

Praise for *Predictive Analytics*

“Littered with lively examples . . .”

—*The Financial Times*

“Readers will find this a mesmerizing and fascinating study. I know I did! . . . I was entranced by the book.”

—*The Seattle Post-Intelligencer*

“Siegel is a capable and passionate spokesman with a compelling vision.”

—*Analytics Magazine*

“A must-read for the normal layperson.”

—*Journal of Marketing Analytics*

“This book is an operating manual for twenty-first-century life. Drawing predictions from big data is at the heart of nearly everything, whether it’s in science, business, finance, sports, or politics. And Eric Siegel is the ideal guide.”

—**Stephen Baker, author, *The Numerati* and *Final Jeopardy: The Story of Watson, the Computer That Will Transform Our World***

“Simultaneously entertaining, informative, and nuanced. Siegel goes behind the hype and makes the science exciting.”

—**Rayid Ghani, Chief Data Scientist, Obama for America 2012 Campaign**

“The most readable (for we laymen) ‘big data’ book I’ve come across. By far. Great vignettes/stories.”

—**Tom Peters, coauthor, *In Search of Excellence***

“The future is right now—you’re living in it. Read this book to gain understanding of where we are and where we’re headed.”

—**Roger Craig, record-breaking analytical *Jeopardy!* champion; Data Scientist, Digital Reasoning**

“A clear and compelling explanation of the power of predictive analytics and how it can transform companies and even industries.”

—**Anthony Goldbloom, founder and CEO, Kaggle.com**

“The definitive book of this industry has arrived. Dr. Siegel has achieved what few have even attempted: an accessible, captivating tome on predictive analytics that is a must-read for all interested in its potential—and peril.”

—**Mark Berry, VP, People Insights, ConAgra Foods**

“I’ve always been a passionate data geek, but I never thought it might be possible to convey the excitement of data mining to a lay audience. That is what Eric Siegel does in this book. The stories range from inspiring to downright scary—read them and find out what we’ve been up to while you weren’t paying attention.”

—**Michael J. A. Berry, author of *Data Mining Techniques, Third Edition***

“Eric Siegel is the Kevin Bacon of the predictive analytics world, organizing conferences where insiders trade knowledge and share recipes. Now, he has thrown the doors open for you. Step in and explore how data scientists are rewriting the rules of business.”

—**Kaiser Fung, VP, Vimeo; author of *Numbers Rule Your World***

“Written in a lively language, full of great quotes, real-world examples, and case studies, it is a pleasure to read. The more technical audience will enjoy chapters on The Ensemble Effect and uplift modeling—both very hot trends. I highly recommend this book!”

—**Gregory Piatetsky-Shapiro, Editor, KDnuggets;
founder, KDD Conferences**

“Exciting and engaging—reads like a thriller! Predictive analytics has its roots in people’s daily activities and, if successful, affects people’s actions. By way of examples, Siegel describes both the opportunities and the threats predictive analytics brings to the real world.”

—**Marianna Dizik, Statistician, Google**

“A fascinating page-turner about the most important new form of information technology.”

—**Emiliano Pasqualetti, CEO, DomainsBot Inc.**

“Succeeds where others have failed—by demystifying big data and providing real-world examples of how organizations are leveraging the power of predictive analytics to drive measurable change.”

—**Jon Francis, Senior Data Scientist, Nike**

“In a fascinating series of examples, Siegel shows how companies have made money predicting what customers will do. Once you start reading, you will not be able to put it down.”

—**Arthur Middleton Hughes, VP, Database Marketing Institute;
author of *Strategic Database Marketing, Fourth Edition***

“Excellent. Each chapter makes the complex comprehensible, making heavy use of graphics to give depth and clarity. It gets you thinking about what else might be done with predictive analytics.”

—**Edward Nazarko, Client Technical Advisor, IBM**

“What is predictive analytics? This book gives a practical and up-to-date answer, adding new dimension to the topic and serving as an excellent reference.”

—**Ramendra K. Sahoo, Senior VP,
Risk Management and Analytics, Citibank**

“Competing on information is no longer a luxury—it’s a matter of survival. Despite its successes, predictive analytics has penetrated only so far, relative to its potential. As a result, lessons and case studies such as those provided in Siegel’s book are in great demand.”

—**Boris Evelson, VP and Principal Analyst, Forrester Research**

“Fascinating and beautifully conveyed. Siegel is a leading thought leader in the space—a must-have for your bookshelf!”

—**Sameer Chopra, Chief Analytics Officer, Orbitz Worldwide**

“A brilliant overview—strongly recommended to everyone curious about the analytics field and its impact on our modern lives.”

—**Kerem Tomak, VP of Marketing Analytics, [Macy's.com](#)**

“Eric explains the science behind predictive analytics, covering both the advantages and the limitations of prediction. A must-read for everyone!”

—**Azhar Iqbal, VP and Econometrician,
Wells Fargo Securities, LLC**

“*Predictive Analytics* delivers a ton of great examples across business sectors of how companies extract actionable, impactful insights from data. Both the novice and the expert will find interest and learn something new.”

—**Chris Pouliot, Director, Algorithms and Analytics, Netflix**

“In this new world of big data, machine learning, and data scientists, Eric Siegel brings deep understanding to deep analytics.”

—**Marc Parrish, VP, Membership, Barnes & Noble**

“A detailed outline for how we might tame the world’s unpredictability. Eric advocates quite clearly how some choices are predictably more profitable than others—and I agree!”

—**Dennis R. Mortensen, CEO of Visual Revenue,
former Director of Data Insights at Yahoo!**

“This book is an invaluable contribution to predictive analytics. Eric’s explanation of how to anticipate future events is thought provoking and a great read for everyone.”

—**Jean Paul Isson, Global VP Business Intelligence and Predictive Analytics, Monster Worldwide; coauthor, *Win with Advanced Business Analytics: Creating Business Value from Your Data***

“Predictive analytics is the key to unlocking new value at a previously unimaginable economic scale. In this book, Siegel explains how, doing an excellent job to bridge theory and practice.”

—**Sergo Grigalashvili, VP of Information Technology,
Crawford & Company**

“Predictive analytics has been steeped in fear of the unknown. Eric Siegel distinctively clarifies, removing the mystery and exposing its many benefits.”

—**Jane Kuberski, Engineering and Analytics,
Nationwide Insurance**

“As predictive analytics moves from fashionable to mainstream, Siegel removes the complexity and shows its power.”

—**Rajeeve Kaul, Senior VP, OfficeMax**

“Dr. Siegel humanizes predictive analytics. He blends analytical rigor with real-life examples with an ease that is remarkable in his field. The book is informative, fun, and easy to understand. I finished reading it in one sitting. A must-read . . . not just for data scientists!”

—**Madhu Iyer, Marketing Statistician, Intuit**

“An engaging encyclopedia filled with real-world applications that should motivate anyone still sitting on the sidelines to jump into predictive analytics with both feet.”

—**Jared Waxman, Web Marketer at LegalZoom,
previously at Adobe, Amazon, and Intuit**

“Siegel covers predictive analytics from start to finish, bringing it to life and leaving you wanting more.”

—**Brian Seeley, Manager, Risk Analytics, Paychex, Inc.**

“A wonderful look into the world of predictive analytics from the perspective of a true practitioner.”

—**Shawn Hushman, VP, Analytic Insights,
Kelley Blue Book**

“A must—*Predictive Analytics* provides an amazing view of the analytical models that predict and influence our lives on a daily basis. Siegel makes it a breeze to understand, for all readers.”

—**Zhou Yu, Online-to-Store Analyst, Google**

“As our ability to collect and analyze information improves, experts like Eric Siegel are our guides to the mysteries unlocked and the moral questions that arise.”

—Jules Polonetsky, Co-Chair and Director, Future of Privacy Forum; former Chief Privacy Officer, AOL and DoubleClick

“Highly recommended. As Siegel shows in his very readable new book, the results achieved by those adopting predictive analytics to improve decision making are game changing.”

—James Taylor, CEO, Decision Management Solutions

“An engaging, humorous introduction to the world of the data scientist. Dr. Siegel demonstrates with many real-life examples how predictive analytics makes big data valuable.”

—David McMichael, VP, Advanced Business Analytics

“An excellent exposition on the next generation of business intelligence—it’s really mankind’s latest quest for artificial intelligence.”

**—Christopher Hornick, President and CEO,
HBSC Strategic Services**

PREDICTIVE ANALYTICS

PREDICTIVE ANALYTICS



**THE POWER TO PREDICT WHO WILL
CLICK, BUY, LIE, OR DIE**

ERIC SIEGEL

WILEY

Cover image: Winona Nelson
Cover design: Wiley
Interior image design: Matt Kornhaas

Copyright © 2016 by Eric Siegel. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Published simultaneously in Canada.

Jeopardy!® is a registered trademark of Jeopardy Productions, Inc.

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 Section 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, or on the Web at www.copyright.com. 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.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor the author shall be liable for damages arising herefrom.

For general information about our other products and services, please contact our Customer Care Department within the United States at (800) 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.

Library of Congress Cataloging-in-Publication Data:

Names: Siegel, Eric, 1968-

Title: Predictive analytics : the power to predict who will click, buy, lie, or die / Eric Siegel.

Description: Revised and Updated Edition. | Hoboken : Wiley, 2016. | Revised edition of the author's Predictive analytics, 2013. | Includes index.

Identifiers: LCCN 2015031895 (print) | LCCN 2015039877 (ebook) |

ISBN 9781119145677 (paperback) | ISBN 9781119145684 (pdf) |

ISBN 9781119153658 (epub)

Subjects: LCSH: Social sciences—Forecasting. | Economic forecasting |

Prediction (Psychology) | Social prediction. | Human behavior. | BISAC:

BUSINESS & ECONOMICS / Consumer Behavior | BUSINESS & ECONOMICS / Econometrics. | BUSINESS & ECONOMICS / Marketing / General.

Classification: LCC H61.4 .S54 2016 (print) | LCC H61.4 (ebook) | DDC

303.49—dc23

LC record available at <http://lccn.loc.gov/2015031895>

Printed in the United States of America

*This book is dedicated with all my heart to my mother,
Lisa Schamberg, and my father, Andrew Siegel.*

Contents

Foreword Thomas H. Davenport	xvii
Preface to the Revised and Updated Edition	xxi
<i>What's new and who's this book for—the Predictive Analytics FAQ</i>	
Preface to the Original Edition	xxix
<i>What is the occupational hazard of predictive analytics?</i>	
Introduction	
The Prediction Effect	1
<i>How does predicting human behavior combat risk, fortify healthcare, toughen crime fighting, boost sales, and cut costs? Why must a computer learn in order to predict? How can lousy predictions be extremely valuable? What makes data exceptionally exciting? How is data science like porn? Why shouldn't computers be called computers? Why do organizations predict when you will die?</i>	
Chapter 1	
Liftoff! Prediction Takes Action (<i>deployment</i>)	23
<i>How much guts does it take to deploy a predictive model into field operation, and what do you stand to gain? What happens when a man invests his entire life savings into his own predictive stock market trading system?</i>	

Chapter 2

With Power Comes Responsibility: Hewlett-Packard, Target, the Cops, and the NSA Deduce Your Secrets (*ethics*) 47

How do we safely harness a predictive machine that can foresee job resignation, pregnancy, and crime? Are civil liberties at risk? Why does one leading health insurance company predict policyholder death?

Two extended sidebars reveal: 1) Does the government undertake fraud detection more for its citizens or for self-preservation, and 2) for what compelling purpose does the NSA need your data even if you have no connection to crime whatsoever, and can the agency use machine learning supercomputers to fight terrorism without endangering human rights?

Chapter 3

The Data Effect: A Glut at the End of the Rainbow (*data*) 103

We are up to our ears in data, but how much can this raw material really tell us? What actually makes it predictive? What are the most bizarre discoveries from data? When we find an interesting insight, why are we often better off not asking why? In what way is bigger data more dangerous? How do we avoid being fooled by random noise and ensure scientific discoveries are trustworthy?

Chapter 4

The Machine That Learns: A Look inside Chase's Prediction of Mortgage Risk (*modeling*) 147

What form of risk has the perfect disguise? How does prediction transform risk to opportunity? What should all businesses learn from insurance companies? Why does machine learning require art in addition to science? What kind of predictive model can be understood by everyone? How can we confidently trust a machine's predictions? Why couldn't prediction prevent the global financial crisis?

Chapter 5

- The Ensemble Effect: Netflix, Crowdsourcing, and Supercharging Prediction (*ensembles*)

185

To crowdsource predictive analytics—outsource it to the public at large—a company launches its strategy, data, and research discoveries into the public spotlight. How can this possibly help the company compete? What key innovation in predictive analytics has crowdsourcing helped develop? Must supercharging predictive precision involve overwhelming complexity, or is there an elegant solution? Is there wisdom in nonhuman crowds?

Chapter 6

- Watson and the *Jeopardy!* Challenge (*question answering*)

207

*How does Watson—IBM’s *Jeopardy!*-playing computer—work? Why does it need predictive modeling in order to answer questions, and what secret sauce empowers its high performance? How does the iPhone’s Siri compare? Why is human language such a challenge for computers? Is artificial intelligence possible?*

Chapter 7

- Persuasion by the Numbers: How Telenor, U.S. Bank, and the Obama Campaign Engineered Influence (*uplift*)

251

What is the scientific key to persuasion? Why does some marketing fiercely backfire? Why is human behavior the wrong thing to predict? What should all businesses learn about persuasion from presidential campaigns? What voter predictions helped Obama win in 2012 more than the detection of swing voters? How could doctors kill fewer patients inadvertently? How is a person like a quantum particle? Riddle: What often happens to you that cannot be perceived and that you can’t even be sure has happened afterward—but that can be predicted in advance?

Afterword **291***Eleven Predictions for the First Hour of 2022***Appendices**

- | | |
|--|-----|
| A. The Five Effects of Prediction | 295 |
| B. Twenty Applications of Predictive Analytics | 296 |
| C. Prediction People—Cast of “Characters” | 300 |

Hands-On Guide **303***Resources for Further Learning***Acknowledgments** **307****About the Author** **311****Index** **313**

Also see the Central Tables (color insert) for a cross-industry compendium of 182 examples of predictive analytics.

This book’s Notes—120 pages of citations and comments pertaining to the chapters above—are available online at www.PredictiveNotes.com.

Foreword

This book deals with quantitative efforts to predict human behavior. One of the earliest efforts to do that was in World War II. Norbert Wiener, the father of “cybernetics,” began trying to predict the behavior of German airplane pilots in 1940—with the goal of shooting them from the sky. His method was to take as input the trajectory of the plane from its observed motion, consider the pilot’s most likely evasive maneuvers, and predict where the plane would be in the near future so that a fired shell could hit it. Unfortunately, Wiener could predict only one second ahead of a plane’s motion, but 20 seconds of future trajectory were necessary to shoot down a plane.

In Eric Siegel’s book, however, you will learn about a large number of prediction efforts that are much more successful. Computers have gotten a lot faster since Wiener’s day, and we have a lot more data. As a result, banks, retailers, political campaigns, doctors and hospitals, and many more organizations have been quite successful of late at predicting the behavior of particular humans. Their efforts have been helpful at winning customers, elections, and battles with disease.

My view—and Siegel’s, I would guess—is that this predictive activity has generally been good for humankind. In the context of healthcare, crime, and terrorism, it can save lives. In the context of advertising, using predictions is more efficient and could conceivably save both trees (for direct mail and catalogs) and the time and attention of the recipient. In politics, it seems to reward those candidates who respect the scientific method (some might disagree, but I see that as a positive).

However, as Siegel points out—early in the book, which is admirable—these approaches can also be used in somewhat harmful ways. “With great power comes great responsibility,” he notes in quoting *Spider-Man*. The implication is that we must be careful as a society about how we use predictive models, or we may be restricted from using and benefiting from them. Like other powerful technologies or disruptive human innovations, predictive analytics is essentially amoral and can be used for good or evil. To avoid the evil applications, however, it is certainly important to understand what is possible with predictive analytics, and you will certainly learn that if you keep reading.

This book is focused on predictive analytics, which is not the only type of analytics, but the most interesting and important type. I don’t think we need more books anyway on purely descriptive analytics, which only describe the past and don’t provide any insight as to why it happened. I also often refer in my own writing to a third type of analytics—“prescriptive”—that tells its users what to do through controlled experiments or optimization. Those quantitative methods are much less popular, however, than predictive analytics.

This book and the ideas behind it are a good counterpoint to the work of Nassim Nicholas Taleb. His books, including *The Black Swan*, suggest that many efforts at prediction are doomed to fail because of randomness and the inherent unpredictability of complex events. Taleb is no doubt correct that some events are black swans that are beyond prediction, but the fact is that most human behavior is quite regular and predictable. The many examples that Siegel provides of successful prediction remind us that most swans are white.

Siegel also resists the blandishments of the “big data” movement. Certainly some of the examples he mentions fall into this category—data that is too large or unstructured to be easily managed by conventional relational databases. But the point of predictive analytics is not the relative size or unruliness of your data, but what you do with it. I have found that “big data often equals small math,” and many big data practitioners are content just to use their data to create some appealing visual analytics. That’s not nearly as valuable as creating a predictive model.

Siegel has fashioned a book that is both sophisticated and fully accessible to the non-quantitative reader. It's got great stories, great illustrations, and an entertaining tone. Such non-quants should definitely read this book, because there is little doubt that their behavior will be analyzed and predicted throughout their lives. It's also quite likely that most non-quants will increasingly have to consider, evaluate, and act on predictive models at work.

In short, we live in a predictive society. The best way to prosper in it is to understand the objectives, techniques, and limits of predictive models. And the best way to do that is simply to keep reading this book.

—**Thomas H. Davenport**

Thomas H. Davenport is the President's Distinguished Professor at Babson College, a fellow of the MIT Center for Digital Business, Senior Advisor to Deloitte Analytics, and cofounder of the International Institute for Analytics.

He is the coauthor of *Competing on Analytics*, *Big Data @ Work*, and several other books on analytics.

Preface to the Revised and Updated Edition

What's New and Who's This Book for— The Predictive Analytics FAQ

Data Scientist: The Sexiest Job of the Twenty-first Century

—Title of a *Harvard Business Review* article by Thomas Davenport and DJ Patil, who in 2015 became the first U.S. Chief Data Scientist

Prediction is booming. It reinvents industries and runs the world.

More and more, predictive analytics (PA) drives commerce, manufacturing, healthcare, government, and law enforcement. In these spheres, organizations operate more effectively by way of predicting behavior—i.e., the outcome for each individual customer, employee, patient, voter, and suspect.

Everyone's doing it. Accenture and Forrester both report that PA's adoption has more than doubled in recent years. Transparency Market Research projects the PA market will reach \$6.5 billion within a few years. A Gartner survey ranked business intelligence and analytics as the current number one investment priority of chief information officers. And in a [Salesforce.com](#) study, PA showed the highest growth rate of all sales tech trends, more than doubling its adoption in the next 18 months. High-performance sales teams are four times more likely to already be using PA than underperformers.

I am a witness to PA's expanding deployment across industries. Predictive Analytics World (PAW), the conference series I founded, has hosted over 10,000 attendees since its launch in 2009 and is expanding well beyond its original PAW Business events. With the expert assistance of industry partners, we've launched the industry-focused events PAW Government, PAW Healthcare, PAW Financial, PAW Workforce, and PAW Manufacturing, events for senior executives, and the news site *The Predictive Analytics Times*.

Since the publication of this book's first edition in 2013, I have been commissioned to deliver keynote addresses in each of these industries: marketing, market research, e-commerce, financial services, insurance, news media, healthcare, pharmaceuticals, government, human resources, travel, real estate, construction, and law, plus executive summits and university conferences.

Want a future career in futurology? The demand is blowing up. McKinsey forecasts a near-term U.S. shortage of 140,000 analytics experts and 1.5 million managers "with the skills to understand and make decisions based on analysis of big data." LinkedIn's number one "Hottest Skills That Got People Hired" is "statistical analysis and data mining."

PA is like *Moneyball* for . . . money.

FREQUENTLY ASKED QUESTIONS ABOUT *PREDICTIVE ANALYTICS*

Who is this book for?

Everyone. It's easily understood by all readers. Rather than a how-to for hands-on techies, the book serves lay readers, technology enthusiasts, executives, and analytics experts alike by covering new case studies and the latest state-of-the-art techniques.

Is the idea of predictive analytics hard to understand?

Not at all. The heady, sophisticated notion of *learning from data to predict* may sound beyond reach, but breeze through the short Introduction chapter and you'll see: The basic idea is clear, accessible, and undeniably far-reaching.

Is this book a how-to?

No, it is a conceptually complete, substantive introduction and industry overview.

Not a how-to? Then why should techies read it?

Although this mathless introduction is understandable by any reader—including those with no technical background—here's why it also affords value for would-be and established hands-on practitioners:

- **A great place to start**—provides prerequisite conceptual knowledge for those who will go on to learn the hands-on practice or will serve in an executive or management role in the deployment of PA.
- **Detailed case studies**—explores the real-world deployment of PA by Chase, IBM, HP, Netflix, the NSA, Target, U.S. Bank, and more.
- **A compendium of 182 mini-case studies**—the Central Tables, divided into nine industry groups, include examples from BBC, Citibank, ConEd, Facebook, Ford, Google, the IRS, [Match.com](#), MTV, PayPal, Pfizer, Spotify, Uber, UPS, Wikipedia, and more.
- **Advanced, cutting-edge topics**—the last three chapters introduce subfields new even to many senior experts: *Ensemble models*, *IBM Watson's question answering*, and *uplift modeling*. No matter how experienced you are, starting with a conceptually rich albeit non-technical overview may benefit you more than you'd expect—especially for *uplift modeling*. The Notes for these three chapters then provide comprehensive references to technically deep sources (available at www.PredictiveNotes.com).
- **Privacy and civil liberties**—the second chapter tackles the particular ethical concerns that arise when harnessing PA's power.
- **Holistic industry overview**—the book extends more broadly than a standard technology introduction—all of the above adds up to a survey of the field that sheds light on its societal, commercial, and ethical context.

That said, burgeoning practitioners who wish to jump directly to a more traditional, technically in-depth or hands-on treatment of this topic should

consider themselves warned: This is not the book you are seeking (but it makes a good gift; any of your relatives would be able to understand it and learn about your field of interest).

As with introductions to other fields of science and engineering, if you are pursuing a career in the field, this book will set the foundation, yet only whet your appetite for more. At the end of this book, you are guided by the Hands-On Guide on where to go next for the technical how-to and advanced underlying theory and math.

What is the purpose of this book?

I wrote this book to demonstrate why PA is intuitive, powerful, and awe-inspiring. It's a book about the most influential and valuable achievements of computerized prediction and the two things that make it possible: the people behind it and the fascinating science that powers it.

While there are a number of books that approach the how-to side of PA, this book serves a different purpose (which turned out to be a rewarding challenge for its author): sharing with a wider audience a complete picture of the field, from the way in which it empowers organizations, down to the inner workings of predictive modeling.

With its impact on the world growing so quickly, it's high time the predictive power of data—and how to scientifically tap it—be demystified. Learning from data to predict human behavior is no longer arcane.

How technical does this book get?

While accessible and friendly to newcomers of any background, this book explores “under the hood” far enough to reveal the inner workings of *decision trees* (Chapter 4), an exemplary form of predictive model that serves well as a place to start learning about PA, and often as a strong first option when executing a PA project.

I strove to go as deep as possible—substantive across the gamut of fascinating topics related to PA—while still sustaining interest and accessibility not only for neophyte users, but even for those interested in the field avocationally, curious about science and how it is changing the world.

Is this a university textbook?

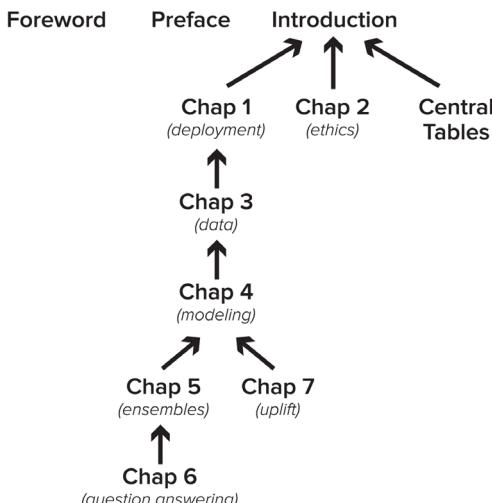
This book has served as a textbook at more than 30 colleges and universities. A former computer science professor, I wrote this introduction to be conceptually complete. In the table of contents, the words in parentheses beside each chapter’s “catchy” title reveal an outline that covers the fundamentals: (1) *model deployment*, (2) *ethics*, (3) *data*, (4) *predictive modeling*, (5) *ensemble models*, (6) *question answering*, and (7) *uplift modeling*. To guide reading assignments, see the diagram under the next question below.

However, this is not written in the formal style of a textbook; rather, I sought to deliver an entertaining, engaging, relevant work that illustrates the concepts largely via anecdotes.

For instructors considering this book for course material, additional resources and information may be found at www.teachPA.com.

How should I read this book?

The chapters of this book build upon one another. Some depend only on first reading the Introduction, but others build cumulatively. The figure below depicts these dependencies—read a chapter only after first reading the one it points up to. For example, Chapter 3 assumes you’ve already read Chapter 1, which assumes you’ve read the Introduction.



Dependencies between chapters. An arrow pointing up means, “Read the chapter above first.”

Note: If you are reading the e-book version, be sure not to miss the Central Tables (a compendium of 182 mini-case studies), the link for which may be less visibly located toward the end of the table of contents.

What's new in the “Revised and Updated” edition of *Predictive Analytics*?

- **The Real Reason the NSA Wants Your Data: Automatic Suspect Discovery.** A special sidebar in Chapter 2 (on ethics in PA) presumes—with much evidence—that the National Security Agency considers PA a strategic priority. Can the organization use PA without endangering civil liberties?
- **Dozens of new examples from Facebook, Hopper, Shell, Uber, UPS, the U.S. government, and more.** The Central Tables’ compendium of mini-case studies has grown to 182 entries, including breaking examples.
- **A much-needed warning regarding bad science.** Chapter 3, “The Data Effect,” includes an in-depth section about an all-too-common pitfall and how we avoid it, i.e., how to successfully tap data’s potential without being fooled by random noise, ensuring sound discoveries are made.
- **Even more extensive Notes, updated and expanded to 120 pages, now moved online.** Now located at www.PredictiveNotes.com, the Notes include citations and comments that pertain to the above new content, as well as updated citations throughout chapters.

Where can I learn more after this book, such as a how-to for hands-on practice?

- **The Hands-On Guide at the end of this book**—reading and training options that guide getting started
- **This book’s website**—videos, articles, and more resources: www.thepredictionbook.com

- **Predictive Analytics World**—the leading cross-vendor conference series in North America and Europe, which includes advanced training workshop days and the industry-specific events PAW Business, PAW Government, PAW Healthcare, PAW Financial, PAW Workforce, and PAW Manufacturing: www.pawcon.com
- **The Predictive Analytics Guide**—articles, industry portals, and other resources: www.pawcon.com/guide
- **Predictive Analytics Applied**—the author’s online training workshop, which, unlike this book, *is* a how-to. Access immediately, on-demand at any time: www.businessprediction.com
- ***The Predictive Analytics Times***—the premier resource: industry news, technical articles, videos, events, and community: www.predictiveanalyticstimes.com

Preface to the Original Edition

Yesterday is history, tomorrow is a mystery, but today is a gift. That's why we call it the present.

—Attributed to A. A. Milne, Bil Keane, and Oogway,
the wise turtle in *Kung Fu Panda*

People look at me funny when I tell them what I do. It's an occupational hazard.

The Information Age suffers from a glaring omission. This claim may surprise many, considering we are actively recording Everything That Happens in the World. Moving beyond history books that document important events, we've progressed to systems that log every click, payment, call, crash, crime, and illness. With this in place, you would expect lovers of data to be satisfied, if not spoiled rotten.

But this apparent infinity of information excludes the very events that would be most valuable to know of: *things that haven't happened yet*.

Everyone craves the power to see the future; we are collectively obsessed with prediction. We bow to prognostic deities. We empty our pockets for palm readers. We hearken to horoscopes, adore astrology, and feast upon fortune cookies.

But many people who salivate for psychics also spurn science. Their innate response says "yuck"—it's either too hard to understand or too boring. Or perhaps many believe prediction by its nature is just impossible without supernatural support.

There's a lighthearted TV show I like premised on this very theme, *Psych*, in which a sharp-eyed detective—a modern-day, data-driven Sherlock Holmesian hipster—has perfected the art of observation so masterfully, the cops believe his spot-on deductions must be an admission of guilt. The hero gets out of this pickle by conforming to the norm: He simply informs the police he is psychic, thereby managing to stay out of prison and continuing to fight crime. Comedy ensues.

I've experienced the same impulse, for example, when receiving the occasional friendly inquiry as to my astrological sign. But, instead of posing as a believer, I turn to humor: "I'm a Scorpio, and Scorpions don't believe in astrology."

The more common cocktail party interview asks what I do for a living. I brace myself for eyes glazing over as I carefully enunciate: *predictive analytics*. Most people have the luxury of describing their job in a single word: doctor, lawyer, waiter, accountant, or actor. But, for me, describing this largely unknown field hijacks the conversation every time. Any attempt to be succinct falls flat:

I'm a business consultant in technology. They aren't satisfied and ask, "What kind of technology?"

I make computers predict what people will do. Bewilderment results, accompanied by complete disbelief and a little fear.

I make computers learn from data to predict individual human behavior. Bewilderment, plus nobody wants to talk about data at a party.

I analyze data to find patterns. Eyes glaze over even more; awkward pauses sink amid a sea of abstraction.

I help marketers target which customers will buy or cancel. They sort of get it, but this wildly undersells and pigeonholes the field.

I predict customer behavior, like when Target famously predicted whether you are pregnant. Moonwalking ensues.

So I wrote this book to demonstrate for you why predictive analytics is intuitive, powerful, and awe-inspiring.

I have good news: *A little prediction goes a long way.* I call this The Prediction Effect, a theme that runs throughout the book. The potency of prediction is

pronounced—as long as the predictions are better than guessing. This effect renders predictive analytics believable. We don’t have to do the impossible and attain true clairvoyance. The story is exciting yet credible: Putting odds on the future to lift the fog just a bit off our hazy view of tomorrow means pay dirt. In this way, predictive analytics combats risk, boosts sales, cuts costs, fortifies healthcare, streamlines manufacturing, conquers spam, toughens crime fighting, optimizes social networks, and wins elections.

Do you have the heart of a scientist or a businessperson? Do you feel more excited by the very idea of prediction, or by the value it holds for the world?

I was struck by the notion of *knowing the unknowable*. Prediction seems to defy a law of nature: You cannot see the future because it isn’t here yet. We find a workaround by building machines that learn from experience. It’s the regimented discipline of using what we *do* know—in the form of data—to place increasingly accurate odds on what’s coming next. We blend the best of math and technology, systematically tweaking until our scientific hearts are content to derive a system that peers right through the previously impenetrable barrier between today and tomorrow.

Talk about boldly going where no one has gone before!

Some people are in sales; others are in politics. I’m in prediction, and it’s awesome.

predictive analytics

never Researchers public book computers take things decisions call years see examples common predictive common people + just technology order company best predictive well credit human annual Data case growing estimated person healthcare name used sham benefit know almost across customers detect email day human annual benefit yet target hospital times individual bad every learn new

prediction

never Researchers public book computers take things decisions call years see examples common predictive common people + just technology order company best predictive well credit human annual Data case growing estimated person healthcare name used sham benefit know almost across customers detect email day human annual benefit yet target hospital times individual bad every learn new

Introduction

The Prediction Effect

I'm just like you. I succeed at times, and at others I fail. Some days good things happen to me, some days bad. We always wonder how things could have gone differently. I begin with seven brief tales of woe:

1. In 2009 I just about destroyed my right knee downhill skiing in Utah. The jump was no problem; it was landing that presented an issue. For knee surgery, I had to pick a graft source from which to reconstruct my busted ACL (the knee's central ligament). The choice is a tough one and can make the difference between living with a good knee or a bad knee. I went with my hamstring. *Could the hospital have selected a medically better option for my case?*
2. Despite all my suffering, it was really my health insurance company that paid dearly—knee surgery is expensive. *Could the company have better anticipated the risk of accepting a ski jumping fool as a customer and priced my insurance premium accordingly?*
3. Back in 1995 another incident caused me suffering, although it hurt less. I fell victim to identity theft, costing me dozens of hours of bureaucratic baloney and tedious paperwork to clear up my damaged credit rating. *Could the creditors have prevented the fiasco by detecting*

that the accounts were bogus when they were filed under my name in the first place?

4. With my name cleared, I recently took out a mortgage to buy an apartment. Was it a good move, or *should my financial adviser have warned me the property could soon be outvalued by my mortgage?*
5. While embarking on vacation, I asked the neighboring airplane passenger what price she'd paid for her ticket, and it was much less than I'd paid. *Before I booked the flight, could I have determined the airfare was going to drop?*
6. My professional life is susceptible, too. My business is faring well, but a company always faces the risk of changing economic conditions and growing competition. *Could we protect the bottom line by foreseeing which marketing activities and other investments will pay off, and which will amount to burnt capital?*
7. Small ups and downs determine your fate and mine, every day. A precise spam filter has a meaningful impact on almost every working hour. We depend heavily on effective Internet search for work, health (e.g., exploring knee surgery options), home improvement, and most everything else. We put our faith in personalized music and movie recommendations from Spotify and Netflix. After all these years, my mailbox wonders why companies don't know me well enough to send less junk mail (and sacrifice fewer trees needlessly).

These predicaments matter. They can make or break your day, year, or life. But what do they all have in common?

These challenges—and many others like them—are best addressed with *prediction*. Will the patient's outcome from surgery be positive? Will the credit applicant turn out to be a fraudster? Will the homeowner face a bad mortgage? Will the airfare go down? Will the customer respond if mailed a brochure? By predicting these things, it is possible to fortify healthcare, combat risk, conquer spam, toughen crime fighting, boost sales, and cut costs.

PREDICTION IN BIG BUSINESS—THE DESTINY OF ASSETS

There's another angle. Beyond benefiting you and me as consumers, prediction serves the organization, empowering it with an entirely new form of competitive armament. Corporations positively pounce on prediction.

In the mid-1990s, an entrepreneurial scientist named Dan Steinberg delivered predictive capabilities unto the nation's largest bank, Chase, to assist with their management of millions of mortgages. This mammoth enterprise put its faith in Dan's predictive technology, deploying it to drive transactional decisions across a tremendous mortgage portfolio. What did this guy have on his résumé?

Prediction is power. Big business secures a killer competitive stronghold by predicting the future destiny and value of individual assets. In this case, by driving mortgage decisions with predictions about the future payment behavior of homeowners, Chase curtailed risk, boosted profit, and witnessed a windfall.

INTRODUCING . . . THE CLAIRVOYANT COMPUTER

Compelled to grow and propelled to the mainstream, predictive technology is commonplace and affects everyone, every day. It impacts your experiences in undetectable ways as you drive, shop, study, vote, see the doctor, communicate, watch TV, earn, borrow, or even steal.

This book is about the most influential and valuable achievements of computerized prediction, and the two things that make it possible: the people behind it, and the fascinating science that powers it.

Making such predictions poses a tough challenge. Each prediction depends on multiple factors: The various characteristics known about each patient, each homeowner, each consumer, and each e-mail that may be spam. How shall we attack the intricate problem of putting all these pieces together for each prediction?

The idea is simple, although that doesn't make it easy. The challenge is tackled by a systematic, scientific means to develop and continually improve prediction—to literally *learn* to predict.

The solution is *machine learning*—computers automatically developing new knowledge and capabilities by furiously feeding on modern society's greatest and most potent *unnatural* resource: data.

“FEED ME!”—FOOD FOR THOUGHT FOR THE MACHINE

Data is the new oil.

—European Consumer Commissioner Meglena Kuneva

The only source of knowledge is experience.

—Albert Einstein

In God we trust. All others must bring data.

—William Edwards Deming (a business professor famous for work in manufacturing)

Most people couldn't be less interested in data. It can seem like such dry, boring stuff. It's a vast, endless regimen of recorded facts and figures, each alone as mundane as the most banal tweet, “I just bought some new sneakers!” It's the unsalted, flavorless residue deposited en masse as businesses churn away.

Don't be fooled! The truth is that data embodies a priceless collection of experience from which to learn. Every medical procedure, credit application, Facebook post, movie recommendation, fraudulent act, spammy e-mail, and purchase of any kind—each positive or negative outcome, each successful or failed sales call, each incident, event, and transaction—is encoded as data and warehoused. This glut grows by an estimated 2.5 quintillion bytes per day (that's a 1 with 18 zeros after it). And so a veritable Big Bang has set off, delivering an epic sea of raw materials, a plethora of examples so great in number, only a computer could manage to learn from them. Used correctly, computers avidly soak up this ocean like a sponge.

As data piles up, we have ourselves a genuine gold rush. But data isn't the gold. I repeat, data in its raw form is boring crud. The gold is what's discovered therein.

The process of machines learning from data unleashes the power of this exploding resource. It uncovers what drives people and the actions they take—what makes us tick and how the world works. With the new knowledge gained, prediction is possible.



This learning process discovers insightful gems such as:¹

- Early retirement decreases your life expectancy.
- Online daters more consistently rated as attractive receive *less* interest.
- Rihanna fans are mostly political Democrats.
- Vegetarians miss fewer flights.
- Local crime increases after public sporting events.

Machine learning builds upon insights such as these in order to develop predictive capabilities, following a number-crunching, trial-and-error process that has its roots in statistics and computer science.

¹ See Chapter 3 for more details on these examples.

I KNEW YOU WERE GOING TO DO THAT

With this power at hand, what do we want to predict? Every important thing a person does is valuable to predict, namely: *consume, think, work, quit, vote, love, procreate, divorce, mess up, lie, cheat, steal, kill, and die*. Let's explore some examples.²

PEOPLE CONSUME

- Hollywood studios predict the success of a screenplay if produced.
- Netflix awarded \$1 million to a team of scientists who best improved their recommendation system's ability to predict which movies you will like.
- The Hopper app helps you get the best deal on a flight by recommending whether you should buy or wait, based on its prediction as to whether the airfare will change.
- Australian energy company Energex predicts electricity demand in order to decide where to build out its power grid, and Con Edison predicts system failure in the face of high levels of consumption.
- Wall Street firms trade algorithmically, buying and selling based on the prediction of stock prices.
- Companies predict which customer will buy their products in order to target their marketing, from U.S. Bank down to small companies like Harbor Sweets (candy) and Vermont Country Store (“top quality and hard-to-find classic products”). These predictions dictate the allocations of precious marketing budgets. Some companies literally predict how to best influence you to buy more (the topic of Chapter 7).
- Prediction drives the coupons you get at the grocery cash register. U.K. grocery giant Tesco, the world’s third-largest retailer, predicts which discounts will be redeemed in order to target more than

² For more examples and further detail, see this book’s Central Tables.

100 million personalized coupons annually at cash registers across 13 countries. Similarly, Kmart, Kroger, Ralph's, Safeway, Stop & Shop, Target, and Winn-Dixie follow in kind.

- Predicting mouse clicks pays off massively. Since websites are often paid per click for the advertisements they display, they predict which ad you're mostly likely to click in order to instantly choose which one to show you. This, in effect, selects more relevant ads and drives millions in newly found revenue.
- Facebook predicts which of the thousands of posts by your friends will interest you most every time you view the news feed (unless you change the default setting). The social network also predicts the suggested “people you may know,” not to mention which ads you’re likely to click.

PEOPLE LOVE, WORK, PROCREATE, AND DIVORCE

- The leading career-focused social network, LinkedIn, predicts your job skills.
- Online dating leaders [Match.com](#), OkCupid, and eHarmony predict which hottie on your screen would be the best bet at your side.
- Target predicts customer pregnancy in order to market relevant products accordingly. Nothing foretells consumer need like predicting the birth of a new consumer.
- Clinical researchers predict infidelity and divorce. There’s even a self-help website tool to put odds on your marriage’s long-term success ([www.divorceprobability.com](#)).

PEOPLE THINK AND DECIDE

- Obama was reelected in 2012 with the help of voter prediction. The Obama for America campaign predicted which voters would be positively persuaded by campaign contact (a call, door knock, flier, or TV ad), and which would actually be inadvertently influenced to

(continued)

(continued)

vote adversely by contact. Employed to drive campaign decisions for millions of swing state voters, this method was shown to successfully convince more voters to choose Obama than traditional campaign targeting. Hillary for America 2016 is positioning to apply the same technique.

- “What did you mean by that?” Systems have learned to ascertain the intent behind the written word. Citibank and PayPal detect the customer sentiment about their products, and one researcher’s machine can tell which [Amazon.com](#) book reviews are sarcastic.
- Student essay grade prediction has been developed for possible use to automatically grade. The system grades as accurately as human graders.
- There’s a machine that can participate in the same capacity as humans in the United States’ most popular broadcast celebration of human knowledge and cultural literacy. On the TV quiz show *Jeopardy!*, IBM’s Watson computer triumphed. This machine learned to work proficiently enough with English to predict the answers to free-form inquiries across an open range of topics and defeat the two all-time human champs.
- Computers can literally read your mind. Researchers trained systems to decode a scan of your brain and determine which type of object you’re thinking about—such as certain tools, buildings, and food—with over 80 percent accuracy for some human subjects.

PEOPLE QUIT

- Hewlett-Packard (HP) earmarks each and every one of its more than 300,000 worldwide employees according to “Flight Risk,” the expected chance he or she will quit their job, so that managers may intervene in advance where possible and plan accordingly otherwise.
- Ever experience frustration with your cell phone service? Your service provider endeavors to know. All major wireless carriers

predict how likely it is you will cancel and switch to a competitor—possibly before you have even conceived a plan to do so—based on factors such as dropped calls, your phone usage, billing information, and whether your contacts have already defected.

- FedEx stays ahead of the game by predicting—with 65 to 90 percent accuracy—which customers are at risk of defecting to a competitor.
- The American Public University System predicted student dropouts and used these predictions to intervene successfully; the University of Alabama, Arizona State University, Iowa State University, Oklahoma State University, and the Netherlands' Eindhoven University of Technology predict dropouts as well.
- Wikipedia predicts which of its editors, who work for free as a labor of love to keep this priceless online asset alive, are going to discontinue their valuable service.
- Researchers at Harvard Medical School predict that if your friends stop smoking, you're more likely to do so yourself as well. Quitting smoking is contagious.

PEOPLE MESS UP

- Insurance companies predict who is going to crash a car or hurt themselves another way (such as a ski accident). Allstate predicts bodily injury liability from car crashes based on the characteristics of the insured vehicle, demonstrating improvements to prediction that could be worth an estimated \$40 million annually. Another top insurance provider reported savings of almost \$50 million per year by expanding its actuarial practices with advanced predictive techniques.
- Ford is learning from data so its cars can detect when the driver is not alert due to distraction, fatigue, or intoxication and take action such as sounding an alarm.
- Researchers have identified aviation incidents that are five times more likely than average to be fatal, using data from the National Transportation Safety Board.

(continued)

(continued)

- All large banks and credit card companies predict which debtors are most likely to turn delinquent, failing to pay back their loans or credit card balances. Collection agencies prioritize their efforts with predictions of which tactic has the best chance to recoup the most from each defaulting debtor.

PEOPLE GET SICK AND DIE

I'm not afraid of death; I just don't want to be there when it happens.

—Woody Allen

- In 2013, the Heritage Provider Network handed over \$500,000 to a team of scientists who won an analytics competition to best predict individual hospital admissions. By following these predictions, proactive preventive measures can take a healthier bite out of the tens of billions of dollars spent annually on unnecessary hospitalizations. Similarly, the University of Pittsburgh Medical Center predicts short-term hospital readmissions, so doctors can be prompted to think twice before a hasty discharge.
- At Stanford University, a machine learned to diagnose breast cancer better than human doctors by discovering an innovative method that considers a greater number of factors in a tissue sample.
- Researchers at Brigham Young University and the University of Utah correctly predict about 80 percent of premature births (and about 80 percent of full-term births), based on peptide biomarkers, as found in a blood exam as early as week 24 of pregnancy.
- University researchers derived a method to detect patient schizophrenia from transcripts of their spoken words alone.
- A growing number of life insurance companies go beyond conventional actuarial tables and employ predictive technology to establish mortality risk. It's not called *death insurance*, but they calculate when you are going to die.

- Beyond life insurance, one top-five *health* insurance company predicts the probability that elderly insurance policyholders will pass away within 18 months, based on clinical markers in the insured's recent medical claims. Fear not—it's actually done for benevolent purposes.
- Researchers predict your risk of death in surgery based on aspects of you and your condition to help inform medical decisions.
- By following one common practice, doctors regularly—yet unintentionally—sacrifice some patients in order to save others, and this is done completely without controversy. But this would be lessened by predicting something besides diagnosis or outcome: healthcare *impact* (impact prediction is the topic of Chapter 7).

PEOPLE LIE, CHEAT, STEAL, AND KILL

- Most medium-size and large banks employ predictive technology to counter the ever-blooming assault of fraudulent checks, credit card charges, and other transactions. Citizens Bank developed the capacity to decrease losses resulting from check fraud by 20 percent. Hewlett-Packard saved \$66 million by detecting fraudulent warranty claims.
- Predictive computers help decide who belongs in prison. To assist with parole and sentencing decisions, officials in states such as Oregon and Pennsylvania consult prognostic machines that assess the risk a convict will offend again.
- Murder is widely considered impossible to predict with meaningful accuracy in general, but within at-risk populations predictive methods can be effective. Maryland analytically generates predictions as to which inmates will kill or be killed. University and law enforcement researchers have developed predictive systems that foretell murder among those previously convicted for homicide.
- One fraud expert at a large bank in the United Kingdom extended his work to discover a small pool of terror suspects based on their

(continued)

(continued)

banking activities. While few details have been disclosed publicly, it's clear that the National Security Agency also considers this type of analysis a strategic priority in order to automatically discover previously unknown potential suspects.

- Police patrol the areas predicted to spring up as crime hot spots in cities such as Chicago, Memphis, and Richmond, Va.
- Inspired by the TV crime drama *Lie to Me* about a microexpression reader, researchers at the University at Buffalo trained a system to detect lies with 82 percent accuracy by observing eye movements alone.
- As a professor at Columbia University in the late 1990s, I had a team of teaching assistants who employed cheating-detection software to patrol hundreds of computer programming homework submissions for plagiarism.
- The IRS predicts if you are cheating on your taxes.

THE LIMITS AND POTENTIAL OF PREDICTION

An economist is an expert who will know tomorrow why the things he predicted yesterday didn't happen.

—Earl Wilson

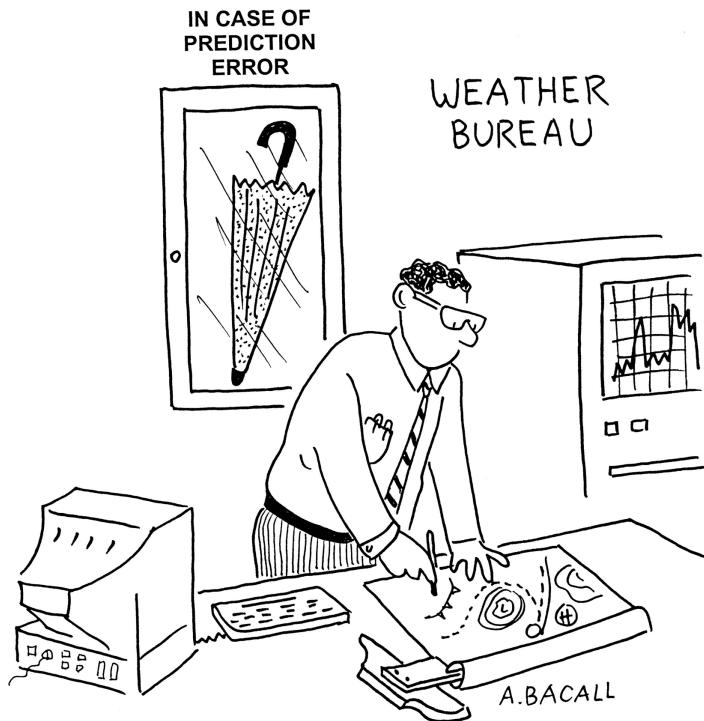
How come you never see a headline like “Psychic Wins Lottery”?

—Jay Leno

Each of the preceding accomplishments is powered by prediction, which is in turn a product of machine learning. A striking difference exists between these varied capabilities and science fiction: They aren't fiction. At this point, I predict that you won't be surprised to hear that those examples represent

only a small sample. You can safely predict that the power of prediction is here to stay.

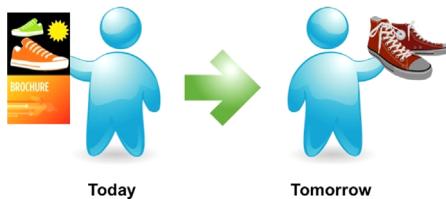
But are these claims too bold? As the Danish physicist Niels Bohr put it, “Prediction is very difficult, especially if it’s about the future.” After all, isn’t prediction basically impossible? The future is unknown, and uncertainty is the only thing about which we’re certain.



Let me be perfectly clear. It’s fuzzy. Accurate prediction is generally not possible. The weather is predicted with only about 50 percent accuracy, and it doesn’t get easier predicting the behavior of humans, be they patients, customers, or criminals.

Good news! Predictions need not be accurate to score big value. For instance, one of the most straightforward commercial applications of

predictive technology is deciding whom to target when a company sends direct mail. If the learning process identifies a carefully defined group of customers who are predicted to be, say, three times more likely than average to respond positively to the mail, the company profits big-time by preemptively removing likely *nonresponders* from the mailing list. And those non-responders in turn benefit, contending with less junk mail.



Prediction—A person who sees a sales brochure today buys a product tomorrow.

In this way the business, already playing a sort of numbers game by conducting mass marketing in the first place, tips the balance delicately yet significantly in its favor—and does so without highly accurate predictions. In fact, its utility withstands quite poor accuracy. If the overall marketing response is at 1 percent, the so-called hot pocket with three times as many would-be responders is at 3 percent. So, in this case, we can't confidently predict the response of any one particular customer. Rather, the value is derived from identifying a group of people who—in aggregate—will tend to behave in a certain way.

This demonstrates in a nutshell what I call *The Prediction Effect*. Predicting better than pure guesswork, even if not accurately, delivers real value. A hazy view of what's to come outperforms complete darkness by a landslide.

The Prediction Effect: *A little prediction goes a long way.*

This is the first of five Effects introduced in this book. You may have heard of the butterfly, Doppler, and placebo effects. Stay tuned here for the *Data*, *Induction*, *Ensemble*, and *Persuasion Effects*. Each of these Effects encompasses the fun part of science and technology: an intuitive hook that reveals how it works and why it succeeds.

THE FIELD OF DREAMS

People . . . operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage.

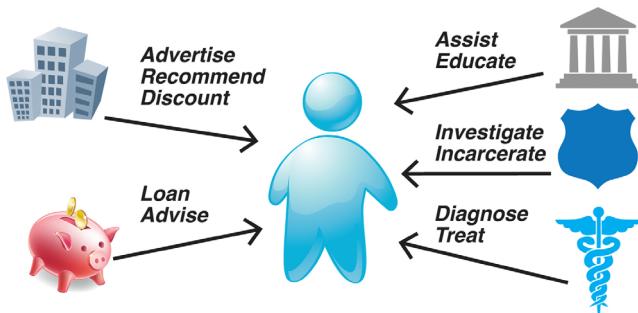
—Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*

What field of study or branch of science are we talking about here? Learning how to predict from data is sometimes called *machine learning*—but it turns out this is mostly an academic term you find used within research labs, conference papers, and university courses (full disclosure: I taught the Machine Learning graduate course at Columbia University a couple of times in the late 1990s). These arenas are a priceless wellspring, but they aren’t where the rubber hits the road. In commercial, industrial, and government applications—in the real-world usage of machine learning to predict—it’s called something else, something that in fact is the very topic of this book:

Predictive analytics (PA)—*Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.*³

³ In this definition, *individuals* is a broad term that can refer to people as well as other organizational elements. Most examples in this book involve predicting people, such as customers, debtors, applicants, employees, students, patients, donors, voters, taxpayers, potential suspects, and convicts. However, PA also applies to individual companies (e.g., for business-to-business), products, locations, restaurants, vehicles, ships, flights, deliveries, buildings, manholes, transactions, Facebook posts, movies, satellites, stocks, *Jeopardy!* questions, and much more. Whatever the domain, PA renders predictions over scalable numbers of individuals.

Built upon computer science and statistics and bolstered by devoted conferences and university degree programs, PA has emerged as its own discipline. But beyond a field of science, PA is a movement that exerts a forceful impact. Millions of decisions a day determine whom to call, mail, approve, test, diagnose, warn, investigate, incarcerate, set up on a date, and medicate. PA is the means to drive *per-person* decisions empirically, as guided by data. By answering this mountain of smaller questions, PA may in fact answer the biggest question of all: *How can we improve the effectiveness of all these massive functions across government, healthcare, business, nonprofit, and law enforcement work?*



Predictions drive how organizations treat and serve an individual, across the frontline operations that define a functional society.

In this way, PA is a completely different animal from *forecasting*. Forecasting makes aggregate predictions on a macroscopic level. How will the economy fare? Which presidential candidate will win more votes in Ohio? Whereas forecasting estimates the total number of ice cream cones to be purchased next month in Nebraska, PA tells you which *individual* Nebraskans are most likely to be seen with cone in hand.

PA leads within the growing trend to make decisions more “data driven,” relying less on one’s “gut” and more on hard, empirical evidence. Enter this fact-based domain and you’ll be attacked by buzzwords, including *analytics*, *big data*, *data science*, and *business intelligence*. While PA fits

underneath each of these umbrellas, these evocative terms refer more to the culture and general skill sets of technologists who do an assortment of creative, innovative things with data, rather than alluding to any specific technology or method. These areas are broad; in some cases, they refer simply to standard Excel reports—that is, to things that are important and require a great deal of craft, but may not rely on science or sophisticated math. And so they are more subjectively defined. As Mike Loukides, a vice president at the innovation publisher O'Reilly, once put it, "Data science is like porn—you know it when you see it." Another term, *data mining*, is often used as a synonym for PA, but as an evocative metaphor depicting "digging around" through data in one fashion or another, it is often used more broadly as well.

ORGANIZATIONAL LEARNING

The powerhouse organizations of the Internet era, which include Google and Amazon . . . have business models that hinge on predictive models based on machine learning.

—Professor Vasant Dhar, Stern School of Business,
New York University

A breakthrough in machine learning would be worth 10 Microsofts.

—Bill Gates

An organization is sort of a "megaperson," so shouldn't it "megalearn"? A group comes together for the collective benefit of its members and those it serves, be it a company, government, hospital, university, or charity. Once formed, it gains from division of labor, mutually complementary skills, and the efficiency of mass production. The result is more powerful than the sum of its parts. Collective learning is the organization's next logical step to further leverage this power. Just as a salesperson learns over time from her positive and negative interactions with sales leads, her successes, and failures, PA is the process by which an organization learns from the experience it has

collectively gained across its team members and computer systems. In fact, an organization that doesn't leverage its data in this way is like a person with a photographic memory who never bothers to think.

With only a few striking exceptions, we find that organizations, rather than individuals, benefit by employing PA. Organizations make the many, many operational decisions for which there's ample room for improvement; organizations are intrinsically inefficient and wasteful on a grand scale. Marketing casts a wide net—junk mail is marketing money wasted and trees felled to print unread brochures. An estimated 80 percent of all e-mail is spam. Risky debtors are given too much credit. Applications for government benefits are backlogged and delayed. And it's organizations that have the data to power the predictions that drive improvements in these operations.

In the commercial sector, profit is a driving force. You can well imagine the booming incentives intrinsic to rendering everyday routines more efficient, marketing more precisely, catching more fraud, avoiding bad debtors, and luring more online customers. Upgrading how business is done, PA rocks the enterprise's economies of scale, optimizing operations right where it makes the biggest difference.

THE NEW SUPER GEEK: DATA SCIENTISTS

The alternative [to thinking ahead] would be to think backwards . . . and that's just remembering.

—Sheldon, the theoretical physicist on *The Big Bang Theory*

Opportunities abound, but the profit incentive is not the only driving force. The source, the energy that makes it work, is Geek Power! I speak of the enthusiasm of technical practitioners. Truth be told, my passion for PA didn't originate from its value to organizations. I am in it for the fun. The idea of a machine that can actually learn seems so cool to me that I care more about what happens inside the magic box than its outer usefulness.

Indeed, perhaps that's the defining motivator that qualifies one as a geek. We love the technology; we're in awe of it. Case in point: The leading free, open-source software tool for PA, called R (a one-letter, geeky name), has a rapidly expanding base of users as well as enthusiastic volunteer developers who add to and support its functionalities. Great numbers of professionals and amateurs alike flock to public PA competitions with a tremendous spirit of "coopetition." We operate within organizations, or consult across them. We're in demand, so we fly a lot. But we fly coach, at best Economy Plus.

THE ART OF LEARNING

Whatcha gonna do with your CPU to reach its potentiality?

Use your noggin when you log in to crank it exponentially.

The endeavor that will render my obtuse computer clever:

Self-improve impeccably by way of trial and error.

Once upon a time, humanity created The Ultimate General Purpose Machine and, in an inexplicable fit of understatement, decided to call it "a computer" (a word that until this time had simply meant a person who did computations by hand). This automaton could crank through any demanding, detailed set of endless instructions without fail or error and with nary a complaint; within just a few decades, its speed became so blazingly brisk that humanity could only exclaim, "Gosh, we really cranked that!" An obviously much better name for this device would have been the appropriately grand *La Machine*, but a few decades later this name was hyperbolically bestowed upon a food processor (I am not joking). *Quel dommage.* "What should we do with the computer? What's its true potential, and how do we achieve it?" humanity asked of itself in wonderment.

A computer and your brain have something in common that renders them both mysterious, yet at the same time easy to take for granted. If while

pondering what this might be you heard a pin drop, you have your answer. They are both silent. Their mechanics make no sound. Sure, a computer may have a disk drive or cooling fan that stirs—just as one’s noggin may emit wheezes, sneezes, and snores—but the mammoth grunt work that takes place therein involves no “moving parts,” so these noiseless efforts go along completely unwitnessed. The smooth delivery of content on your screen—and ideas in your mind—can seem miraculous.⁴

They’re both powerful as heck, your brain and your computer. So could computers be successfully programmed to think, feel, or become truly intelligent? Who knows? At best these are stimulating philosophical questions that are difficult to answer, and at worst they are subjective benchmarks for which success could never be conclusively established. But thankfully we do have some clarity: There is one truly impressive, profound human endeavor computers *can* undertake. They can learn.

But how? It turns out that learning—generalizing from a list of examples, be it a long list or a short one—is more than just challenging. It’s a philosophically deep dilemma. Machine learning’s task is to find patterns that appear not only in the data at hand, but in general, so that what is learned will hold true in new situations never yet encountered. At the core, this ability to generalize is the magic bullet of PA. There is a true art in the design of these computer methods. We’ll explore more later, but for now I’ll give you a hint. The machine actually learns more about your next likely action by studying *others* than by studying *you*.

While I’m dispensing teasers that leave you hanging, here’s one more. This book’s final chapter answers the riddle: *What often happens to you that*

⁴ Silence is characteristic to solid state electronics, but computers didn’t have to be built that way. The idea of a general-purpose, instruction-following machine is abstract, not affixed to the notion of electricity. You could construct a computer of cogs and wheels and levers, powered by steam or gasoline. I mean, I wouldn’t recommend it, but you could. It would be slow, big, and loud, and nobody would buy it.

cannot be witnessed, and that you can't even be sure has happened afterward—but that can be predicted in advance?

Learning from data to predict is only the first step. To take the next step and *act on predictions* is to fearlessly gamble. Let's kick off Chapter 1 with a suspenseful story that shows why launching PA feels like blasting off in a rocket.

Acknowledgments

I could not have written this book without the love, guidance, and encouragement of my family: Lisa Schamberg (mother), Andrew Siegel (father), Maria de Fatima Callou (wife), Rachel Siegel (sister), Ene Piirak (stepmother), Patrick Robins (stepfather), and Anita King (grandmother).

A few cherished advisers provided extensive guidance and feedback on my writing. More than a mother to me—if such a thing is possible—Lisa Schamberg is also a natural scholar and brilliant former high school teacher who took the time to provide insightful input for every single page. John Elder, a top industry leader (and the subject of Chapter 1), has patiently enhanced my knowledge and writing in astounding ways, and placed his company’s resources at my disposal to support the development of this book. Jim Sterne’s patience and dedication assisting with my writing are matched only by his profound talent; he is also the godfather of Predictive Analytics World, which would not exist without him.

Shannon Vargo, my editor at John Wiley & Sons, Inc., provided resources, flexibility, and encouragement to nurture this project to life. The dynamic publishing duo Lee Thompson (throughout) and Myles Thompson (early in the process) gave me the extensive feedback and worldly context needed to bring this topic into the realm of the relevant.

Several people went above and beyond with extensive feedback and guidance: my father, Andrew Siegel; David Waltzer; Adam Cohen; Gary Miner; Dean Abbott; Paul Hofmann; and David Dimas.

I would like to thank the subjects of my writing—all pioneers of prediction—most of whom endured extensive interviews filled with

many probing questions: John Elder, Rayid Ghani, Daniel Porter, Gitali Halder, Anindya Dey, Andrew Pole, Dan Steinberg, Martin Chabbert, Martin Piotte, David Gondek, and Eva Helle. See this book’s Cast of “Characters” for more information about these practitioners and where they fit into the book.

Thanks to Eleazar Eskin and Ben Bullard for generous technical guidance on some hairy topics.

These proficient reviewers provided critical feedback on my writing, keeping it up to snuff: Mark Abdollahian, Vivek Ajmani, Laura Bahr, Femi Banjo, Anasse Bari, Carsten Boers, Richard Boire, Renee-Marie Brewster, Alexander Chaffee, Erin Cowan, Roger Craig, Kenny Darrell, Udi Dotan, Andrew Ferguson, Anthony Goldbloom, Mikhail Golovnya, Michael Grundhoefer, Paul Hofmann, Jason Howard, Hevel Jean-Baptiste, Tracy Jiang, Elise Johnson, Kathleen Kane, Karrie Karahalios, Joseph Kaszynski, Max Kuhn, Kim Larsen, Victor Lo, Vijay Mehrotra, Anne Milley, Linda Miner, Hamid Nemati, Al Nigl, Jesse Parelius, James Plotkin, Nicholas Radcliffe, Karl Rexer, Jacques Robin, Anne Schamberg, Jay Schamberg, Bill Simmon, Neil Skilling, Daniel Sokolov, Adam Sullivan, Patrick Surry, Astro Teller, Dana vanderHeyden, Marc vanderHeyden, Geert Verstraeten, Matthew Wagner, Phil Wagner, Maria Wang, Ezra Werb, Wlodek Zadrozny, and Margit Zwemer.

I gained access to this book’s extensive case study examples primarily through the productive and vibrant industry community that is Predictive Analytics World. I’d like to thank all the speakers and attendees of this conference series, as well as my business partners in its production, Matthew Finlay and Adam Kahn and their crackerjack team at Rising Media, who know how to make events excite and unite. This conference has helped catalyze and solidify the field, following predictive analytics from a nascent industry to a commercial movement.

Thanks to my assistant, Barbara Cook, for her endless efforts setting up this book’s extensive Notes, and to the supremely gifted designer Matt Kornhaas for the figures throughout these chapters.

Here’s a shout-out to the extra-special educators, of whom I had more than my fair share: Thomas McKean (kindergarten), Chip Porter (grades

4–6), Margaret O’Brien (Burlington High School, Vermont), Harry Mairson (Brandeis University), Richard Alterman (Brandeis University), James Pustejovsky (Brandeis University), and Kathleen McKeown (Columbia University).

The aforementioned have molded me and bolstered this book. Nevertheless, I alone take responsibility for errors or failings of any kind in its contents.

CHAPTER 1

Liftoff! Prediction Takes Action

How much guts does it take to deploy a predictive model into field operation, and what do you stand to gain? What happens when a man invests his entire life savings into his own predictive stock market trading system? Launching predictive analytics means to act on its predictions, applying what's been learned, what's been discovered within data. It's a leap many take—you can't win if you don't play.

In the mid-1990s, an ambitious postdoc researcher couldn't stand to wait any longer. After consulting with his wife, he loaded their entire life savings into a stock market prediction system of his own design—a contraption he had developed moonlighting on the side. Like Dr. Henry Jekyll imbibing his own untested potion in the moonlight, the young Dr. John Elder unflinchingly pressed “go.”

There is a scary moment every time new technology is launched. A spaceship lifting off may be the quintessential portrait of technological greatness and national prestige, but the image leaves out a small group of spouses terrified to the very point of psychological trauma. Astronauts are in essence stunt pilots, voluntarily strapping themselves in to serve as guinea pigs for a giant experiment, willing to sacrifice themselves in order to be part of history.

From grand challenges are born great achievements. We've taken strolls on our moon, and in more recent years a \$10 million Grand Challenge prize was awarded to the first nongovernmental organization to develop a reusable manned spacecraft. Driverless cars have been unleashed—“Look, Ma, no hands!” Fueled as well by millions of dollars in prize money, they navigate autonomously around the campuses of Google and BMW.

Replace the roar of rockets with the crunch of data, and the ambitions are no less far-reaching, “boldly going” not to space but to a new final

frontier: predicting the future. This frontier is just as exciting to explore, yet less dangerous and uncomfortable (outer space is a vacuum, and vacuums totally suck). Millions in grand challenge prize money go toward averting the unnecessary hospitalization of each patient and predicting the idiosyncratic preferences of each individual consumer. The TV quiz show *Jeopardy!* awarded \$1.5 million in prize money for a face-off between man and machine that demonstrated dramatic progress in predicting the answers to questions (IBM invested a lot more than that to achieve this win, as detailed in Chapter 6). Organizations are literally keeping kids in school, keeping the lights on, and keeping crime down with predictive analytics (PA). And success is its own reward when analytics wins a political election, a baseball championship, or . . . did I mention managing a financial portfolio?

Black-box trading—driving financial trading decisions automatically with a machine—is the holy grail of data-driven decision making. It's a black box into which current financial environmental conditions are fed, with buy/hold/sell decisions spit out the other end. It's black (i.e., opaque) because you don't care what's on the inside, as long as it makes good decisions. When working, it trumps any other conceivable business proposal in the world: Your computer is now a box that turns electricity into money.

And so with the launch of his stock trading system, John Elder took on his own personal grand challenge. Even if stock market prediction would represent a giant leap for mankind, this was no small step for John himself. It's an occasion worthy of mixing metaphors. By putting all his eggs into one analytical basket, John was taking a healthy dose of his own medicine.

Before continuing with the story of John's blast-off, let's establish how launching a predictive system works, not only for black-box trading but across a multitude of applications.

GOING LIVE

Learning from data is virtually universally useful. Master it and you'll be welcomed nearly everywhere!

—John Elder

New groundbreaking stories of PA in action are pouring in. A few key ingredients have opened these floodgates:

- wildly increasing loads of data;
- cultural shifts as organizations learn to appreciate, embrace, and integrate predictive technology;
- improved software solutions to deliver PA to organizations.

But this flood built up its potential in the first place simply because predictive technology boasts an inherent generality—there are just so many conceivable ways to make use of it. Want to come up with your own new innovative use for PA? You need only two ingredients.

EACH APPLICATION OF PA IS DEFINED BY:

1. **What's predicted:** the kind of behavior (i.e., action, event, or happening) to predict for each individual, stock, or other kind of element.
2. **What's done about it:** the decisions driven by prediction; the action taken by the organization in response to or informed by each prediction.

Given its open-ended nature, the list of application areas is so broad and the list of example stories is so long that it presents a minor data-management challenge in and of itself! So I placed this big list (182 examples total) into nine tables in the center of this book. Take a flip through to get a feel for just how much is going on. That's the sexy part—it's the “centerfold” of this book. The Central Tables divulge cases of predicting: stock prices, risk, delinquencies, accidents, sales, donations, clicks, cancellations, health problems, hospital admissions, fraud, tax evasion, crime, malfunctions, oil flow, electricity outages, approvals for government benefits, thoughts, intention, answers, opinions, lies, grades, dropouts, friendship, romance, pregnancy, divorce, jobs, quitting, wins, votes, and more. The application areas are growing at a breakneck pace.

Within this long list, the quintessential application for business is the one covered in the Introduction for mass marketing:

PA APPLICATION: TARGETING DIRECT MARKETING

- 1. What's predicted:** Which customers will respond to marketing contact.
- 2. What's done about it:** Contact customers more likely to respond.

As we saw, this use of PA illustrates *The Prediction Effect*.

The Prediction Effect: *A little prediction goes a long way.*

Let's take a moment to see how straightforward it is to calculate the sheer value resulting from The Prediction Effect. Imagine you have a company with a mailing list of a million prospects. It costs \$2 to mail to each one, and you have observed that one out of 100 of them will buy your product (i.e., 10,000 responses). You take your chances and mail to the entire list.

If you profit \$220 for each rare positive response, then you pocket:

$$\begin{aligned}\text{Overall profit} &= \text{Revenue} - \text{Cost} \\ &= (\$220 \times 10,000 \text{ responses}) - (\$2 \times 1 \text{ million})\end{aligned}$$

Whip out your calculator—that's \$200,000 profit. Are you happy yet? I didn't think so.

If you are new to the arena of direct marketing (welcome!), you'll notice we're playing a kind of wild numbers game, amassing great waste, like one million monkeys chucking darts across a chasm in the general direction of a dartboard. As turn-of-the-century marketing pioneer John Wanamaker famously put it, "Half the money I spend on advertising is wasted; the trouble is I don't know which half." The bad news is that it's actually more than half; the good news is that PA can learn to do better.

A FAULTY ORACLE EVERYONE LOVES

The first step toward predicting the future is admitting you can't.

—Stephen Dubner, Freakonomics Radio, March 30, 2011

The “prediction paradox”: The more humility we have about our ability to make predictions, the more successful we can be in planning for the future.

—Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—but Some Don’t*

Your resident “oracle,” PA, tells you which customers are most likely to respond. It earmarks a quarter of the entire list and says, “These folks are three times more likely to respond than average!” So now you have a short list of 250,000 customers of whom 3 percent will respond—7,500 responses.

Oracle, shmoracle! These predictions are seriously inaccurate—we still don’t have strong confidence when contacting any one customer, given this measly 3 percent response rate. However, the overall IQ of your dart-throwing monkeys has taken a real boost. If you send mail to only this short list then you profit:

$$\begin{aligned}\text{Overall profit} &= \text{Revenue} - \text{Cost} \\ &= (\$220 \times 7,500 \text{ responses}) - (\$2 \times 250,000)\end{aligned}$$

That’s \$1,150,000 profit. You just improved your profit 5.75 times over by mailing to *fewer* people (and, in so doing, expending fewer trees). In particular, you predicted who wasn’t worth contacting and simply left them alone. Thus you cut your costs by three-quarters in exchange for losing only one-quarter of sales. That’s a deal I’d take any day.

It’s not hard to put a value on prediction. As you can see, even if predictions themselves are generated from sophisticated mathematics, it takes only simple arithmetic to roll up the plethora of predictions—some accurate, and others not so much—and reveal the aggregate bottom-line effect. This isn’t just some abstract notion; The Prediction Effect means business.

PREDICTIVE PROTECTION

Thus, value has emerged from just a little predictive insight, a small prognostic nudge in the right direction. It's easy to draw an analogy to science fiction, where just a bit of supernatural foresight can go a long way. Nicolas Cage kicks some serious bad-guy butt in the movie *Next*, based on a story by Philip K. Dick. His weapon? Pure prognostication. He can see the future, but only two minutes ahead. It's enough prescience to do some damage. An unarmed civilian with a soft heart and the best of intentions, he winds up marching through something of a war zone, surrounded by a posse of heavily armed FBI agents who obey his every gesture. He sees the damage of every booby trap, sniper, and mean-faced grunt before it happens and so can command just the right moves for this Superhuman Risk-Aversion Team, avoiding one calamity after another.

In a way, deploying PA makes a Superhuman Risk-Aversion Team of the organization just the same. Every decision an organization makes, each step it takes, incurs risk. Imagine the protective benefit of foreseeing each pitfall so that it may be avoided—each criminal act, stock value decline, hospitalization, bad debt, traffic jam, high school dropout . . . and each ignored marketing brochure that was a waste to mail. *Organizational risk management*, traditionally the act of defending against singular, macrolevel incidents like the crash of an aircraft or an economy, now broadens to fight a myriad of microlevel risks.

Hey, it's not all bad news. We win by foreseeing good behavior as well, since it often signals an opportunity to gain. The name of the game is “Predict ‘n’ Pounce” when it pops up on the radar that a customer is likely to buy, a stock value is likely to increase, a voter is likely to swing, or the apple of one’s online dating eye is likely to reciprocate.

A little glimpse into the future gives you power because it gives you options. In some cases the obvious decision is to act in order to avert what may not be inevitable, be it crime, loss, or sickness. On the positive side, in the case of foreseeing demand, you act to exploit it. Either way, prediction serves to drive decisions.

Let's turn to a real case, a \$1 million example.

A SILENT REVOLUTION WORTH A MILLION

When an organization goes live with PA, it unleashes a massive army, but it's an army of ants. These ants march out to the front lines of an organization's operations, the places where there's contact with the likes of customers, students, or patients—the people served by the organization. Within these interactions, the ant army, guided by predictions, improves millions of small decisions. The process goes largely unnoticed, under the radar . . . until someone bothers to look at how it's adding up. The improved decisions may each be ant-sized, relatively speaking, but there are so many that they come to a powerful net effect.

In 2005, I was digging in the trenches, neck deep in data for a client who wanted more clicks on their website. To be precise, they wanted more clicks on their sponsors' ads. This was about the money—more clicks, more money. The site had gained tens of millions of users over the years, and within just several months' worth of tracking data that they handed me, there were 50 million rows of learning data—no small treasure trove from which to learn to predict . . . *clicks*.

Advertising is an inevitable part of media, be it print, television, or your online experience. Benjamin Franklin forgot to include it when he proclaimed, "In this world nothing can be said to be certain, except death and taxes." The flagship Internet behemoth Google credits ads as its greatest source of revenue. It's the same with Facebook.

But on this website, ads told a slightly different story than usual, which further amplified the potential win of predicting user clicks. The client was a leading student grant and scholarship search service, with one in three college-bound high school seniors using it: an arcane niche, but just the one over which certain universities and military recruiters were drooling. One ad for a university included a strong pitch, naming itself "America's leader in creative education" and culminating with a button that begged to be clicked: "Yes, please have someone from the Art Institute's Admissions Office contact me!" And you won't be surprised to hear that creditors were also placing ads, at the ready to provide these students another source of funds: loans. The sponsors would pay up to \$25 per lead—for each

would-be recruit. That's good compensation for one little click of the mouse. What's more, since the ads were largely relevant to the users, closely related to their purpose on the website, the response rates climbed up to an unusually high 5 percent. So this little business, owned by a well-known online job-hunting firm, was earning well. Any small improvement meant real revenue.

But improving ad selection is a serious challenge. At certain intervals, users were exposed to a full-page ad, selected from a pool of 291 options. The trick is selecting the best one for each user. The website currently selected which ad to show based simply on the revenue it generated on average, with no regard to the particular user. The universally strongest ad was always shown first. Although this tactic forsakes the possibility of matching ads to individual users, it's a formidable champion to unseat. Some sponsor ads, such as certain universities, paid such a high bounty per click, and were clicked so often, that showing any user a less powerful ad seemed like a crazy thing to consider, since doing so would risk losing currently established value.

THE PERILS OF PERSONALIZATION

By trusting predictions in order to customize for the individual, you take on risk. A predictive system boldly proclaims, "Even though ad A is so strong overall, for this particular user it is worth the risk of going with ad B." For this reason, most online ads are not personalized for the individual user—even Google's AdWords, which allows you to place textual ads alongside search results as well as on other Web pages, determines which ad to display by Web page context, the ad's click rate, and the advertiser's bid (what it is willing to pay for a click). It is not determined by anything known or predicted about the particular viewer who is going to actually see the ad.

But weathering this risk carries us to a new frontier of customization. For business, it promises to "personalize!," "increase relevance!," and "engage one-to-one marketing!" The benefits reach beyond personalizing marketing treatment to customizing the individual treatment of patients and suspected criminals as well. During a speech about satisfying our widely varying preferences in choice of spaghetti sauce—chunky? sweet? spicy?—Malcolm

Gladwell said, “People . . . were looking for . . . universals, they were looking for one way to treat all of us[;] . . . all of science through the nineteenth century and much of the twentieth was obsessed with universals. Psychologists, medical scientists, economists were all interested in finding out the rules that govern the way all of us behave. But that changed, right? What is the great revolution of science in the last 10, 15 years? It is the movement from the search for universals to the understanding of variability. Now in medical science we don’t want to know . . . just how cancer works; we want to know how your cancer is different from my cancer.”

From medical issues to consumer preferences, individualization trumps universals. And so it goes with ads:

PA APPLICATION: PREDICTIVE ADVERTISEMENT TARGETING

- 1. What’s predicted:** Which ad each customer is most likely to click.
- 2. What’s done about it:** Display the best ad (based on the likelihood of a click as well as the bounty paid by its sponsor).

I set up PA to perform ad targeting for my client, and the company launched it in a head-to-head, champion/challenger competition to the death against their existing system. The loser would surely be relegated to the bin of second-class ideas that just don’t make as much cash. To prepare for this battle, we armed PA with powerful weaponry. The predictions were generated from machine learning across 50 million learning cases, each depicting a microlesson from history of the form, “User Mary was shown ad A and she did click it” (a positive case) or “User John was shown ad B and he did not click it” (a negative case).

The learning technology employed to pick the best ad for each user was a Naïve Bayes model. Rev. Thomas Bayes was an eighteenth-century mathematician, and the “Naïve” part means that we take a very smart man’s ideas and compromise them in a way that simplifies yet makes their application feasible, resulting in a practical method that’s often considered good enough at prediction and scales to the task at hand. I went with this method for its relative simplicity, since in fact I needed to generate 291 such models, one for each ad. Together, these models predict which ad a user is most likely to click on.

DEPLOYMENT'S DETOURS AND DELAYS

As with a rocket ship, launching PA looks great on paper. You design and construct the technology, place it on the launchpad, and wait for the green light. But just when you're about to hit "go," the launch is scrubbed. Then delayed. Then scrubbed again. The Wright brothers and others, galvanized by the awesome promise of a newly discovered wing design that generates lift, endured an uncharted, rocky road, faltering, floundering, and risking life and limb until all the kinks were out.

For ad targeting and other real-time PA deployments, predictions have got to zoom in at warp speed in order to provide value. Our online world tolerates no delay when it's time to choose which ad to display, determine whether to buy a stock, decide whether to authorize a credit card charge, recommend a movie, filter an e-mail for viruses, or answer a question on *Jeopardy!* A real-time PA solution must be directly integrated into operational systems, such as websites or credit card processing facilities. If you are newly integrating PA within an organization, this can be a significant project for the software engineers, who often have their hands full with maintenance tasks just to keep the business operating normally. Thus, the *deployment* phase of a PA project takes much more than simply receiving a nod from senior management to go live: It demands major construction. By the time the programmers deployed my predictive ad selection system, the data over which I had tuned it was already about 11 months old. Were the facets of what had been learned still relevant almost one year later, or would prediction's power peter out?

IN FLIGHT

*This is Major Tom to Ground Control
I'm stepping through the door
And I'm floating in a most peculiar way . . .*

—“Space Oddity” by David Bowie

Once launched, PA enters an eerie, silent waiting period, like you're floating in orbit and nothing is moving. But the fact is, in a low orbit around Earth you're actually screaming along at over 14,000 miles per hour. Unlike the drama of launching a rocket or erecting a skyscraper, the launch of PA is a

relatively stealthy maneuver. It goes live, but daily activities exhibit no immediately apparent change. After the ad-targeting project's launch, if you checked out the website, it would show you an ad as usual, and you could wonder whether the system made any difference in this one choice. This is what computers do best. They hold the power to silently enact massive procedural changes that often go uncredited, since most aren't directly witnessed by any one person.

But, under the surface, a sea change is in play, as if the entire ocean has been reconfigured. You actually notice the impact only when you examine an aggregated report.

In my client's deployment, predictive ad selection triumphed. The client conducted a head-to-head comparison, selecting ads for half the users with the existing champion system and the other half with the new predictive system, and reported that the new system generated at least 3.6 percent more revenue, which amounts to \$1 million every 19 months, given the rate at which revenue was already coming in. This was for the website's full-page ads only; many more (smaller) ads are embedded within functional Web pages, which could potentially also be boosted with a similar PA project.

No new customers, no new sponsors, no changes to business contracts, no materials or computer hardware needed, no new full-time employees or ongoing effort—solely an improvement to decision making was needed to generate cold, hard cash. In a well-oiled, established system like the one my client had, even a small improvement of 3.6 percent amounts to something substantial. The gains of an incremental tweak can be even more dramatic: In the insurance business, one company reports that PA saves almost \$50 million annually by decreasing its loss ratio by *half a percentage point*.

So how did these models predict each click?

ELEMENTARY, MY DEAR: THE POWER OF OBSERVATION

Just like Sherlock Holmes drawing conclusions by sizing up a suspect, prediction comes of astute observation: What's known about each individual provides a set of clues about what he or she may do next. The chance a user will click on a certain ad depends on all sorts of elements, including the individual's current school year, gender, and e-mail domain

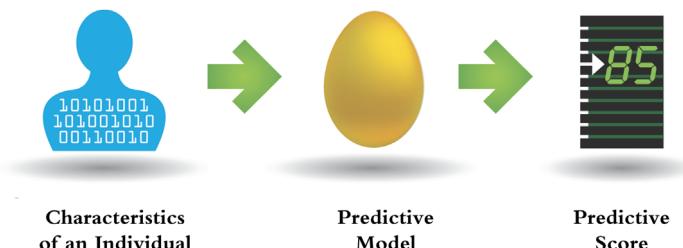
(Hotmail, Yahoo, Gmail, etc.); the ratio of the individual's SAT written-to-math scores (is the user more a verbal person or more a math person?), and on and on.

In fact, this website collected a wealth of information about its users. To find out which grants and scholarships they're eligible for, users answer dozens of questions about their school performance, academic interests, extracurricular activities, prospective college majors, parents' degrees, and more. So the table of learning data was long (at 50 million examples) and was also wide, with each row holding all the information known about the user at the moment the person viewed an ad.

It can sound like a tall order: *harnessing millions of examples in order to learn how to incorporate the various factoids known about each individual so that prediction is possible*. But we can break this down into a couple of parts, and suddenly it gets much simpler. Let's start with the contraption that makes the predictions, the electronic Sherlock Holmes that knows how to consider all these factors and roll them up into a single prediction for the individual.

Predictive model—*a mechanism that predicts a behavior of an individual, such as click, buy, lie, or die. It takes characteristics of the individual as input and provides a predictive score as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior.*

A predictive model (depicted throughout this book as a “golden” egg, albeit in black and white) scores an individual:



A predictive model is the means by which the attributes of an individual are factored together for prediction. There are many ways to do this. One is to weigh each characteristic and then add them up—perhaps females boost their score by 33.4, Hotmail users decrease their score by 15.7, and so on.

Each element counts toward or against the final score for that individual. This is called a *linear model*, generally considered quite simple and limited, although usually much better than nothing.

Other models are composed of *rules*, like this real example:

IF the individual
is still in high school
AND
expects to graduate college within three years
AND
indicates certain military interest
AND
has not been shown this ad yet
THEN the probability of clicking on the ad for the Art Institute is
13.5 percent.

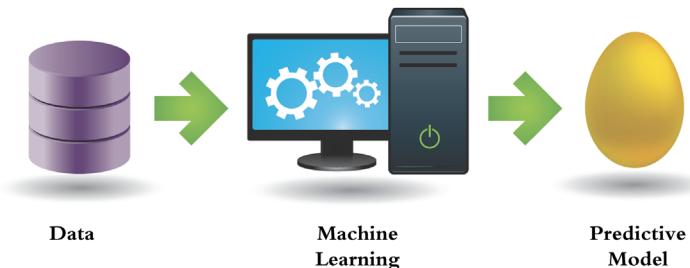
This rule is a valuable find, since the overall probability of responding to the Art Institute's ad is only 2.7 percent, so we've identified a pocket of avid clickers, relatively speaking.

It is interesting that those who have indicated a military interest are more likely to show interest in the Art Institute. We can speculate, but it's important not to assume there is a *causal* relationship. For example, it may be that people who complete more of their profile are just more likely to click in general, across all kinds of ads.

Various types of models compete to make the most accurate predictions. Models that combine a bunch of rules like the one just shown are—relatively speaking—on the simpler side. Alternatively, we can go more “supermath” on the prediction problem, employing complex formulas that predict more effectively but are almost impossible to understand by human eyes.

But all predictive models share the same objective: They consider the various factors of an individual in order to derive a single predictive score for that individual. This score is then used to drive an organizational decision, guiding which action to take.

Before using a model, we've got to build it. Machine learning builds the predictive model:



Machine learning crunches data to build the model, a brand-new prediction machine. The model is the product of this learning technology—it is itself the very thing that has been learned. For this reason, machine learning is also called *predictive modeling*, which is a more common term in the commercial world. If deferring to the older metaphorical term *data mining*, the predictive model is the unearthed gem.

Predictive modeling generates the entire model from scratch. All the model's math, weights, or rules are created automatically by the computer. The machine learning process is designed to accomplish this task, to mechanically develop new capabilities from data. This automation is the means by which PA builds its predictive power.

The hunter returns back to the tribe, proudly displaying his kill. So, too, a data scientist posts her model on the bulletin board near the company ping-pong table. The hunter hands over the kill to the cook, and the data scientist cooks up her model, translates it to a standard computer language, and e-mails it to an engineer for integration. A well-fed tribe shows the love; a psyched executive issues a bonus.

TO ACT IS TO DECIDE

Knowing is not enough; we must act.

—Johann Wolfgang von Goethe

Once you develop a model, don't pat yourself on the back just yet. Predictions don't help unless you do something about them. They're just thoughts, just

ideas. They may be astute, brilliant gems that glimmer like the most polished of crystal balls, but displaying them on a shelf gains you nothing—they just sit there and look smart.

Unlike a report sitting dormant on the desk, PA leaps out of the lab and takes action. In this way, it stands above other forms of analysis, data science, and data mining. It desires deployment and loves to be launched—because, in what it foretells, it mandates movement.

The predictive score for each individual directly informs the decision of what action to take with that individual. Doctors take a second look at patients predicted to be readmitted, and service agents contact customers predicted to cancel. Predictive scores issue imperatives to *mail, call, offer a discount, recommend a product, show an ad, expend sales resources, audit, investigate, inspect for flaws, approve a loan, or buy a stock*. By acting on the predictions produced by machine learning, the organization is now applying what's been learned, modifying its everyday operations for the better.

To make this point, we have mangled the English language. Proponents like to say that PA is *actionable*. Its output directly informs actions, commanding the organization about what to do next. But with this use of vocabulary, industry insiders have stolen the word *actionable*, which originally meant *worthy of legal action* (i.e., “sue-able”), and morphed it. They did so because they’re tired of seeing sharp-looking reports that provide only a vague, unsure sense of direction.

With this word’s new meaning established, “your fly is unzipped” is *actionable* (it is clear what to do—you can and should take action to remedy), but “you’re going bald” is not (there’s no cure; nothing to be done). Better yet, “I predict you will buy these button-fly jeans and this snazzy hat” is actionable to a salesperson.

Launching PA into action delivers a critical new edge in the competitive world of business. One sees massive commoditization taking place today as the faces of corporations appear to blend together. They all seem to sell pretty much the same thing and act in pretty much the same ways. To stand above the crowd, where can a company turn?

As Thomas Davenport and Jeanne Harris put it in *Competing on Analytics: The New Science of Winning*, “At a time when companies in many industries offer similar products and use comparable technology, high-performance business

processes are among the last remaining points of differentiation.” Enter PA. Survey results have in fact shown that “a tougher competitive environment” is by far the strongest reason why organizations adopt this technology.

But while the launch of PA brings real change, it can also wreak havoc by introducing new risk. With this in mind, we now return to John’s story.

A PERILOUS LAUNCH

Dr. John Elder bet it all on a predictive model. He concocted it in the lab, packed it into a black box, and unleashed it on the stock market. Some people make their own bed in which they must then passively lie. But John had climbed way up high to take a leap of faith. Diving off a mountaintop with newly constructed, experimental wings, he wondered how long it might take before he could be sure he was flying rather than crashing.

The risks stared John in the face. His and his wife’s full retirement savings were in the hands of an experimental device, launched into oblivion and destined for one of the same two outcomes achieved by every rocket: glory or mission failure. Discovering profitable market patterns that sustain is the mission of thousands of traders operating in what John points out is a brutally competitive environment; doing so automatically with machine learning is the most challenging of ambitions, considered impossible by many. It doesn’t help that a stock market scientist is completely on his own, since work in this area is shrouded in secrecy, leaving virtually no potential to learn from the successes and failures of others. Academics publish, marketers discuss, but quants hide away in their Batcaves. What can look great on paper might be stricken with a weakness that destroys or an error that bankrupts. John puts it plainly: “Wall Street is the hardest data mining problem.”

The evidence of danger was palpable, as John had recently uncovered a crippling flaw in an existing predictive trading system and personally escorted it to its grave. Opportunity had come knocking on the door of a small firm called Delta Financial in the form of a black-box trading system purported to predict movements of the Standard & Poor’s (S&P) 500 with 70 percent accuracy. Built by a proud scientist, the system promised to make millions, so stakeholders were flying around all dressed up in suits, actively lining up

investors prepared to place a huge bet. Among potential early investors, Delta was leading the way for others, taking a central, influential role. The firm was known for investigating and championing cutting-edge approaches, weathering the risk inherent to innovation. As a necessary precaution, Delta sought to empirically validate this system. The firm turned to John, who was consulting for them on the side while pursuing his doctorate at the University of Virginia in Charlottesville. John's work for Delta often involved inspecting, and sometimes debunking, black-box trading systems.

How do you prove a machine is broken if you're not allowed to look inside it? Healthy skepticism bolstered John's resolve, since the claimed 70 percent accuracy raised red flags as quite possibly too darn good to be true. But he was not granted access to the predictive model. With secrecy reigning supreme, the protocol for this type of audit dictated that John receive only the numerical results, along with a few adjectives that described its design: *new, unique, powerful!* With meager evidence, John sought to prove a crime he couldn't even be sure had been committed.

Before each launch, organizations establish confidence in PA by "predicting the past" (aka backtesting). The predictive model must prove itself on historical data before its deployment. Conducting a kind of simulated prediction, the model evaluates across data from last week, last month, or last year. Feeding on input that could only have been known at a given time, the model spits out its prediction, which then matches against what we now already know took place thereafter. Would the S&P 500 go down or up on March 21, 1991? If the model gets this retrospective question right, based only on data available by March 20, 1991 (the day just before), we have evidence the model works. These retrospective predictions—without the manner in which they had been derived—were all John had to work with.

HOUSTON, WE HAVE A PROBLEM

Even the most elite of engineers commit the most mundane and costly of errors. In late 1998, NASA launched the Mars Climate Orbiter on a daunting nine-month trip to Mars, a mission that fewer than half the world's launched probes headed for that destination have completed successfully. This \$327.6 million

calamity crashed and burned, due not to the flip of fate's coin, but rather a simple snafu. The spacecraft came too close to Mars and disintegrated in its atmosphere. The source of the navigational bungle? One system expected to receive information in metric units (newton-seconds), but a computer programmer for another system had it speak in English imperial units (pound-seconds). Oops.

John stared at a screen of numbers, wondering if anything was wrong and, if so, whether he could find it. From the long list of impressive—yet retrospective—predictions, he plainly saw the promise of huge profits that had everyone involved so excited. If he proved there was a flaw, vindication; if not, lingering uncertainty. The task at hand was to reverse engineer: Given the predictions the system generated, could he infer how it worked under the hood, essentially eking out the method in its madness? This was ironic, since all predictive modeling is a kind of reverse engineering to begin with. Machine learning starts with the data, an encoding of things that have happened, and attempts to uncover patterns that generated or explained the data in the first place. John was attempting to deduce what the other team had deduced. His guide? Informal hunches and ill-informed inferences, each of which could be pursued only by way of trial and error, testing each hypothetical mess-up he could dream up by programming it by hand and comparing it to the retrospective predictions he had been given.

His perseverance finally paid off: John uncovered a true flaw, thereby flinging back the curtain to expose a flustered Wizard of Oz. It turned out that the prediction engine committed the most sacrilegious of cheats by looking at the one thing it must not be permitted to see. It had looked at the future. The battery of impressive retrospective predictions weren't true predictions at all. Rather, they were based in part on a three-day average calculated across yesterday, today . . . and tomorrow. The scientists had probably intended to incorporate a three-day average leading up to today, but had inadvertently shifted the window by a day. Oops. This crippling bug delivered the dead-certain prognosis that this predictive model would not perform well if deployed into the field. Any prediction it would generate today could not incorporate the very thing it was designed to foresee—tomorrow's stock price—since, well, it isn't known yet. So, if foolishly deployed, its accuracy could never match the exaggerated performance falsely demonstrated across

the historical data. John revealed this bug by reverse engineering it. On a hunch, he handcrafted a method with the same type of bug and showed that its predictions closely matched those of the trading system.

A predictive model will sink faster than the *Titanic* if you don't seal all its "time leaks" before launch. But this kind of "leak from the future" is common, if mundane. Although core to the very integrity of prediction, it's an easy mistake to make, given that each model is backtested over historical data for which prediction is not, strictly speaking, possible. The relative future is always readily available in the testing data, easy to inadvertently incorporate into the very model trying to predict it. Such temporal leaks achieve status as a commonly known gotcha among PA practitioners. If this were an episode of *Star Trek*, our beloved, hypomanic engineer Scotty would be screaming, "Captain, we're losing our temporal integrity!"

It was with no pleasure that John delivered the disappointing news to his client, Delta Financial: He had debunked the system, essentially exposing it as inadvertent fraud. High hopes were dashed as another fairy tale bit the dust, but gratitude quickly ensued as would-be investors realized they'd just dodged a bullet. The wannabe inventor of the system suffered dismay but was better off knowing now; it would have hit the fan much harder postlaunch, possibly including prosecution for fraud, even if inadvertently committed. The project was aborted.

THE LITTLE MODEL THAT COULD

Even the young practitioner that he was, John was a go-to data man for entrepreneurs in black-box trading. One such investor moved to Charlottesville, but only after John Elder, PhD, new doctorate degree in hand, had just relocated to Houston in order to continue his academic rite of passage with a postdoc research position at Rice University. He'd left quite an impression back in Charlottesville, though; people in both the academic and commercial sectors alike referred the investor to John. Despite John's distance, the investor hired him to prepare, launch, and monitor a new black-box mission remotely from Houston. It seemed as good a place as any for the project's Mission Control.

And so it was time for John to move beyond the low-risk role of evaluating other people's predictive systems and dare to build one of his own. Over several months, he and a small team of colleagues built upon core insights from the investor and produced a new, promising black-box trading model. John was champing at the bit to launch it and put it to the test. All the stars were aligned for liftoff except one: The money people didn't trust it yet.

There was good reason to believe in John. Having recently completed his doctorate degree, he was armed with a fresh, talented mind, yet had already gained an impressively wide range of data-crunching problem-solving experience. On the academic side, his PhD thesis had broken records among researchers as the most efficient way to optimize for a certain broad class of system engineering problems (machine learning is itself a kind of optimization problem). He had also taken on predicting the species of a bat from its echolocation signals (the chirps bats make for their radar). And in the commercial world, John's pregrad positions had dropped him right into the thick of machine learning systems that steer for aerospace flight and that detect cooling pipe cracks in nuclear reactors, not to mention projects for Delta Financial looking over the shoulders of other black-box quants.

And now John's latest creation absolutely itched to be deployed. Backtesting against historical data, all indications whispered confident promises for what this thing could do once set in motion. As John puts it, "A slight pattern emerged from the overwhelming noise; we had stumbled across a persistent pricing inefficiency in a corner of the market, a small edge over the average investor, which appeared repeatable." Inefficiencies are what traders live for. A perfectly efficient market can't be played, but if you can identify the right imperfection, it's payday.

PA APPLICATION: BLACK-BOX TRADING

- 1. What's predicted:** Whether a stock will go up or down.
- 2. What's done about it:** Buy stocks that will go up; sell those that will go down.

John could not get the green light. As he strove to convince the investor, cold feet prevailed. It appeared they were stuck in a stalemate. After all, this

guy might not get past his jitters until he could see the system succeed, yet it couldn't succeed while stuck on the launchpad. The time was now, as each day marked lost opportunity.

After a disconcerting meeting that seemed to go nowhere, John went home and had a sit-down with his wife, Elizabeth. What supportive spouse could possibly resist the seduction of her beloved's ardent excitement and strong belief in his own abilities? She gave him the go-ahead to risk it all, a move that could threaten their very home. But he still needed buy-in from one more party.

Delivering his appeal to the client investor raised questions, concerns, and eyebrows. John wanted to launch with his own personal funds, which meant no risk whatsoever to the client and would resolve any doubts by field-testing John's model. But this unorthodox step would be akin to the dubious choice to act as one's own defense attorney. When an individual is without great personal means, this kind of thing is often frowned upon. It conveys overconfident, foolish brashness. Even if the client wanted to truly believe, it would be another thing to expect the same from coinvestors who hadn't gotten to know and trust John. But with every launch, proponents gamble something fierce. John had set the rules for the game he'd chosen to play.

He received his answer from the investor: "Go for it!" This meant there was nothing to prevent moving forward. It could have also meant the investor was prepared to write off the project entirely, feeling there was nothing left to lose.

HOUSTON, WE HAVE LIFTOFF

Practitioners of PA often put their own professional lives a bit on the line to push forward, but this case was extreme. Like baseball's Billy Beane of the Oakland A's, who literally risked his entire career to deploy and field-test an analytical approach to team management, John risked everything he had. It was early 1994, and John's individual retirement account (IRA) amounted to little more than \$40,000. He put it all in.

“Going live with black-box trading is really exciting and really scary,” says John. “It’s a roller coaster that never stops. The coaster takes on all these thrilling ups and downs, but with a very real chance it could go off the rails.”

As with baseball, he points out, slumps aren’t slumps at all—they’re inevitable statistical certainties. Each one leaves you wondering, “Is this falling feeling part of a safe ride, or is something broken?” A key component to his system was a cleverly designed means to detect real quality, a measure of system integrity that revealed whether recent success had been truly deserved or had come about just due to dumb luck.

From the get-go, the predictive engine rocked. It increased John’s assets at a rate of 40 percent per year, which meant that after two years his money had doubled.

The client investor was quickly impressed and soon put in a couple of million dollars himself. A year later, the predictive model was managing a \$20 million fund across a group of investors, and eventually the investment pool increased to a few hundred million dollars. With this much on tap, every win of the system was multiplicatively magnified.

No question about it: All involved relished this fiesta, and the party raged on and on, continuing almost nine years, consistently outperforming the overall market all along. The system chugged, autonomously trading among a dozen market sectors such as technology, transportation, and healthcare. John says the system “beat the market each year and exhibited only two-thirds its standard deviation—a home run as measured by risk-adjusted return.”

But all good things must come to an end, and just as John had talked his client up, he later had to talk him down. After nearly a decade, the key measure of system integrity began to decline. John was adamant that they were running on fumes, so with little ceremony the entire fund was wound down. The system was halted in time, before catastrophe could strike. In the end, all the investors came out ahead.

A PASSIONATE SCIENTIST

The early success of this streak had quickly altered John’s life. Once the project was cruising, he had begun supporting his rapidly growing family

with ease. The project was taking only a couple of John's hours each day to monitor, tweak, and refresh what was a fundamentally stable, unchanging method within the black box. What's a man to do? Do you put your feet up and sip wine indefinitely, with the possible interruption of family trips to Disney World? After all, John had thus far always burned the candle at both ends out of financial necessity, with summer jobs during college, part-time work during graduate school, and this black-box project, which itself had begun as a moonlighting gig during his postdoc. Or do you follow the logical business imperative: Pounce on your successes, using all your free bandwidth to find ways to do more of the same?

John's passion for the craft transcended these self-serving responses to his good fortune. That is to say, he contains the spirit of the geek. He jokes about the endless insatiability of his own appetite for the stimulation of fresh scientific challenges. He's addicted to tackling something new. There is but one antidote: a growing list of diverse projects. So, two years into the stock market project, he wrapped up his postdoc, packed up his family, and moved back to Charlottesville to start his own data mining company.

And so John launched Elder Research, now the largest predictive analytics services firm (pure play) in North America. A narrow focus is key to the success of many businesses, but Elder Research's advantage is quite the opposite: its diversity. The company's portfolio reaches far beyond finance to include all major commercial sectors and many branches of government. John has also earned a top-echelon position in the industry. He coauthors massive textbooks, frequently chairs or keynotes at Predictive Analytics World conferences, takes cameos as a university professor, and served five years as a presidential appointee on a national security technology panel.

LAUNCHING PREDICTION INTO INNER SPACE

With stories like John's coming to light, organizations are jumping on the PA bandwagon. One such firm, a mammoth international organization, focuses the power of prediction introspectively, casting PA's keen gaze on its own employees. Read on to witness the windfall and the fallout when scientists dare to ask: Do people like being predicted?

CHAPTER 2

With Power Comes Responsibility

Hewlett-Packard, Target, the Cops, and the NSA Deduce Your Secrets

How do we safely harness a predictive machine that can foresee job resignation, pregnancy, and crime? Are civil liberties at risk? Why does one leading health insurance company predict policyholder death? Two extended sidebars explore: (1) Does the government undertake fraud detection more for its citizens or for self-preservation, and (2) for what compelling purpose does the National Security Agency (NSA) need your data even if you have no connection to crime whatsoever, and can the agency use machine learning supercomputers to fight terrorism without endangering human rights?

Predictive analytics . . . is right at the fulcrum point of utopian and dystopian visions of the future.

—Andrew Frank, Research Vice President, Gartner

What would happen if your boss were notified that you’re allegedly going to quit—even though you had said this to no one? If you are one of the more than 300,000 who work at Hewlett-Packard (HP), your employer has tagged you—and all your colleagues—with a “Flight Risk” score. This simple number foretells whether you’re likely to leave your job. As an HP employee, there’s a good chance you didn’t already know that. Postpone freaking out until you finish reading the full explanation in this chapter.

This story about HP arrived in the wake of media outcry against Target in 2012 after learning the big-box retailer had taken to predicting customer pregnancy. The media firestorm invoked misleading accusations, fear of corporate power, postulations by television personalities, and, of course, predictive analytics (PA). To my surprise, I ended up in the thick of it.

TV news programs strike like a blunt instrument, but often in the right general direction. The media assault was reactionary and chose to misinform, yet legitimate quandaries lurk below the surface. Target's and HP's predictive power brings to focus an exceptionally challenging and pressing ethical question. Within the minefield that is the privacy debate, the stakes just rose even higher.

Why? Because prediction snoops into your private future. These cases involve the corporate deduction of previously unknown, sensitive facts: Are you considering quitting your job? Are you pregnant? This isn't a case of mishandling, leaking, or stealing data. Rather, it is *the generation of new data*, the indirect discovery of unvolunteered truths about people. Organizations predict these powerful insights from existing innocuous data, as if creating them out of thin air. Are they equipped to manage their own creation?

While we come to terms with the sheer magnitude of prediction's power, we've only begun to fathom the privacy concerns it introduces. A chain reaction triggers and surprises even the experts: Organizations exert newfound capabilities, consumers rise up, the media stir the pot, and scientists dodge bullets and then reexamine scruples.

The journey eventually takes us to a particularly uncomfortable dilemma. Beyond expectant moms and departing employees, PA also flags potential criminals and actively helps law enforcement decide who stays in prison and who goes free.

This tale follows my journey from carefree technologist to unwitting talking head and the journey of organizations from headstrong to humbled. The asocial domain of data and analytics is not so irrelevant after all.

THE PREDICTION OF TARGET AND THE TARGET OF PREDICTION

In 2010, I invited an expert at Target, Andrew Pole, to keynote at Predictive Analytics World, the conference series I founded. Pole manages dozens of analytics professionals who run various PA projects at Target. In October of that year, Pole delivered a stellar speech on a wide range of PA deployments at Target. He took the stage and dynamically engaged the audience,

revealing detailed examples, interesting stories, and meaningful business results that left the audience clearly enthused. Free to view, here it is: www.pawcon.com/Target.

Toward the end, Pole described a project to predict customer pregnancy. Given that there's a tremendous sales opportunity when a family prepares for a newborn, you can see the marketing potential.

But this was something pointedly new, and I turned my head to scan the audience for any reactions. Nothing. Nada. Zilch. Normally, for marketing projects, PA predicts buying behavior. Here, the thing being predicted was not something marketers care about directly, but rather, something that could itself be a strong predictor of a wide range of shopping needs. After all, the marketer's job is to discover demand and pounce on it. You can think of this predictive goal as a "surrogate" (sorry) for the pertinent shopping activities a retail marketer is paid to care about.

PA APPLICATION: PREGNANCY PREDICTION

1. **What's predicted:** Which female customers will have a baby in coming months.
2. **What's done about it:** Market relevant offers for soon-to-be parents of newborns.

From what data did Target learn to predict pregnancy, given that predictive modeling requires a number of known cases from which to learn? Remember, the predictive modeling process is a form of automated data crunching that learns from *training examples*, which must include both positive and negative examples. An organization needs to have positively identified in the past some cases of what it would like to predict in the future. To predict something like "will buy a stereo," you can bet a retailer has plenty of positive cases. But how can you locate Target customers known to be pregnant?

You may be surprised how simple it is to answer this puzzle. Can you guess? Let's assume no medical information or pharmaceutical data is employed for this project. Why does a customer inform Target she is pregnant? The answer: the Target baby registry. Registrants not only

disclose they're pregnant, but they also reveal their due date. In addition, Target has indicated there are other marketing programs through which more moms-to-be identify themselves, thus also serving as positive learning examples.

Target pulled together training data by merging the baby registry data with other retail customer data and generated a "fairly accurate" predictive model. The store can now apply the model to customers who have *not* registered as pregnant. This identifies many more pregnant customers, since we can assume most such customers in fact do not register.

The model predictively evaluates a customer based on what she has purchased, which can include baby-related products, but may include combinations of other products not necessarily directly related to babies. Deriving the model is an automated act of trend spotting that explores a broad range of factors. I doubt Target's system confirmed that buying pickles and ice cream turns out to be a good indicator of pregnancy, but any and all product categories were analyzed and considered. The model identified 30 percent more customers for Target to contact with pregnancy-oriented marketing material—a significant marketing success story.

A PREGNANT PAUSE

Strutting charismatically across the conference stage, Pole boldly lauded this unorthodox endeavor, which he led at Target. The business value was clear, the story entertaining. It's likely he was delivering what had gone over well for internal Target presentations, but now to an open forum. It made for great material and engaged the audience.

I wondered for a moment if there had been any concerns but assumed, as one engrossed in the core technology itself may tend to do, that this project had been vetted, that concerns had been allayed and put to rest by folks at Target. Emerging from inside the PA practitioner's dark data cave, squinting at the world outside, it can be hard to imagine how unsuspecting folks walking down the street might respond to such a project. In fact, Pole reassured the audience that Target carefully adheres to all privacy and data-use laws. "Target wants to make sure that we don't end up in the newspaper or on TV because

we went out and we used something that we're not supposed to be using.” Little did we know where this was headed.

MY 15 MINUTES

Because the ensuing media storm around Target’s pregnancy prediction pulled me into its wake, I witnessed from a front-row seat how, if one reporter sets off just the right spark, the pundits will obediently burn and the news cycle will fan the flames.

Who spilled the beans in the first place? A few months after Pole’s presentation, *New York Times* reporter Charles Duhigg interviewed me. Exploring, he asked for interesting new ways PA was being used. I rattled off a few and included pregnancy prediction, pointing him to the online video of Pole’s talk, which had thus far received no media attention, and connecting him to Pole via e-mail. I must admit that by now the privacy question had left my mind almost entirely.

One year later, in February 2012, Duhigg published a front-page *New York Times Magazine* article, sparking a viral outbreak that turned the Target pregnancy prediction story into a debacle. The article, “How Companies Learn Your Secrets,” conveys a tone that implies wrongdoing is a foregone conclusion. It punctuates this by alleging an anonymous story of a man discovering his teenage daughter is pregnant only by seeing Target’s marketing offers to her, with the unsubstantiated but tacit implication that this resulted specifically from Target’s PA project. The *Times* even produced a short video to go with the article, which features dramatic, slow-motion, color-muted images of Target shoppers checking out, while creepy, suspenseful music plays and Duhigg himself states, “If they know when [your life is changing], then they can . . . manipulate you . . . so that your habits put dollars in their pockets.” He refers to the practice of data-driven marketing as “spying on” customers.

This well-engineered splash triggered rote repetition by press, radio, and television, all of whom blindly took as gospel what had only been implied—that the teen’s story stemmed from Target’s pregnancy prediction—and ran with it. Not incidentally, the article was excerpted from and helped launch

Duhigg's book, *The Power of Habit: Why We Do What We Do in Life and Business* (Random House, 2012), which hit the *New York Times* bestseller list.

The tornado sucked me in because the article quoted me in addition to Pole who, along with Target as a whole, had now unsurprisingly clammed up. As an independent consultant, I enjoyed unfettered freedom to make public appearances. I had no prudent employer that might hold me back.

THRUST INTO THE LIMELIGHT

This techie transmogrified into a pundit, literally overnight, as I raced to New York City on a red-eye to appear on Fox News. But placing my talking head on millions of TVs does not magically prepare me for such a role. Thriving in an abstract pool of data, the PA professional occasionally surfaces for air, usually only by accident. For the most part, this work is an exercise in math and algorithms to discover patterns that promise to hold true tomorrow—a strange, magical game to almost defy whatever laws of physics prohibit time travel. Inside this petri dish, you're insulated, knowing nothing of the visceral angst of broken hearts or broken privacy. In asking me to shed my lab coat for a suit and tie, the powers that be declared that our previously esoteric activities buried beneath these murky depths of data are truly important after all.

The morning news program *Fox & Friends* positioned me behind a desk, and I struggled to sit still in what was clearly the hot seat. Celebrity host Gretchen Carlson looked over and raised her voice to greet me from across the studio just before we started: "Hi, Eric!" I greeted her back as if it were just another day in the studio: "Hi, Gretchen!"

Then we were live to an estimated two million viewers. Falling in line behind the *Times*, Carlson read Target the riot act for exposing a girl's pregnancy, neglecting to mention the story was only an unsubstantiated allegation and implying this kind of collateral damage is innate to PA's application. A third talking head, a professor of medical ethics, reinforced the theme that all applications of PA ought best be shut down, at least pending further investigation. The millions of TVs tuned to Fox at that moment

displayed a Target store, overlaid with the question, “Are stores spying on you?” Later the screen proclaimed, “Target has got you in its aim.”

It quickly became clear I was to serve as a foil as the news show demonized my profession. For the moment, I was the face of PA, and I had to fight back. If there is a certain carelessness in how organizations wield the increasing power to predict, so too is there carelessness in misleading media coverage. I took a deep breath and asserted that the *New York Times* article was misleading because it implied Target has a “supernatural” ability to accurately predict who is pregnant, and because it established an unsubstantiated connection to the pregnant teen’s alleged story. Target’s predictions are not medical diagnosis and are not based on medical information. Finally, I managed to squeeze into my allotted seconds the main point: It is really important that PA not be universally stigmatized. You can watch the televised clip at www.pawcon.com/target-on-fox.

In another interview, I was confronted with a quote from privacy advocate Katherine Albrecht, who said, “The whole goal [of retailers] is to figure out everything you can learn about your customer. We’re creating a retail zoo, where customers are the exhibits.” My reply? Unlike the social sciences, PA’s objective is to improve operational efficiency rather than figure people out for its own sake—and, either way, just because you’re observing a person does not mean that person is being treated like an animal.

The media coverage was broad and, within a few weeks, it seemed like everyone I spoke with both inside and outside my work life had at least caught wind of the Target pregnancy story. Even comedian Stephen Colbert covered it, suggesting Target’s next move will be to predict from your spouse’s shopping habits that she is having an affair, and therefore send you a coupon for a hot plate that will go perfectly with your new studio apartment (more than just a joke, divorce prediction is included in this book’s Central Table 1).

As the dust settles, we’re left with a significant challenge: How can true privacy concerns be clearly defined, even as media overblows and confuses?

YOU CAN'T IMPRISON SOMETHING THAT CAN TELEPORT

Information about transactions, at some point in time, will become more important than the transactions themselves.

—Walter Wriston, former chairman and CEO of Citicorp

Information wants to be free.

—Stewart Brand to Steve Wozniak at the first Hackers Conference, 1984

Data matters. It's the very essence of what we care about.

Personal data is not equivalent to a real person—it's much better. It takes no space, costs almost nothing to maintain, lasts forever, and is far easier to replicate and transport. Data is worth more than its weight in gold—certainly so, since data weighs nothing; it has no mass.

Data about a person is not as valuable as the person, but since the data is so much cheaper to manage, it's a far better investment. Alexis Madrigal, senior editor at *The Atlantic*, points out that a user's data can be purchased for about half a cent, but the average user's value to the Internet advertising ecosystem is estimated at \$1,200 per year.

Data's value—its power, its meaning—is the very thing that also makes it sensitive. The more data, the more power. The more powerful the data, the more sensitive. So the tension we're feeling is unavoidable. If nobody cared about some piece of data, nobody would try to protect it, and nobody would want to access it or even bother to retain it in the first place. John Elder reflected, "The fact that it's perceived as dangerous speaks to its power; if it were weak, it wouldn't be a threat."

Ever since the advent of paper and pen, this has been the story. A doctor scribbled a note, and the battle to establish and enforce access policies began.

But now, digital data travels so far, so fast, between people, organizations, and nations. Combine this ability of data to go anywhere at almost no cost with the intrinsic value of the stuff that's traveling, and you have the makings of a very fickle beast, a swarm of gremlins impressively tough to control. It's like trying to incarcerate the *X-Men*'s superhero Nightcrawler, who has the ability to teleport. It's not confined to our normal three dimensions of movement, so you just can't lock it up.

Data is such a unique thing to ship, we have a special word for its telekinetic mode of transport. We call it telecommunication.

Data wants to spread like wildfire. As privacy advocate David Sobel put it, “Once information exists, it’s virtually impossible to limit its use. You have all this great data lying around, and sooner or later, somebody will say, ‘What else can I do with it?’”

This new, powerful currency proves tough to police. A shady deal to share consumer records is completed with the press of a button—no covert physical shipment of goods required.

LAW AND ORDER: POLICIES AND POLICING OF DATA

[Privacy is] the most comprehensive of all rights and the one most cherished by a free people.

—Supreme Court Justice Louis Brandeis, 1928

And yet, we must try our darnedest to tame this wild creature. An open free-for-all is surely not an option. The world will continue struggling to impose order on the distribution of medical facts, financial secrets, and embarrassing photos. Consternation runs deep, with an estimated one in four Facebook users posting false data due to privacy concerns.

Each organization must decide data’s who, what, where, when, how long, and why:

Retain—What is stored and for how long.

Access—Which employees, types of personnel, or group members may retrieve and look at which data elements.

Share—What data may be disseminated to which parties within the organization, and to what external organizations.

Merge—What data may be joined together, aggregated, or connected.

React—How may each data element be acted upon, determining an organization’s response, outreach, or other behavior.

To make it even more complicated, add to each of these items “. . . under which circumstances and for what type of intention or purpose.”

Pressing conundrums ensue. Which data policies can and should be established via legislation, and which by industry best practices and rules of etiquette? For which data practices may the organization default the consumer in, in which case she must take explicit action to opt out if so desired? How are policies enforced: What security standards—encryption, password integrity, firewalls, and the like—promise to earn Fort Knox’s reputation in the electronic realm?

We have our work cut out for us.

THE BATTLE OVER DATA

The Internet of free platforms, free services, and free content is wholly subsidized by targeted advertising, the efficacy (and thus profitability) of which relies on collecting and mining user data.

—Alexander Furnas, writer for *The Atlantic*

The stakes increase and the opponents’ resolve hardens like cooling lava.

In one corner we have privacy advocates, often loath to trust organizations, racing to squeeze shut data’s ebb and flow: Contain it, delete it, or prevent it from being recorded in the first place.

In the other corner we have the data hustlers, salivating: the hoarders and opportunists. This colorful group ranges from entrepreneurs to managers, techies, and board members.

Data prospectors see value, and value is exciting—from more than just a selfish or economic standpoint. We love building the brave new world: increasing productivity and efficiency, decreasing junk mail and its environmental impact, improving healthcare, and suggesting movies and music that will better entertain you. And we love taking on the scientific challenges that get us there.

And yet, even the data hustlers themselves can feel the pain. I was at Walgreens a few years ago, and upon checkout an attractive, colorful coupon spit out of the machine. The product it hawked, pictured for all my fellow

shoppers to see, had the potential to mortify. It was a coupon for Beano, a medication for flatulence. I'd developed mild lactose intolerance but, before figuring that out, had been trying anything to address my symptom. Acting blindly on data, Walgreens' recommendation system seemed to suggest that others not stand so close.

Other clinical data holds a more serious and sensitive status than digestive woes. Once, when teaching a summer program for talented teenagers, I received data I felt would have been better kept away from me. The administrator took me aside to inform me that one of my students had a diagnosis of bipolar disorder. I wasn't trained in psychology. I didn't want to prejudge the student, but there is no "delete" button in the brain's memory banks. In the end, the student was one of my best, and his supposed disorder never seemed to manifest in any perceivable way.

Now we are witnessing the increasing use of location data from cell phones and cars. Some people are getting into serious trouble with their bosses, spouses, and other law enforcement agencies. Tom Mitchell, a professor at Carnegie Mellon University and a world leader in the research and development of machine learning capabilities, wrote in a *Science* article: "The potential benefits of mining such data [from cell phones that track location via GPS] are various; examples include reducing traffic congestion and pollution, limiting the spread of disease, and better using public resources such as parks, buses, and ambulance services. But risks to privacy from aggregating these data are on a scale that humans have never before faced."

These camps will battle over data for decades to come. Data hustlers must hone their radar for land mines, improving their sensitivity to sensitivity. Privacy advocates must see that data-driven technology is a tool that can serve both good and evil—like a knife. Outlawing it completely is not an option. There's no objectively correct resolution; this is a subjective, dynamic arena in which new aspects of our culture are being defined. Dialogue is critical, and a "check here to agree to our lengthy privacy policy that you are too busy to read" does not count as dialogue. Organizations and consumers are not speaking the same language. Striking a balance, together, is society's big new challenge. We have a long way to go.

DATA MINING DOES NOT DRILL DOWN

Exonerate the data scientists and their darling invention. PA in and of itself does not invade privacy—its core process is the *opposite* of privacy invasion. Although it's sometimes called *data mining*, PA doesn't "drill down" to peer at any individual's data. Instead, PA actually "rolls up," learning patterns that hold true in general by way of rote number crunching across the masses of customer records. Data mining often appears to be a culprit when people misunderstand and completely reverse its meaning.

But PA palpably intensifies the battle over data. Why? It ignites fire under data hustlers across the world with a greater and more urgent hunger for more data. Having more data elements per customer means better odds in number crunching's exploration for what will prove most predictive. And the more rows of customer records, the better the predictive model resulting from PA's learning process.

Don't blame the sun when a thirsty criminal steals lemonade. If data rules are fair and right, PA activities that abide by them cannot contribute to abuse or privacy invasion. In this case, PA will be deemed copacetic and be greeted with open arms, and all will be well in our happy futuristic world of prediction. Right?

Fade to black and flash forward to a dystopia. You work in a chic cubicle, sucking chicken-flavored sustenance from a tube. You're furiously maneuvering with a joystick, remotely operating a vehicle on a meteor digging for precious metals. Your boss stops by and gives you a look. "We need to talk about your loyalty to this company."

The organization you work for has deduced that you might be planning to quit. It predicts your plans and intentions, possibly before you have even conceived them.

HP LEARNS ABOUT ITSELF

In 2011, two crackerjack scientists at HP broke ground by mathematically scrutinizing the loyalty of each and every one of their more than 300,000

colleagues. Gitali Halder and Anindya Dey developed predictive models to identify all “Flight Risk” employees, those with a higher expected chance of quitting their jobs.

Retaining employees is core to protecting any organization. After all, an organization’s defining characteristic is that it’s a collection of members. One of five ideological tenets set forth by a founder of HP is: “We achieve our common objectives through teamwork.” Employees contribute complementary skills and take on complementary roles. They learn how to work together. It’s bad news when a good one goes. The management of employee turnover is a significant challenge for all companies. For example, another multinational corporation looked to decrease turnover among customer service agents at a call center in Barcelona. Folks would come just to spend the summer in that beautiful city and then suddenly give notice and split. It would help to identify such job applicants in advance.

In this endeavor, the organization is aiming PA inwardly to predict its own staff’s behavior, in contrast to the more common activity of predicting its patrons’ behavior. As with predicting which customers are most likely to leave in order to target retention efforts, HP predicts which of its staff are likely to leave in order to do the same. In both cases, it’s like identifying leaks in a boat’s hull in order to patch them up and keep the ship afloat.¹

PA APPLICATION: EMPLOYEE RETENTION

1. **What’s predicted:** Which employees will quit.
2. **What’s done about it:** Managers take the predictions for those they supervise into consideration, at their discretion. This is an example of decision *support* rather than feeding predictions into an automatic decision process.

¹ This and related workforce applications of PA are emerging rapidly enough that the field warranted the 2015 launch of its own annual conference: Predictive Analytics World for Workforce.



Reproduced with permission.

INSIGHT OR INTRUSION?

HP is the iconic success story. It literally started in the proverbial garage and now leads the worldwide manufacturing of personal computers. The company came in as the twenty-seventh largest employer of 2011, amassing \$127 billion in revenue, which makes it one of the highest-earning technology companies in the world.

HP is an empire of sorts, but by no means a locked-up citadel. Some working groups report turnover rates as high as 20 percent. On a ship this big, there are bound to be some leaks, especially given the apparent short attention span of today's technology worker.

HP is a progressive analytics leader. Its analytics department houses 1,700 workers in Bangalore alone. They boast cutting-edge analytical capabilities across sales, marketing, supply chain, finance, and human resources (HR)

domains. Their PA projects include customer loss prediction, sales lead scoring, and supplier fraud detection.

Gitali Halder leads HP's analytics team in Bangalore focused on HR applications. With a master's in economics from the Delhi School of Economics and several years of hands-on experience, Halder is your true PA powerhouse. Confident, well spoken, and gregarious, she compels and impresses. Having teamed with HP consultant Anindya Dey, also in Bangalore, the two shine as a well-presented dynamic duo, as evidenced by their polished presentation on this project at the Predictive Analytics World conference in November 2011 in London.

Halder and Dey compiled a massive set of training data to serve as learning material for PA. They pulled together two years of employee data such as salaries, raises, job ratings, and job rotations. Then they tacked on, for each of these employee records, whether the person had quit. Thus, HP was positioned to learn from past experience to predict a priceless gem: which combinations of factors define the type(s) of employees most likely to quit their jobs.

If this project helps HP slow its employee turnover rate, Halder and Dey may stand above the crowd as two of its most valuable employees—or become two of the most resented, at least by select colleagues. Some devoted HP workers are bound to be uncomfortable that their Flight Risk score exists. What if your score is wrong, unfairly labeling you as disloyal and blemishing your reputation?

A whole new breed of powerful HR data emerges: speculative data. Beyond personal, financial, or otherwise private data about a person, this is an estimation of the future and thus speaks to the heart, mind, and intentions of the employee. Insight or intrusion?

It depends on what HP does with it.

FLIGHT RISK: I QUIT!

On the other side of the world, Alex Beaux helps Halder and Dey bring the fruits of their labor to bear upon a select niche of HP employees. It's 10.5 hours earlier in Houston, where Beaux sits as a manager for HP's

internal Global Business Services (GBS). With thousands of staff members, GBS provides all kinds of services across HP to departments that have something they'd like to outsource (even though "outsourcing" to GBS technically still keeps the work within HP).

Beaux, Halder, and Dey set their sights on GBS's Sales Compensation team, since its roughly 300 employees—spread across a few countries—have been exhibiting a high attrition rate of up to 20 percent. A nicely contained petri dish for a pilot field test of Flight Risk prediction, this team provides support for calculating and managing the compensation of salespeople internationally.

The message is clear: Global enterprises are complex! This is not a team of salespeople. It isn't even a regular HR team that supports salespeople. Rather, it is a global team, mostly in Mexico, China, and Poland, that helps various HR teams that support salespeople. And so this project is multilevel: It's the analytical HR management of a team that helps HR (that supports salespeople).

Just read that paragraph five more times and you'll be fine. I once worked on an HP project that predicted the potential demand of its corporate clients—how many computers will the company need to buy, and how much of that need is currently covered by HP's competitors? Working on that project for several months, I was on conference calls with folks from so many working groups named with so many acronyms and across so many time zones that it required a glossary just to keep up.

This organizational complexity means there's great value in retaining sales compensation staff. A lot of overhead must be expended to get each new hire ramped up. Sales compensation team members boast a very specific skill set, since they manage an intricate, large-scale operation. They work with systems that determine the nitty-gritty as to how salespeople are compensated. A global enterprise does not follow an orderly grid designed by a city planner—it takes on a patchwork quality since so much organizational growth comes of buying smaller companies, thus absorbing new sales teams with their own compensation rules. The GBS Sales Compensation team handles an estimated 50 percent of the work to manage sales compensation across the entire global organization.

INSIGHTS: THE FACTORS BEHIND QUITTING

The data showed that Flight Risk depends on some of the things you would expect. For example, employees with higher salaries, more raises, and increased performance ratings quit less. These factors pan out as drivers that decrease Flight Risk. Having more job rotations also keeps employees on board; Beaux conjectures that for the rote, transactional nature of this work, daily activities are kept more interesting with periodic change.

One surprise is that getting a promotion is not always a good thing. Across all of HP, promotions do decrease Flight Risk, but within this Sales Compensation team, where a number of promotions had been associated with relatively low raises, the effect was reversed: Those employees who had been promoted more times were more likely to quit, unless a more significant pay hike had gone along with the promotion.

The analysis is only as good as the data (garbage in, garbage out). In a similar but unrelated project for another company, I predictively modeled how long new prospective hires for a Fortune 1000 business-to-business (B2B) provider of credit information would stay on if hired for call center staffing. Candidates with previous outbound sales experience proved 69 percent more likely to remain on the job at least nine months. Other factors included the number of jobs in the past decade, the referring source of the applicant, and the highest degree attained. This project dodged a land mine, as preliminary results falsely showed new hires without a high school degree were 2.6 times as likely to stay on the job longer. We were only days away from presenting this result to the client—and recommending that the company hire more high school dropouts—when we discovered an unusual combination of errors in the data the client had delivered.² Error-prone data—noise—usually just means fewer conclusions will be drawn, rather than strong false ones, but this case was an exceptional perfect storm—a close call!

² Encodings for the highest degree attained were inconsistent and the inconsistency corresponded with non-random portions of the dataset. Discovering this was largely serendipitous; with less luck it could easily have continued to go unnoticed.

As for any domain of PA, the predictive model zips up these various factors into a single score—in this case, a Flight Risk score—for each individual. Even if many of these phenomena seem obvious or intuitive, the model is where the subtle stuff comes in: how these elements weigh in relative to one another, how they combine or interact, and which other intuitive hunches that don’t pan out should be eliminated. A machine learning process automates these discoveries by crunching the historical data, literally learning from it.

Halder and Dey’s Flight Risk model identified \$300 million in estimated potential savings with respect to staff replacement and productivity loss across all HP employees throughout all global regions. The 40 percent of HP employees with highest Flight Risk scores included 75 percent of the quitters (a predictive *lift* of 1.9).

I asked the two, who themselves are HP employees, what their own Flight Risk scores were. Had they predicted themselves likely to quit? Halder and Dey are quick to point out that they like their jobs at HP very much, but admit they are in fact members of a high-risk group. This sounds likely, since analytics skills are in high demand.

DELIVERING DYNAMITE

When chemists synthesize a new, unstable element, they must *handle with care*.

HP’s Flight Risk scores deploy with extreme caution, under lock and key. Beaux, Halder, and Dey devised a report delivery system whereby only a select few high-level managers who have been trained in interpreting Flight Risk scores and understanding their limitations, ramifications, and confidentiality may view individual employee scores—and only scores for employees under them. In fact, if unauthorized parties got their hands on the report itself, they would find there are no names or identifying elements for the employees listed there—only cryptic identifiers, which the authorized managers have the key to unscramble and match to real names. All security systems have vulnerabilities, but this one is fairly bulletproof.

For the GBS Sales Compensation team of 300 employees, only three managers see these reports. A tool displays the Flight Risk scores in

a user-friendly, nontechnical view that delivers supporting contextual information about each score in order to help explain why it is high or low. The consumers of this analytical product are trained in advance to understand the Flight Risk scores in terms of their accompanying explanations—the factors about the employee that contributed to the score—so that these numbers aren’t deferred to as a forceful authority or overly trusted in lieu of other considerations.

A score produced by any predictive model must be taken with a very particular grain of salt. Scores speak to trends and probabilities across a large group; one individual probability by its nature oversimplifies the real-world thing it describes. If I were to miss a single credit card payment, the probability that I’d miss another in the same year may quadruple, based on that factor alone. But if you also take into account that my roof caved in that month (this is a fictional example), your view will change. In general, the complete story for an individual is in fact more than we can ever know. You can see a parallel to another scrutinized practice: diagnosing someone with a psychological disorder and thus labeling them and influencing how they’re to be treated.

Over time, the Flight Risk reports sway management decisions in a productive direction. They serve as early warning signals that guide management in planning around loss of staff when it can’t be avoided, and working to keep key employees where possible. The system informs what factors drive employee attrition, empowering managers to develop more robust strategies to retain their staffs in order to reduce costs and maintain business continuity.

THE VALUE GAINED FROM FLIGHT RISK

And the results are in. GBS’s Sales Compensation staff attrition rates that were above 20 percent in some regions have decreased to 15 percent and continue to trend downward. This success is credited in large part to the impact of Flight Risk reports and their well-crafted delivery.

The project gained significant visibility within HP. Even HP’s worldwide vice president of sales compensation heartily applauded the project. Flight

Risk reports continue to make an impact today, and their underlying predictive models are updated quarterly over more recent data in order to remain current.

These pioneers may not realize just how big a shift this practice is from a cultural standpoint. The computer is doing more than obeying the usual mechanical orders to retain facts and figures. It's producing new information that's so powerful, it must be handled with a new kind of care. We're in a new world in which systems not only divine new, potent information but must carefully manage it as well.

Managed well and delivered prudently, Flight Risk scores can perhaps benefit an organization without ruffling too many feathers. Given your established relationship with your boss, perhaps you'd be comfortable if he or she received a Flight Risk score for you, assuming it was considered within the right context. And perhaps it's reasonable and acceptable for an employer to crunch numbers on employee patterns and trends, even without the employees necessarily knowing about it. There's no universally approved ethical framework yet established—the jury is still out on this new case.

But, moving from employment record to criminal record, what if law enforcement officers appeared at your door to investigate you, Future Crime Risk report in hand?

PREDICTING CRIME TO STOP IT BEFORE IT HAPPENS

What if you could shift the intelligence paradigm from “sense, guess, and respond” to “predict, plan, and act”?

—Sgt. Christopher Fulcher, Chief Technology Officer of the Vineland, New Jersey, Police Department

Cops have their work cut out for them. Crime rates may ebb and flow, but law enforcement by its nature will always face the impossible challenge of optimizing the deployment of limited resources such as patrolling officers and perusing auditors.

Police deploy PA to predict the location of crime and to direct cops to patrol those areas accordingly. One system, backtested on two years of

data from Santa Cruz, California, correctly predicted the locations of 25 percent of burglaries. This system directs patrols today, delivering 10 hot spots each day within this small city to send police vehicles to. The initiative was honored by *Time* magazine as one of the 50 best inventions of 2011.

PA APPLICATION: CRIME PREDICTION (AKA PREDICTIVE POLICING)

- 1. What's predicted:** The location of a future crime.
- 2. What's done about it:** Police patrol the area.

Another crime prediction system, revealed at a 2011 conference by Chief Information Officer Stephen Hollifield of the Richmond, Virginia, police department, serves up a crime-fighting display that marks up maps by the risk of imminent crime and lists precincts, neighborhoods, and crime types by risk level. Since this system's deployment, Richmond crime rates have decreased. Similar systems are in development in Chicago; Los Angeles; Vineland, New Jersey; and Memphis, where prediction is credited with reducing crime by 31 percent. In 2009, the U.S. National Institute of Justice awarded planning grants to seven police departments to create crime prediction capabilities.

Lightning strikes twice. The predictive models leverage discoveries such as the trend that crimes are more—not less—likely to soon reoccur in nearby locations, as detected in Santa Cruz. In Richmond, the predictive model flags for future crime based on clues such as today's city events, whether it's a payday or a holiday, the day of the week, and the weather.

What's not to like? Law enforcement gains a new tool, and crime is defrayed. Any controversy over these deployments appears relatively tame. Even the American Civil Liberties Union gave this one a nod of the head. No harm, no foul.

In fact, there's one type of crime that elicits loud complaints when predictive models *fail* to detect it: fraud. To learn more, see the sidebar on fraud detection. After the sidebar, we continue on to explore how crime-predicting computers inform how much time convicts spend in prison.

SPECIAL SIDEBAR ON FRAUD DETECTION

Criminals can be such nice guys. I became friends with one in 1995. I was pursuing my doctorate in New York City and he was the new boyfriend of my girlfriend's sister. Extremely charismatic and supposedly a former professional athlete, the crook wooed, wowed, and otherwise ingratiated himself into our hearts and home. I'll never forget the really huge, fun dinner he treated us to at the famous Italian restaurant Carmine's. I didn't think twice about letting him use my apartment when I went on a vacation.

A year or two later I discovered he had acquired my Social Security number, stolen my identity, and soiled my sparkly clean credit rating. He had started a small water bottling business in the Los Angeles area, posing as me. Despite being a decade older than I, on the wrong coast, and not even attempting to emulate my signature, he had attained numerous credit accounts, including credit cards and leases on water bottling equipment. After building considerable debt, he abandoned the business and defaulted on the payments. It took a couple of years of tedious paperwork to clear my name and clean up my credit rating.

Where's a good predictive model when you need one? Why couldn't these credit applications have been flagged or quarantined, checking with me by way of the contact information established in my credit files? After all, once all the evidence was gathered and submitted, most auditors immediately perceived the case as obvious fraud.

While some deployments of PA give rise to concern, the absence thereof does as well. Enter *fraud detection*.

A WOLF IN SHEEP'S CLOTHING

Fraud, defined as "intentional deception made for personal gain," is the very act of a wolf dressing up in sheep's clothing. It's when someone pretends to be someone else or to be authorized to do something the fraudster is not authorized to do. A student copies another's homework, a sumo wrestler throws a match, an online

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

gambler cheats with illegal information as part of an inside job, inauthentic Twitter accounts spread misinformation about a political candidate, or a death is faked in order to make a claim against a life insurance policy. All such crimes have been detected analytically.

It's a good time to be a fraudster since they enjoy a massive, expanding stomping ground: the Internet, a transaction infrastructure for global commerce. But by connecting to everybody, we've connected to folks with malicious intent. The easier it is to conduct consumer and business transactions, the easier it is to fake them as well. And with the buyer, seller, goods, and payment spread across four different physical locations, there is an abundance of vulnerabilities that may be exploited.

As transactions become increasingly numerous and automated, criminal opportunities abound. Fraudulent transactions such as credit card purchases, tax returns, insurance claims, warranty claims, consumer banking checks, and even intentionally excessive clicks on paid ads incur great cost. The National Insurance Crime Bureau says that insurance criminals steal over \$30 billion annually, making such fraud the second most costly white-collar crime in the United States—behind tax evasion—resulting in \$200 to \$300 of additional insurance premiums per U.S. household; we are paying these criminals out of our pockets.

“It is estimated that the nation’s banks experience over \$10 billion per year in attempted check fraud,” says former Citizens Bank Vice President Jay Zhou, now a data mining consultant. Credit card fraud losses approach \$5 billion annually in the United States, and Medicaid fraud is estimated to be the same amount for New York State alone. According to the most recent report published by the Federal Trade Commission, 2011 brought over 1.8 million complaints of fraud, identity theft, or other intentional deceit in business, about 40 percent more than in 2010.

(continued)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

Aggregate fraud loss in the United States sees estimates from \$100 billion to \$1 trillion.

Prediction helps. Predictively scoring and ranking transactions dramatically boosts fraud detection. A team of enforcement workers can inspect only a fixed number of suspected transactions each week. For example, Progressive Insurance employs about 200 “special investigations professionals” on this task. Delivering a more precisely identified pool of candidate transactions—fewer false alarms (false positives)—renders their time more effectively spent; more fraud is detected, and more losses are prevented or recouped.

PA APPLICATION: FRAUD DETECTION³

- 1. What's predicted:** Which transactions or applications for credit, benefits, reimbursements, refunds, and so on are fraudulent.
- 2. What's done about it:** Human auditors screen the transactions and applications that are predicted most likely to be fraudulent.

Math is fighting back. Most large—and many medium-sized—financial institutions employ fraud detection. For example, Citizens Bank developed a fraud prediction model that scores each check, predicting well enough to decrease fraud loss by 20 percent. One automobile insurance carrier showed that PA delivers 6.5 times the fraud detection capacity of that attained with no means to rank or score insurance claims. Online transaction giant PayPal suffered an almost 20 percent fraud rate soon after it was launched, a primary threat to its

³ Rather than performing prediction in the conventional sense of the word, this application of PA performs detection. As with predicting the future, such an application imperfectly infers an unknown.

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

success. Fraud detection methods brought the rate down to a reported less than 1 percent. The people behind each of these stories have spoken at the Predictive Analytics World conference, as have those telling similar stories from 1-800-FLOWERS, Activision, the Belgian government, the U.S. Postal Service, the Internal Revenue Service (IRS), administrators of Medicare and Medicaid, and a leading high-tech company that catches warranty claims from repair shops that didn't actually do the service at all.

GOVERNMENT, PROTECT THYSELF

The government is working hard on fraud management—but unlike its efforts enforcing against crimes like theft and assault, most of this effort isn't focused on protecting you, or even any business. When it comes to fraud, the U.S. government is fighting to protect its own funds. In fact, fraud detection is the most evident government application of PA, providing a means to decrease loss in the face of tightening budgets.

Elder Research (John Elder's company) headed a fraud modeling project for the IRS that increased the capacity to detect fraudulent returns by a factor of 25 for a certain targeted segment. A similar effort has been reported by the Mexican Tax Administration, which has its own Risk Models Office.

The U.S. Defense Finance and Accounting Service, responsible for disbursing nearly all Department of Defense funds, executes millions of payments on vendor invoices. Dean Abbott, a top PA consultant who has also consulted for the IRS, led the development of a predictive model capable of detecting 97 percent of known cases of fraudulent invoices. The model scores invoices based on factors such as the time since the last invoice, the existence of other payees at the same postal address, whether the address is a P.O. box, and whether the vendor submitted invoices out of order.

(continued)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

Beyond these possible signs of fraud, other innovative clues turbocharge the predictive model, helping determine which cases are flagged. 1-800-FLOWERS improved its ability to detect fraud by considering the social connections between prospective perpetrators. In fact, one fraud scheme can't be detected without this kind of social data. (Oxymoron, anyone?) A group of criminals open financial accounts that improve their respective credit ratings by transferring funds among themselves. Since the money transfers take place only between these accounts, the fraudsters need not spend any real money in conducting these transactions; they play their own little zero-sum game. Once each account has built up its own supposedly legitimate record, they strike, taking out loans, grabbing the money, and running. These schemes can be detected only by way of social data to reveal that the network of transactors is a closed group.

Naturally, criminals respond by growing more creative.

THE FRAUD DETECTION ARMS RACE

The fraudsters were also good, and nimble, too, devising new scams as soon as old ones were compromised.

—Steven Levitt and Stephen Dubner, *SuperFreakonomics*

Just as competing businesses in the free market push one another to better themselves, fraud detection capabilities drive criminals toward self-improvement by the design of smarter techniques. The act of fraud strives to be stealthy, sneaking under the predictive model's radar. As with the possibility of superbacteria emerging from the overuse of antibiotics, we are inadvertently creating a stronger enemy.

But there's good news. The white hats sustain a great advantage. In addition to exerting human creativity like our opponents, we have the data with which to hone fraud detection models. A broad set of data containing historical examples of both fraudulent and legitimate

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

transactions intrinsically encodes the inherent difference between the two. PA is the very means by which to discover this difference from data. And so, beyond storing and indexing a table of “signatures” that betray the perpetration of known fraud schemes, the modeling process generates detection schemes that cast a wider net. It predicts forthcoming forms of fraud by generalizing from previously observed examples. This is the defining characteristic of a learning system.

THIS MEANS WAR

It’s a war like any other. In fact, cyberwarfare itself follows the same rules. PA bolsters information security by detecting hackers and viruses that exploit online weaknesses, such as system bugs or other vulnerabilities. After all, the Internet’s underlying networking technology, TCP/IP, is a platform originally designed only for interactions between mutually entrusted parties. As the broad, commercial system it evolved to be, the Internet is, underneath the hood, something of a slapped-together hack with regard to security. Like an unplanned city, it functions, but like a Social Security number awaiting discovery in an unlocked drawer, it holds intrinsic weaknesses.

PA APPLICATION: NETWORK INTRUSION DETECTION

- 1. What’s predicted:** Which low-level Internet communications originate from imposters.
- 2. What’s done about it:** Block such interactions.

PA boosts detection by taking a qualitatively new step in the escalating arms race between white and black hats. A predictive detection system’s field of vision encompasses a broad scope of potential attacks that cannot be known by perpetrators, simply because they don’t

(continued)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

have access to the same data used to develop the predictive model. Hackers can't know if their techniques will be detected. PA's deployment brings a qualitative change in the way we compete against malicious intent.

But beware! Another type of fraud attacks you and every one of us, many times a day. Are you protected?

LIPSTICK ON A PIG

An Internet service cannot be considered truly successful until it has attracted spammers.

—Rafe Colburn, Internet development thought leader

Alan Turing (1912–1954), the father of computer science, proposed a thought experiment to explore the definition of what would constitute an “intelligent” computer. This so-called *Turing test* allows people to communicate via written language with someone or something hidden behind a closed door in order to formulate an answer to the question: Is it human or machine? The thought experiment poses this tough question: If, across experiments that randomly switch between a real person and a computer, subjects can’t correctly tell human from machine more often than the 50 percent correctness one could get from guessing, would you then conclude that the computer, having thereby passed the test by proving it can trick people, is intelligent? I’ll give you a hint: There’s no right answer to this philosophical conundrum.

In practice, computers attempt to fool people for money every day via e-mail. It’s called spam. As with androids in science fiction movies like *Aliens* and *Blade Runner*, successful spam makes you believe. Spam’s cousin, *phishing*, persuades you to divulge financial secrets. *Spambots* take the form of humans in social networks and dating sites in

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

order to grab your attention. And spammy Web pages trick search engines into pointing you their way.

Spam filters, powered by PA, are attempting their own kind of Turing test every day at an e-mail in-box near you.

PA APPLICATION: SPAM FILTERING

1. **What's predicted:** Which e-mail is spam.
2. **What's done about it:** Divert suspected e-mails to your spam e-mail folder.

Unfortunately, in the spam domain, white hats don't exclusively own the arms race advantage. The perpetrators can also access data from which to learn, by testing out a spam filter and reverse engineering it with a model of their own that predicts which messages will make it through the filter. University of California, Berkeley researchers showed how to do this to render one spam filter useless.

ARTIFICIAL ARTIFICIAL INTELLIGENCE

In contrast to these precocious computers, we sometimes witness a complete role reversal: a person pretends to be a machine. The Mechanical Turk, a hoax in the eighteenth century, created the illusion of a machine playing chess. The Turk was a desk-sized box that revealed mechanical gears within and sported a chessboard on top. Seated behind the desk was a mannequin whose arm would reach across the board and move the pieces. A small human chess expert who did not suffer from claustrophobia (chess is a long game) hid inside the desk, viewing the board from underneath and manipulating the mannequin's arm. Napoleon Bonaparte and Benjamin Franklin had the pleasure of losing to this wonder of innovation—I mean, this crouching, uncomfortable imposter.

(continued)

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

In the modern-day equivalent, human workers perform low-level tasks for the Amazon Mechanical Turk, a crowdsourcing website by [Amazon.com](#) that coordinates hundreds of thousands of workers to do “things that human beings can [still] do much more effectively than computers, such as identifying objects in a photo . . . [or] transcribing audio recordings.” Its slogan is “Artificial Artificial Intelligence.” (This reminds me of the vegetarian restaurant with “mock mock duck” on the menu—I swear, it tastes exactly like mock duck.) As NASA put it in 1965 when defending the idea of sending humans into space, “Man is the lowest-cost, 150-pound, nonlinear, all-purpose computer system which can be mass-produced by unskilled labor.”

But for some tasks, we don’t have to pretend anymore. Everything changed in 1997 when IBM’s Deep Blue computer defeated then world chess champion Garry Kasparov. Predictive modeling was key. No matter how fast the computer, perfection at chess is impossible, since there are too many possible scenarios to explore. Various estimates agree there are more chess games than atoms in the universe, a result of the nature of exponential growth. So the computer can look ahead only a limited number of moves, after which it needs to stop enumerating scenarios and evaluate game states (boards with pieces in set positions), predicting whether each state will end up being more or less advantageous.

PA APPLICATION: PLAYING A BOARD GAME

1. **What’s predicted:** Which game board state will lead to a win.
2. **What’s done about it:** Make a game move that will lead to a state predicted to lead to a win.

Upon losing this match and effectively demoting humankind in its standoff against machines, Kasparov was so impressed with the strategies Deep Blue exhibited that he momentarily accused IBM of

SPECIAL SIDEBAR ON FRAUD DETECTION (CONTINUED)

cheating, as if IBM had secretly hidden another human grandmaster chess champion, squeezed in there somewhere between a circuit board and a disk drive like a really exorbitant modern-day Mechanical Turk. And so IBM had passed a “mini Turing test” (not really, but the company did inadvertently fool a pretty smart guy).

From this upset emerges a new form of chess fraud: humans who employ the assistance of chess-playing computers when competing in online chess tournaments. And yet another arms race begins, as tournament administrators look to detect such cheating players. This brings us full circle, back to computers that pose as people, as is the case with spam.

So computer “intelligence” has flipped the meaning of fraud on its head, reversing it. Rather than a chess-playing person pretending to be a machine (the Mechanical Turk), we have a machine masking as a person (cheating in human chess tournaments). It’s rather like *Star Trek*’s Commander Data, an emotionally stunted android afflicted with the Pinocchio Syndrome of wanting to be more human.

THE DATA OF CRIME AND THE CRIME OF DATA

PA has taken on an enormous crime wave. It is central to tackling fraud and promises to bolster street-level policing as well.

In these efforts, PA’s power optimizes the assignment of resources. Its predictions dictate how enforcers spend their time—which transactions auditors search for fraud and which street corners cops search for crime.

But how about giving PA the power to help decide who belongs in prison?

To help make these tough decisions, judges and parole boards consult predictive models. To build these models, Philadelphia’s Adult Probation and Parole Department enlisted a professor of statistics and criminology from the University of Pennsylvania. The parole department’s research director,

Ellen Kurtz, told *The Atlantic*, “Our vision was that every single person, when they walked through the door [of a parole hearing], would be scored by a computer” as to his or her risk of recidivism—committing crime again.

Oregon launched a crime prediction tool to be consulted by judges when sentencing convicted felons. The tool is on display for anyone to try out. If you know the convict’s state ID and the crime for which he or she is being sentenced, you can enter the information on the Oregon Criminal Justice Commission’s public website and see the predictive model’s output: the probability the offender will be convicted again for a felony within three years of being released.

PA APPLICATION: RECIDIVISM PREDICTION FOR LAW ENFORCEMENT

- 1. What’s predicted:** Whether a prosecuted criminal will offend again.
- 2. What’s done about it:** Judges and parole boards consult model predictions when making decisions about an individual’s incarceration.

The predictive model behind Oregon’s tool performs admirably. Machine learning generated the model by processing the records of 55,000 Oregon offenders across five years of data. The model then validated across 350,000 offender records across 30 years of history. Among the least risky tenth of criminals—those for whom the model outputs the lowest predictive scores—recidivism is just 20 percent. Yet among the top fifth receiving the highest scores, recidivism will probably occur; over half of these offenders will commit a felony again.

Law enforcement’s deployment of PA to predict for individual convicts is building steam. In these deployments, PA builds upon and expands beyond a longstanding tradition of crime statistics and standard actuarial models. Virginia’s and Missouri’s sentencing guidelines also prescribe the consideration of quantitative risk assessment, and Maryland has models that predict murder. The machine is a respected adviser that has the attention of judges and parole boards.

Humans could use some help with these decisions, so why not introduce an objective, data-driven voice into the process? After all, studies have shown that arbitrary extraneous factors greatly affect judicial decisions. A joint study

by Columbia University and Ben Gurion University (Israel) showed that hungry judges rule negatively. Judicial parole decisions immediately after a food break are about 65 percent favorable, but then drop gradually to almost zero percent before the next break. If your parole board judges are hungry, you're much more likely to stay in prison.

With this reasoning accepted, the convict's future now rests in nonhuman hands. Given new power, the computer can commit more than just prediction errors—it can commit injustice, previously a form of misjudgment that only people were in a position to make. It's a whole new playing field for the machine, with much higher stakes. Miscalculations in this arena are more costly than for other applications of PA. After all, the price is not as high when an e-mail message is wrongly incarcerated in the spam folder or a fraud auditor's time is wasted on a transaction that turns out to be legitimate.

MACHINE RISK WITHOUT MEASURE

In the movie *Minority Report*, Tom Cruise's science fiction cop tackles and handcuffs individuals who have committed no crime (yet), proclaiming stuff like: "By mandate of the District of Columbia Precrime Division, I'm placing you under arrest for the future murder of Sarah Marks and Donald Dubin." Rather than the punishment fitting the crime, the punishment fits the precrime.

Cruise's bravado does not go unchecked. Colin Farrell's Department of Justice agent confronts Cruise, and the two brutes stand off, mano a mano. "You ever get any false positives?" accuses Farrell.

A *false positive*, aka *false alarm*, is when a model incorrectly predicts yes when the correct answer is no. It says you're guilty, convicting you of a crime you didn't (or in this case, won't) commit.

As self-driving cars emerge from Google and BMW and begin to hit the streets, a new cultural acceptance of machine risk will emerge as well. The world will see automobile collision casualty rates decrease overall and eventually, among waves of ire and protest, will learn to accept that on some occasions the computer is to blame for an accidental death.

But when a criminal who would not reoffend is kept in prison because of an incorrect prediction, we will never have the luxury of knowing. You can

prove innocent a legitimate transaction wrongly flagged as fraudulent, but an incarcerated person has no recourse to disprove unjust assumptions about what his or her future behavior outside prison would have been. If you prevent something, how can you be certain it was ever going to happen?

We're entrusting machines to contribute to life-changing decisions for which there can be no accountability: We can't measure the quality of these decisions, so there's no way to determine blame. We've grown comfortable with entrusting humans, despite their cherished fallibility, to make these judgment calls. A culture shift is nigh as we broaden this sacred circle of trust. PA sometimes makes wrong predictions but often proves to be less wrong than people. Bringing PA in to support decision making means introducing a new type of bias, a new fallibility, to balance against that of a person.

The development of computerized law enforcement presents extraordinarily tough ethical quandaries:

- Does the application of PA for law enforcement fly in the face of the very notion of judging a person as an individual? Is it unfair to predict a person's risk of bad behavior based on what other people—who share certain characteristics with that person—have done? Or, isn't the prediction by a human (e.g., a judge) of one's future crimes also intrinsically based only on prior observations of others, since humans learn from experience as well?
- A crime risk model dehumanizes the prior offender by paring him or her down to the extremely limited view captured by a small number of characteristics (variables input to a predictive model). But, if the integration of PA promises to lower the overall crime rate—as well as the expense of unnecessary incarceration—is this within the acceptable realm of compromises to civil liberties (on top of incarceration) that convicts endure?
- With these efforts under way, should not at least as much effort go into leveraging PA to improve offender rehabilitation; for example, by targeting those with the highest risk of recidivism? (In one groundbreaking case, the Florida Department of Juvenile Justice does just this—see Central Table 5.)

PA threatens to attain too much authority. Like an enchanted child with a Magic 8 Ball toy (originated in 1950), which is designed to pop up a *random* answer to a yes/no question, insightful human decision makers could place a great deal of confidence in the recommendations of a system they do not deeply understand. What may render judges better informed could also sway them toward less active observation and thought, tempting them to defer to the technology as a kind of crutch and grant it undue credence. It's important for users of PA—the judges and parole board members—to keep well in mind that it bases predictions on a much more limited range of factors than are available to a person.

THE CYCLICITY OF PREJUDICE

Yet another quandary lurks. Although science promises to improve the effectiveness and efficiency of law enforcement, when you formalize and quantify decision making, you inadvertently instill existing prejudices against minorities. Why? Because prejudice is cyclic, a self-fulfilling prophecy, and this cycling could be intensified by PA's deployment.

Across the United States, crime prediction systems calculate a criminal's probability of recidivism based on things like the individual's age, gender, and neighborhood, as well as prior crimes, arrests, and incarcerations. No government-sponsored predictive models explicitly incorporate ethnic class or other minority status.

However, ethnicity creeps into the model indirectly. Philadelphia's recidivism prediction model incorporates the offender's ZIP code, known to highly correlate with race. For this reason, redlining, the denying of services by banks, insurance companies, and other businesses by geographical region, has been largely outlawed in the United States.

Similarly, terrorist prediction models factor in religion. Levitt and Dubner's book *SuperFreakonomics* (HarperCollins, 2009) details a search for suspects among data held by a large UK bank. Informed in part by attributes of the September 11 perpetrators, as well as other known terrorists, a fraud detection analyst at the bank pinpointed a very specific group of customers to forward to the authorities. This *microsegment* was defined by factors such as

the types of bank accounts opened, existence of wire transfers and other transactions, record of a mobile phone, status as a student who rents, and a lack of life insurance (since suicide nullifies the policy). But to get the list of suspects down to a manageable size, the analyst filtered out people with non-Muslim names, as well as those who made ATM withdrawals on Friday afternoons—admittedly a proxy for practicing Muslims. Conceptually, this may not be a huge leap from the internment of suspected enemies of the state, although it should be noted that this was not a government-sponsored analysis. While this work has been criticized as an “egregious piece of armchair antiterrorism,” the bank analyst who delivered the suspect list to the authorities may exert power by way of his perceived credibility as a bank representative.

But even if such factors are disallowed for prediction, it’s still a challenge to avoid involving minority status.

Bernard Harcourt, a professor of both political science and law at the University of Chicago and author of *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, told *The Atlantic* that minority group members discriminated against by law enforcement, such as by way of profiling, are proportionately more likely to show a prior criminal record (e.g., since they may be screened more often), which artificially inflates the minority group’s incidence of criminal records. Rather than race being a predictor of prior offenses, prior offenses are indicative of race. By factoring in prior offenses to predict future crimes, “you just inscribe the racial discrimination you have today into the future.” It’s a cyclic magnification of prejudice’s already self-fulfilling prophecy.

Even Ellen Kurtz, who champions the adoption of the crime model in Philadelphia, admits, “If you wanted to remove everything correlated with race, you couldn’t use anything. That’s the reality of life in America.”

But don’t make data a scapegoat. It isn’t solely a petri dish in which racial discrimination grows—it’s also a tool that serves the fight against discrimination. Government departments outside law enforcement, such as the Federal Housing Finance Agency, the Education Department, and the Department of Housing and Urban Development, collect data for the very purpose of detecting discriminatory practices in banking loans, public education, affordable housing, and employment opportunities.

Within law enforcement, the math getting us in trouble could also remedy the problem by quantifying prejudice. However, that could be done only by introducing the very data element that—so far—remains outside the analysis, albeit inside the eye of every profiling police officer: race. Technically, there could be an analytical means to take this on if race were input into the system. This would require addressing new questions and debates analogous to those that arise with the implementation of equal-opportunity practices.

GOOD PREDICTION, BAD PREDICTION

Privacy is a compromise between the interests of the government and the citizen.

—Eric Schmidt, former Executive Chairman and CEO, Google

Information technology has changed just about everything in our lives. . . . But while we have new ethical problems, we don't have new ethics.

—Michael Lotti

When we think in terms of power, it is clear we are getting a raw deal: We grant private entities—with no interest in the public good and no public accountability—greater powers of persuasion than anyone has ever had before and in exchange we get free e-mail.

—Alexander Furnas, writer for *The Atlantic*

With great power comes great responsibility.

—Spider-Man's wise uncle (paraphrasing the Bible, Voltaire, and others)

Pregnancy prediction faces the opposite dilemma of that faced by crime prediction. Crime prediction causes damage when it predicts *wrong*, but predicting sensitive facts like pregnancy can cause damage when it's *right*. Like X-ray glasses, PA unveils new hot-button data elements for which all the fundamental data privacy questions must be examined anew. Sherlock Holmes, as well as his modern-day doppelganger Dr. Gregory House, size you up and embarrass you: A few scuff marks on your shoe and the detective knows you're having an affair. Likewise, no one wants her pregnancy unwittingly divulged; it's safe to assume organizations generally don't wish to divulge it, either.

It's tempting to write off these matters as benign in comparison to the qualms of crime prediction. KDnuggets, a leading analytics portal, took a poll: "Was Target wrong in using analytics to identify pregnant women from changes in their buying behavior?" The results were 17 percent "Yes," 74 percent "No," and 9 percent "Not sure" among the analytics community. One written comment pointed out that intent is relevant, asking, "When I yield a seat on a train to elderly people or a pregnant woman, am I 'trying to infer sensitive personal data such as pregnancy or elderliness'? Or just trying to provide the person with her needs?"

But knowledge of a pregnancy is extremely potent, and leaking it to the wrong ears can be life-changing indeed. As one online pundit proclaimed, imagine the pregnant woman's "job is shaky, and your state disability isn't set up right yet, and, although she's working on that, to have disclosure could risk the retail cost of a birth (\$20,000), disability payments during time off (\$10,000 to \$50,000), and even her job."

As with pregnancy, predictive models can also ascertain minority status—from behavior online, where divulging demographics would otherwise come only at the user's discretion. A study from the University of Cambridge shows that race, age, sexual orientation, and political orientation can be determined with high levels of accuracy based on one's Facebook likes. This capability could grant marketers and other researchers access to unvolunteered demographic information.

Google itself appears to have sacrificed a significant boon from predictive modeling in the name of privacy by halting its work on the automatic recognition of faces within photographs. When he was Google's CEO, Eric Schmidt stated his concern that facial recognition could be misused by organizations that identify people in a crowd. This could, among other things, ascertain people's locations without their consent. He acknowledges that other organizations will continue to develop such technology, but Google chose not to be behind it.

Other organizations agree: Sometimes it's better not to know. John Elder tells of the adverse reaction from one company's HR department when the idea of predicting employee death was put on the table. Since death is one

way to lose an employee, it's in the data mix. In a meeting with a large organization about predicting employee attrition, one of John's staff witnessed a shutdown when someone mentioned the idea. The project stakeholder balked immediately: "Don't show us!" Unlike healthcare organizations, this HR group was not meant to handle and safeguard such prognostications.

Predicting death is so sensitive that it's done secretly, keeping it on the down low even when done for benevolent purposes. One top-five health insurance company predicts the likelihood an elderly insurance policyholder will pass away within 18 months, based on clinical markers in the insured's recent medical claims. On the surface, this sounds potentially dubious. With the ulterior motives of health insurance often under scrutiny, one starts to imagine the terrible implications. Might the insurance company deny or delay the coverage of treatment based in part on how likely you are to die soon anyway? Not in this case. The company's purposes are altruistic. The predictions serve to trigger end-of-life counseling (e.g., regarding living wills and palliative care). An employee of the company told me the predictive performance is strong, and the project is providing clear value for the patients. Despite this, those at the company quake in their boots that the project could go public, agreeing only to speak with me under the condition of anonymity. "It's a very sensitive issue, easily misconstrued," the employee said.

The media goes too far when it sounds alarms that imply PA ought to be sweepingly indicted. To incriminate deduction would be akin to outlawing thought. It's no more than the act of figuring something out. If I glance into my friend's shopping cart and, based on certain items, draw the conclusion that she may be pregnant, have I just committed a *thoughtcrime*—the very act enforced against by Big Brother in George Orwell's *Nineteen Eighty-Four*? And so the plot twists, since perhaps critics of Target who would compare this kind of analysis to that of Big Brother are themselves calling the kettle black by judging Target for thoughtcrime. Pregnancy prediction need not be viewed as entirely self-serving—as with any marketing, this targeting does have potential to serve the customer. In the end, with all his eccentricities,

Sherlock Holmes is still our hero, and his revealing deductions serve the greater good.

“Privacy and analytics are often publicly positioned as mortal enemies, but are they really?” asks Ari Schwartz of the U.S. Department of Commerce’s National Institute of Standards and Technology. Indeed, some data hustlers want a free-for-all, while others want to throw the baby out with the bathwater. But Schwartz suggests, “The two worlds may have some real differences, but can probably live a peaceful coexistence if they simply understand where the other is coming from.”

It’s not what an organization comes to know; it’s what it *does* about it. Inferring new, powerful data is not itself a crime, but it does evoke the burden of responsibility. Target does know how to benefit from pregnancy predictions without actually divulging them to anyone (the alleged story of the pregnant teen is at worst an individual albeit significant gaffe). But any marketing department must realize that if it generates quasimedical data from thin air, it must take on, with credibility, the privacy and security practices of a facility or department commonly entrusted with such data. *You made it, you manage it.*

PA is an important, blossoming science. Foretelling your future behavior and revealing your intentions, it’s an extremely powerful tool—and one with significant potential for misuse. It’s got to be managed with extreme care. The agreement we collectively come to for PA’s position in the world is central to the massive cultural shifts we face as we fully enter and embrace the information age.

THE SOURCE OF POWER

New questions arise as we move from predicting the repeat offenses of convicts to the discovery of new potential suspects within the general populace of civilians. The following sidebar on *automatic suspect discovery* brings these questions to the surface, after which the next chapter turns to the source of predictive power—data—and explores the most bizarre insights it reveals, and how easy it is to be fooled by it.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA

Synopsis: It's a foregone conclusion that the world's largest spy organization running the country's largest surveillance data center and employing the world's largest number of PhD mathematicians considers predictive analytics (PA) a strategic priority. Can the NSA use machine learning supercomputers to fight terrorism—and can other agencies do so to fight crime in general—without endangering civil liberties?

Today's data privacy debate falters, because both sides are under-informed.

The NSA has endured intense scrutiny and suffered heavy backlash over its mass data collection that was unveiled in detail by whistleblower Edward Snowden in 2013. But don't give too much credence to the news or even the books—public discourse leaves out the greatest power law enforcement stands to gain from this data.

SUMMARY OF THE MAINSTREAM DEBATE REGARDING NSA DATA COLLECTION:

Privacy advocates: The NSA is violating civil liberties by collecting data on a massive scale about private citizens, including the majority who are not even suspected of any wrongdoing. Access to this data, whether in-house or by proxy via telecom companies, facilitates arbitrary snooping.

The NSA (and supportive legislators): We require comprehensive data in-house so we can rapidly investigate specific individuals when they become of interest. We do not inspect the activities of ordinary civilians in general.

This contentious dialogue only touches on half the story. Both sides fail to address what's really at stake for law enforcement: *Data empowers*

(continued)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

not only the investigation of established suspects, but also the discovery of new suspects. I would like to propose the following term for this emerging form of data-driven law enforcement:

Automatic Suspect Discovery (ASD)—The identification of previously unknown potential suspects by applying PA to flag and rank individuals according to their likelihood to be worthy of investigation, either because of their direct involvement in, or relationship to, criminal activities.

A note on automation: ASD flags new persons of interest who may then be elevated to suspect by an ensuing investigation. By the formal law enforcement definition of the word, an individual would not be classified as a suspect by a computer, only by a law enforcement officer.

ASD provides a novel means to unearth new suspects. Using it, law enforcement can hunt scientifically, more effectively targeting its search by applying PA, the same state-of-the-art, data-driven technology behind fraud detection, financial credit scoring, spam filtering, and targeted marketing.

THE SPY WHO LOVED MY DATA

To harness this potential, law enforcement needs the whole haystack. The government doesn't desire data about you just to spy at will—on the off chance you turn out to be a suspect. Rather, they actually require this data as a baseline in order to pursue their greater objective with ASD. This approach relies on wide-scale data access, even including data about both you and me—a full regimen of data about normal, innocent civilian activity unrelated to crime of any sort. Mathematically speaking, the broader a swath of noncriminal cases fed into the analysis, the better it works.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY:
THE REAL REASON THE NSA WANTS YOUR DATA
(CONTINUED)



Given this, ASD only amplifies the stakes of the contentious security-versus-privacy debate; both sides are bound to dig in and redouble their conviction. The promise of a novel technique for suspect discovery emboldens law enforcement's rationale for collecting as much data as possible. On the other side, privacy advocates perceive law enforcement's now stronger incentive as an even greater cause for alarm. Viewing the bulk collection of personal data itself as a violation of civil liberties, they argue the price is too high—especially given that any quantitative approach such as ASD cannot guarantee results *a priori*.

HOW IT WORKS: SHRINK THE HAYSTACK

Law enforcement (antiterrorism or otherwise) is a numbers game, a quest to find needles in the haystack that is the general population.

(continued)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

The working hours spent by agents, officers, and analysts constitute a precious, finite resource that must be allocated as effectively as possible. As staff collects evidence, follows leads, and studies forensics, there is no magic oracle to focus these efforts and ensure the quest is efficient. But PA can better target a portion of the work.

PA APPLICATION: AUTOMATIC SUSPECT DISCOVERY (ASD)

- 1. What's predicted:** Whether an individual is a "person of interest."
- 2. What's done about it:** Individuals with a sufficiently high predictive score are considered or investigated.

As with fraud detection, prediction shrinks the haystack to be searched. This multiplies the effectiveness of available human resources. By focusing time on the top echelon, those with the highest predictive scores, an investigator is more likely to come across worthy suspects. While it is reasonable to assume ASD pays off over time, investigators must understand the odds have only shifted; it's not a magic crystal ball. Most targets of investigation still turn out to be innocent—that is to say, the false positive rate will be lowered but by no means eliminated; the haystack is smaller but still large.

For best results, ASD may be applied repeatedly over a batch of predictive objectives. Its success depends on creatively defining *person of interest*, that is, the class of potential suspect being sought. For example, to predict known perpetrators of a rare crime such as terrorism, there may be too few known positive examples—"needles"—with which to train the predictive model. Therefore, ASD may be more likely to succeed when targeting instead a broader category of "interesting" persons, which could be defined as, for example, members

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

of an active surveillance group or persons with certain links to key criminal networks.⁴

As with all application areas, PA learns from data that encodes both positive and negative cases—in the case of ASD, both the known needles and the vast haystack, respectively. The analytical number-crunching process builds models (e.g., patterns or other formulations) to distinguish needles from hay. Models are then used to score each individual according to the probability of being a person of interest. This is the very purpose and function of core PA methods, such as *decision trees* and *ensemble models* (covered in Chapters 4 and 5, respectively).

EXAMPLE PATTERNS: WHAT IT COULD DISCOVER

Data brims with predictive potential. Even when the data about each individual is limited—such as with *metadata*, which characterizes e-mail and telephone communications by their time, date, destination, and the like—there's a lot to work with. These are the nuts and bolts of behavior that are often at least as revealing as, not to mention much easier to process than, communication content, that is, the typed message of an e-mail or spoken words during a phone call.

The experts see the predictive potential. Dean Abbott, a senior hands-on consultant who's applied PA for fraud detection for both the private and public sectors, agrees that ASD is a worthy application of

(continued)

⁴ However, as a counterexample, note that the pattern (aka microsegment) designed by a UK bank analyst covered earlier in this chapter (from *SuperFreakonomics*) was formed vis-à-vis a target set of known terrorism perpetrators.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

PA. “Yes, I absolutely think it would be worth the effort to build a predictive model based on metadata that identifies new leads for a given hotlist—especially one that incorporates link data of who’s called whom,” he said.

A predictive model acts as a choosy, discriminating fishing net. It may include patterns that capture a wide yet precisely defined spectrum of possibilities, arbitrarily abstract and multidimensional. Investigation activities target the individuals who match such patterns; those matches define the now smaller haystack to be searched.

For example, Defense Department-funded university research identified certain circumstances—characterized by the following pattern—that present an 88 percent probability of an attack by the South Asian terrorist organization Lashkar-e-Taiba:

- **PATTERN:** *Between five and 24 of the organization’s operatives have been arrested and operatives are on trial in India or Pakistan.*

In a similar vein, such patterns could serve to identify attackers rather than impending attacks. Here is the controversial pattern designed by a UK bank analyst to discover terrorism suspects covered earlier in this chapter (from the book *SuperFreakonomics*—due to its intentional religious discrimination, I consider this example ethically prohibitive; despite that, I include this rare, public, data-driven example to illustrate the mechanics of patterns):

- **PATTERN:** *The individual has opened a certain type of bank account, has placed certain types of wire transfers or other transactions,*

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

has a mobile phone (this example is from the early 2000s), is listed as a student who rents, shows no life insurance policy (suicide would nullify the policy), and holds certain attributes that indicate a likelihood to be Muslim.

To further illustrate the concept, here are three simple example patterns to identify possible suspects that could be generated by PA (fictional, for illustrative purposes only):⁵

- **PATTERN:** *The caller has placed calls from at least two countries per week for eight months, calls from an average of four countries per week, has placed two calls to numbers two degrees of separation from a hotlist of numbers, and received a call from a hotlist number within the last four hours (such a rule could trigger a real-time alert to analysts).*
- **PATTERN:** *The caller shows typical calling patterns (regarding frequency, variance of call durations, and the number of both frequent and infrequent correspondents), but with the addition of calls to more than four never-before-called government phone numbers per week on most weeks, across more than seven countries, for three months.*
- **PATTERN:** *The e-mail address, logged into at a flagged Internet café, is likely a proxy for another e-mail address that has second-degree*

(continued)

⁵ Patterns like these could be derived by decision trees in combination with specialized data preparation (predictor variables designed for call pattern detection). The adeptness of such patterns improves by combining a larger number of such pattern-matching rules—hundreds rather than only several—as achieved by *ensemble models*.

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

ties to a hotlist of e-mail addresses. The proxy pairing is based on the frequency of forwards between the two that are not replied to, the overlap in the sets of correspondents, and similar geolocation login patterns.

Although a particular pattern may “catch no fish” and come up empty, when a number of even the most arcane patterns are applied across a large population of civilians, there’s an opportunity to eventually find matches. Applied with tactical panache, I believe that iteratively running ASD projects that incorporate human creativity and law enforcement expertise is bound to deliver.

Law enforcement has an unfair advantage. Criminals lack one key resource required to compete against this form of intelligence: the data. Criminal organizations generally cannot recreate law enforcement’s surveillance of persons of interest, let alone the much larger dataset of negative examples, the civilians. So they have no means to ascertain the predictive patterns that crime fighters derive from this data, which leaves them with no insight to evade being detected by such patterns. As with *network intrusion detection*, ASD achieves a qualitatively unparalleled advancement in this escalating arms race, the ongoing competition between detection and evasion.

PRESUMPTION: THE NSA USES PREDICTIVE ANALYTICS

It’s a foregone conclusion the NSA considers PA a strategic priority. Any use of PA by the NSA is necessarily a secret; the lack of public examples is the nature of the beast. However, wondering whether they use it is like speculating whether a chef who bought flat pasta, meat sauce, mozzarella, and ricotta is making lasagna. Beyond a reasonable doubt, the

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

world's largest spy organization running the country's largest surveillance data center and employing the world's largest number of PhD mathematicians strives to analytically learn from data.

There's much supporting the assumption that the NSA has worked with PA and will continue to do so (see the corresponding section in this book's Notes at www.PredictiveNotes.com for details pertaining to the following summary list):

- NSA documents and official documents about the NSA explicitly indicate established capabilities in machine learning and pattern discovery.
- The NSA has purchased intelligence software solutions that include PA capabilities from two companies, Palantir and Cognito.
- NSA job postings for "data scientists" seek candidates experienced with machine learning and other related technologies.
- The NSA's domestic counterpart conducts ASD: The U.S. Department of Justice released a report describing how the FBI applies PA to assign terrorism "risk scores" to possible suspects.
- PA stands clear as an increasingly common practice for law enforcement of all kinds, including U.S. Armed Forces-funded terrorism prediction, predictive policing, recidivism prediction, and fraud detection, arguably the leading government application of PA.
- Data-driven suspect discovery is a publicly established concept. The popular book *SuperFreakonomics* even covers a specific example of iteratively redefining a pattern to discern terrorism suspects (summarized earlier in this chapter).

(continued)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

THE ARGUMENT FOR COLLECTING THE WHOLE HAYSTACK

Law enforcement is intrinsically destined to apply PA, which serves to discover potential suspects who would otherwise continue undetected. Just as Santa Claus defies scale and visits every single household overnight with lightning speed, this virtual cop sizes up every civilian by matching against scientifically established patterns. And just as companies screen each transaction for fraud and each employee for propensity to quit their job, so too does a government strive to screen each civilian for connection to crime.

Without an understanding of ASD, privacy advocates trip up on fallacies. Wisconsin Rep. James Sensenbrenner, who himself introduced the Patriot Act in the House, argued, “The bigger haystack makes it harder to find the needle.” It’s a common misconception. Even with regard to the private sector, journalists warn of “drowning” in too much data. But PA practitioners recognize that the data glut is not a problem—it’s an opportunity.

Public figures overlook an irony intrinsic to ASD: Wide-scale data collection can serve to identify the few who should be actively surveilled, rather than spy on the many. But some pundits presume the opposite necessarily holds true, in part because ASD is not widely known. Robert Scheer, author of *They Know Everything about You: How Data-Collecting Corporations and Snooping Government Agencies Are Destroying Democracy*, inadvertently invoked ASD when he wrote, “Intelligence should be about learning what you need to know and don’t already, not just about sucking up unmanageable gigabytes of minutiae everywhere in the world, which has been the NSA’s enormously costly and ineffectual game of choice.”

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

A comparable controversy plays out in the field of medicine, where the potential for lifesaving insights also compels open data. Healthcare data-sharing proponent John Wilbanks argues that privacy protections on clinical research data slow down research. “These are tools that we created to protect us from harm, but what they’re doing is protecting us from innovation now,” he said in a TED talk. “When I tell cancer survivors that this tool we created to protect them is actually preventing their data from being used, . . . their reaction is not, ‘Thank you, God, for protecting my privacy.’ It’s outrage that we have this information and we can’t use it.”

And so law enforcement by its nature lusts for ever-growing surveillance, just as users of PA for all purposes across all sectors perpetually crave bigger data.

THE COUNTERARGUMENT: CURTAIL MONITORING TO PROTECT CIVIL LIBERTIES

For all its promise, mass government surveillance risks civil liberties and therefore cannot go unrestrained. Those civilians whose data is considered up close by law enforcement personnel, although constituting a minority of cases, are vulnerable to high degrees of potentially unfounded scrutiny and other enforcement activities. With data collection capabilities growing in scope, an agent of the law is armed with more information about the person of interest than he or she may reasonably have known was being tracked. This data then becomes the subject of the particular prejudices of the agent, who considers the demographic profile in combination with perceived aspects of the suspect’s private online and telecommunication activities. Over large

(continued)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

numbers of cases, for some this inevitably leads to grave inconveniences, further invasion into their personal life, or even harassment and unjust prosecution.

The presence of this potential infliction upon the few curtails liberties for the many. Glenn Greenwald, author of *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State* and lead journalist on the 2013 disclosures, wrote that “it is in the realm of privacy where creativity, dissent, and challenges to orthodoxy germinate. A society in which everyone knows they can be watched by the state—where the private realm is effectively eliminated—is one in which those attributes are lost, at both the societal and the individual level. . . . Mass surveillance by the state is therefore inherently repressive.”

Brazilian President Dilma Rousseff brought this reasoning to its natural conclusion, following revelations that the NSA had monitored Brazilian citizens and allegations that the intelligence organization had even intercepted official e-mail and telephone communications of the president herself. She declared, “In the absence of the right to privacy, there can be no true freedom of expression and opinion, and therefore no effective democracy.”

Besides requiring wide-scale data collection to feed as input, ASD’s outputs—the predictions it generates—also incur risk to liberty. Potential suspects flagged by ASD face the risk of invasive treatment. Innocent civilians, the inevitable false positives among ASD’s targets, could fall subject to unjust scrutiny, drilling down into the data collected about them. When ASD flags an individual, this does not necessarily mean reasonable suspicion has been established by way of specific evidence. However, the personal data previously collected about the individual will not continue to lay dormant—a law

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

enforcement officer will access and leverage it, which may in turn lead to unwarranted acts of search, seizure, or detention.

The ACLU calls this profiling. In a discussion of ASD with Allen Gilbert, the executive director at the American Civil Liberties Union of Vermont, he told me: “Predictive analytics is in essence a form of profiling. It provides an excuse rather than evidence to target someone as a criminal suspect. It short-circuits the Fourth Amendment’s protections against search and seizure without reasonable suspicion of crime. A civil libertarian gasps that such pre-judging—prejudice—is considered justified in modern-day crime fighting.”

CONCLUSION: A SMARTER DEBATE

Want a productive debate? Then learn more—whichever side you’re on. Any simple, sweeping resolution put forth overlooks a great depth of multilayered gray area. Sound bites don’t cut it.

We face two extremely challenging tasks:

- To balance the great value aggregated data bears against the danger it holds. The agreed-upon extent of active government surveillance can range across a continuum. At one extreme, at least some minimal level of tracking is broadly accepted without controversy, such as each time we drive through a tollbooth or tunnel, every flight we take, and each time we cross an international border. At the other extreme, there’s also general agreement that particularly high levels of monitoring would be too much, for example if the government required a video feed for every room in every building.

(continued)

SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY: THE REAL REASON THE NSA WANTS YOUR DATA (CONTINUED)

- To determine whether and how ASD may safely target law enforcement activities. By design, data-driven investigation more effectively targets and could *decrease* the prevalence of inaccurate human discrimination that accompanies investigations driven by “gut” or “hunch.” Is it possible for law enforcement to investigate analytically-derived leads in a prudent manner, or does ASD entail an unacceptably high intrinsic risk for law enforcement abuse?

The position of agreed compromise on these two questions—destined to continuously evolve, by the way—must be set by *more deeply informed debate and negotiation*. Both opposing sides must learn more about the other’s concerns:

- Data hustlers** who support increased data collection by law enforcement must become deeply familiar with the philosophy, practicalities, and political history that illustrate how compromised privacy brings a loss of liberty and incurs the risk of abuse by law enforcement officers. Furthermore, to inform the risks at hand, law enforcement must render data collection practices and internal data access regulations publicly transparent.
- Privacy advocates** who support decreased data collection by law enforcement must come to understand why ASD presents a much stronger incentive for broad-scale data collection than if data were only to serve for investigating individuals: It provides a means to unearth new suspects who might otherwise go undetected, a key capability for the war on terror as well as other law enforcement efforts.

**SPECIAL SIDEBAR ON AUTOMATIC SUSPECT DISCOVERY:
THE REAL REASON THE NSA WANTS YOUR DATA
(CONTINUED)**

Whatever the extent of data collection, as ASD continues to develop it must be carefully managed. I contacted an expert on the ramifications PA holds for *reasonable suspicion*, a legal standard for everyday law enforcement activity. His name is Andrew Ferguson, a law professor of the University of the District of Columbia. He put it this way: “Predictive analytics is clearly the future of law enforcement. The problem is that the forecast for transparency and accountability is less than clear.”

© 2016 by Eric Siegel. Published 2016 by John Wiley & Sons, Inc.



CHAPTER 3

The Data Effect

A Glut at the End of the Rainbow

We are up to our ears in data, but how much can this raw material really tell us? What actually makes it predictive? What are the most bizarre discoveries from data? When we find an interesting insight, why are we often better off not asking why? In what way is bigger data more dangerous? How do we avoid being fooled by random noise and ensure scientific discoveries are trustworthy?

Spotting the big data tsunami, analytics enthusiasts exclaim, “Surf’s up!”

We’ve entered the golden age of predictive discoveries. A frenzy of number crunching churns out a bonanza of colorful, valuable, and sometimes surprising insights:¹

- People who “like” curly fries on Facebook are more intelligent.
- Typing with proper capitalization indicates creditworthiness.
- Users of the Chrome and Firefox browsers make better employees.
- Men who skip breakfast are at greater risk for coronary heart disease.
- The demand for Pop-Tarts spikes before a hurricane.
- Female-named hurricanes are more deadly.
- High-crime neighborhoods demand more Uber rides.

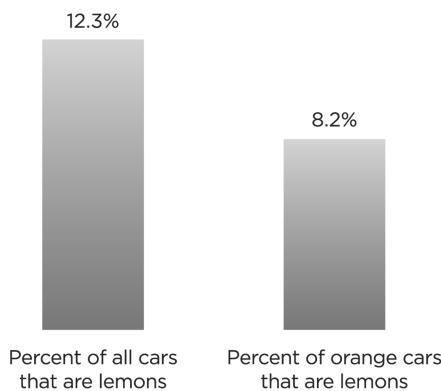
¹ For more details on these findings, see the “Bizarre and Surprising Insights” tables later in this chapter; for the specific citations, see the corresponding Notes (at www.PredictiveNotes.com).

A CAUTIONARY TALE: ORANGE LEMONS

Look like fun? Before you dive in, be warned: This spree of data exploration must be tamed with strict quality control. It's easy to get it wrong and end up with egg on your face.

In 2012, a *Seattle Times* article led with an eye-catching predictive discovery: “An orange used car is least likely to be a lemon.”² This insight came from a predictive analytics (PA) competition to detect which used cars are bad buys (*lemons*). While insights also emerged pertaining to other car attributes—such as make, model, year, trim level, and size—the apparent advantage of being orange caught the most attention. Responding to quizzical expressions, data wonks offered creative explanations, such as the idea that owners who select an unusual car color tend to have more of a “connection” to and take better care of their vehicle.

Examined alone, the “orange lemon” discovery appeared sound from a mathematical perspective. Here’s the specific result:



This shows orange cars turn out to be lemons one third less often than average. Put another way, if you buy a car that’s *not* orange, you increase your risk by 50 percent.

Well-established statistics appeared to back up this “colorful” discovery. A formal assessment indicated it was *statistically significant*, meaning that the

² This discovery was also featured by *The Huffington Post*, *The New York Times*, *National Public Radio*, *The Wall Street Journal*, and the *New York Times* Bestseller *Big Data: A Revolution That Will Transform How We Live, Work, and Think*.

chances were slim this pattern would have appeared only by random chance. It seemed safe to assume the finding was sound. To be more specific, a standard mathematical test indicated there was less than a 1 percent chance this trend would show up in the data if orange cars weren't actually more reliable.

But something had gone terribly wrong. The “orange car” insight later proved inconclusive. The statistical test had been applied in a flawed manner; the press had run with the finding prematurely. As data gets bigger, so does a potential pitfall in the application of common, established statistical methods. We'll dive into this dilemma later—but for now here's the issue in a nutshell: *Testing many predictors means taking many small risks of being fooled by randomness, adding up to one big risk.*

This chapter first establishes just how important an opportunity data represents, and then shows how to securely tap it—here's the flow of topics:

The source: where data comes from.

- Why logs of transactions aren't boring
- Why *social data* isn't always an oxymoron
- Estimating the mass mood of the public
- The massive recycling effort that supplies data for PA

The enormousness: how much there is and what the *big* in big data actually means.

The excitement: why data is so predictive—The Data Effect.

The gold rush: what data tells us—46 fascinating discoveries.

Caveat #1: why causality is generally an unknown.

Caveat #2: what went wrong with the “orange lemons” case and how to tap data's potential without drawing false conclusions.

THE SOURCE: OTHERWISE BORING LOGS FUEL PREDICTION

Today's predictive gold mine occurred by happy accident. Most data accumulates not to serve analytics, but as the by-product of routine tasks.

Consider all the phone calls you make. Your wireless provider logs your communications for billing and other transactional purposes. Boring! And

yet these logs also reveal a wellspring of behavioral trends that characterize you and your contacts (and serve law enforcement activities, as discussed in the previous chapter). Companies leverage the predictive power of such consumer behavior to, for example, keep consumers around. By predicting who's going to leave, companies target offers—such as a free phone—in order to retain would-be defectors.

"Social data" may sound like an oxymoron to many, but data about social behavior predicts like nobody's business. Optus, a leading cell phone carrier in Australia, doubled the precision of predicting whether a customer will cancel by incorporating the behavior of each customer's social contacts: If the people you regularly call defect to another wireless provider, there's an up to sevenfold greater risk you will also do so, as more than one telecom has discovered.³

Beyond the telecom industry, another immense sector of modern society stockpiles records of person-to-person interactions: social media sites like Facebook, Twitter, and an endless assortment of blogs. Seeing the potential, the financial industry taps these sites to help assess the creditworthiness of would-be debtors, and the Internal Revenue Service taps them to check out taxpayers. City health departments predict restaurant health code violations via Yelp reviews.

In short, what you've posted online may help determine whether your application for a credit card is approved, whether your tax return is audited, and whether a restaurant is inspected.

SOCIAL MEDIA AND MASS PUBLIC MOOD

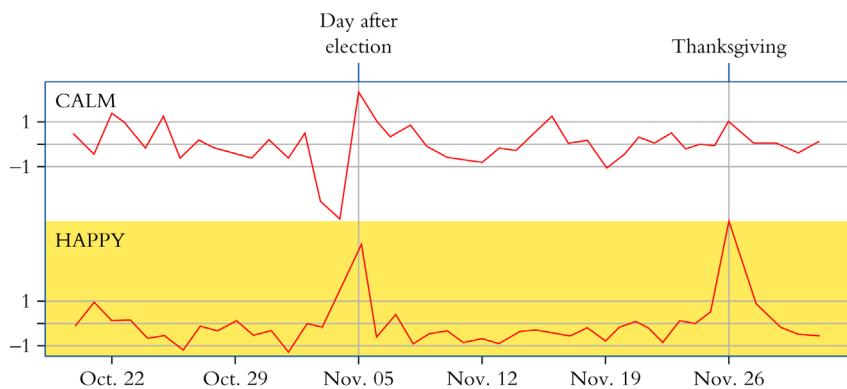
Can a population's overall average mood predict mass behavior? Many bet yes. A trending area of research taps social media posts to gauge the aggregate mood of the public. Researchers evaluate these readings of mass mood for their ability to predict all kinds of population-level behaviors, including the stock market, product sales, top music hits, movie box-office revenue, Academy Award and Grammy winners, elections, and unemployment statistics.

³ As with the law enforcement examples in Chapter 2's sidebar on automatic suspect discovery, this represents another application area where cellphone *metadata* alone proves to be predictively valuable.

Emotions don't usually fall within the domain of PA. Feelings are not concrete things easily tabulated in a spreadsheet as facts and figures. They're ephemeral and subjective. Sure, they may be the most important element of our human condition, but their subtleties place them outside the reach of most hard science. While a good number of neuroscientists are wiring up the noggins of undergraduate students in exchange for free pizza, many data scientists view this work as irrelevant, far removed from common applications of PA.

But social media blares our emotions. Bloggers, tweeters, and posters broadcast their thoughts, thereby transforming from private, introverted "Dear Diary" writers into vocal extroverts. A mass chorus expresses freely, unfettered by any preordained purpose or restriction. Bloggers alone render an estimated 864,000 posts per day, and in so doing act as an army of volunteers who express sentiment on the public's behalf.

Take a look at how our collective mood moves. Here's sample output of a word-based measure of mood by researchers at Indiana University. Based on a feed from Twitter, it produces daily readings of mass mood for the dimensions *calm* versus *anxious*, and *happy* versus *unhappy* (shown from October 2008 to December 2008):⁴



⁴ Johan Bollen, Huina Mao, and Xiao-Jun Zeng, "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2, no. 1 (March 2011). Figure reproduced with permission.

As we oscillate between elation and despair, this jittery movement reveals that we are a moody bunch. The time range shown includes a U.S. presidential election as well as Thanksgiving. Calmness rebounds once the voting of Election Day is complete. Happiness spikes on Thanksgiving.

A tantalizing prospect lingers for black-box trading if the mass mood approach bears fruit predicting the stock market. While there's not yet publicly known proof that it could predict the market well enough to make a killing, optimistic pioneers believe mass mood will become a fundamental component of trading analysis, alongside standard economic gauges. Entrepreneurial quant Randy Saaf said, "We see 'sentiment' as a diversified asset class like foreign markets, bonds, [and] gold."

RECYCLING THE DATA DUMP

One man's trash is another man's treasure.

By leveraging social media in a new way, researchers discover newfound value in oversharing. People tweet whatever the heck suits their fancy. If someone tweets, "I feel awesome today! Just wanted to share," you might assume it interests only the tweeter's friends and family, and there's no value for the rest of the world. As with most applications of PA, though, the data at hand is readily repurposed.

This repurposing signifies a mammoth recycling initiative: the discovery of new value in the data avalanche. Like the millions of chicken feet the United States has realized it can sell to China rather than throw away, our phenomenal accumulation of 1's and 0's surprises us over and over with newfound applications. Calamari was originally considered junk, as was the basis for white chocolate. My mom, Lisa Schamberger, makes photographic art of compost, documenting the beauty inherent in organic waste. Mad scientists want to make use of nuclear waste. I can assure you, data scientists are just as mad.

Growing up watching *Sesame Street*, I got a kick out of the creature Oscar the Grouch, who lives in a garbage can and sings a song about how much he loves trash. It turns out Oscar isn't so crazy after all.

If social media amounts to large-scale, unregulated graffiti, there's a similar phenomenon with the millions of encyclopedias' worth of organizational data scrawled onto magnetic media for miscellaneous operational functions. It's a zillion tons of human refuse that does not smell. What do ScarJo, Iceland, and borscht have in common with data? They're all beautiful things with unwelcoming names.

Most data is not accumulated for the purpose of prediction, but PA can learn from this massive recording of events in the same fashion that you can learn from your accumulation of life experience. As a simple example, take a company's record of your e-mail address and membership status—utilitarian, yet also predictive. During one project, I found that users who signed up with an [Earthlink.com](#) (an Internet provider) e-mail address were almost five times more likely to convert from a free trial user level to the premium paid level than those with a [Hotmail.com](#) e-mail address. This could be because those who divulged only a temporary e-mail account—which is the intent for some users of free e-mail services like Hotmail—were, on average, less committed to their trial membership. Whatever the reason, this kind of discovery helps a company predict who will be acquired as a paying customer.

THE INSTRUMENTATION OF EVERYTHING WE DO

Count what is countable, measure what is measurable, and what is not measurable, make measurable.

—Galileo

Intangibles that appear to be completely intractable can be measured.

—Douglas Hubbard, *How to Measure Anything*

Some historians assert that we are now experiencing the information revolution, following the agricultural and industrial revolutions. I buy it.

Colin Shearer, a PA leader at IBM, eloquently states that the key to the information revolution is “the instrumentation of everything.” More and more, each move you make, online and offline, is recorded, including transactions conducted, websites visited, movies watched, links clicked, friends called, opinions posted, dental procedures endured, sports games won (if you’re a professional athlete), traffic cameras passed, flights taken, Wikipedia articles edited, and earthquakes experienced. Countless sensors deploy daily. Mobile devices, robots, and shipping containers record movement, interactions, inventory counts, and radiation levels. Personal health monitors watch your vital signs and exercise routine. The mass migration of online applications from your desktop to the cloud (aka *software as a service*) makes even more of your computer use recordable by organizations.

Free public data is also busting out, so a wealth of knowledge sits at your fingertips. Following the *open data* movement, often embracing a not-for-profit philosophy, many data sets are available online from fields like biodiversity, business, cartography, chemistry, genomics, and medicine. Look at one central index, www.kdnuggets.com/datasets/, and you’ll see what amounts to lists of lists of data resources. The Federal Chief Information Officer of the United States launched Data.gov “to increase public access to high value, machine readable datasets generated by . . . the Government.” Data.gov sports over 390,000 data sets, including data about marine casualties, pollution, active mines, earthquakes, and commercial flights. Its growth is prescribed: A directive in 2009 obliged all U.S. federal agencies to post at least three “high-value” data sets.

Far afield of government activities, a completely different accumulation of data answers the more forbidden question, “Are you having *fun* yet?” For a dating website, I predicted occurrences of *online flirtation*. After all, as data shows, you’re much more likely to be retained as a customer if you get some positive attention. When it comes to recording and predicting human behavior, what’s more fundamental than our mating rituals? For this project, actions such as a virtual “wink,” a message, or a request to connect as “friends” counted as “flirtatious.” Working up a sort of digital tabloid magazine, I produced reports such as the average waiting times before a

flirt is reciprocated, depending on the characteristics of the customer. For example:

Sexual orientation:	Average hours before reciprocal flirt (if any):
Man seeking man	40
Woman seeking man	33
Man seeking woman	43
Woman seeking woman	55

For your entertainment, here's an actual piece of code from a short 175-line computer program called "Flirtback" that I wrote (in the computer language AWK, an oldie but goodie):

```
sex = sexuality[flirt_to]; # sexual orientation
sumbysex[sex] += (delta/(60*60));
nPairsSex[sex]++
```

Come on, you have to admit that's some exciting stuff—enough to keep any computer programmer awake.

Data expresses the bare essence of human behavior. What it doesn't capture is the full dimension and innuendo of human experience—and that's just fine for PA. Because organizations record the aspects of our actions important to their function, one extraordinarily elusive, daunting task has already been completed in the production of raw materials for PA: abstracting the infinite complexity of everyday life and thereby defining which of its endless details are salient.

A new window on the world has opened. Professor Erik Brynjolfsson, an economist at the Massachusetts Institute of Technology, compares this mass instrumentation of human behavior to another historic breakthrough in scientific observation. "The microscope, invented four centuries ago, allowed people to see and measure things as never before—at the cellular level," said *The New York Times*, explaining Brynjolfsson's perspective. "It was a revolution in measurement. Data measurement is the modern

equivalent of the microscope.” But rather than viewing things previously too small to see, now we view things previously too big.

BATTEN DOWN THE HATCHES: TMI

There are over 358 million trillion gallons of water on Earth.

—A TV advertisement for Ice Mountain Spring Water

The world now contains more photographs than bricks.

—John Szarkowski, Director of Photography,
Museum of Modern Art (back in 1976)

All this tracking dumps upon us a data glut. Six hundred blog posts are published per minute; by 2011, there were over 100 million blogs across WordPress and Tumblr alone. As for Twitter, “Every day, the world writes the equivalent of a 10-million-page book in Tweets or 8,163 copies of Leo Tolstoy’s *War and Peace*,” says the official Twitter blog. Stacking that many copies of the book “would reach the height of about 1,470 feet, nearly the ground-to-roof height of Taiwan’s Taipei 101, the second tallest building in the world.”

YouTube gains an hour of video each second. Estimates put the World Wide Web at over 8.32 billion Web pages. Millions of online retail transactions take place every hour. More photos are taken daily than in the first 100 years of photography, more in two minutes than in all of the 1800s, with 200 million uploaded to Facebook every day. Femto-photography takes a trillion frames per second to capture light in motion and “see around corners.” Over 7 billion mobile devices capture usage statistics. More than 100 things per second connect to the Internet, and this rate is increasing; by 2020 the “Internet of Everything” will connect 50 billion things, Cisco projects.

Making all this growth affordable, the cost of data storage is sinking like a rock. The cost per gigabyte on a hard drive has been exponentially decaying since the 1980s, when it approached \$1 million. By 2014, it reached 3 cents. We can afford to never delete.⁵

⁵ When first released, Google’s free e-mail service, Gmail, had no option to delete a message, only to archive it.

Government intelligence aims to archive vast portions of all communication. The U.S. National Security Agency's \$2 billion Utah Data Center, a facility five times the size of the U.S. Capitol, is designed to store mammoth archives of human interactions, including complete phone conversations and e-mail messages.

Scientific researchers are uncovering and capturing more and more data, and in so doing revolutionizing their own paradigms. Astronomers are building a new array of radio telescopes that will generate an exabyte of data per day (an exabyte is a quintillion bytes; a byte is a single value, an integer between 0 and 255, often representing a single letter, digit, or punctuation mark). Using satellites, wildlife conservationists track manta rays, considered vulnerable to extinction, as the creatures travel as far as 680 miles in search of food. In biology, as famed futurist Ray Kurzweil portends, given that the price to map a human genome has dropped from \$1 billion to a few thousand dollars, information technology will prove to be the domain from which this field's greatest advances emerge.

Overall, data is growing at an incomprehensible speed, an estimated 2.5 quintillion bytes (exabytes) of data per day. A quintillion is a 1 with 18 zeros. In 1986, the data stored by computers, printed on double-sided paper, could have covered the Earth's landmasses; by 2011, it could have done so with two layers of books.

The growth is exponential. Data more than doubles every three years. This brought us to an estimated 8 zettabytes in 2015—that's 8,000,000,000,000,000,000,000 (21 zeros) bytes. Welcome to Big Bang 2.0.

The next logical question is: What's the most valuable thing to do with all this stuff? This book's answer: *Learn from it how to predict.*

WHO'S YOUR DATA?

Good, better, best, bested. How do you like that for a declension, young man?

—Edward Albee, *Who's Afraid of Virginia Woolf?*

Bow your head: The hot buzzword *big data* has ascended to royalty. It's in every news clip, every data science presentation, and every advertisement for analytics solutions. It's a crisis! It's an opportunity! It's a crisis of opportunity!

Big data does not exist. The elephant in the room is that there is no elephant in the room. What's exciting about data isn't how much of it there is, but how quickly it is growing. We're in a persistent state of awe at data's sheer quantity because of one thing that does not change: There's always so much more today than yesterday. Size is relative, not absolute. If we use the word *big* today, we'll quickly run out of adjectives: "big data," "bigger data," "even bigger data," and "biggest data." The International Conference on Very Large Databases has been running since 1975. We have a dearth of vocabulary with which to describe a wealth of data.⁶

"Big data" is also grammatically incorrect. It's like saying "big water." Rather, it should be "a lot of data" or "plenty of data."

What's big about data is the excitement—about its rate of growth and about its predictive value.

THE DATA EFFECT: IT'S PREDICTIVE

*The leg bone connected to the knee bone,
and the knee bone connected to the thigh bone,
and the thigh bone connected to the hip bone.*

—From the song "Dry Bones"

There's a ton of it—so what? What guarantees that all this residual rubbish, this by-product of organizational functions, holds value? It's no more than an extremely long list of observed events, an obsessive-compulsive enumeration of things that have happened.

The answer is simple. Everything in the world is affected by connections to other things—things touch and cause one another in all sorts of ways—and this is reflected in data. For example:

- Your purchases relate to your shopping history, online behavior, and preferred payment method, and to the actions of your social

⁶ Other buzzwords also have their issues. Calling this work *data science* is like calling a librarian a "book librarian." Calling it *data mining* is like calling gold mining "dirt mining."

contacts. Data reveals how to predict consumer behavior from these elements.

- Your health relates to your life choices and environment, and therefore data captures connections predictive of health based on type of neighborhood and household characteristics.
- Your job satisfaction relates to your salary, evaluations, and promotions, and data mirrors this reality.

Data always speaks. It always has a story to tell, and there's always something to learn from it. Data scientists see this over and over again across PA projects. Pull some data together and, although you can never be certain what you'll find, you can be sure you'll discover valuable connections by decoding the language it speaks and listening. That's *The Data Effect* in a nutshell.

The Data Effect: *Data is always predictive.*

This is the assumption behind the leap of faith an organization takes when undertaking PA. Budgeting the staff and tools for a PA project requires this leap, knowing not what specifically will be discovered and yet trusting that something will be. Sitting on an expert panel at Predictive Analytics World, leading UK consultant Tom Khabaza put it this way: “Projects never fail due to lack of patterns.” With The Data Effect in mind, the scientist rests easy, secure the analysis will be fruitful.

Data is the new oil. It's this century's greatest possession and often considered an organization's most important strategic asset. Several thought leaders have dubbed it as such—“the new oil”—including European Consumer Commissioner Meglena Kuneva, who also calls it “the new currency of the digital world.” It's not hyperbole. In 2012, Apple, Inc. overtook Exxon Mobil Corp., the world's largest oil company, as the most valuable publicly traded company in the world. Unlike oil, data is extremely easy to transport and cheap to store. It's a bigger geyser, and this one is never going to run out.

THE BUILDING BLOCKS: PREDICTORS

Prediction starts small. PA's building block is the *predictor variable*, a single value measured for each individual (known informally as a *factor*, *attribute*, *feature*, or *predictor*, and more formally as an *independent variable*). For example, *recency*, the number of weeks since the last time an individual made a purchase, committed a crime, or exhibited a medical symptom, often reveals the chances that individual will do it again in the near term. In many arenas, it makes sense to begin with the most *recently* active people first, whether for marketing contact, criminal investigation, or clinical assessment.

Similarly, *frequency*—the number of times the individual has exhibited the behavior—is also a common, fruitful measure. People who have done something a lot are more likely to do it again.

In fact, it is usually what individuals *have done* that predicts what they *will do*. And so PA feeds on data that extends past dry yet essential demographics like location and gender to include *behavioral predictors* such as recency, frequency, purchases, financial activity, and product usage such as calls and Web surfing. These behaviors are often the most valuable—it's always a *behavior* that we seek to predict, and indeed behavior predicts behavior. As Jean-Paul Sartre put it, “[A man’s] true self is dictated by his actions.”

PA builds its power by combining dozens—or even hundreds—of predictors. You give the machine everything you know about each individual, and let 'er rip. The core learning technology to combine these elements is where the real scientific magic takes place. That learning process is the topic of the next chapter; for now, let's look at some interesting individual predictors.

FAR OUT, BIZARRE, AND SURPRISING INSIGHTS

Some predictors are more fun to talk about than others.

Are customers more profitable if they don't think? Does crime increase after a sporting event? Does hunger dramatically influence a judge's life-altering decisions? Do online daters more consistently rated as attractive receive *less* interest? Can promotions *increase* the chance you'll quit your job? Do vegetarians miss fewer flights? Does your e-mail address reveal your intentions?

Yes, yes, yes, yes, yes, yes, and yes!

Welcome to the *Ripley's Believe It or Not!* of data science. Poring over a potpourri of prospective predictors, PA's aim isn't only to assess human hunches by testing relationships that seem to make sense, but also to explore a boundless playing field of possible truths beyond the realms of intuition. And so, with The Data Effect in play, PA drops onto your desk connections that seem to defy logic. As strange, mystifying, or unexpected as they may seem, these discoveries help predict.

Here are some colorful discoveries, each pertaining to a single predictor variable (for each example's citation, see the Notes at www.PredictiveNotes.com).

Bizarre and Surprising Insights—Consumer Behavior

Insight	Organization	Suggested Explanation ⁷
Guys literally drool over sports cars. Male college student subjects produce measurably more saliva when presented with images of sports cars or money.	Northwestern University Kellogg School of Management	<i>Consumer impulses are physiological cousins of hunger.</i>
If you buy diapers, you are more likely to also buy beer.	Osco Drug	<i>Daddy needs a beer.</i>

(continued)

⁷ Warning: Do not give much credence to the “Suggested Explanation” column’s attempt to answer “why” for each insight. For each one, there are also other plausible explanations, and, in most cases, only intuition rather than scientific evidence behind the particular answer provided. This issue is explored in the next section immediately after these tables of “Bizarre and Surprising Insights.”

Bizarre and Surprising Insights—Consumer Behavior (continued)

Insight	Organization	Suggested Explanation
Dolls and candy bars. Sixty percent of customers who buy a Barbie doll buy one of three types of candy bars.	Walmart	<i>Kids come along for errands.</i>
Pop-Tarts before a hurricane. Prehurricane, Strawberry Pop-Tart sales increased about sevenfold.	Walmart	<i>In preparation before an act of nature, people stock up on comfort or nonperishable foods.</i>
Staplers reveal hires. The purchase of a stapler often accompanies the purchase of paper, waste baskets, scissors, paper clips, folders, and so on.	A large retailer	<i>Stapler purchases are often a part of a complete office kit for a new employee.</i>
Higher crime, more Uber rides. In San Francisco, the areas with the most prostitution, alcohol, theft, and burglary are most positively correlated with Uber trips.	Uber	<i>"We hypothesized that crime should be a proxy for nonresidential population. . . . Uber riders are not causing more crime. Right, guys?"</i>
Mac users book more expensive hotels. Orbitz users on an Apple Mac spend up to 30 percent more than Windows users when booking a hotel reservation. Orbitz applies this insight, altering displayed options according to your operating system.	Orbitz	<i>Macs are often more expensive than Windows computers, so Mac users may on average have greater financial resources.</i>
Your inclination to buy varies by time of day. For retail websites, the peak is 8:00 PM; for dating, late at night; for finance, around 1:00 PM; for	Survey of websites	<i>The impetus to complete certain kinds of transactions is higher during certain times of day.</i>

Bizarre and Surprising Insights—Consumer Behavior (*continued*)

Insight	Organization	Suggested Explanation
travel, just after 10:00 AM. This is not the amount of website traffic, but the propensity to buy of those who are already on the website.		
Your e-mail address reveals your level of commitment. Customers who register for a free account with an Earthlink.com e-mail address are almost five times more likely to convert to a paid, premium-level membership than those with a Hotmail.com e-mail address.	An online dating website	<i>Disclosing permanent or primary e-mail accounts reveals a longer-term intention.</i>
Banner ads affect you more than you think. Although you may feel you've learned to ignore them, people who see a merchant's banner ad are 61 percent more likely to subsequently perform a related search, and this drives a 249 percent increase in clicks on the merchant's paid textual ads in the search results.	Yahoo!	<i>Advertising exerts a subconscious effect.</i>
Companies win by not prompting customers to think. Contacting actively engaged customers can backfire—direct mailing financial service customers who have already opened several	U.S. Bank	<i>Customers who have already accumulated many credit accounts are susceptible to impulse buys (e.g., when they walk into a bank branch) but, when contacted at home, will respond by</i>

(continued)

Bizarre and Surprising Insights—Consumer Behavior (*continued*)

Insight	Organization	Suggested Explanation
accounts decreases the chances they will open more accounts (<i>more details in Chapter 7</i>).		<i>considering the decision and possibly researching competing products online. They would have been more likely to make the purchase if left to their own devices.</i>
Your Web browsing reveals your intentions. Wireless customers who check online when their contract period ends are more likely to defect to a competing cell phone company.	A major North American wireless carrier	<i>Adverse to early termination fees, those intending to switch carriers remind themselves when they'll be free to change over.</i>
Friends stick to the same cell phone company (a social effect). If you switch wireless carriers, your contacts are in turn up to seven times more likely to follow suit.	A major North American wireless carrier; Optus (Australian telecom) saw a similar effect.	<i>People experience social influence and/or heed financial incentives for in-network calling.</i>

Bizarre and Surprising Insights—Finance and Insurance

Insight	Organization	Suggested Explanation
Low credit rating, more car accidents. If your credit score is higher, car insurance companies will lower your premium, since you are a lower driving risk. People with poor credit ratings are charged more for car	Automobile insurers	<i>"Research indicates that people who manage their personal finances responsibly tend to manage other important aspects of their life with that same level of responsibility, and that would include being responsible behind the wheel of their car," Donald</i>

Bizarre and Surprising Insights—Finance and Insurance (*continued*)

Insight	Organization	Suggested Explanation
insurance. In fact, a low credit score can increase your premium more than an at-fault car accident; missing two payments can as much as double your premium.		<i>Hanson of the National Association of Independent Insurers theorizes.</i>
Your shopping habits foretell your reliability as a debtor. If you use your credit card at a drinking establishment, you're a greater risk to miss credit card payments; at the dentist, lower risk; buy cheap, generic rather than name-brand automotive oil, greater risk; buy felt pads that affix to chair legs to protect the floor, lower risk.	Canadian Tire (a major retail and financial services company)	<i>More cautionary activity such as seeing the dentist reflects a more conservative or well-planned lifestyle.</i>
Typing with proper capitalization indicates creditworthiness. Online loan applicants who complete the application form with the correct case are more dependable debtors. Those who complete the form with all lower-case	A financial services startup company	<i>Adherence to grammatical rules reflects a general propensity to correctly comply.</i>

(continued)

Bizarre and Surprising Insights—Finance and Insurance (*continued*)

Insight	Organization	Suggested Explanation
letters are slightly less reliable payers; all capitals reveals even less reliability.		
Small businesses' credit risk depends on the owner's behavior as a consumer. Unlike business loans in general, when it comes to a small business, consumer-level data about the owner is more predictive of credit risk performance than business- level data (and combining both data sources is best of all).	Creditors to the leasing industry	<i>A small business's behavior largely reflects the choices and habits of one individual: the owner.</i>

Bizarre and Surprising Insights—Healthcare

Insight	Organization	Suggested Explanation
Genetics foretell cheating wives. Within a certain genetic cluster, having more genes shared by a heterosexual couple means more infidelity by the female.	University of New Mexico	<i>We're programmed to avoid inbreeding, since there are benefits to genetic diversity.</i>
Early retirement means earlier death. For a certain working category of males in Austria, each additional year of early	University of Zurich	<i>Unhealthy habits such as smoking and drinking follow retirement. Voltaire said, "Work spares us from three evils: boredom, vice, and</i>

Bizarre and Surprising Insights—Healthcare (continued)

Insight	Organization	Suggested Explanation
retirement decreases life expectancy by 1.8 months.		<i>need.” Malcolm Forbes said, “Retirement kills more people than hard work ever did.”</i>
Men who skip breakfast get more coronary heart disease. American men 45 to 82 who skip breakfast showed a 27 percent higher risk of coronary heart disease over a 16-year period.	Harvard University medical researchers	<i>Besides direct health effects—if any—eating breakfast may be a proxy for lifestyle: People who skip breakfast may lead more stressful lives and “were more likely to be smokers, to work full time, to be unmarried, to be less physically active, and to drink more alcohol.”</i>
Google search trends predict disease outbreaks. Certain searches for flu-related information provide insight into current trends in the spread of the influenza virus.	Google Flu Trends	<i>People with symptoms or in the vicinity of others with symptoms seek further information.</i>
Smokers suffer less from repetitive motion disorder. In certain work environments, people who smoke cigarettes are less likely to develop carpal tunnel syndrome.	A major metropolitan newspaper, conducting research on its own staff’s health	<i>Smokers take more breaks.</i>
Positive health habits are contagious (a social effect). If you quit smoking, your close contacts become 36 percent less likely to smoke. Your chance of	Research institutions	<i>People are strongly influenced by their social environment.</i>

(continued)

Bizarre and Surprising Insights—Healthcare (*continued*)

Insight	Organization	Suggested Explanation
becoming obese increases by 57 percent if you have a friend who becomes obese.		
Happiness is contagious (a social effect). Each additional Facebook friend who is happy increases your chances of being happy by roughly 9 percent.	Harvard University	<i>“Waves of happiness . . . spread throughout the network.”</i>
Knee surgery choices make a big difference. After ACL-reconstruction knee surgery, walking on knees was rated “difficult or impossible” by twice as many patients who donated their own patellar tissue as a graft source rather than hamstring tissue.	Medical research institutions in Sweden	<i>The patellar ligament runs across your kneecap, so grafting from it causes injury in that location.</i>
Music expedites poststroke recovery and improves mood. Stroke patients who listen to music for a couple of hours a day more greatly improve their verbal memory and attention span and improve their mood, as measured by a psychological test.	Cognitive Brain Research Unit, Department of Psychology, University of Helsinki, and Helsinki Brain Research Centre, Finland	<i>“Music listening activates a widespread bilateral network of brain regions related to attention, semantic processing, memory, motor functions, and emotional processing.”</i>
Yoga improves your mood. Long-term yoga practitioners showed benefits in a psychological test for mood in comparison to nonyoga practitioners, including a higher “vigor” score.	Research institutions in Japan	<i>Yoga is designed for, and practiced with the intent for, the attainment of tranquility.</i>

Bizarre and Surprising Insights—Crime and Law Enforcement

Insight	Organization	Suggested Explanation
Suicide bombers do not buy life insurance. An analysis of bank data of suspected terrorists revealed a propensity to not hold a life insurance policy.	A large UK bank	<i>Suicide nullifies a life insurance policy.</i>
Unlike lightning, crime strikes twice. Crime is more likely to repeat nearby, spreading like earthquake aftershocks.	Departments of math, computer science, statistics, criminology, and law in California universities	<i>Perpetrators “repeatedly attack clusters of nearby targets because local vulnerabilities are well-known to the offenders.”</i>
Crime rises with public sporting events. College football upset losses correspond to a 112 percent increase in assaults.	University of Colorado	<i>Psychological theories of fan aggression are offered.</i>
Crime rises after elections. In India, crime is lower during an election year and rises soon after elections.	Researchers in India	<i>Incumbent politicians crack down on crime more forcefully when running for reelection.</i>
Phone card sales predict danger in the Congo. Impending massacres in the Congo are presaged by spikes in the sale of prepaid phone cards.	CellTel (African telecom)	<i>Prepaid cards denominated in U.S. dollars serve as in-pocket security against inflation for people “sensing impending chaos.”</i>
Hungry judges rule negatively. Judicial parole decisions	Columbia University and Ben Gurion University (Israel)	<i>Hunger and/or fatigue leave decision makers feeling less forgiving.</i>

(continued)

Bizarre and Surprising Insights—Crime and Law Enforcement

(continued)

Insight	Organization	Suggested Explanation
immediately after a food break are about 65 percent favorable, which then drops gradually to almost zero percent before the next break. If the judges are hungry, you are more likely to stay in prison.		

Bizarre and Surprising Insights—Miscellaneous

Insight	Organization	Suggested Explanation
Music taste predicts political affiliation. Kenny Chesney and George Strait fans are most likely conservative, Rihanna and Jay-Z fans liberal. Republicans can be more accurately predicted by music preferences than Democrats because they display slightly less diversity in music taste. Metal fans can go either way, spanning the political spectrum.	The Echo Nest (a music data company)	<i>Personality types entail certain predilections in both musical and political preferences (this is the author's hypothesis; the researchers do not offer a hypothesis).</i>
Online dating: Be cool and unreligious to succeed. Online dating messages that initiate first contact and include the word <i>awesome</i> are more than twice as likely to elicit a	OkCupid (online dating website)	<i>There is value in avoiding the overused or trite; video games are not a strong aphrodisiac.</i>

Bizarre and Surprising Insights—Miscellaneous (continued)

Insight	Organization	Suggested Explanation
<p>response as those with <i>sexy</i>. Messages with “your pretty” get fewer responses than those with “you’re pretty.” “Howdy” is better than “Hey.” “Band” does better than “literature” and “video games.” “Atheist” far surpasses most major religions, but “Zeus” is even better.</p>	<p>Hot or not? People consistently considered attractive get less attention. Online daters rated with a higher variance of attractiveness ratings receive more messages than others with the same average rating but less variance. A greater range of opinions—more disagreement on looks—results in receiving more contact.</p>	<p><i>People often feel they don’t have a chance with someone who appears universally attractive. When less competition is expected, there is more incentive to initiate contact.</i></p>
<p>Users of the Chrome and Firefox browsers make better employees. Among hourly employees engaged in front-line service and sales-based positions, those who use these two custom Web browsers perform better on employment assessment metrics and stay on longer.</p>	<p>A human resources professional services firm, over employee data from Xerox and other firms</p>	<p><i>“The fact that you took the time to install [another browser] shows . . . that you are an informed consumer . . . that you care about your productivity and made an active choice.”</i></p>

(continued)

Bizarre and Surprising Insights—Miscellaneous (continued)

Insight	Organization	Suggested Explanation
A job promotion can lead to quitting. In one division of HP, promotions increase the risk an employee will leave unless accompanied by sufficient increases in compensation; promotions without raises hurt more than help.	Hewlett-Packard	<i>Increased responsibilities are perceived as burdensome if not financially rewarded.</i>
More engaged employees have fewer accidents. Among oil refinery workers, a one percentage-point increase in team employee engagement is associated with a 4 percent decrease in the number of safety incidents per employee.	Shell	<i>More engaged workers are more attentive and focused.</i>
Higher status, less polite. Editors on Wikipedia who exhibit politeness are more likely to be elected to “administrative” status that grants greater operational authority. However, once elected, an editor’s politeness decreases.	Researchers examining Wikipedia behavior	<i>“Politeness theory predicts a negative correlation between politeness and the power of the requester.”</i>
Vegetarians miss fewer flights. An airline		
Airline customers who preorder a vegetarian meal are more likely to make their flight.		<i>The knowledge of a personalized or specific meal awaiting the customer provides an incentive or establishes a sense of commitment.</i>
Smart people like curly fries. Liking “Curly Fries” on	Researchers at the University of	<i>An intelligent person was the first to like this Facebook</i>

Bizarre and Surprising Insights—Miscellaneous (continued)

Insight	Organization	Suggested Explanation
Facebook is predictive of high intelligence.	Cambridge and Microsoft Research	<i>page, “and his friends saw it, and by homophily, we know that he probably had smart friends, and so it spread to them . . . ,” and so on.</i>
A photo’s quality is predictable from its caption. Even without looking at the picture itself, key words from its caption foretell whether a human would subjectively rate the photo as “good.” The words <i>Peru, tombs, trails, and boats</i> corresponded with better photos, whereas the words <i>graduation</i> and <i>CEO</i> tend to appear with lower-quality photos.	(Not available)	<i>Certain events and locations are conducive to or provide incentive for capturing more picturesque photos.</i>
Female-named hurricanes are more deadly. Based on a study of the most damaging hurricanes in the United States during six recent decades, the ones with “relatively feminine” names killed an average of 42 people, almost three times the 15 killed by hurricanes with “relatively male” names.	University researchers	<i>This may result from “a hazardous form of implicit sexism.” Psychological experiments in a related study “suggested that this is because feminine- versus masculine-named hurricanes are perceived as less risky and thus motivate less preparedness. . . . Individuals systematically underestimate their vulnerability to hurricanes with more feminine names.”</i>

(continued)

Bizarre and Surprising Insights—Miscellaneous (*continued*)

Insight	Organization	Suggested Explanation
Men on the <i>Titanic</i> faced much greater risk than women. A woman on the <i>Titanic</i> was almost four times as likely to survive as a man. Most men died and most women lived.	Miscellaneous researchers	<i>Priority for access to lifeboats was given to women.</i>
Solo rockers die younger than those in bands. Although all rock stars face higher risk, solo rock stars suffer twice the risk of early death as rock band members.	Public health offices in the UK	<i>Band members benefit from peer support, and solo artists exhibit even riskier behavior.</i>

CAVEAT #1: CORRELATION DOES NOT IMPLY CAUSATION

Satisfaction came in the chain reaction.

—From the song “Disco Inferno,” by The Trammps

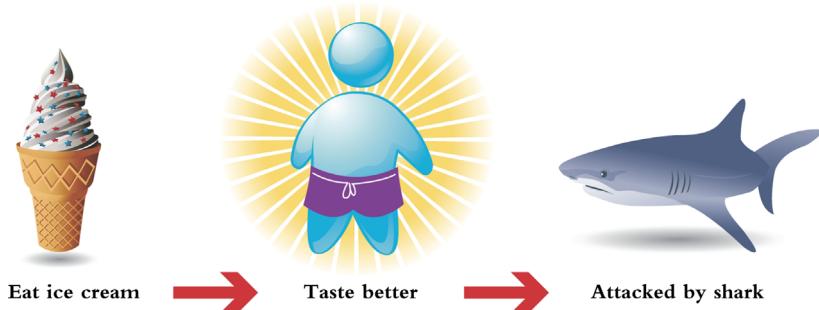
The preceding tables, packed with fun-filled facts, do not explain a single thing.

Take note, the third column is headed “Suggested Explanation.” While the left column’s discoveries are validated by data, the reasons behind them are unknown. Every explanation put forth, each entry in the rightmost column, is pure conjecture with absolutely no hard facts to back it up.

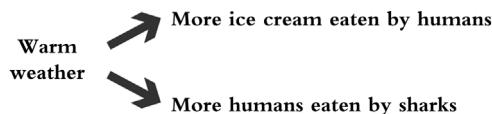
The dilemma is, as it is often said, *correlation does not imply causation.*⁸ The discovery of a predictive relationship between A and B does not mean one causes the other, not even indirectly. No way, no how.

⁸ The Latin phrase *Post hoc, ergo propter hoc* (“After this, therefore because of this”) is another common expression that references the issue at hand; it refers to the unwarranted act of concluding a causal relationship.

Consider this: Increased ice cream sales correspond with increased shark attacks. Why do you think that is? A causal explanation could be that eating ice cream makes us taste better to sharks:



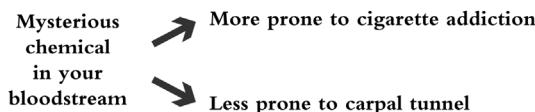
But another explanation is that, rather than one being caused by the other, they are both caused by the same thing. On cold days, people eat less ice cream and also swim less; on warm days, they do the opposite:



Take the example of smokers getting less carpal tunnel syndrome, from the table of healthcare examples. One explanation is that smokers take more breaks:



But another could be that there's some mysterious chemical in your bloodstream that influences both things:



I totally made that up. But the truth is that finding the connection between smoking and carpal tunnel syndrome in and of itself provides no evidence that one explanation is more likely than the other. With this in mind, take another look through the tables. The same rule applies to each example. We know the *what*, but we don't know the *why*.

When applying PA, we generally don't have firm knowledge about causation, and we often don't necessarily care. For many PA projects, the value comes from prediction, with only an avocational interest in understanding the world and figuring out what makes it tick.

Causality is elusive, tough to nail down. We naturally assume things do influence one another in some way, and we conceive of these effects in physical, chemical, medical, financial, or psychological terms. The noble scientists in these fields have their work cut out for them as they work to establish and characterize causal links.

In this way, data scientists have it easier with PA. It just needs to work; prediction trumps explanation. PA operates with extreme solution-oriented intent. The whole point, the “ka-ching” of value, comes in driving decisions from many individual predictions, one per patient, customer, or person of any kind. And while PA often delivers meaningful insights akin to those of various social sciences, this is usually a side effect, not the primary objective.

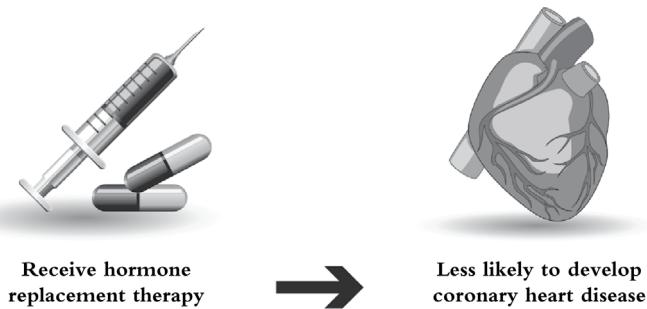
This makes PA a kind of “metascience” that transcends the taxonomy of natural and social sciences, abstracting across them by learning from any and all data sources that would typically serve biology, criminology, economics, education, epidemiology, medicine, political science, psychology, or sociology. PA’s mission is to engineer solutions. As for the data employed and the insights gained, the tactic in play is: “Whatever works.”

And yet even hard-nosed scientists fight the urge to overexplain. It's human nature, but it's dangerous. It's the difference between good science and bad science.

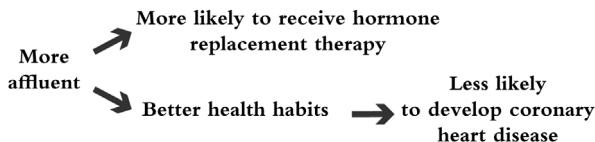
Stein Kretsinger, founding executive of [Advertising.com](#) and a director at Elder Research, tells a classic story of our overly interpretive minds. In the early 1990s, as a graduate student, Stein was leading a medical research meeting, assessing the factors that determine how long it takes to wean off a

respirator. As this was before the advent of PowerPoint projection, Stein displayed the factors, one at a time, via graphs on overhead transparencies. The team of healthcare experts nodded their heads, offering one explanation after another for the relationships shown in the data. After going through a few, though, Stein realized he'd been placing the transparencies with the wrong side up, thus projecting mirror images that depicted the *opposite* of the true relationships. After he flipped them to the correct side, the experts seemed just as comfortable as before, offering new explanations for what was now the very opposite effect of each factor. Our thinking is malleable—people readily find underlying theories to explain just about anything.

In another case, a published medical study discovered that women who happened to be receiving hormone replacement therapy showed a lower incidence of coronary heart disease. Could it be that a new treatment for this disease had been discovered?



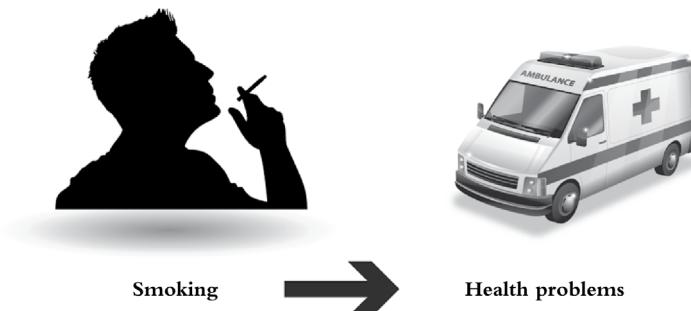
Later, a proper control experiment disproved this false conclusion. Instead, the currently held explanation is that more affluent women had access to the hormone replacement therapy, and these same women had better health habits overall:



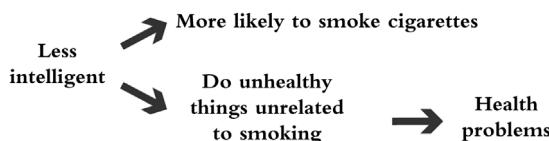
Prematurely jumping to conclusions about causality is bad science that leads to bad medical treatment. This kind of research snafu is not an isolated case.

According to *The Wall Street Journal*, the number of retracted journal publications has surged in recent years.

But, in this arena, the line between apt and inept sometimes blurs. Twenty years ago, while in graduate school, I befriended a colleague, a chain smoker who was nevertheless brilliant with the mathematics behind probability theory. He would hang you out to dry if you attempted to criticize his bad smoking habit on the basis of clinical studies. “Smoking studies have no control group,” he’d snap.⁹ He was questioning the common causal conclusion:



One day in front of the computer science building, as I kept my distance from his cloud of smoke, he drove this point home. New to the study of probability, I suddenly realized what he was saying and, looking at him incredulously, asked, “You mean to say that it’s possible smoking studies actually reflect that stupid people smoke, and that these people also do other stupid things, and only those other things poorly affect their health?” By this logic, I had been stupid for not considering him quite possibly both stupid and healthy.



⁹ This is because it’s not reasonable to instruct one clinical group to smoke, nor to expect another to uniformly resist smoking. To statistically prove in this way that something kills, you would need to kill some people.

He exhaled a lungful of smoke triumphantly as if he'd won the argument and said with no irony, "Yes!" The same position had also been espoused in the 1950s by an early founder of modern statistics, Ronald Fisher. He was a pipe-smoking curmudgeon who attacked the government-supported publicity about tobacco risks, calling it egregious fearmongering.

In addressing the effects of tobacco, renowned healthcare statistician David Salsburg wrote that the very meaning of cause and effect is "a deep philosophical problem . . . that gnaws at the heart of scientific thought." Due in part to our understanding of how inhaled agents actively lead to genetic mutations that create cancerous cells, the scientific community has concluded that cigarettes are causal in their connection to cancer. While I implore scientists not to overinterpret results, I also implore you not to smoke.

CAVEAT #2: SECURING SOUND DISCOVERIES

The trouble with the world is that the stupid are cocksure and the intelligent are full of doubt.

—Bertrand Russell

Even before suggesting any causal explanation for a correlation observed in data, you had better verify it's actually a real trend rather than misleading noise.

At the beginning of this chapter, we saw that data can lead us astray, tempting us—and several mass media outlets—to believe orange cars last longer. In that data, used cars sporting this flashy color turned out to be lemons 33 percent less often. However, subsequent analysis has severely weakened the confidence in this discovery, relegating it to inconclusive. What went wrong?

Warning! Big data brings big potential—but also big danger. With more data, a unique pitfall often dupes even the brightest of data scientists. This hidden hazard can undermine the process that evaluates for statistical significance, the gold standard of scientific soundness. And what a hazard it is! A bogus discovery can spell disaster. You may buy an orange car—or

undergo an ineffective medical procedure—for no good reason. As the aphorisms tell us, bad information is worse than no information at all; misplaced confidence is seldom found again.

This peril seems paradoxical. If data's so valuable, why should we suffer from obtaining more and more of it? Statistics has long advised that having more examples is better. A longer list of cases provides the means to more scrupulously assess a trend. Can you imagine what the downside of more data might be? As you'll see in a moment, it's a thought-provoking, dramatic plot twist.

The fate of science—and sleeping well at night—depends on deterring the danger. The very notion of empirical discovery is at stake. To leverage the extraordinary opportunity of today's data explosion, we need a surefire way to determine whether an observed trend is real, rather than a random artifact of the data.

Statistics approaches this challenge in a very particular way. It tells us the chances the observed trend could randomly appear even if the effect were not real. That is, it answers this question:¹⁰

Question that statistics can answer: *If orange cars were actually no more reliable than used cars in general, what would be the probability that this strong a trend—depicting orange cars as more reliable—would show in data anyway, just by random chance?*

With any discovery in data, there's always some possibility we've been *Fooled by Randomness*, as Nassim Taleb titled his compelling book. The

¹⁰ Mini statistics lesson: The notion of the trend being untrue—the notion that orange cars have no advantage—is called the *null hypothesis*. And the probability the observed effect would occur in data if the null hypothesis were true (i.e., the answer to the question above) is called the *p-value*, a foundational concept brought to popularity in the 1920s by the same Ronald Fisher who criticized anti-tobacco propaganda in the 1950s. If the p-value is low enough—e.g., below 1 percent or 5 percent—then a researcher will typically reject the null hypothesis as too unlikely, and view this as support for the discovery, which is thereby considered *statistically significant*. This evaluation process is standard practice for executing on the scientific method itself.

book reveals the dangerous tendency people have to subscribe to unfounded explanations for their own successes and failures, rather than correctly attributing many happenings to sheer randomness. The scientific antidote to this failing is probability, which Taleb affectionately dubs “a branch of applied skepticism.”

Statistics is the resource we rely on to gauge probability. It answers the orange car question above by calculating the probability that what’s been observed in data would occur randomly if orange cars actually held no advantage. The calculation takes data size into account—in this case, there were 72,983 used cars varying across 15 colors, of which 415 were orange.¹¹

Calculated answer to the question: 0.68 percent

Looks like a safe bet. Common practice considers this risk acceptably remote, low enough to at least tentatively believe the data. But don’t buy an orange car just yet—or write about the finding in a newspaper for that matter.

WHAT WENT WRONG: ACCUMULATING RISK

In China when you’re one in a million, there are 1,300 people just like you.

—Bill Gates

So if there had only been a 1 percent long shot that we’d be misled by randomness, what went wrong?

The experimenters’ mistake was to not account for running many small risks, which had added up to one big one. In addition to checking whether being orange is predictive of car reliability, they also checked each of the other 14 colors, as well as the make, model, year, trim level, type of transmission, size, and more. For each of these factors, they repeatedly ran the risk of being fooled by randomness.

¹¹ The applicable statistical method is a *1-sided equality of proportions hypothesis test*, which produced a p-value under 0.0068. The p-value is the estimated chance we would have ended up with this data if in fact the observed effect were not real; that is, if being colored orange had no correlation with whether the car is a good or bad buy.

Probability is relative, affected entirely by context. With additional background information, a seemingly unlikely event turns out to be not so special after all. Imagine your friend calls to tell you, “I won the jackpot at hundred-to-one odds!” You might get a little excited. “Wow!”

Now imagine your friend adds, “By the way, I’m only talking about one of 70 times that I spun the jackpot wheel.” The occurrence that had at first seemed special suddenly has a new context, positioned alongside a number of less remarkable episodes. Instead of exclaiming wow, you might instead do some arithmetic. The probability of losing a spin is 99 percent. If you spin twice, the chances of losing both is 99 percent \times 99 percent, which is about 98 percent. Although you’ll probably lose both spins, why stop at two? The more times you spin, the lower the chances of never winning once. To figure out the probability of losing 70 times in a row, multiple 99 percent times itself 70 times, aka 0.99 raised to the power of 70. That comes to just under 0.5. Let your friend know that nothing special happened—the odds of winning at least once were about 50/50.

Special cases aren’t so special after all. By the same sort of reasoning, we might be skeptical about the merits of the famed and fortuned. Do the most successful elite hold talents as elevated as their singular status? As Taleb put it in *Fooled by Randomness*, “I am not saying that Warren Buffett is not skilled; only that a large population of random investors will almost necessarily produce someone with his track records just by luck.”

Play enough and you’ll eventually win. Likewise, press your luck repeatedly and you’ll eventually lose. Imagine your same well-intentioned friend calls to tell you, “I discovered that orange cars are more reliable, and the stats say there’s only a 1 percent chance this phenomenon would appear in the data if it weren’t true.” You might get a little impressed. “Interesting discovery!”

Now imagine your friend adds, “By the way, I’m only talking about one among dozens of car factors—my computer program systematically went through and checked each one.” Both of your friend’s stories enthusiastically led with a “remarkable” event—a jackpot win or a predictive discovery. But the numerous other less remarkable attempts—that often go unmentioned—are just as pertinent to each story’s conclusion.

Wake up and smell the probability. Imagine we test 70 characteristics of cars that in reality are not predictive of lemons. But each test suffers a, say, 1 percent risk the data will falsely show a predictive effect just by random chance. The accumulated risk piles up. As with the jackpot wheel, there's a 50/50 chance the unlikely event will eventually take place—that you will stumble upon a random perturbation that, considered in isolation, is compelling enough to mislead.

THE POTENTIAL AND DANGER OF AUTOMATING SCIENCE: VAST SEARCH

The most exciting phrase to hear in science, the one that heralds new discoveries, is not “Eureka!” but rather “Hmm . . . that’s funny . . .”

—Isaac Asimov

A tremendous potential inspires us to face this peril: Predictive modeling automates scientific discovery. Although it may seem like an obvious thing to do in this computer age, trying out each predictor variable is a dramatic departure from the classic scientific method of developing a single hypothesis and then testing it. Your computer essentially acts as hundreds or even thousands of scientists by conducting a broad, exploratory analysis, automatically evaluating an entire batch of predictors. This aggressive hunt for any novel source of predictive information leaves no stone unturned. The process is key to uncovering valuable, unforeseen insights.

Automating this search for valuable predictors empowers science, lessening its dependence on ever-elusive serendipity. Instead of waiting to inadvertently stumble upon revelations or racking our brains for hypotheses, we rely less on luck and hunches by systematically testing many factors. While necessity is the mother of invention, historically speaking, serendipity has long been its daddy. It was only by happy accident that Alexander Fleming happened upon the potent effects of penicillin, by noticing that an old bacteria culture he was about to clean up happened to be contaminated with some mold—which was successfully killing it. Likewise, Minoxidil was

inadvertently discovered as a baldness remedy in an unexpected, quizzical moment: “Look, more hair!”

But as exciting a proposition as it is, this automation of data exploration builds up an overall risk of eventually being fooled—at one time or another—by randomness. This inflation of risk comes as a consequence of assessing many characteristics of used cars, for example. The power of automatically testing a batch of predictors may serve us well, but it also exposes us to the very real risk of bogus discoveries.

Let’s call this issue *vast search*—the term that industry leader (and Chapter 1’s predictive investor) John Elder coined for this form of automated exploration and its associated peril. Repeatedly identified anew across industries and fields of science, this issue is also called the *multiple comparisons problem* or *multiple comparisons trap*. John warns, “The problem is so widespread that it is the chief reason for a crisis in experimental science, where most journal results have been discovered to resist replication; that is, to be wrong!”

Statistics darling Nate Silver jumped straight to the issue of vast search when asked generally about the topic of big data on *Freakonomics Radio*. With a lot of data, he said, “you’re going to find lots and lots of correlations through brute force . . . but the problem is that a high percentage of those, maybe the vast majority, are false correlations, are false positives. . . . They [appear] statistically significant, but you have so many lottery tickets when you can run an analysis on a [large data set] that you’re going to have some one-in-a-million coincidences just by chance alone.”

The casual “mining” of data—analysis of one sort or another to find interesting tidbits and insights—often involves vast search, making it all too easy to dig up a false claim. With this misstep so commonplace, there’s a real possibility that some of the predictive discoveries listed in the tables earlier in this chapter could face debunking, depending on whether the researchers have taken proper care. As we’ll see in the next chapter, one mischievous professor illustrated the problem of searching and “re-searching” too far and wide when he unearthed a cockamamie relationship between dairy products in Bangladesh and the U.S. stock market.

Bigger data isn’t the problem—more specifically, it’s *wider* data. When prepared for PA, data grows in two dimensions—it’s a table:

**One column per predictor variable:
wider means more predictors—vaster search**

The diagram shows a rectangular data table with a double-headed horizontal arrow above it labeled "One column per predictor variable: wider means more predictors—vaster search". To the right of the table is a double-headed vertical arrow labeled "One row per training case: longer means more cases—better".

DATA						
Dodge	Neon	2004	compact	silver	OK	
Mitsubishi	Galant	2004	medium	white	OK	
Mercury	Sable	2004	medium	white	BAD	
Ford	Focus	2005	compact	silver	OK	
Kia	Spectra	2004	medium	black	OK	
Dodge	Caravan	2005	van	red	BAD	
Ford	Explorer	2002	medium	blue	BAD	
Chrysler	Pacifica	2004	crossover	silver	OK	
Pontiac	Vibe	2004	medium	orange	OK	
...						

**One row
per training case:
longer means more
cases—better**

**A small sample of data for predicting bad buys among used cars.
The complete data is both wider and longer.**

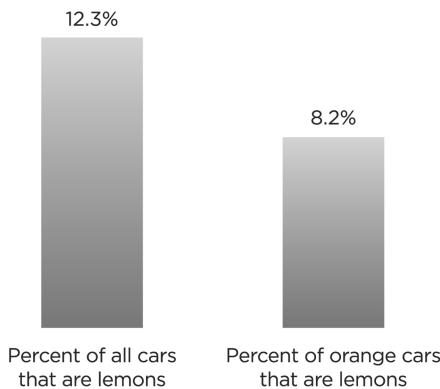
As you accrue more examples of cars, people, or whatever you're predicting, the table grows longer (more rows, aka *training cases*). That's always a good thing. The more training cases to analyze, the more statistically sound.¹² Expanding in the other dimension, each row widens (more columns) as more factors—aka predictor variables—are accrued. A certain factor such as car color may only amount to a single column in the data, but since we look at each possible color individually, it has the virtual effect of adding 15 columns to the width, one per color. Overall, the sample data in the figure above is not nearly as wide as data often gets, but even in this case the vast

¹² This only holds true under the assumption you have a representative sample, e.g., an unbiased, random selection of cases.

search effect is at play. With wider and wider data, we can only tap the potential if we can avoid the booby trap set by vast search.

A FAILSAFE FOR SOUND RESULTS

To understand what sort of failsafe mechanism we need, let's revisit the misleading "orange lemons" discovery.



This 12.3-versus-8.2 result is calculated from four numbers:

There were 72,983 cars, of which 8,976 were lemons.

There were 415 orange cars, of which 34 were lemons.

The standard method—the one that misled researchers as well as the press—evaluates for statistical significance based *only* on those four numbers. When fed these as input, the test provides a positive result, calculating there was only a 0.68 percent chance we would witness that extreme a difference in orange cars if they were in actuality no more prone to be lemons than cars of other colors.

But these four numbers alone do not tell the whole story—the *context* of the discovery also matters. How vast was the search for such discoveries? How many other factors were also checked for a correlation with whether a car is a lemon?

In other words, if a data scientist hands you these four numbers as "proof" of a discovery, you should ask what it took to find it. Inquire, "How many other things did you also try that came up dry?"

With the breadth of search taken into account, the “orange lemon” discovery collapses. Confidence diminishes and it shows as inconclusive. Even if we assume the other 14 colors were the only other factors examined, statistical methods estimate a much less impressive 7.2 percent probability of stumbling by chance alone upon a bogus finding that appears this compelling.¹³ Although 7.2 percent is lower odds than a coin toss, it’s no long shot; by common standards, this is not a publishable result. Moreover, 7.2 is an optimistic estimate. We can assume the risk was even higher than that (i.e., worse) since other factors such as car make, model, and year were also available, rendering the search even wider and the opportunities to be duped even more plentiful.

Inconclusive results are no results at all. It may still be true that orange cars are less likely to be lemons, but the likelihood this appeared in the data by chance alone is too high to put a lot of faith in it. In other words, there’s not enough evidence to rigorously support the hypothesis. It is, at least for now, relegated to “a fascinating possibility,” only provisionally distinct from any untested theories one might think up.

Want conclusive results? Then get *longer* data, i.e., more rows of examples. Adequately rigorous failsafe methods that account for the breadth of search set a higher bar. They serve as a more scrupulous filter to eliminate inconclusive findings before they get applied or published. To compensate for this strictness and increase the opportunity to nonetheless attain conclusive results, the best recourse is elongating the list of cases. If the search is vast—that is, if the data is wide—then findings will need to be more compelling in order to pass through the filter. To that end, if there are ample examples with which to confirm findings—in other words, if the data makes up for its width by also being longer—then legitimate findings will have the empirical support they need to be validated.

¹³ This probability was estimated with a method called *target shuffling*, which does take the vastness of search into account. For details, see “Are Orange Cars Really not Lemons?” by John Elder and Ben Bullard of Elder Research, Inc. (elderresearch.com/orange-car)

The Data Effect will prevail so long as there are enough training examples to correctly discern which predictive discoveries are authentic.

A PREVALENT MISTAKE

Despite the seriousness of this mistake, the vast search pitfall regularly trips up even the most well-intentioned data scientists, statisticians, and other researchers. A perfect storm of influences leads to its prevalence:

- **It's elusive.** You have to think outside a certain box. The classic application of statistical methods has traditionally focused on evaluating for significance based entirely on the result itself. There's a conceptual leap in moving beyond that to also account for the breadth of search, the full suite of other predictors also considered.
- **It's new.** Since the advent of big data—to be specific, wide data—has more recently intensified this problem, awareness across the data science community still needs to catch up.
- **Simplicity can deceive.** Ironically, although bite-sized anecdotes are more likely to make compelling headlines and draw public attention, they're less likely to be properly screened against failure. It's widely understood that a predictive model, whose job is to combine variables in order to fit the data, can go too far and *overfit*—a primary topic of the next chapter. Since single-variable insights—such as the “orange lemons” claim and the many examples listed earlier in this chapter's tables—are so much simpler than multivariate models, their potential to hold a spurious aberration is underestimated and so they're often subjected to less rigorous scrutiny.
- **Falsehoods don't look wrong.** Without realizing that a pattern may only in actuality be random noise, people creatively formulate compelling causal explanations. This is human nature, but on many occasions it only increases one's attachment to a false discovery.
- **It's a buzzkill.** Given the strong incentives to make predictive discoveries, the temptation is there to be less than scrupulous, either intentionally—

or, more commonly, with a certain convenient forgetfulness—neglecting to account for the full scope of the search that led to the discovery.

In the big data tsunami, you've got to either sharpen your skills or get out of the water.

PUTTING ALL THE PREDICTORS TOGETHER

There ought to be a rock band named after this chapter's explosive topic, "The Predictors."¹⁴

The number of predictors at our disposal grows along with an unbridled trend: Exploding quantities of increasingly diverse data are springing forth, and organizations are innovating to turn all this unprocessed sap into maple syrup.

The next step is a doozy. To fully leverage predictor variables, we must deftly and intricately combine them with a predictive model. To this end, you can't just stir the bowl with a big spoon. You need an apparatus that learns from the data itself how best to mix and combine it.

Holy combinatorial explosion, Batman! This will make the vast search problem worse—much worse. By combining two predictors, as in, "Are cars with the color black and the make Audi liable to be lemons?" for example—or even more than two—we will build up a much larger batch of relationships to evaluate. This also means a much greater number of opportunities to be fooled by randomness.

Concerned? Overwhelmed? What if I told you there's an intuitive, elegant method for building a predictive model, as well as a simple way to confirm a model's soundness, *without the need to mathematically account for the vastness of search?* The next chapter shows you how it's done—20 pages on how machine learning works, plus another 12 covering the most practical yet philosophically intriguing question of data science: How can we ensure that what the machine has learned is real, that the predictive model is sound?

¹⁴ I spoke too soon; there is one! They're in Australia. See www.thepredictors.com.au. I told you data rocks.



CHAPTER 4

The Machine That Learns

A Look inside Chase's Prediction of Mortgage Risk

What form of risk has the perfect disguise? How does prediction transform risk to opportunity? What should all businesses learn from insurance companies? Why does machine learning require art in addition to science? What kind of predictive model can be understood by everyone? How can we confidently trust a machine's predictions? Why couldn't prediction prevent the global financial crisis?

This is a love story about a man named Dan and a bank named Chase, and how they learned to persevere against all odds—more precisely, how they deployed *machine learning* to empower prediction, which in turn mitigates risk. Exploring this story, we'll uncover how machine learning really works under the hood.¹

BOY MEETS BANK

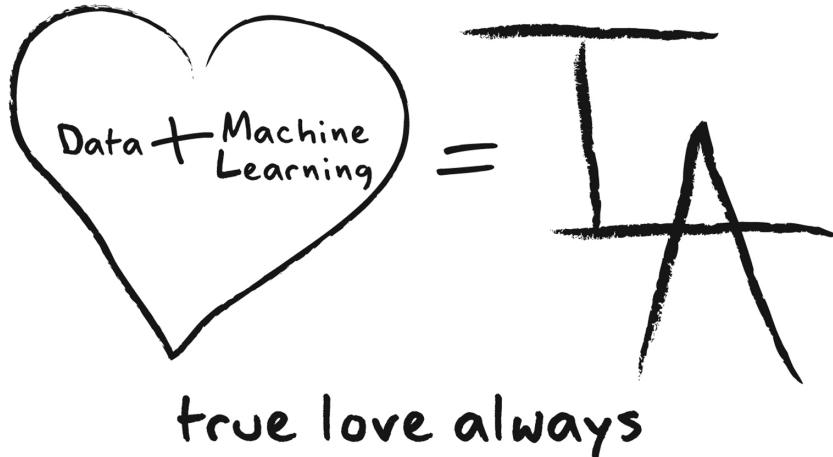
Once upon a time, a scientist named Dan Steinberg received a phone call because the largest U.S. bank faced new levels of risk. To manage risk, they were prepared to place their bets on this man of machine learning.

'Twas a fortuitous engagement, as Dan had just the right means and method to assist the bank. An entrepreneurial scientist, he had built a commercial predictive analytics (PA) system that delivered leading research

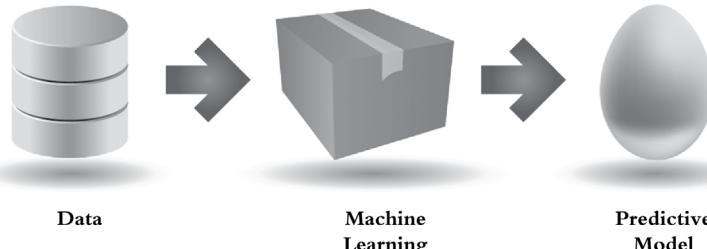
¹ Further technical details for the Chase case study are available in a 2005 conference presentation referenced in this chapter's Notes (www.PredictiveNotes.com).

from the lab into corporate hands. The bank held as dowry electronic plunder: endless rows of 1's and 0's that recorded its learning experience.

The bank had the fuel, and Dan had the machine. It was a match made in heaven. Daydreaming, I often doodle in the margins:



A more adult business professional might open his heart in a more formal way, depicting this as we did in a previous chapter:



Machine learning processes data to produce a predictive model.

BANK FACES RISK

For any organization, financial risk creeps up stealthily, hidden by a perfect, simple disguise: Major loss accumulates from many small losses, each of which alone seems innocuous. Minor individual misfortunes, boring and utterly undramatic, slip below the radar. They're practically invisible.

Soon after a megamerger in 1996 that rendered Chase Bank the nation's largest, the bank's home finance team recognized a new degree of risk. Their pool of mortgage holders had grown significantly. It was now composed of what had originally been six banks' worth of mortgages: millions of them. Each one represented a bit of risk—a *microrisk*. That's when Dan received the call.

Ironically, there are two seemingly opposite ways a mortgage customer can misbehave. They can fail to pay you back, or they can pay you back in full but too quickly:

Microrisk A: Customer defaults on the mortgage payments.

Microrisk B: Customer *prepays* the mortgage all at once, early, due to refinancing with a competing bank or selling the house. Prepayment is a loss because the bank fails to collect the mortgage's planned future interest payments.

These losses are demoted to “micro” because, for a bank, any one mortgage customer just isn’t that big a deal. But microlosses threaten to add up. In the financial world, the word *risk* most often refers to *credit risk*; that is, microrisk A, wherein an outstanding debt is never recovered and is lost forever. But when your bread and butter is interest payments, microrisk B is no picnic either. To put it plainly, your bank doesn’t want you to get out of debt.²

PREDICTION BATTLES RISK

Most discussions of decision making assume that only senior executives make decisions or that only senior executives' decisions matter. This is a dangerous mistake.

—Peter Drucker, an American educator and writer born in 1909

Chase’s mortgage portfolio faced risk factors amounting to hundreds of millions of dollars. Every day is like a day at the beach—each grain of sand is

² Similarly, credit card issuers aren’t pleased if you always pay in full and never pay interest.

one of a million microrisks. Once a mortgage application is stamped “low-risk” and approved, the risk management process has actually only just begun. The bank’s portfolio of open mortgages must be tended to like cows on a dairy farm. The reason? Risk lurks. Millions of mortgages await decisions as to which to sell to other banks, which to attempt to keep alive, and which to allow refinancing for at a lower interest rate.

PA serves as an antidote to the poisonous accumulation of microrisks. PA stands vigil, prospectively flagging each microrisk so the organization can do something about it.

It’s nothing new. The notion is mainstream and dates back to the very origin of PA. Predicting consumer risk is well known as the classic *credit score*, provided by FICO and credit bureaus such as Experian. The credit score’s origin dates back to 1941, and the term is now a part of the common vernacular. Its inception was foundational for PA, and its success has helped propel PA’s visibility. Modern-day risk scores are often built with the same predictive modeling methods that fuel PA projects.

The benefits of fighting risk with PA can be demonstrated with ease. While prediction itself may be an involved task, it only takes basic arithmetic to calculate the value realized once prediction is working. Imagine you run a bank with thousands of outstanding loans, 10 percent of which are expected to default. With one of every 10 debtors an impending delinquent account, the future drapes its usual haze: You just don’t know which will turn out to be bad.

Say you score each loan for risk with an effective predictive model. Some get high-risk scores and others low-risk scores. If these risk scores are assigned well, the top half predicted as most risky could see almost twice as many as average turn out to be defaulters—to be more realistic, let’s say 70 percent more than the overall default rate. That would be music to your ears. A smidgeon of arithmetic shows you’ve divided your portfolio into two halves, one with a 17 percent default rate (70 percent more than the overall 10 percent rate), and the other with a 3 percent default rate (since 17 and 3 average out to 10).

High-risk loans: 17 percent will default.

Low-risk loans: 3 percent will default.

You've just divided your business into two completely different worlds, one safe and one hazardous. You now know where to focus your attention.

Following this promise, Chase took a large-scale, calculated *macrorisk*. It put its faith in prediction, entrusting it to drive millions of dollars' worth of decisions. But Chase's story will earn its happy ending only if prediction works—if what's learned from data pans out in the great uncertainty that is the future.

Prediction presents the ultimate dilemma. Even with so much known of the past, how can we justify confidence in technology's vision of the unknowable future?

Before we get into how prediction works, here are a few words on risk.

RISKY BUSINESS

The revolutionary idea that defines the boundary between modern times and the past is the mastery of risk: the notion that the future is more than a whim of the gods and that men and women are not passive before nature. Until human beings discovered a way across that boundary, the future was a mirror of the past or the murky domain of oracles and soothsayers who held a monopoly over knowledge of anticipated events.

—Peter Bernstein, *Against the Gods: The Remarkable Story of Risk*

There's no such thing as bad risk, only bad pricing.

—Stephen Brobst, Chief Technology Officer, Teradata

Of course, banks don't bear the entire burden of managing society's risk. Insurance companies also play a central role. In fact, their core business is the act of data crunching to quantify risk so it can be efficiently distributed. Eric Webster, a vice president at State Farm Insurance, put it brilliantly: "Insurance is nothing but management of information. It is pooling of risk, and whoever can manipulate information the best has a significant competitive advantage." Simply put, these companies are in the business of prediction.

The insurance industry has made an art of risk management. In his book *The Failure of Risk Management*, Douglas Hubbard points out what is poignant for all organizations that aren't insurance companies: "No certified, regulated

profession like the actuarial practice exists outside of what is strictly considered insurance.”

Despite this, any and all organizations can master risk the way insurance does. How? By applying PA to predict bad things. For any organization, a predictive model essentially achieves the same function as an insurance company’s *actuarial* approach: rating individuals by the chance of a negative outcome. In fact, we can define PA in these very terms.³

Here’s the original definition:

Predictive analytics (PA)—*Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.*

What an organization effectively learns with PA is *how to decrease risk by way of anticipating microrisks*. Here’s an alternative, risk-oriented definition:

Predictive analytics (PA)—*Technology that learns from experience (data) to manage microrisk.*

Both definitions apply, since each one implies the other.

Like the opportunistic enterprise Tom Cruise’s adolescent entrepreneur launches in his breakout movie of 1983, *Risky Business*, all businesses are risky businesses. And, like insurance companies, all organizations benefit from measuring and predicting the risk of bad behavior, including defaults, cancellations, dropouts, accidents, fraud, and crime. In this way, PA transforms risk to opportunity.

For the economy at large, where could risk management be more important than in the world of mortgages? The mortgage industry, measured in the trillions of dollars, serves as the financial cornerstone of homeownership, the hallmark of family prosperity. And, as important as mortgages are,

³ It works in the other direction as well: While standard actuarial methods involve manual steps such as tabulation and analysis, insurance companies are widely augmenting these practices with predictive modeling in order to better predict outcome. Predictive modeling methods, the topic of this chapter, are more automated and souped up.

risky ones are generally considered a central catalyst to the recent financial crisis or Great Recession.

Microrisks matter. Left unchecked, they threaten to snowball. Our best bet is to learn to predict.

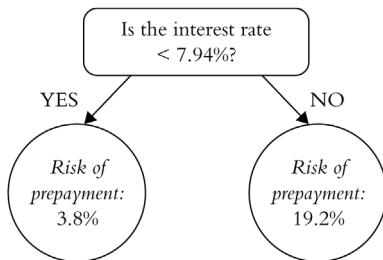
THE LEARNING MACHINE

To learn from data: the process isn't nearly as complex as you might think.⁴

Start with a modest question: What's the *simplest* way to begin distinguishing between high- and low-risk mortgages? What single factor about a mortgage is the most telling?

Dan's learning system made a discovery within Chase's data: *If a mortgage's interest rate is under 7.94 percent, then the risk of prepayment is 3.8 percent; otherwise, the risk is 19.2 percent.*⁵

Drawn as a picture:



⁴ Hands-on familiarity with machine learning is on the rise. Stanford University's computer science department (one of the top three in the United States) first made its Machine Learning course available online for free in 2011, drawing an international enrollment of over 100,000. This success inspired its professor, Andrew Ng, to cofound Coursera, offering free online courses across subject areas.

⁵ The study detailed in this chapter is across 21,816 fixed-rate mortgages with terms of at least 15 years that are relatively new, between one and four years old, and therefore face a higher prepayment risk than average, since borrowers who have been paying off their mortgages for more than four years are more likely to stick with their mortgage as is. Note that interest rates are relative to those in the late 1990s when this project took place and its data was collected.

What a difference! Based only on interest rate, we divide the pool of mortgages into two groups, one five times riskier than the other with respect to the chances of *prepayment* (a customer making an unforeseen payoff of the entire debt, thereby denying the bank future earnings from interest payments).

This discovery is valuable, even if not entirely surprising. Homeowners paying a higher interest rate are more inclined to refinance or sell than those paying a lower rate. If this was already suspected, it's now confirmed empirically, and the effect is precisely quantified.

Machine learning has taken its first step.

BUILDING THE LEARNING MACHINE

You're already halfway there. Believe it or not, there is only one more step before you witness the full essence of machine learning—the ability to generate a predictive model from data, to learn from examples and form an electronic Sherlock Holmes that sizes up an individual and predicts.

You're inches away from the key to one of the coolest things in science, the most audacious of human ambitions: *the automation of learning*.

No sophisticated math or computer code required; in fact, I can explain the rest in two words. But first, let's take a moment to fully define the scientific challenge at hand.

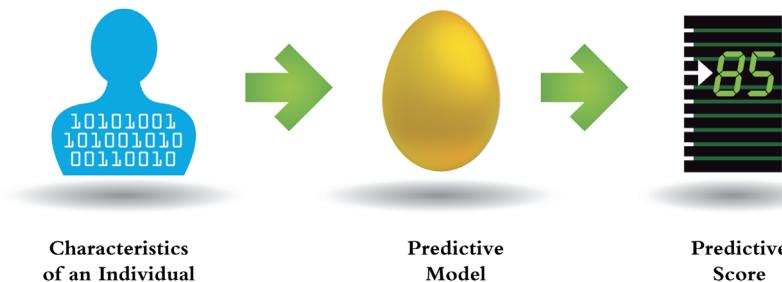
The insight established so far, that interest rate predicts risk, makes for a crude predictive model. It puts each individual mortgage into one of two predictive categories: high risk and low risk. Since it considers only one factor, or *predictor variable*, about the individual, we call this a *univariate* model. All the examples in the previous chapter's tables of bizarre and surprising insights are univariate—they each pertain to one variable such as your salary, your e-mail address, or your credit rating.

We need to go *multivariate*. Why? An effective predictive model surely must consider multiple factors at once, instead of just one. And therein lies the rub. As a refresher, here's the definition:

Predictive model—A mechanism that predicts a behavior of an individual, such as click, buy, lie, or die (or prepay a mortgage). It takes characteristics (variables) of the

individual as input and provides a predictive score as output. The higher the score, the more likely it is that the individual will exhibit the predicted behavior.

Once created with machine learning, a predictive model predicts the outcome for one customer at a time:



Consider a mortgage customer who looks like this:

Borrower: Sally Smithers

Mortgage: \$174,000

Property value: \$400,000

Property type: Single-family residence

Interest rate: 8.92 percent

Borrower's annual income: \$86,880

Net worth: \$102,334

Credit score: Strong

Late payments: 4

Age: 38

Marital status: Married

Education: College

Years at prior address: 4

Line of work: Business manager

Self-employed: No

Years at job: 3

Those are the predictor variables, the characteristics fed into the predictive model. The model's job will be to consider any and all such variables and squeeze them into a single predictive score. Call it the calculation of a new *über-variable*. The model spits out the score, putting all the pieces together to proclaim a singular conclusion.

That's the challenge of machine learning. Your mission is to program your mindless laptop to crunch data about individuals and automatically build the multivariate predictive model. If you succeed, your computer will be learning how to predict.

LEARNING FROM BAD EXPERIENCES

Experience is the name everyone gives to his mistakes.

—Oscar Wilde

My reputation grows with every failure.

—George Bernard Shaw

There's another requirement for machine learning. A successful method must be designed to gain knowledge from a bittersweet mix of good and bad experience, from both the positive and the negative outcomes listed in the data. Some past mortgages went smoothly, whereas others suffered the fate of prepayment. Both of these flavors of data must be leveraged.

To predict, the question we strive to answer is: "How can you distinguish between positive and negative individuals ahead of time?" Learning how to replicate past successes by examining only the positive cases won't work.⁶ Negative examples are critical. Mistakes are your friend.

HOW MACHINE LEARNING WORKS

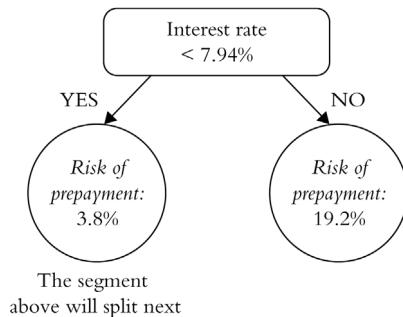
And now, here's the intuitive, elegant answer to the big dilemma, the next step of learning that will move beyond univariate to multivariate predictive modeling, guided by both positive and negative cases: *Keep going.*

⁶ Analyzing only positive cases is sometimes called *profiling* and *cloning* customers.

So far, we've established two risk groups. Next, in the low-risk group, find another factor that best breaks it down even further, into two subgroups that themselves vary in risk. Then do the same thing in the high-risk group. And then keep going within the subgroups. Divide and conquer and then divide some more, breaking down to smaller and smaller groups. And yet, as we'll discover, don't go too far.

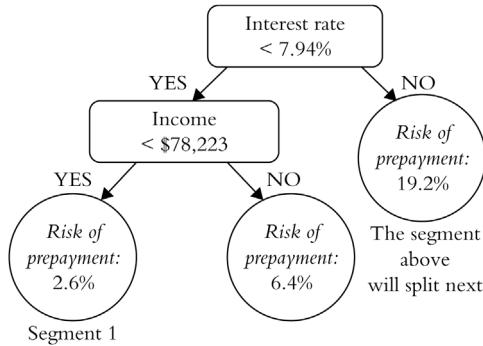
This learning method, called *decision trees*, isn't the only way to create a predictive model, but it's consistently voted as the most or second most popular by practitioners, due to its balance of relative simplicity with effectiveness. It doesn't always deliver the most precise predictive models, but since the models are easier on the eyes than impenetrable mathematical formulas, it's a great place to start, not only for learning about PA, but at the outset of almost any project that's applying PA.

Let's start growing the decision tree. Here's what we had so far:



Now let's find a predictor variable that breaks the low-risk group on the left down further. On this data set, Dan's decision tree software picks the debtor's income:⁷

⁷ The decision tree shown, as well as the decision trees shown later that also predict mortgage prepayment, are simplified depictions in that they don't show how unknown values are handled. For example, some mortgage holders' income levels are unknown. For such missing values, an alternative *surrogate variable* is referenced by the decision tree method in order to decide whether to go left or right at that decision point. Although built from real data, the example decision trees in this chapter are not from the deployed Chase mortgage project.

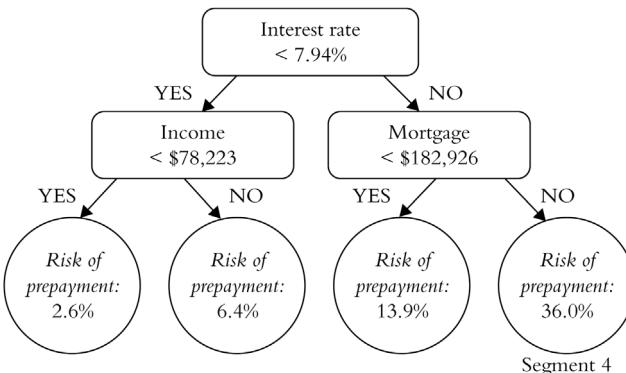


You can see the tree is growing downward. As any computer scientist will tell you, trees are upside down and the *root* is on the top (but if you prefer, you may turn this book upside down).

As shown, the mortgage holder's income is very telling of risk. The lower-left *leaf* (end point of the tree) labeled "Segment 1" corresponds with a subgroup of mortgage holders for whom the interest rate is under 7.94 percent and income is under \$78,223. So far, this is the lowest-risk group identified, with only a 2.6 percent chance of prepayment.

Data trumps the gut. Who would have thought individuals with lower incomes would be less likely to prepay? After all, people with lower incomes usually have a higher incentive to refinance their mortgages. It's tough to interpret; perhaps those with a lower income tend to pursue less aggressive financial tactics. As always, we can only conjecture on the causality behind these insights.

Moving to the right side of the tree, to further break down the high-risk group, the learning software selects mortgage size:



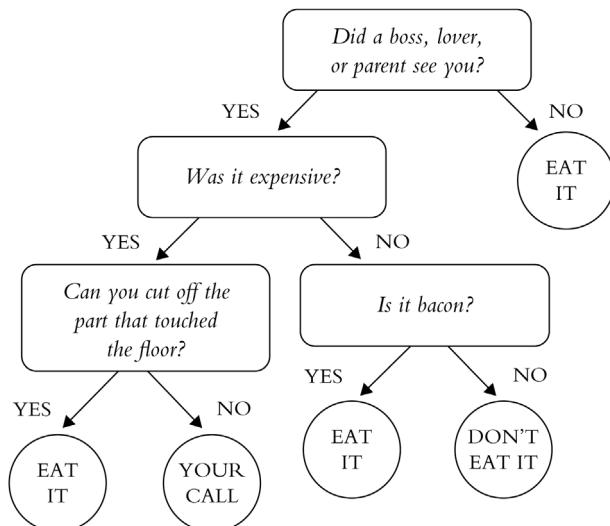
With only two factors taken into consideration, we've identified a particularly risky pocket: higher-interest mortgages that are larger in magnitude, which show a whopping 36 percent chance of prepayment (Segment 4).

Before this model grows even bigger and becomes more predictive, let's talk trees.

DECISION TREES GROW ON YOU

It's simple, elegant, and precise. It's practically mathless. To use a decision tree to predict for an individual, you start at the top (the root) and answer yes/no questions to arrive at a leaf. The leaf indicates the model's predictive output for that individual. For example, beginning at the top, if your interest rate is not less than 7.94 percent, proceed to the right. Then, if your mortgage is under \$182,926, take a left. You end up in a leaf that says, based on these two factors, the risk that you will prepay is 13.9 percent.

Here's an example that decides what you should do if you accidentally drop your food on the floor (excerpted from "The 30-Second Rule: A Decision Tree" by Audrey Fukuman and Andy Wright)—this one was not, to my knowledge, derived from real data:

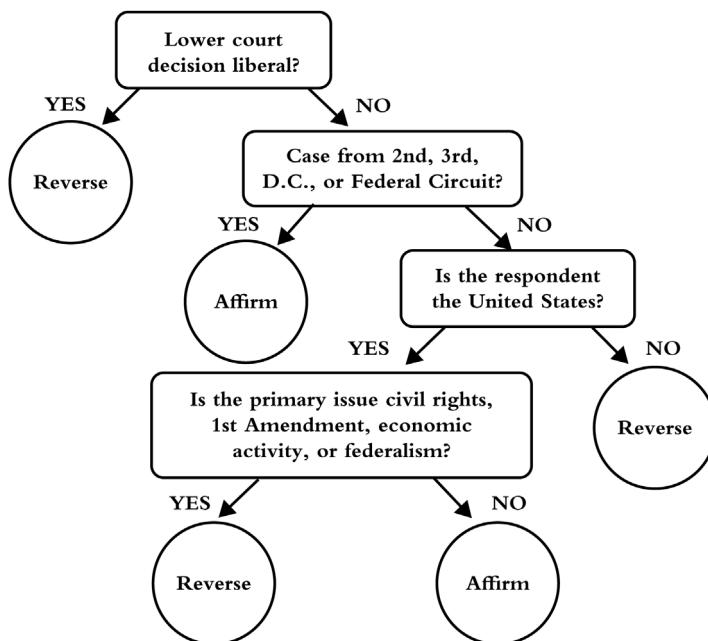


Imagine you just dropped an inexpensive BLT sandwich in front of your mom. Follow the tree from the top and you'll be instructed to eat it anyway.

A decision tree grows upon the rich soil that is data, repeatedly dividing groups of individuals into subgroups. Data is a recording of prior events, so this procedure is learning from the way things turned out in the past. The data determines which variables are used and at what splitting value (e.g., “Income < \$78,223” in the mortgage decision tree). Like other forms of predictive modeling, its derivation is completely automatic—load the data, push a button, and the decision tree grows, all on its own. It’s a rich discovery process, the kind of data mining that strikes gold.

Extending far beyond the business world, decision trees are employed to decide almost anything, whether it is medical, legal, governmental, astronomical, industrial, or you name it. The learning process is intrinsically versatile, since a decision tree’s area of specialty is determined solely by the data upon which it grows. Provide data from a new field of study, and the machine is learning about an entirely new domain.

One decision tree was trained to predict votes on U.S. Supreme Court rulings by former Justice Sandra Day O’Connor. This tree, built across several hundred prior rulings, is from a research project by four university professors in political science, government, and law (“Competing Approaches to Predicting Supreme Court Decision Making,” by Andres D. Martin et al.):



It's simple yet effective. The professors' research shows that a group of such decision trees working together outperforms human experts in predicting Supreme Court rulings. By employing a separate decision tree for each justice, plus other means to predict whether a ruling will be unanimous, the gaggle of trees succeeded in predicting subsequent rulings with 75 percent accuracy, while human legal experts, who were at liberty to use any and all knowledge about each case, predicted at only 59 percent. Once again, data trumps the gut.⁸

COMPUTER, PROGRAM THYSELF

Find a bug in a program, and fix it, and the program will work today. Show the program how to find and fix a bug, and the program will work forever.

—Oliver Selfridge

The logical flow of a decision tree amounts to a simple computer program, so, in growing it, the computer is literally programming itself. The decision tree is a familiar structure you have probably already come across, if you know about any of these topics:

- **Taxonomy.** The hierarchical classification of species in the animal kingdom is in the form of a decision tree.
- **Computer programs.** A decision tree is a nested if-then-else statement. It may also be viewed as a flow chart with no loops.
- **Business rules.** A decision tree is a way to encode a series of if-then business rules; each path from the root to a leaf is one rule (aka a *pattern*, thus the data mining term *pattern discovery*).
- **Marketing segmentation.** The time-honored tradition of segmenting customers and prospects for marketing purposes can be conceived in the form of a decision tree. The difference is that marketing segments are usually designed by hand, following the marketer's intuition,

⁸ Ian Ayres provides an informative overview of the fundamental *intuition versus data* debate in the Chapter "Experts versus Equations" of *Super Crunchers: Why Thinking-by-Numbers Is the New Way to Be Smart* (Bantam, 2007).

whereas decision trees generated automatically with machine learning tend to drill down into a larger number of smaller, more specific subsegments. Also, decision trees usually have a larger group of candidate variables to select from than does handmade segmentation. We could call it *hyper-segmentation*.

- **The game “20 Questions.”** To pass time during long car rides, you think of something and your opponent tries to guess what it is, narrowing down the possibilities by asking up to 20 yes/no questions. Your knowledge for playing this game can be formed into a decision tree. In fact, you can play “Guess the Dictator or Sitcom Character” against the computer at www.smalltime.com/Dictator; if, after asking yes/no questions, it comes up dry, it will add the person you were thinking of to its internal decision tree by saying “I give up” and asking for a new yes/no question (a new *variable*) with which to expand the tree. My first computer in 1980 came with this game (“Animal,” on the Apple][+). It kept the decision tree saved on my 5½” floppy disk.

LEARN BABY LEARN

Old statisticians never die; they just get broken down by age and sex.

—Anonymous

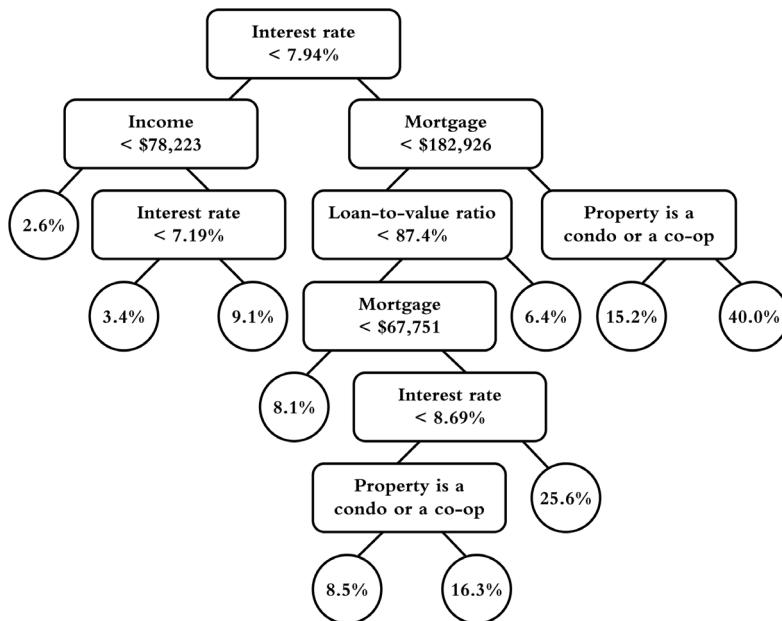
Let’s keep growing on Chase’s data. This is the fun part: pushing the “go” button, which feels like pressing the gas pedal did the first time you drove a car. There’s a palpable source of energy at your disposal: the data, and the power to expose discoveries from it. As the tree grows downward, defining smaller subsegments that are more specific and precise, it feels like a juice squeezer that is crushing out knowledge juice. If there are ways in which human behavior follows patterns, the patterns can’t escape undetected—they’ll be squeezed out into view.

Before modeling, data must be properly arranged in order to access its predictive potential. Like preparing crude oil, it takes a concerted effort to prepare this digital resource as *learning data* (aka *training data*). This involves organizing the data so two time frames are juxtaposed: (1) stuff we knew in the past, and (2) the outcome we’d like to predict, which we came to find out

later. It's all in the past—history from which to learn—but pairing and relating these two distinct points in time is an essential mechanical step, a prerequisite that makes learning to predict possible. This *data preparation* phase can be quite tedious, an involved hands-on technical process often more cumbersome than anticipated, but it's a small price to pay.⁹

Given this potent load of prepared training data, PA software is ready to pounce. “If only these walls could speak . . .” Actually, they can. Machine learning is a universal translator that gives a voice to data.

Here's the tree on Chase mortgage data after several more learning steps (this depiction has less annotation—per convention, go left for “yes” and right for “no”):



⁹ For most all PA software tools, the training data must be a two-dimensional table (or database *view*) with one row per individual (*training case*) and one column per predictor variable, as well as a column corresponding to the *dependent variable*—the outcome being predicted. Although conceptually simple, transforming an organization's data into this form is commonly estimated as 80 percent of the hands-on hours of a PA project.

Learning has now discovered 10 distinct segments (tree leaves), with risk levels ranging from 2.6 percent all the way up to 40 percent. This wide variety means something is working. The process has successfully found groups that differ greatly from one another in the likelihood the thing being predicted—prepayment—will happen. Thus, it has learned how to rank by future probabilities.

To be predicted, an individual tumbles down the tree from top to bottom like a ball in the pinball-like game Pachinko, cascading down through an obstacle course of pins, bouncing left and right. For example, Sally Smithers, the example mortgage customer from earlier in this chapter, starts at the top (tree root) and answers yes/no questions:

Q: *Interest rate < 7.94 percent?*

A: No, go right.

Q: Mortgage < \$182,926?

A: Yes, go left.

Q: Loan-to-value ratio < 87.4 percent?

A: Yes, go left (*the loan is less than 87.4 percent of the property value*).

Q: Mortgage < \$67,751?

A: No, go right.

Q: Interest rate < 8.69 percent?

A: No, go right.

Thus, Sally comes to a landing in the segment with a 25.6 percent propensity. The average risk overall is 9.4 percent, so this tells us there is a relatively high chance she will prepay her mortgage.

Business rules are found along every path from root to leaf. For example, following the path Sally took, we derive a rule that applies to Sally as well as many other homeowners like her (the path has five steps, but the rule can be summarized in fewer lines because some steps revisit the same variable):

IF:

the mortgage is greater than or equal to \$67,751 and less than \$182,926

AND:

the interest rate is greater than or equal to 8.69 percent

AND:

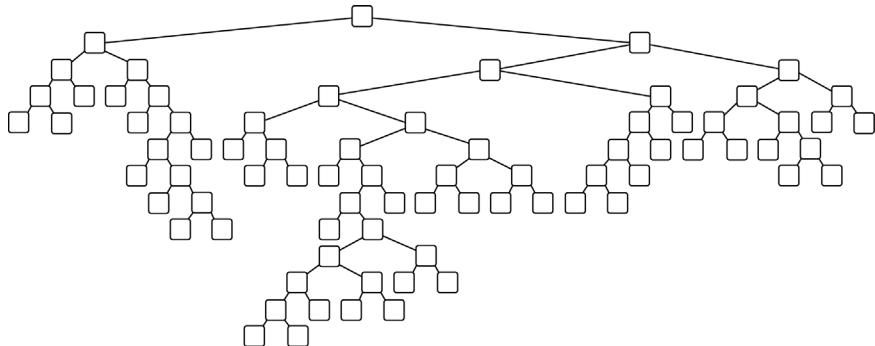
the loan-to-value ratio is less than 87.4 percent

THEN:

the probability of prepayment is 25.6 percent.

BIGGER IS BETTER

Continuing to grow the mortgage risk model, the learning process goes further and lands on an even bigger tree, with 39 segments (leaves), that has this shape to it:



A decision tree with 39 segments.

As the decision tree becomes bigger and more complex, the predictive performance continues to increase, but more gradually. There are diminishing returns.

A single metric compares the performance of predictive models: *lift*. A common measure, lift is a kind of *predictive multiplier*. It tells you how many more target customers you can identify with a model than without one.

Think of the value of prediction from the perspective of the bank. Every prepayment is the loss of a profitable customer. More broadly, the departure of customers is called *customer attrition*, *churn*, or *defection*. Predicting customer attrition helps target marketing outreach designed to keep customers around. Offers designed to retain customers are expensive, so instead of contacting every borrower, the bank must target very precisely.

PA APPLICATION: CUSTOMER RETENTION WITH CHURN MODELING

1. What's predicted: Which customers will leave.

2. What's done about it: Retention efforts target at-risk customers.

Suppose the bank stands to lose 10 percent of its mortgage borrowers. Without a predictive model, the only way to be sure to reach all of them is to contact every single borrower. More realistically, if the marketing budget will allow only one in five borrowers to be contacted, then by selecting randomly without a model, only one in five of those customers soon to be lost will be contacted (on average). Of course, with a crystal ball that predicted perfectly, we could zero in on just the right customers—wouldn't that be nice! Instead, with a less fantastical but reasonably accurate predictive model, we can target much more effectively.

Three times more effectively, to be precise. With the full-sized decision tree model shown previously, it turns out that the 20 percent scored as most high risk includes 60 percent of all the would-be defectors. That is 300 percent as many as without the model, so we say that the model has a *lift* of three at the 20 percent mark. The same marketing budget now has three times as many opportunities to save a defecting customer as before. The bank's bang for its marketing buck just tripled.

The trees we've seen achieve various lifts at the 20 percent mark:

Decision Tree	Lift at 20 Percent
4 segments	2.5
10 segments	2.8
39 segments	3.0

As the tree gets bigger, it keeps getting better, so why stop there? Shall we keep going? Slow down, Icarus! I've got a bad feeling about this.

OVERLEARNING: ASSUMING TOO MUCH

If you torture the data long enough, it will confess.

—Ronald Coase, Professor of Economics, University of Chicago

There are three kinds of lies: lies, damned lies, and statistics.

—British Prime Minister Benjamin Disraeli

(quote popularized by Mark Twain)

An unlimited amount of computational resources is like dynamite: If used properly, it can move mountains. Used improperly, it can blow up your garage or your portfolio.

—David Leinweber, *Nerds on Wall Street*

A few years ago, Berkeley Professor David Leinweber made waves with his discovery that the annual closing price of the S&P 500 stock market index could have been predicted from 1983 to 1993 by the rate of butter production in Bangladesh. Bangladesh's butter production mathematically explains 75 percent of the index's variation over that time. Urgent calls were placed to the Credibility Police, since it certainly cannot be believed that Bangladesh's butter is closely tied to the U.S. stock market. If its butter production boomed or went bust in any given year, how could it be reasonable to assume that U.S. stocks would follow suit? This stirred up the greatest fears of PA skeptics and vindicated nonbelievers. Eyebrows were raised so vigorously, they catapulted Professor Leinweber onto national television.

Crackpot or legitimate educator? It turns out Leinweber had contrived this analysis as a playful publicity stunt, within a chapter entitled “Stupid Data Miner Tricks” in his book *Nerds on Wall Street*. His analysis was designed to highlight a common misstep by exaggerating it. It’s dangerously easy to find ridiculous correlations, especially when you’re “predicting” only 11 data points (annual index closings for 1983 to 1993). By searching through a large number of financial indicators across many countries, something or other

will show similar trends, just by chance. For example, shiver me timbers, a related study showed buried treasure discoveries in England and Wales predicted the Dow Jones Industrial Average a full year ahead from 1992 to 2002.

Predictive modeling can worsen this problem. If, instead of looking at how one factor simply shadows another, you apply the dynamics of machine learning to create models that combine factors, the match can appear even more perfect. It's a catchphrase favored by naysayers: "Hey, throw in something irrelevant like the daily temperature as another factor, and a regression model gets *better*—what does that say about this kind of analysis?" Leinweber got as far as 99 percent accuracy predicting the S&P 500 by allowing a regression model to work with not only Bangladesh's butter production, but Bangladesh's sheep population, U.S. butter production, and U.S. cheese production. As a lactose-intolerant data scientist, I protest!

Leinweber attracted the attention he sought, but his lesson didn't seem to sink in. "I got calls for years asking me what the current butter business in Bangladesh was looking like and I kept saying, 'Ya know, it was a joke, it was a joke!' It's scary how few people actually get that." As *Black Swan* author Nassim Taleb put it in his suitably titled book, *Fooled by Randomness*, "Nowhere is the problem of induction more relevant than in the world of trading—and nowhere has it been as ignored!" Thus the occasional overzealous yet earnest public claim of economic prediction based on factors like women's hemlines, men's necktie width, Super Bowl results, and Christmas day snowfall in Boston.

The culprit that kills learning is *overlearning* (aka *overfitting*). Overlearning is the pitfall of mistaking noise for information, assuming too much about what has been shown within data. You've overlearned if you've read too much into the numbers, led astray from discovering the underlying truth.

Decision trees can overlearn like nobody's business. Just keep growing the tree deeper and deeper—a clear temptation—until each leaf narrows down to just one individual in the training data. After all, if a rule in the tree (formed by following a path from root to leaf) references many variables, it can eliminate all but one individual. But such a rule isn't general; it applies to

only one case. Believing in such a rule is accepting *proof by example*. In this way, a large tree could essentially memorize the entire training data. You've only rewritten the data in a new way.

Rote memorization is the antithesis of learning. Say you're teaching a high school class and you give your students the past few years of final exams to help them study for the exam they'll take next week. If a student simply memorizes the answer to each question in the prior exams, he hasn't actually learned anything and he won't do well on a new exam with all-new questions. Our learning machine has got to be a better student than that.

Even without going to that extreme, striking a delicate balance between learning and overlearning is a profound challenge. For any predictive model a pressing question persists: Has it learned something true that holds in general, or has it discovered patterns that only hold within this data set? *How can we be confident a model will work tomorrow when it is called upon to predict under unique circumstances never before encountered?*

THE CONUNDRUM OF INDUCTION

It must be allowed that inductive investigations are of a far higher degree of difficulty and complexity than any questions of deduction.

—William Stanley Jevons, economist and logician, 1874

To understand God's thoughts, we must study statistics, for these are the measure of his purpose.

—Florence Nightingale

Though this be madness, yet there is method in it.

—*Hamlet*, by William Shakespeare

Life would be so much easier if we only had the source code.

—Hacker aphorism

The objective of machine learning is *induction*:

Induction—Reasoning from detailed facts to general principles.

This is not to be confused with *deduction*, which is essentially the very opposite:

Deduction—Reasoning from the general to the particular (or from cause to effect).

Deduction is much more straightforward. It's just applying known rules. If all men are mortal and Socrates is a man, then deduction tells us Socrates is mortal.

Induction is an art form. At our disposal we have a detailed manifestation of how the world works: data's recording of what happened. From that we seek to generalize, to draw grand conclusions, to ascertain patterns that will hold true in situations not yet seen. We attempt to *reverse engineer* the world's laws and principles. It's the discovery of the method in the madness.

Although a kind of reasoning, induction always behaves unreasonably. This is because it must be based on overly simplistic assumptions. Assumptions are key to the inductive leap we strive to take. You simply cannot design a learning method without them. We don't know enough about the way the world works to design perfect learning. If we did, we wouldn't need machine learning to begin with. For example, with decision trees, the implicit assumption is that the rules within a decision tree, as simple as they may be, are an astute way to capture and express true patterns.

Carnegie Mellon professor Tom Mitchell, founding chair of the world's first machine learning department and the author of the first academic textbook on the subject, *Machine Learning*, calls this kind of assumption an *inductive bias*. Establishing these foundational assumptions—part and parcel to inventing new induction methods—is the art behind machine learning. There's no one best answer, no one learning method that always wins above all others. It depends on the data.¹⁰

Machine induction and the induction of birth have something in common. In both cases, there's a genesis.

¹⁰ This nonexistence of a universal solution to machine learning is put into formal terms by the “no free lunch” theorem.

THE ART AND SCIENCE OF MACHINE LEARNING

The method is to modify your model incrementally.

Tweak the technique, geek, improving it incessantly.

Each step is taken to improve prediction on the training cases.

One small step for man; one giant leap—the human race is going places!

Modeling methods vary, but they all face the same challenge: to learn as much as possible, yet not learn too much. Among the competing approaches to machine learning, decision trees are often considered the most user friendly, since they consist of rules you can read like a long (if cumbersome) English sentence, while other methods are more mathy, taking the variables and plugging them into equations.

Most learning methods *search* for a good predictive model, starting with a trivially simple and often inept model and tweaking it repeatedly, as if applying “genetic mutations,” until it evolves into a robust prediction apparatus. In the case of a decision tree, the process starts with a small tree and grows it. In the case of most mathematical equation-based methods, it starts with a random model by selecting random parameters and then repeatedly nudges the parameters until the equation is predicting well. For all learning techniques, the training data guides each tweak as it strives to improve prediction across that data set. To put names on the mathy methods that compete with decision trees, they include *artificial neural networks*, *loglinear regression*, *support vector machines*, and *TreeNet*.

Machine learning’s ruthless, incessant adaptation displays an eerie power. It even discovers and exploits weaknesses or loopholes inadvertently left in place by the data scientist. In one project with my close friend Alex Chaffee (a multitalented software architect), we set up the computer to “evolve” a Tetris game player, learning how to decide where to drop each piece while playing the game. In one run of the system, we accidentally reversed the objective (a single errant minus sign instead of a plus sign within thousands of lines of computer code!) so that, instead of striving to tightly pack the game pieces, it was rewarded for packing *less* densely by creating big, vacant holes. Before we realized it was chasing a bug, we were perplexed to see the

resulting game player stacking pieces up diagonally from the bottom left of the game board to the top right, a creative way to play as *poorly* as possible.¹¹ It conjures the foreboding insight of brooding scientist Ian Malcolm in Michael Crichton's dinosaur thriller *Jurassic Park*: "Life finds a way."

Regardless of the learning technique and its mathematical sophistication, there's always the potential to overlearn. After all, commanding a computer to learn is like teaching a blindfolded monkey to design a fashion diva's gown. The computer knows nothing. It has no notion of the meaning behind the data, the concept of what a mortgage, salary, or even a house is. The numbers are just numbers. Even clues like "\$" and "%" don't mean anything to the machine. It's a blind, mindless automaton stuck in a box forever enduring its first day on the job.

Every attempt to predictively model faces this central challenge to establish general principles and weed out the noise, the artifacts peculiar only to the limited data at hand. It's the nature of the problem. Even if there are millions or billions of examples in the data from which to learn, it's still a limited portion compared to how many conceivable situations could be encountered in the future. The number of possible combinations that may form a learning example is exponential. And so, architecting a learning process that strikes the balance between learning too much and too little is elusive and mysterious to even the most hard-core scientist.

In solving this conundrum, art comes before science, but the two are both critical components. Art enables it to work, and science proves it works:

- 1. Artistic design:** Research scientists craft machine learning to attempt to avert overlearning, often based on creative ideas that sound just brilliant.
- 2. Scientific measure:** The predictive model's performance is objectively evaluated.

In the case of number 2, though, what method of evaluation could possibly suffice? If we can't entirely trust the design of machine learning, how can

¹¹ To view a nonbuggy, proficient Tetris player we evolved, see www.predictionimpact.com/tetris.

we trust a measure of its performance? Of course, all predictions could be evaluated by simply waiting to see if they come true. But since we plan to pay heed to a model's predictions and take actions accordingly, we must establish confidence in the model immediately. We need an almost instantaneous means to gauge performance so that, if overlearning takes place, it can be detected and the course of learning corrected by backtracking and trying again.

FEELING VALIDATED: TEST DATA

The proof is in the pudding.

There's no fancy math required to test for true learning. Don't get me wrong; they've tried. Theoretical work abounds—these deep thinkers have even met for their 28th Annual Conference on Learning Theory. But the results to date are limited. It seems impossible to design a learning method that's guaranteed not to overlearn. It's a seriously hard scientific problem.

Instead, a clever, mind-numbingly simple trick is employed to test for overlearning: *Hold aside some data to test the model.* Randomly select a *test set* (aka *validation* or *out-of-sample set*) and quarantine it. Use only the remaining portion of data, the *training set*, to create the model. Then, evaluate the resulting model across the test set. Since the test set was not used to create the model, there's no way the model could have captured its esoteric aspects, its eccentricities. The model didn't have the opportunity to memorize it in any way. Therefore, however well the model does on the test set is a reasonable estimation of how well the model does in general, a true evaluation of its ability to predict. For evaluating the model, the test set is said to be *unbiased*.

Training data	Testing data
Used by machine learning to generate a predictive model	Used to evaluate the predictive model

No mathematical theory, no advanced science, just an elegant, practical solution. This is how it's done, always. It's common practice. Every predictive modeling software tool has a built-in routine to hold aside and evaluate over test data. And

every research journal article reports predictive performance over test data (unless you’re poking fun at the industry’s more egregious errors with a humorous example about Bangladesh’s butter and the stock market).¹²

There’s one downside to this approach. You sacrifice the opportunity to learn from the examples in the test set, generating the model only from the now-smaller training set. Typically this is a loss of 20 percent or 30 percent of the training data, which is held aside as test data. But the training set that remains is often plenty big, and the sacrifice is a small price to pay for a true measure of performance.

Following this practice, let’s take a look at the true test performance of the decision tree models we’ve looked at so far. Recall that the *lift* performance of our increasingly larger trees, as evaluated over the 21,816 cases in the training set, was:

Decision Tree	Lift at 20 Percent on the Training Set
4 segments	2.5
10 segments	2.8
39 segments	3.0

It turns out, for these trees, no overlearning took place. As evaluated on another 5,486 examples that had been held aside all along as the test set, the lifts for these three models held at 2.5, 2.8, and 3.0, respectively. Success!

Decision Tree	Lift on the Training Set	Lift on the Test Set
4 segments	2.5	2.5
10 segments	2.8	2.8
39 segments	3.0	3.0

¹² By evaluating over an unseen test set, we eliminate the messy requirement to explicitly account for *vast search*, as covered in the prior chapter. In that chapter, we saw how the “orange lemon” finding was debunked by way of a statistical method that took vast search into consideration, but that option was applied only because no separate test set was available. Normally, you don’t have to work so hard in this respect.

Keep going, though, and you'll pass your limit. If the tree gets even bigger, branching out to a greater number of smaller segments, learning will become overlearning. Taking it to an extreme, once we get to a tree with 638 segments (i.e., end points or leaves), the lift on the training set is 3.8, the highest lift yet. But the performance on that data, which was used to form the model in the first place, is a biased measure. Trying out this large tree on the test set reveals a lift of 2.4, lower than that of the small tree with only four segments.

Decision Tree	Lift on the Training Set	Lift on the Test Set
638 segments	3.8	2.4 (<i>overlearning</i>)

The test data guides learning, showing when it has worked and when it has gone too far.

CARVING OUT A WORK OF ART

In every block of marble I see a statue as plain as though it stood before me, shaped and perfect in attitude and action. I have only to hew away the rough walls that imprison the lovely apparition to reveal it to the other eyes as mine see it.

—Michelangelo

Everything should be made as simple as possible, but not simpler.

—Albert Einstein (as paraphrased by Roger Sessions)

The decision tree fails unless we tame its wild growth. This presents a tough balance to strike. Like a parent, we strive to structure our progeny's growth and development so they're not out of control, and yet we cannot bear to quell creativity. Where exactly to draw the line?

When they first gained serious attention in the early 1960s, decision trees failed miserably, laughed out of court for their propensity to overlearn. “They were called ‘a recipe for learning something wrong,’” says Dan

Steinberg. “This was a death sentence, like a restaurant with *E. coli*. Trees were finished.”

For those researchers who didn’t give up on trees, formally defining the line between learning and overlearning proved tricky. It seemed as though, no matter where you drew the line, there was still a risk of learning too little or too much. Dramatic tension mounted like an unresolvable tug-of-war.

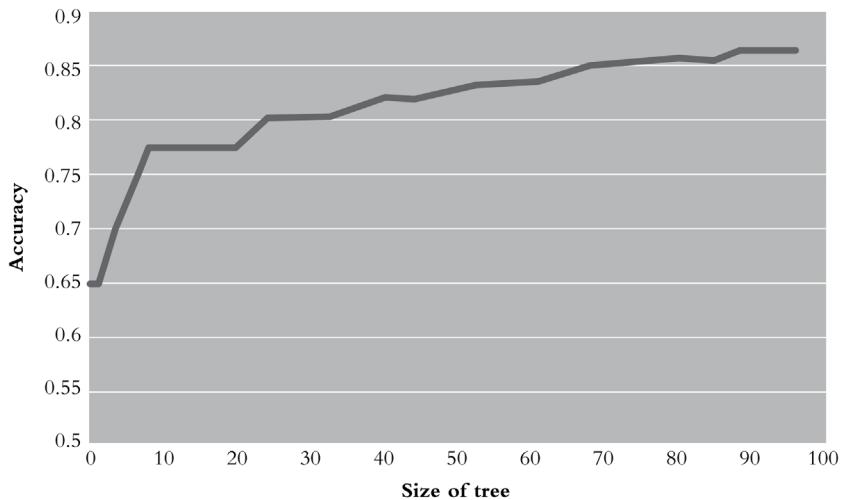
As with the theater, irony eases the tension. The most popular solution to this dilemma is ironic. Instead of holding back so as to avoid learning too much, don’t hold back at all. Go all the way—learn way too much . . . and then take it all back, piece by piece, unlearning until you’re back to square one and have learned too little. Set forth and make mistakes! Why? Because the mistakes are apparent only after you’ve made them.

In a word, grow the tree too big and bushy, and then prune it back. The trick is that pruning is guided not by the training data that determined the tree’s growth, but by the testing data that now reveals where that growth went awry. It’s an incredibly elegant solution that strikes the delicate balance between learning and overlearning.

To prune back a tree is to backtrack on steps taken, undoing some of machine learning’s tweaks that have turned out to be faulty. By way of these undo’s that hack and chop tree branches, a balanced model is unearthed, not timidly restricted and yet not overly self-confident. Like Michelangelo’s statue, revealed within his block of marble by carving away the extraneous material that shrouds it, an effective predictive model is discovered within.

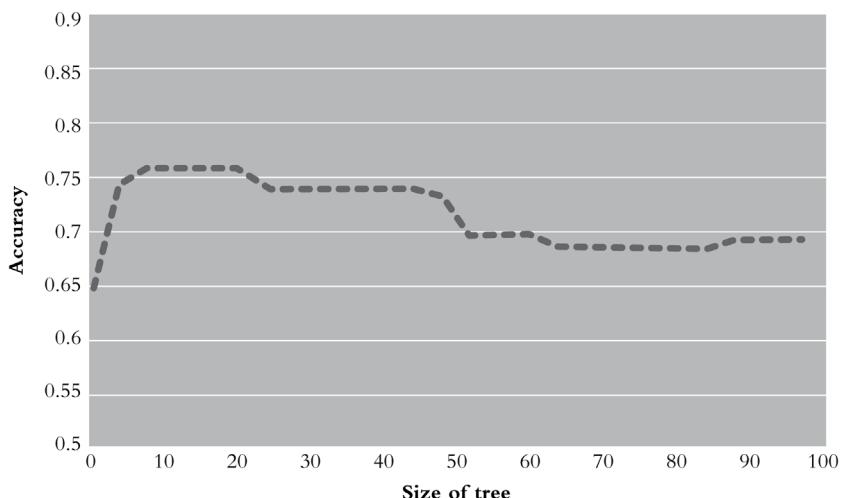
It’s easy to take a wrong turn while building a predictive model. The important thing is to ensure that such steps are undone. In a training workshop I lead, trainees build predictive models by hand, following their own process of trial and error. When they try out a change that proves to hurt a model rather than improve it, they’ve been heard to exclaim, “Go back, go back—we should go back!”

To visualize the effect, consider the improvement of a decision tree during the training process:



From *Machine Learning*, by Tom Mitchell

As shown, while the tree grows, the accuracy—as measured over the training data used to grow it—just keeps improving. But during the same growth process, if we test for true, unbiased accuracy over the test set, we see that it peaks early on, and then further growth makes for overlearning, only *hurting* its predictions:



From *Machine Learning*, by Tom Mitchell

Hedging these bushes follows the principle known as *Occam's razor*: Seek the simplest explanation for the data available. The philosophy is, by seeking more parsimonious models, you discover better models. This tactic defines part of the inductive bias that intentionally underlies decision trees. It's what makes them work. If you care about your model, give it a KISS: "Keep it simple, stupid!"

The leading decision tree modeling standard, called *Classification and Regression Trees* (CART), employs this elegant form of pruning, plus numerous other bells and whistles in its routines.¹³ CART was established by a 1984 book of the same name by four legendary researchers from Berkeley and Stanford: Leo Breiman, Jerome Friedman, Charles Stone, and Richard Olshen. I call them the "Fab Four." As with most major inventions such as the television and the airplane, other parties released competing decision tree-based techniques around the same time, including researchers in Australia (ID3) and South Africa (CHAID). CART is the most commonly adopted; PA software tools from the likes of IBM and Dell include a version of CART. Dan Steinberg's company, Salford Systems, sells the only CART product codeveloped by the Fab Four, who are also investors.

An entrepreneurial scientist, Dan earned the Fab Four's trust to deliver CART from their research lab to the commercial world. Dan hails from Harvard with a PhD in econometrics. Not to put the CART before the horse, he founded his company soon after CART was invented.

The validation of machine learning methods such as CART is breaking news: "Human Intuition Achieves Astounding Success." The fact that machine learning works tells us that we humans are smart enough—the hunches and intuitions that drive the design of methods to learn yet not overlearn pan out. I call this *The Induction Effect*:

¹³ CART® is a registered trademark licensed exclusively to Salford Systems.

The Induction Effect: *Art drives machine learning; when followed by computer programs, strategies designed in part by informal human creativity succeed in developing predictive models that perform well on new cases.*

PUTTING DECISION TREES TO WORK FOR CHASE

Dan agreed to help Chase with mortgage prediction (in a collaborative effort alongside a large consulting company), and the rubber hit the road. He pulled together a small team of scientists to apply CART to Chase's mortgage data.

Chase had in mind a different use of churn prediction than the norm.¹⁴ Most commonly, a company predicts which customers will leave (aka *churn* or *defect*—in the case of mortgage customers, *prepay*) in order to target *retention* activities meant to convince them to stay.

But Chase's plans put a new twist on the value of predicting churn. The bank intended to use the predictive scores to estimate the expected future value of individual mortgages in order to decide whether it would be a good move to sell them to other banks. Banks buy and sell mortgages at will. At any time, a mortgage could be sold based on its current market price, given the profile of the mortgage. But the market at large didn't have access to these predictive models, so Chase held a strong advantage. It could estimate the future value of a mortgage based on the predicted chance of prepayment. In a true manifestation of prediction's power, Chase could calculate whether selling a mortgage was likely to earn more than holding on to it. Each decision could be driven with prediction.

¹⁴ However, examples of the more typical process of customer retention with churn modeling are included elsewhere in this book—see Central Table 2 and Chapter 7.

PA APPLICATION: MORTGAGE VALUE ESTIMATION

1. **What's predicted:** Which mortgage holders will prepay within the next 90 days.
2. **What's done about it:** Mortgages are valued accordingly in order to decide whether to sell them to other banks.

Chase intended to drive decisions across many mortgages with these predictions. Managing millions of mortgages, Chase was putting its faith in prediction to drive a large-scale number of decisions.

PA promised a tremendous competitive edge for Chase in the mortgage marketplace. While the market price for a mortgage depended on only several factors, a CART model incorporates many more variables, thus serving to more precisely predict each mortgage's future value.

With prediction, risk becomes opportunity. A mortgage destined to suffer the fate of prepayment is no longer bad news if it's predicted as such. By putting it up for sale accordingly, Chase could tip the outcome in its favor.

With data covering millions of mortgages, the amount available for analysis far surpassed the training set of about 22,000 cases employed to build the example decision trees depicted in this chapter. Plus, for each mortgage, there were in fact *hundreds* of predictor variables detailing its ins and outs, including summaries of the entire history of payments, home neighborhood data, and other information about the individual consumer. As a result, the project demanded 200 gigabytes of storage. You could buy this much today for \$25 and place it into your pocket, but in the late 1990s, the investment was about \$250,000, and the storage unit was the size of a refrigerator.

The Chase project required numerous models, each specialized for a different category of mortgage. CART trees were grown separately for fixed-rate versus variable-rate mortgages, for mortgages of varying terms, and at different stages of tenure. After grouping the mortgages accordingly, a separate decision tree was generated for each group. Since each tree addressed a different type of situation, the trees varied considerably, employing their own particular group of variables in divergent ways. Dan's team

delivered these eclectic decision trees to Chase for integration into the bank's systems.

MONEY GROWS ON TREES

The undertaking was an acclaimed success. People close to the project at Chase reported that the predictive models generated millions of dollars of additional profit during the first year of deployment. The models correctly identified 74 percent of mortgage prepayments before they took place, and drove the management of mortgage portfolios successfully.¹⁵

As an institution, Chase was bolstered by this success. To strengthen its brand, it issued press releases touting its competency with advanced analytics.

Soon after the project launch, in 2000, Chase achieved yet another mammoth milestone to expand. It managed to buy JPMorgan, thus becoming JPMorgan Chase, now the largest U.S. bank by assets.

THE RECESSION—WHY MICROSCOPES CAN'T DETECT ASTEROID COLLISIONS

Needless to say, PA didn't prevent the global financial crisis that began several years later, in late 2007. That wasn't its job. Applied to avert microrisks, PA packs a serious punch. But tackling macroscopic risk is a completely different ballgame. PA is designed to rank individuals by their *relative* risk, but not to adjust the *absolute* measurements of risk when a broad shift in the economic environment is nigh. The predictive model operates on

¹⁵ There is some disagreement among sources regarding the degree to which prepay prediction was employed by Chase to help retain mortgages and/or to price mortgages; the latter use was considered potentially damaging to Chase's reputation in the eyes of partner banks.

variables about the individual, such as age, education, payment history, and property type. These factors don't change even as the world around the individual changes, so the predictive score for the individual doesn't change, either.¹⁶

Predicting macroscopic risk is a tall order, with challenges surpassing those of microrisk prediction. The pertinent factors can be intangible and human. As the *New York Times*'s Saul Hansell put it, "Financial firms chose to program their risk-management systems with overly optimistic assumptions. . . . Wall Street executives had lots of incentives to make sure their risk systems didn't see much risk." Professor Bart Baesens of the University of Southampton's Centre for Risk Research adds, "There's an inherent tension between conservative instincts and profit-seeking motives." If we're not measuring and reporting on the truth, there's no analytical cure.

Efforts in economic theory attempt to forecast macroscopic events, although such work in forecasting is not usually integrated within the scope of PA. However, Baesens has suggested, "By incorporating macroeconomic factors into a model, we can perform a range of data-driven stress tests." Such work must introduce a new set of variables in order to detect worldwide shifts and requires a different analytical approach, since there are no sets of training data replete with an abundance of Black Swan events from which PA may learn. The rarest things in life are the hardest to predict.

AFTER MATH

Decision trees vanquish, but do they satisfy the data scientist's soul? They're understandable to the human eye when viewed as rules, each one an interpretable (albeit clunky) English sentence. This is surely an advantage

¹⁶ The form of microrisk relevant to the economic crisis is that of delinquent debtors, rather than the microrisk predicted in this chapter's case study, mortgage prepayments. But predicting delinquent accounts with PA is also subject to these same limitations.

for some organizations, but on other occasions we'd gladly exchange simplicity for performance.

In the next chapter, we pursue Netflix's heated public competition to outpredict movie ratings. Fine-tuning predictive performance is the name of the game. Must souping up model precision involve overwhelming complexity, or is there an elegant way to build and scale?



CHAPTER 5

The Ensemble Effect

Netflix, Crowdsourcing, and Supercharging Prediction

To crowdsource predictive analytics—outsource it to the public at large—a company launches its strategy, data, and research discoveries into the public spotlight. How can this possibly help the company compete? What key innovation in predictive analytics has crowdsourcing helped develop? Must supercharging predictive precision involve overwhelming complexity, or is there an elegant solution? Is there wisdom in nonhuman crowds?

CASUAL ROCKET SCIENTISTS

A buddy and I are thinking of building a spaceship next year. The thing is, we have absolutely no training or background. But who cares? I want to go to outer space.

This may sound outlandish, but in the realm of predictive analytics (PA), it is essentially what Martin Chabbert and Martin Piotte did. In 2008, this pair of Montrealers launched a mission to win the \$1 million Netflix Prize, the most high-profile analytical competition of its time. Incredibly, with no background in analytics, these casual part-timers became a central part of the story.

The movie rental company Netflix launched this competition to improve the movie recommendations it provides to customers. The company challenged the world by requiring that the winner improve upon Netflix's own established recommendation capabilities by 10 percent. Netflix is a prime example of PA in action, as a reported 70 percent of Netflix movie choices arise from its online recommendations. Product recommendations are increasingly important for the retail industry in general. More than a sales

ploy, these tailored recommendations provide relevancy and personalization that customers actively seek.

PA APPLICATION: MOVIE RECOMMENDATIONS

- 1. What's predicted:** What rating a customer would give to a movie.
- 2. What's done about it:** Customers are recommended movies that they are predicted to rate highly.

PA contests such as the Netflix Prize leverage competitive spirit to garner scientific advancement. Like a horse race, a competition levels the playing field and unambiguously singles out the best entrant. With few limitations, almost anyone in the world—old or young, tall or short—can participate by downloading the data, forming a predictive model, and submitting.

It's winner take all. To ensure submissions are objectively compared, prediction competitions employ a clever trick: The competitor must submit not a predictive model, but its predictive scores, as generated for an evaluation data set within which the correct answers—the target values that the model is meant to infer—are withheld. Netflix Prize models predict how a customer would rate a movie (based on how he or she has rated other movies). The true ratings are suppressed in the publicly posted evaluation data, so submitters can't know exactly which examples they're getting right and which they're getting wrong at the time of submission. All said, to launch the competition, Netflix released to the public over 100 million ratings from some 480,189 customers (anonymized for privacy considerations, with names suppressed).¹

The model's ability to predict is all that matters, not the modeler's background, experience, or academic pedigree. Such a contest is a hard-nosed, objective bake-off—whatever can cook up the solution that best handles the predictive task at hand wins kudos and, usually, cash.

¹ PA contests do include the target values in the main data set provided for competitors to train their models. It is up to competitors to split that data into training and testing sets during model development, as discussed in the prior chapter.

DARK HORSES

And so it was with our two Montrealers, Martin and Martin, who took the Netflix Prize by storm despite their lack of experience—or perhaps *because* of it. Neither had a background in statistics or analytics, let alone recommendation systems in particular. By day, the two worked in the telecommunications industry developing software.

But by night, the two-member team plugged away at home for 10 to 20 hours per week apiece, racing ahead in the contest under the team name PragmaticTheory. The “pragmatic” approach proved groundbreaking. The team wavered in and out of the number one slot; during the final months of the competition, the team was often in the top echelons.

There emerges an uncanny parallel to SpaceShipOne, the first privately funded human spaceflight, which won the \$10 million Ansari X Prize. According to some, this small team, short on resources with a spend of only \$25 million, put the established, gargantuan NASA to shame by doing more for so much less. PA competitions do for data science what the X Prize did for rocket science.

MINDSOURCED: WEALTH IN DIVERSITY

[Crowdsourcing is] a perfect meritocracy, where age, gender, race, education, and job history no longer matter; the quality of the work is all that counts.

—Jeff Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*

When pursuing a grand challenge, from where will key discoveries appear? If we assume for the moment that one cannot know, there’s only one place to look: everywhere. Contests tap the greatest resource, the general public. A common way to enact crowdsourcing, an open competition brings together scientists from far and wide to compete for the win and cooperate for the joy. With crowdsourcing, a company outsources to the world.

The \$1 million Netflix Prize attracted a white-hot spotlight and built a new appreciation for the influence crowdsourcing holds to rally an international

wealth of bright minds. In total, 5,169 teams formed to compete in this contest, submitting 44,014 entries by the end of the event.

PA crowdsourcing reaps the rewards brought by a diverse *brainshare*. Chris Volinsky, a member of a leading Netflix Prize team named BellKor from AT&T Research, put it to me this way: “From the beginning, I thought it was awesome how many people in the top of the leaderboard were what could be called ‘amateurs.’ In fact, our group had no experience with [product recommendations] when we started, either. . . . It just goes to show that sometimes it takes a fresh perspective from outside the field to make progress.”

One mysterious, highly competitive team came out of the woodwork, calling itself “Just a guy in a garage.” The team was anonymous but rose at one point to sixth place on the competition’s leaderboard. Later, going public, it turned out to be a one-member team, a former management consultant who went to college for psychology and graduate school for operations research (he lists himself as unemployed and has revealed he was working out of a second bedroom in his house rather than an actual garage).

So too did our pair of dark horse laymen, team PragmaticTheory, circumvent established practices, effortlessly thinking outside a box they knew nothing of in the first place. Unbounded, they could *boldly go* . . . in new directions that no one had gone before. As Martin Chabbert told me in an interview, they “figured that a more pragmatic and less dogmatic approach might yield some good results.” Ironically, their competitive edge appeared to hinge less on scientific innovation and more on their actual expertise: adept software engineering. Martin provided this striking lesson:

Many people came up with (often good) ideas . . . but translating those words into a mathematical formula is the complicated part. . . . Our background in engineering and software was key. In this contest, there was a fine line between a bad idea and a bug in the code. Often you would think that the model was simply bad because it didn’t yield the expected results, but in fact the problem was a bug in the code. Having the ability to write code with few bugs and the skill to actually find the bugs before giving up on the model is something that definitely helped a lot. . . . Compared to what most people think, this was more of an engineering contest than a mathematical contest.

Cross-discipline competitors thrive, as revealed by many PA contests beyond the Netflix Prize. One competition concerned with educational applications witnessed triumph by a particle physicist (Great Britain), a data analyst for the National Weather Service (Washington, D.C.), and a graduate student (Germany); \$100,000 in prize money sponsored by the Hewlett Foundation (established by a founder of Hewlett-Packard) went to these winners, who developed the best means to automatically grade student-written essays. Their resulting system grades essays as accurately as human graders, although none of these three winners had backgrounds in education or text analytics.

And guess what kind of expert excelled at predicting the distribution of dark matter in the universe? Competing in a contest sponsored by NASA and the Royal Astronomical Society, Martin O’Leary, a British PhD student in *glaciology*, generated a method the White House announced has “outperformed the state-of-the-art algorithms most commonly used in astronomy.” For this contest, O’Leary provided the first major breakthrough (although he was not the eventual winner). As he explains it, aspects of his work mapping the edges of glaciers from satellite photos could extend to mapping galaxies as well.

CROWDSOURCING GONE WILD

Given the right set of conditions, the crowd will almost always outperform any number of employees.

—Jeff Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*

The organizations I’ve worked with have mostly viewed the competition in business as a race that benefits from sharing, rather than a fight where one’s gain can come only from another’s loss. The openness of crowdsourcing aligns with this philosophy.

—Stein Kretsinger, Founding Executive, [Advertising.com](#)

One small groundbreaking firm, Kaggle, has taken charge and leads the production of PA crowdsourcing. Kaggle has launched more than 175 PA competitions, including the essay-grading and dark matter ones mentioned

above. Over 50,000 registered competitors are incentivized by prizes that usually come to around \$10,000 to \$20,000, but climb as high as \$500,000. These diverse minds from more than 200 universities and 100 countries, about half of them academics, have submitted over 144,000 attempts for the win.²

An enterprise turns research and development completely on its head in order to leverage PA crowdsourcing. Instead of protecting strategy, plans, data, and research discoveries as carefully guarded secrets, a company must launch them fully into the public spotlight. And instead of maintaining careful control over its research staff, the organization gets whoever cares to take part in the contest and join in on the fun (for fully public contests, as is the norm). Crowdsourcing must be the most ironic, fantastical way for a business to compete.

Crowdsourcing forms a match made in heaven. Kaggle's founder and CEO, Anthony Goldbloom (a *Forbes* "30 Under 30: Technology" honoree), spells out the love story: "On one hand, you've got companies with piles and piles of data, but not the ability to get as much out of it as they would like. On the other hand, you've got researchers and data scientists, particularly at university, who are pining for access to real-world data" in order to test and refine their methodologies.

With strong analytics experts increasingly tough to find, seeding your talent pool by reaching out to the masses starts to sound like a pretty good idea. A McKinsey report states, "By 2018, the United States alone could face

² Moving beyond PA to the broader category of science and business problems, InnoCentive is the analogue to Kaggle, with over 1,300 crowdsourcing challenges posted to date. The cover illustration of this book's first edition was developed by the winner of a "crystal ball" design contest the author hosted on 99designs. By hosting a competitive 3-D video game puzzle that anyone can learn to play, Foldit broke ground in protein folding to produce three discoveries that have been published in *Nature*. Noncompetitive crowdsourcing also bears great fruit, including the advent of Wikipedia and open source software such as the Linux operating system and R, the most popular free software for analytics, which itself is employed by more Kaggle competitors than any other tool.

a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions.” To leave no analytical stone unturned, innovative organizations turn by necessity to the crowd at large. As Kaggle pitches, “There are countless strategies that can be applied to any predictive modeling task, and it is impossible to know at the outset which technique or analyst will be most effective.”

Until a few years ago, most PA competitions were held by academic institutions or research conferences. Kaggle has changed this. With the claim that it “has never failed to outperform a preexisting accuracy benchmark, and to do so resoundingly,” Kaggle has brought commercial credibility to the practice. For example, across this book’s Central Tables of 182 PA examples, 14 come from Kaggle competitions—namely:³

Organization	What Is Predicted	Central Table to See for More Info
Facebook	Friendship	1
dunnhumby	Supermarket visits	2
Allstate	Bodily harm from car crashes	3
Heritage Health Prize	Days spent in the hospital	4
Researchers	HIV progression	4
New South Wales, Australia	Travel time vis-à-vis traffic	6
University of Melbourne	Awarding of grants	7
Hewlett Foundation	Student grades	7
Grockit	Student knowledge	7
Imperium	Insults	8
Ford Motor Co.	Driver inattentiveness	8
Online Privacy Foundation	Psychopathy	8
Wikipedia	Editor attrition	9
CareerBuilder	Job applications	9

³ Kaggle competitions have also advanced the state of the art in HIV research and chess ratings. For a list of even more PA competitions than those found on Kaggle’s website, see this chapter’s Notes at www.PredictiveNotes.com.

YOUR ADVERSARY IS YOUR AMIGO

Competition paradoxically breeds cooperation. Kaggle's tagline is "making data science a sport." But these lab coat competitors don't seem to exhibit the same fierce, cutthroat voraciousness as sweaty athletes out on the field. Despite the cash incentive to come out on top, participants are often driven by the love of science. They display a spirited tendency to collaborate and share. It's the best of *coopetition*. Netflix Prize leader Martin Chabbert told me the prize's public forum "was also a place where people proposed new ideas; these ideas often inspired us to come up with our own creative innovations." And *Wired* magazine wrote, "The prize hunters, even the leaders, are startlingly open about the methods they're using, acting more like academics huddled over a knotty problem than entrepreneurs jostling for a \$1 million payday." When John Elder took part in the competition, he took pause. "It was astonishing how many people were openly sharing and cooperating," John says. "It comes of what people do out of camaraderie."

And so a community emerges around each contest, catalyzing a petri dish of great ideas. But John Elder recognizes that disclosure can cost a competitive edge. John and some staff at Elder Research were part of a Netflix Prize team during earlier phases of the contest when much of the major headway was still being made. At one point the team held third place, having employed a key analytical method before any other competitor. The method, you will soon see, was a key ingredient both to winning the Netflix Prize and to building IBM's Watson, the *Jeopardy!* player. In a collegial spirit, John's team went as far as displaying this choice of method as their very name, thereby revealing their secret weapon. The team was called "Ensemble Experts."

UNITED NATIONS

As competitors rounded the final bend of the horse race known as the Netflix Prize (which launched before Kaggle all but took over the field), a handful of key leaders held a tense dead heat. Ironically, the race moved along at the

speed of a snail, as if watching a sporting event's slow-motion instant replay on TV. Because there were diminishing returns on the teams' efforts and their predictive models' complexity, the closer they got to their objective—a 10 percent improvement over Netflix's established method that would qualify for the \$1 million win—the more slowly they progressed.

Despite the glacial pace, it was gripping. The leaders faced dramatic upsets every week as they leapfrogged one another on the contest's public leader-board. The teams jockeyed for position by way of minuscule improvements.

While nobody, including Netflix, knew if the 10 percent mark was even possible, there was the constant sense that, at any moment, a team could find a breakthrough and catapult into the win zone.

A promising breakthrough popped in September 2008, temporarily leaving our heroic lay competitors in the dust. Two other teams, BellKor (from AT&T Research) and BigChaos (a strikingly young-looking team from a small analytics start-up in Austria), formed an alliance. They joined forces and blended predictive models to form an über-team. With all the communal cooperation already taking place informally, it was time to make it official.

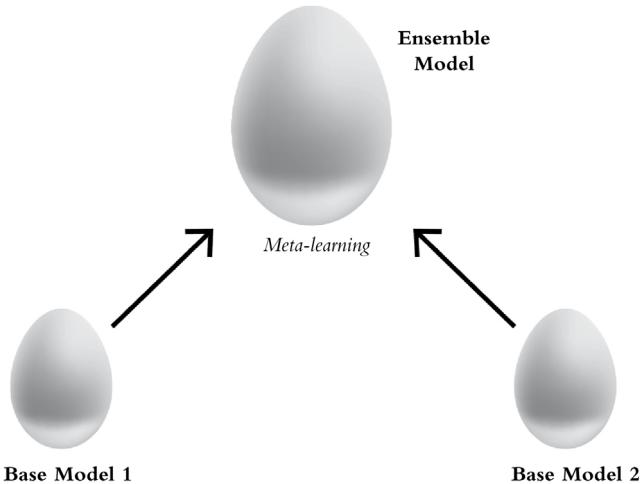
It was risky to team up. By sharing technology, the teams lost their mutual competitive edge against one another. If they won, they'd need to split the winnings. But if they didn't team up quickly enough, other teams could try the same tactic for the win.

It worked. The teams' predictive models were quite different from one another and, as hoped, the strengths of one model compensated for the weaknesses of the other. By integrating the models, they achieved a performance that enjoyed the best of both models. Only by doing so did the new über-team, BellKor in BigChaos, spring ahead far enough to qualify for—and win—the contest's annual progress prize of \$50,000.

META-LEARNING

Here's where the power to advance PA begins. Combining two or more sophisticated predictive models is simple: Just apply predictive modeling to

learn how to combine them. Since each model comes about from machine learning, this is an act of “learning on top of learning”—*meta-learning*.



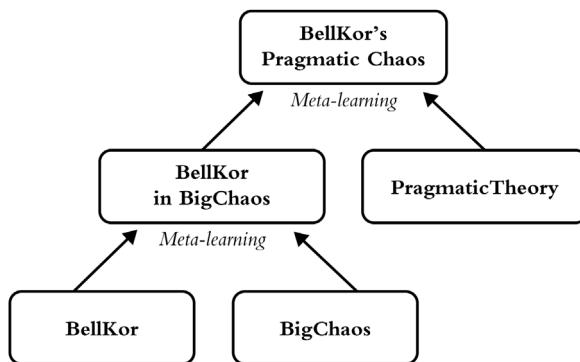
An ensemble of two predictive models.

Therefore, competitors turned collaborators with two distinct, intricate models that have been developed in very different ways don’t necessarily need to work that hard to combine them. Instead of digging in and thinking intensely to compare and contrast their theories and techniques, BigChaos team member Andreas Töscher told me, they let predictive modeling do the blending. They trained a new model that sits above the two existing models like a manager. This new *ensemble model* then considers both models’ predictions on a case-by-case basis. For certain cases, it can give more credence to model A rather than B, or the other way around. By so doing, the ensemble model is trained to predict which cases are weak points for each component model. There may be many cases where the two models are in agreement, but where there is disagreement, teaming the models together provides the opportunity to improve performance.

For the Netflix Prize, the dynamics of the gameplay had now changed, triggering a new flurry of merging and blending as teams consolidated, rolling up into bigger and better competitors. It was like the mergers and

acquisitions that take place among companies in a nascent, quickly developing industry.⁴

This merging and blending outplayed the ingenuity of our heroic lay team PragmaticTheory (the two Montrealers named Martin). But the team's success had gained the attention of its adversaries, and an invitation was extended by über-team BellKor in BigChaos to join and form an über-über-team. And so BellKor's Pragmatic Chaos came to be:



BellKor's Pragmatic Chaos made the grade. On June 26, 2009, it broke through the 10 percent barrier that qualified the super team for the \$1 million Netflix Prize.

A BIG FISH AT THE BIG FINISH

But it wasn't over yet. Per contest rules, this accomplishment triggered a 30-day countdown, during which all teams could continue to submit entries.

⁴ We've seen such corporate rollups in the PA industry itself, among software companies; for example, IBM bought SPSS, which had bought Integral Solutions Limited; SAS bought Teragram (text analytics); and Pitney Bowes bought Portrait Software, which had bought Quadstone.

An archnemesis had emerged, called none other than The Ensemble (not to be confused with the team that included John Elder, Ensemble Experts, which employed ensemble methods internally but did not involve combining separate teams). This rival gave BellKor's Pragmatic Chaos a serious run for the money by rolling together teams like mad. By the end, it was an amalgam of over 20 teams, one of which openly absorbed any and all teams that wished to join. By uploading its predictions, a joining team would be rewarded in proportion to the resulting improvement, the bump it contributed to the growing ensemble—but only if the overarching team won, of course. It was like the Borg from *Star Trek*, an abominable hive-like force that sucks up entire civilizations after declaring menacingly, “You will be assimilated!” A number of teams allowed themselves to be swallowed by this fish that ate the fish that ate the fish. After all, if you can't beat 'em, join 'em.

Although it combined the efforts of only three teams, BellKor's Pragmatic Chaos rallied to compete against this growing force. The 30 days counted down. Neck and neck, the two über-teams madly submitted new entries, tweaking, retweaking, and submitting again, even into the final hours and minutes of this multiple-year contest. Crowdsourcing competitions cultivate a heated push for scientific innovation, engendering focus and drive sometimes compared to that attained during wartime.

Time ran out. The countdown was over and the dust was settling. The contest administrators at Netflix went silent for a few weeks as they assessed and verified. They held yet another undisclosed set of data with which to validate the submissions and determine the final verdict. Here is the top portion of the final leaderboard:

Rank	Team Name	Best Test Score	Percentage Improvement	Best Submit Time
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22

Rank	Team Name	Best Test Score	Percentage Improvement	Best Submit Time
3	Grand Prize Team	0.8582	9.90	2009–07–10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009–07–10 01:12:31
5	Vandelay Industries!	0.8591	9.81	2009–07–10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009–06–24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009–05–13 08:14:09

BellKor's Pragmatic Chaos won by a nose. Its performance was so close to The Ensemble's that it was considered a quantitative tie in accord with the contest's posted rules. Because of this, the determining factor was *which of the tied entries had been submitted first*. At the very end of a multiple-year competition, BellKor's Pragmatic Chaos had uploaded its winning entry just 20 minutes before The Ensemble. The winning team received the cash and the other team received nothing. Netflix CEO Reed Hastings reflected, "That 20 minutes was worth a million dollars."

COLLECTIVE INTELLIGENCE

With most things, the average is mediocrity. With decision making, it's often excellence.
—James Surowiecki, *The Wisdom of Crowds*

Even competitions much simpler than a data mining contest can tap the wisdom held by a crowd. The magic of *collective intelligence* was lightheartedly demonstrated in 2012 at the Predictive Analytics World (PAW) conference. Charged with drawing attention to his analytics company on the

event's exposition floor, Gary Panchoo held a money-guessing contest. Here he is, collecting best guesses as to how many dollar bills are in the container:



The guessers as a group outsmarted every individual guess. The winner was only \$10 off the actual amount, \$362. But the average of the 61 guesses, \$365, was off by just \$3.

With no coordinated effort among the guessers, how could this be a common phenomenon? One way to look at it is that all people's overestimations and underestimations even out. If we assume that people guess too high as much as

they do too low, averaging cancels out these errors in judgment. No one person can overcome his or her own limited capacity—unless you’re a superhero, you can’t look at the container of dollars and be superconfident about your estimation. But across a group, the mistakes come out in the wash.

Uniting endows power. By coming together as a group, our limited capacities as individuals are overcome. Moreover, we no longer need to take on the challenging task of identifying the best person for the job. It doesn’t matter which person is smartest. A diverse mix best does the trick.

The *collective intelligence* of a crowd emerges on many occasions, as explored thoroughly by James Surowiecki in his book *The Wisdom of Crowds*. Examples include:

- *prediction markets*, wherein a group of people together estimate the prospects for a horse race, political event, or economic occurrence by way of placing bets (unfortunately, this adept forecasting method cannot usually scale to the domain of PA, in which thousands or millions of predictions are generated by a predictive model);
- the audience of the TV quiz show *Who Wants to Be a Millionaire?*, whom contestants may poll to weigh in on questions; and
- Google’s PageRank method, by which a Web page’s value and importance are informed by how many links people have created to point to the page.

Human minds aren’t the only things that can be effectively merged together. It turns out the aggregate effect emerging from a group extends also to *nonhuman* crowds—of predictive models.

THE WISDOM OF CROWDS . . . OF MODELS

The “wisdom of crowds” concept motivates ensembles because it illustrates a key principle of ensembling: Predictions can be improved by averaging the predictions of many.

—Dean Abbott, Abbott Analytics

Like a crowd of people, an ensemble of predictive models benefits from the same “collective intelligence” effect.⁵ Each model has its strengths and weaknesses. As with guesses made by people, the predictive scores produced by models are imperfect. Some will be too high and some too low. Averaging scores from a mix of models can wipe away much of the error. Why hire the best single employee when you can afford to hire a team whose members compensate for one another’s weaknesses? After all, models work for free; a computer uses practically no additional electricity to apply 100 models rather than just one.

Ensemble modeling has taken the PA industry by storm. It’s often considered the most important predictive modeling advancement to come to fruition in the first decade of this century. While its success in crowdsourcing competitions has helped bolster its credibility, the craft of ensembling pervades beyond that arena, both in commercial application and in research advancement.

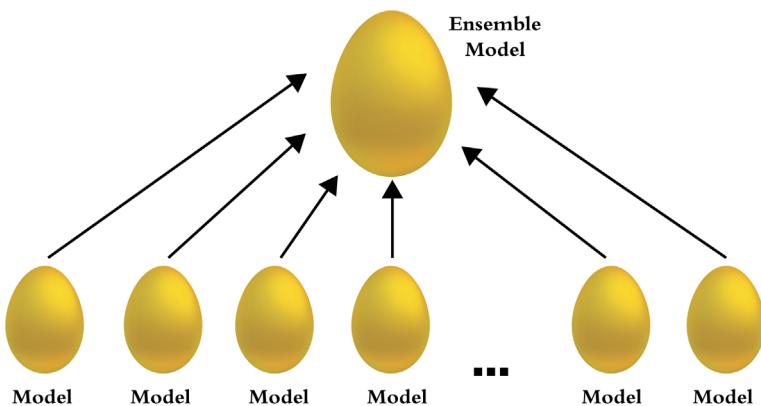
But increasing complexity is paradoxical to improved learning. An ensemble of models—which can grow to include thousands—is much more involved than a single model, so it’s a move away from the “keep it simple, stupid” (KISS) principle (aka Occam’s razor) that’s so critical to avoiding overlearning, as discussed in Chapter 4. Before ironing out this irony, let’s take a closer look at how ensemble models work.

A BAG OF MODELS

Leo Breiman, one of the Fab Four inventors of CART decision trees (detailed in Chapter 4), developed a leading method for ensemble models called *bagging* (short for *bootstrap aggregating*). The way it operates is practically self-evident. Make a bunch of models, a bagful. To predict, have each model make its prediction, and tally up the results. Each model gets to vote (voting is similar to averaging and in some cases is equivalent). The models are endowed with a key characteristic: diversity. Diversity is ensured by building

⁵ Dean Abbott, a leading PA consultant and frequent writer on the topic of ensemble models, brought this analogy to my attention.

each model on a different subset of the data, in which some examples are randomly duplicated so that they have a stronger influence on the model's learning process, and others are left out completely. Reflecting this random element, one variation on bagging that assembles a number of CART decision trees is dubbed *random forests*. (Doesn't this make a single tree seem "à la *CART*"?)

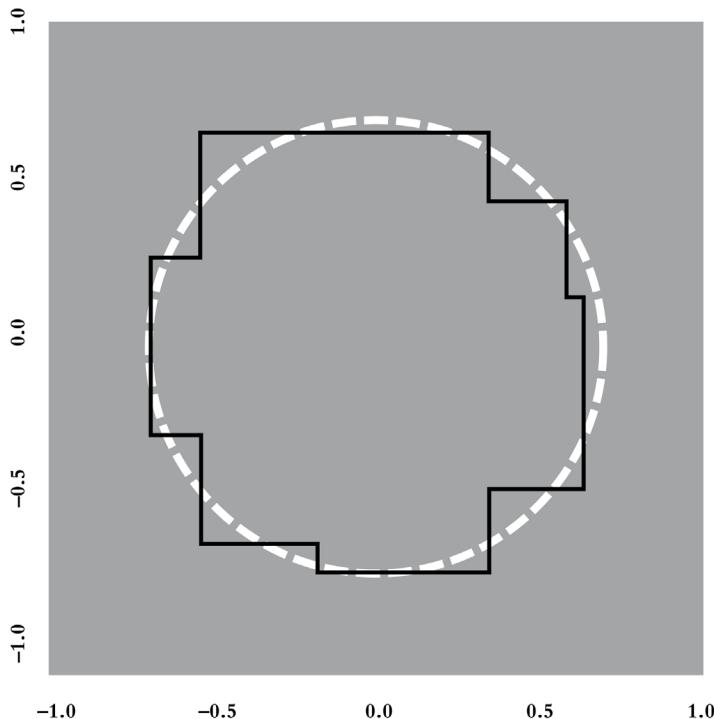


A group of models comes together to form an ensemble.

The idea of collecting models and having them vote is as simple and elegant as it sounds. In fact, other ensemble methods, all variations on the same theme, also sport friendly, self-descriptive names, including *bucket of models*, *bundling*, *committee of experts*, *meta-learning*, *stacked generalization*, and *TreeNet* (some employ voting and others meta-learn as for the Netflix Prize).

The notion of assembling components into a more complex, powerful structure is the very essence of engineering, whether constructing buildings and bridges or programming the operating system that runs your iPhone. Nobody must conceive of the entire massive structure at once—indeed, nobody can anyway. Tiered assembly makes architecting manageable.

An ensemble usually kicks a single model's butt. Check out this attempt by a single decision tree to model a circle:

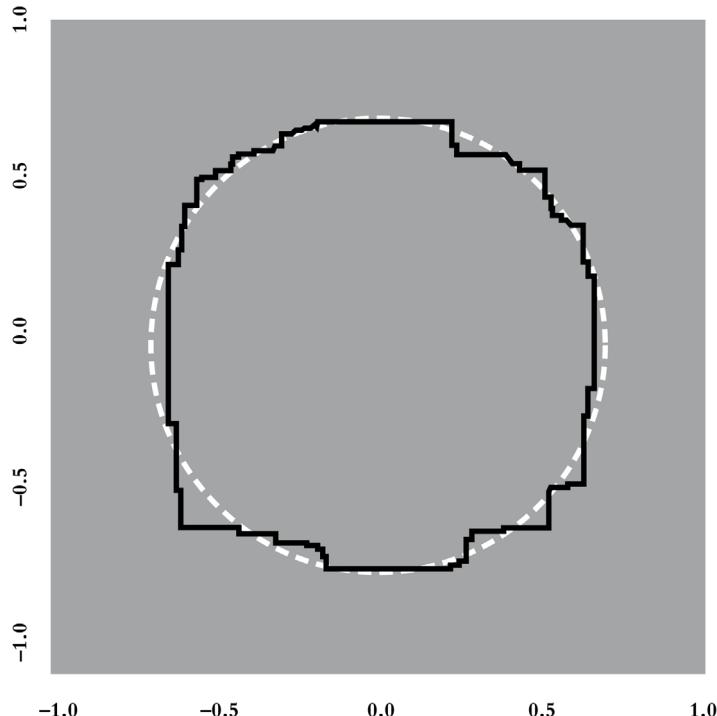


This and the following figure are reproduced with permission.⁶

In this experiment, a CART decision tree was trained over a data set that was manufactured to include positive and negative examples, inside and outside the circle, respectively. Because a decision tree can only compare the predictor variables (in this case, the x and y coordinates) to a fixed value and cannot perform any math on them, the tree's *decision boundary* consists only of horizontal and vertical lines. No diagonal or curvy boundaries are allowed. The resulting model does correctly label most points as to whether they're inside or outside the circle, but it's clearly a rough, primitive approximation.

⁶ John Elder and Greg Ridgeway, "Combining Estimators to Improve Performance," KDD Tutorial Notes, 1999.

Bagging a set of 100 CART trees generates a smoother, more refined model:⁷



Reproduced with permission.

⁷ While these visuals provide an intuitive view, real PA applications are usually difficult or impossible to view in this way. These examples are two-dimensional, since each case is defined only by the x and y coordinates. Predictive models normally work with dozens or hundreds of variables, in which case the decision boundary cannot be viewed with a two-dimensional diagram. Further, the reality behind the data that the predictive model is attempting to ascertain—in this manufactured example, a single circle—is unknown (if it were known, there would be no need for data analysis in the first place), and is generally more complex than a circle.

ENSEMBLE MODELS IN ACTION

Teams often use an ensemble model to win Kaggle contests.

—Anthony Goldbloom, founder and CEO of Kaggle

Whether assembled by the thousands or pasted together manually (as in the case when Netflix Prize teams joined forces), ensemble models triumph time after time. Research results consistently show that ensembles boost a single model’s performance in the general range of 5 to 30 percent, and that integrating more models into an ensemble often continues to improve it further. “The ensemble of a group of models is usually better than most of the individual models it’s made up of, and often better than them all,” says Dean Abbott.

Commercial deployment is expanding. Across this book’s Central Tables of PA examples, at least eight employed ensemble models: IBM (*Jeopardy!*-playing Watson computer), the IRS (tax fraud), the Nature Conservancy (donations), Netflix (movie recommendations), Nokia-Siemens (dropped calls), University of California, Berkeley (brain activity, to construct a moving image of what you’re seeing), U.S. Department of Defense (fraudulent government invoices), and U.S. Special Forces (job performance).

It seems too good to be true. With ensembles, we are consistently rewarded with better predictive models, often without any new math or formal theory. Is there a catch?

THE GENERALIZATION PARADOX: MORE IS LESS

Ensembles appear to increase complexity . . . so, their ability to generalize better seems to violate the preference for simplicity summarized by Occam’s Razor.

—John Elder, “The Generalization Paradox of Ensembles”

In Chapter 4 we saw that pursuing the heady goal of machine learning, *to learn without overlearning*, requires striking a careful balance. Building up a predictive model’s complexity so that it more closely fits the training data can go only so far. After a certain point, true predictive performance, as measured over a held-aside test set, begins to suffer.

Ensembles remain robust even as they become increasingly complex. They seem to be immune to this limitation, as if soaked in a magic potion against overlearning. John Elder, who humorously calls ensemble models a “secret weapon,” identified this phenomenon in a research paper and dubbed it “the generalization paradox of ensembles.”

John resolves the apparent paradox by redefining *complexity*, measuring it “by function rather than form.” Ensemble models look more complex—but, he asks, do they *act* more complex? Instead of considering a model’s structural complexity—how big it is or how many components it includes—he measures the *complexity of the overall modeling method*. He employs a measure called *generalized degrees of freedom*, which shows how adaptable a modeling method is, how much its resulting predictions change as a result of small experimental changes to the training data. If a small change in the data makes a big difference, the learning method may be brittle, susceptible to the whims of randomness and noise found within any data set. It turns out that this measure of complexity is *lower* for an ensemble of models than for individual models. Ensembles overadapt less. In this way, ensemble models exhibit *less* complex behavior, so their success in robustly learning without overlearning isn’t paradoxical after all.

Enter *The Ensemble Effect*. By simply joining models together, we enjoy the benefit of cranking up our model’s *structural complexity* while retaining a critical ingredient: robustness against overlearning.

The Ensemble Effect: *When joined in an ensemble, predictive models compensate for one another’s limitations so the ensemble as a whole is more likely to predict correctly than its component models are.*

THE SKY’S THE LIMIT

With the newfound power of ensemble models and the fervor to tackle increasingly grand challenges, what’s next? In the following chapter, PA takes on a tremendous one: competing on the TV quiz show *Jeopardy!*



CHAPTER 6

Watson and the *Jeopardy!* Challenge

How does Watson—IBM’s Jeopardy!-playing computer—work? Why does it need predictive modeling in order to answer questions, and what secret sauce empowers its high performance? How does the iPhone’s Siri compare? Why is human language such a challenge for computers? Is artificial intelligence possible?

January 14, 2011. The big day had come. David Gondek struggled to sit still, battling the butterflies of performance anxiety, even though he was not the one onstage. Instead, the spotlights shone down upon a machine he had helped build at IBM Research for the past four years. Before his eyes, it was launched into a battle of intellect, competing against humans in this country’s most popular televised celebration of human knowledge and cultural literacy, the quiz show *Jeopardy!*

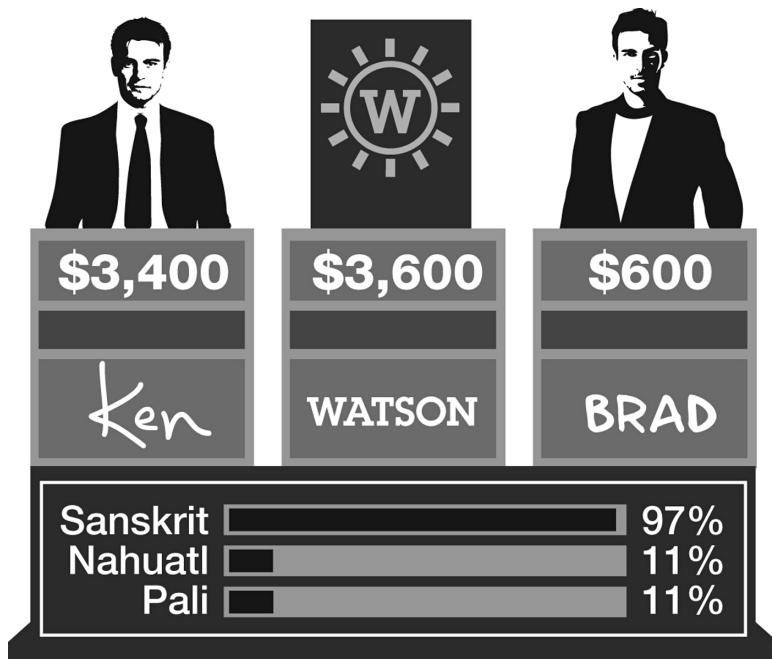
Celebrity host Alex Trebek read off a clue, under the category “Dialing for Dialects:”

VEDIC, DATING BACK AT LEAST
4,000 YEARS, IS THE EARLIEST
DIALECT OF THIS CLASSICAL
LANGUAGE OF INDIA

*

* *Jeopardy!* questions stamped with an asterisk were posed during Watson’s televised match.

Watson,¹ the electronic progeny of David and his colleagues, was competing against the two all-time champions across the game show's entire 26-year televised history. These two formidable opponents were of a different ilk, holding certain advantages over the machine, but also certain disadvantages. They were human.



Watson competes against two humans on *Jeopardy!*

¹ In this chapter, *Watson* refers to the highly specialized IBM computer that competed on *Jeopardy!* in 2011. Although the name *Watson* referred only to that specific system at that time, IBM has subsequently broadened its use of the word in its corporate branding strategy. *Watson* now also refers to at least three other loosely related initiatives: 1) IBM's promising research efforts applying some of the same analytical approaches developed for *Jeopardy!* within healthcare and other application areas; 2) *Watson Analytics*, a cloud-based business tool for predictive analytics and data visualization available in wide release; and 3) the technology one IBM partner credits for the function of its product, the CogniToys Dino, a toy dinosaur designed to conduct educational dialogues with children.

Watson buzzed in ahead of its opponents. Deaf and unable to hear Trebek's professional, confident voice, it had received the *Jeopardy!* clue as a transmission of typed text. The audience heard Watson's synthesized voice respond, phrasing it according to the show's stylistic convention of posing each answer in the form of a question. "What is Sanskrit?"²

For a computer, questions like this might as well be written in Sanskrit. Human languages like English are far more complex than the casual speaker realizes, with extremely subtle nuance and a pervasive vagueness we non-machines seem completely comfortable with. Programming a computer to work adeptly with human language is often considered the ultimate challenge of artificial intelligence (AI).

TEXT ANALYTICS

It was Greek to me.

—William Shakespeare

I'm completely operational, and all my circuits are functioning perfectly.

—HAL, the intelligent computer from *2001: A Space Odyssey* (1968)

Science fiction almost always endows AI with the capacity to understand human tongues. Hollywood glamorizes a future in which we chat freely with the computer like a well-informed friend. In *Star Trek IV: The Voyage Home* (1986), our heroes travel back in time to a contemporary Earth and are confounded by its primitive technology. Our brilliant space engineer Scotty, attempting to make use of a Macintosh computer, is so accustomed to computers understanding the spoken word that he assumes its mouse must be a microphone. Patiently picking up the mouse as if it were a quaint artifact, he jovially beckons, "Hello, computer!"

² In this chapter, I refer to each *Jeopardy!* clue as a *question* and each contestant response as an *answer*. It is a game of question answering, despite its stylistic convention of phrasing each contestant response in the form of a question beginning "what is" or "who is."

2001: A Space Odyssey's smart and talkative computer, HAL, bears a legendary, disputed connection in nomenclature to IBM (just take each letter back one position in the alphabet); however, author Arthur C. Clarke has strenuously denied that this was intentional. Ask IBM researchers whether their question-answering Watson system is anything like HAL, which goes famously rogue in the film, and they'll quickly reroute your comparison toward the obedient computers of *Star Trek*.

The field of research that develops technology to work with human language is *natural language processing* (NLP, aka *computational linguistics*). In commercial application, it's known as *text analytics*. These fields develop analytical methods especially designed to operate across the written word.

If data is all Earth's water, textual data is the part known as "the ocean." Often said to compose 80 percent of all data, it's everything we the human race know that we've bothered to write down. It's potent stuff—content-rich because it was generated with the intent to convey not just facts and figures, but human knowledge.

But text, data's biggest opportunity, presents the greatest challenge.

OUR MOTHER TONGUE'S TRIALS AND TRIBULATIONS

It is difficult to answer, when one does not understand the question.

—Sarek, Spock's father, in *Star Trek IV: The Voyage Home*

Let's begin with the relatively modest goal of grammatically deconstructing the Sanskrit question, repeated here:

**VEDIC, DATING BACK AT LEAST
4,000 YEARS, IS THE EARLIEST
DIALECT OF THIS CLASSICAL
LANGUAGE OF INDIA**



For example, consider how "of India" fits in. It's a prepositional phrase that modifies "this classical language." That may seem obvious to you, human reader, but if the final two words had been "of course," that phrase would

instead modify the main verb, “is” (or the entire phrase, depending on how you look at it).

Determining how each component such as “of India” fits in relies on a real understanding of words and the things in the world that they represent. Take the classic linguistic conundrum, “Time flies like an arrow.” Which is the main verb of the sentence? It is *flies* if you interpret the sentence as: “Time moves quickly, just as an arrow does.” But it could be *time* if you read it as the imperative, ordering you to “Measure the speed of flies as you would measure that of an arrow.”

The preferred retort to this aphorism, often attributed to Groucho Marx, is: “Fruit flies like a banana.” It’s funny and grammatically revealing. Suddenly *like* is now the verb, instead of a preposition.

“I had a car.” If the duration of time for which this held true was one year, I would say, “I had a car *for* a year.” But change one word and everything changes. “I had a baby.” If the duration of labor was five hours, you would say, “I had a baby *in* five hours,” not “*for* five hours.” The word choice depends on whether you’re describing a situation or an event, and the very meaning of the object—*car* or *baby*—makes the difference.

“I ate spaghetti with meatballs.” Meatballs were part of the spaghetti dish.

“I ate spaghetti with a fork.” The fork was instrumental to eating, not part of the spaghetti.

“I ate spaghetti with my friend Bill.” Bill wasn’t part of the spaghetti, nor was he instrumental to eating, although he was party to the eating event.

“I had a ball.” Great, you had fun.

“I had a ball but I lost it.” Not so much fun! But in a certain context, the same phrase goes back to being about having a blast:

Q: “How was your vacation and where is my video camera?”

A: “I had a ball but I lost it.”

In language, even the most basic grammatical structure that determines which words directly connect depends on our particularly human view of

and extensive knowledge about the world. The rules are fluid, and the categorical shades of meaning are informal.³

ONCE YOU UNDERSTAND THE QUESTION, ANSWER IT

How can a slim chance and a fat chance be the same, while a wise man and wise guy are opposites?

—Anonymous

Why does your nose run, and your feet smell?

—George Carlin

Beyond processing a question in the English language, a whole other universe of challenge lurks: *answering it*. Assume for a moment the language challenges have been miraculously met and the computer has gained the ability to “understand” a *Jeopardy!* question, to grammatically break it down, and to assess the “meaning” of its main verb and how this meaning fuses with the “meanings” of the other words such as the subject, object, and prepositional phrases to form the question’s overall meaning. Consider the following question, under the category “Movie Phone:”

**KEANU REEVES HAD A NOKIA
PHONE, BUT IT TOOK A LAND LINE
TO SLIP IN & OUT OF THIS, THE
TITLE OF A 1999 SCI-FI FLICK**

³ We face yet another “Mission Impossible” trying to get the computer to write instead of read. Generating human language trips up the naïve machine. I once received a voice-synthesized call from Blockbuster (a video rental chain of its day) reminding me of my rented movie’s due date. “This is a message for Eric the Fifth Siegel,” it said. My middle initial is V. Translation between languages also faces hazards. An often-cited example is that “The spirit is willing, but the flesh is weak,” if translated into Russian and back, could end up as “The vodka is good, but the meat is rotten.”

A perfect language-understanding machine could invoke a routine to search a database of movies for one starring Keanu Reeves in which a plot element involves using a land-line telephone to “get out of” something—that something also being the title of the movie (*The Matrix*). Even if the reliable transformation of question to database lookup were possible, how could any database be sure to include coverage of these kinds of abstract movie plot elements, which are subjective and open ended?

As another example that would challenge any database, consider this *Jeopardy!* question under the category “The Art of the Steal:”

THE ANCIENT “LION OF NIMRUD”
WENT MISSING FROM THIS
CITY’S NATIONAL MUSEUM IN
2003 (ALONG WITH A LOT OF
OTHER STUFF)



First, to succeed, the system must include the right information about each art piece, just as movie plot elements were needed for the *Matrix* question. IBM would have needed the foresight to include in a database of artworks whether, when, and where each item was stolen (for this item, the answer is Baghdad). Second, the system would also need to equate “went missing” with being stolen. That may be a reasonable interpretation regarding artwork, but if I said that my car keys went missing, we wouldn’t reach the same conclusion. How endlessly involved would a mechanical incarnation of human reason need to be in order to automatically make such distinctions? Written sources such as newspaper articles did in fact use a diverse collection of words to report this art carving’s *disappearance, looting, theft, or being stolen*.

Movies and artworks represent only the tip-top of a vast iceberg. *Jeopardy!* questions could fall within any domain, from the history of wine to philosophy to literature to biochemistry, and the answer required could be a person, place, animal, thing, year, or abstract concept. This unbounded challenge is called *open question answering*. Anything goes.

The old-school AI researcher succumbs to temptation and fantasizes about building a Complete Database of Human Knowledge. That researcher is fun

to chat with. He holds a grandiose view regarding our ability to reach for the stars by digging deep, examining our own inner cognitions, and expressing them with computer programs that mimic human reason and encode human knowledge. But someone has to break it to the poor fellow: This just isn't possible. As more pragmatic researchers concluded in the 1980s and 1990s, it's too large and too ill defined.

In reality, given these challenges, IBM concluded only 2 percent of *Jeopardy!* questions could be answered with a database lookup. The demands of open question answering reach far beyond the computer's traditional arena of storing and accessing data for flight reservations and bank records. We're going to need a smarter robot.

THE ULTIMATE KNOWLEDGE SOURCE

We are not scanning all those books to be read by people. We are scanning them to be read by an AI.

—A Google employee regarding Google's book scanning, as quoted by George Dyson in *Turing's Cathedral: The Origins of the Digital Universe*

A bit of good news: IBM didn't need to create comprehensive databases for the *Jeopardy!* challenge because the ultimate knowledge source already exists: *the written word*. I am pleased to report that people like to report; we write down what we know in books, Web pages, Wikipedia entries, blogs, and newspaper articles. All this *textual data* composes an unparalleled gold mine of human knowledge.

The problem is that these things are all encoded in human language, just like those confounding *Jeopardy!* questions. So the question-answering machine must overcome not only the intricacies and impossibilities of the question itself, but the same aspects of all the millions of written documents that may hold the question's answer.

Googling the question won't work. Although it's a human's primary means of seeking information from the Internet's sea of documents, Google doesn't hone down to an answer. It returns a long list of Web pages, each with hundreds or thousands of possible answers within. It is not designed for

the task at hand: identifying the singular answer to a question. Trying to use Google or other Internet search solutions to play *Jeopardy!*—for example, by doing a search on words from a question and answering with the document topic of the top search result—does not cut it. If only question answering were that easy to solve! This kind of solution answers only 30 percent of the questions correctly.

APPLE'S SIRI VERSUS WATSON

How does the iPhone personal assistant Siri compare with Watson? First introduced as the main selling point to distinguish the iPhone 4S from the preceding model, Siri responds to a broad, expanding range of voice commands and inquiries directed toward your iPhone.

Siri handles simpler language than Watson does: Users tailor requests for Siri knowing that they're speaking to a computer, whereas Watson fields *Jeopardy!*'s clever, wordy, information-packed questions that have been written with only humans in mind, without regard or consideration for the possibility that a machine might be answering. Because of this, Siri's underlying technology is designed to solve a different, simpler variant of the human language problem.

Although Siri responds to an impressively wide range of language usage, such that users can address the device in a casual manner with little or no prior instruction, people know that computers are rigid and will constrain their inquiries accordingly. Someone might request, "Set an appointment for tomorrow at 2 o'clock for coffee with Bill," but will probably not say, "Set an appointment with that guy I ate lunch with a lot last month who has a Yahoo! e-mail address," and will definitely not say, "I want to find out when my tall, handsome friend from Wyoming feels like discussing our start-up idea in the next couple weeks."

Siri flexibly handles relatively simple phrases that pertain to smartphone tasks such as placing calls, text messaging, performing Internet

(continued)

APPLE'S SIRI VERSUS WATSON (CONTINUED)

searches, and employing map and calendar functions (she's your *social techretary*).

Siri also fields general questions, but it does not attempt full open question answering. Invoking a system called WolframAlpha (accessible for free online), it answers simply phrased, fact-based questions via database lookup; the system can only provide answers calculated from facts that appear explicitly within its impressive, curated collection of structured, tabular database, such as:

The birthdates of famous people—How old was Elton John in 1976?

Astronomical facts—How long does it take light to go to the moon?

Geography—What is the biggest city in Texas?

Healthcare—What country has the highest average life expectancy?

One must phrase questions in a simple form, since WolframAlpha is designed first to compute answers from tables of data, and only secondarily to attempt to handle complicated grammar.

Siri processes spoken inquiries, whereas Watson processes transcribed questions. Researchers generally approach processing speech (*speech recognition*) as a separate problem from processing text. There is more room for error when a system attempts to transcribe spoken language before also interpreting it, as Siri does.

Siri includes a dictionary of humorous canned responses. If you ask Siri about its origin with, “Who’s your daddy?” it will respond, “I know this must mean something . . . everybody keeps asking me this question.” This should not be taken to imply adept human language processing. You might also ask, “What does the fox say?”

Siri and WolframAlpha’s question answering performance is continually improved by ongoing research and development efforts, guided in part by the constant flow of incoming user queries.

ARTIFICIAL IMPOSSIBILITY

*I'm wondering how to automate my wonderful self—
a wond'rrous thought that presupposes my own mental health.
Maybe it's crazy to think thought's so tangible, or that I can sing.
Either way, if I succeed, my machine will attempt the very same thing.*

—What artificial intelligence researchers sing in the shower

It is irresistible to pursue this because, as we pursue understanding natural language, we pursue the heart of what we think of when we think of human intelligence.

—David Ferrucci, Watson Principal Investigator, IBM Research

There's a fine line between genius and insanity.

—Oscar Levant

Were these IBM researchers certifiably nuts to take on this grand challenge, attempting to programmatically answer any *Jeopardy!* question? They were tackling the breadth of human language that stretches beyond the phrasing of each question to include a sea of textual sources, from which the answer to each question must be extracted. With this ambition, IBM had truly doubled down.

I would have thought success impossible. After witnessing the world's best researchers attempting to tackle the task through the 1990s (during which I spent six years in natural language processing research, as well as a summer at the same IBM Research center that bore Watson), I was ready to throw up my hands. Language is so tough that it seemed virtually impossible even to program a computer to answer questions within a limited domain of knowledge such as movies or wines. Yet IBM had taken on the unconstrained, open field of questions across any domain.

Meeting this challenge would demonstrate such a great leap toward humanlike capabilities that it invokes the “I” word: intelligence. A computer pulling it off would appear as magical and mysterious as the human mind. Despite my own 20-odd years studying, teaching, and researching all things artificial intelligence (AI), I was a firm skeptic. But this task required a leap so great that seeing it succeed might leave me, for the first time, agreeing that the term *AI* is justified.

AI is a loaded term. It blithely presumes a machine could ever possibly qualify for this title. Only with great audacity does the machine-builder bestow the honor of “intelligence” upon her own creations. Invoking the term comes across as a bit self-aggrandizing, since the inventor would have to be pretty clever herself to pull this off.

The *A* isn’t the problem—it’s the *I*. Intelligence is an entirely subjective construct, so AI is not a well-defined field. Most of its various definitions boil down to “making computers intelligent,” whatever that means! AI ordains no one particular capability as the objective to be pursued. In practice, AI is the pursuit of philosophical ideals and research grants.

What do God, Groucho Marx, and AI have in common? They’d never be a member of a club that would have them as a member. AI destroys itself with a logical paradox in much the same way God does in Douglas Adams’s *Hitchhiker’s Guide to the Galaxy*:⁴

“I refuse to prove that I exist,” says God, “for proof denies faith, and without faith I am nothing.”

“But,” says Man, “The Babel fish [which translates between the languages of interplanetary species] is a dead giveaway isn’t it? It could not have evolved by chance. It proves that you exist, and so therefore, by your own arguments, you don’t. QED.”

“Oh dear,” says God, “I hadn’t thought of that,” and promptly disappears in a puff of logic.

AI faces analogous self-destruction because, once you get a computer to do something, you’ve necessarily trivialized it. We conceive of as yet unmet “intelligent” objectives that appear big, impressive, and unwieldy, such as transcribing the spoken word (*speech recognition*) or defeating the world chess champion. They aren’t easy to achieve, but once we do pass such benchmarks, they suddenly lose their charm. After all, computers can manage only mechanical tasks that are well understood and well specified. You might be impressed by its lightning-fast speed, but its electronic

⁴ Watson’s avatar, its visual depiction shown on *Jeopardy!*, consists of 42 glowing, crisscrossing threads as an inside joke and homage that references the significance this number holds in Adams’s infamous *Hitchhiker’s Guide*.

execution couldn't hold any transcendental or truly humanlike qualities. If it's possible, it's not intelligent. Conversely, as famed computer scientist Larry Tesler succinctly put it, "Intelligence is whatever machines haven't done yet."

Suffering from an intrinsic, overly grandiose objective, AI inadvertently equates to "getting computers to do things too difficult for computers to do"—artificial impossibility.

LEARNING TO ANSWER QUESTIONS

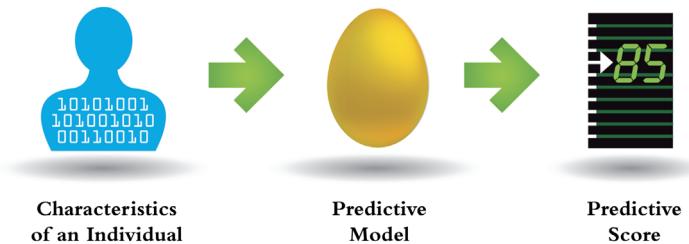
But in fact, IBM did face a specific, well-defined task: answering *Jeopardy!* questions. And if the researchers succeeded and Watson happened to appear intelligent to some, IBM would earn extra credit on this homework assignment.

As a rule, anticipating all possible variations in language is not possible. NLP researchers derive elegant, sophisticated systems to deconstruct phrases in English and other natural languages, based on deep linguistic concepts and specially designed dictionaries. But, implemented as computer programs, the methods just don't scale. It's always possible to find phrases that seem simple and common to us as humans, but trip up an NLP system. The researcher, in turn, broadens the theory and knowledge base, tweaking the system to accommodate more phrases. After years of tweaking, these hand-engineered methods still have light-years to go before we'll be chatting with our laptops just the same as with people.

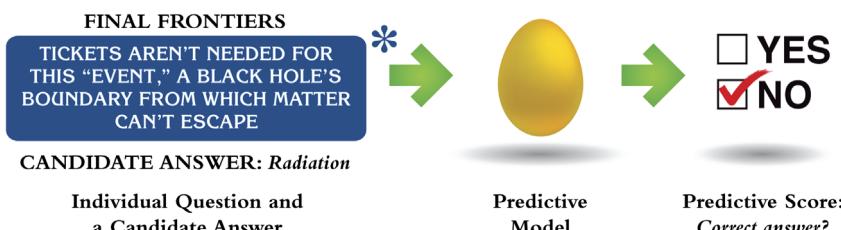
There's one remaining hope: Automate the researchers' iterative tweaking so it explodes with scale as a *learning* process. After all, that is the very topic of this book:

Predictive analytics (PA)—*Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions.*

Applying PA to question answering is a bit different from most of the examples we've discussed in this book. In those cases, the predictive model foretells whether a human will take a certain action, such as *click*, *buy*, *lie*, or *die*, based on things known about that individual:



IBM's Watson computer includes models that predict whether human experts would consider a *Jeopardy!* question/answer pair correct:



If the model is working well, it should give a low score, since *event horizon*, not *radiation*, is the correct answer (*Star Trek* fans will appreciate this question's category, "Final Frontiers"). Watson did prudently put a 97 percent score on *event horizon* and scored its second and third candidates, *mass* and *radiation*, at 11 percent and 10 percent, respectively. This approach frames question answering as a PA application:

PA APPLICATION: OPEN QUESTION ANSWERING

- What's predicted:** Given a question and one candidate answer, whether the answer is correct.
- What's done about it:** The candidate answer with the highest predictive score is provided by the system as its final answer.

Answering questions is not *prediction* in the conventional sense—Watson does not predict the future. Rather, its models "predict" the correctness of an answer. The same core modeling methods apply—but unlike other applications of predictive modeling, the unknown thing being "predicted" is

already known by some, rather than becoming known only when witnessed in the future. Through the remainder of this chapter, I employ this alternative use of the word *predict*, meaning, “to imperfectly infer an unknown.” You could even think of Watson’s predictive models as answering the predictive question: “Would human experts agree with this candidate answer to the question?” This semantic issue also arises for predicting clinical diagnosis (Central Table 4), fraud (Central Table 5), human thought (Central Table 8) and other areas—all marked with \mathcal{D} (for “detect”) in the Central Tables.

WALK LIKE A MAN, TALK LIKE A MAN

IBM needed data—specifically, example *Jeopardy!* questions—from which to learn. Ask and ye shall receive: Decades of televised *Jeopardy!* provide hundreds of thousands of questions, each alongside its correct answer (IBM downloaded these from fan websites, which post all the questions). This wealth of learning data delivers a huge, unprecedented boon for pushing the envelope in human language understanding. While most PA projects enjoy as data a good number of example individuals who either did or did not take the action being predicted (such as all those behaviors listed in the left columns of this book’s Central Tables of PA applications), most NLP projects simply do not have many previously solved examples from which to learn.

With this abundance of *Jeopardy!* history, the computer could learn to become humanlike. The questions, along with their answer key, contribute examples of human behavior: how people answer these types of questions. Therefore, this form of data fuels machine learning to produce a model that mimics how a human would answer, “Is this the right answer to this question?”—the learning machine models the human expert’s response. We may be too darn complex to program computers to mimic ourselves, but the model need not derive answers in the same manner as a person; with predictive modeling, perhaps the computer can find some innovative way to program itself for this human task, even if it’s done differently than by humans.

As Alan Turing famously asked, would a computer program that exhibits humanlike behavior qualify as AI? It's anthropocentric to think so, although I've been called worse.

But having extensive *Jeopardy!* learning data did not in itself guarantee successful predictive models, for two reasons:

1. Open question answering presents tremendous unconquered challenges in the realms of language analysis and human reasoning.
2. Unlike many applications of PA, success on *Jeopardy!* requires high predictive *accuracy*; The Prediction Effect from Chapter 1—*a little prediction goes a long way*—does not apply here.

When IBM embarked upon the *Jeopardy!* challenge in 2006, the state of the art fell severely short. The most notable source of open question answering data was a government-run competition called TREC QA (Text REtrieval Conference—Question Answering). To serve as training data, the contest provided questions that were much more straightforward and simply phrased than those on *Jeopardy!*, such as, “When did James Dean die?” Competing systems would pore over news articles to find each answer. IBM had a top-five competitor that answered 33 percent of those questions correctly, and no competing system broke the 50 percent mark. Even worse, after IBM worked for about one month to extend the system to the more challenging arena of *Jeopardy!* questions, it could answer only 13 percent correctly, substantially less than the 30 percent achieved by just using Internet search.

PUTTING ON THE PRESSURE

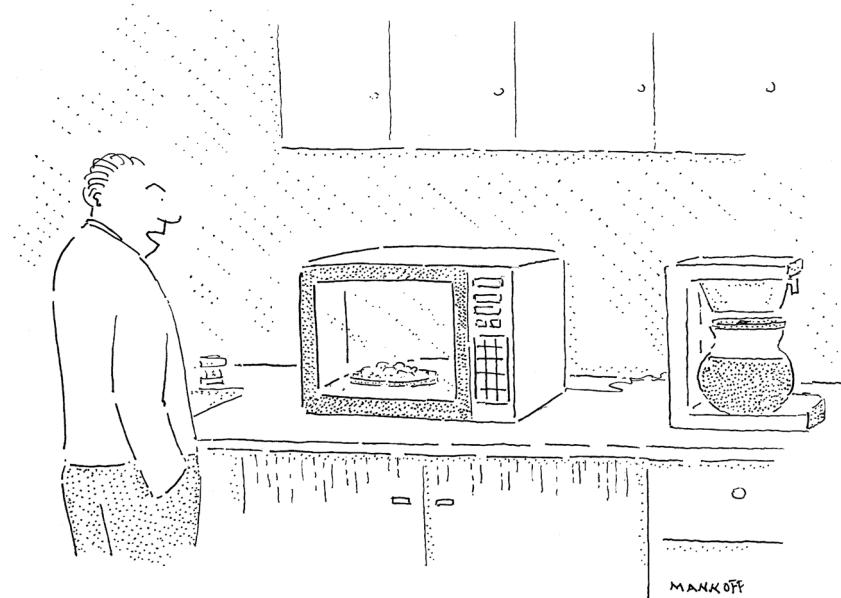
Scientists often set their own research goals. A grand challenge takes this control out of the hands of the scientist to force them to work on a problem that is harder than one they would pick to work on themselves.

—Edward Nazarko, Client Technical Advisor, IBM

Jumping on the *Jeopardy!* challenge, IBM put its name on the line. Following the 1997 chess match in which IBM's Deep Blue computer defeated then

world champion Garry Kasparov, the 2011 *Jeopardy!* broadcast pitted man against machine just as publicly, and with a renewed, healthy dose of bravado. A national audience of *Jeopardy!* viewers waited on the horizon.

As with all grand challenges, success was not a certainty. No precedent or principle had ensured it would be possible to fly across the Atlantic (Charles Lindbergh did so to win \$25,000 in 1927); walk on the moon (NASA's Apollo 11 brought people there in 1969, achieving the goal John F. Kennedy set for that decade); beat a chess grandmaster with a computer (IBM's Deep Blue in 1997); or even improve Netflix's movie recommendation system by 10 percent (2009, as detailed in the previous chapter).



"No, I don't want to play chess. I just want you to reheat the lasagna."

Reproduced with permission.

In great need of a breakthrough, IBM tackled the technical challenge with the force only a megamultinational enterprise can muster. With over \$92 billion in annual revenue and more than 412,000 employees worldwide, IBM is the third-largest U.S. company by number of employees. All told, its investment to develop Watson is estimated in the tens of millions of dollars,

including the dedication of a team that grew to 25 PhD's over four years at its T. J. Watson Research Center in New York (which, like the *Jeopardy!*-playing computer, was named after IBM's first president, Thomas J. Watson).

The power to push really hard does not necessarily mean you're pushing in the right direction. From where will scientific epiphany emerge? Recall the key innovation that the crowdsourcing approach to grand challenges helped bring to light, *ensemble models*, introduced in the prior chapter. It's just what the doctor ordered for IBM's *Jeopardy!* challenge.

THE ANSWERING MACHINE

David Gondek and his colleagues at IBM Research could overcome the daunting *Jeopardy!* challenge only with *synthesis*. When it came to processing human language, the state of the art was fragmented and partial—a potpourri of techniques, each innovative in conception but severely limited in application. None of them alone made the grade.

How does IBM's Watson work? It's built with ensemble models. Watson merges a massive amalgam of methodologies. It succeeds by fusing technologies. There's no secret ingredient; it's the overall recipe that does the trick. Inside Watson, ensemble models select the final answer to each question.

Before we more closely examine how Watson works, let's look at the discoveries made by a PA expert who analyzed *Jeopardy!* data in order to "program himself" to become a celebrated (human) champion of the game show.

MONEYBALLING *JEOPARDY!*

On September 21, 2010, a few months before Watson faced off on *Jeopardy!*, televisions across the land displayed host Alex Trebek speaking a clue tailored to the science fiction fan.

ZACHARY QUINTO SHOWED US
THE LOGIC AS THIS CHARACTER
IN 2009'S "STAR TREK"

Contestant Roger Craig avidly buzzed in. Like any technology PhD, he knew the answer was Spock.

As Spock would, Roger had taken studying to its logical extreme. *Jeopardy!* requires inordinate cultural literacy, the almost unattainable status of a Renaissance man, one who holds at least basic knowledge about pretty much every topic. To prepare for his appearance on the show, which he'd craved since age 12, Roger did for *Jeopardy!* what had never been done before. He *Moneyballed* it.

Roger optimized his study time with prediction. As a mere mortal, he faced a limited number of hours per day to study. He rigged his computer with *Jeopardy!* data. An expert in predictive modeling, he developed a system to learn from his performance practicing on *Jeopardy!* questions so that it could serve up questions he was likely to miss in order to efficiently focus his practice time on the topics where he needed it most. *He used PA to predict himself.*

PA APPLICATION: EDUCATION—GUIDED STUDYING FOR TARGETED LEARNING

- 1. What's predicted:** Which questions a student will get right or wrong.
- 2. What's done about it:** Spend more study time on the questions the student will get wrong.

This bolstered the brainiac for a breakout. On *Jeopardy!*, Roger set the all-time record for a single-game win of \$77,000 and continued on, winning more than \$230,000 during a seven-day run that placed him as the third-highest winning contestant (regular season) to date. He was invited back a year later for a “Tournament of Champions” and took its \$250,000 first place award. He estimates his own ability to correctly answer 90 percent of *Jeopardy!* questions, placing him among a small handful of all-time best players.

Analyzing roughly 211,000 *Jeopardy!* questions (downloaded as IBM did from online archives maintained by fans of the game show), Roger gained perspective on its knowledge domain. If you learn about 10,000 to 12,000 answers, he told me, you've got most of it covered. This includes countries,

states, presidents, and planets. But among many categories, you only need to go so far. Designed to entertain its audience, *Jeopardy!* doesn't get too arcane. So you only need to learn about the top cities, elements, movies, and flowers. In classical music, knowing a couple of dozen composers and the top few works of each will do the trick.

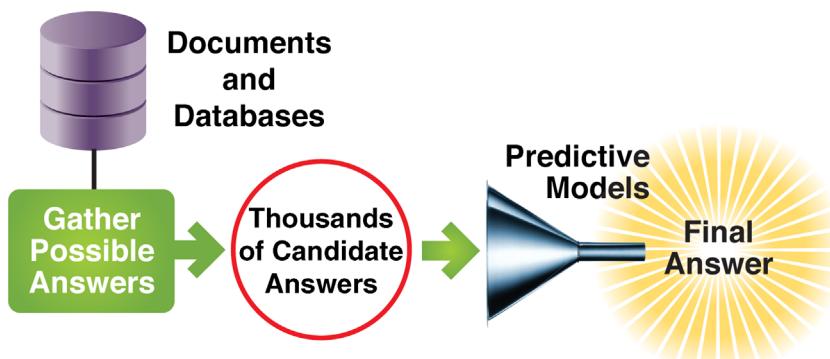
These bounds are no great relief to those pursuing the holy grail of open question answering. Predictive models often choose between only two options: *Will the person click, buy, lie, or die—yes or no?* As if that's not hard enough, for each question, Watson must choose between more than 10,000 possible answers.

The analytical improvement of human competitors was more bad news for Watson. Allowed by Roger to access his system, Watson's soon-to-be opponent Ken Jennings borrowed the study-guiding software while preparing for the big match, crediting it as "a huge help getting me back in game mode."

AMASSING EVIDENCE FOR AN ANSWER

Here's how Watson works. Given a question, it takes three main steps:

1. Collect thousands of *candidate answers*.
2. For each answer, amass *evidence*.
3. Apply predictive models to *funnel down*.



Predictive modeling has the final say. After gathering thousands of candidate answers to a question, Watson funnels them down to spit out the single answer scored most highly by a predictive model.

Watson gathers the answers and their evidence from sources that IBM selectively downloaded, a snapshot of a smart part of the Internet that forms Watson's base of knowledge. This includes 8.6 million written documents, consisting of 3.5 million Wikipedia articles (i.e., a 2010 copy of the entire English portion thereof), the Bible, other miscellaneous popular books, a history's worth of newswire articles, entire encyclopedias, and more. This is complemented by more structured knowledge sources such as dictionaries, thesauri, and databases such as the Internet Movie Database.

Watson isn't picky when collecting the candidate answers. The system follows the strategy of casting a wide, ad hoc net in order to ensure that the correct answer is in there somewhere. It rummages through its knowledge sources in various ways, including performing search in much the same way as Internet search engines like Google do (although Watson searches only within its own internal store). When it finds a relevant document, for some document types such as Wikipedia articles, it will grab the document's title as a candidate answer. In other cases, it will nab "answer-sized snippets" of text, as Watson developers call them. It also performs certain lookups and reverse lookups into databases and dictionaries to collect more candidate answers.

Like its fictional human namesake, the partner of Sherlock Holmes, Watson now faces a classic whodunit: Which of the many suspected answers is "guilty" of being the right one?⁵ The mystery can only be solved with diligent detective work in order to gather as much evidence as possible for or against each candidate. Watson pounds the pavement by once again surveying its sources.

With so many possible answers, uncertainty looms. It's a serious challenge for the machine to even be confident what *kind* of thing is being asked for. An actor? A movie? State capital, entertainer, fruit, planet, company, novel, president, philosophical concept? IBM determined that *Jeopardy!* calls for

⁵ Watson was not named after this fictional detective—it was named after IBM founder Thomas J. Watson.

2,500 different types of answers. The researchers considered tackling a more manageable task by covering only the most popular of these answer types, but it turned out that even if they specialized Watson for the top 200, it could then answer only half the questions. The range of possibilities is too wide and evenly spread for a shortcut to work.

ELEMENTARY, MY DEAR WATSON

Evidence counterattacks the enemy: *uncertainty*. To this end, Watson employs a diverse range of language technologies. This is where the state of the art in NLP comes into play, incorporating the research results from leading researchers at Carnegie Mellon University, the University of Massachusetts, the University of Southern California, the University of Texas, Massachusetts Institute of Technology, other universities, and, of course, IBM Research itself.

Sometimes, deep linguistics matters. Consider this question:

IN MAY 1898 PORTUGAL
CELEBRATED THE 400TH
ANNIVERSARY OF THIS
EXPLORER'S ARRIVAL IN INDIA

When David Gondek addressed Predictive Analytics World with a keynote, he provided an example phrase that could threaten to confuse Watson:

In May, Gary arrived in India after he celebrated his anniversary in Portugal.

So many words match, the system is likely to include *Gary* as a candidate answer. Search methods would love a document that includes this phrase. Likewise, Watson's evidence-seeking methods built on the comparison of words would give this phrase a high score—most of its words appear in the question at hand.

Watson needs linguistic methods that more adeptly recognize how words relate to one another so that it pays heed to, for example:

On the 27th of May 1498, Vasco da Gama landed in Kappad Beach.

Other than *in*, *of*, and *the*, only the word *May* overlaps with the question. However, Watson recognizes meaningful correspondences. Kappad Beach is in India. *Landed in* is a way to paraphrase *arrived in*. A 400th anniversary in 1898 must correspond to a prior event in 1498.

These matches establish support for the correct answer, Vasco da Gama. Like all candidate answers, it is evaluated for compatibility with the answer type—in this case, *explorer*, as determined from *this explorer* in the question. Vasco da Gama is indeed famed as an explorer, so support would likely be strong.

These relationships pertain to the very meaning of words, their *semantics*. Watson works with databases of established semantic relationships and seeks evidence to establish new ones. Consider this *Jeopardy!* question:

**IN CELL DIVISION, MITOSIS
SPLITS THE NUCLEUS AND
CYTOKINESIS SPLITS THIS LIQUID
CUSHIONING THE NUCLEUS**

Watson's candidate answers include organelle, vacuole, cytoplasm, plasma, and mitochondria. The type of answer sought being a liquid, Watson finds evidence that the correct answer, cytoplasm, makes the cut. It looks up a record listing cytoplasm as a fluid, and has sufficient evidence that fluids are often liquids to boost cytoplasm's score on that basis.

Here, Watson performs a daredevil stunt of logic. Reasoning as humans do in the wide-open domain of *Jeopardy!* questions is an extreme sport. Fuzziness pervades—for example, most reputable sources Watson may access would state all liquids are fluids, but some are ambiguous as to whether glass is definitely solid or liquid. Similarly, all people are mortal, yet infamous people have attained immortality. Therefore, a strict hierarchy of concepts just can't

apply. Because of this, as well as the vagueness of our languages' words and the difference context makes, databases of abstract semantic relationships disagree madly with one another. Like political parties, they often fail to see eye to eye, and a universal authority—an absolute, singular truth—to reconcile their differences simply does not exist.

Rather than making a vain attempt to resolve these disagreements, Watson keeps all pieces of evidence in play, even as they disagree. The resolution comes only at the end, when weighing the complete set of evidence to select its final answer to a question. In this way, Watson's solution is analogous to yours. Rather than absolutes, it adjusts according to context. Some songs are both a little bit country and a little bit rock and roll. With a James Taylor song, you could go either way.

On the other hand, keeping an “open mind” by way of this sort of flexible thinking can lead to embarrassment. Avoiding absolutes means playing fast and loose with semantics, leaving an ever-present risk of gaffes—that is, mistaken answers that seem all too obvious to us humans. For example, in Watson's televised *Jeopardy!* match, it faced a question under the category “U.S. Cities”:



ITS LARGEST AIRPORT IS
NAMED FOR A WORLD WAR II
HERO; ITS SECOND LARGEST,
FOR A WORLD WAR II BATTLE

Struggling, Watson managed to accumulate only scant evidence for its candidate answers, so it would never have buzzed in to attempt the question. However, this was the show's “Final *Jeopardy!*” round, so a response from each player was mandatory. Instead of the correct answer, Chicago, Watson answered with a city that's not in the United States at all, Toronto. Canadian game show host Alex Trebek poked a bit of fun, saying that he had learned something new.

English grammar matters. To answer some questions, phrases must be properly deconstructed. Consider this question:

**HE WAS PRESIDENTIALLY
PARDONED ON SEPT. 8, 1974**

In seeking evidence, Watson pulls up this phrase, which appeared in a *Los Angeles Times* article:

Ford pardoned Nixon on Sept. 8, 1974.

Unlike you, a computer won't easily see the answer must be Nixon rather than Ford. Based on word matching alone, this phrase provides equal support for Ford as it does for Nixon. Only by detecting that the question takes the passive voice, which means the answer sought is the receiver rather than the issuer of a pardon, and by detecting that the evidence phrase is in the active voice, is this phrase correctly interpreted as stronger support for Nixon than Ford.⁶

NLP's attempts to grammatically deconstruct don't always work. Complementary sources of evidence must be accumulated, since computers won't always grok the grammar. Language is tricky. Consider this question:

*

**MILORAD CAVIC ALMOST
UPSET THIS MAN'S PERFECT 2008
OLYMPICS, LOSING TO HIM BY
ONE HUNDREDTH OF A SECOND**

A phrase like this could be stumbled upon as evidence:

Sam was upset before witnessing the near win by Milorad Cavic.

If *upset* is misinterpreted as a passively voiced verb rather than an adjective, the phrase could be interpreted as evidence for Sam as the question's answer.

⁶ Watson employs as its main method for grammatical parsing the *English Slot Grammar*, by IBM's own researcher Michael McCord (I had the pleasure to use this tool for my doctoral research in the mid-1990s).

However, it was swimmer Michael Phelps who held on to his perfect 2008 Olympics performance. Even detecting the simplest grammatical structure of a sentence depends on the deep, often intangible meaning of words.

MOUNTING EVIDENCE

There's no silver bullet. Whether interpreting semantic relationships between words or grammatically deconstructing phrases, language processing is brittle. Even the best methods are going to get it wrong a lot of the time. This predicament is exacerbated by the clever, intricate manner in which questions are phrased on *Jeopardy!* The show's question writers have adopted a playful, informative style in order to entertain the TV viewers at home.

The only hope is to accumulate as much evidence as possible, searching far and wide for support of, or evidence against, each candidate answer. Every little bit of evidence helps. In this quest, diversity is the name of the game. An aggregate mass of varied evidence stands the best chance, since neither the cleverest nor the simplest method may be trusted if used solo. Fortunately, diversity comes with the territory: As with scientific research in general, the NLP researchers who developed the methods at hand each worked to distinguish their own unique contribution, intentionally differentiating the methods they designed from those of others.

Watson employs an assorted number of evidence routines that assess a candidate answer, including:

- **Passage search.** After inserting the candidate answer into the question to try it on for size (e.g., “*Nixon* was presidentially pardoned on Sept. 8, 1974”) and searching, do many matches come up? How many match word for word, semantically, and after grammatical deconstruction? What’s the longest similar sequence of words that each found phrase has in common with the question?
- **Popularity.** How common is the candidate answer?
- **Type match.** Does the candidate match the answer type called for by the question (e.g., entertainer, fruit, planet, company, or novel)? If it’s a person, does the gender match?

- **Temporal.** Was the candidate in existence within the question's time frame?
- **Source reliability.** Does the evidence come from a source considered reliable?

For each question, you never know which of these factors (and the hundreds of variations thereof that Watson measures) may be critical to arriving at the right answer. Consider this question:



CHILE SHARES ITS
LONGEST LAND BORDER
WITH THIS COUNTRY

Although the correct answer is Argentina, measures of evidence based on simple search show overwhelming support for Bolivia due to a certain border dispute well covered in news articles. Fortunately, enough other supporting evidence such as from logically matched phrases and geographical knowledge sources compensates and wins out, and Watson answers correctly.

Some may view this ad hoc smorgasbord of techniques as a hack, but I do not see it that way. It is true that the most semantically and linguistically intricate approaches are brittle and often just don't work. It can also be said that the remaining methods are harebrained in their oversimplicity. But a collective capacity *emerges* from this mix of components, which blends hundreds of evidence measurements, even if each alone is crude.⁷

⁷ Watson and PA in general are not designed to simulate how people think, predict, learn language, or answer questions. But it may be worth considering that, although as a human you experience a feeling of confidence and certainty in your answer to some questions, some components of the cognition that lead you there may be just as harebrained in isolation as Watson's components. Sometimes you have a specific recollection of the answer, as Watson does in certain cases of strong singular evidence. At other times, your confidence may only feel like a strong hunch, possibly based on a large number of weak factors.

The Ensemble Effect comes into full play: The sheer count and diversity of approaches make up for their individual weaknesses. As a whole, the system achieves operational proficiency on a previously unachievable, far-off goal: open human language question answering.

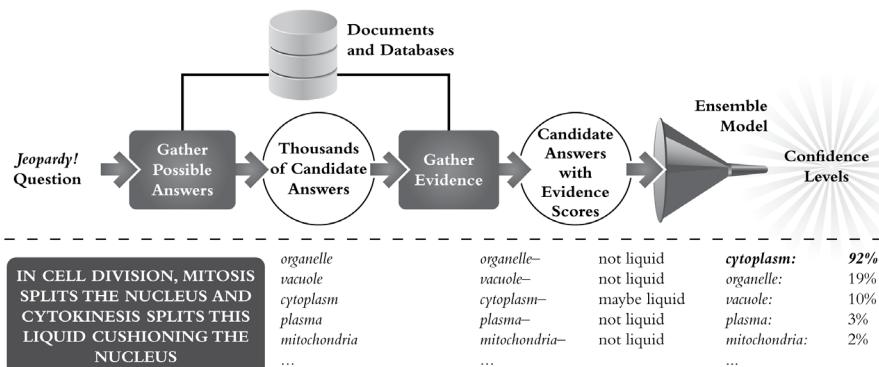
WEIGHING EVIDENCE WITH ENSEMBLE MODELS

There are two ways of building intelligence. You either know how to write down the recipe, or you let it grow itself. And it's pretty clear that we don't know how to write down the recipe. Machine learning is all about giving it the capability to grow itself.

—Tom Mitchell, founding chair of the world's first Machine Learning Department (at Carnegie Mellon University)

The key to optimally joining disparate evidence is machine learning. Guided by the answer key for roughly 25,000 *Jeopardy!* questions, the learning process discovers how to weigh the various sources of evidence for each candidate answer. To this end, David Gondek led the application of machine learning in developing Watson. He had his hands on the very process that brings it all together.

Synthesizing sources of evidence to select a single, final answer propels Watson past the limits of Internet search and into the formerly unconquered domain of question answering. Here's a more detailed overview:



An overview of key steps Watson takes for each question, with an example question and its candidate answers along the bottom.

An ensemble model selects the final answer from thousands of candidates.

As shown, Watson gathers candidate answers and then evidence for each candidate. Its ensemble model then scores each candidate answer with a confidence level so that it may be ranked relative to the other candidates. Watson then goes with the answer for which it holds the highest confidence, speaking it out loud when prompted to do so on *Jeopardy!*

PA APPLICATION: OPEN QUESTION ANSWERING

1. **What's predicted:** Given a question and one candidate answer, whether the answer is correct.
2. **What's done about it:** The candidate answer with the highest predictive score is provided by the system as its final answer.

AN ENSEMBLE OF ENSEMBLES

David led the design of Watson's innovative, intricate machine learning components, of which the ensembling of models is part and parcel. Moving from document search to open question answering demands a great leap, so the design is a bit involved. Watson incorporates ensembling in three ways:

1. **Combining evidence.** Hundreds of methods provide evidence scores for each candidate answer. Instead of tallying a simple vote across contributing evidence scores, as in some work with ensembles described in the prior chapter, the method takes it a step further by training a model to decide how best to fuse them together.⁸
2. **Specialized models by question type.** Watson has separate specialized ensemble models for specific question types, such as puzzle, multiple choice, date, number, translation, and etymology (about the

⁸ *Ensemble model* commonly refers to the combination of trained predictive models. However, many of Watson's evidence-scoring methods themselves were hand-designed by experts rather than developed by learning over data, so I am using the term a bit more broadly. But The Ensemble Effect is at play; the strengths of cooperating methods make up for one another's weaknesses.

history and origin of words) questions. In this way, Watson consists of *an ensemble of ensembles*.

3. Iterative phases of predictive models. For each question, Watson iteratively applies several phases of predictive models, each of which can compensate for mistakes made by prior phases. Each phase filters candidates and refines the evidence. The first phase filters down the number of candidate answers from thousands to about one hundred, and subsequent phases filter out more. After each phase's filtering, the evidence scores are reassessed and refined relative to the now-smaller list of candidate answers. A separate predictive model is developed for each phase so that the ranking of the shrinking list of candidates is further honed and refined. With these phases, Watson consists of *an ensemble of ensembles of ensembles*.

MACHINE LEARNING ACHIEVES THE POTENTIAL OF NATURAL LANGUAGE PROCESSING

Despite this complexity, Watson's individual component models are fairly straightforward: they perform a weighted vote of the evidence measures. In this way, some forms of evidence count more, and others count less. Although David tested various modeling methods, such as decision trees (covered in Chapter 4), he discovered that the best results for Watson came from another modeling technique called *logistic regression*, which weighs each input variable (i.e., measure of evidence), adds them up, and then formulaically shifts the resulting sum a bit for good measure.⁹

Since the model is made up of weights, the modeling process learns to literally *weigh the evidence* for each candidate answer. The predictive model filters out weak candidate answers by assigning them a lower score. It doesn't

⁹ After the weighted sum, logistic regression transforms the result with a function called an *S-curve* (aka *sigmoid squashing function*). The S-curve is designed to help predictive models with binary (twofold) target outputs, such as answering a yes/no question: *Is this answer correct, given the cumulative evidence?*

help Watson derive better candidate answers—rather, it cleans up the bulky mass of candidates, narrowing down to one final answer.

To this end, the predictive models are trained over 5.7 million examples of a *Jeopardy!* question paired with a candidate answer. Each example includes 550 predictor variables that summarize the various measures of evidence aggregated for that answer (therefore, the model is made of 550 weights, one per variable). This large amount of training data was formed out of 25,000 *Jeopardy!* questions. Each question contributes to many training examples, since there are many incorrect candidate answers. Both the correct and incorrect answers provide experience from which the system learns how to best weigh the evidence.

Watson leverages The Ensemble Effect, propelling the state of the art in language processing to achieve its full potential and conquer open question answering. Only by learning from the guidance provided by the archive of *Jeopardy!* questions was it possible to successfully merge Watson's hundreds of language-processing methods. Predictive modeling has the effect of measuring the methods' relative strengths and weaknesses. In this way, the system quantifies how much more important evidence from linguistically and semantically deep methods can be, and just how moderately simpler word-matching methods should be weighed so that they, too, may contribute to question answering.

With this framework, the IBM team empowered itself to incrementally refine and bolster Watson in anticipation of the televised *Jeopardy!* match—and moved the field of question answering forward. The system allows researchers to experiment with a continually growing range of language-processing methods: Just throw in a new language-processing technique that retrieves and reports on evidence for candidate answers, retrain the system's ensemble models, and check for its improved performance.

As David and his team expanded and refined the hundreds of evidence-gathering methods, returns diminished relative to efforts. Performance kept improving, but at a slower and slower pace. However, they kept at it, squeezing every drop of potential out of their brainshare and data, right up until the final weeks before the big match.

CONFIDENCE WITHOUT OVERCONFIDENCE

Both experts and laypeople mistake more confident predictions for more accurate ones. But overconfidence is often the reason for failure. If our appreciation of uncertainty improves, our predictions can get better too.

—Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail—but Some Don’t*

You got to know when to hold ’em, know when to fold ’em.

—Don Schlitz, “The Gambler” (sung by Kenny Rogers)

Jeopardy! wasn’t built for players with no self-doubt.

—Chris Jones, *Esquire* magazine

Besides answering questions, there’s a second skill each *Jeopardy!* player must hone: assessing self-confidence. Why? Because you get penalized by answering incorrectly. When a question is presented, you must decide whether to attempt to buzz in and provide an answer. If you do, you’ll either gain the dollar amount assigned to the question or lose as much.

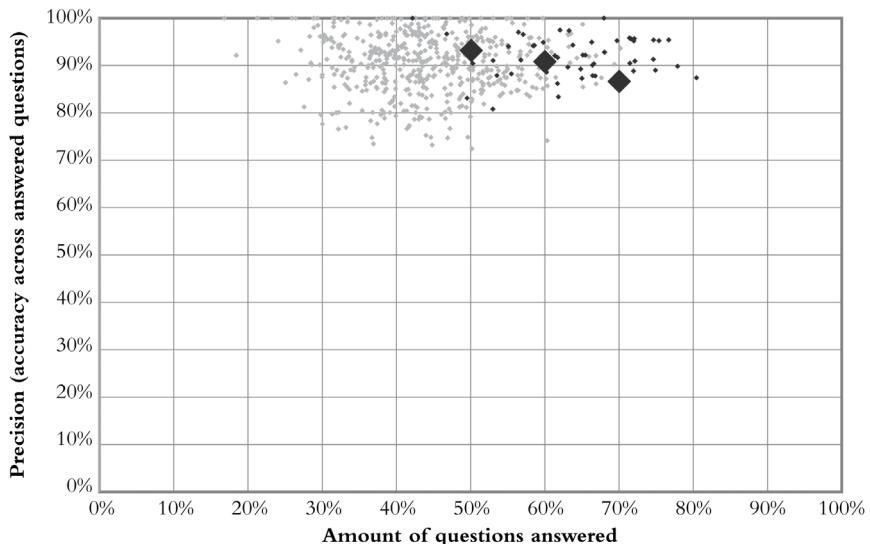
In this way, *Jeopardy!* reflects a general principle of life and business: *You need not do everything well; select the tasks at which you excel.* It’s the very practice of putting your best foot forward. In fact, many commercial uses of PA optimize on this very notion. Just as Watson must predict which questions it can successfully answer, businesses predict which customers will be successfully sold to—and therefore are worth the expenditure of marketing and sales resources.

Calculating a measure of self-confidence in each answer could be a whole new can of worms for the system. Is it a tall order to require the machine to “know thyself” in this respect?

David Gondek showed that this problem could be solved “for free.” The very same predictive score output by the models that serves to select the best answer also serves to estimate confidence in that answer. The scores are probabilities. For example, if a candidate answer with a score of 0.85 has a higher score than every other candidate, it will be Watson’s final answer, and Watson will consider its chance of being correct at 85 percent. As the IBM team put it, “Watson knows what it knows, and it knows what it doesn’t know.”

Watching Watson's televised *Jeopardy!* matches, you can see these self-confidence scores in action. For each question, Watson's top three candidate answers are displayed at the bottom of your TV screen along with their confidence scores (for example, see the second figure in this chapter). Watson bases its decision to buzz in on its top candidate's score, plus its position in the game relative to its opponents. If it is behind, it will play more aggressively, buzzing in even if the confidence is lower. If ahead in the game, it will be more conservative, buzzing in to answer only when highly confident.

A player's success depends not only on how many answers are known, but on his, her, or its ability to assess self-confidence. With that in mind, here's a view that compares *Jeopardy!* players:



Jeopardy! player performances. Each dot signifies a winner's game (the dark dots represent Ken Jennings's games). The three large diamonds represent the per-game performance Watson can achieve.¹⁰

¹⁰ Graph adapted from D. Ferrucci et al., "Building Watson: An Overview of the DeepQA Project," *AI Magazine* 31, no. 3 (2010), 59–79.

Players strive for the top right of this graph. Most points on the graph depict the performance of an individual human player. The horizontal axis indicates what proportion of questions they successfully buzzed in for, and the vertical axis tells us, for those questions they answered, how often they were correct. Buzzing in more would put you further to the right, but would challenge you with the need to know more answers.

Human *Jeopardy!* winners tend toward the top, since they usually answer correctly, and some also reach pretty far to the right. Each light gray dot represents the performance of the winner of one game. The impressively positioned dark gray dots that stretch further to the right represent the outstanding performance of champion player Ken Jennings, whose breathtaking streak of 74 consecutive wins in 2004 demonstrated his prowess. He is one of the two champions against whom Watson was preparing to compete.

Watson performs at the level of human experts. Three example points (large diamonds) are shown to illustrate Watson's potential performance. When needed, Watson sets itself to buzz in more often, assuming an aggressive willingness to answer even when confidence is lower. This moves its performance to the right and, as a result, also a bit down. Alternatively, when playing more conservatively, fewer questions are attempted, but Watson's answer is more often correct—precision is higher (unlike politics, on this graph left is more conservative).

Human sweat empowered Watson's human level of performance. The machine's proficiency is the product of four painstaking years of perseverance by the team of researchers.¹¹

THE NEED FOR SPEED

There was one more requirement. Watson had to be fast.

¹¹ The industry is taken with Watson. A Predictive Analytics World keynote address by Watson's machine learning leader, David Gondek, dazzled a ballroom of industry insiders, who on average rated the speech's content at an unmatched 4.7 out of 5 in a subsequent poll.

A *Jeopardy!* player has only a few seconds to answer a question, but on a single computer (e.g., 2.6 gigahertz), determining an answer can take a couple of hours. It's a lengthy process because Watson employs hundreds of methods to search a huge number of sources, both to accrue candidate answers and to collect evidence measurements for each one. It then predictively scores and ranks the candidates by applying the series of predictive models (I refer here only to the deployed use of Watson to play *Jeopardy!*, after the machine learning process is completed and the models are being employed without further learning).

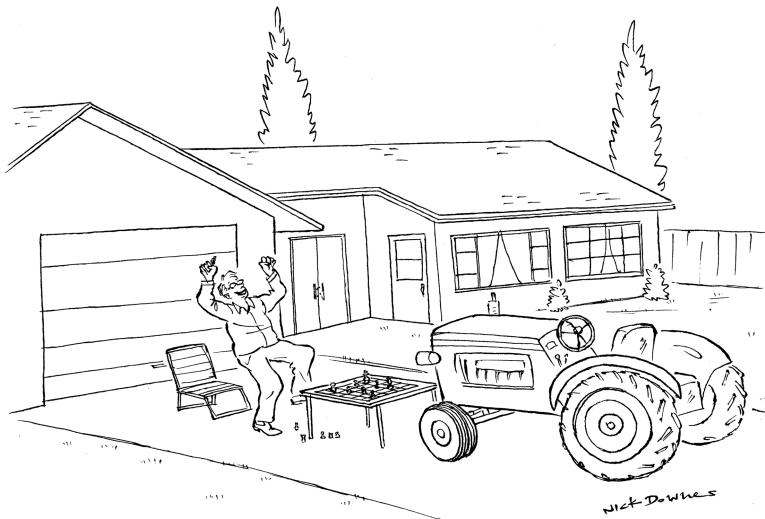
To make it thousands of times faster, Watson employs thousands of CPUs. This supercomputer clobbers bottlenecks and zips along, thanks to a cluster of 90 servers consisting of 2,800 core processors. It handles 80 trillion operations per second. It favors 15 terabytes of RAM over slower hard-drive storage. The cost of this hardware brawn is estimated to come to \$3 million, a small fraction of the cost to develop its analytical software brains.

Having thousands of CPUs means that thousands of tasks can be done simultaneously, in parallel. Watson's process lends itself so amenably to taking advantage of this hardware by way of distribution into contemporaneous subtasks that the research team considers it *embarrassingly parallel*. For example, each evidence-seeking, language-processing routine can be assigned to its own processor.

Better is bigger. To assemble Watson, IBM crated in a mammoth configuration of hardware, about 10 refrigerators' worth. Watson didn't go to *Jeopardy!*; *Jeopardy!* came to Watson, setting up a temporary game show studio within IBM's T. J. Watson Research Center.

DOUBLE *JEOPARDY!*—WOULD WATSON WIN?

Watson was not sure to win. During sparring games against human champions, Watson had tweaked its way up to a 71 percent win record. It didn't always win, and these trial runs didn't pit it against the lethal competition it was preparing to face on the televised match: all-time leading *Jeopardy!* champions Ken Jennings and Brad Rutter.



"Once again, man beats machine!"

Reproduced with permission.

The *Jeopardy!* match was to gain full-scale media publicity, exposing IBM's analytical prowess or failure. The top-rated quiz show in syndication, *Jeopardy!* attracts nearly 9 million viewers every day and would draw an audience of 34.5 million for this special man-versus-machine match. If the massive popularity of *Jeopardy!* put on the pressure, so too was it the only reason this grand challenge might be doable. As the United States' greatest pop culture institution of human knowledge, *Jeopardy!*'s legacy provided the treasure trove of historical question/answer pairs from which Watson learns.

Beyond impressing or disappointing your average home viewer, Watson's impending performance held enormous professional ramifications. Within both the practical realm of information technology and the research world of artificial intelligence, IBM had loudly proclaimed that it was prepared to run a three-and-a-half-minute mile. After the immense investment, one can only imagine the seething pressure the research team must have felt from the powers at IBM to defend the corporate image and ensure against public humiliation. At this juncture, the researchers saw clear implications for their scientific careers as well as for science itself.

During its formative stages, Watson's most humorous mistakes entertained, but threatened to embarrass IBM on national TV. Under the category "The Queen's English":

GIVE A BRIT A TINKLE
WHEN YOU GET INTO TOWN AND
YOU'VE DONE THIS

Watson said: *urinate* (correct answer: call on the phone).

Under the category "New York Times Headlines":

AN EXCLAMATION POINT
WAS WARRANTED FOR THE
“END OF” THIS! IN 1918

Watson said: *a sentence* (correct answer: World War I).

Under the category "Boxing Terms":

RHYMING TERM FOR
A HIT BELOW THE BELT

Watson said: *wang bang* (correct answer: low blow).

The team rallied for the home stretch. Watson principal investigator David Ferrucci, who managed the entire initiative, moved everyone from their offices into a common area he considered akin to a war room, cultivating a productive but crisislike level of eustress. Their lives were flipped on their

heads. David Gondek moved temporarily into a nearby apartment to eliminate his commuting time. The team lived and breathed open question answering. “I think I dream about *Jeopardy!* questions now,” Gondek said. “I have nightmares about *Jeopardy!* questions. I talk to people in the form of a question.”

JEOPARDY! JITTERS: DEPLOYING A PROTOTYPE

There's no such thing as human error. Only system error.

—Alexander Day Chaffee, software architect

Core Watson development team member Jennifer Chu-Carroll tried to stay calm. “We knew we probably were gonna win, but . . . what if we did the math wrong for some reason and lost by a dollar instead of won by a dollar?” There were provisions in their agreement with the *Jeopardy!* producers for do-overs in the case of a hardware crash (the show was taped, not broadcast live, and like any computer, sometimes you need to turn off Watson and then start it back up again). However, if Watson spat out an embarrassing answer due to a software bug without crashing, nothing could be done to take it back. This was going to national television.

Groundbreaking deployments of new technology—whether destined to be in orbit or intelligent—risk life and limb, not only because they boldly go where no one has gone before, but because they launch a prototype. Moon-bound *Apollo 11* didn’t roll off the assembly line. It was the first of its kind. The Watson system deployed on *Jeopardy!* was beta. Rather than conducting the established, sound process of “productizing” a piece of software for mass distribution, this high-speed, real-time behemoth was constructed not by software engineers who build things, but by the same scientific researchers who designed and developed its analytical capabilities. On the software side, the deployed system and the experimental system were largely one and the same. There was no clear delineation between some of the code they used for iterative, experimental improvement with machine learning and code within the deployed system. Of course, these were world-class researchers, many

with software design training, but the pressure mounted as these scientists applied virtual hammer to nail to fashion a vessel that would propel their laboratory success into an environment of high-paced, unforeseen questions.

Shedding their lab coats for engineering caps, the team members dug in as best they could. As David Gondek told me, changes in Watson's code continued even until and including the very day before the big match, which many would consider a wildly unorthodox practice in preparing for a mission-critical launch of software. Nobody on the team wanted to be the programmer who confused metric and English imperial units in their code, thus crashing NASA's Mars Climate Orbiter, as took place in 1998 after a \$327.6 million, nine-month trip to Mars. Recall the story of the Netflix Prize (see Chapter 5), which was won in part by two nonanalysts who found that their expertise as professional software engineers was key to their success.

The brave team nervously saw Watson off to meet its destiny. The training wheels were off. Watson operates on its own, self-contained and disconnected from the Internet or any other knowledge source. Unlike a human *Jeopardy!* player, the one connection it does need is an electrical outlet. It's scary to watch your child fly from the nest. Life has no safety net.

As a machine, Watson was artificial. The world would now witness whether it was also intelligent.

FOR THE WIN

You are about to witness what may prove to be an historic competition.

—Alex Trebek

If functional discourse in human language qualifies, then the world was publicly introduced to the greatest singular leap in artificial intelligence on February 14, 2011.

As the entertainment industry would often have it, this unparalleled moment in scientific achievement was heralded first with Hollywood cheese, and only secondarily with pomp and circumstance. After all, this

was a populist play. It was, in a sense, the very first conversant machine ever, and thus potentially easier for everyone to relate to than any other computer. Whether perceived as *Star Trek*-ian electronic buddy or HAL-esque force to be reckoned with, 34.5 million turned on the TV to watch it do its thing.

The *Jeopardy!* theme song begins to play,¹² and a slick, professional voice manically declares, “From the T. J. Watson Research Center in Yorktown Heights, New York, this is *Jeopardy!*, the IBM Challenge!”

When colleagues and I watch the footage, there’s a bit of culture shock: We’re looking for signs of AI, and instead see glitzy show business. But this came as no surprise to the members of Team Watson seated in the studio audience, who had been preparing for *Jeopardy!* for years.

Once the formalities and introductions to Watson pass, the show moves along jauntily as if it’s just any other episode, as if there is nothing extraordinary about the fact that one of the players spitting out answer after answer is not an articulate scholar with his shirt buttoned up to the top, but instead a robot with a synthetic voice straight out of a science fiction movie.

But for David Gondek and his colleagues it was anything but ordinary. The team endured a nail-biting day during the show’s recording, one month before its broadcast. Watching the two-game match, which was televised over a three-day period, you see dozens of questions fly by. When the camera turns for audience reactions, it centers on the scientists, David Ferrucci, David Gondek, Jennifer Chu-Carroll, and others, who enjoy moments of elation and endure the occasional heartache.

On this day, Machine triumphed over Man. Watson answered 66 questions correctly and eight incorrectly. Of those eight, only the answer that categorized Toronto as a U.S. city was considered a gaffe by human standards. The example questions covered in this chapter marked with an asterisk (“*”) were fielded by Watson during the match (all correctly except

¹² This well-known tune is a simple exercise in major fifths composed by Merv Griffin, *Jeopardy!*’s creator. In contradiction with what some consider a mind-numbing quality, the song’s title is the same as the IBM motto coined by the company’s founder, Thomas Watson: “Think.”

the one answered with Toronto). The final scores, measured in *Jeopardy!* as dollars, were Watson: \$77,147, Jennings: \$24,000, and Rutter: \$21,600.¹³

Prompted to write down his answer to the match's final question, Ken Jennings, invoking a *Simpsons* meme originating from an H. G. Wells movie, appended an editorial: "I, for one, welcome our new computer overlords." He later ruminated, "Watson has lots in common with a top-ranked human *Jeopardy!* player: It's very smart, very fast, speaks in an uneven monotone, and has never known the touch of a woman."

AFTER MATCH: HONOR, ACCOLADES, AND AWE

I would have thought that technology like this was years away, but it's here now. I have the bruised ego to prove it.

—Brad Rutter

This was to be an away game for humanity, I realized.

—Ken Jennings

Maybe we should have toned it down a notch.

—Sam Palmisano, then CEO, IBM

One million-dollar first place award for the *Jeopardy!* match? Check (donated to charities). American Technology Awards' "Breakthrough Technology of the Year Award"? Check. *R&D* magazine "Innovator of the Year" award? Check.

Webby "Person of the Year" award? Unexpected, but check.

Riding a wave of accolades, IBM is working to reposition components of Watson and its underlying question-answering architecture, which the company calls *DeepQA*, to serve industries such as healthcare and finance. Consider medical diagnosis. The wealth of written knowledge is so great, no doctor could read it all; providing a ranked list of candidate diagnoses for each

¹³ This strong lead was due at least in part to the speed with which Watson could buzz in to answer questions, although that issue is involved and debated; it is complicated to truly level the playing field when human and machine compete.

patient could mean doctors miss the right one less often. Guiding the analysis of knowledge sources by learning from training data—answers in the case of *Jeopardy!* and diagnoses in the case of healthcare—is a means to “capture and institutionalize decision-making knowledge,” as Robert Jewell of IBM Watson Solutions put it to me.

IAMBIC IBM AI

Is Watson intelligent? The question presupposes that such a concept is scientific in the first place. The mistake has been made, as proselytizers have often “over-souled” AI (credit for this poignant pun goes to Eric King, president of the consultancy he dubbed with the double entendre The Modeling Agency). It’s easy to read a lot into the thing. Case in point: I once designed a palindrome-generation system (a palindrome reads the same forward and backward) when teaching the AI course at Columbia University that spontaneously derived “Iambic IBM AI.” This one is particularly self-referential in that its meter is iambic.

Some credit Watson with far too much smarts. A guard working at IBM’s research facility got David Gondek’s attention as he was leaving for the day. Since this was a machine that could answer questions about any topic, he suggested, why not ask it who shot JFK?

Strangely, even technology experts tend to answer this philosophical question with a strong opinion in one direction or the other. It’s not about right and wrong. Waxing philosophical is a dance, a wonderful, playful pastime. I like to join in the fun as much as the next guy. Here are my thoughts:

Watching Watson rattle off one answer after another to diverse questions laced with abstractions, metaphors, and extraneous puns, I am dumbfounded. It is the first time I've felt compelled to anthropomorphize a machine in a meaningful way, well beyond the experience of suspending disbelief in order to feel fooled by a magic trick. To me, Watson looks and feels adept, not just with information but with knowledge. My perceptions endow it with a certain capacity to cogitate. It's a sensation I never thought I'd have cause to experience in my lifetime. To me, Watson is the first artificial intelligence.

If you haven't done so, I encourage you to watch the *Jeopardy!* match (see the Notes at www.PredictiveNotes.com for a YouTube link).

PREDICT THE RIGHT THING

Predictive models are improving and achieving their potential, but sometimes predicting what's going to happen misses the point entirely. Often, an organization needs to decide what next action to take. One doesn't just want to predict what individuals will do—one wants to know what to do about it. To this end, *we've got to predict something other than what's going to happen—something else entirely.* Turn to the next chapter to find out what.



CHAPTER 7

Persuasion by the Numbers

How Telenor, U.S. Bank, and the Obama Campaign Engineered Influence

What is the scientific key to persuasion? Why does some marketing fiercely backfire? Why is human behavior the wrong thing to predict? What should all businesses learn about persuasion from presidential campaigns? What voter predictions helped Obama win in 2012 more than the detection of swing voters? How could doctors kill fewer patients inadvertently? How is a person like a quantum particle? Riddle: What often happens to you that cannot be perceived and that you can't even be sure has happened afterward—but that can be predicted in advance?

In her job in Norway, Eva Helle stood guard to protect one of the world's largest cell phone carriers from its most dire threat. Her company, Telenor, had charged her with a tough assignment because, as it happens, the mobile business was about to suddenly turn perilous.

A new consumer right exerted new corporate strain: Mobile phone numbers became portable. Between 2001 and 2004, most European countries passed legislation to mandate that, if you switch to another wireless service provider, you may happily bring your phone number along with you—you need not change it (the United States did this as well; Canada, a few years later).

As customers leaped at the chance to leave, Eva faced an old truth. You just never know how fickle people are until they're untied. The consumer gains power, and the corporation pays a price.

But, as Eva and her colleagues would soon learn, the game had changed even more than they realized. Their method to woo customers and convince

them to stay had stopped working. A fundamental shift in how customers respond to marketing forced Eva to reconsider how things were done.

CHURN BABY CHURN

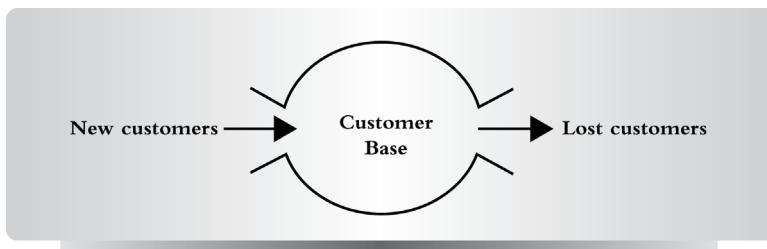
Before this change, Telenor had been successfully applying the industry's leading technique to hold on to its cell phone subscribers—a technique that applies predictive analytics (PA):

PA APPLICATION: CUSTOMER RETENTION WITH *CHURN MODELING*

- 1. What's predicted:** Which customers will leave.
- 2. What's done about it:** Retention efforts target at-risk customers.

Churn modeling may be the hottest marketing application of PA, and for good reason. Any seasoned executive will tell you retention is all-important because it's usually cheaper to convince a customer to stay than to acquire a new one.

Picture customer turnover as air flowing into and out of a balloon:



Retaining more customers is akin to clamping down on the nozzle on the right. Lessening the rate of loss just a bit, the balloon blossoms, magnifying its rate of expansion—that is, the growth rate of the company's customer base. This growth is the *raison d'être* of business.

Prediction and proaction are musts. Persuading someone to stay often sets a mobile carrier back a free phone or a hefty discount. A company must target this generosity where it's needed: those customers predicted to leave. Like most major cell phone carriers, Telenor had been enjoying a clear win with churn modeling.¹

What could possibly go wrong?

¹ This book's Central Table 2 lists several more examples of applied churn modeling, and Chapter 4 reveals how Chase applied the prediction of customer departure in a unique way.

SLEEPING DOGS

*If I leave here tomorrow
Would you still remember me?
For I must be traveling on, now
'Cause there's too many places I've got to see.*

—From “Free Bird” by Lynyrd Skynyrd

Imagine you received an alluring brochure from your cell phone company that says:



Tantalized? Imagining a higher-tech toy in your pocket?

Now imagine you are newly emancipated, recently granted the liberty to take your phone number with you to another carrier. You've been aching to change to another carrier to join your friends who say they love it over there. In fact, your provider may have sent you this offer only because it predicted your likely departure.

Big mistake. *The company just reminded you that your contracted commitment is ending and you're free to defect.*



Contacting you backfired, increasing instead of decreasing the chance you'll leave. If you are a sleeping dog, they just failed to let you lie.

Bad news piled on. While already struggling against rising rates of defection, Eva and her colleagues at Telenor detected this backfiring of their efforts to retain, a detrimental occurrence that was now happening more often. More customers were being inadvertently turned away, triggered to leave when they otherwise, if not contacted, might have stayed. It was no longer business as usual.

A NEW THING TO PREDICT

You didn't have to be so nice; I would have liked you anyway.

—The Lovin' Spoonful, 1965

D'oh!

—Homer Simpson

This newly dominant phenomenon brought up for Telenor the question of what PA should be used to predict in the first place. Beyond predicting departure, must a company secondarily predict how customers will respond when contacted? Must we predict the more complicated, two-part question, “Who is leaving but would stay if we contacted them?” This sounds pretty

convoluted. To do so, it seems like we'd need data tracking when people *change their minds!*

This question of integrating a secondary prediction also pertains to another killer app of PA, the utterly fundamental targeting of marketing:

PA APPLICATION: TARGETED MARKETING WITH RESPONSE MODELING

- 1. What's predicted:** Which customers will purchase if contacted.
- 2. What's done about it:** Contact those customers who are more likely to do so.

Despite *response modeling*'s esteemed status as the most established business application of PA (see the 12 examples listed in this book's Central Table 2), it falls severely short because it predicts the outcome for those we *do* contact, but not for those left uncontacted. Assume we have contacted these individuals:



If the dark gray individuals made a purchase, we may proceed with patting ourselves on the back. We must have done a great job of targeting by way of astute predictions about who would buy if contacted, since so many actually did so—relative to how direct marketing often goes, achieving response rates of a few percent, 1 percent, or even less.

One simple question jolts the most senior PA expert out of a stupor: *Which of the dark gray individuals would have purchased anyway, even if we hadn't contacted them?* In some cases, up to half of them—or even more—are so prone to purchasing, they would have either way.

Even an analytics practitioner with decades of experience tweaking predictive models can be floored and flabbergasted by this. She wonders to herself, “Have I been predicting the wrong thing the whole time?” Another bonks himself on the head, groaning, “Why didn’t I ever think of that?” Analytics labs echo with the inevitable Homer Simpson exclamation, “D’oh!”

Let's step back and look logically at an organization's intentions:

- The company wants customers to stay and to buy.
- The company does not intend to *force* customers (they have free will).
- Therefore, the company needs to *convince* customers—to influence, to persuade.

If persuasion is what matters, shouldn't that be what's predicted? Let's try that on for size.

Prediction goal: *Will the marketing brochure persuade the customer?*

Mission accomplished. This meets the company's goals with just one predictive question, integrating within it both whether the customer will do what's desired and whether it's a good idea to contact the customer.

Predicting impact impacts prediction. PA shifts substantially, from predicting a behavior to predicting *influence on behavior*.

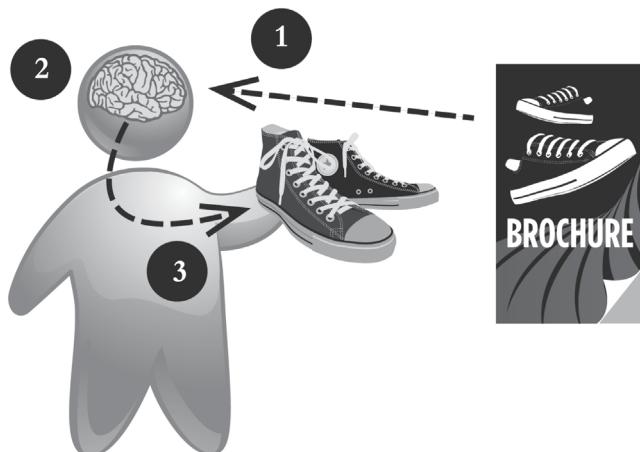
Predicting influence promises to boost PA's value, since an organization doesn't just want to know what individuals will do—it wants to know *what it can do about it*. This makes predictive scores actionable.

I know I asked this earlier but, what could possibly go wrong?

EYE CAN'T SEE IT

Houston, we have another problem.

How can you know something happened if you didn't see it? Take a look at this possible instance of influence:



1. The individual perceives the sales brochure.
2. Something happens inside the brain.
3. The individual buys the product.

Is it safe to assume influence took place? How do we know the brochure made a difference? *Perhaps the individual would have purchased anyway.*

The brain's a black box into which we cannot peek. Even if we were conducting neuroscience, it's not clear if and when that field of science will progress far enough to detect when one changes one's mind (and even if it could, we'd need brain readings from each consumer to employ it!).

Introspection doesn't work, either. You cannot always report on how your own decision making took place. You just can't be certain what made a difference, whether your friend, client, sister, or even you yourself would have made a purchase if circumstances had been different.

To observe influence, we'd need to detect *causality*: Did the brochure *cause* the individual to purchase? As explored in Chapter 3, our knowledge about causality is limited. To truly know causality would be to fully understand how things in the world affect one another, with all the detail involved, the chain reactions that lead one event to result in another. This is the domain of physics, chemistry, and other sciences. It's How the World Works. Ultimately, science tells us only a limited amount.

Therefore, *influence cannot be observed*. We can never witness an individual case of persuasion with complete certainty.

How, then, could we ever predict it?

PERCEIVING PERSUASION

No man ever steps in the same river twice.

—Heraclitus

Good grief. The most valuable thing to predict can't even be detected in the first place.

The desire to influence drives every move we make. As organizations or individuals, out of self-interest or altruistically, almost everything we do is meant to produce a desired effect, including:

- Send a brochure to a customer (or voter).
- Prescribe a medication to a patient.
- Provide social benefits intended to foster self-sufficiency.

Each action risks backfiring: The customer cancels, the patient suffers an adverse reaction, or the beneficiary becomes dependent on assistance. So we make choices not only to pursue what will work, but also to avoid what would do more harm than good.

In one arena in particular, do we feel the pangs of misstep and failure: dating. In courtship, you are both the director of marketing and the product. You're not in the restaurant for food—rather, it is a sales call. Here are some tips and pointers to persuade. Don't be overly assertive, too frequently contacting your prospect. Yet don't remain overly passive, risking that a competitor will swoop in and steal your thunder. Try to predict what you think is the right message, and avoid communicating the wrong thing.

In the movie *Groundhog Day*, our hero Bill Murray acquires a kind of superpower: the coveted ability to perceive influence. Stuck in a magical loop, reliving the same dull day over and over, he faces a humbling sort of purgatory, apparently designed to address the character's flamboyant narcissism. He cannot escape, and he becomes despondent.

Things turn around for Bill when he recognizes that his plight in fact endows him with the ability to *test different marketing treatments on the same subject under exactly the same circumstances*—and then observe the outcome. Desperate to win over the apple of his eye (Andie MacDowell) and immune to the fallout and crush of failure, he endeavors in endless trial and error to eventually learn just the right way to woo her.

Only in this wonderful fantasy can we see with certainty the difference each choice makes. That's life. You never know for sure whether you made the optimal choice about anything. Should I have admitted I love the Bee

Gees? Should we have sent that brochure? Would the other surgical treatment have gone better? Woulda, coulda, shoulda.

In real life, there are no do-overs, so our only recourse is to predict beforehand as well as possible what will work. But, in real life, what's real? If we can't observe influence, how do we know it ever really happens at all?

PERSUASIVE CHOICES

Think before you speak.

Even in dating, there's science to persuasion. Dating website OkCupid showed that messages initiating first contact that include the word *awesome* are more than twice as likely to elicit a response as those with *sexy*. *Howdy* is better than *hey*. *Band* does better than *literature* and *video games* (go figure).

Psychology professor Robert Cialdini persuaded people to commit less crime, and proved it worked. Visitors regularly steal a precious resource from Arizona's Petrified Forest National Park: chunks of petrified wood. Cialdini measured the result of posting the following sign:



With that sign in place, the rate of theft was 1.67 percent. Next he tested another message that more strongly emphasizes the negative effect of theft:



You might expect that would further reduce theft, but it backfired. This message has the effect of destigmatizing theft, since it implies the act is common—“Everybody does it.” Possibly for that reason, it resulted in more than four times as much theft as the first sign, 7.92 percent. Regardless of the psychological interpretation and whether the result is a surprise, persuasion has been proven. We can safely conclude that *relaying the first message rather than the second influences people to steal less*. Similar effects have been shown in the persuasion of hotel room towel recycling and decreasing home energy usage, as explored in Cialdini’s coauthored book, *Yes! 50 Scientifically Proven Ways to Be Persuasive*.²

These studies prove influence takes place across a group but ascertain nothing about any one individual, so the choice of message still cannot be individually selected according to what’s most likely to influence each person.

² Although psychological interpretations such as this destigmatizing effect are not conclusively supported by the data analysis, it is also true that persuasion “by the numbers”—the focus of this chapter—depends on the creative design of messages (more generally, treatments) to test in the first place. As always, human creativity, such as that in the field of psychology, and number crunching—the soft and the hard sciences—complement one another and are mutually interdependent.

In the field of medicine, most clinical studies do this same thing—compare two treatments and see which tends to work better overall. For knee surgery after a ski accident, I had to select a graft source from which to reconstruct my busted anterior cruciate ligament (ACL, the knee's central ligament—previously known to me as the Association for Computational Linguists). I based my decision on a study that showed subsequent knee walking was rated “difficult or impossible” by twice as many patients who donated their own patellar tissue rather than hamstring tissue.³

It’s good, but it’s not personalized. I can never know if my choice for knee surgery was the best for my particular case (although my knee does seem great now). The same holds true for any treatment decision based on such studies, which provide only a one-size-fits-all result. We’re left with uncertainty for each individual patient. If you take a pill and your headache goes away, you can’t know for sure that the medicine worked; maybe your headache would have stopped anyway.

More generally, if you prevent something bad, how can you be sure it was ever going to happen in the first place?

BUSINESS STIMULUS AND BUSINESS RESPONSE

Many of your everyday clicks contribute to the Web’s constant testing of how to improve overall persuasiveness. Google has compared 41 shades of blue to see which elicits more clicks. Websites serve the ads that get clicked the most and run random AB tests to compare which Web page design and content lead to the most buying. Facebook conducts controlled experiments to see how changes to the rules driving which friends’ posts get displayed influence your engagement and usage of their website (see Central Table 1).

³ The decision was mine alone, with no personalized guidance from a physician. I found each knee surgeon to be almost entirely devoted to one graft source or another and therefore unable to provide balanced guidance for my choice. My only option was to first select a surgical procedure and then choose a doctor who focused on that procedure.

I tested titles for this book, following in the footsteps of *SuperCrunchers* and *The 4-Hour Workweek*. Placed as ads on Google Adwords, *Predictive Analytics*, when displayed on tens of thousands of screens of unsuspecting experimental subjects across the country, was clicked almost twice as often as *Geek Prophecies* and also beat out *I Knew You Were Going to Do That* and *Clairvoyant Computers*, plus six other book titles that I also entered into this contest. It was convenient that the field's very name came out as the top contender, an unquestionably fitting title for this book.

In both medicine and marketing, this scheme to test *treatments* reveals the impact of selecting one outward action over another—but only as a trend across the group of subjects as a whole. After this sort of experiment, the best an organization can do is run with the one most effective treatment, applying it uniformly for all individuals.

In this practice, the organization is employing a blunt instrument. Looking back, we still don't know for whom the treatment was truly effective. Looking forward, we still don't know how to make personalized choices for each individual.

THE QUANTUM HUMAN

Here's the thing about the future. Every time you look at it, it changes. Because you looked at it.

—Nicolas Cage's clairvoyant in *Next*

Heisenberg might have slept here.

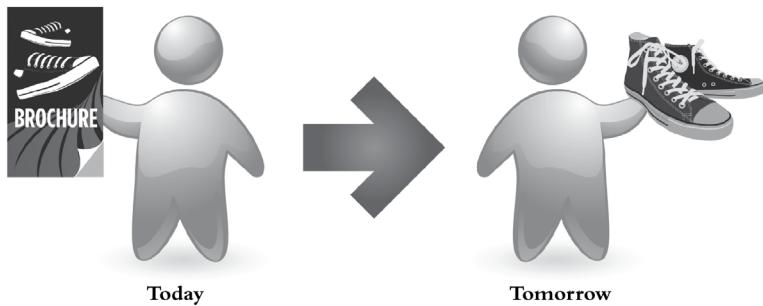
—Anonymous

As in quantum physics, some things are unknowable. Although you may protest being reduced to a quantum particle, there's a powerful analogy to be drawn between the uncertainty about influence on an individual and *Heisenberg's uncertainty principle*. This principle states that we can't know everything about a particle—for example, both its position and speed. It's a trade-off. The more precisely you measure one, the less precisely you can measure the other.

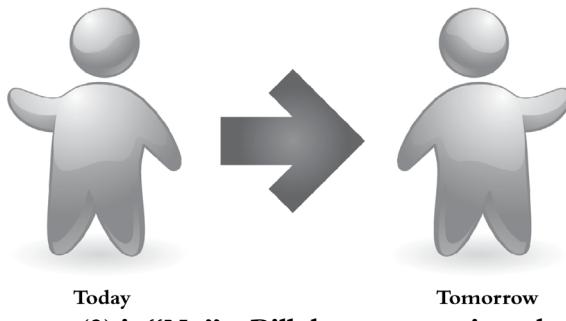
Likewise, we can't know everything about a human. In particular, we can't know both things that we'd need to know in order to conclude that a person could be influenced. For example:

1. Will Bill purchase if we send him a brochure?
2. Will Bill purchase if we *don't* send him a brochure?

If we did know the answer to both, we'd readily know this most desired fact about Bill—whether he's *influenceable*. In some cases, the answers to the two questions disagree, such as:



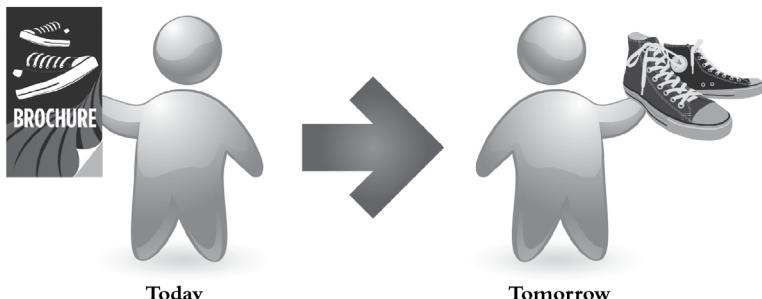
The answer to (1) is “Yes”—Bill receives a brochure and then purchases.



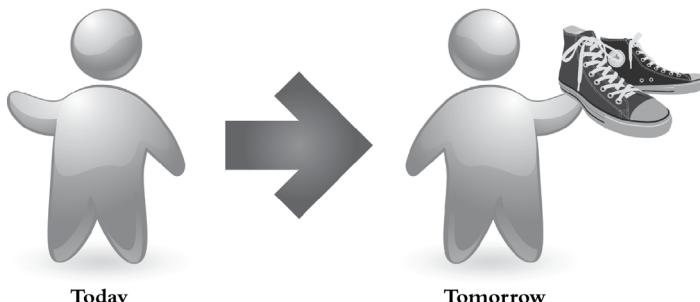
The answer to (2) is “No”—Bill does not receive a brochure and does not purchase.

In this case, we would conclude that the choice of treatment does have an influential effect on Bill; he is persuadable.

In other cases, the answers to the questions agree, such as:



The answer to (1) is “Yes”—Bill receives a brochure and then purchases.



The answer to (2) is also “Yes”—Bill does not receive a brochure but then purchases anyway.

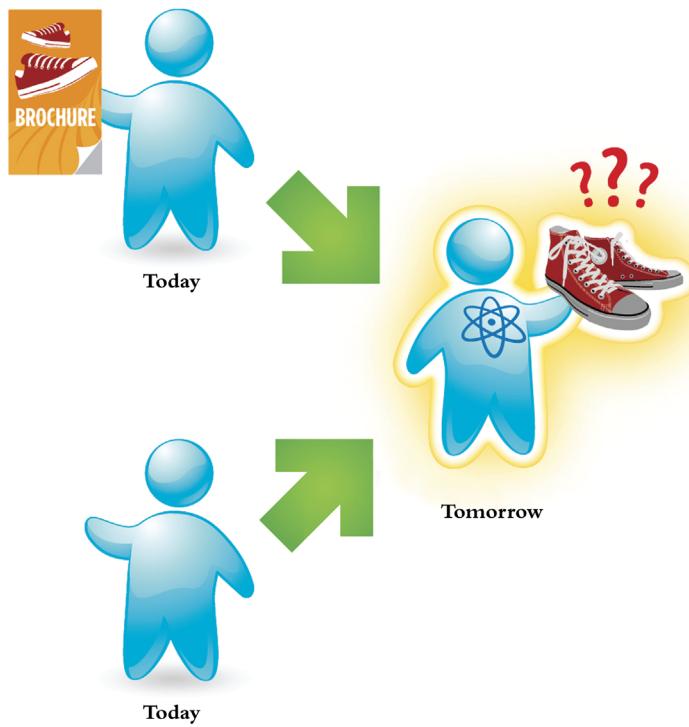
In this case, we conclude the choice of treatment has no influence; he would buy either way. This type of customer is called a *sure thing*.

Other scenarios exist. Sometimes a brochure backfires and adversely influences a customer who would otherwise buy not to.

But this is a fantasy—we *can't* know the answer to both questions. We can find out (1) by sending Bill a brochure. We can find out (2) by not sending him a brochure. But we can't both contact and not contact Bill. We can't administer medicine and not administer medicine. We can't try two different forms of surgery at once. In general, you can't test an individual with both treatments.

This uncertainty leaves us with philosophical struggles akin to those of quantum physics. Given that we could never know both, does a particle ever

really have both a true position and a true speed? Similarly, do answers to both of the previous questions about a person truly exist? Answering one renders the other purely hypothetical. It's like the tree falling in the forest with no one to perceive the sound, which becomes only theoretical. This most fundamental status of a human as influenceable or not influenceable holds only as an ethereal concept. It's only observable in aggregate across a group, never established for any one person. Does the quality of influenceability exist only in the context of a group, emergently, defying true definition for any single individual? If influenceable people do walk among us, you can never be certain who they are.



The quantum human—is he or she influenceable?

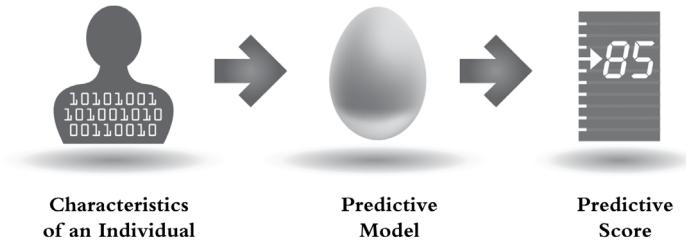
This unknowability equates the past and the future. We don't know whether a person *was* influenced, and we don't know whether the person *could be* influenced—whether he or she is *influenceable*. It's kind of a refreshing change that prediction is no more difficult than retrospection, that tomorrow

presents no greater a challenge than yesterday. Both previous and forthcoming influence can only at best be estimated. Clearly, the future is the more valuable one to estimate. If we can know *how likely* each person is to be influenced, we can drive decisions, treating each individual accordingly.

But how can you predictively model influence? That is, how could you train a predictive model when there are no learning examples—no individual known cases—of the thing we want to predict?

PREDICTING INFLUENCE WITH UPLIFT MODELING

A model that predicts influence will be a predictive model like any other:



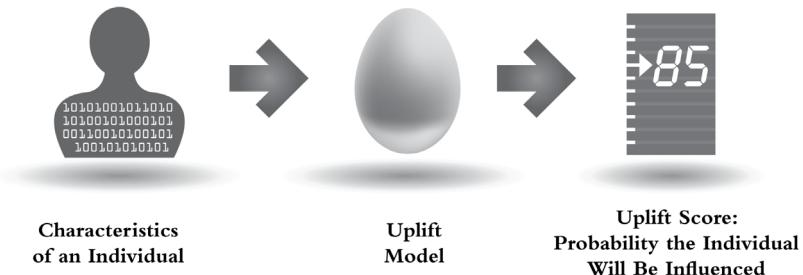
Like all the models we've covered in this book, it takes characteristics of the individual as input and provides a predictive score as output.

But it will be a special case of predictive models. Instead of predicting an outright behavior, we need *a model that scores according to the likelihood an individual's behavior will be influenced*. We need an *uplift model*:

Uplift model—*A predictive model that predicts the influence on an individual's behavior that results from applying one treatment over another.*⁴

The uplift score answers the question, “*How much more likely is this treatment to generate the desired outcome than the alternative treatment?*” It guides an organization’s

⁴ Not to be confused with the *lift* of a predictive model covered in Chapter 4, uplift modeling is also known as *differential response*, *impact*, *incremental impact*, *incremental lift*, *incremental response*, *net lift*, *net response*, *persuasion*, *true lift*, or *true response modeling*.



choice of treatment or action, what to do or say to each individual.⁵ The secondary treatment can be the passive action of a *control set*—for example, make no marketing contact or administer a placebo instead of the trial drug—in which case an uplift model effectively decides whether or not to treat.

How do you learn about something you can't see? We never have at our disposal learning examples of the very thing we want to predict: *influenceable individuals*. We don't have the usual training data from which to directly learn.

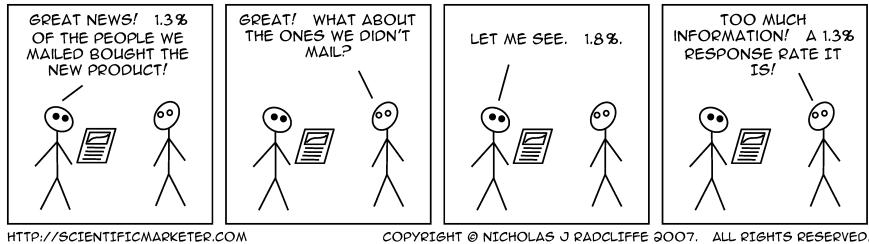
To do the seemingly impossible, *uplift modeling* needs a clever work-around. To see how it works, let's explore a detailed example from U.S. Bank.

BANKING ON INFLUENCE

U.S. Bank Assistant Vice President Michael Grundhoefer isn't satisfied with good. In the mid-1990s, the bank's direct marketing efforts to sell financial products such as lines of credit fared well. Most mail campaigns turned a satisfactory profit. Michael, who headed up the analytics behind many of these campaigns, kept a keen eye on the underlying response models and how they could be improved.

Companies often misinterpret marketing campaign results. Here's where they go terribly wrong: They look at the list of customers contacted and ask, "How many responded?" That's the *response rate*. One of the original inventors of uplift modeling, Nicholas Radcliffe (now an independent consultant and sometimes visiting professor in Edinburgh), drew a cartoon about that measure's drawbacks:

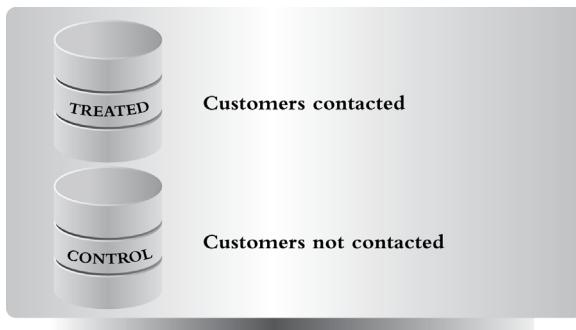
⁵ While *prescriptive analytics* might be a suitable synonym for uplift modeling, it is not usually used this way. Ill-defined, it is a problematic term; see the Notes for more.



Cartoon reproduced with permission.

The response rate completely overlooks how many would buy anyway, even if not contacted. Some products just fly off the shelves and sell themselves. For business, that's a good thing—but if so, it's important not to credit the marketing. You could be wasting dollars and chopping down trees to send mail that isn't actually helping.

Just as with medicine, marketing's success—or lack thereof—is revealed by comparing to a control set, a group of individuals suppressed from the treatment (or administered a placebo, in the case of medicine). Therefore, we need to collect two sets of data:



If the treated customers buy more than the control customers, we know the campaign successfully persuades. This proves some individuals were influenced, but, as usual, we don't know which.

PREDICTING THE WRONG THING

If you come to a fork in the road, take it.

—Yogi Berra

To target the marketing campaigns, Michael and his team at U.S. Bank were employing the industry standard: response models, which predict who will

buy if contacted. That's not the same thing as predicting who will buy *because* they were contacted; it does not predict influence. Compared to a control set, Michael showed the campaigns were successful, turning a profit. But he knew the targeting would be more effective if only there were a way to predict which customers would be *persuaded* by the marketing collateral.

Standard response models predict the wrong thing and are in fact falsely named. Response models don't predict response *caused by* contact; they predict buying *in light of* contact. But predicting for whom contact will be the cause of buying is more pertinent than predicting buying in general. Knowing who your "good" customers are—the ones who will buy more—may be nice to know, but it takes second place.⁶

For some projects, conventional response models have it backward. By aiming to increase response rate, they complacently focus on the metric that's easiest to measure. As former U.S. Secretary of Defense Robert McNamara said, "We have to find a way of making the important measurable, instead of making the measurable important." A standard response model will gladly target customers who would buy anyway, doing little to address how much junk mail we as consumers receive. Instead, it's only a small sliver of persuadable customers who are actually worth mailing to, if we can identify them.

Standard response modeling predicts:

1. Will the customer buy if contacted?

Uplift modeling changes everything by adding just one word:

2. Will the customer buy **only** if contacted?

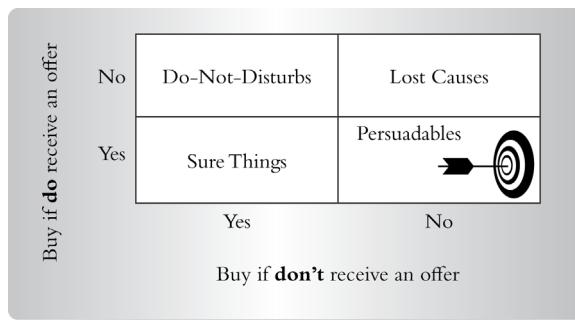
⁶ Driving decisions by only predicting the outcome of one treatment without predicting the result of the other is a form of *satisficing*. It's a compromise. Instead of compromising, marketing needs all the help it can get to better target. As a data miner, I actually receive e-mail inquiries from drilling supply vendors. I'm not that kind of miner. Eric King of The Modeling Agency receives job inquiries from (human) models seeking opportunities in the fashion industry.

Although the second question may appear simple, it answers the composite of two questions: “Will the customer buy if contacted and not buy otherwise?” This two-in-one query homes in on the difference that will result from one treatment over another. It’s the same as asking, “Would contacting the customer *influence* him or her to buy?”

RESPONSE UPLIFT MODELING

Weigh your options.

By addressing a composite of two questions, each individual belongs in one of four conceptual segments that distinguish along two dimensions:



Conceptual response segments. The lower-right segment is targeted with uplift modeling.⁷

This quad first distinguishes from top to bottom which customers will buy in light of marketing contact, which is the job of conventional response modeling. But then it further distinguishes along a second dimension: Which customers will make a purchase even if not contacted?

⁷ Table derived from Nicholas Radcliffe, “Generating Incremental Sales: Maximizing the Incremental Impact of Cross-Selling, Up-Selling and Deep-Selling through Uplift Modeling,” Stochastic Solutions Limited, February 16, 2008, and Suresh Vittal, “Optimal Targeting through Uplift Modeling: Generating Higher Demand and Increasing Customer Retention While Reducing Marketing Costs,” Forrester Research white paper, 2008.

Michael at U.S. Bank wanted to target the lower-right quadrant, those worthy of investing the cost to contact. These persuadables won't buy if not contacted, but will buy if they are. These are the individuals an uplift model aims to flag with the affirmative prediction.

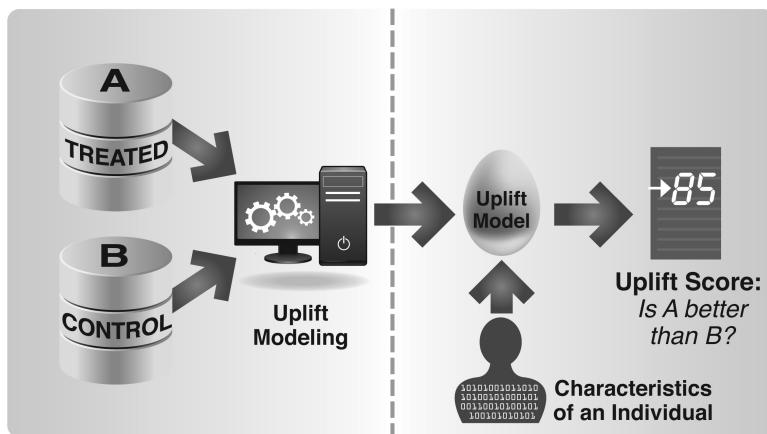
PA APPLICATION: TARGETED MARKETING WITH *RESPONSE UPLIFT MODELING*

- 1. What's predicted:** Which customers will be persuaded to buy.
- 2. What's done about it:** Target persuadable customers.

An uplift model provides the opportunity to reduce costs and unnecessary mail in comparison to a traditional response model. This is achieved by suppressing from the contact list those customers in the lower-left quadrant, the so-called sure things who will buy either way.

THE MECHANICS OF UPLIFT MODELING

Uplift modeling operates simultaneously on two data sets—both the treated set and the control set—learning from them both:

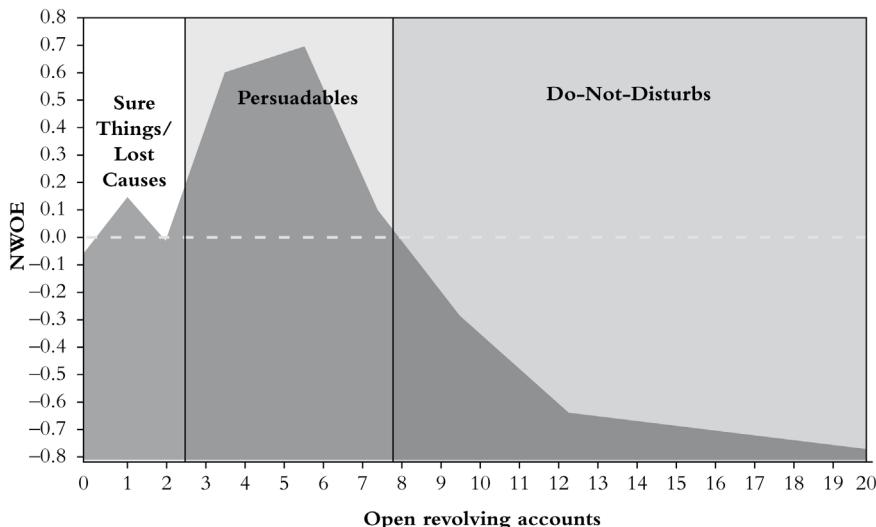


Two training sets are used together to develop an uplift model.

To learn to distinguish influenceables—those for whom the choice of treatment makes a difference—uplift modeling learns from both customers

who were contacted and others who weren't. Processing two data sets represents a significant paradigm shift after decades of predictive modeling and machine learning research almost entirely focused on tweaking a modeling process that operates across a single data set.

Starting first with a single-variable example, we can see that it is possible to predict uplift by comparing behavior across the two data sets:



Net weight of evidence (NWOE, a measure of uplift) varies by a customer's number of open revolving accounts. Graph courtesy of Kim Larsen.

This fictional but typical example of a financial institution's direct-marketing results illustrates that mildly engaged customers are hot, readily persuadable by direct mail. The vertical axis represents *net weight of evidence* (NWOE), a measure of uplift, and the horizontal axis represents the number of open revolving accounts the customer already holds. In this case, it turns out that customers in the middle region, who don't already hold too many or too few open revolving accounts, will be more likely to be persuaded by direct mail.

Less engaged customers on the left are unmoved—whether they were already destined to open more accounts or not, their plans don't change if contacted. This includes both sure things and lost causes—either way, it isn't worth contacting them.

Avoid at all costs contacting customers on the right—they are “do-not-disturbs.” Contacting these individuals, who already hold a good number of accounts, actually decreases the chance they’ll buy. The curve dips down into negative numbers—a veritable *downlift*. The explanation may be that customers with many accounts are already so engaged that they are more sensitive to, aware of, and annoyed by what they consider to be unnecessary marketing contact. An alternative explanation is that customers who have already accumulated so many credit accounts are susceptible to impulse buys (e.g., when they come into a bank branch), but when contacted at home will be prone to respond by considering the decision more intently and researching competing products online.

This shows the power of one variable. How can we leverage PA’s true potential by considering multiple variables, as with the predictive models of Chapter 4? Let’s turn back to Michael’s story for a detailed example.

HOW UPLIFT MODELING WORKS

Despite their marketing successes, Michael at U.S. Bank had a nagging feeling things could be better. Unlike many marketers, he was aware of the difference between a campaign’s response rate and the sales generated by it. Inspecting reports, he could see the response models were less than ideal. He tried out some good ideas of his own to attempt to model persuasion, which provided preliminary yet inconsistent and unstable success.

One time, Michael noted failure for a certain group within a direct mail campaign selling a home-equity line of credit to existing customers. For that group, the campaign not only failed to cover its own printing and mailing costs, it in fact had the detrimental effect of decreasing sales, a slight *downlift* overall.

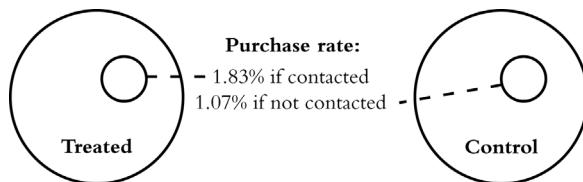
Michael was beginning to collaborate with a small company called Quadstone (now Pitney Bowes Software) that provided a new commercial approach to uplift modeling. The system could derive marketing segments that reveal persuadable customers, such as:⁸

⁸ Thanks to Patrick Surry at Pitney Bowes Software for this example segment derived across U.S. Bank data. The segment is simplified for this illustration.

**Has paid back more than 17.3% of current loan
-AND-
Is using more than 9.0% of revolving credit limit
-AND-
Is designated within a certain set of lifestyle segments**

A segment of persuadable individuals.

This is not your usual marketing segment. It doesn't designate customers more likely to buy. It doesn't designate customers less likely to buy. It is customers *more likely to be influenced by marketing contact*. The difference marketing makes for this segment can be calculated only by seeing how its purchase rate differs between the treated and control sets:⁹



Purchase rates of the persuadable segment described above differ, depending on whether marketing contact is received.

Success! Within this segment, the direct mail elicits more responses from customers who were contacted (the treated set) than those not contacted (the control set). By automatically deriving its defining characteristics, uplift modeling has discovered a segment of customers for which this direct mail campaign succeeds after all.

⁹ A simpler alternative to analyzing both sets at once is to make a separate predictive model for each treatment, as was the approach behind the online ad-selection case study described in Chapter 1. Michael at U.S. Bank evaluated this simpler method and concluded that the tree-based approach to uplift modeling provided stronger and more consistent results.

The uplift modeling method that discovers such segments is an expansion of decision trees (see Chapter 4) called *uplift trees*. Normal decision trees strive to identify segments extreme in their response rates—many responses or few responses. Uplift trees use variables to mechanically “segment down” in the same way, but seek to find segments extreme in the difference treatment makes—segments that are particularly influenceable. A single uplift tree is composed of a number of segments such as the one shown above.¹⁰

For U.S. Bank, response uplift modeling delivered an unprecedented boost, increasing the marketing campaign’s return on investment (ROI) by a factor of five in comparison with standard response model targeting. This win resulted from reducing both the amount of direct mail that commanded no impact (sent to lost causes or sure things) and the amount that instigated an adverse response (sent to sleeping dogs, aka do-not-disturbs).

CASE STUDY: U.S. BANK

Business case: Direct mail for a home-equity line of credit to existing customers.

Approach: Target campaign with an uplift model.

Resulting improvements over prior conventional analytical approach:

- Return on investment (ROI) increased five times over previous campaigns (from 75 percent to 400 percent).
- Campaign costs cut by 40 percent.
- Revenue gain increased by over 300 percent.

Uplift practitioners at Fidelity Investments also see the light: *Spend less, earn more*. By avoiding sure things and do-not-disturbs, “Uplift modeling empowers your organization to capture more than 100 percent of responses

¹⁰ Ensemble models (see Chapter 5) of decision trees are recommended when employing this analytical approach to uplift modeling to help ensure stable results. Although predicting influence rather than outright behavior, The Ensemble Effect still applies (as do The Prediction, Data, and Induction Effects).

by contacting less than 100 percent of the target population,” says Kathleen Kane, Fidelity’s principal decision scientist.

THE PERSUASION EFFECT

Uplift modeling conquers the imperceivable—*influence*—by newly combining two well-trodden, previously separate paradigms:

1. comparing treated and control results; and
2. predictive modeling (machine learning, statistical regression, etc.).

Only by cleverly combining these two practices does the newfound ability to predict persuasion for each individual become possible. I call this *The Persuasion Effect*:

The Persuasion Effect: *Although imperceptible, the persuasion of an individual can be predicted by uplift modeling, predictively modeling across two distinct training data sets that record, respectively, the outcomes of two competing treatments.*

If you haven’t already figured it out, this answers the riddle posed at the beginning of this chapter. *Being influenced* is the thing that often happens to you that cannot be witnessed and that you can’t even be sure has happened afterward—but that can be predicted in advance. In this way, PA transcends human perception.

INFLUENCE ACROSS INDUSTRIES

Uplift modeling applies everywhere: marketing, credit risk, electoral politics, sociology, and healthcare. The intent to influence is common to almost all organizations, so The Persuasion Effect is put into play across industry sectors.

Application of Uplift Modeling	Treatment Decision	Objective
Targeted marketing with response uplift modeling	<i>Should we contact the customer or not (active or passive treatment)?</i>	Positive impact of direct marketing campaigns
Customer retention with churn uplift modeling	<i>Should we provide the customer a retention offer or not (active or passive treatment)?</i>	Positive impact of retention campaigns
Content selection	<i>With which ad, illustration, choice of words, or product should we solicit the customer?</i>	Response rate of direct marketing, cross-sell, and online and offline ads
Channel selection	<i>Through which channel should we contact the customer (e.g., mail, e-mail, or telephone)?</i>	Positive impact of direct marketing campaigns
Dynamic pricing and discounting	<i>Should we offer the customer a higher price or a lower price?</i>	Revenue of sales
Collections	<i>Should we offer the debtor a deeper write-off?</i>	Revenue of accounts receivable
Credit risk	<i>Should we offer the customer a higher or lower credit limit? A higher or lower APR?</i>	Revenue from interest payments and savings from fewer defaults
Electoral politics	<i>Should we market to the constituent/in the state (swing voter/swing state)?</i>	Positive votes resulting from political election campaigns (see this chapter's sidebar for how Obama's 2012 campaign employed uplift modeling)
Sociology	<i>Should we provide benefits to this individual?</i>	Improved social program outcome: long-term self- sufficiency

(continued)

(continued)

Application of Uplift Modeling	Treatment Decision	Objective
Personalized medicine	<i>Which medical treatment should the patient receive?</i>	Favorable patient outcome in clinical healthcare

This chapter covers in detail the first two areas on marketing in the table above, as well as a case study in electoral politics (in the sidebar about Obama's 2012 presidential campaign at the end of this chapter). Here's a bit more about the rest of them (note that for some of these application areas, no public case studies or proofs of concept yet exist—uplift modeling is an emerging technology).

Content and channel selection. Uplift modeling selects for each user the ad, offer, content, product, or channel of contact (phone, e-mail, etc.) most likely to elicit a response. In these cases, there is no passive option and therefore no control set—both data sets test an active treatment.

Dynamic pricing and collections. As for any decision, a certain risk is faced for each treatment option when pricing: The higher price may turn a customer away, but the lower price (or deeper discount or write-off, for collections) unnecessarily sacrifices revenue if the customer would have been willing to pay more.

Credit risk. The balance between risk and upside profitability for each debtor is influenced by both the credit limit and the APR offered. Raising one or both may result in higher revenue in the form of interest payments, but may also increase the chance of the debtor defaulting on payments and an ensuing write-off.

Electoral politics. As a resident of California, I see few if any ads for presidential campaigns—the state is a lock; depending on your political affiliation, it could be viewed as either a sure thing or a lost cause. Just as so-called swing clients (influenceables) are potentially persuaded by marketing contact, the same benefit is gained where this term originates: political campaigns that target swing voters. The constituents with the most potential

to be influenced by campaign contact are worth the cost of contact. Analogously, only the swing states that could conceivably be persuaded as a whole are worth expending great campaign resources. For more on elections and uplift modeling, see this chapter's sidebar, "Beyond Swing Voters: How Persuasion Modeling Revolutionized Political Campaigns for Obama and Beyond."

Sociology: targeting social programs. Speaking of politics, here is a concept that could change everything. Social programs such as educational and occupational support endure scrutiny as possibly more frequently awarded to go-getters who would have succeeded anyway. For certain other beneficiaries, skeptics ask, does support backfire, leaving them more dependent rather than more self-sufficient? Only by predicting how a program will influence the outcome for each individual prospect can programs be targeted in a way that addresses these questions. In so doing, *might such scientifically based, individualized economic policies help resolve the crippling government deadlock that results from the opposing fiscal ideologies currently held by conservative and liberal policymakers?*

Personalized medicine. While one medical treatment may deliver better results on average than another, this one-size-fits-all approach commonly implemented by clinical studies means treatment decisions that help many may in fact hurt some. In this way, healthcare decisions backfire on occasion, exerting influence opposite to that intended: They hurt or kill—although they kill fewer than following no clinical studies at all. *Personalized medicine* aims to predict which treatment is best suited for each patient, employing analytical methods to predict treatment impact (i.e., medical influence) similar to the uplift modeling techniques used for marketing treatment decisions. For example, to drive beta-blocker treatment decisions for heart failure, Harvard University researchers "use two independent data sets to construct a systematic, subject-specific treatment selection procedure." A certain HIV treatment is shown to be more effective for younger children. Treatments for various cancers are targeted by genetic markers—a trend so significant the Food and Drug Administration is increasingly requiring for new pharmaceuticals, as *The New York Times* puts it, "a companion test that could reliably detect the [genetic] mutation so that

the drug could be given to the patients it is intended to help,” those “most likely to benefit.”

IMMOBILIZING MOBILE CUSTOMERS

It wasn’t long after phone number portability came, raising a hailstorm in the telecommunications industry, that Quadstone spoke with Eva at Telenor about the new uplift modeling technique. It was a revelation. Eva had already confirmed that Telenor’s retention efforts triggered some customers to leave rather than persuading them to stay, but she wasn’t aware of any established technique to address the issue. The timing was fortuitous, as Quadstone was just starting out, seeking its first few clients to prove uplift modeling’s value.

PA APPLICATION: CUSTOMER RETENTION WITH *CHURN UPLIFT MODELING*

- 1. What’s predicted:** Which customers can be persuaded to stay.
- 2. What’s done about it:** Retention efforts target persuadable customers.

Customers can be as easily scared away as a skittish bunny. Traditional churn models often inadvertently frighten these rabbits, since customers most likely to leave are often those most easy to trigger—sleeping dogs easy to wake up. This includes, for example, the health club member who never gets to the gym and the Netflix subscriber who rarely trades in each DVD movie rental—both just need a reminder before they get around to canceling (it would be more ideal to reengage them). Someone once told me that, when he received an offer to extend his camera’s warranty, it reminded him that coverage was soon ending. He promptly put his camera into the microwave to break it so he could return it. It would inevitably be more cost-effective to avoid triggering such criminal activity than to prosecute for it after the fact.

Prompting a cell phone customer to leave can be especially costly because it may trigger a social domino effect: People tend to stick with the same wireless carrier as their friends. One major North American carrier showed that a customer is seven times more likely to cancel if someone in the person’s calling network cancels.

For Telenor, churn uplift modeling delivered an astonishing boost to the effectiveness of its retention initiatives: The ROI increased by a factor of 11, in comparison with its prior, established practice of targeting with standard churn models. This came from decreasing the number of sleeping dog customers the company had been inadvertently waking, and secondarily from reducing the total number of mail pieces sent—like U.S. Bank, Telenor got more for less.

CASE STUDY: TELENO, THE WORLD'S SEVENTH-LARGEST MOBILE OPERATOR

Business case: Retention campaign for cell phone subscribers.

Approach: Target campaign with an uplift model.

Resulting improvements over the conventional approach to analytical retention:

- Campaign ROI increased by a factor of 11.
- Churn reduced a further 36 percent.
- Campaign costs reduced by 40 percent.

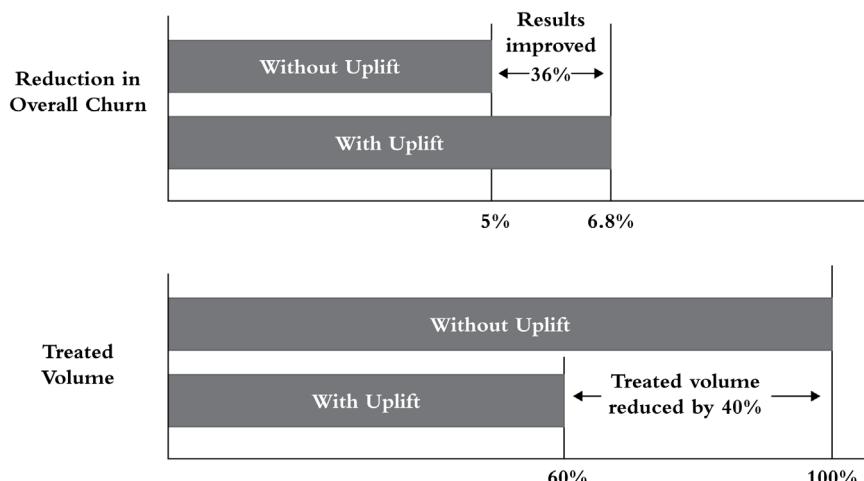


Figure permission of Pitney Bowes Software.

For the international mobile carrier, which serves tens of millions of cell phone subscribers across 11 markets, this was a huge win. Beyond addressing the new business challenges that came of phone number portability, it alleviated the systematic “sleeping dog” problem inherent to churn modeling, one Telenor likely had suffered from all along. Even when there’s a net benefit from marketing, offers could be triggering some customers to leave who would have otherwise stayed.

For Eva, who has since been promoted to head of customer analytics, and for the rest of the world, this only marks the beginning of the emerging practice of inducing influence and predicting persuasion.

BEYOND SWING VOTERS

No other presidential campaign [besides Obama's] has relied so heavily on the science of analytics, using information to predict voting patterns. Election day may have changed the game.

—Christi Parsons and Kathleen Hennessey, *Los Angeles Times*,
November 13, 2012

Elections hang by a thinner thread than you think.

You may know that President Barack Obama's 2012 campaign for a second term *Moneyballed* the election, employing a team of over 50 analytics experts.

You may also know that the tremendous volume of any presidential campaign's activities, frenetically executed into the eleventh hour in pursuit of landing the world's most powerful job, ultimately serves only to sway a thin slice of the electorate: swing voters within swing states.

But what most people do not realize is that presidential campaigns must focus *even more narrowly than that*, taking microtargeting to a whole new level. The Obama campaign got this one right, breaking ground for election cycles to come by applying uplift modeling to

BEYOND SWING VOTERS (CONTINUED)

drive millions of per-voter campaign decisions, thereby boosting persuasive impact.

However, the buzz in 2012 was about something else. Rather than learning about campaign targeting, when it came to the math behind the election, we heard a great deal about Nate Silver. Silver emerged as the media darling of poll analyzers, soaring past the ranks of guru quant or sexy scientist to become the very face of prediction itself. If mathematical “tomorrowvision” had a name, it was Nate. Even before his forecasts were vindicated by the election results, it was hard to find a talk show host—at least among the left—who hadn’t enjoyed a visit from Silver, probing him with evident, slack-jawed fascination.

An election poll does not constitute prognostic technology—it does not endeavor to calculate insights that foresee human behavior. Rather, a poll is plainly the act of voters explicitly telling you what they’re going to do. It’s a minielection dry run. There’s a craft to aggregating polls, as Silver has mastered so adeptly, but even he admits it’s no miracle of clairvoyance. “It’s not really that complicated,” he told late night talk show host Stephen Colbert the day before the election. “There are many things that are *much* more complicated than looking at the polls and taking an average . . . right?”

You want power? *True power comes in influencing the future rather than only speculating on it.* Nate Silver publicly competed to win election forecasting, while Obama’s analytics team discreetly competed to win the election itself.

This reflects the very difference between forecasting and predictive analytics (PA). Forecasting calculates an aggregate view for each U.S. state, while PA delivers more detailed insights that drive action: predictions for each individual voter.

THE RARE BIRD: PERSUADABLE VOTERS

Swing voters are a myth. The concept is ill-defined and subjective. In one approach, the Democratic National Committee (DNC) labels as

(continued)

BEYOND SWING VOTERS (CONTINUED)

“not very partisan” those voters who self-reported as independent, or for whom their party is (for any reason) unknown. Despite being labeled “swing,” many such voters have indeed made up their minds and are unswingable.

Instead of mythical swing voters—or unicorns, for that matter—what matters to a campaign is concrete and yet quite narrow: *Who will be influenced to vote for our candidate by a call, door knock, flier, or TV ad?*

Presidential campaigns must hold themselves to a higher standard than most marketing campaigns. In this unparalleled, ruthless competition of optimal tweaking, the notion of expending resources—such as a paid mailing or a campaign volunteer’s precious time—to contact a constituent who was already going to vote your way is abhorrent. Even worse, it is known that, for some cases, contact will inadvertently change the voter’s mind in the wrong direction—it can backfire and cause the person to vote for the other candidate.

In the business world, marketing campaigns often withstand such cases without wincing. They inadvertently hit some “sure thing” and “do-not-disturb” customers, yet carry on happily with high profits. As long as the overall campaign is doing more good than harm, taking on the more sophisticated methods needed to smooth these imperfect edges is often seen as too high an investment relative to the expected payoff (although this determination is often just inertia speaking; *uplift modeling* is new and not yet widely practiced).

But a presidential campaign comes along only once every four years. Its extraordinarily high stakes demand that all stops be pulled out. It was only a matter of time before campaigns began predicting their potential to influence each voter in order to optimize that influence.

ANOTHER RARE BIRD: PERSUASION MODELING EXPERTS

Enter Rayid Ghani, chief data scientist of the presidential campaign Obama for America 2012. He was the man for the job. With a master’s

BEYOND SWING VOTERS (CONTINUED)

degree in machine learning from Carnegie Mellon (the first university to have a machine learning department), plus 10 years of research work at the labs of consulting behemoth Accenture, Rayid had rare, sought-after experience in uplift modeling—which the campaign called *persuasion modeling*. His background included research determining which medical treatment will provide the best outcome for each patient, and which price will provide the best profit from each retail customer. At Obama for America, he helped determine whether campaign contact would provide the right vote from each constituent.

It's a deep analytical challenge. A predictive model that foresees the ability to persuade is not your average predictive model. Beyond identifying voters who would come out for Obama if contacted, the persuasion models developed by Rayid's staff needed to distinguish those voters who would come out for Obama in any case (the sure things), as well as those who in fact were at risk of being turned off by campaign contact and switching over to vote for the other guy, Mitt Romney (the do-not-disturbs). If you think it through, you'll see the single idea of "can be positively persuaded" actually involves all these distinctions.

PA APPLICATION: POLITICAL CAMPAIGNING WITH VOTER PERSUASION MODELING

- 1. What's predicted:** Which voter will be positively persuaded by political campaign contact such as a call, door knock, flier, online ad, or TV ad.
- 2. What's done about it:** Persuadable voters are contacted, and voters predicted to be adversely influenced by contact are avoided.

For this project, the campaign needed to collect not donations but data. No matter how smart, the brains on Obama's staff could only tackle the persuasion problem with just the right data sets. To this end, they tested across thousands of voters the very actions they would later

(continued)

BEYOND SWING VOTERS (CONTINUED)

decide on for millions. Batches of voters received campaign contact—door knocks, fliers, and phone calls—and, critically, other batches received no contact at all (the control groups). All the batches were then later polled to see whether they would support Obama in the voting booth.

ACTIVELY CAMPAIGNING ON PERSUASION

[The Obama campaign job listing for “predictive modeling”] read like politics as done by Martians.

—Peggy Noonan, *The Wall Street Journal*, July 30, 2011

The Martians have landed.

—Christi Parsons and Kathleen Hennessey, *Los Angeles Times*, November 13, 2012

The data proved that campaigning generally helps, which was good news for the team—but then, analysis had only just begun. Rayid’s team faced the ultimate campaign imperative: Learn to discriminate, voter by voter, whether contact would persuade. This is where persuasion modeling (the technique described in the rest of this chapter as *uplift modeling*) came in and took over by storm.

“Our modeling team built persuasion models for each swing state,” Rayid said. “The models were then employed to predict the potential to persuade for each of millions of individuals in swing states. It told us which were most likely to be won over to Obama’s side, and which we should avoid contacting entirely.”* A small group of only three quants led the hands-on execution of persuasion modeling for the campaign.

* Beyond persuasion modeling, the team also employed predictive modeling to gauge the propensity to vote for Obama regardless of campaign contact, the probability of voting at all (turnout), and the probability of donating in order to target fund-raising efforts.

BEYOND SWING VOTERS (CONTINUED)

Persuasion modeling identified nonpartisan voters, according to Director of Statistical Modeling Daniel Porter, one of the three-member unit. Daniel and his colleagues tweaked the models, experimenting extensively across avant-garde techniques designed to identify which factors predict whether a voter is persuadable. The process delivered, pinpointing certain behaviors that seemed to reveal a voter is not strictly partisan, such as supporting Bush in 2004 but Obama in 2008, or being registered as a Democrat in combination with having voted Republican or living in a highly Republican location.

The available data sources were rich. Although campaign staff have not disclosed many other details about the data elements available to discern persuadability, their related effort predicting a constituent's propensity to vote for Obama (regardless of campaign contact) employed more than 80 fields, including demographics, voting history, and magazine subscriptions. The campaign's most cherished data source was the DNC's database, which includes notes regarding each voter's observed response to door knocks—welcoming or doorslamming—during prior presidential election cycles.

The potential persuadability of each voter predicted by these models guided the massive army of campaign volunteers as they pounded the pavement and dialed phone numbers. When knocking on a door, the volunteer wasn't simply canvassing the local neighborhood—this very voter had been predictively targeted as persuadable. As Daniel Wagner, the campaign's chief analytics officer, told the *Los Angeles Times*, "White suburban women? They're not all the same. The Latino community is very diverse with very different interests." This form of microtargeting delved deeper, even bringing volunteers to specific houses within the thick of strongly Republican neighborhoods, and in so doing, moved beyond protocols that had become standard during prior election cycles.

Fliers also targeted the persuadables. As with door knocks, a voter received the flier only if predicted to be influenced, if that voter's mind

(continued)

BEYOND SWING VOTERS (CONTINUED)

was likely to be changed. Traditional marketing sends direct mail to those expected to buy *in light of* contact, rather than *because of* it. It's a subtle difference, but all the difference in the world. Putting it another way, rather than determining whether contact is *a good idea*, persuasion modeling determines whether contact is *a better idea* than not contacting.

Persuasion modeling worked. This method was shown to convince more voters to choose Obama than traditional campaign targeting. "These models showed significant lift over just targeting voters who were undecided or had registered as nonpartisan," Rayid said.

This relative boost came in part by avoiding those voters for whom contact was predicted to backfire (the "do-not-disturb"). As one might expect, for certain voters, campaign contact hurt more than helped.[†] So, during the full-scale efforts ultimately guided by the persuasion models, many such voters were predictively identified and shrewdly left uncontacted.

Persuasion modeling also targeted the campaign's TV ad buying, which delivered a dramatic improvement. A TV spot—such as Fox News in Tampa during evening hours—sells its ad slots by providing a demographic breakdown of its viewers. Team Obama viewed these breakdowns through the filter of their persuasion models in order to decide which spots to hit. Their postcampaign analysis showed this made the TV ad buy 18 percent more effective—they could persuade 18 percent more voters with the same level of investment, which is a meaningful effect given the TV budget magnitude with which they were working: \$400 million.

[†] Even during the analysis of results collected from campaign testing, this is not self-evident from the data, since no individual voter could be both contacted and not contacted to determine which would lead to a better outcome. Detecting the influence of campaign contact, be it positive influence or negative influence, requires modeling, even retrospectively.

BEYOND SWING VOTERS (CONTINUED)

Unsurprisingly, 2016 presidential campaigns are gearing up to apply persuasion modeling. The specifics are well-guarded secrets, but the trend is undeniable. Even as early as July 2015, Hillary Clinton's "analytics team is looking for data nerds," said her campaign website. Shown as one of 11 campaign job categories, analytics included five types of open roles. Analytics job postings for the campaign on relevant industry portals enlisted staff for "helping the campaign determine which voters to target for persuasion." Bernie Sanders' campaign website included "Director of Data and Analytics" as one of only five posted job listings.

Years after the 2012 election, Daniel Porter's perspective hasn't changed. "It remains clear that persuasion modeling is extraordinarily valuable for political campaigns. In fact, after the experience accrued last time around, it's sure to be done by 2016 campaigns even more effectively than in 2012." There's also going to be better data for this work, at least on the Democratic side. "The DNC is building out further its data infrastructure about voters in battleground states."

It's advanced, it's analytical, but it's not arcane. Persuasion modeling is the final chapter of this book, and has begun a whole new chapter for politics.

Afterword

Eleven Predictions for the First Hour of 2022

What's next is what's next. . . . Predictive analytics is where business intelligence is going.

—Rick Whiting, *InformationWeek*

Good morning. It's January 3, 2022, the first workday of the year. As you drive to the office, the only thing predictive analytics (PA) *doesn't* do for you is steer the car (but that's coming soon as well).

1. **Antitheft.** As you enter your car, a predictive model establishes your identity based on several biometric readings, rendering it virtually impossible for an imposter to start the engine.
2. **Entertainment.** Spotify plays new music it predicts you will like.
3. **Traffic.** Your navigator pipes up and suggests alternative routing due to predicted traffic delays. Because the new route has hills and your car's battery—its only energy source—is low, your maximum acceleration is decreased.
4. **Breakfast.** An en route drive-through restaurant is suggested by a recommendation system that knows its daily food preference predictions must be accurate or you will disable it.
5. **Social.** Your Social Techretary offers to read you select Facebook feeds and [Match.com](#) responses it predicts will be of greatest interest.

Inappropriate comments are filtered out. CareerBuilder offers to read postings of jobs for which you’re predicted to apply. When playing your voice mail, solicitations such as robocall messages are screened by a predictive model just like e-mail spam.

6. **Deals.** You accept your smartphone’s offer to read to you a text message from your wireless carrier. Apparently, they’ve predicted you’re going to switch to a competitor, because they are offering a huge discount on the iPhone 12.
7. **Airfare alert.** A notification pipes up informing you that now is the best time to book your flight for a planned trip, since the price is likely to only go up.
8. **Internet search.** As it’s your colleague’s kid’s birthday, you query for a toy store that’s en route. Siri, available through your car’s audio, has improved—better speech recognition and proficiently tailored interaction.
9. **Driver inattention.** Your seat vibrates as internal sensors predict your attention has wavered—perhaps you were distracted by a personalized billboard a bit too long.
10. **Collision avoidance.** A stronger vibration plus a warning sound alert you to a potential imminent collision—possibly with a child running toward the curb or another car threatening to run a red light.
11. **Reliability.** Your car says to you, “Please take me in for service soon, as I have predicted my transmission will fail within the next three weeks.”

PA not only enhances your commute—it was instrumental to making this drive possible in the first place:

Car loan. You could afford this car only because a bank scored you as a low credit risk and approved your car loan.

Insurance. Motion sensors you volunteered to have installed in your car transmit driving behavior readings to your auto insurance company, which in turn plugs them into a predictive model in order to continually

adjust your premium. Your participation in this program will reduce your payment by \$30 this month.

Wireless reliability. The wireless carrier that serves to connect to your phone—as well as your car—has built out its robust infrastructure as guided by demand prediction.

Cybersecurity. Unbeknownst to you, your car and phone avert crippling virus attacks by way of analytical detection.

Road safety. Impending hazards such as manhole explosions, large potholes, and bridge failures have been efficiently discovered and preempted by government systems that predictively target inspections.

No reckless drivers. Dangerous repeat moving violation offenders have been scored as such by a predictive model to help determine how long their licenses should be suspended.

Your health. Predictive models helped determine the medical treatments you have previously received, leaving you healthier today.

TOMORROW'S JUST A DAY AWAY

All the preceding capabilities are available now or have similar incarnations actively under development. Many are delayed more by the (now imminent) integration of your smartphone with your car than by the development of predictive technology itself. The advent of mobile devices built into your watch or glasses will provide yet another multiplicative effect on the moment-to-moment integration of prediction, as well as further accelerating the accumulation of data with which to develop predictive models.

Today, PA's all-encompassing scope already spans the functions that define a society. Organizations—be they companies, governments, law enforcement, charities, hospitals, or universities—undertake many millions of operational decisions in order to enact services. Prediction is key to guiding these decisions, improving the efficiency of mass-scale operations.

Several mounting ingredients promise to spread prediction even more pervasively: bigger data, better computers, wider familiarity, and advancing

science. A growing majority of interactions between the organization and the individual will be driven by prediction.

THE FUTURE OF PREDICTION

Of course, the details and timing of these developments are up to conjecture; PA has not conquered itself. But we can confidently predict more prediction. Every few months another big story about PA rolls off the presses. We're sure to see the opportunities continue to grow and surprise. Come what may, only time will tell what we'll tell of time to come.

Hands-On Guide

Resources for Further Learning

Although this book covers conceptual knowledge required for those interested in becoming a hands-on user, it is not a “how-to.” The next step for a would-be practitioner is to engage with reading and training options that guide getting started hands-on. Below are resources that cover the technical how-to as well as the more advanced underlying theory and math.

FIRST-STOP RESOURCES FOR BUSINESS USERS AND HANDS-ON PRACTITIONERS:

- The Predictive Analytics Guide—articles, industry portals, and other resources: www.pawcon.com/guide
- *The Predictive Analytics Times*—industry news, technical articles, videos, events, and community: www.predictiveanalyticstimes.com
- This book’s website—videos, articles, and other resources: www.thepredictionbook.com

RELATIVELY FRIENDLY HOW-TO BOOKS THAT MANAGE TO BE ACCESSIBLE DESPITE THE TECHNICAL NATURE OF EXECUTING ON PREDICTIVE ANALYTICS:

- Dean Abbott, *Applied Predictive Analytics: Principle and Techniques for the Professional Data Analyst* (Wiley, 2014).
- John W. Foreman, *Data Smart: Using Data Science to Transform Information into Insight* (Wiley, 2013).
- Gordon S. Linoff and Michael J. A. Berry, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (Wiley, 2011).

- Anasse Bari, Mohamed Chaouchi, and Tommy Jung, *Predictive Analytics For Dummies* (For Dummies, a Wiley Brand, 2014).
- Jeffrey Strickland, *Predictive Modeling and Analytics* (lulu.com, 2014).
- Vijay Kotu and Bala Deshpande, *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner* (Morgan Kaufmann, 2014).
- John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (The MIT Press, 2015).

LEADING FOUNDATIONAL TEXTBOOKS FOR PRACTITIONERS AND RESEARCHERS OF PREDICTIVE MODELING (TECHNICAL):

- Robert Nisbet, John Elder, and Gary Miner, *Handbook of Statistical Analysis and Data Mining Applications* (Academic Press, 2009).
- Tom M. Mitchell, *Machine Learning* (McGraw-Hill Science/Engineering/Math, 1997).
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., corr. 3rd printing, 5th printing (Springer, 2009).

TRAINING OPTIONS FOR BUSINESS USERS AND PROSPECTIVE PRACTITIONERS OF PREDICTIVE ANALYTICS:

Note: Although all are how-tos, only some training programs are hands-on.

- Predictive Analytics Applied, the online course instructed by the author. Available on demand at any time: www.businessprediction.com.
- Full-day training workshops alongside the conference Predictive Analytics World. Several international events annually. www.pawcon.com.
- Complete list of degree programs in analytics, data mining, and data science: www.kdnuggets.com/education.
- Sameer Chopra, GVP & Chief Analytics Officer, Orbitz Worldwide, “Sameer Chopra’s Hotlist of Training Resources for Predictive Analytics,” *Predictive Analytics Times*, July 6, 2015. www.predictiveanalyticsworld.com/patimes/hotlist-of-training-resources-for-predictive-analytics0706153.

CONFERENCES FOR BOTH BUSINESS USERS AND PRACTITIONERS OF ANALYTICS:

- **Predictive Analytics World (PAW)**—Founded by this book’s author, PAW is the leading cross-vendor conference series in North America and Europe, which includes advanced training workshop days and the industry-specific events PAW Business, PAW Government, PAW Healthcare, PAW Financial, PAW Workforce, and PAW Manufacturing. See www.pawcon.com.
- **The Predictive Analytics Times Executive Breakfast**—Attendance is free for qualified professionals. See www.PredictiveExecutive.com.
- **Text Analytics World**—The sister event to PAW covering how to make best use of unstructured data, i.e., the majority of data. See www.tawcon.com.

LEADING BOOKS FOR THE BUSINESS USER OF ANALYTICS:

- Thomas H. Davenport and Jeanne G. Harris, *Competing on Analytics: The New Science of Winning* (Harvard Business School Press, 2007).
- James Taylor, *Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics* (IBM Press, 2011).
- Richard Boire, *Data Mining for Managers: How to Use Data (Big and Small) to Solve Business Challenges* (Palgrave Macmillan, 2014).
- Bill Franks, *The Analytics Revolution: How to Improve Your Business by Making Analytics Operational in the Big Data Era* (Wiley, 2014).

A UNIQUE CONCEPTUAL OVERVIEW THAT ACCESSIBLY INTRODUCES THE SCIENTIFIC CONCEPTS, YET ALSO INTERESTS EXPERTS WITH A FRESH PERSPECTIVE—NAMELY THAT MACHINE LEARNING COULD ADVANCE TO AUTOMATICALLY EXTRACT FROM DATA ALL FUTURE HUMAN KNOWLEDGE, ACROSS ALL FIELDS OF SCIENCE:

- Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (Basic Books, 2015).

BROADER, LAY-READER “POP SCIENCE” BOOKS THAT PROVIDE FURTHER INDUSTRIAL AND CULTURAL PERSPECTIVES ON ANALYTICS AND BIG DATA IN GENERAL:

- Patrick Tucker, *The Naked Future: What Happens in a World that Anticipates Your Every Move?* (Current, 2015).
- Luke Dormehl, *The Formula: How Algorithms Solve All Our Problems . . . and Create More* (Perigee Books, 2014).
- Stephen Baker, *The Numerati* (Mariner Books, 2008).
- Ian Ayres, *Super Crunchers: Why Thinking-By-The Numbers is the New Way to Be Smart* (Bantam, 2007).
- Christian Rudder, *Dataclysm: Who We Are (When We Think No One's Looking)* (Crown, 2014).
- Steve Lohr, *Data-sim: The Revolution Transforming Decision Making, Consumer Behavior, and Almost Everything Else* (HarperBusiness, 2015).

DESPITE THE FINAL WORD OF THIS BOOK’S TITLE, *DIE*, NOWHERE DOES THE BOOK GATHER ALL DEATH PREDICTION CASES TOGETHER IN ONE PLACE. FOR A SUMMARY OF THIS SURPRISINGLY DIVERSE RANGE OF APPLICATIONS, SEE THIS ARTICLE:

- Eric Siegel, PhD, “Deathwatch: Five Reasons Organization Predict When you Will Die,” *Predictive Analytics Times*, July 15, 2013. www.predictiveanalyticsworld.com/patimes/deathwatch-five-reasons-organizations-predict-when-you-will-die/.

For extensive further reading, see this book’s Notes—120 pages of citations and comments—available online at www.PredictiveNotes.com.

APPENDIX A

The Five Effects of Prediction

1. The Prediction Effect: *A little prediction goes a long way*. See the Introduction and Chapter 1.
2. The Data Effect: *Data is always predictive*. See Chapter 3.
3. The Induction Effect: *Art drives machine learning; when followed by computer programs, strategies designed in part by informal human creativity succeed in developing predictive models that perform well on new cases*. See Chapter 4.
4. The Ensemble Effect: *When joined in an ensemble, predictive models compensate for one another's limitations so the ensemble as a whole is more likely to predict correctly than its component models are*. See Chapter 5.
5. The Persuasion Effect: *Although imperceptible, the persuasion of an individual can be predicted by uplift modeling, predictively modeling across two distinct training data sets that record, respectively, the outcomes of two competing treatments*. See Chapter 7.

APPENDIX B

Twenty Applications of Predictive Analytics

These applications—ways in which predictive analytics is employed—are covered within the chapters noted. Others beyond these 20 are listed in this book’s Central Tables.

TARGETING DIRECT MARKETING (*SEE CHAPTERS 1 AND 7*)

1. **What’s predicted:** Which customers will respond to marketing contact.
2. **What’s done about it:** Contact customers more likely to respond.

PREDICTIVE ADVERTISEMENT TARGETING (*SEE CHAPTER 1*)

1. **What’s predicted:** Which ad each customer is most likely to click.
2. **What’s done about it:** Display the best ad (based on the likelihood of a click as well as the bounty paid by its sponsor).

BLACK-BOX TRADING (*SEE CHAPTER 1*)

1. **What’s predicted:** Whether a stock will go up or down.
2. **What’s done about it:** Buy stocks that will go up, and sell those that will go down.

PREGNANCY PREDICTION (*SEE CHAPTER 2*)

1. **What’s predicted:** Which female customers will have a baby in coming months.

2. **What's done about it:** Market relevant offers for soon-to-be parents of newborns.

EMPLOYEE RETENTION (SEE CHAPTER 2)

1. **What's predicted:** Which employees will quit.
2. **What's done about it:** Managers take the predictions for those they supervise into consideration, at their discretion. This is an example of decision *support* rather than feeding predictions into an automatic decision process.

CRIME PREDICTION (AKA PREDICTIVE POLICING) (SEE CHAPTER 2)

1. **What's predicted:** The location of a future crime.
2. **What's done about it:** Police patrol the area.

FRAUD DETECTION (SEE CHAPTER 2)

1. **What's predicted:** Which transactions or applications for credit, benefits, reimbursements, refunds, and so on are fraudulent.
2. **What's done about it:** Human auditors screen the transactions and applications predicted most likely to be fraudulent.

NETWORK INTRUSION DETECTION (SEE CHAPTER 2)

1. **What's predicted:** Which low-level Internet communications originate from imposters.
2. **What's done about it:** Block such interactions.

SPAM FILTERING (SEE CHAPTER 2)

1. **What's predicted:** Which e-mail is spam.
2. **What's done about it:** Divert suspected e-mails to your spam e-mail folder.

PLAYING A BOARD GAME (SEE CHAPTER 2)

1. **What's predicted:** Which game board state will lead to a win.
2. **What's done about it:** Make a game move that will lead to a state predicted to lead to a win.

RECIDIVISM PREDICTION FOR LAW ENFORCEMENT (SEE CHAPTER 2)

1. **What's predicted:** Whether a prosecuted criminal will offend again.
2. **What's done about it:** Judges and parole boards consult model predictions when making decisions about an individual's incarceration.

AUTOMATIC SUSPECT DISCOVERY (SEE CHAPTER 2)

1. **What's predicted:** Whether an individual is a "person of interest."
2. **What's done about it:** Individuals with a sufficiently high predictive score are considered or investigated.

CUSTOMER RETENTION WITH CHURN MODELING*(SEE CHAPTERS 4 AND 7)*

1. **What's predicted:** Which customers will leave.
2. **What's done about it:** Retention efforts target at-risk customers.

MORTGAGE VALUE ESTIMATION (SEE CHAPTER 4)

1. **What's predicted:** Which mortgage holders will prepay within the next 90 days.
2. **What's done about it:** Mortgages are valued accordingly in order to decide whether to sell them to other banks.

MOVIE RECOMMENDATIONS (SEE CHAPTER 5)

1. **What's predicted:** What rating a customer would give to a movie.
2. **What's done about it:** Customers are recommended movies that they are predicted to rate highly.

OPEN QUESTION ANSWERING (SEE CHAPTER 6)

1. **What's predicted:** Given a question and one candidate answer, whether the answer is correct.
2. **What's done about it:** The candidate answer with the highest predictive score is provided by the system as its final answer.

**EDUCATION—GUIDED STUDYING FOR TARGETED LEARNING
(SEE CHAPTER 6)**

1. **What's predicted:** Which questions a student will get right or wrong.
2. **What's done about it:** Spend more study time on the questions the student will get wrong.

**TARGETED MARKETING WITH RESPONSE UPLIFT MODELING
(SEE CHAPTER 7)**

1. **What's predicted:** Which customers will be persuaded to buy.
2. **What's done about it:** Target persuadable customers.

**CUSTOMER RETENTION WITH CHURN UPLIFT MODELING
(SEE CHAPTER 7)**

1. **What's predicted:** Which customers can be persuaded to stay.
2. **What's done about it:** Retention efforts target persuadable customers.

**POLITICAL CAMPAIGNING WITH VOTER PERSUASION MODELING
(SEE CHAPTER 7)**

1. **What's predicted:** Which voter will be positively persuaded by political campaign contact such as a call, door knock, flier, or TV ad.
2. **What's done about it:** Persuadable voters are contacted, and voters predicted to be adversely influenced by contact are avoided.

APPENDIX C

Prediction People—Cast of “Characters”

ERIC SIEGEL, PhD—THIS BOOK’S AUTHOR

- Founder of the Predictive Analytics World conference series.
- Executive editor of *The Predictive Analytics Times*.
- Instructor of the online, on-demand training workshop Predictive Analytics Applied (www.businessprediction.com).
- Former computer science professor at Columbia University.
- *For more information, see About the Author.*

JOHN ELDER, PhD

Invested his entire life savings into his own predictive stock market trading system (see Chapter 1).

- CEO and founder, Elder Research, Inc.
- Founding conference chair of Predictive Analytics World for Government, and a frequent keynote speaker at other Predictive Analytics World events.
- Coauthor, *Handbook of Statistical Analysis and Data Mining Applications*.
- Adjunct professor at the University of Virginia.

GITALI HALDER AND ANINDYA DEY

Led a staff-retention project that earmarks each of Hewlett-Packard's almost 350,000 worldwide employees according to "Flight Risk," the expected chance they will quit their jobs (see Chapter 2).

- Analytics practitioners at Hewlett-Packard.
- Backgrounds in statistics and economics.

ANDREW POLE

Led a marketing project at Target that predicts customer pregnancy (see Chapter 2).

- Analytics manager at Target.
- Previously a lead consumer analyst at Hallmark Cards.
- Master's degrees in statistics and economics.
- View his original newsbreaking presentation on pregnancy prediction: www.pawcon.com/Target.

DAN STEINBERG, PHD

Led the prediction of outcome for millions of mortgages at Chase Bank (see Chapter 4).

- President and founder, Salford Systems.
- Entrepreneur who delivers state-of-the-art predictive modeling from the research lab to commercial deployment.
- PhD in economics from Harvard University.
- Former University of California professor.

MARTIN CHABBERT AND MARTIN PIOTTE

With no background in analytics, they took Netflix's \$1 million predictive contest by storm (see Chapter 5).

- Software engineers in the telecommunications industry.

DAVID GONDEK, PHD

Led the design of machine learning integration for IBM's *Jeopardy!*-playing computer, Watson (see Chapter 6).

- Research scientist at IBM Research.
- PhD in computer science from Brown University.

ROGER CRAIG, PHD

Prepared to compete on the TV quiz show *Jeopardy!* by developing a model that predicted which practice questions he'd get wrong in order to target his many hours of studying (see Chapter 6).

- Attained the highest ever *Jeopardy!* one-day winning total and won the show's 2011 Tournament of Champions.
- Over 10 years of experience applying predictive analytics across multiple industries.

EVA HELLE

At a large telecom, Telenor, she predictively optimized how best to persuade each cell phone customer to stay (see Chapter 7).

- Customer analytics lead at Europe's Telenor, the world's seventh-largest mobile operator.
- Master's degree in statistics and marketing.

RAYID GHANI AND DANIEL PORTER

Helped Barack Obama's 2012 presidential campaign employ *persuasion modeling* in order to predict which individual voters would be positively influenced by campaign contact (a call, door knock, flier, or TV ad), and which would be driven to vote adversely by contact (See Chapter 7's sidebar).

- Chief data scientist and director of statistical modeling, respectively, Obama for America 2012 Campaign.
- Experts in persuasion modeling, aka uplift modeling.

About the Author



Eric Siegel, PhD, founder of the Predictive Analytics World conference series and executive editor of *The Predictive Analytics Times*, makes the how and why of predictive analytics understandable and captivating. Eric is a former Columbia University professor—who used to sing educational songs to his students—and a renowned speaker, educator, and leader in the field.

Eric has appeared on Al Jazeera America, Bloomberg TV and Radio, Business News Network (Canada), Fox News, Israel National Radio, NPR Marketplace, Radio National (Australia), and TheStreet. He and this book have been featured in *Businessweek*, *CBS MoneyWatch*, *The Financial Times*, *Forbes*, *Forrester*, *Fortune*, *The Huffington Post*, *The New York Review of Books*, *Newsweek*, *The Seattle Post-Intelligencer*, *The Wall Street Journal*, *The Washington Post*, and *WSJ MarketWatch*.

Eric Siegel is available for select lectures. To inquire: www.ThePredictionBook.com

Interested in employing predictive analytics at your organization?

- Access the author's online, on-demand training workshop, Predictive Analytics Applied: www.businessprediction.com
- Get started with the Predictive Analytics Guide: www.pawcon.com/guide
- Follow Eric Siegel on Twitter: [@predictanalytic](https://twitter.com/predictanalytic)

Index

Notes:

- *PA* stands for *predictive analytics*.
- Page numbers in *italics* followed by “*i*” refer to pages within the Central Tables insert.
- The “n” after a page number refers to an entry that appears in a footnote on that page.

A

Abbott, Dean, 71, 91, 199, 200, 204, 303

AB testing, 261

Accident Fund Insurance, 7*i*

accommodation bookings, 2*i*, 8*i*

actuarial approach, 152

advertisement targeting, predictive,
31, 296

advertising. *See* marketing and advertising

Against Prediction: Profiling, Policing, and

Punishing in an Actuarial Age
(Harcourt), 82

AI. *See* artificial intelligence (AI)

Airbnb, 2*i*, 8*i*

Air Force, 18*i*

airlines and aviation, predicting in, 8, 14*i*,
128

Albee, Edward, 113

Albrecht, Katherine, 53

algorithmic trading. *See* blackbox trading

Allen, Woody, 10

Allstate, 7*i*, 9, 191

Amazon.com

employee security access needs, 12*i*

machine learning and predictive models, 8

Mechanical Turk, 76

personalized recommendations, 5*i*

sarcasm in reviews, 8

American Civil Liberties Union (ACLU),
67, 99

American Public University System, 9, 16*i*
analytics, 16

Analytics Revolution, The (Franks), 305

Ansari X Prize, 187

Apollo 11, 223, 244

Apple, Inc., 115, 223

Apple Mac, 118

Apple Siri, 215, 292

Applied Predictive Analysis (Abbott), 303

“Are Orange Cars Really not Lemons?”
(Elder, Bullard), 143n

Argonne National Laboratory, 13*i*

Arizona Petrified Forest National Park,
259–260

Arizona State University, 9, 16*i*

artificial intelligence (AI)

[Amazon.com](#) Mechanical Turk, 76

mind-reading technology, 8, 19*i*

possibility of, the, 209, 217–219, 248

Watson computer and, 213–219

Asimov, Isaac, 139

astronomy, 113, 189

- AT&T Research BellKor Netflix Prize teams, 188–189, 192–197
- Australia, 14*i*, 16*i*, 106, 120, 178, 191
- Austria, 122, 193
- automatic suspect discovery (ASD), 87–101 approach to, 88–91 arguments for and against, 96–99 assumptions about NSA’s use of, 94–95 challenges of, 99–100 defined, 88 example patterns, 91–94 privacy issues and, 87–88, 100
- automobile insurance crashes, predicting, 9, 191 credit scores and accidents, 120–121 driver inattentiveness, 9, 18*i*, 191 fraud predictions for, 11*i*, 70
- Averitt, 18*i*
- aviation incidents, 9, 14*i*
- Aviva Insurance (U.K.), 11*i*
- AWK computer language, 111
- Ayres, Ian, 306
- B**
- backtesting, 39–41. *See also* test data
- Baesens, Bart, 182
- bagging (bootstrap aggregating), 200–203
- Baker, Stephen, 306
- Baltimore, 11*i*
- Bangladesh, 140, 167, 174
- Barbie dolls, 118
- Bari, Anasse, 304
- Bayes, Thomas (Bayes Network), 31
- Beane, Billy, 43
- Beano, 57
- Beaux, Alex, 61–62, 63
- behavioral predictors, 115
- Bella Pictures, 5*i*
- BellKor, 188–189, 193, 195–197
- BellKor Netflix Prize teams, 188–189, 192–197
- Ben Gurion University (Israel), 79, 125
- Bernstein, Peter, 151
- Berra, Yogi, 268
- Berry, J. A., 303
- Big Bang theory, 4, 113
- Big Bang Theory, The* (TV show), 18
- Big Brother, 85
- BigChaos team, 192–197
- big data about, xviii, 16, 105, 114 The Data Effect and, 115, 135–145, 295 “wider” data and, 140–141
- billing errors, predicting, 10*i*
- black-box trading, 8*i*, 24, 38–44
- Black Swan, The* (Taleb), xviii, 168
- Blue Cross Blue Shield of Tennessee, 7*i*, 10*i*
- BMW, 23, 79
- BNSF Railway, 13*i*
- board games, predictive play of, 76, 297
- Bohr, Niels, 13
- Boire, Richard, 305
- Bollen, Johan, 107n
- books, about Predictive Analytics, 303–304, 305, 306
- book titles, testing, 262
- Bowie, David, 32
- brain activity, predicting, 19*i*
- Brandeis, Louis, 55
- Brazil, 98
- breast cancer, predicting, 9*i*, 10
- Breiman, Leo, 178, 200
- Brigham Young University, 10, 10*i*
- British Broadcasting Corporation, 16*i*, 18*i*
- Brobst, Stephen, 151
- Brooks, Mel, 38
- Brynjolfsson, Eric, 111–112
- buildings, predicting fault in, 13*i*
- Bullard, Ben, 143n
- burglaries, predicting, 11*i*, 67
- business rules, decision trees and, 161, 164
- buying behavior, predicting, 3*i*, 48–49, 121
- C**
- Cage, Nicolas, 28, 262
- Canadian Automobile Association, 14*i*
- Canadian Tire, 7*i*, 121

- car crashes and harm, predicting, 9, 10*i*, 191
CareerBuilder, 20*i*, 191
Carlin, George, 212
Carlson, Gretchen, 52
Carnegie Mellon University, 57, 228, 285
cars, “orange lemons,” 104–109, 142–143
car services passenger destination, predicting, 14*i*
CAR T decision trees, 178–181, 200–203
causality, 35, 131–135, 158, 257
cell phone industry
 consumer behavior and, 120
 dropped calls, predicting, 8–9, 14*i*, 204
 GPS data and location predicting, 2*i*, 57, 84
 Telenor (Norway), 5*i*, 252–254, 281
CellTel (African telecom), 125
Centers for Medicare and Medicaid Services, 11*i*
Cerebellum Capital, 8*i*
Chabbert, Martin, 185–189, 192, 301
Chaffee, Alexander Day, 171, 244
CHAID, 178
Chaouchi, Mohamed, 304
Charlotte Rescue Mission, 15*i*
Chase Bank
 actuarial approach used by, 152
 churn modeling, 166, 252*n*
 data, learning from, 153–156
 economic recession and risk, 181–182
 learning from data, 162–163
 mergers and growth of, 149, 181
 microrisks faced by, 149–153
 mortgage risk decision trees, 163–165, 179–181
 mortgage risks, predicting, 5*i*, 7*i*, 149–150
 mortgage value estimation, 180–181
 predictive models and success, 181
 risks faced by, 149–153, 181
 Steinberg as leader, 3, 147–148, 153–156, 175–176, 178–181
cheating, predicting, 11–12, 18*i*, 76
check fraud, predicting, 11, 11*i*
chess-playing computers, 76, 222–223
Chicago Police Department, 12, 12*i*
Chopra, Sameer, 304
Chu-Carroll, Jennifer, 244, 246
churn modeling, 166, 252–254, 277, 281, 298, 299
Cialdini, Robert, 259–260
Citibank, 8, 18*i*
Citigroup, 7*i*
Citizens Bank, 11, 11*i*, 69
City of Boston, 15*i*
City of Chicago, 14*i*, 15*i*
City of New York, 12*i*, 14*i*, 16*i*
city power, predicting fault in, 13*i*
city regulations, crime predictions and, 11*i*
civil liberties, risks to, 47–48, 80
Clarke, Arthur C., 209
clicks, predicting, 7, 20*i*, 25, 29, 262
clinical trial recruitment, predicting, 10*i*
Clinton, Hillary, 8, 289
cloning customers, 156*n*
Coase, Ronald, 167
Colbert, Stephen, 53, 283
Colburn, Rafe, 74
collection agencies, PA for, 10, 13*i*
collective intelligence, 197–199
collective learning, 17–18
Columbia University, 12, 79, 125, 248
Commander Data, 77
company networks, predicting fault in, 13*i*
Competing on Analytics: The New Science of Winning (Davenport and Harris), 37–38, 305
competitions. *See* PA competitions
computational linguistics, 210. *See also*
 natural language processing (NLP)
computer “intelligence.” *See* artificial intelligence (AI)
computer programs, decision trees and, 161–162

- computers
 creation of, 18–19
 Deep Blue computer, 76, 222–223
 human brain *versus*, 19–20
 viruses on, predicting, 12*i*, 73
See also artificial intelligence; machine learning; Watson computer *Jeopardy!* challenge
 computer science, 74
 Con Edison, 6, 13*i*
 conferences, 305
 Congo massacres, 125
 consumer behavior, insights on, 118–119
 contests. *See PA competitions*
 Continental Airlines, 14*i*
 control sets, 267, 268
 corporate roll-ups, 195n
 couponing, predictive, 6–7
 Coursera, 153n
 Cox Communications, 4*i*
 Craig, Roger, 17*i*, 225, 302
 credit cards
 fraud detection for, 69
 payment systems, predicting fault in, 10, 13*i*
 credit risk
 credit scores and, 149–151
 typing and, 121–122
 Crichton, Michael, 172
 crime fighting and fraud detection
 cheating in board games, 76
 credit card fraud, 69
 in financial institutions, 69–71
 fraud, defined, 68
 fraudulent transactions, 69
 IRS tax fraud, 11*i*, 12, 71, 204
 machine injustice, 79–80
 network intrusion, 12*i*, 72
 PA application, 70, 297
 PA for, 8*i*–9*i*, 11, 68–77
 prejudice in, risk of, 81–82
 spam filtering, 74
 crime prediction for law enforcement
 automatic suspect discovery (ASD) and NSA, 87–101
 cybercrime, 73
 ethical risks in, 79–82
 fighting, PA for, 8*i*–9*i*, 11, 67
 judicial decisions and, 77–80, 160–161
 murder, predicting, 11, 11*i*, 12*i*, 18*i*, 78
 PA application, 66–67, 297
 PA examples and insights, 125
 prediction models for, pros and cons of, 83–86
 recidivism, 12*i*, 77–80
 Uber and behavior, 118
 crowdsourcing
 collective intelligence and, 197–199
 Kaggle PA crowdsourcing contests, 189–191, 204
 noncompetitive crowdsourcing, 190n
 PA and, 185, 187–191, 197–200, 224
 Cruise, Tom, 79, 152
 customer need, predicting fault in, 14*i*
 customer retention
 cancellations and predicting, 4*i*, 8
 with churn modeling, 166, 252–254, 277, 298
 with churn uplift modeling, 277, 299
 contacting customers and, 119–120
 feedback and dissatisfaction, 18*i*
 customization, perils of, 30–31
 cybercrime, 73
 cybernetics, xvii
- D**
- D'Arcy, Aoife, 304
 data, sharing and using
 automatic suspect discovery (ASD) and NSA, 87–101
 battle over mining, 56–57
 choosing what to collect, 6–12, 56–57, 249, 254–256
 growth and rates of expansion, 109–114
 policies for, 55

- privacy concerns and, 52–54, 58–60, 83–86
- data, types and sources of
big data, xviii, 16, 105, 114
buzzwords for, 114
consumption and, 6–7
employee data, 61
fake data, 68–69
free public data, 110
learning data, 162–163
location data, 2*i*, 57, 84
medical data, 85
personal data, 54–55, 57
social media posts, 9*i*
textual data, 210, 214
- data, value of
about, 4–5
learning from data, 153–156
machine learning and, 4–5
personal data, 54–55, 57
as “the new oil,” 115
- Dataclysm* (Rudder), 306
- Data Effect, The, 115, 135–145, 295
- data gluts, 112
- Data.gov, 110
- data hustlers and prospectors, 57, 85
- data mining, 17, 36, 58, 114
- Data Mining for Managers* (Boire), 305
- Data Mining Techniques* (Linoff, Berry), 303
- data preparation phase, 162–163
- data science, 16–17, 114n
- Data-sim* (Lohr), 306
- Data Smart* (Foreman), 303
- data storage, 56–57, 112
- dating websites
attractiveness ratings and success, 127
consumer behavior on, 118–119
predicting on, 3*i*, 7, 110, 127, 259, 291–292
- Davenport, Thomas H., xix, 37–38, 305
- death predictions, 9*i*, 10*i*, 10–11, 84–85, 306
- “Deathwatch” (Siegel), 306
- deception, predicting, 11–12, 18*i*
- Decision Management Systems* (Taylor), 305
- decision trees
CART decision trees, 178–181, 200–203
circle single model, 201–202
decision boundaries in, 202
getting data from, 175–179
in machine learning, 156–162
methods competing with, 171–172
mortgage risk decision trees, 163–165, 180–181
overlearning and assuming, 167–169
random forests, 201
uplift trees, 275
- deduction *versus* induction, 169–170
- Deep Blue computer, 76, 222–223
- DeepQA, 247
- deliveries, predicting, 14*i*
- Delta Financial, 38–41
- Deming, William Edwards, 4
- Democratic National Committee (DNC), 283–288
- Deshpande, Bala, 304
- Dey, Anindya, 59, 61, 301
- Dhar, Vasant, 17
- diapers and beer, behavior and, 117
- Dick, Philip K., 28
- differential response modeling. *See* uplift modeling
- discrimination, risks of, 81–82
- disease, predicting, 9*i*
- Disraeli, Benjamin, 167
- divorce, 7, 53–54
- dolls and candy bars, behavior and, 118
- Domingoes, Pedro, 305
- donations and giving, predicting, 15*i*, 204
- do-overs, 257–259
- Dormehl, Luke, 306
- downlift, 273
- driver inattentiveness, predicting, 9, 18*i*, 191
- driverless cars, 23, 79
- Drucker, Peter, 149
- drugs effects and use, predicting, 10*i*

- “Dry Bones” (song), 114
 DTE Energy, 8*i*
 Dubner, Stephen J., 27, 72, 81, 91n, 92, 95
 Duhigg, Charles, 51–52
 dunnhumby, 191
 Dyson, George, 214
- E**
[Earthlink.com](#), 119
 Echo Nest, 126
 economic recession, 181–182
 education
 grades, predicting, 16*i*, 191
 guided studying for targeted learning, 225, 299
 PA for, 8, 15*i*–17*i*
 student dropout risk, predicting, 9, 16*i*
 student knowledge and performance, predicting, 17*i*, 191
 eHarmony, 7
 Eindhoven University, 16*i*
 Eindhoven University, Netherlands, 9
 Einstein, Albert, 4, 175
 Elder, John, 8*i*, 143n, 300, 304
 about, 23, 39, 43–44
 “Are Orange Cars Really not Lemons?,” 143n
 black-box trading systems, 24, 38–44
 on employee death predictions, 84
 on generalization paradox, 204–205
 in Netflix Prize competition, 192
 on passion for science, 44–45
 on power of data, 54
 risks taken by, 38–41
 on “vast search,” 140
 Elder Research, Inc., 45, 71, 132, 143n, 192
 elections, crime rates and, 125
 electoral politics
 Hillary for America 2016 Campaign, 15*i*
 musical taste political affiliation, 125
 Obama for America 2012 Campaign, 15*i*
 Obama for America Campaign, 8, 282–288, 286–288
 uplift modeling applications for, 277, 286–288
 voter persuasion, predicting, 15*i*, 160, 282–288
 electronic equipment, predicting fault in, 13*i*
Elements of Statistical Learning, The
 (Hastie, Tibshirani, Friedman), 304
 Elie Tahari, 4*i*
 e-mail
 consumer behavior and addresses for, 119–120
 government storage of, 113
[Hotmail.com](#), 34, 109, 119
 phishing e-mails, 74
 spam filtering for, 74
 emotions
 in blog posts, 107
 human behavior and, 106–108
 mood predictions and, 106–108
 See also human behavior
 employee longevity, predicting, 20*i*
 employees and staff
 job applications and positions, 20*i*, 191
 job performance, predicting, 20*i*, 204
 job promotions and retention, 128
 job skills, predicting, 20*i*
 LinkedIn for career predictions, 7, 20*i*
 privacy concerns and data on, 84
 quitting, predicting, 9, 20*i*, 58–64, 128
 Energex (Australia), 6, 16*i*
 energy consumption, predicting, 16*i*
Ensemble Effect, The, 205, 275n, 295
Ensemble Experts, 192
 ensemble models
 about, 204
 automatic suspect discovery (ASD) and, 93n
 CART decision trees and bagging, 178–181, 200–203
 collective intelligence in, 199–200
 complexity in, 204

- crowdsourcing and, 185, 187–191, 197–200, 224
generalization paradox and, 204
IBM Watson question answering computer and, 18*i*, 204
IRS (tax fraud), 11*i*, 12, 71, 204
meta-learning and, 193–195
Nature Conservancy (donations), 15*i*, 204
Netflix (movie recommendations), 5*i*, 6, 183, 186, 204
Nokia–Siemens Networks (dropped calls), 14*i*, 204
University of California, Berkeley (brain activity), 19*i*, 204
for uplift modeling, 275n
U.S. Department of Defense (fraudulent invoices), 11*i*
U.S. Department of Defense Service (fraudulent invoices), 204
U.S. Special Forces (job performance), 20*i*, 204
Ensemble team, 197
Epagogix, 8*i*
erectile dysfunction, 10*i*
Experian, 150
Exxon Mobil Corp., 115
- F**
“Fab Four” inventors, 178, 200
Facebook
 data glut on, 112
 data on, 4, 29, 55
 fake data on, 55
 friendships, predicting, 2*i*, 191
 happiness as contagious on, 124
 job performance and profiles on, 20*i*
 social effect of, 124
 student performance PA contest, 17*i*
facial recognition, 2*i*
Failure of Risk Management, The (Hubbard), 151–152
false conclusions, avoiding, 104–109, 142–143
false positives (false alarms), 79, 140
family and personal life, PA for, 2*i*–3*i*, 7, 54, 124
Farrell, Colin, 79
fault detection for safety and efficiency, PA for, 13*i*–14*i*
Federal Trade Commission, 69
FedEx, 4*i*, 9
Femto-photography, 112
Ferguson, Andrew, 101
Ferrucci, David, 217, 243, 246
FICO, 10*i*, 150
Fidelity Investments, 275
finance and accounting, fraud detection in, 11, 69–71
finance websites, behavior on, 118–119
financial risk and insurance, PA for, 7*i*–8*i*, 11
Fingerhut, 4*i*
Finland, 124
fire, predicting, 16*i*
First Tennessee Bank, 3*i*, 4*i*
Fisher, Ronald, 135, 136n
Fleming, Alexander, 139
flight delays, predicting fault in, 14*i*
Flight Risks, predicting, 8, 47, 58–64
Flirtback computer program, 111
Florida Department of Juvenile Justice, 12*i*
fMRI brain scans, 19*i*
Foldit, 190n
Food and Drug Administration (FDA), 279
Fooled by Randomness (Taleb), 136–137, 138, 168
Ford Motor Co., 9, 18*i*, 191
forecasting, 17, 284
Foreman, John W., 303
Formula, The (Dormehl), 306
Fox & Friends (TV show), 52
Franklin, Benjamin, 29, 75
Franks, Bill, 305
fraud, defined, 68
fraud detection, 48, 61, 68–77, 297. *See also* crime fighting and fraud detection

- Freakonomics Radio, 27, 72, 81, 140
Freakonomics Radio, 140
- frequency, 116
- Friedman, Jerome, 178, 304
- friendships, predicting, 3*i*, 191
- Fukuman, Audrey, 159
- Fulcher, Christopher, 66
- Fundamentals of Machine Learning for Predictive Data Analytics* (Kelleher, MacNamee, D'Arcy), 304
- fund-raising, predicting in, 16*i*
- Furnas, Alexander, 56, 83
- future, views on
- human nature and knowing about, xxi
 - predictions for 2022, 291–293
 - uncertainty of, 12–13
- G**
- Galileo, 109
- Gates, Bill, 137
- generalization paradox, 204
- Ghani, Rayid, 284–285, 302
- Gilbert, Allen, 99
- Gladwell, Malcolm, 30–31
- GlaxoSmithKline (U.K.), 10*i*
- Goethe, Johann Wolfgang von, 36
- Goldbloom, Anthony, 190, 204
- Gondek, David, 207, 224, 227–231, 234–237, 240n, 245–248, 302
- Google
- mouse clicks, measuring for predictions, 7, 29, 262
 - privacy policies, 84
 - Schmidt, 83, 84
 - searches for playing *Jeopardy!*, 214
 - self-driving cars, 23, 79
 - spam filtering, 74
- Google Adwords, 30, 262
- Google Flu Trends, 9*i*, 123
- Google Page Rank, 199
- government
- data storage by, 110
 - fraud detection for invoices, 11*i*, 71, 204
- PA for, 15*i*–17*i*
- public access to data, 110
- See also individual names of U.S. government agencies*
- GPS data, 2*i*, 57, 84
- grades, predicting, 16*i*, 191
- grant awards, predicting, 16*i*, 191
- Greenwald, Glenn, 98
- Grockit, 17*i*, 191
- Groundhog Day* (film), 258
- Grundhoefer, Michael, 267, 268–269
- H**
- hackers, predicting, 12*i*, 73
- HAL (intelligent computer), 209–210
- Halder, Gitali, 59, 61, 301
- Handbook of Statistical Analysis and Data Mining Applications* (Nisbet, Elder, Miner), 304
- Hansell, Saul, 182
- happiness, social effect and, 107, 124
- Harbor Sweets, 4*i*, 6
- Harcourt, Bernard, 82
- Harrah's Las Vegas, 4*i*
- Harris, Jeanne, 37–38, 305
- Harvard Medical School, 9
- Harvard University, 123
- Hastings, Reed, 197
- healthcare
- death predictions in, 9*i*, 10–11, 84–85
 - health risks, predicting, 10*i*
 - hospital admissions, predicting, 10*i*
 - influenza, predicting, 9*i*, 124
 - insurance companies, predicting, 7*i*, 9*i*
 - medical research, predicting in, 124
 - medical treatments, risks for wrong predictions in, 267–269, 277, 279
 - medical treatments, testing persuasion in, 262
- PA for, 9*i*–10*i*, 10–11, 124
- personalized medicine, uplift modeling applications for, 277, 279
- health insurance companies, PA for, 9*i*–10*i*, 10–11, 84–85

- Hebrew University, 18*i*
Heisenberg, Werner Karl, 262
Helle, Eva, 280, 281, 282, 302
Helsinki Brain Research Centre, 124
Hennessey, Kathleen, 282, 286
Heraclitus, 257
Heritage Health Prize, 191
Heritage Provider Network, 10, 10*i*
Hewlett Foundation, 16*i*, 189, 191
Hewlett-Packard (HP)
 employee data used by, 61
 financial savings and benefits of PA, 61, 64
Global Business Services (GBS), 62
quitting and Flight Risks, predicting, 8, 20*i*, 48, 58–64, 128
sales leads, predicting, 5*i*
turnover rates at, 60
warranty claims and fraud detection, 11, 11*i*
Hillary for America 2016 Campaign, 8, 15*i*
HIV progression, predicting, 9*i*, 191
HIV treatments, uplift modeling for, 279
Hollifield, Stephen, 67
Holmes, Sherlock, 33, 34, 83, 86
Hopper, 8*i*
hormone replacement, coronary disease
 and, 133
hospital admissions, predicting, 10*i*, 10–11
Hotmail.com, 34, 109, 119
House (TV show), 83
“How Companies Learn Your Secrets”
 (Duhigg), 51–52
Howe, Jeff, 187, 189
HP. *See* Hewlett-Packard (HP)
Hubbard, Douglas, 109, 151–152
human behavior
 collective intelligence, 197–199
 consumer behavior insights, 119–120
 emotions and mood prediction, 199–200
 mistakes, predicting, 11–12
 social effect and, 11–12, 120, 124
human genome, 113
human language
 natural language processing (NLP), 210, 219–221, 227–231
PA for, 18*i*–19*i*
persuasion and influence in, 259–260
human resources. *See* employees and staff
- I**
- IBM
 corporate roll-ups, 195n
 Deep Blue computer, 76, 224
 DeepQA project, 247
 Iambic IBM AI, 248
 mind-reading technology, 8
 natural language processing research, 227
 sales leads, predicting, 5*i*
 student performance PA contest, 17*i*
T. J. Watson Research Center, 224
value of, 222–223
See also Watson computer *Jeopardy!*
challenge
ID3, 178
impact modeling. *See* uplift modeling
Imperium, 18*i*, 191
inappropriate comments, predicting, 18*i*
incremental impact modeling. *See* uplift modeling
incremental response modeling. *See* uplift modeling
India, 125
Indiana University, 107
Induction Effect, The, 179, 295
induction *versus* deduction, 169–170
inductive bias, 170
ineffective advertising, predicting, 20*i*
infidelity, predicting, 122
Infinity Insurance, 7*i*, 16*i*
influence. *See* persuasion and influence
influenza, predicting, 9*i*, 124
information technology (IT) systems,
 predicting fault in, 13*i*
InnoCentive, 190n
insults, predicting, 18*i*, 191

- insurance claims
 automobile insurance fraud, predicting, *7i*, 10–11, 70, 121, 191
 death predictions and, *7i*, 10–11, 84–85
 financial risk predicting in, *8i*
 health insurance, *9i*, 10–11, 84–85
 life insurance companies, 11
 life insurance companies, *7i*
 Integral Solutions Limited, 195n
 Internal Revenue Service (IRS), *11i*, 12, 71, 106, 204
 International Conference on Very Large Databases, *114*
 Iowa State University, 9, *16i*
 iPhone. *See* Apple Siri
 Iran, *13i*
 Israel Institute of Technology, *12i*
- J**
 Japan, 124
 Jennings, Ken, 226, 239, 240, 246
Jeopardy! (TV show). *See* Watson computer
Jeopardy! challenge
 Jevons, William Stanley, 169
 Jewell, Robert, 248
 jobs and employment. *See* employees and staff
 Jones, Chris, 238
Journal of Computational Science, 107n
 JPMorgan Chase. *See* Chase Bank
 judicial decisions, crime prediction and, 79, 160–161
 Jung, Tommy, 304
Jurassic Park (Crichton), 172
 Just Giving, *15i*
- K**
 Kaggle, 189–191, 204
 Kane, Katherine, 276
 Kasparov, Garry, 76, 223
 KDnuggets, 84, 110
 Keane, Bil, xxix
 “keep it simple, stupid” (KISS) principle, 178, 200
 Kelleher, John D., 304
 Khabaza, Tom, 115
 killing, predicting, 11, *11i*, *18i*
 King, Eric, 248, 269n
 Kiva, *8i*
 Kmart, 7
 knee surgery choices, 124
 knowledge (for education), predicting, *13i*
 Kotu, Vijay, 304
 Kretsinger, Stein, 132, 133, 189
 Kroger, 7
 Kuneva, Meglena, 4, 115
 Kurtz, Ellen, 78, 82
 Kurzweil, Ray, 113
- L**
 language. *See* human language
 Lashkar-e-Taiba, 92
 law enforcement. *See* crime prediction for law enforcement
 lead poisoning from paint, predicting, *16i*
 learning
 about, 17–21
 collective learning, 17–18
 education—guided studying for targeted learning, 225, 299
 learning from data, 162–163
 memorization *versus*, 169
 overlearning, avoiding, 167–169, 175, 200, 204
 Leinweber, David, 167
 Leno, Jay, 12
 Levant, Oscar, 217
 Levitt, Stephen, 72, 81, 91n, 92, 95
 Lewis, Michael, 15
 lies, predicting. *See* deception, predicting
Lie to Me (TV show), 12
 life insurance companies, PA for, *7i*, 11
 Life Line Screening, *4i*
 lift, 166, 266n
 Lindbergh, Charles, 223
 LinkedIn
 friendships, predicting, *3i*

- job skills, predicting, 7, 20*i*
Linoff, Gordon S., 303
Linux operating systems, 190n
Lloyds TSB, 5*i*
loan default risks, predicting, 7*i*, 8*i*
location data, 2*i*, 57, 84
logistic regression, 236
Lohr, Steve, 306
London Stock Exchange, 8*i*
Los Angeles Police Department, 12*i*, 67
Lotti, Michael, 83
Loukides, Mike, 17
love, predicting, 7, 110–111, 126–127, 259, 291–292
Lynyrd Skynyrd (band), 253
- M**
- MacDowell, Andie, 258
machine learning
about, 4–5, 19–21, 147–148, 156–159, 171–173, 204
courses on, 153n
in crime prediction, 79–80
data preparation phase for, 162–163
decision trees in, 156–162
induction and, 169
induction *versus* deduction, 169–170
learning data, 162–163
learning from mistakes in, 156
learning machines, building, 153–156
overlearning, 167–169, 175, 200, 204
predictive models, building with, 36, 41
silence, concept of, 20n
testing and validating data, 173–175
univariate *versus* multivariate models, 154–156, 156–157
See also Watson computer *Jeopardy!*
challenge
machine risk, 79–80
MacNamee, Brian, 304
macroscopic risks, 182
Mac *versus* Windows users, 118
Madrigal, Alexis, 54
- Magic 8 Ball toy, 81
Mao, Huina, 107n
maritime incidents, predicting, 14*i*
marketing and advertising
banner ads and consumer behavior, 119
mouse clicks and consumer behavior, 7, 29, 262
targeting direct marketing, 26, 296
marketing models
do-overs in, 257–258
messages, creative design for, 259–262
Persuasion Effect, The, 276
quantum humans, influencing, 262–266
response uplift modeling, 270–271, 275
marketing segmentation, decision trees and, 161–162
marriage and divorce, predicting, 7, 53–54, 122
Mars Climate Orbiter, 39–40, 245
Martin, Andres D., 160
Maryland, crime predictions in, 11, 11*i*, 78
Massachusetts Institute of Technology (MIT), 227
Master Algorithm, The (Domingos), 305
Match.com, 3*i*, 7, 291–292
Matrix, The (film), 213
McCord, Michael, 231n
McKinsey reports, 190–191
McNamara, Robert, 269
Mechanical Turk, 76
medical claims, fraudulent, 11*i*
medical treatments. *See* healthcare
memorization *versus* learning, 169
Memphis (TN) Police Department, 12, 12*i*, 67
metadata, 91–92, 106, 106n
meta-learning, 193–195
Mexican Tax Administration, 71
Miami Police Department, 12*i*
Michelangelo, 175
microloan defaults, predicting, 16*i*
Microsoft, 2*i*, 223
Milne, A. A., xxix

- Mimoni (Mexico), 7*i*
 mind-reading technology, 8, 19*i*
 Miner, Gary, 304
Minority Report (film), 79
 Missouri, 79
 Mitchell, Tom, 57, 170, 177, 178, 234, 304
 mobile operators. *See* cell phone industry
 moneyballing, concept of, 15, 224–226, 282
 mood labels, 106–108
 mood prediction, blogs and, 107
 mortgage prepays and risk, predicting, 7*i*, 149–151
 mortgage risk decision trees, 163–165, 180–181
 mortgage value estimation, 180–181, 298
 mouse clicks, predicting, 7, 29, 262
 movie hits, predicting, 8*i*, 20*i*
 movie recommendations, 6, 183, 186, 204, 298
 movies, 20*i*
MTV, 21*i*
 MultiCare Health System (Washington State), 10*i*
 “multiple comparisons problem”/multiple comparisons trap”, 140
 multivariate models, 154–156, 156–157
 murder, predicting, 11, 11*i*, 12*i*, 18*i*, 78
 Murray, Bill, 258
 music, stroke recovery and, 124
 musical taste, political affiliation and, 126
 Muslims, 81–82
- N**
- Naïve Bayes, 31
Naked Future, The (Tucker), 306
 NASA
 Apollo 11, 223, 244
 Mars Climate Orbiter, 39–40, 245
 PA contests sponsored by, 187, 189
 on space exploration, 76
 National Insurance Crime Bureau, 69
 National Security Agency (NSA), 12*i*, 87–101
 National Transportation Safety Board, 9
 natural language processing (NLP), 210, 219–221, 227–231
 Nature Conservancy, 15*i*, 204
 Nazarko, Edward, 222
Nerds on Wall Street (Leinweber), 167
 Netflix movie recommendations, 5*i*, 6, 183, 186, 204
 Netflix Prize
 about, 185–187
 competition and winning, 192
 crowdsourcing and PA for, 185–187, 187–191, 197–200, 223
 meta-learning and ensemble models in, 193–195, 204
 Pragmatic Theory team, 188–189, 196–197
 Netherlands, 16*i*
 net lift modeling. *See* uplift modeling
 net response modeling. *See* uplift modeling
 net weight of evidence (NWOE), 272
 network intrusion detection, 12*i*, 72
 New South Wales, Australia, 191
 New York City Medicaid, 11*i*
 New York State, 11*i*
New York Times, The, 279
Next (Dick), 28, 262
 Ng, Andrew, 153n
 Nightcrawler (superhero), 54
Nineteen Eighty-Four (Orwell), 85
 99designs, 190n
 Nisbet, Robert, 304
 “no free lunch” theorem, 170n
 Nokia, 2*i*
 Nokia-Siemens Networks, 14*i*, 204
 nonprofit organizations, PA for, 15*i*–17*i*
 Noonan, Peggy, 286
No Place to Hide (Greenwald), 98
 Northwestern University Kellogg School of Management, 117
 nuclear reactors, predicting fault in, 13*i*
 null hypothesis, 136n
Numerati, The (Baker), 306

O

Obama for America 2012 Campaign, 8, 15*i*, 282–288
observation, power of, 33–36
Occam’s razor, 178, 200
O’Connor, Sandra Day, 160
office equipment, predicting fault in, 13*i*
Oi (Brasil Telecom), 8*i*
oil flow rates, predicting, 13*i*
oil refinery safety incidents, predicting, 13*i*
OkCupid, 3*i*, 7, 126–127, 259
Oklahoma State University, 9, 16*i*
O’Leary, Martin, 189
Olshen, Richard, 178
1–800-FLOWERS, 71
1-sided equality of proportions hypothesis test, 137n
Online Privacy Foundation, 18*i*, 191
Oogway, xxix
open data movement, 110
open question answering, 213–222, 226, 238
open source software, 190n
Optus (Australia), 4*i*, 106, 120
“orange lemons” (cars), 104–109, 142–143
Orbitz, 118
Oregon, crime prediction in, 11, 12*i*, 78
organizational learning, 17–18
organizational risk management, 28
Orwell, George, 85
Osco Drug, 117
overfitting. *See* overlearning
overlearning, 167–169, 175, 200, 204
Oz, Mehmet, 27

P

PA (predictive analytics)
about, xxx–xxxii, 15–17, 116
assumption about NSA’s use of, 94–96
choosing what to predict, 6–12, 55–56, 249, 254–256
in crime fighting and fraud detection, 8*i*–9*i*

crowdsourcing and, 185, 187–191, 197–200, 224
defined, 15, 152
in family and personal life, 2*i*–3*i*
in fault detection for safety and efficiency, 13*i*–14*i*
in finance and accounting fraud detection, 11, 69–71, 121
in financial risk and insurance, 7*i*–8*i*
forecasting *versus*, 17, 284
frequently asked questions about, xxii–xxvii
in government, politics, nonprofit, and education, 15*i*–17*i*
in healthcare, 9*i*–10*i*, 10–11, 124
in human language understanding, thought, and psychology, 18*i*–19*i*
launching and taking action with PA, 23–26
in law enforcement and fraud detection, 11*i*–12*i*
in marketing, advertising, and the Web, 4*i*–6*i*
“orange lemons” and, 104–109, 142–143
overview, xxi–xxii
risk-oriented definition of, 152
text analytics, 209, 214
in workforce (staff and employees), 20*i*
PA (predictive analytics) applications
black-box trading, 24, 43
blog entry anxiety detection, 107
board games, playing, 76, 292
credit risk, 149–151, 277
crime prediction, 66–67, 297
customer retention with churn modeling, 166, 252–254, 277, 298, 299
customer retention with churn uplift modeling, 277, 299
defined by, 26
education—guided studying for targeted learning, 226, 299
employee retention, 58–64, 297
fraud detection, 70–71, 297

- PA (predictive analytics) applications
(continued)
- mortgage value estimation, 180, 298
 - movie recommendations, 186, 298
 - network intrusion detection, 72, 297
 - open question answering, 220, 238
 - political campaigning with voter persuasion modeling, 285, 299
 - predictive advertisement targeting, 31, 296
 - pregnancy prediction, 48, 296–297
 - recidivism prediction for law enforcement, 78, 298
 - spam filtering, 74, 297
 - targeting direct marketing, 26–27, 296
 - uplift modeling applications, 277
- See also* Central Tables insert
- PA (predictive analytics) competitions and contests
- in astronomy and science, 189
 - for design and games, 190n
 - for educational applications, 189
 - Kaggle crowdsourcing contests, 189–191, 204
 - Netflix Prize, 185–187, 204
- PA (predictive analytics) insights
- consumer behavior, 119–120
 - crime and law enforcement, 125
 - finance and insurance, 121
 - miscellaneous, 126–130
- Palmisano, Sam, 247
- Panchoo, Gary, 198
- Pandora, 21*i*
- parole and sentencing, predicting, 11, 12*i*, 77–78
- Parsons, Christi, 282, 286
- PAW (Predictive Analytics World)
- conferences, xxii, xxvii, 48–49, 61, 70–71, 115, 197–199, 227, 240n, 305
- payment processors, predicting fault in, 13*i*
- PayPal, 8, 18*i*, 70
- penicillin, 139
- Pennsylvania, 12*i*
- personalization, perils of, 30–31
- persuasion and influence
- observation and, 255–257
 - persuadable individuals, identifying, 262–266, 274, 286–288
 - predicting, 255–256, 255–276
 - scientifically proving persuasion, 259–260
 - testing in business, 262
 - uplift modeling for, 266–267
 - voter persuasion modeling, 285, 299
- Persuasion Effect, The, 276, 295
- persuasion modeling. *See* uplift modeling
- Petrified Forest National Park, Arizona, 259–260
- Pfizer, 10*i*
- Philadelphia (PA) Police Department, 77
- photographs
- caption quality and likability, 129
 - growth of in data glut, 112
- Piotte, Martin, 185–189, 301
- Pitney Bowes, 195n, 273
- Pittsburgh Science of Learning, 17*i*
- Pole, Andrew, 48–49, 301
- police departments. *See* crime prediction for law enforcement
- politics, PA for, 15*i*–17*i*. *See also* electoral politics
- Porter, Daniel, 287, 289, 302
- Portrait Software, 195n
- Post hoc, ergo propter hoc*, 130
- Power of Habit: Why We Do What We Do in Life and Business* (Duhigg), 52
- Pragmatic Theory team, 188–189, 196–197
- prediction
- benefits of, 3, 28
 - choosing what to predict, 6–12, 55–56, 249, 254–256
 - collective obsession with, xxi
 - future predictions, 291–293
 - good *versus* bad, 83–86
 - limits of, 12–14
 - organizational learning and, 17–18

- prediction, effects of and on
about, 12–14, 295
- Data Effect, The, 115, 135–145, 295
- Ensemble Effect, The, 205, 275n, 295
- Induction Effect, The, 179, 295
- Persuasion Effect, The, 276, 295
- Prediction Effect, The, xxx–xxxi, 1–21,
26–27, 295
- prediction markets, 199
- predictive analytics. *See PA (predictive analytics)*
- Predictive Analytics* (Siegel), website of, 303
- Predictive Analytics and Data Mining*
(Kotu, Deshpande), 304
- Predictive Analytics Applied (training program), 304
- Predictive Analytics for Dummies* (Bari, Chaouchi, Jung), 304
- Predictive Analytics Guide, xxvii, 303
- Predictive Analytics Times*, xxvii, 303, 304,
305, 306
- Predictive Analytics World (PAW), xxvii,
305
- Predictive Analytics World (training programs), 304
- Predictive Analytics World (PAW)
conferences, xxii, xxvii, 48–49, 61,
70–71, 115, 197–199, 227, 240n,
305
- Predictive Marketing and Analytics* (Strickland),
304
- predictive models
about, 23–24
action and decision making, 36–38
causality and, 35, 131–135, 158,
257
defined, 34, 154–155
deployment phase, 32
Elder’s success in, 41–45
going live, 24–26, 30–31
machine learning and building, 36, 41
marketing models, 257–267, 270–271,
273–276
- observation and, 33–36
overlearning and assuming, 167–169
- personalization and, 30–31
- response modeling, 255–256, 268–270,
270–271
- response uplift modeling, 270–271, 275,
277, 299
- risks in, 38–41
- univariate *versus* multivariate, 154–156,
156–157
- uplift modeling, 266–267
- See also ensemble models*
- [PredictiveNotes.com](#), xvi, xxiii, xxvi, 95,
103, 147, 191, 249, 306
- predictive technology, 3–4. *See also machine learning*
- predictor variables, 116
- pregnancy and birth, predicting
customer pregnancy and buying behavior, 7, 48–52, 296–297
- premature births, 10, 10*i*
- prejudice, risk of, 81–82
- PREMIER Bankcard, 4*i*, 7*i*
- prescriptive analytics, xviii, 267n
- privacy, 47–48, 56–57, 83–86, 186
Google policies on, 84
- insight *versus* intrusion regarding,
60–61
- predicted consumer data and, 47–52
- probability, The Data Effect and,
137–139
- profiling customers, 156n
- Progressive Insurance, 70
- Psych* (TV show), xxx
- psychology
predictive analysis in, 18*i*
schizophrenia, predicting, 10, 18*i*
- psychopathy, predicting, 18*i*, 191
- purchases, predicting, 4*i*, 48–49, 121
- p-value, 136n
- Q**
- Quadstone, 195n, 273, 280

R

- Radcliffe, Nicholas, 267
 Radica Games, 19*i*
 Ralph's, 7
 random forests, 201
 Rebellion Research, 8*i*
 recency, 116
 recidivism prediction for law enforcement, 11–12, 12*i*, 77–78, 298
 recommendation systems, 7, 57, 185–187, 223
 Reed Elsevier, 5*i*, 17*i*
 reliability modeling, 13*i*
 REO Speedwagon (band), 58
 response modeling
 drawbacks of, 268–270
 examples of, 4*i*
 targeted marketing with, 255–256, 270–271, 277, 299
 response rates, 268, 276–277
 response uplift modeling, 255–256, 270–271, 275, 277
 restaurant health code violations, predicting, 16*i*
 retail websites, behavior on, 118–119
 retirement, health and, 122
 Richmond (VA) Police Department, 12, 12*i*, 67, 78
 RightShip, 14*i*
 Rio Salado Community College, 16*i*
 risk management, 28, 137–139, 149–151, 152
Riskprediction.org.uk, 9*i*
 risk scores, 149–151
Risky Business (film), 152
 Romney, Mitt, 285
 Rousseff, Dilma, 98
 Royal Astronomy Society, 189
 R software, 190n
 Rudder, Christian, 306
 Russell, Bertrand, 135
 Rutter, Brad, 241, 247

S

- safety and efficiency, PA for, 13*i*–14*i*
 Safeway, 7
 sales leads, predicting, 7*i*
 Salford Systems, 178
 “Sameer Chopra’s Hotlist of Training Resources for Predictive Analytics” (*Predictive Analytics Times*), 304
 Sanders, Bernie, 289
 Santa Cruz (CA) Police Department, 12*i*, 67
 sarcasm, in review, 18*i*
 Sartre, Jean-Paul, 116
 SAS, 195n
 satellites, predicting fault in, 13*i*
 satisficing, 269n
 Schamberg, Lisa, 108
 Scheer, Robert, 96
 schizophrenia, predicting, 10, 18*i*
 Schlitz, Don, 238
 Schmidt, Eric, 83, 84
 Schwartz, Ari, 86
Science magazine, 57
 scientific discovery, automating, 139–141
Seattle Times, 104
 security levels, predicting, 12*i*
 self-driving cars, 23, 79
 Selfridge, Oliver, 161
 Semisonic (band), 41
 sepsis, predicting, 9*i*
Sesame Street, 109
 Sesenbrenner, James, 96
 Sessions, Roger, 175
 Shakespeare, William, 169, 209
 Shaw, George Bernard, 156
 Shearer, Colin, 110
 Shell, 13*i*, 20*i*
 shopping habits, predicting, 4*i*, 48–49, 121
 sickness, predicting, 9*i*, 10–11
 Siegel, Eric, 300, 303, 306, 311
 silence, concept of, 20n
 Silver, Nate, 27, 140, 238, 283
Simpsons, The (TV show), 247, 254
 Siri, 202, 215

- Sisters of Mercy Health Systems, 9*i*
- smoking and smokers
- health problems and causation for, 135
 - motion disorders and, 123, 131–132
 - social effect and quitting, 9, 123
- Snowden, Edward, 87, 98
- Sobel, David, 55
- social effect, 10–11, 120, 123
- social media networks
- data glut on, 112
 - happiness as contagious on healthcare, 124
- LinkedIn, 3*i*, 7, 20*i*
- PA for, 3*i*
- spam filtering on, 74
- Twitter, 69, 112
- YouTube, 112
- See also* Facebook
- sociology, uplift modeling applications for, 277
- SpaceShipOne, 187
- spam, predicting, 20*i*
- spam filtering, 74, 297
- Spider-Man* (film), xviii, 83
- sporting events, crime rates and, 125
- sports cars, 119
- Spotify, 21*i*
- Sprint, 4*i*
- SPSS, 195n
- staff behavior. *See* employees and staff
- Standard & Poor's (S&P) 500, 39
- Stanford University, 9*i*, 10, 153n, 178
- staplers, hiring behavior and, 118
- Star Trek* (TV shows and films), 41, 77, 196, 209, 220
- statistics, 5, 16, 78, 112, 125, 135–145, 136–139, 167, 186
- StatSoft, 178
- stealing, predicting, 11–12
- Steinberg, Dan, 3, 147–148, 149, 153, 175–176, 178–181, 301
- stock market predictions
- black-box trading systems, 8*i*, 24, 38–44
 - Standard & Poor's (S&P) 500, 39
- Stone, Charles, 178
- Stop & Shop, 6
- street crime, predicting, 11*i*
- Strickland, Jeffrey, 304
- student dropout risks, predicting, 9, 16*i*
- student performance, predicting, 16*i*, 191
- suicide bombers, life insurance and, 125
- Sun Microsystems, 5*i*
- Super Crunchers* (Ayres), 306
- SuperFreakonomics* (Levitt and Dubner), 72, 81, 91n, 92, 95
- supermarket visits, predicting, 4*i*, 191
- surgical site infections, predicting, 9*i*
- Surowiecki, James, 197, 199
- Sweden, 124
- system failures, 13*i*
- Szarkowski, John, 112
- T**
- Taleb, Nassim Nicholas, xviii, 136–137, 138, 168
- Talking Heads (band), 52
- Target
- baby registry at, 49–50
 - couponing predictively, 6
 - customer pregnancy predictions, 3*i*, 7, 48–52, 296–297
 - privacy concerns in PA, 52, 85
 - product choices and personalized recommendations, 6*i*
 - purchases and target marketing predictions, 4*i*
- targeted marketing with response uplift modeling, 255–256, 270–271, 277, 299
- targeting direct marketing, 26–27, 296
- target shuffling, 143n
- taxonomies, 161
- tax refunds, 11*i*, 12, 71, 204
- Taylor, James, 305
- TCP/IP, 73
- Telenor (Norway), 5*i*, 252–254, 281
- Teragram, 195n
- terrorism, predicting, 11*i*, 11–12, 81–82

- Tesco (U.K.), 5*i*, 6
 test data, 173–175
 test preparation, predicting, 17*i*
 text analytics, 189, 195n, 209
 Text Analytics World, 305
 textbooks, 304
 text data, 209, 214
 text mining. *See* text analytics
They Know Everything about You (Scheer), 96
 thought and understanding, PA for, 8,
 18*i*–19*i*
 thoughtcrimes, 85
 Tibshirani, Robert, 304
Titanic (ship), 130
 T. J. Watson Research Center, 224
 tobacco. *See* smoking and smokers
 Tolstoy, Leo, 122
 traffic, predicting, 14*i*, 191
 training data, 49, 61, 148–149, 162–163,
 170, 276. *See also* learning
 training programs, 304
 train tracks, predicting fault in, 13*i*
 Trammps (band), 130
 travel websites, behavior on, 118–119
 Trebek, Alex, 207, 209, 224, 230, 245
 TREC QA (Text Retrieval Conference—
 Question Answering), 222
 truck driver fatigue, predicting, 18*i*
 true lift modeling. *See* uplift modeling
 true response modeling. *See* uplift modeling
 TTX, 13*i*
 Tucker, Patrick, 306
 Tumblr, 112
 Turing, Alan and the Turing test, 74, 77,
 222
 Twenty Questions game, decision trees and,
 162
Twilight Zone (TV show), 262
 Twitter
 2001: A Space Odyssey (film), 209
 data glut on, 112
 fake accounts on, 69
 mood prediction research via, 107n
 person-to-person interactions saved by,
 106
 2degrees (New Zealand), 5*i*
 typing, credit risk and, 121–122
- U**
- Uber, 2*i*, 118
 uncertainty principle, 262
 understanding and thought, predictive
 analysis in, 18*i*–19*i*
 univariate models, 154–156, 156–159
 University of Alabama, 9, 16*i*
 University of Buffalo, 12, 18*i*
 University of California, Berkeley, 19*i*, 204
 University of Colorado, 125
 University of Helsinki, 124
 University of Iowa Hospitals & Clinics, 9*i*
 University of Massachusetts, 227
 University of Melbourne, 16*i*, 191
 University of New Mexico, 122
 University of Phoenix, 16*i*
 University of Pittsburgh Medical Center,
 10, 10*i*
 University of Southern California, 227
 University of Texas, 227
 University of the District of Columbia, 101
 University of Utah, 10, 10*i*, 15*i*
 University of Zurich, 122
 uplift modeling
 customer retention with churn uplift
 modeling, 252–254, 276–279, 281,
 299
 downlift in, 273
 influence across industries, 276–279
 mechanics and workings of, 266–267,
 270–275
 Obama for America Campaign and, 277,
 286–288
 The Persuasion Effect, and, 276
 response uplift modeling, 270–271, 275
 targeted marketing with response uplift
 modeling, 255–256, 277
 Telenor using, 252, 254, 281

- U.S. Bank using, 6, 7*i*, 119–120, 266–276
uplift trees, 275
UPS, 14*i*
U.S. Armed Forces, 12*i*
U.S. Bank, 5*i*, 6, 119–120, 266–276, 273n, 274n, 275
U.S. Department of Defense, 92
U.S. Department of Defense Finance and Accounting Service, 11*i*, 71, 204
U.S. Food and Drug Administration (FDA), 279
U.S. government. *See* government
U.S. National Institute of Justice, 67
U.S. National Security Agency, 113
U.S. Naval Special Warfare Command, 20*i*
U.S. Postal Service, 11*i*, 16*i*, 71
U.S. Social Security Administration, 16*i*
U.S. Special Forces, 20*i*, 204
U.S. Supreme Court, 160–161
Utah Data Center, 113
- V**
- variables. *See* predictor variables
“vast search,” 140
Vermont Country Store, 4*i*, 6
Vineland (NJ) Police Department, 12*i*, 67
viral tweets/posts, predicting, 20*i*
Virginia, crime prediction in, 12, 12*i*, 67, 78
Volinsky, Chris, 188
voter persuasion, predicting, 15*i*, 160, 282–288
- W**
- Wagner, Daniel, 287
Walgreens, 57
Wall Street Journal, The, 134
Walmart, 118
Wanamaker, John, 26
warranty claim fraud, predicting, 11, 11*i*
washing machines, fault detection in, 13*i*
Watson, Thomas J., 224, 227n
Watson computer *Jeopardy!* challenge about, 18*i*, 24, 207–209
artificial intelligence (AI) and, 217–219
candidate answer evidence routines, 232
confidence, estimation of, 238–240
Craig’s question predictions for, 17*i*, 225
creation and programming of, 207–215, 225, 226–227
ensemble models and evidence, 234–235
Jeopardy! questions as data for, 207–209, 222
language processing and machine learning, 236–237
language processing for answering questions, 18*i*, 204, 210, 219–234
moneyballing *Jeopardy!*, 224–226
natural language processing (NLP) and, 210, 219–221, 227–231
open question answering, 213–222, 226, 238, 244
playing and winning, 8, 18*i*, 24, 204, 241–247
praise and success of, 247–248
predictive models and predicting answers, 226–227, 231–234
predictive models for predicting answers, 17*i*, 18*i*
predictive models for question answering, 17*i*, 18*i*, 219–221, 226–227, 231–234
Siri versus Watson, 215–216
speed in answering for, 241
Web browsing, behavior and, 120
Webster, Eric, 152
Wells Fargo, 20*i*
Whiting, Rick, 291
Who Wants to Be a Millionaire? (TV show), 199
“wider” data, 140–141
Wiener, Norbert, xvii
Wikipedia, 20*i*
editor attrition predicting, 9, 20*i*, 191
entries as data, 214, 227
noncompetitive crowdsourcing in, 190n

- Wilde, Oscar, 156
Wilson, Earl, 12
Windows versus Mac users,
 118
Winn-Dixie, 6
Wired magazine, 191
Wisdom of Crowds, *The* (Surowiecki), 197,
 199
WolframAlpha, 216
WordPress, 112
workplace injuries, predicting, 7*i*
Wright, Andy, 159
Wright brothers, 32
Wriston, Walter, 54
- X**
- X Prize, 187
- Y**
- Yahoo!, 119
Yahoo! Labs, 19*i*
*Yes! 50 Scientifically Proven Ways to Be
Persuasive* (Cialdini et al.), 260
yoga, mood and, 124
YouTube, 112
- Z**
- Zeng, Xiao-Jun, 107n
Zhou, Jay, 69