ROBERT GORDON UNIVERSITY ABERDEEN

Calling Bullsh*t on Data Visualisations!

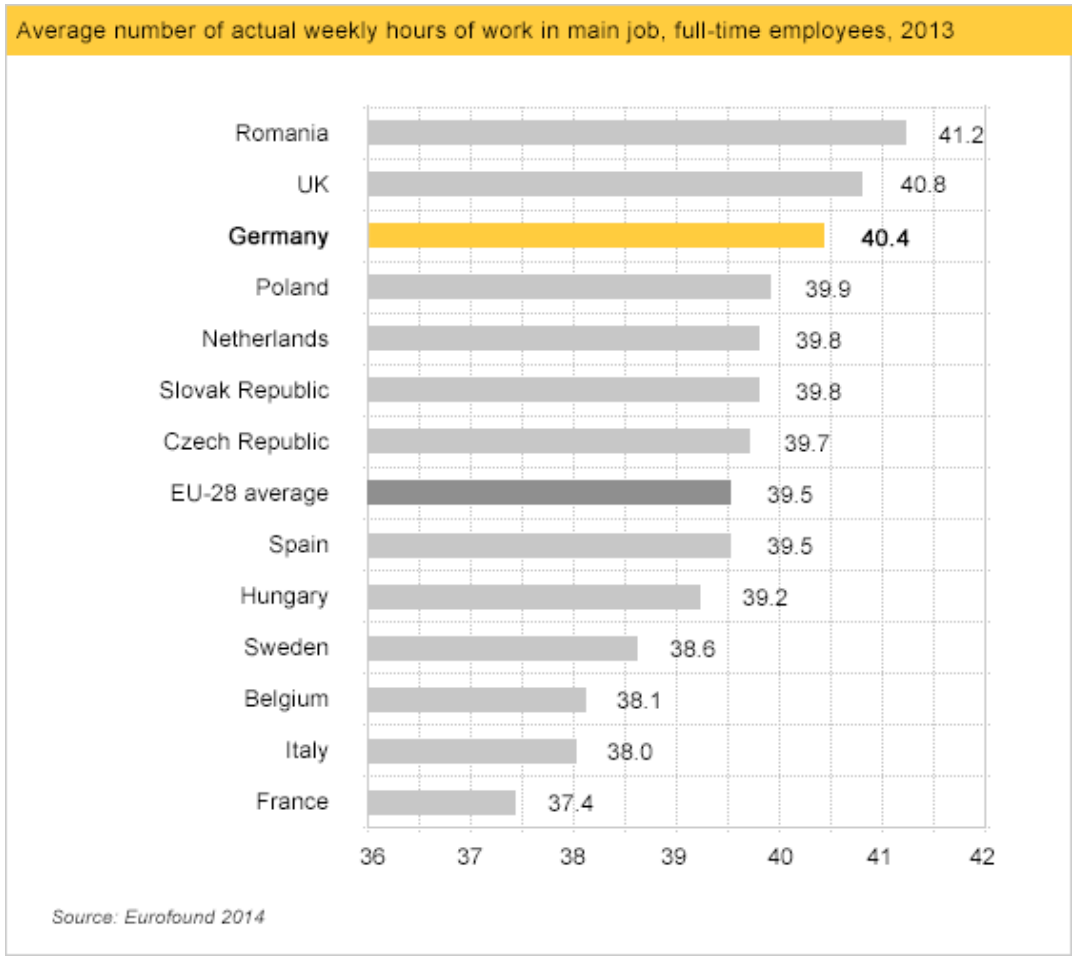
CM4125 – Topic 3

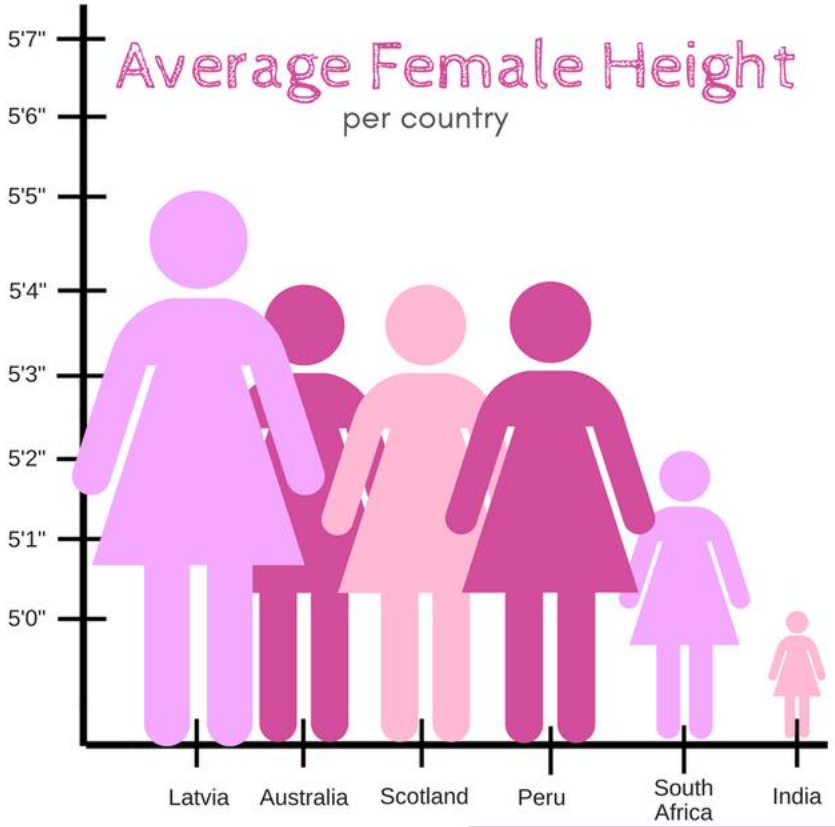
What this lecture is about?

- Based on Week 6 of the Calling Bullsh*t course of the University of Washington by Carl T. Bergstrom and Jevin West
  - https://www.callingbullshit.org/syllabus.html#Visual

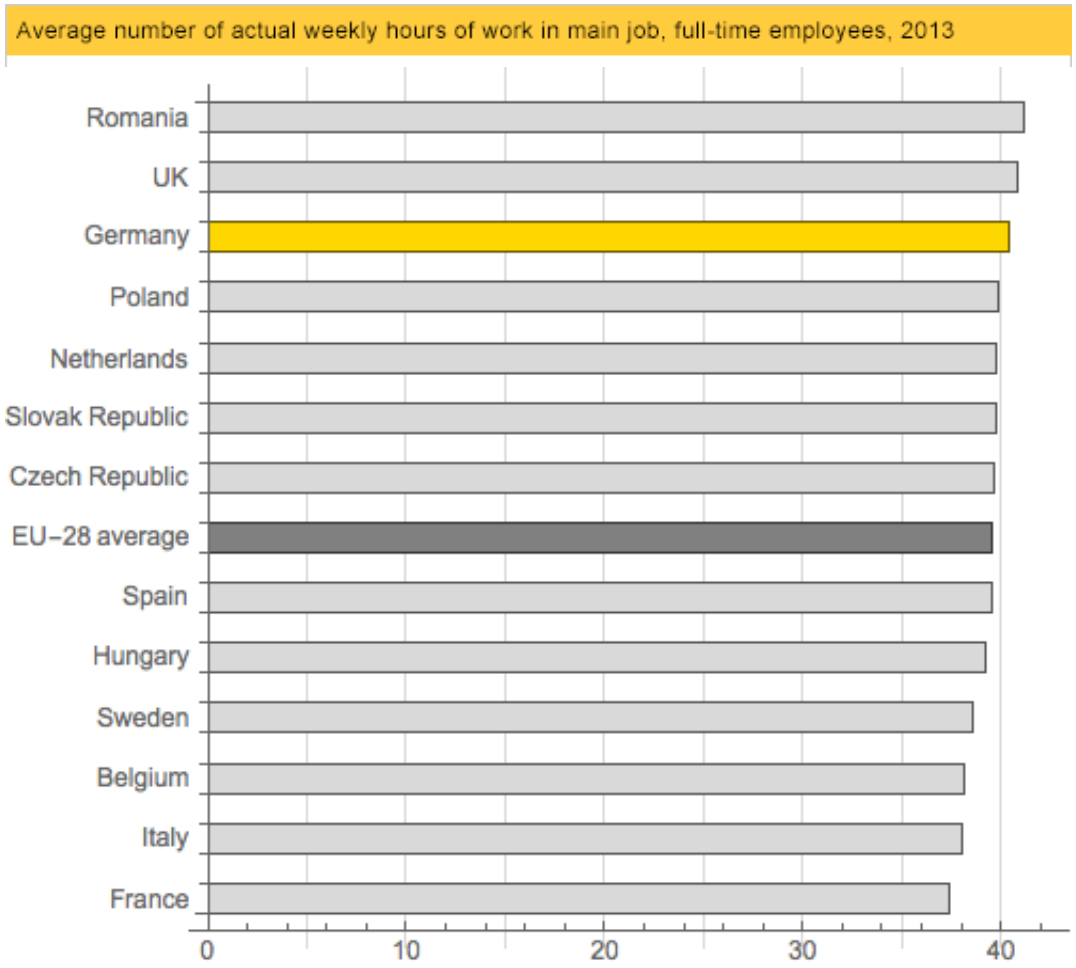
- Supplementary reading:
  - Alberto Cairo (2019) *How Charts Lie: Getting Smarter about Visual Information*. W.W. Norton and Company.
  - Edward Tufte (1983) *The Visual Display of Quantitative Information*. Chapters 2 (Graphical integrity) and 5 (Chartjunk: vibrations, grids, and ducks).

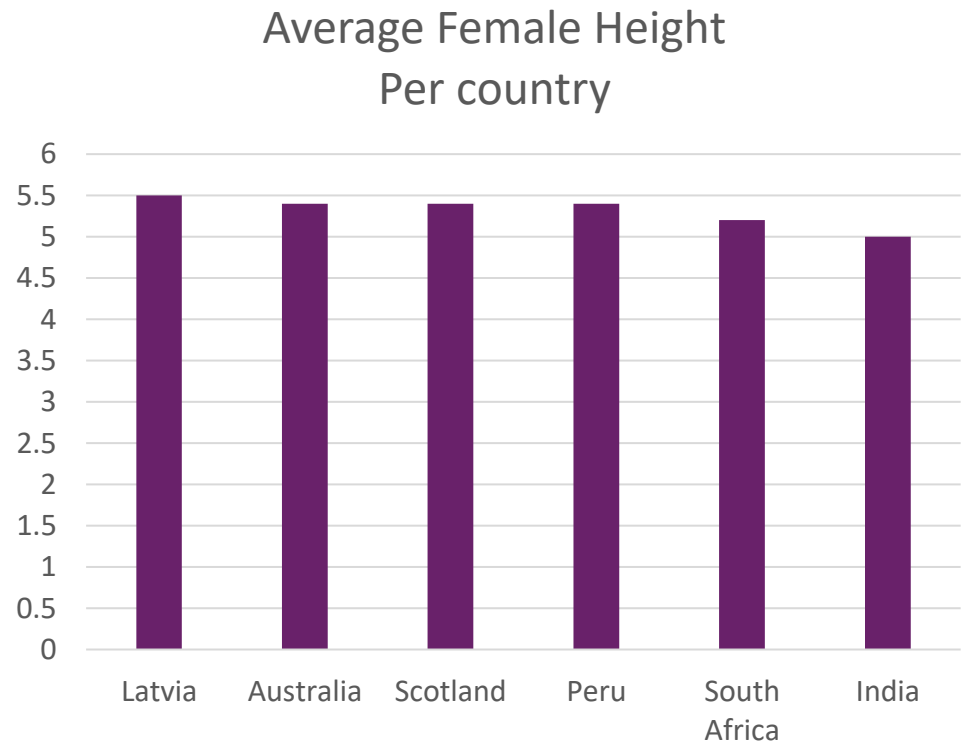
1. Bar chart y-axes **need to** include zero



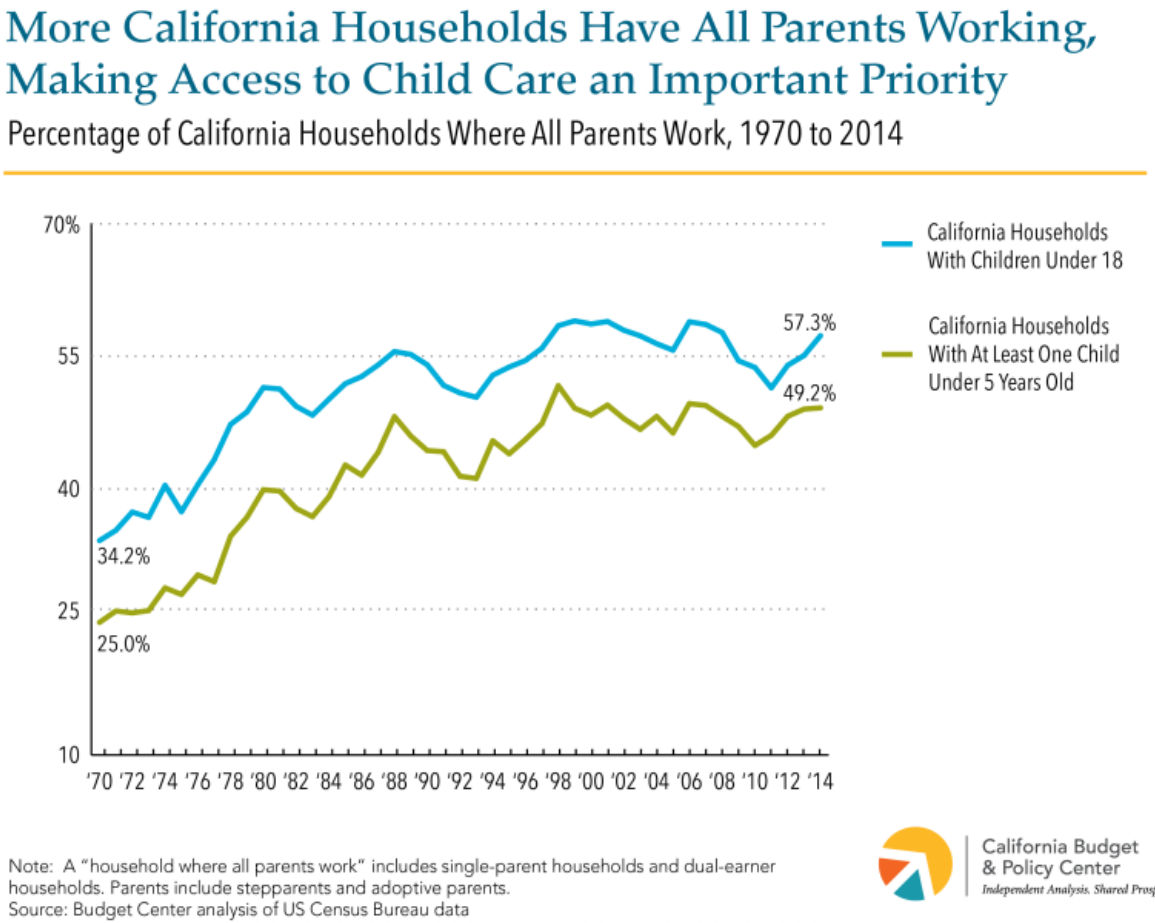


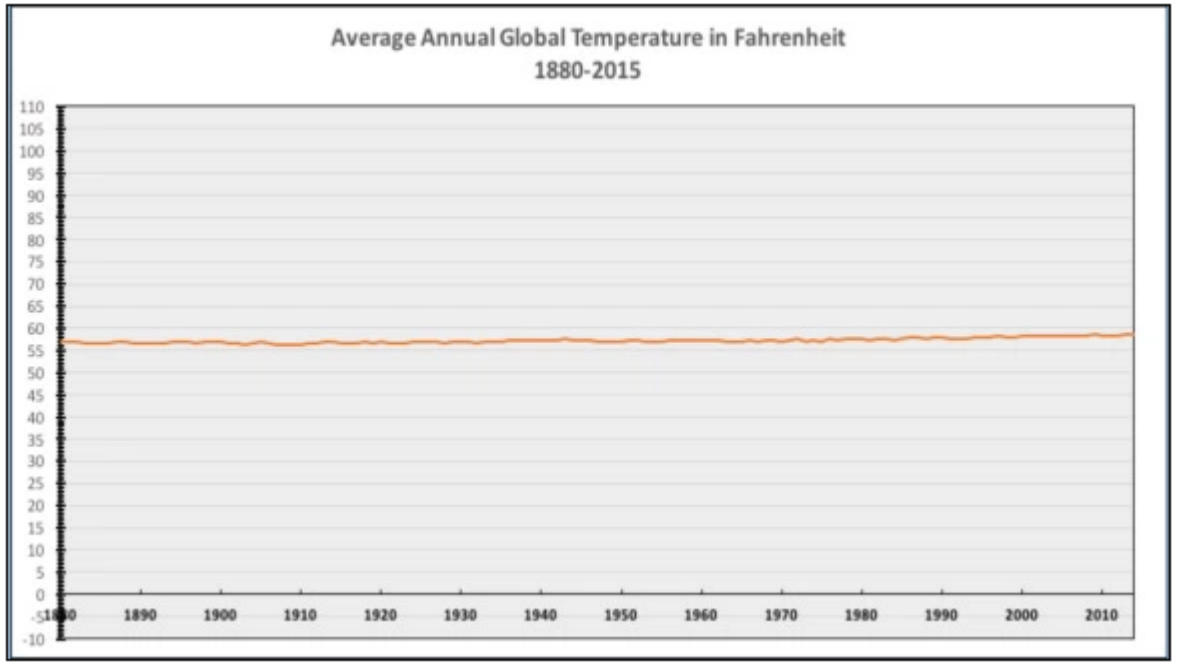
1. Bar chart y-axes **need to** include zero



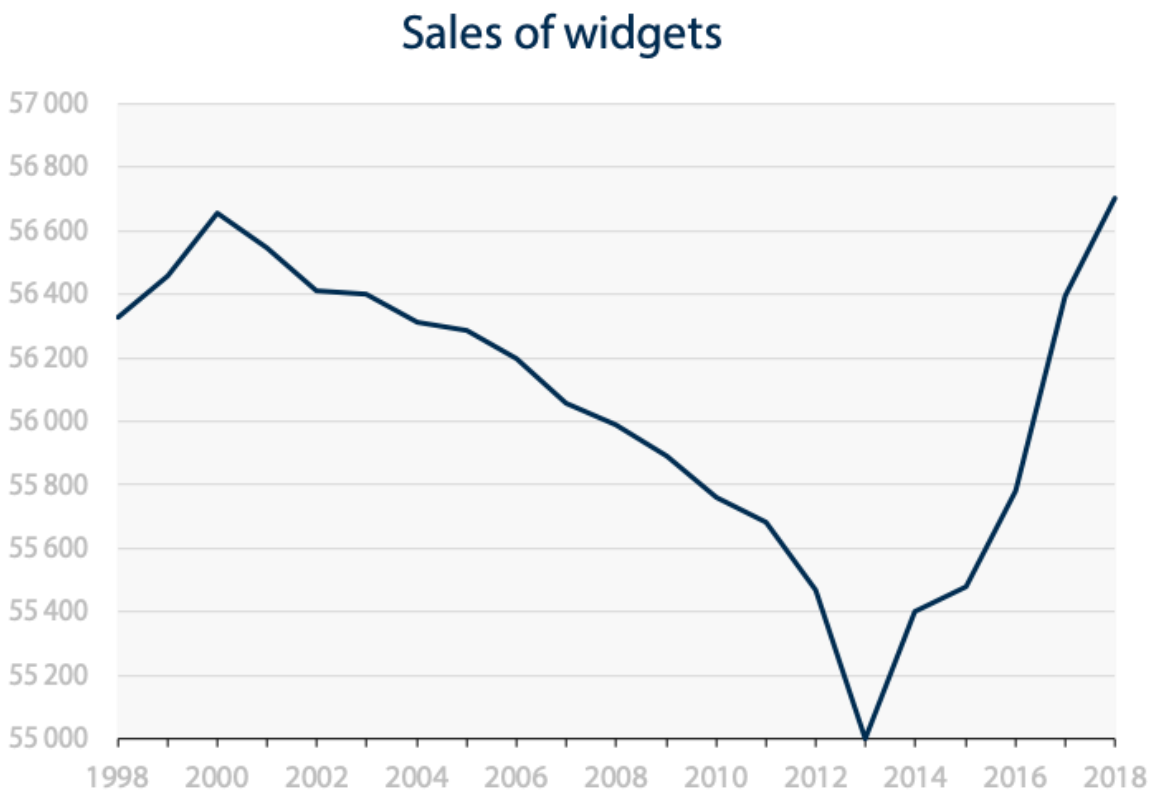


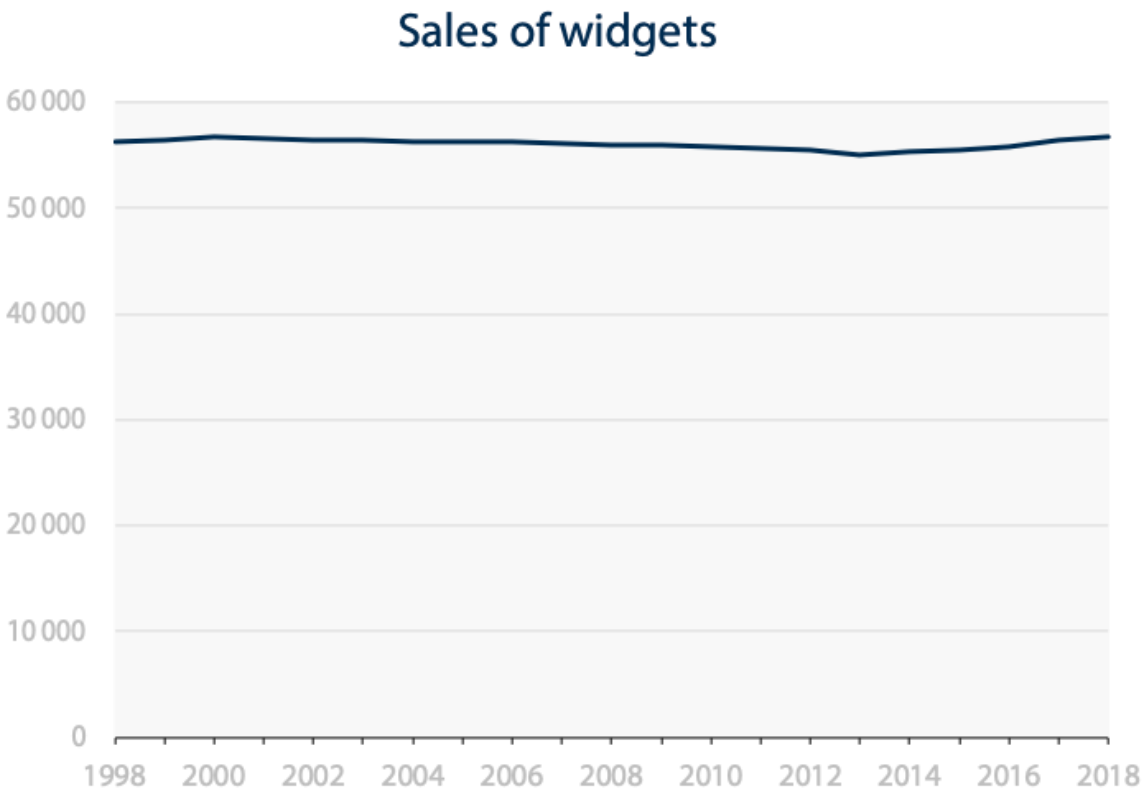
2. Line graph y-axes **don't need** to include zero





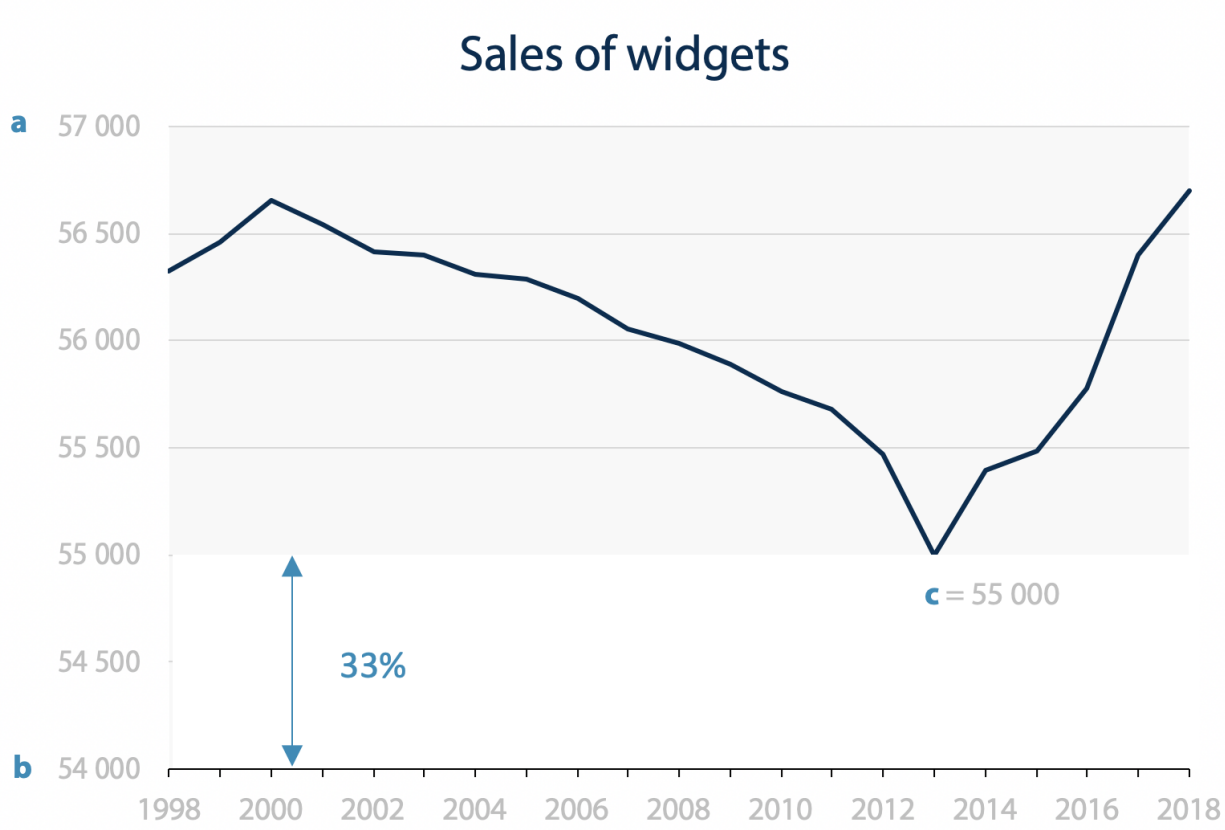
Well…


Sales of widgets
57 000
56 800
56 600
56 400
56 200
56 000
55 800
55 600
55 400
55 200
55 000
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018


Sales of widgets
60 000
50 000
40 000
30 000
20 000
10 000
0
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
2018

Solution

Empty space = $\frac{\text{Minimum data value - Minimum scale value}}{\text{Maximum scale value - Minimum scale value}}$



$$x = \frac{(c-b)}{(a-b)}$$

$$b = \frac{3c-a}{2}$$

$$b = \frac{3 * 55{,}000 - 57000}{2} = 54\,000$$

3. **No** multiple axes on a single graph

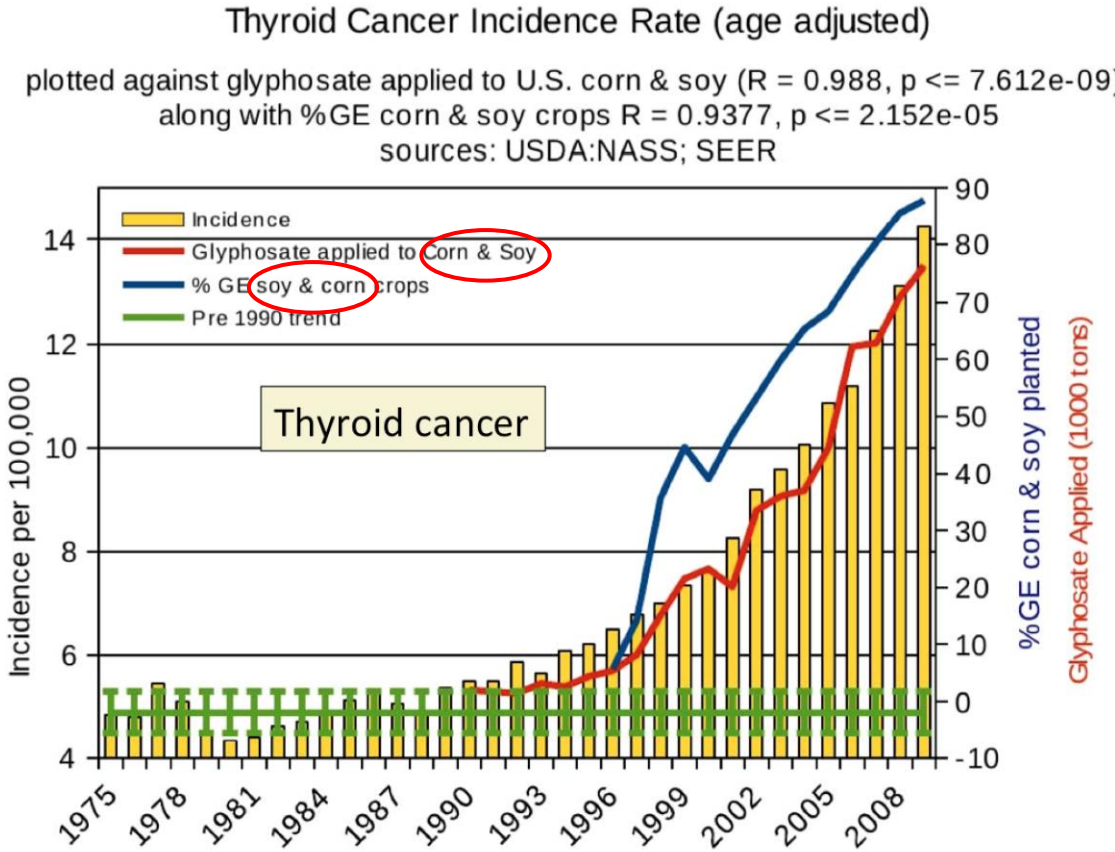
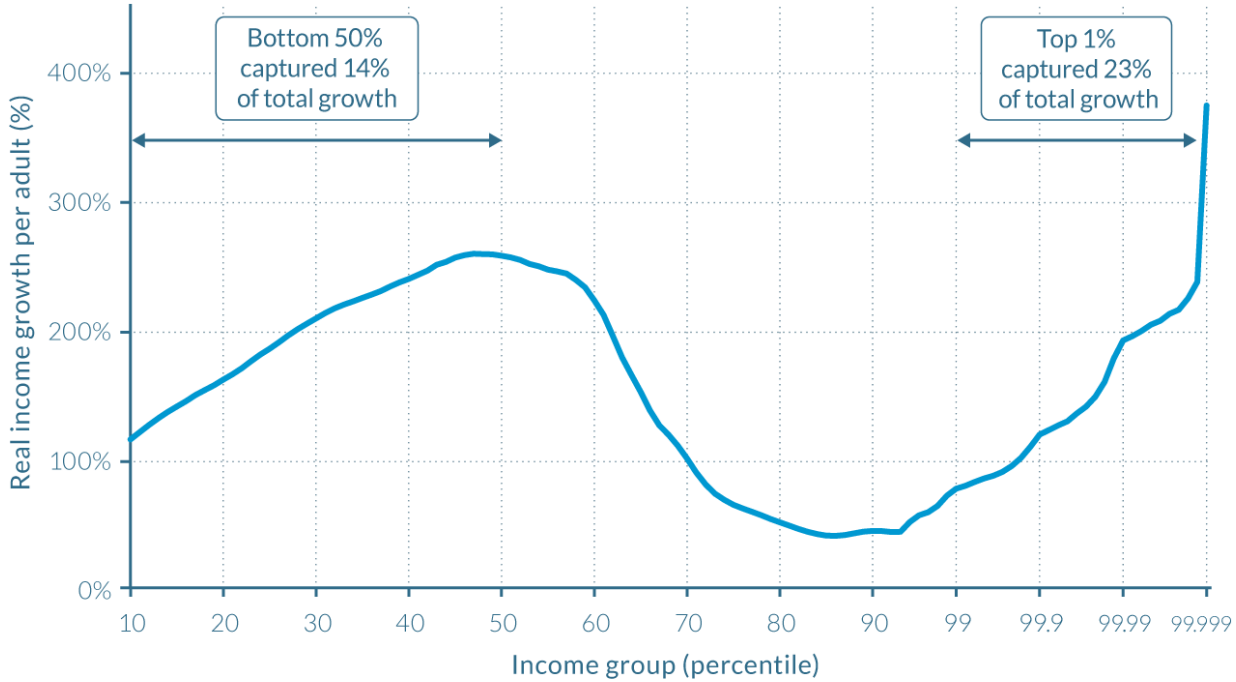


*Figure 10, Swanson et al. Journal of Organic Systems 2014; 9(2):6-37.

How to solve this?

4. An axis should **not** change scales midstream



Source: WID.world (2017). See wir2018.wid.world/methodology.html for data series and notes.

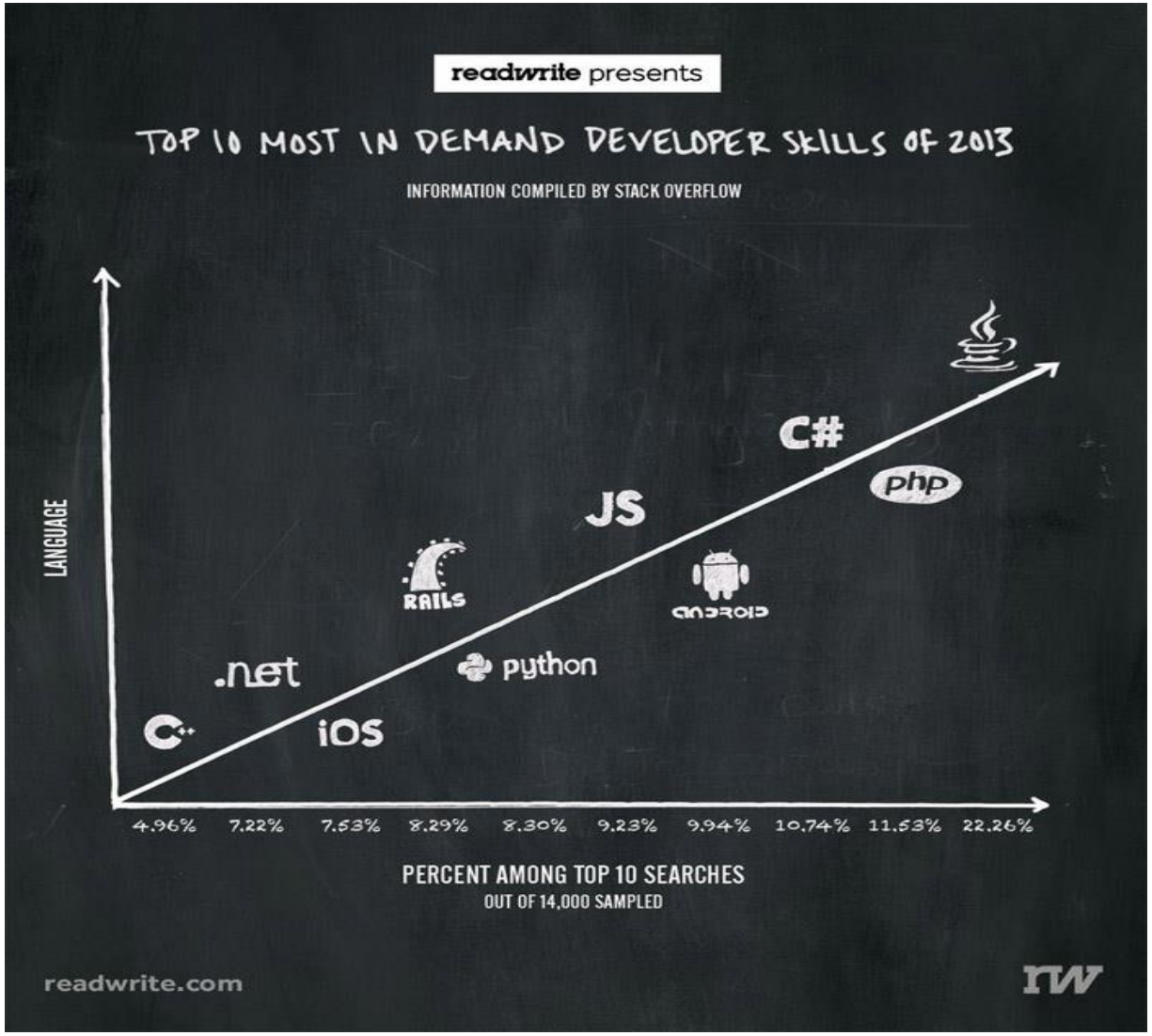
On the horizontal axis, the world population is divided into a hundred groups of equal population size and sorted in ascending order from left to right, according to each group's income level. The Top 1% group is divided into ten groups, the richest of these groups is also divided into ten groups, and the very top group is again divided into ten groups of equal population size. The vertical axis shows the total income growth of an average individual in each group between 1980 and 2016. For percentile group p99p99.1 (the poorest 10% among the world's richest 1%), growth was 77% between 1980 and 2016. The Top 1% captured 23% of total growth over this period. Income estimates account for differences in the cost of living between countries. Values are net of inflation.

Use log scales!

Why not use the golden ratio?!

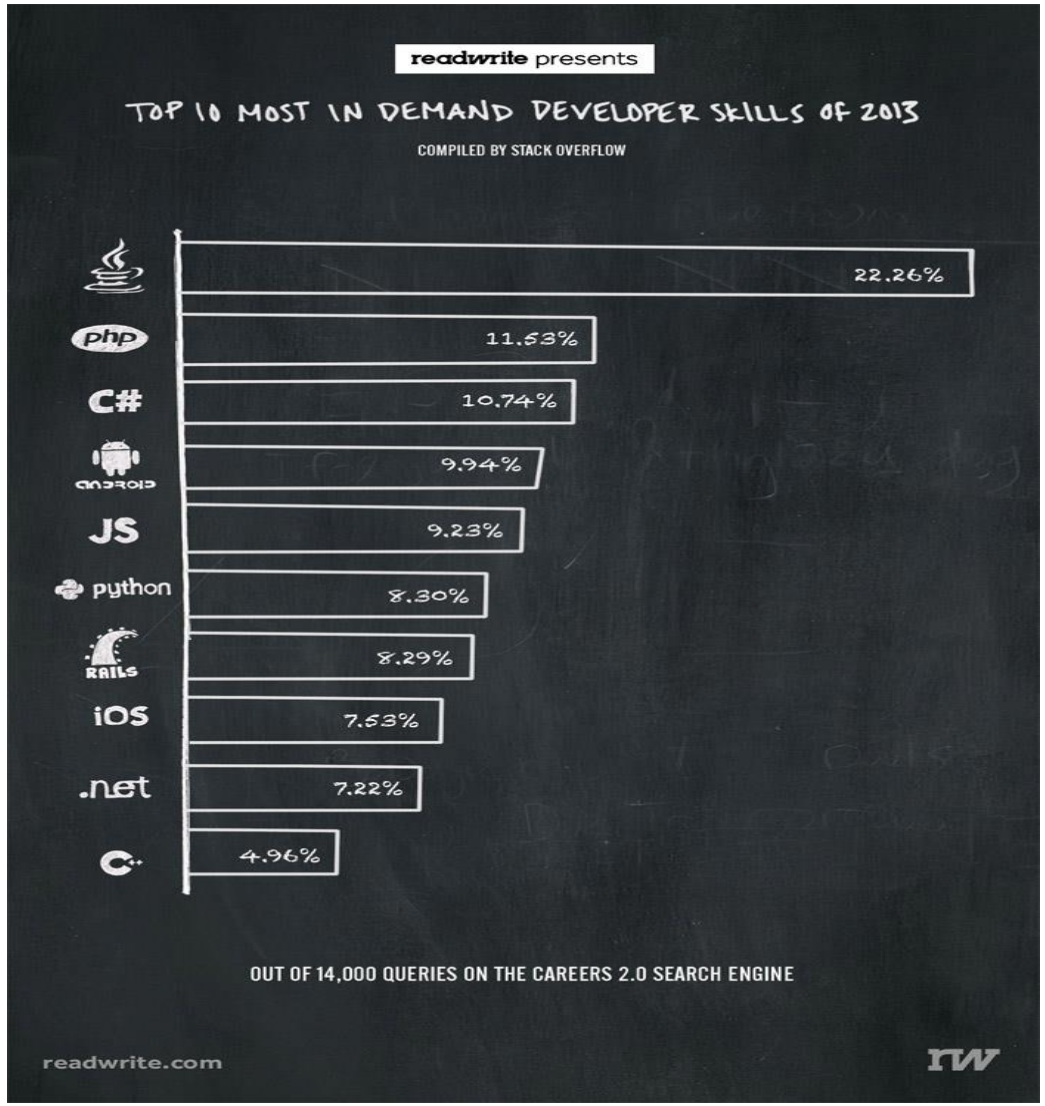
Example (of what NOT to do)

5. An axis should have *something* on it



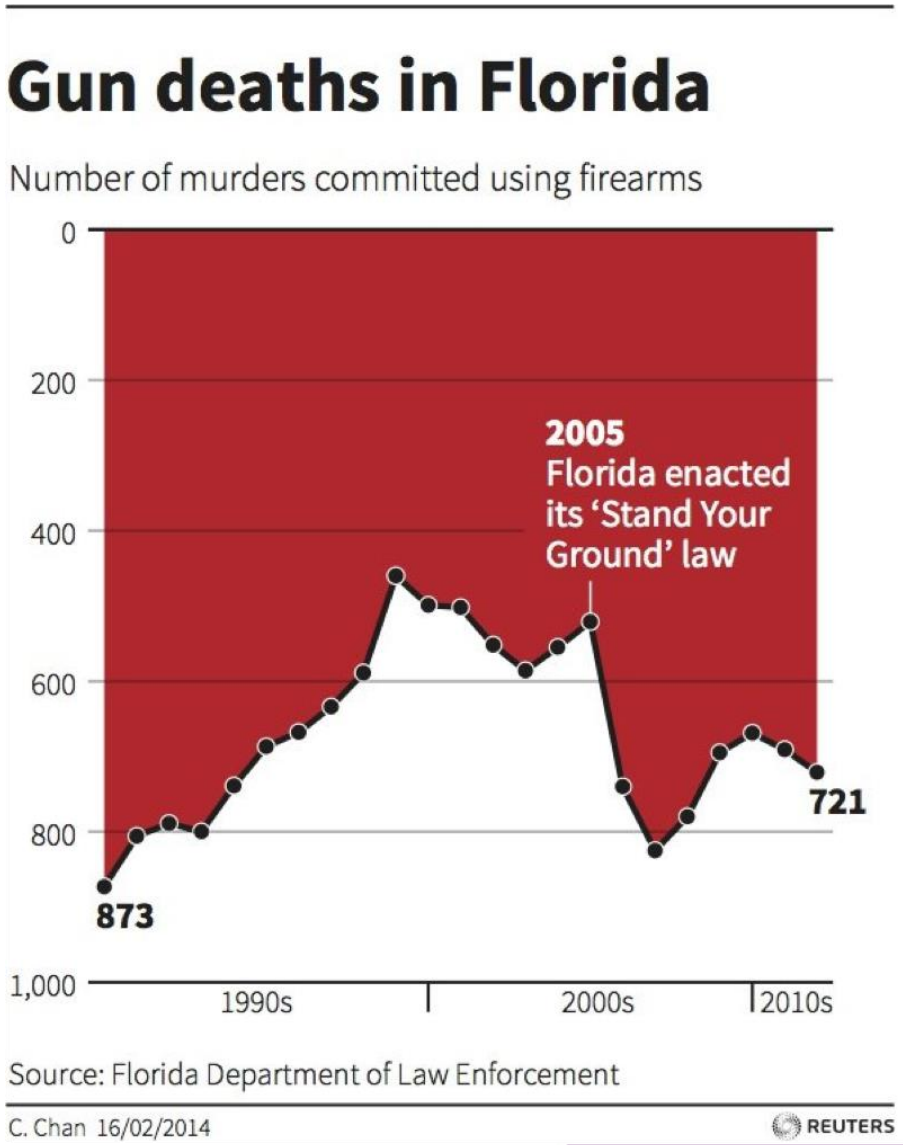
Like this

Tools such as WebPlotDigitizer can help you get the missing info

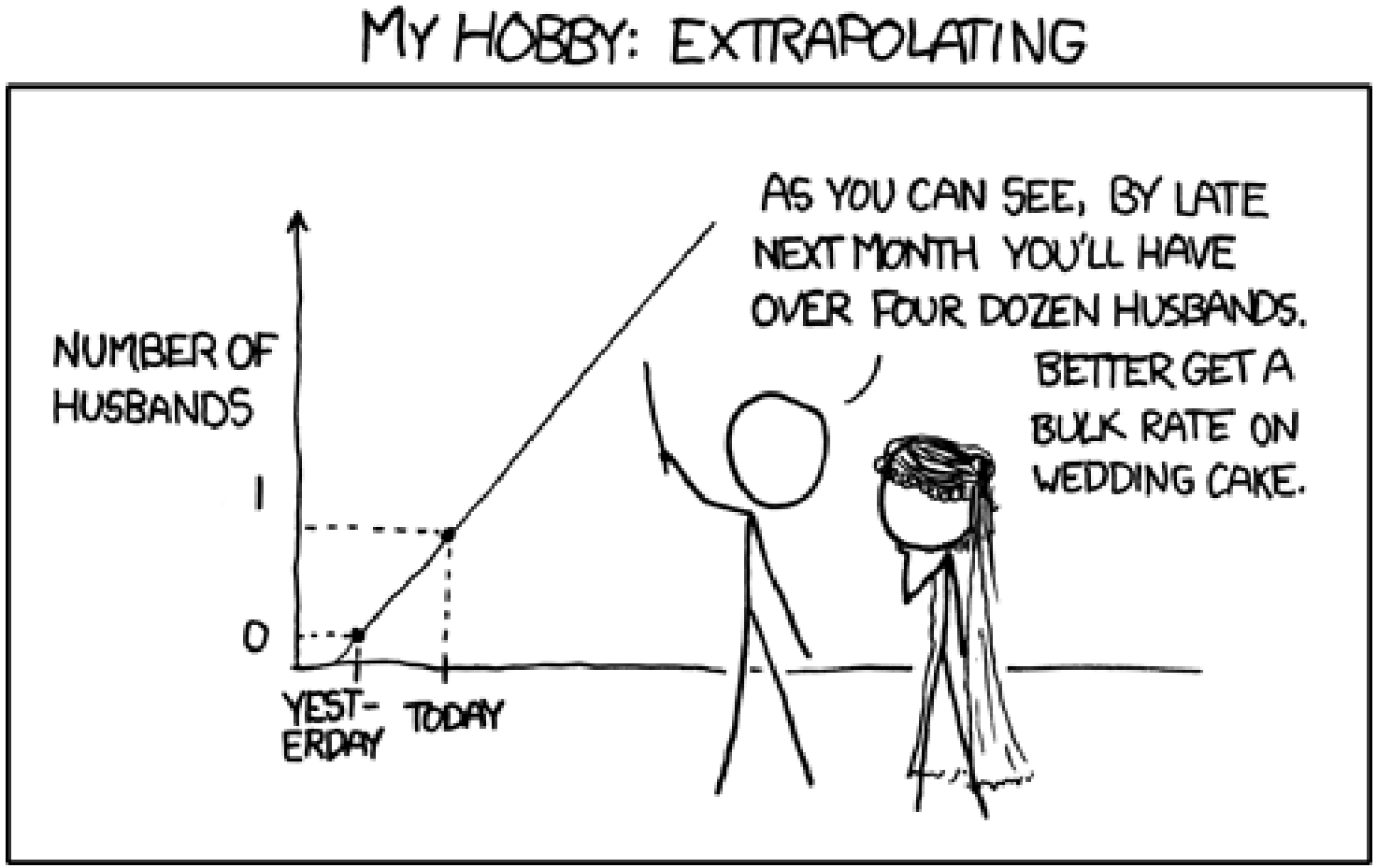


6. **Don't** invert the axis(?)

The author defended it!
What's your stance?



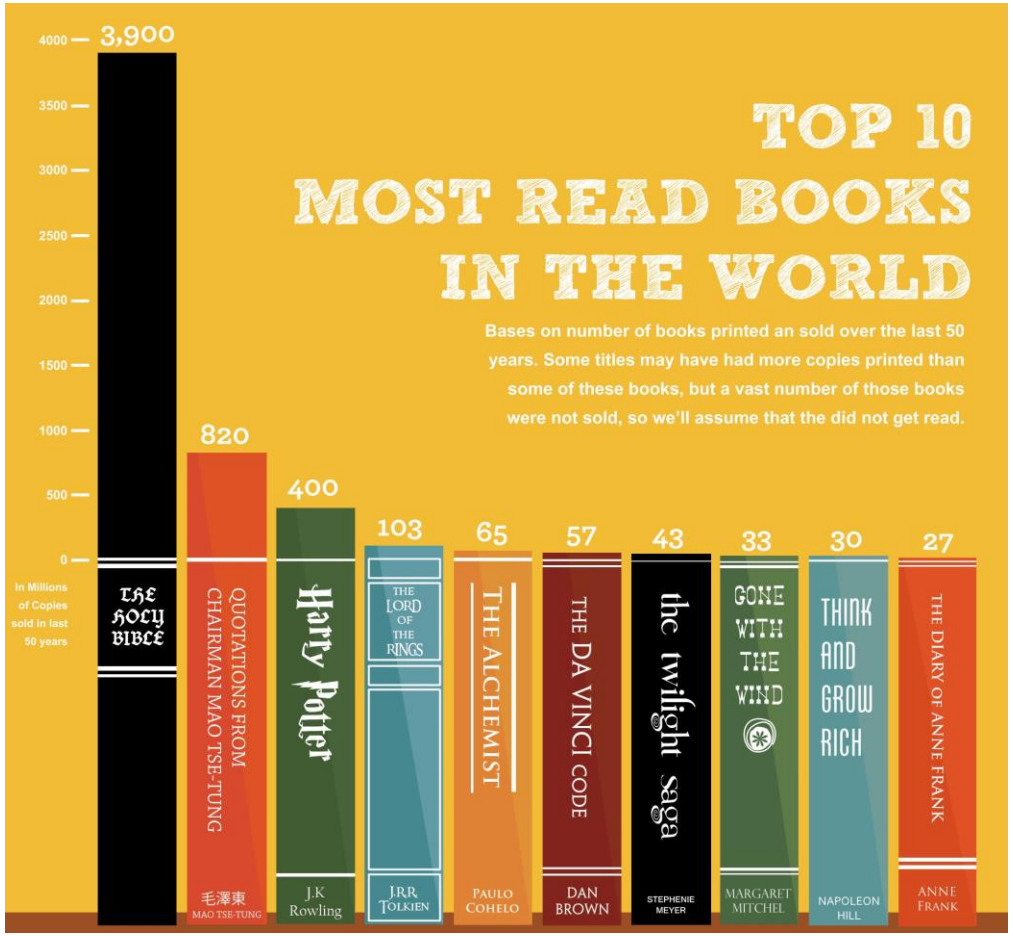
7. Avoid extrapolation!

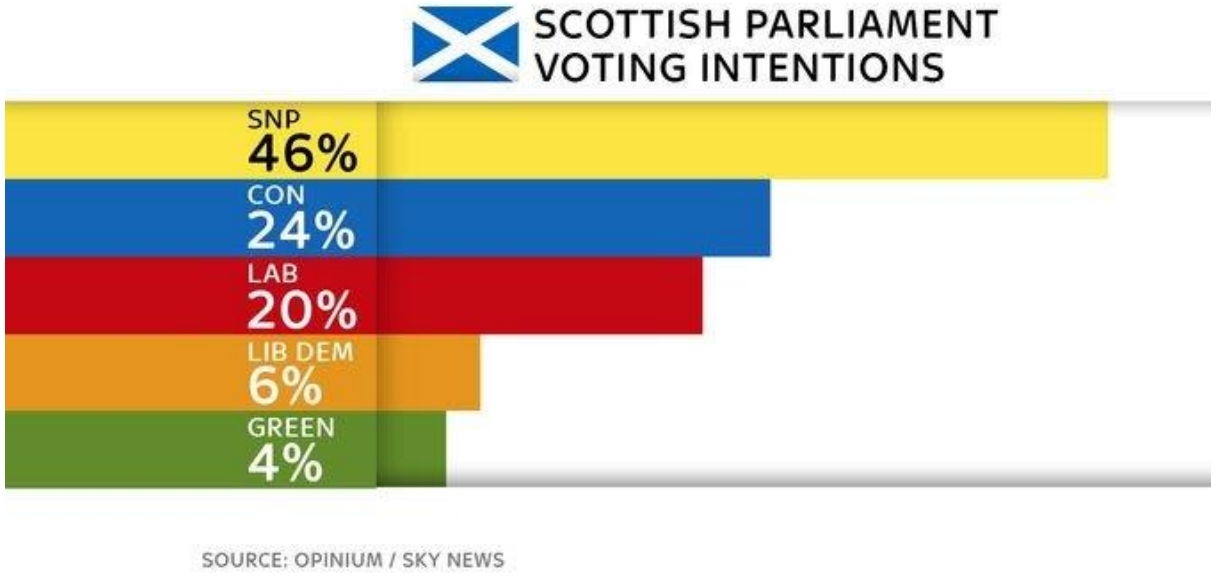


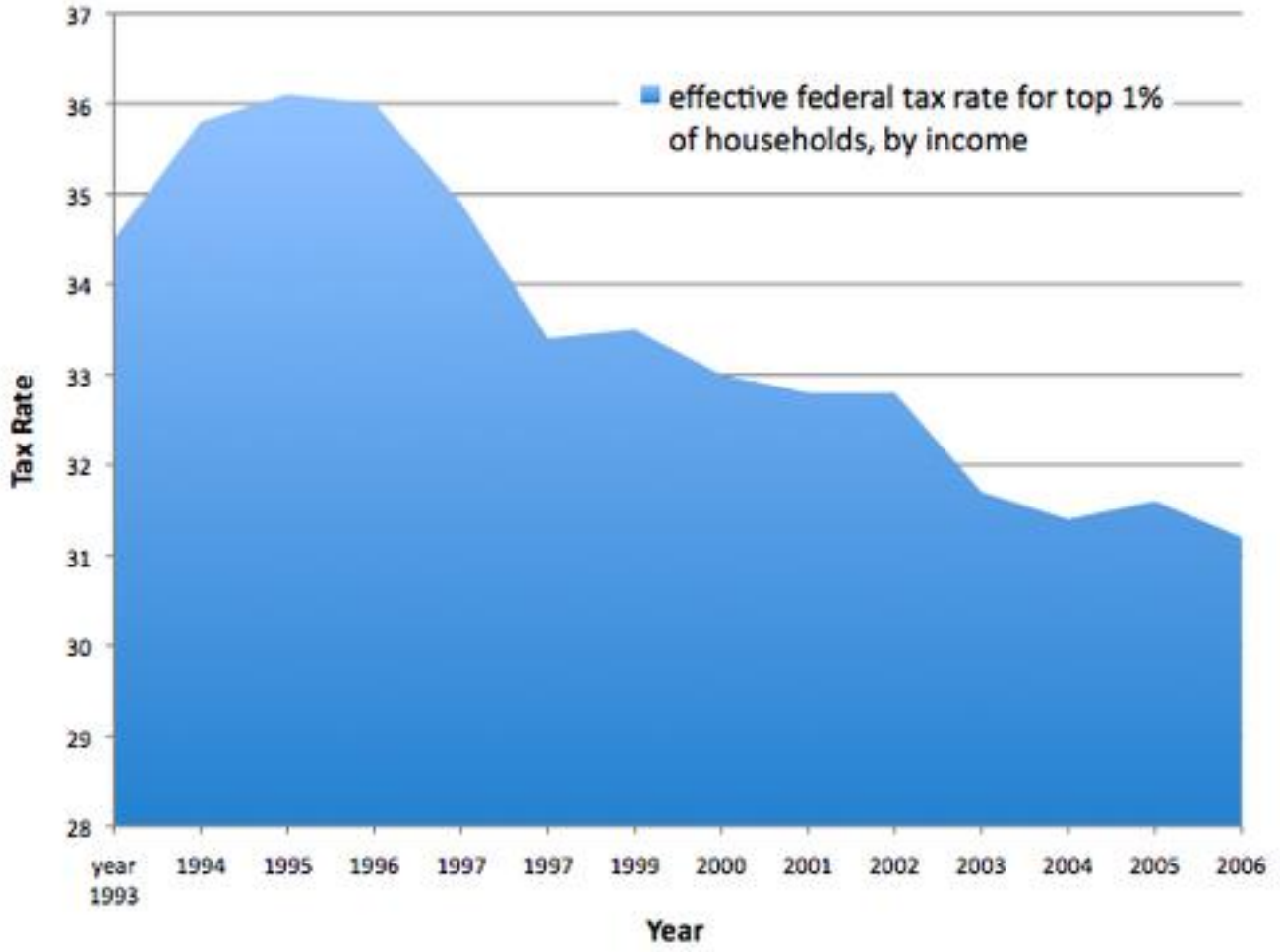
8. Proportional Ink

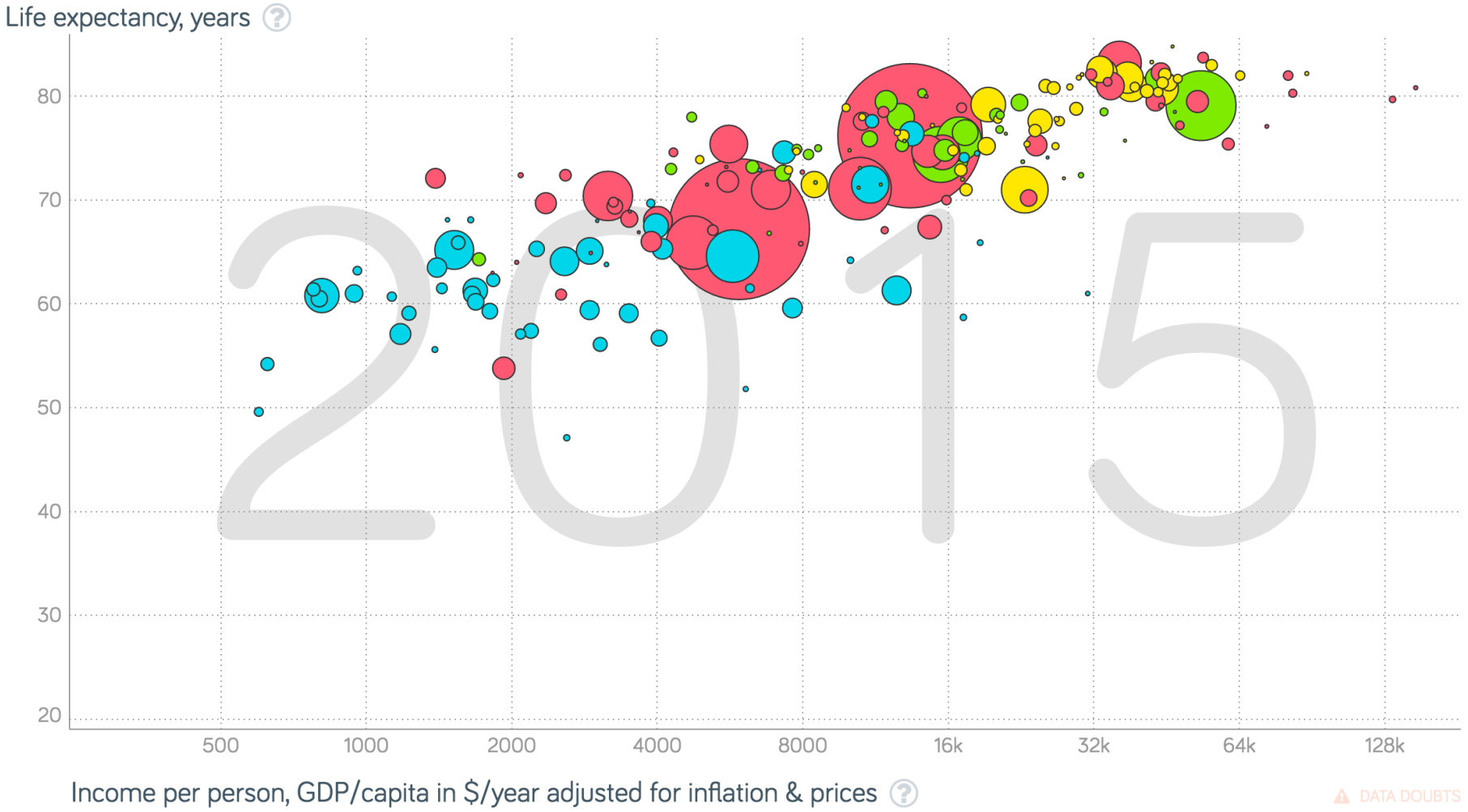
"When a shaded region is used to represent a numerical value, the area of that shaded region should be directly proportional to the corresponding value"

Extends the argument for misleading axes


TOP 10
MOST READ BOOKS
IN THE WORLD
Bases on number of books printed an sold over the last 50 years. Some titles may have had more copies printed than some of these books, but a vast number of those books were not sold, so we'll assume that the did not get read.
In Millions of Copies sold in last 50 years
4000
3500
3000
2500
2000
1500
1000
500
0
3,900
THE HOLY BIBLE
820
QUOTATIONS FROM CHAIRMAN MAO TSE-TUNG
毛澤東 MAO TSE-TUNG
400
Harry Potter
J.K Rowling
103
THE LORD OF THE RINGS
J.R.R TOLKIEN
65
THE ALCHEMIST
PAULO COHELO
57
THE DA VINCI CODE
DAN BROWN
43
the twilight saga
STEPHENIE MEYER
33
GONE WITH THE WIND
MARGARET MITCHEL
30
THINK AND GROW RICH
NAPOLEON HILL
27
THE DIARY OF ANNE FRANK
ANNE FRANK

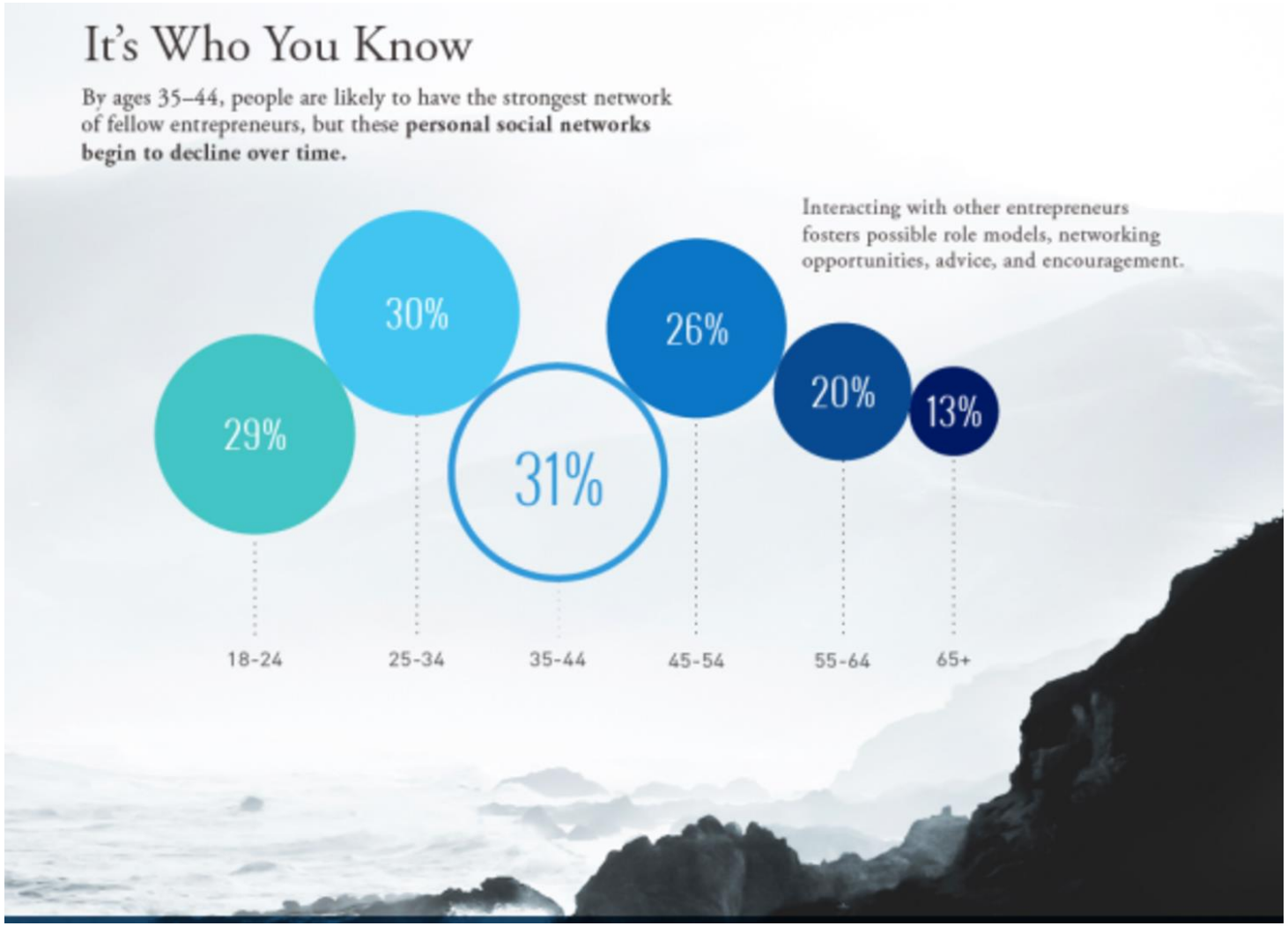

SCOTTISH PARLIAMENT
VOTING INTENTIONS
SNP
46%
CON
24%
LAB
20%
LIB DEM
6%
GREEN
4%
SOURCE: OPINIUM / SKY NEWS

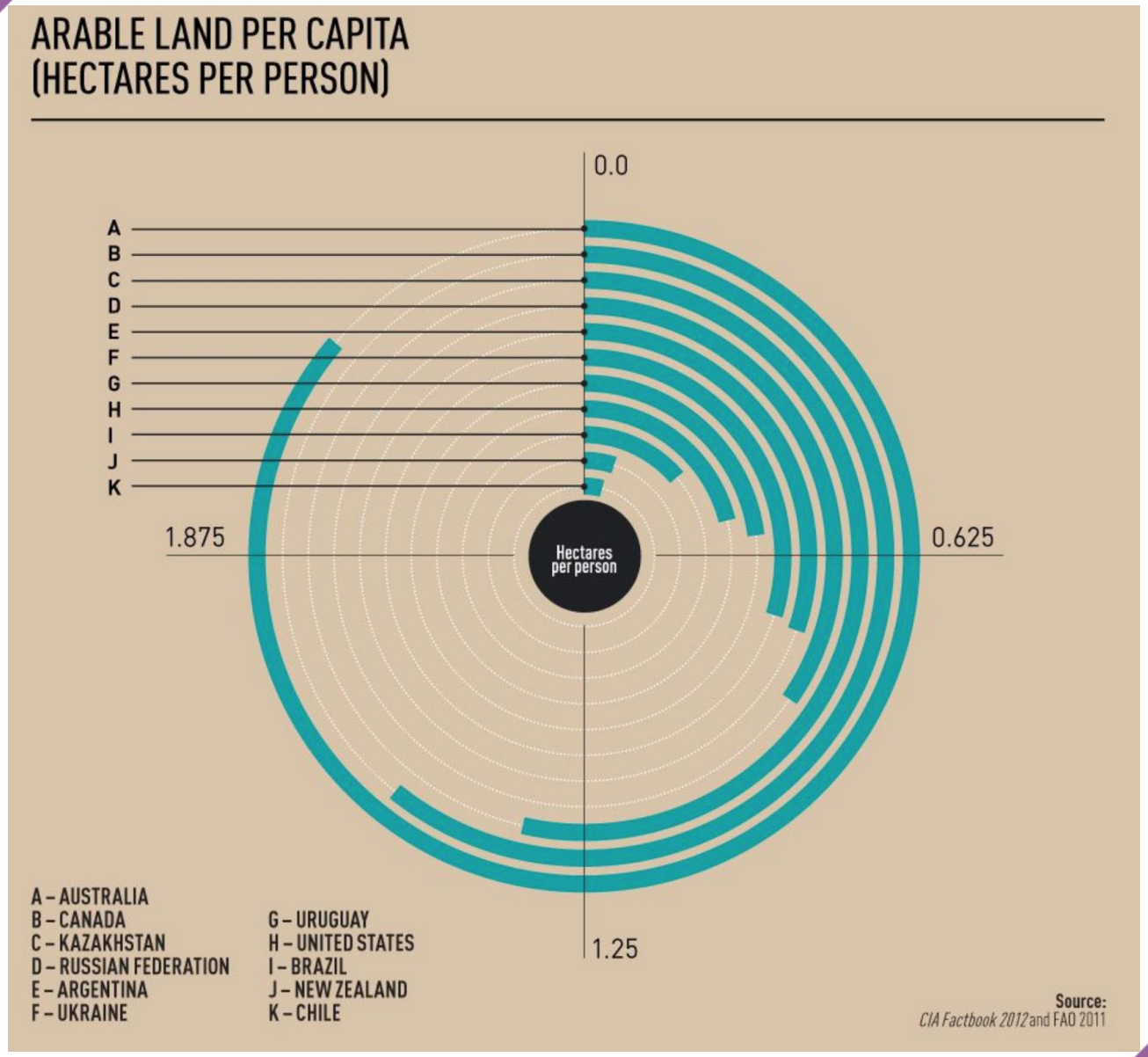

37
36
35
34
33
32
31
30
29
28
Tax Rate
effective federal tax rate for top 1% of households, by income
year 1993
1994
1995
1996
1997
1997
1999
2000
2001
2002
2003
2004
2005
2006
Year

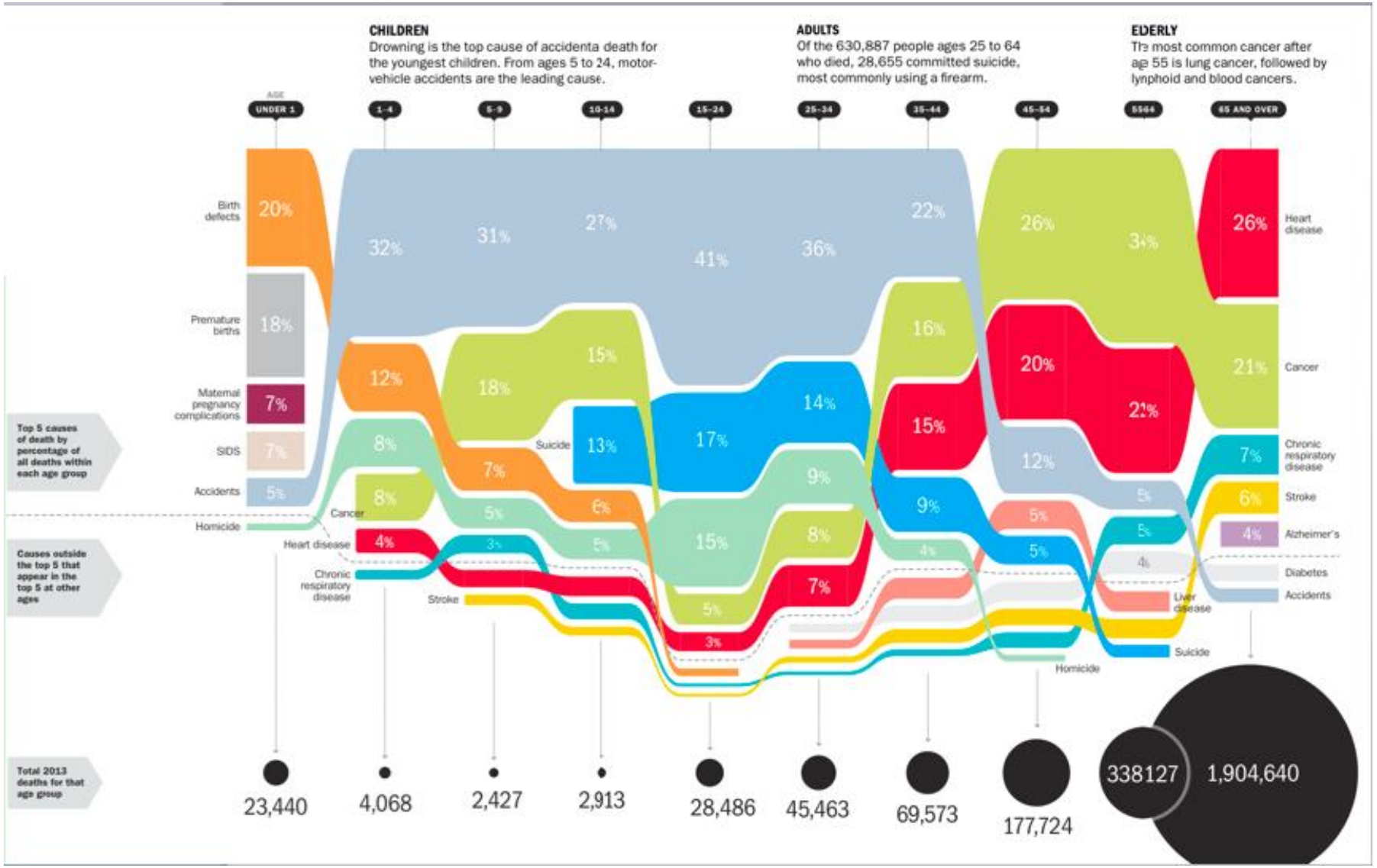


Should we violate the principle?

Radius ≠ Area

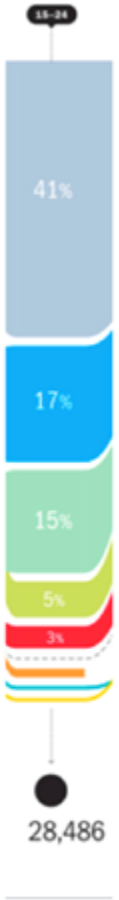

It's Who You Know
By ages 35–44, people are likely to have the strongest network of fellow entrepreneurs, but these personal social networks begin to decline over time.
Interacting with other entrepreneurs fosters possible role models, networking opportunities, advice, and encouragement.
29%
30%
31%
26%
20%
13%
18-24
25-34
35-44
45-54
55-64
65+


ARABLE LAND PER CAPITA
(HECTARES PER PERSON)
0.0
A
B
C
D
E
F
G
H
I
J
K
1.875
Hectares per person
0.625
1.25
A – AUSTRALIA
B – CANADA
C – KAZAKHSTAN
D – RUSSIAN FEDERATION
E – ARGENTINA
F – UKRAINE
G – URUGUAY
H – UNITED STATES
I – BRAZIL
J – NEW ZEALAND
K – CHILE
Source:
CIA Factbook 2012 and FAO 2011

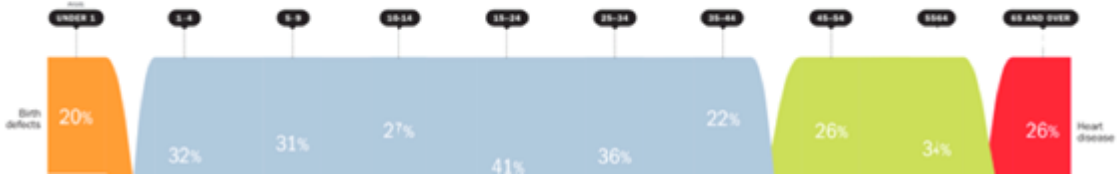


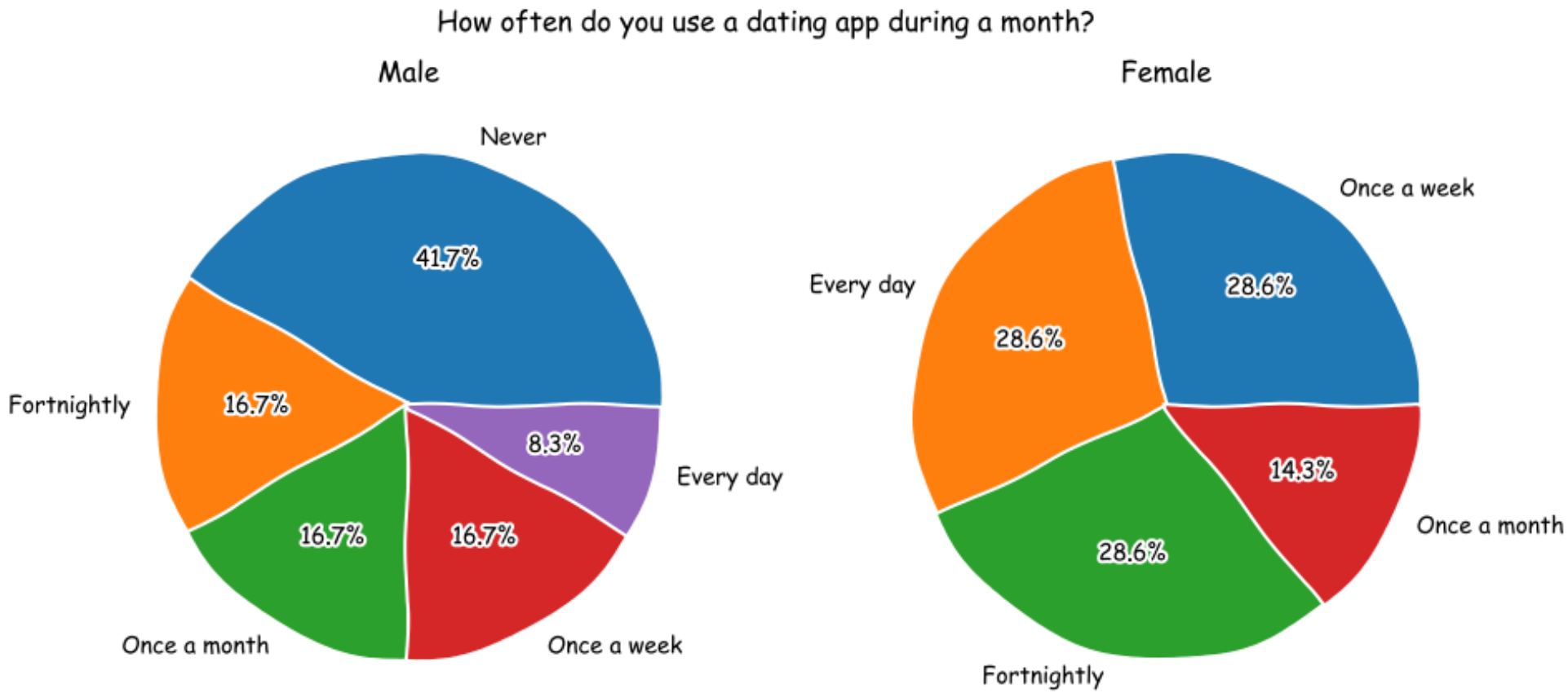
https://time.com/3935281/what-are-my-health-risk-factors/

Along vertical slices, ink is proportional to value because shaded areas represent the fraction of a fixed number of deaths (here 28,486) from each cause.



Along horizontal slices, ink is not proportional because total deaths differ widely by age group. Far more people 65 and older died of heart disease (red) than children age1-4 die of accidents (blue-gray), but the latter takes more ink because it represents a larger percentage of the (relatively few) total deaths at that age.





What is wrong in this one?

9. Perspective

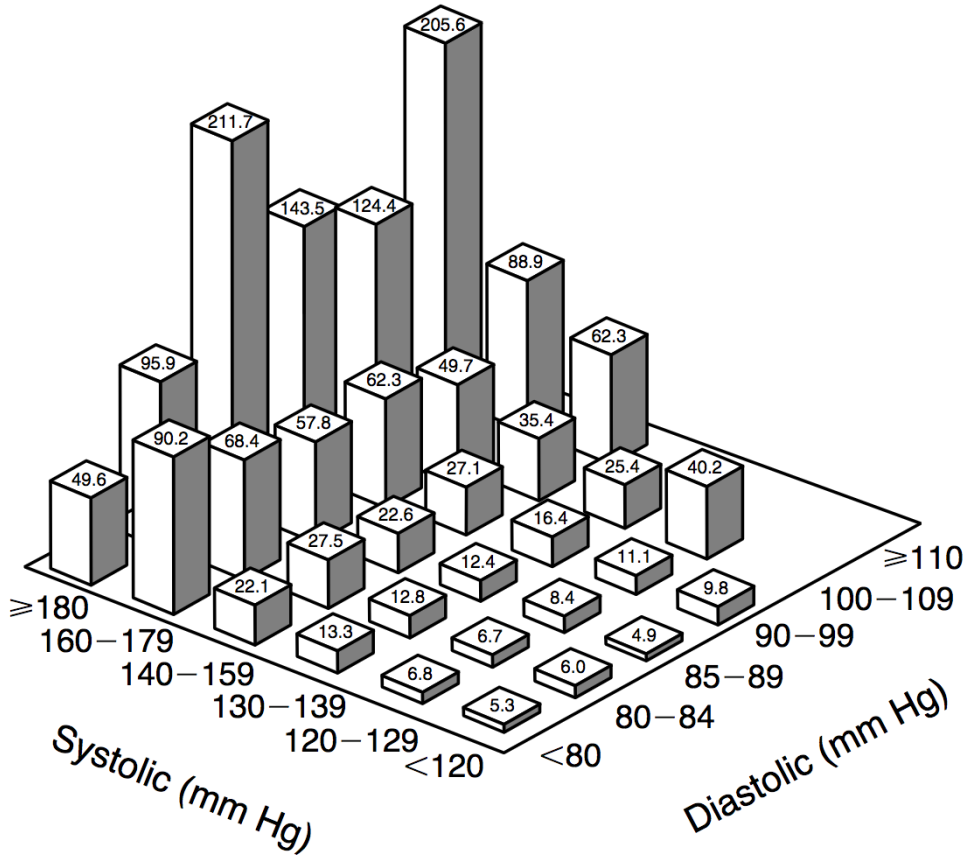
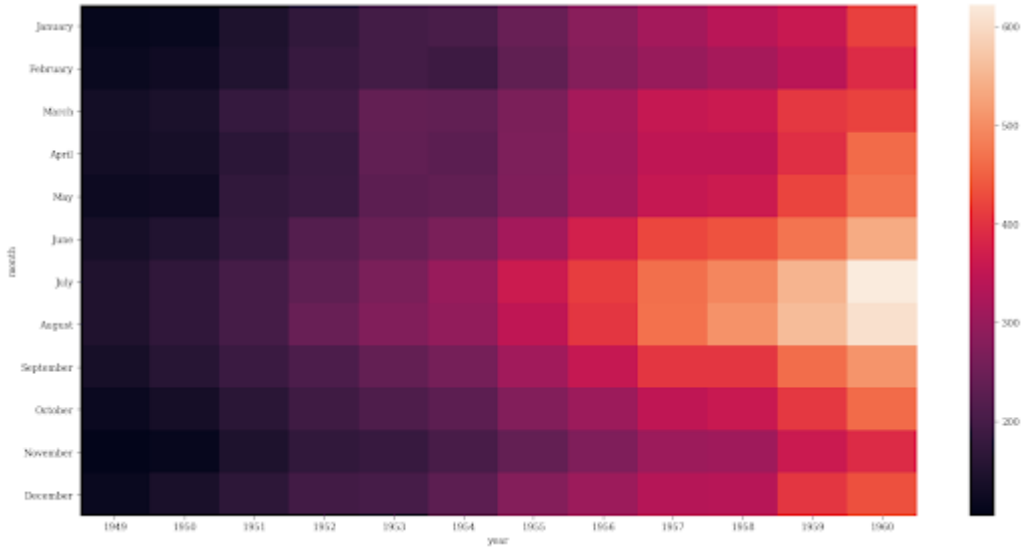
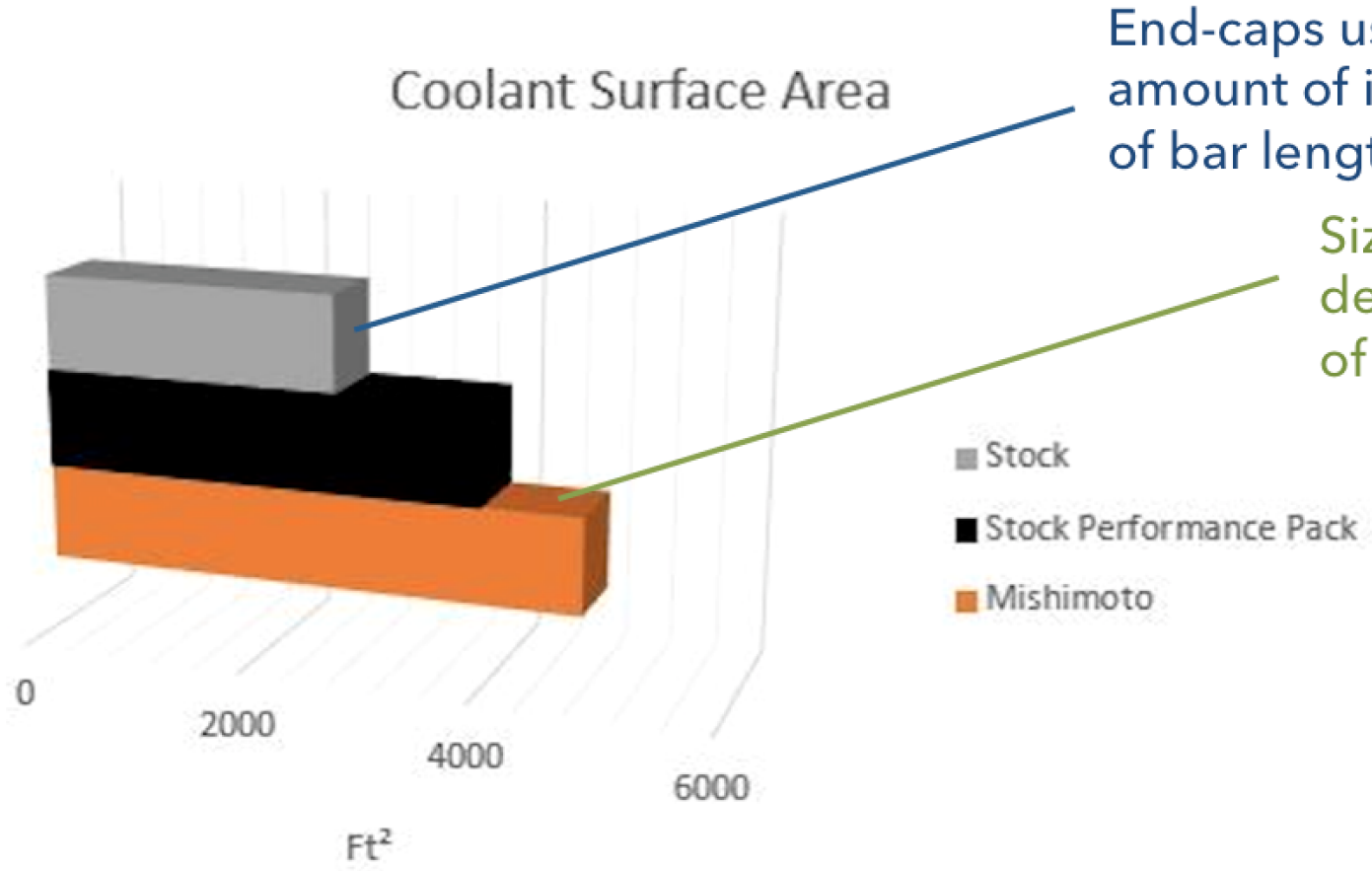


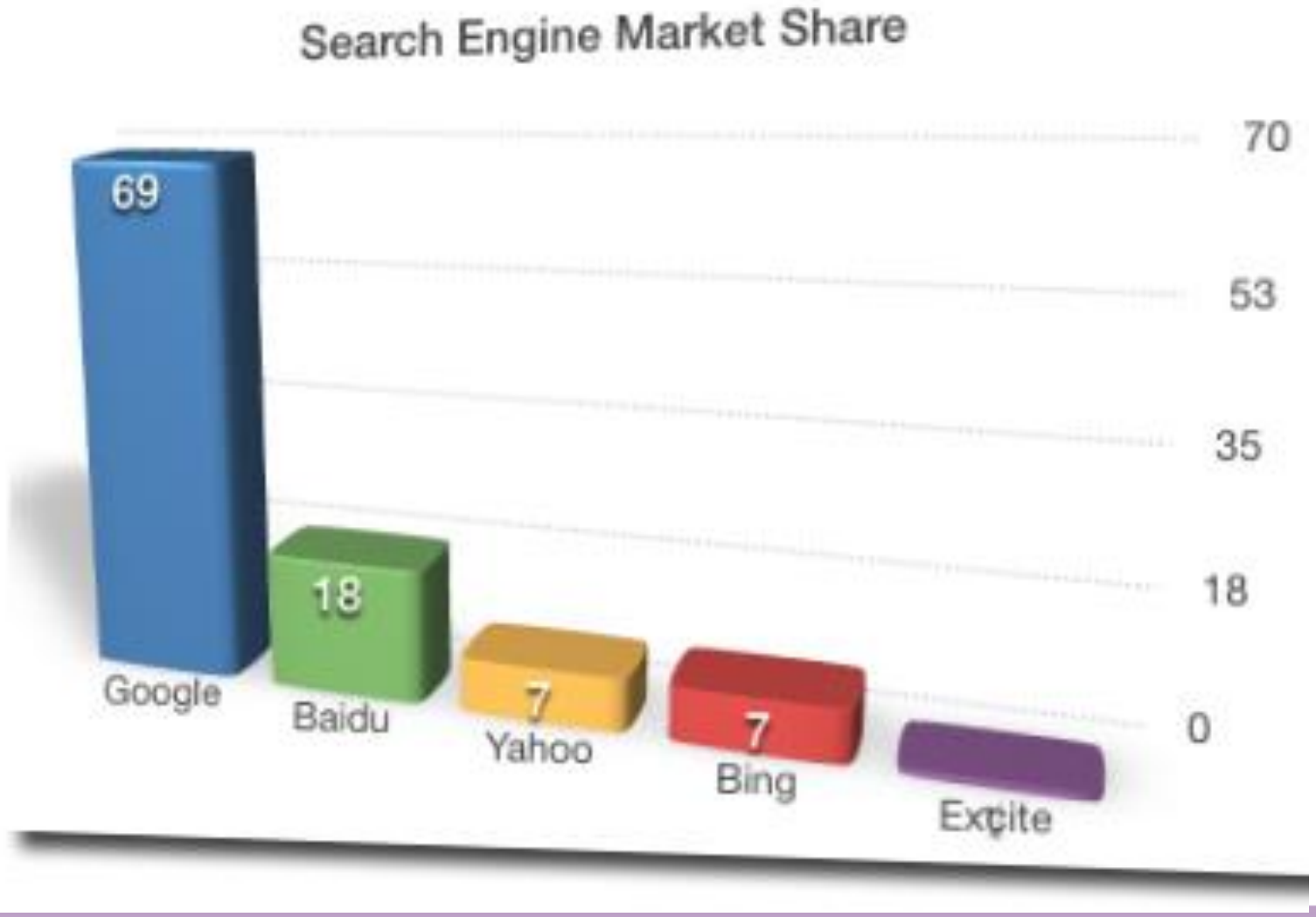
Figure 2. Age-Adjusted Rate of End-Stage Renal Disease Due to Any Cause per 100,000 Person-Years, According to Systolic and Diastolic Blood Pressure in 332,544 Men Screened for MRFIT.

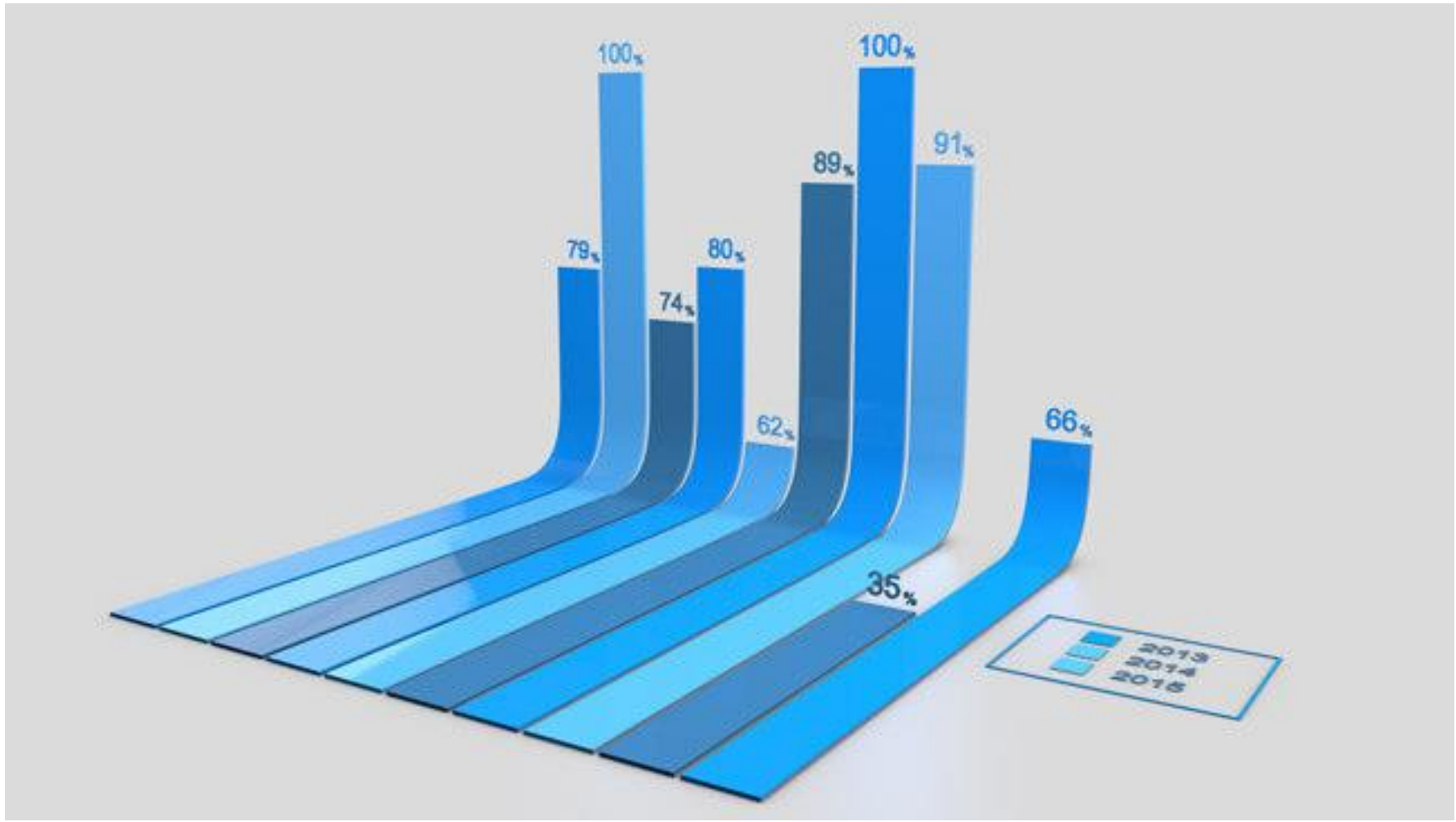


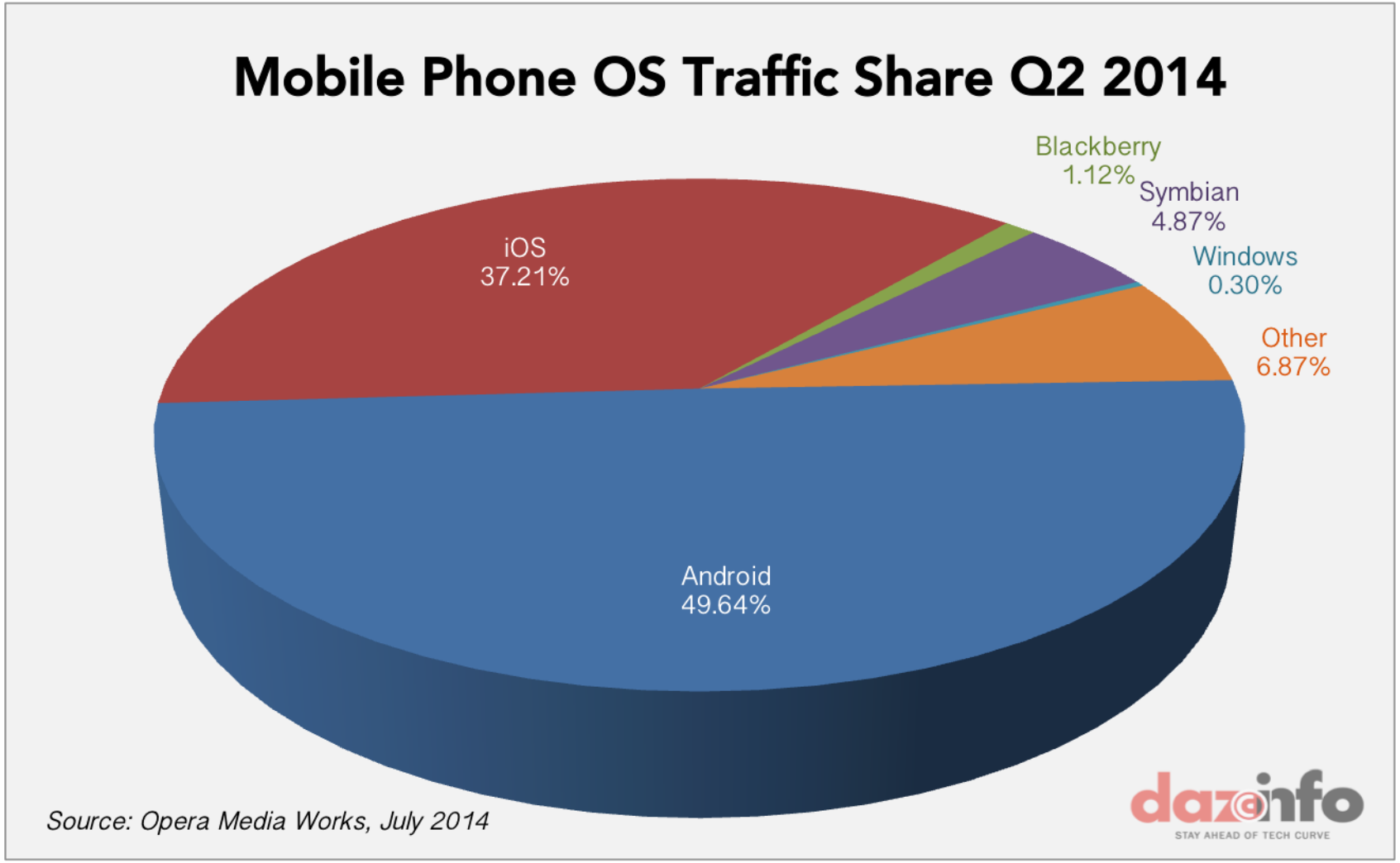


End-caps use the same amount of ink irrespective of bar length.

Size of this upper face depends on the length of the black bar.


Search Engine Market Share
70
53
35
18
0
69
18
7
7
Google
Baidu
Yahoo
Bing
Excite


100%
100%
91%
89%
79%
80%
74%
62%
66%
35%
2013
2014
2015


Mobile Phone OS Traffic Share Q2 2014
Blackberry
1.12%
Symbian
4.87%
Windows
0.30%
Other
6.87%
iOS
37.21%
Android
49.64%
Source: Opera Media Works, July 2014
dazeinfo
STAY AHEAD OF TECH CURVE

Is there a way to choose the "right" plot?

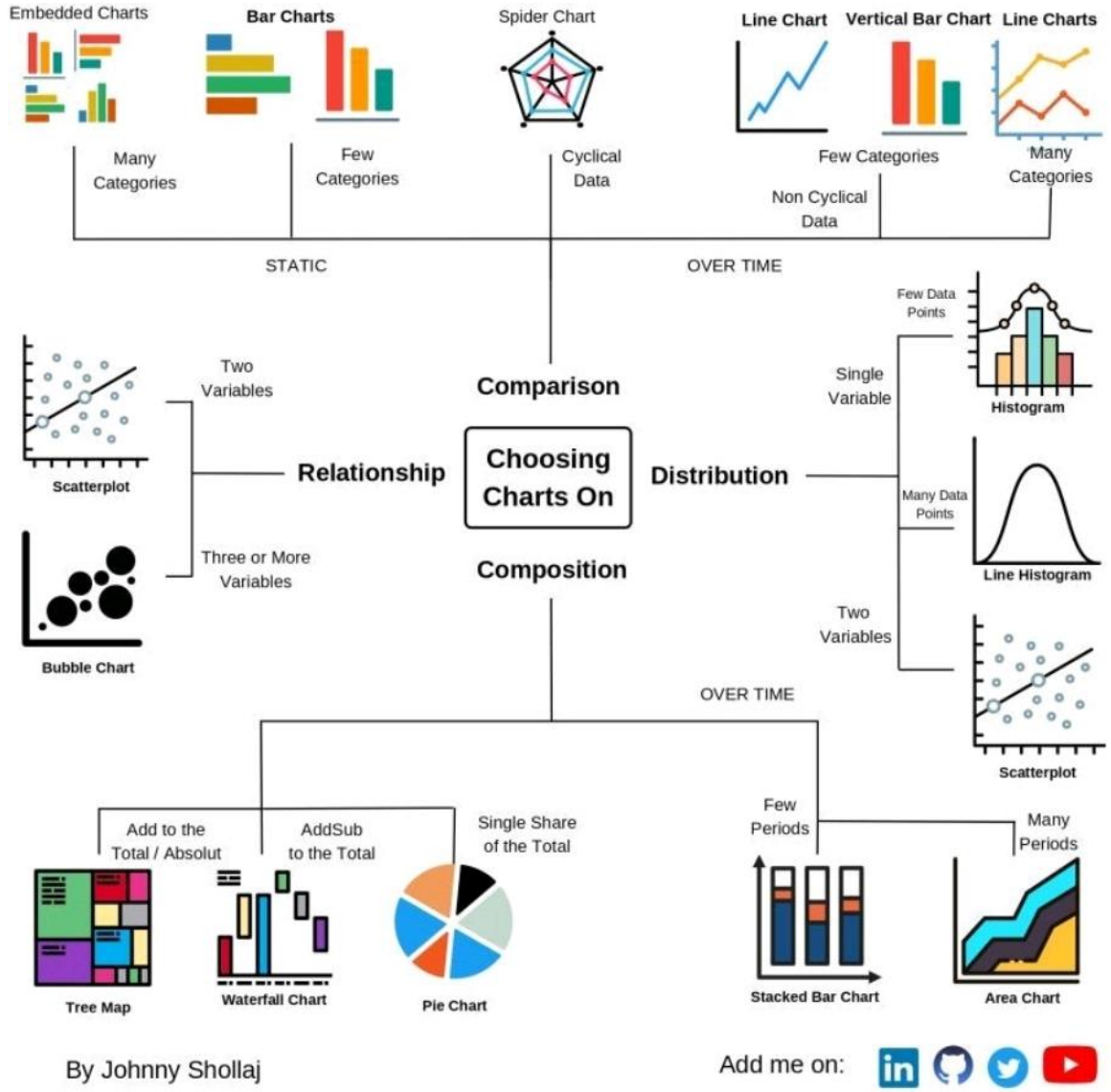
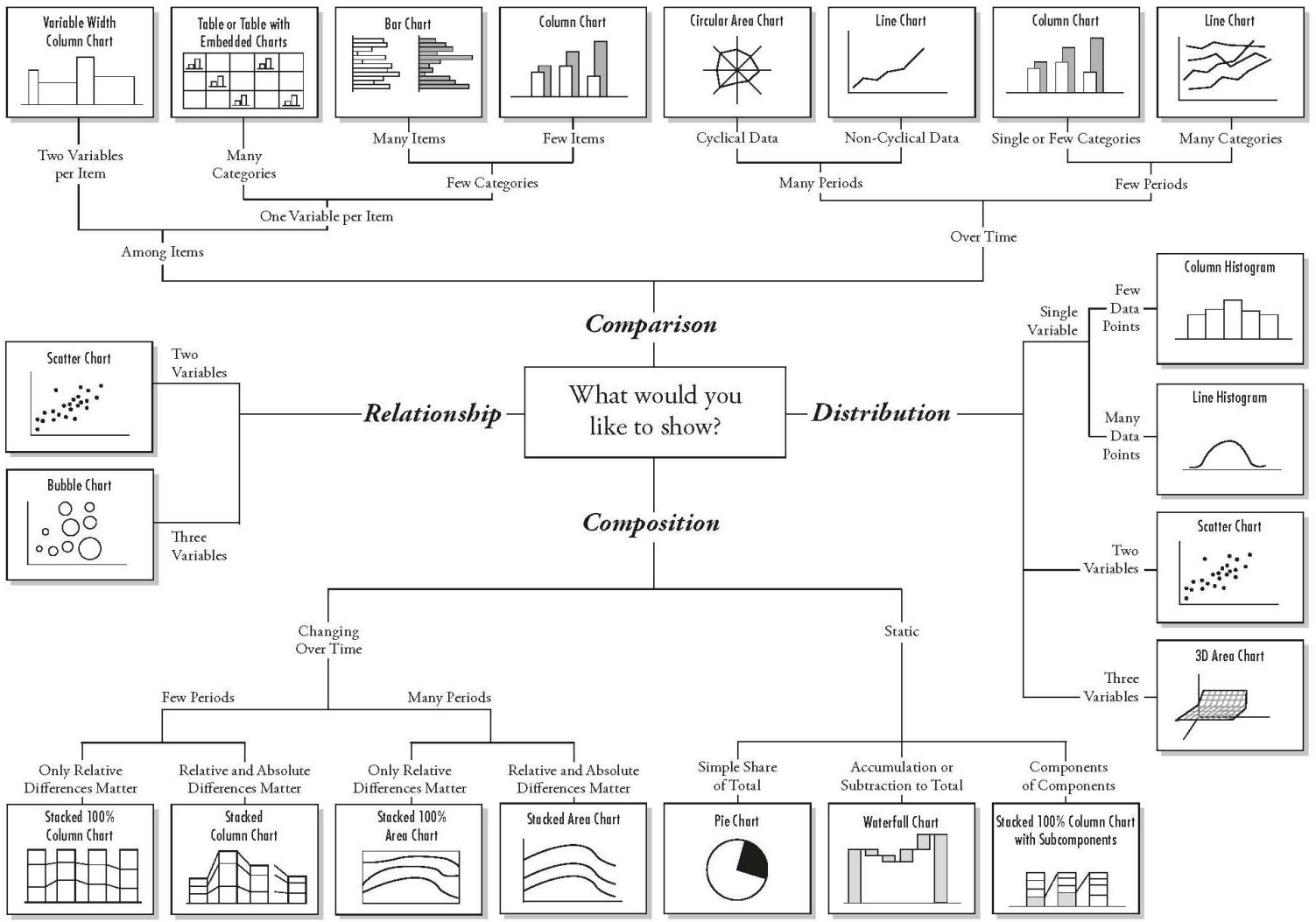

Embedded Charts
Bar Charts
Spider Chart
Line Chart
Vertical Bar Chart
Line Charts
Many Categories
Few Categories
Cyclical Data
Few Categories
Many Categories
Non Cyclical Data
STATIC
OVER TIME
Few Data Points
Histogram
Two Variables
Comparison
Single Variable
Scatterplot
Relationship
Choosing Charts On
Distribution
Many Data Points
Line Histogram
Three or More Variables
Composition
Two Variables
Bubble Chart
OVER TIME
Scatterplot
Add to the Total / Absolut
AddSub to the Total
Single Share of the Total
Few Periods
Many Periods
Tree Map
Waterfall Chart
Pie Chart
Stacked Bar Chart
Area Chart
By Johnny Shollaj
Add me on:

Chart Suggestions—A Thought-Starter



https://extremepresentation.com/design/7-charts/

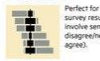
Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

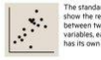
Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

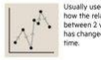
Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

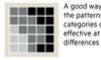
Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

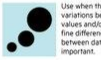
Example FT uses
Wealth, deprivation, league tables, constituency election results

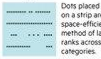
Ordered bar

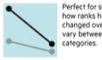
Standard bar charts display the ranks of values much more easily when sorted into order.

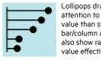
Ordered column

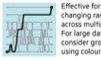
See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

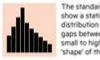
Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

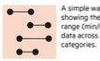
Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

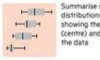
Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

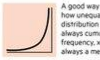
Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

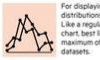
Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

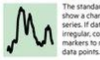
Beeswarm

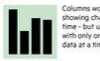
Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time – but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

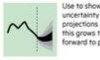
Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

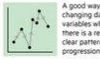
Area chart

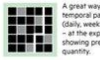
Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

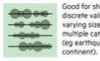
Fan chart (projections)

Use to show the uncertainty in future projections - usually this grows the further forward to projection.

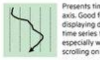
Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

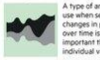
Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline

Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Streamgraph

A type of area chart; use when seeing changes in proportions over time is more important than individual values

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger/ or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space-efficient way of showing value of multiple variables– but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet

Good for showing a measurement against the context of a target or performance range.

Grouped symbol

An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

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

FT

If there's time…

Check some examples from last lab