Robert Ball
Brian Rague

The Beginner's Guide to Data Science

Springer

The Beginner's Guide to Data Science

Robert Ball • Brian Ragué

The Beginner's Guide to Data Science

Springer

Robert Ball
Weber State University
Ogden, UT, USA

Brian Rague
Weber State University
Ogden, UT, USA

ISBN 978-3-031-07864-4 ISBN 978-3-031-07865-1 (eBook)
https://doi.org/10.1007/978-3-031-07865-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Laura, with all my heart – R.B.
To my family Gina, Justin, Maria, Alexis, and Roanna for their enduring love and support – B.W.R.

Preface

Data science does not broker in abstractions or theories. One of the primary objectives of data science is to make sense of a large number of observations, and consequently to make sense of the real world. The type of analysis performed in data science is deeply rooted in the art of acquiring, wrangling, and visualizing information. Information either stored on servers, in the cloud, or in our brains can be said to have substance. Bits and bytes have weight both metaphorically and physically.

Our ability to decipher and filter large amounts of sensory data allows us to navigate successfully through our busy and sometimes dangerous world. Similarly, data science provides the methodology and tools to accurately interpret an increasing volume of incoming information in order to discern patterns, evaluate trends, and make the right decisions. The results of data science analysis provide real-world answers to real-world questions.

This book is about real-world questions and about arriving at the answers as expeditiously and precisely as possible. On the surface, these inquiries may appear to require simple, straightforward responses – yes or no. However, it is the depth and breadth of the question and the ramifications inherent in the answer that demand our focus and respect.

In one possible scenario, responding to the question in the affirmative may take us down a path involving significantly greater personal or financial commitments and resources than if the answer were a succinct and definitive "no." As with any true scientific investigation, answers should lead to further questions. The bottom line is that within the data science realm, answers and questions are equivalent in importance and the tools we use to derive the end results must rise to the level of proficiency and accuracy required by the analytical demands imposed by both question and answer.

Given a well-posed question, the topics in this book provide concise descriptions of the techniques and tools used in data science to generate viable answers. Data science as a discipline leverages strategies and technologies derived from computer science, statistics, and various business domains. Data science is the progeny of these fields but does not belong to any particular one.

The data scientist is an inter-disciplinarian, utilizing the core knowledge from several areas to make assessments of current data for the purposes of determining future directions in research and analysis. This book is intended for the beginning application-oriented data scientist who wishes to learn the essential methods necessary to extract meaning from numbers.

Is This a Textbook or a Practitioner's Book?

Yes – to both.

We believe that a modern textbook should cover the materials and applications that a student needs. Unlike traditional textbooks in which the student begins at "Chap. 1" and reads sequentially to the last chapter, we have made every effort to present and package the essential topics contained in this book so that the student is free to utilize select chapters based solely on their interests.

As a result, the chapters contained in this book are not necessarily meant to be read in the order in which they appear.

Programming Examples and Images

There are many Python example programs used throughout this book. We have done our best to make these examples readily available at a separate downloadable location. In addition, almost every image used in this book was created by Python code. For these plots and diagrams, the source code designed to create the image is cited and available. The current download location for the code included in this book is the URL: https://github.com/robertball/Beginners-Guide-Data-Science

There are three main reasons for posting the Python code at this separate location:

- There are many instances when having the code occupy unnecessary space in the book is neither meaningful nor instructive, so we have maintained the code separately to be perused, executed, and evaluated at your leisure.
- Python and its various associated libraries change. As a result, there may come a time when certain sample programs in the book no longer run efficiently or successfully on modern computers. Under these special circumstances we will respond by updating the relevant programs so they remain functional and informative. We cannot easily update the content of this book, but we can easily revise and refine downloadable code.
- Copying directly from this book to another resource such as a development environment often inadvertently transfers various formatting issues. Access to a download of the original code can avoid these formatting and program structure issues altogether.

Code Formatting
We differentiate code examples from the narrative text of the book. The following is an example of a code snippet:

```
File: an_example_file_name.py (does not actually exist - just an example)
print('This prints out text.')
print('This does too!')
```

Definitions
We highlight definitions in the book by *italicizing* the term being described.

Ship Names, Movie Names, Book Names, and Latin
Following modern conventions, we have also italicized ship names like the *HMS Titanic* (i.e., *The Titanic*), movie names like *Back to the Future*, book names like *Wizard of Oz*, and Latin words or phrases like *a priori*.

Emphasis
We designate special emphasis for specific points and statements by **bolding** them.

Ogden Robert Ball
UT, USA

Ogden Brian Ague
UT, USA

Contents

Chapter 1
Introduction to Data Science

***

The purpose of this book is very simple: to help you make money either directly or indirectly through data science. Before we provide our definition of data science, let us be clear about your potential motivations for reading this book.

If you wish to read a book about the theoretical foundations of data science (or big data/data analytics) then we recommend you do not waste your time with this book. This book at its core is a practical book that will help you make money. There are many other academic books on data science that are filled with mathematical symbols and expressions that explain the principles and algorithms behind data science in greater detail than this book. We do in fact use mathematical symbols and expressions from time to time; however, they are intended to help further clarify the topic under consideration, but you may ignore them without compromising your understanding of the material.

This book will help you make money directly if you are an entrepreneur seeking to produce a product such as a recommendation engine to move your business forward. In addition, this book will help you make money indirectly by launching your career into the expanding field of data science where you will assist your organization (e.g., business, government, or charity) to increase their revenue and you will consequently maintain a stable, growing income by retaining a prestigious position with a rewarding salary.

Regardless of which path you choose, you will want to read this book if you are a motivated person willing to learn a range of different skills to generate revenue, directly or indirectly, in data science. The path is not always obvious, but for the motivated person, it is well worth it.

If you wish to make money directly then the possibility exists that millions to billions of dollars can be earned by learning and effectively applying the concepts of data science. For example, consider Amazon (the company). The core part of their business is twofold: (1) recommending the right products to people by leveraging a recommendation system so that people buy what they want when they want and, (2) determining the best methods to reduce costs by shipping the products that people purchase in a fast and efficient way through optimized logistics and warehousing governed by processes related to supply chain management. These two main business strategies either originate from or are heavily influenced by data science and analytics.

If you wish to make money indirectly then data science *can* help your organization in making well founded decisions and predictions that cover all facets of business operations. The more valuable *you* are to your organization in designing and distributing new products, reducing costs, and discovering new markets, for example, the more secure and extensive your employment opportunities will be and the more money you will be able to earn.

Either way you choose, this book will help you succeed in a practical way by combining business (domain) knowledge and common sense with the reliable foundations of statistics and computer science.

However, whereas most data science books focus primarily on statistics and computer science, we realize that without fully considering and appreciating the business aspect of data science your end result is primarily of theoretical, academic value. In other words, without serious examination of a practical business purpose, your efforts and results will belong to the

R. Ball, B. Rague, *The Beginner's Guide to Data Science*, https://doi.org/10.1007/978-3-031-07865-1_1

nominal category of “Oh! That is neat! I am sure your mother is very proud of what you did.” In contrast, our hope is that your data science results will produce the following reaction, “Wait... Are you saying that if we do that (or build that) that we can make that much money!! Woah... This is huge! Let me call the VP and see how fast we can get your results and recommendations into action right now.”

Regardless of your level of interest in making money, it is important to have a revenue-generating mindset. No matter what kind of project you undertake, you will need money to operate. Your project might be especially noble, altruistic, and perfectly aligned with your inner values and priorities, and we fully support and congratulate those objectives. However, you will need money to keep the lights on in the building, to buy food, to hire additional people for the project, and in general to keep the motivational fire blazing. The more your results help finance either yourself or your organization the more likely both your short-term and long-term goals will be realized.

1.1 Superpowers

If you could have a superpower, what would it be?

Would you like the power to fly like Superman or Peter Pan? What about the capacity to predict the future? Alternatively, would you prefer the ability to walk through walls?

To many people data science is a superpower. For example, techniques and strategies related to data science allow you to predict the future with some level of confidence. Data science grants you the ability to project who may die and who may live and allows you greater insights about investing into the stock market.

Data science enables you to figure out who wrote a particular “anonymous” book based on word frequencies and helps you determine if the author of a book is male or female and the country where they most likely lived during their developing years as a child and teenager.

The principles and practices of data science allow you to detect fraud, scams, and other deceptive practices.

Data science enables you to peer into the far future and to see how many people will populate the earth during any given future year.

The analysis associated with data science also allows you to accomplish more mundane tasks such as understanding why students fail or succeed in their chosen major. It answers the questions of why someone was elected to office and why someone else was not.

Data science practice delves deeper and inspects the fundamental reasons why one house sells immediately at a given location while the exact same house would never sell somewhere else.

What about our ongoing fascination with past events? Data Science allows us to investigate historical data such as the passengers of the *Titanic*, a large ship that sank in the Atlantic Ocean in 1912 and allows us to determine with over 95% accuracy if a particular passenger was destined to live or die. More importantly, data science techniques clarify *why* one passenger died and *why* another passenger survived.

1.2 What Is Data Science?

If you invite 100 data scientists into a room and ask them to define “data science” then you will likely hear 100 different definitions.

For the purposes of this book, we will define *data science* as an inter-disciplinary field of investigation that is concerned with obtaining accurate and reliable insights about data. This insight is often obtained by using domain knowledge (i.e., the business side), statistics, linear algebra, machine learning, visualization, programming, cluster computing, and creativity. In more straightforward terms, data science is a field that involves primarily the domain (what the topic is about), statistics, and computer science.

Another way to view data science is as an exciting and expanding field of endeavor where we do whatever is necessary to fully comprehend and gain insight into data.

Insight, the ability to gain an accurate and intuitive understanding of data, is the key to success for modern businesses. Insight into what customers want and why they want it defines the profitable and rewarding pathways into the foreseeable future for many businesses.

Actionable insight enables people and businesses to react to what is going on around them in the invisible, data-driven world. The methodologies and techniques developed to discover "what we know we don't know" are vital to the success of a business.

Although we can sense the many physical, tangible properties of the world around us, such as smell, color, sound, and temperature, there are many things that we cannot detect. For example, we cannot observe with our eyes what the probability will be for snow next week nor see how traffic patterns are related to economies nor sense so many other millions of patterns that are hidden inside complex data.

Due to the complex nature of life and the dynamic processes associated with daily experience, data science itself is complex and involves many diverse topics that can never be fully covered in a single book. After you read this book, we encourage you to continue your lifelong journey of further exploring data science topics.

Data science is *interdisciplinary*, which by definition means that data science does not belong sequestered and siloed within a single field. By its very nature data science does not exist in isolation or separate from the human experience.

Does anything really exist in isolation? No subject exists by itself, even the seemingly independent axioms and theorems that define mathematics. Can a person study mathematics without enlisting their brain in the analysis of abstract concepts? A brain is made up of cells whose study is part of biology. Also, the study of the mind and behavior is psychology. Also, mathematicians write to describe their work. A mathematician cannot write without pencils, pens, chalk, or some other writing instrument. If mathematics could simply be thought of and not transcribed to paper, then it would have to be communicated through sound which involves physics.

To use the field of geography as a resource, the first law of geography states, "All things are related. The closer they are the more related they are to each other." That is true with all fields and all disciplines in life.

The main thing to remember about data science is that both statistics and computer science are **tools**. For example, programming languages such as R and Python are **used for a specific purpose** when applied to data science problems. A data scientist usually does not take a compiler design course to create programming languages. Similarly, although machine learning is **utilized in and integral** to data science, data scientists typically do not create new machine learning algorithms.

This conceptual overlapping organization resembles how e-commerce (buying and selling products online) **uses** advanced cryptography algorithms to ensure that sensitive information such as credit card numbers are not stolen during an online transaction. Although advanced mathematics was used to create the algorithms it is not necessary to have an advanced degree in mathematics to **employ** them.

Using a more widely familiar example, we all recognize that we do not need to know how to build a car to drive one. Automotive engineers require years of experience in mathematics and engineering to design and manufacture a car. However, to drive a car, you only need to complete a driver's education class and obtain your driver's license. Although it can be argued that knowing more about the detailed electrical and mechanical functions of your car can be a significant advantage when traveling from place to place, it is neither a necessary nor sufficient condition for driving the vehicle.

Figure 1.1 displays a Venn diagram of where data science resides in relation to computer science, statistics, and business, a visual confirmation that data science is truly an interdisciplinary endeavor.

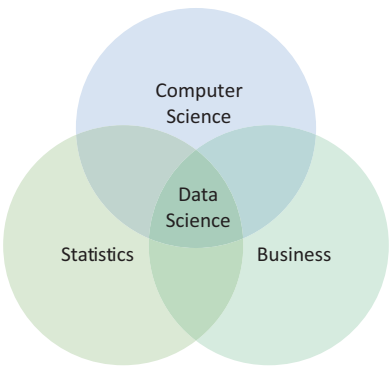

Fig. 1.1 The interdisciplinary nature of data science. In reality, nothing exists in isolation

1.3 Predicting the Future

Think about the temperature of air outside right now at your current location. What will be the recorded temperature at your location tomorrow? Let us be even more specific and state that we only want to know what tomorrow's low temperature will be.

This particular fact-finding task is framed as a simple question but figuring out the answer can be very challenging. *Meteorology*, the branch of science that is concerned with the atmosphere and forecasting of weather, is a complex field, covering metrics and topics such as pressure, temperature, jet streams, moisture, atmospheric layers, mesoscale processes, pollution, volcanic activity, and energetics.

Do you need a degree or advanced knowledge in meteorology to forecast tomorrow's temperature at your location? The answer is no.

If you allow a prediction within a specific range, then we could state that tomorrow's low temperature will be today's low temperature plus or minus (+/−) 50° Fahrenheit (F). For example, if it is currently 25° outside then tomorrow will most likely be between −25° and 75°. How would you assess the quality of this prediction? If you wait until tomorrow this guess will most likely be 100% correct. Our prediction is so broad and inclusive that it does not even matter if we measure temperature using the Fahrenheit or Celsius scale, as either outcome will likely fall within the predicted range.

However, can we narrow down this range of temperatures? Of course! We can predict that tomorrow's low temperature will most likely be today's low temperature +/− 10° F. We might not achieve 100% accuracy, but we will still be in close proximity to the actual temperature.

What level of accuracy can be achieved? Predicting tomorrow's temperature is not that hard because even a modest level of data analysis indicates that it is rare for tomorrow's temperature to be extremely different from today's temperature, although this may occur on rare occasions.

Let us construct a more complex query: What will the temperature at your current location be 100 days from now? More abstractly, what will the temperature be at your current location X days from now where X is any natural number? For example, X could be 120 or 200 or even 1000.

This question is essentially about a forecast, i.e. predicting the future. If we could accurately estimate the temperature and general weather (e.g., rain, snow, wind) of any location on earth for any day in the future to 100% accuracy then we will easily be the source of information for most everyone on the planet for all future weather forecasts. Even if we were only 95% accurate then we could open a weather forecasting business and effectively shut down all other weather predicting businesses across the globe.

In other words, predicting the future, more commonly known as forecasting, ultimately reduces to the essentials of business, providing value over a sustained period of time.

Data science helps businesses become and/or stay relevant, provides customers what they want in a timely manner, and most importantly, seeks to understand why the customer wants what they want even if they do not entirely understand it themselves.

Data science can be used in any area where data exists such as health care, finance, law, government, education, sports, and other areas of work or play. If data are generated in that field, then data science can be effectively utilized. In fact, what current field or topic does not utilize data in some form?

1.4 Understand the Process by Focusing on the End

One way to understand what data scientists do is by examining the end product. The following is a list of items and objectives that most data science projects include:

- Provide actionable insights
- Tell stories with data
- Communicate complex results in a clear, understandable manner
- Create consumable predictive products
- Align business goals with the data science process

1.4.1 Actionable Insights

One of the most important results of any data science project is to formulate *actionable insights*. For example, after analyzing publicly available campaign donations, key factors of donors might be found that shed light on the industries typically associated with each political party.

For example, if a person is employed in finance then it has been generally demonstrated they are more likely to vote for a Republican candidate in the United States. For most people, an insight that links careers and political parties may barely register. However, for potential political candidates (e.g., people trying to get elected or reelected), such information could establish a good return on investment regarding donor identification.

To illustrate, a political candidate canvassing a random cross-section of people for donations would likely receive a low overall ROI for their time and effort. A *return on investment (ROI)* is a key performance metric for determining the effectiveness in expending resources toward a specific business objective. For example, given the amount you invested in this book, what did you actually gain from reading it and applying the principles described?

Regarding the ROI for a political candidate's general survey of a random cross-section of people, did the hours of requesting financial backing justify the few donations received? Conversely, if the candidate knew exactly the industry and individuals to pursue for donations, like the finance sector mentioned previously, then a much higher ROI would be achieved by targeting only those people that evince a high probability of donating.

Another example of effectively applying actionable insights could involve the type of information discussed in a presentation about climate change. Given the same data and results, one presentation might produce actionable results while another does not.

For example, one presentation might emphasize only how humans have affected the global climate. The audience might find that presentation interesting, but not constructively respond by altering any behavior after the presentation has concluded. Another presentation based on the same information might instead offer actionable results and provide a list of measures the audience can perform to counteract any negative influence on climate. Specifically, the presentation might illustrate how switching from incandescent to LED light bulbs in a household is an impactful, actionable item for the audience to consider.

1.4.2 Tell Stories with Data

One important skillset that data scientists usually learn very early is that telling a story is instrumental in crystallizing the purpose of data analysis and the insights to be gained by exploring the data. Questions that should be addressed when working with data sets include:

- What does the data mean?
- Where did the data come from?
- What current insights are available from the data?
- What future predictions are available from the data?

For example, given all the data from the standardized test scores of elementary students in a given state, people have found that elementary students who take the standardized tests earlier in the day *before* lunch attain higher scores than peers in the same school who are administered the test in the afternoon.

An actionable insight would be to recommend that the elementary schools only administer the standardized tests in the morning. This approach would likely result in overall higher scores for the elementary school, which often translates into additional funding for these schools.

The descriptive storyline of these results and recommendations is vitally important though. Numerous psychological studies have found that people gravitate and respond to stories. People love stories and will pay attention if you relate an abstract idea as a story. However, simply listing facts will likely result in your audience becoming distracted and thinking about other pressing concerns, like their next meal or family responsibilities.

Consequently, a data scientist should present the insights gained from studying the standardized tests by recounting the personal stories and experiences of either actual or fictional students.

For example, which do you find more interesting: (1) listening to someone lecture at length about concepts related to circadian rhythm (the reason why students perform better on tests in the morning) or (2) listening to a person who introduces Sally and Sarah, identical twins from a loving home with well-adjusted parents in a middle-class neighborhood.

Sally and Sarah have similar IQ's, and on a particular weekday both ate the same thing for breakfast, and both rode their bikes to school. However, Sally earned a statistically higher score on a standardized test administered at school that day when compared with her twin sister, Sarah. Why? Now that the data scientist has engaged the audience's attention actionable insights can be more effectively shared and understood.

Without the audience's full attention, important information and explanations may be ignored or misinterpreted.

Stories provide a unique opportunity to bring the data to life! The data scientist learns there are many visualization techniques, some static and some interactive, that can provide significant insights into the data. You, the data scientist, will be able to literally show through storytelling how the data interacts on different dimensions and what it means in the broader context of the question.

Staring at spreadsheets or tables filled with thousands or millions of rows of data can simultaneously overwhelm and demotivate most individuals. Creating visualization examples and live animations of that same data can edify and energize those same people.

There are few things more exciting in life than sharing insights that you invested months to discover and then witnessing the metaphoric lightbulbs illuminate people's expressions as you lay out your data story. An isolated table of data is boring. Giving the numbers within that table a background, a present existence, and a potential future is a storywriter's dream. The crucial difference is the story behind your data will be firmly rooted in reality and not fiction.

1.4.3 Communicate Complex Results

One of the biggest challenges with data science projects is the results can be *very* complex. For example, we will find in Chap. 5 that multiple factors interact with each other to best predict the survival or death of passengers during the sinking of the *HMS Titanic* in the early morning hours of April 15, 1912. However, no matter the level of complexity of the relationships among the different factors, if you cannot effectively communicate the results to your target audience then you have wasted your efforts.

Some people believe that complex ideas cannot be communicated in simple ways. For example, the concept of integrals in calculus can be difficult to understand, since it involves understanding areas, curves, and infinitesimal intervals. Although integration can appear complex on the surface, elementary school children are able to sufficiently grasp the essential concept when taught carefully and appropriately. Although integrals are a form of advanced mathematics, they are merely an extension of a simple idea.

For elementary school children, if the area of a fence is represented by the number of dogs that can comfortably be enclosed within that fence then we can leverage that proposition to understand integrals. For example, we can state the premise that a one-foot by one-foot fence has a total area of one square foot and only one dog can be placed comfortably in that enclosure. We can extend this concept to understand rectangular areas. For instance, six dogs can be comfortably situated in a two-foot by three-foot fence. But what about a fence with a curved shape enclosing a specified area? By visualizing a large fence with a curved shape, we can demonstrate that an integral is simply adding up small dog-sized rectangles that fit within the area of interest -- we are really just counting dogs.

In other words, if elementary school children can be taught complex ideas using simple, relatable explanations then you can similarly communicate fundamental concepts to reasonably intelligent stakeholders who may not necessarily understand (or *wish* to understand) the intricate details.

The best teachers, communicators, and data scientists are not the ones who appear clever and intelligent but are those who express the most complex ideas in terms their audience can comprehend. It is easy to create material that appears complicated and is not well understood. What is often difficult and takes special effort is to present problematical ideas in a way that encourages and elevates the awareness and understanding of the target audience or stakeholders.

1.4.4 Create Consumable Predictive Products

The end product of a data science project depends significantly on the data and the project.

While the results of one data science project might simply be a presentation to stakeholders, another result might include the design, development, or application of wholesale predictive software systems. For example, one project might explain why a company should move from one location to another to cut costs. Conversely, a second project might create results in software that help people regulate their insulin levels.

Some examples of software utilized in data science projects could be any of the following:

- A suite of software that performs real-time analysis of news feeds and stock feeds to predict which stocks to buy or sell immediately based on current events.
- NLP (Natural Language Processing) software that analyzes all the current tweets from Twitter to create a summary of trends of what is currently taking place in the world. This information could be used by politicians, businesspeople, or the media.
- A recommendation engine that suggests what you should eat on any given day.
- A News feed that automatically provides local, state, and national news based on the user's current location.

Data science is a broad subject generating results for a wide variety of applications.

1.4.5 Aligning Business Goals with the Data Science Process

Whether you present new business opportunities or investigate ways to optimize current business endeavors, your focus is to maintain and grow your business.

One important metric in business is the *key performance indicator (KPI)*. Although we will not go through a long list of all the KPI's, we will bring them to your attention. For example, we previously mentioned ROI (Return on Investment).

The core measure behind ROI is that if you or your organization invest something, such as time or money, what will be gained in return? For example, if your organization compensates you for a period of two years to uncover actionable insights about when and how children write letters to Santa Claus, what do your employers expect to receive in return? Your time is valuable both to you and to your organization. Will this project help bring in more money or more customers? In other words, regardless of the project focus, how will your efforts ultimately help your organization achieve its goals?

We list just of few examples of KPI's to help you think about business goals in general:

- Customer lifetime value
- Customer acquisition cost
- Net profit margin
- Number of new contracts signed per period
- Average time for conversion
- Click-through rate
- Operational cash flow
- Inventory turnover
- Monthly website traffic

In the end, if your project moves your organization in the direction of its business goals then you will be retained and your organization will generate revenue. However, if your project is simply an interesting but ineffective endeavor and does not align with your organization's goals then eventually either your job, your project, or both will be reassessed.

Ultimately, aligning your data science projects with the business goals of your organization through various KPI's will help both you and your organization. If you do not align your data science projects with the goals of your organization then you will most likely not be involved with the organization very much longer.

1.5 It Is All About the Question!

In any data science project we pursue there must be some goal or key question we are trying to answer. Without a guiding question we seek to resolve we often get distracted by other interesting peripheral questions that may be noteworthy but will ultimately detract from our objective. Throughout the project duration, staying on target and remaining focused on the main objective is of primary importance.

This strategy does not preclude recognizing and recording other salient insights and avenues of investigation along the way. However, if your goal is to ultimately help your organization be profitable, then diversions that lead away from that goal should be carefully evaluated.

For example, Machine Learning (ML) is often used in many data science projects, which benefit from selecting and implementing correct and efficient ML algorithms. However, if your focus is to invest most of your time tweaking ML

algorithms to run slightly faster then you are inadvertently creating a distraction whereas your primary goal should be to obtain the answer or result. After all, an answer to a question, even if generated by an inefficient algorithm, is better than never obtaining that answer because you are spending all your time fine-tuning the algorithm. The answer to your question is ultimately more important than the efficiency of the algorithm used to secure that answer.

There are many types of questions that you might answer for your projects. Regardless of the *type* of question, the question itself is the most important thing because it guides and supports us in our investigation.

The question should be specific and measurable. A purposely vague question like "How do people generally feel about the economy?" is hard to understand and it will be difficult to know when you have truly arrived at the answer to the question. However, a specific and measurable question like "How do people regard taxes on the day that taxes are due?" is an improvement.

1.5.1 Classification Questions

One of the common challenges that data scientists confront is classification. *Classification* problems are those that label data. For example, given an email, is it spam or a legitimate email? Or, given a set of books and 5 possible authors, who wrote which book?

Generally, you can think of classification problems as answering the following fundamental question: **Is this A or B?** Like the above email example, this choice involves a simple *binary answer*: spam or legitimate email. In some cases the response when attempting to assign an observations to a single category may be either positive (*yes*) or negative (*no*).

The solution to a classification problem, such as possible authors, might be a *multinomial answer*, an answer that has more than two possible choices. Instead of restricting the question to A or B, a multinomial question asks: **Is this A or B or … or Z?**

Whether your classification problem is binary or multinomial, classification problems are fairly common and may be expressed through various forms. For example, the following are different types of classification problems:

- Will this tire fail in the next 1000 miles: Yes or no?
- Which brings in more customers: a $5 coupon or a 25% discount?
- If there is a sale at a store, will people spend more at an in-person store (also known as a brick and mortar store) or an online store?
- The classic machine learning (ML) problem: Is the image a cat or a dog?
- Of the three possible ad types, which one are people more likely to click on to improve the click through rate (CTR)? (*Click through rate (CTR)* is a measure that many online businesses use to measure ad effectiveness.)
- Which of the types of marketing provides more customer loyalty?
- Given two texts (e.g., a tweet, blogs, articles, or book.) which one is more positive in terms of tone?

Humans are wired to classify things. We classify students based on grades. We classify teachers based on how well they teach. We classify our children, our pets, our neighbors, the areas of the world, and so on. Whether we do it consciously or unconsciously we are always classifying the world around us. In data science we get paid to do what everyone around us is doing anyway.

1.5.2 Anomaly Detection

"Is this weird?"…that is the basic question behind detecting anomalies. *Anomaly detection* can be applied in situations such as plagiarism, stock market trends, population trends, price spikes, and election results.

Your question might be one in which the new thing you just observed is clearly outside what would be considered normal or expected. If this is the case, is a new trend indicated or is this simply an outlier? An *outlier* is a data point or observation that is outside or at the extreme ends of the expected range.

To answer questions of this nature, we should know how to measure the similarities and differences between data points. Chapters 4 and 5 will focus heavily on these ideas and strategies.

An example of anomaly detection is to determine the best salesperson in your company. You might do this to identify the top salesperson to both reward that person and to understand their methods so that other salespeople can recognize and apply better sales techniques.

Another use for anomaly detection is to investigate election patterns to establish an occurrence of election fraud. For example, if a state traditionally reports approximately 40% of the eligible population voting every election and then suddenly

this measure jumps to over 75% in the current election cycle, can this increase be attributed to election fraud or simply a quite successful "get-out-the-vote" campaign?

In another related example, assume you have data from a sports organization with statistics about all the players. If there are a few players noticeably and abnormally better in specific athletic categories (e.g., stronger, faster.), is it because these players practiced more or instead ingesting prohibited performance enhancing substances like steroids?

Regardless of the context, all anomaly detection questions highlight unexpected singular items or events that are substantially different from the remainder of the data.

1.5.3 Prediction/Forecasting

Prediction is one of the essential tasks for most data scientists. *Prediction* is the process of learning from a set of data, called training data, then estimating the outcome based on different factors. Prediction often answers the question **how much** or **how many?**

For example, if you receive approval by a bank for a loan then the bank researched your application and financials and formulated a prediction from the data that you will be able to pay back the loan. The factors included data from your financial history and the resulting prediction is that you will be able to pay back the loan. This is an example of classification.

For a prediction example that does not involve classification, consider a company that projects their sales amount for the next quarter. (A *quarter* is a business term that divides the year into four equal parts: January – March, April – June, July – September, and October – December.) The prediction of sales for the next quarter is not a classification prediction, but a regression problem.

A *regression* prediction is one that predicts continuous numerical values instead of discrete classifications. For example, examples of regression include predicting the population of people for a country in a given year, predicting the sales of a company, or predicting the number of goals a hockey team will tally over a season.

Predictions based on a time component are called *forecasting*. For example, all weather predictions are forecasts because the answer depends on a future point in time – the forecast next Tuesday might be sunny, but on Wednesday the forecast might be rainy.

1.5.4 Clustering

"To which group does this belong?" is the basic question addressed by clustering. *Clustering* is a form of classification that assigns the individual data points to specific categories.

There are two general types of clustering depending on whether the classification labels are known beforehand or not. For example, given a dataset of purebred dogs and dog breed labels (e.g., great Dane, golden retriever, chihuahua) you could place the dogs into clusters – the precise dog group – based on their breed physical characteristics such as weight, height, size, and shedding. However, what if you were given a set of dogs that were not purebred – a mix of different breeds – and thus did not possess distinctive labels? Mixed breeds do not have existing published statistics, so we need to cluster them based on similar weight and height without breed labels.

Since we do not have an established number of labels we might decide arbitrarily on three resulting clusters: big, medium, and small dogs. Alternatively, we might target a different number of groups such as four or five clusters based on other distinguishing characteristics. Regardless of the final number of groups, the clustering process is the same.

Clustering can also provide clarity on the following question: **How is this organized?** Sometimes you want to understand the underlying structure of a data set. For this question, there are no pre-established outcomes or labels associated with the observations in the dataset. The following provides example questions for this type of clustering problem:

- Which viewers like the same types of movies?
- Which printer models fail the same way?
- What categories of buyers do we have?
- How many distinct categories can we find in the (Republican/Democrat/Libertarian/etc.) party?
- Are most of our buyers early adapters, early majority, late majority, or laggards?

Clustering your dataset into recognizable categories can provide a significant degree of insight into the structure and meaning of the data.

1.5.5 Recommendations

Recommendations, generated from algorithms called recommendation engines, are a form of prediction. Fundamentally, a *recommendation engine* is a prediction engine that appraises and estimates people's opinions and preferences, answering the question **which option should be taken?** Chapter 7 explains the fundamentals of modern recommendation engines.

If you have ever utilized a search engine on the Internet then you have used a recommendation engine. In this case, a search engine's job is to predict the most preferred websites based on a few keywords.

For instance, if you simply entered "cat" into a search engine it has to assess if you would like to know more about the small mammal many people have as a pet, the UNIX operating system command "cat" (as in concatenate), a Broadway play named "Cats," or the company Caterpillar that produces construction equipment with stock symbol CAT. There are also dozens of other related but less widely-known concepts that "cat" could represent.

In a way, the software must look into a proverbial crystal ball and parse the dozens of potential meanings for the target keywords entered into the search field.

Recommendation engines are also closely connected with online stores. Virtually all major online retailers now have recommendation engines that assist the user/consumer. Familiar online shopping pitches such as "other people that bought this product also bought…" originate from recommendation engines.

There are many different types of recommendation engines but they all have the same core objective: Which option among many should be recommended to ensure the highest probability that the user will remain engaged? A recommendation engine adheres to the same operational philosophy as click through rate (CTR) discussed above.

1.5.6 Data Science Project Examples

There are many different data science projects that can be pursued and investigated. We provide the following brief list to help you understand the focus for many professionals in this discipline:

- We are losing approximately 5% of our customers per month. Why?
- What is the real cost, which involves all the parts of the company, of complying with a new regulation?
- Where should we set up another store, ATM, new distribution center, etc.?
- There is a cost associated with hiring and training new employees. It seems that once employees are trained sufficiently they tend to leave for higher salaries.
  - When do most of these employees leave? Why?
  - How can we increase loyalty to our company and retain them?
  - What type of employee, based on background, leaves most frequently? Can we pre-screen for this type of employee and not hire them in the first place?
- If we increase/decrease the quality of our product and also adjust our selling price higher/lower, how will that affect our total sales?
- Based on our products and our most loyal customers, is our marketing team targeting and offering our products to the right people?

The following is also a general list of questions that you might ask for any given data science project to ensure that the goals of the project are business oriented:

- Does the project support the core business or is it peripheral to the business mission?
- Why are we doing this?
- Can I explain the situation and context to a non-expert?
- Do I understand the business requirements to grow the company?
- Are my models shiny and fast or goal-oriented for the business?
- What are the concrete and actionable results?
- Is there an enthusiastic champion in upper management who will act on the insights provided by our results?
- Will the stakeholders accept the change(s) we propose?
- Once the project is finished will we have subject matter experts available to support the finished product or results?

There is an abundance of data science projects any company can undertake. Once the project is identified, the key question then becomes, **should you do it?** If the results of the effort ultimately build business value then you generally should pursue the project. However, it is very easy to become consumed and distracted by the exciting peripherals of the data science milieu which could result in costing your organization money with no appreciable, measurable return.

1.6 Understanding vs Specific Tools

Understanding the essential concepts of data science is more important than learning a specific tool.

For example, Python is currently the most widely used programming language in data science followed by R and VBA (the programming language used in Microsoft Excel). In Sect. 3.3 we explain how to use dataframes with Pandas, currently a popular library that is employed extensively by many data scientists. The Pandas tool is very useful and we have many examples demonstrating how this library will save you time with statistics in Chaps. 4 and 5 as well as with time series analysis in Chap. 10.

However, if you only learn how to use the **tool** then you will be doing yourself a great disservice. For example, in Chap. 10 it is much more important to learn the concepts related to smoothing and moving averages than to learn how to perform these operations specifically in Pandas.

We have found that it is relatively easy to learn a novel way of writing and implementing software if you have the core foundational programming knowledge. Twenty years from now most people may have moved away from Python to another programming language that does not yet exist. If you only know how to accomplish your tasks based on rote memorized functions from Python and the Pandas library, then you will have a hard time transitioning to the new language. Alternatively, if you have focused on the core foundational programming knowledge, then learning how to calculate moving averages in the new language will be a trivial adaptation in the future.

In summary, if you focus on the key principles of data science then you will soon discover that they never change. The current tools, such as Microsoft Excel, Python, and Pandas, will change, but not the principles upon which they are founded. Focus on the principles and you will always have a job. Focus only on the tools and you will find that you may be left behind relentlessly trying to catch up with the latest technology.

1.7 Data Science Life Cycle

As with any established methodology, data science has a predominant life cycle.

The beginning of every data science project defines the question that you hope to answer. This is both the most important and likely the most difficult step. Knowing what question to ask typically involves experience and domain-specific knowledge.

Once the question is identified and validated you then need to gather the data. Gathering the data comes in many forms and is explained in detail in Chap. 2. Whether you get data directly from surveys or indirectly from databases or downloads, you cannot do any type of analysis without the data. As in all research projects, data may be quantitative, qualitative, or a combination of both.

However, once you have the data you may discover that your guiding question is no longer relevant and needs to be modified. Realizing that your question requires revision because of the data under investigation is often as important as coming up with the initial question itself.

The process of data exploration is paramount in the life cycle. You must explore the data to understand the meaning and purpose embodied by this information. However, the data is often not clean. *Clean data* is data that is ready to be used immediately for exploration and analysis. To clean your data you will often need to wrangle, transform, or reconfigure your data, which is the topic of Chap. 3.

Once the data has been wrangled into a usable or clean form then you are ready to engage in a full exploratory analysis to achieve greater insights about the data. We describe the statistics necessary to examine our data in Chaps. 4 and 5.

Part of the exploration process is to visualize the data. *Data visualization* is the vital step of displaying or viewing the data in a form that enables a more comprehensive understanding of the context and trends inherent in the data science problem. In data science, the phrase "A picture is worth a thousand words" may not be accurate. Sometimes a picture or visualization provides instantaneous insight so much greater than simply looking at textual or numerical data that it may easily be worth far more than a thousand words.

Visualization is such an important aspect of the process that it is intertwined throughout the discussions in this book. You will come upon different methods of visualizations frequently throughout this text. Reading a data visualization book will be well worth your time.

You will find that several parts of the life cycle are iterative. Often the process of transforming or wrangling the data followed by visualization will lead to additional questions that bring about additional transformations and visualizations. This part of the data science life cycle is aptly described as "not knowing what you don't know." In other words, many questions arise from exploring and understanding the data that could not be conceived before the process started.

Data reduction is integral to the process. *Data reduction* is the transformation of the data from one form to another for the purposes of simplifying and decreasing the volume of data while simultaneously preserving the essential informational content. For example, given a date timestamp that includes the year, month, day, minute, and second of an event, it may be sufficient to reduce these data to simply the month when identifying quarterly business trends.

The end product of any data science life cycle is either a model that can be used for prediction, a presentation to clarify the meaning and future research of the data, or both. For example, you might develop a recommendation engine for your organization. After completing the software a presentation of the results to stakeholders and interested people will likely follow.

As emphasized above, the presentation is about telling the story of the data, about bringing the data to life, then sharing actionable insights with the audience. No matter how much behind-the-scenes work you put into your project, properly communicating its impact and influence is vitally important.

1.8 Python vs R

Python and R are both the most common programming languages used in data science and often listed as the top languages in job postings. R followed SAS (an older programming language) and emerged initially as the language utilized primarily for statistics and ML (Machine Learning). R is rarely used outside of statistics and data science. R is generally easier to learn for people that have no previous experience with programming languages.

Python followed later and is a general programming language that is one of the most popular in the world. Python is considered easier to learn for anyone that already has a background in programming languages. Also, Python is often one of the first programming languages introduced in university computer science programs.

All the code examples in this book are written in Python for one simple reason: according to years of extensive surveys and studies, Python is gradually gaining popularity over R in data science circles. In a nutshell, the older a person is the more likely they are to use R in industry, unless they are even older and use SAS. Conversely, the younger a person is the more likely they are to use Python in industry. Of course, familiarity with both Python and R provides advantages for individuals engaged in data science.

Industry is a term intended to mean working environments outside of a conventional educational setting. For example, employment or consulting with Apple, Google, or Walmart would be an example of working in industry. If a person does not work in industry, then they likely work in *Academia* or affiliated research institutions. Academia is a term intended to mean the workplace is predominantly an educational or research-oriented setting, such as a university, college, policy think tank, or FFRDC (federally funded research and development center).

1.9 Big Data, Data Analytics, and Data Science

Big data, data analytics, and data science are all terms used interchangeably by some and more precisely by others.

In our experience, universities and professors are more careful about defining the differences and overlapping concepts between the three terms. In contrast, employers of data scientists in the field usually are not overly concerned about these differences. People in industry usually care more about obtaining actionable results to their data science questions to make their organizations more profitable and impactful. They usually are not concerned about whether a “data analyst” or “data scientist” performs these tasks.

Consider the scenario in which you are being interviewed for a position and the interviewer asks if you can perform predictive forecasting work with linear regressions. You can safely assume the interviewer does not really care if you call yourself a “data analyst,” a “data scientist,” or even a “pink balloon.” (Although they might be confused about why you call yourself a “pink balloon.” However, some companies appreciate having eccentric employees, so that strategy might get you hired.)

In this book we distinguish between a data analyst and a data scientist by examining the level of involvement during the lifespan of the project. Data analysts typically identify trends and patterns, create visualizations, and help people make sense of the existing datasets to help them make decisions. These professionals focus on the things we know that we don’t know.

On the other hand, the data scientist typically considers the larger context of the existing problem. This includes culling out new problems and creating additional projects with specific and measurable questions, gathering data that will help answer these questions, analyzing data (which includes identifying trends and patterns, and creating visualizations), writing

software for new software prototypes, and providing presentations. Data scientists are generally interested in the things we don't know that we don't know.

However, identifying the differences between the fields of data analytics and data science is very difficult because few people agree on exactly what each area encompasses. In practical terms, instead of arguing semantics it does not matter what the difference is between data analytics and data science because what is more important is generating the actionable insights described earlier that help your organization become profitable.

Big data is more of a marketing term than a scientific term, similar in spirit to how non-technical people talk about "cloud computing" instead of the more technical term, "distributed computing." "Cloud computing" conjures up ideas of computing taking place mysteriously and magically in the ether while "distributed computing," although more technically accurate, sometimes confuses non-technical people. After all, what does "distributed" suggest to most people?

Big data is a term fundamentally implying that to make sense of the extremely large volumes of data currently available for consumption because of technological advances, visualizations and algorithms are necessary. In other words, to understand the vast amounts of data at our fingertips professionals should apply tools and techniques such as transformations, statistics, visualizations, and many other analytical algorithms to these data to discover relevant patterns and insights.

Many academics refer to the *"four V's" of big data*: volume, variety, velocity, and veracity.

Volume simply refers to the size of your data, with the commonly used modifier "big" implying an inordinate, unmanageable amount. However, in practical relative terms, how much is a lot? For some organizations big data is any data set that is larger than a gigabyte (10^9 bytes). For some, it is greater than a terabyte (10^{12} bytes). For others it might be the next level of a petabyte (10^{15} bytes). The amount of volume or size of a dataset will vary from project to project.

Variety means that the data is packaged in various formats. Your data might be clean and easy to use and derives directly from a database, or it might be unstructured and fragmented, meaning that you derive it from a variety of sources. For example, you might have a project that includes sources of data from video, sound, and text. In contrast, your data might be only stock symbols and numbers. The exact type of data that you use can be drastically different from project to project or it might always be sales data year over year. The data source and structure depend on the project and the kind of data your organization uses.

Velocity is the frequency at which new data is generated. For example, is your updated weather data available at 5-minute increments or daily increments? When analyzing stock data, are you only looking at historical data? For this particular case, the data never changes. However, if you are performing *real-time* stock analysis you might have to work with stock data that changes every second.

Veracity is the level of trust you place in your data. How much do you trust your data? For example, if you are using census data do you blindly accept the precision of all the information that you obtain from the government? Putting aside any conspiracy theories about the government and the census, can we expect the data to be 100% accurate? Tens of thousands of people work on every census. Did any of those hard-working people make a mistake? To accept the census data as 100% accurate is likely both naïve and wrong. We should assume people are fallible and make mistakes, and simple logic compels us to conclude we cannot abide 100% certainty for most data sets.

From a practical standpoint, "big data" is simply data that is too large for meaningful insights to be gained without methodically applying algorithms and expert analysis. To reduce this concept even further, big data is the name given to the datasets you examine in your projects as a data scientist.

Exercises

1. What is the difference between insight and actionable insight?
2. Explain how using stories in presentations can engage your audience.
3. What is the importance of explaining complex things in simple ways?
4. Why do KPI's matter to a data scientist?
5. In your own words, what is a data scientist? How does that differ from a computer scientist, mathematician, or statistician?
6. Is detecting credit card fraud an example of classification, anomaly detection, or both?
7. What is the point of clustering?
8. How does the size of your data impact your analysis?
9. There are many types of recommendation engines. However, it often appears that search engines are different from online retail stores. Compare and contrast general purpose search engines to recommendation engines utilized by online retail stores.

Chapter 2
Data Collection

***

Although there may be many points of contention regarding data science, one thing universally understood about the field is that the practice of data science is impossible without data. You may be given data, or you may have to go out and collect it. The bottom line is that any data science project begins with data.

As discussed in the previous chapter, the difference between a "data scientist" versus a "data analyst" is confusing at best. However, some people would claim that the main difference is a "data analyst" primarily analyzes pre-packaged data whereas a "data scientist" needs to identify, seek out, and gather the data as an important first step.

There are generally two ways to acquire the data of interest:

- Data Creation – Generate raw observations
- Data Gathering – Obtain it from an accessible source

2.1 Data Creation

Data creation is by far the harder approach when compared to obtaining the data from an immediately available and accessible resource.

There are many fields of study and research that collect data. To simplify, we will separate these fields into two types: those that collect information involving people and those that collect information about everything else in the universe.

The following fields are a sample of the many disciplines that collect information about people: psychology, anthropology, sociology, economics, political science, and human-computer interaction (HCI). These fields focus on the study of human behavior. For example, psychology researches the human mind, economics studies how people manage money, and HCI studies how people interact with computers and machines. These fields are interested in specific aspects of human nature and collect information about the motivations and theories that sufficiently describe human **behavior**.

There are other fields that also produce data based on measuring and representing human **activity**. For example, literature, art, theater, and sports, all produce their own data, but are different from the other fields in the kind of data they produce.

There are many other fields that are interested in the natural world and are not specifically concerned with large scale human systems and behavior, such as physics, biology, meteorology, geology, seismology and chemistry. These fields are often more concerned with **natural processes**, such as how forces interact, how weather works, and how the earth is formed and evolves.

The type of data you acquire from different fields will need to be studied and analyzed in different ways. For example, a sociology project analyzing migration patterns of immigrants will necessarily involve time series data, which we discuss in

R. Ball, B. Rague, *The Beginner's Guide to Data Science*, https://doi.org/10.1007/978-3-031-07865-1_2

Chap. 10. Other examples of time series data are analyzing the progression of stock market data, population trends, and election results based on year.

However, analyzing other kinds of data such as literature requires a completely different approach. Analyzing literature will most likely include natural language processing (NLP), a methodology we discuss in Chap. 9. Other examples of natural language processing include analyzing blogs, Twitter data, and emails.

Other data applications, like analyzing mineral deposits or predicting weather patterns, also demand different approaches based on the data. So, it is important to understand what kind of data you have before settling on the proper analysis strategy. Possessing or accessing some context and domain-specific knowledge (i.e., the kind of business you are dealing with) is vital to achieving any meaningful actionable insight.

Context is the area, domain, or situation that explains the data. For example, is it major league baseball data? Weather data? Astronomy data?

Unless you are an expert in a particular field then you most likely will not be gathering your own data. If you are interested in examining climate change then you will likely not start collecting your own dataset of barometric pressure, wind speeds, and precipitation amounts. You will probably work with experts in the area that have already completed some level of data acquisition from established weather stations around the world or at least have easy access to the data from these weather stations.

Examples of data creation include measuring interesting events, like how much rain fell yesterday, how much stress can be put on a pipe before leaks occur, how fast people memorize parts in a play, and other observable phenomena. Anything that is measurable and recorded can be captured as data.

2.2 IRB Approval

Generally, whenever you acquire data related to observations of **people** then you will be expected to secure IRB approval. The *IRB (Institutional Review Board)* is a group of people (or board) that evaluates the ethics of data collection involving people as research subjects. Any type of direct data retrieval from individuals, whether in the form of surveys, psychology tests, blood tests, or other means must be carefully examined and approved by the IRB associated with your organization **before** you begin your data collection. Even surveys, which may appear harmless on the surface but still are designed to extract personal information or opinions, must be approved.

The point of IRB approval is to protect people in the role of research subjects. The IRB is intended to protect people's privacy, their well-being, their identity, and their lives. The origins of the IRB dates back to when psychologists would apply electric shocks to people during experiments. The Stanford Prison experiment is a famous example of how a seemingly simple experiment conducted at Stanford University in 1971 got out of hand. In the experimental design, student volunteers were assigned roles as either fictional prisoners or fictional guards. The purpose and tenor of the experiment were derailed badly when the "guards" took their premise of authority too far. This famous experiment and other similar precarious research investigations led to the federal requirement that studies involving people need to be evaluated and approved by an objective oversight board prior to any experiment taking place.

Your organization will either have its own Institutional Review Board or will have defined a clear and consistent method to obtain approval from a review board that is external to your organization.

However, if you are collecting data that does not directly or indirectly affect people then you do not need to seek IRB approval. Some examples of this type of data collection include measuring the amount of rain, monitoring wind speed, or recording the speed of birds. In some instances, conducting analysis on anonymized data about people may not require IRB approval. However, we strongly recommend that you contact your IRB representative before conducting any type of analysis that involves the use of people as research subjects.

2.3 HCI: A Case Study

The myriad of details of data creation are beyond the scope of this book. However, you may find that reading a book on a particular field of study will be sufficient to enhance your knowledge of how best to accomplish your analysis. In the specific case of HCI (Human-Computer Interaction: the field where people are studied on how they interact with computers), you may find you need to gather additional information about how people navigate a particular website or piece of software to acquire a full understanding of the domain.