Ideas in Ecology and Evolution 10: 1–5, 2017
doi:10.4033/iee.2017.10.1.e

Received 13 February 2017; Accepted 14 February 2017

Editorial

Open Sesame: R for Data Science is Open Science

Christopher J. Lortie

Christopher J. Lortie (lortie@yorku.ca), Department of Biology, York University, 4700 Keele St. Toronto ON Canada. M3J1P3 and *The National Center for Ecological Analysis and Synthesis (NCEAS). UCSB. Suite 300, 735 State St., Santa Barbara, CA. 93101*

Data science context

Data science is a critical component of many domains of research, including the domain within which I primarily function, ecology. However, in teaching biostatistics within the university context, we typically focus on the statistics, and less on the science of data (i.e. handling, understanding, and manipulating data) (Dhar 2013). This is unfortunate; however, the teaching landscape is now rapidly evolving to include offerings of numerous institutional Master's of Data Science degrees (a comprehensive list of offerings is provided here: http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/). Failure to appreciate the differences between data science and statistics is common, and it is an easy distinction to blur because the disciplines support one another. Most data science and statistical processes are iterative. Typically, there is feedback from the handling and manipulating of data to the formal analyses or model building, depending on the outcomes and insights provided by the data visualization and exploratory analyses. University-level science teaching is steadily embracing open science and shares many of the skills that students need to be scientifically-literate citizens. However, data-literate citizens are important, too, if we want the next generation to make informed, evidence-based decisions about human health, the economy, and the health of our ecosystems. We need to understand data directly—its potential value as an evidence asset, but also its challenges. Critical thinking tools for data are non-trivial concepts, and statistics are absolutely needed. However, the science of data, big or little, is critical in appreciating the decisions, steps, and workflow needed to prepare, share, analyze, collaborate, and evaluate quantitative and qualitative data. This should be transparent and reinforced in both teaching and communicating science. There are extensive resources online including blogs, training courses, and webinars that directly and indirectly discuss the value of data science for other domains of science. There are also many books to this effect that both appreciate and communicate the value of data science thinking and improve the skill set that one can apply to teaching, research, and collaboration in general (kaggle provides an excellent list of resources here: https://www.kaggle.com/wiki/DataScienceBooksAndCourses). Recently, I completed my latest adventure, 'R for Data Science,' which describes several sets of packages for R, and general workflows for tidy data science (Grolemund and Wickham 2016). Few sets of data science packages are so cleanly aligned and bundled in philosophy and implementation. This book thus serves as an excellent opportunity to illustrate how the importance of reproducible, transparent data science leads to similar thinking and communication.

R for Data Science book

The book was written in R Markdown, compiled using bookdown, and it is free online at http://r4ds.had.co.nz. Appropriately, it thus embodies both open science and data science in how it is written. Open science is a movement that promotes sharing the *process* and not just the final product of evidence-based inquiry common in science (Wolkovich, Regetz, and O'Connor 2012, Michener and Jones 2012). Data science includes wrangling, handling, and formatting data for subsequent analyses and visualization (Dhar

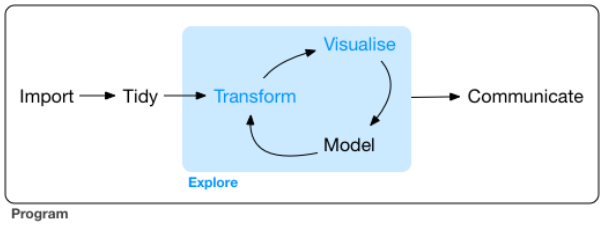
2013, Hardin et al 2015, Baumer 2015). The decisions involved in data science are important and frame the conclusions and implications for a particular body of evidence. The key elements often include transparency, reproducibility, and access. Bookdown is an R package that knits a set of R Markdown files together into a book using RStudio (Xie 2017). The capacity to publish a book using the same tools that one can use to wrangle, visualize, and analyze data is a profound innovation. Access to books published using these tools, then shared on GitHub, is a *new form of publishing*. R for Data Science was written and shared using this workflow. Reading a book written using one of the most powerful open science/data science tools, i.e. R (language and environment) implemented through RStudio, and being able to access it online with code, you immediately appreciate the trickle effects of 'open data science' thinking on writing, collaboration, and even professional publishing. This is all incredible, and it is a peek into a very different future of scholarly communication and publishing. The book was officially published December 12th, 2016 in print form. However, because it was available online as a knitted book, it could be read as it developed and evolved, and, the authors capitalized on the versioning workflow, transparency, and collaborative philosophy of many data scientists, while soliciting specific feedback on sections in writing. Again, the development of this book was a fantastic example of a novel scholarly communication approach to writing and sharing. I read what was available in real time. It confirmed and advanced my understanding and skill set for data science immensely, and provided an obvious open science/data science implication for publishing: that a book can continue to evolve and version within this context, and that this process can be facile with tools such as GitHub and RStudio. Here is a brief summary, without spoilers, of some of the dimensions I used to conclude that this book is an excellent example of fantastic data science 'beasts and where to find them.'

Language & clarity

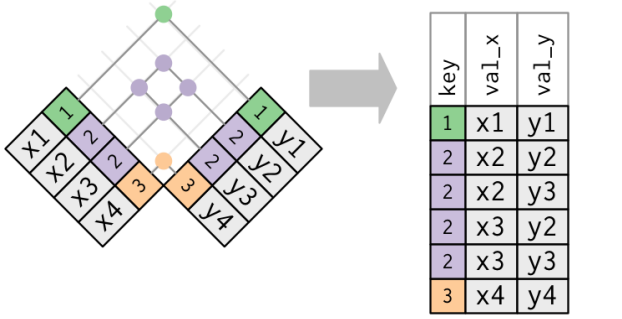
In reading R statistics, statistics (in general), or data science books, one expects/hopes that—like literate coding (Knuth 1984, 1992)—the prose will be accessible, pleasant, and appropriately pitched. This book was ideal in this respect. It was more formal than conversational, but not too technical. The structure facilitated reading and comprehension because it was clear and logical. The visual elements added a dimension of attractive clarity to the writing that was not just code, prose, R, or data visualizations. Many of the visuals were also excellent heuristics. Some were a reminder to the reader of the big picture in data science while others highlighted a workflow or approach. In discussing data science and specifically, its tools, it is sometimes easy to get too detailed, too rapidly, and get 'lost in the weeds.' The tension between big picture and sufficient detail is well navigated here in writing style, and the graphics facilitated a higher-level of understanding from code to explanatory text to conceptualization within a figure. This mapping process of ideas emulates one specific package, ggplot2 (Wickham 2009), which invokes the grammar of graphics and proposes that mapping data is an integral process to visualization. Here, many aspects of data science are mapped onto conceptual explanations. Code and package use change over time, but learning data science concepts promotes both more extended retention and a deeper level of understanding. The writing of this book sets the reader up for better future thinking of how to use R and how to code more effectively in general. This is like many excellent offerings that provide insights into efficiency in R (Gillespie and Lovelace 2017), but R for Data Science is focused more on the specifics of certain packages, and less on general coding practices in R. These figures (Figure 1) were, for instance, particularly useful herein for illustrating the purpose of specific functions and the respective outcomes. I could have even used more of these general visual guides here and there, but in digging into the examples, I recognize that they were likely not always needed (and too much can be a bad thing too if poorly executed). The clarity was very high in almost every chapter of the book. I struggled with some of the more complex chapters (for me) such as relational data, or some elements of the model building, but the flow of the writing and logic kept me rolling through these even if some of the details remained elusive.

The expectation that data science or statistics books should be only read once is a challenging notion. Many of the chapters in this book certainly satisfy that criterion, but it depends on the purpose. Some of the more challenging chapters that you identify can be re-read for better comprehension and one could also follow along/experiment with in RStudio. Sometimes, it is nonetheless good to get the message from alternate sources described or explained a little differently. In recently exploring a wide range of data science books in R, some of the contemporary offerings will not be revisited. My feeling for 'R for Data Science' is that the clean style and direct writing do not conflate the message, and, re-reads would likely be beneficial when needed for reframing general approaches, teaching, or re-examining a specific workflow for some of the packages described. The message in many chapters is also unique, and even a brief revisit would highlight some of the handling elements and assumptions associated with best practices for data science.

(a)



(b)



(c)

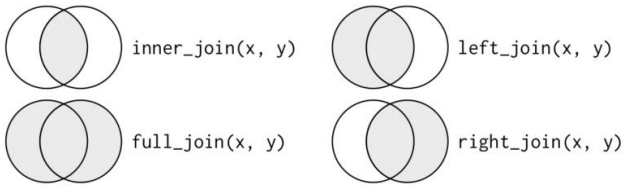


Figure 1. A sample of the conceptual figures from 'R for Data Science' (Grolemund and Wickham 2016). The first instance (a) visualizes the general data science workflow proposed throughout the book and served as a consistent visual guide for the reader framing the purpose of each chapter. The second instance provided (b) illustrated the general process of joining different dataframes that is then specifically mapped out using the package dplyr with the key functions depicted using Venn diagrams with simplified code (c). These images (available at http://r4ds.had.co.nz) are © 2016 Grolemund and Wickham, licensed under the Creative Commons license, CC BY-NC-NC 3.0 US.

Philosophy

Welcome to the tidyverse. Enough said to all who follow and read up within the R community (https://twitter.com/hashtag/tidyverse). This universe is logical and feels natural. Tidy data are data frames with a specific structure; namely, variables are organized by columns, observations by rows, and each cell is a value (Wickham 2014, Grolemund and Wickham 2016). The packages are designed to take advantage of this consistent structure both in terms of ease of writing and reading code. The forthcoming ggvis will help further align the grammar and semantics that parallel the code and flow with pipes versus '+' of ggplot2. Tibbles are a pleasant surprise. The wrangle readings satisfy. Tidiness is next to high orderedness. Subscribing to the philosophy of readable code, consistent data structures, and logical workflows will promote better open science and reproducibility. This is never explicitly stated, or, if it was, I missed it. I suspect that this is a good thing. We can approach open science, open data, and more transparency in science from top-down or bottom-up efforts. By not repeatedly banging that drum per se, but directly providing and describing the tools to handle data cleanly and consistently, this book provides a solid bottom-up pillar for the open science movement. Tidy data and readable code are shareable *and* useable. Finally—and aligned with this tools-first approach—the value of models and epistemology of hypotheses are stated later in the book (Chapter 19). This organization worked for me as reader of this book, but likely not when teaching students. I like the hypothesis/model philosophy of 'knowing data' developed here. It was big data in origins, balanced, and emphasized bias and non-independence in exploring and testing models. What you can learn from a model also depends on how it is applied. This was well described. Split. Build. Think. Test. Know. Your own personal variation would likely fit within a similar framework even with little data. I did wonder a bit how I could adapt some of the model fitting ideas to more of the little data common in some the ecological inquiries. Solutions can include the following approaches: (i) pilot field experiments can provide the training data, and (ii) resampling/ bootstrapping using modelr to populate larger datasets for more independent EDA. The reminder to avoid repetition is repeated. Not ironically.

Skills

Many books do not need to adapt. Most R statistics books likely do. Packages are often a game changer. Grammar changes. Base R is a must know, of course (Helmreich 2016), but streamlining and specifics often live in the libraries that the community develops. This book is available for sale on Amazon.com, and I assume the version will adapt, but more slowly than the bookdown version. The frame-rate of change in no way precludes reading the book now or revisiting it at some later point in time. Model building, wrangling basics, functions, and iterations chapters are solid reading that provide a skill set needed right now. The data visualization and perhaps data transformation chapters are most likely to change soon. Read now and capture those skills, but expect change. There are also some nice examples of intermediate to advanced tricks in plotting that reading now will provide. Certainly, this is the case

in the iteration and model chapters too—good intermediate-skill building blocks for advanced coding data science. This skill set is pretty darn awesome (PDA), and the strings chapter was also very rich in new skills, as well as a launchpad to text mining with other packages (it inspired me to try it right after I finished reading the book). Skills abound, and the context of application is described appropriately.

The bottom line (of code) review for most readers

```
> high.returns <- c("basic.R.users",
"intermediate.R.users")

> tidy.data.science <- philosophy of
consistent structures %>% visualize with
models %>% share
```

Implications and conclusions

There are many tools for open science (data management plans, slide-sharing repositories, data repositories, GitHub, preprints, sharing meta-data, social media, blogs, and data publications to name a few). However, effective data science in R can also be a powerful ally if you include the final steps of communication described in this book (Chapters 21–23). For instance, sharing code via different formats is well described, and now simple with a few packages and RStudio. Communication of code is promoted and articulated clearly in the final chapters. The best data science is certainly open science, and open science should not be a static one-way process. The capacity for RStudio connected to GitHub to allow versioning is thus an immediate mechanism to make use of R as a bridge between data science and open science (Figure 2). Code review, workflows, and versioning are not a new set of concepts to programmers. However, the target audience of this book is likely very broad, including domain-level scientists like myself, and the description of these ideas in the context of communication and collaboration was very compelling. In summary, both the proximate elements of the tools described, and the concepts needed to deploy and use them, are well framed in this book. Reading the book from a 'how-to-do what I need to do' perspective, concurrent with a 'how can this book advance many domains of science that use data science within their process of discovery' approach, worked well. Shifting between the how and why was possible because of the general organization of the book. Other readers can decide whether this larger story-arc worked for them in this specific book, but I am hopeful we can all agree that sharing the process of handling data to make decisions, statistical and otherwise, is critical now more than ever. We now have at least one magical phrase to make a difference in sharing the process of these aspects of science. Open Sesame.

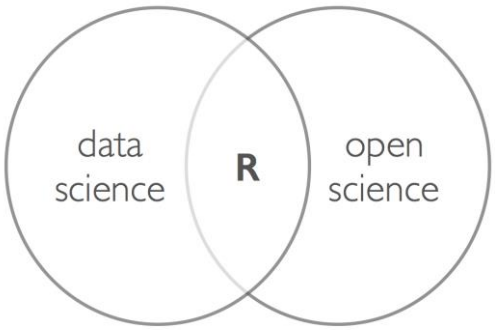


Figure 2. The coding language and environment R can be a literal bridge between data science and open science. In the book 'R for Data Science', the capacity for R as foundation to incorporate transparency, reproducibility, and access into processes associated with data science are well described and illuminate capacities for novel, open scientific communication. This figure is freely available for reuse in presentations and other formats, and can be cited and downloaded at https://figshare.com/articles/Data_science_and_open_science_Venn_Diagram/3652539. To cite the figure directly, please use the format: Lortie, Christopher (2016): Data science and open science Venn Diagram. figshare.https://doi.org/10.6084/m9.figshare.3652539.v1

References

Baumer, B.A. 2015. Data science course for undergraduates: Thinking with data. The American Statistician 69: 334-342. CrossRef

Dhar, V. 2013. Data science and prediction. Communications of the ACM 56: 64. CrossRef

Gillespie, C and R. Lovelace. 2017. Efficient R programming. O'Reilly Media.

Grolemund, G. and H. Wickham. 2016. R for Data Science, Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media.

Grolemund, G. and H. Wickham. 2016. R for Data Science. O'Reilly Media.

Hardin, J. Hoerl, R., Horton, N.J., and D. Nolan. 2015. Data science in statistics curricula: Preparing students to "Think with data". The American Statistician 69: 343-353. CrossRef

Helmreich, J.E. 2016. Learning Base R. Journal of Statistical Software 69, Book Review 4. CrossRef

Knuth, D.E. 1984. Literate Programming. The Computer Journal 27: 97-111.

Knuth, D.E. 1992. Literate Programming. University of Chicago Press.

Michener, W.K. and M.B. Jones. 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27: 85-93. CrossRef
Wickham, H. 2009. ggplot2. Springer.
Wickham, H. 2014. Tidy data. Journal of Statistical Software 59: 1-23.
Wolkovich, E.M., Regetz, J. and M.I. O'Connor. 2012. Advances in global change research require open science by individual researchers. Global Change Biology 18: 2102-2110. CrossRef
Xie, Y. 2017. Bookdown: authoring books and technical documents with R Markdown. CRC Press.