



UNIVERSITAT  
DE  
BARCELONA



# Data Visualization

Postgraduate Course on DataScience and BigData

Santi Seguí | 2018-2019

# What is data Visualization?

A primary goal of data visualization is to **communicate information clearly, precisely** and **efficiently** to users via the selected information graphics, such as tables and charts.

The main goal of any graph or visualization is to be a tool for your eyes and brain to perceive what lies beyond the data.

# The simplest Data Visualization

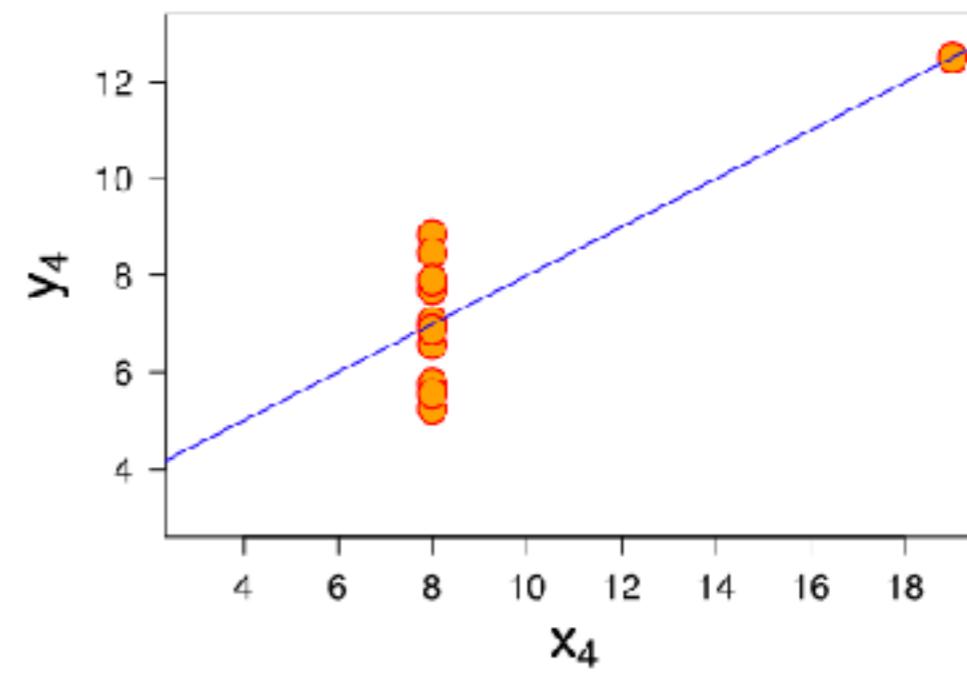
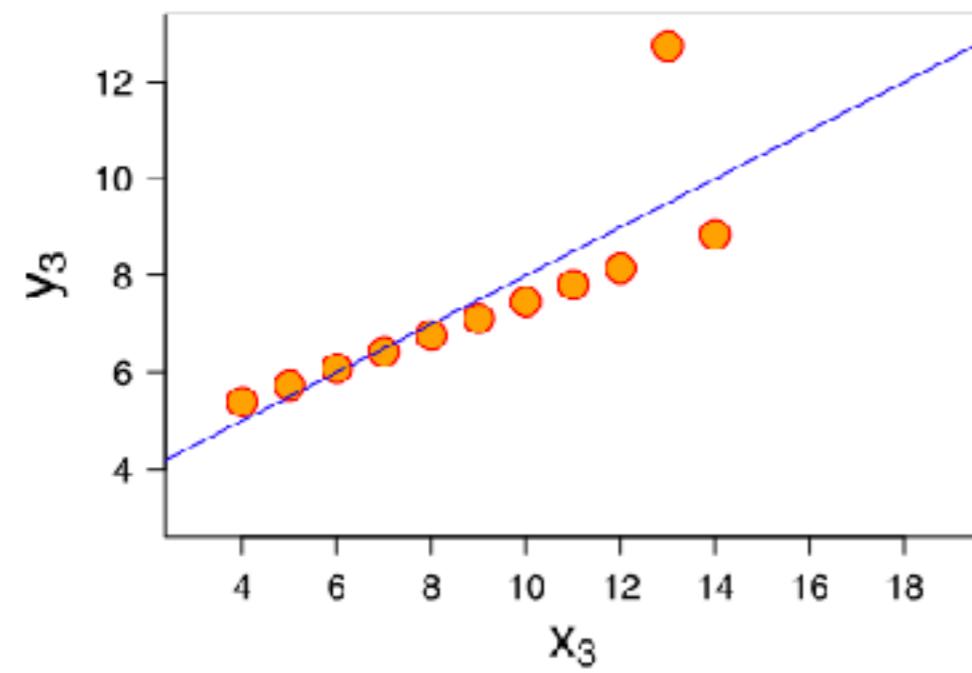
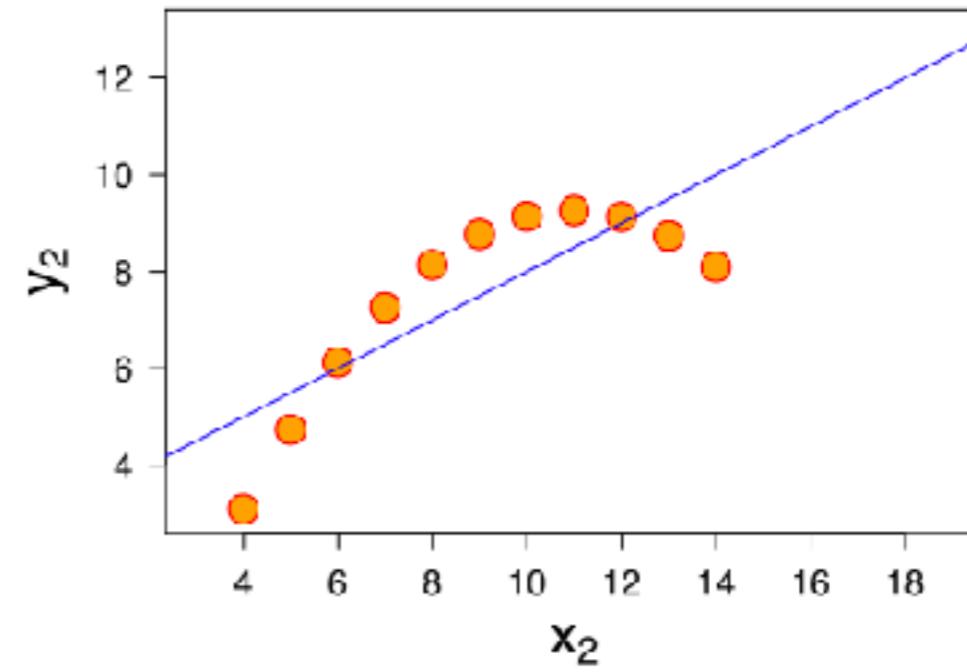
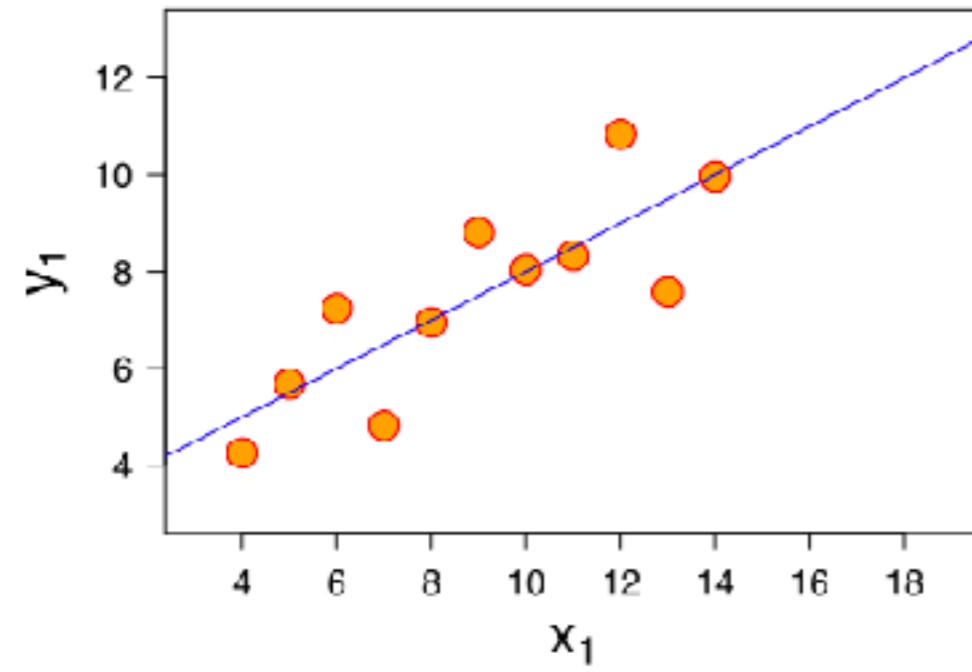
Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

### Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

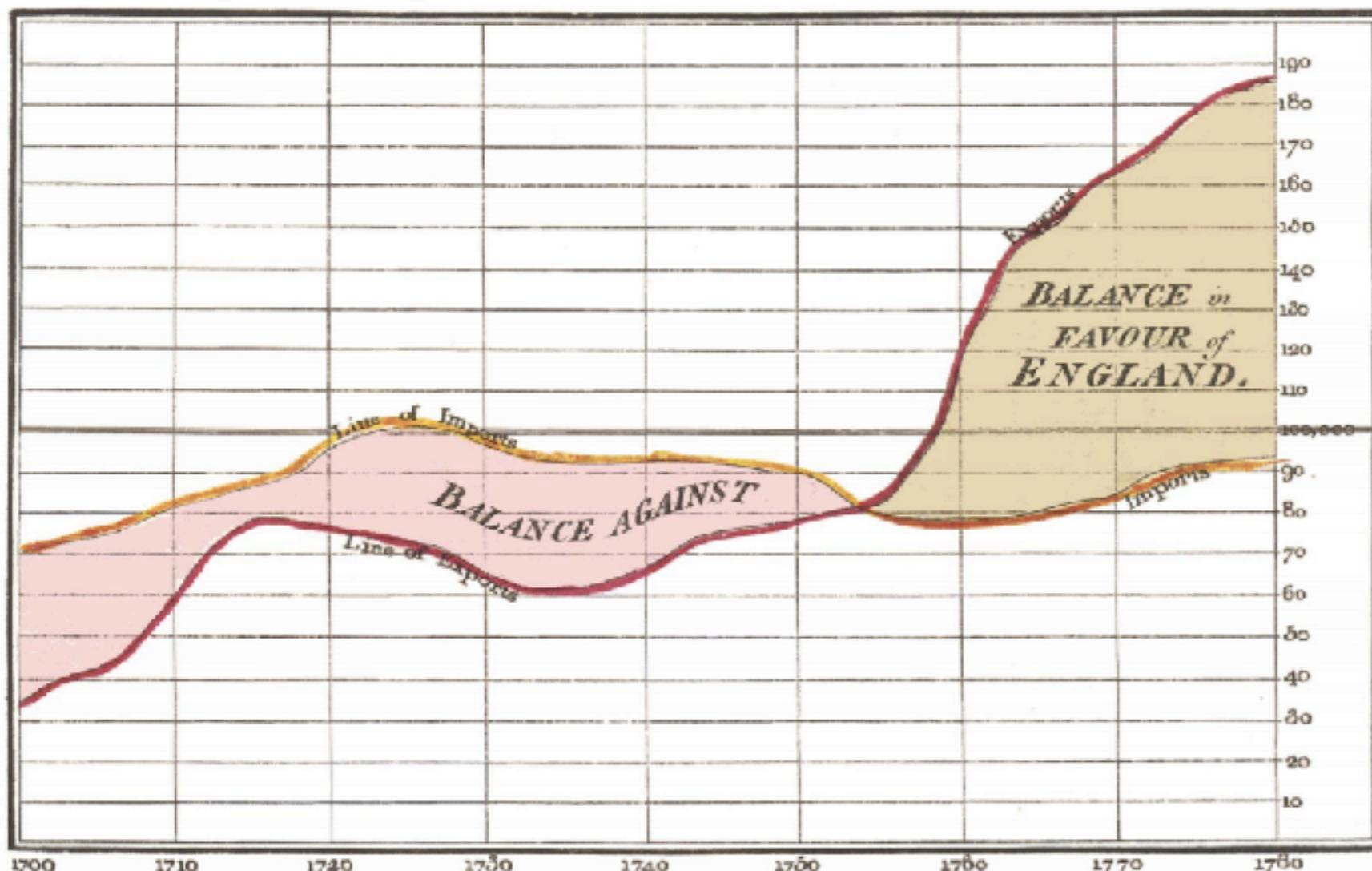
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places



# What is a **good** (or excellent) Visualization?

# William Playfair (1786)

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



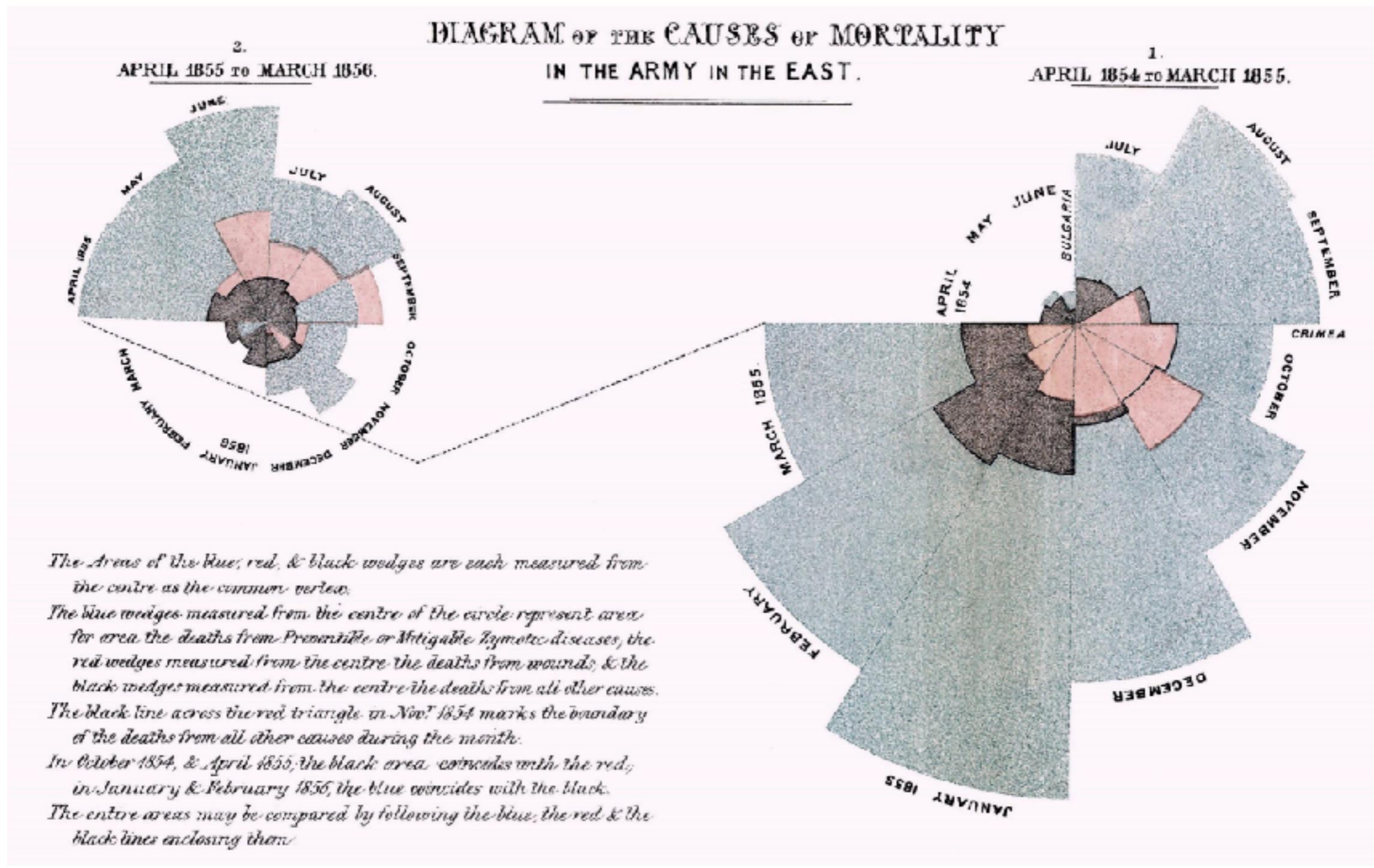
Line, bar, pie charts were all invented by the same person!

Aside from revolutionizing graphics, Playfair was an economist, engineer, and even a secret agent.

(Image: Wikipedia)

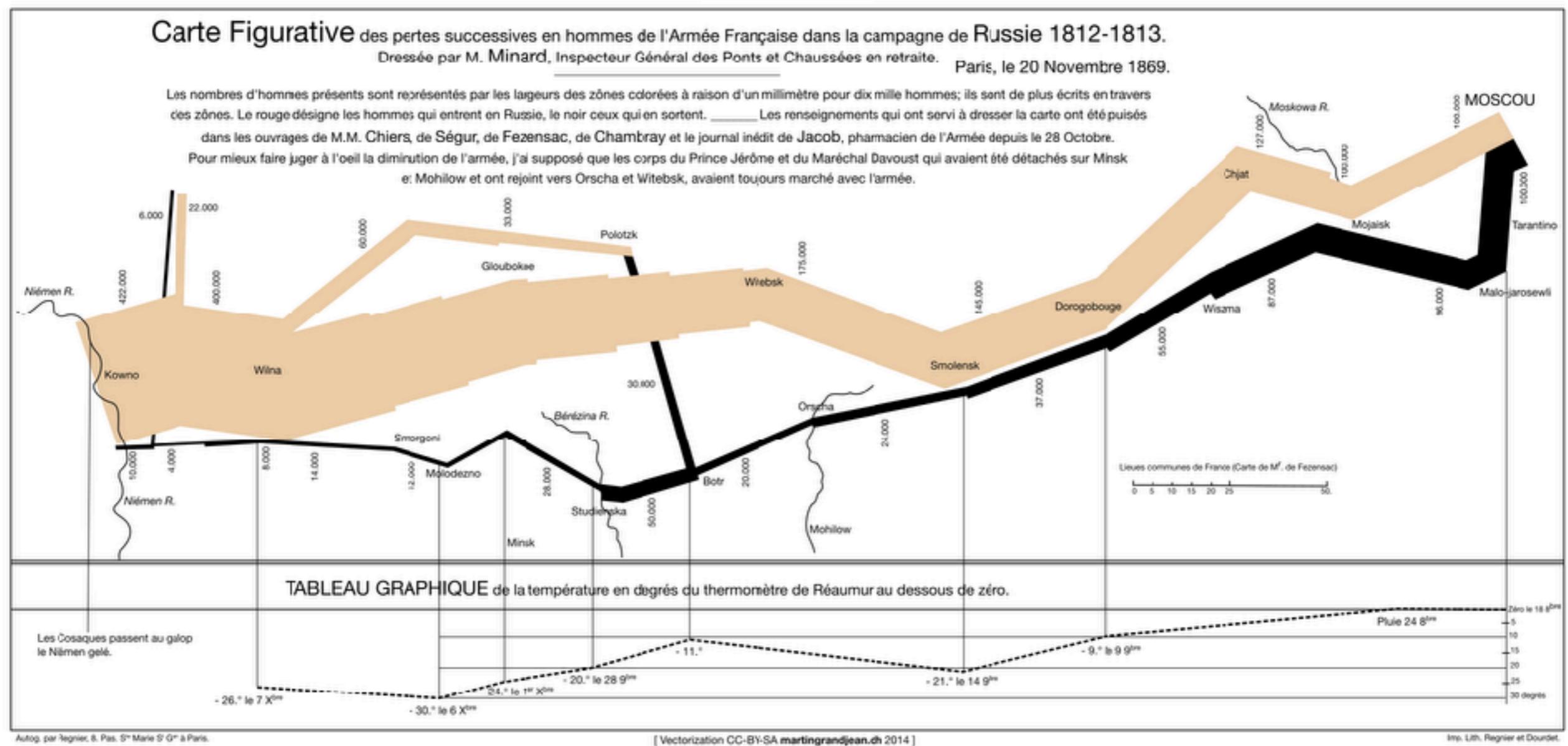
Published at the Act library, 10<sup>th</sup> May 1786, by W<sup>m</sup> Playfair.  
N<sup>o</sup> 100, Strand, London.

# Florence Nightingale (1858)



# Napoleon Invasion by Charles Joseph Minard

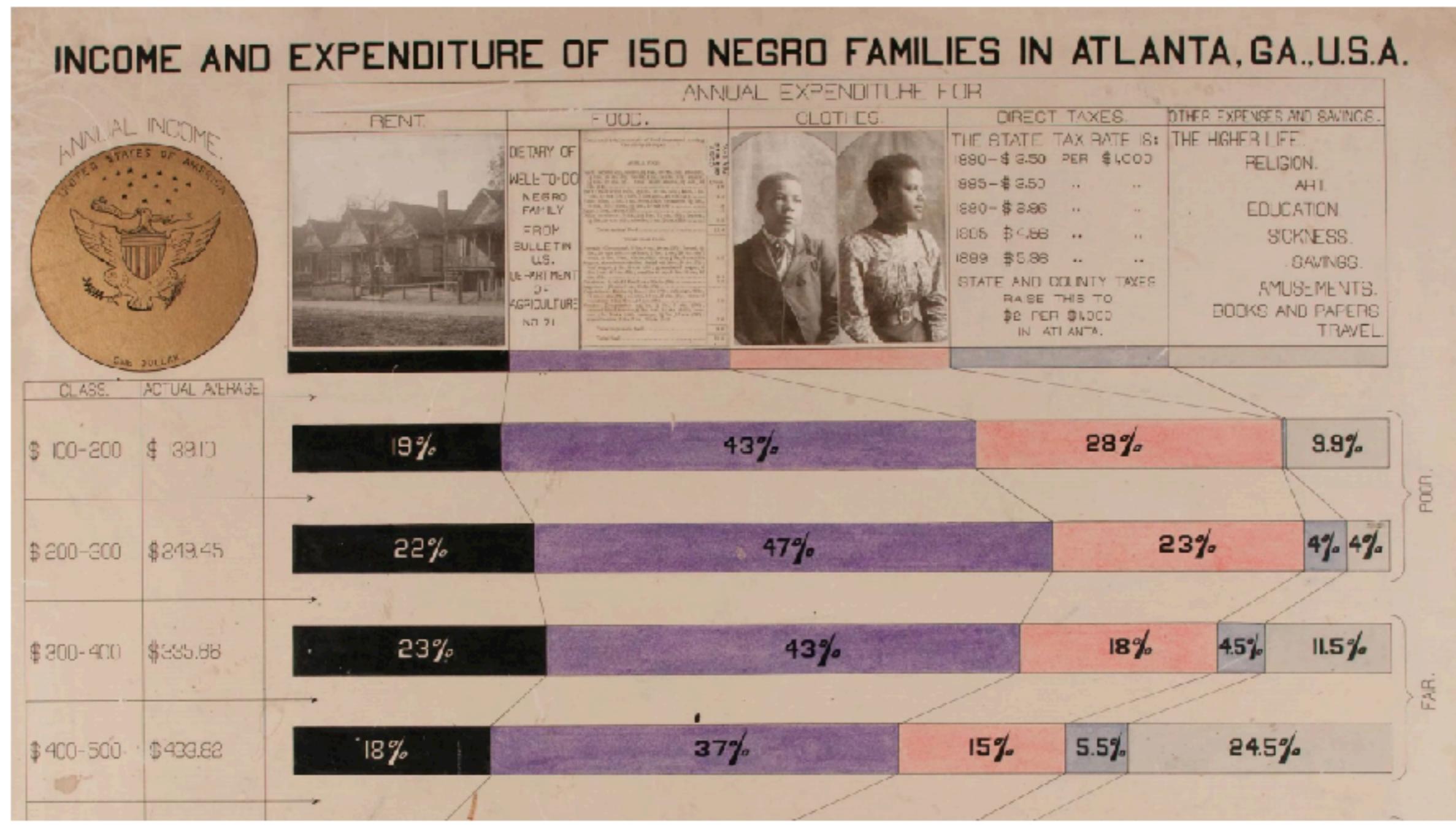
Edward Tufte said it "may well be the best statistical graphic ever drawn"



[http://en.wikipedia.org/wiki/Charles\\_Joseph\\_Minard#mediaviewer/File:Minard.png](http://en.wikipedia.org/wiki/Charles_Joseph_Minard#mediaviewer/File:Minard.png)

Variables displayed in the graph: 1) the number of Napoleon's troops; 2) the distance traveled; 3) temperature; 4) latitude and longitude; 5) direction of travel; and 6) location relative to specific dates

# W.E.B Du Bois (1900)



**Asking the right questions  
& making the right grafics**

# What is a **good** (or excellent) Visualization?

Accuracy — Story Telling — Knowledge Discovery

# Accuracy?

The results have to be accurate!!

**Let's see some examples of what you should not do!**

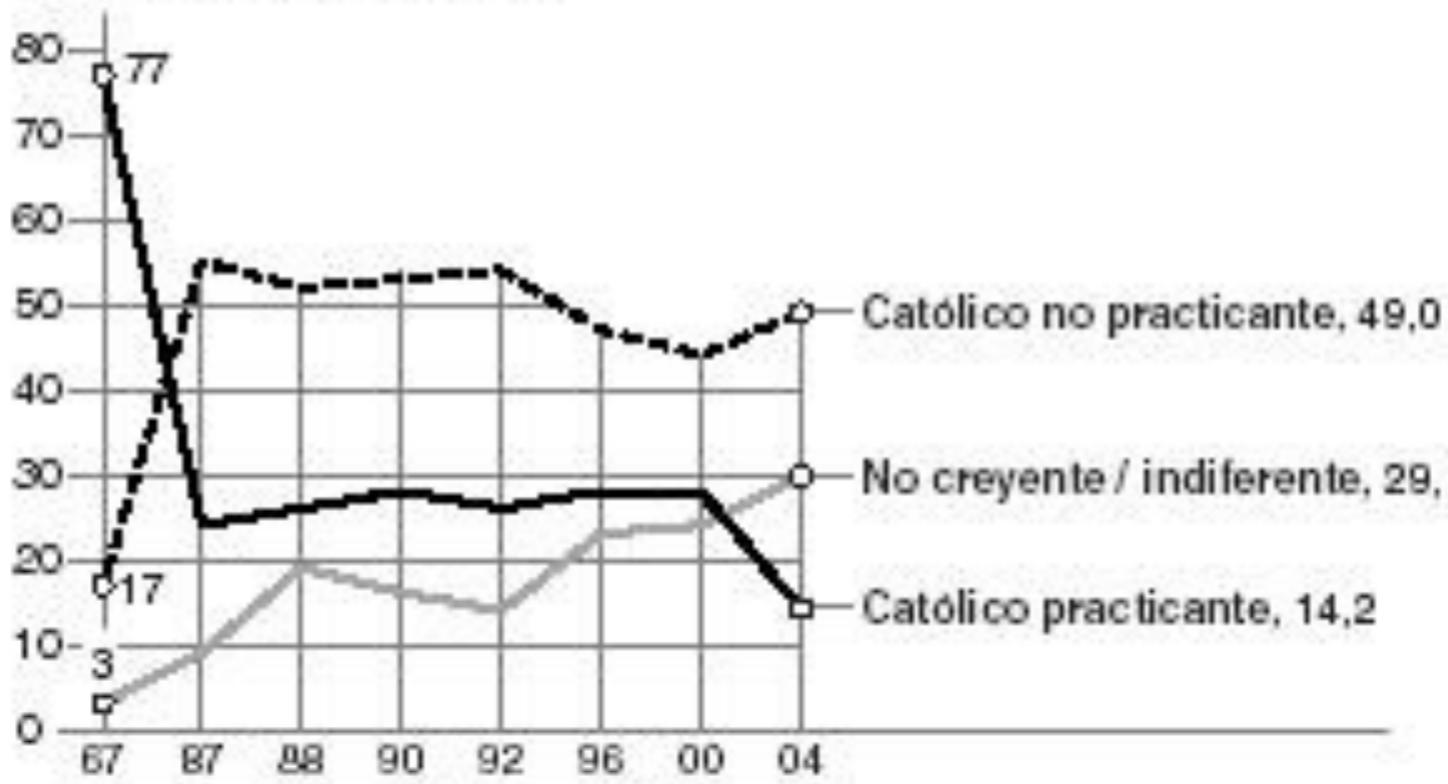
*The following graphs have been obtained from different websites and I use them here as illustrative samples of bad visualizations, however, I did not check the veracity of them.*

# Case 1: What is wrong?

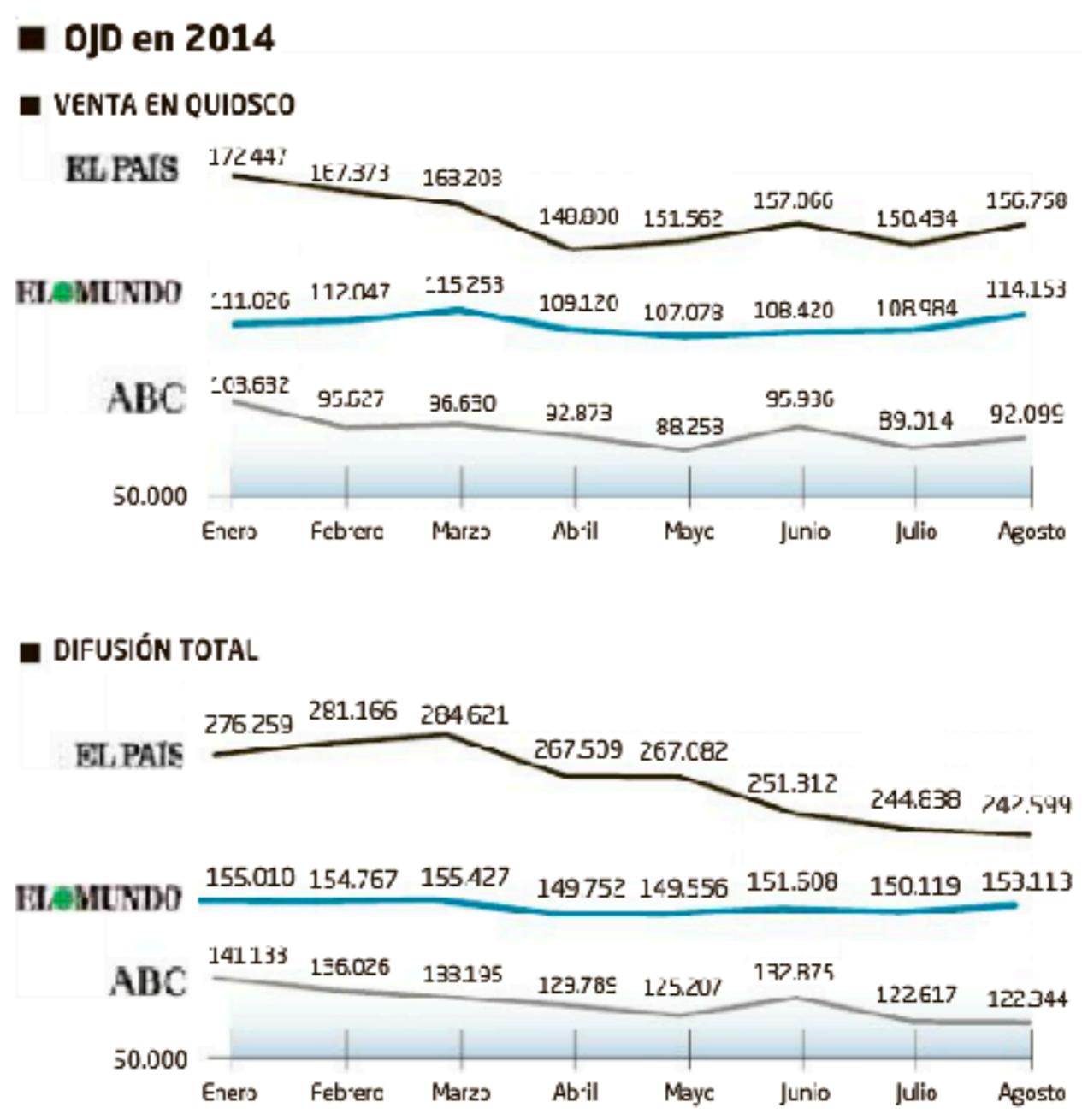
## Encuesta a los jóvenes españoles

En porcentaje.

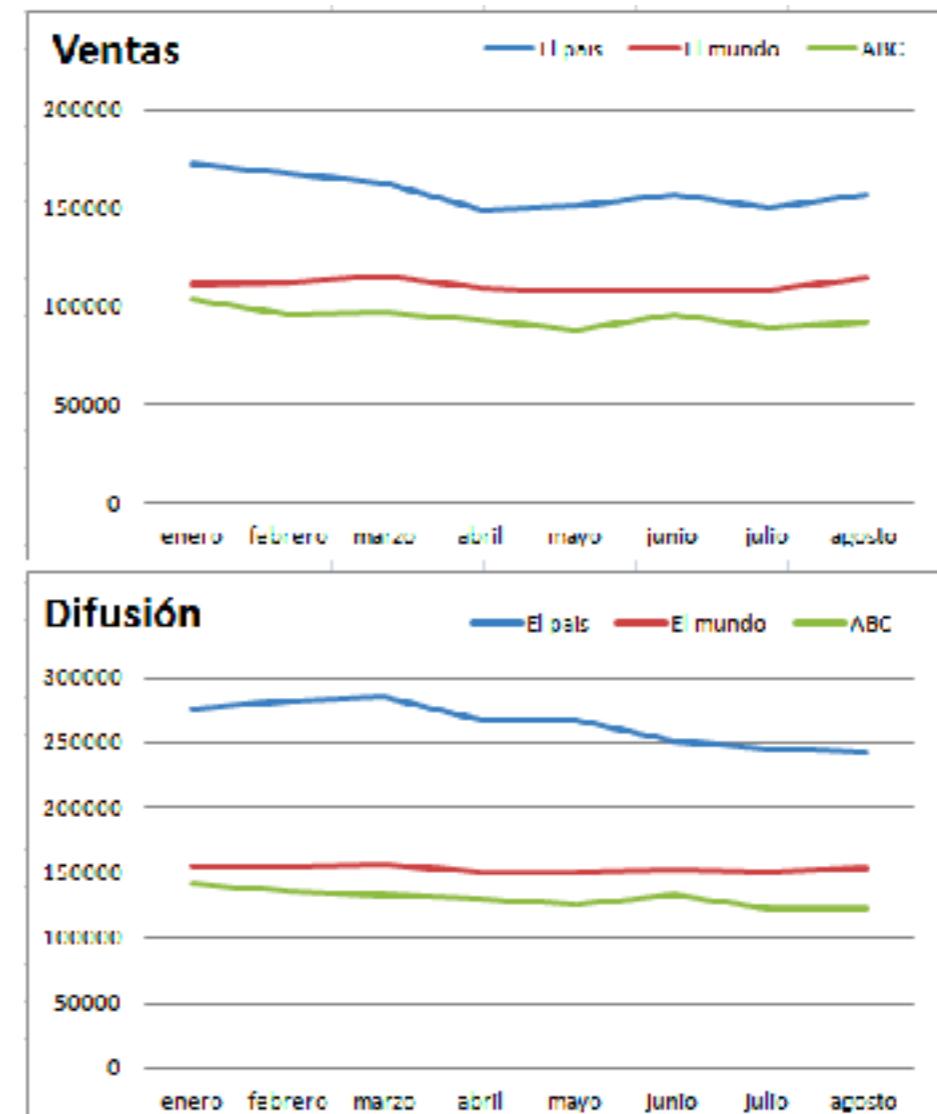
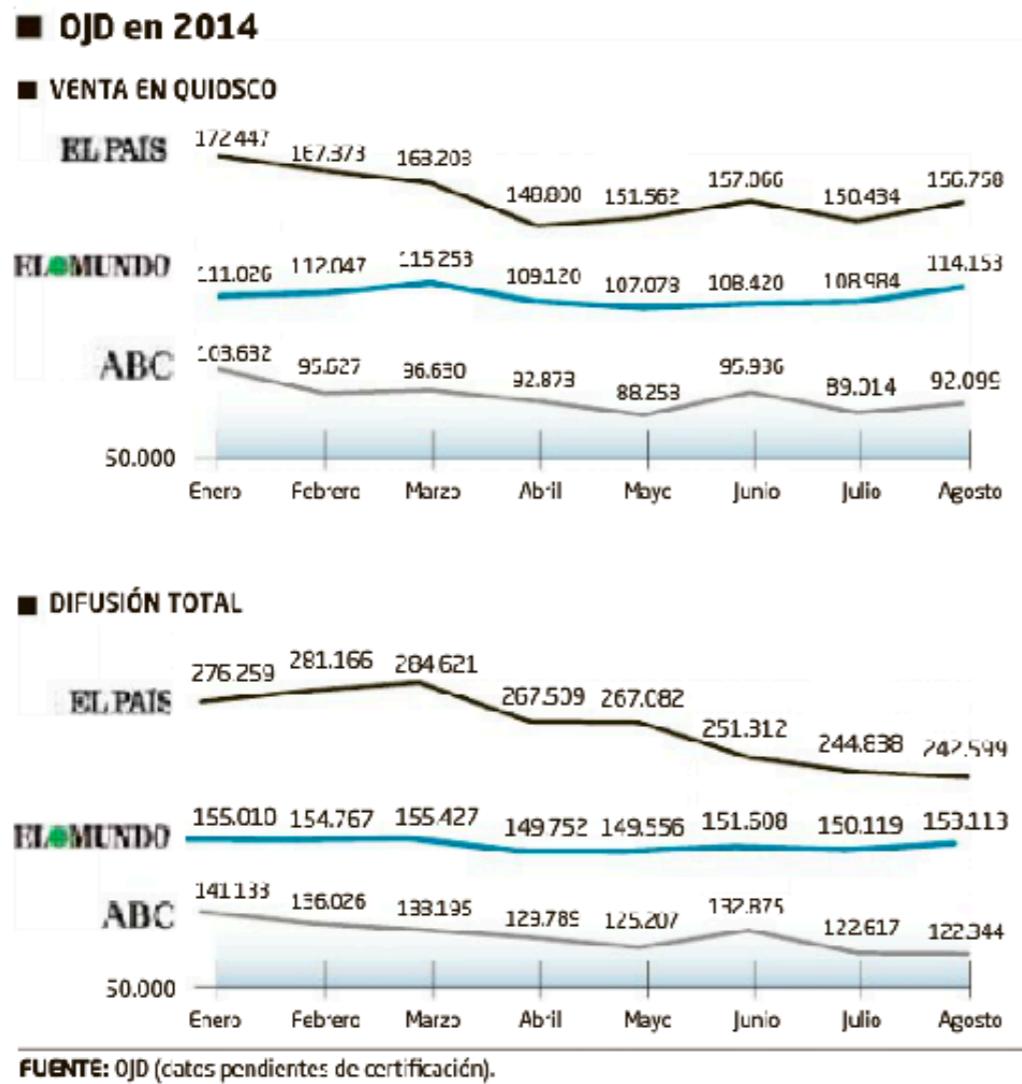
- Evolución de la identificación religiosa de los jóvenes  
Población entre 15 y 29 años.



# Case 2: Newspapers sells



# Case 2: Newspapers sells



# Case 3: Hotel Prices

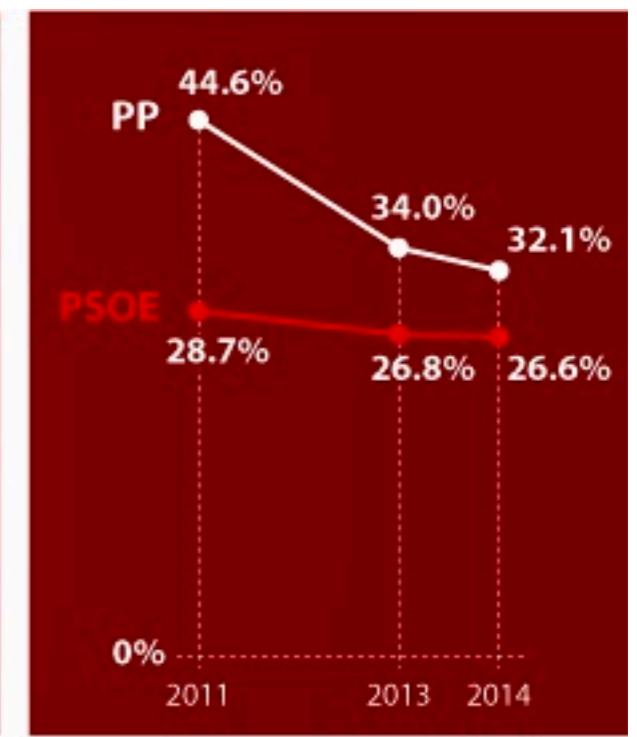


"Los precios de los hoteles bajan el 94% en Nueva York. La caída más acusada en el sector hotelero de todo el mundo se registró en Nueva York, con un desplome del 93,57 por ciento en sus tarifas, hasta los 154 euros de media por noche, frente a los 218 euros del mismo mes de 2008."

# Case 4: Y-Axis



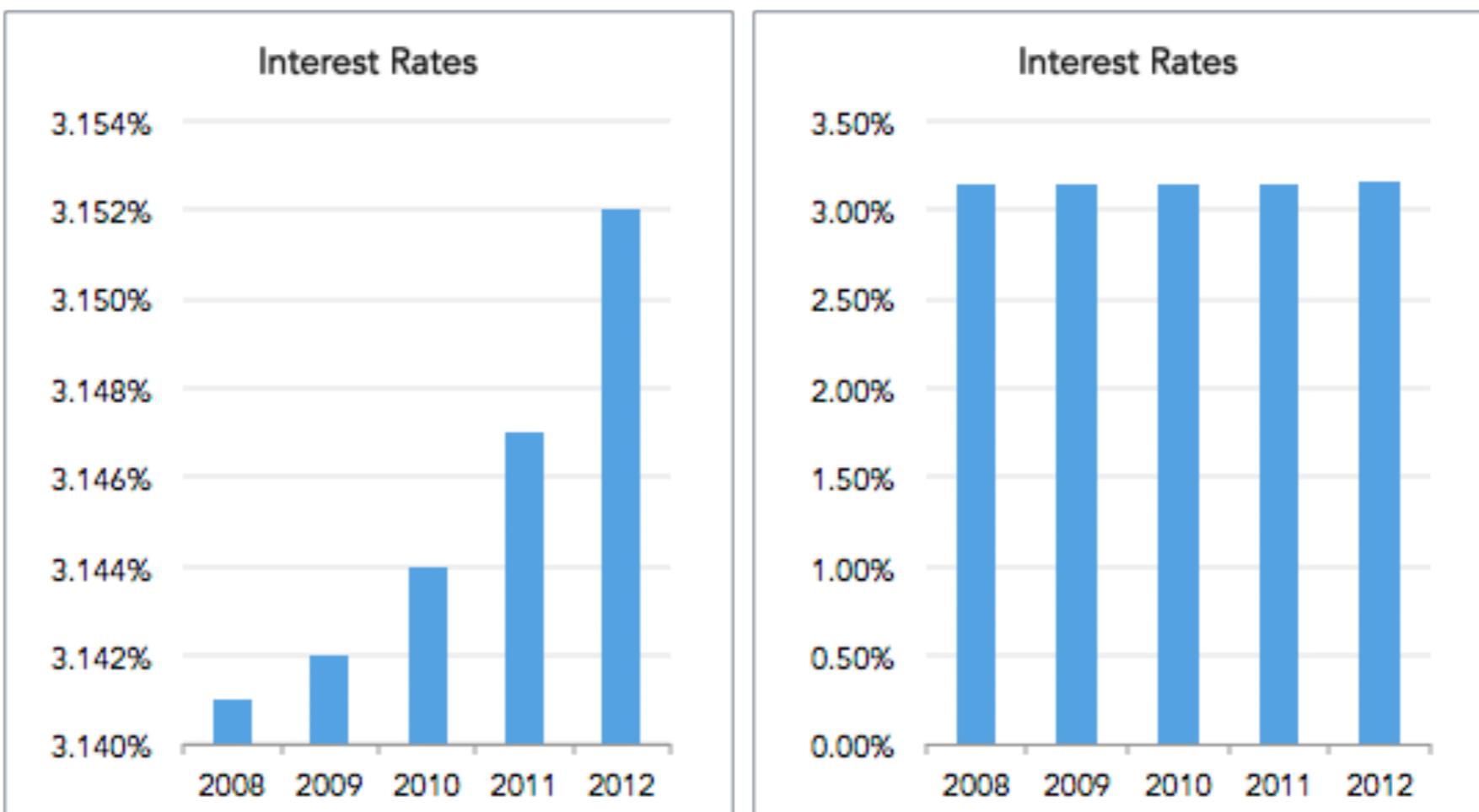
This is a lie



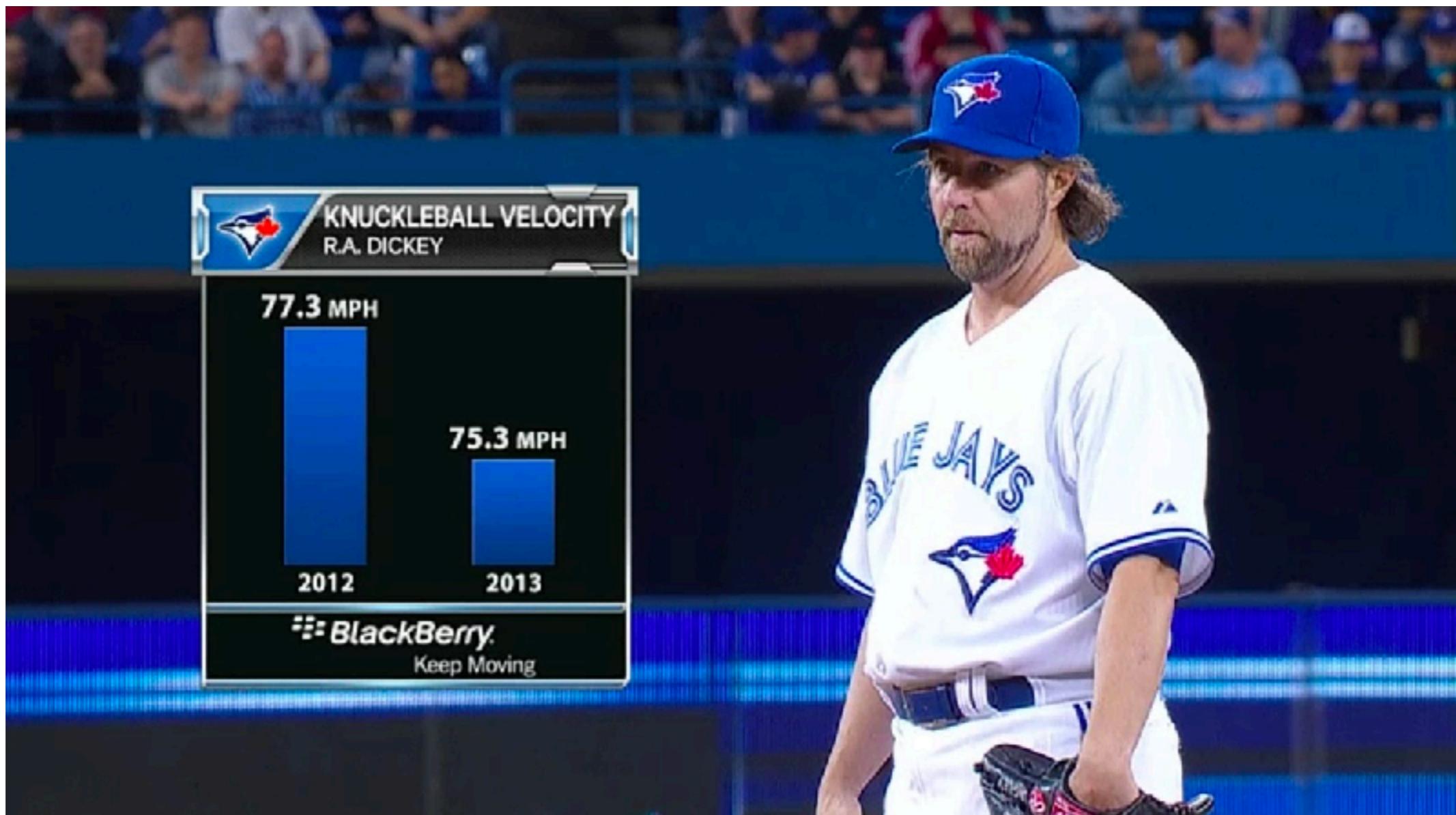
This is the truth

# Case 5: Y-Axis

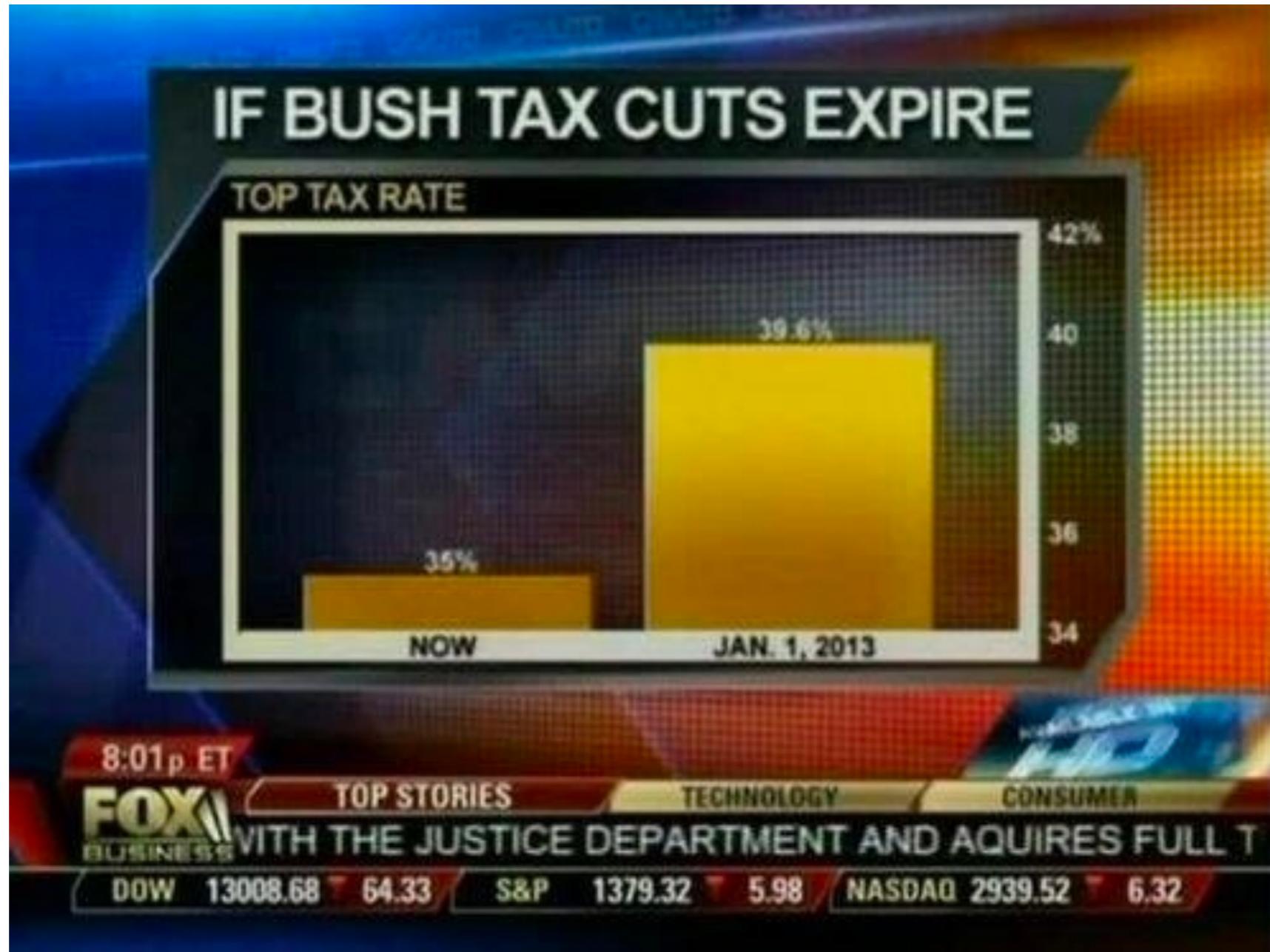
## Same Data, Different Y-Axis



# Case 5: Y-Axis

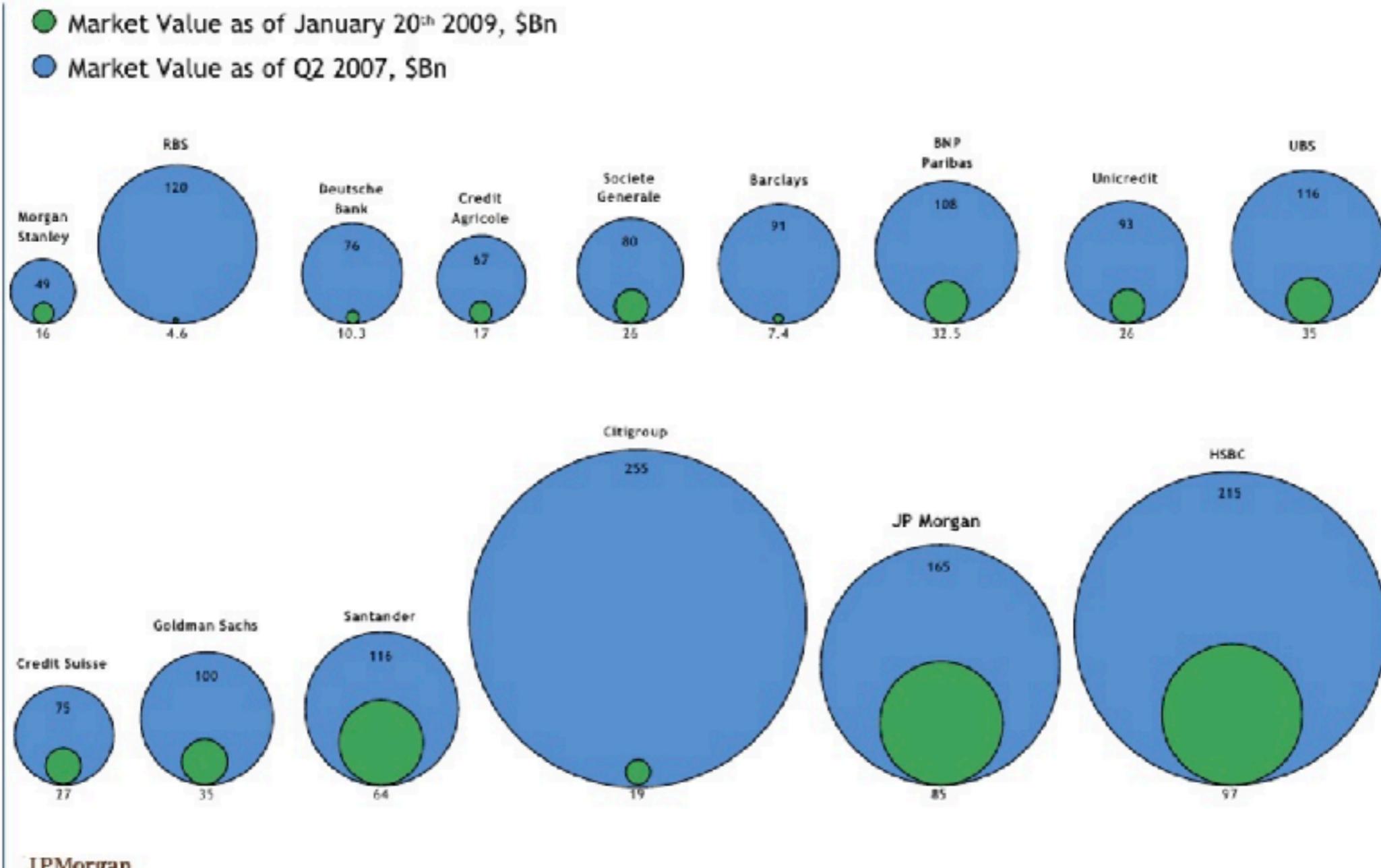


# Case 5: Y-Axis



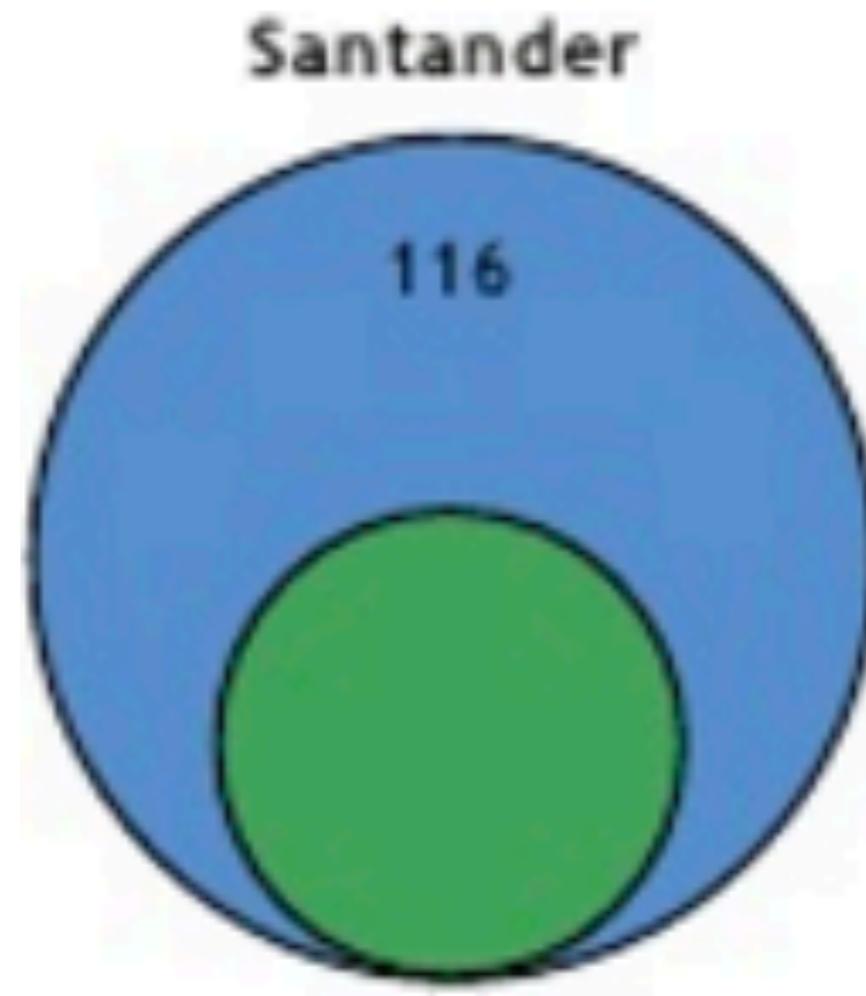
# Case 6: The bubble Plague

## Banks: Market Cap (the Original Slide)



This display compares the market capitalization (what it would cost to buy all of a company's stock at the current price) of 15 major banks as of January 20th, 2009 to their market capitalization in Q2 2007, before the banking crisis hit. People are supposed to compare the small green circles to the larger blue circles to see how much the market capitalization of each bank declined.

# Case 6: The bubble Plague

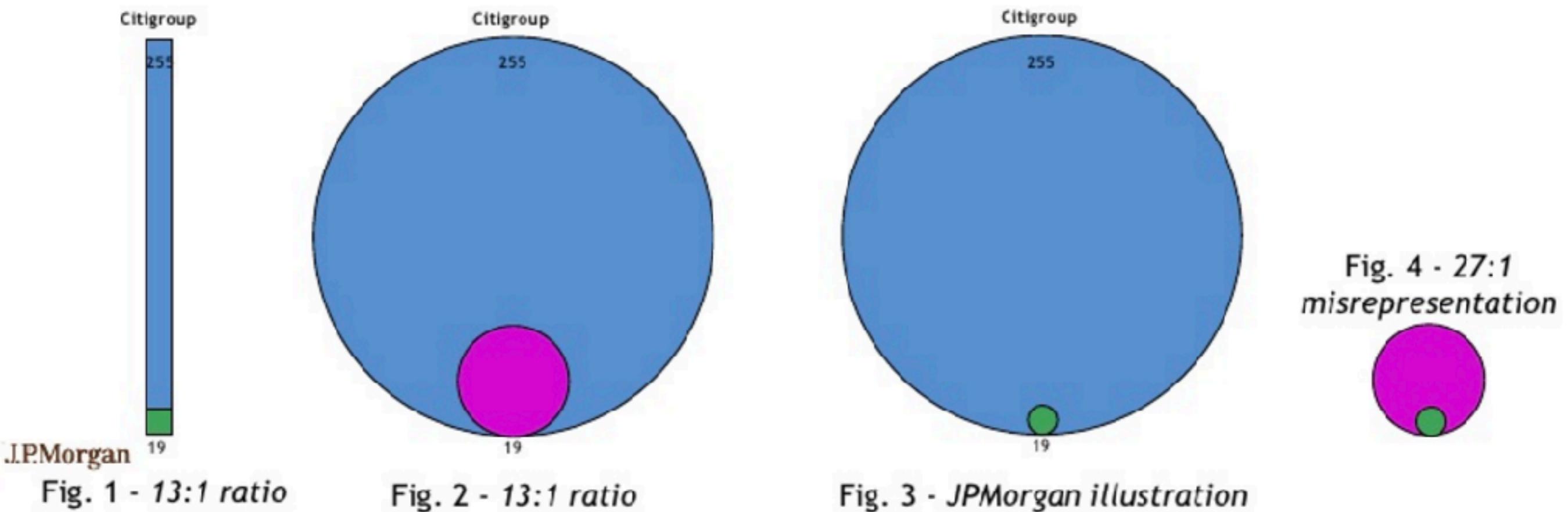


Can you say the the percentage of capital loss observed from 2007 to 2009?

Can you say which was the value in 2009 if the value in 2007 was 116?

# Case 6: The bubble Plague

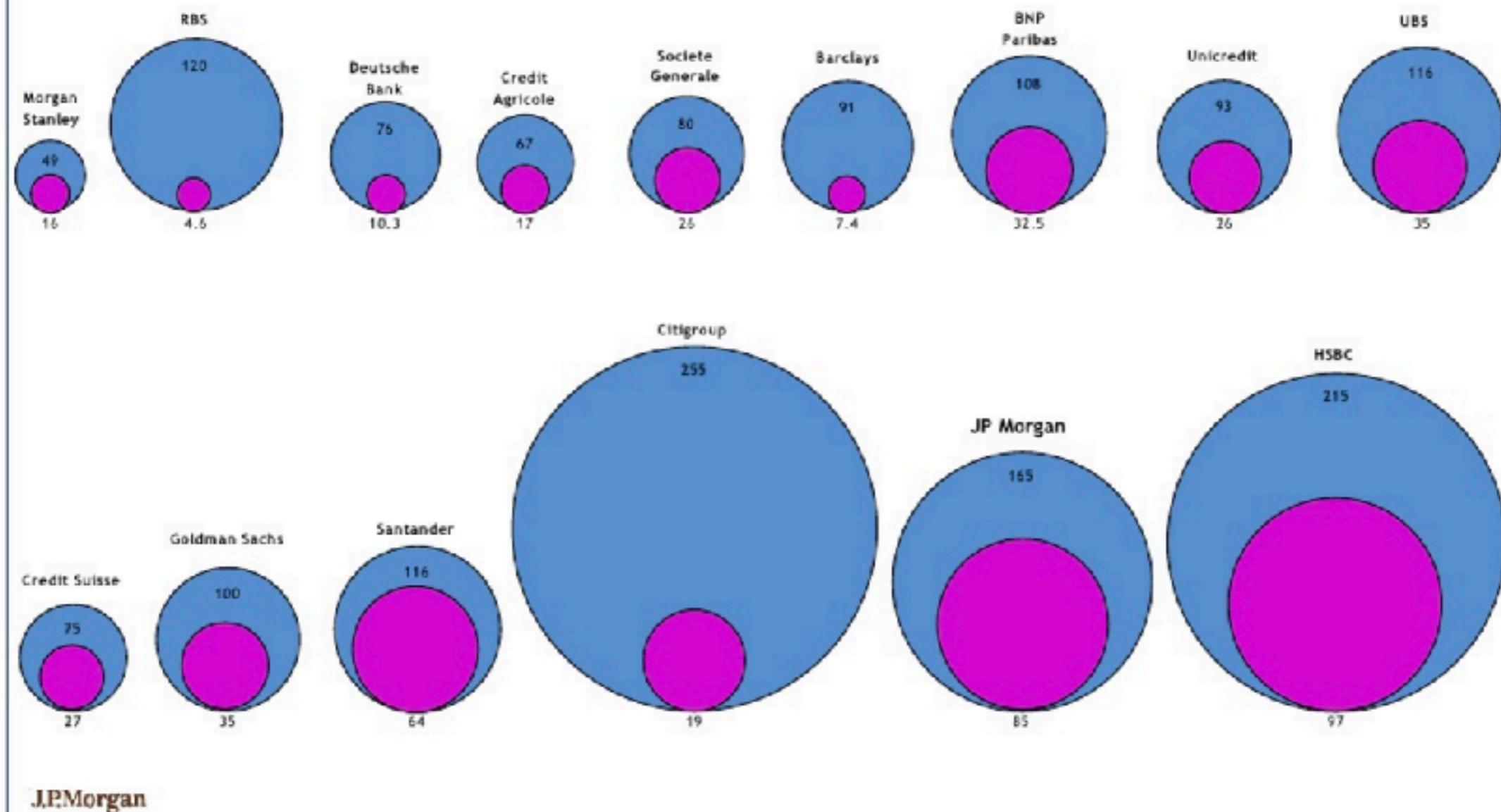
The creators of this graph encoded the values using the diameter of the circles, but when the diameter of the circle changes, the area changes even more quickly, which causes us to dramatically over-estimate the difference in values if we do what comes naturally—attempt to compare their areas.



# Case 6: The bubble Plague

## Banks: Market Cap (Corrected)

- Market Value as of January 20<sup>th</sup> 2009, \$Bn (Honestly Represented)
- Market Value as of January 20<sup>th</sup> 2009, \$Bn (Misrepresented)
- Market Value as of Q2 2007, \$Bn

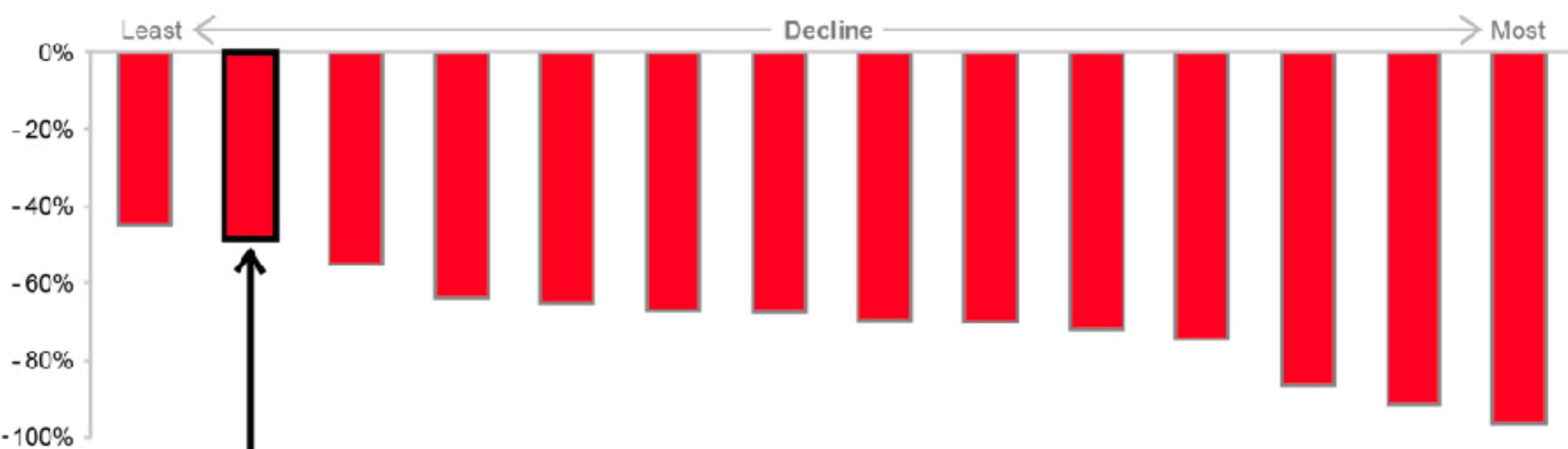
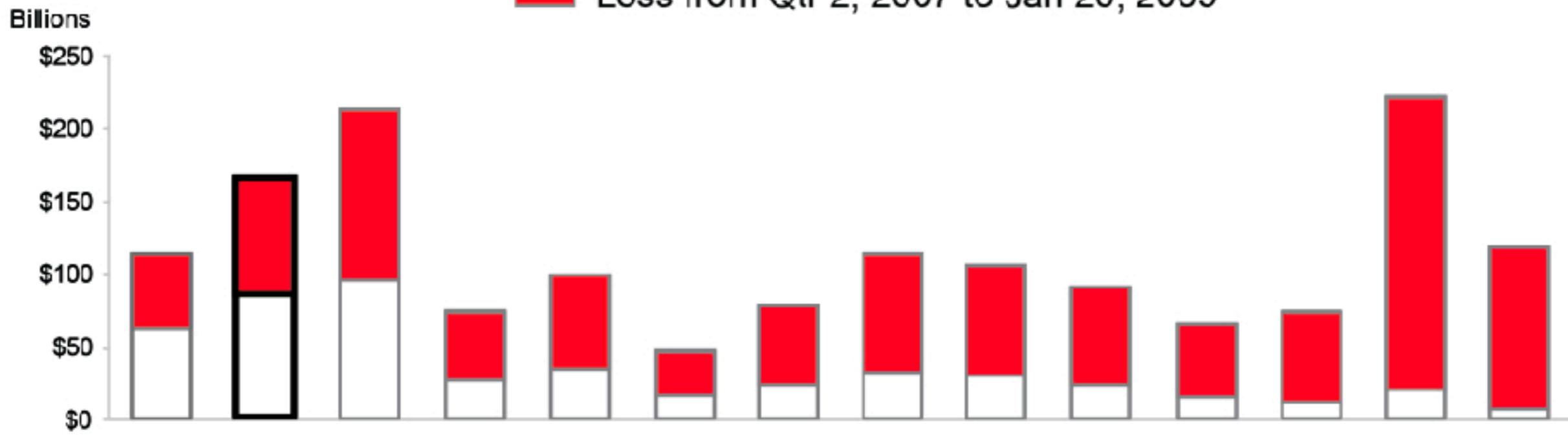


While JPMorgan considers this information to be reliable, we cannot guarantee its accuracy or completeness

Source: Bloomberg, Jan 20<sup>th</sup> 2009

# Declines in Bank Market Values Since the Financial Crisis Began

Loss from Qtr 2, 2007 to Jan 20, 2009



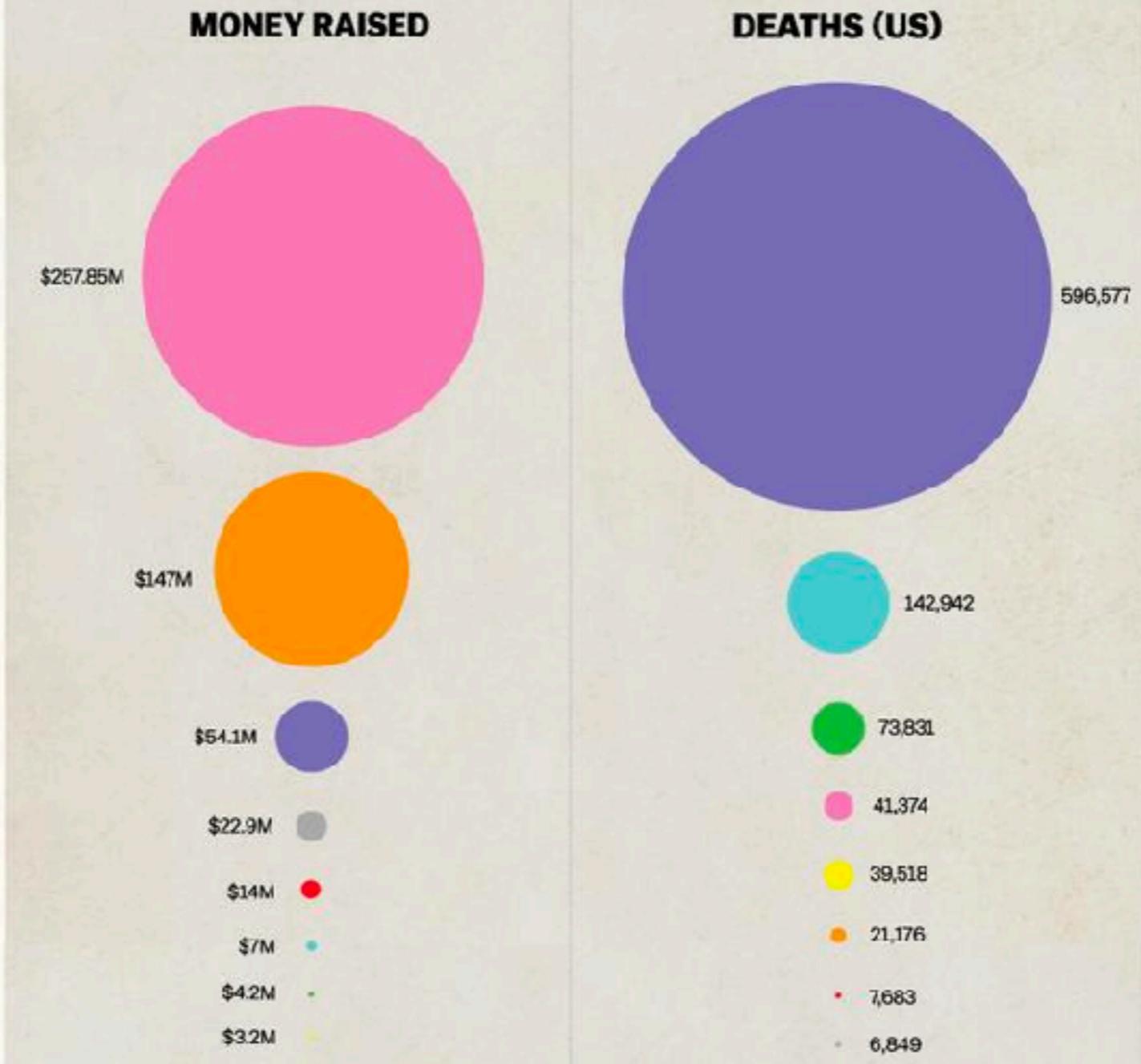
Among major banks, J.P. Morgan had the second least percentage decline in market value.

# WHERE WE DONATE VS. DISEASES THAT KILL US

**Heart Disease**  
Jump Rope for Heart (2013)      **Suicide**  
Out of Darkness Overnight Walk (2014)

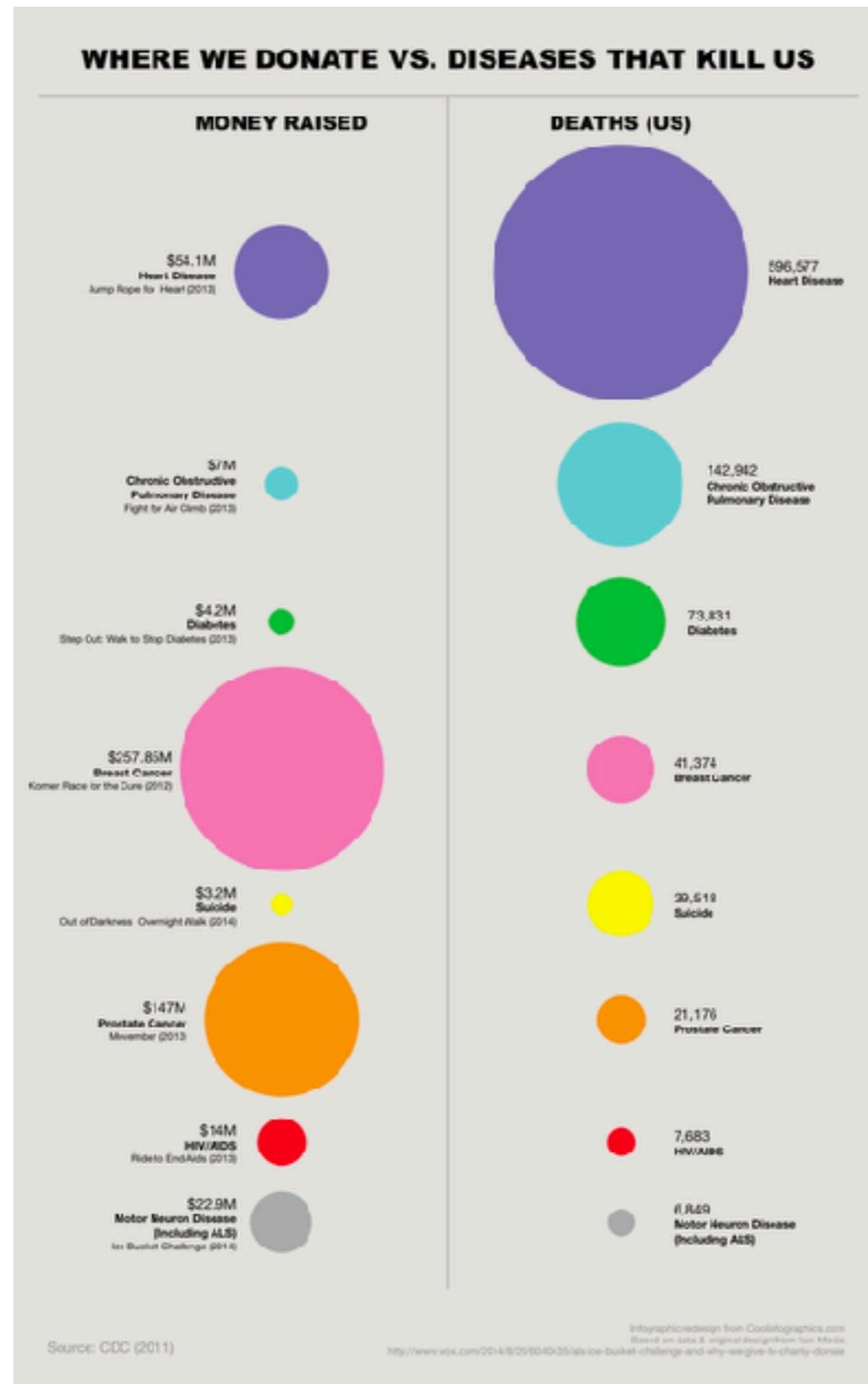
**Diabetes**  
Step Out: Walk to Stop Diabetes (2013)      **HIV / AIDS**  
Ride to End AIDS (2013)      **Breast Cancer**  
Komen Race for the Cure (2012)

**Motor Neuron Disease (including ALS)**  
ALS Ice Bucket Challenge (2014)      **Chronic Obstructive Pulmonary Disease**  
Fight for Air Climb (2013)      **Prostate Cancer**  
Movember (2013)

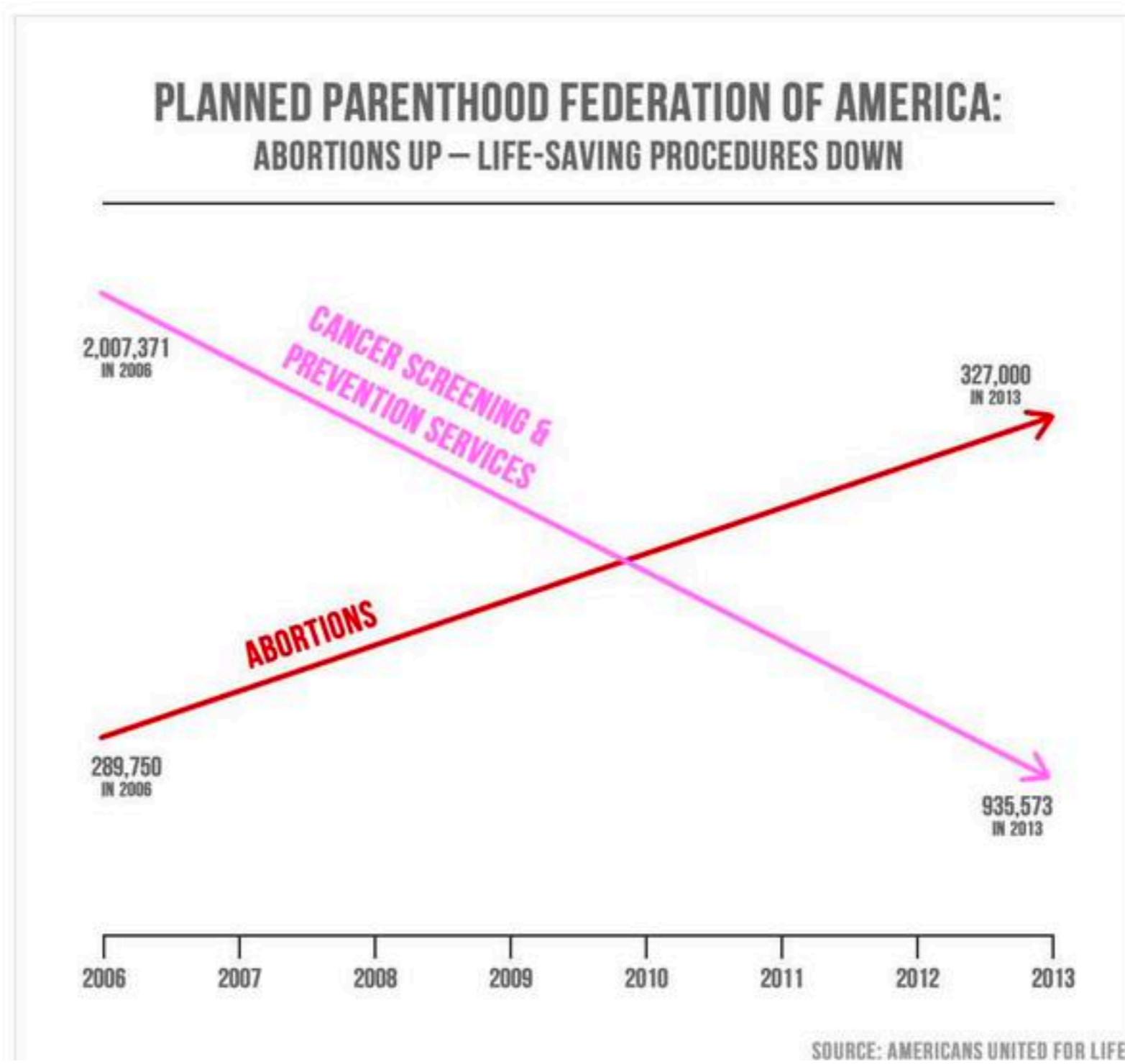


Source: CDC (2011)

# Much better but not perfect

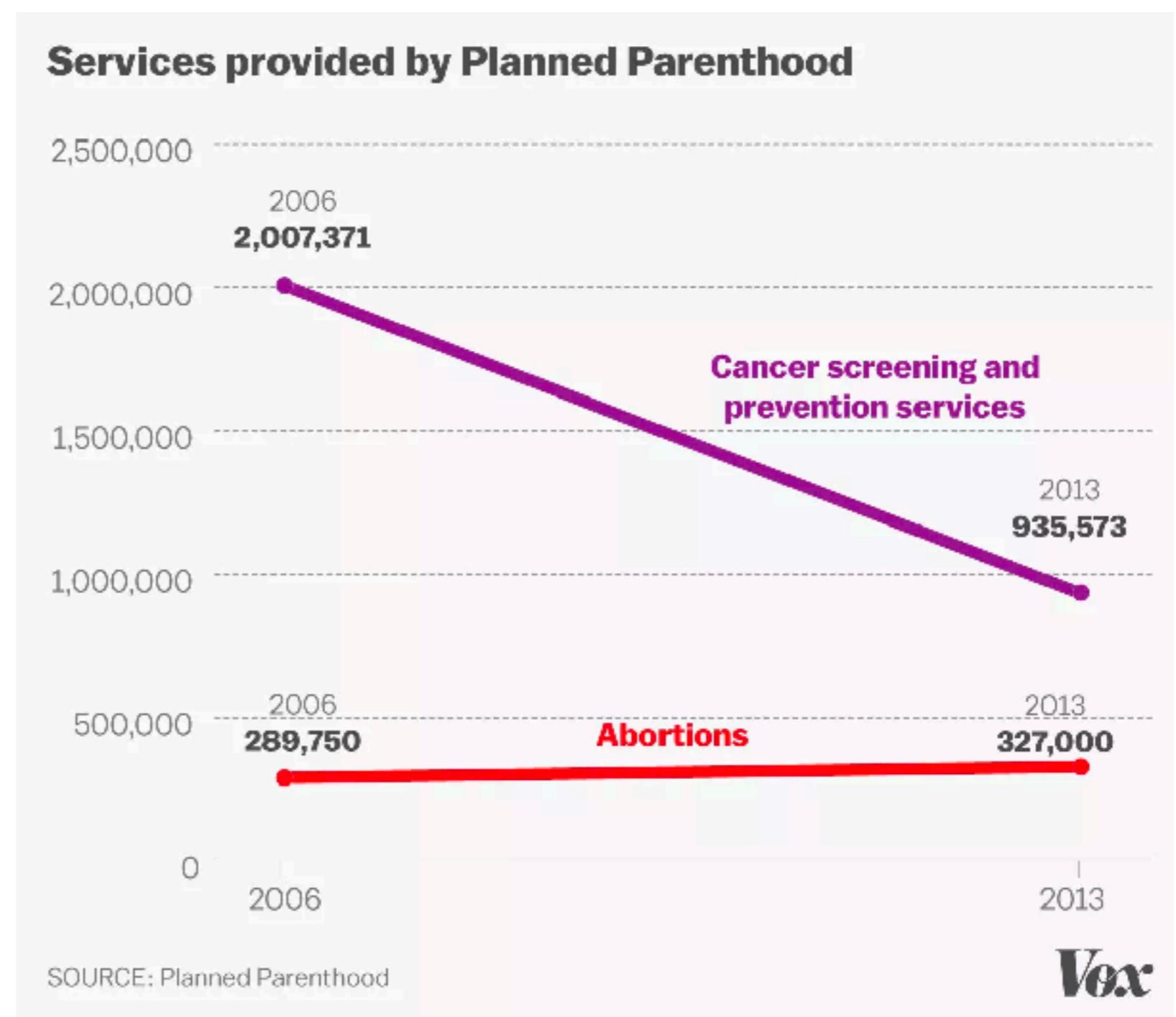
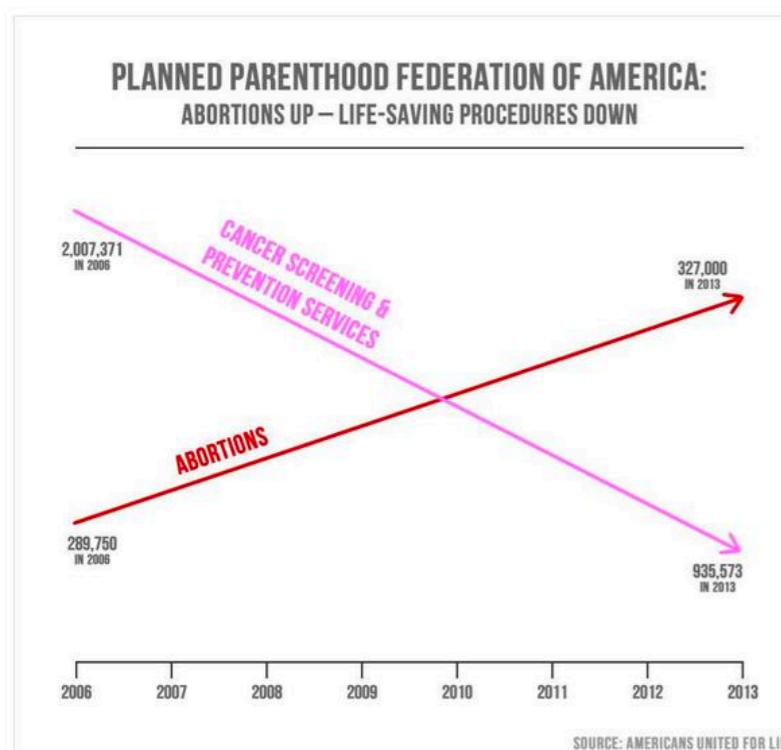


# Correlations

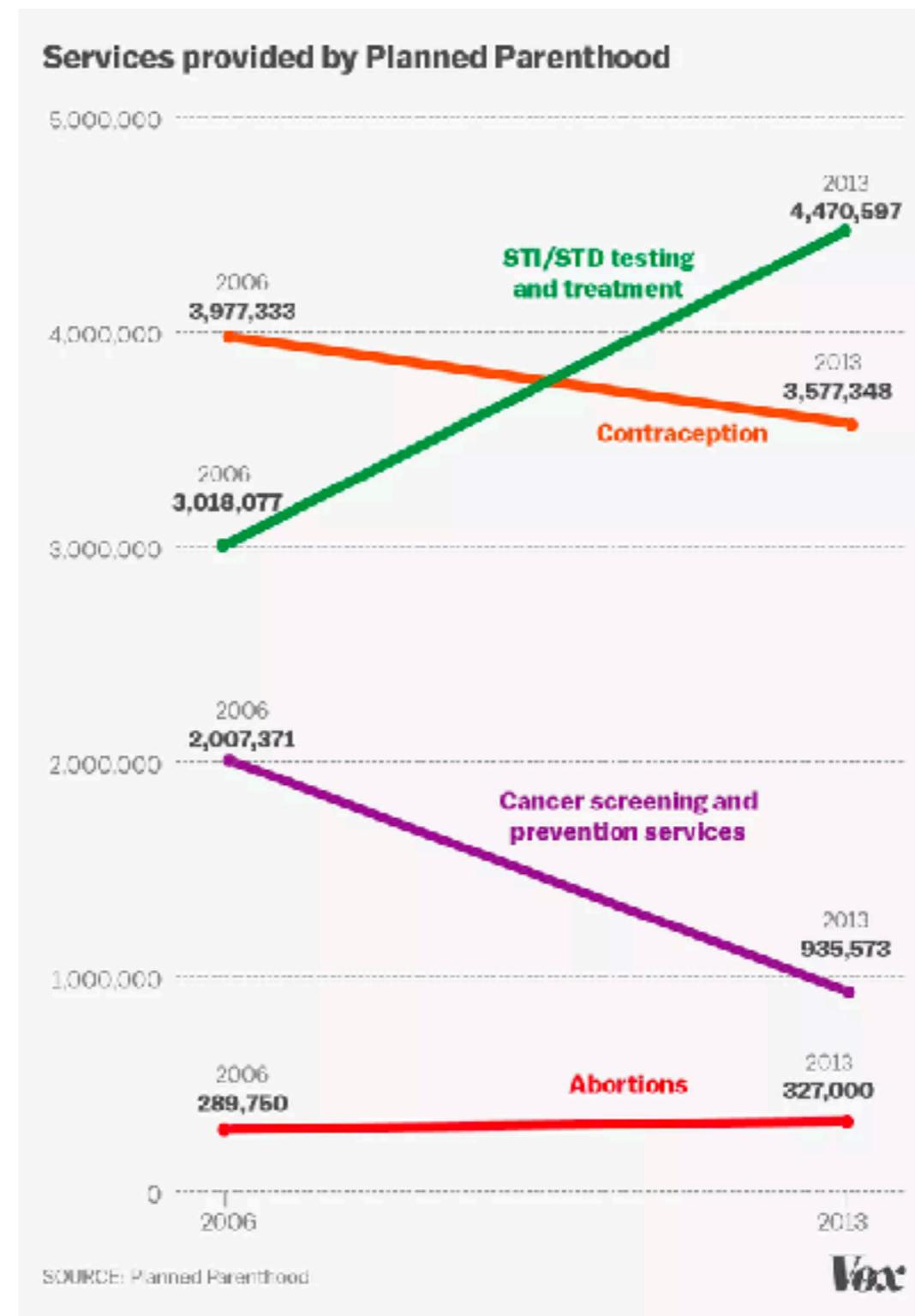
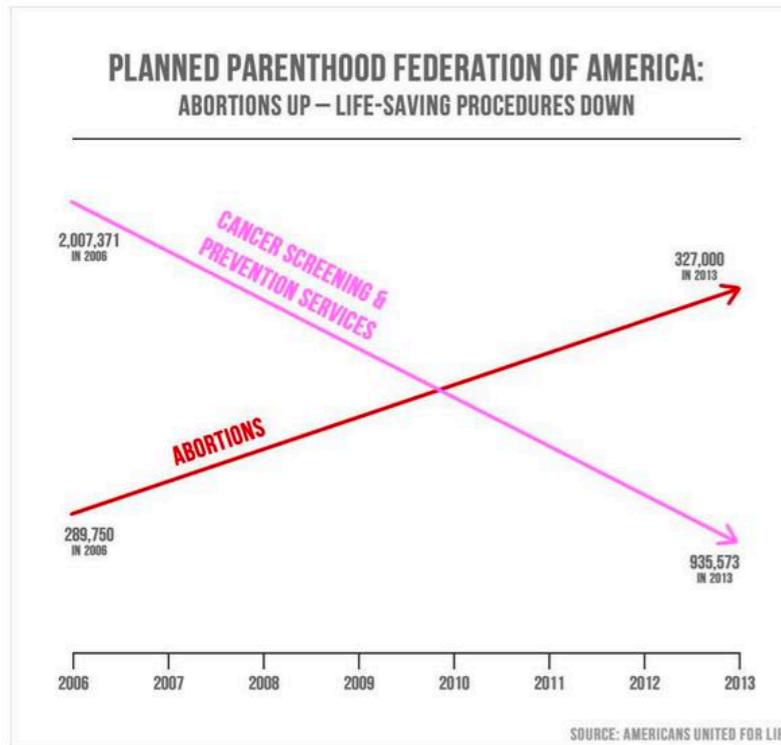


This chart was used by politicians willing to defund Planned Parenthood last September 2015  
(<http://www.vox.com/2015/9/29/9417845/planned-parenthood-terrible-chart>)

# The real graph

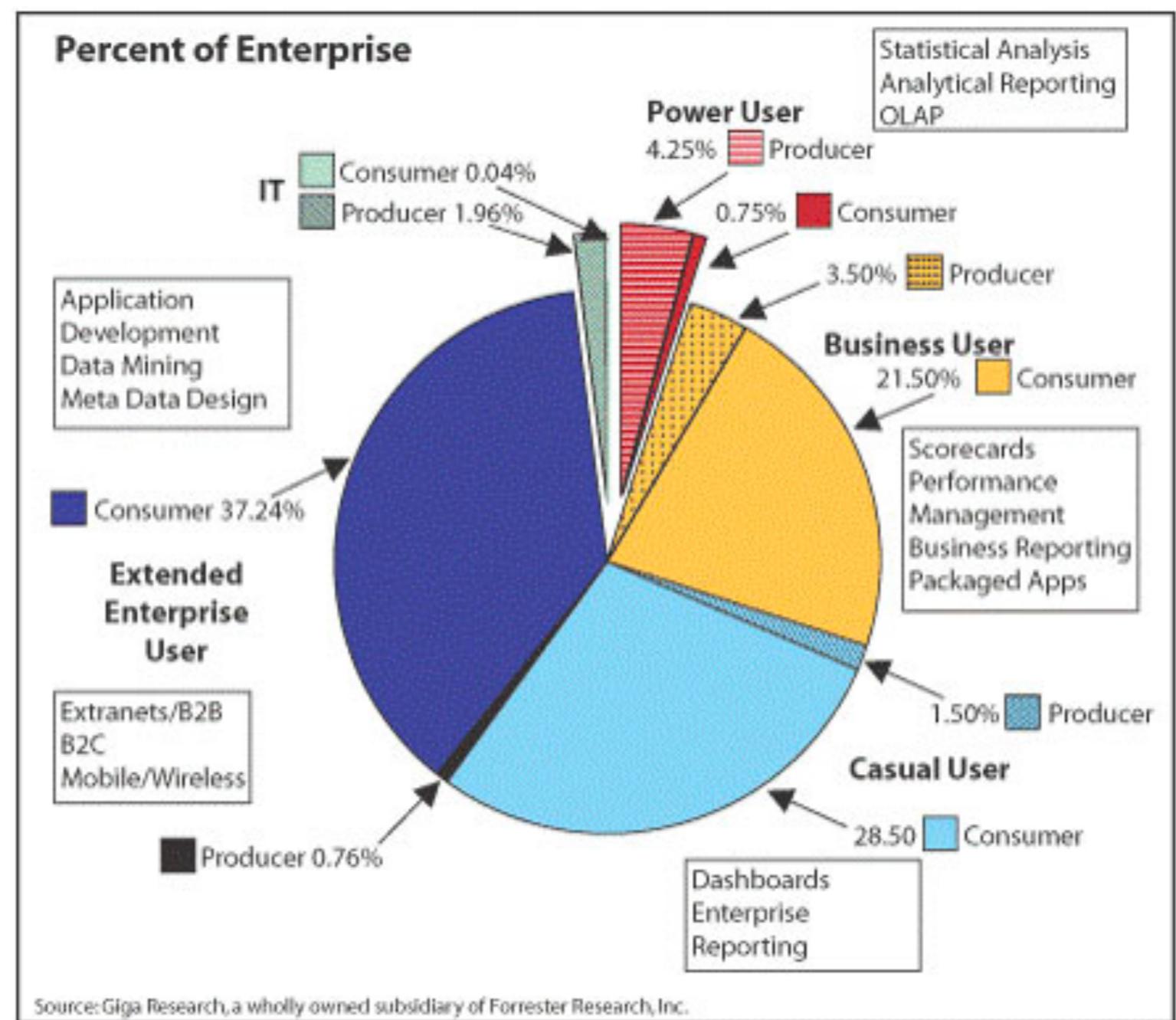


# The real graph



As a summary:  
Typical error on Data Visualization

**<http://tcagency.es/errores-de-visualizacion-de-datos/>**



**Too much information!**

# AVERAGE COST PER QUALIFIED LEAD

---



According to research done by Statistic Land and CMO

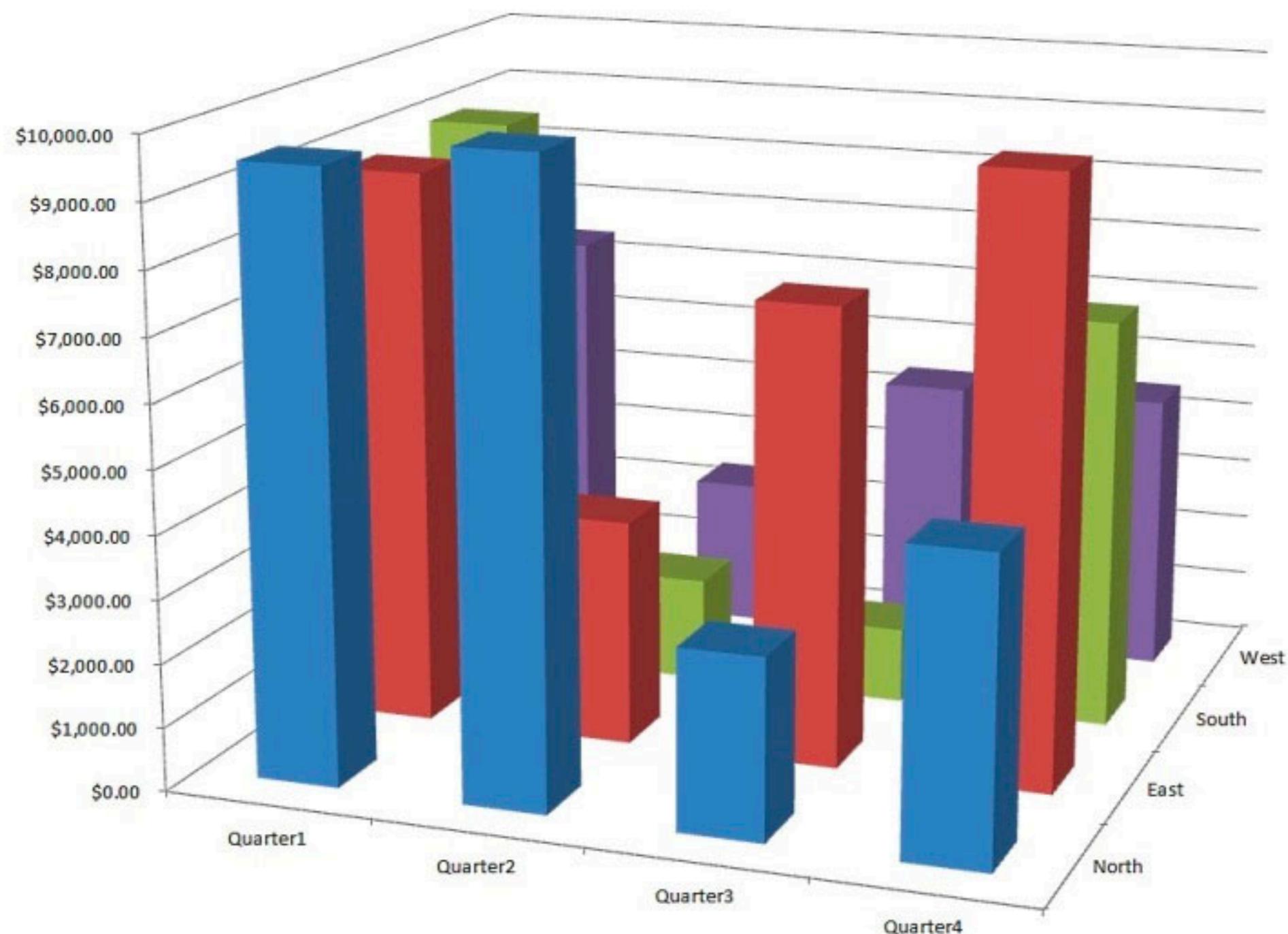
**The wrong Scale!**

## THE SMALL BUSINESS ADMINISTRATION REPORTS: NEW BUSINESS SURVIVAL RATES:



**Maths!**

## Quarterly Report by Region

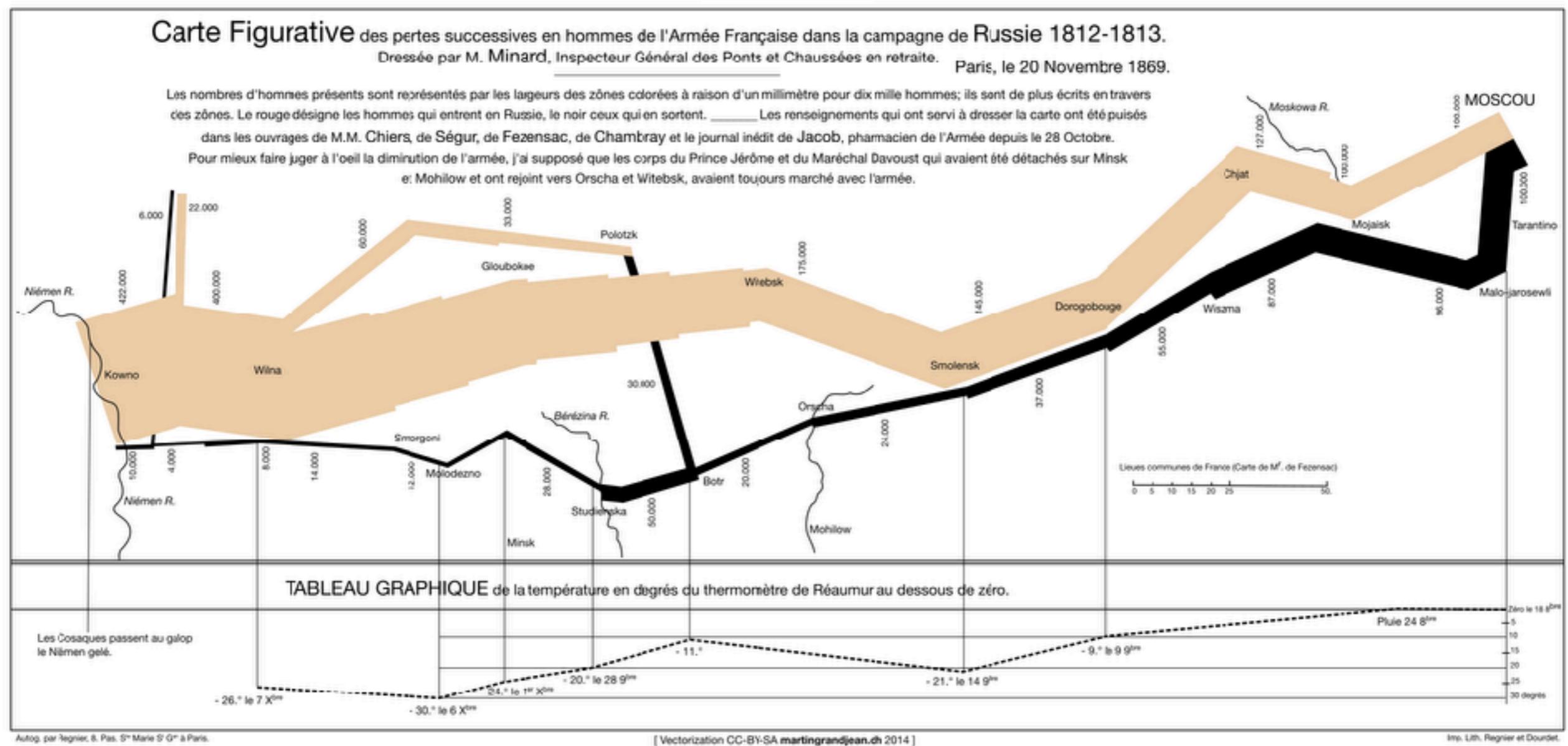


The choice of graph

# Story Telling

# Napoleon Invasion by Charles Joseph Minard

Edward Tufte said it "may well be the best statistical graphic ever drawn"



[http://en.wikipedia.org/wiki/Charles\\_Joseph\\_Minard#mediaviewer/File:Minard.png](http://en.wikipedia.org/wiki/Charles_Joseph_Minard#mediaviewer/File:Minard.png)

Variables displayed in the graph: 1) the number of Napoleon's troops; 2) the distance traveled; 3) temperature; 4) latitude and longitude; 5) direction of travel; and 6) location relative to specific dates

The story of the world in 200 countries over 200 years  
using 120,000 numbers by Hans Rosling



[https://www.youtube.com/watch?time\\_continue=203&v=jbkSRLYSoj](https://www.youtube.com/watch?time_continue=203&v=jbkSRLYSoj)

Is the world overpopulated? Will the population increase?  
by Hans Rosling



[\*\*https://www.youtube.com/watch?v=-UbmG8gtBPM\*\*](https://www.youtube.com/watch?v=-UbmG8gtBPM)

# Guns in America?

**FiveThirtyEight**

Share

Another third of all gun deaths — about 12,000 in total each year — are [homicides](#).

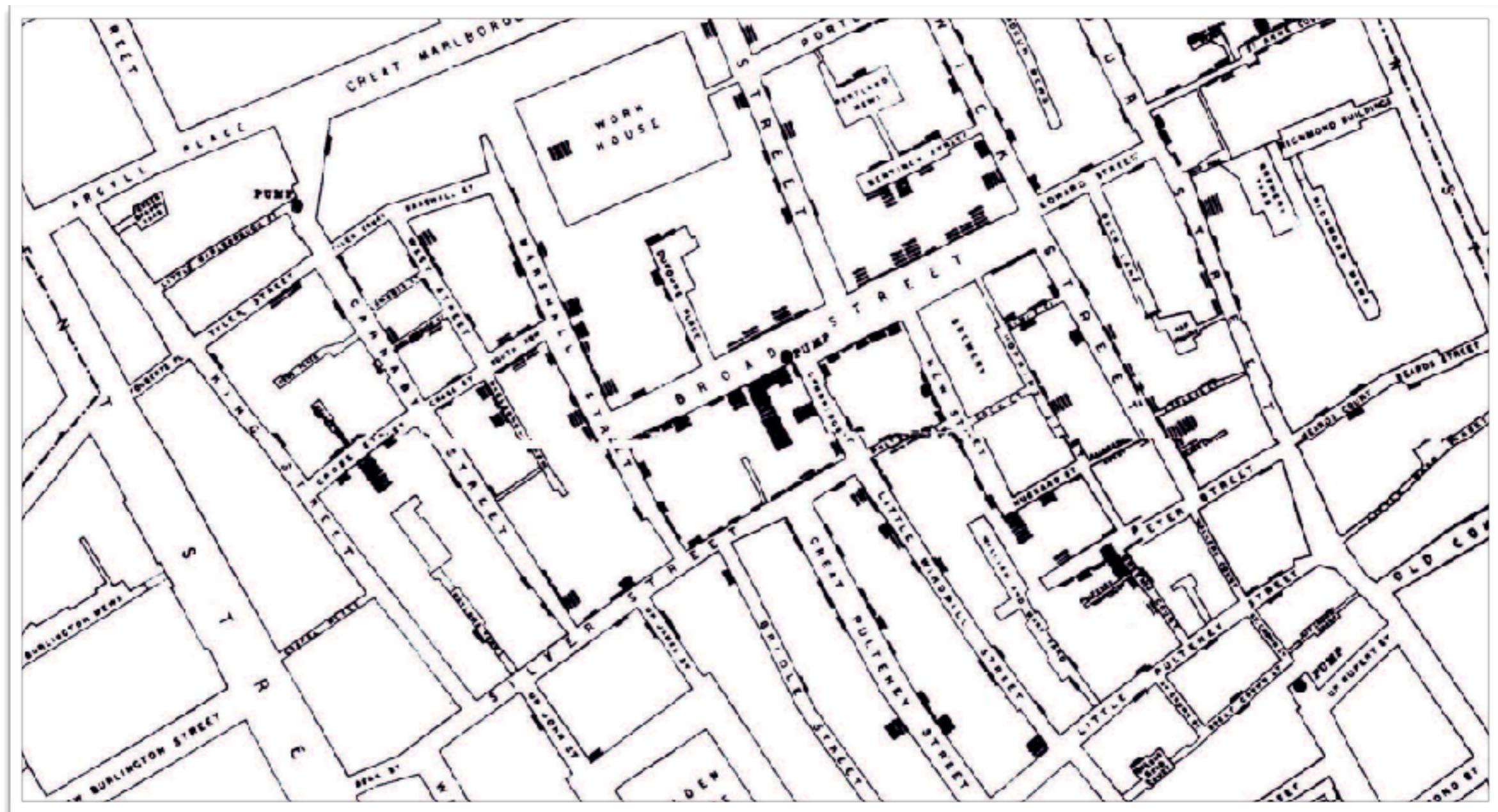


1 2 3 4 5 6 7 8 9 10 11 12 | Explore the data for yourself »

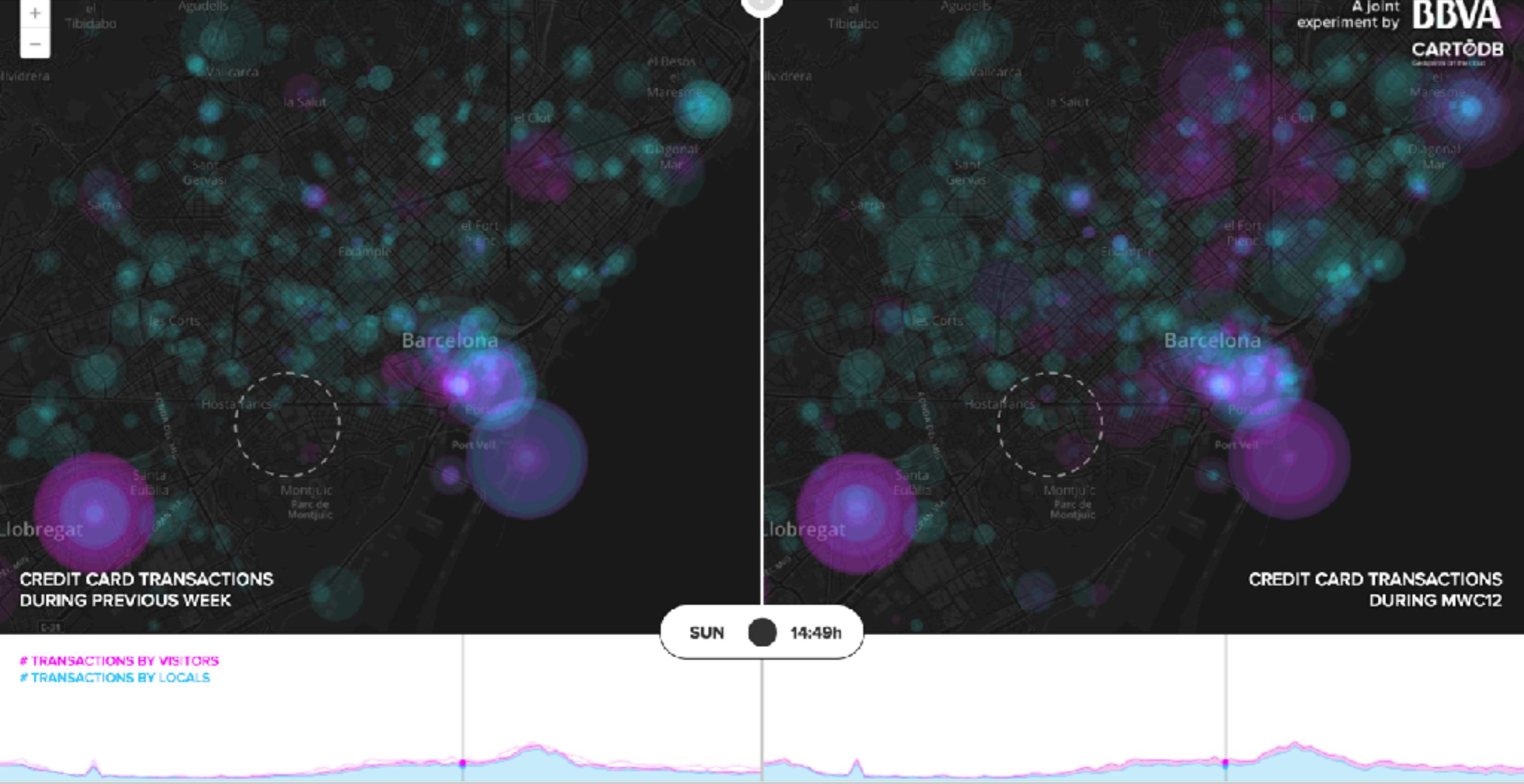
<https://fivethirtyeight.com/features/gun-deaths/>

# Knowledge Discovery

London 1854. Difficult times for London population, specially in Soho, when a cholera outbreak killed thousands. A map saved the city. How?



Map by John Snow. [https://en.wikipedia.org/wiki/John\\_Snow](https://en.wikipedia.org/wiki/John_Snow)



**Economic Impact** The economic impact of the MWC on Barcelona created with CartoDB and BBVA. The week before vs. the 2012 MWC week <http://mwcimpact.com/>

Wednesday, Nov 15

00:00 am

Slower 30 mins Faster

### Terminals

- JFK T1
- JFK T23 Park
- JFK T4
- JFK T5n
- JFK T7
- JFK T8
- LGA A
- LGA B
- LGA C
- LGA D

JFK LGA  
Scotch Plains Westfield Cranford

Plainfield

Taxi trips

Livingston	0
Caldwell	0
West Orange	0
Irvine	0
Millburn	0
Summit	0
Union	0

Elizabeth Rahway Carteret Woodbridge Perth Amboy

Clark Linden

Bayonne

NY 27

# One Dataset - Lots of interesting stuff to learn

**Taxi GPS data helps researchers study Hurricane Sandy's effect on NYC traffic**

<https://engineering.illinois.edu/news/article/9717>

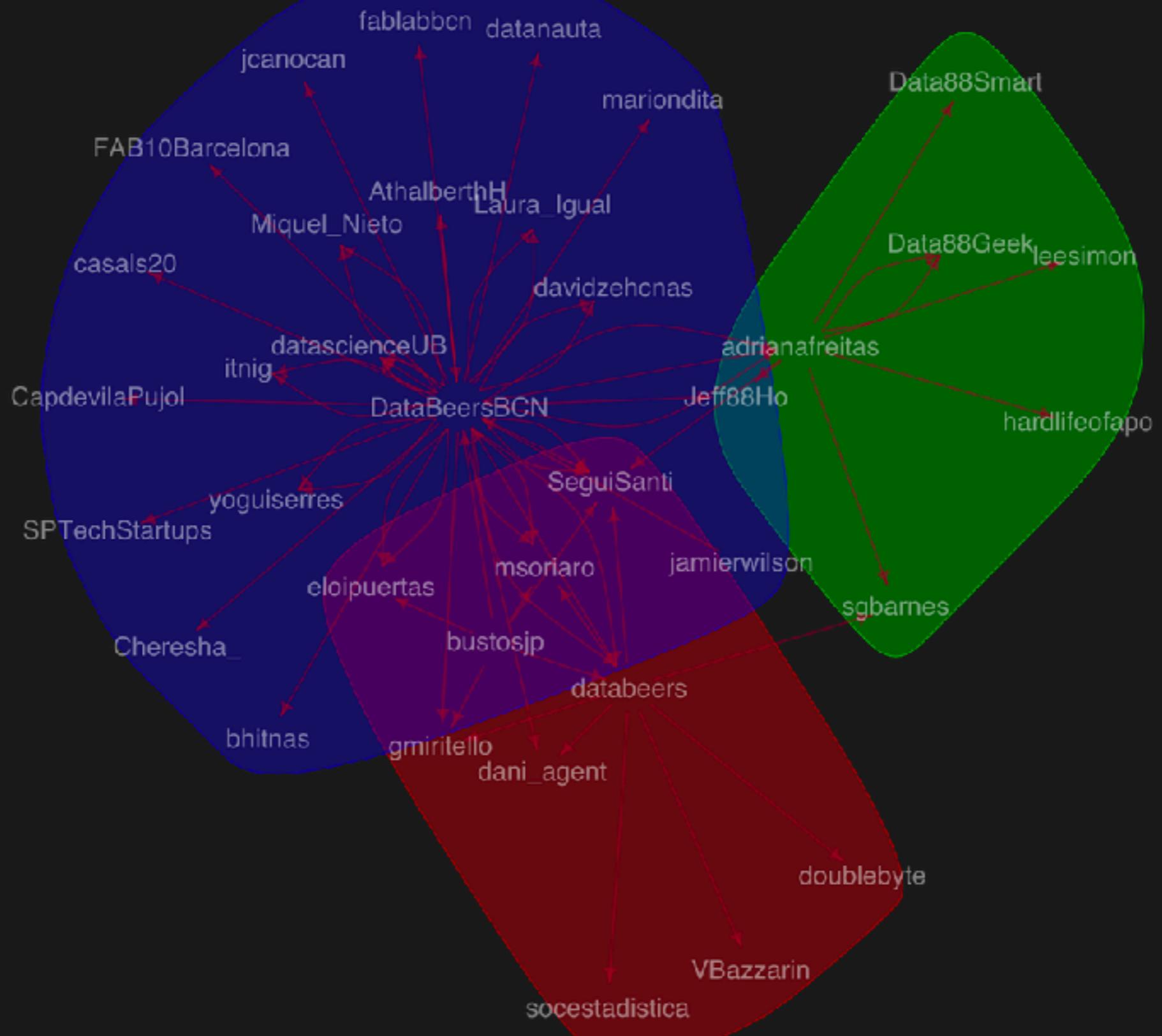
**Exploring New York City taxi trails and sharing our way to a more sustainable urban future**

<http://hubcab.org/#13.00/40.7219/-73.9484>

**Public NYC Taxicab Database Lets You See How Celebrities Tip**

<http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546>

## Tweets with 'DataBeersBCN': Who retweets whom



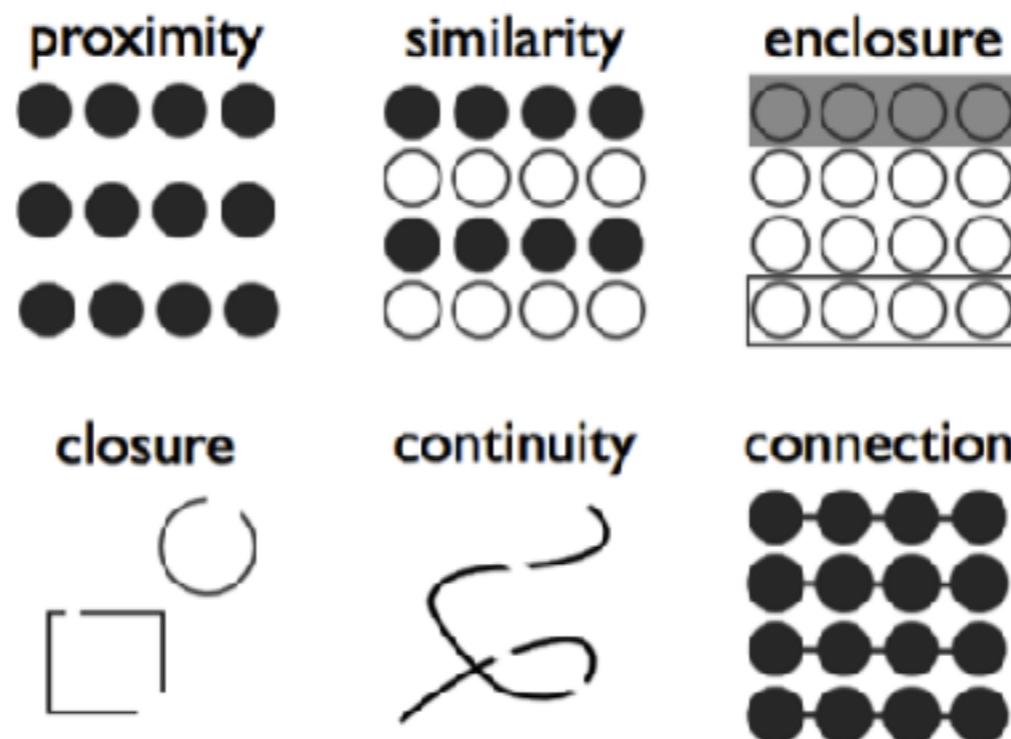
# The Gestalt Principles

**Similarity:** When items look the same, people perceive them to be of the same type. We naturally assume that shapes that look the same are related. When you create a data viz and you keep items together that look the same, you make it easy for someone to understand that those items represent a group.

**Closure:** Our eyes tend to add any missing pieces of a familiar shape. If two sections are taken out of a circle, as shown in the following figure, people still perceive the whole circle.

**Continuation:** If people perceive objects as moving in a certain direction, they see them as continuing to move that way. The figure below shows an example.

**Figure/ground:** Depending on how people look at a picture, they see either the figure (foreground) or the ground (background) as standing out, as shown in the following figure.

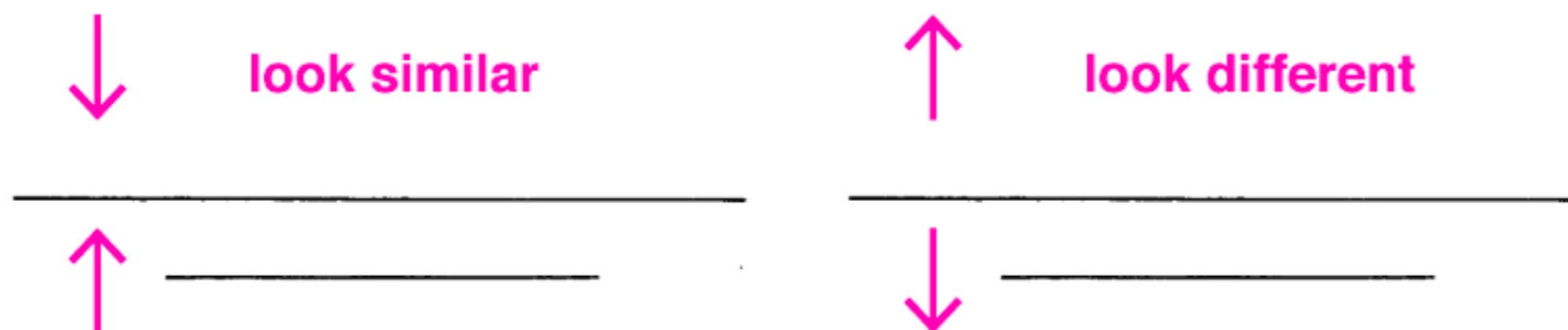


# More about perception

Great post about Perception: <https://medium.com/@kennelliott/39-studies-about-human-perception-in-30-minutes-4728f9e31a73#.rwk4sut5h>

One Example: **Reference point:**

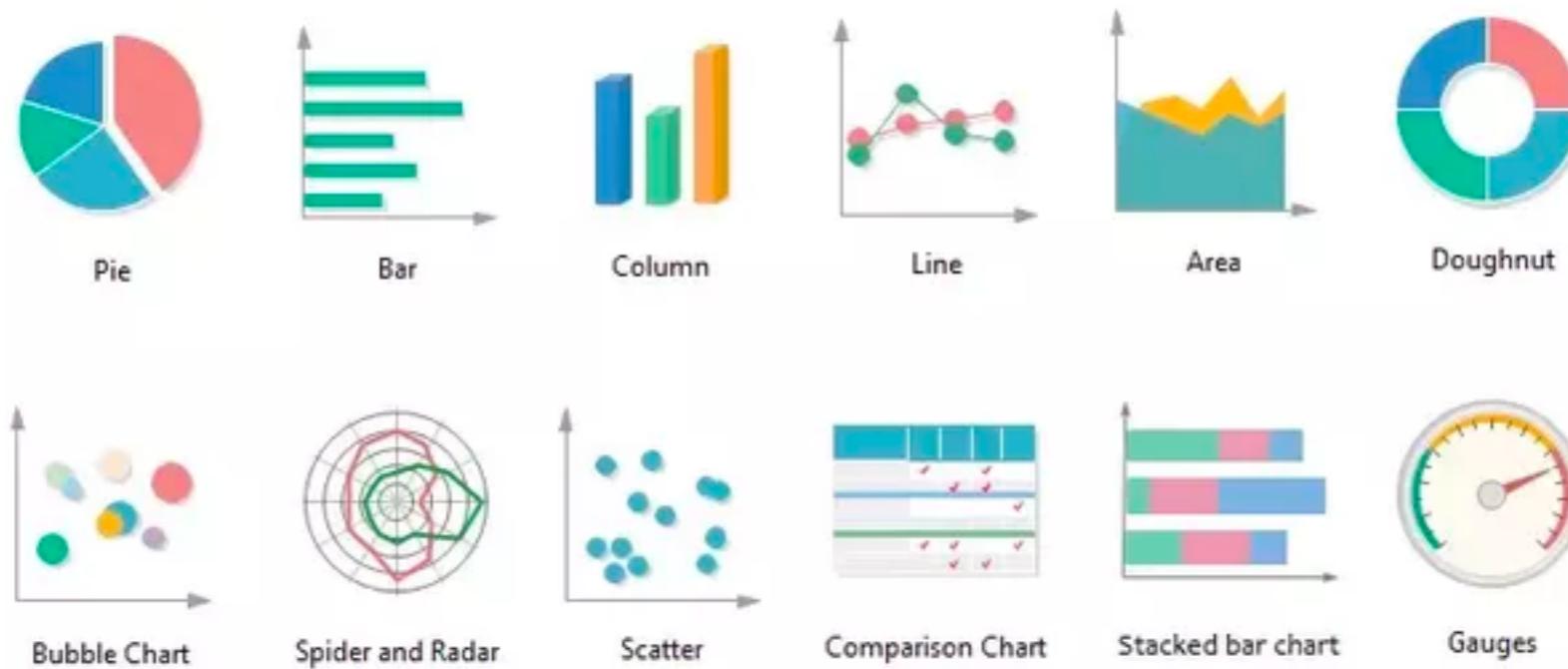
In the previous link, they mentioned: "Three studies show that we have inherent biases related to the types of graphs we see and the objects we see in them. These biases may distort the information we retrieve from a graph. We know that from Steven's law, when an object is seen in context of other larger objects, it appears larger itself. When it is seen in context with smaller objects, it appears smaller. Jordan and Schiano (3) found that increasing spatial separation between lines produced the opposite effect. If lines were close enough, a line's length was more similar to the length of the line around it (this is also known as assimilation). If the lines were far apart, long lines appeared longer, and short lines appeared shorter (also known as contrast)."



**Figure 1. Example of the parallel-lines illusion configuration. Relative to a no-context control line, the test line is overestimated when presented with the longer contextual line.**

# Things to take into account when designing a visual data representation:

**The Minimum Ink Principle**  
**Colormap**  
**Type of chart or graph**



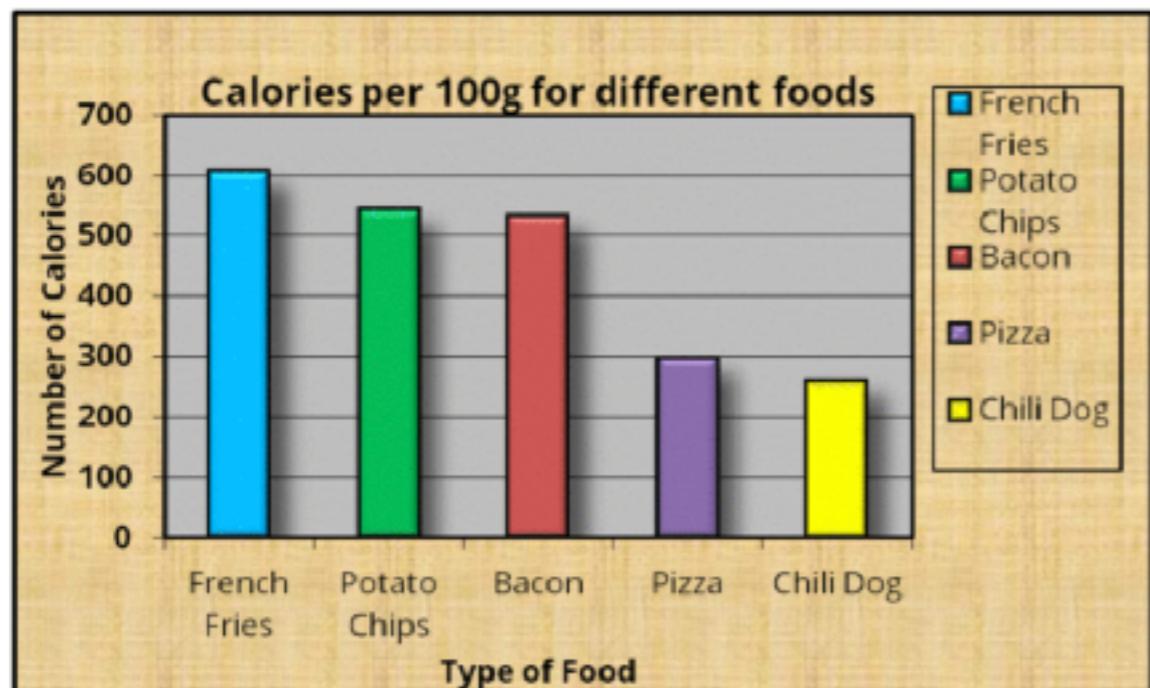
# The Minimum Ink Principle

**Remove**  
to improve  
(the **data-ink** ratio)

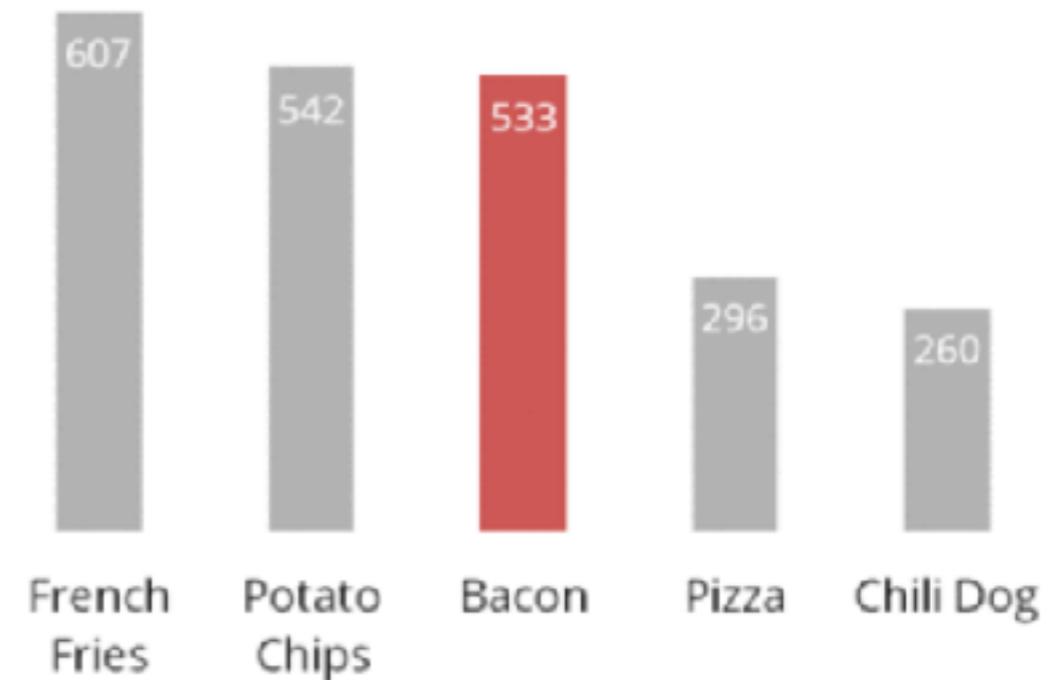
Created by Darkhorse Analytics

[www.darkhorseanalytics.com](http://www.darkhorseanalytics.com)

# The Minimum Ink Principle



Calories per 100g



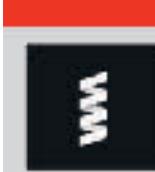


Colors:  
How do I choose the proper colormap?

**First Advise:** Picking colors is really important task, you don't have to reinvent the wheel. A number of great websites out there will help you choose the right palette for your infographic

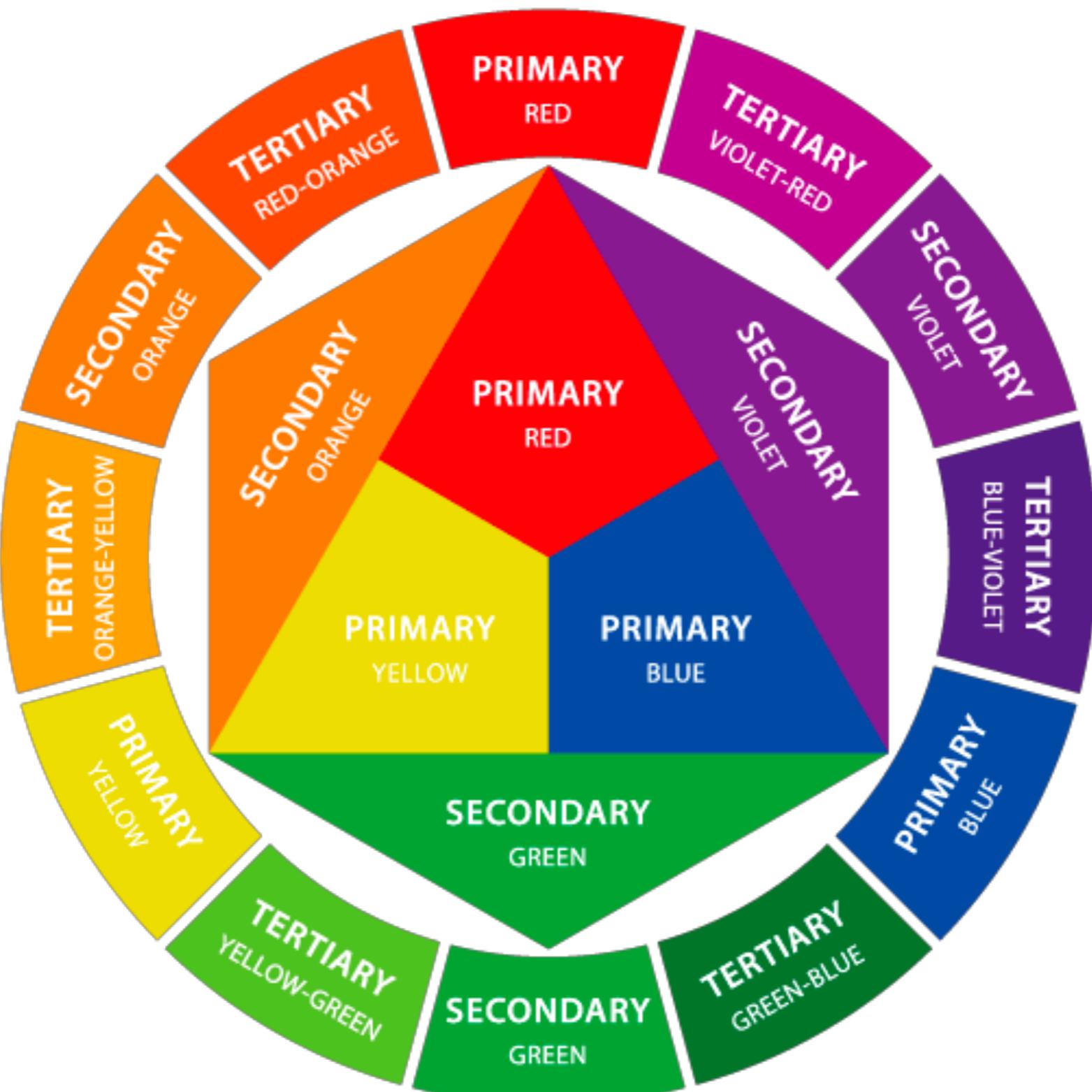
**<https://color.adobe.com/create/color-wheel/>**

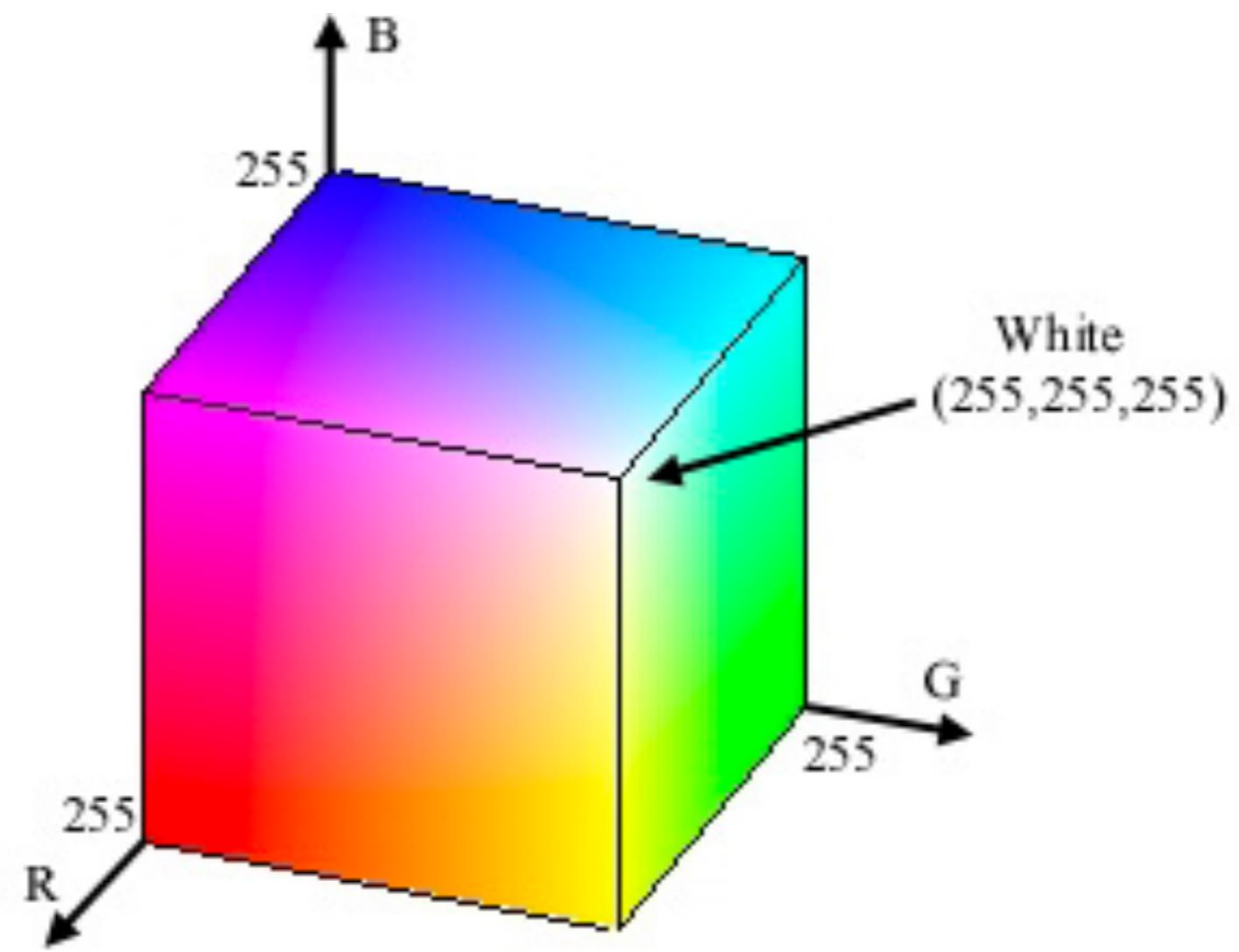
**<http://www.colourlovers.com/>**



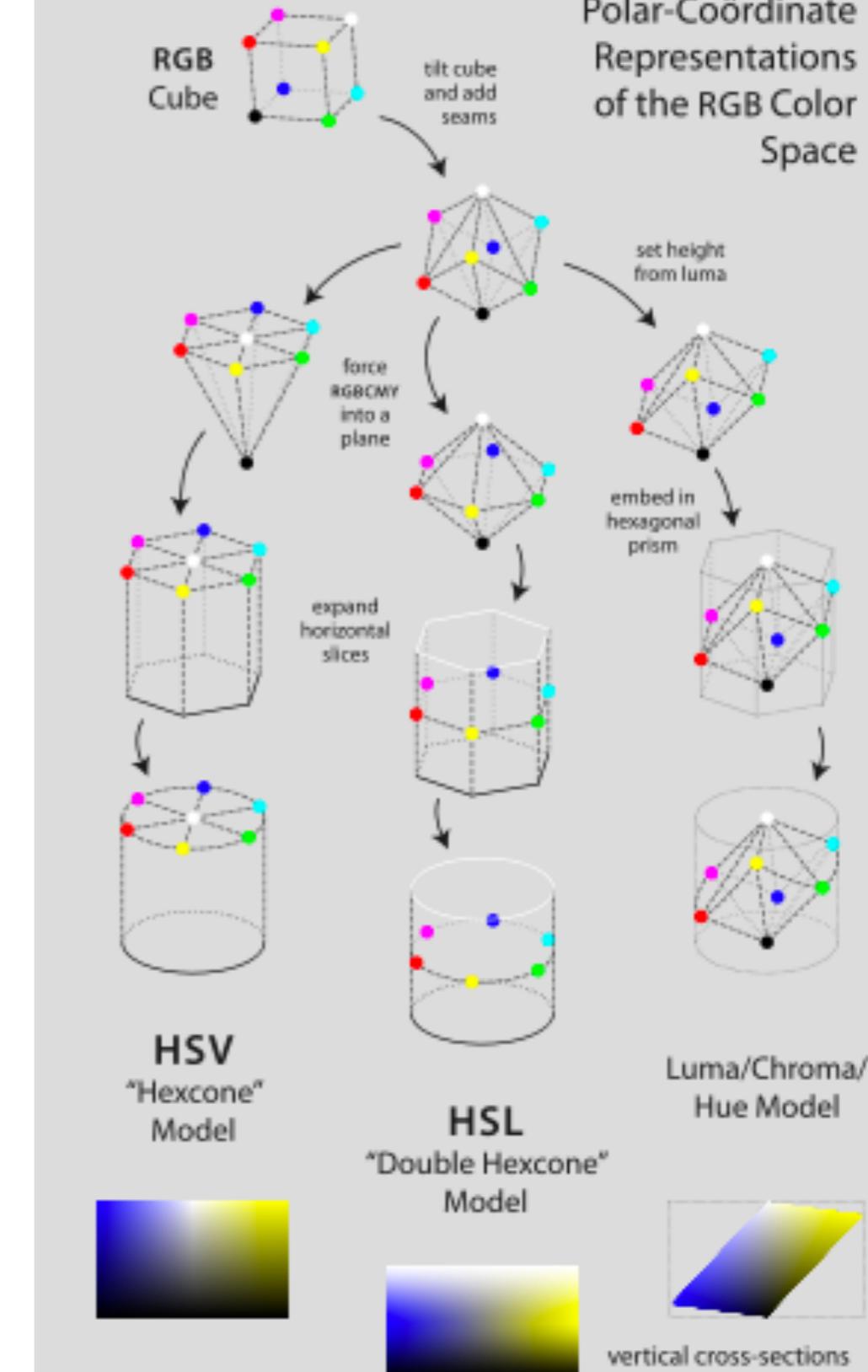
# More tips

- For **Categorical** variables:
  - People have trouble differentiating between more than 5-7 hues (colors)
- For **numerical** variables:
  - People have trouble differentiating between more than 5-7 shades
  - Rainbow colors gradients are very problematic
- For **highest** contrast, only us color to highlight

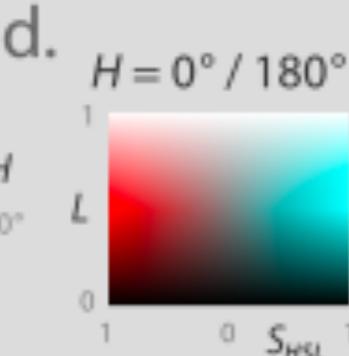
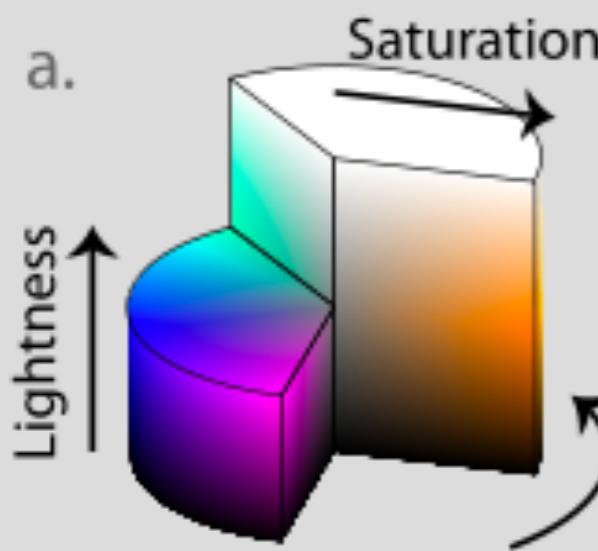




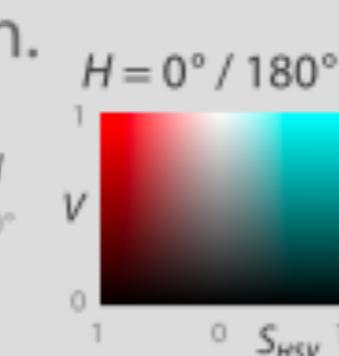
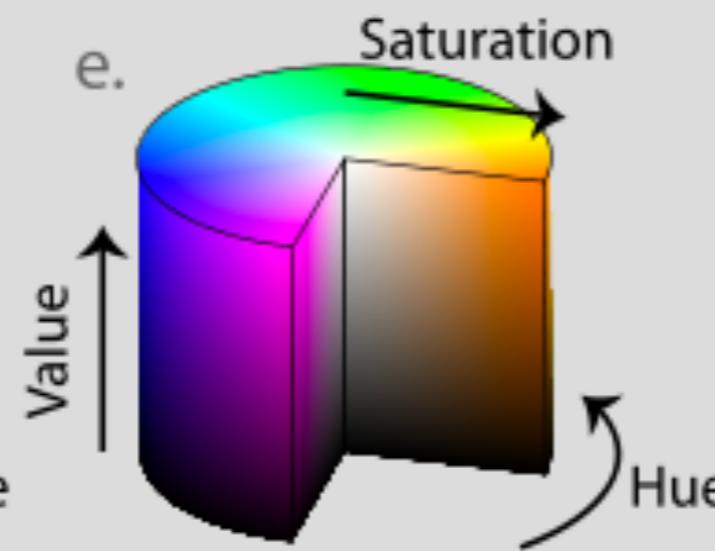
## Polar-Coördinate Representations of the RGB Color Space



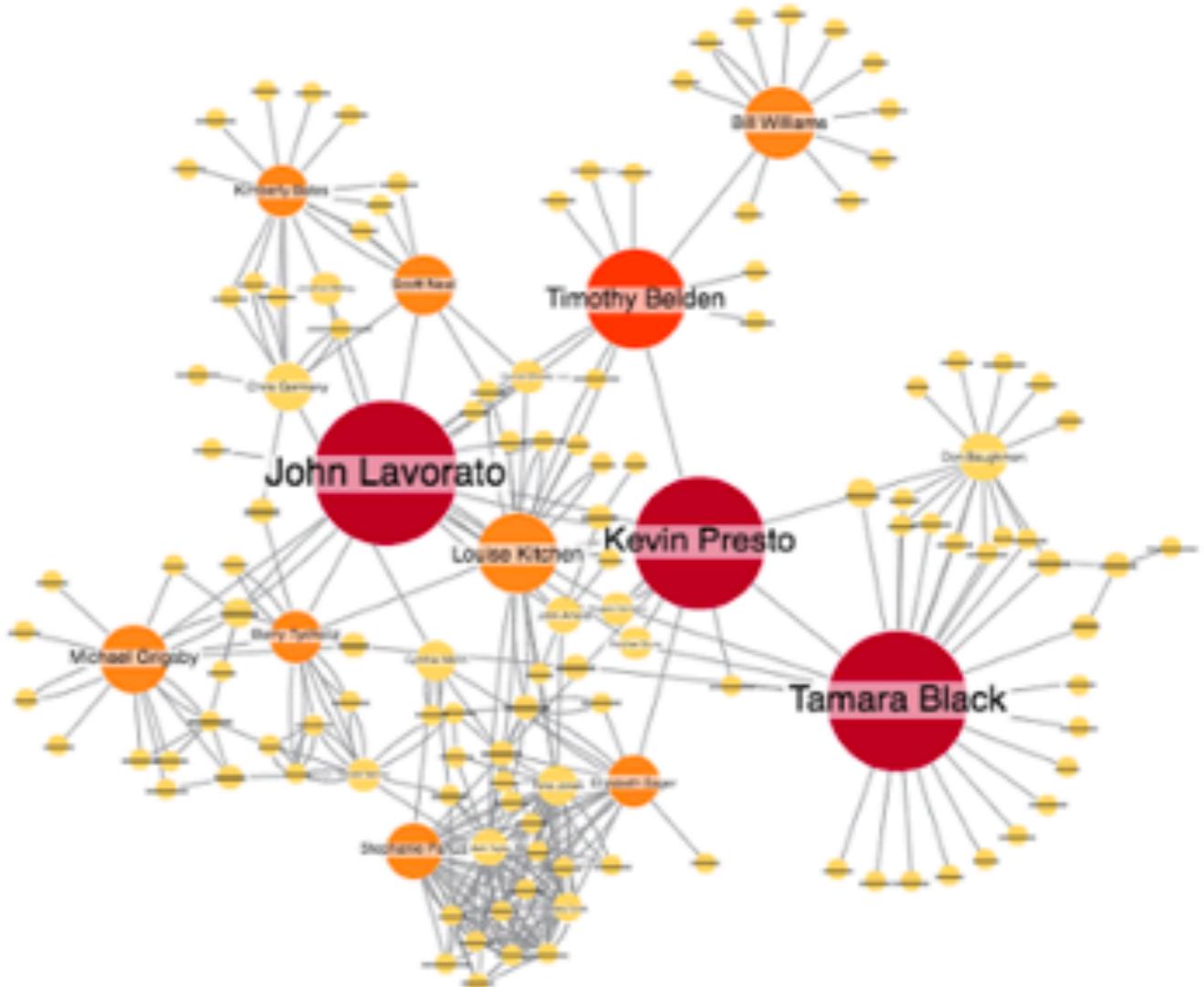
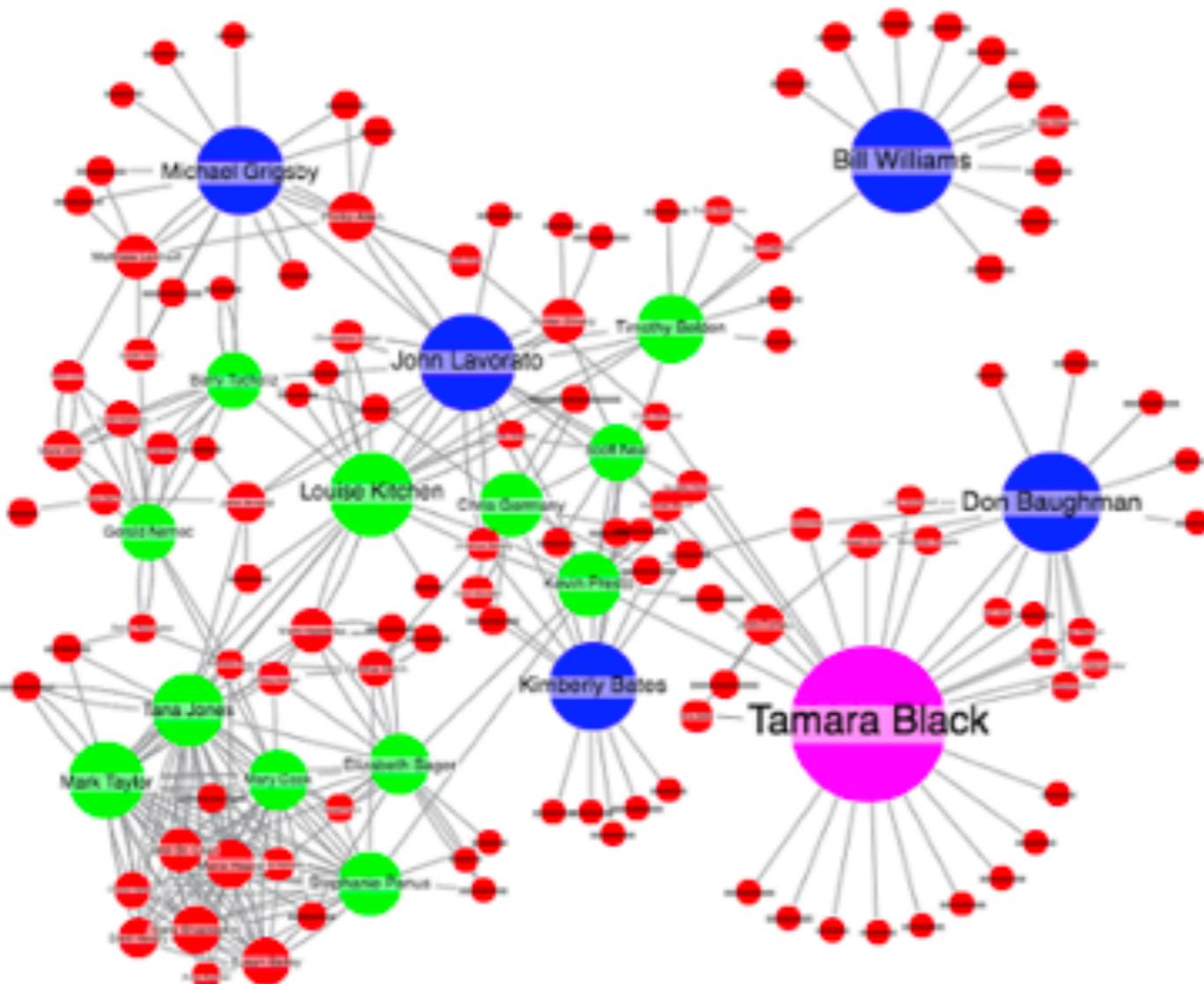
## HSL



## HSV



# Which one do you prefer?



# How to choose colors for your graphs?

- Step 1: Decide what the colors will represent
- Step 2: Understand your data scale
- Step 3: Decide how many hues you need
- Step 4: Look for obvious options
- Step 5: Create your palette

# Colormap?

Colormaps are often split into several categories based on their function:

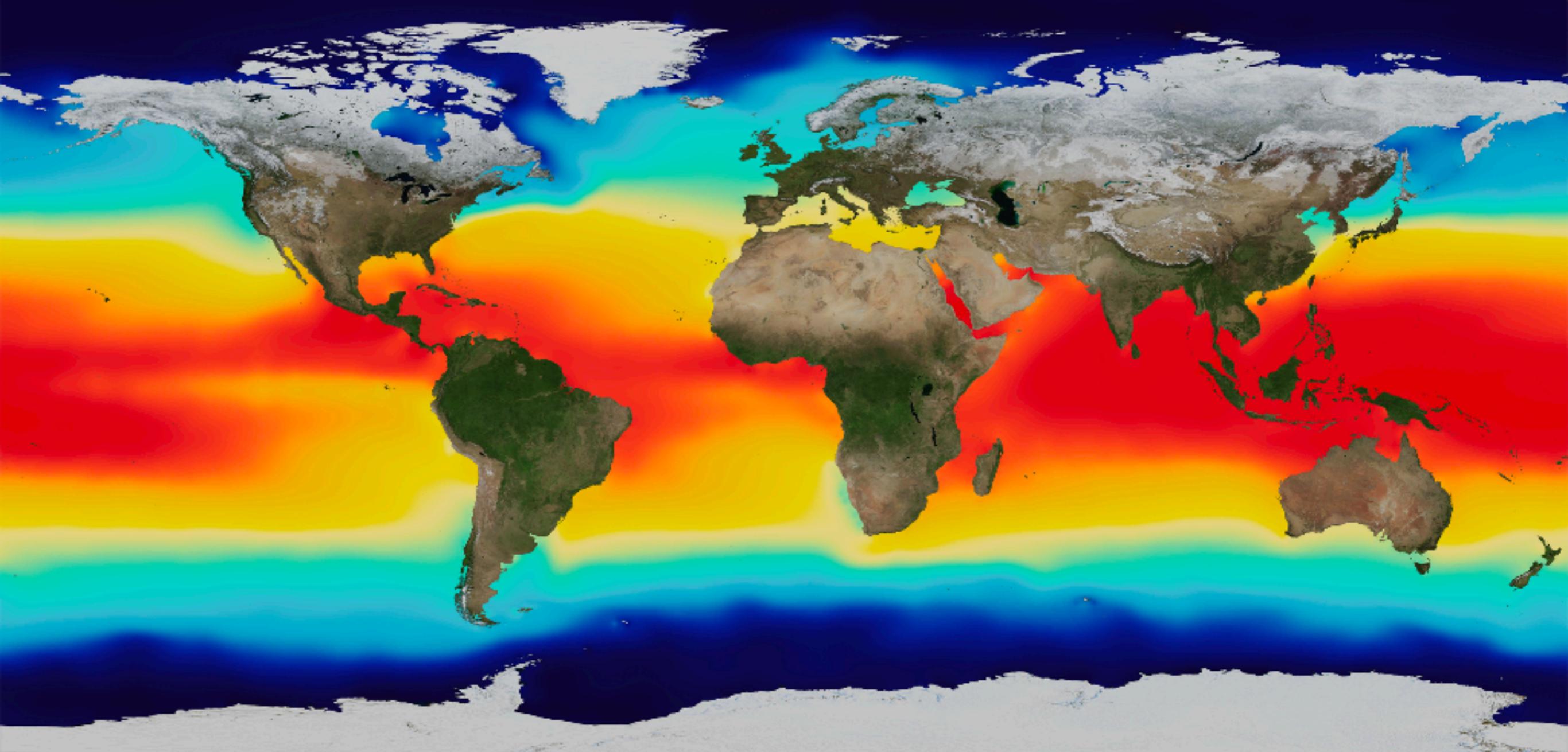
- Sequential;
- Diverging;
- Categorical



Usually only one color is used, changing light intensity or saturation. Should be used for representing information that has ordering.

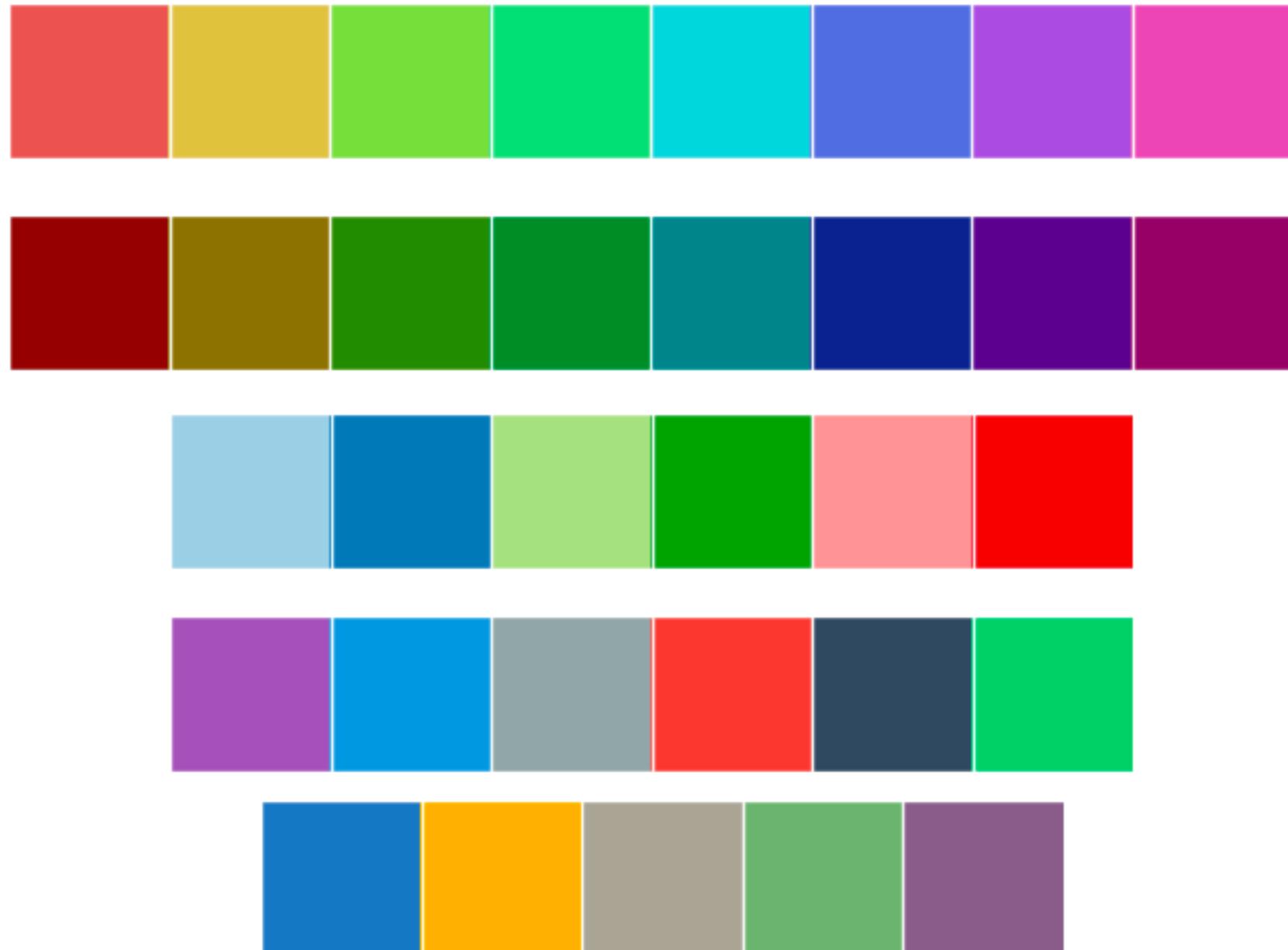
# Colormap?

**Diverging:** Usually two colors and the transition from one to the other are used. Should be used when the information being plotted has a critical middle value. For instance, negative temperatures (blue) and positive temperatures (red).



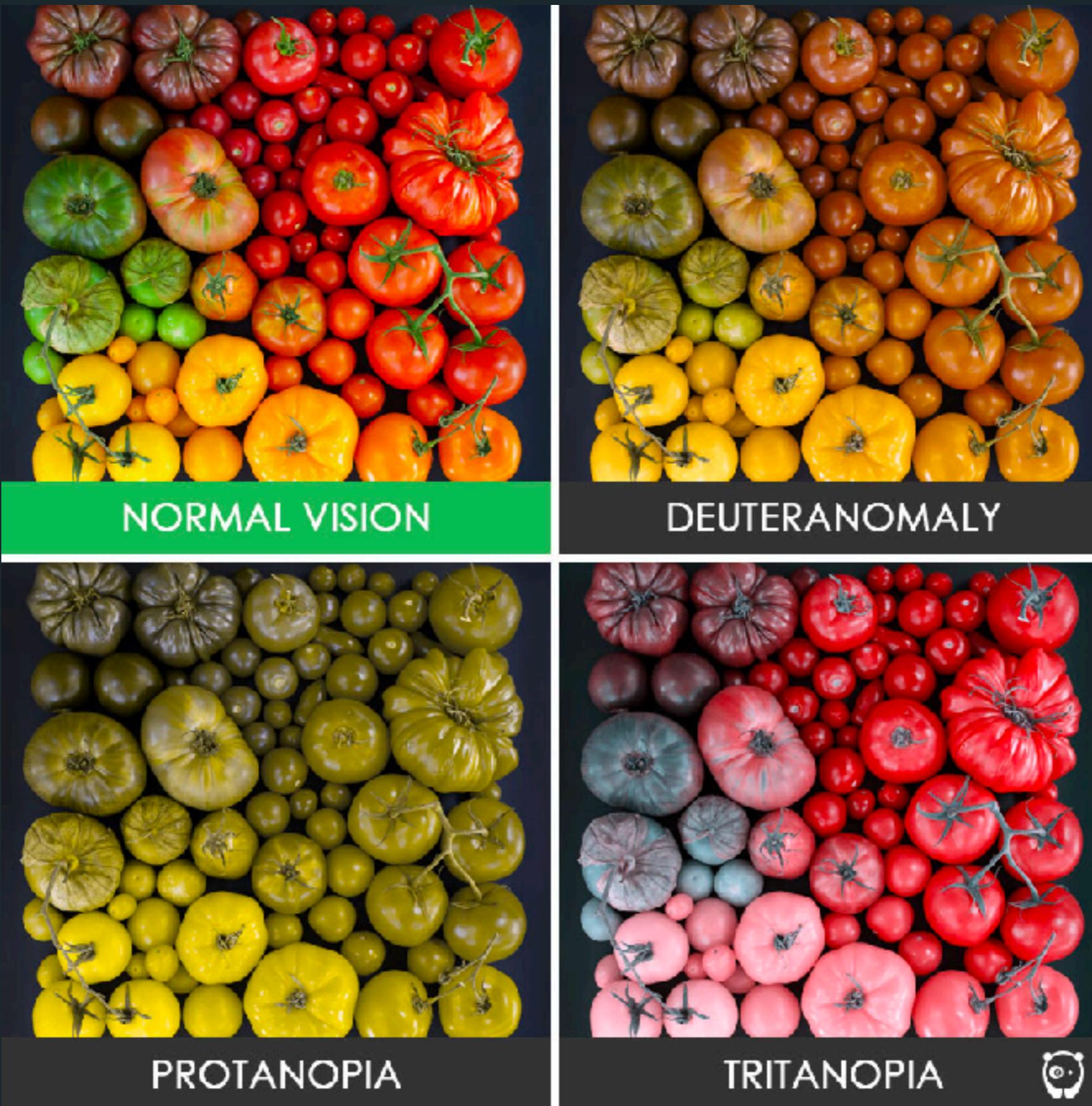
# Colormap?

**Categorical:** often are miscellaneous colors; should be used to represent information which does not have ordering or relationships.



[http://seaborn.pydata.org/tutorial/color\\_palettes.html](http://seaborn.pydata.org/tutorial/color_palettes.html)

# Color blindness



# Color blind



## indistinguishable colors in color blindness



<http://mkweb.bcgsc.ca/colorblind>

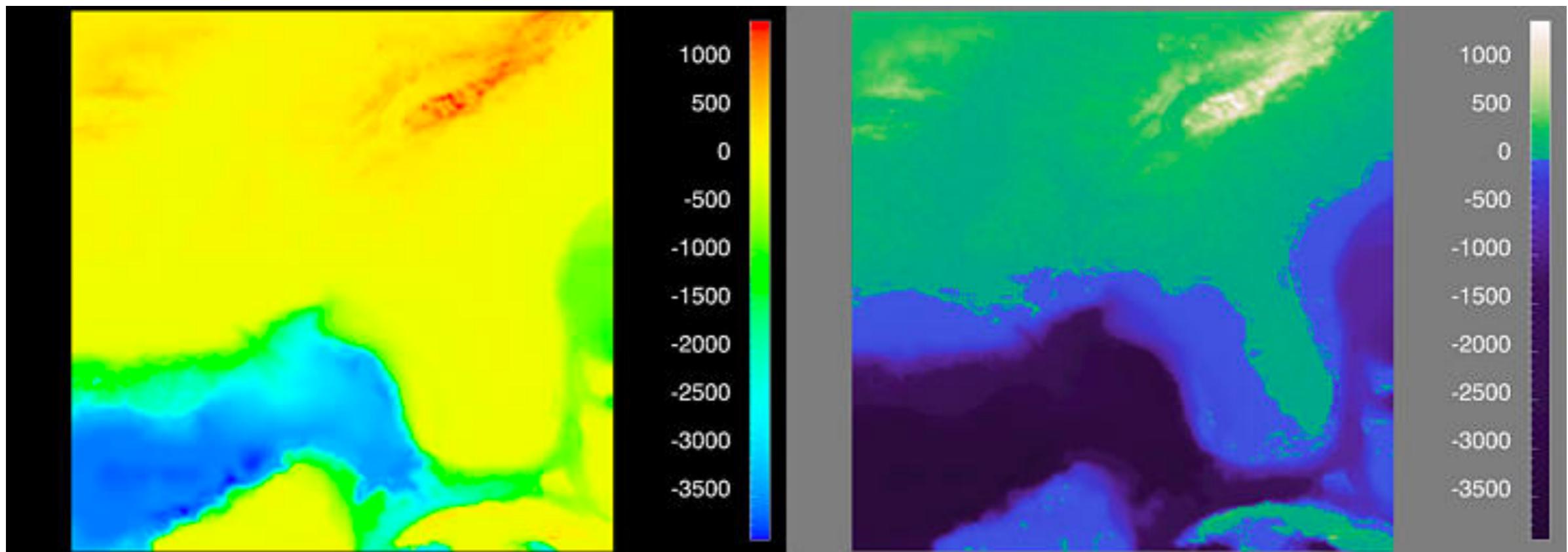
## The issue:

"Ten percent of men are colorblind and mostly red/green issues."

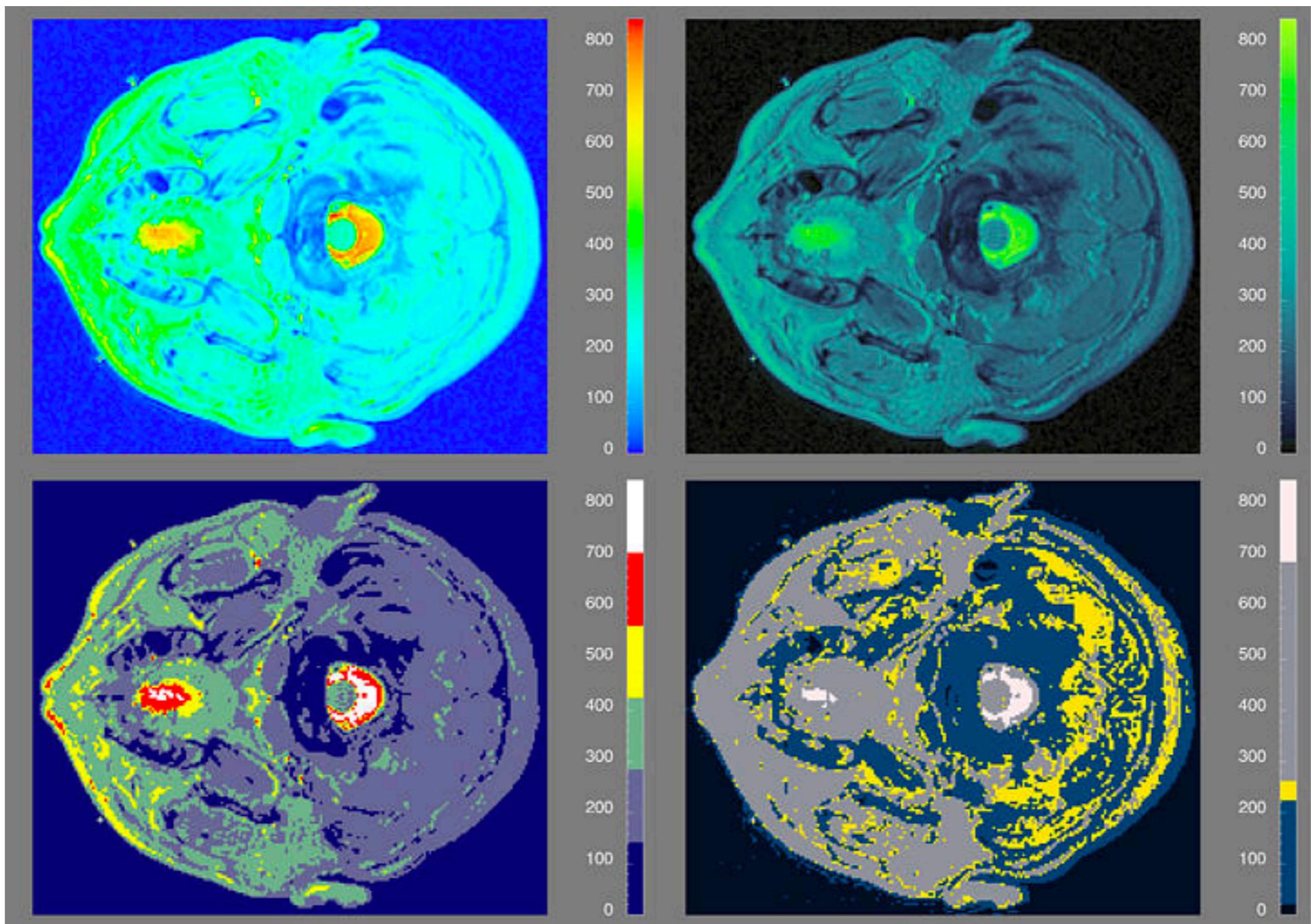
## The reaction?

"Don't use red and green together."

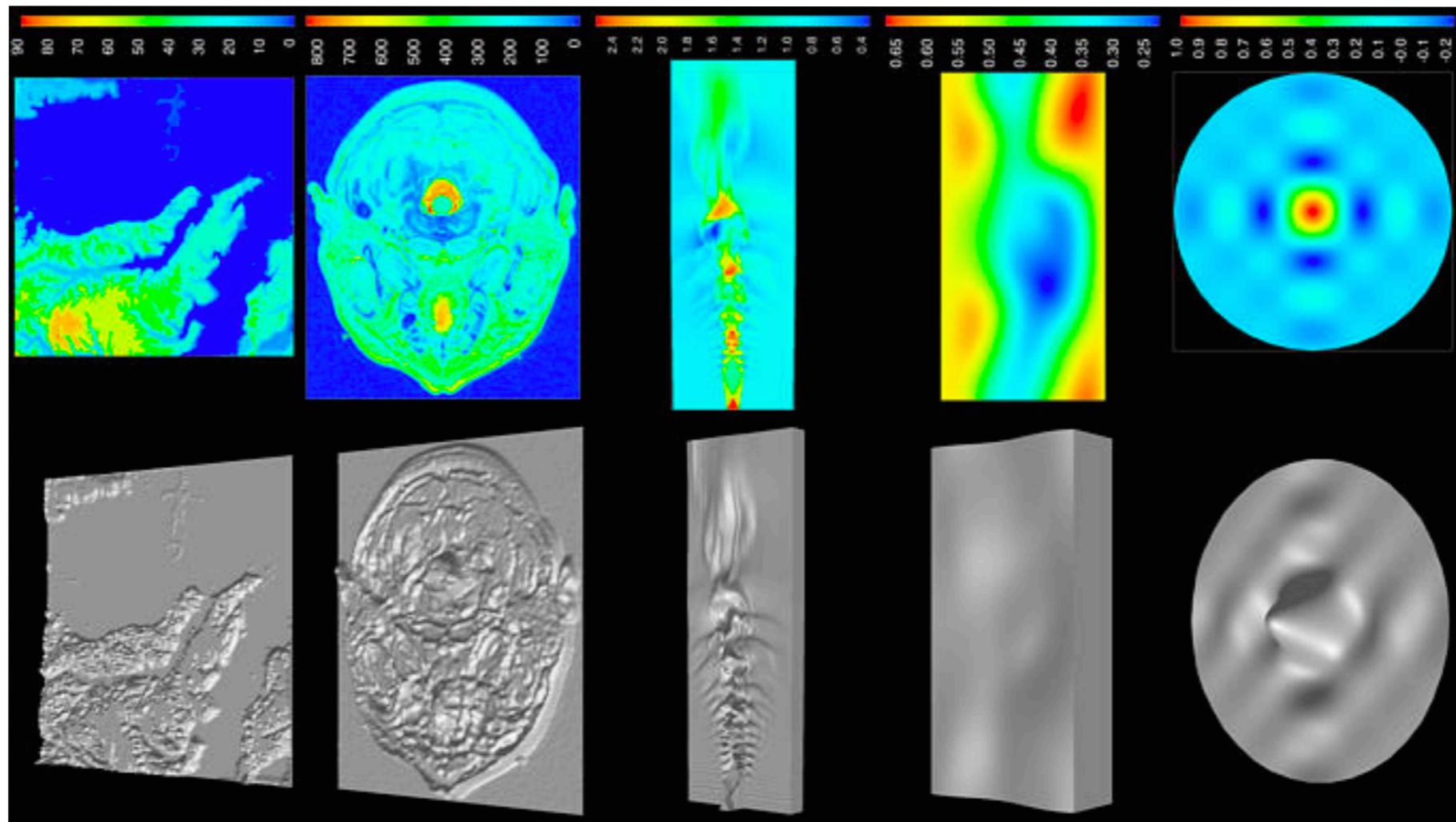
# ColorMap



# ColorMap



# ColorMap



# The type of graphs

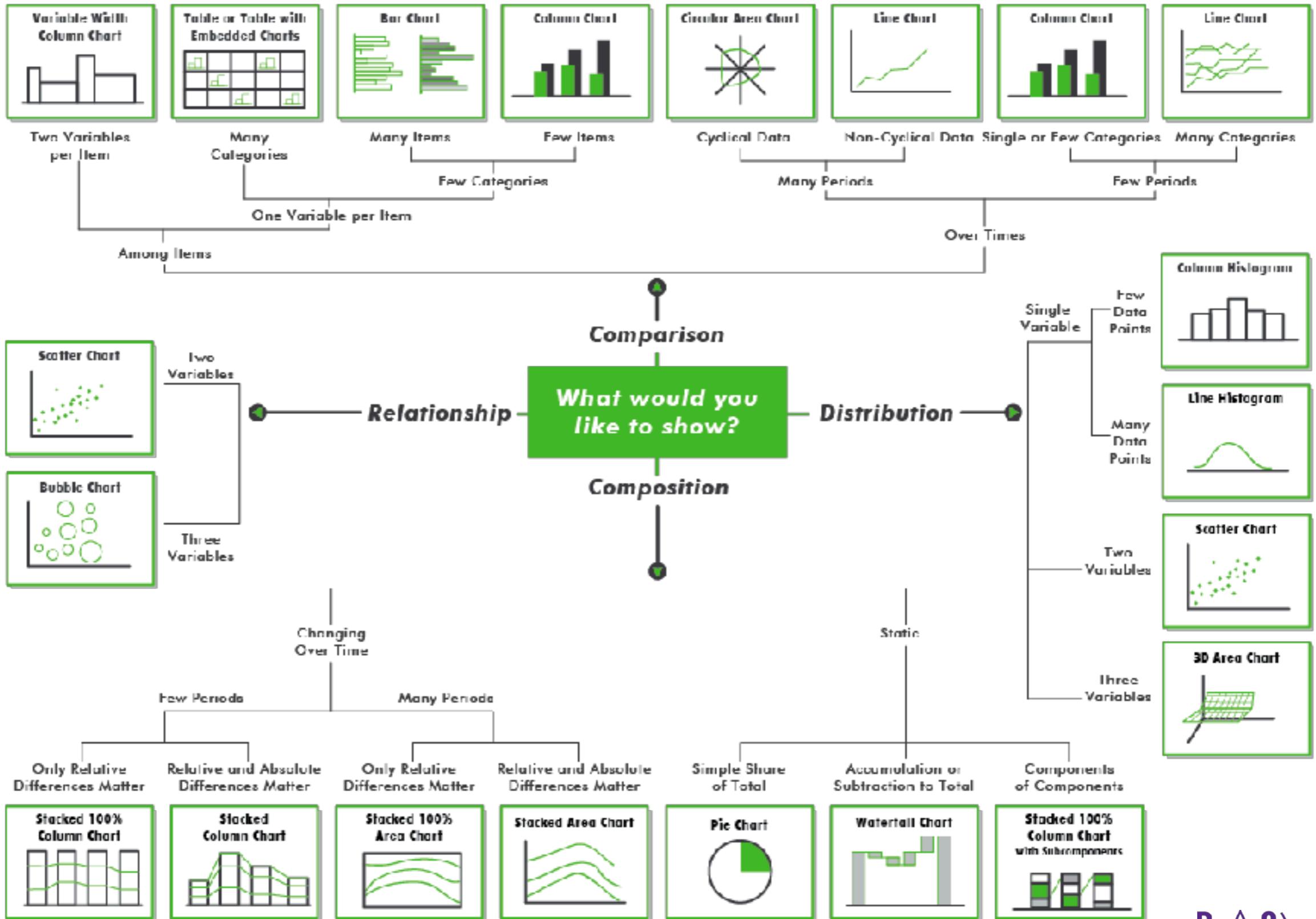
# How do I choose which type of chart or graph to use?

**First, we have to understand the message we want to present with the data!**

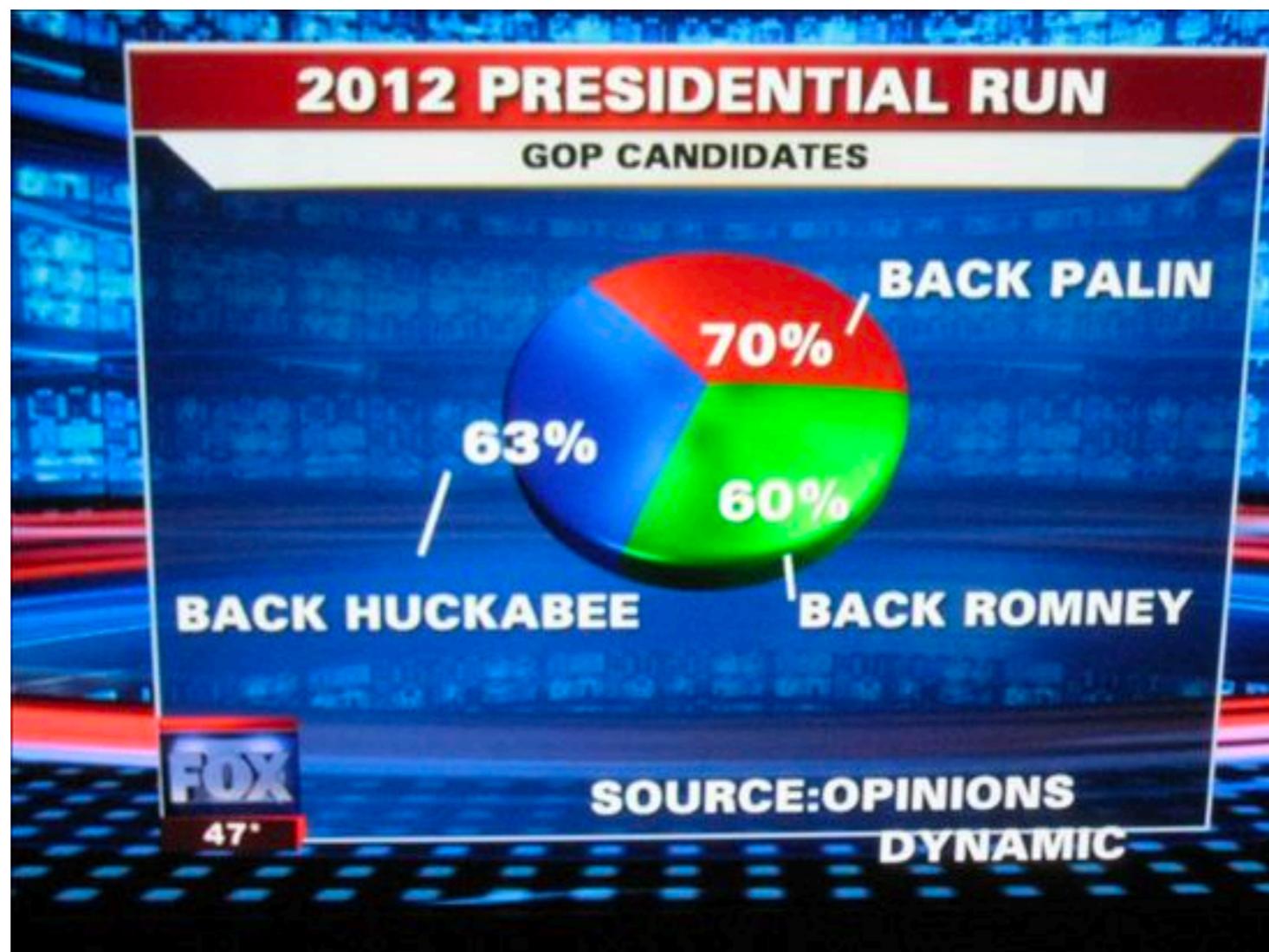
When you're putting together a chart, you're trying to show one of four things with the data you have: a **relationship** between data points, a **comparison** of data points, a **composition** of data, or a **distribution** of data.

- A relationship tries to show a connection or correlation between two or more variables through the data presented, like the market cap of a given stock over time versus overall market trend.
- A comparison tries to set one set of variables apart from another, and display how those two variables interact, like the number of visitors to five competing web sites in a single month.
- A composition tries to collect different types of information that make up a whole and display them together, like the search terms that those visitors used to land on your site, or how many of them came from links, search engines, or direct traffic.
- A distribution tries to lay out a collection of related or unrelated information simple to see how it correlates, if at all, and to understand if there's any interaction between the variables, like the number of bugs reported during each month of a beta

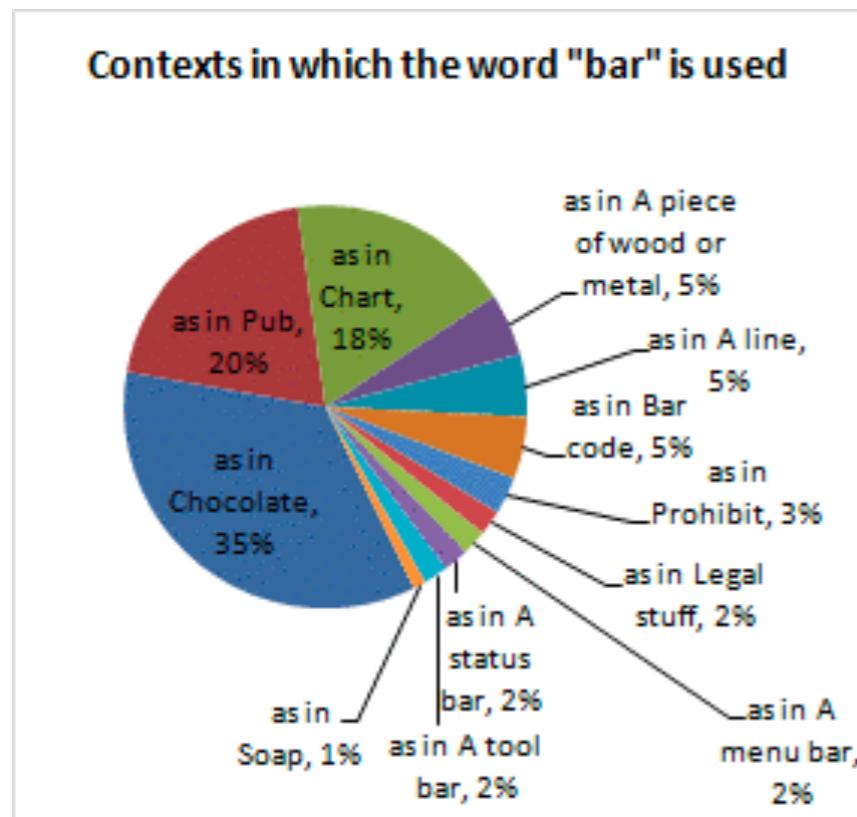
# Data Visualization - Choosing the right Charts



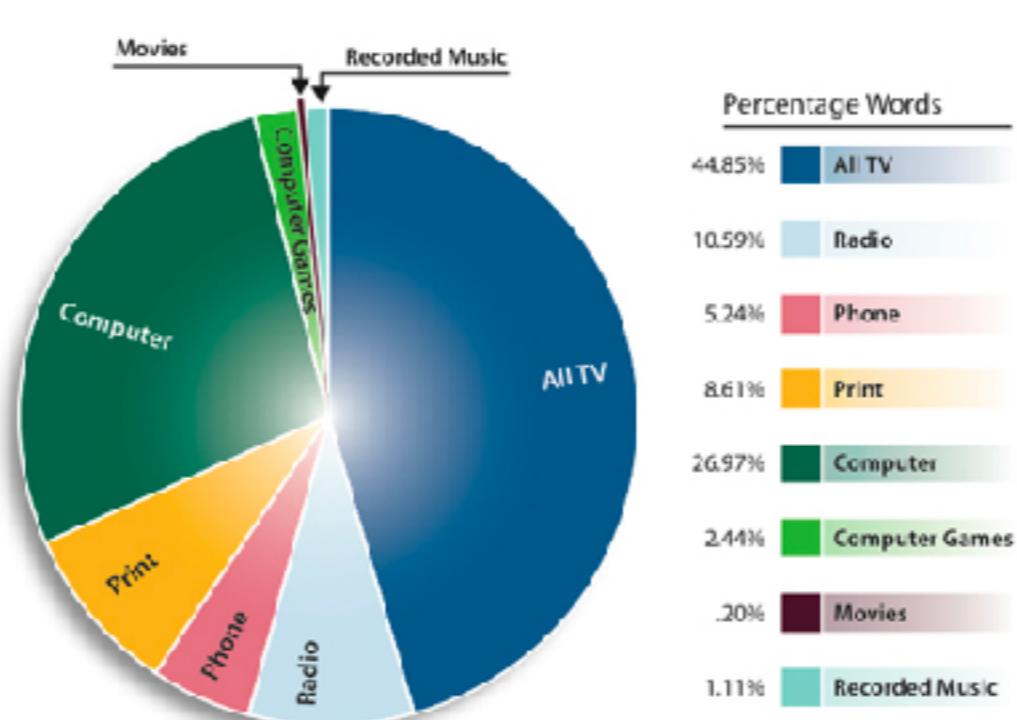
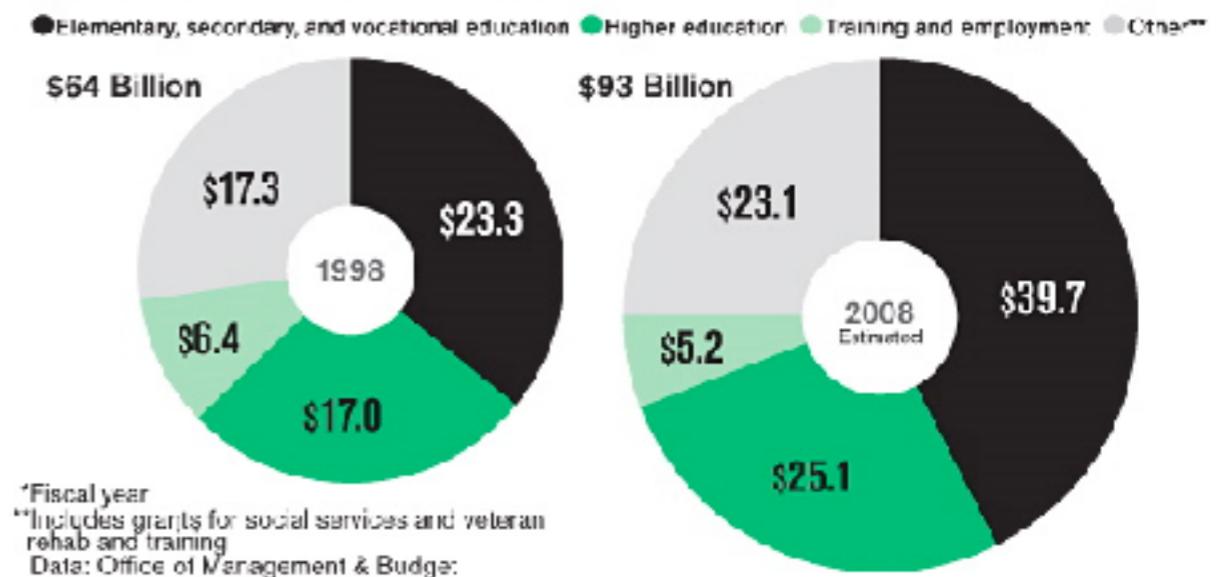
# PieChart?



# PieChart?



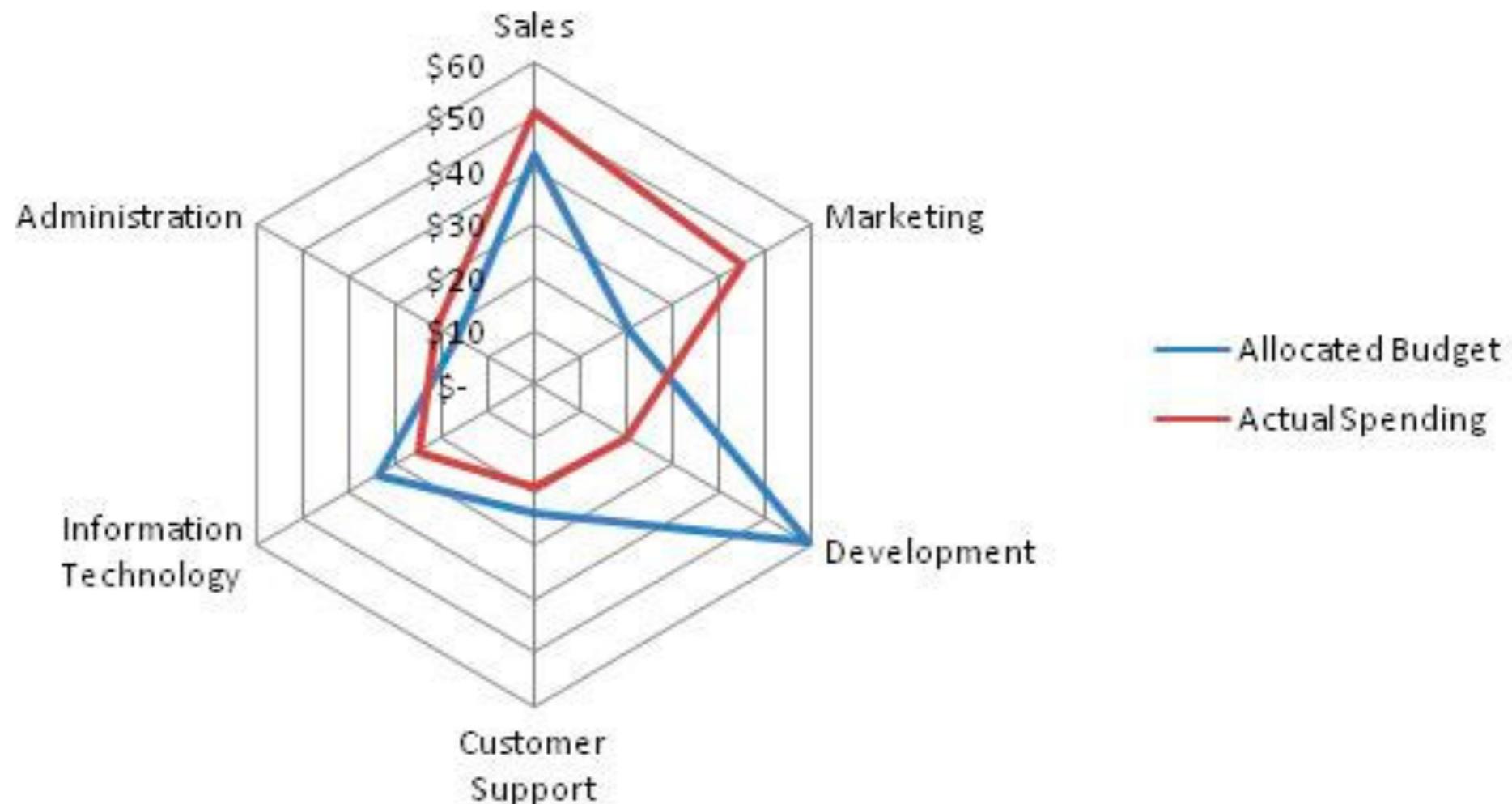
## FEDERAL SPENDING ON EDUCATION AND TRAINING, 2008 DOLLARS\*



# Pie Charts

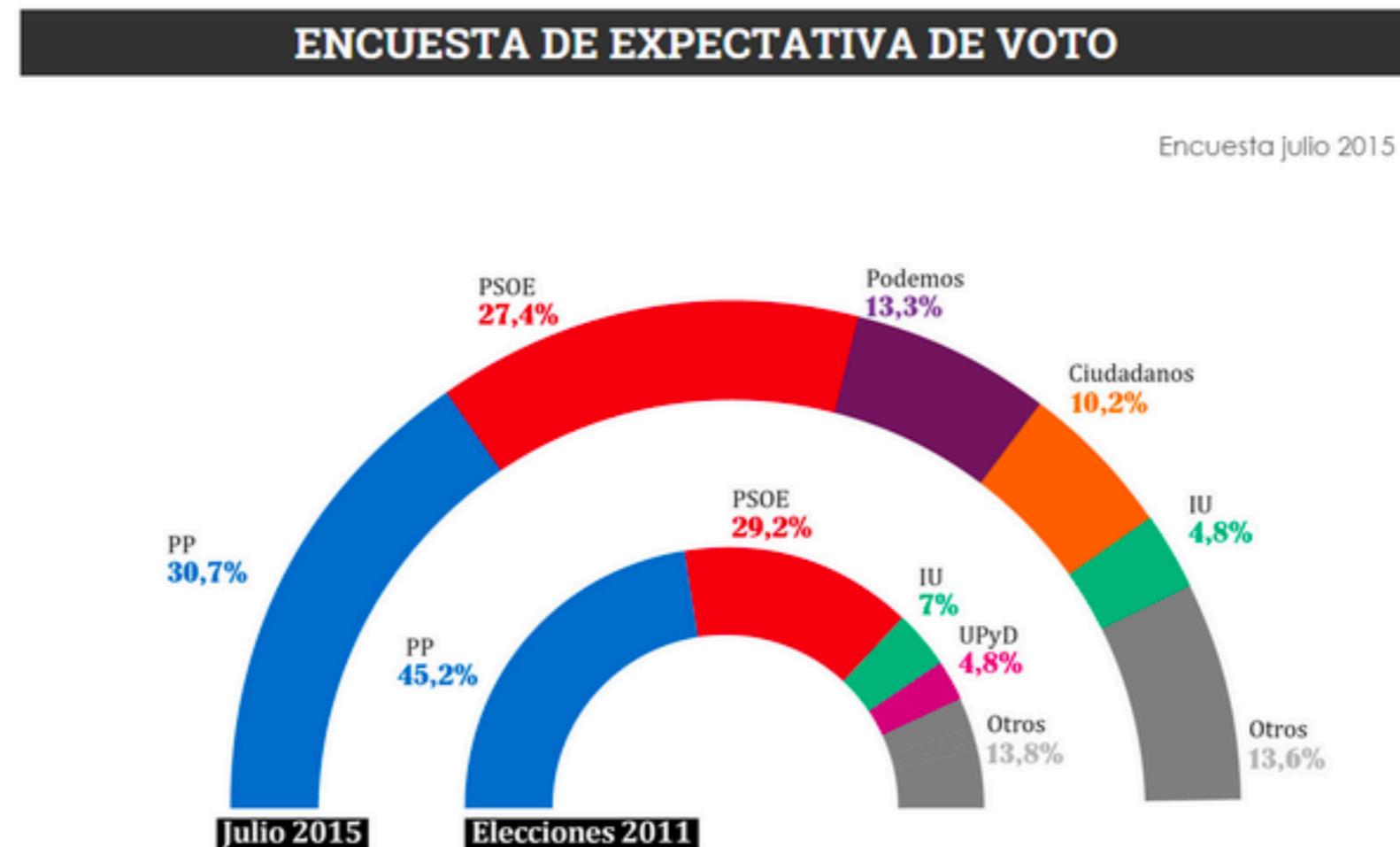
We are **not very good** at measuring **angles**,  
but we recognize **90 and 180 degree angles** with very  
**high precision**

# Radar Graph?

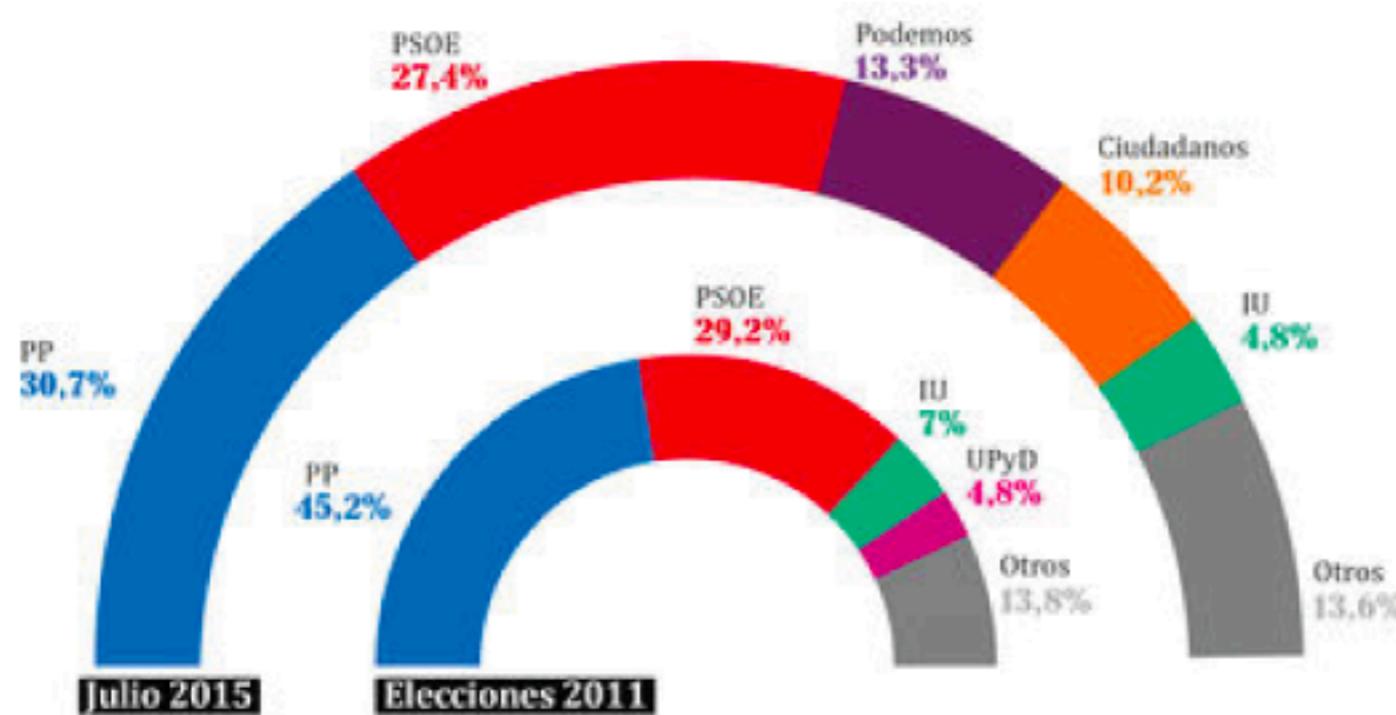


**Do not use this graph!!!**

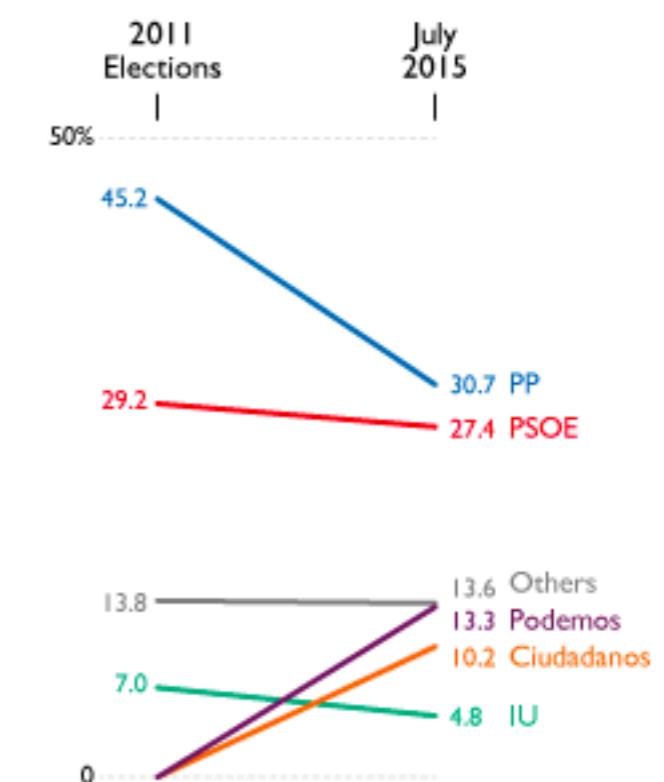
# Is this the proper chart?



## BEFORE



## AFTER



Can we do it better? <http://4.bp.blogspot.com/-ZsRWe33yhRY/VaO3ZVvoJ4I/AAAAAAAFAF8/qlh8n6oJmrs/s1600/BeforeAfterParlamento.png>

**<https://visme.co/blog/best-information-graphics-2016/>**