1

WHAT IS DATA SCIENCE?

Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets. Many of the elements of data science have been developed in related fields such as machine learning and data mining. In fact, the terms *data science*, *machine learning*, and *data mining* are often used interchangeably. The commonality across these disciplines is a focus on improving decision making through the analysis of data. However, although data science borrows from these other fields, it is broader in scope. Machine learning (ML) focuses on the design and evaluation of algorithms for extracting patterns from data. Data mining generally deals with the analysis of structured data and often implies an emphasis on commercial applications. Data science takes all of these considerations into account but also takes up other challenges,

such as the capturing, cleaning, and transforming of unstructured social media and web data; the use of big-data technologies to store and process big, unstructured data sets; and questions related to data ethics and regulation.

Using data science, we can extract different types of patterns. For example, we might want to extract patterns that help us to identify groups of customers exhibiting similar behavior and tastes. In business jargon, this task is known as *customer segmentation*, and in data science terminology it is called *clustering*. Alternatively, we might want to extract a pattern that identifies products that are frequently bought together, a process called *association-rule mining*. Or we might want to extract patterns that identify strange or abnormal events, such as fraudulent insurance claims, a process known as *anomaly* or *outlier detection*. Finally, we might want to identify patterns that help us to classify things. For example, the following rule illustrates what a classification pattern extracted from an email data set might look like: *If an email contains the phrase "Make money easily," it is likely to be a spam email.* Identifying these types of classification rules is known as *prediction*. The word *prediction* might seem an odd choice because the rule doesn't predict what will happen in the future: the email already is or isn't a spam email. So it is best to think of prediction patterns as predicting the missing value of an attribute rather than as predicting

If a human expert can easily create a pattern in his or her own mind, it is generally not worth the time and effort of using data science to “discover” it.

the future. In this example, we are predicting whether the email classification attribute should have the value "spam" or not.

Although we can use data science to extract different types of patterns, we always want the patterns to be both nonobvious and useful. The example email classification rule given in the previous paragraph is so simple and obvious that if it were the only rule extracted by a data science process, we would be disappointed. For example, this email classification rule checks only one attribute of an email: Does the email contain the phrase "make money easily"? If a human expert can easily create a pattern in his or her own mind, it is generally not worth the time and effort of using data science to "discover" it. In general, data science becomes useful when we have a large number of data examples and when the patterns are too complex for humans to discover and extract manually. As a lower bound, we can take a large number of data examples to be defined as more than a human expert can check easily. With regard to the complexity of the patterns, again, we can define it relative to human abilities. We humans are reasonably good at defining rules that check one, two, or even three attributes (also commonly referred to as *features* or *variables*), but when we go higher than three attributes, we can start to struggle to handle the interactions between them. By contrast, data science is often applied in contexts where we want to look for patterns among tens,

hundreds, thousands, and, in extreme cases, millions of attributes.

The patterns that we extract using data science are useful only if they give us insight into the problem that enables us to do something to help solve the problem. The phrase *actionable insight* is sometimes used in this context to describe what we want the extracted patterns to give us. The term *insight* highlights that the pattern should give us relevant information about the problem that isn't obvious. The term *actionable* highlights that the insight we get should also be something that we have the capacity to use in some way. For example, imagine we are working for a cell phone company that is trying to solve a customer *churn* problem—that is, too many customers are switching to other companies. One way data science might be used to address this problem is to extract patterns from the data about previous customers that allow us to identify current customers who are churn risks and then contact these customers and try to persuade them to stay with us. A pattern that enables us to identify likely churn customers is useful to us only if (*a*) the patterns identify the customers early enough that we have enough time to contact them before they churn and (*b*) our company is able to put a team in place to contact them. Both of these things are required in order for the company to be able to act on the insight the patterns give us.

A Brief History of Data Science

The term *data science* has a specific history dating back to the 1990s. However, the fields that it draws upon have a much longer history. One thread in this longer history is the history of data collection; another is the history of data analysis. In this section, we review the main developments in these threads and describe how and why they converged into the field of data science. Of necessity, this review introduces new terminology as we describe and name the important technical innovations as they arose. For each new term, we provide a brief explanation of its meaning; we return to many of these terms later in the book and provide a more detailed explanation of them. We begin with a history of data collection, then give a history of data analysis, and, finally, cover the development of data science.

A History of Data Gathering

The earliest methods for recording data may have been notches on sticks to mark the passing of the days or poles stuck in the ground to mark sunrise on the solstices. With the development of writing, however, our ability to record our experiences and the events in our world vastly increased the amount of data we collected. The earliest form of writing developed in Mesopotamia around 3200 BC and was used for commercial record keeping. This type

of record keeping captures what is known as *transactional data*. Transactional data include event information such as the sale of an item, the issuing of an invoice, the delivery of goods, credit card payment, insurance claims, and so on. *Nontransactional data*, such as demographic data, also have a long history. The earliest-known censuses took place in pharaonic Egypt around 3000 BC. The reason that early states put so much effort and resources into large data-collection operations was that these states needed to raise taxes and armies, thus proving Benjamin Franklin's claim that there are only two things certain in life: death and taxes.

In the past 150 years, the development of the electronic sensor, the digitization of data, and the invention of the computer have contributed to a massive increase in the amount of data that are collected and stored. A milestone in data collection and storage occurred in 1970 when Edgar F. Codd published a paper explaining the *relational data model*, which was revolutionary in terms of setting out how data were (at the time) stored, indexed, and retrieved from databases. The relational data model enabled users to extract data from a database using simple queries that defined what data the user wanted without requiring the user to worry about the underlying structure of the data or where they were physically stored. Codd's paper provided the foundation for modern databases and the development of *structured query language* (SQL), an

international standard for defining database queries. Relational databases store data in tables with a structure of one row per instance and one column per attribute. This structure is ideal for storing data because it can be decomposed into natural attributes.

Databases are the natural technology to use for storing and retrieving structured transactional or *operational* data (i.e., the type of data generated by a company's day-to-day operations). However, as companies have become larger and more automated, the amount and variety of data generated by different parts of these companies have dramatically increased. In the 1990s, companies realized that although they were accumulating tremendous amounts of data, they were repeatedly running into difficulties in analyzing those data. Part of the problem was that the data were often stored in numerous separate databases within the one organization. Another difficulty was that databases were optimized for storage and retrieval of data, activities characterized by high volumes of simple operations, such as SELECT, INSERT, UPDATE, and DELETE. In order to analyze their data, these companies needed technology that was able to bring together and reconcile the data from disparate databases and that facilitated more complex analytical data operations. This business challenge led to the development of *data warehouses*. In a data warehouse, data are taken from across the organization

and integrated, thereby providing a more comprehensive data set for analysis.

Over the past couple of decades, our devices have become mobile and networked, and many of us now spend many hours online every day using social technologies, computer games, media platforms, and web search engines. These changes in technology and how we live have had a dramatic impact on the amount of data collected. It is estimated that the amount of data collected over the five millennia since the invention of writing up to 2003 is about 5 exabytes. Since 2013, humans generate and store this same amount of data *every day*. However, it is not only the amount of data collected that has grown dramatically but also the variety of data. Just consider the following list of online data sources: emails, blogs, photos, tweets, likes, shares, web searches, video uploads, online purchases, podcasts. And if we consider the metadata (data describing the structure and properties of the raw data) of these events, we can begin to understand the meaning of the term *big data*. Big data are often defined in terms of the three Vs: the extreme *volume* of data, the *variety* of the data types, and the *velocity* at which the data must be processed.

The advent of big data has driven the development of a range of new database technologies. This new generation of databases is often referred to as "*NoSQL databases*." They typically have a simpler data model than

traditional relational databases. A NoSQL database stores data as objects with attributes, using an object notation language such as the *JavaScript Object Notation* (JSON). The advantage of using an object representation of data (in contrast to a relational table-based model) is that the set of attributes for each object is encapsulated within the object, which results in a flexible representation. For example, it may be that one of the objects in the database, compared to other objects, has only a subset of attributes. By contrast, in the standard tabular data structure used by a relational database, all the data points should have the same set of attributes (i.e., columns). This flexibility in object representation is important in contexts where the data cannot (due to variety or type) naturally be decomposed into a set of structured attributes. For example, it can be difficult to define the set of attributes that should be used to represent free text (such as tweets) or images. However, although this representational flexibility allows us to capture and store data in a variety of formats, these data still have to be extracted into a structured format before any analysis can be performed on them.

The existence of big data has also led to the development of new data-processing frameworks. When you are dealing with large volumes of data at high speeds, it can be useful from a computational and speed perspective to distribute the data across multiple servers, process queries by calculating partial results of a query on each server,

and then merge these results to generate the response to the query. This is the approach taken by the *MapReduce* framework on Hadoop. In the MapReduce framework, the data and queries are mapped onto (or distributed across) multiple servers, and the partial results calculated on each server are then reduced (merged) together.

A History of Data Analysis

Statistics is the branch of science that deals with the collection and analysis of data. The term *statistics* originally referred to the collection and analysis of data about the state, such as demographics data or economic data. However, over time the type of data that statistical analysis was applied to broadened so that today statistics is used to analyze all types of data. The simplest form of statistical analysis of data is the summarization of a data set in terms of *summary* (*descriptive*) *statistics* (including measures of a central tendency, such as the *arithmetic mean*, or measures of variation, such as the *range*). However, in the seventeenth and eighteenth centuries the work of people such as Gerolamo Cardano, Blaise Pascal, Jakob Bernoulli, Abraham de Moivre, Thomas Bayes, and Richard Price laid the foundations of probability theory, and through the nineteenth century many statisticians began to use probability distributions as part of their analytic tool kit. These new developments in mathematics enabled statisticians to move beyond descriptive statistics and to start

doing *statistical learning*. Pierre Simon de Laplace and Carl Friedrich Gauss are two of the most important and famous nineteenth-century mathematicians, and both made important contributions to statistical learning and modern data science. Laplace took the intuitions of Thomas Bayes and Richard Price and developed them into the first version of what we now call *Bayes' Rule*. Gauss, in his search for the missing dwarf planet Ceres, developed the *method of least squares*, which enables us to find the best model that fits a data set such that the error in the fit minimizes the sum of squared differences between the data points in the data set and the model. The method of least squares provided the foundation for statistical learning methods such as *linear regression* and *logistic regression* as well as the development of *artificial neural network* models in artificial intelligence (we will return to least squares, regression analysis, and neural networks in chapter 4).

Between 1780 and 1820, around the same time that Laplace and Gauss were making their contributions to statistical learning, a Scottish engineer named William Playfair was inventing statistical graphics and laying the foundations for modern *data visualization* and *exploratory data analysis*. Playfair invented the *line chart* and *area chart* for time-series data, the *bar chart* to illustrate comparisons between quantities of different categories, and the *pie chart* to illustrate proportions within a set. The advantage of visualizing quantitative data is that it allows us to

use our powerful visual abilities to summarize, compare, and interpret data. Admittedly, it is difficult to visualize large (many data points) or complex (many attributes) data sets, but data visualization is still an important part of data science. In particular, it is useful in helping data scientists explore and understand the data they are working with. Visualizations can also be useful to communicate the results of a data science project. Since Playfair's time, the variety of data-visualization graphics has steadily grown, and today there is research ongoing into the development of novel approaches to visualize large, multidimensional data sets. A recent development is the *t-distributed stochastic neighbor embedding* (t-SNE) algorithm, which is a useful technique for reducing high-dimensional data down to two or three dimensions, thereby facilitating the visualization of those data.

The developments in probability theory and statistics continued into the twentieth century. Karl Pearson developed modern hypothesis testing, and R. A. Fisher developed statistical methods for *multivariate analysis* and introduced the idea of *maximum likelihood estimate* into statistical inference as a method to draw conclusions based on the relative probability of events. The work of Alan Turing in the Second World War led to the invention of the electronic computer, which had a dramatic impact on statistics because it enabled much more complex statistical calculations. Throughout the 1940s and

subsequent decades, a number of important computational models were developed that are still widely used in data science. In 1943, Warren McCulloch and Walter Pitts proposed the first mathematical model of a *neural network*. In 1948, Claude Shannon published "A Mathematical Theory of Communication" and by doing so founded *information theory*. In 1951, Evelyn Fix and Joseph Hodges proposed a model for *discriminatory analysis* (what would now be called a *classification* or *pattern-recognition* problem) that became the basis for modern *nearest-neighbor models*. These postwar developments culminated in 1956 in the establishment of the field of *artificial intelligence* at a workshop in Dartmouth College. Even at this early stage in the development of artificial intelligence, the term *machine learning* was beginning to be used to describe programs that gave a computer the ability to learn from data. In the mid-1960s, three important contributions to ML were made. In 1965, Nils Nilsson's book titled *Learning Machines* showed how neural networks could be used to learn linear models for classification. The following year, 1966, Earl B. Hunt, Janet Marin, and Philip J. Stone developed the concept-learning system framework, which was the progenitor of an important family of ML algorithms that induced decision-tree models from data in a top-down fashion. Around the same time, a number of independent researchers developed and published early versions of the *k-means* clustering algorithm,

now the standard algorithm used for data (customer) segmentation.

The field of ML is at the core of modern data science because it provides algorithms that are able to automatically analyze large data sets to extract potentially interesting and useful patterns. Machine learning has continued to develop and innovate right up to the present day. Some of the most important developments include *ensemble models*, where predictions are made using a set (or committee) of models, with each model voting on each query, and *deep-learning neural networks*, which have multiple (i.e., more than three) layers of neurons. These deeper layers in the network are able to discover and learn complex attribute representations (composed of multiple, interacting input attributes that have been processed by earlier layers), which in turn enable the network to learn patterns that generalize across the input data. Because of their ability to learn complex attributes, deep-learning networks are particularly suitable to high-dimensional data and so have revolutionized a number of fields, including *machine vision* and *natural-language processing*.

As we discussed in our review of database history, the early 1970s marked the beginning of modern database technology with Edgar F. Codd's relational data model and the subsequent explosion of data generation and storage that led to the development of data warehousing in the 1990s and more recently to the phenomenon of big data.

However, well before the emergence of big data, in fact by the late 1980s and early 1990s, the need for a field of research specifically targeting the analysis of these large data sets was apparent. It was around this time that the term *data mining* started to be used in the database communities. As we have already discussed, one response to this need was the development of data warehouses. However, other database researchers responded by reaching out to other research fields, and in 1989 Gregory Piatetsky-Shapiro organized the first workshop on *knowledge discovery in databases* (KDD). The announcement of the first KDD workshop neatly sums how the workshop focused on a multidisciplinary approach to the problem of analyzing large databases:

> Knowledge discovery in databases poses many interesting problems, especially when databases are large. Such databases are usually accompanied by substantial domain knowledge which can significantly facilitate discovery. Access to large databases is expensive—hence the need for sampling and other statistical methods. Finally, knowledge discovery in databases can benefit from many available tools and techniques from several different fields including expert systems, machine learning, intelligent databases, knowledge acquisition, and statistics.[1]

In fact, the terms *knowledge discovery in databases* and *data mining* describe the same concept, the distinction being that data mining is more prevalent in the business communities and KDD more prevalent in academic communities. Today, these terms are often used interchangeably,[2] and many of the top academic venues use both. Indeed, the premier academic conference in the field is the International Conference on Knowledge Discovery and Data Mining.

The Emergence and Evolution of Data Science

The term *data science* came to prominence in the late 1990s in discussions relating to the need for statisticians to join with computer scientists to bring mathematical rigor to the computational analysis of large data sets. In 1997, C. F. Jeff Wu's public lecture "Statistics = Data Science?" highlighted a number of promising trends for statistics, including the availability of large/complex data sets in massive databases and the growing use of computational algorithms and models. He concluded the lecture by calling for statistics to be renamed "data science."

In 2001, William S. Cleveland published an action plan for creating a university department in the field of data science (Cleveland 2001). The plan emphasizes the need for data science to be a partnership between mathematics and computer science. It also emphasizes the need for data science to be understood as a multidisciplinary endeavor

and for data scientists to learn how to work and engage with subject-matter experts. In the same year, Leo Breiman published "Statistical Modeling: The Two Cultures" (2001). In this paper, Breiman characterizes the traditional approach to statistics as a data-modeling culture that views the primary goal of data analysis as identifying the (hidden) stochastic data model (e.g., *linear regression*) that explains how the data were generated. He contrasts this culture with the algorithmic-modeling culture that focuses on using computer algorithms to create prediction models that are accurate (rather than explanatory in terms of how the data was generated). Breiman's distinction between a statistical focus on models that explain the data versus an algorithmic focus on models that can accurately predict the data highlights a core difference between statisticians and ML researchers. The debate between these approaches is still ongoing within statistics (see, for example, Shmueli 2010). In general, today most data science projects are more aligned with the ML approach of building accurate prediction models and less concerned with the statistical focus on explaining the data. So although data science became prominent in discussions relating to statistics and still borrows methods and models from statistics, it has over time developed its own distinct approach to data analysis.

Since 2001, the concept of data science has broadened well beyond that of a redefinition of statistics. For

example, over the past 10 years there has been a tremendous growth in the amount of the data generated by online activity (online retail, social media, and online entertainment). Gathering and preparing these data for use in data science projects has resulted in the need for data scientists to develop the programming and hacking skills to scrape, merge, and clean data (sometimes unstructured data) from external web sources. Also, the emergence of big data has meant that data scientists need to be able to work with big-data technologies, such as Hadoop. In fact, today the role of a data scientist has become so broad that there is an ongoing debate regarding how to define the expertise and skills required to carry out this role.[3] It is, however, possible to list the expertise and skills that most people would agree are relevant to the role, which are shown in figure 1. It is difficult for an individual to master all of these areas, and, indeed, most data scientists usually have in-depth knowledge and real expertise in just a subset of them. However, it is important to understand and be aware of each area's contribution to a data science project.

Data scientists should have some domain expertise. Most data science projects begin with a real-world, domain-specific problem and the need to design a data-driven solution to this problem. As a result, it is important for a data scientist to have enough domain expertise that they understand the problem, why it is important, and how a data science solution to the problem might fit into

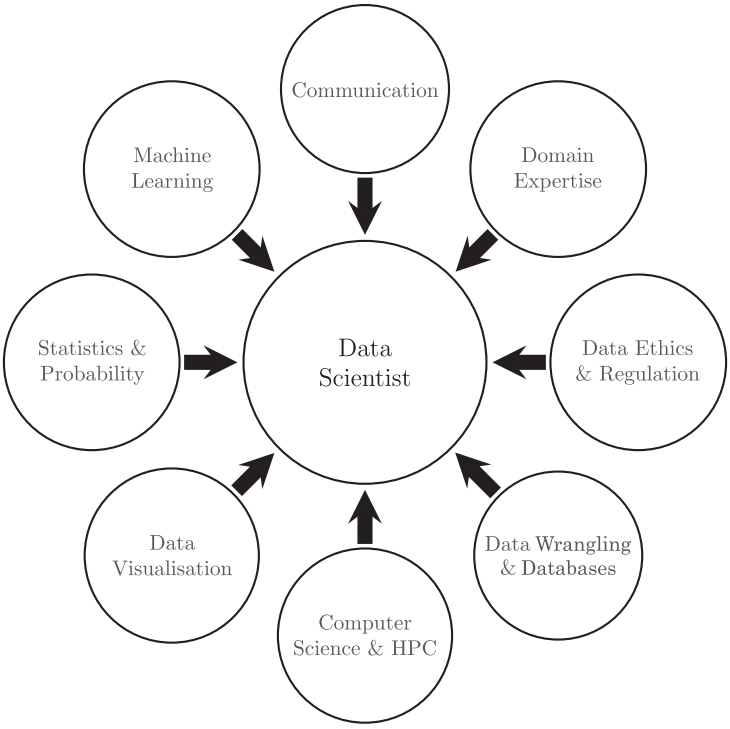

Figure 1 A skills-set desideratum for a data scientist.

an organization's processes. This domain expertise guides the data scientist as she works toward identifying an optimized solution. It also enables her to engage with real domain experts in a meaningful way so that she can illicit and understand relevant knowledge about the underlying problem. Also, having some experience of the project domain allows the data scientist to bring her experiences from working on similar projects in the same and related domains to bear on defining the project focus and scope.

Data are at the center of all data science projects. However, the fact that an organization has access to data does not mean that it can legally or should ethically use the data. In most jurisdictions, there is antidiscrimination and personal-data-protection legislation that regulates and controls the use of data usage. As a result, a data scientist needs to understand these regulations and also, more broadly, to have an ethical understanding of the implications of his work if he is to use data legally and appropriately. We return to this topic in chapter 6, where we discuss the legal regulations on data usage and the ethical questions related to data science.

In most organizations, a significant portion of the data will come from the databases in the organization. Furthermore, as the data architecture of an organization grows, data science projects will start incorporating data from a variety of other data sources, which are commonly referred to as "big-data sources." The data in these data

sources can exist in a variety of different formats, generally a database of some form—relational, NoSQL, or Hadoop. All of the data in these various databases and data sources will need to be integrated, cleansed, transformed, normalized, and so on. These tasks go by many names, such as *extraction, transformation, and load*, "data munging," "data wrangling," "data fusion," "data crunching," and so on. Like source data, the data generated from data science activities also need to be stored and managed. Again, a database is the typical storage location for the data generated by these activities because they can then be easily distributed and shared with different parts of the organization. As a consequence, data scientists need to have the skills to interface with and manipulate data in databases.

A range of computer science skills and tools allows data scientists to work with big data and to process it into new, meaningful information. *High-performance computing* (HPC) involves aggregating computing power to deliver higher performance than one can get from a stand-alone computer. Many data science projects work with a very large data set and ML algorithms that are computationally expensive. In these situations, having the skills required to access and use HPC resources is important. Beyond HPC, we have already mentioned the need for data scientists to be able to scrap, clean, and integrate web data as well as handle and process unstructured text and images. Furthermore, a data scientist may also end up

writing in-house applications to perform a specific task or altering an existing application to tune it to the data and domain being processed. Finally, computer science skills are also required to be able to understand and develop the ML models and integrate them into the production or analytic or back-end applications in an organization.

Presenting data in a graphical format makes it much easier to see and understand what is happening with the data. Data visualization applies to all phases of the data science process. When data are inspected in tabular form, it is easy to miss things such as outliers or trends in distributions or subtle changes in the data through time. However, when data are presented in the correct graphical form, these aspects of the data can pop out. Data visualization is an important and growing field, and we recommend two books, *The Visual Display of Quantitative Information* by Edward Tufte (2001) and *Show Me the Numbers: Designing Tables and Graphs to Enlighten* by Stephen Few (2012) as excellent introductions to the principles and techniques of effective data visualization.

Methods from statistics and probability are used throughout the data science process, from the initial gathering and investigation of the data right through to the comparing of the results of different models and analyses produced during the project. Machine learning involves using a variety of advanced statistical and computing techniques to process data to find patterns. The

data scientist who is involved in the applied aspects of ML does not have to write his own versions of ML algorithms. By understanding the ML algorithms, what they can be used for, what the results they generate mean, and what type of data particular algorithms can be run on, the data scientist can consider the ML algorithms as a gray box. This allows him to concentrate on the applied aspects of data science and to test the various ML algorithms to see which ones work best for the scenario and data he is concerned with.

Finally, a key aspect of being a successful data scientist is being able to communicate the story in the data. This story might uncover the insight that the analysis of the data has revealed or how the models created during a project fit into an organization's processes and the likely impact they will have on the organization's functioning. There is no point executing a brilliant data science project unless the outputs from it are used and the results are communicated in such a way that colleagues with a nontechnical background can understand them and have confidence in them.

Where Is Data Science Used?

Data science drives decision making in nearly all parts of modern societies. In this section, we describe three case

studies that illustrate the impact of data science: consumer companies using data science for sales and marketing; governments using data science to improve health, criminal justice, and urban planning; and professional sporting franchises using data science in player recruitment.

Data Science in Sales and Marketing

Walmart has access to large data sets about its customers' preferences by using point-of-sale systems, by tracking customer behavior on the Walmart website, and by tracking social media commentary about Walmart and its products. For more than a decade, Walmart has been using data science to optimize the stock levels in stores, a well-known example being when it restocked strawberry Pop-Tarts in stores in the path of Hurricane Francis in 2004 based on an analysis of sales data preceding Hurricane Charley, which had struck a few weeks earlier. More recently, Walmart has used data science to drive its retail revenues in terms of introducing new products based on analyzing social media trends, analyzing credit card activity to make product recommendations to customers, and optimizing and personalizing customers' online experience on the Walmart website. Walmart attributes an increase of 10 to 15 percent in online sales to data science optimizations (DeZyre 2015).

The equivalent of up-selling and cross-selling in the online world is the "recommender system." If you have

watched a movie on Netflix or purchased an item on Amazon, you will know that these websites use the data they collect to provide suggestions for what you should watch or buy next. These recommender systems can be designed to guide you in different ways: some guide you toward blockbusters and best sellers, whereas others guide you toward niche items that are specific to your tastes. Chris Anderson's book *The Long Tail* (2008) argues that as production and distribution get less expensive, markets shift from selling large amounts of a small number of hit items to selling smaller amounts of a larger number of niche items. This trade-off between driving sales of hit or niche products is a fundamental design decision for a recommender system and affects the data science algorithms used to implement these systems.

Governments Using Data Science

In recent years, governments have recognized the advantages of adopting data science. In 2015, for example, the US government appointed Dr. D. J. Patil as the first chief data scientist. Some of the largest data science initiatives spearheaded by the US government have been in health. Data science is at the core of the Cancer Moonshot[4] and Precision Medicine Initiatives. The Precision Medicine Initiative combines human genome sequencing and data science to design drugs for individual patients. One part of the initiative is the All of Us program,[5] which is

gathering environment, lifestyle, and biological data from more than one million volunteers to create the world's biggest data sets for precision medicine. Data science is also revolutionizing how we organize our cities: it is used to track, analyze, and control environmental, energy, and transport systems and to inform long-term urban planning (Kitchin 2014a). We return to health and smart cities in chapter 7 when we discuss how data science will become even more important in our lives over the coming decades.

The US government's Police Data Initiative[6] focuses on using data science to help police departments understand the needs of their communities. Data science is also being used to predict crime hot spots and recidivism. However, civil liberty groups have criticized some of the uses of data science in criminal justice. In chapter 6, we discuss the privacy and ethics questions raised by data science, and one of the interesting factors in this discussion is that the opinions people have in relation to personal privacy and data science vary from one domain to the next. Many people who are happy for their personal data to be used for publicly funded medical research have very different opinions when it comes to the use of personal data for policing and criminal justice. In chapter 6, we also discuss the use of personal data and data science in determining life, health, car, home, and travel insurance premiums.

Data Science in Professional Sports

The movie *Moneyball* (Bennett Miller, 2011), starring Brad Pitt, showcases the growing use of data science in modern sports. The movie is based on the book of the same title (Lewis 2004), which tells the true story of how the Oakland A's baseball team used data science to improve its player recruitment. The team's management identified that a player's on-base percentage and slugging percentage statistics were more informative indicators of offensive success than the statistics traditionally emphasized in baseball, such as a player's batting average. This insight enabled the Oakland A's to recruit a roster of undervalued players and outperform its budget. The Oakland A's success with data science has revolutionized baseball, with most other baseball teams now integrating similar data-driven strategies into their recruitment processes.

The moneyball story is a very clear example of how data science can give an organization an advantage in a competitive market space. However, from a pure data science perspective perhaps the most important aspect of the moneyball story is that it highlights that sometimes the primary value of data science is the identification of informative attributes. A common belief is that the value of data science is in the models created through the process. However, once we know the important attributes in a domain, it is very easy to create data-driven models. The key to success is getting the right data and finding

The key to success is getting the right data and finding the right attributes.

the right attributes. In *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, Steven D. Levitt and Stephen Dubner illustrate the importance of this observation across a wide range of problems. As they put it, the key to understanding modern life is "knowing what to measure and how to measure it" (2009, 14). Using data science, we can uncover the important patterns in a data set, and these patterns can reveal the important attributes in the domain. The reason why data science is used in so many domains is that it doesn't matter what the problem domain is: if the right data are available and the problem can be clearly defined, then data science can help.

Why Now?

A number of factors have contributed to the recent growth of data science. As we have already touched upon, the emergence of big data has been driven by the relative ease with which organizations can gather data. Be it through point-of-sales transaction records, clicks on online platforms, social media posts, apps on smart phones, or myriad other channels, companies can now build much richer profiles of individual customers. Another factor is the commoditization of data storage with economies of scale, making it less expensive than ever before to store data. There has also been tremendous growth in computer power. Graphics

cards and graphical processing units (GPUs) were originally developed to do fast graphics rendering for computer games. The distinctive feature of GPUs is that they can carry out fast matrix multiplications. However, matrix multiplications are useful not only for graphics rendering but also for ML. In recent years, GPUs have been adapted and optimized for ML use, which has contributed to large speedups in data processing and model training. User-friendly data science tools have also become available and lowered the barriers to entry into data science. Taken together, these developments mean that it has never been easier to collect, store, and process data.

In the past 10 years there have also been major advances in ML. In particular, deep learning has emerged and has revolutionized how computers can process language and image data. The term *deep learning* describes a family of neural network models with multiple layers of units in the network. Neural networks have been around since the 1940s, but they work best with large, complex data sets and take a great deal of computing resources to train. So the emergence of deep learning is connected with growth in big data and computing power. It is not an exaggeration to describe the impact of deep learning across a range of domains as nothing less than extraordinary.

DeepMind's computer program AlphaGo[7] is an excellent example of how deep learning has transformed a

field of research. Go is a board game that originated in China 3,000 years ago. The rules of Go are much simpler than chess; players take turns placing pieces on a board with the goal of capturing their opponent's pieces or surrounding empty territory. However, the simplicity of the rules and the fact that Go uses a larger board means that there are many more possible board configurations in Go then there are in chess. In fact, there are more possible board configurations for Go than there are atoms in the universe. This makes Go much more difficult than chess for computers because of its much larger search space and difficulty in evaluating each of these possible board configurations. The DeepMind team used deep-learning models to enable AlphaGo to evaluate board configurations and to select the next move to make. The result was that AlphaGo became the first computer program to beat a professional Go player, and in March 2016 AlphaGo beat Led Sedol, the 18-time Go world champion, in a match watched by more than 200 million people worldwide. To put the impact of deep learning on Go in context, as recently as 2009 the best Go computer program in the world was rated at the low end of advanced amateur; seven years later AlphaGo beat the world champion. In 2016, an article describing the deep-learning algorithms behind AlphaGo was published in the world's most prestigious academic science journal, *Nature* (Silver, Huang, Maddison, et al. 2016).

Deep learning has also had a massive impact on a range of high-profile consumer technologies. Facebook now uses deep learning for face recognition and to analyze text in order to advertise directly to individuals based on their online conversations. Both Google and Baidu use deep learning for image recognition, captioning and search, and machine translation. Apple's virtual assistant Siri, Amazon's Alexa, Microsoft's Cortana, and Samsung's Bixby use speech recognition based on deep learning. Huawei is currently developing a virtual assistant for the Chinese market, and it, too, will use deep-learning speech recognition. In chapter 4, "Machine Learning 101," we describe neural networks and deep learning in more detail. However, although deep learning is an important technical development, perhaps what is most significant about it in terms of the growth of data science is the increased awareness of the capabilities and benefits of data science and organization buy-in that has resulted from these high-profile success stories.

Myths about Data Science

Data science has many advantages for modern organizations, but there is also a great deal of hype around it, so we should understand what its limitations are. One of the biggest myths is the belief that data science is an autonomous

process that we can let loose on our data to find the answers to our problems. In reality, data science requires skilled human oversight throughout the different stages of the process. Humans analysts are needed to frame the problem, to design and prepare the data, to select which ML algorithms are most appropriate, to critically interpret the results of the analysis, and to plan the appropriate action to take based on the insight(s) the analysis has revealed. Without skilled human oversight, a data science project will fail to meet its targets. The best data science outcomes occur when human expertise and computer power work together, as Gordon Linoff and Michael Berry put it: "Data mining lets computers do what they do best—dig through lots of data. This, in turn, lets people do what people do best, which is to set up the problem and understand the results" (2011, 3).

The widespread and growing use of data science means that today the biggest data science challenge for many organizations is locating qualified human analysts and hiring them. Human talent in data science is at a premium, and sourcing this talent is currently the main bottleneck in the adoption of data science. To put this talent shortfall in context, in 2011 a McKinsey Global Institute report projected a shortfall in the United States of between 140,000 and 190,000 people with data science and analytics skills and an even larger shortfall of 1.5 million managers with the ability to understand data

science and analytics processes at a level that will enable them to interrogate and interpret the results of data science appropriately (Manyika, Chui, Brown, et al. 2011). Five years on, in their 2016 report, the institute remained convinced that data science has huge untapped value potential across an expanding range of applications but that the talent shortfall will remain, with a predicted shortfall of 250,000 data scientists in the near term (Henke, Bughin, Chui, et al. 2016).

The second big myth of data science is that every data science project needs big data and needs to use deep learning. In general, having more data helps, but having the *right* data is the more important requirement. Data science projects are frequently carried out in organizations that have significantly less resources in terms of data and computing power than Google, Baidu, or Microsoft. Examples indicative of the scale of smaller data science projects include claim prediction in an insurance company that processes around 100 claims a month; student dropout prediction for a university with less than 10,000 students; membership dropout prediction for a union with several thousand members. So an organization doesn't need to be handling terabytes of data or to have massive computing resources at its disposal to benefit from data science.

A third data science myth is that modern data science software is easy to use, and so data science is easy to

do. It is true that data science software has become more user-friendly. However, this ease of use can hide the fact that doing data science properly requires both appropriate domain knowledge and the expertise regarding the properties of the data and the assumptions underpinning the different ML algorithms. In fact, it has never been easier to do data science badly. Like everything else in life, if you don't understand what you are doing when you do data science, you are going to make mistakes. The danger with data science is that people can be intimidated by the technology and believe whatever results the software presents to them. They may, however, have unwittingly framed the problem in the wrong way, entered the wrong data, or used analysis techniques with inappropriate assumptions. So the results the software presents are likely to be the answer to the wrong question or to be based on the wrong data or the outcome of the wrong calculation.

The last myth about data science we want to mention here is the belief that data science pays for itself quickly. The truth of this belief depends on the context of the organization. Adopting data science can require significant investment in terms of developing data infrastructure and hiring staff with data science expertise. Furthermore, data science will not give positive results on every project. Sometimes there is no hidden gem of insight in the data,

and sometimes the organization is not in a position to act on the insight the analysis has revealed. However, in contexts where there is a well-understood business problem and the appropriate data and human expertise are available, then data science can (often) provide the actionable insight that gives an organization the competitive advantage it needs to succeed.