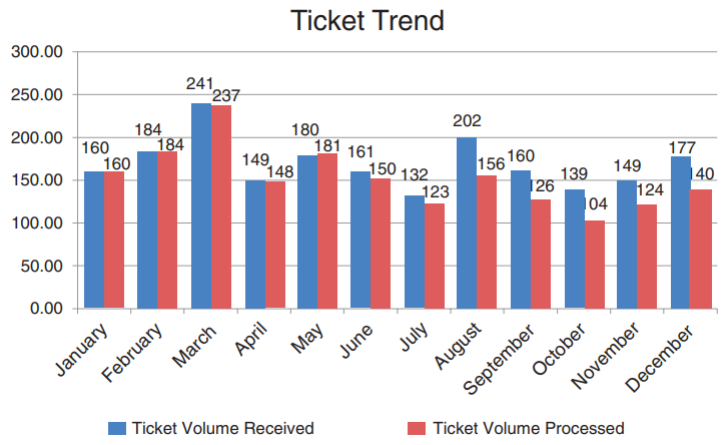
Quality of Visual Representations of Data

Álvaro Figueira, PhD.
MSc in Data Science - Data Visualization

[dcc] DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO U.PORTO

Example 1 (before)



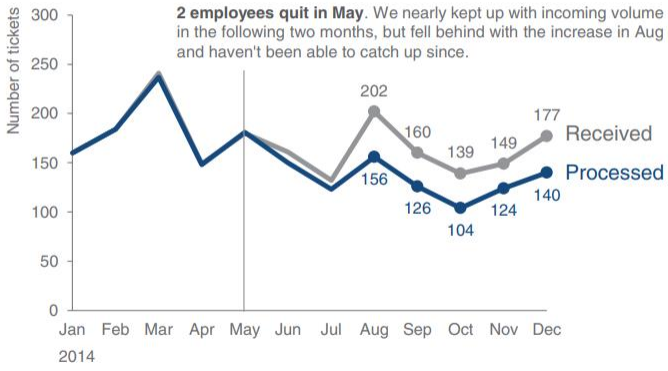
Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

Example 1 (after)

Please approve the hire of 2 FTEs

to backfill those who quit in the past year

Ticket volume over time



Data source: XYZ Dashboard, as of 12/31/2014 | A detailed analysis on tickets processed per person and time to resolve issues was undertaken to inform this request and can be provided if needed.

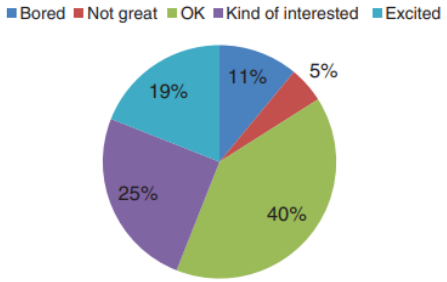
Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

180

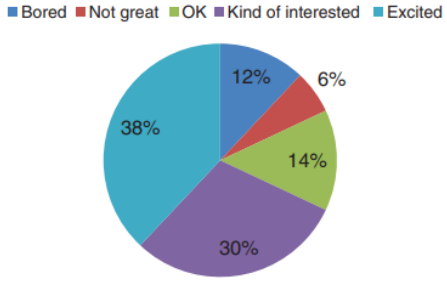
Example 2 (before)

Survey Results

PRE: How do you feel about doing science?



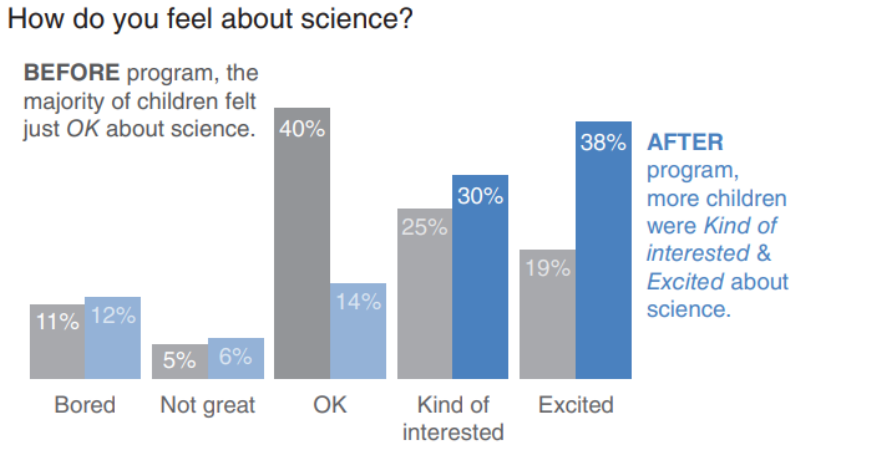
POST: How do you feel about doing science?



Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

181

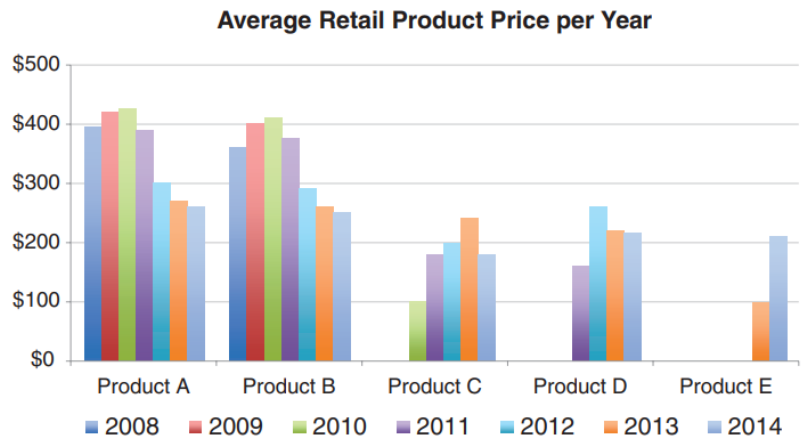
Example 2 (after)



Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

182

Example 3 (before)

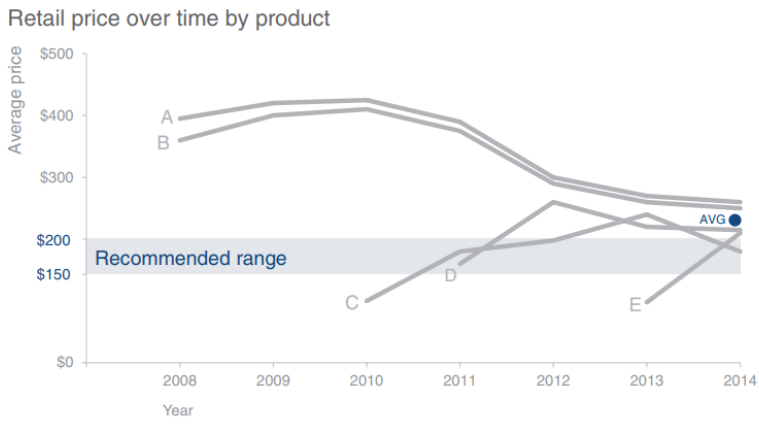


Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

183

Example 3 (after)

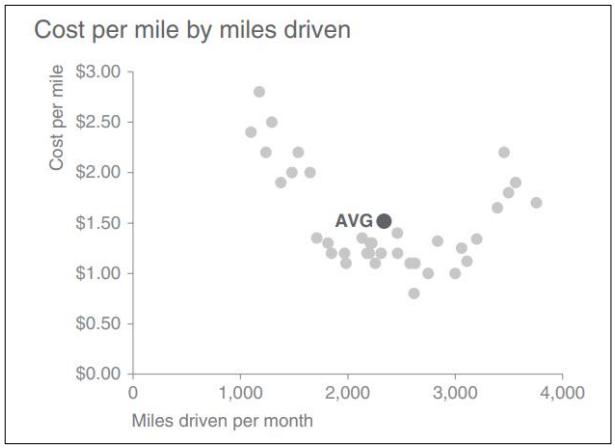
To be competitive, we recommend introducing our product *below the $223 average* price point in the **$150–$200 range**



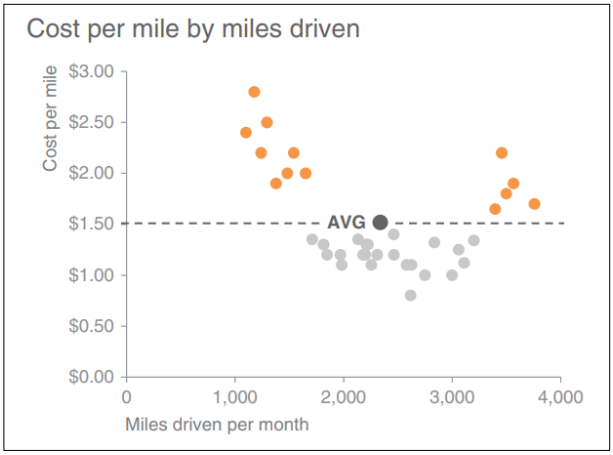
Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

184

Example 4



before



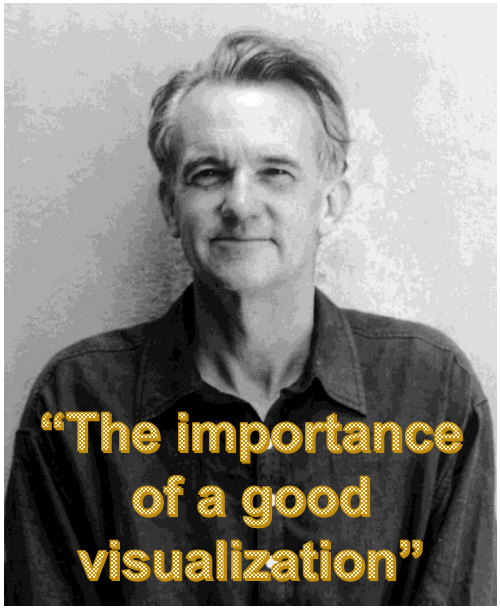
after

Adapted from "storytelling with data", Cole Knaflic, Wiley, 2015.

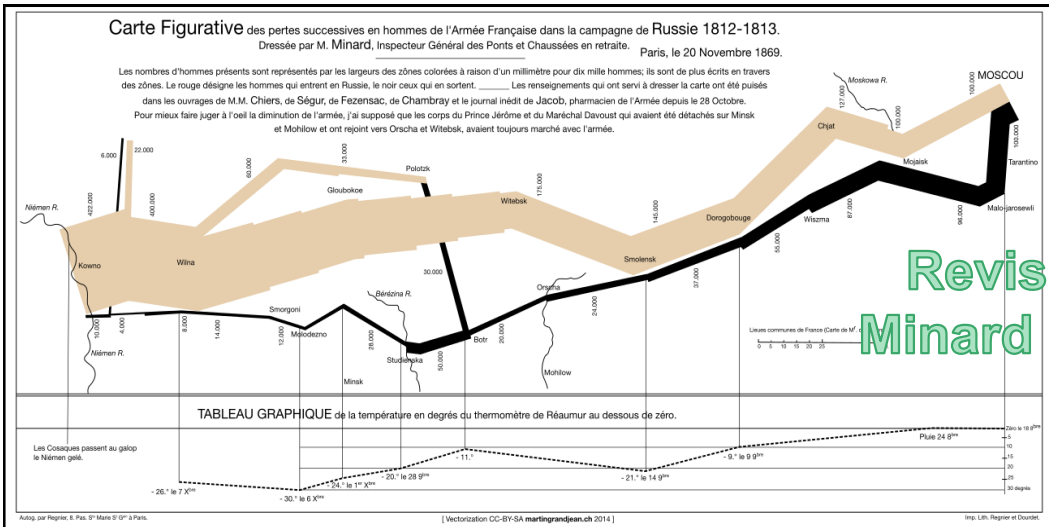
185

Edward Rolf Tufte

An American statistician and professor emeritus of political science, statistics, and computer science at Yale University.
He is noted for his writings on information design and as a pioneer in the field of **data visualization**



186



Revisiting the Minard diagram

Tufte wrote in 1983 that: "It may well be the best statistical graphic ever drawn."

A chart created by French engineer Charles Minard showing Napoleon's 1812 invasion of Russia. The tan bar shows the army marching from the Poland/Russia border. The bar branches and decreases thickness showing parts of the army splitting off to fight various battles, gradually loosing troops on their way to Moscow. The black bar shows the defeated army marching back from Moscow, eventually dwindling to only 10,000 by the time of their return (down from 422,000 at the start of the march).

187

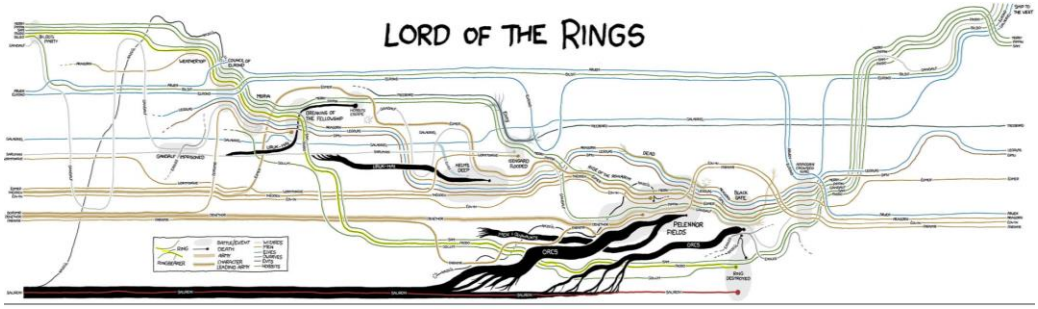
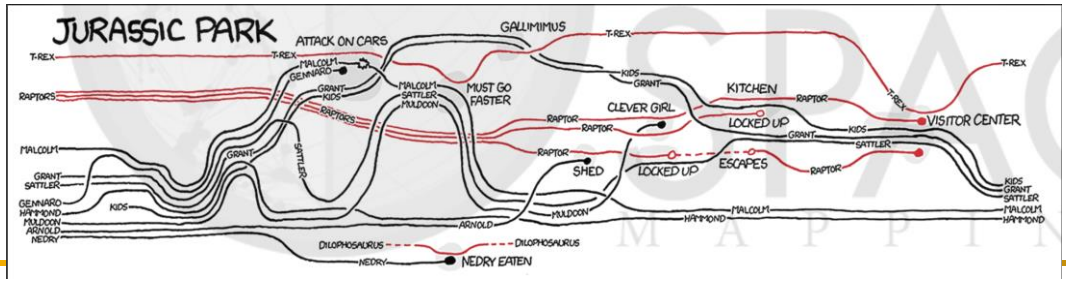
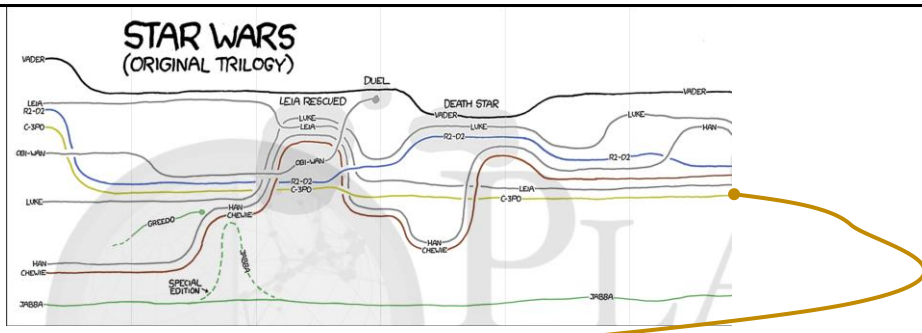
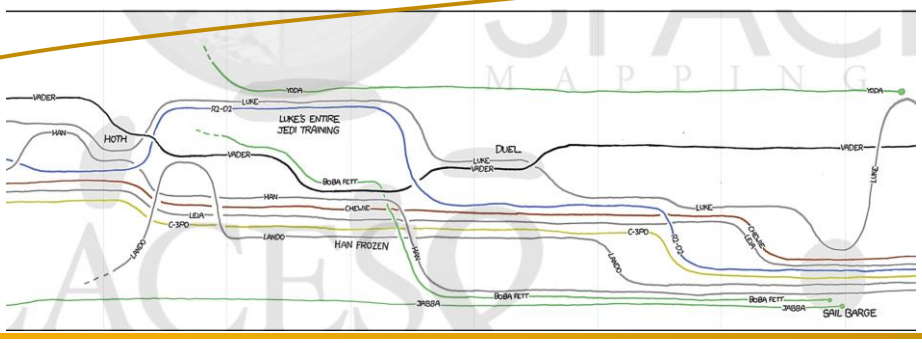


Figure 1. A web comic showing character interactions as they progressed through time in Lord of the Rings. The thick black bar along the bottom represents Sauron's army. The bar branches and decreases thickness showing parts of the army splitting off to fight various battles, gradually loosing troops and dwindling to nothing by the time the ring is destroyed.



188





189

Visualizing the Broad Street Cholera Outbreak

- A severe outbreak of cholera occurred in 1854 near Broad Street in the City of Westminster, London, England, unknowing to people causing over 600 deaths
- Physician Jon Snow identified the source of the outbreak as the public water pump on Broad Street
- Snow used a dot map to illustrate the cluster of cholera cases around the pump
- He also used statistics to illustrate the connection between the quality of the water source and cholera cases.

190

Aspects to account for building good graphs

1. Graphical integrity
2. Data-ink
3. Chartjunk
4. Data density
5. Small multiples

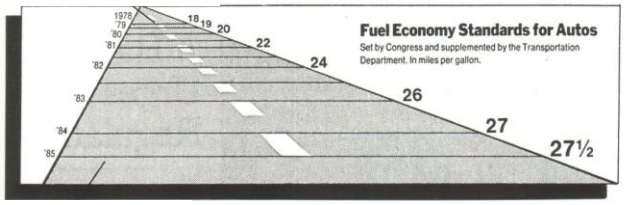
The Visual Display of Quantitative Information, Tufte, 1983

191

Graphical integrity

Visual representations of data must tell the truth!

Tufte shows a whole range of graphs that misrepresent information. He does this by suggesting and calculating a graph's **Lie Factor** which is the **ratio between the size of the effect shown in the graphic by the size of the effect in the data**. If the Lie Factor is greater than 1 the graph overstates the effect.

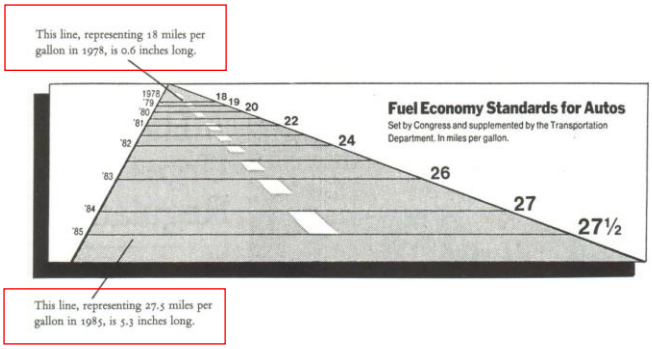


192

Graphical integrity

Visual representations of data must tell the truth!

Tufte shows a whole range of graphs that misrepresent information. He does this by suggesting and calculating a graph's **Lie Factor** which is the **ratio between the size of the effect shown in the graphic by the size of the effect in the data**. If the Lie Factor is greater than 1 the graph overstates the effect.



193

Graphical integrity

Tufte enunciates five principles of graphical integrity:

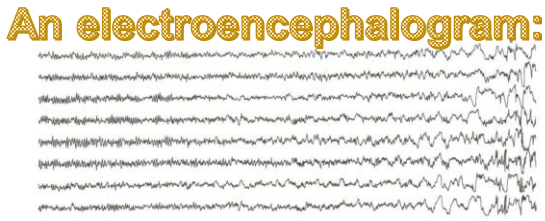
1. The **representation of numbers**, as physically measured on the surface of the graph itself, should be directly proportional to the numerical quantities represented.
2. Show **data variation**, not design variation.
3. The number of information carrying dimensions depicted should not exceed the **number of dimensions in the data**.
4. Clear, detailed and thorough **labeling** should be used to defeat graphical distortion and ambiguity.
5. Graphics must **not refer to data out of context**.

194

Data-ink

Data Ink is the ink on a graph that represents data.Tufte asserts that good graphical representations **maximize data-ink** and **erase as much non-data-ink as possible**. He defines the **data-ink ratio** by being the **proportion of a graphic that cannot be erased without loss of data-information**. He then advises to:

- above all else, show data;
- maximize the data-ink ratio;
- erase non-data-ink;
- erase redundant data-ink;
- revise and redo.



An electroencephalogram would have a data-ink ratio of 1.

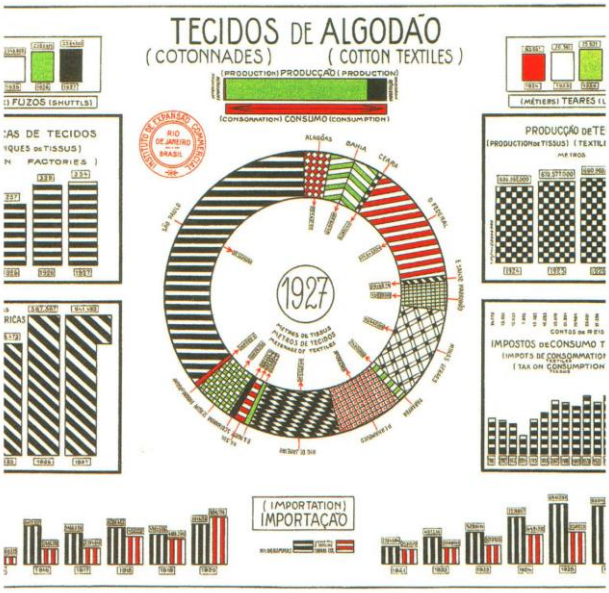
195

Chartjunk

Tufte devises upon what he calls **chartjunk**—the **excessive and unnecessary use of graphical effects** in graphs. He exemplifies this by pointing defects in graphics such as:

- *moiré patterns*
- heavy grids
- self-promoting graphs

that are used to *demonstrate the graphic ability of the designer rather than to display data*.

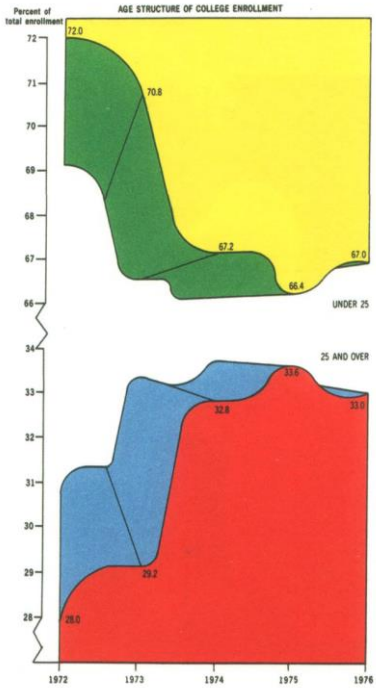


196

Chartjunk

- Usage of additional colors for non existent dimensions in data.
- A series of weird three-dimensional displays appearing in the magazine of American Education in the 1970's delighted the connoisseurs of the graphically preposterous.

"This may well be the worst graphic ever to find its way into print." (Tufte, 1983)

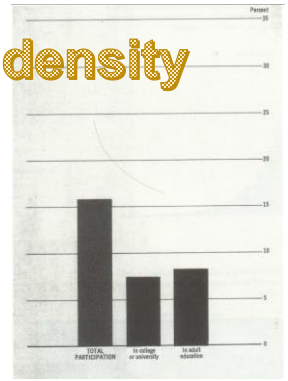


What are the problems with this chart?

197

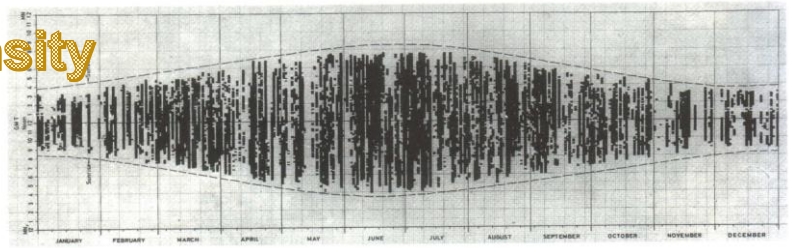
Data density

Low density



The **data density** of a graph is the **proportion of the total size of the graph that is dedicated to display data**. Tufte argues about the advantages of high data density graphs—data density should be maximized within reason. He states that most graphs can be shrunk way down without losing legibility or information (The *Shrink Principle*), economizing space and bringing room to the portrayal of more information.

High density



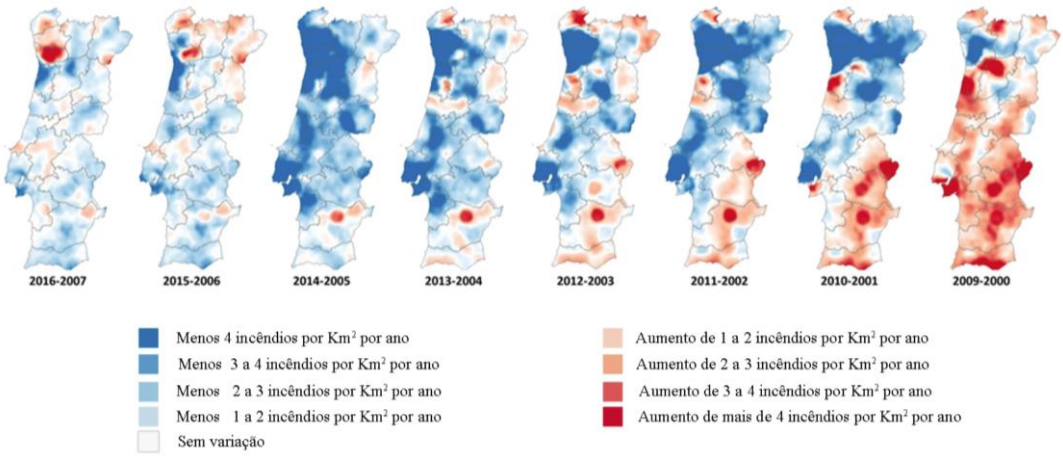
198

Small Multiples

Small multiples are **series of the same high-density graphic repeated in one layout**. Tufte states that small multiples are a great tool to visualize large quantities of data and with a **high number of dimensions**, enabling **to rapidly compare**, for example, the nature of a whole dataset across one selected dimension.

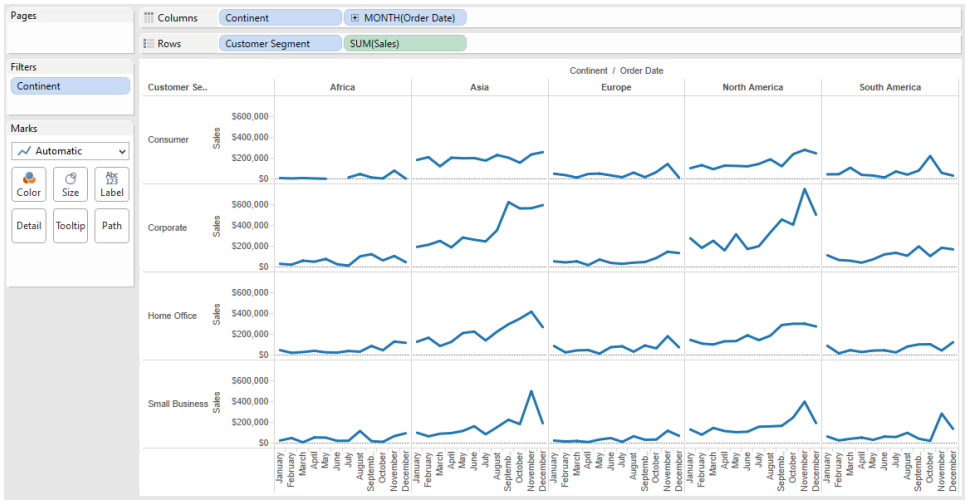
199

Small Multiples



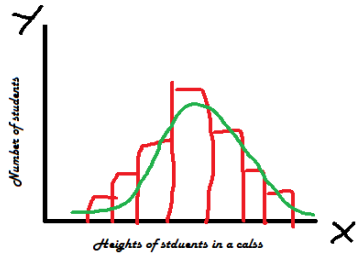
200

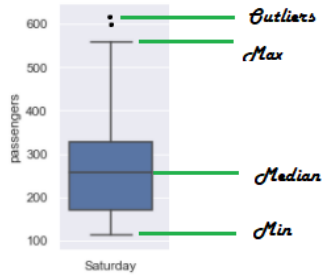
Small multiples



201

The four pillars of data visualization (1)





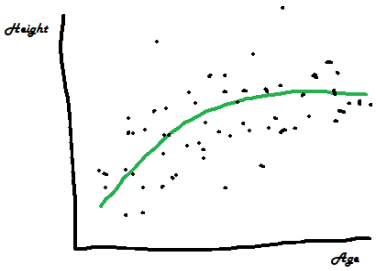
1) Distribution

An important concept in statistics and data science is distribution. Distribution generally refers to the probability of occurrence of an outcome. In a distribution of 100-coin flips how many will get heads and how many tails? Frequency distributions like this are presented in histograms or curves.

Above is a representation of students' heights distribution in a swimming class. The x-axis shows different height categories and y-axis has the number of students in each category. The boxplot above represents the distribution of the number of air passengers on Saturdays over several years. This single plot reveals so much information — the mean/median number of passengers on Saturdays, the minimums and maximums, the outliers and more!

207

The four pillars of data visualization (2)



2) Relationship

Trees grow taller as they get older in the early years. That's a relationship between two variables — height and age.

height = f(age)

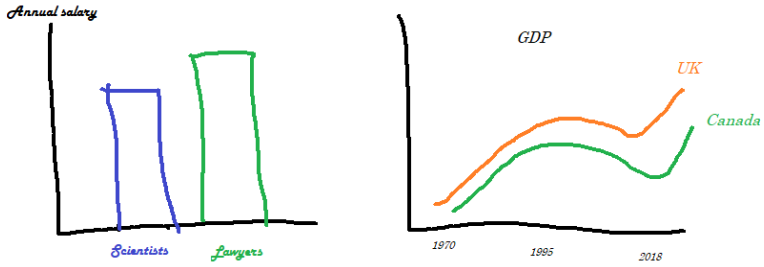
In another example, the price of a house depends on the number of beds, number of bathrooms, location, square footage etc. This is a relationship between one dependent and many explanatory variables.

price = f(beds, baths, location, area)

If you look at a dataset just as numbers, there is no way to identify these relationships. But in fact, you can, without going into complex statistical analysis, with the help of a good visualization.

208

The four pillars of data visualization (3)



3) Comparison

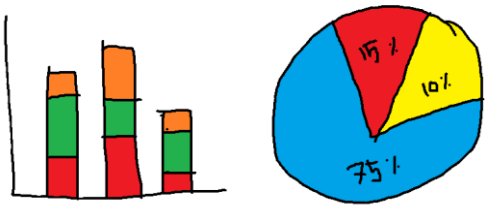
The third cornerstone of data visualization is *Comparison.* This kind of visual material compares multiple variables in datasets or multiple categories within a single variable.

The one on left compares a variable (salary) between two groups of observations (scientists vs lawyers) on a bar chart. The right panel is also a comparison chart — in this case, comparing a variable (GDP) between two groups (UK and Canada) but along a time dimension.

209

209

The four pillars of data visualization (4)



TIP:
it is important to understand what is the purpose of the visualization and what information you want to communicate

4) Composition

The purpose of these charts is to show the composition of one or more variables in absolute numbers and in normalized forms (e.g. percentage).

Composition charts are some of visualization techniques that nowadays have limited use cases (do you really need a pie chart to show a composition of yellow 10% and red 15%?). Nevertheless, sometimes they can present information in a visually aesthetic and familiar, vintage fashion.

210

210

Gestalt principles of visual perception

When it comes to identifying which elements in our visuals are signal (the information we want to communicate) and which might be noise (clutter), we will consider the Gestalt Principles of Visual Perception.

The Gestalt School of Psychology set out in the early 1900s to understand how individuals perceive order in the world around them.

They proposed "the principles of visual perception", which are still accepted today, and that define how people interact with and create order out of visual stimuli.

We'll discuss six principles here: proximity, similarity, enclosure, closure, continuity, and connection.

211

Proximity

We tend to think of objects that are physically close together as belonging to part of a group.

We naturally see the dots as three distinct groups because of their relative proximity to each other

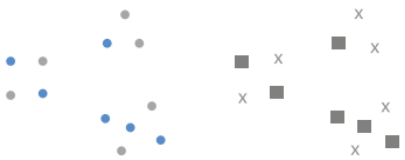
We can leverage this way that people see in table design.

Simply by differentiating the spacing between the dots, our eyes are drawn either down the columns in the first case or across the rows in the second case.

212

212

Similarity

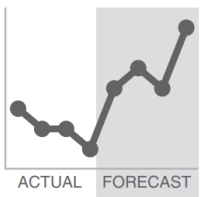


Objects that are of similar color, shape, size, or orientation are perceived as related or belonging to part of a group.

We naturally associate the blue circles together on the left or the grey squares together on the right.

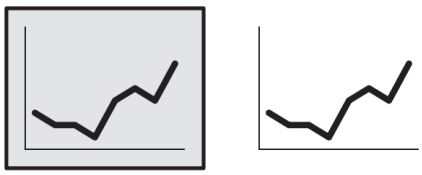
Enclosure

We think of objects that are physically enclosed together as belonging to part of a group. It doesn't take a very strong enclosure to do this: light background shading is often enough.



Closure

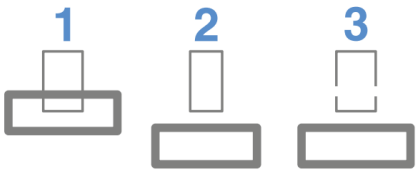
The closure concept says that people like things to be simple and to fit in the constructs that are already in our heads. Because of this, people tend to perceive a set of individual elements as a single, recognizable shape when they can—when parts of a whole are missing, our eyes fill in the gap. For example, the elements in the figure will tend to be perceived as a circle first and only after that as individual elements



215

215

Continuity

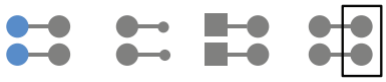


The principle of continuity is similar to closure: when looking at objects, our eyes seek the smoothest path and naturally create continuity in what we see even where it may not explicitly exist. In the example, if we take the objects (1) and pull them apart, most people will expect to see what is shown next (2), whereas it could as easily be what is shown after that (3).

216

216

Connection

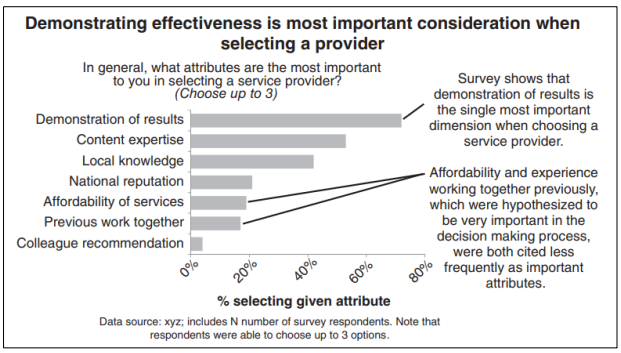


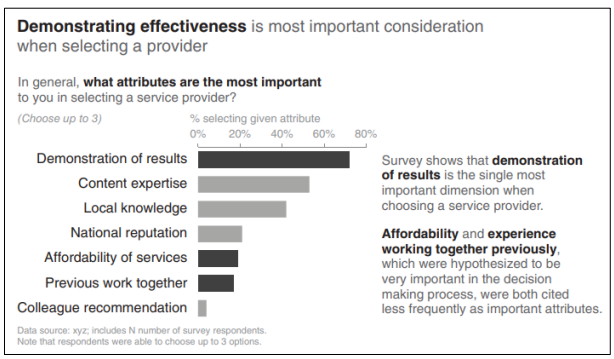
We tend to think of objects that are physically connected as part of a group. The connective property typically has a stronger associative value than similar color, size, or shape.

Note when looking at the figure, our eyes probably pair the shapes connected by lines (rather than similar color, size, or shape): that's the connection principle in action.

The connective property isn't typically stronger than enclosure, but you can impact this relationship through thickness and darkness of lines to create the desired visual hierarchy.

Example of the use of the Gestalt Principles





Wrap up

It is important in a visualization, when wanting to communicate:

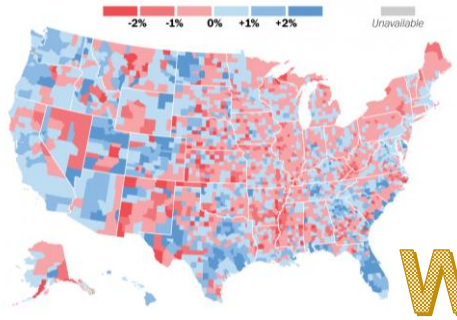
- Make the point of the visualization clear
- Do not confuse or distract the viewer
- Create an honest visualization
- Check the Gestalt principles
- Check Tufte's principles of good design

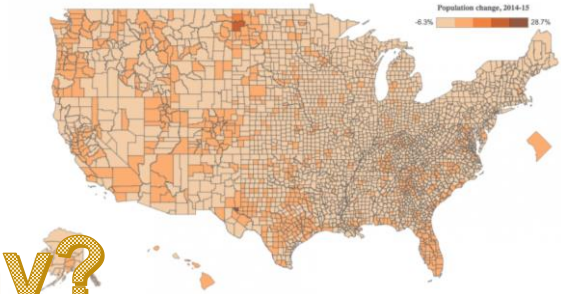
Exercise

The two graphics are built exactly from the same source. Which one do you prefer?

A year of population change

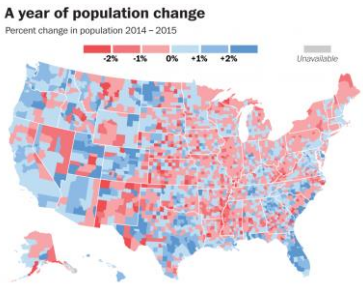
Percent change in population 2014 - 2015

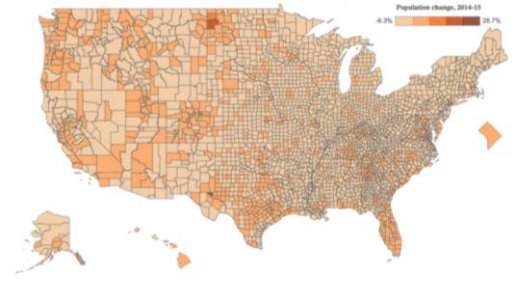




Why?

https://www.washingtonpost.com/news/wonk/wp/2016/04/11/the-dirty-little-secret-that-data-journalists-arent-telling-you/





The map above tells the story of a year's population change in the United States, according to the latest census data.
It shows **where the population is growing**, like the coasts, the Sun Belt and the oil fields of western North Dakota.
It shows, too, **where numbers are in decline** -- along the Mississippi River and in much of the rural Northeast, from northern Maine down through New York, western Pennsylvania and into the heart of the Appalachians.

"Population growth slowed last year in some of the nation's most expensive counties, like those in California's Silicon Valley, and picked up in more affordable counties in the Sun Belt," according to the text associated with the map.

However, it's next to impossible to discern that just from looking at the map. In fact, it's hard to discern just about anything.

Visualizing data is as much an art as a science and where to set a color threshold, how many thresholds to set, etc. -- can radically alter how numbers are displayed and perceived.
On the Right, data was sliced up the full range -- from minus-6.3 percent (Terrell County, Tex.) to plus-28.7 percent (Loving County, Tex.) -- into five buckets of equal size: -6.3 to 0.7, 0.7 to 7.7, 7.7 to 14.7, etc.
Assigning a color to each bucket, color each county according to which bucket its population falls into, and *voila*! A map, right?
The problem is that while the buckets are nice and evenly distributed, the numbers are not. There are 3,141 counties in the Census Bureau's data set, and 3,138 of them fall into either the first or second buckets. Only three counties had a population gain greater than 7.7 percent. **So those three darker shades of the scale are essentially unused, and the entire map gets washed out into two similar colors**.
But there's a potentially bigger problem, too. Some counties lost population, while others gained. That's a pretty big distinction. Mapmakers often respect big distinctions like that by using a *bivariate* color scale -- say, one set of colors for positive values (like blue), and another set of colors for negative ones (like red).

221