Unsupervised Learning of Correlational Structure *

Andrew Chalnick
Computer and Information Science
University of Pennsylvania

Dorrit Billman
Department of Psychology
University of Pennsylvania

May 23, 1988

Abstract

People can learn simply from observation, without explicit feedback. Natural language acquisition is perhaps the most spectacular example, but unsupervised learning occurs in many domains. We present 1) a task analysis of a broad class of unsupervised learning problems and 2) an initial simulation based on the task analysis which successfully learns all the rule types identified in the analysis. Our task analysis characterizes systems of interpredictive correlational rules which could be the basis for category formation in unsupervised learning. For example, observation of various animals could lead to abstracting covariation rules among wings, feathers, and flight, and also among fins, scales, and swimming. These rules in turn could form the basis for the categories *bird* and *fish*. Our analysis identifies three types of predictive features and three types of rules which may be available in input: *universal*, *contrastive*, and *exception-based rules*. This analysis guided design of our learning procedures. Our simulation succeeds in learning all three rule types. This is difficult because procedures which facilitate learning one rule type may inhibit learning another. Further, our simulation is restricted in psychologically motivated ways and succeeds despite these requirements. We know of no other simulation or modeling project which addresses exactly this class of learning problems. Our results demonstrate the existence of successful procedures. However, we believe our most valuable contributions are our task analysis and framework for testing the power and limits of domain-general learning procedures applied to unsupervised learning problems.

Introduction

What procedures allow learners to discover categories from observation of instances? When contrastive feedback is provided people (and computers) can use this to guide learning. When discriminative feedback about category membership is not provided, a greater burden is placed on the learner to discover structure in input and set up sensible categories. Our project investigates what learning procedures are sufficient for successful learning under these conditions. One goal of the project is to develop a successful learning program; the more fundamental goal is to develop a model of the learning task which will allow us to investigate the effects of component learning procedures on different aspects of the learning problem.

Our core principle for explaining observational learning is internally generated feedback: learners compare predicted and observed properties and use the match or mismatch to guide learning. When input provides interpredictive relations among feature values, this structure can be discovered using internal feedback. Prior work began exploration of learning under these constraints [Billm87b]. This paper presents a more sophisticated model.

Systems of correlations are central to learning categories from observation. Furthermore, representation of correlational structure is intimately linked to category use: The primary purpose of categories is to organize knowledge to allow sensible inferences. Given enough information to classify something as a bird, this licenses additional inferences: for example, it hatched from an egg and will lay eggs if female. If we observe a new property for some particular bird, say, eating worms, we may generalize that property to other members of the same category. While these inferences are certainly not correct all the time, they provide us with a valuable way of extending our knowledge. Rosch [Rosch78] has pointed to the importance of correlations in defining category structure. She argued, however, that correlations are important in leading a cultural group to discover or rely on a category, but not that individuals use this correlational structure in category learning. We are interested in how an individual might use discov-

*We thank Richard Billington and Lyle Ungar for valuable input and discussion about this research.

ery of correlations as the basis for category learning. Prior work has found that individuals do use correlational structure in learning categories [Garne74, Billm87b, Billm87c].

In unsupervised learning, correlational structure is even more important than when categories are designated for the learner and explicit feedback about membership is provided. In learning to distinguish categories A and B with explicit feedback, finding any attribute which predicts category membership is sufficient. The learner has no need to notice whether various predictor features covary with one another. When no feedback is specified externally, concept learning will be largely driven by discovery of interpredictive relations among feature values. That is, subject's may discover coherent patterns in the observed examples and use this to set up categories.

We can understand the learning problem better by analyzing the structure which is potentially available in input. Specifically, we can identify three types of predictive rules and three classes of features. The three types of predictive rules are universal, contrastive, and exception-based. *Universal* rules apply to all instances in the domain; if the domain of learning is animals, then the system should discover that all animals eat and breathe. *Contrastive* rules could be used to divide the domain into major, contrasting classes. In learning about a wide range of animals, we want the system to discover that fins, scales, and swimming all go together, as do wings, feathers, and flying. Discovery of these rules was the initial goal in designing the system. Clusters of such rules can then form the basis for contrasting categories such as fish and birds. We would like the system to learn these regularities even when exceptions occur. Finally, *exception-based* rules represent information about the exceptions themselves; if an animal has wings, but looks like it is dressed in a tuxedo, it will not fly. The system should be able to learn about bats, whales, and penguins.

Procedures which facilitate discovery of one class of these rules may often inhibit learning rules of another class. Learning universal rules requires good sensitivity to very general patterns. Learning rules about exceptions requires good sensitivity to quite specific patterns. Learning rules which would form the basis of maximally coherent categories requires good sensitivity to features which are related to many other features. Thus designing a system capable of learning all three classes requires accommodating conflicting demands. This problem has not been highlighted in prior work, because the issue is much less important in learning with feedback. Specifically, there is no pressure to learn universal properties when learning categories from contrastive feedback. However, when a system is designed to seek good predictions, it must be prevented from focusing too much on vacuous predictions of universal properties.

Parallel to the three types of rules are three types of features. *Universal* features do not vary across the entire set of objects in the problem domain. *Contrasting* features are those which both vary across the domain and covary with other features. These are the features which we intuitively think of a defining separate classes. Features such as mode of locomotion, type of body covering, type of limb, and distinctive location or habitat divide up major classes of animals. *Idiosyncratic* features are highly variable across the domain, but do not generally covary with other features. Their predictive value is primarily in conjunction with several other features, to identify an individual or exception. Learning about penguins may require attention to distinctive coloring, even though color may not be widely predictive. Our learning model uses this analysis of universal, contrasting, and idiosyncratic features in the procedures which control learning.

Problem Definition

Our modeling begins with three fundamental assumptions about learning correlational patterns. First, people and other cognitive systems do learn about patterns of feature correlations, even in unsupervised learning. Second, this information is represented directly and locally, as in classifier [Holla75, Holla86] and production [AKB79] systems, not indirectly in a set of weights distributed across a system [Rumel86]. Direct representation of rules allow other mental procedures, as in inference and transfer, to selectively operate on representations of different regularities. Third, all three types of rules, universal, contrasting, and exception-based are important components of learning about the correlational structure in input. In addition, we place several constraints on the available information and resources.

1. The information available to the learner is limited. No feedback is provided and learning takes place from unsupervised observation of examples. Much natural learning is informal and untutored. Feedback is often sporadic,

unreliable, or unavailable. By modeling learning with no designated feedback we investigate the most difficult case; models for learning with feedback can be set up as an easier, special case [Billm87c]. Most researchers have addressed learning with feedback. Whether the learning criteria is predicting category membership [AKB79] or earning as much of a target resource as possible [Holla86], it is directly specified for the learner. In our task, the learner must discover which features are predictable as well as which features are predictive.

2. Memory, either storage or retrieval, is limited. Some specific information may be preserved. However, we do not allow learning procedures which operate over the set of previously seen objects. Rather, an observation affects the learner's state of knowledge but no representation of the object as an individual need be retained.

3. The learner's initial knowledge is limited. Learning should not depend heavily on the initial state of the learner. First, the learning procedure should be sufficiently robust so that the order of presentation of examples does not profoundly change the course of learning. Second, learning should not depend on extensive prior knowledge of the domain. We are interested in specifying general learning procedures which can apply even where the learner lacks much knowledge initially.

4. In general, we want our learning procedures to apply homogeneously across rules without reference to rule content. This approach contrasts with those where the strength of the learning procedures depends on the content of old knowledge.

There are undoubtedly many circumstances where the conditions of learning are not so austere. We are interested in investigating this difficult class of learning problems because we believe that these circumstances will tell us most about the strengths and limits of general, data-driven learning procedures. We investigate learning of systems of structured representations, given minimal initial knowledge and minimal information in input. Our simulation operates under these constraints.

Representation

The representation format is described here. This, together with the procedure description to follow, specifies all the apriori information built into the system. The learning model is independent of the particular domain. However, for a learner to succeed, input must provide at least some contrastive or universal features and the learner must represent these. For this implementation, the domain consists of birds, fish, mammals, whales, penguins and bats. Input objects are represented as vectors of feature values, where all features must be specified. Thus a typical object might be a red, furry land animal with legs which weighs 200 lbs., and eats and breathes.

Representation of Features

Both observed objects and internal rules are represented in terms of the same set of features. This feature set is not currently altered by learning. The simulation runs reported here used seven features: breathes (t or f), eats (t or f), color (blue, yellow, red, white, black, green, tuxedo, brown, pink), weight (coded into 7 distinct ranges), locomotion mechanism (legs, wings, fins), habitat (land, air, water), and body covering (hair, feathers, scales). Each feature has two associated parameters, salience and variability. Feature *salience* is a function of predictive success across all the rules in which the feature participates as a predictor. The feature *variability* is estimated from storing the set of recently observed values of that feature. This requires maintaining minimal information about the distributions of feature values. Contrastive features are variable and salient. Universal features are homogeneous. Idiosyncratic features are variable and have low salience. This information about features is used in the learning procedure.

Representation of Rules

Knowledge of regularities is represented in conditional rules. Each conditional rule consists of a condition and an implication. The condition specifies the values of a proper subset of the observable features. The implication specifies the value of one predicted feature. A conditional rule might specify that, if something has scales and fins, then it travels around in the water. Each conditional rule has an associated strength estimate. Rule strength is a function of the rule's

predictive validity, the salience of the features in its condition, and the variablity of its predicted feature. The present work differs from an earlier project [Billm87] in that we now allow multiple features in the condition. This is a fundamental change. First, when only single feature conditions are specified, it is feasible to enumerate all the representable rules. Then the learning procedure need only select good rules from among an initially instantiated set. When multiple features are allowed, the combinatorial explosion means that the learner must not only decide which of a number of rules is the best, but it must also generate these rules. Second, it changes the representational power of the system. Learning higher order regularities, subpatterns, and exceptions requires use of multiple features.

Procedure

The learning procedure consists of major and minor cycles. In the minor cycle, objects are presented to the learner and tested. The major cycle removes weak rules and replaces them with new, potentially stronger rules. The rules compete with one another, with their level of success based on their ability to explain the domain. The rules which provide the best model will survive while the others will be removed.

The Minor Cycle

The learner first picks a random object to examine, samples a set of features from the object, and then picks a rule to test given the sampled feature set. *Focused sampling* alters selection probability of the features sampled in observing an object. It directs attention to features which have proved predictive in the past. When a feature is sampled, its variability is updated.

With multiple features in the condition, it is unlikely that a particular set of feature values will find an exact match in the condition of some rule. Thus, we need a partial match value. The value is a function of 1) the number of conflicts between the rule and the set of sampled features, 2) the number of matches, and 3) the variability of the action feature of the rule (this helps avoid making vacuous predictions). A rule is then picked probabilistically as a function of its match score.

After a single rule is picked it can either be tested or generalized. A rule is generalized if some part of the rule's condition conflicts with some feature values of the object. To generalize, we remove the conflicts and add the generalized rule back into the rule base. The generalized rule is now free to compete with the parent rule. Only a limited expansion of the rule base is allowed in one major cycle. If a generalized rule is formed, its prediction is tested; if not, the orginal rule is tested. If the prediction is correct, the rule's predictive validity and the salience of the features in the condition are increased. If unsuccessful, they are decreased. The modification is done in accord with a *delta* rule. The delta rule revises a parameter by moving its current value a certain percentage closer to minimum or maximum value, depending on whether the rule or feature is being increased or decreased.

The Major Cycle

A major cycle follows a block of many minor cycles. Weak rules are dropped from the rule base and new ones inserted for testing. New rules are generated by sampling random objects and creating rules with conditions containing values for all but one of the features with the remaining feature in the implication part of the rule. Thus the new rules are maximally specific. The size of the rule base is expanded by the number of strong rules found in the prior block of learning cycles. This expansion allows new rules to compete more successfully. Finally, each rule's predictive validity is decremented a fixed percentage. This tax helps weaken and eliminate irrelevant rules.

Simulation Evaluation

The primary goal for the initial development phase of the system was a sufficiency demonstration for a test problem. We wanted to find a set of procedures and parameter values within our specified constraints which succeeded in learning a substantial amount of the structure available in the input. Given this reference point, we could then use the system to explore the effects of varying the learning parameters. This report summarizes our success in constructing a system which meets this initial design goal. Our primary success criteria is the number of target rules the system has learned over a given time period. In addition, we are interested in attentional learning, that is, discovery of the predictive features.

Descriptive Analysis

The first method of evaluating rule learning is examination of the set of rules discovered after a sig-

nificant learning exposure. We can ask how many universal, contrastive, and exception-based rules were discovered and what the number and nature of other, non-target rules were. Below, we describe results from one simulation run. It had 40% exceptional objects in the object base and a minor cycle of 400 iterations. We report the strong rules found by the system during learning, from among the 20 rules stored.

After **20 cycles**, the system had already learned rules about birds, fish and whales. It had also learned that all things breathe. Of the nine rules then, one is universal, five are contrastive, and three are exceptional. There are no overly specific rules in the top set.

(Feathers) ⇒ Wings
(Water) ⇒ Fins
(Scales Water) ⇒ Fins
Fins) ⇒ Water
(Fur Fins Water 1000) ⇒ White
(White Fur Water 1000) ⇒ Fins
(Feathers Wings) ⇒ Air
(White Fur Fins Water) ⇒ 1000
Nil ⇒ Breathes

After **40 major cycles**, the system learned about penguins. While the system had previously learned that feathers and wings imply air, it now knows that if the animal is dressed in what appears to be a tuxedo, we can expect to find it on land. Also notice that there is now an extraneous piece of information — (blue color) — in the second new rule. The frequency of blue fish hasn't changed, but the salience of color has increased, giving strength to overly specific rules such as this one.

(Tuxedo Feathers Wings) ⇒ Land
(Scales Water Blue) ⇒ Fins

By **50 cycles**, the system had learned about relations among all the contrastive features, had predictive rules for each type of exception, and knew the universal property of breathing. Though the system had not learned all possible predicitively valid rules, it had learned universal, contrastive, and exception-based rules.

(Legs Land) ⇒ Fur
(Brown Wings Land) ⇒ Fur

Two new rules were added at 100 cycles.

(Wings Air) ⇒ Feathers
(Fins) ⇒ Water

Good mastery of exceptions and contrasting classes was widespread across runs with different levels of exceptions and various parameters. Within the present system configuration, universal rules were only learned with 40% exceptions in input. Exceptions are important for successful generalization since our conservative generalization mechanism is driven by dropping conflicting features, and conflicts are only found in exceptions. This conservatism also means that incorrect generalizations are never found.

Performance Statistics

During the course of each learning experiment we collected data on how many of the contrastive and universal rules the system has learned. We do not include exception-based rules here because the set of possible exception-based rules is so large. There are a total of 27 contrastive and 2 universal rules. The 27 contrastive rules include all combinations of predictive relations among body covering, habitat, and limb type. These form an interlocking and redundant set of predictions, for example, predictions that wings and feathers imply flying; wings imply flying; and feathers implies flying. Thus, percent of rules learned provides a quite conservative measure; with respect to the training domain, predictive success could be perfect with 9 of the 27 rules.

Figure 1 shows the results of this data collection from runs with 0%, 20%, 30%, and 40% exceptions. Each line here is the average of two runs with identical parameters but different random factors in selection of objects to observe and rules to test. Given that there are exceptions at all, learning is slowed with a higher proportion of exceptions. However, a different pattern holds in the runs with no exceptions. In these runs, the system can only learn a third of the rules, and performance quickly moves to this level. Our generalization mechanism is very conservative. Generalizations are only introduced when a more specific rule is wrong. When there are no exceptions, rules about a pair of features, such as wings and feathers implying air, are never generalized further; wings implying air is never produced. Thus, without exceptions, the system quickly learns each of these nine predictive contrasting rules. This is sufficient to correctly predict all the contrastive regularities

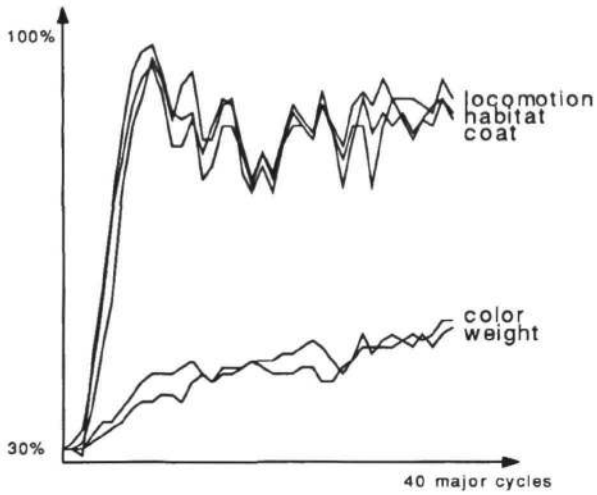

Figure 2: 20% Exceptions

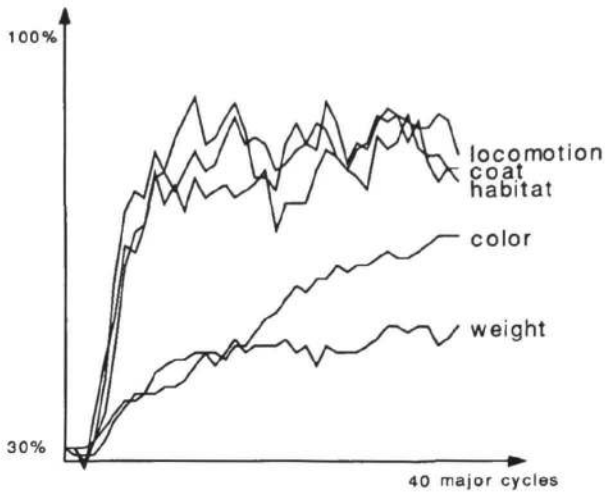

Figure 3: 30% Exceptions

found in the learning set and is another indicator of the conservative nature of our scoring rule as percent of all contrastive rules.

Attentional learning

In addition to learning about predictive rules, the system also learns what features are predictive. This information is an important part of making knowledge about the structure of the observed objects explicit. It is important information to use in explaining transfer (negative or positive) in learning new problems and may also facilitate learning within one problem. Figure 2 and Figure 3 show the salience of features in the 20% and 30% exception conditions whose rule learning curves appear in Figure 1. The constant (low) salience of the two universal features is not shown. Attentional learning is fast and produces sharp separation of the contrastive and idiosyncratic features with 0% (not shown) and with 20% exceptions. The system quickly learns which features are the best predictors. As the exception level increases, the contrastive rules are less and less reliable and the exceptional rules become the best predictors. Since color is a distinctive feature for each exception and is required (in combination with other features) for predictive success, color salience rises over the course of learning. With 40% exceptions (not shown) color ranks as the best predictor after about 40 learning cycles.

Summary

We summarize first the most important weaknesses and then the strengths of the current project. Our system requires that all objects be representable in the same set of features. However, in many learning problems the features relevant to objects in one subdomain are not relevant in another. Hence, the system cannot learn domains where features, not just feature values, vary. More fundamentally, there are limits to a representation based solely on conjunctions of observable feature values; while sets of conditional rules using such feature vectors are a fairly powerful form of representation, we do not believe it is sufficient to account for induction and inference using category knowledge. We need to add explicit representations of categories, not just the corresponding predictive rules. Our current work includes some procedures for gathering the information collected in sets of contrastive rules and building a semantic network of explicitly represented categories. These procedures for adding categories are very ad hoc and will require much more rigorous development. The most serious flaw in the present system in achieving its initial goals is our generalization mechanism. The system relies heavily on exceptions for learning general rules — this results in overly conservative learning. Stronger generalization procedures will be particularly important in learning a system of hierarchical categories with more than one level of contrastive categories.

One major strength of the current project is the problem analysis: systems of predictive rules and categories can be learned from observation of ex-

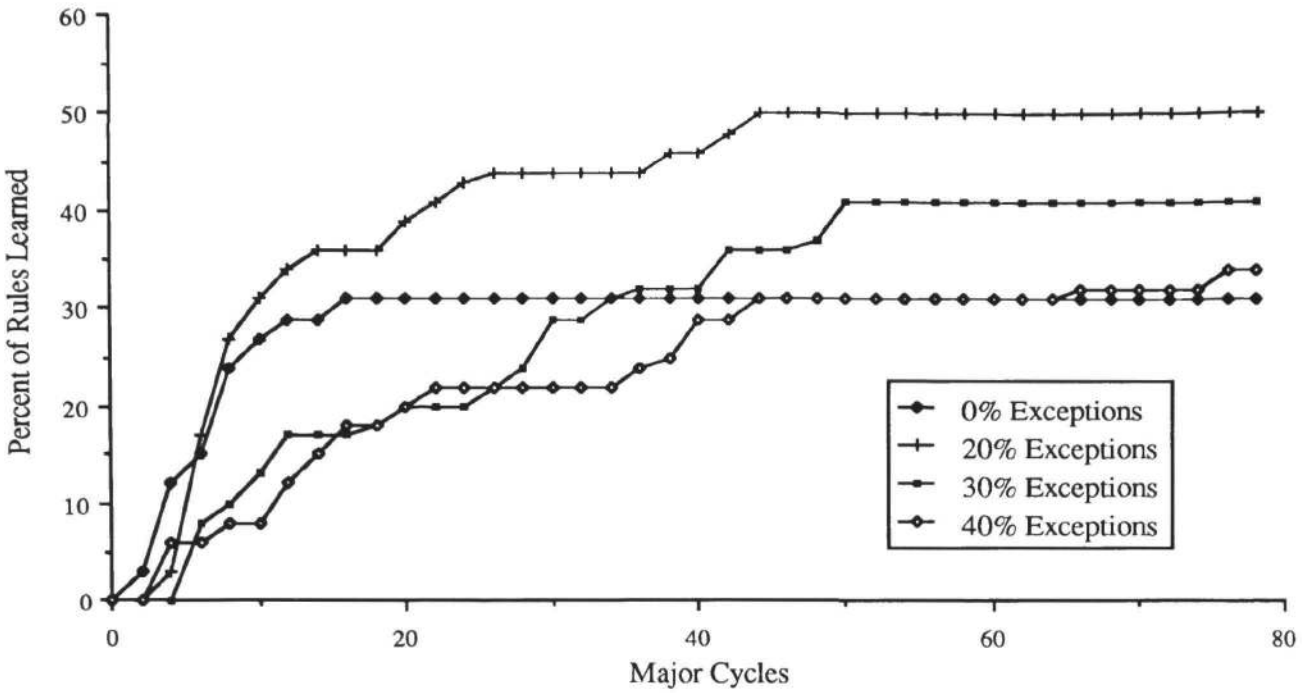

Figure 1: Rule Learning Curves with Focused Sampling

amples by comparing predicted and expected feature values. We identified three types of predictive rules and three possible ways that features can covary. We use our classification of features into universal, contrastive, and idiosyncratic categories to guide the design of our learning procedure. Specifically, we use feature salience and variability as well as predictive success and specificity to guide learning. Second, in applying this task analysis in the current simulations, we have a learning system which meets the design criteria specified initially and which successfully learns contrastive, exception-based, and universal rules. Finally, our simulation provides a flexible tool for further research. It allows us to modify components of the learning system and test their effects on different aspects of learning.

References

[AKB79] Anderson, J.R., Kline, P.J., & Beasely, C.M. (1979). *A general learning theory and its application to schema abstraction.* In G.H. Bower (Ed.), The Psychology of Learning and Motivation, Vol.13. New York: Academic Press.

[Billm87] Billman, D.O., Richards, J., & Heit, E. (1987). Abstraction of correlational rules in implicit concept learning tasks. in review.

[Billm87b] Billman, D.O., Heit, E., & Dorfman, J. (1987). *Facilitation from clustered features: Using correlations in observational learning.* In The Ninth Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

[Billm87c] Billman, D.O. & Heit, E. (1987). Observational Learning From Internal Feedback: A simulation of an adaptive learning method. Cognitive Science. In press.

[Garne74] Garner, W.R. (1974). *The Processing of information and structure.* Hillsdale, NJ: Erlbaum.

[Holla75] Holland, J.H. (1975) *Adaptation in Natural and Artificial Systems.* Ann Arbor: The University of Michigan Press.

[Holla86] Holland, J.H. (1986). *Escaping brittleness: The possibilities of general-purpose learning algorithms applied to parallel rule-based systems.* In R.S. Michalski, J.G. Carbonell, & T.M.Mitchell (Ed.), Machine Learning. Palo Alto: Tioga Press.

[Rosch78] Rosch, E.H. (1978). *Principles of categorization.* In E.H. Rosch & B.B.Lloyd (Eds.), Cognition and Categorization Hillsdale, N.J.: Erlbaum Publishers.

[Rumel86] Rumelhart, D. & McClelland, J. (1986). Parallel distributed Processing. Cambridge, MA: Bradford Books/MIT Press.