Take a dream. You are walking through the woods. In a clearing you come upon a marble statue of a human figure. There is an inscription on the pedestal: “Behold one possessed of mind without idea, form without sensation.” You continue walking. Evening comes on, and the woods are full of sounds and shadows. Suddenly, to your right, you see a hulking form. You leap back, ready to run—is this a bear? No, there is no danger; the “bear” is only a bush. The night grows darker. The path is rising now, and at the top of a dimly outlined hill, you can
see the lights of a château. By the time you reach it and take shelter, all is dark outside, and from your curtained room inside you have no idea what lies outside the walls. Morning comes, the curtains are thrown back, you see. . . .

These imaginary experiences, and their resolution, illustrate the essential problems of perception and its relation to cognition. This chapter is a discussion of what perception is and how it works. We specifically address six questions:

1. What is perception and why is it a difficult ability to understand?
2. What general principles help us to understand perception?
3. How do we put together parts to recognize objects and events?
4. How do we recognize objects and events?
5. How does our knowledge affect our perception?
6. Finally, how do our brains put together the many and varied cues we use to perceive?

1. WHAT IT MEANS TO PERCEIVE

The “sculptor” of the mysterious statue was the French philosopher Etienne Bonnot de Condillac (1715–1780), who created it in his Treatise on Sensations (1754a). The statue, he imagined, had in working order what we would call the “mental hardware” and “software” of a normal human being, but no senses. Condillac believed that such a being would have no mental life, that no ideas were possible in the absence of sensation.

Pursuing his thought experiment, he imagined opening up the nose of the statue so that it could now smell. “If we offer the statue a rose,” wrote Condillac, “it will be, in its relation to us, a statue which smells a rose; but, in relationship to itself, it will be merely the scent of the flower.” That is, Condillac thought that if the statue had only a single sensation, then that sensation would be the whole content of its mind.

Even if we adopt a position less absolute than Condillac’s, we can agree that the mental life of an organism without senses would be unimaginably different from the mental life we experience. Sensation and perception provide the raw material for cognition, certainly, but this assessment underplays their role. Our perceptions are not a simple registration of sensory stimuli. Sophisticated cognitive processes begin to work on this material almost immediately, producing the brain’s interpretation of the external world as incoming stimuli are analyzed, and existing knowledge guides these dynamic processes.

The second and third parts of your dream are illustrations that make clear why perception is much more than the straightforward registration of sensory stimuli. In your second dream experience, the menacing shape in the forest seems familiar but only faintly so. This is because the images appear outside their original context of Shakespeare’s A Midsummer Night’s Dream: “In the night,” says Theseus, Duke of Athens, “imagining some fear, how easy is a bush supposed a bear.” Shakespeare understood that sensory stimuli typically are ambiguous, open to multiple interpretations; this is the first problem of perception.
1. What It Means to Perceive

What do you see in Figure 2–1? Probably a cube. Does it seem to be floating in front of a black background with white dots on it? Or, rather, to be lying behind a black sheet with holes punched into it? As to the cube itself, is the surface that seems closest to you angled up and to the left or angled down and to the right? Why see a cube at all? The image, of course, is actually flat on the page. You might swear that you can see the lines of the cube crossing the black region, but they are not present in the image. There are only eight carefully positioned white dots, each containing a carefully positioned set of three line segments. Nonetheless we see the cube, even though the image doesn’t have all the properties of a real cube, even one drawn on a two-dimensional surface, but only a sparse subset of those properties. We fill in the missing pieces and perceive more than just what you sense of the image properties.

What do you see in Figure 2–1? Probably a cube. Does it seem to be floating in front of a black background with white dots on it? Or, rather, to be lying behind a black sheet with holes punched into it? As to the cube itself, is the surface that seems closest to you angled up and to the left or angled down and to the right? Why see a cube at all? The image, of course, is actually flat on the page. You might swear that you can see the lines of the cube crossing the black region, but they are not present in the image. There are only eight carefully positioned white dots, each containing a carefully positioned set of three line segments. Nonetheless we see the cube, even though the image doesn’t have all the properties of a real cube, even one drawn on a two-dimensional surface, but only a sparse subset of those properties. We fill in the missing pieces and perceive more than is actually there. So the first problem is that sensory input does not contain enough information to explain our perception. When you looked at and interpreted Figure 2–1, for example, you had to infer an object from mere hints.

In the last part of your dream, you get out of bed, walk to the window, and throw open the heavy curtains. In an instant, you are exposed to a panorama of mountains, fields, houses, towns. What do you perceive? Condillac thought you would see only a patchwork of colored regions, an experience full of sensations but without the organization that makes perception (1754b). In fact, we know that you could understand the gist of the scene after it had been exposed to your visual sense for only a small fraction of a second: studies have shown that you can look at pictures on a computer screen at a rate of eight per second, monitor that stream, and find, for example, the picnic scene in the series (Potter & Levy, 1969) or even the scene in a series that does not contain an animal (Intraub, 1980). Still, Condillac was right in noting a problem: this second problem is that the world presents us with too much sensory input to include into our coherent perceptions at any single given moment.

Figure 2–2 is a scene of a beautiful summer afternoon in a park with lots going on. The image is not difficult, but it has many elements: although you can see and understand it, you cannot fully process it in one rapid step. Quick—is there a dog in it? Because it is impossible to process everything in the image at one go, you may not
know whether there is a dog until you have searched for it. Moving your eyes over the image, you pause at different parts of it and fixate your gaze, bringing the center of the retina, the region with sharpest vision, over the area you wish to examine. There’s a dog! Even though we can see over a large region at one time, we can see relatively fine detail only in a small region—at the point of fixation. Searching is one way to deal with the excess of input.

Much information, for example information about the precise differences in the intensity of light at each point in space, is thrown away at the very start of the journey from sensation to understanding. One of the dogs on Grand Jatte, however, is not small. You certainly can see it without moving your eyes from the center of the image. But it is very likely that you could not determine whether there is a dog in the painting until you selected that portion of it for further consideration. Our ability to engage in selective attention allows us to choose part of the current sensory input for further processing at the expense of other aspects of that input; we will consider the nature of attention in detail in Chapter 3.

The two problems of perception in relation to the sensory world, then, are “not enough” and “too much.” In both cases, cognitive mechanisms are necessary to provide the means to interpret and understand the material our senses bring to us.
Comprehension Check:

1. Why is perception important for cognition?
2. What are the two main problems that make perception difficult?

2. HOW IT WORKS: THE CASE OF VISUAL PERCEPTION

The goal of perception is to take in information about the world and make sense of it. Condillac’s statue tells us that our mental life depends on meeting this goal. Theseus’s bear reminds us that the information available to us may be ambiguous and therefore insufficient for the determinative interpretation that only cognitive processes and background knowledge can make. The view from Condillac’s château reveals that there is too much information for us to process and we need to select.

An analogous action of selection needs to be made right now: all our senses are vitally important and no sense acts in isolation from the others. For example, consider the interplay of vision, hearing, taste, smell, and touch in your most recent dining experience. Sadly, all that richness cannot be adequately captured in a single chapter, so, sacrificing breadth for a modicum of depth, we will select vision to discuss and we will further select a restricted set of examples within the visual domain.

Vision, like hearing, is a distance sense, evolved to sense objects without direct contact. It can tell us what is out there and where it is. If we think of humans and other creatures as organisms that must interact with the world, we see that our senses also provide something further: a nudge toward action. What is out there, where is it, what can I do about it? (Oh, look, a lovely low-hanging apple—I’ll pick it!) Visual perception takes in information about the properties and locations of objects so that we can make sense of and interact with our surroundings.

2.1. The Structure of the Visual System

The main visual pathways in the brain can be thought of as an intricate wiring pattern that links a hierarchy of brain areas (Figure 2–3). Starting at the bottom, the pattern of light intensity, edges, and other features in the visual scene forms an image on the retina, the layer of cells that respond to light, called photoreceptors, and nerve cells at the back of each eye. There light is converted into electrochemical signals, which are transmitted to the brain via the optic nerves (one from each eye); each optic nerve is a bundle of the long axon fibers of the ganglion cells in the retina. The axons make contact with the neurons of the lateral geniculate nucleus (LGN) in the thalamus, a structure lying under the surface of the brain. From there, axons of LGN neurons send signals up to the primary visual cortex (which is also called V1 for “visual area 1,” or “striate cortex” because when stained it has the appearance of a stripe across it that can be seen with a microscope). Output from the striate cortex feeds a host of visual areas (V2, V3, V4, and others) as well as areas that are not exclusively visual in function.

Beyond the primary visual cortex, two main pathways can be identified. A dorsal pathway reaches up into the parietal lobes and is important in processing information about where items are located and how they might be acted on, guiding
A “wiring diagram” of the visual system, showing connections among brain areas. Note that there are two types of retinal ganglion cells (magnocellular, abbreviated m, and parvocellular, abbreviated p); these cells project axons to different portions of areas V1 and V2.)

movements such as grasping. A ventral pathway reaches down into the temporal lobes; this pathway processes information that leads to the recognition and identification of objects. This two-pathways story is valid, but as Figure 2–3 shows, it is a great simplification of an extremely complicated network.

2.2. Top-Down and Bottom-Up Processing

The daunting complexity of the visual system is functional as well as structural, as is shown in Figure 2–3. The pathways and their many ramifications are not one-way streets. Most visual areas that send output to another area also receive input from that area; that is, they have reciprocal connections—for example, LGN provides input to V1 and V1 provides other input to LGN. This dynamic arrangement reflects an important principle of visual perception: visual perception—in fact, all perception—is the product of bottom-up and top-down processes. Bottom-up processes are driven by sensory information from the physical world. Top-down processes actively seek and extract sensory information and are driven by our knowledge, beliefs, expectations, and goals. Almost every act of perception involves both bottom-up and top-down processing.

One way to experience the distinction consciously is to slow part of the top-down contribution. Look at Figure 2–4. There is certainly something there to be seen: bottom-up processes show you lines and define regions. But if you play with the image mentally and consider what the regions might signify, you can feel a top-down contribution at work. The image could be . . . a bear climbing up the other side of a tree! Whether or not you came up with this solution yourself, your appreciation of it depends on top-down knowledge: your knowledge of what a tree and a bear’s paws look like, your knowledge of how bears climb trees. This kind of knowledge not only organizes what you see, but also can even modulate the processes that created the representations of the lines and regions.

Another example that points to the distinction between bottom-up and top-down processing and the relationship between them can be seen in visual search tasks. If you’re told to find the target in Figure 2–5a, you have no problem. Bottom-
up processing quickly identifies the white star as the stand-out. But bottom-up processing isn’t enough to guide you to the target in Figure 2–5b. There you see a number of items that differ in various ways—in shape, color, orientation. To find the target you need information—“the target is the horizontal black bar”—and thus top-down processing. Now you have the means to search for the target.

Both of these examples demonstrate that perceptions (this is a bear, this is a target) are interpretations of what we see, representations produced by the interaction of bottom-up and top-down processing.

2.3. Learning to See

Our interpretations of the world around us are determined by the interaction of two things: (1) the biological structure of our brains and (2) experience, which modifies that structure. The visual system in newborn infants is nearly fully developed at birth, and most of the major structural changes are complete in the first year of life (Huttenlocher, 1993, 2002). Babies open their eyes almost immediately after birth and soon they begin to look around, moving their eyes to investigate their surroundings and to fixate on objects of interest. Typically fixations last about half a second, so babies have on the order of 10 million glimpses of the world in their first year of life. That’s an enormous amount of information. A baby may see a parent’s face, the surrounding crib, a nursing bottle many thousand times, often from different viewpoints, at different times of day, and in different contexts. As the lingering memory of each occurrence combines with each new instance, the cascade of information somehow accumulates to form lasting mental representations of the people, places, and things in the environment. These representations form the basis for the subsequent recognition of objects.

Research on the development of visual perception in newborn animals has shown that the characteristics of the infant’s environment at particular times strongly influence some of the capabilities of the adult. The early stages of life include biologically determined critical periods, periods during which the animal must develop particular responses. If exposure to the natural environment is limited during the critical period...
for a particular response, the animal will fail to develop that ability properly, even with normal exposure during adulthood. For example, a kitten reared with a patch over one eye for 6 months may grow into a cat with two normal eyes, but with impairments in the perception of depth that depends on integrating information from both eyes (Wiesel & Hubel, 1963). In such a cat, a greater area of visual cortex is devoted to analyzing input from the unpatched eye than the patched eye. Interestingly, a kitten with patches over both eyes during the same period will not have deficits in the perception of depth as an adult and will have more balanced cortical organization (Wiesel & Hubel, 1965). Different aspects of sensory processing have different critical periods.

In addition, different sources and different modalities of sensory input seem to compete for representation in cortex (Le Vay et al., 1980). If one channel, such as input from one eye, is more active than another, cortical resources are redeployed in that direction and, once assigned in infancy, such resources are not easily modified in adulthood. Competition for neural representation has been demonstrated throughout the brain and for many different abilities: there is competition between auditory and visual perception (Cynader, 1979; Gyllensten et al., 1966); competition to register sensation from different fingers (Jenkins et al., 1990; Merzenich & Kaas, 1982); and competition between different languages in bilingual people (Neville & Bavelier, 1998).

Because it is known that experience alters the course of visual development, programs have been developed for stimulating the fetus with lights and sounds not normally present in the womb in an attempt to speed or enhance development. Normal prenatal stimulation, such as the sound of the mother's voice, can lead to better perception in infants. However, our knowledge in this area is far from complete, and it is possible that abnormal stimulation can lead to impoverished rather than superior development. Indeed, some studies have shown that some prenatal stimulation can impair normal perceptual development later in life (Lickliter, 2000). Although we know that our environment shapes the brain structures that support our capacity for normal cognition, we do not yet know how to control that process.

Comprehension Check:

1. In what ways is the brain structured like a hierarchy? In what ways is it not?
2. What is the difference between bottom-up and top-down processing?
3. How does visual experience influence what we see?

3. BUILDING FROM THE BOTTOM UP: FROM FEATURES TO OBJECTS

Condillac’s statue had all the machinery for cognition but no sensory input, so its brain never used its tremendous capability for representation and processing of the physical world. The brain’s ingenious techniques for combining perceived features, so that we can understand the complexity surrounding us by resolving it into objects familiar and unfamiliar, lay idle and useless. If the statue’s eyes were open to the
world, they would let in a flood of information through neural pathways, and a remarkable amount of sophisticated analysis would be performed to detect important aspects of the environment. And we, who have access to the world through our senses, have very busy brains. In the modality of vision, starting from the bottom up, let’s discuss what happens.

3.1. Processing Features, the Building Blocks of Perception

Visual features include spots and edges, colors and shapes, movements and textures. These are all attributes that are not in themselves objects, but in combination they can define the objects we see. They are the building blocks of perception.

In the eyes, photoreceptor cells in the retina convert light energy (photons) reflected from the various objects in the physical world into an electrochemical signal that can travel through the nervous system. The more light, the more signal. Varying intensities of light fall on the array of photoreceptors, so the input at any given moment might be conceived of as a set of numbers, each number equivalent to an intensity of light, one number per photoreceptor, such as the array of numbers shown in Figure 2–6. The task of the bottom-up processes in the visual system is to extract from the physical equivalent of this mass of numbers the features that will permit the subsequent processes to figure out what is out there in the world.

3.1.1. Spots and Edges

We can see progress toward this goal of feature extraction if we look at a ganglion cell, one of those neurons in the retina whose axon fibers form the optic nerve. Each ganglion cell is connected, through a series of other cells, to a collection of photoreceptors that are neighbors to each other. This means that the ganglion cell will respond only to light that lands on those receptors and, thus, to light in one specific region in the visual field, the portion of the world that is visible at the present moment. Look at Figure 2–7. There is a spot of light out in the world, the stimulus. The receptors, in this example, respond with 100 units of signal where the light is bright and just 10 where the light is dimmer. Our ganglion cell gets input from the receptors that lie in its receptive field, the region shown in color at the bottom of the figure. In vision, the receptive field of a cell is the area of the visual field in which a stimulus will affect the activity of the cell. If we were talking about a cell that responds to touch, the receptive field would be a patch of skin.

Most important, the connections from photoreceptors to ganglion cell are not all the same. Light in some portions of the receptive field excites the cell, that is, makes it more active. Light elsewhere inhibits the cell, making it less active. Specifically, the wiring is arranged so that input in the central zone (white) excites the ganglion cell, whereas input in the surrounding region (gray) inhibits the cell. Since we have arranged for the spot of light to fall on that excitatory central portion, this ganglion cell will be quite strongly excited. If the center region was stimulated by a gray area, the cell would not be very excited. And if the whole field were 100 units bright, the cell would not be very excited either, because the strong excitation of the
center would be offset by strong inhibition from the surround. So this cell is maximally excited when a bright spot of the size of that central region falls on the central region.

Something interesting happens if a collection of photoreceptors organized into these center–surround receptive fields receives input across an edge in the image in the visual scene, such as the border between light and dark rectangles in Figure 2–8. Assume that maximum stimulation of the center of each receptive field produces 10 units of excitation and stimulation of the surround produces 5 units of inhibition. A spot falling just on the center would produce 10 units of response. A
bright field, filling the whole receptive field (as is happening on the left in Figure 2–8) produces only 5 units; this is the light rectangle. The area on the right is dark; say that absolutely no light is falling here, and the value is 0. Now look what happens at the edge between the light and dark areas. Here one receptive field is mostly on the light side, another mostly on the dark side. When the center is on the bright side and a bit of the surround is in darkness, the response goes up, perhaps to 7 units. When the center is on the dark side and only a bit of the surround is on the bright side, the response could go down to a “darker-than-dark” level, quantified here as $-2$. In this way, the structure and arrangement of photoreceptors can serve to enhance the contrast at edges.
Building from the Bottom Up: From Features to Objects

3. Building from the Bottom Up: From Features to Objects

Figure 2–8 analyzes the effect; Figure 2–9 demonstrates it. The gray areas (the bars, or rectangles) are in fact each uniform, but each lighter bar appears a bit lighter on the right side, where it abuts a darker bar, and each darker bar appears a bit darker on the corresponding left side. This phenomenon was described by the Austrian physicist Ernst Mach in the mid-nineteenth century (Mach, 1865; Ratliff, 1965), and bars like those in Figure 2–9 are called Mach bands. This perceptual phenomenon is predicted by responses of ganglion cell neurons. The center–surround organization of ganglion cells is well designed to pick out edges in the visual environment.

Figure 2–8 analyzes the effect; Figure 2–9 demonstrates it. The gray areas (the bars, or rectangles) are in fact each uniform, but each lighter bar appears a bit lighter on the right side, where it abuts a darker bar, and each darker bar appears a bit darker on the corresponding left side. This phenomenon was described by the Austrian physicist Ernst Mach in the mid-nineteenth century (Mach, 1865; Ratliff, 1965), and bars like those in Figure 2–9 are called Mach bands. This perceptual phenomenon is predicted by responses of ganglion cell neurons. The center–surround organization of ganglion cells is well designed to pick out edges in the visual environment.

---

**FIGURE 2–8** How we detect edges

Ganglion cell receptive fields (large, outer circles) with +10 excitatory regions and −5 inhibitory regions are shown over a visual display of a light rectangle next to a dark rectangle. Interesting responses happen at the edge between the two rectangles. The graph at the bottom of the figure plots the amount of response over the different regions of the display.

**FIGURE 2–9** A demonstration of Mach bands

Six uniform rectangles are shown abutting one another, ordered from lightest to darkest. Even though the level of gray in each rectangle is uniform, it looks as if each one is a bit lighter on its right edge than its left edge and darker on its left edge. These edge effects come from the neighboring rectangle, and are predicted from the responses of ganglion cell neurons shown in Figure 2–8.
3.1.2. Throwing Away Information

The visual system seems to be designed to collect information about features, such as spots and edges, and not to spend unnecessary energy on nearly uniform areas where nothing much is happening. This bias is demonstrated by the Craik–O’Brien–Cornsweet illusion (Cornsweet, 1970; Craik, 1940; O’Brien, 1958), shown in Figure 2–10. Part (a) of the figure appears to be a lighter and a darker rectangle, each of them shading from darker to lighter. But if we cover the edge at the center between the two rectangles, as in part (b), we see that most of the areas of the two rectangles is the same gray level. The visual system found the light–dark edge at the center and, in effect, made the not unreasonable assumption that the image was lighter on the light side of the edge than on the darker side. Because edge information is important for defining the shape of objects and providing cues for where to direct action, it makes sense that the visual system is tuned to pick out edges. Throwing away information about the intensity of lightness at every point in space—information that would have enabled you to see that the extreme left and right parts of the figure are the same shade of gray—demonstrates that visual perception efficiently extracts visual features by ignoring some data.

The human visual system processes its input in great detail, but not in every part of the visual field. Here again information is thrown away. For example, when you read you point your eyes at word after word, fixating at a succession of points on the page. When you do that, the image of the word falls on the fovea, a part of the retina that is served by many ganglion cells with tiny receptive fields, each sometimes so small that its entire center region takes input from only a single photoreceptor. The result is that this area is capable of high resolution, and fine details (like the distinguishing squiggles of letters and numbers) can be perceived. Farther out from the point of fixation, the receptive fields get bigger and bigger, so that hundreds of

\[\text{FIGURE 2–10} \quad \text{The Craik-Cornsweet-O’Brien illusion}\]

(a) A gray rectangle is shown with a special edge in the middle. The rectangle seems to be divided with a lighter region on the left and darker region on the right. If you look closely, you will see that actually the two regions are not uniform. There is a gradual transition in each side, producing a sharp change from light to dark in the middle. (b) This is the same figure as in (a), but with the middle region covered by a black rectangle. Now you can see that the gray regions are actually the same. Try putting your finger over the middle of (a) to reveal the illusion.
receptors may be lumped together into a single receptive-field center. These big receptive fields cannot process fine detail and, as a result, neither can you in those portions of the field. Look at the letter “A” in the display below:

A B C D E F G H I J

How far up the alphabet can you read without moving your gaze from the “A”? If you say you can get beyond “E” or “F,” you’re probably cheating. Why throw away all this information? Because it would be wasted: your brain simply could not process the whole image at the detailed resolution available at the fovea.

3.1.3. Neural Processing of Features

The route to the brain from the ganglion cells is via the optic nerves, which meet just before entering the brain to form the *optic chiasm*, so called from the shape of the Greek letter “χ,” or chi. Here some of the fibers from each optic nerve cross to the opposite hemisphere of the brain, sending information from the left side of each eye’s visual field to the right hemisphere and information from the right sides to the left hemisphere. Various pathways carry the information to the lateral geniculate nucleus and thence to the primary visual cortex.

In the primary visual cortex, the extent of the whole visual field is laid out across the surface of the cortex. Cells in primary visual cortex (V1), striate cortex, respond to variations in basic features such as orientation, motion, and color. Output from V1 via the dorsal or ventral pathway feeds a collection of visual areas known collectively as extrastriate cortex (and individually as V2, V3, V4, and so on). Extrastriate cortex contains areas whose cells appear to be specialized for the further processing of these basic features and of more elaborate representations, such as of faces.

Neurons are organized functionally in depth as well as along the surface of cortex. The visual cortex is divided up into *hypercolumns*, chunks of brain with a surface area about 1 millimeter by 2 millimeters and a thickness of about 4 millimeters. All the cells in a hypercolumn will be activated by stimuli in one small part of the visual field. Cells in the next hypercolumn will respond to input from a neighboring portion of visual space. Many more hypercolumns are devoted to the detailed processing of input to the fovea than to the cruder processing of more peripheral parts of the visual field. Within a hypercolumn there is further organization. Here cells are ordered by their sensitivity to specific aspects of the visual feature, such as edges at a specific orientation. Thus, if one cell within a hypercolumn sensitive to edge orientation responds the most to vertical lines, the next cell over will respond most to lines tilted a bit off vertical, and the next to those even a bit more tilted.

It is worth examining the response to orientation a little more closely to appreciate the fine discrimination of neural processing. We are very sensitive to variation in orientation. Under good viewing conditions (with good lighting and nothing blocking the view), we can easily tell the difference between a vertical line and a line tilted 1 degree off the vertical. Does this mean that each hypercolumn needs 180 or more precisely tuned, orientation-detecting neurons, at least one for each degree of tilt from vertical through horizontal (at 90 degrees) and continuing the tilt further to vertical again at 180 degrees? (Think of the tilt of the second hand on a clock dial as
it sweeps from 0 to 30 seconds.) No, the system appears to work differently. Individual neurons respond to a fairly broad range of orientation. A neuron might respond best to lines tilted 15 degrees to the left of vertical and also respond to vertical lines and lines tilted 30 degrees. Precise assessments of orientation are made by comparing activity across a population of neurons. Thus, simplifying for the sake of argument, if some neurons are optimally tuned for tilt 15 degrees to the left and others for the same amount of tilt to the right, a line perceived as vertical would be one that stimulates both these populations of neurons equally.

How do we know this is how it works? One way to demonstrate the differential orientation tuning of neurons is to fixate your gaze on a pattern of lines that have the same tilt, which soon will tire out some of the neurons. Suppose that “vertical” is defined as equal output from neurons sensitive to left and to right tilt; further, suppose we tire out the right-tilt neurons. Now a line that is actually vertical appears to be tilted to the left. The line, which would normally produce equal activity in the left- and the right-sensing neurons, produces more activity in the left-sensing neurons because the right-sensing ones have been fatigued. The comparison of left and right will be biased to the left, resulting in the perception of a tilted line. This bias in perceiving orientation is known as the tilt aftereffect (Figure 2–11)—try it yourself. Similar effects occur in color, size, and (most dramatically) direction of motion. In all cases, the principle is the same: the value of a particular feature is determined by comparison between two or more sets of neurons—with different sensitivities—

![Figure 2–11: The tilt aftereffect](image)

*First, notice that the patterns on the right are both identical and vertical. Now, adapt your visual neurons to the patterns at the left by fixating on each of the black bars between the two patterns. Slowly go back and forth between these two bars 20 times. Immediately thereafter, move your eyes to the circle between the two patterns on the right. Notice that the patterns no longer look perfectly vertical, but seem to tilt. The illusory tilt you see is in the opposite direction of the tilt that you adapted to, so the top will look tilted to the left and the bottom will look tilted to the right.*
responding to that stimulus. If you change the relative responsiveness of the sets of neurons that are being compared, you change the perception of the feature.

Motion is detected in area V5 (also known as MT, for “middle temporal” visual area), an area on the lateral sides of the extrastriate cortex (Dubner & Zeki, 1971). Cells in this area respond to an object moving in a particular direction, such as up or down, or perhaps moving toward or away from the observer. How is it known that this particular brain area is crucial for representing and processing motion in the human brain? Transcranial magnetic stimulation (TMS, see Chapter 1) of this area can temporarily prevent people from seeing motion or induce them to see motion that does not occur (Beckers & Homberg, 1992; Beckers & Zeki, 1995; Cowey & Walsh, 2000). In addition, damage to this area results in akinetopsia, or motion blindness—the loss of the ability to see objects move (Zihl et al., 1983). Those affected report that they perceive a collection of still images. They have difficulty making judgments about moving things: When will that moving car pass me? When do I stop pouring water into the glass?

Other TMS studies have found a specialized region for color perception. Moreover, brain damage to this specific part of the extrastriate cortex, in V4, cause achromatopsia, or cortical color blindness (Zeki, 1990). All color vision is lost and the world appears in shades of gray. And in achromatopsia, unlike as in blindness caused by damage to the eyes or optic nerve, even memory of color is gone.

The existence of these specialized areas suggests that perception starts by breaking down the visual scene into features that are processed separately.

3.2. Putting It Together: What Counts, What Doesn’t

Earlier, we noted that the world does not look like a collection of brightness values (as in Figure 2–6). Nor does it look like a collection of visual properties such as orientation, motion, and so forth. We see a world of objects and surfaces. Those perceived objects and surfaces represent our best guesses about the meaning of the particular visual properties that we are seeing right now. A large set of rules governs the complex process by which we infer the contents of the visual world. In the next sections we will offer a few illustrative examples.

3.2.1. Grouping Principles

To begin, the system must determine which features go together (Derlach et al., 2005). What features are part of the same object or surface? In Germany in the early twentieth century, researchers known collectively as the Gestalt psychologists (Gestalt is the German for “form” or “shape”) began to uncover some of the grouping principles that guide the visual system and produce our perception of what goes with what. Some of these are shown in Figure 2–12. Figure 2–12a is a $4 \times 4$ set of identical, evenly spaced dots. In Figure 2–12b, the effect of proximity, one of the most basic of the grouping principles, groups those dots into rows because, all else being equal, things that are closer to one another are more likely to grouped together—that is, perceived as a whole—than things that are farther apart (Chen & Wang, 2002; Kubovy & Wagemans, 1995; Kubovy et al., 1998). Figure 2–12c shows what happens when all is not equal. Here, the principle of uniform connectedness forms a
vertical organization that overrules proximity. Other principles, shown in Figures 2–12d–f, include properties drawn from topology (for example, does an element have a “hole” in it? Chen et al., 2002).

In the center of Figure 2–13a, you can see a potato-shaped ring of line segments grouping together. Why do these lines group whereas others do not? Here the principle is colinearity: lines group when their orientations are close to that of a neighbor’s. Colinearity is a particular case of relatability (Kellman & Shipley, 1991). The basic idea of relatability is embodied in Figure 2–13b. If line 1 is part of an extended contour in the world, which of the other lines in the neighborhood is likely to be part of the same contour? Line 3 is a good bet. Lines 2 and 4 are plausible. Line 5 is unlikely. Neurons that detect each oriented edge in the image also play a role in computing the extension of that edge into neighboring portions of the image. The potato in Figure 2–13a is the result of a computation performed by a set of those feature detectors. Grouping that occurs between line segments links parts that likely belong to the same contour, helping us move from information about a local edge to information about the shapes of objects.

3.2.2. Filling in the Caps
The grouping principles hold even when only parts of objects are visible, which is very useful when making sense of the confusion of stimuli in the real world. With the proper cues, something not there at all can be interpreted as something that’s there but
3. Building from the Bottom Up: From Features to Objects

hidden—a big difference to our perception, and recognition, of objects. Figure 2–14a shows a mixture of apparently unconnected shapes. When the horizontal bars are drawn between them, as in Figure 2–14b, the shapes cohere into recognizable forms (Bregman, 1981). The edges of the shapes alone have insufficient relatability to suggest how they should connect, but when the bars are added, a new interpretation is possible. The bars reveal that the white areas in Figure 2–14a can be perceived not as voids but as hidden, or occluded, elements of the display. With that additional information

![Image of Figure 2–13](attachment:image.png)
**FIGURE 2–13** Grouping by colinearity and relatability
(a) Some of the lines within this scattered array form a potato-shaped figure. Lines group if their orientation is close to that of a neighbor’s (colinearity), and if it is easy to connect one line to the next (relatability). (b) If line 1 is part of an extended contour in the world, which of the other lines is likely to be part of the same contour?

![Image of Figure 2–14](attachment:image.png)
**FIGURE 2–14** Putting parts together
(a) Shapes and parts with no apparent meaning. (b) The same parts shown with “occluding” black bars. Now the shapes (“B”s) are visible because you can connect the pieces.

the visible edges are relatable and the shapes can be inferred as the visible parts of larger forms. This demonstration shows that perceptual processes can help us fill in the gaps to infer a coherent visual world even when not all the information is given.

Such processing can also lead us to see things that are not in fact there. If a black rectangle is placed across the two white rectangles in Figure 2–15a, we infer that there is just one long white rectangle instead of two short ones. The black rectangle is seen as occluding part of a single white one. (c) The two white rectangles have open ends. One interpretation is, again, one long white rectangle partly occluded by an invisible shape. (d) With more lines added, the invisible rectangle is visible: you see a “subjective” or “illusory” contour.

![Figure 2–15 Illusory contours](a) Two white rectangles. (b) A black rectangle is added. The interpretation changes so that now the figure looks like one long white rectangle instead of two short ones. The black rectangle is seen as occluding part of a single white one. (c) The two white rectangles have open ends. One interpretation is, again, one long white rectangle partly occluded by an invisible shape. (d) With more lines added, the invisible rectangle is visible: you see a “subjective” or “illusory” contour.

Neuroscientific research has discovered the mechanisms that fill in missing contours. Neurons in the primary visual cortex respond to the location and orientation of edges in the sensory world. Connections among the different neurons that respond to edges in different orientations allow them to compare their inputs. Using
3. Building from the Bottom Up: From Features to Objects

Some simple circuitry, neurons that respond to actual edges induce responses in neighboring neurons (Francis & Grossberg, 1996). The end result is that neurons respond to illusory edges in a way similar to the way they respond to a real line in the same space (Bakin et al., 2000; Grosof et al., 1993; Sugita, 1999). The perception of the illusory line is supported by the interaction among neighboring neurons in the primary visual cortex. Construction of perception from sparse cues in the environment is built into the earliest stages of information processing.

3.2.3. The Binding Problem

The examples considered so far have all dealt with the grouping of the same sort of feature—does line 1 go with line 2? What happens when we need to determine whether line 1 goes with color A? This question illustrates the binding problem; that is, how do we associate different features, say, shape, color, and orientation, so that we perceive a single object? The binding problem arises in part because of the way that information processing is carried out by the brain, where one system analyzes color, another shape, and another motion. How do we combine this information so that we see a red ball flying through the air? Part of the answer is that spatial location can serve as the required “glue.” If the roundness, the redness, and the particular motion all occupy the same point in space at the same time, then it seems reasonable that they would be bound together. However, there are limits to the utility of simple spatial co-occurrence. Look for the white vertical bar (or the gray horizontal one) in Figure 2–16; you won’t find it easily. Until you deploy attention to a specific “plus” in the pattern, you cannot tell whether or not grayness goes with verticalness (Wolfe et al., 1990). Although many grouping processes can occur at the same time across the visual field, some—notably binding of different sorts of features to the same object—require attention (Treisman, 1996).

Comprehension Check:

1. What are some of the building blocks of visual perception?
2. What principles are followed when the brain puts it all together?
4. ACHIEVING VISUAL RECOGNITION:
HAVE I SEEN YOU BEFORE?

To understand the world, visual information is not enough. Theseus’s bear is the em-
blem of the problem of recognition, which is to compare current visual information
(large, round, dark, rough edged) with knowledge of the world. (A previously seen
object has a certain shape and color, a bush has another shape, another color.)
Recognition is the process of matching representations of organized sensory input to
stored representations in memory. Determining what is out there in the world—and
reacting to it safely and efficiently if it turns out to be a bear—depends on our abil-
ity to find a correspondence between input from our eyes at the moment and earlier
input that we organized and stored in memory.

4.1. A Brain That Cannot Recognize

Most of the time we don’t even think about what it means to recognize objects.
Sighted people look into a room, see the chairs, tables, books, and ornaments that
may be there and know essentially what these things are, quickly and effortlessly.
Blind people recognize objects by touch or by sound. Recognition is not dependent
on a particular sensory modality. But there are people who have no sensory deficit
at all who nonetheless cannot readily recognize the objects around them. This con-
dition, which is called agnosia (literally, “without knowledge”), results from dam-
age to the brain, not to the sensory organs. When sight is unimpaired and yet
recognition fails, the deficit is known as visual agnosia. The experience of a patient
known as John illustrates the cause and effects of visual agnosia (Humphreys &
Riddoch, 1987).

John, who grew up in England, was a pilot in World War II. After the war, he
married and then worked for a company that manufactured windows for houses, in
time becoming head of marketing for Europe. Following an emergency operation for a
perforated appendix, he suffered a stroke: a small blood clot traveled to his brain
and blocked the arteries that sustained tissue in the occipital lobes. After his stroke,
although he was perfectly able to make out the forms of objects about him and nav-
gate through his room, John was unable to recognize objects. He didn’t know their
names or purposes. He was unable to read. Even after recovering from surgery and
returning home, his ability to recognize objects did not fully return. He even had
difficulty recognizing his wife.

When shown a line drawing of a carrot (Figure 2–17a), John remarked, “I
have not even the glimmerings of an idea. The bottom point seems solid and the
other bits are feathery. It does not seem logical unless it is some sort of brush.” When
shown the drawing of an onion (Figure 2–17b), he said, “I’m completely lost at the
moment. . . . It has sharp bits at the bottom like a fork. It could be a necklace of
sorts.” Shown a set of line drawings like these, John recognized fewer than half. He
was better at naming real objects than drawings, but nonetheless correctly named
only two-thirds of the objects shown to him, even though they were very common
objects such as a book and an apple. Asked to name the same objects by touch,

John’s recognition was much better, establishing that he did not have a general difficulty understanding or speaking the name of the objects, but instead a selective difficulty with visual recognition.

One remarkable aspect of John’s experience is that his impairment did not include failure to detect features or groups. As evident from his descriptions above, he could accurately see features such as pointy edges and shapes. Further, he was fairly good at copying pictures and even drawing objects from memory (although he didn’t recognize what he drew). What he is missing since his stroke is the ability to take the organized visual information he has access to and match it to his visual memories of objects. The selective impairment produced by visual agnosia demonstrates that there are at least some processes used for visual recognition that are not used to extract or organize visual features.

4.2. Models of Recognition

Recognition seems simple for us as we go about the world. But even with an intact brain, recognition is not a trivial act. It remains extremely difficult for even the most sophisticated computer programs. Work in developing computer recognition systems and models by which recognition is achieved has led to remarkable advances during the last 20 years in our understanding of human recognition systems.

Powerful challenges face both computer and brain in the effort to recognize objects. One is viewpoint dependence: an object can be viewed from an infinite combination of possible angles and possible distances, each of which projects a slightly different two-dimensional image on a plane (and on the retina), varying in size or orientation or both. Recognition of an object viewed from different angles presents a particular challenge: the projected two-dimensional image of each three-dimensional part (for example, the seat and the various legs of a chair) changes in size, appearance, and position as a function of rotation (Figure 2–18), yet we have little difficulty recognizing the object as a chair. This challenge is very similar to one of the fundamental problems in perception we discussed earlier, namely, that the sensory input does not contain enough information. All that is
available from any one viewpoint is the two-dimensional projection, so how do we determine the object’s three-dimensional structure?

Then there is the challenge of exemplar variation: there are many different instances of each object category. (The chair in Figure 2–18 is not the only chair in the world.) Any object category consists of many possible examples, yet we readily recognize dining chairs, beach chairs, office chairs, and rocking chairs as all being chairs. This challenge is very similar to the other fundamental problem discussed earlier, namely, that the sensory input contains too much information. How does a computer (and how do we) manage this abundance? One solution would be to store each of these views and each of these examples of chairs as independent representations, but this would make it difficult to generalize our perception of objects to new views or examples. Another way would be to capitalize on the regularities and redundancies of the world by identifying salient features or their underlying structure—in other words, the discriminating features of “chair”—to be able efficiently to match sensory input with stored representations of objects. Understanding how computer systems are designed to overcome these challenges of recognition can help us understand how the human brain might be performing the same feat.

Four types of models have been proposed, each with a different approach to overcoming the challenges of recognition. Template-matching models match the whole image to a stored representation of the whole object. Feature-matching models extract important or discriminating features from the image and match these with
known features of objects. The recognition-by-components model represents the three-dimensional structure of objects by specifying their parts and the spatial relations among those parts. Configural models distinguish among objects that share the same basic parts and overall structure by coding each exemplar according to how it deviates from the average or prototypical object. Each model has advantages and disadvantages that make it suitable for recognition of some objects and not others. It is entirely possible that the human recognition system uses multiple sets of representations and processes, which may be more or less effective for different types of objects.

4.2.1. Template-Matching Models

A template is a pattern, like a cookie cutter or a stencil. It can be used to compare individual items to a standard. A batch of cookies can be compared to a cookie cutter; a broken cookie is rejected (or immediately eaten) because it does not match the specifications of the template cookie cutter. The template-matching method as initially conceived is straightforward and useful as long as the item to be recognized and the template to which the system compares it are almost identical and different from others. However, models based on the traditional idea of a template cannot accommodate variations in object size and orientation—variation that, as we’ve seen, occurs in our sensory life. A template that’s doing its job would reject such apparently different versions.

However, the template-matching models used in modern computer programs are more sophisticated and flexible. These models adjust a scanned image by transformations of size and rotation, stretching it and warping it, to provide a view that is the best possible fit to the templates. Template matching is the method used to recognize bar codes and fingerprints. When the object to be identified is well specified and unique, template matching is a quick and reliable method.

That’s how computers typically do it. Similarly, for humans and other animals representations of objects in memory could be used as templates to match with the sensory input for the recognition of objects. In theory, you could recognize letters of the alphabet by comparing the shape you see with your memory of the shape of each letter of the alphabet until you come up with a match (Figure 2–19a). This method would work reasonably well for printed text because, although type fonts differ, each letter style has a characteristic design that is identical every time that letter appears. But the main disadvantage of the template-matching method is that recognition often demands great flexibility; think of the variety in handwritten letters from one person to another and in different circumstances. No rigid template would reliably match everybody’s “A,” sometimes scrawled in a hasty note, sometimes carefully drawn (Figure 2–19b). Some computer programs designed to recognize handwriting use flexible templates with algorithms that take into consideration factors such as the direction of strokes of the pen and the context of the word. Flexibility is further provided by templates that are constructed from a hierarchy of component templates that each detect a part of the pattern of interest. Computers use flexible hierarchical templates to recognize people from the unique pattern in the iris of the eye (Daugman, 1993). It is still unclear whether or in what circumstances the human brain uses stored representations as templates to recognize objects.
4.2.2. Feature-Matching Models

In some circumstances, accurate recognition does not require that the whole object be fully specified, only some discriminating “features.” Note that we are using the term *features* here in a more general sense than in the discussion of, for example, edges and colors, so it can mean any attribute that distinguishes one object from others. How do you know that’s a tree you’re looking at? You don’t know the exact locations of the branches or the measurements of the trunk, but that doesn’t matter: if you can determine that the thing has those two features—branches and a trunk—it’s a tree.
Feature-matching models search for simple but characteristic features of an object; their presence signals a match. What constitutes a feature in these models? That varies with the type of object. The first stage of visual analysis detects edges and colors, and some models use these simple attributes as features: a feature-matching model could recognize printed letters with a limited set of features that are line segments of different orientations and degrees of curvature. The letter “A” has three such features: a right-slanted line, a left-slanted line, and a horizontal line. No other letter of the roman alphabet has this combination of features. The model would detect these line segments (and only these), and the letter “A” would be accurately recognized (Selfridge, 1955, 1959). Other models require more complex features: models designed for face recognition use eyes, nose, and mouth as features, and models for animal recognition use head, body, legs, and tail. This type of model is more flexible than template-matching models because as long as the features are present it will work, even if the object has parts that may be rearranged. Feature-matching models may also require less storage space than template models because relatively few features would render recognizable many objects of the same category that are not identical.

The feature-matching approach also lends itself well to the idea that processing of information in the brain is parallel (that is, happening at the same time) and distributed (that is, happening in different neural areas). The brain is a network of interconnected neurons with largely interactive components arranged in a loose hierarchy. Such an architecture, diagrammed in the neural-network model discussed in Chapter 1 (see Figure 1–13), has been used to model letter and word recognition as a feature-matching model such as the one shown in Figure 2–20. Recognition is mimicked by a set of simple processing elements, the units of a neural-net model, that interact with one another through excitatory and inhibitory connections. Excitatory connections increase the activity of a unit, inhibitory connections decrease it. In a letter-recognition model, units representing different line segments are connected to units in the next level that represent letters. A connection is excitatory if the letter has the feature specified by that line segment, inhibitory if it does not. When the letter “A” is presented to the network, the right-slant, left-slant, and horizontal line segments become active and excite the units in the letter level that have those features. Some letter units have no additional features beyond what an “A” has but lack some feature of an “A” (for example, neither “V” nor “X” has a horizontal line), so these letter units will become only partially active. Other letter units share some of those features and also have another feature (both “K” and “Y” have slanted lines and also a vertical line); these too will become only partially active. Only the representation of the letter that matches all the features will be maximally active, and go on to influence recognition at the next level of the net, where units representing individual letters excite or inhibit units representing words. By representing those features in an interactive, distributed network, models such as this can recognize any object that has the right features.

For a feature-matching model to be a plausible explanation of how we recognize objects, neurons or populations of neurons should show selectivity to parts of the input similar to the features in the model. Whereas there is much evidence (see the
accompanying *A Closer Look*) to show that neurons in the visual cortex are tuned to lines of specific orientation and degree of curvature (Ferster & Miller, 2000; Hubel & Wiesel, 1959), we do not know whether there are neurons tuned to specific letters or words. Selectivity has been found for other features, such as color, size, texture, and shape (Desimone et al., 1984; Tanaka et al., 1991). Neurons have even shown selectivity to features that are specific parts of objects, such as the eyes of a face (Perrett et al., 1982), and they can become more selective for specific features of objects through experience. Animals that are trained to classify objects as

---

**FIGURE 2–20 A feature net model**

Each circle is a unit of the model that may correspond to groups of neurons in the brain. Lines between units show the connections between units. Connections are excitatory (arrowheads) or inhibitory (dots). Presentation of a stimulus to the network excites the feature-level units in the bottom row. These influence activity in the letter units (middle row), which in turn influence the word units (top row).

Achieving Visual Recognition: Have I Seen You Before?

Space limitations prevent us from describing each experiment in detail, but to get an idea of the logic of experimentation it is useful to look at the details of at least one study cited in the text. For this purpose, we consider a ground-breaking experiment done by David Hubel and Torsten Wiesel (reported in 1959 in “Receptive Fields of Single Neurons in the Cat’s Striate Cortex,” Journal of Physiology, 148, 574–591), which was part of the work that won these researchers a Nobel Prize in Physiology or Medicine in 1981.

Introduction
The investigators were interested in how the neurons in the occipital cortex may be responsible for visual perception. What types of things make neurons respond, and how are the neurons organized?

Method
To test responses of individual neurons, the investigators implanted an electrode into neurons in the occipital lobes of anesthetized cats. By recording the change in voltage on the electrode, they recorded the activity of each neuron and could determine when the neuron was responding. To test what types of things the neurons would respond to, they set the cats up to look at a large projection screen and shined spots of light on the screen. Previous research had successfully used this method to elicit specific responses from photoreceptors and ganglion cells in the eye and map out their receptive fields. The investigators used the same method, but they were recording responses from the primary visual cortex in the occipital lobe.

Results
Unlike the responses of photoreceptors and ganglion cells, most neurons in the primary visual cortex did not respond very much when spots of light were shown to the cat. Diffuse light was also not effective. Instead, the investigators discovered that responses were much stronger to bars of light of a specific orientation. For example, one neuron might respond best when a horizontal bar of light was shown, where as another neuron would respond best when a vertical bar of light was shown. Testing many neurons adjacent to one another in the occipital lobe, they discovered a regular organization of the responses of neurons. The orientation that elicited the strongest response in one neuron, also called the “preferred” orientation, was only slightly different from that of a neighboring neuron. Across a row of adjacent neurons, the preferred orientation varied systematically to map out all orientations.

Discussion
The finding that neurons in the primary visual cortex respond to bars of different orientation demonstrates that these neurons perform a much more sophisticated analysis of the visual world than the photoreceptors or ganglion cells. These cortical neurons can detect lines and may be responsible for detecting the boundaries or edges of objects.
members of different categories (for example, deciding whether an object is—using human terms—a dog or a cat) have neural populations that increase in selectivity for the features that best distinguish the categories (in this case, long neck and short tail) (Freedman et al., 2001, 2002).

The fact that neurons are selective for an array of different features may suggest that the particular features that are important for recognition may vary with the level of detail required at the moment. In the state of high emotional alert described by Theseus—“In the dark, imagining some fear”—a rough outline and round shape can be enough to “recognize” a bear. Our use of feature matching rather than template matching may depend also on how difficult it is to see, and how closely the object matches the “canonical,” or traditional, picture of it. For example, a robin is a canonical bird shape and might be recognized by a template, whereas an emu is not a typical bird and might be recognized by feature matching (Kosslyn & Chabris, 1990; Laeng et al., 1990). Feature matching seems to be a mechanism for recognition that can be used by the brain to recognize categories of objects rather than individual entities.

A major difficulty with early feature models was that they could not distinguish objects with the same component features but arranged in a different spatial relationship, for example, the letters “V” and “X.” Modern computer models, however, encode not only the features in the object but also the spatial relations among them. Thus the representation of “V” might include termination of the lines after meeting at the vertex at the base, and the representation of “X” would include the property of intersection. These more flexible models are fairly successful at recognizing objects in a specific category such as two-dimensional handwritten letters and words, and some models can even recognize exemplars from a particular category of three-dimensional objects such as faces seen across a limited range of views (Penev & Atick, 1996).

### 4.2.3. Recognition-by-Components Model

Although templates and simple features might work in building models for recognition of two-dimensional objects, it is not easy to see how they can solve the problems inherent in recognition of three-dimensional objects across different views, or in recognition of some objects as being different exemplars of the same type of object. Perhaps one clue to how the brain solves these problems is that we may describe objects according to their parts and the spatial relations among those parts (Cave & Kosslyn, 1993; Laeng et al., 1999). The utility of many objects is contingent on the correct arrangement of parts (Figure 2–21). To explain our ability to recognize objects in the varying circumstances presented by the real world, we require a model built on something more flexible than a template and that matches structural information beyond features.

The recognition-by-components (RBC) model provides a possible method for recognizing three-dimensional objects across variations in viewpoint or exemplars (Biederman, 1987). The model assumes that any three-dimensional object can be generally described according to its parts and the spatial relations among those parts. The current model proposes that a set of 24 geometrical three-dimensional
shapes, such as cylinders and cones, can be used to represent just about any object; in the language of the model, these shapes are called geons (Figure 2–22a) (Biederman, 1995). In addition, the spatial relations among geons must be defined: a cone might be “on top of” or “attached to the side of” a cylinder. Almost any object can be specified by its structural description, that is, its components and their spatial relations. A bucket, for example, is a cylinder with a curved rod on the top; a mug is a cylinder with a curved rod on the side (Figure 2–22b). The RBC model detects the geons and their spatial relations and attempts to match the assembled parts to a stored three-dimensional representation of a known object (Hummel & Biederman, 1992).

Geons are useful units for describing objects because their properties are viewpoint invariant (the opposite of viewpoint dependent); that is, they are in the image regardless of the direction from which the object is viewed. Viewpoint-invariant properties include straight lines, corners, and vertices. A straight line, such as the edge of a rectangle, will project to a straight line on any two-dimensional image plane, regardless of the viewpoint (as do the chair legs in Figure 2–18). Each geon is associated with a set of viewpoint-invariant properties that uniquely specify it from the other geons. Thus,
the structural description of an object is viewpoint invariant even when the perceived shape of the object as a whole changes dramatically with viewing conditions.

There is some evidence in support of the RBC model. Participants in behavioral studies can easily recognize geon renditions of man-made objects, suggesting that these simplified representations may have some validity. Evidence also comes from studies making use of visual priming, which produces faster recognition observed when an object is seen for the second time. In general, the effect of priming occurs when a stimulus or task facilitates processing a subsequent stimulus or task—priming “greases the wheels,” so to speak. Using this technique, Irving Biederman (1995) created complementary pairs of images of a given object (say, a flashlight) with some contours deleted (Figure 2–23). Each image in a pair had half the contours of the entire object, and the two images had no contours in common. A second pair of contour-deleted images presented the same object, but one of a different design and, therefore, described by different geons. Participants were shown one member of a pair, then either its partner (built with the same geons) or a member of the other pair (the same object, but described by different geons). Recognition was faster when the second image presented had the same geons as the first.

There is some evidence that neurons in inferior (i.e., bottom) temporal cortex are sensitive to properties that are viewpoint invariant (Vogels et al., 2001), but many neurons respond to an object from only a limited set of views, such as the front view but not the side view of a head (Logothetis et al., 1995; Perrett et al., 1991). The observation that many neurons fail to generalize across all possible views seems to contradict what the RBC model would seem to predict. In addition, although the
RBC theory may account for our recognition of man-made objects, it is less clear how it can be applied to our recognition of natural objects such as animals or plants. Faces are a good illustration of the problem. Faces generally include two eyes, a nose, and a mouth in the same arrangement. The RBC model would construct the same arrangement of geons for every face, and so would not detect individual differences between one face and another—the very way we often, and easily, recognize people. RBC-style models can be good at finding the most commonly used category name of an object (mug, dog), but they have more trouble identifying the specific exemplar (my special coffee mug, my neighbor’s standard poodle).

4.2.4. Configural Models

Configural models often can deal with the limitations of RBC models. They propose that objects that share the same parts and a common structure are recognized according to the spatial relations among those parts and the extent to which those spatial relations deviate from the prototype, or “average,” object. Configural models of recognition help explain how we recognize different individual examples of a category; they have been especially successful in the domain of face recognition (Diamond & Carey, 1986; Rhodes et al., 1987).

In a configural model, specific faces are described by their deviations from the prototypical face, as defined by quantified average proportions in a population. All faces would have the same component parts in the same spatial arrangement, but their relative sizes and distances make each unique.
Several lines of evidence support the configural theory of face recognition. For one thing, we are somewhat better at recognizing caricatures of famous faces, which accentuate the differences from the average face, than more veridical line drawings; this finding suggests that we code faces according to such deviations (Rhodes et al., 1987). Studies have also shown that participants instructed to stare at a particular face and then look at an average face may briefly experience a visual aftereffect in which they perceive the “opposite” or “anticaricature” of the original face (Leopold et al., 2001; Webster & MacLin, 1999; Webster et al., 2004; Zhao and Chubb, 2001). Try it yourself with Figure 2–24.

Several lines of evidence also suggest that only upright faces are processed in this special way. If participants are shown a set of pictures of faces and objects, they are better at recognizing upright faces than a variety of different upright objects, but poorer at recognizing upside-down faces than inverted objects (Yin, 1969). Other studies have shown that inverted faces, like objects that are not faces, are processed in a piecemeal manner, whereas upright faces—the view usually seen in life—elicit more configural or holistic processing (Young et al., 1987). Participants are better at learning the difference between two upright faces that differ only by the shape of a single element, such as the nose, than at learning the difference between two noses shown in isolation (Tanaka & Farah, 1993; Tanaka & Sengco, 1997). Moreover, even though the facial context provides no additional information about the shape of the nose, participants do better at encoding and remembering the nose shape in the context of an upright face. However, no such benefit of holistic processing was found for inverted faces. We are also apparently better at evaluating the overall configuration or spatial relationships among facial features, such as the distance between the eyes or between the nose and the eyes, for upright than for inverted faces (Searcy & Bartlett, 1996).
Neuroscientific research also provides support for the configural model of face recognition. Single-unit recordings from face-selective neurons in the monkey temporal lobes suggest that many neurons respond to the configuration of multiple features rather than to any single face part (Young & Yamane, 1992). In humans, damage to the fusiform face area, a part of the temporal lobes, produces the disorder known as prosopagnosia, the inability to recognize different faces. The deficit is specific; patients have no trouble recognizing that something is a face as opposed to, say, a pumpkin, but have difficulty telling one face from another. The configuration of parts of the face seems to be particularly difficult for them to discern, lending support to the idea that configural processing is important for face recognition. The discovery of a specialized area of the brain for face recognition has ignited debate between scientists who study object recognition, as discussed in the accompanying Debate box.

A variant of this view is the expertise hypothesis, which proposes that a specialized neural system develops that allows expert visual discrimination, and is required to judge subtle differences within any particular visual category (Gauthier et al., 2000). We probably spend more time looking at faces than at any other object. We are face experts—in a single glance we can quickly process the identity, sex, age, emotional expression, viewpoint, and gaze-direction of a face. It is possible that the specialized neural system in the fusiform gyrus is responsible for any recognition process for which we have expertise. Research shows that, while looking at pictures of birds, bird experts show stronger activity in the fusiform gyrus than do other people (Gauthier et al., 2000).

A contrasting view is that many—if not most—visual representations are spatially distributed throughout the ventral pathway. Perhaps the ventral temporal cortex often serves as an all-purpose recognition area for telling apart all different types of objects. Indeed, typically patients with damage to the inferior temporal cortex have difficulty recognizing all categories of objects. In addition, neuroimaging studies of normal object recognition have found that regions outside the fusiform face area that respond suboptimally to faces still show differential responses to faces and to other types of stimuli (Haxby et al., 2001). This means that sufficient visual information is analyzed outside the fusiform face area to distinguish faces from other objects. However, neuropsychological evidence of double dissociations between face recognition and object recognition are difficult to explain if representations are completely distributed. One possible reconciliation is that all ventral areas are involved in object recognition and provide useful information for categorization, but certain distinct systems are necessary for performing fine-tuned discriminations within a category. This is an active area of ongoing research, and no doubt more will be learned about the organization of visual recognition in time.

Comprehension Check:

1. What is visual agnosia?
2. What are the four types of models of object recognition?
Looking at the brain from below, we see the location of the fusiform face area (marked with color ovals) on the inferior side of the cortex. Cortex that is responsive to faces can be found in both hemispheres, as depicted, but for most people the area in the right hemisphere is larger and more responsive.
5. INTERPRETING FROM THE TOP DOWN: WHAT YOU KNOW GUIDES WHAT YOU SEE

Perception is not a one-way flow of information; we are predisposed to understand new information in relation to what we already know. As bottom-up information comes in from sensory organs and is passed up the hierarchy of analysis, concurrent information moves top down (in accordance with your knowledge, beliefs, goals and expectations) and affects earlier processes. Theseus’s bear is more likely to be perceived as the bush it is if you are in the middle of a well-tended garden, not “imagining some fear” in a dark forest where the appearance of a bear is more likely. We use knowledge to make perception more efficient, accurate, and relevant to the current situation, filling in the missing parts of sensory input on the basis of information previously stored in memory. Context counts.

5.1. Using Context

The things we see are not perfect reflections of the world—how can they be? What is the “real” color of a brick wall, the part in sunlight or the part in shadow? Our perception of the basic components of the world, such as colors and objects, is simply inaccurate, as has been demonstrated by psychological experiments and observations during the past hundreds of years (Wade, 1998). So how do we manage in a world so rich with sensory stimuli? We manage because information is interpreted relative to context across all levels of perceptual representation and processing. Our perceptual system has heuristics—problem-solving short-cuts, as opposed to exhaustive algorithms—for making sense of the world by making inferences from the information it receives. Perception is the result of these inferences.

5.1.1. Context Effects for Feature and Group Processing

Visual illusions demonstrate how perception can inter properties that do not exist in the image; a good example is the illusory white rectangle in Figure 2–15. The readily perceived edges of the rectangle are in fact not present in the image; our perceptual systems supply them from the context of black edges and lines. The phenomenon of illusory contours is one way in which perception fills in the missing pieces to make an understandable interpretation of the world.

Studies of visual illusions have revealed that context—including our knowledge, beliefs, goals and expectations—leads to a number of different assumptions about visual features. We expect the brick wall to be “in reality” the same color throughout, so despite the evidence before us caused by changes in illumination across its surface, we believe it to be all the same color; this effect is known as the brightness illusion (Figure 2–25). Similarly, size illusions demonstrate that we assume objects maintain their “true” size across changes in apparent distance from the observer (Figure 2–26). If we did not make these assumptions, and saw “literally” rather than perceived inferentially, the world would be very confusing indeed.
Grouping is an automatic process, and a context of many items can make it difficult to perceive a single item independently. In the design by M. C. Escher shown in Figure 2–27, grouping creates an interestingly ambiguous figure. On the periphery of the image we have clearly defined birds (top) and fish (bottom). As we move toward the center from top and bottom, the objects become respectively less like birds and less like fish. The bird context strives to maintain the avian identity while the fish context supports a piscine interpretation. The result is a region in the middle where you can see either the birds or the fish, but it is hard to see both simultaneously. Grouping makes it difficult to see each item independently, but it allows us to see common attributes of many items at once—here we see the birds as one flock, the fish as one school. We can then perform operations on the group as a whole. It will be less demanding, for instance, to follow the motion of the group than to track each bird independently.

We assimilate all the birds in the Escher design into a single flock because they all look similar. What if one were different? The context effects produced by a group can also be contrasting, making one odd item in a group look even more unusual than in
5. Interpreting from the Top Down: What You Know Guides What You See

FIGURE 2–26  A size illusion

Most of these people in a corridor of the U.S. Capitol appear to be similarly sized in life. The anomaly is the tiny foreground couple (arrow), who are in fact duplicates of, and therefore the same actual size as, the couple in the background. Moving the background couple out of their context reveals the illusion.

(From *Perception* [p. 37] by I. Rock, 1984, Scientific American Library. © Bettmann/CORBIS. Reprinted with permission.)

fact it is. A classic example, the Ebbinghaus illusion (named for its discoverer, the German psychologist Hermann Ebbinghaus, 1850–1913), is shown in Figure 2–28. The central circles in each group are the same size, but the one in the context of the smaller circles looks larger. This illusion is strongest when all the shapes are similar and are perceived as belonging together (Coren & Enns, 1993; Shulman, 1992).
Inferring the motion of the flock of birds as a whole may make it easier for us to see a deviation within that common motion.

5.1.2. Context Effects for Object Recognition

Recognition is dependent on our previous experience with the world and the context of that experience. Recognition of an object may be improved if it is seen in an expected context (you’d agreed to meet your friend at the diner) or a customary one (your friend often eats at that diner), and impaired if the context is unexpected (what’s my cousin from Australia doing at this U.S. diner?) or inconsistent with previous experience (I’ve never seen you here before!). Experiments have shown that the influence of context on recognition of simple objects may be based on allocation of attention (Biederman et al., 1982), or strategies for remembering or responding to objects in scenes (Hollingworth & Henderson, 1998). Context effects in object recognition reflect the information that is important for and integral to the representation of objects.
Research has demonstrated that top-down processing can influence our perception of parts of objects. For example, the context of surrounding letters can manipulate the perception of a target letter in an effect known as word superiority, demonstrated in Figure 2–29 (Selfridge, 1955). The middle letter of each word is actually the identical arrangement of lines, but it is seen as either an “H” or an “A” to fit the context provided. In behavioral studies, participants are better at identifying a briefly flashed letter (for example, “A”) if it is shown in the context of a word (“FAT”) rather than in isolation (“A”) or in a nonword (“XAQ”) (Reicher, 1969; Wheeler, 1970). This is surprising because participants are asked only to identify a single letter and do not need to read the word. You might think that the correct identification of the letters of the word is required before the word can be recognized.

**FIGURE 2–28** The Ebbinghaus size illusion
The central circles in the two sets are the same size. However, the central one on the left looks larger than the central one on the right. In the context of the smaller circles, the central one looks bigger—and vice versa.

You may find it easy to read these words, but a simple feature detector would not. The letters in the middle of each word are actually the identical set of lines. The context of the surrounding letters and their suggestion of a meaningful word let us interpret the central letter as an “H” in the first word and an “A” in the second, so we read “THE CAT.”

because words are made up of letters. So how can the word context help, if you already see the letters? Research like this demonstrates that the recognition of objects is not strictly a matter of putting together the pieces via bottom-up processing. The whole word is recognized by the combined influence of all the letters, thus supporting the identification of each letter because of its context. Later on in this chapter, we will see how an interactive model of recognition can explain the influence of words on the perception of letters and the word superiority effect.

Similar results are obtained when participants are asked to make judgments about components of objects. When asked to judge the color of line segments, participants do better if the line is in a recognizable letter or shape than if it appears in an unusual arrangement (Reingold & Jolicoeur, 1993; Weisstein & Harris, 1974; Williams & Weisstein, 1978). The processing of faces also illustrates the power of context: as we have noted, participants are better at distinguishing faces that differ only by the configuration of the nose than they are at distinguishing various noses presented in isolation. However, the context effect on nose identification disappears if the faces are inverted. This effect, known as face superiority, demonstrates that the parts of an upright face are not processed independently, but rather are recognized in the context of the whole face. These context effects with words and objects demonstrate that our recognition of one part of an image is often dependent on our processing of other aspects of that image. Figure 2–30 shows a striking example of the effect of face context (after Thompson, 1980). The two pictures are faces, one right-side up and the other upside-down. The upside-down one may look a bit strange, but not extraordinarily so. However, if you look at the picture upside-down, so you see the strange image upright, you’ll see that it is actually quite gruesome. The context of the face in its upright position makes it easier for you to see how strange it really is.

5.2. Models of Top-Down Processing

As we have seen, perception is a product of top-down and bottom-up processing. Working from the bottom up, features are combined into some representation of the object and then the object is matched to representations in memory. Working from the other direction, how can we model the effects of context on object recognition?

5.2.1. Network Feedback Models

One of the models proposed for recognition discussed earlier is the network-based feature-matching model, diagrammed in Figure 2–20. In that discussion, our concern was the linking of features to form larger recognizable entities. Because in network models the units at different levels of representation process information at different, and interacting, levels of organization, this same architecture can be used to understand how information at the higher levels (for example, words) can influence information at earlier stages (for example, letters or features of letters). This direction of information flow is feedback because it
presumably is a reaction to incoming, bottom-up information that in turn tunes earlier stages of the system for better performance (Mesulam, 1998).

The feature net model of word recognition demonstrates the mechanics of top-down effects such as word superiority. The feature net model can detect a particular letter from its characteristic line features, such as the curves in the letter “O.” So far, so good—but our visual environment is much more cluttered, variable, and unpredictable than a perfect white printed page. What if ink spilled over part of a letter, as possibly happened in Figure 2–31? Without top-down knowledge, it might be impossible to identify the letter “O”; the visible portions are compatible with “C,” “G,” “O,” and “Q.” However, at the word level of representation, there are only a few three-letter words that begin with the letter “C” and end with the letter “T.” The letters “C” and “T” would make each of these words—“CAT,” “COT,” and “CUT”—partially active. This word unit then feeds information back to the letter representations “A,” “O,” and “U,” while the incoming bottom-up information from the features weakly activates the letters “C,” “G,” “O,” and “Q.” When the
top-down influence is added to the feature information, the “O” receives the strongest facilitation and the word “COT” emerges as the most active unit in the top layer. Feedback facilitation from the top layer resolves the problem of recognition of imperfect input by using stored information about words to guide processing.

The other models of recognition can also use feedback methods between different types of representations to model top-down influences; for example, the recognition of the configuration of an upright face, influenced by our top-down knowledge of what faces usually “look like,” can similarly explain why we perceive parts better in upright faces.

5.2.2. Bayesian Approaches
A different approach to modeling the influence of top-down effects is based on the observation that the influence of stored information is probabilistic; that is, it reflects what has often happened in the past and is therefore likely to happen again. Is it possible that our perceptual systems store information about the likelihood of different events in the perceptual world? If so, the problem of identifying objects becomes similar to the mathematical problem of estimating probabilities. Consider this example: There is a high probability that a banana is yellow, curved, and elongated. A summer squash is also likely to be yellow, curved, and elongated. If you are
looking at something yellow, curved, and elongated, is the object a banana, a squash, or something else? Well, it’s hard to say; a number of things in the world—balloons, for one—could be yellow, curved, and elongated. The probability that a banana has these properties doesn’t really help. Recognition would be easier if we knew the reverse probability, the odds that something yellow, curved, and elongated is a banana. It is possible to estimate reverse probability from the available probabilities by a mathematical rule known as Bayes’s theorem (after the eighteenth-century English mathematician, Thomas Bayes). Bayesian methods use information from previous experience to make guesses about the current environment. Thus, by application of Bayes’s theorem, if you have seen lots of bananas and only a few summer squash, it is a reasonable guess that the present yellow, curved, and elongated object is a banana.

Researchers use the Bayesian approach to demonstrate that previous experience can determine people’s current perceptions. As with context effects, we build up expectations of what we will see based on what we have seen before. During learning of simple tasks, such as detecting black and white patterns, Bayesian models have correctly predicted participants’ abilities (Burgess, 1985). Knowing more, from experience, about which pattern is likely to appear improves the accuracy of perception at the rate Bayesian theory predicts. In more demanding tasks, such as judging the shades of gray of boxes under different lighting, Bayesian models capture our ability to judge the shades and our tendency to assume that the brightest box in any display is painted white (Brainard & Freeman, 1997; Land & McCann, 1971). Bayesian probabilities are even successful in describing much more complicated perceptual judgments such as how we see objects move and change shape (Weiss & Adelson, 1998). Recognition of many other attributes and objects also has been modeled with this approach (Knill & Richards, 1996) because it is a powerful and quantifiable method of specifying how previously stored information is included in the interpretation of current experiences.

Comprehension Check:

1. In what ways does context affect perception of objects?
2. How can top-down processing change how features are perceived?

A broader view of perceptual processes is provided by the view from the window of your room in Condillac’s hilltop château. Remember the situation: you arrived in the dark, with no idea of your surroundings. Now it is morning; someone brings you your brioche and café au lait and opens the curtains. What do you see in your first glance out the window? The panoramic view is too much information to perceive all at once. Yet, immediately, your perceptual processes begin to detect the features and put together the parts, and simultaneously your knowledge about the environment—about trees, fields, mountains, whether or not you’ve seen these particular ones before—gives you some context for shaping the incoming sensory information.
Bottom-up processing is determined by information from the external environment; top-down processing is guided by internal knowledge, beliefs, goals, and expectations. What method do we usually use? That is not a useful question. At any given moment, and for the various interpretations of different stimuli—which in life arrive constantly, and in multitudes—we rely more on one process than the other, but both are essential for perception. Many top-down context effects result from interactions between bottom-up processing and top-down knowledge.

6.1. Refining Recognition

Most of the time, bottom-up and top-down processes work together—and simultaneously—to establish the best available solution for object recognition. Information does not percolate upward through the visual system in a strictly serial fashion, followed by a trickling down of information from processes operating on stored representations. The essence of perception is dynamic interaction, with feed-forward and feedback influences going on all the time. Interactive models of recognition, such as the feature net model (McClelland & Rumelhart, 1981), assume that units influence one another between all layers. Line-orientation units and word-level units influence letter-level units at the same time to specify the degree of activation of the letter-level units.

Similar interaction is observed in the brain. Some visual areas in the dorsal pathway, including MT and attention-related areas in the parietal and frontal lobes, respond soon after the fastest neurons in V1 fire and well before neurons in the ventral pathway can respond (Schmolesky et al., 1998). These fast-reacting high-level areas may be preparing to guide activity in lower level areas.

Interactions between processes can be implemented in the brain because the connections between visual areas are reciprocal. Visual structures (such as the lateral geniculate nucleus, LGN) that process input at earlier stages feed information forward to areas (such as V1) that process later stages; there is also substantial feedback from later stages to earlier stages. Reciprocal connections between different visual areas generally occur between groups of neurons that represent similar locations in the visual field, so these neurons can rapidly exchange information about what features or objects are in that location (Rockland, 2002; Salin & Bullier, 1995). Some of this information processing involves building from center–surround units in the LGN to orientation detectors in V1. Feedback connections from high-level areas to low-level areas help to guide processing in low-level areas. The visual system invests a lot of biologically expensive wiring in these feedback connections. Area V1 sends more projections back to the LGN that it receives from the LGN, and receives more feedback projections from area V2 than it sends upward to V2. “No man is an island,” and no visual area operates independently of its neighbors. These reciprocal connections allow for iterative processing, that is, processing in which information is repeatedly exchanged between visual areas, each time with additional data, to refine the representation of the stimulus and extend the duration of its representation (Di Lollo et al., 2000). The brain appears to be organized in a way that promotes the interaction of top-down and bottom-up processing.
6.2. Resolving Ambiguity

Information from any single vantage point is fundamentally ambiguous. Because we can never be sure what is out there in the real world, the brain must analyze incoming information to provide the most likely result. Usually there is only one best solution, but sometimes there is more than one. Take the Necker cube (Figure 2–32), for example, named after the nineteenth-century Swiss crystallographer Louis Albert Necker, who observed that some of his line drawings of crystal structures seemed spontaneously to reverse their orientation. This famous figure can be perceived as a three-dimensional cube seen either from above or from below (or occasionally as a flat two-dimensional figure). When looking at such ambiguous stimuli, we typically experience bistable perception—that is, we can perceive both interpretations, but only one at a time. We can’t see both interpretations at once, even though we know that both exist and in fact have seen them. Bistable perception leads to spontaneous alternations between the two interpretations, even when we keep our eyes focused on a fixation point so that the bottom-up input is held constant. The phenomenon is a demonstration that the visual system is highly dynamic and continuously recalculates the best possible solution when two are initially in equilibrium.

Neural networks can model these spontaneous alternations by relying on two principles, competition and adaptation. If one of two possible interpretations produces a stronger pattern of activation, it will suppress the other, producing a single winning interpretation. However, the ability of the “winner” to suppress the “loser” gradually adapts or weakens over time, until the “winner” can dominate. The process is similar to a contest between two wrestlers. As they roll on the mat trying to pin each other, the one who is on top seems to be succeeding but is vulnerable to attacks from the one below. If their skills are equal, or nearly so, they will do many turns, with each wrestler enjoying a series of momentary, and alternating, successes and failures. Perceptual interpretations compete in similar fashion to be the “winner,” when two possibilities can both fit the incoming information, so there is no clear winner, we see, as it were, the wrestling match in progress (Levelt, 1965).

Bistability can occur at many levels in the visual system, as demonstrated by the different types of ambiguous figures that can vex us. Some, such as the Rubin vase

![An ambiguous figure: the Necker cube](image-url)

The cube (a) has two possible interpretations. You can see either a cube facing down to the left (b) or one facing up to the right (c). This ambiguous figure will appear to flip back and forth between the two interpretations spontaneously.
(Figure 2–33a), named for the Danish psychologist Edgar Rubin (1886–1951), and Escher’s birds and fish, present an ambiguity of figure–ground relations. In these cases the two interpretations differ according to the part of the image that appears to be the figure, “in front,” and the part that appears to be the background. Other ambiguous figures, such as the duck–rabbit figure (Figure 2–33b), show a competition between two representations that correspond to different interpretations. Parts of the ventral extrastriate cortex involved in object recognition become active during the spontaneous reversals of these ambiguous figures (Kleinschmidt et al., 1998), suggesting that these extrastriate object areas are probably linked to our conscious experience of objects.

Further hints regarding the nature and origins of consciousness are provided by a form of bistable perception called binocular rivalry, a state in which individual images to each eye compete (Andrews et al., 2005). If a different monocular image—that is, an image seen by only one eye—is viewed in the fovea of each eye, we spontaneously alternate between the two images, reversing every few seconds and never seeing both at the same time. This is a particularly interesting phenomenon because a clear distinction can be made between what is presented and what is consciously perceived. The images are there, in front of a healthy visual system. When they are presented together but perceived only alternately, what neural activity is taking place in the brain beyond the retina? Neurophysiological studies in monkeys have found neural activity that is correlated with awareness in higher visual areas (Leopold & Logothetis, 1996). Human neuroimaging studies have found corresponding alternation of activation of high-level face- and place-selective brain areas during rivalry between a face image and a house image (Tong et al., 1998). More important, such effects have been found in the primary visual cortex (Polonsky et al., 2000; Tong & Engel, 2001), suggesting that this form of perceptual competition occurs at the earliest stage of cortical processing. Studies of rivalry provide evidence for the locus of the neural correlate of consciousness, and the results suggest that activity even as early as the primary visual cortex may be involved in consciousness.
But these neurophysiological studies and neural-net models do not explain the essential element of bistable perception—mutual exclusivity. Why can’t we have multiple perceptual interpretations at once? The full answer is not yet known, but one explanation is that bistability is a by-product of the inhibition that is necessary for the successful functioning of our brains and neural networks. Seeing both stimuli in binocular rivalry would not be helpful to the human organism. If you hold your hand in front of one eye so that one eye sees the hand and the other sees a face in front of you, it would be a mistake—that is, very far from the reality of the stimuli—for the visual system to create a fused hand-face. One percept has to win and inhibit the other possibilities. Most of the time, there is one clear winner. The conditions that produce strong rivalry and bistability arise in the lab more often than in our everyday perceptual lives.

6.3. Seeing the “What” and the “Where”

Vision is about finding out what is where. To guide our actions, we need to be able to identify objects and know their precise spatial position. As previously mentioned, the processes for determining what and where are implemented in separate pathways in the brain (Figure 2–34). Spatial processing of location relies on the dorsal “where” pathway, which consists of many visual areas that lead from V1 to the parietal lobes. Object recognition relies on the ventral visual pathway, which projects from V1 to ventral areas such as V4 and the inferior temporal cortex. In a classic study, Ungerleider and Mishkin (1982) demonstrated that these two anatomical
pathways perform these specific functions by lesioning the brains of monkeys that were trained to do both a recognition and a localization task. Monkeys with damage to the inferior temporal cortex, in the ventral pathway, had selective impairments in object recognition. They were no longer able to distinguish between blocks of different shapes, such as a pyramid and cube. Monkeys with damage to the posterior parietal cortex, in the dorsal pathway, had impaired ability for localizing objects. They were no longer able to judge which two of three objects were closer together. Neuroimaging of normal brain function in humans also shows this dissociation: there is more activity in dorsal areas with localization tasks and more activity in ventral areas with recognition tasks.

“What” and “where” may have separable neural substrates, but we experience a visual world in which “what” and “where” are integrated. Information about what an object is must interact with information about where an object is to be combined into our perception of the world. Very little is understood about how the brain accomplishes this feat; so far, research has been able only to describe the responsibilities of the two visual pathways. One proposal is that the dorsal pathway may be involved in planning visually guided actions as well as in localizing objects (Goodale & Milner, 1992).

Investigators tested a patient who had diffuse damage throughout the ventral stream as a result of carbon monoxide poisoning. She had severe apperceptive agnosia, that is, impairment in judging even basic aspects of the form or shape of objects (Goodale et al., 1990, 1991). She could not even describe a line as vertical, horizontal, or tilted. However, if she was asked to “post” a card through a slot tilted at a particular angle, she could do so accurately (Figure 2–35; A. D. Milner et al., 1991), but could not say
which way the slot was oriented. Her deficit could not be attributed to impaired language ability or to an inability to understand the task, because when she was asked to rotate the card to the same angle as the slot seen at a distance, she could; but she couldn’t report which way (or whether) the slot was oriented. These findings suggest that she had access to the information about orientation of the slot only through action.

By contrast, damage to the dorsal pathway can lead to apraxia, the inability to make voluntary movements even though there is no paralysis (for a review, see Goodale et al., 1990; Koski et al., 2002). Patients with apraxia can perform actions from memory and have no difficulty describing what they see; they would not have difficulty reporting the orientation of the card slot. However, they have tremendous difficulty performing new actions on what they see, such as posting the card through the slot. These and other findings support the notion that the dorsal and ventral pathways can be doubly dissociated and therefore support separate functions. Models of recognition and spatial localization suggest that the separation of these functions leads to better performance of each, as long as enough resources (that is, nodes and connections) are available (Rueckl et al., 1989). Just what types of functions each pathway supports and how the two interact are still being explored.

**Comprehension Check:**

1. Does perception come from bottom-up or top-down processing?
2. What are the “what” and “where” pathways?

**Revisit and Reflect**

1. **What is perception and why is it a difficult ability to understand?**
   
The senses are our window into the world, and they provide the raw material for building an understanding of the environment. The primary goals of perception are to figure out what is out there and where it is. But perception is not a simple registration of sensations: it involves interpretation of often ambiguous, insufficient, or overwhelming information in the light of your knowledge, beliefs, goals and expectations. Ambiguous: Is it a bear or a bush, a rabbit or a duck? The context of the night scared you—it’s only a bush. And bistability lets you see duck–rabbit–duck–rabbit and protects you from the confusion of duck–rabbit.

   Not enough: Sensory input does not contain enough information to specify objects precisely, so we must make unconscious assumptions and guesses. Too much: Too much sensory input is available at any given moment, so processing must, again unconsciously, capitalize on redundancies and expectations to select the important data for detailed analysis.

   **Think Critically**

   - Do you think it is possible that aliens from another planet might have better perceptual systems than ours? Why or why not?
1. Is what constitutes “too much” information always the same, from moment to moment, or does this depend on context? If the latter, how do perceptual systems alter their performance depending on context to take in more or less information?

2. **What general principles help us to understand perception?**

In the brain, bottom-up processes and top-down processes continuously interact, enabling the development and refinement of useful percepts. Bottom-up processes detect the features of sensory stimuli—such as edges, spots, color, and motion. The visual system makes conscious and (as when supplying missing portions of a form) unconscious inferences from these groupings. Occasionally the inferences are “incorrect,” as in the case of illusory contours, but often are nonetheless useful, enabling us to navigate the sensory world. Top-down processes rely on knowledge, beliefs, goals and expectations to guide perceptual exploration and interpretation. Perceptual mechanisms in the brain throw away some information that is redundant so they can pare down input to the essential features and fill in missing information from stored information about how things usually look and expectations about what is of interest at the moment.

**Think Critically**

- When is perception more or less demanding in everyday life? How might actions such as driving a car in traffic or reading in a noisy environment rely more or less on top-down processing?
- How might adults and children be different in their perception of common objects, such as bottles and faces? How about rare objects, such as a wing nut and a platypus?

3. **How do we put together parts to recognize objects and events?**

The building blocks of visual processing are detected at early stages of visual analysis and then combined to bring about object recognition. Feature detectors, such as the neurons that respond to lines and edges, can have local interactions that can promote a global interpretation, such as a long line or edge. Grouping principles are rules that perception uses to put together features that likely belong together, for example because they are close together (grouping by proximity) or alike (grouping by similarity). Various other principles also underlie how we organize features into patterns that are likely to correspond to objects.

**Think Critically**

- Why do we say that two things are “similar” or “dissimilar”? It has sometimes been said that in order to understand the nature of similarity we would need to understand most of visual perception. Why might this be true?
- Say you were magically transported to the planet Ziggatat in a different dimension and when you looked around you didn’t see any object you recognized. How would you describe what you saw? How could you tell where one object ended and another one started?
4. **How do we recognize objects and events?**

Models of ways in which the brain may recognize objects and events include template-matching models, which match sensory information in its entirety to a mental template; feature-matching models, which match discriminating features of the input to stored feature descriptions of objects; recognition-by-components models, which match parts arranged in a specified structure to stored descriptions of objects; and configurational models, which match the degree of deviation from a prototype to a stored representation. Objects may be broken into three-dimensional parts (such as geons) that lead to recognition through their arrangement; configurations of object parts may be the key element that allows recognition of some objects, such as faces. It is likely that the brain recognizes objects by a combination of these representations and processes to maximize reliability of perception and make recognition faster and more economical. Visual perception seems to capitalize on the best method to recognize objects depending on the object to be recognized.

**Think Critically**

- What are the relative advantages and disadvantages of the main methods of recognizing objects?
- “Recognition” is sometimes distinguished from “identification.” When this distinction is made, recognition consists of simply matching the perceptual input to stored perceptual information, so that you know the stimulus is familiar; in contrast, identification consists of activating information that is associated with the object (such as its name and categories to which it belongs). Do you think this distinction is useful? What predictions might it make about possible effects of brain damage on perception?

5. **How does our knowledge affect our perception?**

Knowledge about objects provides the basis for recognition. Knowledge also guides perception to the most likely interpretation of the current environment; this interpretation allows us to compensate for missing segments of an edge by extending the detected edges to fill in our perception. In addition, the context surrounding a feature, group, or object helps to determine perception; context can facilitate recognition when it is complementary, or impair recognition when it is misleading. Interactions between knowledge and current perceptual input bring about perception.

**Think Critically**

- How might people from different parts of the world perceive things differently? What types of surroundings would improve or impair recognition for different peoples?
- Back on the planet Ziggatat, say you’ve figured out what parts belong to what and have come up with names for the objects. What problems will remain as you learn this new environment?
6. Finally, how do our brains put together the many and varied cues we use to perceive?

Reciprocal neural connections between brain areas play a key role in integrating cues that are processed in different pathways—no visual area operates independently of its neighbors—ensuring that information can be fed forward and back between levels of representation. The essence of perception is dynamic interaction, with feed-forward and feedback influences going on all the time; interactive models of recognition assume that units influence one another between all layers. Moreover, perceptual systems find a single interpretation of the input in which all of the pieces fit together simultaneously, even if another interpretation is possible. Interpretations are achieved and changed in accordance with the principles of competition and adaptation: if one of two (or more) possible interpretations produces a stronger pattern of activation, this interpretation suppresses the other(s); however, the “winner” gradually adapts and weakens over time, until a “loser” can dominate. Thus, if the stimulus is ambiguous, your perception of it will change over time. Finally, in some cases, distinct systems—such as those used to determine “what” and “where”—operate simultaneously and relatively independently, and are coordinated in part by the precise time when specific representations are produced; this coordination process relies on attention, which is the subject of the following chapter.

Think Critically

- Why does it make sense that processes are always interacting, instead of only after each has “finished” its own individual job?
- Is it better to have one interpretation of an ambiguous stimulus than to try to keep in mind all the ways the stimulus could be interpreted? Why do you think the brain “wants” to find a single interpretation?