THE PARALLEL DISTRIBUTED PROCESSING APPROACH TO SEMANTIC COGNITION

James L. McClelland and Timothy T. Rogers‡*

How do we know what properties something has, and which of its properties should be generalized to other objects? How is the knowledge underlying these abilities acquired, and how is it affected by brain disorders? Our approach to these issues is based on the idea that cognitive processes arise from the interactions of neurons through synaptic connections. The knowledge in such interactive and distributed processing systems is stored in the strengths of the connections and is acquired gradually through experience. Degradation of semantic knowledge occurs through degradation of the patterns of neural activity that probe the knowledge stored in the connections. Simulation models based on these ideas capture semantic cognitive processes and their development and disintegration, encompassing domain-specific patterns of generalization in young children, and the restructuring of conceptual knowledge as a function of experience.

SYLLOGISM
A formal structure for deduction in argument, consisting of a major and a minor premise from which a conclusion logically follows.

**Center for the Neural Basis of Cognition and Department of Psychology, Carnegie Mellon University, 4400 Fifth Avenue, Pittsburgh, Pennsylvania 15213-2683, USA.*
‡Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, UK.
e-mails: jlm@cnbc.cmu.edu; tim.rogers@mrc-cbu.cam.ac.uk
doi:10.1038/nrn1076

How do we know that Socrates is mortal? Aristotle suggested that we reason from two propositions, in this case: Socrates is a man; and all men are mortal. This classical SYLLOGISM forms the basis of many theories of how we attribute properties to individuals. First we categorize them, then we consult properties known to apply to members of the category. Another answer — the one that we and a growing community of researchers would give — is that the knowledge that Socrates is mortal is latent in the connections among the neurons in the brain that process semantic information. In this article, we contrast this approach with other proposals, including a hierarchical propositional approach that grows out of the classical tradition. We show how it can address several findings on the acquisition of SEMANTIC KNOWLEDGE in development and its disintegration in dementia. It can also capture a set of phenomena that have motivated the idea that semantic cognition rests on innately specified, intuitive, domain-specific theories. Although challenges remain to be addressed, this approach provides an integrated account of a wide range of phenomena, and provides a promising basis for addressing the remaining issues.

The hierarchical propositional approach

In the early days of computer simulation models, researchers assumed that human semantic cognition was based on the use of categories and propositions. Quillian[1] proposed that if the concepts were organized into a hierarchy progressing from specific to general categories, then propositions true of all members of a superordinate category could be stored only once, at the level of the superordinate category. For example, propositions true of all living things could be stored at the top of the tree (FIG. 1). Other propositions, true of all animals but not of plants, could be stored at the next level down, and so on, with specific facts about an individual concept stored directly with it. To determine whether a proposition were true of a particular concept, one could access the concept, and see whether the proposition was stored there. If not, one could search at successively higher levels until the property was found, or until the top of the hierarchy was reached.

Quillian's proposal was appealing in part for its economy of storage: propositions true of many items could often be specified just once. The proposal also allowed for immediate generalization of what is known

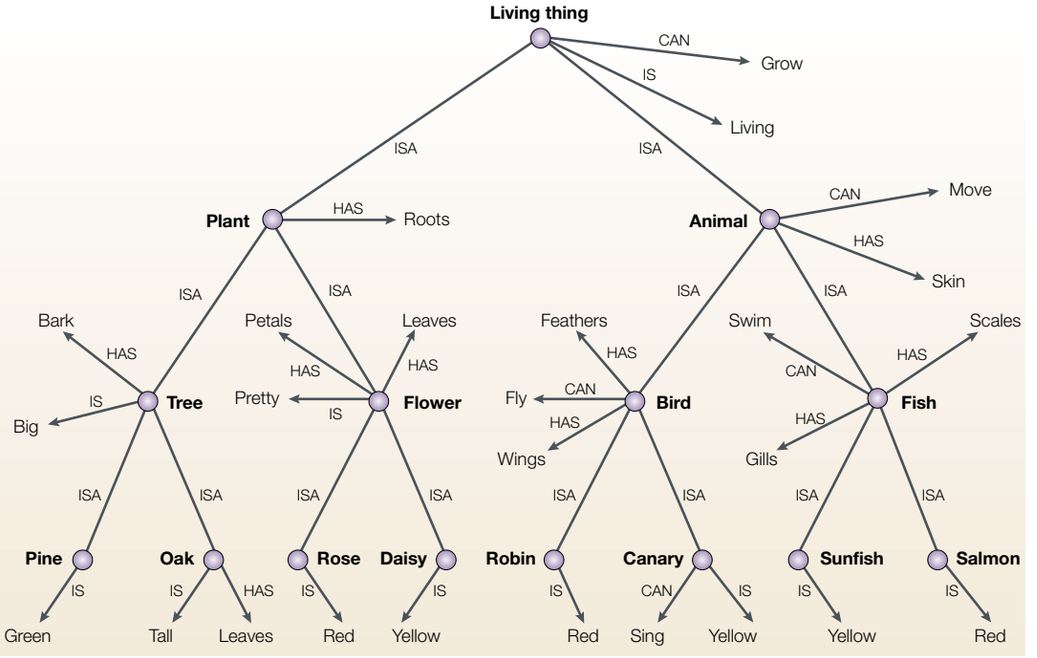


Figure 1 | **The hierarchical propositional model of Quillian[1] applied to the domain of living things, as adapted by Rumelhart[60,61].** Each arrow represents a proposition, with the subject argument at the tail of the arrow, the relation written along the shaft, and the predicate argument at the head. The 'ISA' relations express propositions of the form '*X* ISA *Y*', such as 'robin ISA bird', and are arranged in a hierarchical taxonomy. The relations 'CAN' and 'HAS' specify actions and parts, whereas 'IS' primarily captures superficial appearance properties. In addition to 'robin ISA bird' the network also encodes the proposition 'bird CAN fly': it therefore follows that 'robin CAN fly'.

about stored members of a category to new members, and immediate application of newly learned propositions about the general properties of a concept to all of the more specific subordinate concepts. In a seminal article in 1975, Warrington[2] used Quillian's hierarchical model as the basis for capturing a progressive neurological condition that has come to be called SEMANTIC DEMENTIA[3]. Patients suffer a progressive deterioration in semantic tasks such as naming objects, sorting them into taxonomic categories or verifying their properties[4]. Such patients do not lose all information about a concept at once, but seem first to lose more specific distinguishing information (for example, that a tiger has stripes), with general properties (for example, that a tiger has fur) remaining relatively spared even late in the illness's progression (FIG. 2). Patients also tend to attribute general properties of a superordinate category to individuals that lack them (for example, to add an extra pair of legs to animals such as swans and ducks)[5,6].

Warrington[2] also suggested that young children first acquire very general conceptual distinctions, and then progress to finer and finer ones. There is considerable evidence that is consistent with an overall general-to-specific progression[7–12], although debate surrounds just how general children's first distinctions are and whether there is a PERCEPTUAL-TO-CONCEPTUAL SHIFT[13–15]. Warrington suggested that development proceeds from the top of the TAXONOMIC HIERARCHY and works its way down, whereas disintegration starts at the bottom and works its way up.

In spite of its initial appeal to Warrington and others, Quillian's model is confronted with problems by psychological findings. For example, the model predicts that people will be faster to verify idiosyncratic, specific properties of objects than their shared properties, as specific properties are stored directly with the concept but the general ones are stored further away. Once potential confounds[16] are controlled for, however, no such effect is found[17,18]. And there is something paradoxical about the model; the essential message from development and disintegration is that the general properties of concepts are more strongly bound to an object than its more specific properties, but in Quillian's model the specific properties are stored closest and are therefore most strongly associated with a concept.

A more fundamental problem arises from the reliance on storing knowledge at the superordinate category level rather than with individual concepts. The question arises, just which superordinates should be included, and which properties should be stored with them? At what point in development are they introduced? What are the criteria for creating such categories? And how does one deal with the fact that properties that are shared by many items, which could be treated as members of the same category, are not necessarily shared by all members? For example, many plants have leaves, but not all do — pine trees have needles. If we store 'has leaves' with all plants, then we must somehow ensure that it is negated for those plants that do not have leaves.

SEMANTIC KNOWLEDGE
Knowledge about objects and their properties, and of relationships between and among them, including knowledge of word meanings. General encyclopaedic knowledge is sometimes also included.

SEMANTIC DEMENTIA
A degenerative neuropathological condition that results in the progressive loss of semantic knowledge as revealed through naming, description and non-verbal tests of semantic knowledge, resulting from disease of the anterior and lateral aspects of the temporal lobes.

PERCEPTUAL-TO-CONCEPTUAL SHIFT
A hypothesized developmental transition whereby infants initially categorize objects on the basis of their directly perceived visual properties, but later come to categorize them on the basis of deeper relationships.

TAXONOMIC HIERARCHY
A structured set of concepts linked together with class-inclusion relationships.

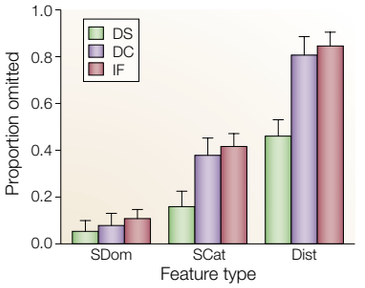
PARALLEL DISTRIBUTED PROCESSING
(PDP). A computational modelling framework in which cognitive and other mental processes arise from the interactions of simple, neuron-like units through their weighted connections. PDP models are a subset of connectionist or artificial neural network models that use distributed representations (a scheme in which the representation of an item is distributed as a pattern of activity across a pool of units also used for the representation of other items) and that treat any act of information processing as involving the simultaneous participation of many units.

a

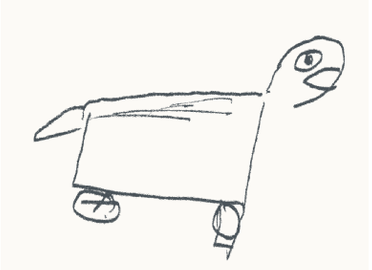
Picture naming responses for JL

Item	Sept. 91	March 92	March 93
Bird	+	+	Animal
Chicken	+	+	Animal
Duck	+	Bird	Dog
Swan	+	Bird	Animal
Eagle	Duck	Bird	Horse
Ostrich	Swan	Bird	Animal
Peacock	Duck	Bird	Vehicle
Penguin	Duck	Bird	Part of animal
Rooster	Chicken	Chicken	Dog

b



c **IF's delayed copy of a camel**



d **DC's delayed copy of a swan**

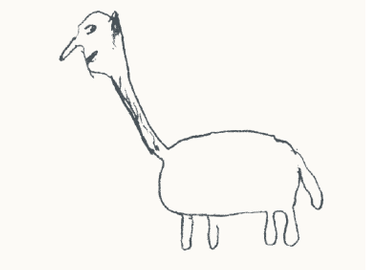


Figure 2 | **Evidence of conceptual disintegration in semantic dementia. a** | Naming responses given by patient JL to pictures of birds (drawn from a set of line drawings for which control subjects consistently provide the name given in the left column[117]) at three times during the progression of his illness. '+' indicates correct responses. **b** | Proportion of features of different types omitted from drawings by three other semantic dementia patients. Patients were shown a picture of the object including all of the tested properties and were asked to copy the picture from memory after a 10-s delay. All patients copied the picture accurately while it remained in view, but had difficulty in reproducing the distinctive but not the domain-general properties of the pictured objects after a delay. SDom, properties shared by typical members of the general domain (for example, eyes, shared by animals) of the test item; SCat, properties shared by typical members of the superordinate category (for example, wings, shared by birds); Dist, distinctive features of the test item itself (for example, stripes, distinctive attribute of tiger). **c** | Delayed copy of a camel; no hump is evident. **d** | Delayed copy of a swan. A long neck is present, indicating some preserved representation of specific information, but there are four legs, illustrating the tendency these patients have to fill in properties that are generally present in items within the overall domain (animals) even if not present in the specific item (swan) or its immediate superordinate (bird). IF and DC, patients with semantic dementia. Part **a** reproduced, with permission, from REF. 4 © (1995) Taylor & Francis. Part **b** reproduced, with permission, from REF. 6 © (1999) Cognitive Neuroscience Society.

If instead we store it only with plants that have leaves, we cannot exploit the generalization.

Graded category membership

A related consideration is that typicality, instead of the number of intervening 'ISA' links in Quillian's hierarchy (FIG. 1), seems to be a better predictor of human performance in category verification and other semantic tasks. For example, subjects verify the statement 'robin is a bird' faster than they verify 'chicken is a bird', and they verify 'chicken is an animal' faster than 'chicken is a bird'[19,20]. These findings are better captured by models (like the one in REF. 19) in which category verification occurs by comparing representations of the item and the category, and responding on the basis of similarity, rather than by models in which verification occurs by searching a hierarchical tree.

These and other limitations of hierarchical propositional models[21] spurred a movement that began in the 1970s to explore alternatives. Several models sprang up in this period that were designed to capture the notion that category membership is graded and depends on patterns of feature values[19] or proximity in a multidimensional representational space[22]. More recently, an explicitly Bayesian approach to categorization has been developed on the basis of principles of optimal inference in the face of a probabilistic relationship between categories and their properties[23,24]. These approaches have all been used to address basic aspects of categorization, and to develop models of induction (if a dog has an omentum, does a duck have one too?)[25]. Another development[26,27] was the suggestion that categories might be represented not by a single summary representation, but by the full set of previously experienced exemplars, with category membership being determined by assessing the summed similarity of a test item to known exemplars. Models based on this idea have been successfully elaborated to address many findings from artificial category learning experiments[28–30]. But in spite of these developments, no successful, integrated theory of categorization and other aspects of semantic cognition has emerged from these efforts[31].

Parallel distributed processing

In this article, our goal is to describe progress towards such a theory within the framework of PARALLEL DISTRIBUTED PROCESSING (PDP)[32]. PDP models share some properties with several of the models mentioned above, but pose an even more radical alternative to hierarchical propositional models. Although the field of semantic cognition is broad and contains many aspects that no extant theory can fully address, there has been considerable progress in the use of PDP models to address several aspects that were not previously encompassed by a single account.

In PDP models, processing takes place by the propagation of activation among simple, neuron-like processing units (BOX 1). Semantic information is not stored as such, but instead is reconstructed in response to probes, in a process called pattern completion. In Hinton's early model[33], two of the consituents of a three-item proposition (such as 'canary ISA —') could be presented, with the task of filling in the third constituent ('bird'). Filling in occurs through the propagation of activation among units through their connections, and the outcome depends on the strengths (or weights) of the connections, which are shaped by experience. When Hinton introduced the model, only primitive algorithms existed for learning the connection weights. Research in the early 1980s produced more powerful learning algorithms[34,35] with the ability to assign useful INTERNAL REPRESENTATIONS to items, including algorithms for recurrent network architectures[36] and biologically plausible implementations[37–39].

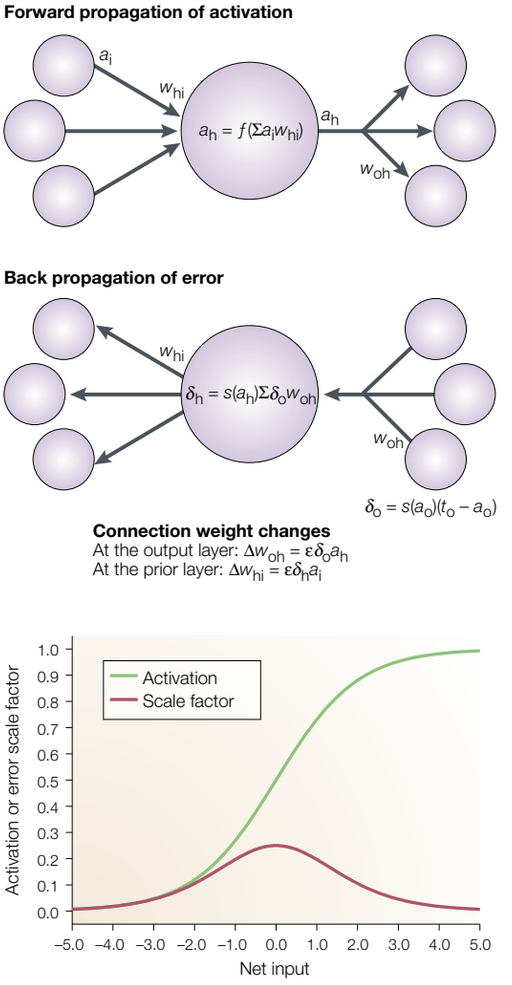
A growing body of work addresses semantic cognition within the PDP framework[6,40–51]. Much of this work has focused on deficits in semantic cognition or on semantic errors that result from brain disorders (for example, reading 'apricot' as 'peach'), reflecting the suitability of PDP models for addressing the graded nature

INTERNAL REPRESENTATION
In a PDP network, a pattern of activity that arises across a layer of hidden units. When a network is presented with a given input, the pattern of activity arising across its hidden layer is the internal representation of that input.

Box 1 | Processing and learning in a feedforward parallel distributed processing network

A FEEDFORWARD NETWORK[35] consists of input units, one or more groups of hidden units, and output units. Input units project through weighted connections to hidden units and/or directly to output units. Hidden units project to hidden units in other groups and/or to output units. There are no connections within groups or return connections to previous groups. A network with only one hidden unit is illustrated. Subscripts i, h and o are used for input, hidden and output units, respectively.

Connection weights are initially small and random. The network is trained with a series of examples that specify activations for input units (a_i) and target activations for output units (t_o). One sweep through the examples constitutes an epoch. For each example, specified values are assigned to the input units, and activation propagates forward (top panel). Each unit calculates its net input — the sum over each incoming connection of the activation of the sending unit multiplied by the weight w on the connection. Activations of the units are set using a smooth monotonic function (green curve in bottom panel), and are propagated forward. Learning depends on the differences between the target and obtained activations of the output units. Adjustments to the weights are made to reduce the error E — the sum of the squares of the differences. To derive the weight changes, one asks, how would an increment to each weight influence E, given the existing weights and activations? This can be broken down into (a) the effect of changing the weight on the input to the receiving unit, and (b) the effect of changing the input to the receiving unit on E. (a) depends on the activation of the sending unit; if the activation is 0, changing the weight does nothing. (b), called δ, depends on (c) the effect of changing the unit's input on its activation, and (d) the effect of changing its activation on E. Term (c) is a scaling factor $s(a)$ that depends on the unit's input (red curve in bottom panel). For an output unit, term (d) depends on the difference between the target and the current activation; if positive, E decreases with a increment in activation; if negative, E decreases with a decrement in activation. Changing the activation of a hidden unit will affect the error at each output unit. The amount depends on the weight from the hidden unit to the output unit, multiplied by the effect of changing its activation on E, which is its δ term. Accordingly, δ for each output unit is scaled by the weight from the hidden unit to the output unit and propagated back to the hidden unit, where these terms are summed and scaled to obtain the hidden unit's δ (middle panel). Once δ has been computed for a hidden unit, it can be propagated back further. The weight adjustments are scaled by a constant ε that must be small to ensure progress[63]. Some adjustments cumulate across examples, whereas others cancel out. Overall the weights adapt slowly, yielding gradual evolution of the patterns of activation and gradual reduction of error.



of such deficits. There have also been many applications to cognitive development[52–59], reflecting the sensitivity of PDP models to structure in experience, and their corresponding ability to capture patterns of change in cognitive abilities during childhood.

This article focuses on a simple example of the general class of PDP models that addresses core issues in semantic cognition, while also encompassing conceptual development and semantic disorders. The model (FIG. 3), which was introduced by Rumelhart[60,61], has a feedforward structure, so that activation flows only in one direction — from units that represent items (such as 'canary') and relations (such as ISA) through intermediate or 'hidden' layers, to an output layer containing units corresponding to possible completions of three-constituent propositions. This simplifies Hinton's model, in which activation could flow in all directions. Such recurrent networks are more fully consistent with our (and Rumelhart's) view of the nature of cognitive processes, and the approach has been used in many PDP models of semantic cognition. We focus on the feedforward case because it shows most simply how the PDP approach can address many aspects of the psychological findings.

Rumelhart addressed the processing and learning of the specific body of information stored in the hierarchical propositional network shown in FIG. 1. The units in one of the two groups on the left (input) side of the network stand for the concepts at the bottom of the hierarchy. The units in the other group on the left stand for the relations. The units on the right (output) side stand for all possible completions of three-term propositions true of the concepts. However, connections are initially set to small random values so that activations produced by a particular input are weak and undifferentiated. The network is trained by presenting it with experiences based on the information contained in Quillian's hierarchy. For example, the hierarchy specifies that a canary can grow, move, fly and sing, so one of the training examples specifies 'canary' and 'CAN' for the input and 'grow', 'move', 'fly' and 'sing' as the target output. For this case, 'canary' and 'CAN' are activated on the input units; activity propagates forward through the HIDDEN UNITS to the output units; and the activations resulting there are

HIDDEN UNITS
Units in a neural network that mediate the propagation of activity between input and output layers. The activations or target values of such units are not specified by the environment, but instead arise from the application of a learning procedure that sets the connection weights into and out of the unit.

FEEDFORWARD NETWORK
A class of neural networks wherein activation propagates only in one direction, from a set of inputs towards a set of output units, possibly through one or more layers of hidden units.

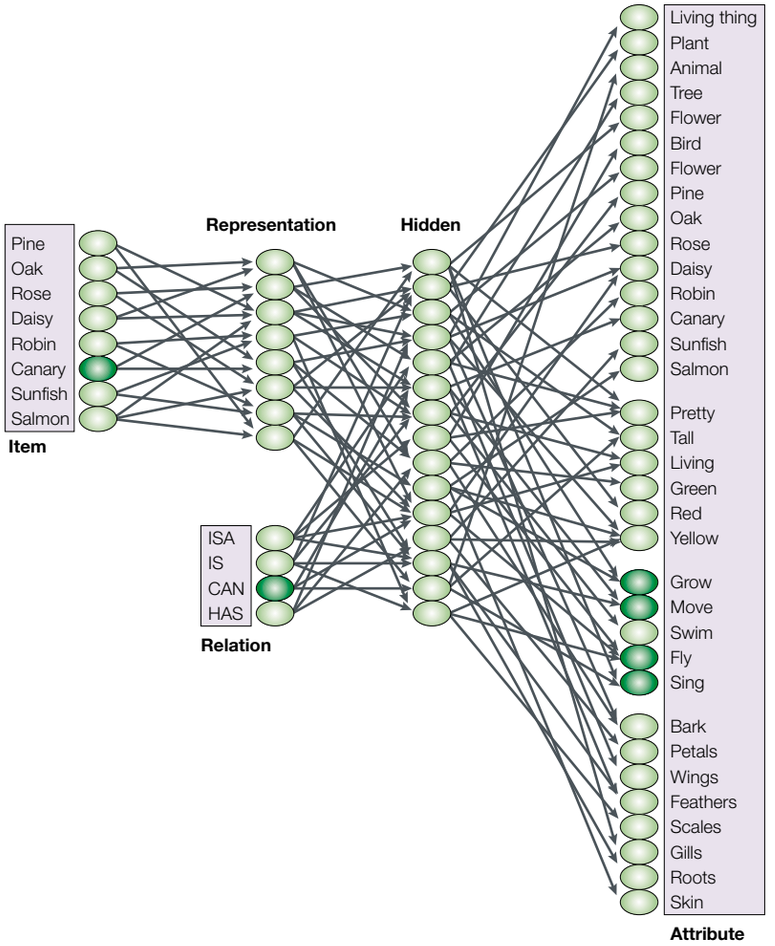


Figure 3 | **Our depiction of the connectionist network used by Rumelhart**[60,61]. The network is used to learn propositions about the concepts shown in FIG. 1. The entire set of units used in the network is shown. Inputs are presented on the left, and activation propagates from left to right. Where connections are indicated, every unit in the pool on the left (sending) side projects to every unit on the right (receiving) side. An input consists of a concept–relation pair; the input 'canary CAN' is represented by darkening the active input units. The network is trained to turn on all those output units that represent correct completions of the input pattern. In this case, the correct units to activate are 'grow', 'move', 'fly' and 'sing'. Subsequent analysis focuses on the concept representation units, the group of eight units to the right of the concept input units. Adapted, with permission, from REF. 61 © (1993) MIT Press.

compared to the correct output (activation of 'grow', 'move', 'fly' and 'sing' should be 1, and activation of other output units should be 0). The connection weights are then adjusted to reduce the difference between the target and the obtained activations. The set of training experiences includes one for each concept–relation pair, with the target specifying all valid completions consistent with FIG. 1.

The network is trained through many epochs or successive sweeps through the set of training examples. Only small adjustments to the connection weights are made after each example is processed, so that learning is very gradual — akin to the process we believe occurs in development, as children experience items and their properties through day-to-day experience. Of course, the tiny training set used is not fully representative of children's experience, and the coding of experience for the network finesses some important issues. However, we argue that the training data capture two essential features. First, many types of naturally occurring things have a hierarchical similarity structure, as Quillian noticed; and second, from exposure to examples of objects children learn just what the similarities are and how they can be exploited.

The Rumelhart model can show how learning can shape not only overt responses, but also internal representations. A special set of internal or hidden units, labelled 'representation' units, was included between the input units for the individual concepts and the large group of hidden units that combine the concept and relation information. When the network is initialized, the patterns of activation on the representation units are weak and random, owing to the random initial connection weights, but gradually these patterns become differentiated, recapitulating the general-to-specific progression seen in many developmental studies. The simulation results in FIG. 4 show that patterns representing the different concepts are similar at the beginning of training, but gradually become differentiated in waves. One wave of differentiation separates plants from animals. The next waves differentiate birds from fish, and trees from flowers. Later waves differentiate the individual objects. The process is continuous, but there are periods of stability punctuated by relatively rapid transitions also seen in many other developmental models[54,56,59], reminiscent of the seemingly stage-like character of many aspects of cognitive development[62].

Rumelhart focused on showing how this network recapitulates Quillian's hierarchical representation of concepts, but in a different way than Quillian envisioned it — in the pattern of similarities and differences among the internal representations of the various concepts, rather than in the form of explicit 'ISA' links. This characteristic of the model is clearly brought out in the hierarchical clustering analysis of the representations of the concepts (FIG. 4b). Rumelhart also showed how the network could generalize what it knows about familiar concepts to new ones. He introduced the network to a new concept, 'sparrow', by adding a new input unit with 0-valued connections to the representation units. He then presented the network with the input–output pair 'sparrow–ISA–bird/animal/living thing'. Only the connection weights from 'sparrow' to the representation units were allowed to change. As a result, 'sparrow' produced a pattern of activation similar to that already used for the robin and the canary. Rumelhart then tested the responses of the network to other questions about the sparrow, by probing with the inputs 'sparrow–CAN', 'sparrow–HAS' and 'sparrow–IS'. In each case the network activated output units corresponding to shared characteristics of the other birds in the training set (CAN grow, CAN move, CAN fly; HAS skin, HAS wings, HAS feathers), and produced very low activation of output units corresponding to attributes not characteristic of any animals. Attributes varying between the birds and attributes possessed by other animals received intermediate degrees of activation. This behaviour is a

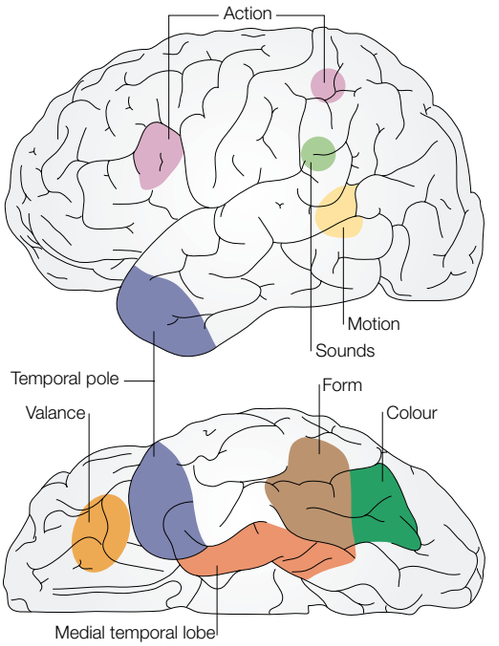
CATASTROPHIC INTERFERENCE
The loss of information previously stored in a PDP network that can occur as a result of later learning. Reducing overlap among representations or ensuring that learning is very gradual and interleaved with ongoing exposure to material already known are two ways of avoiding this problem.

graded version of the pattern of generalization that would be seen in the Quillian hierarchical semantic network if the proposition 'sparrow–ISA–bird' were explicitly added to it.

Our own interest in the Rumelhart model[63] was intially sparked by a dilemma. On the one hand, the progressive differentiation seen in the network captures a corresponding process seen in cognitive development[7–11]. This 'progressive penetration into the nature of things'[62] is a general property of child development and PDP networks and has been expoited in many other models[54–58,64]. So, the models provide an appealing explanation of how experience gradually shapes the way we think as we develop. But the gradual learning seen in such models flies in the face of the fact that children (and adults) can also learn quickly, acquiring new object names and other information in one or a very few exposures[65]. Worse still, any attempt to force the network to acquire new information quickly (by turning up the learning rate or massive repetition) can lead to CATASTROPHIC INTERFERENCE[63,66], in which the process of learning new information results in such large changes in the connection weights that much previous information is destroyed.

To resolve this dilemma, McClelland, McNaughton and O'Reilly[63] introduced the complementary learning systems theory, in which a slow-learning semantic knowledge system is complemented by a second, fast-learning system in the medial temporal lobes. New information is acquired rapidly in connections within this second system. It provides the initial basis for reconstruction of newly formed memories, and is also one source of teaching signals for the semantic system, thought to reside in the neocortex (BOX 2). This and related proposals[67–70] account for the effects of extensive damage to the medial temporal lobes[71,72], which can result in a

Box 2 | **Semantic cognition in the brain**



The models described in the main text provide an abstract theory about the representation and processing of semantic information; how might it be instantiated in the brain? Both the effects of brain damage and functional imaging studies support the idea that semantic processing is widely distributed. One view is that the act of bringing to mind any particular type of semantic information about an object evokes a pattern of neural activity in a part of the brain dedicated to that type of information[103–109]. The brain areas that become activated when thinking about action on an object are near those directly involved in action[105,106], and similarly for the form, movement, colours[105] and sounds[107] of objects. The abstract model can be brought in line with these findings by supposing that the output units for each type of information are located in distinct brain regions. In addition to these units, however, the model calls for representation units that tie together all of an object's properties across different information types. Such units might lie in the temporal pole, which is profoundly affected in semantic dementia[76,77]. Others[103,104] have emphasized the potential role of this region as a repository of addresses or tags for conceptual representations. We suggest that the patterns of activation in these areas are themselves 'semantic' in two respects. First, their similarity relations capture the semantic similarities among concepts, thereby fostering semantic induction. Second, damage or degeneration in these areas disrupts the ability to activate elsewhere the more specific properties of concepts (while still supporting activation of properties shared by semantically similar objects). The complementary, fast learning system is thought to be located in medial temporal lobe[63,72].

There are many applications of parallel distributed processing models to semantic disorders[6,42,45,47,50,51], but as yet no unified account for the full variety of different patterns of semantic deficit[108]. Many patients show deficits that are specific to a particular superordinate category (such as living things) rather than to a particular information type, but others do not. One class of models[50,51,111] indicates that apparent category specificity might reflect differences in the pattern of covariation of features in different categories. Another possibility[41,109,112] is that category specificity arises from lesions affecting neurons that represent the type of information most relevant to the affected category, where this type of information is central to the representation of category members. It has been argued that these approaches cannot account for the full range of category-specific cases[113]. Category-specific organization might emerge over the course of development[114]; this might be part of the developmental process of conceptual differentiation[2]. It is likely that there are forces at work in the brain that tend to cause neighbouring neurons to represent similar things; this might help to minimize the lengths of axons and dendrites needed to connect neurons that communicate with each other[115]. Many neural network models incorporate such forces[47,58,115,116], leading to progressive topographic differentiation in development.

severe inability to acquire new arbitrary factual information and a loss of recently acquired information, but complete sparing of general semantic knowledge and (more controversially[73,74]) relatively preserved memory for remote over recent information.

In the complementary learning systems theory[63,75], the gradual learning characteristics of PDP networks are thought to capture essential properties of the semantic learning system in the neocortex. With this idea in mind, we inquired whether the Rumelhart model could also provide a basis for understanding the progressive loss of semantic knowledge in semantic dementia, which results from the degeneration of the anterior and lateral regions of the temporal neocortex[76,77]. We found that as the model's internal representations are increasingly degraded, it shows a pattern of knowledge disintegration that is strikingly similar to that seen in the patients. In considering how this arises in the model, it is useful to visualize the relative locations of the different concepts in the network's representational space (FIG. 5a). The dimensionality of this space is large (equal to the number of representational units, which in this case was eight), so for visualization we reduce the number of dimensions using a projection into two dimensions that preserves, as well as possible, the relative distances among the representations of the different concepts.

Similar concepts tend to be near each other in this space, and unrelated concepts are far apart, so we can find regions of the space that are associated with concepts at different levels of generality. Each concept (such as 'canary') occupies its own small region; its immediate superordinate occupies a larger region encompassing it and other members of the same superordinate (such as 'bird'); and the more general category (such as 'animal') extends over a larger region, encompassing the birds as well as the fish. With this in mind we can consider what will happen if we degrade the representations of individual concepts.

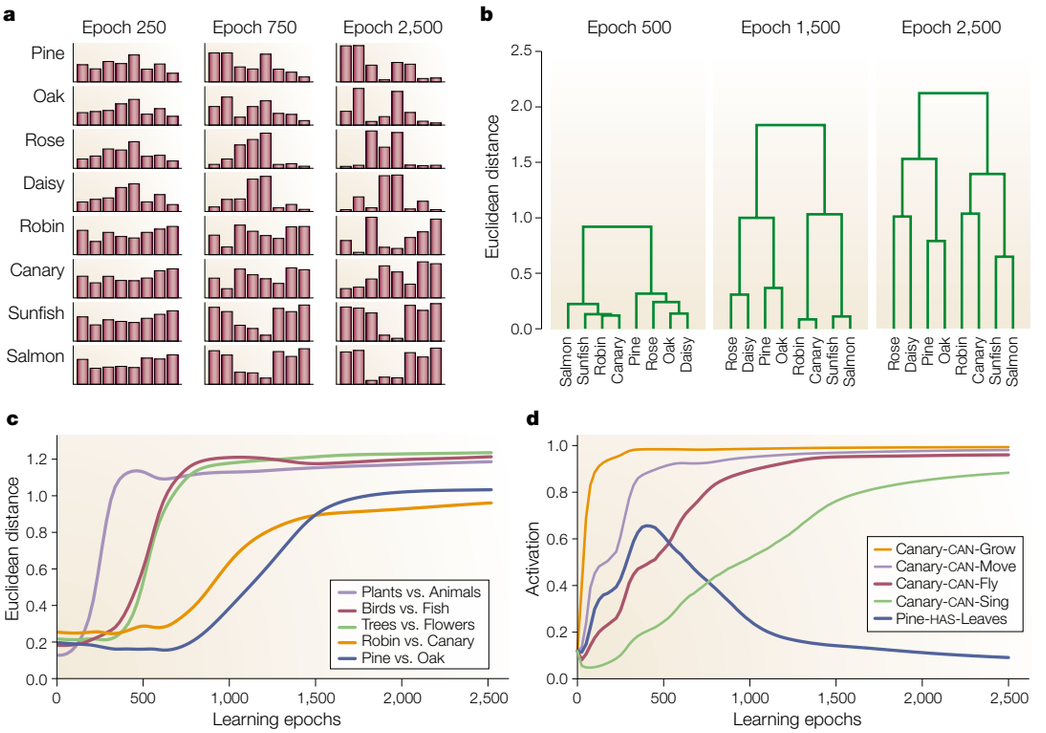


Figure 4 | **The process of differentiation of conceptual representations.** The representations are those seen in the feedforward network model shown in FIG. 3. **a** | Acquired patterns of activation that represent the eight objects in the training set at three points in the learning process (epochs 250, 750 and 2,500). Early in learning, the patterns are undifferentiated; the first difference to appear is between plants and animals. Later, the patterns show clear differentiation at both the superordinate (plant–animal) and intermediate (bird–fish/tree–flower) levels. Finally, the individual concepts are differentiated, but the overall hierarchical organization of the similarity structure remains. **b** | A standard hierarchical clustering analysis program has been used to visualize the similarity structure in the patterns shown in **a**. The algorithm searches the patterns to find the two that are the closest, according to a Euclidean distance measure, creates a node in the tree at a vertical position corresponding to the distance between them, replaces the two patterns with their average, and then iterates until one grand average pattern remains. **c** | Pairwise distances between representations of groups of concepts or individual concepts, illustrating the continuous but stage-like character of progressive differentiation. **d** | The network's performance in activating various properties of the canary, indicating that correct performance is acquired in a general-to-specific manner, and tracks the differentiation of concepts shown in **c**. Also shown is the activation of 'leaves' when the network is probed with 'pine-HAS'. This shows an inverted 'U'-shaped developmental course, capturing the 'illusory correlations' or incorrect attributions of typical properties that have been cited in support of children's use of innately constrained naive domain theories[85,89,118]. (Based on simulations reported in REF. 78.)

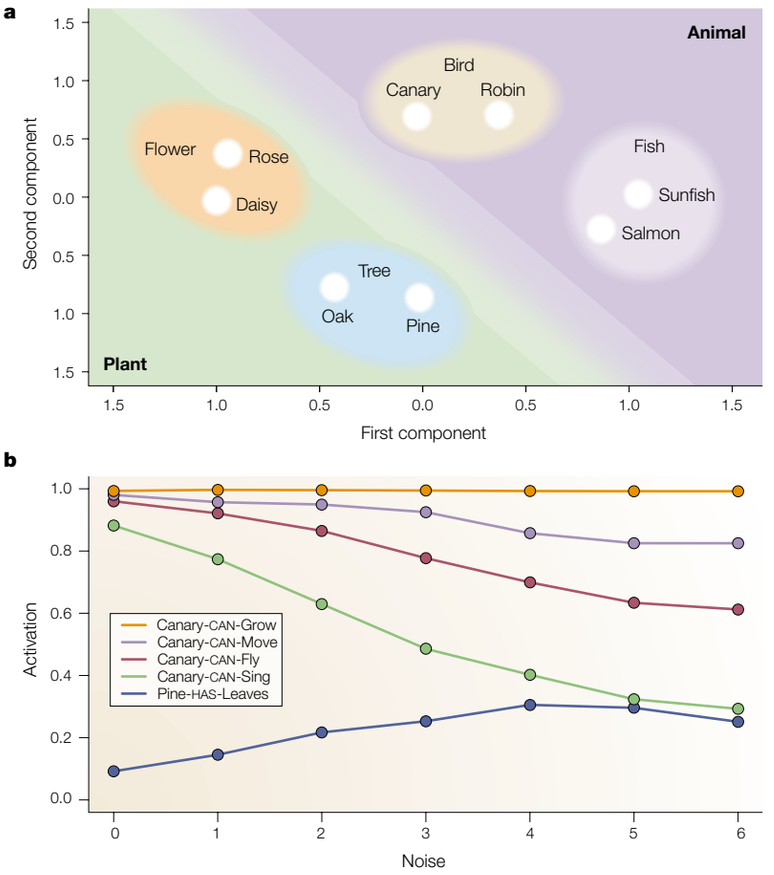


Figure 5 | **Two-dimensional projection of the representations of the various concepts after learning in the Rumelhart[60,61] network, and effects of damage on activations of properties of individual concepts. a** | The relative positions of the eight concepts after learning, with shading used to suggest the regions spanned by various concepts, which have fuzzy boundaries. **b** | The effect of adding random noise of increasing magnitude into the inputs to the representation units. Noise tends to perturb representations in a way that approximates the expected effects of destruction of the neurons involved in the representations of concepts in the brain. In a network with a small number of hidden units like those used here, any perturbation tends to have a specific and idiosyncratic effect not characteristic of what would occur with a larger number of units[119]. Simulation of the effect of neuron loss, therefore, requires averaging over many instances of the same degree of perturbation. Shared properties tend to be preserved whereas idiosyncratic ones tend to be lost, and properties that a concept does not share with other similar concepts (leaves on a pine tree) tend to come back as the representations become less differentiated. (Based on simulations reported in REF. 78.)

One form of degradation, which might resemble the loss of neurons in semantic dementia[6,78], is random perturbation of the representation of each concept, with the amount of perturbation increasing to represent increasing neuronal loss. Increasing degrees of perturbation degrade the network's ability, first to activate specific information about the item (specific name, object-specific properties) and later to activate more general properties, recapitulating the pattern of progressive deterioration of conceptual knowledge seen in semantic dementia (FIG. 5b). In addition, properties that a concept does not have, but which are characteristic of its superordinate domain (for example, leaves on a pine tree), are applied to it when they should not be, just as in semantic dementia.

BASIC LEVEL
The level of a taxonomic hierarchy at which normal participants typically identify a given object. For most concepts, the basic level is at an intermediate level of specificity, such as bird rather than animal or canary. So, when shown a photograph of a canary, people will be more likely to identify it as a bird then as a canary or an animal.

Basic level and item-frequency effects

Although the model shows a general-to-specific process of conceptual differentiation, the names that children[79] and adults[20,80] use to describe objects are generally at an intermediate level of categorization, often called the BASIC LEVEL[81]. For English-speaking city dwellers, the words 'tree', 'bird' and 'dog' tend to be acquired earlier than the superordinate terms 'plant' or 'animal', or than more specific terms such as 'canary', 'pine' or 'poodle'. Why, if concepts are initially differentiated at the superordinate level, does naming first emerge at an intermediate level? An additional curious property of chidren's early naming behaviour is that the names they acquire early (such as 'dog') tend to be over-extended to other items (especially other four-legged animals).

Our simulations[78] indicate that these phenomena might arise from a combination of factors, including the clustering of objects in the world into tight-knit intermediate-level groups within superordinate categories[81], the tendency for parents to use intermediate-level words more frequently than more general or more specific words when speaking to children, and the fact that a few items — such as dogs — are discussed far more frequently in such speech than are related items — such as other land animals, birds or fish[82,83].

We have addressed these issues in an extension of the model described earlier, using a larger number of concepts (21), including four trees, four flowers, four fish, four birds and five four-legged land animals (dog, cat, pig, goat and mouse[78]). In the runs of the simulation considered here, the dog occurred eight times more frequently than any of the other land animals. Also, training experiences in which the network was given exposure to names typically occuring in spoken input to children (tree, flower, bird and fish for the members of these categories, and the names dog, cat, pig, goat and mouse for the land animals) occurred more frequently than training with names at more specific or more general levels. The results of this simulation (FIG. 6) indicated several important points. First, the network still shows progressive differentiation, beginning with the basic distinction between plants and animals, and progressing to the second-level distinction between the different types of animals and plants before the third-level distinction between specific examples. We illustrate this process by tracing the movement over time of a subset of the concepts in a two-dimensional projection of the network's state space (thin lines in FIG. 6a). Second, the network's naming responses at different points (FIG. 6b) indicate earlier mastery of the frequently encountered intermediate names than more general and more specific names. Finally, the network's naming responses reflect the tendency to overextend, for a time, the frequently encountered name 'dog' to other animals, but not to unrelated concepts such as the pine (FIG. 6c). This tendency is eliminated first for birds and fish, and later for other land animals, such as the goat. So, like children, the network differentiates concepts progressively, but names first at an intermediate level, overextending frequent names during an intermediate stage of development.

What are the reasons for these aspects of the network's behaviour? Progressive differentiation arises from the COHERENT COVARIATION of attributes across many different items. All animals share several properties (have skin, eyes, mouths; can see, eat) that differentiate them from all plants, and vice versa. In the model, the connection weights that determine the representation and processing of all the animals tend to be driven in the same direction by their shared properties; the corresponding weights for the plants tend to be driven in the opposite direction. This accounts for the first wave of differentiation that separates animals from plants. Similarly, birds share properties that differentiate them from fish and land animals. The shared properties that distinguish each of these groups from the others drive the second wave of differentiation. The timing of the waves of differentiation are jointly determined by the number of coherently covarying properties and the number of concepts that exemplify them; the superordinate differentiation of plants and animals arises first primarily because all of the concepts fall on one side or the other of this split. The differentiation of individual concepts from each other tends to be late, since individual concepts tend to be differentiated by a few properties that do not coherently covary with other properties. Note that progressive differentiation does not depend on the fact that the network is trained with the names of concepts: the same progression is seen when training is restricted to the 'IS', 'CAN', and 'HAS' properties of the items. This is consistent with the emergence of early signs of conceptual differentiation in children as young as seven months[9].

The tendency to produce intermediate level names before names at other levels arises as a consequence of both differential exposure frequency and the pattern of covariation of properties across concepts; we have shown this in simulations in which exposure frequency and covariation of properties are independently manipulated[78]. The tendency to name dogs correctly before other animals arises from the high frequency of occurrence of dog experiences in the training set. The tendency for 'dog' to be overextended to other animals early in learning reflects the interaction of this frequency effect with the progressive differentiation process. When the network is beginning to learn to activate the name dog when the dog is presented, the representations of the dog and all the other animals are differentiated from the representations of the plants, but the representation of the dog is not well differentiated from the representations of the other animals. So the weights that allow the representation of dog to activate the name 'dog' will produce the same result for the other animals. It is only as the semantic patterns for the animals become differentiated that the tendency to apply the name dog to other animals falls off.

Patients with dementia also tend to overextend the names of very common objects to similar objects. Patient JL[4] reached a point in his deterioration where he correctly named only 3 out of 24 land animals: cat, dog and horse. He used these three names for 20 of the remaining 21 cases. Similarly, when our network is trained with dogs occurring more frequently than other land animals, it over-generalizes the name dog to other animals when its semantic representations are degraded. This tendency reflects the fact (FIG. 6a) that when it occurs frequently, a concept's features (including its name) will tend to be associated with a relatively large region of the semantic space, surounding other related concepts. Degrading the representation of a similar concept will tend to cause it to fall into the space associated with the more common concept.

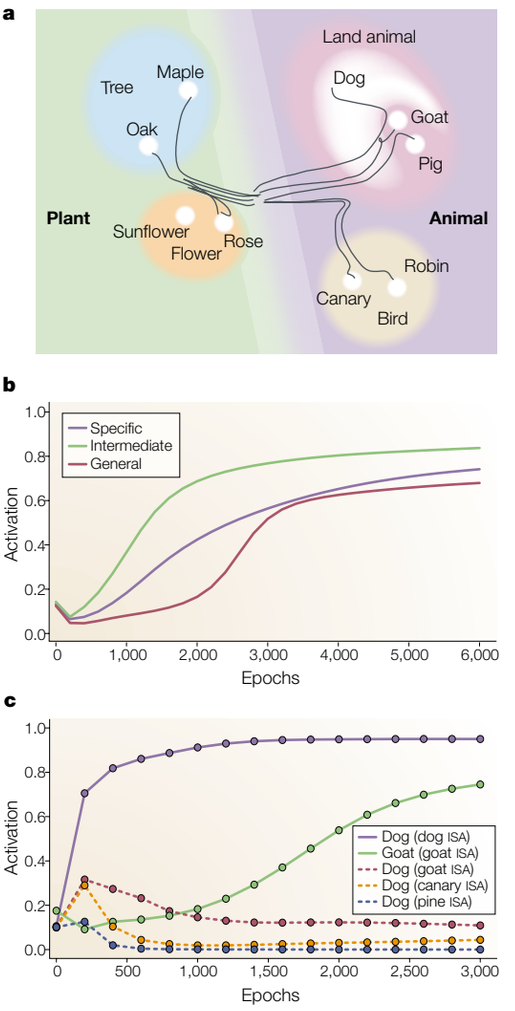


Figure 6 | **Trajectories of learning concepts and their names in a network trained with 21 concepts in five categories (trees, flowers, birds, fish and land animals).** **a** | The end-points of a subset of the concepts, with the trajectories followed during learning overlayed on the diagram. **b** | Initial acquisition of the intermediate level names 'bird', 'fish', 'flower' and 'tree' before names at either more general ('animal', 'plant') or more specific ('robin', 'sunfish', 'rose', 'oak') levels. **c** | The network quickly learns to apply the name 'dog' to the dog input. Early on, it also tends to apply this name to other animals (goat, canary) but not plants (pine). Eventually, correct names for goat (illustrated) and other concepts (not illustrated) are learned. (Based on simulations reported in REF. 78.)

COHERENT COVARIATION
Consistent co-occurrence of a set of properties across different objects. The concept is distinct from simple correlation in that it generally refers to the co-occurrence of more than two properties. For example, having wings, having feathers, having hollow bones and being able to fly all consistently co-occur in birds.

PDP and theory theory

The success of our model in addressing these issues led us to consider whether it could be extended to address additional issues raised by a viewpoint on semantic cognition that is sometimes called 'THEORY THEORY'[84–86]. Proponents of this view suggest that semantic knowledge is built on naive, domain-specific, causal knowledge about objects and their properties — knowledge that is thought to play the part for the child that a theory plays for a scientist[87]. Such knowledge determines what concepts are good ones, which properties are central to a concept, and which properties are only incidental. For example, the category of birds, and the intuitive importance for this category of the ability to fly and having wings, are viewed as flowing from a naive DOMAIN THEORY of the causal structure underlying flight[85]. A naive domain theory can be vague and need not be one that its holder can explicitly articulate[88], but it is thought to guide us to view wings and other properties such as feathers and hollow bones as central (because they enable flight). Some proponents of this and related approaches have suggested that known learning methods are too weak to explain concept acquisition, and that some initial domain knowledge (or constraints leading to that knowledge) must therefore be innate[85,89]. Other proponents (including Carey[84]), pointing to the reorganization of knowledge that occurs through development in some domains, have argued that it is possible to reformulate domain knowledge, but even these theory theorists have had relatively little to say about what drives reoganization.

We consider several classes of findings raised by proponents of theory theory. The first is why some concepts might be better or more coherent than others, and why some properties might be more central to a concept than other properties. For the theory theorist, good concepts are those whose properties are linked by causal relations (having wings and hollow bones enables flight), and the CENTRAL PROPERTIES are those that are so linked (CAN fly, HAS wings, HAS hollow bones). Naive domain theories render these properties easier to learn, and might lead us to impute them to objects that do not seem to have them[85,89]. We suggest instead that causal structure in the physical world (wings and hollow bones do enable flight) leads to coherent covariation among observed properties (many objects have wings and hollow bones, and can fly). By virtue of mechanisms similar to those that operate in the network, we as cognizers come to treat properties that covary coherently as central in importance, and to represent different objects that share these central properties as similar to one another — even though they might differ in other respects.

We have already seen how coherent covariation drives progressive differentiation. Here we discuss its role in determining the relative importance of different features, by contrasting two features of the canary: 'HAS wings' and 'IS yellow'. Even though the network receives direct training on each of these properties equally often, the network learns that the canary has wings far more quickly than it learns that it is yellow. This is not due to greater frequency of learning about having wings than being yellow; in the eight-concept corpus only one other thing has wings (robin) but two other things are yellow (sunfish and daisy). Instead, it is because having wings coherently covaries with other properties that the canary shares with the robin. The connection weight changes that support correct activation of one of these properties tend to support activation of the others, so the learning that occurs for each is mutually beneficial. By contrast, the set of objects that are all yellow share no coherently covarying attributes, so the network must learn about yellowness individually for each concept. In our model, coherent covariation of properties has two other important effects. First, it leads to the overextension of properties to objects that do not have them. For example, 'HAS leaves' covaries with other properties that differentiate plants from animals, and as a result the network tends to over-extend these properties to plants (such as the pine) that do not have leaves, but instead have needles (FIG. 4c). Second, it determines the strength with which a given feature contributes to representational change in a single learning episode. Properties that covary together generate larger weight changes throughout the network, and therefore exert more of an influence on the development of internal representations. The consequence is that the similarities represented by the network are primarily determined by the properties that covary coherently[78]. The model's sensitivity to coherent covariation can therefore explain why some sets of properties are easier to learn and remember than others, why such properties are sometimes incorrectly attributed to members of the domain, and why coherent properties are more central to category membership than other properties.

Second, we consider the fact that people, including young children, generalize properties differently, depending on the type of property and the type of concept to which it is applied. In one relevant study[90], children were shown an alligator puppet called Allie and were told that Allie liked to eat a particular object (thereby suggesting that the object was a food) or that he liked to play with the object (suggesting that it was a kind of toy). The children were then asked to indicate which of two other objects might be the same kind of thing. Children who were induced to treat the objects as food tended to choose another object with the same colour but a different shape, whereas children induced to treat the objects as toys tended to choose another with the same shape but a different colour. This indicates that shape might be more important for defining toys and colour for foods. In another study[91], children saw unfamiliar target objects (such as a triceratops) named at a superordinate level (dinosaur) to which they assigned either a biological property (has cold blood) or a physical property (weighs 1 ton). Children tended to generalize biological properties to objects with the same superordinate category name but a different appearance (they generalized cold blood to a brontosaurus more than to a rhinocerous) but they tended to generalize physical properties to objects with a similar appearance (they generalized weighing 1 ton to a rhinocerous more than to a brontosaurus). Though the effects in these studies

THEORY THEORY
A class of theories that take as their main premise the proposition that human cognition is underpinned by naive domain theories. Under this view, naive theories help the learner to determine which concepts are good ones, and which properties are important for determining conceptual relations among objects; and conceptual development is likened to the process of theory change in science.

DOMAIN THEORY
Knowledge of the causal explanations that are appropriate to a particular kind of object. For example, gravity and momentum are relevant for inanimate objects, whereas beliefs and desires are relevant for human beings (and might be extended to other animals by young children).

CENTRAL PROPERTIES
Properties of an object that are understood to be most important for determining what kind of thing it is.

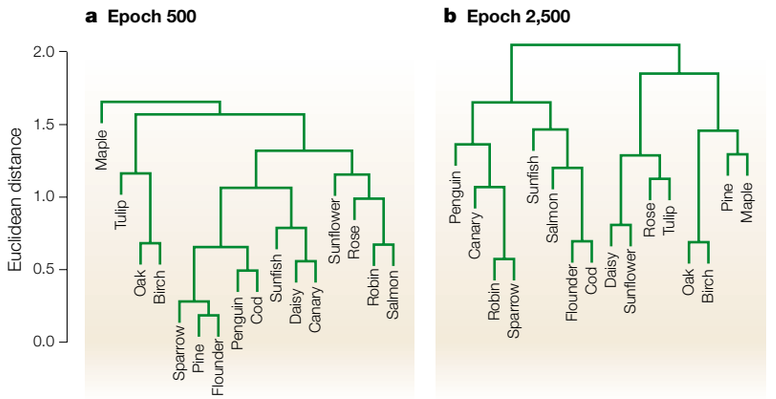


Figure 7 | **Hierarchical clustering of the 21 concepts used in the simulation capturing the coalescence of underlying conceptual structure after initial acquisition of superficial, appearance-based structure. a** | The similarity structure acquired early in learning, when the superficial appearance ('IS') properties, available on every learning trial, have been learned. The groupings are strongly affected by these properties ('IS yellow' applies to 'sunfish', 'daisy' and 'canary'; 'IS red' applies to 'robin', 'rose' and 'salmon'). **b** | The similarity structure acquired later in learning, when the network has become sensitive to the coherent covariation of properties that occur less frequently and in different relational contexts. The concepts of plant and animal, and the subordinate concepts of trees, flowers, birds, fish and land animals, have coalesced at this point. (Based on simulations reported in REF. 78.)

are weak, it is interesting that the 3–4-year-old children used show an effect at all, and stronger effects are seen in older individuals.

The PDP model also tends to generalize properties differently depending on the type of concept and property[78]. Differential generalization by type of concept arises in our model when there is domain-specific covariation of properties in experience. Shape tends to covary more than colour with type of toy (cars can be any colour but must have four wheels) whereas colour tends to covary more than shape with type of food. The model's learning algorithm is sensitive to this domain-specific covariation. To illustrate this, we carried out simulations in which two properties (IS large and IS small) vary coherently with those that distinguish trees from flowers, but vary randomly with respect to those that distinguish fish from birds. As a result, the network learns more readily about the relative sizes of plants than animals. It also assigns different internal representations to plants that differ only in size, but similar internal representations to animals that differ only in size. In other words, size becomes 'important' for representing plants, but not animals. This outcome results from the same forces of coherent covariation that underlie the network's sensitivity to having wings versus being yellow; the simulation shows that the network is sensitive to differences in coherent covariation of the same feature in different domains. Importantly, no innate domain-specific knowledge is required for this learning — only exposure to training experiences exemplifying the relevant domain-specific covariation. The model also shows differential generalization of different properties of the same concept: newly acquired 'HAS' properties tend to generalize in accordance with the taxonomic hierarchy, whereas new 'IS' properties generalize much more idiosyncratically, reflecting the fact that such properties do not covary reliably with the structure of the taxonomic hierarchy.

Third, we consider the reorganization of conceptual knowledge in development. Carey[84] found that young children tend to treat plants and animals very differently, essentially failing to appreciate their conceptual unity as living things. By contrast, older children come to understand what it means to be a living thing, and appreciate the common structure between the domains of plants and animals (both require nutrition, both have self-sustaining processes that can be terminated but not re-started, both involve reproduction of offspring from products that pass on the traits of the parent). Information about these sorts of properties might not be available as frequently as other information that tends to support a distinction between plants and animals, but eventually experience with these commonalities accumulates. Carey suggests that the overarching concept of 'living thing' coalesces as a result of learning about these commonalities, several of which arise in seemingly distinct contexts, but which she nevertheless sees as related to the concept of being alive as we intuitively understand it. A key element of Carey's argument is that the reorganization that occurs represents the replacement of one 'theory' in the child's mind with another, similar to the replacement of one theory with another in the progress of science[92].

A similar process of coalescence can occur in our model, leading to a reorganization of its internal representations (FIG. 7). Here we consider the coalescence of the concepts of plant and animal, and within these of the different basic types (trees, flowers, birds and fish). This kind of coalescence might follow an earlier stage in which object repesentations are based on their more superficial appearance properties. Such a restructuring can occur if the model is more frequently exposed to superficial appearance properties of concepts (the 'IS' properties) than to other properties (the 'CAN', 'HAS' and 'ISA' properties) that covary more coherently. In the simulation, the network masters the more frequently available appearance information first (FIG. 7a), and initially establishes an organization based on this information. Gradually, as it acquires information about the other three types of relation, the coherent covariation of HAS, CAN and ISA information comes to dominate learning, and the internal representations reorganize to capture the underlying taxonomic organization rather than the appearance information (FIG. 7b).

The role of causal information

Finally, we consider more direct evidence that causal information has a role in semantic cognition. Children and adults place more weight on the CAUSAL PROPERTIES of certain objects than on their appearance, or on other properties that are seen as effects of underlying causal properties[93–95]. If a coloured block seems to cause another object to flash and emit noise, children will group it with other blocks that exert similar effects, ignoring shape or colour differences. Furthermore, semantic judgements are influenced by information about the causal mechanisms that give rise to an object's

CAUSAL PROPERTIES
The properties of objects that give rise to predictable outcomes in event sequences in which the object is observed to participate. For example, when pressing a button on the remote control consistently precedes the TV turning on, the remote can be said to have the causal property of turning on the TV.

observed properties. In one study[96], children were told about a raccoon that has come to look like a skunk. In one case, the raccoon looked like a skunk because it was wearing a costume; in another it was given an injection right after birth, and when it grew up it looked like a skunk. Other stories were also used. Very young children thought that the animal had become a skunk in all cases, whereas older children tended to accept some transformations (such as the injection) as changing the raccoon into a skunk but did not accept others (such as the costume). Thus, for older children, whether the raccoon has become a skunk depended on the causal mechanism that gave rise to the object's appearance.

Under the theory theory, naive domain theories about causal mechanisms produce the effects seen in these studies. We suggest that the principles of PDP can explain how we learn to be sensitive to the causal properties of objects. Although PDP models have not been applied to modelling these studies, we have suggested[78] that this would be possible using recurrent networks that learn representations of items in event sequences[55,97–99] or verbal descriptions of such sequences[100]. In these networks, events are broken into a series of steps, and the task is to learn to predict each step from information extracted from previous steps. In such networks the representations assigned to items are affected by the consequences of the item's occurrence and by the influences of other items. Knowledge about the causal properties of an object such as a remote control device could be acquired in such networks through learning from events in which pressing a button on the device is followed by the TV switching on or off. The tendency to generalize on the basis of shared causal powers, rather than superficial appearance properties such as shape or colour, would arise for such objects because their causal powers covary coherently with other properties (such as having buttons to press, requiring batteries inside, depending on a clear line of sight). Similarly, we suggest, our understanding of what it means to be in a costume arises from experiences of oneself and others in costume, in which we learn that the wearer can still feel, sound and behave like himself while wearing the costume and will return to his former appearance upon removing it. On the basis of experience with costumes, children and networks could come to know that the wearer is unchanged by the costume, in spite of current appearances, and so they would know that a raccoon in skunk's clothing is still a raccoon underneath.

In summary, work within the theory-theory framework has increased appreciation for the subtlety and complexity of human semantic cognition. For some this work has suggested that something more structured than a PDP network — with built-in symbol-manipulation abilities[101] or built-in domain specific constraints[85,102] — must be required to capture the power of human semantic cognition. But simulation models have shown that the very simple Rumelhart network can capture many of the relevant findings. Other models, in which object representations are learned from event sequences, indicate how sensitivity to causal information might be acquired through experience. Further research is certainly needed to establish the viability of the PDP approach to address these issues and to determine just how much built-in structure might be required, but we are optimistic that the principles of learning at work in the models will allow them to address the full range of relevant findings.

1. Quillian, M. R. in *Semantic Information Processing* (ed. Minsky, M.) 227–270 (MIT Press, Cambridge, Massachusetts, 1968).
2. Warrington, E. K. The selective impairment of semantic memory. *Q. J. Exp. Psychol.* **27**, 635–657 (1975).
Classic paper first reporting the syndrome that is now known as semantic dementia.
3. Snowden, J. S., Goulding, P. J. & Neary, D. Semantic dementia: a form of circumscribed cerebral atrophy. *Behav. Neurol.* **2**, 167–182 (1989).
4. Hodges, J. R., Graham, N. & Patterson, K. Charting the progression in semantic dementia: implications for the organisation of semantic memory. *Memory* **3**, 463–495 (1995).
5. Bozeat, S., Lambon-Ralph, M. A., Patterson, K. & Hodges, J. R. When objects lose their meanings: what happens to their use? *Cogn. Affect. Behav. Neurosci.* (in the press).
6. Rogers, T. T., McClelland, J. L. Patterson, K., Lambon-Ralph, M. A. & Hodges, J. R. A recurrent connectionist model of semantic dementia. *Cogn. Neurosci. Soc. Annual Meeting* <http://cognet. mit. edu/posters/poster. tcl?publication_id=3664> (1999).
7. Keil, F. C. *Semantic and Conceptual Development: An Ontological Perspective* (Harvard Univ. Press, Cambridge, Massachusetts, 1979).
8. Mandler, J. M. How to build a baby: II. Conceptual primitives. *Psychol. Rev.* **99**, 587–604 (1992).
9. Mandler, J. M. Perceptual and conceptual processes in infancy. *J. Cogn. Dev.* **1**, 3–36 (2000).
Reviews 10 years of research on progressive differentiation of conceptual knowledge in infancy.
10. Mandler, J. M., Bauer, P. J. & McDonough, L. Separating the sheep from the goats: differentiating global categories. *Cogn. Psychol.* **23**, 263–298 (1991).
11. Mandler, J. M. & McDonough, L. Concept formation in infancy. *Cogn. Dev.* **8**, 291–318 (1993).
12. Pauen, S. The global-to-basic shift in infants' categorical thinking: first evidence from a longitudinal study. *Int. J. Behav. Dev.* **26**, 492–499 (2002).
13. Rakison, D. in *Early Concept and Category Development: Making Sense of the Blooming, Buzzing Confusion* (eds Oakes, L. M. & Rakison, D. H.) (Oxford Univ. Press, New York, in the press).
14. Mareschal, D. Infant object knowledge: current trends and controversies. *Trends Cogn. Sci.* **4**, 408–416 (2000).
15. Quinn, P. C. in *Blackwell Handbook of Childhood Cognitive Development* (ed. Goswami, U.) 84–101 (Blackwell, Oxford, UK, 2002).
16. Collins, A. M. & Quillian, M. R. Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* **8**, 240–248 (1969).
17. McCloskey, M. & Glucksberg, S. Decision processes in verifying category membership statements: implications for models of semantic memory. *Cogn. Psychol.* **11**, 1037 (1979).
18. Murphy, G. L. & Brownell, H. H. Category differentiation in object recognition: typicality constraints on the basic category advantage. *J. Exp. Psychol. Learn. Mem. Cogn.* **11**, 70–84 (1985).
19. Rips, L. J., Shoben, E. J. & Smith, E. E. Semantic distance and the verification of semantic relations. *J. Verbal Learn. Verbal Behav.* **12**, 1–20 (1973).
20. Rosch, E. Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* **104**, 192–233 (1975).
21. Rumelhart, D. E. & Ortony, A. in *Schooling and the Acquisition of Knowledge* (eds Anderson, R. C., Sprio, R. J. & Montague, W. E.) 99–135 (Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1976).
22. Rumelhart, D. E. & Abrahamson, A. A. A model of analogical reasoning. *Cogn. Psychol.* **5**, 1–28 (1973).
23. Anderson, J. R. *The Adaptive Character of Thought* (Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1990).
24. Heit, E. in *Rational Models of Cognition* (eds Oaksford, M. & Chater, N.) 248–274 (Oxford Univ. Press, Oxford, 1998).
25. Heit, E. Properties of inductive reasoning. *Psychon. Bull. Rev.* **7**, 569–592 (2000).
26. Medin, D. L. & Shaffer, M. M. Context theory of classification learning. *Psychol. Rev.* **85**, 207–238 (1978).
27. McClelland, J. L. Retrieving general and specific information from stored knowledge of specifics. *Proc. Third Annu. Conf. Cogn. Sci. Soc.* 170–172 (1981).
28. Nosofsky, R. M. Attention, similiarity and the identification-categorization relationship. *J. Exp. Psychol. Learn. Mem. Cogn.* **115**, 39–57 (1986).
29. Kruschke, J. K. ALCOVE: an exemplar-based connectionist model of category learning. *Psychol. Rev.* **99**, 22–44 (1992).
30. Nosofsky, R. M. & Palmeri, T. J. An exemplar-based random-walk model of speeded classification learning. *Psychol. Rev.* **104**, 266–300 (1997).
31. Murphy, G. L. *The Big Book of Concepts* (MIT Press, Cambridge, Massachusetts, 2002).
An up-to-date review of the literature on concepts and semantic cognition.
32. Rumelhart, D. E., McClelland, J. L. & the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations* (MIT Press, Cambridge, Massachusetts, 1986).
Introduces the parallel distributed processing framework. Chapter 6 describes back-propagation, which is briefly presented in reference 35.
33. Hinton, G. E. in *Parallel Models of Associative Memory* (eds Hinton, G. E. & Anderson, J. A.) 161–187 (Erlbaum, Hillsdale, New Jersey, 1981).
34. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9**, 147–169 (1985).
35. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
36. Williams, R. J. & Zipser, D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.* **1**, 270–280 (1989).
37. Hinton, G. E. & McClelland, J. L. in *Neural Information Processing Systems* (ed. Anderson, D. Z.) 358–366 (American Institute of Physics, New York, 1988).
38. Mazzoni, P., Andersen, R. A. & Jordan, M. I. A more biologically plausible learning rule for neural networks.

Proc. Natl. Acad. Sci. USA **88**, 4433–4437 (1991).
39. O'Reilly, R. C. Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Comput.* **8**, 895–938 (1996).
40. Devlin, J. T., Gonnerman, L. M., Andersen, E. S. & Seidenberg, M. S. Category specific semantic deficits in focal and widespread brain damage: a computational account. *J. Cogn. Neurosci.* **10**, 77–94 (1998).
41. Farah, M. J. & McClelland, J. L. A computational model of semantic memory impairment: modality-specificity and emergent category-specificity. *J. Exp. Psychol. Gen.* **120**, 339–357 (1991).
42. Gotts, S. & Plaut, D. C. The impact of synaptic depression following brain damage: a connectionist account of 'access/refractory' and 'degraded-store' semantic impairments. *Cogn. Affect. Behav. Neurosci.* **2**, 187–213 (2002).
43. Hinton, G. E. in *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (ed. Morris, R. G. M.) 46–61 (Clarendon, Oxford, 1989).
44. Hinton, G. E. & Shallice, T. Lesioning a connectionist network: investigations of acquired dyslexia. Technical Report No. CRG-TR-89–3 (University of Toronto, Department of Computer Science, Toronto, Ontario, Canada, 1989).
45. Lambon Ralph, M. A., McClelland, J. L., Patterson, K., Galton, C. J. & Hodges, J. R. No right to speak? The relationship between object naming and semantic impairment: neuropsychological evidence and a computational model. *J. Cogn. Neurosci.* **13**, 341–356 (2001).
46. McRae, K., de Sa, V. R. & Seidenberg, M. S. On the nature and scope of featural representations of word meaning. *J. Exp. Psychol. Gen.* **126**, 99–130 (1997).
47. Plaut, D. C. Graded modality-specific specialization in semantics: a computational account of optic aphasia. *Cogn. Neuropsychol.* **19**, 603–639 (2002).
48. Plaut, D. C. & Booth, J. R. Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychol. Rev.* **107**, 786–823 (2000).
49. Plaut, D. C. & Shallice, T. Deep dyslexia: a case study of connectionist neuropsychology. *Cogn. Neuropsychol.* **10**, 377–500 (1993).
50. Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R. & Levy, J. P. Conceptual structure and the structure of concepts: a distributed account of category-specific deficits. *Brain Lang.* **75**, 195–231 (2000).
51. Tyler, L. K. & Moss, H. E. Towards a distributed account of conceptual knowledge. *Trends Cogn. Sci.* **5**, 244–252 (2001).
52. Mareschal, D., French, R. M. & Quinn, P. C. A connectionist account of asymmetric category learning in early infancy. *Dev. Psychol.* **36**, 635–645 (2000).
53. Mareschal, D., Plunkett, K. & Harris, P. A computational and neuropsychological account of object-oriented behaviours in infancy. *Dev. Sci.* **2**, 306–317 (1999).
54. McClelland, J. L. in *Parallel Distributed Processing: Implications for Psychology and Neurobiology* (Morris, R. G. M.) 8–45 (Oxford Univ. Press, New York, 1989).
55. Munakata, Y., McClelland, J. L., Johnson, M. H. & Siegler, R. S. Rethinking infant knowledge: toward an adaptive process accout of successes and failures in object permanence tasks. *Psychol. Rev.* **104**, 686–713 (1997).
56. Munakata, Y. & McClelland, J. L. Connectionist models of development. *Dev. Sci.* (in the press).
57. Quinn, P. C. & Johnson, M. H. The emergence of perceptual category representations in young infants: A connectionist analysis. *J. Exp. Child Psychol.* **66**, 236–263 (1997).
58. Schyns, P. G. A modular neural network model of concept acquisition. *Cogn. Sci.* **15**, 461–508 (1991).
59. Shultz, T. R., Mareschal, D. & Schmidt, W. C. Modeling cognitive development of balance scale phenomena. *Machine Learn.* **16**, 57–86 (1994).
60. Rumelhart, D. E. in *An Introduction to Neural and Electronic Networks* (eds Zornetzer, S. F., Davis, J. L. & Lau, C.) 405–420 (Academic, New York, 1990).
61. Rumelhart, D. E. & Todd, P. M. in *Attention and Performance XIV: Synergies in Experimental Psychology, Artifical Intelligence, and Cognitive Neuroscience* (eds Meyer, D. E. & Kornblum, S.) 3–30 (MIT Press, Cambridge, Massachusetts, 1993).
62. Flavell, J. *The Developmental Psychology of Jean Piaget* (Van Nostrand, New York, 1963).
63. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* **102**, 419–457(1995).
Elaborates the complementary learning system theory that is described briefly in the current article.
64. Quinn, P. C. & Johnson, M. H. Global-before-basic object categorization in connectionist networks and 2-month-old infants. *Infancy* **1**, 31–46 (2000).
65. Bloom, P. *How Children Learn the Meaning of Words* (MIT Press, Cambridge, Massachusetts, 2000).
66. McCloskey, M. & Cohen, N. J. in *The Psychology of Learning and Motivation* (ed. Bower, G. H.) 109–165 (Academic, New York, 1989).
67. Marr, D. Simple memory: a theory for archicortex. *Phil. Trans. R. Soc. Lond. B* **262**, 23–81 (1971).
68. Milner, P. A cell assembly theory of hippocampal amnesia. *Neuropsychologia* **27**, 23–30 (1989).
69. Rolls, E. T. in *An Introduction to Neural and Electronic Networks* (eds Zornetzer, S. F., Davis, J. L. & Lau, C.) 73–90 (Academic, San Diego, California, 1990).
70. Alvarez, P. & Squire L. R. Memory consolidation and the medial temporal lobe: a simple network model. *Proc. Natl Acad. Sci. USA* **91**, 7041–7045 (1994).
71. Scoville, W. B. & Milner, B. Loss of recent memory after bilateral hippocampal lesions. *J. Neurosurg. Neurol. Psychiatry* **20**, 11–21 (1957).
72. Squire, L. R. Memory and the hippocampus: a synthesis from findings with rats, monkeys and humans. *Psychol. Rev.* **99**, 195–231 (1992).
A useful overview of amnesia resulting from lesions to the medial temporal lobes.
73. Nadel, L. & Moscovitch, M. Memory consolidation, retrograde amnesia and the hippocampal complex. *Curr. Opin. Neurobiol.* **7**, 217–227 (1997).
74. Cipolotti, L. *et al.* Long-term retrograde amnesia: the crucial role of the hippocampus. *Neuropsychologia* **39**, 151–172 (2001).
75. Norman, K. A. & O'Reilly, R. C. Modeling hippocampal and neocortical contributions to recognition memory: a complementary learning systems approach. *Psychol. Rev.* (in the press).
76. Hodges, J. R., Patterson, K., Oxbury, S. & Funnell, E. Semantic dementia: progressive fluent aphasia with temporal lobe atrophy. *Brain* **115**, 1783–1806 (1992).
77. Garrard, P. & Hodges, J. R. Semantic dementia: clinical, radiological, and pathological perspectives. *J. Neurol.* **247**, 409–422 (2000).
78. Rogers, T. T. & McClelland, J. L. *Semantic Cognition: A Parallel Distributed Processing Approach* (MIT Press, Cambridge, Massachusetts, in the press).
A full presentation of the PDP approach.
79. Mervis, C. B. in *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization* (ed. Neisser, U.) 201–233 (Cambridge Univ. Press, Cambridge, 1987).
80. Jolicoeur, P., Gluck, M. & Kosslyn, S. M. Pictures and names: making the connection. *Cogn. Psychol.* **19**, 31–53 (1984).
81. Rosch, E., Mervis, C. B., Gray, W., Johnson, D. & Boyes-Braem, P. Basic objects in natural categories. *Cogn. Psychol.* **8**, 382–439 (1976).
82. Brown, R. *A First Language* (Harvard Univ. Press, Cambridge, Massachusetts, 1973).
83. MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk* (Lawrence Erlbaum Associates, Mahwah, New Jersey, 2000).
84. Carey, S. *Conceptual Change in Childhood* (MIT Press, Cambridge, Massachusetts, 1985).
Seminal studies of conceptual reorganization in children.
85. Keil, F. C. in *The Epigenesis of Mind: Essays on Biology and Cognition* (eds Carey, S. & Gelman, R.) 237–256 (Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1991).
86. Murphy, G. L. & Medin, D. L. The role of theories in conceptual coherence. *Psychol. Rev.* **92**, 289–316 (1985).
87. Gopnik, A. & Wellman, H. M. in *Mapping the Mind: Domain Specificity in Cognition and Culture* (eds Hirschfeld, L. A. & Gelmans, S. A.) (Cambridge Univ. Press, Cambridge, 1994).
One of the most explicit attempts to articulate the nature of 'theory theory'.
88. Wilson, R. A. & Keil, F. C. in *Explanation and Cognition* (eds Keil, F. C. & Wilson, R. A.) 87–114 (MIT Press, Boston, Massachusetts, 2000).
89. Gelman, R. First principles organize attention to and learning about relevant data: Number and the animate/inanimate distinction as examples. *Cogn. Sci.* **14**, 79–106 (1990).
90. Macario, J. F. Young children's use of color in classification: Foods and canonically colored objects. *Cogn. Dev.* **6**, 17–46 (1991).
91. Gelman, S. A. & Markman, E. M. Categories and induction in young children. *Cognition* **23**, 183–209 (1986).
92. Kuhn, T. *The Structure of Scientific Revolutions* (Univ. Chicago Press, Chicago, 1962).
93. Ahn, W.-K. Why are different features central for natural kinds and artifacts?: The role of causal status in determining feature centrality. *Cognition* **69**, 135–178 (1998).
94. Ahn, W., Gelman, S., Amsterlaw, J. A., Hohenstein, J. & Kalish, C. W. Causal status effect in children's categorization. *Cognition* **76**, B35–B43 (2000).
95. Gopnik, A. & Sobel, D. M. Detecting blickets: how young children use information about novel causal powers in categorization and induction. *Child Dev.* **71**, 1205–1222 (2000).
96. Keil, f. *Concepts, Kinds, and Cognitive Development* (MIT Press, Cambridge, Massachusetts, 1989).
Presents key findings relevant to the role of causal mechanisms in determining what kind of thing an object is.
97. Elman, J. L. Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
98. Cleeremans, A., Servan-Schreiber, D. & McClelland, J. L. Finite state automata and simple recurrent networks. *Neural Comput.* **1**, 372–381 (1989).
99. Rogers, T. T. & Griffin, R. in *Proc. 13th Biennial Int. Conf. Infant Studies* (eds Schmuckler, M. A. Trehub, S. E. & Kosarev, M. F.) 625 (International Society for Infant Studies, 2002).
100. St. John, M. F. The story Gestalt: a model of knowledge-intensive processes in text comprehension. *Cogn. Sci.* **16**, 271–306 (1992).
101. Marcus, G. F. *The Algebraic Mind* (MIT Press, Cambridge, Massachusetts, 2001).
Critique of the parallel distributed processing framework, offering an alternative perspective on the basic nature of human cognitive and semantic processing.
102. Carey, S. & Spelke, E. in *Domain-Specific Knowledge and Conceptual Change* (ed. Hirschfeld, L. A.) 169–200 (Cambridge Univ. Press, New York, 1994).
103. Barsalou, L. W., Simmons, W. K., Barbey, A. & Wilson, C. D. Grounding conceptual knowledge in modality-specific systems. *Trends Cogn. Sci.* (in the press).
104. Damasio, A. R. The brain binds entities and events by multiregional activation from convergence zones. *Neural Comput.* **1**, 123–132 (1989).
105. Martin, A., Haxby, J. V., Lalonde, F. M., Wiggs, C. L. & Ungerleider, L. G. Discrete cortical regions associated with knowledge of color and knowledge of action. *Science* **270**, 102–105 (1995).
106. Kellenbach, M. L., Brett, M. & Patterson, K. Actions speak louder than functions: the importance of manipularity and action in tool representation. *J. Cogn. Neurosci*. **15**, 30–46 (2003).
107. Kellenbach, M. L., Brett, M. & Patterson, K. Larger, colorful, or noisy? Attribute- and modality-specific activations during retrieval of perceptual attribut knowledge. *Cogn. Affect. Behav. Neurosci*. **1**, 207–221 (2001).
108. Martin, A. & Chao, L. L. Semantic memory in the brain: structure and processes. *Curr. Opin. Neurobiol.* **11**, 194–201 (2001).
A useful contemporary review of findings on semantic representation in the brain.
109. Warrington, E. K. & Shallice, T. Category specific semantic impairments. *Brain* **107**, 829–854 (1984).
Seminal report of category-specific deficits in semantic knowledge.
110. Rogers, T. T. & Plaut, D. C. in *Category Specificity in Mind and Brain* (eds Forde, E. & Humphreys, G.) 251–289 (Psychology Press, East Sussex, UK, 2002).
111. Gonnerman, L. M., Andersen, E. S., Devlin, J. T., Kempler, D. & Seidenberg, M. S. Double dissociation of semantic categories in Alzheimer's disease. *Brain Lang.* **57**, 254–279 (1997).
112. Allport, D. A. in *Current Perspectives in Dysphasia* (eds Newman, S. K. & Epstein, R.) 207–244 (Churchill Livingstone, Edinburgh, 1985).
113. Caramazza, A. & Shelton, J. R. Domain-specific knowledge systems in the brain: the animate-inanimate distinction. *J. Cogn. Neurosci.* **10**, 1–34 (1998).
114. Warrington, E. K. & McCarthy, R. Categories of knowledge: further fractionation and an attempted integration. *Brain* **110**, 1273–1296 (1987).
115. Jacobs, R. A. & Jordan, M. I. Computational consequences of a bias toward short connections. *J. Cogn. Neurosci.* **4**, 323–336 (1992).
116. Kohonen, T. The self-organizing map. *Proc. IEEE* **78**, 1464–1480 (1990).
117. Snodgrass, J. G. & Vanderwart, M. A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Learn. Mem. Cogn.* **6**, 174–215 (1980).
118. Gelman, R. & Williams, E. M. in *Handbook of Child Psychology, Volume II: Cognition, Perception and Development* (eds Kuhn, D. & Siegler, R.) 575–630 (John Wiley and Sons, New York, 1997).
119. Plaut, D. C. Double dissociation without modularity: evidence from connectionist neuropsychology. *J. Clin. Exp. Neuropsychol.* **17**, 291–321 (1995).

Acknowledgements
Preparation of this article was supported by the National Institute of Mental Health (USA) and the Medical Research Council (UK).