SCIENCE AS AN ANOMALY-DRIVEN ENTERPRISE: A COMPUTATIONAL APPROACH TO GENERATING ACCEPTABLE THEORY REVISIONS IN THE FACE OF ANOMALOUS DATA

by

Will Bridewell

M.S. in Computer Science,

University of Pittsburgh, 2002

B.S. in Computer Science

B.S. in Mathematics

B.S. in Psychology,

Northern Kentucky University, 1998

Submitted to the Graduate Faculty of

Faculty of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2004

UNIVERSITY OF PITTSBURGH

FACULTY OF ARTS AND SCIENCES

This dissertation was presented

by

Will Bridewell

It was defended on

September 17, 2004

and approved by

Bruce G. Buchanan, Ph. D., Professor Emeritus

Milos Hauskrecht, Ph. D., Assistant Professor

Greg F. Cooper, Ph. D., M. D., Associate Professor

Kevin D. Ashley, Ph. D., J. D., Professor

Dissertation Advisors: Bruce G. Buchanan, Ph. D., Professor Emeritus,

Milos Hauskrecht, Ph. D., Assistant Professor

SCIENCE AS AN ANOMALY-DRIVEN ENTERPRISE: A COMPUTATIONAL APPROACH TO GENERATING ACCEPTABLE THEORY REVISIONS IN THE FACE OF ANOMALOUS DATA

Will Bridewell, PhD

University of Pittsburgh, 2004

Anomalous data lead to scientific discoveries. Although machine learning systems can be forced to resolve anomalous data, these systems use general learning algorithms to do so. To determine whether anomaly-driven approaches to discovery produce more accurate models than the standard approaches, we built a program called Kalpana. We also used Kalpana to explore means for identifying those anomaly resolutions that are acceptable to domain experts. Our experiments indicated that anomaly-driven approaches can lead to a richer set of model revisions than standard methods. Additionally we identified semantic and syntactic measures that are significantly correlated with the acceptability of model revisions. These results suggest that by interpreting data within the context of a model and by interpreting model revisions within the context of domain knowledge, discovery systems can more readily suggest accurate and acceptable anomaly resolutions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

When I was an undergraduate studying psychology, I took the required course on experimental methods. In this course, I had the chance to perform my very first scientific experiment. I gathered a group of volunteer students and had them perform a task and fill out a test. Upon examining the results, I noticed a nice linear trend relating task performance to test score. In fact, all the points were clustered neatly together—except one. I spent days staring at the data, trying to understand that outlier, that anomaly to the trend. Mostly, I was fascinated in how poor our understanding of the process of science actually is. How could there be no definitive method for analyzing and explaining, or explaining away, anomalous results? I have no idea who that student was, but that person likely shares responsibility for the topic of this thesis.

I would like to say that scientific research is inherently collaborative, so although this work has, in a sense, been my personal vision quest, I was aided by kind spirits along the way. I would like to thank John Dowling and Wendy Chapman for their time, expertise, and data. Their generous donation of resources made my work much easier. I would also like to thank Bruce Buchanan, who kindly nudged me away from less interesting topics and gave me guidance when needed. My committee, composed of Bruce Buchanan, Milos Hauskrecht, Kevin Ashley, and Greg Cooper, also deserves my gratitude for keeping me on track and giving me support, encouragement, and insight along the way. Additionally, Kurt Van Lehn and Micki Chi provided invaluable support and advice.

In addition to those individuals who guided my research, I should thank those individuals who provided social support along the way. If it were not for Kathy O'Connor's initial act of hospitality, I am sure that I would be somewhere less appealing. All of my friends deserve my thanks, those in academia for listening to my rantings and ravings while smiling politely

and those beyond the pale for keeping me grounded. And, lest I forget, I should extend thanks to the staff of the Department of Computer Science who helped me maintain my sanity underneath a massive mound of bureaucracy.

Finally, I thank my family. They know why.

1.0 INTRODUCTION

Anomalies drive scientific discovery. In *The Structure of Scientific Revolutions*, Thomas Kuhn wrote, "Discovery commences with the awareness of anomaly ... [and] it closes only when the paradigm has been adjusted so that the anomalous has become the expected [31]." That is, when faced with an inexplicable event, scientists conclude that their theory is either incomplete or incorrect. Then, by concentrating their attention on the anomalous case, the scientists may discover alterations to their current theory or a new theory to replace the old. For example, in an earlier work, Kuhn tells how the retrograde motion of Mars and related anomalies led to a series of major upheavals in Renaissance cosmology [30]. We seek to create a system that, like Kepler, Copernicus, and others, exploits the discovery potential of anomalies.

Unlike Kuhn, we engage in computational system building, therefore we must clarify and limit our definition of "anomaly." While the term can apply to any unexpected event, we restrict its usage to either single observations or a collection of similar observations that contradict a specific theory. For example, when Galileo observed mountains on the moon, he presented believers in the prevailing theory of celestial bodies with an anomaly. How can a perfect, crystalline sphere have mountains and ridges? In addition to our specific usage, the general use of the term "anomaly" refers to unexpected trends in a data set, observations at the edge of acceptability, and so on. Although these broader extensions of the term identify future directions of inquiry, we always use "anomaly" in its restricted sense. As we shall see, this restriction brings with it specific advantages.

Anomaly-driven discovery, including anomaly-driven theory revision, yields benefits not exploited by general inductive systems (e.g. RL [46], C4.5 [48]). Foremost, the anomaly pinpoints the error site in the theory. After all, how can we identify an anomalous observation

if we lack specific expectations? Thus the problem reduces from improving the theory as a whole to addressing a deviation from a particular belief. That is, anomaly-driven discovery refines questions such as, "Is the prevailing theory of cosmology correct?" into those like, "Why does Mars appear to move backwards in the sky at particular intervals?"

As a result of error localization, anomaly-driven discovery produces additional advantages. First, the contested section of the theory lends context to both the observation and possible remedies. Thus, knowing that, at times, Mars appears to move backwards in the sky, a scientist can relate this behavior to the planet's "normal" trajectory as well as the retrograde motion of other planets. As a second advantage, the anomaly constrains the search for new knowledge. That is, the revised theory must account for all previously explained observations in addition to the anomalous case. So, the anomaly specifies where to look, what information to consider, and to a small degree what form the discovery must ultimately take.

As an example of anomaly-driven discovery, consider the identification of the louse's role in the transmission of typhus. Before Charles Nicolle's work, typhus was assumed to be transmitted as most other contagious diseases (i.e., through direct physical contact). However, Nicolle noticed that in a crowded hospital only certain individuals tended to contract the disease. More importantly he found that "typhus patients continued to spread infection up to the point when they entered the hospital waiting-room . . ., [but] they became completely inoffensive as soon as they had been bathed and dressed in the hospital uniform. [26]" This geographically based transmission, where patients were contagious in the waiting room but not contagious once admitted, was anomalous to the general theory of disease transmission.

The discovery of the mechanism of transmission for typhus illustrates the benefits of an anomaly-driven approach. By identifying an anomalous condition, Nicolle could localize the fault in the theory. That is, the dominant theory of medicine held that most diseases spread through individual contact. Nicolle observed that this was not the case with typhus, thereby localizing the problem to one of disease transmission. Next, both context and constraints could be identified. The context consisted of the need for a new mode of transmission, the conditions of the physical locations where typhus spread, and the conditions where it did not. The constraint was that the mode of transmission had to explain how patients infected with typhus could suddenly stop being contagious without ignoring how highly contagious the

disease can be. Nicolle's discovery, that typhus was being transmitted by lice and that these lice were removed after thorough cleaning, repaired the theory, satisfied the constraints, and made use of the anomaly's context.

1.1 THE CRITERIA FOR ACCEPTABLE ANOMALY RESOLUTION

Although anomalies identify weaknesses within a theory, not all revisions that they engender are acceptable. For instance, a revision should repair the fault in the theory, otherwise the anomaly is not resolved. Additionally, the repair should not cause other faults to appear. And finally, the repair should be justifiable within the context of the domain. This characteristic helps prevent cluttering a theory with poorly supported special cases. We claim that for a revision to be acceptable it must meet these three criteria: rehabilitation, monotonicity, and defensibility.

1.1.1 Rehabilitation

An anomaly resolution must remove the original contradiction. Whether the removal results from an alteration of the theory or the discarding of data, this condition must be met. For example, consider Galileo's observation of mountains on the supposedly spherical moon. The most ready resolution of this anomaly presents itself as a claim of methodological error. Using this approach, Galileo's opponents cast doubts upon the reliability of the telescope. Here the data are rejected, and the contradiction disappears. Ludovico delle Colombe chose another tactic. By conjecturing that the mountains existed beneath a crystalline sphere, he both saved the original theory and accounted for the data. Again, the contradiction disappears—regardless of the veracity of the resolution. Finally, Galileo hypothesized that the spherical nature of the moon was a myth. His claim also resolves the contradiction differing from the others in its general acceptance by the scientific community.

1.1.2 Monotonicity

In addition to removing the contradiction, acceptable anomaly resolutions must not degrade the theory. That is, new anomalies should not arise from old observations once the resolution is applied. This condition provides a monotonicity constraint. The assumption behind this criterion holds that the current theory is mostly correct and that the anomaly indicates an error far from the base axioms. Occasionally this assumption may be revealed as erroneous, as with the shift from a geocentric to heliocentric model of the solar system, but even these exceptional instances must account for prior observations.[1]

1.1.3 Defensibility

Finally, an acceptable anomaly resolution must also be defensible. A revision that cannot be justified within the context of the current domain must be rejected. For example, suppose that Nicolle conjectured that the odd pattern of typhus infection resulted from a natural immunity in those workers inside the hospital. Such an explanation rehabilitates the contradiction without producing new anomalies. However, support for this revision is circumstantial. Unless the revision adequately explains why there should be such an odd distribution of workers within the environment, it is unacceptable. In general, wide acceptance of a belief should be contingent on the presence of a thorough supporting rationale. Without such a criterion, we risk reducing our theories to statements of apparent associations and *ad hoc* excuses.

1.2 HYPOTHESIS

We claim that anomaly-driven discovery will lead to acceptable revisions not considered by traditional learning approaches. That is, emphasizing the anomaly, along with its relationship to the data and the original theory, will lead to a larger number of revisions that are

[1] Sometimes there are exceptions even to this rule where the new theory reduces the scope of the domain. In such cases, the new theory need not explain those observations now outside of its range. Nevertheless, those old observations remaining within the theory's domain must not now be construed as anomalous.

rehabilitative, monotonic, and defensible compared to approaches that ignore the context of the anomaly. Additionally, we claim that heuristics can be devised that identify defensible revisions and that these heuristics must employ background knowledge instead of, or in addition to, syntactic analysis of the revisions. In particular, we believe that relatively shallow knowledge of the domain will lead to substantial improvement in the ability to assess the acceptability of a given revision. Finally, we claim that acceptability serves as a more refined measure of a theory's generalizability than predictive accuracy. More specifically, the intersection of accurate and acceptable anomaly resolutions leads to the best revised theory.

We will assess our claims for the acceptability of a revision using the judgment of a domain expert. In the ideal case, a system employing our heuristics, which we will later clarify, will produce only acceptable revisions. Therefore the domain expert can choose from a collection of repairs without sifting through those that are either implausible or irrelevant. The expert's success in this task requires the ability not only to rule out unacceptable revisions but also to determine whether an acceptable revision applies to a meaningful class of anomalies or serves only to explain away a singularity. The distinction made within this latter task is beyond the scope of this work.

1.3 METHODS OF RESOLVING ANOMALIES

Although anomalies can point toward new knowledge, they do not always lead to discoveries. Work in cognitive science has identified eight responses to anomalous data [6, 9]: (1) ignoring the data, (2) holding the data in abeyance, (3) maintaining uncertainty about the data, (4) excluding the data from the theory's scope, (5) rejecting the data, (6) reinterpreting the data, (7) altering the periphery of the theory, and (8) replacing the theory. Of these responses, only (7) and (8) lead to alterations in the theory, while responses 1–4 are relatively uninteresting as they do not involve any form of explanation. Of those that do require an explanation, the explanation provided when rejecting the data requires the least detail. Here the scientist claims that the data were produced from error, random effect, or fraud and therefore need not be addressed. At the other extreme of explanatory detail, the

scientist alters a core belief upon which the theory rests. Between these two responses lie the reinterpretation of data and peripheral alteration of the theory.

When a scientist reinterprets an anomaly, he accepts its validity, but provides an explanation that does not alter the relevant theory. As an example, consider the Allais effect [3], which has been reproduced with varying success. During a solar eclipse, the oscillation of a Foucault pendulum deviates slightly from its expected trajectory. One explanation for these deviations suggests a fundamental flaw in the theory of gravitation. However, other proposed explanations reinterpret the data so that the anomalous observations no longer contradict the theory [15]. For example, cooling in the upper atmosphere may result in enough increased mass to account for the deviations. This explanation reinterprets the data as the product of a plausible effect that can account for the measured change in gravity. However, if further observations fail to support the suggested reinterpretations, the theory itself may be altered or even replaced so that the anomaly will be resolved.

For an illustration of the difference between altering the periphery of a theory and changing core beliefs, consider astronomy at the time of Kepler. Prior to Kepler's changes, the predominant astronomical theories, either the Ptolemaic or the Copernican, employed circular orbits. To account for any discrepancy between the predicted orbits of the planets and the recorded data, scientists introduced epicycles—in abundance. That is, the typical approach to resolving anomalies involved the addition of minor circular orbits that were centered on larger orbits. These alterations were based on the assumption that planetary motion is circular.[2] Within this context, all inconsistencies had to be accounted for by introducing new circles into the system. The addition of these new circles exemplifies the concept of peripheral theory change. In contrast, Kepler's actions demonstrate the changing core of beliefs—what Kuhn calls a paradigm shift. By the time that he developed a satisfactory theory of his own, Kepler had dispensed with circular orbits and epicycles altogether, placing the planets on elliptical orbits. His new theory shared components from prior astronomical conceptualizations, but he excised a fundamental assumption to create a more accurate and parsimonious explanation.

[2]This assumption was fundamental both intellectually and, perhaps more importantly at the time, religiously.

Darden's work goes beyond the list of eight responses as she describes actual strategies for anomaly resolution [11, 12]. She groups rejection and reinterpretation of the data together, describing them as methods of "monster-barring," meaning that when an anomaly can be explained away, it no longer threatens the theory. However, Darden gives no indication of how such monster anomalies can be differentiated, prior to successful explanation, from what she calls model anomalies. Unlike monster anomalies, model anomalies effect change in the original theory. That is, the explanation of model anomalies requires either an alteration, addition, or removal of some theoretical component, which Darden defines as a part of the theory that changes over time (e.g., an equation, a condition). These approaches either generalize or specialize the theory.[3]

Generalizing the theory involves adding a new theoretical component or generalizing a current component. In either case the scope of the theory, which includes all cases that match the antecedent of at least one rule, expands. For example, consider a rule-based model that predicts whether an individual has an infectious disease of the lower respiratory system (RS). Let this model contain the rule, "IF wheezing is present and tachycardia is present, THEN RS is present." This component can be generalized by removing one of the conjuncts in the antecedent. For instance, the generalization, "IF wheezing is present THEN RS is present," reduces the number of restrictions to the applicability of the rule. Alternatively, adding the rule, "IF cough is present, THEN RS is present," expands the model to cover all cases when an individual has a cough.

In contrast, deleting or specializing a theoretical component specializes the theory, thus restricting its scope. To illustrate, consider the model containing the rules, "IF wheezing is present, THEN RS is present," and, "IF cough is present, THEN RS is present." Removal of the latter rule restricts the theory by preventing it from classifying cases that it previously could (i.e., when cough is present and wheezing is absent). Similarly, adding the conjunct, "tachycardia is present" to the antecedent of the former rule also restricts the model. In particular, patients who are wheezing, but do not exhibit tachycardia may no longer be classified. Thus both alterations reduce the model's scope.

[3] In [11], Darden describes other methods of altering components such as tweaking parameters and proposing the opposite of the failing component. We have limited our discussion to the four common approaches that are relevant to the current work.

The version-space approach to inductive concept learning [41] accessibly characterizes the use of some of these strategies. In this representation, the learner stores two sets of hypotheses that describe a target concept. For example, let the target concept be the presence of RS. One hypothesis could state, "IF cough is present, THEN RS is present." The sets of hypotheses, G and S, respectively consist of the most general and the most specific descriptions consistent with all observed data. Therefore, if a new instance matches all members of S, then the learner classifies it as a positive instance of the target concept, whereas if the instance fails to match any members of G, then the case is classified as a negative instance.

To update its concept description, the version-space learner applies anomaly-resolution strategies when given a new training example. Returning to the RS domain, where all the concept descriptions indicate an infection in the lower respiratory system, let G contain the description, "cough is present," and S contain, "dyspnea is present and cough is present." When it incorrectly classifies an example, the learner will react in one of two ways. A positive example causes the learner to specialize G by removing those descriptions that fail to match the new instance. Thus, if the observed patient lacks a cough, the description, "cough is present," will be removed from G. At the same time, S will be updated by generalizing all hypotheses that exclude the case. So, the example description becomes, "dyspnea is present." Negative examples alter the boundaries of G and S in similar ways. Using the original description of the concept, suppose now that the patient does not have respiratory syndrome, but does exhibit both cough and dyspnea. Now the description in S, "dyspnea is present and cough is present," will be excised from the set so that no member of S matches a negative example. In addition, the system specializes the hypothesis in G so that it reads, "cough is present and dyspnea is absent," thereby ensuring that the new description will fail to match the negative example. Thus misclassified positive and negative examples lead to revisions that remove the contradiction.

As shown in the examples above, the version-space approach to learning involves both specialization and generalization of the theory itself and of the descriptions composing the theory. Removal or specialization of descriptions in G specializes the theory. In the first case, the learner reduces the restrictions to being a negative instance, and in the second case, the specialization increases the difficulty of being a positive case (i.e., more of the case's observed

values must match). In contrast, the removal or generalization of descriptions in S generalizes the theory. That is, fewer members of S and fewer conditions imposed by those members lead to fewer requirements for an observation to match this lower boundary and to be classified as positive.

Version spaces serve as a starting point for understanding how Darden's strategies can be implemented. Though limited in utility the formalism shows that anomalies can play a crucial role in developing our knowledge of concepts. From here, we can explore further system implementations to understand how anomalies drive learning and how resolutions can be identified.

1.4 THEORY REVISION SYSTEMS

Most of the relevant research on anomalies in artificial intelligence falls under the category of theory revision. Theory revision systems consist of programs that alter a knowledge base when faced with contradicting data. These systems tend to be autonomous, and primarily alter symbol-based theories. While differing in structure, they each embody six primary tasks of the revision process:

1. anomaly detection
2. fault localization
3. revision generation
4. revision assessment
5. revision application
6. expectation evaluation

Although all incremental learners update their beliefs, TEIRESIAS [14] was one of the first systems to address the six tasks of theory revision in detail. Working with the program, a domain expert presents cases and evaluates the resulting classification. If the system produces an incorrect (i.e., anomalous) prediction, the expert queries the program about its rationale to detect the fault in the knowledge base. The expert and system then collaborate

to generate an anomaly resolution, with the expert having the final say in its suitability. Application of the repair can consist of a complex combination of rule additions, deletions, and alterations, the result of which TEIRESIAS evaluates within the context of its own expectations. The program communicates any unmet expectations to the expert, who then determines whether and how they should be resolved.

EITHER [42] contrasts with TEIRESIAS in that it automates the tasks of theory revision. Instead of relying on a domain expert to pose cases and evaluate the output, the program examines a set of supervised data, identifying anomalies as cases where the prediction fails to match the observed outcome. When EITHER fails to prove that a datum belongs to the observed class, its abduction component backtracks from the example to determine the facts required for a correct prediction. These facts are collected into sets that indicate how to generalize the rules in the theory. After creating the generalizations, or after incorrectly classifying a datum, the program generates, assesses, and applies its own repairs. Unlike TEIRESIAS, where these steps were distinct, EITHER performs them atomically due to the strict constraints on its search. In general, the system applies minimal alterations to the original theory until it correctly classifies all the given data. Since it possesses no meta-level expectations (as TEIRESIAS does), which is typical of automated theory revisers, EITHER does not address the final step of expectation evaluation.

EITHER and its successor [5] are representative of automated revision systems. Beginning with a theory, these programs identify anomalies by locating either unclassified or incorrectly classified data. While most systems address both cases, a few specifically target the former [10, 21, 62]. Methods for fault localization, which depend highly on the representation of the original theory, often emphasize abduction, although some programs employ other means. For example, PTR [29] represents the domain theory as a graph and adjusts edge weights to identify the point of revision while RAPTURE [38] employs backpropagation to alter a network representation of the rules. As with EITHER, automated revision programs combine the fault localization task with the generation, assessment, and application of revisions. That is, given an anomaly, these programs search for and apply the first minimal alteration that they find.

FORTE, STALKER, and CLARUS serve as exceptions to the minimal-alteration approach to revision assessment. FORTE [49] considers a large number of revisions at once, selecting the one that resolves the most anomalies in the data. The process of evaluating the theory, identifying anomalies, and applying revisions then repeats until all anomalies are resolved. STALKER [8], on the other hand, examines the anomalies individually. For each anomaly, the program generates all possible revisions, applies each revision to the original theory, and counts the number of examples that the altered theory correctly classifies. The program then applies the revision that classifies the most examples without introducing new anomalies. CLARUS [7] differs from both FORTE and STALKER in that it introduces semantics into its performance metric. Instead of relying on predictive accuracy alone, the system also attempts to minimize linguistic heterogeneity. That is, unless substantial gains in predictive accuracy can be obtained, CLARUS prefers to introduce predicates using familiar terms and concepts. To determine familiarity, the program examines linguistic connections using WordNet [17].

HYPGENE and KEKADA are two more theory revision systems of note. Unlike the others, HYPGENE [28] takes a domain specific approach to theory revision, operating on models of gene regulation mechanisms. This system extends the scope of the revision process, considering alterations in experimental conditions as well as changes to the knowledge base. KEKADA [33] couches revision within a closed-loop scientific reasoner. That is, the program not only analyzes data, but also suggests experiments, interprets results from those experiments, and updates its knowledge or working hypothesis to reflect its findings. These two systems emphasize the importance of theory revision to scientific discovery and indicate directions in which the list of primary tasks can be extended.

Apart from CLARUS and KEKADA, the systems mentioned above assess and select revisions based upon a criterion of conservatism. That is, they attempt to enact the minimal possible change to the original theory that accounts for the anomalous data. For instance, a conservative system would add another epicycle to the Ptolemaic model as opposed to placing the Sun at the center, or, in terms of a version-space learner, a single condition would be added or removed from a hypothesis even though both data and domain knowledge may support a more drastic change. In principle, this strategy reduces future backtracking due

to poor choice, thereby reducing risk. However, a conservative mindset could not produce a heliocentric model from a geocentric one.

Unfortunately, conservatism not only keeps us from new ideas, but also fails to prevent needless backtracking as it promises. The number of data required to fully expresses the feature space increases exponentially with the number of features. For instance, the space defined by two binary features can be completely covered using four cases. However, given n features, full exploration of the space of possible alterations requires 2^n data. Data sets large enough to account for the size of their feature space tend to be rare. For example, in scientific domains such as biology, data sets extracted from DNA microarrays possess thousands of features, but usually contain fewer than one hundred observations. This relatively sparse data set can support several equally conservative rehabilitative and monotonic revisions. Choosing among these revisions becomes arbitrary, and guarantees of safety in minimal change fail to hold.

Both CLARUS and KEKADA employ strategies to overcome the weakness of conservatism as a selection criterion. By considering semantic relationships among features, CLARUS strengthens its notion of conservatism. That is, the system avoids knowledge unrelated to the portion of the theory under revision. However, as the authors note, lexical tags constructed manually lead to better performance than that obtained from the semantic information gathered from WordNet. KEKADA's approach involves the inclusion of several domain-specific and domain-general heuristics that further constrain the set of admissible revisions. This program's major limitation lies in the ad hoc nature of these heuristics. Specifically, the description of KEKADA lacks a detailed analysis of the contributions of each heuristic, likely due to their loose, yet tangled, structure within the program. Since we would like to explore heuristics such as those used by KEKADA and CLARUS, and since KEKADA provides a poor test bed, we designed a system that would better facilitate our goal.

Revisiting the six components of theory revision, we introduce our system, Kalpana[4], to explore aspects ranging from anomaly detection to revision assessment. For Kalpana, anomaly detection occurs, not unlike the systems above, when it makes an incorrect clas-

[4]Kalpana is a Sanskrit word that roughly translates to "inventing" or "fashioning."

sification. Since the program's models are based on single-step classifiers (i.e., the chain of reasoning for a classification consists of a single IF-THEN rule), fault localization consists of identifying a classifier that incorrectly matches the observation. Revision generation occurs in a manner similar to that of STALKER. In particular, Kalpana generates multiple revisions, each consistent with all noncontradictory observations, in response to each anomaly. These revisions may serve as hypotheses for future experimentation or as indicators of erroneous data. The program then assesses these revisions based on syntactic and semantic cues.

1.5 ENSURING ACCEPTABLE REVISIONS

With Kalpana, we hope to address one particular limitation of prior theory revision systems: the identification of defensible hypotheses. Since it is a theory revision system, any anomaly resolutions produced by Kalpana satisfy our criterion of rehabilitation. In addition, the program enforces monotonicity by comparing all revisions to the noncontradictory data available. However, we do not require the system to produce a revision for every anomaly. Instead we ask Kalpana to examine those revisions that are produced and to determine which of those are defensible. To this end we explore several approaches to defensibility, some of which have been considered in previous research.

Of the systems described in Section 1.4, few attempt to judge their revisions in a manner related to defensibility. To this end, CLARUS [7] measures the distance in a semantic network between new terms and those present within the theory. This strategy requires the assumption that semantic distance correlates well with the plausibility of the revisions.[5] Intuitively, CLARUS expects researchers to prefer those revisions that employ language similar to what already exists within the theory. Although we expect that defensible revisions will contain words or features well-used within the theory's domain, and we incorporate a similar measure in Kalpana, semantic distance alone fails to capture other important characteristics of defensibility.

[5]The approach taken by CLARUS relates well to Nelson Goodman's concepts of entrenchment and projectibility [24], which we discuss in Section 5.2.2.

Pazzani, a coauthor of the CLARUS system, and colleagues [44] have explored another measure related to defensibility. In particular, they found that their rule-learning system produced classifiers that contradicted expert knowledge. The domain experts evaluating these classifiers had difficulty understanding and accepting the contradictory rules. As an example, Pazzani cites a learned rule for classifying normal patients and those with Alzheimer's disease. This rule states, "If the years of education of the patient is > 5, and *the patient does not know the date* and *the patient does not know the name of a nearby street*, then the patient is normal." In this statement, the italics indicate the two conjuncts more often associated with the alternative classification.

To ensure that only those features normally associated with a consequent appear in the antecedent of a rule, the authors introduced monotonicity constraints. These constraints differ from our definition of monotonicity in that they introduced semantics, in essence giving another measure of defensibility. Thus while CLARUS attempts to ensure the meaningfulness of newly introduced terms, monotonicity constraints ensure that newly asserted knowledge does not violate background knowledge relevant to the theory. To this end, the rule learner was modified to understand expert-specified relationships between attribute values and classifications, such as the association of a patient's forgetfulness with the presence of Alzheimer's disease. Then, when building the classifier, the system can avoid features in violation of the stated relationships, with the exception that substantial evidence can override a constraint.

As evidence of the generality of monotonicity constraints, consider the use of fact polarization in predicting legal outcomes. Lawlor [36], in his effort to define a science of law, suggests that legally relevant facts should be considered with respect to the effect of their presence on a court's decision. For instance, in a trade secrets case, the fact, "product easy to reverse engineer," would be negatively polarized for the plaintiff, meaning that its presence should only be used in predictors favoring the defendant. Due to the apparent usefulness of these types of constraints, we added them to the heuristics for defensibility that we explore with Kalpana.

Apart from the work by Pazzani and his colleagues, most measures of a revision's acceptability are syntactic in nature. These syntactic measures are tied more closely to formal learning theory than to the inherently semantic notion of defensibility. The two most fully

explored syntactic approaches emphasize conservatism and simplicity. Conservatism arises primarily in the area of belief revision, the logical basis of which derives from work by Alchourrón, Gärdenfors, and Makinson [2] (the AGM theory). Unfortunately, in practical application, equally conservative revisions are plentiful even when we constrain which alterations we will consider. As for simplicity, Kolmogorov complexity [57] influences many of the measures. However, Kolmogorov complexity requires a fixed terminology to avoid Nelson Goodman's paradox of simplifying a theory by introducing single terms that are artificial constructs of many other terms [23]. That is, we can always generate syntactically simpler hypotheses by altering our vocabulary, thereby hiding the complexity of the data with a smaller number of terms.

1.6 OUR APPROACH

In Kalpana we introduce measures of defensibility related to the work by Pazzani and colleagues as well as newly developed measures. We base our collection of measures on the six virtues of hypotheses discussed by Quine and Ullian in *Web of Belief* [47]:

1. conservatism—preservation of prior beliefs
2. modesty—use of familiar terminology
3. simplicity—lack of unnecessary information
4. generality—applicable to a wide range of events
5. refutability—capable of being disproved
6. precision—statement of clear, distinct boundaries

Kalpana accounts for the final three virtues implicitly. A general-to-specific search leads to the most general revision that meets all the constraints of the system. Additionally, each anomaly resolution is encoded as an IF-THEN statement with testable conditions, thereby meeting the requirement of refutability. And finally, the precision of the revisions derives from external constraints on the values allowed within the revisions themselves.

Though the last three virtues are implicit within Kalpana's design, we explicitly implemented the first three. Kalpana's measure of conservatism relates closely to Pazzani's monotonicity constraints. Specifically, expert-provided knowledge enables the system to prefer revisions consistent with prior beliefs. Modesty, which resembles Nelson's Goodman's notion of projectibility [24], keeps the program from suggesting unlikely relationships. For instance, a claim that subluxations caused a patient's pneumonia would be immodest. In a sense, CLARUS sought modest revisions by limiting the terms available to a semantically related subset. And finally, simplicity should keep the system from forming a theory by memorizing the data. By using these virtues as guidelines, we have worked to endow Kalpana with the knowledge required to recognize acceptable revisions to theories.

To explore anomaly-driven theory revision, we performed three sets of experiments. First, we tested whether an anomaly-driven approach would lead to the production of acceptable revisions. Second, we explored the effect of applying acceptable revisions to a theory in need of repair. And third, we developed and tested measures of defensibility to better identify the acceptable revisions. Before giving the results of these experiments, we introduce Kalpana's architecture and provide an annotated example of the system's performance.

2.0 KALPANA: THE PROGRAM

We built Kalpana to test our hypotheses while exploring the practice of anomaly-driven theory revision. As with most systems, Kalpana expects input and produces output. The input consists of data, a model,[1] and background knowledge. The program outputs a collection of anomaly resolutions paired with acceptability ratings. After describing the format of Kalpana's input and output, we describe the respiratory syndrome (RS) domain and give an annotated example of Kalpana operating within that domain.

2.1 INPUT AND OUTPUT

Kalpana expects the data to be represented as labeled feature vectors. A feature vector, and hence a datum, consists of one or more values for some preselected attributes, which may reflect either an observation (e.g., age, temperature) or a theoretical concept (e.g., body mass index, density). In the latter case, the values are calculated in advance and are represented no differently from observed values. All the values taken together constitute a single case. For example, consider the vector *[(color, red), (has_rings, no), (has_water, no), (oxygen_rich_atmosphere, no) (supports_life, no)]*, which could describe the planet Mars. The first three attributes are directly observable, whereas *supports_life* and *oxygen_rich_atmosphere* are theoretical constructs.

[1] Little terminological conformity exists within the machine learning community when it comes to the terms "theory" and "model." However, some choose to reserve the word "theory" for explanatory knowledge structures. While the collection of sentences used within our system may have explanatory power, we would also like to consider statements that are solely descriptive. Thus, we opt for the use of the less contentious term "model."

Table 1: Three truncated data from the respiratory syndrome domain. In the data set used for our experiments, 65 attributes were included in the feature vector for each of 282 cases.

Data					
ID	**Cough**	**Wheezing**	**...**	**Pharyngitis**	**Respiratory Syndrome (RS)**
1	present	absent	...	absent	absent
2	present	absent	...	present	absent
3	absent	present	...	present	absent

Kalpana explores the data in the context of a model, which consists of a collection of statements describing a set of target concepts. These statements or rules are represented as propositional Horn clauses where the positive term always indicates the target classification. Thus the model contains no intermediary concepts. For example, we might have a rule of the form, "IF *oxygen_rich_atmosphere* is *no* and *has_water* is *no*, THEN *supports_life* is *no*," where the ample presence of oxygen determines whether life can exist. This restriction may lead to models with both more and longer rules, but it does not alter which classifiers can be described by the modeling language.

In addition to the collection of rules, a model consists of a method for resolving internal conflicts. That is, if two rules predict different classes for a datum, examining the partial ordering of the rules will determine that case's final class.[2] Kalpana expects the rules in the model to be partially ordered by specificity. So, given two clauses such that the antecedent of the first is a direct subset of the antecedent of the second, the consequent of the second will be asserted. Or, more intuitively, a direct specialization of a particular rule defines an exception to that rule. For instance, if a datum matches the antecedents of both "IF *headache* is *present*, THEN *RS* is *absent*" and "IF *headache* is *present* and *pneumonia diagnosis* is *present*, THEN *RS* is *present*," then the model will classify the datum using the latter,

[2]Apart from the rule conflicts resolved by the imposed partial ordering, Kalpana assumes that the rules classifying a particular case all make the same prediction.

more specific rule. If an internal conflict cannot be resolved, which occurs when one of the conflicting rules is not a direct subset of the other, then the case becomes an anomaly.

Anomalies can surface in two ways given the described model representation.[3] Either a non-empty set of rules predicts the incorrect class for a datum, or two or more sets of rules assign mutually exclusive classes to the datum such that the conflict cannot be resolved using the partial ordering of rules. For example, consider the data in Table 1 and the model composed of the following four rules:

1. IF *cough* is *present*, THEN *RS* is *present*.
2. IF *wheezing* is *present*, THEN *RS* is *present*.
3. IF *pharyngitis* is *present*, THEN *RS* is *absent*.
4. IF *wheezing* is *present* and *pharyngitis* is *present*, THEN *RS* is *absent*.

The first datum is anomalous because the sole applicable rule, Rule (1), gives an incorrect classification. The same rule identifies the second datum as an anomaly. Although Rule (3) produces the correct classification, the model lacks a partial ordering that would allow us to favor Rule (3) over Rule (1). In contrast, Rule (2) incorrectly classifies the third datum, but since Rule (4) directly specializes the more general rule and produces the correct classification, this datum is not anomalous. Anomalies such as the first and second cases lead Kalpana to specialize all the overly general rules predicting the incorrect class. So, to rehabilitate the model with respect to the second datum, Kalpana would produce an exception rule of the form, "IF *cough* is *present* and *pharyngitis* is *present*, THEN *RS* is *absent*."

In addition to data and a model, Kalpana accepts background knowledge. If present, this background knowledge aids in judging the acceptability of all generated revisions. In our current implementation, Kalpana requires both the declarative knowledge and the means for interpreting that knowledge. For instance, we can state which attributes, when present, normally indicate the presence of respiratory syndrome. We must also provide a function that produces an exception-rule's acceptability-score based on the presence of these attributes. For example, the knowledge may be that an emergency department diagnosis of pneumonia is highly correlated with the presence of respiratory syndrome. So that Kalpana can apply

[3]We do not assume that a model classifies all the data that it sees. Though some data may remain unclassified, Kalpana does not consider them anomalous.

Anomalies Resolved: (237)
Original Rule:
(SPUTUM is PRESENT) implies (RESPIRATORY_SYNDROME is PRESENT)

Exception Rule:
(FEVER is ABSENT) and
(DYSPNEA is ABSENT) and
(COUGH is PRESENT) and
(CHEST_PAIN is PRESENT) and
(SPUTUM is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)

Method of Generation: Method of Difference: Compared to all data
Semantic Defensibility: 1.0
Syntactic Defensibility: 0.38

Figure 1: An example revision produced by Kalpana.

this knowledge, we would provide a function that gives a low score to rules similar to "IF X and *a diagnosis of pneumonia* is *present*, THEN RS is *absent*," where X stands for zero or more conjuncts.

Once provided with the described input, Kalpana produces a collection of revisions, which are sorted by both the anomalies that they resolve and their acceptability. Figure 1 shows an example revision. Kalpana first prints a list of the data that the revision resolves (e.g., datum 237) followed by the overly general rule responsible for the anomaly. Next, the program lists the anomaly resolution, the generation heuristic that produced it, and the revision's semantic and syntactic defensibility. In this case, we see that the rule "IF *sputum* is *present*, THEN RS is *absent*" has been specialized to include references to fever, dyspnea, cough, and chest pain (other revisions include other ways of specializing the original rule). Chapter 3 describes the various generation heuristics, and Chapter 5 details the functions that produced the defensibility scores. In both cases, a higher score indicates a more defensible revision, with the semantic score ranging freely over the reals and the syntactic score falling between zero and one.

2.2 EXAMPLE

For this example, and for subsequent experiments, we used medical data gathered from emergency department (ED) reports. Using these reports, we intend to develop a model of RS, which serves as a general class encompassing disease in the lower respiratory system. Such a model can be used to detect localized outbreaks of related illnesses. For instance, bioterrorist attacks (e.g., anthrax dissemination) or natural causes (e.g., severe acute respiratory syndrome), can lead to an increase in lower respiratory complaints. The cause for these complaints may initially be misdiagnosed due to its rare or unknown nature, but by grouping the related cases under a more general class, a previously hidden trend may surface, leading to quicker response. Since the resulting model should be both sensitive and defensible, this task fits well with our goals for Kalpana.

2.2.1 Input

The data used in this study were extracted from 282 free text ED reports. Due to the relatively low prevalence of respiratory disease, even in an ED environment, these reports were fortified with positive examples. Half of the reports were randomly selected from a collection of cases describing a respiratory ailment, while the other half were randomly taken from nonrespiratory cases. The presence of an International Classification of Diseases, 9th Revision [1] diagnostic code (ICD9 code) indicating a respiratory ailment distinguished the sets of cases. Of the 282 data, 190 were used for training purposes with the remaining 92 set aside for testing. Sixty-five relevant attributes were identified by Dr. John Dowling, an expert in infectious diseases, through his examination of a separate set of reports. A separate physician then read through the reports and assigned values for the selected attributes. In addition, this physician as well as two others assessed whether the patient described by the reports was a positive case for respiratory syndrome.

The 65 attributes selected by our domain expert, which are listed in Table 2, belong to four classes: signs and symptoms, physical findings, chest radiograph findings, and diagnoses. The first class consists of patient-reported complaints such as cough, dyspnea (diffi-

Table 2: The attributes selected for the identification of respiratory syndrome.

Signs and Symptoms	Physical Findings	Chest Radiograph Findings	Diagnoses
congestion	breath sounds decreased	pneumonia x-ray	acute coronary syndrome
cough	cyanosis	pulmonary edema x-ray	anxiety
dyspnea	dullness	widened mediastinum	aortic dissection
hemoptysis	oxygen desaturation	atelectasis	cardiomyopathy
pleuritic pain	rales/crackles	hyperinflated lungs	chest trauma
sputum	rhonchi	mass	hiatal hernia
chest pain	stridor	mediastinal shift	pharyngitis
conjunctivitis	tachypnea	pericardial effusion	pneumothorax diagnosis
stomatitis	wheezing	pleural effusion	pulmonary edema from congestive heart failure
upper abdominal pain	abdominal distension	poor inspiration	pulmonary embolus
chills	cervical adenopathy	pneumothorax x-ray	sarcoidosis
headache	chest tenderness		sepsis
pneumonia history	pleural rub		viral syndrome diagnosis
flu symptoms	subcutaneous edema of chest or neck		asthma
sweats	tachycardia		bronchitis
	fever		chronic obstructive pulmonary disease (COPD)
			cystic fibrosis
			empyema
			HIV/AIDS
			influenza
			lung tumor
			musculoskeletal chest pain
			pneumonia diagnosis

culty breathing), and headache, whereas the attending physician observes the attributes in the second class. Chest radiograph findings come from descriptive reports of patient x-rays, which may indicate the presence of pneumonia, a mass within the lungs, or other features, and the diagnoses record the ED physician's classification of the specific case given limited exposure to the patient. The diagnosis category includes both respiratory and nonrespiratory conditions.

When extracting features from the text reports, a physician assigned nominal values to each of the attributes. In particular, he marked attributes as *absent* when the report explicitly listed them as such and *missing* when they were not mentioned. Additionally, those attributes listed within the report as present were labeled *present* for physical and chest-radiograph findings, and one of *acute*, *chronic*, or *indeterminate* for diagnoses, signs, and symptoms. For the purposes of this study, an *acute* condition must be present for less than two weeks, otherwise the condition is labeled as *chronic*. When the physician could not determine the duration of a condition from the ED report, he labeled the attribute *indeterminate*. To reduce the risk of overfitting, we made the attributes binary by considering *present*, *acute*, *chronic*, and *indeterminate* features to be *present*, and both *absent* and *missing* features to be *absent*.

To determine the target classification of each case, we merged the ratings of three physicians. After reading the original ED reports, these physicians labeled each case as *acute*, *chronic*, or *absent* with respect to RS. As with the core features, we interpreted both *acute* and *chronic* to indicate presence, leaving us with two mutually exclusive concepts. To produce a gold standard, we used the majority opinion of the three physicians. So, a case rated as *acute* by one physician, *chronic* by a second, and *absent* by a third is considered *present*.

Figure 2 shows one simple model used for this example. (We introduce other models later.) The antecedent of each rule consists of a feature that is positively correlated with the consequent. We applied this model to the training data and removed any contradictions, which left us with 153 observations. We then injected four known anomalies into this data set and used it, along with the model, as input for Kalpana.

In addition to data and a model, we provided the program with background knowledge so that it could calculate the defensibility of its revisions. We compiled the relevant in-

- IF *cough* is *present*, THEN RS is *present*.
- IF *wheezing* is *present*, THEN RS is *present*.
- IF *sputum* is *present*, THEN RS is *present*.
- IF a *positive pneumonia x-ray* is *present*, THEN RS is *present*.
- IF *dyspnea* is *present*, THEN RS is *present*.

Figure 2: M1: A simple model of respiratory syndrome.

formation, described in Chapter 5, from a two hour directed discussion with our domain expert. In short, the collected knowledge addressed three of the virtues of hypotheses mentioned in Section 1.6: conservatism, modesty, and simplicity. Kalpana combines individual measures of these virtues into a defensibility score with a higher value indicating a more defensible revision.

2.2.2 Output

Kalpana produced 12 revisions for the 4 anomalies. Figure 3 shows two of these revisions while Appendix A lists the entire set. When reporting a revision, the program first gives the list of resolved anomalies followed by the original rule upon which the revision was based, the revision itself, the method used for generating the rule, and two scores of defensibility. For instance, the first revision in Figure 3 resolves the anomaly with case number six. This datum had been incorrectly classified as having RS due to the presence of dyspnea. By searching for differences between the anomaly and those positive cases of RS correctly classified by the model, Kalpana determines that the presence of acute coronary syndrome justifies an alternative consequent. The resulting rule has a semantic defensibility score of 1.5 and a syntactic score of 0.6, both of which are the highest values out of all 12 revisions.

Once Kalpana produced the revisions, we sent truncated versions containing only the anomalies resolved, the original rule, and the exception rule to our domain expert. He deemed both revisions in Figure 3 acceptable. For the first, he gave the reason, "Acute

coronary syndrome explains dyspnea," and for the second he wrote, "Musculoskeletal chest injury explains dyspnea," with the implication that the injury caused the chest tenderness. Here the use of the word "explains" captures an intended interpretation of these anomaly resolutions. That is, although the rules produced by Kalpana merely describe the data, their connection to the expert's knowledge of causation within the domain gives them explanatory power. Thus the new features explain away the features that normally indicate the incorrect class. In particular, dyspnea alone signifies the presence of RS. However, when acute coronary syndrome causes dyspnea, the patient's difficulty breathing has an alternative cause that effectively nullifies it as evidence of RS.

Anomalies Resolved: (6)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)

Exception Rule:
(ACUTE_CORONARY_SYNDROME is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)

Method of Generation: Method of Difference: Compared to data of the originally assigned class
Semantic Defensibility: 1.5
Syntactic Defensibility: 0.6

Anomalies Resolved: (7)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)

Exception Rule:
(CHEST_TENDERNESS is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)

Method of Generation: Method of Difference: Compared to data of the originally assigned class
Semantic Defensibility: 1
Syntactic Defensibility: 0.6

Figure 3: Example revisions produced by Kalpana.

3.0 GENERATING RESOLUTIONS TO ANOMALIES

To test whether an anomaly-driven approach leads to the production of acceptable revisions, we designed a generator of model revisions. As input, the generator takes both data, in the form of previously classified feature-vectors with noisy records and missing values, and a model, which consists of a disjunctive set of single-step classifiers in propositional Horn clause form (see 2.1). We assume that continuous attributes have been sectioned into intervals or replaced with nominal values during the generation of the original model. To resolve contradictions within the model, direct specializations (i.e., exception rules) have priority over their more general base rules (see p. 18).

The revision generator begins by identifying anomalies, which are contradictions between the data and the model. Once Kalpana identifies the anomalies, it partitions the data into subsets that reveal salient aspects of the anomalies. Then the program searches for both differences between the anomaly and subsets of nonanomalous[1] data and similarities among specific partitions of anomalies. In each case, the generator applies versions of John Stuart Mill's methods of induction [39], producing zero or more revisions for each anomaly. Finally, Kalpana outputs the resulting revisions along with an evaluation based on our criteria of acceptability, with rehabilitation and monotonicity being strictly enforced.

[1]We coin the distinction of *nonanomalous* (versus anomalous) data to refer to data that are wholly consistent with a specified model.

3.1 IMPLEMENTATION

3.1.1 Foundation

To generate explanations of anomalous data, Kalpana applies modifications of old inductive principles. In his classic work on the scientific method, John Stuart Mill [39] identified four methods of induction that have been rediscovered and recast numerous times[2]. Our work concerns the first two methods: the *method of agreement*

> "If two or more instances of the phenomenon under investigation have only one circumstance in common, the circumstance in which alone all the instances agree is the cause (or effect) of the given phenomenon"

and the *method of difference*

> "If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon."

In a feature vector representation, a "circumstance" is an attribute-value pair, or feature. Using the method of agreement, we ask how each anomaly resembles specific subsets of other anomalies, and using the method of difference we ask how each anomaly differs from various subsets of nonanomalous data. By using Mill's methods to guide our revision generators, and by applying the generators to subsets of data defined by the anomaly, we conjecture that acceptable revisions will result.

3.1.2 Method of Agreement

We based the first generator for revisions on the method of agreement. Strict application of this method requires a researcher either to record a complete description of the universe for each datum or to ensure that the subset of recorded features captures the complete set of causally relevant events or properties—both of which are impossible to meet. The former option requires complete state knowledge of the universe, whereas the latter requires

[2]While we recognize that Mill's methods are overly simplistic, we believe that they provide a useful starting point for any work in explanation.

omniscience about the studied phenomenon. Therefore, to make progress in the process of induction, we assume that the researcher records at least some of those features constituting the known and suspected causes of the studied phenomenon (or more accurately, that the researcher has selected those features with a high prior probability of predicting or influencing the outcome of the experiment). This assumption does not claim that all relevant features are present or conversely, that all present features are relevant, just that the researcher does not purposefully sabotage the learning system. With this assumption in mind, we set the phenomenon or class to be studied to "anomaly" and explore the causes by asking the following questions:

A1. How is this anomaly similar to other anomalous data?

A2. How is this anomaly similar to other anomalies with the same observed outcome?

A3. How is this anomaly similar to other anomalies incorrectly classified by the same rule?

A4. How is this anomaly similar to other anomalies incorrectly classified by the same rule that share the same observed outcome?

Each of these four questions requires the existence of multiple anomalies, and in all but the first case, the anomalies under comparison must have aspects in common apart from being an anomaly. With Question A1, Kalpana attempts to determine how the new datum resembles prior observations that are also anomalous. Finding similarities in A1 can be understood as an attempt to create a classifier for the concept "anomalous" using only positive examples. Question A2 limits the domain of the first question to those anomalies that with the same observed class as the anomaly under question. For example, Kalpana might compare all anomalies positive for respiratory syndrome (RS). The next question in the list implies that a particular classifier may, by being too general, cause all the anomalies. For instance, consider a set of anomalies grouped by the rule, "IF *wheezing* is *present*, THEN *RS* is *present*." If each member of the set had the feature "*asthma* is *present*," (not an RS in our case), then the anomalies could be resolved with the exception rule, "IF *wheezing* is *present* and *asthma* is *present*, THEN *RS* is *absent*." Finally, Question A4 significantly restricts the subset of anomalies under consideration to the intersection of the data satisfying Questions A2 and A3. In the context of the previous example, we would require the anomalies to

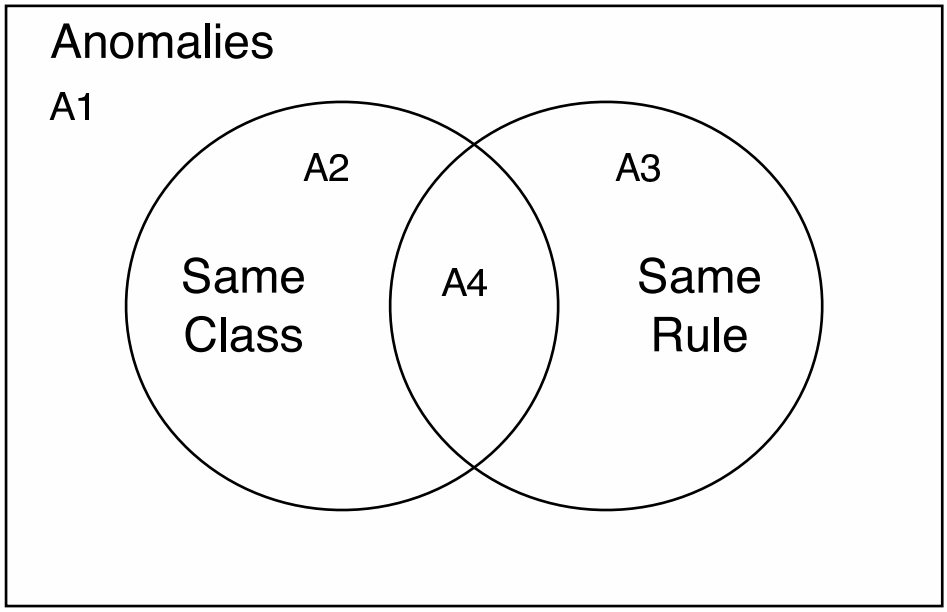

Figure 4: The relationships among subsets of data as defined by Questions A1–A4.

all be incorrectly classified by the rule about wheezing and be negative for RS (a trivial requirement in a dichotomous domain). Figure 4 displays the graphical relationship among the subsets of data defined by these questions.

To answer questions A1–A4, we define agreement, or similarity, to be a function of the features shared among the anomalies[3]. In particular, determining how two cases agree involves collecting the set of features that they have in common. This approach resembles the behavior of a specific-to-general concept learner. That is, if an entire set of anomalies shares the same feature, such as the presence of a cough, then that feature may indicate the cause of anomaly. While Mill's method of agreement specifies that we must find a single circumstance to imply causality, identifying all similarities leads to a revision that may be pared down using other knowledge or methods.

The revision generator presented in Figure 5 begins by identifying the pool of features common among the current group of anomalies (see Appendix B.4 for pseudocode describing

[3]Although we use a function that emphasizes matching features, others may wish to define similarity using a different function.

```
Method-of-Agreement(anomalies, data):
  // remove the anomalies from the rest of the data
  nonanomalies = data - anomalies
  // collect all features shared by the group of anomalies
  pool = shared-features(anomalies)
  // initialize a list of revisions
  revisions = [ ]
  // separately consider each anomaly
  for each a in anomalies
    // consider each rule that incorrectly classifies the anomaly
    for each i in (incorrect-classifiers(a))
      // create the root revision from the antecedent of the
      // incorrect classifier and the correct classification of
      // the anomaly
      r = create-root(antecedent(i), class(a))
      // add all the features that keep the new revision from
      // creating any new anomalies
      push(necessary-features(r, pool, nonanomalies),
           antecedent(r))
      // while the new revision continues to create new anomalies
      // and there are features that the program can use to
      // specialize the revision
      while (overly-general(r, nonanomalies) and
             features-left(pool, r))
        // add an unused feature that best separates the anomaly from
        // nonanomalous data incorrectly classified by the current
        // revision
        push(best-separator(r, pool - antecedent(r), nonanomalies),
             antecedent(r))
      // push the resulting revision onto the list
      push(r, revisions)
  return revisions
```

Figure 5: The algorithm for Kalpana's method of agreement generator.

the nontrivial subprocedures). Kalpana defines these groups using Questions A1–A4. For instance, suppose that the program identifies several anomalies, one of which contradicts the rule, "IF *wheezing* is *present*, THEN *RS* is *present*." Four groups of anomalies will be created: one composed of all the anomalies, a second composed of all the anomalies with RS absent, the third consisting of all anomalies with wheezing present and RS not present, and the fourth with all anomalies where wheezing is present and RS is absent. Since the target class in the RS domain is dichotomous, the third and fourth groups will be identical.

Next, the generator creates revisions for each incorrect classifier of an anomaly. So if the anomaly given above were also misclassified by the rule, "IF *cough* is *present*, THEN *RS* is *present*," the program would address this contradiction as separate from the one related to wheezing. When creating the revision, Kalpana begins with a root rule, which is unique for a specific anomaly–incorrect classifier pair, composed of the incorrect classifier's antecedent and a consequent that matches the anomaly's observed outcome. For instance, the root rule for the wheezing example would be, "IF *wheezing* is *present*, THEN *RS* is *absent*." The conflict resolution mechanism that we chose stipulates the form of this root because it requires exception rules, which are direct specializations of the incorrect classifiers, to resolve contradicting predictions. To create direct specializations, the program extends the consequent of the root revision with features from the pool.

If the nonanomalous data contain true positives for the original classifier, the root revision will introduce new anomalies, requiring further refinement to meet the monotonicity criterion. Refinement of the root revision begins with the identification of all necessary features within the pool of common features, where a necessary feature *uniquely* prevents the root revision from classifying a nonanomalous datum. To identify the necessary features, Kalpana creates one rule for each feature in the pool, such that the antecedent contains every feature from the pool except for that one. If the resulting revision incorrectly classifies a previously nonanomalous datum, then the program considers the missing feature to be necessary. For example, if the pool contains the features "*asthma* is *present*" and "*cough* is *present*," Kalpana will form the rules, "IF *wheezing* is *present* and *asthma* is *present*, THEN *RS* is *absent*," and, "IF *wheezing* is *present* and *cough* is *present*, THEN *RS* is *absent*." If the latter rule

produces a new anomaly (and the former does not), then "*asthma* is *present*" will be regarded as necessary.

After Kalpana adds all the necessary features to the antecedent of the root revision, the resulting rule can still be overly general. That is, the revision continues to produce contradictions in the nonanomalous data. When this happens, the agreement generator creates an order among the remaining features in the pool based upon the number of prevented contradictions. While the exception rule remains overly general, Kalpana selects unused features that best distinguish the anomaly from the nonanomalies from the pool, adding them to the rule's antecedent. If the pool of features is exhausted before the monotonicity criterion can be met, the revision as a whole is discarded.

3.1.3 Method of Difference

Our final two revision generators implement the method of difference. As with the agreement generator, we assume that the researcher records some, but not necessarily all, relevant features. Additionally, we loosen another restriction stated by Mill. In particular, he wrote that the method of difference requires two examples where one example must possess the characteristic under study (here the characteristic is "anomalous"), while the other must not. To infer causality, both examples must share all features except one. Meeting this requirement, especially when working with observational data, can be too difficult. For example, a research scientist has little control over which patients visit the hospital (except when that scientist is the admitting physician).

To enable the generation of revisions using the method of difference, we assume that any subset of features may in itself be considered a single "circumstance" or feature in accord with Mill. So, suppose that two patients, one anomalous and one nonanomalous resemble each other with the exception that the anomalous patient has asthma and a cough, while the nonanomalous patient does not. In this case, we consider "*asthma* is *present* and *cough* is *present*" to be a single unshared feature. Even when we can collect the necessary data, the existence of feature interactions (cancellations, feedback, etc.) renders this assumption both

reasonable and necessary. For instance, the anomaly may be caused (or explained away) by the presence of exactly two features.

Similar to questions A1–A4, the following four questions help us concentrate on the anomalous case when generating our revisions:

D1. How does the anomaly differ from all nonanomalous examples?

D2. How does the anomaly differ from nonanomalous examples with the same observed class?

D3. How does the anomaly differ from nonanomalous examples with the predicted class?

D4. How does the anomaly differ from nonanomalous examples with the predicted class that were classified by the mispredicting rule?

The first question asks which features of the anomaly separate it from all the nonanomalous data. To answering this question we could label the single case as *anomalous* and the correctly classified data as *nonanomalous* and then apply a general-to-specific rule learner to this relabeled data. Question D2 narrows the search to differences within the same observed class. As an example, suppose that a model incorrectly classified an anomaly as positive for RS (i.e., the case does not have RS although the model predicts that RS is present). Kalpana answers Question D2 by comparing the anomaly to the nonanomalous data where RS is observed to be absent. The idea that the anomaly indicates a previously unknown expression of the class motivates this question.

The final two questions compare the anomaly to data that the model claims it resembles. Question D3 uses all the nonanomalous data of the predicted class, with the intuition that features identified from this subset of data may be strong dividers between the two classes. Question D4 limits this group of data further, comparing the anomaly to only those cases that match the overly general classifier. The antecedent of the incorrect classifier used for grouping this subset of data can be viewed as a traversal down a single branch of a decision tree. The anomaly represents a case that the tree incorrectly classifies at a leaf node and spurs growth of that branch. While this approach has been studied in detail (e.g., [48]), research usually emphasizes how it affects predictive accuracy, whereas our interest lies in whether the approach leads to acceptable (and more specifically, defensible) hypotheses. Figure 6 shows the relationships among the defined subsets of data.

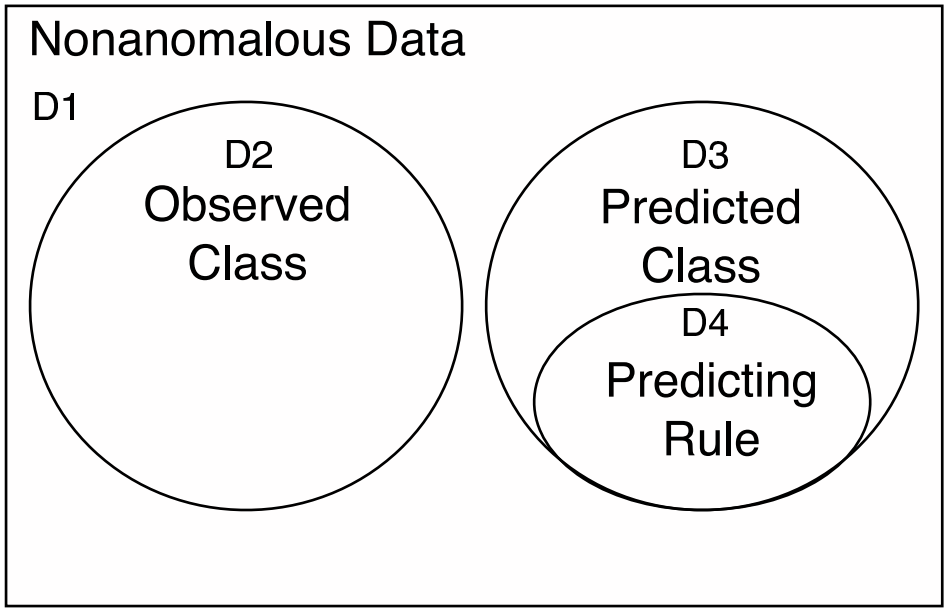

Figure 6: The relationships among subsets of data as defined by Questions D1–D4.

When answering these four questions, we define a difference to be a feature or set of features of the anomaly not shared by any nonanomalous data in the current subset. For instance, if the anomaly matches the set of features, "*cough* is *present* and *dyspnea* is *present*," and the nonanomalies do not, then that set of features constitutes a difference. When identifying differences, Kalpana employs one of two algorithms depending on whether it is searching for a single feature or a set of features. Figure 7 shows the algorithm for the former search, and Figure 8 illustrates the latter approach (see Appendix B.4 for pseudocode describing the nontrivial subprocedures).

The algorithm shown in Figure 7, called the basic method of difference, generates revisions by adding a single feature to the root revision, which is created in the same manner as in the method of agreement (see p. 32). For the method of difference, Kalpana considers each anomaly independently, identifying the separators (i.e., single-feature differences) that will be added to the root revision. Finding separators involves viewing each attribute individually, identifying the feature that corresponds to the attribute's observed value within the anomaly. The program retains only those features that do not match any nonanomalous data. Once

```
Method-of-Difference-Basic(anomalies, data):
  // begin with an empty set of revisions
  revisions = [ ]
  // remove the anomalies from the rest of the data
  nonanomalies = data - anomalies
  // individually consider each anomaly
  for all a in anomalies
    // consider each feature expressed in the anomaly that is
    // not expressed in any of the nonanomalous data
    for all s in separators(feature-set(a), nonanomalies)
      // consider each rule that incorrectly classifies the
      // anomaly
      for all i in incorrect-classifiers(a)
        // create the root revision from the antecedent of the
        // incorrect classifier and the correct classification of
        // the anomaly
        r = create-root(antecedent(i), class(a))
        // add the separating feature to the antecedent of the
        // root revision
        push(s, antecedent(r))
        // when the root revision meets the monotonicity
        // criterion, add it to the collection of revisions
        unless(overly-general(r, nonanomalies))
          push(r, revisions)
  return revisions
```

Figure 7: The algorithm for Kalpana's basic method of difference generator.

identified, Kalpana appends each feature to the antecedent of each root revision, creating one complete revision for each separating-feature–incorrect-classifier pair. If a resulting revision creates new anomalies when applied to the entire collection of nonanomalous data (by a rule-matching procedure), then it is discarded according to the monotonicity criterion.

Figure 8 displays the algorithm used by Kalpana to identify a set of features constituting a difference. This procedure executes only when the basic method of difference generator from Figure 7 fails to produce any revisions. While revisions produced by the basic method extend a decision branch one step in several directions, this generator lengthens the branch in a single direction using the anomaly as its guide. To do this, the program examines each anomaly independently in the context of one of its incorrect classifiers. Kalpana creates a root revision from the antecedent of the classifier and the observed class of the anomaly, and follows this with a search for the feature that best separates the anomaly from the current subset of nonanomalous data. For example, if the addition of "*cough* is *absent*" to the revision's antecedent places the anomaly in a group with three members of the nonanomalous subset and the addition of "*wheezing* is *absent*" places the anomaly in a group with five, then the program will select the first feature. If multiple features separate equally well, then Kalpana makes an arbitrary choice among them. This procedure continues until either the anomaly is completely segregated or all possible features have been added to the revision. In the latter case, the antecedent of the revision contains all the features of the anomaly. As with the other generators, if the resulting revision fails to meet the monotonicity criterion, Kalpana discards it.

3.1.4 Time Complexity of the Revision Generators

The runtime of Kalpana varies depending on which generators a particular problem requires, but we can analyze the running times of the individual generators with respect to their worst-case performance. Even with gross estimates, each of the three algorithms employed has polynomial time complexity in terms of the input. To establish the runtime bounds we begin with two assumptions. In particular, we consider the act of matching a feature to a datum to take constant time, and we assert that rules may contain at most one value per

```
Method-of-Difference-Decision-Branch(anomalies, data):
  // begin with an empty set of revisions
  revisions = [ ]
  // remove the anomalies from the rest of the data
  nonanomalies = data - anomalies
  // individually consider each anomaly
  for all a in anomalies
    // consider each rule that incorrectly classifies the anomaly
    for all i in incorrect-classifiers(a)
      // create the root revision from the antecedent of the
      // incorrect classifier and the correct classification of
      // the anomaly
      r = create-root(antecedent(i), class(a))
      // while the revision fails to meet the monotonicity
      // criterion and while the features matching the observed
      // values in the anomaly have not all been used in the
      // antecedent of the revision
      while(overly-general(r, nonanomalies) and
            (feature-set(a) - features(antecedent(r)) > 0)
        // find the one feature that separates the anomaly from the
        // greatest number of nonanomalous data in the current
        // subset
        // add that feature to the antecedent of the revision
        push(best-separator(r, feature-set(a) - antecedent(r),
                            nonanomalies),
             antecedent(r))
      // if the revision meets the monotonicity criterion, keep it
      unless(overly-general(r, nonanomalies))
        push(r, revisions)
  return revisions
```

Figure 8: The algorithm for Kalpana's decision-branch method of difference generator.

attribute (i.e., attribute values are mutually exclusive). In addition to these assumptions, we introduce following notation for the input so that we can effectively discuss complexity: α represents the number of anomalies, γ the number of nonanomalies, τ the number of attributes, ϕ the number of features (attribute-value pair), and ρ the number of rules.

We begin by analyzing the worst-case complexity of the helper functions, which may be found as pseudocode in Appendix B.4 when their implementations are nontrivial. Determining which features occur in all the anomalies requires time $O(\phi\alpha)$. Identifying the necessary features within the set of shared features requires $O(\gamma\phi\tau\rho)$ time in the worst case. That is, given a set of rules and features, we build all possible one-feature extensions of all the rules and match each rule to all the nonanomalous data. Finding the best separator within a pool of features takes time $O(\gamma\phi\tau)$ because each rule gets extended by each feature in the pool and is then matched to the data. Finally, we require $O(\gamma\tau)$ time to test a rule for over-generality, which entails matching the features within that rule to all the given data.

Having analyzed the helper functions within our algorithm that do not operate in constant time, we can proceed with the analysis of the main methods. The majority of the time spent within the method of agreement occurs within a nested loop involving the anomalies and the incorrect classifiers of each anomaly. Within this loop, we identify the necessary features for the explanation of the anomaly, and if required, we search for high quality separators until the revision completely separates the anomalies from the nonanomalies. Of these operations, the former takes the most time. Thus the method of agreement algorithm runs in time $O(\alpha\gamma\tau\phi\rho^2)$.[4] That is, it is polynomial in the size of the data set, the number of features, and the size of the model. We note that Kalpana applies this algorithm to each defined group of anomalies.

The method of difference algorithms differ significantly in their complexity due to the relatively simple search performed by the basic approach. The basic method of difference spends most of its time determining the generality of a specific revision. To clarify, after identifying all separators for each anomaly the algorithm appends those separators to each relevant rule, and then matches that rule to the nonanomalous data. So, the total runtime

[4]We have ignored the action of separating those features used within the revision from the features left within the pool of shared features. While this takes longer than constant time, $O(\tau)$ actually, its effect is additive and can be overestimated using the given runtime polynomial.

is $O(\alpha\gamma\rho\tau^2)$ because the number of attributes limits the number of separators as well as the number of features that can exist within a single rule. The decision-branch method of difference takes slightly longer time in the worst case since it can attempt to improve a rule beyond adding a single feature. In fact, the runtime for this method is $O(\alpha\gamma\rho\phi\tau^2)$, where the addition of ϕ indicates that the method may examine all features when building the revision.

We emphasize that the complexity provided for the hypothesis generation algorithms likely overestimates the actual worst-case bound. However, we wished to show that even in the extreme case, which may in fact be impossible, the algorithms still operate within polynomial bounds with respect to their input. Although we assumed that matching occurs in constant time, we note that the time required may be significant, thereby adversely affecting the observed runtime of Kalpana because of the sheer number of matches required. We have not attempted to optimize the algorithms in terms of matching the features to the data, but we suspect that a few minor changes would reduce redundancy and lower the provided bounds (e.g., see [4]).

3.2 EVALUATION OF KALPANA'S REVISION GENERATORS

3.2.1 Method

We evaluated Kalpana's revision generators using the acceptability of their proposed revisions in the domain of bioterrorism surveillance. We expected that the anomaly-driven approach of these generators would lead to a greater number of acceptable revisions than an approach not driven by anomalies. That is, the latter approach would miss acceptable revisions found by the former, indicating that context of the anomaly contains information useful when revising models. To test this hypothesis, we compared the numbers of total and acceptable revisions produced by all of Kalpana's generators on all subsets to the output from two subsets in particular. To clarify, suppose that we treat the problem as one of classification with the target concept being "anomaly." This situation resembles Kalpana's approach to answering Question D1. If, on the other hand, we retained the original target class, a decision-tree

- If *cough* is *present*, then *RS* is *present*.
- If *wheezing* is *present*, then *RS* is *present*.
- If *sputum* is *present*, then *RS* is *present*.
- If a *pneumonia x-ray* is *positive*, then *RS* is *present*.
- If *dyspnea* is *present*, then *RS* is *present*.

Figure 9: M1: Overly general model of respiratory syndrome (RS). (Same as Figure 2.)

generator would view the subset related to Question D4 when faced with the anomaly. Thus, Kalpana's results from these subsets approximates the behavior of a well-studied class of symbolic learners, and we use this behavior later (Chapter 5) as a baseline to measure performance. Additionally, this experiment allows us to explore the sort of revisions that Kalpana produces and to see whether acceptable revisions are relatively easy or difficult to find.

For our evaluation, we chose the task of identifying patients with respiratory syndrome (RS) using the data described in Section 2.2.1. Kalpana applied each of the three models shown in Figures 9, 10, and 11 to the 190 training data. Anomalies resulted in all applications, and Kalpana's revision generators were invoked to resolve them. The resulting revisions were paired with their respective overly general rules and shown to our domain expert. Using the instructions shown in Section 2.2.1, the expert judged whether each revision was acceptable.

- If *chest tenderness* is *present*, then *RS* is *absent*.
- If a *positive pulmonary edema x-ray* is *present*, then *RS* is *present*.
- If a *positive pneumonia x-ray* is *present*, then *RS* is *present*.
- If a *pneumonia diagnosis* is *present*, then *RS* is *present*.
- If *rhonchi* are *present*, then *RS* is *present*.
- If a *positive pleural effusion x-ray* is *present*, then *RS* is *present*.
- If *sputum* is *absent* and *headache* is *present*, then *RS* is *absent*.
- If *cough* is *absent* and *chest pain* is *present*, then *RS* is *absent*.
- If *dyspnea* is *present* and *chest pain* is *absent*, then *RS* is *present*.
- If *dyspnea* is *absent*, a *positive pneumonia x-ray* is *absent*, and *tachycardia* is *present*, then *RS* is *absent*.
- If *dyspnea* is *absent*, a *positive pneumonia x-ray* is *absent*, and *oxygen desaturation* is *absent*, then *RS* is *absent*.
- If *cough* is *absent*, *oxygen desaturation* is *absent*, and *rales/crackling* is *absent*, then *RS* is *absent*.
- If a *positive pneumonia x-ray* is *absent*, and *oxygen desaturation* is *absent*, and *a positive pulmonary edema x-ray* is *absent*, then *RS* is *absent*.
- If *wheezing* is *absent*, a *positive pneumonia x-ray* is *absent*, a *positive pulmonary edema x-ray* is *absent*, and a *positive pleural effusion x-ray* is *absent*, then *RS* is *absent*.

Figure 10: M2: A plausible model of respiratory syndrome (RS) extracted from the training data described in Section 2.2.1.

- If *chest tenderness* is *present*, then *RS* is *absent*.
- If *headache* is *present*, then *RS* is *absent*.
- If a *pulmonary edema-congestive heart failure diagnosis* is *present*, then *RS* is *present*.
- If a *positive pulmonary edema x-ray* is *present*, then *RS* is *present*.
- If a *positive pneumonia x-ray* is *present*, then *RS* is *present*.
- If a *positive pleural effusion x-ray* is *present*, then *RS* is *present*.
- If *wheezing* is *present*, then *RS* is *present*.
- If *dyspnea* is *absent* and a *positive pneumonia x-ray* is *absent*, then *RS* is *absent*.
- If a *positive pneumonia x-ray* is *absent*, *oxygen desaturation* is *absent*, and *chest pain* is *present*, then *RS* is *absent*.
- If *sputum* is *absent*, a *positive pneumonia x-ray* is *absent*, and *oxygen desaturation* is *absent*, then *RS* is *absent*.
- If *dyspnea* is *present*, *bronchitis* is *absent*, and *chest pain* is *absent*, then *RS* is *present*.
- If *sputum* is *present*, a *positive pneumonia x-ray* is *absent*, a *positive pleural effusion x-ray* is *absent*, and *asthma* is *absent*, then *RS* is *absent*.

Figure 11: M3: A second plausible model of respiratory syndrome (RS) extracted from the training data described in Section 2.2.1.

Table 3: The number of unique revisions generated by Kalpana for models M1, M2, and M3 along with the number of those revisions that were acceptable.

Model	Unique Revisions	Acceptable Revisions
M1	26	15
M2	89	22
M3	67	12
Total	182	49

Table 4: The number of revisions generated by Kalpana using Model M1.

Revisions	Method of Agreement Subsets					Method of Difference Subsets				
	A1	A2	A3	A4	All	D1	D2	D3	D4	All
All	0	0	3	3	6	14	2	15	12	43
Acceptable	0	0	3	3	6	8	1	10	10	29

3.2.2 Results

Our primary goal during evaluation was to determine the effectiveness of examining specific subsets of data when generating revisions. Tables 4, 5, and 6 show the results of Kalpana's runs on each of the three models respectively. These tables give the number of total revisions along with the number of acceptable revisions generated in response to Questions A1–A4 and D1–D4. Note that revisions could be duplicated within the same model and between models (though the latter never occurred in these experiments), so Table 3 lists the number of unique revisions and the number of those that were acceptable

Table 4 lists the results of applying the model in Figure 9 to the 190 data. In this table, the columns indicate which questions the subset of nonanomalous data relates to. The first row gives the total number of revisions generated by Kalpana from each subset, and the second row tallies the acceptable revisions. As shown in Table 3, the program generated 26 unique revisions, 15 of which were acceptable. The repeated revisions stemmed primarily from the generation of the same exception rule for the same anomaly when viewing separate

Table 5: The number of revisions generated by Kalpana using Model M2.

Revisions	Method of Agreement Subsets					Method of Difference Subsets				
	A1	A2	A3	A4	All	D1	D2	D3	D4	All
All	0	3	1	1	5	30	11	53	63	157
Acceptable	0	1	1	1	3	1	0	18	19	38

Table 6: The number of revisions generated by Kalpana using Model M3.

Revisions	Method of Agreement Subsets					Method of Difference Subsets				
	A1	A2	A3	A4	All	D1	D2	D3	D4	All
All	0	0	2	2	4	12	4	40	51	107
Acceptable	0	0	2	2	4	2	0	6	8	16

subsets of data. However, in some cases, different anomalies would lead to the same revision. Although most of the overlapping rules came from the method of difference generator, all three of the rules produced by the method of agreement were duplicated when Kalpana answered Questions A3 and A4. Of the total set of revisions, the domain expert judged all three unique exception rules produced by the method of agreement and many of the revisions produced from the method of difference to be acceptable.

Table 5 results from applying the model in Figure 10 to the 190 data. Here 22 of the 89 unique revisions were acceptable. The most fruitful subsets in terms of acceptable anomaly resolutions came from Questions D3 and D4. In this case, it appears that only one of the subsets defined by Questions D3 and D4 need be examined by the program since all 18 of the acceptable exception rules extracted from D3 were repeated using D4. The single acceptable revision produced via Question D1 was unique to that subset (i.e., no other subsets generated from the various anomalies led to the production of that revision), but the subset's signal to noise ratio appears to be low. As in the first set of results, the method of agreement produces a limited number of revisions that are almost always acceptable—and always so when we constrain Kalpana to viewing subsets from Questions A3 and A4.

Table 7: A summary of all the revisions generated by Kalpana from Models M1, M2, and M3.

Revisions	Method of Agreement Subsets					Method of Difference Subsets				
	A1	A2	A3	A4	All	D1	D2	D3	D4	All
All	0	3	6	6	15	56	17	108	126	307
Acceptable	0	1	6	6	13	11	1	34	37	83

Table 8: The number of unique revisions attributable to Question D4 along with the number of revisions that occur uniquely when addressing the other seven questions. The number of acceptable revisions appears in parentheses.

Model	Revisions from Question D4	Revisions from Other Subsets
M1	12 (10)	14 (5)
M2	63 (19)	26 (3)
M3	51 (8)	16 (4)

Table 6 shows how the various subsets contributed to the development of acceptable exception rules for the model in Figure 11. As with the former two experiments Questions A3 and A4, while producing a limited number of revisions, tended to result in acceptable rules. In contrast to the previous results, the method of difference yielded much fewer acceptable revisions in relation to the total number produced. Out of the 67 unique anomaly resolutions, only 12 were deemed acceptable.

Finally, Table 7 summarizes the number of revisions generated by Kalpana for all three models. The trends shown in this table mirror the other results. In particular, the subsets addressing Questions A3 and A4 produce the highest fraction of acceptable revisions, while D4 leads to the greatest number. Interestingly, A3 and A4 led to the generation of identical anomaly resolutions such that for our experiment, using both subsets was redundant. However, this need not always be the case. Table 8 illustrates the fruitfulness of D4 by showing how many revisions, total and unique, that Kalpana found looking just at the D4 subset ver-

sus all the other subsets. To generate this table, we looked at all the revisions, picking out those that stemmed from D4 and removing redundancies produced from other the subsets. All revisions not attributable to D4 were then grouped together to form the numbers in the second column, revealing that D4 was over twice as fruitful in terms of acceptable anomaly resolutions as the remaining seven subsets combined.

3.2.3 Discussion

At a glance, our results weakly support our claim that anomaly-driven theory revision leads to a greater number of acceptable revisions. Although Kalpana generated more revisions overall using our sets of questions as a guide, most of these came from Question D4, which served as a baseline. However, the results reveal three interesting findings. First, the method of agreement produced a surprisingly high proportion of acceptable revisions. Second, the excessively general theory (M1 in Figure 9) led to fruitfulness in several of the subsets. And, third, the most productive subsets of data, in terms of the number of acceptable anomaly resolutions produced, contained the fewest members.

Even though 307 of the 322 total revisions came from applying the method of difference, those 15 produced by the method of agreement were more likely to be acceptable. In fact, the anomaly resolutions generated using Questions A3 and A4 were always acceptable. We suspect that viewing groups of anomalies contributes to this effect. In particular, by comparing multiple anomalies (in these experiments from 2 to 23 cases), Kalpana reduces the likelihood of emphasizing coincidental differences between any single anomaly and the nonanomalous data. By avoiding some coincidental differences, Kalpana produces revisions that relate better to an expert's knowledge of the domain.

The next finding indicates that anomaly-centered subsets of data may be more useful when working with a very general initial model. The first model, M1, shown in Figure 9 contained only five rules and was obviously too general to accurately characterize RS. In this case, all but one of the subsets produced an acceptable revision. In contrast, the more specific models benefited most from more restricted subsets of data. That is, the characteristic of being an anomaly was less important than other factors, such as the rule that the anomaly

violated. This finding implies that the most fruitful subset of data for specific models comes from Question D4, which roughly corresponds to the data used by a decision-tree generator. Our approach differs slightly in that it separately considers each anomaly, whereas a decision-tree generator would consider multiple anomalies at one as long as they cluster at the same leaf. Thus we suspect that as a decision tree becomes more specific, its learning algorithm becomes more focused with respect to anomalies, leading to meaningful revisions.

Finally, we notice that Kalpana generates acceptable revisions from relatively few non-anomalous data. The subset defined by Question A4 was not only the best subset for the method of agreement but also the most restrictive of the group. The same can be said of Question D4's subset with respect to the method of difference. This preliminary finding implies that model revision depends primarily on the relationships among the data as opposed to the amount of data available. That is, blindly gathering more data to improve a model will not prove as helpful as collecting very specific data centered on an anomaly. So, when data production can be controlled, the best strategy involves sampling the data space around an anomalous result (in terms of a similarity measure). Thus, the task of theory revision should be anomaly-driven not only when we find weaknesses in the model, but also when we analyze data to correct that model.

3.3 CONCLUSION

In this chapter, we examined the hypothesis that anomaly-driven theory revision would produce a larger number of acceptable exception rules than traditional methods. After introducing the idea of anomaly-focused subsets of data and describing some induction procedures, we explored this hypothesis using data from the medical domain. Although the results only weakly supported the hypothesis, we identified a few interesting findings. In particular, our results supported the approach of current symbolic learning systems. That is, the most fruitful subsets when using the method of difference correspond to the actual subset examined by rule and decision-tree learners. Additionally, we found that relatively few data are necessary for anomaly revision.

Regardless of which methods or subsets generate the revisions, we still wish to determine their acceptability. Since most of the generated revisions were unacceptable, we made little progress toward the identification of those that are acceptable and, in particular, defensible. Before determining the characteristics of defensible revisions, we explore whether the application of such revisions lead to more accurate models.

4.0 APPLYING ACCEPTABLE REVISIONS

As we saw in Chapter 3, anomaly-driven revision generators do not guarantee the acceptability of the revisions. Even though Kalpana ensures that its revisions are both rehabilitative and monotonic, several anomaly resolutions failed to be defensible. Regardless, we could apply the revisions indiscriminately, making an arbitrary selection if multiple exception rules repair the same anomaly. However, work by Pazzani and colleagues [44] indicates that domain experts are less likely to accept such models. Therefore we should consider applying only the acceptable revisions.

Suppose that we limit ourselves to acceptable revisions, how will such a limitation affect predictive accuracy? Intuitively, we expect the accuracy of the resulting models to be just as high or higher than comparable models that contain unacceptable revisions. Two assumptions support this intuition. First we assume that our domain knowledge better approximates the true state of the world than a random collection of generalizations. Second we assume that generalizations sharing terms and relationships with our background knowledge better represent truth than arbitrary relationships[1]. These are strong assumptions, so to confirm our intuition we answer the question empirically.

To determine the correctness of our intuition, we performed three experiments of increasing complexity. The first two experiments used synthetic data whereas the final experiment used data from the respiratory syndrome (RS) domain. The first experiment links defensibility to probability, suggesting that when one knows only the statistical characteristics of

[1]Here we invoke J.G. Frazer's first principle of magic: like produces like, or the Law of Similarity [19]. That is, if our revision looks like our domain knowledge, and our domain knowledge provides predictive accuracy, then our revision does as well. Of course scientists are not shamans, so we alter the fallacious claim that the revision *does* provide predictive accuracy to the less controversial claim that such a revision *is more likely to* provide accuracy.

a domain, defensibility and probability equate. The second experiment extends the first by adding an irrelevant attribute to the data, thereby slightly complicating the domain. The third experiment tests the experimental question on a real-world domain. For this, we use an expert's assessment of defensibility to determine how this property affects accuracy in the absence of complete domain knowledge.

4.1 EXPERIMENTS IN SYNTHETIC DOMAINS

4.1.1 The Average of Two Attributes

The first test uses synthetic data generated from a domain theory that is difficult to represent with propositional Horn clauses. The data consist of two attributes with integer values ranging between 1 and 100. We assigned a final class of *low*, *mid*, or *high* depending on the average of these two attributes. As shown in Figure 12, an average of less than or equal to 30 results in a classification of *low*, an average greater than 70 is considered *high*, and we labeled the rest of the values *mid*. For testing purposes, we created 30 data sets each containing 100 feature vectors, randomly selecting the values for the data from a uniform distribution and assigning the class using the described theory.

We created four models for the purposes of this experiment. The first model, shown on the bottom in Figure 12, serves as the flawed base upon which the other models are built, using only the first of the two attributes to classify the data. The symbolic values of the attribute employ the intervals defined for classification. The remaining models include revisions from Figure 13.[2] The highly defensible rules appear in the second, third, and fourth models, with the third model incorporating the moderately defensible revisions, and the fourth model including the four least defensible rules.

For this straightforward domain, where full knowledge is available, we judged defensibility based on the probability that a hypothesis would be true. Therefore, we know that the

[2]When adding revisions to the models in both of our synthetic domains, we ignore the monotonicity constraint. That is, we assume that all data that we have seen before support all of the revisions that we introduce.

True domain model

- If $(a_1 + a_2)/2 \leq 30$ THEN the *class* is *low*.
- If $30 < (a_1 + a_2)/2 \leq 70$ THEN the *class* is *mid*.
- If $70 < (a_1 + a_2)/2 \leq 100$ THEN the *class* is *high*.

B: Base model

- IF a_1 is *low*, THEN the *class* is *low*.
- IF a_1 is *mid*, THEN the *class* is *mid*.
- IF a_1 is *high*, THEN the *class* is *high*.

Figure 12: Actual and approximate domain models for the two-attribute, synthetic data set.

combination of a *low* value with a *high* value results in a classification of *mid*. That is, since the values are integers, the average of the lowest possible *low* value and the lowest possible *high* is 36. Checking that the highest values for both ranges also average to a *mid* result indicates the correctness of the two rules capitalizing on this knowledge. Therefore, we consider these rules to be highly defensible. This same information allows us to ignore rules where both attributes have *low* values but the consequent yields a *high* classification. Such occurrences have probabilities of zero and are therefore not defensible. Due to the impossibility of such cases, they can be safely ignored as they would not be induced from any noise-free data.

The other eight rules in Figure 13 require extra knowledge about the sampling process to assess their defensibilities. Since we generated the data uniformly at random, we know that a single attribute with the value *mid* increases the likelihood that the actual class will be *mid*. As an example, if the attributes have values of *mid* and *high*, the average can range between 51 and 85. Sampling the integer values uniformly at random means that 3/7 of the time we expect that the average will fall within the *high* range and 4/7 of the time we expect it to fall within the *mid* range. Therefore, the rule "IF a_1 is *mid* and a_2 is *high*, THEN the *class* is *mid*" is more defensible than when the same antecedent predicts a class of *high*.

HD: Revisions with High Defensibility

- IF a_1 is *low* and a_2 is *high*, THEN the *class* is *mid*.
- IF a_1 is *high* and a_2 is *low*, THEN the *class* is *mid*.

MR: Revisions with Moderate Defensibility

- IF a_1 is *low* and a_2 is *mid*, THEN the *class* is *mid*.
- IF a_1 is *mid* and a_2 is *low*, THEN the *class* is *mid*.
- IF a_1 is *mid* and a_2 is *high*, THEN the *class* is *mid*.
- IF a_1 is *high* and a_2 is *mid*, THEN the *class* is *mid*.

LR: Revisions with Low Defensibility

- IF a_1 is *low* and a_2 is *mid*, THEN the *class* is *low*.
- IF a_1 is *mid* and a_2 is *low*, THEN the *class* is *low*.
- IF a_1 is *mid* and a_2 is *high*, THEN the *class* is *high*.
- IF a_1 is *high* and a_2 is *mid*, THEN the *class* is *high*.

Figure 13: Revisions for the base model (Figure 12). Note that only two of the MR rules are necessary because the second rule in the base model gives the same classification as those that assign a class of *mid* when the first attribute is *mid*.

Table 9: The average number of anomalies resulting from each model.

Model	Mean Number of Anomalies (out of 100) (99% Confidence Interval)	Standard Deviation
B[a]	42.33 (40.31, 44.36)	4.03
BHD[b]	23.73 (21.48, 25.98)	4.47
BHD-MR[c]	17.13 (15.28, 18.98)	3.67
BHD-LR[d]	30.93 (27.87, 34.00)	6.10

[a] Base model.
[b] Base model with highly defensible revisions (HD).
[c] BHD with the moderately defensible revisions (HD + MR).
[d] BHD with the barely defensible revisions, but not the moderately defensible ones (HD + LR).

We applied our four models to the 30 data sets, determining predictive accuracy by dividing the number of nonanomalous data by the total size of the data set. Since each data set contains 100 elements, we instead report the number of anomalies, which indicates the error rate for each model when the data are sampled uniformly at random. To assess the performance of the models, we first calculated the mean number of anomalies produced by each model. We also calculated the difference in model performance, testing the significance using paired t-tests.

Before discussing the results, we qualify our findings by indicating the simplicity of the experiment. That is, when we assume that defensibility results from some understanding of the relative probabilities of events, then these findings lose some force. No one would argue that incorporating rules tied directly to more probable events would decrease our model's performance. However, this experiment was not designed to confirm a theoretically evident result, but to cast defensibility as an estimate of probability and thus make it easy to assess. Additionally this experiment elucidates the degree of improvement that can result from emphasizing defensible revisions.

Table 9 gives the average number of anomalies created by each model when applied to the data, the 99% confidence interval around that mean, and the standard deviation. Keeping

Table 10: The expected value of the difference in the number of anomalies between each pair of models (column − row) with 99% confidence intervals.[a]

Models	B	BHD	BHD-MR	BHD-LR
B	–	-18.60	-25.20	-11.40
		(-20.24, -16.96)	(-27.81, -22.59)	(-15.03, -7.77)
BHD	–	–	-6.60	7.20
			(-9.38,-3.82)	(4.35, 10.05)
BHD-MR	–	–	–	13.80
				(9.59, 18.01)

[a] All differences are significant with $p < 0.001$.

in mind that fewer anomalies indicates better performance, these results show that the base model (B) produces 42.33 anomalies on average (i.e., 42.33% of our data are anomalous). Adding the two highly defensible rules to the base model (BHD) reduced the number of anomalies by roughly 44% to 23.73. When we add the moderately defensible revisions to the improved model (BHD-MR), the number of anomalies decreases to 40% of the baseline. And, as expected, the addition of the barely defensible rules to BHD (BHD-LR) hurts the performance of BHD, reducing its improvement over B to 27%. In fact, most of this model's improvement over the baseline is attributable to the highly defensible rules.

Table 10 shows the expected value of the difference in the number of anomalies generated, which indicates how the addition of various revisions alters the model's performance, along with the endpoints of a 99% confidence interval around the value. These results indicate that BHD-LR should produce more anomalies than all but the base model. Comparing BHD-LR to BHD indicates the number of anomalies attributable to the barely defensible rules. Rounding up, we expect between 5 and 11 anomalies to result from choosing a revision that is not defensible. This increase almost mirrors the decrease in anomalies seen when we choose the moderately defensible rules. There are no surprises here, and again, we note that

- IF a_1 is *low*, THEN the *class* is *low*.
- IF a_1 is *mid*, THEN the *class* is *mid*.
- IF a_1 is *high*, THEN the *class* is *high*.
- IF a_1 is *low* and a_3 is *high*, THEN the *class* is *mid*.
- IF a_1 is *high* and a_3 is *low*, THEN the *class* is *mid*.
- IF a_1 is *low* and a_3 is *mid*, THEN the *class* is *mid*.
- IF a_1 is *mid* and a_3 is *low*, THEN the *class* is *mid*.
- IF a_1 is *mid* and a_3 is *high*, THEN the *class* is *mid*.
- IF a_1 is *high* and a_3 is *mid*, THEN the *class* is *mid*.

Figure 14: BHD3-LR: The base model with additional rules that rely erroneously upon the third, irrelevant attribute for classification.

this initial experiment primarily serves to tie defensibility to probability and to give us a context in which we can understand the remaining experiments.

4.1.2 The Introduction of an Irrelevant Attribute

The next experiment continues our examination of the link between probability and defensibility. For this experiment we expanded the data to contain three attributes. We assign the class by averaging the values of the first two attributes as with the prior experiment, which means that the third attribute is irrelevant. We generated values for the three attributes in 30 sets of 100 data each by selecting integer values uniformly at random between 1 and 100. As with the prior experiment, we used the model in Figure 12 to create the final classification.

This test employs four models, two of which come from the prior experiment. Model B (see Figure 12) again serves as the first model, and BHD-MR, which contains the moderately and highly defensible revisions from Figure 13, serves as the second. We retain BHD-MR for this experiment due to its performance during the earlier experiment, and since we used identical sampling and classification procedures for the new data, we expect BHD-MR to

perform identically well in this case. Figure 14 shows the third model (BHD3-LR), which resembles the second except that it uses the first and third attributes for classification. As a result, and since we know the true classifier, this model is the least defensible. However, BHD3-LR represents the model produced when we mistakenly apply incorrect revisions. To construct the final model (B-ALL), we take the union of the second and third models, which could occur if a system applied all possible revisions indiscriminately.

Since we have full knowledge of the domain for this experiment, defensibility once again comes from the probability that a revision will give a correct classification. We selected the defensible anomaly resolutions in BHD-MR because they improved the performance of the base model by the greatest amount in the prior experiment. In contrast, the revisions in BHD3-LR are obviously not defensible since they appeal to the information in an irrelevant attribute. However, if this attribute actually contributed to the class assignment in place of the second one, then these would be the most defensible revisions (i.e., for the same reason that the revisions in BHD-MR actually are the most defensible). So, supposing that the class was assigned as in the first experiment, but the attribute that was paired with the first was unknown, then the revisions added to both BHD-MR and BHD3-LR would be equally defensible. In addition, the revisions in BHD3-LR referring to the third attribute would be the most defensible ones within the selected expression of the domain.

After applying our models to the data sets, we once again used the reduction in the number of anomalies to judge performance. Table 11 gives the mean number of anomalies resulting from each model with a 99% confidence interval, as well as the standard deviation. As in the prior experiment, BHD-MR again reduces the number of anomalies by just under 60%. In comparison BHD3-LR yields an 18% improvement over B, and B-ALL reduces the number of anomalies that the base model produces by 33%.

We should note that the models cannot produce any internal conflicts. The unacceptable revisions, being more specific than all the rules in B, take precedence. Additionally, all the revisions predict the same outcome—*mid*. Therefore each newly introduced anomaly in B-ALL was a case where the revisions employing the irrelevant attribute predicted *mid* and the observed class was either *low* or *high*. So, although the revisions added to BHD3-LR

Table 11: The average number of anomalies resulting from each model.

Model	Mean Number of Anomalies (out of 100) (99% Confidence Interval)	Standard Deviation
B[a]	42.80 (40.27, 45.33)	5.03
BHP-MR[b]	17.17 (14.82, 19.51)	4.65
BHP3-LR[c]	34.90 (32.20, 37.60)	5.36
B-ALL[d]	28.73 (26.47, 30.99)	4.49

[a] Base model.
[b] Base model with the highly and moderately defensible revisions.
[c] Base model with the with revisions referring to the irrelevant attribute.
[d] Base model with revisions from BHP-MR and BHP3-LR.

resolved some anomalies produced by B, they also created new anomalies in the test sets. These new anomalies explain why B-ALL performs worse than BHD-MR.

Table 12 shows the expected value of the difference in the number of anomalies generated as well as the associated 99% confidence interval. As with the prior experiment, these results indicate how the addition of various revisions alters the model's performance. Here, BHD3-LR, though improved over the base model, produces the most anomalies when compared to both BHD-MR and BHD-ALL. In particular, we expect BHD-MR to produce 17.73% fewer anomalies on average than BHD3-LR, whereas BHD-ALL should produce 6.17% fewer.

That BHD3-LR should produce fewer anomalies (-7.9%) than B is an oddity. The reason for such an improvement is twofold. First, the majority of the data has the class *mid* due to uniform sampling of the attributes. Second, although we based the rules in BHD3-LR on an irrelevant attribute, they do make the prediction of *mid* more likely than when applying B. Thus we see that due to the effects of sampling bias we may select errant revisions. If we accept a revision that improves performance without regard for its defensibility, we could easily create a model similar to B-ALL or even BHD3-MR by overfitting the data. In such a case, we would carry forward rules unrelated to the true domain theory and could even overlook those rules closest to the truth.

Table 12: The expected value of the difference in the number of anomalies between each pair of models (column − row) with 99% confidence intervals.[a]

Models	B	BHD-MR	BHD3-LR	B-ALL
B	–	-25.63	-7.90	-14.07
		(-28.44, -22.83)	(-11.14, -4.65)	(-17.21, -10.9)
BHD-MR	–	–	17.73	11.57
			(15.04, 20.42)	(9.85, 13.28)
BHD3-LR	–	–	–	-6.17
				(-8.16, -4.18)

[a] All differences are significant with $p < 0.001$.

Following the form of the earlier experiment on plausibility, we have used a simple domain theory to identify some of the perils of ignoring defensibility during model revision. Primarily, we found that sampling bias can lead to the acceptance of incorrect, but well-performing, revisions. We are unsurprised with this result, but take it as a firm indicator that predictive accuracy should not be the sole measure used when evaluating a revision. Specifically, we can rely on wise judgments of defensibility to reduce the effects of both sample bias and, although we have not dealt with it, data noise.

4.2 AN EXPERIMENT IN A REAL-WORLD DOMAIN

To determine how well our results with synthetic data map to real, complex domains in which defensibility means more than "more probable," we explored the use of acceptability within the field of biomedical informatics. More specifically, we worked to identify a model of the RS domain introduced in Section 2.2. While the small sample sizes limit our ability to make effective claims of statistical significance, they can give an indication of what we might expect in future work.

To classify the data, we developed two models generated using the rule-learning program RL [46]. We separated the 190 training data described in Section 2.2.1 into training and test sets so that during each learning trial RL extracted rules from 80% of the data, using the other 20% to determine the effectiveness of the learned model. We created seven models with this method, giving each model a different inductive bias and a different data partition. We then presented the resulting rule sets to our domain expert, Dr. John Dowling, who judged the two presented in Figures 15 and 16 to be the most plausible.

We used the models selected by the domain expert to test our assumption that the application of defensible revisions improves predictive accuracy in contrast to the application of indefensible ones. In fact, we altered our conjecture slightly from the prior two experiments in that we required the revisions that we applied for this experiment to match all three acceptability criteria.[3] To create the revisions, Kalpana first classified the 190 training data using one of the models, passing any resulting anomalies to the revision generators described in Chapter 3. These generators created at least one revision for each incorrect classifier of each anomaly. Thus, if Kalpana classified the anomalous case as RS-*present*, but the case matched the first two rules in Figure 16, then the generators produced at least two revisions. We presented the resulting revisions, as shown in Figure 17, to Dr. Dowling for analysis along with the instructions given in Appendix C.

Therefore each revision was determined to be either defensible or not based on the opinion of an expert in infectious diseases[4]. The information about the difficulty of classification was collected to distinguish between easily classified revisions and those that are borderline, possibly requiring further assumptions on the part of the expert. In some cases, Dr. Dowling volunteered his own rationale for the ratings. For example, he labeled the second revision in Figure 17 plausible, and mentioned that it was a difficult case to decide. As an explanation of the difficulty, he wrote, "[I] must assume that the source of bleeding [hemoptysis, which specifically refers to bleeding from the respiratory tract] is not visible on x-ray." Thus he

[3]With the experiments on synthetic data, all but two of the revisions would lead to new anomalies, so we relaxed the monotonicity constraint.

[4]The instructions given to Dr. Dowling refer to the plausibility of rules as opposed to their defensibility. Due to the generality of the term "plausibility," we later chose to refine our terminology within this thesis to reflect the specificity of the concept we are attempting to understand.

- IF *chest tenderness* is *present*, THEN *RS* is *absent*.
- IF a *positive pulmonary edema x-ray* is *present*, THEN *RS* is *present*.
- IF a *positive pneumonia x-ray* is *present*, THEN *RS* is *present*.
- IF a *pneumonia diagnosis* is *present*, THEN *RS* is *present*.
- IF *rhonchi* are *present*, THEN *RS* is *present*.
- IF a *positive pleural effusion x-ray* is *present*, THEN *RS* is *present*.
- IF *sputum* is *absent* and *headache* is *present*, THEN *RS* is *absent*.
- IF *cough* is *absent* and *chest pain* is *present*, THEN *RS* is *absent*.
- IF *dyspnea* is *present* and *chest pain* is *absent*, THEN *RS* is *present*.
- IF *dyspnea* is *absent*, a *positive pneumonia x-ray* is *absent*, and *tachycardia* is *present*, THEN *RS* is *absent*.
- IF *dyspnea* is *absent*, a *positive pneumonia x-ray* is *absent*, and *oxygen desaturation* is *absent*, THEN *RS* is *absent*.
- IF *cough* is *absent*, *oxygen desaturation* is *absent*, and *rales/crackling* is *absent*, THEN *RS* is *absent*.
- IF a *positive pneumonia x-ray* is *absent*, and *oxygen desaturation* is *absent*, and *a positive pulmonary edema x-ray* is *absent*, THEN *RS* is *absent*.
- IF *wheezing* is *absent*, a *positive pneumonia x-ray* is *absent*, a *positive pulmonary edema x-ray* is *absent*, and a *positive pleural effusion x-ray* is *absent*, THEN *RS* is *absent*.

Figure 15: M4: A plausible model of respiratory syndrome (RS) extracted from training data.

- IF *chest tenderness* is *present*, THEN *RS* is *absent*.
- IF *headache* is *present*, THEN *RS* is *absent*.
- IF a *pulmonary edema-congestive heart failure diagnosis* is *present*, THEN *RS* is *present*.
- IF a *positive pulmonary edema x-ray* is *present*, THEN *RS* is *present*.
- IF a *positive pneumonia x-ray* is *present*, THEN *RS* is *present*.
- IF a *positive pleural effusion x-ray* is *present*, THEN *RS* is *present*.
- IF *wheezing* is *present*, THEN *RS* is *present*.
- IF *dyspnea* is *absent* and a *positive pneumonia x-ray* is *absent*, THEN *RS* is *absent*.
- IF a *positive pneumonia x-ray* is *absent*, *oxygen desaturation* is *absent*, and *chest pain* is *present*, THEN *RS* is *absent*.
- IF *sputum* is *absent*, a *positive pneumonia x-ray* is *absent*, and *oxygen desaturation* is *absent*, THEN *RS* is *absent*.
- IF *dyspnea* is *present*, *bronchitis* is *absent*, and *chest pain* is *absent*, THEN *RS* is *present*.
- IF *sputum* is *present*, a *positive pneumonia x-ray* is *absent*, a *positive pleural effusion x-ray* is *absent*, and *asthma* is *absent*, THEN *RS* is *absent*.

Figure 16: M5: A second plausible model of respiratory syndrome (RS) extracted from training data.

ID: 16
Original Rule:
(CHEST_PAIN is ABSENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)

Exception Rule:
(FLU_SYMPTOMS is PRESENT) and
(CHEST_PAIN is ABSENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)

Is the exception rule plausible?
Was this a difficult case to decide?

ID: 111
Original Rule:
(X_RAY_PULMONARY_EDEMA is ABSENT) and
(X_RAY_PNEUMONIA is ABSENT) and
(X_RAY_PLEURAL_EFFUSION is ABSENT) and
(WHEEZING is ABSENT)
implies (RESPIRATORY_SYNDROME is ABSENT)

Exception Rule:
(HEMOPTYSIS is PRESENT) and
(X_RAY_PULMONARY_EDEMA is ABSENT) and
(X_RAY_PNEUMONIA is ABSENT) and
(X_RAY_PLEURAL_EFFUSION is ABSENT) and
(WHEEZING is ABSENT)
implies (RESPIRATORY_SYNDROME is PRESENT)

Is the exception rule plausible?
Was this a difficult case to decide?

Figure 17: Two example revisions as presented to the domain expert.

Table 13: The performance of M4 and M5 before and after applying revisions.

Model	Revisions Added	Resolved Anomalies	New Anomalies	Total Anomalies
M4	None	0	19	19
(Figure 15)	All	3	5	21
	Defensible	9	1	11
	Indefensible	11	6	14
M5	None	0	20	20
(Figure 16)	All	12	1	9
	Defensible	7	0	13
	Indefensible	8	1	13

employed knowledge beyond the features and rules present within the model to ascertain the defensibility of the exception rules.

From each original model, we created three new models using the resulting revisions suggested by Kalpana and rated by the domain expert. As with the prior experiments, one model contained all the revisions to the base model, while the other two used either of the defensible or indefensible ones. We applied these three models, along with the associated original model, to the 92 cases originally set aside for testing purposes. The number of anomalies remaining within the test set determined each model's performance. A perfect theory should encounter no anomalies, so fewer anomalies indicates better performance.

Table 13 displays the results of applying the models to the test data. The first base model shown in Figure 15 produces 19 anomalies. Adding the indefensible rules resulted in 21 anomalies, of which 16 remained from the base model and 5 came from the new rules. The model revised with the defensible exception rules fared better with only 10 anomalies remaining from the base model. Additionally, the defensible revisions resulted in the addition of only one new anomaly. Incorporating all the anomaly resolutions into the original model left us with 14 anomalies, of which 8 came from the original rules.

The second base model shown in Figure 16 led to slightly different results. The original model yields 20 anomalies, while adding either the defensible or indefensible revisions reduces this number to 13. While the model containing indefensible revisions produces a new anomaly, that model also resolves one anomaly from the original 20 not resolved by the defensible revisions. Combining both sets of revisions with the original rules leaves nine anomalies, of which, only the one introduced by the indefensible revisions is not included in the original 20.

Although the small sample size of this data set and the examination of only two models precludes significance testing, we can interpret the results in light of our previous findings with synthetic data. As in those experiments, the addition of defensible revisions to the base model improves that model's performance. However, in one instance, the set of indefensible revisions appears to do just as well. Also, it seems that at times, those revisions enhance the defensible ones. Since the results differ based on the original model, we examined them more closely.

First, we consider M4, the model given in Figure 15. In this case, the defensible revisions performed much better than those not judged to be defensible. The defensible exception rules not only led to the resolution of more anomalies, but also created fewer new anomalous data. The poor performance of the unacceptable revisions carried over to the model containing all the rules. That is, although the unacceptable rules accounted for anomalies unresolved by the acceptable ones, the addition of the new anomalies nullified the benefit of combining all the revisions. These results aligned with both our intuition and the results found with the synthetic data. While we consider it somewhat interesting that the defensible revisions can lead to new anomalies, we are not surprised due to the complexity of the domain and the likely presence of noise within the data set.

Second, we examine M5, shown in Figure 16. The results in this experiment surprised us given our prior findings. While the defensible revisions did quite well, the unacceptable rules actually resolved more of the original model's anomalies. Here performance does not reflect the benefits of defensibility. However, by using the acceptable revisions, we avoided introducing rules that result in new anomalies. Thus we presumably incorporated fewer incorrect revisions than if we had selected the unacceptable rules. The same effect existed

when looking at the prior model, and we may ultimately find that choosing defensible revisions leads to fewer future revisions. That is, since defensibility helps us introduce fewer incorrect rules, a revision system need not spend time correcting for those rules.

Perhaps the most surprising result from this final experiment was that applying all the revisions improved performance over the application of just the defensible ones. This finding leads us to consider the secondary benefit of defensible revisions: justifiability. First, when no discernible difference in the performance of two models exists, we should select the one that "makes sense." This reasoning allows us to select among the models incorporating one or the other of the sets of revisions. That is, since the defensible anomaly resolutions lead to a model more consistent with background knowledge, we should select it. Second, we may find, as in this experiment, that adding all the revisions leads to the best performance; however, we must analyze the tradeoff between accuracy and this secondary characteristic.

The pattern, if it can be called such in this limited study, shows that the cautious approach is to accept only the defensible revisions. That is, when we added the new rules to Models M4 and M5, the defensible revisions created the fewest new anomalies and always improved the performance of the model. The indefensible rules tended to introduce more anomalies and in some cases decreased the model's performance. Even when the performance is increased, the indefensible revisions will contradict our domain knowledge and will force additional revisions. Currently, we leave the decision of which revisions to apply with the user, assuming that he will best know when to sacrifice understandability for accuracy.

4.3 CONCLUSION

The results of this chapter, though exploratory, indicate that defensible revisions at times improve predictive accuracy over indefensible ones. From an intuitive standpoint, defensible revisions will better fit with our knowledge, presumably giving us a more favorable opinion of the final model. Experimentally, it seems that defensible anomaly resolutions allow us to cautiously improve our model. That is, we can avoid the addition of incorrect knowledge. This finding is also intuitive given that when we consider a revision's defensibility, we esti-

mate its prior probability given a large network of unrepresented information. In the next chapter, we examine ways to identify defensible hypotheses so that we might capitalize on their advantages.

5.0 DEFENSIBILITY

In Chapter 3 we showed that an anomaly-driven approach leads to the production of acceptable revisions. We followed this by finding that in some situations, the identification and application of defensible revisions can increase predictive accuracy. We now turn our attention to the task of identifying defensible revisions automatically. That is, we wish to identify heuristics that will enable us to segregate the acceptable from the unacceptable. While we could use an estimate of the predictive accuracy of the rule, our experiments in Chapter 4 showed that it was not always the most desirable measure. Therefore, we introduce syntactic and semantic measures of defensibility guided by the virtues of hypotheses discussed by Quine and Ullian [47].

5.1 CONSERVATISM

5.1.1 What Is Conservatism

The conservatism of a hypothesis (in our case, a revision), according to Quine and Ullian [47], reflects the amount by which the hypothesis conflicts with current beliefs. These beliefs may consist of the collective knowledge of a scientific domain, one person's individual convictions, or the expectations of a committee. Regardless, some decision-making entity must assess the degree of conflict of the new belief with some combination of implicit or explicit knowledge and rule upon its acceptability. This degree of belief may involve a strict measure of conflict based upon violations of the law of the excluded middle, or as is more likely, it

could incorporate flexibility to allow for uncertainty. More interesting than an exact, formal definition of conservatism is how it can be and has been used in the process of altering beliefs.

Beginning with an example, suppose that we have, by our faith in the church, decided that all celestial bodies are perfect, crystalline spheres. After peering through our newly invented telescope, we notice that the Moon has what appear to be mountainous regions. One approach to resolving the anomaly involves claiming that no celestial bodies are perfect spheres. This hypothesis is minimally conservative in that it contradicts our original theory in every possible case. Alternatively we might claim that the mountainous objects are, in fact, an illusion created by the telescope. This statement fully conserves our original theory, but it may not conserve our theory of optics (if we have one).

Between the two extremes of reinterpreting the data and casting away our beliefs lie several revisions, each with its own level of conservatism with relationship to our beliefs. To explain our awkward observation of the Moon, we could claim that the Moon is a singular exception to our theory of celestial bodies. This new revision contradicts a small portion of our original theory, since all other celestial bodies remain perfect spheres. Additionally, we need not question the nature of observations produced by the telescope, therefore retaining most of our original theory without sacrificing our knowledge of optics.[1]

Although conservatism is often discussed in the context of a single theory, our example shows that an anomaly resolution impinges upon other beliefs as well. That is, contextual beliefs of which we are not always aware bind our thoughts. The complexity of our belief networks makes a complete analysis of conservatism difficult if not impossible. Even adjusting for those expectations specific to a single individual, we encounter assumptions such as visual acuity that are difficult to formalize but general enough to deserve attention.

In some circumstances, we can reduce the complexity of our belief network. For instance, we may make assumptions about the correctness of our data. If we consider the data to be perfect, then we bar from thought all anomaly resolutions questioning that data. Alternatively, we can associate a degree of belief with each datum. With this approach, revisions

[1]Note that naming specific individual objects as exceptions violates Goodman's principle that statements of a theory be lawlike [24].

that question the validity of a value begin with a level of conservatism corresponding to the degree of belief.

Simplifying assumptions, such as whether the data are perfect, serve the purpose of abstracting away some of the multitudinous beliefs that influence us. By finding the right level of abstraction, we can better approximate the conservatism of a hypothesis without resorting to a full belief-mapping of any particular individual. Thus, when calculating conservatism, we attempt to identify those beliefs relevant to the anomaly resolution, situated in the particular domain, and generalizable across domain experts. In effect, we limit the scope of the beliefs that must be checked against our revision.

After defining the scope of beliefs that may be revised, we can evaluate the conservatism of a particular revision. Returning to the earlier example, we considered the hypothesis that the mountains on the Moon were fabrications of the telescope. To evaluate the conservatism of this hypothesis, we need to consider the theory of optics as well as our own prior experiences with the same telescope. However, we can simplify our analysis of the hypothesis by assuming that the telescope conveys an accurate image of the Moon. Thus we reduce the difficulty of calculating conservatism by condensing a subnetwork of theories into a single, stated belief.

5.1.2 Conservatism's Role in Discovery

Before using conservatism to judge the defensibility of a revision, we should determine whether the measure serves any useful purpose. In apparent support, Kuhn [31, 32] states that the majority of scientific research consists of puzzle-solving work that brings an established theory closer to observed effects. He claims that this strict adherence to the current model, which embodies itself in the act of fine-tuning the model, enables the scientist to identify nontrivial anomalies. That is, "their recognition and evaluation ... depend upon a firm commitment to the contemporary scientific tradition.[32]"

Shapere [55] also notes the conservatism inherent in science, writing, "science builds on what it has found it can trust, what it has least specific reason to doubt and what it has found most broadly applicable." He continues to write that in the face of an unexpected observation, "we begin by suspecting those [ideas] which are, in light of our previous well-

founded beliefs, most likely to be at fault, and least costly to give up." Shapere's words are more descriptive than explanatory, but they are positioned in an essay espousing a progressive view of science. Additionally, he implies that progress results in part from this rough ordering in which beliefs scientists examine their beliefs. This order is, as mentioned above, guided by the likelihood of fault and the intellectual cost of modifying one's beliefs.

Although Kuhn's view of science as an arational activity contradicts Shapere's progressive view, both philosophers recognize that conservatism plays a primary role in the modification of theories. For Kuhn, the scientist mostly solves puzzles within the context of the prevailing theory. Conservatism holds sway until a mounting collection of anomalies forces scientists in the field to switch to the revolutionary mode of discovery. Kuhn claims that this mode leads to a radical change in the theoretical structure. However, he does not assert that the change will bring us closer to universal truth.

For Shapere, there is only one mode of science. In his model, the scientist's conservatism leads to a strategy of theory change based on progressive deepening. That is, using a tuning operation, scientists always resolve anomalies at the shallowest level possible. This constant tuning results in the removal of doubt about our understanding of the world. So, on the one hand, conservatism results in the gradual accumulation of anomalies, and on the other hand, it keeps us from altering our theory too drastically. In either case, conservation of beliefs is a core concept that enables theory change.

Kuhn and Shapere's models of science imply that conservatism leads to scientific discovery. However, discovery may occur in spite of conservatism instead of as its result. What, then, keeps us from approaching each anomaly as a challenge to the very core of our beliefs? Quine and Ullian [47] write that extreme conjectures yield more room for error and that when we make "a leap in the dark the likelihood of a happy landing is severely limited."[2] Additionally taking small steps helps us create a more detailed map of our surroundings. Thus we build a better understanding of how the anomaly affects our beliefs and how we might best respond.

[2]For Popper [45], extreme conjectures that are easier to refute become stronger as attempts to refute them fail. So in his view, bold conjectures make good science.

5.1.3 The Logic of Belief Revision and Conservatism

The theory of belief revision established by Alchourrón, Gärdenfors, and Makinson (the AGM theory) [2] relies on conservatism in the learning process. This theory attempts to capture the properties of operators that expand, retract, or revise a current set of beliefs. Expansion occurs when we incorporate a new belief that does not contradict any prior beliefs. Retraction involves the removal of a belief. And revision happens when we introduce a belief that conflicts with our current expectations. When conflict arises, we must alter our beliefs to maintain consistency.

When revision is necessary, there may be a large number of possible alterations. To guide selection, we appeal to epistemic entrenchment [20]. Epistemic entrenchment defines the properties of an ordering among beliefs that captures their interdependencies. Pagnucco's [43] overview summarizes the key aspects of these properties. Included is the dominance postulate that states that when one formula entails another, the first formula is less entrenched than the one entailed. That is, we cannot retract the entailed formula alone—the entailing formula must also be removed. Therefore, removing only the entailing formula results in less change to the original belief set.

This postulate apparently embodies conservatism. We are expected to prefer the removal of less entrenched beliefs—those beliefs that require fewer additional beliefs to be retracted. Thus we conserve more of our original belief set. This reflects what has been called "the hallmark of the AGM postulates" [13]—the general notion that we should emphasize conservatism during both retraction and revision.

This intuition has remained largely unchallenged in the AGM literature. However, Rott [51] argues that any dependence on conservatism is illusory. First, he claims that the logic of belief revision fails to express conservatism as traditionally understood. That is, on the one hand, no clear method for determining the "minimal mutilation" of a belief set exists when two revisions are not strict subsets. Thus, the best we can hope for is a partial ordering of revisions. Second, Rott asserts that the AGM theory actually allows us to keep a less entrenched belief over a more entrenched belief when a successful revision requires the

retraction of just one of the two. Regardless of these comments, Rott continues to appeal to conservatism [50, 52], although his context has changed.

5.1.4 Problems with Conservatism

While Rott limits his critique to the logical difficulties of conservatism, in practical use, we claim that such a measure is too coarsely grained. In the most straightforward case, a set of possible revisions may simply extend a theory. Thus we can append beliefs that do not contradict our theory. Conservatism has nothing interesting to say about these completely additive changes because each alteration is maximally conservative. In actuality, the measure's real weakness comes not from the lack of retraction, but from the equivalence in the measures of different revisions that all preserve conservatism.

Multiple revisions may be equally conservative even in the presence of a preference. For instance, consider adding the statement that Earth-like entities can have mountains to the original theory, which says that all celestial bodies are perfect spheres. Seeing mountains on the Moon, we can conjecture both that the Moon is an Earth-like entity and that large, yellow entities in the night sky have mountains. Either way, we contradict our prior beliefs. The first revision challenges the Moon's status as a celestial body, and the second revision creates a special class of celestial bodies. Both revisions can be viewed as equally conservative since they only alter our perception of the moon. However, intuitively we prefer the first over the second.

As a more mundane example, consider the theory that all basketballs bounce. Now add the belief that no flat objects bounce. If we then see a small, flat basketball fail to bounce, we may entertain two equally conservative conjectures: that all non-flat basketballs bounce and that all non-small basketballs bounce. Here the first conjecture is preferable in the context of our theory, but again we cannot express the difference in terms of conservation of beliefs. Thus conservatism may not be detailed enough to fully measure a revision's acceptability.

Another indication that conservatism does not solely influence the acceptability of a revision comes from Kuhn's concept of essential tension [32]. Conservatism can be the enemy of discovery, and a scientist must know when to sacrifice an explanation that retains prior

beliefs in place of a revolutionary conjecture. After years of investigation, Kepler ceased adding circles to the Copernican view of the solar system [30]. Instead, he replaced the complex orbits of the planets with simpler, elliptical paths. Thus, while Kepler could have continued building up the theory that orbits consist of various circles, thereby conserving prior beliefs, he chose to pursue a hypothesis that rewrote centuries of misconceptions. This example alone indicates that conservatism, at least in science, is not the only measure used when selecting among competing hypotheses. To address these problems, we once again turn to *The Web of Belief* [47], and note that other virtues of acceptable hypotheses interact with conservatism.

5.2 MODESTY

After discussing conservatism, Quine and Ullian suggest modesty as a virtue of hypotheses, presenting it in two forms. On one hand, they base modesty in logical implication. That is, given two hypotheses X and Y such that $X \rightarrow Y$, Y is the more modest of the two. On the other hand, the authors relate modesty to familiarity. Thus suppose someone walks outside to collect his newspaper, only to find it missing. He might entertain the belief that the newspaper was never delivered or he might make the less modest conjecture that the editor-in-chief refused him service. Both this interpretation of modesty as familiarity as well as its relationship to logical implication address the weaknesses of conservatism.

5.2.1 Modesty via logical implication

When describing the logical interpretation of modesty, Quine and Ullian provide a single, opaque example that deserves elaboration: "A hypothesis A is more modest than A and B as a joint hypothesis.[47]" Compare the statements "The Moon is a perfect, crystalline sphere" and "The Moon is spherical in shape." Here, the former statement implies the latter, thus the latter is the more modest of the two. Figure 18 gives a graphical representation of the relationship. The circle labeled B denotes the various situations where B can be true.

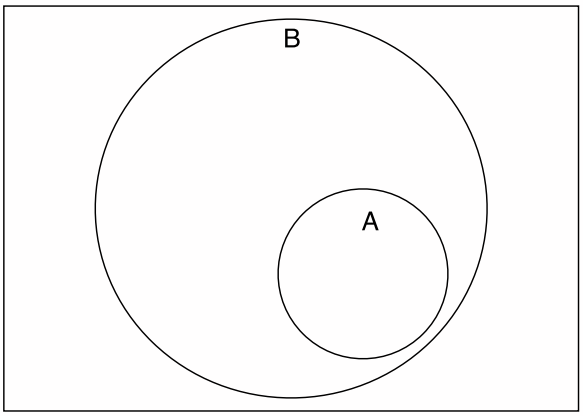

Figure 18: The relationship between two statements A and B when A implies B. According to Quine and Ullian [47], B is the more modest of the two since it is true in at least the same situations where A holds and may be true in others.

The smaller circle, A, denotes those cases where A can be true. This relationship illustrates that the consequent must occur at least as frequently as the antecedent. The consequent may obtain in cases where the antecedent does not, but not vice versa. Therefore the act of asserting the consequent is more modest because the consequent is more often true.

While this approach to modesty seems reasonable, its use in the assessment of the acceptability of a hypothesis requires caution. Suppose we have decided to take the modest route and claim that the Moon is spherical in shape. We can instantly create a more modest hypothesis via disjunction. For example, the conjecture "The Moon is spherical in shape or the Moon is made of green cheese," is logically weaker than the first disjunct alone ($A \rightarrow (A \vee B)$). In fact, we can continue to increase the modesty of our hypothesis by multiplying the disjuncts with arbitrary claims (e.g., "or life is like a teapot"). Such a shortcut to acceptability degrades the utility of modesty.

5.2.2 Modesty as Familiarity

The logical approach to modesty fails because it lacks a method for selecting meaningful disjuncts. As a result, we introduce a semantic method such that the more familiar the

events or terms within a conjecture, the more modest that conjecture. This approach yields a more experiential version of modesty than that of logical implication. In particular, modest revisions will refer to events or properties that have been noted to occur frequently as opposed to those that are unique or rare. In the same vein, the use of familiar terms or concepts just seems more modest than appealing to newly created terms such as "grue."

Grueness was used by Nelson Goodman [24] in his classic work on the familiarity of terms to illustrate the concept of entrenchment (not to be confused with epistemic entrenchment, which refers specifically to held beliefs). Entrenched terms are those that are more often used to describe a particular domain. For example, in the domain of physics, 'mass,' 'gravity,' and 'charge' are well entrenched terms while 'bitter' and 'cagey' are not. Goodman argued that the entrenchment of a term indicates its projectibility, where projectibility describes how useful a term is when predicting future events. A term's entrenchment can be measured and may be equivalent to its familiarity.

The ideas of entrenchment and projectibility are situated within a larger argument presented in *Fact, Fiction, and Forecast*[24]. Here, Goodman replaces Hume's riddle of induction [27] with a new one. This new riddle can be summarized as follows. Given some observation that we wish to generalize, we must choose those observed features that seem best suited for this purpose. What is the nature of this process of selection, and when do we find generalization useful? Goodman proceeds to argue that there is no simple answer to these questions. However, he does present a reasonable solution for choosing features for use in a new hypothesis. His method requires the recollection of those features used in prior hypotheses. That is, despite the true utility of a feature, when creating new hypotheses, we tend to (and should) favor those features that have been used in the past. Such a hypothesis possesses modesty because its included terms will not add new, untested concepts to the original theory.

The familiarity of terms extends to the familiarity of events. To illustrate the interaction of familiar events with the modesty of a hypothesis, suppose that this winter the local news reported an increase in cases of respiratory syndrome (RS) within the country. The various strains of the influenza virus are common causes of RS. Since an increase in influenza typically occurs during the winter months, the hypothesis "The rise of RS is due to the normal temporal cycle of influenza outbreaks" is highly modest. In comparison, the hypothesis "The rise of

RS is due to the spread of anthrax by a bioterrorist" is riskier. The latter hypothesis hinges on an event that is both far less frequent than that described by the former and relatively less familiar.

The modesty of the RS hypotheses also relates to the number of assumptions that they require. The former hypothesis asks that we assume that the past rise in influenza infection serves as evidence to the current rise. The latter requires us to hold beliefs about an individual with malicious intent able to gain access to a controlled substance and distribute it successfully. Without any further evidence than the news report, the virtue of modesty tells us that the first hypothesis is more plausible.

An additional aspect of modesty related to familiar events involves the activity required by the hypothesis. That is, in our example, the cycling of influenza outbreaks appears to occur naturally without the need for explicit action. On the contrary, the malicious spread of an infectious agent presumably requires an intentional act and substantial effort. While the latter hypothesis may reflect the truth of the situation, "the counsel of modesty [is] that the lazy world is the likely world.[47]"

5.2.3 The Weakness of Modesty

In summary, logically weaker hypotheses are preferable because they likely refer to more events than their antecedents. In addition, hypotheses using well-entrenched terms are preferable to those that do not. That is, referring to events and concepts expected in situations similar to the current context yields a more acceptable hypothesis. While the justification for the use of entrenched terms may appear unconvincing, it is, itself, based upon the principle of induction—namely, it has worked in the past.

Given this new measure of defensibility, recall that we introduced modesty as a criterion that would allow us to choose between two hypotheses that equally conserve the original theory. Returning to the example theory about mountains on the moon, remember that we had difficulty choosing between the two revisions "The Moon is an Earth-like entity" and "Large, yellow entities in the night sky have mountains." Conservatism provided no help. However, modesty comes to the rescue. The concept of being Earth-like (as in "celestial

bodies are not Earth-like") is more familiar in this context than that of being large and yellow. While other celestial bodies may be large and yellow, we do not expect that to be an important, differentiating (i.e., projectible) characteristic. Thus we accept the first revision.

Suppose that we had no entrenched terms for describing the variation of the moon. For instance, we may have to choose between the revisions "Yellow entities in the night sky have mountains" and "Large objects in the night sky have mountains." Should we form a disjunction of these two revisions creating a more modest conjecture? Possibly, but why not add "Objects made of green cheese have mountains" to the disjunction as well? After all, the resulting statement would be logically weaker. To combat this peculiarity, we will introduce a third virtue of hypotheses: simplicity.

5.3 SIMPLICITY

Simplicity has attracted much attention, yet it remains poorly understood. Originally, it was thought to be a quality of nature[3]. However current opinion ascribes simplicity to theories and hypotheses themselves[4]. Since this shift of attribution, numerous authors have worked to define simplicity [22, 24, 45, 53, 56], but a formal description remains evasive. Most recently, researchers have appealed to computational theory and Kolmogorov complexity for a solution [18, 37, 60], but these formalisms fail to solve the fundamental problems addressed most notably by Goodman [23] (see Section 5.3.3) . Guided by the description given in [47], we sidestep the difficulty of establishing a complete definition and instead show how even incomplete notions of simplicity, when mediated by modesty, can be useful when assessing hypothesis plausibility.

[3]Some still hold this belief [63].

[4]Interestingly, this ontological shift is in itself a sacrifice of conservatism for simplicity. While it may be the case that nature is, in fact, simple in design, the process of identifying such simplicity on all fronts is far more daunting than the task of showing an initial psychological preference for ideas of limited complexity as in [40].

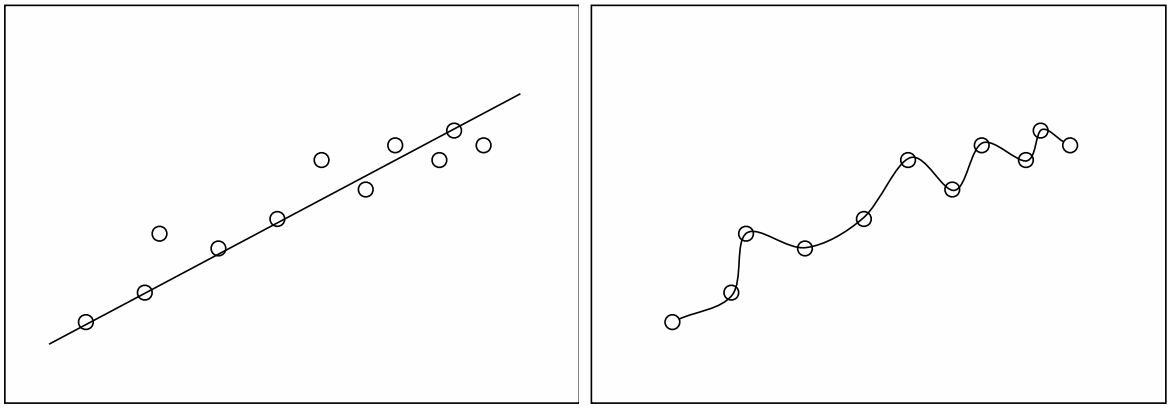

Figure 19: In Karl Popper's view [45], the line in the right image is simpler than that in the left due to its high falsifiability. In contrast, Herb Simon emphasized [56] that the line in the left provides a simpler model of the data.

5.3.1 Simplicity of Equations

In [47] Quine and Ullian do not so much define simplicity as they explore various descriptions that have arisen in the philosophy of science. First, they describe simplicity in terms of fitting a curve to data. For example, given observations plotted on a Cartesian plane, the curve that fits the data reasonably well while having the least amount of curvature is considered the simplest generalization. While not the first to take this perspective, Karl Popper [45] expounded on it and justified his approach through an appeal to falsifiability. Here, the more probable the hypothesis, the less falsifiable it is. Thus, since $y = ax$ describes a wider range of cases than $y = 3x$, the latter is simpler and should be preferred as it is therefore more falsifiable.

Simon [56] revisits Popper's work, drawing a distinction between the simplicity of a hypothesis about data and a hypothesis about the world. Popper made the claim that simpler hypotheses are less probable. In Simon's view, this stance makes sense only when generalizing to future instances. When creating hypotheses about data under examination, the plot on the left in Figure 19 is both simpler and more probable *within the context of those data*. As Simon mentions, there is no conflict among the separate claims of simplicity

attributed to these situations. Therefore, in both cases, a preference for simplicity appears reasonable, but as Popper mentions (referring to [61] and [64]), there are no established "*logical or epistemological advantages* the simpler law is supposed to possess, compared with one that is more complex.[45]"

5.3.2 Simplicity of Programs

Before continuing our examination of Quine and Ullian's treatment of simplicity, we need to look closer at this virtue in the context of computer programs. Simon's informal treatment [56] presents two programs that are able to generate data that are consistent with a particular theory. The first program is stochastic, requires two parameters, and was implemented in 15 program statements. The second program requires no parameters, is deterministic, and was implemented in 27 statements. Simon asserts that the latter program, regardless of the parsimony of the former, is simpler. Here, he again sides with Popper, stating that the number of parameters is the real determiner of simplicity, not the length of the program.

In contrast to Simon's claim, some proponents of Kolmogorov complexity insists that the shortest program that will generate the data is the simplest [37, 54, 60]. In this context, given some initial data, we identify the programs in some universal language that output that data. These programs (i.e., hypotheses) are given a prior probability based on their length such that shorter programs are considered more probable. In the case that several programs of the same length are capable of producing the observed data, we are enjoined to keep all such hypotheses on the table. As for the introduction of parameters, Bosch [60] suggests that we should sum the probabilities of each possible instantiation of the program where the parameter is specified.

Consolidating the perspectives of Simon, Popper, and the Kolmogorov pundits with respect to program length seems daunting at best. While Simon and Popper agree that the number of parameters increases the complexity of the program,[5] Popper's definition of simplicity is equivalent to falsifiability. Simon acknowledges such a connection, but he refuses to

[5]Note that a vector or matrix may be considered a single parameter, thereby confounding the dependency on the number of parameters as a measure of simplicity. Nelson Goodman gives the most thorough treatment of this problem, which we discuss in Section 5.3.3.

refine simplicity beyond an intuitive notion related to the degrees of freedom in a hypothesis. Additionally, Simon allows for simplicity to result in an increase of prior probability when the hypothesis has been extracted from data. This view corresponds to the use of Kolmogorov complexity. That is, programs discussed in the context of Kolmogorov complexity are always extracted from the data, and simplicity directly correlates with probability. Simon criticized Popper and others for not treating this method of hypothesis generation, claiming that they assert that "hypotheses spring full-blown from the head of Zeus.[56]" With this distinction in mind, Simon elaborated his position, writing that Popper's notion of probability refers to the actual state of nature described by a hypothesis and not to the hypothesis itself.

5.3.3 Problems of Simplicity

Taking Simon's perspective, we ignore Popper's attribution of probability for our purposes, but further problems lie ahead. If we are to employ the Kolmogorov measure to capture simplicity, we must decide how to interpret a program that accepts parameters. If we were to sum the probabilities of the program such that every parameter is fully specified in all possible ways, functions with free variables would be more probable. This method captures Popperian simplicity, unfortunately there seems to be no intuitive transformation that would convert the value into the measure of simplicity desired by Simon. Even if such a transformation were found, Kolmogorov complexity presents another stumbling block.

A basic assumption of all simplicity measures based on Kolmogorov complexity is that the set of programs is enumerable. If our programs can employ real numbers as constants or parameters, then such an assumption is too strong. We might claim that in general, the precision of our parameters or constants is limited by the precision of our measurements. So, for any number, we can establish a fixed precision making our programs enumerable again. As a counterexample, consider a program that takes the area of a circle as a parameter. Here our level of precision is limited by two factors: our ability to measure the radius of the circle, and the number of digits of π that we use. While we may not be able to improve our rulers, we can constantly add a more exact approximation of π. Thus measurement is not a limiting factor to the level of precision we may achieve. The point here is that

requiring our set of programs to be enumerable imposes limits on the programs that we can consider, and these limits are fixed by neither logical nor natural concerns. That is not to say that Kolmogorov-based measures of simplicity are useless, but they contain (beyond their selection of representation) a component of arbitrariness that needs recognition when the measures are used.

Finally, whenever we discuss simplicity with relation to the number or form of parameters of some formalized sentence, we must acknowledge the problem addressed by Goodman. In [23], he writes, "We can always, by a calculated selection of vocabulary, translate any hypothesis into one of minimal length." That is, our terminology can mask the complexity of our statements. Goodman continues to write, "To reject unfamiliar predicates wholesale in favor of familiar ones would be to disallow the introduction of needed new terms into scientific language." So he enjoins us to be cautious in our description of the world, but he gives no logical or well-defined guide when to posit new terms. At best, he might suggest that simpler explanations favor entrenched predicates (those in common usage), though we may introduce projectible traits (those apparently good for induction) if they can lead to a general simplification of a theory as a whole.[6]

5.3.4 The Subjectivity of Simplicity

Beyond their discussion of simplicity as related to the order of an equation, Quine and Ullian state that simplicity is, more than conservatism and modesty, a matter of personal preference. As Kuhn claims [30], Copernicus developed his model of the solar system because of his belief that the Ptolemaic model failed to satisfy his Neoplatonic notion of harmony. Although Copernicus's model failed to be either substantially simpler or more accurate than the Ptolemaic, he was driven by his beliefs about the simplicity of the circle to reformulate cosmological theory.

Since simplicity possesses such a subjective flair and there is little to suggest a deep relationship between simplicity and hypotheses, we have chosen to adopt a straightforward

[6]Compare Goodman's advice to Einstein's statement, "It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.[16]"

interpretation. For our purposes, the more easily testable an anomaly resolution, the simpler it is. While the vagueness of "easier" in this context may rightfully cause unease, we promise to quantify it once we discuss the details of our implementation. Naively, we might say that the simpler revision requires fewer observations than its alternatives. This claim fails to address why hypotheses referring to density appear simpler than those requiring measurement of mass and volume, but it gives us a foundation for developing a solution.

With our, admittedly vague, restatement of simplicity in mind, we now address how simplicity balances modesty's affinity for disjunctions. We mentioned that, when choosing among the anomaly resolutions A and $A \vee B$, modesty would have us prefer the latter. The problem is that the formal definition of modesty would have us multiply disjuncts indefinitely to improve the acceptability of our revision. Assuming that each disjunct describes an observable feature, we can call on simplicity to limit the ultimate size of the conjecture. Thereby we avoid giving preference to the infinitely long revision.

5.4 THE IMPLEMENTATION OF DEFENSIBILITY

Having explored possible components of a defensibility measure and discussed their interactions, we now describe their implementation within the context of Kalpana. To investigate the benefit of external domain knowledge (i.e., that not contained within the model), we developed both syntactic and semantic measures. The syntactic measures use only the explicit model of the domain. Within Kalpana this means that these measures analyze the given rule set. The semantic measures use domain knowledge acquired from an expert in the field relating to the features in the data set. For example, we asked our expert to estimate feature entrenchment using his full knowledge of the domain of which the model is a small component. This information was gathered with our understanding of conservatism, modesty, and simplicity in mind.

- IF *cough* is *present*, THEN *RS* is *present*.
- IF *wheezing* is *present*, THEN *RS* is *present*.
- IF *sputum* is *present*, THEN *RS* is *present*.
- IF a *pneumonia x-ray* is *positive*, THEN *RS* is *present*.
- IF *dyspnea* is *present*, THEN *RS* is *present*.

Figure 20: M1: Overly general model of respiratory syndrome (RS).

5.4.1 Conservatism

We begin the description of our implementation with conservatism. We gave an abstract definition of this concept in Section 5.1, stating that conservatism is the amount by which a hypothesis conflicts with current beliefs. For Kalpana in its uninformed state, the current beliefs consist of the given model. The hypothesis is the generated anomaly resolution. Since for Kalpana a revision must be an exception rule, it already conflicts with one rule in the model. Additionally, the revision may be an exception rule (either intentionally or unintentionally) to other rules. For instance, consider the basic model of respiratory syndrome (RS) given in Figure 20. The revision "IF *sputum* is *present* and *cough* is *present* THEN *RS* is *absent*" conflicts with both the first and third rules. In this case, we would want our measure of conservatism to reflect the full degree of syntactic conflict to accurately capture the defensibility of the revision.

To measure conservatism, Kalpana counts the number of rules that a particular revision specializes. In the prior example the revision directly specializes two rules. To ensure that Kalpana's measure of conservatism corresponds with our intuitive notion of the concept, we can chose to normalize the raw score by dividing by the total number of rules in the model and subtract the result from one. Thus conservative revisions have scores closer to one. To illustrate, the model in Figure 20 contains five rules, two of which conflict with the exception rule from the previous paragraph. As a result, that exception rule's conservatism equals

0.6 (three-fifths of the rules in the theory do not conflict with the explanation). While this measure seems reasonable, it can weaken in utility as the size of the model increases.

The limitations of Kalpana's uninformed approach to conservatism become evident when the data contain more features than represented in the model. Within the data for RS, there exist 65 binary attributes (giving a total of 132 features) plus the binary target class (*RS* is *present* or *absent*). The model in Figure 20 refers to five of the 132 total features and cannot indicate the conservatism of a revision based on any of the other features. Additionally, given a rule with multiple features in the antecedent, we are limited to evaluating conservatism of revisions that contain all of the original rule's features. For example, the rule "IF *cough* is *present* and *congestion* is *present*, THEN *RS* is *present*" can only be used to determine the conservatism of revisions matching the form "IF *cough* is *present* and *congestion* is *present* and (some conjunction of features), THEN *RS* is *absent*." To increase the utility of conservatism, we must know something more about all features, regardless of their presence in the model.

We designed the semantic measure of conservatism to address the limitations of the syntactic one. To this end, we encoded knowledge consisting of suspected relationships between each feature and each target class. For example, the feature "*cough* is *present*" supports the outcome, "*RS* is *present*." Using this information, Kalpana assigns a score of conservatism by examining all the features in the rule. When a particular feature present in the revision supports a contradictory outcome, then the program lowers the conservatism score of that revision. An advantage to this approach is that the system can now reason about individual features added to a revision as opposed to considering the revision as an indivisible whole. That is, while the syntactic measure compares rules to other rules, the semantic measure compares features to knowledge.

As an example, consider the relationship between chills and RS. The presence of chills is positively associated with the presence of RS, however this link is not evident from the model in Figure 20. With this new information, Kalpana can determine that "IF *cough* is *present* and *chills* are *present*, THEN *RS* is *present*" conserves prior beliefs, while the hypothesis "IF *cough* is *present* and *chills* are *present*, THEN *RS* is *absent*" is contentious. That is, using

the presence of both cough and chills to support the absence of RS when both features give evidence for RS's presence makes the second revision indefensible.

Although in the above example, we are able to reason about the feature "*chills* are *present*," we cannot assume the inverse. That is, knowing that the presence of chills is associated with the presence of RS tells us little or nothing about how the absence of chills is associated. So the revision "IF *cough* is *present* and *chills* are *absent*, THEN *RS* is *present*" cannot be judged based on our known relationship between chills and RS. This limitation is a property of the domain, and may not apply in some domains. More specifically, the absence of a symptom in general does not contain the same amount of information as its presence [7].

Kalpana's implementation of the semantic measure of conservatism penalizes revisions that violate explicit relationships. Through a conversation with a domain expert, we were able to assign 63 of the 65 attributes to two categories (positive diagnoses of bronchitis or musculoskeletal chest pain were considered ambiguous). The expert placed all attributes that, when positive, indicate the presence of RS into the first category and all attributes that, when positive, indicate the absence of RS into the second category. Thus of the 130 features, we have extra information about 63 for which the specified attribute is present. Appendix D.1 lists these categories.

Kalpana calculates semantic conservatism as follows. Given a revision, and the base rule from which that revision was created, the program examines the newly added features, checking a single condition for each feature in the revision and not in the base rule. Does that feature belong to a category that asserts a conflicting consequence? To demonstrate, let the original rule be "IF *cough* is *present*, THEN *RS* is *present*," and let the anomaly resolution be "IF *cough* is *present* and *chills* are *present*, THEN *RS* is *absent*." When evaluating the conservatism of this revision, Kalpana looks at the new feature, "*chills* are *present*," and determines whether the use of this feature conforms to prior beliefs. In this case, the added feature contradicts the consequence of the explanation. That is, the presence of chills gives

[7] The underlying issue is the paradox of confirmation [25]. We address the paradox by assuming that positive confirmation is necessary in medicine (and other domains), not just the absence of negative evidence. The semantics of each domain determine which relationships are positive.

- IF *cough* is *present*, THEN *RS* is *present*.
- IF *cough* is *present* and *lung tumor diagnosis* is *present*, THEN *RS* is *absent*.
- IF *dyspnea* is *present*, THEN *RS* is *present*.

Figure 21: M6: An incomplete, hypothetical model of respiratory syndrome (RS)

positive support for RS. Since one of the new features of the rule violates our current beliefs, Kalpana reduces the conservatism of the hypothesis by two.

5.4.2 Modesty

As mentioned in Section 5.1.4, limitations arise when conservatism is used to calculate defensibility. For example, depending on our background knowledge, a revision that refers to the presence of pneumonia can be just as conservative as one that refers to the color of the patient's hair. Additionally conservatism confers penalties, but cannot award praise. That is, the measure leads us away from the absurd (e.g., the presence of pneumonia indicates the absence of RS), but it fails to direct us toward the defensible. To address these limitations, we developed syntactic and semantic measures of modesty. Though we mentioned two characterizations of modesty in Section 5.2, modesty via implication and modesty via familiarity, both of our measures were designed with the latter perspective in mind. That is, those features more strongly associated with a particular class are deemed more valuable.

To gauge the modesty of an exception rule based solely on syntax, Kalpana examines each new feature in the revision and evaluates that feature's use within the model. If the feature exists within the theory, then the system calculates the proportion of the rules in which the feature supports the revision's consequent to all rules in which the feature appears. For instance, consider the model in Figure 21. The revision "IF *dyspnea* is *present* and *cough* is *present*, THEN *RS* is *absent*" is less modest than "IF *dyspnea* is *present* and *a lung tumor diagnosis* is *present*, THEN *RS* is *absent*." The feature "*cough* is *present*" is associated with the absence of RS in half of the rules in which it appears, while the feature "*lung tumor*

diagnosis is *present*" indicates RS's absence in all rules where it exists. Kalpana divides the number of statements that use the new feature to predict the same consequent by the total number of statements that contain the feature regardless of predicted class. So, more discriminatory features are judged to be more modest[8].

When a feature does not appear in the model, Kalpana calculates the proportion of rules containing the feature's attribute to the total number of rules in the model. Thereby the program expresses a preference for rules that use attributes commonly associated with the target attributes. This method of determining modesty penalizes the introduction of new features since Kalpana divides the number of the attribute's occurrences by the total number of rules in the model. In the extreme case where the attribute as well as the feature is new to the model, Kalpana will assign a modesty value of zero. Thus the system prefers to introduce features in a familiar context followed by well recognized attributes before we attempt to introduce entirely new terms (i.e., either attributes or features not present within the model).

After assessing the modesty of each feature in a revision, Kalpana calculates the total modesty of the rule using the mean modesty of the features. This approach introduces a weakness in the measure. Specifically, a feature or attribute that is new to the model can substantially decrease the overall modesty of the revision. This characteristic of the measure contradicts some methods of scientific discovery. For example, a scientist may include several features in a data set, suspecting them to be beneficial to the classification task although they are unexpressed in the model. That is, each attribute possesses some degree of modesty not accounted for in the syntactical structure of the model, which may lead Kalpana's purely syntactical measure to underestimate the modesty of a revision.

When designing Kalpana's semantic measure of modesty, we addressed both the limitation of the syntactical measure of modesty and the weakness of the semantic measure of conservatism. The domain knowledge that we chose for this measure relates to the amount of information carried by particular attributes. That is, the presence of pneumonia yields more information about the presence of RS than the color of the patient's hair. While the semantic conservatism measure could tell us which of the presence or absence of pneumonia

[8]We assume that those features that better discriminate between classes are more often used when discussing the revision's consequent.

(or which color of hair) most strongly indicates the presence of RS, all the features were assumed to convey an equal amount of defensibility. In contrast, modesty gives the means to decide which features better indicate a particular classification. This ability helps Kalpana more accurately quantify defensibility by assuming that the entrenched, or more familiar, terms are more likely to reoccur in defensible revisions.

As an illustration of this measure, consider the two attributes headache and cough. The presence of a headache is associated with the absence of RS, while the presence of a cough is associated with the presence of RS. The semantic measure of conservatism uses this information to ensure that the relationships within a revision correspond to domain knowledge. Thus the two anomaly resolutions "IF *headache* is *present* and *dyspnea* is *present*, THEN *RS* is *absent*" and "IF *cough* is *present* and *dyspnea* is *present*, THEN *RS* is *present*" have the same defensibility. However, in contrast to the encoded knowledge, headaches may occur as a symptom of RS. Headaches appear in numerous situations, but since they are also part of a general response to an infection, their presence alone gives very little evidence toward ruling out RS. Cough, on the other hand, typically specifies a respiratory ailment which may be within the lower respiratory system. Thus, the attribute "cough" carries more explanatory weight than headache.

Kalpana rewards revisions for using these more valuable attributes and penalizes revisions that use attributes that are less valuable. Our expert in infectious diseases chose, from the 65 original attributes, a subset of 44 attributes that he considered more valuable for the current task. His selection was made within the context of identifying patients with RS. Appendix D.2 lists these attributes. As with informed conservatism, the informed modesty measure examines those features new to the revision, altering its defensibility based upon the attribute's presence in the list.

Kalpana's specific implementation of the semantic measure adds two points to the revision's modesty score for each valuable attribute and subtracts one point for an invaluable attribute. As an example, let the base rule be "IF *cough* is *present* THEN *RS* is *present*." Suppose the following two rules are created to resolve a particular anomaly. The first, "IF *cough* is *present* and *chest tenderness* is *present*, THEN *RS* is *absent*," uses an attribute listed as valuable. The second, "IF *cough* is *present* and *chest pain* is *present*, THEN *RS* is

absent," uses a less valuable attribute. The added feature in both rules is conservative, so each rule has a base defensibility of 0. The modesty of the former rule is 2, while that of the latter rule is -1. By these scores, Kalpana considers the first rule to be the more defensible.

5.4.3 Simplicity

Simplicity counterbalances the aberrant behavior of overly specific rules. Given our current measure of defensibility, conservatism would allow us to add as many features as we wish so long as they come from the correct set. Also, at least in the case of the semantic measure, modesty rewards a similar action. In contrast, a preference for comprehensible models requires that the set of rules composing a model possess a low average number of features per antecedent. Our syntactic measure takes this approach, while the semantic measure tries to minimize the difficulty of matching the rule's antecedent.

We adopted a heuristic used by general-to-specific rule learners, such as RL [46], for our syntactic measure of simplicity. Specifically, the measure favors revisions with the fewest new features ϕ in their antecedents using

$$\sqrt{\frac{1}{\phi}}$$

as a score for each revision (note that ϕ will never equal 0 since a revision must always have at least one additional feature). As an example, consider the revision, "IF *cough* is *present* and *dyspnea* is *absent* and *wheezing* is *absent*, THEN *RS* is *absent*," to the base rule, "IF *cough* is *present*, THEN *RS* is *present*." The revision has two more features in its antecedent then the base rule and would thus have a simplicity score of 0.71. Although this measure has intuitive appeal and helps protect against overfitting data, it treats all features as equal. In some cases we may wish to rank the simplicity of the individual features according to domain knowledge, thereby fine-tuning the measure's effect.

Though related to the number of features, the simplicity of a revision or hypothesis is actually the simplicity of making the observations necessary to evaluate that hypothesis. While this distinction remains vague about what exactly simplicity is, it nevertheless refines our notion of the concept and indicates how we should introduce semantics. For instance,

the cost of measurement, the difficulty of accurate measurement, and the general difficulty of obtaining a measurement all, in some sense, capture the notion of simplicity and relate to individual features. When applying Kalpana to the RS domain, we chose the last of those three to help approximate the actual simplicity of a revision. In particular, we had the domain expert divide the attributes into the three categories easy, moderate, and difficult, referring to the ease of obtaining the associated values. Appendix D.3 lists these groups.

Once Kalpana determines which categories new features in a revision belong to, it must convert that information into a numerical score. Easy attributes are given a score of 0.5, moderates a score of 1.0, and attributes that are difficult to observe a score of 1.5. Returning to the example given for the syntactic simplicity measure, dyspnea falls into the easy category and wheezing belongs to the moderate class, which gives the revision a total simplicity score of 1.5. Although we suspect that collecting information on three easy attributes remains simpler than gathering measurements for a single difficult attribute, we wanted to retain a bias toward shorter antecedents. For instance, we would rather introduce a rule that mentions one difficult attribute, such as pulmonary embolus, than introduce five of the easy attributes, even if the required effort is lower in the latter case. Thus the scores related to the classes are closer in value than some more precise measure of difficulty might suggest.

5.5 TESTING MEASURES OF DEFENSIBILITY

To determine whether syntactic or semantic defensibility relates to the actual acceptability of a revision, we tested our measures in the respiratory syndrome (RS) domain. We used the results from our experiment in Section 4.2 to evaluate the efficacy of our measures. In particular, the measures were applied to the revisions produced by the models in Figures 15 and 16. Dr. Dowling's assessment of defensibility was our gold standard. We conjectured that the semantic measure of defensibility as a whole as well as the individual components would be significantly correlated with Dr. Dowling's ratings. We also suspected that the syntactic measures would fare poorly both individually and in combination.

Before evaluating the system, we acquired the domain knowledge relevant to Kalpana's measures from Dr. Dowling. This process lasted roughly two hours with the following three questions driving most of the discussion. With respect to conservatism, we asked for the most likely classification given that an individual feature was present. That is, does knowing that cough is present lead us to believe that RS is present or that it is absent. Next we asked, in relation to modesty, which of the 65 features provide the most valuable information in the diagnosis of RS. Thus we claim that a revision that introduces *cough* as an attribute for predicting RS is more modest than one that introduces *headache*, since the former attribute is more strongly associated with RS than the latter. With our final question, we had Dr. Dowling rate the attributes based on the difficulty of observation. So, noticing a patient's cough is much simpler than making a diagnosis of pulmonary embolus. The former attribute can be observed by the physician or reported by the patient. In contrast, the latter requires several diagnostic steps that are limited in their reliability (i.e., even x-ray results can yield false negatives for the condition). The complete knowledge acquired from our expert is given in Appendix D.

In addition to background knowledge, our measures require combining rules so that each measure contributes to the defensibility of the revision. For the syntactic measures, which all fell in the interval $[0, 1]$, we took the unweighted mean of the rule's conservatism, modesty, and simplicity. With the semantic measures, we used the sum of the scores. We also explored, using a separate set of revisions, a threshold for defensibility. In this experiment, we set the threshold to two. Thus we have nine measures of defensibility to test against our gold standard—the scores for the six individual measures, the two combined scores for the syntactic and semantic versions, and the version of the semantic measurement that was given a threshold.

Kalpana produced 90 unique revisions for the model in Figure 15 and 68 unique revisions for the model in Figure 16. These revisions were merged into a single data set of 158 rules, in which there were no repeated exception rules. Table 14 shows the correlation between each of the measures and the gold standard. P-values and 95% confidence intervals are also given.

Table 14: The correlation of each measure of defensibility with the gold standard.

Measure	r	p	95% CI
syntactic conservatism	-0.1395	0.0402	(-0.289, 0.017)
syntactic modesty	0.3160	< 0.0001	(0.169, 0.449)
syntactic simplicity	0.1348	0.0456	(-0.021, 0.284)
syntactic defensibility	0.0560	0.2423	(-0.101, 0.210)
semantic conservatism	0.2947	< 0.0001	(0.146, 0.431)
semantic modesty	0.0222	0.3909	(-0.134, 0.177)
semantic simplicity	-0.0759	0.1716	(-0.229, 0.081)
semantic defensibility	0.1370	0.0431	(-0.020, 0.287)
semantic defensibility (threshold of 2)	0.1117	0.0812	(-0.045, 0.263)

Of the nine measures, syntactic modesty and semantic conservatism correlated moderately with plausibility, both of which were highly significant. Additionally the thresholded and nonthresholded semantic measures of defensibility as well as syntactic simplicity and syntactic conservatism correlated weakly with the gold standard. Although we expected stronger correlations, we have confirmed that the semantic measure of defensibility associated more strongly with the underlying concept than the syntactic measure. From these results, we note several interesting findings.

Contrary to our expectations, the results indicate that syntactic conservatism inversely correlates with defensibility. That is, those revisions that conflict with more rules tend to be more defensible. One explanation for this is that when a revision conflicts with multiple rules, its new features are more likely to be present within the model. So, given two rules with differing antecedents, and an exception rule to both, some of the features in the antecedent of the first rule, but not in the antecedent of the second rule will be present in the revision. The exception is when the second rule is a direct specialization of the first rule. In addition,

syntactic modesty is moderately, positively correlated with defensibility. Therefore, the nonconservative revision, which happens to be more modest, appears more defensible.

Interestingly, syntactic modesty correlated the most highly with the gold standard. We conjecture that the performance of syntactic modesty relates to its underlying function, which leads to a preference for a condensed vocabulary. In the strictest sense, a modest revision introduces no new attributes or features into the model. If we assume that the vocabulary of the model suffices for the correct expression of a classifier, then we would expect syntactic modesty to be an accurate estimator of a revision's defensibility. This finding leads us to argue that the models used in this example contain an important subset of the relevant features defining respiratory syndrome. The poor performance of semantic modesty supports this argument in that the use of a larger set of familiar (i.e., modest) terms significantly reduced the efficacy of the measure.

Following syntactic modesty, semantic conservatism is the strongest indicator of defensibility. Due to the relationship between this measure and Pazzani's monotonicity constraints [44], our finding builds on that work. Whereas Pazzani evaluated these constraints with regards to the predictive accuracy of the generated rule sets, we focused our attention on the acceptability of such rules. We found that conservative rules need not always be defensible and that room exists for improving upon this measure. Combining monotonicity constraints with a strict language bias, as we introduced in syntactic modesty, seems most promising of the options we considered.

The two semantic measures of defensibility along with the syntactic measure of simplicity round out the remaining, weakly correlated estimators of defensibility. Although the combined semantic measures both performed better than the combined syntactic measure, their correlations were low. We believe that a poor choice for both modesty and simplicity measures was the primary cause. That is, a measure of modesty must more effectively constrain the language of the model than our semantic measure did, and the measure of simplicity should presumably focus more on cost efficiency, as in [65], than the difficulty of observation. Finally, we note that syntactic simplicity, which often appears as a preference during model building, seems inadequate as a measure of revision acceptability.

5.6 SUMMARY

In this chapter we built an understanding of defensibility from both a syntactic and semantic viewpoint. Our skeleton consisted of the virtues of hypotheses described by Quine and Ullian in *The Web of Belief*. After implementing the measures of defensibility and incorporating them into Kalpana, we tested them on data from the medical domain. Results showed that, to a slight degree, semantic measures of defensibility outperform straightforward versions of syntactic measures, although we were surprised to find that syntactic modesty correlated so highly with the gold standard. Given that a little knowledge garnered from a two hour session with a domain expert led to an increase in our ability to quantify defensibility, we conjecture that more and finer-grained knowledge will lead to more accurate measures. In particular, identifying a small set of well-entrenched attributes appears most promising.

6.0 CONCLUSION

Two fundamental themes converged within this exploratory work: anomaly-driven theory revision and the identification of acceptable revisions. Anomaly-driven revision involves the use of an incorrectly predicted example to select appropriate data and knowledge to generate and select a repair to an original model. We found that experts do not consider all revisions equal and that their judgment of the revisions can help increase the predictive accuracy of the model. This finding led us to search for a means of efficiently capturing the relevant domain knowledge of the expert to identify the set of acceptable anomaly resolutions. We were, to some extent, successful, and our work makes several contributions, with a few caveats, while suggesting interesting next steps.

6.1 CONTRIBUTIONS

A major contribution of this work is the identification of a method for performing anomaly-driven theory revision within the rule-based classification framework. That is, given an anomaly, we create four subsets of data each for the identification of both differences and similarities. We then either compare the data within each subset (in the method of agreement) or compare the anomaly to each subset (in the method of difference). Since each subset of data relates to the anomaly in a meaningful way, we can gain insight into why a particular case is anomalous, which pushes our revisions away from mere description and closer to the realm of explanation. We can move the revisions even further in the direction of explanation by assuring that they are acceptable to domain experts.

Providing operational definitions of the concepts of acceptability and defensibility is another main contribution of the current work. Acceptability consists of three criteria. The first, rehabilitation, is a definitional feature of any anomaly resolution. The second, monotonicity, plays a precautionary role, hopefully keeping a single example from causing a reduction in the future predictive accuracy of the model. The third, defensibility, requires the revision to be justifiable within the context of extended domain knowledge (i.e., knowledge relevant to, but not contained within the model). Using suggestions from Quine and Ullian [47], we explored the nature of defensibility, separating it into three components that guided both the development of measures and the acquisition of knowledge: conservatism, modesty, and simplicity.

We were not the first to consider some manner of acceptability as a core component of the rules within a model. In particular, monotonicity constraints [44] and the CLARUS system [7] capture versions of both conservatism and modesty, whereas many systems employ a measure of simplicity similar to the syntactic measure used in our system, Kalpana. The differences between our semantic measure of conservatism and monotonicity constraints are negligible, but the measures of efficacy differ. That is, while Pazzani showed that monotonicity constraints improve the predictive accuracy of a model, we examined their ability to capture defensibility. To this end, we found a positive correlation between semantic conservatism and an expert's judgment. These results fit well together, yielding stronger support for this relatively simple method of incorporating domain knowledge into the revision process. With respect to the relationship between CLARUS and our modesty measures, we also see mutual support. The main result is that establishing a preferred subset of features with which to discuss class membership assists in the identification of acceptable revisions. Or, less formally and much more generally, having an expert specify his or her preferences makes it easier to match those preferences.

In addition to the above results, we made minor discoveries regarding productive subsets of data for model revision. First, we found that when using the method of difference, the most fruitful subset of data for resolving the anomaly corresponds to the subset used in traditional decision-tree and rule-learning systems. To our knowledge this study yields the first explicit, empirical justification for preferring that particular subset. Second, we found

that comparing anomalies of the same class that are misclassified by the same rule leads to a high proportion of acceptable anomaly resolutions. Our approach, based on Mill's method of agreement, serves as a suitable basis for any program that processes multiple anomalies at once. Third, we found that applying acceptable revisions results in fewer new anomalies in unseen data, versus arbitrary selection, and this directly translates to improved predictive accuracy.

6.2 LIMITATIONS

Though this work makes several contributions, we note that it also possesses a number of limitations. In particular, the study itself was one of exploration. So, we concerned ourselves mostly with defining a new method of model revision and testing it on a limited set of data. As a result, our findings should be interpreted as guidelines for future research as opposed to a finished product. Apart from the relatively limited evaluation, testing within the medical domain carries its own caveats.

We note that the data used for this study came from a domain full of conflict. Although the three physicians that produced our table of attribute values often agreed, this was not always the case. Even though they examined the same medical reports, they brought their own training and skills into the data extraction process. In fact, disagreement among medical experts is both normal and expected. Selecting the majority opinion should control for some of the disagreement, but there are no guarantees. Additionally, the concept of respiratory syndrome is somewhat abstract, so errors in class labeling should also be expected—a condition that is exacerbated by having only one physician's opinion on the class.

Finally, by converting our attributes into Boolean form, we traded information for a simplified model space. In particular, the difference between a missing attribute and the same attribute specifically noted as absent can be significant. For example, knowing that a patient's x-ray was negative for pneumonia gives strong evidence against pneumonia. However, if that patient never had a chest radiograph, then we must retain our uncertainty. Interest-

ingly, knowing that a particular attribute is missing sometimes carries information of its own, indicating that the attending physician may have deemed the measurement unnecessary.

Regardless of these limitations, we believe that our results are strong enough to support future research. Most of the problems mentioned above are inherent to work within the medical domain and are not rare or specific to this specific study. Applying our methods to other domains should help overcome these problems while determining the generality of the approach.

6.3 FUTURE WORK

While expanding into other domains is a reasonable future step for our research, other extensions also present themselves. For instance, anomaly resolutions possess more than mere descriptive power. To a certain extent they explain the anomaly since they isolate those conditions present only within the anomaly. We hope to determine how to exploit this extra information to extend beyond flat exception rules into the realm of deep explanations. We also hope to explore anomaly-driven revision in the context of different formalisms such as those used for equation discovery [34, 35, 58, 59]. Finally, we intend to investigate cases where revisions actually *explain away* anomalies, allowing us to ignore those data with respect to a specific model.

We began with a claim that anomalies drive scientific discovery. To address this claim, we developed methods for analyzing anomalies within a concept learning framework. By centering our work around the anomaly we create new possibilities for the development of discovery algorithms. In particular, we emphasize a revision's justifiability and what it claims about an anomaly as opposed to its complexity or its effect upon the original, flawed model. We end by saying that we have only scratched the surface of anomaly-driven techniques and the identification of acceptable revisions.

APPENDIX A

REVISIONS FROM CHAPTER 2

The following twelve revisions were generated for the example in Chapter 2. The comments were provided by our domain expert, Dr. John Dowling, and have been subjected to minor editing to preserve consistency and clarity.

Anomalies Resolved: (6 7 229 87)
Original Rule:
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(DYSPNEA is PRESENT) and
(OXYGEN_DESATURATION is ABSENT) and
(X_RAY_PNEUMONIA is ABSENT) and
(X_RAY_HYPERINFLATED_LUNGS is ABSENT) and
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Agreement: Compared to anomalies of the same observed class
Semantic Defensibility: -0.5
Syntactic Defensibility: 0.38333338
Domain Expert's Comment: Not acceptable. All those negative attributes don't rule out respiratory syndrome (RS) when cough and dyspnea are present. In particular, x-ray pneumonia could be absent, yet there is a lesion or foreign body within the tracheobronchial tree.

Anomalies Resolved: (6 7 229 87)
Original Rule:
(SPUTUM is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)

Exception Rule:
(WHEEZING is ABSENT) and
(X_RAY_PLEURAL_EFFUSION is ABSENT) and
(X_RAY_PNEUMONIA is ABSENT) and
(X_RAY_HYPERINFLATED_LUNGS is ABSENT) and
(SPUTUM is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Agreement: Compared to anomalies of the same observed class
Semantic Defensibility: -2
Syntactic Defensibility: 0.4666667
Domain Expert's Comment: Not acceptable. Similar reasoning. All those negative attributes don't rule out RS when sputum is present. In particular, x-ray pneumonia could be absent, yet there is sputum coming from the tracheobronchial tree, as in bronchitis. [Editorial note: Chronic bronchitis was not classified as an RS.]

Anomalies Resolved: (87 229)
Original Rule:
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(TACHYCARDIA is PRESENT) and
(OXYGEN_DESATURATION is ABSENT) and
(X_RAY_PNEUMONIA is ABSENT) and
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Agreement: Compared to anomalies misclassified by the same rule of the same class
Semantic Defensibility: 0.5
Syntactic Defensibility: 0.48133898
Domain Expert's Comment: Not acceptable. Cough is indicative of some sort of RS with 99% assurity. Tachycardia is non-specific and can just mean that the patient is sick rather than indicating a cardiac cause.

Anomalies Resolved: (229)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(X_RAY_PNEUMONIA is ABSENT) and
(PLEURITIC_PAIN is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to data of the predicted class matching

the incorrect classifier
Semantic Defensibility: 0.5
Syntactic Defensibility: 0.5357023
Domain Expert's Comment: Not acceptable. Pleuritic pain adds to the likelihood that RS is present.

Anomalies Resolved: (229)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule: (CHILLS is PRESENT) and
(ASTHMA is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to all data
Semantic Defensibility: -1.0
Syntactic Defensibility: 0.5023689
Domain Expert's Comment: Not acceptable. Chills doesn't detract from the likelihood that RS is present. (I believe we are not calling asthma an RS [that we are interested in]). [Editorial note: See Dr. Dowling's comment on the next case for an explanation of why asthma, which is not considered an RS and which could explain dyspnea, is not sufficient for ruling out respiratory syndrome.]

Anomalies Resolved: (229)
Original Rule:
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(CHILLS is PRESENT) and
(ASTHMA is PRESENT) and
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to all data
Semantic Defensibility: -1.0
Syntactic Defensibility: 0.5023689
Domain Expert's Comment: Not acceptable. Asthma per se shouldn't give chills (i.e., fever). So, there is likely some RS complicating the asthma (e.g., pneumonia).

Anomalies Resolved: (87)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:

(FEVER is PRESENT) and
(BRONCHITIS is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to all data
Semantic Defensibility: -1.5
Syntactic Defensibility: 0.5023689
Domain Expert's Comment: Not acceptable. Bronchitis and fever strengthen the likelihood of RS.

Anomalies Resolved: (87)
Original Rule:
(SPUTUM is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(FEVER is PRESENT) and
(BRONCHITIS is PRESENT) and
(SPUTUM is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to all data
Semantic Defensibility: -1.5
Syntactic Defensibility: 0.5023689
Domain Expert's Comment: Not acceptable. Bronchitis and fever strengthen likelihood of RS.

Anomalies Resolved: (87)
Original Rule:
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(FEVER is PRESENT) and
(BRONCHITIS is PRESENT) and
(COUGH is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to all data
Semantic Defensibility: -1.5
Syntactic Defensibility: 0.5023689
Domain Expert's Comment: Not acceptable. Fever and bronchitis strengthen the likelihood of RS.

Anomalies Resolved: (7)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)

Exception Rule:
(ACUTE_CORONARY_SYNDROME is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to data of the originally assigned class
Semantic Defensibility: 1.5
Syntactic Defensibility: 0.6
Domain Expert's Comment: Acceptable. Acute coronary syndrome explains dyspnea in the absence of RS.

Anomalies Resolved: (7)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(CHEST_TENDERNESS is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to data of the originally assigned class
Semantic Defensibility: 1
Syntactic Defensibility: 0.6
Domain Expert's Comment: Acceptable. Musculoskeletal chest injury explains dyspnea.

Anomalies Resolved: (6)
Original Rule:
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is PRESENT)
Exception Rule:
(ACUTE_CORONARY_SYNDROME is PRESENT) and
(DYSPNEA is PRESENT)
implies (RESPIRATORY_SYNDROME is ABSENT)
Method of Generation: Method of Difference: Compared to data of the originally assigned class
Semantic Defensibility: 1.5
Syntactic Defensibility: 0.6
Domain Expert's Comment: Acceptable. Acute coronary syndrome explains dyspnea.

APPENDIX B

PSEUDOCODE

B.1 METHOD OF AGREEMENT

```
Method-of-Agreement(anomalies, data):
  // remove the anomalies from the rest of the data
  nonanomalies = data - anomalies
  // collect all features shared by the group of anomalies
  pool = shared-features(anomalies)
  // initialize a list of revisions
  revisions = [ ]
  // separately consider each anomaly
  for each a in anomalies
    // consider each rule that incorrectly classifies the anomaly
    for each i in (incorrect-classifiers(a))
      // create the root revision from the antecedent of the
      // incorrect classifier and the correct classification of
      // the anomaly
      r = create-root(antecedent(i), class(a))
      // add all the features that keep the new revision from
      // creating any new anomalies
      push(necessary-features(r, pool, nonanomalies),
           antecedent(r))
      // while the new revision continues to create new anomalies
      // and there are features that the program can use to
      // specialize the revision
      while (overly-general(r, nonanomalies) and
             features-left(pool, r))
        // add an unused feature that best separates the anomaly from
        // nonanomalous data incorrectly classified by the current
        // revision
        push(best-separator(r, pool - antecedent(r), nonanomalies),
```

```
            antecedent(r))
        // push the resulting revision onto the list
        push(r, revisions)
  return revisions
```

B.2 METHOD OF DIFFERENCE: BASIC

```
Method-of-Difference-Basic(anomalies, data):
  // begin with an empty set of revisions
  revisions = [ ]
  // remove the anomalies from the rest of the data
  nonanomalies = data - anomalies
  // individually consider each anomaly
  for all a in anomalies
    // consider each feature expressed in the anomaly that is
    // not expressed in any of the nonanomalous data
    for all s in separators(feature-set(a), nonanomalies)
      // consider each rule that incorrectly classifies the
      // anomaly
      for all i in incorrect-classifiers(a)
        // create the root revision from the antecedent of the
        // incorrect classifier and the correct classification of
        // the anomaly
        r = create-root(antecedent(i), class(a))
        // add the separating feature to the antecedent of the
        // root revision
        push(s, antecedent(r))
        // when the root revision meets the monotonicity
        // criterion, add it to the collection of revisions
        unless(overly-general(r, nonanomalies))
          push(r, revisions)
  return revisions
```

B.3 METHOD OF DIFFERENCE: DECISION BRANCH

```
Method-of-Difference-Decision-Branch(anomalies, data):
  // begin with an empty set of revisions
  revisions = [ ]
  // remove the anomalies from the rest of the data
  nonanomalies = data - anomalies
  // individually consider each anomaly
```

```
  for all a in anomalies
    // consider each rule that incorrectly classifies the anomaly
    for all i in incorrect-classifiers(a)
      // create the root revision from the antecedent of the
      // incorrect classifier and the correct classification of
      // the anomaly
      r = create-root(antecedent(i), class(a))
      // while the revision fails to meet the monotonicity
      // criterion and while the features matching the observed
      // values in the anomaly have not all been used in the
      // antecedent of the revision
      while(overly-general(r, nonanomalies) and
            (feature-set(a) - features(antecedent(r)) > 0)
        // find the one feature that separates the anomaly from the
        // greatest number of nonanomalous data in the current
        // subset
        // add that feature to the antecedent of the revision
        push(best-separator(r, feature-set(a) - antecedent(r),
                            nonanomalies),
             antecedent(r))
      // if the revision meets the monotonicity criterion, keep it
      unless(overly-general(r, nonanomalies))
        push(r, revisions)
  return revisions
```

B.4 NONTRIVIAL SUBPROCEDURES

```
shared-features(anomalies):
  // select an arbitrary anomaly
  a = first(anomalies)
  // create a list of the features that match the anomaly
  shared-feature-list = [ ]
  for each f in feature-list
    when matches(f, a)
      push(f, shared-feature-list)
  // walk through the collection of anomalies, matching
  // each feature to each anomaly
  for each a in anomalies
    for each f in shared-feature-list
      // when the selected feature does not match an
      // anomaly (and therefore is not shared among the
      // set), remove it from the list of shared features
      when not(matches(f,a))
        remove(f, shared-feature-list)
```

```
  return shared-feature-list


necessary-features(root, pool, nonanomalies):
  // initialize the necessary features to an empty list
  necessary-feature-list = [ ]
  // visit each feature in the pool of features
  for each f in pool
    // create a new revision from the root revision, adding
    // all features except the one currently selected to the
    // antecedent of the root
    // note that new-rule takes a collection of features that
    // it assembles into a conjunctive antecedent and a single
    // feature that becomes the consequent
    new-revision = new-rule(antecedent(root) + pool - f,
                   consequent(root))
    // match the new revision against all the nonanomalous data
    // if a nonanomaly matches the antecedent of the rule, then
    // the feature that was left out is considered necessary for
    // avoiding the misclassification of the nonanomaly
    for d in nonanomalies
      if (match(new-revision, d))
        push(f, necessary-feature-list)
        break to 'for each f in pool'
  return necessary-feature-list


best-separator(revision, features, nonanomalies):
  // keep track of the feature that best separates
  // the anomaly classified by the current revision
  // from all the nonanomalies
  least-number-matched = infinity
  best-separator
  // examine each feature not already used within the
  // antecedent of the revision
  for f in features - antecedent(revision)
    // create a new rule from the antecedent of the revision,
    // the current feature, and the consequent of the revision
    new-revision = new-rule(antecedent(revision) + f,
                            consequent(revision))
    // count how many nonanomalies match the new revision
    number-matched = count-matches(new-revision, nonanomalies)
    if (number-matched < least-number-matched)
      best-separator = f
      least-number-matched = number-matched

  return best-separator
```

```
separators(features, nonanomalies):
  // keep track of all the features that do not match any
  // of the observed values in the nonanomalies
  separator-list = [ ]
  // match each feature to every nonanomaly
  for f in features
    is-separator = true
    for d in nonanomalies
      // if the feature matches a nonanomaly, move on to the
      // next feature, otherwise add it to the list of separators
      if (match-feature(f, d))
        is-separator = false
        break;
    if (is-separator)
      push(f, separator-list)

  return separator-list
```

APPENDIX C

INSTRUCTIONS FOR THE DOMAIN EXPERT

There are 158 items consisting of an original rule that misclassified one or more cases and an exception rule that should explain (away) the misclassification. The exception rule will include all the features (e.g., *rhonchi* are *present*) contained within the original rule in addition to new features. Those new features are supposed to carry the explanatory power in the context of the features from the original rule.

I would like you to judge whether the exception rules are plausible based on the new features alone. That is, can the new features in the exception rule justify a change in the target class? As an example of how to interpret the raw presentation of the rules, the original rule in the first rule pair states, "IF rhonchi are present, THEN the patient has respiratory syndrome." Some cases were found that contradict that rule, leading to the creation of the exception rule, "EVEN IF rhonchi are present, when fever is absent and there is no x-ray result that is positive for pneumonia, THEN the patient does not have respiratory syndrome."

Once you have judged the plausibility of the rule, please indicate whether the judgment was difficult to make. In this data, acute and chronic were grouped as RS is *present*. If you would like further clarification, please ask.

APPENDIX D

DOMAIN KNOWLEDGE IN KALPANA

D.1 THE CONSERVATISM OF ATTRIBUTES

The following attributes, when present, are associated with the *presence* of respiratory syndrome.

Signs and Symptoms: congestion, cough, dyspnea, hemoptysis, pleuritic pain, sputum, chills, history of pneumonia

Physical Findings: breath sounds decreased, cyanosis, dullness, fever, oxygen desaturation, wheezing, stridor, tachypnea, rales/crackling, rhonchi

Chest Radiograph Findings: pulmonary edema, pneumothorax, widened mediastinum, pneumonia

Diagnoses: asthma, chronic obstructive pulmonary disease, cystic fibrosis, empyema, influenza, pneumonia, HIV/AIDS

The following attributes, when present, are associated with the *absence* of respiratory syndrome.

Signs and Symptoms: chest pain, conjunctivitis, stomatitis, upper abdominal pain, headache, flu symptoms, sweats

Physical Findings: abdominal distension, cervical adenopathy, chest tenderness, pleural rub, subcutaneous edema chest/neck, tachycardia

Chest Radiograph Findings: atelectasis, hyperinflated lungs, mass, mediastinal shift, pericardial effusion, pleural effusion, poor inspiration

Diagnoses: acute coronary syndrome, anxiety, aortic dissection, cardiomyopathy, chest trauma, hiatal hernia, pharyngitis, pneumothorax, pulmonary edema CHF, pulmonary embolus, sarcoidosis, sepsis, viral syndrome, lung tumor

D.2 THE MODESTY OF ATTRIBUTES

The following attributes were considered more valuable than the rest when identifying a case of respiratory syndrome.

Signs and Symptoms: congestion, cough, dyspnea, hemoptysis, pleuritic pain, sputum, chills, pneumonia history

Physical Findings: breath sounds decreased, cyanosis, dullness, fever, oxygen desaturation, wheezing, rhonchi, stridor, tachypnea, rales/crackling, chest tenderness, pleural rub

Chest Radiograph Findings: pneumonia, mass, pulmonary edema, pneumothorax,

Diagnoses: acute coronary syndrome, anxiety, aortic dissection, cardiomyopathy, chest trauma, hiatal hernia, pharyngitis, pneumothorax, chronic obstructive pulmonary disease, cystic fibrosis, empyema, HIV/AIDS, influenza, lung tumor, musculoskeletal chest pain, pneumonia, asthma, pulmonary edema CHF, pulmonary embolus, sepsis, viral syndrome

D.3 THE SIMPLICITY OF ATTRIBUTES

The domain expert considered the following attributes to be relatively easy to observe.

Signs and Symptoms: congestion, cough, dyspnea, hemoptysis, pleuritic pain, sputum, chills, history of pneumonia, chest pain, conjunctivitis, stomatitis, upper abdominal pain, headache, flu symptoms, sweats

Physical Findings: stridor, tachypnea, cyanosis, abdominal distension, tachycardia, fever

Diagnoses: acute coronary syndrome, chest trauma, pharyngitis, sarcoidosis, viral syndrome, asthma, chronic obstructive pulmonary disease, HIV/AIDS

The domain expert considered the following attributes to be moderately difficult to observe.

Physical Findings: breath sounds decreased, dullness, wheezing, rhonchi, rales/crackling, oxygen desaturation, cervical adenopathy, chest tenderness, pleural rub, subcutaneous edema chest neck

Chest Radiograph Findings: pneumonia, pulmonary edema, widened mediastinum, atelectasis, mass, poor inspiration, pneumothorax, hyperinflated lungs, mediastinal shift, pericardial effusion, pleural effusion

Diagnoses: anxiety, hiatal hernia, pneumothorax, pulmonary edema CHF, sepsis, empyema, influenza, lung tumor, musculoskeletal chest pain, pneumonia

The domain expert considered the following attributes to be relatively difficult to observe.

Diagnoses: aortic dissection, cardiomyopathy, pulmonary embolus

BIBLIOGRAPHY

[1] 1989. *International Statistical Classification of Diseases and Related Health Problems*, 9th revision. Geneva, Switzerland: World Health Organization.

[2] Alchourrón, Carlos E., Peter Gärdenfors, and David Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, **50**(2), 510–530.

[3] Allais, Maurice F. C. 1959. Should the laws of gravity be reconsidered. *Aero/space Engineering*, September, 46–52.

[4] Aronis, John M. and Foster J. Provost. 1997. Increasing the efficiency of data mining algorithms with breadth-first marker propagation. *Pages 119–122 of:* David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy (eds.), *Proceedings of the Third International Conference on Knowledge Discovery*, Newport Beach, CA. Menlo Park, CA: AAAI Press.

[5] Baffes, Paul T. and Raymond J. Mooney. 1993. Extending theory refinement to m-of-n rules. *Informatica*, **17**, 387–397.

[6] Brewer, William F. and Clark A. Chinn. 1994. Scientists' responses to anomalous data: Evidence from psychology, history, and philosophy of science. *Pages 304–313 of: The Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol. 1, East Lansing, MI.

[7] Brunk, Clifford and Michael J. Pazzani. 1995. A lexical based semantic bias for theory revision. *Pages 81–89 of:* Armand Prieditis and Stuart J. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, CA. Los Altos, CA: Morgan Kaufmann.

[8] Carbonara, Leonardo and Derek Sleeman. 1999. Effective and efficient knowledge base refinement. *Machine Learning*, **37**, 143–181.

[9] Chinn, Clark A. and William F. Brewer. 1998. An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, **35**(6), 623–654.

[10] Danyluk, Andrea Phoreckyj. 1991. Gemini: An integration of analytical and empirical learning. *Pages 191–206 of:* Ryszard S. Michalski and Gheorghe Tecuci (eds.), *Proceedings of the First International Workshop on Multistrategy Learning*, Harpers Ferry, WV. Fairfax, VA: George Mason University Press.

[11] Darden, Lindley. 1991. *Theory Change in Science: Strategies from Mendelian Genetics*. New York City, NY: Oxford University Press.

[12] Darden, Lindley. 1992. Strategies for anomaly resolution. *Pages 251–273 of:* Ronald N. Giere (ed.), *Cognitive Models of Science*. Minnesota Studies in the Philosophy of Science, vol. 15. Minneapolis, MN: University of Minnesota Press.

[13] Darwiche, Adnan and Judea Pearl. 1997. On the logic of iterated belief revision. *Artificial Intelligence*, **89**, 1–29.

[14] Davis, Randall and Douglas B. Lenat. 1982. *Knowledge-Based Systems in Artificial Intelligence*. New York City, NY: McGraw-Hill.

[15] Duif, Chris P. 2004. A review of conventional explanations of anomalous observations during solar eclipses. ArXiv:gr-qc/0408023, http://arxiv.org/abs/gr-qc/0408023 (accessed August 21, 2004).

[16] Einstein, Albert. 1934. On the method of theoretical physics. *Philosophy of Science*, **1**(2), 163–169.

[17] Fellbaum, Christiane (ed.). 1998. *Wordnet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

[18] Fox, Dirk, Martin Schmidt, Misha Koshelev, Vladik Kreinovich, Luc Longpré, and Jeff Kuhn. 1997. We must choose the simplest physical theory: Levin-Li-Vityani theorem and its potential physical applications. Technical Report UTEP-CS-97-21. University of Texas at El Paso.

[19] Frazer, Sir James George. 1922. *The Golden Bough: A Study in Magic and Religion*. abridged edn. New York City, NY: Macmillan Co.

[20] Gärdenfors, Peter and David C. Makinson. 1988. Revisions of knowledge systems and epistemic entrenchment. *Pages 83–95 of:* M. Vardi (ed.), *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*. Los Altos, CA: Morgan Kaufmann.

[21] Genest, Jean, Stan Matwin, and Boris Plante. 1990. Explanation-based learning with incomplete theories: A three-step approach. *Pages 286–294 of:* Bruce W. Porter and Raymond J. Mooney (eds.), *Proceedings of the Seventh International Conference of Machine Learning*, Austin, TX. Los Altos: Morgan Kaufmann.

[22] Goodman, Nelson. 1958. The test of simplicity. *Science*, **128**(3331), 1064–1069.

[23] Goodman, Nelson. 1961. Safety, strength, simplicity. *Philosophy of Science*, **28**(2), 150–151.

[24] Goodman, Nelson. 1965. *Fact, Fiction, and Forecast*. 2nd edn. Indianapolis, IN: The Bobbs-Merrill Company, Inc.

[25] Hempel, Carl G. 1945. Studies in the logic of confirmation (i.). *Mind*, **54**, 1–26.

[26] Henschen, F. 1965. The Nobel Prize in physiology or medicine 1928, presentation speech. *In: Nobel Lectures, Physiology or Medicine 1922-1941*. Amsterdam, Holland: Elsevier Publishing Company.

[27] Hume, David. 1739. *A Treatise of Human Nature*. Project Gutenberg. http://www.gutenberg.org/dirs/etext03/trthn10.txt (accessed November 27, 2004).

[28] Karp, Peter D. 1989. *Hypothesis Formation and Qualitative Reasoning in Molecular Biology*. Ph.D. thesis, Department of Computer Science, Stanford University, Stanford, CA.

[29] Koppel, Moshe, Ronen Feldman, and Alberto Maria Segre. 1994. Bias-driven revision of logical domain theories. *Journal of Artificial Intelligence Research*, **1**, 159–208.

[30] Kuhn, Thomas S. 1957. *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*. Cambridge, MA: Harvard University Press.

[31] Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. second edn. Chicago, IL: The University of Chicago Press.

[32] Kuhn, Thomas S. 1977. The essential tension: Tradition and innovation in scientific research. *Chap. 9, pages 225–239 of: The Essential Tension*. Chicago, IL: Chicago University Press.

[33] Kulkarni, Deepak. 1988. *The Processes of Scientific Research: The Strategy of Experimentation*. Ph.D. thesis, Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

[34] Langley, Pat, Javier Sánchez, Ljupčo Todorovski, and Sašo Džeroski. 2002. Inducing process models from continuous data. *Pages 347–354 of:* Claude Sammut and Achim Hoffmann (eds.), *Proceedings of the Nineteenth International Conference on Machine Learning*, Sydney, Australia. Los Altos, CA: Morgan Kaufmann.

[35] Langley, Pat, Dileep George, Stephen Bay, and Kazumi Saito. 2003. Robust induction of process models from time-series data. *Pages 432–439 of:* Tom Fawcett and Nina Mishra (eds.), *Proceedings of the Twentieth International Conference on Machine Learning*, Washington, D.C. Menlo Park, CA: AAAI Press.

[36] Lawlor, Reed C. 1969. Axioms of fact polarization and fact ranking—their role in stare decisis. *Villanova Law Review*, **14**, 703–726.

[37] Li, Ming and Paul Vitanyi. 1992. Inductive reasoning and Kolmogorov complexity. *Journal of Computer System Sciences*, **44**(2), 343–384. http://www.math.uwaterloo.ca/~mli/ (accessed December 13, 2003).

[38] Mahoney, J. Jeffrey and Raymond J. Mooney. 1994. Comparing methods for refining certainty-factor rule-bases. *Pages 173–180 of:* William W. Cohen and Haym Hirsh (eds.), *Proceedings of the Eleventh International Conference of Machine Learning*, New Brunswick, NJ. Los Altos, CA: Morgan Kaufmann.

[39] Mill, John Stuart. 1900. *A System of Logic Ratiocinative and Inductive Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation.* 8th edn. London, UK: Longmans, Green, and Co.

[40] Miller, George A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, **63**, 81–97.

[41] Mitchell, Tom M. 1982. Generalization as search. *Artificial Intelligence*, **18**, 203–226.

[42] Mooney, Raymond J. and Dirk Ourston. 1994. A multistrategy approach to theory refinement. *Pages 141–164 of:* Ryszard S. Michalski and Gheorghe Teccuci (eds.), *Machine Learning: A Multistrategy Approach*, vol. 4. San Mateo, CA: Morgan Kaufmann.

[43] Pagnucco, Maurice. 1996 (February). *The Role of Abductive Reasoning Within the Process of Belief Revision.* Ph.D. thesis, University of Sydney, Sydney, Australia.

[44] Pazzani, M. J., S. Mani, and W. R. Shankle. 2001. Acceptance by medical experts of rules generated by machine learning. *Methods of Information in Medicine*, **40**(5), 380–385.

[45] Popper, Karl. 2002. *The Logic of Scientific Discovery.* London, UK: Routledge.

[46] Provost, Foster J. and Bruce G. Buchanan. 1995. Inductive policy: The pragmatics of bias selection. *Machine Learning*, **20**(1), 35–61.

[47] Quine, Willard V. and Joseph S. Ullian. 1978. *The Web of Belief.* second edn. New York City, NY: McGraw-Hill.

[48] Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann.

[49] Richards, Bradley L. and Raymond J. Mooney. 1995. Automated refinement of first-order horn-clause domain theories. *Machine Learning*, **19**(2), 95–131.

[50] Rott, Hans. 1999. Coherence and conservatism in the dynamics of belief, part i: Finding the right framework. *Erkenntnis*, **50**, 387–412.

[51] Rott, Hans. 2000. Two dogmas of belief revision. *Journal of Philosophy*, **97**, 503–522.

[52] Rott, Hans. 2003. Coherence and conservatism in the dynamics of belief, part ii: Iterated belief change without dispositional coherence. *Journal of Logic and Computation*, **13**, 111–145.

[53] Rudner, Richard S. 1961. An introduction to simplicity. *Philosophy of Science*, **28**(2), 109–119.

[54] Schmidhuber, Jurgen. 1995. Discovering solutions with low Kolmogorov complexity and high generalization capability. *Pages 488–496 of:* Armand Prieditis and Stuart J. Russell (eds.), *Proceedings of the Twelfth International Conference on Machine Learning*, Tahoe City, CA. Los Altos: Morgan Kaufmann.

[55] Shapere, Dudley. 1980. The character of scientific change. *Pages 61–101 of:* Thomas Nickles (ed.), *Scientific Discovery, Logic, and Rationality*. Boston Studies in the Philosophy of Science. Boston, MA: D. Reidel Publishing Company.

[56] Simon, Herbert. 1968. On judging the plausibility of theories. *Pages 439–459 of:* B. Van Rootselaar and J. F. Staal (eds.), *Logic, Methodology, and Philosophy of Science III*. Amsterdam, Holland: North-Holland Publishing, Co.

[57] Solomonoff, Ray J. 1964. A formal theory of inductive inference, part 1. *Information and Control*, **7**(1), 1–22. http://world.std.com/~rjs/1964pt1.pdf (accessed December 13, 2003).

[58] Todorovski, Ljupčo and Sašo Džeroski. 1997. Declarative bias in equation discovery. *Pages 376–384 of:* Douglas H. Fisher (ed.), *Proceedings of the Fourteenth International Conference on Machine Learning*, Nashville, TN. Los Altos: Morgan Kaufmann.

[59] Todorovski, Ljupčo, Sašo Džeroski, Pat Langley, Christopher Potter. 2003. Using equation discovery to revise an earth ecosystem model of the carbon net production. *Ecological Modelling*, **170**, 141–154.

[60] van den Bosch, Alexander P.M. 1994. *Simplicity and Prediction*. Unpublished Manuscript. http://citeseer.nj.nec.com/vandenbosch94simplicity.html (accessed December 8, 2003).

[61] Weyl, Hermann. 1949. *Philosophy of Mathematics and Natural Science*. Princeton, NJ: Princeton University Press.

[62] Whitehall, Bradley L., Stephen C-Y, and Robert E. Stepp. 1991. Theory completion using knowledge-based learning. *Pages 144–159 of:* Ryszard S. Michalski and Gheorghe Tecuci (eds.), *Proceedings of the First International Workshop on Multistrategy Learning*, Harpers Ferry, WV. Fairfax, VA: George Mason University Press.

[63] Wolfram, Stephen. 2002. *A New Kind of Science*. Champaign, IL: Wolfram Media, Inc.

[64] Wrinch, Dorothy and Harold Jeffreys. 1921. On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, **42**(249), 369–390.

[65] Zupan, Blaž, Ivan Bratko, Janez Demšar, Peter Juvan, Tomaž Curk, Urban Borštnik, J. Robert Beck, John Halter, Adam Kuspa, and Gad Shaulsky. 2003. Genepath: A system for inference of genetic networks and proposal of genetic experiments. *Artificial Intelligence in Medicine*, **29**, 107–130.