

# Introduction to Data Visualization

Morine Amutorine  
Benjamin Akera  
Elaine Nsoesie





# Instructor introductions ...

Contacts:

Morine - [morine.amutorine@one.un.org](mailto:morine.amutorine@one.un.org) / Twitter:  
@M\_moryn

Elaine - [onelaine@bu.edu](mailto:onelaine@bu.edu) / Twitter: @ensoesie

Ben - [akeraben@gmail.com](mailto:akeraben@gmail.com)



# Resources

## **Github Repo**

[https://github.com/ensoesie/DSA\\_Visualization](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

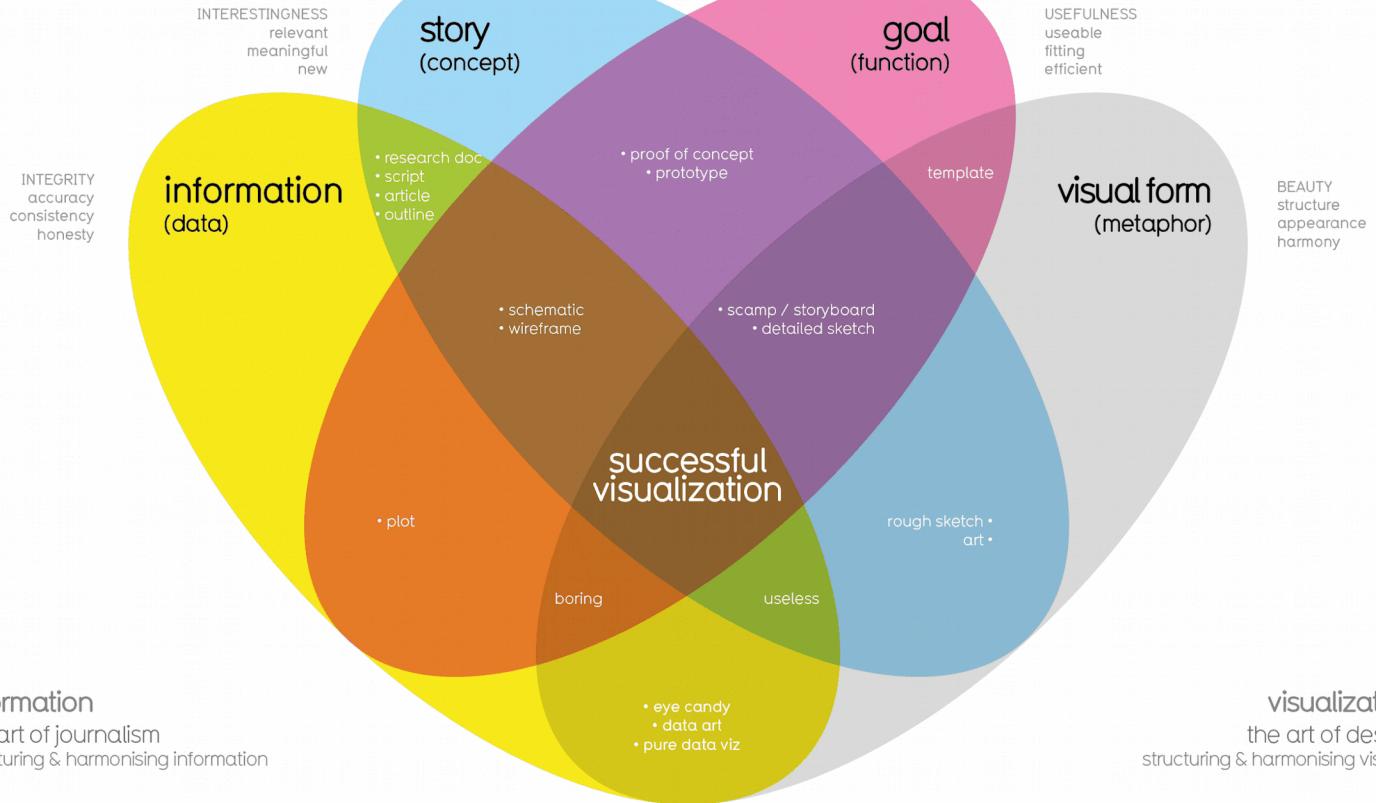
It is easier to remember pictures than text

Useful for understanding data

Can summarize large amounts of complex data

## What Makes a Good Visualization?

explicit (implicit)





# **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

# Use color and size to highlight and suppress information

East Asia and Pacific

South Asia

Europe and Central Asia

Middle East and North Africa

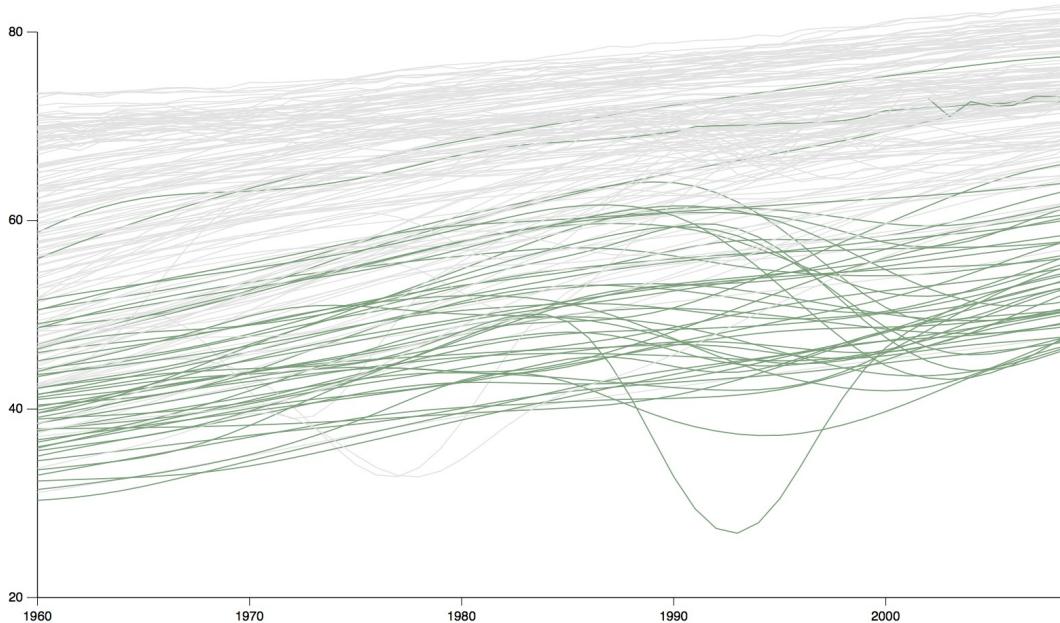
Sub-Saharan Africa

Latin America and Caribbean

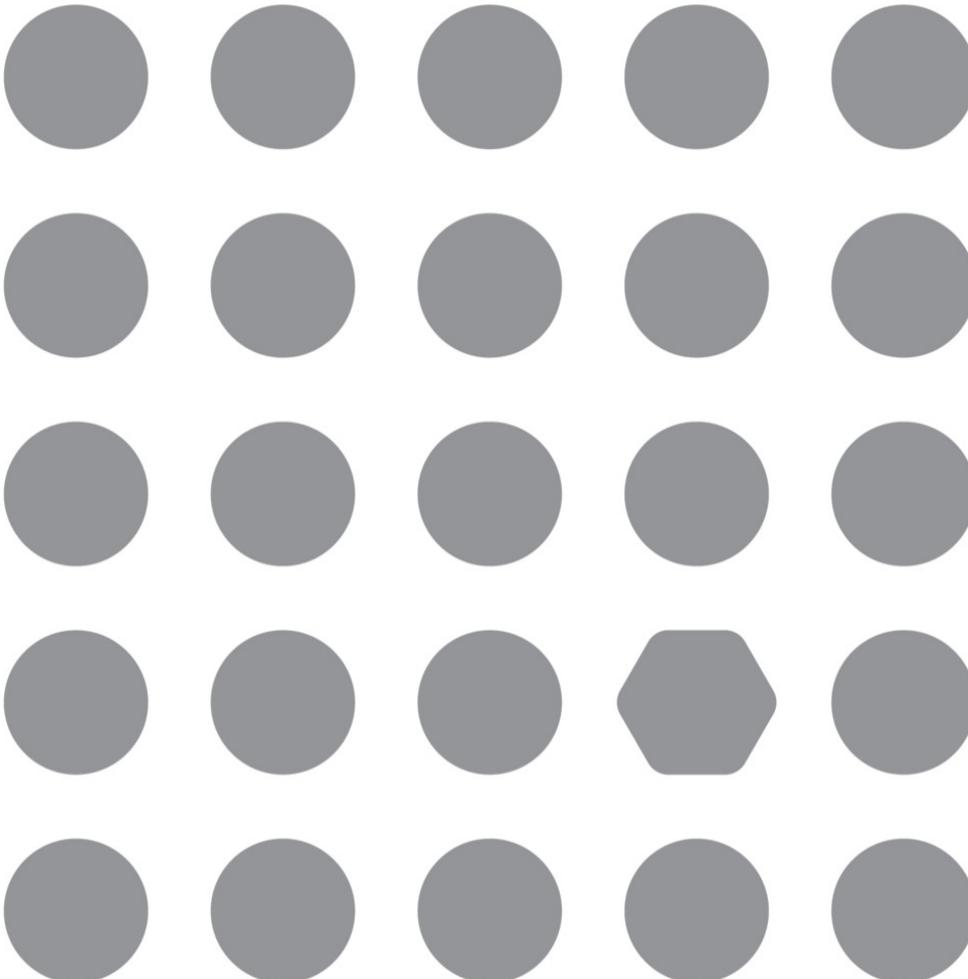
North America

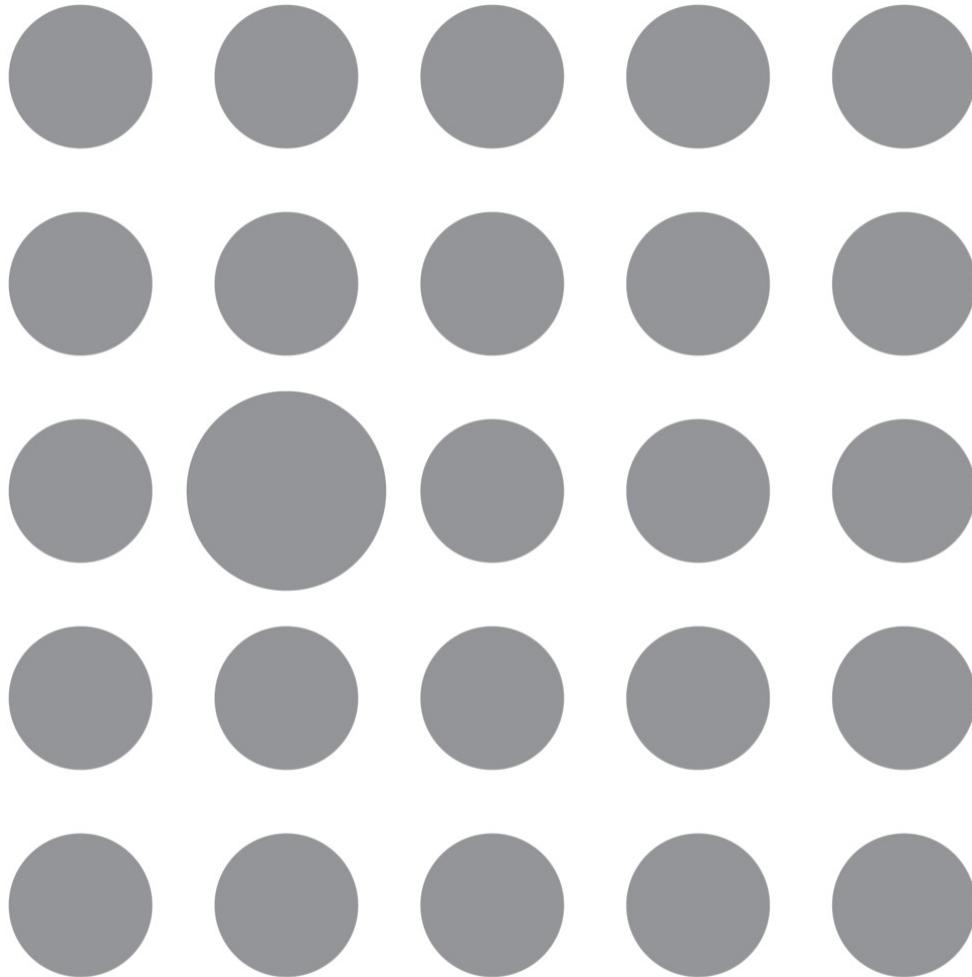
## WORLD

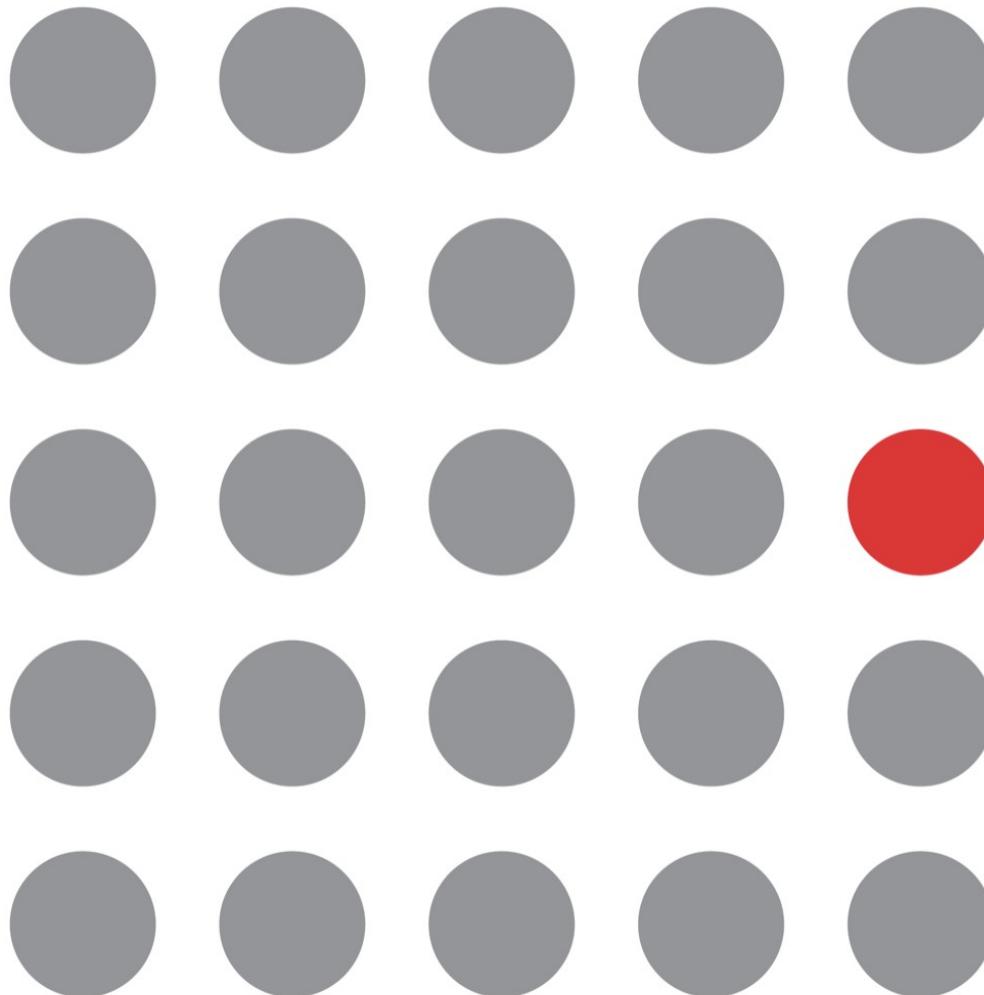
The average life expectancy in the world in 2009 was 69 years.



Source: The World Bank; Graphic by: Nathan Yau



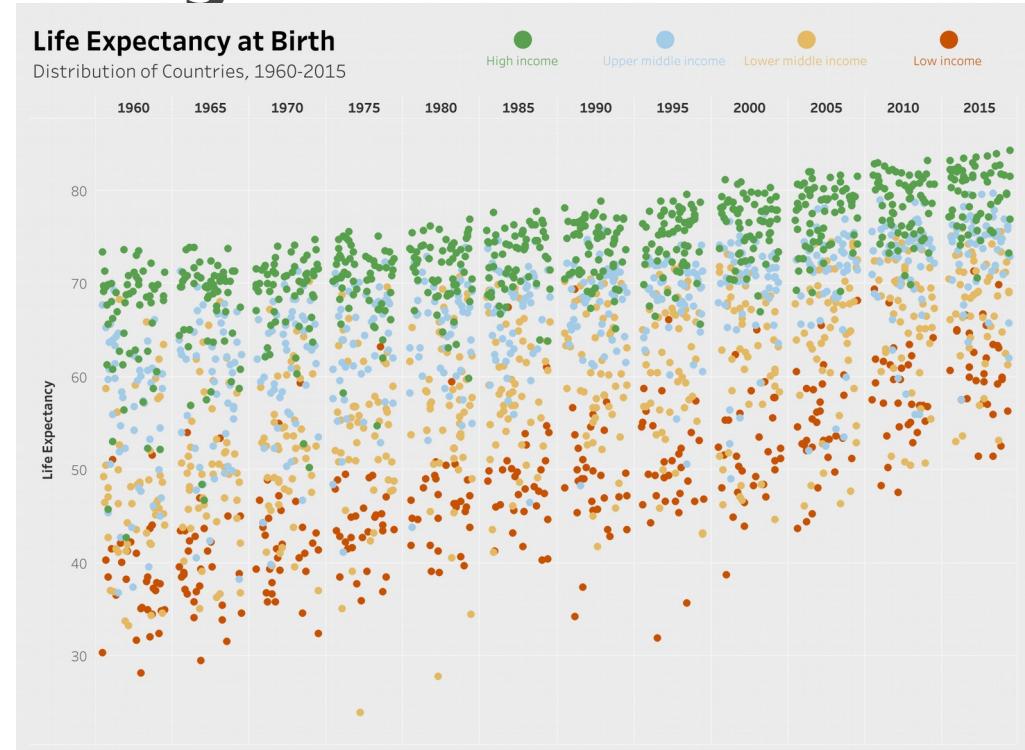




3

# 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**



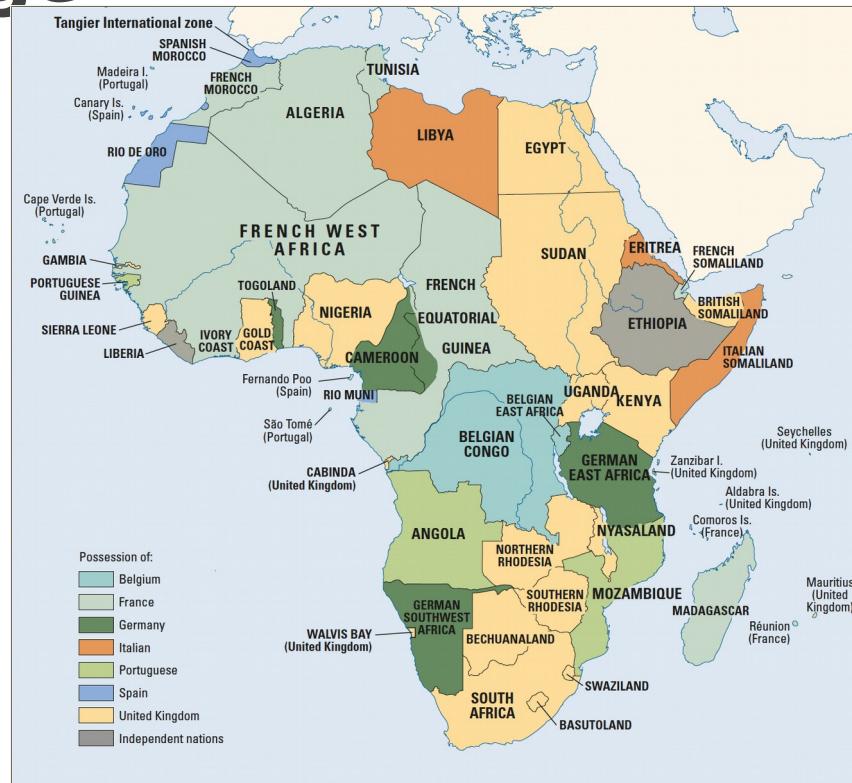
## 4

# 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



4

# Think carefully about color selection and usage

Some colors have pre-established meanings

Consider those with color blindness

**Red**

Stop

Dangerous

Hot

**Green**

Moving

Money

Plants

**Blue**

Water

Cool

Safe

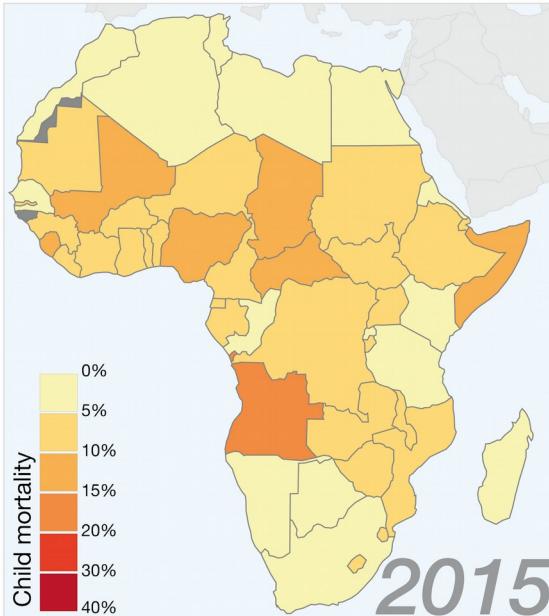
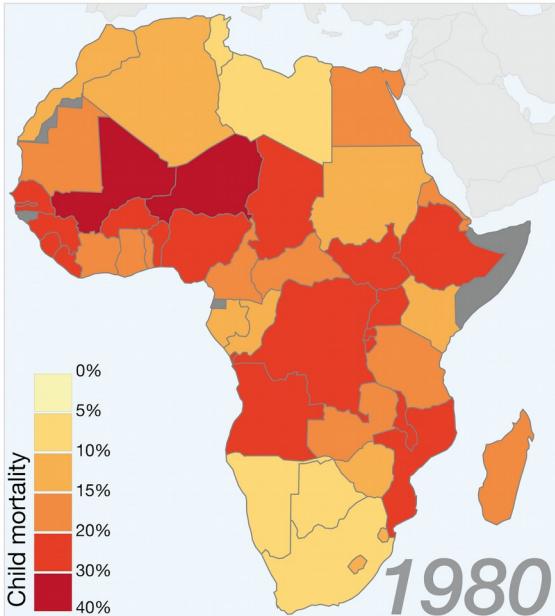
## 5

# Use all available space and proper scales

## Child Mortality in 1980 and 2015

Child mortality is the probability that a newborn will die before reaching the age of 5.

Our World  
in Data



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

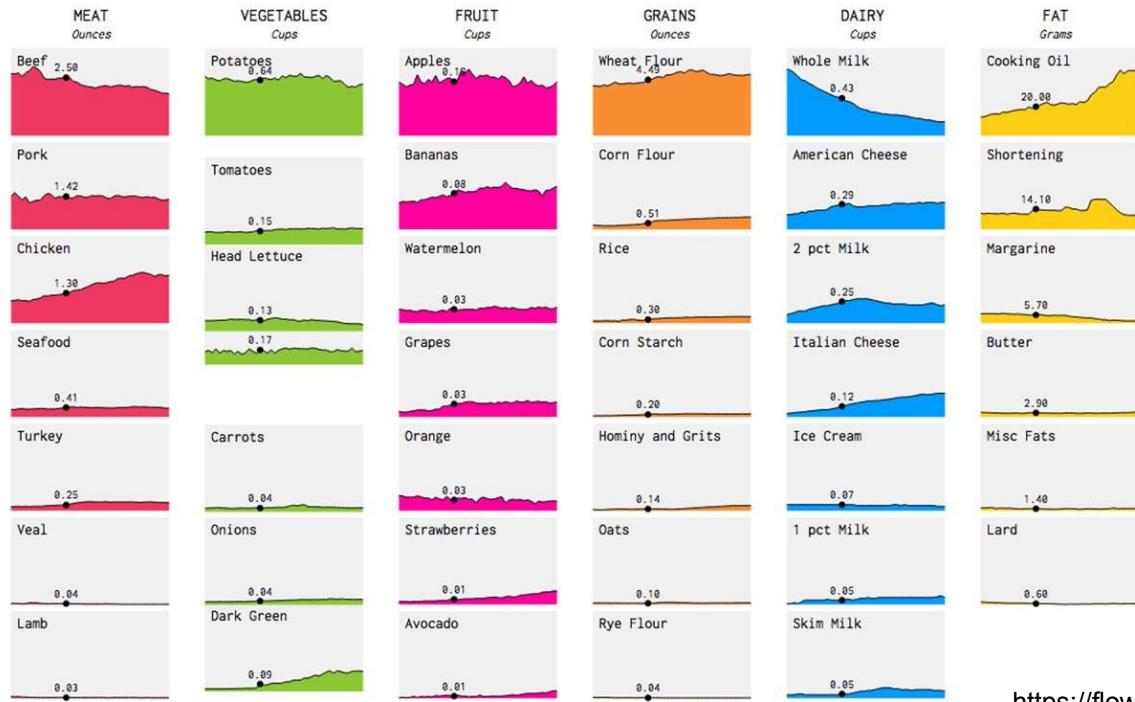
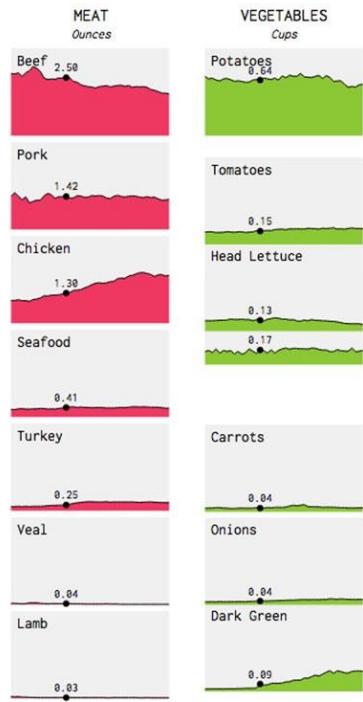
## 6

# Use text and labels to improve interpretation

AVERAGE DAILY CONSUMPTION, PER PERSON

SPEED Pause Slow Medium Fast

SCALE BY Group Item



YEAR  
**1985**

Use meaningful titles

Label axis, as needed

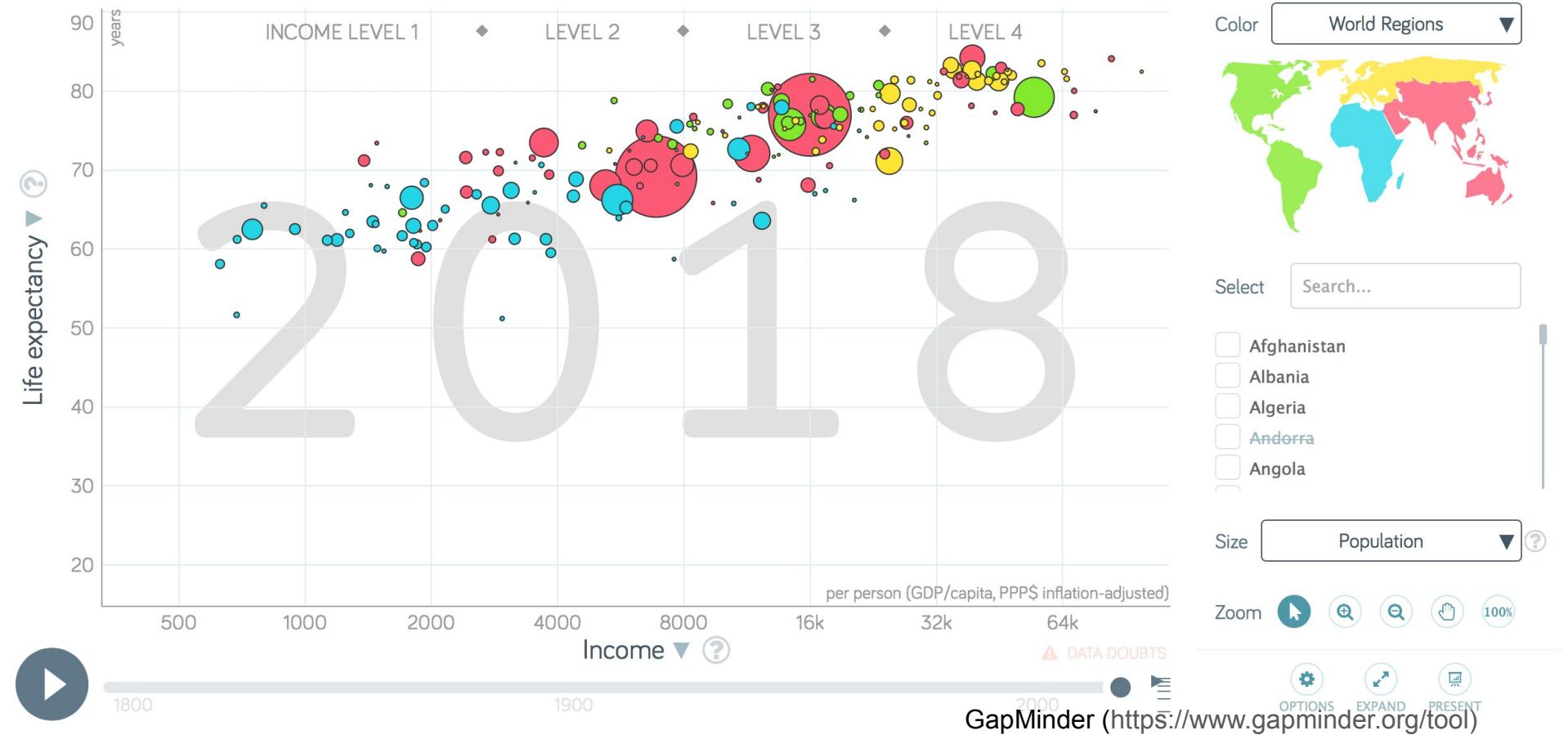
Add texts directly to the image - do not always rely on legends

Lines should not obstruct points

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

7

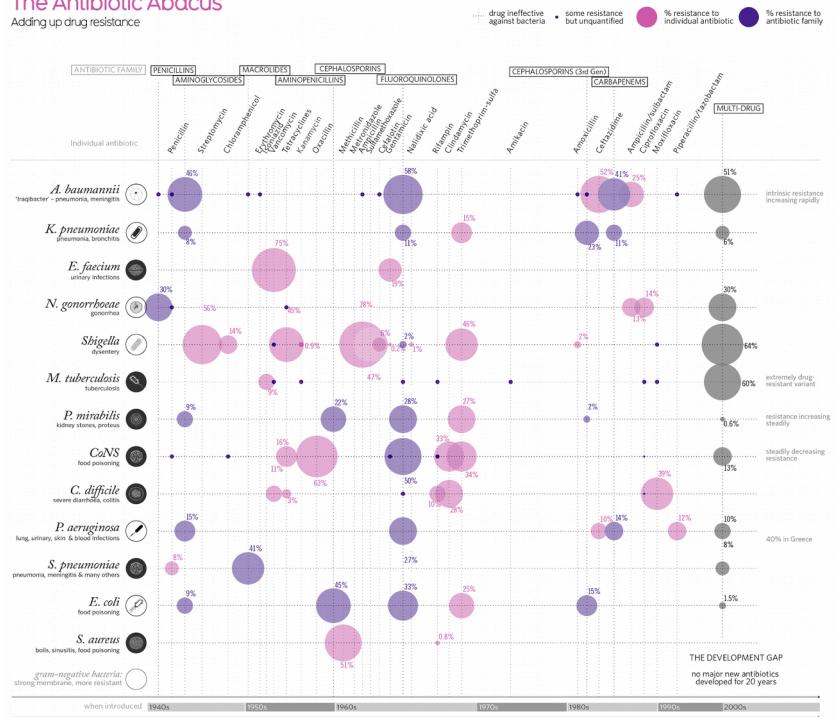
# **Balance complexity and clarity**



# Balance complexity and clarity (infographics)

## The Antibiotic Abacus

Adding up drug resistance



Templates and examples available online

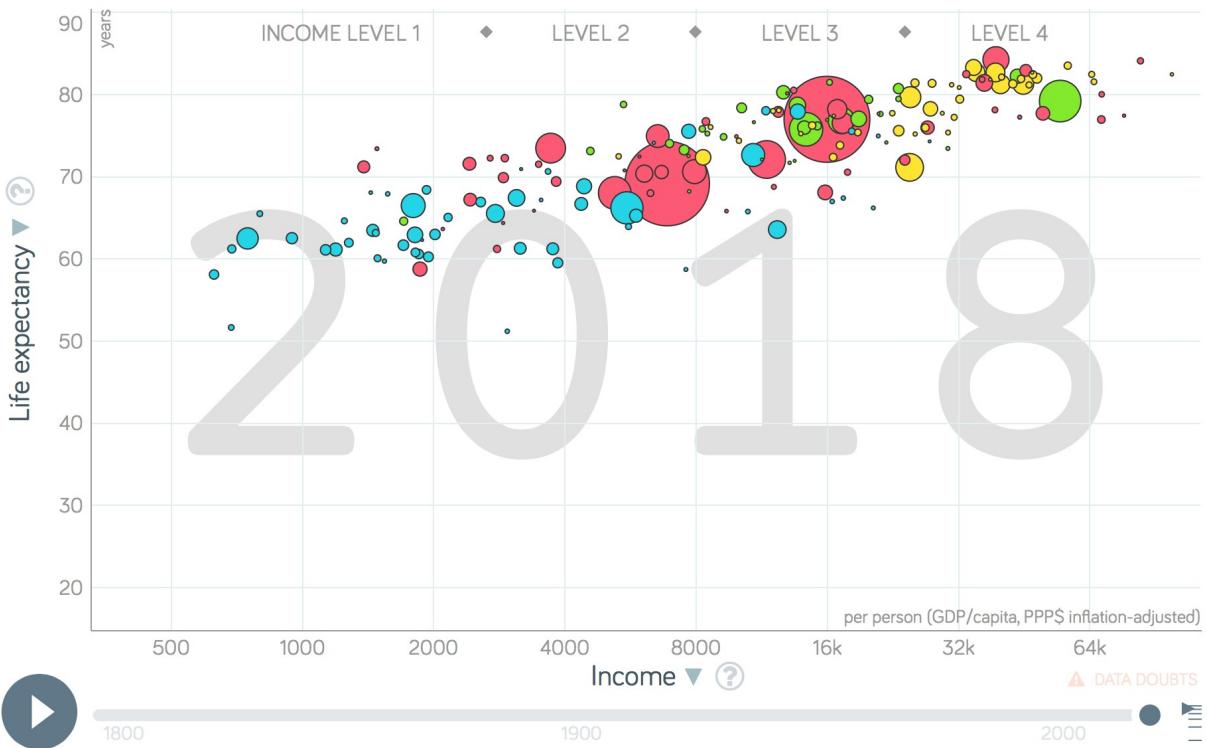
Can combine visualizations from python with manual editing

# Examples

# When to use?

Visualize  
correlation/association

## Bubbles



Color World Regions ▾



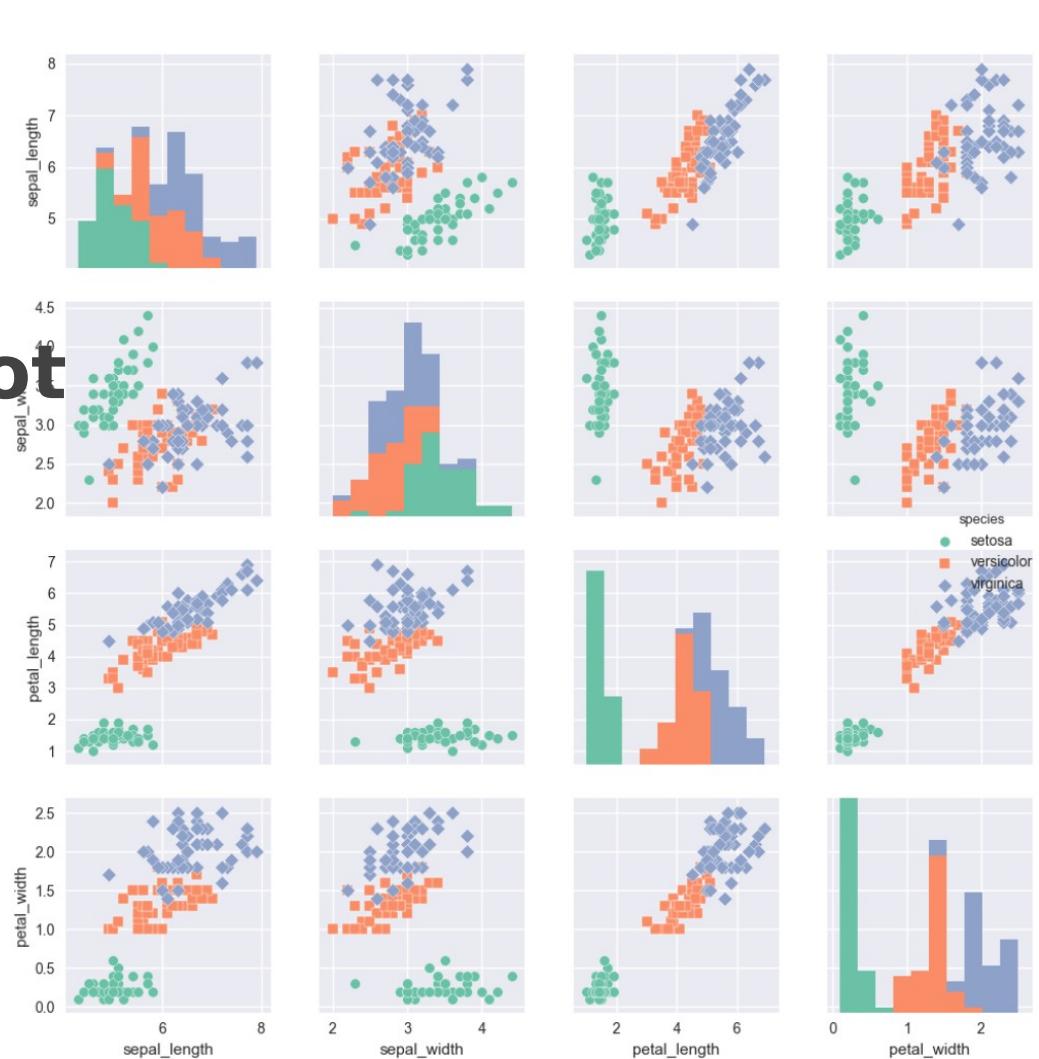
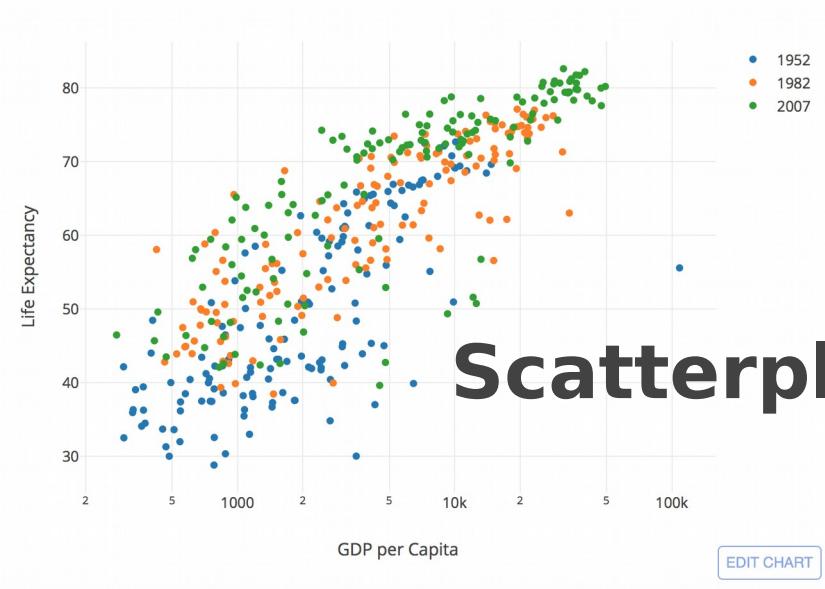
Select Search...

- Afghanistan
- Albania
- Algeria
- Andorra
- Angola

Size Population ▾

Zoom 100%

OPTIONS EXPAND PRESENT



- Connected scatter  
Correlogram  
Heatmap

# When to use?

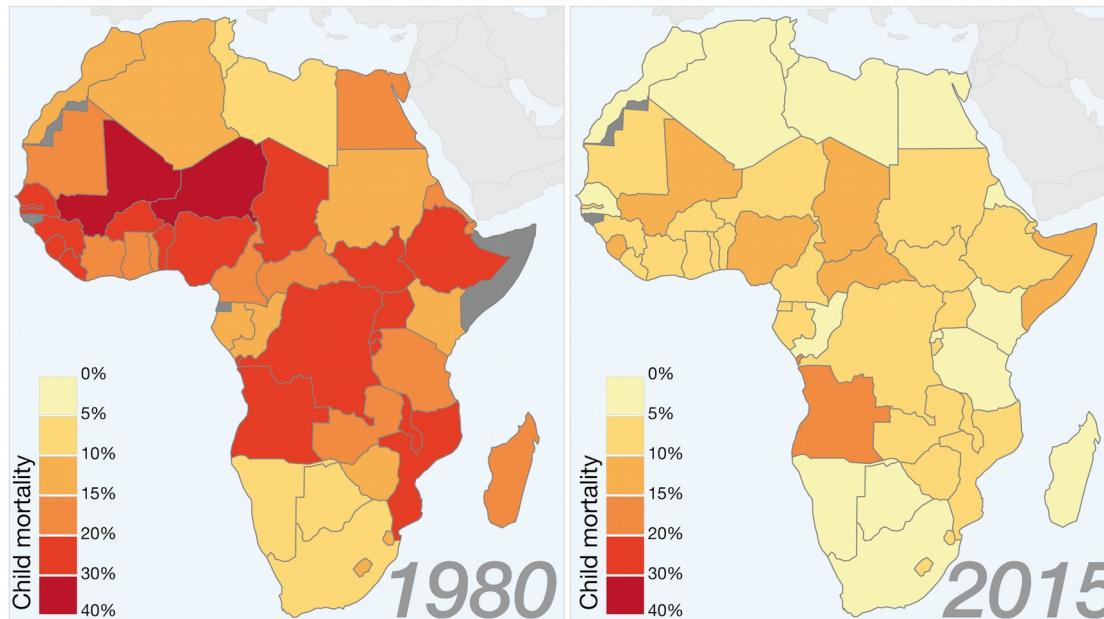
Useful for spatial  
visualizations

## Maps

### Child Mortality in 1980 and 2015

Child mortality is the probability that a newborn will die before reaching the age of 5.

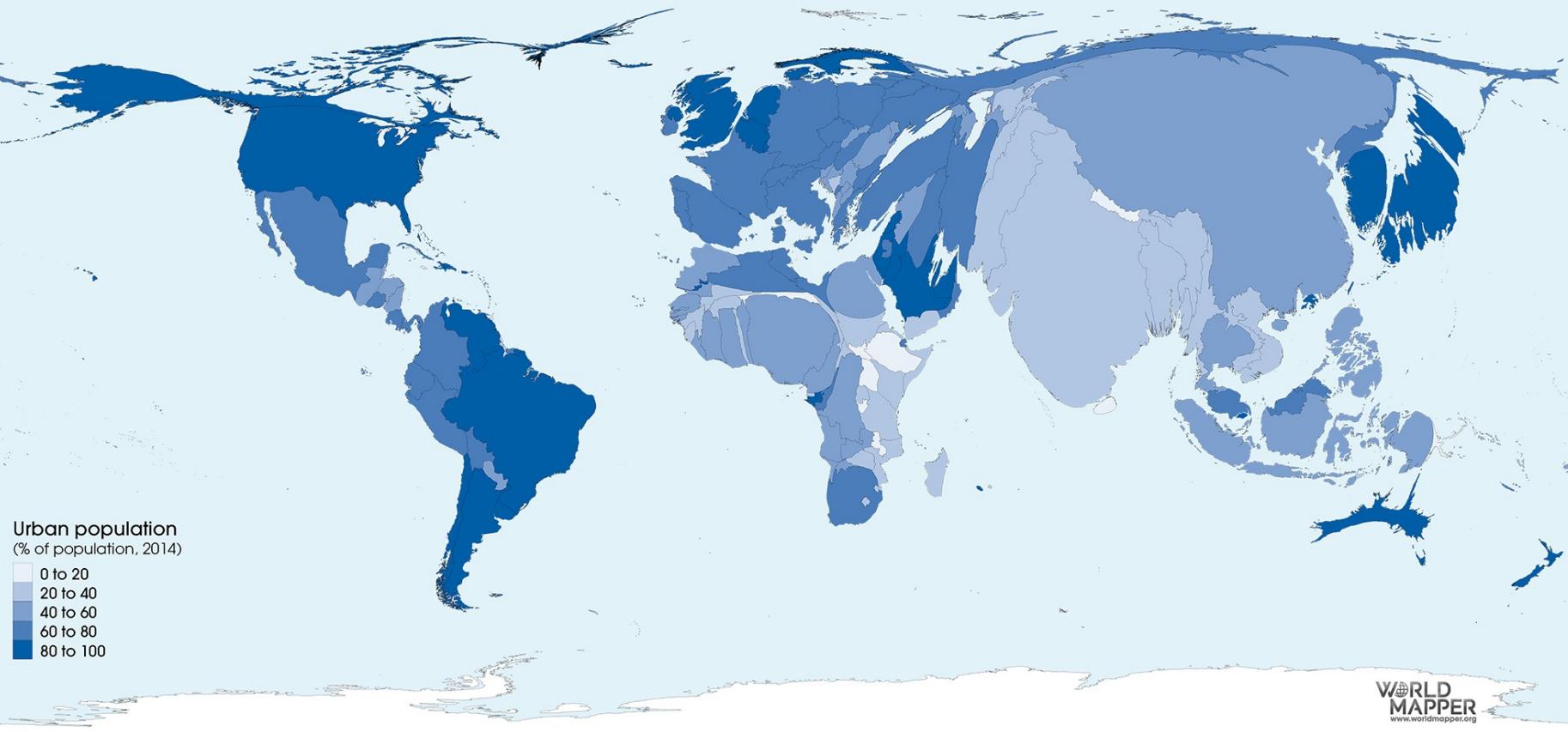
Our World  
in Data

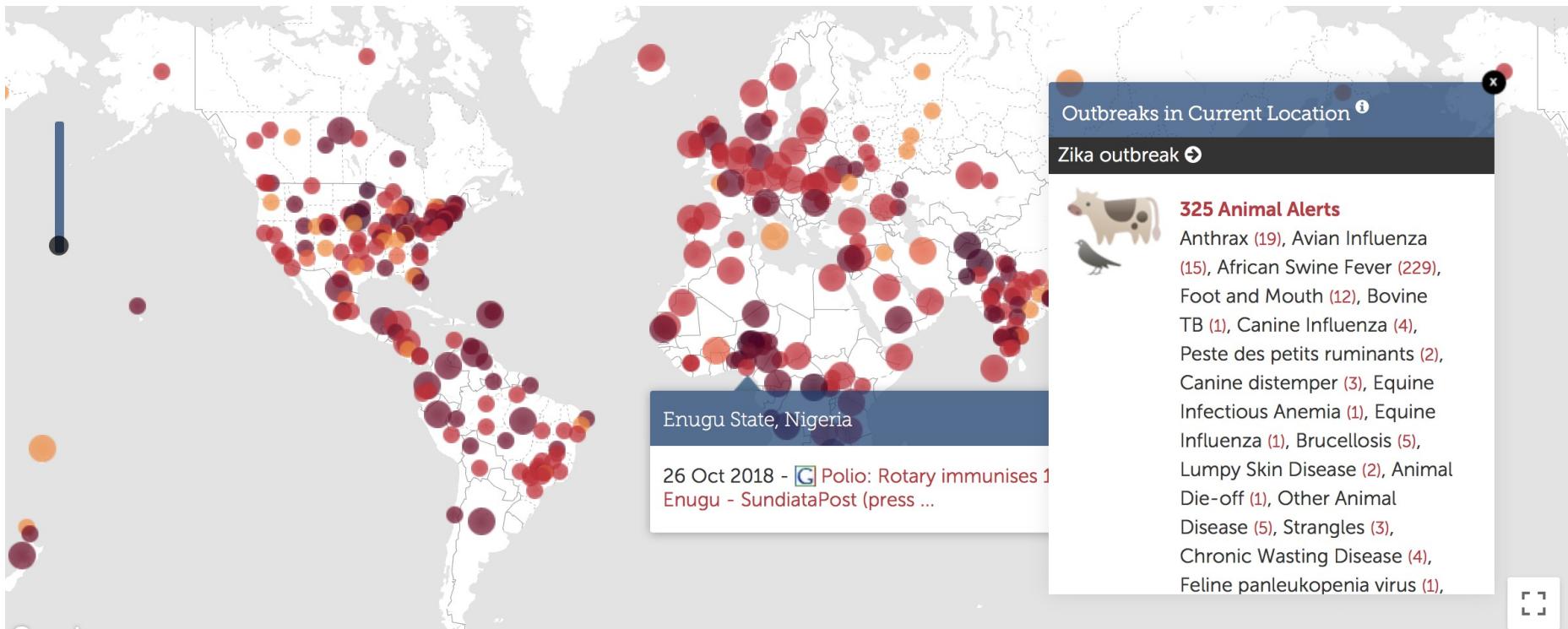


Data source: UN Child Mortality Estimates

This data visualization is part of [AfricanInData.org](#) – an Our World in Data project.

Licensed under CC-BY-SA by the author Max Roser.





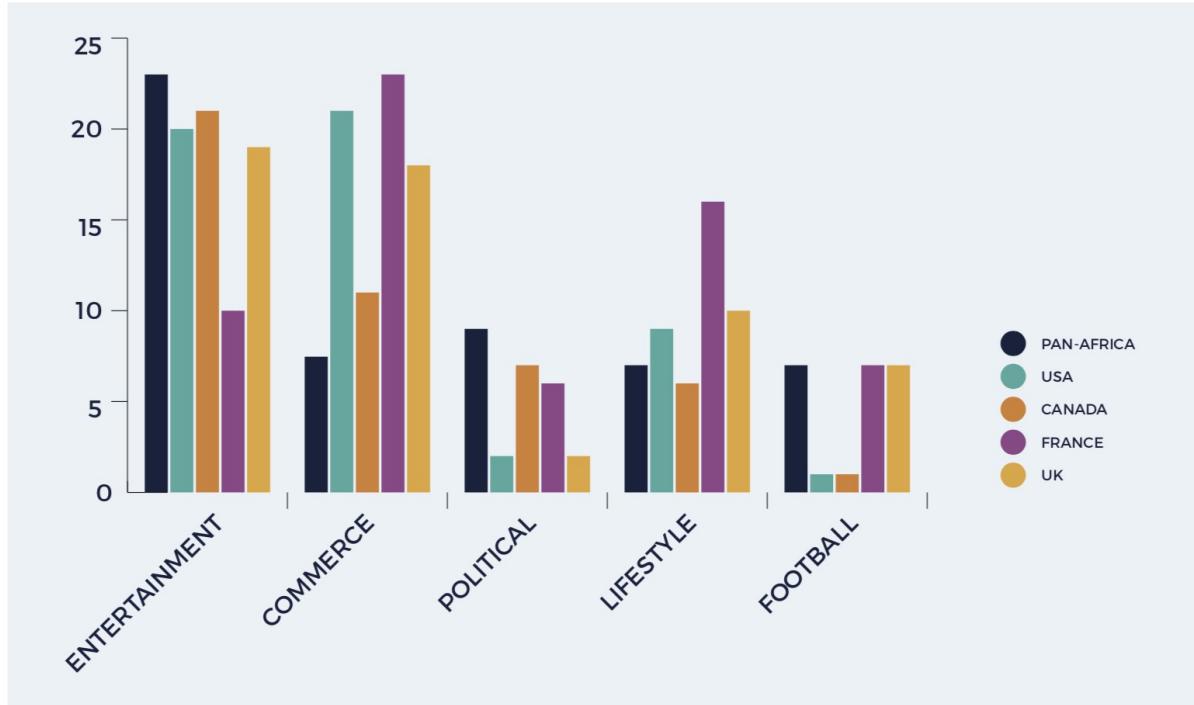
- Maps with bubbles
- Maps with pins

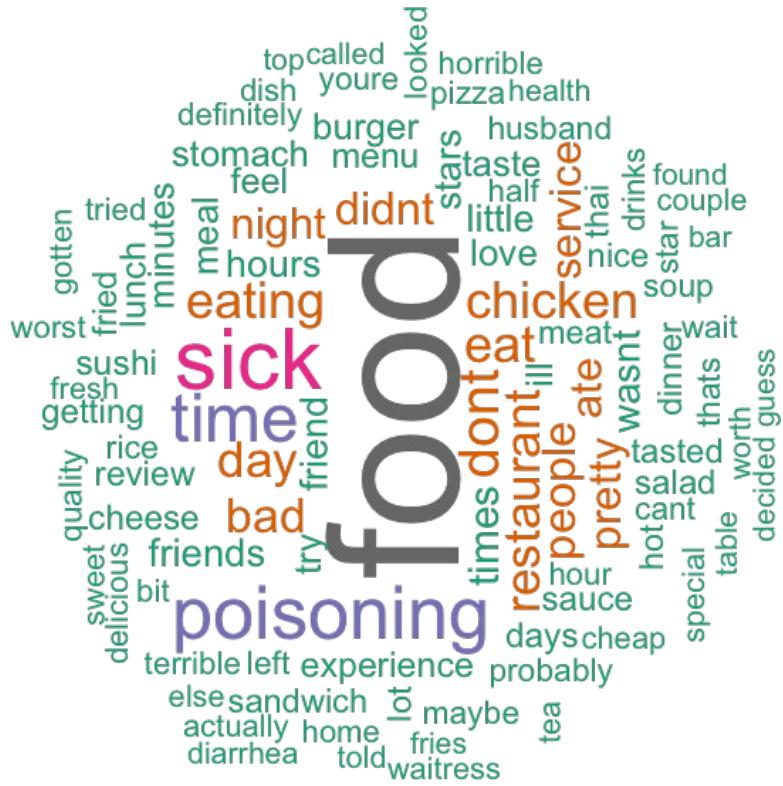
# When to use?

Useful for rankings

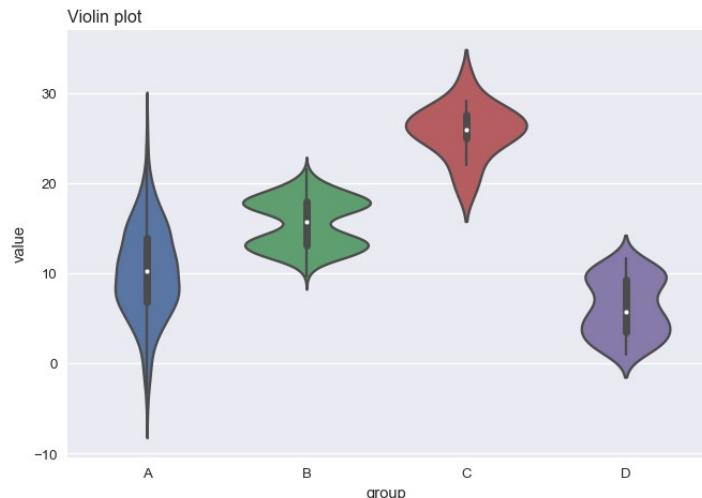
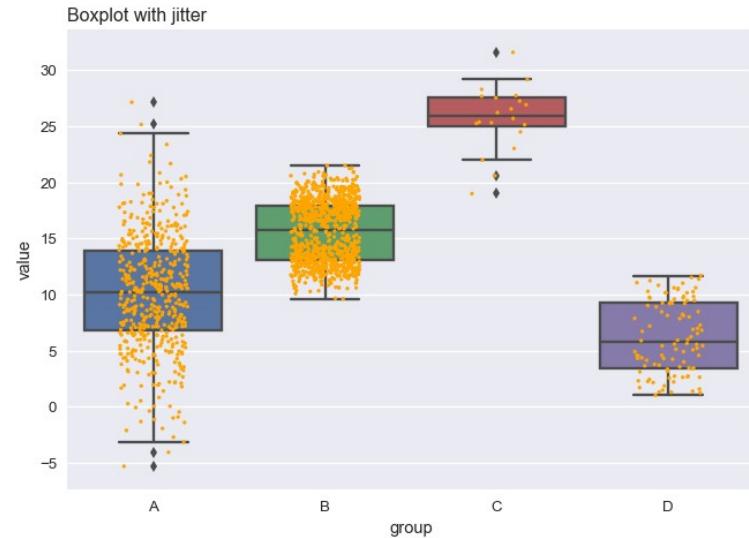
## Bar plots

Top five themes of hashtags around the world





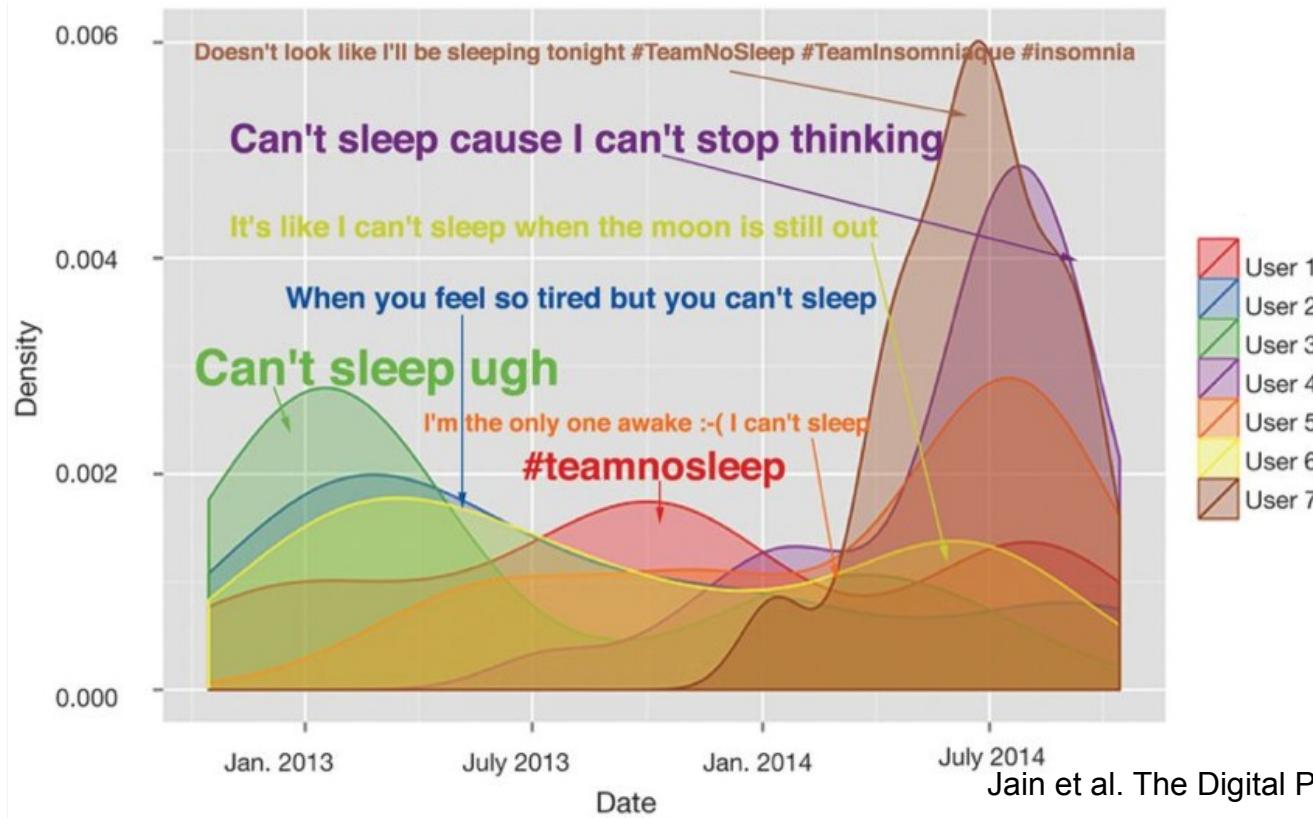
- Box plot
  - Lollipop plot
  - Word cloud

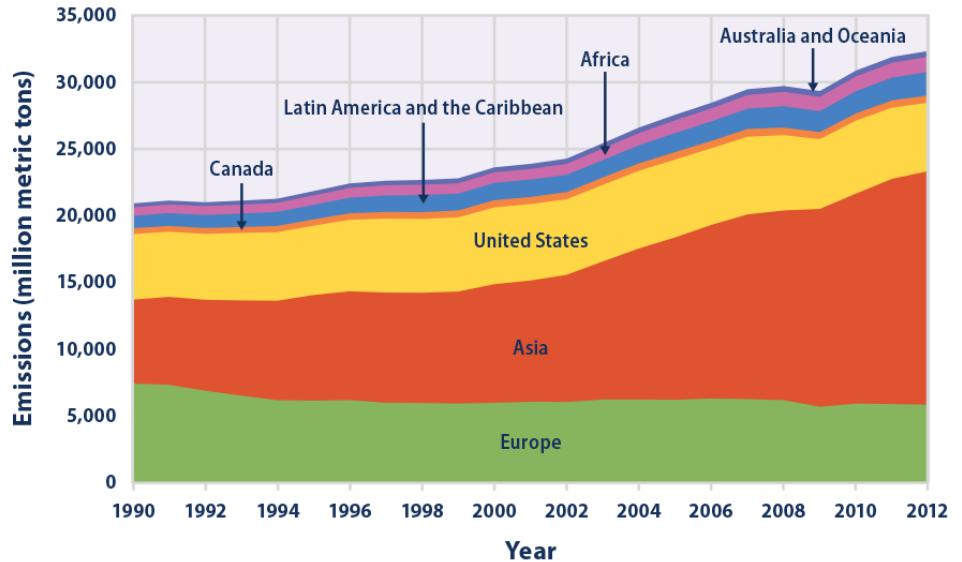


# When to use?

Useful for showing evolution

## Area/density plots

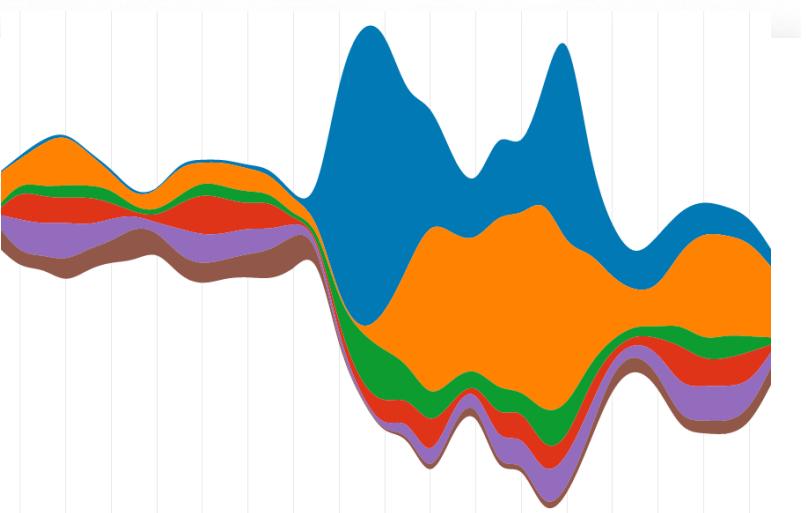
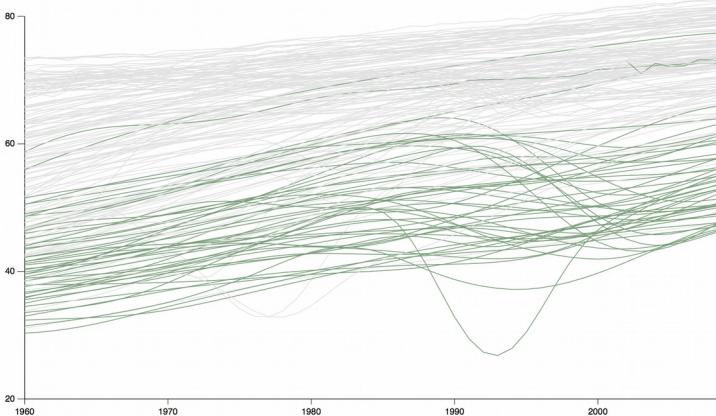




East Asia and Pacific | South Asia | Europe and Central Asia | Middle East and North Africa | Sub-Saharan Africa | Latin America and Caribbean | North America

## WORLD

The average life expectancy in the world in 2009 was 69 years.



- Line plot  
(Stacked) area plot  
Stream chart

# When to use?

Useful for information flow



# Networks

if 13 TRAXLER, GAVRIN, and LINDELL

PHYS. REV. PHYS. EDUC. RES. 14, 020107 (2018)

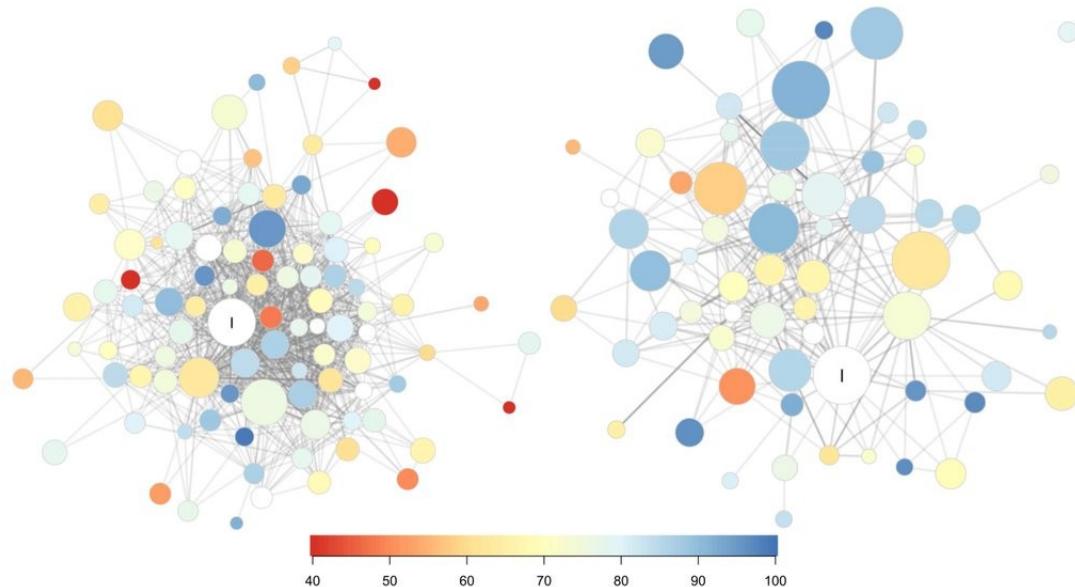


FIG. 4. Forum networks from weeks 7–8 in semester 1 (left) and semester 2 (right). Line opacity is scaled by edge weight, so darker lines indicate more threads in common for a student pair. Nodes are sized by total contributions over the semester and colored by grade (red low, yellow medium, blue high).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

## Flows of Global Health Financing

Total for 2016: \$37.5 billion in 2017 US dollars

Source: All

Channel: All

Region: Sub-Saharan Africa

Australia

BMGF

Canada

France

Germany

Other governments

Other sources

Private philanthropy

United Kingdom

United States

Australia

BMGF

Canada

Development banks

European Commission

France

Gavi

Germany

Global Fund

NGOs & foundations

Other bilateral aid agencies

UN agencies

United Kingdom

United States

Sub-Saharan Africa

Year: 2016



Target

Region

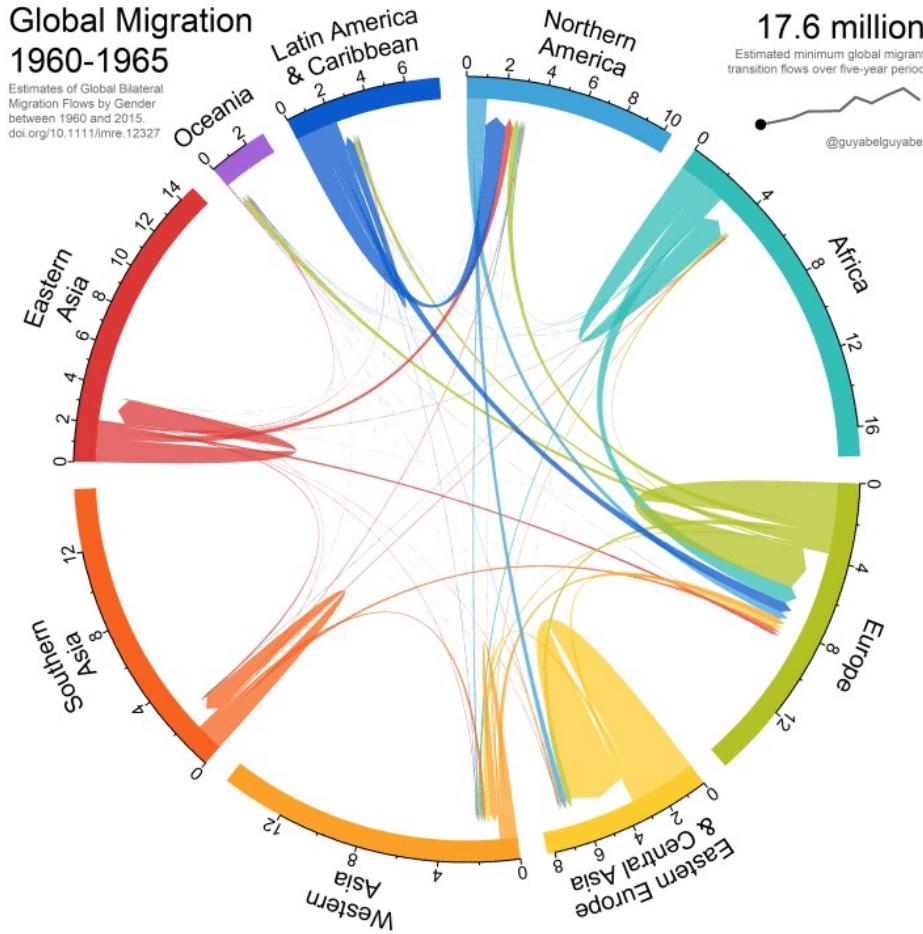
Health Focus Area

Reset

Code available from:

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

# Chord diagram



# Bad visualizations

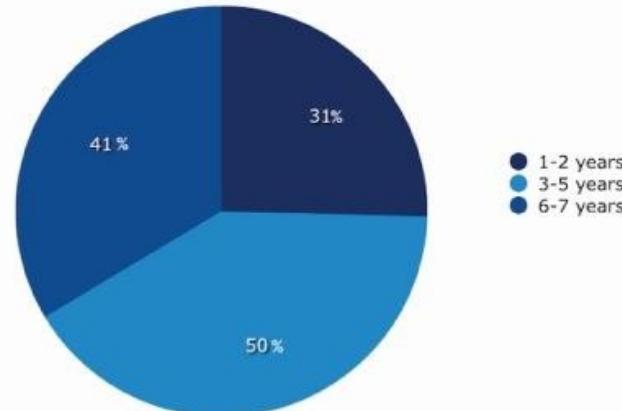
Which of these images has issues?

Have You Ever Liked a Brand on Facebook?



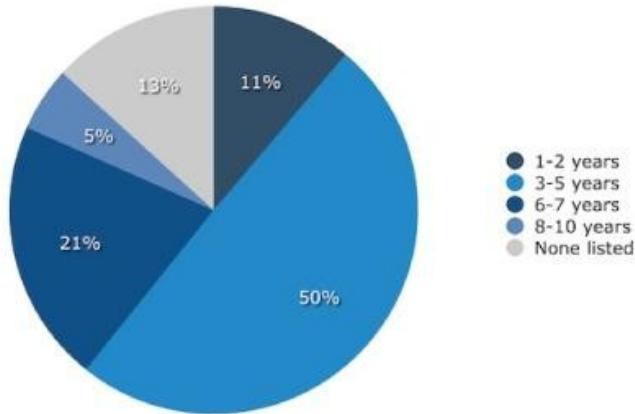
A

Years Experience Required by Employers



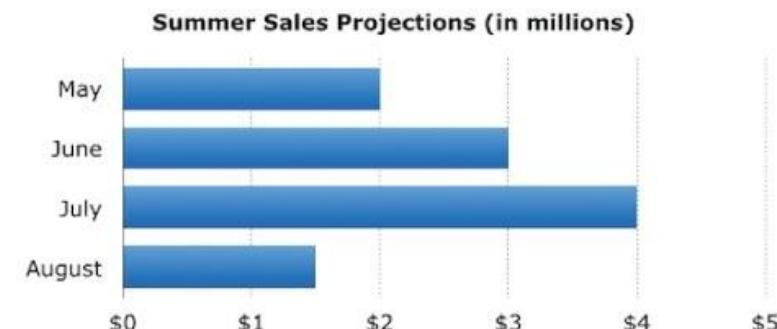
C

Which of these images has issues?



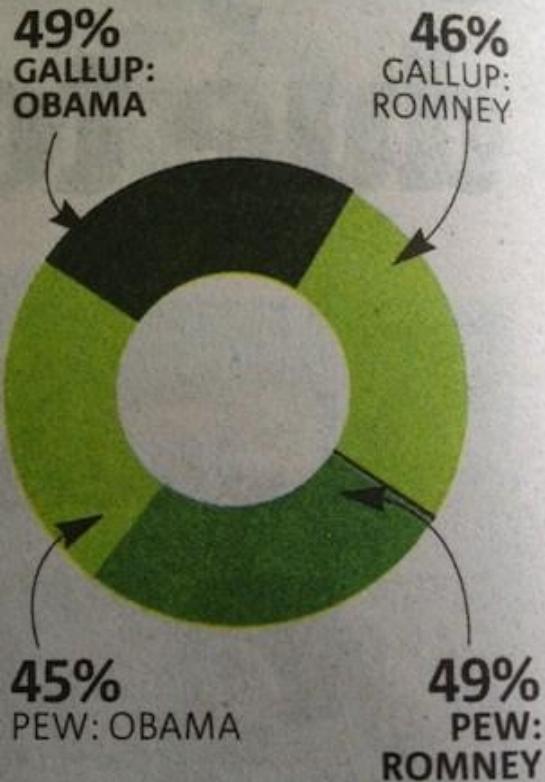
**B**

---

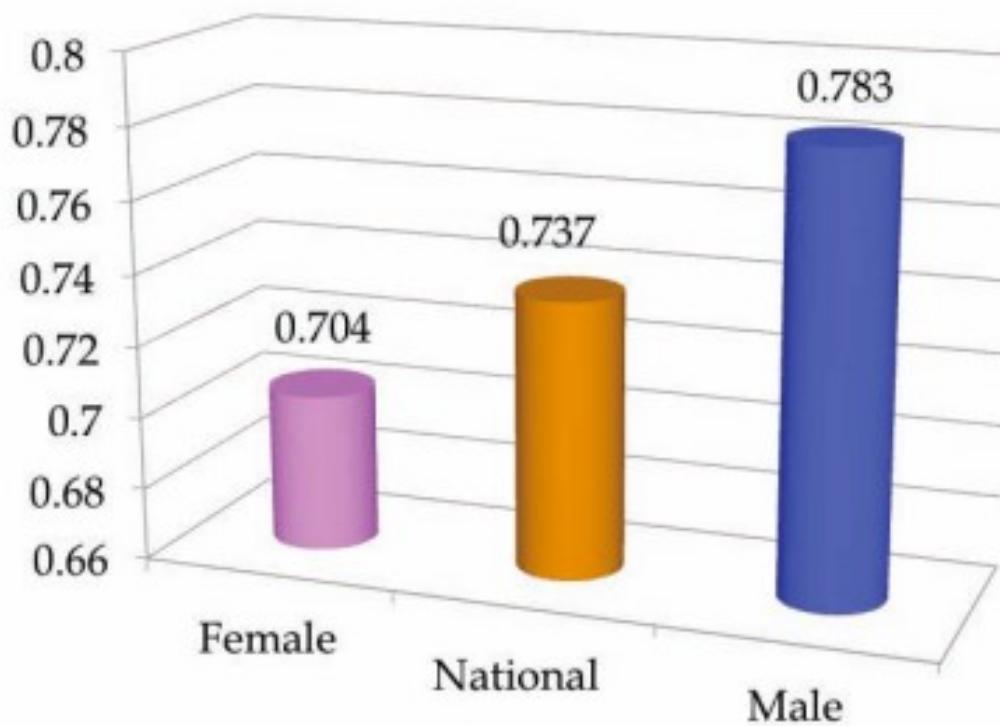


**D**

## Conflicting polls



# What's wrong with these images?



# Tools and Resources



# Python libraries

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

# The Chart Doctor

## 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: Alex Saxon; Data: Capital One UK; Economic Data: Bank of England; Financial Data: PwC; Health Data: NHS; Home Data: Rightmove; Manufacturing Data: FT; Retail Data: Ocado; Technology Data: Gartner; Transport Data: TfL

 [ft.com/vocabulary](http://ft.com/vocabulary)

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

### Deviation

Explain what variables (x) have a feed influence on the outcome (y). For example, if you want to show the effect of a variable over time or a long-term average. Can also be used to highlight outliers (see also the sheet).

Example FT uses

Home loan interest rates change

### Correlation

Show the relationship between two or more variables. Be careful, correlation does not mean causation. If you want to show causation, then choose the chart that's most appropriate for that (see also the sheet).

Example FT uses

House price vs income, income and the experience

### Ranking

Use when an ordering position is as critical as the value itself. For example, when showing the top 100 companies in the world, it's vital to highlight the points of interest.

Example FT uses

World's dimensions, league tables, international rankings

### Distribution

Show values in a dataset and how often they occur. For example, the frequency of a word in a text, or the frequency of a disease in a population. Distinguishing between discrete or continuous distributions is crucially important to provide suitable context for the chart.

Example FT uses

House price distribution, market capitalisation, revealing inequality

### Change over Time

Shows how things change over time. Trends can be linear or non-linear. It's important to highlight the lack of uniformity or equality in the data.

Example FT uses

House price trends, economic time series, sectoral changes in a market

### Magnitude

Show size comparisons. There can be many ways to do this, for example, by area, volume, count, or price. Comparing discrete or continuous magnitudes is crucially important to provide context for the chart.

Example FT uses

Market capitalisation, market capitalisation, volumes in general

### Part-to-whole

Show how a single entity can be broken down into its components. For example, the parts of a car, the parts of a company, or the parts of a sentence. Usually there are three components, though there could be more. Consider the number of components when choosing the chart.

Example FT uses

Fiscal budget, company structures, revenue breakdowns

### Spatial

Shows how locations map onto a variety of other variables. For example, more likely to be healthy in areas with better resources or geographical locations.

Example FT uses

Healthcare access, health migrants, securities, information relationship, general

### Flow

Show the value volumes or intensity of movement between locations. More likely to be used for flows between locations or regions.

Example FT uses

Trade routes, traffic, migration, securities, information relationship, general

### Diverging bar

A simple standard bar chart with both negative and positive values.

### Scatterplot

The standard way to show the relationship between two continuous variables.

### Ordered bar

Used when data starts with a value of zero and increases sequentially.

### Histogram

The standard way to show the distribution of a single variable. It's good for showing the shape of the data.

### Line

The standard way to show a changing trend. Good for showing the direction of the data, but less good for showing the magnitude.

### Column

The standard way to show the magnitude of things that don't have a natural order.

### Stacked bar/column/bar

A simple way of showing whole values while highlighting their individual components.

### Basic choropleth (map)

The standard approach for mapping data onto a map - should always be used unless it's difficult to read or looks like a dot map.

### Proportional symbol (map)

Use for dot maps where each dot's size corresponds to a value. Data will be hard to read.

### Waterfall

Designed to show the sequencing of data flows. Good for showing processes, typically sequential or complex processes.

### Diverging stacked bar

Perfect for presenting data that has a natural ordering, such as years, months, days, etc.

### Column + line timeline

A good way of showing the relationship between two variables over time.

### Ordered volume

Used when there are big differences between values.

### Dot plot

A simple way of showing multiple data points. Good for showing multiple categories.

### Dot strip plot

Good for showing individual values in a large dataset when there are too many to plot all at once.

### Bar

Good for showing the magnitude of data across multiple categories.

### Bar

Good for showing the magnitude of data across multiple categories.

### Bar

Good for showing the magnitude of data across multiple categories.

### Bar

Good for showing the magnitude of data across multiple categories.

### Spine

Shows a single value or range of values across multiple categories.

### Bubble

Like a scatterplot, but with additional detail about the size of the bubbles.

### Dot strip

Shows data in order, on a strip or a grid, with each row or column showing a different category.

### Barcode plot

Like a strip plot, but for drawing all the bars at once. Good for highlighting individual categories.

### Slope

Good for showing a changing rate over time, or a slope between two points without showing a lot of detail.

### Paired column

For an interrelated column, but allows for a wider range of values than a standard column.

### Paired bar

For an interrelated bar chart, but allows for a wider range of values than a standard bar.

### Paired pie

A common way of showing two related data sets at the same time - as long as it's not too hard to read.

### Bar chart

Similar to a pie chart - but the counts for each category are shown as bars. Good for showing the magnitude of the data.

### Surface/depth line

The shaded area of the chart shows a depth of field, allowing the user to see the difference between two series.

### XY heatmap

A good way of showing the relationship between two categories, but less good for showing the differences in amounts.

### Dot

Good for showing how data points relate to one another over time or space.

### Violin plot

Similar to a box plot but with a violin shape showing the density of data points for each category.

### Population pyramid

A standard way for showing the distribution of a population by age and sex.

### Area chart

Similar to a line chart, but showing the area under the line, or the area between two lines.

### Mathematic

A good way of showing the proportion of data at different levels, but the data needs to be very specific.

### Propotional symbol

Used when there are big differences in values and seeing the components is more important than the total.

### Chord

Similar to a pie chart - but the counts for each category are shown as chords.

### Frequency polygon

Another way of showing the distribution of data, but less good for showing the magnitude of the data.

### Calender heatmap

A good way of showing the relationship between two variables.

### Connected scatterplot

A good way of showing data points that are related to each other. Good for showing the relationship between data points.

### Radar

A good way of showing multiple data points on a circular grid.

### Parallel coordinates

An alternative to radar charts - good for showing multiple data points from different categories.

### Circle reader

Good for showing the magnitude of data across multiple categories.

### Ground symbol

An alternative to a scatterplot when there are many points, especially well suited for geographical data.

### Vertical treemap

Presents data on the Y axis, good for showing detailed data across multiple categories.

### Streamgraph

Another alternative to a scatterplot when there are many points, especially well suited for geographical data.

### Heatmap

Used for showing the relationship between two variables.

### Dot density

Stretching and shrinking dots so that each one is scaled to a particular value.

### Dot density

Used to show the location of individual observations.

### Gridplot

Good for showing % information, works well with categorical variables and small numbers of items.

### Venn

Generally only used for schematic representation.

### Waterfall

Can be useful for showing cumulative effects, especially when the components are negative.



© Financial Times



# Other tools

- Tableau
- R ggplot2 and others
- D3

**Next ... ipython tutorial**