

INST0065

Data Visualization and GIS

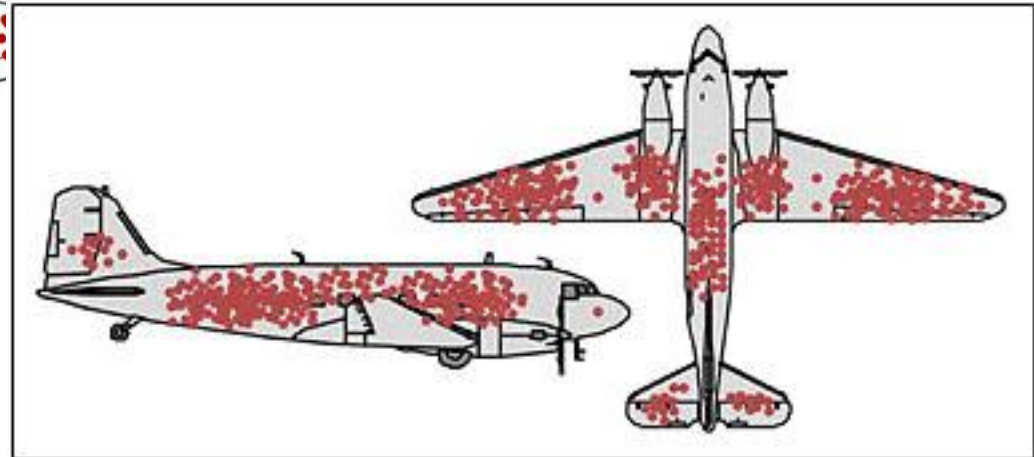
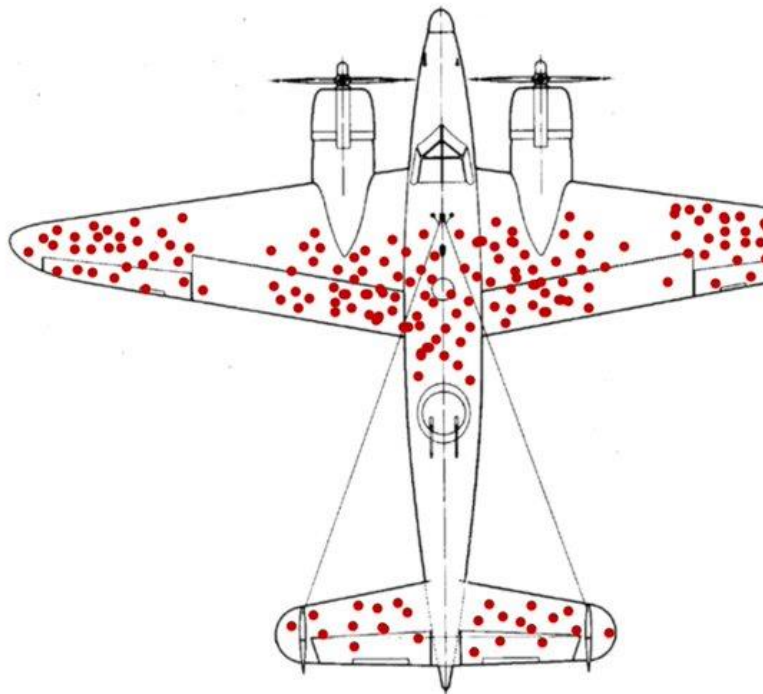
Week 1: Introduction

Dr. Oliver Duke-Williams

o.duke-williams@ucl.ac.uk

(Please use Moodle forums for messages about this module)

Twitter: @oliver_dw



Credit: Cameron Moll

Source: https://en.wikipedia.org/wiki/Survivorship_bias

Source: <https://www.motherjones.com/kevin-drum/2010/09/counterintuitive-world/>

Typical retelling: <https://twitter.com/garius/status/1347490994443464704>

See also: Casselman (2016)

Objectives of this module

- Principles of data visualization
- Types of visualization
- Meaning and mis-representation in visualisations
- Critical assessment of existing visualisations
- Acquiring, processing and preparing data
- Creating basic and more advanced visualisations using appropriate software
- Geographic Information Systems – principles
- Geographic Information Systems – creating maps using appropriate software

Contents

- About this module
- Data visualization and GIS
- Data visualization examples (good and bad)

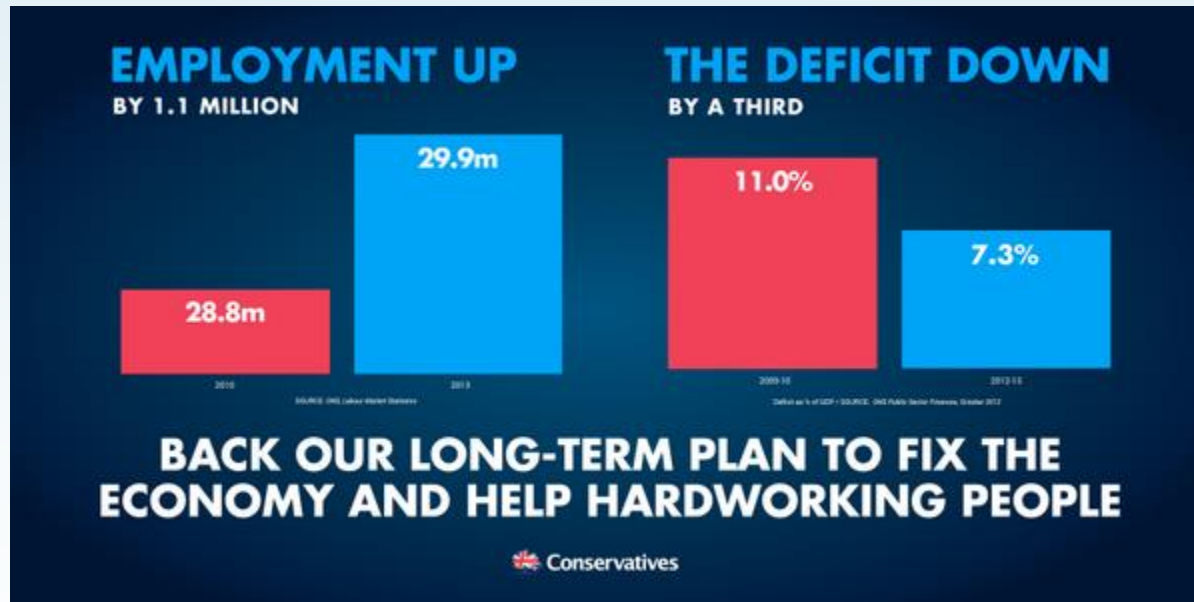
Module structure

- Each week
 - 'Lecture' slot – introduced by ODW
 - Time to discuss and ask questions about previous week's material
 - From week 2
 - Groups of students will discuss a data visualization
 - Chosen because it is good?
 - Chosen because it is bad?
 - Chosen because it is interesting?
 - Chosen because it is topical?
 - Practical session

Data Visualization and GIS

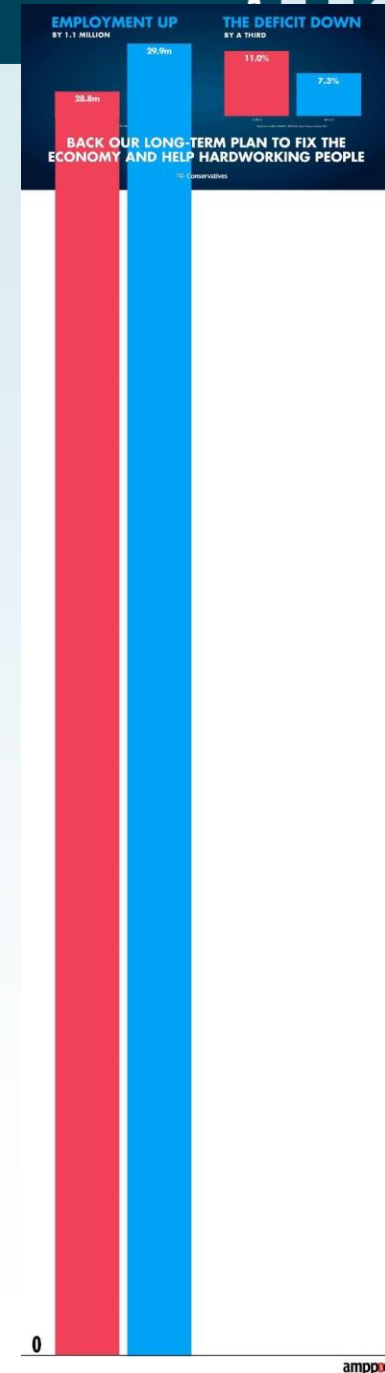
- There are many types of data visualization
 - GIS is a (specialised) subset of dataviz
- Catalogues of many types of dataviz exist, e.g.
 - <https://datavizproject.com/>
 - <https://datavizcatalogue.com/>
- The 'right' dataviz to use will depend on what input data you have, and what you are trying to learn or communicate about it

A data visualisation (Conservative party, 1983)



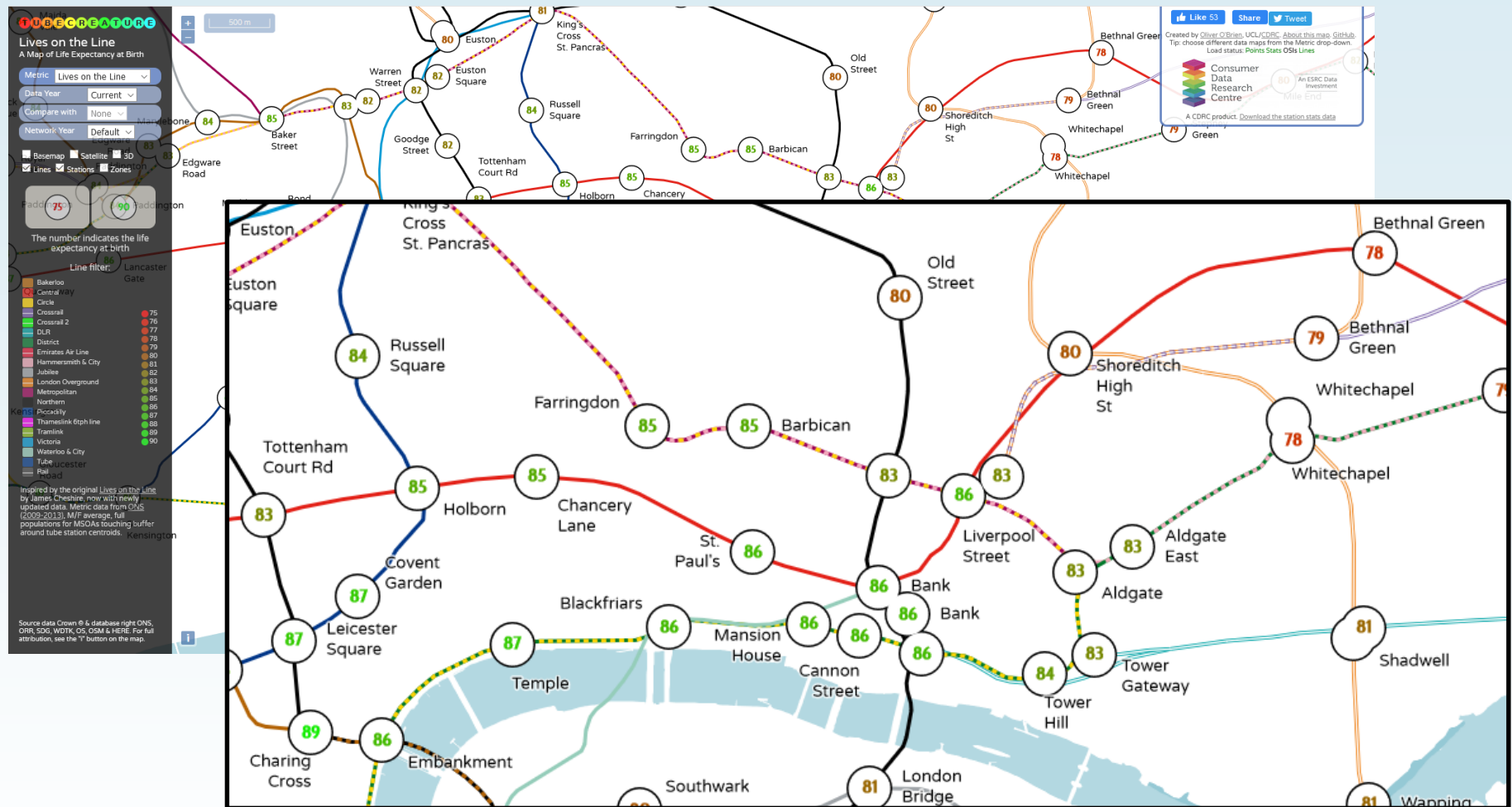
Source: <https://twitter.com/Conservatives/status/407945493700812800>

Another view of the same data



Source: <https://twitter.com/ampp3d/status/408198815733137409>

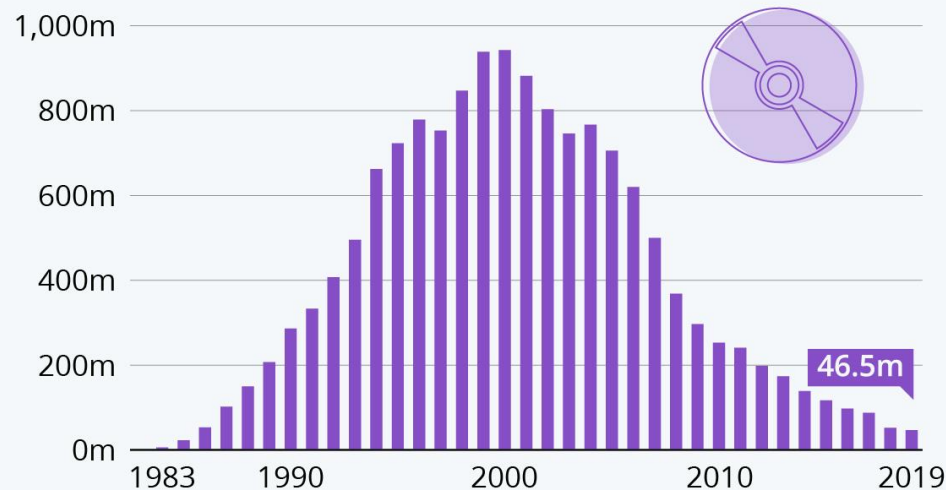
A selection of examples...



Source: <https://jcheshire.com/featured-maps/lives-on-the-line/>

The Rise and Fall of the Compact Disc

CD album sales in the United States since 1983
(in million units)

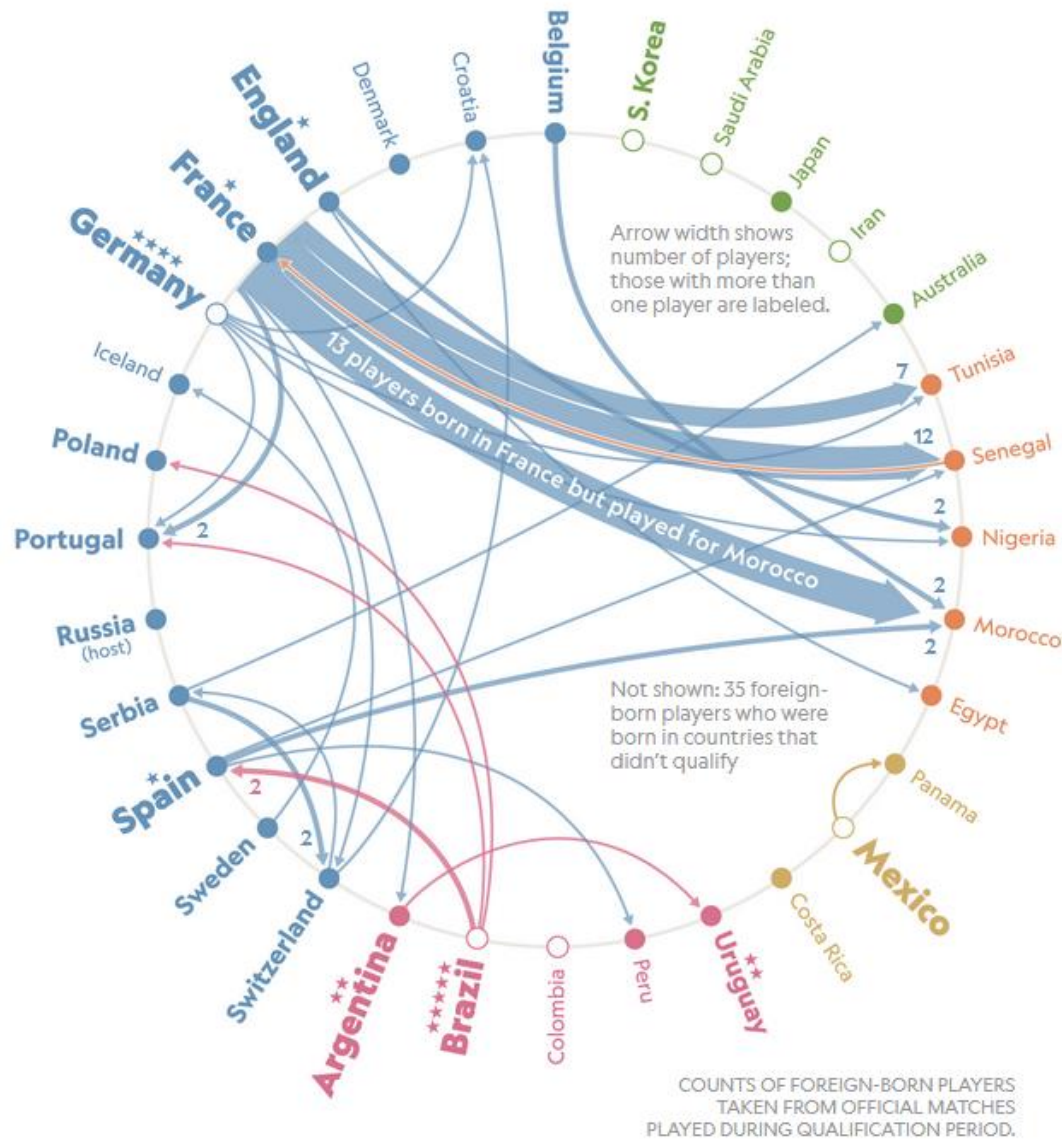


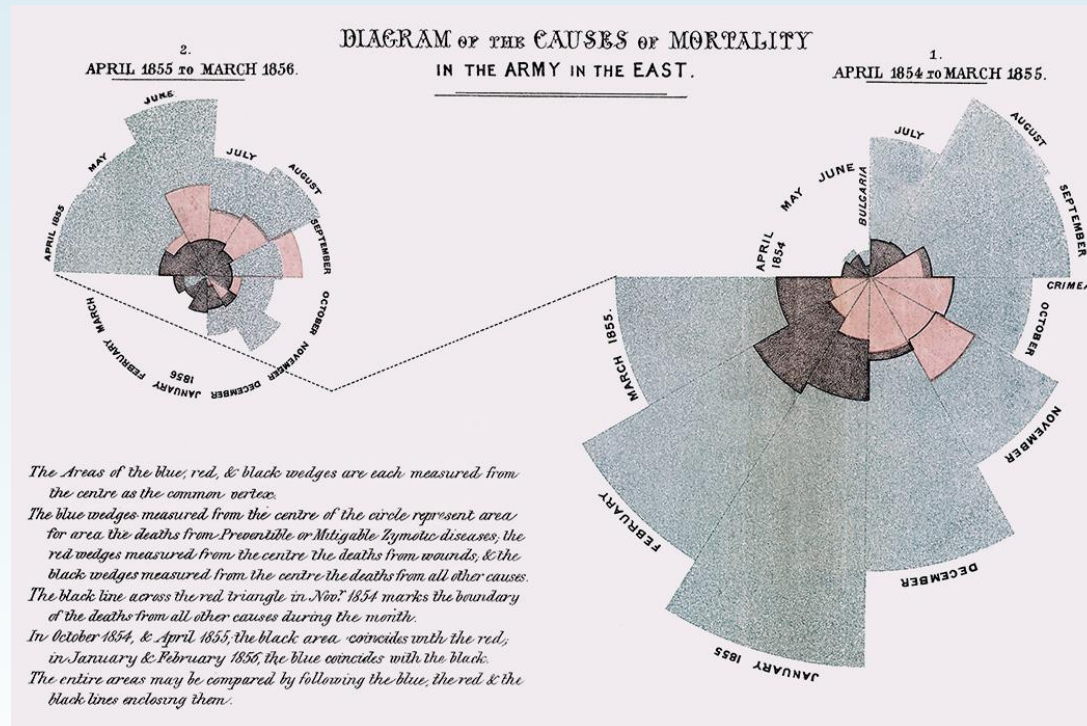
Source: RIAA



statista

Source: <https://www.statista.com/chart/12950/cd-sales-in-the-us/>





Info: <https://daily.jstor.org/florence-nightingale-data-visualization-visionary/>

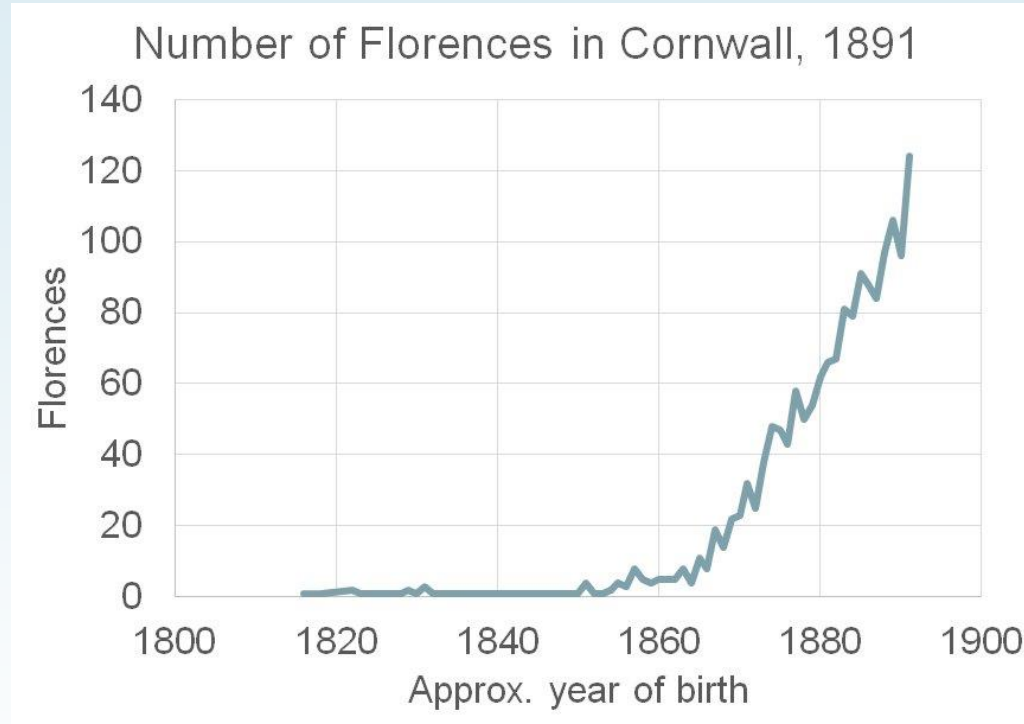
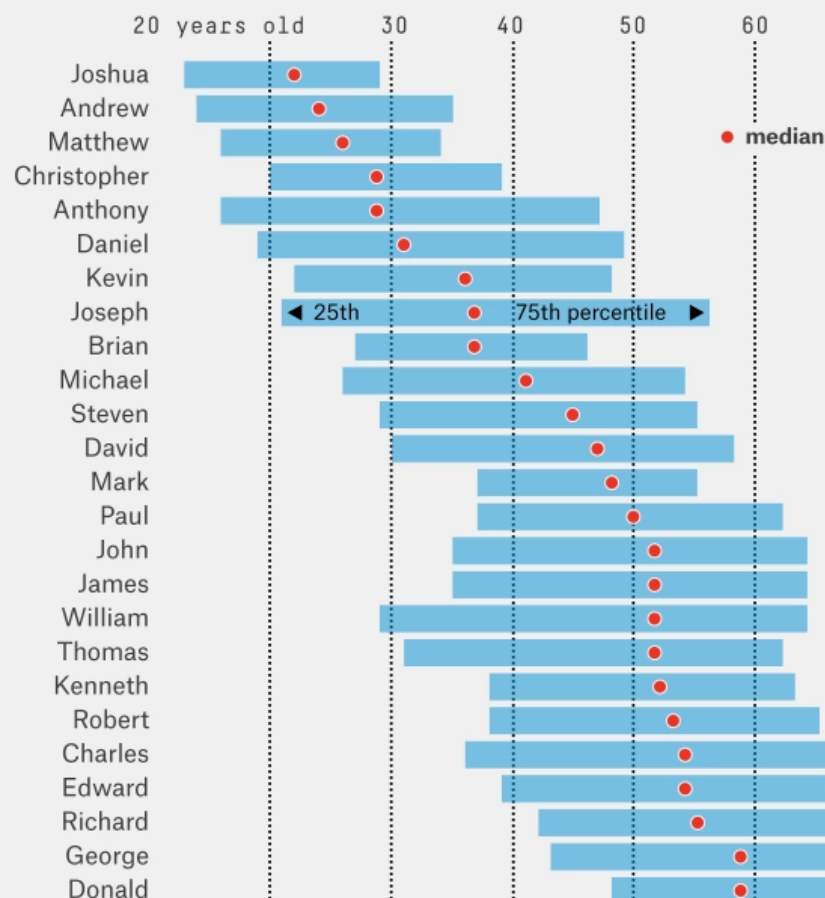


Image source: https://twitter.com/oliver_dw/status/1177942909762572290?s=20

Median Ages For Males With the 25 Most Common Names

Among Americans estimated to be alive as of Jan. 1, 2014

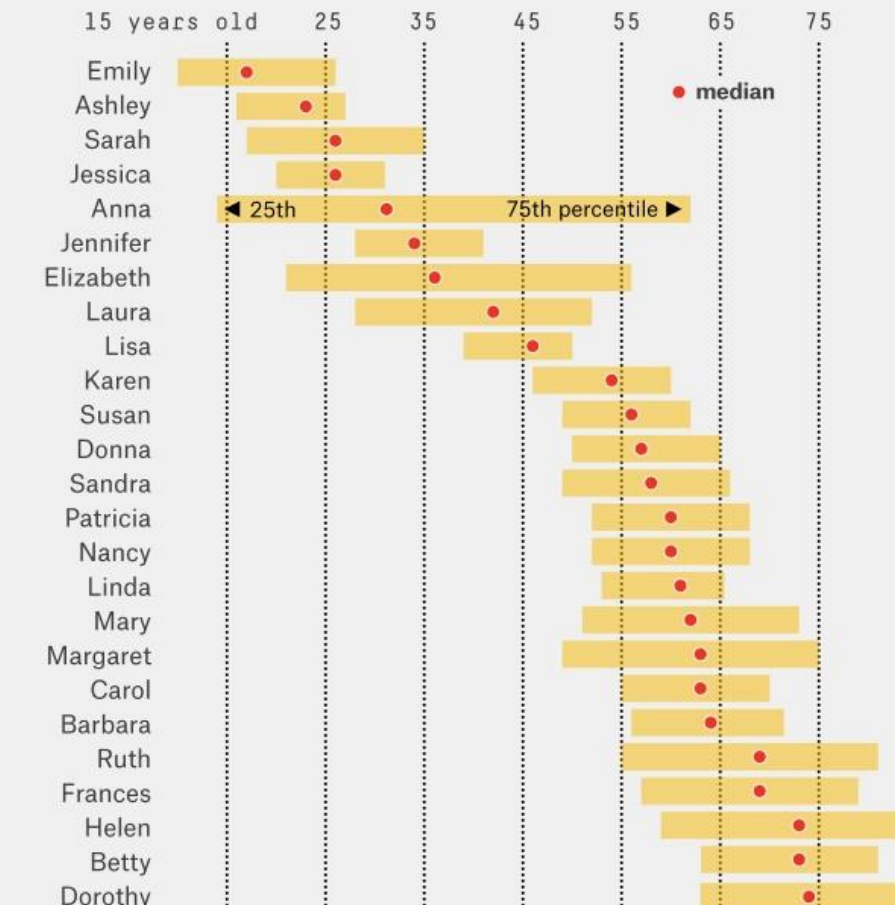


FIVETHIRTYEIGHT

SOURCE: SOCIAL SECURITY ADMINISTRATION

Median Ages For Females With the 25 Most Common Names

Among Americans estimated to be alive as of Jan. 1, 2014



FIVETHIRTYEIGHT

SOURCE: SOCIAL SECURITY ADMINISTRATION

Data visualizations don't have to be on screen



Source: <https://blog.mattwaite.com/post/108885953514/using-lego-to-teach-data-visualization>



American artist [Cory Imig](http://dataphys.org/list/psychogeographical-mapping-travel-logging-with-lego-bricks/) reconstructed the layout of the city of Savannah using LEGO bricks, and over the course of one month she added a colored brick every time she went to a particular place. Each color is a different day of the week.

Source: <http://dataphys.org/list/psychogeographical-mapping-travel-logging-with-lego-bricks/>



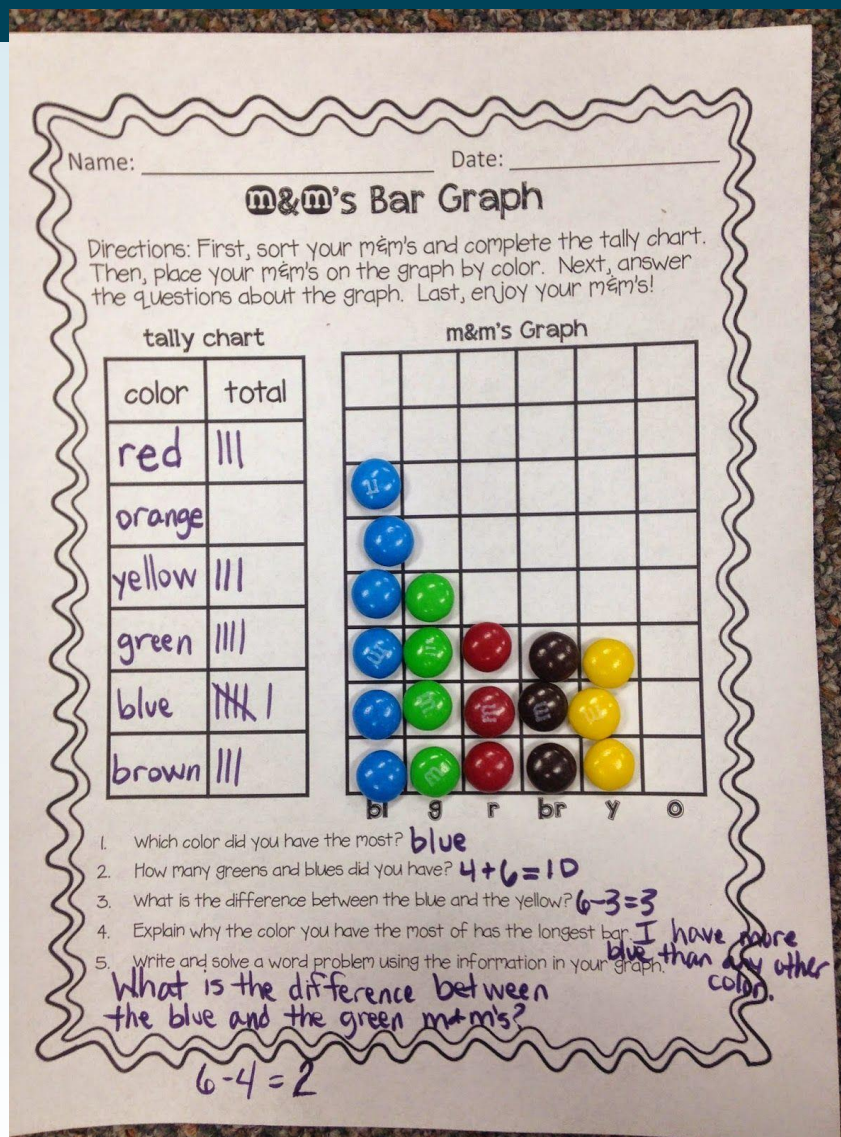
My mother is a commuter in the Munich area. And an enthusiastic knitter. In 2018 she knitted a "train delay scarf". Two rows per day: gray if you are less than 5 minutes late, pink if you are 5 to 30 minutes late, red if you are delayed on both journeys or once over 30 minutes.

Source: https://twitter.com/sara_weber/status/1081950904671240192?s=20

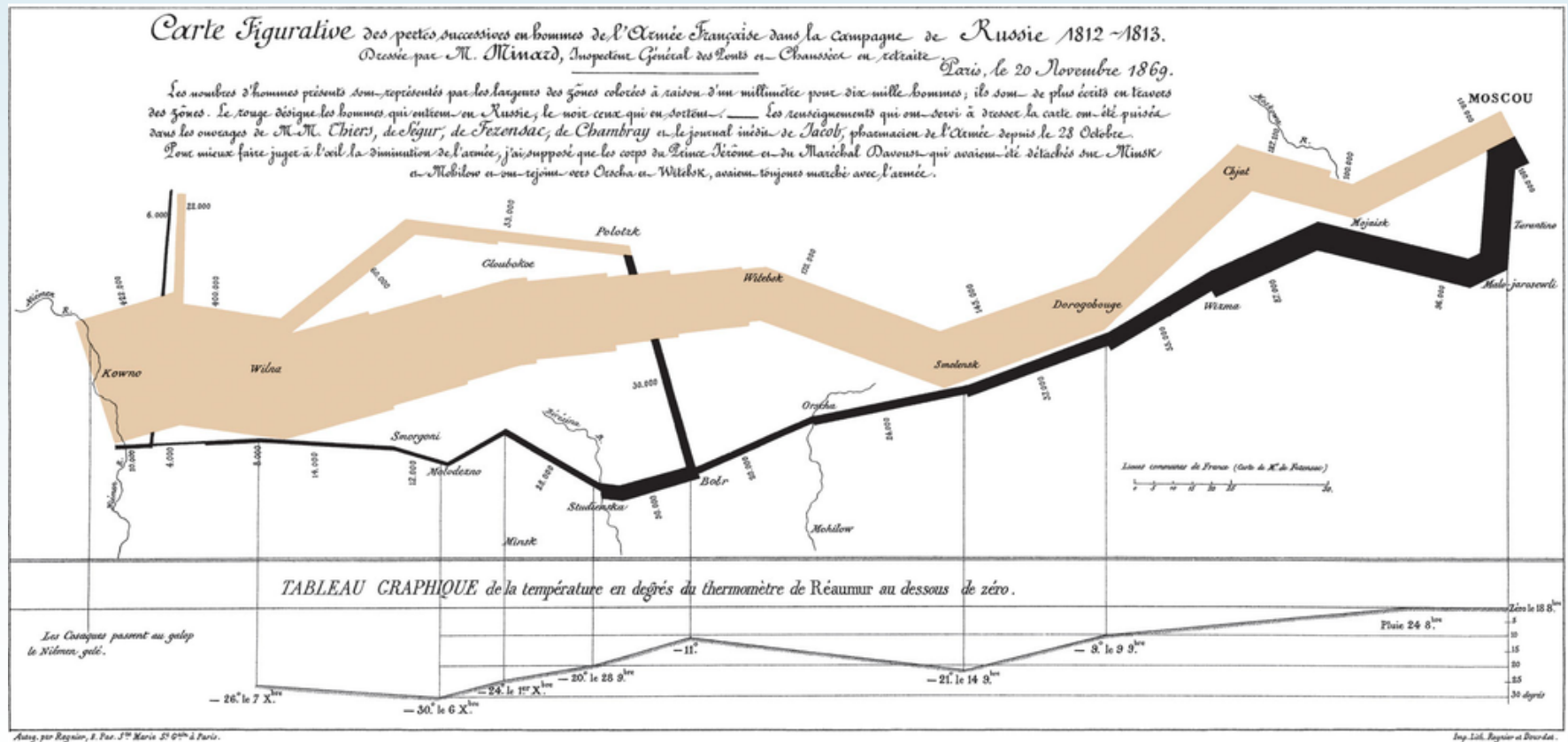


A visualization of my son's sleep pattern from birth to his first birthday. Crochet border surrounding a double knit body. Each row represents a single day. Each stitch represents 6 minutes of time spent awake or asleep

Source: <https://twitter.com/Lagomorpho/status/11497545925796003>



The 'Minard map'



Translation of text

Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812–1813.

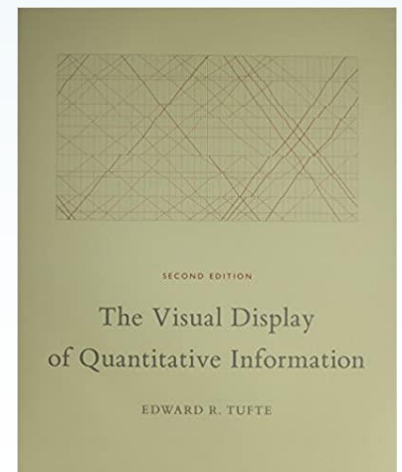
Drawn up by M. Minard, Inspector General of Bridges and Roads in retirement.
Paris, 20 November 1869.

The numbers of men present are represented by the widths of the coloured zones at a rate of one millimetre for every ten-thousand men; they are further written across the zones. The red designates the men who enter into Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M. M. Thiers, of Segur, of Fezensac, of Chambray, and the unpublished diary of Jacob, pharmacist of the army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of prince Jerome and of Marshal Davoush who had been detached at Minsk and Moghilev and have rejoined around Orcha and Vitebsk, had always marched with the army.

Edward Tufte's praise for the map

"It may well be the best statistical graphic ever drawn."

Tufte E, (1983) The Visual Display of Quantitative Information



The first is the classic of Charles Joseph Minard (1781–1870), the French engineer, which shows the terrible fate of Napoleon's army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence,¹² this combination of data map and time-series, drawn in 1869, portrays a sequence of devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at left on the Polish-Russian border near the Niemen River, the thick tan flow-line shows the size of the Grand Army (422,000) as it invaded Russia in June 1812. The width of this band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. At upper right we see Minard's French original, which was printed as a two-color lithograph in the form of a small poster. And at lower right, our English translation.

It may well be the best statistical graphic ever drawn.

¹² E. J. Marey, *La méthode graphique* (Paris, 1885), p. 73. For more on Minard, see Arthur H. Robinson, "The Thematic Maps of Charles Joseph Minard," *Imago Mundi*, 21 (1967), 95–108.

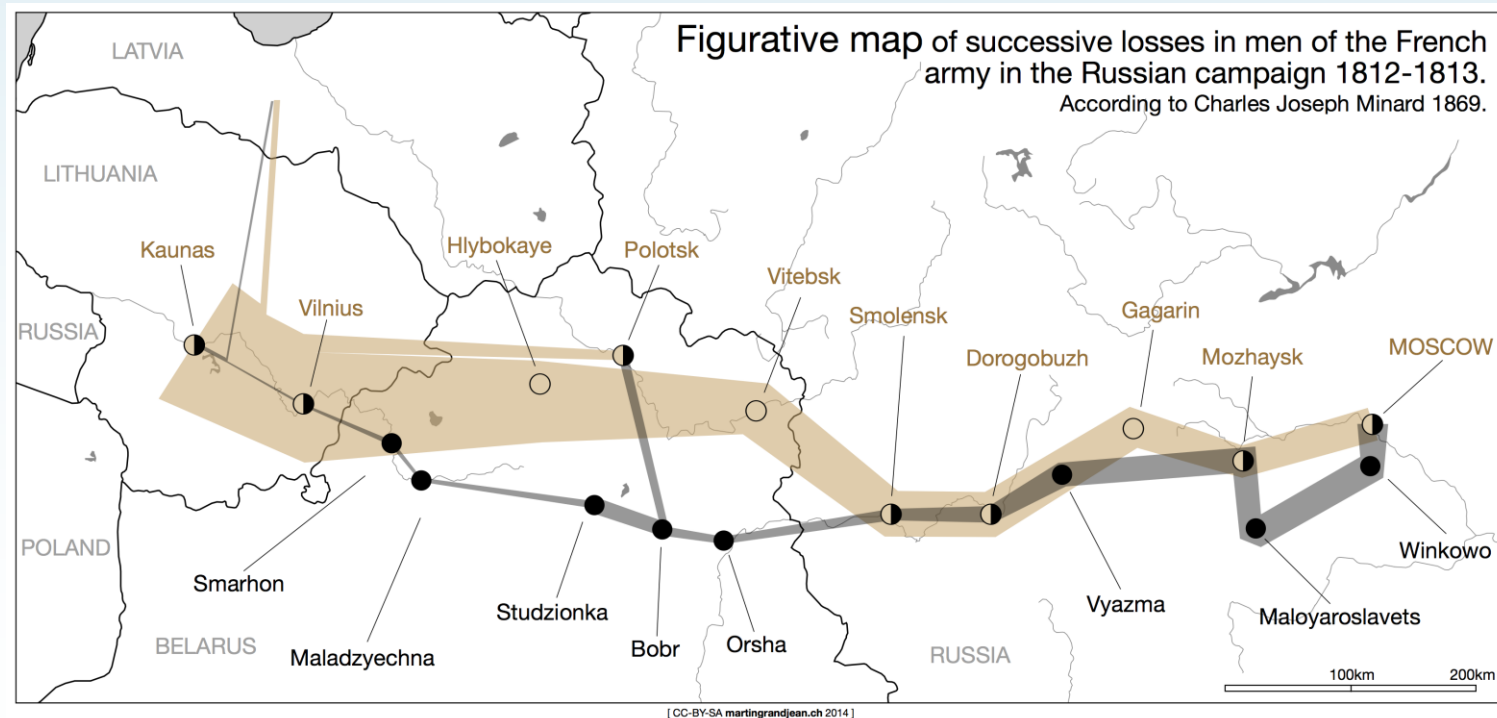
Upper image from Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard, 1845–1869*, Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris, item 28 (62 by 25 cm, or 25 by 10 in). English translation by Dawn Finley and redrawing by Elaine Morse, completed August 2002.

Map content

- Tufte identifies the following features
 1. Line width = size of army
 2. Latitude
 3. Longitude
 4. Direction of travel
 5. Location by date
 6. Temperature during the retreat

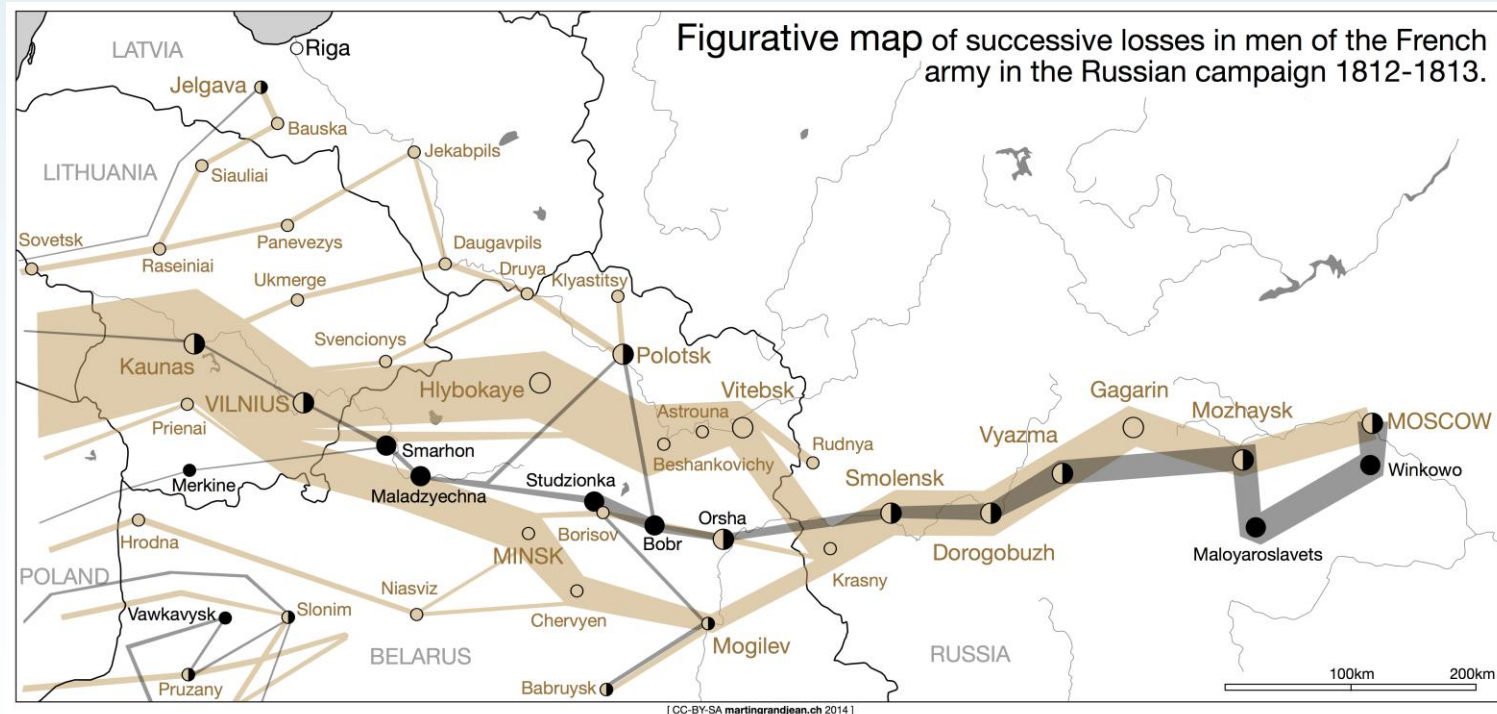
Adaptations to the map (1)

Martin Grandjean: Historical Data Visualization: Minard's map vectorized and revisited



Adaptations to the map (2)

Martin Grandjean: Historical Data Visualization: Minard's map vectorized and revisited

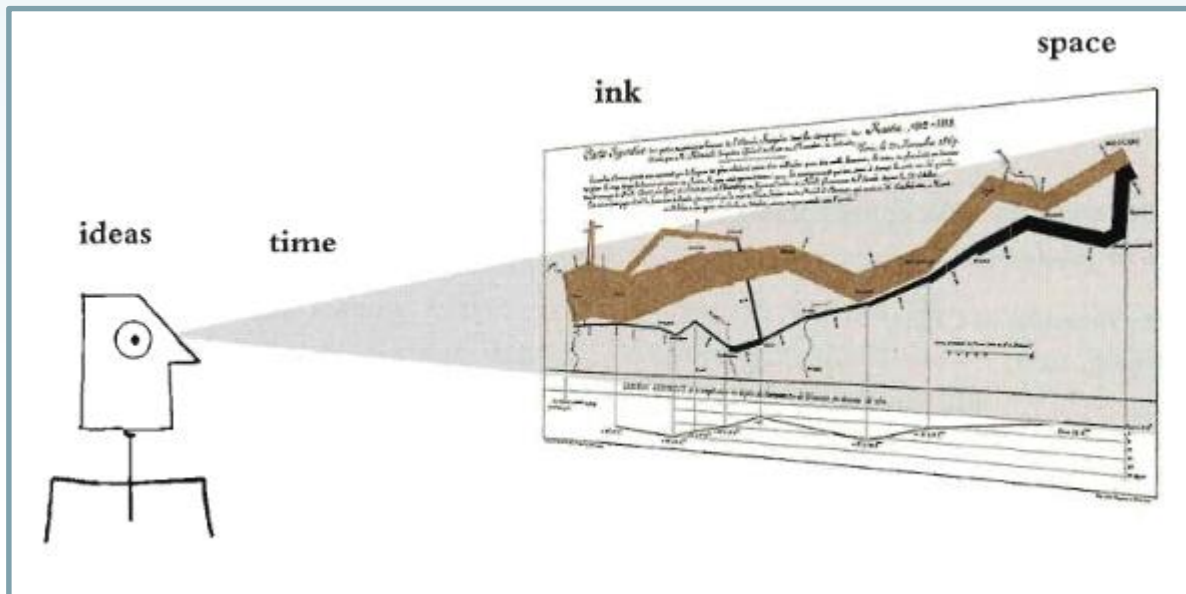


Principles of Graphical Excellence

Graphical excellence is the well-designed presentation of interesting data – a matter of *substance*, of *statistics* and of *design*.

Graphical excellence consists of complex ideas communicated with clarity, precision and efficiency.

Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

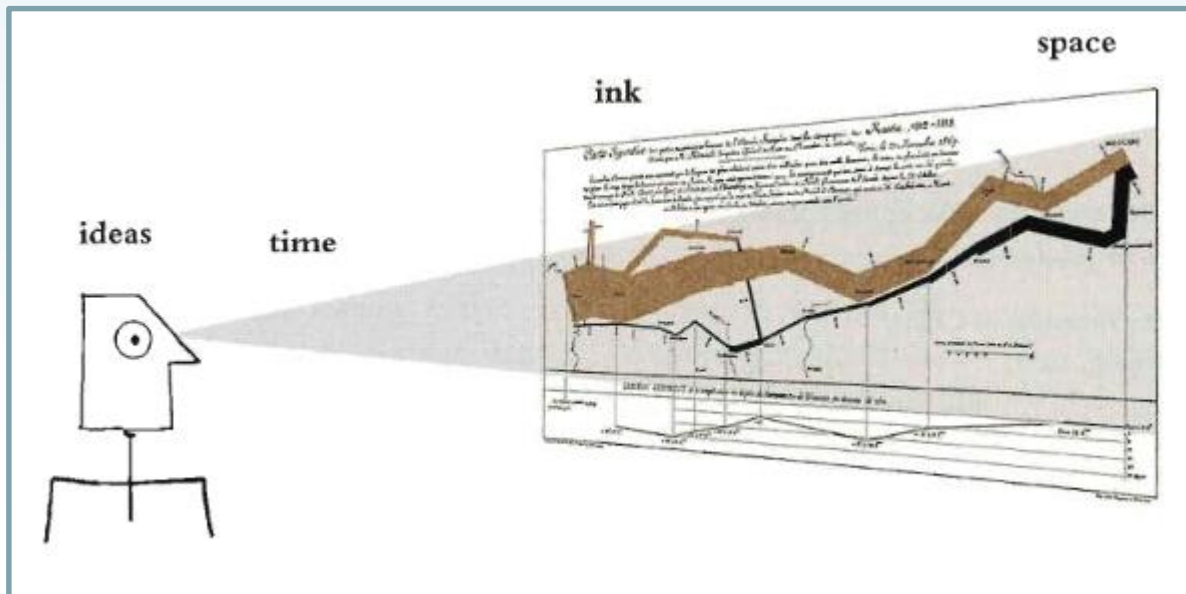


Source: Tufte (1983)

Principles of Graphical Excellence

Graphical excellence is nearly always multivariate.

And graphical excellence requires telling the truth about the data.



Source: Tufte (1983)

Tools that we will use

- The language R
- The desktop package QGIS
- Both are open source and can be downloaded, and are also available on the UCL desktop

R

- We will learn more formally about R in the practical
- A brief background:
 - 1977: John Tukey publishes 'Exploratory Data Analysis'
 - 1980: The language S is created, designed to enable EDA
 - 1980s and 1990s, S-PLUS is commercially developed as an implementation of S
 - 1995: First release of R, an open-source implementation of S

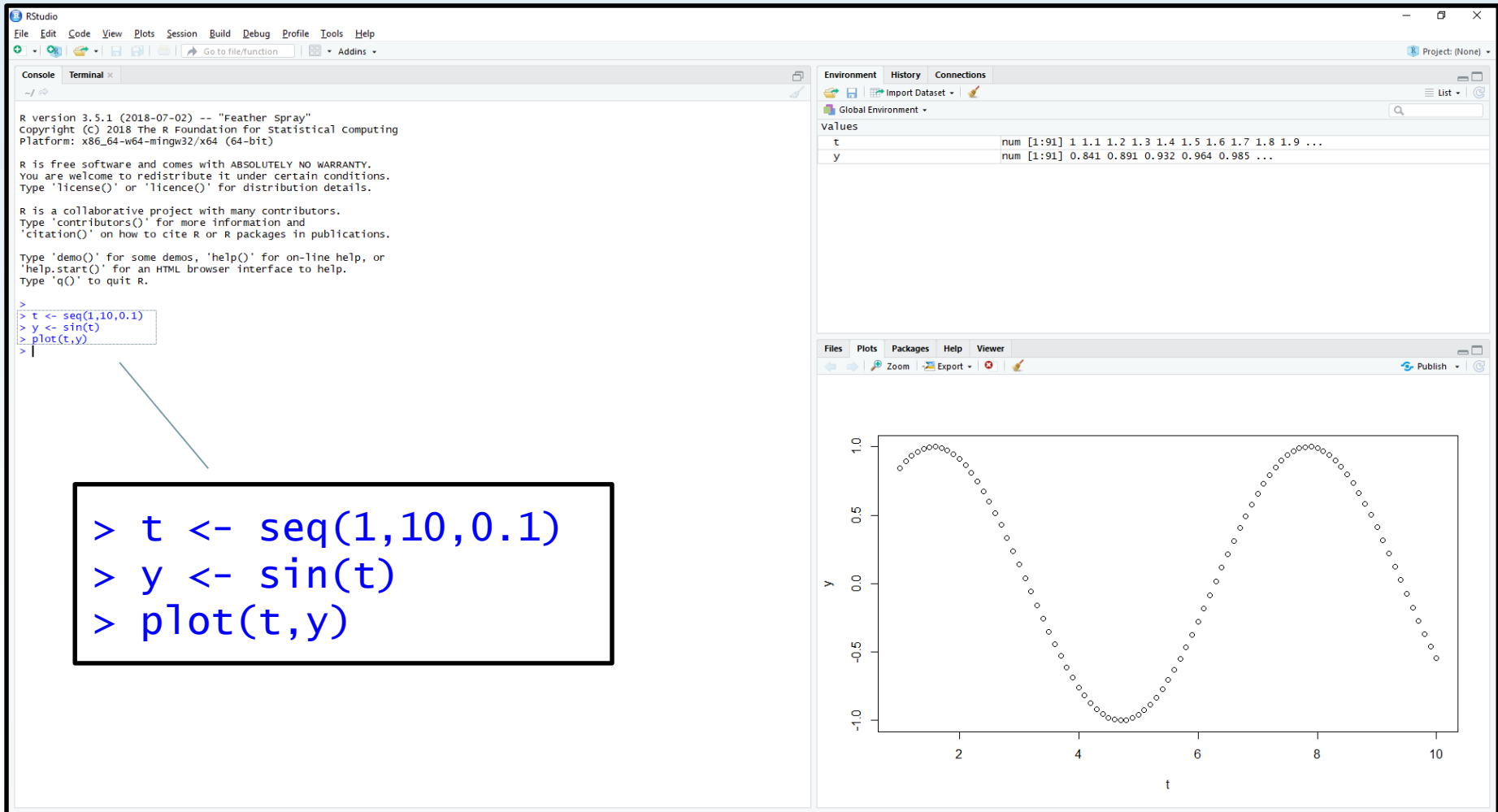
R

- R commands are typed at a console prompt
 - A sequence of commands can be assembled into scripts
- R Studio provides a console plus supporting tools and output windows etc
- R commands consist of expressions and assignment

```
> 1 + 2          # This is an expression
```

```
> x <- 1 + 2      # This is an assignment
```

R Studio



The screenshot displays the R Studio environment. The console on the left shows the R version (3.5.1) and the execution of three commands: `t <- seq(1,10,0.1)`, `y <- sin(t)`, and `plot(t,y)`. The Environment pane on the right shows the variables `t` and `y` as numeric vectors. The Plots pane on the bottom right shows a scatter plot of `y` versus `t`, which is a sine wave.

Console Output:

```
R version 3.5.1 (2018-07-02) -- "Feather Spray"
Copyright (c) 2018 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
> t <- seq(1,10,0.1)
> y <- sin(t)
> plot(t,y)
> 
```

Environment Pane:

Variable	Class	Values
t	num [1:91]	1 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 ...
y	num [1:91]	0.841 0.891 0.932 0.964 0.985 ...

Plots Pane:

A scatter plot of `y` versus `t` showing a sine wave. The x-axis (`t`) ranges from 1 to 10, and the y-axis (`y`) ranges from -1.0 to 1.0. The plot consists of open circles representing the data points.

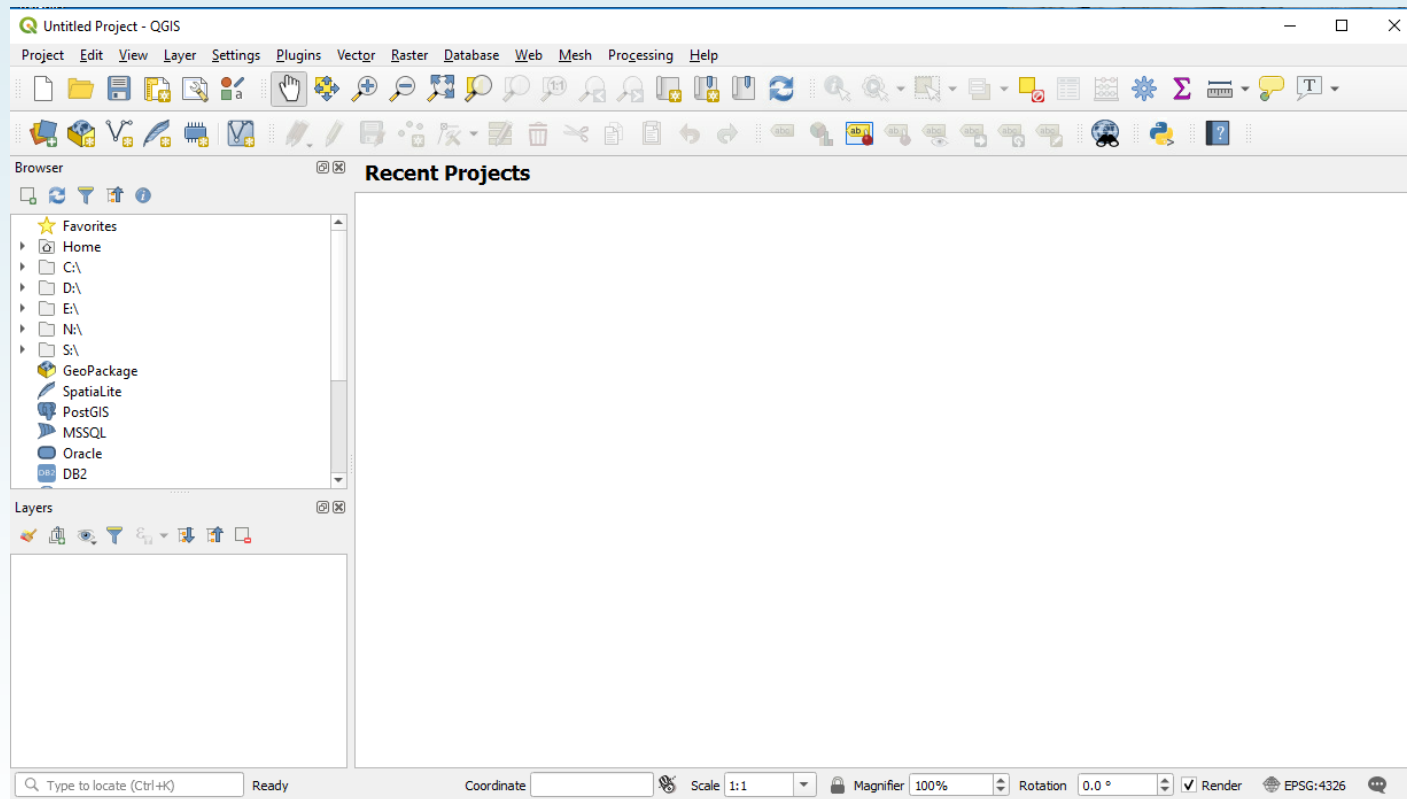
QGIS

- Quantum GIS (QGIS)
 - Available on Desktop@UCL
 - Open source
 - Broadly similar to ArcGIS
 - Very powerful

Background

- Quantum GIS (QGIS)
 - Open source GIS system
 - Builds on and integrates a diverse set of separately produced software
 - www.qgis.org
 - Available for Windows / Mac / Linux / Android

QGIS – starting it for the first time...



- The initial desktop is typical of many specialist applications
 - Many buttons / icons
 - Impossible to know where to start for new users

