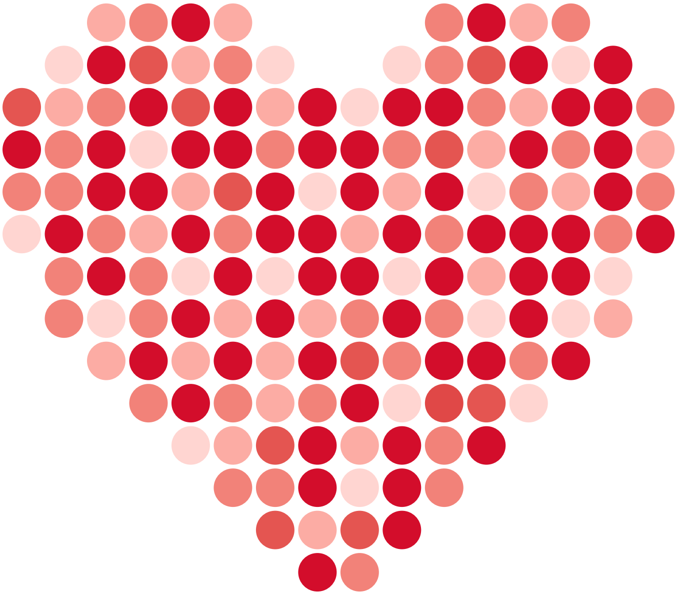
O'REILLY®

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

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.

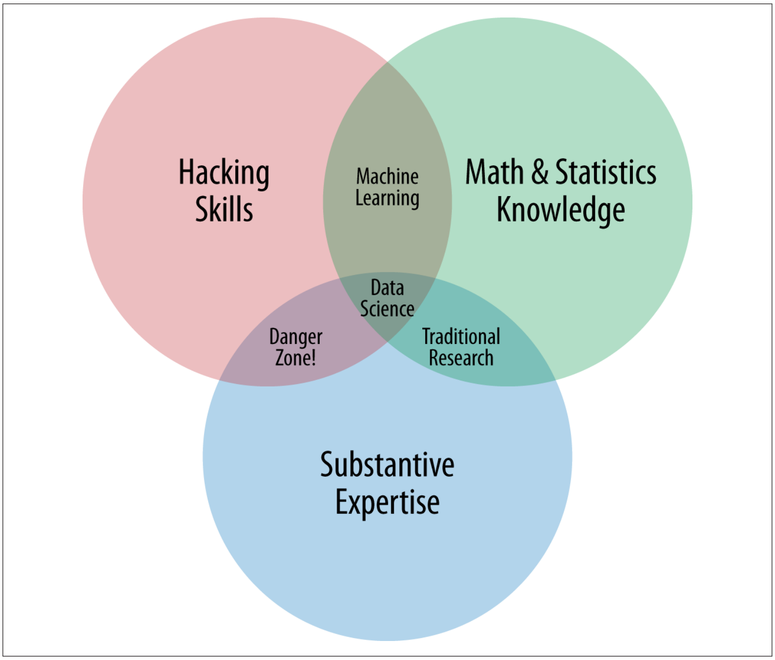


Figure 1-1. Conway's Venn diagram of a data scientist's skill sets

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.

The advantage of planting analysts in individual business units is that it puts the analysts closer to customers and end users. The downside of spreading analytic expertise among various units includes poor communication, lack of collaboration, and the tendency to reinvent the wheel to solve local problems instead of seeking help from other parts of the enterprise.

Another problem with the decentralized analytics model is lack of governance. Today, it is unusual to find the words "governance" and "analytics" in the same sentence. As big data takes on a higher profile in modern corporations, governance will almost certainly become an issue.

For example, very few data analysts save code or models that do not result in practical solutions to immediate problems. As a consequence, analysts can waste an incredible amount of effort making the same or similar mistakes. Unlike, say, chemistry or biology, in which the results of all experiments are duly noted and logged whether or not they are successful, the precise details of data science experiments are usually captured when the analyst succeeds at solving the particular problem at hand.

Another issue that arises from using Hadoop and other frameworks for handling large amounts of unstructured data is the preservation of documentation and potentially important details about the data.

Sean Kandel, a coauthor of the study referenced earlier, sees the "impulse to dump data into an HDFS" as a growing cultural challenge. "When you have to have a traditional data warehousing environment, there is more of a culture around governance and making sure the data that comes in is well structured and fits the global schema," says Kandel. "When you get away from those established practices, it becomes harder to work with the data."

As Kandel and his coauthors write in their paper:

> With relational databases, organizations typically design a database schema and structure incoming data upon load. This process is often time-consuming and difficult, especially with large complex data sets. With Hadoop, analysts typically take advantage of its ability to operate on less structured data formats. Instead of structuring the data up front during ingest, organizations commonly dump data files into the Hadoop Distributed File System (HDFS) with little documentation. Analysis of this data then requires parsing the data during Map-Reduce jobs or bulk reformatting to load

into relational databases. While remaining unstructured, the data may be difficult to search and profile due to the lack of a defined schema. In some cases, the analysts who originally imported and understood the data may no longer work at the company or may have forgotten important details.

"In a large company," says Kandel, "those people might be hard to find. Now you have some interesting questions: Who is responsible for annotating data? How do you structure the data warehouse? How do you convince people to take the time to label the data properly?"

The lack of a disciplined process—what some would call governance—for handling data at every stage of the analytics process suggests the need for automated systems that capture keystrokes or create audit trails that would make it possible for data scientists to replicate or re-examine the work of other data scientists.

Fitting In

Paul Kent is vice president of big data at SAS, one of the earliest and best-known makers of data analytics. He sees a sort of natural "give and take" between traditional analysts working with limited sets of structured data and a newer generation of analysts who seem comfortable handling an endless deluge of unstructured data.

"I think you have to give the newer analysts their own space. They'll need to preserve some of their independence. They won't be happy playing by the old-school rules," says Kent. "Big data has changed the way we look at data. It's messy, and it's not expensive to save. So we save as much as we can. And when we have questions in the future, we'll map those questions to the data that we've saved."

In the past, data infrastructures were designed around a known set of questions. Today, it's much harder to predict the questions that will be asked. That uncertainty makes it nearly impossible to build traditional-style infrastructures for handling big data.

"We really can't design the perfect structure for data and then just pour data into it," says Kent. "So you have to think about it the other way around. We don't even know the questions we're going to ask tomorrow or next month. So we keep as much data as we can, and we try to be as flexible as possible so we can answer questions when they come up."

The "old-school" perspective was that "if you think real hard, you can design a nice structure for your data and then fill it up whenever you get your data—every week, every day, or every hour," says Kent. If the structure you designed was good enough, it could be tweaked or modified over time to keep up with the changing needs of the market.

"The new school says, 'Nope, that won't work. Let's just save the data as it comes in. We'll merge it and join it and splice it on a case-by-case basis.' The new-school approach doesn't necessarily need a relational database. Sometimes they'll just work with raw files from the originating system," says Kent.

Andreas Weigend teaches at Stanford University and directs the Social Data Lab. The former chief scientist at Amazon, he helped the company build the customer-centric, measurement-focused culture that has become central to its success. Weigend sees data-driven companies following an evolutionary path from "data set to tool set to skill set to mindset." He suggests eight basic rules for organizations in search of a big data strategy:

1. Start with the problem, not with the data.
2. Share data to get data.
3. Align interests of all parties.
4. Make it trivially easy for people to contribute, connect, collaborate.
5. Base the equation of your business on customer-centric metrics.
6. Decompose the business into its "atoms."
7. Let people do what people are good at, and computers what computers are good at.
8. Thou shalt not blame technology for barriers of institutions and society.

Weigend's list of rules focuses entirely on the cultural side of big data. In some ways, it's like the driver's manual you read in high school: heavy on driving etiquette and light on auto mechanics. The miracle of the internal combustion engine is taken for granted. What matters now is traveling safely from point A to point B.

Conversations about big data have moved up the food chain. People seem less interested in the technical details and more interested in how big data can help their companies become more effective, more nimble, and more competitive. As Marcia Tal puts it, “The C-suite wants to know what big data is worth to the organization. They want to see the revenue it generates. They want to understand its value and measure the return on their investment.”

CHAPTER 2

Data and Social Good

Topline Summary

Can advances in data science be leveraged for social good? Is there a natural intersection between philanthropy and data science? The early hype around data science emphasized its power to make businesses more profitable and efficient. Is that the only plausible narrative?

Most data scientists are not looking to become software zillionaires. Most of them would prefer to use their skills and knowledge to make the world a better place. This report, written in the winter and spring of 2015, focuses on data scientists and statisticians who improve communities and help humanity in a variety of ways, from preventing teen suicides in American suburbs to raising funds for impoverished farmers in Africa.

Hearts of Gold

Several years ago, large management consulting firms began describing data as the "new oil"—a magically renewable and seemingly inexhaustible source of fuel for spectacular economic growth. The business media rapidly picked up on the idea, and reported breathlessly about the potential for data to generate untold riches for those wise enough to harness its awesome power.

At the same time, another story was unfolding. That story wasn't about a few smart guys getting rich. It was about people using data to improve lives and make the world a better place.

For many of us, it's an alluring narrative, perhaps because it supports our hope that deep down, data scientists and statisticians are nice people who value social good over crass materialism.

Megan Price, for example, is director of research at the Human Rights Data Analysis Group. She designs strategies and methods for using data to support human rights projects in strife-torn countries like Guatemala, Colombia, and Syria. "I've always been interested in both statistics and social justice," Price says. In college, she started off as a math major, switched to statistics, and later studied public health in grad school. "I was surrounded by people who were all really invested in using their math and science skills for social justice. It was a great environment for bringing those interests together."

In Guatemala, Price serves as lead statistician on a project in which she analyzes documents from the National Police Archive. She helped her colleagues prepare evidence for high-profile court cases involving Guatemalan officials implicated in kidnappings. By rigorously analyzing data from government records, Price and her colleagues revealed clear links between the officials and the crimes. In Syria, she was lead statistician and author on three reports commissioned by the United Nations Office of the High Commissioner for Human Rights (OHCHR).

"I'd like to think that many statisticians and data scientists would do that kind of work if they had the chance," she says. "But it can be difficult to find the right opportunities. Doing pro bono work is a lovely idea, but there are limits to what you can accomplish by volunteering a few hours on nights and weekends. Many projects require full-time commitment."

Price hopes to see an increase in "formal opportunities" for data scientists to work on noncommercial, socially relevant projects. "Right now, there are very few organizations hiring full-time data scientists for social justice. I'm hoping that will change over the next 10 to 15 years."

Structuring Opportunities for Philanthropy

In many ways, DataKind is a harbinger of the future that Price envisions. DataKind is nonprofit that connects socially minded data scientists with organizations working to address critical humanitarian issues. “We’re dedicated to tackling the world’s greatest problems with data science,” says Jake Porway, DataKind’s founder and executive director. “We connect people whose day jobs are on Wall Street or in Silicon Valley with mission-driven organizations that can use data to make a positive impact on the world.”

DataKind’s programs range from short-term engagements done over a weekend to long-term, multimonth projects. All programs bring together data scientists and social-change organizations to collaborate on meaningful projects that move the needle on humanitarian challenges.

For example, when data scientists at Teradata were looking for new and improved ways to apply their skills to philanthropy, they teamed up with DataKind. The two organizations co-hosted a weekend “DataDive” that provided an opportunity for data scientists from DataKind and Teradata to work collaboratively with nonprofits and humanitarian organizations such as iCouldBe, HURIDOCS, GlobalGiving, and the Cultural Data Project on a wide range of data challenges, from improving an online mentoring program for at-risk youth to tracking human rights cases in Europe.

“One thing we found is there is no lack of demand for these services. We have over 200 organizations that have submitted applications to receive some sort of data science services,” Porway says. “On the other side, we should mention, we have more than 5,000 people who have signed up to volunteer. There is demand on both sides.”

In many instances, the challenge is combining or integrating data from disparate sources. In London, DataKind UK, one of the organization’s six chapters worldwide, helped St Mungo’s Broadway, a charity that helps people deal with issues leading to homelessness, link its data with data from Citizens Advice, a national charity providing free information on civil matters to the public. Linking the data yielded a trove of new insights that made it easier for St Mungo’s Broadway to predict which clients were more likely to benefit from its support.

In India, DataKind works with Simpa Networks, a venture-backed technology company in India that sells solar-as-a-service to energy-poor households and small businesses. Simpa's mission is making sustainable energy "radically affordable" to the 1.6 billion people at the "base of the pyramid" who currently lack access to affordable electricity.

In a six-month project financially underwritten by MasterCard, a team of DataKind volunteers is using Simpa Networks' historical data on customer payment behavior to predict which new applicants are likely to be a good fit for its model. That will enable Simpa Networks to best serve its customers and better assess new customers to offer the most appropriate services and support.

"Our goal is offering energy services to everyone, which includes customers who otherwise would be 'unbankable' according to mainstream financial institutions," says Paul Needham, Simpa Networks' chairman and CEO.

Data analytics plays a major role in supporting Simpa's ambitious mission. "Customer usage and payment behaviors are constantly tracked, and the data is fed into our proprietary credit-scoring model. That helps us get smarter about selecting customers and allows us to take risks on rural farmers that some banks would be uncomfortable financing," Needham says.

The energy situation is especially dire in India, where 75 million families have no access to electricity, and enormous sums are spent on unclean fuels such as kerosene for lanterns. "The good news is that effective decentralized energy solutions already exist. Solar photovoltaic solutions such as solar home systems can be sized appropriately to meet the energy needs of rural households and small businesses," Needham says. With data analytics, Simpa can make the case for loaning money that can be applied to clean-energy systems.

"Having learned from our past impact evaluation results, we have sufficient evidence to support the fact that Simpa's clean energy service will significantly reduce the time needed to conduct farming work, household chores, cooking, and cleaning," Needham says. "We anticipate that overall health standards will improve in these households due to the improved quality of light and will encourage the move away from kerosene and other hazardous forms of energy usage. In our midline impact evaluation study, we have seen that 80 percent of customers surveyed suffered eye irritation due to smoke; after Simpa's intervention, this figure dropped to 28 percent. Similarly, 10 percent of customers surveyed experienced fire accidents; after Simpa's intervention, this figure dropped to zero. We also believe that shop owners in these energy-poor areas will be able to stay open longer hours, which is likely to increase their sales and overall productivity."

Telling the Story with Analytics

DataKind also has collaborated with Crisis Text Line (CTL), a free service providing emotional support and information for anyone in a crisis. The process for accessing help is simple and efficient: people in need of help send texts to CTL, and trained specialists respond to the texts with support, counseling, information, and referrals.

CTL is staffed by volunteers, and like all volunteer organizations, its resources are constrained. CTL's mission is providing potentially life-saving support services for people in need—but it's also critical for the organization to avoid overwhelming its volunteers.

"Repeat callers have posed a challenge for crisis centers since the 1970s," explains Bob Filbin, CTL's chief data scientist. "When you read through the academic literature, you see that repeat callers are a big difficulty for crisis centers."

It's not that CTL's counselors don't want to help everyone who texts them—it's just that some people who contact CTL need a rapid intervention to avert a tragedy. The hard part is figuring out which people are experiencing acute, short-term crises requiring immediate attention and which people are suffering from less acute problems that can be dealt with over a slightly longer time frame.

After analyzing data from thousands of texts and examining patterns of usage from academic literature, Filbin and his colleagues were able to make suggestions for managing the problem of repeat texters. "We realized that our counselors were spending 34 percent of their time with 3 percent of our texters. By rolling out new policies and new technical products, we were able to reduce the portion of time our counselors spent with repeat texters from 34 percent to 8 percent. It was a huge win for us because it allowed more people to use the service."

In addition to freeing up more time for volunteers to interact with people experiencing acute problems, CTL was able to improve service for the repeat texters by guiding them toward helpful long-term resources.

Using data analysis to boost CTL's ability to deliver potentially life-saving services to people in need is especially gratifying, Filbin says. "It's very exciting when we can use data to overturn existing assumptions or drive meaningful change through an organization. Bringing data to bear on the problem, measuring our progress, and evaluating the effectiveness of our policies and products—it all makes an enormous difference."

From Filbin's perspective, it all comes down to good storytelling. "Data is only valuable when people act on it. Framing the data in terms of saving time was an emotional trigger that helped people understand its value," he says. "By reducing the conversation minutes with repeat texters from 34 percent down to 8 percent, we suddenly saved a quarter of our volunteers' time. That's a powerful story."

The idea of using data as a tool for storytelling is a recurring theme among data scientists working in philanthropic organizations. Most of the data scientists interviewed for this report mentioned storytelling as an important output of their work. Essentially, a good story makes it easier for managers and executives to make decisions and to take action on the insights generated by the data science team.

Data as a Pillar of Modern Democracy

Emma Mulqueeny, who writes a popular blog on data science, sees a larger trend evolving. Mulqueeny is the founder of Rewired State and Young Rewired State, a commissioner for the Speaker's Commission on Digital Democracy in the UK, a Google Fellow, and a digital tech entrepreneur. Earlier in her career, while working for the UK government on digital communication strategies, she noticed a sea change in the way people responded to statements made by government officials.

"There was a huge scandal over expenses," she recalls, "and suddenly it seemed as though everybody lost their trust in everything the government was saying. Suddenly, everybody wanted facts. They didn't want your interpretation of facts; they just wanted facts."

Government officials were aghast. But as a result of the scandal, efforts were made to increase transparency. Data that previously had been off-limits or difficult to obtain was made available to the public. Data.Gov.UK and Data.Gov, both launched in 2009, are prime examples of the "open data" trend in democratic societies. It's almost as if governments are saying, "You want data? We got your data right here!"

Mulqueeny sees those kinds of efforts as steps in the right direction, but she's adamant about the need for doing more. "The way people are operating online, the way they're learning, sharing, and influencing is very much dependent on what's pushed into their space," she says. "We're all familiar with Google's machine learning algorithms. You search for 'blue trousers' and suddenly everywhere you go after that, you're seeing little adverts for blue trousers and other items to buy. Marketers know how to mark up data so it can be used for marketing."

Democratically elected governments, on the other hand, are still struggling with data. "Let's say you feel passionate about chickens. If the information is properly marked up, you are more likely to see when the government is discussing matters related to chickens," Mulqueeny says. "Now let's say the government decides to outlaw chickens in London. If the information is marked up, you'll probably see it. But if it's not properly marked up, you won't. Which means that you won't find out the government is considering banning chickens until you read about it in a newspaper or some other media outlet."

From Mulqeeny's perspective, real democracy requires more than just sharing data—it requires making sure that data is properly tagged, annotated, and presented to people when they are online. In effect, she is raising the bar for governments and saying they need to be as good as—or better than—online marketers when it comes to serving up information.

"People have expectations that their interests will be served in the space in which they choose to be online and that they will find out what's happening when they are online," she says. "That's the heart of everything at the moment."

No Strings Attached, but Plenty of Data

For as long as most of us can remember, charities have worked like this: people or organizations make donations to charities, and charities distribute the donations to people or organizations that need support. Recently, and for a variety of legitimate reasons, the validity of that model has been called into question. As a result, new models for charitable giving have emerged.

GiveDirectly is an organization that channels donations directly to the extreme poor in Kenya and Uganda. The money is distributed via mobile phones, which makes it relatively easy to keep precise digital records of who's getting what from whom. GiveDirectly's model was inspired by programs initiated by the Mexican government in the 1990s. Those programs showed that direct cash transfers to poor people were often more helpful than benefits that were distributed indirectly.

The "secret formula" behind GiveDirectly's success is scientific discipline. Two of the group's cofounders, Michael Faye and Paul Niehaus, describe the differences between GiveDirectly and traditional charities:

> From the very beginning, we took a principled stand and decided to run randomized trials, which are the gold standard for discovering whether something works or doesn't. Some people can always find excuses for not running randomized trials. They will say they're too expensive or they take too much time or they might jeopardize the business model.
>
> Our response to those excuses is to ask, 'Would you buy drugs from a pharmaceutical company that doesn't run randomized trials of its drugs?' Of course you wouldn't. So why would you donate money to a charity that doesn't test its programs?

Although GiveDirectly distributes donations with no strings attached, its approach is the antithesis of just throwing money at problems. True to their roots as trained economists, Faye and Niehaus have devised an excruciatingly detailed system for making sure donations are used properly. After choosing a village or area to receive donations, GiveDirectly sends a team to the location. The team goes from house to house, creating a highly detailed, data-rich map of the location. Then a second team is dispatched to register local inhabitants and verify the data assembled by the first team.

No money is distributed until a third team has verified the information provided by the first two teams, and even then, only token payments are made to make absolutely sure the money winds up in the right hands. When all the tests are complete, additional payments are authorized, flowing directly to the local residents via mobile banking or other forms of digital cash transfer.

It's a rigorous approach, but it's an approach that can be scaled and audited easily. Transparency, redundancy, and continual analysis are crucial to the success of the overall process. "We think it's the future of charity in the developing world. In fact, we don't see ourselves as a charity—we see ourselves as service providers," Faye says.

GiveDirectly draws a distinction between data and evidence. "We emphasize that understanding impact requires not just knowing what happened, but what knowing *would* have happened if we hadn't intervened," Faye says. "We do that with randomized controlled trials."

Faye and Niehaus urge donors to ask basic questions of all charitable organizations:

- Where exactly does a donated dollar go? Who are the beneficiaries and how much money ultimately winds up in their hands?
- Beyond data alone, do the organizations have evidence showing the impact of their interventions?
- Are the organizations doing more good per dollar than the poor could do by themselves?

Collaboration Is Fundamental

When the New York City Department of Health and Mental Hygiene (DOHMH) realized that restaurant reviews posted on Yelp could be a source of valuable information in the ongoing battle to prevent foodborne illnesses, the department reached out to Yelp and to data scientists at Columbia University for help.

Over a nine-month period, roughly 294,000 Yelp reviews were screened by software that had been "trained" to look for potential cases of foodborne disease. According to an article posted on the Centers for Disease Control and Prevention (CDC) website, "the software flagged 893 reviews for evaluation by an epidemiologist, resulting in the identification of 468 reviews that were consistent with recent or potentially recent foodborne illness."

The article notes that only 3 percent of flagged reviews described events that had been reported to the health department. While the absolute numbers involved were relatively small, the project represents a major victory for data science.

Expending all of that effort to identify a handful of potentially dangerous restaurants in New York City might not seem like a big deal, but imagine scaling the process and offering it to every health department in the world.

"Data is everywhere now, more so than ever before in history," says Luis Gravano, a professor of computer science at Columbia University who worked with the health department on the Yelp project. "Regular people now are leaving a rich trail of incredibly valuable information, through the content they post online and via their mobile devices." Increasingly, data that people generate over the course of their daily lives is picked up by sensors. That kind of passively generated data is "less explicit, but also potentially quite valuable," Gravano says.

The data generated by "regular people" represents a unique opportunity for data scientists. "Collectively, the data is a great resource for all of us who analyze data," he says. "But the challenge is finding the gold nuggets of information in these mountains of data."

Dr. Sharon Balter, an epidemiologist at the health department, says data science was the key to finding the important pieces of information hidden in the reviews. "The team from Columbia helped us focus on the small number of restaurant reviews that might indicate real problems. The challenge is sifting through thousands of reviews. We don't have the resources to investigate every one of them," Balter says. "The algorithms developed by the Columbia team helped us determine which leads to investigate, and that was incredibly helpful."

Here's how the process worked, according to the CDC article:

> Beginning in April 2012, Yelp provided DOHMH with a private data feed of New York City restaurant reviews. The feed provided data publicly available on the website but in an XML format, and text classification programs were trained to automatically analyze reviews. For this pilot project, a narrow set of criteria were chosen to identify those reviews with a high likelihood of describing food-

> borne illness. Reviews were assessed retrospectively, using the following criteria: 1) presence of the keywords "sick," "vomit," "diarrhea," or "food poisoning" in contexts denoting foodborne illness; 2) two or more persons reported ill; and 3) an incubation period ≥10 hours.
>
> Ten hours was chosen because most foodborne illnesses are not caused by toxins but rather by organisms with an incubation period of ≥10 hours (*1*). Data mining software was used to train the text classification programs (*2*). A foodborne disease epidemiologist manually examined output results to determine whether reviews selected by text classification met the criteria for inclusion, and programs with the highest accuracy rate were incorporated into the final software used for the pilot project to analyze reviews prospectively.
>
> The software program downloaded weekly data and provided the date of the restaurant review, a link to the review, the full review text, establishment name, establishment address, and scores for each of three outbreak criteria (i.e., keywords, number of persons ill, and incubation period), plus an average of the three criteria. Scores for individual criteria ranged from 0 to 1, with a score closer to 1 indicating the review likely met the score criteria.
>
> Reviews submitted to Yelp during July 1, 2012–March 31, 2013 were analyzed. All reviews with an average review score of ≥0.5 were evaluated by a foodborne disease epidemiologist. Because the average review score was calculated by averaging the individual criteria scores, reviews could receive an average score of ≥0.5 without meeting all individual criteria.
>
> Reviews with an average review score of ≥0.5 were evaluated for the following three criteria: 1) consistent with foodborne illness occurring after a meal, rather than an alternative explanation for the illness keyword; 2) meal date within 4 weeks of review (or no meal date provided); 3) two or more persons ill or a single person with symptoms of scombroid poisoning or severe neurologic illness. Reviews that met all three of these criteria were then investigated further by DOHMH. In addition, reviews were investigated further if manual checking identified multiple reviews within 1 week that described recent foodborne illness at the same restaurant.

Gravano and Balter agree that the availability of "nontraditional" data was critical to the success of their endeavor. "We're no longer relying solely on traditional sources of data to generate useful insights," Gravano says. As a result, groups of people that were previously "uncounted" can now benefit from the work of data scientists. "We're setting up an infrastructure that will make those kinds of projects more routine. Our hope, moving forward, is that our work will become a continuous process and that we will continually refine our algorithms and machine learning tools," he says.

Recently, another group of researchers at Columbia used machine learning tools to better understand and predict preterm births, a healthcare issue affecting 12–13 percent of infants born in the US. That study relied on a clinical-trial data set collected by the National Institute of Child Health and Human Development (NICHD) and the Maternal-Fetal Medicine Units Network (MFMU).

Conclusion

Most of the sources interviewed for this report highlighted the multidisciplinary and inherently collaborative nature of data science, and several expressed the belief that at some level, most data scientists see their roles as beneficial to society. That said, there still appears to be a clear need for organizations that provide structures and processes for enabling the collaboration and teamwork necessary for effective pro bono data science projects. In other words, doing data science for the good of humankind requires more than good intentions—it requires practical frameworks, networks of qualified people, and sources of funding.

Applying data science principles to solve social problems and improve the lives of ordinary people seems like a logical idea, but it is by no means a given. Using data science to elevate the human condition won't happen by accident; groups of people will have to envision it, develop the routine processes and underlying infrastructures required to make it practical, and then commit the time and energy necessary to make it all work.

Columbia University has taken a step in the right direction by launching the Data Science Institute, an interdisciplinary learning and research facility with dedicated faculty and six specialized centers: Cybersecurity, Financial and Business Analytics, Foundations of Data Science, Health Analytics, New Media, and Smart Cities.

"Whatever good you want to do in the world, the data is there to make it possible," says Kathleen McKeown, director of the Data Science Institute. "Whether it's finding new and unexpected treatments for disease or techniques for predicting the impact of natural disasters, data science has tremendous potential to benefit society."

Author's Note

Crisis Text Line is looking for volunteers. If you are interested in becoming a crisis counselor, please visit www.crisistextline.org/join-our-efforts/volunteer/ for more information.

DataKind is also seeking volunteers. If you're a data scientist looking to use your skills to give back, you can apply to volunteer with DataKind at www.datakind.org/getinvolved/ or learn more at an upcoming event in your area: http://www.datakind.org/community-events/.