Wiley CIO Series

Foreword by
JIM STOGDILL
General Manager,
Radar,
O'Reilly Media

BIG DATA
BIG ANALYTICS

EMERGING BUSINESS INTELLIGENCE AND ANALYTIC TRENDS FOR TODAY'S BUSINESSES

Michael Minelli • Michele Chambers • Ambiga Dhiraj

BIG DATA, BIG ANALYTICS

WILEY CIO SERIES

Founded in 1807, John Wiley & Sons is the oldest independent publishing company in the United States. With offices in North America, Europe, Asia, and Australia, Wiley is globally committed to developing and marketing print and electronic products and services for our customers' professional and personal knowledge and understanding.

The Wiley CIO series provides information, tools, and insights to IT executives and managers. The products in this series cover a wide range of topics that supply strategic and implementation guidance on the latest technology trends, leadership, and emerging best practices.

Titles in the Wiley CIO series include:

The Agile Architecture Revolution: How Cloud Computing, REST-Based SOA, and Mobile Computing Are Changing Enterprise IT by Jason Bloomberg

Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses by Michele Chambers, Ambiga Dhiraj, and Michael Minelli

The Chief Information Officer's Body of Knowledge: People, Process, and Technology by Dean Lane

CIO Best Practices: Enabling Strategic Value with Information Technology by Joe Stenzel, Randy Betancourt, Gary Cokins, Alyssa Farrell, Bill Flemming, Michael H. Hugos, Jonathan Hujsak, and Karl D. Schubert

The CIO Playbook: Strategies and Best Practices for IT Leaders to Deliver Value by Nicholas R. Colisto

Enterprise IT Strategy, + Website: An Executive Guide for Generating Optimal ROI from Critical IT Investments by Gregory J. Fell

Executive's Guide to Virtual Worlds: How Avatars Are Transforming Your Business and Your Brand by Lonnie Benson

Innovating for Growth and Value: How CIOs Lead Continuous Transformation in the Modern Enterprise by Hunter Muller

IT Leadership Manual: Roadmap to Becoming a Trusted Business Partner by Alan R. Guibord

Managing Electronic Records: Methods, Best Practices, and Technologies by Robert F. Smallwood

On Top of the Cloud: How CIOs Leverage New Technologies to Drive Change and Build Value Across the Enterprise by Hunter Muller

Straight to the Top: CIO Leadership in a Mobile, Social, and Cloud-based (Second Edition) by Gregory S. Smith

Strategic IT: Best Practices for IT Managers and Executives by Arthur M. Langer

Strategic IT Management: Transforming Business in Turbulent Times by Robert J. Benson

Transforming IT Culture: How to Use Social Intelligence, Human Factors and Collaboration to Create an IT Department That Outperforms by Frank Wander

Unleashing the Power of IT: Bringing People, Business, and Technology Together by Dan Roberts

The U.S. Technology Skills Gap: What Every Technology Executive Must Know to Save America's Future by Gary Beach

BIG DATA, BIG ANALYTICS

EMERGING BUSINESS INTELLIGENCE AND ANALYTIC TRENDS FOR TODAY'S BUSINESSES

Michael Minelli
Michele Chambers
Ambiga Dhiraj

John Wiley & Sons, Inc.

Cover image: © nobeastsofierce/Alamy
Cover design: John Wiley & Sons, Inc.

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

Library of Congress Cataloging-in-Publication Data

Minelli, Michael, 1974-
Big data, big analytics : emerging business intelligence and analytic trends for today's businesses / Michael Minelli, Michele Chambers, Ambiga Dhiraj.
pages cm
Includes bibliographical references and index.
ISBN 978-1-118-14760-3 (cloth); ISBN 978-1-118-22583-7 (ebk); ISBN 978-1-118-23915-5 (ebk); ISBN 978-1-118-26381-5 (ebk)
1. Business intelligence. 2. Information technology. 3. Data processing. 4. Data mining. 5. Strategic planning. I. Chambers, Michele. II. Dhiraj, Ambiga, 1975-III. Title.
HD38.7.M565 2013
658.4'72—dc23

2012044882

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

To my wife Jenny and our three incredible children, Jack, Madeline, and Max. Also to my parents, who have always been there for me.

—Mike

To my son Cole, who is the light of my life and the person who taught me empathy. Also to my adopted family and support system, Lisa Patrick, Pei Yee Cheng, and Patrick Thean. Finally, to my colleagues Bill Zannine, Brian Hess, Jon Niess, Matt Rollender, Kevin Kostuik, Krishnan Parasuraman, Mario Inchiosa, Thomas Baeck, Thomas Dinsmore, and Usama Fayyad, for their generous support.

—Michele

To Mu Sigmans all around the world for their passion toward building the decision sciences industry.

—Ambiga

CONTENTS

FOREWORD: BIG DATA AND CORPORATE EVOLUTION

When my friend Mike Minelli asked me to write this foreword I wasn't sure at first what I should put on paper. Forewords are often one part book summary and one part overview of the field. But when I read the draft Mike sent me I realized that this is a really good book, and it doesn't need either of those. Without any additional help from me it will give you plenty of insight into what is happening and why it's happening now, and it will help you see the possibilities for your industry in this transition to a data-centric age. Also, the book is just full of practical suggestions for what you can do about them. But perhaps there's an opportunity to establish a wider context. To explore what Big Data means across a broad arc of technological advancement. So rather than bore you with a summary of a book you're going to read anyway, I'll try to daub a bit of paint onto the big picture of what it all might mean.

This foreword is based on the thesis that Big Data isn't merely another technology. It isn't just another gift box en route to the world's systems integrators via the conveyor belt of Gartner hype cycles. I believe Big Data will follow digital computing and internetworking to take its place as the third epoch of the information age, and in doing so it will fundamentally alter the trajectory of corporate evolution. The corporation is about to undergo a change analogous to the rise of consciousness in humans.

So let's start at the beginning. The Industrial Age was an era of vast changes in society. We harnessed first steam and then electricity as prime movers to unleash astonishing increases in productivity. The result was the first sustained growth of wealth in human history.

Those early industrial concerns required vast pools of labor that gradually grew more specialized. To coordinate the efforts of all of those people, management developed systems of rules and hierarchy of authority. At massive scale the corporation was no longer the direct exercise of an owner's will, it was a kind of organism.

It was an organism whose systems of control were born out of the Napoleonic bureaucracy of the French State and its emphasis on specialized function, fixed rules, and rigid hierarchy. The "bureau" in bureaucracy literally means desk, and paper was both the storage mechanism in them and the signaling mechanism between them.

The bureaucracy was a form of organization that could process stimuli at scale and coordinate masses of participants, but it was, and remains today, severely limited in its evolutionary progress. Bureaucracy is the nematode of human industrial organization.

With over 24,000 species the nematode is a plentiful and adaptable round worm whose nervous system typically consists of 302 neurons. A mere 20 of those neurons are in its pharyngeal nervous system, the part that serves as a rudimentary brain. Yet it is able to maintain homeostasis, direct movement, detect information in its environment, create complex responses, and even manage some basic learning. So, it's a nice approximation for the bureaucratic corporation.

Despite its display of complex behaviors the nematode is of course completely unaware of them in any conscious sense. Its actions, like those of a bureaucracy, are reactive and dispositional. A worm bumps into something and is stimulated. Neurons fire. Worm reacts. It moves away or maybe eats what it bumped. Likewise shelves go empty and an order is placed. Papers move between desks. Trucks arrive. Shelves get replenished.

Worms and corporations are both complex event-processing engines, but they are largely deterministic. The corporation is evolving though, becoming more aware of its surroundings and emergent in its reactions. The information age, or the second industrial age, has been a major part of that.

In 1954 Joe Glickauf of Arthur Andersen implemented a payroll system for the General Electric Corporation on a UNIVAC 1 digital electronic computer. He thus introduced the computational epoch of the information age to the American corporation. (Incidentally, also creating the IT consulting industry.) Throughout the 1950s other corporations rapidly adopted systems like it to serve a wide spectrum of corporate processes. The corporation was still a nematode but we were wiring the worm and aggressively digitizing its nervous system.

Yet it remained basically the same worm. Sure, it became more efficient and could react faster but with basically the same dispositions, because as we automated those existing systems with computers we mimicked the paper. Invoices, accounts, and customer master files all simply migrated into the machine as we dumped file cabinets into database tables. We were wiring the worm, but we weren't re-wiring it.

So it remained a bureaucracy, just a more efficient, responsive, and scalable one. Yet this was the beginning of a symbiotic evolution between corporation and information age technology and it became a departure point in the corporation's further evolutionary history. This digital foundation is the substrate on which further evolutionary processes would occur.

Then about thirty years ago, Leonard Kleinrock, Lawrence Roberts, Robert Kahn, and Vint Cerf invented the Internet and ushered in the second epoch of information age, the network era.

Suddenly our little worm was connected to its peers and surrounding ecosystem in ways that it hadn't been before. Messaging between companies became as natural as messaging between desks and with later pushes by Jack Welch and others who understood the revolution that was at hand, those messages finally succumbed to the pull of digitization. The era of the paper purchase order and invoice finally died. The first 35 years of digitization had focused on internal processes; now the focus was more on interactions with the outside world. (I say more, because EDI had been around for a while. But it was with the cost structure of the Internet that it really took off.) For the worm it was like the evolution of a sixth sense. It could see further, predict deeper into the future, and respond faster.

But those new networks didn't just affect the way our corporations interacted with the outside world. They also began to erode the very foundation of bureaucracy: its hierarchy.

While the strict hierarchy of bureaucracy had been a force multiplier for labor during the industrial age, in practice it meant that a company could never be smarter than the smartest person at its head. Restrained by hierarchy, rigid rules, and specialized functions, the sum total of a corporation's intelligence was always much less than the sum of the intelligence of its participants.

With globalization, complex connections, and faster market cycle times the complexity of the corporation's environment has increased rapidly and has long since exceeded the complexity that any single person can understand. There has after all only been one Steve Jobs. Something had to give.

So corporations have (slowly) begun the journey toward more agile, network-enabled, learning organizations that can crowd source intelligence both within their ranks and from inside their customer bases. They are beginning to exhibit locally emergent behaviors in response to that learning. This is what is behind corporate mottos like Facebook's "Move fast and break stuff." It's just another way of saying that initiative is local and that the head can't know everything.

Of course companies in the network era still have organization charts. But they don't tell the whole story anymore. These days we need to analyze email patterns, phone records, instant messaging and other evidence of actual human connection to determine the real organizational model that emerges like an interstitial lattice within the official org chart.

So corporate evolution is no longer just incremental improvement along an efficiency and productivity vector. The very form of the corporation is changing, enabled by technology and spurred by the necessity of complexity and cycle times. The corporation is growing external sensors and the necessary neurons to deal with what it discovers. It is changing from dispositional and reactive to complex and emergent in order to better impedance match with the post-industrial world it occupies.

So here we are, at the doorstep of the Information Age's Big Data epoch. The corporation has already taken advantage of the computing and internetworking epochs to evolve significantly and adapt to a more complex world. But even bigger changes are ahead.

This book will take you through the entire Big Data story, so I'm not going to expound much on the meaning of Big Data here. I'll just describe enough to set the stage for the next phase of corporate evolution. And this is a key point: Big Data isn't Business Intelligence (BI) with bigger data.

We are no longer limited to the structured transactional world that has been the domain of corporate information technology for the last 55 years. Big Data represents a transition-in-kind for both storage and analysis. It isn't just about size.

The data your corporation does "BI" with today is mostly internally generated highly-structured transactional data. It's like a record of the neurons that fired. All too often the role of the business intelligence analyst really boils down to corporate kinesthesis. Reports are generated to tell the head of a hierarchy what its limbs are doing, or did.

But Big Data has the potential to be different. For one, often the data being analyzed will come from somewhere else, and in its original unstructured form. And two, we won't just be analyzing what we did; we'll be analyzing what is happening in the world around us, with all of the richness and detail of the original sensation.

Now we can think of web logs, video clips, voice response unit recordings, every document in every SharePoint repository, social data, open government data, partner data sets, and many more as part of our analytical corpus. No longer limited to mere introspection, analysis can be about more deeply detailed external sensing. What do my customers do? Who do they know? Were they happy or angry when they called? What are their network neighbors like and when and how much will they be influenced by them? Which of my customers are most similar? What are they saying about our competitors? What are they buying from our competitors? Are my competitors' parking lots full? And on and on…

Perhaps more importantly, how can this mass of data be turned directly into product, or at least an attribute of our products? Can we close the loop: from what we sense in our environment, to what to know, and to what we do?

The term data science speaks to the notion that we are now using data to apply the scientific method to our businesses. We create (or discover) hypotheses, run experiments, see if our customers react the way we predict and then build new products or interactions based on the results. Forward thinking companies are closing the loops so that the entire process runs without human intervention and products are updated in real time based on customer behavior or other inputs.

Put another way, the corporation's OODA Loop (Observe, Orient, Decide, Act. The work of USAF Col Boyd, the OODA loop describes a model for action in the face of uncertainty) is being implemented, at least in the tactical time scale, directly in the machinery of the corporation. Humans design the algorithms, but their participation isn't necessary beyond that. And unlike traditional BI, which focused on the OO of the OODA loop, the modern corporation has to directly integrate the Decide and Act phases to keep up with the dynamics of the modern market. It's not enough to be more analytical, future corporations will require greater product and organizational agility to act in real time.

As analogy, we humans experience our world in real time via internally rendered maps of our sensory perceptions, and we store those maps as memory. Maps are the scaffolding on which mind and our processes of self unfold. They are the evolutionary portal through which we passed from disposition to reasoning, when along the way we evolved from reactive worm to reasoning human.

By storing rich complex interactions, the corporation is beginning to create and store map-like structures as well. Instead of reducing complex interactions into the cartoonish renderings of summarized transactions, we are beginning to store the whole map, the pure bits from every sensor and touch point. And with the network and relationship data we are capturing now, corporate memories are beginning to look like the associative model of the human brain. The corporation isn't becoming a person, but it is becoming more than a worm. (I realize that as of this writing the Supreme Court disagrees with my assessment.) It's becoming intelligent.

The big data epoch will be one of a major transition. For the past 55 years the focus of information technology has been on wiring the worm for automation, efficiency, and productivity. Now I think we'll see that shift to support of the very intelligence of the corporation.

Until now we measured projects mostly on the ROI inherent in their potential cost savings. But we'll soon begin to think in terms of intelligentization—a made up word that means making something smarter. Our goal in business and IT will be the application of data and analytics to increasing corporate intelligence. Something like $IQ_{corp} = f(data, algorithms)$. That's an altogether different framing goal for technology, and it will mean new ways of organizing and conceptualizing how it is funded and delivered.

How does the data we capture and the algorithms we develop increase the intelligence of our organization? Can we begin to think in terms of something like an IQ for our companies—a combination of its sensory perception, recall, reasoning, and ability to act? Will we go from return on investment to acquisition of intelligence? Regardless, we will be building companies that are smarter and faster-reacting than the humans that run them.

Of course, this isn't the end of transactional IT. The corporation will have "vestigial IT" too just like the human brain still has regions remaining from our dispositional evolutionary past. After all, we still pull our hands away from a hot stove without thinking about it first, and companies will continue to automatically resupply empty shelves. But an intelligent corporation will be one with a seamlessly integrated post-dispositional reasoning mind wired for action. One that is more intelligent as a collection of people and as a set of systems than any member of its management, and one whose OODA loop often runs without human intervention.

Big Data is an epoch in the information age, and on the other side of this discontinuity in corporate evolution the companies you work for are going to be smarter.

Jim Stogdill
General Manager, Radar, O'Reilly Media

PREFACE

Big Data, *Big Analytics* is written for business managers and executives who want to understand more about "Big Data." In researching this book, we realized that there were many texts about high-level strategy and some that went deep into the weeds with sample code. We have attempted to create a balance between the two, making the topic accessible through stories, metaphors, and analogies even though it's a technical subject area.

We've started out the book defining Big Data and discussing why Big Data is important. We illustrate the value of Big Data through industry examples in Chapter 2 and then move into describing the enabling technology in Chapters 3 through 5. While we introduce the people working with Big Data earlier in the book, in Chapter 6 we dive deeper into the organization and the roles it takes to make Big Data successful in an organization. We wrap up the book with a thorough summary of the ethical and privacy issues surrounding Big Data in Chapter 7. *Big Data*, *Big Analytics* concludes with an entertaining lecture by Avinash Kaushik of Google.

We welcome feedback. If you have ideas on how we can make this book better—or what topics you'd like covered in a new edition, we'd love to hear from you. Please visit us at www.BigDataBigAnalytics.com.