A Model of Similarity: Metric In a Patch

Yinon Nachshon[1], Haim Cohen[1,2] , Matania Ben-Artzi[3*] ,Anat Maril[1,4*]

[1]Department of Cognitive Science, The Hebrew University of Jerusalem, Israel

[2] Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel

[3]Department of Mathematics, The Hebrew University of Jerusalem, Israel

[4]Department of Psychology, The Hebrew University of Jerusalem, Israel

*As PhD advisors to the first author, Matania Ben-Artzi and Anat Maril contributed equally to this work.

Abstract

We introduce a novel model of similarity. Following previous models, we espouse the metric approach, namely, (dis)similarity between objects is represented as distance. Unlike previous models, we incorporate a distinction between a long term memory (LTM)-like probability space that functions as a data-base, and a short term memory (STM)-like space in which similarity is calculated. STM in our model, is a manifold-like metric space, which contains only a subset of LTM at a time and changes constantly with respect to both its content and metric. On top of this structure of STM, we assume another discrete layer. This layer represents the limited sensitivity of our "measuring tools", namely, our limited capability of distinguishing between similar objects. It results in a "pixalization" of STM representing limited resolution. We associate this seemingly shortcoming of STM to the very important function of abstraction. Finally, we show how probability (that exists in LTM) dictates varying concept representations and similarity in STM.

Keywords: semantic space, semantic similarity, metric, manifold, attention, resolution, varying representation

1. Scientific Background: Similarity Modelling

1.1. Similarity modelling: Overview

The psychological notion of similarity has been studied and modelled extensively over the past decades. Similarity underlies many aspects of our most basic daily activities and forms the basis for several higher cognitive functions, such as classification and abstraction. Every day, we encounter objects with which we are unfamiliar and, on the basis of the similarity of these objects to others with which we are familiar, we decide how to respond to them. For example, it is due to similarity that we are able to classify the large moving objects on the road as cars despite seeing each individual car for the first time. A successful model of similarity should provide an understanding and make predictions regarding all of the different functions that depend on similarity. The current research offers a novel model of similarity which, unlike previous models, allows for a representation's dependence on different attributes to vary over time, as well as across psychological space. As detailed in the following sections, the proposed model successfully addresses the two major challenges facing any metric similarity model: (a) to be psychologically accurate and explanatory and (b) to account for violations of metric axioms (a major challenge for metric models of similarity).

As suggested above, models of similarity typically refer to the notion of psychological space (e.g. Tenenbaum & Griffiths, 2001; Hahn, 2014; Nosofsky, 1992). An individual's psychological space consists of all of the objects and concepts that the individual has come to know. Generally speaking, in metric models of similarity, the distance between objects in psychological space represents the dissimilarity of those objects (Shepard, 1987; Krumhansl, 1978; Tversky & Krantz, 1970; Borg & Groenen, 2005). Accordingly, an object is usually represented as a point in psychological space while a concept is represented as a subset of psychological space comprising

all of the points that represent objects relevant to that concept. For example, the pen I hold in my hand (an object) is represented as a point, while the concept of a pen is represented as the set of points representing all possible pens.

Several metric models of similarity can be found in the literature: (a) Euclidean models, in which the space is $\mathbb{R}^n$, an object is a point and dissimilarity is represented by Euclidean distance (Shepard, 1987); (b) Euclidean probabilistic models, in which objects are represented as sets of possible manifestations (with a probability density function attached to each one) and the dissimilarity between two objects is inversely proportional to the degree to which they overlap (Ashby & Townsend, 1986); (c) models of metric space with infinite dimensions (with a continuum cardinality), which have the advantage of being able to represent sets in the traditional finite dimensional psychological space as points and thus can be used to measure distances between sets (Townsend, Burns & Pei, 2013) and, finally, (d) models of psychological space as a manifold, which allow for varying contributions of attributes to the metric (Nosofsky, R. M. 1986)).

Common to all these models is the idea that coordinates represent attributes and that an object is, therefore, represented solely by the values of its attributes. In some models, an attribute's contribution to the metric, the attribute's "weight," may differ across different similarity judgments.

1.1.1. The problem: Psychological accuracy versus mathematical correctness

While the intuition underlying the modeling of (dis)similarity as a metric seems obvious, the application of a mathematical tool to such an abstract entity as psychological space is not at all trivial. As noted above, the primary difficulty with the metric approach in this context stems

from the fact that violations of metric axioms appear to be involved in similarity judgment tasks (Laub, Müller, Wichmann & Macke, 2006; Medin, D. L., Goldstone, & Gentner, 1993; Tversky, 1977; Tversky and Gati, 1982; Voorspoels et al., 2011; Yearsley et al., 2017; Jäkel et al., 2008; Aguilar & Medin, 1999). Before further considering this difficulty, we introduce the metric axioms below.

Let *x*, *y* and *z* be points in a metric space, which is a set *M* together with a distance function $d: M\ X\ M: \rightarrow \mathbb{R}$ obeying the metric axioms presented in Table 1:

Table 1. Metric axioms and their psychological meanings.

Axiom		Psychological Meaning
$\boldsymbol{d(x,y) \geq 0}$	**Non-negativity**	An object *x* can't be more similar to object *y* than to itself
$\boldsymbol{d(x,y) = d(y,x)}$	**Symmetry**	An object *x* is similar to an object *y* exactly as much as the object *y* is similar to the object *x*.
$\boldsymbol{d(x,y) = 0 \leftrightarrow x = y}$	**Identity of indiscernibles**	An object is identical only to itself.
$\boldsymbol{d(x,z) \leq d(x,y) + d(y,z)}$	**Triangle inequality**	If *x* is similar to *y* and *y* is similar to *z*, then *x* and *z* cannot be very dissimilar.

One major axiom violation that have been discussed in the literature involve the triangle-inequality axiom; we illustrate it using the following well-known example that deals with concepts rather than objects, but, nevertheless, does demonstrate the main idea: `Flame` is similar to `sun` and `sun` is similar to `ball` but `flame` is not at all similar to `ball`. The second

documented major axiom violation involves the symmetry axiom. For example, Cuba is similar to the USSR, since like the USSR used to be, Cuba is also Communist, but the opposite is not true due to the abundance of information about the USSR, which distinguishes it from Cuba (Tversky 1977, Tversky & Gatti 1978, Tversky & Gatti 1981). We address both of these axiom violations below. A successful model of similarity ought to account for such violations while at the same time being psychologically accurate and explanatory.

1.1.2. Our solution: A patch of scaled resolution

The model of similarity proposed here successfully handles these challenges by assuming an ever-changing patch, which represents the psychological active subspace. This patch has two main properties, which we will later discuss in detail: a) finite resolution and b) a manifold-like structure. These two properties allow, respectively: a) taking into consideration the limited accuracy of our measuring tools (which has important consequences that we will discuss later) and b) varying dependence of representations (of objects) on attributes across the patch and over time. Varying representation over time is achieved by the continuous changing of the patch's metric and content. That is, not only does the representation of objects vary over time, but the objects composing the set of reference vary as well.

We believe that these ideas are in line with the psychological reality of an object's representation and, at the same time, provide a clear mathematical solution to various difficulties stemming from metric descriptions of the psychological space. We see attention as the mediator of the process of 'translating' context into a metric: Attention focused on the difference between two

objects decreases their similarity. The resolution of the sub-space, in turn, heavily depends on the metric; greater distance between two objects increases their separability[1].

The idea that the representation of an object may vary over time means that representations are always context-dependent. For example, I see a lion in a cage at the zoo. In this context, the attributes that draw my attention may include the lion's elegance, graceful gait and shiny fur. In other words, these are the attributes that are likely to contribute the most to my representation of the lion and, accordingly, to any similarity judgment involving the lion. But, suppose the cage suddenly breaks open. Now, the dominant attributes would most likely be quite different and include the lion's speed, its strength and the sharpness of its teeth. These are also the attributes that will gain dominance in 'calculating' the lion's similarity to other objects. (e.g., I do not know much about the speed of lions, but I do know that other big cats are very fast).

With respect to variance in object representation across psychological space, we propose that the contribution of a given attribute to a representation (and, as a result, to similarity) differs for different objects across the psychological space. To continue with the example above, the speed of a bird flying in the sky would likely be much less relevant and draw less attention than the speed of the free-roaming lion. Thus, in this context, speed is an attribute that would influence similarity judgments involving the lion to a much greater degree than it would influence similarity judgments involving the bird.

We assume that the metric changes **continuously** over time and across the patch. In general, this means that an attribute that is very important at Time t for the representation of an object A (and thus for the similarity of A to other objects) will also be important, firstly, for other similar

[1] In our view, as will be discussed later in this paper, diminished resolution is sometimes beneficial.

objects, and secondly, for the object A in temporal proximity to t (a formalization of this assumption is presented later in this work).

1.1.2.1. Advantages

The principles of our model were discussed briefly in the introduction. These principles yield three main advantages for our model, as described in the sections below.

a. Psychological accuracy

Most metric similarity models use a flat metric space[2]. These models are not suitable for describing large portions of the psychological space, in which the identity of the attributes and even their number (which is interpreted as the dimension of the space) may vary. For example, many attributes that are relevant for the description of a face are irrelevant for the description of a pen. A model of varying representation overcomes this difficulty.

Allowing, however, for variance in "dimension" across psychological space raises the problem of how to compute distance (i.e., dissimilarity) between objects that are represented by different numbers of attributes. For example, clothing adds a whole new set of attributes to a person's appearance (thereby increasing the dimension), yet, when identifying a person, the relevant similarity judgment is between the person's pre-established representation in psychological space, independent of any specific clothing, and the person in the real world, who happens to be clothed one way or another. In other words, the relevant similarity judgment involves a comparison between an object of lower dimension (the person irrespective of clothing) and an object of higher dimension (the person in particular clothing).

[2] In our case, it can be described as a space across which the metric's dependence on an attribute remains constant.

Our model addresses this problem by allowing representations to vary over time. That is, we allow representations to be context-dependent. This is where diminished resolution becomes handy: In an effective process of identifying a person, the metric is not influenced by the clothing-related attributes. As a result, dimension is reduced by diminished resolution along the clothing-related attributes. The resultant space is 'spanned' by the person's identification-relevant attributes (e.g., the shape of the face). In other words, attributes relevant to the identification of the person gain dominance in the representation, while attributes related to clothing, being irrelevant to the task at hand, can be ignored. Note that difficulty ignoring clothing-related attributes may impair the identification task, which means that the diminished resolution here is functional.

b. Similarity in varying levels of abstraction

Note that the way our model accounts for tasks involving representations of different dimensions (like the identification task) is in line with the claim (discussed in detail below) that the distinction between objects and concepts in practice is not definite, but rather gradual. As such, we would like to be able to use the same mechanism to describe similarity for any level of abstraction. Scaled resolution allows us to do so. An evaluation of the similarity of different concepts uses the same mechanism, but with rougher resolution.

c. Accounting for metric-axiom violations

As detailed below, another advantage of allowing varying representation over time in our model is that it provides a reasonable explanation for violations of metric axioms in similarity-judgment tasks. Note that (a) the very execution of a similarity judgment dictates the specific representation activated and (b) at any one definite point in time, metric axioms may hold with respect to a given patch, but (c) that patch changes with time. Given that metric axioms are

applied to similarity judgments carried out at different times (e.g., `sun-flame; sun-ball`), violations of metric axioms that necessarily stem from rigid representations cease to exist when changing ad hoc representations are permitted.

1.1.3. Patch: Justification

In psychological research (Baddeley, & Warrington, 1970; Buchsbaum, Padmanabhan, and Berman, 2011; Corkin, 2002; Squire, 2009; Vallar & Baddeley, 1984; Warrington & Shallice, 1969), memory is traditionally divided into long-term memory (LTM) and short-term memory (STM or working memory). The similarity model proposed here, assuming a patch characterized by specific content and metric, is essentially a model of an online active subspace of the psychological space, akin to STM[3] (though the values of the so called "continual" attributes must be preserved in LTM). Previous models of similarity have not incorporated the STM — LTM distinction and, consequently, committed themselves to invariable representations compatible with LTM. While "continual" representations do exist only in LTM, we believe that STM (the patch mentioned above) is the primary locus for similarity judgments, which rely on temporary, incomplete and varying object representations and involve other resources (e.g., perception) besides LTM. We claim that similarity is not information about objects, but rather a momentary impression about relations between objects.

1.1.3.1. Why a manifold?

We argued above that metric and object representations in STM change continuously, over time and across the patch. The possibility of representing continuous change across space and time is

[3] The terms STM and LTM as used here do not necessarily conform to the STM-LTM full characterization as analysed in the memory literature. Rather, they are used to express an active, transitory, limited-capacity subspace of the psychological space, akin to STM, and a relatively stable, long-lasting, (practically) infinite-capacity knowledge repository, akin to LTM.

best served by a (suitable geometric object) manifold-like model. The continuity of change is important, to limit the flexibility of the model and produce predictions.

To summarize, the core of our model of similarity involves a manifold-like, finite-resolution patch, which becomes possible when we consider a model of similarity in STM. A manifold-like structure allows for varying representation across the patch. The three main advantages of a patch model are that: (a) it is psychologically accurate, considering context and attention influences; (b) it is capable of describing similarity at varying levels of abstraction and (c) it avoids the problem of metric-axiom violations.

We present the proposed model below, beginning with necessary definitions and a general description of similarity rules. We then move on to discuss the notion of finite resolution, which, together with varying representation, makes our model psychologically accurate. We then consider concept representation and similarity judgments at varying levels of abstraction. Finally, we discuss probability and classification. The issue of metric-axiom violations is discussed elsewhere (Nachshon, Cohen and Maril, 2022).

2. The model: Overview

2.1. Definitions

We define an **attribute** relevant to a set S of objects in some patch of the psychological space (such as STM) as a function φ_i from the set S to the real line. This permits a full ordering of objects in S along this attribute. Namely, an attribute is a variable according to which objects can be ordered, which serves as a component in the description of the objects in this set. For example, for an attribute X, an object A is more/equally/less X than object B**.** In other words, an attribute comprises information that can be used to compare two objects (e.g., one tree is taller or

greener than another). Note that an attribute defined in this way does not have to carry any objective or absolute (let alone physical) meaning.

An ensemble of relevant attributes in a set S of a semantic space induces a topology as the minimal (coarsest) topology for which all the relevant attributes are continuous, which is to say that two objects in *S* are "close" if and only if they are "close" in terms of every relevant attribute, meaning, if $|\varphi_i(x) - \varphi_i(y)|$ is small for every i. We say that an ensemble of *n* relevant attributes in a neighborhood around a point *x* is valid if the image of the function $f: S \to \mathbb{R}^n$ defined as $f(x) = (\varphi_1(x), .., \varphi_n(x))$ contains an open set around $f(x)$.

The **dimension** of a neighborhood around a point *x* in the set *S* is now (well) defined as the size of a valid ensemble of attributes, which means the smallest number of attributes needed to give a full description of this neighborhood. Note that f defines the topology on *S*, so that if f is a local homeomorphism between the neighborhood of x and its image under f, then, locally, the set *S* will have the structure of a topological manifold.

2.2. Manifold-like structure of STM: Intuition

As noted above, our model is a model of similarity in STM. This section is dedicated to an intuitive discussion of this idea, which is followed by a more formal discussion in the following section.

We look at STM as a limited-capacity space, which we call $V(t)$[4], and which holds at a certain time a set of concepts[5] and a local metric (dissimilarity function). Both the set and the metric vary continuously over time. An illustration of the need for local settings would be as follows:

[4] The finite capacity of STM is generally agreed upon. In our model, this is expressed by the finite volume of $V(t)$.
[5] A set of concepts from LTM alongside other resources (e.g., perception).

Two rectangles drawn on a piece of paper may be similar because they are "close" in length, in width, in color, etc. Usually, we can tell how much each attribute affects similarity. However, a problem arises when we try to formulate a similarity rule that is more general than a mere collection of local similarities. As discussed above, different attributes are relevant at different places across $V(t)$, over time and to different extents and, therefore, the similarity rule changes accordingly. Here is where a manifold-based model becomes handy. Instead of relying on a global distance rule, in manifold-based models, we can calculate the length of small intervals using a local distance rule and we can then add up these distances to compute path lengths. The distance between two objects is then defined as the length of the shortest possible path between two points[6]. We argue that this is the way in which dissimilarity is evaluated.

In addition, the limited capacity (finite volume) of $V(t)$ means that any attention paid to an attribute always comes at the expense of attention paid to other attributes and attention to certain semantic domains comes at the expense of attention that could be paid to other semantic domains. Below, we suggest a formal version of these general ideas.

2.2.1. Similarity formalization and definitions: Metric manifold

The topology of a semantic set was defined above (as the coarsest topology for which all of the attributes are continuous). Locally, the set with this topology has the structure of an n-dimensional manifold (see page 9 above). As mentioned above, we define STM as a set we call $V(t)$, which is the set of objects attended to at Time t, with the topology defined above. $V(t)$, then, is locally homeomorphic to $\mathbb{R}^n$ (for some $n \in \mathbb{N}$) and, locally, has the structure of a topological manifold. For the sake of simplicity, we start by looking at $V(t)$ as having one

[6] Note that a global metric defined in this manner obeys the metric axioms.

coordinate system, in which each coordinate represents an attribute[7]. We assume that this set has a metric[8] that represents dissimilarity. We further assume that $V(t)$ has the structure of a Riemannian manifold[9]. The general idea is as follows: Locally, a small neighborhood in $V(t)$[10] around a point $p \epsilon V(t)$ can be approximated as a linear vector space (the tangent space at p) in which an inner product and thus a norm can be defined and, therefore, distances can be calculated. Let q and p be two points in $V(t)$ and let $p*$ and $q*$ be their respective coordinates $v = q* - p*$ where v is an array of differences along attributes. The matrix $g_{p,t}$[11] , then, describes the similarity between p and q in the following way: For a sufficiently small $|v|$, we get $d(p,q) \approx \sqrt{g_{p,t}(v,v)}$. Now, a length of a path $\theta: (a,b) \to V(t)$ can be approximated in the following way: For a sufficiently small $\varepsilon > 0$, we choose a set of points $S_\varepsilon = (p_0, p_1, \dots, p_l)$ along the path θ such that $p_i = \theta(\varepsilon \cdot i)$, $p*_i$ its representative coordinates and $p*_{i+1} = p*_i + v_i$. The approximated length of θ is given by $\sum_{i=0}^{l-1} \sqrt{g_{p_i,t}(v_i, v_i)} = \sum_{i=0}^{l-1} \varepsilon \cdot \sqrt{g_{p_i,t}(\frac{v_i}{\varepsilon}, \frac{v_i}{\varepsilon})}$. The smaller ε gets, the more accurate the calculated path length, when for infinitesimal ε we get $\lim_{\varepsilon \to 0} \frac{v_i}{\varepsilon} = \theta'(\varepsilon \cdot i)$ and a path length of $\int_a^b \sqrt{g_{(\theta(s),t)}\big(\theta'(s), \theta'(s)\big)} ds$ for the (differentiable)

[7] Namely, a single chart, therefore, the differentiability of $V(t)$ derives naturally from the differentiability of $\mathbb{R}^n$.
[8] Which is to say that similarity is not information about objects, but rather context-dependent.
[9] For now, we ignore singularities, that is, places where the dimension of $V(t)$ changes. As mentioned above, for convenience, we assume a single chart (a single coordinate system). Therefore, both the differentiability of paths and the tangent vector to a path in a point can be defined.
[10] For now, we ignore the dependence of similarity on time and concentrate on the metric at a certain moment in time within $V(t)$.
[11] The matrix g represents the inner product, which is a generalization of the dot product of two vectors $v = (x_1, y_1),\ u = (x_2, y_2), \langle v, u \rangle = \langle (x_1, y_1), (x_2, y_2) \rangle = x_1 \cdot x_2 + y_1 \cdot y_2 = |v| \cdot |u| \cdot cos\theta$ where $cos\theta$ is the angle between the two vectors. The inner product induces a metric in the following way: $\langle u, u \rangle = |u|^2$ when $|u|$ is the norm (the 'size') of u. The distance between two close enough points, A and B, in an inner product vector space is now $|A - B|$. Note that the notion of an inner product exists in most of the metric models of similarity, as a measure of interdependence between two attributes (variables). g is represented as a matrix, which in classic metric manifolds should be *positive definite*. As we shall see later, we allow the matrix to degenerate, but we do not allow negative eigenvalues, which means that any difference along an attribute can have only a positive effect on dissimilarity (g is positive **semidefinite**).

path $\theta: (a, b) \to V(t)$[12], and the minimal path length between the two endpoints, $\min_{\theta}\left\{\int_a^b \sqrt{g_{(p,t)}\big(\theta'(s), \theta'(s)\big)ds}\right\}$, is the subjective dissimilarity between $\theta(a)$and-$\theta(b)$.

We have then:

$$d(p,q) = \min_{\theta:\theta(a)=p \text{ and } \theta(b)=q}\left\{\int_a^b \sqrt{g_{(p,t)}\big(\theta'(s), \theta'(s)\big)ds}\right\} \qquad (1)$$

As we discussed above, $V(t)$ is a metric space, whose metric represents dissimilarity. We will now discuss the changes in $V(t)$ over time. To do so, we will first define a topology on $V = \bigcup_{t=t_1}^{t_2} V(t)$ as a which is as follows: a subset U of V is open, if and only if any projection of U on V(t) is open in V(t) and the projection of U on the time axis is open. In addition for$(p, t_0) \in U$ there is $\delta > 0$ such that for t for which $|t - t_0| < \delta$, there is an open neighborhood W_p(t) in V(t) of p such that $\bigcup_{t=t-\delta}^{t+\delta} W_p$ is also in U. With this topology, V is locally homeomorphic to $\mathbb{R}^{n+1}$. Note that points (p, t) in V are events rather than objects. Several studies have shown that the similarity between events (p_1, t_1) and (p_2, t_2) depends also on the size of the time interval $t_2 - t_1$ between them (Day, Bartels 2004). Therefore, it is natural to define a metric G in V, such that, when limited to $t = t_0$ we get the metric g_{t_0} in $V(t_0)$[13]. Note that since we have $V(t)$ for every t, an event (p, t_0) (an object p in $V(t_0)$) gives rise to a (smooth) curve P in V, which

[12] We can say then that $\sqrt{g_{(\vec{p},t)}(\vec{v_i}, \vec{v_i})}$ is the rate of change of the dissimilarity in the direction of the unit vector $\vec{v}$ at Time t at Point p. In particular, $\sqrt{g_{ii(p,t)}} = \sqrt{g_{(\vec{p},t)}(e_i, e_i)}$ is the change in the dissimilarity in the direction of the attribute e_i.

[13] At a point (p, t) for two vectors u, v with zero projection on the time axis, $G_{(p,t)}(u, v) = g_{(p,t)}(u, v)$.

intersects any $V(t)$ at exactly one point, $P(t)$. The curves of the type $P(t)$[14]carry the attributes of an object p as they vary in time (e.g. an object may change its color).

Once we look at the set V as a topological space, we add the assumption that the manifold V (t) and the metric g (t) vary smoothly in time. This enables us to regard the metric G, mentioned above, as a smooth metric on V.

An **attention-density function** will be defined as:

$$Ad(p,t) = \sqrt{|det\ G\ (p,t)|} \qquad (2)$$

(which is the size of a volume element) for $p \in V$. The consensus about finite attention load can thus be translated as $\iint_{\overline{V}(t)} Ad_t(p)dp \leq M < \infty$, where $Ad_t(p) = Ad(p,t)$. For the sake of simplicity, we will assume $\iint_{\overline{V}(t)} Ad(p,t)dp = vol\big(V(t)\big) = const$. The attention to a subset A of $V(t)$ will be-$\iint_A\ 1\ dx$, meaning the volume of A in $V(t)$.

2.3. <u>Scale of resolution: Toward a concepts-as-points representation</u>

In this section, we introduce the notion of scaled resolution, which states that in some situations, not every two points can be distinguished. Scaled resolution expresses the fact that the ability to distinguish between two very similar, yet different, objects is limited. For example, when I look at a lion at the zoo, the lion is an object. It is an object in the sense that I know that the lion has a certain value for each and every one of the attended attributes. However, I do not know the exact values of these attributes. Better evaluation of attributes' values requires effort. Another example: Sometimes I may mistakenly identify someone, at first glance, as a person whom I

[14]Any such curve is related to a specific object

know. A closer look, however, may reveal that the person I know is actually taller, heavier, etc. In such a situation, it is my ‘measuring tools’ that are limited and not the space.

Scaled resolution is necessary for two main reasons, which are discussed below in greater detail. The first is that finite resolution enables modelling of uncertainty in identification, a common psychological reality that must be taken into consideration in any model of similarity (e.g., it may underlie mental pathologies). The second reason involves the consideration of concepts as clusters of points, which allows for elegant modelling of the similarity between concepts in the very same manner in which the similarity between objects is modeled — as the distances between such clusters of points. It is also intuitively appealing: When we consider the extent of the similarity between wolves and dogs, we do not need to consider the dissimilarity between different dogs and the dissimilarity between different wolves, rather, we "compute the distance" between a `wolf` and a `dog`.

In general, previous metric models of similarity are continuous and have no associated discrete structure. As such, they latently assume infinite resolution. The meaning of resolution in this context is the ability to distinguish between two different points (objects). In continuous models of similarity, every object is represented differently and, therefore, any two objects can be distinguished (Townsend, Burns & Pei, 2013; Nosofsky, R. M. 1986).

Nevertheless, many similarity models (F. G. Ashby and J. T. Townsend’s (1986); Nosofsky (1986); Medin & Schaffer, 1978) espouse a probabilistic approach, which is in line with psychological reality (since there is often a level of uncertainty in object classification), but not in line with infinite resolution: If any two objects are distinguishable, no element of uncertainty should exist. Previous models of perception didn’t overlook the issue of discrimination between

stimuli and also the discrimination- distance connection (Dzhafarov & Colonius 1999). However, they didn't derive discreteness of the space from the distance- discrimination connection, nor did they apply this connection to semantics (or abstraction).

Note that, in our model, the underlying layer also has a continuous structure. However, we have incorporated into our model a similarity notion applied to a scale of coarser levels. This allows the consideration of the scale of resolutions (emanating from the continuous model). We will discuss the consequences of this structure later in this work.

Our approach to scaled resolution is a way to state that two very close objects should be unified (i.e., clustered together) for certain purposes. Representation is conducted in a finite space; therefore, the acceptance of the existence of the multidimensional set $V(t)$ imposes *pixelization*, that is, it imposes a clustered structure on $V(t)$, in which every cluster represents a set of points joined together by proximity. The c*lusterization* methodology is carried out by assembling together objects that are separated from one another by distances that are smaller than a given small, positive constant. We refer to this constant as the *distinction constant*. This small constant should represent, in our view, some fixed capacity for distinguishing between objects. As such, this constant is fixed and is not subject to change.

On the basis of our manifold $V(t)$, we construct a finite set Y of elements, each representing a fixed-sized cluster of objects of $V(t)$. This means that two objects of $V(t)$ may belong to the same element in Y (making them indistinguishable) only if the distance between them is smaller than the distinction constant. It is easy to see, then, that the number of elements of Y intersecting a curve between two points in $V(t)$is proportional to its length. As discussed above, the length of a trajectory is proportional to the attention focused on it.

It should be noted that the clustering changes over time (as the metric changes). The following subsection is devoted to the formalization of this pixelization.

2.3.1. Scaled resolution formalization

A topological structure $V(t)$ representing the content of STM together with its metric was discussed above. In our view, the fact that the brain is a finite system suggests that we need to impose some form of discrete approximation on $V(t)$. Such an approximation is possible since $V(t)$ is totally bounded. For $V(t)$ with its metric and for some $\varepsilon > 0$ (the distinction constant), there is some finite minimal cover of $V(t)$ by ε-balls Y that we see as the set of representations at time t. $Y = \{Y_i(t)\}$ has a natural metric inherited from $V(t)$, which is known as the Hausdorff distance[15]. Note that the size of a 'mistake' in distance calculation between two points p, q of $V(t)$, that is, the difference between the two distances $d(p, q)$ and $d_H(Y_p, Y_q)$ is at most 2ε.

Above, we defined attention density as the size of a volume element in $V(t)$. Meaning that the more a certain set is attended, the greater is its volume. As a result, the more a set is attended, the more clusters of Y are needed to cover it. Therefore, the greater the attention that is accorded to a set, the more Y elements will be included in it; that is, the greater the resolution will be. In cases in which every two points in $V(t)$ that differ **only** along a certain attribute e_i are included in the same $\varepsilon - ball$ (the length of the projection of $V(t)$ on e_i is smaller than ε), dimension reduction occurs (the clustered metric no longer depends on e_i).

[15] The distance between two ε-balls is the maximal distance between a point on one of the balls and the other ball.

2.4. Concept representation

2.4.1. Concept representation: Intuition and motivation

As noted above, most metric-similarity models treat objects as points, while concepts are viewed as sets containing all possible manifestations of the relevant objects. In these models, both concepts and objects, (like the similarity itself), exist in LTM. Generally, these models do not specify a mechanism for computing distances between concepts (e.g. Shepard, 1987; Krumhansl, 1978; Tversky & Krantz, 1970; Borg & Groenen, 2005). In our view, however, psychology-wise, the distinction between objects and concepts in LTM is not well defined. Many specific objects (psychology-wise) have attributes for which only a range of possible values is known (though not an exact value) or have a variety of manifestations (e.g., the same face with different make-up, a different hair style, etc.) and should, therefore, be represented as sets.

We express this idea by:

1. Looking at LTM as a probability space (see "the role of probability" section below). According to this perspective, instead of a sharp distinction between concepts and objects in LTM, it is more appropriate to talk about a hierarchy of concepts (levels of abstraction), where even "specific objects" in LTM are characterized by a distribution of attributes values. This distribution represents a 'chronic' uncertainty (lack of knowledge) - even the most familiar 'objects' (e.g. my dog) are not totally familiar (e.g. I don't know my dog's height to the millimeter). Therefore, we refer only to concepts in LTM, with varying level of specificity.

2. When concepts are retrieved to STM, the uncertainty (regarding the attributes values) is expressed in our model by the fact that the elements of the set Y (the discrete approximation, see subsection 2.3.1) can never be single points. Another level of 'acute' uncertainty, is a function of

attention as described above (a single cluster may include my dog only, or all the dogs in the world).

In general, a concept will be any well-defined subset C of LTM. On top of this, when we consider the metric in STM and its discrete approximation, Y, it follows that any concept (at any level of abstraction) may be represented in STM by a single element of the discrete approximation Y.

2.4.2. Concept representation in STM: Identification

In this section, we explain in greater detail how a concept is represented during an identification task, that is, during any task requiring identification of a concept rather than differentiation between specific objects within a concept. Examples of such tasks include concept-instantiation searches (e.g., looking for any pen on your desk, or looking for the telephone in a hotel room) and similarity judgments between two concepts (e.g., the similarity between dogs and wolves). In all of these tasks, the important attributes are those that unite rather than differentiate individual instances of the concepts involved.

Below, we briefly explain why the assumptions made in our model (particularly the assumption that two different points can be regarded as indistinguishable at a certain moment, i.e., finite resolution) lead to the conclusion that, in a concept identification tasks, the representation of the concept tends to be reduced to a single basic cluster (a single element of Y). In addition, we provide a more accurate characterization of the effects of different attributes on the metric in concept-identification tasks. We then elaborate on and formalize this discussion.

In STM, finite resolution leads to uncertainty. Thus, accurate identification of concepts requires resource allocation. As attentive load is finite, attention to a domain in the psychological space

comes at the expense of other domains and attention to an attribute comes at the expense of other attributes. Due to the connection between attention and resolution, the same assertion can be applied to resolution quality. During a concept-identification task, resolution between two objects that undoubtedly belong to the concept is redundant. Thus, all of the ensemble of the concept's objects will be represented together as a single cluster (within a single 'pixel', that is, an element of Y).

Below, we discuss probability in the semantic space. For the sake of simplicity, we first use the term in an intuitive manner. Later on, we will discuss this term in greater detail. As for attributes, attention to an attribute of a certain object in $V(t)$ is determined by its influence on the probability or the plausibility that the object represents the concept, given the values of the other attributes [coordinates in $V(t)$]. The more an attribute influences the probability that an object represents the concept, the better the resolution along that attribute will be. This probability, in turn, depends heavily on the distribution of the attribute's values within the concept. In general, the more the attribute varies between the different objects of the concept (e.g., the attribute 'color' for cars), the smaller its influence on the identification task.

Finally, to avoid missing any concept manifestations, $V(t)$ must include all of the points with a positive probability of representing the concept. The volume of the set composed of these points is a function of the attention allocated to the identification task.

2.4.2.1. <u>The role of probability</u>

We start this subsection with a look at the metric during an identification task. Note that LTM as one's database contains a lot of statistical information. For example, most dogs are between 30 and 60 cm tall, pens are around 17 cm long and there are no blue mammals or mountains higher

than Mount Everest. These pieces of data result from statistical accumulation of information and eventually comprise a person's LTM. These facts are not influenced by attention. Nevertheless, statistical information has a strong influence on the way that objects are represented in STM. For instance, when I look for a pen on my desk, I search for an object that is between 10 and 20 cm long, because the probability of a pen being shorter or longer than those limits is very low. Since probability plays a key role in the representation of concepts, which, in turn, determines their similarity, we begin by laying the foundations for probability in a semantic space. Since LTM is the primary locus for probability (probability is information about objects and concepts), we begin by defining probability in LTM. We will then describe how probability is evaluated in STM and how it influences representation in STM.

So far, we defined and discussed attributes only in STM. Note, however, that STM attributes often have counterparts in LTM. For example, I know that my father (a concept in LTM) is taller than my mother. This is information I have about my parents, therefore, it is part of my LTM. In LTM, attributes are random variables, functions from LTM to the real line (e.g., I know that about 70% of all people are shorter than my father).

We look at LTM as a space-carrying probability distribution for a set of random variables (LTM attributes). This space consists of many basic concepts (my LTM contains all of the concepts I have come to know). The basic concepts of LTM are the ones that cannot be separated into more basic ('smaller') concepts. Therefore, they are disjoint (measurable) sets of LTM. Any LTM concept (whether basic, such as `my neighbor's dog` or more abstract such as `dog`) inherits a probability distribution, which is the conditional probability (the random variables distribution given that the concept is C). This distribution is based on knowledge. A total lack of knowledge about the values of a certain variable for a concept C is represented by a uniform marginal

probability distribution of this variable (I have no idea about the speed of my neighbor's dog, it may be any speed within a reasonable range); coming to know the concept better is represented as a 'narrowing' of the distribution (having seen that dog chasing a car, I'm 90% certain that its maximal speed is less than 30 mph).

As was discussed above, both LTM and perception are sources of STM. Many tasks, such as categorization and identification, deal specifically with integration of these two sources. Note however, that elements of LTM (basic as they may be) when retrieved to $V(t)$ form sets in $V(t)$. For instance, I know the neighbor's dog pretty well (in the framework of my LTM), but I have never noticed that it has a white spot on its tail (when considered in the framework of my STM). Nevertheless, this newly found spot does not prevent me from identifying the dog as the one belonging (in high probability) to my neighbor when I see it. Namely, the concept labeled as `my neighbor's dog` is actually a set in $V(t)$ when retrieved to STM (for example, it contains dogs with and without a white spot on their tails).

As noted, we define a concept as a well-defined subset of LTM (including single basic concepts). However, not all subsets of a concept have the same status. For example, most people find it easier to identify a German Shepherd as a dog, as compared to a Miniature Pinscher. In our construction, this is a result of probability considerations: For the concept `dog`, we have a random variable that yields probabilities of various dog breeds. Thus the probability that it is a German Shepherd is higher than the probability it is a Miniature Pinscher. Later on, this connection (between identification and probability) will be discussed and justified, but for now, we start by introducing the notion of conditional probabilities in LTM.

As noted, given a concept, there is a distribution of random variables values (LTM attributes). This distribution can be interpreted as a conditional probability on the concept. Thus given the concept C (e.g., dog), we may ask what the probability of a certain subset of C is (e.g., German Shepherd), as defined by attributes' values ranges (e.g., height between 50 and 60 cm, weight between 30 and 40 kg, etc.).

Note that the conditional distribution that is relevant for identification tasks is not the probability of a subset of C given the concept is C (in LTM), but actually the opposite one. Namely, given a cluster cl in $V(t)$, what is the probability that this cluster represents a concept C?

As discussed above, LTM is a source of STM. At Time t, we consider a subset S of LTM that is retrieved to STM. We define a retrieval (set valued) function $\theta: S \to V(t)$ as follows: For $s \in S$ where s is a basic concept, $\theta(s)$ is a subset of $V(t)$. Now, we can define another function Ω on Y (the collection of $V(t)$'s clusters) that identifies, for a cluster cl in $V(t)$, all of the basic concepts in S whose images intersect with cl, namely:

$$\Omega(cl) = \{s \; for \; s \in S \; such \; that \; \theta(s) \cap cl \neq \emptyset\} .$$

Generally, when there is no identification difficulty, $\Omega(cl)$ consists of a single concept. We define identification ambiguity as $|\Omega(cl)| > 1$. Namely, cl intersects the image of more than one concept. Note that when there is ambiguity about the cluster's identity, the cluster may intersect the image of an LTM concept without including the concept's image or being included in it.

For example, I see an animal out of the corner of my eye (cl). It has the size and the color of my neighbor's dog (s_1) and I do notice that the animal has a white spot on its tail. It may be my neighbor's dog, so the intersection between the two sets is not empty ($\theta(s_1) \cap cl \neq \emptyset$).

However, the cluster cl representing the animal in $V(t)$ is not included in the image of my neighbor's dog [e.g., the animal may be another dog (s_2)]. It also does not include the image of my neighbor's dog (e.g., my neighbor's dog may not have a white spot on its tail).

Let $x_1, x_2, .., x_k, .., x_n$ be the coordinates (attributes) of $V(t)$. Without loss of generality, we assume that $x_1, x_2, \dots, x_k$ are the attributes of $V(t)$ with counterparts in LTM as random variables $X_1, X, \dots, X_k$. The attributes $x_{k+1}, x_{k+2}, \dots, x_n$ are coordinates representing the perceptual attributes (e.g., the white spot on the tail of my neighbor's dog).

We define a projection $\psi: V(t) \rightarrow \mathbb{R}^k$ as follows: For $x = (x_1, x_2, \dots, x_k, \dots, x_n)$, we define $(x) = (x_1, x_2, \dots, x_k)$. For a cluster cl, we get a set $\psi(cl)$ in $\mathbb{R}^k$.

Let S_{cl} be the union of the basic concepts in $\Omega(cl)$ as a probability space (the conditional probability): $S_{cl} = \bigcup_{\theta(C_i) \cap cl \neq \emptyset} C_i$ (3)

We define $P(cl)$ as the joint probability of $X_1, X_2, .., X_k$ on $\psi(cl)$ (i.e., $P((X_1, X_2, .., X_k) \in \psi(cl))$), when the probability is the one defined in the probability space S_{cl} and $P\left(C/_{cl}\right)$ is the joint probability of $X_1, X_2, .., X_k$ defined for $\psi(cl) \cap \psi\big(\theta(C)\big)$ divided by $P(cl)$, which is $P(X \in \psi(cl) \cap \psi\big(\theta(C)\big))/\, P(cl)$ where $X = (X_1, X_2, .., X_k)$.

Given that the concept is C, the probability of getting a value within the cluster = $P\left(cl/_C\right)$ is the joint probability of $X_1, X_2, .., X_k$ defined for $\psi(cl) \cap \psi\big(\theta(C)\big)$ divided by $P(C)$, which is $P(X \in \psi(cl) \cap \psi\big(\theta(C)\big))/P(C)$.

We have then:

$$P\left(cl/_C\right) = P(X \in \psi(cl) \cap \psi\left(\theta(C)\right))/P(C) \qquad (4)$$

$$P\left(C/_{cl}\right) = P(X \in \psi(cl) \cap \psi\left(\theta(C)\right))/\ P(cl) \qquad (5)$$

2.4.2.2. <u>The metric in an identification task</u>

Any uncertainty about whether an object in $V(t)$ is a member of a certain category (i.e., whether or not it represents the concept) comes from one of two possible sources: (1) uncertainty related to the point - **object**, which may be due to objective inaccessibility of relevant information (e.g., we see a person from behind) or due to lack of attention and (2) uncertainty about **the concept**, which may be objective, as when the category is not well-defined (has "fuzzy boundaries," Zadeh LA, Fu KS, Tanaka K, Shimura M, eds. (1975); e.g., the concept "game") or subjective, as when the subject is not sure about the category's boundaries.

Information about concepts is not attention-dependent, so uncertainty about the concept may be a property of LTM [as opposed to $V(t)$]. Therefore, we will deal only with uncertainty about the point (inaccessibility or lack of attention as in Item 1 above). Namely, we consider only concepts that are 'sharp' in LTM. This uncertainty is expressed in our model by finite resolution and the clusters structure. For example, imagine that I am walking down my street at night when I briefly see a four-legged animal crossing the street. I can only see its size and get a vague idea of its form and gait. I do not know its color, whether it has a long or a short tail, or how big its ears are. Thus, essentially, it is represented by a cluster. Relying on the little information I have collected, I can now try to figure out what the animal may be (to associate it with a concept). My decision regarding its identity may be based on my knowledge about the cluster that represents it – a four-legged animal with a flowing gait that is the size of a large dog that is often seen in my neighborhood. Note that once the cluster boundaries have been determined, the probabilities that

it represents certain concepts are not attention-dependent (but rather statistically known). Thus, the contribution of attention to determining the probability that a cluster represents a candidate concept lies only in determining the boundaries of the cluster.

Put more formally, for a concept C in LTM and a cluster cl in $V(t)$, we are interested in two conditional probabilities: the probability of getting the attributes' values within the cluster cl, given that it is in the image of C - $P(cl/C)$, and the probability that a given cluster cl will represent the concept C - $P(C/cl)$. There is a natural connection between these two conditional probabilities that can be represented, in general, as:

$$P(C/cl) = {P(cl/C) \cdot P(C)}\Big/{P(cl)} \qquad (6)$$

where $P(C)$ is the probability of the concept C and $P(cl)$ is the probability of the cluster in the probability space S_{cl} defined above (formula (3) above).

Next, we would like to discuss the optimal metric needed to identify a concept C. This metric, which is attention-dependent, will determine the boundaries of a cluster cl and, subsequently, the probability that it represents the concept C.

Recall the example above, of the animal crossing the street. The cluster boundaries are built based on both inaccessibility of attributes' values (it was dark, so I could not determine the animal's color) and lack of attention (I could have made an effort to determine the animal's size more accurately). Metric manipulation can improve resolution only along accessible attributes. For an inaccessible attribute, any two objects that differ only along this attribute will lie in the same cluster. Namely, the metric cannot be influenced by this attribute (regardless of attention).

For example, if I am looking at pictures of flowers in a book, I cannot rely on my sense of smell to differentiate between two different flowers.

Our discussion regarding concept representation during identification tasks is based on the assumption that the representation is dictated by two conflicting challenges: (1) conserving resources and (2) avoiding mistakes in identification.

Our resource is resolution (recall that the number of clusters is limited). Therefore, in an identification task, resolution should be kept for distinguishing between objects that are associated with the concept and those that are not. For example, if we are looking for someone in a crowd and we know that that person is wearing a uniquely colored shirt, it would be an advisable strategy to focus on the color of the shirt. This is true even if we know what the person we are searching for looks like.

We look at concept-instantiation search as the prototype of an identification task. For example, if I am on a safari and looking for an elephant, I do not know which attributes are going to be accessible. Nevertheless, while I am looking for an elephant, I keep a representation of the concept ‘elephant’ in mind, with which to compare candidate objects.

Apparently, identification is often achieved based on very little information. For example, the presence of a trunk is almost sufficient to identify an animal as an elephant. However, such formative attributes are sometimes inaccessible (e.g., an elephant might turn its back to us). For this reason, each and every piece of information that can differentiate relevant objects from irrelevant objects is incorporated in an identification task (e.g., we do attend to size when trying to identify an elephant).

As mentioned, the metric challenge during an identification task is to be economical, but still accurate. Namely, sharp resolution should be kept for instances in which it is necessary. The set of accessible attributes is not foretold. Therefore, a good strategy in an identification task would be metric dependence on an attribute regardless of the values of other attributes.

Taking into account the discussion above, to achieve good identification, the general idea is as follows: The dependence of the metric on an attribute x_i (e.g., height) at a point a with a value a_i (e.g., 50 cm) on x_i should be proportional to how much a small change in x_i changes the probability of an object with a value $x_i = a_i$ (and unknown values of other attributes) to represent the concept C (e.g., dog).

Let us assume that x_i is the only accessible attribute during a task in which the concept C has to be identified. For a small $\delta > 0$, we define a set $\delta_{i,a_i} = \{x \in V(t) | \, x_i \in (a_i - \delta, a_i + \delta)\}$[16]. The set δ_{i,a_i} includes all of the objects in $V(t)$ with a value close to a_i on the coordinate x_i (e.g., all the animals that are around 50 cm tall). Now, we can calculate the probability that the set δ_{i,a_i} represents the concept C (e.g., given that an animal has a height of about 50 cm, what is the probability that it is a dog). This is achieved in the same way it was done for the cluster cl (see Formula 5). Namely, we refer to δ_{i,a_i} as a cluster and the probability is given as $P\left(C/\delta_{i,a_i}\right) = P\left(\delta_{i,a_i}/C\right) \cdot P(C) \Big/ P(\delta_{i,a_i})$. Next, we define a function $FC_i: V(t) \to \mathbb{R}$ as follows:

$$FC_i(a) = \lim_{\delta \to 0} P\left(C/\delta_{i,a_i}\right) \qquad (7)$$

[16] Note that the value δ is not metric-dependent, since it is at the level of the chart; namely, it is the Euclidean measure of $(a_i - \delta, a_i + \delta)$. In the example above, for instance, δ may be 1 cm. It does not say anything about the metric (whether or not 1 cm has much of an effect on the similarity).

(e.g., what is the probability that an animal whose height 'approaches' 50 cm is a dog).

The quantity we tried to express in words earlier (the effect of a small change in x_i on the probability that an object belongs to C) can now be expressed as $\frac{\partial FC_i}{\partial x_i}$. The desirable metric can now be optimally described by the **concept-identification metric** (involving the *k* attributes with counterparts in LTM). This metric is defined in a small neighborhood and we take a as any point in that neighborhood.

$$g_C(a) = I_C \cdot \begin{pmatrix} \left(\frac{\partial FC_1(a)}{\partial x_1}\right)^2 & \cdots & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \left(\frac{\partial FC_k(a)}{\partial x_k}\right)^2 & 0 & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \cdots & 0 \end{pmatrix} \quad (8)$$

This metric roughly represents the allocation of attention to attributes, pertaining to their effect on the probability of an association with C. The distance between a point A and a point B that differ only in their values of x_i represents the difference in the probability to be associated with C. I_C represents the degree of attention toward the identification task.

Note that there is a set CS included in $\theta(C)$ for which this matrix vanishes, namely, all diagonal elements in the above matrix are zero. We introduced above a set of k attributes in $V(t)$ with counterparts in LTM. Often only a subset of those is accessible (e.g., in the dark, I cannot determine the color of the animal crossing my street). For any subset of accessible attributes, there is a set of points that are associated with the concept C with maximal certainty. For any such subset of accessible attributes, this maximal certainty set includes CS (the set of vanishing metric).

As stated above, the concept-identification metric is active in any task requiring identification of a concept (e.g., `dog`) rather than differentiation between specific sub-concepts (`German Shepherd` vs. `Miniature Pinscher`) within a concept. We look at CS as the representation of the concept C in such tasks.

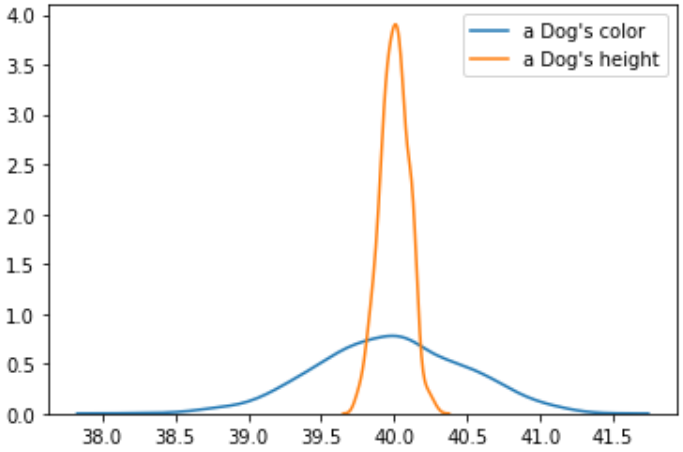
Recall that:

$$FC_i(a) = \lim_{\delta \to 0} P\left(C/\delta_{i,a_i}\right) = \lim_{\delta \to 0} {P\left(\delta_{i,a}/C\right) \cdot P(C)} \Big/ {P(\delta_{i,a})} = P(C) \cdot {\lim_{\delta \to 0} \frac{P(\delta_{i,a}/C)}{\delta}} \Big/ {\lim_{\delta \to 0} \frac{P(\delta_{i,a})}{\delta}}.$$

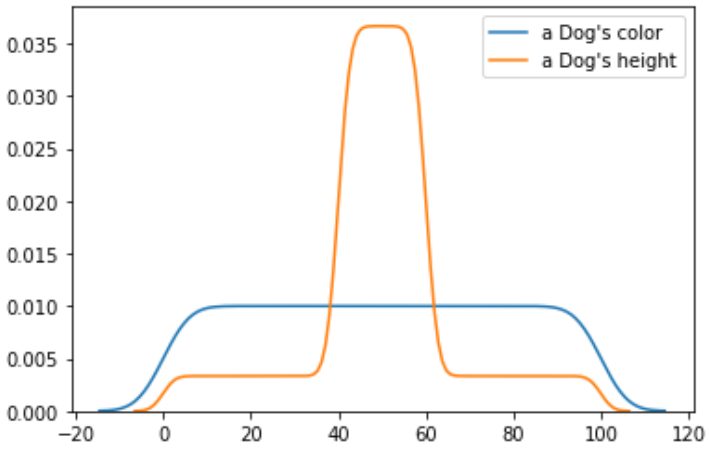
The limit $\lim_{\delta \to 0} \frac{P(\delta_{i,a}/C)}{\delta}$ is similar to the notion of a probability-density function of X_i in C.

From the definition of the concept-identification metric, it follows that, in general, the larger the range of X_i values for the concept C, the smaller the attribute's influence on the metric and the rougher its resolution. In cases in which any value is possible for a certain attribute, the attribute will not have any influence on the metric.

Look, for example, at the following illustration describing two distributions:



Dogs come in a wide variety of colors, so this attribute's distribution is flat; whereas the distribution of the height is rather concentrated around some typical height. Accordingly, the graph of the functions FC_i and FC_j, where C stands for the concept dog and i and j are color and height respectively, are described in the following illustration.



For a small range of height values (40 to 60 cm), the probability that an object is associated with the concept dog (given height only) does not change, even with respect to the possibility of other animals. Outside of this range, however, the probability depends heavily on height. For the attribute color, on the other hand, for practically the whole range, the probability does not change. It is clear that, in order to identify a dog, it is advisable to concentrate on height rather than color.

Here is another example: Suppose we are looking for baby shoes. Since the size of the shoes has a limited range (i.e., baby shoes), a lot of attention is paid to this attribute and the metric, accordingly, depends sharply on this attribute. On the other hand, the color of the shoes has no particular value and, consequently, little or no attention needs to be paid to that attribute.

Note that an interesting consequence of this process is dimension reduction. We have discussed this connection in detail, in identifying the concept *C,* in terms of the resolution along an attribute x_i (which is closely related to the attribute's influence on the metric) and the way the function $FC_i(a)$ (roughly, the probability that a point with a value a_i on x_i represents the concept C) varies along that attribute. It follows that for an attribute x_i for which $\frac{\partial F_C}{\partial x_i} = 0$ anywhere (an attribute not bearing on the concept's identification), we get $g_{Cii} = 0$, which means that the metric does not depend on x_i. Thus, we see dimension reduction of a vicinity of STM. The implication is that the concept is associated with a vicinity of lower dimension in STM.

2.5. Classification

Above, we dealt in detail with identification tasks. We now move on to a discussion of classification tasks.

Classification tasks can be carried out in a laboratory (Vriezen, Moscovitch & Bellos, 1995; Deng, & Sloutsky, 2015; Hoffman & Rehder, 2010; Yamauchi & Markman, 1998) or in a real-world setting. In a typical laboratory classification task, the subject is presented with an object and is asked to classify it into one of several predefined categories. Note that a real-world classification task is quite different. In a real-world classification task, the presented object is either recognized or is not recognized by the subject as representing a concept. This means that in real-world classification tasks: (a) there are no predefined choices (to check the object against); (b) we are dealing with concepts rather than arbitrary categories and (c) any number of "right" choices is possible (including zero). In the discussion below, we focus on real-world classification tasks.

We discussed above the representation of a concept in an identification task. Note that classification and identification, though closely related, are different tasks. In an identification task, the concept is represented a priori (e.g., I am on a safari looking for an elephant). In a classification task, on the other hand, the candidate categories are chosen in response to a perceived object and are assessed against it (e.g., on the safari, I see an animal and have to decide what it is). In particular, this means that the set of accessible attributes is given.

For example, I enter a room and find there a new object, which I have never seen before. Nevertheless, I know that this is a chair. I was not expecting to see the chair and so the metric was not modified to identify a chair.

Yet, another scenario is also possible: I entered the room, saw the chair and could not recognize it as a chair because it was not at all similar to the other chairs with which I am familiar. On the other hand, if I had looked for a chair (identification task), maybe I would have noticed that this object is good to sit on (I would have used the identification metric discussed above) and could have recognized it as a chair.

This example highlights two things about classification. First, classification depends on the metric in use, at the moment, in STM. Second, it is widely accepted that classification depends on the similarity of the classified object to other objects representing the candidate concept (exemplars and/or prototype).

2.5.1. Classification as a process

Classification, then, is clearly a process. This process is a composition of two sub-processes. One sub-process involves testing the suitability of the classified object to a certain concept with the metric at hand (e.g., to what degree is the object I see now similar to chairs I have known). The

first concept to test the object against is the one with the highest probability of being associated with the object, with respect to the given metric. The second sub-process involves the constant changing of the metric to suit the identification of the candidate concept (e.g., I see an object and modify the metric to determine whether that object is a chair).

During the process, the similarity of the classified object to other objects representing the candidate concept determines how the metric changes. In turn, the metric is used to compute the similarity of the classified object to other objects representing the candidate concept.

Staying with the example above, let us consider an object that I see. If it looks similar to chairs I have seen before, I will tend to pay more attention to those attributes that define `chair`; meaning that the metric changes toward being the one used for the identification of the concept `chair`. Suppose that the relevant attributes are: how comfortable the object is to sit on, the extent to which the object seems to have been made for the purpose of sitting and how portable the object is. Now, I am examining the object in the light of these attributes. If it is compatible with the representation of a chair, I will classify it as a chair. In other words, it will be included in the cluster representing the abstract concept of a chair (see the identification metric; Formula 8 above). In contrast, it is probable that by considering the attributes above, I will conclude that, while looking like a chair, the object should not be classified as a chair (e.g., it is not suitable to sit on or is not man-made). Since, in our construction, the metric changes continuously over time, a process is considered as evolving in time, rather than as a time-independent, discrete assessment.

2.5.2. Classification as an asymptotic process

We look to classify an observed Object x in terms of familiar concepts. First, a candidate concept C is chosen, according to the probability that it is associated with the object x. Then comes a process in which the proximity of the object to samples of (namely basic concepts associated with) the candidate concept C determines how the metric will evolve. In turn, the evolving metric signifies the above proximity. Therefore, it is natural to describe this process as a time variation of the metric imposed in $V(t)$. The process of classification should be performed within a limited time frame T_0. Therefore, it can be best described as an asymptotic process in which the evolution of g as $t \to T_0$ is given by $F\left(g_{(x,t),x}\right)$, where the independent variable is Time, t, and x (the object) is a parameter. When the observed relevant attribute values are fed into the function F, they generate a modification of the metric. Thus, if the distances between x and certain exemplars of the candidate concept are diminishing, then $g_{(x,t)}$ is approaching a 'verification metric' that will allow the object to be identified as being associated with the concept C.

To be more precise, we see the classification process as composed of a) choosing a candidate concept C (the concept with the highest probability of being associated with the object x) and b) verification of whether the object x really is associated with C. A classification task may include several of these verification cycles.

We now address in greater detail the way in which the attribute values of x are fed into F in the verification process. Recall our reference to the fact that the proximity of the classified object to exemplars of the concept C, at the beginning of the classification process, determines how the metric will evolve. However, the proximity of x to these exemplars is rather obscure (i.e., which

exemplars to consider, how the distances between *x* and these exemplars are summed, etc.). Therefore, it makes sense to also consider the probability that the cluster cl_x that includes the object x represents the concept C (see Formula 6 above). The probability we are interested in, then, is $P(C/cl_x)$. Note that this probability strongly depends on the metric $g_{(x,t)}$, since it is determined by the cluster boundaries which, in turn, are determined by the metric.

For example, suppose that at a certain moment, while seeing a Miniature Pinscher, I am more focused on its size and speed than on other features, such as its gait and shape. This means that different animals, which are similar in their size and speed, may be indistinguishable to me at that moment[17]. This may lead to the false classification of the Miniature Pinscher as a cat, since based on speed and size alone, the probability that the animal is a cat is greater than the probability that it is a dog. A closer look is often needed for better classification in cases like this one.

Note that in a classification task, as opposed to an identification task (as defined above), the set of accessible attributes is predetermined. By *set of accessible attributes* we mean those accessible attributes $x_1, x_2, \dots, x_m$ with counterparts in LTM (e.g., the size and the color of the object that may be a chair), namely the random variables $X_1, X_2, \dots, X_m$ about which I have information. The counterparts $x_{m+1}, x_{m+2}, \dots, x_k$ of the random variables $X_{m+1}, X_{m+2}, \dots, X_k$ are inaccessible (e.g., the weight of the object before I tried to lift it) and, therefore, every two points that differ only along attributes of the latter set lie within the same cluster.

We defined (in Section 2.4.2.1) the projection $\psi: V(t) \to \mathbb{R}^k$ as $\psi(x) = (x_1, x_2, \dots, x_k)$ for $x = (x_1, x_2, \dots, x_n)$, which maps x to an array of random variable values. We now define similarly a

[17] Namely, their images in $V(t)$ lie in the same cluster.

projection $\psi' : V(t) \to \mathbb{R}^m$ as $\psi'(x) = (x_1, x_2, .., x_m)$ for the set of accessible attributes (a subset of the *k* variables with counterparts in LTM). Namely, the image of the object x (e.g., the object that might be a chair) is an array indicating x's values on the accessible attributes only (e.g., color, shape, size, etc.). Here, x is the object that is being classified, which determines the cluster cl_x ($x \in cl_x \subset V(t)$). The sets $\psi'(\theta(C_i))$ for $C_i \in \Omega(cl_x)$ are the images in $\mathbb{R}^m$ of candidate basic concepts to be represented by x. Note that these images are not necessarily disjoint any more (as opposed to the sets $\psi(\theta(C_i))$). It forces us to consider the probabilistic structure that has been established above (see Section 2.4.2.1). We use Formula 5 for the probability of the cluster cl_x to represent the concept C. Now, $\psi'(cl_x)$ replaces $\psi(cl_x)$. The only difference is that now we have fewer random variables, meaning that the $\psi'(\theta(C_i))$ are not necessarily disjoint any more.

Since $\psi'(\theta(C_i))$ are not disjoint, even for a single point x of $V(t)$ there is no certainty with which basic concept x is associated. We define then, for a concept C, a function $F_C : V(t) \to \mathbb{R}$ as $F_C(x) = \lim_{\delta \to 0} P(C / B_\delta(x))$ [18]. [$F_C(x)$ is the limit probability that a small set around the point x of $V(t)$ is associated with the concept C.]

Next, we define a verification metric that is closely related to the identification metric above (Formula 8). The difference is that now the set of accessible attributes has been predetermined. Going back to to the example above, I enter a room and see an object that might be a chair. I do

---

[18] Note that $F_C(x) = \lim_{\delta \to 0} P(C / B_\delta(x))$ does not depend on the metric, since $\lim_{\delta \to 0} P(C / B_\delta(x)) = P(C) \cdot \lim_{\delta \to 0} \left(\frac{P(B_\delta / C) / \mu(B_\delta)}{P(B_\delta) / \mu(B_\delta)} \right) = \frac{\lim_{\delta \to 0} P(B_\delta / C) / \mu(B_\delta)}{\lim_{\delta \to 0} P(B_\delta) / \mu(B_\delta)}$. Here, $\mu(B_\delta)$ is the Euclidean measure of B_δ. The limits $\lim_{\delta \to 0} P(B_\delta / C) / \mu(B_\delta)$ dnd $\lim_{\delta \to 0} P(B_\delta) / \mu(B_\delta)$ do exist and do not depend on the metric. (This is our assumption when we consider the joint probability.)

have information about the object: its size, its shape and so forth. The probability that the object is a chair can now be calculated (see Formula 6 above) based on the cluster's boundaries. These boundaries are determined based on the values of the object's attributes and the metric at hand. The metric can now be modified to represent probability differences between objects. Namely, the greater the difference between two objects of $V(t)$ in terms of their probability of being associated with the concept C, the greater the distance between them. In particular, the metric should clearly distinguish between objects that are undoubtedly associated with the concept C and objects that are undoubtedly not associated with that concept. If the object x is a chair, judged with the verification metric, it should be 'distant' from the images of other concepts (e.g., `table` or `closet`).

Here again, for a point x, the dependence of the metric on an attribute x_i should express how much $F_C(x)$ changes along the attribute x_i around the point x. Recall that $F_C(x)$ is the limit probability for a small neighborhood around an object x to represent the concept C. This is expressed as $\frac{\partial F_C(x)}{\partial x_i}$, which represents the rate of change of F_C along the attribute x_i.

The verification metric, then, is given as:

$$g_{V_C}(x) = \begin{matrix} \frac{\partial F_C(x)}{\partial x_1}^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial F_C(x)}{\partial x_m}^2 \end{matrix} \qquad (10)$$

This metric is not time-dependent. Like the identification metric, it is an ad hoc metric used for the verification task. Note that g_{V_C} defined in this way degenerates in the interior of $\theta(C)$. More specifically, if z and y are in the interior of $\theta(C)$, then the distance between them, as determined

by g_{V_C}, is very close to zero. On the other hand, if z is in the interior of $\theta(C)$ while y is on the exterior, then the distance between them is very large.

As mentioned above, a classification task may include several cycles of verification, in cases where $\Omega(cl_x)$ (the set of concepts that their images intersect the cluster cl_x) contains more than one concept. For example, when I see the Miniature Pinscher, at first, I think it is a cat. I look more attentively, which means that I activate the verification metric for the concept `cat`. I then reject the concept `cat` and check the object against the concept `dog` and find a match, which means that with the verification metric of the concept `dog` I get a probability close to 1 that the cluster cl_x, which includes the object x, is associated with the concept `dog`.

Below, we describe the verification process in more detail.

For the sake of simplicity, we start with a case in which $\Omega(cl_x)$ contains a single concept C (e.g., the object that may or may not be a chair, but is definitely not any other concept I know). The classification process can be described by the rate of vanishing or divergence of $\left|g_{(x,t)} - g_{V_C}(x)\right|$ as t approaches some upper time limit T_0[19] (i.e., how fast I move to judge the object in view of the relevant attributes, for example, the degree to which the object is comfortable to sit on).

The convergence rate of $g_{(x,t)i,j} - g_{V_{Cij}}(x)$ is represented as: $F_{ij}\left(P\left(C/_{cl_x(g(x,t))}\right), g_{V_C}(x)\right)$, where F_{ij} is a complex function that cannot be defined should be characterized empirically.

[19] T_0 is the upper limit of the timeframe for classification, mentioned above.

Here, cl_x is the cluster that contains the object x. The functions F_{ij} (for any element ij of the matrix) express faster convergence of g to $g_{V_{Cij}}$ as $P(C/cl_x(g(x,t)))$ grows. Namely, the larger the probability that the classified object x is associated with C, the faster $g_{(x,t)}$ will converge to g_{V_C}. If x is associated with C, as $g_{(x,t)}$ converges to g_{V_C}, $P\left(C/cl_x(g(x,t))\right)$ will grow[20] and the convergence of $g_{(x,t)}$ to g_{V_C} will be accelerated via a positive feedback mechanism. If x is not associated with C, the movement of $g_{(x,t)}$ toward g_{V_C} will cause $P\left(C/cl_x(g(x,t))\right)$ to decrease and, as a result, the movement of $g_{(x,t)}$ toward g_{V_C} will be reversed and $g_{(x,t)}$ will diverge from g_{V_C}.

Note that the greater the similarity (the shorter the distance) of x to other objects representing C (exemplars), the larger $P\left(C/cl_x(g(x,t))\right)$ will get.

The process above can be described as asymptotic convergence:

$\left|g_{(x,t)i,j} - g_{V_{Cij}}(x)\right| = O((T_0 - t)^{\alpha_{ij}})$ as t approaches T_0 where α_{ij} is determined by F_{ij}.

[20] Note that the boundaries of cl_x and thus $P(C/cl_x(g(x,t)))$ depend on the metric.

2.5.3. <u>Termination of the process of verification</u>

Let the beginning of the verification process be at Time $t = 0$. If at $t = t_0 > 0$, $g(x,t)$ gets close enough to g_{V_C} while $P\left(C/_{cl_x(g(x,t))}\right)$ crosses some certainty threshold, then x will be classified as associated with C. It follows that in a case in which x is associated with C, t_0 is determined by $P\left(C/_{cl_x(g(x,t))}\right)$ and g at Time 0 (the beginning of the verification process). Namely, the larger $P\left(C/_{cl_x(g(x,t))}\right)$ is, given the metric at the beginning of the process, and the closer g is to g_{V_C}, the smaller t_0 will be. This means that the more 'typical' x is (the greater its similarity to exemplars of C), the faster the verification process will be.

If at any time t, $P\left(C/_{cl_x(g(x,t))}\right)$ decreases, then the verification process will terminate, x will not be classified as associated with C and another concept will be tested for verification. In a classification task involving several candidate concepts, the concept C for which $P\left(C/_{cl_x(g(x,0))}\right)$ is the largest is the one chosen to initiate the verification process for the object x.

Let us look at a simplified example of a verification process: I see a ball and I am not sure whether it is a soccer ball or a volleyball. I think that it is the type of ball used to play soccer. This type of ball is slightly bigger and its surface is smoother than that of a volleyball. And so the verification process begins. For the sake of simplicity, we ignore objective uncertainty and assume that uncertainty is caused only by poor resolution. So, in this example, C is a soccer ball, x is the ball I hold in my hand and $x \in \theta(C)$.

For the sake of simplicity, we make the following assumptions: First, since I know that the object I hold in my hand is a ball, $\Omega(cl_x)$ contains only balls, so the relevant probability subspace of LTM - S_{cl} (see Formula 3) is the subspace of balls (which means that shape is not an attribute in S_{cl}). For convenience, we assume that $V(t)$ consists of balls only at Time t, so the shape is a degenerate attribute, which means that only one value for this attribute exists in $V(t)$. Second, the only attributes relevant to determining whether the ball is a soccer ball are its size and the smoothness of its surface. Note that since there is no objective uncertainty (but only limited resolution expressed by the clusters structure), the image of a soccer ball and the image of a volleyball are disjointed.

Recall that:

1. The verification metric g_{V_C} defined in this way degenerates in the interior of $\theta(C)$. More specifically, if z and y are in the interior of $\theta(C)$, then the distance between *z* and *y*, as determined by g_{V_C}, is very close to zero. On the other hand, if z is in the interior of $\theta(C)$ while y is on the exterior, then, because there is no objective uncertainty, by definition, the distance between *z* and *y* approaches infinity as g approaches g_{V_C}.
2. cl_x is an ε −ball with respect to the metric $g(x,t)$.
3. By Formula 5: $P\left(C/_{cl}\right) = \frac{P\left(X\in\psi(cl)\cap\psi(\theta(C))\right)}{P\left(X\in\psi(cl)\right)}$.

For the limit metric g_{V_C}, for any point *p* such that $p \in V(t)\backslash\theta(C)$, we get $p \notin B_\varepsilon\left(g_{V_C}(x)\right) = cl_x(g_{V_C}(x))$, while, on the other hand, $\theta(C) \subseteq cl_x(g_{V_C}(x))$, which indicates perfect separation. We assume, then, that $\theta(C)$ for any g close enough to g_{V_C} is also a ball with x as its center, so $P\left(\psi\left(cl_x\left(g(x,t)\right)\right)\right) = P\left(\psi\left(\theta(C)\right)\right) + P\left(\psi\left(cl_x\left(g(x,t)\right)\right)\backslash\psi\left(\theta(C)\right)\right)$. Therefore,

$P\left(C/_{cl_x}\right) = 1 - P\left(\psi\left(cl_x(g(x,t))\right)\backslash\psi(\theta(C))\right)$. We assume that there is some density function f and a measure μ in $\mathbb{R}^2$ such that $P\left(\psi\left(cl_x(g(x,t))\right)\backslash\psi(\theta(C))\right) = \iint_{\psi\left(cl_x(g(x,t))\right)\backslash\psi(\theta(C))} f\, d\mu = \int_{cl_x(g(x,t))\backslash\theta(C)} f/\sqrt{|det\, g\ (x,t\ )|}\, dx$

for any $\varepsilon' > 0$ there is $\delta > 0$ such that when

$|g - g_{V_C}| < \delta$, we get $1 - P\left(C/_{cl_x(g(x,t))}\right) < \varepsilon'$.

Therefore, $P\left(C/_{cl_x(g(x,t))}\right)$ increases as $|g - g_{V_C}|$ decreases. It is reasonable, then, for the sake of our simple example, to assume that[21] $P\left(C/_{cl_x(g(x,t))}\right) = 1 - |g(x,t) - g_{V_C}(x)|$. The convergence of $g_{(x,t)i,j}$ to $g_{V_{Cij}}$ at a rate that is inversely related to $P\left(C/_{cl_x(g(x,t))}\right)$ can be depicted in the following specific equation $\frac{\partial|g_{(x,t)i,j} - g_{V_{Cij}}(x)|}{\partial t} = -\frac{1}{1-P\left(C/_{cl_x(g(x,t))}\right)} = 1/-|g - g_{V_C}|$ where $F_{ij}\left(P\left(C/_{cl_x(g(x,t))}\right), g_{V_C}(x)\right) = -\frac{1}{1-P\left(C/_{cl_x(g(x,t))}\right)} = 1/-|g - g_{V_C}|$.

In this equation, the closer $P\left(C/_{cl_x(g(x,t))}\right)$ is to 1, the more negative the time derivative of $\left|g_{(x,t)i,j} - g_{V_{Cij}}(x)\right|$. For further simplicity, we take $g_{(x,t)}$ to be diagonal with $g_{(x,t)1,1} = g_{(x,t)2,2}$ at any t.

[21] To make things simpler, we included many approximations and inaccuracies in this example. Note that according to the way we developed the formula for the probability $P\left(C/_{cl_x(g(x,t))}\right)$, it does depend on the metric. But, here, we take the liberty to assume that it depends only on the metric at x. Nevertheless, in principle, this formula seems reasonable.

The solution of this equation is $\left|g_{(x,t)i,j} - g_{V_{Cij}}(x)\right| = \sqrt{2T - 2t}$, where T is a constant smaller than T_0. T is determined by the initial conditions (i.e., the metric and the probability that x is associated with C, at the beginning of the verification process). As can be seen, as t approaches T, $g_{(x,t)i,j}$ approaches $g_{V_{Cij}}(x)$ at an increasing rate. In turn, $P\left({}^{C}/_{cl_x(g(x,t))}\right)$ approaches 1. The smaller T is, the faster the verification process will progress.

3. The Model: Summary

We propose a novel model of similarity in which STM can be visualized as a (multidimensional) flexible sheet (the 'patch') with coordinates drawn on it. This flexible sheet can be stretched over different domains and in different directions, resulting in varying representations. The sheet moves continuously with respect to time. Finally, the accessible information is a discrete approximation of this flexible sheet. In particular, we tried to demonstrate how the task at hand dictates the representation and, thus, the metric.

We believe that this view of similarity may serve to defend the metric approach. Criticisms of the metric (Tversky 1977, Tversky & Gatti 1978, Tversky & Gatti 1981) approach for similarity modeling derive mostly from the evident violations of metric axioms, which, in general, make the metric approach psychologically inaccurate. We tackle this problem by incorporating the LTM- STM distinction. We assume that varying representation occurs in STM, which is, in our view, the metric environment in which similarity operates, while LTM is a (non-metric) probability space which functions as database and in which similarity is irrelevant. In this way, we avoid previously studied violations of those metric axioms that deal with multiple distances, 'calculated' in different contexts.

References

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. *Psychonomic Bulletin & Review*, *6*(2), 328-337.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological review*, *93*(2), 154.

Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long-and short-term memory. *Journal of verbal learning and verbal behavior*, *9*(2), 176-189.

Borg, I., & Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media.

Buchsbaum, B. R., Padmanabhan, A., & Berman, K. F. (2011). The neural substrates of recognition memory for verbal information: spanning the divide between short-and long-term memory. *Journal of Cognitive Neuroscience*, *23*(4), 978-991.

Corkin, S. (2002). What's new with the amnesic patient H.M.? Nature Reviews Neuroscience, 3, 153–160.

Deng, W. S., & Sloutsky, V. M. (2015). The development of categorization: effects of classification and inference training on category representation. *Developmental Psychology*, *51*(3), 392.

Dzhafarov, E. N., & Colonius, H. (1999). Fechnerian metrics in unidimensional and multidimensional stimulus spaces. *Psychonomic bulletin & review*, *6*(2), 239-268.

Fung, L. W., & Fu, K. S. (1975). An axiomatic approach to rational decision making in a fuzzy environment. In *Fuzzy sets and their applications to Cognitive and decision processes* (pp. 227-256). Academic Press.

Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 271-280.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*(2), 319.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297-303.

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density.

Laub, J., Müller, K. R., Wichmann, F. A., & Macke, J. H. (2006). Inducing metric violations in human similarity judgements. *Advances in neural information processing systems*, *19*.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological review*, *100*(2), 254.

Nachshon, Y., Cohen, H. & Maril, A. (2022). Semantic distance- justification and limitations. *In preparation.*

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39. review, 93(2), 154.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, *43*(1), 25-53.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Squire, L. R. (2009). Memory and brain systems: 1969–2009. Journal of Neuroscience, 29, 12711–12716.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(4), 629-640.

Townsend, J. T., Burns, D., & Pei, L. (2013). The prospects for measurement in infinite-dimensional psychological spaces. *Measurement with Persons*, 143-174.

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

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Loyd (Eds.), Cognition and categorization (pp. 79-98). Hillsdale, NJ: L. Erlbaum.

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

Tversky, A., & Krantz, D. H. (1970). The dimensional representation and the metric structure of similarity data. *Journal of mathematical psychology*, *7*(3), 572-596.

Vallar, G., & Baddeley, A. (1984). Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. Journal of Verbal Learning and Verbal Behavior, 23, 151–161.

Voorspoels, W., Vanpaemel, W., & Storms, G. (2011). A formal ideal-based account of typicality. *Psychonomic bulletin & review*, *18*(5), 1006-1014.

Vriezen, E. R., Moscovitch, M., & Bellos, S. A. (1995). Priming effects in semantic classification tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 933.

Warrington, E., & Shallice, T. (1969). Selective impairment of auditory verbal short-term memory. Brain, 92, 885–896.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and language*, *39*(1), 124-148.

Yearsley, J. M., Barque-Duran, A., Scerrati, E., Hampton, J. A., & Pothos, E. M. (2017). The triangle inequality constraint in similarity judgments. *Progress in biophysics and molecular biology*, *130*, 26-32.