# *Week 5 : Visualisation and Plotting*[1]

*Ramzi Saouma*

*June 21, 2020*

Making informative visualisation is one of the most important elements of machine learning. It is often done before the modelling phase as a preparatory step to understand the data and it is also used post-modelling to inspect or present the results as evidence for an audience. Excellence in numerical visualisation comprises complex ideas presented to a viewer with clarity, efficiency and integrity.

## *Examining The Data*

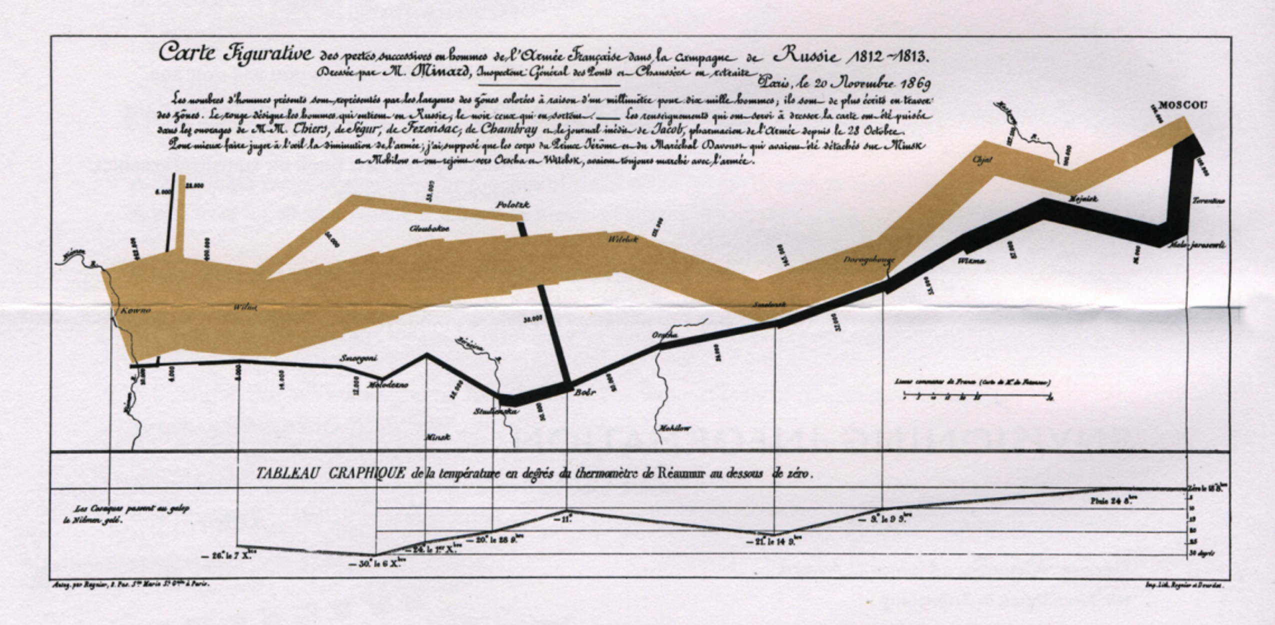
If your goal is to apply machine learning analysis to a data set, it is imperative to look at the data set and apply basic visualization methods. We can begin by a rudimentary scanning of the dataset to look at the various values in the different rows before proceeding. We can think of several reasons for doing this. First, it's helpful to simply get a sense of what's actually in the data set. because it may be the case that by inspecting the features of each object, you might get a better idea of what type of cleaning or pre-processing that is required for this dataset, as well as the range of values or the distribution of values that is typical for each attribute or each feature. This initial exploration can be especially valuable when you are dealing with complex objects, such as text that may be represented by many features that are extracted using several pre-processing steps. For example, you might discover that the data set you got has a single column with person's name that still needs to be split into two separate first and last name columns (this might be important if you are using the name as one of the prediction feature). Second, you might notice missing or noisy data, or perhaps some specific inconsistencies, such as the wrong data type being used for a column. You may also discover incorrect or inconsistent units of measurement for a particular column, particular feature. Perhaps, you might also notice that there are not enough examples of a particular labeled class. For instance, say you are doing a health application with a patient record for each row - some might accidentally have recorded the weight in grams instead of kilograms and so forth. This can make obviously a huge difference in how accurate your results are. So inspecting and visualizing the data will help you detect and understand these potential source of noise or errors. And finally, it might turn out that for your data set, your problem is actually solvable without machine learning. Now, this does not happen all that often, but if it does, you can save yourself considerable time by simply looking at what data

[1] This document has been written in LaTeX using the work of The Tufte-LaTeX Developers which is Inspired by Edward R. Tufte!

exists in your data set. In some scenarios, your data set might actually contain a feature that is clearly a strong indicator of the label that you want to predict.

## *The Fundamental Principles of Analytical Design*

### *Case I: The Russian Campaign. Content Above All*


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869
Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Wixma
Chjat
Mojaisk
Tarutino
Malo-jarosewli
Smorgoni
Molodezno
Minsk
Studienska
Bobr
Orscha
Mohilow
422.000
400.000
6.000
22.000
60.000
33.000
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
Pluie 24 8.bre
Zéro le 18 8.bre
— 26° le 7 X.bre
— 30° le 6 X.bre
— 24° le 1er X.bre
— 20° le 28 9.bre
— 11°
— 21° le 14 9.bre
— 9° le 9 9.bre

*Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812-1813.* `https://upload.wikimedia.org/wikipedia/commons/2/29/Minard.png`

Introduction- The above statistical map, drawn by Minard, reveals the terrible fate of Napoleon's army during the Russian campaign.[2] The drawing combines data map and time series with a lot of eloquence and clarity. Despite being two centuries old, this maps is considered to be one of the best statistical graphics ever [3]. In his famous drawing, Minard describes the sequence of the losses in men of the French army during the French conquests of Russia in 1812. The title and the captions on top of the drawing introduces the type (figurative map) and the context (French invasions). After introducing himself, Minard explains the color code and the 3 scales of measurement. According to Tufte [4], Minard's maps "exemplifies many of the fundamental principles of analytical design.

Comparison- "The fundamental analytical act in statistical reasoning is to answer the question 'Compared to what?' Whether we

[2] Charles Joseph Minard ( 27 March 1781 – 24 October 1870) French engineer, famous for his contribution to information graphics.

[3] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7

[4] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7

are evaluating changes over space or time, searching big data bases, adjusting and controlling for variables, designing experiments, specifying multiple regressions, or doing just about any kind of evidence-based reasoning, the essential point is to make intelligent an appropriate comparisons. Thus visual displays, if they are to assist thinking, should show comparisons".[5] Minard's drawing leverages few revealing comparisons. The most obvious one is the size of his army that is dramatically dropping as it advances in space and time passes while temperature is dropping.

CAUSALITY, MECHANISM, STRUCTURE, EXPLANATION-How did the Grand army lose 98 percent of its forces? The drawing specifies the locations of the large drops, but it is not specific when it comes to the reasons. Drawing the temperature during the retreat is alluding to the severe cold as a major reason for defeat. We also notice large drops around areas where the army had to cross a river such as the Kowno river.

MULTIVARIATE-Minard's piece, draws on 6 times variables. Size of the army, the two spacial coordinates, the direction of the army, the temperature and finally time. Minard's multivariate drawing shows six dimensions with a lot of clarity and ease. Show multivariate data; that is, show more than 1 or 2 variable.

INTEGRATION OF EVIDENCE- Minard brings together various modes of information in order to describe troop movements and war consequences. The author uses the Niemen river as a benchmark to contrast the size of the army at the beginning versus at the end of the campaign. He does not shy at leaving a mocking remark referring to the Cossack horsemen. His choice of color is also by design when you think about the notes he leaves on those of the soldiers flow. "Words, numbers, pictures, diagrams, graphics, charts, tables belong together. Excellent maps, which are the heart and soul of good practices in analytical graphics, routinely integrate words, numbers, line-art, grids, scales."[6] Hence it is always a good practice to annotate your visuals with words and numbers to reveal what's going on. Images and art can also deliver a punchy message on top of your data.

DOCUMENTATION- The credibility of the evidence you are presenting in your visualisation is highly depending on the data sources. Documentation is an essential mechanism of quality control for displays of evidence. Thus authors must be mentioned, sponsors noted, interests and agenda revealed, sources described, dates annotated and scales

[5] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7

[6] Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7

labeled.

CONTENT COUNTS MOST OF ALL- Analytical presentations ultimately stand or fall depending on the quality, relevance and integrity of their content. In this regard, the first question you need to ask yourself before engaging into analytical displays is not "which color should i use?" but "what are the content and reasoning tasks this display is supposed to achieve?".

*Case II: Defense in South America - Forms and Functions* [7]*:*

"...What does the designer want me to do with this graphic? In other words: If we accept that an info-graphic is, at its core, a tool, then what task is this intended to help me with? Here is my personal list for the Brazilian defense graphic:

1. The graphic must present several variables—armed forces personnel, population to be defended, defense budget, and so forth—so that I have the proper information in front of me.
2. It should allow comparisons. At a glance, I should be able to tell which country has the biggest and the smallest army, is more or less populated, or invests more heavily or lightly in its military.
3. It should help me organize countries, from the biggest to the smallest, based on the variables and the comparisons.
4. It should make correlations or relationships evident to me. For instance, are population and size of defense forces directly and perfectly proportional?

Of these four possible tasks: present, compare, organize, correlate, the graphic accomplishes just one satisfactorily. It presents tons of variables and values. However, it does not show them in proportion to one another. This makes it impossible for readers to interpret the data..."[8]

*Case III: World Biggest Banks- The Bubble Trick* [9]

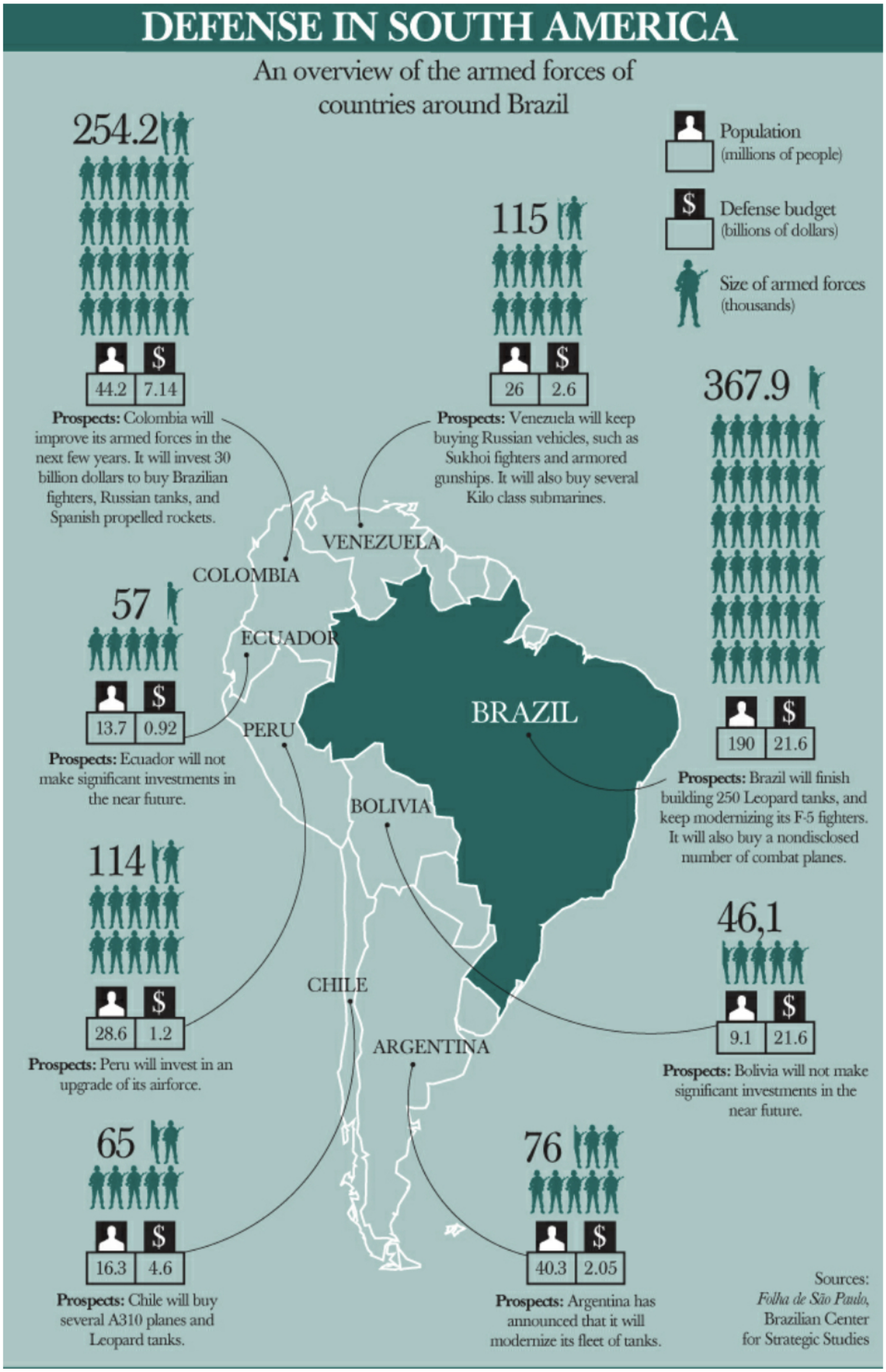
The data in this figure 3 is based on a Bloomberg news project. The aim of the visual is to show the impact of the financial crisis on the capitalisation of the biggest international banks. The initial reaction of the reader might be "I do not see a problem. It is evident that all the big banks took a big hit on their capital."

That is correct. However let's take a close look at one of the banks, Société Générale (figure 2.9) . What is the percentage decrease in their capital between January 2007 and January 2009? The majority

[7] Alberto Cairo. *The Functional Art*. New Riders, Berkeley, California, 2013. ISBN 978-0-321-83473-7

[8]

[9] Alberto Cairo. *The Functional Art*. New Riders, Berkeley, California, 2013. ISBN 978-0-321-83473-7



**Prospects:** Colombia will improve its armed forces in the next few years. It will invest 30 billion dollars to buy Brazilian fighters, Russian tanks, and Spanish propelled rockets.

**Prospects:** Venezuela will keep buying Russian vehicles, such as Sukhoi fighters and armored gunships. It will also buy several Kilo class submarines.

**Prospects:** Ecuador will not make significant investments in the near future.

**Prospects:** Brazil will finish building 250 Leopard tanks, and keep modernizing its F-5 fighters. It will also buy a nondisclosed number of combat planes.

**Prospects:** Peru will invest in an upgrade of its airforce.

**Prospects:** Bolivia will not make significant investments in the near future.

**Prospects:** Chile will buy several A310 planes and Leopard tanks.

**Prospects:** Argentina has announced that it will modernize its fleet of tanks.

Sources:
*Folha de São Paulo,*
Brazilian Center
for Strategic Studies

.

Figure 1: "The Defense of the Neighbors: An overview of the armed forces of countries surrounding Brazil."

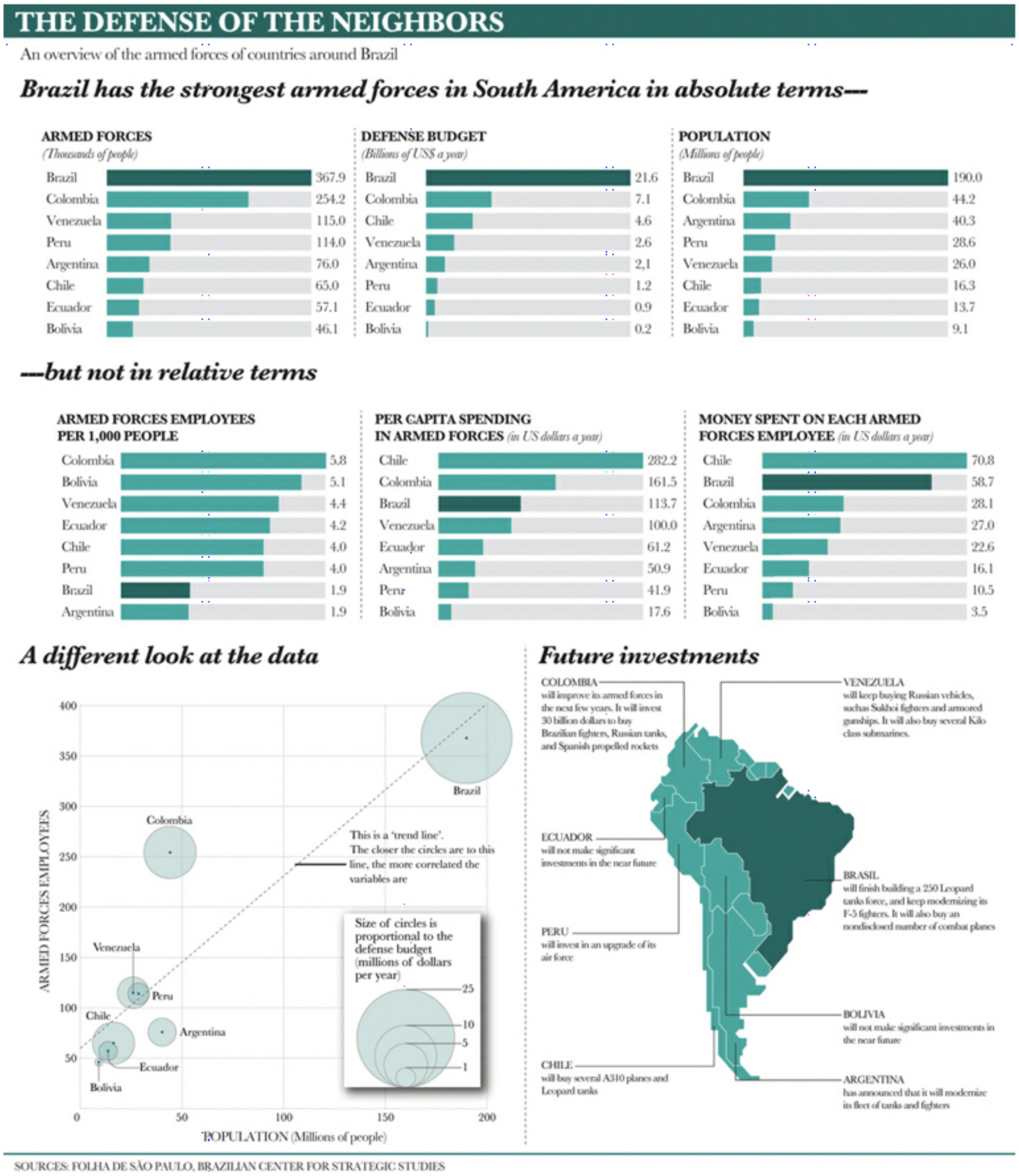
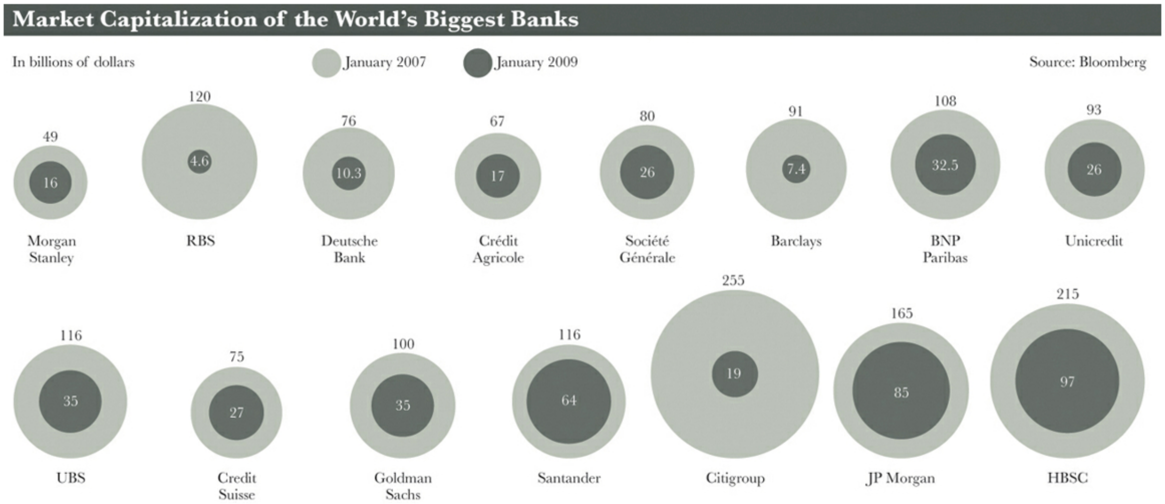


Figure 2: A different take on the defense info-graphic.

of reader will see that drop as less than 50%. What if we visualise the same data using a bar chart as per Figure 2.10. Human brains are not good at calculating surface sizes - they are much better a estimating change in one dimensional space. Hence if we want to see the impact on bank's capital and compare that within the major global banks we are better off using the visual in Figure 2.10. 4.



.

Figure 3: The Financial Crisis impact on Global Banks

## References

Alberto Cairo. *The Functional Art*. New Riders, Berkeley, California, 2013. ISBN 978-0-321-83473-7.

Edward R. Tufte. *Beautiful Evidence*. Graphics Press, LLC, first edition, May 2006. ISBN 0-9613921-7-7.

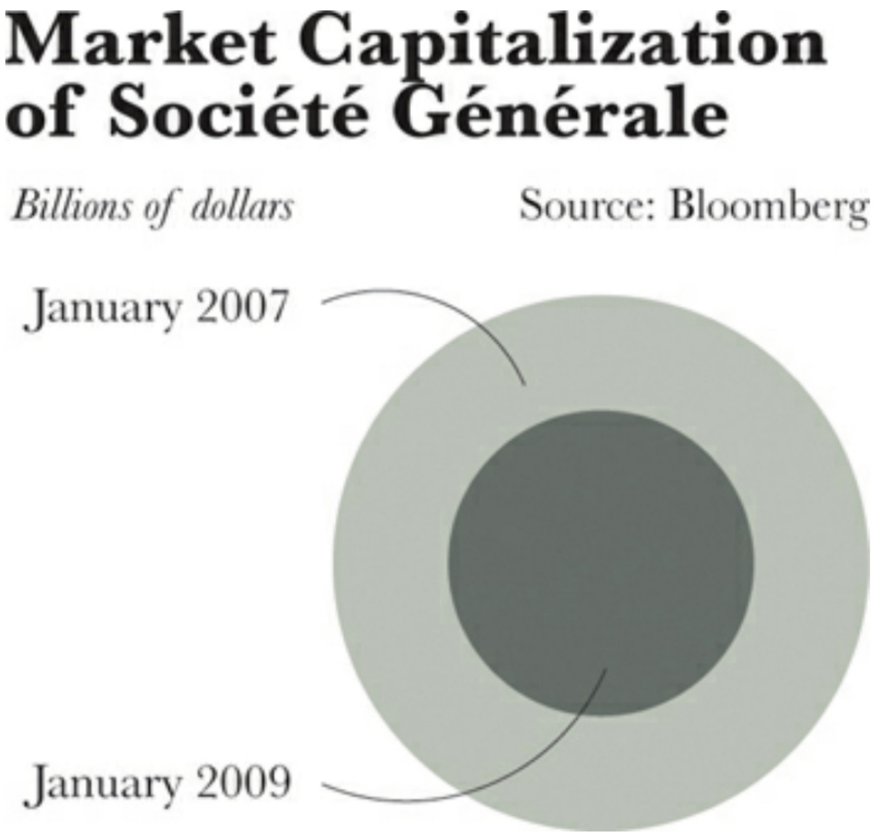


**Figure 2.9. The first bubble represents $80 billion. What percentage of that does the second bubble represents? Half, perhaps?**

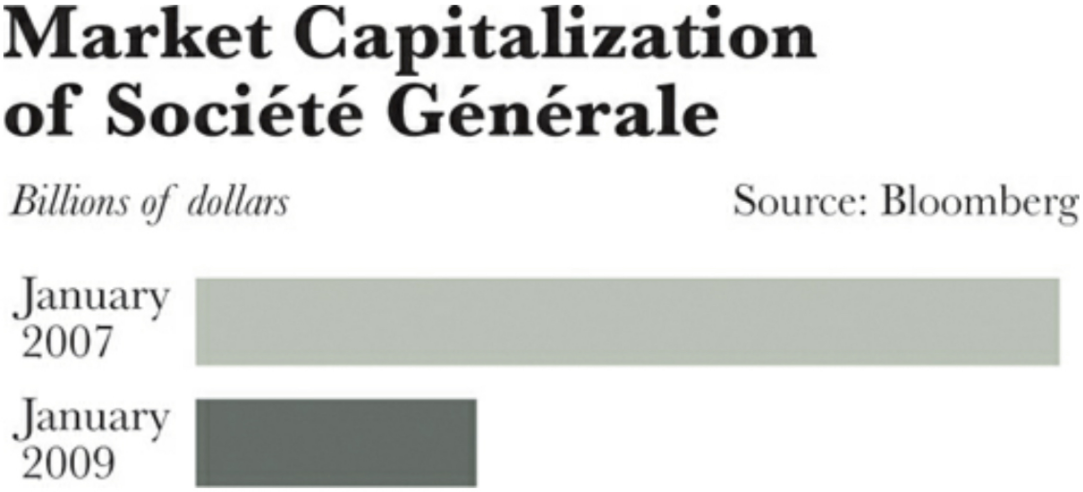
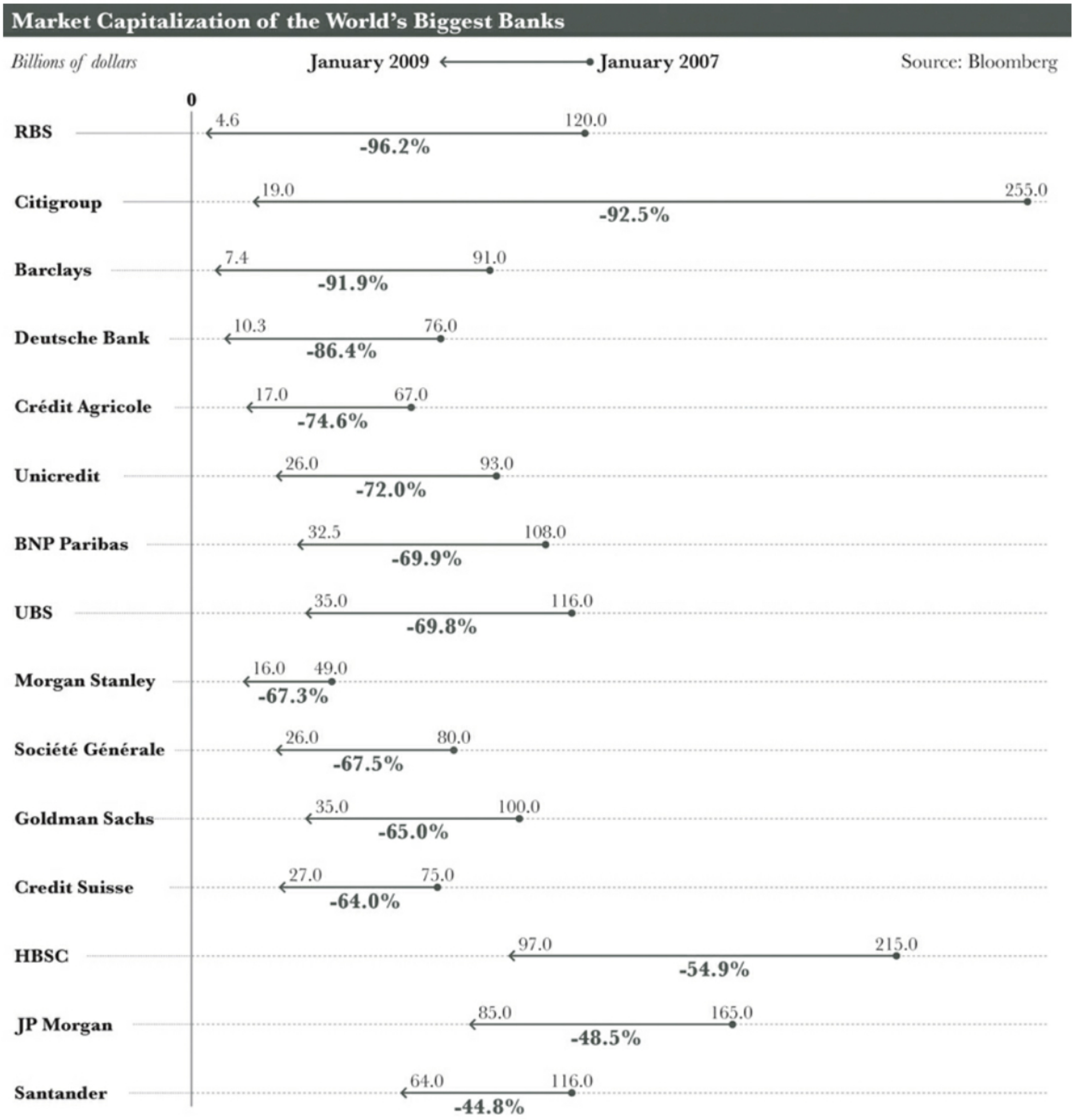


**Figure 2.10. Our friend, the bar chart, comes to the rescue.**

.



.

Figure 4: An example of a visualisation that is more in line with the goals of the author