Box 6–1 Brain Imaging as Data Science

Compared to many areas of science, the basic methods of brain imaging have enjoyed remarkable standardization. A major reason for this has been the availability of widely adopted software packages since the earliest days of fMRI in the mid-1990s. These packages were created and released by research groups, and—before it was fashionable—most were open-source.

At first, they included tools for preprocessing, alignment, analysis models, and statistical corrections. They have since incorporated new tools developed by researchers, including nonlinear alignment, field map correction, nonparametric statistics, and parallelization.

As a result, virtually all fMRI researchers use one or more of these packages, at least for part of their analysis pipeline. The following are popular free software packages for fMRI analysis:

AFNI: https://afni.nimh.nih.gov

FSL: https://fsl.fmrib.ox.ac.uk

SPM: https://www.fil.ion.ucl.ac.uk/spm

Beyond these specialized packages, fMRI is increasingly being viewed through the more general lens of data science. There are two reasons for this. First, fMRI produces a huge amount of data, both within each session but also aggregated across the thousands of studies that have been conducted. Making sense of fMRI data can thus be considered a big-data problem. Second, the data are incredibly complex and noisy, and the cognitive signals of interest are weak and hard to find. This creates a data mining challenge that has inspired many computer scientists.

The most concrete manifestation of this trend is the rise of machine learning in fMRI analysis. Other points of contact with data science include the challenges associated with the real-time analysis of streaming data, the application of network analysis and graph theoretic approaches, the use of high-performance computing clusters and cloud systems, and the growing practice of researchers publicly sharing data (eg, https://openneuro.org), code (on services such as GitHub), and educational materials (eg, https://brainiak.org/tutorials). Thus, the field of brain imaging will continue to benefit from advances in computer science, engineering, applied math, and statistics.

steps referred to as *preprocessing*. Preprocessing seeks to remove known sources of noise in the data, caused by either the subject or the MRI machine. Standard practice includes five basic steps known as motion correction, slice-time correction, temporal filtering, spatial smoothing, and anatomical alignment.

Motion correction seeks to address inevitable noise in the data due to a subject's head movement. Even the best subjects move their heads a few millimeters over the course of a scan, such that the voxels across three-dimensional brain volumes become somewhat misaligned. This movement can be corrected for using a spatial interpolation algorithm that lines up all of the volumes within each run. This algorithm quantifies the amount of movement at each point during the scan, including the translation in the x, y, and z dimensions, and the amount of rotation about these axes (*pitch*, *roll*, and *yaw*, respectively). These six time courses can later be included in the data analysis as *regressors*, to further remove motion artifacts.

Slice-time correction is applied to deal with differences in the timing of the acquisition of samples across different slices. EPI sequences collect the slices that make up each brain volume sequentially, often in an interleaved order to avoid contamination of adjacent slices. Thus, there is a large difference in the timing of the first- and last-acquired slices of the same volume, which are closer in time to the preceding and subsequent volumes, respectively, than to each other. Correcting for this difference in the timing of the slices can be accomplished with temporal interpolation to estimate what the signal would have been if all slices were acquired simultaneously.

Temporal filtering and *spatial smoothing* aim to increase the signal-to-noise ratio. Temporal filtering removes components of the time course in each voxel that are highly likely to be noise rather than meaningful variance, such as very low frequencies (>100-second period) that typically result from scanner drift. Spatial smoothing applies a kernel (typically 4–8 mm wide) to blur individual volumes, averaging out noise across adjacent voxels and improving the odds that functions will overlap across subjects after anatomical alignment.

This *anatomical alignment* is accomplished by registering data across runs and subjects, usually with simple transformations (eg, shift, rotate, scale), to a standard template such as Montreal Neurological Institute or Talairach space. Typically, fMRI data are

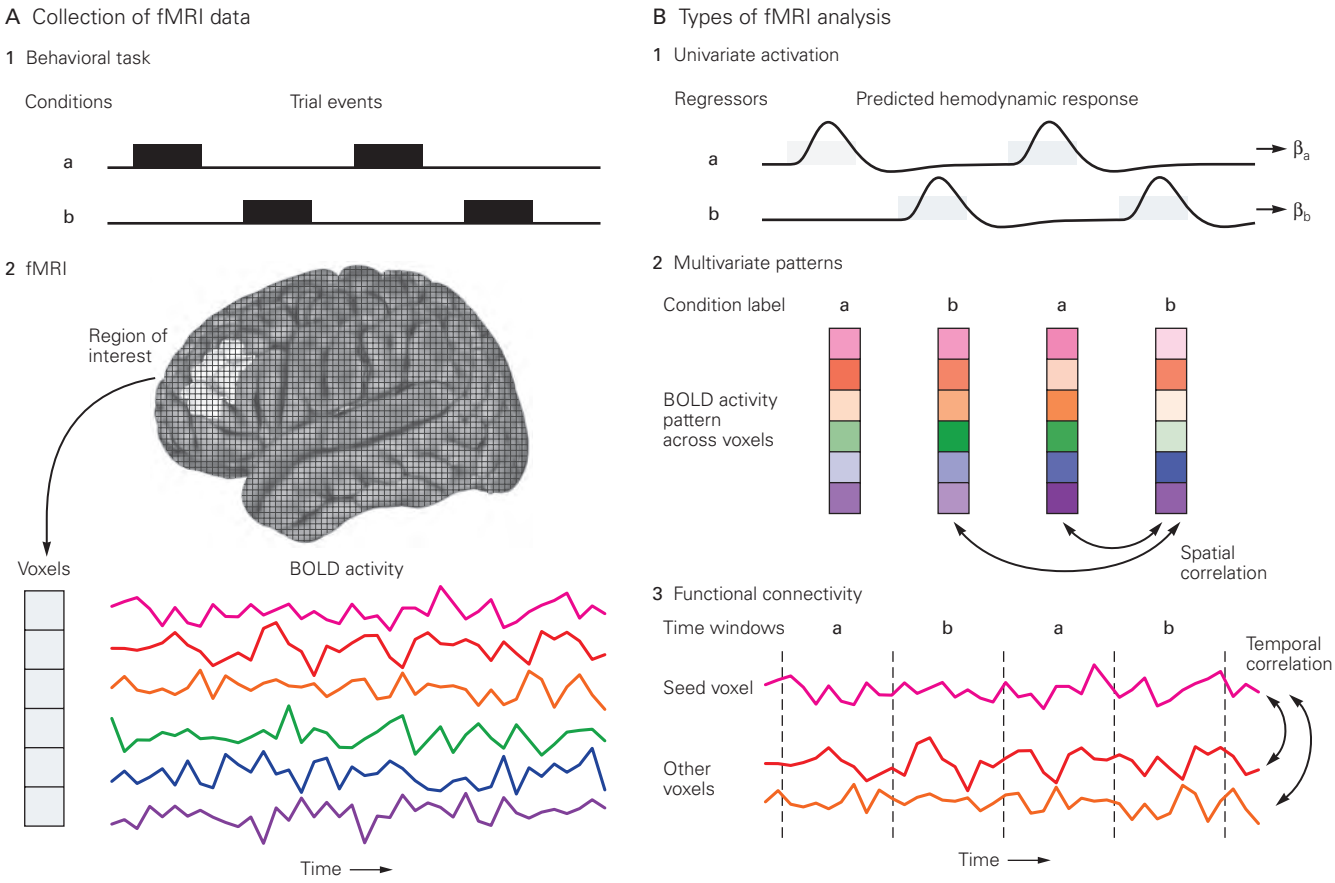

Figure 6–2 Collecting and analyzing fMRI data.

A. An fMRI experiment typically involves subjects performing a behavioral task while BOLD activity is measured from the brain.

1. The example task consists of two conditions (**a**, **b**) that alternate in time, each with two events depicted (**black rectangles**).

2. The time course of BOLD activity in six example voxels (different colors) from a region of interest (**ROI**) during the task. Analysis often focuses on an ROI or other subset of voxels in the brain to reduce the number of statistical tests performed. When all voxels in the brain are analyzed, statistical corrections are applied to reduce the number of false positives. The results of such analyses are often overlaid on a structural MRI as a color-coded heat map. The map is the result of extensive preprocessing and analysis and does not directly reflect neuronal activity or even blood oxygenation. Rather, voxels are colored to indicate that they have passed the threshold of being considered significant in a statistical test.

B. Three analysis approaches are often used in fMRI experiments such as the one depicted in **A**.

1. Univariate activation analysis attempts to explain the BOLD activity of each individual voxel in terms of what happened in the task. This is accomplished using a statistical model that contains a regressor for each task condition specifying the predicted hemodynamic response (**bell curves**) for trial events from that condition (**gray rectangles**). The result of fitting the model to BOLD activity is a beta value for each regressor in every voxel, quantifying the average response of the voxel to trials of that condition. The beta values for a voxel can be subtracted to measure whether there is a greater response in one condition than another. To determine statistical significance, this difference in activation between conditions in each voxel is compared across subjects.

2. Multivariate pattern analysis considers the pattern of BOLD activity across voxels. These spatial patterns are extracted for each trial from a subset of voxels (six depicted) and at a particular moment in time, often the peak of the predicted hemodynamic response (color saturation indicates amplitude of BOLD activity in each voxel on that trial). There are two common ways of analyzing these patterns. The first (shown) involves calculating the spatial correlation of patterns from a pair of trials to explore how similarly voxels responded to the trials. If a brain region represents different information across conditions, this pattern similarity should be higher for pairs of trials from the same versus different conditions. The second type of multivariate pattern analysis (not shown) uses a type of machine learning known as pattern classification. Some of the patterns and their corresponding condition labels are used to train a classifier model, assigning weights to voxels based on how useful they are at distinguishing between conditions. The model is then tested with other patterns on which it was not trained. If a brain region represents different information across conditions, the model should be able to correctly guess from which condition the patterns were extracted. To determine statistical significance, spatial correlations or classification accuracies in a region are compared across subjects.

3. Functional connectivity analysis examines how BOLD activity is correlated between voxels over time. Typically, a seed voxel or ROI is chosen and its time course (**pink curve**) is correlated with the time courses of other voxels (two shown here). This can be performed while the subject is resting, resulting in a correlation value for every voxel that can be used to identify brain networks in a baseline state. Functional connectivity can also be calculated in different time windows of a task (**dashed lines**), resulting in a correlation value for each trial that can be used to understand the dynamics of these networks. To determine statistical significance, temporal correlations for each voxel are compared across subjects between conditions or against zero.

first aligned to a structural scan from the same subject, and then this structural scan is aligned to the standard template.

Once these five steps are completed, the data are ready for analysis.

fMRI Can Be Used to Localize Cognitive Functions to Specific Brain Regions

The first kind of fMRI analysis seeks to localize functions in the brain and to determine what brain regions are associated with a behavior. This is based on having subjects complete a task during fMRI and then examining the relationship between different phases of the experiment and changes in BOLD activity in different parts of the brain. Based on researchers' knowledge of what happened at different times in the experiment, the function of the regions can be inferred.

A series of statistical analyses are performed to quantify this relationship and to determine its significance. Typically, this is accomplished using a statistical regression method known as a *general linear model* (GLM). The GLM attempts to explain observed data (here, the time course of the BOLD activity in each voxel) as a linear combination of regressors that reflect independent variables (eg, task conditions) and covariates (eg, movement parameters).

The regressors that model task conditions serve as a hypothesis about how a voxel should respond if involved in the cognitive function manipulated by that task. The regressor for each condition is generated by marking the onset and duration of each trial of that condition in the experimental time line, corresponding to the expected neuronal activity, and then accounting for the delayed hemodynamic response. All regressors are fit simultaneously to the fMRI activity in each voxel, and the result is a parameter estimate (or "beta") for each condition and voxel, reflecting how much of the temporal variance of the voxel is uniquely explained by that condition's trials on average.

To localize a function, betas from two or more conditions are compared in a contrast. The most basic form of contrast is to subtract one beta (eg, control condition) from another (eg, experimental condition). Contrasts are typically averaged over runs within each subject and then entered into a t-test to assess reliability across subjects. Because statistics are calculated for every voxel, there is a high risk of false positives, and a correction for multiple comparisons is required (eg, by giving voxels more credence if they cluster together with other significant voxels). Alternatively, a more constrained analysis can be performed, focusing on a limited number of regions of interest (ROIs) that are defined a priori. Contrast values can then be averaged over the voxels in an ROI to produce regional estimates, rather than examining all voxels in the brain, thereby reducing the number of comparisons.

This general family of approaches is often described as measuring *univariate activation*—"univariate" because each voxel or region is treated independently and "activation" because the result is a measure of the relative activity evoked by one condition versus another. This kind of analysis is typically used to localize a cognitive function to a set of voxels or regions in the brain.

However, univariate activation can be used for more than localization. For example, a GLM can make quantitative predictions about BOLD activity by assigning a continuous weight, rather than a categorical one, to each trial in a regressor based on an experimental parameter (eg, working memory load), behavioral measurement (eg, response time), or computational model (eg, prediction error in reinforcement learning). The resulting beta reflects how much a voxel correlates with the variable of interest.

Another use of univariate activation is for measuring changes in BOLD activity as a function of repeating a stimulus. Such studies take advantage of *adaptation* (or *repetition suppression*)—the tendency of stimulus-selective neurons to respond less to repeated versus new stimuli. This fact allows the tuning of a brain region to be inferred by conducting an experiment in which related and unrelated stimuli are presented sequentially. In some trials, one stimulus is followed by a near-repetition of the same stimulus, but with a feature changed (eg, its location or size). A univariate analysis tests whether BOLD activity in voxels from the region is lower on these trials compared to other trials in which either (1) the first stimulus is followed by an unrelated second stimulus or (2) the changed stimulus is preceded by an unrelated stimulus. If such a BOLD reduction is observed, the region can be interpreted as not tuned for the changed feature (eg, the region could be considered location or size invariant).

fMRI Can Be Used to Decode What Information Is Represented in the Brain

The second category of fMRI analysis seeks to characterize what kinds of information are represented in different regions of the brain to guide behavior. Rather than analyze voxels independently or average over voxels within an ROI, these analyses examine the information carried by spatial *patterns* of BOLD activity over multiple voxels. This is typically referred to as *multivariate pattern analysis* (MVPA). There are two

types of MVPA, based on the similarity or classification of activity patterns.

Similarity-based MVPA tries to understand what information is contained or "represented" in a brain region. This is accomplished by examining how similarly the region processes different conditions or stimuli in an experiment. This similarity is calculated from the pattern of activation across voxels in an ROI, defined as either the pattern of beta values from a GLM or the pattern of raw BOLD activity from preprocessed data. Once these patterns have been defined for multiple conditions or stimuli, the correlation or distance of each pair of patterns is calculated. This produces a matrix of the pairwise similarities between conditions or stimuli within the ROI. With this matrix, it is possible to infer to what information the ROI is most sensitive. For example, if subjects are shown photos of different objects (eg, a banana, canoe, taxi), a matrix of distances between the activity patterns evoked by these objects can be computed for different brain regions. An ROI in which there is less distance between banana and canoe than between either of them and taxi could be interpreted to mean that the region represents shape (ie, concavity); another region in which the lowest distance is between banana and taxi might represent color (ie, yellow); or one with the lowest distance between canoe and taxi might be interpreted as representing function (ie, transportation).

Neural similarity from fMRI can also be compared with similarity calculated in other ways for the same conditions or stimuli, including from human judgments, computational models, or neural measures in other species. For example, if human subjects rate a large set of stimuli in terms of how similar they look to each other, a brain region with a matching similarity structure could be considered a candidate source of this behavior. This approach of calculating *second-order* correlations between neural and behavioral similarity matrices, or between neural similarity matrices from two sources, is called *representational similarity analysis* (RSA).

Classifier-based MVPA uses techniques from machine learning (discussed in Chapter 5) to decode what information is present in a brain region. The first step is to train a classifier model on a subset of the fMRI data to discriminate between conditions or stimulus classes from patterns of BOLD activity across voxels in an ROI. These patterns are usually obtained from individual trials, and each is labeled according to the condition or stimulus on the corresponding trial. This training set thus contains several brain pattern examples of each class. Classifier training can use many different algorithms, the two most common being support vector machine and regularized logistic regression. The result is typically a weight for each voxel reflecting how activity in that voxel contributes to classification collectively with the other voxels. The second step after training is to test the classifier by examining how well it can decode patterns from a held-out and independent subset of fMRI data (eg, from a different run or subject). The pattern of BOLD activity on each test trial is multiplied by the learned classifier weights and summed to produce a guess about how the pattern should be labeled. Classification accuracy is quantified as the proportion of these guesses that match the correct labels. Importantly, this approach can be used to understand how different brain regions give rise to behavior, such as by attempting to classify which action was performed, which decision was made, or which memory was retrieved.

fMRI Can Be Used to Measure Correlated Activity Across Brain Networks

The third category of fMRI analysis seeks to understand the organization of the brain as a network. Knowing what brain regions do individually does not fully explain how the brain as a whole generates behavior. It is additionally critical to know how brain regions relate to each other—that is, where do the inputs to a region come from and where do the outputs go? This requires an understanding of which regions communicate with each other and when and how they transmit information. This is difficult to determine definitively with fMRI but can be estimated by measuring the correlation of BOLD activity between voxels or regions over time. If two parts of the brain have correlated activity, they may be sharing the same information or participating in the same process. Such correlations are interpreted as measures of *functional connectivity*.

One way to study functional connectivity with fMRI is to measure BOLD correlations in a resting state. Subjects are scanned while they lie still without performing a task, and then the time course of BOLD activity from one "seed" ROI is extracted and correlated with the time courses from other ROIs or from all voxels in the brain. Alternatively, clustering or component analyses can be used without a seed to identify collections of voxels with similar temporal profiles. Resting functional connectivity defined in these ways has helped reveal that the brain contains several large-scale networks of regions. The most widely studied of these networks is referred to as the default mode network, which includes the posterior medial cortex, lateral parietal cortex, and medial prefrontal cortex.

By definition, resting connectivity cannot be linked to concurrent behavior. Nor is it static, as telling subjects not to do anything does not restrict what they think about. Nevertheless, resting connectivity can be linked to behavior indirectly by examining how it goes awry in disease or disorders and how it relates to cognitive differences between people.

Functional connectivity can be linked more directly to behavior if it is measured during tasks rather than at rest. One difficulty in interpreting such correlations between regions is that two regions might be correlated during a task not because they are communicating with each other, but because of a third variable. For example, the regions might be responding independently but coincidentally to the same stimulus. Thus, task-based functional connectivity is typically calculated after removing, or otherwise accounting for, BOLD responses evoked by stimuli. This approach allows functional connectivity to be manipulated experimentally and compared across task conditions. These comparisons provide insight into how the involvement and interaction of brain regions in a network change dynamically to support different behaviors. This has proven useful for understanding cognitive functions such as attention, motivation, and memory, which depend on some brain regions modulating others.

Functional connectivity can also be viewed as a pattern (of correlations rather than activity) and submitted to MVPA. Correlation patterns are larger in scale than activity patterns: If there are n voxels in an activity pattern, there are on the order of n^2 voxel pairs in a correlation pattern. Thus, it can be helpful to summarize the properties of correlation patterns using graph theory, where individual voxels or regions are treated as the nodes in the graph and the functional connectivity between these nodes determines the edge strengths.

Functional MRI Studies Have Led to Fundamental Insights

Functional MRI has changed our understanding of the basic neurobiological building blocks of human behavior. Combining experimental manipulations and computational models from cognitive psychology with precise neurobiological measurements has expanded existing theories of the mind and brain and has stimulated new ideas. Discoveries from fMRI have impacted not just our understanding of behaviors presumed to be uniquely human, but also behaviors that have long been investigated in animals.

In this section, we review three examples of this progress. The study of face perception reveals how human fMRI studies have inspired research in animals. The study of memory illustrates how fMRI has challenged theories from cognitive psychology and systems neuroscience. The study of decision-making shows how animal studies and computational models have advanced fMRI research.

fMRI Studies in Humans Have Inspired Neurophysiological Studies in Animals

Our understanding of how the brain perceives faces has grown tremendously over the past two decades (Chapter 24). The advances described below provide an example of how findings from fMRI in humans inspired follow-up studies with neuronal recordings and causal interventions in nonhuman primates. This synergy across species and techniques led to a more complete understanding of the fundamental process by which faces are recognized.

Some classes of stimuli are more important for survival than others. Does the brain have dedicated machinery for the processing of such stimuli? Faces are an obvious case in humans. The development of fMRI combined with careful and systematic experimental designs led to important insights into how and where faces are processed in the human brain. One region in the fusiform gyrus, often referred to as the fusiform face area (FFA), was found to show robust and selective BOLD activity when humans view faces.

Early fMRI studies that led to this discovery relied on simple designs in which subjects were presented with a series of different types of visual stimuli. To measure the face selectivity of brain areas, the BOLD response to faces was compared with the BOLD responses for the other categories (eg, places, objects). An area of the lateral fusiform gyrus, most reliably in the right hemisphere, was strongly activated by faces. These findings fit with earlier findings of individual neurons in nonhuman primates that respond to faces, but inspired a new wave of animal studies to examine a larger-scale network of brain regions. These newer animal studies, borrowing experimental designs from the human studies, first used fMRI to find orthologs of the FFA. The resulting face patches were then probed invasively with neuronal recording and stimulation. This revealed insights into the distributed neural circuitry for face processing in primates.

In addition to responding selectively to face stimuli, does the FFA contribute to the behavior of face recognition? This question has been addressed using stimulus variations that are known to affect face recognition

(eg, presenting faces that are inverted or presenting parts of faces). Initial fMRI studies using simple comparisons of stimulus categories (inverted versus upright faces) produced weak and mixed results. Follow-up studies used an adaptation design to determine how BOLD activity changes when a face is repeated intact or altered. The findings suggested that the FFA represents intact faces differently than when the same visual features are reconfigured in a way that disrupts behavioral recognition.

Another way to examine the behavioral significance of a region is to study patients who have behavioral deficits—in this case, an impairment of face recognition known as prosopagnosia. Surprisingly, some fMRI studies found an intact FFA in these patients, casting doubt on its necessity for face perception. However, here too follow-up studies using an adaptation design proved informative: The otherwise intact FFA of prosopagnosics did not adapt when the same face was repeated. This suggests that the FFA responds differently in people with prosopagnosia, consistent with its importance to face recognition.

The finding that visual categories, or mental processes more generally, can be mapped to one or a small number of regions like the FFA was important for thinking about the relationship between mind and brain. Whether specific functions are localized or broadly distributed has been a central question regarding brain organization throughout the history of neuroscience (Chapter 1). The discovery of the FFA and the face patch system provided new evidence of localization, and encouraged researchers to pursue the hypothesis that other complex cognitive functions might be localized in specific brain areas or small sets of nodes, but also to question whether localization is the right way to think about brain organization. For example, further studies showed that faces produce widely distributed responses over visual cortex and that the FFA can be co-opted for recognition of other kinds of objects with which we have expertise. These debates reflect the transformative nature of this original work, both for studies of the human brain and for related questions in animal models.

fMRI Studies Have Challenged Theories From Cognitive Psychology and Systems Neuroscience

Many theoretical models from cognitive psychology were originally agnostic about the brain. However, there are now several examples of fMRI findings that changed our understanding of the organization and mechanisms of cognition.

One prominent example is the study of memory. The overall goal of memory research, beginning in the 19th century, has been to understand how a memory is created, retrieved, and used, and whether these processes differ across types of memory. A key discovery came from research on patient H.M. and the realization that damage to the hippocampus causes a loss of the ability to form new autobiographical memories but does not impact the ability to learn certain skills (Chapter 52). These findings led to the idea that memory can be divided into two broad classes, conscious versus unconscious (also known as declarative versus procedural or explicit versus implicit). In the tradition of localization, these and other types of memory were mapped onto distinct brain regions, based on where in the brain a patient had damage and which behavioral symptoms they exhibited.

Later fMRI studies of the healthy human brain helped reveal that this dichotomy was oversimplified. First, several studies using what came to be known as the *subsequent memory task* showed that regions beyond the hippocampus are implicated in the successful formation of declarative memory. In such studies, subjects are presented with a series of stimuli (pictures or words) while being scanned. Later, usually outside of the MRI machine, their memory for these stimuli is tested. The BOLD responses from when a stimulus was initially encoded are then sorted based on whether it was subsequently remembered or forgotten. These conditions are contrasted to reveal which brain regions show more (or less) activity during successful memory formation. In addition to finding such differences in the hippocampus and surrounding medial temporal lobe, BOLD activity in prefrontal and parietal cortices is also predictive of later memory. By measuring the whole brain of healthy individuals, fMRI revealed that declarative memory is served by more than one brain system—processes linked to prefrontal cortex (eg, semantic elaboration) and parietal cortex (eg, selective attention) are also involved in encoding.

The traditional taxonomy of memory organization was challenged in another way by fMRI studies. fMRI revealed that a wide range of tasks that were previously assumed to not involve the hippocampus (or declarative memory) in fact do consistently engage this region. These studies often use learning tasks that would classically be considered unconscious, in which subjects have the opportunity to learn but are never asked to report their memories and, in some instances, are unable to do so if prompted. For example, in the *probabilistic classification task*, subjects learn by trial and error to sort visual cues into categories, even when the relationship between cues and categories is sometimes unreliable. BOLD activity during such learning trials is estimated and compared to

a baseline task that does not involve trial-and-error learning (eg, studying cues with their categories provided). Such comparisons generally reveal activation in the striatum, but also reliably in the hippocampus (see Chapter 52).

In summary, fMRI studies of tasks thought to rely on declarative memory often recruit regions outside of the hippocampus, and tasks thought to rely on procedural memory can recruit the hippocampus. In both cases, these discoveries were serendipitous and made possible only because data were obtained from the whole brain with fMRI. Although these began as unexpected results, they led to systematic follow-up studies that have updated our understanding of the organization of memory. Chiefly, they challenged the original emphasis on conscious awareness as the defining characteristic of hippocampal processing. This in turn helped relate the findings from human studies to those from animal studies, where the notion of conscious memory is less central and where tasks that engage the hippocampus often involve spatial navigation. Thus, fMRI findings in humans have been transformative for our understanding of theoretical models of memory, in terms of both neural structures and cognitive behaviors.

fMRI Studies Have Tested Predictions From Animal Studies and Computational Models

The integration of computational models with fMRI has been an important development in cognitive neuroscience. One example of this comes from studies of how the brain learns to predict and obtain rewards, combined with models of reinforcement learning that formalize this process. These models co-evolved with studies of reward-based decision-making in animals, which also inspired later human studies.

Central to these studies and theories, midbrain dopaminergic neurons increase their firing in response to unexpected rewards, such as juice (Chapter 43). Once a predictive cue has been reliably paired with a reward, the neurons shift their response in time to this predictive cue. If a predicted reward fails to occur, firing decreases. This pattern of responses suggests that midbrain dopaminergic neurons signal the difference between expected and actual rewards. This difference is commonly known as *reward prediction error* and has been modeled using equations based on reinforcement learning theory. When this model is applied to human tasks involving rewards, hypothesized reward prediction errors can be estimated on a trial-by-trial basis. These estimates can then be used to predict BOLD activity and identify voxels and regions that may be involved in reinforcement learning in the human brain.

In a typical study of this type, subjects perform a learning task during fMRI, making a series of choices about visual cues to predict possible rewards. They learn the outcome immediately after each choice. For example, a subject might view two shapes (eg, circle, triangle), choose one by pressing a button, and then learn whether the choice led to a monetary reward. The key feature of such tasks is that the association between shapes and rewards is probabilistic and changes over the course of the experiment. Because of this noisy relationship, subjects must learn to track the likelihood of reward for each shape. Reward prediction error can be calculated on each trial based on the history of the subject's choices and rewards and then included in the analysis of their fMRI data. Many studies using this approach have found that trial-by-trial reward prediction error correlates with BOLD activity in the ventral striatum, an area that receives input from midbrain dopaminergic neurons.

Other computational models, such as *deep neural networks*, which integrate cognitive psychology, computer science, and neuroscience, have also served an important theoretical purpose by generating novel hypotheses about brain activity. Because these models are often inspired by the architecture and functions of the brain, they help bridge levels of analysis, from physiological recordings in animals to fMRI in humans. They also serve a useful purpose in data analysis by simulating variables of psychological and neurobiological interest that can be sought in the brain, an approach often referred to as model-based analysis.

Functional MRI Studies Require Careful Interpretation

The examples provided earlier illustrate how fMRI can improve our understanding of the links between brain and behavior. At the interface with psychology, fMRI can complement purely behavioral measurements. Many complex human behaviors (eg, memory recall, decision-making) depend on multiple processing stages and components. Measuring these processes with fMRI can provide richer and more mechanistic explanations of behavior than those based on simple behavioral measurements such as accuracy or response time alone. At the interface with systems neuroscience, fMRI complements direct neuronal recordings. Most brain areas (eg, hippocampus) support multiple behaviors and do so in concert with other regions. The ability to image the whole brain with fMRI makes it

possible to arrive at a more complete understanding of neural mechanisms at the network level.

What does it mean then to find BOLD activity in a region during a task? The multiplicity of mappings between brain and behavior poses serious challenges to interpretation of fMRI results (Figure 6–3). One fundamental consideration is the type of inference. Most fMRI studies use *forward inference*, in which an experiment compares BOLD activity between task conditions that manipulate the engagement of a particular mental process (eg, comparing the effects of face versus nonface stimuli to study face recognition). Brain regions that differ between these conditions can be inferred to take part in the manipulated process. Forward inference relies on a task manipulation and therefore allows a researcher to infer that differences in brain activity are related to the mental process of interest.

With *reverse inference*, differences in neural activity are the basis for inferring which specific mental process is active, even when the conditions that gave rise to the differences were not designed to manipulate that process. For example, in the previous face versus nonface contrast, a researcher might interpret differential activity in the striatum as evidence that faces are rewarding. This kind of reverse inference is often unjustified, as reward was neither measured nor manipulated—the interpretation is based on other studies that manipulated reward and found striatal activity. The problem arises because each brain region generally supports more than one function, meaning that it is unclear from the observation of activity alone which function(s) were engaged. Indeed, the striatum is also strongly implicated in movement, so perhaps faces are engaging motor rather than reward processes? The logically sound conclusion in this example, reflecting forward inference, is that the striatum is involved in some (as yet unresolved) aspect of face recognition.

One solution therefore is not to use reverse inference in fMRI studies. However, there are some situations in which reverse inference can be desirable or even necessary. For example, reverse inference can allow researchers to perform exploratory analyses and generate new hypotheses, even from data that were collected for other purposes. This may be especially important for getting the most out of fMRI data that are hard to collect, such as from children, the elderly, and patients (Box 6–2). Motivated by this need, statistical tools have been developed to support reverse inference. For example, the web-based tool Neurosynth uses a large database of published studies to assign a probability that a specific mental process (eg, reward) is involved given that BOLD activity has been observed in a particular region (eg, striatum).

It is also important to make a distinction between a correlation of brain activity with behavior versus a cause-and-effect relationship between the brain activity and the behavior. If a brain region is selectively and consistently involved in a specific mental process, this correlation does not license the conclusion that it plays a necessary or sufficient role in the process. With respect to sufficiency, the brain region might (and likely does) work with one or more other brain regions to accomplish the process. With respect to necessity, activity in the region might be a secondary by-product of processing elsewhere.

One approach to bolstering the interpretation of an fMRI study is to evaluate how the findings converge with those from more invasive methods, such as electrical stimulation in epilepsy patients. Because every tool has limitations, including other correlational measures such as neuronal recordings, this principle of converging evidence is central to advancing understanding of how the brain supports behavior. In addition to converging evidence across studies and tools, there are also efforts to manipulate brain function simultaneously with fMRI, using transcranial magnetic stimulation or real-time neurofeedback.

Future Progress Depends on Technological and Conceptual Advances

Functional MRI is the best technology we have so far for probing the healthy human brain. It allows measurement of the whole brain at reasonably high resolution as well as many aspects of the mind in large subject samples without harm. However, in other ways, it is far from what we ultimately need if we are to obtain a deeper and more precise understanding of how the brain works. When compared to tools available in animals, fMRI provides relatively noisy, slow, and indirect measurements of neuronal activity and circuit dynamics.

Efforts are underway to address these limitations, both technically and biologically. On the technical front, multiband imaging sequences can enhance the temporal and spatial resolution of fMRI data by enabling the acquisition of multiple slices through the brain in parallel. However, faster measurements are inherently limited by the slow speed of the hemodynamic response and smaller voxels still average across hundreds of thousands of neurons.

On the biological front, we have a rudimentary understanding of how BOLD activity emerges from physiological mechanisms in the brain, such as single neuron activity, population activity, the function of

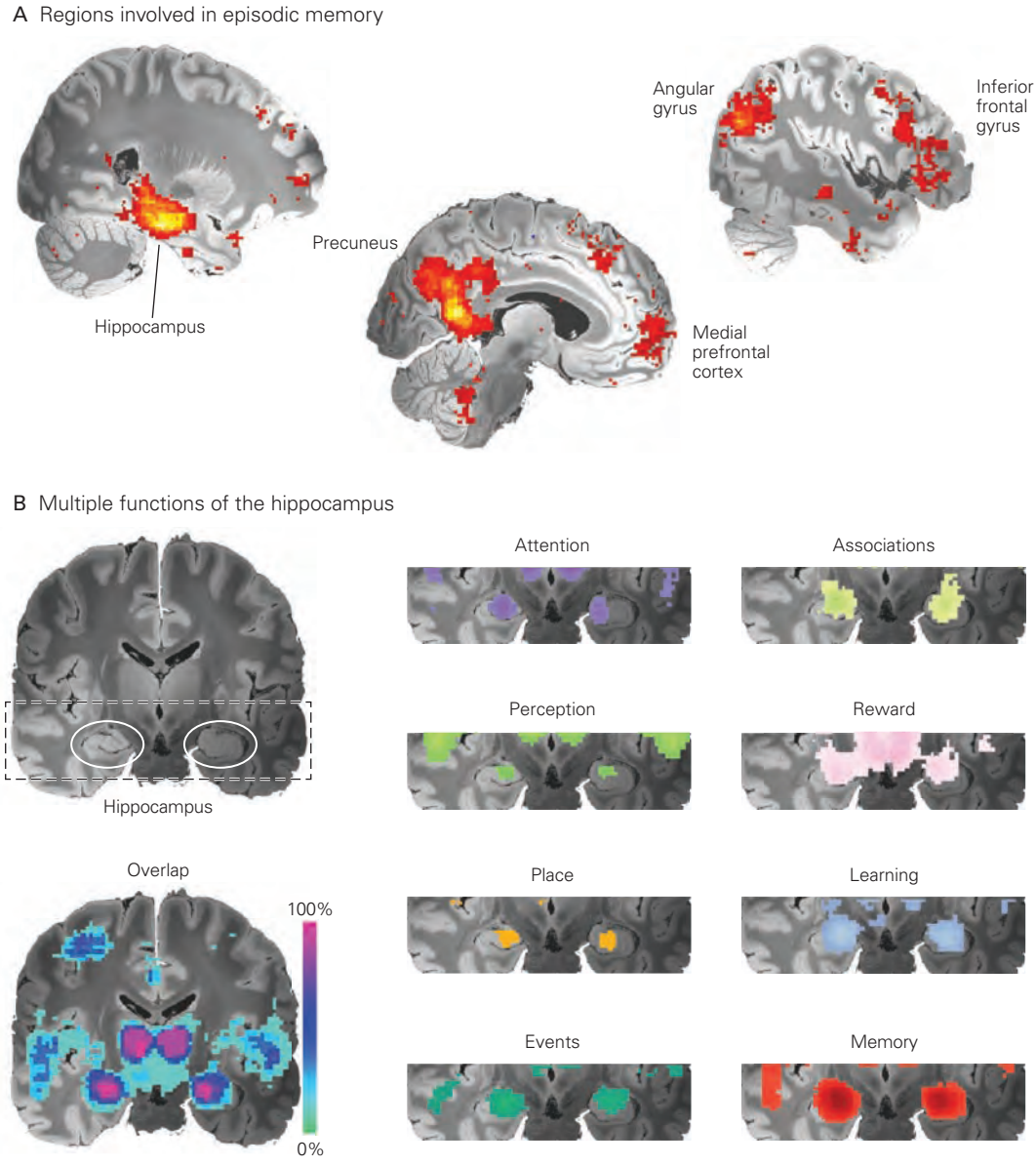

Figure 6–3 Challenges of mapping mind and brain. Any interpretation of data from fMRI must consider the complexity of the relationship between cognitive functions and brain regions. This complexity is illustrated here with a meta-analysis from a database containing more than 14,000 published fMRI studies. (Data retrieved in 2019 from http://neurosynth.org, displayed on brain from Edlow et al. 2019; figure updated and adapted from Shohamy and Turk-Browne 2013 by Tristan Yates.)

A. This map shows that multiple brain regions are engaged by episodic memory—that is, encoding and retrieval of specific events from one's past. Colored voxels indicate a high probability of the term "episodic" in studies that reported activation in these voxels (reverse inference). This example illustrates how a single cognitive function can be associated with multiple brain regions (one-to-many mapping).

B. These maps show that multiple cognitive functions engage the hippocampus (circled in white in each hemisphere). Colored voxels in each inset brain indicate a high probability that these voxels were activated in studies that examined the corresponding term (forward inference). The overlap map shows the percentage of these terms that activated each voxel. This example illustrates how a single brain region can be associated with multiple cognitive functions and behaviors (many-to-one mapping).