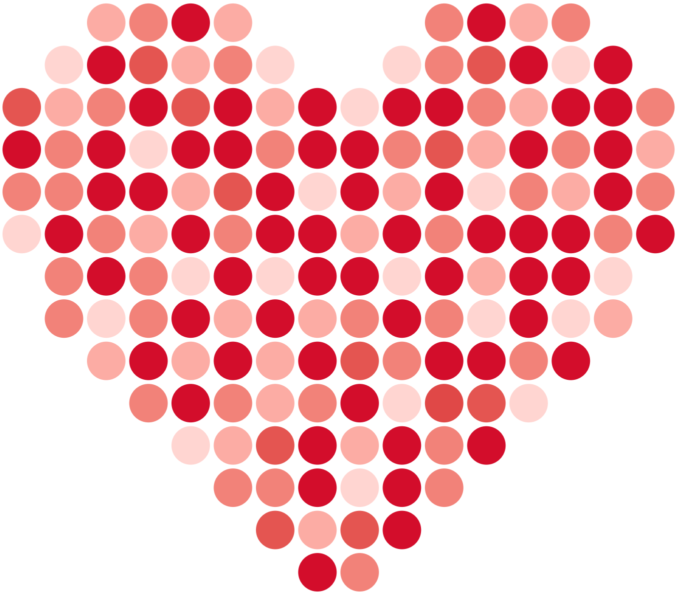
O'REILLY®

Learning to Love Data Science

Exploring Predictive Analytics, Machine Learning, Digital Manufacturing, and Supply Chain Optimization



Mike Barlow

O'REILLY®

Learning to Love
Data Science

Until recently, many people thought big data was a passing fad. "Data science" was an enigmatic term. Today, big data is taken seriously, and data science is considered downright sexy. With this anthology of reports from award-winning journalist Mike Barlow, you'll appreciate how data science is fundamentally altering our world, for better and for worse.

Barlow paints a picture of the emerging data space in broad strokes. From new techniques and tools to the use of data for social good, you'll find out how far data science reaches.

With this anthology, you'll learn how:

- Big data is driving a new generation of predictive analytics, creating new products, new business models, and new markets
- New analytics tools let businesses leap beyond data analysis and go straight to decision-making
- Indie manufacturers are blurring the lines between hardware and software products
- Companies are learning to balance their desire for rapid innovation with the need to tighten data security
- Big data and predictive analytics are applied for social good, resulting in higher standards of living for millions of people
- Advanced analytics and low-cost sensors are transforming equipment maintenance from a cost center to a profit center

Mike Barlow is an award-winning journalist, author, and communications strategy consultant. Since launching his own firm, Cumulus Partners, he has represented major organizations in a number of industries.

DATA

US $19.99 CAN $22.99

ISBN: 978-1-491-93658-0

Twitter: @oreillymedia
facebook.com/oreilly

Learning to Love Data Science

Explorations of Emerging Technologies and Platforms for Predictive Analytics, Machine Learning, Digital Manufacturing, and Supply Chain Optimization

Mike Barlow

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Learning to Love Data Science

by Mike Barlow

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Marie Beaugureau
Production Editor: Nicholas Adams
Copyeditor: Sharon Wilkey
Proofreader: Sonia Saruba
Interior Designer: David Futato
Cover Designer: Ellie Volckhausen
Illustrator: Rebecca Demarest

November 2015: First Edition

Revision History for the First Edition

2015-10-26: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781491936580* for release details.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-93658-0

[LSI]

For Darlene, Janine, and Paul

Table of Contents

Foreword . . . . **ix**

Editor's Note . . . . **xiii**

Preface . . . . **xv**

1. The Culture of Big Data Analytics **1**
It's Not Just About Numbers 1
Playing by the Rules 3
No Bucks, No Buck Rogers 5
Operationalizing Predictability 7
Assembling the Team 9
Fitting In 13

2. Data and Social Good **17**
Hearts of Gold 17
Structuring Opportunities for Philanthropy 19
Telling the Story with Analytics 21
Data as a Pillar of Modern Democracy 22
No Strings Attached, but Plenty of Data 24
Collaboration Is Fundamental 25
Conclusion 28

3. Will Big Data Make IT Infrastructure Sexy Again? **31**
Moore's Law Meets Supply and Demand 31
Change Is Difficult 33
Meanwhile, Back at the Ranch… 36

Throwing Out the Baby with the Bathwater? 37
API-ifiying the Enterprise 40
Beyond Infrastructure 41
Can We Handle the Truth? 42

4. When Hardware Meets Software. 45
Welcome to the Age of Indie Hardware 45
Mindset and Culture 48
Tigers Pacing in a Cage 49
Hardware Wars 51
Does This Mean I Need to Buy a Lathe? 53
Obstacles, Hurdles, and Brighter Street Lighting 54

5. Real-Time Big Data Analytics. 59
Oceans of Data, Grains of Time 59
How Fast Is Fast? 61
How Real Is Real Time? 64
The RTBDA Stack 66
The Five Phases of Real Time 68
How Big Is Big? 70
Part of a Larger Trend 71

6. Big Data and the Evolving Role of the CIO. 73
A Radical Shift in Focus and Perspective 74
Getting from Here to There 76
Behind and Beyond the Application 77
Investing in Big Data Infrastructure 79
Does the CIO Still Matter? 81
From Capex to Opex 81
A More Nimble Mindset 82
Looking to the Future 84
Now Is the Time to Prepare 85

7. Building Functional Teams for the IoT Economy. 87
A More Fluid Approach to Team Building 90
Raising the Bar on Collaboration 92
Worlds Within Worlds 93
Supply Chain to Mars 95
Rethinking Manufacturing from the Ground Up 97
Viva la Revolución? 98

8. **Predictive Maintenance: A World of Zero Unplanned Downtime. . 101**
Breaking News 102
Looking at the Numbers 102
Preventive Versus Predictive 105
Follow the Money 106
Not All Work Is Created Equal 107
Building a Foundation 109
It’s Not All About Heavy Machinery 110
The Future of Maintenance 111

9. **Can Data Security and Rapid Business Innovation Coexist?. 113**
Finding a Balance 113
Unscrambling the Eggs 115
Avoiding the “NoSQL, No Security” Cop-Out 117
Anonymize This! 119
Replacing Guidance with Rules 122
Not to Pass the Buck, but… 124

10. **The Last Mile of Analytics: Making the Leap from Platforms to Tools. 125**
Inching Closer to the Front Lines 126
The Future Is So Yesterday 126
Above and Beyond BI 128
Moving into the Mainstream 131
Transcending Data 135

Foreword

I met Mike Barlow a couple of years ago at an industry conference in New York. Our mutual interest in the Industrial Internet of Things (IIoT) has led to many interesting conversations, and I have observed some parallels in our experiences as authors.

We have both written about the convergence of key trends such as big data analytics, digital manufacturing, and high-speed networks. We both believe in the IIoT's potential to create new jobs, open new markets, and usher in a new age of global prosperity.

And both of us are glad he landed on the name *Learning to Love Data Science* for his book. He easily could have named it *How Data Science Is Helping Us Build a Better, Safer, and Cleaner World.*

Mike and I agree that information captured from machines, fleets of vehicles, and factories can be harnessed to drive new levels of efficiency and productivity gains. As much as I love data science, what I love even more is how it can unleash the power of innovation and creativity across product development, manufacturing, maintenance, and asset performance management.

We're not talking about ordinary analytics, like the kind that serve up recommendations when you use a search engine, but the complex physics-based analytics that detect meaningful patterns before they become an unforeseen problem, pitfall, or missed opportunity. This enables us to deliver positive outcomes like predicting service disruptions before they occur, across a wider spectrum of industries, affecting more people in more places than we could have dreamed of even three years ago.

Recently, I've read about how data science and advanced analytics are replacing traditional science. Commentary like, "All you need to do is look at the data," or "The data will tell you everything you need to know," is espoused without really understanding or appreciating what is happening in the background.

Data science isn't "replacing" anything; to the contrary, data science is adding to our appreciation of the world around us. Data science helps us make better decisions in a complex universe. And I cannot imagine a scenario in which the data itself will simply tell you everything you need to know.

In the future, I envision a day in which data science is so thoroughly embedded into our daily routines that it might seem as though the data itself is magically generating useful insights. As Arthur C. Clarke famously observed, "Any sufficiently advanced technology is indistinguishable from magic." Perhaps in the future, data science will indeed seem like magic.

Today, however, heavy lifting of data science is still done by real people. Personally, I believe human beings will always be in the loop, helping us interpret streams of information and finding meaning in the numbers. We will move higher up in the food chain, not be pushed out of the picture by automation. The future of work enhanced by data will enable us to focus on higher-level tasks.

From my perspective, data is a foundational element in a new and exciting era of connected devices, real-time analytics, machine learning, digital manufacturing, synthetic biology, and smart networks. At GE, we're taking a leadership role in driving the IIoT because we truly believe data will become a natural resource that ignites the next industrial revolution and helps humanity by making a positive difference in communities around the world.

How much will the IIoT contribute to the global economic picture? There's a range of estimates. The McKinsey Global Institute estimates it will generate somewhere between $3.4 trillion and $11.1 trillion annually in economic value by 2025. The World Economic Forum (WEF) predicts it will generate $14.2 trillion in 2030. I think it's safe to say we're on the cusp of something big.

Of course, it involves more than just embracing the next wave of disruptive innovation and technology. The people, processes, and culture around the technology and innovation also have to change. Frankly, the technology part is easy.

Standing up a couple of Hadoop clusters and building a data lake doesn't automatically make your company a data-driven enterprise. Here's a brief list of what you'll really need to think about, understand, and accept:

- How the cultural transformation from analogue to digital impacts people and fundamentally changes how they use data.
- Why it's imperative to deliver contextually relevant insights to people anywhere in the world, precisely when those insights are needed to achieve real business outcomes.
- Creating minimally viable products and getting them to market before your competitors know what you're doing.
- Understanding how real machines work in the real world.
- Rewarding extreme teamwork and incenting risk-takers who know how to create disruptive innovation while staying focused on long-term strategic goals.

The Industrial Internet of Things isn't just about data and analytics. It's about creating a new wave of operational efficiencies that result in smarter cities, zero unplanned outages of power and critical machinery, enormous savings of fuel and energy, and exponentially better management of natural resources. Achieving those goals requires more than just programming skills—you also need domain expertise, business experience, imagination, and the ability to lead. That's when the real magic begins.

This collection of reports will expand your understanding of the opportunities and perils facing us at this particular moment in history. Consider it your head start on a journey of discovery, as we traverse the boundary zone between the past, present, and future.

—William Ruh,
Chief Digital Officer,
GE Software

Editor's Note

This book is a collection of reports that Mike Barlow wrote for O'Reilly Media in 2013, 2014, and 2015. The reports focused on topics that are generally associated with data science, machine learning, predictive analytics, and "big data," a term that has largely fallen from favor.

Since Mike is a journalist and not a scientist, he approached the reports from the perspective of a curious outsider. The reports betray his sense of amused detachment, which is probably the right way to approach writing about a field like data science, and his ultimate faith in the value of technology, which seems unjustifiably optimistic.

At any rate, the reports provide valuable snapshots, taken almost randomly, of a field whose scale, scope, and influence are growing steadily. Mike's reports are like dispatches from a battlefield; they aren't history, but they provide an interesting and reasonably accurate picture of life on the front lines.

—Michael Loukides,
Vice President, Content
Strategy, O'Reilly Media

Preface

I first heard the term "data science" in 2011, during a conversation with David Smith of Revolution Analytics. David led me to Drew Conway, whose data science Venn diagram (reproduced with his permission in Figure 1-1) has acquired the legendary status of an ancient rune or hieroglyph.

Like its cousin, "big data," data science is a fuzzy and imprecise term. But it gets the job done, and there's something appealing about appending the word "science" to "data." It takes the sting out of both words. As a bonus, it enables the creation of another wonderful and equally confusing term, "data scientist."

Confusing is the wrong word. Redundant is a better choice. Science is inseparable from data. There is no science without data. Calling someone a "data scientist" is like calling someone a "professional Major League Baseball player." All the players in Major League Baseball are paid to play ball. Therefore, they are professionals, no matter how poorly they perform on any given day at the ballpark.

That said, the term "data scientist" suggests a certain raffish quality. Indeed, the early definitions of data science usually included hacking as a foundational element in the process. Maybe that's why so many writers think the term "data science" is sexy—it conveys a sense of the unorthodox. It requires ingenuity, fearlessness, and deep knowledge of arcane rituals. Like big data, it's shrouded in mystery.

That's exactly the sort of thinking that gets writers excited and drives editors crazy. Imprecise definitions aside, there's an audience for stories about data science. That's the reason why books like this one are published: They feed our need for understanding something that seems important and yet resists easy explanations.

I certainly hope you find the contents of this book interesting, entertaining, and educational. This book won't teach you how to become a data scientist, but it will give you fairly a decent idea of the ways in which data science is fundamentally altering our world, for better and for worse.

As you might have already guessed, the main audience for this book isn't data scientists, per se. I think it's safe to assume they already love data science, to one degree or another. This book is written primarily for people who want to learn a bit about data science but would rather not sign up for an online class or attend a lecture at their local library.

Careful readers will notice that I rather carelessly use the terms "data science" and "big data" interchangeably, like the way some people use the terms "Middle Ages" and "Medieval Period" interchangeably. I am guilty as charged, and I hope you can forgive me.

Safari® Books Online

Safari Books Online is an on-demand digital library that delivers expert content in both book and video form from the world's leading authors in technology and business.

Technology professionals, software developers, web designers, and business and creative professionals use Safari Books Online as their primary resource for research, problem solving, learning, and certification training.

Safari Books Online offers a range of plans and pricing for enterprise, government, education, and individuals.

Members have access to thousands of books, training videos, and prepublication manuscripts in one fully searchable database from publishers like O'Reilly Media, Prentice Hall Professional, Addison-Wesley Professional, Microsoft Press, Sams, Que, Peachpit Press, Focal Press, Cisco Press, John Wiley & Sons, Syngress, Morgan Kaufmann, IBM Redbooks, Packt, Adobe Press, FT Press, Apress, Manning, New Riders, McGraw-Hill, Jones & Bartlett, Course Technology, and hundreds more. For more information about Safari Books Online, please visit us online.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *http://bit.ly/learningtolovedatascience*.

To comment or ask technical questions about this book, send email to *bookquestions@oreilly.com*.

For more information about our books, courses, conferences, and news, see our website at *http://www.oreilly.com*.

Find us on Facebook: *facebook.com/oreilly*

Follow us on Twitter: *twitter.com/oreillymedia*

Watch us on YouTube: *www.youtube.com/oreillymedia*

Acknowledgments

This book is a work of journalism, not science. It's based on the aggregated wisdom of many sources, interviewed over the course of several years. All the sources cited in the original reports had the opportunity to review what I'd written about them prior to publication, which I think is a fair practice.

A long time ago, journalists invented an early form of crowdsourcing. We called it "multiple sourcing." Back in the old days, our gruff editors would reflexively spike "one-source" stories. As a result, we learned quickly to include quotes and supporting information from as many sources as possible. Multiple sourcing was also a great CYA (cover your ass) strategy: if you wrote something in a story that turned out to be incorrect, you could always blame the sources.

This book would not have been possible without the cooperation of many expert sources, and I thank them profusely for generously sharing their time and knowledge.

I owe special thanks to Mike Loukides and the wonderfully talented group of editors at O'Reilly Media who worked with me on this project: Holly Bauer, Marie Beaugureau, Susan Conant, and Timothy McGovern. Additionally, I am grateful for the support and guidance provided by Edith Barlow, Greg Fell, Holly Gilthorpe, Cornelia Lévy-Bencheton, Michael Minelli, William Ruh, Joseph Salvo, and Amy Sarociek. Thank you all.

CHAPTER 1

The Culture of Big Data Analytics

Topline Summary

Hollywood loves the myth of a lone scientist working late nights in a dark laboratory on a mysterious island, but the truth is far less melodramatic. Real science is almost always a team sport. Groups of people, collaborating with other groups of people, are the norm in science—and data science is no exception to the rule.

When large groups of people work together for extended periods of time, a culture begins to emerge. This paper, written in the spring of 2013, was an early attempt at describing the people and processes of the emerging culture of data science.

It's Not Just About Numbers

Today's conversational buzz around big data analytics tends to hover around three general themes: technology, techniques, and the imagined future (either bright or dystopian) of a society in which big data plays a significant role in everyday life.

Typically missing from the buzz are in-depth discussions about the people and processes—the cultural bedrock—required to build viable frameworks and infrastructures supporting big data initiatives in ordinary organizations.

Thoughtful questions must be asked and thoroughly considered. Who is responsible for launching and leading big data initiatives? Is it the CFO, the CMO, the CIO, or someone else? Who determines the success or failure of a big data project? Does big data require corporate governance? What does a big data project team look like? Is it a mixed group of people with overlapping skills or a handpicked squad of highly trained data scientists? What exactly is a data scientist?

Those types of questions skim the surface of the emerging cultural landscape of big data. They remind us that big data—like other so-called technology revolutions of the recent past—is also a cultural phenomenon and has a social dimension. It’s vitally important to remember that most people have not considered the immense difference between a world seen through the lens of a traditional relational database system and a world seen through the lens of a Hadoop Distributed File System.

This paper broadly describes the cultural challenges that invariably accompany efforts to create and sustain big data initiatives in a global economy that is increasingly evolving toward the Hadoop perspective, but whose data-management processes and capabilities are still rooted firmly in the traditional architecture of the data warehouse.

The cultural component of big data is neither trivial nor free. It is not a list of “feel-good” or “fluffy” attributes that are posted on a corporate website. Culture (that is, people and processes) is integral and critical to the success of any new technology deployment or implementation. That fact has been demonstrated repeatedly over the past six decades of technology evolution. Here is a brief and incomplete list of recent “technology revolutions” that have radically transformed our social and commercial worlds:

- The shift from vacuum tubes to transistors
- The shift from mainframes to client servers and then to PCs
- The shift from written command lines to clickable icons
- The introduction and rapid adoption of enterprise resource planning (ERP), e-commerce, sales-force automation, and customer relationship management (CRM) systems
- The convergence of cloud, mobile, and social networking systems

Each of those revolutions was followed by a period of intense cultural adjustment as individuals and organizations struggled to capitalize on the many benefits created by the newer technologies. It seems unlikely that big data will follow a different trajectory. Technology does not exist in a vacuum. In the same way that a plant needs water and nourishment to grow, technology needs people and processes to thrive and succeed.

According to Gartner, 4.4 million big data jobs will be created by 2014, and only a third of them will be filled. Gartner's prediction evokes images of "gold rush" for big data talent, with legions of hardcore quants converting their advanced degrees into lucrative employment deals. That scenario promises high times for data analysts in the short term, but it obscures the longer-term challenges facing organizations that hope to benefit from big data strategies.

Hiring data scientists will be the easy part. The real challenge will be integrating that newly acquired talent into existing organizational structures and inventing new structures that will enable data scientists to generate real value for their organizations.

Playing by the Rules

Misha Ghosh is a global solutions leader at MasterCard Advisors, the professional services arm of MasterCard Worldwide. It provides real-time transaction data and proprietary analysis, as well as consulting and marketing services. It's fair to say that MasterCard Advisors is a leader in applied data science. Before joining MasterCard, Ghosh was a senior executive at Bank of America, where he led a variety of data analytics teams and projects. As an experienced practitioner, he knows his way around the obstacles that can slow or undermine big data projects.

"One of the main cultural challenges is securing executive sponsorships," says Ghosh. "You need executive-level partners and champions early on. You also need to make sure that the business folks, the analytics folks, and the technology folks are marching to the same drumbeat."

Instead of trying to stay "under the radar," Ghosh advises big data leaders to play by the rules. "I've seen rogue big data projects pop up, but they tend to fizzle out very quickly," he says. "The old adage that it's better to seek forgiveness afterward than to beg for permission doesn't really hold for big data projects. They are simply too expensive and they require too much collaboration across various parts of the enterprise. So you cannot run them as rogue projects. You need executive buy-in and support."

After making the case to the executive team, you need to keep the spark of enthusiasm alive among all the players involved in supporting or implementing the project. According to Ghosh, "It's critical to maintain the interest and attention of your constituency. After you've laid out a roadmap of the project so everyone knows where they are going, you need to provide them with regular updates. You need to communicate. If you stumble, you need to let them know why you stumbled and what you will do to overcome the barriers you are facing. Remember, there's no clear path for big data projects. It's like *Star Trek*—you're going where no one has gone before."

At present, there is no standard set of best practices for managing big data teams and projects. But an ad hoc set of practices is emerging. "First, you must create transparency," says Ghosh. "Lay out the objectives. State explicitly what you intend to accomplish and which problems you intend to solve. That's absolutely critical. Your big data teams must be 'use case-centric.' In other words, find a problem first and then solve it. That seems intuitive, but I've seen many teams do exactly the opposite: first they create a solution and then they look for a problem to solve."

Marcia Tal pioneered the application of advanced data analytics to real-world business problems. She is best known in the analytics industry for creating and building Citigroup's Decision Management function. Its charter was seeking significant industry breakthroughs for growth across Citigroup's retail and wholesale banking businesses. Starting with three people in 2001, Tal grew the function into a scalable organization with more than 1,000 people working in 30 countries. She left Citi in 2011 and formed her own consulting company, Tal Solutions, LLC.

"Right now, everyone focuses on the technology of big data," says Tal. "But we need to refocus our attention on the people, the processes, the business partnerships, revenue generation, P&L impact, and business results. Most of the conversation has been about generating insights from big data. Instead, we should be talking about how to translate those insights into tangible business results."

Creating a sustainable analytics function within a larger corporate entity requires support from top management, says Tal. But the strength and quality of that support depends on the ability of the analytics function to demonstrate its value to the corporation.

“The organization needs to see a revenue model. It needs to perceive the analytics function as a revenue producer, and not as a cost center. It needs to see the value created by analytics,” says Tal. That critical shift in perception occurs as the analytics function forms partnerships with business units across the company and consistently demonstrates the value of its capabilities.

“When we started the Decision Management function at Citi, it was a very small group and we needed to demonstrate our value to the rest of the company. We focused on specific business needs and gaps. We closed the gaps, and we drove revenue and profits. We demonstrated our ability to deliver results. That’s how we built our credibility,” says Tal.

Targeting specific pain points and helping the business generate more revenue are probably the best strategies for assuring ongoing investment in big data initiatives. “If you aren’t focusing on real pain points, you’re probably not going to get the commitment you need from the company,” says Tal.

No Bucks, No Buck Rogers

Russ Cobb, vice president of marketing and alliances at SAS, also recommends shifting the conversation from technology to people and processes. “The cultural dimension potentially can have a major impact on the success or failure of a big data initiative,” says Cobb. “Big data is a hot topic, but technology adoption doesn’t equal ROI. A company that doesn’t start with at least a general idea of the direction it’s heading in and an understanding of how it will define success is not ready for a big data project.”

Too much attention is focused on the cost of the investment and too little on the expected return, says Cobb. “Companies try to come up with some measure of ROI, but generally, they put more detail around the ‘I’ and less detail around the ‘R.’ It is often easier to calculate costs than it is to understand and articulate the drivers of return.”

Cobb sees three major challenges facing organizations with big plans for leveraging big data. The first is not having a clear picture of the destination or desired outcome. The second is hidden costs, mostly in the area of process change. The third and thorniest challenge is organizational. "Are top and middle managers ready to push their decision-making authority out to people on the front lines?" asks Cobb. "One of the reasons for doing big data is that it moves you closer to real-time decision making. But those kinds of decisions tend to be made on the front lines, not in the executive suite. Will management be comfortable with that kind of cultural shift?"

Another way of phrasing the question might be: Is the modern enterprise really ready for big data? Stephen Messer, cofounder and vice chairman of Collective[i], a software-as-a-service business intelligence solution for sales, customer service, and marketing, isn't so sure. "People think this is a technological revolution, but it's really a business revolution enabled by technology," says Messer. Without entrepreneurial leadership from the business, big data is just another technology platform.

"You have to start with the business issue," says Messer. "You need a coalition of people inside the company who share a business problem that can be solved by applying big data. Without that coalition, there is no mission. You have tactics and tools, but you have no strategy. It's not transformational."

Michael Gold, CEO of Farsite, a data analytics firm whose clients include Dick's Sporting Goods and the Ohio State University Medical Center, says it's important to choose projects with manageable scale and clearly defined objectives.

"The questions you answer should be big enough and important enough for people to care," says Gold. "Your projects should create revenue or reduce costs. It's harder to build momentum and maintain enthusiasm for long projects, so keep your projects short. Manage the scope, and make sure you deliver some kind of tangible results."

At a recent Strata + Hadoop World conference in New York, Gold listed three practical steps for broadening support for big data initiatives:

1. Demonstrate ROI for a business use case.
2. Build a team with the skills and ability to execute.
3. Create a detailed plan for operationalizing big data.

"From our perspective, it's very important that all of the data scientists working on a project understand the client's strategic objectives and what problems we're trying to solve for them," says Gold. "Data scientists look at data differently (and better, we think) when they're thinking about answering a business question, not just trying to build the best analytical models."

It's also important to get feedback from clients early and often. "We work in short bursts (similar to a scrum in an Agile methodology) and then present work to clients so they can react to it," says Gold. "That approach ensures that our data scientists incorporate as much of the clients' knowledge into their work as possible. The short cycles require our teams to be focused and collaborative, which is how we've structured our data science groups."

Operationalizing Predictability

The term "data scientist" has been used loosely for several years, leading to a general sense of confusion over the role and its duties. A headline in the October 2012 edition of the *Harvard Business Review*, "Data Scientist: The Sexiest Job of the 21st Century," had the unintended effect of deepening the mystery.

In 2010, Drew Conway, then a Ph.D. candidate in political science at New York University, created a Venn diagram showing the overlapping skill sets of a data scientist (Figure 1-1). Conway began his career as a computational social scientist in the US intelligence community and has become an expert in applying computational methods to social and behavioral problems at large scale.

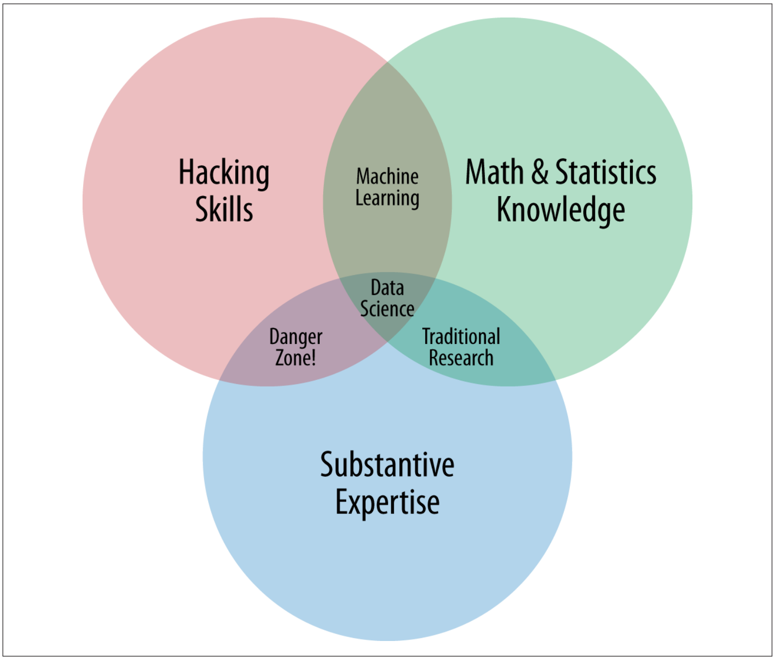


Figure 1-1. Conway's Venn diagram of a data scientist's skill sets

From Conway's perspective, a data scientist should possess the following:

- Hacking skills
- Math and statistical knowledge
- Substantive expertise

All three areas are important, but not everyone is convinced that one individual has to embody all the skills of a data scientist to play a useful role on a big data analytics team.

The key to success, as Michael Gold suggested earlier, is operationalizing the processes of big data. Taking it a step further, it is also important to demystify big data. While the *Harvard Business Review* certainly meant no harm, its headline had the effect of glamorizing rather than clarifying the challenges of big data.

Zubin Dowlaty, vice president of innovation and development at Mu Sigma, a provider of decision science services, envisions a future in which big data has become so thoroughly operationalized and automated that humans are no longer required.

"When I walk into an enterprise today, I see the humans are working at 90 percent capacity and the machines are working at 20 percent capacity," says Dowlaty. "Obviously, the machines are capable of handling more work. Machines, unlike humans, scale up very nicely."

Automation is a necessary step in the development of large-scale systems that feed on big data to generate real-time predictive intelligence. "Anticipation denotes intelligence," says Dowlaty, quoting a line from the science-fiction movie *The Fifth Element*. "Operationalizing predictability is what intelligence is all about."

Assembling the Team

At some point in the future, probably sooner rather than later, Dowlaty's vision of automated big data analytics will no doubt become reality. Until then, however, organizations with hopes of leveraging the potential of big data will have to rely on humans to get the work done.

In a 2012 paper,[1] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer presented the results of interviews with 35 data analysts working in commercial organizations in healthcare, retail, finance, and social networking. Hellerstein, a professor at UC Berkeley, summarized key findings of the paper at a recent Strata conference. The paper includes insights and models that will likely prove useful to anyone tasked with assembling a big data analytics team. Based on their interviews, the researchers perceive three basic analyst archetypes:

- *Hacker*
- *Scripter*
- *Application user*

The hacker is typically a fluent programmer and manipulator of data. The scripter performs most of his work within an existing software package and works mostly on data that has been retrieved

[1] "Enterprise Data Analysis and Visualization: An Interview Study."

from a data warehouse by information technology (IT) staff. The application user relies on spreadsheets or highly specialized applications and typically works on smaller data sets than hackers and scripters.

It is important for management to understand the differences between those types of analysts when staffing a data analytics team. Hackers are more likely to have a background in computer science. "They are folks who have good facility with programming and systems, but less facility with stats and some of the more 'scientific' aspects of data science. They also tend to have less contextual knowledge of the domain-specific questions being explored in the data," explains Hellerstein.

Scripters, on the other hand, are more likely to be trained statisticians, and app users are more likely to be business people. At the risk of oversimplification, a chart showing the three kinds of analysts and their typical academic backgrounds might look something like this:

Analyst type	Training or academic background
Hacker	Computer science major
Scripter	Statistics major
Application user	MBA

"No (single) one of these categories is more likely than another to succeed on its own," says Hellerstein. "You can teach stats and business to a hacker, or you can teach computer science and business to a scripter, or you can teach stats and computer science to an app user."

Scripters and app users would likely require some sort of self-service software to function without help from IT. Similar software might also be useful for hackers, sparing them the drudgery of data prep.

The good news is that several companies are working hard at developing self-service tools that will help analysts become more self-reliant and less dependent on IT. As the tools become more sophisticated and more widely available, it is possible that the distinctions between the three types of analysts might fade or at least become less problematic.

Even when a full suite of practical self-service tools becomes available, it might still make sense to hire a variety of analyst types. For instance, an analytics group that hired only hackers would be like a baseball team that signed only pitchers. Successful teams—whether in business or in sports—tend to include people with various skills, strengths, and viewpoints. Or to put it more bluntly, good luck trying to manage an analytics team made up solely of hackers.

The paper also describes five high-level tasks of data analysis:

- Discovery
- Wrangling
- Profiling
- Modeling
- Reporting

Each of the five tasks has a different workflow, presents a different set of challenges or pain points, and requires a different set of tools. Clearly, the universe of practical analytics is a blend of various tasks, tools, and workflows. More to the point, each stage of the analytics process requires an analyst or analysts with particular skills and a particular mindset.

Not all data analysts are created equal, nor are they likely to share the same zeal for different parts of the process. Some analysts will be better at some aspects of analysis than others. Putting together and managing teams that can handle all the necessary phases of data analysis is a major part of the cultural challenge facing organizations as they ramp up big data initiatives.

Team leadership is another challenge. MasterCard's Ghosh recommends that big data projects "be led by passionate and creative data scientists, not by bureaucrats or finance professionals." Others argue that big data initiatives should be led by seasoned corporate executives with boardroom negotiating skills and a keen understanding of how the C-suite operates.

Some companies have hired a chief analytics officer or created an enterprise analytics group that functions as a shared service, similar to an enterprise IT function. Most companies, however, embed analysts within separate business units.