International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 17, No. 5 (2003) 885–901

World Scientific
www.worldscientific.com

HUMAN UNSUPERVISED AND SUPERVISED LEARNING AS A QUANTITATIVE DISTINCTION

TODD M. GURECKIS* and BRADLEY C. LOVE†

Department of Psychology, University of Texas at Austin,
1 University Station A8000, Austin, TX 78712-0187, USA
**gureckis@love.psy.utexas.edu*
†*love@love.psy.utexas.edu*
http://love.psy.utexas.edu/

SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network) is a network model of human category learning. SUSTAIN initially assumes a simple category structure. If simple solutions prove inadequate and SUSTAIN is confronted with a surprising event (e.g. it is told that a bat is a mammal instead of a bird), SUSTAIN recruits an additional cluster to represent the surprising event. Newly recruited clusters are available to explain future events and can themselves evolve into prototypes/attractors/rules. SUSTAIN has expanded the scope of findings that models of human category learning can address. This paper extends SUSTAIN to account for both supervised and unsupervised learning data through a common mechanism. The modified model, uSUSTAIN (unified SUSTAIN), is successfully applied to human learning data that compares unsupervised and supervised learning performances.[18]

Keywords: Category; learning; unsupervised; supervised; psychology.

1. Introduction

Categories provide a crucial function underlying the cognitive abilities of humans. They allow us to generalize our knowledge to novel situations and to infer unknown properties of the environment. These abilities are indispensable to any intelligent system.

Researchers studying human categorization have traditionally focused on human performance in supervised learning tasks (see Refs. 2, 4 and 7 for some exceptions). In this experimental paradigm, subjects learn to classify stimuli as members of contrastive categories through trial by trial learning with corrective feedback. Theories (and models) of learning are favored that can account for the relative difficulty of acquiring different category structures.[25,30]

Although classification learning does capture aspects of human learning, others are not addressed by this paradigm. For instance, humans can spontaneously construct categories in the absence of feedback. As an example, many of us have created the categories "interesting" email and "junk" email in the absence of

explicit feedback. Such learning is referred to as unsupervised learning. Supervised and unsupervised learning are often seen as being qualitatively different. Supervised learning is characterized as intentional, in that learners actively search for rules (perhaps by hypothesis testing) and are explicitly aware of the rule they are considering.[26] On the other hand, unsupervised learning is seen as an incidental, undirected, stimulus driven, and incremental accrual of information.[3,8,13,14,17]

In contrast to this view, Love[18] has found that intentional unsupervised learning performance is more similar to supervised learning performance than it is to incidental unsupervised learning performance. This result suggests that the unsupervised/supervised dichotomy may not be valid. Gureckis and Love[10] have argued that unsupervised and supervised learning can be modeled through a common mechanism. However, our account has yet to model to a direct comparison between supervised and unsupervised learning. Here, we apply Gureckis and Love's[10] variant of the SUSTAIN (Supervised and Unsupervised STratified Adaptive Incremental Network) model, referred to as uSUSTAIN (unified SUSTAIN), to the Love[18] data uSUSTAIN differs from other models that seek to unify unsupervised and supervised learning, such as Anderson's[1] rational model, in that uSUSTAIN is applicable to both unsupervised and supervised learning tasks while not predicting that these tasks lead to equivalent performance (which they do not). In the remainder of this paper, we overview SUSTAIN and uSUSTAIN. We then fit uSUSTAIN to the Love[18] data and consider the implications of the simulations.

2. The Modeling Approach: SUSTAIN and uSUSTAIN

SUSTAIN has been successfully applied to an array of challenging human data sets spanning a variety of category learning paradigms including classification learning,[21] learning at different levels of abstraction,[20] inference learning,[19] and unsupervised learning.[11,22]

In the following sections, we discuss SUSTAIN's operation, its underlying principles, and the mathematical equations that follow from these principles. We then introduce a modification to SUSTAIN that enables it to account for supervised and unsupervised learning data through a single recruitment mechanism. This mechanism makes use of an intuitive and general notion of *surprise* to facilitate learning. This modified version of SUSTAIN is referred to as uSUSTAIN.

2.1. *Overview of model*

SUSTAIN is a network model of human category learning. On each learning trial, SUSTAIN takes as input a description of the current stimulus item represented to the model by a set of perceptual feature dimensions. For example, a stimulus item depicting a large, purple square will be represented to the model by the feature dimensions color, size and stripe. Like other models of category learning (such as Ref. 1), SUSTAIN treats the category membership (or category label) of a stimulus item as another stimulus feature dimension. SUSTAIN maintains a selective

attention mechanism which allows it to learn to focus attention on stimulus dimensions that are particularly useful for the current categorization task (similar to Ref. 16).

The internal representations in the model consist of a set of clusters. Categories are represented in the model as one or more associated clusters. Initially, the network has only one cluster that is centered upon the first input pattern. As new stimulus items are presented, the model attempts to assign these new items to an existing cluster. This assignment is done through an unsupervised procedure based on the similarity of the new item to the stored clusters. When a new item is assigned to a cluster, the cluster updates its internal representation to become the average of all items assigned to the cluster so far.

However, if SUSTAIN discovers through feedback that this similarity-based assignment is incorrect, a new cluster is created to encode the current item as an exception (for a concrete example of this see Principle 3 in the following section). In unsupervised learning tasks there is no corrective feedback, so instead SUSTAIN creates a new cluster if the current stimulus item is not sufficiently similar to any existing clusters (the threshold for this sufficiency is controlled by a parameter in the model). Both of these cluster recruitment strategies are unified under the principe of "adaptation to surprise".[10] In supervised learning, SUSTAIN creates a new cluster in response to a surprising misclassification, whereas in unsupervised learning, a new cluster is created when the model encounters a surprisingly novel stimulus item.

Clusters compete with each other to respond to the current stimulus item. The cluster that wins this competition passes its activation over connection weights to a set of output units. These output units replicate the structure of the input dimensions. The connection weights are adjusted over the course of learning so that the association between each cluster and the appropriate response for members of that cluster is strengthened. For example, a cluster whose members are mostly in category "A" would develop over the course of learning a stronger connection to the category "A" output unit than to the category "B" output unit. The activation of an output unit is proportional to the strength of the activation passed from the winning cluster and the strength of the connection weight. SUSTAIN's ultimate response is biased towards the most activated output unit. In this way, classification decisions are ultimately based on the cluster to which an instance is assigned.

2.2. *The key principles of SUSTAIN*

With this general understanding of the operation of the model in mind, we now examine the five key principles of SUSTAIN. These principles highlight the important features of the model and provide the foundation for the model's formalism.

2.2.1. *Principle 1, SUSTAIN is biased towards simple solutions*

SUSTAIN is initially directed towards simple solutions. At the start of learning, SUSTAIN has only one cluster which is centered on the first input item. It then

adds clusters (i.e. complexity) only as needed to accurately describe the category structure. Like other models of category learning (e.g. Ref. 16), SUSTAIN learns to selectively attend to stimulus feature dimensions that are most useful for categorization. This focus on a subset of stimulus dimensions also serves to bias SUSTAIN towards simple solutions.

2.2.2. *Principle 2, similar stimulus items tend to cluster together*

In learning to classify stimuli as members of two distinct categories, SUSTAIN will cluster similar items together. For example, different instances of a bird subtype (e.g. sparrows) could cluster together and form a sparrow cluster instead of leaving separate traces in memory for each instance. Clustering is an unsupervised process because cluster assignment is done on the basis of similarity, not feedback.

2.2.3. *Principle 3, SUSTAIN learns in both a supervised and unsupervised fashion*

In learning to classify the categories "birds" and "mammals", SUSTAIN relies on both unsupervised and supervised learning processes. Consider a learning trial in which SUSTAIN has formed a cluster whose members are small birds, and another cluster whose members are four-legged mammals. If SUSTAIN is subsequently asked to classify a bat, it will initially predict that a bat is a bird on the basis of overall similarity (bats and birds are both small, have wings, fly, etc.). Upon receiving feedback from the environment (supervision) indicating that a bat is a mammal, SUSTAIN will recruit a new cluster to represent the bat as an exception to the mammal category. The next time SUSTAIN is exposed to the bat or another similar bat, SUSTAIN will correctly predict that a bat is a mammal. This example also illustrates how SUSTAIN can entertain more complex solutions when necessary through cluster recruitment (see Principle 1).

2.2.4. *Principle 4, the pattern of feedback matters*

As the example used above illustrates, feedback affects the inferred category structure. Prediction failures result in a cluster being recruited, thus different patterns of feedback can lead to different representations being acquired. This principle allows SUSTAIN to predict different acquisition patterns for different learning modes that are informationally equivalent but differ in their pattern of feedback. The learning conditions in the Love[18] study considered in this paper are informationally equivalent, but differ in their pattern of feedback.

2.2.5. *Principle 5, cluster competition*

Clusters can be seen as competing explanations of the input. The strength of the response from the winning cluster (the cluster the current stimulus is most similar

to) is attenuated in the presence of other clusters that are somewhat similar to the current stimulus (see Ref. 31, account of competing explanations in reasoning).

2.3. *Mathematical formulation of SUSTAIN*

This section of the paper explains how the general principles that govern SUSTAIN's operation are implemented in an algorithmic model.

2.3.1. *Input representation*

Stimuli are represented in the model as vector frames where the dimensionality of the vector is equal to the dimensionality of the stimuli. The category label is also included as a stimulus dimension. Thus, stimuli that vary on three perceptual dimensions (e.g. size, shape and color) and are members of one of two categories would require a vector frame with four dimensions. A four-dimensional binary-valued stimulus (three perceptual dimensions plus the category label) can be thought of as a four character string (e.g. **1 2 1 1**) in which each character represents the value of a stimulus dimension. For example, the first character could denote the size dimension with a **1** indicating a small stimulus and a **2** indicating a large stimulus.

Of course, a learning trial usually involves an incomplete stimulus representation. For instance, in classification learning all the perceptual dimensions are known, but the category label dimension is unknown and queried. After the learner responds to the query, corrective feedback is provided. Assuming the fourth stimulus dimension is the category label dimension, the classification trial for the above stimulus is represented as **1 2 1 ?** → **1 2 1 1**.

On every classification trial, the category label dimension is queried and corrective feedback indicating the category membership of the stimulus is provided. In contrast, on inference learning trials, subjects are given the category membership of the item, but must infer an unknown stimulus dimension. Possible inference learning trials for the above stimulus description are **? 2 1 1** → **1 2 1 1**, **1 ? 1 1** → **1 2 1 1**, and **1 2 ? 1** → **1 2 1 1**. Notice that inference and classification learning provide the learner with the same stimulus information after feedback (though the pattern of feedback varies).

Unsupervised learning does not involve informative feedback. In unsupervised learning, every item is considered to be a member of the same global category. Thus, the category label dimension is unitary valued and uninformative for differentiating between stimuli. However, the degree to which any particular stimulus activates this category dimension indicates the degree to which the network recognizes the stimulus.

In order to represent a nominal stimulus dimension that can display multiple values, SUSTAIN devotes multiple input units. To represent a nominal dimension containing k distinct values, k input units are utilized. All the units forming a dimension are set to zero, except for the one unit that denotes the nominal value

of the dimension (this unit is set to one). For example, the stimulus dimension of marital status has three values ("single", "married", "divorced"). The pattern [0 1 0] represents the dimension value of "married". A complete stimulus is represented by the vector $I^{\text{pos}_{ik}}$ where i indexes the stimulus dimension and k indexes the nominal values for dimension i. For example, if marital status was the third stimulus dimension and the second value was present (i.e. married), then $I^{\text{pos}_{32}}$ would equal one, whereas $I^{\text{pos}_{31}}$ and $I^{\text{pos}_{33}}$ would equal zero. The "pos" in I^{pos} denotes that the current stimulus is located at a particular position in a multidimensional representational space.

2.3.2. *Receptive fields*

Each cluster has a receptive field for each stimulus dimension. A cluster's receptive field for a given dimension is centered at the cluster's position along that dimension. The position of a cluster within a dimension indicates the cluster's expectations for its members.

The tuning of a receptive field (as opposed to the position of a receptive field) determines how much attention is being devoted to the stimulus dimension. All the receptive fields for a stimulus dimension have the same tuning (i.e. attention is dimension-wide as opposed to cluster-specific). A receptive field's tuning changes as a result of learning. This change in receptive field tuning implements SUSTAIN's selective attention mechanism. Dimensions are highly attended to develop peaked tunings, whereas dimensions are not well attended to develop broad tunings. Dimensions that provide consistent information at the cluster level receive greater attention.

Mathematically, receptive fields have an exponential shape with a receptive field's response decreasing exponentially as distance from its center increases. The activation function for a dimension is:

$$\alpha(\mu) = \lambda e^{-\lambda\mu} \tag{1}$$

where λ is the tuning of the receptive field, μ is the distance of the stimulus from the center of the field, and $\alpha(\mu)$ denotes the response of the receptive field to a stimulus falling μ units from the center of the field. The choice of exponentially shaped receptive fields is motivated by Shepard's[29] work on stimulus generalization.

Although receptive fields with different λ have different shapes (ranging from a broad to a peaked exponential), for any λ, the area "underneath" a receptive field is constant:

$$\int_0^\infty \alpha(\mu)d\mu = \int_0^\infty \lambda e^{-\lambda\mu}d\mu = 1\,. \tag{2}$$

For a given μ, λ that maximizes $\alpha(\mu)$ can be computed from the derivative:

$$\frac{\partial\alpha}{\partial\lambda} = e^{-\lambda\mu}(1-\lambda\mu)\,. \tag{3}$$

These properties of exponentials prove useful in formulating SUSTAIN.

2.3.3. *Cluster activation*

With nominal stimulus dimensions, the distance μ_{ij} (from 0 to 1) between the ith dimension of the stimulus and cluster j's position along the ith dimension is:

$$\mu_{ij} = \frac{1}{2} \sum_{k=1}^{v_i} |I^{\text{pos}_{ik}} - H_j^{\text{pos}_{ik}}| \tag{4}$$

where v_i is the number of different nominal values on the ith dimension, I is the input representation (as described in a previous section), and $H_j^{\text{pos}_{ik}}$ is cluster j's position on the ith dimension for value k (the sum of all k for a dimension is 1). The position of a cluster in a nominal dimension is actually a probability distribution that can be interpreted as the probability of displaying a value given that an item is a member of the cluster. For example, a cluster in which 20% of the members are single, 45% are married, and 35% are divorced will converge to the location [0.20 0.45 0.35] within the marital status dimension. The distance μ_{ij} will always be between 0 and 1 (inclusive).

The activation of a cluster is given by:

$$H_j^{\text{act}} = \frac{\sum_{i=1}^{m} (\lambda_i)^r e^{-\lambda_i \mu_{ij}}}{\sum_{i=1}^{m} (\lambda_i)^r} \tag{5}$$

where H_j^{act} is the activation of the jth cluster, m is the number of stimulus dimensions, λ_i is the tuning of the receptive field for the ith input dimension, and r is an attentional parameter (always non-negative). When r is large, input units with tighter tunings (units that seem relevant) dominate the activation function. Dimensions that are highly attended have larger λs and will have greater importance in determining the clusters' activation values. Increasing r simply accentuates this effect. If r is set to zero, every dimension receives equal attention. Equation (5) sums up the responses of the receptive fields for each input dimension and normalizes the sum (again, highly attended dimensions weigh heavily). Cluster activation is bound between 0 (exclusive) and 1 (inclusive). Unknown stimulus dimensions (e.g. the category label in a classification trial) are not included in the above calculation.

2.3.4. *Competition*

Clusters compete to respond to input patterns and in turn inhibit one another. When many clusters are strongly activated, the output of the winning cluster H_j^{out} is less:

$$\begin{aligned} &\text{For the winning } H_j \text{ with the greatest } H^{\text{act}},\ H_j^{\text{out}} = \frac{(H_j^{\text{act}})^\beta}{\sum_{i=1}^{n} (H_i^{\text{act}})^\beta} H_j^{\text{act}} \\ &\text{For all other } H_j,\ H_j^{\text{out}} = 0\,, \end{aligned} \tag{6}$$

where n is the number of clusters and β is the lateral inhibition parameter (always non-negative) that regulates cluster competition. When β is small, competing clusters strongly inhibit the winner. When β is large the winner is weakly inhibited.

Clusters other than the winner have their output set to zero. Equation (6) is a straightforward method for implementing lateral inhibition. It is a high level description of an iterative process where units send signals to each other across inhibitory connections. Psychologically, Eq. (6) signifies that competing alternatives will reduce confidence in a choice (reflected in a lower output value).

2.3.5. *Response*

Activation is spread from the clusters to the output units of the queried (the unknown) stimulus dimension z:

$$C_{zk}^{\text{out}} = \sum_{j=1}^{n} w_{j,zk} H_j^{\text{out}} \tag{7}$$

where C_{zk}^{out} is the output of the output unit representing the kth nominal value of the queried (unknown) zth dimension, n is the number of clusters, and $w_{j,zk}$ is the weight from cluster j to category unit C_{zk}. A winning cluster (especially one that did not have many competitors and is similar to the current input pattern) that has a large positive connection to an output unit will strongly activate the output unit. The summation in the above calculation is not really necessary given that only the winning cluster has a nonzero output, but is included to make the similarities between SUSTAIN and other models more apparent.

The probability of making response k (the kth nominal value) for the queried dimension z is

$$Pr(k) = \frac{e^{(d \cdot C_{zk}^{\text{out}})}}{\sum_{j=1}^{v_z} e^{(d \cdot C_{zj}^{\text{out}})}} \tag{8}$$

where d is a response parameter (always non-negative) and v_z is the number of nominal units (and hence output units) forming the queried dimension z. When d is high, accuracy is stressed and the output unit with the largest output is almost always chosen. The Luce choice rule is conceptually related to this decision rule.[23]

2.3.6. *Learning*

After responding, feedback is provided to SUSTAIN. The target value for the kth category unit of the queried dimension z is:

$$t_{zk} = \left\{ \begin{array}{l} \max(C_{zk}^{\text{out}}, 1), \text{ if } I^{\text{pos}_{zk}} \text{ equals } 1 \\ \min(C_{zk}^{\text{out}}, 0), \text{ if } I^{\text{pos}_{zk}} \text{ equals } 0 \end{array} \right\}. \tag{9}$$

Kruschke[16] refers to this kind of teaching signal as a "humble teacher" and explains when its use is appropriate. Basically, the model is not penalized for predicting the correct response more strongly than is necessary.

A new cluster is recruited if the winning cluster predicts an incorrect response. In the case of a supervised learning situation, a cluster is recruited according to the following procedure:

$$\begin{aligned}&\text{For the queried dimension } z, \text{ if } t_{zk} \text{ does not equal 1 for the } C_{zk}\\ &\text{with the largest output } C_{zk}^{\text{out}} \text{ of all } C_{z*}, \text{ then recruit a new cluster}.\end{aligned} \tag{10}$$

In other words, the output unit representing the correct nominal value must be the most activated of all the output units forming the queried stimulus dimension.

In the case of an unsupervised learning situation, SUSTAIN is self-supervising and recruits a cluster when the most activated cluster H_j's activation is below the threshold τ:

$$\text{if } (H_j^{\text{act}} < \tau), \text{ then recruit a new cluster}. \tag{11}$$

Unsupervised recruitment in SUSTAIN bears a strong resemblance to recruitment in Adaptive Resonance Theory,[5] Clapper and Bower's qualitative model,[6] and Hartigan's leader algorithm.[12]

When a new cluster is recruited it is centered on the misclassified input pattern and the clusters' activations and outputs are recalculated. The new cluster then becomes the winner because it will be the most highly activated cluster (it is centered upon the current input pattern — all μ_{ij} will be zero). Again, SUSTAIN begins with a cluster centered on the first stimulus item.

The position of the winner is adjusted:

$$\text{For the winning } H_j\,,\ \Delta H_j^{\text{pos}_{ik}} = \eta(I^{\text{pos}_{ik}} - H_j^{\text{pos}_{ik}}) \tag{12}$$

where η is the learning rate. The centers of the winner's receptive fields move towards the input pattern according to the Kohonen learning rule.[15] This learning rule centers the cluster amidst its members.

Using our result from Eq. (3), receptive field tunings are updated according to:

$$\Delta\lambda_i = \eta e^{-\lambda_i \mu_{ij}}(1 - \lambda_i \mu_{ij}) \tag{13}$$

where j is the index of the winning cluster.

Only the winning cluster updates the value of λ_i. Equation (13) adjusts the peakedness of the receptive field for each input so that each input dimension can maximize its influence on the clusters. Initially, λ_i is set to be broadly tuned with a value of 1. The value of 1 is chosen because the maximal distance μ_{ij} is 1 and the optimal setting of λ_i for this case is 1 (i.e. Eq. (13) equals zero). Under this scheme, λ_i cannot become less than 1, but can become more narrowly tuned.

When a cluster is recruited, weights from the unit to the output units are set to zero. The one layer delta learning rule[32,28] is used to adjust these weights:

$$\Delta w_{j,zk} = \eta(t_{zk} - C_{zk}^{\text{out}})H_j^{\text{out}}\,, \tag{14}$$

where z is the queried dimension. Note that only the winning cluster will have its weights adjusted since it is the only cluster with a nonzero output.

2.4. *uSUSTAIN: a unified approach to supervised and unsupervised learning*

SUSTAIN can model both supervised and unsupervised learning, but it relies on different recruitment mechanisms. In both cases, a cluster is recruited in response to a surprising event (i.e. the existing cluster structure does not properly characterize the current stimulus), but how a surprising event is defined differs. In the supervised case, the surprising event is a prediction error, whereas in the case of unsupervised learning the surprising event is an unfamiliar stimulus.

Although the two separate recruitment procedures have been successful, a single recruitment procedure is preferable. Beyond parsimony, a unified account could prove useful in clarifying the relationship between unsupervised and supervised learning. A simple way to integrate the two recruitment strategies is to generalize the unsupervised procedure so that it is applicable to supervised learning situations. Under this scheme, a new cluster is recruited when the current stimulus is not sufficiently similar to any cluster in its category:

$$\begin{aligned} &\text{For the queried dimension } z\,, \\ &\text{If } \mathrm{Max}(\{H_j^{\mathrm{act}} \mid \mu_{zj} = 0\}) < \tau, \text{ then recruit a new cluster}\,, \end{aligned} \tag{15}$$

where H_j^{act} is the activation of cluster j, μ_{zj} is the distance [as defined in Eq. (4)] along the zth dimension of the current stimulus and cluster j's position along the zth dimension, and τ is a constant between 0 and 1 (a parameter). The requirement that μ_{zj} be zero specifies that only clusters associated with the category of the current stimulus are considered. In unsupervised learning, all items belong to the same global category which represents items the network has seen before. Thus, $\mathrm{Max}(\{H_j^{\mathrm{act}} \mid \mu_{zj} = 0\})$ refers to the most activated cluster overall. In supervised learning, the most activated cluster predicting the correct category may not be the most activated cluster overall.

Besides providing a unified framework, this recruitment strategy has a number of other virtues over SUSTAIN's original recruitment rule [Eq. (10)] for supervised learning. For example, the unified procedure will recruit a new cluster when an unusual item is encountered that does not result in a prediction error whereas the previous error-driven recruitment scheme would not recruit a new cluster to encode the unusual item. Assigning a very unusual item to an existing cluster (a cluster the item is not very similar to) could result in catastrophic interference (see Ref. 27) as the cluster must undergo radical change to accommodate its newest member.

3. Evaluating uSUSTAIN

In order to evaluate this unified formulation of the model we applied uSUSTAIN to the studies previously accounted for using separate cluster recruitment mechanisms for supervised and unsupervised learning.[10] It is important to recognize that the recruitment procedure that uSUSTAIN uses is, in fact, a generalization of unsupervised recruitment procedure used by the original SUSTAIN model. Thus,

uSUSTAIN and SUSTAIN provide equivalent accounts of unsupervised learning.[9] uSUSTAIN and SUSTAIN have fit a number of unsupervised learning studies and have generated novel predictions that have been subsequently tested and confirmed with human subjects.[11]

A true test of generality of the uSUSTAIN approach lies in its ability to fit supervised learning data. Gureckis and Love[9] applied uSUSTAIN to a number of supervised learning studies and found that uSUSTAIN approximated SUSTAIN's successes. Despite its simplicity, the unified recruitment procedure in uSUSTAIN has proven remarkably successful in this domain.

Although uSUSTAIN has demonstrated the ability to account for human learning performance across a wide range of category learning paradigms, it has never been applied to a study specifically designed to compare unsupervised and supervised learning. Given the past successes of the model, it would be informative to apply the model to a direct comparison between supervised and unsupervised learning. The following section examines uSUSTAIN's account of Love's[18] study that compares incidental unsupervised learning, intentional unsupervised learning, and supervised classification learning in a controlled manner.

4. Comparing Supervised and Unsupervised Learning

The Love[18] study is unique in that it specifically allows for a direct comparison of supervised and unsupervised learning. In supervised learning, the common dependent measure used to assess learning difficulty is training accuracy.[25,30] However, there is no measure of training accuracy in unsupervised learning (there is no right or wrong response on each study trial). In order to directly compare learning performance across these two types of learning, a comparable dependent measure was developed.

To accomplish this, stimuli were created by embedding the category label (which is typically a verbal label such as category "A" or "B") into each stimulus as a fourth binary-valued perceptual dimension (see Table 1). On supervised classification study phase trials, subjects were shown the value of the first three perceptual

Table 1. The logical structure of Types I, II, IV and VI classification problems tested in Ref. 30.

Stimulus	I	II	IV	VI
1 1 1	**1**	**1**	**1**	**1**
1 1 2	**1**	**1**	**1**	**2**
1 2 1	**1**	**2**	**1**	**2**
1 2 2	**1**	**2**	**2**	**1**
2 1 1	**2**	**2**	**1**	**2**
2 1 2	**2**	**2**	**2**	**1**
2 2 1	**2**	**1**	**2**	**1**
2 2 2	**2**	**1**	**2**	**2**

dimensions and were queried on the fourth. After responding, the correct value of the fourth dimensions was shown. In the Love[18] study, the fourth dimension (i.e. the "category" dimension) was the border color (either yellow or white) of a geometric figure. Subjects indicated whether they believed the border color (not shown on the display) was yellow or white based on the three other perceptual dimensions (which were shown on the display). After responding, the complete figure was displayed.

On unsupervised study phase trials, all four perceptual dimensions were shown on study phase trials (the fourth dimension was not queried). In the intentional unsupervised learning condition, subjects were aware they were in a learning task and were instructed to actively search for patterns that characterized the training items. In contrast, subjects in the incidental unsupervised learning condition were not aware that they were in a learning task and were instructed to simply rate how pleasant they found each stimulus item.

In each of the three study conditions (supervised classification learning, intentional unsupervised learning, incidental unsupervised learning), subjects were trained on either Types I, II, IV, or VI category structures (see Table 1) defined by Shepard, Hovland and Jenkins.[30] Type I problem only requires attention along one input dimension, whereas Type II problem requires attending to two dimensions (Type II is XOR on the first two dimensions with an irrelevant third dimension). The categories in Type II problem have a highly nonlinear structure. Type IV requires attention along all three perceptual dimensions with each dimension serving as an imperfect predictor. Type IV is notable because it displays a linear category structure. Type VI also requires attention to all three perceptual dimensions and has no regularities across any pair of dimensions. In all conditions, subjects completed ten study blocks (a block consists of the presentation of each stimulus item in a random order).

Category learning performance was measured in a test phase which followed the study phase. Subjects viewed a pair of stimuli that varied only on the fourth dimensions (i.e. the category dimension). Subjects were instructed to choose the item that appeared during the study phase (a familiarity or recognition judgment). As in traditional supervised classification learning studies, subjects could base this judgment on their knowledge of the relationship between the category dimension and other dimensions (e.g. rules, correlations, etc.) as well as on memorized exemplars. Love[18] verified that this testing procedure yields performance scores that correlate highly with study phase accuracy in the supervised condition. Thus, test phase accuracy can be used to compare the ability of subjects to learn in each of the three study conditions.

The results are shown in Table 2. The acquisition patterns for the three learning conditions differ significantly. Subjects in the unsupervised conditions did not show a preference for Type II category structure relative to Type IV structure. This effect was most pronounced in the incidental unsupervised learning condition. One explanation for this difference between the incidental and intentional unsupervised learning conditions is that intentional unsupervised learning task encouraged sub-

Table 2. The study phase and test phase results from Ref. 18. uSUSTAIN's fit is shown in parentheses.

Problem	Study Accuracy	Test Accuracy
Supervised Classification Learning		
Type I	0.86 (0.74)	0.89 (0.90)
Type II	0.67 (0.63)	0.73 (0.75)
Type IV	0.65 (0.60)	0.70 (0.65)
Type VI	0.59 (0.56)	0.61 (0.58)
Intentional Unsupervised Learning		
Type I	NA	0.84 (0.86)
Type II	NA	0.64 (0.57)
Type IV	NA	0.67 (0.66)
Type VI	NA	0.54 (0.50)
Incidental Unsupervised Learning		
Type I	NA	0.85 (0.81)
Type II	NA	0.56 (0.51)
Type IV	NA	0.67 (0.63)
Type VI	NA	0.56 (0.50)

jects to form explicit rules which is an efficient strategy for category structures that are describable by a compact rule (Type II has an XOR structure, see Table 1)

4.1. *Modeling results*

uSUSTAIN was trained in a manner analogous to how subjects were trained using 10 randomly ordered study blocks. In the supervised condition, the model was asked to predict the value of the fourth feature dimension (the category label), while in both unsupervised conditions, the value of all four stimulus dimensions was available to the model.

All study phase stimuli were encoded as belonging to a global category. In order for uSUSTAIN to mimic the forced choice nature of the test phase, the activation of this global category unit was calculated for each of the two items presented as a pair. The degree to which this unit was activated indicates the level of familiarity the model has for the item. The ultimate response of the network was towards the item in the forced choice that had the strongest response activity using a probabilistic decision procedure analogous to Eq. (8).

uSUSTAIN's fit of the data (averaged of 5000 runs) is shown in Table 2 in parentheses and the best-fit modeling parameters used are shown in Table 3. uSUSTAIN correctly demonstrates different patterns of acquisition for the three study conditions. In the supervised classification condition, uSUSTAIN captures the correct ordering of problem difficulty (Type I is easiest, followed by Type II, then Type IV, and finally Type VI). In both unsupervised conditions, the model correctly shows a decrease in test phase accuracy for Type II category structures relative to the supervised classification condition. Note that uSUSTAIN predicts

Table 3. uSUSTAIN's best fitting parameters for Ref. 18 studies.

Function/Adjusts	Symbol	Supervised	Intentional/Incidental Unsupervised
Learning rate	η	0.0172	0.0186
Cluster competition	β	2.90	0.608
Decision consistency	d	14.474	14.442
Attentional focus	r	0.475	2.209
Threshold	τ	0.568	0.553/0.487

lower accuracy for category structures that are nonlinear (Types II and VI) in both unsupervised conditions.

The clusters that uSUSTAIN creates were analyzed. The modal number of clusters recruited in the supervised learning condition exactly matches with previously reported results using the original version of SUSTAIN to account for the Shepard, Hovland, Jenkins[30] problems.[22] Two clusters were created in Type I condition, which divided the stimulus items on the basis of the first stimulus dimension. Four clusters were recruited in Type II condition, which captures the nature of the XOR on the first two dimensions. Six clusters were created in Type IV problem and eight clusters were recruited in Type VI condition where each item had to be memorized. In both of unsupervised learning conditions, the modal number of clusters was two for all problems except Type II under incidental conditions, which required four clusters.

Using essentially the same set of parameters uSUSTAIN was able to qualitatively approximate performance in both the intentional and incidental unsupervised learning conditions. Only the threshold parameter, τ, was different for the two studies ($\tau = 0.533$ for intentional unsupervised learning and $\tau = 0.487$ for incidental unsupervised learning). The increased threshold parameter in the intentional condition allowed uSUSTAIN to create four clusters in Type II problem which improves its test phase performance for this problem.

5. Discussion and Conclusions

uSUSTAIN holds that supervised and unsupervised human learning engage common mechanisms. In both cases, a new cluster is recruited when a stimulus is not sufficiently similar to any existing cluster belonging to the appropriate category. In order to evaluate uSUSTAIN's account of human learning, the model was applied to Love's[18] data set that compared human performance in supervised classification learning, incidental unsupervised learning and intentional unsupervised learning.

uSUSTAIN successfully fit these data. An analysis of uSUSTAIN's clustering solutions suggests that humans are more likely to aggregate stimulus items in memory in unsupervised learning whereas they are more likely to segregate stimulus items and store separate memory traces in supervised learning. Within unsupervised learning, the tendency to collapse items into common clusters is stronger in incidental learning.

uSUSTAIN's fit of the Love[18] data suggests that unsupervised learning, particularly incidental unsupervised learning, is best matched with linear category structures because the optimal clustering solution for a linear category structure involves one cluster per category. On the other hand, nonlinear category structures are not well matched to an unsupervised induction task because nonlinear category structures can only be captured with multiple clusters per category. While the linear/nonlinear distinction has not proved critical in supervised classification learning,[24] Love[18] suggested that the distinction may be meaningful in unsupervised learning. uSUSTAIN's account of the data supports this conjecture.

One counterintuitive prediction that uSUSTAIN makes is that incidental unsupervised learning may be the preferred induction task for some tasks. In other words, sometimes humans may be better off not trying to master the learning problem. One such situation is when numerous stimulus dimensions are weakly correlated with one another. Under such circumstances, uSUSTAIN predicts that supervised classification learning and intentional unsupervised learning will lead to clustering solutions that over-differentiate items and therefore do not fully capture the intercorrelated structure of the categories. In contrast, incidental unsupervised learning tends to aggregate items in common clusters and is more likely to capture the underlying category structure. uSUSTAIN's lower setting of τ parameter (which increases uSUSTAIN's tendency to cluster items together) for incidental unsupervised learning drives this prediction.

Despite the apparent differences between supervised classification learning, intentional unsupervised learning and incidental unsupervised learning, all three induction tasks are modeled through a common mechanism in uSUSTAIN. Beyond the current project, an important goal of our efforts is to model human learning across a range of situations and induction tasks. Doing so highlights theoretical connections across data sets and should lead to a more general understanding of human learning.

Acknowledgments

This work was supported by AFOSR Grant F49620-01-1-0295 to B. C. Love. Correspondence concerning this research should be addressed to Todd M. Gureckis, gureckis@love.psy.utexas.edu.

References

1. J. Anderson, "The adaptive nature of human categorization," *Psychol. Rev.* **98** (1991) 409–429.
2. F. Ashby, S. Queller and P. M. Berretty, "On the dominance of unidimensional rules in unsupervised categorization," *Percep. Psychophys.* **61** (1999) 1178–1199.
3. D. C. Berry and Z. Dienes, *Implicit Learning*: *Theoretical and Empirical Issues*, Erlbaum, Hillsdale, NJ, 1993.

4. D. Billman and J. Knutson, "Unsupervised concept learning and value systematicity: a complex whole aids learning the parts," *J. Experim. Psychol. Learn. Mem. Cogn.* **22**, 2 (1996) 458–475.
5. G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a selforganizing neural pattern recognition machine," *Comput. Vis. Graph. Imag. Proc.* **37** (1987) 54–115.
6. J. P. Clapper and G. H. Bower, "Learning and applying category knowledge in unsupervised domains," *Psychol. Learn. Motiv.* **27** (1991) 65–108.
7. J. P. Clapper and G. H. Bower, "Category invention in unsupervised learning," *J. Experim. Psychol. Learn. Mem. Cogn.* **20** (1994) 443–460.
8. A. Cleermans, *Mechanisms of Implicit Learning*: *Connectionist Models of Sequence Processing*, MIT Press, Cambridge, MA, 1993.
9. T. Gureckis and B. C. Love, "Modeling unsupervised learning with sustain," *Proc. 15th Annual FLAIRS Conf.*, 2002, pp. 163–167.
10. T. Gureckis and B. C. Love, "Towards a unified account of supervised and unsupervised learning," *J. Experim. Th. Artif. Intell.* **15** (2003) 1–20.
11. T. Gureckis and B. C. Love, "Who says models can only do what you tell them? unsupervised category learning data, fits, and predictions," *Proc. 24th Ann. Conf. Cognitive Science Society*, Hillsdale, NJ, Lawrence Erlbaum Associates, 2002, pp. 399–404.
12. J. A. Hartigan, *Clustering Algorithms*, Wiley, NY, 1975.
13. N. Hayes and D. E. Broadbent, "Two modes of learning for interactive tasks," *Cognition* **28** (1988) 249–276.
14. H. S. Hock, L. Malcus and L. Hasher, "Frequency discrimination: assesing global elemental letter units in memory," *J. Experim. Psychol. Learn. Mem. Cogn.* **12** (1986) 232–240.
15. T. Kohonen, *Self-Organization and Associative Memory*, Springer, Berlin, Heidelberg, 3rd edn., 1989.
16. J. Kruschke, "ALCOVE: an exemplar-based connectionist model of category learning," *Psychol. Rev.* **99** (1992) 22–44.
17. P. Lewicki, *Nonconscious Social Information Processing*, Academic Press, NY, 1986.
18. B. C. Love, "Comparing supervised and unsupervised category learning," *Psychol. Bull. Rev.* **9**, 4 (2002) 829–835.
19. B. C. Love, A. B. Markman and T. Yamauchi, "Modeling classification and inference learning," *Proc. Fifteenth Nat. Conf. Artificial Intelligence*, 2000, pp. 136–141.
20. B. C. Love and D. L. Medin, "Modeling item and category learning," *Proc. 20th Ann. Conf. Cognitive Science Society*, Mahwah, NJ, Lawrence Erlbaum Associates, 1998, pp. 639–644,
21. B. C. Love and D. L. Medin, "SUSTAIN: a model of human category learning," *Proc. Fifteenth Nat. Conf. Artificial Intelligence*, Cambridge, MA, MIT Press, 1998, pp. 671–676.
22. B. C. Love, D. L. Medin and T. Gureckis, "SUSTAIN: a network model of human category learning," *Psychol. Rev.* (2002) in press.
23. R. D. Luce, *Individual Choice Behavior*: *A Theoretical Analysis*, Greenwood Press, Westport, CN, 1959.
24. D. L. Medin and P. J. Schwanenflugel, "Linear separability in classification learning," *J. Experim. Psychol.*: *Human Learn. Mem.* **7** (1981) 355–368.
25. R. M. Nosofsky, M. A. Gluck, T. J. Palmeri, S. C. McKinley and P. Glauthier, "Comparing models of rule based classification learning: a replication and extension of Shepard, Hovland, and Jenkins (1961)," *Mem. Cogn.* **22** (1994) 352–369.

26. R. M. Nosofsky, T. J. Palmeri and S. C. McKinley, "Rule-plus-exception model of classification learning," *Psychol. Rev.* **101**, 1 (1994) 53–79.
27. R. Ratcliff, "Connectionist models of recognition memory: constraints imposed by learning and forgetting functions," *Psychol. Rev.* **97** (1990) 285–308.
28. D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representations by back-propagating errors," *Nature* **323** (1986) 533–536.
29. R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science* **237** (1987) 1317–1323.
30. R. N. Shepard, C. L. Hovland and H. M. Jenkins, "Learning and memorization of classifications," *Psychol. Monogr.* **75**, 13, Whole No. 517 (1961).
31. S. A. Sloman, "Explanatory coherence and the induction of properties," *Thinking & Reasoning* **3** (1997) 81–110.
32. B. Widrow and M. E. Hoff, "Adaptive switching circuits," *IRE WESCON Convention Record*, NY, 1960, pp. 96–104.

Todd M. Gureckis is a graduate student in psychology at the University of Texas at Austin. He received a B.S. in electrical and computer engineering from the University of Texas at Austin in 2001.

His main research interest is computational modeling of human categorization behavior, particularly in unsupervised contexts.

Bradley C. Love is an Assistant Professor in cognitive psychology at the University of Texas at Austin. He received his Ph.D. in cognitive psychology from Northwestern University and his B.S. in cognitive and linguistic sciences from Brown University.

His main research interest is empirical and computational explorations of human category learning.