Attribute Theories of Memory[*]

Sean M. Polyn

August 31, 2022

Abstract

Attribute theories propose that a cognitive system constructs representations with constituent elements that reflect the attributes, features, properties, or characteristics of the world and the things in it. These theories are flexible; Attributes may be used to describe tangible physical things in one's environment, but also intangible things like plans and concepts. In the domain of memory, attribute theories describe how representations of one's experience are stored, retrieved, and manipulated by the cognitive system. The mathematical language of linear algebra is central to these theories. Vector operations are used to define various measures of representational distance and similarity. Matrices are used to define associative structures like synaptic weights in neural network models, and sets of memories in instance theory models. We examine some foundational attribute-based approaches to memory theory, including early attempts to define psychological similarity, and the influential stimulus-sampling theory. We review important empirical phenomena involving memory attributes, including proactive interference, encoding specificity, and source memory. Finally, we consider how attribute-based approaches have influenced modern cognitive neuroscientific investigations and theory.

[*] This work was supported by a grant from the National Science Foundation to SMP (#1756417). Cite as: Polyn, S. M. (2022) Attribute theories of memory. In Wagner, A. D. and Kahana, M. J. (Ed.). Oxford Handbook of Human Memory.

Keywords: representation; multidimensional; feature; similarity; distance; semantic; vector; matrix

1 Introduction

An attribute theory of memory focuses on the structure of cognitive representations, on the nature of the elements composing them, and on the mechanisms, operations, and processes necessary to store and retrieve them. But what is an attribute, and what is a representation? These terms are deliberately broad, and are used in a variety of ways in the literature. Perhaps it is easiest to start concretely, by considering the perceptual system. At any given waking moment, your perceptual system is busily processing your surroundings, constructing internal representations that reflect the structure of the external world (Churchland and Sejnowski, 1992). During breakfast, your gaze lingers on a mug, and a representation is constructed: a code specifying the attributes of the mug, in other words, its characteristics, its features, its properties, its qualities. These are potentially concrete things: The color of the mug, the shape of its handle, the scent of the coffee filling it. But a cognitive theory needs to also deal with things that are markedly less concrete: The intention to drink the coffee, the verbal response to a query from a loved one, the plans for the day ahead. In an attribute-based framework, representations can be constructed for all these things and more; any thought, any wish, desire, hope, or belief can be treated as a representation, and can be stored, retrieved, and manipulated by the cognitive system as part of its everyday processing.

The key idea here is that representations have internal structure; they are composed of elements, which are a kind of theoretical building block. Each element has an associated state, which means it has a value associated with it at a given moment. A particular set of values across a set of elements forms a pattern, and the pattern specifies the attributes of whatever is being represented. Associations (another theoretical building block) allow us to link the elements comprising a particular representation to one another, and to link elements of one representation to elements of another representation. A wide variety of attribute-based theoretical frameworks have been developed over the decades, and as we will see, different frameworks take dramatically different approaches regarding the nature of both elements and associations.

These basic ideas regarding elements and associations form the foundation of attribute theories of memory, and can be traced back to the earliest theories of memory. For example, in 1859, Hamilton wrote about the law of redintegration, whereby thoughts that are part of the same act of cognition become linked, and may suggest one another in the future (OED Online, 2019). Hollingsworth (1928) defined redintegration in somewhat more modern terms as “the type of process in which a part of a complex stimulus provokes the complete reaction that was previously made to the complex stimulus as a whole.” These statements contain within them the seeds of theoretical concepts suggesting a dynamic approach to memory. Thoughts, stimuli, and memories may be made of elemental parts that become associated with one another, and these associations may support retrieval and reactivation of other elements, whether they correspond to other memories, or past responses.

We begin the chapter by introducing some of the mathematical concepts commonly used in attribute theories, and then review some of the interesting ways that these concepts have been used to develop theories of memory. Some of the approaches are more computationally or mechanistically explicit, and others are more abstract. Some attempt to specify the neural substrate of particular representations or processes, while others remain agnostic about the neural underpinnings of memory. Each of these varied approaches has made a substantial contribution to the literature.

2 The mathematical language of attribute theories

The language of linear algebra is a natural one for attribute-based cognitive theories, in that it defines mathematical constructs that can be used as representations and associations in formal and informal models of cognitive processes. For most of this chapter, we will need

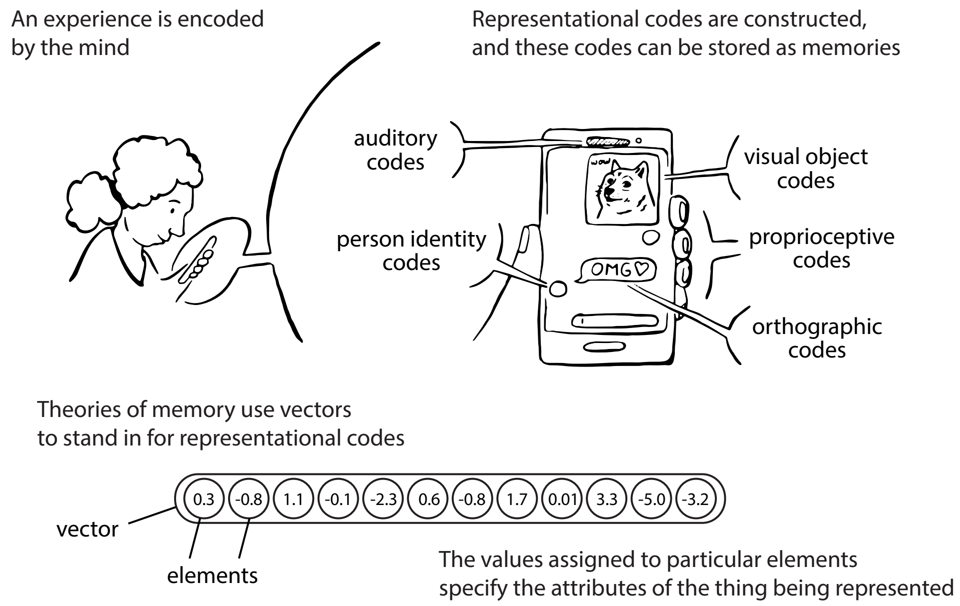
An experience is encoded
by the mind
Representational codes are constructed,
and these codes can be stored as memories
auditory
codes
visual object
codes
person identity
codes
proprioceptive
codes
orthographic
codes
wow
OMG♡
Theories of memory use vectors
to stand in for representational codes
0.3
-0.8
1.1
-0.1
-2.3
0.6
-0.8
1.7
0.01
3.3
-5.0
-3.2
vector
elements
The values assigned to particular elements
specify the attributes of the thing being represented

Figure 1: A schematic overview of some of the main ideas behind attribute theories of memory.

only the basic tools of linear algebra: Vectors, vector spaces, and matrices (Perlis, 1991). The basic ideas will be set up in this section, more advanced mathematical tools will be described as they arise in particular sections.

2.1 Vectors.

A vector is a ordered set of numbers, either a row or a column. Each one of these numbers is an element (a representational element, using the language established above). If a particular vector z has 3 elements, an index variable (say, i) can be used to refer to the value associated with each element (e.g., z_1, z_2, z_3). A particular vector exists within a vector space, just as a particular integer exists along a number line. As you're getting your bearings thinking about these ideas, it can be useful to think about a Cartesian x-y plane. Each point on the plane can be described as an ordered set of 2 numbers, the x-coordinate and the y-coordinate. The vector space is the plane itself, and can be thought of as the set of all possible vectors that could be represented within the space. So, a vector is simultaneously a set of numbers and a point in a space. Thinking about vectors as points in a 2- or 3-dimensional space is useful because you can visualize these low-dimensional spaces. However, care should be taken about generalizing your intuitions about low-dimensional spaces to high-dimensional spaces, where some common intuitions may be violated. Some excellent work by Kanerva describes the nature and properties of high-dimensional representational spaces (Kanerva, 1988, 2009).

In an attribute theory, a particular vector can correspond to a particular cognitive representation. Each element is assigned a specific number, which can be thought of as specifying an attribute or feature. In some models each attribute may have an obvious meaning. Consider how computers represent color. A 3-dimensional vector can be used to represent a wide variety of colors in an RGB color space. Here each attribute has a clear meaning: The values of the three elements describe the amount of red, green, and blue present in the color. A particular shade of blue could have the coordinates (54, 140, 203). This vector specifies a point in RGB color space. However, there are other ways to represent this particular piece of information. For example, the same color can be represented in a 4-dimensional CMYK color space, with coordinates (73, 31, 0, 20). For these color spaces, each attribute has a well-defined meaning, but as we will see, one can construct many theoretically useful kinds of representations where there is no obvious meaning for a given attribute.

As mentioned above, a vector space is the set of all possible vectors that can be represented within the space. In other words, the vector space defines what values are allowed to be used in constructing a particular vector. An 8-bit digital RGB space is discrete, each element can only take on integer values from 0 to 255 (yielding 256^3 possible colors). The binary vector space of a Hopfield network is also discrete, each element can only take the states +1 or -1 (or +1 and 0, depending on which paper you're reading). Vector spaces can also be continuous. In many linear associative models, each element in a representational vector is a real number, making the set of possible vectors infinite. There are no hard rules for what kind of vector space to use in a particular attribute model, only traditions. This is not to say that one's choice of vector space has no consequence, or that different kinds of vector spaces are necessarily interchangeable, but it seems reasonable to say that the consequences are not always well understood.

2.2 Distance and similarity.

A variety of vector-based mathematical operations have been incorporated into attribute theories of memory, standing in for hypothesized cognitive operations carried out by the memory system. It will often be useful to compare two representations, either in terms of the distance between the two points in a vector space, or in terms of the similarity of the two representations. A variety of mathematical operations are often brought to bear on the calculation of distance and similarity, which of course are deeply related.

A common feature to measures of distance is that if the two vectors under consideration are identical, the distance between them is zero. Apart from that, there are a wide variety of techniques available to characterize distance, each potentially having its own benefits, drawbacks, or quirks. Of course, two vectors must exist within the same vector space if the distance between them is to be calculated. Sometimes the choice of vector space will make one kind of distance measure more sensible. For example, Hamming distance is often used to calculate distance in a binary or discrete vector space: One iterates through the indices of the two vectors, and counts up the number of positions where the two vectors have different values.

In a continuous Cartesian vector space, Euclidean distance may be the most familiar distance measure for most students; it is the length of a straight line connecting the two points. In 2-d space the Pythagorean theorem can be used to calculate the Euclidean distance between two points, and an extension of this theorem allows one to do the same for an n-dimensional space. Other distance functions are also possible, including what is sometimes referred to as city-block (or Manhattan) distance, which is defined as the sum of the absolute differences of each of the Cartesian coordinates of the points.

This brings us to the notion of psychological similarity, which is usually defined as a function of some distance measure. A unitary definition of similarity is difficult, as the word gets used in many ways in the scientific literature. Just as there are many ways to characterize distance (even just considering Cartesian spaces), there are many ways to characterize similarity. In a sense, similarity is the inverse of distance, in that the similarity of two vectors will tend to increase as the distance between them decreases. However, whether this inverse relationship will actually hold depends on which distance measure and which similarity measure are being used. The mapping from a distance measure to a similarity measure is not always straightforward. A variety of functions are used to characterize the similarity of two vectors. For many similarity measures, the similarity of two identical vectors is equal to one, and similarity falls as distance between the vectors increases.

In a number of models described in upcoming sections, similarity decays as a negative exponential function of the Euclidean distance between two points. If the distance between 2 points is indicated by D(a,b), then:

$$similarity = e^{-\tau D(a,b)} = \frac{1}{e^{\tau D(a,b)}} \tag{1}$$

With this functional form, as Euclidean distance increases, similarity decreases in a nonlinear fashion: A distance of 0 yields a similarity of 1, and as distance begins to increase, similarity drops sharply before leveling off. The term τ commonly appears as a scaling parameter. This negative exponential function allows a model to capture nonlinear effects in generalization gradients across a number of experiments (see Section 3.1). The nonlinear mapping from distance to similarity means that small differences in distance are more important when the two things being compared are quite similar than when they are already quite distinct. As a fanciful example to illustrate this point, imagine an animal has a memory for the shade of red indicating a perfectly ripe berry. While foraging, it encounters a series of berries with small deviations from that ideal shade of red, and these differences are important, as they will determine the quality of the animal's meal. Accordingly, the nonlinear similarity rule will ensure that these small differences in color correspond to substantial differences in similarity. However, if it encounters a set of berries that are already quite far from the target color (say, different shades of green), minor variations in this distance are less important (as the berries are already clearly inedible), and accordingly correspond to small differences in similarity.

A cosine similarity score provides an alternative way to characterize psychological similarity, and finds widespread use in modeling applications and machine learning. In calculating a cosine similarity score we consider two vectors as lines projecting from the origin of the coordinate system. Cosine similarity is the cosine of the angle between those two lines at the origin. Because cosine similarity depends only on the angle between the two lines projecting from the origin, it does not have a monotonic relationship to Euclidean distance. Two vectors can have an arbitrary Euclidean distance separating them, but if they are collinear with respect to the origin of the coordinate system, they have a cosine of 1; in other words they are perfectly identical with regard to cosine similarity. Figure 2 demonstrates how cosine and Euclidean distances can diverge in this regard. This example is not meant to advocate for either approach, it is simply meant to illustrate how the cosine similarity measure has a form of scaling built into it. The magnitude of the vector doesn't matter, just the angle it takes with respect to the origin. Indeed, this property makes it more appropriate than Euclidean distance for a wide variety of applications.

2.3 Matrices.

Models of memory need a way to store the information contained within cognitive representations. For this, many models use matrices as associative structures. Whereas a vector is a 1-dimensional list of numbers (a row or a column) with a single index, a matrix is more like 2-dimensional table, with 2 indices (the first specifying the row and the second specifying the column of a given element). Instance theories and neural network theories both use matrices to store memories, though in different ways.

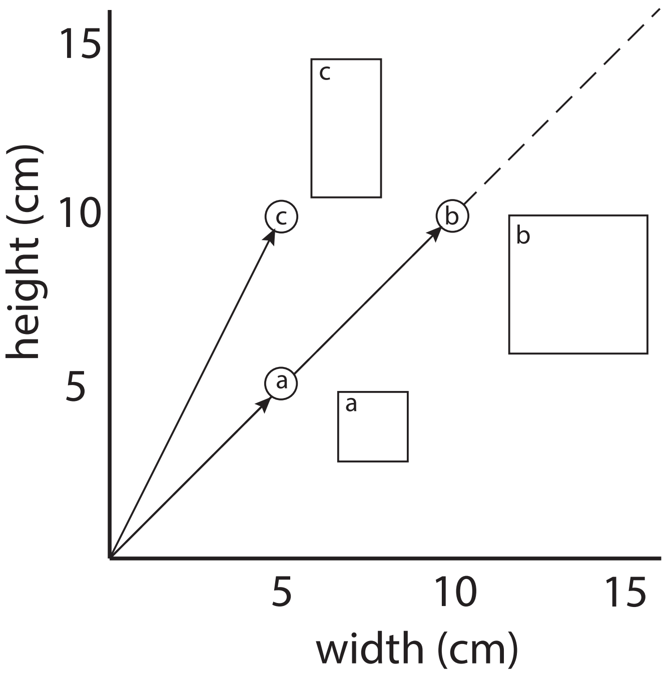

Figure 2: A coordinate system for a representational space describing different rectangles, extending an example developed by Kahana (2012). The x-coordinate of a point specifies the width of a rectangle while the y-coordinate describes its height. Points along the diagonal depict different sizes of squares. Three points correspond to three rectangles: **a** has coordinates (5,5) and **b** (10,10), making them both squares, and **c** has coordinates (5,10), making it a rectangle with width 5 cm and height 10 cm. In this representational space, the Euclidean distance between **a** and **c** is 5, and for **a** and **b** this is $\sqrt{50} \approx 7.1$. The angle separating **a** and **c** at the origin is $\approx 18°$, yielding a cosine similarity of ≈ 0.95. The angle separating **a** and **b** at the origin is 0°, yielding a cosine similarity of 1.0.

3 Early computational theories emphasizing attributes

One might think that the earliest attribute-based theories of memory would deal with quantifiable attributes with clear relations to observable properties of studied material. If anything, the opposite is true: Early theories in this domain grappled with seemingly intractable and often unobservable psychological concepts ranging from codes representing time to codes for the meaning of particular words. Only a few cases dealt with scenarios in which it was possible to characterize stimulus differences in terms of hand-coded attribute values.

3.1 The nature of psychological similarity.

Attneave (1950) reviewed research going back to the 1930s (with some discussion of earlier philosophical approaches) grappling with the question of whether a “dimensional” approach to psychological research could be fruitful, and whether the concept of similarity could be defined in a satisfactory way. He notes that in very particular cases, such as a series of sinusoidal tones, one could plausibly characterize the stimuli along a particular physical dimension. However, he raised strong reservations about whether such an approach could be generally useful, noting that the nature of psychological dimensions was “utterly obscure in complex visual forms” (Attneave, 1950, pg. 519).

Research into the nature of psychological attributes around this time was dominated by a few types of experimental paradigms. Many studies used a paired-associates design in which a relatively small set of stimuli (usually 9 or so) are designed to vary on a small number of dimensions (e.g., a set of circles with varying diameter, or a set of swatches of the same color that vary in saturation and brightness). Each stimulus is associated with a unique, distinct response (sometimes a verbal label, e.g., the digits “one” through “nine”, with the assignment of stimulus to label randomized and often counterbalanced across participants). The stimuli are then presented to the participant, and the verbal responses are collected and used to create confusion matrices containing the conditional probabilities of making a given verbal response upon the presentation of a particular stimulus.

Shepard (1957, 1958, 1987), working with human and non-human animal data from experiments using this paired-associates design, developed a technique to infer the coordinates of the stimuli in an Euclidean representational space, on the basis of the confusion errors made by the participants in these experiments. In this context a confusion error is observed when the participant produces the incorrect response term for a given stimulus. For example, a participant might see a swatch of color and say ”seven”. If the correct label for that swatch was ”three”, this would be taken as evidence that the swatches labeled ”seven” and ”three” are close to one another in the representational space. Shepard proposed that the Euclidean distances between these coordinate points could provide the underpinnings for a well-defined measure of similarity. He promoted a framework in which two identical stimuli (with distance 0 between them) would be assigned a similarity of 1, and that similarity would decline as a negative exponential function as the distance between two stimuli increased (as described in the previous section).

This technique gave impressively consistent results across a variety of stimulus types including shapes, spectral hues, phonemes, free-form figures, and Morse code signals (Shepard, 1987). Over the decades, the assumption that similarity declines as a negative exponential function of distance has been embedded in a variety of modern memory models (Nosofsky, 1986; Kahana and Sekuler, 2002; Nosofsky et al., 2011). However, the general utility of the paired-associates design for characterizing the psychological coordinates of a given set of stimuli is limited, as one is limited by the number of stimulus-to-label associations a given participant can remember.

Other early work used a judgment-based approach to characterizing similarity. In these studies participants were asked directly to make judgments of similarity on a set of stimuli using some variant of the method of triads (Torgerson, 1958). Stimuli would be pulled from the larger set three at a time, and the participant would indicate which two were more closely related. These judgments were used to estimate the pairwise distances between all the stimuli in the set. These values were organized into a distance matrix, which could then be converted into a set of coordinate representations using a multidimensional scaling technique (Shepard, 1980). These multidimensional scaling techniques allow one to specify how many dimensions the coordinate representations should have, and the algorithm finds coordinates for the stimuli that preserves the original inter-stimulus distances as well as possible.

The coordinate representations constructed by this judgment-based approach generally showed good agreement with those constructed using Shepard's confusion-based approach, despite the methodological differences between the two. Unfortunately, the general utility of the method of triads is also limited in terms of the number of stimuli that can be characterized: The number of comparisons necessary to properly estimate the full set of pairwise distances explodes combinatorially as the number of stimuli increases. Despite this limitation, the method of triads is still used to characterize representational similarity, when the stimulus set is small enough to make the approach tractable (Hutchinson and Lockhead, 1977; Yotsumoto et al., 2007).

Around the same time, Osgood et al. (1957) developed a variant of the method of triads that is worthy of separate note, as it presaged a number of modern techniques designed to characterize the meanings of words in terms of their coordinates in a high-dimensional representational space. Participants in one of these experiments were given a set of target words, and were asked to rate each word on a series of scales, where the two words on either side of the scale were chosen to be opposite in meaning. The participant used the scale to indicate which of the two scale words the target word was more closely related to. In one of the group's initial experiments, 100 participants made judgments on 20 target words (example target words included symphony, baby, father, god, and sword) and 50 scales (example judgment scales included good-bad, large-small, beautiful-ugly, calm-agitated, sharp-dull, brave-cowardly, and hard-soft). A variety of factor analysis techniques were used to estimate the coordinates of each target word in a semantic representational space, where differences in meaning between two words was a function of the distance of their coordinates in this space. As such, this is perhaps the earliest vector space model of semantic meaning; more recent approaches will be described in a later section, and in Chapter 3.6.

3.2 Stimulus sampling theory.

Of all these early attribute-based theories, perhaps the most influential was stimulus sampling theory (SST), as developed by Estes and colleagues in the 1950s (Neimark and Estes, 1967). SST was developed by a number of theorists applying it to a variety of experimental scenarios, which led to a healthy number of variants of the basic theory. Some core aspects of the theory can be communicated by describing how it might be applied to a simple conditioning experiment. Consider an experiment in which a tone is played for a rat, following which a shock is delivered through the floor of a cage, eliciting a freezing behavior. SST envisions the cognitive system of the rat as comprised of a great many representational elements, each of which can be in an active or inactive state (sometimes referred to as *available* and *unavailable* states). Some of these elements might correspond to the perceptual experience of the tone. There is a stochastic component to SST. When the tone is played on a given trial, a randomly chosen subset of the tone elements will become active, but others will be inactive. On another trial, when the same tone is played, a different random (but likely overlapping) subset of tone elements will become active. Other aspects of the experience are assigned different sets of elements: shock-related elements, cage-related elements, and even response-related elements. Each of these populations behaves similarly; on a given trial only a subset of each type of element will be in an active state, reflecting natural variability in how the tone is perceived, how the response is executed, and what aspects of the environment are attended to. On each trial, learning occurs in an all-or-none fashion: Elements in an active state become associated with one another. Returning to the conditioning aspect of the experiment, the probability of observing freezing behavior (if the tone is presented without the shock) is simply the proportion of total tone elements that were successfully associated with the freezing-response elements.

Stimulus sampling theory gave theorists a framework in which to think about the structure and dynamic properties of cognitive representations. By this theory, all aspects of inner mental life could be thought of as bundles of attributes, simple representational elements with simple dynamical properties. In contrast to the other ideas reviewed in this section, SST was not generally used to model the representational structure of particular stimuli. Rather, it was designed to simulate and predict the observable behavior of the organism (sometimes rodent, sometimes human, sometimes other animals) in a variety of experimental conditions. Rules regarding the activation or availability of elements were paired with rules regarding the formation of associations between different classes of elements. Together, these rules specified a powerful learning system which could be applied to a wide variety of experimental phenomena.

Later work by Bower developed a number of these representational ideas. He developed the idea that studying an item elicits a multiattribute representational code, and derived some basic predictions of a memory model with such representations at its heart (Bower, 1967). He also built upon the idea that different subsets of attributes could be activated upon different encounters with a stimulus or situation (Bower, 1972). Here, he proposed that if some contextual process caused a gradual change in which representational elements were in the active state, the system could create a gradually changing representational code that could be stored as part of a memory, and used to make judgments about the temporal order and recency of memories. The nature of temporal memory attributes will be explored further in Chapter 3.2, Memory for time.

4 Empirical approaches to characterizing memory attributes

Underwood (1969) made an early attempt to enumerate some of the different classes of attributes that might comprise a given memory, and to review the proliferation of empirical studies examining memories in terms of these multidimensional characteristics. Evidence was accruing for the theoretical importance of temporal attributes and spatial attributes, attributes specifying frequency, modality, and orthographic properties of studied items (as the experimental literature of the time was dominated by list-learning studies of verbal materials), as well as nonverbal attributes specifying, e.g., affective properties of studied material (see Chapter 3.7, Affective memory). Underwood left open the possibility that contextual attributes were also stored as part of a given memory, a theoretical seed that certainly blossomed over the following decades (see Chapter 5.12, Context reinstatement).

4.1 Release from proactive interference.

Over the years, a number of approaches have used patterns of behavioral performance on memory tasks to infer the representational structure of studied material. Shepard's work, described above, examined the patterns of errors (confusions) in a paired-associates task to infer this structure. In later decades, researchers used the pattern of recall failures on a delayed recall task (known as the Brown-Peterson paradigm; Brown, 1958; Peterson and Peterson, 1959) to similar effect (Wickens, 1970; Wickens et al., 1976). In a representative version of this task, a participant studies a short list of words sampled from a particular taxonomic category. After a short retention interval (during which a distracting task is performed), the participant recalls the items from the most recent list. If a series of study lists all sample study items from the same taxonomic category, a proactive interference effect is observed. Performance decreases as a function of the number of preceding lists containing items from the same category, reaching an asymptote after several same-category lists have been studied. If the next list samples items from a new category, performance increases, a phenomenon referred to as *release from proactive interference*. The drop in recall performance across trials was proposed to be due to interference from other similar memory traces, though there was debate as to whether the interference arose during study or recall (Greene, 1992).

Wickens and others showed that the degree of release was sensitive to the similarity of the items on successive lists, such that a larger drop in similarity would show a greater performance benefit following the shift. For example, a shift from farm animals to vegetables would yield a greater performance increase than a shift from farm animals to wild animals. A number of studies used the technique to characterize the representational similarity of different classes of study materials, by examining the relative effectiveness of shifts in characteristics like taxonomic category, grammatical category, and physical properties like background color on memory performance (as reviewed by Wickens, 1970). In these studies, if a shift in a particular characteristic led to a substantial release from proactive interference, this was taken as evidence that the property being shifted was stored as part of the relevant memory trace, and the degree of release was taken to reflect the psychological distance traversed when that characteristic was altered.

4.2 Encoding variability.

Stimulus sampling theory, as described above, proposed that different encounters with the same stimulus could elicit different encoded representations of the stimulus. This idea of encoding variability was developed by Martin (1968) to explain item-level effects in paired associates learning. He proposed that the representations elicited by a less meaningful stimulus could vary quite a bit from encounter to encounter, which explained (among other things) why these stimuli were harder to learn than more meaningful stimuli. Both Light and Carter-Sobell (1970) and Tulving and Thompson (1973) examined the idea that one could experimentally manipulate the encoded attributes of a to-be-remembered word in ways that would affect its later memorability. For example, consider an experiment in which a participant studies a set of paired associates where the first term biases the semantic interpretation of the second term (e.g., *traffic-jam*). The participant is told that they only have to remember the second term. Later, when they are tested, some of the pairs have been modified, with a new first term that emphasizes a different meaning of the second term (e.g., *strawberry-jam*). Even if the participant is told to only focus on the second term and ignore the first, they will still be worse off trying to recognize the second term, relative to a scenario in which the second term is presented alone, or alongside a word that emphasizes the original meaning. Tulving and Thompson (1973) referred to this as *encoding specificity*: The circumstances of study create a particular stored representation, and that memory will be more accessible if the encoded test item is representationally similar to the stored memory. Encoding specificity is related to the idea of transfer appropriate processing (TAP; Morris et al., 1977; Blaxton, 1989), whereby an item will be better remembered if the mental operations engaged at test are similar to those engaged at study.

4.3 Source memory and context change.

The law of redintegration, described in the introduction, proposes that all of the aspects of an experience become associated with one another. A consequence of this is that memories are *content addressable*: One can retrieve a given memory given just a subset of its attributes as a prompt. A corollary of this is that stored memories contain a variety of details from the original experience, often referred to as source characteristics (Hilgard, 1965; Schacter et al., 1984; Johnson et al., 1993). When a person retrieves a memory of a particular experience, source characteristics are often retrieved as well, even if they are not specifically demanded by the task at hand. Memory for many types of source characteristics have been examined in the literature. Visual source characteristics might specify the font a word was presented in (Kirsner, 1973), while spatial source characteristics might specify where on the page an interesting detail was encountered in a book (Rothkopf, 1971). Auditory source characteristics might specify the specific sound of a person's voice (Geiselman and Bjork, 1980). Johnson et al. (1993) provide an excellent review of the variety of source characteristics studied in the literature, and the cognitive processes engaged when attempting to determine the source of a given memory.

The profligate nature of association formation when a given memory is formed has another consequence: The accessibility of a given memory can be strongly context dependent. The context dependence of memories can be powerfully experienced when you revisit a town or city formerly lived in, after an absence of several years. The flood of memories elicited by the sights and sounds of your former home can be quite powerful (Smith, 1988). The context-dependence of memories has been used experimentally to determine which contextual attributes are actually stored as part of a memory. If the presence or absence of a particular contextual feature affects the accessibility of a given memory, this is taken as evidence that this feature was stored as part of the memory trace. A classic experiment by Godden and Baddeley (1975) demonstrated the context dependence of memories nicely. Participants were SCUBA divers. They studied a list of words in one of two environments, on dry land or at a depth of 20 feet underwater. After a suitable retention interval, participants were tested, either in the same environment, or in the other environment. Memory performance was reliably context dependent. The participants remembered the list better when the study environment matched the test environment, suggesting that participants had formed memories that included environmental attributes. A sizable literature explores the varied implications of the context dependence of memories, and the methodological conditions under which these effects are more or less likely to be observed (Godden and Baddeley, 1980; Smith, 1988; Bjork and Richardson-Klavehn, 1989).

5 Memory attributes in computational models.

The language of various attribute theories has suffused nearly all aspects of memory theory, and even non-computational approaches tend to use terminology drawn from these theories. For example, the term 'encode' has become practically synonymous with 'study' in the memory literature, with the former term suggesting that a representational code has been constructed reflecting the identity of that item. Computational models can still show strong influences from attribute theories without explicitly simulating the structure of individual representations. As noted, many implementations of stimulus sampling theory involve the derivation of equations to predict behavior on the basis of an assumed underlying attribute-based representation, but the representations themselves are usually not simulated. However, many computational models of memory are explicitly and fundamentally attribute theories. They define a vector space for the representations of studied items, and the cognitive processes implemented by the model operate on these representations. Many of the models covered in upcoming chapters (e.g. Chs. 5.1–5.3, among others) are of this type. In this section, we give a few illustrative examples of how attributes are used in different modeling frameworks.

5.1 Instance theories.

In the clinical literature on memory disorders, the episodic memory system is often likened to a file cabinet (Budson and Solomon, 2015). Each memory corresponds to a file, and the memory storage operation involves adding new files to the file drawer. Instance theory models of memory operate much in this way. Influential instance theoretical models of memory include MINERVA & MINERVA II (Hintzman, 1984, 1988), REM (Shiffrin and Steyvers, 1997), ITAM (Logan, 2002), and GCM (Nosofsky, 1987). Here, we describe general properties of instance theories that tend to be true of many specific models.

As in any attribute-based memory model, a vector representation is constructed for a to-be-remembered experience, this is the file in the file cabinet analogy. In terms of the theoretical primitives described in the introduction, the numbers composing this vector are the elements of an instance theory. Once stored in the file cabinet, the vector becomes a memory trace. The file drawer itself is a matrix, where each memory trace is assigned a row. The storage operation simply involves copying the representational vector into this memory matrix. In some models this operation can be error-prone, allowing certain features to be stored incorrectly (Shiffrin and Steyvers, 1997). Regardless of the fidelity of the storage operation, each memory trace is given its own row, so the number of rows in the storage matrix grows as new memories are formed.

There are many ways to probe memory in different instance theoretical models; this topic receives further attention in Chapter 5.2 (Global Matching Models). A retrieval cue is simply a representational vector that is used to probe the memory matrix. This probing operation involves comparing the retrieval cue to each of the memory traces stored in the file drawer and calculating their representational distance or similarity. The degree of match between the probe and the contents of memory can then be used to determine whether the probe is familiar, retrieve a particular memory or a blend of memories, or inform other memory-based decisions.

Above, we refer to the set of numbers comprising a representational vector as elements, which raises the question: Where are the associations in an instance theory? In one sense, all of the elements that comprise a single memory trace are automatically associated with one another when they are stored as part of the same memory trace. As such, an instance theory can capture the association between a pair of items in a paired associates task by creating a single memory trace that contains the representations of both items (Hintzman, 1988; Shiffrin and Steyvers, 1997). Another approach involves using subsets of elements to specifically represent the associative features of a study event. In a paired associates scenario, these associative features would be unique to the specific pair of items being studied, and can be generated using a variety of mechanisms (Murdock, 1982; Metcalfe, 1985; Criss and Shiffrin, 2005).

5.2 Connectionist models.

Generally speaking, connectionist models draw inspiration from neuroscientific theories of neural network function (Rumelhart et al., 1986). The elements in these models are often meant to correspond to neurons, or populations of neurons, and the associations between these elements are usually referred to as synapses. These synaptic connections allow the neurons to communicate with one another. Neural network models can vary substantially in terms of how closely they attempt to capture the biophysical properties of the nervous system. Many models are highly abstract, with the activation state of a neuron and the strength of a synapse each specified by a single number.

Connectionist models are fundamentally representational, in that the pattern of activation states across the neurons in a network is itself a kind of representation (although whether those activation states are meant to be interpreted as attributes will depend on the case in question). Subsets of neurons with distinct functions are often referred to as layers, each of which defines a vector space of possible activation states. For example, an input layer may represent the perceptual characteristics of stimuli, and an output layer may represent the different responses a participant might make. The connections between neurons can be autoassociative (connecting the neurons in a layer to one another) or heteroassociative (connecting the neurons in one layer to those in another layer). Autoassociative connections can allow a particular representation to become stable, whereby activating an incomplete or noisy version of the representation will allow the network to recover the original version (Hopfield, 1982). Heteroassociative connections between two layers create a mapping between the two vector spaces–a particular pattern A in one layer will elicit a partner pattern A' in a connected layer, given a particular configuration of the synapses connecting the two layers (Anderson et al., 1977).

Associations in connectionist models are most obviously implemented in the synapses, but as with instance theoretical models, associative information can also be stored in the representations themselves (i.e., in the activation patterns of particular neurons). An example of this is the conjunctive encoding explored in neural network models of hippocampus, where the activation of certain neurons can reflect the co-occurrence of particular combinations of environmental features (O'Reilly and McClelland, 1994).

Memory traces in neural networks are stored in the synaptic weights connecting the neurons to one another. The equations determining how synaptic weights change as a function of experience are often referred to as learning rules, and a wide variety of learning rules have been proposed and examined over the years. A particular synapse connects a pre-synaptic neuron to a post-synaptic neuron. For most learning rules, the strength of the synapse is influenced by the activation states of both neurons. Simple learning rules include varieties of unsupervised Hebbian learning, where the synaptic strength increases if the pre- and post-synaptic neurons are simultaneously active, and decreases otherwise.

In some cases, learning involves associating a particular stimulus representation (in the input layer of a network) with the representation of a particular response (in the output layer of the network). In this scenario, a supervised or error-driven learning rule may be used. If the activation of the neurons in the output layer matches the target output representation, there is no error, and no learning takes place. However, if there is a discrepancy between the representation in the output layer and the target representation, this creates an error signal that is used in the synaptic learning rule. If successful, the changes in synaptic weights will cause the output representation to more closely resemble the target representation (Rumelhart et al., 1986; Trappenberg, 2009).

An interesting case arises when the input layer (representing a presented stimulus) and the output layer (representing a target response) are not connected directly, but rather are separated by one or more intervening layers (often referred to as hidden layers). An error-driven learning rule known as backpropagation can be used to learn an appropriate set of associative weights to map from input patterns to target output patterns. This causes the network to develop internal representations in the hidden layers that enact the transformation of input pattern to output pattern. These hidden representations contain internal attributes that are not specified by the modeler, but which develop over the course of learning. Influential connectionist models of the development of semantic knowledge relate the dynamics of these internal representations to differences in behavioral performance on semantic tasks across the lifespan and in populations with memory disorders (McClelland et al., 1995; Rogers et al., 2004).

5.3 Semantic models.

Some of the earliest attempts to characterize the structure of human semantic memory treated stored knowledge as a network, using mathematical formalisms from graph theory. In this approach, each word is assigned to a node, and the attributes of the word are embedded in the links between nodes (Collins and Quillian, 1969; Collins and Loftus, 1975). These links were often referred to as labeled relations, in that they had to specify the kind of relationship between the two nodes (e.g., *bird* and *wing* would be connected by a "has a" link). The network approach can be contrasted with the vector-based approach described earlier in this chapter, in which each word is assigned a representational vector composed of a number of attributes which specify its meaning. The vector-based approach can be roughly divided into two classes of models: those using hand-coded representations (Smith et al., 1974; Tversky, 1977), and those deriving the representations from a large corpus of text. Early development of these models focused on capturing choice and response time data from simple semantic judgment tasks, including judgments of category membership (e.g., *A butterfly is an insect*; Smith et al. 1974) and judgments as to whether a given proposition is true or false (e.g., *Many arrows are sharp*; Glass et al. 1974).

Latent Semantic Analysis (LSA; Landauer and Dumais, 1997) provides a representative example of a corpus-based model of semantic knowledge. The research team used a large text corpus pulled from an encyclopedia meant for American students. The corpus was comprised of thousands of documents (encyclopedia entries), each of which contained a set of words. This was used to create a large co-occurrence matrix, where each row corresponded to a unique word from the corpus, and each column corresponded to a document. Each entry in the matrix was a number, which was a function of the number of times the word appeared in that particular document. Then a matrix algebraic technique called singular value decomposition was used to reduce the dimensionality of this matrix (specifically, reducing the number of columns) while preserving to some extent the similarity structure of the row vectors to one another. The result of this process is a word embedding, a set of vector representations of (potentially) every word in a lexicon. Landauer and colleagues showed that the word representations produced by LSA could be used to perform a synonym-based subtest of the *Test of English as a Foreign Language* (TOEFL), by choosing the answer with the largest cosine similarity to the target word. The algorithm performed at similar levels as applicants to U.S. colleges from non-English-speaking countries.

As in Osgood's approach, described above, these models treat words as vector representations, points in a high-dimensional representational space, given meaning only with respect to their relationships to other words. Indeed, for many purposes, the representational vectors themselves are replaced with a matrix containing the pairwise similarities between the vectors, often using a cosine similarity score to calculate similarity. Many modern natural language processing frameworks have the same basic structure as LSA, in the sense that a large text corpus is processed in some way to construct vector representations (Lund and Burgess, 1996; Jones and Mewhort, 2007; Griffiths et al., 2007; Recchia and Jones, 2009; Pennington et al., 2014).

These corpus-based vector space models of semantic structure allow one to construct representational vectors for any word that appears in the text corpus. As with hidden layers in multi-layer connectionist networks, the individual attributes may have no obvious meaning. Rather, the pattern of pairwise distances between word representations carries the important information. A key question that arises is whether these representations correspond in a meaningful, systematic way to the unobservable cognitive representations of words that dwell within a given person's semantic memory. One way to address this question is to test whether one can use the representational vectors to predict aspects of the person's behavioral performance on tasks that involve those words. This question has been answered affirmatively; meaningful and reliable variability in behavioral performance on a variety of tasks can be predicted, as reviewed in the citations referenced above. Chapter 3.6 (Semantic attributes) will delve deeper into these issues.

6 Representational analysis in cognitive neuroscience.

Attneave (1950) in his early review of psychological similarity noted that similar stimuli "might achieve identical, or partially identical, neural representations." He based this statement in part on early work on neural network theory by Hebb (1949) and McCullough and Pitts (1943), which developed the idea that neurons could be thought of as simple processing elements sensitive to specific attributes of a stimulus, and capable of altering their connectivity structure to learn from experience. These ideas were buttressed by neurophysiological studies establishing that individual neurons could be selectively sensitive to particular details of an animal's sensory environment (Lettvin et al., 1959; Hubel and Wiesel, 1962), as well as to more complex higher-order properties of stimuli (Gross et al., 1969).

The ability to record brain activity, be it the vascular response of neural tissue in response to metabolic demands, or the electrical fluctuations of neural circuitry, opened up new possibilities in our ability to find the neural correlates of unobservable memory representations. Certainly not all neuroscientific investigations took a representational mindset. Early lines of work using functional magnetic resonance imaging (fMRI), for example, focused on the differential engagement of particular regions in different variants or conditions of a memory task (e.g., Buckner et al., 2000). Work examining event-related responses in scalp EEG took a similar approach, identifying time-locked voltage fluctuations sensitive to task characteristics (Luck, 2005). These approaches allowed researchers to infer the engagement of particular regions and particular waveforms in memory-related processes, but only rarely gave insight into the nature of neural representations or the coding of particular attributes.

Early attempts to characterize neural signals in terms of their multivariate representational structure includes work by Freeman and colleagues on the topographical analysis of electrophysiological responses in the olfactory system to different odors (Freeman, 1975, 1978), and work by Suppes and colleagues classifying individual word identities from scalp EEG signal (Suppes et al., 1997, 1999). It should be noted that in later work, Freeman rejected a representational interpretation of these topographic patterns, on the basis of a number of inconsistencies with the predictions of representational theories (Freeman and Skarda, 1985, 1992).

A groundbreaking study by Haxby et al. (2001) showed that the taxonomic category of visual images could be reliably decoded from the distributed pattern of brain activity across ventral temporal cortex, recorded using fMRI. This demonstration that fMRI signal could be profitably characterized by multivariate analysis spurred a great many studies examining the representational characteristics of brain activity (Carlson et al., 2003; Kamitani and Tong, 2005; Haynes and Rees, 2005; O'Toole et al., 2007; Kriegeskorte et al., 2008). The analysis techniques brought to bear in these studies were drawn from machine learning and statistics (Hastie et al., 2001; Duda et al., 2001), and these approaches are usually referred to as multivariate pattern analysis (MVPA) and representational similarity analysis (RSA) (Norman et al., 2006; Kriegeskorte et al., 2008; Haxby et al., 2014). MVPA applications often assign distinct category labels to different brain states, and use pattern classification algorithms (e.g. logistic regression) to learn the mapping from brain state to category label. In contrast, RSA applications often take brain states and calculate the pairwise distances between them, constructing a representational distance matrix. Analysis can then focus on the structure of the representational distance matrix, for example examining whether matrices derived from different species have similar structure (Kriegeskorte et al., 2008), or examining whether matrices show meaningful correspondence to matrices derived from different theoretical models of semantic structure (Clarke and Tyler, 2014).

The development of neural representational analysis techniques has been beneficial to cognitive neuroscientific studies of memory, as this approach allows better contact between attribute-based cognitive theories of memory, and the multivariate neural signals recorded during memory task performance. Polyn et al. (2005), using fMRI, showed that category-specific patterns of neural activity observed while a participant studied a list were reactivated during memory search, in the seconds leading up to the successful retrieval of one of the studied items. This work extends a substantial body of studies establishing that neural activity patterns observed when an episode is encoded are reactivated when that episode is later retrieved (Danker and Anderson, 2010). Lewis-Peacock and Postle (2008) used similar categorized stimuli to examine the dynamics of these representations during a working memory task, finding evidence that long-term memory representations support the short-term retention of information.

A wealth of other studies have used representational techniques to examine neural activity in terms of its attributes, and some of these have made impressive contact with attribute-based computational models. For example, a study by Mack et al. (2013) examined neural activity patterns in a categorization task and found that the structure of these representations were more consistent with an instance theory model that stored separate memory traces for different studied category exemplars, as opposed to a model that stored prototype representations averaging together the instances from a given category. A study by LaRocque et al. (2013) examined neural representational structure in relation to predictions from connectionist models of medial temporal lobe cortex and hippocampus (McClelland et al., 1995; Norman and O'Reilly, 2003). These models propose that representational codes in cortical regions should preserve the similarity structure of studied materials, but that pattern separation mechanisms in hippocampus should distort this representational structure. Consistent with these models, LaRocque et al. (2013) found that across-item similarity in cortex predicted subsequent memory performance, but this pattern was flipped in hippocampus, where more distinctive item patterns predicted successful performance. Other studies have delved into the semantic structure of neural activity, the dynamics of category learning, and the formation and utilization of associations between arbitrarily paired items (Weber et al., 2009; Chadwick et al., 2016; Davis and Poldrack, 2014; Schlichting and Preston, 2015; Rissman and Wagner, 2012).

7 Conclusion.

Attributes and representations are central to modern theories of memory. There is wide concordance amongst otherwise heterogeneous theoretical approaches regarding the utility of attributes in constructing a theory of memory. Experiences are encoded into representations that reflect the characteristics and circumstances of the surrounding world. These representations are stored, retrieved, and manipulated by the cognitive system. Mathematical tools drawn from linear algebra have been of great utility in the development of attribute theories of memory. In these approaches, representations are treated as vectors, where the particular configuration of values in a vector establishes the attributes of the memory. These basic principles have been used to establish mechanisms for fundamental memory phenomena, such as the sensitivity of memory to the circumstances of encoding, and the context-dependence of memory retrieval. These representational approaches have opened new doors in cognitive neuroscientific investigations of memory, allowing theorists to make contact between attribute theories of memory and the neural signals recorded during a wide variety of memory tasks.

References

Anderson, J. A., Silverstein, J. W., Ritz, S. A., and Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, 84:413–451.

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, 63:516–556.

Bjork, R. A. and Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In Izawa, C., editor, *Current issues in cognitive processes: The Tulane Flowerree symposium on cognition*, chapter 9. Lawrence Erlbaum Associates, New Jersey.

Blaxton, T. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4):657–668.

Bower, G. H. (1967). A multicomponent theory of the memory trace. In Spence, K. W. and Spence, J. T., editors, *The Psychology of Learning and Motivation : Advances in Research and Theory*, volume 1, pages 229–325. Academic Press, New York.

Bower, G. H. (1972). Stimulus-sampling theory of encoding variability. In Melton, A. W. and Martin, E., editors, *Coding Processes in Human Memory*, chapter 5, pages 85–121. John Wiley and Sons, New York.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10:12–21.

Buckner, R. L., Logan, J., Donaldson, D. I., and Wheeler, M. E. (2000). Cognitive neuroscience of episodic memory encoding. *Acta psychologica*, 105(2–3):127–139.

Budson, A. E. and Solomon, P. R. (2015). *Memory Loss, Alzheimer's disease, and dementia: A practical guide for clinicians*. Elsevier Health Sciences.

Carlson, T. A., Schrater, P., and He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of cognitive neuroscience*, 15(5):704–717.

Chadwick, M. J., Anjum, R. S., Kumaran, D., Schacter, D. L., Spiers, H. J., and Hassabis, D. (2016). Semantic representations in the temporal pole predict false memories. *Proceedings of the National Academy of Sciences*, 113(36):10180–10185.

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

Clarke, A. and Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, 34(14):4766–4775.

Collins, A. M. and Loftus, E. F. (1975). Spreading activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal Of Verbal Learning And Verbal Behavior*, 8(2):240–247.

Criss, A. and Shiffrin, R. (2005). List discrimination in associative recognition and implications for representation. *JOURNAL OF EXPERIMENTAL PSYCHOLOGY LEARNING MEMORY AND COGNITION*, 31(6):1199.

Danker, J. F. and Anderson, J. R. (2010). The ghosts of brain states past: Remembering reactivates the brain regions engaged during encoding. *Psychological Bulletin*, 136(1):87–102.

Davis, T. and Poldrack, R. A. (2014). Quantifying the internal structure of categories using a neural typicality measure. *Cerebral Cortex*, 24(7):1720–1737.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern classification, second edition.* Wiley, New York.

Freeman, W. (1975). *Mass action in the nervous system.* Academic Press.

Freeman, W. J. (1978). Spatial properties of an EEG event in the olfactory bulb and cortex. *Electroencephalogr. Clin. Neurophysiol.*, 44:586–605.

Freeman, W. J. and Skarda, C. A. (1985). Spatial eeg patterns, non-linear dynamics and perception: the neo-sherringtonian view. *Brain Research Reviews*, 10(3):147–175.

Freeman, W. J. and Skarda, C. A. (1992). Representations: Who needs them? In McGaugh, J. L., Weinberger, N. M., and Lynch, G., editors, *Brain Organization and Memory: Cells, Systems and Circuits*, pages 375–380. Oxford University Press, New York.

Geiselman, R. E. and Bjork, R. A. (1980). Primary versus secondary rehearsal in imagined voices: Differential effects on recognition. *Cognitive Psychology*, 12:188–205.

Glass, A. L., Holyoak, K. J., and O'Dell, C. (1974). Production frequency and the verification of quantified statements. *Journal of Verbal Learning & Verbal Behavior*, 13(3):237–254.

Godden, D. R. and Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology*, 66:325–331.

Godden, D. R. and Baddeley, A. D. (1980). When does context influence recognition memory? *British Journal of Psychology*, 71:99–104.

Greene, R. L. (1992). *Human memory: Paradigms and paradoxes.* Lawrence Erlbaum and Associates, Hillsdale, New Jersey.

Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psycholgical Review*, 114(2):211–44.

Gross, C. G., Bender, D. B., and Rocha-Miranda, C. E. (1969). Visual receptive fields of neurons in inferotemporal cortex of the monkey. *Science*, 166(910):13030–1306.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning.* Springer-Verlag, New York.

Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37:435–456.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425–2429.

Haynes, J. D. and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5):686–691.

Hebb, D. O. (1949). *Organization of Behavior.* New York: Wiley.

Hilgard, E. R. (1965). *Hypnotic Susceptibility.* Harcourt, Brace and World.

Hintzman, D. L. (1984). MINERVA 2: A simulation of human memory. *Behavioral Research Methods, Instrumentation, and Computers*, 26:96–101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in multiple-trace memory model. *Psychological Review*, 95:528–551.

Hollingsworth, H. L. (1928). *Psychology: Its Facts and Principles.* D. Appleton and Company.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79:2554–2558.

Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106.

Hutchinson, J. W. and Lockhead, G. R. (1977). Similarity as distance: A structural principle for semantic memory. *Journal of Experimental Psychology: Human Learning and Memory*, 3(6):660–678.

Johnson, M. K., Hashtroudi, S., and Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1):3–28.

Jones, M. N. and Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1):1–37.

Kahana, M. J. and Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research*, 42:2177–2192.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8:679–685.

Kanerva, P. (1988). *Sparse distributed memory*. MIT Press.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive computation*, 1(2):139–159.

Kirsner, K. (1973). An analysis of the visual component in recognition memory for verbal stimuli. *Memory & Cognition*, 1:449–453.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60:1126–1141.

Landauer, T. K. and Dumais, S. T. (1997). Solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spector, K., and Wagner, A. D. (2013). Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. *The Journal of Neuroscience*, 33(13):5466–5474.

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

Lewis-Peacock, J. A. and Postle, B. R. (2008). Temporary activation of long-term memory supports working memory. *Journal of Neuroscience*, 28(35):8765–8771.

Light, L. L. and Carter-Sobell, L. (1970). Effects of changed semantic context on recognition memory. *Journal of verbal learning and verbal behavior*, 9(1):1–11.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, 109:376–400.

Luck, S. J. (2005). *An introduction to the event-related potential technique*. MIT Press.

Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments & Computers*, 28(2):203–208.

Mack, M. L., Preston, A. R., and Love, B. C. (2013). Decoding the brain's algorithm for categorization from its neural implementation. *Current Biology*, 23(20):2023–2027.

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: an encoding variability hypothesis. *Psychological Review*, 75(5):421–441.

McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–57.

McCullough, W. S. and Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133.

Metcalfe, J. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review*, 92:1–38.

Morris, C. D., Bransford, J. D., and Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, 16:519–533.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89:609–626.

Neimark, E. D. and Estes, W. K. (1967). *Stimulus sampling theory.* Holden-Day.

Norman, K. A. and O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review*, 110:611–646.

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–430.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13:87–108.

Nosofsky, R. M., Little, D. R., Donkin, C., and Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psycholgical Review*, 118(2):280–315.

O'Reilly, R. C. and McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus*, 4(6):661–682.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The measurement of meaning.* University of Illinois Press.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., and Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, 19(11):1735–1752.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, volume 12.

Perlis, S. (1991). *Theory of matrices*. Courier Corporation.

Peterson, L. R. and Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58:193–198.

Polyn, S. M., Natu, V. S., Cohen, J. D., and Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, 310:1963–1966.

Recchia, G. and Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, 41(3):647–656.

"redintegration, n" (June 2019). *OED Online*. Oxford University Press.

Rissman, J. and Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. *Annual Review of Psychology*, 63:101–128.

Rogers, T., Lambon Ralph, M., Garrard, P., Bozeat, S., McClelland, J., Hodges, J., and Patterson, K. (2004). Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review*, 111(1):205.

Rothkopf, E. Z. (1971). Incidental memory for location of information in text. *Journal of Verbal Learning and Verbal Behavior*, 10(6):608–613.

Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986). *Parallel distributed processing.* MIT Press.

Schacter, D. L., Harbluk, J. L., and McLachlan (1984). Retrieval without recollection: an experimental analysis of source amnesia. *Journal of Verbal Learning and Verbal Behavior*, 23:593–611.

Schlichting, M. L. and Preston, A. R. (2015). Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22(4):325–345.

Shepard, R. N. (1958). Stimulus and response generalization: tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 55(6):509–522.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Shiffrin, R. M. and Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin and Review*, 4:145–166.

Smith, E. E., Shoben, E. J., and Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81(3):214–241.

Smith, S. M. (1988). Environmental context-dependent memory. In Davies, G. M. and Thomson, D. M., editors, *Memory in context: Context in memory.*, pages 13–34. John Wiley & Sons, Oxford, England.

Suppes, P., Han, B., Epelboim, J., and Lu, Z.-L. (1999). Invariance of brain-wave representations of simple visual images and their names. *Proceedings of the National Academy of Science*, 96(25):14658–14663.

Suppes, P., Lu, Z.-L., and Han, B. (1997). Brain wave recognition of words. *Proceedings of the National Academy of Science*, 94(26):14965–14969.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley, New York.

Trappenberg, T. (2009). *Fundamentals of computational neuroscience*. Oxford University Press.

Tulving, E. and Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80:352–373.

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

Underwood, B. J. (1969). Attributes of memory. *Psychological Review*, 76(6):559–573.

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., and Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia*, 47(3):859–868.

Wickens, D. D. (1970). Encoding categories of words: An empirical approach to meaning. *Psychological Review*, 77(1):1–15.

Wickens, D. D., Dalezman, R. E., and Eggemeier, F. T. (1976). Multiple encoding of word attributes in memory. *Memory & Cognition*, 4(3):307–310.

Yotsumoto, Y., Kahana, M. J., Wilson, H. R., and Sekuler, R. (2007). Recognition memory for realistic synthetic faces. *Memory & Cognition*, 35(6):1233–1244.