

VINCENT GRANVILLE, PH.D.  
CO-FOUNDER OF DATA SCIENCE CENTRAL

DEVELOPING

---

**ANALYTIC  
TALENT**

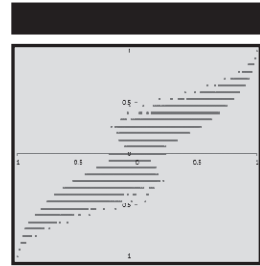
BECOMING A DATA SCIENTIST

WILEY



# **Developing Analytic Talent**





# Developing Analytic Talent

Becoming a Data Scientist

Vincent Granville, Ph.D.

WILEY

## Developing Analytic Talent: Becoming a Data Scientist

Published by  
John Wiley & Sons, Inc.  
10475 Crosspoint Boulevard  
Indianapolis, IN 46256  
[www.wiley.com](http://www.wiley.com)

Copyright © 2014 by John Wiley & Sons, Inc., Indianapolis, Indiana

Published simultaneously in Canada

ISBN: 978-1-118-81008-8  
ISBN: 978-1-118-81004-0 (ebk)  
ISBN: 978-1-118-81009-5 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Limit of Liability/Disclaimer of Warranty:** The publisher and the author make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained herein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publisher nor the author shall be liable for damages arising herefrom. The fact that an organization or Web site is referred to in this work as a citation and/or a potential source of further information does not mean that the author or the publisher endorses the information the organization or website may provide or recommendations it may make. Further, readers should be aware that Internet websites listed in this work may have changed or disappeared between when this work was written and when it is read.

For general information on our other products and services please contact our Customer Care Department within the United States at (877) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

**Library of Congress Control Number:** 2013958300

**Trademarks:** Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates, in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

*For my loving wife, Paris, my wonderful daughter, Capri, and son, Zinal, for their constant support. In loving memory of my father, Roger, who introduced me to mathematics when I was a kid.*







## About the Author

**Vincent Granville, Ph.D.**, is a visionary data scientist with 15 years of big data, predictive modeling, digital, and business analytics experience. Vincent is widely recognized as the leading expert in scoring technology, fraud detection, and web traffic optimization and growth. Over the last 10 years, he has worked in real-time credit card fraud detection with Visa, advertising mix optimization with CNET, change point detection with Microsoft, online user experience with Wells Fargo, search intelligence with InfoSpace, automated bidding with eBay, and click fraud detection with major search engines, ad networks, and large advertising clients. Vincent also manages the largest big data and analytics data science group on LinkedIn, with more than 100,000 members.

Most recently, Vincent launched Data Science Central, the leading community for big data, business analytics, and data science practitioners. Vincent is a former post-doctorate of Cambridge University and the National Institute of Statistical Sciences. He was among the finalists at the Wharton Business Plan Competition and at the Belgian Mathematical Olympiad. Vincent has published 40 papers in statistical journals and is an invited speaker at international conferences. He also developed a new data mining technology known as hidden decision trees, owns multiple patents, published the first data science book, and raised \$6M in start-up funding. Vincent is one of the top 20 big data influencers according to Forbes, and was featured on VentureBeat, MarketWatch, and CNN. Vincent can be reached on Twitter @Analyticbridge.



## About the Technical Editor

**Joni Ngai** is a Digital Evangelist who works with senior executives at Fortune 500 companies to develop digital vision and leverage technology and data to intelligently engage customers in today's connected world. She has extensive experience leading agencies and customers to develop new practices across digital, CRM, online media, analytics, and technology development. Joni started her digital consulting career with Razorfish in New York in 2000. Since then, she has worked with a number of top digital agencies, such as MRM Worldwide and Havas Digital, across the Asia-Pacific region for many global brands like Intel, Microsoft, and P&G. Joni was appointed as Vice-Chair at I-COM China, an industry-backed global forum in digital measurement, to facilitate participation in standardizing online measurements to help industries grow.

Joni graduated from the University of Waterloo majoring in Electrical Engineering with an option in Management Science. She also received an Executive MBA degree from the Kellogg School of Management at Northwestern University and Hong Kong University of Science and Technology. Joni also teaches graduate courses for the Master of Science in New Media program at the Chinese University of Hong Kong.



**Executive Editor**

Carol Long

**Project Editor**

Christina Haviland

**Technical Editor**

Joni Ngai

**Production Editor**

Christine Mugnolo

**Copy Editor**

San Dee Phillips

**Editorial Manager**

Mary Beth Wakefield

**Freelancer Editorial Manager**

Rosemarie Graham

**Associate Director of Marketing**

David Mayhew

**Marketing Manager**

Ashley Zurcher

**Business Manager**

Amy Knies

**Vice President and  
Executive Group Publisher**

Richard Swadley

**Associate Publisher**

Jim Minatel

**Project Coordinator, Cover**

Todd Klemme

**Composer**

Cody Gates,

Happenstance Type-O-Rama

**Proofreaders**

Shannon Coohill, Word One New York

Louise Watson, Word One New York

Candace Cunningham

**Indexer**

Robert Swanson

**Cover Designer**

Ryan Sneed/Wiley





# Acknowledgments

I would like to thank Chris Haviland and Carol Long from Wiley for making this book happen; for taking the risk; and for turning a vast amount of valuable but unstructured and scattered text, published online, into a coherent, comprehensive, and useful book. In many ways, this complex process is similar to turning unstructured into structured data, a challenge many data scientists face on a regular basis, with solutions provided in this book. Also, I would like to thank my business partner and co-founder, Tim Matteson, who helped grow Data Science Central to a point where it has become not only the leading data science community, but also a modern, lean start-up focused on delivering value. Finally, I want to thank all the members of our community for their numerous comments and support. Without them, this book would not exist.





# Contents at a Glance

Introduction		xxi
<b>Chapter 1</b>	<b>What Is Data Science?</b>	<b>1</b>
<b>Chapter 2</b>	<b>Big Data Is Different</b>	<b>41</b>
<b>Chapter 3</b>	<b>Becoming a Data Scientist</b>	<b>73</b>
<b>Chapter 4</b>	<b>Data Science Craftsmanship, Part I</b>	<b>109</b>
<b>Chapter 5</b>	<b>Data Science Craftsmanship, Part II</b>	<b>151</b>
<b>Chapter 6</b>	<b>Data Science Application Case Studies</b>	<b>195</b>
<b>Chapter 7</b>	<b>Launching Your New Data Science Career</b>	<b>255</b>
<b>Chapter 8</b>	<b>Data Science Resources</b>	<b>287</b>
Index		299







# Contents

<b>Introduction</b>	<b>xxi</b>
<b>Chapter 1 What Is Data Science?</b>	<b>1</b>
Real Versus Fake Data Science	2
Two Examples of Fake Data Science	5
The Face of the New University	6
The Data Scientist	9
Data Scientist Versus Data Engineer	9
Data Scientist Versus Statistician	11
Data Scientist Versus Business Analyst	12
Data Science Applications in 13 Real-World Scenarios	13
Scenario 1: DUI Arrests Decrease After End of State Monopoly on Liquor Sales	14
Scenario 2: Data Science and Intuition	15
Scenario 3: Data Glitch Turns Data Into Gibberish	18
Scenario 4: Regression in Unusual Spaces	19
Scenario 5: Analytics Versus Seduction to Boost Sales	20
Scenario 6: About Hidden Data	22
Scenario 7: High Crime Rates Caused by Gasoline Lead. Really?	23
Scenario 8: Boeing Dreamliner Problems	23
Scenario 9: Seven Tricky Sentences for NLP	24
Scenario 10: Data Scientists Dictate What We Eat?	25
Scenario 11: Increasing Amazon.com Sales with Better Relevancy	27
Scenario 12: Detecting Fake Profiles or Likes on Facebook	29
Scenario 13: Analytics for Restaurants	30

	Data Science History, Pioneers, and Modern Trends	30
	Statistics Will Experience a Renaissance	31
	History and Pioneers	32
	Modern Trends	34
	Recent Q&A Discussions	35
	Summary	39
<b>Chapter 2</b>	<b>Big Data Is Different</b>	<b>41</b>
	Two Big Data Issues	41
	The Curse of Big Data	41
	When Data Flows Too Fast	45
	Examples of Big Data Techniques	51
	Big Data Problem Epitomizing the Challenges of Data Science	51
	Clustering and Taxonomy Creation for Massive Data Sets	53
	Excel with 100 Million Rows	57
	What MapReduce Can't Do	60
	The Problem	61
	Three Solutions	61
	Conclusion: When to Use MapReduce	63
	Communication Issues	63
	Data Science: The End of Statistics?	65
	The Eight Worst Predictive Modeling Techniques	65
	Marrying Computer Science, Statistics, and Domain Expertise	67
	The Big Data Ecosystem	70
	Summary	71
<b>Chapter 3</b>	<b>Becoming a Data Scientist</b>	<b>73</b>
	Key Features of Data Scientists	73
	Data Scientist Roles	73
	Horizontal Versus Vertical Data Scientist	75
	Types of Data Scientists	78
	Fake Data Scientist	78
	Self-Made Data Scientist	78
	Amateur Data Scientist	79
	Extreme Data Scientist	80
	Data Scientist Demographics	82
	Training for Data Science	82
	University Programs	82
	Corporate and Association Training Programs	86
	Free Training Programs	87
	Data Scientist Career Paths	89
	The Independent Consultant	89
	The Entrepreneur	95
	Summary	107

---

<b>Chapter 4</b>	<b>Data Science Craftsmanship, Part I</b>	<b>109</b>
	New Types of Metrics	110
	Metrics to Optimize Digital Marketing Campaigns	111
	Metrics for Fraud Detection	112
	Choosing Proper Analytics Tools	113
	Analytics Software	114
	Visualization Tools	115
	Real-Time Products	116
	Programming Languages	117
	Visualization	118
	Producing Data Videos with R	118
	More Sophisticated Videos	122
	Statistical Modeling Without Models	122
	What Is a Statistical Model Without Modeling?	123
	How Does the Algorithm Work?	124
	Source Code to Produce the Data Sets	125
	Three Classes of Metrics: Centrality, Volatility, Bumpiness	125
	Relationships Among Centrality, Volatility, and Bumpiness	125
	Defining Bumpiness	126
	Bumpiness Computation in Excel	127
	Uses of Bumpiness Coefficients	128
	Statistical Clustering for Big Data	129
	Correlation and R-Squared for Big Data	130
	A New Family of Rank Correlations	132
	Asymptotic Distribution and Normalization	134
	Computational Complexity	137
	Computing $q(n)$	137
	A Theoretical Solution	140
	Structured Coefficient	140
	Identifying the Number of Clusters	141
	Methodology	142
	Example	143
	Internet Topology Mapping	143
	Securing Communications: Data Encoding	147
	Summary	149
<b>Chapter 5</b>	<b>Data Science Craftsmanship, Part II</b>	<b>151</b>
	Data Dictionary	152
	What Is a Data Dictionary?	152
	Building a Data Dictionary	152
	Hidden Decision Trees	153
	Implementation	155
	Example: Scoring Internet Traffic	156
	Conclusion	158

Model-Free Confidence Intervals	158
Methodology	158
The Analyticbridge First Theorem	159
Application	160
Source Code	160
Random Numbers	161
Four Ways to Solve a Problem	163
Intuitive Approach for Business Analysts with Great Intuitive Abilities	164
Monte Carlo Simulations Approach for Software Engineers	165
Statistical Modeling Approach for Statisticians	165
Big Data Approach for Computer Scientists	165
Causation Versus Correlation	165
How Do You Detect Causes?	166
Life Cycle of Data Science Projects	168
Predictive Modeling Mistakes	171
Logistic-Related Regressions	172
Interactions Between Variables	172
First Order Approximation	172
Second Order Approximation	174
Regression with Excel	175
Experimental Design	176
Interesting Metrics	176
Segmenting the Patient Population	176
Customized Treatments	177
Analytics as a Service and APIs	178
How It Works	179
Example of Implementation	179
Source Code for Keyword Correlation API	180
Miscellaneous Topics	183
Preserving Scores When Data Sets Change	183
Optimizing Web Crawlers	184
Hash Joins	186
Simple Source Code to Simulate Clusters	186
New Synthetic Variance for Hadoop and Big Data	187
Introduction to Hadoop/MapReduce	187
Synthetic Metrics	188
Hadoop, Numerical, and Statistical Stability	189
The Abstract Concept of Variance	189
A New Big Data Theorem	191
Transformation-Invariant Metrics	192
Implementation: Communications Versus Computational Costs	193
Final Comments	193
Summary	193

<b>Chapter 6</b>	<b>Data Science Application Case Studies</b>	<b>195</b>
	Stock Market	195
	Pattern to Boost Return by 500 Percent	195
	Optimizing Statistical Trading Strategies	197
	Stock Trading API: Statistical Model	200
	Stock Trading API: Implementation	202
	Stock Market Simulations	203
	Some Mathematics	205
	New Trends	208
	Encryption	209
	Data Science Application: Steganography	209
	Solid E-Mail Encryption	212
	Captcha Hack	214
	Fraud Detection	216
	Click Fraud	216
	Continuous Click Scores Versus Binary Fraud/Non-Fraud	218
	Mathematical Model and Benchmarking	219
	Bias Due to Bogus Conversions	220
	A Few Misconceptions	221
	Statistical Challenges	221
	Click Scoring to Optimize Keyword Bids	222
	Automated, Fast Feature Selection with Combinatorial	
	Optimization	224
	Predictive Power of a Feature: Cross-Validation	225
	Association Rules to Detect Collusion and Botnets	228
	Extreme Value Theory for Pattern Detection	229
	Digital Analytics	230
	Online Advertising: Formula for Reach and Frequency	231
	E-Mail Marketing: Boosting Performance by 300 Percent	231
	Optimize Keyword Advertising Campaigns in 7 Days	232
	Automated News Feed Optimization	234
	Competitive Intelligence with Bit.ly	234
	Measuring Return on Twitter Hashtags	237
	Improving Google Search with Three Fixes	240
	Improving Relevancy Algorithms	242
	Ad Rotation Problem	244
	Miscellaneous	245
	Better Sales Forecasts with Simpler Models	245
	Better Detection of Healthcare Fraud	247
	Attribution Modeling	248
	Forecasting Meteorite Hits	248
	Data Collection at Trailhead Parking Lots	252
	Other Applications of Data Science	253
	Summary	253

<b>Chapter 7</b>	<b>Launching Your New Data Science Career</b>	<b>255</b>
	Job Interview Questions	255
	Questions About Your Experience	255
	Technical Questions	257
	General Questions	258
	Questions About Data Science Projects	260
	Testing Your Own Visual and Analytic Thinking	263
	Detecting Patterns with the Naked Eye	263
	Identifying Aberrations	266
	Misleading Time Series and Random Walks	266
	From Statistician to Data Scientist	268
	Data Scientists Are Also Statistical Practitioners	268
	Who Should Teach Statistics to Data Scientists?	269
	Hiring Issues	269
	Data Scientists Work Closely with Data Architects	270
	Who Should Be Involved in Strategic Thinking?	270
	Two Types of Statisticians	271
	Using Big Data Versus Sampling	272
	Taxonomy of a Data Scientist	273
	Data Science's Most Popular Skill Mixes	273
	Top Data Scientists on LinkedIn	276
	400 Data Scientist Job Titles	279
	Salary Surveys	281
	Salary Breakdown by Skill and Location	281
	Create Your Own Salary Survey	285
	Summary	285
<b>Chapter 8</b>	<b>Data Science Resources</b>	<b>287</b>
	Professional Resources	287
	Data Sets	288
	Books	288
	Conferences and Organizations	290
	Websites	291
	Definitions	292
	Career-Building Resources	295
	Companies Employing Data Scientists	296
	Sample Data Science Job Ads	297
	Sample Resumes	297
	Summary	298
<b>Index</b>		<b>299</b>



# Introduction

This book is a type of “handbook” on data science and data scientists, and contains information not found in traditional statistical, programming, or computer science textbooks. The author has compiled what he considers some of the most important information you will need for a career in data science, based on his 20+ years as a leader in the field. Much of the text was initially published over the last three years on the Data Science Central website, which is read by millions of website visitors. The book shows how data science is different from related fields and the value it brings to organizations using big data.

This book has three components: a multi-layer discussion of what data science is and how it relates to other disciplines; technical applications of and for data science including tutorials and case studies; and career resources for practicing and aspiring data scientists. Numerous career and training resources are included (such as data sets, web crawler source code, data videos, and how to build APIs) so you can start practicing data science today and quickly boost your career. If you’re a decision maker, you will find information to help you make decisions on how to build a better analytic team, whether and when you need specialized solutions, and which ones will work best for your need.

## Who This Book Is For

---

This book is intended for data scientists and related professionals (such as business analysts, computer scientists, software engineers, data engineers, and statisticians) who are interested in shifting to big data science careers. It is also for the college student studying a quantitative curriculum with the goal of becoming a data scientist. Finally, it is for managers of data scientists, and people interested in creating a startup business or consultancy around data science.

These readers will find valuable information throughout the book, and specifically in the following chapters:

- **Data science practitioners** will find Chapters 2, 4, 5, and 6 particularly valuable because they contain material on big data techniques (clustering and taxonomy creation) and modern data science techniques such as combinatorial feature selection, hidden decision trees, analytic APIs, and when MapReduce is useful. A number of case studies (fraud detection, digital analytics, stock market strategies, and more) are detailed enough to allow the reader to replicate the analyses when facing similar data in the real world when doing their jobs. However, it is also explained in simple words, not spending too much time on technicalities, code, or formulas, to make it accessible to high level managers.
- **Students** attending computer science, data science, or MBA classes will find Chapters 2, 4, 5, and 6 valuable for their purposes. In particular, they will find more advanced material in Chapters 2, 4, and 5, such as practical data science methods and principles, most of it not found in textbooks or taught in typical college curricula. Chapter 6 also provides real life applications and case studies, including more in-depth technical details.
- **Job applicants** will find resources about data science training and programs in Chapter 3. Chapters 7 and 8 provide numerous resources for job seekers including interview questions, sample resumes, sample job ads, a list of companies that routinely hire data scientists, and salary surveys.
- **Entrepreneurs** who want to launch a data science startup or consultancy will find sample business proposals, startup ideas, and salary surveys for consultants in Chapter 3. Also, throughout the book, consultants will find discussions on improving communication in data science work, lifecycles of data science projects, book and conference references, and many other resources.
- **Executives** trying to assess the value of data science, where it most benefits enterprise projects, and when architectures such as MapReduce are useful will find valuable information in Chapters 1, 2, 6 (case studies), and 8 (sample job ads, resumes, salary surveys). The focus of these chapters is usually not technical, except, to a limited extent, in some parts of Chapters 2 and 6, where new analytic technologies are introduced.

## What This Book Covers

---

The technical part of this book covers core data science topics, including:

- Big data and the challenges of applying traditional algorithms to big data (Solutions are provided, for instance in the context of big data clustering or taxonomy creation.)



- A new, simplified, data science–friendly approach to statistical science, focusing on robust model-free methods
- State-of-the-art machine learning (hidden decision trees and combinatorial feature selection)
- New metrics for modern data (synthetic metrics, predictive power, bumpiness coefficient)
- Elements of computer science that are needed to build fast algorithms
- MapReduce and Hadoop, including numerical stability of computations performed with Hadoop

The focus is on recent technology. You will not find material about old techniques such as linear regression (except for anecdotal references), since such are discussed at length in all standard books. There is some limited discussion on logistic-like regression in this book, but it's more about blending it with other classifiers and proposing a numerically stable, approximate algorithm (approximate solutions are often as good as the exact model, since no data fits perfectly with a theoretical model).

Besides technology, the book provides useful career resources, including job interview questions, sample resumes, and sample job ads. Another important component of this book is case studies. Some of the case studies included here have a statistical/machine learning flair, some have more of a business/decision science or operations research flair, and some have more of a data engineering flair. Most of the time, I have favored topics that were posted recently and very popular on Data Science Central (the leading community for data scientists), rather than topics that I am particularly attached to.

## How This Book Is Structured

---

The book consists of three overall topics:

- What data science and big data is, and is not, and how it's different from other disciplines (Chapters 1, 2, and 3)
- Career and training resources (Chapters 3 and 8)
- Technical material presented as tutorials (Chapters 4 and 5, but also the section on Clustering and Taxonomy Creation for Massive Data Sets in Chapter 2, and the section on New Variance for Hadoop and Big Data in Chapter 8), and in case studies (Chapters 6 and 7)

The book provides valuable career resources for potential and existing data scientists and related professionals (and their managers and their bosses), and generally speaking, to all professionals dealing with increasingly bigger, more complex, and faster flowing data. The book also provides data science recipes,

craftsmanship, concepts (many of them, original and published for the first time), and cases studies illustrating implementation methods and techniques that have proven successful in various domains for analyzing modern data — either manually or automatically.

## What You Need to Use This Book

---

The book contains a small amount of sample code, either in R or Perl. You can download Perl from <http://www.activestate.com/activeperl/downloads> and R from <http://cran.r-project.org/bin/windows/base/>. If you use a Windows machine, I would first install Cygwin, a Linux-like environment for Windows. You can get Cygwin at <http://cygwin.com/install.html>. Python is also available as open source and has a useful library called Pandas.

For most of the book, one or two years of college with some basic quantitative courses is enough for you to understand the content. The book does not require calculus or advanced math — indeed, it barely contains any mathematical formulas or symbols.

Yet some quite advanced material is described at a high level. A few technical notes spread throughout the book are for those who are more mathematically inclined and interested in digging deeper. Two years of calculus, statistics, and matrix theory at the college level are needed to understand these technical notes. Some source code (R, Perl) and data sets are provided, but the emphasis is not on coding.

This mixture of technical levels offers the opportunity for you to explore the depths of data science without advanced math knowledge (a bit like the way Carl Sagan introduced astronomy to the mainstream public).

## Conventions

---

To help you get the most from the text and keep track of what’s happening, we’ve used a number of conventions throughout the book.

**NOTE** Notes, tips, cross-references, and asides to the current discussion are offset and placed in features like this.

As for styles in the text:

- We *highlight* new terms and important words when we introduce them.
- We show keyboard strokes like this: Ctrl+A.

- We show filenames, URLs, and code within the text like so:

`persistence.properties`.

- We present code like this:

We use a monofont type with no highlighting for most code examples.



# What Is Data Science?

Sometimes, understanding what something *is* includes having a clear picture of what it *is not*. Understanding data science is no exception. Thus, this chapter begins by investigating what data science is not, because the term has been much abused and a lot of hype surrounds big data and data science. You will first consider the difference between true data science and fake data science. Next, you will learn how new data science training has evolved from traditional university degree programs. Then you will review several examples of how modern data science can be used in real-world scenarios.

Finally, you will review the history of data science and its evolution from computer science, business optimization, and statistics into modern data science and its trends. At the end of the chapter, you will find a Q&A section from recent discussions I've had that illustrate the conflicts between data scientists, data architects, and business analysts.

This chapter asks more questions than it answers, but you will find the answers discussed in more detail in subsequent chapters. The purpose of this approach is for you to become familiar with how data scientists think, what is important in the big data industry today, what is becoming obsolete, and what people interested in a data science career don't need to learn. For instance, you need to know statistics, computer science, and machine learning, but not everything from these domains. You don't need to know the details about complexity of sorting algorithms (just the general results), and you don't need to know how

to compute a generalized inverse matrix, nor even know what a generalized inverse matrix is (a core topic of statistical theory), unless you specialize in the numerical aspects of data science.

**TECHNICAL NOTE**

This chapter can be read by anyone with minimal mathematical or technical knowledge. More advanced information is presented in “Technical Notes” like this one, which may be skipped by non-mathematicians.

**CROSS-REFERENCE** You will find definitions of most terms used in this book in Chapter 8.

---

## Real Versus Fake Data Science

---

Books, certificates, and graduate degrees in data science are spreading like mushrooms after the rain. Unfortunately, many are just a mirage: people taking advantage of the new paradigm to quickly repackage old material (such as statistics and R programming) with the new label “data science.”

Expanding on the R programming example of fake data science, note that R is an open source statistical programming language and environment that is at least 20 years old, and is the successor of the commercial product S+. R was and still is limited to in-memory data processing and has been very popular in the statistical community, sometimes appreciated for the great visualizations that it produces. Modern environments have extended R capabilities (the in-memory limitations) by creating libraries or integrating R in a distributed architecture, such as RHadoop (R + Hadoop). Of course other languages exist, such as SAS, but they haven’t gained as much popularity as R. In the case of SAS, this is because of its high price and the fact that it was more popular in government organizations and brick-and-mortar companies than in the fields that experienced rapid growth over the last 10 years, such as digital data (search engine, social, mobile data, collaborative filtering). Finally, R is not unlike the C, Perl, or Python programming languages in terms of syntax (they all share the same syntax roots), and thus it is easy for a wide range of programmers to learn. It also comes with many libraries and a nice user interface. SAS, on the other hand, is more difficult to learn.

To add to the confusion, executives and decision makers building a new team of data scientists sometimes don’t know exactly what they are looking for, and they end up hiring pure tech geeks, computer scientists, or people lacking proper big data experience. The problem is compounded by Human Resources