too BIG to IGNORE

THE BUSINESS CASE FOR BIG DATA

PHIL SIMON Award-Winning Author of THE AGE OF THE PLATFORM

WILEY

Contents

“Today Big data affects everybody and will continue to do so for the foreseeable future. In *Too Big to Ignore*, Phil Simon makes the topic accessible and relatable. This important book shows people how to put Big Data to work for their organizations.”

–William McKnight, President, McKnight Consulting Group

“Simon has an uncanny ability to connect business cases with complex technical principles, and most importantly, clearly explain how everything comes together. In this book, Simon demystifies Big Data. Simon’s vision helps the rest of us understand how this evolving and pervasive subject affects businesses today.”

—Dalton Cervo, co-author of *Master Data Management in Practice—Achieving True Customer MDM* and president of Data Gap Consulting.

“From Twitter feeds to photo streams to RFID pings, the Big Data universe is rapidly expanding, providing unprecedented opportunities to understand the present and peer into the future. Tapping its potential while avoiding its pitfalls doesn’t take magic; it takes a map. In *Too Big to Ignore,* Phil Simon offers businesses a comprehensive, clear-eyed, and enjoyable guide to the next data frontier.”

—Chris Berdik, author of *Mind over Mind: The Surprising Power of Expectations*

“Business leaders are drowning in data, and the deluge has only just begun. In *Too Big to Ignore,* Simon delves into the world of Big Data, and makes the business case for capturing, structuring, analyzing, and visualizing the immense amount of information accessible to businesses. This book gives your organization the edge it needs to turn data into intelligence, and intelligence into action.”

—Paul Roetzer, Founder & CEO, PR 20/20; author of *The Marketing Agency Blueprint*

"Phil Simon's *Too Big to Ignore* clearly demonstrates the increasing role and value of Big Data. His illustrative case studies and engaging style will dispel any doubts executives may have about how Big Data *is driving* success in today's economy."

—Adrian C. Ott, award-winning author of *The 24-Hour Customer*

Wiley & SAS Business Series

The Wiley & SAS Business Series presents books that help senior-level managers with their critical management decisions.

Titles in the Wiley and SAS Business Series include:

Activity-Based Management for Financial Institutions: Driving Bottom-Line Results by Brent Bahnub

Big Data Analytics: Turning Big Data into Big Money by Frank Ohlhorst

Branded! How Retailers Engage Consumers with Social Media and Mobility by Bernie Brennan and Lori Schafer

Business Analytics for Customer Intelligence by Gert Laursen

Business Analytics for Managers: Taking Business Intelligence Beyond Reporting by Gert Laursen and Jesper Thorlund

The Business Forecasting Deal: Exposing Bad Practices and Providing Practical Solutions by Michael Gilliland

Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy by Olivia Parr Rud

CIO Best Practices: Enabling Strategic Value with Information Technology, Second Edition by Joe Stenzel

Connecting Organizational Silos: Taking Knowledge Flow Management to the Next Level with Social Media by Frank Leistner

Credit Risk Assessment: The New Lending System for Borrowers, Lenders, and Investors by Clark Abrahams and Mingyuan Zhang

Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring by Naeem Siddiqi

The Data Asset: How Smart Companies Govern Their Data for Business Success by Tony Fisher

Demand-Driven Forecasting: A Structured Approach to Forecasting by Charles Chase

The Executive's Guide to Enterprise Social Media Strategy: How Social Networks Are Radically Transforming Your Business by David Thomas and Mike Barlow

Executive's Guide to Solvency II by David Buckham, Jason Wahl, and Stuart Rose

Fair Lending Compliance: Intelligence and Implications for Credit Risk Management by Clark R. Abrahams and Mingyuan Zhang

Foreign Currency Financial Reporting from Euros to Yen to Yuan: A Guide to Fundamental Concepts and Practical Applications by Robert Rowan

Human Capital Analytics: How to Harness the Potential of Your Organization's Greatest Asset by Gene Pease, Boyce Byerly, and Jac Fitz-enz

Information Revolution: Using the Information Evolution Model to Grow Your Business by Jim Davis, Gloria J. Miller, and Allan Russell

Manufacturing Best Practices: Optimizing Productivity and Product Quality by Bobby Hull

Marketing Automation: Practical Steps to More Effective Direct Marketing by Jeff LeSueur

Mastering Organizational Knowledge Flow: How to Make Knowledge Sharing Work by Frank Leistner

The New Know: Innovation Powered by Analytics by Thornton May

Performance Management: Integrating Strategy Execution, Methodologies, Risk, and Analytics by Gary Cokins

Retail Analytics: The Secret Weapon by Emmett Cox

Social Network Analysis in Telecommunications by Carlos Andre Reis Pinheiro

Statistical Thinking: Improving Business Performance, Second Edition by Roger W. Hoerl and Ronald D. Snee

Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics by Bill Franks

The Value of Business Analytics: Identifying the Path to Profitability by Evan Stubbs

Visual Six Sigma: Making Data Analysis Lean by Ian Cox, Marie A. Gaudard, Philip J. Ramsey, Mia L. Stephens, and Leo Wright

Win with Advanced Business Analytics: Creating Business Value from Your Data by Jean Paul Isson and Jesse Harriott

For more information on any of the above titles, please visit www.wiley.com.

Too Big to Ignore

The Business Case for Big Data

Phil Simon

WILEY

Cover image: © Baris Simsek/iStockphoto
Cover design: John Wiley & Sons, Inc.

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

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

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

Library of Congress Cataloging-in-Publication Data:

ISBN 9781119217848 (paper)
ISBN 9781118638170 (Hardcover)
ISBN 9781118642108 (ebk)
ISBN 9781118641682 (ebk)
ISBN 9781118641866 (ebk)

Other Books by Phil Simon

Why New Systems Fail: An Insider's Guide to Successful IT Projects

The Next Wave of Technologies: Opportunities in Chaos

The New Small: How a New Breed of Small Businesses Is Harnessing the Power of Emerging Technologies

The Age of the Platform: How Amazon, Apple, Facebook, and Google Have Redefined Business

101 Lightbulb Moments in Data Management: Tales from the Data Roundtable (Editor)

The fact that we can now begin to actually look at the dynamics of social interactions and how they play out, and are not just limited to reasoning about averages like market indices is for me simply astonishing. To be able to see the details of variations in the market and the beginnings of political revolutions, to predict them, and even control them, is definitely a case of Promethean fire. Big Data can be used for good or bad, but either way it brings us to interesting times. We're going to reinvent what it means to have a human society.

—Sandy Pentland, Professor, MIT

Knowledge is good.

—Motto of fictitious Faber College, *Animal House*

List of Tables and Figures

Preface

Errors using inadequate data are much less than those using no data at all.

—Charles Babbage

It's about 7:30 a.m. on October 26, 2011, and I'm driving on The Strip in Las Vegas, Nevada. No, I'm not about to play craps or see Celine Dion. (While very talented, she's just not my particular brand of vodka.) I'm going for a more professional reason. Starting sometime in mid-2011, I started hearing more and more about something called *Big Data.* On that October morning, I was invited to IBM's Information on Demand (IOD) conference. It was high time that I learned more about this new phenomenon, and there's only so much you can do in front of a computer.

Beyond my insatiable quest for knowledge on all matters technology, truth be told, I went to IOD for a bunch of other reasons. First, it was convenient: The Strip is a mere fifteen minutes from my home. Second, the price was right: I was able to snake my way in for free. It turns out that, since I write for a few high-profile sites, some people think of me as a member of the media. (Funny how I never would have expected that ten years ago, but far be it from me to look a gift horse in the mouth.) Third, it was a good networking opportunity and my fourth book, *The Age of the Platform*, had just been published. I am familiar enough with the book business to know that authors have to get out there if they want to generate a buzz and move copies. These were all valid reasons to hop in my car, but for me there was an extra treat. I had the opportunity to meet and listen firsthand to the conference's two keynote speakers: Michael Lewis (one of my favorite writers) and a man by the name of Billy Beane.

For his part, Lewis wasn't at IOD to promote his latest opus like I was. On the contrary, he was there to speak about his 2003 book *Moneyball: The Art of Winning an Unfair Game.* The book had been enjoying a huge commercial resurgence as of late, thanks in no small part to the recent film of the same name starring some guy named Brad Pitt. I hadn't read *Moneyball* in some years, but I remember breezing through it. Lewis's writing style is nothing if not engaging. (He even made subprime mortgages and synthetic collateralized debt obligations [CDOs] interesting in *The Big Short*.)

I've always been a bit of a stats geek, and *Moneyball* instantly hit a nerve with me. It told the story of Beane, the general manager (GM) of the budget-challenged Oakland A's. Despite his team's financial limitations, he consistently won more games than most other mid-market teams—and even franchises like the New York Yankees that effectively printed their own money. The obvious question was how? Beane bucked convention and routinely ignored the advice of long-time baseball scouts, often earning their derision in the process. Instead, Beane predicated his management style on a rather obscure, statistics-laden field called *sabermetrics*. He signed free agents who he believed were undervalued by other teams. That is, he sought to exploit market inefficiencies.

One of Beane's favorite bargains: a relatively cheap player with a high on-base percentage (OBP).[5] In a nutshell, Beane's simple and irrefutable logic could be summarized as follows: players more likely to get on base are more likely to score runs. By extension, higher-scoring teams tend to win more games than their lower-scoring counterparts. But Beane didn't stop there. He was also partial to players (again, only at the right price) who didn't swing at the first pitch. Beane liked hitters who consistently made opposing pitchers work deep into the count. These patient batters

were more likely to make opposing pitches tired—and then give *everyone* on the A's better pitches to hit. (Again, more runs would result, as would more wins.)

Figure P.1 Michael Lewis and Billy Beane with Katty Kay at IBM Information on Demand 2011[1]

Source: Todd Watson

Back then, evaluating players based on unorthodox stats like these was considered heresy in traditional baseball circles. And that resistance was not just among baseball outsiders. In the late 1990s and early 2000s, a conflict within the A's organization was growing between Beane and his most visible employee: manager Art Howe. A former infielder with three teams over twelve years, Howe for one wasn't on board with Beane's unconventional program, to put it mildly. As Lewis tells it in *Moneyball*, Howe was nothing if not old school. He certainly didn't need some

newfangled, stat-obsessed GM telling him the X's and O's of baseball.

Oakland's internal conflict couldn't persist; a GM and manager have to be on the same page in all sports, and baseball is no exception. Rather than fire Howe outright (with the A's eating his $1.5 million salary), Beane got creative, as he is wont to do. He cajoled the New York Mets into taking him off their hands, not that the Mets needed much convincing. The team soon signed its new leader to a then-bawdy four-year, $9.4 million contract. After all, Howe had won a more-than-respectable 53 percent of his games with the small-market A's and *he just looks managerial*. The man has a great jaw. Imagine what Howe could do for a team with a big bankroll like the Mets?

Howe's tenure with the Mets was ignominious. The team won only 42 percent of its games on Howe's watch. After two seasons, the Mets realized what Beane knew long ago: Howe and his managerial jaw were much better in theory than in practice. In September 2004, the Mets parted ways with their manager.

While Beane may have been the first GM to embrace sabermetrics, he soon had company. His success bred many disciples in the baseball world and beyond. Count among them Theo Epstein, currently the President of Baseball Operations for the Chicago Cubs. In his previous role as GM of the Boston Red Sox, Epstein even hired Bill James, the godfather of sabermetrics. And it worked. Epstein won two World Series for the Sox, breaking the franchise's 86-year drought. Houston Rockets's GM Daryl Morey is bringing *Moneyball* concepts to the NBA. As a November 2012 *Sports Illustrated* article points out, the MIT MBA takes a radically different approach to player acquisition and development compared to his peers.[2]

And then there's the curious case of Kevin Kelley, the head football coach at the Pulaski Academy, a high school in Little

Rock, Arkansas. Kelley isn't your average coach. The man "stopped punting in 2005 after reading an academic study on the statistical consequences of going for the first down versus handing possession to the other team."[3] Coach Kelley simply refuses to punt. Ever. Even if it's fourth and 20 from his own ten-yard line. But it gets even better. Ever the contrarian, after Pulaski scores, Kelley has his kicker routinely try on-side kicks to try to get the ball right back. In one game, Kelley's team scored twenty-nine points before the opponent even touched the football![4] The results? The Bruins have won multiple state championships using their coach's unconventional style.

So why were Lewis and Beane the keynote speakers at IOD, a corporate information technology (IT) conference? Because, as *Moneyball* demonstrates so compellingly, today new sources of data are being used across many different fields in very unconventional and innovative ways to produce astounding results—and a swath of people, industries, and established organizations are finally starting to realize it.

This book explains why Big Data is a big deal. For example, residents in Boston, Massachusetts, are automatically reporting potholes and road hazards via their smartphones. Progressive Insurance tracks real-time customer driving patterns and uses that information to offer rates truly commensurate with individual safety. HR departments are using new sources of information to make better hiring decisions. Google accurately predicts local flu outbreaks based on thousands of user search queries. Amazon provides remarkably insightful, relevant, and timely product recommendations to its hundreds of millions of customers. Quantcast lets companies target precise audiences and key demographics throughout the Web. NASA runs contests via gamification site TopCoder, awarding prizes to those with the most innovative and cost-effective

solutions to its problems. Explorys offers penetrating and previously unknown insights into health care behavior.

How do these organizations and municipalities do it? Technology is certainly a big part, but in each case the answer lies deeper than that. Individuals at these organizations have realized that they don't have to be statistician Nate Silver to reap massive benefits from today's new and emerging types of data. And each of these organizations has embraced Big Data, allowing them to make astute and otherwise impossible observations, actions, and predictions.

It's time to start thinking big.

This book is about an unassailably important trend: Big Data, the massive amounts, new types, and multifaceted sources of information streaming at us faster than ever. Never before have we seen data with the volume, velocity, and variety of today. Big Data is no temporary blip of a fad. In fact, it is only going to intensify in the coming years, and its ramifications for the future of business are impossible to overstate.

Put differently, Big Data is becoming too big to ignore. And that sentence, in a nutshell, summarizes this book.

Phil Simon
Henderson, NV
March 2013

NOTES

1. Watson, Todd, "Information on Demand 2011: A Data-Driven Conversation with Michael Lewis & Billy Beane," October 26, 2011, http://turbotodd.wordpress.com/2011/10/26/information-

on-demand-2011-a-data-driven-conversation-with-michael-lewis-billy-beane/, retrieved December 11, 2012.
2. Ballard, Chris, “Lin’s Jumper, GM Morey’s Hidden Talents, More Notes from Houston,” November 30, 2012, http://sportsillustrated.cnn.com/2012/writers/chris_ballard/11/30/houston-rockets-jeremy-lin-james-harden-daryl-morey/index.html, retrieved December 11, 2012.
3. Easterbrook, Gregg, “New Annual Feature! State of High School Nation,” November 15, 2007, http://sports.espn.go.com/espn/page2/story?page=easterbrook/071113, retrieved December 11, 2012.
4. Wertheim, Jon, “Down 29-0 Before Touching the Ball,” September 15, 2012, http://sportsillustrated.cnn.com/2011/writers/scorecasting/09/15/kelley.pulaski/index.html, retrieved December 11, 2012.
5 For those of you not familiar with the term, *OBP* represents the true measure of how often a batter reaches base. It includes hits, walks, and times hit by a pitch. Beane also sought out those with high on-base plus slugging percentages. OPS equals the sum of a player’s OBP and slugging percentage (total bases divided by at bats).

Acknowledgments

Kudos to the Wiley team of Tim Burgard, Shelly Sessoms, Karen Gill, Johnna VanHoose Dinse, Chris Gage, and Stacey Rivera for making this book possible so quickly. You all were a “big” help.

I am grateful to smart cookies Charlie Lougheed, Jim McKeown, Jason Crusan, Jag Duggal, Jim Kelly, Clinton Bonner, William McKnight, Scott Kahler, and Seth Grimes for their time and expertise. Talking to these folks made research fun. A tip of the hat to Hope Nicora, Andy Havens, Adrian Ott, Brad Feld, Chris Berdik, Terri Griffith, Jim Harris, Dalton Cervo, Jill Dyché, Todd Hamilton, Tony Fisher, Ellen French, Dick and Bonnie Denby, Kristen Eckstein, Bob Charette, Andrew Botwin, Thor and Keri Sandell, Clair Byrd, Jay and Heather Etchings, Karlena Kuder, Luke “Heisenberg” Fletcher, Michael, Penelope, and Chloe DeAngelo, Shawn Graham, Chad Roberts, Sarah Terry, Jeff Lee, Mark Cenicola, Brenda Blakely, Colin Hickey, Bruce Webster, Alan Berkson, Michael West, John Spatola, Marc Paolella, Angela Bowman, and Brian and Heather Morgan and their three adorable kids.

Next up are the usual suspects: my longtime Carnegie Mellon friends Scott Berkun, David Sandberg, Michael Viola, Joe Mirza, and Chris McGee.

My heroes from Rush (Geddy, Alex, and Neil), Dream Theater (Jordan, John, John, Mike, and James), Marillion (h, Steve, Ian, Mark, and Pete), and Porcupine Tree (Steven, Colin, Gavin, John, and Richard) have given me many years of creative inspiration through their music. Keep on keepin’ on!

Vince Gilligan, Aaron Paul, Bryan Cranston, Dean Norris, Anna Gunn, Betsy Brandt, RJ Mitte, and the rest of the cast

and team of *Breaking Bad* make me want to do great work.

Next up: my parents. I’m not here without you.

Introduction: This Ain’t Your Father’s Data

Throughout history, in one field after another, science has made huge progress in precisely the areas where we can measure things—and lagged where we can’t.

—Samuel Arbesman

Car insurance isn’t a terribly sexy or dynamic business. For decades, it has essentially remained unchanged. Nor is it an egalitarian enterprise: while a pauper and a millionaire pay the same price for a stamp ($0.45 in the United States as of this writing), the car insurance world works differently. Some people just pay higher rates than others, and those rates have at least initially very little to do with whether one is a “safe” driver, whatever that means. Historically, many if not most car insurance policies were written based on very few independent variables: age, gender, zip code, previous speeding tickets and traffic violations, documented accidents, and type of car. As I found out more than twenty years ago, a newly licensed, seventeen-year-old guy in New Jersey who drives a sports car has to pay a boatload in car insurance for the privilege—even if he rarely drives above the speed limit, always obeys traffic signals, and has nary an accident on his record. Like just about every kid my age, I wasn’t happy about my rates. After all, I was an “above average” driver, or at least I liked to think so. Why should I have to pay such exorbitant fees?

Of course, we all can’t be above average; it’s statically impossible. Truth be told, I’m sure that back then I occasionally didn’t come to a complete stop at every red light. While I’ve never been arrested for DUI, to this day I don’t always obey the speed limit. (Shhh . . . don’t tell

anyone.) When I’m driving faster than the law says I should, I sometimes think of the famous George Stigler picture of Milton Friedman taken in the mid-twentieth-century. Friedman was paying a speeding ticket with, paradoxically, a big smile on this face. Why such joy? Because Friedman was an economist and, as such, he was rational to a fault. In his view of the world, the time that he regularly saved by exceeding the speed limit was worth more to him than the risk and fine of getting caught. To people like Friedman and me, speeding is only a simple expected value calculation: Friedman sped because the rewards outweighed the risks. When a cop pulled him over, he was glad to pay the fine. But I digress.

So why do most car insurance companies base their quotes and rates on relatively simple variables? The answer isn’t complicated, especially when you consider the age of these companies. Allstate opened its doors in 1931. GEICO was founded in 1936, and the Progressive Casualty Insurance Company set up shop only one year later. Think about it: seventy-five years ago, those primitive models represented the best that car insurance companies could do. While each has no doubt tweaked its models since then, old habits die hard, as we saw with Art Howe and Billy Beane in the Preface. For real change to happen, somebody needs to upset the applecart. In this way, car insurance is like baseball.

BETTER CAR INSURANCE THROUGH DATA

The similarities between the ostensibly unrelated fields of baseball and car insurance don’t end there. Much like the baseball revolution pioneered by Billy Beane, car insurance today is undergoing a fundamental transformation. Just ask

Joseph Tucci. As the CEO at data storage behemoth EMC Corporation, he knows a thing or thirty about data. On October 3, 2012, Tucci spoke with Cory Johnson of Bloomberg Television at an Intel Capital event in Huntington Beach, California. Tucci talked about the state of technology, specifically the impact of Big Data and cloud computing on his company—and others.[1] At one point during the interview, Tucci talked about advances in GPS, mapping, mobile technologies, and telemetry, the net result of which is revolutionizing many businesses, including car insurance. No longer are rates based upon a small, primitive set of independent variables. Car insurance companies can now get much more granular in their pricing. Advances in technology are letting them answer previously unknown questions like these:

- Which drivers routinely exceed the speed limit and run red lights?
- Which drivers routinely drive dangerously slow?
- Which drivers are becoming less safe—even if they have received no tickets or citations? That is, who used to generally obey traffic signals but don't anymore?
- Which drivers send text messages while driving? (This is a big no-no. In fact, texting while driving [TWD] is actually considerably more dangerous than DUI.[2] As of this writing, fourteen states have banned it.)
- Who's driving in a safer manner than six months ago?
- Does a man with two cars (a sports car and a station wagon) drive each differently?
- Which drivers and cars swerve at night? (This could be a manifestation of drunk driving.)
- Which drivers checked into a bar using FourSquare or Facebook and drove their own cars home (as opposed to taking a cab or riding with a designated driver)?

Thanks to these new and improved technologies and the data they generate, insurers are effectively retiring their

decades-old, five-variable underwriting models. In their place, they are implementing more contemporary, accurate, dynamic, and data-driven pricing models. For instance, in 2011, Progressive rolled out Snapshot, its Pay As You Drive (PAYD) program.[3] PAYD allows customers to voluntarily install a tracking device in their cars that transmits data to Progressive and possibly qualifies them for rate discounts. From the company's site:

> How often you make hard brakes, how many miles you drive each day, and how often you drive between midnight and 4 a.m. can all impact your potential savings. You'll get a Snapshot device in the mail. Just plug it into your car and drive like you normally do. You can go online to see your latest driving details and projected discount.

Is Progressive the only, well, progressive insurance company? Not at all. Others are recognizing the power of new technologies and Big Data. As Liane Yvkoff writes on CNET, "State Farm subscribers self-report mileage and GMAC uses OnStar vehicle diagnostics reports. Allstate's Drive Wise goes one step further and uses a similar device to track mileage, braking, and speeds over 80 mph, but only in Illinois."[4]

So what does this mean to the average driver? Consider two fictional people, both of whom hold car insurance policies with Progressive and opt in to PAYD:

- Steve, a twenty-one-year-old New Jersey resident who drives a 2012, tricked-out, cherry red Corvette
- Betty, a forty-nine-year-old grandmother in Lincoln, Nebraska, who drives a used Volvo station wagon

All else being equal, which driver pays the higher car insurance premium? In 1994, the answer was obvious: Steve. In the near future, however, the answer will be much less certain: *it will depend on the data*. That is, vastly different driver profiles and demographic information will mean less and less to car insurance companies. Traditional