Cognitive Neuroscience

DOMINIC STANDAGE[1] AND THOMAS TRAPPENBERG[2]

[1]Centre for Neuroscience Studies, Queen's University, Kingston, Canada

[2]Faculty of Computer Science, Dalhousie University, Halifax, Canada

standage@biomed.queensu.ca | tt@cs.dal.ca

1 Introduction

The field of cognitive neuroscience addresses the question *How does the brain think?* . Although simplistic, this short definition captures the primary characteristics of the field. The word *how* addresses the physical mechanisms underlying our thought processes, the word *brain* stresses that our explanations must be rooted in structures and functions of the brain, and the word *think* addresses the high level, systemic nature of the processes under study. By its common usage, thinking refers to things like problem solving, decision-making and the recall of personal memories.

To cognitive neuroscientists, providing a mechanistic account of cognitive phenomena means linking experimental results obtained at the levels of neuroscience and behavior. Crucially, these mechanisms must be rooted in brain function. We may interpolate between physiological and behavioral data at a number of levels of abstraction, but in all cases, the mechanisms proposed must have structural or functional correlates with the brain. Below, we describe a number of approaches to this research goal. In all cases, our usage of the term *mechanism* refers to brain-related processes. Our examples show that

methods in cognitive neuroscience make different assumptions and focus on different levels of abstraction, but they all have this fundamental trait in common. Most examples in this chapter reflect our computational background, but other approaches fall under the umbrella of cognitive neuroscience if they mechanistically link cognitive function and brain activity.

Despite much recent interest and the popularization of several prominent methods, cognitive neuroscience is not as new as one might think. The term *cognitive neuroscience* was coined in the late 1970's, but the field has its origins in the work of philosophers and psychologists considering the 'mind-brain' problem. The first cognitive neuroscientist was arguably Franz Joseph Gall, who believed that the human brain was compartmentalized according to psychological function. Around 1800, Gall proposed that personality traits were localized in the cortex and that the most active locations, reflecting a person's dominant personality traits, should grow to be the most prominent. His doctrine of phrenology claimed that the growth of these cortical 'organs' lead to bumps on the skull, so that a person's personality could be determined by the shape of his or her head. With the benefit of hindsight, phrenology seems at best quaint and at worst absurd, but it may represent the first attempt to address the physical underpinnings of behavior.

Cognitive neuroscience may be roughly divided into four subfields: clinical studies of neurological patients, non-invasive brain imaging methods, electrophysiology, and computational modeling. The first of these approaches is by far the oldest. Perhaps the best example of early lesion studies of cognitive phenomena comes from the work of neurologist Paul Broca in 1861. Broca discovered a clinical patient who could make only one verbal sound, but whose language comprehension was unaffected. Following the patient's death, damage was found to his left frontalparietal lobe, now known as Broca's area. Several years later, neurologist Carl Wernicke discovered a stroke victim who made grammatically well-formed utterances that were seemingly devoid of meaningful content. After the patient's death, a lesion was discovered along the border of his parietal and temporal lobes, now known as Wernicke's area. Together,

these clinical cases showed that the cognitive phenomenon of language was mediated by locally specific cortical regions, contributing to different aspects of speech generation and comprehension.

Lesion studies have provided a wealth of scientific data showing the cortical and sub-cortical localization of various aspects of cognitive phenomena. Clinical methods face several challenges, however, including the requirement of patients with similar lesions. Slight differences in the location of brain lesions can lead to major differences in cognitive performance. More recently, neuroimaging techniques have provided a powerful means of studying the respective functions of brain structures, including electroencephalogram (EEG) recordings, positron emission tomography (PET), and functional magnetic resonance imaging (fMRI). The non-invasiveness of these methods enables researchers to record brain activity in healthy, behaving human subjects. Electrophysiological methods, such as single- and multi-electrode recordings, provide much higher resolution, albeit with non-human animal subjects. Clearly, these approaches are highly complementary.

In particular, fMRI has made a huge impact in cognitive neuroscience in recent years, where the blood oxygen level dependent (BOLD) signal reveals brain structures correlated with the performance of cognitive tasks. Note the reference here to measurements of brain activity and behavior. Without data, there can be no cognitive neuroscience. On their own, though, data are not enough. We must have a theory to link brain activity with behavioral measurements. In the case of fMRI, the BOLD signal tells us that a region of the brain contributes to a cognitive function, but *how* does this region contribute to that function? Computational models are a powerful way to address this question. These models are mathematical descriptions of brain systems, programmed on a computer. Here, a mathematical description is necessary to make quantifiable hypotheses, as interactions between subsystems can lead to complex and sometimes unexpected system behaviour. The role of models in identifying the causal relationship between brain activity and behavior is the central theme of this chapter and is depicted in Figure 1.

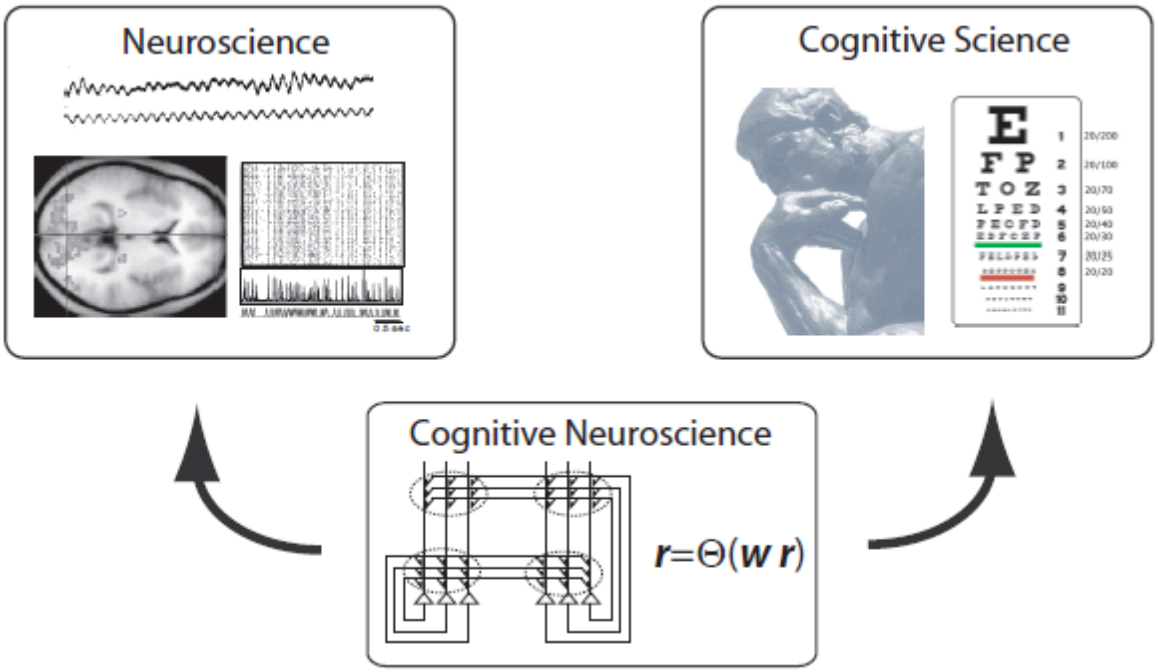

Figure 1: Cognitive neuroscience links activity in the brain with cognitive function.

We have chosen three cognitive phenomena for discussion: episodic memory, decision-making, and category learning. We believe these three phenomena exemplify different challenges to cognitive neuroscientists and their methods, facilitating discussion of a wide spectrum of computational, cognitive research. We also believe they provide a well-rounded chronological and historical perspective. Episodic memory has been the subject of research interest for the longest of the three, but is arguably the most poorly understood. Research on decision theory in many ways exemplifies current computational cognitive methods, and research on category learning exemplifies a budding research area within the computational, cognitive community. Episodic memory is poorly understood for good reasons, and we

discuss this cognitive phenomenon first to illustrate the sorts of challenges faced by cognitive neuroscientists. We believe that decision theory is in many ways more tractable, both experimentally and computationally. This belief in no way suggests that episodic memory is ‘too hard’ or that research methods in this field are misguided. Decision-making processes simply provide a more direct example of how computational methods can link physiology and behavior. Finally, cortical models of category learning provide an example of an exciting, rapidly expanding field. This field is by no means new, but is receiving long-overdue interest and recognition.

2 Computational modeling of cognitive phenomena

Before presenting some examples of cognitive neuroscience research, we introduce some of our methods. We have stated that the aim of cognitive neuroscience is to bridge brain activity and cognitive function with structural or mechanistic properties of the brain. As computationalists, our approach is to devise models that touch both sides of this gap. Our models of behavior must therefore be grounded in structures or activity found in the brain. A challenge raised by this requirement stems from the many scales of these structures and the mechanisms they generate. Of course, this challenge facilitates methodological flexibility and diversity, and may rightly be considered a strength of the field. A range of structural scales is depicted in Figure 1 where examples span the molecular and personal levels. With this dilemma in mind, on which level should we describe the brain? A subatomic level is probably too detailed, but what about the neural level? Should our models address intracellular mechanisms?

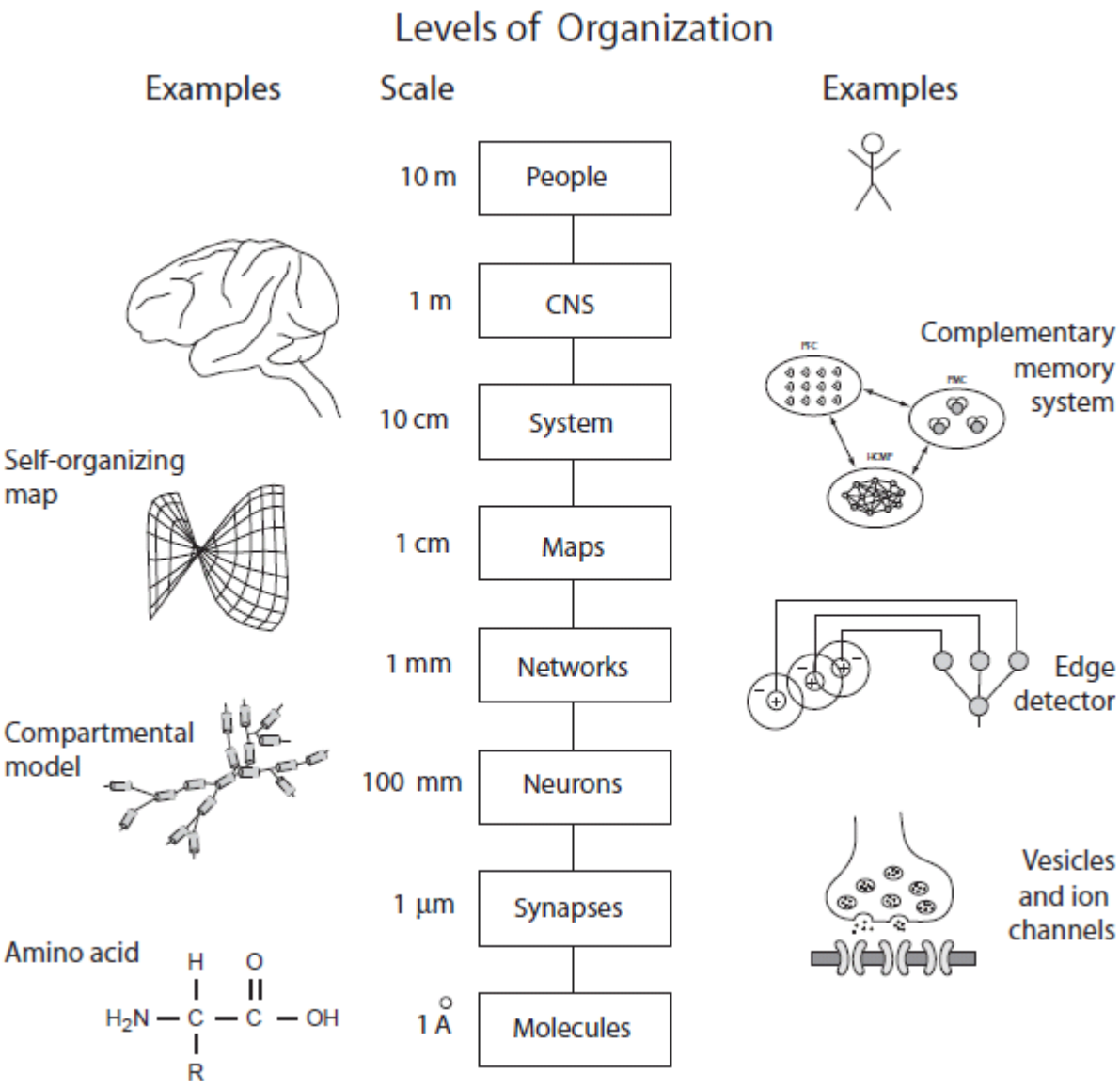

Figure 2: Structural scales of the brain and behavior.

The answer is that it depends on what we are studying and how we study it. While molecular effects can have direct consequences on behavior, we do not expect a single neuron to assume full responsibility for a behavioral phenomenon. Such a scenario would defy stability in neural information processing. We discuss several approaches to modeling in cognitive neuroscience at different levels of abstraction. There

are tradeoffs in these cases. For instance, we can record brain activity in non-human animals with neural resolution, but our interpretation of their behavior is highly constrained. Alternatively, the behavior of human subjects is much easier to assess, but resolution of data such as fMRI is lower. Finding an appropriate level of abstraction is a fundamental challenge to scientific enquiry. A theory must be simple enough to provide insight and constrained enough to make verifiable predictions. Cognitive neuroscience is consequently found in many forms, linking many levels of structure in the brain with the behavior of organisms.

In the following, we focus on abstractions underlying so-called *neural network* models. Models under this umbrella have long had behavioral correlates and have evolved into descriptions of brain structures and activity. We refer to the basic units in these models as nodes, as these units are not representative of neurons per se. As depicted in Figure 3A, each node receives input from potentially many sources, where each input channel has its own strength or *weight*. All inputs are multiplied by the weight of their respective channels and then summed together. The net input is run through a function (g in the figure) determining the output of the node.

If a node doesn't represent a neuron, what does it represent? As a mathematical construct, it can represent a lot of things. The most common interpretation in the present context is that it represents a *Hebbian cell assembly*, a collection of neurons involved at a specific stage of a cognitive task. The output of the node represents the population activity or *rate* of the neurons in the assembly. Later, we show a case where the rate represents the average spike frequency of a neuron within the population, but often, individual neurons in an assembly exhibit very different response characteristics, so the node only represents the population average. In these cases, the global activity of collections of nodes may be more appropriately compared to larger scale brain activity such as that revealed by fMRI.

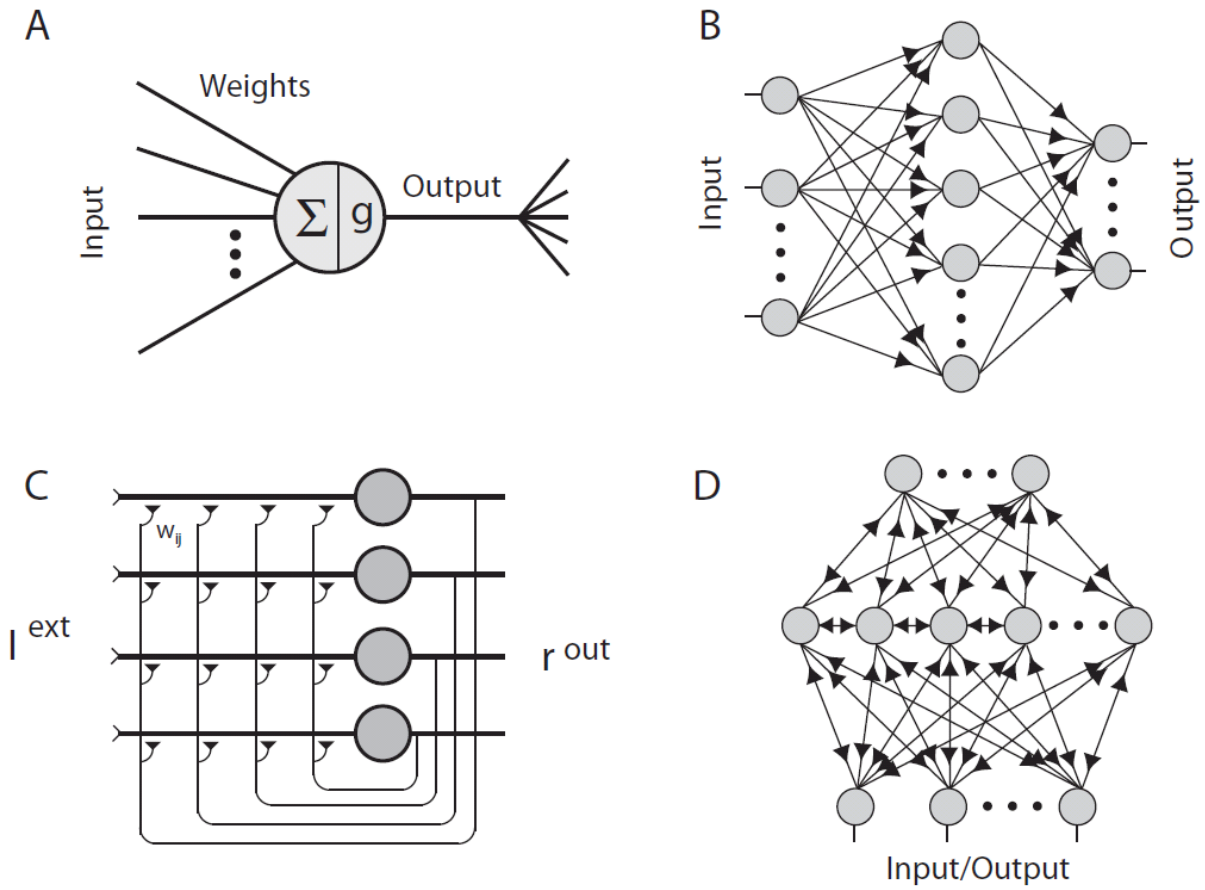

Figure 3: (A) In each node in a neural network, output is a function of the sum of weighted inputs. (B, C, D) Feedforward, recurrent and bidirectional architectures respectively.

Clearly, the computational abilities of nodes are very limited, but they can be very powerful in combination. Examples of neural network architectures are shown in Figure 3. In Figure 3B, activity is fed from left to right in a multi-layer *feedforward* network. These networks are often called multi-layer perceptrons and have been widely used in the connectionist movement in cognitive science, also known as parallel distributed processing (PDP). The fundamental difference between connectionist neural networks and those in cognitive neuroscience is that structural and/or mechanistic correlates with the brain are required of the latter. Their realization is left unspecified in PDP architectures. For example, the backpropagation algorithm (Rumelhart and McClelland 1986) is a common method to establish weight

values in neural networks like the one in Figure 3B. Such an algorithm is a practical way to specify weights to achieve many tasks, but only the final weights may be assigned biological meaning. That is, the final weights allow the network to behave in a biologically plausible manner, but the learning algorithm is biologically unrealistic. Additionally, connectionist models usually employ finely tuned, task-specific architectures, while a major trend in cognitive neuroscience models is to capture more general brain mechanisms and relate them to specific tasks. Feedforward architectures certainly have functional relevance to brain processes, but they represent just one of many architectures found in the brain.

A network architecture we discuss at length is shown in Figure 3C. In this *recurrent* network, the output from each node is fed back as input to all other nodes in the network[1]. Recurrent networks are formally dynamic systems and their behavior is well understood mathematically. For instance, recurrent networks can function as fast-learning, content addressable memory systems, where the network retrieves learned patterns from partial or noisy versions of these patterns. These memories are believed to be crucial to several forms of information processing in the brain, notably episodic memory. Recurrent networks dominate much of the discussion in this chapter.

Finally, we believe networks that model general brain functions are likely to be of a more general type. For instance, we expect the same type of network to mediate learning in sensory and association cortices. An example of such a general architecture is provided by Figure 3D. We believe this architecture captures one of the brain's fundamental information processing principles, as discussed in Section 5. As shown in the figure, this network has a hierarchical structure with bidirectional flow of information. The dynamic between bottom-up and top-down processing is an essential component of a number of cognitive phenomena and a major *modus operandi* of the brain. Through this and related mechanisms, we learn to represent the likelihood that sensory information matches learned expectations of the world, facilitating a flexible representational and predictive framework.

3 Episodic memory

Before considering episodic memory, or the memory of personal experience, we first consider memory formation at the (sub-cognitive) level of neurons and synapses. It is widely believed that memory is accomplished by the change in strength of synapses, or *synaptic plasticity*. Moreover, plasticity is believed to be activity-dependent. Consider two neurons *A* and *B*, where *A* synapses onto *B*. When *A* contributes to firing *B*, the synapse from *A* to *B* increases in strength. When *B* fires without *A*'s help, the synapse decreases in strength. This principle was first proposed by the great neuropsychologist Donald Hebb (Hebb 1949) and has long been established experimentally.

3.1 The SPM hypothesis and the need for models

While Hebbian plasticity is near-universally accepted as a fundamental learning mechanism, direct evidence is a difficult proposition. Martin et al. (2000) suggest several criteria for establishing the synaptic plasticity and memory (SPM) hypothesis. If an animal displays memory for some experience, synaptic change should be detectable. If so, then imposing these changes without the actual experience should lead to the same memory. Conversely, preventing the synapses from changing should prevent the memory from being formed. These are very difficult criteria to establish experimentally, not least because of the immense difficulty in identifying the synapses participating in learned representations, the difficulty of measuring these synapses individually, and the complicated timescales and stages of learning and plasticity. Some studies show compelling evidence in support of the SPM hypothesis, but the difficulty of providing direct evidence for this theory is foreboding of the sorts of challenges facing more cognitive-level research and highlights the usefulness of computational methods. Models allow us to explore the system-level consequences of more detailed experimental observations. For example, neural

network models of memory have long used Hebbian rules to demonstrate the SPM hypothesis. We come back to this point briefly in Section 5.

Moving from the level of neurons and synapses to the levels of systems of neuronal networks is a huge step up. What would constitute a definitive test of a theory about episodic memory? We'd need to precisely identify the brain structures responsible for representations of personal experience and monitor these representations as they progress through a series of transformations. Theories abound as to the structures and representational forms involved, but identification of specific networks, neurons and synapses within these macroscopic structures is an unenviable task.

There is overwhelming evidence that multiple brain structures are involved in the encoding and recall of episodic memories. Some of these structures support sustained activity for short periods, some support rapid encoding of representations over an intermediate-length period, some are believed to transfer these representations to structures responsible for more long-term encoding, and some are believed to compare expected representations with incoming neural activity. Within networks of networks of networks, where each network contains (at least) millions of neurons and billions of synapses, and where each synapse is associated with multiple learning states, we must identify and measure individual synapses before and after some experience, showing how they move through a series of states and how these states contribute to neural activity within each network.

Even if we ignore the overwhelming technical challenges to our definitive test, current technologies addressing this level of resolution are invasive, so we do not expect volunteers. We must turn instead to non-human animal subjects, raising another major challenge to cognitive neuroscience methods. Episodic memory involves the mental 'reliving' of some part of a subject's personal history, an extremely difficult phenomenon to assess non-verbally. The current approach to experiments on non-human animals is to demonstrate *episodic-like* memory. That is, to show that an animal remembers *what* happened, *where* it

happened, and *when* it happened, the three qualities generally regarded as essential to episodic memory. Many of these experiments are brilliantly designed and very convincing, but nonetheless, there is no direct way to assess whether an animal is mentally reliving a previous experience. In this case, if you accept the three criteria, you are free to interpret experimental results within the scope of this assumption.

3.2 Computational models of episodic memory

Now look at a modeling approach. Arguably, the first computational cognitive neuroscience model belongs to David Marr (Marr 1971). Marr was less concerned with episodic memory per se than with the function of the hippocampus more generally, but the two have been inextricably linked since Scoville and Milner's report of the memory deficits of patient HM in 1957. Following bilateral removal of his hippocampus and nearby cortical structures, HM was unable to form new experiential memories, despite being able to learn new motor skills. Marr's hippocampal model exemplifies computational cognitive neuroscience because it provides an anatomically-grounded, mechanistic explanation of these and other behavioral data.

The basic network structure underlying these models is shown in Figure 3C. Each node is connected to all other nodes, and the strength of the connection between any two nodes is a function of their states, as proposed by Hebb and described at the beginning of this section. During a learning phase, nodes are 'clamped' to values that represent internal representations of external events. This clamping may be equated with the response of the hippocampus to cortical input. During a retrieval phase, recurrent networks such as Marr's have an important property known as attractor dynamics. Given a noisy version of a state from the training set, the original state is retrieved by summing the input arriving at each node via all connections in the network. That is, the network can be cued by partial or noisy input to retrieve previous states. This ability is referred to as pattern completion and is a form of *autoassociative* memory,

so-called because the memory is associated with itself. In more cognitive terms, the model recalls its experience following exposure to single events, essential properties of episodic memory.

To Marr, hippocampal subfield CA3 was reminiscent of a recurrent architecture because of the unusually high density of collateral interactions between neurons in this region. In addition to proposing a role for the extensive collateral interactions of CA3 neurons, Marr was aware of the sparseness of neural activity in the dentate gyrus (DG), a hippocampal region providing input to CA3. He hypothesized that sparse representations (fewer active neurons) in DG reduce the overlap between representations entering CA3, allowing rapid encoding. The importance of this hypothesis cannot be overstated. Computationally, overlapping patterns are suited to slow extraction of central tendencies, such as the learning of object categories or motor skills. Conversely, the extraction of statistical regularities must be *avoided* in a structure that learns specific, individual patterns, such as those representing specific, personal experiences. In brief, overlap between representations leads to competition between them in the recall process. By reducing overlap, Marr's DG allowed CA3 to perform one-shot learning, a requirement of episodic memory.

3.3 Hippocampal models in the tradition of Marr

A thorough account of hippocampal modeling in the tradition of Marr would fill this entire volume, but we believe the basic model of Alvarez and Squire (1994) captures the dominant theory of the most important hippocampal data. A crucial aspect of these data concerns cortical memory consolidation. Above, we briefly mentioned patient HM and his inability to form new experiential memories following removal of his hippocampus. This *anterograde* amnesia is common to many hippocampal patients, though HM's case is certainly extreme. HM also exhibits a *graded retrograde* amnesia. Among his experiences before surgery, he's less likely to remember events occurring closer to the time of excision.

His memories of events more than around two years before surgery appear to be unaffected. Like anterograde amnesia, graded retrograde amnesia is common to many hippocampal patients. This phenomenon suggests that memories are stored in the hippocampus for an intermediate period, but are ultimately encoded in neocortex.

In Alvarez and Squire's model, shown in Figure 4, the hippocampus plays the role of a rapid learner, serving to consolidate cortical memories. Under this general framework, the hippocampus responds to a unique configuration of cortical activity with a unique internal representation. Because cortical-hippocampal pathways are bidirectional, this representation serves as a key or index to the cortical activity that created it. Upon presentation of a subset of the original cortical activity, the key is retrieved by the pattern completion abilities of recurrent networks, as in Marr's model. The key reactivates the full cortical representation and thereby the memory. With repeated activation of the hippocampal key and consequent reactivation of the composite cortical pattern, cortical representations gradually learn to activate each other and eventually no longer need the hippocampus as intermediary. Learning is achieved by a simple Hebbian rule, where changes in weight depend on the correlation between the firing rates of the nodes they connect. The rule simply uses a higher learning rate with their hippocampal weights than their cortical weights. In summary, the fast-learning hippocampus serves memory consolidation in the slower-learning cortex.

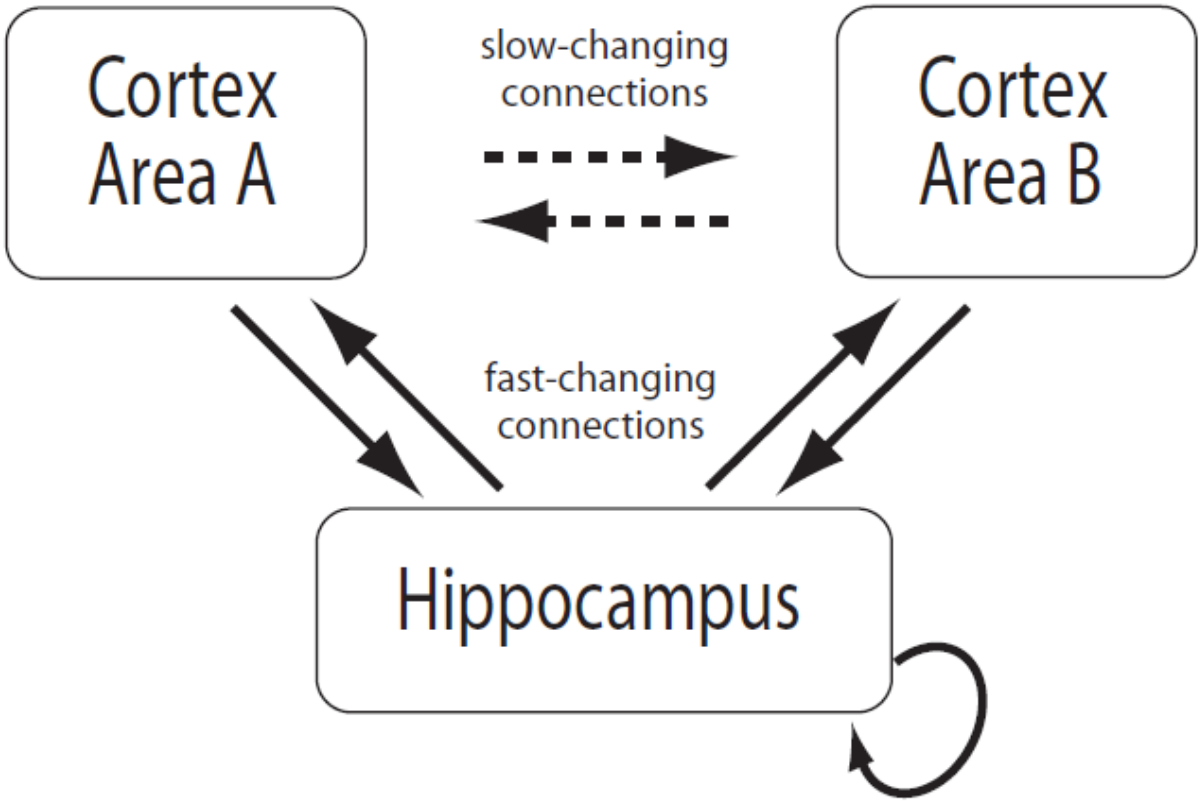

Figure 4: The hippocampal model of Alvarez and Squire (1994).

We have noted that in general, the purpose of non-human animal experiments in cognitive neuroscience is to allow researchers to invasively study brain activity in behaving subjects. Most of these data reflect activity at the level of neurons, but the models described here do not provide this level of resolution. Spatial constraints prevent us from discussing so-called *spiking neuron models* of the hippocampus and episodic memory, but suffice to say, these models offer cellular resolution because each node may be equated with a single neuron. As such, their output may be directly compared with intracellular recordings from behaving non-human animal subjects. These models embody different sets of assumptions than the models we have discussed and provide a powerful means of bridging high resolution activity and structure with behavioral data.

4 Dynamic neural field models and decision-making

Our discussion of episodic memory highlights many of the challenges faced by cognitive neuroscientists and the importance of memory systems to the understanding of brain function. Next, we introduce dynamic neural field (DNF) models. These basic models capture fundamental characteristics of neural organization, describing both neural and behavioral data. To demonstrate these important characteristics, we show how a DNF model captures basic neural response properties, comparing output from the model with *tuning curves* in cat visual cortex. We then discuss the model in the context of perceptual choice, showing how it explains neural responses in monkey inferior temporal cortex (IT). We also demonstrate the model in a closely related decision-making task, where the strength of visual evidence is easily controlled and psychophysical measurements such as reaction times and response accuracy are easily determined. The model captures many details of neural activity in monkey lateral intra-parietal cortex (LIP) and corresponding psychometric functions.

4.1 Dynamic neural field models and tuning curves

To understand DNF models, consider again a recurrent network where each node is connected to all other nodes, depicted in Figure 3C. In Marr's model, we considered a cognitive task in which connection weights where determined by Hebbian learning. Here, our nodes have the same structure and function, but the weight between any two nodes is solely a function of the distance between them. Consider the nodes in Figure 5A. The weight from node 3 to node 2 is the same as the weight from node 3 to node 4 because the distance between them is the same. Similarly, the weight from 3 to 1 is the same as the weight from 3 to 5. This arrangement holds for the connections emanating from any node, so the weight from 2 to 1 is

the same as the weight from 2 to 3 and so on. Furthermore, the same weights holds for any node-centric perspective. For instance, the weight from 1 to 4 is the same as the weight from 2 to 5.

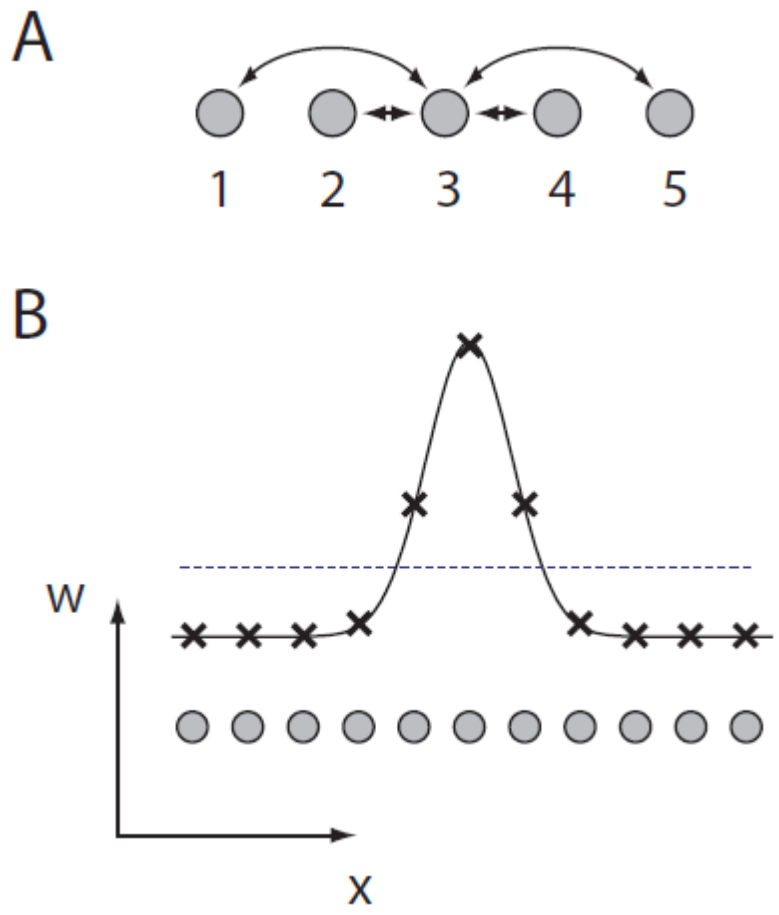

Figure 5: (A) A dynamic neural field model is a recurrent network (Figure 3C) where weights depend on the distance between nodes. (B) Subtracting a constant from a Gaussian function of this distance provides negative (inhibitory) weights below the dashed line.

For this illustrative case, we use Gaussian weights, depicted in Figure 5B. This connectivity may be genetically encoded in some structures, but can also result from Hebbian learning. This arrangement initially renders all weights positive, approaching 0 with increasing distance between nodes. With all

positive weights, we have no competition because all nodes excite one another, leading to an explosion of network activity. We therefore include a global, activity-dependent inhibition by subtracting a constant value from all weights, creating negative (inhibitory) weights. These weights are represented in the figure by values below the dashed line. The inhibition mimics the activity of a pool of inhibitory neurons. The resulting network allows nodes to support each other locally and inhibit each other distally.

Now we look at some physiological data. Figure 6A shows data from the experiments of Henry et al. (1974). These data show electrophysiological recordings from a neuron in cat primary visual cortex (V1) while moving line segments are shown to the cat at different orientations. This neuron is maximally responsive to horizontal lines, responding with decreasing activity as the stimulus deviates from this preferred orientation. The neuron's response to specific feature values is called its *tuning curve*. In this case, the tuning curve is approximately Gaussian.

Now look at Figure 6C. In the surface plot, the x-axis shows time and the y-axis shows the position of 100 nodes in the model. Node activity is shown on a grey scale, where dark areas depict high-rate activity. Over time, we provide inputs centred on a succession of nodes, corresponding to the presentation of line segments with orientations from -45 to 45 degrees. Consider node 50, maximally-responsive to an orientation of 0 degrees (shown on the right). Initially, the stimulus does not evoke a noticeable response from node 50, evidenced by the white (low) activity at this location in the network. With different orientations over time, node 50 becomes more active and then decreases it response again, as shown in Figure 6B. This figure is remarkably similar to the electrophysiological recordings shown directly above. For this reason, DNF models are a dominant model of cortical hypercolumns (Hansel and Sompolinsky 1998).

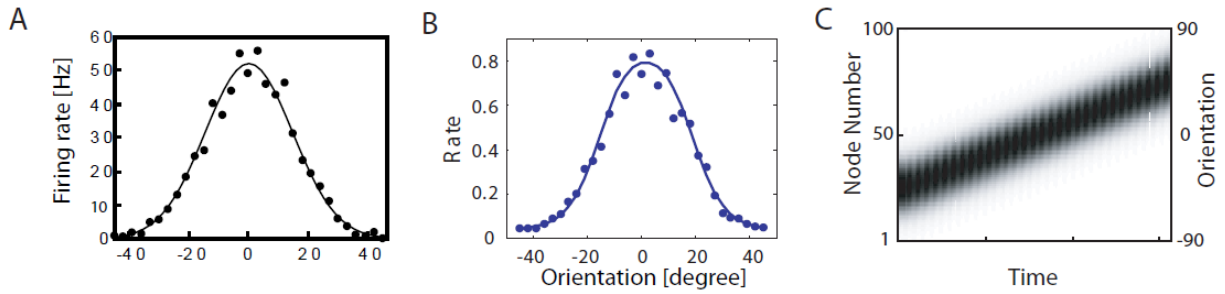

Figure 6: Tuning curves of cortical neurons are explained by a dynamic neural field model.

4.2 Perceptual choice

We now apply the above model to a more cognitive task, that of perceptual choice. Perceptual choice is a decision-making process. We routinely make decisions based on sensory data, much of which may be uncertain or may conflict with other sources of information such as our memories and expectations. We weigh-up all these sources and decide accordingly. The two crucial aspects of models of perceptual choice are that a decision is made by accumulating evidence over time and that this evidence is uncertain.

Historically, two types of models have been discussed in the literature: accumulator models and diffusion models. With accumulator models, evidence for stimuli accumulates over time and a decision is reached when evidence for a particular stimulus exceeds a threshold, or when a response is required and the stimulus with the greatest sub-threshold evidence is chosen. With diffusion models, it is not the absolute evidence that is accumulated, but the difference between the evidence for each stimulus. Traditional models of perceptual choice provide a good example of cognitive science methods because they make useful predictions about behavior, but they do not come under the umbrella of cognitive *neuro*science because they are not concerned with the biological mechanisms underlying their implementation. With a DNF model, evidence for competing stimuli is modelled by inputs to the network, demonstrated by the

tuning curves above. Competition is provided by mutual inhibition between active regions of the network, where the first region to reach a threshold level of activity corresponds to the perceptual choice.

We now demonstrate a DNF model of perceptual choice. In the experiments of Chelazzi et al. (1993), electrophysiological recordings were made from monkey inferior temporal cortex (IT), an area correlated with higher-level visual processing. Monkeys were shown numerous images from magazines until a 'good' and 'poor' stimulus were identified for each of several neurons. A good stimulus is one that elicits a strong neuronal response. A poor one does not. As depicted in Figure 7, monkeys first fixated on a central dot on a computer screen. Subsequently, either a good or poor visual stimulus was presented for fraction of a second. After a further delay, the two images were simultaneously presented and the monkey's task was to move its eyes to the image that was previously shown.

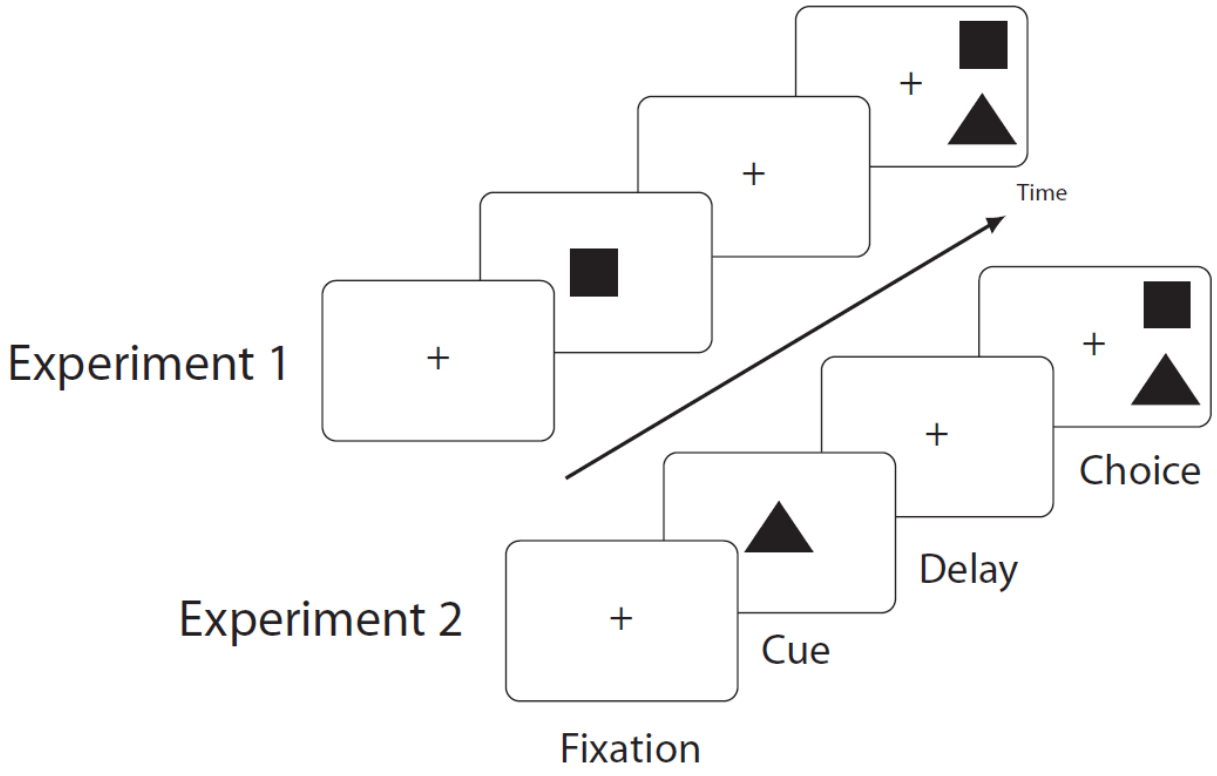

Figure 7: Experimental paradigm of Chelazzi et al. (1993).

Figure 8A shows the response of an IT neuron during the experiment. The left shaded area shows the time of cue presentation. When the good image is presented, the neuron responds with a pronounced increase in rate, plotted by the solid line. The poor image leads to a decrease in firing rate, plotted by the dashed line. This behavior is consistent with the DNF model, shown for this task in the panel below. In computer simulations, presentation of the good stimulus alone leads to the high-rate response shown by the blue line in the left shaded area. Because the network implements local excitation and distal inhibition, this stimulus leads to a decrease in activity in a node that responds to a different (poor) stimulus, plotted by the red, dashed line.

As alluded to in Section 2, we are now comparing the activity of a node to the response of an individual neuron, not to a population of neurons. This comparison is appropriate because the node represents a group of neurons with similar response characteristics. No single neuron in the group is solely responsible for the network's response to the preferred feature, but in this case, all members exhibit similar activity.

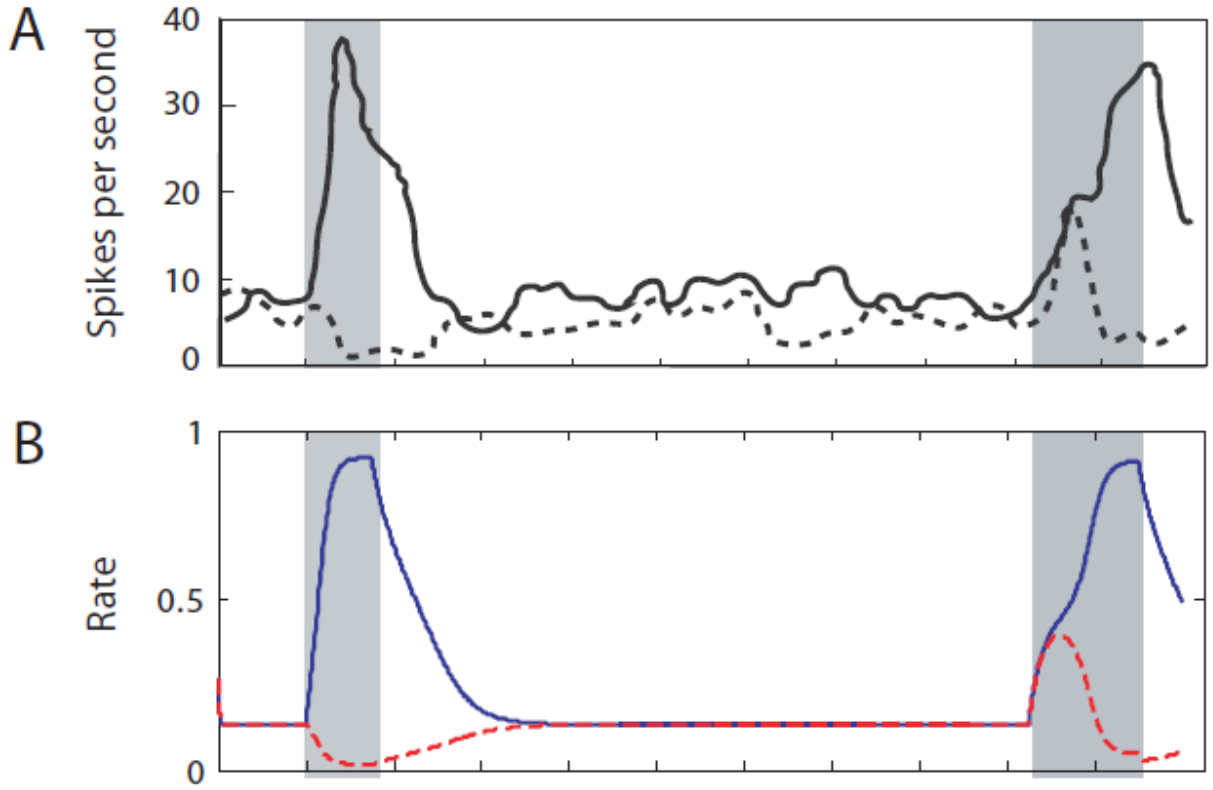

Figure 8: Data from Chelazzi et al. (1993) is shown in *A*. Output from a dynamic neural field model is shown in *B*.

The most stunning result by Chelazzi et al. is the response of the system when both the poor and the good images are presented after 3.3 seconds. If the initial cue was the good image, the neuron again responds strongly, but when the cue was the poor image, an initial increases in rate is followed by a marked decrease in activity. This effect is easily understood with the help of the DNF model. Presentation of both stimuli leads to strong initial responses by nodes selective for these inputs, shown by the blue and red curves in the right shaded area. We give a small bias of 1% to the cued image, corresponding to the memory of the cue. Mutual inhibition does the rest, as the remembered stimulus dominates processing.

4.3 Decision-making, psychophysics and uncertainty

Our discussion of perceptual choice has thus far centered on electrophysiological data, addressing animal behavior with neural resolution, but with very crude behavioral resolution. To provide a cognitive perspective, we need finer grained behavioral measurements. Motion discrimination experiments provide such a tool. In a typical motion discrimination task, a display of dots is pre-sented on screen and a fraction of the dots are repeatedly displaced at a fixed offset. This offset provides a kind of step frame animation, so the dots appear to move in the direction of displacement, typically to the left or to the right. Monkeys are trained to move their eyes in the direction of dominant movement to indicate their perceptual choice. Experimentalists manipulate the *coherence* of movement by controlling the percentage of displaced dots. This simple manipulation allows experimentalists to finely control the strength of evidence. Recording response time and accuracy provides psychophysical measurements in response to the controlled parameter (coherence in this case).

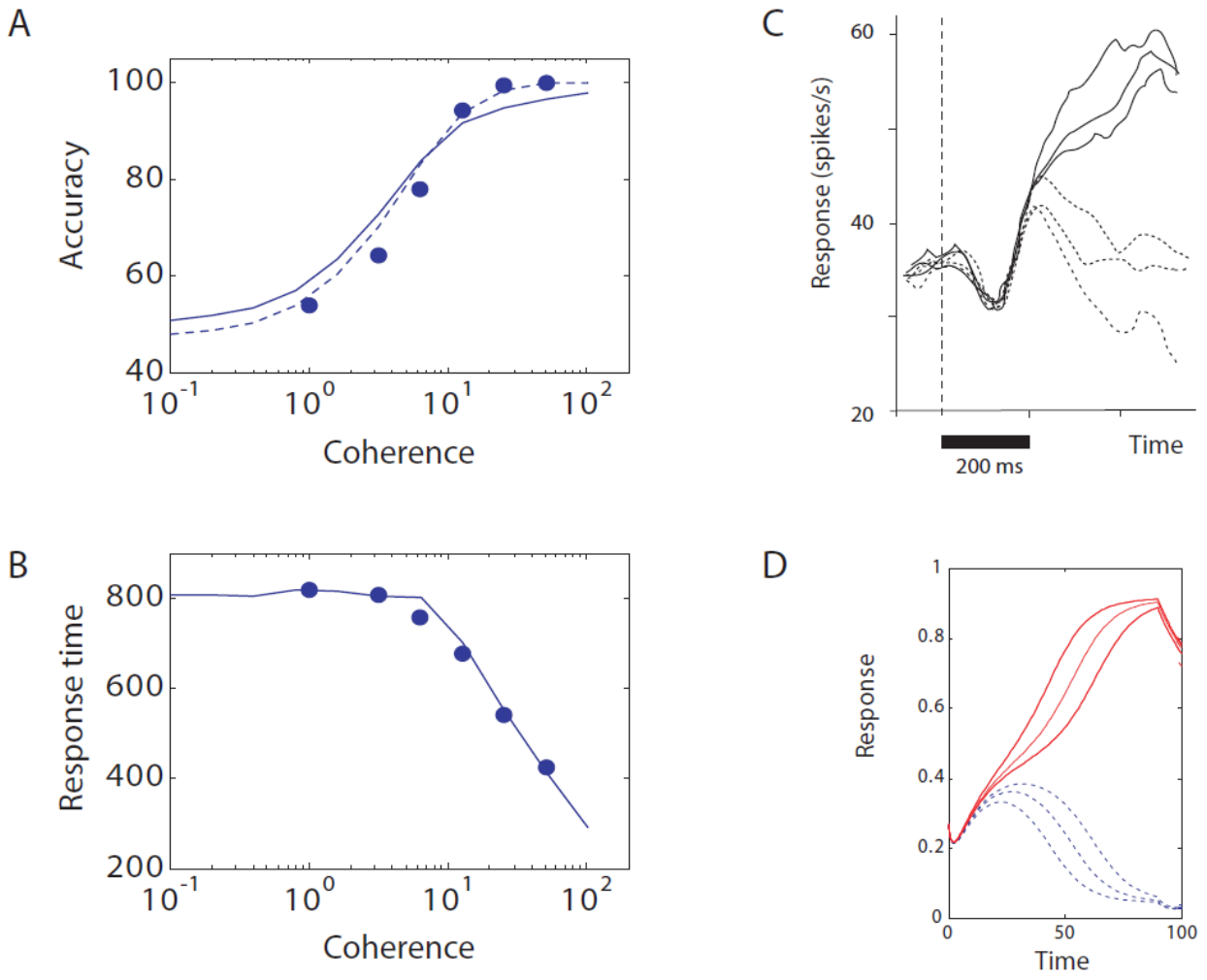

Figure 9: Data from Roitman and Shadlen (2002) is explained by the dynamic neural field model. (A) The *psychometric* function. Dots show data. Curves show simulations. (B) The *chronometric* function. (C) Neural response over time for different strengths of evidence. (D) Simulations of the task in *C*.

Psychophysical measurements from Roitman and Shadlen (2002) are shown as dots in Figures 9A and B. These figures show the *psychometric* and *chronometric* functions respectively, depicting accuracy and reaction time as a function of the strength of evidence. At low levels of coherence, the monkey makes many errors and reaction times are long. With increasing coherence, accuracy increases and reaction times become shorter. Output from the model fits these data well (blue curves) where coherence and

response time are modelled by input strength and the time to threshold respectively. Results from corresponding neural recordings in monkey lateral intraparietal area (LIP) are shown in Figure 9C. These data also compare favorably with output from simulations, shown in Figure 9D.With increasing coherence, there is an increase in activity among neurons with a preferred response to the direction of motion (solid lines): the higher the coherence, the steeper the slope of the lines. The response of neurons preferential to other stimuli increases initially before decreasing (dashed lines): the stronger the evidence against their preferred direction, the steeper the slope of the decrease. This effect is equivalent to the competition between good and poor stimuli in the Chelazzi experiment, described above. For the interested reader, more details of the above simulations can be found in Trappenberg (2008). For a very thorough treatment of the experiments of the Shadlen group, see the biologically-motivated, systems-level model of Grossberg and Pilly (2008).

As a final word on DNF models, we wish to emphasize their explanatory power at multiple levels of abstraction. In the examples above, DNF models provide mechanistic explanations for neurophysiological and behavioural data, suggesting that they capture a fundamental principle of information processing in the brain. Specificially, this princinple is the competition between neural populations by mutual inhibition. Dynamic neural field theory addresses this competition continuously in space (neural tissue) and time, enabling the model to capture the real-time competition between feature values that differ along a continuum.

5 Hierarchical bidirectional memory

Models of episodic memory and decision-making are exemplary of a vast computational literature within cognitive neuroscience, but in isolation, the models we have described thus far are limited in scope. In this section, we provide a brief overview of a set of models we believe will play a prominent role in future

theories of cortex and cognitive function, returning to a generic architecture shown earlier in Figure 3D. At this level of abstraction, there are two main characteristics of these networks: hierarchical structure and bidirectional connectivity. Well known to anatomists, these characteristics are found in all processing pathways in the brain and are especially prominent in cortex. Hierarchical processing is required to represent the incredible complexity of the world and the concepts we use to function within it. Bidirectional processing generates expectations, essential to cope with the high-volume processing demands of even the simplest cognitive functions.

In the following, our usage of the term ‘representation’ simply refers to neural activity. For instance, in sensory processing, neurons fire in response to things in the world, so their activity represents those very things in the brains internal model of the world. In hierarchical neural processing, representations at low levels of a hierarchy are combined to form composite representations at the next level up. These representations are in turn combined so that higher-level constructs are represented at increasingly advanced levels of the hierarchy. For example, in visual processing, neurons in early visual cortex respond to points of light. Signals from these cells converge on neurons at the next stage of processing, causing them to fire and thereby strengthening these connections by Hebbian plasticity. These cells can represent edges due to the co-appearance of points of light in natural objects. The compositional process continues such that edges are combined to form contours and so on until representations of objects are achieved. Hebbian learning facilitates this process by ensuring the same features are combined to represent the same objects as sets of features are repeatedly encountered together in the world. This fundamental, compositional process is depicted in Figure 10.

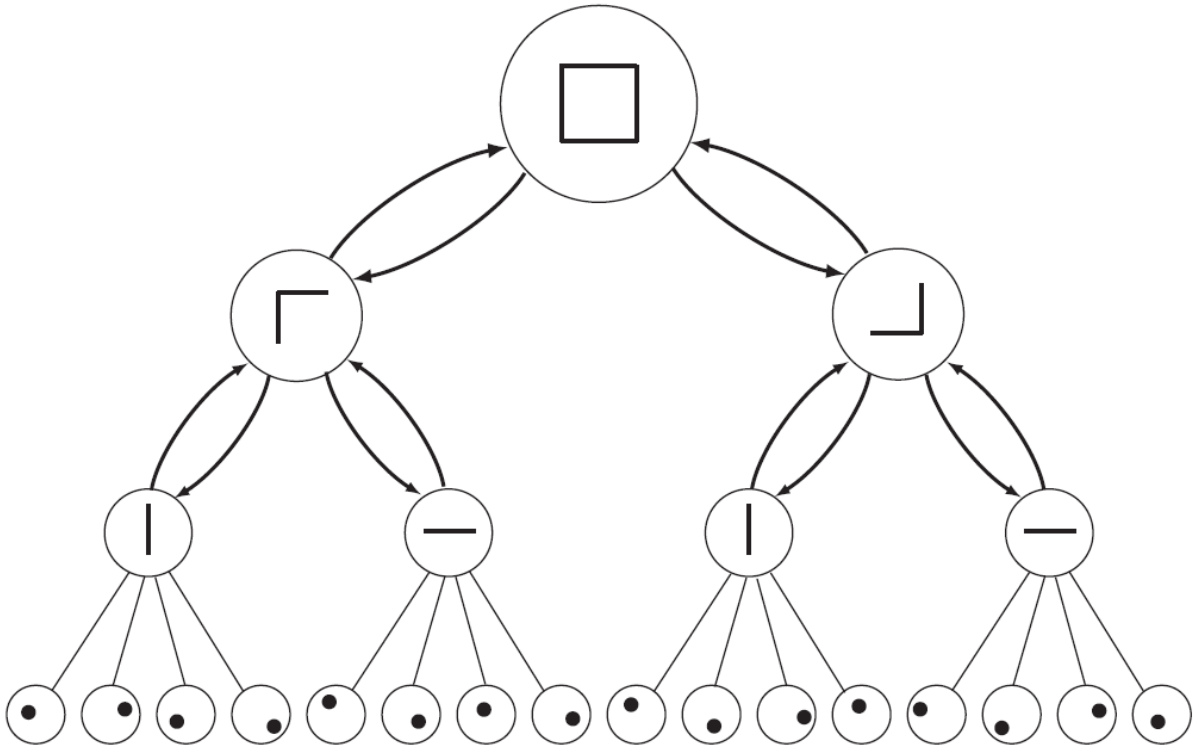

Figure 10: Hierarchical, bidirectional neural processing.

Described in this way, feature detection is a bottom-up process, but that is only half the story. Feedback connectivity provides top-down processing. It's well known that cortical regions are commonly reciprocally connected. That is, if an area *A* projects to area *B*, then area *B* commonly projects to *A*. It's hard to imagine this arrangement on a neuron by neuron basis, but remember, the nodes depicted in Figure 10 represent cell assemblies. Coactive representations at consecutive levels of a bidirectional hierarchy strengthen synapses in *both* directions, as depicted in the figure. As such, when low-level features are combined to form a composition, not only do the connections driving the composition become stronger, but the composition learns to activate the lower-level features. To us, this synergy of bottom-up and top-down processing is one of most powerful ideas in neuroscience, explaining imagination, expectation, and perhaps most importantly, prediction.

To understand why it explains imagination, consider the square at the top. To visualize the square, we need to activate its features. In the simplified example in the figure, activating this node leads to the propagation of activity to the next level down, where the two 'corner nodes' are activated. These nodes in turn propagate their activity to the nodes representing each side of the square, but they also re-excite the node above. It's as if they're saying "we're still working on your square features, but we need you to stay active to get it done". This propagation of activity continues up and down the hierarchy until representations at all levels are co-active. In short, a square, like anything else, is defined by its features, so to imagine the square, we need to activate these lower-level feature representations. You can think of this process as a kind of pattern completion, starting at the top of the hierarchy. Activation of the complete, hierarchical pattern *is* imagining the square.

To understand why hierarchical, bidirectional processing explains expectations and predictions, we turn to Stephen Grossberg's Adaptive Resonance Theory (ART) (see Carpenter and Grossberg 2003). Continuing with our simplified example of visual processing, imagine that our hierarchical network has learned the square representation and that a subset of the square's features arrives at our feature detection nodes (perhaps our view of the square is partially blocked by another object). This subset of 'point-of-light' nodes is enough to drive some of the 'line' nodes above, which propagate activity back down to all point-of-light nodes that have consistently driven them in the past, and propagate activity back up to the 'corner' nodes. Activity moving up and down the hierarchy like this is said to *resonate* as long as top-down and bottom-up representations match. That is, we see a square if our partial or noisy input provides a reasonable match with our expectations of squares. It's a beautiful idea, but there's more.

Now imagine we've been told to look out for squares. Higher cortical areas bias our top-down activity to favor square sightings, so that any square-like features are more quickly and reliably composed by virtue of this bias. In this case, top-down activity reflecting an expectation of squares is stronger than usual and

the balance between top-down and bottom-up activity is dynamically adjusted. We thus more readily see squares, but only when there are squares to be seen. This stipulation may sound obvious, but it leads to an important point. Under ART, resonance breaks down when top-down expectations and bottom-up sensory information *do not* match. In this case, further network activity amounts to a search for another learned object category. If none is found (resonance is not achieved) the activity leads to the learning of a new category by Hebbian plasticity.

We believe hierarchical, bidirectional networks provide one of the most promising tools in cognitive neuroscience, with the potential to provide cortical explanations for many aspects of cognition. A recent example of work in this field is the model of George and Hawkins (2005) demonstrating these principles in a *Bayesian* framework. Such models exemplify some of the most advanced methods in machine learning. A more biologically realistic implementation of many of these foundations has been advanced by Geoffrey Hinton for many years. His networks are based on so-called Boltzmann machines or *deep belief networks*. These models have probabilistic nodes that more closely resemble the incredibly noisy processing of real neurons. While many models in this area are quite abstract, continued focus on related theoretical and experimental research has the potential to make great strides over the coming decades.

6 Summary

Research in cognitive science has provided a wealth of knowledge about the underlying properties of cognitive phenomena. Related cognitive models have provided useful tools for understanding human behavior and have played an important role scientifically and in technical applications. The field of human-computer interface, for instance, has benefited immensely from behavioral modeling. In contrast, neuroscience focuses on physical mechanisms. A mechanistic understanding of the brain is essential for scientific advancements in many research areas. To give but one example, understanding the function of

neuronal ion-channels is crucial to the design of drugs combatting neuro-degenerative diseases. Cognitive neuroscience bridges cognitive science and neuroscience, grounding cognitive functions in the underlying mechanisms of the brain.

The brain has been assumed to be the seat of the mind since the age of Enlightenment and not surprisingly, the boundaries between cognitive science and neuroscience are often blurry. The field of cognitive neuroscience provides many tools for understanding this relationship in more detail. Functional brain imaging methods such as EEG and fMRI are perhaps the best examples of these tools, with the power to directly measure brain activity in humans performing cognitive tasks, but we need more than just data. How are behavioral data and brain imaging data related? Theories that address this question must concisely describe these data and make testable predictions and are by no means limited to sets of mathematical equations.

Our discussion of episodic memory in Section 3 illustrates some of the difficulties of experimental work in cognitive neuroscience. In large part, these difficulties stem from the invasiveness of current electrophysiological methods and the consequent need for non-human subjects. The behavior of these subjects requires interpretation, as cognitive phenomena do not typically have strong behavioral correlates and brain anatomy is far from uniform across species. Recent technological advances such as multi-voxel analysis of fMRI have great potential in this regard. The combination of this technology with the high temporal resolution of EEG provides a powerful tool for studying humans engaged in cognitive tasks. These experiments are providing a wealth of data to guide theories of the processing mechanisms of structures in the brain.

The models we have discussed make numerous assumptions, including abstractions at the levels of physiology, anatomy, and interacting dynamic systems. Without assumptions, the predictive power of a model is lost among the details of its implementation. The DNF model introduced in Section 4

exemplifies assumptions about structure in the brain. In the implementation here, cooperation and competition within the network are implemented by short range excitation and long range inhibition. The model explains a remarkable variety of physiological and psychophysical data, but excitatory and inhibitory connections in the brain tend to have the opposite arrangement, *ie*. local inhibition and more distal excitation. This apparent anatomical inconsistency has lead to investigations of the function of individual layers in cortex, where specific roles are proposed for cell types in columnar cortical architectures. Alternatively, the *effect* of local excitation and long range inhibition can emerge dynamically from a number of architectures. Finally, neural field models are used by some researchers to capture the effect of interacting systems in the brain, independent of the details of implementation. All three levels of enquiry are equally valuable.

In Section 5, we briefly described an abstract, yet system wide model of neocortex that we believe may ultimately unify theories in cognitive neuroscience. This framework offers a concrete path to the theoretical investigation of exciting experimental discoveries. The integration of experimental and theoretical methods in cognitive neuroscience is arguably still in its infancy, and interest and enthusiasm for this unified approach is rapidly growing. We believe the systematic integration of experiment and theory will lead to great advances in cognitive neuroscience in the coming decades.

7 Recommended reading

The first book to broadly address computational mechanisms of brain and mind was The Computational Brain (Churchland and Sejnowski 1992). The combined expertise of the authors spans philosophy of mind and computational neuroscience, and results in a readable, thorough and authoritative book that remains highly relevant to cognitive neuroscience. Hawkins (Jeff Hawkins with Sandra Blakeslee 2004) provids an accessible, enjoyable account of material alluded to in Section 5. Written for a general

audience, his book beautifully captures principles of cortical information processing, offering a unified theory of cortex and a recipe for building intelligent machines. More detailed explanations of modeling techniques and their interpretation and application in cognitive neuroscience can be found in (Trappenberg 2002). This chapter is mostly computational in its scope. For a broader perspective on cognitive neuroscience, see Gazzaniga et al. (2002).

References

Carpenter, G. A. and Grossberg, S. 2003. 'Adaptive resonance theory', in Arbib (ed.), *The handbook of brain theory and neural networks, second edition,* pp. 87–90. Cambridge, MA: MIT Press.

Chelazzi, L., Miller, E. K., Duncan, J., and Desimone, R. 1993. A neural basis for visual search in inferior temporal cortex. *Nature* 363: 345–347.

Churchland, P. S. and Sejnowski, T. J. 1992. *The computational brain*. MIT Press.

Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. 2002. *Cognitive neuroscience*. W. W. Norton and Company.

George, D., & Hawkins, J. 2005. A hierarchical bayesian model of invariant pattern recognition in the visual cortex. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2005)*.

Grossberg, S. and Pilly, P. K. 2008. Temporal dynamics of decision-making during motion perception in the visual cortex. *Vision Research* 48: 1435 - 1473.

Hansel, D. and Sompolinsky, H. 1998. 'Modeling feature selectivity in local cortical circuits', in Koch and Segev (eds.), *Methods in neural modeling, from ions to networks* (2 ed.). MIT Press.

Hebb, D. O. 1949. *The organisation of behaviour*. John Wiley, New York.

Henry, G. H., Dreher, B., and Bishop, P. O. 1974. Orientation specificity of cells in cat striate cortex. *Journal of Neurophysiology 37*: 1394–1409.

Jeff Hawkins with Sandra Blakeslee 2004. *On intelligence*. Owl Books.

Marr, D. 1971. Simple memory: A theory of archicortex. *Philosophical Transactions of the Royal Society of London*, *262*(841): 23–81.

Martin, S. J., Grimwood, P., and Morris, R. G. M. 2000. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience*, 23: 649–711.

Roitman, J. D. and Shadlen, M. N. 2002. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience*, *22*(21): 9475–9489.

Rumelhart, D. E. and McClelland, J. L. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Trappenberg, T. 2010. *Fundamentals of computational neuroscience* (2 ed.). Oxford University Press.

Trappenberg, T. 2008. 'Decision making and population decoding with strongly inhibitory neural field models', in Heinke and Mavritsaki (eds.), *Computational modelling in behavioural neuroscience: Closing the gap between neurophysiology and behaviour*. Psychology Press, London.

Notes

1 Feedback activity between nodes in the same neural substrate should not be confused with the backpropagation of error signals mentioned above.