# Category learning without labels—A simplicity approach

**Emmanuel Minos Pothos** (e.pothos@ed.ac.uk)
Department of Psychology, University of Edinburgh; 7 George Square
Edinburgh, EH8 9JZ UK

**Nick Chater** (nick.chater@warwick.ac.uk)
Department of Psychology, University of Warwick;
Coventry, CV4 7AL UK

**Abstract**

In an extensive research tradition in categorization, researchers have looked at how participants will classify new objects into existing categories; or the factors affecting learning to associate category labels with a set of objects. In this work, we examine a complementary aspect of categorization, that of the spontaneous classification of items into categories. In such cases, there is no "correct" category structure that the participants must infer. We argue that the this second type of categorization, unsupervised categorization, can be seen as some form of perceptual organization. Thus, we take advantage of theoretical work in perceptual organization to use simplicity as a principle suitable for a model of unsupervised categorization. The model applied directly to similarity ratings about the objects to be categorized successfully predicted participants' spontaneous classifications. Moreover, we report evidence whereby perceived similarity is affected by spontaneous classification; this supplements the already substantial literature on such effects, but in categorization situation where the objects' classification is not pre-determined.

There are several situations in real life where novel objects can be spontaneously organized into groups. Consider a set of pebbles taken from a beach, or cloud patterns on a particular day, or just meaningless shapes shown onto a computer screen. This spontaneous classification can be appropriately labeled "unsupervised" because there are no "correct" categories the observer need to infer. By contrast, in supervised categorization, the learner (e.g., a child or someone learning a new language), has to infer what a category is by observing exemplars of the category and guessing their category membership (e.g., a child could be corrected for calling an apple an orange; through a process of corrective feedback, she would eventually learn to associate the appropriate objects with the category label "orange").

## Supervised vs. unsupervised categorization

While there has been very little theoretical work on unsupervised categorization, this has not been the case for supervised categorization. Several models have been put forward, covering different intuitions about the cognitive mechanisms of supervised categorization. For example, in definitional accounts of concepts (e.g., Katz & Fodor, 1963), categories are characterized by necessary and sufficient conditions for an item to be a category member (see Pothos & Hahn, 2000, for a recent evaluation). In exemplar theories (e.g., Nosofsky, 1989), a concept is represented by a set of known instances of that concept; new instances are therefore assigned to different categories in terms of their similarity to the members of each category. In prototype theories assignment is also determined by a similarity process, but this time to the prototype of each category, where a category prototype encapsulates some measure of central tendency across the exemplars of the category (e.g., Homa, Sterling, & Trepel, 1981).

Despite the technical sophistication of this research, it does not cover the whole scope of categorization processes. Models such as the exemplar model or the prototype one could never be used to predict how a person would spontaneously classify a set of items. In fact, in an influential paper Murphy and Medin (1985) criticized models such as the above for failing to explain category coherence—why it is the case that certain groupings of items make better categories than others; for example, the categories of birds or cups are coherent, but a category consisting of dolphins born on Tuesdays together with pink tulips within 20 miles of London, and the Eiffel Tower would be nonsensical. Given that the exemplar or prototype models could not explain such observations, Murphy and Medin concluded that they are inadequate models of categorization (and thus made a case for the importance of general knowledge in categorization).

However, under the light of the present distinction between supervised and unsupervised categorization, it is not the case that the exemplar or the prototype modes are inadequate in that they fail to capture general knowledge effects. Rather, category coherence is a problem of unsupervised categorization, as it relates to how categories originate–a process which, necessarily, cannot be guided by a 'supervisor.'

To summarize this section, the distinction of categorization models into supervised and unsupervised serves the useful purpose of enabling a closer specifi-

cation of the type of results that we expect each model to be able to capture. Unsupervised models of categorization will fail in predicting how participants will classify a new instance into a set of existing categories; but such models could probably be used to ground a theory of category coherence. The converse applies to models of supervised categorization.

## Previous work on unsupervised categorization

There has been an extensive experimental tradition on spontaneous classification, under the name of free sorting. However, the objective of free classification research is to identify the factors that appear to influence performance in sorting tasks, such as different types of instructions / experimental procedures and the structure of the stimuli (e.g., whether they are made of integral or separable dimensions, and the extent to which this affects the number of dimensions used in the classification task; e.g., Handel & Preusser, 1970; Wills & McLaren, 1998; Kaplan & Murphy, 1999). Thus, results from free sorting do not bear directly on the study of spontaneous classification, in the sense of actually predicting the classifications people are likely to come up with.

Trying to predict how objects are divided into groups has been a very frequently researched topic. While an exhaustive review of the different accounts by far exceeds the scope of the present work, we next discuss some of the qualifying factors of previous work with respect to its appropriateness for modeling unsupervised categorization.

Within machine learning and statistics, there is a long literature on clustering. There are two broad classes of clustering algorithms, agglomerative models and K-means ones. In the former case, for a set of N objects a hierarchy of clusters is produced whereby in the bottom level there are N clusters (a cluster for each object) and in the top level only one cluster (which includes all the objects). In the latter case, the number of clusters in which a set of objects is to be divided is set externally (this is why this approach is called "K-means"; for a review see Krzanowski & Marriott, 1995). In both approaches, knowledge of the number of groups sought is assumed; it must be pre-determined by the researcher. However, for a psychological model of unsupervised categorization we need to be able to predict both the number of categories and how the objects to be categorized are portioned into these categories within the same formalism.

This turns out to be an important limitation in terms of applying previous relevant modeling work in psychology directly to the problem of unsupervised categorization, as well. This applies, for example, to Ahn and Medin's Two Stage Model of Category Construction (Ahn & Medin, 1992), Michalski and Stepp's (1983) CLUSTER/2, and Anderson's rational categorization work (1991; additionally, Anderson's model is sensitive to order of presentation of the items to be categorized, so that his work is directed more towards dynamic aspects of categorization). This is not to criticize any of the excellent work cited above, but rather attempt to specify more precisely its modeling objective, with respect to how well it applies to unsupervised categorization.

Perhaps more directly relevant is Fisher's COBWEB (e.g., Fisher, 1996), which is based on the psychologically motivated principle of category utility (e.g., Corter & Gluck, 1992). Variants of the model can indeed determine the number of categories, as well as the way the items should be partitioned into the categories. However, three factors prevent its direct comparability to the present model. Firstly, category utility has been put forward to explain basic level categorization (e.g., Rosch & Mervis, 1985); the relation between basic level categorization is presently unknown. Secondly, COBWEB has been investigated—and to a large extent validated—as a statistical model, not a psychological algorithm. One of the differences between the two is that a psychological model is supposed to be founded on computational principles that make some statement about cognition. Finally, category utility assumes a representation of objects in terms of features; categorization predictions in this work are derived on the basis of empirically derived similarity information.

## Perceptual organization and simplicity

Categorization and perceptual organization, albeit superficially dissimilar processes, are nevertheless quite interlinked. Clearly categorization depends on perceptual organization, as how we perceive a set of objects will by necessity determine how we will categorize them. However, there is also a very extensive research tradition on effects of categorization on perceptual organization, showing that the way we categorize a set of objects is likely to affect how we perceive them (e.g., Goldstone, 1994; Harnad, 1987; Schyns & Oliva, 1998). Thus, we could maybe usefully look for a principle in perceptual organization to ground our model of unsupervised categorization.

A very influential approach in perceptual organization is the simplicity principle (e.g., Pomerantz and Kubovy, 1986; Chater, 1999), according to which the perceptual system is viewed as finding the simplest perceptual organization consistent with the sensory input. In fact, the simplicity principle has been recently shown to be equivalent to the most influential alternative, the likelihood principle (Chater, 1996).

In a simplicity framework, the notions of "interpretation" and "encoding" are central. At an intuitive level, encoding of information results in some data; simplicity is just a strategy for choosing an interpretation for the data. If we have a sequence like "abababababab" we could interpret it as "5 x (ab)"; but, clearly, there are many alternative interpretations (e.g., "a, 2 x (baba), b"). According to simplicity, the preferred theory / in-

terpretation is the one that minimizes the sum of the (1) complexity of the theory and (2) the complexity of the data when encoded with the theory.

## The simplicity model of unsupervised categorization

Full details of the model are given in Pothos & Chater (1998) and Pothos & Chater (in press). Here, we only attempt to qualitatively discuss the main features of the model.

There has been extensive research on the importance of information in categorization, other than similarity. However, there must be an important component of categorization research that is driven primarily by similarity as well. This would be particularly evident in the case of grouping novel objects, since there would be no a priori expectations for such objects. Also, incorporating general knowledge influences in models of categorization has been notoriously difficult. Thus, in this work we will restrict the simplicity model to a version whereby general knowledge effects are not taken into account.

We assume that the information encoded for a set of objects is information about how similar each object is to each other. A possible "interpretation" for this information is in terms of groups of categories; in other words, the cognitive system could attempt to recognize structure in the encoded similarity information that is best captured by dividing the objects into groups.

To determine which grouping is most suitable we need to consider the following terms:

code length for similarities in terms of grouping + code length for grouping (1)

code length for similarities without groups (2)

The simplicity principle will support the classification such that (1) is a lot less than (2).

Translating the above intuition into a computational model, we consider similarity information of the form (object A, object B) more or less similar to (object X, object, Y). The advantage of this approach is that the applicability of our categorization model is not restricted by representational assumptions for the objects to be categorized. For example, we can equally well apply the model, whether the items to be categorized are represented as bundles of features, points in some multidimensional space, or even simply in terms of pairwise similarities.

We define a group or a category as a collection of objects such that the similarities between any two objects in the group are greater than the similarities between any two objects between groups. In this way, the similarity relations that would have had to be specified without groups are reduced. For example, if we have objects A, B, and C, and we put objects A and B in one group, while object C is on its own, then by the above, this is equivalent to saying that the similarity between A and B is greater than between A and C, and B and C. In this way, we have an "operative" definition of a category.

Thus, with groups we have some information gain, or reduction in code length, since we do not need to specify as many similarity relations; this would be the "gain" associated with a classification. However, it will rarely be the case that all the specified similarity relations will be correct; in other words, a particular grouping might specify that objects A, B are more similar to objects X, Y, when in fact it is the other way around. Thus, the overall classification gain will be reduced by the costs of correcting the errors; there is an additional cost required to specify which is the particular classification used (for the actual formulae and derivations see Pothos & Chater, 1998).

## Experimental investigation

We wish to illustrate the applicability of the model with empirically derived similarity information about the items to be categorized. This approach is consistent with a growing trend in categorization research to take into account the well documented similarity structure changes that take place as a result of categorization.

The simplicity model can be used to predict the classification that should be most psychologically intuitive to naïve observers for a set of objects. We can thus examine the extent to which the classifications spontaneously produced by naïve observers are compatible with the simplicity model predictions.

### Materials

We used 11 items that varied along two dimensions (the physical space representation is shown is Figure 1; a 12th item had to be eliminated from analyses as it was not the same in the ratings and categorization tasks). The two dimensions defined the size of a square and the size of the filled-circles texture inside the square (see Figure 2 for an example). The stimuli were presented in a folder, printed individually on A4 paper in black ink for the categorization task, and on a 15'' Macintosh computer screen when participants were asked for similarity ratings.

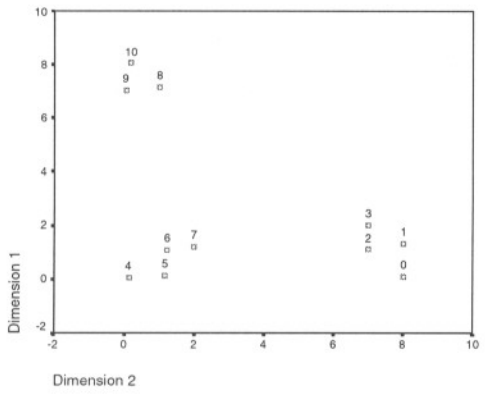


Figure 1: The parameter space representation of the stimuli.

Figure 2: An example of the stimuli used.

## Procedure

29 University of Oxford students were paid for their participation. In the first part of the study, they received instructions saying that they were about to receive a set of items and that they would have to divide them into groups "in a way that seems intuitive and natural, so that more similar items end up in the same group." They were also told that although there was no limit on how many groups they could use, they should not use more that what they thought necessary. The order in which the stimuli were arranged in the folder was randomized for each participant.

After participants had classified the items, they performed the ratings task on a computer. They were instructed that they were about to see the items of the first part in pairs and that their task was to indicate the similarity between the items in each pair on a 1 to 9 scale, where a "1" would correspond to most similar items and a "9" to items that were most different. In particular, for each pair, the first item was presented for one and a half seconds, then there was a fixation point for 250ms, the second item appeared for one and a half seconds, a blank screen for 250ms, and a 1–9 ratings scale. The order in which each item appeared in a pair was counterbalanced so that we had two ratings per participant for each pair. Two randomized different orders were used for the ratings part of the experiment.

## Results and Discussion

The similarity ratings were averaged into a large similarity matrix for all the items. This matrix was made symmetrical across the diagonal by using the arithmetic mean and also self-similarities were set to 0 (corresponding to maximum similarity). The simplicity model predictions were computed on the basis of these ratings. The best compression categorization involved three groups, with items 0–3, 4–7, and 8–10 in each group (item labels correspond to Figure 1).

In order to determine whether some of the observed categorizations were more likely than others we identified all the distinct categorizations produced by participants ("distinct" solutions), as well as the number of times participants divided the items in the way predicted by the simplicity principle. If there had been no preference for any particular categorization, we assumed that all distinct solutions would have been produced with a roughly equal frequency, given by the ratio (total number of groupings) / (number of distinct groupings). Using chi-square tests we can then examine whether the frequency of any of the classifications produced would be different from that computed by chance. This was the case only for the classification predicted by the model ($chi2(1) = 84.8$, $p<.001$; the frequency of this categorization was 11 times, out of 29).

To obtain some insights into participants' performance, we employed a non-metric MDS procedure to construct a putative internal spatial representation of the items; such a procedure is not related to the application of the simplicity model (which operates directly on the similarity ratings). Figure 3 shows the resulting MDS solution (all MDS procedures run with Euclidean metric). The three groups in the MDS solution correspond exactly to the three groups in Figure 1–but the items within each cluster are effectively indistinguishable in the internal space.

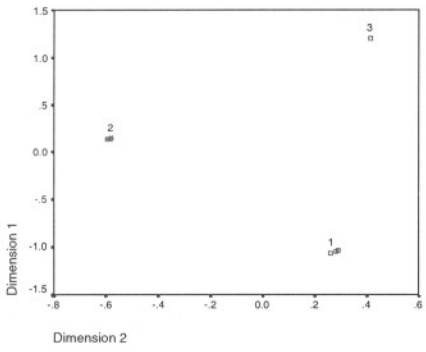


Figure 3: Labels items 0–3, "2" items 4–7, and "3" items 8–10, where item labels refer to Figure 1.

We next divided participants into homogeneous groups, and examined these groups individually. To do this, we looked at the groupings produced in the first part of the experiment, and then classified these groupings themselves (using the Rand index as a measure of the similarity between pairs of groupings, and the simplicity model as the clustering procedure). There were two main groups of categorizations, call them Group A (which contained the best compression solution; five different categorizations, that were produced by 14 out of the 29 participants) and Group B (nine solutions from 13 participants), as well as a smaller group which we shall not consider further (two other categorizations, from two participants).

We then separately considered the similarity ratings of participants whose groupings were in Groups A and B. The MDS procedure for Group A resulted in a spatial arrangement of the stimuli, identical to that shown in Figure 3. Figure 4, the MDS solution for Group B participants, is dramatically different; although some aspects of the nearest-neighbors structure seem to have been somewhat preserved (so that, for instance, points that were close to each other originally are still close to each other) the overall arrangement has been distorted so as to no longer reflect the obvious three groups category structure present in the Group A representation of the stimuli. In conclusion, it looks as if people who identified the best compression categorization (Group A), subsequently rated the similarity of different stimuli with each other in a way fully compatible with this category structure. This finding constitutes the first evidence that unsupervised classification affects the perceived similarity structure of a set of objects (see, e.g., Goldstone, 1995; Goldstone, Steyvers & Larimer, 1996 for corresponding evidence in supervised classification, that is categorization processes whereby categories are pre-specified). Future research will extend the present methodology to examine the extent to which simplicity might always be optimized with respect to how different individuals perceive the similarity structure of a set of objects.

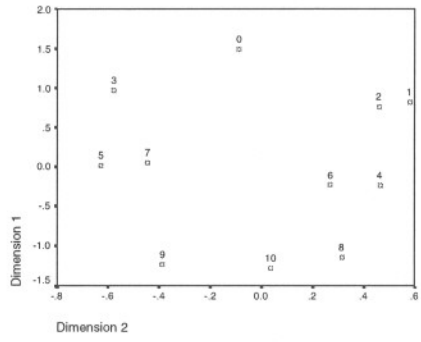


Figure 4: MDS solution for Group B participants.

To summarize, analysis of the similarity ratings of the stimuli confirmed the predictions of the simplicity model. Moreover, inspection of the MDS solutions showed that the categorization appears to have influenced similarity judgments, implying that perceived similarity may be affected by unsupervised classification.

## References

Ahn, W. & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, 16, 81-121.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.

Chater, N. (1996). Reconciling Simplicity and Likelihood Principles in Perceptual Organization. *Psychological Review*, 103, 566-591.

Chater, N. (1999). The Search for Simplicity: A Fundamental Cognitive Principle? *Quarterly Journal of Experimental Psychology*, 52A, 273-302.

Corter, J. E. & Gluck, M. A. (1992). Explaining Basic Categories: Feature Predictability and Information. *Psychological Bulletin*, 2, 291-303.

Fisher, D. (1996). Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4, 147-179.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, 123, 178-200.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Handel, S. & Preusser, D. (1970). The free classification of hierarchically and categorically related stimuli. *Journal of Verbal Learning and Verbal Behavior*, 9, 222-231.

Harnad, S. (Ed.) (1987). *Categorical Perception*. Cambridge: Cambridge University Press.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, 7, 418-439.

Kaplan, A. & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. *Memory & Cognition*, 27, 699-712.

Katz, J. & Fodor, J. A. (1963). The Structure of a Semantic Theory. *Language*, 39, 170-210.

Krzanowski, W. J. & Marriott, F. H. C. (1995). *Multivariate Analysis, Part 2: Classification, Covariance Structures and Repeated Measurements*. Arnold: London.

Michalski, R. & Stepp, R. E. (1983). Automated construction of classifications: conceptual clustering versus numerical taxonomy. *IEEE Transactions on pattern analysis and machine intelligence*, Vol. PAMI-5, 396-410.

Murphy, G. L. & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review*, 92, 289-316.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Journal of Experimental Psychology: Perception and Psychophysics*, 45, 279-290.

Pomerantz, J. R. & Kubovy, M. (1986). Theoretical Approaches to Perceptual Organization: Simplicity and Likelihood principles. In: K. R. Boff, L. Kaufman & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance, Volume II: Cognitive Processes and Performance*, 1-45. New York: Wiley.

Pothos, E. M. & Chater, N. (1998). Rational Categories. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, 848-853, LEA: Mahwah, NJ.

Pothos, E. M. & Hahn, U. (2000). So concepts aren't definitions, but do they have necessary *or* sufficient features?. *British Journal of Psychology*, 91, 439-450.

Pothos, E. M. & Chater, N. (in press). Basic Categories by Simplicity. In M. Ramscar, U. Hahn, E. Cambouropoulos, & H. Pain (Eds.) *Similarity and Categorization*. Oxford: Oxford University Press.

Rosch, E. & Mervis, B. C. (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology*, 7, 573-605.

Schyns, P. G. & Oliva, A. (1999). Dr. Angry and Mr. Smile: When categorization flexibly modifies the perception of faces in rapid visual presentations. *Cognition*, 69, 243-265

Wills, A. J. & McLaren, I. P. L. (1998). Perceptual learning and free classification. *Quarterly Journal of Experimental Psychology*, 51B, 235-270.