Texts in Computer Science

More information about this series at http://www.springer.com/series/3191

Steven S. Skiena

The Data Science Design Manual

Springer

Steven S. Skiena
Computer Science Department
Stony Brook University
Stony Brook, NY
USA

ISSN 1868-0941 ISSN 1868-095X (electronic)
Texts in Computer Science
ISBN 978-3-319-55443-3 ISBN 978-3-319-55444-0 (eBook)
https://doi.org/10.1007/978-3-319-55444-0

Library of Congress Control Number: 2017943201

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Making sense of the world around us requires obtaining and analyzing data from our environment. Several technology trends have recently collided, providing new opportunities to apply our data analysis savvy to greater challenges than ever before.

Computer storage capacity has increased exponentially; indeed remembering has become so cheap that it is almost impossible to get computer systems to forget. Sensing devices increasingly monitor everything that can be observed: video streams, social media interactions, and the position of anything that moves. Cloud computing enables us to harness the power of massive numbers of machines to manipulate this data. Indeed, hundreds of computers are summoned each time you do a Google search, scrutinizing all of your previous activity just to decide which is the best ad to show you next.

The result of all this has been the birth of *data science*, a new field devoted to maximizing value from vast collections of information. As a discipline, data science sits somewhere at the intersection of statistics, computer science, and machine learning, but it is building a distinct heft and character of its own. This book serves as an introduction to data science, focusing on the skills and principles needed to build systems for collecting, analyzing, and interpreting data.

My professional experience as a researcher and instructor convinces me that one major challenge of data science is that it is considerably more subtle than it looks. Any student who has ever computed their grade point average (GPA) can be said to have done rudimentary statistics, just as drawing a simple scatter plot lets you add experience in data visualization to your resume. But meaningfully analyzing and interpreting data requires both technical expertise and wisdom. That so many people do these basics so badly provides my inspiration for writing this book.

To the Reader

I have been gratified by the warm reception that my book *The Algorithm Design Manual* [Ski08] has received since its initial publication in 1997. It has been recognized as a unique guide to using algorithmic techniques to solve problems that often arise in practice. The book you are holding covers very different material, but with the same motivation.

In particular, here I stress the following basic principles as fundamental to becoming a good data scientist:

- *Valuing doing the simple things right*: Data science isn't rocket science. Students and practitioners often get lost in technological space, pursuing the most advanced machine learning methods, the newest open source software libraries, or the glitziest visualization techniques. However, the heart of data science lies in doing the simple things right: understanding the application domain, cleaning and integrating relevant data sources, and presenting your results clearly to others.

  Simple doesn't mean easy, however. Indeed it takes considerable insight and experience to ask the right questions, and sense whether you are moving toward correct answers and actionable insights. I resist the temptation to drill deeply into clean, technical material here just because it is teachable. There are plenty of other books which will cover the intricacies of machine learning algorithms or statistical hypothesis testing. My mission here is to lay the groundwork of what really matters in analyzing data.

- *Developing mathematical intuition*: Data science rests on a foundation of mathematics, particularly statistics and linear algebra. It is important to understand this material on an intuitive level: why these concepts were developed, how they are useful, and when they work best. I illustrate operations in linear algebra by presenting pictures of what happens to matrices when you manipulate them, and statistical concepts by examples and *reducto ad absurdum* arguments. My goal here is transplanting intuition into the reader.

  But I strive to minimize the amount of formal mathematics used in presenting this material. Indeed, I will present exactly one formal proof in this book, an incorrect proof where the associated theorem is obviously false. The moral here is not that mathematical rigor doesn't matter, because of course it does, but that genuine rigor is impossible until after there is comprehension.

- *Think like a computer scientist, but act like a statistician*: Data science provides an umbrella linking computer scientists, statisticians, and domain specialists. But each community has its own distinct styles of thinking and action, which gets stamped into the souls of its members.

  In this book, I emphasize approaches which come most naturally to computer scientists, particularly the algorithmic manipulation of data, the use of machine learning, and the mastery of scale. But I also seek to transmit the core values of statistical reasoning: the need to understand the application domain, proper appreciation of the small, the quest for significance, and a hunger for exploration.

  No discipline has a monopoly on the truth. The best data scientists incorporate tools from multiple areas, and this book strives to be a relatively neutral ground where rival philosophies can come to reason together.

Equally important is what you will not find in this book. I do not emphasize any particular language or suite of data analysis tools. Instead, this book provides a high-level discussion of important design principles. I seek to operate at a conceptual level more than a technical one. The goal of this manual is to get you going in the right direction as quickly as possible, with whatever software tools you find most accessible.

To the Instructor

This book covers enough material for an *"Introduction to Data Science"* course at the undergraduate or early graduate student levels. I hope that the reader has completed the equivalent of at least one programming course and has a bit of prior exposure to probability and statistics, but more is always better than less.

I have made a full set of lecture slides for teaching this course available online at `http://www.data-manual.com`. Data resources for projects and assignments are also available there to aid the instructor. Further, I make available online video lectures using these slides to teach a full-semester data science course. Let me help teach your class, through the magic of the web!

Pedagogical features of this book include:

- *War Stories*: To provide a better perspective on how data science techniques apply to the real world, I include a collection of "war stories," or tales from our experience with real problems. The moral of these stories is that these methods are not just theory, but important tools to be pulled out and used as needed.

- *False Starts*: Most textbooks present methods as a fait accompli, obscuring the ideas involved in designing them, and the subtle reasons why other approaches fail. The war stories illustrate my reasoning process on certain applied problems, but I weave such coverage into the core material as well.

- *Take-Home Lessons*: Highlighted "take-home" lesson boxes scattered through each chapter emphasize the big-picture concepts to learn from each chapter.

- *Homework Problems*: I provide a wide range of exercises for homework and self-study. Many are traditional exam-style problems, but there are also larger-scale implementation challenges and smaller-scale interview questions, reflecting the questions students might encounter when searching for a job. Degree of difficulty ratings have been assigned to all problems.

  In lieu of an answer key, a Solution Wiki has been set up, where solutions to all even numbered problems will be solicited by crowdsourcing. A similar system with my *Algorithm Design Manual* produced coherent solutions,

or so I am told. As a matter of principle I refuse to look at them, so let the buyer beware.

- *Kaggle Challenges*: Kaggle (`www.kaggle.com`) provides a forum for data scientists to compete in, featuring challenging real-world problems on fascinating data sets, and scoring to test how good your model is relative to other submissions. The exercises for each chapter include three relevant Kaggle challenges, to serve as a source of inspiration, self-study, and data for other projects and investigations.

- *Data Science Television*: Data science remains mysterious and even threatening to the broader public. *The Quant Shop* is an amateur take on what a data science reality show should be like. Student teams tackle a diverse array of real-world prediction problems, and try to forecast the outcome of future events. Check it out at `http://www.quant-shop.com`.

  A series of eight 30-minute episodes has been prepared, each built around a particular real-world prediction problem. Challenges include pricing art at an auction, picking the winner of the Miss Universe competition, and forecasting when celebrities are destined to die. For each, we observe as a student team comes to grips with the problem, and learn along with them as they build a forecasting model. They make their predictions, and we watch along with them to see if they are right or wrong.

  In this book, *The Quant Shop* is used to provide concrete examples of prediction challenges, to frame discussions of the data science modeling pipeline from data acquisition to evaluation. I hope you find them fun, and that they will encourage you to conceive and take on your own modeling challenges.

- *Chapter Notes*: Finally, each tutorial chapter concludes with a brief notes section, pointing readers to primary sources and additional references.

Dedication

My bright and loving daughters Bonnie and Abby are now full-blown teenagers, meaning that they don't always process statistical evidence with as much alacrity as I would I desire. I dedicate this book to them, in the hope that their analysis skills improve to the point that they always just agree with me.

And I dedicate this book to my beautiful wife Renee, who agrees with me even when she doesn't agree with me, and loves me beyond the support of all creditable evidence.

Acknowledgments

My list of people to thank is large enough that I have probably missed some. I will try to do enumerate them systematically to minimize omissions, but ask those I've unfairly neglected for absolution.

First, I thank those who made concrete contributions to help me put this book together. Yeseul Lee served as an apprentice on this project, helping with figures, exercises, and more during summer 2016 and beyond. You will see evidence of her handiwork on almost every page, and I greatly appreciate her help and dedication. Aakriti Mittal and Jack Zheng also contributed to a few of the figures.

Students in my Fall 2016 *Introduction to Data Science* course (CSE 519) helped to debug the manuscript, and they found plenty of things to debug. I particularly thank Rebecca Siford, who proposed over one hundred corrections on her own. Several data science friends/sages reviewed specific chapters for me, and I thank Anshul Gandhi, Yifan Hu, Klaus Mueller, Francesco Orabona, Andy Schwartz, and Charles Ward for their efforts here.

I thank all the *Quant Shop* students from Fall 2015 whose video and modeling efforts are so visibly on display. I particularly thank Jan (Dini) Diskin-Zimmerman, whose editing efforts went so far beyond the call of duty I felt like a felon for letting her do it.

My editors at Springer, Wayne Wheeler and Simon Rees, were a pleasure to work with as usual. I also thank all the production and marketing people who helped get this book to you, including Adrian Pieron and Annette Anlauf.

Several exercises were originated by colleagues or inspired by other sources. Reconstructing the original sources years later can be challenging, but credits for each problem (to the best of my recollection) appear on the website.

Much of what I know about data science has been learned through working with other people. These include my Ph.D. students, particularly Rami al-Rfou, Mikhail Bautin, Haochen Chen, Yanqing Chen, Vivek Kulkarni, Levon Lloyd, Andrew Mehler, Bryan Perozzi, Yingtao Tian, Junting Ye, Wenbin Zhang, and postdoc Charles Ward. I fondly remember all of my Lydia project masters students over the years, and remind you that my prize offer to the first one who names their daughter Lydia remains unclaimed. I thank my other collaborators with stories to tell, including Bruce Futcher, Justin Gardin, Arnout van de Rijt, and Oleksii Starov.

I remember all members of the General Sentiment/Canrock universe, particularly Mark Fasciano, with whom I shared the start-up dream and experienced what happens when data hits the real world. I thank my colleagues at Yahoo Labs/Research during my 2015–2016 sabbatical year, when much of this book was conceived. I single out Amanda Stent, who enabled me to be at Yahoo during that particularly difficult year in the company's history. I learned valuable things from other people who have taught related data science courses, including Andrew Ng and Hans-Peter Pfister, and thank them all for their help.

> If you have a procedure with ten parameters, you probably missed some.
>
> – Alan Perlis

Caveat

It is traditional for the author to magnanimously accept the blame for whatever deficiencies remain. I don't. Any errors, deficiencies, or problems in this book are somebody else's fault, but I would appreciate knowing about them so as to determine who is to blame.

Steven S. Skiena
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-2424
`http://www.cs.stonybrook.edu/~skiena`
skiena@data-manual.com
May 2017

Contents