Towards a Method of Searching a Diverse Theory Space for Scientific Discovery

Joseph Phillips

University of Pittsburgh
Computer Science Dept.
Pittsburgh, PA 15260, USA
josephp@cs.pitt.edu

Abstract. Scientists need *customizable* tools to help them with discovery. We present an adjustable heuristic function for scientific discovery. This function may be considered in either a Minimum Message Length (MML) or a Bayesian Net manner. The function is approximate because the default method of specifying theory prior probabilities is a gross estimate and because there is more to theory choice than maximizing probability. We do, however, effectively capture some user preferences with our technique. We show this for the qualitatively different domains of geophysics and sociology.

1 Introduction

Our ultimate goal is to write a general program to assist scientists in creating and improving scientific models. Realizing this goal requires progress in machine learning, knowledge discovery in databases, data visualization and search algorithms. It also requires progress in scientific model preferencing. The scientific model preference problem is compounded by the fact that several scientists with very similar background knowledge may see the same data but may prefer different models. This paper is the first in an on going study to address scientific model preferencing issue.

Scientific discovery can be viewed as a parameter search in a large and extremely inhomogeneous space. Physicists, for example, prefer strong relationships between numeric values (*e.g.*, equations) when they can be found. They also, however, use knowledge that is more conveniently expressed hierarchically in decision trees and semantic nets. This is exemplified by the classification of, and the assigning of fundamental properties to subatomic particles.

The minimum message length (MML) criterion is a mathematically well-grounded approach for choosing the most probable theory given data [21][8][24][5]. Inspired by information theory, the criterion states that the most probable model has the smallest encoding of both the theory and data. Ideally, the theory's encoding results from a domain expert's estimation of its prior probability and is language independent. The encoding of the data should also be probabilistic: as a function of a given theory.

Despite its generality and power for finding parameters in single classes of models (*e.g.*, the class of polynomials), many have expressed skepticism about whether MML may meaningfully be applied to finding parameters in inhomogeneous model spaces (*e.g.*, general scientific discovery). Cheeseman, for example, states "although finding the most probable domain model is often regarded as the goal of scientific investigation, in general, it is not the optimal means of making predictions." [5]

Our immediate, limited goal is to devise a heuristic function that can help users in large and inhomogeneous model spaces. Ideally, a search algorithm that is informed with our heuristic will return several regions in the model space that contain promising models, some known and some novel. Our approach is to adapt MML in a customizable manner.

1. We make MML applicable to a larger set of scientific discovery by mapping its terms onto those used by scientists: theory, laws and data. The MML theory is mapped to scientific theory. The MML data is split into scientific laws and data.
2. We make our heuristic function adjustable, but in a principled manner, by giving the user only two calibration parameters. These parameters directly correspond to the relationship between scientific theory and law, and scientific theory and data. It would be nice if we could ignore differences between theories and pretend that there is one "best" theory for all scientists. This, however, ignores significant evidence that scientists differ in opinion, *e.g.*, see [10][15].

We judge our function based on criteria for heuristic functions: generality, ease of computation, simplicity and smoothness.

We do *not* claim that we have "solved" this problem. The feature set by which to judge theories and the identification of the "best" model remain unsolved problems.

1. We offer no good guidance in developing the theory's prior probability. Cheeseman and others have stressed the importance of using domain knowledge to specify the theory's prior probability. They have also stated that syntactic features are often a poor substitute. We are aware of no general algorithm for the estimation of a theory's prior probability. Although our technique is not limited to syntactic features, we use them in this paper. Our approach is compatible with more principled prior probability specifying techniques.
2. We make no claim that the "best" theory will result from this approach. This is due to (1) the unsolved prior probability problem, (2) to the difficulty in searching a large and inhomogeneous model space, and (3) the fact that the most probable model may or may not be the best model.

We have developed a useful heuristic function despite these two major limitations. Its generality is tested by analyzing its performance in two completely different domains: sociology and geophysics.

This paper is organized as follows. Section 2 discusses previous approaches to automated scientific discovery. Section 3 briefly introduces MML. Our approach is detailed in section 4. Section 5 presents and discusses our experiments. Section 6 concludes.

2 Scientific Discovery

Several criteria have been proposed by philosophers of science for comparing competing hypotheses [3]. Among them are accuracy/empirical support, simplicity, novelty and cost/utility. Most automated approaches consider accuracy and simplicity.

IDS by Nordhausen and Langley was perhaps the first *general* program for scientific discovery [18][19]. IDS takes as input an initial hierarchy of abstracted states and a sequential list of "histories" (qualitative states, see [6]). Using each history IDS

modifies the affected nodes of the abstracted state tree to incorporate any new knowledge gained from that history. Its output is a fuller, richer hierarchy of nodes representing history abstractions.

Thagard introduced Processes of Induction (or PI), to propose a computational scheme for scientific reasoning and discovery, but not as a working discovery tool [23]. PI represents models as having theories, laws and data. It evaluates scientific models by multiplying a simplicity metric by a data coverage metric. The simplicity metric is a function of how many facts have been explained and of how many co-hypotheses were needed to help explain them. The evaluation scheme is fixed and has no notion of degree of inaccuracy.

Zytkow and Zembowicz developed 49er, a general knowledge discovery tool [27][26]. It has a two stage process for finding regularities in databases. The first stage creates contingency tables (counts of how often values of one attribute co-occur with those of another) for pairings of database attributes. The second stage uses the contingency tables to constrain the search for other, higher order, regularities (*e.g.* taxonomies, equations, subset relations, *etc*.)

Valdes-Perez has suggested searching the space of scientific models from the simplest to ones with increasingly more complexity, stopping at the first that fits the data. MECHEM uses this approach to find chemical reaction mechanisms [25]. Such orderings would be easy to encode as heuristic functions.

We extend these approaches by using an adjustable, explicitly mentioned heuristic function that does not require enumerating all possible models. Our approach is to generalize Thagard's scheme and place it on sounder theoretical footing.

3 Information Theory and Diverse Model Discovery

The MML criterion is to minimize the sum of the length of a theory and data given the theory. Some data will have a smaller combined compressed length than the original message. For example, the pitch and relative durations of some bird calls may be written in musical notation. This notation dramatically reduces the information from the original time-dependent air-pressure signal that the bird produced. However, many sounds are not appropriately described by musical notation (*e.g.*, human speech). The original time-dependent air-pressure signal will be a better representation than musical notation.

The equation that relates these terms for data set D; context c; discrete, mutually exclusive and exhaustive hypotheses $\{H_0, H_1 .. H_n\}$ with assigned prior probabilities $p(H_i|c)$; and computed conditional data probabilities $p(D|H_i,c)$ is:

$$-\log p(H_i|D, c) = -\log p(H_i|c) - \log p(D|H_i, c) + const \quad (1)$$

which is equation (2) of [5]. Recall that the *-log(p(choice))* is the Shannon lower bound on the information needed to distinguish *choice* from other possibilities. The constant term serves to "normalize" the probabilities and may be ignored if you only want their relative order. Cheeseman gives this iterative process for applying MML:

1. Define the theory space.
2. Use domain knowledge to assign prior probabilities to the theories.

3. Use Bayes' theorem to obtain the posterior probabilities of the theories given the data from adequate descriptions of the theories (*i.e.*, from descriptions that let you compute $p(D|H_i,c)$).
4. Search the space with an appropriate algorithm.
5. Stop the search when a probable enough theory has been found (subject to computational constraints), or to redefine the theory space or prior probabilities.

Several obstacles hamper efforts to apply MML to general scientific discovery. Among them are the specification of the initial theory prior probabilities, the inherently iterative nature of MML, and the difficulty in searching this space for a true "highest probability" theory.

Like other MML efforts, there is no good rule for specifying an initial set of prior probabilities. Although Cheeseman and others warn about using syntactic features, this may be the easiest approach to try in a new domain.

MML is an inherently iterative process of redefining theory spaces and prior probabilities. This complicates the usage of any function that needs calibration.

The scientific theory search space is expected to be highly irregular, hampering the search for the "best" model. This is true of other domains. Cheeseman suggests simulated annealing and the EM algorithm as potential search mechanisms.

4 Our Approach

We do *not* claim to have an optimal heuristic function in terms of returning the truly "best" model. Rather, our goal is to create a decent heuristic function that may help scientists on their initial searches with large, inhomogeneous spaces.

Good heuristics for real-world problems are often tricky to design [16]. We evaluate our function based on four criteria:

1. Generality over different sciences: We seek a function that is applicable to both primarily conceptual models as well as primarily numeric.
2. Ease of computation: The function should not rely too heavily on values that are computationally difficult to obtain. And, once it has its values, it should be rapidly computable.
3. Simplicity of form: There are several competing beliefs for how scientific models should be evaluated. The function's design should be as transparent as possible so that its assumptions are readily comprehended.
4. Smoothness: The function should give similar models similar scores.

We chose these criteria because they are important to our long-term goal of creating a general program to assist a variety of scientists.

Our contributions are the improvements in generality and ease of computation over Thagard's function. Generality is improved in three ways. First, it is adjustable to the tastes of a particular scientist. Second, it is able to handle degrees of inaccuracy. Lastly, it may use statistical arguments as well as proofs. Statistical arguments also improve the ease of computation: the function does not *have* to try to formally prove laws or data using perhaps an undecidable theory. The form of our function, however, is a little more detailed than Thagard's. The smoothness of both of our approaches critically depends upon how the user designs models.

Following Thagard, models have three components: a theory that specifies the details of the model, the data to predict, and a set of laws found from the data and predicted by the theory. The theory and the law set are both composed of assertions in some language. We use first order predicate logic with the data structure extensions of Prolog as our language in this paper. The distinction between which assertions are theory and which are laws is given by Lakatos. He distinguishes between commonly accepted knowledge (the "hard core", *i.e.*, theory) and between more tentatively held knowledge (the "auxiliary hypotheses", *i.e.*, laws) of a given research program [12][13]. The auxiliary hypotheses are the statements that are not commonly held (*i.e.*, have lower prior probability), and are the main objects that are manipulated during Kuhnian normal scientific discovery [10]. The data is assumed to be in tabular form with associated uncertainties and error bars.

It is simplest to assume that:

1. all measurements are independent of each other,
2. the data influence the choice of law set, and
3. the law set influences the choice of theory assertions.

Figure 1 depicts these assumptions graphically as a Bayesian network.

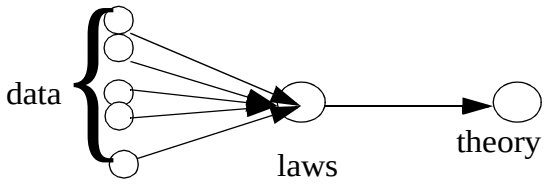


FIGURE 1. Bayesian network underlying the relationship between data, laws and theory

We are interested in the most probable total model. We derive the following starting from the Bayesian network of figure 1. Let T denote theory, LS denote a set of laws, and D denote data below:

$$p(T, D) = p(T|D) \cdot p(D) \tag{2}$$

Using Bayes' rule we may re-write this as:

$$= \sum_i p(T) \cdot p(LS_i|T) \cdot p(D|LS_i) \tag{3}$$

The last expression sums over all law sets and is appropriate when there may be disagreement over which law set is best (*e.g.*, several scientists combining their beliefs). However, for an individual scientist, a particular law set may appear much more probable than any of its competitors. In this case we may simplify the expression to:

$$p(T, D) = p(T) \cdot p(LS|T) \cdot p(D|LS) \tag{4}$$

Now we consider the meaning of each term.

The first term of equation 4 tells us the *a priori* probability of a theory, without reference to the law set or data. It encodes the biases on theories. It may be used, for example, to prefer one type of assertion over another. A commonly mentioned bias in science is one for syntactic *simplicity*, which is often measured as the length of an expression in a given language. This first term is the natural place to encode such a bias because this common measure of simplicity is only a function of the length of the expression.

$$p(T) = -\log 2(s(T)) \tag{5}$$

The function s(T) returns a measure of the size of T in some language. The function p(T) uses Shannon information theory to convert from a size to a probability.

We admit that the syntactic length metric is crude. We welcome scientists to redefine p(T) as they choose based upon their own domain knowledge. In defense of this initial estimate of p(T) we note that syntactic metrics: (1) are easy to compute, (2) are well agreed upon as being relevant (if not completely correct), (3) are common to many or all sciences (as opposed to symmetry, for example, which enjoys larger support among physicists than among other scientists), and, (4) would favor syntactically simple theories, which may be easier to comprehend. The last point is especially relevant for *initial* probability distributions, which may return several interesting model space regions that scientists must understand before determining if they warrant further exploration.

The second term tells us how likely the assertions of the law set are given the theory that we have chosen. At one extreme, if all laws are logically entailed by the theory, the term is 1.0 because they must be true (given the theory as premises). It is also 1.0 if the law set is empty because the theory is used to directly compute the data. At the other extreme, the term must be 0.0 if the theory contradicts any statement of the law set. Values in between signify that the law set may or may not follow, depending on specific values of free parameters in the theory. Free parameters are values that the theory refer to that do not have definite values, but distributions over sets of values. Examples include coefficients with standard deviations, and random numbers used during stochastic experiments. In these cases, the second term is set equal to the fraction of the free parameter space in which all of the statements of the law set are found to hold. For random numbers it will be more practical to estimate this value by sampling the space. Laws are limited to refer to the theoretical terms introduced in theories.

The third term measures empirical support and the degree of data coverage by telling us how likely the data are given the statements of the law set and theory. The same extremes hold when all of the data are logically entailed or some of it is contradicted by the law set or theory. Again, values in between 0.0 and 1.0 represent the fraction of the free parameter space in which the data are observed. Statistical assertions have an implicit free parameter that tells from which data set the statistic was collected. For example, consider two integers, each in the set [0..9], with an average value of 1. The implicit free parameter must denote one of three sets: (1,1), (0,2) or (2,0).

Please consider this (propositional) example. Let our theory be the assertion "a->b", our law be "a" and our data be two occurrences of "b." We would pay the appropriate (perhaps syntactic) price for the theory. The law is not derivable from the theory, so we set its probability to p(a) (the *a priori* probability that free variable A which ranges over "a" and "not(a)" actually is "a"). From our theory and law we may deduce our data with probability 1. If, however, we add assertions "c->a" and "c" to our theory then we have (perhaps) increased theory cost, but the law is now deducible from theory. Thus, the law has probability 1 and has no cost.

A problem with the heuristic function as given is that it has no parameters to be tuned to a particular scientist's preferences. This implies that it always returns the same value for the same arguments. This contradicts our goal of not imposing one ideal form on all scientific models.

Scientists should be able to fine tune the heuristic function, but any adjustment should be general enough to be applicable to all models. Further, we want the number of parameters to be relatively small, both because it will make the function easier to calibrate and because we want to guard against potential abuse by choosing a set of parameters that happen to make one model score well and a similar one score poorly. Our solution was to generalize the function in the following manner:

$$h_{tm+}(T, LS, D) = p(T)^{A} \cdot p(LS|T)^{B} \cdot p(D|LS)^{C} \tag{6}$$

The "tm" signifies that the function is over total models (*i.e*. theory, law set and data) and the "+" reminds us that this a function to maximize (*i.e*., larger values are better). The three parameters A, B and C allow us to independently vary the relative weights of the *a priori* model probability, the law set probability and the data probability.

Instead of maximizing probability, we may view it as minimizing information:

$$h_{tm-} = A \cdot s(T) - (B \cdot \log 2(p(LS|T))) - (C \cdot \log 2(p(D|LS)) \tag{7}$$

The "-" subscript denotes that this function should be minimized.

Equation 7 generalizes original MML equation 1 in two ways. First, equation 1's $-\log p(D|H_i,c)$ has been split into two terms, one for both the law set and the data. Both are graded probabilistically. Second, the coefficients A, B and C act as linear weights for the information terms. The linear weights may seem to grossly over generalize equation 1, but it really depends on how they are used. This is discussed in more detail in the next section.

There are two advantages to this weighing approach. First, it conforms to our notions that some sciences value theory conciseness and hard predictions more than others. Set the values of A and C higher in these sciences. Second, it does not allow arbitrary and contrived exceptions to make two similar total models score significantly differently.

Although we have offered a syntactic feature-based approach to specifying a theory's a prior probability, we have not limited scientists to use our function. Further, we admit that this is an iterative approach where probabilities are refined.

Revisiting our criteria we find:

1. Generality is achieved with the adjustable weights, the usage on probabilities of laws instead of counts of "explained facts", the usage of prior distributions instead of "co-hypotheses", and the potential use of proofs or statistical arguments.
2. The ease of computation is limited by our proof or statistical argument method, not by the heuristic.
3. Simplicity is achieved because the form is of a weighted sum with terms for theory, law and data.
4. Smoothness is achieved because lumping all theory together, all laws together and all data together hampers a user's ability to create one model that scores well and another very similar one that scores poorly.

Further generalizations of h_{tm+} and h_{tm-} may be envisioned. Each of the coefficients A, B and C may split into several coefficients $A[1..n_1]$, $B[1..n_2]$ and $C[1..n_3]$. These finer-grained coefficients may be used to weigh specific aspects of the theory (*e.g.* A[1] for equations, A[2] for decision trees, *etc*.), specific laws of the laws set (*e.g.* B[1] for equations, B[2] for simple logical assertions, *etc*.), and specific types of data (*e.g*. C[1] for spatial measurements, C[2] for temporal measurements, *etc*.)

Using the finer-grained coefficients is justifiable in some cases, like when there are large differences in the precision. For example, in seismology, earthquake times are known with very high precision: to within a few seconds per century. Earthquake locations are known with less precision: to only within tens of kilometers per 40,000 km (the Earth's circumference). Earthquake energies are known with far less precision, frequently only to an order of magnitude. We may want to weigh each type of data separately, taking into consideration how much precision is given and how much we want this data fit at the expense of other data.

Parameters A, B and C from equations 6 and 7 were not subdivided to simplify analysis and presentation.

5 Experiments and discussion

This section discusses the rough calibration of the heuristic function to models in two sciences. Geophysics and sociology were chosen because they cover a broad spectrum of acceptable scientific models.

We do *not* evaluate this function by comparing its output with that of IDS, PI, 49er, or Mechem. Which model a scientist believes in given specific data is, at least to some degree, subjective. Rather, we seek a method of calibrating our heuristic such that if it is given examples of models that users like then it can prefer similar models in the future.

The heuristic function's parameters may be calibrated for each science by analyzing its accepted models. Although there are three parameters, we only care about are their relative values. Accordingly, we may set A to 1 and let B and C vary. Equivalently, borrowing from physical chemistry, we can plot B/A versus C/A to create a "phase diagram" that tells which of the various total models are preferred by the heuristic. Each phase diagram constrains the area of each scientific model. This in turn constrains B/A and C/A for all models.

Comparing B/A with C/A makes the linear weights of equation 7 a conservative generalization of equation 1. The plots are primarily a comparison between B and C, and represent a value judgement on how much scientists want their uncertainty in the laws rather than in the data. There is no "correct" answer to this question. As we will see, it varies from scientist to scientist. This also strengthens our argument for an adjustable heuristic function.

If a scientist prefers model X then that scientist should set the parameters to where X is preferred. If the scientist is strongly tempted by model Y, then the scientist should adjust the parameters to be in the region of X but leaning towards that of Y. The scientist may iteratively update the parameter values as new models are evaluated by both the scientist and the heuristic.

Please recall our limited goal: to do an initial search in a large and inhomogeneous space for areas that contain potentially promising models. We do not promise the best models. Also, this may be an iterative process where theory prior probabilities are revised according to previous results.

The Knowledge Base and How It Predicts

The experiments were designed for a variant of the knowledge base discussed in [20]. The knowledge base has two lists of assertions, one for the theory and one for the laws. These assertions describe a standard **is_a** frame hierarchy of knowledge. Assertions may be frame inheritance statements, equations or Prolog-like logic sentences. A Prolog-like resolution engine drives inference, but dedicated code handles frame inheritance and equations for efficiency.

The output of the knowledge base to a given query is either an answer, or FAILURE, signifying no prediction is possible. An information cost accrued by the data when a prediction is wrong or missing. For symbolic values this cost is the Shannon information cost of the prior probability of the recorded answer. Thus, the default model to try to beat is the product of the prior probabilities of each datum. For integers and fixed and floating point values the cost is:[1]

(8)

$$-\log 2(DistinctValDiff(predict, record) + 1)$$

where **DistinctValDiff**() returns the number of distinct, representable values between the predicted and recorded values in the attribute's given precision. (For example, if an attribute was limited to multiples of 0.1 then **DistinctValDiff**(0.2,0.4) is 2.) When **predict** is missing then the function is set to its highest value for that attribute.

Sociology Data

1. Equation 8 corresponds to the last term of equation 7. It defines a maximal probability at the recorded value, and exponential decaying probability above and below that value. This distribution may be replaced by others and is not a critical aspect of this approach.

This technique requires large amounts of calibration data. We focused on models of family structure because United States Census data on family structure are readily available [4].

Data are not available for specific individuals, but they are summarized in several tables. From these summaries the number of families with 1, 2, 3, 4, 5, and 6 or more "own children" may be calculated for each family type. The family types are married family, male-householder family, female-householder family, married subfamily, male-householder subfamily and female-householder subfamily. Additionally, the number of childless families (but not subfamilies) may be calculated. The term "own children" means children related by birth, marriage or adoption. The U.S. Census Bureau switched from "head of house" to "householder" to emphasize the sharing of responsibilities prevalent in modern American families. The term "subfamily" refers to parent(s) who live with other adult(s) who are the householder(s) (*e.g.* their own parent(s).)

We randomly created a database of 10,000 people in proportion to the distribution of household types and number of children computed from the U.S. Census data. This database under represents the number of children a little because the U.S. Census data does not distinguish between 6 or more children. We treated such cases as exactly 6 children. It under represents the number of adults more because we made no attempt to include all cases of adults living with other adults. Our interest is only in predicting where children live as a function of their parents. The database lists each person, their address, and, when the person is a child, their mother and father. Children who did not live with their father got illegal values as their father attribute. This was also done for the mother attribute. All attributes are symbolic.

Sociology Models

After surveying ethnographic reports on 250 societies, Murdock came to the anti-climatic conclusion that the form of families in all societies is of ". . . a married man and woman with their offspring. [17]" (This is a *minimal* family structure because that unit may be embedded in larger structures.)

We take this statement as the theory. We encode it in the structure of the virtual relations of figure 2, augmented with some extra semantics. For example, from the structure of the database we may deduce that all families have one address, one childset, one mother, one father, that a set of children may have 0 or more children, *etc*. The additional rules allow members to inherit selected properties of their families. Predicate **prop(frame,attribute,value)** notes that property **attribute** of **frame** has value **value**.

family	address	childset	mother	father

child	childset	family

$$\forall(child(C) \wedge fam(F) \wedge prop(C, family, F) \wedge prop(F, A, V) \rightarrow prop(C, A, V))$$

$$\forall(fam(F) \wedge prop(F, mother, M) \wedge prop(F, addr, ADDR) \rightarrow prop(M, addr, ADDR))$$

$$etc$$

FIGURE 2. Codification of Murdock's theory

The laws operationalize the theory by making direct predictions about recorded values. For example, assume the child database included address information. We may then note a correlation between a child's address and that of their parent's.

$$\forall(child(C) \wedge mom(M) \wedge fam(F) \wedge prop(C, mom, M) \wedge prop(M, fam, F) \wedge prop(F, addr, A) \rightarrow prop(C, addr, A))$$

$$\forall(child(C) \wedge dad(P) \wedge fam(F) \wedge prop(C, dad, P) \wedge prop(P, fam, F) \wedge prop(F, addr, A) \rightarrow prop(C, addr, A))$$

FIGURE 3. Codification of potential Murdock laws (atoms *mother, father* and *family* have been abbreviated as *mom, dad* and *fam*)

The competing sociological model is due to Adams [1]. After examining Latin American and some ethnic societies, Adams concluded that the evidence for the nuclear families as described by Murdock was "marginal at best" [14]. Instead he proposed the mother-child dyad as the primary unit. This new model is created by removing the **father** attribute, or merely disallowing its use in proofs. We also delete the **father** law mentioned in figure 3 from the law set.

We bound the parameters by considering two unacceptable models at opposite extremes. The first is the "data" model. It uses neither theory nor laws to predict values. It merely reflects the prior probability of any one value. The second is the "theory" model. It explicitly memorizes each value individually as a statement in the theory. It has neither general statements nor laws, and overfits the data.

Table 1 gives the sizes of the each component of each total model. Both Murdock's and Adams' models must memorize adult addresses. Adams' must also memorize those of children who live with their fathers but not mothers. The law sentences in figure 3 logically follow from theory so they have size 0. Unfortunately, the zero size forbids the constraining of the B parameter by this experiment.

TABLE 1. Sizes of sociological models

Model	Abbr	Theory	Law	Data
data	d	0	0	107637
Adams	a	240	0	79582
Murdock	m	480	0	77739
theory	t	960960	0	0
Adams'	A	240	23429	77739

Figure 4a gives the "phase diagram" plot of data. Where a model out scores all others its abbreviating letter appears in the parameter space. log2(C/A) is plotted on the X axis and log2(B/A) on the Y.

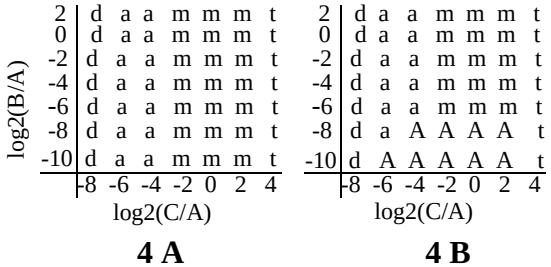


FIGURE 4. Sociology model "phase diagrams"

To place bounds on B we consider adding the **father** sentence to Adams' law set. However, we cannot prove it from our theory. Therefore, we accept **father** in the model as a free variable with its (data-specified) prior probabilities. This results in a model with the equivalent predictive power of Murdock's. It can now predict the addresses of children living with only their fathers. The price we pay is Shannon information cost of the prior probability of each usage of the **prop(Child,father,Father)** predicate for these predictions. See Adams' in table 1. The revised "phase diagram" with Adams' new model is plot in figure 4b.

Geophysics Data

We obtained data from the United States Geological Survey's National Earthquake Information Center. We retrieved all recorded earthquakes in the catalog in a rectangular box from 139E to 162E and from 41N to 55N from 1976 to 2000. The Kuril subduction zone, the Japanese island of Hokkaido, and the Kuril island chain are the most prominent geophysical features in this area. Non-tectonic events were removed and the remaining ones were fit to a great circle. This great circle was taken to be the "length" of the fault and events greater than 512 km from it were removed. The time, distance-along-fault, (signed) distance-from-fault and depth of the remaining 11031 events were entered into our earthquake database.

Geophysics Model

In the theory of plate tectonics, a *subduction zone* is a region where one (oceanic) plate sinks beneath another (continental) plate. A *Wadati-Benioff zone* is the seismically active portion of this interface [2][23].

A Wadati-Benioff zone may be modeled as a plane that increases in depth the further one goes into the continental plate. We did so by stating the assertions of figure 5 in the theory where the slope and intercept were found by least-squares fit.

$$DistFromfault = slope \times depth + intercept$$

inherit(kuril_quakes,slope,1.05682).
inherit(kuril_quakes,intercept, -85.9936 km).

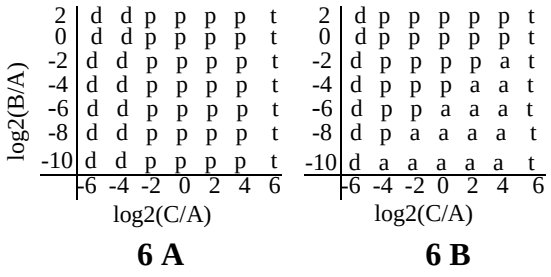
FIGURE 5. The theory of the planar Wadati-Benioff zone model.

The law set was left empty. As before, the "data" model did not try to predict, and the "theory" model overfit by memorization. The results are given in Table 2 and are plotted in Figure 6a.

TABLE 2. Sizes of geophysical models.

Model	Abbr	Theory	Law	Data
data	d	0	0	97750
planar	p	618	0	63904
theory	t	1369230	0	9775
aftershock	a	618	13759	63103

The non-zero entry for the theory model's for data size is due to round off error. That is, there is a slight difference between the decimal recording of the values logical assertions that comprise the theory (which have a fixed number of significant digits given by the precision of the values), and the binary recording of the values in the database.



6 A **6 B**

FIGURE 6. Geophysics model "phase diagrams"

To place bounds on B we add a law to the planar model. When a particular aftershock labelling procedure is used there is an average of a 43.5 km distance between an aftershock and its mainshock. Encoding this as a law permits better predictions of some distances. We include no theory to predict aftershocks, only an empirical procedure for labeling them after the fact. Therefore, we let **mainshock** be a free variable. The aftershock model results are given in Table 2 and in Figure 6B.

We now evaluate our heuristic with the criteria in section 4. Recall, they were (1) generality, (2) ease of computation, (3) simplicity of form, and (4) smoothness. The function is general because it was applied to symbolic sociology and numeric geophysics with equal ease, and because it has been applied to a domain where predictions have varying degrees of accuracy. Its ease of computation is limited by the ability to predict data, prove (or argue for) laws, and know data distributions. Also, its weighted sum form is simple.

The function's "smoothness," its ability to give similar models similar scores, is limited by how honest people are with the law set. When some condition is true over the whole parameter space one could move it from theory to laws to avoid paying the syntactic cost. This is against the philosophy of this approach. Also, trying to estimate data distributions when there is little data may be a serious problem. Distributions may be used as "fudge factors" to vary a model's score on the B/A axis. However, a potential advantage is that it will force such assumptions to be explicitly stated.

We do not argue for one particular ratio for C/A or B/A. Rather, we seek a method for calibration. That said, we note that both geophysics and sociological had similar C/A bounds. Having B be too great may lead to "overfitting" the laws to the theory and ruling out yet unknown secondary effects. For discovery it may be best to fix A and C and let B vary as the model becomes more refined. This is another study.

Note that this was truly a test of scientific *rediscovery.* Both the sociology and the geophysics theories were applied to new data. Neither Adams nor Murdock were trying to fit U.S. demographics for 1998. Benioff stated his hypothesis after examining events from S. America and Hindu-Kish, not the Kurils. (Wadati probably had data for Honshu, not the Kurils.)

6 Conclusion

Scientists have different opinions on what the same data entails. To ignore that is to ignore the history of science. We have developed a heuristic function that takes some of these differences into account, and may be calibrated to a particular scientist, along our given axes. This heuristic function is a generalization of single model family parameter finding MML. It generalizes MML in a principled fashion to consider how much faith to put in laws versus data. Our approach also extends [23] to be applied to scientific discovery. It is general and has been applied to both symbolic and numeric scientific models.

We do not claim to have solved the whole scientific model preferencing problem. Serious limitations remain including (1) the specification of the original model prior probability, (2) the inhomogeneity of the search space, and (3) the fact that the "most probable" model is not necessarily the best one. The purpose of this heuristic is to help scientists identify interesting regions in the model space, *i.e.*, models that are the immediate neighbors of their favorite models in the B/A-C/A plots. This is an initial step of an iterative process.

Computer scientists might believe that a heuristic function could not sufficiently constrain search in a domain as rich as scientific discovery. However, the heuristic function is only part of the search algorithm. The search algorithm may employ rules to suggest when to apply scientific operators (*e.g.*, [11]), or may use metalearning to discover which operators are best in a particular domain. Preliminary results from rediscovery in geophysics show that rules and metalearning may be combined or employed separately to significantly speed scientific discovery [20].

Acknowledgments

I thank my geophysicist Larry Ruff for his patience, my former advisors John Laird and Nandit Soparkar, and the National Physical Science Consortium and the Rackham Merit Fellowship for funding.

References

1. Adams, R.N. 1960. An inquiry into the nature of the family. p 30-49 in Dole, G. and Carneiro, R.L. (eds.), *Essays in the Science of Culture: In Honor of Leslie A. White*. Thomas Y. Crowell. New York.
2. Benioff, H., 1948. Earthquakes and rock creep. *Geol. Soc. Am. Bull.*, 59, p. 1391.
3. Buchanan, B., Phillips, J. 2001. Towards a computational model of hypothesis formation and model building in science. *Model Based Reasoning: Scientific Discovery, Technological Innovation, Values*. Kluwer.
4. Casper, L., Bryson, K. 1998. *Current Population Reports: Population Characteristics: Household and Family Characteristics. March 1998 (Update)*. United States Census Bureau.
5. Cheeseman, P. 1995. On Bayesian model selection. In Wolpert, D. (ed.) *The Mathematics of Generalization: Proceedings of teh SFI/CNLS Workshop on Formal Approaches to Supervised Learning*. Addison-Wesley: Reading, MA.
6. Forbus, K., 1985, Qualitative process theory, in *Qualitative reasoning about physical system*s, D. Bobrow, ed., MIT Press: Cambridge, Mass.

7. Fuller, S. 1993. *Philosophy of Science and its Discontents, Second Edition*. Guilford Press, New York.
8. Georgeff, M.P. and Wallace, C.S. 1984. A general selection criterion for induction inference. In *Proceedings of the European Conference on Artificial Intelligence*, p. 473-482. Elsevier: Amsterdam.
9. Korf, R.E. 1988. Search: A Survey of recent results. In H.E. Shrobe (Ed.), *Exploring Artificial Intelligence: Survey Talks from the National Conferences on Artificial Intelligence* (pp. 197-237). Morgan Kaufman.
10. Kuhn, T. 1962. *The Structure of Scientific Revolutions*. University of Chicago: Chicago.
11. Kulkarni, D. and Simon, H. 1988. The processes of scientific discovery: the strategy of experimentation, *Cognitive Science*, vol. 12, p. 139-175.
12. Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In Lakatos, I. and Musgrave, A. (ed.) *Criticism and the growth of knowledge*. Cambridge University Press: Cambridge.
13. Lakatos, I. 1971. History of science and its rational reconstructions. In Buck, R.C. and Cohen, R.S. (ed.) *Boston Studies in the Philosophy of Science*. vol 8, p 91-135. Reidel: Dordrecht.
14. Lee, G. 1977. *Family Structure and Interaction: A Comparative Analysis*. J.B. Lippincott. Philadelphia.
15. McAllister, J. 1996. *Beauty and Revolution in Science*. Cornell University: Ithaca.
16. Michalewicz, Z., Fogel, D. 2000. *How to Solve It: Modern Heuristics*. Springer-Verlag. Berlin.
17. Murdock, G.P. 1949. *Social Structure*. The Free Press. New York.
18. Nordhausen, B., Langley, P., 1987, Towards an integrated discovery system, in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Milan, Italy.
19. Nordhausen, B., Langley, P., 1990, An integrated approach to empirical discovery, in Shrager J, and Langley, P. (ed.) *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, San Mateo.
20. Phillips, J. 2000. *Representation Reducing Heuristics for Semi-Automated Scientific Discovery.* Ph D. Thesis, University of Michigan.
21. Rissanen, J. 1978. Modeling by shortest data description. *Automatica*, 14, p. 45-471.
22. Sleep, N., Fujita, K. 1997. *Principles of Geophysics*. Blackwell Science. Malden.
23. Thagard, P. 1988. *Computational Philosophy of Science*, MIT Press, Cambridge MA.
24. Wallace, C.S., and Freeman, P.R. 1987. Estimation and inference by compact encoding. *J. Roy. Stat. Soc., Series B*, 49, p 233-265.
25. Valdes-Perez, R. 1995. Machine discovery in chemistry: new results. *Artificial Intelligence*, 74(1), p 191-201.
26. Zembowicz, R. and Zytkow, J. 1996. From contingency tables to various forms of knowledge in databases, in: *Advances in Knowledge Discovery and Data Mining*, Fayyad et al (eds.) AAAI Press, San Mateo.
27. Zytkow, J. and Zembowicz, R. 1993. Database exploration in the search for regularities, *J. Intelligent Information Systems*, 2:39-81.