Examples of the Lifecycle

Several case studies that address the entire data science lifecycle are placed throughout this book. These cases serve double duty. They focus on one stage in the lifecycle to provide a specific example of the topics in the part of the book where they are located, and they also demonstrate the entire cycle.

The focus of Chapter 5 is on the interplay between a question of interest and how data can be used to answer the question. The simple question “Why is my bus always late?” provides a rich case study that is basic enough for the beginning data scientist to track the stages of the lifecycle, and yet nuanced enough to demonstrate how we apply both statistical and computational thinking to answer the question. In this case study, we build a simulation study to inform us about the distribution of wait times for riders. And we fit a simple model to summarize the wait times with a statistic. This case study also demonstrates how, as a data scientist, you can collect your own data to answer questions that interest you.

Chapter 12 studies the accuracy of mass-market air sensors that are used across the United States. We devise a way to leverage data from highly accurate sensors maintained by the Environmental Protection Agency to improve readings from less expensive sensors. This case study shows how crowdsourced, open data can be improved with data from rigorously maintained, precise, government-monitored equipment. In the process, we focus on cleaning and merging data from multiple sources, but we also fit models to adjust and improve air quality measurements.

In Chapter 18 our focus is on model building and prediction. But we cover the full lifecycle and see how the question of interest impacts the model that we build. Our aim is to enable veterinarians in rural Kenya, who have no access to a scale to weigh a donkey, to prescribe medication for a sick animal. As we learn about the design of the study, clean the data, and balance simplicity with accuracy, we assess the predictive capabilities of our model and show how scientists can partner with people facing practical problems and assist them with solutions.

Finally, in Chapter 21 we examine hand-classified news stories in an effort to algorithmically detect fake news from real news. In this case study, we again see how readily accessible information creates amazing opportunities for data scientists to develop new technologies and investigate today’s important problems. These data have been scraped from new stories on the web and classified as fake or real news by people reading the stories. We also see how data scientists thinking creatively can take general information, such as the content of a news article, and transform them into analyzable data to address topical questions.