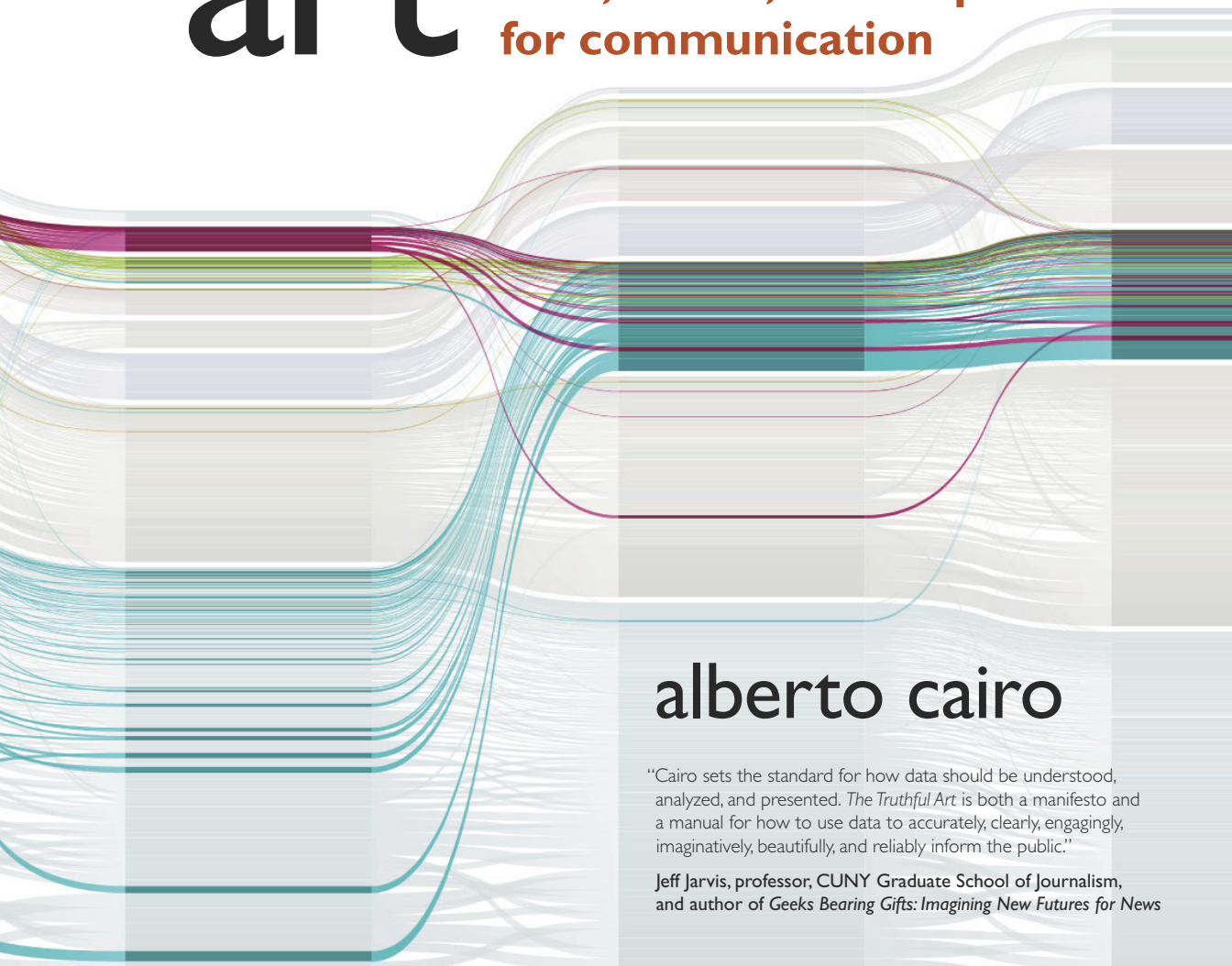
the truthful art

data, charts, and maps for communication



alberto cairo

"Cairo sets the standard for how data should be understood, analyzed, and presented. *The Truthful Art* is both a manifesto and a manual for how to use data to accurately, clearly, engagingly, imaginatively, beautifully, and reliably inform the public."

Jeff Jarvis, professor, CUNY Graduate School of Journalism, and author of *Geeks Bearing Gifts: Imagining New Futures for News*

Praise for *The Truthful Art*

"Alberto Cairo is widely acknowledged as journalism's preeminent visualization wiz. He is also journalism's preeminent data scholar. As newsrooms rush to embrace data journalism as a new tool—and toy—Cairo sets the standard for how data should be understood, analyzed, and presented. *The Truthful Art* is both a manifesto and a manual for how to use data to accurately, clearly, engagingly, imaginatively, beautifully, and reliably inform the public."

—Jeff Jarvis, professor at CUNY Graduate School of Journalism and author of *Geeks Bearing Gifts: Imagining New Futures for News*

"A feast for both the eyes and mind, Alberto Cairo's *The Truthful Art* deftly explores the science—and art—of data visualization. The book is a must-read for scientists, educators, journalists, and just about anyone who cares about how to communicate effectively in the information age."

—Michael E. Mann, Distinguished Professor, Penn State University and author of *The Hockey Stick and the Climate Wars*

"Alberto Cairo is a great educator and an engaging storyteller. In *The Truthful Art* he takes us on a rich, informed, and well-visualized journey that depicts the process by which one scrutinizes data and represents information. The book synthesizes a lot of knowledge and carefully explains how to create effective visualizations with a focus on statistical principles. *The Truthful Art* will be incredibly useful to both practitioners and students, especially within the arts and humanities, such as those involved in data journalism and information design."

—Isabel Meirelles, professor at OCAD University (Canada) and author of *Design for Information*

"As soon as I started immersing myself in *The Truthful Art,* I was horrified (and somewhat ashamed) to realize how much I didn't know about data visualization. I've spent most of my career pursuing a more illustrative way to present data, but Alberto Cairo's clarifying prose superbly explained the finer points of data viz. Since Alberto warns us that "[data is] always noisy, dirty, and uncertain," everyone in this business had better read his book to find out how to properly construct visualizations that not only tell the truth, but also allow us to interact meaningfully with them."

—Nigel Holmes, founder of Explanation Graphics

These are the requirements of any rational conjecture. **Conjectures first need to make sense** (even if they eventually end up being wrong) based on existing knowledge of how nature works. The universe of stupid conjectures is infinite, after all. Not all conjectures are born equal. Some are more plausible *a priori* than others.

My favorite example of conjecture that doesn't make sense is the famous *Sports Illustrated* cover jinx. This superstitious urban legend says that appearing on the cover of *Sports Illustrated* magazine makes many athletes perform worse than they did before.

To illustrate this, I have created **Figure 4.1**, based on three different fictional athletes. Their performance curve (measured in goals, hits, scores, whatever) goes up, and then it drops after being featured on the cover of *Sports Illustrated*.

Saying that this is a curse is a bad conjecture because we can come up with a much more simple and natural explanation: athletes are usually featured on magazine covers when they are at the peak of their careers. Keeping yourself in the upper ranks of any sport is not just hard work, it also requires tons of good luck. Therefore, after making the cover of *Sports Illustrated*, it is more probable that the performance of most athletes will worsen, not improve even more. Over time, an athlete is more likely to move closer to his or her average performance rate than away from it. Moreover, aging plays an important role in most sports.

What I've just described is **regression toward the mean**, and it's pervasive.[4] Here's how I'd explain it to my kids: imagine that you're in bed today with a cold. To cure you, I go to your room wearing a tiara of dyed goose feathers and a robe made of oak leaves, dance Brazilian samba in front of you—feel free to picture this scene in your head, dear reader—and give you a potion made of water, sugar, and an infinitesimally tiny amount of viral particles. One or two days later, you

4 When playing with any sort of data set, if you randomly draw one value and obtain one that is extremely far from the mean (that is, the average value), the next one that you draw will probably be closer to the mean than even further away from it. Regression toward the mean was first described by Sir Francis Galton in the late nineteenth century, but under a slightly different name: regression toward mediocrity. Galton observed that parents who were very tall tended to have children who were shorter than they were and that parents who were very short had children who were taller than them. Galton said that extreme traits tended to "regress" toward "mediocrity." His paper is available online, and it's a delight: http://galton.org/essays/1880-1889/galton-1886-jaigi-regression-stature.pdf.

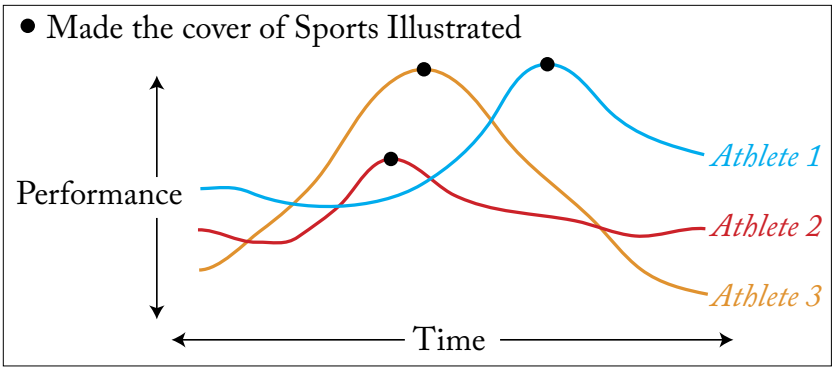


Figure 4.1 Athletes tend to underperform after they've appeared on the cover of *Sports Illustrated* magazine. Does the publication cast a curse on them?

feel better. Did I cure you? Of course not. It was your body regressing to its most probable state, one of good health.[5]

For a conjecture to be good, it also needs to be testable. In principle, you should be able to weigh your conjecture against evidence. Evidence comes in many forms: repeated observations, experimental tests, mathematical analysis, rigorous mental or logic experiments, or various combinations of any of these.[6]

Being testable also implies being *falsifiable*. A conjecture that can't possibly be refuted will never be a good conjecture, as rational thought progresses only if our current ideas can be substituted for better-grounded ones later, when new evidence comes in.

Sadly, we humans love to come up with non-testable conjectures, and we use them when arguing with others. Philosopher **Bertrand Russell** came up with a splendid illustration of how ludicrous non-testable conjectures can be:

> If I were to suggest that between the Earth and Mars there is a china teapot revolving about the sun in an elliptical orbit, nobody would be able to

5 Think about this next the time that anyone tries to sell you an overpriced "alternative medicine" product or treatment. The popularity of snake oil-like stuff is based on our propensity to see causality where there's only a sequence of unconnected events ("follow my unsubstantiated advice—feel better") and our lack of understanding of regression toward the mean.

6 If you read any of the books recommended in this chapter, be aware that many scientists and philosophers of science are more stringent than I am when evaluating if a particular procedure really qualifies as a test.

> disprove my assertion provided I were careful to add that the teapot is too small to be revealed even by our most powerful telescopes. But if I were to go on to say that, since my assertion cannot be disproved, it is intolerable presumption on the part of human reason to doubt it, I should rightly be thought to be talking nonsense. (*Illustrated* magazine, 1952)

Making sense and being testable alone don't suffice, though. **A good conjecture is made of several components, and these need to be hard to change without making the whole conjecture useless**. In the words of physicist David Deutsch, a good conjecture is "hard to vary, because all its details play a functional role." The components of our conjectures need to be logically related to the nature of the phenomenon we're studying.

Imagine that a sparsely populated region in Africa is being ravaged by an infectious disease. You observe that people become ill mostly after attending religious services on Sunday. You are a local shaman and propose that the origin of the disease is some sort of negative energy that oozes out of the spiritual aura of priests and permeates the temples where they preach.

This is a bad conjecture not just because it doesn't make sense or isn't testable. It *is* testable, actually: when people gather in temples and in the presence of priests, a lot of them get the disease. There, I got my conjecture tested and corroborated!

Not really. This conjecture is bad because we could equally say that the disease is caused by invisible pixies who fly inside the temples, the souls of the departed who still linger around them, or by any other kind of supernatural agent. Changing our premises keeps the body of our conjecture unchanged. Therefore, a flexible conjecture is always a bad conjecture.

It would be different if you said that the disease may be transmitted in crowded places because the agent that provokes it, whether a virus or a bacterium, is airborne. The closer people are to each other, the more likely it is that someone will sneeze, spreading particles that carry the disease. These particles will be breathed by other people and, after reaching their lungs, the agent will spread.

This is a good conjecture because all its components are naturally connected to each other. Take away any of them and the whole edifice of your conjecture will fall, forcing you to rebuild it from scratch in a different way. After being compared to the evidence, this conjecture may end up being completely *wrong*, but it will forever be a *good* conjecture.

Hypothesizing

A conjecture that is formalized to be tested empirically is called a **hypothesis**.

To give you an example (and be warned that not all hypotheses are formulated like this): if I were to test my hunch that using Twitter for too long reduces writers' productivity, I'd need to explain what I mean by "too long" and by "productivity" and how I'm planning to measure them. I'd also need to make some sort of prediction that I can assess, like "each increase of Twitter usage reduces the average number of words that writers are able to write in a day."

I've just defined two variables. A **variable** is something whose values can change somehow (yes-no, female-male, unemployment rate of 5.6, 6.8, or 7.1 percent, and so on). The first variable in our hypothesis is "increase of Twitter usage." We can call it a **predictor** or **explanatory** variable, although you may see it called an **independent** variable in many studies.

The second element in our hypothesis is "reduction of average number of words that writers write in a day." This is the **outcome** or **response** variable, also known as the **dependent** variable.

Deciding on what and how to **measure** is quite tricky, and it greatly depends on how the exploration of the topic is designed. When getting information from any source, sharpen your skepticism and ask yourself: do the variables defined in the study, and the way they are measured and compared, reflect the reality that the authors are analyzing?

An Aside on Variables

Variables come in many flavors. It is important to remember them because not only are they crucial for working with data, but later in the book they will also help us pick methods of representation for our visualizations.

The first way to classify variables is to pay attention to the scales by which they're measured.

Nominal

In a nominal (or categorical) scale, values don't have any quantitative weight. They are distinguished just by their identity. Sex (male or female) and location (Miami, Jacksonville, Tampa, and so on) are examples of nominal variables. So

are certain questions in opinion surveys. Imagine that I ask you what party you're planning to vote, and the options are Democratic, Republican, Other, None, and Don't Know.

In some cases, we may use numbers to describe our nominal variables. We may write "0" for male and "1" for female, for instance, but those numbers don't represent any amount or position in a ranking. They would be similar to the numbers that soccer players display on their back. They exist just to identify players, not to tell you which are better or worse.

Ordinal

In an ordinal scale, values are organized or ranked according to a magnitude, but without revealing their exact size in comparison to each other.

For example, you may be analyzing all countries in the world according to their Gross Domestic Product (GDP) per capita but, instead of showing me the specific GDP values, you just tell me which country is the first, the second, the third, and so on. This is an ordinal variable, as I've just learned about the countries' rankings according to their economic performance, but I don't know anything about how far apart they are in terms of GDP size.

In a survey, an example of ordinal scale would be a question about your happiness level: 1. Very happy; 2. Happy; 3. Not that happy; 4. Unhappy; 5. Very unhappy.

Interval

An interval scale of measurement is based on increments of the same size, but also on the lack of a true zero point, in the sense of that being the absolute lowest value. I know, it sounds confusing, so let me explain.

Imagine that you are measuring temperature in degrees Fahrenheit. The distance between 5 and 10 degrees is the same as the distance between 20 and 25 degrees: 5 units. So you can add and subtract temperatures, but you cannot say that 10 degrees is twice as hot as 5 degrees, even though 2×5 equals 10. The reason is related to the lack of a real zero. The zero point is just an arbitrary number, one like any other on the scale, not an absolute point of reference.

An example of interval scale coming from psychology is the intellectual quotient (IQ). If one person has an IQ of 140 and another person has an IQ of 70, you can say that the former is 70 units larger than the latter, but you cannot say that the former is *twice* as intelligent as the latter.

Ratio

Ratio scales have all the properties of the other previous scales, plus they also have a meaningful zero point. Weight, height, speed, and so on, are examples of ratio variables. If one car is traveling at 100 mph and another one is at 50, you can say that the first one is going 50 miles faster than the second, and you can also say that it's going twice as fast. If my daughter's height is 3 feet and mine is 6 feet (I wish), I am twice as tall as her.

Variables can be also classified into discrete and continuous. A **discrete** variable is one that can only adopt certain values. For instance, people can only have cousins in amounts that are whole numbers—four or five, that is, not 4.5 cousins. On the other hand, a **continuous** variable is one that can—at least in theory—adopt any value on the scale of measurement that you're using. Your weight in pounds can be 90, 90.1, 90.12, 90.125, or 90.1256. There's no limit to the number of decimal places that you can add to that. Continuous variables can be measured with a virtually endless degree of precision, if you have the right instruments.

In practical terms, the distinction between continuous and discrete variables isn't always clear. Sometimes you will treat a discrete variable as if it were continuous. Imagine that you're analyzing the number of children per couple in a certain country. You could say that the average is 1.8, which doesn't make a lot of sense for a truly discrete variable.

Similarly, you can treat a continuous variable as if it were discrete. Imagine that you're measuring the distance between galaxy centers. You could use nanometers with an infinite number of decimals (you'll end up with more digits than atoms in the universe!), but it would be better to use light-years and perhaps limit values to whole units. If the distance between two stars is 4.43457864... light-years, you could just round the figure to 4 light-years.

On Studies

Once a hypothesis is posed, it's time to test it against reality. I wish to measure if increased Twitter usage reduces book-writing output. I send an online poll to 30 friends who happen to be writers, asking them for the minutes spent on Twitter today and the words they have written. My (completely made up) results are on **Figure 4.2**. This is an **observational study**. To be more precise, it's a **cross-sectional study**, which means that it takes into account data collected just at a particular point in time.