The Journal of Open Source Education

Data Carpentry for Biologists: A semester long Data Carpentry course using ecological and other biological examples

Ethan P. White[1, 2, 3], **Zachary T. Brym**[4], **Andrew J. Marx**[1], **Kristina Riemer**[5], **Sergio Marconi**[6], **David J. Harris**[1], **Virnaliz Cruz**[7], **and S. K. Morgan Ernest**[1, 2]

1 Department of Wildlife Ecology and Conservation, University of Florida **2** Biodiversity Institute, University of Florida **3** Informatics Institute, University of Florida **4** Tropical Research and Education Center, Department of Agronomy, University of Florida **5** College of Agriculture and Life Sciences, University of Arizona **6** School of Forest Resources & Conservation, University of Florida **7** School of Natural Resources and Environment, University of Florida

DOI: 10.21105/jose.00139

Software

- Review
- Repository
- Archive

Submitted: 26 August 2021
Published: 15 April 2022

Summary

Data Carpentry for Biologists is a semester-long course in best practices for storing, loading, manipulating, and visualizing data using R. The course material includes video demonstrations, lecture notes for live coding demonstrations, links to openly available reference readings, coding practice exercises, and the output expected from completed exercises. The course is structured in topics that combine sets of learning materials covering a week of college level material on a single subject. The lessons and exercises focus on biological examples with a particular focus on ecological examples. The course material is designed to be used in two ways. First, it can be used in a self-paced online format for individual learners. This is achieved by having all of the necessary material to understand and complete the course present on the website along with instructions for self-guided learning. Second, the course is designed to be modified and remixed to be taught in college and university classrooms. This is achieved by a modular design that allows modifying all aspects of the course and by detailed documentation for course customization. The website is viewed by thousands of users each month and the material and infrastructure has been used in courses at multiple colleges and universities.

Statement of Need

Being able to work within a computational environment to store, load, manipulate, analyze, and visualize data has become a key component of many areas of biology (Hampton et al., 2017; Jones et al., 2006). Despite this, many biologists lack access to formal training in computation and are expected to pick up these skills on their own (Williams et al., 2019). When training is available it is often not focused on a particular research domain, creating significant barriers to novices learning these important skills (Teal et al., 2015). This lack of training has significant costs for the progress of biology because researchers spend more time learning the necessary skills and often develop less optimal approaches to solving problems (Teal et al., 2015). As a result, surveys have identified a major need for more computational training among biologists (Barone et al., 2017). Part of the reason for a lack of domain specific courses at colleges and universities is a lack of faculty with either the time or expertise to develop a new course in scientific computing. For example, a recent survey identified lack of expertise, training, and time as important barriers to faculty for developing undergraduate training in bioinformatics (Williams et al., 2019).

The 'Data Carpentry for Biologists course' is designed to help overcome a variety of these training impediments. The materials can be used as is or modified by instructors developing a scientific computing course for either in-person or online courses. The website is also designed to be used by students as an independent self-guided course.

Features

General instructional design

The course is focused on teaching practical computational skills in a domain specific context that is directly applicable to ecologists and biologists more broadly. Many general computing courses start with the foundations of computer programming, but this can be demotivating for domain specialists who are learning computing to accomplish specific data management and analysis tasks. In addition, some learning theories, such as Constructivism, suggest that students learn by incorporating new ideas into their existing knowledge (Bada & Olusegun, 2015), thus general computing courses could also make learning inherently more difficult for students. Therefore this course follows the broader Data Carpentry philosophy of focusing on teaching the skills students need using familiar data and common computational challenges encountered within their scientific domains (Teal et al., 2015). To accomplish this the course uses a number of real ecological datasets and the coding demonstrations and exercises focus on common tasks in the analysis of biological data.

The course is built around the "I do, we do, you do" approach to teaching where first the instructor demonstrates how to do something, then the students work on an example with the instructor present to help and answer questions, and finally the students work on additional examples independently. This approach is based on explicit instruction principles, which leverage the benefits of active-learning without leaving students who are less comfortable with the material feeling lost (Archer & Hughes, 2010; Rosenshine, 1987). This approach, described as "a systematic method of teaching with emphasis on proceeding in small steps, checking for student understanding, and achieving active and successful participation by all students" (Rosenshine, 1987, p. 34), is useful for teaching introductory computing to scientists because it gradually builds comfort and competence for all students in this essential foundation of research.

A standard lesson starts with a brief introduction to the concept being taught and why it is important. This is followed in the classroom by a live coding demonstration of the first piece of the material. During and following the demonstration students have the opportunity to ask questions. The class is then assigned a short exercise designed to reinforce and check understanding on the material. While working on the exercise students ask additional questions and the instructor looks for students who are struggling with the concept and engages them to help work through whatever challenge they are facing. The next small chunk of material is then presented. Additional exercises are provided for the students to work on outside of class to further reinforce the material and help the students engage with more complex and integrated approaches to what they are learning.

Self-guided online learning

All course material is available online at https://datacarpentry.org/semester-biology and the site includes instructions for self-guided learning. All of the video coding demonstrations are available on a dedicated YouTube playlist. To mimic explicit instruction as much as possible in an online format, online lessons are broken down into a series of short videos and exercises, just like in a classroom environment. Students watch a short video and then are provided an exercise that reinforces and checks the understanding of material in that video. An entire lesson is made up of a series of videos and exercises followed by a link to the remaining exercises for that lesson.

In order to allow self-guided students to check to make sure they understand the exercises, the expected output for each exercise is included on the website. This allows the learner to self-evaluate and troubleshoot their solution to the exercise, while still keeping the code solutions private for use in grading in traditional classroom settings. This is also beneficial to students using this material in traditional courses because it helps remove any uncertainty about exactly what each exercise expects the learner to do.

Basis for other college and university courses

The course is designed to be modified and remixed to meet the needs of the particular instructor and students and allow using the material broadly across college and university classrooms. This allows faculty who have enough expertise to teach an introductory R course, but who lack either the time or expertise to develop the course, to be able to teach this course at their institutions. To facilitate this, the course site has a modular design that allows modifying and remixing exercises, assignments, readings, and coding lessons, along with detailed documentation for doing so (https://datacarpentry.org/semester-biology/docs/).

Access to a separate private repository that includes the code solutions for all of the exercises is available to instructors at colleges and universities by opening an issue in the main course repository (https://github.com/datacarpentry/semester-biology/). This allows the code solutions to the exercises to be shared among instructors while still enabling the use of the exercises for summative evaluation purposes.

The course material and infrastructure is and has been used for a number of college and university courses including:

- Data Science for Biologists at Virginia Commonwealth University
- Data Science for Agriculture at Oklahoma State University
- Data Visualization for Plant Pathologists at the University of Florida
- Data Science for SAFS at the University of Washington
- Data Carpentry for Pharmacists at the University of Health Sciences and Pharmacy in St. Louis
- R Programming for Biologists at Stonehill College
- Data Carpentry for Ecologists at the University of Georgia
- Introduction to Data Analysis for Aquatic Sciences at the University of Washington
- Data Science in Omics Introduction at Oklahoma State University
- Ecoinformatics at Kenyon College
- Data Management for Biologists at the University of Minnesota
- Introducing Agroecology: The Basics of Agroecology for Practitioners at the University of Florida
- Data Science with R

Course Infrastructure

The course website is built using the Jekyll static site generator with a customized version of the Lanyon theme. Development is conducted in the course's GitHub repository and the site is automatically deployed from this repository using GitHub Pages.

The modular design of the course is supported by "assignments" that automatically combine "readings" and "lecture notes" on a topic with "exercises" and associated point values provided by YAML lists that indicate the exercises to include and the point values to be assigned to them. The course schedule is automatically generated based on selected assignments, also provided in a YAML list. This allows both exercises and assignments to be reused and remixed by changing the selections in YAML lists.

The online materials for the course are designed to be accessible to all learners. The site has been reviewed using the WAVE web accessibility evaluation tool and pa11y, and all videos have been manually captioned by the instructor presenting the material. In order to ensure that accessibility is maintained as the site is continuously updated, we use pa11y-ci and continuous integration (GitHub Actions) to automatically scan all pull requests for accessibility.

This general infrastructure for modular, collaborative, and accessible course development is useful regardless of the content of the course. For example, one of the authors (Zachary Brym) uses the same infrastructure for extension programs on agroecology at the University of Florida. We hope that open courses on other topics can be built on this infrastructure to support the broader adoption of collaborative course development for college and university classrooms.

Course Development Background

The initial development of this material and infrastructure was started in 2008 by Ethan White (with help and advice from Morgan Ernest) at Utah State University as part of an NSF CAREER award to develop a Programming for Biologists course to address the need for more computational training in biology. It was then converted to R and the current Data Carpentry for Biologists course structure by Zachary Brym and Ethan White for use at the University of Florida. While being taught at the University of Florida, the TA for the course, Andrew Marx, contributed substantially to improving the materials, and a number of members of White and Ernest's Weecology Research Group (including Kristina Riemer, Sergio Marconi, David Harris, and Virnaliz Cruz) contributed new material and improved existing material. Weecology is an interdisciplinary research group that conducts research at the intersection of computation, technology, and ecology, with a long history of involvement in computational training including founding, leadership, maintainer, and instructor roles within The Carpentries. The authors of this material have extensive background in both computational biology and teaching computational tools in workshop and classroom environments.

Acknowledgements

The development of this material was supported by the University of Florida, the National Science Foundation through CAREER award 0953694 to E.P. White, and the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through grant GBMF4563 to E.P. White. Thanks to Tracy Teal and The Carpentries for supporting the development and Data Carpentry branding of this course and associated material. Thanks to Claire Williams for her exceptional assistance in course management at the University of Florida and Eric Hellgren for supporting the open development of the course. A huge thanks to the many students who have taken this and earlier versions of this course for their enthusiasm for the material and essential feedback that dramatically improved how the course is structured and taught.

References

Archer, A. L., & Hughes, C. A. (2010). *Explicit instruction: Effective and efficient teaching*. Guilford Publications.

Bada, S. O., & Olusegun, S. (2015). Constructivism learning theory: A paradigm for teaching and learning. *Journal of Research & Method in Education*, *5*(6), 66–70.

Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Computational Biology*, *13*(10), e1005755. https://doi.org/10.1371/journal.pcbi.1005755

Hampton, S. E., Jones, M. B., Wasser, L. A., Schildhauer, M. P., Supp, S. R., Brun, J., Hernandez, R. R., Boettiger, C., Collins, S. L., Gross, L. J., & others. (2017). Skills and knowledge for data-intensive environmental research. *BioScience*, *67*(6), 546–557. https://doi.org/10.1093/biosci/bix025

Jones, M. B., Schildhauer, M. P., Reichman, O., & Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annu. Rev. Ecol. Evol. Syst.*, *37*, 519–544. https://doi.org/10.1146/annurev.ecolsys.37.091305.110031

Rosenshine, B. (1987). Explicit teaching and teacher training. *Journal of Teacher Education*, *38*(3), 34–36. https://doi.org/10.1177/002248718703800308

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., & Pawlik, A. (2015). Data carpentry: Workshops to increase data literacy for researchers. *International Journal of Digital Curation*, *10*, 135–143. https://doi.org/10.2218/ijdc.v10i1.351

Williams, J. J., Drew, J. C., Galindo-Gonzalez, S., Robic, S., Dinsdale, E., Morgan, W. R., Triplett, E. W., Burnette III, J. M., Donovan, S. S., Fowlks, E. R., & others. (2019). Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. *PloS One*, *14*(11), e0224288. https://doi.org/10.1371/journal.pone.0224288