1

A Higher Calling

It is a great time for data science! The *Economist* proudly proclaims that data is "the world's most valuable resource,"[1] and Hal Varian and Tom Davenport[2] have variously called statistics and data science "the sexiest job of the twentieth century." In searching the web for the term *data scientist*, we find the following definition, *"'Data Scientist' means a professional who uses scientific methods to liberate and create meaning from raw data.*"[3] Similar definitions have been offered for statisticians and data analysts.[4] Yet we believe the work is more involved and requires skills far beyond those needed to create meaning from raw data.

This book expands and clarifies what it takes to succeed in this job, within the organizational ecosystem in which it takes place. It builds on years of experience in a wide range of organizations, all over the world. Our goal is to share this experience and some retrospective insights learned in doing real work. Specifically, we propose that the real work of data scientists and statisticians involves helping people make better decisions on the important issues in the near term and building stronger organizations and capabilities in the long term. By "people" we mean, among others, managers in organizations and professionals in service and production industries. This perspective is also relevant to educators in schools and colleges and researchers in laboratories and academic institutions. It is a far higher, and more demanding, calling. For example, you don't get to contribute on the really important decisions unless you're trusted.

Thus, the real work requires total involvement: helping to formulate the problems and opportunities in crisp business or scientific terms; understanding which data to consider and the strengths and limitations in the data; determining when new data is needed; dealing with quality issues; using the data to reduce uncertainty; making clear where the data ends and intuition must take over; presenting results in simple, powerful ways; recognizing that all important decisions involve political realities; working with others; and supporting decisions in practice. This real work is not taught enough in statistics or data science courses.

[1] Cover of the May 6, 2017, issue.

[2] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century (Davenport and Patil 2012).

[3] http://www.datascienceassn.org/code-of-conduct.html (Data Science Association 2018).

[4] Herein, we use the terms *data science*, *data analytics*, and *statistics* interchangeably, fully recognizing that many people see fine distinctions. But these are not central to this book.

The Real Work of Data Science: Turning Data into Information, Better Decisions, and Stronger Organizations, First Edition. Ron S. Kenett and Thomas C. Redman.

Companion website: www.wiley.com/go/kenett-redman/datascience

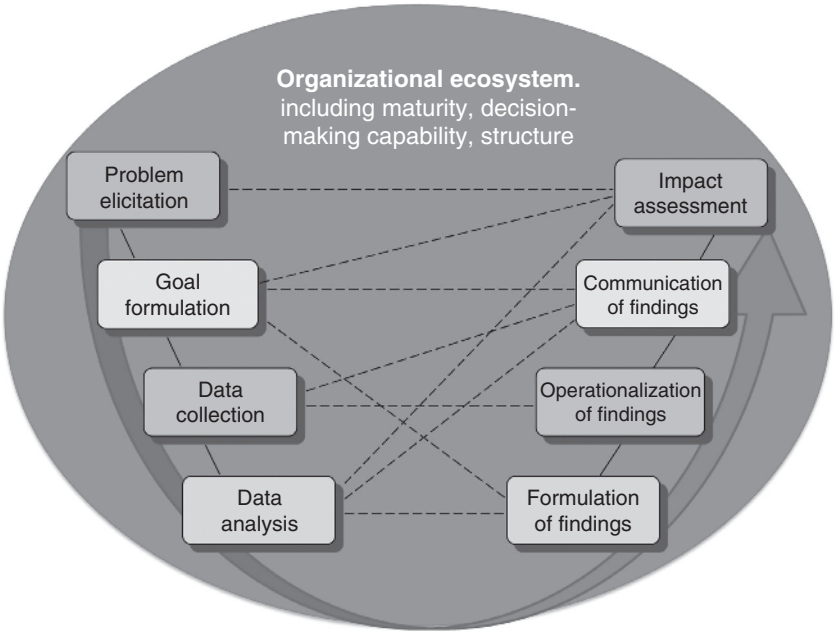

Figure 1.1 The life-cycle view of data analytics, in the context of the organizational ecosystem in which the work takes place.

The unpleasant reality is that many/most companies derive only a fraction of the value that their data, data science, and statistics offer (see, for example, Henke et al. 2016). Data scientists and their managers, including chief analytics officers (CAOs), chief data scientists, heads of data science, and other professionals who employ data scientists,[5] must learn how to address the barriers that get in the way. Thus, the real work also involves raising everyone's ability to conduct simple analyses and understand more complex ones, understand the power of data, understand variation, and integrate data and their intuitions; putting the right data scientists and statisticians in the right spots; educating senior leadership on the power of data; helping them become good consumers of data science; teaching them their roles in advancing the effort; and creating the organizational structures needed to do all of the above effectively and (reasonably) efficiently. This is what this book is about.

Providing the added value we are talking about requires a wide perspective. Figure 1.1 presents the life cycle of data analytics in the context of an organization aiming to profit from data science (adapted from Kenett 2015). As the figure illustrates, the work is highly iterative (for more on this process, see Box 1997).

The Life-Cycle View

The life-cycle view is designed to help data scientists help decision-makers. Let's consider each step of the cycle in turn.

[5] We recognize once again that many people see distinctions in these roles as well, but we will also use them interchangeably, as the distinctions are not central to this book.

Problem Elicitation: Understand the Problem
Observe what happens when you go to a dentist: you give a dentist a hint about your symptoms, you are placed in the chair, the dentist looks into your mouth, diagnoses and (hopefully) solves the problem, and tells you when to come back, all in less than an hour.

The seasoned data scientist knows better. We describe these data scientists in Chapter 2. They listen carefully and ask probing questions, keeping the customers (e.g. the decision-makers) focused and obtaining the relevant details to understand their needs. It may be an operations manager experiencing huge costs because of rework, a marketing manager trying to enter a new market, or a human resources (HR) manager who wants to reduce employee turnover. The experienced data scientist also reads the customer's body language for unspoken clues: does the customer have a hidden agenda, is he or she trying to make someone else look bad or build support for a political squabble?

Like many others, we can't stress this enough – you simply must understand the real problem if you hope to help solve it. The quality of analytic work depends on it (Kenett and Shmueli 2016a). More in Chapters 3 and 4.

Goal Formulation: Clarify the Short-term and Long-term Goals
Don't expect that the decision-maker has clearly formulated the problem. Bill Hunter, a famous statistician from the University of Wisconsin in Madison, tells the story of two chemists who sought his advice. When he asked them to describe their problem, they entered a lengthy discussion that led them to reformulate their problem. This one was much simpler, and they did not need further help from Bill. They left his office after thanking him profusely (Hunter 1979). While Bill's role may seem small, it was essential!

The main point is that a full understanding of the problem requires a full understanding of the context in which it occurs, including the overarching goal. More in Chapter 4.

Data Collection: Identify Relevant Data Sources and Collect the Data
Cobb and Moore (1997) point out that "Statistics requires a different kind of thinking, because data are not just numbers, they are numbers with a context." The context helps identify relevant data sources and their interpretation.

To illustrate, consider this story from Denmark from Kenett and Thyregod (2006). It involves an exercise in a fourth-grade textbook and shows the importance of context and how numbers turn into data. In this exercise, the numbers presented in Figure 1.2 record the number of ice creams sold each day, without any indication of the actual day of the week. In July, it was very hot for nine consecutive days. Students were asked to (i) identify the hot days and (ii) determine which days were Sundays.

By itself, the graph just presents 31 numbers. But Danish schoolchildren know their parents are more inclined to offer ice cream on weekends and on hot days. With this context, it was easy for these young children to complete their assigned tasks.

Context is revealed where data is generated, from the shop floor, to the laboratory, to a social media setting. Data scientists must understand this context and identify the data relevant to the problem. More on this in Chapter 5.

Data Analysis: Use Descriptive, Explanatory, and Predictive Methods
This is the work of "creating meaning from data," "separating the signal from the noise," "turning data into information," and so forth. There are, of course, literally thousands of

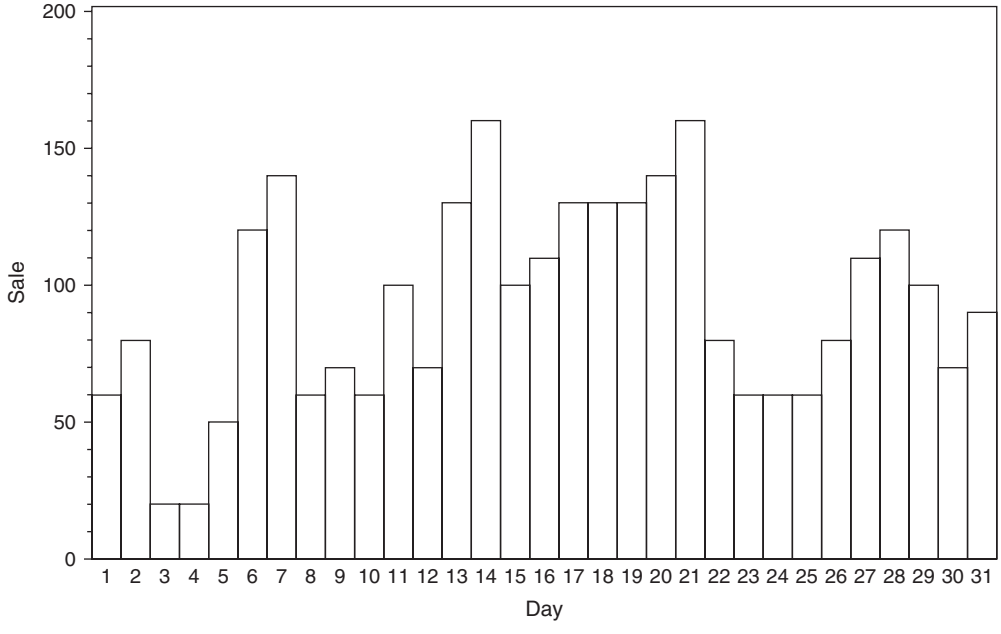

Figure 1.2 The number of ice creams sold in a Danish locality, by day in July.

examples. As one, consider eBay auctions. When you sell an item on eBay, you are asked to specify a "reserve price," a value you set to start the auction. If the final price does not exceed the reserve price, the auction does not transact. On eBay, sellers can choose to place a *public* reserve price that is visible to bidders or a *secret* reserve price (bidders only see that there is a reserve price but do not know its value).

Katkar and Reiley (2006) investigated the effect of this choice. Their data came from an experiment selling 25 identical pairs of Pokémon cards, where each card was auctioned twice, once with a public reserve price and once with a secret reserve price, and consists of complete information on all 50 auctions. They used linear regression and significance tests to quantify the effect, if any, of private/public reserve on the final price. They concluded that "a secret-reserve auction generates a $0.63 lower price on average," a simple statement everyone can understand.

We are less concerned with this work here, except for one critical area usually not well covered in data science training. The cold, brutal reality is that too much data is unfit for analysis (Nagle et al. 2017), and data scientists spend far more of their time on data quality issues than they do on analysis. High-quality data is critical for all analyses and especially so for cognitive technologies (Redman 2018b). So data scientists must deal with the issue. More in Chapter 6.

Formulation of Findings: State Results and Recommendations

Analytics produces outputs such as descriptive statistics, p-values, regression models, analysis of variance (ANOVA) tables, control charts, trees, forests, neural networks, dendrograms, and

so forth. Many are beyond the scope for decision-makers. So, it is essential that data scientists translate their results into language the decision-maker understands.

Further, data scientists must explore the implications of their results and, oftentimes, recommend specific courses of action. Said differently, data scientists cannot simply "throw results and recommendations over the wall." Rather, they must ensure that the decision-maker understands the findings in their proper context. Because many people are involved in important decisions, this may mean several distinct presentations and interactions with senior managers, middle managers, and knowledge workers. All may require different levels of detail, in different forms.

Concepts and notation from mathematical statistics turn many people off. Instead, well-thought-out graphical displays are the communication tools of choice. Findings that cannot be presented in a graph are probably not worth communicating. Keep graphs and slides simple, and keep the "ink-to-information ratio" low, avoiding fancy symbols etc. (Tufte 1997). For a simple example, see Chapter 7.

A great example of this involves an analyst who realized that senior decision-makers did not understand the technical terms associated with the network robustness problem they assigned him. So, he crystallized the problem and formulated his results using a well-known fairy tale: "The first thing we must decide is what kind of network we want: a 'baby bear network,' a 'mama bear network,' or a 'papa bear' network. Roughly this means …" Everyone got it.

While the actual decision is made by others, in the life-cycle model we expect the analysis to support a decision, even a tentative one, as the conclusion to this step.

Operationalization of Findings: Suggest Who, What, When, and How

The data scientist's job does not end with a decision. Rather, he or she should follow the data-based decisions into execution, helping define how results are put into practice (e.g. operational procedures), answering questions that are sure to come up, evaluating new data as it comes in, and advising on situations beyond the scope of the original analysis.

It is tempting to skip this step. But the value of data science only accrues when an analysis and decision are put into practice, not before. More in Chapter 8.

Communication of Findings: Communicate Findings, Decisions, and Their Implications to Stakeholders

Until now, a relatively small number of people have been involved in the work we referred to. But important decisions can impact thousands, even millions, of people. At this step, findings must be communicated to all who may be impacted, a much wider audience than those involved in the decision. While the lion's share of this work is the purview of the decision-maker, the data scientist must play an active role in support.

Impact Assessment: Plan and Deploy an Assessment Strategy

Although it is beyond the scope of helping decision-makers per se, data scientists should assess their impact. Wherever possible, get hard numbers. Of course, as the vignette featuring Bill Hunter illustrates, this is not always possible. And even when you can get hard numbers, solicit feedback from decision-makers.

Then be brutally honest in assessing how you can do better next time. More in Chapter 9.

The Organizational Ecosystem

The work of data science takes place in complex organizational settings, which can both promote and limit its effectiveness. Sometimes simultaneously. Data scientists and CAOs must be aware of and, over time, improve several components of the overall "organizational ecosystem."

The term *data-driven* has invaded the lexicon. One sees extravagant claims for data-driven marketing, data-driven HR, and data-driven technologies. Beyond the hype, and more deeply, is a powerful core that leads to better decisions and stronger organizations. At that core, the more data-driven the organization, the more demanding decision-makers are of data scientists, the more seriously they take sophisticated analyses, and the more they invest in high-quality data, clear decision rights, and the decision-making capabilities. Thus, smart data scientists and CAOs invest considerable time in educating themselves and decision-makers at all levels about this powerful concept and working together to advance it across the organizations.

We will discuss what it means to be data-driven in some detail in Chapter 10. It will come as no surprise that bias, in any form, is diametrically opposed to data-driven decision-making. Step one for data scientists is to remove bias from their own work – a subject we will take up in Chapter 11. The focus of Chapters 12–14 is education. First, Chapter 12 advises data scientists to start with the basics with their peers and other decision-makers. Chapter 13 takes a slightly different tack. It recognizes that demanding customers (e.g. decision-makers) will do as much to advance a data-driven culture and data science as anything. So, the chapter provides a list of questions to help decision-makers know what to ask.

With big data, AI, security concerns, the General Data Protection Regulation (GDPR), digitization, and so much more all over the news, it is hard for senior leaders to see the data space in perspective. Chapter 14 considers the big picture, advising CAOs to develop a wide and deep perspective on the data space and to help their organization's most senior leaders understand the risks and opportunities.

Organizational Structure

The unfortunate truth is that where data scientists sit in an organization dictates what they can do. For example, a data scientist sitting in the maintenance department may be denied access to relevant data from the operations department, for no other reason than the heads of each are competing for the same promotion. While data scientists may like to believe they are above it all, there is no escaping politics. Better for data scientists and CAOs to embrace this reality and strive to get into the right spots. More on this in Chapter 15.

Organizational Maturity

Finally, organizations have different needs of data science, based on their maturity. These run the gamut from those in fire-fighting mode, with basic, immediate needs, to learning organizations with needs for deep, penetrating analyses and predictions. More in Chapters 16 and 17.

Once Again, Our Goal

With this background, our goal is to help data scientists and CAOs become more effective. This means helping data scientists contribute to better decisions and CAOs to stronger organizations, without being too strict about it. We have organized the material as 18 narrowly

focused chapters, loosely tied to the life cycle of analytics and the ecosystem. We have also included some materials not directly tied to either. For example, the next chapter explores the difference between a good data scientist and a great one. Each chapter is short and to the point.

Overall, this book presents a wide-angle view of the real work of data science. Our goal is to expand your perspective, trigger deep thinking, and help you develop insights that make you more effective as a developer, participant, or consumer of the data science experience.