Chapter 8: Connectionism As Synthetic Psychology

8.1 INTRODUCTION

In Chapter 6, we introduced the synthetic approach with the "thoughtless walker" examples. In Chapter 7, we turned to a historical review of more serious robotics research to examine why researchers might be attracted to the synthetic approach. We saw that one of the main attractions was the possibility of generating interesting and surprising behaviors from the interaction between a fairly simple system of components and the environment in which this system was embedded.

One concern raised at the end of Chapter 6, and *not* addressed in Chapter 7, involved the relevance of the synthetic approach (as depicted to this point) to the study of psychological processes. In particular, the modern renaissance of the synthetic approach that was pioneered by such researchers as Ashby and Grey Walter is strongly associated with the movements of behavior-based robotics (Brooks, 1999) and embodied cognitive science (Pfeifer & Scheier, 1999). These research traditions are strongly anti-representational, and are largely dedicated to removing the "think" component from the sense-think-act cycle. This is strongly reminiscent of a failed tradition in experimental psychology, called behaviorism, that attempted to limit psychological theory to observables (namely, stimuli and responses), and which viewed as unscientific any theories that attempted to describe internal processes that mediated relationships between sensations and actions. I believe we can write a psychology, define it as Pillsbury, and never go back upon our definition: never use the terms consciousness, mental states, mind, content, introspectively verifiable, imagery, and the like. "I believe that we can do it in a few years without running into the absurd terminology of Beer, Bethe, Von Uexküll, Nuel, and that of the so-called objective schools generally. It can be done in terms of stimulus and response, in terms of habit formation, habit integrations and the like" (Watson, 1913).

Modern cognitive psychology emerged from a strong reaction against behaviorism's anti-representational stance (Leahey, 1987). In psychology, there is a long history of powerful theoretical and empirical arguments against behaviorism, and as a result behaviorism is no longer an accepted position (but see Leahey, 1987, pp. 461-463). The standard view in psychology is that many phenomena cannot be adequately explained without appealing to mental representations. Given this situation, and given that we have only considered the synthetic approach in the context of anti-representational research, this leads to an obvious question: is there anything in the synthetic approach that can be applied to the study of representational processes?

The purpose of this chapter is to consider one version of the synthetic approach that can be applied in this way, and which as a result can truly be considered to be synthetic psychology. The SEA methodology that we have been discussing in the last two chapters can be employed in a research tradition that is interested in exploring representational issues. This position will be supported in this chapter as follows. First, we will consider the properties of connectionist simulations in the context of the synthetic approach. This will be done to argue that connectionism offers one – though not the only – medium in which representational, synthetic research can be conducted. Second, we will discuss one case study that has recently appeared in the literature (Dawson, Boechler, & Valsangkar-Smyth, 2000). This case study examines how connectionist simulations can be used to investigate issues related to one "higher-order processing" topic: spatial cognition. We will then use this case study as a motivator to step back and consider a variety

of techniques for performing synthetic psychology using connectionism, which is the topic for Part II of this book.

8.2 BEYOND SENSORY REFLEXES

The complexity of the behaviors of all of the machines that were surveyed in Chapter 7 was rooted in a set of simple sensorimotor reflexes that were embedded in a complicated environment. For example, the behavior of all the Lego Dacta robots that were demonstrated in the movies at the end of the chapter were based upon simple routines in which a particular sensation (e.g., a value detected by a light sensor, or a switch depressed on a touch sensor) was immediately converted into a particular response (e.g., a particular motor speed, or a change in motor direction). The extent to which the behavior of these robots was complex, surprising, or interesting was due to the interaction of these simple reflexes with the environments in which the robots were placed.

The purpose of this section is to briefly consider the extent to which sensorimotor reflexes can be relied upon to form the basis of synthetic psychology. First, some evidence supporting the existence of visuomotor modules in humans will be described. This evidence indicates that sensorimotor reflexes should be plausibly considered as a component of synthetic psychology. Second, the limitations of such reflexes will also be considered. The claim that will be made is that synthetic psychology cannot rely exclusively on such reflexes, and should therefore explore other foundations – some of which might be representational.

8.2.1 Visuomotor modules

One of the most influential ideas that has been proposed in cognitive science is that of the modularity of perceptual processing (Fodor, 1983). While “perception is smart like cognition in that it is typically inferential, it is nevertheless dumb like reflexes in that it is typically encapsulated” (p. 2). A module is a domain-specific perceptual system that solves a very particular problem, and is incapable of solving other information processing problems. The operations performed by a module are rapid, mandatory, and run to completion once they are initiated. Fodor argues that all of these characteristics are achieved by associating each module with fixed neural architecture -- modularity is physically built into the brain. The corollary of this position is that general inferential processing, which is by definition nonmodular, is *not* going to be associated with a fixed neural architecture. It is because of this that Fodor (p. 119) is not surprised that we have a neuroscience of sensory systems, but that we do not have a neuroscience of thought.

The modularity proposal is usually portrayed as being part of the “sense-think-act” cycle that defines much of the status quo in cognitive science (Dawson, 1998, Chapter 7). Specifically, many problems in early vision are solved by informationally encapsulated modules (sense). The output of these modules is then passed on to visual cognition or higher-order cognition for inferential or semantic processing (think). The results of this higher-order processing are then used to generate actions. However, this is not the only way in which modularity has been incorporated into cognitive science.

In some of the earliest work on the neuroscience of vision, Lettvin, Maturana, McCulloch, and Pitts (1959) identified neurons in the visual system of the frog that only responded to specific visual stimuli, and which in some sense were modular feature detectors. For instance, one type of cell appeared to be a “bug detector”, because it only responded to a stimulus that could be described as a small, moving black spot. However, such feature detectors in the frog do not appear to feed into a higher-order thinking mechanism. Instead, the frog’s visual system appears to be organized into a system of “sense-act” or visuomotor modules. Not only do these modules detect a specific visual stimulus, but they also generate a specific motor response.

The existence of visuomotor modules in the frog was first demonstrated by Ingle (1973). In a seminal experiment, Ingle surgically removed one hemisphere of the optic tectum of a frog.

This lesion produced a particular form of blindness in which the frog pursued prey presented to the eye that was connected to the remaining tectum, but did not respond to prey presented to the eye that would have been connected to the ablated tectum. The lesion did not affect the frog's ability to avoid a stationary barrier placed between it and its prey. Importantly, the amphibian brain is very plastic, and Ingle found that 6 to 8 months after surgery, the nerve fibers from the "bad eye" regenerated, and became connected to the remaining optic tectum on the "wrong" side of the animal's head. In this case, when a prey target was presented to the "bad eye", the frog was no longer blind to it, and attempted to catch it. However, because of the tectal rewiring, the animal's responses were in the wrong direction. The frog always moved toward a location that was mirror-symmetrical to the actual location of the target, and this incorrect response was shown to be due to the topography of the regenerated nerve fibers. In other words, one role of the optic tectum in the frog is to mediate a visuomotor module that converts a visual sensation directly into a motor response.

Perhaps surprisingly, studies of brain-injured patients have demonstrated that the human visual system may also be organized into visuomotor modules (Goodale, 1988, 1995; Goodale & Humphrey, 1998). For instance, Goodale and his colleagues have studied one patient, DF, who suffered irreversible brain damage as a result of carbon monoxide poisoning. One result of this brain damage was that DF's ability to recognize visual shapes or patterns was severely impaired. She "was unable to describe the orientation and form of any visual contour, no matter how that contour was defined" (Goodale, 1995, p. 167). However, DF's visuomotor abilities were not impaired at all. "Even though she cannot recognize a familiar object on the basis of its visual form, she can grasp that object under visual control as accurately and as proficiently as people with normal vision" (p. 169). Another patient, VK, had the exact opposite pattern of dysfunction after a series of strokes. VK had normal form perception, but her visuomotor control – in particular, her ability to form her hand to grasp objects of different shapes – was severely impaired.

8.2.2 Reflexes Vs. Representations

The evidence that there exists, even in humans, modular systems that involve direct linkages between sensation and action is consistent with behavior-based robotics and embodied cognitive science. Specifically, research in these fields is based upon the assumption that intelligence emerges situating a system in the world, and is not a result of representational processing. The existence of visuomotor modules is strongly suggestive of a human information processing architecture that is similar in many ways to Brook's (1989, 1999) subsumption architecture. However, even researchers of visuomotor modules in humans would agree that such reflexes are not the sole foundations of psychological processing.

For example, Goodale and Humphrey (1998) point out that "while there is certainly plenty of evidence to suggest that visuomotor modularity of the kind found in the frog also exists in the mammalian brain, the very complexity of day-to-day living in many mammals, particularly in higher primates, demands much more flexible organization of the circuitry" (p. 184). They propose a reformulation of Ungerleider and Mishkin's (1982) proposal of two separate anatomical streams of visual processing. Ungerleider and Mishkin proposed a ventral stream from primary visual cortex to inferotemporal cortex for the processing of visual appearances, and a dorsal stream from primary visual cortex to posterior parietal cortex for the processing of visual locations – the so-called what-where distinction. Goodale and Humphrey distinguish these two streams in terms of the kinds of representations that they construct, and their purpose. The dorsal stream computes representations of object locations and shapes in an egocentric frame of reference. These representations are components of visuomotor modules, and are used to control a variety of movements (e.g., saccades, grasps, etc.). The ventral stream computes representations of object features in an allocentric frame of reference. These representations become part of later semantic processing.

Furthermore, the dorsal and ventral streams as described by Goodale and Humphrey (1998) are not independent, but are required to interact with one another. For instance, "certain

objects such as tools demand that we grasp the object in a particular way so that we can use it properly. In such a case both streams would have to interact fairly intimately in mediating the final output" (p. 203). The fact that the two systems can interact is supported by theoretical arguments and anatomical evidence (DeYoe & van Essen, 1988) that shows that they are far more interconnected than was originally proposed by Ungerleider and Mishkin (1982). These interactions are, of course, the source of the flexibility and control that Goodale and Humphrey note is required by higher-order visual systems to deal with complicated environmental demands.

That stimulus-response reflexes are not sufficient to account for many higher-order psychological phenomena is a theme that has dominated cognitivism's replacement of behaviorism as the dominant theoretical trend in experimental psychology. In the study of language, this theme was central to Chomsky's (1959) critical review of Skinner (1957). Many of the modern advances in linguistics were the direct result of Chomsky's proposal that generative grammars provided the representational machinery that mediated regularities in language (Chomsky, 1965, 1995; Chomsky & Halle, 1991). Similar arguments were made against purely associationist models of memory and thought (Anderson & Bower, 1973). For example, Bever, Fodor, and Garrett (1968) formalized associationism as a finite state automaton, and demonstrated that such a system was unable to deal with the clausal structure that typifies much of human thought and language. Paivio (1969, 1971) used the experimental methodologies of the verbal learners to demonstrate that a representational construct – the imageability of concepts – was an enormously powerful predictor of human memory. The famous critique of "old connectionism" by Minsky and Papert (1988) could be considered proofs about the limitations of visual systems that do not include mediating representations. These examples, and many more, have lead to the status quo view that representations are fundamental to cognition and perception (Dawson, 1998; Fodor, 1975; Jackendoff, 1992; Marr, 1982; Pylyshyn, 1984).

Some robotics researchers also share this sentiment, although it must be remembered that behavior-based robotics was a reaction against their representational work (Brooks, 1999). Moravec (1999) suggests that the type of situatedness that characterizes behavior-based robotics (for example, the simple reflexes that guided Grey Walter's tortoises) probably provides an accurate account of insect intelligence. However, at some point systems built from such components will have at best limited abilities. "It had to be admitted that behavior-based robots did not accomplish complex goals any more reliably than machines with more integrated controllers. Real insects illustrate the problem. The vast majority fail to complete their life cycles, often doomed, like moths trapped by a streetlight, by severe cognitive limitations. Only astronomical egg production ensures that enough offspring survive, by chance" (p. 46). Internal representations are one obvious medium for surpassing such limitations.

Interestingly, the view that representations provide an adaptive advantage for an organism, as well as flexibility and control of processing, are both central to the philosophical views of Karl Popper. Popper proposed an evolutionary theory in which organisms are constantly engaged in a process of problem solving, a process that Popper viewed as always being resolved through trial and error. "Error-elimination may proceed either by the complete elimination of unsuccessful forms (the killing-off of unsuccessful forms by natural selection) or by the (tentative) evolution of controls which modify or suppress unsuccessful organs, or forms of behavior, or hypotheses" (Popper, 1979, p. 242). Popper viewed consciousness as an evolved system of "plastic control", a system that could be used to control behavior, but which was also subject to changes via feedback. The purpose of representations was argued to supply "controls which can eliminate errors without killing the organism; and it makes it possible, ultimately, for our hypotheses to die in our stead" (p. 244).

In summary, the synthetic models developed in behavior-based robotics and embodied cognitive science can be described as systems of sensorimotor reflexes or visuomotor modules which, when embedded in a complicated environment, can generate surprising or interesting behavior. These models are consistent with the anti-representational motivation of this research trend, namely, the elimination of the "think" component of the "sense-think-act" cycle. These

models are also consistent with evidence of the existence of visuomotor modules in highly complex organisms, including humans. However, theoretical and empirical arguments would suggest that not all psychological phenomena are equivalent to sensorimotor reflexes. Some representational processes must exist as well, and it is these processes that are of keen interest to psychologists. The question that this leads to is this: can the synthetic approach be conducted in a way that provides the advantages that have been raised in previous chapters, but that also provides insight into representational processing?

8.2.3 Synthesis And Representation

Of course, the answer to the question that was just raised is a resounding yes. There is nothing in the synthetic approach per se that prevents one from constructing systems that use representations. Describing a model as being synthetic or analytic is using a dimension that it is completely orthogonal to the one used when describing a model as being representational or not. This is illustrated in Table 8-1, which categorizes some examples of research programs in terms of these two different dimensions.

	Analytic	**Synthetic**
Representational	• Production system generated from analysis of verbal protocols • e.g. (Newell & Simon, 1972)	• Multilayer connectionist network for classifying patterns using abstract features • e.g. (Dawson, Boechler & Valsangkar-Smyth, 2000)
Non-Representational	• Mathematical model of associative learning based upon analysis of learning behavior of simple organisms • e.g. (Rescorla & Wagner, 1972)	• Behavior-based robotics system constructed from a core of visuomotor reflexes • e.g. (Brooks, 1989)

Table 8-1. Classification of some example research programs according to two separate dimensions, analytic vs. synthetic and representational vs. non-representational.

The placing of most of the research examples in Table 8-1 should be clear from discussions that we have had in preceding chapters. For example, production system research is designated as being both analytic and representational. It is analytic because production systems are almost always derived from an intensive analysis of the verbal protocols of human problem solvers (Ericsson & Simon, 1984; Newell & Simon, 1972). It is representational in the sense that production systems define a set of definite rules that detect, and modify, data structures that are stored in a working memory. Indeed, production systems are one of the prototypical examples of the power of symbolic representations in classical cognitive science (Newell, 1980, 1990).

Behavior-based robotics is designated as being both synthetic and non-representational. As we have seen in Chapter 7, it is explicitly synthetic in the sense that researchers build robots from fairly simple subsystems, and then examine the interesting kinds of behaviors that emerge when the robots are situated in an environment (Pfeifer & Scheier, 1999). It is also an attempt to be as anti-representational as possible. “In particular I have advocated situatedness, embodiment, and highly reactive architectures with no reasoning systems, no manipulable representations, no symbols, and totally decentralized computation” (Brooks, 1999, p. 170). One of the foundational assumptions of behavior-based robotics is that if a system can sense its environment, then it should be unnecessary for the system to build an internal model of the world.

Mathematical models of associative learning, such as the Rescorla-Wagner model (Rescorla & Wagner, 1972), are designated as being both analytic and non-representational.

Such models are described as being analytic because they are usually based upon an analysis of behavioral regularities (see Chapters 3 and 4). They are described as being non-representational because such models do not appeal to representational content to explain behavior, and frequently model direct relationships between stimuli and responses.

8.3 CONNECTIONISM, SYNTHESIS, AND REPRESENTATION

Connectionism was placed in the final cell of table 8-1. In my view, modern multi-layer PDP networks permit research that is both synthetic and representational, and therefore offers one plausible avenue for conducting synthetic psychology. The following subsections will elaborate on why connectionism can be viewed in this way. Specifically, we will briefly discuss connectionism in the context of the three hallmarks of the synthetic approach: synthesis, emergence, and analysis.

8.3.1 Connectionism And Synthesis

In adopting the synthetic approach, a researcher is committed to identifying a basic set of building blocks. Each of these building blocks defines a primitive element. The set of all of the available primitives defines an entire architecture. For a cognitive scientist, an architecture dictates "what operations are primitive, how memory is organized and accessed, what sequences are allowed, what limitations exist on the passing of arguments and on the capacities of various buffers, and so on. Specifying the functional architecture of a system is like providing a manual that defines some particular programming language" (Pylyshyn, 1984, p. 92). The goal of synthetic research is to see what variety of systems can be constructed from a particular architecture.

In cognitive science, an architecture is usually a kind of programming language. However, this is not a necessary property. In some cases, there may not be any programming environment at all. For example, in building our "thoughtless walkers" in Chapter 6, the architecture that we restricted ourselves to was a set of K'NEX rods, connectors, and motors. In other cases, an architecture might involve a combination of hardware and software elements. For example, n building our simple Braitenberg-like vehicles in Chapter 7, the architecture that we restricted ourselves to were the components of Lego Dacta, which included sensors, motors, Lego bricks, and which also included the RCX brick and the primitive operations provided by the NQC programming language. This kind of combined architecture is typical of research in embodied cognitive science (Pfeifer & Scheier, 1999).

The architecture is a foundational idea in cognitive science, and therefore it is not surprising that many different research programs revolve around proposals for the architecture of cognition. In some cases, researchers present a particular architecture as a candidate proposal for the "language of thought". For instance, Newell and Simon (1972) made very strong claims that production systems defined the functional architecture of the mind. Dawson (1998, p. 170) provides (an incomplete) table of proposed cognitive architectures that lists 24 different examples. In other cases, theoretical and empirical debates in cognitive science revolve around whether particular properties are part of the architecture or not. For example, in the 1970s and 80s the imagery debate was about whether the visual properties of mental images were built directly into the architecture (Block, 1981). A more recent debate concerns whether the architecture of mind is analogous to the architecture of a digital computer (Bechtel & Abrahamsen, 1991; Churchland, Koch, & Sejnowski, 1990; Clark, 1989, 1993; Fodor & Pylyshyn, 1988; Pylyshyn, 1991; Smolensky, 1988), and has spawned a new architectural proposal, connectionism (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986).

Parallel distributed processing (PDP) models, or connectionism, are based on general assumptions about the kind of information processing carried out by the brain. First, it is assumed that the primitives for this type of information processing are individual neurons. Second, it is assumed that the pattern of connections between neurons is analogous to the program in a

conventional computer, because these connections define the causal interactions between neurons (Smolensky, 1988). Third, it is assumed because the brain is composed a set of primitive units that operate in parallel, and because representations are distributed across a wide array of neurons and synapses, the kind of information processing carried out by the brain must be quite different from that to found in a digital computer. "The analogy between the brain and a serial digital computer is an exceedingly poor one in most salient respects, and the failures of similarity between digital computers and nervous systems are striking" (Churchland et al., 1990, p. 47).

PDP models represent the embodiment of these general assumptions in a computer simulation environment that permits the construction of networks that can solve problems in an incredibly diverse set of domains (Dawson, 1998). Essentially, the building blocks of PDP models represent abstract mathematical descriptions of the kind of information processing that neurons do. "Because of the 'all-or-none' character of nervous activity, neural events and the relations among them can be treated by means of propositional logic" (McCulloch, 1988, p. 19). This functional approach ignores many of the biological properties of neurons, and attempts to simplify information processing as much as possible.

The first major component of a connectionist architecture is the processing unit. As far as information processing goes, connectionists view individual neurons as computing three different functions. First, they sum the total incoming signal from dendrites. Second, they convert this total incoming signal into internal activity. Third, this internal activity may be converted into an output signal to be communicated to other neurons. As we saw briefly in Chapter 5, processing units in a PDP architecture compute three different equations to mimic these three neural information processing activities. First, a net input function determines the total signal coming into the unit. Second, an activation function converts the net input into an internal level of activity. Third, an output function converts the internal activity into an output signal that can be sent to other processing units. As we will see in Part II of this book, choosing different equations for computing net input, activation, and output can create different "flavors" of connectionism.

The second major component of a PDP architecture is the connection between units. A processor sends a signal to another through a weighted connection, which is a functional description of a synapse (Dawson, 1998). The connection is a communication channel that amplifies or attenuates a numerical signal being sent through by multiplying the signal by the weight associated with the connection. The weight defines the nature and strength of the connection. For example, inhibitory connections are defined with negative weights, and excitatory connections are defined with positive weights. Strong connections have strong weights (i.e., the absolute value of the weight is large), while weak connections have near-zero weights. Different kinds of connectionist networks permit different patterns of connections between processing units. For example, feedforward networks connect one layer of processing units to another in such a way that signals only get sent in one direction. In contrast, recurrent networks permit signals to travel in both directions between sets of processing units. The degree of connectivity might also be varied from one connectionist architecture to another. In some simulations, the networks might be massively parallel, which means that every processing unit in one layer is connected to every processing unit in another layer. In other simulations, the pattern of connectivity might be of limited order, which means that every processor is not connected to every processor in another layer.

The third major component of a connectionist architecture is the learning rule. While the pattern of connections in a network is analogous to a program in a conventional computer, a PDP network is usually not programmed in any traditional sense. Instead, the network is usually taught to perform a task of interest. The purpose of this training is to determine the appropriate value for each connection weight in the network. Prior to training, most networks will have a set of small, randomly assigned weights. At the end of training, the network will have a very specific pattern of connectivity (in comparison to its random start), and will have learned to perform a particular stimulus response pairing. Training is accomplished by using one of a variety of possible learning rules. Many learning rules result in a supervised change in a network's weights. What this means is that for each example pattern that is presented to the network, there is a known

response, and the difference between the desired and actual responses can be used to guide weight modification. Other learning rules are unsupervised, which means that the network is not instructed about what the desired responses are, but instead "carves up" the example patterns into a self-generated set of categories. In Part II of this book we will discuss in detail a variety of different learning rules that can be used to train PDP networks.

With respect to synthesis, connectionist research typically proceeds as follows: First, a researcher identifies a problem of interest, and then translates this problem into some form that can be presented to a connectionist network. Second, the researcher selects a general connectionist architecture, which involves choosing the kind of processing unit, the possible pattern of connectivity, and the learning rule. Third, a network is taught the problem. This usually involves making some additional choices specific to the learning algorithm – choices about how many hidden units to use, how to present the patterns, how often to update the weights, and about the values of a number of parameters that determine how learning proceeds (e.g., the learning rate, the criterion for stopping learning). If all goes according to plan, at the end of the third step the research will have constructed a network that is capable of solving a particular problem. The next subsection illustrates this aspect of connectionist research by describing an example network that was trained to make judgments about the distances between cities on a map of Alberta (Dawson, Boechler & Valsangkar-Smyth, 2000).

8.3.2 Connectionism And Synthesis: An Example

One of my former PhD students, Patricia Boechler, did her thesis on how navigation through a hypertext document was affected by different navigational aids (e.g., Boechler & Dawson, 2001). One of the central themes of her research was the validity using the spatial metaphor to describe this kind of "virtual navigation" (Boechler, 2001). One issue that needs to be addressed whenever such questions are raised concerns what is meant by the term "spatial". We used a PDP network to provide a synthetic framework for exploring the properties of space (Dawson, Boechler & Valsangkar-Smyth, 2000). This subsection describes the creation of the network; later in this chapter we will consider some of its other properties in relation to the issues of emergence and analysis.

8.3.2.1 Metric Representations Of Space

Our everyday interactions with the visual and spatial world are grounded in the essential experience that space is metric. Mathematically speaking, a space is metric if relationships between locations or points in the space conform to three different principles (Blumenthal, 1953). The first is the *minimality principle*. According to this principle, the shortest distance in the space is between a point *x* and itself. The second is the *symmetry principle*. According to this principle, the distance in the space between two points *x* and *y* is equal to the distance between points *y* and *x*. The third is the *triangle inequality*. According to this principle, the shortest distance in the space between two points *y* and *x* is a straight line.

One recurring theme in the study of cognition, perception, and action is that intelligent agents have internalized the metric properties of the space in which they find themselves situated. As a result, the mental representations used by these agents are thought by some researchers to have metric properties in their own right. The paragraphs below briefly introduce three different examples of such proposals: similarity spaces, mental images, and cognitive maps.

Similarity is one of the most important theoretical constructs in cognitive psychology (Medin, Goldstone, & Gentner, 1993). The notion of similarity is central to theories of learning, perception, reasoning, and metaphor comprehension. One of the goals of cognitive psychology has been to determine the mental representations that enable similarity relationships to affect this wide range of psychological phenomena. One proposal that received a great deal of attention in the 1970s was that concepts were represented as points in a multidimensional space, where the

dimensions of the space stood for either simple or complicated featural properties (Romney, Shepard, & Nerlove, 1972; Shepard, Romney, & Nerlove, 1972). In this kind of representation, the similarity between two different concepts was reflected in the distance between their locations in the multidimensional space. Researchers conducted a number of different studies in which ratings of concepts were used to position a set of concepts in the metric space. This empirically derived space was then used to predict behavior on a variety of different tasks, including analogical reasoning (Rumelhart & Abrahamson, 1973) and judgments of the aptness of metaphor (Tourangeau & Sternberg, 1981; Tourangeau & Sternberg, 1982). Importantly, one of the main assumptions underlying the similarity space proposal was that this space was metric.

On the basis of this assumption, one would expect that the metric properties of the space would be reflected in the behaviors that were governed by the space. For example, if a subject used the similarity space to rate the similarity between two concepts A and B, then one would expect these ratings to be symmetric: the similarity between A and B should be the same as the similarity between B and A, because the distance between A and B in the similarity space is presumed to be symmetric.

A second example of a proposed representation that preserves the metric properties of space is mental imagery. Mental imagery is a visual experience that is usually elicited when people solve visuospatial problems. Not only does mental imagery provide a visual or pictorial experience, but mental images give the sense of being manipulated in a spatial manner -- for instance, by being scanned, rotated, or zoomed in to (Kosslyn, 1980). Early behavioral studies of the manipulation of mental images have provided data that suggest that they are indeed spatial in nature. For example, many studies recorded the reaction times of subjects as they used mental images to perform some task, and found, for instance, that latencies increased linearly as a function of increases in the distance that an image had to be scanned or of increases in the amount that an image had to be rotated (Kosslyn, 1980; Shepard & Cooper, 1982).

More recent research has turned to cognitive neuroscience in an attempt to explore the representations responsible for mental imagery. Kosslyn and others have used a variety of modern brain imaging techniques to show that when people generate mental images, they use many of the same brain areas that are also used to mediate visual perception (Farah, Weisberg, Monheit, & Peronnet, 1989; Kosslyn, 1994; Kosslyn et al., 1999; Kosslyn, Thompson, & Alpert, 1997; Kosslyn, Thompson, Kim, & Alpert, 1995; Thompson, Kosslyn, Sukel, & Alpert, 2001). In particular, mental imagery elicits activity in the primary visual cortex, a brain area that is organized topographically. Kosslyn has used this kind of evidence to propose an information processing system that is responsible for the generation and manipulation of images. He argues that mental images are patterns of activity in a visual buffer that is a spatially organized structure in the occipital lobe.

A third example of a proposed representation that preserves the metric properties of space is the cognitive map. Beginning with Tolman's (1932, 1948) proposal that the spatial abilities of the rat were mediated by cognitive maps, representations that preserve the metric properties of space have been fundamentally important to the study of how humans and animals navigate (Kitchin, 1994). Behavioral studies have demonstrated that animal representations of space do indeed appear to preserve a good deal of its metric nature (for introductions, see Cheng & Spetch, 1998; Gallistel, 1990, Chap. 6). Many researchers are now concerned with identifying the biological substrates that encode metric space. Single-cell recordings of neurons in the hippocampus of a freely moving animal have provided compelling biological evidence that one function of the hippocampus is to instantiate a metric cognitive map (O'Keefe & Nadel, 1978). In particular, neuroscientists have discovered *place cells* in the hippocampus that respond only when a rat's head is in a particular location in the environment (O'Keefe & Nadel, 1978). These place cells can be driven by visual information (e.g., by the presence of objects or landmarks in the environment), and appear to be sensitive to some of the metric attributes of space. For example, O'Keefe and Burgess (1996) found evidence that the receptive field of a place cell can be de-

scribed as the sum of two or more Gaussian tuning curves sensitive to the distance between an animal and a wall in the environment.

8.3.2.2 Are Spatial Representations Metric?

While research on each of these three proposals for spatial representations has provided evidence that the metric properties of space can be internalized, this evidence is not univocal. With respect to similarity spaces, Tversky and his colleagues conducted a number of experiments that demonstrated that similarity judgments were not metric, because in different situations it could be shown that these judgments were not always symmetric, did not always conform to the minimality principle, and did not always conform to the triangle inequality (Tversky, 1977; Tversky & Gati, 1982).

With respect to mental imagery, it has been shown that by manipulating the tacit beliefs of subjects (Bannon, 1980), or by altering the complexity of the image being used (Pylyshyn, 1979), the linear relationship between reaction time and image properties could be eradicated. These findings were used to argue that our experience of mental images is based upon more primitive and non-spatial representational components (Pylyshyn, 1980, 1981, 1984). Even the evidence from neuroscience is not without controversy. In a detailed review of the literature, Mellet, Petit, Mazoyer, Denis, and Tzourio, (1998) cite several studies that have found that some mental imagery tasks do not produce activity in primary visual cortex.

With respect to cognitive maps, it has been argued that place cell circuitry by itself does not provide a cognitive map that can be considered to be metric in the mathematical sense. First, place cells are not organized topographically; the arrangement of place cells in the hippocampus is not isomorphic to the arrangements of locations in an external space (Burgess, Recce, & O'Keefe, 1995; McNaughton et al., 1996). Second, it has been argued that place cell receptive fields are at best *locally* metric (Touretzky, Wan, & Redish, 1994), and that as a result a good deal of spatial information (e.g., information about bearing) cannot be derived from place cell activity. Some researchers have argued that place cells make up only a part of the cognitive map, and that the neural representation of metric space requires the coordination of a number of different subsystems (McNaughton et al., 1996; Redish & Touretzky, 1999; Touretzky et al., 1994).

8.3.2.3 A Synthetic Approach To Spatial Representation

The three examples that were briefly reviewed above all involve proposals for metric spatial representations that mediate spatial behavior. However, in each example it was shown that such proposals are not without controversy. In some instances, behavior that is presumably guided by the representation can violate the metric properties of space. In other instances, inspections of the representational or neural structures that mediate spatial behavior or experience reveal regularities that are inconsistent with the notion that the underlying structure is metric in nature.

One reason that such inconsistencies emerge may be because these representational proposals were the product of an analytic research strategy. Cognitive psychologists typically develop theories about underlying representations by decomposing complex behavior into more basic functions (Cummins, 1983; Dawson, 1998). While this approach, called functional analysis, has been extremely successful, it can be dangerous to use. One problem with it that we saw in Chapter 7 is that it can lead to theories that are more complicated than necessary, because the decomposition can fail to partition behavior appropriately into three different categories (behavior caused by the organism, behavior elicited by a complex environment, and behavior that emerges at the interface between an agent and its environment) (Braitenberg, 1984; Simon, 1996). A second problem is that the decomposition is theory-driven, and as a result can miss regularities that are real, but not intuitively obvious. “The tendency will be to break different capacities down into different constituent processes. As a result, explanations that are given of the capabilities in

question will rest on a false and artificial theory, one that is, in effect, *engineered* to account for data but that is not a realistic model of human neuropsychology" Rollins (2001).

The synthetic approach is one alternative to functional analysis. Dawson, Boechler and Valsangkar-Smyth (2000) decided to explore the notion of spatial representations synthetically by building a PDP network that could make judgments that preserved the metric properties of space. Could a simple network learn to make such judgments? If so, then what kind of internal representation would it use? Would the representation be metric or non-metric?

8.3.2.3.1 Defining The Problem

As was noted earlier, the first step in synthesizing a connectionist network is to choose a problem of interest, and to translate this problem into a form that could be dealt with by a PDP model. Dawson, Boechler and Valsangkar-Smyth (2000) wanted to create a network that could perform a behavior that was complicated enough to be of psychological interest, and which also preserved the metric properties of space. The task that they selected was a ratings task, in which a network was presented a pair of cities, and had to rate the distance between the two cities on a scale from 0 to 10. This kind of task is of psychological interest, because it is often used to collect distance-like data from human subjects (Shepard, 1972). By basing the ratings on distances measured between cities on a map, one can also ensure that a system that can make such judgments is preserving the metric properties of space as well.

	BANFF	CALGARY	CAMROSE	DRUMHELLER	EDMONTON	FORT MCMURRAY	GRANDE PRAIRIE	JASPER	LETHBRIDGE	LLOYDMINSTER	MEDICINE HAT	RED DEER	SLAVE LAKE
BANFF	0	2	4	3	5	9	7	3	4	7	5	3	7
CALGARY	2	0	3	2	3	8	8	5	3	6	3	2	6
CAMROSE	4	3	0	2	1	6	6	5	5	3	5	2	4
DRUMHELLER	3	2	2	0	3	8	8	6	3	5	3	2	6
EDMONTON	5	3	1	3	0	5	5	4	6	3	6	2	3
FORT MCMURRAY	9	8	6	8	5	0	8	8	10	6	10	6	5
GRANDE PRAIRIE	7	8	6	8	5	8	0	4	10	8	10	6	4
JASPER	3	5	5	6	4	8	4	0	7	7	8	5	5
LETHBRIDGE	4	3	5	3	6	10	10	7	0	7	2	4	8
LLOYDMINSTER	7	6	3	5	3	6	8	7	7	0	5	4	5
MEDICINE HAT	5	3	5	3	6	10	10	8	2	5	0	5	8
RED DEER	3	2	2	2	2	6	6	5	4	4	5	0	4
SLAVE LAKE	7	6	4	6	3	5	4	5	8	5	8	4	0

Table 8-2. Ratings of distances between cities of Alberta. The number in each cell represents the answer to the question "What is the rating of the distance between City 1 and City 2?", where City 1 is the row label and City 2 is the column label.

Dawson, Boechler and Valsangkar-Smyth (2000) chose thirteen different locations in the province of Alberta: Banff, Calgary, Camrose, Drumheller, Edmonton, Fort McMurray, Grande Prairie, Jasper, Lethbridge, Lloydminster, Medicine Hat, Red Deer, and Slave Lake. They took all possible pairs from this set to create a set of 169 different stimuli, each of which could be described as the question "On a scale from 0 to 10, how far is City 1 from City 2?" The desired ratings for each stimulus were created as follows. First, from a map of Alberta they determined the shortest distance in kilometers between each pair of locations. Second, they then converted these distances into ratings. If a stimulus involved rating the distance from one place to itself, the rating was assigned a value of 0. Otherwise, if the distance was less than 100 kilometers, then it was assigned a value of 1; if the distance was between 100 and 199 kilometers, then it was assigned a value of 2; if the distance was between 200 and 299 kilometers, then it was assigned a

value of 3; and so on up to a maximum value of 10 which was assigned to distances of 900 kilometers or more. The complete set of ratings that were used is provided in Table 8-2.

These ratings were designed to preserve the metric properties of the map of Alberta. For instance, the ratings are symmetric – any cell (x,y) in the table holds the same rating as the corresponding cell (y,x). As well, the minimality principle is upheld, because the rating of the distance from any cell to itself is equal to 0, as can be seen by examining the diagonal of Table 8-2. To confirm that a system that could generate the ratings in this table must have, in some sense, internalized the map of Alberta, Dawson, Boechler and Valsangkar-Smyth (2000) analyzed Table 8-2 with a statistical technique called multidimensional scaling (MDS). MDS is designed to take proximity information as input, and to then convert this information into a geometric configuration of points from which the proximities can be derived (Kruskal & Wish, 1978). For example, if one were to give MDS a table of distances between cities (e.g., a table commonly found on a roadmap), MDS would produce a map with each city situated in the correct location. When Table 8-2 is analyzed using MDS, it generates a plot in which each of the 13 cities are located very near the position in which they would be found if one examined a road map of Alberta, as is shown in Figure 8-1.

Figure 8-1. Results of analyzing Table 8-2 with MDS.

8.3.2.3.2 Choosing The Network Architecture

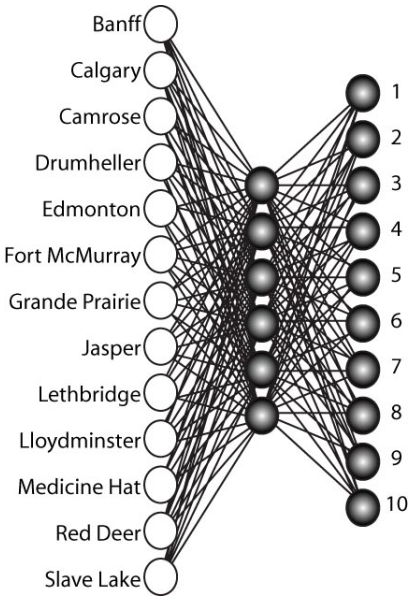

Figure 8-2. The spatial judgement network

The second step in synthesizing a connectionist network is to choose a particular architecture, and to train this architecture to solve the problem of interest. Part of this step involves making fairly general architectural choices. Dawson, Boechler and Valsangkar-Smyth (2000) decided to train a feedforward network to solve this spatial judgment task. This network is illustrated in Figure 8-2. The first layer of this network was a set of 13 different input units. The input units used a very simple unary notation to represent pairs of places to be compared. Each input unit represented one of the thirteen place names. Pairs of places were presented as stimuli by turning two of the input units on (that is, by activating them with a value of 1). For example, to ask the network to rate the distance between Banff and Calgary, the first input unit would be turned on (representing Banff), as would the second input unit (representing Calgary). All of the other input units would be turned off (that is, were activated with a value of 0). This unary representational scheme was chosen because it contains absolutely no information about the location of the different places on a map of Alberta. In other words, the input units themselves did not provide any metric information that the network could use to perform the ratings task.

As can be seen from Figure 8-2, ten output units were used to represent the network's rating of the distance between the two place names presented as input. To represent a rating of 0, the network was trained to turn all of its output units off. To represent any other rating, the network was trained to turn on one, and only one, of its output units. Each of these output units

represented one of the ratings from 1 to 10. For example, if the network turned output unit 5 on, this indicated that it was making a distance rating of 5.

The middle layer of the spatial judgment network was a set of six hidden units. Dawson, Boechler and Valsangkar-Smyth (2000) selected this number of hidden units because pilot simulations had shown that this was the smallest number of hidden units that could be used by the network to discover a mapping from input to output. When fewer than six hidden units were used, the network was never able to completely learn the task. Previous research has suggested that forcing a network to learn a task with the minimum number of hidden units produces a network that is much easier to interpret, in comparison to a network that has more hidden units than are required to solve a problem (Berkeley, Dawson, Medler, Schopflocher, & Hornsby, 1995).

In addition to making decisions about the input representation, the output representation, and the number of hidden units, Dawson, Boechler and Valsangkar-Smyth (2000) had to make specific decisions about the properties of the hidden and output units, and about how the network was to be trained. They decided that the hidden and output units should all be value units. A value unit uses a particular type of Gaussian activation function to convert net input into internal activity that ranges between 0 and 1. Such units generate a maximum response of 1 when their net input is equal to the mean of the Gaussian. These units were used because one of the primary goals of the research was to interpret the internal representations discovered by the network. As will be discussed in more detail below, a number of different studies have demonstrated that networks of value units permit their internal structure to be interpreted in great detail (Berkeley et al., 1995; Dawson, 1998; Dawson & Medler, 1996; Dawson, Medler, & Berkeley, 1997; Dawson, Medler, McCaughan, Willson, & Carbonaro, 2000; Leighton, 1999; Zimmerman, 1999).

Because Dawson, Boechler and Valsangkar-Smyth (2000) decided to use value units in the spatial judgment network, they were committed to using a learning rule that was specifically designed for value unit networks (Dawson & Schopflocher, 1992). In using this rule, they had to make a number of additional decisions about parameters that affected the specific course of learning. Prior to training, all of the connection weights were randomly assigned values ranging from –0.10 to +0.10. The biases of processing units (i.e., the means of the Gaussian activation functions, which are analogous to thresholds) were randomly assigned values ranging from –0.50 to +0.50. The network was trained with a learning rate of 0.10. During each sweep of training, each of the 169 stimuli was presented to the network. The learning rule was used to update connection weights in the network after each stimulus presentation. Prior to each sweep of training, the order of stimulus presentation was randomized. Training proceeded until the network generated a “hit” for every output unit on every pattern. As is our typical practice, a hit was operationalized as an activation of 0.90 or higher when the desired activation was 1.00, and as an activation of 0.10 or lower when the desired activation was 0.00. All of these decisions were made on the basis of previous experience with training the value unit architecture.

8.3.2.3.3 Training The Network

The final step in synthesizing a connectionist network is to actually carry out the training, and create a network that is capable of generating the correct response for every pattern in the training set. Dawson, Boechler and Valsangkar-Smyth (2000) found that their spatial judgment network converged on a solution to the problem – generating a “hit” for each of the 169 patterns – after 10,907 sweeps of training. With this successful training, the research now turns to the concerns of emergence and analysis.

8.3.3 Connectionism And Emergence: A Prelude

In the robot examples of Chapters 6 and 7, after a robot was synthesized, the next step was to place it in an environment and observe its behavior. The point of this observation was to

identify interesting and surprising actions that emerged from the interaction between the robot and its world.

In the early stages of what is now known as New Connectionism, the immediate emphasis of research was also on emergence. As we will discuss in more detail in Section 8.3.4.1, connectionist researchers were interested in developing systems that processed information in a radically different manner than that used by classical or symbolic models. Symbolic models can be characterized as requiring a set of symbolic data structures and a set of explicitly defined rules for manipulating these data structures in memory (Newell, 1980). These rules were explicit, but not consciously available. In contrast, it was argued that connectionist models did not use explicit symbols or rules. "Parallel distributed processing models may provide a mechanism sufficient to capture lawful behavior, without requiring the postulation of explicit but inaccessible rules" (Rumelhart & McClelland, 1986, p. 218). PDP networks were viewed instead as dynamic data structures, responsible for storing information as well as transforming it (Hillis, 1985).

In order to support such claims, connectionist researchers took phenomena that were prototypical examples of symbolic cognitive science, and demonstrated that these phenomena could be mediated by PDP networks. For example, the regularities found in language are typically used as evidence to support the classical approach in cognitive science (Dawson, 1998). Pioneering connectionist researchers would choose to explore problems in language with their networks because such problems were firmly entrenched in the symbolic paradigm. In one influential instance, Rumelhart and McClelland (1986) trained a network to convert English verbs from the present tense into the past tense. This problem was explicitly selected because it exhibits "a phenomenon that is often thought of as demonstrating the acquisition of a linguistic rule" (p. 219). The network was in particular interesting because it went through three stages of development – patterns of successes and errors – that mimicked the stages that human children go through as part of their natural course of language development (but see Pinker & Prince (1988) for a critique of this view). They concluded that they had provided "a distinct alternative to the view that children learn the rules of English past-tense formation in any explicit sense."

Hanson and Olson (1991, p. 332) once noted "the neural network revolution has happened. We are living in the aftermath." At the time when the neural network revolution was in full swing, it was important to demonstrate that PDP models were capable of dealing with domains that were prototypically symbolic. I tell my students that this practice can be called "Gee Whiz connectionism", because its main goal was to allow researchers to exclaim "Gee whiz – PDP networks can do *x*, so *x* can be done without explicit rules." Classical researchers did take note of such results, acknowledging that it was surprising that models built from such simple components were capable of providing accounts of complex phenomena (Fodor & Pylyshyn, 1988).

However, in the aftermath of the neural network revolution, there really is no role for Gee Whiz connectionism. As is discussed in slightly more detail below, modern analyses have demonstrated conclusively that a broad variety of PDP architectures have the same computational power as the architectures that have been incorporated into symbolic accounts of cognition (Dawson, 1998). What this means is that a connectionist network can learn to perform any task that can be accomplished by a classical model. In the heyday of Gee Whiz connectionism, the mere demonstration that a network could do something of interest to classical cognitive science was by itself an emergent phenomenon of considerable interest. Now, with a better understanding of connectionist power, it is expected that networks can perform these tasks. As a result, the fact that a network can learn a task is no longer an emergent phenomenon of any interest to researchers.

Where, then, does emergence enter a synthetic psychology that uses PDP models? The answer to this question is that while it is neither interesting nor surprising to demonstrate that a network can learn a task of interest, it can be extremely interesting, surprising, and informative to determine what regularities the network exploits. What kinds of regularities in the input patterns has the network discovered? How does it represent these regularities? How are these regulari-

ties combined to govern the response of the network? In many instances, the answers to these questions can reveal properties of problems, and schemes for representing these properties, that were completely unexpected. In short, this means that before connectionist modelers can take advantage of the emergent properties of a PDP network that is being used as paradigm for synthetic psychology, the modelers must analyze the internal structure of the networks that they train. In the next section, we will consider several issues related to analyzing connectionist models. Later in this chapter we will return to the issue of emergent phenomena that are revealed from network analyses.

8.3.4 Connectionism And Analysis

In most cases, the identification of interesting emergent properties in a modern PDP network requires a detailed analysis of the internal structure of a trained network. In particular, after a network has learned to solve some problem of interest, a researcher will take the network apart and examine the properties of the internal representations that it has developed. In many cases, it is expected that this kind of analysis will reveal that the network has discovered interesting and surprising regularities in the problem. These surprises are one of the main ways in which connectionist simulations can push research in new directions.

However, if the analysis of connectionist representations is to provide a vehicle for synthetic psychology, then there are two general criticisms that have to be faced first. The first criticism is the general view that the kinds of representations that one will find in PDP networks are not the kinds of representations that will provide accounts of psychological phenomena. The second criticism is that even if these representations were of potential interest, they are nearly impossible to uncover in a trained network. We will consider each of these points below.

8.3.4.1 Connectionism And Representation

One major debate in cognitive science concerns potential differences (and similarities) between symbolic models and connectionist networks (Dawson, 1998). For example, it has been argued that, in contrast to symbolic theories, PDP networks are *subsymbolic* (Smolensky, 1988). To say that a network is subsymbolic is to say that the activation values of its individual hidden units do not represent interpretable features that could be represented as individual symbols. Instead, each hidden unit is viewed as indicating the presence of a *microfeature*. Individually, a microfeature is unintelligible, because its “interpretation” depends crucially upon its context (i.e., the set of other microfeatures which are simultaneously present (Clark, 1993)). However, a collection of microfeatures represented by a number of different hidden units can represent a concept that could be represented by a symbol in a classical model.

One consequence of the proposal that PDP networks use subsymbolic representations is further proposal that they process information in a completely different way than one would find in a symbolic model such as a production system. “Subsymbols are not operated upon by symbol manipulation: they participate in numerical – not symbolic – computation” (Smolensky, 1988). The kinds of numerical operations that are carried out are formal descriptions of the kind of energy minimization that we used to characterize the “thoughtless walkers” in Chapter 6. For example, Smolensky puts forth a “connectionist dynamical system hypothesis” as a proposed account of connectionist information processing. According to this hypothesis, at any state in time a connectionist network can be described as a vector of numbers, with each number representing the state of activity of a processing unit. In some instances, such as an account of learning, the vector might also include the values of a network’s weights. The system is dynamic, in the sense that this vector changes over time. Differential equations precisely describe such changes, which in most cases can be thought of as defining a trajectory in some multidimensional space through which the system travels to minimize some energy or cost value. For example, in Part II we will see several examples of learning rules that change a network’s vector state (i.e., its weights) over time to move the network into a state that minimizes an error term. We will also see examples of networks that modify the activity levels of their processors in an attempt to mini-

mize a mathematical function that describes a quantity that is analogous to energy. In short, the kind of numerical computation envisioned in connectionist networks by Smolensky is some form of statistical mechanics.

The claims that PDP networks represent and process information in completely different ways than symbolic models has led to strong criticisms about their role in cognitive science and psychology. Specifically, some researchers have made strong arguments that the kinds of (non-symbolic) representations that are found in connectionist models are not adequate to account for many of the regularities of human cognition (Fodor & McLaughlin, 1990; Fodor & Pylyshyn, 1988). In particular, Fodor and Pylyshyn argue that connectionist information processing does not involve a combinatorial syntax and semantics, and does not involve processes that are sensitive to constituent structure. They go on to argue that connectionist information processing shares many of the properties (and limitations) of the associationist theories that cognitivism reacted against in the 1950s (see also Bechtel, 1985). In short, their position is that connectionism doesn't provide the kind of representational account that psychology needs. "The problem with connectionist models is that all the reasons for thinking that they might be true are reasons for thinking that they couldn't be *psychology*" (Fodor & Pylyshyn, 1988, p. 66).

There are both theoretical and empirical reasons to believe that this dismissal of connectionism is premature. The symbolic paradigm in cognitive science is based upon the assumption that whatever the architecture of cognition is, it must have the computational power of a universal Turing machine (UTM) (Dawson, 1998). It would appear that connectionist networks also have this level of computational power. In some of the earliest work on neural networks, McCulloch and Pitts (1943) examined finite networks whose components could perform simple logical operations like AND, OR, and NOT. They were able to prove that such systems could compute any function that required a finite number of these operations. From this perspective, the network was only a finite state automaton (see also Hopcroft & Ullman, 1979; Minsky, 1972). However, McCulloch and Pitts went on to show that a UTM could be constructed from such a network, by providing the network the means to move along, sense, and rewrite an external "tape" or memory. "To psychology, however defined, specification of the net would contribute all that could be achieved in that field" (McCulloch & Pitts, 1943/1988, p. 25).

More modern results have validated and extended the pioneering research of McCulloch and Pitts (1943/1948). One common kind of recurrent Connectionist network has been popularized by Elman (e.g., 1990). In this network, there is a bank of "context units" that are used to remember the current activations of the output units. As a result, network output at time *t+1* is a function of input at time *t+1* and the network's previous output at time *t*. Williams and Zipser (1989) used this kind of network to construct the machine head of a UTM. This network learned to use five different output neurons to perform the basic operations of a Turing machine (e.g., move left, no change, write "1", write "0", and move right) on a tape that was used to activate a single input processor. Several formal analyses of this kind of network have also been performed. One theme of this work has been to determine whether or not in principle one could build a finite network to perform the computations of a UTM (Siegelmann & Sontag, 1991; Siegelmann, 1999; Siegelmann & Sontag, 1995). Early work developed a proof that such a network was possible in principle, but this proof limited the absolute size of this network to a relatively large value (a maximum of 10^5 processing units). Later research refined this result, and proved that Minsky's (1972) well known 4-symbol, 7-machine state UTM could be built from a recurrent network that used 1058 processing units. Kilian and Siegelmann (1993) have developed a general proof that recurrent networks of the type used by Elman (e.g., 1990) are indeed equivalent to Turing machines. They concluded that "Turing universality is a relatively common property of recurrent neural network modes" (p. 137).

Empirical evidence also supports the view that the distinction between connectionist and classical models is fairly blurred. For example, in one study (Dawson et al., 1997) my students and I trained a network of value units on a logic problem developed by Bechtel and Abrahamsen (1991). When we analyzed the internal structure of the network, we found evidence for network

states that represented standard rules of logic. A second study provided even stronger evidence of the representational equivalence of the two types of models. Dawson, Medler et al. (2000) used a technique called extra-output learning to train a network to classify mushrooms as being either edible or poisonous on the basis of a vector of feature descriptors. The network was trained to use its extra output units to assert a "reason" for making a particular judgment, where the "reason" was taken from a standard decision tree that could also be used to classify the mushrooms. When the internal states of the network were examined, they found that the network's representations could be reduced to a small number of state vectors, where each vector defined a pattern of activity over the network's hidden units. It turned out that each state vector mapped precisely onto a production from a production system that was created by translating the decision tree. In other words, they found that when the network's hidden units were in a particular state of activity, this was equivalent to saying that the network was executing a particular production. The extra-output learning technique had essentially translated a symbolic representation directly into a connectionist network.

8.3.4.2 Connectionism And Bonini's Paradox

In the section above, we briefly reviewed the criticism that connectionist representations are not appropriate for psychology, and presented theoretical and empirical arguments against this position. It would appear, then, that examining the internal representations of PDP networks is an appropriate activity for synthetic psychology. Unfortunately, connectionist researchers freely admit that it is extremely difficult to determine how their networks accomplish the tasks that they have been taught. "If the purpose of simulation modeling is to clarify existing theoretical constructs, connectionism looks like exactly the wrong way to go. Connectionist models do not clarify theoretical ideas, they obscure them" (Seidenberg, 1993, p. 229). There are a number of reasons that PDP networks are difficult to understand as algorithms, and are thus plagued by what we called Bonini's paradox in Chapter 2.

First, general learning procedures can train networks that are extremely large; their sheer size and complexity makes them difficult to interpret. For example, Seidenberg and McClelland's (1989) network for computing a mapping between graphemic and phonemic word representations uses 400 input units, up to 400 hidden units, and 460 output units. Determining how such a large network maps a particular function is an intimidating task. This is particularly true because in many PDP networks, it is very difficult to consider the role that one processing unit plays independent from the role of the other processing units to which it is connected (see also Farah, 1994).

Second, most PDP networks incorporate nonlinear activation functions. This nonlinearity makes these models more powerful than those that only incorporate linear activation functions (e.g. Jordan, 1986), but it also results requires particularly complex descriptions of their behavior. Indeed, some researchers choose to ignore the nonlinearities in a network, substituting a simplified (and often highly inaccurate) qualitative account of how it works (e.g., Moorhead, Haig, & Clement, 1989).

Third, connectionist architectures offer (too) many degrees of freedom. One learning rule can create many different networks -- for instance, containing different numbers of hidden units – that can each compute the same function. Each of these systems can therefore be described as a different algorithm for computing that function. One does not have any *a priori* knowledge of which of these possible algorithms might be the most plausible as a psychological theory of the phenomenon being studied.

8.3.4.3 Interpreting Connectionist Networks

Difficulties in understanding how a particular connectionist network accomplishes the task that it has been trained to perform has raised serious doubts about the ability of connectionists to provide fruitful theories about cognitive processing. Because of the problems of network interpre-

tation, McCloskey (1991) suggested that “connectionist networks should not be viewed as theories of human cognitive functions, or as simulations of theories, or even as demonstrations of specific theoretical points” (p. 387). Fortunately, connectionist researchers are up to this kind of challenge. Several different approaches to interpreting the algorithmic structure of PDP networks have been described in the literature.

One approach to interpreting a network’s algorithm involves studying its connection weights. For example, Hanson and Burr (1990) review a number of techniques for doing this, including compiling frequency distributions of connection strengths, quantifying global patterns of connectivity with descriptive statistics, illustrating local patterns of connectivity with “star diagrams”, and performing cluster analyses of hidden unit activations. Hinton (1986) provides an excellent example of how an examination of connection weights in a trained network can reveal the regularities that the network is using to solve a difficult pattern recognition task.

A second approach involves using algorithms to translate the pattern of connections in a network into a production rule that describes how the response of one processing unit depends on the responses of a subset of the processing units that feed into it (e.g., Gallant, 1993, Chap. 17)). In general, this is problematic, because as one increases the number of connection weights in the network, the number of possible productions encoded in a network increases exponentially. As a result, the researcher must use additional assumptions about the nature of the rules in order to help constrain the interpretation (i.e., to help limit the number of proposed productions).

A third strategy is to map out the response characteristics of each processor in the network. For example, Moorhead et al. (1989) used the generalized delta rule to train a PDP network to identify the orientation of line segments presented to an array of input units. Their primary research goal was to determine whether the hidden units in this system developed center-surround receptive fields analogous to those found in the primate lateral geniculate nucleus. They chose to answer this question by stimulating each input element individually, and plotting the resulting activation in each hidden unit.

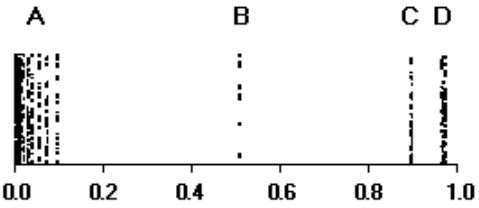

Figure 8-3. An example jittered density plot for a hidden unit from a network of value units.

In my own lab, my students and I have focused on interpreting the structure of networks of value units, which is one of the reasons that Dawson, Boechler and Valsangkar-Smyth (2000) decided to use such units in the spatial judgment network that was described above. Value unit networks have many advantages over traditional PDP networks (Dawson, 1990; Dawson & Schopflocher, 1992; Evans, 1989; Medler & Dawson, 1994a, 1994b; Shamanski & Dawson, 1994). In our earlier research, and as was briefly mentioned in Section 2.2.5.2, we discovered one property of value units that permits the internal structure of a trained network to be interpreted (Berkeley et al., 1995; Dawson, Berkeley, Medler, & Schopflocher, 1994). After a value unit network is trained to solve a problem, the training set is re-presented to the network, and we record the activity produced in each hidden unit by each pattern. These data are then graphed in *jittered density plots* (Chambers, Cleveland, Kleiner, & Tukey, 1983). One density plot is drawn for each hidden unit, and each "dot" in a plot represents the unit’s response to one of the patterns. We discovered that such plots for value units are usually very structured – they are organized into distinct bands (Figure 8-3). Furthermore, all the patterns falling into the same band in the plot share particular features. By statistically describing the properties of the patterns that fall into each band, we were able to identify the input features detected by each hidden unit. We were also able to establish how these features are combined to yield a network’s response to a stimulus (Dawson et al., 1997).

Our early research was quite successful in developing value unit networks that produced banded jittered density plots, and in interpreting the internal structure of the network by focusing

on the local properties of individual hidden units – namely, the features associated with individual bands. One of my former students has used this finding to argue that bands are subsymbolic, in the sense of Smolensky (1988), but that subsymbols are more like symbols than one might expect at first (Berkeley, 2000).

Our later research has brought such conclusions into question. In many cases, we have found networks with hidden units that demonstrate marked banding, but the local interpretations of the bands are of little use (Dawson & Piercey, 2001). Instead, the interpretation of how the network is solving the problem requires considering the content of one band in the context of other bands in other hidden units. In other words, the interpretation of the network depends on examining distributed representations.

For example, Dawson, Medler et al. (2000) used a training technique called extra output learning to insert a symbolic theory into a network of value units. When the network was interpreted (to verify theory insertion), a precise relationship between the network and the symbolic theory only emerged when cluster analysis was performed on the basis of activations across all the hidden units in the network. Similarly, Leighton and Dawson (2001) trained a network of value units to solve the (Wason, 1966) card selection task, in which a subject must select cards to be examined to test a logical argument. In one version of the network trained to generate the logically correct responses to the task, pairs of hidden units cooperated to activate an output unit (i.e., to select a card for examination). The network's behavior could not be understood by examining individual hidden units. Finally, Zimmerman (1999) trained a network of value units to generate solutions to the balance scale task used to study cognitive development (Inhelder & Piaget, 1958). In this task, the network is presented a configuration of weights on either side of a balance scale, and has to judge whether the scale would tip to the left, tip to the right, or balance. Zimmerman found that the network solved this task by performing a function approximation that required a coarse-coding combination of activities from all of the network's hidden units. In all three of these examples, banding was found for the hidden units of the networks. However, the local interpretations of these bands could not adequately explain the network's behavior. Interpretation was possible, but only after considering regularities in hidden unit activity when the hidden units were considered simultaneously.

This finding is exactly in accordance with the standard view of subsymbolic representations. According to Smolensky (1988), subsymbols are constituents of traditional symbols. "Entities that are typically represented in the symbolic paradigm are typically represented in the subsymbolic paradigm by a large number of subsymbols" (p. 3). As a result, "it is often important to analyze connectionist models at a higher level; to amalgamate, so to speak, the subsymbols into symbols". Of course, the problem is that in some instances the internal network structure might be fairly local. What this means is that when a network is being interpreted, one has to be open to the possibility of either a local or a distributed interpretation.

There are several conclusions that can be drawn from the research that has been conducted in my lab. First, it is possible to interpret and understand the internal representations of multi-layer connectionist networks. In some instances, a local interpretation approach works; in others, one has to examine regularities distributed across hidden units. Second, no one interpretative strategy will work every time. Network analysis requires a great deal of patience, and the ability to consider a variety of statistical approaches to network analysis is definitely an asset. Third, and a consequence of the first two points, network analysis is not easy. This is likely one reason that detailed network interpretations are rarely found in the literature. Nevertheless, network interpretation is both necessary and possible. This is demonstrated in the next section with a detailed description of how Dawson, Boechler and Valsangkar-Smyth (2000) "cracked" the internal structure of the spatial judgment network.

8.3.5 Connectionism And Analysis: An Example

To provide an example of connectionism and analysis, let us return to the spatial judgment network of Dawson, Boechler and Valsangkar-Smyth (2000). After they were able to successfully construct the network, their interest turned to the kinds of internal representations used by the network to generate its metric behavior. In what way do the hidden units of this network represent the metric structure of a two-dimensional map of Alberta? Have the hidden units developed a metric representation of space? Or have the hidden units instead developed some complex nonmetric representation from which metric behavior can be derived?

8.3.5.1 Relating The Map Of Alberta To Hidden Unit Connection Weights

Dawson, Boechler and Valsangkar-Smyth (2000) began by exploring the possibility that the network might have developed internal representations similar in nature to those that have been attributed to cells in the hippocampus. For example, consider the possibility that each hidden unit occupies a position in the map of Alberta, and uses its connection weights to represent the distances from the hidden unit to each of the Albertan cities. If this hypothesis is correct, then one should be able to find a position for each hidden unit on the map of Alberta such that there is a substantial correlation between the unit's connection weights and the distances from each city to the hidden unit location.

Dawson, Boechler and Valsangkar-Smyth (2000) used the Solver tool in Microsoft Excel to move each hidden unit to a latitude and longitude on the map of Alberta. The spreadsheet that they designed computed the distance between the (current) position of the hidden unit and each of the 13 Albertan cities. The spreadsheet then computed the correlation between these 13 distances and the 13 connection weights feeding into the hidden unit. The Solver tool in the spreadsheet then changed the position of the hidden unit, finally stopping when it identified the position on the map that produced the highest correlation between map distances and connection weights. There was a very strong relationship between distances and connection weights. The correlations were 0.88 for H0, 0.59 for H1, 0.72 for H2, -0.54 for H3, 0.79 for H4, and –0.48 for H5. Figure 8-3 illustrates the positions of the six hidden units on the map that produced these correlations.

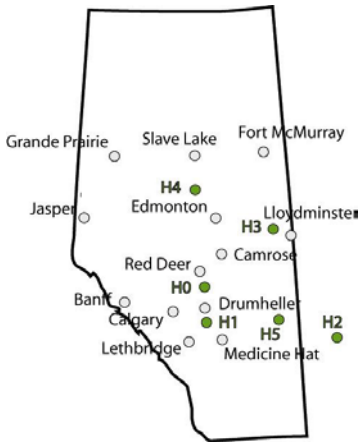

Figure 8-3. The positions of the six hidden units on the map of the Alberta.

8.3.5.2 Relating Connection Weights To Hidden Unit MDS Spaces

Dawson, Boechler and Valsangkar-Smyth's (2000) first analysis indicated that each hidden unit could be viewed as occupying a position on the map of Alberta, and that its connection weights were related to distances between the hidden unit and the 13 cities on the map. However, while the correlation between map distances and connection weights were substantial, they were not as strong as might be expected. They noted that one problem with the first analysis was that it imposes our notion of the space in question (i.e., the map of Alberta) onto the behavior of the hidden units. It does not permit the possibility that the hidden units are spatial, but the space that they are sensitive to is quite different from the map in Figure 8-2. There are at least two reasons to expect that the hidden units have a distorted representation of the map.

The first reason is theoretical. If connection weights leading into a hidden unit represent distance, then these distances are dramatically transformed by the Gaussian activation function of the hidden unit when connection weight signals are converted into hidden unit activity. This kind of transformation would be equivalent to a distortion of the Figure 8-2 map.

The second reason is empirical. For any input pattern, a hidden unit's activity can be viewed as being analogous to that hidden unit's rating of the distance between cities. If we exam-

ine hidden unit activity to various pairs of cities, then we can see that the hidden unit's "ratings" do not seem particularly accurate. Consider, for example, hidden unit 2. When the network is asked to rate the distance between Red Deer and Jasper, this unit generates an activation value of 0.69. On the map of Alberta, the distance between Jasper and Red Deer is 413 km. However, nearly identical behavior is produced in the unit by two other cities, Edmonton and Lloydminster, which are much closer together on the map (251 km). When these two cities are presented to the network an activation of 0.71is produced in hidden unit 2.

If the hidden units are spatial in nature, but are dealing with a space that is quite different from the one that we might expect (i.e., Figure 8-2), then how should their behavior be analyzed? One approach would be to consider each hidden unit as being a subject in a distance rating experiment. For each stimulus, the rating generated by the hidden unit is the hidden unit's activity. If all of these ratings are taken and organized them into a table like Table 8-2, then MDS can be applied to this data. This analysis determines the structure of the space that underlies the hidden units behavior, which can then be related to the connection weights that feed into each unit.

Dawson, Boechler and Valsangkar-Smyth (2000) performed this analysis on the 13 X 13 "activity matrix" for each hidden unit, in which each row and each column corresponded to an Albertan city, and each matrix entry a_{ij} was the hidden unit's activation value when the network was asked to rate the distance between city i and city j. They found that a two-dimensional plot provided a nearly perfect account of the activity matrix of each hidden unit. They then repeated the analysis that was reported in section 8.3.5.2.1. However, instead of using the map of Alberta, for each hidden unit they used the coordinates of the cities obtained from the MDS analysis of the unit's activity matrix. With these analyses, for each hidden unit we found a location in the MDS space that produced a near perfect correlation between distances and connection weights, as is reported in Table 8-3.

Hidden Unit	X-Coordinate	Y-Coordinate	Correlation Between Distances And Weights
H0	1.53	-1.88	-0.95
H1	-0.07	0.10	0.96
H2	-1.33	0.31	-0.88
H3	-0.13	-0.14	0.93
H4	0.07	0.16	-0.95
H5	3.20	0.19	-0.96

Table 8-3.
Results of relating connection weights to city distances from the MDS solutions obtained from the activity matrix for each hidden unit. The table provides the maximum correlation, as well as the coordinates of the hidden unit in the space that produces this maximum correlation.

8.3.5.3 Coarse Coding From Hidden Unit Activations To Distance Ratings

Up to this point, we have seen evidence that the spatial judgment network developed a spatial representation of map locations, in which the weights that fed into a hidden unit encoded information about the distance between the hidden unit's position in a 2D space and city locations in the same space. However, we have not yet discussed how the network exploits the features detected by the hidden units to produce the desired ratings as output.

We saw earlier that an individual unit's responses to different stimuli were not necessarily accurate. For instance, when presented two cities that were relatively close together, a unit might generate internal activity very similar in value to that generated when presented two other cities that were much further apart. To verify this claim quantitatively, Dawson, Boechler and Valsangkar-Smyth (2000) took the activity of each hidden unit and correlated it with the desired rating for the input patterns. For units H0 through H5, these correlations were –0.32, 0.04, 0.04, -0.10,

0,04, and 0.16. It would appear that the activities of individual hidden units were at best weakly related to the desired distance ratings. How is it possible for such inaccurate responses to result in accurate outputs from the network?

The answer to this question is that the hidden unit activations in the network are a form of representation called *coarse coding*. In general, coarse coding means that an individual processor is sensitive to a broad range of features, or at least to a broad range of values of an individual feature (e.g., Churchland & Sejnowski, 1992). As a result, individual processors are not particularly useful or accurate feature detectors. However, if different processors have overlapping sensitivities, then their outputs can be pooled, which can result in a highly useful and accurate representation of a specific feature. Indeed, the pooling of activities of coarse-coded neurons is the generally accepted account of hyperacuity, in which the accuracy of a perceptual system is substantially greater than the accuracy of any of its individual components (e.g., Churchland & Sejnowski, 1992).

The coarse coding that is used in the spatial judgment network can be thought of as follows: Each hidden unit occupies a different position on the map of Alberta. When presented a pair of cities, each unit generates an activation value that reflects a rough estimate of the combined distance from the two cities to the hidden unit. While each hidden unit by itself generates only a rough estimate, when all six hidden units are considered at the same time, a much more accurate estimate of the distance between the two cities is possible. To demonstrate this, Dawson, Boechler and Valsangkar-Smyth (2000) used multiple linear regression to predict the distance rating (an integer ranging from 0 to 10) from the activations generated in 6 of the hidden units by each of the 169 stimuli that were presented to the network during training. The regression equation produced an R^2 of 0.71 ($F[6,163] = 66.81, p < 0.0001$). In other words, a linear combination of the hidden unit activities can by itself account for over 70% of the variance of the distance ratings. After being trained to solve the problem, the network, in virtue of the nonlinear transformations performed by the Gaussian activation functions of its output units, can combine the hidden unit activities to account for 100% of the distance ratings.

8.3.6 Connectionism And Emergence: An Example

Several different analyses of the internal structure of the spatial judgment network were reported above, and all of these analyses converged on one general finding: the hidden units of the network developed metric representations of space. First, two-dimensional MDS analyses accounted for almost all of the variance in the activation matrix that was created for each hidden unit. Second, if one assumed that each hidden unit occupied a location on the map of Alberta, one could find a location for each hidden unit that produced a high correlation between the connection weights feeding into the hidden unit and the distances on the map between cities and the position of the hidden unit. Third, if one replaced the map of Alberta with a customized two-dimensional space for each hidden unit (a space revealed by the MDS analyses), near perfect correlations between connection weights and distances in the space were revealed.

With these analyses completed, we can now return to the issue of connectionism, analysis, and emergence. Specifically, now that a spatial judgment network has been synthesized, and now that its internal structure has been thoroughly analyzed, what are the implications of this simulation? Dawson, Boechler and Valsangkar-Smyth (2000) discussed two general insights that were provided by their research. The first had to do with a controversy about how the hippocampus represents space. The second had to do with the relationship between metric representations and nonmetric behaviors. These two issues are discussed in the sections that follow.

8.3.6.1 Implications For The Hippocampal Cognitive Map

The strong interest that neuroscientists have taken in the study of spatial behavior and cognitive maps can largely be traced back to the discovery of place cells in the hippocampus (O'Keefe & Dostrovsky, 1971). The properties of place cells have been used as evidence for the

neural basis of a cognitive map in the hippocampus (O'Keefe & Nadel, 1978). This map was argued to be a Euclidean description of the environment based on an allocentric frame of reference. In other words, locations in this map were defined in terms of the world, and not in terms of a coordinate system based upon (and moving with) the animal. Additional support for this proposal came from the fact that lesions to the hippocampus produce deficits in a variety of spatial tasks (for an introduction, see Sherry & Healy, 1998). Furthermore, robots that use a representational scheme based upon the properties of place cells can navigate successfully in their environment, indicating that the place cell architecture is a plausible proposal for a cognitive map (Burgess, Donnett, Jeffery, & O'Keefe, 1999).

One common analogy used by researchers is that a cognitive map is like a graphical map (Kitchin, 1994). "This does not mean that there must be a region in the brain onto which the environment is physically mapped, but rather that there will be a correspondence between input-output behaviors of the storage and retrieval functions of the two representations" (p. 4). The aforementioned properties of place cells would appear to support this analogy. One might plausibly expect that the cognitive map is a two-dimensional array in which each location in the map (i.e., each place in the external world) is associated with the firing of a particular place cell.

However, anatomical evidence does not support this analogy. First, there does not appear to be any regular topographic organization of place cells relative to either their positions within the hippocampus or to the positions of their receptive fields with respect to the environment (Burgess et al., 1995; McNaughton et al., 1996). Second, place cell receptive fields are at best *locally* metric (Touretzky et al., 1994). This is because one cannot recover information about bearing from place cell representations, and one cannot measure the distance between points that are more than about a dozen body lengths apart because of a lack of place cell receptive field overlap. Some researchers now propose that the metric properties of the cognitive map emerge from the coordination of place cells with cells that deliver other kinds of spatial information, such as head direction cells which fire when an animal's head is pointed in a particular direction, regardless of the animal's location in space (McNaughton et al., 1996; Redish & Touretzky, 1999; Touretzky et al., 1994).

Dawson, Boechler and Valsangkar-Smyth (2000) observed that the hidden units in the spatial judgment network also appear to be subject to the same limitations that have brought into question the ability of place cells to provide a metric representation of space. First, because the hidden units were all connected to all of the input units, the network had no definite topographic organization. Second, each hidden unit appeared to be at best locally metric. While the input connections were correlated with distances on the map, the responses of individual hidden units did not provide an accurate spatial account of the map. Nevertheless, the fact that the network could be trained to accurately generate the ratings indicated that the responses of these locally metric, inaccurate processors represented accurate spatial information about the entire map of Alberta. This was possible because the network did not base its output on the behavior of a single hidden unit. Instead, it relied on coarse coding, and generated its response on the basis of the activities of all six hidden units considered simultaneously.

Dawson, Boechler and Valsangkar-Smyth (2000) noted that one implication of this coarse coding is that spatial relationships amongst locations in Alberta can be captured by a representational scheme that is not isomorphic to a graphical map. In particular, if one views the hidden units as being analogous to place cells, then the network demonstrates that spatial relationships among 13 different landmarks can be represented by a system which assigns place cells to only 6 different map locations.

The reason that this is possible is because the representational scheme discovered by the network is allocentric, but in a fashion that might not be immediately expected. Taken literally, the term allocentric means "centered on another", but there are at least four distinct kinds of representations for which this would be true (Grush, 2000). In two of these, the locations of objects are either specified with respect to one object in the environment (an object-centered refer-

ence frame) or with respect to a position in the environment at which no object is located (a virtual or neutral point of view). The representation used in the PDP network is allocentric in this latter sense, because the positions of cities are represented relative to the positions of hidden units, and the hidden units are not positioned at city locations. However, the network representation extends this notion of allocentric, because city locations are not encoded with respect to a single virtual location, but instead with respect to a set of six different virtual positions, all of which have to be considered at the same time to accurately retrieve spatial information from the network (i.e., to judge the distance between cities). Dawson, Boechler and Valsangkar-Smyth (2000) called this a *coarse allocentric code*.

The major hypothesis about the hippocampus that was suggested by the spatial judgment network is that place cells also implement a coarse allocentric code. As a result, the place cells need not be organized topographically, because they don't represent the environment in the same way as a graphical map. Instead, locations of landmarks in the environment could be represented as a pattern of activity distributed over a number of different place cells. If this were the case, then in spite of their individual limitations, coarse coding of place cell activities could be used to represent a detailed cognitive map without necessarily being coordinated with other neural subsystems. In other words, Dawson, Boechler and Valsangkar-Smyth's (2000) discovery of coarse allocentric coding in their network provides one plausible manner in which the spatial abilities of the hippocampus can be reconciled with its non-maplike organization.

8.3.6.1 Coarse Allocentric Coding And Nonmetric Judgments

The spatial nature of the network's internal representations is perhaps not surprising, given that the network was trained to internalize a metric space. However, as was noted earlier in this chapter, there does exist a tension between the metric properties of a representation and the properties of the behavior that the representation mediates. Specifically, is it possible for a metric representation to mediate nonmetric behavior?

This issue is important, because the discovery that human similarity judgments were nonmetric had a severe impact on proposals about the representations that mediated this behavior. Tversky and his colleagues conducted a number of experiments that demonstrated that similarity judgments were not metric, because in different situations it could be shown that these judgments were not always symmetric, did not always conform to the minimality principle, and did not always conform to the triangle inequality (Tversky, 1977; Tversky & Gati, 1982). As a result, many researchers completely abandoned the notion of the similarity space, and instead moved to feature based comparison models that could easily handle nonmetric regularities. This was in spite of the fact that it is possible to elaborate a perfectly metric representational space in such a way that it can be used to mediate nonmetric judgments. For example, Krumhansl (1978, 1982) demonstrated that if one took a metric space and augmented the kind of operations that were applied to it one could easily account for asymmetric similarity judgments.

Dawson, Boechler and Valsangkar-Smyth's (2000) discovery of the coarse allocentric code was exciting because it raised the possibility of a metric representation that might be flexible enough to mediate spatial judgments that were not completely metric. In other words, they were interested in the possibility that coarse allocentric coding could support nonmetric judgments without the need for additional rules or processes.

One of the reasons for the rise in the popularity of connectionist networks over symbol-based models is that network models degrade gracefully and are damage resistant (McClelland, Rumelhart, & Hinton, 1986). To say that a network degrades gracefully is to say that as noise is added to its inputs, its output responses become poorer, but it does not stop responding (Dawson, 1998). The model deals as best it can with less than perfect signals. To say that a network is damage resistant is to say that as noise is added to its internal structure (e.g., by damaging connections or by ablating hidden units), its output responses become poorer, but it still

functions as well as it can. Traditional symbol-based models do not degrade gracefully, and are not damage resistant.

The damage resistance and graceful degradation of PDP networks is due to the redundancy of their internal representations when they employ coarse coding. One further advantage that this kind of representation can provide, which is related to graceful degradation, is generalization. When presented with a new stimulus – one that the network was never trained on – a network often can generate a plausible response, taking advantage of the similarity between the new stimulus and old stimuli, and the fact that such similarity can be easily exploited in redundant representations. In fact, if too many hidden units are used, and if these units start to pay attention to specific stimuli, then generalization will be poorer. This is one aspect of what is called "the three bears" problem (Seidenberg & McClelland, 1989).

In a second simulation, Dawson, Boechler and Valsangkar-Smyth (2000) were concerned with a different type of generalization – the generalization of representation type from one problem to another. Specifically, imagine if the spatial judgment network's task was changed in such a way that the distance ratings violated one of the metric properties of space. Could allocentric coarse coding still be used to represent a solution to the problem? Or would a change in task result in a completely different representational approach?

The problem that Dawson, Boechler and Valsangkar-Smyth (2000) trained a network to solve in the second simulation was a distance estimation task that was identical to the one that we have described above, with the exception that the network was trained to make different judgments when asked to judge the distance between a city and itself. In the first simulation, such judgments obeyed the minimality principle of metric space, and the network was trained to make a judgment of 0 when presented such stimuli. In the second simulation, the minimality principle was violated. Instead of making a judgment of 0 when rating the distance of a city to itself, the network was trained to make a rating of 0, 1, or 2 depending upon the city.

Figure 8-4. The position of the seven hidden units in the second spatial judgement network.

When the minimality constraint was violated in this way, Dawson, Boechler and Valsangkar-Smyth (2000) found that the ratings task became more difficult. In particular, the problem could not be solved when the network had six hidden units. An additional hidden unit was required. When seven hidden units were used in the network, it learned to solve the problem in 2057 learning epochs.

In spite of the task being more difficult, though, there was no evidence that the network created a qualitatively different representation to solve the problem. Dawson, Boechler and Valsangkar-Smyth (2000) analyzed this second network in the same fashion that they used to analyze the first network, and which was described above. They found that the second network used allocentric coarse coding to make distance judgments. Each hidden unit could be considered as occupying a position on the map of Alberta, and the weights feeding into each unit were correlated with the distances between the hidden units and the Albertan cities. The responses of individual hidden units provided relatively inaccurate sensitivity to distance information. However, when the responses of all seven hidden units were pooled, very accurate distance judgments were possible. Finally, and most importantly, there was no evidence that any one of the hidden units had a special role in making the subset of judgments that defined the violation of the minimality principle.

In particular, one possibility that Dawson, Boechler and Valsangkar-Smyth (2000) considered was that six of the hidden units in the new network were performing the same function as were the six in the first network, and that the seventh hidden unit was a special purpose unit designed to deal with the nonmetric judgments taken from the diagonal of the new ratings matrix. This was not the case – all seven hidden units could be described in the same general way, and all seven were involved in coarse allocentric coding. Figure 8-4 illustrates the positions of the seven hidden units of this network on the map of Alberta.

8.3.6.3 Implications

Earlier in this chapter, we briefly considered three different research areas related to spatial cognition: similarity spaces, mental imagery, and cognitive maps. For each of these areas, it was argued that there existed a tension between behavioral regularities and representational properties. For example, consider the relationship between similarity judgments (which are strongly related to the distance judgments used in the current study) and representational proposals. In the beginning, similarity judgments were assumed to obey the metric properties of space, and as a result researchers proposed that these judgments were mediated by a metric spatial representation (Romney et al., 1972; Shepard et al., 1972). However, later research revealed that the judgments that subjects made were not always metric. What were the representational implications due to these behavioral observations?

One alternative was to completely abandon metric spatial representations, and to adopt representations that were less structured. For example, some researchers replaced the similarity space with a proposal in which concepts were represented as sets of features, and nonmetric behavioral regularities emerged from the procedures used to compare feature sets (Malgady & Johnson, 1976; Ortony, 1979; Tversky, 1977; Tversky & Gati, 1982). This approach has the advantage of being able to account for nonmetric behavioral regularities. However, it has disadvantages as well. The ability to fit nonmetric behavior emerges from manipulating constants in feature comparison equations. These constants provide additional degrees of freedom that must be fit from study to study to predict human judgments. Because of these additional degrees of freedom, this kind of theory is less powerful -- less constrained -- than the similarity space that it replaced (Pylyshyn, 1984).

A second alternative was to modify the similarity space proposal in such a way that this metric space could mediate nonmetric behaviors. For instance, Krumhansl (1978, 1982) modified the similarity space by including new rules that measured the density of points in the space, where density reflected the number of neighbors that were close to a point in the space. Krumhansl included density calculations in addition to distance in the rules that were used to compare different points in the space. The inclusion of density permitted nonmetric judgments to emerge from the space. This approach has the advantage of maintaining some of the attractive properties of the similarity space. However, the density calculations also introduce new degrees of freedom that reduce the explanatory power of theory.

A third example is provided by the synthetic approach taken by Dawson, Boechler and Valsangkar-Smyth (2000). A model based on relatively simple building blocks, with few underlying representational hypotheses, was trained to generate metric spatial judgments. Once the model had been synthesized, they took great pains to analyze its internal structure. The result was the discovery of a particular kind of representation, allocentric coarse coding, that would not have been an obvious proposal had our starting point been the analysis of behavior. A second study demonstrated that this kind of representation was also capable of mediating spatial judgments that violated the minimality principle of metric space. In other words, the synthetic approach utilized by Dawson, Boechler and Valsangkar-Smyth (2000) has shown how a connectionist representation can account for both metric and nonmetric regularities.

8.4 SUMMARY AND CONCLUSIONS

In Chapters 6 and 7, the synthetic approach was illustrated with examples that used robots, toy and otherwise. Much of this research, which is now known as behavior-based robotics and embodied cognitive science, is aimed at challenging the assumption that cognition and intelligence is based upon mental representations. While it is of considerable interest that many complicated behaviors can be produced by systems that only exploit visuomotor reflexes, many domains of cognitive science and psychology are still likely to need to appeal to representations. One question addressed in this chapter was whether the synthetic approach could be employed in a fashion that still permitted representations to be explored.

It was argued in this chapter that PDP models offered one plausible method for conducting psychological research that was both synthetic and representational. The synthetic component of this kind of research involves using components defined by a selected connectionist architecture to construct a (multi-layer) network capable of solving some problem of interest. Once the network has been constructed, its internal structure is analyzed in detail. The purpose of this analysis is to discover the regularities in the training patterns that are used by the network to solve the problem, as well as the manner in which these regularities are represented in the network's connections. Once this analysis is complete, it is expected that the discovered regularities and representations will lead to unexpected insights into the problem. In other words, in a connectionist synthetic psychology emergence will follow analysis.

This chapter also presented one case study in synthetic psychology, the spatial judgment network of Dawson, Boechler and Valsangkar-Smyth (2000). One reason for choosing this example was to show that fairly simple components could be used to construct a system capable of performing a task of psychological interest. A second reason for this example was to illustrate an instance of "representational emergence". When the spatial judgment network was originally created, the only general issue in mind was building a PDP system that could respond as if it had internalized a spatial map. We were interested in identifying how such a map was internalized, but had no pet theory about its structure. At the end of the analysis, when we had identified the coarse allocentric coding in the hidden units, we found that we had something to say about spatial representation in the hippocampus and about the ability to generate judgments that were nonmetric. These insights were surprising to us, and demonstrate some of the power that can emerge from adopting a synthetic paradigm.

We have now come to the end of Part I of this book. We have discussed different types of models, and have contrasted analytic and synthetic approaches. We have ended with a case study that shows how connectionism can contribute to synthetic psychology. In Part II of the book, we will turn to a more careful study of different connectionist architectures, and use this information to get a richer sense of what a connectionist synthetic psychology might look like.

8.5 REFERENCES

Anderson, J. R., & Bower, G. H. (1973). *Human Associative Memory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bannon, L. J. (1980). *An investigation of image scanning.* Unpublished Unpublished doctoral dissertation, University of Western Ontario, London, ON.

Bechtel, W. (1985). Contemporary connectionism: Are the new parallel distributed processing models cognitive or associationist? *Behaviorism, 13*, 53-61.

Bechtel, W., & Abrahamsen, A. (1991). *Connectionism And The Mind*. Cambridge, MA: Basil Blackwell.

Berkeley, I. S. N. (2000). What the #$*%! is a subsymbol? *Minds and Machines, 10*, 1-13.

Berkeley, I. S. N., Dawson, M. R. W., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden value unit activations reveal interpretable bands. *Connection Science, 7*, 167-186.

Bever, T. G., Fodor, J. A., & Garrett, M. (1968). A formal limitation of associationism. In T. R. Dixon & D. L. Horton (Eds.), *Verbal Behavior And General Behavior Theory* (pp. 582-585). Englewood Cliffs, NJ: Prentice-Hall.

Block, N. (1981). *Imagery*. Cambridge, MA: MIT Press.
Blumenthal, L. M. (1953). *Theory And Applications Of Distance Geometry*. London: Oxford.
Boechler, P. M. (2001). How spatial is hyperspace? Interacting with hypertext documents: Cognitive processes and concepts. *Cyberpsychology & Behavior, 4*, 23-46.
Boechler, P. M., & Dawson, M. R. W. (2001). The effects of navigational tool information on hypertext navigation behavior: A configural analysis of page-transition data. *Journal of Educational Multimedia and Hypermedia, in press*.
Braitenberg, V. (1984). *Vehicles: Explorations In Synthetic Psychology*. Cambridge, MA: MIT Press.
Brooks, R. A. (1989). A robot that walks; emergent behaviours from a carefully evolved network. *Neural Computation, 1*, 253-262.
Brooks, R. A. (1999). *Cambrian Intelligence: The Early History Of The New AI*. Cambridge, MA: MIT Press.
Burgess, N., Donnett, J. G., Jeffery, K. I., & O'Keefe, J. (1999). Robotic and neuronal simulation of the hippocampus and rat navigation. In B. N. & K. J. Jeffery & J. O'Keefe (Eds.), *The Hippocampal And Parietal Foundations Of Spatial Cognition*. Oxford: Oxford University Press.
Burgess, N., Recce, M., & O'Keefe, J. (1995). Spatial models of the hippocampus. In M. A. Arbib (Ed.), *The Handbook Of Brain Theory And Neural Networks*. Cambridge, MA: MIT Press.
Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphic methods for data analysis*. Belmont, CA: Wadsworth International Group.
Cheng, K., & Spetch, M. L. (1998). Mechanisms of landmark use in mammals and birds. In S. Healy (Ed.), *Spatial Representation In Animals*. Oxford: Oxford University Press.
Chomsky, N. (1959). A review of B.F. Skinner's Verbal Behavior. *Language, 35*, 26-58.
Chomsky, N. (1965). *Aspects Of The Theory Of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.
Chomsky, N., & Halle, M. (1991). *The Sound Pattern Of English*. Cambridge, MA: MIT Press.
Churchland, P. S., Koch, C., & Sejnowski, T. J. (1990). What is computational neuroscience? In E. L. Schwartz (Ed.), *Computational Neuroscience*. Cambridge, MA: MIT Press.
Churchland, P. S., & Sejnowski, T. J. (1992). *The computational brain*. Cambridge, MA: MIT Press.
Clark, A. (1989). *Microcognition*. Cambridge, MA: MIT Press.
Clark, A. (1993). *Associative engines*. Cambridge, MA: MIT Press.
Cummins, R. (1983). *The Nature Of Psychological Explanation*. Cambridge, MA.: MIT Press.
Dawson, M. R. W. (1990). Training networks of value units: Learning in PDP systems with non-monotonicactivation functions. *Canadian Psychology, 31*(4), 391.
Dawson, M. R. W. (1998). *Understanding Cognitive Science*. Oxford, UK: Blackwell.
Dawson, M. R. W., Berkeley, I. S. N., Medler, D. A., & Schopflocher, D. S. (1994). Density plots of hidden value unit activations reveal interpretable bands and microbands.
Dawson, M. R. W., Boechler, P. M., & Valsangkar-Smyth, M. (2000). Representing space in a PDP network: Coarse allocentric coding can mediate metric and nonmetric spatial judgements. *Spatial Cognition and Computation, 2*, 181-218.
Dawson, M. R. W., & Medler, D. A. (1996). Of mushrooms and machine learning: Identifying algorithms in a PDP network. Canadian Artificial Intelligence, 38, 14-17.
Dawson, M. R. W., Medler, D. A., & Berkeley, I. S. N. (1997). PDP networks can provide models that are not mere implementations of classical theories. *Philosophical Psychology, 10*, 25-40.
Dawson, M. R. W., Medler, D. A., McCaughan, D. B., Willson, L., & Carbonaro, M. (2000). Using extra output learning to insert a symbolic theory into a connectionist network. *Minds And Machines, 10*, 171-201.
Dawson, M. R. W., & Piercey, C. D. (2001). On the subsymbolic nature of a PDP architecture that uses a nonmonotonic activation function. *Minds and Machines, 11*, 197-218.
Dawson, M. R. W., & Schopflocher, D. P. (1992). Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification. *Connection Science, 4*, 19-31.

DeYoe, E. A., & van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. *Trends in Neuroscience, 11*, 219-226.

Elman, J. (1990). Finding structure in time. *Cognitive science, 14*, 179-211.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal Reports As Data*. Cambridge, MA: MIT Press.

Evans, J. S. B. T. (1989). *Bias In Human Reasoning: Causes And Consequences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Farah, M. J. (1994). Neuropsychological evidence with an interactive brain: A critique of the "locality" assumption. *Behavioral and brain sciences, 17*, 43-104.

Farah, M. J., Weisberg, L. L., Monheit, M., & Peronnet, F. (1989). Brain activity underlying mental imagery: Event-related potentials during mental image generation. *Journal of Cognitive Neuroscience, 1*, 302-316.

Fodor, J. A. (1975). *The Language Of Thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. (1983). *The Modularity Of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition, 35*, 183-204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture. *Cognition, 28*, 3-71.

Gallant, S. I. (1993). *Neural network learning and expert systems*. Cambridge, MA: MIT Press.

Gallistel, C. R. (1990). *The Organization Of Learning*. Cambridge, MA: MIT Press.

Goodale, M. A. (1988). Modularity in visuomotor control: From input to output. In Z. W. Pylyshyn (Ed.), *Computational Processes In Human Vision: An Interdisciplinary Perspective* (pp. 262-285). Norwood, NJ: Ablex.

Goodale, M. A. (1995). The cortical organization of visual perception and visuomotor control. In S. M. Kosslyn & D. N. Osherson (Eds.), *An Invitation To Cognitive Science: Visual Cognition* (Vol. 2, pp. 167-213). Cambridge, MA: MIT Press.

Goodale, M. A., & Humphrey, G. K. (1998). The objects of action and perception. *Cognition, 67*, 181-207.

Grush, R. (2000). Self, world and space: The meaning and mechanisms of ego- and allocentric spatial representation. *Brain And Mind, 1*, 59-92.

Hanson, S. J., & Burr, D. J. (1990). What connectionist models learn: Learning and representation in connectionist networks. *Behavioral and brain sciences, 13*, 471-518.

Hanson, S. J., & Olson, C. R. (1991). Neural networks and natural intelligence: Notes from Mudville. *Connection science, 3*, 332-335.

Hillis, W. D. (1985). *The Connection Machine*. Cambridge, MA: MIT Press.

Hinton, G. E. (1986). *Learning distributed representations of concepts.* Paper presented at the The 8th Annual Meeting of the Cognitive Science Society, Ann Arbor, MI.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction To Automata Theory, Languages, And Computation*. Reading, MA: Addison-Wesley.

Ingle, D. (1973). Two visual systems in the frog. *Science, 181*(4104), 1053-1055.

Inhelder, B., & Piaget, J. (1958). *The Growth Of Logical Thinking From Childhood To Adolescence*. New York, NY: Basic Books.

Jackendoff, R. (1992). *Languages Of The Mind*. Cambridge, MA: MIT Press.

Jordan, M. I. (1986). An introduction to linear algebra in parallel distributed processing. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing, Volume 1*. Cambridge, MA: MIT Pressv.

Kilian, J., & Siegelmann, H. T. (1993). *On the power of sigmoid neural networks.* Paper presented at the Proceedings of the Sixth ACM Workshop On Computational Learning Theory.

Kitchin, R. M. (1994). Cognitive maps: What are they and why study them? *Journal Of Environmental Psychology, 14*, 1-19.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1994). *Image and Brain*. Cambridge, MA: MIT Press.

Kosslyn, S. M., Pascual-Leone, A., Felican, O., Camposano, S., Keenan, J. P., Thompson, W. L., Ganis, G., Sukel, K. E., & Alpert, N. M. (1999). The role of area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science, 284*, 167-170.

Kosslyn, S. M., Thompson, W. L., & Alpert, N. M. (1997). Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *Neuroimage, 6*, 320-334.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in area 17. *Nature, 378*, 496-498.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*, 445-463.

Krumhansl, C. L. (1982). Density versus feature weights as predictors of visual identifications: Comment on Appelman and Mayzner. *Journal Of Experimental Psychology: General, 111*, 101-108.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.

Leahey, T. H. (1987). *A History Of Psychology* (Second ed.). Englewood Cliffs, NJ: Prentice-Hall.

Leighton, J. P., & Dawson, M.R.W. (2001). A parallel distributed processing model of Wason's selection task. *Cognitive Systems Research, 2,* 207-231..

Lettvin, J. Y., NMaturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE, 47*(11), 1940-1951.

Malgady, R. G., & Johnson, M. G. (1976). Modifiers in metaphor: Effect of constituent phrase similarity on the interpretation of figurative sentences. *Journal Of Psycholinguistic Research, 5*, 43-52.

Marr, D. (1982). *Vision*. San Francisco, Ca.: W.H. Freeman.

McClelland, J. L., Rumelhart, D. E., & Group, t. P. (1986). *Parallel Distributed Processing, V.2*. Cambridge, MA: MIT Press.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing* (Vol. 1). Cambridge, MA: MIT Press.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological science, 2*, 387-395.

McCulloch, W. S. (1988). *Embodiments Of Mind*. Cambridge, MA: MIT Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

McNaughton, B., Barnes, C. A., Gerrard, J. L., Gothard, K., Jung, M. W., Knierim, J. J., Kudrimoti, H., Qin, Y., Skaggs, W. E., Suster, M., & Weaver, K. L. (1996). Deciphering the hippocampal polyglot: The hippocampus as a path integration system. *The Journal of Experimental Biology, 199*, 173-185.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100*(2), 254-278.

Medler, D. A., & Dawson, M. R. W. (1994a). Training redundant artificial neural networks: Imposing biology on technology. *Psychological Research, 57*, 54-62.

Medler, D. A., & Dawson, M. R. W. (1994b). Using redundancy to improve the performance of artificial neural networks.

Mellet, E., Petit, L., Mazoyer, B., Denis, M., & Tzourio, N. (1998). Reopening the mental imagery debate: Lessons from functional anatomy. *Neuroimage, 8*, 129-139.

Minsky, M. (1972). *Computation: Finite And Infinite Machines*. London: Prentice-Hall International.

Minsky, M., & Papert, S. (1988). *Perceptrons, 3rd Edition*. Cambridge, MA: MIT Press.

Moorhead, I. R., Haig, N. D., & Clement, R. A. (1989). An investigation of trained neural networks from a neurophysiological perspective. *Perception, 18*, 793-803.

Moravec, H. (1999). *Robot*. New York, NY: Oxford University Press.

Newell, A. (1980). Physical symbol systems. *Cognitive Science, 4*, 135-183.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurones. *Nature, 381*, 425-428.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research, 34*, 171-175.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus As A Cognitive Map*. Oxford: Clarendon Press.
Ortony, A. (1979). Beyond literal similarity. *Psychological Review, 86*, 161-180.
Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological review, 76*, 241-263.
Paivio, A. (1971). *Imagery And Verbal Processes*. New York: Holt, Rinehart & Winston.
Pfeifer, R., & Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press.
Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*, 73-193.
Popper, K. R. (1979). *Objective Knowledge*. London: Oxford University Press.
Pylyshyn, Z. W. (1979). The rate of 'mental rotation' of images: A test of a holistic analogue hypothesis. *Memory and Cognition, 7*, 19-28.
Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences, 3*, 111-169.
Pylyshyn, Z. W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological Review, 88*(1), 16-45.
Pylyshyn, Z. W. (1984). *Computation And Cognition*. Cambridge, MA.: MIT Press.
Pylyshyn, Z. W. (1991). The role of cognitive architectures in theories of cognition. In K. VanLehn (Ed.), *Architectures For Intelligence* (pp. 189-223). Hillsdale, NJ: Lawrence Erlbaum Associates.
Redish, A. D., & Touretzky, D. S. (1999). Separating hippocampal maps. In B. N. & K. J. Jeffery & J. O'Keefe (Eds.), *The Hippocampal And Parietal Foundations Of Spatial Cognition*. Oxford: Oxford University Press.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research And Theory* (pp. 64-99). New York, NY: Appleton-Century-Crofts.
Rollins, M. (2001). The strategic eye: Kosslyn's theory of imagery and perception. *Minds and Machines, 11*, 267-286.
Romney, A. K., Shepard, R. N., & Nerlove, S. B. (1972). *Multidimensional Scaling: Theory And Applications In The Behavioral Sciences. Volume II: Applications*. New York, NY: Seminar Press.
Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology, 5*, 1-28.
Rumelhart, D. E., & McClelland, J. (1986). On learning the past tenses of English verbs. In J. McClelland & D. E. Rumelhart (Eds.), *Parallel Distributed Processing. Volume 2: Psychological And Biological Models* (pp. 216-271). Cambridge, MA: MIT Press.
Rumelhart, D. E., McClelland, J. L., & Group., t. P. (1986). *Parallel Distributed Processing, V.1*. Cambridge, MA: MIT Press.
Seidenberg, M. (1993). Connectionist models and cognitive theory. *Psychological science, 4*, 228-235.
Seidenberg, M., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological review, 97*, 447-452.
Shamanski, K. S., & Dawson, M. R. W. (1994). *Problem type by network type interactions in the speed and transfer of connectionist learning.* Paper presented at the Machine Learning Workshop at AI/GI/VI'94, Banff, AB.
Shepard, R. N. (1972). A taxonomy of some principal types of data and of multidimensional methods for their analysis. In R. N. Shepard & A. K. Romney & S. B. Nerlove (Eds.), *Multidimensional scaling: Theory and applications in the behavioral sciences. Vol 1:Theory* (pp. 21-47). New York, NY: Seminar Press.
Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.
Shepard, R. N., Romney, A. K., & Nerlove, S. B. (1972). *Multidimensional Scaling: Theory And Applications In The Behavioral Sciences. Volume I: Theory*. New York, NY: Seminar Press.
Sherry, D., & Healy, S. (1998). Neural mechanisms of spatial representation. In S. Healy (Ed.), *Spatial Representation In Animals*. Oxford: Oxford University Press.

Siegelman, H. T., & Sontag, E. D. (1991). Turing computability with neural nets. *Applied mathematics letters, 4*, 77-80.

Siegelmann, H. T. (1999). *Neural Networks And Analog Computation: Beyond The Turing Limit*. Boston, MA: Birkhauser.

Siegelmann, H. T., & Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and System Sciences, 50*, 132-150.

Simon, H. A. (1996). *The Sciences Of The Artificial* (Third ed.). Cambridge, MA: MIT Press.

Skinner, B. F. (1957). *Verbal Behavior*. New York, NY: Appleton-Century-Crofts.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Thompson, W. L., Kosslyn, S. M., Sukel, K. E., & Alpert, N. M. (2001). Mental imagery of high- and low-resolution gratings activates Area 17. *Neuroimage, 14*, 454-464.

Tolman, E. C. (1932). *Purposive behavior in animals and men*. New York: Century Books.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review, 55*, 189-208.

Tourangeau, R., & Sternberg, R. J. (1981). Aptness in metaphor. *Cognitive psychology, 13*, 27-55.

Tourangeau, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition, 11*, 203-244.

Touretzky, D. S., Wan, H. S., & Redish, A. D. (1994). Neural representation of space in rats and robots. In J. M. Zurada & R. J. Marks & C. J. Robinson (Eds.), *Computational Intelligence: Imitating Life.* New York, NY: IEEE Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89*, 123-154.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. Ingle & M. A. Goodale & R. J. W. Mansfield (Eds.), *Analysis Of Visual Behavior*. Cambridge, MA: MIT Press.

Wason, P. C. (1966). *Reasoning*. New York: Penguin.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158-177.

Williams, R., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation, 1*, 270-280.

Zimmerman, C. L. (1999). *A network interpretation approach to the balance scale task.* Unpublished Ph.D., University of Alberta, Edmonton.