Data Visualisation and Dashboarding

UNIVERSITY OF WESTMINSTER

www.westminster.ac.uk

Welcome

Weekly lecture covers theory of data communication

Tutorial: Creating visualisations in R, Power BI and other tools

Assessment: ICT 30%, CW 70%

Rolf Bänziger
r.banziger@westminster.ac.uk

Dr Deepika
dd@westminster.ac.uk

Edmund Gall
e.gall@westminster.ac.uk

What will we learn?

Why data visualisation is important

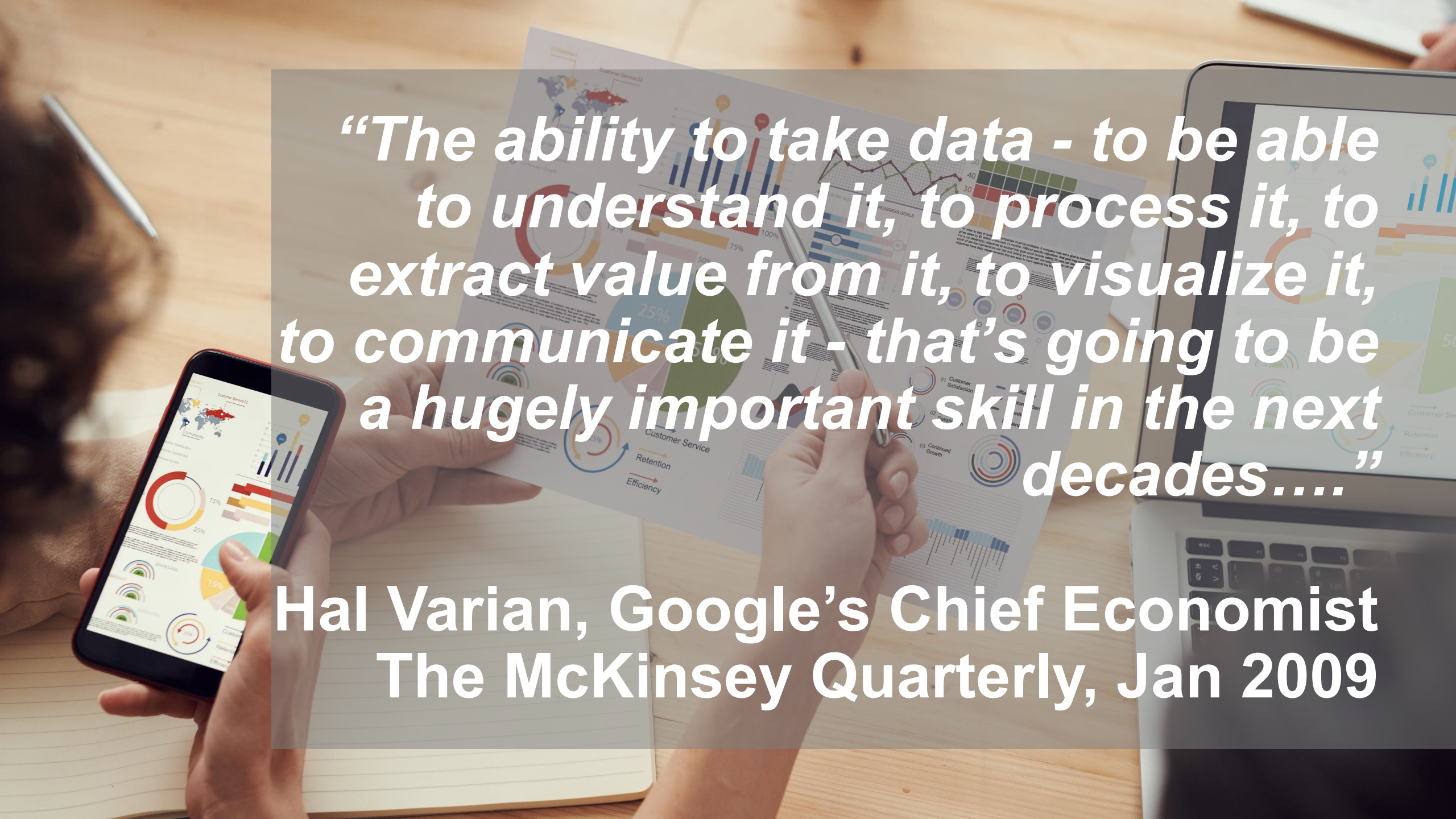
What makes a good data visualisation?

Tools to explore and create quality data visualisations

Evaluate and critique data graphics

How to communicate data insights

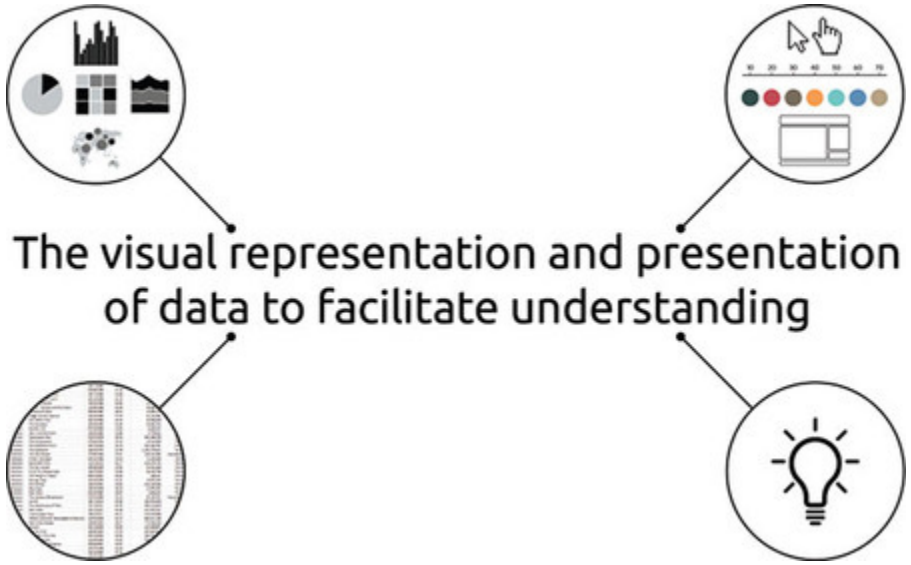
Design rich, interactive dashboards


“The ability to take data - to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it - that’s going to be a hugely important skill in the next decades….”
Hal Varian, Google’s Chief Economist
The McKinsey Quarterly, Jan 2009

What is data visualisation?

- Data visualization is multidisciplinary with aspects of science, usability and art.
  - The important point is that there is a need to ensure that the balance of all aspects is correct, that for example we do not focus on the aesthetics at the risk of the clarity of data.
- A data visualization first and foremost has to accurately convey the data. It must not mislead or distort which is often done on purpose.
- At the same time, it needs to be aesthetically pleasing.
- Data visualization is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying trends and clusters, spotting local patterns, evaluating modelling output, and presenting results.

Defining Data Visualisation



Kirk, A. (2019) Data visualisation : a handbook for data driven design. 2nd edition. Los Angeles: SAGE.

Data visualization timeline

1644
Flemish astronomer Michael Florent van Langren provides the first representation of statistical data.

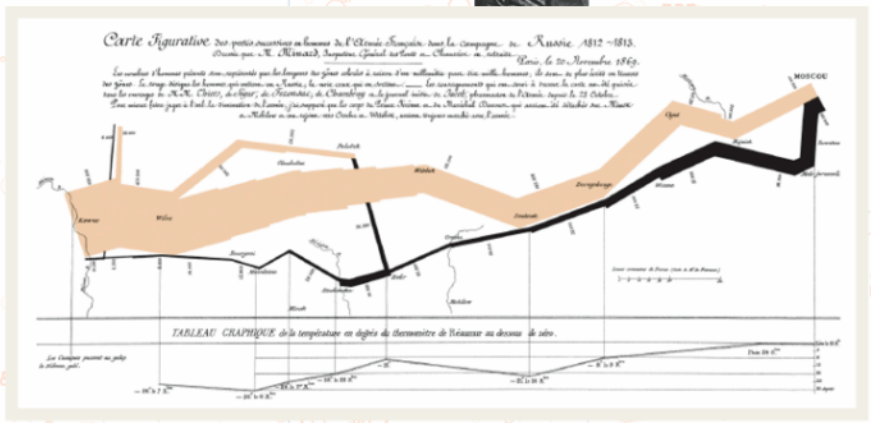
1600s

1700s
Thematic mapping emerged and abstract graphs of functions, measurement errors and the collection of empirical data were introduced.

1700s

▲ 1800s
William Playfair, among others, introduced some of today's most popular graphs and various statistical chart types were invented.

◄ 1854
Physician John Snow maps the outbreaks of cholera that occurred across London during the 1854 epidemic.



▲ 1869
Charles Joseph Minard charts the number of men in Napoleon's 1812 Russian army.

PIE CHART FROM WILLIAM PLAYFAIR'S "STATISTICAL BREVIARY"

1800s

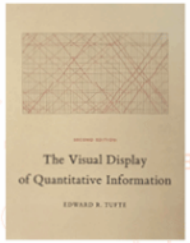
1900s

Early 1900s
Statisticians are less concerned with data visualization and more focused on exact numbers. Simultaneously, data visualization gains public popularity, and charts and graphs start appearing in textbooks and business applications.

Late 1900s
The emergence of computer processing allows statisticians to collect, store and efficiently visualize larger volumes of data.

1960s-1970s
Researchers John W. Tukey and Jacques Bertin develop the science of data visualization in statistics and cartography, respectively.

▼ Early 1980s
Edward Tufte publishes *The Visual Display of Quantitative Information*, which is currently used in university courses.



Purpose of data visualisation

To explore and analyse.

Get closer to data.

Discover patterns

Enable decisions

Visualisations are very helpful in communicating contextual information to a variety of stakeholders and interested parties.

Of course, a bad visualisation can distort this discovery and communication.

Explore and analyse data

Anscombe quartet

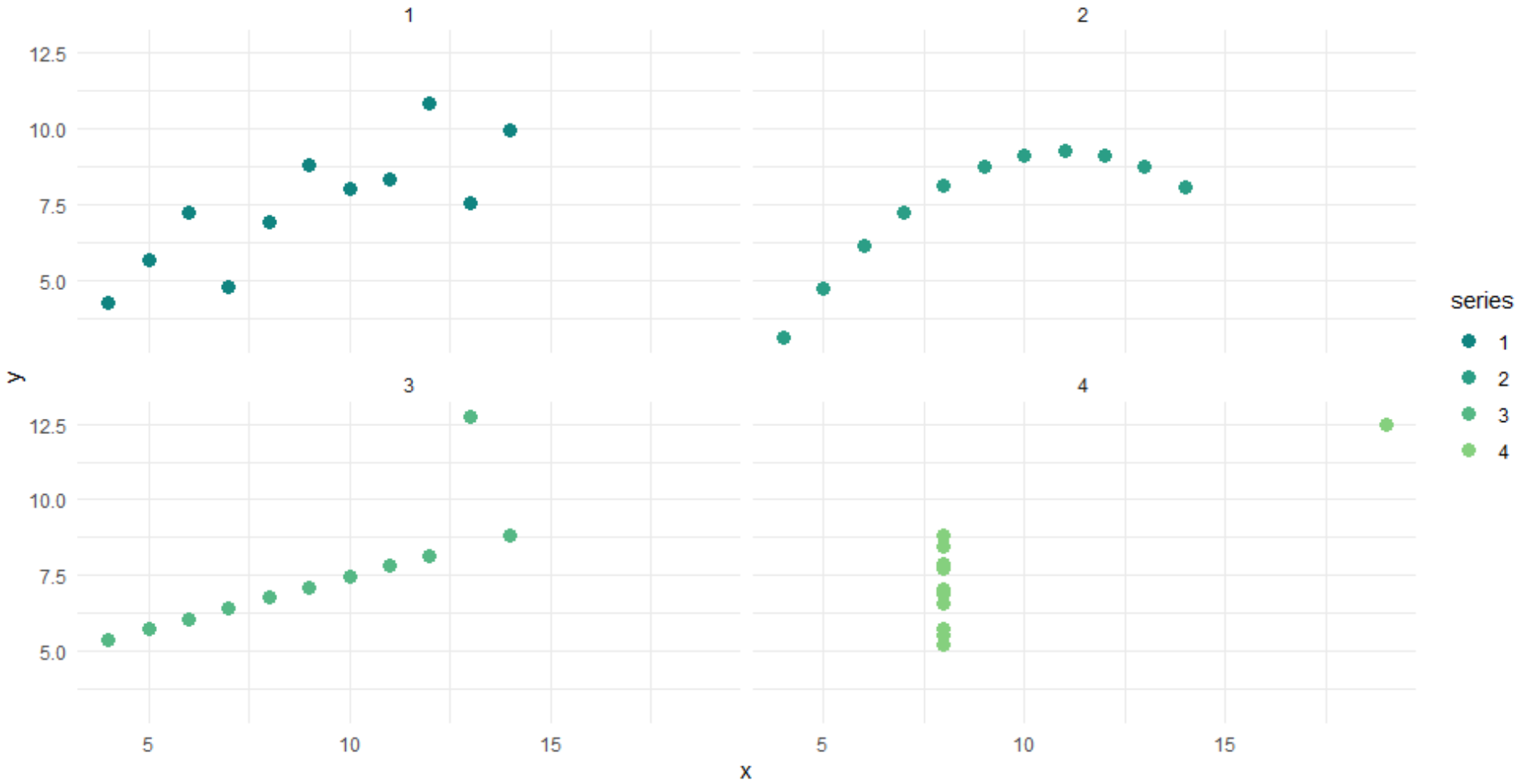
Four series (x_n and y_n)

Same or similar average, standard deviation, variance, etc.

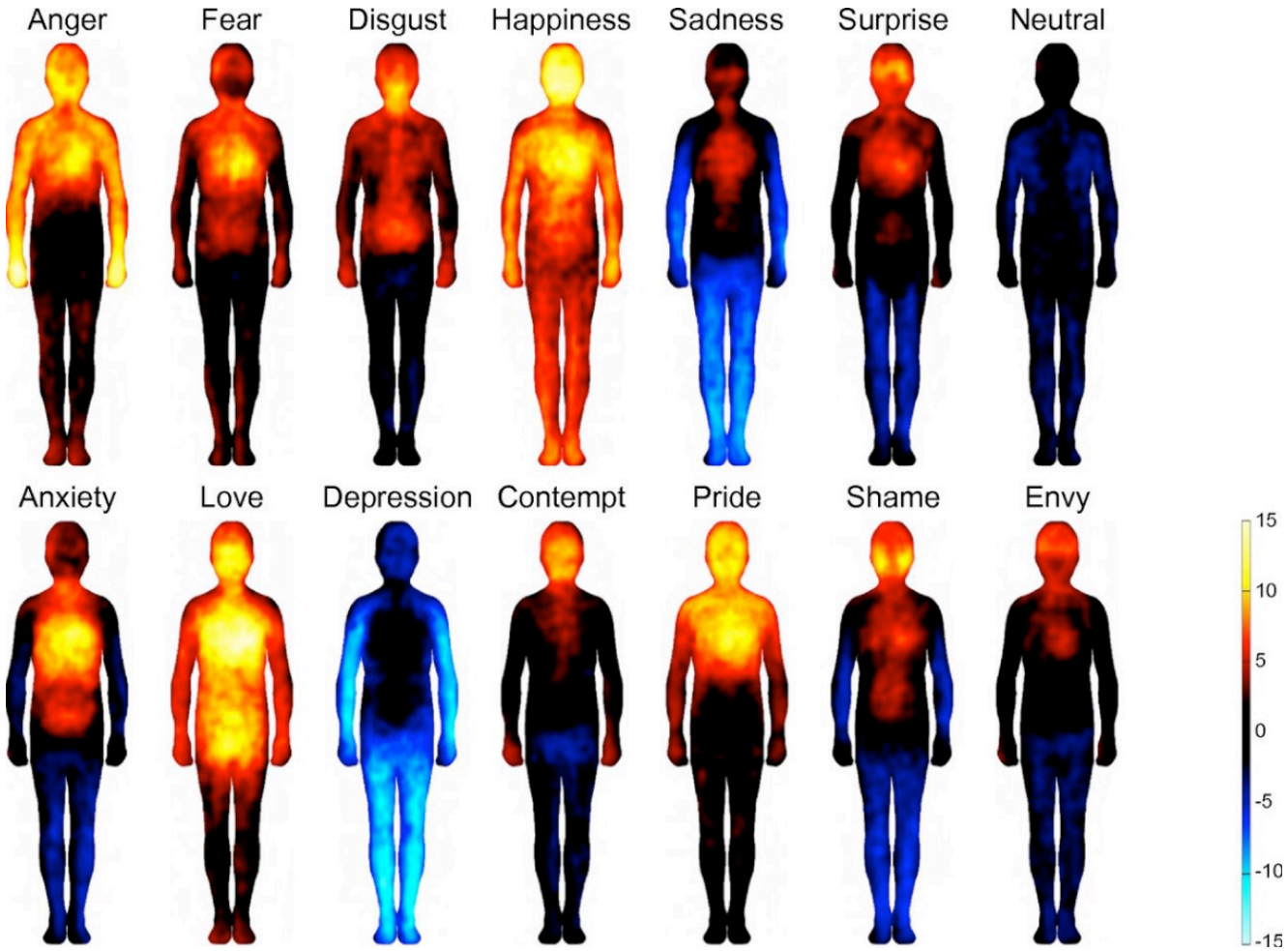
Same data?

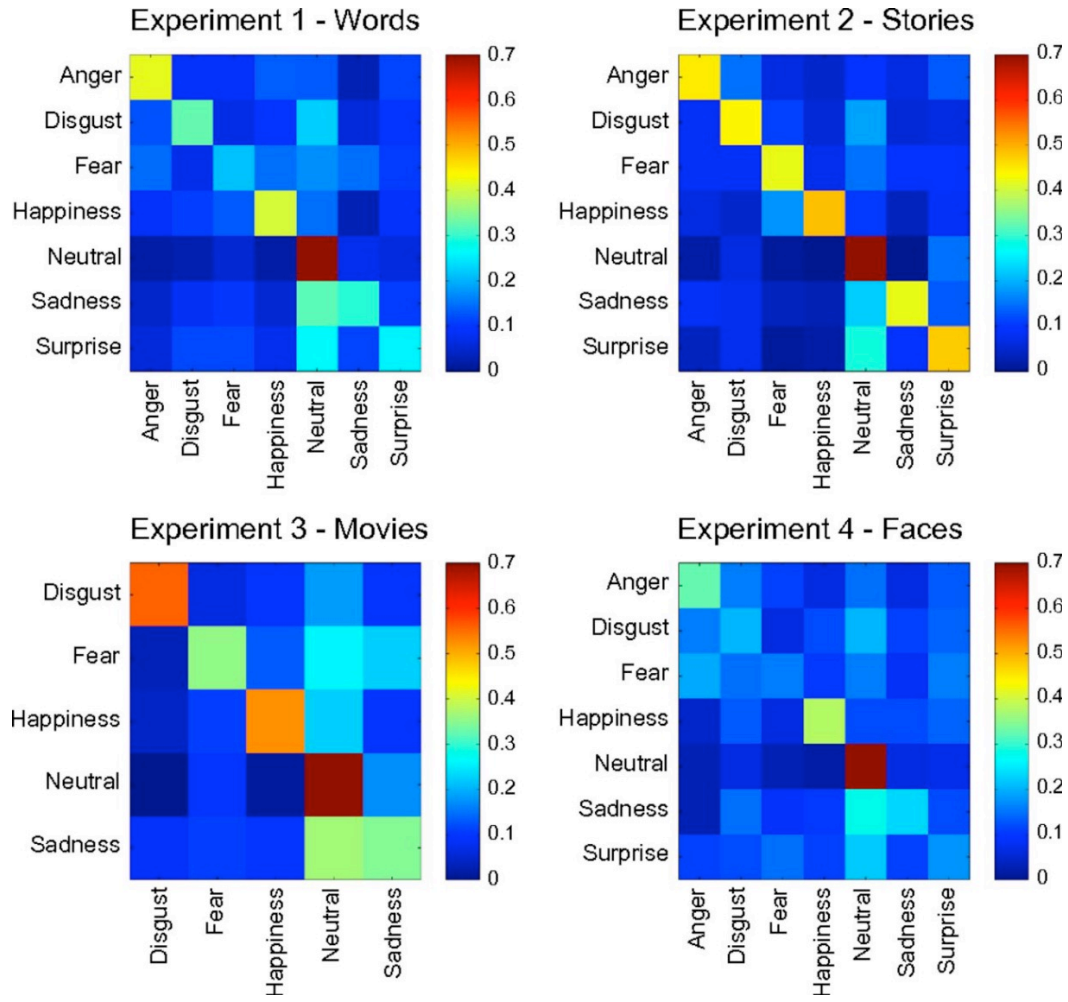
	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.1	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.1	5.39	12.5
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89
Avg	**9**	**9**	**9**	**9**	**7.501**	**7.501**	**7.501**	**7.501**
Sd	**3.32**	**3.32**	**3.32**	**3.32**	**2.03**	**2.03**	**2.03**	**2.03**
Variance	**11**	**11**	**11**	**11**	**4.1**	**4.1**	**4.1**	**4.1**

Anscombe’s quartet



Closer to data - Bodily maps of emotions





UNIVERSITY OF WESTMINSTER

Closer to data - Bodily maps of emotions

Study (2013) of physical responses to emotional cues based on 700 participants.

Participants were shown two body outlines and asked to note areas of their body that they felt activity was increasing or decreasing based on the emotion they were focusing on.

Summarising the results as a spatial density heatmap, we can clearly see how the spatial data tells a clear story for each emotion.

Nummenmaa *et al* https://www.pnas.org/content/111/2/646

Understanding influence

‘On the mode of the communication of cholera’ John Snow, 1854

(a redrawing appearing in ‘The Visual Display of Quantitative Information’ Edward R Tufte)

Understanding influence

August 1854 –Cholera outbreak in Soho, London –one of many such outbreaks in the city at the time caused by population density combined with poor sanitation and sewage.

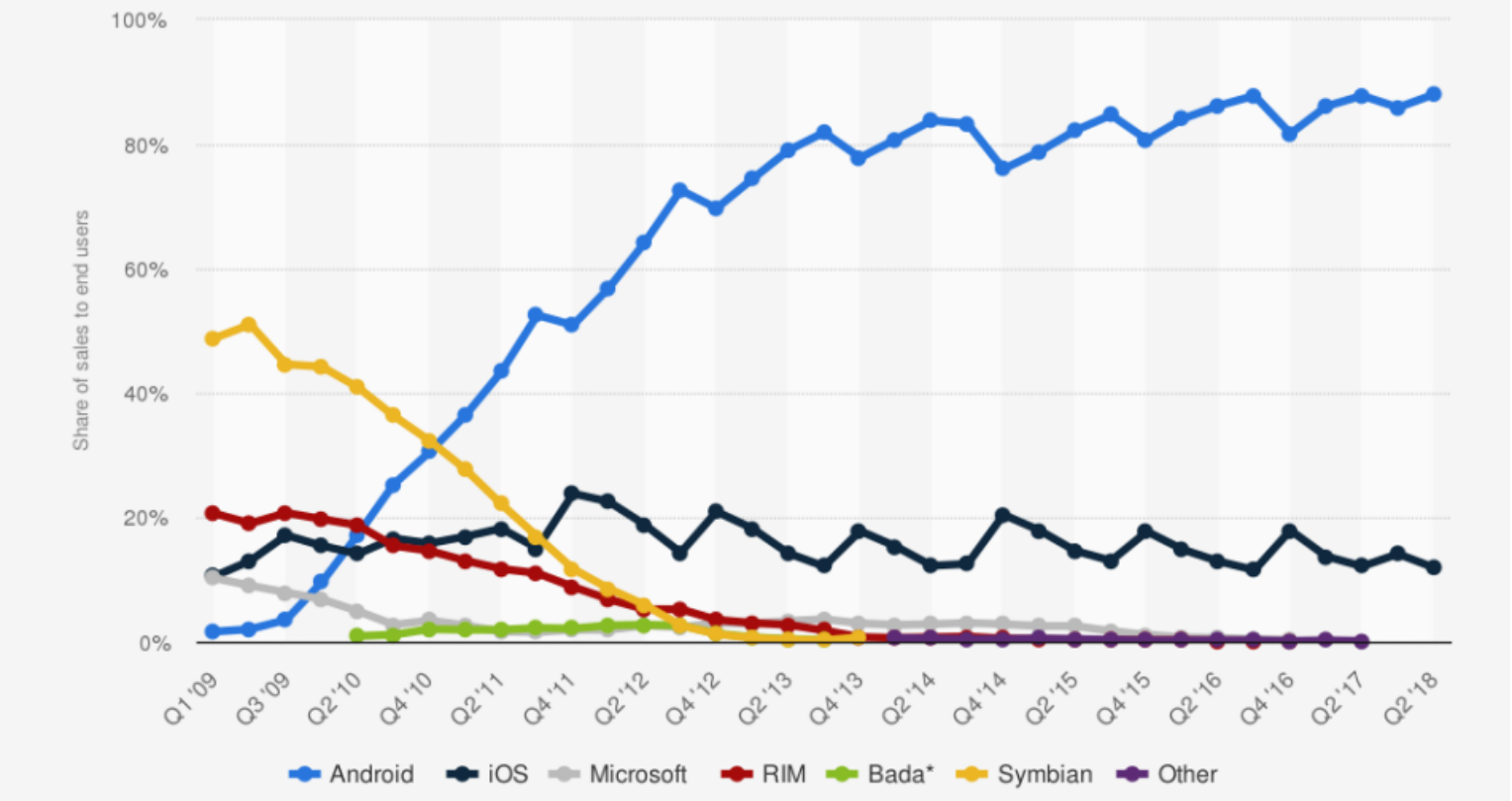
Over 1 in 10 residents being killed by cholera.

Dr John Snow, in conjunction with others, studied those affected by cholera to try to determine a source. The spread of cholera was not clearly understood at the time as ‘germ theory’ would not be described by Pasteur until 1861.

His work identified the Broad Street pump as the probably cause and he successfully argued for the removal of the pump handle to combat the outbreak. Some regard this as the beginning of the science of epidemiology.

This dot map, slightly adapted here, is a clear visualisation of the data and how it relates to the Broad Street Pump.

Discovering patterns



Global market share held by leading smartphone operating systems in sales to end users 1st quarter 2009 to 2nd quarter 2018

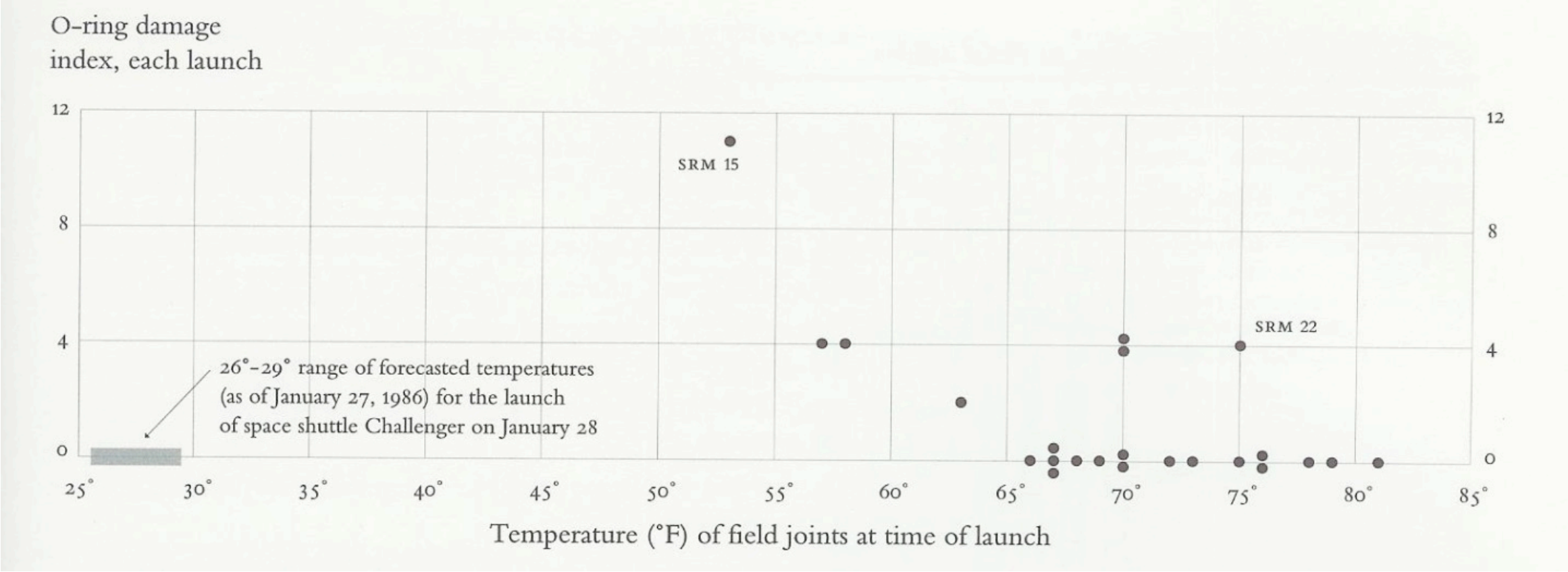
Discovering patterns

Chart shows trends for global mobile operating system market share, in terms of sales to end users from 2009 to 2018

In 2nd quarter of 2018, 88% of smartphones sold to end users were phones with the Android operating system

An app developer may want to build their apps prioritising the Android operating system

Enabling decisions



Enabling decisions

January 1986: In -1°C temperature (31°F), much colder than ever launched before, the NASA space shuttle orbiter Challenger took off and exploded shortly afterwards killing all 7 crew.

The resulting inquiry of the worst disaster in Nasa's history up to then determined the rubber 'O-ring' that sealed the sections of the solid rocket boosters (SRBs) and prevented combustion gases leaking out –had failed.

Engineers working for Morton Thiokol, the company who built the SRBs, had argued that the O-rings were not tested at such low temperatures. In a crucial meeting, they produced two slides -one noting SRBS were untested below 53°F and the second, a table of the conditions that had seen o-ring failures occur. These data points are highlighted by the blue box.

Some have argued, such as Edward Tufte, that by plotting the data points, there was a much stronger chance of persuading NASA management to delay the launch.

However, we cannot say this for certain. The NASA 'go-fever' was a result of many external and internal organisational pressures to launch. What we can say is that by visualising the data, it is much easier to understand the potential risk of launching at low temperatures.

Inspiring actions

This Rose graphic indicates the number of deaths that occured from preventable diseases (in blue), those that were the results of wounds (in red), and those due to other causes (in black).

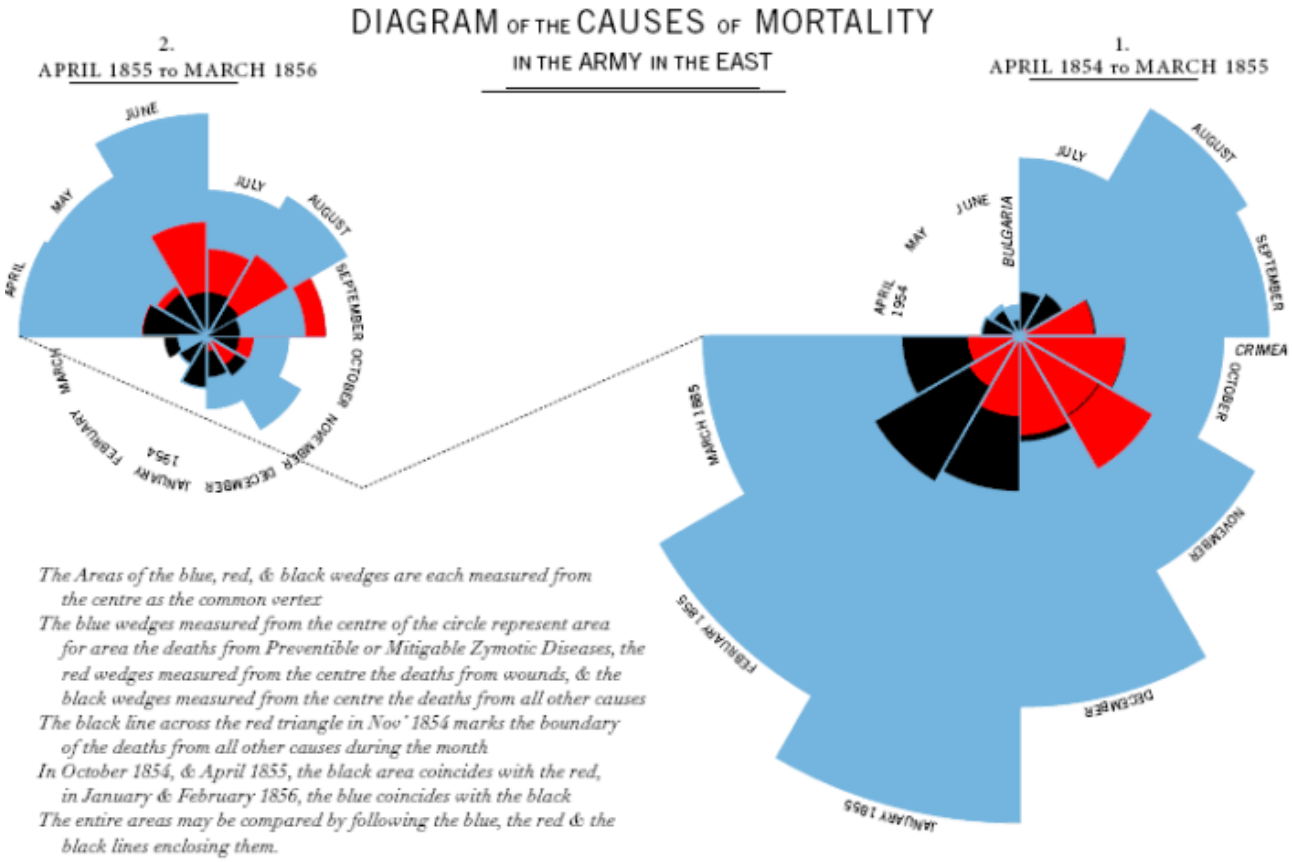
The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.

The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.

The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.

In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black.

The entire areas may be compared by following the blue, the red, & the black lines enclosing them.



Inspiring actions

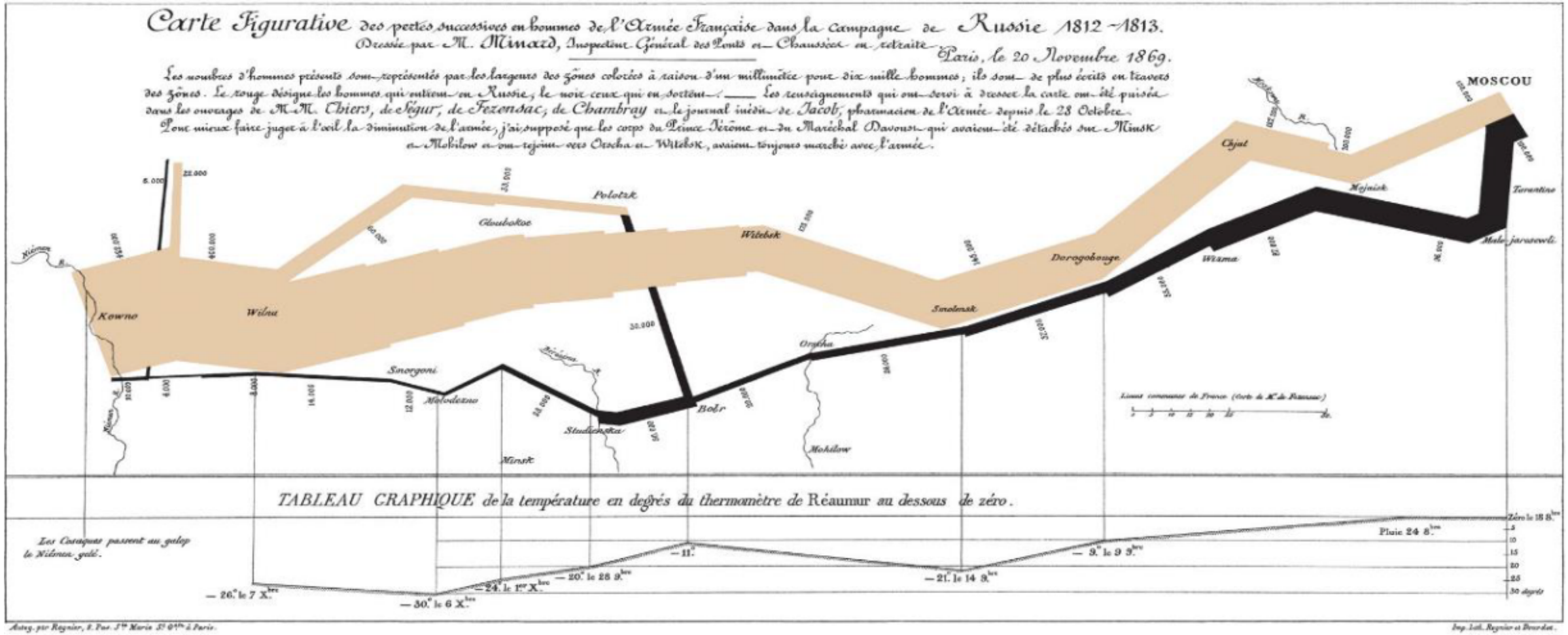
Born 1820, Florence Nightingale was a social reformer and statistician, and the founder of modern nursing.

Shocked by her experiences of British soldiers dying in hospitals during the Crimean War (1953-1856) from preventable diseases, she campaigned for better care.

The Nightingale ‘Rose’ is an example of her published work to visually demonstrate to non-statisticians the statistics behind soldier deaths.

Her work influenced governments to improve their military hospital care –in particular the reduction of death from preventable diseases.

Contextualise information



Contextualise information

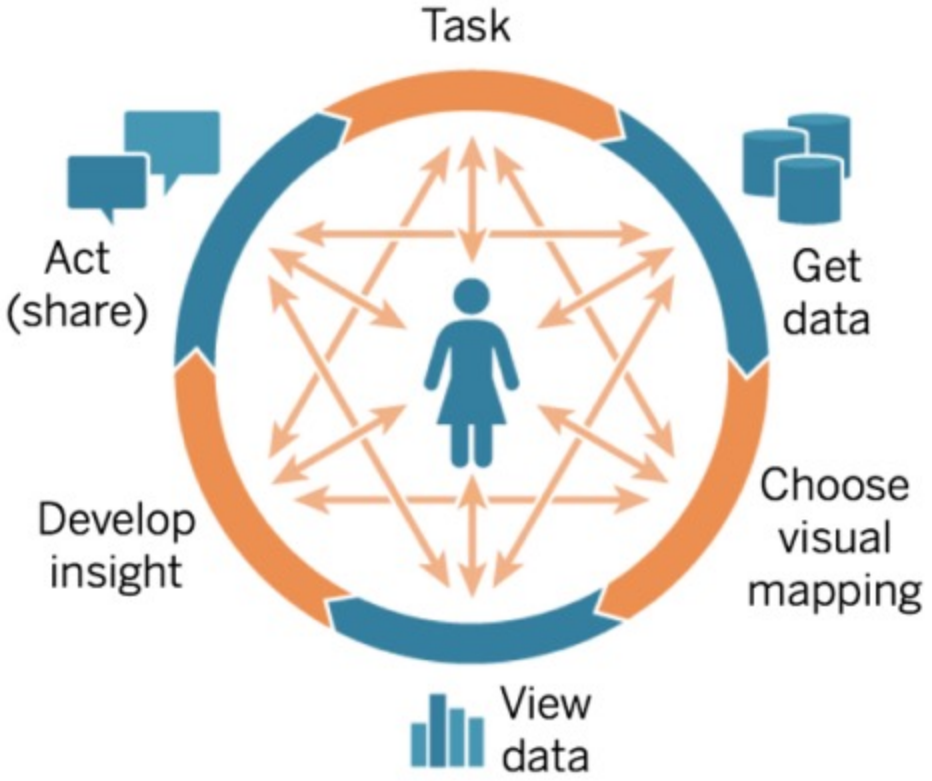
This visualisation shows a flow map of the disastrous 1812-13 Russian Campaign by Napoleon Bonaparte.

Drawn by Charles Minard (published in 1869) who was a civil engineer and pioneer of the use of graphics in engineering and statistics.

Edward Tufte ranks this as the best statistical graphic ever produced. But this is arguable….

What it does very well is encode many variables into a single graphic. It shows the size of the army, the geographic position, the direction travelled, how groups broke off on certain dates and lastly it depicts the very cold weather of the retreat.

Making sense of data – the basic visualisation process



From [The Cycle of Visual Analysis – Tableau]

Challenges to overcome

Increasing data volumes

Data preparations

Understanding the expected data characteristics

Choosing the right tool and technique

Correctly using the tool

Precisely presenting data

Making the right interpretations

The good, the bad and the ugly

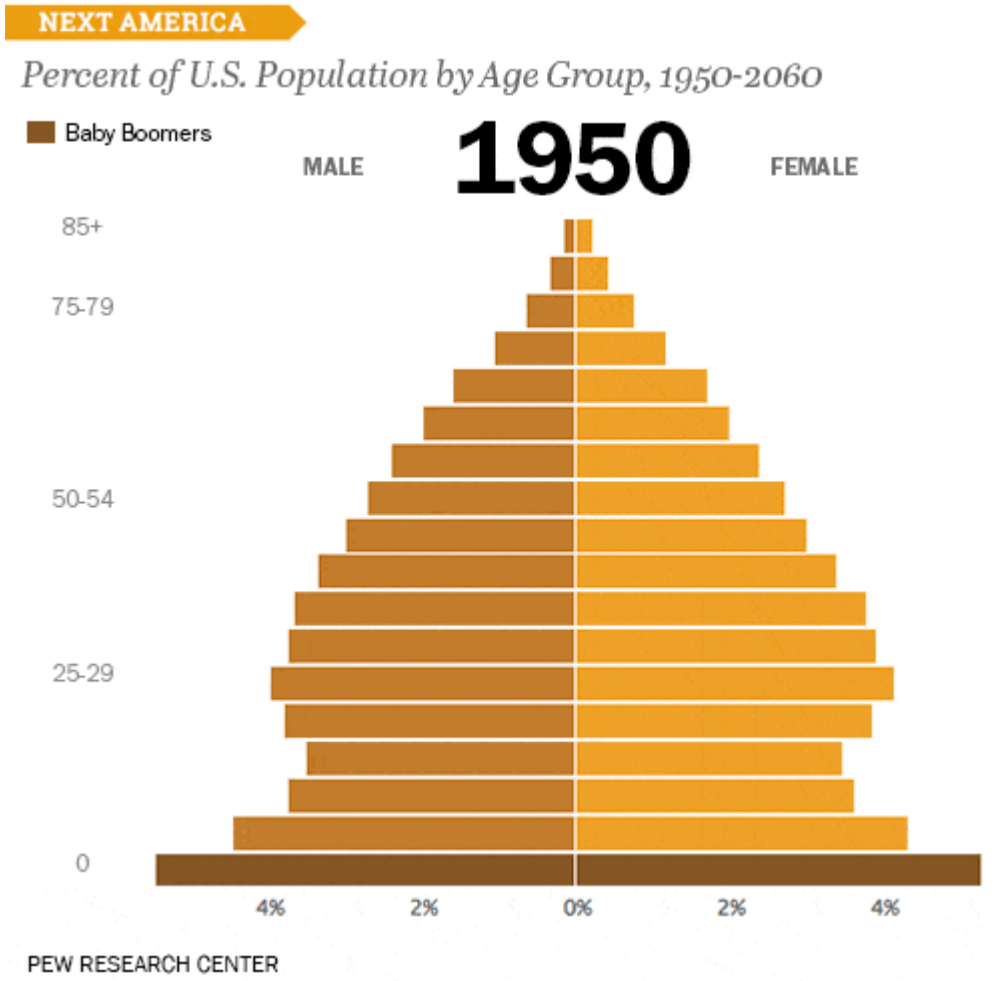
—

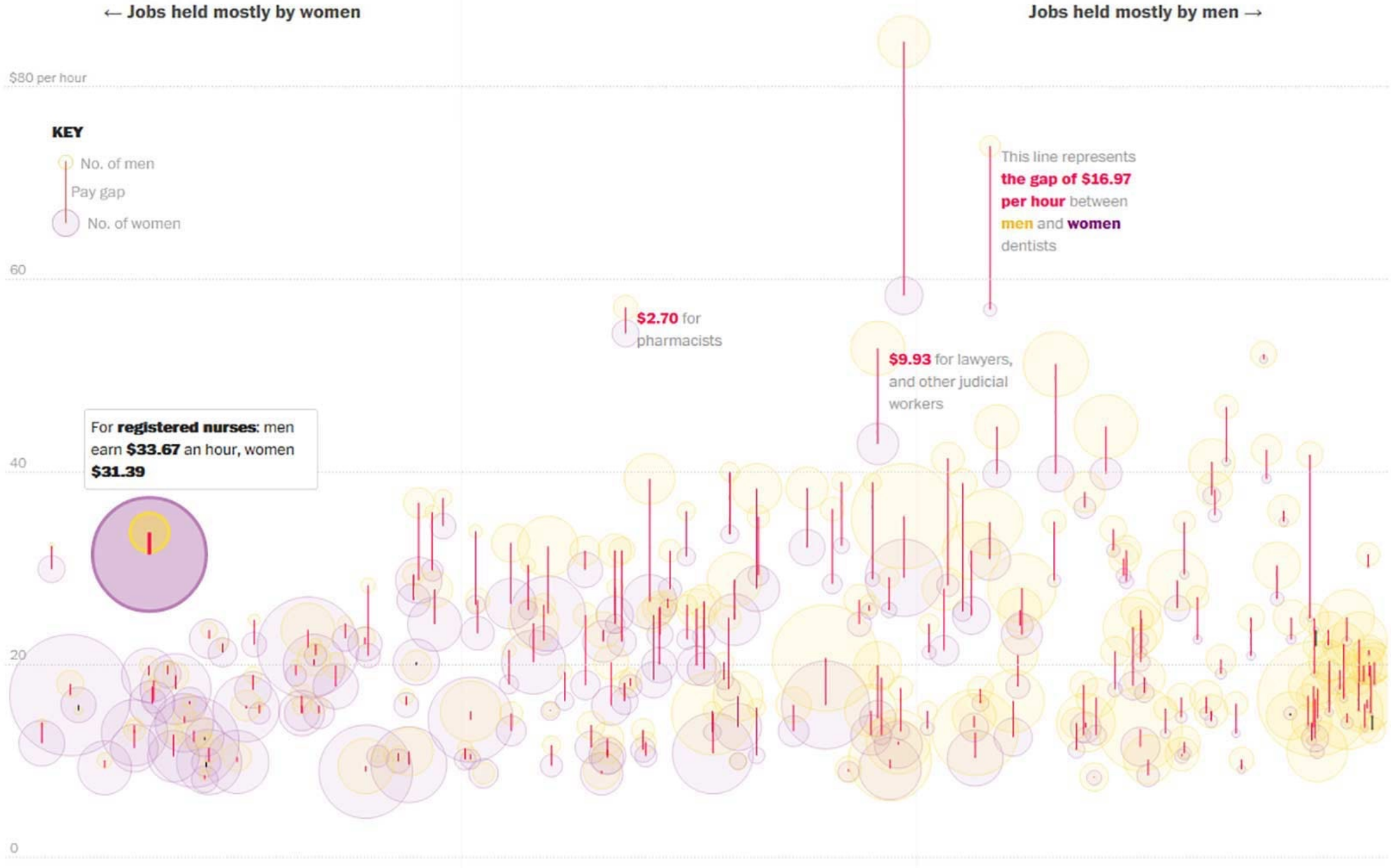
Good examples

Aninmated age and gender demographic.

Shows demographic shift since 1950s.

Pyramid truns into rectangle due to improvements in healthcare and decrease of birthrates.



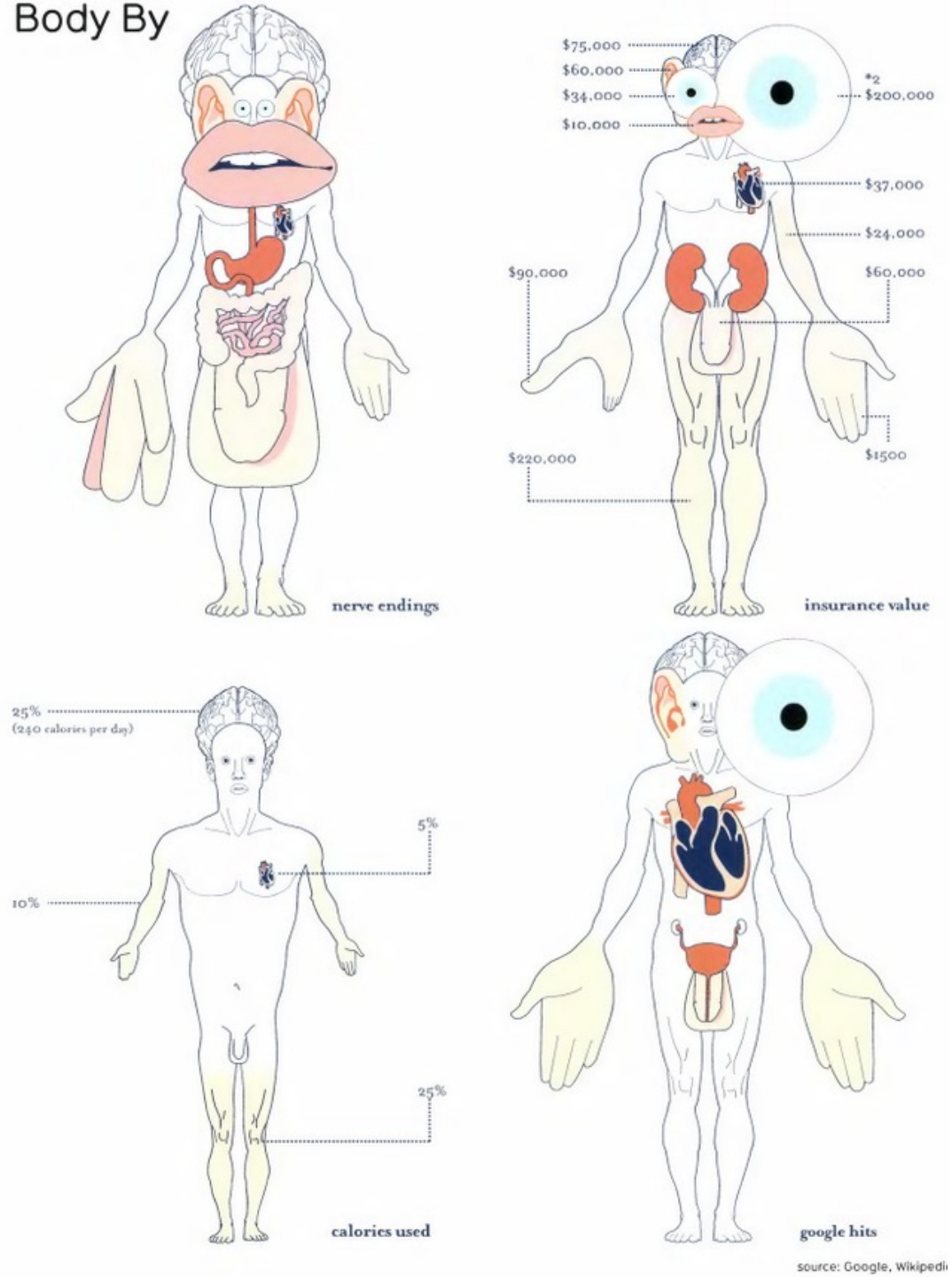

← Jobs held mostly by women
Jobs held mostly by men →
$80 per hour
KEY
No. of men
Pay gap
No. of women
60
This line represents the gap of $16.97 per hour between men and women dentists
$2.70 for pharmacists
$9.93 for lawyers, and other judicial workers
For registered nurses: men earn $33.67 an hour, women $31.39
40
20
0

Body by…

Information is beautiful

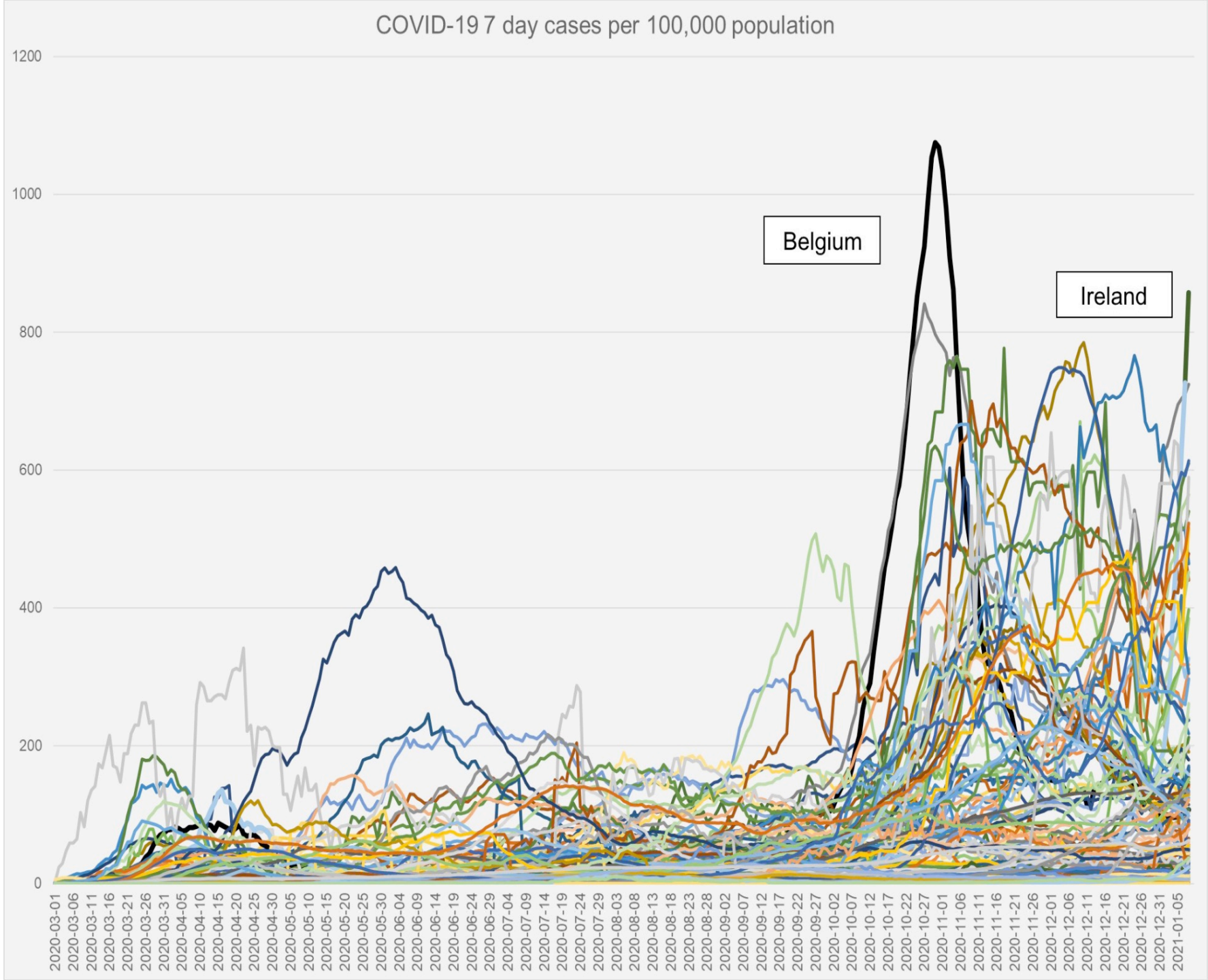
McCandless, David.; 2012

London : William Collins; New ed.



Weak example

Hard to distinguish lines



UNIVERSITY OF WESTMINSTER

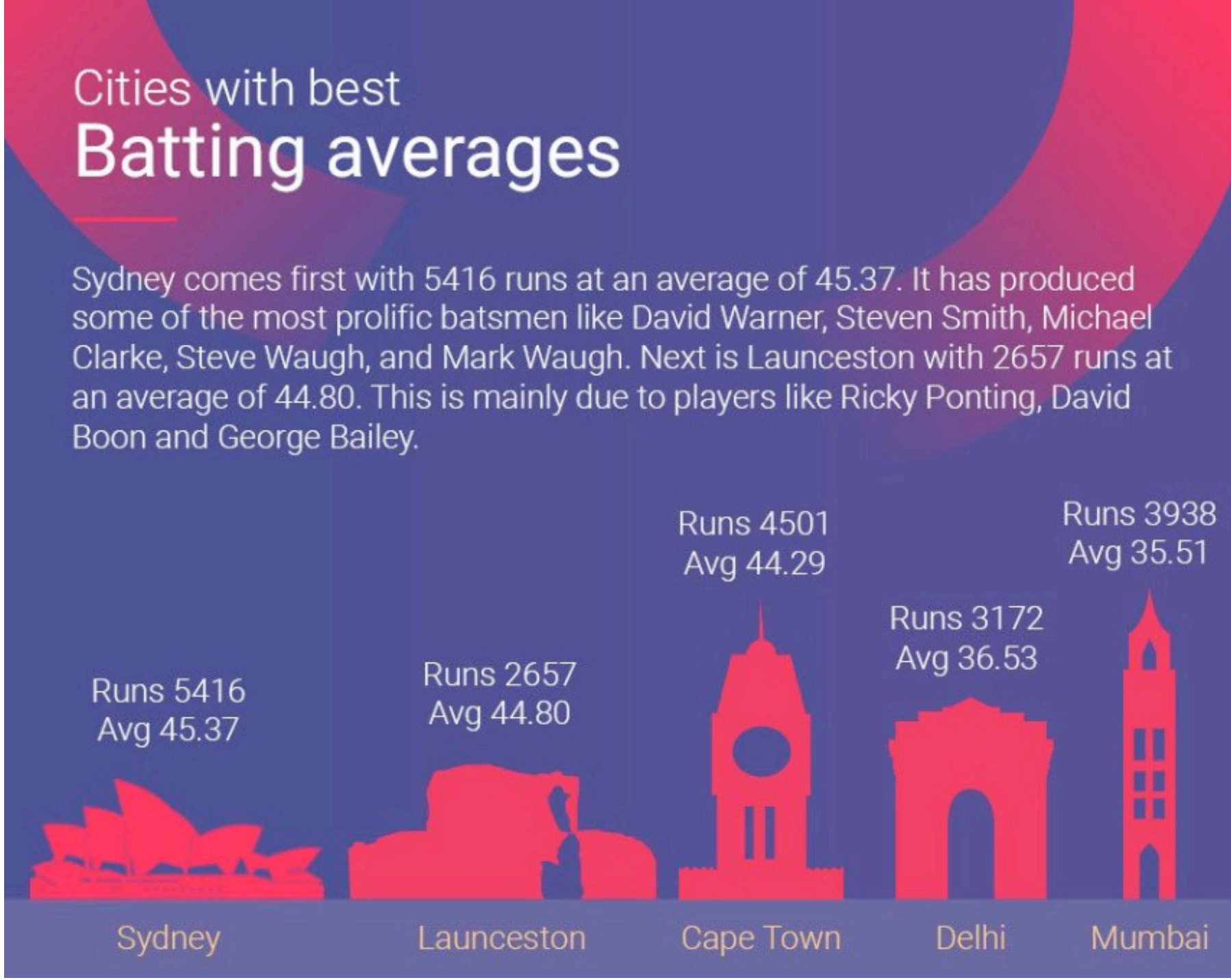
Weak example

Same colours

No axis

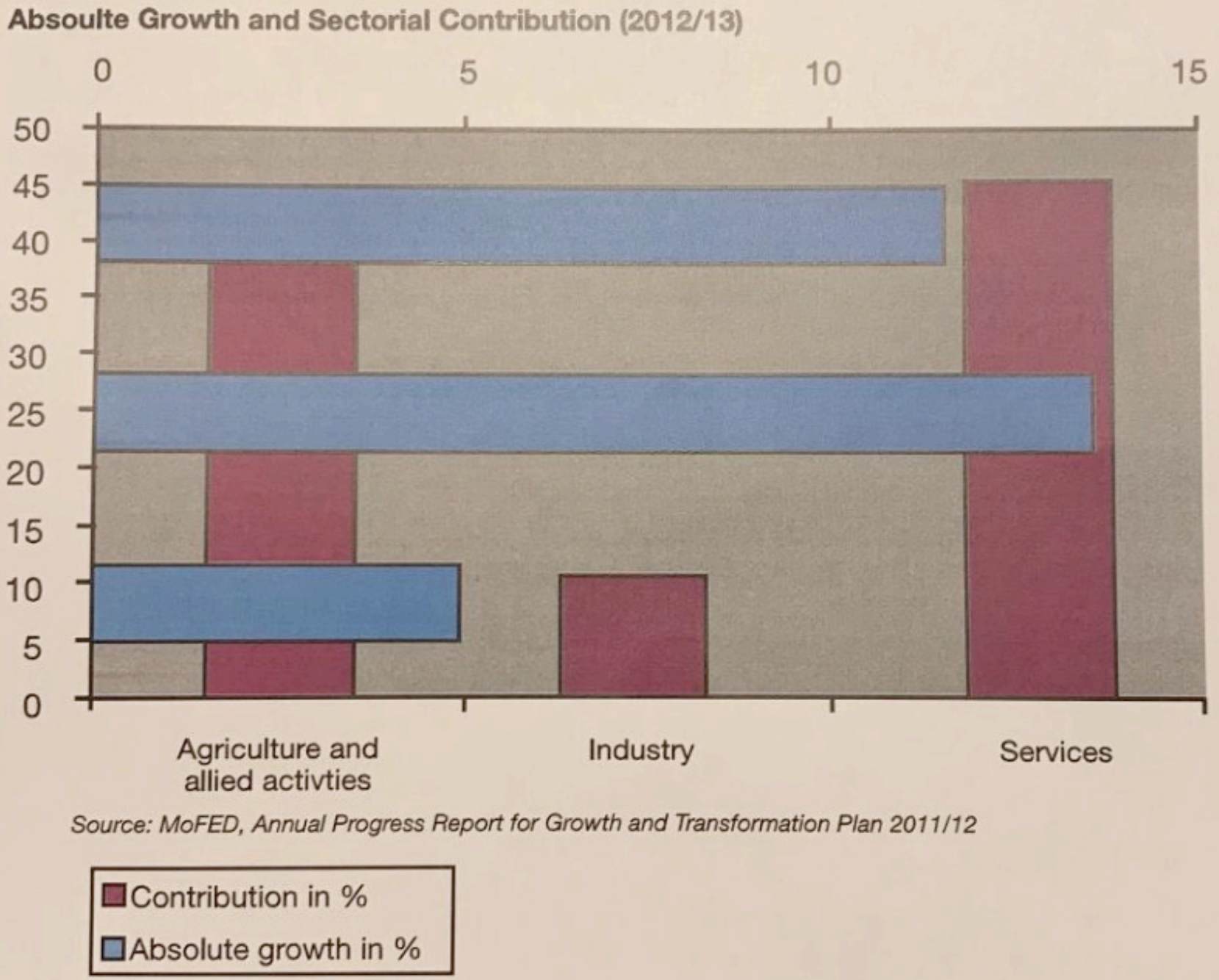
Weak example

What are the graphics for?



Weak example

Can anyone make sense of this?



Pioneers

UNIVERSITY OF WESTMINSTER

William Playfair

‘The commercial and political atlas’ 1786

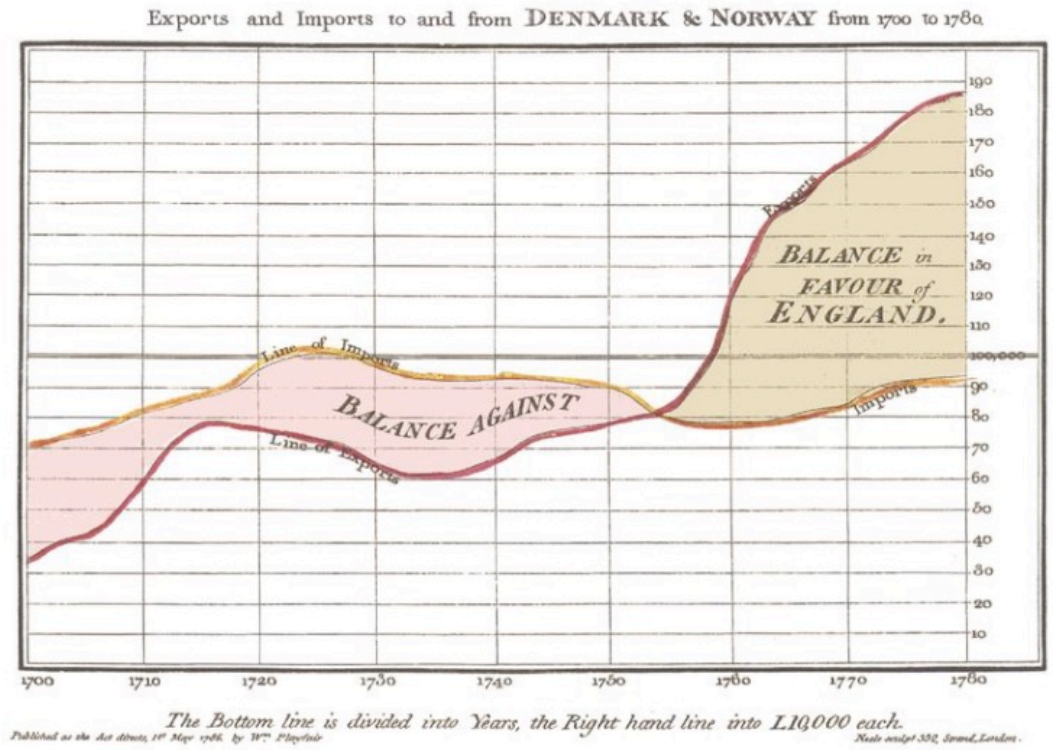
Scottish political economist (born 1759) he is considered one of the great data visualisation thinkers.

Playfair held a variety of jobs in his life but the ones that stand out are as a draftsman of engineering drawings and in his later work as an economist.

In his later work as an economist, Playfair wanted to escape from the tabulated data that governments communicated their statistics with.

Using his experience as a draftsman and knowledge of cartographic printing, he published ‘The Commercial and Political Atlas’ in 1786

He is credited with forever changing the way we look at data –he formalised the bar, line and pie charts.

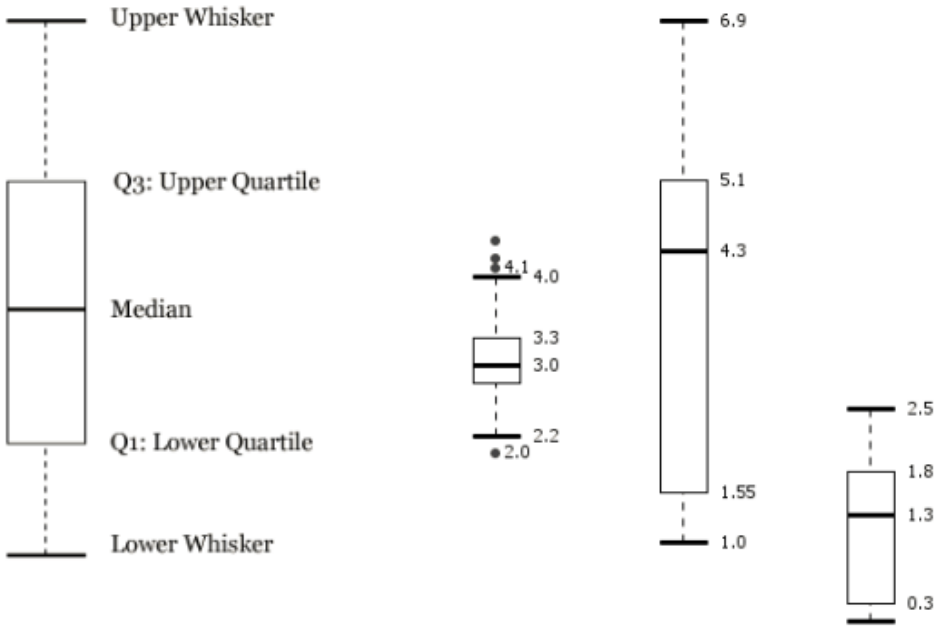


John W. Tukey

Born 1915, he was a legendary American statistician – he contributed a lot to the field of statistics but we are interested in his approach to data analysis and visualisation.

He strongly argued for the exploratory approach i.e. data analysis to form a conclusion -over the confirmatory approach i.e. data used to confirm a hypothesis. In particular he argued for the use of visualisation to explore data.

He wrote the book on exploratory data analysis (‘Exploratory Data Analysis’ 1977) and popularised the boxplot.



‘Exploratory Data Analysis’ 1977

Edward Tufte

Born 1942, he is a professor at Yale University. In 1975 he was asked to teach statistics to a group of visiting journalists at Princeton. He found the available statistical graphics lacking in visual quality.

This initial course became a series of seminars with John W. Tukey and by 1982 a self-published book ‘The Visual Display of Quantitative Information’.

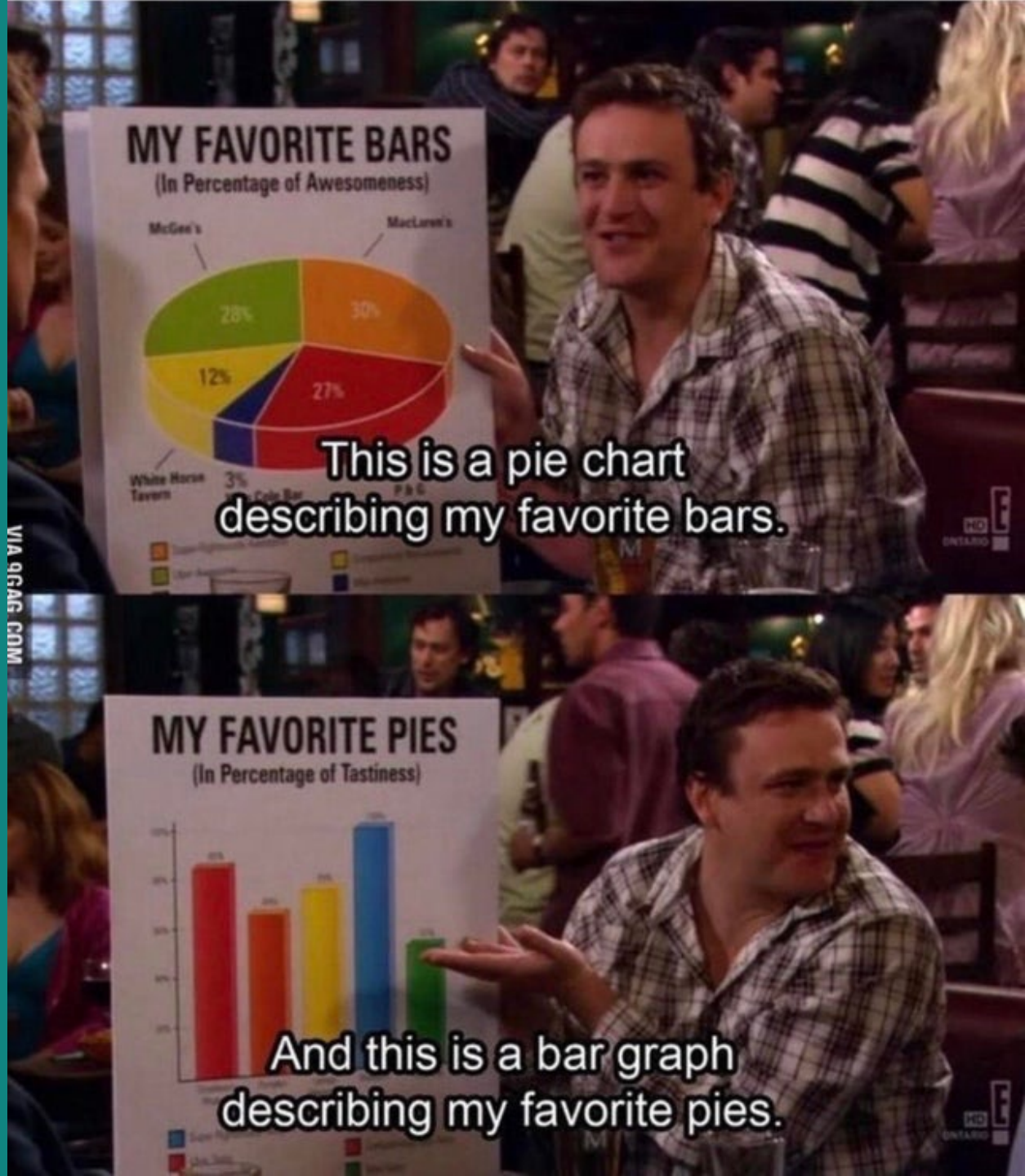
Tufte is considered by some as the ‘godfather’ of data visualisation.

His concept of data ink is crucial to the consideration of the re-design process.

$$Data\ ink\ ratio = \frac{data\ - ink}{ink}$$

“Above all else, show the data.”
‘The Visual Display of Quantitative Evidence’, 1983

Chart of the week



Bar chart

Displays quantitative values for different categories

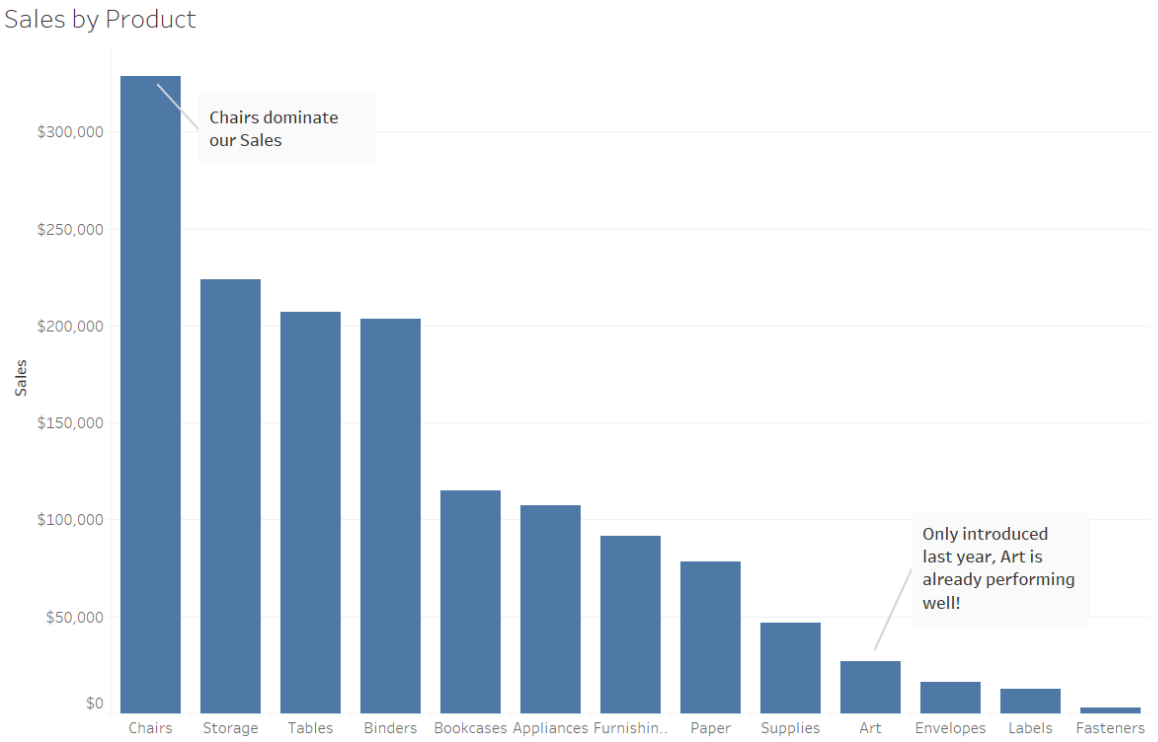
Value is represented in length of bar

Bars always need to start from 0

Order of bars can be important

Easy to see outliers

Tips: Use annotations to highlight interesting data points!



What did we learn today?

Module schedule and assignment

Why visualise data?

The basics of the data visualisation process

Bar chart

Schedule and assessment

	Lecture	Tutorial
1	Introduction to Module	Introduction to R
2	Perception & Gestalt concepts	Introduction to ggplot2 with R
3	Design Principles and graphical encodings	Advanced ggplot2 functionality
4	Colours and typography	Data tidying and transformation with R
5	Data preparation & Exploratory Data Analysis Coursework details issued	Introduction to Power BI
6	Engagement week – no teaching this week	
7	Dashboard design	Advanced PowerBI
8	Spatial data and animation	R: Spatial graphics & animations
9	Storytelling with data	Storytelling exercise
10	Processes, network graphs and unstructured data	In-class test during your allocated tutorial slot (13/07/23)
11	Infographics	Introduction to Tableau
12	Visualisation platforms & Inspirational resources	Formative coursework feedback

Assessment

In-class Test

Multiple-choice

Mostly theory, some practical questions

In tutorial in week 10

30%

Coursework

Analyse and visualise a dataset and present findings

Deadline on 3rd August 2023, 1 pm

70%

Thank you

UNIVERSITY OF WESTMINSTER

www.westminster.ac.uk