Data Science Resources

Doug Joubert

Invalid Date

Table of Contents

# Data Science Resources

## Data and Analytics for Research Training Program

Resource [URL](https://arcus.github.io/education_modules/educators)

Educational materials created on this site. You can tailor which modules you give to you learners based on their needs, or adapt them to fit your purposes. Please read the section [Remixing Our Materials](https://arcus.github.io/education_modules/educators#remixing-our-materials) to learn about our Creative Commons Attribution-ShareAlike license and suggestions of ways to use this curriculum. I have included some specific lessons.

## Data analysis for the life sciences with r

Irizarry R. A. & Love M. I. (2017). *Data analysis for the life sciences with r*. CRC Press Taylor & Francis Group.

This title covers several of the statistical concepts and data analytic skills needed to succeed in data-driven life science research. The authors proceed from relatively basic concepts related to computed p-values to advanced topics related to analyzing high-throughput data. They include the R code that performs this analysis and connect the lines of code to the statistical and mathematical concepts explained.

## Elements of Data Analytics Style

Leek, J. (2015). The Elements of Data Analytic Style. Leanpub. https://leanpub.com/datastyle

This book is focused on the details of data analysis that sometimes fall through the cracks in traditional statistics classes and textbooks. It is based in part on the authors blog posts, lecture materials, and tutorials such as:

* [10 things statistics taught us about big data analysis](http://simplystatistics.org/2014/05/22/10-things-statistics-taught-us-about-big-data-analysis/)
* [The Leek Group Guide to R packages](https://github.com/jtleek/rpackages)
* [How to share data with a statistician](https://github.com/jtleek/datasharing)

## Hitchhiker’s Guide to Data Science for Social Good

Author: Contributors listed on GitHub

Resource:

GitHub [URL](https://github.com/dssg/hitchhikers-guide.git)

The primary audience for this guide is the set of fellows coming to DSSG but we want everything we create to be open and accessible to larger world. We hope this is useful to people beyond the summer fellows coming to DSSG.

**If you are applying to the program or have been accepted as a fellow,** [check out the manual](https://github.com/dssg/hitchhikers-guide/blob/master/dssg-manual) to see how you can prepare before arriving, what orientation and training will cover, and what to expect from the summer.

**If you are interested in learning at home,** check out the [tutorials and teach-outs](https://github.com/dssg/hitchhikers-guide/blob/master/curriculum) developed by our staff and fellows throughout the summer, and to suggest or contribute additional resources.

\*Another one of our goals is to encourage collaborations. Anyone interested in doing this type of work, or starting a DSSG program, to build on what we’ve learned by **using and contributing to** these resources.

## R for Data Science

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). R for data science: Import, tidy, transform, visualize, and model data (2nd edition). O’Reilly Media, Inc.

Use R to turn data into insight, knowledge, and understanding. With this practical book, aspiring data scientists will learn how to do data science with R and RStudio, along with the tidyverse–a collection of R packages designed to work together to make data science fast, fluent, and fun. Even if you have no programming experience, this updated edition will have you doing data science quickly. You’ll learn how to import, transform, and visualize your data and communicate the results. And you’ll get a complete, big-picture understanding of the data science cycle and the basic tools you need to manage the details. Updated for the latest tidyverse features and best practices, new chapters show you how to get data from spreadsheets, databases, and websites. Exercises help you practice what you’ve learned along the way. The preprint version of the book is free and available [online](https://r4ds.hadley.nz/).

## R in Action

Kabacoff, R. I. (2022). R in action: Data analysis and graphics with R and Tidyverse (Third edition). Manning Publications Co.

R is the most powerful tool you can use for statistical analysis. This definitive guide smooths R’s steep learning curve with practical solutions and real-world applications for commercial environments. R in Action, Third Edition makes learning R quick and easy. That’s why thousands of data scientists have chosen this guide to help them master the powerful language. Far from being a dry academic tome, every example you’ll encounter in this book is relevant to scientific and business developers, and helps you solve common data challenges. R expert Rob Kabacoff takes you on a crash course in statistics, from dealing with messy and incomplete data to creating stunning visualizations. This revised and expanded third edition contains fresh coverage of the new tidyverse approach to data analysis and R’s state-of-the-art graphing capabilities with the ggplot2 package.

## Tiny Modeling with R: a framework for modeling in the tidyverse

Kuhn M. & Silge J. (2022).*Tidy modeling with r : a framework for modeling in the tidyverse*. O’Reilly Media.

Get going with tidymodels, a collection of R packages for modeling and machine learning. Whether you’re just starting out or have years of experience with modeling, this practical introduction shows data analysts, business analysts, and data scientists how the tidymodels framework offers a consistent, flexible approach for your work. RStudio engineers Max Kuhn and Julia Silge demonstrate ways to create models by focusing on an R dialect called the tidyverse. Software that adopts tidyverse principles shares both a high-level design philosophy and low-level grammar and data structures, so learning one piece of the ecosystem makes it easier to learn the next. You’ll understand why the tidymodels framework has been built to be used by a broad range of people.

## Tidyverse Skills for Data Science

Wright, C., Ellis, S., Hicks, S., & Peng, R. (2021). *Tidyverse Skills for Data Science in R*. Leanpub. <http://leanpub.com/tidyverseskillsdatascience>

This book sets out to introduce the conceptual framework behind tidy data and introduce the tidyverse and tidyverse-adjacent packages. Mastery of these fundamental concepts and familiarity with what can be accomplished using the tidyverse will be critical throughout the more technical Resources ahead. So, be sure you are familiar with the vocabulary provided and have a clear understanding of the tidy data principles introduced here before moving forward.

## Welcome to Text Mining with R

Author: Julia Silge and David Robinson

Resource [URL](https://www.tidytextmining.com/index.html)

GitHub [URL](https://github.com/dgrtwo/tidy-text-mining.git)

This book serves as an introduction of text mining using the tidytext package and other tidy tools in R. The functions provided by the tidytext package are relatively simple; what is important are the possible applications. Thus, this book provides compelling examples of real text mining problems.

# Git and GitHub

## Version Control with GitHub

Author:

Resource [URL](https://carpentries-incubator.github.io/git-novice-branch-pr/)

GitHub [URL](https://github.com/carpentries-incubator/git-novice-branch-pr.git)

Version control is the lab notebook of the digital world: it’s what professionals use to keep track of what they’ve done and to collaborate with other people. Every large software development project relies on it, and most programmers use it for their small jobs as well. And it isn’t just for software: books, papers, small data sets, and anything that changes over time or needs to be shared can and should be stored in a version control system.

## Getting started with GitHub Pages

Author: [Leonard Gram](http://xlson.com/)

Resource [URL](http://xlson.com/2010/11/09/getting-started-with-github-pages.html#disqus_thread)

[Github Pages](http://pages.github.com/) is something as simple as a GitHub service for serving up static HTML from a GitHub repository. It’s the perfect place for project documentation, like [Javadoc](http://en.wikipedia.org/wiki/Javadoc), and that’s what I’ve been using it for. It can also be used for blogging with [Jekyll](http://jekyllrb.com/) or even a complete site for your project.

## Happy Git and GitHub for the useR

Author: Jenny Bryan

Resource [URL](https://happygitwithr.com/)

Happy Git provides opinionated instructions on how to:

* Install Git and get it working smoothly with GitHub, in the shell and in the [RStudio IDE](https://www.rstudio.com/products/rstudio/).
* Develop a few key workflows that cover your most common tasks.
* Integrate Git and GitHub into your daily work with R and [R Markdown](https://rmarkdown.rstudio.com/).

The target reader is someone who uses R for data analysis or who works on R packages, although some of the content may be useful to those working in adjacent areas.

The first two parts, [Installation](https://happygitwithr.com/install-intro.html#install-intro) and [Connect Git, GitHub, RStudio](https://happygitwithr.com/connect-intro.html#connect-intro), provide a quick start to verify your setup.

# DataViz

## Fundamentals of Data Visualization

Wilke C. (2019). *Fundamentals of data visualization : a primer on making informative and compelling figures* (First). O’Reilly Media.

Author: Wilke

Resource [URL](https://clauswilke.com/dataviz/)

GitHub [URL](https://github.com/clauswilke/dataviz.git)

This is the website for the book “Fundamentals of Data Visualization,” published by O’Reilly Media, Inc. The website contains the complete author manuscript before final copy-editing and other quality controls.

## The truthful art: Data Charts and Maps for Communication.

Cairo A. (2016). *The truthful art : data charts and maps for communication*. New Riders.

Alberto Cairos foundational guide to understanding information graphics and visualization, the respected data visualization professor explains in clear terms how to work with data, discover the stories hidden within, and share those stories with the world in the form of charts, maps, and infographics. In The Truthful Art, Cairo transforms elementary principles of data and scientific reasoning into tools that you can use in daily life to interpret data sets and extract stories from them.

# Markdown and Pandoc

## A Brief History of R Markdown (Slide Deck)

Author: Yihui Xie

Resource [URL](https://slides.yihui.org/2021-Brazilian-R-Day.html)

A Brief History of R Markdown presented as part of the 2021 Brazilian R-Day

## Do Yourself a Favor: Learn Markdown for Academic Writing

Author: [Biochemistry Computational Research Facility (BCRF)](https://bcrf.biochem.wisc.edu/)

Resource [URL](https://bcrf.biochem.wisc.edu/2021/06/10/do-yourself-a-favor-learn-markdown-episode-3/)

A short guide to using markdown in academic writing.

## R Markdown: The Definitive Guide

Author: Yihui Xie, J. J. Allaire, and Garrett Grolemund

Resource [URL](https://bookdown.org/yihui/rmarkdown/)

Online version of the printed book, which consists of four parts.

* Chapter [1](https://bookdown.org/yihui/rmarkdown/installation.html#installation) introduces how to install the relevant packages
* Chapter [2](https://bookdown.org/yihui/rmarkdown/basics.html#basics) is an overview of R Markdown, including the possible output formats, the Markdown syntax, the R code chunk syntax, and how to use other languages in R Markdown
* Part II is the detailed documentation of built-in output formats in the rmarkdown package, including document formats and presentation formats.
* Part III lists about ten R Markdown extensions that enable you to build different applications or generate output documents with different styles.

## R Markdown Cheat Sheets

Author: Posit

Resource [URL](https://github.com/rstudio/cheatsheets/raw/main/rmarkdown-2.0.pdf)

RStudio has created a large number of [cheat sheets](https://github.com/rstudio/cheatsheets/raw/main/rmarkdown-2.0.pdf), including the one-page R Markdown cheat sheet. Also, contains contributed cheat sheets.

## R Markdown Cookbook

Author: Yihui Xie, Christophe Dervieux, Emily Riedere

Resource [URL](https://bookdown.org/yihui/rmarkdown-cookbook/)

Online version of the printed book, broken down into small “recipes” that aim to demonstrate a single concept at a time.

## RMarkdown for Scientists

Author: Nicholas Tierney

Resource [URL](https://rmd4sci.njtierney.com/)

Book on rmarkdown, aimed for scientists. It was initially developed as a 3 hour workshop.

## Scientific Writing with Markdown

Author: [Jaan Tollander de Balsch](https://jaantollander.com/)

Resource [URL](https://jaantollander.com/post/scientific-writing-with-markdown/)

Great guide (with videos) on getting started with markdown.

# Project Management

## Data Management Best Practices

Author: Axiom Data Science

Resource [URL](https://www.axiomdatascience.com/best-practices/DataManagementBestPractices.html)

Axiom guide on best practices for managing scientific data.

## File Structure in Projects

recommendations for successful file structure organization and management.

Author: Broad Research Communication Lab, [brcl@broadinstitute.org](mailto:brcl@broadinstitute.org)

Resource [URL](https://mitcommlab.mit.edu/broad/commkit/file-structure/#Contents)

## Good enough practices in scientific computing

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, et al. (2017) Good enough practices in scientific computing. PLOS Computational Biology 13(6): e1005510. <https://doi.org/10.1371/journal.pcbi.1005510>

Article that presents a set of good computing practices that every researcher can adopt, regardless of their current level of computational skill.

Author: Wilson

Resource [URL](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005510)

## Managing Open and Reproducible Computational Projects (Beta)

Author: Carpentries

Resource [URL](https://carpentries-incubator.github.io/managing-computational-projects/01-introduction.html)

GitHub [URL](https://github.com/carpentries-incubator/managing-computational-projects)

Material covers best practices for managing and supervising computational projects in biology and related fields through data science methods, analysis, interpretation, and reporting processes.

## Quick Guide to Organizing Computational Biology Projects

Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLOS Computational Biology 5(7): e1000424. <https://doi.org/10.1371/journal.pcbi.1000424>

Author: Noble

Resource [URL](https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424)

The purpose of this article is to describe one good strategy for carrying out computational experiments. Author focuses on issues such as organizing files and directories and documenting progress. These are simply the principles and practices that I have developed over 12 years of bioinformatics research, augmented with various suggestions from other researchers with whom I have discussed these issues.

## Reproducible Research Data and Project Management in R

Author: Dr Anna Krystalli

Resource [URL](https://annakrystalli.me/rrresearchACCE20/)

This course focuses on **data and project management through R and Rstudio**, will introduce students to best practice and equip them with modern tools and techniques for managing data and computational workflows to their full potential. The course is designed to be relevant to students with a wide range of backgrounds, working with anything from relatively small sets of data collected from field or experimental observations, to those taking a more computational approach and bigger datasets.

## Using RStudio Projects

RStudio blog post on managing RStudio projects.

Author:

Resource [URL](https://support.posit.co/hc/en-us/articles/200526207-Using-RStudio-Projects)

# Quarto

## Official Documentation & Quick Start Guides

* [Documentation: Quarto documentation](https://quarto.org/) - Official Quarto Documentation
* [GitHub: Quarto GitHub repository](https://github.com/quarto-dev) - Official Quarto GitHub repository
* [Tutorial: Hello, Quarto](https://quarto.org/docs/get-started/hello/) - Official “Hello, Quarto” tutorial
* [Tutorial: Computations](https://quarto.org/docs/get-started/computations/) - Official “Computations” tutorial
* [Tutorial: Authoring](https://quarto.org/docs/get-started/authoring/) - Official “Authoring” tutorial

## Tutorials and Workshops

* [Communicate Chapters](https://r4ds.hadley.nz/communicate.html): R4DS has a set of 3 chapter on Quarto, which are part of the updated Communication section. The chapters focus on Quarto, other publishing frameworks, and an introduction to the “lab notebook” framework. Given that this content is part of the work-in-progress second edition of R for Data Science, you might want to consult the complete [first edition](https://r4ds.had.co.nz)
* [Quarto for Scientists](https://qmd4sci.njtierney.com/): This is a book on rmarkdown, aimed for scientists. It was initially developed as a 3 hour workshop, but is now developed into a resource that will grow and change over time.
* [Reproducible authoring with Quarto](https://www.youtube.com/watch?v=6p4vOKS6Xls) - 2022 Toronto Workshop on Reproducibility with Mine Çetinkaya-Rundel (slides: <https://mine-cetinkaya-rundel.github.io/2022-repro-toronto/>).
* [Tutorial: Making shareable docs with Quarto](https://openscapes.github.io/quarto-website-tutorial/) - A tutorial to make a website with Quarto.
* [Workshop: From R Markdown to Quarto](https://rstudio-conf-2022.github.io/rmd-to-quarto/) - A workshop for those who want to take their R Markdown skills and expertise and apply them in Quarto, the next generation of R Markdown.
* [Workshop: Getting started with Quarto](https://rstudio-conf-2022.github.io/get-started-quarto/) - “Get started with Quarto” workshop materials for rstudio::conf(2022). Note: [Mickaël Canouil](https://github.com/mcanouil) maintains a [page](https://github.com/mcanouil/awesome-quarto#official-documentation--quickstarts) of Quarto resources

# Reproducibility

## Data Set References

Author: APA Style

Resource [URL](https://apastyle.apa.org/style-grammar-guidelines/references/examples/data-set-references)

Provides an example citation using APA format. Data set references are covered in Section 10.9 of the [APA Publication Manual, Seventh Edition](https://apastyle.apa.org/products/publication-manual-7th-edition).

## Citing Research Objects

Author: Turing Way Community

Resource [URL](https://the-turing-way.netlify.app/communication/citable/citable-cite.html)

Recommendations for citing “research objects” directly in the paper in places where it is relevant. This is a commonly practised way of citing publications and is valid for citing other research components like data and software. A citation includes the following information:

* Author
* Title
* Year of publication
* Publisher (for data, this is often the data repository where it is housed)
* Version (if indicated)
* Access information (a URL or DOI)

## Coding Mindset

Author: **Broad Research Communication Lab,** [brcl@broadinstitute.org](mailto:brcl@broadinstitute.org)

Resource [URL](https://mitcommlab.mit.edu/broad/commkit/coding-mindset/)

This article covers a coding mindset, and outlines three high-level goals of a programmer that is writing code:

1. Solves a specific problem
2. Is easy to read and understand
3. Is maintainable and extendable

## Coding and Comment Style Guide

Author: **Broad Research Communication Lab,** [brcl@broadinstitute.org](mailto:brcl@broadinstitute.org)

Resource [URL](https://mitcommlab.mit.edu/broad/commkit/coding-and-comment-style/)

This article discusses how to use effective naming, structuring, context, and comments to communicate your logic in an easy-to-use code. Coding styles come in many shapes and sizes, but good ones derive from the same fundamental principles and possess a few key properties. Understanding these shared properties will help you choose a style that improves your code development, while also making it easier for others to implement

## Reproducibility, Generalizability, and Reuse

Author:

Resource [URL](https://liascript.github.io/course/?https://raw.githubusercontent.com/arcus/education_modules/main/reproducibility/reproducibility.md#1)

GitHub [URL](https://github.com/carpentries-incubator/git-novice-branch-pr.git)

This module provides learners with an approachable introduction to the concepts and impact of **research reproducibility**, **generalizability**, and **data reuse**, and how technical approaches can help make these goals more attainable.

This material will help you understand much of the current literature and debate around how research should be conducted, and will provide you with a starting point for understanding why some practices (like writing code, even for researchers who have never programmed a computer) are gaining traction in the research field.

## Turning Guide for Reproducible Research

Author: Becky Arnold, Louise Bowler, Sarah Gibson et.

Resources [URL](https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html)

Zenodo [URL](http://doi.org/10.5281/zenodo.3233986)

The Turing Way handbook provides an overview to reproducible, ethical and collaborative data science. The Turing Way project is open source, open collaboration, and community-driven. We involve and support a diverse community of contributors to make data science accessible, comprehensible and effective for everyone.

The book started in January 2019 as a guide for reproducibility, covering [Version Control](https://the-turing-way.netlify.app/reproducible-research/vcs.html#rr-vcs), [Code Testing](https://the-turing-way.netlify.app/reproducible-research/testing.html#rr-testing), and [Continuous integration](https://the-turing-way.netlify.app/reproducible-research/ci.html#rr-ci). However, technical skills are just one aspect of making data science research “open for all” and so in February 2020, *The Turing Way* expanded into a series of guides: [Guide for Reproducible Research](https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html#rr), [Guide for Project Design](https://the-turing-way.netlify.app/project-design/project-design.html#pd), [Guide for Communication](https://the-turing-way.netlify.app/communication/communication.html#cm), [Guide for Collaboration](https://the-turing-way.netlify.app/collaboration/collaboration.html#cl), and [Guide for Ethical Research](https://the-turing-way.netlify.app/ethical-research/ethical-research.html#er). [The Turing Way December Latest](https://github.com/alan-turing-institute/the-turing-way/releases/latest) release is available via GitHub.

## Introduction to Reproducible Publications with RStudio (Alpha)

Author

Resource [URL](https://carpentries-incubator.github.io/Reproducible-Publications-with-RStudio/)

Participants will build on their research and analytically skills with R Studio by learning how to create reproducible documents that combine code, analysis and narrative. RStudio makes it possible to work on a complete research project in a more efficient, integrated and organized manner: from data upload, cleaning, and analysis, to writing and styling a manuscript in R Markdown. It allows users to add code, figures, citations, and bibliography, as well as to generate outputs in various formats (e.g., html, pdf, doc). Rstudio also connects with Git and Github and learners will have a chance to experiment with this integration and understand its advantages for collaboration and version control. This workshop will involve lectures and demos alternating with hands-on exercises and challenges, so by the end of the workshop, learners will have “authored” a reproducible paper of your own with the data and narrative we provide.

## R for Reproducible Scientific Analysis

Author:

Resource [URL](https://swcarpentry.github.io/r-novice-gapminder/)

The goal of this lesson is to teach novice programmers to write modular code and best practices for using R for data analysis. The emphasis of these materials is to give attendees a strong foundation in the fundamentals of R, and to teach best practices for scientific computing: breaking down analyses into modular units, task automation, and encapsulation.

## Tao of open science for ecology (Resource)

Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, W. C. Lenhardt, A. MacDonald, W. K. Michener, J. Mudge, A. Pourmokhtarian, M. P. Schildhauer, K. H. Woo, and N. Zimmerman. 2015. The Tao of open science for ecology. Ecosphere 6(7):120.

Resource [URL](https://doi.org/10.1890/ES14-00402.1)

Presents an overview of tools and best practices that can enable these shifts in mindset at each stage of the research process, including tools to support data management planning and reproducible analyses, strategies for soliciting constructive feedback throughout the research process, and methods of broadening access to final research products.

# R and RStudio

## Command Line Essentials

Resource [URL](https://datacarpentry.org/shell-genomics/02-the-filesystem/index.htm)

A short primer on how to navigate through a file structure at the command line, from the [Introduction to the Command Line for Genomics Workshop](https://datacarpentry.org/shell-genomics/)

## Data Analysis and Visualization in R for Ecologists (Beta Alternative)

Author:

Resource [URL](https://carpentries-incubator.github.io/R-ecology-lesson-alternative/)

GitHub [URL](https://github.com/carpentries-incubator/R-ecology-lesson-alternative.git)

Designed for participants with no programming experience. It starts with information about the R programming language and the RStudio interface. It then moves to loading in data and exploring how to visualize it with ggplot2. The next episode takes learners through an exploration of data frames and some common data cleaning operations, before discussing vectors and factors. The final episode introduces the flow of data in R, and how to combine operations to select, filter, and mutate a data frame.

## R for Social Scientists

Author: Carpentries

Resource [URL](https://datacarpentry.org/r-socialsci/)

GitHub [URL](https://github.com/datacarpentry/r-socialsci.git) (I have forked)

Data Carpentry’s aim is to teach researchers basic concepts, skills, and tools for working with data so that they can get more done in less time, and with less pain. The lessons below were designed for those interested in working with social sciences data in R.

This is an introduction to R designed for participants with no programming experience. These lessons can be taught in a day (~ 6 hours). They start with some basic information about R syntax, the RStudio interface, and move through how to import CSV files, the structure of data frames, how to deal with factors, how to add/remove rows and columns, how to calculate summary statistics from a data frame, and a brief introduction to plotting.

## Using the RStudio Terminal in the RStudio IDE

Resource [URL](https://support.posit.co/hc/en-us/articles/115010737148-Using-the-RStudio-Terminal-in-the-RStudio-IDE)

The RStudio terminal provides access to the system shell from within the RStudio IDE. It supports full-screen terminal applications such as vim, Emacs, and tmux as, well as regular command-line operations with line-editing and shell history.

Potential uses of the shell including advanced source control operations, execution of long-running jobs, remote logins, system administration of RStudio Workbench or RStudio Server, and the aforementioned full-screen terminal programs.

# Statistics

## Open Statistics

Diez, D., Cetinkaya-Rundel, M., & Barr, C. (2022). OpenIntro Statistics.

OpenIntro Statistics covers a First course in statistics, providing a rigorous introduction to applied statistics that is clear, concise, and accessible. Appendix B, provides additional information for each of the data sets used in the main text and is new in the Fourth Edition. Online guides to each of these data sets are also provided at openintro.org/data and through a companion [R package](http://openintrostat.github.io/openintro/).

## Practical Statistics for Data Scientists, 2nd Edition

Bruce P. Bruce A. Gedeck P. & Safari an O’Reilly Media Company. (2020). Practical statistics for data scientists 2nd edition (2nd ed.). O’Reilly Media.

GitHub [URL](https://github.com/gedeck/practical-statistics-for-data-scientists.git)

The second edition of this popular guide adds comprehensive examples in Python, provides practical guidance on applying statistical methods to data science, tells you how to avoid their misuse, and gives you advice on what’s important and what’s not.

With this book, you’ll learn:

* Why exploratory data analysis is a key preliminary step in data science
* How random sampling can reduce bias and yield a higher-quality dataset, even with big data
* How the principles of experimental design yield definitive answers to questions
* How to use regression to estimate outcomes and detect anomalies
* Key classification techniques for predicting which categories a record belongs to
* Statistical machine learning methods that “learn” from data
* Unsupervised learning methods for extracting meaning from unlabeled data

Code repository for O’Reilly book.

# Other Resources

## Bash and Command Line 101

Author: [Nicole Feldman and Elizabeth Drellich](mailto:feldmanna@chop.edu%20drelliche@chop.edu)

Resource [URL](https://liascript.github.io/course/?https://raw.githubusercontent.com/arcus/education_modules/main/bash_command_line_101/bash_command_line_101.md#1)

Git [URL](https://github.com/arcus/education_modules.git)

This course is designed to be both an introduction to bash / command line for those who are total newbies as well as refresher for those some with experience running code who want a more solid command of the basics.

Introduction to Bioconductor: This is a course that is offered through [Harvard EdX](https://learning.edx.org/course/course-v1:HarvardX+PH525.5x+2T2022/home).

Introduction to Bioconductor and Standard Classes: Charlotte Soneson provides a nice overview of datatypes in R and compares these [datatypes](https://ivanek.github.io/analysisOfGenomicsDataWithR/02_IntroToBioc_html.html#standard-bioconductor-data-structures) in Bioconductor.

## Data-Carpentry-style lesson on some ML techniques in R

Author:

Resource [URL](https://carpentries-incubator.github.io/r-ml-tabular-data/)

GitHub [URL](https://github.com/carpentries-incubator/r-ml-tabular-data.git)

Lesson introduces a selection of machine learning techniques for analyzing tabular data, including random forests and gradient boosted trees. No experience in machine learning is necessary, but learners should be familiar with data analysis and visualization in R.

This lesson assumes some familiarity with R, including dplyrand ggplot. Learners who have completed an introductory Data Carpentry lesson in R should be able to follow the presentation. For a good refresher on prerequisite material, consider the lessons [Data Analysis and Visualization in R for Ecologists](https://datacarpentry.org/R-ecology-lesson/) or [R for Social Scientists](https://datacarpentry.org/r-socialsci/).

## Good Enough Practices in Scientific Computing

Author:

Resource [URL](https://carpentries-incubator.github.io/good-enough-practices/)

GitHub [URL](https://github.com/carpentries-incubator/good-enough-practices.git)

This lesson is targeted at a broad audience of researchers who want to learn how to be more efficient and effective in their data analysis and computing, whatever their career stage. Examples of our target audience are found in [learner profiles](https://carpentries-incubator.github.io/good-enough-practices/learner-profiles/index.html).

The lesson is inspired by and based on the paper, [Good Enough Practices in Scientific Computing (Wilson et al., 2017)](https://doi.org/10.1371/journal.pcbi.1005510): “a set of good computing practices that every researcher can adopt, regardless of their current level of computational skill”.

These practices, which encompass data management, programming, collaborating with colleagues, organizing projects, tracking work, and writing manuscripts, are drawn from a wide variety of published sources, from the daily lives of contributors, and from work with volunteer organizations that have delivered workshops to thousands of researchers since 2010.

## Introduction to Bioinformatics

Author: Laurent Gatto

Resource [URL](https://uclouvain-cbio.github.io/WSBIM1207/)

GitHub [URL](https://github.com/UCLouvain-CBIO/WSBIM1207.git) (I have forked)

This is an English language adaptation of [WSBIM1207](#X6589fc6ab0dc82cf12099d1c2d40ab994e8410c), an introduction to bioinformatics (and data science) for biology and biomedical students. It introduces bioinformatics methodology and technologies without relying on any prerequisites. The aim of this course is for students to be in a position to understand important notions of bioinformatics and tackle simple bioinformatics-related problems in R, in particular to develop simple R analysis scripts and reproducible analysis reports to interrogate, visualize and understand data in a tidy tabular format.

## Introduction to Data Analysis with R and Bioconductor

Author:

Resource [URL](https://carpentries-incubator.github.io/bioc-intro/)

GitHub [URL](https://github.com/carpentries-incubator/bioc-intro) (I have forked)

This is a beta incubator lesson. It appears to be modeled after the ecology or the R for Reproducible Scientific Analysis workshop. Lessons are pretty similar in scope until you get the the [Starting With Data](https://carpentries-incubator.github.io/bioc-intro/25-starting-with-data.html), which uses genomic data.

## Learning How to Learn Data Science

Resource [URL](https://liascript.github.io/course/?https://raw.githubusercontent.com/arcus/education_modules/main/learning_to_learn/learning_to_learn.md#1)

Git URL

The process of learning data science can be different from that of learning other subjects. This module goes over some of those differences and provides advice for navigating this potentially unfamiliar territory.

**Learning Objectives**

After completion of this module, learners will be able to:

* recognize ways in which learning data science and coding may be different than other educational experiences
* identify ways to extend their learning beyond module content
* recognize how to understand when to ask for help

## **Biomedical Data Science**

Authors: Rafael Irizarry and Michael Love, [MIT License](https://github.com/genomicsclass/labs/blob/master/LICENSE)

Resource [URL](#X6589fc6ab0dc82cf12099d1c2d40ab994e8410c)

# Licenses

Licensed under [CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/) 2022 by [the authors](https://carpentries-incubator.github.io/Reproducible-Publications-with-RStudio/CITATION).

# References