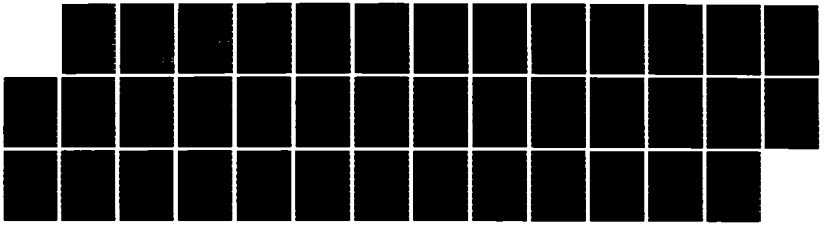
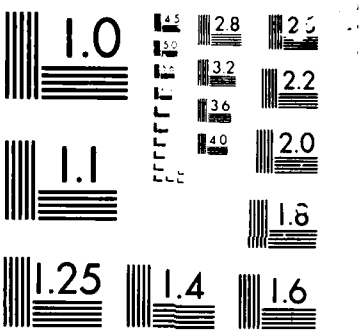
AD-A169 317 NEURAL REPRESENTATION OF CONCEPTUAL KNOWLEDGE(U) 1/1
ROCHESTER UNIV NY DEPT OF COMPUTER SCIENCE J A FELDMAN
JUN 86 TR-189 N00014-84-K-0655

UNCLASSIFIED F/G 6/4 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.8
3.2
3.6
4.0

Neural Representation of Conceptual Knowledge

Jerome A. Feldman
Department of Computer Science
The University of Rochester
Rochester, New York 14627

TR189
June 1986

Department of Computer Science
University of Rochester
Rochester, New York 14627

Neural Representation of Conceptual Knowledge

Jerome A. Feldman
Department of Computer Science
The University of Rochester
Rochester, New York 14627

TR189
June 1986

Abstract

The neural encoding of memory is a problem of great interest and importance. Traditional proposals have taken one of two extreme views: The one-concept, one-neuron, punctate view and the fully distributed, holographic alternative. Major advances in the behavioral, biological and computational sciences have greatly increased our understanding of the question and its potential answers. There is now good reason to reject both extreme views, but a compact encoding that derives from the punctate model appears to fit well with all the facts. Much of the work espousing holographic models is reinterpreted as studying system properties of neural networks and, as such, considered to be of great importance. Some suggestions for directions of further research are discussed.

This work was supported in part by ONR grant No. N00014-84-K-0655, and in part by NSF Coordinated Experimental Research grant No. DCR-8320136.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: TR 189	2. GOVT ACCESSION NO.: ADA169317
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): Neural Representation of Conceptual Knowledge	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Jerome A. Feldman	8. CONTRACT OR GRANT NUMBER(s): N00014-84-K-0655
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Computer Science, University of Rochester, Rochester, New York 14627	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Information Systems, Arlington, Virginia 11127	12. REPORT DATE: June 1986
	13. NUMBER OF PAGES: 35 pages
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Office of Naval Research, Information Systems, Arlington, Virginia 22217	15. SECURITY CLASS. (of this report): -
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report):

Distribution of this document is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Connectionist, neural model, distributed, conceptual knowledge

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The neural encoding of memory is a problem of great interest and importance. Traditional proposals have taken one of two extreme views: The one-concept, one-neuron, punctate view and the fully distributed, holographic alternative. Major advances in the behavioral, biological and computational sciences have greatly increased our understanding of the question and its potential answers. There is now good reason to reject both extreme views, but a compact encoding that derives from the punctate model appears to fit well with all the facts, Much of the work espousing holographic models is reinterpreted.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

How it is that the brain codes, stores and retrieves memories is among the most important and baffling questions in science. [Thompson 1986]

1. Introduction and Punctate Theories

In addition to its compelling scientific interest, the neural substrate of memory is a question of considerable practical importance. There are obvious applications in neurology, but it goes well beyond that. Our understanding of how human minds incorporate and process information has a profound influence on many aspects of social intercourse including formal and informal education, psychotherapy, public information and inter-personal communication [Roediger 1980]. Research in the behavioral and brain sciences entrails implicit assumptions about neural encoding.

Recent advances in the behavioral, biological and computational sciences may yield a major improvement in our understanding of the neural coding of memory. Neurobiology is making remarkable strides in elucidating the structure of the nervous system and the details of its functioning. The behavioral sciences are developing deep structural models of language, vision, etc. and employing increasingly sophisticated experimental and simulation techniques to refine them. Computer science has produced powerful devices and profound theories of representation and computation which are supporting the study of how the complex structural theories of the behavioral scientists could be carried out by the information processing mechanism revealed by neurobiology. The neural encoding of conceptual knowledge is one central question in this enterprise.

The entire paper is predicated on the direct neural encoding of conceptual structure -- could it be otherwise? In a conventional digital computer, the wiring diagrams tell us virtually nothing about the information structure stored in the system; the computer is a general purpose interpreter. Almost all its hardware is passive memory which is idle except when accessed by the interpreter. Many cognitive scientists still like to think of the brain in this way and, for them, the study of neural encodings makes no sense. One explores the high level structure of cognition in a way that is independent of any embodiment. A great deal can be learned in this way and this paper relies heavily on the results of such studies. The problem with this stance is that the speed of the system (brain) relative to that of its elements (neurons) places very severe constraints on the possible organization of knowledge. Human reactions, over a wide range of tasks, take a few hundred milliseconds or about 100 times the switching time of individual neurons. Thus, for example, an interpreter of the standard kind would be too slow by many orders of magnitude. While there are in principle a wide range of possible realizations of intelligence, we know that our brains are constrained to use one that is quite direct. One can also view this paper as an exploration of the relationships among three kinds of structure: the structure of the brain, the structural characterization of behavior and the conceptual structure of knowledge. The questions addressed in this paper are 1) What do we now know about the range of possible encodings of knowledge in the brain? and 2) What does this suggest for experimental and theoretical research? All of this falls within a research domain characterized by terms like those in Table 1.

Common Terms
Connectionist
Neural Model
Massively Parallel
Cell Assembly
Pattern of Activation
Parallel Distributed Processing

Table 1

We will pursue these questions following what has become the paradigmatic method of cognitive science -- converging constraints. Neither behavioral, biological nor computational results alone greatly restrict theorizing, but taken together these results preclude many models that look plausible from a narrow perspective. The 100-step rule for simple behaviors is an elementary example of how converging constraints can lead to insights.

For concreteness, let us suppose that the problem is to describe how a human brain could represent and exploit the information in a standard encyclopedia. We assume that any representation will employ some notion of "concept" which corresponds roughly to a dictionary word-sense. The only other primitive needed is the "relation." We assume that all the required knowledge can be captured as a collection of relations among concepts. The encyclopedia contains pictures so some appearance models are required, but these can also be expressed in terms of concepts and relations. There is no attempt to define formally the notion of concept and relation because the various representational schemes discussed will entail properties of the primitives. What we will do is show how different neural encoding models relate to constraints arising from various branches of science. To begin, we will focus only on the representation of concepts, ignoring relations as most of the literature continues to do.

Even leaving relations aside, the literature contains a wide range of notions of what a concept is and how it might be represented neurally. The simplest view restricts consideration to very concrete nouns such as "horse" or "chair," ignoring more complex concepts such as "game," "yesterday," "active" or "love." Even in this quite restricted context, the representation of concepts has proven to be a deep problem for philosophers, linguists, psychologists, etc. [Smith & Medin 1981]. Our point of departure is the computational models of Artificial Intelligence, which have made explicit many of these issues in conceptual knowledge [Charniak & McDermott, 1985]. A minimum requirement is that the representation support the answering of questions about concepts (and later about their relations). One particularly simple kind of question concerns the structure of the object itself, e.g., how many legs has some chair. Models which treat concepts as unstructured collections of attributes will need to provide additional mechanisms for answering even these simple questions. Many studies of neural concept representations ignore

entirely the issue of how concepts are used, but no serious attempt to understand real animals can afford to do so.

The range of possible representations of concepts in the brain is constrained by a combination of computational considerations and neurobiological findings. The 100-step rule for simple tasks already eliminates any conventional computer model. Other pertinant facts include the relatively small number of neurons (about 10^{11} or one-hundred billion), the large number of connections between neurons (about 10^4 per unit), and the low rate of information transfer. It may seem that 10^{11} is not a small number, but when one considers the 10^6 input fibres from each eye, a computer scientist immediately detects a major constraint. For example, dedicating one unit to test for a possible line between any pair of points in the retina would take more neurons ($(10^6)^2$) then there are. The information rate between individual neurons at a firing rate of 100 spikes per second is about five bits or enough to encode one letter of the alphabet. (If complex messages are being conveyed, it is not by individual neurons.) Much of the computational power of the brain derives from its great connectivity and a challenge to theory is to explain how this power is realized. Although this paper concentrates on representation more than learning, there are also constraints on plasticity that are pertinent. The major finding is that the growth of new fibers in adults is much too slow and constrained to account for learning and that there is no generation of new neurons. There is sufficient converging evidence [Lynch 1986] to allow us to assume that long-term concept and relation memory comes about through changes in connection strength, but rapid connection change is problematical. There is also reason to believe that skill learning may involve significantly different mechanisms than concept learning [Thompson 1986].

The best way to begin a serious discussion of neural encoding of concepts is to tie down two simple theories that embody the extreme ends of the range of possible answers. The most compact representation possible would be to have a unique unit dedicated to each concept. If we assume that a unit corresponds to one neuron then this is the "grandmother cell" or "pontifical cell" theory. The other extreme would be to have each concept representated as a pattern of activity in all of the units in the system. This is well known as the "holographic" model of memory and is the most highly distributed theory that we will consider. In addition to the pure theory based on optical holograms, we will call holographic any model that has all the units in a system encoding each concept [Willshaw 1981]; most of these are matrix formulations. Nothing much would change in either theory if a "unit" corresponded to a dendritic sub-tree instead of an entire neuron. The discussion will proceed by moving in from the extremes to examine the range of plausible encoding models and some of their properties. A great deal of excellent work has been done employing the end-case assumptions, but neither of them could actually be right.

The extreme opposing models of neural representation lead to radically different views of many aspects of Cognitive Science. Table 2 presents a number of contrasting terms that arise, respectively, from the punctate and fully distributed views of neural coding. Not all of these items will be meaningful to every reader, but everyone should recognize some striking contrasts in perspective. All of these contrasting notions will find their way into the discussion. One term that is possibly misleading in Table 2 is "distributed representation." The problem is that people have been using this term to denote everything from a fully holographic model to one where two units help code a concept, thus the term has lost its usefulness. The various contrasting terms often accompany significant differences in research goals and strategies. One intriguing idea that we will pursue is that many of these

Contrasting Terms

punctate	diffuse
local	distributed representation
grandmother neuron	hologram, spin-glass
disjoint codes	homogeneous code
detector	filter
labelled line	pattern of activity
active memory	passive memory
reduction	emergence
hierarchy	complete connectivity
recruiting	adapting
general computation	correlation

Table 2

apparent conflicts are basically alternative ways of looking at the same set of phenomena, vaguely reminiscent of atomic physics and thermodynamics.

We should first dispense with an abstract argument that equates the punctate and hologram models. It is true, in a sense, that an encoding having one active unit per concept is a pattern of activity in the mass. But this identification is too abstract to be meaningful. Another proposed way to identify holographic and punctate representations comes from linear algebra and the idea of alternate coordinate axes for a vector space. If, as in many models, the output of a unit is the (thresholded) linear combination of its inputs, one can view this unit as a (very large) vector, v , whose coordinates are the outputs of each predecessor unit. There is, in principle, another set of bases for the vector space for which this vector, v, is an axis and can therefore be represented by one non-zero coordinate. This argument fails for three reasons. Even for strictly linear input combinations, the output threshold destroys the applicability of linear algebra and there is no biologically plausible way to eliminate non-linearity. In addition, no single transform would work unless the set of concepts were independent and therefore small (cf. Section 2). More importantly, the computational properties of the two representations are radically different as with the example just above.

Punctate Models

The next step is to show why neither the pure punctate nor the holographic model are consistant with the facts. We will start with the punctate model because the story here is simpler.

The extreme end of the compact representation position is to assume each concept is represented by exactly one neuron. This view received considerable support from single unit recording research, which found that units in sensory areas responded best to a relatively narrow class of stimuli [Hubel & Weisel 1979]. The punctate encoding is also called "labelled lines" emphasizing the fact that, in this encoding, each axon will be conveying a specific message when it is active. The most influential expression of this position was Horace Barlow's "neuron doctrine" which is worth quoting in its outline form [Barlow, 1972].

> The following five brief statements are intended to define which aspect of the brain's activity is important for understanding its main function, to suggest the way that single neurons represent what is going on around us, and to say how this is related to our subjective experience. The statements are dogmatic and incautious because it is important that they should be clear and testable.
>
> *First dogma*
>
> A description of that activity of a single nerve cell which is transmitted to and influences other nerve cells, and of a nerve cell's response to such influences from other cells, is a complete enough description for functional understanding of the nervous system. There is nothing else 'looking at' or controlling this activity, which must therefore provide a basis for understanding how the brain controls behaviour.
>
> *Second dogma*
>
> At progressively higher levels in sensory pathways information about the physical stimulus is carried by progressively fewer active neurons. The sensory system is organized to achieve as complete a representation as possible with the minimum number of active neurons. ([cf. Figure 7])
>
> *Third dogma*
>
> Trigger features of neurons are matched to the redundant features of sensory stimulation in order to achieve greater completeness and economy of representation. This selective responsiveness is determined by the sensory stimulation to which neurons have been exposed, as well as by genetic factors operating during development.
>
> *Fourth dogma*
>
> Just as physical stimuli directly cause receptors to initiate neural activity, so the active high-level neurons directly and simply cause the elements of our perception.

Fifth dogma

The frequency of neural impulses codes subjective certainty: A high impulse frequency in a given neuron corresponds to a high degree of confidence that the cause of the percept is present in the external world.

The five dogmas are much less controversial in their original form than in most of the strawman characterizations derived from them. Obviously enough, much of their force is in defense of the direct encoding position shared by all connectionist models. The most relevant point in our discussion is that individual neurons are the appropriate object for study in determining how the brain does its work. This is commonly identified with the most compact, punctate, style of neural model and is an appropriate introduction to this end of the spectrum.

Figure 1 (after Shastri [1985]) presents a vastly oversimplified version of how a punctate system might encode and exploit conceptual knowledge. The memory network is a category-based hierarchy with each concept and property-name represented by a rectangular unit. The triangular nodes stand for intermediate units which become active when two of their three input lines are active. Suppose the system has a routine that retrieves its knowledge of food tastes as an aid to ordering wine, such as that cartooned in the lower half of the figure. If activation is spread simultaneously to the "main course" of the meal and to the desired property "has-taste," exactly one triangular evidence node, b_1, will receive two active inputs and this will lead to the activation of the concept "salty." This is the required answer, but for technical reasons an intervening clean-up network is needed where the answer is acutally put to use. One interesting feature of Shastri's model is that the same memory network is able to classify a salty, pink food as ham -- the triangular evidence nodes work in both directions.

While oversimplified, Figure 1 does convey much of the flavor of punctate (and other compact) connectionist models and their appeal to some scientists and rejection by others. The main point is that everything is quite explicit; the concepts, properties and even the rules of operation are simple and direct. This makes it relatively easy to express and test specific models either at the neural level or more abstractly as in our example. No one believes that the brain uses exactly the structure of Figure 1, but any highly compact concept representation could behave in essentially the same way. The very explicitness of all this is what makes many scientists reject punctate models either for abstract study or as a theory of brain structure. While any particular structure or theory can be encoded (just like circuit design), nothing may be learned about the general properties of intelligence. Moreover, how could such a hard-wired sytem develop and adapt in living brains? We will discuss this last issue in Section 3; the others are questions of scientific taste and judgment. But the answer to how the brain represents knowledge is not a matter of taste and we next explore some arguments that it can not be in the punctate style of Figure 1.

The first point is that a large number of neurons ($\sim 10^5$) die each day and these are distributed throughout the brain. If each concept were represented by exactly one neuron, one would expect to lose at least some concepts (at random) each day. This argument is often taken to be conclusive evidence against the compact model, but there is a slight variant of the punctate view that is proof against the death of

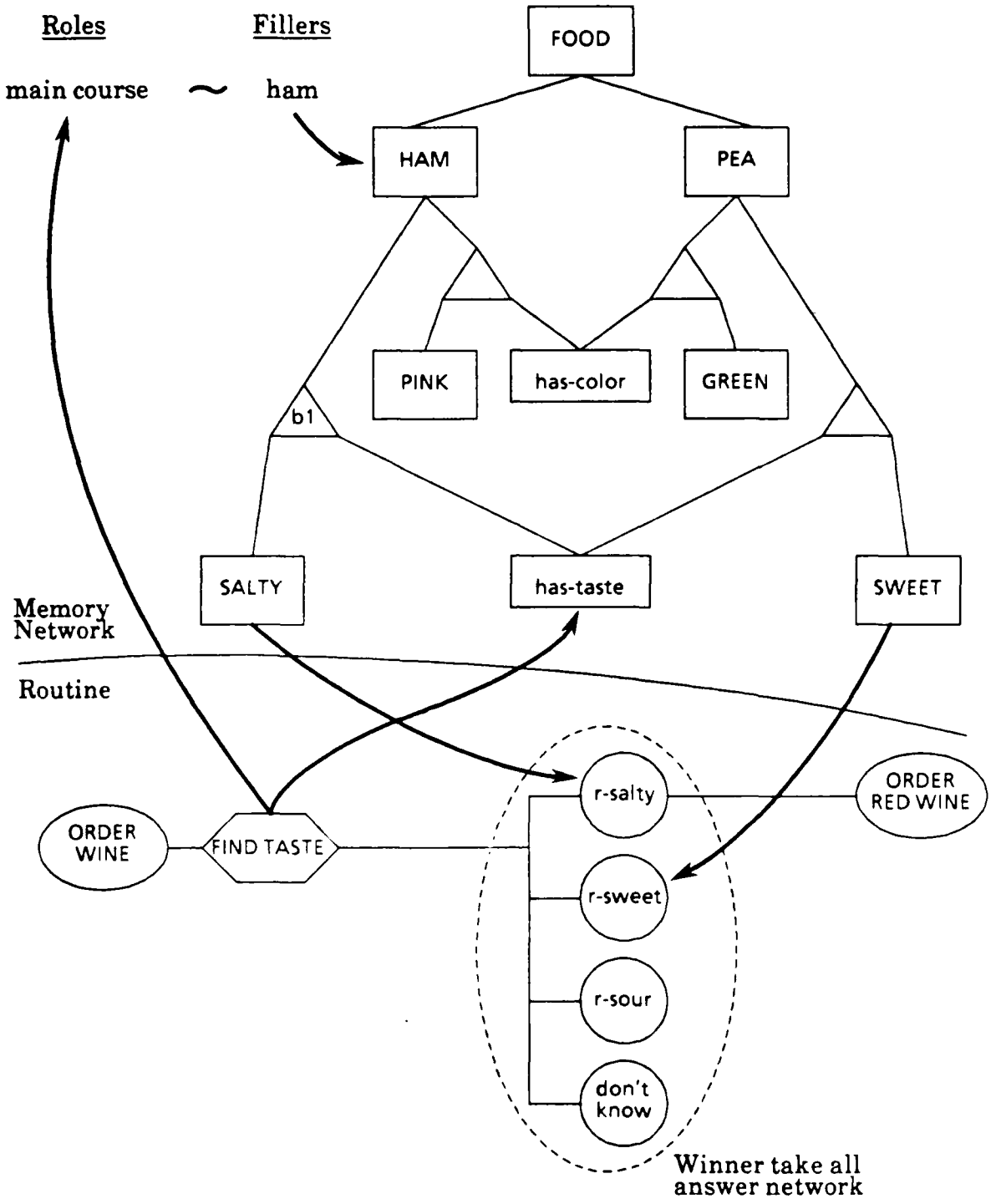


Figure 1: Connectionist Retrieval System

individual units. Suppose that instead of one unit per concept, the system dedicated three or five, distributed somewhat physically. All of the theories and experiments would look the same as in the one unit case, but the redundancy would solve the problem of neuron death. While the number of neurons dying is large, the fraction is quite small ($\sim 10^{-6}$) so the probability of losing two of the representatives of a concept in a lifetime is quite low ($\sim 10^{-7}$). There is also considerable redundancy at the system level, several separate ways to perform a given function. As we will see, there are other reasons for rejecting the punctate model, but the death of neurons is not critical.

Another, and more telling, argument against a purely punctate model is that there would not be enough units to represent everything. It is fine to envision a unit that is active when you recognize your grandmother's picture in the encyclopedia, but what about all your other memories of her. Is there a separate unit for each age, each outfit, each body position, etc.? There are two aspects to this argument, the sheer number of things to be represented and the relational structure among them. Merely representing all the possible concepts is not a problem if efficient encodings are used. For example, a group of one hundred binary units can represent 2^{100} ($\sim 10^{30}$) distinct patterns which is far more than the number of concepts required. A million such groups is ample memory and is a small fraction of the neurons available ($\sim$ .1%). The most efficient coding information is not the main design criterion, but there are constraints which preclude some encodings as too wasteful. Supporters of the compact end of the spectrum have conceded this and have developed coding techniques which permit the encoding of significantly more information without a major change in computational architecture.

In parallel with these theoretical developments, experimentalists have been reframing their view of the activity of single units. For some time, the idea of a neuron as a "detector" of one kind of event has been declining but no alternative term has yet evolved. The psychophysicist's idea of a "filter" is the diffuse equivalent (all units filter all signals) and equally misleading. Although no new word has been established, many experimentalists now (correctly, in my view) view sensory neurons as having responsivity of different fineness to a variety of stimulus dimensions. Single unit neurophysiology studies are finding effects of stimuli beyond the classical receptive field [Allman *et al.* 1985] as is evitable in an interacting system of units. And, of course, the idea that all of intelligence could be understood simply by finding the neuron for each concept was always silly. It is clear that Barlow's dogmas could use some revision and Appendix A suggests one variant which seems reasonable. For example, dogma five should reflect the fact that firing frequency encodes some stimulus feature information as well as confidence. Figure 2 shows the firing rate of a visual system neuron as a function of binocular temporal and spatial offset of a target and is typical. When one includes the firing rate (instead of just on/off) as part of the information code, the rage of distinct entities representable goes up significantly [Ballard 1986b]. For example, the relative activity rate of three color "channels" is enough to specify a wide range of hue and intensity combinations. Summarizing the discussion to this point: While the purely punctate view is unsupportable, there is no numerical problem with a theory that has each concept represented by the activity of a few units. Such an encoding shares many of the properties of the punctate model and is consistant with single unit experiments in a wide range of brain structures.

Another argument used against compact models of neural representation arises from bulk activity experiments. A single small stimulus can give rise to activity in a significant fraction of the total population. There are several reasons why this fact

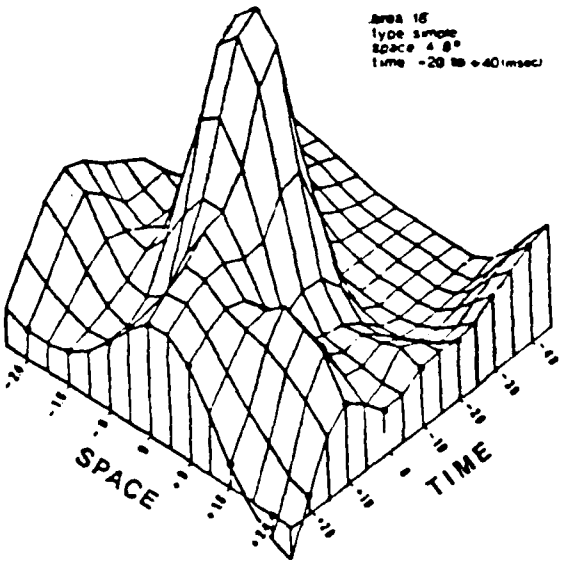


Figure 2: Response of a cat visual (area 18) cell to a binocular stimulus as a function of temporal and spatial offset. Spatial offset is from -2.4 to 2.4 degrees, temporal from −20 to +40 m.sec. (from [Gardner *et al.* 1985]).

does not preclude compact representations. For one thing, the "simplicity" of a stimulus looks different from various encoding schemes. A small dot, in a Fourier encoding, activates receptors for all spatial frequencies. In a parameter-space encoding [Ballard 1986a], a single feature provides evidence for many higher-level features and these may be suppressed slower than when more information is present. Finally, there does appear to be some non-specific gating [Barber 1980] which activates an entire structure when any input appears. None of this is to say that the punctate position is correct, only that arguments from bulk activity do not preclude its viability.

While the cell-death and bulk activities arguments against punctate representation are easily accommodated, it is easy to see that there really could not be enough neurons to have one for each concept of interest. One clear example arises in early vision. It is well known that the visual system is sensitive to at least the following local stimulus properties: orientation, intensity, hue, depth, motion direction and size. A system that could resolve 10 values for each of the six dimensions would require 10^6 units to represent all combinations of values. But there are about 10^6 separate points at the narrowest (retinal ganglion) level that must be represented so the total requirement would be 10^{12} which is too many. Similar combinatorial arguments can be made at higher level conceptual levels, as in all the memories of one's grandmother. At these higher levels, the structure of the knowledge itself provides considerable encoding economy (Section 3), but for early

processing additional mechanisms are required. In fact, quite a lot is known about the response properties of cells in visual cortex and there is a clean computational story that both helps organize the experimental data and contributes to our goal of constraining the possible models of neural encoding.

The basic idea is depicted in Figure 3 for the case of two stimulus dimensions. Suppose (as turns out to be the case) that units respond non-uniformly to various stimulus dimensions. For example, the vertical rectangle in the lower left depicts the responsiveness of a cell that is five times more sensitive to size than to orientation and the horizontal rectangle the opposite. The nice point is that the joint activity of two such cells can code the stimulus space as finely as the finest dimension of either (cf. the crossed rectangles), while requiring significantly fewer units. This computational mechanism goes by the name "coarse-fine coding" and appears to describe a good deal of neural computation. In general, given K stimulus dimensions, each with a desired resolution of N values, the punctate encoding requires N^K units. A coarse-fine encoding with the coarse dimensions D requires a total of

$$T = K \cdot N\left(\frac{N}{D}\right)^{K-1}$$

units. This is because there will be K separate tilings (covers) of the K-dimensional space, but each will be covered coarsely and this requires (N/D) units in all dimensions but one, which has N units. This formula still grows exponentially, but is significantly smaller for the cases of interest. For example, our early vision example had K = 6 and N = 10. With D = 5, this yields 1920 units per point instead of the 1,000,000 for the pure punctate encoding. Since I believe these ideas to be central, a number of related issues will be discussed next.

The critical point in the construction is the overlap of receptive fields, not their assymmetric shape. Essentially the same arguments can be made for symmetric overlapping fields and this is known as "coarse-coding" [Hinton 1981]. Computational ideas of this kind have been known for some time to provide a nice account of hyper-accuity, the ability of people to resolve details finer than the spacing of their receptors. In both coarse and coarse-fine coding there is a price to be paid for saving all these units. If two stimuli that overlap co-occur, the system will be unable to resolve them. This situation is depicted in Figure 4. Suppose one stimulus is encoded by the two rectangles labelled X and the other by the two rectangles labelled Y. The intersections labelled X and Y encode the desired information, but the "ghosts" labelled G would be equally active. One of these expresses the conjunction of Y-size and X-orientation and the other the converse. This is an instance of "cross-talk" in neural encodings. Cross-talk is the fundamental problem of shared encodings and appears to be a critical limiting factor on distributed models, as we will see. For the simple case of coarse coding, some analysis of the trade-offs has been carried out [Sullins 1985]

A final point on coarse coding is that the multiple-dimension, fine-grained information is left implicit in the representation. The joint activity of several units encodes the desired information, but how can subsequent computation make use of it? This is another critical issue in shared encodings and a major reason why no holographic model has gone more than one level. In the case where the information is carried by the activity of a small number of units (K ~ 10), there is a simple and biologically plausible solution. Suppose that the inputs to a unit (neuron) were not all treated uniformly, but were grouped into "sites" each of which computed separately a combined input value. If each site computed the logical AND of all its

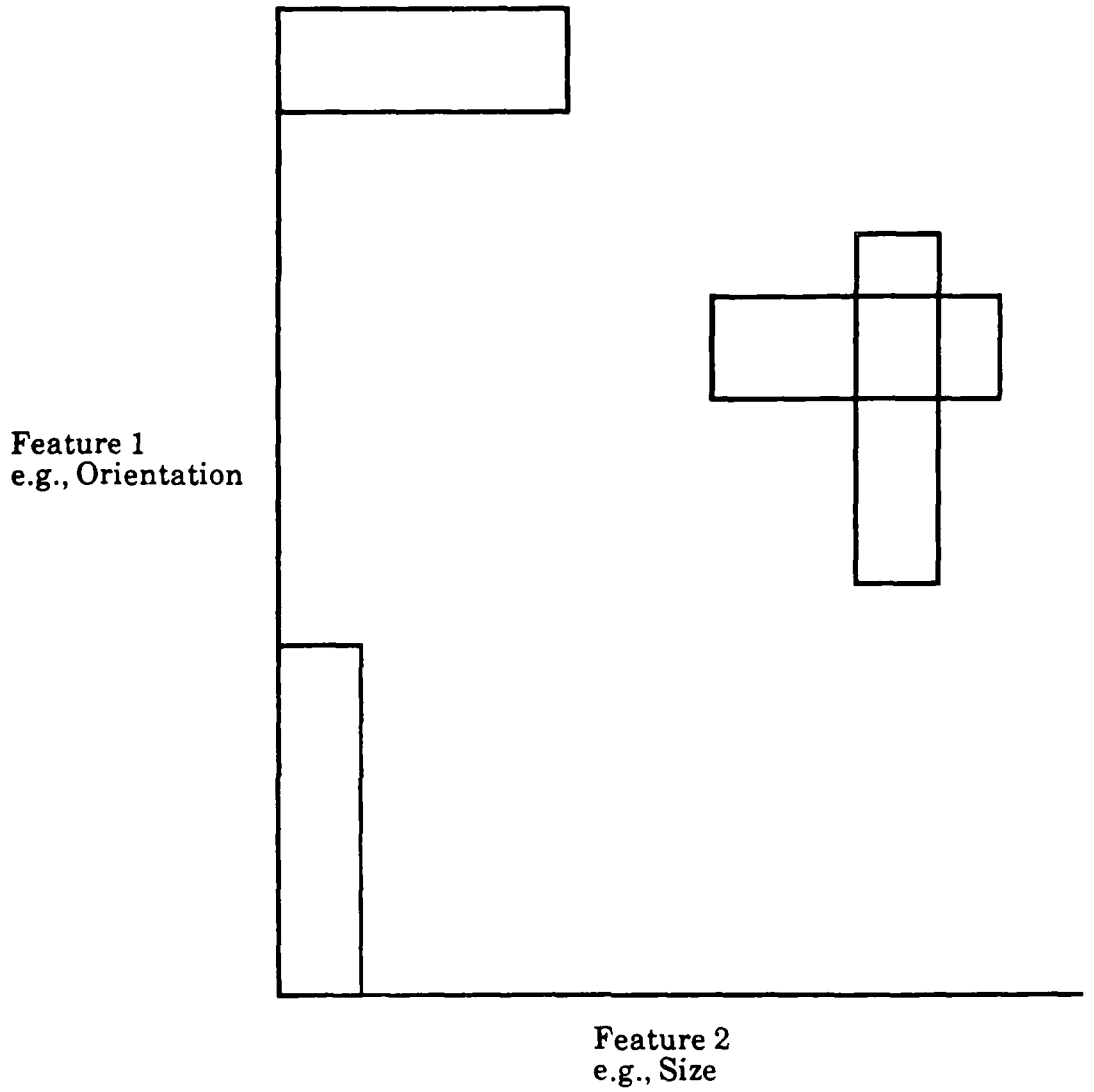


Figure 3: **Coarse-Fine Coding**

inputs, this would be an ideal receiver for the joint output of coarse-coding. A cartoon of how this might go is shown in Figure 5, which shows a fragment of an abstract multiplication table for integers between 1 and 9. Only one inhibitory link is shown. Units representing possible answers (e.g., 12) have separate conjunctive receiving sites and take on the activation of the maximum (or logical OR) of the sites. There are two reasons why this general idea provides a major saving over punctate encodings. One reason is that one computational role of units has been brought down to sites, significantly increasing the feasible numbers (by perhaps 1000). The other is that not all combinations of outputs need to be explicitly represented. Although it is well beyond the current story, one can envision how the coarse-fine coding units could combine to represent the feature space with different resolutions.

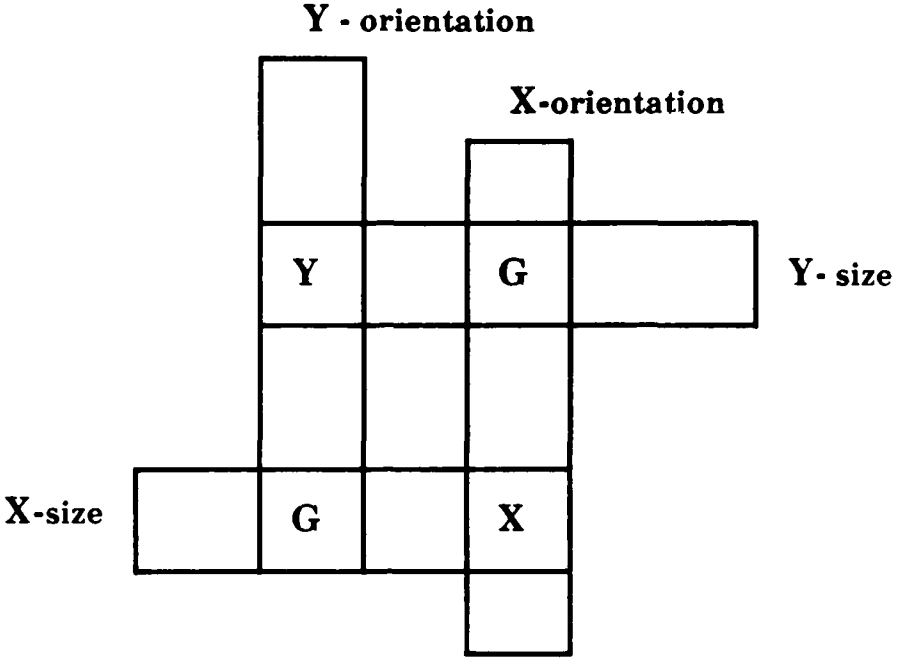


Figure 4: Ghosts in the Machine

It is interesting to consider if we could eliminate the circular nodes carrying the punctate number representations and deal directly with the distributed representation based on number range and odd-even parity. The obvious solution is to link directly the appropriate combinations to the receiving sites. This works fairly well, but does encounter some problems. Suppose 1 and 4 were simultaneously active. Our distributed representation of 2 is <3 and even, both of which are active in this case and this could lead to activation of the answer 8. This problem can be fixed, e.g. by having separate sites for the first and second arguments, but it does point out the delicacy of doing computation with distributed representations.

One final discussion will close off this path of consideration. The coarse-fine coding examples used overlapping encodings, but were based on the minimum possible number of cells to cover some feature space. Suppose instead we allowed for redundancy in the coverage, say three separate tesselations of the space. In terms of Figure 3, a second covering could be similar bars at 45° and 315° to the axes [Ballard 1986b]. We could still use separate receiving sites for each desired stimulus, but the computation would be not just a logical AND. In fact, a thresholded sum of activity might be quite plausible as the sites' way of computing the likelihood of its combination being present. This would combine the error tolerance and information reduction ideas in a simple and plausible way. Edelman has come to essentially the

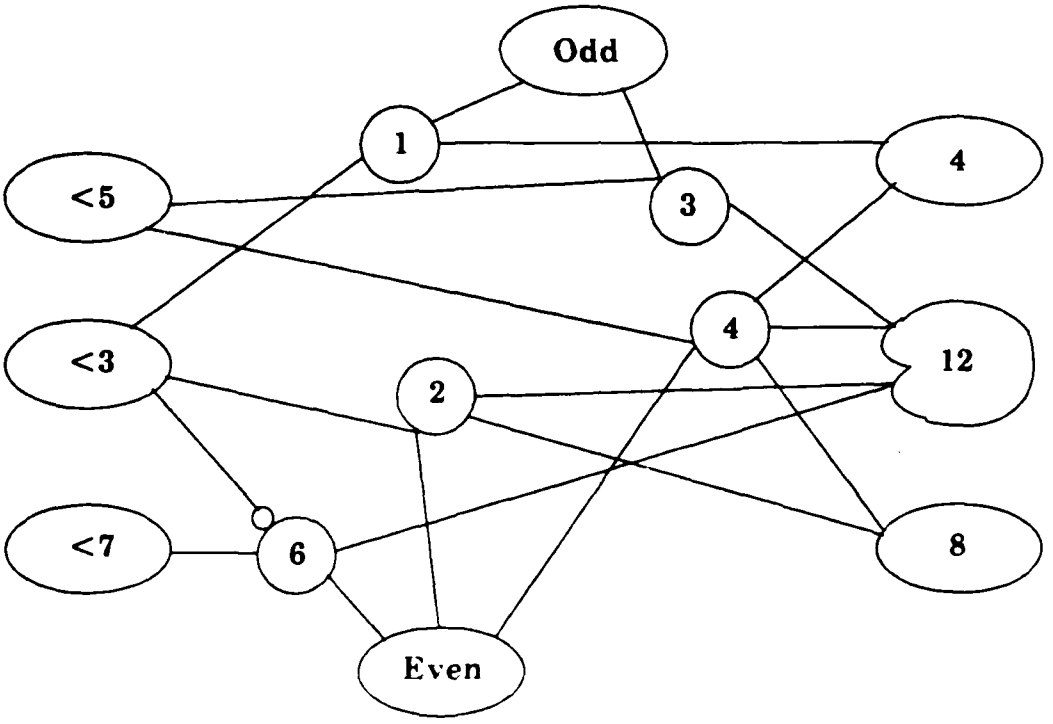


Figure 5: Multiplication Table for Integers

same kind of model through a very different route, starting from his expertise in immunology [Edelman 198?]. Since this is how I believe the brain really works, I will next attempt to show how the holographic approach narrows down (sic) to the same solution.

2. Holographic Models

Holographic models have been fervently supported by biologists, psychologists and theoreticians. There is no comparable fervor for compact models. In fact, several researchers preach fully distributed models while employing punctate ones, often in the same paper. To cite one well known example, the elegant travelling salesman model of Hopfield and Tank [1986] is purely punctate and unmappable to Hopfield's holographic memory proposals [1982]. One major contributing factor in the popularity of this view was the early work of Karl Lashley who found that for a variety of tasks, the deficit exhibited by lesioned rats was best explained by the total amount of cortex removed -- the "law of mass action." Lashley summarized his view of memory representation in the classic 1950 paper "In Search of the Engram." The following quotation continues to motivate much current work [Lashley 1950]:

> It is not possible to demonstrate the isolated localization of a memory trace anywhere within the nervous system. Limited regions may be essential for learning or retention of a particular activity, but within such regions the parts are functionally equivalent. The engram is represented throughout the area. . . . Briefly, the characteristics of the nervous network are such that when it is subject to any pattern of excitation, it may develop a pattern of activity, reduplicated throughout an entire functional area, by spread of excitations, juch as the surface of a liquid develops an interference pattern of spreading waves when it is disturbed at several points. . . . Consideration of the numerical relations of sensory and other cells in the brain makes it certain, I believe, that all of the cells of the brain must be in almost constant activity, either firing or actively inhibited. There is no great excess of cells which can be reserved as the seat of special memories. The complexity of the fucntions involved in reproductive memory implies that every instance of recall requires the activity of literally millions of neurons. The same neurons which retain the memory traces of one experience must also participate in countless other activities.
>
> Recall involves the syngergic action or some sort of resonance among a very large number of neurons. . . . From the numerical relations involved, I believe that even the reservation of individual synapses for special associative reaction is impossible.

Since Lashley's time, the intricate specialized structure of mammalian cortex has been elucidated and no one currently holds the view that all of cortex is holographic. (Reading Lashley's article makes it clear that he would reach very different conclusions on current experimental evidence.) Since all of the primary sensory and motor areas have been found to have specialized structure, the holographic hypothesis is currently restricted to "higher" brain areas whose functional organization is not yet understood. Another historical source of motivation for diffuse models was the wonderful 1949 book, *The Organization of Behavior*, by Donald Hebb. Hebb introduced the notion of cell assemblies, but was (appropriately for the time) vague about how they actually encode knowledge. Hebb definitely envisioned a dynamic pattern of activity, but there are two interpretations of this. The most literal would be to assume concepts are purely dynamic and are not tied to any particular tissue; this idea has been pursued a bit [von der Marlsberg 1985; Bienenstock 1985] but without much success and the experimental data is not encouraging. A more general notion of dynamic activity of cell assemblies is inherent in all current connectionist theories; the compact - diffuse question is about what fraction of a system is involved. A closely related question is structure; it does not seem useful to think of a computer as a "chip assembly." We will return to this after examining theoretical hologram models.

Holographic models are theoretically attractive because of two properties: fault-tolerance and generalization. The brain clearly has these properties and it is easy to see informally why holographic models do also. If all of the units of a large system are involved in coding one concept, the failure of some of them can be tolerated. Furthermore, two concepts which share much of their activity patterns will tend to

behave similarly. More technically, the theoretical hologram models have all been based on one form or another of mathematical correlation. An excellent compendium of holographic-style models can be found in Hinton and Anderson [1981], and I will often cite this rather than the original papers.

The purest holographic model is, unsurprisingly, the optical hologram itself. A nice presentation can be found in [Willshaw 1981] where it is also shown that that the purely linear hologram is undesirable on both computational and biological gounds. The models typically studied are large rectangular matrices representing all the possible connections between m input units and n output units, which need not be distinct (cf. Figure 6). The most common case, and one of the easiest to consider, is where each unit compares the sum of its weighted inputs to its threshold and emits 1 if the sum of inputs is greater and 0 otherwise. A concept is represented as a binary vector over all the input lines. Each element of the vector may be uninterpretted or can be thought of as the presence or absence of a microfeature characterizing the concept. (A punctate view of this would be to imagine each output unit as being responsible for computing one bit of the output, using all of the inputs according to its weights on them.) If the input and output are identified, the matrix becomes a pattern completion machine. The basic idea is simple and derives from the mathematical notion of correlation. The correlation of two binary vectors u and v is simply:

$$\sum_{j=1}^{n} u_j v_j$$

and this is obviously maximized when u = v. In a well distributed set of vectors, an input distorted by modest noise or omission will correlate best with the appropriate complete vector. This is the basic source of error resistance and generalization in holographic-style models. What makes them interesting is that a connection matrix with the appropriate weights can be learned with a simple, local procedure. Most of the learning discussion is in Section 3, but this case is so central that it will be done here.

For the auto-associative, pattern completion case we want weights w_{ij} such that

$$X_j = \sum w_{ij} X_i \Big|_0^1$$

assuming the threshold is included as another weight. The natural local weight-change rule is to increase the weight between two units when they both are active (output = 1). This is essentially Hebb's rule and can be interpretted as increasing the weight of units whose firing is correlated. Symbolically:

$$\Delta w_{ij} = (X_i \cdot X_j) \cdot \delta \qquad \text{where } \delta \text{ is a constant.}$$

If all the weights are initially set to zero and the auto-association matrix is trained on some P sequences then

$$w_{ij} = \delta \cdot \sum_{k=1}^{P} X_i^{(k)} \cdot X_j^{(k)} .$$

The resulting matrix can be shown to be optimal in some cases (orthonormal vectors) and is a good estimator in many others [Kohonen 1984]. Thus a simple and

biologically plausible local learning rule yields a matrix which reliably finds the best match to an input pattern under some distortion conditions (but not others). The binary, linear auto-associator is the simplest of wide variety of essentially similar models. Many variations have been tried on all aspects of the model [Hinton & Anderson 1981; Kohonen 1984; Barto & Anandan 1984]. The more refined models use a variant of Hebb's rule which allows feedback from some external result to affect the weight change process [Kohonen 1984; Barto *et al.* 1981; Anderson & Murphy 1986]. The shortcomings of the methodology do not lie in the learning rules, they demonstrate that essentially anything that can be represented as a linear threshold matrix can be learned by correlation -- the problem is that the representation itself is much too weak. There is a well-known relation between correlation matrices and classical pattern recognition and the same basic inability to deal with structure, occlusion and invariance applies in both cases.

There are a number of computational considerations which rule out any existing holographic encoding and make it extremely unlikely that one will be discovered. Before examining these in detail, we must eliminate a trivial path of escape sometimes sought by defenders of the holographic models. It is a truism of computation theory that any universal mechanism can simulate any other if computational costs are ignored. Since a large enough (associative or other) memory can be made universal, one can find a way to encode any computation as a holographic memory plus a small interpreter. But ignoring computational costs is precisely what we can not do in the current enterprise -- computational costs are one of the principle constraints on the viability of neural representation schemes. Much of the holographic work has been directed to purely passive memory networks, and we will examine these, but it is important not to forget that no passive memory will satisfy the basic time and competence constraints for conceptual knowledge. A sequential machine with a fast associative memory is still orders of magnitude off the required performance.

When computational issues are taken seriously, the holographic model is fatally deficient, even as a passive associative memory. The basic problems with any holographic representational scheme are cross-talk, communication, invariance and the inability to capture structure. Essentially the same problems have prevented the development of holographic computer memories or recognition systems despite considerable effort. Consider the problem of representing a concept like "grandmother" as a pattern of activity in all the units in some memory network. We can assume that different parts of the network represent various input modalities (e.g., vision) and they all lead to the same overall pattern. But notice what would happen if two (or more) concepts were presented at the same time, e.g., grandmother at the White House. The encodings for the two concepts would normally overlap and the system would get garbled. This is a massive instance of the cross-talk problem of Figure 4. Of course, we can reduce the probability of cross-talk by having fewer units active for each pattern. Suppose that there are punctate output detectors for each pattern in the holographic memory (nothing becomes easier if there are not). Willshaw [1981] addressed the question of how one could arrange the coding and detector thresholds so that only the desired detector would respond to each pattern. If one assumes that the cross-talk is randomly distributed (the best case), the system will be reliable only if the number of units active for each pattern is proportional to the logarithm of the number of units in the diffuse memory. This means that a network of 1,000,000 units should use an encoding with about 20 active units per concept -- this is suspiciously close to the number that would arise from a redundant, coarse-coded compact approach.

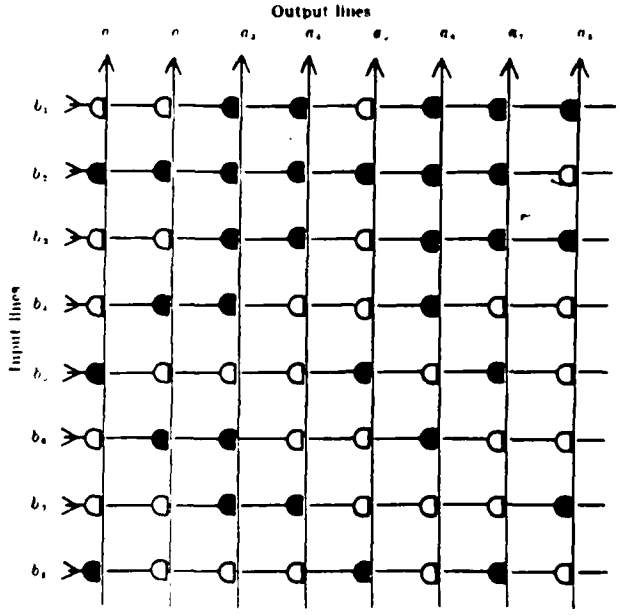


Figure 6: An Associative Net. The nodes that have been activated in the storage process are colored black (from [Willshaw 1981]).

If a fully distributed encoding is used for concepts, there is an unsolved problem of how conceptual information could move from one subsystem, e.g., vision, to another, such as speech. The obvious encoding would be with a "bus" or group of links as wide as the holographic representation, which is unrealistic. In fact, there are many fewer long distance than local connectionions in the cortex. If only some of the units were linked, they would constitute a more compact representation of the concept, which violates the holographic assumption. Even so, unless these encodings were largely disjoint, communication over the bus would have to be sequential to avoid cross-talk. If cross-talk were present, people would make mistakes like seeing a horse and a chair and mistakenly saying something irrelevant like "apple sauce." If microfeatures were the basis of the representation, then something like "rocking horse" should result. The idea that visual information is conveyed to other parts of the brain one concept at a time clearly violates timing constraints by orders of magnitude. The notion that "attention" can restrict the system to one concept at a time is not tenable. Consider how a cowboy movie is mapped from vision to concepts. What is the concept for each frame: the posse, the horse and rider and saddle and hat, etc.? How do they fit together to keep a gestalt of the scene and scenario? What is happening with all the other visual information? How does attention know what to transmit unless the structure of the scene and story line are already known at the vision end? And, of course, we would be back at a sequential machine model. There

appears to be no alternative to assuming that, at least for communication, representation of concepts must be largely disjoint and thus compact.

The same communication problem would arise within the concept memory itself if one tried to build a knowledge structure like Figure 1 with a diffuse representation. If a concept like "salty" were represented only by a large pattern, the links for this entire pattern would have to go to all the places that related to saltiness -- and be treated correctly at each of these. If a concept encoded by N units needed to be linked to M other concepts, a total of M*N links would be needed. One punctate intermediate unit can reduce the requirement to N+M connections. The more distributed the representation, the more serious this problem becomes. Again, any serious reduction in this wiring requirement would constitute a compact representation. And, as in the inter-modal case, unless these representations were largely disjoint, concept processing woud have to be sequential to avoid cross-talk. This eliminates the spreading activation and massively parallel processing which was the motivation for the whole idea.

Even so, no one has suggested how to represent any but the simplest concepts in the holographic style. For example, the problem of all the different historical views of one's grandmother is as difficult for the hologram as for the punctate model. In fact, it is far from obvious how to make the same distributed pattern active for alternative views of a chair, even without occlusion. One could, I suppose, assume that there are essential invariant features of one's grandmother that are derived (magically) from all the different pictures, stories etc. involving her. Even so, how would one express the structure of concepts, like the fact that grandma has two legs, the left of which is slightly shorter. All concepts in any holographic structure that I know of are totally without internal structure. There is an idea of associating the components of a representation with microfeatures (and this will be discussed later) but these are still unstructured. Nor do any of the holographic proposals provide a way of answering even simple questions like the color of grandmother's hair (let alone at different ages). There are a number of other problems with holographic models, but this should suffice to show that this end of the compact-diffuse spectrum is no more viable than the punctate end.

The biological evidence against anything vaguely like a holographic model is equally compelling. This is only fair, since the holographic hypothesis denies any relevance to neuroanatomy and physiology. Since intricate specificity and detailed visuotopic, tonotopic, etc. maps have been discovered everywhere in the brain that has been examined, the only hope for the hologram is for some higher association areas. Even there, the anatomical structure has been found to be much like the sensory areas [Goldman - Rakic & Schwartz 1982] and nothing like the connection of each input to each output required by matrix models. Simple counting arguments show that at most 10,000 neurons could be in a matrix-like network and the local connectivity precludes this possibility as well. Neurosurgery relies upon precision stimulation and recording to localize lesion sites. There may be a point to studying fully distributed encodings, but direct mapping to the brain is not among them. Even Kohonen [1981, p. 132] who is a major figure in holographic-style modelling says:

> There are good experimentally verified reasons to assume that for a particular sensory experience or other occurrence, the pattern of activity over the complete memory field consists of only a few activated local areas . . .

Another argument for highly distributed representations derives from the large number of input fibres ($\sim 10^4$) to cortical neurons. If all of these fibers participate actively then, ipso facto, the representation is diffuse. There are three reasons why this argument fails. While there has been no definitive study on the number of pre-synaptic events required for neural firing, estimates gleaned from papers and conversations run from one event to a few dozen. No one has suggested that several thousand synapses must fire at once for an action potential. Also, we can see from Figures 5, 7 that many of the connections could represent alternative ways of activating the same concept (e.g. from different points in space). Another way of looking at this is that the thousand-fold connectivity is capturing an "OR" of activation conditions rather than an "AND." Finally, as we show in Section 3, learning in a connectionist system requires the potential for many more connections than are ever made functional. The most striking physiological demonstration of this principle is in the neural reorganization studies of Merzenich *et al.* [1984].

All of this may seem to be flogging a dead horse model, but purely holographic theories continue to be seriously proposed. There has been a very recent flurry of interest in spin-glasses as holographic memory models in the theoretical physics community [Hopfield 1982; Toulouse 1985]. Any physical system will, in isolation, reach some stable state and each of these states could be looked upon as encoding a different concept; this is obviously a massively parallel system. The spin-glass story is beyond the scope of this article, but the key idea is that spin-glasses are idealized materials which can take on many stable states in ways that are mathematically interesting. The hope is that analysis of these will provide insights into the behavior of the brain and/or the design of parallel hardware. This might well happen, but the current spin-glass models share the inherent deficiencies of all holographic representations and have no direct applicability to the neural representation of concepts.

There are two basic ways of adding the required structure to hologram-like models. One can either add structure to the collection of units (cf. Figure 7) or one can try to construct structures out of components, each of which is an unstructured cell asessembly. Both approaches lead to systems quite like the redundant, coarse-coded approach. Consider the latter idea: Nothing as concrete as Figure 1 has been attempted, but one envisions about 100 (Palm) units in an assembly jointly encoding some unspecified number of concepts. But we know that the cross-talk and coarse-coding constraints limit the range of possible number of concept nodes to be a few hundred. Of course, the units in the assembly can not represent the detailed structure of the concept because they are shared by unrelated concepts.The structure is represented by connections (also never specified) between cell assemblies which presumably play a role equivalent to the evidence links in Figure 1. This move preserves the sovereignty of the cell assembly, but only in a titular fashion -- the work is all done by the connections among assemblies.

The alternative move, adding structure to the assembled masses, comes in many variants and is treated in the next section.

3. The Middle Ground and Beyond

> It is difficult to interpret such findings, but I think that they point to the conclusions that the associative connections or memory traces of the conditioned reflex do not extend across the cortex as well-defined arcs or paths. Such arcs are either diffused through all parts of the cortex, pass by relay through lower centres, or do not exist. [Lashley 1950, p. 461]

> It is clear that various forms and aspects of learning and memory involve particular systems, networks and circuits in the brain and it now appears possible to identify these circuits, localize the sites of memory storage and analyze the cellular and molecular mechanisms of memory. [Thompson 1986].

We have seen that neither the purely punctate nor the fully holographic model are at all plausible as theories of how the brain represents conceptual knowledge. A variety of arguments have all converged on the idea that concepts are represented by overlapping activity among a modest number (3-100) of units and that the structure within and across these groups can not be uniform or arbitrary. This is all fairly close to the compact end of the original spectrum, but the story is more complex than that. There are important lessons to be learned from the work on highly distributed models and many difficult problems which have eluded all theories. To simplify the discussion, we will adopt the standard convention and treat all compact representations as punctate. The first section showed both that the purely punctate theory can not be biologically valid and also that the redundant coarse-coded variant is plausible and computationally very close to punctate version. It is simply much easier to understand punctate systems and all current models, however distributed, use punctate encodings of some elements for clarity. It is understood that direct neurobiological applications will have to take the detailed encoding more seriously.

From my perspective, there is only one conceptual difference between the current compact and diffuse models of neural knowledge representation and this difference can be seen as one of research strategy. Consider the punctate model of the appearance of horses shown in Figure 7. This is taken from a paper of mine that attempts to show how vision yields such hierarchical conceptual descriptions [Feldman 1985]. At the lowest level are feature pairs, which we assume are derived by early vision networks. (The fact that there are feature pairs rather than individual features or n-tuples of them arises from technical considerations not relevant here (cf. [Feldman 1985, McClelland & Rumelhart 1986]). The remaining structure is organized as a hierarchy (cf. Table 1 and dogma 2) where activation propagates towards the top pontifical cell which is active when a horse is recognized. Even in this unrealistically punctate version, the recognition of a horse is a "pattern of activation" in many of these units. Since the connections are two-way, mentioning the name of a horse will cause (some) activation in all the nodes comprising visual and other descriptions of horses. Looking at this another way, we see that it makes no sense to talk about the activation of a single concept in such structure -- activation automatically spreads to encompass a sub-network. Notice also that this structure will recognize a horse even when some features are missing or garbled if the other features plus context are sufficiently strong. This captures the error

tolerance and some of the generalization ability that were most attractive about correlation matrices. Computer experiments of moderate complexity along these lines have been successfully carried out [Sabbah 1985; Shastri 1985]. The current version of holographic correlation theory uses microfeatures (and feature pairs) essentially identical to the bottom row of Figure 7. The difference is that a correlation model would not have the hierarchical structure but would represent horse exclusively by the pattern of activity of the feature-pair units. But consider how one might expand the correlation model to represent horse heads, eyes, etc., as well. No one has done this, but it is hard to imagine any solution that did not separate out groups of features and combine them to form higher groups. Now, we have seen that combining diffuse encodings is a problem whose only known solution is to focus the representation. So why are all these smart people so excited about highly distributed representations?

There are two important problems that have been effectively studied using diffuse encodings -- generalization and learning. Some insight into the situation can be gained by examining the reading model of Figure 8, taken from McClelland and Rumelhart [1981]. This is one of the earliest and most successful applications of connectionist modelling to psychology. The model is basically punctate, incorporating units for specific shapes, letters and words in a highly structured excitation - inhibition network. The model was used to successfully model a wide range of experimental data, including the word superiority effect previously considered paradoxical. Subjects can more reliably recognize a letter, e.g., "A" in the context of a word, e.g., "FAST", than in isolation, and the model shows clearly how feedback from the word level can aid decisions at the letter level. But the model did more than this, and therein lies our tale. It turns out that the model (and the subjects) can also recognize letters better in the context of pronounceable non-words than in isolation. There is no explicit representation of these non-words, but the units for the similarly spelled words combine to provide enough boost to the target letter. Anderson and Hinton [1981, p. 28] describe this in the following way:

> Thus, the pronounceable non-words are represented by distributed patterns of activity at the word level.

There is much to be learned from an examination of this statement. It is, of course, literally not true. There are many aspects of the non-words (their length, pronounciation, etc.) which are not represented at all and one would not expect someone to recognize repeated occurrences of a non-word from this model. What is true is that *for this task*, the network behaves as if it had a representation of the pronounceable non-words. In fact, it is not just pronounceable nonwords that show the enhancement effect. The concepts with the distributed representation would have to be something like "collections of common pieces of words." It seems to me to be much better to understand this "emergence" as a general property of evidential networks, than to postulate extra concepts having a distributed representation. In general, the research that has been described as studying diffuse representations can be understood as really concerned with system properties of neural networks. Structured networks of punctate elements (like an electronic circuit) have "emergent properties" as anyone who has tried to understand or repair one can testify.

What actually happens in the demonstrations of distributed representations is the following. A problem is picked in which certain properties (e.g., surface syntactic categories) are the ones required for a particular class of general answers (e.g., case roles). The distributed representation is chosen to be the required properties and

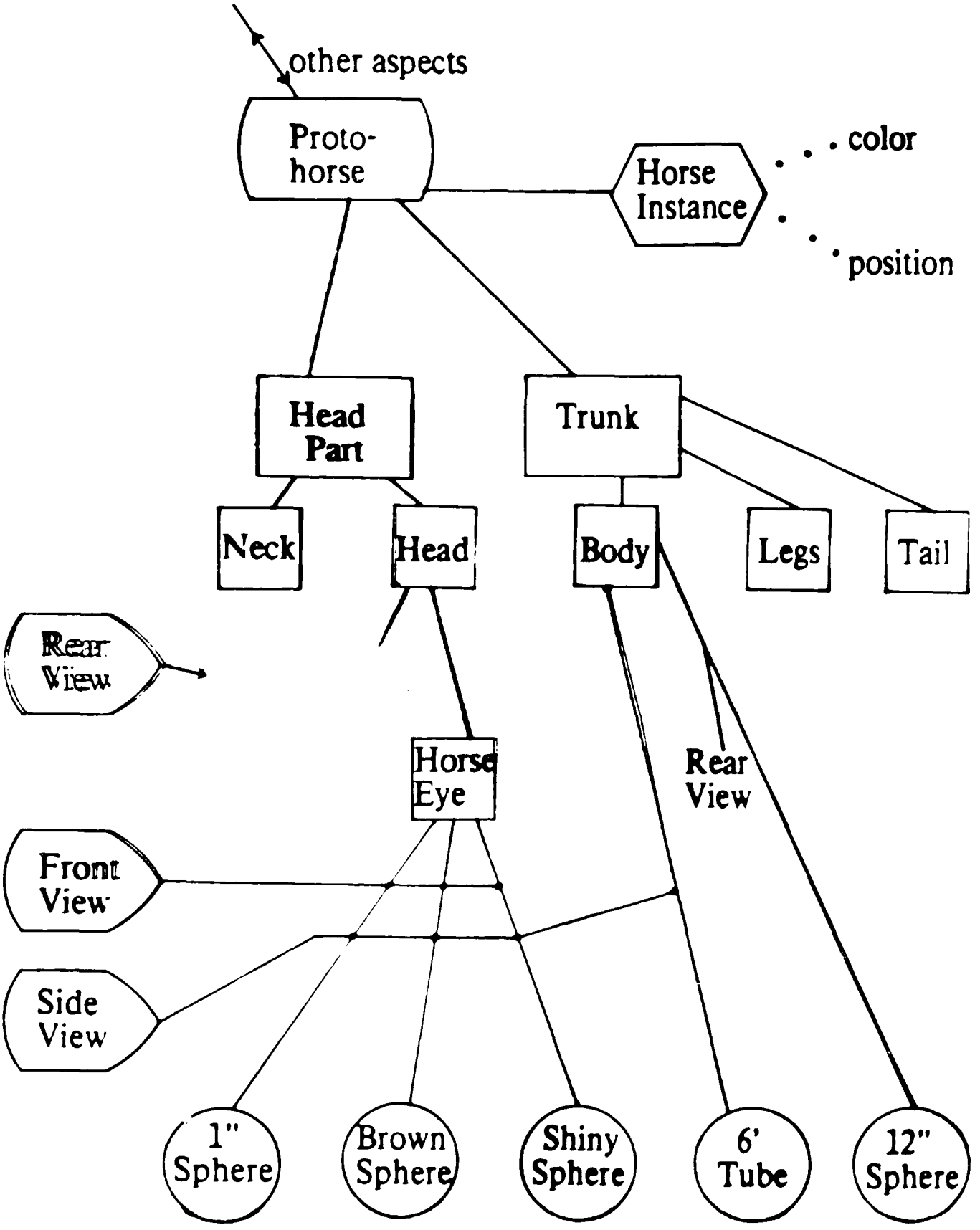


Figure 7: Punctate Model of a Horse

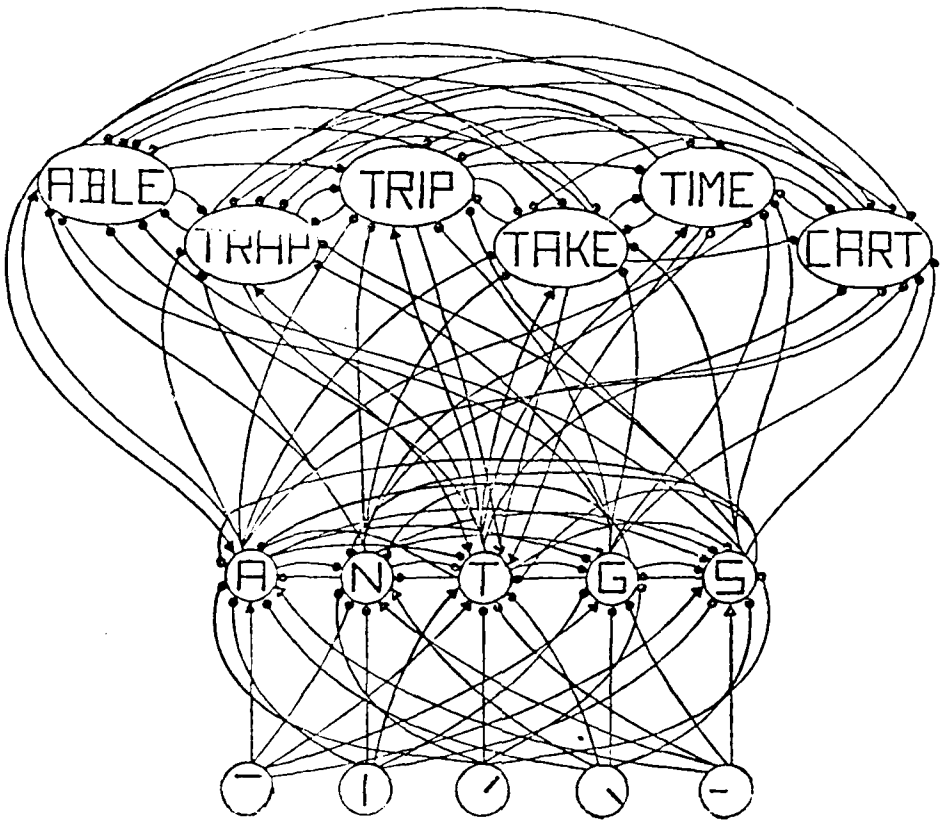


Figure 8: Reading Model after [McClelland &Rumelhart 1981]. A few of the neighbors of the node for the letter "T" in the first position in a word, and their interconnections.

thus the system shows (or learns) the appropriate generalization. Since the combination rule is always linear threshold, care must be taken with the choice of primitives and some precoding may be necessary. Other generalizations over the same domain (e.g., past tense formation) are based on different properties and are thus studied in separate experiments using entirely different distributed representations. A great deal of elegant and important work has been done in this style, but the basic relation between properties and generalizations seems to have been misunderstood.

The generalizations based on property vectors do capture relations that would not arise naturally from hierarchical representations such as Figure 7. In the horse example one might want to capture the generalization that animals with hair are mammals. We could link horse to mammal, but this is of no direct use in classifying novel hairy animals. The point is that if all uses of a concept were forced to go through some fixed taxonomy, the system would be unable to yield many crucial generalizations. There are systematic relationships among properties that are best

captured directly. The relations between syntactic and case roles, between pronunciation and morphology and between hairiness and mammalhood are systematic and most compactly encoded by circumventing particular instances of words or mammals. There is an interesting duality here. If one knew enough hairy mammals, it is likely that activation of "hairy" would lead (indirectly) to the activation of "mammal." One could view this as all the individual mammals constituting a distributed representation of the hairy-mammal relation, but this is a bad way to view network properties which are in general quite complex evidential relations [Shastri 1985].

Learning

All of the preceding discussion has explored the ways in which conceptual knowledge might be represented in neural networks, ignoring the critical problem of learning. The question of the relative role of learning in intelligent behavior is as old and as basic as they come. It is now clear that humans come with an enormous amount of pre-wired structure, develop a great deal more in environmentally driven ways, and (except for administrators) continue to learn throughout life. The problem this presents for connectionist learning studies is a severe methodological one.

The basic difficulty is that we know human learning is based on an elaborate existing structure, but we know very little about the exact nature of this structure, particularly in the area of conceptual knowledge. Any study of learning either assumes some existing structure (at the risk of trivializing the learning aspects) or assumes no existing structure and is restricted to quite simple problems. The correlation matrix models of memory have presented an attractive research vehicle. The total initial connectivity and linear threshold rule are a minimal *a priori* structure, and the correlation method of updating is an easily analyzed learning rule. The results of such studies often take on a structure-from-chaos aspect which many people find attractive.

Obviously enough, these correlation matrices employ a diffuse encoding and the appeal of the two ideas is strongly connected. But the ideas of correlation, feedback and weight change are not restricted to diffuse representations, linear threshold rules or complete connectivity matrices. In fact, complete connectivity is not biologically plausible and the other assumptions also have problems. A great deal of recent work [Belew 1986; Rumelhart & McClelland 1986; Parker 1985] involves using correlation type ideas on networks with much richer structure, while preserving domain independence.

It has been known for some time that any effective learning rule would have to include an input from the ultimate result of the computation; Hebb's correlation rule has no way to punish a connection that leads to a disaster. In a network that directly links inputs and outputs, like a matrix model or a 1-layer perceptron, it is easy to punish the offending links. For a system with more structure and indirect links, it has not been obvious how to assess who deserves the credit and who deserves the blame. Recent developments in learning theory, while not totally solving this problem, have enabled the study of learning in much richer structures [Rumelhart 1986].

The basic idea is to "back-propagate" error signals from the output (which is directly corrected) to successively lower levels of input. A simple version would be to have the weight change at layer n, $\Delta w_{ij}^{(n)}$, given by

$$\Delta w_{ji}^{(n)} = \varepsilon \, \delta_i^{(n+1)} \, X_j^{(n)}$$

where $\delta_i^{(n+1)}$ is the error attributed to unit i in layer $n+1$, $X_j^{(n)}$ is the output of units in layer n (now continuous valued) and ε is a learning constant. One can view this as changing w_{ji} in proportion to how the output of unit j *correlates* with the error attributed to unit i. The back-propagation algorithm has been used in a variety of studies and has been quite successful [McClelland & Rumelhart 1986]. There are still problems with its range of application, convergence rate and biological plausibility, but it does provide the best way known to study learning in structured networks. For all the reasons discussed above, this is a critical problem for our enterprise.

Learning in compact connectionist systems involves some additional considerations. A major problem in this formulation is "recruiting" the compact representation for new concepts [Wickelgren 1979; Feldman 1982; Shastri 1985; Fanty 1986]. It is all very well to show the advantages of representational schemes like Figures 1, 7 and 8, but how could they come about? This question is far from settled, but there are some encouraging preliminary results. The central question is how a system that grows essentially no new connections could find (recruit) compact groups of units to capture new concepts and relations. One relevant result concerns the probability of finding compact groups that link nodes in a random graph [Feldman 1981]. It turns out that, for biologically reasonable parameters, the probability of a compact cluster ($\sim$ 20 units) is quite high. (Incidentally the probability of finding a single unit is essentially zero, providing another argument against punctate theories.) The brain is not random, except perhaps very locally, but that actually is good news for recruiting concept representations. Current work on learning in more structured networks [Fanty 1986; Hinton forthcoming] is examining this more closely.

Structure and Relations

Questions of structure have played a major role throughout the paper. The structure of the nervous system, of domains of knowledge, and of concepts themselves have been central concerns. Despite the initial disclaimer, some consideration of relations has also appeared. Structure, in my opinion, is the dominant issue in neural knowledge representation. Most of the shortcomings of holographic style models can be traced to their unstructured nature. This has been recognized for some time and people have made attempts to correct the problem. One idea is to identify the components of the giant vector with microfeatures of concepts. The cleverest choice is to make the microfeatures follow a conceptual hierarchy so that more specific concepts (e.g. horse) share features with more general ones (mammal, animal) [Hinton; Anderson]. This implicitly captures one kind of structural relation among concepts, but leaves all the others untouched. And if properties are all microfeatures of the representation how could one incorporate new information like the fact that early horses were a few inches tall or that a particular horse had broken a leg. This move, like the generalization demos, can be seen as another way of exploiting relationships that exist among properties (cf. [Shastri 1985]) and has nothing to do with holography.

When is comes to more general relations, there are two ways to try. One, which was mentioned earlier, assumes that relations among concepts are represented by

links (axons) connecting the representations of the concepts involved. This inherently requires compact representations for the concepts, although many authors who talk about the idea in general terms would be repelled by the thought. I believe that resource considerations make this the only viable option, but no one has figured out how to make it work except in simple cases. One thing we do know is that hierarchical compact representations like Figure 7 can capture internal structural relations defining concepts and that this goes far towards solving the questions of how to treat all the different views of a horse or your grandmother. Briefly, the conceptual knowledge about, e.g. horses, is a largely shared structure deriving much of its content from its relationship to other concepts. This is a kind of distributed representation, but not at all what is usually meant by the term. There is also strong evidence that we do not retain all the detailed memories of our grandmothers, but recreate them with a significant tendency towards regularization [Neisser 1982].

If one is committed to a diffuse, unstructured representation for concepts there is no direct way to realize relationships among them. Ignoring computational costs, one can design a holographic memory to store relational information in symbolic form e.g. as triples like (Brother Billy Jimmy) [Hinton 1981; Kohonen 1980]. But this is a move of desperation in neural modelling, since all of the problems of sequential symbolic computing re-enter the scene. A sequential machine with a fast relation store will show none of the performance or context sensitivity that motivates connectionist models.

In summary, connectionist studies of pure learning minimize the pre-existing structure and tend to study diffuse models. This has turned out to be very valuable and will continue to be. Like the studies of overall convergence, e.g., [Wilson & Cowan 1972; Hopfield 1982; Cohen & Grossberg 1983], these are best done assuming no particular structure of the network. But the structure is there and only compact representations can capture it.

Let us return to the analogous contrast between atomic physics and thermodynamics. Atomic physics (really chemistry) is concerned with precise structures and their interactions. Because of the complexity involved, cell-biology would be an even better analog for the detailed structural concerns of compact models. Modern thermodynamics (statistical mechanics) derives its power from abstracting, from enormous systems of units, a very small number of state variables that characterize certain questions of interest. While the abstraction is only strictly true for systems without structure, the results of the theory have much wider application. Similarly, bulk models of neural activity are likely to continue to yield valuable insights into neural functions, but it is fatal to ignore the detailed structure present. The challenge is to get global results into a form where they can be used to advantage in working out how conceptual knowledge is represented and exploited by the brain.

Structured networks of evidence-combining simple units have a number of attractive computational properties. Error resistance, context dependence, and the ability to assimilate conflicting information are natural properties of such systems. A reasonable degree of generalization follows from these properties. Simple weight-change rules can enable these systems to improve their performance significantly. The collective behavior of such systems can produce powerful computations not easily anticipated, as is true of any complex circuit. While certain concepts are represented explicitly, the system is not restricted to dealing only with those. A properly structured network can behave, in certain situations, as if it had concepts and relations implicitly represented. An explicit, compact handle on concepts is

required when they have disjunction or internal structure or when they participate in relations or are communicated among sub-systems. In addition, the nature of any emergent system properties depends heavily on which concepts are explicitly represented and the detailed structure of the representation.

This suggests the following research priorities for the question of neural representation of knowledge. The detailed anatomy and physiology of nervous systems remains a top priority. The computational properties of connectionist models must be better understood both in specific circuits and as mass systems. Plasticity and learning, of course, remain in their central place, but the central problem is change of existing structure. The most difficult problems, however, appear to lie at the higher conceptual levels. All of the concepts and relations treated in this paper are extremely simplistic. The technical level of research on concept and knowledge representation among linguists, philosophers and symbolic AI types is enormously more sophisticated (e.g. [Wilensky 1986]). If it is true that our brains do these things directly with neural nets, connectionist formulations should yield better characterizations of higher-level thought than symbolic logic, perhaps through the mediation of new scientific languages. Expectations that this will happen without detailed consideration of the structure of the tasks and of the underlying hardware should be based on a time-frame of evolutionary scale.

Acknowledgements

Many of the ideas in this paper arose in discussions with Geoff Hinton over the years and were crystallized in working with Mark Fanty on his dissertation proposal. Also providing valuable feedback on even earlier drafts were Dana Ballard, Gary Dell, Horst Greilich, Brian Madden and John Maunsell.

4. References

Abeles, M., *Local Cortical Circuits*, Springer-Verlag, Berlin, 1982.

Allman, J., F. Miezen and E. McGuiness, "Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons," *Annual Review of Neuroscience*, 8, 407-430, 1985.

Amari, S. and M.A. Arbib, *Competition and Cooperation in Neural Networks*, Vol. 45 of Lecture Notes in Biomathematics, S. Levin (ed.), Springer-Verlag, 1982.

Andersen, R.A., "Value, variable, and coarse coding by posterior periatal neurons," *The Behavioral and Brain Sciences*, Open Peer Commentary, 9, 1, 90-91, March 1986.

Anderson, J.A., "Cognitive and psychological computation with neural models," *IEEE Transactions: Systems, Man and Cybernetics*, 13, 799-815, 1983.

Anderson, J.A. and G.E. Hinton, "Models of information processing in the brain," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale NJ, 1981.

Anderson, J.A. and M.C. Mozer, "Categorization and selective neurons," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale NJ, 1981.

Anderson, J.A. and G.L. Murphy, "Comments on concepts," *Proceedings*, NSF Connectionism Workshop, February 1986.

Ballard, D.H., "Cortical connections and parallel processing: Structure and function," *The Behavioral and Brain Sciences*, 9, 1, 67-120, March 1986a.

Ballard, D.H., "Interpolation coding: A representation for numbers in neural models," TR175, Dept. of Computer Science, University of Rochester, May 1986b.

Barber, C., *Evoked Potentials*, University Park Press, Baltimore, 1980.

Barlow, H.B., "Single units and sensation: A neuron doctrine for perceptual psychology?" *Perception*, 1, 371-392, 1972.

Barnden, J.A., "On short term information-processing in connectionist theories," *Cognition and Brian Theory*, 7, 1, 25-59, 1984

Barto, A.G., R.S. Sutton, and P.S. Brouwer, "Associative search network: A reinforcement learning associative memory," *Biological Cybernetics*, 40, 201-211, 1981.

Barto, A.G. and P. Anandan, "Pattern recognizing stochastic learning automata," Tech. Report 84-30, Computer and Information Science, Univ. Massachusetts at Amherst, December 1984.

Belew, R.K., "Learning to use symbols in connectionist network," submitted to AAAI-86, March 1986.

Beurle, R.L., "Properties of a mass of cells capable of regenerating pulses," *Philosophical Transactions of the Royal Society, Series B*, **240**, 55-94, 1956.

Bienenstock, E., "Dynamics of central nervous system," *Proceedings*, Workshop on Dynamics of Macrosystems, Laxenburg, Austria, J.P. Aubin and K. Sigmund (eds.), Springer-Verlag, 1985.

Charniak, E. and D. McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley Publishing Company, Reading, MA., 1985.

Cohen, M.A. and S. Grossberg, "Absolute stability of global pattern formation and parallel memory storage by competitive neural networks," *IEEE Transactions: Systems, Man and Cybernetics*, **13**, 815-825, 1983.

Derthick, M. and D.C. Plaut, "Is distributed connectionism compatible with the physical symbol system hypothesis?" submitted for publication, April 1986.

Desimone, R., S.J. Schein, J. Moran and L.G. Ungerleider, "Contour, color, and shape analysis beyond the striate cortex," *Vision Research*, **25**, 441-452, 1985.

Edelman, G.M., "Group selection as the basis for higher brain function," in *The Organization of Cerebral Cortex*, F.O. Schmitt, F.G. Worden, G. Adelman and S.G. Dennis (eds.), MIT Press, Cambridge, MA, 1981.

Edelman, G.M. and V.B. Mountcastle, *The Mindful Brain: Cortical Organization and the Group-Selective Theory of Higher Brain Function*, MIT Press, Cambridge, MA, 1978.

Fanty, M.A., "Concept learning in connectionist networks," internal publication, Computer Science Dept., Univ. Rochester, May 1986.

Feldman, J.A., "Four frames suffice: A provisional model of vision and space," *The Behavioral and Brains Sciences*, **8**, 265-289, 1985a.

Feldman, J.A., "Dynamic connections in neural networks," *Biological Cybernetics*, **46**, 27-39, 1982.

Feldman, J.A., "A connectionist model of visual memory," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale NJ, 1981.

Feldman, J.A. and D.H. Ballard, "Connectionist models and their properties," *Cognitive Science*, **6**, 205-254, 1982.

Feldman, J.A. and L. Shastri, "Evidential inference in activation networks," *Proceedings*, Cognitive Science Conference, Boulder, Colorado, 1984.

Gardner, J.C., R.M. Douglas and M.S. Cynader, "A time-based stereoscopic depth mechanism in the visual cortex," *Brain Research*, **328**, 154-157, 1985.

Geman, S., "Notes on a self-organizing machine," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Associates, Inc., Publishers, Hillsdale NJ, 1981.

Gluck, M.A. and J.E. Corter, "Information and category utility," preliminary draft, 1986.

Goldman-Rakic, P.S. and M.L. Schwartz, "Interdigitation of contralateral and ipsilateral columnar projectsions to frontal association cortex in primates," *Science*, **216**, 755-757, 1982.

Gross, C.G., C.E. Rocha-Miranda and D.B. Bender, "Visual properties of neurons in inferotemporal cortex of the macaque," *Journal of Neurophysiology*, **35**, 96-111, 1972.

Hebb, D.O., *The Organization of Behavior*, John Wiley, New York, 1949.

Hillis, W.D., *The Connection Machine*, MIT Press, Cambridge, 1985.

Hinton, G.E., "Distributed representations," CMU-CS-84-157, Dept. of Computer Science, Carnegie Mellon University, October 1984.

Hinton, G.E., "Implementing semantic networks in parallel hardware," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Assoc., Hillsdale, NJ, 1981a.

Hinton, G.E., "Shape representation in parallel systems," in *Proceedings of the 7th International Joint Congress on Artificial Intelligence, B.C.*, 1088-1096, 1981b.

Hinton, G.E. and J.A. Anderson, *Parallel Models of Associative Memory*, Lawrence Erlbaum Assoc., Hillsdale, NJ, 1981.

Hopfield, J.J., "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America* **79**, 2554-2558, 1982.

Hopfield, J.J. and D.W. Tank, "Neural computation in optimization problems," *Biological Cybernetics*, 1985.

Hubel, D.H. and T.N. Wiesel, "Brain mechanisms of vision," *Scientific American*, **241**, 3, 150-162, 1979.

Just, M.A. and P.A. Carpenter, "Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability," *Psychological Review*, **92**, 137-72, 1985.

Knapp, A.G. and J.A. Anderson, "Theory of categorization based on distributed memory storage," *Journal of Experimental Psychology: Learning, Memory and Cognition*, **10**, 616-637, 1984.

Kohonen, T., *Content-Addressable Memories*, Springer-Verlag, Berlin, 1980.

Kohonen, T., *Associative Memory: A System-Theoretical Approach*, Springer-Verlag, Berlin, 1977.

Kohonen, T., E. Oja and P. Lehtio, "Storage and processing of information in distributed associative memory systems," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Associates, Inc., Hillsdale NJ, 1981.

Lashley, K., "In search of the engram," in Symposia of the Society for Experimental Biology, No. 4, *Physiological Mechanisms in Animal Behavior*, Academic Press, New York, 454-483, 1950.

Lashley, K.S., *Brain Mechanisms and Ingtelligence*, University of Chicago Press, Chicago, 1929.

Lynch, G., The Neurobiology of Learning and Memory, MIT Press, Cambridge, MA, 1986.

McClelland, J.L. and D.E. Rumelhart (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol 2: Applications.* Bradford Books/ MIT Press, Cambridge, Mass., 1986.

McClelland, J.L. and D.E. Rumelhart, "An interactive activation model of the effect of context on language learning (Part I)," *Psychological Review*, **88**, 375-401, 1981.

Merzenich, M.M., R.J. Nelson, M.P. Stryker, M.S. Cynader, A. Schoppmann and J.M. Zook, "Somatosensory cortical map changes following digit amputation in adult monkeys," *The Journal of Comparative Neurology*, **224**, 591-605, 1984.

Minsky, M., "K-Lines: A theory of memory," *Cognitive Science*, 117-133, 1980.

Minsky, M. and S. Papert, *Perceptrons*, Second edition, MIT Press, Cambridge, MA, 1972.

Mumford, D., "Two tests for the value unit model: Multicell recordings and pointers," *The Behavioral and Brain Sciences*, Open Peer Commentary, **9**, 1, 102-103, March 1986.

Neisser, U. (ed.), *Memory Observed*, W. H. Freeman Publishers, San Francisco, 1982.

Newell, A., "Intellectual issues in the history of artificial intelligence," in *The Study of Information: Interdisciplinary Messages*, F. Machlup and U. Mansfield (eds.), John Wiley & Sons, Inc., 1983.

Palm, G., "What are the units of neural representation?" submitted for publication, 1986.

Palm, G., "Associative networks and their information storage capacity," *Cognitive Systems*, 1, 2, 107-118, June 1985.

Palm, G., "On associative memory," *Biological Cybernetics*, **36**, 19-31, 1980.

Palm, G. and T. Bonhoeffer, "Parallel processing for associative and neuronal networks," *Biological Cybernetics*, **51**, 201-204, 1984.

Parker, D.B., "Learning-Logic," TR-47, Center for Computational Research in Economics and Management Science, MIT, April 1985.

Phillips, C.G., S. Zeki and H.B. Barlow, "Localization of function in the cerebral cortex," *Brain*, **107**, 328-361, 1984.

Pribram, K.H., M. Nuwer and R. Baron, "The holographic hypothesis of memory structure in brain function and perception," in *Contemporary Developments in Mathematical Psychology*, R.C. Atkinson, D.H. Krantz, R.C. Luce and P. Suppes (eds.), W.H. Freeman, San Francisco, 1974.

Roediger, H.L. III, "Memory metaphors in cognitive psychology," *Memory and Cognition*, **8**(1), 231-246, 1980.

Rumelhart, D.E. and J.L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol 1: Foundations.* Bradford Books/ MIT Press, Cambridge, Mass., 1986.

Rumelhart, D.E. and J.L. McClelland, "An interactive activation model of the effect of context in language learning (Part 2)," *Psychological Review*, **89**, 60-94, 1981.

Sabbah, D., "Computing with connections in visual recognition of Origami objects," *Cognitive Science*, **9**, 25-50, 1985.

Schank, R.C., G.C. Collins and L.E. Hunter, "Transcending inductive category formation in learning," *The Behavioral and Brian Sciences*, to appear, 1986.

Sejnowski, T.J., "Open questions about computation in cerebral cortex," to appear, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Applications*, J.L. McClelland and D.E. Rumelhart (eds.), Bradford Books/ MIT Press, Cambridge, Mass., 1986.

Shastri, L. "Evidential reasoning in semantic networks: A formal theory and its parallel implementation," Ph.D. thesis and TR166, Computer Science Dept., Univ. Rochester, September 1985.

Shaw, G.L., D.J. Silverman and J.C. Pearson, "Model of cortical organization embodying a basis for a theory of information processing and memory recall," *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 2364-2368, 1985.

Smith, E.E. and D.L. Medin, *Categories and Concepts.* Harvard University Press, Cambridge, MA, 1981.

Sullins, J., "Value cell encoding strategies," TR165, Computer Science Dept., Univ. Rochester, August 1985.

Sutton, R.S. and A.G. Barto, "Toward a modern theory of adaptive networks: Expectation and prediction," *Psychological Review*, **88**, 2, 135-170, 1981.

Toulouse, G., S. Dehaene and J-P. Changeux, "A spin glass model of learning by selection," preprint, 1985.

van Heerden, P.J., "A new optical method of storing and retrieving information," *Applied Optics*, **2**, 387-392 , 1963.

von der Heydt, R., E. Peterhans and G. Baumgartner, "Illusory contours and cortical neuron responses," *Science*, **224**, 1260-1262, 1984.

von der Malsburg, C., "Nervous structures with dynamical links," *Berichte der Bunsen-Gesellschaft fur Physikalische Chemie*, in press.

Weinberger, N.M. and D.M. Diamond, "Dynamic modulation of the auditory system by associative learning," to appear, *Auditory Function*, Neurosciences Institute Publication, Edelman, Cowan and Gall, (eds.), John Wiley & Sons, 1986.

Wickelgren, W.A., "Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting the amnesic syndrome, and the hippocampal arousal system," *Psychological Review*, **86**, 44-60, 1979.

Wilensky, R., "Some problems and proposals for knowledge representation," Report No. UCB/CSD 86/294, Computer Science Division, Univ. of Calif. at Berkeley, May 1986.

Willshaw, D., "Holography, associative memory, and inductive generalization," in *Parallel Models of Associative Memory*, G.E. Hinton and J.A. Anderson (eds.), Lawrence Erlbaum Assoc., Hillsdale, NJ, 1981.

Wilson, H.R. and J.D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical Journal*, **12**, 1-24, 1972.

APPENDIX

Interim Neuron Doctrine

"The following five brief statements are intended to define which aspect of the brain's activity is important for understanding its main function, to suggest the way that single neurons represent what is going on around us, and to say how this is related to our subjective experience. The statements are dogmatic and incautious because it is important that they should be clear and testable.

First dogma

A description of that activity of a single nerve cell which is transmitted to and influences other nerve cells, and of a nerve cell's response to such influences from other cells, is a complete enough description for functional understanding of the nervous system. There is nothing else 'looking at' or controlling this activity, which must therefore provide a basis for understanding how the brain controls behaviour. **Since significant behaviors involve many individual nerve cells, functional understanding of the nervous system will require scientific languages for characterizing the behavior of networks of neurons.**

Second dogma

Efficient coding of information is a central problem of the sensory system. At progressively higher levels in the sensory pathways information about the physical stimulus is more abstract and is represented by progressively fewer active neurons.

Third dogma

Trigger features of neurons are matched to the redundant features of sensory stimulation in order to achieve greater completeness and economy of representation. This selective responsiveness is determined by the sensory stimulation to which neurons have been exposed, as well as by genetic factors operating during development.

Fourth dogma

Just as physical stimuli directly cause receptors to initiate neural activity, so the active **networks of intermediate and** high-level neurons directly and simply cause the elements of our perception.

Fifth dogma

Frequency coding is the primary basis of neural communication. Sensory neurons respond with high frequency discharge to external stimuli which fit into a narrow range of possibilities; the higher the discharge the more narrow the range of possible causes.

Zeroth dogma

Intelligent behavior and its neural realization are incredibly complex. A functional understanding of this will require organizational principles from the behavioral and computational sciences as well as biology.

END

DTIC

8-86