Who says models can only do what you tell them? Unsupervised category learning data, fits, and predictions

Todd M. Gureckis (gureckis@love.psy.utexas.edu)
Department of Psychology; The University of Texas at Austin
Austin, TX 78712 USA

Bradley C. Love (love@psy.utexas.edu)
Department of Psychology; The University of Texas at Austin
Austin, TX 78712 USA

Abstract

How do people learn and organize examples in the absence of a teacher? This paper explores this question through a examination of human data and computational modeling results. The SUSTAIN (Supervised and Unsupervised STratified Incremental Network) model successfully fits human learning data drawn from two published studies. The first study examines how correlations between features can facilitate unsupervised learning. The second set of studies examines the role that similarity and attention play in unsupervised category construction (i.e., sorting) tasks. Importantly, SUSTAIN suggests two novel behavioral predictions that are confirmed.

Introduction

The study of human category learning has focused on supervised learning. Researchers typically utilize a experimental procedure in which the participant must learn to classify a set of stimuli while receiving corrective feedback on every trial. Certainly, there are many other ways to learn about the world. Our environment does not always provide us with explicit feedback and thus, some learning is better characterized as unsupervised. For example, we routinely categorize incoming email as "junk mail" in the absence of a teacher. A great deal of human learning may be unsupervised. The goal of this paper is to expand our understanding of how humans learn from examples without supervision.

To achieve this goal, we fit the SUSTAIN model of category learning to Billman and Knutson's (1996) studies concerning how humans learn correlations through observation and to Medin, Wattenmaker, and Hampson's (1987) data on unsupervised category construction (i.e., sorting) behavior. SUSTAIN successfully accounts for human performance in both of these studies with one set of parameters. Importantly, SUSTAIN's account of these studies suggests novel predictions which are subsequently tested (and confirmed) with human subjects.

The Modeling Approach

SUSTAIN has been successfully applied to an array of challenging human data sets spanning a variety of category learning paradigms including supervised classification (Love & Medin, 1998), inference learning (Love, Markman, & Yamauchi, 2000), and unsupervised learning (Gureckis & Love, 2002). One primary goal of our modeling approach is to address multiple forms of category learning (both supervised and unsupervised) with one consistent set of principles. After a brief introduction to the operation of SUSTAIN, these core principles will be discussed.

Introduction to SUSTAIN

SUSTAIN is a clustering model of human category learning. The internal representation of the model consists of a set of clusters. Category representations consist of one or more associated clusters. At the start of learning, the network has a single cluster that is centered in this representational space upon the first input pattern.

When a new stimulus item is presented, SUSTAIN attempts to assign the item to the most similar existing cluster. This assignment is unsupervised since it is based only on the similarity between item and cluster. If a *surprising* event occurs, such as a misprediction in supervised learning or a stimulus is encountered in unsupervised learning that is not similar to any existing cluster, SUSTAIN creates a new cluster to encode the current stimulus. This new cluster is centered in the representational space on the misclassified item.

When a stimulus is not surprising, the item is assigned to the most similar existing cluster and this cluster updates its internal representation to become more similar to the current item (a process somewhat analogous to prototype formation). Classification decisions are based on the cluster to which a stimulus instance is assigned. Like other models of category learning (e.g., Kruschke, 1992), SUSTAIN's selective attention mechanism learns to selectively weight stimulus feature dimensions that are most useful for categorization.

The Principles of SUSTAIN

With this general understanding of the operation of the model, we now examine the six key principles that underly SUSTAIN.

Principle 1, SUSTAIN is directed 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 of the learning task. Its selective attention mechanism further serves to bias SUSTAIN towards simple solutions by focusing the model on the stimulus dimensions that provide consistent information.

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.

Principle 3, SUSTAIN relies on both unsupervised and supervised learning processes As discussed above, SUSTAIN can cluster based on similarity (an unsupervised process). SUSTAIN's operation is also affected by supervision (when available). Consider the example of SUSTAIN learning to classify stimuli as members of the category mammals or birds. Let's assume that a cluster representing four-legged, hairy, land creatures is already acquired, as well as another cluster representing small, winged, creatures that fly. The first time SUSTAIN is asked to classify a bat, the model will predict that a bat is a bird because the bat stimulus will be more similar to the existing bird cluster than to the existing mammal cluster. Upon receiving corrective feedback (supervision), SUSTAIN will note its error and create a new cluster to store the anomalous bat stimulus. Now, when this bat or one similar to it is presented to SUSTAIN, SUSTAIN will correctly predict that the bat is a mammal. This example also illustrates how SUSTAIN can entertain more complex solutions when necessary through cluster recruitment (see Principle 1).

Principle 4, Clusters are recruited in response to surprising events As the previous example illustrates, surprising events lead to new clusters being recruited. In unsupervised learning, a surprising event is simply exposure to a stimulus that is not sufficiently similar to any existing cluster (i.e., a very novel stimulus).

Principle 5, the pattern of feedback matters As the bird-mammal example 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 (e.g., inference versus classification learning) that are informationally equivalent but differ in their pattern of feedback.

Principle 6, 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 (compare to Sloman's, 1997, account of competing explanations in reasoning).

Model Fits and Predictions

In the following sections, Billman and Knutson's (1996) results are described, fit, and SUSTAIN's novel predictions are tested. Following Billman and Knutson, Medin et al.'s (1987) work is given similar consideration.

Modeling Billman and Knutson's (1996)

Billman and Knutson's experiments tested the prediction that category learning is easier when certain stimulus feature dimensions are predictive of other feature dimensions (e.g., "has wings", "can fly", "has feathers" are all inter-correlated features of birds) than when correlations are unrelated or are not numerous. Their studies evaluate how relations among stimulus feature dimensions affect learning in an unsupervised task. SUSTAIN has successfully fit Billman and Knutson's (1996) Experiment 2 and 3 (Gureckis & Love, 2002). Here, we focus on Experiment 3.

Fitting Billman and Knutson's (1996) data Subjects studied stimulus items that depicted imaginary animals made up of of seven feature dimensions: type of head, body, texture, tail, legs, habitat, and time of day pictured. Each dimension could take on one of three values. For example, the time of day could be "sunrise", "nighttime", or "midday". The correlational structure of the feature dimensions varied according to which of two conditions (either the Structured or the Orthogonal condition) the subject was randomly assigned. The abstract structure of the two conditions is shown in Table 1. In the Structured condition, the first three stimulus dimensions are intercorrelated (for a total of three correlations), while the remaining four dimensions vary freely. The Orthogonal condition's structure also contains three correlations (the first and second dimensions are correlated, as are the third and fourth, and the fifth and the sixth), but the correlations are isolated (e.g., the first and third dimension are not correlated).

In the learning phase for both conditions, subjects were told that they were participating in a visual memory experiment and viewed 27 stimulus items for four blocks (a block is a single pass through all training items). Each of the 27 items appeared once

Table 1: The logical structure of the stimulus items for the Orthogonal and Structured conditions in Experiment 3 of Billman and Knutson (1996). The seven columns denote the seven stimulus dimensions. Each dimension can display one of three different values, indicated by a 1, 2, or 3. An x indicates that the dimension was free to assume any of the three possible values.

Structured Condition		
1 1 1 x x x x	2 2 2 x x x x	3 3 3 x x x x
Orthogonal Condition		
1 1 1 1 1 1 x	2 2 1 1 1 1 x	3 3 1 1 1 1 x
1 1 1 1 2 2 x	2 2 1 1 2 2 x	3 3 1 1 2 2 x
1 1 1 1 3 3 x	2 2 1 1 3 3 x	3 3 1 1 3 3 x
1 1 2 2 1 1 x	2 2 2 2 1 1 x	3 3 2 2 1 1 x
1 1 2 2 2 2 x	2 2 2 2 2 2 x	3 3 2 2 2 2 x
1 1 2 2 3 3 x	2 2 2 2 3 3 x	3 3 2 2 3 3 x
1 1 3 3 1 1 x	2 2 3 3 1 1 x	3 3 3 3 1 1 x
1 1 3 3 2 2 x	2 2 3 3 2 2 x	3 3 3 3 2 2 x
1 1 3 3 3 3 x	2 2 3 3 3 3 x	3 3 3 3 3 3 x

per block in a random order. The only difference between the Structured and Orthogonal conditions was the abstract structure of the stimuli that were shown during the learning phase.

In the test phase of the experiment, subjects viewed a novel set of 54 stimulus pairs. Each member of the pair had two of the seven feature dimensions obscured (e.g., the locations where the tail and head should have been were blacked out) so that information about only one correlation was available for each item in test pair. One item in the pair preserved the studied correlation, while the other item violated the correlation. Subjects were asked to choose the stimulus item in the pair that seemed most similar to the items studied in the learning phase (a forced choice procedure). The item that preserved the studied correlation was considered the correct choice. For example, in the isolating condition the correct item of the pair might have the abstract structure [1 1 m 1 m 1 2] because it preserves the correlation between the first and second dimensions (the 'm' represents a dimension that was blocked). The incorrect item of the pair might then be [1 2 m 1 m 1 2] which breaks the correlation present in the training items between the first and second dimension.

The basic result from Experiment 3 was that the "correct" item was chosen more often in the Structured condition than in the Orthogonal condition (77% vs. 66% from Table 2). This finding supports the hypothesis that extracting a category's structure is facilitated by intercorrelated dimensions.

Table 2: The mean accuracy for humans and SUSTAIN in Billman and Knutson's (1996) Experiment 3.

	Orthogonal	Structured
Human	.66	.77
SUSTAIN	.60	.77

Table 3: SUSTAIN's best fitting parameters for the studies considered. SUSTAIN's parameters are not discussed in this paper, but this table is included for readers who wish to replicate our results.

function/adjusts	symbol	value
learning rate	η	0.0966
cluster competition	β	6.40
decision consistency	d	1.98
attentional focus	r	10.0
threshold	τ	0.5

Modeling Results SUSTAIN was trained in a manner analogous to how subjects were trained by using four randomly ordered learning blocks. No feedback was provided as all stimulus items were encoded as being members of the same category. New clusters were recruited according to the unsupervised notion of surprise. In order for SUSTAIN to mimic the forced choice nature of the test phase, a response probability was calculated for each of the two items. The ultimate response of the network was biased towards the item in the forced choice that had the strongest response probability.

SUSTAIN was run numerous times on both conditions in both experiments and the results were averaged. The best fitting parameters are shown in Table 3. SUSTAIN correctly predicts greater accuracy in the Structured condition than in the Orthogonal condition (see Table 2).

In Experiment 3, SUSTAIN's most common solution in the Orthogonal condition was to partition the studied items into three clusters. However, the nature of the three partitions varied across runs. SUSTAIN tended to focus on one of three correlations present in the Isolated condition and ignored the other two. For instance, during training SUSTAIN might create three clusters organized around the first two input dimensions (one cluster for each correlated value across the two dimensions) and ignore the correlation between the third and fourth dimensions and the fifth and sixth dimensions.

SUSTAIN also recruited three clusters in the Structured condition. The same dynamics that lead SUSTAIN to focus on only one correlation in the Orthogonal condition leads SUSTAIN to focus on all of the interrelated correlations in the Structured condition. When SUSTAIN learns one correlation in the

Structured condition, SUSTAIN necessarily learns all of the pairwise correlations because of the way clusters are updated (i.e., three clusters are formed that capture the three basic subtypes of stimuli). This type of learning in the Structured condition is what lead to the higher accuracy levels.

SUSTAIN's solution to Experiment 3 suggests some novel predictions: (a) When correlations are not interrelated, learning one correlation should block the learning of other correlations, and (b) When correlations are interrelated, either all of the correlations are learned or none of the correlations are learned. These predictions are explored in the following section.

Testing the Predictions In the original Billman and Knutson article, accuracy was considered in aggregate for all three correlations. Here, we reanalyze Billman and Knutson's data by considering each subjects' performance on each correlation (i.e., each subject contributes three scores to the analysis instead of one). SUSTAIN predicts that human subjects will learn only one of the three correlations in the Orthogonal condition, but will learn either all or none of the correlations in the Structured condition. If this is true, the mean variance of subjects' accuracies for the three correlations should be higher in the Orthogonal condition than in the Structured condition. This was indeed the case. The mean variance of each subject's three accuracy scores was 0.030 for the Orthogonal condition, but only 0.010 in the Structured condition ($t(46) = 2.76$, $p < .001$).

Discussion Due to the way SUSTAIN organizes its clusters, it predicts that learning one correlation in the Orthogonal condition blocks the learning of other correlations (which should result in a high within subject variance), whereas in the Structured condition learning one correlation is tied to learning all three correlations (which should result in a low within subject variance). These predictions were made prior to obtaining access to Billman and Knutson's data. The combined results of the original Billman study and the subsequent analysis, suggest that people find categories that are organized around highly correlated features to be easier to learn because correlations enable the transfer of knowledge across features. The mechanism that supports this operation may bare a strong resemblance to SUSTAIN.

Modeling Sorting Behavior with SUSTAIN

Billman and Knutson's (1996) studies suggest that subjects prefer stimulus organizations in which the perceptual dimensions are intercorrelated. However, studies in category construction reveal a contrasting pattern — subjects tend to sort stimuli along a single dimension. This behavior persists despite the fact that alternate organizations exist that respect the intercorrelated nature of the stimuli, such as an intercorrelated family resemblance structure (Medin, Wattenmaker, & Hampson, 1987).

Table 4: The logical structure of the perceptual dimensions in Medin et al. (1987) sorted in two ways. In the family resemblance table, the stimuli with a preponderance of 1's can be seen as forming one family, while the stimuli with a preponderance of 2's can be seen as forming a second family or covert category. In the one-dimensional sort table, the same stimuli items are grouped on the basis of a single dimension (the first dimension).

Family Resemblance		One-dimensional Sort	
1 1 1 1	2 2 2 2	1 1 1 1	2 2 2 2
1 1 1 2	2 2 2 1	1 1 1 2	2 2 2 1
1 1 2 1	2 2 1 2	1 1 2 1	2 2 1 2
1 2 1 1	2 1 2 2	1 2 1 1	2 1 2 2
2 1 1 1	1 2 2 2	1 2 2 2	2 1 1 1

SUSTAIN was applied to the sorting data from Medin et al.'s (1987) Experiment 1 in hopes of reconciling the apparently contradictory findings. In Experiment 1, subjects were instructed to sort ten stimuli into two equal sized piles. Stimuli were cartoon-like animals that varied on four binary-valued perceptual dimensions (head shape, number of legs, body markings, and tail length). The logical structure of the items is shown in Table 4. The basic finding is that subjects sort along a single dimension (the one-dimensional sort in Table 4) as opposed to sorting stimuli according to their intercorrelated structure (i.e., the family resemblance structure shown in Table 4).

In these simulations, SUSTAIN was constrained to create only two piles (i.e., clusters) like Medin at al.'s subjects. This was accomplished by preventing SUSTAIN from recruiting a third cluster. SUSTAIN was presented with the items from Table 4 for 10 random training blocks to mirror subjects' examination of the stimulus set and their ruminations as to how to organize the stimuli. To evaluate the performance of the model, we looked at how SUSTAIN's two clusters were organized. Using the same parameters that were used in the Billman and Knutson (1996) studies listed in Table 3, SUSTAIN correctly predicted that the majority of sorts (99%) are organized along one stimulus dimension.

SUSTAIN's natural bias to focus on a subset of stimulus dimensions (which is further stressed by the selective attention mechanism) led it to predict the predominance of one-dimensional sorts. Attention is directed towards stimulus dimensions that consistently match at the cluster level. This leads

to certain dimensions becoming more salient over the course of learning (i.e., the model's attention value along that dimension becomes larger). The dimension that develops the greatest salience over the course of learning becomes the basis for the one-dimensional sort.

Which dimension provides consistent information during the course of learning will, in part, be determined by the order in which the stimulus items are presented to the model. Thus, SUSTAIN predicts that the order of card consideration in a sorting task might constrain which dimension human subjects focus their sort on. If card ordering has no effect and subjects randomly choose a dimension to sort on or choose due to individual differences in the salience of a particular dimension, then SUSTAIN's account should be insufficient.

Testing the Prediction

The following study tests this prediction by creating a modified version of the Medin, et al. sorting experiment in which the order that subject may consider cards is manipulated. Our interest was to test if the dynamics that led SUSTAIN to choose a particular dimension to sort on were the same dynamics that constrained subjects' sorting strategies.

Procedure Stimuli in our experiment were geometric shapes, printed on laminated cards, that varied on four of five binary valued dimensions (one dimensions value was held constant and thus had no influence on subjects sorting decisions). The dimensions were size (big or small), color of border (white or yellow), main color (blue or purple), a slash across the shape (present or absent), and texture (smooth or rough). Each dimension is independent and equally salient (as verified by multi-dimensional scaling of subjects' pairwise similarity ratings).

Participants were given a large board that was divided in half with a dark line (see Figure 1). Each side of the board had five positions in which to place cards. Before the start of an experiment trial, two "guide" cards were placed on the board that had opposing values along each dimension. Figure 1 shows an empty board with the abstract structure of these two guide cards. The particular values and meaning of each stimulus dimension was random for each subject (i.e., the values of the stimulus dimensions such as size and color were randomly assigned to one column of the abstract structure shown).

During the experiment, participants were given one new card at a time by the experimenter and were asked to place the card in an empty position on one side of the board according to what seemed most natural or sensible given the other cards on that side. The first two cards actually handed to subjects were constrained so that they mismatched on one dimension from the guide cards already on the board. For example, given the two cards in Figure 1 the abstract structure of the first two cards actually handed to the subjects might be [0 0 0 1] and [1 1 1 0].

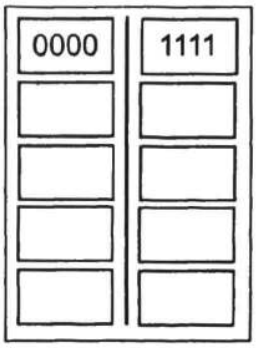

Figure 1: The layout and initial configuration of the board given to subjects is shown.

The final 6 cards given to subjects were drawn from the remaining possible. Cards were randomly chosen but came in pairs of opposing values. For example, if the fourth card had the abstract structure [0 0 1 0], the fifth might be [1 0 1 1]. This manipulation also helped to encourage subjects to fill the board up in a more or less even fashion rather than filling up one side completely, then having no choice as where to place the remaining cards.

Our hypothesis was that subjects would, like SUSTAIN, place the first two cards on the board on the basis of overall similarity to the guide cards as opposed to randomly choosing a single dimension on which to focus their sorting strategy. Thus in our example, [0 0 0 1] would be placed under the [0 0 0 0] prototype and [1 1 1 0] would be placed under the [1 1 1 1] prototype. If subjects allocated attention to dimensions that provide consistent information like SUSTAIN, then attention would be increased on only the dimensions that matched the guide cards (all but the fourth dimension in this case). This initial attentional disadvantage on the fourth, mismatched dimension would prevent subjects from sorting on that dimension.

Results Twenty-eight psychology undergraduate students participated in the study for course credit. The results collected for this study are show in Table 5. Of the 28 subjects, 23 subjects performed a one-dimensional sort while 5 used an alternate sorting strategy. Of the 23 subjects that performed a one dimensional sort, only 2 of these 23 subjects sorted the cards using the mismatched dimension as their basis for organization. If subjects had no particular preference for any dimension and the manipulation of the cards had no effect, then the probability of getting 21 out of 23 subjects to sort on a dimension other than the one mismatching dimension is

Table 5: The results of the sorting study.

	Number of Subjects
Subjects using a 1D sort	23
--*Mismatched Dimension*	2
—*Other Dimensions*	21
Subjects using a non 1D sort	5
—*Family Resemblance*	3
—*Unknown Strategy*	2
Total Subjects	28

less than .05 as given by a two-tailed binomial trial (n=23, p = .25). Of the five subjects that did not perform a one-dimensional sort, three performed a family resemblance sort and two performed a sort using an undecipherable sorting strategy.

SUSTAIN was simulated using the same parameters used for the Billman and Knutson studies (Table 3) and using the same conditions from the Medin, et al. sorting simulation, but with the specific card orderings that subjects were given in our experiment. In 100% of the trials, the model used a dimension other than the mismatched dimension as the basis for a one-dimensional sort.

Discussion

The dimension that subjects choose to sort in this task cannot be explained as random choice. The results presented in our experiment provide evidence that the order of card presentation plays a role in influencing subjects to sort on a particular dimension.

Specifically, sorting behavior is influenced by the way we perceive similarity between stimuli. In this unsupervised task, attention is allocated such that the similarity space changes during the course of learning. At the start of learning, each dimensions is more or less equally important, but as learning proceeds, certain dimensions become more salient (because they are more informative) while others become less. This warping of the similarity space is what ultimately causes judgments in this type of task to become based on a single dimension, rather than on the overall similarity between items. The fact that SUSTAIN predicted this behavior gives additional support to the notion that it's principles reflect some of the true operational principles of human learning.

Conclusions and Implications

SUSTAIN's combined account of Billman and Knutson's (1996) studies and Medin et al. (1987) suggest that the salience of stimulus dimensions change as a result of unsupervised learning and that the correlated structure of the world is more likely to be respected when there are numerous intercorrelated dimensions that are strong. In cases where the total number of correlations is modest, and the correlations are weak and not interrelated (such as in the Medin et al. stimuli), SUSTAIN predicts that stimuli will be organized along a single dimension.

The ability of SUSTAIN to account for two diverse unsupervised learning data sets with a single set of parameters demonstrate how it's formulation positions it as a robust model of category learning. In addition to the studies reported here, SUSTAIN's principles has been shown to generalize across a number of other forms of category learning (such as supervised learning and inference learning). It is these well-defined principles and the transparent operation of SUSTAIN that allow it to make the two predictions which have been successfully confirmed here.

Acknowledgments

We would like to offer our sincere thanks to Dorrit Billman for providing access to her data and to Rob Goldstone for motivating the sorting study. This work was supported by AFOSR Grant F49620-01-1-0295.

References

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: A complex whole aids learning the parts. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*(2), 458-475.

Gureckis, T. M., & Love, B. C. (2002). *Modeling unsupervised learning with sustain.* (In Press, *FLAIRS 2002 Special Track: Categorization and Concept Representation: Models and Implications*)

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Love, B. C., Markman, A. B., & Yamauchi, T. (2000). Modeling classification and inference learning. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 136-141.

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. In *Proceedings of the fifteenth national conference on artificial intelligence* (p. 671-676). Cambridge, MA: MIT Press.

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive Psychology, 19*, 242-279.

Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking & Reasoning, 3*, 81-110.