REVISED AND UPDATED

PREDICTIVE ANALYTICS

AN INTRODUCTION FOR EVERYONE

"Mesmerizing & fascinating..."
—*The Seattle Post-Intelligencer*

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

ERIC SIEGEL

WILEY

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, Macys.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

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

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

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

10 9 8 7 6 5 4 3 2 1

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

Contents

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?

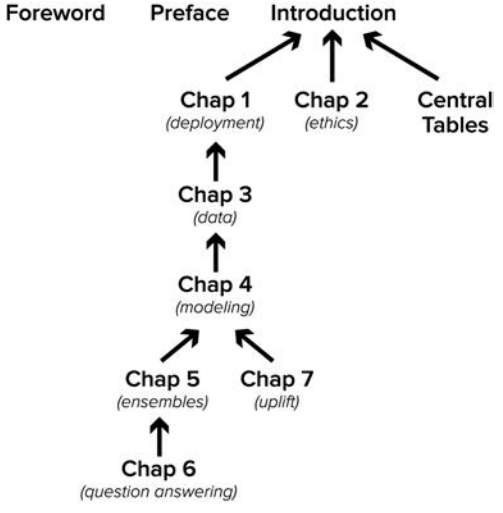
This book has served as a textbook at more than 30 colleges and universities. A former computer science professor, I wrote this introduction to be conceptually complete. In the table of contents, the words in parentheses beside each chapter's "catchy" title reveal an outline that covers the fundamentals: *(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.predictive analyticstimes.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 Scorpios 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.

prediction
data
predictive analytics
computer
like
better
make
machine learning
way
work
less
marketing
companies
insurance
likely
science
predictive
technology
business
within
accuracy
may
risk
knee
decisions
things
People
credit
company
well
across
customer
organization
organizations
just
best
power
next
first
possible
future
take
book
Researchers
computers
public
never
Learning
even
life
hand
card
million
predicting
order
human
name
others
day
new
learn
times
individual
bad
think
every
called
based
mail
drive
experience
mortgage
system
something
makes
often
developed
Effects
value
turn
learns
researchers
respond
government
surgery
process
general
time
learning
team

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.

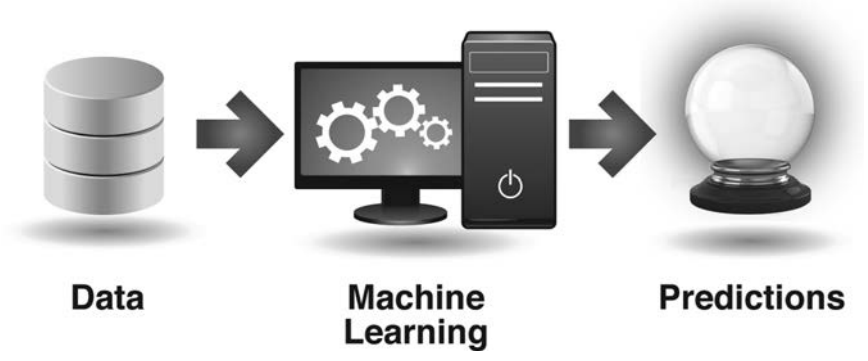
—William Edwards Deming (a business professor famous for work in manufacturing)

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

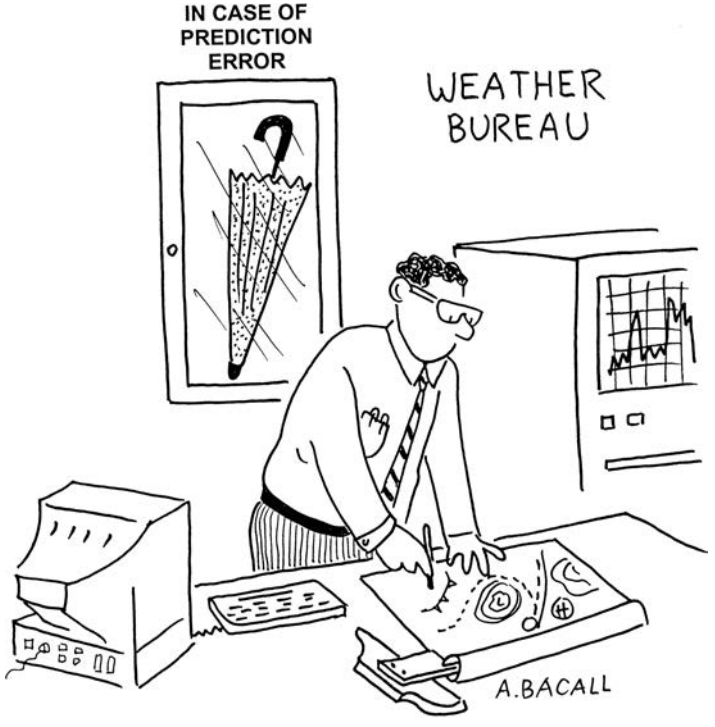
How come you never see a headline like "Psychic Wins Lottery"?

—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.

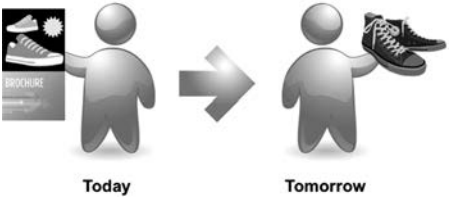
But are these claims too bold? As the Danish physicist Niels Bohr put it, "Prediction is very difficult, especially if it's about the future." After all, isn't prediction basically impossible? The future is unknown, and uncertainty is the only thing about which we're certain.



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 nonresponders 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

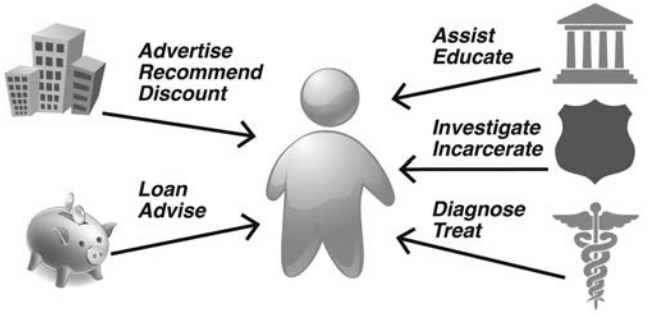
> *People . . . operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage.*
>
> —Michael Lewis, *Moneyball: The Art of Winning an Unfair Game*

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.

system
data
deployment
launch
model
individual
predictive
ad
John Elder
organization
learning
million
stock
blackbox
new
like
trading
predictive model
predicted
machine
likely
market
take
click
technology
business
much
investor
predict
ads
time
project
way
small risk
people
Prediction
client
since
user
go
go
yet
list
good
across
future
place
certain
thing
customers
company
marketing
decisions
nothing
money
many
now
profit
well
world
two
Dr
predicting
rules
generated
respond
meant
science
live
every
know
within
shown
Delta
another
action
score
value
kind
retrospective
long
buy
life
means
high
average
website
make
computer
given
takes
case
mail
little
years
even
whether
also
done
organizations
makes
school
use
based
half
overall
last
act
real
decision
win
first
often
predictive analytics
users
millions
look
part
Even
side
day
right
One
may
put
clicks
known
investors
something
scientist
entire
order
method
best
firm
man
application
back
learn
goes
see
online
going
great
year
revenue

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

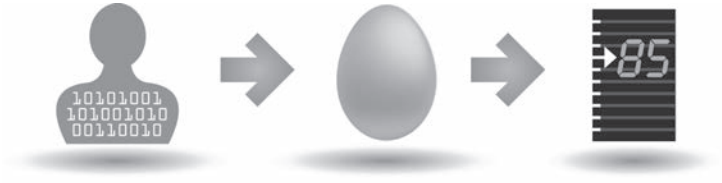
(Hotmail, Yahoo, Gmail, etc.); the ratio of the individual's SAT written-to-math scores (is the user more a verbal person or more a math person?), and on and on.

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:



Characteristics of an Individual **Predictive Model** **Predictive Score**

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?

About the Author

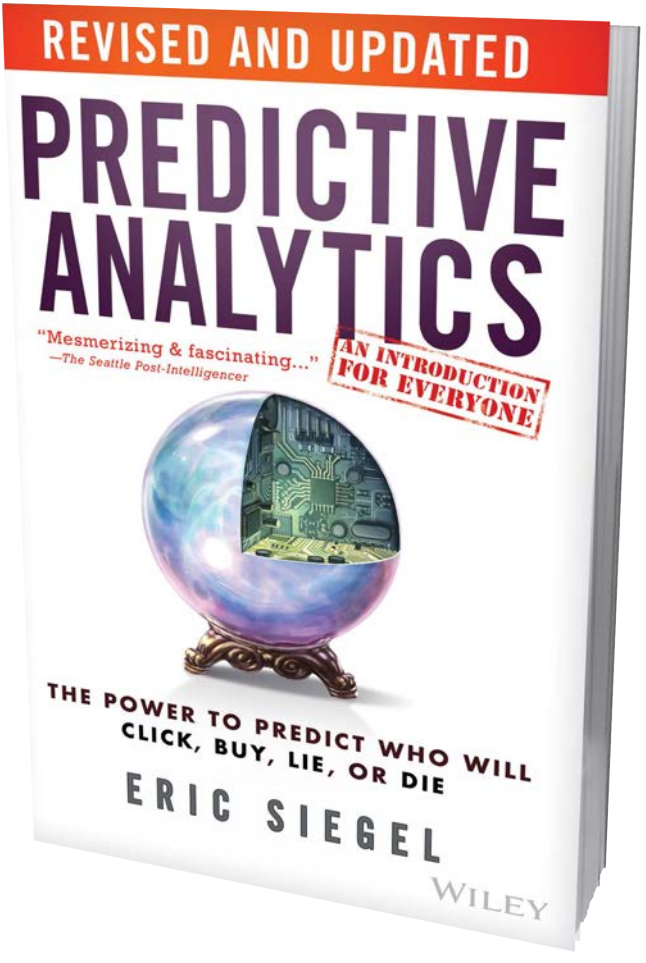
Eric Siegel, PhD, founder of the Predictive Analytics World conference series and executive editor of *The Predictive Analytics Times*, makes the how and why of predictive analytics understandable and captivating. Eric is a former Columbia University professor—who used to sing educational songs to his 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



amazon

BARNES & NOBLE
BOOKSELLERS

BAM!
BOOKS-A-MILLION

TRANSLATED INTO 9 LANGUAGES
USED IN COURSES AT MORE THAN 30 UNIVERSITIES

In this rich, fascinating—and surprisingly accessible—introduction, leading expert Eric Siegel reveals how predictive analytics works, and how it affects everyone every day.

Trendsetters like Chase, Facebook, Google, Hillary for America, HP, IBM, Match.com, Netflix, the NSA, Pfizer, Target, and Uber are seizing upon the power of big data to predict human behavior—including yours.

Why? Predictive analytics reinvents industries and runs the world. Read on to discover how it combats risk, boosts sales, fortifies healthcare, optimizes social networks, toughens crime fighting, and wins elections.

"What Nate Silver did for poker and politics, this does for everything else."

—David Leinweber, author of *Nerds on Wall Street*

"The *Freakonomics* of big data."

—Stein Kretsinger, founding executive, Advertising.com

"A deeply informative dive into a topic that is critical to virtually every sector of business today."

—Geoffrey Moore, author of *Crossing the Chasm*

"*Moneyball* for business, government, and healthcare."

—Jim Sterne, founder, eMetrics Summit

Photo Credit: Dana Patrick

ERIC SIEGEL, PhD, is the founder of Predictive Analytics World and executive editor of *The Predictive Analytics Times*. A former Columbia University professor, he is a renowned speaker, educator, and leader in the field.

Learn more: www.ThePredictionBook.com

BUSINESS & ECONOMICS/
Popular Culture

Cover Design: Wiley
Cover Image: Winona Nelson

Subscribe to our free Business eNewsletter at wiley.com/enewsletters

Visit wiley.com/business

WILEY

Also available as an e-book

ISBN 978-1-119-14567-7

Online Paper Review Analysis

Article *in* International Journal of Advanced Computer Science and Applications · September 2015

DOI: 10.14569/IJACSA.2015.060930

3 authors:

Doaa Mohey El-Din
Cairo University
28 PUBLICATIONS **678** CITATIONS
SEE PROFILE

Hoda Mokhtar
Cairo University
84 PUBLICATIONS **666** CITATIONS
SEE PROFILE

Osama Ismael
Cairo University
14 PUBLICATIONS **97** CITATIONS
SEE PROFILE

Online Paper Review Analysis

Doaa Mohey El-Din
Information Systems Department
Faculty of Computers and Information, CU
Cairo, Egypt

Hoda M.O. Mokhtar
Information Systems Department
Faculty of Computers and Information, CU
Cairo, Egypt

Osama Ismael
Information Systems Department
Faculty of Computers and Information, CU
Cairo, Egypt

***Abstract*—Sentiment analysis or opinion mining is used to automate the detection of subjective information such as opinions, attitudes, emotions, and feelings. Hundreds of thousands care about scientific research and take a long time to select suitable papers for their research. Online reviews on papers are the essential source to help them. The reviews save reading time and save papers cost. This paper proposes a new technique to analyze online reviews. It is called sentiment analysis of online papers (SAOOP). SAOOP is a new technique used for enhancing bag-of-words model, improving the accuracy and performance. SAOOP is useful in increasing the understanding rate of review's sentences through higher language coverage cases. SAOOP introduces solutions for some sentiment analysis challenges and uses them to achieve higher accuracy. This paper also presents a measure of topic domain attributes, which provides a ranking of total judging on each text review for assessing and comparing results across different sentiment techniques for a given text review. Finally, showing the efficiency of the proposed approach by comparing the proposed technique with two sentiment analysis techniques. The comparison terms are based on measuring accuracy, performance and understanding rate of sentences.**

Keywords—Sentiment analysis; Opinion Mining; Reviews; Text analysis; Bag of words; sentiment analysis challenges

I. INTRODUCTION

World Wide Web (www) has become the most popular communication platforms to the public reviews, opinions, comments and sentiments about products, places, scientific books or papers and to daily text reviews. The number of active user bases and the size of their reviews created daily on online websites are massive. There are 2.4 billion active online users, who write and read online and Internet usage around the world [1]. Scientific research domain has a big world in journals and conferences, there are more than 4000 rated conferences and 5000 ranked journals [2]. Each one of them has thousand number of papers such as ACM, Springer and Science direct. Notably, a large fragment of WWW researchers makes their content public, allowing researchers, societies, universities, corporations to use and analyze data. According to a new survey conducted by Dimensional Research, April 2013: 90% of customer's decisions depends on Online Reviews [3]. According to 2013 Study [4]: 79% of customer's confidence is based on online personal recommendation reviews. As the result, a large number of studies and research have monitored the trending new research increasing year by year. In this work, trying to achieve trusted scientific reviews evaluation to be useful for researchers and facilitate the selection of the suitable papers.

Recently, several websites encourage researchers to express and exchange their views, suggestions and opinions related to scientific papers. Sentiment analysis [5] depends on two issues sentiment polarity and sentiment score. Sentiment polarity [6] is a binary value either positive or negative. On the other hand, sentiment score which relies on one of three models [7]. Those models are Bag-of-words model (BOW) [8], part of speech (POS) [9], and semantic relationships [10]. BOW [11] is the most popular for researchers and based on the representation of word but BOW neglects language grammar. POS [12] which is grammatically tagging especially verbs, adjectives and adverbs [13]. For example, (The research is not good.) declaring in (The/DT research/NN is/VBZ not/RB good/JJ. /.). In the example DT refers to "Determiner", NN refers to "Noun", singular or mass, VBZ refers to "Verb", RB refers to "Adverb", and JJ refers to "Adjective". But a semantic relationship method is the most complex method, which is based on the relationship between concepts or meanings for example antonym, synonym, homonym etc.

There is a research gap the sentiment analysis accuracy because of sentiment evaluation drawbacks and sentiment analysis challenges [14].The evaluation sentiment drawbacks that Reflected in language coverage. This paper focuses on understanding text reviews and introduces solutions for some sentiment challenges. The sentiment analysis challenges summarized in ten challenges [15]. They are spam and fake reviews Detection, Limitation of classification filtering, Asymmetry in availability of opinion mining software Incorporation of opinion with implicit and behavior data, Incorporation of opinion with implicit and behavior data, and Natural language processing overheads (ambiguity), Generation of highly content lexicon database, handling of bipolar sentiments, dealing with short Sentence like abbreviations, Requirement of World Knowledge, Negation. All challenges have a bad effect on the understanding of reviews.

In this paper, the research aims to fill this research gap by proposing the new technique for sentiment analysis of online scientific papers reviews (SAOOP). The technique also measures efficiency by making a comparison between SAOOP, and other two sentiment analysis techniques [16]. Namely "Natural Language Toolkit-Text processing" (NLTK) and "recursive deep models for semantic" (NLPS). The results depend on comparing accuracy, performance and rate of coverage of language through two datasets.

The rest of this paper is organized as follows: Section 2 represents related works. Section 3, the presentation of the

new technique "SAOOP". In Section 4, outlines of the Experiment as well as the sample used for comparison. Section 5 highlights the comparison results. Finally, Section 6 concludes and proposes directions for future work.

II. Related Works

The purpose of this paper is sentiment evaluation which means to find the sentiment polarity (positive, negative, or neutral) of a text reviews data and evaluate the sentiment score of the text review. Generally a text review is divided into single sentences ("sentence-based") and words ("words-based") or very short texts from a single source.

A. Sentiment Analysis: An Overview

The author in (Sentiment analysis of document based on annotation) presented a tool which judges the quality of text based on annotations on scientific papers [17]. The authors's methodology declares in collective's sentiment of annotations in two approaches. This methodology counts all the annotation produces the documents and calculates total sentiment scores. The problem of this methodology appears in a relationship between annotations that is complex. The technique needs to have a big query knowledge base containing metadata. The notion declares in that the values are not accurate enough such as the value of "Good=0.875" has greater value than the value of "Best=0.75" although the result is wrong in logical meaning. Nevertheless, believing that collecting metadata and evaluating them could be useful to achieve higher analysis quality.

The researchers proposed a "Web Based Opinion Mining system" for hotel reviews [18]. They introduced an evaluation system for online user's reviews and comments to support quality controls into hotel management. The research is capable of detecting and retrieving reviews on the web and deals with German reviews. The multi-topic/multi-polarity is the method of this research; the system would recognize the neutral e.g., "don't know" to "classify sentiment polarity that as neutral" and the multi-topic cases identified in their corpus. The major weakness illustrates in not handling some cases in multi-topic segments. The authors [19] analyzed sentiments reviews of mobile devices products. Their Machine learning (ML) [20] system investigates the classification accuracy of Naïve Bayes algorithm. In addition to Judge the product quality and status in the market is advantageous. They use three machine learning algorithms (Naïve Base Classifier, K-nearest neighbor [21], and random forest [22] to calculate the sentiments accuracy. The random forest improves the performance of the classifier.

B. Sentiment Analysis Techniques

This section provides a brief description of the two sentiment analysis techniques investigated in this paper. These techniques are the most popular in the literature and they cover diverse techniques such as the use of Natural Language Processing (NLP) [23] in assigning polarity and sentiment score.

1) Natural Language Toolkit: The authors aim at an evaluation sentiment scores and polarity. They produce the Natural Language Toolkit (NLTK) [12]. NLTK is a text analysis technique that evaluates cognitive and constitutional components of a given text reviews based on using lexicon including words. They use hierarchal sentiment classification level with two levels (Neutral, Positive, and Negative). The drawback of this technique illustrated in low accuracy and some logical errors. Because the technique needs to increase handling of language coverage [24].

2) NLP Stanford sentiment (NLPS): The researchers introduce recursive neural models have in common: word vector representations and classification [25]. The authors released a tool named "NLP Stanford" NLPS [26], which develops an integration of learning techniques that produces better results and higher accuracy training model empirically. Their goal is based on Semantic word spaces have been very beneficial but NLPS cannot express the meaning of longer phrases in a primary way. So they improve this technique by detection the sentiment requires wider supervised training and evaluation resources.

III. Sentiment Analysis of Online Papers (SAOOP)

In this section, Sentiment analysis of online papers "SAOOP" will be presented. SAOOP is used in opinion mining [27] and based on a new English lexical dictionary [28]. This lexical dictionary groups adjectives, nouns, verbs, adverbs, adjectives, prefixes, suffixes and other grammatical classes into synonym. The proposed technique is an enhancement on Bag-Of-Words (BOW) model [29] in sentiment analysis to achieve high accuracy, which depends on word weight replacing term frequency of each word. The proposed technique solves two important Bag-of-words weaknesses.

The standard bag of words is not automatic in classification and creating polarity lexicons because BOW model needs to create manual lists of 'positive' and 'negative' words [30]. That means the review judgment is based on the probability of positive or negative words. The second is low accuracy because the standard BOW model neglects text grammatically. Sentiment classification levels will be divided into five classes (very positive, positive, negative, very negative and Neutral).

The proposed technique makes the sentiment classification levels are more detailed and easy by word percentage of each class. The goal of SAOOP is for inferring the polarity of common meaning and polarity concepts from natural language text at a word level, rather than at the syntactic level. SAOOP also classifies reviews into some categorizations based on papers parameters. In addition, the estimation rank of each paper based on evaluation some parameters.

A. SAOOP Overview

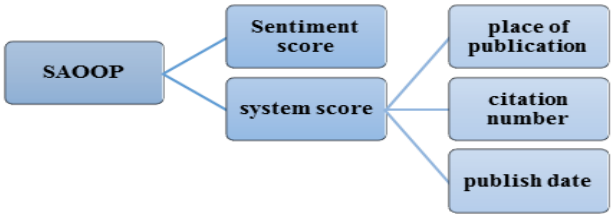


Fig. 1. SAOOP Overview

"Fig.1" shows that SAOOP model consists of two components sentiment score and system score. SAOOP can evaluate any paper based on the components. Sentiment score depends on total reviews evaluation score. And system score which depends on the sum of total scores of three parameters of paper (place of publication), citation number of paper and paper publishing date. SAOOP technique helps researchers to select the suitable paper with the total paper score.

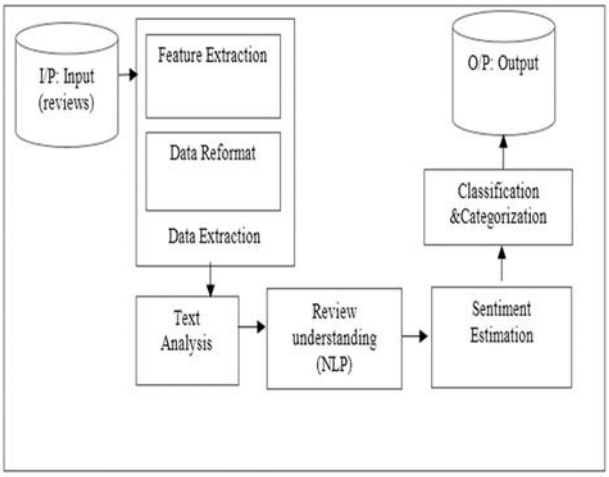


Fig. 2. SAOOP Technique overview

"Fig.2" declares SAOOP Technique overview. The input is scientific paper website link. In data extraction [31] level two parts: first, using Easy web extract tool which is web scraping tool to extract data of paper from scientific papers website online. Part two is data reformat from Excel sheet which is one output of EasWebExtract tool [32] suitable with SAOOP database format. In text analysis level, SAOOP applies some functions of text analysis on reviews of each paper. In the first, applying the splitting sentences function, tokenizing words function, and checking of stop list and removing them [33].

In review understanding (NLP) level, the proposed technique understands the sentences meaning and check words in vocabulary lexicon with similarity and differences algorithms. In estimation phase, showing the evaluation sentiment score for each word into text review and the polarity detection for each one and each sentence and calculate the total score of sentiment review score. In classification phase, that's splitting into two parts, first the reviews classification into five sentiment classification levels (very positive, positive, negative, very negative and objective (neutral), also having degrees of each sentiment level with scale from [-1, 0, 1]. There is also another classification which declares each review categorization based on five meaning classes (topic, date, author, citation, and place of publication). The benefit from the extracted data to memorize them and make relationships between evaluated papers and reviews and categorize reviews logically based on topic domain parameters. Output is the sentiment evaluation score of all reviews with all papers with caring of number of reviews parameter, and evaluation of scientific paper parameters score which is based on metadata of each paper (place of publication, publishing date, and number of citation). So the consequent result is ranking to each paper with the total score of sentiment and system scores with graphical reports of results.

B. SAOOP Methodology

SAOOP can assign polarity based on this approach, considering the words weight replacing term frequency, by assuming each word has two values and polarity with this assumption equation,

$$V(w) = \sum(W(p) + W(n)) = 1. \quad (1)$$

V (w) is value of word, W (p) refers to positive value and W (n) refers to negative word, the selection between positive or negative polarity Influenced by the meaning of words and each other polarity. But the sentence contains negative that differs in the word value. If the word is positive, convert to negative polarity and the negative score will be as in the equation,

$$V\ (w)\ = W(p) - 0.2. \quad (2)$$

And if the word is negative, the score will be calculated by V (w) =W(n) + 0.2. The selection of 0.2 because this disison is suitable for the five sentiment class's levels [18]. The proposed technique also creates papers ranks with calculating sentiment and measuring domain parameters. By assuming,

$$\mathrm{P(TS)} = \sum_{\mathrm{R=0}}^{\mathrm{n}} (\mathrm{T(SA)} + \mathrm{T(SS)}). \quad (3)$$

In the equation, P (TS) refers to a total score of each paper, T (SA): is a total score of sentiment score of all reviews on each paper with caring of number of positive reviews. In the next equation,

$$(SA) = \sum_{i=1}^{n} \frac{P(SA(R))}{n}. \quad (4)$$

The calculation of the total score of all reviews depends on the score of each review. There is a difficult problem between large number of reviews and evaluating sentiment polarity of each one, this problem is improper the most review number having assessment higher score. For example, one paper publishing in 2013 that's mean from 2 years and this paper has twenty reviews, not equal evaluation one paper publishing in 2005, that's mean from 10 years and the second paper has twenty reviews. The first one is the top rated because the evaluation number of reviews in short time. In other example, one paper publishing from 2 years and having twenty negative

reviews, not equal evaluation other one publishing from 10 years and having positive twenty reviews. The second one has maximum rated because the evaluation numbers of positive reviews is larger than the one, although the second is the oldest. As mentioned before double trouble with reviews number and the relationship between date and other relation between sentiment polarity of reviews and number of reviews. That interprets difficulty of evaluation domain parameters.

The proposed technique faces these challenges and evaluates the percentage of positive reviews over total scores. But still there is a problem in relationship between date and number of reviews, for example: one paper publishing from 2 years which has twenty positive reviews, not equal evaluation other one publishing from 10 years which has positive twenty reviews. Actually that is not equal their selves because the recent has bigger reviews number. So SAOOP presents a solution for date relation with reviews number, according with two parameters number of positive reviews and the recent paper. T (SS): is a total score of system score parameters that are evaluated logically of paper parameters according to this equation,

$$V(SS) = \sum(\frac{S(pp)}{\lambda} + \frac{S(C)}{2\lambda} + \frac{S(D)}{2\lambda}). \quad (5)$$

V (SS) expresses the value of systems score. S (PP) means the score of publication place, S(C) refers to the score of paper citation number, and S (D) means the score of paper publish date. Assuming λ is a constant equal 2, dividing into λ and 2λ to determine the priority of evaluation of the parameter. The evaluation topic parameters process does not ease because of depending on the logical meaning of each one. So the research focuses on scientific papers domain to put the foundation of evaluation parameters to achieve the fact value of each paper to support researcher with sentiment analysis by ranking papers based on total score of them. There is inverse relationship between publishing date and number of citation of the paper, which declare in this equation,

$$S(C) = \frac{current\ year - publish\ date\ year}{number\ of\ Citation}. \quad (6)$$

The result is not true the highest citation number having the highest evaluation score of it. For example, one paper publishing in 2013 that's mean from 2 years which has ten citations, not equal evaluation one paper publishing in 2005, that's mean from 10 years which has ten citations. The first one is the highest score because number of citations in shortly is high, this first paper will be predicated if the paper has the same time 10 years, it mostly has 50 reviews not 10 reviews such as the second paper. In other words, the first paper has 5 papers into each year but the second has 1 into each year. To evaluate score of publishing place conference which depends on ACM conferences tiers with a sample into computer science conferences, such as "VLDB: Very Large Data Bases is in the top tier: tier 1", "ER: Intl Conf. on Conceptual Modeling (Conf. on the entity Relationship Approach)" is in next tier which is in lower tier: Tier 2, and "IDEAS: Intl Database Engineering and Application Symposium" is in a lower tier: Tier 3" [34].

C. SAOOP & Sentiment Challenges

SAOOP enables to make solutions to most significance sentiment analysis challenges [35]. The proposed technique can produce some solutions for main challenges to reach to higher accuracy. The discussion of the solutions in the following:

1) Topic domain independence

Domain-dependent [36] is a difficult challenge to recognize topic nature. There are some words have many meanings and different sentiment values relevant to the topic. There is also a problem shows in extracting keyword or features and how to evaluate words based on each topic. One feature set may give very good performance in one domain, at the same time it performs very poor in some other domain. The produced solution suitable with a small scale by applying the proposed technique on one topic domain and examine domain parameters evaluation by categorization reviews because they also give different meaning with the same word. This research presents a technique to recognize topic nature automatically. The proposed technique is based on extracting keywords and relevant features of each topic. In addition, to produce a solution for some words have many meanings and different sentiment values relevant to the topic. The proposed technique is based on Classification review of each domain features and keywords.

For example, "IEEE is [great +] publication for your paper", SAOOP can put IEEE is in a place of publication classification (based on feature name of publication) and the polarity is positive. "The publishing conference is [great+]", this review refers to the place of publication classification (based on keywords) and the polarity refers to positive. In other example, "The paper publishing date is [old-]", this reviews refers date classification (based on date attribute) and "Old" having the negative score. "The author is [old$^-$] in this field", but SAOOP can categorize the last review in author classification that is meaning the author is expert in this field so "Old" will be had positive score.

SAOOP improves the sentiment score to be more accurate and fair. By assuming some words have 0 value because of depending on classifications of each sentence of each review, there are some groups of words having a polarity and score to relate with the detected classification.

2) Negation

Negation is the biggest challenge in sentiment analysis [37]. The new technique produces a solution to improve evaluation negative with the enhanced bag of words technique. This research handles the two techniques: explicitly and implicitly negative [38]. First: explicitly is deliberately formed and are easy to self-report and by keywords. Second implicitly [5] is the unconscious level, are involuntarily formed and are typically unknown to us without any keywords of negative. In addition, the handling the negative meaning of some conjunctions such as "not only", and "But". The dual negative is the most important case which cares to achieve the total sentiment polarity. Reverses polarity of mid-level terms: great V.S not great.

A method often followed in handling negation explicitly in sentences like:

"I do [not like$^+$] $^-$ the paper", is to detect the negative polarity because the word (not) and convert the sentence

operator to negative. But this does not work for "I do [not like^{+}] $^{-}$ this research but I [like^{+}] the field". But still there can be problems.

Other example, "I find the functionality of this new methodology [less $^{-}$] practical", this review refers to explicit comparative negative. "This algorithm is [not great^{+}]$^{-}$]", the proposed technique handles in this review the positive and negative evaluation which declares in [not great! = bad] but [not great = good]. Implicitly negative such as "This research is [very [complex $^{-}$] $^{-}$]" this example does not have any negative keyword, but the meaning has negative and the polarity will be negative of this sentence.

There are sentences having keywords of negation and they don't have the negative polarity such as "[Not only^{+}] I [like^{+}] this algorithm, but also [easy^{+}] to understand and apply." the polarity is not reversed after "not" due to the presence of "only". So this type of combinations of "not" with other words like "only" has to be kept in mind while designing the algorithm.

There is a difference between "not only" and not because not only strengths the meaning (more positive or less negative) based on the polarity of the sentence. In this example other case of implicit negative, I [wish $^{-}$] to work [harder $^{-}$]". In the last review, the new technique presents future words e.g. wish refers to the negative polarity but first must check the polarity of the next sentence polarity because maybe changed the polarity depends on meaning.

3) Creation lexicon

The proposed SAOOP yields an improvement over prior published bag of words built lexicons. This technique also provides an improvement in calculation technique used in reviews sentiment analysis. SAOOP technique presents a solution to take care of grammar (which is one of limitation of Bag-Of-Words) and to save time took is N-gram algorithm to create subsequences of terms. There are two phases that will be produced:

- Phase 1. Data Preparation Phase

Less number of words in vocabulary lexicon to fast search based on similarity and differences algorithms. SAOOP neglects verbs tenses or word formula (singular or plural), that's meaning neglecting English grammar and syntax because of the comparison and differentiation with the infinitive verbs, and singular words with most letters similarity.

- Phase 2. Lexicon Development Phase

Evaluation words /terms: is based on enhanced bag of words: the proposed technique doesnot depend on term frequency. This phase is based on assuming each word has two values and the total of them equal 1. Each term has 2 polarities (+/-).

Negative Algorithm

1. **For each review R in paper P sets**
2. **For each sentence Sent in Review R**
3. **Apply Pre-Processing: Remove the stop words**
4. **Convert all words to Upper case**
5. **Check on expressions have "No or Not" e.g Not Only**
6. **Check on first Negation keywords list, which effects on the polarity of the words e.g "Not" I don't like this paper.**
7. **By assuming, the negative value for positive word ,**
   $$V(w) = W(p) - 0.2$$
8. **And assume the positive value for the negative word,**
   $$V(w) = W(N) + 0.2$$
9. **Check on the next word after negative e.g "like" has positive value and polarity but here it will take negative polarity and value.**
10. **Detect the polarity and value.**
11. **Check on the second list of negation keywords, which effects on the polarity of sentence e.g. never, yet, neither.**
12. **Convert polarity of the sentence by multiplication with(-1) ,** $V(Sent) = S(Sent) * -1$
13. **Check on future words e.g "wish/hope".**
14. **Check on the next sentence polarity.**
15. **End for**
16. **Detect sentence value and polarity.**
17. **End For**
18. **calculate review value and polarity**

(Note: knowing our attention of review classification.)

4) World knowledge requirement

SAOOP technique produces a solution for Knowledge about worlds' facts, events, people are often required to correctly classify the text. Trying to achieve higher accuracy and get the evaluation for some neutral reviews. The World knowledge challenge solution is based on the hierarchical database of nouns. Semantic (hierarchal) relationships between nouns to achieve the polarity, score and meaning. Also to differ between them and keywords or features. Consider the following example, "the author is a [lion^{-}] in this field", the previous review present negative polarity because lion is a name of animal but in real evaluation in the review refers to a positive polarity. In the next review, "Bing is really [Einstein?]" evaluation sentiment analysis without world knowledge classifies above sentence as neutral, but this review is an objective sentence because Einstein is the name of the famous scientist, so it refers a positive polarity also. This review is very hard for software to understand that automatically. SAOOP creates a huge lexicon database to contain the world knowledge especially related to researchers and the most common in the reviews. The solution of world knowledge also assumes values of the words based on the most common meaning. The evaluation of these world knowledge depends on keywords and classification of reviews.

5) Spam and Fake Reviews:

The WWW contains both realistic and spam contents. For effective Sentiment classification, this spam content should be eliminated before processing.

SAOOP can be done by empty or identifying duplicates, by detecting outliers and by considering the reviewer reputation. The proposed Technique enhances reviews spam and fake. SAOOP technique can avoid and cure the most of them by:

- Remove empty reviews: To calculate the real number of reviews.
- Delete duplicate reviews by considering the same reviewer: To Calculate the real sentiment score of the paper.

For example, one paper has 10 reviews, 3 of them for the same text review and with the same user, and 2 is empty reviews , in most sentiment application, if having 10 reviews number and the same repeated reviews will calculate together, the sentiment score is not real because having fake reviews and the results became fake also. And also there are some reviews are general are not related to the paper actually. SAOOP can produce solution for the case study on citeulike.com website, through making quaternary relationship between a set of paper parameters "paper name", "author names", "review" and "Username" (who is a review writer) with taking into consideration review written time, if the review is repeated by the same review writer with ensuring if the review is fake by all parameters and time, the proposed technique will delete the spam review before calculating the sentiment analysis. SAOOP can also deal with fake reviews if it empty and deleted.

In this paper, showing the implementation of SAOOP technique using C# programming language working on Microsoft visual studio 2010 platform. The newly created lexicon is based on SQL Server Management Studio 2008.

IV. EXPERIMENT

In this section, the discussion of the comparison between the proposed technique and two sentiment analysis techniques. This comparison shows the accuracy and performance results based on two datasets. This comparison also compares with the effects and solutions of sentiment analysis challenges.

A. Datasets

The comparison uses two different datasets: 1) real data set: which splits into two data sets with training set (1000 text reviews) and test set (5000 text reviews), 2) verified data set: which is a real set with unknown evaluation around10.000 text reviews (including more than 5.000 positive words, 5.000 negative words).

1) Real dataset

The first sample set is a sample of WWW.citeulike.com papers reviews and Metadata posted by computer science papers branch [39, 34]. The comparison in real data set in computer science scope including two parts: training data and test data [40]. Training data is a set of data to evaluate sentiment around 1000 reviews, knowing the values before. The second part is a test data: which is a set of data to evaluate sentiment with hide class label around 5000 text reviews. Citeulike receives in excess of 200,000 distinct visits (defined by Google Analytics as a group of page views by a unique user with timeout after 30 minutes inactive) monthly, with each visit originating an average of 2.77 page views [41]. Of that 200,000 around distinct users who have previously visited the site on multiple occasions.

There are currently 505,402 items posted in the database (counting n people post the same article); 1,676,130 tags (counting n if there are 'n' tags applied to an article); and 130,548 distinct words used these numbers are growing exponentially. This sample set allows us to study the responses to noticeable past texts. In addition, to evaluate the improved levels of techniques, methodologies in sentiment analysis. SAOOP can handle ten cases to ease to understand text review accurately by CiteULike users they illustrated in table 1. SAOOP can care and evaluate of some English grammar to improve BOW model.

2) Verified dataset

The second dataset which is called verified data set is a real data set but they can't be known the evaluation before. The dataset has around 10.000 text reviews in this sample. This data set is splitting into two parts of verified data reviews as positive and negative. These datasets include a wide range of online papers texts reviews: general reviews. In Table 2, the sample reviews of online scientific papers. SAOOP technique can evaluate sentiment score with the relationship of reviews categorization. With applying on this human-verified sample set [29], by fitting to quantify the range with different sentiment analysis techniques can accurately evaluate polarity of text reviews.

TABLE I. TABLE ENGLISH LANAUGE COVERAGE HANDLING BY SAOOP

Cases	Definition & Examples
1. Expressions	Based on syntax (e.g., Not Only, No one.)
2.Topic objects[3]	Features (e.g., lot of contributions)
3.Negation	Implicit (e.g., Independent)) and explicit (e.g. not bad, does not very good)meaning
4.Suffixes &prefixes	The beginning or end letters of word to have different meaning (e.g., dislike, opposed to, useless)
5.Verbs	Converting verbs tenses into infinitive
6.Adjective& Adverb	e.g., Well, improved, highly
7.Nouns	e.g., algorithms, improving, enhancing
8.Comparative [4] comparisons	e.g., easier convert to→ easy. ("More"; "higher") and ("most", "highest").
9. Phrases	e.g., very good, the professional work
10. Some special (Need, Wish)	(e.g., hope, wish): in the most times, they have negative polarity.

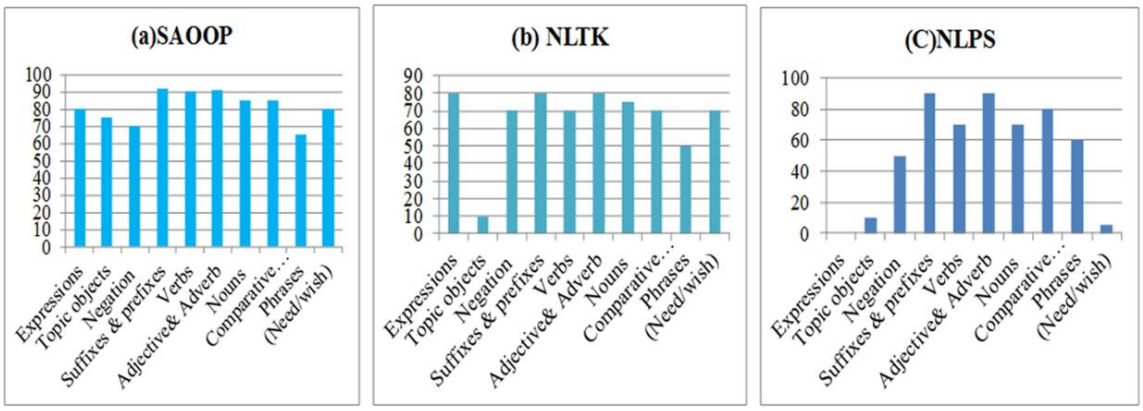


Fig. 3. Coverage Rate of ten cases understanding cases with the three techniques

TABLE II. SAMPLE OF REVIEWS

Sample of reviews
This paper is very well.
It's not great
The best web point
I am interesting in this field
Extremely good
It's not only hot research area but also having new scientific contributions.
This point research is more affected in web mining than using in neural network.
high accuracy
It's not have good value enough
Citation is valuable

B. Comparsion Measures

In order to define the evaluation of accuracy and performance of the three techniques, which will consider in the following table.3:

TABLE III. MEASUREMENT TABLE

		Actual observation Positive	Negative
Predicted expectation	Positive	x	y
	Negative	z	w

Let present True positive (x) was defined when a text was correctly classified as positive, False Positive (y) is a negative text which was classified as positive, False Negative (z) is a positive text but was classified as negative, and the last one True Negative (w) is a correctly classified as negative [42]. In order to compare and evaluate the techniques, by considering the following metrics, commonly used in information retrieval: true positive rate or recall: $R = x/(x + z)$, false positive rate or precision: $P = x/(x + y)$, accuracy: $A = (x + w)/(x + y + z + w)$, and F-measure (performance): $F = 2 \bullet (P \bullet R)/(P + R)$. In many cases simply use the F-measure, as it is a measure of a test's accuracy and relies on both the precision and recall [10]. By reporting, all the measurement mentioned above by practical interpretation. The true positive rate or recall can be understood as the rate at which positive reviews are predicted to be positive (R), whereas the true negative rate is the rate at which negative reviews are predicted to be negative.

The accuracy represents the rate at which the method predicts results correctly (A). The precision also called the positive predictive rate, calculates how close the measured values are to each other (P). The comparison also provides the F-measure results, since it is a standard way of summarizing precision and recall (F). Ideally, a polarity identification method reaches the maximum value of the F- measure, which is 1, meaning that its polarity classification is perfect. The y-axis is a percentage of the understanding sentence rate.

V. COMPARISON RESULTS

In order to facilitate understanding the advantages, disadvantages, and limitations of the various sentiment analysis techniques [43]. This section also presents the comparison results among them.

Understanding of word coverage: in the beginning, the comparison of the coverage of English grammar cases across the representative scientific reviews from CiteULike website. Then examination the intersection of the covered reviews cases across the techniques were in table 1. "Fig. 3 (a)" shows the result for the proposed technique SAOOP, which explain in section 4. "Fig. 3 (b)" declares the NLTK technique. NLTK which is a teaching tool works in, computational linguistics using Python [44]. And "Fig.3 (c)" shows NLP technique. NLPS technique which is predicting the sentiment of reviews based on a recursive model.

As shown in the figure, SAOOP has the highest understanding sentence coverage with 82.5 % with two data sets with three data sets samples, respectively, followed by NLPS which can't evaluate the total sentiment score but with detecting word by word polarity its percentage is 72%.

NLTK can interpret less than 10% of all relevant reviews. In addition, we compare with the percentage of handling sentiment analysis challenges to high accuracy and performance of sentiment analysis of the three techniques of the text reviews depicted in "Fig. 3".

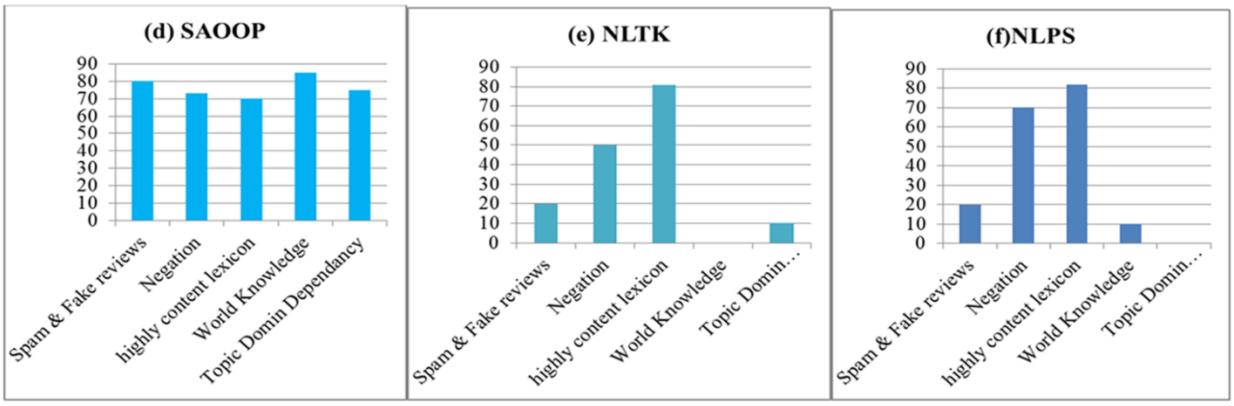


Fig. 4. Perecentage of handling sentiment analysis with the three techniques

According to "Fig.4 (d)" in fact, SAOOP had a new solution for some sentiment challenges but NLPS and NLTK, they and can't produce methodology to solve them expect some cases in negation but they have many logical errors, that shown by "Fig.4 (E) and (F)". The analysis results in table 3, refers to the: Percentage of accuracies between techniques based on different data set size. Also we examine the average result analysis of the two big data set that spirited into three data sets, that illustrate the highest average results with sentiment score of the proposed SAOOP technique then NLPS and the lowest one is a NLTK Technique. Finally, the summarization the results with the average of the three data sets (real and verified sets), we find the average of sentiment score of the proposed technique improve the results. Because of working binary analysis solutions of some important challenges and evaluate some technical cases in the text which have a problem in evaluation to be more accurate. In next section, we discuss the accuracy results of the comparison.

a) Accuracy: With the examination of the percentage degree of different techniques accuracy on text reviews content. In order to compute the accuracy of each technique, by calculating the intersections of the positive or negative proportion given by each technique. Table.4 presents the percentage of accuracy for the three compared techniques. For each technique in the first column, showing the estimation from the two data sets of reviews. Finding that some techniques have a high coefficient as in the case of SAOOP (82.5%), while others have least overlap such as NLTK (62%) and NLPS (70.2%).

The last "column" of the table shows on average to what extent each technique agrees with the other two samples. The last "row" quantifies how other methods agree with a certain technique, on average.

With the results of table 4, they illustrate differences between accuracy and performance of the three techniques. Table 4 shows techniques recall, precision, accuracy and performance.

"Fig.5" is shown the accuracy results of them. In a summary, the result indicates that existing tools vary widely in terms of accuracy about sentiment score, with scores ranging from 60% to 80%.

TABLE IV. AVERAGE RESULTS FOR ALL DATASETS

Metric	SAOOP	NLTK	NLPS
Recall	0.856	0.571	0.253
Precision	0.867	0.845	0.846
Accuracy	0.817	0.629	0.715
F-measure	0.846	0.665	0.729

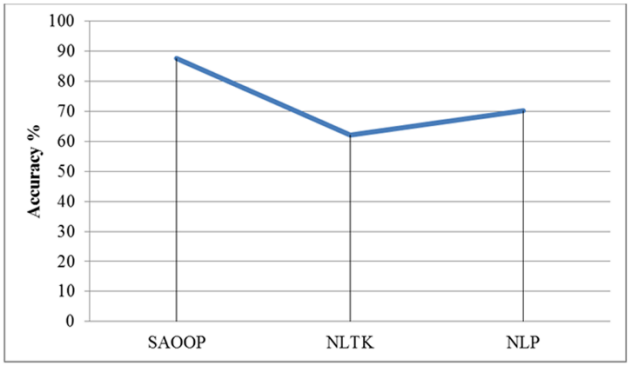


Fig. 5. Differences between Accuracies of three techniques

TABLE V. PRECENTAGE OF ACCURACY BETWEEN TECHNIQUES

Dataset	SAOOP	NLTK	NLPS	Average
Training set 1.000	83.5%	62%	72%	72%
Test set 5.000	81.99%	61%	70%	70.99%
Teal set 10.000	82.5%	60%	71.56%	71.186%
Average	82.5%	61.514%	71.604%	-

b) Perfromance: In this section, showing an evaluation of the performance of the three compared techniques. For comparing the performance results, Table.5 which gives the average of the results obtained for all datasets. For the F-measure, a score of 1 is ideal and 0 is the worst possible.The technique with the highest F-measure was faced sentiment analysis challenges and cover ten cases of each text review (0.846), which had the highest sentiment accurate and understanding text coverage. The second rated technique in the understanding of F-measure is NLPS, which obtained a much higher

coverage than understanding and challenges. It is important to note this problem that it can't be interpreted into of total score of the text review. For observation better performance on data sets that contain more expressed sentiment, such as text reviews (e.g., papers online) and the lowest performance is NLTK technique.

VI. CONCLUSION

Sentiment analysis is the most important source in decision making. Almost people becomes depends on it to achieve the efficient product. Thousands of researchers rapidly year by year that focuses on scientific online reviews for papers to help them. So the researchers introduce a new sentiment technique. In this paper, the researchers create a new technique is called sentiment analysis of online papers "SAOOP". The proposed technique will be a suitable and efficient solution to analyze online reviews. The target of technique to improve accuracy and achieve to accurate review meaning. The proposed SAOOP approach is based on two methods: evaluation and analysis reviews (sentiment analysis) and solve some sentiment analysis challenges. In order to serve researchers in selecting efficient papers. In addition, it evaluates topic domain parameters of scientific papers (place of publication, publishing date, and a number of citation paper) to evaluate the total score of papers. To evaluate SAOOP efficiency, making a comparison between it and two famous techniques. The results have a comparison between the accuracy and performance between the three techniques when the researchers apply the techniques on three data sets (training, test and verified). The comparison results illustrate how proposed technique can increase accuracy and performance with facing many language coverage cases and solving some sentiment analysis challenges. The accuracy results show in NLTK (62%) and NLPS (70%) to 82% (SAOOP) with the proposed technique.

REFERENCES

[1] Suman, D.R, Wenjun, Z., Social Multimedia Signals: A Signal Processing Approach to Social Network Phenonmena, ISBN-13: 978-3319091167, Springer International Publishing Switzerland, 2015.

[2] Larsen, Peder Olesen, and Markus von Ins. "The Rate of Growth in Scientific Publication and the Decline in Coverage Provided by Science Citation Index."Scientometrics 84.3 (2010): 575–603. PMC. Web. 25 Sept. 2015.

[3] Peng, L., Cui, G., Zhuang, M., & Li, C. (2014). What do seller manipulations of online product reviews mean to consumers? (HKIBS Working Paper Series 070-1314). Hong Kong: Hong Kong Institute of Business Studies, Lingnan University.

[4] Thomas, B., Keep Social Honest , "What Consumers Think about brands on social media, and what bunesses need to do about it" Report, 2013.

[5] Namita, M., Basant, A., Garvit, C., Nitin, B., & Prateek, P., "Sentiment analysis of Hindi Review based on Negation and Discourse Relation", International Joint Conference on Natural Language Processing, Nagoya, Japan, 2013.

[6] Theresa,W. ,Janyce, W. ,& Paul, H., "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", proceeding HLT'05 proceedings of the conference of Human Language Technology and Empirical Methods in Natural Language Processing, 2006.

[7] Ian, H.W., Eibe, F., & Mark A. H., Data Mining: Practical machine learning tools and techniques. The Morgan Kaufmann series in data management systems, 3rd Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[8] Yin, Z., Rong, J., & Zhi-Hua, Z., "Understanding Bag-of-Words Model: A Statistical Framework", International Journal of Machine Learning and Cybernetics, 2010.

[9] Diana, M., & John.C.,"Disambiguating Nouns, verbs, and adjectives using automatically acquired Selectional preferences", Association for Computational Linguistics , Vol. 29, 2003.

[10] Presutti, V., Stankovic, M., Cambria, E., Cantador, I., Di Iorio, A., Di Noia, T., Lange, C., Reforgiato Recupero, D., Tordai, A., "Semantic web evaluation challenge. 1st Edition, Series title: Communications in Computer and Information Science, Springer International Publishing, 2014.

[11] Bing, L., "Sentiment Analysis and Subjectivity". In Nitin, I. & Fred, J. (eds). Handbook of Natural Language Processing. 2nd Ed, Machine Learning & pattern recognition series, Chapman& Hall/CRC, 2010.

[12] Samih,Y., Erdogan,Y.,& Halife, K., "Tagging Accuracy Analysis on Part-of-Speech Taggers", Journal of Computer and Communications, 2014.

[13] Mitchell, P.M., Mary A.M., and Beatrice, S., "Building a Large Annotated Corpus of English: The Penn Treebank", Computational Linguistics Journal , Voul. 19, Number 21993.

[14] Saifee, V., & Jay, T., "Applications and Challenges for Sentiment Analysis: A Survey", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue2, 2013.

[15] Hassena, R.P., "Opinion Mining and Sentiment Analysis Challenges and Applications", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 5, 2014.

[16] Bing, L., Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, 2012.

[17] Archana, S., "Sentiment analysis of document based on annotation", CORR Journal, Vol. abs/1111.1648, 2011.

[18] Walter, K., & Mihaela, V., "Sentiment analysis for hotel reviews", proceedings of the computational linguistics-applications, Jacharanka Conference, 2011.

[19] A.Tamilsevi, & M.Parveen, T., "Sentiment analysis of micro blogs using opinion mining classification algorithm", International Journal of Science and Research (IJSR), Vol. 2 Issue 10, 2013.

[20] Nisha.J, & Dr.E. Kirubakaran, "M-Learning sentiment analysis with Data Mining Techniques", International Journal of Computer Science and Telecommunications, Volume 3, Issue 8, 2012.

[21] G¨unes, E., Ahamed, H., Qian, D., & Dragomir, R., "Improved Nearest Neighbor Methods for Text", Technical Report CSE-TR-576-11, University of Michigan, Department of Electrical Engineering and Computer Science, 2011.

[22] Frederick, L., "Implementation of Breiman's Random Forest Machine Learning Algorithm", ECE591Q Machine Learning Journal Paper, 2005.

[23] Christof, M., "Machine learning for query formulation in question and answering", Natural Language Engineering, Cambridge University, 2011.

[24] Steven, B., Ewan, K., and Edward, L., Natural Language Processing with Python, 1st Edition, O'Reilly Media publisher, 2009.

[25] Richard, S., Alex, P., Jean, Y.W., Jason, C., Christopher, D.M., Andrew, Y.N. & Christopher, P., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank", Conference on Empirical Methods in Natural Language Processing, 2013.

[26] I. Hemalatha , Dr. G. P Saradhi V., & Dr. A. Govardhan, "Preprocessing the Informal Text for efficient Sentiment Analysis" , International Journal of Emerging Trends & Technology in Computer Science (IJETTCS),Volume 1, Issue 2, 2012.

[27] Dushyant, B.R., & Samrat, K., "A Review on Emerging Trends of Web Mining and IT's Application", International Journal of Engineering Development and Research | IJEDR, 2013.

[28] Bo, P., & Lillian, L., "Opinion mining and sentiment analysis", Journal Foundations and Trends in Information Retrieval, Vol. 2, 2008.

[29] Bing, L., "Sentiment Analysis: A Multi-Faceted Problem", IEEE Intelligent Systems, 2010.

[30] Hang, C., Vibhu, M., & Mayur, D., "Comparative experiments on sentiment classification for online product reviews", American Association for Artificial Intelligence (AAAI Conference), 2006.

[31] Mr. Akshay, A. Adsod, & Prof. Nitin R.C., "A review on web mining", International Journal of Engineering Trends and Technology (IJETT)", Volume 10 Number 3, 2014.

[32] Alberto, H.F.L., Berthier, A.R-N, Altigran, S. d. S., and Juliana, S. T., "A brief survey of web data extraction tools", ACM SIGMOD, Vol.31, Issue 2,, New York , USA, June 2002

[33] Farah, B., Carmine, C., & Diego.R., "Sentiment analysis: Adjectives and Adverbs are better than Adjectives Alone", International Conference on Weblogs and Social Media -ICWSM , Boulder, CO USA, 2007.

[34] Utkarsh, G., Clemente, I., and Mike P. W.,Understanding Factors Inuencing the Citation Count of Networking Conference Papers, in proc. CHI 09 proceedings of the SIGCHI Conference on human factors in computing systems, 2009.

[35] Sujata, R. & Patreek, K., "Challenges of Sentiment Analysis and Existing State of Art", International Journal of Innovation and Research in Computer Science (IJIRCS), 2014.

[36] Chenghua, L., and Yulan, H., "Joint Sentiment/Topic Model for Sentiment Analysis", CIKM '09 Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, ACM 978-1-60558-512-3/09/11, 2009.

[37] Andreas, H., Gerhard, P., and Andreas, N., "A Brief Survey of Text Mining", LDV Forum - GLDV Journal for Computational Linguistics and Language Technology, 2005.

[38] Bas, H., Paul, v.I., Alexander, H., Flavius, F., & Uzay, K., "Accounting for Negation in sentiment analysis", DIR, Amsterdam, the Netherlands, 2011.

[39] John, M. B., "Counting publications - Journals vs. Conferences in Computer Science", in NIK-2013 conference, 2013.

[40] Quoc, L., & Tomas, M., "Distributed Representations of Sentences and Documents", Proceedings of the 31 st International Conference on Machine Learning, Beijing, China. JMLR: W&CP volume 32, 2014.

[41] Umer, F.,Yang, S., John, M. C., and Lee, G.,"Social Bookmarking for Scholarly Digital Libraries", Published by the IEEE Computer Society, *Pennsylvania State University*, December 2007.

[42] Pollyanna, G., Matheus, A., Fabrico, B., & Meeyoung, C.,"Comparing and Combining Sentiment Analysis Methods", ACM Association for Computing Machinery, 2013.

[43] Mark, C., Oliver, D., & Fatih, U., "Potential and Limitations of Commercial Sentiment Detection Tools", Emotion and Sentiment in Social and Expressive Media-ESSEM, 2013.

[44] Steven, B., Ewan, K., Edward, L., & Jason, B., "Multidisciplinary Instruction with the Natural Language Toolkit", Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL-08), Columbus, Ohio, USA, 2008.