CHAPTER 1

INTRODUCTION TO DATA SCIENCE

1.1 WHY DATA SCIENCE?

Data science is one of the fastest growing fields in the world, with 6.5 times as many job openings in 2017 as compared to 2012.[1] Demand for data scientists is expected to increase in the future. For example, in May 2017, IBM projected that yearly demand for "data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020."[2] http://InfoWorld.com reported that the #1 "reason why data scientist remains the top job in America"[3] is that "there is a shortage of talent." That is why we wrote this book, to help alleviate the shortage of qualified data scientists.

1.2 WHAT IS DATA SCIENCE?

Simply put, *data science* is the systematic analysis of data within a scientific framework. That is, data science is the

- adaptive, iterative, and phased approach to the analysis of data,
- performed within a systematic framework,
- that uncovers optimal models,
- by assessing and accounting for the true costs of prediction errors.

[1]Forbes, https://www.forbes.com/sites/louiscolumbus/2017/12/11/linkedins-fastest-growing-jobs-today-are-in-data-science-machine-learning/#5b3100f051bd

[2]Forbes, https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6b6fde277e3b

[3]http://Infoworld.com, https://www.infoworld.com/article/3190008/big-data/3-reasons-why-data-scientist-remains-the-top-job-in-america.html

Data Science Using Python and R, First Edition. Chantal D. Larose and Daniel T. Larose.

Data science combines the

- data-driven approach of statistical data analysis,
- the computational power and programming acumen of computer science, and
- domain-specific business intelligence,

in order to uncover actionable and profitable nuggets of information from large databases.

In other words, data science allows us to extract actionable knowledge from under-utilized databases. Thus, data warehouses that have been gathering dust can now be leveraged to uncover hidden profit and enhance the bottom line. Data science lets people leverage large amounts of data and computing power to tackle complex questions. Patterns can arise out of data which could not have been uncovered otherwise. These discoveries can lead to powerful results, such as more effective treatment of medical patients or more profits for a company.

1.3 THE DATA SCIENCE METHODOLOGY

We follow the *Data Science Methodology* (DSM),[4] which helps the analyst keep track of which phase of the analysis he or she is performing. Figure 1.1 illustrates the adaptive and iterative nature of the DSM, using the following phases:

1. **Problem Understanding Phase.** How often have teams worked hard to solve a problem, only to find out later that they solved the wrong problem? Further, how often have the marketing team and the analytics team not been on the same page? This phase attempts to avoid these pitfalls.
   a. First, clearly enunciate the project objectives,
   b. Then, translate these objectives into the formulation of a problem that can be solved using data science.
2. **Data Preparation Phase.** Raw data from data repositories is seldom ready for the algorithms straight out of the box. Instead, it needs to be cleaned or "prepared for analysis." When analysts first examine the data, they uncover the inevitable problems with data quality that always seem to occur. It is in this phase that we fix these problems. Data cleaning/preparation is probably the most labor-intensive phase of the entire data science process. The following is a non-exhaustive list of the issues that await the data preparer.
   a. Identifying outliers and determining what to do about them.
   b. Transforming and standardizing the data.
   c. Reclassifying categorical variables.
   d. Binning numerical variables.
   e. Adding an index field.

[4]Adapted from the Cross-Industry Standard Practice for Data Mining (CRISP-DM). See, for example, *Data Mining and Predictive Analytics*, by Daniel T. Larose and Chantal D. Larose, John Wiley and Sons, Inc, 2015.

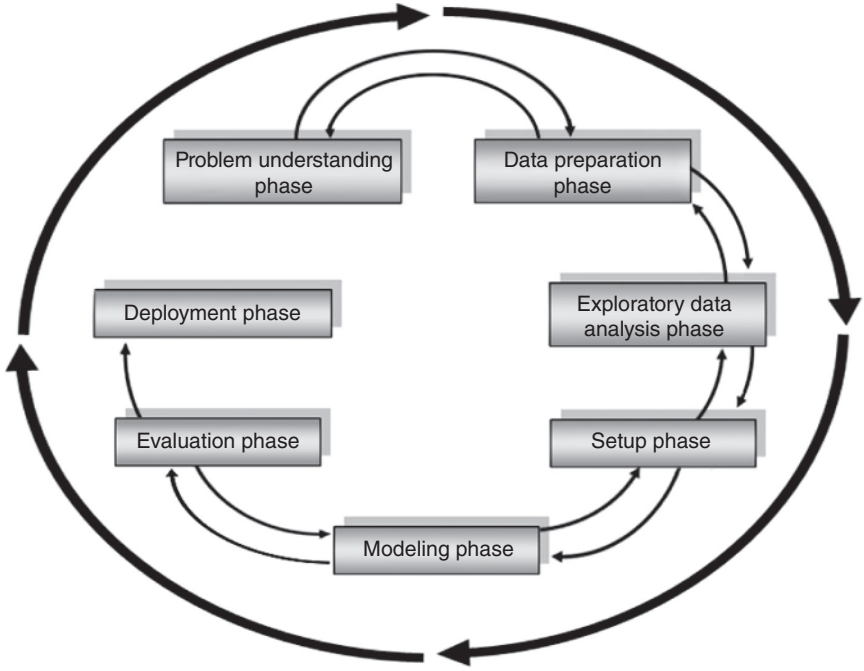

Figure 1.1 Data science methodology: the seven phases.

The data preparation phase is covered in Chapter 3.

3. **Exploratory Data Analysis Phase.** Now that your data are nice and clean, we can begin to explore the data, and learn some basic information. Graphical exploration is the focus here. Now is not the time for complex algorithms. Rather, we use simple exploratory methods to help us gain some preliminary insights. You might find that you can learn quite a bit just by using these simple methods. Here are some of the ways we can do this.
   a. Exploring the univariate relationships between predictors and the target variable.
   b. Exploring multivariate relationships among the variables.
   c. Binning based on predictive value to enhance our models.
   d. Deriving new variables based on a combination of existing variables.

   We cover the exploratory data analysis phase in Chapter 4.
4. **Setup Phase.** At this point we are nearly ready to begin modeling the data. We just need to take care of a few important chores first, such as the following:
   a. Cross-validation, either twofold or *n*-fold. This is necessary to avoid data dredging. In addition, your data partitions need to be evaluated to ensure that they are indeed random.
   b. Balancing the data. This enhances the ability of certain algorithms to uncover relationships in the data.

   c. Establishing baseline performance. Suppose we told you we had a model that could predict correctly whether a credit card transaction was fraudulent or not 99% of the time. Impressed? You should not be. The non-fraudulent transaction rate is 99.932%.[5] So, our model could simply predict that *every* transaction was non-fraudulent and be correct 99.932% of the time. This illustrates the importance of establishing baseline performance for your models, so that we can calibrate our models and determine whether they are any good.

   The Setup Phase is covered in Chapter 5.

5. **Modeling Phase.** The modeling phase represents the opportunity to apply state-of-the-art algorithms to uncover some seriously profitable relationships lying hidden in the data. The modeling phase is the heart of your data scientific investigation and includes the following:

   a. Selecting and implementing the appropriate modeling algorithms. Applying inappropriate techniques will lead to inaccurate results that could cost your company big bucks.

   b. Making sure that our models outperform the baseline models.

   c. Fine-tuning your model algorithms to optimize the results. Should our decision tree be wide or deep? Should our neural network have one hidden layer or two? What should be our cutoff point to maximize profits? Analysts will need to spend some time fine-tuning their models before arriving at the optimal solution.

   The modeling phase represents the core of your data science endeavor and is covered in Chapters 6 and 8–14.

6. **Evaluation Phase.** Your buddy at work may think he has a lock on his prediction for the Super Bowl. But is his prediction any good? That is the question. Anyone can make predictions. It is how the predictions perform against real data that is the real test. In the evaluation phase, we assess how our models are doing, whether they are making any money, or whether we need to go back and try to improve our prediction models.

   a. Your models need to be evaluated against the baseline performance measures from the Setup Phase. Are we beating the monkeys-with-darts model? If not, better try again.

   b. You need to determine whether your models are actually solving the problem at hand. Are your models actually achieving the objectives set for it back in the Problem Understanding Phase? Has some important aspect of the problem not been sufficiently accounted for?

[5]The Alaric Fraud Report, 2015, https://www.paymentscardsandmobile.com/wp-content/uploads/2015/03/PCM_Alaric_Fraud-Report_2015.pdf

c. Apply error costs intrinsic to the data, because data-driven cost evaluation is the best way to model the actual costs involved. For instance, in a marketing campaign, a false positive is not as costly as a false negative. However, for a mortgage lender, a false positive is much more costly.

d. You should tabulate a suite of models and determine which model performs the best. Choose either a single best model, or a small number of models, to move forward to the Deployment Phase.

The Evaluation Phase is covered in Chapter 7.

7. **Deployment Phase.** Finally, your models are ready for prime time! Report to management on your best models and work with management to adapt your models for real-world deployment.

   a. Writing a report of your results may be considered a simple example of deployment. In your report, concentrate on the results of interest to management. Show that you solved the problem and report on the estimated profit, if applicable.

   b. Stay involved with the project! Participate in the meetings and processes involved in model deployment, so that they stay focused on the problem at hand.

It should be emphasized that the DSM is iterative and adaptive. By *adaptive*, we mean that sometimes it is necessary to return to a previous phase for further work, based on some knowledge gained in the current phase. This is why there are arrows pointing both ways between most of the phases. For example, in the Evaluation Phase, we may find that the model we crafted does not actually address the original problem at hand, and that we need to return to the Modeling Phase to develop a model that will do so.

Also, the DSM is *iterative*, in that sometimes we may use our experience of building an effective model on a similar problem. That is, the model we created serves as an input to the investigation of a related problem. This is why the outer ring of arrows in Figure 1.1 shows a constant recycling of older models used as inputs to examining new solutions to new problems.

1.4 DATA SCIENCE TASKS

The most common data science tasks are the following:

- Description
- Estimation
- Classification
- Clustering
- Prediction
- Association

Next, we describe what each of these tasks represent and in which chapters these tasks are covered.

1.4.1 Description

Data scientists are often called upon to *describe* patterns and trends lying within the data. For example, a data scientist may describe a cluster of customers most likely to leave our company's service as those with high-usage minutes and a high number of customer service calls. After describing this cluster, the data scientist may explain that the high number of customer service calls indicates perhaps that the customer is unhappy. Working with the marketing team, the analyst can then suggest possible interventions to explore to retain such customers.

The description task is in widespread use around the world by specialists and nonspecialists alike. For example, when a sports announcer states that a baseball player has a lifetime batting average (hits/at-bats) of 0.350, he or she is describing this player's lifetime batting performance. This is an example of *descriptive statistics*,[6] further examples of which may be found in the Appendix: Data Summarization and Visualization. Nearly every chapter in the book contains examples of the description task, from the graphical EDA methods of Chapter 4, to the descriptions of data clusters in Chapter 10, to the bivariate relationships in Chapter 11.

1.4.2 Estimation

Estimation refers to the approximation of the value of a numeric target variable using a collection of predictor variables. Estimation models are built using records where the target values are known, so that the models can learn which target values are associated with which predictor values. Then, the estimation models can estimate the target values for new data, for which the target value is unknown. For example, the analyst can estimate the mortgage amount a potential customer can afford, based on a set of personal and demographic factors. This estimate is based on a model built by looking at past models of how much previous customers could afford. Estimation requires that the target variable be numeric. Estimation methods are covered in Chapters 9, 11, and 13.

1.4.3 Classification

Classification is similar to estimation, except that the target variable is categorical rather than continuous. Classification represents perhaps the most widespread task in data science, and the most profitable. For instance, a mortgage lender would be interested in determining which of their customers is likely to default on their

[6]For example, see *Discovering Statistics*, by Daniel T. Larose, W.H. Freeman, 2016.

mortgage loans. Similarly, for credit card companies. The classification models are shown lots of complete records containing the actual default status of past customers. The models then learn which attributes are associated with customers who default. Finally, these trained models are then deployed to new data, customers who have applied for a loan or a credit card, with the expectation that the models will help to classify which customers are most likely to default on their loans. Classification methods are covered in Chapters 6, 8, 9, and 13.

1.4.4 Clustering

The clustering task seeks to identify groups of records which are similar. For example, in a data set of credit card applicants, one cluster might represent younger, more educated customers, while another cluster might represent older, less educated customers. The idea is that the records in a cluster are similar to other records in the same cluster, but different from the records in other clusters. Finding workable clusters is useful in at least two respects: (i) your client may be interested in the cluster profiles, that is, detailed descriptions of the characteristics of each cluster, and (ii) the clusters may themselves be used as inputs to classification or estimation models downstream. Clustering methods are covered in Chapter 10.

1.4.5 Prediction

The prediction task is similar to estimation or classification, except that for prediction the forecasts relate to the future. For example, a financial analyst may be interested in predicting the price of Apple stock three months down the road. This would represent estimation, since price is a numeric variable, and prediction, since it relates to the future. Alternatively, a drug discovery chemist may be interested in whether a particular molecule will lead to a profitable new drug for a pharmaceutical company. This represents both prediction and classification, since the target variable is a yes/no variable, whether the drug will be profitable.

1.4.6 Association

The association task involves determining which attributes are associated with each other, that is, which attributes "go together." The data scientist using association seeks to uncover rules for quantifying the relationship between two or more attributes. These association rules take the form, "If *antecedent*, then *consequent*," together with measures of the support and confidence of the association rule. For example, marketers trying to avoid customer churn might uncover the following association rule: "If calls to customer service greater than three, then customer will churn." The support refers to the proportion of records the rule applies to; the confidence is the proportion of times the rule is correct. We cover the association task in Chapter 14.

EXERCISES

CLARIFYING THE CONCEPTS

1. What is data science?
2. Which areas of study does data science combine?
3. What is the goal of data science?
4. Name the seven phases of the DSM.
5. Why is it a good idea to have a Problem Understanding Phase?
6. Why do we need a Data Preparation Phase? Name three issues that are handled in this phase.
7. In which phase does the data analyst begin to explore the data to learn some simple information?
8. Explain in your own words why we need to establish baseline performance for our models. Which phase does this occur in?
9. Which phase represents the heart of your data scientific investigation? Why might we apply more than one algorithm to solve a problem?
10. How do we determine whether our predictions are any good? During which phase does this occur?
11. True or false: The data scientist's work is done with the Evaluation Phase. Explain.
12. Explain how the DSM is adaptive.
13. Describe how the DSM is iterative.
14. List the most common data science tasks.
15. Which of these tasks have many nonspecialists been doing all along?
16. What is estimation? In estimation, what must be true of the target variable?
17. What is the most widespread task in data science? For this task, what must be true of the target variable?
18. What are cluster profiles?
19. True or false: Prediction can only be used for categorical target variables. Explain.
20. For an association rule, what do we mean by support?