Models of Perceptual Learning

Shimon Edelman
Department of Psychology
Cornell University
Ithaca, NY 14853, USA

Nathan Intrator
Department of Computer Science
Tel-Aviv University
Ramat-Aviv 69978, Israel

1 Introduction

A generation ago, mathematical psychology, which at the time was *the* discipline in charge of modeling behavior, appeared to be in poor shape. William Estes, one of the main *dramatis personae* in that field, described it like this: "Look at our present theories... or at the probabilistic models that are multiplying like overexcited paramecia. Although already too complicated for the average psychologist to handle, these theories are not yet adequate to account for the behavior of a rodent on a runway" (Estes, 1957). During the following decades, when the mainstream psychology underwent a major paradigm shift, the modeling of perceptual learning fared better than what one might have expected from the view expressed by Estes. A new theoretical outlook, which encouraged thinking now termed representational or computational, took over the field. At the same time, the models became, if anything, more complex compared to those of 1957.

Encouragingly, the models are now also more successful in explaining behavior (rather than merely predicting the probability of a certain response to a given stimulus), while giving no undue troubles to the psychologists.[1] Insofar as there is progress, it seems to stem mainly from (1) the improvement in the experimental techniques that subserve data collection in behavioral and physiological psychology, and (2) the revision of the theoretical basis from which models are drawn. The "rodent on a runway" example mentioned above serves well to illustrate both these points. On the theoretical or conceptual side, the current explanation takes the route presaged by Tolman and based on the concept of cognitive maps (O'Keefe and Nadel, 1978). On the experimental side, the existence of cognitive maps in the rat brain could not have been demonstrated without modern multi-electrode recording methods and the information-processing tools that accompany them.[2]

In this chapter, we chose to concentrate on approaches to the modeling of perceptual learning, rather than on its phenomenology or on specific models — the standard fare of the reviews one finds in the literature (Gibson, 1969; LaBerge, 1976; Walk, 1978; Barto, 1989; Gallistel, 1990; Gluck and Granger, 1993; Berry, 1994; Gilbert, 1994; Sagi and Tanne, 1994; Ahissar and Hochstein, 1998). In perceptual learning, of course, periodic reviews are as important as in any other discipline blessed with a steady stream of empirical findings. In such reviews, however, the stress is frequently on comparing the relative merits of learning mechanisms, at the expense of the attention devoted to its computational theory (Marr and Poggio, 1977). This preoccupation with mechanisms reflects the classical methodological stance, codified by (Popper, 1992), according to which empirical studies should begin with a discussion of the models about to be tested, and should end by refuting some of the models.

It is indeed easier to refute a specific model, mechanism or wiring diagram than to gain support for a general theory. Nevertheless, a field of study stands to gain more from the latter endeavor:

a good theory provides an *explanation* for the observed phenomena, potentially subsuming an entire range of models of the underlying mechanisms within the same formal framework (Deutsch, 1997). Following this line of reasoning, an understanding of perceptual learning would involve, first and foremost, addressing basic questions such as: *what does it mean, from a general information-processing standpoint, for a system to learn something?* Note that the answer "to learn a perceptual task means to acquire an adequate low-dimensional internal representation of the stimulus set" would be proper in this case (even if it eventually proves to be factually wrong), because it is coached in general information-processing (in the terminology of (Marr and Poggio, 1977), computational) terms. In comparison, the identification of learning with growing some extra dendrites would constitute a category mistake — as would the seemingly more abstract statement that learning is the recruitment of extra memory, unless the need for this memory is explained in functional terms that are algorithm- and implementation-neutral.

In the remainder of this chapter, we attempt to follow the explanation route, by surveying computational theories of learning, with the aim to identify the dimensions along which varieties of learning can be classified. Accordingly, the following four sections discuss (1) the goals of learning, (2) the mechanisms that can support learning, (3) the cues that a learning system can rely upon in attempting to improve its performance, and (4) the paradigms or metaphors used to describe learning computationally. Our hope is that the brief review of the computational underpinnings of learning presented here will have made both the relationships among the existing models and the current trends more readily apparent.

2 Goals of learning

The main task-level distinction that is conceptually prior to any discussion of the mechanisms of learning is the one between merely exercising memory and acquiring the ability to process new stimuli and solve new problems on the basis of the experience with familiar ones.

2.1 Memorization

Early quantitative studies in the psychology of the acquisition of declarative information employed as stimuli lists of items that had to be committed to memory. For example, the subjects could be asked to memorize lists of nonsense syllables, whose recall was subsequently tested by the experimenter. In this setting, popularized by (Ebbinghaus, 1885), rehearsal of the stimulus is certainly a sensible strategy, repetition being the mother of learning.[3] Likewise, in learning that can be described as procedural (i.e., learning to perform a perceptual discrimination or a motor task), repetition was found early on to bring about an improvement of performance; cf. the discussion of Volkman's 1858 study of cutaneous spatial acuity in (Gibson, 1969).

In the century and more that elapsed since the pioneering work of Volkman and Ebbinghaus, repetition was shown to lead to improved performance virtually in every perceptual and motor domain tested. At the same time, the scope of memorization as a paradigm for learning was gradually revealed to be limited. Specifically, it became clear that performance gain stemming from repetition is transferred only partially to novel situations (Gibson, 1941; Ellis, 1965). The extent and the nature of the transfer depends on the relationship between the sets of perceptual stimuli (or the repertoire of movements in motor learning), and between the tasks defined over these stimuli, in the original and novel situations (Osgood, 1949). To cite some relatively recent examples, limited transfer was reported by (Fiorentini and Berardi, 1981), who found that practicing discrimination between spatial contrast gratings at one orientation does not improve the performance at

an orthogonal orientation (see the chapter by Fiorentini and Berardi in this volume). An analogous situation prevails in motor learning; for example, an acquired ability to perform precise elbow flexions transferred only partially from one set of joint angles to another (Gottlieb et al., 1988).

The distinction between memorization and transfer is of crucial importance to any theory of learning. A theory that fails to make this distinction succumbs to the same confusion that surrounds the much-publicized inability of an early neural-network model of learning, the perceptron (Minsky and Papert, 1969), to solve the exclusive-OR (XOR) problem (see Figure 1). This problem is special, because every nearest neighbor of each input belongs to the opposite class, and, therefore, no cluster structure exists here (that is, no similarity structure, where nearby points belong to the same class).

The prospects of perceptrons (and of any models of perceptual learning that share their limitations) would be indeed bleak if real-life scenarios tended to resemble the XOR setup, in which generalization is ill-defined (Bishop, 1995). However, as we shall argue next, such learning scenarios, which focus on memorization — i.e., those which test subjects (1) in a fixed task, and (2) with the same stimuli encountered during the learning phase — cover only a part of the great variety of daily-life situations in which learning is known to occur.

2.2 Generalization

The behavioral importance of transfer of learning, or generalization, to novel conditions (as contrasted with memorization), has been pointed out and discussed by philosophers, psychologists and neurobiologists. In philosophy, the "naturalistic" approach to epistemology (Kornblith, 1985) involves the concept of natural kinds — categories of objects which share sufficiently many features to support associative recall, inference (prediction) of unobserved properties, and valid generalization from one stimulus to another (Quine, 1969; Dretske, 1995).

In psychology, empirical data on stimulus generalization gathered in the 1940's and 50's prompted Guttman (1963) to pose the explanation of generalization as a central theoretical challenge. Shepard attempted to meet this challenge by incorporating a hypothesis about generalization gradients into the foundations of a "universal law" that describes quantitatively the relationship between the likelihood of two stimuli receiving the same response, and their perceived similarity (Shepard, 1987). More precisely, Shepard showed, on the basis of data from a wide range of perceptual experiments, that stimuli in each experiment can be arranged in a low-dimensional metric feature space in such a manner that the probability of generalization between any two stimuli is monotonic in their proximity (i.e., similarity). Shepard's treatment of this issue included a derivation of the monotonic dependence law from some basic assumptions on the probability measure used to quantify generalization.

In theoretical neurobiology, the notion of generalization underlies the "Fundamental Hypothesis" stated in Marr's theory of the cerebral neocortex (Marr, 1970): "Where instances of a particular collection of intrinsic properties (i.e., properties already diagnosed from sensory information) tend to be grouped such that if some are present, most are, then other useful properties are likely to exist which generalize over such instances. Further, properties often are grouped in this way" (pp. 150-151). Although this hypothesis seems at present every bit as convincing as it must have appeared to Marr, it remains, unfortunately, empirically unsubstantiated; its vindication (or refutation) is likely to affect the explanatory power of statistical theories of brain function, such as those of (Uttley, 1959) and (Marr, 1970), and of their recent successors — neural network theories (more on this below).

The foundational status of generalization in visual perception and cognition can be easily illustrated in intuitive terms, on an everyday task. Consider, for example, learning to recognize a

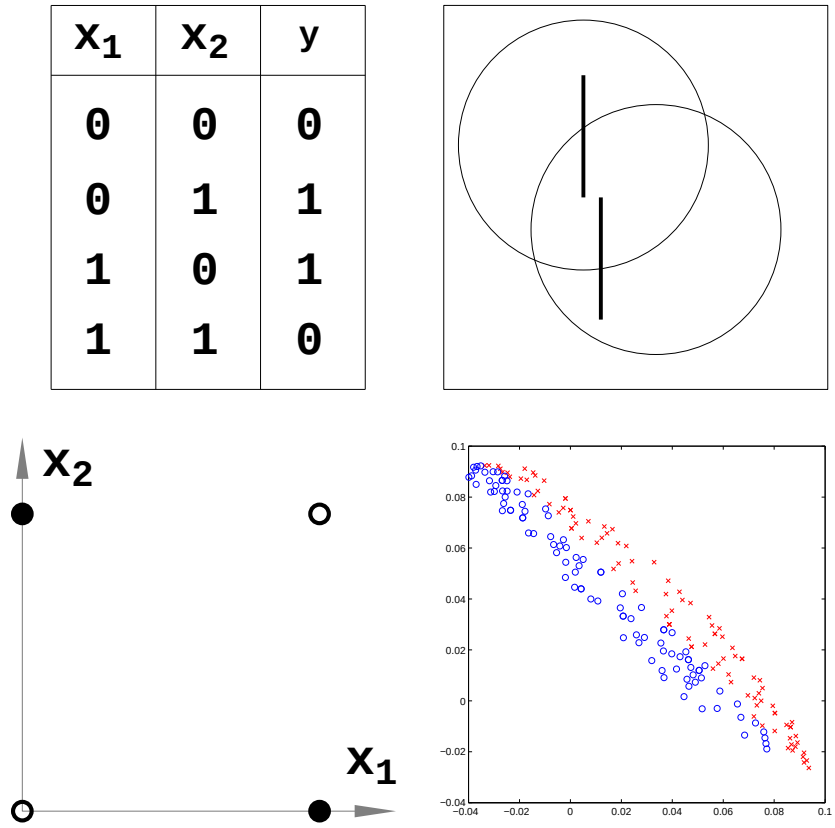

Figure 1: *Left:* the truth table definition of a 2-variable exclusive-OR (XOR) problem. In the lower panel, note how the points belonging to the two classes are interspersed among each other. The common characteristic of this kind of problem, which impedes generalization, is that the probability of two neighboring points belonging to the same class is at chance. *Right:* the vernier discrimination problem. The points in the lower panel are representations of 200 vernier stimuli in the space of the outputs of the two Gaussian receptive fields whose size and position is outlined in the upper panel (cf. Poggio et al., 1992). The simulation that produced this plot used a 100×100 image, and two Gaussian filters with $\sigma = 30$, positioned at $(70, 30)$ and $(30, 70)$. The two symbols, $\circ$ and $\times$, correspond to the two senses of vernier displacement, making this a class-conditional probability density plot of a sort. The vernier displacement in this experiment ranged from 5 to 15 pixels. The crucial characteristic here is the clustering of points that belong to the same class; the clusters, moreover, are simply connected and unimodal (human observers find learning more difficult when the class-conditional distributions are disjunctive or multimodal (Flannagan et al., 1986)). It seems safe to conjecture that in general the class-conditional densities arising from perceptual tasks can be relatively easily made to look like this, facilitating decision-making (unlike the case of the XOR example). In this chapter, we argue that learning can be construed as the formation of a representation space in which the problem at hand is well-behaved in this sense.

face from several snapshots. The observer's ability to recognize the face in this case most probably extends to new images (obtained, e.g., under various combinations of viewpoint and illumination). Moreover, the observer is expected to be able to solve a range of perceptual problems involving that face (e.g., to estimate its direction of gaze, to categorize its various expressions, etc.). These latter abilities effectively require that learning be transferred from one set of stimuli (i.e., images of the many faces previously processed by the subject) to another (i.e., images of the new face). Thus, any theory of perceptual learning must include a component that would account for generalization across stimuli and across tasks, over and above rote memory (see Figure 2). The central role of generalization in learning underscores the importance of experiments such as those of Fiorentini and Berardi, which define the limits of generalization of the human perceptual system, and thereby make a crucial contribution to the discovery of the principles and the mechanisms that support it.

3 Mechanisms of learning

The characteristics of learning that we discussed so far had to do with the nature of the task, where the main distinction is between memorization and generalization. We next raise the issue of the *mechanisms* employed by models of learning to explain the improvement of the performance with practice. At the highest level of abstraction, a common (albeit, as we shall argue, not entirely warranted) distinction is between symbolic mechanisms and neuromorphic ones. Whereas in symbolic learning the building blocks of models are propositions and rules (Carbonell et al., 1983), the components of neural models of learning are activation states of simple computing elements, and their interconnection patterns (Selfridge, 1959; Hinton, 1989; Rumelhart and Todd, 1993).

It is now widely realized that the principles of operation of neural learning models apply also to more traditional computational paradigms and data structures (Omohundro, 1987). Even more importantly, neural networks turn out to be amenable to mathematical analysis that invokes well-established statistical tools dealing with inference and decision-making (Bishop, 1995). For example, Widrow's Adaline (adaptive linear element) networks can be identified with linear discriminant functions, and multi-layer perceptrons — with multivariate multiple nonlinear regression. Further parallels between the neural network terminology and that of statistics can be found in (Sarle, 1994).

Inferential statistics thus constitutes a useful foundation for the understanding of the computational capabilities of neural networks. If this foundation is to be useful in the development of specific models of learning in the nervous system, statistical samples of stimuli must be shown to contain information necessary for learning. Having been downplayed for decades by the work of Chomsky and his school, the notion that statistical inference can support learning even in markedly "symbolic" domains such as language acquisition is now making a comeback. On the one hand, this process is aided by the growing evidence that humans (both adults and infants) are sensitive to statistical cues present in linguistic stimuli. For example, subjects can extract from such cues, implicitly, information about boundaries between the underlying morphological units (Saffran et al., 1996), word meaning (Markson and Bloom, 1997), and even grammar-like rules (Berns et al., 1997). These findings raise doubts concerning the exclusive applicability of symbol-manipulating computational models to learning problems hitherto coached in purely symbolic terms. Consequently, models built around symbolic abstraction (rule inference) now have to compete routinely with models that posit similarity-based processing (Berry, 1994; Goldstone and Barsalou, 1998).[4]

In visual perception, 3D object recognition is one domain where the hegemony of symbol manipulation models is increasingly challenged by statistical/connectionist learning approaches. Visual recognition gives rise to a variety of learning-related tasks, similar to those encountered in the

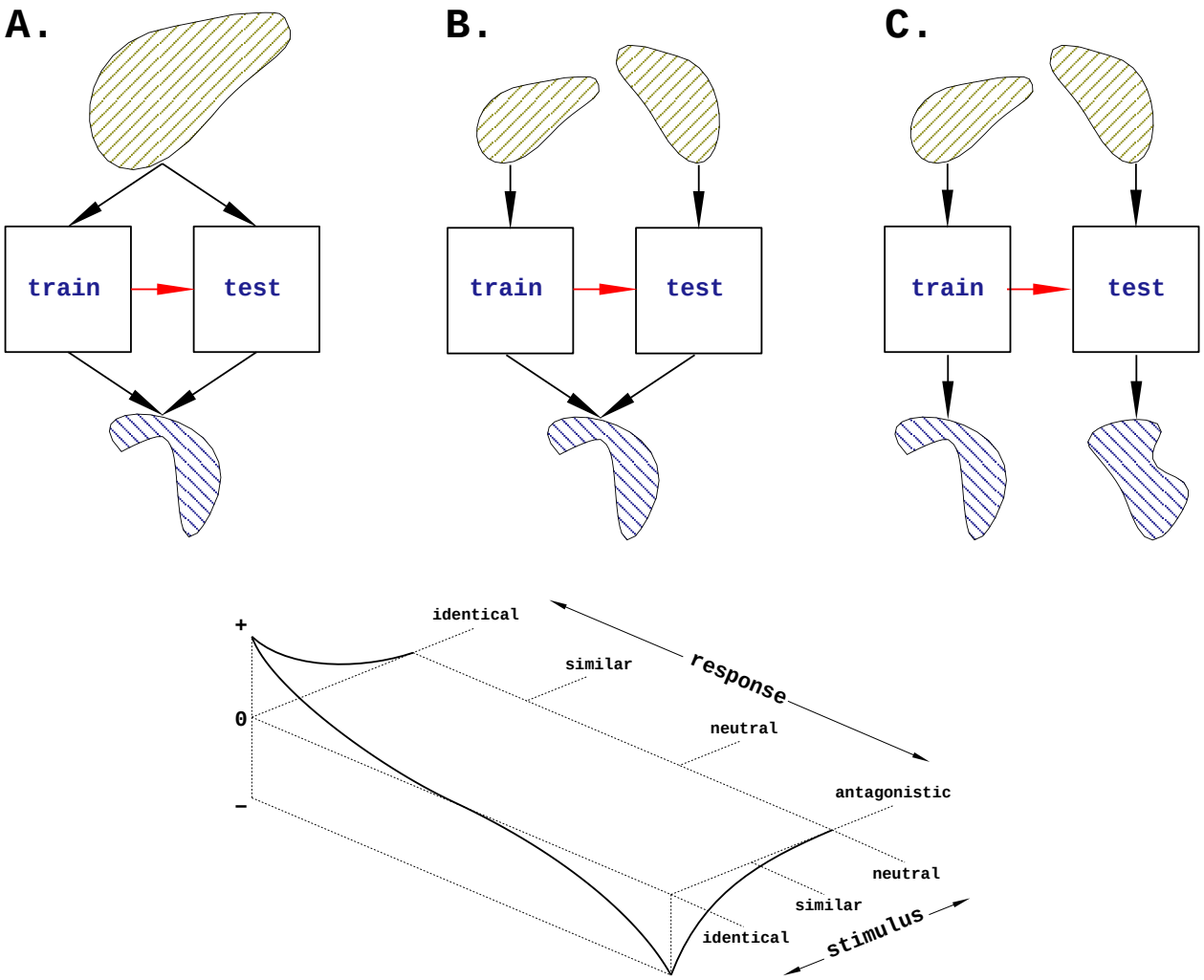

Figure 2: *Top:* A schematic illustration of three varieties of data- and task-related situations in learning. *A (left):* When the inputs and the required outputs are the same in both training and testing phases, learning amounts to a memorization of the input-output association. A simple example here is learning to associate a name with a face. The arrow leading from the `train` to the `test` box represents the parameters acquired by the adaptive mechanism during the learning process. *B (middle):* If new data are to be mapped into the same output space, the system must generalize the previously learned association. Examples: naming a familiar face seen under novel conditions (such as peculiar illumination); categorization of a novel instance of a class of stimuli. *C (right):* The problem turns into that of transfer of learning to a new task, if both the input and the output spaces change between training and testing. Example: matching two views of an unfamiliar face, on the basis of prior experience with other face stimuli. *Bottom:* The dependence of transfer on the relationship between the characteristics of two tasks, adapted from (Osgood, 1949). The degree of transfer grows with the similarity between the stimuli in the two tasks, and, for highly similar stimuli, is reduced if the required responses are different.

context of face processing (mentioned briefly above). In these tasks, the lure of symbolic compositional models stems from the observation that for many object classes the main challenge inherent in learning recognition — achieving invariance over transformations or deformations of the stimulus — disappears if the objects are represented structurally (Biederman, 1987). This observation leads to the postulate that learning to recognize an object entails the identification of its parts and the determination of their spatial relationships. Under this assumption, the possession of a library of generic parts that can be assembled in various ways would also endow the system with the ability to represent and process novel objects — the ultimate kind of generalization.

Recently, a different route to invariance and to the ability to process novel shapes has been proposed and implemented in a series of models (Poggio and Edelman, 1990; Edelman and Duvdevani-Bar, 1997; Riesenhuber and Poggio, 1998; Edelman, 1998a; Edelman and Intrator, 2000). The computational underpinnings of this alternative approach are discussed elsewhere (section 5.2; see also the chapters by Sinha and Poggio, and by Bülthoff and Wallis in the present volume). For now, we shall take the encroachment of alternative learning models into territory hitherto reserved for structural methods as a license to focus our review on neural, rather than symbolic, computation.

4 Cues for learning

A central question to be addressed in the modeling of a perceptual learning task is that of *supervision*: what are the sources and what is the form of the information that guides the learning process? The usual distinction found in the literature is between supervised models, which require each training stimulus to be accompanied by the desired output, and unsupervised ones, which are able to extract some statistical information from the data, without guidance.

Classifying an experimental setup on the basis of supervision available to the learning model is, however, not always as straightforward as it may seem. Even when the learning process is fully and explicitly controlled by the experimenter, the subject has access to, and is likely to make use of, information that transcends the experimental design. For example, when the subject is required to learn the names of some unfamiliar faces, the ensuing confusion rate will be higher among some faces compared to others.

The reasons for the advantage of some stimuli over others are simple, and may have nothing to do with the labels (i.e., names) provided explicitly by the experimenter. Rather, they stem from the choice of faces selected for the experiment, and from the computational make-up of the subject's visual system. The necessary commitment on the part of any system to the use of some features at the expense of others gives rise to representational idiosyncrasies. These, in turn, cause some faces to be mapped into more crowded regions of the internal "face space" (Valentine, 1991; Edelman, 1998b) than others, and to be subsequently more easily confused. Analogous phenomena are observed in situations that require the subject to generalize, say, from one view to another. This generalization is easier to learn for typical than for atypical faces (Nosofsky, 1988; Newell et al., 1999), typicality being defined as proximity to the center of the cluster corresponding to the set of stimuli in the face space.

4.1 Unsupervised learning

The human visual system is clearly able to group stimuli by similarity that is intrinsic to a given test set and is influenced by the context, but does not require that class membership be dictated or revealed by an external supervisor. The information processing task in this case is known

as (unsupervised) clustering (Duda and Hart, 1973); it constitutes one of the most challenging problems in computational learning.

Both "hard" and "soft" or fuzzy versions exist for many clustering algorithms. In hard clustering, each point can only belong to one cluster, while fuzzy algorithms allow a graded membership index. The latter approach is probably better suited to the modeling of perceptual categorization decisions: the boundary between classes in forced-choice tasks is never clear-cut. The dependence of the probability of membership in a given category on the location of the stimulus that falls "in between" clusters is typically sigmoidal. The slope of the sigmoid in the transition zone may be quite steep, a phenomenon that is manifested both in discrimination and in identification tasks, and is known as categorical perception (Harnad, 1987).

In general, clustering algorithms decide whether or not to attribute points from a given data set $\{\mathbf{x}_i\}$ — in the present context, descriptions of stimuli in some feature space — to the same category on the basis of interpoint distances.[5] A typical algorithm employs standard combinatorial optimization techniques to minimize some function of V_w/V_b, where $V_w = \sum_k E\left[\mathbf{x}_k - \bar{\mathbf{x}}_k\right]^2$ is the total within-cluster variance ($\mathbf{x}_k \in C_k$, the latter being the k'th cluster), and V_b – the between-clusters variance, for the given data set $\{\mathbf{x}_i\}$.

In some situations, prior knowledge about the nature of the problem provides hints that make unsupervised clustering easier. Consider, for example, the problem of learning to recognize several faces from their snapshots. The formation of the initial representation here is equivalent to clustering the snapshots (in whatever feature space the system uses as its front end). This clustering will be facilitated if the snapshots appear in a natural order, corresponding to the succession of views seen by the subjects, as the head to which the face belongs revolves in front of them. Such facilitation can be supported by a Hebbian learning mechanism, whereby association between successive views is made stronger in proportion to their temporal proximity (Hinton, 1989; Stone, 1996).

4.2 Supervised learning

Whereas unsupervised algorithms have only the feature-space distances between the stimuli to go by, in the supervised scenario the training data is a set of input-output pairs: $T = \{(\mathbf{x}_i, \mathbf{y}_i)\}$. The model in this case is required to learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ such that $\forall i, \mathbf{y}_i = f\left(\mathbf{x}_i\right)$.

For the purpose of memorization, the functional requirement imposed on the learning algorithm in the fully supervised case is fulfilled by a lookup table, or, equivalently, by an associative network (Willshaw et al., 1969); cf. Marr's work on simple memory (Marr, 1971). If generalization is required (i.e., if the model needs to estimate $\hat{\mathbf{y}} = f(\mathbf{x})$ for some $\mathbf{x} \notin \{\mathbf{x}_i\}$), the problem becomes that of interpolation (or, as it were, extrapolation, depending on the location of the new $\mathbf{x}$ relative to $\{\mathbf{x}_i\}$) of the function f.

An efficient distributed implementation of function interpolation exists in the form of regularization networks (Poggio and Girosi, 1990; Girosi et al., 1995). Of these, radial basis function networks (RBFs) constitute a special case, which has certain biological appeal (Poggio, 1990). The simplest Gaussian RBF model approximates the input-output mapping f as a superposition of basis function values: $y = f\left(\mathbf{x}\right) = \sum_i c_i e^{-\|\mathbf{x}-\mathbf{x}_i\|^2/\sigma_i^2}$; extensions to vector-valued output, other basis functions, and other input-space metrics exist.

4.3 Self-supervised learning

An interesting variation on the supervised learning scenario is found in the idea of auto-association: setting the output of a learning system to be identical to its input, while forcing the input-output

mapping to pass via an intermediate representation stage. This idea can be implemented conveniently by a three-layer perceptron in which the number of units in the middle ("hidden") layer is smaller than in the input/output (Cottrell et al., 1987). If such a "bottleneck" network is trained successfully on a set of data, the activities of its hidden units form a reduced-dimensionality representation of the data. In traditional theories of perception, this coding or feature extraction operation is held to be a central characteristic of learning (LaBerge, 1976).

Intuitively, having access to the right features is expected to facilitate greatly the performance of the system (Gibson, 1969). More formally, the effectiveness (and, indeed, the very feasibility) of learning from examples depends critically on the availability of a low-dimensional description of the problem. Such a description facilitates learning because it avoids the *curse of dimensionality* — the exponential dependence of the number of examples needed for valid generalization on the number of dimensions of the representation space (Bellman, 1961; Stone, 1982; Huber, 1985).

5 Paradigms of learning

Psychologists had been aware of the central role of feature discovery in learning for a long time before any systematic attempts to cast this idea into computational terms were made.[6] To cite a prominent example, Eleanor Gibson (1969) concluded in her review of perceptual learning that the skilled perceiver is able to gain extra information from the stimulus, by detecting features and "higher-order structure" to which the naive viewer is not sensitive. In this section, we shall offer a straightforward (by the present standards) computational formulation of this insight, using the notion of the formation of new representations. Two distinct kinds of representation will then be considered, the first suitable for regression problems, and the second — for problems involving classification (the former are stated in terms of continuous variables, while in the latter the variables are usually discrete; see Figure 3). Finally, a common framework that subsumes both these kinds of tasks will be examined.

5.1 Learning as formation of new representations

The power of forming new representations, or, more specifically, of adjusting the representation to the problem at hand, stems largely from the singular computational advantage conferred by the possibility of a linear solution. In regression tasks, such a solution is possible when a high linear correlation exists between the variables; in classification — when the classes are linearly separable. Thus, a representation is good if it allows a linear solution.

Consider the example illustrated in Figure 4, where the task is to discriminate between the interior and the exterior of a circle. Solving this problem using a linear mechanism (left panel) is difficult, because of the need to create many instances of such a mechanism to approximate the curved decision boundary. Thus, a system that tries to solve this problem in the original representation space will use more resources and is likely to take longer than a system that is confronted with the same problem stated in polar coordinates (Figure 4, right), which can be solved using a linear mechanism. This distinction has implementational parallels: the representation on the left is suitable for interpolation by "bump" functions (e.g., radial basis functions or RBFs), while the representation on the right is suitable for "ridge" functions, such as those implemented by the inner-product unit in a perceptron (or the hidden units in a multilayer perceptron, or MLP).

If a linear solution is ruled out in the original space, and if the system is biased towards using linear mechanisms, it may attempt to learn a new, better representation, that is, to remap the data into another representation space, in which linear regression or separability is possible. It should be

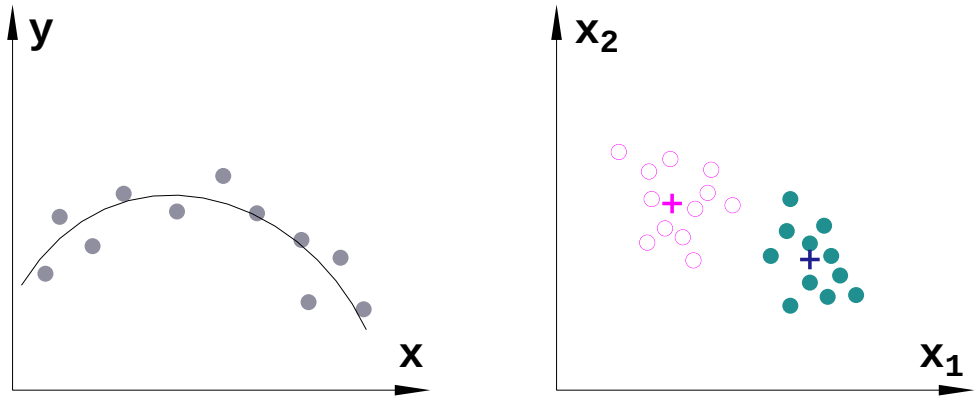

Figure 3: *Left:* a regression problem. When expressed in terms of probabilities, the performance goal of the learning system in this case can be stated as the estimation of the posterior probability $p(y|x)$. *Right:* a classification problem. Here, the performance goal can be formulated as the estimation of $P(C_k|\mathbf{x})$, where C_k signifies membership in class k, and $\mathbf{x} = (x_1,\ x_2)^T$. In either case, the posterior probabilities can be learned directly (as in some neural network models), or computed from class-conditional probabilities estimated during a prior learning stage using the Bayes theorem: $P(C_k|\mathbf{x}) = p(\mathbf{x}|C_k)P(C_k)/p(\mathbf{x})$. See (Bishop, 1995) for an accessible exposition of these approaches.

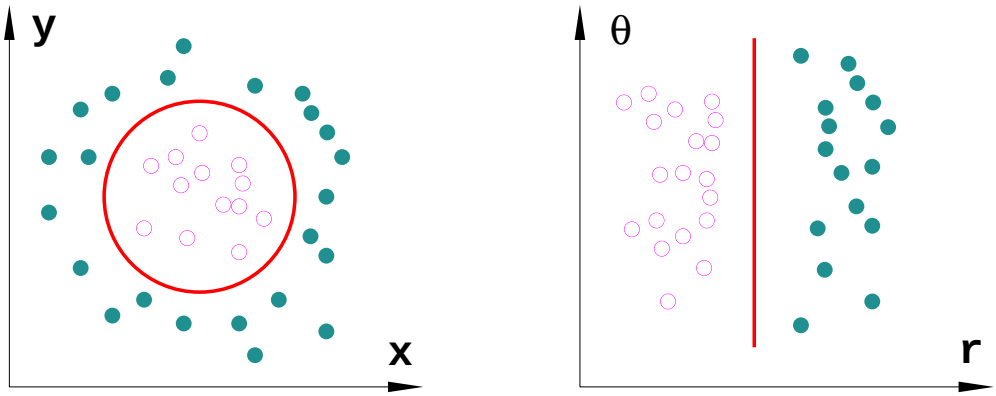

Figure 4: A nonlinear discrimination problem (left) is translated to a linear one via a coordinate transformation (right). The mapping $(x_1, x_2) \leftrightarrow (r, \theta)$ is itself nonlinear ($x_1 = r\cos\theta;\ \ x_2 = r\sin\theta$).

quite obvious that a universal mechanism capable of such a remapping would go a long way towards solving just about any problem posed to it (because of the relative ease of solving the linear version of the problem in the new representation space). Some recently developed linearization methods do claim such universal applicability. For example, the support vector machines for classification and regression remap the original problem into an extremely high-dimensional space where it becomes linear; the curse of dimensionality is avoided by making the solution dependent on a small number of examples (the "support vectors") that lie close to the regression surface or to the class boundary (Vapnik, 1995).[7] If the human visual system uses anything like the support vector algorithm, it should learn faster when given the more informative (and more difficult) examples first. We could not find evidence to this effect in the perceptual learning literature: human vision seems to be rather more limited than a support vector machine in this respect.

Interestingly, the attempts on the part of a learning system to improve its performance under the limitations imposed by its architecture are precisely what may pass as the manifestation of the *process* of learning to an external observer. For example, a system constrained to use RBFs may start by "tiling" the inside and the outside of the circle in Figure 4, using up many basis function units. It may then shift to the more economical representation in which the inside of the circle (and, therefore, its outside) is represented by a single basis — an event that would look like the kind of feature discovery mentioned by E. Gibson. In other words, learning can be defined computationally as the art of creating the most suitable representation of the data, given the constraints of the model at hand.[8]

5.2 Learning visual manifold geometry (regression)

For a resource-constrained system, attempting to solve the learning problem in the original feature space may prove too complicated, as illustrated by the application of RBFs to the circle example. At the same time, attempting to remap the original problem into a new representation space where it would become linear may be equally hard, rendering the problem intractable. Fortunately, it appears that many perceptual problems possess an inherent structure that makes them amenable to learning methods that rely neither on the exhaustive tiling of a high-dimensional space (an intractable undertaking, because of the curse of dimensionality), nor on a sophisticated remapping (which may be beyond the visual system's capabilities). The structure in question is that of a smooth low-dimensional manifold, and it arises because (1) problem spaces in a typical perceptual task are parameterized by only a few variables (Edelman and Intrator, 1997), and (2) the "front end" of a typical visual system — its *measurement space*[9] — largely preserves the local geometry of the distal problem space (Edelman, 1999).

The first of the two observations that we just stated can be illustrated by a series of examples taken from all areas of perception. Consider, for example, the vernier discrimination task, in which the observer is to judge the sense of the relative displacement of two abutting line segments (see Figure 1, right). The solution in this case is parameterized by a single variable, which controls the displacement of the segments perpendicular to their extent. Stepping this variable by small increments through a range of values between, say, $-15''$ and $+15''$ would cause the measurement-space representation of the resulting stimulus to ascribe a 1-dimensional manifold.[10] All that a visual system would have to do to learn vernier discrimination would be to interpolate this manifold from a set of examples (i.e., input-output pairs), as described in section 4.2.

As another illustration, we may consider the problem of learning to recognize an object from examples (i.e., a few of its stored views, each view being construed as a snapshot of the multidimensional measurement space). For a rigid object which is allowed two rotational degrees of freedom, corresponding to the two axes of rotation in depth, the manifold spanned in the measurement space

will be 2-dimensional, and will be amenable to learning from examples, as shown in (Poggio and Edelman, 1990) An entire range of other learning tasks having to do with object recognition can be solved on the basis of related principles, as shown in (Edelman, 1999; Duvdevani-Bar et al., 1998).

In psychology, the well-behavedness of the internal representation space of various visual qualities has been remarked upon repeatedly, beginning with the shift towards representational theories of vision in the 1960's. The concept of an internally represented stimulus space was mentioned by (Guttman, 1963), who pointed out, in a discussion of generalization in animal learning, that the pigeon "knows the spectrum, in an important sense of the word 'know'" — it exhibits the kind of orderly generalization between colors that psychologists routinely observe, and must, therefore, possess an internal "color-space." More than two decades later, Shepard (1987) formulated his law of generalization in terms of proximities in an internal psychological space, observing that structure inherent in various distal "quality spaces" (Clark, 1993), such as the color continuum discussed by Guttman, is faithfully represented internally.

It is interesting to compare the observations concerning the low dimensionality and the smoothness of the internally represented quality spaces to similar observations made by statisticians and neural network researchers. In nonparametric statistics, for example, the surprisingly good performance of nearest-neighbor methods, which rely on raw feature-space distances, has been explained intuitively in terms of the fact that the relevant points in these spaces — that is, the examples — tend to be confined to smooth low-dimensional subspaces (Friedman, 1994). In neural network research, an analogous observation can be found in (Bregler and Omohundro, 1995).

Realizing that the problems at hand typically possess such a convenient structure is, however, only the first step towards their solution. The reduction of dimensionality of the space in which the stimuli are originally encoded is a nontrivial operation; as we pointed out above, in human visual perception the original dimensionality of any stimulus is, nominally, on the order of 10^6, which is the number of fibers in each optic nerve. The illustration in Figure 3, left, in which the manifold is embedded in 2-dimensional space, is, therefore, *highly* simplified. Many of the computational approaches devised for the extraction of low-dimensional manifolds do not scale well with the dimensionality of the embedding space. This includes the Self-Organizing Map algorithm (Kohonen, 1982), and the different varieties of auto-encoders, or bottleneck networks (Cottrell et al., 1987; Leen and Kambhatla, 1994).

The performance of such unsupervised or self-supervised manifold-extracting algorithms can be improved if additional knowledge is brought to bear on the problem. Typically, this is done by making the learning mechanism observe certain invariances known to apply to the problem (Foldiak, 1991; Wiskott, 1998). A particularly simple way to do that is to provide the label of the category to which each stimulus belongs.[11] To see how this information helps the algorithm to isolate the relevant manifold, note that directions orthogonal to it can be effectively specified by forcing stimuli that differ along those directions to be mapped to the same category (Intrator and Edelman, 1997).

5.3 Learning visual category structure (classification)

From the perspective of the task, the main difference between regression and classification is that in the latter, the location of the point within the low-dimensional structure does not matter, while in the former it does. For example, the location of the point representing a face in a face space (the manifold corresponding to the different possible views of the same face) would encode its orientation — a piece of information that should not be discarded. In comparison, in the vernier task, where the problem is that of classification, only the membership in one of the two clusters in the representation space matters to the system.

Despite this difference, the basic considerations identified before in the discussion of regression apply also to classification. In particular, the curse of dimensionality still has to be taken into account. Huber (1985) illustrates this point quantitatively, by showing how difficult it is to find a 3-dimensional Gaussian bump (which could, in terms of Figure 3, right, correspond to one of the class-conditional clusters), when it is embedded in a 10-dimensional space. Although neurally inspired models of learning that are tailored specifically for categorization and mixture estimation do exist (Carpenter and Grossberg, 1990; Carpenter et al., 1992; Williamson, 1997), they are not expected to deal better with realistically high-dimensional cases than the knowledge-based models mentioned earlier (which were designed for manifold extraction, yet should be equally capable of clustering).

5.4 Learning joint input-output probability density

When learning is treated as a problem in statistical inference, the observations we made in the last few paragraphs can be rephrased using the concept of the *underlying generator* of the data — the entity that causes whatever regularities are present in the data set. In visual perception, this entity is the distal stimulus, which gives rise to the observed values of features through a complex process (reflection and scatter of light, propagation in the medium, refraction by the optics of the eye, phototransduction, etc.). In regression-like tasks, the distal stimulus space is continuous by nature, e.g., the continuum of views of an object that undergoes rotation in front of the observer (in Figure 3, left, it is the smooth curve underlying the sausage-like cloud of points). In classification, the distal stimulus space is discrete, e.g., the set of categories to which the viewed object may belong (in Figure 3, right, this space consists of the centroids of the two clusters of data points).

Thus, both regression and classification can be subsumed under a common framework, which calls for estimating the joint probability density of all the variables included in the data set. It is well-known that this information reveals everything that there is to know about stochastic data, such as the measurements performed by a perceptual system upon the world. The underlying generator of the data (and, therefore, quantities needed for regression or classification) can then be estimated optimally from the density function. However, the first step in this process — the inference of an unconstrained density function from data — is prone to the curse of dimensionality, as shown in the seminal work of Stone (1980; 1982).

In view of this problem, researchers typically take two approaches, which are not mutually exclusive. The first is to make some assumptions about the density function. For example, one may assume that the density function is smooth, then estimate it using splines (Wahba, 1979) or radial basis functions (Poggio and Girosi, 1990). Alternatively, one may assume something about the structure of the density. For example, it may be postulated to belong to an additive model, making it expressible as a sum of functions of some low-dimensional projections of the data (Stone, 1985; Stone, 1986). One may also assume that the density is factorial, namely, a product of marginal densities of one variable (Dayan et al., 1995). The latter two methods do not attempt to reduce the dimensionality of the density function, yet they do make the estimation process more efficient and less prone to the curse of dimensionality.

The second general approach bypasses the problem of density estimation, rather than attempting to solve it. It is based on the observation that for many practical problems only a certain function of the density is required. The hope is that such a function can be easily computed directly from the data, without the need to go via the full density estimation. This happens, e.g., when the desired function is defined over a low-dimensional manifold embedded in the original space, or, more generally, when the desired function has a simpler structure compared to the full density. In such cases, the learning system may attempt to extract the low-dimensional representation of the

problem from the data, using an unsupervised approach such as principal component analysis and its generalizations, or using a supervised approach tailored to the desired target function, as it is done in many feed-forward network models.

In all these cases, a model would do well if it applies the methods listed in the preceding sections, which dealt with learning manifold extraction (regression) and clustering (classification). Reverting to those methods means, effectively, that their subsumption under the aegis of density estimation is not practical, unless the estimation algorithm (1) aims for learning a certain target function of the density, which is usually problem-specific, and (2) relies on some prior assumptions about the properties of the desired representation, such as low dimensionality and smoothness (Intrator, 1993).

6 Discussion

The theoretical stance adopted so far in this chapter equates learning with the acquisition of efficient representations, a computational procedure that can be regarded as a kind of statistical inference. It may seem that exploring the implications of this stance have lead us away from the gritty details which must be dealt with by any model that aims to simulate human learning behavior. We believe, however, that a good model starts at the top, with a clear notion of what is being modeled, and why. In this concluding section, we recapitulate the links between computational theory and mechanism-level practice in the modeling of perceptual learning, and speculate about the possible future developments in the modeling of perceptual learning.

6.1 On the levels of explanation of learning

The overarching concern in the modeling of a perceptual phenomenon is, of course, getting the performance right. Beyond that, however, there is a considerable variation in what is deemed acceptable: while some comprehensive models treat both the computational (theoretical) and the implementational aspects of the problem, others tend to concentrate on the issues of implementation and mechanism. Models built around neural networks are particularly likely to belong to the second category, going straight from the phenomenology to a hypothesis about the underlying mechanism, perhaps also attempting to emulate along the way the *real* biological neural network.

We illustrate this observation with an example that involves one of the most striking manifestations of perceptual learning, found in the task of detecting a small low-contrast Gabor patch projected onto a certain retinotopically defined location. The detection threshold in this task depends on whether or not the target patch is flanked at a distance by patches of similar orientation and spatial frequency (Polat and Sagi, 1993). Shortly after the effect of the flanking patches has been demonstrated, it turned out to be amenable to learning: the spatial range of the effect (i.e., the maximum effective distance between the target and the flanking patches) grows with practice (see Sagi and Zenger, this volume). Significantly, learning is only possible if the original, untrained range is extended gradually, by exposing the subject to configurations of progressively larger and larger extent (Polat and Sagi, 1994).

A phenomenon such as this seems positively to demand a mechanism-level explanation in terms of receptive fields of retinotopic "units," linked laterally and exerting facilitatory influence on each other; (Polat and Sagi, 1994) offered precisely this explanation for their psychophysical findings. However, as we claimed in the introduction, models formulated primarily in the language of units and connections achieve less than what a model can and should achieve, because they concentrate on the wiring details at the expense of leaving the master plan — the computational goal of the

system — out of the picture. To support this argument, let us re-consider the "lateral learning" scenario, keeping in mind the taxonomy of learning paradigms discussed earlier.

Assume for the moment that the goal of the system is to detect the faintest possible line element (a real-life counterpart to a Gabor patch) in a given retinal location. Merely lowering the decision threshold for that location will likely just increase the false-alarm rate there; additional information must be brought to bear on the decision, if it is to be reliable. The presence of other line elements in the vicinity would count as the necessary additional support, if they are compatible with the original hypothesis (i.e., if their orientation is consistent with that of the element whose fate they are about to seal). Thus, the task at hand can be reformulated as that of (literal) *interpolation* between the flanking lines (or extrapolation, if the continuation of an "end-stopped" segment is sought).

The value of this formulation is in that it brings about the possibility of a uniform treatment for a range of perceptual learning tasks. Indeed, on an abstract level, learning to detect a Gabor patch flanked by similar patterns is now seen to be the same as learning to recognize an object from a novel viewpoint, which is "sandwiched" between two familiar views. The analogy drawn between these two tasks hinges on a parallel between the *view space* of the object on the one hand, and the *"space" space* — that is, the retinal location space — of the Gabor patch on the other hand. Once this analogy is realized, cross-fertilization may occur in both directions. On the one hand, models of object recognition may benefit from postulating a mechanism that carries out interpolation by growing lateral links between neighboring units in a view representation space. On the other hand, models of line detection may benefit from exploring the possibilities originally developed in the context of object recognition (e.g., interpolation with feedforward basis functions).

An edifying perspective on the issue of levels of modeling is provided by recalling some of the "old-fashioned" models of brain function (and learning) produced by neurobiologists. Two such models that were prominent in their own time, one dealing with "universals" or the problem of invariance (Pitts and McCulloch, 1965) and the other with probabilistic generalization (Marr, 1970), actually *did* link theory and mechanism. Marr's (1970) model of the neocortex, for example, spans the entire possible range of levels. It starts with a general, yet succinctly phrased hypothesis concerning the probabilistic structure of the world (the "Fundamental Hypothesis," which we mentioned in section 2.2), and ends with a detailed explanation of the possible ways in which neuroanatomy and neurophysiology of the cortex may be tuned to put the observed probabilities to work. This style of modeling, which *integrates* different levels of explanation, requires a combination of encyclopaedic knowledge with considerable ingenuity on the part of the modeler. Unfortunately, it is now quite rare, having been replaced by a methodology that allows the levels — computational, algorithmic, and implementational — to be kept separate.

6.2 Prognosis

In visual perception, learning is a pervasive phenomenon, which, when properly studied, offers the researcher a unique searchlight with which the inner workings of the system can be illuminated. It would be rash to try and anticipate what a judicious use of this searchlight will reveal in the future — which is why the following remarks express, mostly, desiderata, not predictions.

Here are, then, a few issues that we would like to see addressed in the context of the modeling of perceptual learning.

- *Integrate past achievements.* Attempts to develop mathematical models of learning date back more than half a century. Much of the work carried out before mid-1960's has now been branded "behaviorist" and effectively buried in the libraries. It has been pointed out that a

re-examination of that work may lead to interesting insights into the nature of the present-day models (Hintzman, 1994).

- *Look at learning differently.* A diametrically opposite trend is that of complete rejection of both the old and the contemporary models of learning, in favor of theories that (to most psychologists) seem rather esoteric, involving concepts such as catastrophes, self-organized criticality, or merely phase transitions in dynamical systems. The drive in that direction stems from *bona fide* challenges, such as the need to explain abrupt learning and related phenomena (see Rubin, Nakayama and Shapley, this volume). Developing ways of coping with these challenges, short of rejecting all current approaches, is quite a challenge in itself.

- *Explain as much as possible.* The "dynamical" models attempt to explain the behavior of the perceptual system by appealing to an isomorphism between its physics (i.e., the differential equations that describe it) and the physics of other systems that exhibit a similar behavior. In that, they resemble the behaviorist models, which skirt the issues of representation, and deal with disembodied equations aimed at mimicking the phenomenology of the target system. Surprisingly, purely representational models too end up dealing only with the phenomenology; a good example is Shepard's law of generalization (mentioned in section 2.2), which does not make any claims as to the reality of the "psychological similarity space" that it postulates. (Shepard, 1987). In contrast to all these, explanations offered by the more daring connectionist models (those that bite the bullet and hope for the best) include parallels both on the level of behavior, and on the level of architecture.

- *Go after the big question.* We have been told repeatedly now what ingredients should go into a comprehensive explanation of an information-processing phenomenon: computation, representation and algorithm, implementation. In the context of understanding the brain, however, a really comprehensive explanation would start from a premise that is concrete, yet spans all the levels of the "hierarchy" just mentioned: a postulate as to what is it that the brain *does*. Several such postulates are on the offer, e.g., Marr's (probabilistic inference; 1970), Barlow's (redundancy reduction; 1990), Poggio's (function approximation; 1990). A more intense competition in this arena is likely to lead to some exciting developments in the modeling of learning.

We conclude this chapter by concurring with Dr. Faust: if offered the choice of treasure, go for broke. The ability to learn, at all levels and under all circumstances, is the most striking attribute of human cognition; what, then, would it take to *really* understand it? Gounod's Faust requested of Mephistopheles youth — the treasure that contains all others. So — we wish for a model of the brain, which, if successful, will make the modeling of perceptual learning as such a matter of the past.

Notes

[1]For an interesting historical perspective on these issues, see (Hintzman, 1994).

[2]Cognitive maps in the rat brain are thought to reside in the hippocampus, a cortical structure implicated in perceptual (spatial) and other kinds of learning. Information about the spatial location of the animal turns out to be represented in the firing patterns of hippocampal "place" cells, whose ensemble activity constitutes an internal cognitive map of the rat's environment (Wilson and McNaughton, 1993). Another class of cells in the hippocampus are the "head direction" cells, which serve as an internal compass to orient the cognitive map. The functional properties of place and head direction cells emerge from a complex and as yet poorly understood interaction between internally generated, self-motion cues (e.g., vestibular information) and external sensory input (e.g., visual landmarks) (Knierim et al., 1995).

[3]As in the old Latin saying, *repetitio est mater studiorum.*

[4]At a certain level of abstraction, the distinction between symbolic and "connectionist" computational paradigms ceases to make sense, as attested by the possibility of implementing rules and variables in neural networks; see, e.g., (Ajjanagadde and Shastri, 1991).

[5]The distance $d(\mathbf{x}_i, \mathbf{x}_j)$ between two points in the feature space can be assumed to vary monotonically with the inverse of the perceived dissimilarity between the corresponding stimuli. For a discussion of this issue, and of the choice of metric to be used in the computation of distance, see (Shepard, 1987).

[6]In fact, one of the most influential figures in the field, James J. Gibson, consistently resisted any attempts to invoke computation as an explanatory tool in perceptual psychology.

[7]Conceptually, the reliance on a small number of support vectors that exemplify the distinction that the system must learn is close to Winston's use of "near miss" examples in his system that learned concepts from symbolic descriptions (Winston, 1975).

[8]Note that this formulation mixes two "levels of understanding" — the abstract and the implementational — which would have been kept separate according to the methodology propounded by (Marr and Poggio, 1977).

[9]This can be defined through the notion of the vector of measurements performed by the visual system on the world, and is well-exemplified by the million-dimensional space of all possible activities of the retinal ganglion cells, whose axons constitute the optic nerve.

[10]In Figure 1, right, both the displacement of the vernier and its location in the visual field

were varied, which is why the clouds of points belonging to each of the two classes are not exactly 1-dimensional manifolds.

[11]This technique can be compared to the "stimulus predifferentiation" method, shown in the past to boost the generalization rate in learning (Ellis, 1965), p.58.

References

Ahissar, M. and Hochstein, S. (1998). Perceptual learning. In Walsh, V. and Kulikowski, J., editors, *Perceptual constancies*, chapter 17, pages 455–498. Cambridge University Press, Cambridge, UK.

Ajjanagadde, V. and Shastri, L. (1991). Rules and variables in neural nets. *Neural Computation*, 3:121–134.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571.

Barto, A. (1989). From chemotaxis to cooperativity: abstract exercises in neuronal learning strategies. In Durbin, R., Miall, C., and Mitchison, G., editors, *The computing neuron*, pages 73–98. Addison Wesley, New York, NY.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Berns, G. S., Cohen, J. D., and Mintun, M. A. (1997). Brain regions responsive to novelty in the absence of awareness. *Science*, 276:1272–1276.

Berry, D. C. (1994). Implicit learning: twenty-five years on. a tutorial. In Umiltá, C. and Moscovitch, M., editors, *Attention and Performance*, volume XV, chapter 30, pages 755–781. MIT Press.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147.

Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press, Oxford.

Bregler, C. and Omohundro, S. M. (1995). Nonlinear image interpolation using manifold learning. In G. Tesauro, D. S. T. and Leen, T. K., editors, *Advances in Neural Information Processing 7*, pages 973–980. MIT Press.

Carbonell, J. G., Michalski, R. S., and Mitchell, T. M. (1983). An overview of machine learning. In Michalski, R. S., Carbonell, J. G., and Mitchell, T. M., editors, *Machine Learning, an Artificial Intelligence Approach*, pages 3–23. Tioga Publishing Company, Palo Alto, CA.

Carpenter, G. A. and Grossberg, S. (1990). Adaptive resonance theory: Neural network architectures for self-organizing pattern recognition. In Eckmiller, R., Hartmann, G., and Hauske, G., editors, *Parallel Processing in Neural Systems and Computers*, pages 383–389. North-Holland, Amsterdam.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Trans. on Neural Networks*, 3:698–713.

Clark, A. (1993). *Sensory qualities*. Clarendon Press, Oxford.

Cottrell, G. W., Munro, P., and Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. In *Ninth Annual Conference of the Cognitive Science Society*, pages 462–473, Hillsdale. Erlbaum.

Dayan, P., Hinton, G. E., and Neal, R. M. (1995). The Helmholtz Machine. *Neural Computation*, 7:889–904.

Deutsch, D. (1997). *The fabric of reality.* Allen Lane.

Dretske, F. (1995). *Naturalizing the mind.* MIT Press, Cambridge, MA. The Jean Nicod Lectures.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis.* Wiley, New York.

Duvdevani-Bar, S., Edelman, S., Howell, A. J., and Buxton, H. (1998). A similarity-based method for the generalization of face recognition over pose and expression. In Akamatsu, S. and Mase, K., editors, *Proc. 3rd Intl. Symposium on Face and Gesture Recognition (FG98)*, pages 118–123, Washington, DC. IEEE.

Ebbinghaus, H. (1885). *Memory: A Contribution to Experimental Psychology.* Dover, New York. reprinted 1964; translated by H. A. Ruger and C. E. Bussenius 1913.

Edelman, S. (1998a). Representation is representation of similarity. *Behavioral and Brain Sciences*, 21:449–498.

Edelman, S. (1998b). Spanning the face space. *Journal of Biological Systems*, 6:265–280.

Edelman, S. (1999). *Representation and recognition in vision.* MIT Press, Cambridge, MA.

Edelman, S. and Duvdevani-Bar, S. (1997). A model of visual recognition and categorization. *Phil. Trans. R. Soc. Lond. (B)*, 352(1358):1191–1202.

Edelman, S. and Intrator, N. (1997). Learning as extraction of low-dimensional representations. In Medin, D., Goldstone, R., and Schyns, P., editors, *Mechanisms of Perceptual Learning*, pages 353–380. Academic Press.

Edelman, S. and Intrator, N. (2000). (coarse coding of shape fragments) + (retinotopy) $\approx$ representation of structure. *Spatial Vision*, -:–. in press.

Ellis, H. (1965). *The transfer of learning.* Macmillan, New York.

Estes, W. K. (1957). Of models and men. *American Psychologist*, 12:609–617.

Fiorentini, A. and Berardi, N. (1981). Perceptual learning specific for orientation and spatial frequency. *Nature*, 287:453–454.

Flannagan, M. J., Fried, L. S., and Holyoak, K. J. (1986). Distributional expectations and the induction of category structure. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 12:241–256.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.

Friedman, J. (1994). Flexible metric nearest neighbor classification. Technical report, Stanford University.

Gallistel, C. R. (1990). *The organization of learning.* MIT Press, Cambridge, MA.

Gibson, E. J. (1941). Retroactive inhibition as a function of the degree of generalization between tasks. *J. Exp. Psychol.*, 28:93–115.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* Appleton Century Crofts, New York.

Gilbert, C. D. (1994). Neuronal dynamics and perceptual learning. *Current Biology*, 4:627–629.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269.

Gluck, M. A. and Granger, R. (1993). Computational models of the neural bases of learning and memory. *Ann. Rev. Neurosci.*, 16:667–706.

Goldstone, R. L. and Barsalou, L. W. (1998). Reuniting perception and cognition: the perceptual bases of similarity and rules. *Cognition*, 65:231–262.

Gottlieb, G. L., Corcos, D. M., Jaric, S., and Agarwal, G. C. (1988). Practice improves even the simplest movements. *Exp. Brain. Research*, 73:436–440.

Guttman, N. (1963). Laws of behavior and facts of perception. In Koch, S., editor, *Psychology: a study of a science*, volume 5, pages 114–178. McGraw-Hill, New York.

Harnad, S., editor (1987). *Categorical Perception: The Groundwork of Cognition.* Cambridge University Press, New York.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, 40:185–234.

Hintzman, D. L. (1994). Twenty-five years of learning and memory: was the cognitive revolution a mistake? In Umiltá, C. and Moscovitch, M., editors, *Attention and Performance*, volume XV, chapter 16, pages 360–391. MIT Press.

Huber, P. J. (1985). Projection pursuit (with discussion). *The Annals of Statistics*, 13:435–475.

Intrator, N. (1993). Combining Exploratory Projection Pursuit and Projection Pursuit Regression. *Neural Computation*, 5:443–455.

Intrator, N. and Edelman, S. (1997). Learning low dimensional representations of visual objects with extensive use of prior knowledge. *Network*, 8:259–281.

Knierim, J. J., Kudrimoti, H. S., and McNaughton, B. L. (1995). Place cells, head direction cells, and the learning of landmark stability. *Journal of Neuroscience*, 15:1648–1659.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kornblith, H. (1985). *Naturalizing epistemology.* MIT Press, Cambridge, MA.

LaBerge, D. (1976). Perceptual learning and attention. In Estes, W. K., editor, *Handbook of learning and cognitive processes*, volume 4, pages 237–273. Erlbaum, Hillsdale, NJ.

Leen, T. K. and Kambhatla, N. (1994). Fast non-linear dimension reduction. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6, pages 152–159. Morgan Kauffman, San Francisco, CA.

Markson, L. and Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, 385:813–815.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London B*, 176:161–234.

Marr, D. (1971). Simple memory: a theory for archicortex. *Phil. Trans. Royal Soc. London*, 262:23–81.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488.

Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge, MA.

Newell, F., Chiroro, P., and Valentine, T. (1999). Recognising unfamiliar faces: The effects of distinctiveness and view. *Q. J. Exp. Psychol.* in press.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:700–708.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press, Oxford.

Omohundro, S. M. (1987). Efficient algorithms with neural network behavior. *Complex Systems*, 1:273–347.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review*, 56:132–143.

Pitts, W. and McCulloch, W. S. (1947/1965). How we know universals: the perception of auditory and visual forms. In *Embodiments of mind*, pages 46–66. MIT Press, Cambridge, MA.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Poggio, T., Edelman, S., and Fahle, M. (1992). Learning of visual modules from examples: a framework for understanding adaptive visual performance. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 56:22–30.

Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.

Polat, U. and Sagi, D. (1993). Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Research*, 33:993–997.

Polat, U. and Sagi, D. (1994). Spatial interactions in human vision: from near to far via experience dependent cascades of connections. *Proceedings of the National Academy of Science*, 91:1206–1209.

Popper, K. R. (1992). *Conjectures and refutations: the growth of scientific knowledge*. Routledge, London. 5th ed.

Quine, W. V. O. (1969). Natural kinds. In *Ontological relativity and other essays*, pages 114–138. Columbia University Press, New York, NY.

Riesenhuber, M. and Poggio, T. (1998). Just one view: Invariances in inferotemporal cell tuning. In Jordan, M. I., Kearns, M. J., and Solla, S. A., editors, *Advances in Neural Information Processing*, volume 10, pages 215–221. MIT Press.

Rumelhart, D. E. and Todd, P. M. (1993). Learning and connectionist representations. In Meyer, D. E. and Kornblum, S., editors, *Attention and Performance XIV*, pages 3–34. MIT Press.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274:1926–1928.

Sagi, D. and Tanne, D. (1994). Perceptual learning: learning to see. *Current opinion in neurobiology*, 4:195–199.

Sarle, W. S. (1994). Neural networks and statistical models. In *Proceedings of the 19th Annual SAS Users Group International Conference*, pages 1538–1550, Cary, NC. SAS Institute.

Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In *The mechanisation of thought processes*. H.M.S.O., London.

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

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8:1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of statistics*, 10:1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13:689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, 14:590–606.

Stone, J. V. (1996). Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8:1463–1492.

Uttley, A. M. (1959). The design of conditional probability computers. *Information and Control*, 2:1–24.

Valentine, T. (1991). Representation and process in face recognition. In Watt, R., editor, *Vision and visual dysfunction*, volume 14, chapter 9, pages 107–124. Macmillan, London.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag, Berlin.

Wahba, G. (1979). Convergence rates of 'thin plate' smoothing splines when the data are noisy. In Gasser, T. and Rosenblatt, M., editors, *Smoothing Techniques for Curve Estimation*, pages 233–245. Springer Verlag, Berlin.

Walk, R. D. (1978). Perceptual learning. In Carterette, E. C. and Friedman, M. P., editors, *Handbook of Perception*, volume IX, pages 257–298. Academic Press, New York, NY.

Williamson, J. R. (1997). A constructive, incremental-learning network for mixture modeling and classification. *Neural Computation*, 9:1517–1543.

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature*, 222:960–962.

Wilson, M. A. and McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, 261:1055–1058.

Winston, P., editor (1975). *The psychology of computer vision.* McGraw-Hill, New York.

Wiskott, L. (1998). Learning invariance manifolds. In Niklasson, L., Bodén, M., and Ziemke, T., editors, *Proc. Int'l Conf. on Artificial Neural Networks, ICANN'98, Skövde*, Perspectives in Neural Computing, pages 555–560. Springer.