CHAPTER 1

Introduction

Data science is a collection of techniques used to extract value from data. It has become an essential tool for any organization that collects, stores, and processes data as part of its operations. Data science techniques rely on finding useful patterns, connections, and relationships within data. Being a buzzword, there is a wide variety of definitions and criteria for what constitutes data science. Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining. However, each term has a slightly different connotation depending on the context. In this chapter, we attempt to provide a general overview of data science and point out its important features, purpose, taxonomy, and methods.

In spite of the present growth and popularity, the underlying methods of data science are decades if not centuries old. Engineers and scientists have been using predictive models since the beginning of nineteenth century. Humans have always been forward-looking creatures and predictive sciences are manifestations of this curiosity. So, who uses data science today? Almost every organization and business. Sure, we didn't call the methods that are now under data science as "*Data Science.*" The use of the term *science* in data science indicates that the methods are evidence based, and are built on empirical knowledge, more specifically historical observations.

As the ability to collect, store, and process data has increased, in line with Moore's Law - which implies that computing hardware capabilities double every two years, data science has found increasing applications in many diverse fields. Just decades ago, building a production quality regression model took about several dozen hours (Parr Rud, 2001). Technology has come a long way. Today, sophisticated machine learning models can be run, involving hundreds of predictors with millions of records in a matter of a few seconds on a laptop computer.

The process involved in data science, however, has not changed since those early days and is not likely to change much in the foreseeable future. To get meaningful results from any data, a major effort preparing, cleaning,

Data Science. DOI: https://doi.org/10.1016/B978-0-12-814761-0.00001-0

scrubbing, or standardizing the data is still required, before the learning algorithms can begin to crunch them. But what may change is the automation available to do this. While today this process is iterative and requires analysts' awareness of the best practices, soon smart automation may be deployed. This will allow the focus to be put on the most important aspect of data science: interpreting the results of the analysis in order to make decisions. This will also increase the reach of data science to a wider audience.

When it comes to the data science techniques, are there a core set of procedures and principles one must master? It turns out that a vast majority of data science practitioners today use a handful of very powerful techniques to accomplish their objectives: decision trees, regression models, deep learning, and clustering (Rexer, 2013). A majority of the data science activity can be accomplished using relatively few techniques. However, as with all 80/20 rules, the long tail, which is made up of a large number of specialized techniques, is where the value lies, and depending on what is needed, the best approach may be a relatively obscure technique or a combination of several not so commonly used procedures. Thus, it will pay off to learn data science and its methods in a systematic way, and that is what is covered in these chapters. But, first, how are the often-used terms artificial intelligence (AI), machine learning, and data science explained?

1.1 AI, MACHINE LEARNING, AND DATA SCIENCE

Artificial intelligence, Machine learning, and data science are all related to each other. Unsurprisingly, they are often used interchangeably and conflated with each other in popular media and business communication. However, all of these three fields are distinct depending on the context. Fig. 1.1 shows the relationship between artificial intelligence, machine learning, and data science.

Artificial intelligence is about giving machines the capability of mimicking human behavior, particularly cognitive functions. Examples would be: facial recognition, automated driving, sorting mail based on postal code. In some cases, machines have far exceeded human capabilities (sorting thousands of postal mails in seconds) and in other cases we have barely scratched the surface (search "artificial stupidity"). There are quite a range of techniques that fall under artificial intelligence: linguistics, natural language processing, decision science, bias, vision, robotics, planning, etc. Learning is an important part of human capability. In fact, many other living organisms can learn.

Machine learning can either be considered a sub-field or one of the tools of artificial intelligence, is providing machines with the capability of learning

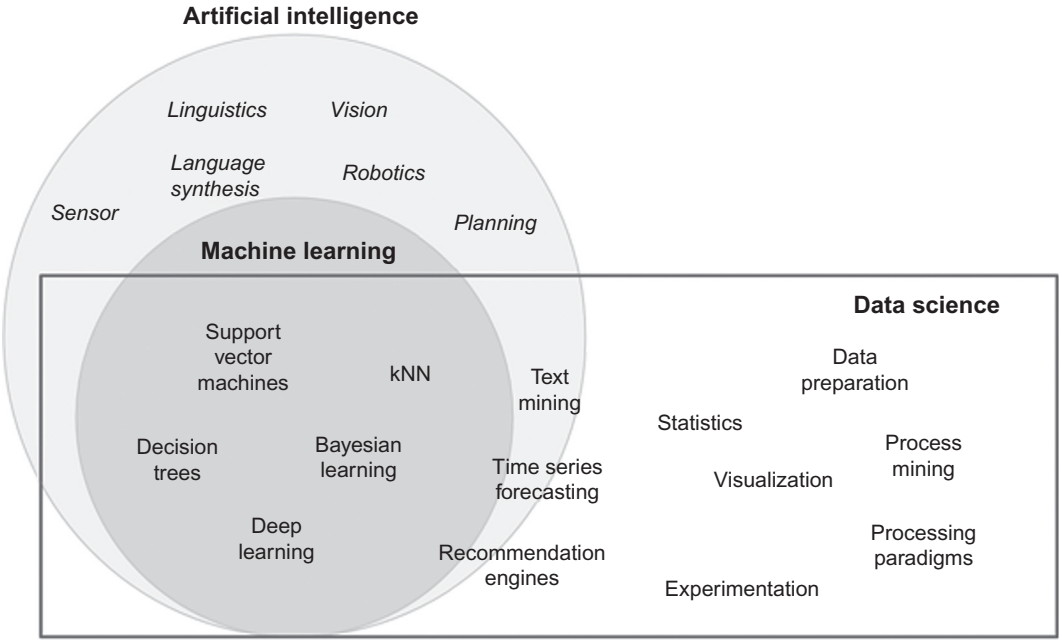

FIGURE 1.1
Artificial intelligence, machine learning, and data science.

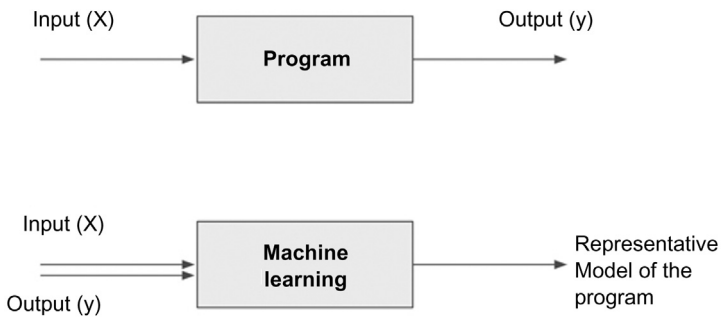

FIGURE 1.2
Traditional program and machine learning.

from experience. Experience for machines comes in the form of data. Data that is used to teach machines is called training data. Machine learning turns the traditional programing model upside down (Fig. 1.2). A program, a set of instructions to a computer, transforms input signals into output signals using predetermined rules and relationships. Machine learning algorithms,

also called "learners", take both the known input and output (training data) to figure out a model for the program which converts input to output. For example, many organizations like social media platforms, review sites, or forums are required to moderate posts and remove abusive content. How can machines be taught to automate the removal of abusive content? The machines need to be shown examples of both abusive and non-abusive posts with a clear indication of which one is abusive. The learners will generalize a pattern based on certain words or sequences of words in order to conclude whether the overall post is abusive or not. The model can take the form of a set of "if—-then" rules. Once the data science rules or model is developed, machines can start categorizing the disposition of any new posts.

Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics. It is an interdisciplinary field that extracts value from data. In the context of how data science is used today, it relies heavily on machine learning and is sometimes called data mining. Examples of data science user cases are: recommendation engines that can recommend movies for a particular user, a fraud alert model that detects fraudulent credit card transactions, find customers who will most likely churn next month, or predict revenue for the next quarter.

1.2 WHAT IS DATA SCIENCE?

Data science starts with *data*, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables. Data science utilizes certain specialized computational *methods* in order to discover meaningful and useful structures within a dataset. The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI). We can further define data science by investigating some of its key features and motivations.

1.2.1 Extracting Meaningful Patterns

Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships within a dataset in order to make important decisions (Fayyad, Piatetsky-shapiro, & Smyth, 1996). Data science involves inference and iteration of many different hypotheses. One of the key aspects of data science is the process of *generalization* of patterns from a dataset. The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data. Data science is also a process with defined steps, each

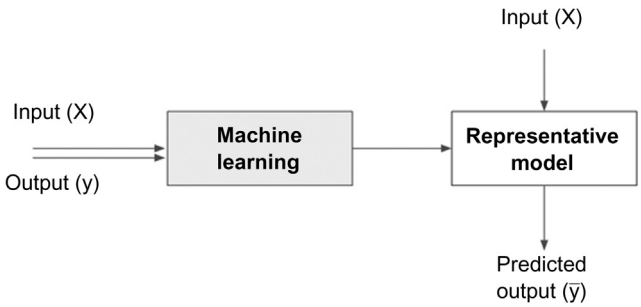

FIGURE 1.3
Data science models.

with a set of tasks. The term *novel* indicates that data science is usually involved in finding previously unknown patterns in data. The ultimate objective of data science is to find potentially useful conclusions that can be acted upon by the users of the analysis.

1.2.2 Building Representative Models

In statistics, a model is the representation of a relationship between variables in a dataset. It describes how one or more variables in the data are related to other variables. Modeling is a process in which a representative abstraction is built from the observed dataset. For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan. For this task, previously known observational data including credit score, income level, loan amount, and interest rate are needed. Fig. 1.3 shows the process of generating a model. Once the representative model is created, it can be used to predict the value of the interest rate, based on all the input variables.

Data science is the process of building a representative model that fits the observational data. This model serves two purposes: on the one hand, it predicts the output (interest rate) based on the new and unseen set of input variables (credit score, income level, and loan amount), and on the other hand, the model can be used to understand the relationship between the output variable and all the input variables. For example, does income level really matter in determining the interest rate of a loan? Does income level matter more than credit score? What happens when income levels double or if credit score drops by 10 points? A Model can be used for both predictive and explanatory applications.

1.2.3 Combination of Statistics, Machine Learning, and Computing

In the pursuit of extracting useful and relevant information from large datasets, data science borrows computational techniques from the disciplines of statistics, machine learning, experimentation, and database theories. The algorithms used in data science originate from these disciplines but have since evolved to adopt more diverse techniques such as parallel computing, evolutionary computing, linguistics, and behavioral studies. One of the key ingredients of successful data science is substantial prior knowledge about the data and the business processes that generate the data, known as *subject matter expertise*. Like many quantitative frameworks, data science is an iterative process in which the practitioner gains more information about the patterns and relationships from data in each cycle. Data science also typically operates on large datasets that need to be stored, processed, and computed. This is where database techniques along with parallel and distributed computing techniques play an important role in data science.

1.2.4 Learning Algorithms

We can also define data science as a process of discovering previously unknown patterns in data using *automatic iterative methods*. The application of sophisticated learning algorithms for extracting useful patterns from data differentiates data science from traditional data analysis techniques. Many of these algorithms were developed in the past few decades and are a part of machine learning and artificial intelligence. Some algorithms are based on the foundations of Bayesian probabilistic theories and regression analysis, originating from hundreds of years ago. These iterative algorithms automate the process of searching for an optimal solution for a given data problem. Based on the problem, data science is classified into *tasks* such as classification, association analysis, clustering, and regression. Each data science task uses specific learning algorithms like decision trees, neural networks, *k*-nearest neighbors (*k*-NN), and *k*-means clustering, among others. With increased research on data science, such algorithms are increasing, but a few classic algorithms remain foundational to many data science applications.

1.2.5 Associated Fields

While data science covers a wide set of techniques, applications, and disciplines, there a few associated fields that data science heavily relies on. The techniques used in the steps of a data science process and in conjunction with the term "data science" are:

- *Descriptive statistics*: Computing mean, standard deviation, correlation, and other descriptive statistics, quantify the aggregate structure of a

dataset. This is essential information for understanding any dataset in order to understand the structure of the data and the relationships within the dataset. They are used in the exploration stage of the data science process.

- *Exploratory visualization*: The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and to comprehend large datasets. Similar to descriptive statistics, they are integral in the pre- and post-processing steps in data science.
- *Dimensional slicing*: Online analytical processing (OLAP) applications, which are prevalent in organizations, mainly provide information on the data through dimensional slicing, filtering, and pivoting. OLAP analysis is enabled by a unique database schema design where the data are organized as dimensions (e.g., products, regions, dates) and quantitative facts or measures (e.g., revenue, quantity). With a well-defined database structure, it is easy to slice the yearly revenue by products or combination of region and products. These techniques are extremely useful and may unveil patterns in data (e.g., candy sales decline after Halloween in the United States).
- *Hypothesis testing*: In confirmatory data analysis, experimental data are collected to evaluate whether a hypothesis has enough evidence to be supported or not. There are many types of statistical testing and they have a wide variety of business applications (e.g., A/B testing in marketing). In general, data science is a process where many hypotheses are generated and tested based on observational data. Since the data science algorithms are iterative, solutions can be refined in each step.
- *Data engineering*: Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage. Database engineering, distributed storage, and computing frameworks (e.g., Apache Hadoop, Spark, Kafka), parallel computing, extraction transformation and loading processing, and data warehousing constitute data engineering techniques. Data engineering helps source and prepare for data science learning algorithms.
- *Business intelligence*: Business intelligence helps organizations consume data effectively. It helps query the ad hoc data without the need to write the technical query command or use dashboards or visualizations to communicate the facts and trends. Business intelligence specializes in the secure delivery of information to right roles and the distribution of information at scale. Historical trends are usually reported, but in combination with data science, both the past and the predicted future data can be combined. BI can hold and distribute the results of data science.

1.3 CASE FOR DATA SCIENCE

In the past few decades, a massive accumulation of data has been seen with the advancement of information technology, connected networks, and the businesses it enables. This trend is also coupled with a steep decline in data storage and data processing costs. The applications built on these advancements like online businesses, social networking, and mobile technologies unleash a large amount of complex, heterogeneous data that are waiting to be analyzed. Traditional analysis techniques like dimensional slicing, hypothesis testing, and descriptive statistics can only go so far in information discovery. A paradigm is needed to manage the massive volume of data, explore the inter-relationships of thousands of variables, and deploy machine learning algorithms to deduce optimal insights from datasets. A set of frameworks, tools, and techniques are needed to intelligently assist humans to process all these data and extract valuable information (Piatetsky-Shapiro, Brachman, Khabaza, Kloesgen, & Simoudis, 1996). Data science is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithms to search for patterns from data. Each key motivation for using data science techniques are explored here.

1.3.1 Volume

The sheer volume of data captured by organizations is exponentially increasing. The rapid decline in storage costs and advancements in capturing every transaction and event, combined with the business need to extract as much leverage as possible using data, creates a strong motivation to store more data than ever. As data become more granular, the need to use large volume data to extract information increases. A rapid increase in the volume of data exposes the limitations of current analysis methodologies. In a few implementations, the time to create generalization models is critical and data volume plays a major part in determining the time frame of development and deployment.

1.3.2 Dimensions

The three characteristics of the Big Data phenomenon are high volume, high velocity, and high variety. The variety of data relates to the multiple types of values (numerical, categorical), formats of data (audio files, video files), and the application of the data (location coordinates, graph data). Every single record or data point contains multiple attributes or variables to provide context for the record. For example, every user record of an ecommerce site can contain attributes such as products viewed, products purchased, user demographics, frequency of purchase, clickstream, etc. Determining the most effective offer for an ecommerce user can involve computing information across

these attributes. Each attribute can be thought of as a dimension in the data space. The user record has multiple attributes and can be visualized in multi-dimensional space. The addition of each dimension increases the complexity of analysis techniques.

A simple linear regression model that has one input dimension is relatively easy to build compared to multiple linear regression models with multiple dimensions. As the dimensional space of data increase, a scalable framework that can work well with multiple data types and multiple attributes is needed. In the case of text mining, a document or article becomes a data point with each unique word as a dimension. Text mining yields a dataset where the number of attributes can range from a few hundred to hundreds of thousands of attributes.

1.3.3 Complex Questions

As more complex data are available for analysis, the complexity of information that needs to get extracted from data is increasing as well. If the natural clusters in a dataset, with hundreds of dimensions, need to be found, then traditional analysis like hypothesis testing techniques cannot be used in a scalable fashion. The machine-learning algorithms need to be leveraged in order to automate searching in the vast search space.

Traditional statistical analysis approaches the data analysis problem by assuming a stochastic model, in order to predict a response variable based on a set of input variables. A linear regression is a classic example of this technique where the parameters of the model are estimated from the data. These hypothesis-driven techniques were highly successful in modeling simple relationships between response and input variables. However, there is a significant need to extract nuggets of information from large, complex datasets, where the use of traditional statistical data analysis techniques is limited (Breiman, 2001)

Machine learning approaches the problem of modeling by trying to find an algorithmic model that can better predict the output from input variables. The algorithms are usually recursive and, in each cycle, estimate the output and "learn" from the predictive errors of the previous steps. This route of modeling greatly assists in exploratory analysis since the approach here is not validating a hypothesis but generating a multitude of hypotheses for a given problem. In the context of the data problems faced today, both techniques need to be deployed. John Tuckey, in his article "We need both exploratory and confirmatory," stresses the importance of both exploratory and confirmatory analysis techniques (Tuckey, 1980). In this book, a range of data science techniques, from traditional statistical modeling techniques like regressions to the modern machine learning algorithms are discussed.

1.4 DATA SCIENCE CLASSIFICATION

Data science problems can be broadly categorized into *supervised* or *unsupervised* learning models. Supervised or directed data science tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data. Supervised techniques predict the value of the output variables based on a set of input variables. To do this, a model is developed from a *training* dataset where the values of input and output are previously known. The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known. The output variable that is being predicted is also called a class label or target variable. Supervised data science needs a sufficient number of labeled records to learn the model from the data. Unsupervised or undirected data science uncovers hidden patterns in unlabeled data. In unsupervised data science, there are no output variables to predict. The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves. An application can employ both supervised and unsupervised learners.

Data science problems can also be classified into tasks such as: classification, regression, association analysis, clustering, anomaly detection, recommendation engines, feature selection, time series forecasting, deep learning, and text mining (Fig. 1.4). This book is organized around these data science tasks. An overview is presented in this chapter and an in-depth discussion of the

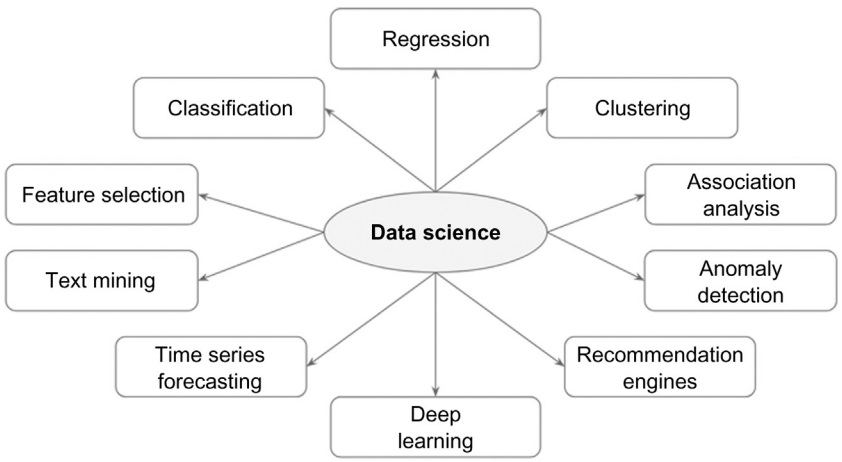

FIGURE 1.4
Data science tasks.

concepts and step-by-step implementations of many important techniques will be provided in the upcoming chapters.

Classification and *regression* techniques predict a target variable based on input variables. The prediction is based on a generalized model built from a previously known dataset. In regression tasks, the output variable is numeric (e.g., the mortgage interest rate on a loan). Classification tasks predict output variables, which are categorical or polynomial (e.g., the yes or no decision to approve a loan). *Deep learning* is a more sophisticated artificial neural network that is increasingly used for classification and regression problems. *Clustering* is the process of identifying the natural groupings in a dataset. For example, clustering is helpful in finding natural clusters in customer datasets, which can be used for market segmentation. Since this is unsupervised data science, it is up to the end user to investigate why these clusters are formed in the data and generalize the uniqueness of each cluster. In retail analytics, it is common to identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other. This task is called market basket analysis or *association analysis*, which is commonly used in cross selling. *Recommendation engines* are the systems that recommend items to the users based on individual user preference.

Anomaly or outlier detection identifies the data points that are significantly different from other data points in a dataset. Credit card transaction fraud detection is one of the most prolific applications of anomaly detection. *Time series forecasting* is the process of predicting the future value of a variable (e.g., temperature) based on past historical values that may exhibit a trend and seasonality. *Text mining* is a data science application where the input data is text, which can be in the form of documents, messages, emails, or web pages. To aid the data science on text data, the text files are first converted into document vectors where each unique word is an attribute. Once the text file is converted to document vectors, standard data science tasks such as classification, clustering, etc., can be applied. *Feature selection* is a process in which attributes in a dataset are reduced to a few attributes that really matter.

A complete data science application can contain elements of both supervised and unsupervised techniques (Tan et al., 2005). Unsupervised techniques provide an increased understanding of the dataset and hence, are sometimes called descriptive data science. As an example of how both unsupervised and supervised data science can be combined in an application, consider the following scenario. In marketing analytics, clustering can be used to find the natural clusters in customer records. Each customer is assigned a cluster label at the end of the clustering process. A labeled customer dataset can now be used to develop a model that assigns a cluster label for any new customer record with a supervised classification technique.

1.5 DATA SCIENCE ALGORITHMS

An algorithm is a logical step-by-step procedure for solving a problem. In data science, it is the blueprint for how a particular data problem is solved. Many of the learning algorithms are recursive, where a set of steps are repeated many times until a limiting condition is met. Some algorithms also contain a random variable as an input and are aptly called *randomized algorithms*. A classification task can be solved using many different learning algorithms such as decision trees, artificial neural networks, *k*-NN, and even some regression algorithms. The choice of which algorithm to use depends on the type of dataset, objective, structure of the data, presence of outliers, available computational power, number of records, number of attributes, and so on. It is up to the data science practitioner to decide which algorithm (s) to use by evaluating the performance of multiple algorithms. There have been hundreds of algorithms developed in the last few decades to solve data science problems.

Data science algorithms can be implemented by custom-developed computer programs in almost any computer language. This obviously is a time-consuming task. In order to focus the appropriate amount of time on data and algorithms, data science tools or statistical programing tools, like R, RapidMiner, Python, SAS Enterprise Miner, etc., which can implement these algorithms with ease, can be leveraged. These data science tools offer a library of algorithms as functions, which can be interfaced through programming code or configurated through graphical user interfaces. Table 1.1 provides a summary of data science tasks with commonly used algorithmic techniques and example cases.

1.6 ROADMAP FOR THIS BOOK

It's time to explore data science techniques in more detail. The main body of this book presents: the concepts behind each data science algorithm and a practical implementation (or two) for each. The chapters do not have to be read in a sequence. For each algorithm, a general overview is first provided, and then the concepts and logic of the learning algorithm and how it works in plain language are presented. Later, how the algorithm can be implemented using RapidMiner is shown. RapidMiner is a widely known and used software tool for data science (Piatetsky, 2018) and it has been chosen particularly for ease of implementation using GUI, and because it is available to use free of charge, as an open source data science tool. Each chapter is

Table 1.1 Data Science Tasks and Examples

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known dataset	Decision trees, neural networks, Bayesian models, induction rules, *k*-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known dataset	Linear regression, logistic regression	Predicting the unemployment rate for the next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the dataset	Distance-based, density-based, LOF	Detecting fraudulent credit card transactions and network intrusion
Time series forecasting	Predict the value of the target variable for a future timeframe based on historical values	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the dataset based on inherit properties within the dataset	*k*-Means, density-based clustering (e.g., DBSCAN)	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data	FP-growth algorithm, a priori algorithm	Finding cross-selling opportunities for a retailer based on transaction purchase history
Recommendation engines	Predict the preference of an item for a user	Collaborative filtering, content-based filtering, hybrid recommenders	Finding the top recommended movies for a user

LOF, *local outlier factor;* ARIMA, *autoregressive integrated moving average;* DBSCAN, *density-based spatial clustering of applications with noise;* FP, *frequent pattern.*

concluded with some closing thoughts and further reading materials and references are listed. Here is a roadmap of the book.

1.6.1 Getting Started With Data Science

Successfully uncovering patterns in a dataset is an iterative process. Chapter 2, Data Science Process, provides a framework to solve the data science problems. A five-step process outlined in this chapter provides guidelines on gathering subject matter expertise; exploring the data with statistics and visualization; building a model using data science algorithms; testing

and deploying the model in the production environment; and finally reflecting on new knowledge gained in the cycle.

Simple data exploration either visually or with the help of basic statistical analysis can sometimes answer seemingly tough questions meant for data science. Chapter 3, Data Exploration, covers some of the basic tools used in knowledge discovery before deploying data science techniques. These practical tools increase one's understanding of data and are quite essential in understanding the results of the data science process.

1.6.2 Practice using RapidMiner

Before delving into the key data science techniques and algorithms, two specific things should be noted regarding how data science algorithms can be implemented while reading this book. It is believed that learning the concepts and implementing them enhances the learning experience. First, it is recommended that the free version of RapidMiner Studio software is downloaded from http://www.rapidminer.com and second, the first few sections of Chapter 15: Getting started with RapidMiner, should be reviewed in order to become familiar with the features of the tool, its basic operations, and the user interface functionality. Acclimating with RapidMiner will be helpful while using the algorithms that are discussed in this book. Chapter 15: Getting started with RapidMiner, is set at the end of the book because some of the later sections in the chapter build on the material presented in the chapters on tasks; however, the first few sections are a good starting point to become familiar with the tool.

Each chapter has a dataset used to describe the concept of a particular data science task and in most cases the same dataset is used for the implementation. The step-by-step instructions on practicing data science using the dataset are covered in every algorithm. All the implementations discussed are available at the companion website of the book at www.IntroDataScience.com. Though not required, it is advisable to access these files to as a learning aid. The dataset, complete RapidMiner processes (*.rmp files), and many more relevant electronic files can be downloaded from this website.

1.6.3 Core Algorithms

Classification is the most widely used data science task in business. The objective of a classification model is to predict a target variable that is binary (e.g., a loan decision) or categorical (e.g., a customer type) when a set of input variables are given. The model does this by learning the generalized relationship between the predicted target variable with all other input attributes

from a known dataset. There are several ways to skin this cat. Each algorithm differs by how the relationship is extracted from the known training dataset. Chapter 4, Classification, on classification addresses several of these methods.

- *Decision trees* approach the classification problem by partitioning the data into purer subsets based on the values of the input attributes. The attributes that help achieve the cleanest levels of such separation are considered significant in their influence on the target variable and end up at the root and closer-to-root levels of the tree. The output model is a tree framework than can be used for the prediction of new unlabeled data.
- *Rule induction* is a data science process of deducing "if–then" rules from a dataset or from the decision trees. These symbolic decision rules explain an inherent relationship between the input attributes and the target labels in the dataset that can be easily understood by anyone.
- *Naïve Bayesian* algorithms provide a probabilistic way of building a model. This approach calculates the probability for each value of the class variable for given values of input variables. With the help of conditional probabilities, for a given unseen record, the model calculates the outcome of all values of target classes and comes up with a predicted winner.
- Why go through the trouble of extracting complex relationships from the data when the entire training dataset can be memorized and the relationship can appear to have been generalized? This is exactly what the *k-NN* algorithm does, and it is, therefore, called a "lazy" learner where the entire training dataset is memorized as the model.
- Neurons are the nerve cells that connect with each other to form a biological neural network in our brain. The working of these interconnected nerve cells inspired the approach of some complex data problems by the creation of *artificial neural networks*. The neural networks section provides a conceptual background of how a simple neural network works and how to implement one for any general prediction problem. Later on we extend this to deep neural networks which have revolutionized the field of artificial intelligence.
- *Support vector machines (SVMs)* were developed to address optical character recognition problems: how can an algorithm be trained to detect boundaries between different patterns, and thus, identify characters? SVMs can, therefore, identify if a given data sample belongs within a boundary (in a particular class) or outside it (not in the class).
- *Ensemble learners* are "meta" models where the model is a combination of several different individual models. If certain conditions are met, ensemble learners can gain from the wisdom of crowds and greatly reduce the generalization error in data science.

The simple mathematical equation $y = ax + b$ is a linear regression model. Chapter 5, Regression Methods, describes a class of data science techniques in which the target variable (e.g., interest rate or a target class) is functionally related to input variables.

- *Linear regression*: The simplest of all function fitting models is based on a linear equation, as previously mentioned. Polynomial regression uses higher-order equations. No matter what type of equation is used, the goal is to represent the variable to be predicted in terms of other variables or attributes. Further, the predicted variable and the independent variables all have to be numeric for this to work. The basics of building regression models will be explored and how predictions can be made using such models will be shown.
- *Logistic regression*: Addresses the issue of predicting a target variable that may be binary or binomial (such as 1 or 0, yes or no) using predictors or attributes, which may be numeric.

Supervised data science or directed data science predict the value of the target variables. Two important *unsupervised* data science tasks will be reviewed: Association Analysis in Chapter 6 and Clustering in Chapter 7. Ever heard of the beer and diaper association in supermarkets? Apparently, a supermarket discovered that customers who buy diapers also tend to buy beer. While this may have been an urban legend, the observation has become a poster child for association analysis. Associating an item in a transaction with another item in the transaction to determine the most frequently occurring patterns is termed *association analysis*. This technique is about, for example, finding relationships between products in a supermarket based on purchase data, or finding related web pages in a website based on clickstream data. It is widely used in retail, ecommerce, and media to creatively bundle products.

Clustering is the data science task of identifying natural groups in the data. As an unsupervised task, there is no target class variable to predict. After the clustering is performed, each record in the dataset is associated with one or more cluster. Widely used in marketing segmentations and text mining, clustering can be performed by a range of algorithms. In Chapter 7, Clustering, three common algorithms with diverse identification approaches will be discussed. The *k-means clustering* technique identifies a cluster based on a central prototype record. *DBSCAN* clustering partitions the data based on variation in the density of records in a dataset. *Self-organizing maps* create a two-dimensional grid where all the records related with each other are placed next to each other.

How to determine which algorithms work best for a given dataset? Or for that matter how to objectively quantify the performance of any algorithm on a dataset? These questions are addressed in Chapter 8, Model Evaluation,

which covers performance evaluation. The most commonly used tools for evaluating classification models such as a confusion matrix, ROC curves, and lift charts are described.

Chapter 9, Text Mining, provides a detailed look into the area of text mining and text analytics. It starts with a background on the origins of text mining and provides the motivation for this fascinating topic using the example of IBM's Watson, the Jeopardy—winning computer program that was built using concepts from text and data mining. The chapter introduces some key concepts important in the area of text analytics such as term frequency—inverse document frequency scores. Finally, it describes two case studies in which it is shown how to implement text mining for document clustering and automatic classification based on text content.

Chapter 10, Deep Learning, describes a set of algorithms to model high level abstractions in data. They are increasingly applied to image processing, speech recognition, online advertisements, and bioinformatics. This chapter covers the basic concepts of deep learning, popular use cases, and a sample classification implementation.

The advent of digital economy exponentially increased the choices of available products to the customer which can be overwhelming. Personalized recommendation lists help by narrowing the choices to a few items relevant to a particular user and aid users in making final consumption decisions. Recommendation engines, covered in Chapter 11, are the most prolific utilities of machine learning in everyday experience. Recommendation engines are a class of machine learning techniques that predict a user preference for an item. There are a wide range of techniques available to build a recommendation engine. This chapter discusses the most common methods starting with *collaborative filtering* and *content-based filtering* concepts and implementations using a practical dataset.

Forecasting is a common application of time series analysis. Companies use sales forecasts, budget forecasts, or production forecasts in their planning cycles. Chapter 12 on Time Series Forecasting starts by pointing out the distinction between standard supervised predictive models and time series forecasting models. The chapter covers a few time series forecasting methods, starting with time series decomposition, moving averages, exponential smoothing, regression, ARIMA methods, and machine learning based methods using windowing techniques.

Chapter 13 on Anomaly Detection describes how outliers in data can be detected by combining multiple data science tasks like classification, regression, and clustering. The fraud alert received from credit card companies is the result of an anomaly detection algorithm. The target variable to be

predicted is whether a transaction is an outlier or not. Since clustering tasks identify outliers as a cluster, distance-based and density-based clustering techniques can be used in anomaly detection tasks.

In data science, the objective is to develop a representative model to generalize the relationship between input attributes and target attributes, so that we can predict the value or class of the target variables. Chapter 14, Feature Selection, introduces a preprocessing step that is often critical for a successful predictive modeling exercise: *feature selection*. Feature selection is known by several alternative terms such as attribute weighting, dimension reduction, and so on. There are two main styles of feature selection: filtering the key attributes before modeling (filter style) or selecting the attributes during the process of modeling (wrapper style). A few filter-based methods such as principal component analysis (PCA), information gain, and chi-square, and a couple of wrapper-type methods like forward selection and backward elimination will be discussed.

The first few sections of Chapter 15, Getting Started with RapidMiner, should provide a good overview for getting familiar with RapidMiner, while the latter sections of this chapter discuss some of the commonly used productivity tools and techniques such as data transformation, missing value handling, and process optimizations using RapidMiner.

References

Breiman, L. (2001). Statistical modeling: Two cultures. *Statistical Science, 6*(3), 199–231.

Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From data science to knowledge discovery in databases. *AI Magazine, 17*(3), 37–54.

Parr Rud, O. (2001). *Data science Cookbook*. New York: John Wiley and Sons.

Piatetsky, G. (2018). Top Software for Analytics, Data Science, Machine Learning in 2018: Trends and Analysis. Retrieved July 7, 2018, from https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html.

Piatetsky-Shapiro, G., Brachman, R., Khabaza, T., Kloesgen, W., & Simoudis, E. (1996). An overview of issues in developing industrial data science and knowledge discovery applications. In: *KDD-96 conference proceedings. KDD-96 conference proceedings*.

Rexer, K. (2013). *2013 Data miner survey summary report*. Winchester, MA: Rexer Analytics. <www.rexeranalytics.com>.

Tan, P.-N., Michael, S., & Kumar, V. (2005). *Introduction to data science*. Boston, MA: Addison-Wesley.

Tuckey, J. (1980). We need exploratory and Confirmatory. *The American Statistician, 34*(1), 23–25.