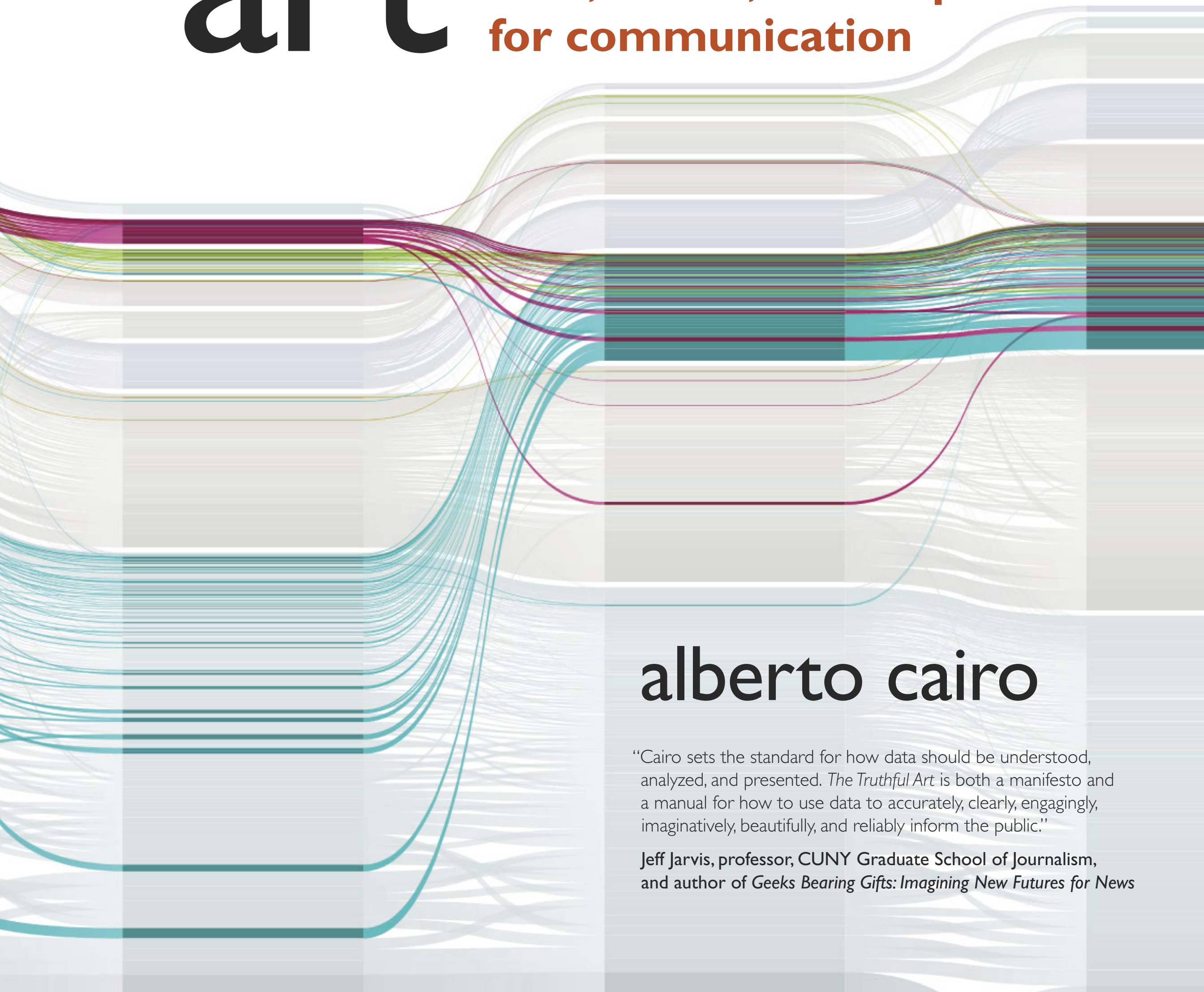
the truthful art

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

"To communicate data clearly, you have to think about it clearly. *The Truthful Art* dives deep and provides an enlightened introduction to the 'power tools' of data experts: science, statistics, and visualization."

—Fernanda Viégas and Martin Wattenberg, Google

"*The Truthful Art* is essential reading for my visual communication students and for anyone (at any level) who cares about telling a story visually. Get this book, read it, act on it. If you're looking for help to put your data visualization on the right track, this is it."

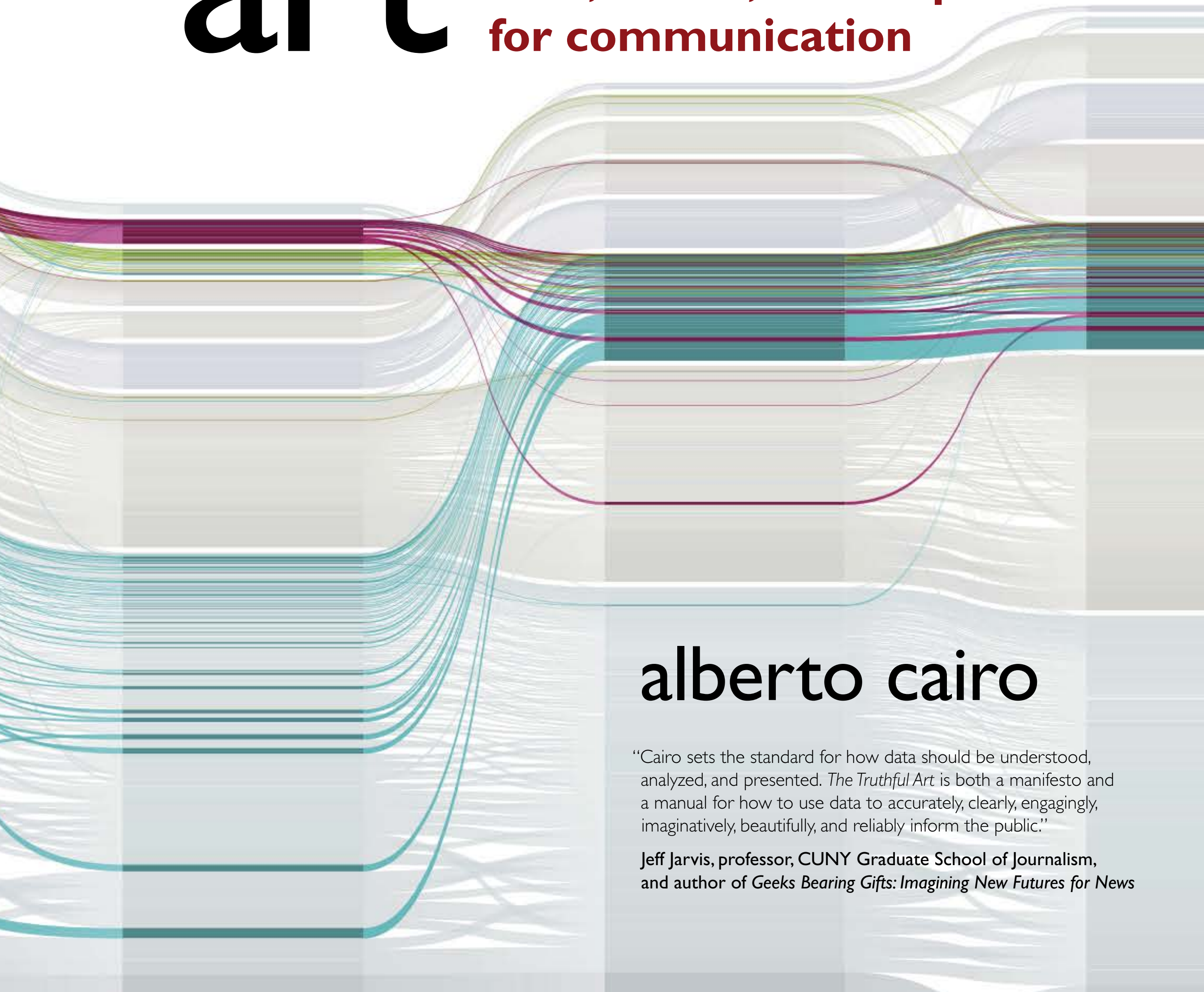
—John Grimwade, School of Visual Communication, Ohio University

"If I were smarter, had more patience with academia, and was more focused, I might turn out to be more like Alberto, closer to the brilliance that he applies to the nature of information architecture. His title explains a lot: truth represents a most fundamental of attitudes, in questions asked, answers given, and journeys taken. This [book] is a must on your thoughtful shelf of understanding."

—Richard Saul Wurman, founder of the TED Conference

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*

The Truthful Art:
Data, Charts, and Maps for Communication

Alberto Cairo

New Riders

Find us on the Web at www.newriders.com
New Riders is an imprint of Peachpit, a division of Pearson Education.
To report errors, please send a note to errata@peachpit.com

Acquisitions Editor: Nikki Echler McDonald
Production Editor: Tracey Croom
Development Editor: Cathy Lane
Copy Editor: Cathy Lane
Proofers: Patricia Pane, Kim Wimpsett
Compositor: Kim Scott, Bumpy Design
Indexer: James Minkin
Cover and Interior Designer: Mimi Heft
Cover Illustration: Moritz Stefaner

Notice of Rights

Notice of Liability

The information in this book is distributed on an “As Is” basis without warranty. While every precaution has been taken in the preparation of the book, neither the author nor Peachpit shall have any liability to any person or entity with respect to any loss or damage caused or alleged to be caused directly or indirectly by the instructions contained in this book or by the computer software and hardware products described in it.

Trademarks

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Peachpit was aware of a trademark claim, the designations appear as requested by the owner of the trademark. All other product names and services identified throughout this book are used in editorial fashion only and for the benefit of such companies with no intention of infringement of the trademark. No such use, or the use of any trade name, is intended to convey endorsement or other affiliation with this book.

ISBN 13: 9780321934079
ISBN 10: 0321934075

9 8 7 6 5 4 3 2 1

Printed and bound in the United States of America

To my father

Acknowledgments

I always chuckle when someone calls me an "expert" on visualization or infographics. As a journalist, I've made a profession of being an amateur, in the two senses of the word: someone who doesn't have a deep understanding of anything, but also someone who does what he does due to unabashed love for the craft.

This book is a tribute to that second kind of amateur, folks who bring good data to the world in a time when society is drowning in tsunamis of spin and misinformation. They know that it is possible to change the world for the better if we repeat the truth often and loud enough.

I'd like to first thank my University of Miami (UM) colleague Rich Beckman. It's not an exaggeration to say that I wouldn't be where I am today without his help, advice, and mentorship.

Seth Hamblin, a friend at *The Wall Street Journal*, passed away while I was writing this book. Seth was in love with infographics and visualization. When I told him about *The Truthful Art*, he got as excited as a kid. He was a beautiful human being, and he'll be missed.

To Greg Shepherd, dean of UM's School of Communication; Nick Tsinoremas, director of UM's Center for Computational Science (CCS); Sawsan Khuri, also from CCS, a great colleague and better friend; Sam Terilli, head of our department of journalism; and Kim Grinfeder, who leads our Interactive Media program. Also at UM, I'd like to thank my colleagues in the departments of Journalism and Interactive Media, and at the Center for Communication, Culture, and Change.

To Phil Meyer, whose book *Precision Journalism* inspired me many years ago. In fact, I wrote *The Truthful Art* with the goal of being a *Precision Journalism* for the new century.

To my past colleagues at *La Voz de Galicia*, *Diario16*, DPI Comunicación, *El Mundo*, and Editora Globo. Among them, Helio Gurovitz, the former managing editor at Globo's *Época* magazine, who combines a deep knowledge of journalism with an unusual sensibility of how to use numbers and graphics.

Also, to my clients and partners worldwide, particularly Jim Friedland, for all his support in the past two years.

Many people read this book while it was in the works. My editor, Nikki McDonald, and my copyeditor, Cathy Lane, kept an eye on me at all times, and did their best to make me meet deadlines (they failed).

Stephen Few sent me detailed notes about each chapter. Steve is both a friend and arguably my most severe critic. I've done my best to incorporate as much from his feedback as possible, but not all. I know that Steve will still disagree with some of my musings, but those disagreements can be great topics to ponder while enjoying some fine wine and cheese.

Erik Jacobsen also provided detailed feedback. His notes have been invaluable.

Three statisticians, Diego Kuonen, Heather Krause, and Jerzy Wieczorek, read the most technical chapters and made sure that I was not writing anything particularly silly. Others who commented on the book are: Andy Cotgreage, Kenneth Field, Jeff Jarvis, Scott Klein, Michael E. Mann, Isabel Meirelles, Fernanda Viégas, Martin Wattenberg, and Sisi Wei. Thank you all.

Thanks also to all the individuals and organizations who let me showcase their work in *The Truthful Art*. You're too numerous to mention, but you'll see your names in pages to come.

Some of my followers on Twitter volunteered for the last round of proofreading. They are: Mohit Chawla, Fernando Cucchietti, Stijn Debrouwere, Alex Lea, Neil Richards, Frédéric Schütz, and Tom Shanley.

Special thanks to Moritz Stefaner, for giving me permission to use one of his amazing graphics on the cover of this book.

To Nancy.

Finally, and above all, thanks to my family.

About the Author

Alberto Cairo is the Knight Chair in Visual Journalism at the School of Communication of the University of Miami (UM), where he heads specializations in infographics and data visualization. He's also director of the visualization program at UM's Center for Computational Science, and Visualization Innovator in Residence at Univisión.

He is the author of the books *Infografía 2.0: Visualización interactiva de información en prensa*, published just in Spain in 2008, and *The Functional Art: An Introduction to Information Graphics and Visualization* (New Riders, 2012).

In the past two decades, Cairo has been director of infographics and visualization at news organizations in Spain and Brazil, besides consulting with companies and educational institutions in more than 20 countries. He also was a professor at the University of North Carolina-Chapel Hill between 2005 and 2009.

Cairo's personal weblog is www.thefunctionalart.com. His corporate website is www.albertocairo.com.

His Twitter handle is @albertocairo.

Additional Materials

I designed many of the charts and maps you're about to see in *The Truthful Art*, but I haven't written much about the software I used to create them. If you're interested in learning about tools, please visit my weblog, www.thefunctionalart.com, and go to the **Tutorials and Resources** section on the upper menu.

There, you will find several articles and video lessons I recorded about programs and languages like R, iNzight, and Yeeron, among others.

Contents

PART III functional

PART IV practice

Preface

It All Begins with a Spark

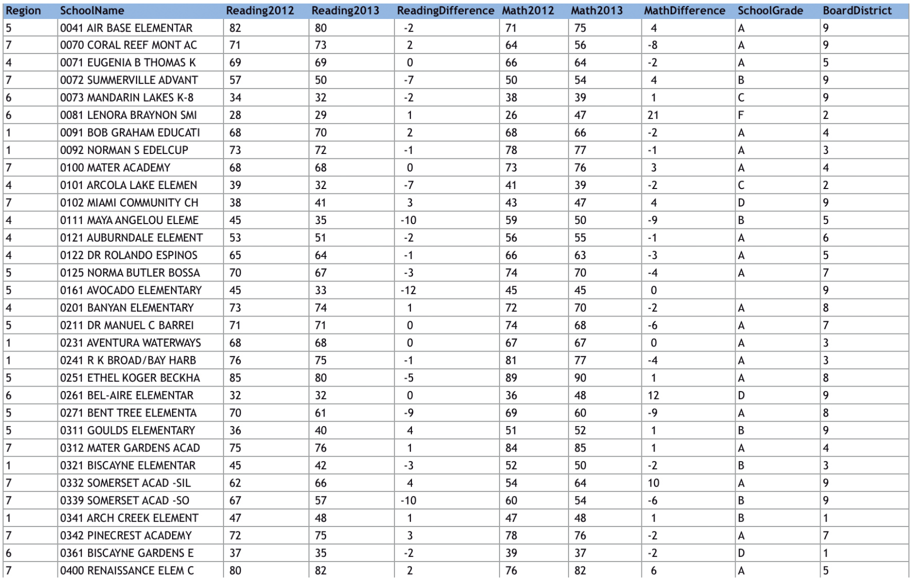
> Why is it that when one man builds a wall, the next man immediately needs to know what's on the other side?
>
> —Tyrion Lannister in George R.R. Martin's *A Game of Thrones*

There's probably something you don't know about college professors: we tend to have peculiar hobbies.

In October 2014, I spent my entire fall recess catching up with R, a programming language for statistical analysis; ggplot2, an R library that creates nice-looking charts; and Tableau, a data visualization program.[1] Learning any software tool without using it is impossible, so I needed some data to play with, and not just any data, but data I could care about.

A few months back, my family and I had moved to a new home, so I had briefly visited the Miami-Dade County Public Schools website (DadeSchools.net) to check the quality of the elementary school, middle school, and high school in our area. Each had a grade of A. I had felt reassured at the time, but also a bit

[1] I hope that this doesn't impress you. I am by no means an advanced user of any of these tools. All graphics in these pages were designed with very little knowledge of how to use them properly. For more information, visit http://www.r-project.org/, http://ggplot2.org/, and http://www.tableau.com.



Region	SchoolName	Reading2012	Reading2013	ReadingDifference	Math2012	Math2013	MathDifference	SchoolGrade	BoardDistrict
5	0041 AIR BASE ELEMENTAR	82	80	-2	71	75	4	A	9
7	0070 CORAL REEF MONT AC	71	73	2	64	56	-8	A	9
4	0071 EUGENIA B THOMAS K	69	69	0	66	64	-2	A	5
7	0072 SUMMERVILLE ADVANT	57	50	-7	50	54	4	B	9
6	0073 MANDARIN LAKES K-8	34	32	-2	38	39	1	C	9
6	0081 LENORA BRAYNON SMI	28	29	1	26	47	21	F	2
1	0091 BOB GRAHAM EDUCATI	68	70	2	68	66	-2	A	4
1	0092 NORMAN S EDELCUP	73	72	-1	78	77	-1	A	3
7	0100 MATER ACADEMY	68	68	0	73	76	3	A	4
4	0101 ARCOLA LAKE ELEMEN	39	32	-7	41	39	-2	C	2
7	0102 MIAMI COMMUNITY CH	38	41	3	43	47	4	D	9
4	0111 MAYA ANGELOU ELEME	45	35	-10	59	50	-9	B	5
4	0121 AUBURNDALE ELEMENT	53	51	-2	56	55	-1	A	6
4	0122 DR ROLANDO ESPINOS	65	64	-1	66	63	-3	A	5
5	0125 NORMA BUTLER BOSSA	70	67	-3	74	70	-4	A	7
5	0161 AVOCADO ELEMENTARY	45	33	-12	45	45	0		9
4	0201 BANYAN ELEMENTARY	73	74	1	72	70	-2	A	8
5	0211 DR MANUEL C BARREI	71	71	0	74	68	-6	A	7
1	0231 AVENTURA WATERWAYS	68	68	0	67	67	0	A	3
1	0241 R K BROAD/BAY HARB	76	75	-1	81	77	-4	A	3
5	0251 ETHEL KOGER BECKHA	85	80	-5	89	90	1	A	8
6	0261 BEL-AIRE ELEMENTAR	32	32	0	36	48	12	D	9
5	0271 BENT TREE ELEMENTA	70	61	-9	69	60	-9	A	8
5	0311 GOULDS ELEMENTARY	36	40	4	51	52	1	B	9
7	0312 MATER GARDENS ACAD	75	76	1	84	85	1	A	4
1	0321 BISCAYNE ELEMENTAR	45	42	-3	52	50	-2	B	3
7	0332 SOMERSET ACAD -SIL	62	66	4	54	64	10	A	9
7	0339 SOMERSET ACAD -SO	67	57	-10	60	54	-6	B	9
1	0341 ARCH CREEK ELEMENT	47	48	1	47	48	1	B	1
7	0342 PINECREST ACADEMY	72	75	3	78	76	-2	A	7
6	0361 BISCAYNE GARDENS E	37	35	-2	39	37	-2	D	1
7	0400 RENAISSANCE ELEM C	80	82	2	76	82	6	A	5

Figure P.1 The top portion of a spreadsheet with data from public schools in Miami-Dade County.

uneasy, as I hadn't done any comparison with schools in other neighborhoods. Perhaps my learning R and Tableau could be the perfect opportunity to do so.

DadeSchools.net has a neat data section, so I visited it and downloaded a spreadsheet of performance scores from all schools in the county. You can see a small portion of it—the spreadsheet is 461 rows tall—in **Figure P.1**. The figures in the Reading2012 and Reading2013 columns are the percentage of students from each school who attained a reading level considered as satisfactory in those two consecutive years. Math2012 and Math2013 correspond to the percentage of students who were deemed reasonably numerate for their age.

While learning how to write childishly simple scripts in R, I created rankings and bar charts to compare all schools. I didn't get any striking insight out of this exercise, although I ascertained that the three public schools in our neighborhood are decent indeed. My job was done, but I didn't stop there. I played a bit more.

I made R generate a scatter plot (**Figure P.2**). Each dot is one school. The position on the X-axis is the percentage of students who read at their proper level in 2013. The Y-axis is the same percentage for math proficiency. Both variables

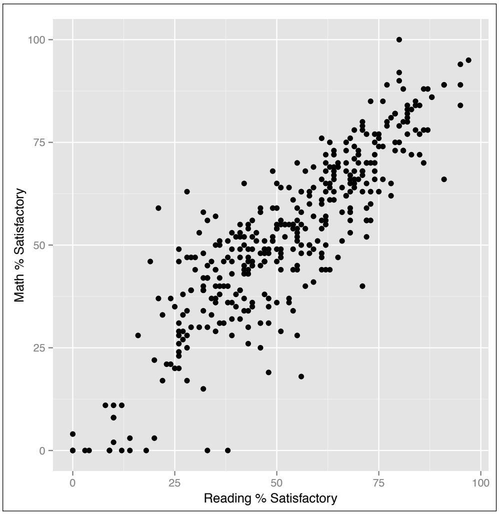


Figure P.2 Each dot on the chart is a school. Reading and math skills are strongly related.

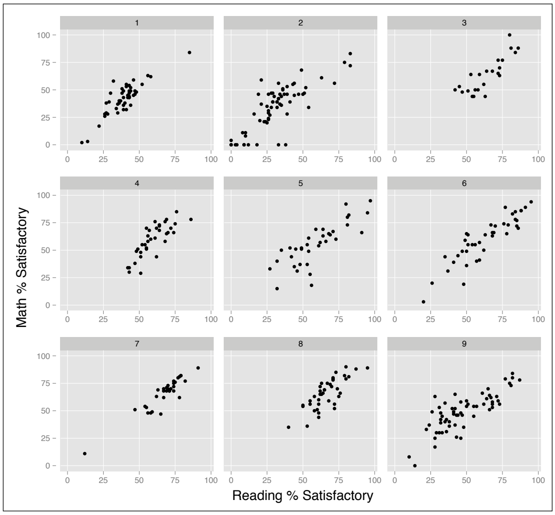


Figure P.3 The same data, divided by school board.

are clearly linked: the larger one gets, the larger the other one tends to become.[2] This makes sense. There is nothing very surprising other than a few outliers, and the fact that there are some schools in which no student is considered proficient in reading and/or math. This could be due to mistakes in the data set, of course.

After that, I learned how to write a short script to design not just one but several scatter plots, one for each of the nine school board districts in Miami-Dade County. It was then that I became really intrigued. See the results in **Figure P.3**.

There are quite a few interesting facts in that array. For instance, most schools in Districts 3, 7, and 8 are fine. Students in Districts 1 and 2, on the other hand, perform rather poorly.

At the time I was not familiar with the geography of the Miami-Dade school system, so I went online to find a map of it. I also visited the Census Bureau website to get a map of income data. I redesigned and overlaid them. (See **Figure P.4**. Warning: I didn't make any adjustment to these maps, so the overlap isn't perfect.) I got what I foresaw: the worst-performing districts, 1 and 2, encompass low-income neighborhoods, like Liberty City, Little Haiti, and Overtown.

[2] In statistics, we may call this a "strong positive correlation." But I'm getting a bit ahead of myself.

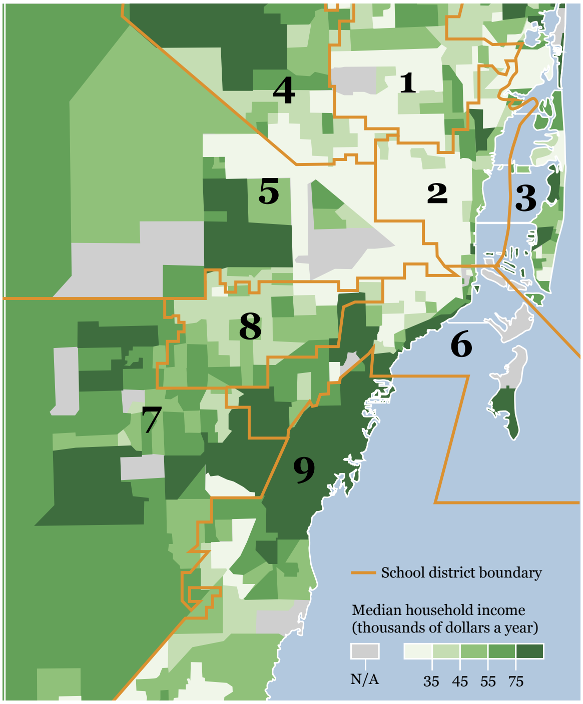


Figure P.4 Median household income in the nine school board districts of Miami-Dade County.

Immediately, questions started piling up in my head. Is the relationship between bad schools and low household income direct? Does a bad education lead to reduced wages? Or do kids coming from low-income families go to school being already disadvantaged, and that worsens the scores of the schools they attend? Am I getting causality right? What are other possible variables that affect both school performance and income?

What about the outliers in those charts, those schools in Districts 1 and 7, for instance, that are so far from their flocks? Or that school in District 3 that got a perfect score in math? And what about District 6? Schools in that plot are much more spread out than in the others. Is that related to the sharp divide between a richer strip on the east (Coconut Grove) and poorer blocks on the west within that school district?

And more: have all these percentages and grades changed substantially in the past few years? If so, is it due to real variation in the quality of our public education or because of changes in the methods researchers use to measure attainment? So many questions.

And so the seeds for many potential stories got planted. I didn't have an idea of what they might be at that point or if any of them would be worth telling. I just got a glimpse, an enticing clue. As most visualization designers and data journalists I know will tell you, sometimes it is not you who finds good ideas when you're seeking them. Instead, good ideas find you in the most unexpected circumstances.

Good ideas are fleeting things, so I feverishly scribbled notes in a computer application called Stickies, short messages for my future self, musings of a mind in a state of joyous flow. I added, "Find some education experts.[3] Ask them. Contact the folks running dadeschools.net. You'll likely need more data from the U.S. Census Bureau's website." And so on and so forth.

As the saying goes, every great story begins with a spark. Fun ensues.

[3] Here's Robert B. Reich—who isn't an expert on education but was Secretary of Labor under President Bill Clinton—in his book *Saving Capitalism* (2015): "A large portion of the money to support public schools comes from local property taxes. The federal government provides only about 10 percent of all funding, and the states provide 45 percent, on average. The rest is raised locally (…) Real estate markets in lower-income communities remain weak, so local tax revenues are down. As we segregate by income into different communities, schools in lower-income areas have fewer resources than ever. The result is widening disparities in funding per pupil, to the direct disadvantage of poor kids." Another possible clue to follow.

This page intentionally left blank

This page intentionally left blank

4

Of Conjectures and Uncertainty

We live in a world with a surfeit of information at our service. It is our choice whether we seek out data that reinforce our biases or choose to look at the world in a critical, rational manner, and allow reality to bend our preconceptions. In the long run, the truth will work better for us than our cherished fictions.

—Razib Khan, "The Abortion Stereotype,"
The New York Times (January 2, 2015)

To become a visualization designer, it is advisable to get acquainted with the language of research. Getting to know how the methods of science work will help us ascertain that we're not being fooled by our sources. We *will* still be fooled on a regular basis, but at least we'll be better equipped to avoid it if we're careful.

Up to this point I've done my best to prove that interpreting data and visualizations is to a great extent based on applying simple rules of thumb such as "compared to what/who/where/when," "always look for the pieces that are missing in the model," and "increase depth and breadth up to a reasonable point." I stressed those strategies first because in the past two decades I've seen that many designers and journalists are terrified by science and math for no good reason.[1]

It's time to get a bit more technical.

The Scientific Stance

Science isn't only what scientists do. Science is a stance, a way to look at the world, that everybody and anybody, regardless of cultural origins or background, can embrace—I'll refrain from writing "should," although I feel tempted. Here's one of my favorite definitions: "Science is a systematic enterprise that builds, organizes, and shares knowledge in the form of testable explanations and predictions."[2] **Science is, then, a set of methods, a body of knowledge, and the means to communicate it.**

Scientific discovery consists of an algorithm that, in a highly idealized form, looks like this:

1. You grow curious about a phenomenon, you explore it for a while, and then you formulate a plausible **conjecture** to describe it, explain it, or predict its behavior. This conjecture is just an informed hunch for now.
2. You transform your conjecture into a formal and testable proposition, called a **hypothesis**.
3. You thoroughly study and measure the phenomenon (under controlled conditions whenever it's possible). These measurements become **data** that you can use to **test** your hypothesis.
4. You draw **conclusions**, based on the evidence you have obtained. Your data and tests may force you to reject your hypothesis, in which case you'll need go to back to the beginning. Or your hypothesis may be tentatively corroborated.

[1] Journalists and designers aren't to blame. The education we've all endured is. Many of my peers in journalism school, back in the mid-1990s, claimed that they weren't "good at math" and that they only wanted to write. I still hear this from some of my students at the University of Miami. I guess that something similar can be seen among designers ("I just want to design!"). My response is usually, "If you cannot evaluate and manipulate data and evidence at all, what are you going to write (design) about?"

[2] From Mark Chang's *Principles of Scientific Methods* (2014). Another source to consult is "Science and Statistics," a 1976 article by George E. P. Box.
http://www-sop.inria.fr/members/Ian.Jermyn/philosophy/writings/Boxonmaths.pdf

5. Eventually, after repeated tests and after your work has been reviewed by your peers, members of your knowledge or scientific community, you may be able to put together a systematic set of interrelated hypotheses to describe, explain, or predict phenomena. We call this a **theory**. From this point on, always remember what the word "theory" really means. A theory isn't just a careless hunch.

These steps may open researchers' eyes to new paths to explore, so they don't constitute a process with a beginning and an end point but a loop. As you're probably guessing, we are returning to themes we've already visited in this book: good answers lead to more good questions. The scientific stance will never take us all the way to an absolute, immutable truth. What it may do—and it does it well—is to move us further to the right in the truth continuum.

From Curiosity to Conjectures

I use Twitter a lot, and on days when I spend more than one hour on it, I feel that I'm more distracted and not as productive as usual. I believe that this is something that many other writers experience. Am I right or am I wrong? Is this just something that I feel or something that is happening to everyone else? Can I transform my hunch into a general claim? For instance, can I say that an X percent increase of Twitter usage a day leads a majority of writers to a Y percent decrease in productivity? After all, I have read some books that make the bold claim that the Internet changes our brains in a negative way.[3]

What I've just done is to notice an interesting pattern, a possible cause-effect relationship (more Twitter = less productivity), and made a conjecture about it. It's a conjecture that:

1. It makes sense intuitively in the light of what we know about the world.
2. It is testable somehow.
3. It is made of ingredients that are naturally and logically connected to each other in a way that if you change any of them, the entire conjecture will crumble. This will become clearer in just a bit.

[3] The most famous one is *The Shallows* (2010), by Nicholas Carr. I am quite skeptical of this kind of claim, as anything that we do, see, hear, and so on, "changes" the wiring inside our skulls.

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.

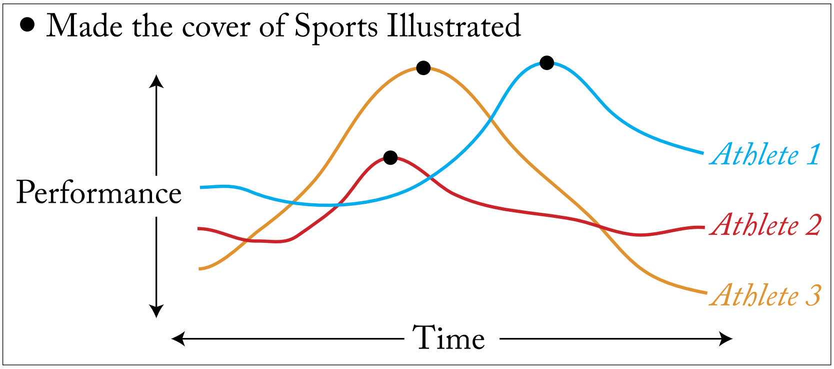


Figure 4.1 Athletes tend to underperform after they've appeared on the cover of *Sports Illustrated* magazine. Does the publication cast a curse on them?

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.

If I carefully document my friends' Twitter usage and the pages they write for a long time (a year, a decade, or since Twitter was launched), I'll have a **longitudinal study**. On **Figure 4.3**, I plotted Twitter usage (X-axis) versus words written (Y-axis) every year by three of my 30 fictional author friends. The relationship becomes clear: on average, the more time they spend on Twitter, the less they write for their own books. That's very unwise!

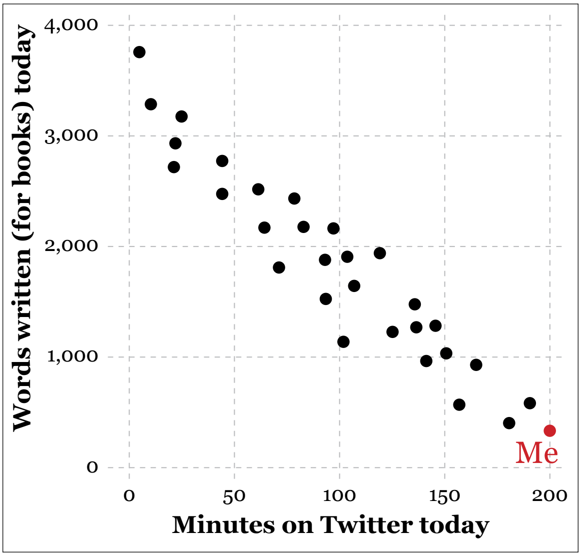


Figure 4.2 Writer friends don't let their writer friends use Twitter when they are on a deadline.

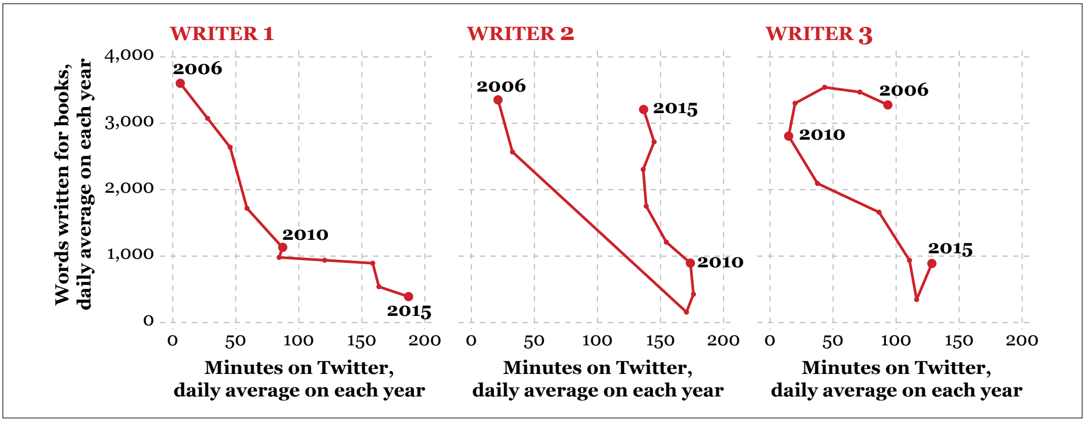


Figure 4.3 The more writers use Twitter, the fewer words they write. Don't forget that this is all bogus data.

The choice of what kind of study to conduct depends on many factors. Doing longitudinal studies is usually much more difficult and expensive, as you'll need to follow the same people for a long time. Cross-sectional studies are faster to build but, in general, their results aren't very conclusive.[7]

Going back to my inquiry, I face a problem: I am trying to draw an inference ("writers can benefit from using Twitter less") from a particular group of writers. That is, I am trying to study something about a **population**, *all* writers, based on a **sample** of those writers, my friends. **But are my friends representative of all writers? Are inferences drawn from my sample applicable to the entire population?**

Always be suspicious of studies whose samples have not been randomly chosen.[8] Not all scientific research is based on random sampling, but analyzing a random sample of writers chosen from the population of all writers will yield more accurate results than a cherry-picked or self-selected sample.

This is why we should be wary of the validity of things like news media online polls. If you ask your audience to opine on a subject, you cannot claim that you've learned something meaningful about what the public in general thinks. You cannot even say that you know the opinion of your own audience! You've just heard from those readers who feel strongly about the topic you asked about, as they are the ones who are more likely to participate in your poll.

Randomization is useful to deal with **extraneous variables**, mentioned in Chapter 3 where I advised you to always try to increase depth and breadth. It may be that the results of my current exploration are biased because a good portion of my friends are quite geeky, and so they use Twitter a lot. In this case,

[7] Different kinds of studies beget different kinds of conclusions. For instance, in a cross-sectional study you might be able to conclude, "In the population we studied, the kind of people who tweet little are also the kind of people who write a lot," but you cannot add anything about time change or causality. If you do a longitudinal study, you might conclude, "In the population studied, the kind of people who choose to start tweeting less are also the kind who start writing more," but you cannot say anything about causality. If you then decide to conduct a controlled experiment, you might be able to say, "In the population studied, whichever kind of person you are, if you start tweeting less, then you'll start writing more." But even in this case you cannot say anything about how many of those people are naturally inclined to tweet or to write. Science is hard!

[8] Many introduction to statistics textbooks include a section about how random sampling is conducted. I recommend that you take a look at a couple of them. Before you do so, though, you may want to read this nice introduction by Statistics Canada: http://tinyurl.com/or47fyr.

the geekiness level, if it could be measured, would distort my model, as it would affect the relationship between predictor and outcome variable.

Some researchers distinguish between two kinds of extraneous variables. Sometimes we can identify an extraneous variable and incorporate it into our model, in which case we'd be dealing with a **confounding variable**. I know that it may affect my results, so I consider it for my inquiry to minimize its impact. In an example seen in previous chapters, we controlled for population change and for variation in number of motor vehicles when analyzing deaths in traffic accidents.

There's a second, more insidious kind of extraneous variable. Imagine that I don't know that my friends are indeed geeky. If I were unaware of this, I'd be dealing with a **lurking** variable. A **lurking variable** is an extraneous variable that we don't include in our analysis for the simple reason that its existence is unknown to us, or because we can't explain its connection to the phenomenon we're studying.

When reading studies, surveys, polls, and so on, always ask yourself: did the authors rigorously search for lurking variables and transform them into confounding variables that they can ponder? Or are there other possible factors that they ignored and that may have distorted their results?[9]

Doing Experiments

Whenever it is realistic and feasible to do so, researchers go beyond observational studies and design **controlled experiments**, as these can help minimize the influence of confounding variables. There are many kinds of experiments, but many of them share some characteristics:

1. They observe a large number of subjects that are representative of the population they want to learn about. Subjects aren't necessarily people. A subject can be any entity (a person, an animal, an object, etc.) that can be studied in controlled conditions, in isolation from external influences.

[9] One of the best quotes about the imperfection of all our rational inquiry methods, including science, comes from former Secretary of Defense Donald Rumsfeld. In a 2002 press conference about using the possible existence of weapons of mass destruction in Iraq as a reason to go to war with that country, he said, "Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult one." It's acceptable to argue that Rumsfeld was being disingenuous, as some of those "unknown unknowns" were actually "known unknowns" or even "known knowns."

2. Subjects are divided into at least two groups, an **experimental** group and a **control** group. This division will in most cases be made blindly: the researchers and/or the subjects don't know which group each subject is assigned to.

3. Subjects in the experimental group are exposed to some sort of condition, while the control group subjects are exposed to a different condition or to no condition at all. This condition can be, for instance, adding different chemical compounds to fluids and comparing the changes they suffer, or exposing groups of people to different kinds of movies to test how they influence their behavior.

4. Researchers measure what happens to subjects in the experimental group and what happens to subjects in the control group, and they compare the results.

   If the differences between experimental and control groups are noticeable enough, researchers may conclude that the condition under study may have played some role.[10]

We'll learn more about this process in Chapter 11.

When doing visualizations based on the results of experiments, it's important to not just read the abstract of the paper or article and its conclusions. Check if the journal in which it appeared is peer reviewed and how it's regarded in its knowledge community.[11] Then, take a close look at the paper's methodology. Learn about how the experiments were designed and, in case you don't understand it, contact other researchers in the same area and ask. This is also valid for observational studies. A small dose of constructive skepticism can be very healthy.

In October 2013, many news publications echoed the results of a study by psychologists David Comer Kidd and Emanuele Castano which showed that reading literary fiction temporarily enhances our capacity to understand other people's

[10] To be more precise, scientists compare these differences to a hypothetical range of studies with the same sample size and design but where the condition is known to have no effect. This check (statistical hypothesis testing) helps to prevent spurious conclusions due to small samples or high variability. This check isn't about whether the effect is large in an absolute/pragmatic sense, as we'll see soon.

[11] You can search for the impact factor (IF) of the publication. This is a measure of how much it is cited by other publications. It's not a perfect quality measure, but it helps.

mental states. The media immediately started writing headlines like "Reading fiction improves empathy!"[12]

The finding was consistent with previous observations and experiments, but reporting on a study after reading just its abstract is dangerous. What were the researchers really comparing?

In one of the experiments, they made two groups of people read either three works of literary fiction or three works of nonfiction. After the readings, the people in the literary fiction group were better at identifying facially expressed emotions than those in the nonfiction group.

The study looked sound to me when I read it, but it left crucial questions in the air: what *kinds* of literary fiction and nonfiction did the subjects read? It seems predictable that you'll feel more empathetic toward your neighbor after reading *To Kill a Mockingbird* than after, say, Thomas Piketty's *Capital in the Twenty-First Century*, a brick-sized treaty on modern economics that many bought—myself included—but just a few read.

But what if researchers had compared *To Kill a Mockingbird* to Katherine Boo's *Behind the Beautiful Forevers*, a haunting and emotional piece of journalistic reporting? And, even if they had compared literary fiction with literary non-fiction, is it even possible to measure how "literary" either book is? Those are the kinds of questions that you need to ask either to the researchers that conducted the study or, in case they cannot be reached for comment, to other experts in the same knowledge domain.

About Uncertainty

Here's a dirty little secret about data: it's always noisy and uncertain.[13]

To understand this critical idea, let's begin with a very simple study. I want to know my weight. I've been exercising lately, and I want to check the results. I step on the scale one morning and I read 192 lbs.

[12] "Reading Literary Fiction Improves Theory of Mind." http://scottbarrykaufman.com/wp-content/uploads/2013/10/Science-2013-Kidd-science.1239918.pdf

[13] Moreover, data sets are sometimes incomplete and contain errors, redundancies, typos, and more. For an overview, see Paul D. Allison's *Missing Data* (2002). To deal with this problem, tools like OpenRefine (http://openrefine.org/) may come in handy.

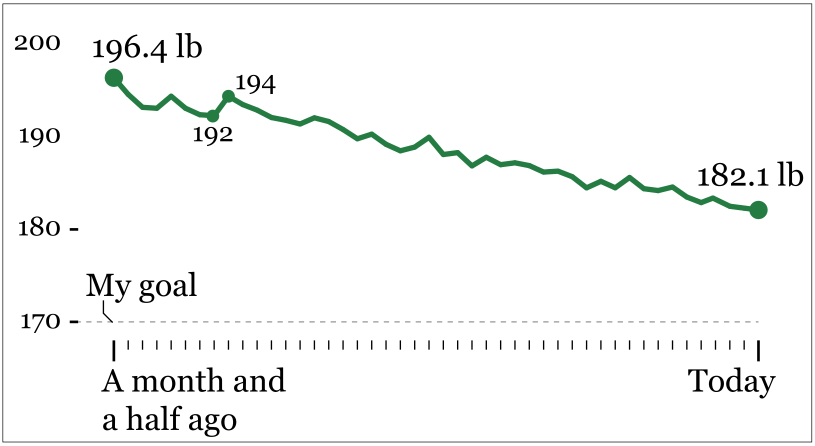


Figure 4.4 Randomness at work—weight change in a month and a half.

Out of curiosity, I decide to weigh myself again the day after. The scale shows 194 lbs. Damn it! How is that even possible? I've been eating better and running regularly, and my weight had already dropped from 196.4 lb. There's surely some sort of discrepancy between the measurements I'm getting and my true weight. I decide to continue weighing myself for more than a month.

The results are in **Figure 4.4**. There's a clear downward trend, but it only becomes visible when I display more than five or six days in a row. If I zoom in too much to the chart and just pay attention to two or three days, I'd be fooled into thinking that the noise in the data means something.

There may be different reasons for this wacky fluctuation to happen. My first thought is that my scale may not be working well, but then I realize that if there were some sort of technical glitch, it would bias *all* my measurements systematically. So the scale is not the source of the fluctuation.

It might be that I don't always balance my weight equally between my feet or that I'm weighing myself at slightly different times on each day. We tend to be a bit heavier in the afternoon than right after we wake up because we lose water while we sleep, and our body has already processed the food we ate the night before. But I was extremely careful with all those factors. I did weigh myself exactly at 6:45 a.m. every single day. And still, the variation is there. Therefore, I can only conclude that it's the result of factors that I can't possibly be aware of. I'm witnessing **randomness**.

Data always vary randomly because the object of our inquiries, nature itself, is also random. We can analyze and predict events in nature with an increasing amount of precision and accuracy, thanks to improvements in our techniques and instruments, but a certain amount of random variation, which gives rise to **uncertainty**, is inevitable. This is as true for weight measurements at home as it is true for anything else that you want to study: stock prices, annual movie box office takings, ocean acidity, variation of the number of animals in a region, rainfall or droughts—anything.

If we pick a random sample of 1,000 people to analyze political opinions in the United States, we cannot be 100 percent certain that they are perfectly representative of the entire country, no matter how thorough we are. If our results are that 48.2 percent of our sample are liberals and 51.8 percent are conservatives, we cannot conclude that the entire U.S. population is exactly 48.2 percent liberal and 51.8 percent conservative.

Here's why: if we pick a completely different random sample of people, the results may be 48.4 percent liberal and 51.6 percent conservative. If we then draw a third sample, the results may be 48.7 percent liberal and 51.3 percent conservative (and so forth).

Even if our methods for drawing random samples of 1,000 people are rigorous, there will always be some amount of uncertainty. We may end up with a slightly higher or lower percentage of liberals or conservatives out of pure chance. This is called **sample variation**.

Uncertainty is the reason why researchers will never just tell you that 51.8 percent of the U.S. population is conservative, after observing their sample of 1,000 people. What they will tell you, with a high degree of confidence (usually 95 percent, but it may be more or less than that), is that the percentage of conservatives seems to be indeed 51.8 percent, but that there's an error of plus or minus 3 percentage points (or any other figure) in that number.

Uncertainty can be represented in our visualizations. See the two charts in **Figure 4.5**, designed by professor **Adrian E. Raftery**, from the University of Washington. As they display projections, the amount of uncertainty increases with time: the farther away we depart from the present, the more uncertain our projections will become, meaning that the value that the variable "population" could adopt falls inside an increasingly wider range.

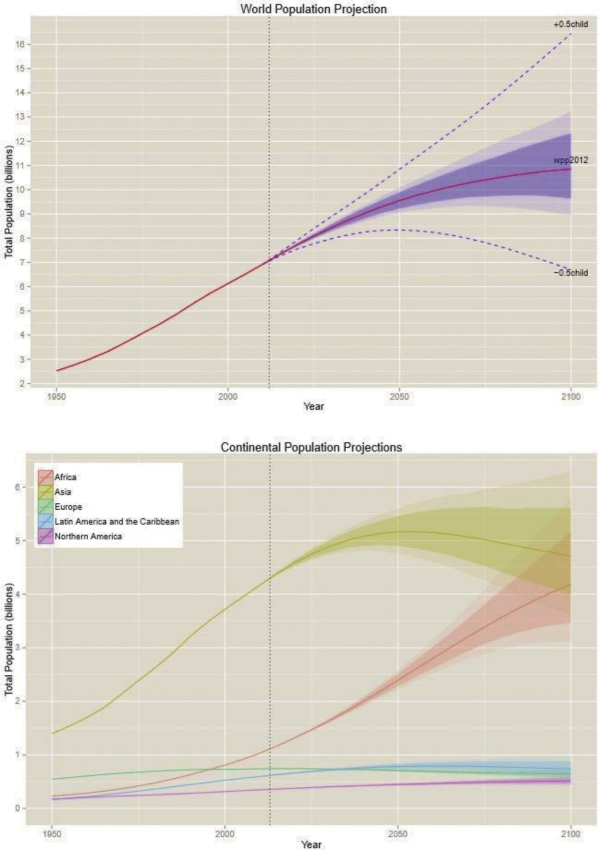


Figure 4.5 Charts by Adrian E. Raftery (University of Washington) who explains, "The top chart shows world population projected to 2100. Dotted lines are the range of error using the older scenarios in which women would have 0.5 children more or less than what's predicted. Shaded regions are the uncertainties. The darker shading is the 80 percent confidence bars, and the lighter shading shows the 95 percent confidence bars. The bottom chart represents population projections for each continent." http://www.sciencedaily.com/releases/2014/09/140918141446.htm

The hockey stick chart of world temperatures, in Chapter 2, is another example of uncertainty visualized. In that chart, there's a light gray strip behind the dark line representing the estimated temperature variation. This gray strip is the uncertainty. It grows narrower the closer we get to the twentieth century because instruments to measure temperature, and our historical records, have become much more reliable.

We'll return to testing, uncertainty, and confidence in Chapter 11. Right now, after clarifying the meaning of important terms, it's time to begin exploring and visualizing data.

To Learn More

- Skeptical Raptor's "How to evaluate the quality of scientific research." http://www.skepticalraptor.com/skepticalraptorblog.php/how-evaluate-quality-scientific-research/
- *Nature* magazine's "Twenty Tips For Interpreting Scientific Claims." http://www.nature.com/news/policy-twenty-tips-for-interpreting-scientific-claims-1.14183
- Box, George E. P. "Science and Statistics." Journal of the American Statistical Association, Vol. 71, No. 356. (Dec., 1976), pp. 791-799. Available online: http://www-sop.inria.fr/members/Ian.Jermyn/philosophy/writings/Boxonmaths.pdf
- Prothero, Donald R. *Evolution: What the Fossils Say and Why It Matters*. New York: Columbia University Press, 2007. Yes, it's a book about paleontology, but, leaving aside the fact that prehistoric beasts are fascinating, the author offers one of the clearest and most concise introductions to science I've read.

This page intentionally left blank

Index