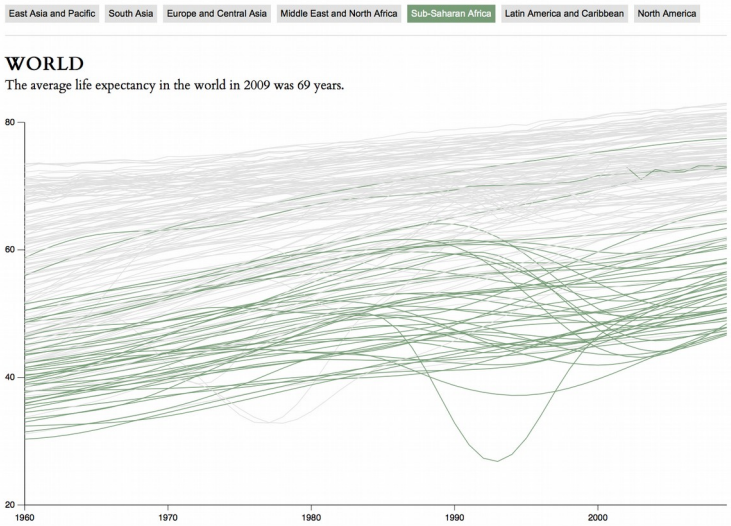
Useful for showing evolution

Area/density plots

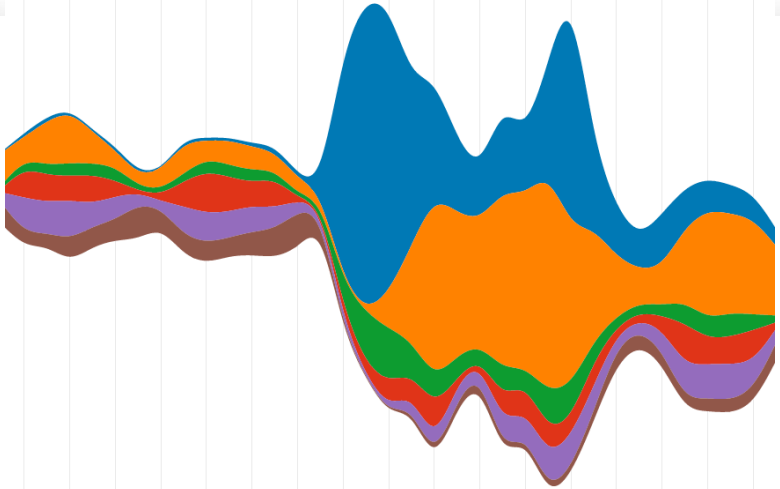


Jain et al. The Digital Phenotype. Nat Biotech





- Line plot
  (Stacked) area plot
  Stream chart



When to use?

Useful for information flow

Networks

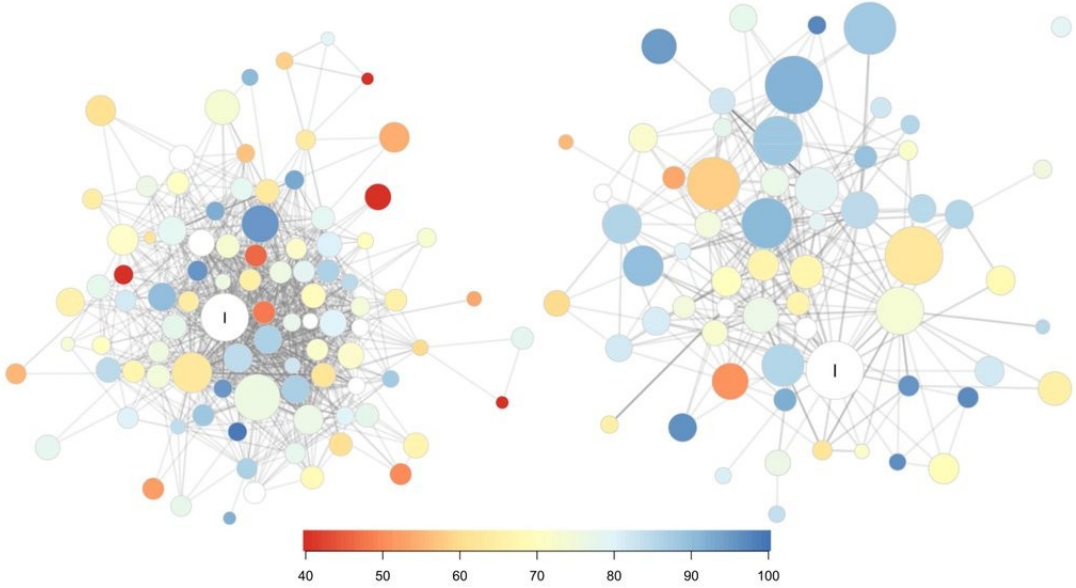
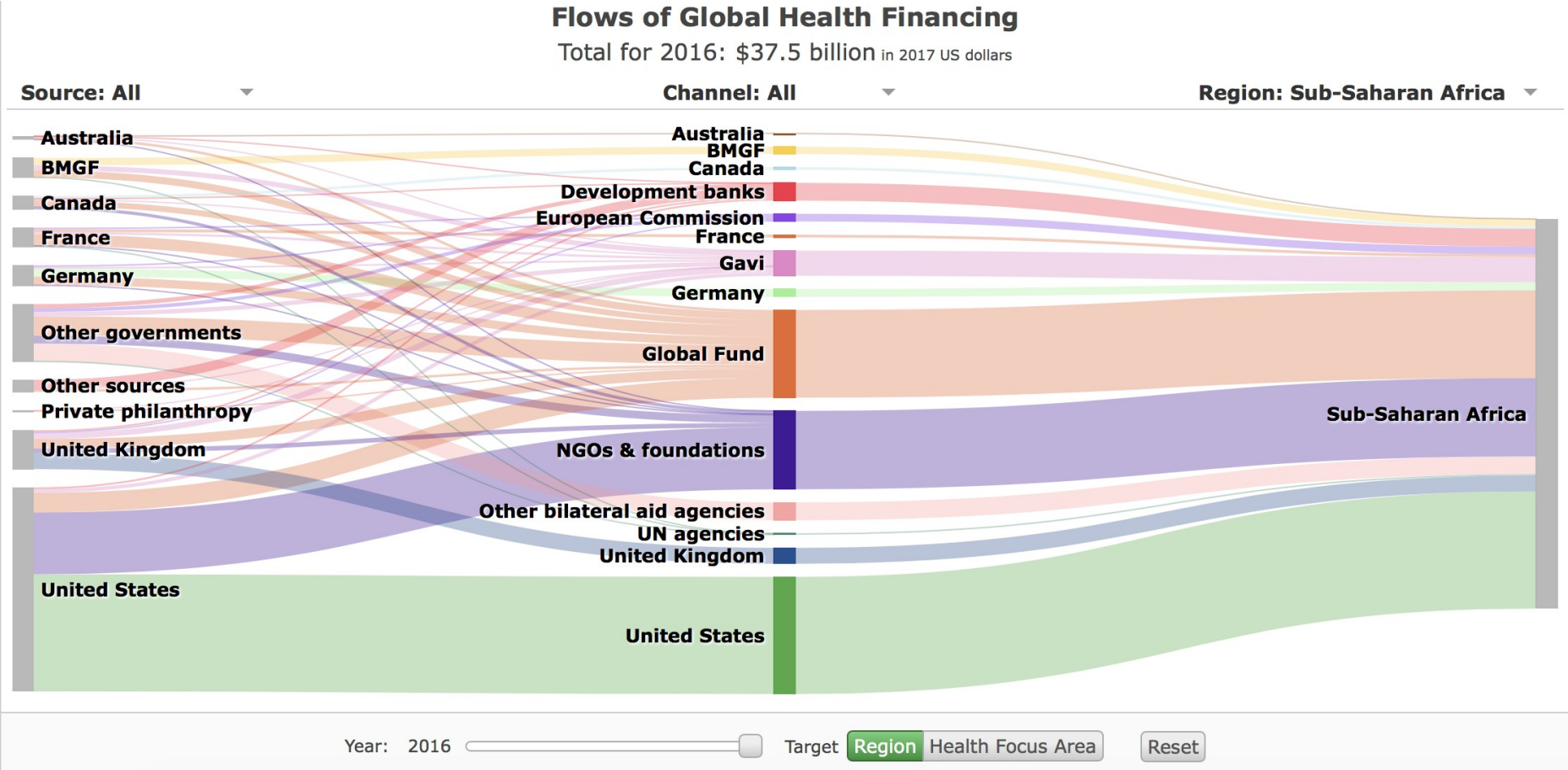


FIG. 4. Forum networks from weeks 7–8 in semester 1 (left) and semester 2 (right). Line opacity is scaled by edge weight, so darker lines indicate more threads in common for a student pair. Nodes are sized by total contributions over the semester and colored by grade (red low, yellow medium, blue high). Nodes without grades (withdrawals and instructor or CN staff accounts) are white, and the instructor's node is labeled "I."



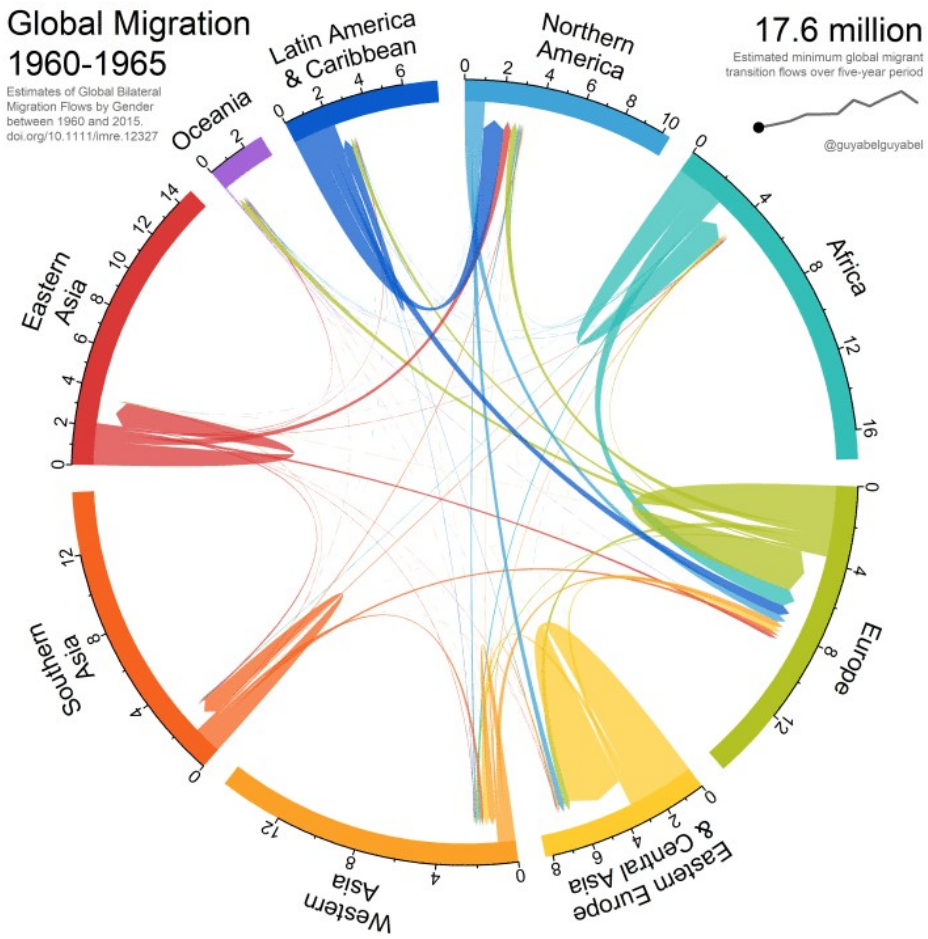
- Sankey diagram

https://vizhub.healthdata.org

Code available from:

https://guyabel.com/post/animated-directional-chord-diagrams/

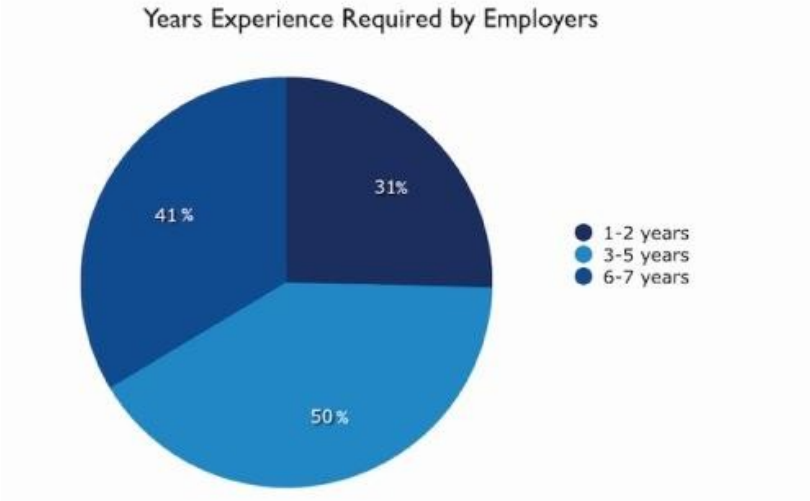
Chord diagram



Bad visualizations

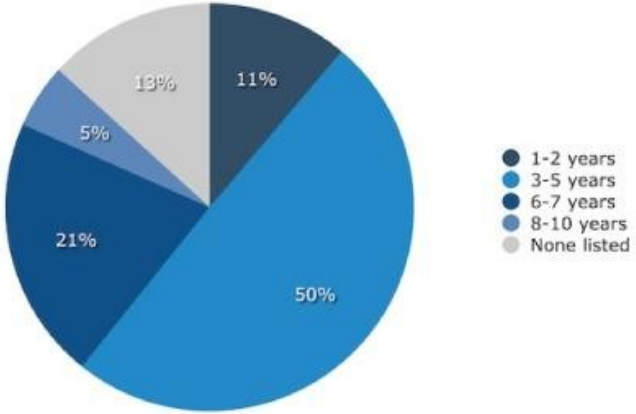
Which of these images has issues?

A

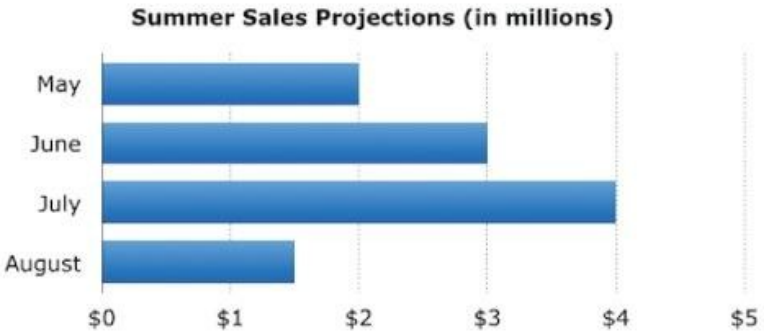


C

Which of these images has issues?

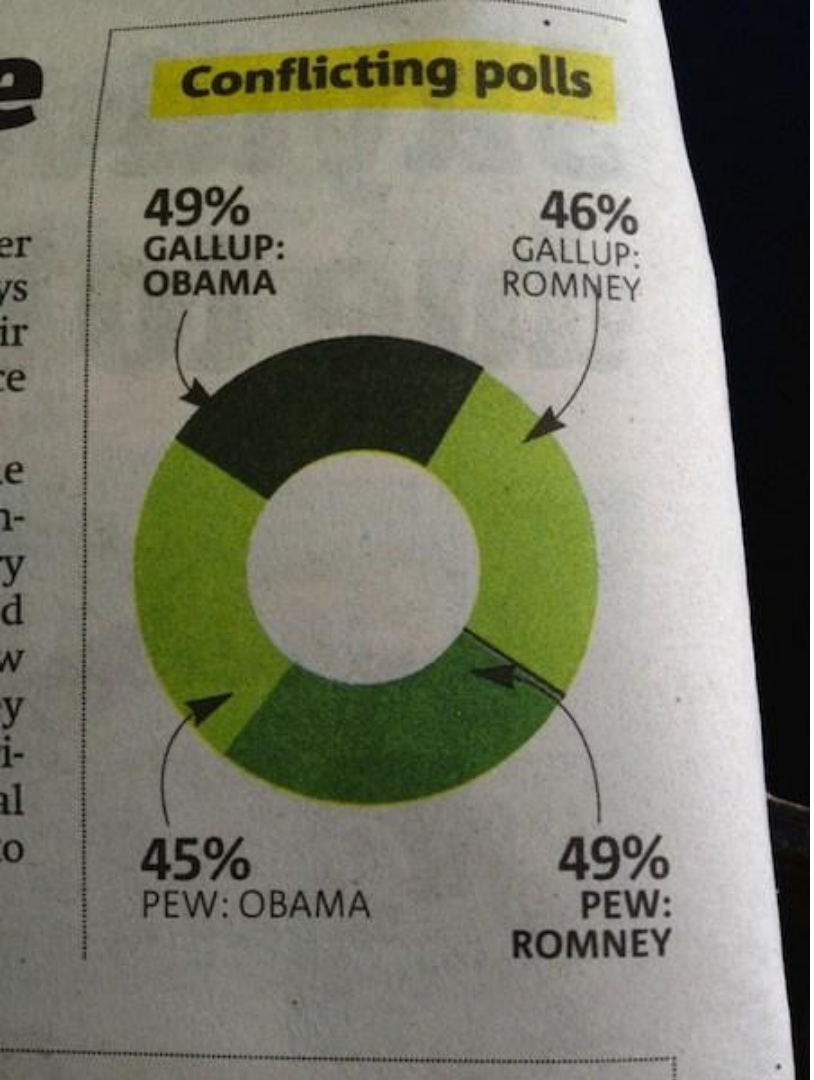


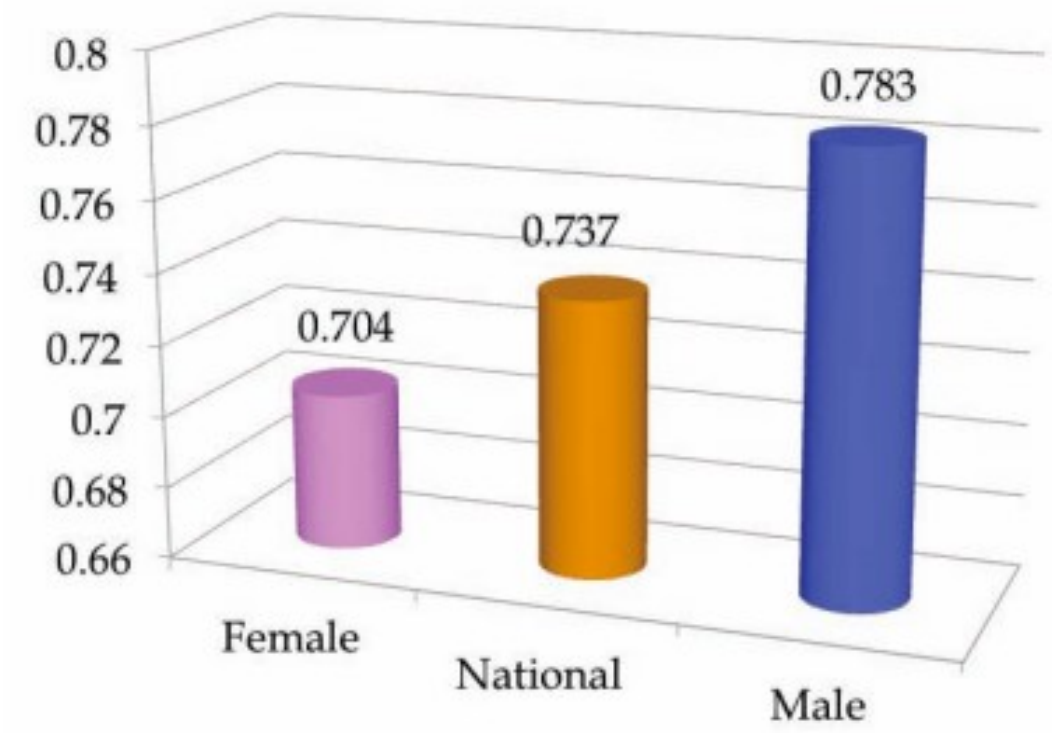
B



D

What’s wrong with these images?

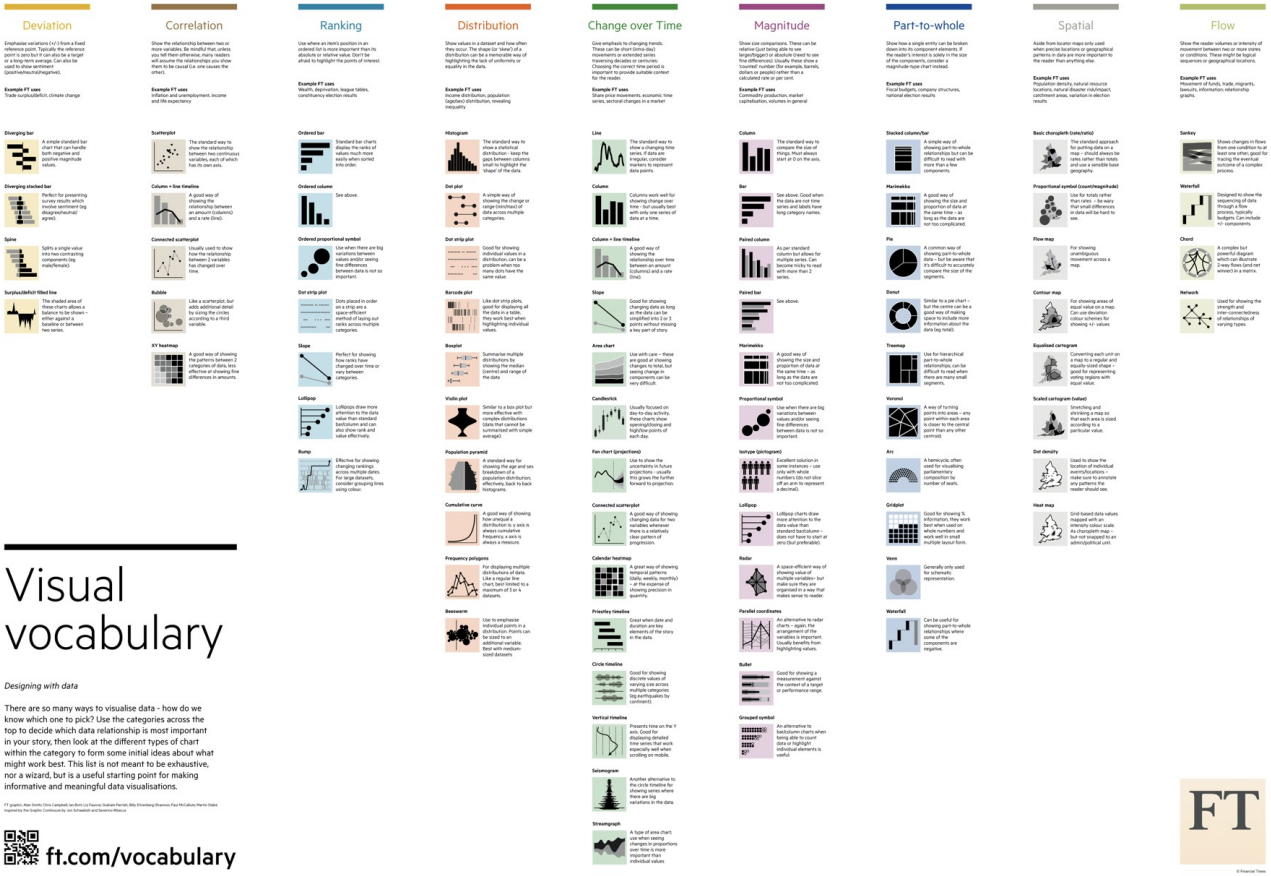




Tools and Resources

Python libraries

- Matplotlib
- ggplot
- Seaborn
- Bokeh
- Pygal
- Plotly
- Geoplotlib
- Gleam
- Missingno
- Leather
- Pydot



Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Ian Bott; Liz Faunce; Graham Parrish; Billy Ehrenberg-Shannon; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

The Chart Doctor

https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary

Other tools

- Tableau
- R ggplot2 and others
- D3

Next … ipython tutorial