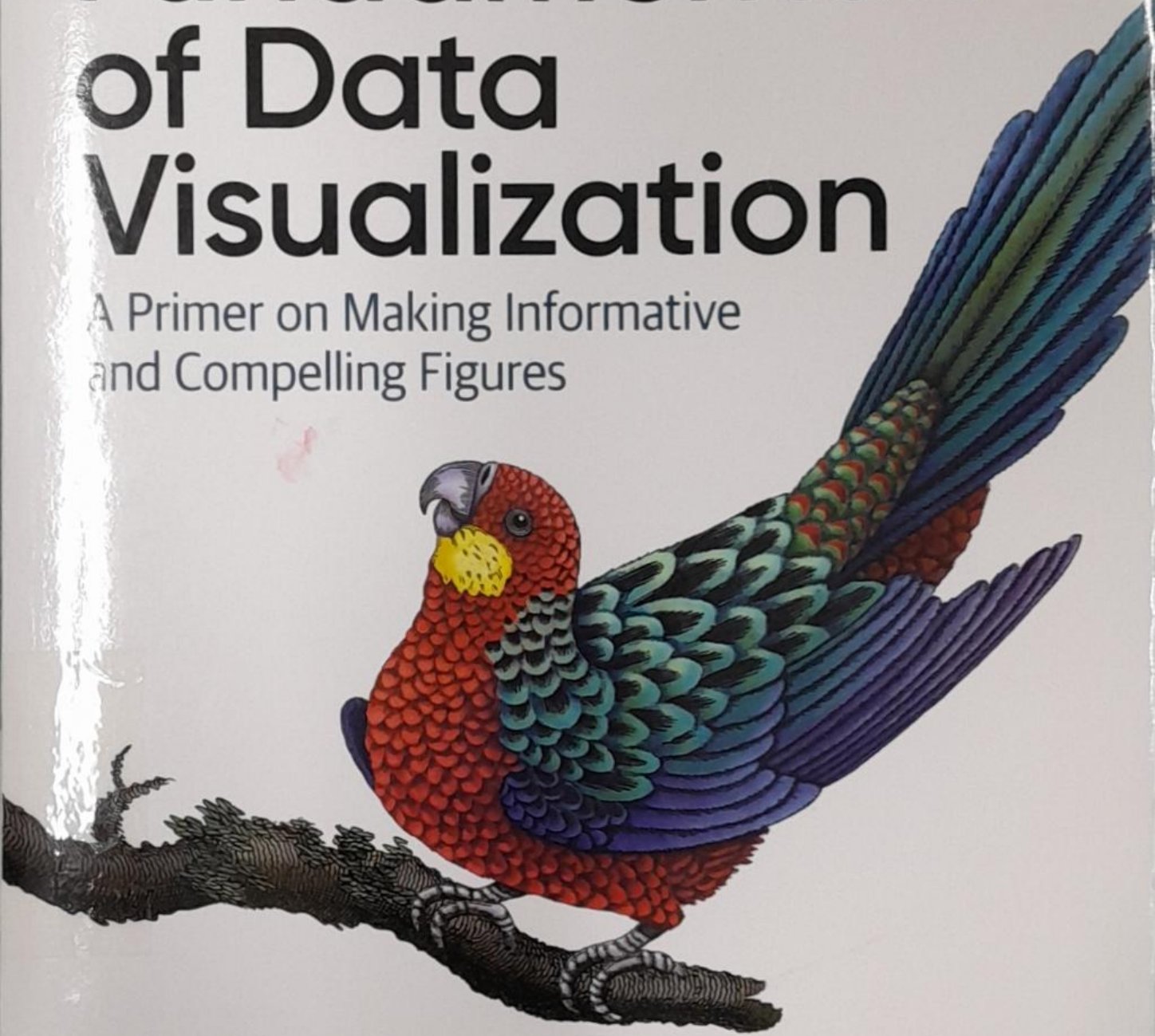


O'REILLY®

Fundamentals of Data Visualization

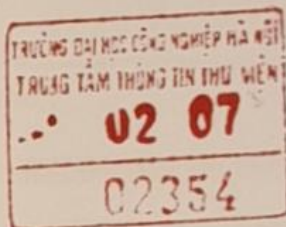
A Primer on Making Informative
and Compelling Figures



Claus O. Wilke

Fundamentals of Data Visualization

*A Primer on Making Informative
and Compelling Figures*



Claus O. Wilke

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Fundamentals of Data Visualization

by Claus O. Wilke

Copyright © 2019 Claus O. Wilke. All rights reserved.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://oreilly.com>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editors: Mike Loukides and
Melissa Potter
Production Editor: Kristen Brown
Copyeditor: Rachel Head
Proofreader: James Fraleigh

Indexer: Ellen Troutman-Zaig
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Claus Wilke

March 2019: First Edition

Revision History for the First Edition

2019-03-15: First Release
2019-05-03: Second Release
2019-06-07: Third Release
2020-03-27: Fourth Release

See <http://oreilly.com/catalog/errata.csp?isbn=9781492031086> for release details.

The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. *Fundamentals of Data Visualization*, the cover image, and related trade dress are trademarks of O'Reilly Media, Inc.

The views expressed in this work are those of the author, and do not represent the publisher's views. While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-492-03108-6

[LSI]

Table of Contents

Preface.....	xi
1. Introduction.....	1
Ugly, Bad, and Wrong Figures	2
<hr/>	
Part I. From Data to Visualization	
2. Visualizing Data: Mapping Data onto Aesthetics.....	7
Aesthetics and Types of Data	7
Scales Map Data Values onto Aesthetics	10
3. Coordinate Systems and Axes.....	13
Cartesian Coordinates	13
Nonlinear Axes	16
Coordinate Systems with Curved Axes	22
4. Color Scales.....	27
Color as a Tool to Distinguish	27
Color to Represent Data Values	29
Color as a Tool to Highlight	33
5. Directory of Visualizations.....	37
Amounts	37
Distributions	38
Proportions	39
x-y relationships	41
Geospatial Data	42

Uncertainty	43
6. Visualizing Amounts	45
Bar Plots	45
Grouped and Stacked Bars	50
Dot Plots and Heatmaps	53
7. Visualizing Distributions: Histograms and Density Plots	59
Visualizing a Single Distribution	59
Visualizing Multiple Distributions at the Same Time	64
8. Visualizing Distributions:	
Empirical Cumulative Distribution Functions and Q-Q Plots	71
Empirical Cumulative Distribution Functions	71
Highly Skewed Distributions	74
Quantile-Quantile Plots	78
9. Visualizing Many Distributions at Once	81
Visualizing Distributions Along the Vertical Axis	81
Visualizing Distributions Along the Horizontal Axis	88
10. Visualizing Proportions	93
A Case for Pie Charts	93
A Case for Side-by-Side Bars	97
A Case for Stacked Bars and Stacked Densities	99
Visualizing Proportions Separately as Parts of the Total	101
11. Visualizing Nested Proportions	105
Nested Proportions Gone Wrong	105
Mosaic Plots and Treemaps	107
Nested Pies	111
Parallel Sets	113
12. Visualizing Associations Among Two or More Quantitative Variables	117
Scatterplots	117
Correlograms	121
Dimension Reduction	124
Paired Data	127
13. Visualizing Time Series and Other Functions of an Independent Variable	131
Individual Time Series	131
Multiple Time Series and Dose-Response Curves	135

Time Series of Two or More Response Variables	138
14. Visualizing Trends	145
Smoothing	145
Showing Trends with a Defined Functional Form	151
Detrending and Time-Series Decomposition	155
15. Visualizing Geospatial Data	161
Projections	161
Layers	169
Choropleth Mapping	172
Cartograms	176
16. Visualizing Uncertainty	181
Framing Probabilities as Frequencies	181
Visualizing the Uncertainty of Point Estimates	186
Visualizing the Uncertainty of Curve Fits	197
Hypothetical Outcome Plots	201

Part II. Principles of Figure Design

17. The Principle of Proportional Ink	207
Visualizations Along Linear Axes	208
Visualizations Along Logarithmic Axes	212
Direct Area Visualizations	215
18. Handling Overlapping Points	219
Partial Transparency and Jittering	219
2D Histograms	222
Contour Lines	225
19. Common Pitfalls of Color Use	233
Encoding Too Much or Irrelevant Information	233
Using Nonmonotonic Color Scales to Encode Data Values	237
Not Designing for Color-Vision Deficiency	238
20. Redundant Coding	243
Designing Legends with Redundant Coding	243
Designing Figures Without Legends	250

21. Multipanel Figures.....	255
Small Multiples	255
Compound Figures	260
22. Titles, Captions, and Tables.....	267
Figure Titles and Captions	267
Axis and Legend Titles	270
Tables	273
23. Balance the Data and the Context.....	277
Providing the Appropriate Amount of Context	277
Background Grids	282
Paired Data	287
Summary	290
24. Use Larger Axis Labels.....	291
25. Avoid Line Drawings.....	297
26. Don't Go 3D.....	305
Avoid Gratuitous 3D	305
Avoid 3D Position Scales	307
Appropriate Use of 3D Visualizations	313

Part III. Miscellaneous Topics

27. Understanding the Most Commonly Used Image File Formats.....	319
Bitmap and Vector Graphics	319
Lossless and Lossy Compression of Bitmap Graphics	321
Converting Between Image Formats	324
28. Choosing the Right Visualization Software.....	325
Reproducibility and Repeatability	326
Data Exploration Versus Data Presentation	327
Separation of Content and Design	330
29. Telling a Story and Making a Point.....	333
What Is a Story?	334
Make a Figure for the Generals	337
Build Up Toward Complex Figures	341

Make Your Figures Memorable	343
Be Consistent but Don't Be Repetitive	345
Annotated Bibliography	351
Technical Notes	355
References	357
Index	361

Preface

If you are a scientist, an analyst, a consultant, or anybody else who has to prepare technical documents or reports, one of the most important skills you need to have is the ability to make compelling data visualizations, generally in the form of figures. Figures will typically carry the weight of your arguments. They need to be clear, attractive, and convincing. The difference between good and bad figures can be the difference between a highly influential or an obscure paper, a grant or contract won or lost, a job interview gone well or poorly. And yet, there are surprisingly few resources to teach you how to make compelling data visualizations. Few colleges offer courses on this topic, and there are not that many books on this topic either. (Some exist, of course.) Tutorials for plotting software typically focus on how to achieve specific visual effects rather than explaining why certain choices are preferred and others not. In your day-to-day work, you are simply expected to know how to make good figures, and if you're lucky you have a patient adviser who teaches you a few tricks as you're writing your first scientific papers.

In the context of writing, experienced editors talk about "ear," the ability to hear (internally, as you read a piece of prose) whether the writing is any good. I think that when it comes to figures and other visualizations, we similarly need "eye," the ability to look at a figure and see whether it is balanced, clear, and compelling. And just as is the case with writing, the ability to see whether a figure works or not can be learned. Having eye means primarily that you are aware of a larger collection of simple rules and principles of good visualization, and that you pay attention to little details that other people might not.

In my experience, again just as in writing, you don't develop eye by reading a book over the weekend. It is a lifelong process, and concepts that are too complex or too subtle for you today may make much more sense five years from now. I can say for myself that I continue to evolve in my understanding of figure preparation. I routinely try to expose myself to new approaches, and I pay attention to the visual and design choices others make in their figures. I'm also open to changing my mind. I might today consider a given figure great, but next month I might find a reason to

criticize it. So with this in mind, please don't take anything I say as gospel. Think critically about my reasoning for certain choices and decide whether you want to adopt them or not.

While the materials in this book are presented in a logical progression, most chapters can stand on their own, and there is no need to read the book cover to cover. Feel free to skip around, to pick out a specific section that you're interested in at the moment, or one that covers a particular design choice you're pondering. In fact, I think you will get the most out of this book if you don't read it all at once, but rather read it piecemeal over longer stretches of time, try to apply just a few concepts from the book in your figuremaking, and come back to read about other concepts or reread sections on concepts you learned about a while back. You may find that the same chapter tells you different things if you reread it after a few months have passed.

Even though nearly all of the figures in this book were made with R and `ggplot2`, I do not see this as an R book. I am talking about general principles of figure preparation. The software used to make the figures is incidental. You can use any plotting software you want to generate the kinds of figures I'm showing here. However, `ggplot2` and similar packages make many of the techniques I'm using much simpler than other plotting libraries. Importantly, because this is not an R book, I do not discuss code or programming techniques anywhere in this book. I want you to focus on the concepts and the figures, not on the code. If you are curious about how any of the figures were made, you can check out the book's source code at its GitHub repository (<https://github.com/clauswilke/dataviz>).

Thoughts on Graphing Software and Figure-Preparation Pipelines

I have over two decades of experience preparing figures for scientific publications and have made thousands of figures. If there has been one constant over these two decades, it's been the change in figure preparation pipelines. Every few years, a new plotting library is developed or a new paradigm arises, and large groups of scientists switch over to the hot new toolkit. I have made figures using `gnuplot`, `Xfig`, `Mathematica`, `Matlab`, `matplotlib` in Python, base R, `ggplot2` in R, and possibly others I can't currently remember. My current preferred approach is `ggplot2` in R, but I don't expect that I'll continue using it until I retire.

This constant change in software platforms is one of the key reasons why this book is not a programming book and why I have left out all code examples. I want this book to be useful to you regardless of which software you use, and I want it to remain valuable even once everybody has moved on from `ggplot2` and is using the next new thing. I realize that this choice may be frustrating to some `ggplot2` users who would like to know how I made a given figure. However, anybody who is curious about my

coding techniques can read the source code of the book. It is available. Also, in the future I may release a supplementary document focused just on the code.

One thing I have learned over the years is that automation is your friend. I think figures should be autogenerated as part of the data analysis pipeline (which should also be automated), and they should come out of the pipeline ready to be sent to the printer, with no manual post-processing needed. I see a lot of trainees autogenerate rough drafts of their figures, which they then import into Illustrator for sprucing up. There are several reasons why this is a bad idea. First, the moment you manually edit a figure, your final figure becomes irreproducible. A third party cannot generate the exact same figure you did. While this may not matter much if all you did was change the font of the axis labels, the lines are blurry, and it's easy to cross over into territory where things are less clear-cut. As an example, let's say you want to manually replace cryptic labels with more readable ones. A third party may not be able to verify that the label replacement was appropriate. Second, if you add a lot of manual post-processing to your figure-preparation pipeline, then you will be more reluctant to make any changes or redo your work. Thus, you may ignore reasonable requests for change made by collaborators or colleagues, or you may be tempted to reuse an old figure even though you've actually regenerated all the data. Third, you may yourself forget what exactly you did to prepare a given figure, or you may not be able to generate a future figure on new data that exactly visually matches your earlier figure. These are not made-up examples. I've seen all of them play out with real people and real publications.

For all these reasons, interactive plot programs are a bad idea. They inherently force you to manually prepare your figures. In fact, it's probably better to autogenerate a figure draft and spruce it up in Illustrator than to make the entire figure by hand in some interactive plot program. Please be aware that Excel is an interactive plot program as well and is not recommended for figure preparation (or data analysis).

One critical component in a book on data visualization is the feasibility of the proposed visualizations. It's nice to invent some elegant new type of visualization, but if nobody can easily generate figures using this visualization then there isn't much use to it. For example, when Tufte first proposed sparklines nobody had an easy way of making them. While we need visionaries who move the world forward by pushing the envelope of what's possible, I envision this book to be practical and directly applicable to working data scientists preparing figures for their publications. Therefore, the visualizations I propose in the subsequent chapters can be generated with a few lines of R code via `ggplot2` and readily available extension packages. In fact, nearly every figure in this book, with the exception of a few figures in Chapters 26, 27, and 28, was autogenerated exactly as shown.