Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

Chapter 28. Choosing the Right Visualization Software

Throughout this book, I have purposefully avoided one critical question of data visualization: what tools should we use to generate our figures? This question can generate heated discussions, as many people have strong emotional bonds to the specific tools they are familiar with. I have often seen people vigorously defend their own preferred tools instead of investing time in learning a new approach, even if the new approach has objective benefits. And I will say that sticking with the tools you know is not entirely unreasonable. Learning any new tool will require time and effort, and you will have to go through a painful transition period where getting things done with the new tool is much more difficult than it was with the old tool. Whether going through this period is worth the effort can usually only be evaluated in retrospect, after one has invested in learning the new tool. Therefore, regardless of the pros and cons of different tools and approaches, the overriding principle is that you need to pick a tool that works for you. If you can make the figures you want to make, without excessive effort, then that's all that matters.

> **NOTE**
>
> The best visualization software is the one that allows you to make the figures you need.

Having said this, I do think there are general principles we can use to assess the relative merits of different approaches to producing visualizations. These principles roughly break down by how reproducible the visualizations are, how easy it is to rapidly explore the data, and to what extent the visual appearance of the output can be tweaked.

Reproducibility and Repeatability

In the context of scientific experiments, we refer to work as *reproducible* if the overarching scientific finding of the work will remain unchanged if a different research group performs the same type of study. For example, if one research group finds that a new pain medication reduces perceived headache pain significantly without causing noticeable side effects and a different group subsequently studies the same medication on a different patient group and has the same findings, then the work is reproducible. By contrast, work is *repeatable* if very similar or identical measurements can be obtained by the same person repeating the exact same measurement procedure on the same equipment. For example, if I weigh my dog and find she weighs 41 lbs and then I weigh her again on the same scales and find again that she weighs 41 lbs, then this measurement is repeatable.

With minor modifications, we can apply these concepts to data visualization. A visualization is reproducible if the plotted data is available and any data transformations that may have been applied before plotting are exactly specified. For example, if you make a figure and then send me the exact data that you plotted, then I can prepare a figure that looks substantially similar. We may be using slightly different fonts or colors or point sizes to display the same data, so the two figures may not be exactly identical, but your figure and mine convey the same message and therefore are reproductions of each other. A visualization is repeatable, on the other hand, if it is possible to recreate the exact same visual appearance, down to the last pixel, from the raw data. Strictly speaking, repeatability requires that even if there are random elements in the figure, such as jitter (Chapter 18), those elements were specified in a repeatable way and can be regenerated at a future date. For random data, repeatability generally requires that we specify a particular random number generator for which we set and record a seed.

Throughout this book, we have seen many examples of figures that reproduce but don't repeat other figures. For example, Chapter 25 shows several sets of figures that each show the same data but that look somewhat different. Similarly, Figure 28-1a is a repeat of Figure 9-7, down to the random jitter that was applied to each data point, whereas Figure 28-1b is only a reproduction of that figure. Figure 28-1b has different jitter than Figure 9-7, and it also uses a sufficiently different visual design that the two figures look quite distinct, even if they convey the same information about the data.

DO NOT COPY
Angela Zoss <azoss@placeholder.18157.edu>

Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

design that the two figures look quite distinct, even if they convey the same information about the data.

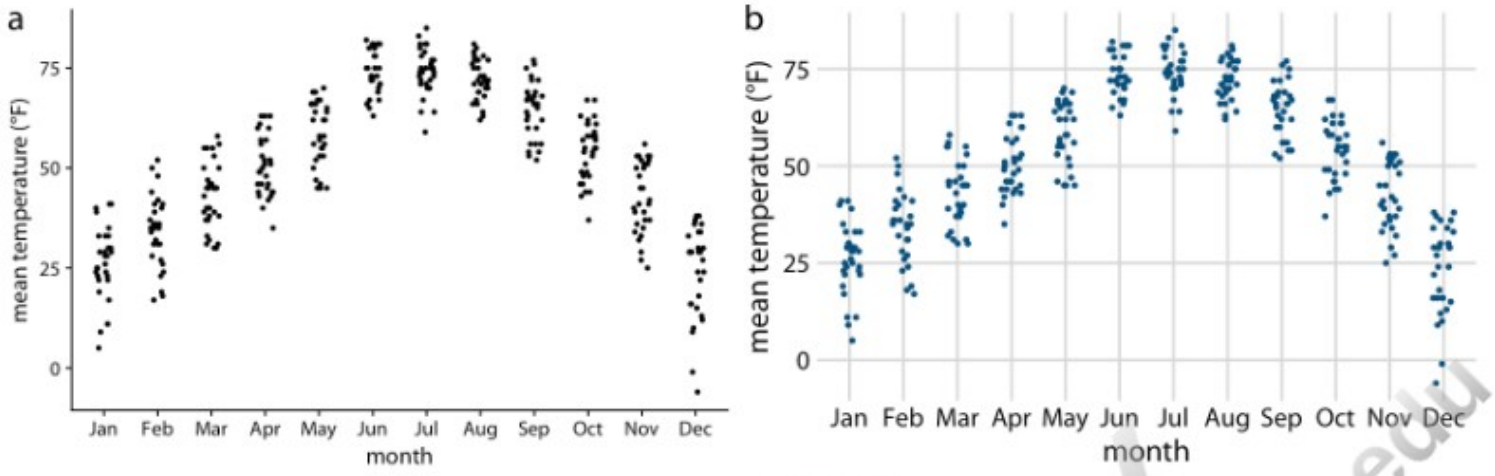


Figure 28-1. Repeat and reproduction of a figure. Part (a) is a repeat of Figure 9-7. The two figures are identical down to the random jitter that was applied to each point. By contrast, part (b) is a reproduction but not a repeat. In particular, the jitter in part (b) differs from the jitter in part (a) or in Figure 9-7. Data source: Weather Underground.

Both reproducibility and repeatability can be difficult to achieve when we're working with interactive plotting software. Many interactive programs allow you to transform or otherwise manipulate the data but don't keep track of every individual data transformation you perform, only of the final product. If you make a figure using this kind of program, and then somebody asks you to reproduce the figure or create a similar one with a different dataset, you might have difficulty doing so. During my years as a postdoc and a young assistant professor, I used an interactive program for all my scientific visualizations, and this exact issue came up several times. For example, I had made several figures for a scientific manuscript. When I wanted to revise the manuscript a few months later and needed to reproduce a slightly altered version of one of the figures, I realized that I wasn't quite sure anymore how I had made the original figure in the first place. This experience has taught me to stay away from interactive programs as much as possible.

DO NOT COPY
434a511a93f9a58@placeholder.edu

Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

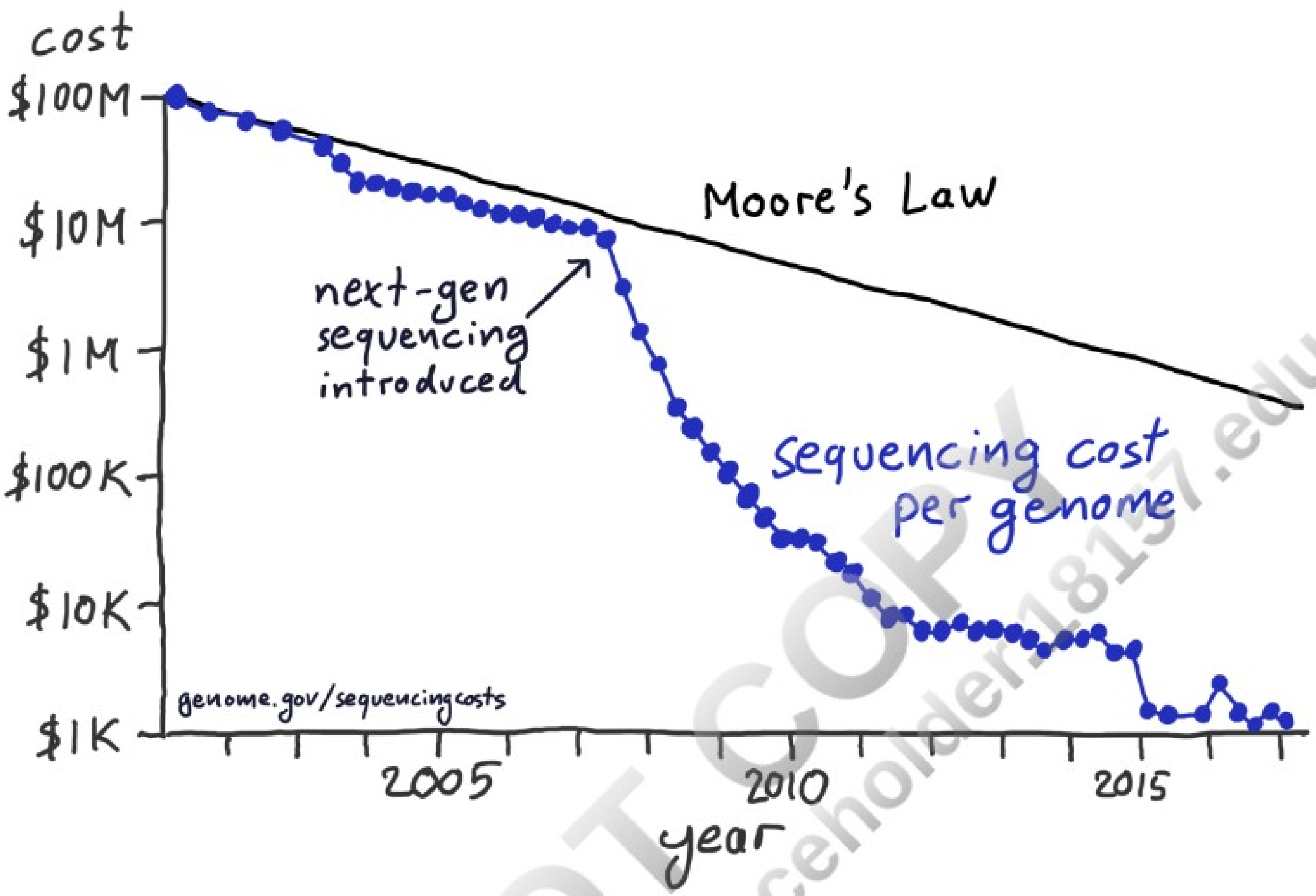


Figure 28-2. After the introduction of next-gen sequencing methods, the sequencing cost per genome has declined much more rapidly than predicted by Moore's law. This hand-drawn figure reproduces a widely publicized visualization prepared by the National Institutes of Health. Data source: National Human Genome Research Institute.

Separation of Content and Design

Good visualization software should allow you to think separately about the content and the design of your figures. By content, I refer to the specific dataset shown, the data transformations applied (if any), the specific mappings from data onto aesthetics, the scales, the axis ranges, and the type of plot (scatterplot, line plot, bar plot, boxplot, etc.). Design, on the other hand, describes features such as the foreground and background colors, font specifications (e.g., font size, face, and family), symbol shapes and sizes, whether or not the figure has a background grid, and the placement of legends, axis ticks, axis titles, and plot titles. When I work on a new visualization, I usually determine first what the content should be, using the kind of rapid exploration described in the previous section. Once the content is set, I may tweak the design, or more likely I will apply a predefined design that I like and/or that gives the figure a consistent look in the context of a larger body of work.

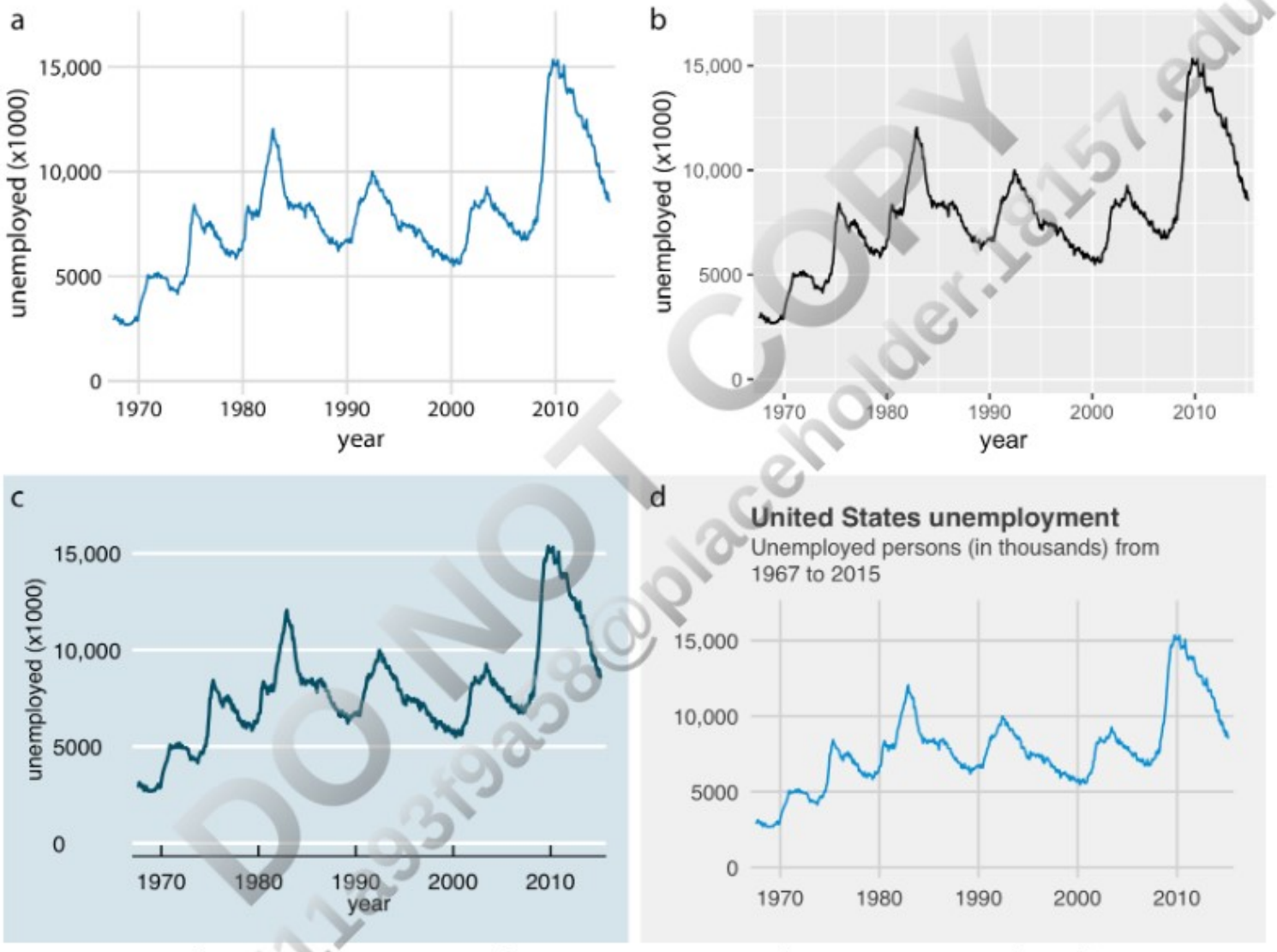
In the software I have used for this book, ggplot2, separation of content and design is achieved via *themes*. A theme specifies the visual appearance of a figure, and it is easy to take an existing figure and apply different themes to it (Figure 28-3). Themes can be written by third parties and distributed as R packages. Through this mechanism, a thriving ecosystem of add-on themes has developed around ggplot2, and it covers a wide range of different styles and application scenarios. If you're making figures with ggplot2, you can almost certainly find an existing theme that satisfies your design needs.

Separation of content and design allows data scientists and designers to each focus on what they do best. Most data scientists are not designers, and therefore their primary concern should be the data, not the design of a visualization. Likewise, most designers are not data scientists, and they should be able to provide a unique and appealing visual language for figures without having to worry about specific data, appropriate transformations, and so on. The same principle of separating content and design has long been followed in the publishing world of books, magazines, newspapers, and websites, where writers provide content but layout and design are handled by a separate group of people who specialize in this area and who ensure that the

Printing is for personal, private use only. No part of this book may be reproduced or transmitted without publisher's prior permission. Violators will be prosecuted.

content but layout and design are handled by a separate group of people who specialize in this area and who ensure that the publication appears in a visually consistent and appealing style. This principle is logical and useful, but it is not yet that widespread in the data visualization world.

In summary, when choosing your visualization software, think about how easily you can reproduce figures and redo them with updated or otherwise changed datasets, whether you can rapidly explore different visualizations of the same data, and to what extent you can tweak the visual design separately from generating the figure content. Depending on your skill level and comfort with programming, it may be beneficial to use different visualization tools at the data exploration and data presentation stages, and you may prefer to do the final visual tweaking interactively or by hand. If you have to make figures interactively, in particular with software that does not keep track of all the data transformations and visual tweaks you have applied, consider taking careful notes on how you make each figure, so that all your work remains reproducible.



DO NOT COPY
4344[illegible]1931f9a58@placeholder.18157.edu

Figure 28-3. Number of unemployed persons in the US from 1970 to 2015. The same figure is displayed using four different ggplot2 themes: (a) the default theme for this book; (b) the default theme of ggplot2, the plotting software I have used to make all the figures in this book; (c) a theme that mimics visualizations shown in the Economist; (d) a theme that mimics visualizations shown by FiveThirtyEight. FiveThirtyEight often foregoes axis labels in favor of plot titles and subtitles, and therefore I have adjusted the figure accordingly. Data source: US Bureau of Labor Statistics.