About this book

Contents

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’re going to try to answer a different question, which we think may actually be even more important, and is the title of this section: why data science? Rather than drawing a clear boundary for you around the topic of data science, it lets us talk about the reasons that we think that data science has become important for neuroimaging researchers and for researchers in other fields, and also to talk about the effects that data science has on 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 great 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 1990’s, they got their first 4 giga-byte (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 MRI measurements) using anything from one 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 500 million 64-bit numbers. Back in the early 1990’s, when this story took place, this was considered a tremendous amount of data, that 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 answer 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. On the other hand, 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 large as possible a set of stimuli from that set. These two examples (and there are others) demonstrate some of the reasons that datasets in neuroimaging are growing. Of course,

neuroimaging is not unique in this respect. Datasets have been growing in many other research fields – arguably in most research fields in which data about the world is being collected.

In addition to the sheer volume of the data, the **dimensionality** of the data is also increasing. This is in part because, as we mentioned above, 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 above, but is not exactly the same. It's 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's a little bit different to understand data that is now collected at a very high resolution, possibly with multiple different complementary measurements at every time point, or at every location or for every individual. This is best captured in the well-known term "the curse of dimensionality", which describes the way in which data of high dimensionality defies our intuitions and expectations based on our experience with low-dimensional data. As with the volume of the data, the issue of dimensionality is common to many research fields as well and tools to understand high-dimensional data have been developed in many of these fields.

Which 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 a lot of interesting methods for dealing with large and high-dimensional datasets, 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 datasets and have provided important insights about a variety of research questions. Interestingly, as we'll see in some of the chapters ahead, the exchange hasn't 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. It means that the data needs to be freely available and that the code that was used to analyze the data needs to be available in a way that others can also run it. This is also 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 it better. It also means that the research can be more easily extended by others, increasing its impact.

Data science is also important because once we start dealing with large and complex datasets, especially if they are collected from human subjects, the **ethical** considerations for use of the data and the way we think about the potential harms to individuals and communities from the research that we do changes quite a bit. For example, considerations of potential harms need to go beyond just issue related to privacy and protection of private information. These are 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. Another potential issue comes from not considering individuals who were not included in the data. The way in which large biomedical datasets are being collected and the implications of the decisions made in designing these studies have profound implications for the ways in which 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 topic – the intended audience for this book.

1.2. Who this book is for

Neuroimaging measurements give us a view into the structure and function of the living human brain, and they've 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 kinds of questions in mind. This book was written with the goal of introducing researchers and students in a variety of research fields to the intersection of data science and neuroimaging. One way to demonstrate who might benefit from this book is to talk about a few of the types of people that we had in mind while we wrote it. Here are a few (fictional!) characters who represent these students and collaborators:

Stennie Soleanna: Stennie is a postdoc in a research institute. In their previous research, they used a lot of Matlab software tools, and some proprietary software. In the new research project that they are undertaking, they would like to use a machine learning approach to classify subjects in a dataset that they are collecting, and to interpret the features that help them make this classification. Their previous training did not include classes in machine learning, and while they have read a lot of papers that used machine learning methods to analyze neuroimaging data, they are not confident that they would be able to execute their planned project. Stennie will receive an overview of the main differences between machine learning techniques and traditional statistical approaches. They will also receive an overview of machine learning techniques and a hands-on introduction to using these machine learning techniques with neuroimaging data.

Mona McHale: Mona is a first-year graduate student in an interdisciplinary neuroscience program. Before she came to graduate school, she was a research assistant in a lab. She worked with an experienced graduate student, who, after graduating from the program decided to go work in industry. The graduate student told Mona that they were grateful for all the technical skills that they acquired during their time in a PhD program. Mona is not sure where she might go after graduate school. She enjoys thinking about basic research problems . In this book, Mona will learn about tools that are useful to her in the projects she will pursue in graduate school, but that are also industry standards in data science outside of academia. Incorporating these tools into her work in graduate school will allow Mona to keep her options open for when she finishes her graduate program.

Rick Rastepappe: Rick is an instructor at Psychology Department in a community college in the US Southwest. During his research training as a graduate student, Rick was in large and well-endowed R1 institutes, and he did experiments studying higher cognitive function in healthy human subjects using functional MRI. Now, that he has changed career stages, and is an institution that is quite different from the ones he trained in, he would like to initiate a new research program that relies on the many open datasets. However, while his research training prepared him very well to design experiments and execute them, and to use standard tools to analyze the data, it did not prepare him for the tasks of wrangling large datasets, and designing novel analysis methods to tackle the questions that he would like to ask in this new stage. Reading this book and working through the exercises will give Rick ideas for the kinds of things that he can do with the large open datasets that he is preparing to analyze. It will also give him new tools to keep track of the computational work that he does and to share the software the he writes.

Gautam Gont: Gautam is a researcher at a big pharma company. He joined his company after a few years in a graduate program in medical and clinical sciences, which he left having completed all but his dissertation. He participates in data collection and data analysis related to studies on drugs that his colleagues in the company develop for treatment of multiple sclerosis. Gautam's work involves managing data and manual quality control of the images. Reading the book, Gautam will learn about standards for neuroimaging data management and best practices in sharing data. He will also learn about software libraries that he can use to build data visualization tools that can speed up the manual quality control work that he does and also automate some aspects of it.

As you can see from these descriptions, we are thinking of a variety of situations in which additional technical knowledge and a fluency in the language of data science can provide some 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. And there are probably other prototypes we did not include here. Notice also that all of the researchers that we describe here have some background in neuroscience and might also have some experience in programming. This is because 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. There are several excellent textbooks that describe the physics of signal formation in different neuroimaging modalities and considerations in neuroimaging data collection and experimental design, a couple of which we mention below. Finally, we will present certain approaches to analysis of neuroimaging data, but this is also not really 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 people like the ones that we described 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 a variety of 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 own research. Many of them are tools that we use on a daily basis in our work. This was important to us for a few reasons: the first is that we want to teach people things that we ourselves 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 its representation to implementing specific answers to these questions. Finally, we find these ideas compelling and fruitful. That's why were drawn to them in the first place. We hope that our enthusiasm about 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, generally applicable whatever particular 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 Section 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 filesystem, 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 Section 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**, next introduced in Section 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 programmming. We think that programming is an important part of doing data science, because it is a really 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. On the other hand, 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 datasets. It also supports the goal of making the research reproducible and extensible. That's because it allows others to see what you did and to repeat what you did in exactly the same way. That means that 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. The 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 Section 5; we will also explain there why we chose specifically the Python programming language for this book), but for a gentler introduction to programming, we will refer you to other resources. On the other hand, we will devote some time to things that are not usually mentioned in books about programming, but are crucially important to data science work: 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 Section 8), data management and exploration (Section 9), and data visualization `viz`). Again, these tools are not neuroimaging-specific, but we will focus in particular on the kinds of tasks that will be useful when doing data science work 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 Nibabel software library (in Section 10), which gives its users the ability to read, write and manipulate data from standard neuroimaging file formats, and the Brain Imaging Data Structure (in Section 11), which is used to organize and describe neuroimaging datasets.

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, we will look at image processing, introducing general tools and ideas for understanding image data (in Section 12), and focusing on tasks that are particularly pertinent for neuroimaging data analysis: image segmentation (in Section 13) and image registration (in Section 14). Finally, the last part of the book will provide an introduction to the broad field of machine learning (`ml`). Both of these applications are taken from fields that could fill entire text-books. We have chosen to provide a path through these that emphasizes intuitive understanding of the main concepts, with code used as a means to explain and to explore these concepts.

Throughout the book, we provide detailed examples that are spelled out in code, with relevant datasets. 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 an analysis can help demystify them and provide clarity. If the description or the (rare) math that describes a particular idea are opaque, we hope that

reading through the code that implements the idea can help understand it better. We really 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 *Exercise* sections that are interspersed throughout the text. 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 datasets available within the books chapters, and we encourage the curious reader to inspect the code that is openly available on GitHub.

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. The Jupyter notebook is an 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 is 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 store 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 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 anlaysis 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 more how to visualize data in `viz`) and interactive elements, the notebooks can also be used to communicate about 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 interleaved with explanatory text. This also means that you can repeat these calculations on your own computer, and start altering them, exploring them and playing with them. All of the notebooks that constitute the various chapters of this book can also be downloaded from the book website

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 acquaintance 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, after 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 for setup at its conclusion.

1.5. Additional resources

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

- McRobbie, D., Moore, E., Graves, M., & Prince, M. (2017). MRI from Picture to Proton (3rd ed.). Cambridge University Press.
- Huettel, S.A., Song, A.W & McCarthy, G. Functional Magnetic Resonance Imaging (2014). Sinauer Associates.

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

- Poldrack R.A., Mumford J.A. & Nichols T.E. (2011). Handbook of Functional MRI Analysis. Cambridge University Press.

The 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:

- D'Ignazio C. & Klein (2020). Data Feminisim. MIT Press; also available at https://data-feminism.mitpress.mit.edu/
- Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (2016). O'Neil C. Broadway Books.

For more about the need for large datasets in neuroimaging, you can read some of the papers that explores 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 in each individual [Naselaris *et al.*, 2021].
