IN THIS CHAPTER

- Deploying data science methods across various industries
- Piecing together the core data science components
- Identifying viable data science solutions to business challenges
- Exploring data science career alternatives

Chapter 1
Wrapping Your Head Around Data Science

For over a decade now, *everyone* has been absolutely deluged by data. It's coming from every computer, every mobile device, every camera, and every imaginable sensor — and now it's even coming from watches and other wearable technologies. Data is generated in every social media interaction we humans make, every file we save, every picture we take, and every query we submit; data is even generated when we do something as simple as ask a favorite search engine for directions to the closest ice cream shop.

If you're anything like I was, you may have wondered, "What's the point of all this data? Why use valuable resources to generate and collect it?" Although even just two decades ago, no one was in a position to make much use of most of the data that's generated, the tides today have definitely turned. Specialists known as *data engineers* are constantly finding innovative and powerful new ways to capture, collate, and condense unimaginably massive volumes of data. Other specialists known as *data scientists* are leading change by deriving valuable and actionable insights from that data.

COPYRIGHTED MATERIAL

In its truest form, data science represents the optimization of processes and resources. Data science produces *data insights* — actionable, data-informed conclusions or predictions that you can use to understand and improve your business, your investments, your health, and even your lifestyle and social life. Using data science insights is like being able to see in the dark. For any goal or pursuit you can imagine, you can find data science methods to help you predict the most direct route from where you are to where you want to be — and to anticipate every pothole in the road between both places.

In this chapter, I explain the difference between data science and data engineering.

Seeing Who Can Make Use of Data Science

The terms *data science* and *data engineering* are often misused and confused, so let me start off by clarifying that these two fields are, in fact, separate and distinct domains of expertise. *Data science* is the computational science of extracting meaningful insights from raw data and then effectively communicating those insights to generate value. *Data engineering*, on the other hand, is an engineering domain that's dedicated to building and maintaining systems that overcome data processing bottlenecks and data handling problems for applications that consume, process, and store large volumes, varieties, and velocities of data.

In both data science and data engineering, you commonly work with the following types of data:

- **Structured data:** Data that is stored, processed, and manipulated in a traditional *relational database management system* (RDBMS). An example of this type of data can be seen in the tabular schema of rows and columns you'd commonly encounter when working with corporate databases.
- **Unstructured data:** Data that is commonly generated from human activities and doesn't fit into a structured database format. Examples of unstructured data are data that comprises email documents, Microsoft Word documents or audio or video files.

- **Semistructured data:** Data that doesn't fit into a structured database system but is nonetheless organizable by tags that are useful for creating a form of order and hierarchy in the data. XML and JSON files are examples of data that comes in semistructured form.

In the past, only large tech companies with massive funding had the skills and computing resources required to implement data science methodologies to optimize and improve their business, but that hasn't been the case for quite a while now. The proliferation of data has created a demand for insights, and this demand is embedded in many aspects of modern culture — from the Uber passenger who expects the driver to show up exactly at the time and location predicted by the Uber app to the online shopper who expects the Amazon platform to recommend the best product alternatives for comparing similar goods before making a purchase. Data and the need for data-informed insights are ubiquitous. Because organizations of all sizes are beginning to recognize that they're immersed in a sink-or-swim, data-driven, competitive environment, data know-how has emerged as a core and requisite function in almost every line of business.

What does this mean for the average knowledge worker? It means that everyday employees are increasingly expected to support a progressively advancing set of technological and data requirements. Why? Because almost all industries are reliant on data technologies and the insights they spur. Consequently, many people are in continuous need of upgrading their data skills, or else they face the real possibility of being replaced by a more data-savvy employee.

The good news is that upgrading data skills doesn't usually require people to go back to college or earn a university degree in statistics, computer science, or data science. The bad news is that, even with professional training or self-teaching, it always takes extra work to stay industry-relevant and tech-savvy. In this respect, the data revolution isn't so different from any other change that has hit industry in the past. The fact is, in order to stay relevant, you need to take the time and effort to acquire the skills that keep you current. When you're learning how to do data science, you can take some courses, educate yourself using online resources, read books like this one, and attend events where you can learn what you need to know to stay on top of the game.

Who can use data science? You can. Your organization can. Your employer can. Anyone who has a bit of understanding and training can begin using data insights to improve their lives, their careers, and the well-being of their businesses. Data science represents a change in the way you approach the world. When determining outcomes, people once used to make their best guess, act on that guess, and then hope for the desired result. With data insights, however, people now have access to the predictive vision that they need to truly drive change and achieve the results they want.

Here are some examples of ways you can use data insights to make the world, and your company, a better place:

- **Develop key performance indicators (KPIs) for your business systems.** Use KPIs to track performance and optimize the return on investment (ROI) for measurable business activities.
- **Develop your marketing strategy.** Use data insights and predictive analytics to identify marketing strategies that work, eliminate underperforming efforts, and test new marketing strategies.
- **Keep communities safe.** Predictive policing applications help law enforcement personnel predict and prevent local criminal activities.
- **Help make the world a better place for those less fortunate.** Data scientists in developing nations are using social data, mobile data, and data from websites to generate real-time analytics that improve the effectiveness of humanitarian responses to disasters, epidemics, food scarcity issues, and more.

Inspecting the Pieces of the Data Science Puzzle

To practice data science, in the true meaning of the term, you need the analytical know-how of math and statistics, the coding skills necessary to work with data, and an area of subject matter expertise. Without this expertise, you may as well call yourself a mathematician or a statistician. Similarly, a programmer without subject matter expertise and analytical know-how may better be considered a software engineer or developer, but not a data scientist.

The need for data-informed business and product strategy has been increasing exponentially for about a decade now, forcing all business sectors and industries to adopt a data science approach. As such, different flavors of data science have emerged. The following are just a few titles under which experts of every discipline are required to know and regularly do data science:

- Clinical biostatistician
- Data and tech policy analyst
- Data scientist–geospatial and agriculture analyst
- Data scientist–health care
- Digital banking product owner
- Director of data science–advertising technology
- Geotechnical data scientist
- Global channel ops–data excellence lead

Nowadays, it's almost impossible to differentiate between a proper data scientist and a subject matter expert (SME) whose success depends heavily on their ability to use data science to generate insights. Looking at a person's job title may or may not be helpful, simply because many roles are titled *data scientist* when they may as well be labeled data strategist or product manager, based on the actual requirements. In addition, many knowledge workers are doing daily data science and not working under the title of *data scientist.* It's an overhyped, often misleading label that's not always helpful if you're trying to find out what a data scientist does by looking at online job boards.

To shed some light, in the following sections I spell out the key components that are part of any data science role, regardless of whether that role is assigned the *data scientist* label.

Collecting, querying, and consuming data

Data engineers have the job of capturing and collating large volumes of structured, unstructured, and semistructured *big data* (an outdated term that's used to describe data that exceeds the processing capacity of conventional database systems because it's too big, it moves too fast, or it lacks the structural requirements of traditional database architectures).

REMEMBER

Data engineering tasks are separate from the work that's performed in data science, which focuses more on analysis, prediction, and visualization. Despite this distinction, whenever data scientists collect, query, and consume data during the analysis process, they perform work similar to that of the data engineer (the role I tell you about earlier in this chapter).

Although valuable insights can be generated from a single data source, often the combination of several relevant sources delivers the contextual information required to drive better data-informed decisions. A data scientist can work from several datasets that are stored in a single database, or even in several different data storage environments. At other times, source data is stored and processed on a cloud-based platform built by software and data engineers.

No matter how the data is combined or where it's stored, if you're a data scientist, you almost always have to *query* data — in other words, write commands to extract relevant datasets from data storage systems. Most of the time, you use Structured Query Language (SQL) to query data. (Chapter 7 is all about SQL, so if the acronym scares you, jump ahead to that chapter now.)

Whether you're using a third-party application or doing custom analyses by using a programming language such as R or Python, you can choose from a number of universally accepted file formats:

- **Comma-separated values (CSV):** Almost every brand of desktop and web-based analysis application accepts this file type, as do commonly used scripting languages such as Python and R.
- **Script:** Most data scientists know how to use Python to analyze and visualize data. These script files end with the extension `.ply` or `.ipynb` (Python).
- **Application:** Excel is useful for quick-and-easy, spot-check analyses on small- to medium-size datasets. These application files have the `.xls` or `.xlsx` extension.

» **Web programming:** If you're building custom, web-based data visualizations, you may be working in D3.js — or data-driven documents, a JavaScript library for data visualization. When you work in D3.js, you use data to manipulate web-based documents using `.html`, `.svg`, and `.css` files.

Applying mathematical modeling to data science tasks

Data science relies heavily on a practitioner's math skills (and statistics skills, as described in the following section) precisely because these are the skills needed to understand your data and its significance. These skills are also valuable in data science because you can use them to carry out predictive forecasting, decision modeling, and hypotheses testing.

REMEMBER

Mathematics uses deterministic methods to form a *quantitative* (or *numerical*) description of the world; *statistics* is a form of science that's derived from mathematics, but it focuses on using a *stochastic* (probabilities) approach and inferential methods to form a quantitative description of the world. (I tell you more about math and statistics in Chapter 4.) Data scientists use mathematical methods to build decision models, generate approximations, and make predictions about the future. Chapter 4 presents many mathematical approaches that are useful when working in data science.

REMEMBER

In this book, I assume that you have a fairly solid skill set in basic math — you'll benefit if you've taken college-level calculus or even linear algebra. I try hard, however, to meet you where you are. I realize that you may be working based on a limited mathematical knowledge (advanced algebra or maybe business calculus), so I convey advanced mathematical concepts using a plain-language approach that's easy for everyone to understand.

Deriving insights from statistical methods

In data science, statistical methods are useful for better understanding your data's significance, for validating hypotheses, for simulating scenarios, and for making predictive forecasts of future events. Advanced statistical skills are somewhat rare, even among quantitative analysts, engineers, and scientists. If you want to go places in data science, though, take some time to get up to speed in a few basic statistical methods, like linear and logistic regression, Naïve Bayes classification, and time series analysis. (These methods are covered in Chapter 4.)

Coding, coding, coding — it's just part of the game

Coding is unavoidable when you're working in data science. You need to be able to write code so that you can instruct the computer on how to manipulate, analyze, and visualize your data. Programming languages like Python are important for writing scripts for data manipulation, analysis, and visualization. SQL, on the other hand, is useful for data querying.

Although coding is a requirement for data science, it doesn't have to be this big, scary *thing* that people make it out to be. Your coding can be as fancy and complex as you want it to be, but you can also take a rather simple approach. Although these skills are paramount to success, you can pretty easily learn enough coding to practice high-level data science. I've dedicated Chapters 6 and 7 to helping you get to know the basics of what's involved in getting started in Python and querying in SQL, respectively.

Applying data science to a subject area

Statisticians once exhibited some measure of obstinacy in accepting the significance of data science. Many statisticians have cried out, "Data science is nothing new — it's just another name for what we've been doing all along!" Although I can sympathize with their perspective, I'm forced to stand with the camp of data scientists who markedly declare that data science is separate, and definitely distinct, from the statistical approaches that comprise it.

My position on the unique nature of data science is based to some extent on the fact that data scientists often use computer languages not used in traditional statistics and take approaches derived from the field of mathematics. But the main point of distinction between statistics and data science is the need for subject matter expertise.

Because statisticians usually have only a limited amount of expertise in fields outside of statistics, they're almost always forced to consult with an SME to verify exactly what their findings mean and to determine the best direction in which to proceed. Data scientists, on the other hand, should have a strong subject matter expertise in the area in which they're working. Data scientists generate deep insights and then use their domain-specific expertise to understand exactly what those insights mean with respect to the area in which they're working.

Here are a few ways in which today's knowledge workers are coupling data science skills with their respective areas of expertise in order to amplify the results they generate:

- **Clinical informatics scientists** combine their health-care expertise with data science skills to produce personalized health-care treatment plans. They use health-care informatics to predict and preempt future health problems in at-risk patients.
- **Marketing data scientists** combine data science with marketing expertise to predict and preempt customer *churn* (the loss of customers from a product or service to that of a competitor's). They also optimize marketing strategies, build recommendation engines, and fine-tune marketing mix models.
- **Data journalists** *scrape* websites (extract data in bulk directly from the pages on a website) for fresh data in order to discover and report the latest breaking-news stories. (I talk more about data storytelling in Chapter 8.)
- **Directors of data science** bolster their technical project management capabilities with an added expertise in data science. Their work includes leading data projects and working to protect the profitability of the data projects for which they're responsible. They also act to ensure transparent communication between C-suite executives, business managers, and the data personnel on their team who actually do the implementation work.
- **Data product managers** supercharge their product management capabilities with the power of data science. They use data science to generate predictive insights that better inform decision-making around product design, development, launch, and strategy.

- **Machine learning engineers** combine software engineering superpowers with data science skills to build predictive applications. This is a classic data implementation role, more of which I discuss in Chapter 2.

Communicating data insights

As a data scientist, you must have sharp verbal communication skills. If a data scientist can't communicate, all the knowledge and insight in the world does *nothing* for the organization. Data scientists need to be able to explain data insights in a way that staff members can understand. Not only that, but data scientists need to be able to produce clear and meaningful data visualizations and written narratives. Most of the time, people need to see a concept for themselves in order to truly understand it. Data scientists must be creative and pragmatic in their means and methods of communication. (I cover the topics of data visualization and data-driven storytelling in much greater detail in Chapter 8.)