In Chapter 10 we introduced the core analysis methods that are used in the vast majority of fMRI studies. Most of those methods use variants of the general linear model to test specific hypotheses about brain function. These hypothesis-driven analyses have accounted for much of the growth of fMRI over the past decade, but they cannot address some potentially important research questions. In this chapter, we outline important and novel techniques for fMRI data analysis that go well beyond multiple regression. These techniques borrow technical concepts from many fields, including statistics and mathematics, economics and other social sciences, engineering, and computer science.

In general, the techniques described in this chapter are used to achieve one of two goals: exploration or prediction. Some techniques explore fMRI data in search of systematic variation, without necessarily adopting an a priori model for that variation. These data-driven analyses often use mathematical algorithms that partition the four-dimensional fMRI time series into a set of components that may reflect distinct aspects of brain functioning. Data-driven analyses may identify regularities that are clearly task-related (Figure 11.1), or they may discover task-unrelated variability that can be eliminated during preprocessing. These analyses are not subject to some of the problems that compromise hypothesis-driven analyses: poorly chosen models, unknown timing of neuronal activity, and variability in the hemodynamic response. Yet, they present many technical and interpretative challenges. Other techniques reverse the traditional direction of fMRI analyses, in that they use fMRI data to predict variations in behavior, perception, or cognition. Researchers now use fMRI to predict many aspects of thought and behavior, including susceptibility to psychiatric disorders, personality traits, complex economic decisions, and even conscious awareness of stimuli.

Some techniques do not fit neatly into one of these two categories. For example, we will also consider techniques for measuring the connectivity between brain regions; these share both exploration and prediction as goals. Yet, all of these techniques differ (sometimes dramatically!) from the simpler regression approaches considered in Chapter 10. In many ways, these new analysis techniques represent the future of BOLD fMRI, more than technological advances in fMRI technology.

**hypothesis-driven analysis** The evaluation of data based on statistical tests of the validity of a null hypothesis.

**multiple regression** A family of statistical approaches that evaluate the relative contributions of several independent variables to a dependent variable.

**data-driven analysis** Exploration of the intrinsic structure of data.
Figure 11.1 Data-driven analyses identify regular variation in fMRI data. This subject participated in a motor movement task with three phases: rotating the right hand at the wrist, rotating the left hand at the wrist, and rotating both hands simultaneously. The blocks of hand movement are shown in blue. By looking only at the raw fMRI data, using a technique called independent components analysis (ICA), the researchers found a set of voxels in the left motor cortex in which activation was evoked by all movements of the right hand. Note that these voxels were identified only by examination of the data themselves, not by hypothesis testing. (Data courtesy of Dr. Martin McKeown, University of British Columbia.)

Data Exploration Approaches

We here consider several techniques that allow researchers to explore their data for potentially meaningful variation. All parse the fMRI time series into sets of common features, or components, that consist of groups of voxels and their temporal properties. Importantly, their algorithms attempt to identify activation that is common to a group of voxels, rather than compare the activation of individual voxels with a hypothesized time course. The components are evaluated based on how much of the variation in the fMRI data they explain. These techniques vary both in the algorithms that they use to parse the time series data and, more importantly, in the rules they use to decide which features are meaningful. Some are completely model-free, in that they extract features without any regard to the underlying experimental paradigm. Others use some information about the task to shape how the components are extracted. These techniques can also differ somewhat in their goals; some are used to identify interesting forms of variability in order to suggest future analyses, while others are used to test simple hypotheses. We will discuss these approaches primarily in the context of identifying important aspects of brain function, however, some of these approaches are now also being used to identify uninteresting and non-task-related variability (e.g., for preprocessing).

Thought Question

Why might researchers use a combination of hypothesis-driven and data-driven analysis approaches?

Principal components analysis (PCA)

The major challenge of fMRI research, as outlined in previous chapters, lies in the detection of relatively small task-related variations against a large back-
ground of non-task-related processes, some representing random noise and others corresponding to unwanted regularities (e.g., breathing). This challenge can be overcome, at least theoretically, by the steps of preprocessing and data analysis: filtering the time series at noise frequencies, smoothing data across voxels, and averaging data from multiple events and sessions. But these steps, especially temporal filtering, require some knowledge of the properties of the noise compared with our signal of interest. What if one does not know the properties of the noise? Or, even more problematic, what if one does not know the properties of the signal? When faced with complex data sets like that of fMRI that contain several underlying components, including both task-related effects and different sources of noise, researchers often apply algorithms for data reduction. As the name implies, data reduction takes a complex, high-dimensional data set and transforms it into a simpler representation that retains most of the variation in the data.

A common technique for data reduction is principal components analysis (PCA). Considered broadly, PCA first identifies the component that accounts for the most variability in a dataset, then identifies a second component that accounts for the largest proportion of the remaining variability, and repeats until the desired number of components have been extracted. Any remaining variation in the data is discarded. (A related technique called factor analysis differs from PCA in that it only includes variation that is shared by all components, or factors, rather than including all variation in the dataset.) Because PCA is iterative, defining each component in relation to the previous components, the components are orthogonal to each other. For four-dimensional data sets like those obtained in fMRI studies, the components extracted by PCA can consist of eigenimages, or spatial patterns whose variations over time are expressed by their eigenvectors. The variance explained by each component is called its eigenvalue; larger eigenvalues mean that more variance is explained. It can be difficult to decide which components to retain and which to discard—that is, to define the set of components that accounts for the original data but does not include unwanted or spurious variability. One oft-used rule (the Kaiser Criterion) is to retain components that have eigenvalues greater than 1, which means that they explain at least as much variance as a variable from the original dataset. Another more graphical approach involves plotting the eigenvalue for each component, which will result in a curve known as a scree plot that resembles a decreasing exponential function. Then, the researchers discard all components beyond where the curve flattens out.

In theory, when researchers apply PCA to fMRI data, the output will consist of a set of components that represent regular activation patterns across voxels. If those components reflect interpretable consequences of an experimental task, then PCA can be used to map out functional systems that are distributed in different parts of the brain, which would be inaccessible when using normal hypothesis-based methods. However, PCA has many limitations. If the task-related signal constitutes only a minor portion of the experimental data, PCA is likely to misrepresent the pattern and timing of task activation. Furthermore, there are not always clear statistics for evaluating the significance of components. The best applications of PCA have involved combining it with other analysis techniques. For example, see the 2007 study by Ecker and colleagues, who initially identified subregions of visual cortex using standard hypothesis-driven techniques, and then used PCA to explore the functional connectivity among those regions.
Independent components analysis (ICA)

Most of the variance in any fMRI dataset will be uninteresting, given the small amplitude of the BOLD signal compared with sources of noise (see Chapter 8). Thus, the first few components extracted by PCA may all reflect noise, and because of the requirement of orthogonality, each of those components influences what is extracted subsequently. A relatively new statistical approach that overcomes this limitation is independent components analysis (ICA), which was introduced to fMRI in 1998 by McKeown and colleagues. An ICA analysis assumes that fMRI data consist of a set of spatially overlapping components, each with an independent spatial pattern and different time course (i.e., they contribute differently at different points in time; see Figure 11.2). The term “independent” means that the algorithm minimizes the overlap between the components; however, unlike PCA, the components do not need to be orthogonal with each other. Effectively, each ICA component is independent of all other components.

The ICA algorithm takes as its input the four-dimensional fMRI data set (i.e., the changes in intensity over time in all voxels) and extracts a set of spatiotemporal patterns such that the differences between the patterns are maxi-
mized. Each component describes the changes in activation within a defined set of voxels, but the activation of those voxels can change dramatically over time. For example, the voxels identified within one component could be activated in a task-related manner during some blocks, but not during others. A given voxel can be a member of more than one component, and its overall time course of activation will be split among those components, which may reflect both task-related activation and sources of noise (Figure 11.3). The most common ICA technique is spatial ICA, which emphasizes spatial independence by minimizing redundancy in the spatial maps of the components. However, researchers can also conduct temporal ICA, which minimizes redundancy in

**spatial ICA** A form of independent component analysis that generates components that have minimal spatial redundancy.

**temporal ICA** A form of independent component analysis that generates components that have minimal temporal redundancy.

---

**Figure 11.3** Extracting task-related and non-task-related components using ICA. In this session, the subject viewed photographs of faces and of photographs of houses in alternating 40 s blocks. A series of components, four of which are shown here, were extracted using ICA. (A) Within voxels in the fusiform gyrus and the dorsal parietal cortex, there was a task-related modulation of activation. Other components identified by ICA were associated with non-task-related variability, such as a transient scanner artifact (B), high-frequency physiological noise around vessels in the base of the brain (C), and head motion (D), which was evident as characteristic changes in activation along the edge of the brain.
the time courses of the components. Calhoun and colleagues demonstrated in 2001 that the choice between spatial or temporal ICA can have dramatic effects on the results, particularly if the underlying processes are correlated. For example, if the original data contain multiple components with a great degree of spatial overlap, then an ICA approach that attempts to maximize spatial independence may be invalid.

The application of ICA to fMRI data can result in the identification of sets of voxels with similar BOLD signal fluctuations over time, even if those voxels are distributed in different parts of the brain, and even if those voxels are influenced by different sources of noise. For example, in 2005 Beckmann and colleagues applied ICA to data collected while subjects were resting quietly in the scanner (see also Box 9.1). They found several independent components that tended to have similar spatial patterns in different subjects (Figure 11.4), suggesting that those components may reflect functional networks in the brain.

By comparing the patterns of activation with prior fMRI work involving active tasks, the authors could identify components associated with vision, audition, executive control, self-directed thought, and other processes. Note that each of these components had a different time course in every subject, because each individual engaged in these processes at different times. Yet, the ICA algorithm could separate those components without any prior input of information about when the processes occurred. Another use of ICA is to identify and remove components whose temporal or spatial properties suggest that they reflect task-unrelated noise. As shown in a 2008 article by Tohka and colleagues, relatively simple and automated decision rules can identify ICA components that are likely to be noise-related. Then, the task-related components can be combined into a new, cleaner time series that is used in standard analyses.

As it is commonly formulated, ICA separates components blindly. No assumptions are made about BOLD signal change or task properties. Nor can standard statistical methods be used for evaluating the significance of any single component. Thus, a major challenge for data analysis using ICA lies in matching components to mental processes (and resulting hemodynamics) evoked by an experimental task. Several approaches combine data-driven ICA with hypothesis-driven statistics to overcome these limitations. A hybrid ICA technique introduced by McKeown in 2000 first uses spatial ICA to identify a set of components and then uses regression analyses to draw conclusions about the task-relatedness of each component. This allows researchers to identify meaningful stimulus-evoked changes, while also exploring their data for potentially unanticipated effects. See the 2004 article by Beckmann and Smith for a discussion of alternative approaches to ICA significance testing. In recent years, the major fMRI analysis packages have begun to incorporate ICA algorithms (e.g., the MELODIC module within FSL), making ICA broadly available to the research community. Despite its limitations, ICA can provide valuable insights into fMRI data, and we recommend that researchers explore its uses for their own studies.

**Partial least squares (PLS)**

Another important method for data-driven fMRI analysis is partial least squares (PLS). This approach was developed in 1996 by McIntosh and colleagues for data collected using positron emission tomography (PET), and was subsequently extended for use with complex fMRI designs in 2004 by the same researchers. Like PCA and ICA, PLS identifies core features that contribute to variability in the fMRI signal. But unlike those techniques, it does not blindly identify components. By incorporating task information at an early stage, it
Figure 11.4 Using ICA to identify brain networks showing high resting-state connectivity. By applying data-driven analyses to fMRI data collected while the subjects were resting quietly, researchers could separate distinct networks whose activation tended to fluctuate similarly over time. The likely functions of these networks can be inferred based on prior work using active tasks and hypothesis-driven analyses. Shown here is a set of eight distinct brain networks, all of which were spatially consistent between subjects. Roughly characterized, these networks correspond to the following types of processing. (Adapted from Beckmann et al., 2005.)

identifies components whose amplitudes change between conditions, and can thus be attributed to the experimental manipulation. It uses a mathematical approach called singular value decomposition, which has conceptual similarities to the approach used in PCA, to identify a series of orthogonal compo-
latent variable A variable whose value is not directly measured, but is inferred based on the values of other variables.

permutation In the context of significance testing, approaches that involve resampling the original data to determine the size of an effect that might be observed with a given alpha level.

hyperscanning The simultaneous collection of fMRI data from two or more subjects who are interacting in an experimental paradigm.

PLS intrareactions Upon first consideration, this results activation. It scan-the experimentill differences in the hemodynamic weights the involvs randomizing what variables, like response time or accuracy, can easily be included in the analyses. Once identified, components can be evaluated statistically using permutation tests. This involves randomizing the assignment of conditions among the events (or blocks) and then repeating the analysis; conducting many such tests generates a distribution of significance values under the null hypothesis, to which the measured significance value can be compared. Thus, PLS provides a rigorous method for evaluating whether specific voxels are influenced by the experimental task. Nevertheless, it shares some limitations with the regression approaches described in the previous chapter. Any errors in the labeling of the experimental conditions will propagate throughout the analysis, leading to results that are difficult to interpret. This and other problems can be minimized by using PLS in conjunction with other approaches. For example, one can follow a traditional regression approach, and use a PLS analysis to target a particular comparison of interest.

Between-Subjects Correlations: Hyperscanning

In fMRI, data are collected from one individual at a time. Many recent studies have involved interactions between individuals: physical actions like tickling and touch, perception of others’ actions, and even competition in multiplayer games. Yet such fMRI data provide information about individuals who are engaging in these social interactions, not about the interactions themselves. Recent work involving a technique called hyperscanning (Figure 11.5) overcomes this limitation by collecting fMRI data from two or more subjects, each in separate scanners, while they interact with each other.

Hyperscanning introduces some technical challenges to fMRI studies. To improve comparisons between individuals, the scanner hardware and pulse sequences are generally kept as similar to each other as possible. Most importantly, both the data collection and stimulus presentation must be tightly synchronized between the scanners; otherwise, little will be gained over individual scanning sessions. A research group led by Read Montague has developed software protocols that coordinate scanning over the internet. Using these protocols, two subjects might lie in scanners thousands of miles apart, and compete with each other in an economic game! Upon first consideration, this might seem more a technical triumph than a novel analysis technique. Most researchers interested in social interactions place one subject in the fMRI scan-
Hyperscanning fMRI. An intriguing approach to studying the neural basis of social interactions involves the collection of fMRI data from two (or more) subjects simultaneously, while they interact with each other (e.g., in an economic game, as in this schematic figure). Hyperscanning requires coordination of the experimental stimuli, so that the actions of one subject can influence the other, and of the fMRI scanners, so that the fMRI data from the two subjects can be compared. The chief advantage of hyperscanning fMRI is that it allows researchers to directly relate the activation in one person's brain to the activation in another person's brain, potentially identifying coordinated activation that might be difficult to find using standard techniques.
showed their trustworthiness, this activation shifted to earlier and earlier in time, and was accompanied by a parallel increase in activation within the cingulate cortex of the investor. The researchers concluded that these two regions constitute a system for evaluating the intentions of others, and thus together mark the progress of social interactions over time.

All interactions between individuals in a hyperscanning experiment are mediated by the intervening behavior; there are, of course, no direct influences between the two brains. Thus, in principle, any conclusion that could be drawn from hyperscanning fMRI could also be reached by scanning individuals one at a time. However, there are practical advantages to using hyperscanning. Because the subjects are participating in a common social setting, many aspects of the experimental task can be matched. Moreover, using one person's brain activation as a predictor of another's activation may result in the identification of temporal regularities that were not expected. In essence, if sufficient technical resources are available (i.e., two or more research scanners) hyperscanning might be used to spark new ideas about the neural mechanisms of social interactions. See Box 11.1 for another approach to combining data from different subjects: using correlation techniques to understand shared aspects of cognition when different people process the same complex visual stimulus.

### Functional Connectivity Approaches

Most studies of brain function build on the concept of localization of function, that different brain regions support different forms of information processing. Yet no brain region exists in isolation. Information flows between regions via the action potentials carried by axons, which are bundled into large fiber tracts. For more than a century, neuroanatomists have mapped the anatomical connections between brain regions, in an attempt to understand the structural connectivity of the brain. Generally, this research has involved dissection and axon staining in brains of dead humans and non-human primates. While much remains to be discovered, the study of the anatomical connections between brain regions has provided a cornerstone for neuroscience research.

Despite the value of this anatomical research, knowledge of structural connections between brain regions can only provide a limited picture of information flow in the brain. Descriptions of functional connectivity, or how the activity of one brain region influences activity in another brain region, are also needed. Many researchers interested in functional connectivity have adopted fMRI techniques, because of their utility for measuring changes in activation throughout the entire brain. Studies of functional connectivity typically involve standard methods of fMRI data collection, but with relatively high temporal resolution. However, different methods for data analysis and statistical inference are used in functional connectivity studies compared with other types of fMRI research. As we will discuss in the following sections, functional connectivity analyses can extend either hypothesis- or data-driven analyses to draw more specific inferences about the direction and timing of information flow across brain regions.

### From coactivation to connectivity: a conceptual overview

Functional MRI holds significant promise for elucidating the functional relationships between brain regions because of its spatial coverage. Most fMRI studies involve the collection of data from the entire brain every few seconds,
BOX 11.1 Intersubject Correlations in Free Viewing

Most fMRI studies treat their participants as independent entities, each performing the same task. Standard hypothesis-driven analyses combine data from different subjects to emphasize common effects and minimize subject-specific effects, thus improving the power of their statistical inferences. Consider, however, an alternative perspective on combining data from different subjects: if a group of individuals all show the same activation, regardless of the experimental hypothesis, then that shared activation likely reflects common mental processing. Here, we introduce the idea of looking for intersubject correlations, or common patterns of activation across subjects. Like other data-driven approaches to fMRI analysis, this involves looking for regular patterns in the data and then interpreting those regularities based on knowledge of the experimental paradigm. But unlike the approaches discussed elsewhere in the chapter, those regular patterns will be between subjects rather than within individual subjects.

An elegant example of intersubject correlations was reported by Hasson and colleagues in 2004. The authors were interested in understanding the brain processes that underlie perception in open-ended, natural settings, such as when people freely perceive and attend to distinct parts of the complex, changing world. This goal poses challenges for conventional hypothesis-driven analyses, given that the freedom of subjects to process visual stimuli idiosyncratically makes it difficult to identify what processes are ongoing at any one time. So, they hypothesized that complex perception, such as that engaged in while watching a movie, contains many processes that are common across individuals, and that those processes should be reflected in similarly common patterns of activation.

They showed five subjects a thirty-minute excerpt from a classic western movie, Sergio Leone’s “The Good, the Bad and the Ugly.” Subjects were not instructed to perform any particular task while watching the movie, but to view it normally and describe the plot at the end of the experiment. Then, the researchers simply collected a time series of fMRI data, using standard protocols for data collection, while their subjects watched the movie. After applying standard preprocessing steps, including normalization of all subjects to the same stereotaxic space, the authors then identified voxels whose time courses were significantly correlated between different pairs of subjects (ten unique pairings). They found significant across-subject correlations in about 30% of the cortical surface, including many parts of the visual cortex and the visual processing streams, as well as in regions within the frontal cortex (Figure 1). For a control condition, the authors compared functional MRI time courses that are shared by different individuals while performing the same experimental tasks or experiencing the same stimuli.

**Figure 1** Regions exhibiting a common pattern of activation in subjects watching the same movie. While in the scanner, subjects watched an extended clip from the movie “The Good, the Bad and the Ugly.” The common visual experience caused multiple brain regions to exhibit changes in activation that were roughly synchronized between subjects. As shown in this figure, these included regions associated with the visual processing of different stimulus types. (After Hasson et al., 2004.)
They repeated the experiment with a separate group of subjects who were lying in the scanner with their eyes closed. This resulted in only chance correlations between subjects.

Further analyses revealed that these intersubject correlations resulted from two different effects. First, activation throughout much of the brain, including most visual regions, rose and fell in a similar pattern across all subjects. To interpret this activation, the authors used a reverse correlation approach: they identified time points when this collection of regions showed maximum activation and then evaluated what was happening in the movie at those times. (This process is conceptually similar to the reverse-inference approach that will be discussed in Chapter 14.) They found that this non-spatially-selective component tended to have highest amplitude during the most surprising and evocative points in the movie (e.g., gunshots, explosions, or unexpected plot twists), and thus it could reflect a broad increase in arousal. Second, there were spatially-selective intersubject correlations that had a unique time course in each of several brain regions. To interpret these activations, the researchers again looked back at the movie. As shown in Figure 2, common activation in the fusiform cortex tended to increase in response to movie scenes that involved a close view of a face, while activation in the anterior collateral sulcus (adjacent to the parahippocampal gyrus) tended to increase in response to scenes that contained views of environments or buildings. While these regularities may seem straightforward given the known functionality of those regions, the authors also found some more subtle functional effects. For example, a region in the anterior parietal lobe exhibited regular increases in activation in response to a wide range of scenes. By looking for common elements in the movie, the authors concluded that this activation was probably evoked by on-camera hand movements. This basic technique has since been applied to other sorts of complex stimuli; for an example with auditory speech.

Figure 2 After identifying brain regions that exhibited significant correlations between subjects, the researchers then examined the movie to identify content that resulted in those correlations. In the region of the fusiform gyrus shown at right, they found that the largest amplitude responses were reliably evoked by scenes that contained a close-up view of a face. (Adapted from Hasson and colleagues, 2004.)
BOX 11.1 (continued)

stimuli, see the 2008 article by Wilson and colleagues. A complementary approach was introduced in a series of studies by Bartels and Zeki, who also investigated natural viewing conditions, but using hypothesis-driven rather than data-driven analyses. In a 2005 study, these authors investigated interactions between brain regions while subjects passively viewed and listened to an extended movie clip—the James Bond movie "Tomorrow Never Dies"—interspersed with infrequent periods of darkness. They first used ICA to identify sets of brain regions whose activation moved together over time, and then examined the functional correlations between those regions to estimate the relative strengths of their connectivity. They found, as expected based on known anatomical structure, that the correlations between homologous pairs of regions in the two hemispheres (e.g., left and right area V3) were generally greater than between non-homologous regions. Interestingly, the non-homologous regions were actually more correlated while the display was blank. This suggests that during movie viewing these non-homologous regions contributed to different aspects of perception, while during the blank screen sessions their joint activation was driven by non-specific arousal processes. Note that this latter approach does not seek common changes in the raw fMRI time courses across subjects, as investigated by Hassen and colleagues, but instead looks for common patterns of connectivity. reverse inference Reasoning from the outcome of a dependent variable to infer the state of an independent variable (or an intervening unobservable variable).

providing complete spatial coverage at moderate temporal resolution. Advanced pulse sequences now allow the collection of more than 20 slices per second, enabling sampling of the entire brain at roughly 1 Hz. Electro-physiological techniques may provide more complete characterizations of single brain regions, because they sample post-synaptic potentials at 1000 Hz or faster or can be used to identify individual action potentials. However, the spatial coverage available with fMRI provides some advantages for describing functional relationships between regions.

The simplest type of relationship between regions is coactivation, in which two or more distinct brain regions show simultaneous activity during an experimental task (Figure 11.6A). A classic example of coactivation can be seen in motor tasks: squeezing your left hand will result in activity in the precentral gyrus in the right hemisphere and the cerebellum in the left hemisphere. Hypothesis-driven fMRI analyses are designed to identify brain regions that show significant activity in response to some experimental manipulation, so by their very nature they provide coactivation data. It is common for fMRI researchers to report that a given experimental task evoked concurrent activity in multiple brain regions (e.g., that a working memory task evoked activity in the prefrontal and parietal cortices), and to infer from this that these regions are part of a single functional system.

However, while coactivation tells us that two brain regions are related, it tells us nothing about the form of that relationship. Consider the simple task of encoding a set of three faces into working memory. Subjects performing this task will show coactivation in the dorsolateral prefrontal cortex, superior parietal cortex, and fusiform gyrus. What can we conclude from this? One possibility is that top-down influences from the prefrontal cortex guide activity in the other regions. Another possibility is that visual processing in the temporal and parietal lobes may lead to prefrontal activity in a bottom-up fashion. Or, activity in all three regions may be triggered by signals from yet another source. Because of the inherent causal uncer-

coactivation The simultaneous activation of two or more brain regions within a single experimental task. Coactivation of brain regions does not imply that the regions are functionally connected.
figure 11.6 From coactivation to connectivity. (a) Any complex experimental task will evoke activity in several brain regions simultaneously. The regions indicated in this color map had significantly greater activity during a rest condition than during a visual search task with similar perceptual characteristics. Even though these regions are coactive, they do not necessarily serve the same function. (b) By looking at patterns of coactivation during different experimental conditions, we can make inferences about connectivity (i.e., the influence of one brain region on another). If one condition activates areas A and C, and another condition activates areas B and C, then we can begin to make inferences about the connectivity between these three regions. (A from Huettel et al., 2001.)

network A description of the relationships among a set of brain regions, including their connectivity and causal relationships.

system A set of regions that are linked by coactivation, but for which the connectivity and causal relationships remain unknown.

double dissociation The demonstration that two experimental manipulations have different effects on two dependent variables. One manipulation affects the first variable but not the second, and the other manipulation affects the second but not the first.

tainty in any single map of the brain, fMRI research has often been dismissed as merely descriptive (i.e., describing patterns of coactivation) rather than mechanistic (i.e., explaining how the brain accomplishes a complex behavior). Patterns of coactive regions are often erroneously summarized as networks of brain activity. This misuses the term "network," which should be restricted to descriptions of structural and functional connectivity that include the causal flow of information. For descriptions of simple associations among regions (i.e., with unknown connectivity and causality), a more appropriate term is system.

Nevertheless, fMRI can be used to build more sophisticated models of connectivity. Some aspects of connectivity can be deduced by measuring the covariance in activities among regions during different experimental conditions. Consider a simple example that consists of three brain regions, A, B, and C, and two experimental conditions, X and Y (Figure 11.6B). Under condition X, brain regions A and C are coactivated, but B is not active. Under condition Y, brain regions B and C are coactivated, but A is not active. These results indicate a double dissociation between regions A and B, since one manipulation has an effect on A but not B, and another on B but not A. From this it can be inferred that these regions have different functional properties. Furthermore, the results for region C provide information about connectivity. The observed pattern of associations and dissociations supports the hypothesis that regions A and B are both functionally related to C but not to each other. Because activation at C does not always predict activation of A or B, the data suggest (but do not prove) directionality. This conclusion would be strengthened if there was a measurable delay between activation in A or B and activation in C. However, the differences in the latency of hemodynamic timing throughout the brain either could reduce the impact of such results or could make such a delay undetectable. Of course, this example simplifies typical fMRI data, which often have a much more complex set of interrelationships.

Thought Question

Many scientists believe that the demonstration of a double dissociation is a necessary step for establishing a clear relationship between independent and dependent variables. Given what you know so far, what do double dissociations tell you that single dissociations do not?

Ideally, we would like models of connectivity to include information about unidirectional connections (i.e., A influences B) as well as feedback connections (i.e., A influences B, which leads to changes in A itself). In order to restrict
the number of possible functional connections within a model, researchers incorporate information about the anatomical connections between brain regions. Although most connections between brain regions are clearly bidirectional, there are a number of relatively unidirectional pathways or loops, especially between cortical and subcortical regions. The connection between the prefrontal cortex and the basal ganglia is one such pathway; the prefrontal cortex has heavy projections to the basal ganglia but receives little direct input in return. Instead, the basal ganglia project to the thalamus, which itself projects back to the prefrontal cortex. By incorporating information about these pathways into models of brain function, researchers can create models that are detailed and directional networks, rather than simple cognitive systems.

Resting-state connectivity

The most straightforward evidence that two (or more) regions share functional connectivity comes from studies of resting-state connectivity. These studies are used to identify synchronous BOLD changes in multiple brain regions while subjects lie in the MRI scanner but do not perform any experimental task. These changes are thought to represent certain intrinsic operations of the brain, or to reflect stimulus-independent processing, rather than the effects of external stimuli. While resting-state connectivity can be found in any aspect of the BOLD signal, it is often evident in periodic and very slow changes in activation (e.g., 0.01 to 0.1 Hz). This was first demonstrated in the sensorimotor cortex by Biswal and colleagues in 1995 (Figure 11.7).

Figure 11.7 Similar patterns of resting-state and task-related functional connectivity. Functional MRI data using BOLD-sensitive (A) and blood-flow-sensitive (B) imaging were collected while the subject performed a bilateral finger tapping task. Significant activity in the motor cortex was found using both techniques. Resting-state data were also collected using each type of imaging. Low-frequency oscillations that signal functional connectivity between regions were identified. The BOLD resting data clearly indicated that the hand areas of the motor cortex were functionally connected (C), matching the activation data, while the flow resting data exhibited much less functional connectivity (D). (From Biswal et al., 1997.)
To identify low-frequency oscillations in the BOLD signal, researchers typically instruct subjects to relax without falling asleep, and then record fMRI data using normal pulse sequences. Because many other physiological processes can cause signal oscillations (e.g., respiratory and cardiac pulsations), it is important to separate oscillations that are potentially associated with neuronal activity from those associated with physiological fluctuations. One effective method is to sample at a very high temporal resolution, fast enough to identify cardiac pulsations (and slower respiration) without aliasing artifacts. Then, physiological variability can be identified and removed. Thereafter, broad patterns of common activation can be identified using data exploration techniques like PCA or ICA. Alternatively, a researcher can identify a seed voxel in a specific region of interest (ROI), and then examine what other regions exhibit similar fluctuations in activation over time. Using this approach, researchers have found coherent activation in sets of regions involved in visual, auditory, memory, and attentional processes, among others. In general, regions that are coactivated during active tasks also show resting-state connectivity. This suggests that brain regions with similar functionality tend to express similar patterns of spontaneous BOLD activation (and presumably neural activity, as well).

Studies of resting-state connectivity can provide insight into the properties of the so-called default network. Imagine that you are a subject in one of these studies. As you enter the scanner, the experimenter instructs you to close your eyes and relax, asking you to remain still without thinking about anything in particular. Perfect adherence to these instructions is, of course, impossible. One cannot shut off active thought in the same way that one closes a running faucet; indeed, the very act of remembering the task instructions violates them. You will spontaneously engage in all sorts of cognition, from remembering recent events to daydreaming about the future, and these thoughts lead to systematic and coordinated activation throughout the brain. However, not all aspects of resting-state connectivity can be attributed to ongoing complex cognitive processes. For example, in a 2007 study, Vincent and colleagues scanned anesthetized monkeys and demonstrated robust resting-state connectivity in various brain systems. These and related findings indicate that coherent low-frequency fluctuations in the BOLD signal may reflect some general property of brain function.

Combining resting-state and task-related connectivity within the same study can strengthen conclusions about the relationships between brain regions. An early example of this type of study was published by Hampson and colleagues in 2002. They collected separate data sets from the same subjects in two types of runs: resting and task. In the resting runs, nothing was presented and no response was required. The task runs consisted of a blocked design in which the subjects listened to either a reading of a story, or silence. The authors used an iterative comparison method that generated connectivity hypotheses based on one data set and tested those hypotheses using the other data set. The basic approach is shown in Figure 11.8. First, the authors used a standard blocked analysis on the task data set to create a task-related activity map. As expected, they found significant activity in Broca’s area and Wernicke’s area. This resulted in the hypothesis that these two regions are functionally related. To test this hypothesis, they used Broca’s area as the seed for a connectivity analysis. The voxel time courses from the resting data set were low-pass filtered and then tested for correlation with the mean activity from Broca’s area. Strong connectivity was found between Broca’s area (in the left hemisphere) and the corresponding anatomical region in the right hemisphere, while weaker connectivity was noted both with Wernicke’s area and a region in the left pre-
motor cortex. In the second iteration of their analysis, the authors defined a new ROI in the left premotor cortex, based on this connectivity map. The hypothesis that Broca’s area and the left premotor cortex are functionally connected was then tested, using both the auditory and silence conditions of the task data set independently. Both conditions revealed significant correlations between these regions. Note that this iterative approach, in which one data set generates hypotheses to be tested with the other data set, could in principle be extended to additional brain regions (e.g., steps 3a and 3b in Figure 11.8).

**Psychophysiological interactions**

Another approach for combining fMRI and behavioral data involves the identification of **psychophysiological interactions (PPI)**, in a method developed by Friston and colleagues in 1997. The core concept of PPI is that mental processes, whether evoked by a stimulus condition or by the subject’s behavior, modulate the influence of one brain region over another. Like other connectivity approaches, PPI uses the time course of activation in one seed region to predict changes in activation in another region. However, the process for identifying PPI sets this method apart from other techniques. First, PPI estimates the neuronal activity in the seed region that would have generated its measured fMRI time course. Second, that estimate is multiplied by the task timing, resulting in a prediction of neuronal activity associated with the experimental manipulation.
This new prediction is then convolved with the fMRI hemodynamic response to generate the predicted PPI BOLD time course, which is then used as a regressor in the analysis design matrix to identify activation in other brain regions. The interactions identified by this approach are named "psychophysiological" because the approach is intended to model interactions between cognitive processes and physiological changes in the connectivity between brain regions.

An intriguing recent example of the power of PPI was reported by Bingel and colleagues in 2007. They were interested in the power of painful stimuli to quickly capture attention away from important tasks. One possible explanation for this effect was that painful stimuli engage the same brain systems as other sorts of distractors. For example, previous work from the same authors had shown that cognitive distraction, achieved by requiring subjects to remember letters presented at the center of a display, attenuated visual cortex activation in response to task-irrelevant background images. Another possibility, which they adopted as their research hypothesis, was that painful stimuli could evoke a similar effect (i.e., attenuating visual cortex activation), but through a different neural pathway. In their fMRI sessions, the authors tested both painful and cognitive distraction using PPI methods. To create painful distractions they shone infrared laser light onto the hands of their fMRI subjects. The associated pain evoked activation in the insular, cingulate, and somatosensory cortices. And, as expected, a difficult cognitive distraction task increased activation in prefrontal and parietal cortices. They focused their PPI analyses on one region related to pain (the anterior cingulate cortex) and one region related to cognitive difficulty (the parietal cortex), and examined whether these regions' activation predicted condition-specific effects in the visual cortex (i.e., decreased activation to passive viewing of low-contrast, compared to high-contrast, background images). As shown in Figure 11.9, their research hypothesis was confirmed: the cingulate cortex activity exerted a pain-related modulation on the lateral visual cortex, whereas the parietal cortex exerted a cognitive-difficulty-related modulation on the same region. Thus, by using PPI, the authors demonstrated that both pain and cognitive difficulty could attenuate visual cortex processing, but via different neural pathways.

Inferring causality from fMRI data

One well-recognized limitation of fMRI research is its difficulty with assigning causality, or determining that a particular pattern of fMRI activation led to a particular behavior or mental state. In Chapter 9, we raised this issue by discussing the idea that fMRI data are epiphenomenal, meaning that they do not reflect information processing itself, and thus cannot be used to produce models of brain function. While this fundamental objection can be overcome by considering that fMRI data provide a valuable index of the underlying information processing, more subtle causal problems remain. Suppose that we observe fMRI activation in two regions, A and B. How can we determine whether activation in region A caused changes in region B, activation in Region B caused changes in Region A, or that both were influenced by some other region? Determining how regions interact is necessary for understanding the flow of information throughout the brain, which in turn allows researchers to construct more biologically plausible models of brain function.

The limited temporal resolution of fMRI data poses many challenges for causal modeling. Nearly all connections between brain regions are bidirectional, such that the flow of information in one direction may be shortly followed by a flow of information in the other direction, and this back-and-forth

causality  The judgment that one event led to another. For fMRI data, measures of causality attempt to determine whether changes in the activation of one brain region led to changes in the activation in another region.

epiphenomenal  A secondary consequence of a causal chain of processes, but playing no causal role in the process of interest.
signaling may occur repeatedly in the repetition time between consecutive image acquisitions. Collecting data with a relatively short TR greatly improves the robustness of causal analyses. Moreover, fMRI does not measure neuronal activity directly, but instead measures a hemodynamic marker of that activity. Thus, any small time differences in activation between regions may simply reflect differences in the timing of hemodynamic responses in those regions (see Chapter 7). Likewise, there exists strong evidence that integrative activity constitutes a major contributor to the BOLD signal, through the metabolic demands associated with restoring dendritic membrane potentials. Therefore, some BOLD signal changes could, at least in principle, represent feedback from downstream regions rather than signaling output, completely skewing assignments of causality. These and related caveats indicate that sophisticated approaches are needed for inferring causality using fMRI data.

A powerful technique for investigating relationships among regions activated in an fMRI study is structural equation modeling (SEM), which identifies the combination of connections between variables that can best account for the observed data. An SEM model consists of a set of nodes, each representing a single measured or latent variable from the data set, and a set of paths between those variables (Figure 11.10). By including both direct and indirect paths between variables, the model can evaluate whether one variable mediates the covariation between two other variables. (Note that when the model includes only measured data and no inferred or latent variables, the approach is a sub-

Figure 11.9 Identification of a psychophysiological interaction (PPI). Researchers hypothesized that processing in the visual cortex could be modulated by input from either of two regions, depending on the task context. (A) Using a PPI analysis, they found that the magnitude of pain-related activation in the anterior cingulate cortex modulated the influence of image contrast on the lateral occipital cortex, a region important for processing complex visual scenes. (B) Under conditions of cognitive distraction, however, the magnitude of working-memory-related activation in the parietal cortex now modulated the image-contrast effects in the visual cortex. Thus, PPI can distinguish modulatory contributions from different regions, which change according to task conditions. (After Bingel et al., 2007.)
Figure 11.10 A conceptual illustration of fMRI structural equation modeling (SEM). The researchers establish a model of functional connections among a set of brain regions, often based on prior knowledge of the structural relationships between the regions. As in the hypothetical example shown, the model can involve both direct (A to B) or indirect (A to C to B) connections between regions. External stimuli or cognitive states can be shown to directly influence the activation of a region (e.g., stimulus inputs influence activation in A) or the connectivity between regions (e.g., increased cognitive control leads to a greater influence between E and D). Note that although the connections are shown as directional for clarity of this example, standard SEM models do not include information about directionality.

Granger causality A form of time series analysis that quantifies the information gained by using the past history of one variable to improve predictions of future values of another variable.

set of SEM called path analysis.) When setting up an SEM model, researchers often postulate connections between brain regions that are suggested by known anatomy, so that the pathways in the model reflect connecting pathways between brain regions. The first use of SEM in fMRI was published in 1997 by Buchel and Friston, who studied the effects of attention on processing in the visual cortex. They found that attention increased the connectivity between the extrastriate visual areas and the posterior parietal cortex. The prefrontal cortex was one potential source of this attentional control effect. More recent studies have used SEM both to evaluate models of connectivity between regions and to measure the effects of other factors on that connectivity. SEM can be used to elucidate many types of relationships between brain regions, and can be used to rule out hypothesized models for interactions. However, it has an important limitation: it does not include information about the relative timing of activation. Instead, it considers interactions between variables (e.g., brain regions) to occur instantaneously, with changes in those variables (e.g., activation in a region) determined by the values of other variables and the strengths of the connecting paths. In other words, in fMRI applications of SEM, possible causal relationships between brain regions are established beforehand within the model, not determined from the data.

To accommodate information about the relative timing of activation, researchers have adapted methods from time series modeling, specifically the idea of Granger causality. This approach, introduced by the Nobel laureate Clive Granger, has been enormously influential in economics and other social sciences. To understand Granger causality, consider two time series A and B, which could represent the BOLD signal measured in two brain regions. Time series A can be said to exert a Granger causal influence on time series B, only if knowledge of the past history of A improves predictions of the future values of B even when the past history of B (and of other variables) is included in the predictive model (Figure 11.11). From the perspective of fMRI, this would mean that knowledge of the activation in one source region (e.g., the left auditory cortex) would allow us to predict changes in the activation in another target region (e.g., Broca's area) at a future time point, with better accuracy than if looking at past activation in the target region alone. In essence, Granger causality quantifies how much, if any, unique information carried by the past history of one variable can be used to predict another variable. Note that
Granger causality is not a measure of true, physical causality. Because it makes the seemingly reasonable assumption that causes precede effects, in fMRI data the sluggishness of the BOLD time course could lead to apparent causal reversals. Moreover, if two time courses are highly redundant, a model that fails to account for possible intervening variables could lead to spurious conclusions. Even with these caveats noted, Granger causality does provide a useful mathematical framework for inferring influences between regions.

A data-driven implementation of Granger causality for fMRI was reported by Roebroeck and colleagues in 2005 (Figure 11.12). First they simulated a causal link between two brain regions, such that changes in the amplitudes of local field potentials in Region A altered the amplitudes of local field potentials in Region B. Through a number of simulations they tested delays of between 0 to 100 ms and variable levels of causal influence. They then convolved these simulated patterns of neuronal activity with a BOLD response function, to generate possible time courses of fMRI activation in those two regions. Their simulations demonstrated that neuronal delays as short as 50 ms could, in principle, be identified within fMRI data sampled at TRs of 500–1000 ms. (As noted above, the lower temporal resolution of fMRI data compared with the actual neuronal signals can lead to misidentification of the direction of causality. Thus, the authors argue that researchers should statistically compare both directions of causality: whether A to B causality is likely to be greater than B to A causality.) They next examined real fMRI data collected at a TR of 1 s while subjects performed a visual processing task that involved viewing photographs of faces and houses. Using a region of the fusiform gyrus known to be face-selective as their reference location, they found evidence for strong causal inputs from the early visual cortex and strong causal outputs to the parietal cortex. Thus, their Granger causality technique could recover known organization within the visual processing stream, even though images were only sampled at 1 s. For comparison, electrophysiological data indicate that the neuronal signals in these regions differ by only about 100 ms in their timing.
Another approach to understanding connectivity between brain regions is dynamic causal modeling (DCM), which was introduced by Friston and colleagues in 2003. Researchers using this method create a model of the functional connections between brain regions, and then evaluate how that connectivity changes as a result of experimental manipulations. Thus, they attempt to infer not just connectivity itself, but task-related changes in connectivity. In principle, knowledge about how experimental manipulations perturb the connectivity between brain regions can lead to better inferences about causality. The key idea of DCM is that experimental stimuli (or task conditions) can change the brain in either of two ways: by modulating a brain region directly (e.g., auditory input evokes activation in the auditory cortex) or by changing connectivity between two regions (e.g., effects on attention). To use DCM, researchers set up a model of relevant brain regions and their hypothesized interactions, and then use the observed fMRI data to evaluate the plausibility of that model. The neuronal activity in each of those brain regions is estimated as a deconvolution of the observed BOLD signal, and the interactions between regions...
regions are inferred according to changes in the neuronal activity, not the BOLD signal itself. In 2008 Marreiros and colleagues extended the DCM approach to include multiple neuronal populations within each brain region, in order to better accommodate limitations in the spatial resolution of fMRI data.

An interesting example of the use of DCM was reported in 2008 by Kasess and colleagues. They were interested in motor imagery, or how individuals visualize physical movements. For example, basketball coaches will often tell players to visualize the sequence of movements before shooting a free throw under pressure. Prior neuroimaging work had clearly shown that brain regions involved with motor planning (e.g., the supplementary motor area, SMA) are activated by motor imagery, but the involvement of the primary motor cortex (M1), if any, remained unclear. The authors limited their imaging volume to four slices encompassing the motor cortex, and used high-speed fMRI with a TR of 300 ms. Their subjects participated in two tasks. In the motor execution task, subjects actually pressed a series of buttons when they heard an auditory cue, whereas in the motor imagery task, the same subjects only imagined pressing the buttons following the auditory cue. The authors tested a set of models of connectivity between the M1 and the SMA, including potential task-related influences on that connectivity. The best-performing model (Figure 11.13) indicated that imagery not only reduces activation in the SMA but also suppresses the connection between the SMA and the M1. This result demonstrates how DCM can extend the conclusions of fMRI studies beyond just activation of different regions, so that researchers can draw conclusions about how one region might suppress or enhance activation in another.

Combining fMRI and DTI

Functional connectivity analyses can be applied to any set of regions, regardless of whether those regions are really connected anatomically. Yet signals from one region to another travel via axons, which are typically organized into large-scale fiber tracts. Activation across broad regions might also reflect widespread changes in brain chemistry (e.g., neurotransmitter concentrations), which generally occur on a slower time scale than the scale typical for fMRI. Decades of neuroanatomical research have resulted in the mapping of the major fiber tracts, and provided a general wiring diagram of the human brain. All neurologically normal individuals share the same fiber tracts; for example, the left and right hemispheres always connect via the corpus callosum, while the frontal and anterior temporal lobes always connect via the uncinate fasciculus. However, the integrity of these tracts (e.g., the myelination of the constituent axons) does vary between individuals. Researchers can collect information about both the location and the integrity of fiber tracts via diffusion tensor imaging (DTI), which was introduced in Chapter 5. Like all other forms of MRI, DTI is non-invasive, and thus it allows researchers to measure properties of fiber tracts in normal human subjects.

By itself, DTI provides no information about brain function—it only provides information about brain structure. DTI data must be combined with other data to draw inferences about function. In one type of approach, DTI can be used to identify the existence and locations of fiber tracts that connect regions activated in an fMRI study. However, this approach has many limitations. Suppose you collect fMRI and DTI data from the same subjects. Your fMRI data show that both the lateral prefrontal cortex and the basal ganglia are activated by your task, and your DTI data reveal a fiber tract connecting these activated regions. What can you infer from these results? Even though the combined...
fractional anisotropy (FA) The preference for molecules to diffuse in an anisotropic manner. An FA value of 1 indicates that diffusion occurs along a single preferred axis, while a value of 0 indicates that diffusion is similar in all directions.

Data suggest that some functional connections exist between these regions, you cannot determine whether information flows from the prefrontal cortex to the basal ganglia, or in the other direction, or in both directions. Nor could you tell whether any such connections are direct, or occur via intervening synapses. Thus, while DTI can show that a connection exists between two regions, it does not necessarily demonstrate that the identified connection contributes to information processing. In another type of approach, the integrity of specific fiber tracts, as assessed through DTI measures like fractional anisotropy, can be used to predict behavioral performance. For example, Madden and colleagues reported in 2004 that older adults who have reduced fractional anisotropy in the anterior limb of the internal capsule, a fiber tract connecting the prefrontal cortex and thalamus, tend to have slower performance on a visual search task. This approach has conceptual similarities to studies of brain lesions (see Chapter 13), in that it relates structural differences between individuals' brains to functional differences in their behavior.

An elegant example of the power of combining DTI with fMRI data was published in 2007 by Andrews-Hanna and colleagues. Prior neuroscience research, initially from postmortem brain measurements and more recently from DTI, had demonstrated that normal aging was associated with degeneration of white matter throughout the brain. Under what is commonly called the “disconnection hypothesis”, this degeneration could reduce the efficacy of information flow between brain regions, leading to the broad and gradual decline in cognitive abilities observed with aging. To test this hypothesis, the researchers collected fMRI and DTI data from large samples of younger (18–40 y) and older adults (60–93 y). Their fMRI analyses evaluated the functional connectivity between the medial prefrontal cortex and medial parietal cortex, two of the central nodes within the default network (see Box 9.1). Consistent with prior research, these regions exhibited a high degree of functional connectivity in the younger adults; their time courses of activation were strongly correlated. This correlation was essentially absent in older adults (Figure 11.14A and B). For comparison, functional connectivity between other sets of regions (e.g., the left and right visual cortices) was the same in younger and older adults. Given these results, the authors tested whether the individuals who had the least functional connectivity also had the lowest fractional anisotropy in the white matter tract connecting these regions. This prediction was validated among the older adults (Figure 11.14C), even after controlling for the overall effects of age upon fractional anisotropy. To link these neural effects back to behavior, the authors demonstrated that those older adults with relatively well preserved functional connectivity between the medial prefrontal and medial parietal cortices tended to score better on a range of cognitive tests (Figure 11.14D). This study provides an excellent example of how fMRI, DTI, and behavioral data can be combined to support a simple conclusion: that age-related changes in the structural connections between regions alter their functional relationships, which in turn leads to broad deficits in cognitive function.

In summary, fMRI provides sufficient spatial and temporal resolution to identify connectivity between brain regions. Simple correlation approaches can be used to reveal regions that are functionally related, even in data sets collected in the absence of an experimental task. More complex hypothesis testing and modeling approaches allow estimation of the relative strength and directionality of the relationships between regions. DTI data can introduce anatomical constraints or suggest ROIs for connectivity models. The success of connectivity mapping rests on a simple principle: that complex behaviors and cognition depend on the computational contributions of a walter of differ-
ent anatomical regions. Even though each brain region projects to many others, considerable specialization remains. In 2002, Toni and colleagues introduced the idea that each region has a unique “connectional fingerprint” that represents its pattern of connectivity with other regions. We believe that this concept is a useful one, because it emphasizes that neuroscience analysis approaches benefit from anatomical considerations. Although all voxels may be treated similarly by standard fMRI analysis methods, they have different anatomical (and functional) connections. Connectivity mapping has the potential to elucidate these different properties of regions, thereby greatly improving the inferences that can be made from fMRI data.

**Prediction Approaches**

Suppose that you are a marketing researcher for a consumer electronics company in the year 2020. You assemble a focus group of individuals and bring them to your laboratory, which happens to contain an MRI

![Figure 11.14](image-url)
scanner. You place each person in your scanner and show them images and descriptions of the latest technological gadgets, while you measure activation evoked by each product. From this activation, and your prior knowledge of the brain regions associated with decision making, you attempt to identify which products (and ways of presenting those products) are likely to be successful in a competitive retail environment. While this example may seem fanciful at best—and fraught with ethical implications at worst (see Chapter 14 for additional discussions)—many fMRI researchers now use similar approaches for decidedly more scientific goals.

Prediction approaches reverse the typical direction of inference in fMRI research. The regression analyses described in the previous chapter explicitly treat task events as independent variables (i.e., forming regressors in a design matrix) and the observed fMRI data as dependent variables. In contrast, prediction approaches treat the fMRI data as independent variables that can predict some aspect of behavior, such as subjective experiences (e.g., looking at attractive versus unattractive images), subsequent memory (e.g., whether a word would be remembered at a memory test two weeks later), and simple purchasing decisions (e.g., whether to buy a new iPod). A variety of prediction approaches have been developed. Some studies label events according to subjects' behavior, not just the stimulus properties. Because these can be analyzed using the regression methods described in the previous chapter, we only briefly describe those approaches here. Others modify the general linear model to include some continuous, behaviorally defined predictor, either across subjects or across trials. Still others attempt to make inferences about the interactions between activation and some behavioral factor, integrating the connectivity methods described above with behavioral methods. And, some approaches use novel computational methods that use the joint changes in activation across sets of voxels to make their predictions.

Remember that none of these approaches provides clear information about causal relationships between brain regions, given the constraint due to the low temporal resolution of fMRI. In many ways, the distinction between "hypothesis testing" approaches and "prediction" approaches is more conceptual than methodological. However, that conceptual difference has important consequences for experimental practice. For example, some studies that attempt to predict behavior from fMRI data use a blinding approach: they create and test a model on part of the fMRI data (e.g., 3 of 4 experimental runs), and then validate the model on completely novel data (e.g., the final run). Whether or not a model is valid depends, in a very literal sense, on whether it can be generalized to new data. In a broad sense, we as fMRI researchers want to be able to predict the functional consequences of brain activation, not merely interpret the activation we observe. However, extreme caution should be taken in making these predictions; see Box 11.2 for examples in which the inaccurate prediction of functional properties could have life-changing effects.

**Predicting variation among individuals**

In the first fMRI studies, like that of Kwong and colleagues discussed in Chapter 7, each participant's data were analyzed individually—in effect, treating each individual as a separate experiment. These studies largely used very robust sensory or motor paradigms, which tend to evoke robust activation in every individual. However, as fMRI research began to address new cognitive phenomena, researchers began combining data from samples of participants. By doing so, the researchers increased their power for detecting effects that were common within the subject sample. In recent years, more and more studies have
**BOX 11.2 Real-Time fMRI**

Most fMRI analyses require substantial computational time. The preprocessing steps outlined in Chapter 8 involve complex signal processing algorithms, from the rigid-body spatial alignment required for motion correction to the temporal filtering used to remove task-unrelated variability. The multiple-regression analysis approaches described in Chapter 10 involve simple matrix operations, but even relatively simple computations can introduce substantial delay when repeated for many thousands of voxels, hundreds of time points, and dozens of subjects. Important analysis steps like correction for multiple comparisons involve additional, subject-specific computations. And, the data-driven approaches described in this chapter vary from relatively simple to remarkably complex. For most research fMRI studies using standard methods, the impressive power of modern computing hardware overcomes these challenges. For example, using a parallel-processing cluster, as is now available within many centers, an entire large-sample fMRI study can be analyzed within several hours. But for some applications, even this seemingly short delay can be unacceptable—those applications require real-time analyses. In this section, we consider two important applications of real-time fMRI: diagnostic scanning for pre-surgical planning and functional biofeedback.

**Mapping function for pre-surgical planning**

One common application of fMRI is the identification of functional areas that need to be preserved during neurosurgery. For many patients with brain tumors, vascular malformations, intractable epilepsy, or other types of pathology, surgical removal of the diseased portion of the brain provides the best clinical outcome. However, surgery may not be appropriate for all subjects, and the amount of tissue to be removed may depend on its functional properties. If tissue removal would result in damage to brain regions essential for language, memory, or primary sensory or motor processing, it could have a devastating effect on the patient’s subsequent quality of life. For example, damage to the left temporal lobe may render someone unable to comprehend speech. Thus, when deciding on the course of treatment for a given patient, neurosurgeons want to be able to evaluate the likely consequences of removing a particular brain region. Cortical mapping plays an important role in this evaluation process.

Since the 1950s, cortical mapping has typically been done during the surgery itself by temporarily probing selected brain regions using direct electrical stimulation of the cerebral cortex (see Chapter 13). This process involves placing an electrode on different parts of the exposed cortical surface. By sending a small electrical current between two adjacent electrodes, the surgeon can inhibit or excite local neuronal activity. In most cases, direct stimulation impairs the function of a brain region; for example, stimulation of language areas may render patients temporarily unable to speak. In other cases, stimulation may actually cause the patient to hear sounds or see colors, motion, or shapes, if the electrodes are placed near primary sensory areas. Although direct electrical stimulation is widely used, it has some disadvantages. The procedure is highly invasive and time-consuming, the patient must be awakened during surgery, the range of functions that can be investigated is limited, and the only cortical areas that can be measured are those near the exposed surface of the brain.

To overcome these disadvantages, a number of hospitals have begun investigating the use of fMRI for cortical mapping in surgical patients (Figure 1). The advantages of fMRI are straightforward. (continued on next page)

**Figure 1 Cortical mapping of function.**

This patient was undergoing cortical mapping in preparation for neurosurgery. The subject underwent both fMRI (A) and direct electrical stimulation (B). On the reconstruction of the patient’s brain in (A), the locations of the electrodes in the grid in (B) are shown as white spots. The electrodes are highlighted according to the effects that direct stimulation had on different tasks: orange, speech arrest; green, induced mouth or face movements; and blue, deficits in auditory comprehension and object naming. The locus of fMRI activity during a language comprehension task is shown on the brain surface in red. Note the overlap, although incomplete, between the fMRI and electrophysiological data.
It can be used to map functions anywhere in the brain, including deep-brain structures difficult to reach using electrodes. It is noninvasive and can be conducted in advance of surgery to aid in surgical planning. Furthermore, a wide range of cognitive, motor, and perceptual functions can be tested. Combining functional and conventional anatomical MRI in the same scanning session can provide maps of important functional areas superimposed on high-resolution images that show detailed anatomical structure and sites of tissue pathology. Multiple brain functions can usually be mapped within a 30- to 40-minute scanning session, providing neurosurgeons with important information to aid in planning the surgical approach. Functional MRI can also be used in conjunction with traditional electrophysiological approaches, by guiding the placement of electrodes for subsequent testing during the operation.

Despite its obvious potential, diagnostic fMRI has not yet been used widely, primarily because its analysis and interpretation is technically demanding.

**real-time analyses** A set of computational steps designed for the rapid analysis of fMRI data, so that statistical tests can be conducted immediately following the acquisition of images.

**pre-surgical planning** The use of fMRI or another neuroscience technique to map particular functions in a single individual, in order to guide clinical decisions about the potential course and consequences of neurosurgery in that individual.

**biofeedback** Providing an explicit indicator of some physiological process, such as the beating of the heart or activation within a particular brain region, so that an individual can attempt to regulate that activation or guide behavior.

Whereas most fMRI research involves pooling results from many subjects, as described in the previous section, diagnostic fMRI must produce interpretable activation maps in a particular individual. It is critical that each scan produce high-quality MR images with good functional resolution. An important factor in achieving reliable diagnostic fMRI capabilities has been the development of real-time analysis methods. In real-time analysis, the fMRI data are reconstructed, preprocessed (if needed), and statistically analyzed immediately following the acquisition of each image. Thus, the desired functional maps are available at the conclusion of the scanning session. Real-time analyses can also be very useful for standard fMRI experiments, because they provide a way to monitor the quality of the results. If the subject moves his head excessively or is unable to perform the experimental task, real-time analyses can catch the problem during the session so that it can be corrected. Real-time generation of statistical maps can also make scanning more efficient by allowing each scan to run only as long as is necessary to reach a satisfactory confidence level. For clinical cortical mapping, the ability to assess task performance and activation map quality while the patient is still in the scanner is crucial. Finally, the goals of significance testing may differ dramatically between experimental and clinical fMRI analyses, since experimental analyses always aim to minimize false positive results, while clinical analyses may instead aim to avoid false negative results.

**fMRI Biofeedback**

Research into biofeedback, or providing data that allows people to influence their own bodily states, has a long and somewhat checkered history. Traditional uses of biofeedback involved relaxation or stress relief: in a typical paradigm, the participant wears sensors that monitor sympathetic nervous system responses (e.g., changes in heart rate or skin conductance) and is shown the output of those sensors on a computer monitor. The subject's task involves minimizing the responses in those sensors through conscious strategies that might involve self-distraction, emotional regulation, or cognitive control.

Modern fMRI biofeedback uses information about the subject's own fMRI activation to influence the presentation of experimental stimuli. By its very nature, biofeedback requires extremely rapid real-time analyses. Whereas the typical goal of pre-surgical planning is to generate a map of activation at the end of each run, studies involving fMRI biofeedback must derive measures of activation that are updated immediately after the collection of each volume. Early studies of fMRI biofeedback indicated that such rapid analyses were possible, with some constraints. In one particularly clever example reported in 2007 by Weiskopf and colleagues, participants in two scanners played a game of "Brain Pong." Like in the old Pong video game, each participant moved a virtual paddle up and down on a computer monitor, attempting to bounce a ball past their opponent's paddle. But in this game, they moved the paddle not with a joystick but by increasing or decreasing activation in a region of interest within the brain.

What effects might biofeedback have on fMRI participants? Some evidence comes from a 2004 article by DeCharms and colleagues, who trained participants to imagine moving their right hands during 30 s imagery blocks, which were separated by 30 s non-task blocks. To ensure that the participants made no muscle movements during the imagery blocks, the researchers recorded electromyographic signals from the
At the end of each task block, the participant was given either accurate or inaccurate feedback about the level of activation in the left motor cortex relative to a target level. As the experiment progressed, subjects given accurate biofeedback both improved the SNR and increased the spatial extent of activation in the motor cortex. Conversely, subjects who were provided with inaccurate biofeedback could not increase the specificity of the BOLD signal over time. Additional, highly intriguing recent evidence comes from a 2007 paper by Bray and colleagues, who rewarded subjects if they were able to successfully activate a particular region of the somatosensory cortex. Over the course of the experiment, subjects learned by trial-and-error which region to activate and were progressively more specific in their activation. Together, these results suggest that subjects can adaptively use information supplied via biofeedback to up- (or down-) regulate specific brain regions, even without prior knowledge of the functions or identities of those regions.

At this point in time, research into fMRI biofeedback remains at a relatively early stage. Most studies have simply supported the concept rather than provided novel information about brain function. Yet, several potentially exciting applications have already arisen. Perhaps most promising are the therapeutic applications, which apply the power and neural specificity of fMRI to the types of problems (e.g., chronic pain) addressed by other forms of biofeedback. In a 2005 study, De Charms and colleagues evaluated whether biofeedback about activation in the anterior cingulate cortex (ACC), a region thought to be critical for the regulation of emotional and painful stimuli, could ameliorate the experience of pain. Their participants included both chronic pain sufferers and control groups of individuals without chronic pain. During 60 s task blocks, subjects were instructed to increase or decrease the BOLD signal in the ACC, whose activation was represented by a virtual fire whose intensity indicated the level of activation. Individuals who were provided accurate biofeedback successfully controlled activation in the ACC over the course of training. But those subjects given inaccurate feedback were unsuccessful in regulating ACC activation. Most strikingly, those chronic pain patients who exhibited the best control over the ACC also expressed the greatest reduction in their pain levels.

Biofeedback research using fMRI still faces several limitations. For analyses to be completed in real-time, the experiments must use relatively straightforward designs and analysis methods. Typical are blocked-design studies in which the subject attempts to increase or decrease activation in a given region during an extended (>30 s) block. For comparison, rapid event-related designs are difficult to analyze in real time, both because of lower SNR and because successive events may generate overlapping time courses of activation. Using a pre-selected ROI can speed biofeedback training. As shown by Laconte and colleagues in a 2007 paper, researchers can use complex analysis approaches like pattern classification, if a training pattern is established earlier in the experimental session. Then, during the testing runs, each time point can be quickly compared with that training pattern for classification. Also, despite the promise of fMRI biofeedback, it lacks the obvious advantages of electrophysiological approaches, which can be conducted using less expensive and more portable equipment. Nevertheless, we expect that new approaches to fMRI biofeedback will lead to exciting new results, some of which will have important clinical consequences.
taken a different tack by examining effects that vary between subjects, thus seeking \textit{individual differences} in brain function. At least three interrelated factors have contributed to the increased emphasis on variation among individuals. First, there has been a steady increase in the number of subjects per study. Whereas early studies often included ten or fewer subjects, many current studies contain samples of more than 20 individuals. Second, the standardization and validation of analysis approaches has made it easier to conduct analyses of individual differences. And third, fMRI researchers have become increasingly interested in new research areas in which individual variation is paramount, such as studies of personality traits, interpersonal attitudes, and decision preferences, among many others.

The most common method for analyzing individual differences involves correlations between fMRI data and behavioral data collected outside the scanner. For example, a researcher might give every subject a questionnaire that assesses some personality trait, like extraversion. The subjects' scores on that questionnaire are then entered into the across-subject analyses as a covariate (see Box 9.1, Figure 3, and Figure 10.21 for examples). In effect, the resulting significance tests identify voxels whose activation in one condition (or difference in activation between conditions) varies in proportion with subjects' trait scores. Other, more complex analysis methods are also possible. If a set of ROIs were implicated in a particular cognitive process (e.g., making risk-averse decisions), their amplitudes of activation might be combined to predict individual differences in behavior (e.g., an individual's overall level of risk aversion). Or, a data-reduction approach like PLS could be used to identify patterns of activation whose amplitude varies systematically between subjects. In principle, any statistical analysis of brain activation, from the simple approaches introduced previously to the advanced approaches outlined in this chapter, could serve to predict or measure individual differences. This sort of correlation approach has become a common tool for fMRI studies. Yet it is an important limitation. If a behavioral measure has high variability or systematic bias, or simply does not map onto the hypothesized ROI, then null or misleading results will be obtained.

So far, we have talked about prediction in a relatively abstract sense: researchers identify two sets of variables, one reflecting measures of behavior and the other reflecting measures of brain function, and then evaluate how well knowledge about one of those sets (brain data) lets us predict the other (behavioral data). However, researchers now use fMRI to make more explicit predictions. Consider the possibility of using some aspect of fMRI activation as a \textit{biomarker} for a psychiatric disorder, such as schizophrenia. If we already know that someone has schizophrenia, then scanning them might seem to provide little new information about their clinical status. But suppose that we scan subjects after they have been diagnosed with that disorder, but before they undergo a particular treatment regimen. Information about their brain function, such as activation in the prefrontal cortex associated with executive function, might predict their subsequent improvement following therapy. If so, then researchers and clinicians could work together to best match treatment options to individuals. Even more promising are studies of asymptomatic individuals who are nonetheless at high risk for a disease. For example, the risk of schizophrenia increases dramatically in young adults who have a close relative with that disorder. By conducting prospective studies in these high-risk individuals, researchers may become better able to predict who will develop the disease, allowing for targeted early interventions that could delay, ameliorate, or prevent development of the disease.

\textbf{individual differences} \textit{Measures of variation among individuals within a population.}

\textbf{biomarker} \textit{A phenotypic feature, whether physical, physiological, or behavioral, that provides a robust predictor of some experimentally or clinically important outcome.}
Predicting variation in behavior

Behavior changes dramatically over time. Even when completing the same simple task over and over again (e.g., “press a button when you see a shape appear”), the same person will respond quickly on some trials and slowly on others. When someone studies a random set of words in advance of a memory test, some of the words will be remembered while others will be forgotten. Most fMRI studies account for such variability in behavior by varying how events are coded via the timing and amplitude of regressors in the design matrix. For example, researchers studying the subsequent memory effect might present a set of words during the scanning session, then test each subject’s memory of those words in a behavioral testing session that could be hours, days, or even weeks later. When analyzing the fMRI data, the researchers create separate regressors for remembered and forgotten words. Similarly, a researcher interested in executive function might code each stimulus according to response time, resulting in regressors for various event durations across trials. (Introducing response time as a regressor in a design matrix can also be useful for minimizing unwanted effects of decision difficulty.) Coding events according to behavior has been particularly important in decision-making research, for which the key experimental conditions are often the choices made by the subject.

Often, researchers want to understand how different aspects of brain function (e.g., activation in different brain regions) predict variability in behavior. A particularly powerful approach has been to combine information about brain function and behavior in a logistic regression model. As introduced in Chapter 10, regression analyses use information from multiple independent variables to predict values of one or more dependent variables. Unlike standard regression analyses that use continuous dependent variables, logistic regression uses a categorical dependent variable. That is, it attempts to predict whether or not some outcome will occur, or to identify the state of some binary process. When using a logistic regression analysis, researchers typically identify a set of variables that might influence the subjects’ behavior. Some of these variables might reflect brain activation, such as activation in each of several regions or even the level of functional connectivity between them, while other variables will reflect aspects of behavior, such as choices made during the previous trial, or the overall decision bias of the subject. Including behavioral data in the logistic regression model minimizes the chances of spurious brain-behavior relationships; i.e., claiming that a brain region predicts a particular behavior, when instead that region’s activation was itself driven by other aspects of behavior.

A notable example of this approach was published in 2005 by Kuhnhe and Knutson. They were interested in how different brain regions might interact to shape choices between risky and safe options in a simple investment game (Figure 11.15A). The subjects made a series of choices between three shapes: two symbolized risky stocks that could either win or lose $10, while one represented the safe bond that was guaranteed to win $1. For each block of 20 trials, one of the stocks was twice as likely to win money as to lose money, whereas the other was twice as likely to lose money as to win money. Initially, the subject did not know which stock was the good one. But over trials, as subjects gained information about the reward histories of each stock, they could make educated guesses in an attempt to earn more money. The optimal strategy, therefore, was to first pick the safe bond and then later, once enough information accumulated, switch over to choosing the presumably good stock. Yet people often made mistakes. They sometimes made risk-seeking mistakes by
choosing one of the stocks even before there was good information about which stock was likely to make money. They also made risk-averse mistakes by choosing the bond even though they had enough information to gamble on the good stock. The authors attempted to predict these mistakes in decision-making using a logistic regression that contained both behavioral information and measures of fMRI activation in key brain regions. As shown in Figure 11.15B, increased activation in the insula predicted that the subject would make a risk-averse mistake following a risky choice, whereas increased activation in the nucleus accumbens (i.e., part of the ventral striatum) decreased the chance of a risk-averse mistake following a safe choice. Most critically, these brain regions were significant predictors of behavior even though information about behavior itself was included in the model.

Pattern classification using machine learning algorithms

Anyone who has examined raw fMRI data has been struck by how nearby and even adjacent voxels may exhibit wildly different time courses (Figure 11.16), which may lead to large differences in significance values following statistical
analyses. Standard fMRI preprocessing and analysis steps ignore inter-voxel differences. Indeed, researchers typically apply spatial smoothing and require clusters of activation, both of which suppress differences between nearby voxels. Yet, we know from more than a half-century of basic neuroscience research that the cortex exhibits substantial local organization, often at a spatial scale of several millimeters or smaller (e.g., ocular dominance columns in the visual cortex). How can researchers use information from individual voxels to draw inferences about the function of a region?

Approaches to solving this problem have largely been based on pattern classification algorithms, building on research from the subfield of machine learning within computer science. Machine learning uses known items from a data set to create rules that can efficiently categorize new items. Suppose that you wanted to create a computer program that could categorize an animal as either a dog or a cat, based on its visual appearance. You show the program some representative dogs (e.g., greyhound, terrier, retriever) and some representative cats (e.g., siamese, persian, calico). From those examples, the program extracts some general rules for categorization, such as “large, long-legged animals with extended snouts are dogs.” These rules should be broad enough to generalize to new examples of the category (e.g., “beagles”), but specific enough to exclude items that are not in that category. Moreover, the algorithm used to create the rules should be computationally tractable. Generally, no classification rule will be both simple and able to classify new items perfectly. For example, some types of dog are small, short-legged, and flat-snouted (e.g.,

\[ \begin{align*}
 & (A) \\
 & (B)
\end{align*} \]

Figure 11.16 Distribution of statistical values across space. While activation maps of the brain (A) typically show in color only those voxels whose statistical values are greater than some threshold (here set to \( t > 3.6 \)), the actual pattern of underlying statistics can be quite complex. Note in (B) that the area within the white box in (A) contains some voxels with very high statistical values, along with others that are only slightly above the threshold. In fact, some voxels that are designated as inactive, and thus not shown on the color map, have significance values quite near the threshold. Other nearby voxels even exhibit negative significance values.

pattern classification An attempt to separate individual examplars into different categories by constructing a set of decision rules based on some combination of their features.

machine learning A subdiscipline within computer science that develops algorithmic rules for relating input data to desirable outputs.
multi-voxel pattern classification An approach for pattern classification in fMRI research that uses as its input data the relative changes in activation across a set of voxels. Therefore, the main challenge of pattern classification is to identify rules that can be generalized to new examples yet are simple and efficient.

For fMRI studies, researchers generally use multi-voxel pattern classification: they predict event categories from various patterns of activation across voxels, rather than from the overall increase or decrease in activation of a large brain region. Although there are several types of pattern classification algorithms used on fMRI data, all involve three main steps (Figure 11.17). First, a

![Feature selection](image)

**Figure 11.17** A conceptual overview of multi-voxel pattern classification of fMRI data. In this example, the researchers want to identify voxels whose activation predicts whether the subject is looking at photographs of animals or plants. At the first stage, feature selection, the researchers identify a subset of voxels for subsequent analyses. A typical feature set consists of the activation intensity for each voxel on each trial. The feature set splits into a training set, from which the pattern classifier will be derived, and a testing set that provides a novel test of the generalization of the classifier. Shown here is a simplified example of pattern classification using two features (i.e., two voxels) and two trial categories (A and B). The activation values of those two voxels on each trial are shown as a two-dimensional plot. Note that fMRI pattern classification involves many more dimensions, and thus a much higher-dimensional space. In the common technique of support vector machines, the pattern classification algorithm attempts to identify the surface that maximally distinguishes the two categories. Here, a linear classifier optimally separates the two stimulus categories. Once a classifier has been identified, it is tested on the novel training set, to ensure that the classification rule can be generalized to untested data.
subset of the fMRI data is extracted in a process called feature selection. In most cases, this involves identifying a subset of $V$ voxels (e.g., all voxels within a predetermined ROI, or voxels within a particular searchlight) and identifying the BOLD amplitude in each voxel at each of $T$ time points. Usually, those time points reflect the stimulus events or behaviors to be sorted, taking into account the delay in the hemodynamic response, for example by selecting a time point 4-6 s following the presentation of a stimulus. In some cases, researchers have used pattern classification to study processes that occur in advance of the subject making a decision, in which case the time points would be earlier than the corresponding behavioral events. This first step reduces the fMRI data to a set of vectors $(N_V)$, each corresponding to the pattern of activation associated with one example from a category. Researchers often subtract the mean signal for the whole region from each vector, so that only relative changes in activation across voxels contribute to the classification algorithm.

**Thought Question**

Why might local patterns of fMRI activation across voxels within a region be more sensitive for detecting cognitive processes than the global change in activation from that entire region?

Second, the researchers partition their data into a training set and a testing set. For example, if the experiment consisted of 50 presentations of male faces and 50 presentations of female faces, the training set might contain 25 faces from each category and the testing set might contain the remaining 25 faces from each category. The vectors from the training set are then entered into a pattern classification algorithm, most commonly (for fMRI studies) using support vector machines (SVMs). Considered generally, an SVM takes the vectors within each of two categories A and B (each vector representing a set of points in a $V$-dimensional space), and attempts to find the surface that maximally distinguishes the two categories within that space. A linear SVM approach uses a hyperplane to distinguish the categories; that is, it finds the combination of voxel weights that best predicts whether a time point belongs to Category A or Category B. A non-linear SVM looks for the complex curved surface that best distinguishes the categories; in effect, it can use the joint movement of pairs or groups of voxels (e.g., increased activation in one voxel at the same time as decreased activation in another) to improve its predictions. The set of voxel weights, whether linear or non-linear, is known as a pattern classifier, and the points that lie along the pattern classifier boundary are called support vectors, hence the name for the technique.

Third, the researchers evaluate whether their pattern classifier can be generalized to new data. Several approaches can be used. One involves splitting the data, as described in the above example, with the success of the classifier defined primarily by its performance on a completely novel testing set. Another approach involves the iterated use of several training sets. For example, researchers could separate their data into five parts, use the first four in a training set for their pattern classifier, and then test the classifier using the fifth part. Then, they could use a different four-fifths of their data to generate a second classifier, and test the new classifier on the remaining one-fifth. By repeating this approach five times, and thus using each fifth of the data once as a testing set, the researchers can calculate the average predictive performance of classifiers using the input voxels. This is known as cross-validation. Further, feature selection An initial step in pattern classification that involves the determination of which input variables should be included in the classification algorithm.

searchlight An approach to feature selection in the pattern classification of fMRI data. As its name implies, a searchlight reflects a geometrically defined ROI (e.g., a sphere of 5-voxel radius) that can be moved throughout the brain.

training set In pattern classification analysis, that part of the data set used to develop the classification algorithm.

testing set In pattern classification analysis, a novel part of the data set used to evaluate the robustness of the classification algorithm.

support vector machine (SVM) A class of algorithm used in pattern classification that attempts to identify the combination of features in the original data set that can most effectively differentiate between two categories.

cross-validation In pattern classification analysis, an approach to evaluating the effectiveness of classification using a given feature set. It involves the iterated generation and testing of classifiers based on different parts of the same training set.
researchers can combine these techniques by first using cross-validation to identify sets of voxels whose activation best distinguishes categories, followed by testing an optimized classifier on an as-yet-unexamined testing set.

Because pattern classification relies on relative changes in intensity in individual voxels, rather than changes in the overall activation of a large region, some aspects of data analysis in pattern classification differ from standard approaches. Notably, researchers do not typically apply spatial smoothing during preprocessing. As discussed in Chapter 8, smoothing spreads information from one voxel across its neighboring voxels, which can improve some aspects of fMRI analyses, such as reducing the number of independent comparisons and minimizing voxel-dependent noise. However, introducing spatial blurring makes individual voxels less predictive. Pattern classification can be improved at the initial, feature-selection stage by excluding some voxels beforehand, so that they do not contribute to the classifier. Features can be based on anatomical criteria (e.g., selecting specific ROIs), functional criteria (e.g., using only activated voxels from another statistical test), or statistical criteria (e.g., removing high variability voxels). Pattern classification studies typically involve analyses at the single-subject level, largely because the fine spatial patterns that distinguish different processes may differ across subjects. The brains of two subjects may share gross functional similarities, such as increased activation in the lateral occipital cortex in response to visual stimuli, and thus may lead to similar outcomes using standard analysis methods. However, the responses of individual voxels to different stimulus categories might differ dramatically. So, researchers usually evaluate whether a given spatial location (e.g., voxels in the primary visual cortex) can be used for pattern classification in each of their subjects. It is less common to derive a spatial pattern in one or more subjects and extend that pattern to novel subjects. See the 2005 paper by Davatzikos and colleagues in the chapter references for a rare example.

**Capabilities and challenges of fMRI pattern classification**

Pattern classification analyses provide several advantages over the regression-based approaches described in Chapter 10. By evaluating changes in the activation of individual voxels, pattern classification incorporates information that is typically discarded or smoothed over. This can greatly improve the sensitivity for detecting small but meaningful changes in fMRI activation. Some task-related effects are detectible using pattern classification, but completely undetectable using standard techniques. Suppose that a process causes opposing changes in two neighboring voxels, such that one increases in activation when the other decreases in activation. A pattern classification approach could detect that those joint movements were task-related, whereas standard analysis techniques would see only minimal changes in overall activation. However, the value of pattern classification extends much farther than simply increasing the sensitivity of analyses. The most powerful applications of this approach address questions about processes that are distributed throughout brain regions. Here we highlight some exciting applications of pattern classification, many of which represent the very cutting edge of fMRI research.

Unquestionably, fMRI pattern classification has had the greatest impact on studies of the neural basis of perception, specifically visual perception. Extensive prior research, mostly using electrophysiological recordings from other species, has demonstrated that our visual perceptions result from the combined output of a hierarchy of neurons, each only processing a small portion of the visual world. While many of the details of visual system function-
ing remain to be discovered, there exists strong evidence for both large- and small-scale organization. Different regions of the cortex process distinct visual features: neurons in the primary visual cortex respond to simple lines and edges, whereas neurons farther along the visual processing pathway respond to objects from specific categories. At least within the early visual regions, individual neurons are arranged spatially according to their input, specifically according to the arrangement of the parts of the retina (and thus the visual field) to which they are most sensitive.

Given how much is already known about the visual system, primarily from electrophysiological research, what can fMRI pattern classification contribute? One potential answer comes from a consideration of the advantages of fMRI. Because fMRI can be conducted in human subjects who are performing any of a wide range of experimental tasks, researchers can use it to study how different visual regions contribute to complex processes: increasing attention, improving the encoding of stimuli into memory, or even imagining a particular object. For example, a 2005 article by Kamitani and Tong used pattern classification to identify voxels whose activation was sensitive to the orientation of a visual stimulus. They found that the pattern of activation in early visual cortical regions reliably distinguished line gratings in two different orientations (e.g., 45° versus 135°), as would be expected based on prior electrophysiological studies. They then showed subjects a complex grid pattern containing line gratings of both orientations, and asked the subjects to monitor one of the two gratings for infrequent changes in its line width. They found that selective attention to one orientation systematically modulated the voxel pattern associated with the passive viewing of that orientation. Conversely, they could reliably predict to which orientation the subject was attending by analyzing activation in specific, local patterns within the visual cortex. This study, along with many others, demonstrates how pattern classification can allow researchers both to identify category-specific patterns of fMRI activation and to evaluate how those patterns are modulated by different cognitive processes.

Another important application for pattern classification has been the perception of object categories. There has been substantial debate about whether higher visual regions that process complex stimulus properties, like object identity and category, contain some internal topographic organization. Cox and Savoy reported an early example of object categorization in 2003. Their subjects viewed examples of ten different object categories, ranging from the commonplace (e.g., “chairs”) to the whimsical (e.g., “garden gnomes”). Different examples of each category were presented within 20 s blocks, and the mean activation during that block was calculated for all voxels within the visual cortex ROIs. Even in their most conservative analysis, one that restricted the feature set to voxels within object-selective regions and that used independent sets of blocks for their training and testing sets, their pattern classifier could correctly identify what object category was being shown about 30–50% of the time. These values are much greater than the chance level of 10%.

It is important to recognize that this study by Cox and Savoy and similar studies show that voxels contain independent information about perception; they do not suggest what form that information might take. Kay and colleagues found one way of bridging this gap in an elegant study published in 2008. First they showed each of their two subjects nearly two thousand different photographs of natural scenes while measuring activation in the visual cortex. From the resulting activation maps, they estimated each voxel’s sensitivity to different spatial locations and line orientations; in effect, this was an attempt to infer the properties of the neurons within that voxel. They next
Figure 11.18 Using fMRI pattern classification to reveal brain regions that predict later conscious decisions. (A) Participants viewed a rapidly changing series of letters, each shown for 500 ms. Whenever the subject desired, he or she could press either the left or right button on a keypad. After that decision, a feedback screen popped up with letters and symbols, and the subject simply indicated which letter had been visible when the decision was made (mean onset time for awareness indicated by the vertical line). (B) The researchers examined time points prior to the reported onset of the decision using pattern classification. In a striking result, activation in the anterior part of the prefrontal cortex (i.e., lateral frontopolar cortex) and in the medial parietal cortex (i.e., posterior cingulate cortex) predicted the subsequent decision as much as 7 s in advance. Red circles indicate time points with significant predictive power. (After Soon et al., 2008.)

recorded fMRI activation while the same subjects viewed 120 new images, each shown 15 times. By comparing the activation pattern during each trial with the predicted activation pattern for each of the 120 images, the authors could predict what image the subject was viewing, based only on the brain activation. Remarkably, their predictions were accurate for 92% and 72% of the images in their two subjects, demonstrating that fMRI could be used to predict the contents of visual experience, based on a priori models of the functional properties of individual voxels.

While studies of visual perception have led the growth of fMRI pattern classification, recent experiments in other topic areas have provided striking examples of its power. One application of fMRI that has attracted considerable and wide-ranging attention is its potential for anticipating our thoughts and actions before they occur. While this may seem impossible, or at best the speculations of bad science fiction, there has been a long history of research into neural signals that precede conscious awareness, following largely from the work of the physiologist Benjamin Libet. In a 2008 article, Soon and colleagues investigated whether fMRI pattern classification could be used to detect subjects' intentions to engage in an action, even before they were consciously aware of those intentions. Subjects viewed a continually changing series of letters, each presented for 500 ms (Figure 11.18A). At any point in time, the subject could decide to press either the left or right button, whereupon a screen popped up
and the subject indicated which letter had been visible when they made their decision. As might be expected, subjects typically reported that their conscious decision occurred less than a second before they actually pressed the button. However, the authors found that activation in an anterior region of the prefrontal cortex (Figure 11.18B) could predict which button would be pressed as much as 7 s before the conscious decision! Given the delay of the hemodynamic response, this region carries information about the nature of an upcoming decision as much as 10 s in advance. While these results fall far short of reading someone's mind (see Chapter 14 for associated ethical issues), they lead to some intriguing possibilities for future research; for example, an in-scanner video game that uses real-time pattern classification analyses to counter subjects’ moves before they occur.

Although powerful, pattern classification approaches will not replace traditional fMRI analyses, at least in the near term. Because of the size of fMRI datasets, pattern classification is a computationally intensive process. Identifying an optimal classifier can require many hours of computer time, and testing that classifier for significance may require many repetitions of the analysis using a permutation approach. Furthermore, even if local patterns are identified, the processes they support may remain obscure, especially if no clear spatial topography exists. Local information may itself be present in several regions. For example, in a 2007 study Hampton and O'Doherty identified a set of three regions that together predicted subjects’ decisions in a reward-learning task. This sort of result cannot by itself distinguish the specific contributions of individual regions; however, it provides a starting point for research using other analysis approaches (e.g., functional connectivity methods). Finally, pattern classification remains an inherently within-subject technique: the specific patterns identified within one subject are unlikely to generalize to other subjects, given the differences in functional organization in different brains. Despite these limitations, we expect that pattern classification will become a popular and even mainstream approach in future fMRI research.

Summary

Until recently, most fMRI studies have used hypothesis-driven approaches for data analysis: the researcher creates a hypothesis about task-related changes in the BOLD signal and then tests how well each voxel matches that hypothesis. However, in recent years, researchers have implemented a variety of data-driven approaches that are used to identify intrinsic variation within an fMRI data set, and then relate that variation to function. Some techniques such as PCA or ICA attempt to separate spatial or temporal patterns within the fMRI timecourse; these can be useful for exploring complex data sets or for removing unwanted variability during preprocessing. Functional connectivity algorithms describe the interrelationships between spatially distant brain regions, which may range from simple correlations between regions to causal influences of one region upon another. Data about the functional connections between regions can be combined with the results of standard fMRI analyses, or with structural information obtained using DTI, to obtain more complete descriptions of functional networks. Finally, researchers now frequently use data about brain function to predict subjects’ behavior or traits. A particularly important approach is fMRI pattern classification, which uses information about the relative changes in intensity across a set of voxels to predict a perceptual or cognitive state. Researchers have used pattern classification to investigate topics as diverse as how objects are
processed in the visual cortex and how decisions arise within the prefrontal cortex. Many data-driven approaches share common limitations—they are often computationally intensive and require complex approaches to determine statistical significance—yet they constitute some of the most exciting new directions for fMRI research.

**Suggested Readings**


*Indicates a reference that is a suggested reading in the field and is also cited in this chapter.*

**Chapter References**


