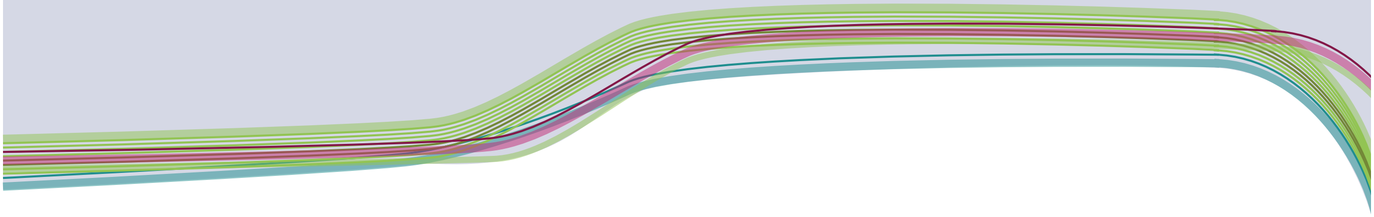


5

Basic Principles of Visualization

"In the course of executing that design, it occurred to me that tables are by no means a good form of conveying such information.... Making an appeal to the eye when proportion and magnitude are concerned is the best and readiest method of conveying a distinct idea."

—William Playfair, *The Statistical Breviary*

There is a time in every class and workshop when someone raises her hand and asks: **How do you know that you have chosen the right graphic form to represent your data?** When is it appropriate to use a bar chart, a line chart, a data map, or a flow diagram? Geez, if I had the answer to all that I'd be rich by now. I invariably reply: "I have no idea, but I can give you some clues to make your own choices based on what we know about why and how visualization works."

In his book *Misbehaving: The Making of Behavioral Economics* (2015), University of Chicago's **Richard H. Thaler** recounts an anecdote that may be useful for any teacher. At the beginning of his career as a professor, Thaler made many of his students mad by designing a midterm exam that was deemed too hard. The average score, on a scale from 0 to 100, was 72. He got a lot of complaints about it.

Thaler decided to run an experiment. In the next exam, he set the maximum score to 137 points. The average ended up being 96 points. His students were thrilled.

Thaler kept the 137 mark in subsequent exams, and also added this line to his syllabus: “Exams will have a total of 137 points rather than the usual 100. This scoring system has no effect on the grade you get in the course, but it seems to make you happier.” It certainly did. After he made this change, Thaler never got any pushback from students again—even if he told them beforehand that they were going to be tricked!

Try to mentally visualize these numbers: 72 versus 100, and 96 versus 137. The first pair is easy. Human brains perform nicely at simple arithmetic with rounded figures. But they are abysmal when forced to manipulate any other kind of number without aid. It’s hard to picture 96 in comparison to 137 in your head. It’d be better if I did it for you on a piece of paper or on a screen. Your wish is my command (**Figure 5.1**; the figures are shown twice, as a dot plot and as a pair of pie charts).[1]

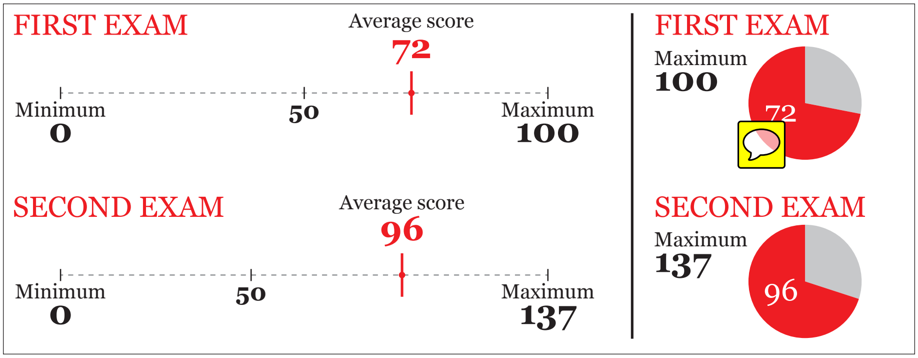


Figure 5.1 Seventy-two over 100 is a better score than 96 over 137. Funny, right?

[1] This is a 2013 Tweet by visualization author Edward R. Tufte, who got things wrong by trying to be too strict: “Pie chart users deserve same suspicion+skepticism as those who mix up its/it’s, there/their. To compare, use little table, sentence, not pies.” I am no fan of pie charts, but in this case, even if they are inferior to the dot plots, the two pie charts work better than a sentence or a table.

It turns out that Thaler's second exam was *harder* than the first one. A score of 96 out of a maximum of 137 is a 70 percent score, in comparison to the 72 percent average of the first exam. But even if you're aware of that—because you know how to transform a raw score into a percentage— **96 over 137 still *feels* higher than 72 over 100**. That's a bug of the wetware sloshing inside your skull. **Most people grasp the truth of an assessment only when they unequivocally *envision* the evidence for it**, something that our kludgy brains alone often can't do well. **That's why visualization works**.

Visually Encoding Data

Vision is the most developed sense in the human species. A huge chunk of our brains is devoted to gather, filter, process, organize, and interpret data sent from the retinas at the back of our eyes. We've evolved to be really fast and effective at detecting visual patterns and exceptions to those patterns. It is only natural, then, that a set of methods consisting on **mapping data into visual properties**—spatial and otherwise—would prove to be so powerful.

"Mapping data into visual properties." That's quite a mouthful, so let me explain. Suppose that you want to compare the unemployment figures of five countries currently in economic recession. Let's call them A, B, C, D, and E.

These figures, which I am withholding for now, are our data. The mapping part consists of choosing the property that will better let readers accomplish a particular goal ("comparing accurately") without being forced to read all numbers. I have encoded them in several ways in **Figure 5.2**. Which one of these graphics would you choose?

I'd go with length, height, or position, and here's why: if you don't know what the numbers are before you see the rest of the charts—the ones based on area, angle, weight, and color—can you quickly identify the highest or lowest unemployment rates, and compare them to the others well? It's harder, isn't it?

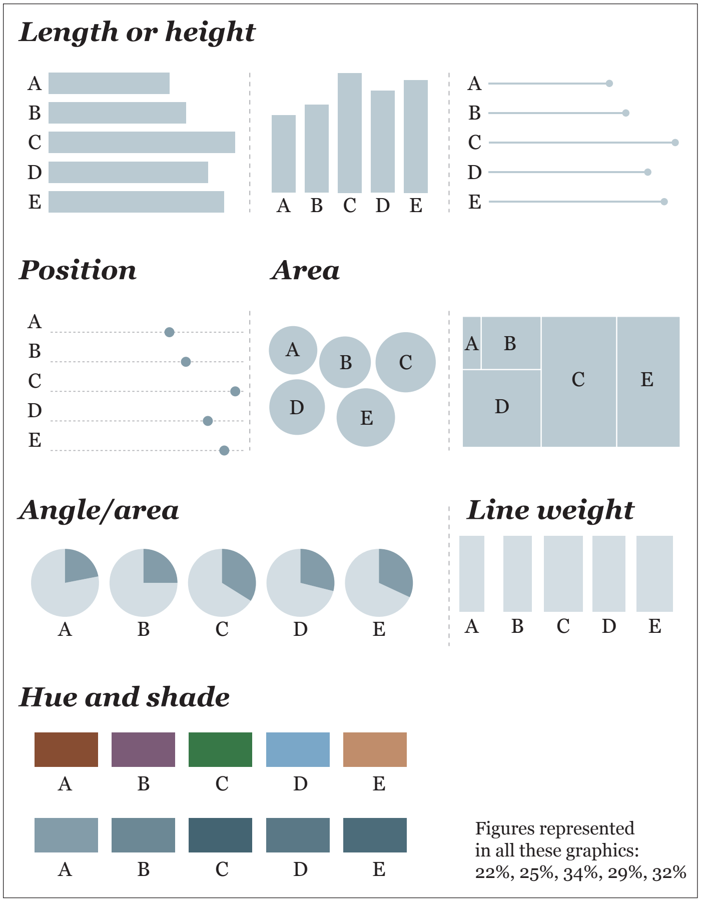


Figure 5.2 Different methods of encoding the same small data set.

Thus, here are some preliminary suggestions to find the right graphic forms for your visualizations:

1. **Think about the task or tasks you want to enable**, or the messages that you wish to convey. Do you want to compare, to see change or flow, to reveal relationships or connections, to envision temporal or spatial patterns and trends? We could summarize this point with a sentence that sounds tautological, but isn't: **Plot what you need to plot.** And if you don't know what it is that you need to plot yet, plot many features of your data until the stories it may hide rise up.

2. **Try different graphic forms** that help readers complete those tasks. If you have more than one task on your wish list, you may need to represent your data in different ways.

3. **Arrange the components of the graphic** so as to make it as easy as possible to extract meaning from it. Whenever it's appropriate, add interactivity to your visualization so people can organize the data at will.

4. **Test the outcomes** yourself, and with people who are representative of your audience—even if it is in a non-scientific, non-systematic manner.

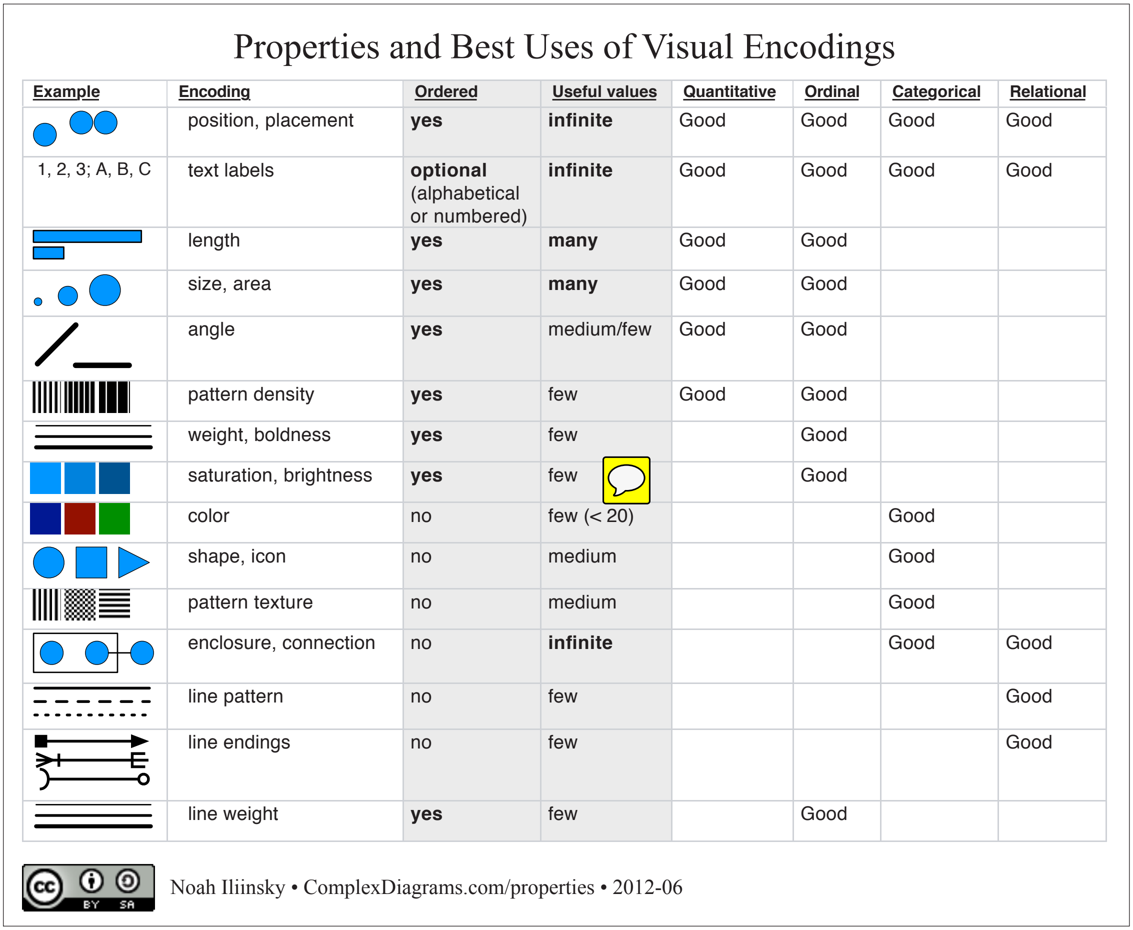
Notice that I haven't said anything yet about making the visualization elegant and appealing. For now, let's just focus on its skeleton, and on minimizing the opportunities for misunderstandings.

Choosing Graphic Forms

Numerous authors have developed sophisticated methods to choose appropriate ways of encoding data depending on what you want to reveal: Willard C. Brinton, Jacques Bertin, William Cleveland, Naomi Robbins, Leland Wilkinson, Isabel Meirelles, Stephen Few, Katy Börner, Nathan Yau, Tamara Munzner, Stephanie Evergreen, Stephen Kosslyn... just to name a few off the top of my head.

In these two pages I am showcasing **Noah Iliinsky's** table of Properties and Best Uses of Visual Encodings (**Figure** 5.3), and **Ann K. Emery's** Essentials website (**Figure** 5.4). Devote some time to explore them.

My favorite tool to make choices on how to present quantitative data, though, is a **hierarchy of elementary perceptual tasks,** or methods of encoding, that was put together in the **80s by two statisticians who were then working at Bell Labs, William S. Cleveland** and **Robert McGill**, and that was later redesigned by Cleveland himself to be included in his magnum opus *The Elements of Graphing Data*. You can see my own version of that scale on **Figure 5.5**, where I added a few examples of of the graphics mainly associated to each step.

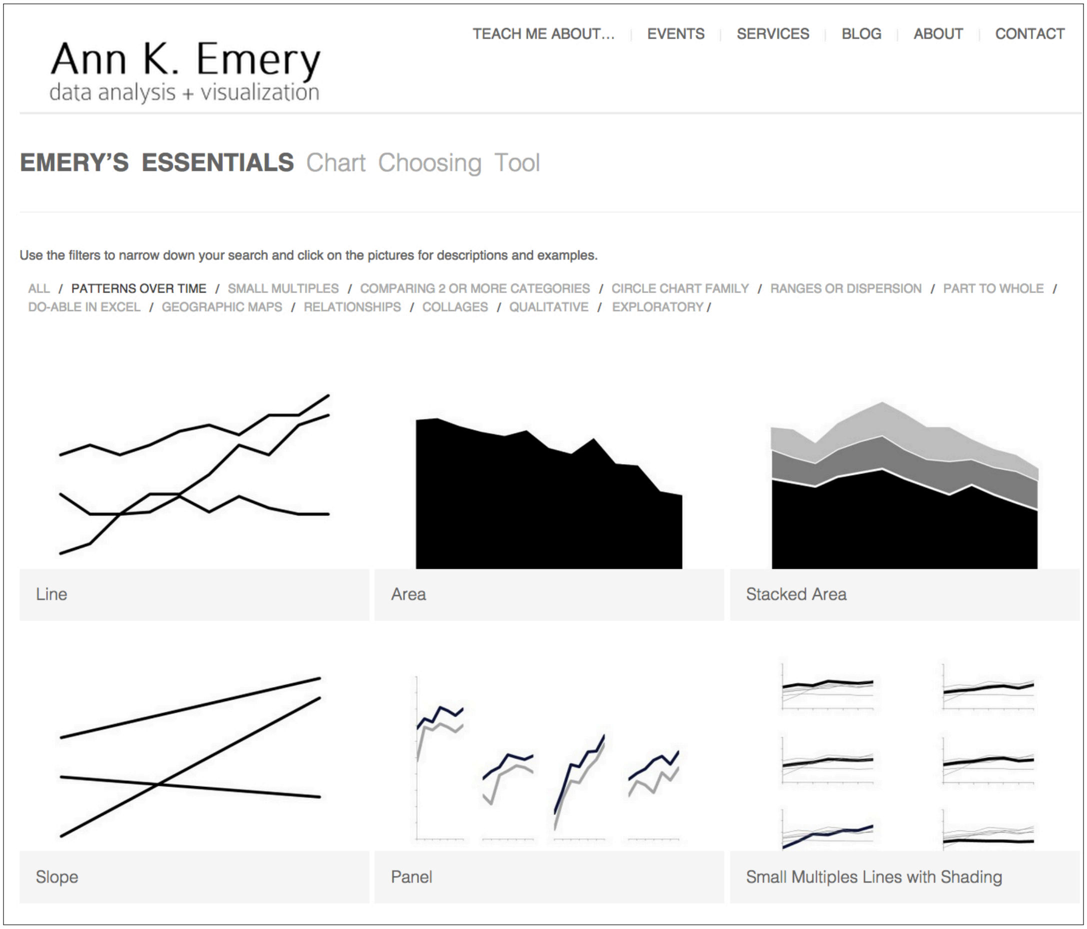


Properties and Best Uses of Visual Encodings

Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
● ●●	position, placement	**yes**	**infinite**	Good	Good	Good	Good
1, 2, 3; A, B, C	text labels	**optional** (alphabetical or numbered)	**infinite**	Good	Good	Good	Good
▬	length	**yes**	**many**	Good	Good		
• ● ⬤	size, area	**yes**	**many**	Good	Good		
/ —	angle	**yes**	medium/few	Good	Good		
▥	pattern density	**yes**	few	Good	Good		
☰	weight, boldness	**yes**	few		Good		
🟦🟦🟦	saturation, brightness	**yes**	few		Good		
🟦🟥🟩	color	no	few (< 20)			Good	
● ■ ▶	shape, icon	no	medium			Good	
▥ ▩ ☰	pattern texture	no	medium			Good	
● ●–●	enclosure, connection	no	**infinite**			Good	Good
– – –	line pattern	no	few				Good
→	line endings	no	few				Good
☰	line weight	**yes**	few		Good		

Noah Iliinsky • ComplexDiagrams.com/properties • 2012-06

Figure 5.3 Table by Noah Iliinsky; http://complexdiagrams.com/properties.



Ann K. Emery
data analysis + visualization
TEACH ME ABOUT... | EVENTS | SERVICES | BLOG | ABOUT | CONTACT

EMERY'S ESSENTIALS Chart Choosing Tool

Use the filters to narrow down your search and click on the pictures for descriptions and examples.

ALL / PATTERNS OVER TIME / SMALL MULTIPLES / COMPARING 2 OR MORE CATEGORIES / CIRCLE CHART FAMILY / RANGES OR DISPERSION / PART TO WHOLE / DO-ABLE IN EXCEL / GEOGRAPHIC MAPS / RELATIONSHIPS / COLLAGES / QUALITATIVE / EXPLORATORY /
Line
Area
Stacked Area
Slope
Panel
Small Multiples Lines with Shading

Figure 5.4 Ann K. Emery's Essentials website: http://annkemery.com/essentials/.

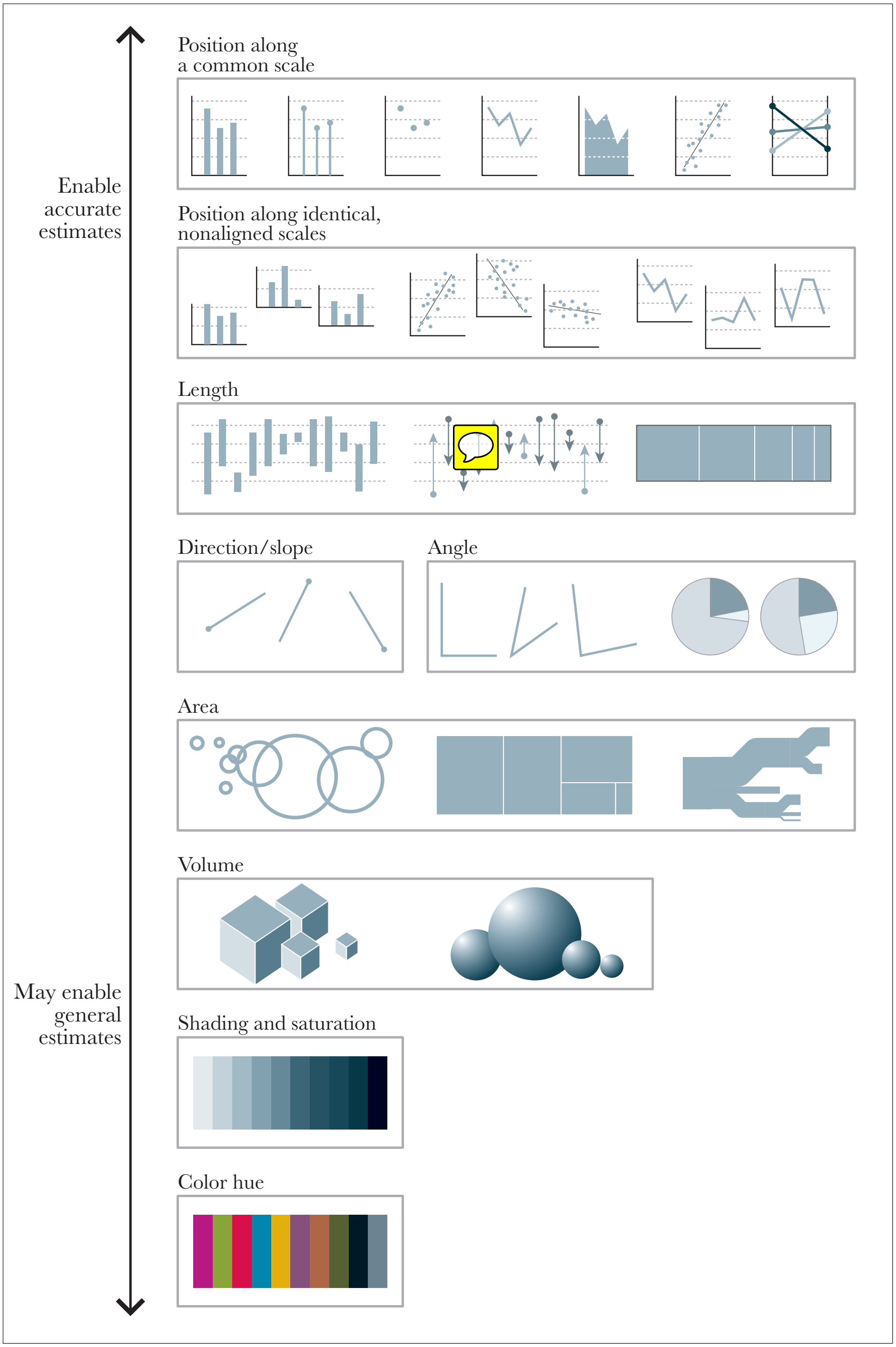


Figure 5.5 Scale of elementary perceptual tasks, inspired by William Cleveland and Robert McGill.

This is how Cleveland and McGill described their hierarchy: “We have chosen the term *elementary perceptual task* because a viewer performs one or more of these mental-visual tasks to extract the values of real variables represented on most graphs.”[2]

In other words, to decode a pie chart, we will try to use the angle or the area of the slices as clues. When seeing a bar chart, we may pay attention to the position of the upper edge of each bar, or to its length or height. When trying to decode a bubble chart, we could try to compare areas (the right choice) or diameters (which would mislead us.)

Cleveland and McGill tested the effectiveness of their perceptual tasks in several experiments. The conclusion was that **if you wish to create a successful statistical chart, you need to construct it based on elementary tasks “as high in the hierarchy as possible.”** The closer you move to the top of the scale, the faster and more accurate the estimates readers can make with your graphic. You can test that yourself going back to Figure 5.2. Area, color, and angle are much less effective than those graphic forms based on positioning objects on common scales.

A Grain of Salt

Two important caveats are in order at this point. First, **Cleveland and McGill were writing just about statistical charts**. What about data maps? After all, maps use many methods of encoding that belong to the bottom half of the hierarchy, such as area, hue, shading, and so on. Is this wrong? Hardly. **Methods of encoding on the bottom half of their scale may be appropriate when the goal isn’t to enable accurate judgments, but to reveal general patterns.**

Figure 5.6 is a **choropleth map** of unemployment rates by U.S. county. Its goal isn’t to let you identify the counties with the highest or lowest rates, or to rank counties in a precise manner. The map’s purpose is to reveal geographic clusters, such as the very low rates in the North-South strip from North Dakota to Texas, or the very high rates in many counties in Southern states.

[2] Cleveland and McGill’s original 1984 paper can be read here: https://www.cs.ubc.ca/~tmm/courses/cpsc533c-04-spr/readings/cleveland.pdf.

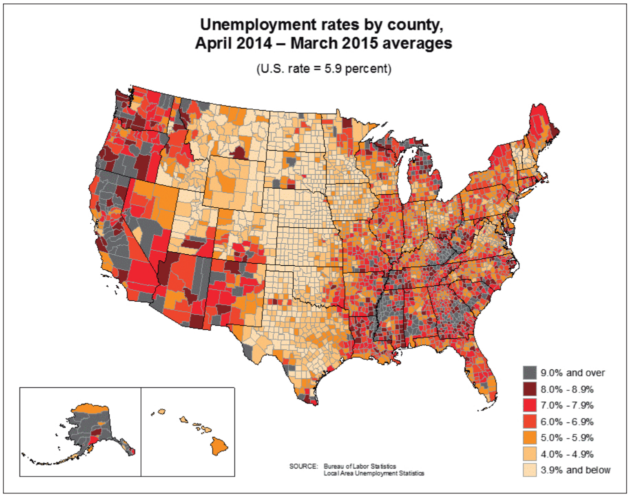

Unemployment rates by county,
April 2014 – March 2015 averages
(U.S. rate = 5.9 percent)
9.0% and over
8.0% - 8.9%
7.0% - 7.9%
6.0% - 6.9%
5.0% - 5.9%
4.0% - 4.9%
3.9% and below
SOURCE: Bureau of Labor Statistics
Local Area Unemployment Statistics

Figure 5.6 From the U.S. Bureau of Labor Statistics, http://www.bls.gov/lau/maps/twmcort.gif.

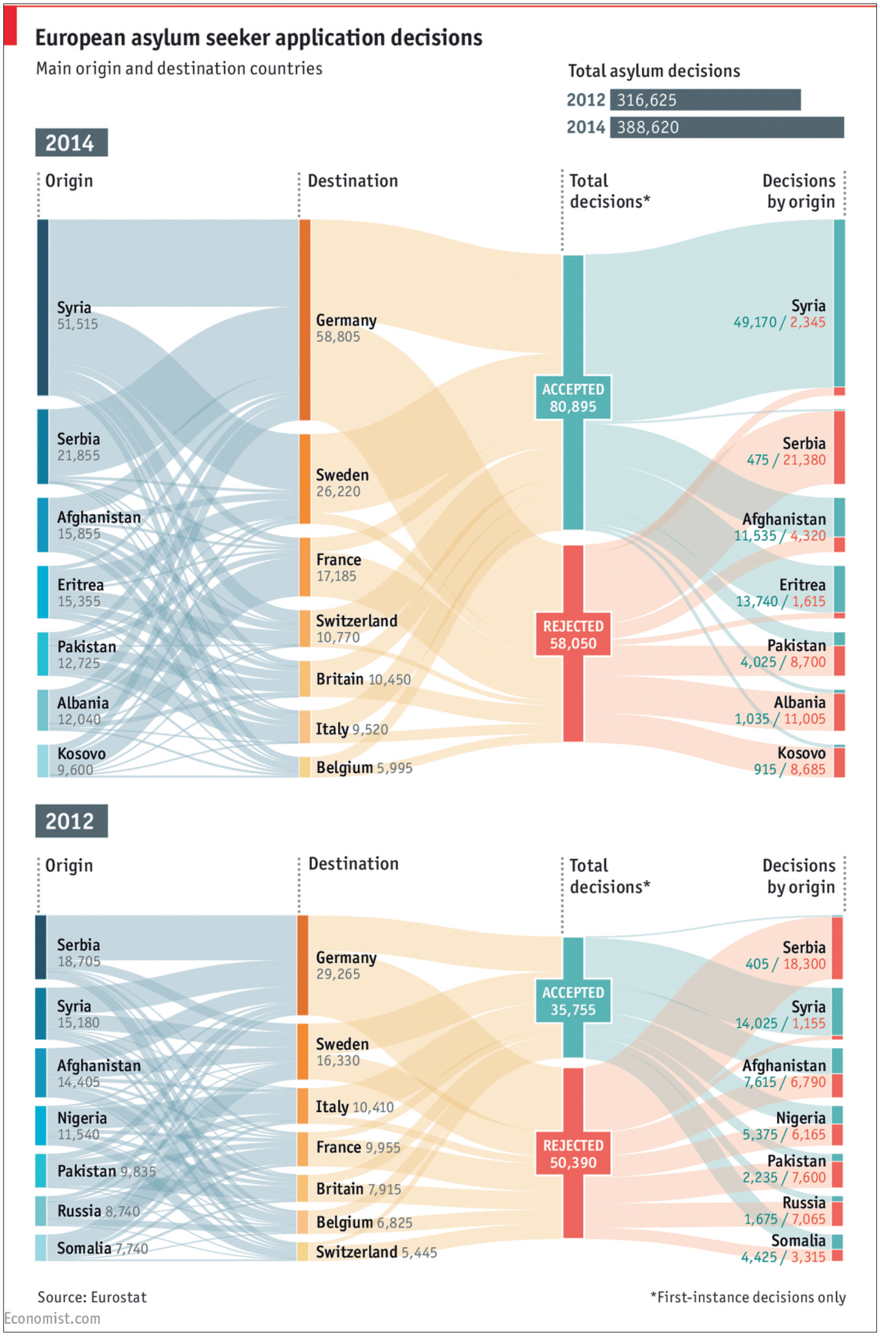
If the goal of this same graphic were to let readers compare counties, then the map wouldn't be the right choice. We'd need to pick a graphic form from the top of Cleveland's and McGill's scale, perhaps a bar chart or a lollipop chart, and then rank and group the counties in a meaningful way—from highest to lowest, alphabetically, per state, and so on.

And **what if our purpose is to show readers *both* the big picture and the details?** Then we'd need *both* the map *and* the chart on the same page or, if this were an interactive visualization, a menu that'd let people switch between them. **Different graphic forms enable different tasks much better than others**.

The second caveat is that you cannot apply anyone's method of choosing the most appropriate graphic form uncritically. Common sense is paramount.

For instance, think of how hard it would be to use a method of encoding from the very top of Cleveland's and McGill's hierarchy to show the same data that **Figure 5.7** displays. Here, readers need to decode length and area, but that's not a huge problem, considering the purpose of the chart.

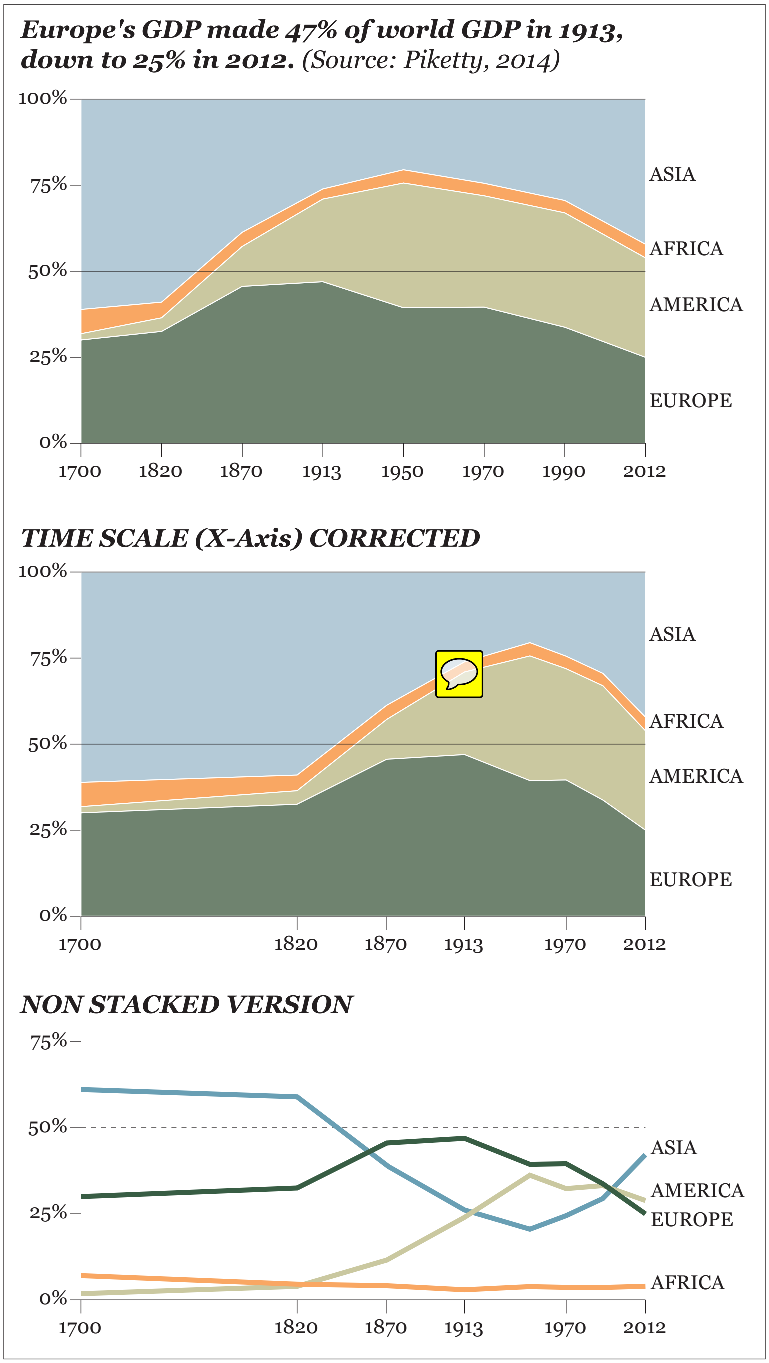
On **Figure 5.8**, I'm comparing several versions of a chart inspired by **Thomas Piketty's** 2014 bestseller *Capital in the Twenty-First Century*. The first one, on top, is similar to one that appears in Piketty's book. The second is my correction, spacing the years on the X-axis correctly. Notice how different the patterns look after doing this.



Source: Eurostat

*First-instance decisions only

Figure 5.7 A Sankey diagram by *The Economist*, http://www.economist.com/blogs/graphicdetail/2015/05/daily-chart-1?fsrc=scn/tw_ec/seeking_safety.



Europe's GDP made 47% of world GDP in 1913, down to 25% in 2012. *(Source: Piketty, 2014)*

TIME SCALE (X-Axis) CORRECTED

NON STACKED VERSION

Figure 5.8 Three charts based on the same data.

Reading Piketty's **stacked area chart** forces you to perform perceptual tasks that belong to the middle of Cleveland's and McGill's scale. You need to either compare areas or distances between the top and bottom edges of each segment. The only changes that can be visualized accurately are Asia's and Europe's, as those two portions are sitting on a horizontal edge, either the 0-baseline or the 100-line on top.

Africa's and America's baselines shift depending on how tall Asia's and Europe's segments are, and that makes detecting changes in those continents difficult. It may well happen that to your eyes it seems that Africa's output grew in the 1950s just because Africa's segment is being pushed up by the increasing size of American economies. But Africa's GDP barely changed in that decade.

This is all fine, though, because the purpose of this chart is explicit in its title: comparing Europe to the rest of the continents, besides making clear that figures add up to 100 percent. That's why on the original chart Europe's segment is emphasized and placed at the bottom, sitting on the 0-baseline. The other continents are shown to provide context.

But what if the goal of the chart was to put all continents on the same footing and compare them in an accurate manner? In that case, the stacked area chart doesn't work well. Can you see, for instance, if America's contribution to the world GDP was larger or smaller than that of Europe in 2012? You can't, unless you use your fingers to measure that last part of the chart. But see how easy this task is if we design a simple, non-stacked **time-series chart**, like the third one at the bottom.

Finally, what if you want to show *both* parts of a whole and all lines as individual entities, sitting on a common 0-baseline? Then, you'd need to design both charts, as **National Public Radio (NPR)** did with its interactive visualization about college majors (**Figure 5.9**.)[3] The designer, **Quoctrung Bui**, decided to first show readers the big picture—all majors together—stacked on top of each other. Then, if a reader decides that she needs more detail about a particular major, she can click on it and see its change on a regular time-series chart.

[3] The organization of majors in this chart is a bit confusing. As the segments are color-coded, I assumed that they were grouped somehow. It turns out that they are organized alphabetically and that colors are assigned somewhat arbitrarily.

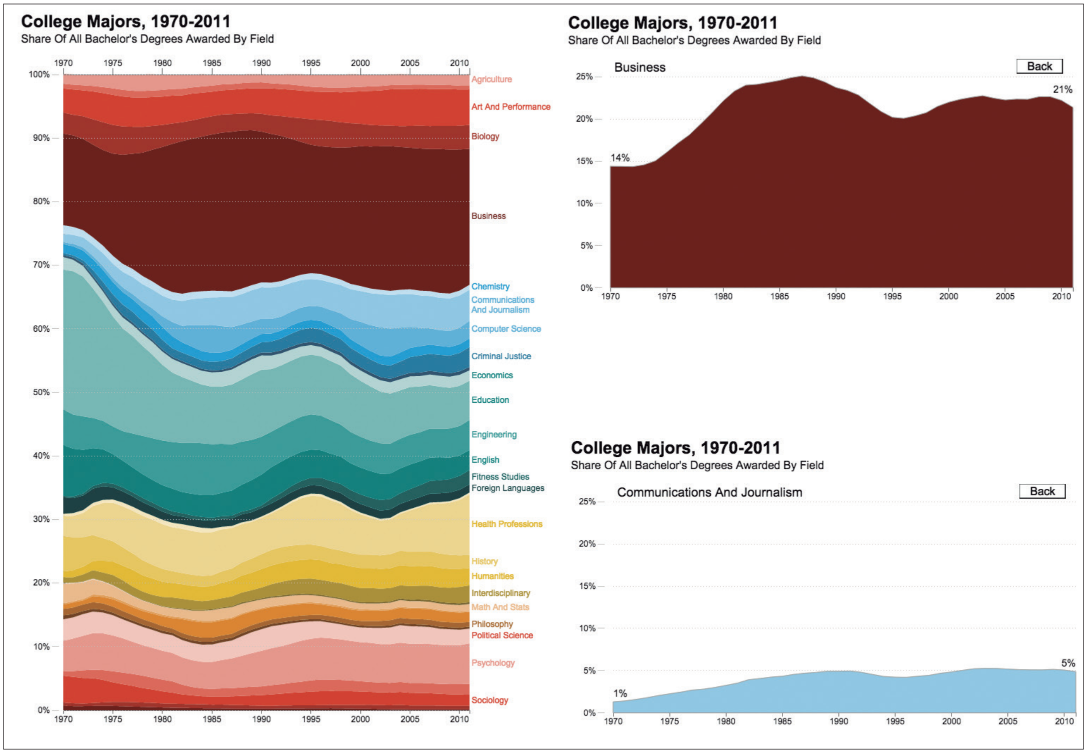


Figure 5.9 Visualization by NPR, http://www.npr.org/sections/money/2014/05/09/310114739/whats-your-major-four-decades-of-college-degrees-in-1-graph.

The examples we've seen in this section will help you understand another important rule of thumb: if you need to show parts of a whole, show them, by all means. But if the purpose of your chart is to show *each* one of those parts individually, do that. Let's rephrase that as a more general rule: **always plot your your data directly**.

In the first chart on **Figure 5.10**, I have chosen the right graphic form from Cleveland's and McGill's scale. All data is plotted on a common axis, so making accurate estimates is quite easy and fast. However, does it really matter to me to plot income and expenses as separate variables? Or does it matter more to see the difference between them? Depending on the answer, you'd need to choose either the first or the second chart. **If the difference matters more, plot the difference**, not income and expenses separately.

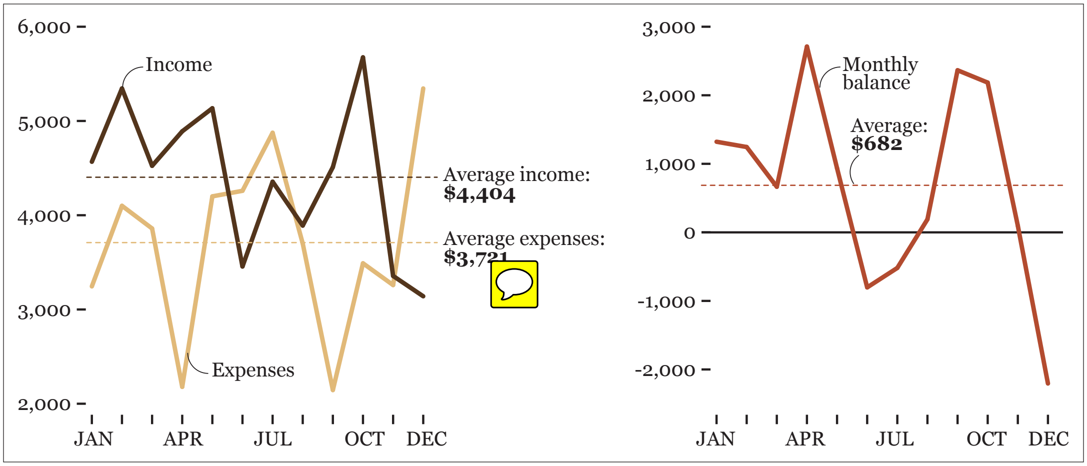


Figure 5.10 Which chart is better? It all depends on if you want to emphasize income versus expenses, or if you wish to display the monthly balance.

Practical Tips About those Tricky Scales

Another factor to consider when deciding how to design a chart is its baseline and the scale on the X-axis (horizontal) and the Y-axis (vertical).

Look at the first two charts in **Figure 5.11** without reading the numbers on the Y-axis. Did you notice how large the differences are? Well, they really aren't! I truncated their Y-axis, so the baseline in both cases is set to 50 percent, rather than to 0 percent. It isn't advisable to do so when using graphic forms in which the length measured from a common baseline is one of the main visual cues

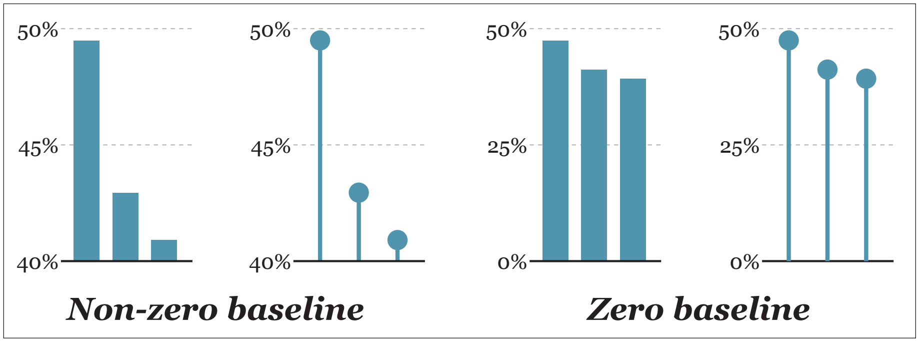


Figure 5.11 Don't truncate the Y-axis in bar charts and lollipop charts.

to interpret the information correctly. Bar charts, lollipop charts, histograms, area charts, and their variants should have a 0-baseline—unless you want to increase the chances of misunderstanding (some people do, unfortunately!).

I should point out an exception: some data sets don't have a natural 0-baseline, though. For instance, in economic analysis and finance, it's common to use index numbers, rather than just raw figures. Indexes often—not always, as we'll see in Chapter 8—have a base value of 100, as in **Figure 5.12**, which compares the cost of a product or service in different countries using the cost in the U.S. as the 100-baseline.

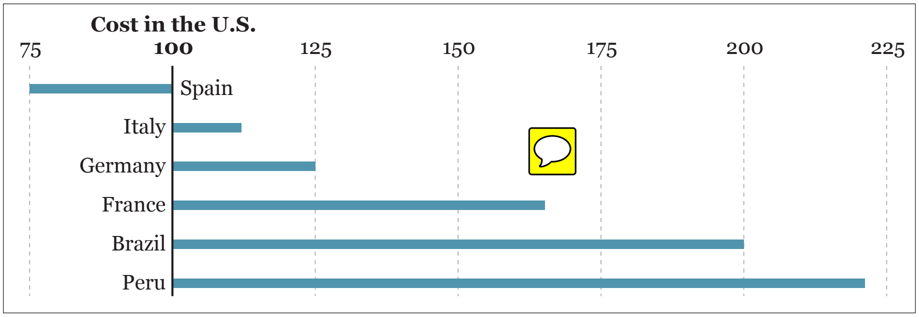


Figure 5.12 The cost of a product in different countries as a ratio of the cost of the same product in the U.S.

You can think of the ratios displayed in this kind of chart in terms of percentages. For instance, 125 means that the cost in Germany is 25 percent higher than in the U.S. The Spanish 75 means that the cost in Spain is 25 percent smaller than in the U.S. The cost in Brazil is double that of the U.S., as 200 corresponds to a 100 percent increase.

We can derive a simple and flexible rule from this discussion: rather than trying to invariably include a 0-baseline in all your charts, **use logical and *meaningful* baselines** instead. This rule should help us decide what to do when designing charts in which length isn't the method of encoding. I am thinking of dot plots, scatter plots, line charts, and so on, which rather rely on position over common axes. For example, if you're talking about the historical unemployment rate in a country, and this variable has never dropped below 5 percent, then 5 percent could be the baseline for your line chart.

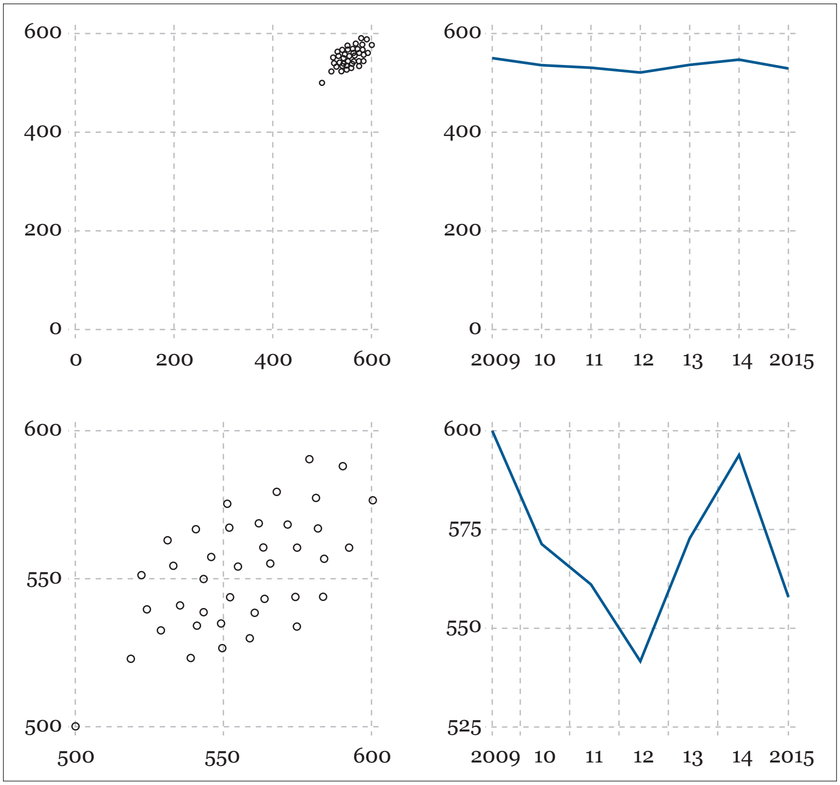


Figure 5.13: Are the 0-baselines really necessary here? Of course they aren't.

Compare the two sets of charts in **Figure 5.13**. It's a bit absurd to waste so much space just to show where the 0 point is, as I did on the two on top.

Another challenging situation appears when comparing widely different variables. See the first row of charts in **Figure 5.14**. The fact that a few data points are so large makes the smaller ones almost impossible to tell apart.

What to do? First, think of the purpose of these charts: is it just to highlight the largest values over the bulk of little ones? If that's what you need, leave the charts as they are. But what if you want readers to be able to clearly see both the large *and* the small values? You'll need at least two charts, each with its own scale, as shown in the second row of the same figure. **If your data varies so much that presenting it all on a single chart renders it useless, plot your data in several charts with dissimilar scales**.

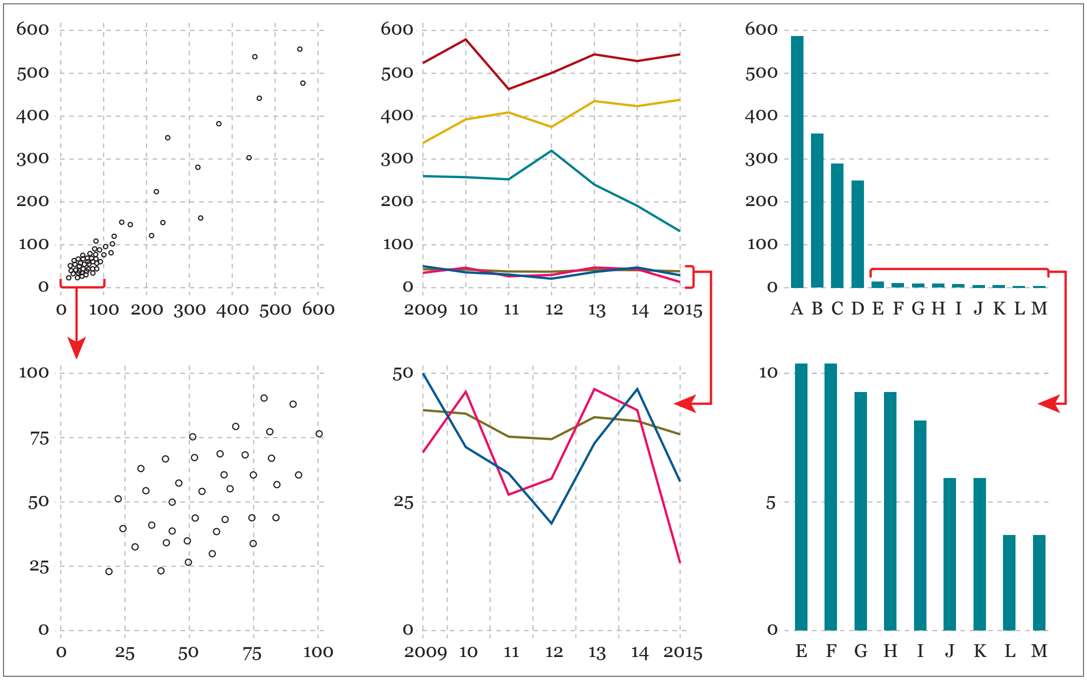


Figure 5.14 Two different scales for subsets of the same data.

Organizing the Display

Choosing the right graphic form isn't enough to design a great visualization. You also need to think of how your variables and categories are going to be organized: from highest to lowest, alphabetically, or by any other criteria. This decision also depends on the critical question I posed before: **what do I need to reveal with this graphic?**

Imagine that you're doing some advertisement market analysis and you wish to know which kind of media influences teenagers and adults the most. You may conduct a survey and display the results as in **Figure 5.15**. This chart lets you compare the different methods of delivering ads *within each age group*.

But what if what you really wish to do is not compare media within age group, but *across age groups*? In other words, what if you want to see which media becomes more or less trustworthy as people age?

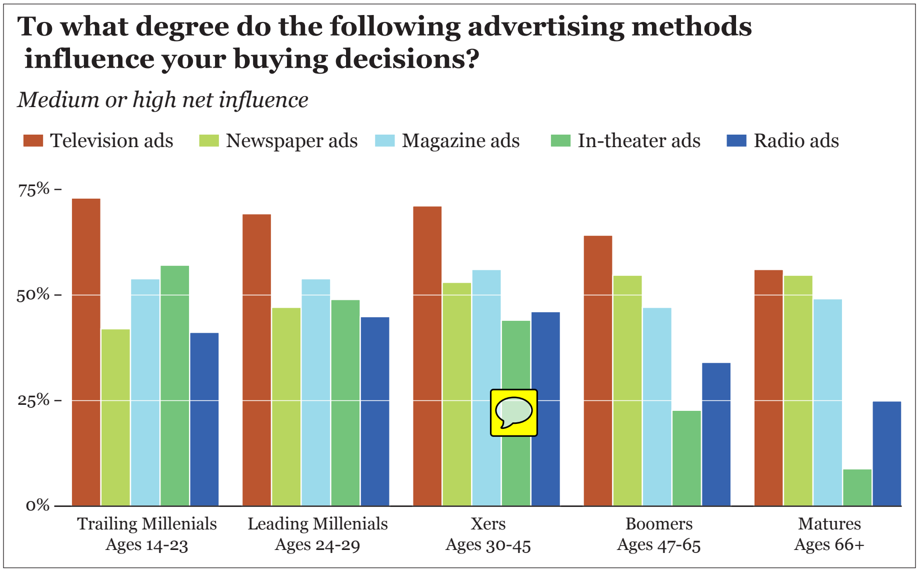


Figure 5.15 Data source, Deloitte's Digital Democracy Survey.

In that case, the current chart isn't that adequate. You can clearly spot TV's downward pattern, but that's just because the bar corresponding to TV ads is the first one of each cluster, and its color stands out over the others. If you want to see if magazine ads become more or less trusted later on in life, your brain will be forced to isolate the blue bars in the middle of each group, and then compare them to each other. That's way too much work. If seeing trends across age groups is the task we want to enable, let's group the bars not by age, but by media (**Figure 5.16**).

We could further improve this chart. I love bar charts, but they tend to look a bit clunky when you have more than 10 bars or so. A good alternative would be a **parallel coordinate plot** (**Figure 5.17**), a kind of line chart which doesn't put time on the X-axis, but some sort of categorical variable. We'll learn more about it in Chapter 9. The beauty of the parallel coordinate chart is that it gives us the best of both worlds: it doesn't just let us see trends across age groups, but it also lets us compare each media within each group, as the dots are stacked on top of each other.

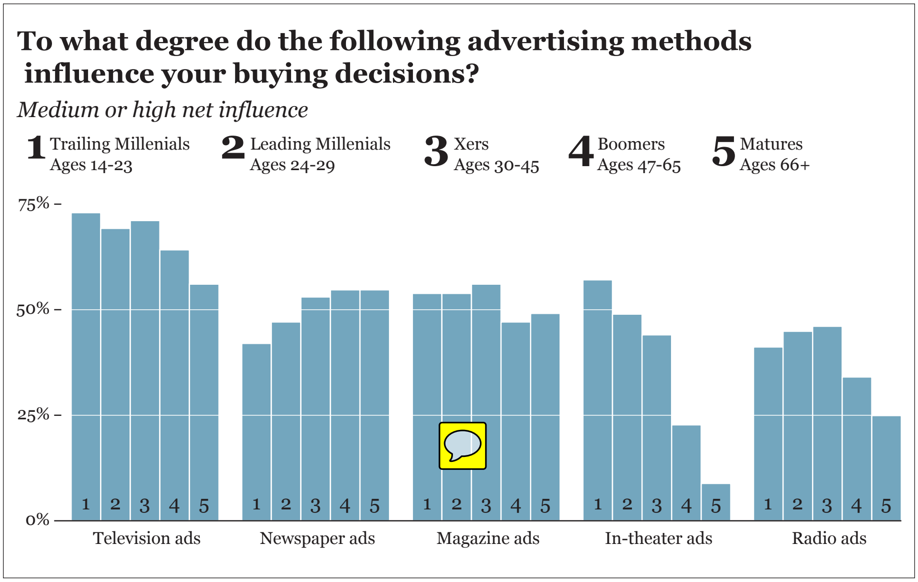


Figure 5.16 Reorganizing the data from Figure 5.15.

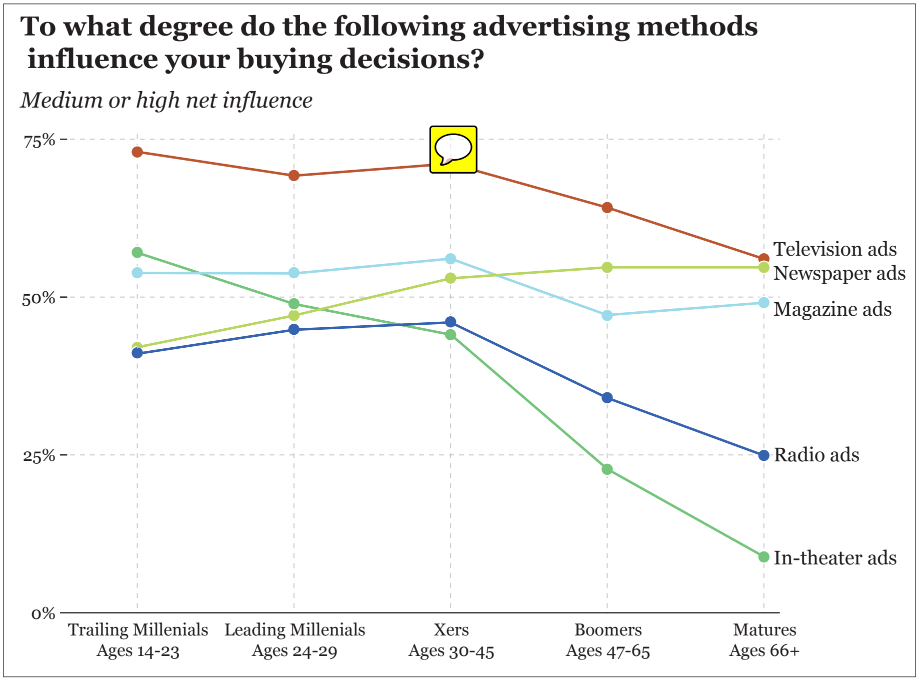
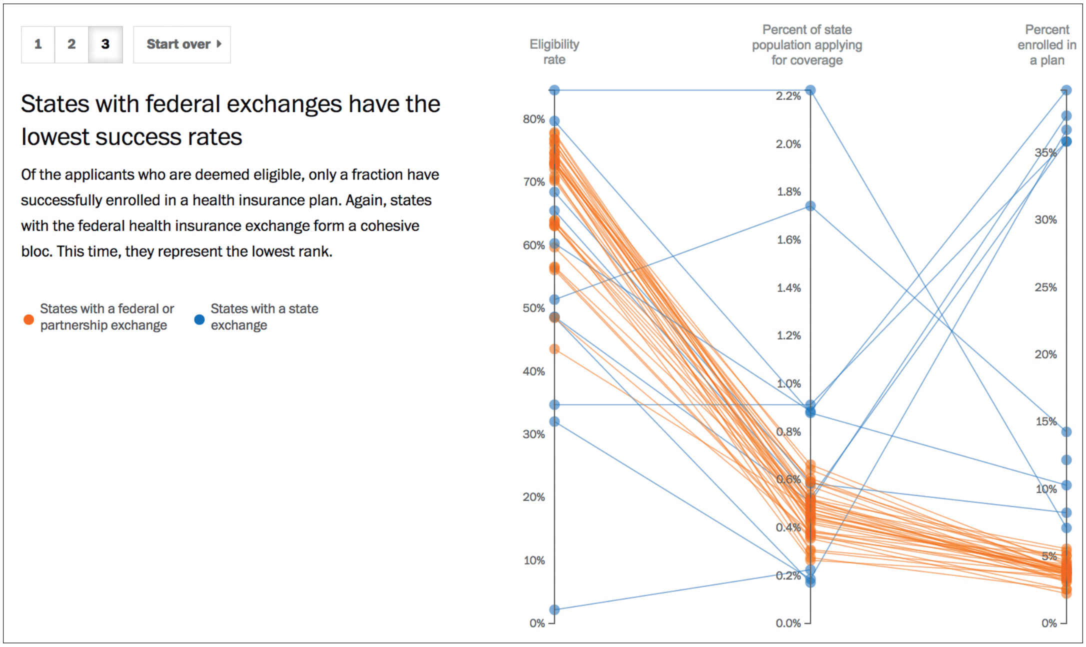


Figure 5.17 Parallel coordinate chart with the same data used in Figure 5.15.



States with federal exchanges have the lowest success rates

Of the applicants who are deemed eligible, only a fraction have successfully enrolled in a health insurance plan. Again, states with the federal health insurance exchange form a cohesive bloc. This time, they represent the lowest rank.

● States with a federal or partnership exchange
● States with a state exchange

Figure 5.18 Visualization by *The Washington Post*, http://www.washingtonpost.com/wp-srv/special/politics/state-vs-federal-exchanges/.

"Wait," you're probably thinking, "aren't line charts intended to display just trends over time intervals?" Many of us learned that rule in primary school. But that's just a convention, and conventions can and should change. Line charts can certainly be used to display time-series data, but time-series charts aren't the only kind of line charts that exist in the visualization designer's repertoire. Parallel coordinate charts are pretty useful to visualize multi-dimensional relationships. See **Figure 5.18** for an example.[4]

Put Your Work to the Test

There are certain graphic forms that I systematically avoid. One is the **radar chart**, as I consider it a feeble way of presenting information. Designers sometimes defend radar charts because they look intriguing and pretty. I am not always against sacrificing a bit of clarity if the payoff in the form of allure is great, but I think that in the case presented in **Figure 5.19** we're sacrificing too much.

[4] Visualization expert Robert Kosara has a good article about parallel coordinate at charts: https://eagereyes.org/techniques/parallel-coordinates.

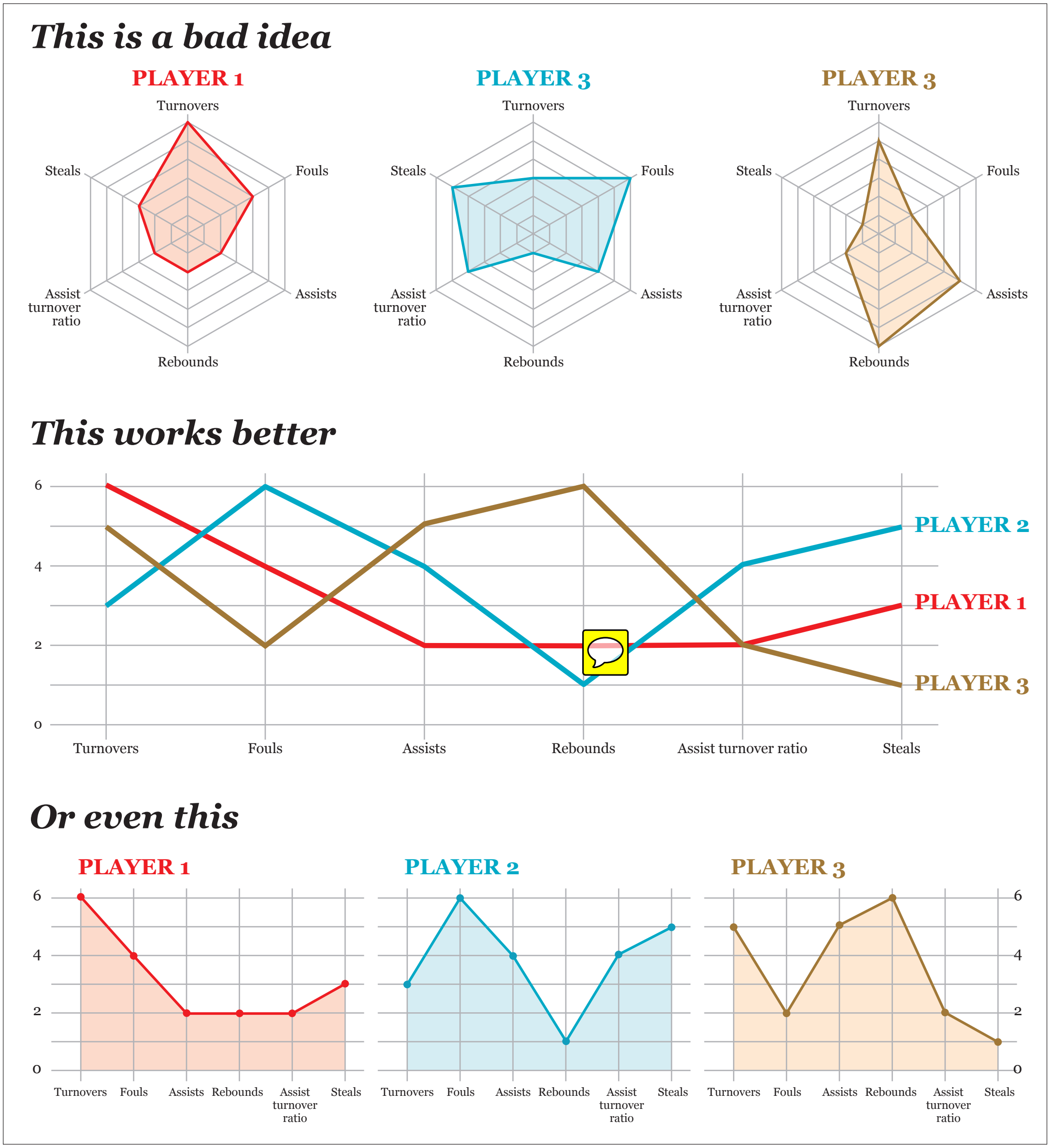


Figure 5.19 Alternatives to radar charts.

On the three radar charts on top, I am presenting metrics from three basketball players. Right underneath the radar charts there are two alternative designs (both parallel coordinates), which make comparing the athletes to each other much easier.

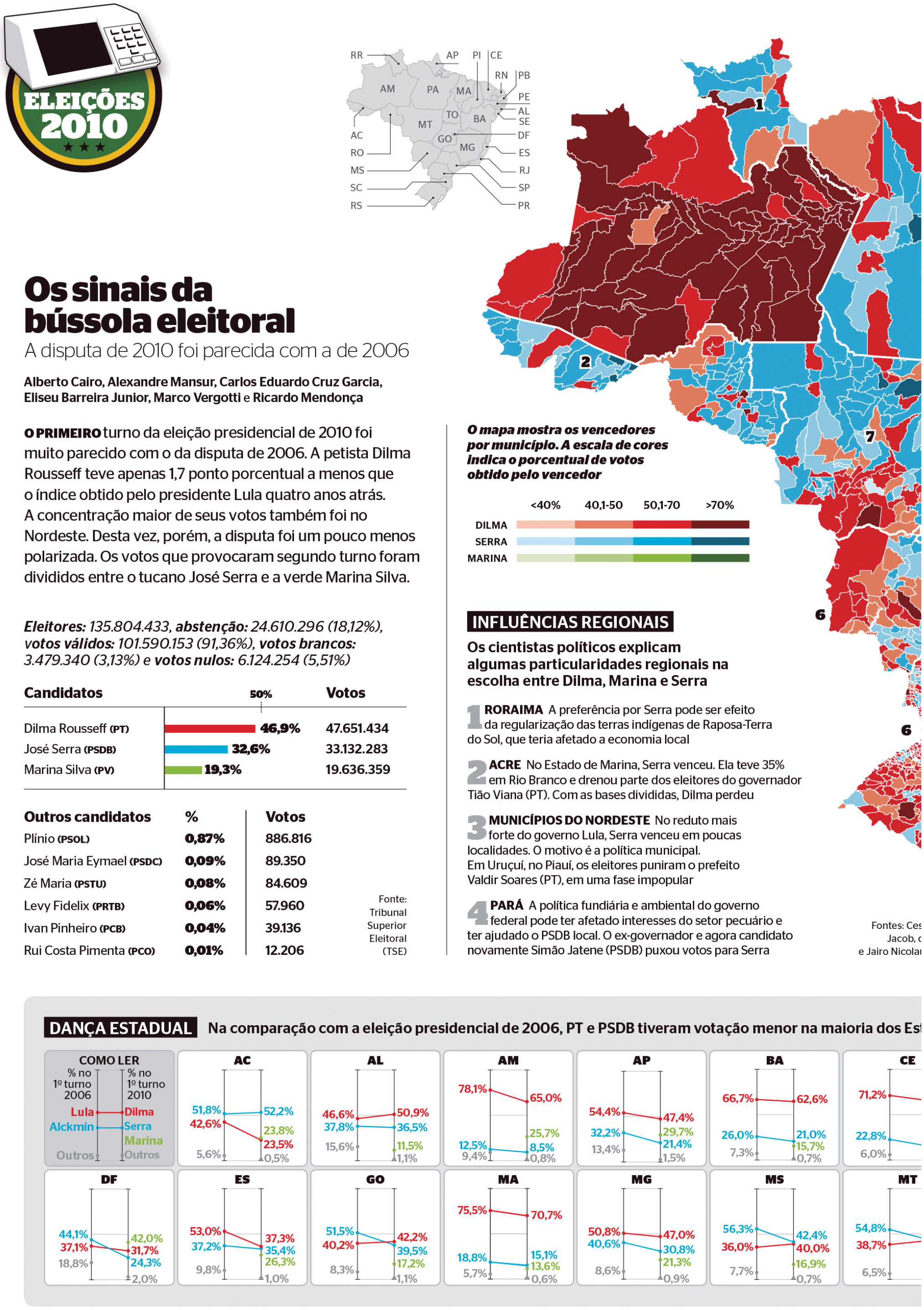
Which one of the two alternative designs I'd choose depends on the number of players shown and on how close they are to each other, as parallel coordinate charts can get easily crowded, and lines may start occluding each other. In that case, it may be better to create a small chart for each athlete. The third option in Figure 5.19 is called a **small multiple array** or trellis chart. A small multiple display is a layout in which small charts or maps with consistent scales are presented side by side or on top of each other.

That said, I have used radar charts myself a couple of times in my career as an infographics and data visualization designer. Why did I break my own rule? Because sometimes a graphic form that is an inept choice in most circumstances may be fruitful in a very specific one.

Figure 5.20 is a poster-size infographic made by my team at the Brazilian weekly news magazine *Época*, where I was graphics director between 2010 and 2012. It shows the results of the 2010 presidential election with a combination of bar charts, **slope charts**—the ones at the bottom, comparing the 2010 results with the results of the previous election, state by state—and a choropleth map.

There's also a large radar chart on the upper-right corner. In this one, each radii corresponds to one of the 27 states Brazil is divided into. There are three color lines, one for each of the candidates. Red is for Dilma Rousseff (who ended up becoming president); blue is for José Serra; and green is for Marina Silva. The center of the radar is the 0 percent point, and the outmost ring corresponds to 100 percent of the vote. The farther away a joint of one of the lines is from the center, the larger the share of the vote that particular candidate got in that state.

Let me admit at the outset that these state-by-state results could also have been displayed using a set or traditional bar charts, but we decided on the radar chart because we wanted to highlight the fact that Dilma Rousseff, the Left candidate, won by a very high margin in northeastern states. Notice that the radii on the radar chart are organized according to their geographic position: northeast on the upper-right corner, southeast on the bottom-right, and so forth. Someone familiar with Brazil's geography will be able to relate the choropleth map to the radar chart when they are displayed side by side, like in this case.



Os sinais da bússola eleitoral

A disputa de 2010 foi parecida com a de 2006

Alberto Cairo, Alexandre Mansur, Carlos Eduardo Cruz Garcia, Eliseu Barreira Junior, Marco Vergotti e **Ricardo Mendonça**

O PRIMEIRO turno da eleição presidencial de 2010 foi muito parecido com o da disputa de 2006. A petista Dilma Rousseff teve apenas 1,7 ponto porcentual a menos que o índice obtido pelo presidente Lula quatro anos atrás. A concentração maior de seus votos também foi no Nordeste. Desta vez, porém, a disputa foi um pouco menos polarizada. Os votos que provocaram segundo turno foram divididos entre o tucano José Serra e a verde Marina Silva.

Eleitores: *135.804.433,* ***abstenção:*** *24.610.296 (18,12%),* ***votos válidos:*** *101.590.153 (91,36%),* ***votos brancos:*** *3.479.340 (3,13%) e* ***votos nulos:*** *6.124.254 (5,51%)*

Candidatos		Votos
Dilma Rousseff (PT)	46,9%	47.651.434
José Serra (PSDB)	32,6%	33.132.283
Marina Silva (PV)	19,3%	19.636.359

Outros candidatos	%	Votos
Plínio (PSOL)	0,87%	886.816
José Maria Eymael (PSDC)	0,09%	89.350
Zé Maria (PSTU)	0,08%	84.609
Levy Fidelix (PRTB)	0,06%	57.960
Ivan Pinheiro (PCB)	0,04%	39.136
Rui Costa Pimenta (PCO)	0,01%	12.206

Fonte: Tribunal Superior Eleitoral (TSE)

O mapa mostra os vencedores por município. A escala de cores indica o porcentual de votos obtido pelo vencedor

INFLUÊNCIAS REGIONAIS

Os cientistas políticos explicam algumas particularidades regionais na escolha entre Dilma, Marina e Serra

1 RORAIMA A preferência por Serra pode ser efeito da regularização das terras indígenas de Raposa-Terra do Sol, que teria afetado a economia local

2 ACRE No Estado de Marina, Serra venceu. Ela teve 35% em Rio Branco e drenou parte dos eleitores do governador Tião Viana (PT). Com as bases divididas, Dilma perdeu

3 MUNICÍPIOS DO NORDESTE No reduto mais forte do governo Lula, Serra venceu em poucas localidades. O motivo é a política municipal. Em Uruçuí, no Piauí, os eleitores puniram o prefeito Valdir Soares (PT), em uma fase impopular

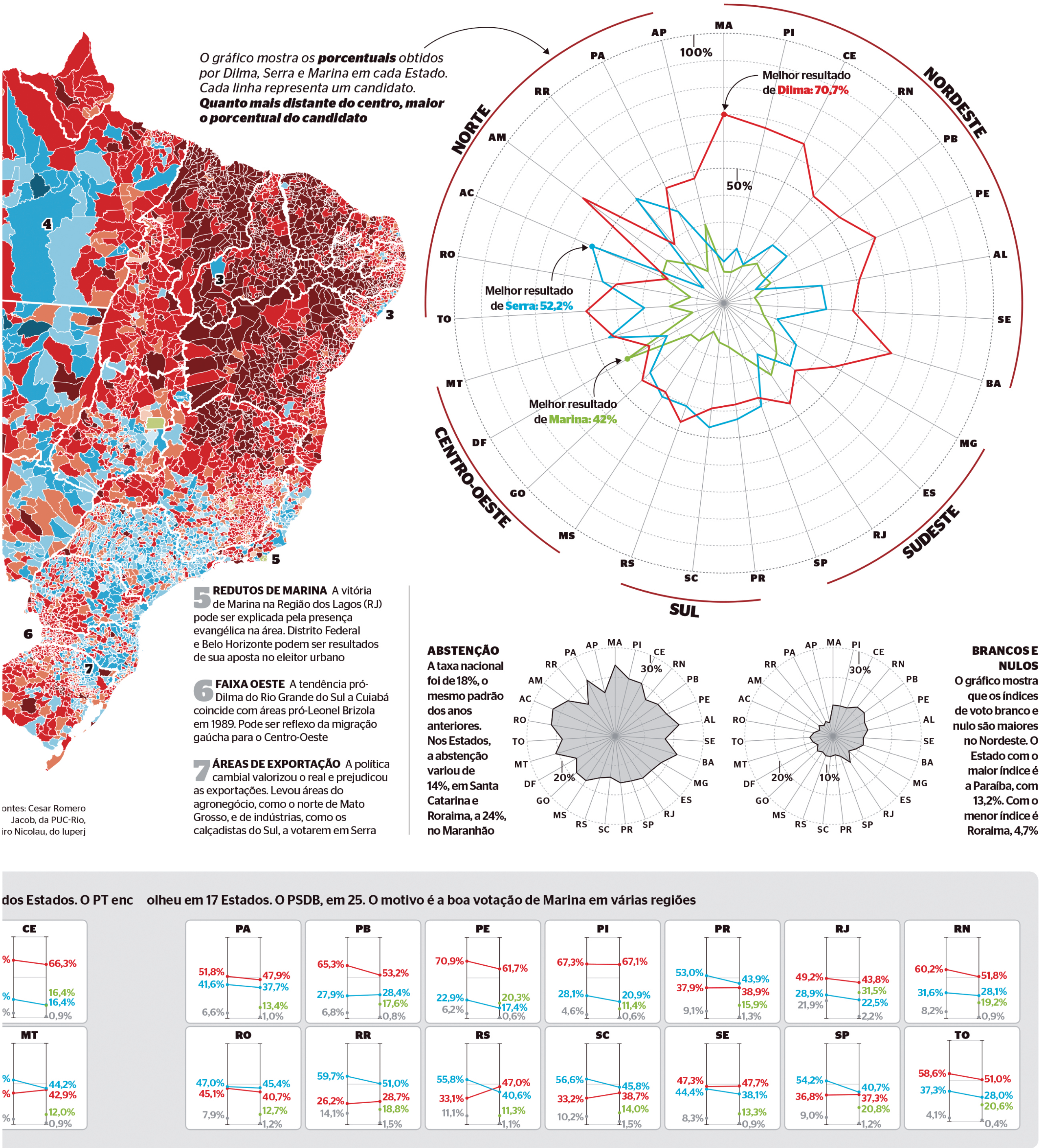
4 PARÁ A política fundiária e ambiental do governo federal pode ter afetado interesses do setor pecuário e ter ajudado o PSDB local. O ex-governador e agora candidato novamente Simão Jatene (PSDB) puxou votos para Serra

Fontes: Ces
Jacob, c
e Jairo Nicolau

DANÇA ESTADUAL

Na comparação com a eleição presidencial de 2006, PT e PSDB tiveram votação menor na maioria dos Est

Figure 5.20 Infographic published by *Época* magazine (Brazil).



O gráfico mostra os ***porcentuais*** *obtidos por Dilma, Serra e Marina em cada Estado. Cada linha representa um candidato.* ***Quanto mais distante do centro, maior o porcentual do candidato***

5 REDUTOS DE MARINA A vitória de Marina na Região dos Lagos (RJ) pode ser explicada pela presença evangélica na área. Distrito Federal e Belo Horizonte podem ser resultados de sua aposta no eleitor urbano

6 FAIXA OESTE A tendência pró-Dilma do Rio Grande do Sul a Cuiabá coincide com áreas pró-Leonel Brizola em 1989. Pode ser reflexo da migração gaúcha para o Centro-Oeste

7 ÁREAS DE EXPORTAÇÃO A política cambial valorizou o real e prejudicou as exportações. Levou áreas do agronegócio, como o norte de Mato Grosso, e de indústrias, como os calçadistas do Sul, a votarem em Serra

ontes: Cesar Romero
Jacob, da PUC-Rio,
iro Nicolau, do Iuperj

ABSTENÇÃO
A taxa nacional foi de 18%, o mesmo padrão dos anos anteriores. Nos Estados, a abstenção variou de 14%, em Santa Catarina e Roraima, a 24%, no Maranhão

BRANCOS E NULOS
O gráfico mostra que os índices de voto branco e nulo são maiores no Nordeste. O Estado com o maior índice é a Paraíba, com 13,2%. Com o menor índice é Roraima, 4,7%

dos Estados. O PT enc olheu em 17 Estados. O PSDB, em 25. O motivo é a boa votação de Marina em várias regiões

It's hard to know if a graphic form will work well until you try it and you compare it to alternatives, so when designing this infographic, I also designed bar charts and a parallel coordinate chart (**Figure 5.21**). We discarded it at the end, in favor of the radar chart, because we tested the latter with some journalists and designers in the newsroom. I also showed it to friends and relatives. All of them got the message the radar chart was intended to convey in a few seconds: Dilma Rousseff's line looks like a rubber band that has been stretched out toward the northeast.

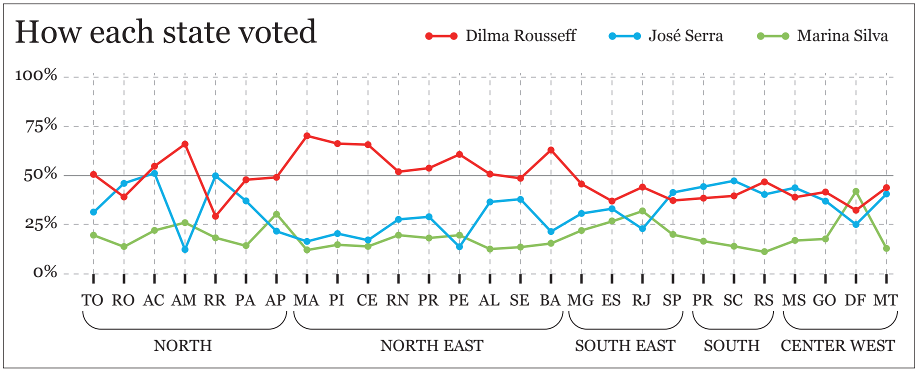


Figure 5.21 Infographic published by *Época* magazine (Brazil).

When the graphic was almost finished, *Época's* managing editor at the time, **Helio Gurovitz**, joked that the radar chart should actually be called a "compass chart," and suggested a title: The Signs of the Electoral Compass ("Os sinais da bússola eleitoral.") That made a lot of sense to me.

What I get from stories like this one is that **rules of visualization matter as much as the results of the tests you may conduct with readers**, even those tests that are as informal as the one I've just described.

Tools like Cleveland's and McGill's hierarchy of methods of encoding are essential for our work, as they are grounded on empirical evidence obtained through experiments. They save time and energy that we can devote to better purposes, like plotting our data several times, giving this or that graphic form a try, putting the results side by side, showing them to as many people as possible, and then asking them to describe what insights they get after exploring the graphic for a bit.

Some testing is critical, as **very often readers don't interpret our visualizations as we want them to**. In *Misbehaving: The Making of Behavioral Economics*, the book I mentioned at the beginning of this chapter, economist Richard H. Thaler describes an experiment he conducted in 1995. He asked employees at the University of Southern California to choose between two imaginary 401(k) retirement plans, a riskier one with higher expected returns (Fund A) and a safer one with lower ones (Fund B).

Thaler showed one group of employees the first two charts in **Figure 5.22**. These represent the distribution of one-year returns. Each bar represents one of 35 possible changes (increase or decrease) from one year to the next.

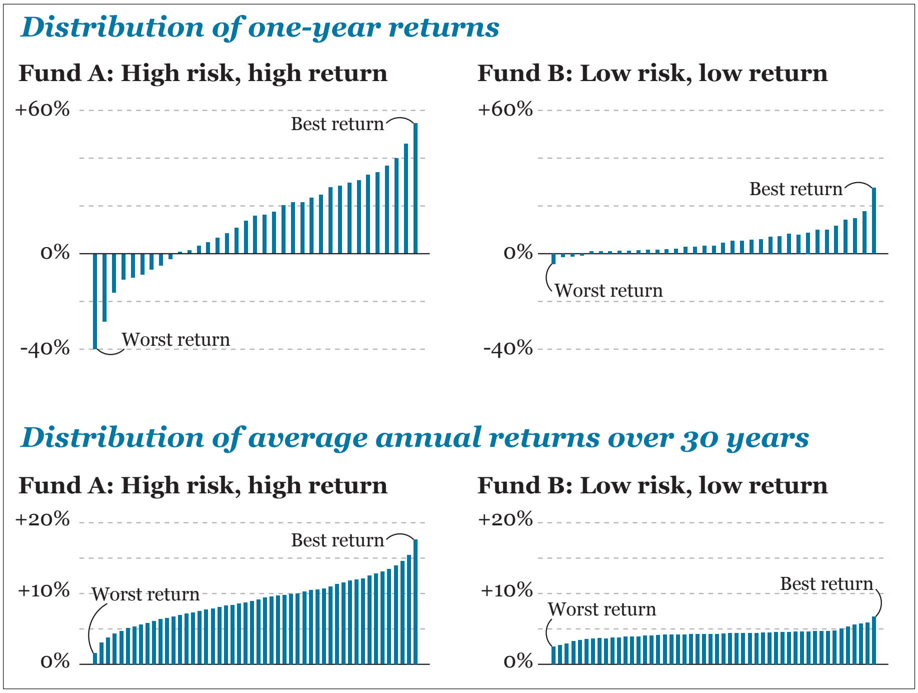


Figure 5.22 Charts based on Richard H. Thaler's *Misbehaving: The Making of Behavioral Economics* (2015).

The worst possible annual return of Fund A, the riskier one, is a –40 percent, and the best one is an increase of nearly 55 percent over the previous year. (Remember that these bars aren't organized chronologically, but from lowest to highest return.) Fund B shows less variation: the worst annual return is loss of –4 percent, while the best return is an increase of around 28 percent in one year.

Another group of subjects were shown the second set of two charts. These are all possible *total* returns over a period of 30 years. If you invest now and only take a look at your returns three decades from now, you may get anything from the lowest to the highest of the returns shown on the charts. There aren't negative returns in this case, as you can see.

The results of the experiments were impressive. People who saw the first two charts said that they weren't willing to take many risks, so they chose to put just 40 percent of their portfolio in Fund A (high risk, high return) and 60 percent in Fund B.

Those who were shown the second two charts said that they would prefer to invest *90 percent of their money in Fund A*, the risky one. The funniest thing of this experiment is that **both sets of charts are based on exactly the same underlying data**, coming from real portfolios made of a mixture of bonds and stocks.

Take notice: The way data is visually presented has very real consequences on the lives of people who read your visualizations.

To Learn More

- Bertin, Jacques. *Semiology of Graphics*. Redlands, CA: Esri Press, 2010. Bertin, a cartographer, was the founding father of modern visualization. This book, originally published in French in 1967, is his magnum opus.

- Börner, Katy. *Atlas of Knowledge: Anyone Can Map*. Boston, MA: MIT Press, 2015. Börner, a professor of Information Science at Indiana University in Bloomington, is the author of two other books about visualization, but this is my favorite one by far. It's full of great examples, and it offers a thorough discussion of methods of encoding data.

- Cleveland, William S. *The Elements of Graphing Data*. Monterey, CA: Wadsworth Advanced and Software, 1985. An absolute classic of visualization.

- Few, Stephen. *Show Me the Numbers: Designing Tables and Graphs to Enlighten.* Oakland, CA: Analytics, 2004. My favorite book about statistical charts for business analytics.

- Meirelles, Isabel. *Design for Information: An Introduction to the Histories, Theories, and Best Practices behind Effective Information Visualizations*. Beverly: MA. Rockport, 2013. This beautiful book doesn't cover just quantitative or data visualization, but it also describes how to represent any kind of information by means of "structures:" hierarchical, relational, temporal, spatial, spatio-temporal, and textual.
- Steele, Julie, and Iliinsky, Noah. *Designing Data Visualizations.* Sebastopol, CA: O'Reilly, 2011. A concise and dense introduction to good visualization practices.