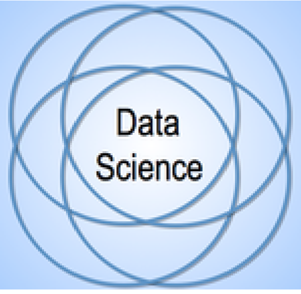
Data Science: An Introduction

Welcome to



Data Science: An Introduction

Wikibook

Preface

This book is a very basic introduction to data science. It is designed for the advanced high school student or average college freshman with a high school-level understanding of math, science, word processing and spreadsheets. No understanding of computer science is assumed. The main emphasis of this book is to help students think about the world in data science terms. While some elementary data science skills will be taught, the point is not skill development, but rather critical thinking and problem solving development. These are skills that can be successfully applied to all phases of life, not just data science.

Data science--as a profession and as an academic discipline unto itself—is new, having been born in the first decade of the 21st century. It is a child born of the mature parental disciplines of scientific methods, data and software engineering, statistics, and visualization. This book is not intended to do justice to any of those disciplines by themselves, but to bring them together in a productive synthesis. As such, the student will be introduced to the parent disciplines and then given exercises that will fuse the parental disciplines into data science. In addition, "hacking" in the original positive sense of the term, is also a contributing parent to the data science child, even though "hacking" is not taught as an academic discipline.

Obviously, a mature data scientist will be proficient in each of the parent disciplines, studying them individually and combining them to solve serious data problems. This text book is but just a first tentative step in that direction.

Data science, as practiced today, arises out of the "big data/cloud computing" world and complexity science. This means data science is an advanced discipline, requiring proficiency in parallel processing, map-reduce computing, petabyte-sized noSQL databases, machine learning, advanced statistics and complexity science. In this sense, "true" data science is more appropriately taught at the Master's and Doctorate level. We believe, however, **that data science is as much about mindset as it is about the skillful use of tools. Thus we want to engage students early in their careers to start thinking holistically about data science**. This textbook will not address the more advanced technologies and techniques of data science. It will, however, help students to start thinking like a data scientist.

In business and government today, data science is performed as teams. We want the students in this class have that experience. Thus, all the homework, assignments, and exercises are designed for teams of 2 to 6 students. We hope the students will have a chance to work with everyone else in the class over the course of the semester. Most data scientists do not get to choose who they work with, but must learn to work with whomever is assigned to their team.

We will do most of our data manipulation, computer programming, and statistical analysis in the open source **R** package. We know that intermediate or advanced students would use other tools such as MySQL, PHP, Python, Java, Hadoop, HBase, AllegroGraph, Mahout, MATLAB, SPSS, SAS, etc. For this introduction, however, we are keeping it simple and sticking to just a single general purpose computing environment.

Finally, we try to use terms and concepts which are already defined in the Wikipedia, Wiktionary, and Wikiversity. This way people can refer to the corresponding Wikipedia/Wiktionary page to get a deeper understanding of the concept.

Stages

In the table of contents on the right side of the page, you will notice there is a little box of four squares. The box indicates the maturity of the chapter. For example,

Note to Instructors

We have designed this text for a 16-week 3-credit class. That is, a class that has three classroom-hours of instruction for 16 weeks—for example, 48 1-hour class periods. There are 32 chapters, which allows for—averaged over the semester—one day a week for student project presentations, for reviews and help sessions, and for testing. We image that there will be more lecture periods toward the beginning of the semester and more presentation and review days toward the end of the semester. The book also assumes 1 to 2 hours of "homework" per class period, which includes readings, assignments, study, and projects. The book's philosophy is that as much will be learned about data science by doing team homework projects as will be learned during the lectures.

In the professional world, data science is a team sport. We designed the difficulty and scope of the homework project for teams. At this level (high school senior or first semester college freshman), it would be difficult for a single individual to complete these assignments alone. We also assume there is a place students can go to get help with the R programming language.

Note to Contributors

First, please register yourself with Wikibooks (and list yourself below), so that we know who our co-contributors are. Also, please abide by the Wikibooks Editing Guidelines, Manual of Style, and Policies and Guidelines. Thank you.

Secondly, we only need basic, clear, straightforward information in each chapter. We are not trying to be exhaustive or complete—the value of this book is in the simple synthesis across subjects. There are other venues in which to wax eloquent on the deepness and complexities of a particular subject. Please place yourself in a "beginner's mind" as you make contributions. Please also scope each chapter so that it can be taught in a one-hour class period. If the chapter requires more than an hour to teach, it is probably too detailed.

- To the extent possible, please use terms and concepts in the way in which they are defined in the Wikipedia and Wiktionary. This way students can refer to the corresponding Wikipedia / Wiktionary page to get a deeper understanding of the concept.

Thirdly, this is a cross-disciplinary book. We want to help people apply data science to all fields. Therefore, we need a wide variety of simple examples and simple exercises.

Fourthly, please adhere to the simple structure of each chapter: Summary of Main Points, Discussion, More Reading, Exercises, and References. We want the More Reading section to link to on-line resources. The References section may contain off-line resources. To start a new page, you should use the wiki markup from **this prototype page**.

Fifthly, as with any Wikibook please feel free to make corrections, expand explanations, and make additions where necessary, even if it is not "your" chapter. Use the discussion page to explain changes that might be controversial.

Sixthly, some syntax rules:

- Please **bold** key terms and phrases the student should learn.
- Put the name of functions and code snippets using the 'code' tags: `<code>lm()</code>`
- Use in-line links `[[ ]]` to the Wikipedia, Wiktionary, WikiCommons, Wikibooks, and other Wikimedia Foundation properties.
- Use references (<ref> </ref>) to "external" sources—both on-line and off-line.
  - Use the citations templates to make citations : Template:Cite book, Template:Cite web, Template:Cite journal
- When inserting R code into a page, please adhere to Google's R Style Guide.[1]
- If you want to add an image or graph, you should load it into the Commons rather than uploading into Wikibooks.
  - If appropriate, add the tag `{{Created with R}}`) when you upload the graph.
- If using a different package than **R** standard packages, put the name of the package in bold in parenthesis after each function : <code>MCMCprobit()</code> ('''MCMCpack''')
- You can use the third chapter Definitions of Data as an example of how to craft a chapter.

Finally, thank you so much for volunteering to be part of our our team!

List of Co-Authors

- Calvin Andrus
- Jon Cook
- Suresh Sood

See also

See the following Wikibooks for good follow-on texts to this introduction:

- The Scientific Method - Scientific Method
- Data Engineering - Relational Database Design, Data Structures, SQL
- Software Engineering - The Science of Programming, R Programming

- Mathematics - High School Mathematics Extensions
- Statistical Analysis - Statistics, Statistical Analysis: An Introduction Using R, Data Mining Algorithms in R
- Visualization -
- Hacking -

References

1. "R Style Guide". Google, Inc.. http://google-styleguide.googlecode.com/svn/trunk/google-r-style.html. Retrieved 6 July 2012.

Copyright Notice