1 Introduction

Emmanuel M. Pothos and Andy J. Wills

Categorization is one of the most fascinating aspects of human cognition. It refers to the process of organizing sensory experience into groups. This is an ability we share to some extent with other animals (e.g. Herrnstein & Loveland, 1964), and is key to our understanding of the world. Humans seem particularly adept at the systematic and productive combination of elementary concepts to develop complex thought. All in all, it is hard to envisage much of cognition without concepts.

The study of categorization has a long history (e.g. Hull, 1920). It is usually considered a particular research theme of cognitive psychology, cognitive science, and cognitive neuroscience. Categorization relates intimately to many other cognitive processes, such as learning, language acquisition and production, decision making, and inductive reasoning. What all these processes have in common is that they are inductive. That is, the cognitive system is asked to process some experience and subsequently extrapolate to novel experience.

A *formal* model of categorization is taken to correspond to any description of categorization processes in a principled, lawful way. Formal models of categorization are theories that allow quantitative predictions regarding the categorization behaviour of participants. Some formal models also make predictions about the underlying neuroscience.

Selecting the models to be discussed in this volume was difficult. Our goal was to create an accessible volume with a reasonably small number of models. As a result, there are many excellent models which we were not able to include. Notable omissions include Heit's (1997) proposal for how to modify exemplar theory to take into account the influences of general knowledge, Kurtz's (2007) auto-associative approach to categorization, Lamberts's (2000) model of the time course of categorization decisions, Rehder's (2003) view of categorization as causal reasoning, Schyns's (1991) model of concept discovery based on Kohonen nets, Vanpaemel and Storms's (2008) attempt to integrate prototype and exemplar theory, several models of statistical clustering (e.g. Fisher, 1996; Fraboni & Cooper, 1989), and interesting

developments based on the mathematics of quantum mechanics (e.g. van Rijsbergen, 2004).

We hope that the models we have selected will help to illustrate some of the key ideas in the formal modelling of categorization. In the next sections, we summarize some of these ideas. We then go on to summarize briefly each of the models presented in this volume, and finish with some thoughts about how these models might be compared.

Supervised and unsupervised categorization

Categorization processes can be distinguished into supervised and unsupervised – in other words, processes that require external feedback versus those that do not. This is an important distinction, and one that has had a substantial influence on the development of categorization research. For example, most categorization models are proposed as either specifically supervised categorization models or as unsupervised ones. The majority of categorization research concerns supervised categorization and, in this volume, we have included the most prominent corresponding models. Equally, we have attempted to include contributions which cover some of the successful unsupervised categorization models.

In brief, supervised categorization concerns the processing of novel experience in relation to a pre-defined set of categories. Simply put, a child might see a round object which looks like it is edible, and wonder how it fits to its existing categories of oranges, lemons, or apples. She might attempt a guess and an adult might point out whether the guess is correct or not; this process of corrective feedback is one of the possible ways in which categories can develop in a supervised way (although it is unclear as to how central this process is in human conceptual development). In the laboratory, supervised categorization often involves creating a set of artificial stimuli, determining how they should be classified (this is done by the experimenter prior to the experiment), asking a participant to guess the classification of each stimulus one by one, and providing corrective feedback.

It seems uncontroversial to say that supervised categorization plays a part in the acquisition of many real-world concepts. However, one can reasonably ask where concepts come from in the first place. A related intuition with respect to real life concepts is that certain concepts are less ambiguous than others (for example, compare ‘chair’ with ‘literature’; with respect to the latter, many naive observers would disagree as to which instances should be considered ‘literature’). Both these problems are problems of unsupervised categorization.

Unsupervised categorization concerns the spontaneous creation of concepts. For example, in the laboratory, participants might be presented with a set of artificial stimuli with instructions to classify them in any way they like. A key goal of unsupervised categorization research is to determine why certain classifications are preferred, compared to others. With respect to real concepts, one can ask what determines the particular division of experience into concepts. Why, for example, do we have separate concepts for 'chairs' and 'armchairs', rather than a single one to encompass all relevant instances? The particular divisions we acquire seem to be affected by the category labels our culture provides (e.g. Roberson *et al.*, 2005; this would correspond to a supervised categorization process), but they must also be influenced by prior intuitions of which groupings are more intuitive. Other things being equal, more intuitive classifications should be easier to learn, and so unsupervised categorization models can also be applied to the problem of predicting which classifications are easier to learn compared to others (of course, some classes of supervised categorization models are suitable for addressing this problem as well).

How fundamental is the distinction between supervised and unsupervised categorization? Consideration of the models of supervised and unsupervised categorization included in this volume reveals several important common features. For example, nearly all the models considered are driven by some function of psychological similarity. Also, some researchers have argued that supervised categorization models are logically equivalent to unsupervised ones (cf. Pothos & Bailey, 2009; Zwickel & Wills, 2005); such an argument for the equivalence of supervised and unsupervised categorization is based on the general computational properties of categorization models. However, even if it is computationally feasible to create a model which can account for both supervised and unsupervised categorization results within the same formalism, psychologically it might be the case that these are separate cognitive processes.

The SUSTAIN model (Love, Medin, & Gureckis, 2004; Chapter 10) was one of the first attempts to account for both supervised and unsupervised categorization within the same model. The model's architecture is specified around a parametric combination of two components. The first component develops category representations as a result of an external supervisory signal and the second component spontaneously generates clusters based on a principle of similarity (that is, more similar items end up in the same cluster). This model is interesting since it embodies a particular assumption about the relation between supervised and unsupervised categorization processes, namely that they are distinct but related.

Exemplars and prototypes

The contrast between exemplar and prototype theory has been at the heart of the development of (supervised) categorization research. Equally, these are the two theories that psychologists without any particular categorization expertise are most likely to recognize. Accordingly, the first two chapters cover a prominent version of exemplar theory, the generalized context model (Nosofsky, 1988; see also Medin & Schaffer, 1978) and prototype theory (Hampton, 2000; Minda & Smith, 2000), respectively. Contrasting these two formalisms is a complicated issue. In principle, it is possible to identify stimulus sets that allow differential predictions (e.g. Medin & Schaffer, 1978; Medin & Schwanenflugel, 1981). In practice, sometimes the comparisons hinge on the role of particular parameters, whose psychological relevance has to be carefully justified. The effort to compare prototype and exemplar theory has led some researchers to examine formal comparisons (e.g., Nosofsky, 1990; see also Ashby & Alfonso-Reese, 1995). Such comparisons and related analyses (e.g., Navarro, 2007; Smith, 2007; Vanpaemel & Storms, 2008) have led to a profound understanding of the formal properties of exemplar and prototype models, to an extent that is rare in psychology.

Unitary and multi-process models

Should categorization be understood as a unitary process (e.g. Nosofsky & Kruschke, 2002) or a combination of independent processes? Chapter 4 covers the COVIS model (COmpetition between Verbal and Implicit Systems; Ashby *et al.*, 1998), which has been built on the assumption that human (supervised) categorization is supported by at least two separate, competing systems. COVIS is also notable as it is currently the only model which has been developed to provide categorization predictions at both the behavioural and neuroscience level. Indeed, COVIS motivated many of the early investigations which have allowed categorization researchers to consider ways in which the impressive recent advances in neuroscience could help the development of categorization theory (e.g. Nomura *et al.*, 2007; Zeithamova & Maddox, 2006).

Parallel distributed processing

Parallel distributed processing (PDP) models are generally considered to have a certain degree of biological plausibility – in other words, the

architecture of the models is said to mimic some aspects of brain architecture. PDP models are often built to describe particular aspects of cognitive development (e.g. Plunkett *et al.*, 1997) or psychopathology (Plaut & Shallice, 1993). McClelland and Rumelhart (1986) have led an extensive connectionist research programme; Chapter 5 covers the extension of this work in categorization behaviour. Unlike most categorization models, which are tested with respect to either the classification of novel instances or the spontaneous generation of categories, the PDP model of Chapter 5 is supported through known developmental aspects of the categorization process and how categorization competence breaks down in specific cases of brain pathologies (such as semantic dementia). Chapter 7 considers the feature-based approach to stimulus representation assumed by PDP models.

Attentional processes

The acquisition of categories seems to result in the direction of attention towards those aspects of the stimuli that are most useful in determining category membership. Most formal models of categorization posit some form of attentional process; the focus of Chapter 6 is these processes. It also extends these ideas to both mixture of experts models (see also Chapter 4) and considers how they might be formulated within a Bayesian framework (see also Chapter 8).

Optimal inference models

Categorization is an example of an inductive problem, which requires the determination of category membership from the limited information provided by the features of a stimulus. The mathematics of Bayes's theorem can be employed to develop accounts of optimal performance on inductive problems. Often, this kind of approach takes a step back from psychological processes to consider how ideal solutions to the inductive problem of categorization might shed light on the behaviour of humans and other animals. Such an approach is embedded in the general effort to understand cognition in terms of Bayesian probabilistic principles (e.g., Griffiths, Steyvers, & Tenenbaum, 2007; Tenenbaum & Griffiths, 2001). Bayesian principles can also be extended to more powerful frameworks (e.g., based on quantum probability; Busemeyer, Wang, & Townsend, 2006). Chapter 8 illustrates the application of Bayesian principles in categorization, in terms of an extension to Anderson's Bayesian model of unsupervised categorization (Anderson, 1991).

Minimum description length

An approach similar to the above is possible if one considers categorization as a process of data reduction. In other words, perhaps one of the reasons we have categories is that they allow a more efficient (less memory intensive) representation of the world. A minimum description length (a.k.a. *simplicity*) framework is basically an algorithmic coding scheme. It allows a researcher to define the codelength for data and hypotheses for the data. Then, the problem of choosing an appropriate hypothesis is translated to a problem of finding the hypothesis which leads to the greatest overall reduction in codelengths. Pothos and Chater (2002) suggested that categories can be considered as hypotheses regarding structure in the similarity relations between a set of stimuli. A particular classification will be preferred if it can simplify the description of similarity information to a greater extent. Thus, simplicity principles naturally lead to a model of unsupervised categorization, which is described in Chapter 9.

It is interesting that the normative computational frameworks of Bayesian probability and MDL can both lead to unsupervised categorization models – perhaps this is because the lack of an external teaching signal in unsupervised categorization is replaced by the assumptions each model makes regarding structure (cf. Chater, 1996).

Machine learning

Categorization research in psychology concerns the organization of objects into categories. Clearly, this process is relevant in many areas of machine learning and statistical clustering. A common problem in such areas is to infer whether it is meaningful to organize some instances into clusters – this is a problem of unsupervised categorization. Chapter 11 covers some related modelling work, in relation to a class of models based on category utility, that is the probability that an instance has certain features given membership to a particular category (i.e., how 'useful' the category is, for the purpose of predicting the features of its members, e.g., Corter & Gluck, 1992). Clearly, category utility is closely related to the Bayesian approach described in Chapter 8.

Considering a machine learning approach to categorization raises several interesting questions. How much convergence should we expect between human and machine learning categorization? Are there categorization methods more efficient or useful than the one employed by the human cognitive system? How domain-dependent is the selection of the optimal categorization strategy?

General knowledge

Murphy and Medin (1985) pointed out that conceptual coherence, the 'glue' that binds the instances of a concept together in a meaningful and intuitive way, has to be more than just, for example, similarity relations. Each concept is an inseparable part of our overall knowledge of the world and, conversely, without this knowledge it is impossible to appreciate the significance of a concept. Compelling as these intuitions about categorization have been, it has proved remarkably difficult to formalize a putative role of general knowledge in categorization (cf. Fodor, 1983). Chapter 12 covers a proposal for a model about how categories develop based in part by some aspects of general knowledge.

Outline of this book

In this section we highlight some of the key aspects of the models covered in this volume. The models are described in detail in their respective chapters. Our purpose is not to repeat this material, rather to draw the attention of the reader to such model features that might enable a better understanding of model differences and commonalities.

Chapter 2 – The generalized context model

The generalized context model (GCM) is an exemplar model of supervised categorization. A novel stimulus is classified into a pre-existing category based on its similarity to known members of that category (and to members of other known categories). Similarity in the GCM is specified in terms of distances in a psychological space, as proposed by, for example, Shepard (1987). So, at the heart of the GCM is a principle of psychological similarity. A fundamental aspect of the GCM is that it computes similarity relations not just on the basis of the original psychological space, but also any transformations of this space that are possible through (graded) attentional selection or compression/stretching of the psychological space as a whole. In this way, the GCM is a very powerful model: it is most often the case that its parameters can be set in a way that human data in a supervised categorization can be closely reproduced.

The GCM makes relatively few prior assumptions about the categorization process. For example, parameters governing attentional weighting, the form of the similarity function, the metric space, the nature of responding (probabilistic versus deterministic) can all be set in response to fitting particular human data. The price for this flexibility is, of course, the relatively large number of free parameters. Some key psychological

assumptions embodied in the GCM (apart from the obvious one, that category representation is based on individual exemplars) are that graded attentional weighting of stimulus dimensions and stretching/compression of psychological space are possible as a result of learning.

Chapter 3 – Prototype models of categorization

The extensive research on the relation between exemplar and prototype theory has led to computational implementations of these ideas in a way that their form is as similar as possible, and differs only with regards to the key psychological assumptions which are unique in each approach. This is a highly desirable situation, as it enables precise comparisons between the two formalisms. According to prototype theory, a novel instance is more likely to be classified into a category if the similarity between the instance and the category prototype is high; prototypes are typically operationalized as averages of category members. As with exemplar theory, more recent versions of prototype theory allow the same transformations of psychological space as the GCM. Another common feature of the two approaches is that they both postulate a single system of categorization.

Prototype theory is very similar to exemplar theory, but for a critical difference. The former is consistent only with linearly separable, convex-shaped categories, but the latter allows any kind of category shape. To see intuitively why this has to be the case, consider that for a category to have a meaningful prototype representation, the prototype (which is the average of the instances) must be included in the area (or volume) of psychological space which is occupied by the category.

Chapter 4 – COVIS

The COVIS model postulates that category learning is mediated by two, competing systems. The first system attempts to develop explicit, verbalizable rules that describe the required categorization. The rule-based system will be favoured to the extent that such rules exist, are simple, and allow accurate classification performance. It is assumed to be supported by the prefrontal cortex, anterior cingulate, the anterior striatum, and the hippocampus. The second system is a procedural learning system, which allows the learning of classifications such that information from all available dimensions is taken into account. Accordingly, the procedural system involves a mechanism of information integration. The brain areas associated with this system are principally the posterior striatum and the inferotemporal cortex. The two systems compete with each other; for any particular stimulus,

preference for one of the two competing systems is determined by confidence in the predicted response and the overall track record of the system.

The unique element of COVIS is that its computational implementation is specified with respect to the known neurophysiology of the brain. For example, the equation determining perseveration for a rule involves a free parameter which is linked to dopamine levels in the striatum. In this way, COVIS can be tested both with behavioural data (e.g., participant performance in a categorization experiment) and neuroscience data (e.g., fMRI studies of how brain activity varies with different categorization tasks).

Chapter 5 – Semantics without categorization

This chapter summarizes the progress in an extensive research programme to model human categorization behaviour with a multi-layer, feedforward, backpropagation network. An underlying hypothesis in this programme is that categories do not exist as distinct representational entities, rather categorization behaviour (of any kind, for example, classification of new instances or inference about the unseen properties of a shown stimulus) arises from the way environmental input affects the connections in a network. A particular feature of the postulated network architecture is the existence of a set of context units, which take into account the particular situation in which the categorization of a new instance takes place (cf. Chapter 12). Different contexts can result in different categorizations for the same instance.

Chapter 6 – Models of attentional learning

This chapter summarizes some of the evidence in support of the idea that categorization involves selective attention, and then discusses the development of models to account for this phenomenon. Starting with approaches related to the global stretching and compression of psychological dimensions implemented in the GCM, a proposal is presented for how attentional allocation may be exemplar specific, and how attentional allocation may be allocated between competing cognitive systems (cf. COVIS). There is also consideration of how attentional allocation might occur within a Bayesian framework (cf. Chapter 8), where multiple hypotheses about category structures are maintained simultaneously.

Chapter 7 – An elemental model of associative learning and memory

This chapter considers a feature-based (a.k.a. *elemental*) approach to modelling categorization (see also Chapter 5). Specifically, the phenomenon

of peak shift is discussed, for which (in both humans and pigeons) an elemental account may be more appropriate than an exemplar-based account (cf. Chapter 2). Peak shift is the phenomenon that, under certain circumstances, classification accuracy may increase with *decreasing* similarity to the members of category into which the item is classified. A formal elemental model of categorization is presented that provides an account of some of the situations where peak shift does, and does not, occur.

Chapter 8 – Nonparametric Bayesian models of categorization

According to this approach to unsupervised categorization, a model of category learning can be developed by considering how one can compute the category membership of a novel stimulus, given the appearance of the stimulus. In other words, the problem of categorization can be reframed as a problem of estimating the probability distribution of different objects with the same category label. Employing a Bayesian probabilistic framework to make this idea more concrete can lead to a number of implementation options, a key difference of which is whether the estimation of the required probability distribution is parametric (some assumptions are made regarding the general form of the distribution) or nonparametric (no assumptions made). This chapter describes a particular categorization model based on the latter approach, so that the prior assumptions about structure in the world are minimal; the model can be seen as an extension of Anderson's (1991) rational model of categorization.

A strength of this Bayesian approach to categorization is that it provides a framework for specifying a family of categorization models, including ones which are analogous to standard exemplar or prototype models (two parameters can determine whether a particular instantiation behaves more like an exemplar or a prototype model).

Chapter 9 – The simplicity model of unsupervised categorization

Chapter 9 describes the second model of (just) unsupervised categorization that is considered in this volume. The simplicity model is based on principles similar to those underlying the Bayesian probabilistic framework explored in Chapter 8. According to the simplicity model, categorization has a functional role, namely that of providing a more efficient description of any encountered stimuli. This 'simplicity' prerogative (informally equivalent to Ockham's razor) is formally implemented in a MDL framework, which is just a set of rules for deciding when a particular description for some data should be preferred. In the case of