
Spatial Cognition and Computation **2:** 181–218, 2000.

Representing space in a PDP network: Coarse allocentric coding can mediate metric and nonmetric spatial judgements

MICHAEL R.W. DAWSON*, PATRICIA M. BOECHLER and MONICA VALSANGKAR-SMYTH
*Biological Computation Project, Department of Psychology, University of Alberta, Edmonton, Alberta, CANADA T6G 2P9 (*Author for correspondence: E-mail: mike@bcp.psych.ualberta.ca)*

Received 22 June 2000; accepted 10 September 2001

Abstract. In one simulation, an artificial neural network was trained to rate the distances between pairs of cities on the map of Alberta, given only place names as input. Distance ratings ranged from 0 (when the network rated the distance between a city and itself) to 10. The question of interest was the nature of the representations developed by the network's six hidden units after it successfully learned to make the desired responses. Analyses indicated that the network used coarse allocentric coding to solve this problem. Each hidden unit could be described as occupying a position on the map of Alberta, and each connection weight feeding into a hidden unit was related to the distance on the map between the hidden unit and one of the stimulus cities. On its own, a single hidden unit's response was a relatively inaccurate distance measure. However, by combining all six hidden unit responses in a coarse coding scheme, accurate responses were generated by the network. In a second simulation, a second network was trained to make similar judgements, but was trained to violate the minimality constraint on metric space when trained to judge the distance between a city and itself. An analysis of this network indicated that it too was using coarse allocentric coding.

Key words: artificial neural network, coarse allocentric coding, minimality principle, symmetry principle, triangle inequality

Introduction

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*, which dictates that the shortest distance in the space is between a point x and itself. The second is the *symmetry principle*, which dictates that 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*, which

dictates that 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. In the sections that follow, we briefly review three examples of this kind of proposal. In each example, we show that proposals for metric spatial representations are not without controversy. Specifically, agents can perform in ways that appear to violate the assumed metric structure of their underlying representations. Furthermore, the neural structures that are presumed to instantiate metric representations do not appear to be organized in any way that maps neatly onto our notions of metric space.

The purpose of this paper is to examine this kind of controversy from the perspective of synthetic psychology. Rather than taking an existing system and analyzing its behaviour and representations, we choose to construct a system capable of generating behaviour that preserves the properties of metric space. In taking this approach, we are particularly interested in two questions. First, if very few constraints are placed on the internal structure of this system, then will it develop metric representations to guide its behaviour? Second, if the same system is designed to generate behaviours that violate metric properties of space, then will it develop nonmetric internal representations? Or can both metric and nonmetric behaviours be mediated by similar (metric) representations?

This paper proceeds as follows: first, we briefly review three proposals for metric representations – similarity space, mental imagery, and the cognitive map. We use this review to motivate our own studies. Second, we describe a simulation in which a parallel distributed processing (PDP) network is trained to make distance judgments that preserve the metric properties of space. We show that this network uses an interesting internal representation that appears to be metric in nature. Third, we describe another study in which a PDP network is trained to make distance judgments that violate one of the properties of metric space. Interestingly, the internal representations of the second network are very similar to those used by the first. Finally, we discuss some implications of the simulation results to proposals concerning metric mental representations.

Similarity space representations

Similarity is one of the most important theoretical constructs in cognitive psychology (Medin et al. 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 et al. 1972; Shepard et al. 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 behaviour on a variety of different tasks, including analogical reasoning (Rumelhart and Abrahamson 1973) and judgments of the aptness of metaphor (Tourangeau and Sternberg 1981, 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 behaviours 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.

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 and Gati 1982). As a result, many researchers 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 use a perfectly metric representational space 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.

The architectural debate about whether similarity spaces mediated similarity judgments illustrates a tension that is central to the current paper. The tension is between regularities observed in behaviour and regularities attributed to underlying mental representations. At issue is whether all of the regularities in a representation assert themselves in behaviours or in

judgments that exploit the representation. If the representation is a metric space, does this imply that behaviour will also be metric? Or can a metric representation be used to mediate behaviour that is nonmetric in nature?

Depictive mental imagery

The second illustration of the tension between metric behaviour and metric representation in cognitive psychology can be found in the study of 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).

While our experience of mental images is definitely spatial in nature, there has been a long-standing debate about whether the underlying representation that supports mental imagery is spatial or not. On the one hand, some researchers have argued that the representational format of mental images is depictive or spatial in nature (Kosslyn 1980). According to this position, images occur in a spatial medium that is equivalent to a coordinate space. The pattern formed in this spatial medium is topographically organized so that each local portion of the image corresponds to a portion of a represented object as seen from a particular point of view, and the distances between the portions of the image implicitly represent the distances between parts of the object being represented. Images not only depict information about spatial extent, but also depict information about the appearance of surface properties of objects. On the other hand, other researchers have argued that while mental images are experienced spatially, their representational format is not spatial at all. For instance, it has been argued that mental images could be represented in a non-depictive format, for instance as a set of statements in propositional logic (Pylyshyn 1973; Pylyshyn 1981). The major papers that explore different sides of this debate have been collected in Block (1981).

In the early stages of the imagery debate, behavioural experiments were used to collect evidence either for or against the depictive position. Many studies recorded the reaction times of subjects as they manipulated 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 and Cooper 1982). Such results provided strong support to the depictive position. However, other researchers found 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 mental imagery is cognitively penetrable, and that our experience of mental images is based upon more primitive and non-spatial representational components (Pylyshyn 1980, 1981, 1984).

Much of this experimental work ground to a halt after the publication of an influential paper that proposed the "indeterminacy of representation" thesis (Anderson 1978). Anderson argued that behavioural evidence could never be used to resolve the imagery debate because different representational systems could lead to identical behavioural predictions. Anderson went on to explore additional constraints or types of evidence that might be useful in resolving representational issues. One of these was physiological data: "if we could open the brain and observed that operating on pictures or on propositions, it seems that the issue would be settled" (p. 271).

Recent research related to the imagery debate has attempted to exploit exactly this type of data. In particular, Kosslyn has turned to evidence from 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 et al. 1989; Kosslyn 1994; Kosslyn et al. 1999; Kosslyn et al. 1997; Kosslyn et al. 1995; Thompson et al. 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.

However, even these neuroscience-based conclusions are not without controversy. In a detailed review of the literature, Mellet et al. (1998) cite several studies that have found that some mental imagery tasks do not produce activity in primary visual cortex. Furthermore, in some cases mental imagery can generate activity in brain areas that are also activated by other (non-spatial) higher order cognitive processes, such as language and memory. Mellet et al. use this kind of evidence to argue that there is no cortical network that is uniquely associated with mental imagery.

In the previous section on similarity spaces, we noted a tension or inconsistency between a proposed spatial representation and observed (nonmetric) behaviour. In the current section, a similar tension is evident. Specifically, in the imagery debate we have on the one hand an experience that is clearly spatial in nature, but on the other hand we have no clear resolution to the debate about whether this experience is mediated by a mental representation that is spatial or not.

A third illustration of this tension between metric behaviour and metric representation can be found in the study of the hippocampus as a 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). Behavioural studies have demonstrated that animal representations of space do indeed appear to preserve a good deal of its metric nature (for introductions, see Cheng and 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 and 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 and 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 described as the sum of two or more Gaussian tuning curves sensitive to the distance of the animal away from a wall in its environment.

However, 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 et al. 1995; McNaughton et al. 1996). Second, it has been argued that place cell receptive fields are at best *locally* metric (Touretzky et al. 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 and Touretzky 1999; Touretzky et al. 1994).

It would appear that how the hippocampus provides a metric representation of space is still an open question (Redish 1999, Chap. 15). One reason that this question has not been satisfactorily answered is because while there is a great deal of functional understanding of the structure of the hippocampus

(i.e., the kinds of stimuli which cause cells to respond), there is much less understanding of how the underlying neural circuitry encodes spatial information. "Despite the remarkable properties of place cells and a number of interesting theoretical proposals, there is no generally accepted account of the function of place cells in spatial orientation" (Sherry and Healy 1998).

Rationale for the current studies

The three examples that were briefly reviewed above all involve proposals for metric, spatial representations that mediate spatial behaviour. However, in each example it was shown that such proposals are not without controversy. In some instances, behaviour 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 behaviour or experience reveal regularities that are inconsistent with the notion that the underlying structure is metric in nature.

One reason that such inconsistencies emerge is because a general strategy in cognitive psychology is to develop theories about underlying representations by decomposing complex behaviour 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 is that it can lead to theories that are more complicated than necessary, because the decomposition can fail to partition behaviour appropriately into three different categories (behaviour caused by the organism, behaviour elicited by a complex environment, and behaviour 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, p. 271).

One alternative to functional analysis has been called synthetic psychology (Braitenberg 1984), and involves taking a fixed set of building blocks and using them to create a system that generates complex or interesting behaviour (Brooks 1999; Minsky 1985; Pfeifer and Scheier 1999). The complex behaviour emerges from the interaction of the system's components with each other and with the environment. One advantage of this approach is that it is often theory-neutral with respect to the behaviour that emerges (Dawson 1998). Because of this, a second advantage of this approach is that

it can lead to the discovery of novel representations for generating behaviour, because preconceived representational theories are not required to guide the construction of the model.

The purpose of the current paper is to illustrate this synthetic approach, by using computer simulations to suggest how a brain-like system might represent the metric properties of space and mediate spatial behaviour. In particular, we describe parallel distributed processing (PDP) networks that were trained to rate the distance between pairs of places in Alberta. After training the networks to perform this task, we examined in detail the internal representations that they had developed. This examination was performed with two different questions in mind. First, when a network is trained to make metric judgements, do its hidden units develop a metric representation of space? Second, when a network is trained to make judgements that are not metric, do its hidden units develop a markedly different internal representation, or is a metric representation still used?

Simulation 1: Metric spatial judgements

This section of the paper proceeds as follows. First, we briefly introduce PDP networks in general, and focus on a particular type of PDP network, called networks of value units. Second, we describe how one network of value units was trained to make spatial judgements. Third, we present a detailed analysis of the structure of the hidden units of this network. Finally, we relate the structure found in these hidden units to research concerning the nature of the cognitive map in the hippocampus.

A brief introduction to PDP networks

A PDP network is a computer simulation of a "brain-like" system of interconnected processing units (see Figure 1). In general, such a network can be viewed as a multiple-layer system that generates a desired response to an input stimulus. The stimulus is provided by the environment, and is encoded as a pattern of activity in a set of *input units*. The response of the system is represented as a pattern of activity in the network's *output units*. Intervening layers of processors in the system, called *hidden units*, detect features in the input stimulus that allow the network to make a correct or appropriate response.

Each *processing unit* in a PDP network is analogous to a neuron. The behaviour of a single processing unit in this system can be characterized as follows: First, the unit computes the total signal being sent to it by other

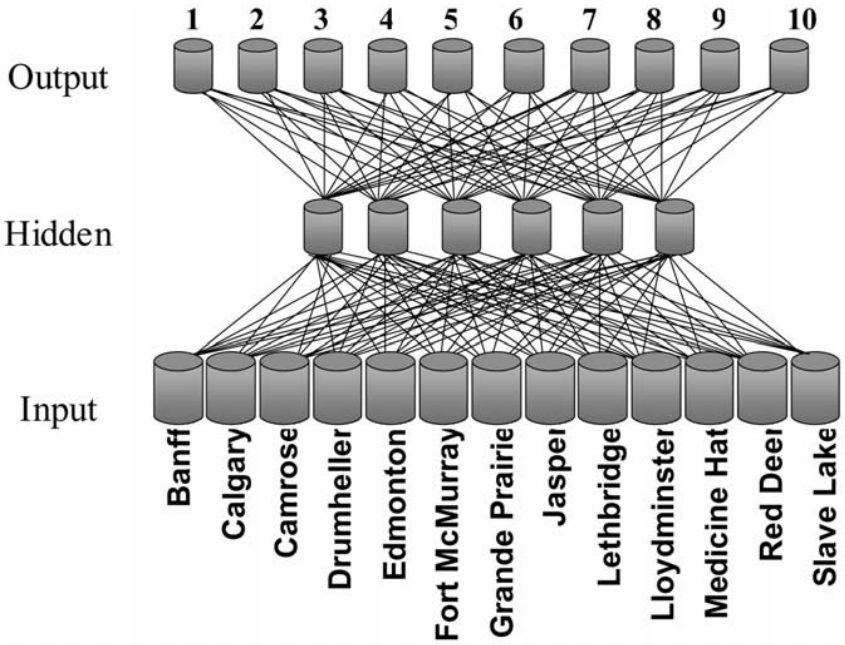

Figure 1. A multilayer perceptron. The environment activates input units, which in turn produce activity in a layer of hidden units. The hidden units then produce a response by sending signals that produce activity in the output units. Links between layers are weighted connections; the values of the weights are determined by having the network learn through examples. This particular network is described later in the paper. Given two input city names, the network was trained to rate the distance between the cities on a map of Alberta.

processors in the network. Second, the unit adopts a particular level of internal activation on the basis of this computed signal. Third, the unit generates its own signal, which is based on its level of internal activity, and sends this signal on to other processors.

The signal sent by one processor to another is transmitted through a *weighted connection.* Such a connection is typically described as being analogous to a synapse, and serves as a communication channel that amplifies or attenuates a numerical signal being sent through it. This is accomplished by multiplying the signal's value 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.

The pattern of connectivity in a PDP network (i.e., the network's entire set of connection weights) defines how signals flow between the processors. As a result, a network's connection weights are analogous to a program in a conventional computer (Smolensky 1988). However, in contrast to a conventional computer, a PDP network is not given an explicit program to perform some desired task. Instead, the network is *taught* to do the task.

For example, consider a popular learning procedure called the generalized delta rule (Rumelhart et al. 1986). One starts with a network that has small, randomly assigned connection weights. The network is then taught by presenting it a set of training patterns, each of which is associated with a known correct response. To train a network on one of these patterns, the pattern is presented to the network's input units, and (on the basis of its existing connection weights) the network generates a response to it. An error term is calculated which is essentially the difference between the desired response of the output unit and its actual response. This error term can be used to modify the network's connections in such a way that the next time this pattern is presented to the network, the amount of error in the network's response will be smaller.

When using the generalized delta rule, error is used to modify connection weights by sending it backwards through the network. Once the error term for each output unit has been calculated, the weights of the connections directly attached to each output unit are modified. Then the output units send their error as a signal through the modified connections to the next layer of hidden units. Each hidden unit computes its overall error by treating the incoming error signals as net input (i.e., a hidden unit's total error is the sum of the weighted error signals that it is receiving from each output unit). Once a hidden unit has computed its overall error, then the weights of the connections that are directly attached to it can be modified. This process can be repeated, if necessary, to send error signals to the next layer of hidden units, and stops once all of the connections in the network have been modified. By repeating this procedure a large number of times for each pattern in the training set, the network's response errors for each pattern can be reduced to near zero. At the end of this 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.

The value unit architecture

One of the main functions of a PDP processing unit is to use an activation function to convert its net input into an internal level of activity that usually is some continuous value that ranges between 0 and 1. For most networks

trained with the generalized delta rule, the activation function is a sigmoid-shaped curve that is defined by the logistic equation (Rumelhart et al. 1986). This function monotonically increases with increases in net input, and for this reason units that use this activation function have been called integration devices (Ballard 1986).

However, PDP architectures can differ from one another along a variety of dimensions (Dawson 1998). One important distinction between different classes of PDP networks involves the activation function. Not all PDP networks are composed of integration devices. Some networks use processors that are tuned to activate to a small range of net inputs, and generate weak responses to net inputs that are either too small or too large to fall in this range. In other words, the response of such a processor is nonmonotonically related to increases in net input, and for this reason it is called a value unit (Ballard 1986).

One example of a network of value units is an architecture designed by Dawson and Schopflocher (1992). The processing units in this architecture use a Gaussian activation function that ranges between 0 and 1, with a standard deviation of 1. These units generate a maximum response of 1 when their net input is equal to the mean of the Gaussian. Networks of value units are trained with a variation of the generalized delta rule.

The research below involved training networks of value units on a distance estimation task. This architecture was used because one of the primary goals of the research was to interpret the internal representations discovered by the network. 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 and Medler 1996; Dawson et al. 1997; Dawson et al. 2000; Leighton 1999; Zimmerman 1999).

A second reason that this architecture was used was because of the existing literature on place cell behaviour. Place cells in the hippocampus behave as though they are value units – they are tuned to respond to particular ranges of spatial measurements in the environment, and their behaviour suggests that their distance sensitivity is modulated by a Gaussian response function (O'Keefe and Burgess 1996). This suggests that the value unit architecture is particularly appropriate for studying how networks of parallel processors represent spatial information, particularly if one goal of the simulation research is to explore representations that might be found in hippocampal circuitry.

Method

Problem definition

Multidimensional scaling (MDS) is a statistical technique that takes proximity information as input, and then converts this information into a geometric configuration of points from which the proximities can be derived (Kruskal and 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.

MDS has been commonly used to construct psychological spaces from proximity data that is analogous to distances between cities. The most common approach to obtaining such data is to present pairs of objects to human subjects, and to have subjects rate the "psychological distance" between the objects. For example, subjects might be asked to judge how similar or related the two objects are (Shepard 1972).

In the first simulation described below, we trained a network to make such judgements about the "crows flight" distances between places on a map. We 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. We 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?". Because all possible pairs of place names were used, some of the stimuli involved rating the distance from one place to itself. As well, for different place names a rating would be obtained for both orders of places (e.g., the distance between Banff and Calgary would be rated, as would the distance between Calgary and Banff).

The desired ratings for each stimulus were created as follows. First, from a map of Alberta we determined the shortest distance in kilometres between each pair of locations (see Table 1a). Second, we 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 kilometres, then it was assigned a value of 1; if the distance was between 100 and 199 kilometres, then it was assigned a value of 2; if the distance was between 200 and 299 kilometres, 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 kilometres or more. The complete set of ratings that were used is provided in Table 1b.

Because this ratings matrix was based on physical distances between geographic locations, and because we enforced the minimality principle by

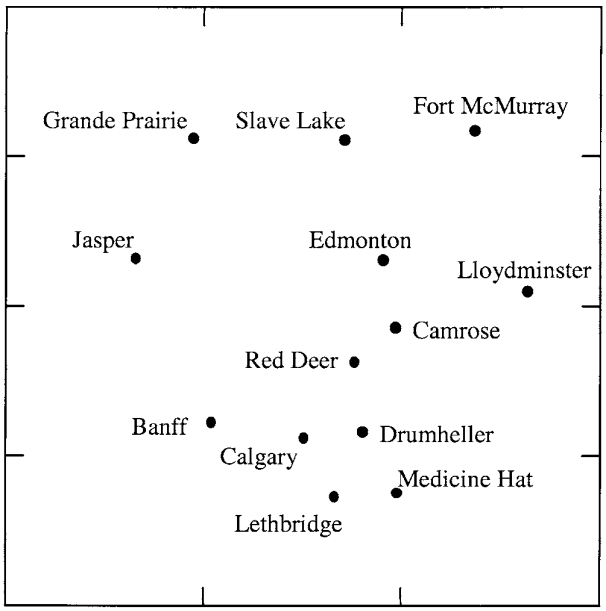

Figure 2. The locations of 13 Albertan cities taken from the MDS analysis of the ratings matrix of Table 1b. The positions of the cities is very similar to their locations on a map of Alberta.

requiring that the distance between a place and itself should be rated as 0, this set of judgements was metric in nature. To confirm this, we applied multidimensional scaling to the ratings data, forcing a solution in which the configuration of points delivered by MDS was restricted to two dimensions. The success of MDS depends upon the assumption that the data being analyzed has properties that conform to the metric constraints upon space. Our solution accounted for 99.4% of the variance in the ratings matrix, and produced a configuration of points that agreed quite nicely with the locations of the thirteen cities on a road map of Alberta (see Figure 2).

Network architecture

The network that was trained to make the distance ratings was a network of value units that had 10 output units, 6 hidden units, and 13 input units. The overall architecture of the network was illustrated earlier in Figure 1.

Input unit representation. 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

Table 1. (A) Distances between cities of Alberta, measured in kilometres. (B) The distances from Table 1a converted into ratings on a 0 to 10 scale

	BANFF	CALGARY	CAMROSE	DRUM-HELLER	EDMONTON	FORT MCMURRAY	GRANDE PRAIRIE	JASPER	LETH-BRIDGE	LLOYD-MINSTER	MEDICINE HAT	RED DEER	SLAVE LAKE
BANFF	0	128	381	263	401	840	682	287	342	626	419	253	652
CALGARY	128	0	274	138	294	733	720	412	216	519	293	145	545
CAMROSE	381	274	0	182	97	521	553	463	453	245	429	129	348
DRUMHELLER	263	138	182	0	279	703	735	547	282	416	247	165	530
EDMONTON	401	294	97	279	0	439	456	366	509	251	526	148	250
FORT MCMURRAY	840	733	521	703	439	0	752	796	948	587	931	587	436
GRANDE PRAIRIE	682	720	553	735	456	752	0	397	935	701	982	586	318
JASPER	287	412	463	547	366	796	397	0	626	613	703	413	464
LETHBRIDGE	342	216	453	282	509	948	935	626	0	605	168	360	760
LLOYDMINSTER	626	519	245	416	251	587	701	613	605	0	480	374	496
MEDICINE HAT	419	293	429	247	526	931	982	703	168	480	0	409	777
RED DEER	253	145	129	165	148	587	586	413	360	374	409	0	399
SLAVE LAKE	652	545	348	530	250	436	318	464	760	496	777	399	0

A

Table 1. Continued

	BANFF	CALGARY	CAMROSE	DRUM-HELLER	EDMONTON	FORT MCMURRAY	GRANDE PRAIRIE	JASPER	LETH-BRIDGE	LLOYD-MINSTER	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

B

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.

It should be noted that this kind of input representation is not sensitive to the order of cities in a question of the form "How far is City 1 from City 2?". For example, the input representation of the question "How far is Banff from Calgary?" is identical to the representation of the question "How far is Calgary from Banff?". Because of this, and because we used all 169 city pairs from the ratings taken from Table 1b, all of the patterns – except those representing the diagonal of Table 1b – were represented twice in the training set. We elected to do this because we intend to compare the results reported below with future networks in which the symmetry constraint is violated, and as a result 169 unique stimuli will be presented to networks. It should also be noted that using this input representation, distance ratings representing the diagonal entries from Table 1b involve activating only one input unit. For example, the input representation of the question "How far is Banff from Banff?" involves turning only the first input unit on, because this unit represents Banff.

Hidden units. Six hidden units were included in this network to solve the problem. Each of these units was a value unit. We selected this number of hidden units because pilot tests 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 et al. 1995).

Output unit representation. Ten output units were used to represent the network's rating of the distance between the two place names presented as input. The output units were also value units. 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.

Network training

The network was trained using the variation of the generalized delta rule that has been developed for networks of value units (Dawson and Schopflocher 1992). 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 and zero momentum. 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" or 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. The network converged on a solution to the problem – generating a correct response for each of the 169 patterns – after 10,907 sweeps of training.

Results

At the end of training, we had produced a network that correctly rated the distances between places in Alberta on a scale from 0 to 10. The ratings generated by the network conform to the metric constraints of space. The question of interest concerns the kinds of internal representations used by the network to generate this metric behaviour. 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 behaviour can be derived?

Relating the map of Alberta to hidden unit connection weights

To begin our analyses, we explored 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.

In order to test this hypothesis, we need some objective method of finding a map location for each hidden unit that optimizes the correlation between map distances and connection weights. The method that we adopted was the Solver tool in Microsoft Excel. The Solver tool is a program that manipulates the values of specified cells in a spreadsheet as it attempts to optimize the value of one other cell (Orvis 1996). Solver uses a particular algorithm, called the generalized reduced gradient method (Fylstra et al. 1998; Gill et al. 1981; Lasdon et al. 1978), to search through different values to find those that produce this optimum value.

For our purposes, we created a spreadsheet that contained the latitude and longitude of each of the 13 Albertan cities. To process one of the hidden units, we initially assigned it a latitude and longitude that placed it in the middle of the map of the province. Using these latitude and longitude coordinates, the spreadsheet was designed to calculate the Euclidean distance between each city and the (current) position of the hidden unit. These 13 distances were then correlated with the 13 connection weights feeding into the hidden unit. This correlation was the value that the Solver tool optimized. The Solver tool searched for the latitude and longitude of the hidden unit that produced the most extreme correlation between distances and connection weights. We found, for each hidden unit, a map location for the unit that produced strong correlations between city distances to that location and the hidden unit's incoming connection weights (see Table 2).

These correlations suggest that the weights feeding into each hidden unit represent the distance between each city and the hidden unit, with the hidden unit occupying a particular location on the map of Alberta. Figure 3 illustrates the map of Alberta with both the locations of the 13 cities and the locations of the 6 hidden units.

Relating connection weights to hidden unit MDS spaces

The previous 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 we expected. One problem with the previous analysis is that it imposes our notion of the space

Table 2. Results of relating Alberta map distances between cities and hidden units the values of the connection weights feeding into the hidden units in Simulation 1

Hidden Unit	Hidden Unit Latitude	Hidden Unit Longitude	Correlation Between Map Distances And Incoming Weights
H0	51.72	113.55	0.88
H1	50.84	113.63	0.59
H2	50.88	117.70	0.72
H3	53.39	115.05	–0.54
H4	53.91	113.42	0.79
H5	51.17	115.57	–0.48

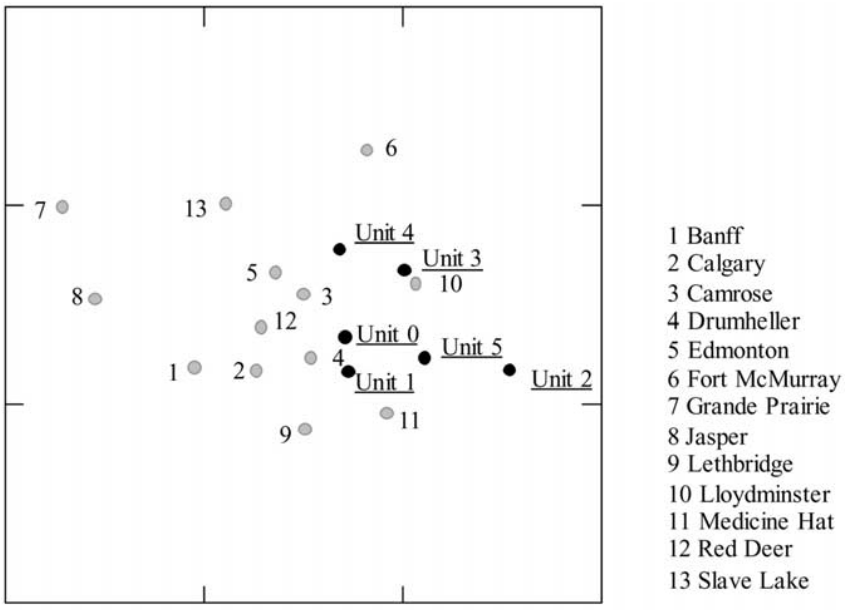

Figure 3. The locations of the six hidden units from the first simulation on the map of Alberta. These locations optimized the correlations between hidden unit connection weights and distances from the cities to the hidden units.

in question (i.e., the map of Alberta) onto the behaviour 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 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. We would expect that this kind of transformation would be equivalent to a distortion of the Figure 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 examine 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 behaviour 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 compared 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 2), then how should we proceed to analyse their behaviour? 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 we take all of these ratings and organize them into a table like Table 1, then we can apply MDS to this data. This analysis will determine the structure of the space that underlies the hidden units behaviour. We can then relate properties of this space to the connection weights that feed into each unit.

In this second analysis, we created a 13 $\times$ 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. We used MDS to analyse each of these activity matrices, in order to create a space that represented each hidden unit's perspective on the map. We found that for each hidden unit a two-dimensional MDS solution accounted for almost all of the variance in the data. This solution generated an R^2 of 0.98 for hidden unit 0, 1.00 for hidden unit 1, 0.99 for hidden unit 2, 1.00 for hidden unit 3, 0.99 for hidden unit 4, and 0.97 hidden unit 5. In each of these two-dimensional solu-

Table 3. Results of relating connection weights to city distances from the MDS solutions obtained from the activity matrix for each hidden unit in Simulation 1. The table provides the maximum correlation, as well as the coordinates of the hidden unit in the space that produces this maximum correlation

Hidden Unit	X-Coordinate Of Hidden Unit	X-Coordinate Of Hidden Unit	Correlation Between MDS Distances And Connection 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

tions, the 13 different cities were arranged in a roughly circular or elliptical pattern around an origin; the specific arrangement varied from one unit to another. Thus, none of the maps resembled the actual Alberta map. However, this analysis did yield a set of x- and y-coordinates for each Albertan city in a two-dimensional space that was customized for each hidden unit.

The second phase of this analysis was to repeat our previous analyses, by determining the coordinates for each hidden unit that optimized the correlation between the unit's connection weights and the distances from each city to the unit. However, instead of using the map of Alberta, for each hidden unit we 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 3.

Coarse coding from hidden unit activations to distance ratings

The previous analyses have indicated that the network has developed a spatial representation of the locations, in which the weights that feed into a hidden unit encode information about the distance between the hidden unit's location in a 2D space and city locations in the same space. However, we have not yet discussed how the features detected by the hidden units are combined to generate the ratings that the network has been trained to produce.

It was pointed out 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, we 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 are at best weakly related to the desired distance ratings. As well, none of the MDS analyses of hidden unit activity matrices revealed a plot that bore any resemblance to the actual map of Alberta. 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 and 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 and Sejnowski 1992).

The coarse coding that is used in the network described 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, we 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.

Discussion

The purpose of the first simulation was to train an artificial neural network to generate ratings of the distances between pairs of Albertan cities. The inputs to the network only represented city names; the network was not provided any information about the actual location of the cities in Alberta. After training was completed, one question of primary interest was whether the internal representations of the network were metric.

Several different analyses of the internal structure of the network were reported above, and all of these analyses converged on an affirmative answer to this question: the hidden units of the network did indeed develop 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.

Implications for the hippocampal cognitive map

The strong interest that neuroscientists have taken in the study of spatial behaviour and cognitive maps can largely be traced back to the discovery of place cells in the hippocampus (O'Keefe and 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 and 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 and 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 et al. 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 behaviours 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 and Touretzky 1999; Touretzky et al. 1994).

Interestingly, the hidden units in the PDP 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 are all connected to all of the input units, the network has no definite topographic organization. Second, each hidden unit appears to be at best locally metric. While the input connections can be correlated with distances on the map, the responses of individual hidden units do not provide an accurate spatial account of the map. Nevertheless, the fact that the PDP network could be trained to accurately generate the ratings given in Table 1b indicates that the responses of these locally metric, inaccurate processors can be used to accurately represent spatial information about the entire map of Alberta. This is possible because the network does not base its output on the behaviour of a single hidden unit. Instead, it relies on coarse coding, and generates its response from the activity of all six hidden units considered simultaneously.

One implication of this coarse coding is that spatial relationships amongst locations in Alberta are being 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 PDP 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 "centred 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-centred reference 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 PDP 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). We call this a *coarse allocentric code*.

The discovery of coarse allocentric coding in the network is not completely surprising, because other network simulations of spatial processing have developed representations that seem similar. Ghiselli-Crippa and Munro (2000) trained a network to navigate through a spatial layout of nodes. When they expressed the representation of a single node as a vector of hidden unit activities, they found high correlations between distances between nodes and distances between hidden unit vectors. The notion of a spatial location being represented as a pattern of activities is central to coarse allocentric coding. Arleo and Gerstner (2000) used reinforcement learning to train a simulation of place cells in the hippocampus that was used to drive a miniature mobile robot. The place cell receptive fields that developed overlapped one another. Such overlapping is characteristic of coarse coding. However, to our knowledge, other simulation researchers have not performed a detailed analysis of hidden unit properties on an individual, connection weight by connection weight basis of the type that we conducted in Simulation 1. As a result, whether the coarse allocentric code discovered in our network is equivalent to the representations used in other simulations remains an open question.

The major hypothesis about the hippocampus that is suggested by the PDP 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.

Simulation 2: Nonmetric spatial judgements

In the Simulation 1, a PDP network was trained to make judgements of distance that preserved the metric properties of space. When the internal representations of this network were examined, it was found that they were basically spatial in nature. Each hidden unit could be assigned a position in a two-dimensional space, and the weights of the connection that provided input to a unit represented the distance between the hidden unit and the city associated with each city. One surprise about this internal representation was that accurate spatial responses by the output units of the network depended upon a coarse allocentric code. Each hidden unit, though spatial in nature, provided distance information that was inaccurate or distorted. Accurate information required considering the response of each hidden unit at the same time.

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 in the introduction, there does exist a tension between the metric properties of a representation and the properties of the behaviour that the representation mediates. Specifically, is it possible for a metric representation to mediate nonmetric behaviour? Our 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 judgements that were not completely metric.

One of the reasons for the rise in the popularity of PDP networks over symbol-based models is that PDP models degrade gracefully and are damage resistant (McClelland et al. 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 and McClelland 1989). This is called the three bears problem because if the number of hidden units is not just right – if there are too few or if there are too many – then generalization will suffer.

In the current paper, we were not concerned with generalization of this particular type. One reason for this was because we were not interested in making an applied tool that would be presented new stimuli, but instead were interested in examining the internal representations of a network trained to accomplish a known task. A second reason was that our pilot tests indicated that a network with six hidden units was "just right" for this problem – with fewer hidden units, the network did not converge, and more hidden units were not required to solve the problem. A third reason was that our detailed analysis of hidden unit representations in Simulation 1 indicated that our network was not falling into the three bears problem, because there was no evidence that individual hidden units were focusing upon specific training patterns.

However, we were concerned with a different type of generalization – the generalization of representation type from one problem to another. Specifically, imagine if the network's task was changed in such a way that the distance ratings violated one of the metric properties of space. Would it be the case that we would observe representational generalization – that allocentric coarse coding could still be used to represent a solution to the problem? Or would a change in task result in a completely different representational approach? Simulation 2 was designed to explore these questions, by having a network learn to make a set of distance judgements that violated the minimality principle of metric space.

Method

Problem definition

The problem that the network was trained to solve in the second simulation was a distance estimation task that was identical to that used in the first simulation, with the exception that the network was trained to make different judgements when asked to judge the distance between a city and itself. In the

first simulation, such judgements obeyed the minimality principle of metric space, and the network was trained to make a judgement of 0 when presented such stimuli. The judgement of 0 reflected the minimality principle, in that a) the shortest distance in the space was from one location to itself, and b) this judgement was made for every location when it was compared to itself.

In this second simulation, the minimality principle was violated. For some of the cities, the network was trained to make a judgement of 2 when rating the distance of a city to itself (Camrose, Grande Prairie, Lloydminster, Red Deer, Slave Lake). For some of the other cities, the network was trained to make a judgement of 1 when rating the distance of a city to itself (Drumheller, Fort McMurray, Lethbridge, Medicine Hat). For the remaining cities, the network was trained to make a judgement of 0 when rating the distance of a city to itself (Banff, Calgary, Edmonton, Jasper). If one takes the diagonal entries of Table 1b and replaces the diagonal with the values listed here, then one has defined the problem for the second simulation.

The logic behind using these values was twofold. First, the point of the second study was to violate the minimality principle, which is accomplished by having different and nonzero values in the diagonal of Table 1b, regardless of what these particular values are. Second, human subjects are more likely to give higher self-similarity ratings (which would be converted into shorter distance ratings) to more familiar items (Tversky 1977). So, when we decided to modify the diagonal entries, we used a rating of 0 for the four most familiar Alberta cities, a rating of 2 for the five smallest and least familiar cities, and a rating of 1 for the remaining four cities.

Network architecture

The input unit and output representations in the second simulation were identical to those used in the first simulation. The only difference between the two simulations was that seven hidden units were used in the second simulation. The seventh hidden value unit was added because repeated testing showed that if only six hidden units were used, then the network would not converge upon a solution to the problem.

Network training

As before, the network was trained using the modified generalized delta rule developed by Dawson and Schopflocher (1992). Connection weights and biases were randomly started in the range from –0.50 to +0.50. The network was trained with a learning rate of 0.01 and with zero momentum; the order of pattern presentation was randomized every sweep through the training set.

The network converged on a solution to the problem (a "hit" for each output unit for every one of the 169 training patterns) after 2057 sweeps of training.

Results

The main issue of interest in the second simulation was the nature of the internal representations used by the network to make the distance ratings that violated the minimality principle of metric space. Given that this network required one extra hidden unit in comparison to the first, it was possible that six of its hidden units were identical to the six hidden units in Simulation 1, and that the seventh hidden unit was unique, and had the task of dealing with the nonmetric diagonal judgements. A second, more interesting, possibility is that all seven of the hidden units were all using allocentric coarse coding, and that this kind of coding was flexible enough to deal with the violation of the minimality principle. To examine the internal representations used by the Simulation 2 network, we approached its analysis in the same fashion as we approached the network in Simulation 1.

Relating the map of Alberta to hidden unit connection weights

Our first analysis involved using the Solver tool in Excel to find a position for each hidden unit on the map of Alberta, such that this position optimized the correlation between each unit's connection weights and the distances between the unit and the 13 Albertan cities. As was the case in Simulation 1, we found that a position could be found for each hidden unit that provided a substantial correlation between distances and hidden units (see Table 4). Furthermore, from an examination of this table, and a comparison between it and Table 2, there did not seem to be any indication that any one of the hidden units was qualitatively different from the other six.

Figure 4 illustrates the positions of the hidden units from Simulation 2 that were obtained by performing the analysis above. In comparison to Figure 3, we can see that the hidden units occupy different positions on the map. However, in both figures the hidden units occupy locations that are not occupied by cities. This property is likely crucial for the allocentric coarse coding to be able to handle the violation of the minimality principle. By being located away from cities, every distance feeding into a hidden unit (represented by connection weights) has to be nonzero.

Table 4. Results of relating Alberta map distances between cities and hidden units the values of the connection weights feeding into the hidden units in Simulation 2

Hidden Unit	Hidden Unit Latitude	Hidden Unit Longitude	Correlation Between Map Distances And Incoming Weights
H0	50.55	112.99	0.69
H1	54.84	115.17	0.54
H2	54.06	113.14	–0.69
H3	50.02	110.72	0.46
H4	54.51	112.20	–0.76
H5	51.17	115.57	–0.57
H6	52.75	115.44	–0.41

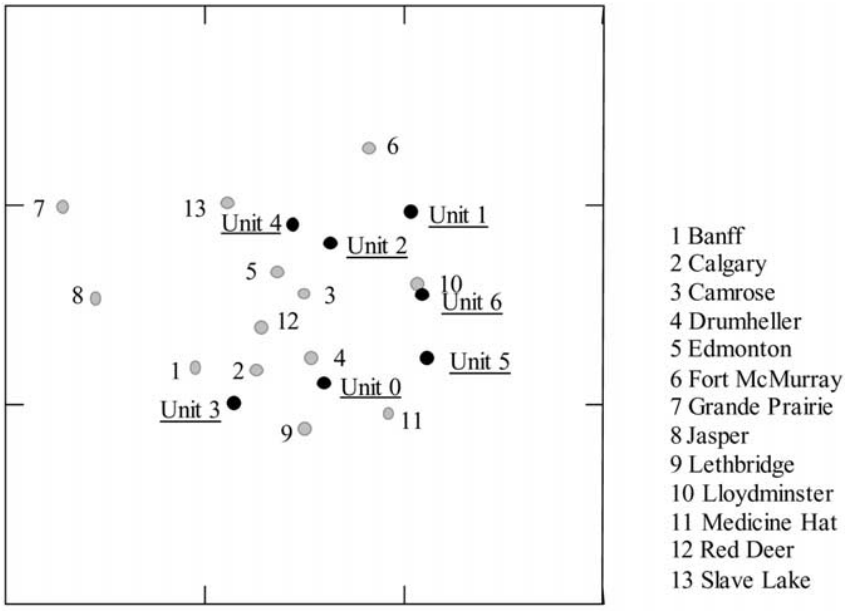

Figure 4. The locations of the seven hidden units from the second simulation on the map of Alberta. These locations optimized the correlations between hidden unit connection weights and distances from the cities to the hidden units.

Table 5. Results of relating connection weights to city distances from the MDS solutions obtained from the activity matrix for each hidden unit in Simulation 2. The table provides the maximum correlation, as well as the coordinates of the hidden unit in the space that produces this maximum correlation

Hidden Unit	X-Coordinate Of Hidden Unit	X-Coordinate Of Hidden Unit	Correlation Between MDS Distances And Connection Weights
H0	–4.08	–0.98	–1.00
H1	–0.27	–0.03	0.84
H2	0.14	–0.11	0.87
H3	0.00	0.12	0.80
H4	0.36	–0.02	0.58
H5	0.38	–0.48	–0.93
H6	–0.90	–0.35	0.88

Relating connection weights to hidden unit MDS spaces

As found in Simulation 1, while the correlations in Table 4 are substantial, they are far from perfect. To test the possibility that connection weights were more related to distances in a "distorted" two-dimensional space, we repeated the second analysis that was carried out in the first study: we took the activity matrix for each hidden unit, performed MDS on each activity matrix to identify a space for each hidden unit (as well as the coordinates for each city in this space). We found, once again, that the two dimensional MDS solution accounted for almost all of the variance in each data matrix (R^2 = 0.981, 0.991, 0.992, 0.998, 0.992, 0.996, 0.998 for the analysis of hidden units 1 through 7). We then used Solver to find the location, in its own space, for each unit that optimized the correlation between connection weights and city distances. As was the case in Simulation 1, this analysis led to very large correlations, which are presented in Table 5.

Coarse coding from hidden unit activations to distance ratings

In Simulation 1, an argument for coarse coding was made because the correlations between individual hidden unit activities and ratings were small, but the set of all activities combined were strongly related to distance ratings. A very similar circumstance was found in Simulation 2. When hidden unit activities were correlated with distance ratings, the correlations (for units H0 through H6 respectively) were –0.51, 0.22, –0.06, –0.03, –0.08, 0.06, and 0.17.

However, when the activities of the seven hidden units were combined in a regression equation that predicted distance ratings (including those ratings that violated the minimality principle), the equation produced an R^2 of 0.74 ($F[6,163] = 66.59,\ p < 0.0001$). In other words, a linear combination of hidden unit activities accounts for over 70% of the variance in the ratings. When this linear combination is transformed by the nonlinear activation functions in the output units, the result is near-perfect prediction (i.e., a converged network).

Coarse coding for violations of minimality

The analyses to this point indicate that the network that was trained in Simulation 2 developed similar types of internal representations to those found in the Simulation 1 network. Furthermore, there does not appear to be any evidence that the second simulation employed one hidden unit as a "specialist" with the task of dealing with the subset of ratings that represent violations of the minimality principle. To provide one additional test of this, we took the diagonal entries of the modified version of Table 1b and correlated them with the connection weights feeding into each hidden unit. Our previous analyses have shown that these connection weights represent distance information. If one of these units uses this information to handle minimality violations, then one would expect that there should be a strong relationship between its incoming connection weights and these violations.

The correlations that were computed for each of the seven hidden units were 0.36, –0.41, 0.22, –0.32, 0.30, –0.12, and –0.32. There are two observations to make about these correlations. First, none of them is large enough to indicate that one hidden unit is capable of dealing with the violations of minimality on its own. Second, many of the hidden units have correlations that are roughly the same size. Together, these observations indicate that no one hidden unit is being used to handle the violations of minimality. Instead, these violations appear to be dealt with in the same manner as the other ratings in the matrix – by coarse coding that involves all seven hidden units.

Discussion

In Simulation 1, a PDP network was trained to make ratings of distances between cities that were consistent with the minimality principle for metric space. This was accomplished by having the network trained to generate a rating of 0 whenever asked to judge the distance between a city and itself. In

Simulation 2, a PDP network was trained to make ratings that violated the minimality principle. When asked to judge the distance between a city and itself, the network was trained to generate a rating of 0 for four cities, a rating of 1 for four other cities, and a rating of 2 for the remaining five cities.

When the minimality constraint was violated, the ratings task became more difficult. In particular, the problem could not be solved when the network had six hidden units (the number used in Simulation 1). An additional hidden unit was required for the network to converge to a solution to the problem.

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. Once again, the network used allocentric coarse coding to make distance judgements. 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 judgements were possible. Finally, there was no evidence that any one of the hidden units had a special role in making the subset of judgements that defined the violation of the minimality principle.

General discussion

In the introduction, we briefly reviewed three different research areas related to spatial cognition: similarity spaces, mental imagery, and cognitive maps. For each of these areas, we argued that there existed a tension between behavioural 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 behavioural 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 behavioural regulari-

ties emerged from the procedures used to compare feature sets (Malgady and Johnson 1976; Ortony 1979; Tversky 1977; Tversky and Gati 1982). This approach has the advantage of being able to account for nonmetric behavioural regularities. However, it has disadvantages as well. The ability to fit nonmetric behaviour 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 behaviours. 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 neighbours 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 in the current manuscript. Rather than analyzing behavioural regularities in detail, and proposing a representational story that attempted to capture these regularities, we took a synthetic perspective. 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, we took great pains to analyze its internal structure. The result was that we found a particular kind of representation, allocentric coarse coding, that would not have been an obvious proposal had our starting point been the analysis of behaviour. 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 in the current paper has shown how a connectionist representation can account for both metric and nonmetric regularities. Our current research involves taking a more detailed look at allocentric coarse coding in order to get a clearer picture of its advantages and disadvantages.

Acknowledgements

This research was supported by NSERC Research Grant A2038 awarded to MRWD. Many thanks to Dr. Richard Harshman of the psychology department at the University of Western Ontario for his helpful suggestions regarding data analysis. Correspondence should be addressed to Dr. Michael Dawson, Department of Psychology, University of Alberta, Edmonton, Alberta, CANADA T6G 2E9.

References

Anderson, J.R. (1978). Arguments Concerning Representations for Mental Imagery, *Psychological Review* 85: 249–277.

Arleo, A. and Gerstner, W. (2000). Spatial Cognition and Neuro-mimetic Navigation: A Model of Hippocampal Place Cell Activity, *Biological Cybernetics* 83: 287–299.

Ballard, D. (1986). Cortical Structures and Parallel Processing: Structure and Function, *The Behavioural And Brain Sciences* 9: 67–120.

Bannon, L.J. (1980). *An Investigation of Image Scanning*. Ontario: Unpublished doctoral dissertation, University of Western Ontario, London.

Berkeley, I.S.N., Dawson, M.R.W., Medler, D.A., Schopflocher, D.P. and Hornsby, L. (1995). Density Plots of Hidden Value Unit Activations Reveal Interpretable Bands, *Connection Science* 7: 167–186.

Block, N. (1981). *Imagery*. Cambridge, MA: MIT Press.

Blumenthal, L.M. (1953). *Theory And Applications Of Distance Geometry*. London: Oxford.

Braitenberg, V. (1984). *Vehicles: Explorations in Synthetic Psychology*. Cambridge, MA: MIT Press.

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. and O'Keefe, J. (1999). Robotic and Neuronal Simulation of the Hippocampus and RNavigation. In B.N., K.J. Jeffery and J. O'Keefe (eds.), *The Hippocampal and Parietal Foundations of Spatial Cognition*. Oxford: Oxford University Press.

Burgess, N., Reece, M. and 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.

Cheng, K. and Spetch, M.L. (1998). Mechanisms of Landmark Use in Mammals and Birds. In S. Healy (ed.), *Spatial Representation In Animals*. Oxford: Oxford University Press.

Churchland, P.S. and Sejnowski, T.J. (1992). *The Computational Brain*. Cambridge, MA: MIT Press.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.

Dawson, M.R.W. (1998). *Understanding Cognitive Science*. Oxford, UK: Blackwell.

Dawson, M.R.W. and 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. and 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.,and 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. and Schopflocher, D.P. (1992). Modifying the Generalized Delta Rule to Train Networks of Nonmonotonic Processors for Pattern Classification, *Connection Science* 4: 19–31.

Farah, M.J., Weisberg, L.L., Monheit, M. and Peronnet, F. (1989). Brain Activity Underlying Mental Imagery: Event-related Potentials During Mental Image Generation, *Journal of Cognitive Neuroscience* 1: 302–316.

Fylstra, D., Lasdon, L., Watson, J. and Waren, A. (1998). Design and Use of the Microsoft Excell Solver, *Interfaces* 28(5): 29–55.

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

Ghiselli-Crippa, T.B. and Munro, P.W. (2000). Effects of Spatial and Temporal Contiguity on the Acquisition of Spatial Information. In S. Solla, T. Leen and K.-R. Muller (eds.), *Advances In Neural Information Processing Systems 12*. Cambridge, MA: MIT Press.

Gill, P.E., Murray, W. and Wright, M.H. (1981). *Practical Optimization*. London: Academic Press.

Grush, R. (2000). Self, World and Space: The Meaning and Mechanisms of Ego- and Allocentric Spatial Representation, *Brain And Mind* 1: 59–92.

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. and 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. and 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. and 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. and Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills, CA: Sage Publications.

Lasdon, L.S., Waren, A.D., Jain, A. and Ratner, M. (1978). Design and Testing of a Generalized Reduced Gradient Code for Nonlinear Programming, *ACM Transactions on Mathematical Software* 4: 34–49.

Leighton, J.P. (1999). *An Alternate Approach to Understanding Formal Reasoning: Thinking According to the Inductive-coherence Model.* Edmonton: Unpublished Ph.D., University of Alberta.

Malgady, R.G. and 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.

McClelland, J.L., Rumelhart, D.E. and Hinton, G.E. (1986). The Appeal of Parallel Distributed Processing. In D. Rumelhart and J. McClelland (eds.), *Parallel Distributed Processing* (Vol. 1). Cambridge, MA: MIT Press.

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. and 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. and Gentner, D. (1993). Respects for Similarity, *Psychological Review* 100(2): 254–278.

Mellet, E., Petit, L., Mazoyer, B., Denis, M., and Tzourio, N. (1998). Reopening the Mental Imagery Debate: Lessons from Functional Anatomy, *Neuroimage* 8: 129–139.

Minsky, M. (1985). *The Society of Mind*. New York: Simon and Schuster.

O'Keefe, J. and Burgess, N. (1996). Geometric Determinants of the Place Fields of Hippocampal Neurones, *Nature* 381: 425–428.

O'Keefe, J. and 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. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press.

Ortony, A. (1979). Beyond Literal Similarity, *Psychological Review* 86: 161–180.

Orvis, W.J. (1996). *Excel for Scientists and Engineers* (Second ed.). San Francisco, CA: Sybex.

Pfeifer, R. and Scheier, C. (1999). *Understanding Intelligence*. Cambridge, MA: MIT Press.

Pylyshyn, Z.W. (1973). What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery, *Psychological bulletin* 80: 1–24.

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, *Behavioural 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.

Redish, A.D. (1999). *Beyond the Cognitive Map*. Cambridge, MA: MIT Press.

Redish, A.D. and Touretzky, D.S. (1999). Separating Hippocampal Maps. In B.N., K.J. Jeffery and J. O'Keefe (eds.), *The Hippocampal And Parietal Foundations Of Spatial Cognition*. Oxford: Oxford University Press.

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. and Nerlove, S.B. (1972). *Multidimensional Scaling: Theory and Applications in The Behavioural Sciences. Volume II: Applications*. New York, NY: Seminar Press.

Rumelhart, D.E. and Abrahamson, A.A. (1973). A Model for Analogical Reasoning. *Cognitive Psychology* 5: 1–28.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). Learning Representations by Back-propagating Errors, *Nature* 323: 533–536.

Seidenberg, M. and McClelland, J. (1989). A Distributed, Developmental Model of Word Recognition and Naming, *Psychological review* 97: 447–452.

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 and S.B. Nerlove (eds.),

Multidimensional Scaling: Theory and Applications in the Behavioural Sciences. Vol 1: Theory (pp. 21–47). New York, NY: Seminar Press.

Shepard, R.N. and Cooper, L.A. (1982). *Mental Images and their Transformations*. Cambridge, MA: MIT Press.

Shepard, R.N., Romney, A.K. and Nerlove, S.B. (1972). *Multidimensional Scaling: Theory and Applications in the Behavioural Sciences. Volume I: Theory*. New York, NY: Seminar Press.

Sherry, D. and Healy, S. (1998). Neural Mechanisms of Spatial Representation. In S. Healy (ed.), *Spatial Representation in Animals*. Oxford: Oxford University Press.

Simon, H.A. (1996). *The Sciences of the Artificial, Third Edition*. Cambridge, MA: MIT Press.

Smolensky, P. (1988). On the Proper Treatment of Connectionism, *Behavioural and Brain Sciences* 11: 1–74.

Thompson, W.L., Kosslyn, S.M., Sukel, K.E. and 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. and Sternberg, R.J. (1981). Aptness in Metaphor, *Cognitive psychology* 13: 27–55.

Tourangeau, R. and Sternberg, R.J. (1982). Understanding and Appreciating Metaphors, *Cognition* 11: 203–244.

Touretzky, D.S., Wan, H.S. and Redish, A.D. (1994). Neural Representation of Space in Rats and Robots. In J.M. Zurada, R.J. Marks and 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. and Gati, I. (1982). Similarity, Separability, and the Triangle Inequality, *Psychological Review* 89: 123–154.

Zimmerman, C.L. (1999). *A Network Interpretation Approach to the Balance Scale Task*. Edmonton: Unpublished Ph.D., University of Alberta.