University of Scranton
Data Boot Camp
An Annotated Select Bibliography

Introductions to Data Science:

1. Doing Data Science: Straight Talk from the Frontline by O'Neil and Schutt

   This book is a survey of essential concepts and topics in data science written for a broad audience. Both authors are practicing data scientists and data science educators. If you are looking for an accessible overview of what data science (and to some extent big data) is all about this is a reasonable place to start. This book is published by O'Reilly press which has put out a plethora of books on all sorts of topics in programming, data science, statistics and similar fields. Data science in its current form has grown out of many different specialized areas such as mathematics/statistics, and computer science. O'Neil and Schutt both arrived in data science going through traditional mathematics (Kathy O'Neil has a PhD in number theory) and statistics. Their respective backgrounds somewhat influence the presentation of the field of data science in this book.

2. Data Science by Kelleher and Tierney

   This book also provides a general overview of data science from a nontechnical perspective. It is sorter than O'Neil and Schutt but lacks some of the flare that is contained in Doing Data Science. Kelleher and Tierney concisely cover the questions what is data sciece? and what is a data set?. They also introduce machine learning, the basic data science workflow and ethical considerations and privacy issues. Their book contains a glossary which it useful when first diving into the field of data science since many terms are not as well-defined as they should be given how commonly used they are.

3. Machine Learning: The New AI by Alpaydin

   Machine learning (or statistical learning or predictive modeling) is one of the basic approaches used in may data science applications. This book, published in the same series as Kelleher and Tierney, provides a nontechnical introduction to machine learning, pattern recognition, and neural networks and deep learning. Similar to Data Science by Kelleher and Tierney, Machine Learning: The New AI by Alpaydin contains a handy glossary.

Highly Recommended to Start:

1. Data Science and Predictive Analytics by Dinov

This text was developed out of courses in data science at the University of Michigan. The key strength of this book is that it guides the reader through many of the most important steps in the typical data science workflow using interesting real-world data and the R programming environment. In addition to being well-written and acessible, the book has an extensive corresponding online accompaniment that greatly enhances the learning experience of the reader.

2. Introduction to Data Science by Irizarry

   This is actually an upcoming book that is currently available as an online text `https://rafalab.github.io/dsbook/`. This text also guides the reader through many of the most important steps in the typical data science workflow using interesting real-world data and the R programming environment. This book is slightly less advanced that Data Science and Predictive Analytics by Dinov and does not have as great of coverage of statistical learning and predictive analytics but does have a careful coverage of basic probability and statistics as they relate to modern data science.

3. Data Visualization by Healy

   This book is simply beautiful. It is highly instructive and also highly useful. Healy describes best practices and provides excellent templates for making high quality and effective data visualizations in R in a very engaging manner. This book is a pleasure to read and delivers more than it promises.

4. Practical Statistics for Data Scientists by Bruce and Bruce

   This book gives a nice and suprisingly accessible overview of the most common "statistical" concepts used in data science. I put statistical in quotes because it goes beyond problems of inference (inference is not as often the goal in data science where prediction is usually the primary interest) and also has a lot of discussion about algorithms. There is hardly any actual math in the book yet the authors still give a responsible treatment of the key tenents

5. Modern Data Science with R by Baumer, Kaplan, and Horton

   This book provides a nice introductory overview of many data science topics including visualization, data wrangling, text analysis, network analysis, spatial data, and some basic modeling. The book provides nice and detailed coverage of tidy data and the "tidyverse" R packages that have been developed especially to exploit the tidy data format in order to optimize performance. There is also a chapters in the book on professioal ethics in data science and database administration.

6. An Introduction to Statistical Learning by James et al.

   This book is currently one of the most popular texts on statistical (machine) learning for data science. The text contains excellent descriptions of the most common statistical learning methods and provides a number of very good examples of the use of the methods. The R statistical programming is used in the book in order

to implement the various methods that are explained in the book. There is also a follow-up text to An Introduction to Statistical Learning titled Elements of Statistical Learning which fully develops the mathematical theory of the methods presented and applied in An Introduction to Statistical Learning.

7. Applied Predictive Modeling by Kuhn and Johnson

   This book is essentially a textbook companion to the documentation on the R package caret. This is an extremely powerful package that makes every aspect of training predictive models in R much easier. If you want to perform high quality machine learning quickly in R then this is a valuable book to read.

8. The Data Science Design Manual by Skiena

   TAn alternative title for this book could be *Data Science for Computer Scientists*. Two things that I like about the book are the overview chapter on what is data science and the concise coverage of mathematics and statistics that are essential for the practicing data scientist. This book is available to be downloaded as a pdf file from Springer Link via the University library. See also the corresponding web page `http://www.data-manual.com/` that also contains lecture videos and a link to a GitHub repository with python code that can be used to reproduce the figures and example in the book.

9. Data Science from Scratch (2nd ed.) by Grus

   Working through this book (more precisely through the code corresponding to this book) is a great way to get a solid foundation in many of the most common machine learning algorithms. The "from scratch" part of the title is a reference to the fact that the code for the book uses (almost) entirely base python and does not make use of sophisticated python modules like numpy and scikit learn. The benefit to this is that you develop a strong intuition for the algorithms but a consequence is that the code is nowhere near optimal in terms of performance. All of the python code for the book is available here: `https://github.com/joelgrus/data-science-from-scratch`.

R Programming:

1. The Art of R Programming by Matloff

   This is the best book that I am aware of that focuses purely on R as a programming language. This book does not contain in a substantial way any applications to statistics or data analysis. There is a chapter about graphics but the spirit of the book is to cover (fairly quickly) in detail the structure, syntax, and practice of programming in R. In this book you will learn about the basic data structures supported in R, object-oriented programming in R, basic input/output, debuggin, and optimizing R program performance. The author Norman Matloff is both a statistician and a computer scientist and is one of those rare people that is comfortable in both theoretical statistics and developing software applications. Some

of the material in this book, *e.g.* object-oriented programming, may seem unmotivated and difficult to apply but in fact the more you work with R, the more you will appreciate the general structures covered in The Art of R Programming. For example, when performing linear regression in R, what is the best way to work with the output from a call to the lm() function? It turns out that this is an object of class lm and understanding this makes it a lot easier to handle the corresponding output and also to quickly adapt what you know about linear regression to more general statistical modeling procedures with R.

2. Programming Skills for Data Science by Freeman and Ross

   Highly recommended. This book doesn't get very far in terms of breadth but it provides a very solid foundation in the basics. In addition to learning R, you also get an introduction to working with the command line (terminal), git and GitHub, markdown, and some other highly valuable tools.

3. The Book of R by Davies

   This book integrates programming (in R) and statistics. If you want to learn basic statistics and basic R programming simultaneously, which is a better idea than you might think at first, then *The Book of R* is a solid option.

4. R for Data Science by Grolemund and Wickham

   R for Data Science is an introductory level text that guides the reader through the use of the R statistical programming language for the most common computing tasks in the typical data science workflow. The main contribution of R for Data Science is that it enforces the use of so-call "tidy" data. This is a formatting protocol championed by Hadley Wickham that has been widely adopted throughout the data science community. Data that has been put into the tidy format is much more convenient to analyze and visualize. This book can be accessed for free online at `https://r4ds.had.co.nz/`.

5. R for Everyone (2nd ed.) by Lander

   If you are competent in the practice of statistics and you know what you want to do but can't remember or don't know how to do it in R then this book can be pretty useful, especially if a google search produces results that are overwhelming.

Python Programming:

1. Think Python (2nd ed.) by Downey

   This book is an introduction to basic computer science using python. You learn the basics of computer science and python programming at the same time. You can read it for free online: `https://greenteapress.com/wp/think-python-2e/`. The book is really well-written and actually pretty fun to read.

2. Introducing Python (2nd ed.) by Lubanovic

   This book is a very straight forward introduction to programming in python.

3. Pandas for Everyone by Chen

   I highly recommend this book if you want to do data wrangling and visualization in python. This book is probably my favorite to get started on doing data science in python but your best served if you already have some experience in general programming in python which you would certainly learn from either *Think Python* or *Introducing Python*.

4. Python Data Science Handbook by VanderPlas

   This is more of a reference. A great place to go if you already know what you want data science to do and just can't remember or don't know how to do it in python.

Excellent Intermediate Level Texts:

1. Statistical Regression and Classification by Matloff

   Matloff demonstrates his computer science side in The Art of R Programmin. In Statistical Regression and Classification he shows his statistical side. I highly recommend this book if you want to get into statiscal (machine) learning but don't have the strongest programming background. The book does assume substantial statistical proficiency and makes use of some basic probability and matrix algebra. Nevertheless, Matloff gives excellent intuitive explanations of many key ideas and blends the right amount of theory with application so that this book is highly useful to the budding practitioner.

2. Linear Models with R (2nd ed.) by Faraway

   Absolutely beautiful book. While the text focused on a specialized topic, it is one of the most important topics in both traditional statistics, applied statistics, and data science. The integration of theory, implementation, and application to extremely well-done. If you have never heard of least squares or hypothesis tests it is not the best place to start but even then the code contained in the book will help you to start applying linear models pretty quickly.

3. Extending the Linear Model with R (2nd ed.) by Faraway

   This is a continuation of Linear Models with R (2nd ed.) by the same author. Again, an absolutely beautiful book. The author explains the theory, implementation, and application of extensions to the basic linear model such as generalized and additive linear models, mixed effect models, and neural networks.

4. A Computational Approach to Statistical Learning by Arnold, Kane, and Lewis

This book can be summed up as "I already understand the maximum likelihood theory of generalized linear models but now I need to know how to implement the details as computer code that can be applied to real data." This book has a lot of usesful R code in it. However, the implementations are not optimized for large-scale application.

Excellent Theoretical Texts:

1. Mathematical Statistics with Resampling and R by Chihara and Hesterberg

   Many techniques in current data science grew out of traditional mathematical statistics. This is a highly recommended textbook on (undergraduate level) mathematical statistics that provides a solid background many techniques and approaches that are commonly used in data science, but that are not typically covered in the older textbooks on mathematical statistics. Notably, Chihara and Hesterberg provide excellent coverage of resampling and bootstrap methods for inference and (interval) estimation. The book requires a fairly strong background in calculus and a moderate background in probability theory. One the authors works at google demonstrating that the theory developed in this book is much more than mental exercise.

2. The Elements of Statistical Learning by Hastie, Tibshirani, and Friedman

   This is probably the most popular graduate level textbook on statistical (machine) learning. The book requires a strong background in calculus, linear algebra, and probability. The authors have made the book available for free: `https://web.stanford.edu/~hastie/ElemStatLearn/`.

3. Pattern Recognition and Machine Learning by Bishop

   This book is something of a classic in machine learning and is written by one of the leaders in the field. The perspective is a little different than that of *The Elements of Statistical Learning* and there is a lot more discussion of neural networks in *Pattern Recognition and Machine Learning*. It's very heavy and full of equations derived using some very nice probability theory. Requires a strong background in calculus, linear algebra, and probability.

4. Machine Learning: A Probabilistic Perspective by Murphy

   Along with *The Elements of Statistical Learning* and *Pattern Recognition and Machine Learning* this is a must read for anyone who is serious about understanding machine learning as an coherent theory. *Machine Learning: A Probabilistic Perspective* requires a strong background in calculus, linear algebra, and probability. The author is now a full-time google employ, further evidence that probability and machine learning theory is much more than mental exercise.

5. Neural Networks and Learning Machines (3rd Ed.) by Haykin

This is a highly mathematical (although not the most mathematical) text on neural networks. Requires a strong background in calculus, linear algebra, and probability.

Interesting Specialized Texts:

1. Humanities Data in R by Arnold and Tilton

   The title of the book really says it all. This book is available to be downloaded as a pdf file from Springer Link via the University library.

2. Text Mining with R A Tidy Approach

   This book gives a really nice (and pretty short) introduction to analyzing text (*e.g.* books from the Gutenberg) Availabel for free online at `https://www.tidytextmining.com/`.

3. Text Analysis with R for Students of Literature by Jockers

   This book is available to be downloaded as a pdf file from Springer Link via the University library.

4. Reproducible Econometrics Using R by Racine

   This is really a book about time-series analysis in R. It's fairly advanced but very well-written. You don't have to be an economist in order to get something useful out of this book. Really anyone intersted in the analysis of temporal data can get a lot out of this one.

5. Political Analysis Using R by Monogan

   The title of the book really says it all. This book is available to be downloaded as a pdf file from Springer Link via the University library. One of several books in the Use R! series published by Springer.

6. A User's Guide to Network Analysis in R by Luke

   This book is available to be downloaded as a pdf file from Springer Link via the University library. One of several books in the Use R! series published by Springer.

7. Statistical Analysis of Network Data with R by Kolaczyk and Csardi

   This book is available to be downloaded as a pdf file from Springer Link via the University library. One of several books in the Use R! series published by Springer.

8. ggplot2: Elegant Graphics for Data Analysis by Wickham

   This book is a full explanation of the ggplot2 R package for data visualization written by the person who developed that package. This book is available to be downloaded as a pdf file from Springer Link via the University library. One of several books in the Use R! series published by Springer.

9. R Packages by Wickham

   If one day you develop enough data and code in R and you want to share it all with the world in the form of your very own R package this book will explain how to do it. Note that there are reasons to develop an R package even if you don't want to share it.