

# ggplot2 & data visualization

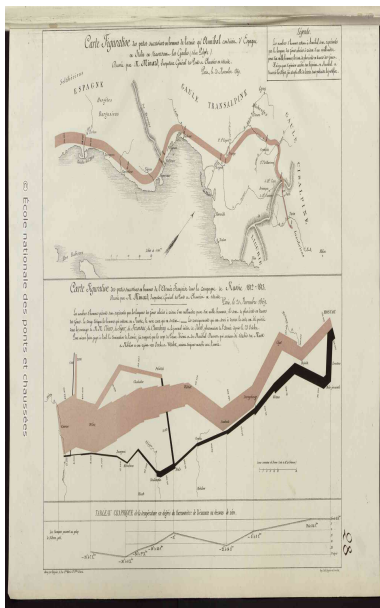
Massimiliano Russo

October 13, 2016

*Of all methods for analyzing and communicating statistical information, well designed data graphics are usually the simplest and at the same time the most powerful.*

- Data visualization is a quick, easy way to convey concepts in a universal manner.
- Turns numbers and letters into aesthetically pleasing visuals, making it easy to recognize patterns and find exceptions.

# Data visualization is an old concept



- Charles Minard in 1869 produced summarized Napoleon's Russian campaign (1812) in a plot.
- thick band illustrates the size of his army at specific geographic points during their advance and retreat.
- He contemporary showed the number of Napoleon's troops; distance; temperature; the latitude and longitude; direction of travel and location relative to specific dates.

An huge amount of software is available to produce plots but we focus on `ggplot2` (Hadley Wickham) R package.

`ggplot2` is an open source implementation of the **layered grammar of graphics**.

It provides some nice features as:

- saving plots (or the beginnings of a plot) as objects
- simplification of multivariate exploration through faceting and coloring
- plot's evolution (or devolution) with minimal changes to code
- great documentation.

# The layered grammar of graphics

## What is it?

Grammar of graphics is a tool that enables us to concisely describe the components of a graphic.

In brief, the grammar tells us that a statistical graphic is a **mapping from data to aesthetic** attributes (colour, shape, size, . . .) of geometric objects (points, lines, bars, . . .).

## Why layers?

The layer system allows to build a plot **step by step** using the same structured thinking that you use to design an analysis, reducing the distance between a plot in your head and one on the page.

# ggplot 's composition I

Any plot that we can create using `ggplot2` share the same composition. We have different elements composing a plot:

- 1 **data** to visualize and a set of **aesthetic mappings** describing how variables in the data are mapped to aesthetic attributes that you can perceive.
- 2 **layers** made up of:
  - geometric elements (**geom**): what you actually see on the plot (points, lines, polygons, ...)
  - statistical transformations (**stats**): summarize data in many useful ways ( binning, counting observations, linear model, ...)

# ggplot 's composition II

- 3 **scales**: map values in the data space to values in an aesthetic space (color,size,shape,...).
- 4 **coord**: coordinates system (Cartesian coordinates, polar coordinates and map projections)
- 5 **faceting**: describes how to break up the data into subsets and how to display those subsets as small multiples (also known as conditioning or latticing/trellising)
- 6 **theme**: controls the finer points of display (font size, background colour,...)

# Some considerations

We stress that `ggplot2`

- is highly flexible (one can easily create new by combining existing elements or even creating new ones)
- produces very nice plots and it is not difficult to learn
- has a great and active community.


However it has its own limitation

- there are cases in which other packages provides easier solutions (e.g. `igraph` for network data)
- exploratory graphs don't have to be pretty and base are usually faster
- although elegant grammar of graphics is sometimes too strict when we already have something in mind.



# Some references

- ▶ A very good book:

 [HADLEY, W. \(2016\)](#)  
ggplot2: Elegant graphics for data analysis.  
*Springer*

- ▶ Sites on ggplot2:

 <http://docs.ggplot2.org/>

 <http://stackoverflow.com>

 <http://groups.google.com/group/ggplot2>

- ▶ Some useful references for colors:

 <http://tools.medialab.sciences-po.fr/iwanthue/index.php>

 <https://colors.co/104547-4b5358-727072-af929d-d2d6ef>