Linking Mind to Brain: The Mathematics of Biological Intelligence

Stephen Grossberg

How our brains give rise to our minds is one of the most intriguing questions in all of science. We are now living in a particularly interesting time to consider this question. This is true because, during the last decade, mathematical models about how the brain works have finally succeeded in quantitatively simulating the experimentally recorded dynamics of individual cells in identified brain circuits and the behaviors that these circuits control. The models that have led to these successes incorporate qualitatively new ideas about how the brain is organized to achieve the remarkable flexibility and power of biological intelligence.

These advances represent significant challenges and opportunities for mathematicians for several reasons. One obvious reason is that the models themselves are interesting mathematical objects. These models are typically defined by high-dimensional dynamical systems in which several types of nonlinear feedback operate across multiple spatial and temporal scales. They represent systems that are capable of autonomously adapting, or self-organizing, in response to a rapidly changing and unpredictable world.

A second reason is that these models represent processes that are of great scientific *and* personal interest, including

- how we perceive a rapidly changing world,
- how we learn from it on our own,
- how we form expectations about what we believe will happen next,
- how we recognize complex events even if they occur in different contexts and from different viewpoints,
- how we join what we *know* about the world with what we *feel* about it to make decisions that are both possible and valued,
- how we selectively attend to events that particularly interest us,
- how we plan what to do next and carry out these plans despite the occurrence of multiple distracting events, and
- how we fuse together multiple types of information to decide where to look with our eyes, where to reach with our arms, and where to move with our legs.

The great variety of these capabilities is mirrored by the heterogeneous organization of the brain into distinct but highly interacting parts, including such regions as the cerebral cortex, hippocampal system, cerebellum, basal ganglia, reticular formation, and so on. Detailed models of how these regions are organized have recently been proposed, notably of how the characteristic organization of neocortex into layered circuits helps it to develop and learn in a stable way, to group distributed information into coherent representations without a loss of analog sensitivity, and to pay attention to interesting events [5, 8].

Stephen Grossberg is Wang Professor of Cognitive and Neural Systems and also professor of mathematics, psychology, and biomedical engineering at Boston University. His e-mail address is steve@cns.bu.edu.

Supported in part by the Defense Advanced Research Projects Agency and the Office of Naval Research (ONR N00014-95-1-0409), the National Science Foundation (NSF IRI-97-20333), and the Office of Naval Research (ONR N00014-95-1-0657).

A third reason is that these models may be applied in both medical research and technology in ways that can have profound effects on society. Applications to medical research include using the models to develop new algorithms for preprocessing and classifying complex medical images and to show how the normal brain can break down in various ways to cause mental disorders, such as schizophrenia and Parkinson's disease. Applications to technology include all the application domains wherein one can use a system that can behave intelligently on its own in response to large amounts of noisy data in an unpredictable environment. Such models are already being used in applications as varied as airplane design, automatic generation of world maps from remote sensing data, classification of macromolecules, and development of adaptive software for the World Wide Web (Carpenter [1]).

A fourth reason is that these models of the mind may clarify our understanding of many historical developments in mathematics, including the sources of mathematical competence itself, and the neural basis of geometrical concepts, logical thinking, imagination, and even consciousness.

A Theoretical Approach to Linking Brain to Mind

The crucial role of mathematical modeling and analysis in these developments can be appreciated by knowing the type of theoretical method that has led to the recent breakthroughs. This sort of model development begins with an analysis of behavioral data, typically scores or even hundreds of parametrically structured behavioral (that is, psychological or cognitive) experiments in a particular problem domain. One begins with behavioral data because the brain has evolved in order to achieve *behavioral* success. Any theory that hopes to link brain to behavior thus needs to discover the computational level on which brain dynamics control behavioral success. As in other scientific disciplines, one works with large amounts of data because otherwise too many seemingly plausible hypotheses cannot be ruled out.

A crucial constraint is to insist upon understanding the behavioral data, which come to us as static numbers or curves on a page, as the emergent properties of a dynamical process that is taking place moment by moment in an individual mind. One also needs to respect the fact that our minds can adapt on their own to changing environmental conditions without being told that these conditions have changed. One thus needs to frontally attack the problem of how an intelligent being can *autonomously adapt to a changing world*. How this happens has led to core new insights.

Such an approach has regularly uncovered new organizational principles and mechanisms, which are typically realized as a "minimal model" that obeys locally defined laws that are capable of operating on their own in real time. A minimal model in this sense is one that loses some important functional property if any of its mechanisms is removed. An important mathematical task is to understand how variations of these mechanisms may have been specialized to deal with different environmental conditions. The remarkable fact is that, whenever such a psychologically derived model has been written down, it has always been interpretable as a neural network that has always included known brain mechanisms. The interpretation of these brain mechanisms has, however, often been novel, because the behavioral analysis clarifies how these brain mechanisms lead to useful, and often unsuspected, functional properties. The networks have also typically predicted the existence of unknown neural mechanisms, and many of these predictions have been supported by subsequent neurophysiological, anatomical, and even biochemical experiments over the years. Once this neural connection has been established by a top-down analysis, one can then work both top-down from behavior and bottom-up from brain to better characterize and refine the model. This merging of behavior and brain in a single model also facilitates their transfer to technological applications; behavior provides the functions and brain the mechanisms that are needed for the technological design.

A fundamental empirical conclusion can be drawn from many experiences of this type: namely, the brain as we know it can be successfully understood as an organ that is designed to achieve autonomous adaptive behavior in response to a changing world. Said in another way, the brain looks the way it does because its networks provide a natural computational framework with which to implement autonomous behavioral adaptation to a changing world.

It has always proved to be the case that the level of brain organization that computes behavioral success is the network or system level; that is why the field of neural networks is so important to this endeavor. Does this conclusion mean that individual nerve cells, or even smaller cellular components, are unimportant? Not at all. Properly defined individual nerve cells and their interactions are needed to correctly define the network and system laws whose interactive, or emergent, properties map onto behavior as we know it. The January 2000 *Notices* article of Nancy Kopell entitled "We Got Rhythm: Dynamical Systems of the Nervous System" provides an excellent example of how important it is to correctly define the individual nerve cells.

These remarks clarify that in order to understand how a brain gives rise to a mind, one must be able to freely move between (at least) the three levels of neuron, network, and behavior, with behavior understood as emergent properties of networks of neurons. Doing this requires that one has a sufficiently powerful theoretical language. The language of dynamical systems has proved to be the relevant tool, namely, those particular classes of nonlinear feedback systems with large numbers of components operating over multiple spatial and temporal scales. Although it requires an interdisciplinary knowledge to derive these models and to test their ability to explain behavioral and brain data, once they are derived they may be studied as mathematical objects with fascinating formal properties. The mathematical study of these systems is still in its infancy, but there are already available many computational studies of these systems that can help to frame such analyses, as well as a core of basic theorems, including global theorems about the limiting and oscillatory behavior of cooperative and competitive systems, and about how we learn to classify complex events in the world and recall them from memory on demand. For an overview see the book [2]. It is in this area that I believe a great opportunity for mathematicians exists that may be on the scale of the opportunities afforded to generations of mathematicians from insights into theoretical physics.

Complementarity: A Global View of Brain Organization

In a brief general article like the present one, it is not easy to survey such a broad field. Perhaps an overview followed by some examples might be most helpful. Let me begin with some comments about how the brain seems to be functionally organized in the large.

In one traditional view, our brains are proposed to possess independent modules, as in a digital computer. For this view, we see by processing perceptual qualities such as visual form, color, and motion using different modules. This view's supporters sometimes turn to the well-known fact that the brain is organized in parallel processing streams. Figure 1 schematizes how at least three such processing streams within the visual cortex are activated by light impinging on the retina. One such stream goes from the retina through a processing stage called LGN Parvo (so named because of its "parvocellular" cell type) to the cortical processing stages V1 Blob, then V2 Thin Stripe, then V4, and then inferotemporal cortex. Another such stream goes from retina through LGN Parvo, then through V1 Interblob, V2 Interstripe, then V4, and again on to inferotemporal cortex. A third stream goes from retina through LGN Magno (so named because of its "magnocellular" cell type) to cortical processing layer 4B in area V1, then to V1 Thick Stripes, then MT, and then parietal cortex. More will be said about the role that these streams play in vision in a moment.

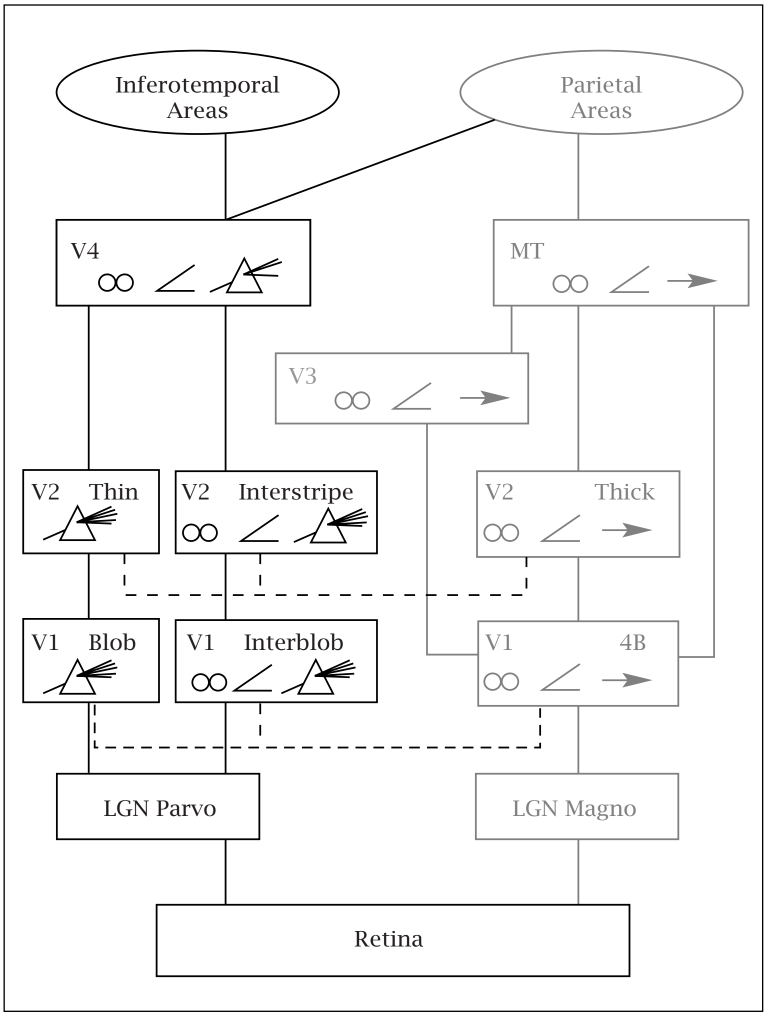

Figure 1. Processing streams for visual data in the early visual areas of the macaque monkey brain. Data flow upward from the retina. In FACADE theory, the two left-hand streams in black are the "What" streams, used in identifying what an object is. The streams in gray are the "Where" streams, used to localize an object. Within the "What" streams, the left-hand or "Blob" stream deals with surfaces, and the middle or "Interblob" stream deals with boundaries. These streams interact with one another. The upper-left symbol in each processing stage (represented by a box) denotes the brain region (e.g., cortical area "V1 Blob"). The icons designate properties of cells in each region (e.g., the prism designates wavelength sensitivity). Adapted with permission from DeYoe and van Essen, *Trends in Neurosciences* 11 (1988), 219.

The existence of such streams certainly supports the idea that brain processing is specialized, but it does not in itself imply that these streams are independent modules that are able to fully compute their particular processes on their own. In fact, many perceptual data argue against the existence of independent modules, because strong interactions are known to occur between perceptual qualities. For example, changes in the perceived form or color of an object can cause changes in its perceived motion and vice versa, while changes in the perceived brightness of an object can cause changes in its perceived depth and vice versa. The existence of such interactions suggests that the mechanisms whereby we perceive the geometry of the world do not obey the classical geometrical axioms on which a large amount of mathematics is based. How and why do these qualities interact? What is the geometry by which we really see the world? An answer to these questions is needed to determine the functional and computational units that govern behavior as we know it.

A great deal of theoretical and experimental evidence suggests that the brain's processing streams compute *complementary* properties. Each stream's properties are related to those of a complementary stream, much as a lock fits its key or two pieces of a puzzle fit together. We are all familiar with complementarity principles in physics, such as the famous Heisenberg Uncertainty Principle of quantum mechanics, which notes that precise measurement of a particle's position forces uncertainty in measuring its momentum and vice versa. As in physics, the mechanisms that enable each stream in the brain to compute one set of properties interfere with its computing a complementary set of properties. Because of the complementarity of the brain's processing streams, each stream exhibits complementary strengths and weaknesses. How, then, do these complementary properties get synthesized into a consistent behavioral experience? It is proposed that *interactions* between these processing streams overcome their complementary deficiencies and generate behavioral properties that realize the unity of conscious experiences. In this sense, *pairs* of complementary streams are the functional units, because only through their interactions can key behavioral properties be competently computed. These interactions may be used to explain many of the ways in which perceptual qualities are known to influence each other. Thus, although analogies such as when a key fits its lock or puzzle pieces fit together are suggestive, they do not fully capture the dynamism of what complementarity means in the brain.

It is also well known that each stream can possess multiple processing stages. For example, in Figure 1 there are distinct processing stages in the LGN, followed by the cortical areas V1, then V2, and then V4 on their way to the inferotemporal and parietal cortices. Why is this so? Accumulating evidence suggests that these stages realize a process of *hierarchical resolution of uncertainty*. In the brain, the uncertainties in question are proposed to be overcome by using more than one processing stage to form a stream. Overcoming informational uncertainty utilizes hierarchical interactions within the stream and also the parallel interactions between streams that overcome their complementary deficiencies. The computational unit is thus not a single processing stage; it is, rather, an ensemble of processing stages that interact within and between complementary processing streams.

According to this view, the organization of the brain obeys principles of uncertainty and complementarity, as does the physical world with which brains interact and of which they form a part. These principles reflect each brain's role as a self-organizing measuring device *in* the world and *of* the world and may better explain the brain's functional organization than the simpler view of computationally independent modules. Illustrative experimental and theoretical evidence for complementary processes and processing streams is described below.

All Boundaries Are Invisible

Visual processing provides excellent examples of parallel processing streams (Figure 1) and of how the dynamics of the brain differ qualitatively from traditional mathematical axioms about geometry. What evidence is there to suggest that these streams compute complementary properties, and how is this done? A neural theory, called FACADE (Form-And-Color-And-DEpth) theory and introduced principally in [4], proposes that perceptual *boundaries* are formed in the LGN-Interblob-Interstripe-V4 stream, while perceptual *surfaces* are formed in the LGN-Blob-Thin Stripe-V4 stream. Many experiments have supported this prediction.

FACADE theory suggests how and why perceptual boundaries and perceptual surfaces compute complementary properties. Figures 2A and 2B illustrate three pairs of complementary properties using a visual illusion that is called a *Kanizsa square*. For example, in response to viewing Figure 2A, our brains construct a percept of a square even though the image contains only four black Pac-Man, or pie-shaped, figures on a white background. As noted below, this percept is due to an interaction between the processing streams that form perceptual boundaries and surfaces.

One might immediately wonder why our brains construct a square where there is none in the image. There are several functional reasons for this. One is that there is a *blind spot* in each retina, namely, a region where no light-sensitive photoreceptors exist. This region is blind because of the way in which the pathways from retinal photoreceptors are collected to form the optic nerve that carries them from the retina to the LGN in Figure 1. We are not usually aware of this blind spot because our brains complete boundary and surface information across it. The actively completed parts of these percepts are visual illusions, because they are not derived directly from visual signals on our retinas. Thus many of the percepts that we believe to be "real" are visual illusions whose boundary and surface representations just happen to look real. I suggest that what we call a visual illusion is just an unfamiliar combination of boundary and surface information. This hypothesis is illustrated by the percepts generated in our brains from the images in Figure 2.

In response to the images in Figures 2A and 2B, our minds tend to form illusory contours *inwardly* between cooperating pairs of colinear edges in the Pac-Man inducers. Four such contours form the boundary of a Kanizsa square. (If boundaries were formed outwardly from a single inducer, then any speck of dirt in an image could crowd all our percepts with an outwardly growing web of boundaries.) This boundary completion process is *oriented* to form only between like-oriented and (almost) colinear inducers. Both of these properties are useful to complete edges in a scene that are not fully detected at the retina due to the blind spot. The square boundary in Figure 2A can be both seen and recognized because of the enhanced illusory brightness of the Kanizsa square relative to its background; see below for an explanation. In contrast, the square boundary in Figure 2B can be recognized even though it is not visible; that is, there is no brightness or color difference on either side of the boundary. Figure 2B shows that *some* boundaries can be recognized even though they are perceptually unseen or invisible. FACADE theory predicts that *all boundaries are invisible* within the boundary stream, which is proposed to be the Interblob cortical processing stream (Figure 1).

Why should all boundaries be invisible? How could this assertion be framed within a theory? The invisible boundary in Figure 2B can be traced to the fact that its vertical boundaries form between black and white inducers that possess opposite contrast polarity with respect to the gray background; that is, the black inducers have a black-to-gray, or dark-to-light, polarity with respect to the background, whereas the white inducers have a white-to-gray, or light-to-dark, polarity with respect to the background. The same is true of the boundary around the gray circular disk in Figure 2C. In this figure the gray disk lies in front of a black and white textured background whose contrasts with respect to the disk reverse across space. In order to build a boundary around the entire disk, despite these contrast reversals, the boundary system pools, or adds, signals from pairs of *simple cells* that are sensitive to the same orientation and position but to opposite contrast polarities. This pooling process occurs in the V1 Interblob stream at the *complex cells.* This is how the square boundary in response to Figure 2B and the circular boundary in response to Figure 2C start to form in our brains. This pooling process renders the boundary system output *insensitive* to contrast polarity. The boundary system hereby loses its ability to represent visible colors or brightnesses, since its output cannot signal the difference between dark and light. It is in this sense that "all boundaries are invisible." The *inward* and *oriented* boundary completion process that forms the illusory square is activated by these pooled signals in the V2 Interstripe area. These three properties of boundary completion are summarized in Table 1.

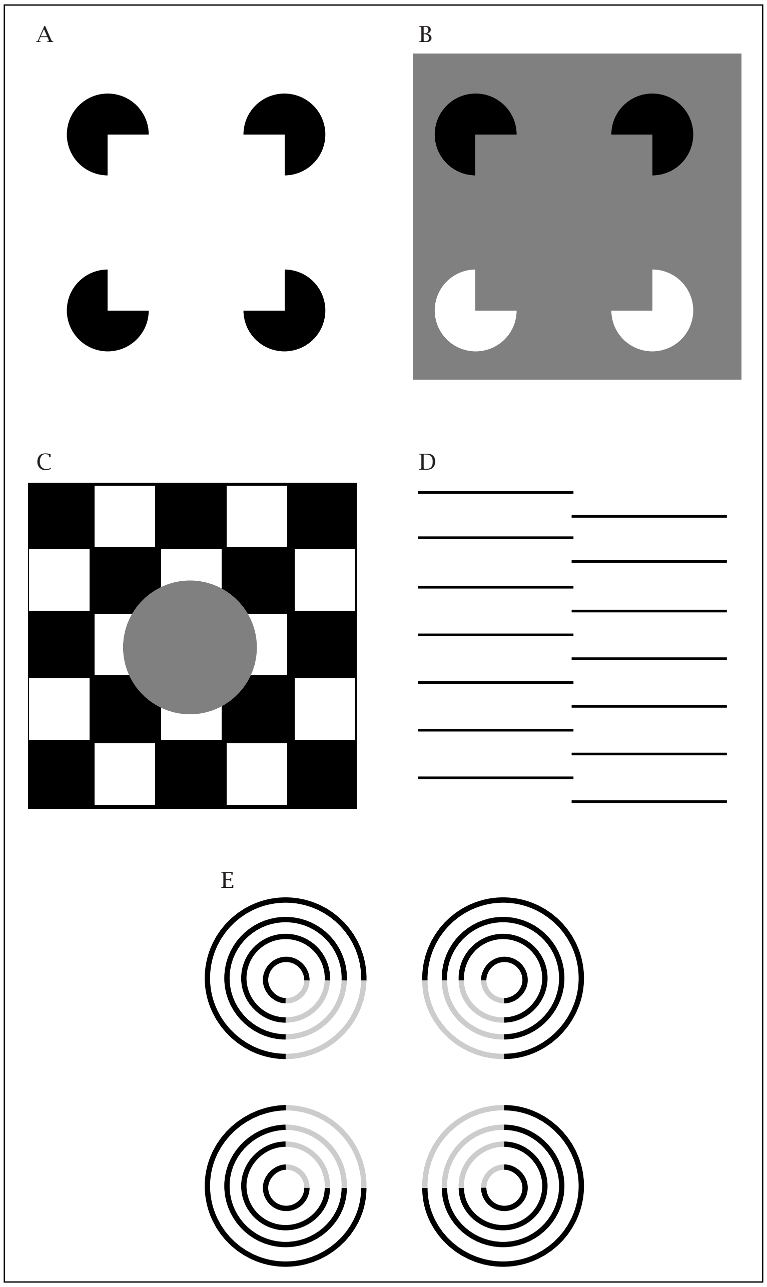

Figure 2. Visual boundary and surface interactions. The five diagrams show how boundaries can be formed in the mind in various circumstances even though they do not appear on the printed page. In (A), (C), and (E) one sees the bounded surface as well as the boundary. In (B) and (D) one can recognize a square or vertical boundary even though it cannot be seen. FACADE theory provides an explanation for what can be perceived in these diagrams.

Boundary Completion	Surface Filling-In
oriented	unoriented
inward	outward
insensitive to contrast polarity	sensitive to contrast polarity

Table 1. A comparison of properties of the processes in the FACADE theory by which the mind fills in surfaces and completes boundaries. Surface filling-in is done by the "Blob" stream in Figure 1, and boundary completion is done by the "Interblob" stream.

Figure 2D illustrates another invisible boundary that can be consciously recognized.

FACADE theory, which is supported by a great deal of anatomical, neurophysiological, and psychophysical evidence, says that such a boundary formation process in the brain is indeed the mechanism whereby we perceive geometrical objects such as lines, curves, and textured objects. Rather than being defined in terms of such classical units as points and lines, these boundaries arise as a coherent pattern of excitatory and inhibitory signals across a mixed cooperative-competitive feedback network that is defined by a nonlinear dynamical system describing the cellular interactions from the retina through LGN and the V1 Interblob and V2 Interstripe areas [8]. In such a network spatially long-range excitatory, or cooperative, interactions try to build the boundaries across space while interacting with shorter-range inhibitory, or competitive, interactions that suppress incorrect boundary groupings. These interactions select the best boundary grouping from among many possible interpretations of a scene. The winning grouping is represented either by an equilibrium point or a synchronous oscillation of the system, depending on how system parameters are chosen. The axioms of classical geometry provided one way to understand geometrical properties of objects in the world. FACADE theory suggests how the brain may actually represent these properties using nonlinear neural networks that do a type of online statistical inference to select and complete the statistically most-favored boundary groupings of a scene while suppressing noise and incorrect groupings. The emerging patterns of boundary excitation obey the three boundary completion properties (inward, oriented, insensitive) that are summarized in Table 1. Although there are global theorems about the stable convergence of competitive and cooperative dynamical systems (e.g., in [3] and Hirsch [10]), there are still no global theorems about the mixed cooperative-competitive dynamical systems that are used to group together distributed information in the brain.

Smale has shown [12] that an arbitrary n-dimensional system of autonomous first-order ODEs can be embedded into an $(n+l)$-dimensional *competitive* system of ODEs. An n-dimensional *autonomous* system in this sense is a system of the form

$$(1) \qquad dx_i/dt = f_i(x), \quad \text{where } x = (x_1, \dots, x_n).$$

Such a system is said to be *competitive* if

$$(2) \qquad \partial f_i / \partial x_j \le 0 \quad \text{for } i \ne j.$$

Thus, increasing the activity x_j of one cell population can only decrease the rate of growth of activity in other cell populations x_i with which it interacts. Interactions within a population can, however, enhance that population's activity. (For a *cooperative* system, the inequalities are reversed.) Thus even the class of competitive systems is too large to be meaningfully classified. The known theorems about cooperative and competitive neural networks, and those conjectured but yet to be proved, characterize and classify that subset of cooperative-competitive networks that appear to have been selected by biology to perform useful tasks, like grouping visual information or deciding which populations' activities should be stored in memory. For example, it has been proved under weak conditions on the functions $a_i(x_i)$, $b_i(x_i)$, and $c(x)$ that every trajectory of n-dimensional competitive systems of the form

$$(3) \qquad dx_i/dt = a_i(x_i)[b_i(x_i) - c(x)]$$

converges to one of possibly infinitely many equilibrium points. The proof of this theorem is based on the intuitive idea that one can analyze the global dynamics of a competitive system by keeping track of which population x_i is winning the competition at any time. This method is made mathematically precise by showing that "every competitive system induces a decision scheme" and by tracking the winning decisions through time while using a Lyapunov functional to determine whether system trajectories will converge or oscillate in prescribed ways. This particular theorem grew out of an analysis of how input patterns are transformed before they are stored in short-term memory, which is the type of memory that enables one to remember a new telephone number for a short time. A perceptual grouping is also a type of short-term memory, but it uses a more complicated dynamical system. Even the simple system (3) has many applications. For example, it can be used to provide sufficient conditions under which a competitive economic market will lead to a stable market price and to balanced books in all the competing firms (see [2], Ch. 2). A significant fraction of all brain processes use cooperative and competitive systems of one sort or another, so the project of classifying them mathematically is of great importance.

Surfaces Are for Seeing

If boundaries are invisible, then how do we see anything? FACADE theory predicts that visible properties of a scene are represented by a surface processing stream, which is predicted to occur within the Blob cortical stream (the left-hand stream in Figure 1). A key step in representing a visible surface is called *filling-in*. What is filling-in, and why and how does it occur? An early stage of surface processing compensates for variable illumination, or *discounts the illuminant*, in order to prevent illuminant variations, which can change from moment to moment, from distorting all percepts. Discounting the illuminant attenuates color and brightness signals except near regions of sufficiently rapid surface change, such as edges or texture gradients, which are relatively uncontaminated by illuminant variations. Later stages of surface formation fill in the attenuated regions with these relatively uncontaminated color and brightness signals. This is the main reason for filling-in. Remarkably, the same process can also allocate brightness and color signals to their perceived depths on a 3-dimensional surface through a process called *surface capture*, whereby the boundaries formed within the V2 Interstripes interact with the V2 Thin Stripes and area V4 (see Figure 1) to trigger depth-selective filling-in processes there. This multistage filling-in process is an example of hierarchical resolution of uncertainty, because the later filling-in stage overcomes uncertainties about brightness and color that were caused by discounting the illuminant at an earlier processing stage.

How do the illuminant-discounted signals fill in an entire region? Filling-in behaves like a diffusion of brightness across space. For an example, consider the percept of *neon color spreading* that is elicited by Figure 2E. This figure consists of circular annuli, part of which are black and part gray. In response to this figure, we can see an illusory square filled with a gray or white color. FACADE theory suggests that this percept is due to an interaction between the boundary and surface systems. In particular, the black boundaries cause small breaks in the gray boundaries where they join; see [4] for further discussion of how this happens. The gray color can thus spread through these breaks from the annuli into the illusory square. In this percept, filling-in spreads *outwardly* from the individual gray inducers in all directions. Its spread is thus *unoriented*. How is this spread of activation contained? FACADE theory predicts that signals from the boundary stream to the surface stream define the regions within which filling-in is restricted. These boundaries surround the annuli (except for their small breaks) and also form the square illusory contour. Thus, filling-in is a form of anisotropic diffusion in which boundary signals nonlinearly gate, or inhibit, the diffusive flow of signal. Without these boundary signals, filling-in would dissipate across space, and no surface percept could form. Invisible boundaries thereby indirectly assure their own visibility through their interactions with the surface stream, within which all visible percepts are predicted to form.

With these comments in mind we can better understand finer aspects of the other percepts that form in response to the images in Figure 2. In Figure 2A the square boundary is induced by four black Pac-Men that are all less luminant than the white background. In the surface stream, discounting the illuminant causes these Pac-Men to induce local brightness enhancements adjacent to the Pac-Men just within the boundary of the square. At a subsequent processing stage, these enhanced brightness signals diffuse within the square boundary, thereby causing the entire interior of the square to look brighter. The filled-in square is visible because the filled-in activity level within the square is higher than the filled-in activity of the surrounding region. Filling-in can in this way lead to visible percepts because it is *sensitive* to contrast polarity. These three properties of surface filling-in (outward, unoriented, sensitive) are summarized in Table 1. They are easily seen to be complementary to the corresponding properties of boundary completion.

In Figure 2B the opposite polarities of the two pairs of Pac-Men with respect to the gray background lead to approximately equal filled-in activities inside and outside the square, so the boundary can be recognized but not seen. In Figure 2D the white background can fill in uniformly on both sides of the vertical boundary by diffusing around the horizontal black lines, so no visible contrast difference is seen.

These remarks just begin the analysis of filling-in. Even in the seemingly simple case of the Kanizsa square, one often perceives a square hovering in front of four partially occluded circular disks, which seem to be completed behind the square even though they are invisible there. FACADE theory predicts how surface filling-in is organized to help such figure-ground percepts to occur in response to both two-dimensional pictures and three-dimensional scenes; the papers [4] and [7] give examples.

In summary, boundary and surface formation illustrate two key principles of brain organization: hierarchical resolution of uncertainty and complementary interstream interactions. Table 1 summarizes three pairs of complementary properties of the boundary and surface streams. Hierarchical resolution of uncertainty is illustrated by surface filling-in: Discounting the illuminant creates uncertainty by suppressing surface color and brightness signals except near surface discontinuities. Higher stages of filling-in complete the surface representation using properties that are complementary to those whereby boundaries are formed, guided by signals from these boundaries.

Boundary-gated surface filling-in is a radically different view of how a surface is formed compared with the classical geometrical view in terms of surface normals or differential forms. The mathematical analysis of this kind of anisotropic diffusion has hardly begun, even though its remarkable properties have already been successfully used in processing complex imagery in technology (Waxman et al. [13]). Another important problem on which a great deal of work remains to be done concerns the origin of the complementarity of boundaries and surfaces. I predict that this property arises through a process of global symmetry-breaking as the embryonic brain bifurcates into its parallel cortical processing streams.

The Link between Learning, Expectation, Attention, and Resonance

Visual and auditory perception have developed into large and multifaceted fields during the past century, at least since the time of Helmholtz. As we ascend higher into the brain, perceptually preprocessed information engages higher cognitive, spatial, and motor processes. Learning occurs in all of these types of processes. As soon as sensory and cognitive learning are considered, a formidable difficulty must be faced: namely, we can learn very quickly about the world, but then why do we not also forget everything that we have previously learned just as quickly? Neural modeling has clarified, as sketched below, how sensory and cognitive processes solve a key problem, called the *stability-plasticity dilemma* [2, 6], whereby the brain can rapidly learn about the world throughout life without catastrophically forgetting our previous experiences. In other words, we remain *plastic* and open to new experiences without risking the *stability* of previously learned memories. This type of fast stable learning enables us to become experts at dealing with changing environmental conditions: Old knowledge representations can be refined by changing contingencies and new ones built up without destroying the old ones due to catastrophic forgetting.

On the other hand, catastrophic forgetting is a *good* property for spatial and motor learning. We have no need to remember all the spatial and motor representations that we used when we were children. In fact, the parameters that controlled our small childhood limbs in space would cause major problems if they continued to control our larger and stronger adult limbs.

It turns out that cognitive learning and motor learning are also realized within parallel processing streams, often called the What and Where streams, that run through the inferotemporal cortex and parietal cortex, respectively; see Figure 1. The inferotemporal cortex learns how to recognize objects and events in the world, namely, What they are. The parietal cortex learns how to localize objects spatially and direct actions towards them, namely, Where they are and How to physically engage them. These distinct What and Where memory properties are proposed to follow from another set of complementary mechanisms, namely, mechanisms whereby these systems *learn* expectations about the world and *match* these expectations against world data. These complementary mechanisms are also predicted to arise through a process of symmetry-breaking during brain development. The present discussion will restrict itself to sensory and cognitive learning, both due to its intrinsic importance and to the fact that a great deal of useful mathematical work remains to be done in this area. Notable work remaining to be done is to classify the resonant states that drive sensory and cognitive learning and that are predicted to support all conscious experiences.

To see how we use a sensory or cognitive expectation and how a resonant state is activated, suppose a person were asked to "find the yellow ball within one-half second, and you will win a $10,000 prize." Activating an expectation of "yellow balls" enables more rapid detection of a yellow ball, with a more energetic neural response, than if one were not looking for it. Sensory and cognitive top-down expectations lead to *excitatory matching* with confirmatory bottom-up data. On the other hand, mismatch between top-down expectations and bottom-up data can suppress the mismatched part of the bottom-up data and thereby start to focus attention upon the matched, or expected, part of the bottom-up data.

This sort of excitatory matching and attentional focusing on bottom-up data using top-down expectations is proposed to generate resonant brain states: When there is a good enough match between bottom-up and top-down signal patterns between two or more levels of processing, their positive feedback signals amplify and prolong their mutual activation, leading to a resonant state. The amplification and prolongation of the system's fast activations is sufficient to trigger learning in the more slowly varying "adaptive weights" that control the signal flow along pathways from cell to cell. Resonance provides a global context-sensitive indicator that the system is processing data worthy of learning. That is why the theory that describes these processes is called *Adaptive* Resonance Theory, or ART.

ART predicts that "all conscious states are resonant states" of the brain. Paradoxical data about conscious perceptual experiences from several modalities have been explained [6] as emergent properties of such resonant states. The mathematical analysis of such resonant states is thus of great importance. When one considers that quantum theory is, at bottom, a resonance theory for which much beautiful mathematics has been created, the fact that the mind also seems to be a resonance machine presents a great opportunity to interested mathematicians. More is said below about how such a resonance develops.

In summary, ART predicts that there is an intimate connection between the mechanisms that enable us to learn quickly and stably about a changing world and the mechanisms that enable us to learn expectations about such a world, test hypotheses about it, and focus attention upon information that we find interesting. ART also proposes that, in order to solve the stability-plasticity dilemma, only resonant states can drive new learning, and this characteristic gives the theory its name.

How Are Learning and Memory Search Related?

Learning within the sensory and cognitive domain is often *match learning*. Match learning occurs only if a good enough match occurs between bottom-up information and a learned top-down expectation that is read out by an active *recognition category*, or *code*. When such an approximate match occurs, previously learned knowledge can be refined. If novel information cannot form a good enough match with the expectations that are read out by previously learned recognition categories, then a memory search, or hypothesis testing, is triggered that leads to selection and learning of a new recognition category rather than catastrophic forgetting of an old one. Figure 3 illustrates how this happens in an ART model; the figure will be discussed in greater detail below. In contrast, learning within spatial and motor processes is proposed to be *mismatch learning* that continuously updates sensory-motor maps or the gains of sensory-motor commands. As a result, we can stably learn what is happening in a changing world, thereby solving the stability-plasticity dilemma, while adaptively updating our representations of where objects are and how to act upon them using bodies whose parameters change continuously through time.

It has been mathematically proved in the context of an ART model that match learning leads to stable memories in response to an arbitrary list of events to be learned [2]. Match learning has a serious potential weakness, however: If one can learn only when there is a good enough match between bottom-up data and learned top-down expectations, then how does one ever learn anything really new? ART proposes that this problem is solved by the brain by using another complementary interaction, this one between processes of resonance and reset, which are predicted to control properties of attention and memory search, respectively. These complementary processes help our brains to balance between the complementary demands of processing the familiar and the unfamiliar, the expected and the unexpected. One of these complementary processes takes place in the What cortical stream that was described above, notably in the inferotemporal and prefrontal cortex (Figure 1). It is here that top-down expectations are matched against bottom-up inputs. When a top-down expectation achieves a good enough match with bottom-up data, this match process focuses attention upon those feature clusters in the bottom-up input that are expected. If the expectation is close enough to the input pattern, then a state of resonance develops as the attentional focus takes hold.

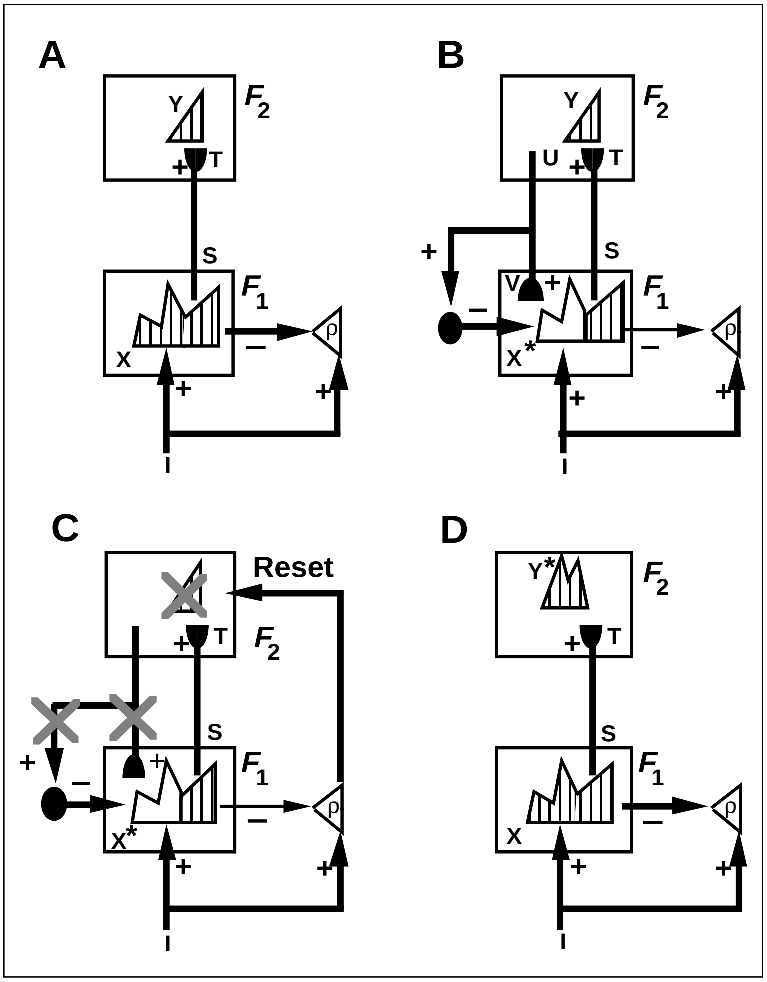

Figure 3. Search for a recognition code within an ART learning circuit. The four diagrams illustrate different aspects of how ART proposes that the brain attempts to match an input pattern, coming from the bottom, with a memory pattern, via a feedback circuit. A match results in a resonant circuit. The diagrams indicate four stages in the search for a correct recognition category at level F2 with which to classify a pattern of activation across the feature detectors at level F1. The hatched areas represent activity levels over a set of cells; the white areas and crosses represent previously active cells and pathways that have been actively inhibited during the search cycle.

Figure 3 illustrates these ART ideas in a simple two-level example. Here, a bottom-up input pattern, or vector, I activates a pattern X of activity across the feature detectors of the first level F_1. For example, a visual scene may be represented by the features comprising its boundary and surface representations. This feature pattern represents the relative importance of different features in the input pattern I. In Figure 3A the pattern peaks represent more activated feature detector cells, the troughs less activated feature detectors. This feature pattern sends signals S through an adaptive filter to the second level F_2, at which a compressed representation Y (also called a *recognition category*, or a *symbol*) is activated in response to the distributed input T. Input T is computed by multiplying the signal vector S by a matrix of adaptive weights that can be altered through learning. The representation Y is compressed by competitive interactions across F_2 that allow only a small subset of its most strongly activated cells to remain active in response to T. The pattern Y in the figure indicates that a small number of category cells may be activated to different degrees. These category cells, in turn, send top-down signals U to F_1. The vector U is converted into the top-down expectation V by being multiplied by another matrix of adaptive weights. When V is received by F_1, a matching process takes place between the input vector I and V that selects that subset X* of F_1 features that were "expected" by the active F_2 category Y. The set of these selected features is the emerging "attentional focus".

If the top-down expectation is close enough to the bottom-up input pattern, then the pattern X* of attended features reactivates the category Y, which, in turn, reactivates X*. The network then locks into a resonant state through a positive feedback loop that dynamically links, or binds, the attended features across X* with their category, or symbol, Y. The individual features at F_1 have no meaning on their own, in the same way that the pixels in a picture are meaningless one-by-one. The category, or symbol, in F_2 is sensitive to the global patterning of these features, but it cannot represent the "contents" of the experience—e.g., the boundaries and surfaces in a picture—because of the very fact that it is a compressed representation. The resonance between these two types of information converts the *pattern* of attended features into a coherent context-sensitive state that can enter consciousness. In particular, such a resonance binds spatially distributed features into either a stable equilibrium or a synchronous oscillation. Such synchronous oscillations have recently attracted much interest after being reported in neurophysiological experiments. The mathematical analysis of such fast-synchronizing networks has just begun ([9] and Somers-Kopell [11]).

In ART, the resonant state, rather than bottom-up activation, is predicted to drive the learning process. The resonant state persists long enough, and at a high enough activity level, to activate the slower learning processes in the adaptive weights that guide the flow of signals between bottom-up and top-down pathways between levels F_1 and F_2. Thus ART-based learning naturally suggests the use of singular perturbation theory, where the fast events are resonating synchronous oscillations and the slow events are learned changes in adaptive weight matrices. This viewpoint helps to explain how adaptive weights that were changed through previous learning can regulate the brain's present information processing without learning about the signals that they are currently processing unless they can initiate a resonant state. Through resonance as a mediating event, one can see from a deeper mathematical viewpoint why humans are "intentional" beings who are continually predicting what may next occur and why we tend to learn about the events to which we "pay attention".

How does a sufficiently bad mismatch between an active top-down expectation and a bottom-up input drive a memory search, say, because the input represents an unfamiliar type of experience? This mismatch within the attentional system is proposed to activate a complementary *orienting system*, which is sensitive to unexpected and unfamiliar events. ART suggests that this orienting system includes the brain region that is called the hippocampus. Output signals from the orienting system rapidly reset the recognition category that has been reading out the poorly matching top-down expectation (Figures 3B and 3C). The cause of the mismatch is thus removed, thereby freeing the system to activate a different recognition category (Figure 3D). The reset event then triggers memory search, or hypothesis testing, which automatically leads to the selection of a recognition category that can better match the input.

If no such recognition category exists—say, because the bottom-up input represents a truly novel experience—then the search process automatically activates an as yet uncommitted population of cells with which to learn about the novel information. This learning process works well under both unsupervised and supervised conditions. Unsupervised learning means that the system can learn how to categorize novel input patterns without any external feedback. Supervised learning lets the system know whether it has categorized the information correctly. Supervision can force a search for new categories that may be culturally determined and are not based on feature similarity alone. For example, separating the letters E and F into separate recognition categories is culturally determined; they are quite similar based on visual similarity alone. In this case, if the input pattern directly represented the pixels of

E and F (which it typically would not), then both E and F might be classified in the same category with category prototype F unless supervised feedback indicated that each pattern needed its own category and category prototype. Taken together, the interacting processes of attentive-learning and orienting-search realize in this way a type of error correction through hypothesis testing that can build an ever-growing, self-refining internal model of a changing world.

A number of global theorems have been proved about how ART learning takes place in specific model systems, including global theorems about how an ART system can stably learn to categorize an arbitrary list of binary or analog input vectors and including how the adaptive weights oscillate, what their equilibrium values are, and how many trials it takes to stabilize learning (see [2] and references 108, 187, 127, 157, 177, and 181 in http://www.cns.bu.edu/Profiles/Grossberg/). Just to give the flavor of these results, here is a sample theorem (ref. 187) about "stable category learning". It represents a synthesis of ideas from neural networks, fuzzy logic, and expert production systems in artificial intelligence. We shall leave some of the terms undefined.

Theorem. *In response to an arbitrary sequence of n-dimensional input vectors, a "Fuzzy ART system" with "complement coding" and "fast learning" can form stable n-dimensional rectangular categories R_j into which the input vectors are partitioned. These rectangles grow during learning to a maximum size $|R_j| \le M(1-\rho)$, where ρ is a "vigilance" parameter that determines how much difference there can be among vectors that are clustered in the same category. In addition, the sizes $|w_j|$ of the adaptive weights that define the category rectangles decrease monotonically with time. In the "conservative limit", one presentation of the input vectors leads to a stable learned classification such that no reset or additional learning occurs on subsequent presentations of any input.*

The theorem asserts that stable learning of an arbitrary list of input vectors can occur on a single learning trial and can form rectangular categories within which to cluster input vectors whose maximal size (and thus coarseness) can be controlled by the vigilance parameter ρ, which is defined within the orienting system of Figure 3.

A large amount of mathematical work remains to be done on proving such learning theorems, particularly towards proving how optimal learning occurs when the input data are noisy, probabilistically defined, and/or self-contradictory (e.g., medical, remote sensing, and World Wide Web databases) and how *distributed* categories are learned whose individual cells, or nodes, can be part of several different categories (see Carpenter [1]). ART models are already being used in many technological applications because of their ability to learn quickly and stably in real-time to categorize and to predict large amounts of information in a rapidly changing world (see [1] for an illustrative list of applications). The vitality of the field may also be seen from the fact that every extension of the mathematical understanding of ART models has led to a corresponding increase in the range of important applications.

Why Does the Cerebral Cortex Have Layers?

How are these ART top-down matching rules actually implemented in the cerebral cortex of the brain? An answer to this question has been recently proposed as part of a rapidly developing theory of why the cerebral cortex is typically organized into six distinct layers of cells [5]. This work proposes an answer to the general question: How does "laminar computing" contribute to biological intelligence? The proposed answer suggests how these layers support circuits that simultaneously realize three types of general properties: (1) self-stabilizing development and learning, i.e., a solution of the "stability-plasticity" problem; (2) seamless fusion of bottom-up automatic processing of information and top-down attentional modulation of information processing based on system goals; and (3) grouping of distributed information into coherent representations that preserve their sensitivity to analog properties of the information, that is, the property of "analog coherence". Properties (1) and (2) suggest how ART mechanisms are instantiated within the known laminar circuits of visual cortex, notably between cortical area V1 and cortical area V2 (Figure 1), and by extension in other sensory and cognitive neocortical circuits. This model is called the LAMINART model because it shows how ART mechanisms are embedded within the laminar circuits of the neocortex. Figure 4 schematizes some of the key LAMINART circuits, all of which are known to occur in the brain.

For present purposes I just want to summarize how the ART matching rule is realized within these laminar circuits. Earlier mathematical work had predicted that such a matching rule would be realized by a *modulatory top-down on-center off-surround network*; see [2] and [6] for reviews. Figure 4 shows how such a circuit may be realized in the cortex and what it is. In Figure 4 the top-down circuit generates outputs from cortical layer 6 of V2 that activate layer 6 of V1 via the vertical pathway between these layers that ends in an open triangle (which designates an excitatory connection). Cells in layer 6 of V1, in turn, activate an "on-center off-surround" circuit to layer 4 of V1. In this circuit an excitatory cell (open circle) in layer 6 excites the excitatory cell (open circle) immediately above it in layer 4 via the vertical pathway from layer 6

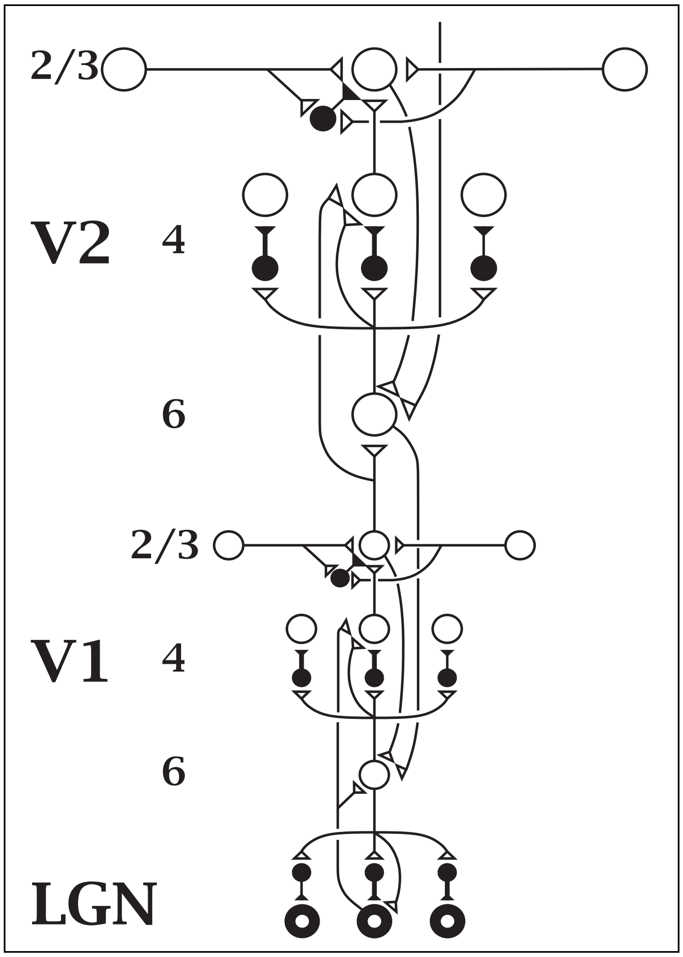

Figure 4. The LAMINART model. The model is a synthesis of feedforward (or bottom-up), feedback (or top-down), and horizontal interactions within and between the LGN and visual cortical areas V1 and V2. Cells and connections with open symbols indicate excitatory interactions, and closed symbols indicate inhibitory interactions.

to 4 that ends in an open triangle. This excitatory interaction constitutes the "on-center". The same excitatory cell in layer 6 also excites nearby inhibitory cells (closed black circles), which, in turn, inhibit cells in layer 4. This spatially distributed inhibition constitutes the "off-surround" of the layer 6 cell. The on-center is predicted to have a modulatory, or sensitizing, effect on layer 4 because of the balancing of excitatory and inhibitory inputs to layer 4 within the on-center. The inhibitory signals in the off-surround can strongly suppress unattended visual features. This arrangement clarifies how top-down attention can sensitize the brain to get ready for expected information that may or may not actually occur without actively firing the sensitized target cells and thereby inadvertently creating hallucinations that the information is already there.

Within the cortex such a top-down circuit is realized by a type of *folded feedback*, whereby feedback inputs from V2 are "folded" back into the feedforward flow of information from layer 6-to-4 of V1. This type of folded feedback has many useful properties that are explained in the original articles [5] and [8]. For present purposes I just want to note that these laminar cortical circuits integrate ART properties of bottom-up adaptive filtering and top-down attentive expectation learning with boundary-grouping properties that are carried out by long-range horizontal connections in layer 2/3. These cortical circuits are marvels of compactness and parsimony that are already starting to get designed into VLSI chips. Their description and mathematical characterization brings us to the threshold of understanding even the cerebral cortex of the brain, that enchanted loom on which so many of our most meaningful experiences, including our mathematical theorems, are played out throughout our lives.

References

[1] G. A. Carpenter, Distributed learning, recognition, and prediction by ART and ARTMAP neural networks, *Neural Networks* **10** (1997), 1473–1494.

[2] G. A. Carpenter and S. Grossberg, *Pattern Recognition by Self-Organizing Neural Networks*, MIT Press, Cambridge, MA, 1991.

[3] S. Grossberg, *Studies of Mind and Brain*, Kluwer/Reidel, Amsterdam, 1982.

[4] ____, 3-D vision and figure-ground separation by visual cortex, *Perception & Psychophysics* **55** (1994), 48–120.

[5] ____, How does the cerebral cortex work? Learning, attention, and grouping by the laminar circuits of visual cortex, *Spatial Vision* **12** (1999), 163–186.

[6] ____, The link between brain learning, attention, and consciousness, *Consciousness and Cognition* **8** (1999), 1–44.

[7] S. Grossberg and N. McLoughlin, Cortical dynamics of 3-D surface perception: Binocular and half-occluded scenic images, *Neural Networks* **10** (1997), 1583–1605.

[8] S. Grossberg and R. D. S. Raizada, Contrast-sensitive perceptual grouping and object-based attention in the laminar circuits of primary visual cortex, *Vision Research* **40** (2000), 1413–1432.

[9] S. Grossberg and D. Somers, Synchronized oscillations during cooperative feature linking in a cortical model of visual perception, *Neural Networks* **4** (1991), 453–466.

[10] M. W. Hirsch, Systems of differential equations that are competitive or cooperative, II: Convergence almost everywhere, *SIAM J. Math. Anal.* **16** (1985), 423–439.

[11] C. Somers and N. Kopell, Waves and synchrony in arrays of oscillators of relaxation and non-relaxation type, *Physica D* **89** (1995), 169–183.

[12] S. Smale, On the differential equations of species in competition, *J. Math. Biol.* **3** (1976), 5–7.

[13] A. M. Waxman, M. C. Seibert, A. Gove, D. A. Fay, A. M. Bernardon, C. Lazott, W. R. Steele, and R. K. Cunningham, Neural processing of targets in visible, multispectral IR and SAR imagery, *Neural Networks* **8** (1995), 1029–1051.