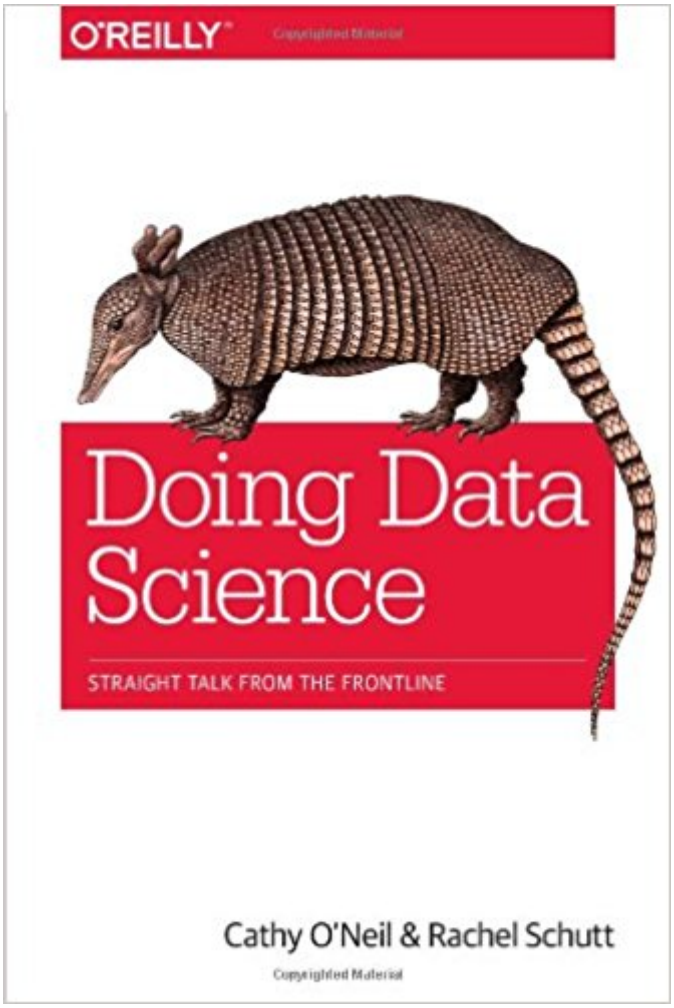
Doing Data Science: Straight Talk From The Frontline



Synopsis

Now that people are aware that data can make the difference in an election or a business model, data science as an occupation is gaining ground. But how can you get started working in a wide-ranging, interdisciplinary field thatâ ™s so clouded in hype? This insightful book, based on Columbia Universityâ ™s Introduction to Data Science class, tells you what you need to know.In many of these chapter-long lectures, data scientists from companies such as Google, Microsoft, and eBay share new algorithms, methods, and models by presenting case studies and the code they use. If youâ ™re familiar with linear algebra, probability, and statistics, and have programming experience, this book is an ideal introduction to data science.Topics include:Statistical inference, exploratory data analysis, and the data science processAlgorithmsSpam filters, Naive Bayes, and data wranglingLogistic regressionFinancial modelingRecommendation engines and causalityData visualizationSocial networks and data journalismData engineering, MapReduce, Pregel, and HadoopDoing Data Science is collaboration between course instructor Rachel Schutt, Senior VP of Data Science at News Corp, and data science consultant Cathy Oâ ™Neil, a senior data scientist at Johnson Research Labs, who attended and blogged about the course.

Book Information

Paperback: 408 pages

Publisher: O’Reilly Media; 1 edition (November 3, 2013)

Language: English

ISBN-10: 1449358659

ISBN-13: 978-1449358655

Product Dimensions: 6 x 0.8 x 9 inches

Shipping Weight: 1.3 pounds (View shipping rates and policies)

Average Customer Review: 3.8 out of 5 stars 53 customer reviews

Best Sellers Rank: #73,071 in Books (See Top 100 in Books) #4 inÂ Books > Science & Math > Mathematics > Applied > Stochastic Modeling #24 inÂ Books > Textbooks > Computer Science > Algorithms #48 inÂ Books > Computers & Technology > Databases & Big Data > Data Mining

Customer Reviews

"Every once in a while a single book comes to crystallize a new discipline. If books still have this power in the era of electronic media, "Doing Data Science: Straight Talk from the Frontline"Â by Rachel Schutt and Cathy O’Neil: O’Reilly, 2013 might just be the book that defines data science."Â -- Joseph RickertRevolutions Blog "I enjoyed Rachel and Cathyâ ™s book, itâ ™s

readable, informative, and like no other book Iâ ™ve read on the topic of statistics or data science."Â â ”Andrew Gelman Professor of statistics and political science, and director of the Applied Statistics Center at Columbia University Â "I got a lot out of Doing Data Science, finding the chapter organization on business problem specification, analytics formulation, data access/wrangling, and computer code to be very helpful in understanding DS solutions." â ”Steve Miller Co-founder, OpenBI, LLC, a Chicago-based business intelligence services firm Â

Cathy Oâ ™Neil earned a Ph.D. in math from Harvard, was postdoc at the MIT math department, and a professor at Barnard College where she published a number of research papers in arithmetic algebraic geometry. She then chucked it and switched over to the private sector. She worked as a quant for the hedge fund D.E. Shaw in the middle of the credit crisis, and then for RiskMetrics, a risk software company that assesses risk for the holdings of hedge funds and banks. She is currently a data scientist on the New York start-up scene, writes a blog at mathbabe.org, and is involved with Occupy Wall Street.Rachel Schutt is the Senior Vice President for Data Science at News Corp. She earned a PhD in Statistics from Columbia University, and was a statistician at Google Research for several years. She is an adjunct professor in Columbiaâ ™s Department of Statistics and a founding member of the Education Committee for the Institute for Data Sciences and Engineering at Columbia. She holds several pending patents based on her work at Google, where she helped build user-facing products by prototyping algorithms and building models to understand user behavior. She has a master’s degree in mathematics from NYU, and a master’s degree in Engineering-Economic Systems and Operations Research from Stanford University. Her undergraduate degree is in Honors Mathematics from the University of Michigan.

Book review - Doing Data Science by O’Neil and Schutt, O’Reilly Media.More breadth than depthWhat is data science? The book Doing Data Science not only explains what data science is but also provides a broad overview of methods and techniques that one must master in order to call one self a data scientist. The book is based on a course about data science given at Columbia University. However it is not to be considered as a text book about data science but more as a broad introduction to a number of topics in data science.In the spring of 2013 I followed two Coursera courses. One about the statistical programming language R and one on Data Analysis. I had for some time been looking for a book that could be used as a follow-up reading on topics in data science. This was the reason I picked up "Doing Data Science".The book begins with a chapter about what data science is all about is followed by four chapters on topics like statistical inference,

explanatory data analysis, various machine learning algorithms, linear and logistic regression, and Naive Bayes. I have a background in both mathematics and statistics and I was able to understand these chapters but the material is covered in such broad terms that I find it hard to believe that a newcomer to this topics will understand or gain much knowledge from reading these chapters. Basic math is presented about the models but without some kind of detailed explanation one cannot develop any deeper intuition for the approach explained.The best parts of the book is definitely chapter 6 to 8 and 10. In here we find interesting discussion about coverage of data science applied to financial modeling, extracting information from data, and social networks. I really enjoyed the examination of time stamped data, the Kaggle Model, feature selection, and case-attribute data versus social network data. The math behind these topics was however once again explained quite superficial. Centrality measures is central to social network analysis but it is very hard to develop intuition for there measures without a more detailed explanation about the underlying math. These chapters contains lots of useful resources for finding additional information about the discussed topics.Data visualization is an integral part of data science for communication results. Beginners in the field of data science needs concrete and easy to follow instruction on how to get started with visualization. Unfortunately the book focuses more on the use of data visualization in modern art projects. The content is simply to abstract for beginners to learn about the usage of visualization in data science.When I was browsing the book before actual buying it I was kind thrilled to see that it covered topics like causality and epidemiology. Topics that I did not found covered in any other book about data science. However the chapter about epidemiology is not about using data science in epidemiology but 'just' about using data science to evaluate the methods used in epidemiology. Likewise there seems to be no link between data science and causality. I later discovered that the authors used an entire blog post ([...] to explain why causality was part of the university course underlying the book. This material or parts of it should have made it into the book. I am still not convinced that causality is a topic in data science.There are several examples in which the book assumes the reader to have knowledge of US government structure and organizations. Examples include page 292 when discussing US health care databases and page 298 where FDA is mentioned without further introduction or explanation about what FDA is.A book than contains programming examples should always make the code accessible to download. Typing in the code yourself is simply waste of time. It is possible to download some of the datasets used in the book through GitHub. But the code does not seem to be available. I also own the electronic version of the book and I tried to copy-paste some of the examples from the e-book but there are several examples of code that hasn't been proof written or tested prior to publication. The sample code

misses references to required R libraries or refers to computer folder structures on some local Columbia University computer. The companion datasets that can be downloaded on GitHub consists of a number of Excel files. The R sample code uses the gdata package to load these Excel files into R for further analysis. It took be quite some time to figure out why this process didn't work on a Windows computer. The gdata package requires Perl to be installed on the computer and this is not default software on Windows. In my opinion one should always publish data in a simple format, e.g. csv files and definitely not proprietary formats like xls for Excel files.Data Science is both science and a lot of practical experience. I guess the title of the book Doing Data Science tries to capture that. You need to do data science in order to learn it. The covered topics are interesting but the material is more breadth than depth. Luckily there are lots of useful links and resources to additional materials. Personally I would prefer more details about the actual data science topics like e.g. extracting meaning from data and social network analysis and less focus on math. The book already requires some knowledge of math, statistics and programming, so why not presume that the reader has the background knowledge and dive straight into the data science discussions.I really like the idea about having a lot of different people present various topics in data science and the book is well written and contains lots of useful resources for further studies of data science. I will recommend to book to people new to the subject but be aware of the fact that source code is not available and that is a major drawback.Disclosure: I review for the O'Reilly Reader Review Program and I want to be transparent about my reviews so you should know that I received a free copy of this ebooks in exchange of my review.

Great text that provides a very informative overview of topics in Data Science. Ideal for someone like me: Generally curious but cautiously skeptical. Description of K-NN/Bayesian classification algorithm encouraged me to study these and related topics further. Obviously not a one-stop shop as the title suggests.My only complain is that the author sometimes digress too much on industry background/providing context. This seemed like a good thing in the beginning, but eventually just kind of became a distraction to the main text. Individual preference may differ on this

The book is well written and provides good insights into how to form a foundational core to further one's education and experience in data analysis and visualization. To truly accomplish the end results that significantly impact informed decisions, one must bring to the table a well rounded background and experience. The book serves as an excellent source to focus on a positive approach to learning and executing data science. Not all will agree, but in my opinion, the secret to

success in this area is never to evolve your skill isolated from the works of those already successful and willing to share their knowledge.

All or most of the key issues of data science are covered. That is really good for reading. However, the presentation of material makes it difficult for a student to quickly follow.

I found this book to be a very odd bird indeed. It is one book you can read from back cover to front cover and not be at a disadvantage. This is because the book is really just a collection of presentations made by various people to a class taught by the primary author Rachel Schutt at Columbia University in the Fall of 2012 Ã¢Â Â“ Introduction to Data Science. It wasnÃ¢Â Â™t entirely clear what content Schutt was directly responsible for since only some of the chapters indicate who the contributors were (one of the chapters was contributed by a group of her students!). The co-author, Cathy OÃ¢Â Â™Neil, IÃ¢Â Â™ve encountered before as an outspoken blogger going by the name Ã¢Â Âœmathbabe Ã¢Â Â• but it wasnÃ¢Â Â™t specifically stated how she became part of the book project, other than to say she was one of the students in SchuttÃ¢Â Â™s class. Chapter 6 was partly written by OÃ¢Â Â™Neil.Both Schutt and OÃ¢Â Â™Neil are Ph.D.s data science appropriate fields, but the book was not Ã¢Â Âœwritten Ã¢Â Â• by the two, rather they seemed to have performed some kind of editing function with the materials submitted by each contributor and added commentaries of their own. As a result, the book is a hodgepodge of anecdotes, factoids, R code snippets, plots, and mathematics, all from the in-class presentations. I enjoy seeing math in data science books, but the equations in this book were sort of just floating there requiring the reader to explore further at another time.Although I have issues with the book as it is not any sort of text for the field, I did enjoy reading it with a number of Ã¢Â Âœ Ah, I didnÃ¢Â Â™t know that! Ã¢Â Â• moments. SchuttÃ¢Â Â™s credentials in data science are considerable, having worked at Google for a few years around the same time that Ã¢Â Âœdata science Ã¢Â Â• was growing up in Silicon Valley. As a result the book has many memorable anecdotes about the early days of the data science industry, and observations about what makes big data tick. I enjoyed the story about the Google software engineer who accidentally deleted 10 petabytes of data, and I think my favorite quote from the book is from the studentÃ¢Â Â™s chapter 15: Kaggle competitions could be described as the dick-measuring contests of data science.With contributorÃ¢Â Â™s chapters on statistical inference, machine learning algorithms, logistic regression, financial modeling, recommendation engines, data visualization, Hadoop, MapReduce, and more, IÃ¢Â Â™d say the book is worth a read, but not

necessarily as a source of learning data science but more as a high-level guide and short historical account of this young industry. You get to learn about the people, companies, technologies that have collectively built the data science arena and youÃ¢Â Â™ll be better for it especially if you are working to become a data scientist yourself.

Doing Data Science: Straight Talk from the Frontline Data Analytics: What Every Business Must Know About Big Data And Data Science (Data Analytics for Business, Predictive Analysis, Big Data Book 1) Data Analytics: Applicable Data Analysis to Advance Any Business Using the Power of Data Driven Analytics (Big Data Analytics, Data Science, Business Intelligence Book 6) Big Data For Business: Your Comprehensive Guide to Understand Data Science, Data Analytics and Data Mining to Boost More Growth and Improve Business - Data Analytics Book, Series 2 Analytics: Data Science, Data Analysis and Predictive Analytics for Business (Algorithms, Business Intelligence, Statistical Analysis, Decision Analysis, Business Analytics, Data Mining, Big Data) How to Talk Dirty : Dirty Talk Examples, Secrets for Women and Men, Straight, Gay and Bi, Spice Up Your Sex Life and Have Mindblowing Sex: Great Sex Book, Series 1 Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data Data Analytics and Python Programming: 2 Bundle Manuscript: Beginners Guide to Learn Data Analytics, Predictive Analytics and Data Science with Python Programming Conversation: The Gentle Art Of Hearing & Being Heard - HowTo "Small Talk", How To Connect, How To Talk To Anyone (Conversation skills, Conversation starters, Small talk, Communication) How to Talk Dirty: Make Him Explode Whispering These 173 Dirty Talk Examples that Will Rock His World & Have Him on His Knees Begging You for Sex (Improve & Spice Up Your Sex Life - Dirty Talk) Small Talk Made EASY!: How to Talk To Anyone Effortlessly and Talk with Confidence and Ease! Analytics: Business Intelligence, Algorithms and Statistical Analysis (Predictive Analytics, Data Visualization, Data Analytics, Business Analytics, Decision Analysis, Big Data, Statistical Analysis) Data Analytics For Beginners: Your Ultimate Guide To Learn and Master Data Analysis. Get Your Business Intelligence Right â " Accelerate Growth and Close More Sales (Data Analytics Book Series) Discovering Knowledge in Data: An Introduction to Data Mining (Wiley Series on Methods and Applications in Data Mining) Straight to Bed: A Gay Man's Guide to Seducing Straight Men Crochet the Corner to Corner and Straight Box Stitch for Beginners: Learn the Basics of Crochet and How to Crochet the Popular C2C and Straight Box Stitch Patterns Doing Business by the Good Book: 52 Lessons on Success Straight from the Bible Life, Liberty, and the Pursuit of Healthiness: Dr. Dean's Straight-Talk Answers to Hundreds of Your

Most Pressing Health Questions Men Issues: Straight Talk About Andropause, Prostate and Erectile Dysfunction