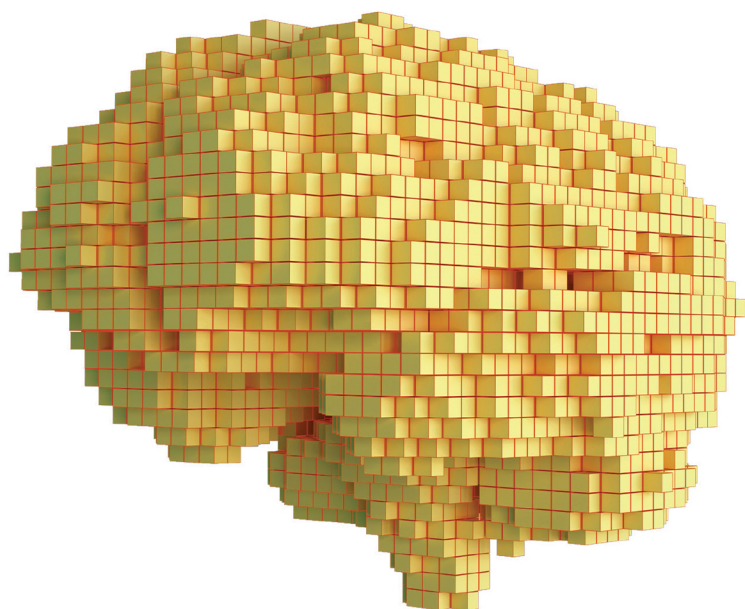


ALEX J. GUTMAN AND JORDAN GOLDMEIER



**BECOMING A**  
**DATA**  
**HEAD**

HOW TO **THINK, SPEAK, AND**  
**UNDERSTAND DATA SCIENCE,**  
**STATISTICS, AND MACHINE LEARNING**



## PRAISE FOR *BECOMING A DATA HEAD*

Big Data, Data Science, Machine Learning, Artificial Intelligence, Neural Networks, Deep Learning . . . It can be buzzword bingo, but make no mistake, everything is becoming “data-fied” and an understanding of data problems and the data science toolset is becoming a requirement for every business person. Alex and Jordan have put together a must read whether you are just starting your journey or already in the thick of it. They made this complex space simple by breaking down the “data process” into understandable patterns and using everyday examples and events over our history to make the concepts relatable.

—**Milen Mahadevan**, President of 84.51°

What I love about this book is its remarkable breadth of topics covered, while maintaining a healthy depth in the content presented for each topic. I believe in the pedagogical concept of “Talking the Walk,” which means being able to explain the hard stuff in terms that broad audiences can grasp. Too many data science books are either too specialized in taking you down the deep paths of mathematics and coding (“Walking the Walk”) or too shallow in over-hyping the content with a plethora of shallow buzzwords (“Talking the Talk”). You can take a great walk down the pathways of the data field in Alex and Jordan’s without fear of falling off the path. The journey and destination are well worth the trip, and the talk.

—**Kirk Borne**, Data Scientist,  
Top Worldwide Influencer in Data Science

The most clear, concise, and practical characterization of working in corporate analytics that I’ve seen. If you want to be a killer analyst and ask the right questions, this is for you.

—**Kristen Kehrner**, Data Moves Me, LLC,  
LinkedIn Top Voices in Data Science & Analytics

THE book that business and technology leaders need to read to fully understand the potential, power, AND limitations of data science.

—**Jennifer L. L. Morgan**, PhD, Analytical Chemist at  
Procter and Gamble

You’ve heard it before: “We need to be doing more machine learning. Why aren’t we doing more sophisticated data science work?” Data science isn’t the magic unicorn that will solve all of your company’s problems. *Data Head* brings this idea to life by highlighting when data science is (and isn’t) the right approach and the common pitfalls to watch out for, explaining it all in a way that a data novice can understand. This book will be my new “pocket reference” when communicating complicated concepts to non-technically trained leaders.

—**Sandy Steiger**, Director, Center for Analytics and  
Data Science at Miami University

Individuals and organizations want to be data driven. They say they are data driven. *Becoming a Data Head* shows them how to actually become data driven, without the assumption of a statistics or data background. This book is for anyone, or any organization, asking how to bring a data mindset to the whole company, not just those trained in the space.

—**Eric Weber**, Head of Experimentation & Metrics Research, Yelp

What is keeping data science from reaching its true potential? It is not slow algorithms, lack of data, lack of computing power, or even lack of data scientists. *Becoming a Data Head* tackles the biggest impediment to data science success—the communication gap between the data scientist and the executive. Gutman and Goldmeier provide creative explanations of data science techniques and how they are used with clear everyday relatable examples. Managers and executives, and anyone wanting to better understand data science will learn a lot from this book. Likewise, data scientists who find it challenging to explain what they are doing will also find great value in *Becoming a Data Head*.

—**Jeffrey D. Camm**, PhD, Center for Analytics Impact,  
Wake Forest University

*Becoming a Data Head* raises the level of education and knowledge in an industry desperate for clarity in thinking. A must read for those working with and within the growing field of data science and analytics.

—**Dr. Stephen Chambal**, VP for Corporate Growth at  
Perduco (DoD Analytics Company)

Gutman and Goldmeier filter through much of the noise to break down complex data and statistical concepts we hear today into basic examples and analogies that stick. *Becoming a Data Head* has enabled me to translate my team's data needs into more tangible business requirements that make sense for our organization. A great read if you want to communicate your data more effectively to drive your business and data science team forward!

—**Justin Maurer**, Engineering and Data Science Manager at Google

As an aerospace engineer with nearly 15 years experience, *Becoming a Data Head* made me aware of not only what I personally want to learn about data science, but also what I need to know professionally to operate in a data-rich environment. This book further discusses how to filter through often overused terms like artificial intelligence. This is a book for every mid-level program manager learning how to navigate the inevitable future of data science.

—**Josh Keener**, Aerospace Engineer and Program Manager

A must read for an in-depth understanding of data science for senior executives.

—**Cade Saie**, PhD, Chief Data Officer

Gutman and Goldmeier offer practical advice for asking the right questions, challenging assumptions, and avoiding common pitfalls. They strike a nice balance between thoroughly explaining concepts of data science while not getting lost in the weeds. This book is a useful addition to the toolbox of any analyst, data scientist, manager, executive, or anyone else who wants to become more comfortable with data science.

—**Jeff Bialac**, Senior Supply Chain Analyst at Kroger

Gutman and Goldmeier have written a book that is as useful for applied statisticians and data scientists as it is for business leaders and technical professionals. In demystifying these complex statistical topics, they have also created a common language that bridges the longstanding communication divide that has — until now — separated data work from business value.

—**Kathleen Maley**, Chief Analytics Officer at datazoom

# Becoming a Data Head



# **Becoming a Data Head**

**How to Think, Speak,  
and Understand  
Data Science, Statistics,  
and Machine Learning**

ALEX J. GUTMAN  
JORDAN GOLDMEIER

**WILEY**

Copyright © 2021 by John Wiley & Sons, Inc., Indianapolis, Indiana

Published simultaneously in Canada

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 [www.wiley.com/go/permissions](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 website 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 [booksupport.wiley.com](http://booksupport.wiley.com). For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

**Library of Congress Control Number:** 2021934226

**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.

ISBN: 978-1-119-74174-9

ISBN: 978-1-119-74176-3 (ebk)

ISBN: 978-1-119-74171-8 (ebk)

Manufactured in the United States of America



*For my children Allie, William, and Ellen.*

*Allie was three when she discovered dad was a “doctor.”  
Puzzled, she looked at me and said, “But, you don’t help people. . . .”  
In that spirit, I also dedicate this book to you, the reader.*

*I hope this helps you.*

*—Alex*

*For Stephen and Melissa*

*—Jordan*



# About the Authors

**Alex J. Gutman** is a data scientist, corporate trainer, Fulbright Specialist grant recipient, and Accredited Professional Statistician® who enjoys teaching a wide variety of data science topics to technical and non-technical audiences. He earned his Ph.D. in applied math from the Air Force Institute of Technology where he currently serves as an adjunct professor.

**Jordan Goldmeier** is an internationally recognized analytics professional and data visualization expert, author, and speaker. A former chief operations officer at Excel.TV, he has spent years in the data training trenches. He is the author of *Advanced Excel Essentials* and *Dashboards for Excel*. His work has been cited by and quoted in the Associated Press, Bloomberg BusinessWeek, and American Express OPEN Forum. He is currently an Excel MVP Award holder, an achievement he's held for six years, allowing him to provide feedback and direction to Microsoft product teams. He once used Excel to save the Air Force \$60 million. He is also a volunteer Emergency Medical Technician.



# About the Technical Editors

**William A. Brenneman** is a Research Fellow and the Global Statistics Discipline Leader at Procter & Gamble in the Data and Modeling Sciences Department and an Adjunct Professor of Practice at Georgia Tech in the Stewart School of Industrial and Systems Engineering. Since joining P&G, he has worked on a wide range of projects that deal with statistics applications in his areas of expertise: design and analysis of experiments, robust parameter design, reliability engineering, statistical process control, computer experiments, machine learning, and statistical thinking. He was also instrumental in the development of an in-house statistics curriculum. He received a Ph.D. in Statistics from the University of Michigan, an MS in Mathematics from the University of Iowa, and a BA in Mathematics and Secondary Education from Tabor College. William is a Fellow in both the American Statistical Association (ASA) and the American Society for Quality (ASQ). He has served as ASQ Statistics Division Chair, ASA Quality and Productivity Section Chair, and as Associate Editor for *Technometrics*. William also has seven years of experience as an educator at the high school and college level.

**Jennifer Stirrup** is the Founder and CEO of Data Relish, a UK-based AI and Business Intelligence leadership boutique consultancy delivering data strategy and business-focused solutions. Jen is a recognized leading authority in AI and Business Intelligence Leadership, a Fortune 100 global speaker, and has been named as one of the Top 50 Global Data Visionaries, one of the Top Data Scientists to follow on Twitter, and one of the most influential Top 50 Women in Technology worldwide.

Jen has clients in 24 countries on 5 continents, and she holds postgraduate degrees in AI and Cognitive Science. Jen has authored books on data and artificial intelligence and has been featured on CBS Interactive and the BBC as well as other well-known podcasts, such as Digital Disrupted, Run As Radio, and her own Make Your Data Work webinar series.

Jen has also given keynotes for colleges and universities, as well as donated her expertise to charities and non-profits as a Non-Executive Director. All of Jen's keynotes are based on her 20+ years of global experience, dedication, and hard work.



# Acknowledgments

I've noticed a trend in acknowledgment sections—the author's spouse is often mentioned at the end. I suppose it's a saving-the-best-for-last gesture, but I promised my wife if I ever wrote a book, I'd mention her first to make it perfectly clear whose contributions mattered most to me. So, to my wife Erin, thank you for your love, encouragement, and smile. As I write this, you are taking our three young children on a bike ride, giving me time to write one final page. (I assure all readers this act is a representative sample of our lives this past year.)

I'd also like to thank my parents, Ed and Nancy, for being the best cheerleaders in whatever I do and for showing me what being a good parent looks like, and to my siblings Ryan, Ross, and Erin for their support.

This book is the culmination of many discussions with friends and colleagues, ranging from whether I should attempt to write a book about data literacy to potential topics that should appear in it. Thank you especially to Altynbek Ismailov, Andy Neumeier, Bradley Boehmke, Brandon Greenwell, Brent Russell, Cade Saie, Caleb Goodreau, Carl Parson, Daniel Uppenkamp, Douglas Clarke, Greg Anderson, Jason Freels, Joel Chaney, Joseph Keller, Justin Maurer, Nathan Swigart, Phil Hartke, Samuel Reed, Shawn Schneider, Stephen Ferro, and Zachary Allen.

I'm also indebted to the hundreds of engineers, business professionals, and data scientists I've interacted with, personally or online, who've taught me how to be a better data scientist and communicator. And to my "students" (colleagues) who have given candid feedback about the courses I've taught, I heard you and I thank you.

I'm fortunate to have many academic and professional mentors who've given me numerous opportunities to find my voice and confidence as a statistician, data scientist, and trainer. Thank you to Jeffery Weir, John Tudorovic, K. T. Arasu, Raymond Hill, Rob Baker, Scott Crawford, Stephen Chambal, Tony White, and William Brennehan (who kindly served as a technical editor on this book). It's impossible not to become wiser hanging around a group like that.

Thanks to the team at Wiley: Jim Minatel for believing in the project and giving us a chance, Pete Gaughan and John Sleeva for guiding us through the process, and the production staff at Wiley for meticulously combing

through our chapters. And to our technical editors, William Brenneman and Jen Stirrup, we appreciate your suggestions and expertise. The book is better because of you.

Special thanks to my coauthor Jordan Goldmeier, for one obvious reason (the book in your hands) and one not so obvious. Early in my career, I complained to Jordan that people didn't share my interest in statistics and statistical thinking. He said if I'm bothered by it, then it's my obligation to change it. I've been working to fulfill that obligation ever since.

Finally, I'd like to thank my wife Erin one final time (because you've got to save the best for last).

—Alex

I would like to acknowledge the many people who brought this book together.

First, and foremost, I would like to acknowledge my coauthor-in-crime, Alex Gutman. For years, we discussed writing a book together. When the moment was right, we pulled the trigger. I couldn't have asked for a better coauthor.

Thanks to the wonderful folks at Wiley who helped put this together, including acquisition editor Jim Minatel, and project editor John Sleeva. Also, I would like to acknowledge our technical editors, William Brenneman and Jen Stirrup for your hard work reviewing the book. We took your comments to heart.

Last but not least, thank you to my partner, Katie Gray, who always believed in this project—and me.

—Jordan



# Contents

<b>Acknowledgments</b>	<b>xiii</b>
<b>Foreword</b>	<b>xxiii</b>
<b>Introduction</b>	<b>xxvii</b>

## **PART ONE**

---

### **Thinking Like a Data Head**

#### **CHAPTER 1**

#### **What Is the Problem? 3**

Questions a Data Head Should Ask	4
<i>Why Is This Problem Important?</i>	4
<i>Who Does This Problem Affect?</i>	6
<i>What If We Don't Have the Right Data?</i>	6
<i>When Is the Project Over?</i>	7
<i>What If We Don't Like the Results?</i>	7
Understanding Why Data Projects Fail	8
<i>Customer Perception</i>	8
<i>Discussion</i>	10
Working on Problems That Matter	11
Chapter Summary	11

#### **CHAPTER 2**

#### **What Is Data? 13**

Data vs. Information	13
<i>An Example Dataset</i>	14
Data Types	15
How Data Is Collected and Structured	16
<i>Observational vs. Experimental Data</i>	16
<i>Structured vs. Unstructured Data</i>	17
Basic Summary Statistics	18
Chapter Summary	19

<b>CHAPTER 3</b>	
<b>Prepare to Think Statistically</b>	<b>21</b>
Ask Questions	22
There Is Variation in All Things	23
<i>Scenario: Customer Perception (The Sequel)</i>	24
<i>Case Study: Kidney-Cancer Rates</i>	26
Probabilities and Statistics	28
<i>Probability vs. Intuition</i>	29
<i>Discovery with Statistics</i>	31
Chapter Summary	33

## **PART TWO**

---

### **Speaking Like a Data Head**

<b>CHAPTER 4</b>	
<b>Argue with the Data</b>	<b>37</b>
What Would You Do?	38
<i>Missing Data Disaster</i>	39
Tell Me the Data Origin Story	43
<i>Who Collected the Data?</i>	44
<i>How Was the Data Collected?</i>	44
Is the Data Representative?	45
<i>Is There Sampling Bias?</i>	46
<i>What Did You Do with Outliers?</i>	46
What Data Am I Not Seeing?	47
<i>How Did You Deal with Missing Values?</i>	47
<i>Can the Data Measure What You Want It to Measure?</i>	48
Argue with Data of All Sizes	48
Chapter Summary	49

<b>CHAPTER 5</b>	
<b>Explore the Data</b>	<b>51</b>
Exploratory Data Analysis and You	52
Embracing the Exploratory Mindset	52
<i>Questions to Guide You</i>	53
<i>The Setup</i>	53
Can the Data Answer the Question?	54
<i>Set Expectations and Use Common Sense</i>	54
<i>Do the Values Make Intuitive Sense?</i>	54
<i>Watch Out: Outliers and Missing Values</i>	58

Did You Discover Any Relationships?	59
<i>Understanding Correlation</i>	59
<i>Watch Out: Misinterpreting Correlation</i>	60
<i>Watch Out: Correlation Does Not Imply Causation</i>	62
Did You Find New Opportunities in the Data?	63
Chapter Summary	63

**CHAPTER 6****Examine the Probabilities 65**

Take a Guess	66
The Rules of the Game	66
<i>Notation</i>	67
<i>Conditional Probability and Independent Events</i>	69
<i>The Probability of Multiple Events</i>	69
Two Things That Happen Together	69
One Thing or the Other	70
Probability Thought Exercise	72
<i>Next Steps</i>	73
Be Careful Assuming Independence	74
<i>Don't Fall for the Gambler's Fallacy</i>	74
All Probabilities Are Conditional	75
<i>Don't Swap Dependencies</i>	76
<i>Bayes' Theorem</i>	76
Ensure the Probabilities Have Meaning	79
<i>Calibration</i>	80
<i>Rare Events Can, and Do, Happen</i>	80
Chapter Summary	81

**CHAPTER 7****Challenge the Statistics 83**

Quick Lessons on Inference	83
<i>Give Yourself Some Wiggle Room</i>	84
<i>More Data, More Evidence</i>	84
<i>Challenge the Status Quo</i>	85
<i>Evidence to the Contrary</i>	86
<i>Balance Decision Errors</i>	88
The Process of Statistical Inference	89
The Questions You Should Ask to Challenge the Statistics	90
<i>What Is the Context for These Statistics?</i>	90
<i>What Is the Sample Size?</i>	91
<i>What Are You Testing?</i>	92

<i>What Is the Null Hypothesis?</i>	92
Assuming Equivalence	93
<i>What Is the Significance Level?</i>	93
<i>How Many Tests Are You Doing?</i>	94
<i>Can I See the Confidence Intervals?</i>	95
<i>Is This Practically Significant?</i>	96
<i>Are You Assuming Causality?</i>	96
Chapter Summary	97

## PART THREE

### Understanding the Data Scientist's Toolbox

#### CHAPTER 8

<b>Search for Hidden Groups</b>	<b>101</b>
Unsupervised Learning	102
Dimensionality Reduction	102
<i>Creating Composite Features</i>	103
Principal Component Analysis	105
<i>Principal Components in Athletic Ability</i>	105
<i>PCA Summary</i>	108
<i>Potential Traps</i>	109
Clustering	110
<i>k</i> -Means Clustering	111
<i>Clustering Retail Locations</i>	111
<i>Potential Traps</i>	113
Chapter Summary	114

#### CHAPTER 9

<b>Understand the Regression Model</b>	<b>117</b>
Supervised Learning	117
Linear Regression: What It Does	119
<i>Least Squares Regression: Not Just a Clever Name</i>	120
Linear Regression: What It Gives You	123
<i>Extending to Many Features</i>	124
Linear Regression: What Confusion It Causes	125
<i>Omitted Variables</i>	125
<i>Multicollinearity</i>	126
<i>Data Leakage</i>	127
<i>Extrapolation Failures</i>	128

<i>Many Relationships Aren't Linear</i>	128
<i>Are You Explaining or Predicting?</i>	128
<i>Regression Performance</i>	130
Other Regression Models	131
Chapter Summary	131

**CHAPTER 10****Understand the Classification Model 133**

Introduction to Classification	133
<i>What You'll Learn</i>	134
<i>Classification Problem Setup</i>	135
Logistic Regression	135
<i>Logistic Regression: So What?</i>	138
Decision Trees	139
Ensemble Methods	142
<i>Random Forests</i>	143
<i>Gradient Boosted Trees</i>	143
<i>Interpretability of Ensemble Models</i>	145
Watch Out for Pitfalls	145
<i>Misapplication of the Problem</i>	146
<i>Data Leakage</i>	146
<i>Not Splitting Your Data</i>	146
<i>Choosing the Right Decision Threshold</i>	147
Misunderstanding Accuracy	147
<i>Confusion Matrices</i>	148
Chapter Summary	150

**CHAPTER 11****Understand Text Analytics 151**

Expectations of Text Analytics	151
How Text Becomes Numbers	153
<i>A Big Bag of Words</i>	153
<i>N-Grams</i>	157
<i>Word Embeddings</i>	158
Topic Modeling	160
Text Classification	163
<i>Naïve Bayes</i>	164
<i>Sentiment Analysis</i>	166
Practical Considerations When Working with Text	167
<i>Big Tech Has the Upper Hand</i>	168
Chapter Summary	169

**CHAPTER 12****Conceptualize Deep Learning****171**

Neural Networks	172
<i>How Are Neural Networks Like the Brain?</i>	172
<i>A Simple Neural Network</i>	173
<i>How a Neural Network Learns</i>	174
<i>A Slightly More Complex Neural Network</i>	175
Applications of Deep Learning	178
<i>The Benefits of Deep Learning</i>	179
<i>How Computers “See” Images</i>	180
<i>Convolutional Neural Networks</i>	182
<i>Deep Learning on Language and Sequences</i>	183
Deep Learning in Practice	185
<i>Do You Have Data?</i>	185
<i>Is Your Data Structured?</i>	186
<i>What Will the Network Look Like?</i>	186
Artificial Intelligence and You	187
<i>Big Tech Has the Upper Hand</i>	188
<i>Ethics in Deep Learning</i>	189
Chapter Summary	190

**PART FOUR****Ensuring Success****CHAPTER 13****Watch Out for Pitfalls****193**

Biases and Weird Phenomena in Data	194
<i>Survivorship Bias</i>	194
<i>Regression to the Mean</i>	195
<i>Simpson’s Paradox</i>	195
<i>Confirmation Bias</i>	197
<i>Effort Bias (aka the “Sunk Cost Fallacy”)</i>	197
<i>Algorithmic Bias</i>	198
<i>Uncategorized Bias</i>	198
The Big List of Pitfalls	199
<i>Statistical and Machine Learning Pitfalls</i>	199
<i>Project Pitfalls</i>	200
Chapter Summary	202

---

<b>CHAPTER 14</b>	
<b>Know the People and Personalities</b>	<b>203</b>
Seven Scenes of Communication Breakdowns	204
<i>The Postmortem</i>	204
<i>Storytime</i>	205
<i>The Telephone Game</i>	206
<i>Into the Weeds</i>	206
<i>The Reality Check</i>	207
<i>The Takeover</i>	207
<i>The Blowhard</i>	208
Data Personalities	208
<i>Data Enthusiasts</i>	209
<i>Data Cynics</i>	209
<i>Data Heads</i>	209
Chapter Summary	210
<b>CHAPTER 15</b>	
<b>What's Next?</b>	<b>211</b>
<b>Index</b>	<b>215</b>





# Foreword

*B*ecoming a Data Head is well-timed for the current state of data and analytics within organizations. Let's quickly review some recent history. A few leading companies have made effective use of data and analytics to guide their decisions and actions for several decades, starting in the 1970s. But most ignored this important resource, or left it hiding in back rooms with little visibility or importance.

But in the early to mid-2000s this situation began to change, and companies began to get excited about the potential for data and analytics to transform their business situations. By the early 2010s, the excitement began to shift toward “big data,” which originally came from Internet companies but began to pop up across sophisticated economies. To deal with the increased volume and complexity of data, the “data scientist” role arose with companies—again, first in Silicon Valley, but then everywhere.

However, just as firms were beginning to adjust to big data, the emphasis shifted again—around about 2015 to 2018 in many firms—to a renewed focus on artificial intelligence. Collecting, storing, and analyzing big data gave way to machine learning, natural language processing, and automation.

Embedded within these rapid shifts in focus were a series of assumptions about data and analytics within organizations. I am happy to say that *Becoming a Data Head* violates many of them, and it's about time. As many who work with or closely observe these trends are beginning to admit, we have headed in some unproductive directions based on these assumptions. For the rest of this foreword, then, I'll describe five interrelated assumptions and how the ideas in this book justifiably run counter to them.

## **Assumption 1: Analytics, big data, and AI are wholly different phenomena.**

It is assumed by many onlookers that “traditional” analytics, big data, and AI are separate and different phenomena. *Becoming a Data Head*, however, correctly adopts the view that they are highly interrelated. All

of them involve statistical thinking. Traditional analytics approaches like regression analysis are used in all three, as are data visualization techniques. Predictive analytics is basically the same thing as supervised machine learning. And most techniques for data analysis work on any size of dataset. In short, a good Data Head can work effectively across all three, and spending a lot of time focusing on the differences among them isn't terribly productive.

**Assumption 2: Data scientists are the only people who can play in this sandbox.**

We have lionized data scientists and have often made the assumption that they are the only people who can work effectively with data and analytics. However, there is a nascent but important move toward the democratization of these ideas; increasing numbers of organizations are empowering “citizen data scientists.” Automated machine learning tools make it easier to create models that do an excellent job of predicting. There is still a need, of course, for professional data scientists to develop new algorithms and check the work of the citizens who do complex analysis. But organizations that democratize analytics and data science—putting their “amateur” Data Heads to work—can greatly increase their overall use of these important capabilities.

**Assumption 3: Data scientists are “unicorns” who have all the skills needed for these activities.**

We have assumed that data scientists—those trained in and focused upon the development and coding of models—are also able to perform all the other tasks that are required for full implementation of those models. In other words, we think they are “unicorns” who can do it all. But such unicorns don't exist at all, or exist only in small numbers. Data Heads who not only understand the rudiments of data science, but also know the business, can manage projects effectively, and are excellent at building business relationships will be extremely valuable in data science projects. They can be productive members of data science teams and increase the likelihood that data science projects will lead to business value.

**Assumption 4: You need to have a really high quantitative IQ and lots of training to succeed with data and analytics.**

A related assumption is that in order to do data science work, a person has to be very well trained in the field and that a Data Head requires a head that is very good with numbers. Both quantitative training and aptitude certainly help, but *Becoming a Data Head* argues—and I agree—that a motivated learner can master enough of data and analytics to be quite useful on data science projects. This is in part because the general principles

of statistical analysis are by no means rocket science, and also because “being useful” on data science projects doesn’t require an extremely high level of data and analytics mastery. Working with professional data scientists or automated AI programs only requires the ability and the curiosity to ask good questions, to make connections between business issues and quantitative results, and to look out for dubious assumptions.

**Assumption 5: If you didn’t study mostly quantitative fields in college or graduate school, it’s too late for you to learn what you need to work with data and analytics.**

This assumption is supported by survey data; in a 2019 survey report from Splunk of about 1300 global executives, virtually every respondent (98%) agreed that data skills are important to the jobs of tomorrow.<sup>1</sup> 81% of the executives agree that data skills are required to become a senior leader in their companies, and 85% agree that data skills will become more valuable in their firms. Nonetheless, 67% say they are not comfortable accessing or using data themselves, 73% feel that data skills are harder to learn than other business skills, and 53% believe they are too old to learn data skills. This “data defeatism” is damaging to individuals and organizations, and neither the authors of this book nor I believe it is warranted. Peruse the pages following this foreword, and you will see that no rocket science is involved!

So forget these false assumptions, and turn yourself into a Data Head. You’ll become a more valuable employee and make your organization more successful. This is the way the world is going, so it’s time to get with the program and learn more about data and analytics. I think you will find the process—and the reading of *Becoming a Data Head*—more rewarding and more pleasant than you may imagine.

Thomas H. Davenport  
Distinguished Professor, Babson College  
Visiting Professor, Oxford Saïd Business School  
Research Fellow, MIT Initiative on the Digital Economy  
Author of *Competing on Analytics*, *Big Data @ Work*,  
and *The AI Advantage*

---

<sup>1</sup>Splunk Inc., “The State of Dark Data,” 2019, [www.splunk.com/en\\_us/form/the-state-of-dark-data.html](http://www.splunk.com/en_us/form/the-state-of-dark-data.html).