CONTENTS

1

Introduction

1.1 Why Data Science?

When you picked up this book to start reading, maybe you were hoping that we would once and for all answer the perennial question: What is data science? Allow us to disappoint you. Instead of providing a short and punchy definition, we are going to try to answer a different question, which we think might be even more important: Why data science? Rather than drawing a clear boundary for you around the topic of data science, this question lets us talk about the reasons that we think that data science has become important for neuroimaging researchers and researchers in other fields, and also talk about the effects that data science has on a broader understanding of the world and even on social issues.

One of the reasons that data science has become so important in neuroimaging is that the *amount* of data that you can now collect has grown substantially in the last few decades. Jack Van Horn, a pioneer of work at the intersection of neuroimaging and data science, described this growth in a paper he wrote a few years ago [Van Horn 2016]. From the perspective of a few years later, he describes the awe and excitement in his lab when, in the early 1990s, they got their first 4 gigabyte (GB) hard drive. A *byte* of data can hold 8 *bits* of information. In our computers, we usually represent numerical data (like the numbers in magnetic resonance imaging [MRI] measurements) using anything from 1 byte (when we are not too worried about the precision of the number) to 8 bytes, or 64 bits (when we need very high precision, a more typical case). The prefix giga-denotes a billion, so the hard drive that Van Horn and his colleagues got could store 4 billion bytes or approximately five hundred million 64-bit numbers. Back in the early 1990s, when this story took place, this was considered a tremendous amount of data, which could be used to store many sessions of MRI data. Today, given the advances that have been made in MRI measurement technology, and the corresponding advances in computing, this would probably store one or maybe two sessions of a high-resolution functional MRI (fMRI) or diffusion MRI (dMRI) experiment (or about half an hour of ultra high-definition video). These advances come hand-in-hand with our understanding that we need more data to answer the kinds of questions that we would like to ask about the brain. There are different ways that this affects the science that we do. If we are interested in answering questions about the

brain differences that explain cognitive or behavioral differences between individuals—for example, where in the neuroanatomy do individual differences in structure correspond to the propensity to develop mental health disorders—we are going to need measurements from many individuals to provide sufficient statistical power. Conversely, if we are instead interested in understanding how one brain processes stimuli that come from a very large set—for example, how a particular person's brain represents visual stimuli—we will need to make many measurements in that one brain with as many stimuli as practically possible from that set. These two examples (and there are many others, of course) demonstrate some of the reasons that data sets in neuroimaging are growing. Of course, neuroimaging is not unique in this respect, and additional examples of large data sets in other research domains are given subsequently.

Data Science Across Domains

This book focuses on neuroimaging, but large, complex, and impactful data sets are not unique to neuroscience. Data sets have been growing in many other research fields, and arguably in almost any research field where data is being collected. Here, we will present just a few examples, with a focus on data sets that contain images, and therefore use analysis tools and approaches that intersect with neuroimaging data science in significant ways.

One example of large data sets comes from *astronomy*. Though it has its roots in observations done with the naked eye, in the last few years, measurement methods and data have expanded significantly. For example, the Vera Rubin Observatory,[1] is one project centered on a telescope that has been built at an altitude of about 2,700 m above sea level on the Cerro Pachón ridge in Chile, and is expected to begin collecting data in early 2024. The experiment conducted by the Vera Rubin telescope will be remarkably simple: scanning about half of the night sky every few days and recording it with the highest-resolution digital camera every constructed. However, despite this simplicity, the data that will be produced is unique, both in its scale—the telescope will produce about 20 terabytes (TB) of data *every night*—and its impact. It is anticipated to help answer a range of scientific questions about the structure and nature of the universe. The data, which will eventually reach a volume of approximately 500 petabytes (PB), will be distributed through a specialized database to researchers all over the world, and will serve as the basis for a wide range of research activity in the decades to come. The project is therefore investing significant efforts to build out specialized data science infrastructure, including software and data catalogs.

Similarly, *earth science* has a long tradition of data collection and dissemination. For example, the NASA Landsat program, which focuses on remote sensing the surface of the earth from earth-orbiting satellites has existed since the 1970s. However, recent Landsat missions (the most recent is Landsat 8) have added multiple new types of

1. https://www.lsst.org/

measurements in addition to the standard remote sensing images produced previously. In total, Landsat and related projects produce more than 4 PB of image data every year. Similar to the efforts that are being made to harness these data in astronomy, there are significant attempts to create a data science ecosystem for analysis of these datasets in earth science. One interesting approach is taken by a project called Pangeo,[2] which has created a community platform for big data geoscience, collecting resources, software, and best practices and disseminating them to the earth science community.

Finally, neuroscience is a part of a revolution that is happening across the *biological sciences*. There are many different sources of information that are creating large and complex data sets in biology and medicine, including very large genomic data. However, one particularly potent data-generation mechanism involves new methods for imaging of tissue at higher and higher spatial and temporal resolution, and with more and more coverage. For example, high-throughput methods now can image an entire brain at the resolution of individual synapses. To share data sets related specifically to neuroscience, the US-based Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative has established several publicly-accessible archives. For example, the Distributed Archives for Neurophysiology Data Integration (DANDI) archive[3] shares a range of neurophysiology data, including electrophysiology, optophysiology (measurements of brain activity using optical methods), and images from immunostaining experiments. At the time of writing, the archive has already collected 440 TB of data of this sort, shared by researchers all over the US and the world.

2. https://pangeo.io/
3. https://dandiarchive.org/

In addition to the sheer volume of data, the *dimensionality* of the data is also increasing. This is in part because the kinds of measurements that we can make are changing with improvements in measurement technologies. This is related to the volume increase that we mentioned previously but is not the same. It is one thing to consider where you will store a large amount of data and how you might move it from one place to the other. It is a little bit different to understand data that are now collected at a very high resolution, possibly with multiple different complementary measurements at every time point, at every location, or for every individual. This is best captured in the well-known term "the curse of dimensionality," which describes how data of high dimensionality defy our intuitions and expectations based on our experiences with low-dimensional data. As with the volume of the data, the issue of dimensionality is common to many research fields and tools to understand high-dimensional data have been developed in many of these fields.

This brings us to another reason that data science is important. This is because we can gain a lot from borrowing methods from other fields in which data has become ubiquitous, or from fields that are primarily interested in data, such as statistics and some

parts of computer science and engineering. These fields have developed many interesting methods for dealing with large and high-dimensional data sets, and these *interdisciplinary* exchanges have proven very powerful. Researchers in neuroscience have been very successful in applying relatively new techniques from these other fields to neuroscience data. For example, *machine learning* methods have become quite popular in analyzing high-dimensional data sets and have provided important insights about a variety of research questions. Interestingly, as we will see in some of the chapters ahead, the exchange has not been completely one-sided, and neuroscientists have also been able to contribute to the conversation about data analysis in interesting and productive ways.

Another way in which data science has contributed to improvements in neuroscience is through an emphasis on *reproducibility*. Reproducibility of research findings requires ways to describe and track the different phases of research in a manner that would allow others to precisely repeat it. Thus the data needs to be freely available, and the code used to analyze the data needs to be available in such a way that others can also run it. This is facilitated by the fact that many of the important tools that are central to data science are *open-source* software tools that can be inspected and used by anyone. This means that other researchers can scrutinize the results of the research from top to bottom, and understand them better. It also means that the research can be more easily extended by others, increasing its impact.

Data science matters because once we start dealing with large and complex data sets, especially if they are collected from human subjects, the *ethical* considerations for use of the data and its potential harm to individuals and communities changes quite a bit. For example, considerations of potential harms need to go beyond just issues related to privacy and the protection of private information. Privacy is of course important, but some of the harms of large-scale data analysis may befall individuals who are not even in the data. For example, individuals who share certain traits or characteristics of the individuals who are included in the data could be affected, as could those individuals who were not included in the data. How large biomedical data sets are being collected and the implications of the decisions made in designing these studies have profound implications for how the conclusions apply to individuals across society.

Taken together, these factors make data science an important and central part of contemporary scientific research. However, learning about data science can be challenging, even daunting. How can neuroimaging researchers productively engage with these topics? This brings us to our next subject: the intended audience for this book.

1.2 Who This Book Is For

This book was written to introduce researchers and students in a variety of research fields to the intersection of data science and neuroimaging. Neuroimaging gives us a view into the structure and function of the living human brain, and it has gained a solid foothold in research on many different topics. As a consequence, people who use neuroimaging to study the brain come to it from many different backgrounds and with many

different questions in mind. We wrote this book thinking of a variety of situations in which additional technical knowledge and fluency in the language of data science can provide a benefit to individuals—whether it be in rounding out their training, in enabling a research direction that would not be possible otherwise, or in facilitating a transition in their career trajectory. The researchers we have written this book for usually have some background in neuroscience and this book is not meant to provide an introduction to neuroscience. When we refer to specific neuroscience concepts and measurements we might explain them, but for a more comprehensive introduction to neuroscience and neuroimaging, we recommend picking up another book (of which there are many); all chapters, including this one, end with an "Additional Resources" section that will include pointers to these resources. We will also not discuss how neuroimaging data comes about. Several excellent textbooks describe the physics of signal formation in different neuroimaging modalities and considerations in neuroimaging data collection and experimental design. Finally, we will present certain approaches to the analysis of neuroimaging data, but this is also not a book about the statistical analysis of neuroimaging data. Again, we refer readers to another book specifically on this topic. Instead, this book aims to give a broad range of researchers an initial entry point to data science tools and approaches and their application to neuroimaging data.

1.3 How We Wrote This Book

This book reflects our own experience of doing research at the intersection of data science and neuroimaging and it is based on our experience working with students and collaborators who come from a variety of backgrounds and have different reasons for wanting to use data science approaches in their work. The tools and ideas that we chose to write about are all tools and ideas that we have used in some way in our research. Many of them are tools that we use daily in our work. This was important to us for a few reasons: the first is that we want to teach people things that we find useful. Second, it allowed us to write the book with a focus on solving specific analysis tasks. For example, in many of the chapters, you will see that we walk you through ideas while implementing them in code, and with data. We believe that this is a good way to learn about data analysis because it provides a connecting thread from scientific questions through the data and their representation to generating specific answers to these questions. Finally, we find these ideas compelling and fruitful. That is why we were drawn to them in the first place. We hope that our enthusiasm for the ideas and tools described in this book will be infectious enough to convince the readers of their value.

1.4 How You Might Read This Book

More important than how we wrote this book, however, is how we envision you might read it. The book is divided into several parts.

Data science operates best when the researcher has comprehensive, explicit, and fine-grained control. The first part of the book introduces some fundamental tools that give

users such control. These serve as a base layer for interacting with the computer and are generally applicable to whatever data analysis task we might perform. Operating with tools that give you this level of control should make data analysis more pleasant and productive, but it does come with a bit of a learning curve that you will need to climb. This part will hopefully get you up part of the way, and starting to use these tools in practice should help you to get up the rest of the way. We will begin with the Unix operating system and the Unix *command line interface* (in Chapter 2). This is a computing tool with a long history, but it is still very well suited for flexible interaction with the computer's operating system and file system, as its robustness and efficiency have been established and honed over decades of application to computationally intensive problems in scientific computing and engineering. We will then (in chapter 3) introduce the idea of *version control*—a way to track the history of a computational project—with a focus on the widely used git version control system. Formal version control is a fundamental building block of data science as it provides fine-grained and explicit control over the versions of software that a researcher works with, and also facilitates and eases collaborative work on data analysis programs. Similarly, computational *environments* and computational *containers*, introduced in chapter 4, allow users to specify the different software components that they use for a specific analysis while preventing undesirable interactions with other software.

The book introduces a range of tools and ideas, but within the broad set of ways to engage with data science, we put a particularly strong emphasis on programming. We think that programming is an important part of data science because it is a good way to apply quantitative ideas to large amounts of data. One of the major benefits of programming over other approaches to data analysis, such as applications that load data and allow you to perform specific analysis tasks at the click of a button, is that you are given the freedom to draw outside the lines: with some effort, you can implement any quantitative idea that you might come up with. Conversely, when you write a program to analyze your data, you have to write down exactly what happens with the data and in what order. This supports the goal of automation; you can run the same analysis on multiple data sets. It also supports the goal of making the research reproducible and extensible. That is because it allows others to see what you did and repeat it in exactly the same way. That means programming is central to many of the topics we will cover. The examples that are provided will use neuroimaging data, but as you will see, many of these examples could have used other data just as well. Note that this book is not meant to be a general introduction to programming. We are going to spend some time introducing the reader to programming in the *Python programming language* (starting in chapter 5; we will also explain specifically why we chose the Python programming language for this book), but for a gentler introduction to programming, we will refer you to other resources. However, we will devote some time to things that are not usually mentioned in books about programming but are crucially important to data science work, such as how to test software and profile its performance, and how to effectively share software with others.

In the next two parts of the book, we will gradually turn towards topics that are more specific to data science in the context of scientific research, and neuroimaging in particular.

First, we will introduce some general-purpose scientific computing tools for numerical computing (in chapter 8), data management and exploration (chapter 9), and data visualization (chapter 10). Again, these tools are not neuroimaging-specific, but we will focus in particular on the kinds of tasks that will be useful when working with neuroimaging data. Then, in the next part, we will describe in some detail tools that are specifically implemented for work with neuroimaging data: the breadth of applications of Python to neuroimaging data will be introduced in chapter 11. We will go into further depth with the NiBabel software library, which gives users the ability to read, write, and manipulate data from standard neuroimaging file formats in chapter 12 and chapter 13.

The last two parts of the book explore in more depth two central applications of data science to neuroimaging data. In the first of these (Part V), we will look at image processing, introducing general tools and ideas for understanding image data (in chapter 14), and focusing on tasks that are particularly pertinent for neuroimaging data analysis; namely, image segmentation (in chapter 15) and image registration (in chapter 16). Finally, the last part of the book will provide an introduction to the broad field of machine learning (Part VI). Both of these applications are taken from fields that could fill entire textbooks. We have chosen to provide a path through these that emphasizes an intuitive understanding of the main concepts, with code used as a means to explain and explore these concepts.

Throughout the book, we provide detailed examples that are spelled out in code, with relevant data sets. This is important because the ideas we will present can seem arcane or obscure if only their mathematical definitions are provided. We feel that a software implementation that lays out the steps that are taken in analysis can help demystify them and provide clarity. If the description or the (rare) math that describes a particular idea is opaque, we hope that reading through the code that implements the idea will help readers understand it better. We recommend that you not only read through the code but also run the code yourself. Even more important to your understanding, try changing the code in various ways and rerunning it with these changes. We propose some variations in the sections labeled "Exercise" that are interspersed throughout the text. Solutions to these exercises are provided in Appendix 1. Some code examples are abbreviated by calling out to functions that we implemented in a companion software library that we named `ndslib`. This library includes functions that download and make relevant data sets available within the book's chapters. We provide a reference to functions that are used in the book in Appendix 2 and we also encourage the curious reader to inspect the code that is openly available on GitHub.[4]

1.4.1 Jupyter

One of the tools that we used to write this book, and that we hope that you will use in reading and working through this book, is called Jupyter.[5] The Jupyter notebook is an

4. https://github.com/neuroimaging-data-science/ndslib
5. https://jupyter.org/

application that weaves together text, software, and results from computations. It is very popular in data science and widely used in scientific research. The notebook provides fields to enter text or code—these are called *cells*. Code that is written in a code cell can be sent to an interactive programming language interpreter for evaluation. This interpreter is referred to as the *kernel* of this notebook. For example, the kernel of the notebook can be an interactive Python session. When you write a code cell and send it to the kernel for evaluation, the Python interpreter runs the code and stores the results of the computation in its memory for as long as the notebook session is maintained (so long as the kernel is not restarted). That means that you can view these results and also use these results in the following code cells. The creators of the Jupyter notebook, Brian Granger and Fernando Pérez, recently explained the power of this approach in a paper that they wrote [Granger and Pérez 2021]. They emphasize something that we hope that you will learn to appreciate as you work through the examples in this book, which is that data analysis is a collaboration between a person and their computational environment. Like other collaborations, it requires a healthy dialogue between both sides. One way to foster this dialogue is to work in an environment that makes it easy for the person to perform a variety of different tasks: analyze data, of course, but also explore the data, formulate hypotheses and test them, and also play. As they emphasize, because the interaction with the computer is done by writing code, in the Jupyter environment a single person is both the author and the user of the program. However, because the interactive session is recorded in the notebook format together with rich visualizations of the data (you will learn how to visualize data in chapter 10) and interactive elements, the notebooks can also be used to communicate their findings with collaborators or publish these results as a document or a webpage. Indeed, most of the chapters of this book were written as Jupyter notebooks that weave together explanations with code and visualizations. This is why you will see sections of code, together with the results of running that code interwoven with explanatory text. This also means that you can repeat these calculations on your computer and start altering, exploring, and playing with them. All of the notebooks that constitute the various chapters of this book can also be downloaded from the book website.[6]

1.4.2 Setting Up

To start using Jupyter and to run the contents of the notebooks that constitute this book, you will need to set up your computer with the software that runs Jupyter and also with the software libraries that we use in the different parts of the book. Setting up your computer will be much easier after you gain some familiarity with the set of tools introduced in the next part of the book. For that reason, we put the instructions for setup and for running the code in Section 4.3, at the end of the chapter that introduces these tools. If you are keen to get started, read through the next chapter and you will eventually reach these instructions.

6. http://neuroimaging-data-science.org

1.5 Additional Resources

For more about the fundamentals of MRI, you can refer to one of the following:

D McRobbie, E Moore, M Graves, and M Prince. *MRI from Picture to Proton* (3rd ed.). Cambridge University Press, 2017.

S A Huettel, A W Song, and G McCarthy. *Functional Magnetic Resonance Imaging*. Sinauer Associates, 2014.

For more about the statistical analysis of MRI data, we refer the readers to the following:

R A Poldrack, J A Mumford, and T E Nichols. *Handbook of Functional MRI Analysis*. Cambridge University Press, 2011.

This book will touch on data science ethics in only a cursory way. This topic deserves further reading and there are fortunately several great books to read. We recommend the following two:

C D'Ignazio and L Klein. *Data Feminism*. MIT Press, 2020.

C O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

For more about the need for large data sets in neuroimaging, you can read some of the papers that explore the statistical power of studies that examine individual differences [Button et al. 2013]. In a complementary opinion, Thomas Naselaris and his colleagues demonstrate how sometimes we don't need many subjects, but instead would rather opt for a lot of data on each individual [Naselaris et al. 2021].

INDEX