Modeling Similarity and Psychological Space

Brett D. Roads[1] and Bradley C. Love[2]

[1]Department of Experimental Psychology, University College London, London, United Kingdom, WC1E; email: b.love@ucl.ac.uk
[2]Department of Experimental Psychology, University College London, London, United Kingdom, WC1E

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–32

https://doi.org/10.1146/((please add article doi))

Keywords

similarity, representation, psychological space, embedding, category learning

Abstract

Similarity and categorization are fundamental processes in human cognition that help complex organisms make sense of the cacophony of information in their environment. These processes are critical for tasks such as recognizing objects, making decisions, and forming memories. In this review chapter, we provide an overview of the current state of knowledge on similarity and psychological spaces, discussing the theories, methods, and empirical findings that have been generated over the years. Although the concept of similarity has important limitations, it plays a key role in cognitive modeling. The review surfaces three key themes. First, similarity and mental representations are merely two sides of the same coin—existing as a similarity-representation duality that jointly defines a psychological space. Second, both the brain's mental representations and the study of mental representations are made possible by exploiting second-order isomorphism. Third, similarity analysis has near-universal applicability across all levels of cognition—providing a common research language.

Contents

1. INTRODUCTION

> Matter is for us not what is primarily given. What is primarily given is, rather, the elements, which, when standing to one another in a certain known relation, are called sensations. (Mach 1914)

Similarity and categorization are fundamental processes in human cognition that help complex organisms make sense of the cacophony of information in their environment. The similarity processes plays a critical role in a variety of tasks, such as recognizing objects, making decisions, and forming memories. In terms of its individual parts, this review is primarily concerned with formal models of similarity. As a whole, this review emphasizes three major themes: the existence of a similarity-representation duality, the investigative beachhead provided by second-order isomorphism, and the applicability of similarity to multiple levels of analysis. These models are broken down into parts that focus on the various data structures used to formalize similarity, how similarity models have been adapted for naturalistic stimuli, and the role of similarity in dynamic cognitive models. This review touches a wide variety of topics and collects evidence from dense specialties, including philosophy, embedding algorithms (e.g., multidimensional scaling) and category learning models. The goal is not to systematically cover these specialties, but to curate a narrative that emphasizes how the notion of similarity is woven throughout research and contributes to diverse research programs.

1.1. Similarity-Representation Duality

On the surface, the notion of psychological similarity is a simple and intuitive idea; similarity evaluates the sameness between two things. For example, two things are more similar the more features that they have in common. The intuitive nature of similarity is supported by the fact that individuals can easily complete tasks based on similarity. If you give an individual a photograph of a basketball, salamander, and a frog and ask them to pick the least similar image—i.e., perform an odd-one-out judgment—most are likely to select the basketball.

However, probe a little deeper and a formal definition of similarity can be challenging to pin down. For example, if individuals are asked to make an odd-one-out judgment for the words “apple”, “penguin”, and “otter”, many are likely to choose based on a fruit-animal distinction, but others may choose based on the icons of major operating systems (MacOS, Linux). While some similarity judgments may have strong consensus, the relative importance of different stimulus features can vary according to context and task (Murphy & Medin 1985). Similarity has repeatedly drawn criticism as a useful construct owing to its extreme flexibility. The philosopher Nelson Goodman succinctly labeled similarity ”invidious, insidious, a pretender, an imposter, a quack” (Goodman 1972, p. 437).

Part of the issue revolves around treating similarity as a unitary and independent concept. The *process* or *function* for outputting a similarity value v requires, at a minimum, two arguments.

$$v = s(\,\cdot\,,\,\cdot\,), \tag{1.}$$

where s is an arbitrary similarity function that operates on arbitrary representations, such as a dot product that operates on vectors. At least two things must be compared—whether they be images, sounds, odors, text—in order to produce a similarity value. The role of some

larger context c, such as an agent's current goals, can be made explicit using a conditional,

$$v = s(\,\cdot\,,\,\cdot\,|c). \tag{2.}$$

Focusing on the functional aspect of similarity, similarity is inextricably linked to the representations similarity operates on. It rarely makes sense to talk about one and not the other. The link between the two is so fundamental, that it is more productive to consider (a) the similarity function and (b) the corresponding representations that serve as operands, as two sides of the same coin—a *similarity-representation duality*. When a similarity judgments are elicited from participants, there are implied mental representations supporting those judgments. But the converse is also true! Given a set of mental representations, there is an implied similarity function for comparing the representations. For example, if one considers representations in early visual cortex, the representations are only *meaningful* with respect to one another. The means of computing relativeness *is* a similarity function. The similarity-representation duality is closely aligned with the views championed by others. For example, (Medin et al. 1993, p. 254) propose that "an important source of constraints derives from the similarity comparison process itself."

The similarity-representation duality is a key theme throughout this review. To promote readability, we refer to a data structure equipped with similarity-representation duality as a *psychological space*. To place the similarity-representation duality and the other themes of this chapter on firmer footing, it is helpful to begin with a broader historical context. We briefly introduce a handful of foundational ideas from philosophical thinkers and early psychological work. The historical progression culminates in a launchpad for the remainder of the chapter. As a side-effect, we will see that contemporary notions of similarity are iterations of older insights.

1.2. The Piers of Similarity

The desire to understand how humans perceive and digest the world has lured thinkers for centuries. Much of this thinking has coalesced around the idea that studying relative differences (i.e., similarities) are tractable, but understanding absolute qualities of the world is problematic. In other words, science may never know if two people perceive the color red in the same way, but it is possible to compare the relative perception of red to other colors.

An understanding of similarity and mental representations has been advanced under various guises across a range of disciplines including philosophy, psychology, neuroscience, and computer science. The broad attraction of this topic is partially due to the fact that perception acts as an inescapable and quirky lens through which all other knowledge is filtered. Analogous to the way that the limitations of a telescope should be understood in order to appropriately process collected light measurements; scientists aim to understand idiosyncrasies of the human mind in order to better understand reality.

The work of ancient philosophers, such as Plato's Republic, Aristotle's Metaphysics, and the Upanishads were some of the earliest to systematically analyze and dissect the nature of perception. One pertinent piece comes from the Allegory of the Cave in Plato's Republic. Inside the cave, prisoners are chained in such a way that they can only look at a wall directly in front of them. On the wall, the prisoners observe moving shadows. The fire creating the light and the objects casting the shadows are unobservable to the prisoners.

> Socrates: So we are! Now, tell me if you suppose it's possible that these captives ever saw

anything of themselves or one another, other than the shadows flitting across the cavern wall before them?
Glaukon: Certainly not, for they are restrained, all their lives, with their heads facing forward only.
...
Socrates: Now, if they could speak, would you say that these captives would imagine that the names they gave to the things they were able to see applied to real things?
Glaukon: It would have to be so.
(Plato, The Republic)

The shadows are the prisoners' reality and the prisoners process that reality accordingly, such as assigning labels and inferring causal relationships. The shadows are a stand-in for the reality that humans can normally perceive and provides an early example the distinguishes between what might be in the world versus what is perceived.

Philosophical debates regarding the distinction between external reality and perceived reality have continued into modern times. Immanuel Kant argued that the only world we can know is the world created by the innate structure of our minds and thus reality "as it is in itself" is unknowable.

What might be said of things in themselves, separated from all relationship to our senses, remains for us absolutely unknown.
(Guyer & Wood 1998)

For Kant, the only reality we know is a cognitively-rendered virtual reality. The philosophical models of Plato and Kant seem to suggest an impasse for studying other's mental representations.

An early researcher to bridge the divide between external reality and the inner psychological reality was Gustav Fechner. Considered by many to be the founder of psychophysics, Fechner approached the problem of ascertaining one's psychological reality by pioneering the use of *relative* judgments. For example, given a change in a particular intensity of light (i.e., lux as measured using a photometer), what is the perceived change according to a human observer? A systematic exploration using relative judgments of this kind led Fechner to posit two related laws: Weber's law (named after his mentor Ernst Heinrich Weber) and Fechner's law. Fechner's law states in order that the intensity of a sensation may increase in arithmetical progression, the stimulus must increase in geometrical progression (Fechner 1860). In other words, Fechner's law posits that psychologically perceived sensation p is proportional to the logarithm of the stimulus intensity s

$$p = c \ln \frac{s}{s_0}, \qquad 3.$$

where s_0 denotes the threshold at which a perceived stimulus becomes zero and c denotes a constant that is determined empirically for a particular stimulus set. Fechner's law is not without its limitations. It is not scale invariant like a power law relationship and has mixed support across different sense modalities. However, the audacity to hypothesize mathematical relationships between objectively measurable differences in the external world and differences in psychological space is a key step in the development of empirical tools for probing the representations of the human mind. Focusing on relative judgments shows

the beginning of a strategy that can be used to circumvent philosophical issues regarding the mapping between external reality and psychological reality.

Heavily influenced by Fechner's work, Ernst Mach also sought to formalize connections between external and inner realities. Mach made many seminal contributions to science, one of which was his work studying the optical illusion that now bears his name: Mach bands. The illusory band is an exaggeration of the contrast between slightly differing color shades due to edge-detection in the visual system. There are two relevant implications from this work. First, information processing begins immediately; the sense organs are not simple conveyors of sensory information but active participants in a process that occurs in successive stages. This insight, combined with others such as lateral inhibition, is one reason why Mach is regarded as a forerunner of the idea of neural nets in perception. Second, to explain the illusion, Mach argued that perception works by focusing on relative differences, not absolute sensory intensity.

> Since every retinal point perceives itself, so to speak, as above or below the average of its neighbors, there results a characteristic type of perception. Whatever is near the mean of the surroundings becomes effaced, whatever is above or below is disproportionately brought into prominence. One could say that the retina schematizes and caricatures. The teleological significance of this process is clear in itself. It is an analog of abstraction and of the formation of concepts.
> (Ratliff 1965)

In Mach's view, relative differences are a primary mechanism that the brain uses during sensation. While experiments with Mach Bands were primarily concerned with relative differences between adjacent patches of the retina, Mach hypothesized that a relative difference operation extended beyond sensation and applied more broadly to all aspects of perception. Nearly a century later, this view would be given strong physiological support from the investigations of cat cortex conducted by Hubel & Wiesel (1959). The relative difference operation was posited as a general principle of cognition, used to assemble increasingly abstract representations of the world.

Relative difference can be computed in many different ways. For example, on could simply compute the difference between the activity of two adjacent photoreceptors $a_i - a_j$, which can also be expressed as the dot product $[a_i, a_j] \cdot [1, -1]$. The second operand ($[1, -1]$) is often referred to as a *kernel*. A dot product of this form can also accommodate an arbitrary number of activations (i.e., a local neighborhood)

$$f(a, k) = a \cdot k, \tag{4.}$$

where a and k are both vectors of the same length. For example, consider the dot product of the kernel $k = [-1, 0, 1]$ and different patterns of neighborhood activation:

1. $[0, 0, 0] \cdot k = 0$
2. $[0, 0, 10] \cdot k = 10$
3. $[0, 10, 10] \cdot k = 10$
4. $[10, 10, 10] \cdot k = 0$
5. $[10, 0, 10] \cdot k = 0$
6. $[0, 10, 0] \cdot k = 0$

This kernel can be viewed as a crude version of a boundary detector, activating when the

incoming activity is not uniform. Stated another way, the kernel is like a template and $f(a, k)$ computes the similarity between the incoming activity a and the template k. The kernel serves as a way of assessing similarity with respect to a particular template or feature. If both input activations match the template, then they are deemed similar. Computing relative differences implies a similarity operation. Computing similarity constitutes a basic step in feature extraction. While this particular operation is simple, kernels form the basis of popular techniques such as convolutional neural networks (CNNs) (Krizhevsky et al. 2017), which enable human-level object recognition in natural images and self-attention (Vaswani et al. 2017), which laid the groundwork for the proliferation of natural language transformers.

Shepard & Chipman (1970) eloquently framed the notion of relative differences in terms of a second-order isomorphism.

> The crucial step consists in accepting that the isomorphism should be sought—not in the first-order relation between (a) an individual object, and (b) its corresponding internal representation—but in the second-order relation between (a) the relations among alternative external objects, and (b) the relations among their corresponding internal representations. Thus, although the internal representation for a square need not itself be square, it should (whatever it is) at least have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of persimmon.
> (Shepard & Chipman 1970, p. 2)

The proposed perspective reiterates Mach's view that it is relative differences—not absolute qualities—that matter. Second-order isomorphism is the basis of popular tools for studying mental representations, such as Representational Similarity Analysis (RSA) (Kriegeskorte et al. 2008). The proposed perspective also brings together the insights of Fechner and Mach, providing a unifying framework for similarity-as-a-research-tool and similarity-as-a-cognitive-mechanism.

Second-order isomorphism is the conceptual pier that stabilizes all theories of similarity. Focusing on second-order isomorphism—rather than first-order isomorphism—allows researchers to rise above the tricky and potentially intractable philosophical quagmire regarding the nature of reality. The price for such a solution is that both the similarity representation and the similarity function must be considered in tandem, otherwise the notion of a psychological space falls apart.

2. PSYCHOLOGICAL SPACES

Since psychological spaces (i.e., mental representations) are not directly observable, they need to be inferred from measured quantities. Using the perception-action cycle as an organizing framework (Von Uexküll 1926, Sperry 1952), one can think of inferring psychological spaces from three qualitatively distinct data sources: stimulus information (e.g., image pixels), neural activity (e.g., fMRI), and behavior (e.g., categorization responses). Data sources that roughly correspond to the beginning, middle, and end of the perception-action cycle, respectively.

An abundance of techniques exist for inferring psychological spaces. The different inference techniques can be broadly categorized based on the characteristics of the data structures they use. Three popular data structures include geometric spaces, set-theoretic spaces, and graph spaces. Each of the three type of data sources (stimulus, neural, behavioral) can be transformed into any one of the three psychological spaces (**Figure 1**). Two additional options add complexity to the possible configurations. First, it is possible to convert from one psychological space to another, for example converting from a graph representation to a geometric representation. Second, one can chain a sequence of transformations, resulting in a different psychological space at each step, such as the neuron activations at successive layers in a deep neural network.

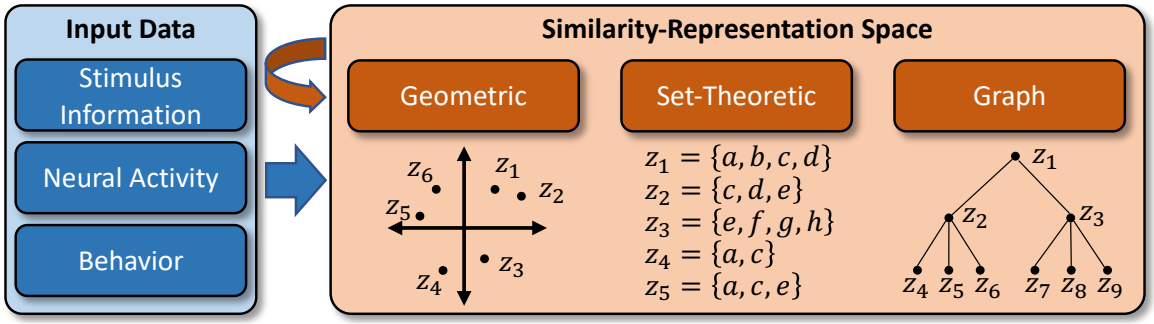

Figure 1

Strategies for inferring psychological spaces. Source input data can be transformed into any one of the three different psychological spaces. One kind of psychological space can also be transformed into a different space.

2.1. Fundamentals

A psychological space is composed of both a similarity function and mental representations. The two aspects are both full-fledged model components—each endowed with the potential to be arbitrarily sophisticated. This means both components may have their own free parameters and exhibit a hierarchical structure. While this fact may be familiar for the representations, it is also true of the similarity function. Similarity functions can be simple and parameter free, as is the case with cosine similarity, or complex and parameterized, as is the case with the generalized exponential (discussed later). In all cases, the choice of a similarity function is a deliberate architectural choice that operates hand-in-glove with the chosen representation.

Techniques for inferring psychological spaces vary, but one generic framing can be achieved by focusing on second-order isomorphism. Shifting the focus to second-order iso-

morphism means that we care about preserving relationships rather than absolute quantities. For example, if people rate an image of a honey badger ($\boldsymbol{x}_i$) as more similar to an image of an American river otter ($\boldsymbol{x}_j$) than an image of an African lion ($\boldsymbol{x}_k$), then the inferred representation should honor those relationships. Loosely speaking, the general inference problem can be framed in the following way:

$$f(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq f(\boldsymbol{x}_i, \boldsymbol{x}_k) \rightarrow s(\boldsymbol{Z}_i, \boldsymbol{Z}_j) \geq s(\boldsymbol{Z}_i, \boldsymbol{Z}_k), \quad 5.$$

where $\boldsymbol{x}$ denotes an initial representation, f is relational measure between the initial representations, $\boldsymbol{Z}$ is the inferred mental representation, and s is a relational measure (e.g., similarity) that operates on the inferred mental representations. Continuing with the example of a badger, otter, and lion; regardless of the psychological space we choose, we expect that the modeled similarity between a honey badger and American river otter ($s(\boldsymbol{Z}_i, \boldsymbol{Z}_j)$) should be greater than the modeled similarity between the honey badger and African lion ($s(\boldsymbol{Z}_i, \boldsymbol{Z}_k)$). The chosen framing also accommodates the case where the function f and representation x are an inferred psychological space instead of measured quantities. Second-order isomorphism focuses on preserving the relationships between things and is less concerned with the absolute values of x and z.

The relational measure f can take many forms. As the example demonstrates, f can be implicitly provided by human participants in the form of similarity ratings or rankings. In a second-order isomorphism paradigm, it is acceptable if the relational measure f is not completely known, so long as the outcomes of the measure can be obtained. The relational measure f can also be explicit, such as a function that outputs the graph geodesic between two nodes on a graph. Alternatively, an explicit measure f may be learned from the data by honing in on statistical regularities, such as a CNN kernel which computes the similarity between an image patch input and a learned template patch.

Given the plethora of potential information to focus on, the relational measure f typically reflects researcher or agent priorities. If people consider a honey badger as more similar to an American river otter, they are prioritizing features like overall size and taxonomic similarity (i.e., both are mustelids). Alternatively, people could have said that African lions and honey badgers are more similar if they placed more emphasis on shared geography. Likewise, researchers may be more interested in relational measures that prioritize neighborhood relationships and deemphasize global relationships; a trick used by many nonlinear dimensionality reduction algorithms such as Locally Linear Embedding (LLE) (Roweis & Saul 2000) and ISOMAP (Balasubramanian & Schwartz 2002, Silva & Tenenbaum 2002, Tenenbaum et al. 2000).

Using an optimization algorithm, the free parameters of a psychological space are found by maximizing goodness of fit (i.e., the loss function) to the observed data. Historically, when referring specifically to the free parameters that correspond to the the representation of stimuli (e.g., coordinates in geometric space) inference algorithms were commonly called "multidimensional scaling" (MDS) or just "scaling" algorithms. Today, some inference algorithms are still called "scaling" algorithms even if they infer a larger set of free parameters beyond the traditional geometric coordinates. In the machine learning literature, analogous inference algorithms are often called *embedding algorithms*. The term *embedding* denotes a higher-dimensional representation that is *embedded* in a lower dimensional space. For that reason, the inferred mental representations of a psychological space could also be called a *psychological embedding*.

The specifics of the optimization algorithm and loss function have evolved over time,

but the general principle remains the same: find a solution that minimizes loss. A general iterative approach to scaling was introduced by Shepard Shepard (1962) and migrated to a standard gradient decent framework by Kruskal Kruskal (1964). As is the case with any machine learning setup, any remaining hyperparameter values—such as the dimensionality of the space—can be determined using an "outer loop" procedure, such as cross-validation.

Inference algorithms leverage the second-order isomorphism trick in order to learn a set of representations where relative differences in inferred psychological space correspond to relative differences in the training data. The details of the algorithm determine which relative differences are emphasized. Some techniques are broadly applicable because they make weak assumptions about the input data, others can only be applied under very specific conditions. In the remainder of this section, we introduce three popular data structures for formalizing mental representations. Along the way, we cover commonly used data and options for similarity function pairings.

2.2. Geometric Representations

Geometric representations, also known as spatial representations, embed mental representations in a multidimensional geometric space. In geometric space, mental representations take the form of multidimensional coordinates $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$, where n indicates the number of mental representations (e.g., stimulus percepts) and d indicates the dimensionality of the space. For improved readability, $\boldsymbol{Z}_i$ is used to denote the ith row vector (e.g., the ith stimulus embedding). Generally speaking, coordinates that are located close together have high similarity while coordinates located far apart have low similarity. Geometric representations have a long history in science, as Roger Shepard pointed out:

> Proposals that stimuli be modeled by points in a space in such a way that perceived similarity is represented by spatial proximity go back to the suggestions of Isaac Newton (3) that spectral hues be represented on a circle, of Helmholtz and Schrödinger (4) that colors in general be represented in a curved Riemannian manifold, of Drobisch (5) that pure tones be represented on a helix, and of Henning (6) that odors and tastes be represented within a prism and a tetrahedron, respectively.
> (Shepard 1980, p. 390)

Early work with geometric spaces used behavioral data to infer psychological spaces, often focusing on identification, categorization and recognition tasks. When modeling identification and categorization performance, the entire process can be framed as a mapping from stimulus representation $\boldsymbol{X}$ (e.g., image pixels) to a categorical outcome y. Given stimulus i, the probability of responding identity/category j is denoted $p(y_j|\boldsymbol{X}_i, \Theta)$, where Θ is a catch-all for any model parameters. Psychologists are particularly interested in the intermediate mental representations of the perception-action cycle, $\boldsymbol{Z}_i = f(\boldsymbol{X}_i, \Theta^{(0)})$, where f denotes some arbitrary transformation of the stimulus, such as a series of convolutional layers. One can think of the transformation function f as a *perception module*.

Early work used a special perception module f that is effectively a lookup table. Instead of using information in the stimulus (e.g., pixel values), all stimulus information is discarded except a stimulus identifier (e.g., an integer that serves as index). The stimulus identifier i is used to look up the appropriate mental representation $\boldsymbol{Z}_i$. In other words, the complete lookup table corresponds to the matrix of values $\boldsymbol{Z}$. In deep learning packages

such as TensorFlow and PyTorch, lookup tables of this sort are called embedding layers, which should not be confused with the broader scope of applications covered by embedding algorithms.

Using a lookup table as a perception module has two advantages. First, allows the inferred mental representations to be almost completely constrained by behavioral data, simplifying the research problem. If using a perception module that performs more sophisticated processing, different architectural choices can impact the inferred psychological representation because they are constrained by both behavior and stimulus data. Second—and more relevant from a historical perspective—a lookup table bypasses the need to implement of a functional perception module, which was intractable for naturalistic stimuli until recently.

In additional to a perception module, one must specify a model component that maps from a psychological space to observed behavior, what could be called a *behavior module*. A popular behavior module comes from the similarity choice model (SCM) (Shepard 1957, Luce 1963), which is sometimes referred to as Luce's ratio of strengths formulation (Luce 1959). The heart of SCM is the response rule

$$P(y_j|i, \boldsymbol{Z}, \boldsymbol{b}) = \frac{b_j s(\boldsymbol{Z}_i, \boldsymbol{Z}_j)}{\sum_{k=1}^{N} b_k s(\boldsymbol{Z}_i, \boldsymbol{Z}_k)}, \quad 6.$$

where y_j denotes a response choosing stimulus j and $\boldsymbol{b}_j$ ($0 \leq \boldsymbol{b}_j \leq 1$, $\sum_k \boldsymbol{b}_k = 1$) is a corresponding response bias.

When originally introduced, SCM did not use $\boldsymbol{Z}$ or specify a similarity function s. Instead the similarity values $s(z_i, z_j)$ were used directly and formulated as free parameters η_{ij} such that $\eta_{ij} \geq 0$ and $\eta_{ij} = \eta_{ji}$. As originally formulated, the free parameters included one bias parameter per possible response (i.e., stimulus) and one "similarity" parameter for each unique stimulus-pair. There are two notable downsides with the original formulation. First, the number of free parameters scales poorly as the number of stimuli increases, growing like $\mathcal{O}(n^2)$. Second, there is no explicit psychological space, although distances are implied by the similarity parameters learned for each unique stimulus-pair.

The complete fusing of a lookup table (perception module) and SCM (behavior module) is demonstrated by the MDS-choice model (Shepard 1957, 1958). An MDS-choice model explicitly infers a psychological space by assuming similarity is functionally related to psychological distance via an exponential-family kernel:

$$s\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j\right) = \exp\left(-\beta||\boldsymbol{Z}_i - \boldsymbol{Z}_j||_\rho^\tau\right), \quad 7.$$

where β, ρ, and τ control the gradient of generalization. The steepness at which similarity decays is largely controlled by β. The most common settings of τ result in a Laplace kernel ($\tau = 1$) and a Gaussian kernel ($\tau = 2$). The most common settings for the Minkowski distance parameter ρ result in Manhattan distance ($\rho = 1$) and Euclidean distance ($\rho = 2$).

In early work that used the SCM, it was unclear whether the exponential family similarity function should take the form of a Laplace or Gaussian distribution (Ashby & Lee 1991, Nosofsky 1988, Shepard 1988). After systematic investigation, it appears that a Laplace distribution best describes the perceptual encoding of a single event, but perceptual and memory noise may result in a stimulus being encoded in slightly different locations in psychological space (Ennis 1988, Ennis et al. 1988). Depending on the relative level of perceptual and memory noise, the effective similarity function will appear to be a Laplace (low relative noise) or Gaussian (high relative noise).

With the pieces in place, inference research boomed. There were multiple breakthroughs in developing inference algorithms with different constraints. For example, embedding with unknown distance (Shepard 1962), nonmetric similarity functions (Kruskal 1964), and asymmetric similarity functions (Krumhansl 1978). Inference algorithms were also extended to accommodate individual differences (Carroll & Chang 1970, Carroll & Wish 1974). Eventually, inference research ran into some problems. First, hardware constraints made it computationally expensive to determine solutions for problems involving more than a few stimuli. Second, the lack of software packages presented a high barrier to entry since models had to be coded from scratch. Third, collecting behavioral data is expensive. All of these factors confined research to relatively small and simplistic stimulus sets. Future hardware and software advances, which we return to in a later section, would blow open the door on studying large-scale, naturalistic stimulus sets. As a preview, one major advancement has been the generalization of geometric spaces to non-Euclidean geometries.

2.3. Set-Theoretic Representations

One criticism of geometric approaches is that they are overly restrictive. A more general in flexible data structure may be needed. One intuitive way to think about similarity is in terms of a feature matching or set-theoretic process. A feature matching process assumes that an item i can be represented as a set of features $\mathcal{Z}_i$. For example, given the concept of bumble bee and butterfly one could list as many features as possible about each of them. These could include attributes like "has wings", "has stinger", "can fly", "is insect", and so on. Continuous features, like specific colors, can be treated as exact matches or the color space could be discretized. The similarity value between two items is then some function of the common and distinctive features. For example, both a butterfly and bumble bee are insects, have wings, and can fly, but only a bumblebee has a stinger.

The feature matching approach was placed on solid theoretical ground by Tversky (1977) in the seminal work "Features of Similarity". Given some stimulus i with features $\mathcal{Z}_i$ and some stimulus j with features $\mathcal{Z}_j$, the similarity value of the two items takes the following general form:

$$s\left(\mathcal{Z}_i, \mathcal{Z}_j\right) = g\left(\mathcal{Z}_i \cap \mathcal{Z}_j, \mathcal{Z}_i - \mathcal{Z}_j, \mathcal{Z}_i - \mathcal{Z}_j\right). \tag{8.}$$

The first operand in the function determines the features common to both items, the second operand determines the set of features unique to item i, and the third operand determines the set of features unique to item j. Various forms of g are possible and Tversky highlighted two of them.

The contrastive model of similarity assumes

$$s\left(\mathcal{Z}_i, \mathcal{Z}_j\right) = f\left(\mathcal{Z}_i \cap \mathcal{Z}_j\right) - \alpha f\left(\mathcal{Z}_i - \mathcal{Z}_j\right) - \beta f\left(\mathcal{Z}_j - \mathcal{Z}_i\right), \tag{9.}$$

where θ, α and β are nonnegative free parameters that determine the relative importance of each term. The function f maps features to an interval scale in a way that takes into account factors that contribute to psychological salience, such as intensity, frequency, familiarity and informational content. The free parameters allow for a number of possibilities, which differentially weights the contribution of shared and distinctive features to the computed similarity value. For example, if $\theta = 1$ and $\alpha = \beta = 0$, then the similarity of the items is entirely determined by the common features. Alternatively, if $\theta = 0$ and $\alpha = \beta = 1$, then you get a similarity function that is entirely determined by distinct features and loosely

resembles a generalized version of Hamming distance. If $\alpha > \beta$, the similarity function more heavily weights the set of features unique to item i.

A second matching function, the ratio model, is partially motivated by the desire to create a similarity function with the codomain $[0, 1]$. The ratio model assumes

$$s\left(\mathcal{Z}_i, \mathcal{Z}_j\right) = \frac{f\left(\mathcal{Z}_i \cap \mathcal{Z}_j\right)}{f\left(\mathcal{Z}_i \cap \mathcal{Z}_j\right) + \alpha f\left(\mathcal{Z}_i - \mathcal{Z}_j\right) + \beta f\left(\mathcal{Z}_j - \mathcal{Z}_i\right)}, \qquad 10.$$

where α and β are nonnegative free parameters that play an analogous role to α and β in the contrastive model of similarity. The ratio model encompasses a number of historical feature-matching approaches as special cases (Tversky 1977).

Both the contrastive model and ratio model are capable of operating on a wide range of representations. The set of features can include an arbitrary mixture of categorical and continuous features. Historically, the stimuli features were hand-coded.

2.4. Graph Representations

A final representation space is realized through graphs or networks. Set-theoretic data structures may be too flexible for some applications. Networks provide an alternative that straddles the mathematical maturity of geometric structures and the flexibility set-theoretic structures. A graph Z is composed of a set of n vertices (or nodes) V and a set of edges $E \subseteq V \prod V$ The nodes represent concepts, and the edges make explicit the relationships between concepts. In general, two nodes are considered similar if the number of edges separating them are low. When using graphs to model mental representations, the set of graphs is typically restricted to those without any closed loops, in which case the focus is on *additive* or *path-length* trees. The distance, and therefore the similarity, between two nodes is given by the sum of the lengths of the edges between them. A variety of algorithms were introduced in the late 1970's for inferring additive trees from behavioral data (Carroll 1976, Cunningham 1978, Sattath & Tversky 1977).

Cluster representations can be thought of as a special case corresponding to a disconnected graph. Given a set of clusters, the nodes of each cluster are disconnected from the nodes of every other cluster. If there is no notion of within-cluster similarity, similarity is simply a binary output that yields 1 if a path exists between two nodes and 0 otherwise.

2.5. Relative Strengths and Weaknesses

Different psychological spaces exhibit various strengths and weaknesses. There is not necessarily one *true* representation. As stated by Shepard,

> It would be a mistake to ask which of these various scaling, tree-fitting, or clustering methods is based on the correct model. As even my small sample of illustrative applications indicates, different models may be more appropriate for different sets of stimuli or types of data. Even for the same set of data, moreover, different methods of analysis may be better suited to bringing out different, but equally informative aspects of the underlying structure.
> (Shepard 1980, p. 397)

Two aspects worth calling out are adherence to metric axioms and the capability to capture hierarchical relationships.

2.5.1. Metric Axioms. Tversky noted that when eliciting human judgments of similarity, people would often provide ratings that violated the standard axioms of metric space. Three key axioms are

- Minimality: $\delta(z_i, z_j) \geq \delta(z_i, z_i) = 0$.
- Symmetry: $\delta(z_i, z_j) = \delta(z_j, z_i)$.
- The triangle inequality: $\delta(z_i, z_j) + \delta(z_j, z_k) \geq \delta(z_i, z_k)$.

It is not unusual for individuals to provide similarity judgments that exhibit asymmetry. For example, people might say that apes are like humans, but they might not say that humans are like apes. The violation of these axioms is technically problematic for geometric representations. Standard geometric representations are metric spaces and can only generate solutions that obey the metric axioms. If people regularly violate these axioms, then models of similarity that assume symmetry, equal self-similarity and triangle inequality may be a poor modeling choice. In fact, a major reason Tversky took the time to articulate a feature matching framework was to create a representation framework that allowed for violations of metric axioms.

2.5.2. Hierarchical Relations. In addition to potential over-adherance to metric axioms, Euclidean spaces also have difficulty accommodating hierarchical similarity relations. The difficulty can be shown a number of ways, but one approach is by analyzing the maximum number of items that can share the same nearest neighbor (Tversky & Hutchinson 1986). To intuit the result, imagine that you have one blue sphere and a infinite set of red spheres. You start by placing blue sphere on the table and then incrementally glue red spheres to the blue sphere. Eventually you will get to a point where you can no longer attach any additional red spheres. This is a problem because the blue sphere may represent an abstract category like "mammal" and the red spheres subcategories like "felines" and "canines". If the data you are trying to embed has more subcategories than will fit around the primary category, then the embedding will fail to adequately capture the hierarchical relationship.

One way to address the hierarchical limitation of geometric spaces is to expand geometric representations beyond Euclidean geometry. One such extension is to use a hyperbolic space which is characterized by negative curvature. A special case of hyperbolic geometry is a Poincaré space. To intuit the properties of a Poincaré space, we can consider the two-dimension case: a Poincaré disk. If one starts at the origin of Poincaré disk and moves towards the edge of the disk, the distance traveled tends to infinity. You can imagine that a Poincaré disk is like a two-dimensional projection of a very deep bowl. As one moves towards the edge of the bowl they also have to move up. Poincaré spaces serve as accommodating representations for embedding hierarchical data. First, the root node of the hierarchical data is placed at the origin. For each successive level, the nodes are place farther from the origin. Since the available space effectively expands as you move away from the origin, Poincaré representations do not suffer from same problem as Euclidean spaces. Such representations have been used to embed knowledge graphs (Wang et al. 2014) and have more recently been integrated into various deep neural network architectures (Peng et al. 2022).

2.6. Potential versus Actual Duality

Before diving into the different psychological spaces, it is import to clarify an aspect of the similarity-representation duality. When architecting a cognitive model, the chosen representation space will only partially constrain the set of compatible similarity functions. Given the previous argument that the similarity-representation duality asserts a tight link between similarity and representations, one may be wondering why there is so much flexibility in selecting a similarity function given a representation space. An essential point to realize is that the link between a similarity function and a representation is *actualized* when a set of representations are learned—either artificially via an algorithm or naturally in the brain. Prior to learning representations, there is only a *potential* link.

Consider the following example: given a set of triplet similarity judgments we would like to use an embedding algorithm to infer the mental stimulus representations of the human participants. Based on other research goals, a researcher might select an $L2$ distance (instead of an alternative Lp distance) to serve as the core of an exponential similarity function. After making this architectural choice, an algorithm is used to learn the most likely set of mental representations given the observed behavior.

After training the model, it would be inappropriate to swap out the L2 distance with an L1 distance. The process of training the model has inextricably linked the representation with the similarity function. In the same vein, it would be inappropriate to substitute the exponential similarity function with a cosine similarity function. Of course, some substitutions will "work"—in the sense that they operate over the same functional domain of operands—but they will be theoretically wrong. Some substitutions may even output highly correlated similarity values. Unless there is a formal justification for a substitution, substitutions create a theoretical mismatch, breaking the similarity-representation link.

Before training, there is a potential link between similarity and representations. After training, the similarity-representation duality link is actualized. Prior to training, there is flexibility in choosing the architectural ingredients—including selecting similarity functions that have free parameters. After training, the architectural choices are baked in and the correct way to read out a similarity value between two items is using the same similarity function that was used to train the space.

3. NATURALISTIC DATA SOURCES

The introduction of psychological spaces was accompanied by relatively simple and well-controlled stimuli. The focus of this section shifts from historical motivators of each representation space to techniques that scale to real-world, naturalistic paradigms. Modern techniques, in particular those driven by machine learning, have unlocked the ability to extract psychological spaces for real-world research paradigms: such as those using naturalistic images. Significant advancements have been made for sourcing data from all parts of the perception-action cycle; from deep neural networks that process naturalistic stimuli to active learning approaches that help scale behavior collection to tens of thousands of stimuli.

3.1. Stimulus Information

A decade ago, options were limited for inferring psychological representations from naturalistic stimuli. The work-around was to use simplistic stimuli and hand-coded feature dimensions. Today, an enormous number of methods exist for extracting rich feature spaces, such as convolutional neural networks (CNNs) and transformers. While enormous progress has been made for all stimulus modalities, two modalities have received particular attention: visual stimuli and natural language text. Image and natural language modules have become sufficiently capable, that a vibrant subfield has emerged to study the correspondence between machine- and human-learned representations.

3.1.1. Natural Images. Algorithms and models can be used to map image pixels to rich psychological spaces. Approaches vary in complexity, from relatively simple domain-general approaches like PCA and ICA (Comon 1994, Bell & Sejnowski 1995), to highly-parameterized domain-specific approaches like CNNs pre-trained on image datasets. The field has seen a general shift from hand-engineered components to data-driven features. Early efforts focused on approaches which can be collectively be called local feature integration. These efforts included scale invariant feature transform (SIFT) (Lowe 1999), histogram of oriented gradients (HoG) (McConnell 1986, Dalal & Triggs 2005), and HMAX (Riesenhuber & Poggio 1999). Collectively, these approaches converged around the general-purpose abilities of convolution layers.

A dominate approach today is the use of stacked convolutional layers in CNNs. Interestingly, the foundation for modern CNNs was established before many alternatives to local feature integration. In 1979, the Neocognitron brought together many of the essential features of modern CNNs (Fukushima 1980). Neocognitron uses a hierarchical, multilayer design that leverages multiple pooling and convolution layers and was inspired by the model proposed by Hubel & Wiesel (1959). Fully formed CNNs burst onto the scene following the success of AlexNet (Krizhevsky et al. 2017). AlexNet is a 1000-category-output model trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). While AlexNet performed admirably well on the ILSVRC task, performance has since increased, leaving a trail of popular models, such as VGG (Simonyan & Zisserman 2015). With minor modifications, such as removing the last layer, most CNN image models can be treated as an encoder or perception module—a function that maps images to a latent representation. Consequently these models can treated as a modular building block and incorporated into a diverse set of cognitive models.

With categorization performance rivaling or exceeding human abilities, focus has shifted

to understanding the differences between machine- and human-learned representations. As part of this effort, a new class of CNNs—such as CorNet (Schrimpf et al. 2020)—have been developed with the key objective of understanding human-learned representations. One of the aim of these models is to provide a perceptual module that yields grounded psychological spaces. Understanding the differences between machine- and human-learned representations requires grounding in another data source, such as neural or behavioral data.

3.1.2. Natural Language Text. Like images, natural language text has seen a meteoric rise in human-like capabilities. Modern efforts got off the ground by leveraging the distribution hypothesis, which assumes that words with similar meaning will occur in similar contexts, and inferring a multidimensional representation that places similar meaning words close together. An early example of this can be seen in latent semantic analysis (LSA) (Dumais et al. 1988), which used "documents" to define a coarse-grained notion of context. More recent research has refined the operational definition of context and is more nuanced. Word embeddings approaches like word2vec (Mikolov et al. 2013) and GloVe (Pennington et al. 2014), define context as a weighted window around every word. Depending on the particular algorithm, similarity between embedding coordinates can be computed using something like cosine similarity or Euclidean distance.

However the meaning of words is not independent of their context, which motivated the introduction of sentence embeddings, which embed entire sentences into a multidimensional space. Sentence embedding models, such as BERT (Devlin et al. 2019), leverage transformers as an extremely flexible mechanism for learning relevant context. Near complete ingestion of all available high-quality natural language text and architectural improvements have culminated in large language models (LLMs) that exhibit an impressive ability to generate well-formed natural language outputs from natural language inputs, what one could think of as "dialogue autocomplete". Some of the most well known LLMs come from OpenAI's family of Generative Pre-trained Transformer models, commonly referred to as GPT (Brown et al. 2020, AI 2023). The encoder component of LLM can be treated as a perceptual model, but mirroring the development of image models, the focus has begun to shift to comparing human- and machine-learned representations.

3.2. Neural Activity

Situated between stimulus and behavior, neural data provides a powerful window into psychological spaces. The universe of techniques for analyzing neural data warrants multiple reviews and cannot be fully covered here. Instead, we briefly cover a few techniques that demonstrate the variety of ways that psychological spaces can be extracted from neural data.

Different recording methodologies (e.g., EEG, MEG, fMRI, microelectrodes, diffusion imaging) for generating neural data can all be used to infer psychological spaces. Off-the-shelf linear dimensionality reduction approaches like PCA and factor analysis can be applied regardless of recording methodology and are particularly common for population responses (Cunningham & Yu 2014). More recently, off-the-shelf nonlinear dimensionality reduction methods have also become popular, such as LLE, ISOMAP, t-SNE (van der Maaten & Hinton 2008), and autoencoder neural networks (Hinton & Zemel 1993, Kingma & Welling 2013). Given the necessity of processing neural data in a consistent way, entire processing

pipelines have also emerged, such as fMRIPrep (Esteban et al. 2019). Collectively, these techniques have revealed characteristics about the psychological space for different functions and regions of the brain. For example, the geometry of visual cortex (Stringer et al. 2019, Guidolin et al. 2022), touch (Nogueira et al. 2023), control (Badre et al. 2021, Gallego et al. 2017), and abstraction in the hippocampus and prefrontal cortex (Bernardi et al. 2020). One can also test which similarity function the brain uses in its internal computation and whether similarity function varies across brain regions and tasks (Bobadilla-Suarez et al. 2020).

3.3. Behavior

Working backwards from observed behavior, inference procedures can uncover the most likely psychological space. Behavior-based embeddings have grown in three key ways: the use of naturalistic stimuli, the diversity of measures, and the scaling up of collection strategies. Relatively simple stimuli have been largely replaced by high-fidelity naturalistic stimuli. The diversity of behavioral measures have proliferated since the seminal psychophysics work of Weber and Fechner. Increasingly accessible tools for creating custom websites has created the ability to design a seemingly limitless number of behavioral paradigms. The integration of modern machine learning approaches, combined with web-based data collection, has enabled behavioral data collection at an unprecendented scale, opening up avenues to new research questions.

While innovative paradigms continue to flourish and push the boundary, a handful of behavioral measures have maintained their status as key players in behavior-based embeddings. These include continuous measures such as similarity ratings, response times. Categorical measures include identification responses, classification responses, same-different judgments, triplet similarity judgments, odd-one-out similarity judgments, and pile sorting. Ordinal measures include generalized similarity rankings.

Behavioral data, like similarity judgments, has been collected for an increasingly diverse set of stimuli. Similarity judgments have been collected for naturalistic images including everyday objects (Hebart et al. 2020, Roads & Love 2021), food (Wilber et al. 2014), birds (Roads & Mozer 2021), rocks (Nosofsky et al. 2018), skin lesions (Roads et al. 2018) images of the reachable world (Josephs et al. 2023). Similarity judgments have also been collected for other modalities, such as odours (Nakayama et al. 2022). Beyond similarity judgments, psychological spaces can also be derived from real-world activity patterns, such as consumer shopping behavior (Hornsby et al. 2020, Hornsby & Love 2022).

In addition to the adoption of naturalistic stimuli, there has also been growth in more naturalistic scale. The real-world is composed of countless concepts. Even if we restrict ourselves to categories encountered on a daily basis, there are thousands of categories (e.g., food, travel, entertainment, work), each with thousands of nonfungible exemplars (my golden retriever versus your golden retriever). Modern approaches have embraced this diversity by developing data collection approaches that scale to a large number of stimuli. Advancements have been made on two fronts: designing tasks that take advantage of human abilities and selecting trial content that maximizes expected information gain.

On the first point for scaling up, the ideal task with heavily depend on the domain. For example, if collecting similarity judgments about odors, a researcher is largely constrained to sequential pair-wise ratings. However, if collecting similarity judgments about images, a researcher can consider a larger set of options since multiple images can be displayed

at the same time. In the case of images, this allows for a generalization of the standard triplet. In a standard triplet task, a participant is given a query image and two reference images. The participant must select the reference image they think is most similar to the query image. This task can be generalized such that a participant is shown a query image and $|\mathcal{R}|$ reference images (Roads & Mozer 2019, Wah et al. 2014, Wilber et al. 2014). The participant must then make S selections that indicate the subset of references that are most similar to the query. These selections can be ranked (Roads & Mozer 2019) or unranked (Wah et al. 2014, Wilber et al. 2014). The advantage of a generalized trial is it exploits a human's ability to quickly process a visual scene. Participants can provide more information about how they perceive similarity in a shorter amount of time (Roads & Mozer 2019, Wilber et al. 2014). This respects the participant's time and reduces data collection costs.

Beyond intelligent task design, the next advancement for scaling up is intelligent content selection. When inferring a psychological space, it is not strictly necessary that all pairwise relations be directly probed. Instead, the collection budget can be allocated where it is needed most. In an active learning paradigm, a three-step iterative procedure is used to allocate the data collection budget as economically as possible (Jamieson et al. 2015, Rau et al. 2016, Sievert et al. 2017, Tamuz et al. 2011, Roads & Mozer 2019, Roads & Love 2021). First, the currently collected data is used to estimate uncertainty associated with the psychological space by computing a posterior distribution. Second, the posterior distribution is used to identify the next batch of trials that are likely to maximize information gain. Third, those selected trials are shown to participants and the process is repeated with the expanded set of behavioral data. The benefit of active learning depends on the application, with some cases showing no benefit (Jamieson et al. 2015) and other cases showing a benefit (Roads & Mozer 2019). In the best case scenario, active learning opens up the possibility of scaling up to larger stimulus sets.

3.4. Comparing Psychological Spaces

With all of these different psychological spaces, it is important to consider the appropriate way to compare them. Given the three different sources of data, there are three purist psychological spaces: stimulus-based, neural-based, and behavior-based. The different sources of data can also be combined to produce hybrid psychological spaces. For example, stimulus information can be combined with human behavior to jointly constrain a stimulus-behavior-based psychological space. All of these options mean multiple types of comparisons are possible, such as stimulus- versus stimulus-based (Kornblith et al. 2019), neural- versus neural-based (Sexton & Love 2022), stimulus- versus neural-based (Kubilius et al. 2019, Sexton & Love 2022), stimulus-behavior- to neural-based (Mack et al. 2016, 2020). These distinctions can be further broken down by participant species, such as stimulus- versus human-neural-based (Jacob et al. 2021) and stimulus- versus macaque-neural-based (Rajalingham et al. 2018). Multiple techniques exist for comparing psychological spaces, but the optimal method for comparing representations is an ongoing research problem.

Numerous techniques exist and each has limitations. Popular techniques for comparing representations include representational similarity analysis (RSA) (Laakso & Cottrell 2000, Kriegeskorte et al. 2008) and canonical correlation analysis (CCA) (Hotelling 1936). Briefly, RSA is a method for comparing two representations that assesses the correlation between the implied pairwise similarity matrices. CCA is a method that compares two rep-

resentations by finding a pair of latent variables (one for each domain) that are maximally correlated. While RSA and CCA remain popular, their limitations have spurred researchers to develop more robust methods. Adjustments to RSA include the similarity metric Centered Kernel Alignment (also known as the RV coefficient) (Cristianini et al. 2006, Cortes et al. 2012, Kornblith et al. 2019), unbiased CKA (Nienborg et al. 2019), feature-reweighted RSA (Kaniuth & Hebart 2022), and extensions for handling noise (Storrs et al. 2021). Numerous CCA variants have been introduced to make the approach more widely applicable, such as probabilistic CCA (Bach & Jordan 2005, Klami et al. 2013), kernel CCA (Hardoon et al. 2004), deep CCA (Andrew et al. 2013), sparse CCA (Witten & Tibshirani 2009), and projection weighted CCA (Morcos et al. 2018). Another family of comparison techniques is pattern component modeling (Diedrichsen et al. 2018).

4. DYNAMIC PSYCHOLOGICAL SPACES

Categorization is a foundational task for agents interacting in the world. As agents go about their lives, they must determine if a given stimulus supports their goals, loosely grouping stimuli into categories such as safe/dangerous, nutritious/toxic, prosocial/antisocial, fact/misinformation, entertaining/boring. While many categories will be fuzzy, one can think of categories as being composed of things with similar consequences, what Shepard refers to as a consequential region in psychological space.

> An object that is significant for an individual's survival and reproduction is never sui generis; it is always a member of a particular class—what philosophers term a "natural kind." Such a class corresponds to some region in the individual's psychological space, which I call a consequential region.
> (Shepard 1987, p. 1319)

The ability to categorize the world provides a framework for executing the appropriate behavior (e.g., toxic → avoid) and equally as important, for generalizing knowledge beyond seen stimuli to novel stimuli.

Category learning provides an interesting testbed for psychological spaces because it surfaces the issue of evolving knowledge and context-dependent behavior. As people move about the world, they are exposed to new experiences and ideally acquire new relevant knowledge and forget other less-useful knowledge. A psychological space needs to accommodate different experiences if it is to fully capture human learning. For example, expert fishermen judge similarities among fish on both functional and morphological criteria while novices judge on morphological criteria alone (Boster & Johnson 1989). Differences between expert and novice perceived similarity is also present when judging trees (Srinivasan Shipman & Boster 2008) and musical pitch (Shepard 1982). In addition to incorporating new experiences, people can also tune their attention to focus on a subset of features in a context-dependent manner. For example, focusing on the irregular outline of a skin lesion to assess if it is malignant. A comprehensive model of psychological space needs to support both of these dynamics.

Early work explored the ability of SCM-like models to predict human categorization behavior (Shepard et al. 1961). Using the now seminal, six category types task, Shepard et al. Shepard et al. (1961) attributed failure of exemplar models to an intervening selective-attention process. These results highlight how psychological spaces are limited in isolation.

A classic human categorization model is the generalized context model (GCM) (Nosofsky 1984, 1986). The model uses a geometric representation of individual stimuli and a summed-similarity behavior module. GCM's summed similarity approach extends Luce's ratio of strengths rule to categories that have many members (i.e., exemplars). The first step defines the aggregate evidence that stimulus i belongs to category j,

$$g(i, j|\boldsymbol{Z}, \boldsymbol{m}) = \sum_{k|y_k=y_i} m_k s(\boldsymbol{Z}_i, \boldsymbol{Z}_k), \qquad 11.$$

where $\boldsymbol{m}_k$ indicates the strength that exemplar k is stored in memory. The second step places the aggregate evidence within a ratio of strengths template:

$$P(y_j|i, \boldsymbol{Z}, \boldsymbol{b}, \boldsymbol{m}) = \frac{b_j g(i, j|\boldsymbol{Z}, \boldsymbol{m})}{\sum_{l \in y} b_l g(i, l|\boldsymbol{Z}, \boldsymbol{m})}, \qquad 12.$$

where b_j is the response bias for category j. The probability of making a particular categorization response is proportional to the aggregate generalization evidence associated with a category.

Perhaps the most crucial aspect of this model is contained within the similarity function s. Instead of using standard Minkowksi distance, the similarity function uses a weighted Minkowski distance as a way to formalize the notion of selective attention:

$$s\left(\boldsymbol{Z}_i, \boldsymbol{Z}_j\right) = \exp\left(-\beta \left\|\boldsymbol{Z}_i - \boldsymbol{Z}_j\right\|_{\rho,\boldsymbol{w}}^{\tau}\right), \tag{13.}$$

where the weighted Minkowksi distance is

$$\left\|\boldsymbol{Z}_i - \boldsymbol{Z}_j\right\|_{\rho,\boldsymbol{w}} = \left(\sum_{d=1}^{D} w_d |Z_{i,d} - Z_{j,d}|^{\rho}\right)^{\frac{1}{\rho}}. \tag{14.}$$

The weights are constrained such that $w_j \geq 0$ and $\sum_d w_d = 1$. The other parameters (β, τ, ρ) have the same interpretation as in SCM. The choice of formalism reflects an adoption of Individual Differences Scaling (Carroll & Wish 1974) for modeling selective attention.

The weighted Minkowksi distance provides a mechanism for stretching and contracting psychological space. Consider a two-dimensional psychological space with four stimuli arranged in a rectangle where the sides of the rectangle are parallel to the axes of the space. Decreasing the attention weights has the effect of bringing all the stimuli closer together in psychological space. If only one weight is changed, opposite sides of the rectangle can be pushed closer or farther apart relative to the other two sides. Manipulations of this form can be used to model changes such as increased discriminability between categories. However, one limitation is that it is not possible to bring stimuli at opposite corners closer together (or father apart) while holding the distance of the other two stimuli constant. This is an appropriate constraint if the psychological dimensions are separable (Nosofsky 1992), but problematic if the psychological dimensions are integral.

As originally formulated, GCM does not specify a mechanism for how the attention weights should change with experience. Instead, the attention weights are treated as free parameters and fit to contrasting sets of behavioral data, such as identification and categorization behavior. Spurred by the success of GCM, other models of category learning have proposed mechanistic update rules for how representations change with experience. CGM and the weighted Minkowski distance has stimulated the development of many additional models of category learning. For example, ALCOVE (Kruschke 1992) proposed rules for updating attention weights based on the errors an agent experiences while performing a task. Similarity can also change on a trial-by-trial basis where features come online one at a time (Lamberts 2000) or based on a sampling process (Braunlich & Love 2022).

Both ALCOVE and GCM use a summed similarity approach that assumes each stimulus is represented independently in psychological space. This assumption defines a class of models called *exemplar models*. Exemplar models are often contrasted with *prototype models*, which explicitly represent each category—instead of each exemplar—in psychological space.

In between the two extremes of exemplar and prototypes, category learning can also be formalized as a rational, cluster-recruitment (Love & Medin 1998, Love et al. 2004), or nonparametric Bayesian model (Navarro & Griffiths 2008) where model complexity is adjusted based on the agent's experience. Such a model begins with a representation with zero or minimal instances in psychological space. As surprising stimuli are encountered, the

representation grows to accommodate the new experiences. The new representations can correspond to a single exemplar or represent a set of exemplars (i.e., a prototype) depending on which representation is more economical. The advantage of nonparametric approaches is that it provides a unifying paradigm for exemplar and prototype models. Past empirical work suggests that some tasks are characterized by a shift from a prototype-based mode to an exemplar-based mode of representation (Smith & Minda 1998). Recent research suggest that the hippocampus works this way for human learning, going from episodes to semantic clusters (Mack et al. 2018). This mirrors the shift that occurs in nonparametric models when modeling nontrivial categorization tasks. While there is general agreement that the information used to support categorization can be modified with experience, there is less agreement on the characterization of this shift (Johansen & Palmeri 2002).

5. GENERAL DISCUSSION

Psychology inherited the tricky philosophical problem of understanding the nature of perceived reality. Modern psychology builds on the relativistic framework of philosophy, accepting that humans cannot know reality in an absolute sense. Focusing on relative comparisons, or similarity, is more than a clever philosophical work-around. Over a hundred years of research, from Fechner's early psychophysics work to modern deep neural networks, have made it clear that similarity is a common currency of perception and cognition. In addition to operating at all levels of cognition, similarity—or more accurately, the second-order isomorphism defined by a set of similarity relations—has been a powerful tool for analyzing and comparing psychological spaces.

It is important to recognize that a computed psychological similarity *value* is dependent on both a psychological similarity *function* and a psychological *representation*. A given similarity function and psychological representation exist in a balance that produces meaningful relationships in a psychological space (e.g., the color red is more similar to orange than to blue). Unilaterally altering one aspect of the similarity-representation duality risks distorting—and potentially destroying—the meaning of the psychological space.

In principle, a psychological space can be formalized using any data structure that permits a similarity computation. In practice, three data structures dominate the literature: geometric, set-theoretic, and network structures. All data structures—along with their permissible similarity functions—exhibit strengths and weaknesses, such as computational efficiency and interpretability. The affordances of different psychological spaces are relevant for the human-led design of cognitive models and nature-led design of brains.

Psychological spaces can be constrained by a wide variety of data. The perception-action cycle provides one way to think about the different sources of data: stimulus information, neural activity, and behavior. Classical approaches to inferring psychological spaces focused on using behavior (e.g., MDS). Modern approaches often take advantage of more that one source of data and leverage recent advances in machine learning (e.g., DNNs). A coherent integration of these pieces will be critical for scaling up the size and scope of cognitive models.

Perceived similarity can change with context and experience. Psychological spaces can be extended to do the same. An early, and still widespread, approach is to use a parameterized similarity function where attention weights stretch and contract the different dimensions of the psychological space. Complementary to a parameterized similarity function, the psychological representation can be updated following proscribed rules. For example, neural networks can be used be used to model sequentially dependent psychological spaces (i.e., network layers) where changing weights between the layers reflects changing experience. While existing approaches are powerful, current models fall short of capturing the full dynamics of human-perceived similarity.

SUMMARY POINTS

1. Relative comparisons and second-order isomorphism are a recurring principle in cognitive science. The principle occurs at multiple levels of analysis, from the neurophysiology of early vision to the analysis of human similarity judgments.
2. The similarity-representation duality states that a representation is only meaningful in the context of a similarity function. Likewise, a similarity function is only mean-

ingful in the context of a defined representation space. Consequently, the choice of a similarity function and representation is not arbitrary.
3. Three popular frameworks for psychological spaces are geometric, set-theoretic, and graph. While these frameworks have different affordances and capabilities, research has resolved some of these differences (such as representing hierarchies) with modern generalizations (e.g., non-Euclidean geometries).

FUTURE ISSUES

1. Scaling up cognitive models by using modern techniques and machine learning best practices.
2. Combat bias present in off-the-shelf modules such as CCNs and LLMs. Without appropriate intervention, these modules can amplify existing bias.
3. Continuing to discover theoretical connections between different types of psychological spaces, potentially resulting in a standard framework for studying psychological spaces.

Moving forward, large datasets, machine learning advancements, and new technologies present significant opportunities and challenges. In particular, there are opportunities to scale up. Scaling up means using increasingly larger datasets at all levels: stimulus, neural, and behavioral data. Scaling up requires the integration of modern techniques for storing and processing large datasets. Fortunately, the ability to use large datasets is baked into standard machine learning frameworks like TensorFlow and PyTorch. Machine learning frameworks pull in state-of-the-art approaches for efficient training, such as the latest gradient decent optimizers, learning rate schedules, regularization techniques, and early stopping. Collectively, all of these tools reduce the barrier to training cognitive models on large datasets.

Scaling up also inherits some challenges. First, Psychology researchers must become well-versed in current software tools and best practices. Second, Models are only as good as the data they are trained on and it is has become abundantly clear that many models—such as LLMs—pass on and potentially amplify historical biases. Likewise, datasets and models can differ in their respect of personal privacy. For example, as of March 11, 2021, the ILSVRC image dataset is available will people's face blurred and it is strongly encouraged that researchers use the new privacy-aware version. Psychology researchers will have to stay vigilant; implementing custom remediations when possible or quickly adopting the remediations shared by others. In the absence of vigilance, the psychological research risks disseminating biased knowledge and making things worse. With these pitfalls in mind, the opportunities for wide-scale, high-impact research is enormous.

Research has generated a diverse zoo of models that depend on psychological spaces. Along the way, interesting connections have been made between seemingly distinct approaches. The debate between similarity functions based on Laplace of Gaussian distribution was resolved by taking into account encoding noise If Euclidean approaches are generalized to include hyperbolic spaces, then the weakness highlighted by Tversky and colleagues (Tversky & Gati 1982, Tversky & Hutchinson 1986) is largely resolved. Shepard's theory of

generalization can be case in a more general Bayesian framework that subsumes a version of Tversky's set-theoretic model of similarity (Tenenbaum & Griffiths 2001). The cognitive models GCM and ALCOVE are closely related to a statistical model called kernel logistic regression (Jäkel et al. 2008). There is even a deep connection between deep neural networks and exemplar-based (i.e., RBF networks) (Maruyama et al. 1992). Future work will continue to resolve differences and assemble stronger, more versatile psychological spaces.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by an ESRC Grant ES/W007347/1, Wellcome Trust Investigator Award WT106931MA, and a Royal Society Wolfson Fellowship 183029 to B.C.L.

LITERATURE CITED

AI O. 2023. Gpt-4 technical report. Tech. rep., Open AI

Andrew G, Arora R, Bilmes J, Livescu K. 2013. Deep canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning* 28(3):1247–1255

Ashby FG, Lee WW. 1991. Predicting similarity and categorization from identification. *Journal of Experimental Psychology: General* 120(2):150–172

Bach FR, Jordan MI. 2005. A probabilistic interpretation of canonical correlation analysis. Tech. Rep. 688, Department of Statistics, University of California, Berkeley

Badre D, Bhandari A, Keglovits H, Kikumoto A. 2021. The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences* 38:20–28

Balasubramanian M, Schwartz EL. 2002. The isomap algorithm and topological stability. *Science* 295(5552):7–7

Bell AJ, Sejnowski TJ. 1995. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation* 7(6):1129–1159

Bernardi S, Benna MK, Rigotti M, Munuera J, Fusi S, Salzman CD. 2020. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* 183(4):954–967.e21

Bobadilla-Suarez S, Ahlheim C, Mehrotra A, Panos A, Love BC. 2020. Measures of neural similarity. *Computational Brain & Behavior* 3(4):369–383

Boster JS, Johnson JC. 1989. Form or function: A comparison of expert and novice judgments of similarity among fish. *American Anthropologist* 91(4):866–889

Braunlich K, Love BC. 2022. Bidirectional influences of information sampling and concept learning. *Psychological Review* 129(2):213–234

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, et al. 2020. Language models are few-shot learners, In *Advances in Neural Information Processing Systems*, eds. H Larochelle, M Ranzato, R Hadsell, M Balcan, H Lin, vol. 33, pp. 1877–1901, Curran Associates, Inc.

Carroll JD. 1976. Spatial, non-spatial and hybrid models for scaling. *Psychometrika* 41:439–463

Carroll JD, Chang JJ. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika* 35(3):283–319

Carroll JD, Wish M. 1974. Models and methods for three-way multidimensional scaling. In *Contemporary Developments in Mathematical Psychology*, eds. DH Krantz, RC Atkinson, RD Luce, P Suppes. San Francisco, CA: W. H. Freeman, 57–105

Comon P. 1994. Independent component analysis, a new concept? *Signal Processing* 36(3):287–314

Cortes C, Mohri M, Rostamizadeh A. 2012. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research* 13(1):795–828

Cristianini N, Kandola J, Elisseeff A, Shawe-Taylor J. 2006. On kernel target alignment. In *Innovations in machine learning*. Springer, 205–256

Cunningham JP. 1978. Free trees and bidirectional trees as representations of psychological distance. *Journal of Mathematical Psychology* 17(2):165–188

Cunningham JP, Yu BM. 2014. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience* 17(11):1500–1509

Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection, In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE

Devlin J, Chang MW, Lee K, Toutanova K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, In *Proceedings of NAACL-HLT*, pp. 4171–4186

Diedrichsen J, Yokoi A, Arbuckle SA. 2018. Pattern component modeling: A flexible approach for understanding the representational structure of brain activity patterns. *NeuroImage* 180:119–133New advances in encoding and decoding of brain signals

Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R. 1988. Using latent semantic analysis to improve access to textual information, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 281–285

Ennis DM. 1988. Confusable and discriminable stimuli: Comment on nosofsky (1986) and shepard (1986). *Journal of Experimental Psychology: General* 117(4):–408–411

Ennis DM, Palen JJ, Mullen K. 1988. A multidimensional stochastic theory of similarity. *Journal of Mathematical Psychology* 32(4):449–465

Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, et al. 2019. fmriprep: a robust preprocessing pipeline for functional mri. *Nature Methods* 16(1):111–116

Fechner GT. 1860. Elemente der psychophysik. Breitkopf u. Härtel

Fukushima K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics* 36:193–202

Gallego JA, Perich MG, Miller LE, Solla SA. 2017. Neural manifolds for the control of movement. *Neuron* 94(5):978–984

Goodman N. 1972. Seven strictures on similarity. In *Problems and Projects*, ed. N Goodman. New York: Bobbs-Merrill., 437–447

Guidolin A, Desroches M, Victor JD, Purpura KP, Rodrigues S. 2022. Geometry of spiking patterns in early visual cortex: a topological data analytic approach. *Journal of The Royal Society Interface* 19(196):20220677

Guyer P, Wood AW. 1998. Critique of pure reason. Cambridge University Press

Hardoon DR, Szedmak S, Shawe-Taylor J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16(12):2639–2664

Hebart MN, Zheng CY, Pereira F, Baker CI. 2020. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behavior* :1173–1185

Hinton GE, Zemel RS. 1993. Autoencoders, minimum description length and helmholtz free energy, In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, p. 3–10, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Hornsby AN, Evans T, Riefer PS, Prior R, Love BC. 2020. Conceptual organization is revealed by consumer activity patterns. *Computational Brain & Behavior* 3:162–173

Hornsby AN, Love BC. 2022. Sequential consumer choice as multi-cued retrieval. *Science Advances* 8(8):eabl9754

Hotelling Harold H. 1936. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika* 28(3-4):321–377

Hubel DH, Wiesel TN. 1959. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology* 148(3):574–591

Jacob G, Pramod R, Katti H, Arun S. 2021. Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications* 12(1):1872

Jäkel F, Schölkopf B, Wichmann FA. 2008. Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review* 15:256–271

Jamieson KG, Jain L, Fernandez C, Glattard NJ, Nowak R. 2015. Next: A system for real-world development, evaluation, and application of active learning, In *Advances in Neural Information Processing Systems*, pp. 2656–2664

Johansen MK, Palmeri TJ. 2002. Are there representational shifts during category learning? *Cognitive Psychology* 45(4):482–553

Josephs EL, Hebart MN, Konkle T. 2023. Dimensions underlying human understanding of the reachable world. *Cognition* 234:105368

Kaniuth P, Hebart MN. 2022. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage* 257:119294

Kingma DP, Welling M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*

Klami A, Virtanen S, Kaski S. 2013. Bayesian canonical correlation analysis. *Journal of Machine Learning Research* 14(1):965–1003

Kornblith S, Norouzi M, Lee H, Hinton G. 2019. Similarity of neural network representations revisited, In *Proceedings of the 36th International Conference on Machine Learning*, eds. K Chaudhuri, R Salakhutdinov, vol. 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, PMLR

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* 2:4

Krizhevsky A, Sutskever I, Hinton GE. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60(6):84–90

Krumhansl CL. 1978. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review* 85(5):445–463

Kruschke JK. 1992. Alcove: an exemplar-based connectionist model of category learning. *Psychological Review* 99(1):22–44

Kruskal JB. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1–27

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, et al. 2019. Brain-like object recognition with high-performing shallow recurrent anns, In *Advances in Neural Information Processing Systems*, eds. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, vol. 32, pp. 12805–12816, Curran Associates, Inc.

Laakso A, Cottrell G. 2000. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology* 13(1):47–76

Lamberts K. 2000. Information-accumulation theory of speeded categorization. *Psychological Review* 107(2):227–260

Love BC, Medin DL. 1998. Sustain: A model of human category learning, In *Proceedings of Fifteenth the National Conference on Artificial Intelligence (AAAI-98)*, pp. 671–676, AAAI

Love BC, Medin DL, Gureckis TM. 2004. Sustain: a network model of category learning. *Psychological Review* 111(2):309–332

Lowe D. 1999. Object recognition from local scale-invariant features, In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157

Luce RD. 1959. Individual choice behavior: A theoretical analysis. New York, NY: Wiley

Luce RD. 1963. Detection and recognition. In *Handbook of Mathematical Psychology*, eds. RD Luce, RR Bush, E Galanter, vol. 1. New York, NY: Wiley, 103–190

Mach E. 1914. The analysis of sensations, and the relation of the physical to the psychical. Open Court Publishing Company

Mack ML, Love BC, Preston AR. 2016. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences* 113(46):13203–13208

Mack ML, Love BC, Preston AR. 2018. Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters* 680:31–38New Perspectives on the Hippocampus and Memory

Mack ML, Preston AR, Love BC. 2020. Ventromedial prefrontal cortex compression during concept learning. *Nature communications* 11(46):1–11

Maruyama M, Girosi F, Poggio T. 1992. A connection between grbf and mlp. Tech. Rep. AIM-1291,

Massachusetts Institute of Technology, Cambridge, Massachusetts
McConnell RK. 1986. Method of and apparatus for pattern recognition. Wayland Res., Inc., Wayland, MA
Medin DL, Goldstone RL, Gentner D. 1993. Respects for similarity. *Psychological review* 100(2):254–278
Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. Distributed representations of words and phrases and their compositionality, In *Advances in neural information processing systems*, pp. 3111–3119
Morcos A, Raghu M, Bengio S. 2018. Insights on representational similarity in neural networks with canonical correlation, In *Advances in Neural Information Processing Systems*, eds. S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, R Garnett, vol. 31. Curran Associates, Inc.
Murphy GL, Medin DL. 1985. The role of theories in conceptual coherence. *Psychological Review* 92(3):289–316
Nakayama H, Gerkin RC, Rinberg D. 2022. A behavioral paradigm for measuring perceptual distances in mice. *Cell Reports Methods* 2(6):100233
Navarro DJ, Griffiths TL. 2008. Latent features in similarity judgments: A nonparametric bayesian approach. *Neural Computation* 20(11):2597–2628
Nienborg H, Poldrack R, Naselaris T, eds. 2019. The effect of task and training on intermediate representations in convolutional neural networks revealed with modified rv similarity analysis
Nogueira R, Rodgers CC, Bruno RM, Fusi S. 2023. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*
Nosofsky RM. 1984. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10(1):104–114
Nosofsky RM. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115:39–57
Nosofsky RM. 1988. On exemplar-based exemplar representations: Reply to ennis (1988). *Journal of Experimental Psychology: General* 117(4):412–414
Nosofsky RM. 1992. Similarity scaling and cognitive process models. *Annual Review of Psychology* 43(1):25–53
Nosofsky RM, Sanders CA, Meagher BJ, Douglas BJ. 2018. Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods* 50:530–556
Peng W, Varanka T, Mostafa A, Shi H, Zhao G. 2022. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(12):10023–10044
Pennington J, Socher R, Manning CD. 2014. Glove: Global vectors for word representation, In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543
Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. 2018. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience* 38(33):7255–7269
Ratliff F. 1965. Mach bands: quantitative studies on neural networks. San Francisco, CA: Holden-Day
Rau MA, Mason B, Nowak R. 2016. How to model implicit knowledge? similarity learning methods to assess perceptions of visual representations. *International Educational Data Mining Society*
Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2(11):1019–1025
Roads BD, Love BC. 2021. Enriching ImageNet with human similarity judgments and psychological embeddings, In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3547–3557
Roads BD, Mozer MC. 2019. Obtaining psychological embeddings through joint kernel and metric

learning. *Behavior Research Methods* 51:2180—-2193

Roads BD, Mozer MC. 2021. Predicting the Ease of Human Category Learning Using Radial Basis Function Networks. *Neural Computation* 33(2):376–397

Roads BD, Xu B, Robinson JK, Tanaka JW. 2018. The easy-to-hard training advantage with real-world medical images. *Cognitive Research: Principles and Implications* 3(38)

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326

Sattath S, Tversky A. 1977. Additive similarity trees. *Psychometrika* 42(3):319–345

Schrimpf M, Kubilius J, Lee MJ, Ratan Murty NA, Ajemian R, DiCarlo JJ. 2020. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* 108(3):413 – 423

Sexton NJ, Love BC. 2022. Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances* 8(28):eabm2219

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

Shepard RN. 1958. Stimulus and response generalization: tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology* 55(6):509–523

Shepard RN. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika* 27(3):219–246

Shepard RN. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science* 210:390–398

Shepard RN. 1982. Geometrical approximations to the structure of musical pitch. *Psychological Review* 89(4):305–333

Shepard RN. 1987. Toward a universal law of generalization for psychological science. *Science* 237(4820):1317–1323

Shepard RN. 1988. Time and distance in generalization and discrimination: Reply to ennis (1988). *Journal of Experimental Psychology: General* 117(4):415–416

Shepard RN, Chipman S. 1970. Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology* 1(1):1 – 17

Shepard RN, Hovland CI, Jenkins HM. 1961. Learning and memorization of classifications. *Psychological monographs: General and applied* 75(13):1–42

Sievert S, Ross D, Jain L, Jamieson K, Nowak R, Mankoff R. 2017. Next: A system to easily connect crowdsourcing and adaptive data collection, In *Proceedings of the 16th Python in Science Conference*, pp. 113–119

Silva V, Tenenbaum J. 2002. Global versus local methods in nonlinear dimensionality reduction, In *Advances in Neural Information Processing Systems*, eds. S Becker, S Thrun, K Obermayer, vol. 15. MIT Press

Simonyan K, Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition, In *International Conference on Learning Representations*

Smith JD, Minda JP. 1998. Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, memory, and cognition* 24(6):1411–1436

Sperry RW. 1952. Neurology and the mind-brain problem. *American Scientist* 40(2):291–312

Srinivasan Shipman AC, Boster JS. 2008. Recall, similarity judgment, and identification of trees: A comparison of experts and novices. *Ethos* 36(2):171–193

Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N. 2021. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience* 33(10):2044–2064

Stringer C, Pachitariu M, Steinmetz N, Carandini M, Harris KD. 2019. High-dimensional geometry of population responses in visual cortex. *Nature* 571:361–365

Tamuz O, Liu C, Belongie S, Shamir O, Kalai AT. 2011. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*

Tenenbaum JB, de Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimen-

sionality reduction. *Science* 290(5500):2319–2323

Tenenbaum JB, Griffiths TL. 2001. Generalization, similarity, and bayesian inference. *Behavioral and brain sciences* 24(4):629–640

Tversky A. 1977. Features of similarity. *Psychological Review* 84(4):327–352

Tversky A, Gati I. 1982. Similarity, separability, and the triangle inequality. *Psychological Review* 89(2):123–154

Tversky A, Hutchinson JW. 1986. Nearest neighbor analysis of psychological spaces. *Psychological Review* 93(1):3–22

van der Maaten L, Hinton G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(86):2579–2605

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need, In *Advances in Neural Information Processing Systems*, eds. I Guyon, UV Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, R Garnett, vol. 30. Curran Associates, Inc.

Von Uexküll J. 1926. Theoretical biology. Harcourt, Brace & Co.

Wah C, Horn GV, Branson S, Maji S, Perona P, Belongie S. 2014. Similarity comparisons for interactive fine-grained categorization, In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Columbus, OH

Wang Z, Zhang J, Feng J, Chen Z. 2014. Knowledge graph embedding by translating on hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence* 28(1)

Wilber MJ, Kwak IS, Belongie SJ. 2014. Cost-effective hits for relative similarity comparisons, In *Second AAAI Conference on Human Computation and Crowdsourcing*

Witten DM, Tibshirani RJ. 2009. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology* 8(1)