CHAPTER 3

fMRI GROWS UP

Measurement is central to scientific research, and breakthroughs have often been driven by the development of new tools to measure the world. However, any new measurement tool has to be validated, to make sure that it is actually measuring what it claims to measure. In the case of fMRI, there was initial evidence that fMRI signals truly reflected brain activity, based on the fact that fMRI results aligned with what we already knew from neurological studies and from animal research: visual stimulation in different parts of the visual field causes activity in the appropriate areas in the visual cortex, motor actions lead to activity in the motor cortex, and so on. However, we still didn't have any direct evidence that linked fMRI signals to the activity of neurons, which meant that we were using the method without really knowing exactly what it was measuring. For this reason, many neuroscientists (particularly those whose research involved recording the activity of neurons directly in animals) were quite dismissive of fMRI—an attitude I experienced first-hand when I started interviewing for faculty jobs in 1998. Until we had evidence for a direct relationship between fMRI signals and the activity of neurons, the road for fMRI research would remain rocky.

Linking fMRI to Neurons

Soon after the discovery of fMRI, researchers started trying to understand the relationship between BOLD fMRI signals and the firing of neurons. If there is one thing that makes a system easier to understand scientifically, it is when the relationship between

the inputs and outputs of the system is straightforward, which scientists call "linearity." A linear system is one in which the output of the system can be described by transforming and adding up the input. For example, my checking account is a linear system, in the sense that the balance is determined by adding up all of the individual transactions. In fact, it is a particular kind of linear system, that we call "time invariant," because it doesn't matter when the individual transactions occur—they still all get added together. Many systems in the world are linear—or at least, as scientists say, "linear to a first approximation," meaning that we can pretend they are linear and still do a pretty good job of explaining them, even if the model is not perfect. On the other hand, there are also many systems that simply can't be treated as linear. An avalanche occurs when too much snow falls on the side of a mountain, but you can't divide that snow into 10 parts and expect to get 10 small avalanches. In a linear system, more is more; in a nonlinear system, more is different.

In the years after the discovery of fMRI, researchers began to try to determine whether the fMRI signal was a linear function of the activity of neurons. That is, if the neurons fire twice as much, is the fMRI response twice as large? This question turns out to be quite challenging to answer because we can't directly measure how much the neurons are firing in a person—all we know is what kind of task we have provided the subject with, and how the fMRI signal changes. To study this question, Geoff Boynton and his colleagues at Stanford took advantage of the fact that neuroscientists have learned a lot about how neurons respond to different types of visual stimuli in the visual system of the monkey, and human brains work a lot like monkey brains. Boynton and his fellow postdoc Steve Engel each spent hours lying in the MRI scanner staring at checkerboard patterns as they flashed eight times a second, in order to measure how the response of their visual cortexes changed with the amount of time the checkerboard was presented for and the relative contrast of the black and white parts of the checkerboard—both factors known to affect the firing of neurons in the visual cortex in a well-characterized way. What they found was that the signal went up as the stimulus got longer and as the contrast was higher, exactly

as they expected from the monkey research. The critical test of linearity was to see whether the response to a longer stimulus could be predicted just by adding together the responses to the shorter stimuli after shifting them in time, which is a critical prediction of the linear model. It worked—not perfectly, but well enough for most people to think that the linear model is a reasonable way to analyze fMRI data. The assumption of linearity is now fundamental to nearly all of the ways in which fMRI data are analyzed.

The work of Boynton and others provided a much tighter link between fMRI signals and the activity of neurons, but there was still a missing link: until someone recorded both single neuron activity and fMRI at the same time in the same brain, it would not be possible to say for sure that fMRI was reflecting the activity of single neurons. This challenge was taken up by Nikos Logothetis, a neuroscientist from the Max Planck Institute of Biological Cybernetics in Tübingen, Germany, who was an expert in studying how neurons respond in different parts of the monkey's visual system. He was becoming increasingly interested in more complex aspects of visual perception, such as our ability to pick out objects from cluttered backgrounds, and he realized that to understand these more complex phenomena, he needed to be able to study the entire system rather than just a few neurons, so when fMRI came around he started working on finding a way to do fMRI studies with monkeys while also recording the electrical activity of neurons. This was in some sense the "holy grail" of fMRI, since it combined the whole-brain breadth of fMRI with the precision of single neuron activity recording. To appreciate just how difficult this is, keep in mind that electrical activity of neurons is recorded using electrodes connected to small wires, which register the electrical activity of the cell. Those changes are fairly small, on the order of microvolts (thousandths of a volt). It's also important to know that when metal is placed in an MRI scanner, the changes in the magnetic field used to create MRI images will cause current to start flowing through the metal, and that current can be much larger than the changes caused by neurons, making it

almost impossible to see the tiny neuronal signals without some very sophisticated signal processing techniques. Logothetis and his team spent several years engineering a solution to these problems, and by 2000 they were able to successfully record from neurons in the brain of an anesthetized monkey while also performing fMRI.

The results from Logothetis's study provided direct evidence for a relationship between neural activity and fMRI,[1] and gave an "all clear" signal to many neuroscientists who had been leery of adopting the new technique without knowing precisely what it measured. The results also provided fMRI researchers like myself a way to answer the lingering questions from some of our colleagues about what fMRI actually measures. What Logothetis and his team did was to measure fMRI signals and neuronal activity simultaneously while the animal was presented with a flashing checkerboard. It may seem surprising that researchers could obtain useful data about brain activity from an anesthetized animal, but in fact the neurons in the visual cortex respond similarly between wakefulness and anesthesia. They saw clear evidence of activity in the visual cortex using fMRI, and this activity increased as the black/white contrast of the checkerboard was turned up (just as it had in Boynton's and Engel's brains). When they measured the firing of individual neurons they saw that it was also directly related to these changes in the BOLD signal; however, they found an even tighter link with something called the "local field potential" (LFP for short). The LFP is a measurement of changes in the electrical signal that happen more slowly than a neuron fires, and are thought to reflect the inputs to neurons rather than their firing. Logothetis's results have been built upon since 2001, most recently using fMRI in rats along with a technique called "optogenetics" that allows researchers to turn on specific kinds of neurons using light.[2] At this point, it is widely accepted that fMRI signals are a direct reflection of the activity of neurons, particularly of the inputs to neurons rather than their firing per se, even though there is still much to learn about exactly how it works.

Finding Modules in the Brain

Since the middle of the twentieth century, most neuroscientists have agreed that at least some psychological functions rely on specific areas in the brain. The advent of fMRI brought with it the promise of mapping out these localized functions with a remarkable degree of accuracy, and one of the first places this was used was in the study of how visual objects are recognized—specifically, faces. We already knew that the recognition of visual objects relies upon the bottom ("inferior") part of the temporal lobe, because damage to this region can leave people unable to recognize objects visually, even though they still have knowledge about the object and can identify it by touch. There was also some reason to believe that faces are processed differently from other objects, because there are rare cases of a syndrome called "prosopagnosia" in which the person is unable to recognize faces but is still able to recognize other kinds of objects. Early work using PET, particularly by the McGill University neuroscientist Justine Sergent,[3] had also shown that there were areas in the temporal lobe that were more involved in face processing than that of other types of objects.

Nancy Kanwisher is a neuroscientist at MIT who was captivated by the potential of fMRI to uncover the biology of the mind.[4] Using the same MRI scanner at the MGH-NMR Center on which Belliveau and Kwong had done their original fMRI studies, Kanwisher (along with her trainees Josh McDermott and Marvin Chun, who have both gone on to have impressive careers themselves) found a brain area that responded much more for faces compared with other types of stimuli.[5] The area was present in almost everyone they looked at (12 out of 15 people) in a location called the "fusiform gyrus" (see color plate 3), which runs along the bottom of the temporal lobe. This area didn't *only* respond to faces, but when they examined the response in this area to many different types of objects, in each case there was at least twice as much activity in the fusiform gyrus for faces compared with the other objects. They also showed that it was remarkably consistent over time—one of the participants (who was actually Nancy Kanwisher herself) was scanned multiple

times over the course of six months, and in each case the pattern of activity was highly similar. They christened this area the "fusiform face area," or FFA for short. In the ensuing two decades, we have learned a lot more about the FFA, including the fact that there is not just one but several regions along the bottom of the temporal lobe that respond to faces (see color plate 3). We also now know that these areas are essential for face processing; this is the same area that was stimulated in the epileptic patient described in chapter 1 that so disrupted his ability to perceive faces. Kanwisher and others have also gone on to show that there are other parts of the temporal lobe that respond in a selective way for other types of stimuli, including body parts, words, and scenes.

Kanwisher and her colleagues made some very strong claims about the localization of face perception, which did not sit well with Isabel Gauthier, a vision researcher who had studied how our ability to recognize visual objects changes with practice. Her studies had trained people to identify a kind of artificial object called "greebles," which look somewhat like alien garden gnomes. Because greebles could be created using computer graphics to have many different kinds of visual features, they could be used to study how people become expert at distinguishing between them. What Gauthier found in her study was that people could get better at recognizing individual greebles with practice, and as this happened, the right fusiform area started to become activated by greebles just like it is for faces. This led her to pose a new theory: the FFA is not actually a "face area," but more like an "expertise area" (or, in keeping with the original acronym, a "flexible fusiform area"), becoming engaged whenever people recognize objects that they have lots of expertise with, especially when they have to distinguish different individuals within the category.

Gauthier set out to test this theory further by examining people who were highly expert at recognizing specific kinds of objects: bird watchers and car experts. There was already some reason to think that visual expertise involves different areas of the brain from those used in recognition of regular objects; in a couple of rare cases, a bird watcher lost the ability to recognize birds and a car aficionado lost the ability to recognize specific

cars, without impairing their ability to recognize other kinds of objects. What Gauthier did in her study was to recruit a set of volunteers who were either bird watchers or self-proclaimed car experts, and then showed them a number of different types of objects, including faces, birds, and cars, along with other objects. The results confirmed the expertise hypothesis: the car experts showed a response to cars in the FFA, while the bird watchers showed a response to birds in that area; importantly, the response to the nonexpert objects was much lower. This seemed to conclusively show that the FFA was specialized for visual expertise, not for faces per se. However, Kanwisher and her collaborators were not convinced, raising a number of critiques of the data and their interpretation. Around the same time, however, another researcher came along whose work would both throw a new wrench into the FFA debate and light the path for a whole new way to think about fMRI data.

First Steps toward Decoding the Brain

Jim Haxby is in some ways the polar opposite of Nancy Kanwisher. Nancy is outwardly intense and energetic, while Jim's intensity lies beneath a soft-spoken and mellow surface. Haxby started his career at the National Institutes of Health, where he did research starting in the early 1990s using PET to study how different types of objects are processed in the brain; in fact, it was partly this work that inspired Kanwisher to look in the temporal lobe for face-related activity. When his group later began to use fMRI, he was captivated by the results from Kanwisher showing a seemingly very specific response to faces in the fusiform gyrus, but he did not believe their interpretation.[6] As his group began to do their own fMRI studies of face perception, he was struck by how these areas that were seemingly "specific" for faces actually showed substantial responses to other types of stimuli as well, bringing into question just how specific they were. Haxby and Kanwisher saw the same data, but came to strikingly different conclusions.

Haxby had an idea that would turn out to revolutionize how we analyze fMRI data. Until that point, fMRI researchers had looked only at "activation"—that is, how much more active a region

was in one condition versus another. This information is usually presented in bright maps showing the regions where there was strong enough signal to be considered "statistically significant"—that is, where we are fairly sure that the difference in activity is not simply due to random fluctuations. Haxby had the idea to instead look at the entire pattern of activity across a region and ask whether it differed between conditions. As an analogy, think of a crowd responding to three candidates in a political debate by clapping (see figure 3.1). One could ask whether there is a section of the crowd that overall claps more for one candidate than the other two; this would be the equivalent of the activation analysis, finding the region of the crowd that is "selective" for one particular candidate. In this example, section 1 of the crowd claps much more strongly for the candidate denoted by the circle than for either of the other two; we would say that it was *selective* for the circle candidate. However, one can also flip this question on its head and ask: Can we tell which candidate is speaking, based on the pattern of clapping across the entire room? We refer to this as "decoding," and it has become a central part of fMRI analysis, as we will discuss throughout much of the rest of this book.

Understanding how decoding works requires a bit more detail about the nature of fMRI data. When we collect brain images using fMRI, those images are made up of a large number of three-dimensional cubes that we call "voxels" (think of pixels on a screen, but in three dimensions), each of which is 1–3 mm on each side. Each of these voxels contains millions of neurons, and the fMRI signal reflects an average of the activity across those neurons. If a voxel includes lots of neurons that fire when a face is present, then that voxel will show a high fMRI signal for faces, but it might also have a smaller number of neurons that prefer other types of objects, say houses or chairs. To push the crowd analogy a bit further, let's pretend that we are measuring the response in each section of the crowd using a microphone, which tells us on average how much clapping there is in that section of the crowd. In this example, each person represents a neuron, and each section of the crowd with a microphone represents a voxel. As in the bottom section of figure 3.1, we will probably see that there are some sections that clap much louder (on average)

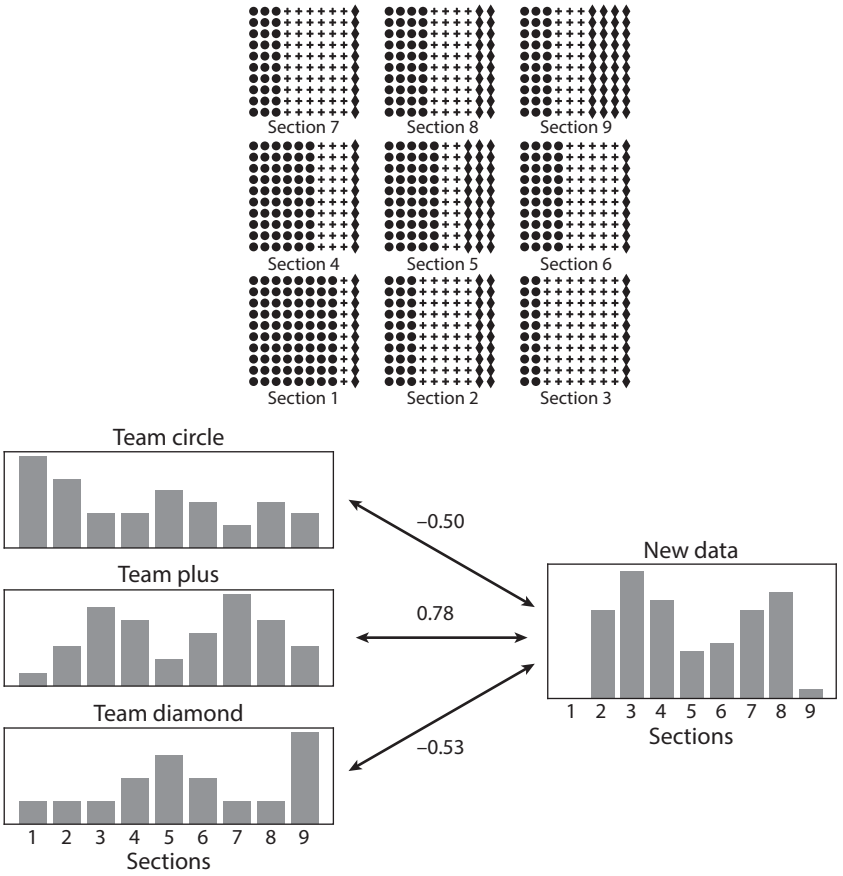

Figure 3.1. An example of how fMRI decoding works, using the analogy of the audience reaction to three different political candidates. The *top panel* depicts nine sections of the audience (which are analogous to voxels in fMRI); each section is made up of many individuals, each of whom claps for only one candidate (denoted by three different shapes: circle, plus, and diamond). These individuals represent the neurons within a voxel in fMRI. The *bottom left panel* shows a graph of the relative amount of clapping for each of the candidates across the nine sections; you can see that each candidate has a very different pattern of clapping across the sections. Now pretend that we have a new measurement of clapping (*bottom right*) and we want to decode which candidate is speaking. We can compute the correlations between the new pattern and each of the known patterns (which are the numbers presented next to the arrows); in this case, the new pattern is most highly correlated with the pattern that we observed when candidate plus was speaking, and thus we would predict that candidate plus was speaking when these data were collected.

for one candidate versus the others; that is the analog of higher activation for faces versus other types of stimuli. You can also imagine, however, that scattered across the crowd are people who have different preferences for the three candidates, and thus even in the areas that are not highly "selective" for one candidate, one might be able to see a difference in the pattern of clapping across those areas that would give a clue as to which candidate was speaking. The figure shows how, given a new measurement of the level of clapping across each section, we could decode who was speaking based on how similar the pattern across sections was to each of the known patterns.

Haxby used this idea to test his hypothesis that the processing of different types of visual information is "distributed" across the temporal lobe; that is, even though there are some areas that are highly responsive to faces versus other objects, those are not the only parts of the visual system that are processing information relevant to faces. In his study, he scanned volunteers while they looked at pictures of many different kinds of objects, including faces, houses, chairs, bottles, cats, shoes, and scissors.[7] Each person went through 10 different scans (which we call "runs") in which they saw each of the different types of object. To test whether he could decode the kind of object that a person was seeing, he first took the data from every other run (say, the odd runs) and measured the response in each voxel to each of the different types of object. Then, he took the data from the even runs, and for each one asked the following question: Given this pattern of activity, which one of the patterns is it most similar to from the odd runs? This allowed him to decode what kind of object the person was seeing. For example, if the current pattern (from one of the even runs) is most similar to the average pattern of activity for cats in the odd runs, then he would predict that the person was viewing a cat when those data were collected on the even run.[8] When Haxby applied this method to his data, he saw that he was able to decode what kind of object the volunteers were looking at, with an accuracy above 90%—in fact, for faces the accuracy was 100%! To test his claim that the processing of objects is distributed across the temporal lobe rather than localized, he looked at whether this decoding still worked even in the areas

where activity was not selective for the specific object type; for example, could he distinguish between faces and other types of objects, even after he removed the voxels that responded more to faces than other object types? The answer was yes—he could still tell with very high accuracy when a person was looking at a face, even when he only looked within the voxels that were *not* part of the face area. As with most scientific debates, no single study provides a conclusive answer, and the controversy about the localization of face processing has continued since Haxby published his original paper. But, most importantly, his paper introduced the field to the idea of decoding mental content from fMRI data.

In chapter 1 I introduced the concept of "reverse inference"—the idea that one can tell what a person is thinking by looking at which brain areas are active—and explained how this kind of inference was problematic when it was applied in the *New York Times* op-ed about the 2008 election. As you have read this chapter, it may have occurred to you that the idea of reverse inference is really not very different from the concept of decoding that was seen in the work of Jim Haxby, and you would be correct: in each case we are using neuroimaging data to try to infer the mental state of an individual. The main difference is that the reverse inference that I ridiculed from the *New York Times* was based not on a formal statistical model but rather on the researcher's own judgment. However, it is possible to develop statistical models that can let us quantify exactly how well we can decode what a person is thinking about from fMRI data, which is the approach that Haxby and his colleagues took. Subsequent research has provided even more evidence of the power of fMRI to decode thoughts, which we will explore in much more detail in the next chapter.

From Modules to Networks

The debate over how faces are recognized was focused on whether the information was localized to one specific area or distributed across the temporal lobe, but there is an important point that this question obscures: the act of recognizing a face

requires that these areas in the temporal lobe communicate with other parts of the brain that are involved in social processing, action, memory, and emotion, to name just a few relevant processes. As fMRI developed, researchers began to characterize how different parts of the brain communicate with one another.

Bharat Biswal arrived as a graduate student at the Medical College of Wisconsin in 1992, just after the team of Bandettini and Wong had performed their first fMRI scans.[9] For his project, he set out to understand the various sources of noise in the fMRI data, such as heartbeats and respiration, both of which can have substantial effects on fMRI signals. One of the tricks that he tried in order to better understand these signals was to take the time series of data from a voxel located in the left motor cortex, and then measure how the signals across the rest of the brain were correlated with this voxel during the scan. He expected that the voxels nearby the one he had selected would be correlated (since they should have neurons that react in a similar way), and indeed he saw this, but he also found something surprising: the motor cortex on the right side of the brain also showed signals that were highly correlated with the voxel in the left motor cortex (see color plate 4), even when the person was just resting in the MRI scanner and making no movements. In fact, the map that he obtained by measuring the correlation between the left and right motor cortexes during resting fMRI looked very much like the map that he obtained when he compared moving both hands to resting, meaning that the motor cortex could be identified using resting-state fMRI even when the person was not moving at all. Biswal published these results in 1995, but it took about a decade for researchers to realize just how important his ideas were. As we will see in chapter 5, the study of the brain in people simply lying in an MRI scanner at rest is now one of the most powerful techniques in human neuroscience.

As researchers began to study how different parts of the brain are connected to one another, they took advantage of several new techniques. When Biswal measured correlations in fMRI signals across different brain regions, he was measuring what we call "functional connectivity"—the degree to which activity in different brain areas moves together over time. This does

not tell us whether the regions are "structurally connected" by the white matter of the brain, which is the cabling that connects different brain areas. It could be that two areas are connected to one another directly by white matter (which we call a "tract")—like Highway 101 that directly connects Los Angeles and San Francisco, but the functional connection could instead go through other areas in more than one step—like driving to San Francisco from Los Angeles via Las Vegas. If we want to understand the wiring diagram of the brain, then this is a crucial question. Historically, the tracing of tracts in the white matter has been studied in animals by injecting a radioactive "tracer" in one region and then looking at where it ends up in the rest of the brain as it moves along the brain's axons. We can't do this in living humans, but once again MRI has come to our rescue, through a technique called "diffusion weighted MRI" (or DWI for short) that images the movement of water molecules in the brain. We can use DWI to image white matter tracts because the axons that make up the tracts are covered with a fatty material (called *myelin*) that insulates them, like the plastic sheath that covers an electrical wire. Because it's difficult for water to pass through the cell membranes and their fatty myelin insulation, it tends to move more easily in the direction of the axon rather than across it. By measuring the diffusion of water in many different directions, we can infer the white matter connections between regions in the brain, using a technique that we call *tractography*.

By putting together information from functional connectivity measurements via fMRI and structural connectivity measurements using DWI, we can start to trace out what we call the "connectome": the catalog of connections between all of the different areas of the brain. Many people are familiar with this term through Sebastian Seung's book by the same name, and his TED talk titled "I Am My Connectome," by which he meant that everything that makes each of us unique is stored in the specific connections between neurons in our brain. Seung's work focuses on specific connections between individual neurons, what we might call the *microscopic* connectome, which can only currently be studied in nonhuman animals. In neuroimaging we instead focus on the large-scale connections between different areas in

the brain, or what we call the *macroscopic* connectome. In the end we hope that these two lines of research will converge, though it will always be difficult (if not impossible) to study the microscopic connectome in humans. The importance of understanding the brain's wiring diagram led the National Institutes of Health (the major funding agency for biomedical research in the United States) to spend US$30 million from 2010 to 2014 on the Human Connectome Project, with the goal of providing a detailed map of human brain connectivity. Over the course of that period, the Human Connectome Project collected MRI data, psychological testing, and genetic material from 1,200 individuals, and the data were made openly available to scientists around the world (just as they had been for the Human Genome Project). These data have led to several important breakthroughs regarding brain function, as well as to a new atlas for the human brain that has discovered new areas and characterized the differences between people in how their brains are organized.

The increased interest in connectomics has coincided with the broader development of what has come to be called "network science"—the science of complex networks, which can range from connectivity in the brain to friendships on Facebook to flights between airports.[10] Many people will be familiar with the idea of "six degrees of separation," made particularly well known by the demonstration that nearly every actor in the Internet Movie Database can be linked to Kevin Bacon by six or fewer costars.[11] What this phenomenon highlights is that complex networks often have a particular kind of structure that makes communication across the network very efficient; these are then known as "small-world" networks. A small-world network is one in which there are a small number of elements (which could be people, brain areas, or airports) that are highly connected, which we refer to as "hubs." For example, Heathrow and Newark International airports are both hubs in the sense that they have flights to many different airports (including other hubs), whereas the airport in Ithaca, New York, or Fresno, California, may only have flights to one or two different airports. Neuroimaging research has shown that the human brain has many of the features of a small-world network, and the tools of network

analysis have been used to make a number of interesting new discoveries about brain function, which we will delve into more deeply when we discuss brain connectivity in chapter 5.

Growing Pains

Whenever a new measurement tool emerges, the scientific community often struggles to understand how to work with the data and what their limits are, and fMRI has been no exception. In fact, the high profile of fMRI research has made it an attractive target for researchers looking to criticize it.

One of the major challenges for the analysis of fMRI is the fact that we collect so many measurements at once. In comparison to a psychology study where we might just measure 5 to 10 different variables, in fMRI we regularly collect data from more than 100,000 locations in the brain. The unique challenges of dealing with such big data were highlighted in one of the most amusing episodes in the history of fMRI.

In 2009 I was a member of the program committee for the Organization for Human Brain Mapping, which is responsible for vetting submissions to make sure that they meet the standards of the organization before accepting them for presentation at the annual meeting. One of the criteria for rejecting a submission is if it is a joke, and one particular submission was flagged for this reason by one of its reviewers. The title of the abstract was "Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument for Multiple Comparisons Correction," which certainly doesn't sound like a very funny joke, but a closer reading of the submission showed why the reviewers had been concerned:

> **Subject**. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.
>
> **Task**. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown

> a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.[12]

What the researchers, Craig Bennett and his colleagues, had done was to put a dead salmon in an MRI scanner, present it with a "task," and record fMRI data. They then analyzed the data in a particular way, and found that there was apparently activation in the salmon's brain in response to the task (see color plate 5). The authors did not do this to demonstrate some kind of after-life mental capacity in the salmon; rather, they did it to prove a critical point about analyzing fMRI data—one which had been known for many years, but had nonetheless been neglected by many researchers in the field.

Remember that fMRI data consist of measurements from many small cubes ("voxels") across the brain. In a standard fMRI scan we would collect data from anywhere between 50,000 and 200,000 voxels. In order to determine which parts of the brain respond to our task, we compute a statistic at each voxel, which quantifies how much evidence there is that the voxel's signal fluctuates in the way we would expect if it were actually responding to the task. We then have to determine which regions show a strong enough response that they cannot be explained by random variability, which we do using a statistical test. If the response in a voxel is strong enough that we don't think it can be explained by chance, then we call it a "statistically significant" response. In order to determine this, we need to determine how willing we are to accept false positive results—that is, results that are called statistically significant even though there is no actual signal in the data (known technically as "type I errors"). There is also another kind of error that we can make, in which we fail to find a statistically significant result even when there is a true effect in the voxel; we call this a "false negative" or "type II error." These two types of statistical errors exist in a delicate balance—holding all else equal, increasing our tolerance for false positives will decrease the rate of false negatives, and vice versa.

The usual rate of false positives that we are willing to accept is five percent. If we use this threshold, then we will make a false positive error on five percent of tests that we perform. If we are doing just a single test, then that seems reasonable—19 out of every 20 times we should get it right. But what if we are doing thousands of statistical tests at once, as we do when we analyze fMRI data? If we just use the standard five percent cutoff for each test, then the number of errors that we expect to make is 0.05 multiplied by the number of tests, which means that across 100,000 voxels we are almost certain to make thousands of false positive errors, and this is in fact what Bennett and his colleagues found. They wrote: "Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the [fMRI] timeseries may yield spurious results if multiple comparisons are not controlled for." Unfortunately, that part of their conclusion was often lost when the results were discussed in the media, resulting in a misleading impression that fMRI data were untrustworthy.

In fact, neuroimaging researchers have understood this problem of "multiple comparisons" since the days of PET imaging, and statisticians have developed many different ways to deal with it. The simplest (named after the mathematician Carlo Bonferroni) is to divide the rate of false positives for each test (known as *alpha*) by the number of tests. This controls the false positive rate, but is often overly conservative, meaning that the actual rate of false positives will be less than the five percent rate. This is problematic because, as I mentioned before, there is a seesaw relation between false positive and false negative rates, so an overly conservative test will also cause a high number of false negative errors, meaning that researchers will fail to find true effects even when they are present. However, there are a number of methods that have been developed that allow researchers to control the level of false positives without being overly conservative. While it was common to see fMRI papers published without appropriate statistical corrections in the early days of imaging, today nearly every paper reporting fMRI results will use a method to correct for multiple comparisons.

Is fMRI Just Voodoo?

Another well-publicized critique of fMRI research centered on the study of how differences between people in their behavior relate to differences in their brain activity, which is commonly used in neuroimaging research. An example of this can be found in a study that my colleagues Sabrina Tom, Craig Fox, and Chris Trepel and I did, which was meant to understand why some people are more willing to take risks than other people;[13] I will discuss this study again in chapter 7. To examine this, we presented the 16 subjects with a number of different gambles (such as a 50/50 chance to win $26 or lose $14) while they were scanned with fMRI, and asked them whether they would take that gamble. To make sure that they treated it like a real decision, after the scan was finished we randomly selected a few of the trials and then, if they had said "yes" to the gamble, we flipped a coin to play that gamble for real money. On average, people are averse to losing, which means that most people won't say "yes" unless the amount they could win is about twice the amount they could lose. However, we also found that there was a great amount of variability across people in their degree of loss aversion: some people would agree to accept gambles where the amount they could win was just barely above the amount they could lose (say a 50/50 chance to either win $14 versus lose $12), whereas other people required the amount to be won to be several times as large as the amount to be lost before they would agree to the gamble.

We set out to understand this by analyzing how their brains responded to increasing gains and increasing losses, and we found that there were some regions in the brain where there was a very close relation between the loss aversion that we saw in their choices, and what we called "neural loss aversion," reflected by the degree that the brain was more turned on by gains than it was turned off by losses. In fact, we saw strikingly strong correlations between behavior and brain activity. The relation between the brain and behavior is defined using a statistic called a "correlation coefficient," which goes from one (meaning that the variables are perfectly related), to zero (meaning that they have no relation), to negative one (meaning that they move

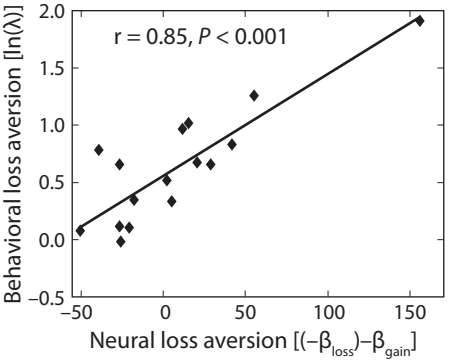

Figure 3.2. Figure from our 2007 paper, showing a puzzlingly high correlation between behavioral and neural loss aversion. The high strength of this correlation was due in part to the nonindependence of the analysis. From "The Neural Basis of Loss Aversion in Decision-Making Under Risk" by Sabrina M. Tom, Craig R. Fox, Christopher Trepel, Russell A. Poldrack, Science, 26 January 2007: 515–518. Copyright © 2007 by AAAS. Reprinted with permission from AAAS.

exactly in the opposite direction). We found that there was a correlation of 0.85 between the brain signals and the subjects' behavior (figure 3.2 reproduces a figure from that paper, showing this high correlation). This should have seemed too good to be true—but as the physicist Richard Feynman once said, "The first principle [of science] is that you must not fool yourself and you are the easiest person to fool."[14] What we should have done was to rerun the study and replicate the finding before publishing it, but running an fMRI study costs many thousands of dollars and could take several months. We let the excitement of discovery overcome our skepticism, and submitted the paper to the high-profile journal *Science*, where it was published in 2007.

Around that same time, Ed Vul and Hal Pashler were working on a paper that would rattle the fMRI world and bring about a serious public crisis. Vul and Pashler had seen other researchers present data just like those from figure 3.2, and it had left them thinking that it must be too good to be true. Their skepticism was driven by the fact that true correlations of that magnitude are only possible if the underlying variables being correlated are highly reliable—where "reliable" means that if we measure them twice, we should get the same number. In fact, it is a statistical

rule that the true correlation between two variables can't be much higher than the reliability of either measurement. We know that the reliability of psychological measurements like our gambling task is rarely above 0.8, so that should have already given us pause. We and others had also done research on the reliability of fMRI measurements across time, and we knew that it also rarely exceeds 0.8 and is often much lower. So how did studies find such high correlations?

Vul and his colleagues intuited that this was because of a statistical error known as "nonindependence" or "circularity." Imagine that I were to tell you that I had made a novel discovery that Stanford students had much higher SAT scores than the general population. You would immediately laugh at me and say "of course they do!" The average SAT score of Stanford students is necessarily higher than that of the general population, because the SAT score is one of the variables that goes into selecting them for admission. In the case of fMRI, we perform a large number of correlation tests across all of the voxels. If we then take the strongest ones and plot them, they will seem impressively large, but that's because we have guaranteed it to be the case. In fact, one can find impressively large correlations from completely random data using this kind of analysis (as we showed in a paper published in 2017).[15] One can imagine a different way of doing the analysis, in which we use one set of data to find the area of interest, and another separate set of data to compute the correlation, which Vul called an "independent" analysis; that approach does not suffer from the circularity that is present in the nonindependent approach.

Vul focused on the domain of social neuroscience, in part because brain-behavior correlation analysis was very common in that area. He requested information from a large number of authors about how they had done their analyses, classifying each analysis as either independent or nonindependent. When he combined the results across studies, he found that the correlations reported in studies using nonindependent analyses were much higher than those using independent analyses, and nearly all of the studies with correlations above 0.8 had used nonindependent analyses. In an aggressive challenge to the field,

Vul and colleagues titled their paper "Voodoo Correlations in Social Neuroscience."[16] I received a copy of their paper some time in 2008, before it had been published. As I read it, I had a sinking feeling; even though our 2007 *Science* paper was not one of the ones included in their list, I could see that we had made exactly the same error of nonindependence. I still thought that the bulk of our results were solid, since the correlation presented in our figure was mostly for illustration, and the underlying results had survived stringent correction for multiple comparisons, but I was still left with unease about the fact that our figure had misled readers. I set out to reanalyze our data in a way that did not suffer from nonindependence, using a technique called "cross-validation" (that we will discuss in more detail in Chapter 4) which allows us to test how well we can make predictions to independent data. What I found was that the correlations still held when using a proper independent analysis, but were smaller by about 40% compared with the nonindependent analyses.[17] Thus, our conclusions still held, but they were less impressive than we originally thought.

It is rare that a debate about data analysis methods in science reaches the pages of *Newsweek*, but that's indicative of the intense firestorm that this paper started. It also spurred a set of spirited published responses from fMRI researchers and statisticians, most of whom agreed with the substance of the Vul critique, if not the alarmist tone. But not all of them. Matt Lieberman (who at the time was a faculty colleague of mine at UCLA) could fairly be considered to be the primary target of the Vul article. It was his 2003 paper that was the first one listed as an example of supposed "voodoo" in Vul's paper, and Vul also claimed to have discovered a statistical error in one of the analyses reported by Lieberman's group. Lieberman and his colleagues Elliot Berkman and Tor Wager responded, stating that "Much of the article's prepublication impact was due to its aggressive tone, which is nearly unprecedented in the scientific literature and made it easy for the article to spread virally in the news."[18] They went on to try to argue that researchers didn't really mean to overstate the strength of their correlations—but this is a difficult argument to make when, as Vul pointed out in

his counter-response,[19] some researchers had referred to their correlations in press releases as "insanely strong."

The critique from Vul and colleagues shook me to the core, and drove me to rethink how we did fMRI research. I had always thought of myself as a reasonably savvy and careful researcher, but the fact that I had let myself be fooled by those seemingly strong correlations showed that I still had a long way to go. In the ensuing years we have become much more attentive to the ways that our data analysis methods can lead us astray, and we think that these improvements have increased the reliability of our research.