Intelligent machines in the 21st century: automating the processes of inference and inquiry

By Kevin H. Knuth

Computational Sciences Div., NASA Ames Research Center, Moffett Field, CA

The last century saw the application of Boolean algebra toward the construction of computing machines, which work by applying logical transformations to information contained in their memory. The development of information theory and the generalization of Boolean algebra to Bayesian inference have enabled these computing machines, in the last quarter of the twentieth century, to be endowed with the ability to learn by making inferences from data. This revolution is just beginning as new computational techniques continue to make difficult problems more accessible. However, modern intelligent machines work by inferring knowledge using only their pre-programmed prior knowledge and the data provided. They lack the ability to ask questions, or request data that would aid their inferences.

Recent advances in understanding the foundations of probability theory have revealed implications for areas other than logic. Of relevance to intelligent machines, we identified the algebra of questions as the free distributive algebra, which now allows us to work with questions in a way analogous to that which Boolean algebra enables us to work with logical statements. In this paper we describe this logic of inference and inquiry using the mathematics of partially ordered sets and the scaffolding of lattice theory, discuss the far-reaching implications of the methodology, and demonstrate its application with current examples in machine learning. Automation of both inference and inquiry promises to allow robots to perform science in the far reaches of our solar system and in other star systems by enabling them to not only make inferences from data, but also decide which question to ask, experiment to perform, or measurement to take given what they have learned and what they are designed to understand.

Keywords: inference, inquiry, probability, entropy, Bayesian methods, lattice theory, machine intelligence

1. Introduction

James Bernoulli (1713) was among the first to realize the difference between deductive logic used in situations of certain knowledge and inductive logic, which is necessary for the uncertain situations found in everyday problems. In *Ars Conjectandi*, (*The Art of Conjecture*), he was the first to quantify uncertainty by identifying a set of equally possible hypotheses. This allowed him to calculate the number of ways in which a given a situation could occur relative to the total number of possible outcomes. He also recognized that what we perceive as chance events could be interpreted as regular predictable events if we were more knowledgeable: "The chance

depends mainly upon our knowledge.". As we will see, this statement reflects what we consider to be a modern view of probability as extended logic.

While Bernoulli became adept at enumerating possibilities and using them to calculate probabilities, he was unable to use the outcomes to make inferences about the way in which an observed situation could have occurred. Reverend Thomas Bayes (1763) turned the situation around and made inferences about the causes using the outcomes. While this rule still carries his name, it was Pierre-Simon Laplace who independently rediscovered Bayes' Theorem (1812), presented it in its modern form, and went on to utilize it to solve problems in astronomy, geodesy, instrumentation, error estimation, population, jurisprudence, and procedures of electoral bodies. Laplace's interpretation of Bayes' Theorem as an extension of logic led directly to his extremely prolific application of the methodology, and can be summed up in a translated quote from Théorie Analytique des Probabilitiés: "Probability theory is nothing but common sense reduced to calculation."

After Laplace, the mathematicians of the nineteenth century worked to rigorously develop probability theory. As a general theory of inference, it was too difficult to derive useful theorems. So the range of applications of the theory was reduced to relatively simple problems involving frequencies of event occurrences, which consequently led to *frequentist statistics*—a field which continues to confuse students with a bewildering array of statistical tests. It is quite amazing how the specificity of modern frequentist methods stands in such stark contrast to the generality of this theory at its conception. While others, most notably Jeffreys (1939), attempted to resurrect the general theory of inference, the renaissance would have to wait until some key insights were made in the mid-twentieth century.

The information technology revolution in the last half of the twentieth century was due in great part to Claude E. Shannon's masterpiece 'The Mathematical Theory of Communication' (1948). In it he single-handedly developed what is now called *information theory*, which has driven telecommunications, coding theory, signal analysis, and computer science as a whole. Key to this development was the concept of *information-theoretic entropy*. It was designed for use in communications, and can be thought of as a measure of the degree of uncertainty of which message, from a set of possible messages, will be received in a communication channel. The name entropy, however, has created much confusion. Myron Tribus (1971) recalls Shannon explaining how von Neumann suggested that he should call his measure 'entropy' because the same function was already employed in statistical mechanics, and more importantly, that "nobody knows what entropy really is, so in a discussion you will always have an advantage." Much confusion ensued due to the fact that Shannon's application of information-theoretic entropy was so specific, yet it was so similar to the poorly-understood entropy in use for over 60 years in physics. The great insights of the next character in our story clears up these mysteries.

While Shannon demonstrated how one could use the probabilities of a set of messages to compute the degree of uncertainty, Edwin Thompson Jaynes computed probabilities based on a maximal degree of uncertainty. In this way he developed the *Principle of Maximum Entropy* (Jaynes 1957, 1979), which allows one to assign probabilities in the event that one possesses some knowledge in the form of constraints. This allows entropy to be used as an inference tool. By maximizing the entropy subject to these constraints, one obtains a set of probabilities that are as noncommittal as possible while agreeing with what is known. Furthermore, Jaynes

(1957) showed that this is precisely the situation encountered in statistical mechanics, where the solutions are those which maximize the entropy subject to constraints such as the total energy of the system. Application to thermodynamics was further developed and established by Myron Tribus (1961, 1971). Thus not only is the information-theoretic entropy related to the thermodynamic entropy, but the laws of statistical mechanics were demonstrated to be processes of inferential reasoning rather than physical laws descriptive of the system itself.

The last key insight, on which we will expound in this paper, was made by Richard Threlkeld Cox, who saw that the sum and product rules of probability could be derived from Boolean logic (Cox 1947, 1961). This was done by generalizing Boolean implication among logical statements to degrees of implication represented by real numbers. Cox's insight was key as it provided the first rigorous proof of probability theory as an extension of logic. Jaynes recognized this and became a strong proponent of probability theory as extended logic and the basis for scientific reasoning, which he advocated in his tome, *Probability Theory: The Logic of Science* (Jaynes 2003) to be posthumously published this year. This new perspective on information-theoretic entropy and probability theory implies that the true information revolution (Solana-Ortega 2001) has only just begun.

While the necessary framework laid in the middle of the twentieth century led to developments other than statistical physics, such as the Burg algorithm for spectral analysis (Burg 1967), application of the Bayesian methodology to more general scientific problems had to wait until the availability of sufficient computing power in the late 1970's and early 1980's. It was not until this time that the methodology could truly prove its worth by outperforming standard techniques in many areas of research. Furthermore, inspired by Cox's success in deriving the sum and product rules of probability from Boolean logic, much effort has gone into better understanding the foundations of probability theory and its relationship to another uncertainty-based area of physics—quantum mechanics. The following sections will introduce a modern picture of this foundation, demonstrate its application toward the automation of inference and inquiry, and look forward to the possibilities this methodology affords.

2. Posets and Lattices

In this section we review the ideas behind partially ordered sets and lattices. This modern viewpoint (Davey & Priestley 2002), will allow us to easily relate the study of inference to the study of inquiry. The key concept required for this development is that we can take a set of objects and an appropriate ordering relation, and *partially order* the objects in the set forming what is called a *partially ordered set* or *poset*. We say *partially* order the objects, because it *may* be that some of the objects in the set are *incomparable*—like apples and oranges. As an example, consider the powerset of $\{a, b, c\}$, which is the set of all possible subsets, written as $\wp(\{a, b, c\}) = \{\varnothing, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$. We can order this set nicely with the ordering relation '*is a subset of*', written for example as $\{a\} \subseteq \{a, b\}$. As we mentioned, this is a partial ordering as some elements, such as $\{a\}$ and $\{b, c\}$, are incomparable as neither one is a subset of the other.

An important insight here is that for any given set of objects, there may be different ordering relations that can be used giving rise to different posets. To
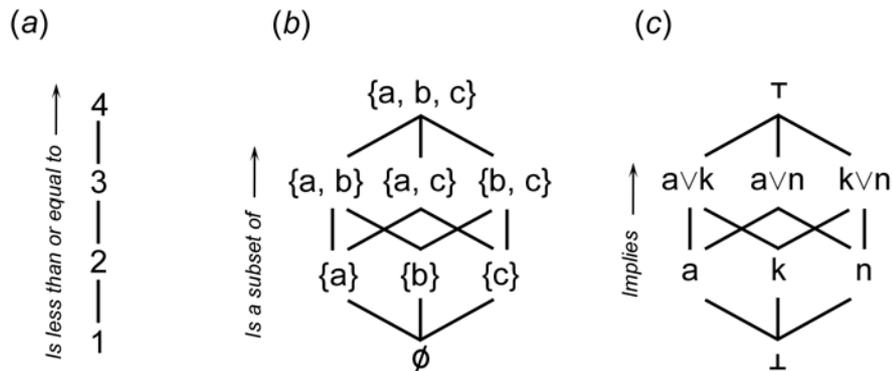
Figure 1. Diagrams of lattices described in the text. a. Natural numbers 1, 2, 3, 4 ordered by 'less than or equal to', b. All subsets of $\{a, b, c\}$ ordered by 'is a subset of', c. Logical statements ordered by implication.

generally express an ordering relation, we use the symbol $\leq$ so that $a \leq b$ is read as '*b includes a*'. In our powerset example, $\leq$ represents $\subseteq$. In the event that $a \leq b$ and $a \neq b$, we write $a < b$ and read '*a is properly contained in b*'. Last, we can think of this partial ordering as imposing a hierachy on the set of elements. When $a < b$, but there is *no element* x such that $a < x < b$, then we can write $a \prec b$, which is read as '*b covers a*'. In this case b is an immediate superior to a in the hierarchy generated by the ordering relation. Another example is the set of numbers $\{1, 2, 3, 4\}$ ordered by the usual '*less than or equal to*'. In this poset, 3 covers 2 as $2 < 3$, but there is no number x in the set where $2 < x < 3$.

The concept of *covering* allows us to illustrate the structure of the poset. First if $a < b$ then b is drawn *higher* than a in the diagram. Second, if b covers a, $a \prec b$, then we *connect* a and b with a line. Figure 1 shows the diagrams for our posets $(\{1, 2, 3, 4\}, \leq)$ and $(\wp(\{a, b, c\}), \subseteq)$. Picking any element on the diagram of $(\wp(\{a, b, c\}), \subseteq)$, one can immediately identify which elements contain it as a subset by following all of the lines upward from that element. Similarly, one can find all of the elements that it contains by following the lines downward.

In figure 1b, if you choose two elements in the diagram, say $\{a\}$ and $\{b\}$, and follow the lines upward, the first common element that includes both $\{a\}$ and $\{b\}$ is $\{a, b\}$, which is called their *join*. The join of elements x and y is written generally as $x \vee y$, thus $\{a\} \vee \{b\} = \{a, b\}$. Dually, if we choose two elements, say $\{a, b\}$ and $\{b, c\}$, we find that the first common element that they both include is $\{b\}$, which we call their *meet*. The meet of two elements x and y is written $x \wedge y$. In the powerset example, the join of two elements can be found by taking their set union, and their meet can be found by taking their set intersection. However, join and meet correspond to other operations in other posets.

If the meet and join always exist, and are commutative, associative, idempotent, and obey the absorbtion law, then the poset is called a *lattice*. Associated with each lattice is an algebra. By focusing on the hierarchical arrangement of the elements in the poset, one sees the structure as a lattice. Whereas by focusing on the join and meet as operations applied to its elements, one sees the structure as an algebra.

3. Boolean Lattices

We now examine George Boole's contribution to logic (1854) from the perspective of lattices. Consider a lattice based on a set of logical statements and ordered with the relation 'implies', which we write as $a \to b$. Thus in this lattice $\leq$ represents $\to$. The statements in the set will be generated from a smaller set of exhaustive, mutually exclusive statements. By exhaustive we mean that at least one of them is true; whereas by mutually exclusive we mean that if one of them is true, then the others are false. This gives us a basis set of statements (generally called *atomic elements*) where we are assured that one and only one is true. As an example consider the possible atomic hypotheses representing accusations of who stole the tarts made by the Queen of Hearts. †

$$a = \text{'Alice stole the tarts!'}$$
$$k = \text{'The Knave of Hearts stole the tarts!'}$$
$$n = \text{'No one stole the tarts!'}$$

We can construct new logical statements by combining two statements in two different ways. First, the *disjunction* of two logical statements is a proposition that says what the two say jointly. One can think of the disjunction as being represented by the word 'or'. By disjoining a and k above, we obtain a new statement, which can be shown to be the join $a \vee k$, that says '*Either Alice or the Knave of Hearts stole the tarts!*'. Notice that if '*Alice stole the tarts!*' is true, then it implies that the disjunction is also true. Thus '*Alice stole the tarts!*' is included in '*Either Alice or the Knave of Hearts stole the tarts!*' so that $a \to a \vee k$. The second operation, the *conjunction*, is a statement that tells what the two statements tell in common. This is represented by the word 'and', and is given in the lattice by the meet operation. Thus the conjunction of two logical statements a and b is $a \wedge b$. Because the logical symbols for disjunction $\vee$ and conjunction $\wedge$ are identical to, and in *this lattice* signify the same operations as, the join and the meet respectively, it is important to remember that the meaning of the symbols for the join $\vee$ and the meet $\wedge$ depend on the particular lattice.

Figure 1c shows the lattice diagram for all the possible disjunctions of the three atomic statements. Notice that implication is directed upward, so that if any lower element is known to be true all the connected elements above it are also known to be true. Working with the truth values of propositions on this lattice is called deductive logic. This type of lattice structure is called a *Boolean lattice*, and its associated algebra is a *Boolean algebra*. Boolean lattices have another interesting property—for every element x, there exists another unique element x' called its *complement*, such that $x \vee x' = \top$, where $\top$ is the top element formed by the disjunction of all the atomic elements, and $x \wedge x' = \bot$, which is the bottom element formed by their conjunction. Note also, that our powerset lattice has the same structure as the lattice of logical statements with the ordering relation $\to$. They are both Boolean lattices and both have operations which follow a Boolean algebra.

† Chapters XI and XII in *Alice's Adventures in Wonderland*, Lewis Carroll, 1865. Lewis Carroll is of course a pseudonym for Charles Lutwidge Dodgson also a logician who did important work in symbolic logic.

4. Derivation of Probability from Logic

Cox's contribution (1947, 1961) was to generalize logical implication to degrees of implication represented by real numbers. This allows one to talk about the degree to which a statement a implies a statement b, written as $(a \rightarrow b)$. To ease the transition to probability, we will write $(a \rightarrow b)$ as a function $p(b|a)$ of the assertions a and b. The first goal is to figure out how to calculate the degree to which a premise i implies a conjunction of two statements $a \wedge b$. Cox assumes that this is a function of the degree to which i implies a and the degree to which $a \wedge i$ implies b

$$p(a \wedge b|i) = F[p(a|i), p(b|a \wedge i)] \tag{4.1}$$

where $F[\cdot, \cdot]$ is a function to be determined. This function will tell us how to do the calculation. It is found by maintaining consistency with Boolean logic. If we consider a statement formed from the conjunction of three propositions $a \wedge b \wedge c$, we can use associativity of the lattice to write this two ways: $(a \wedge b) \wedge c$ or $a \wedge (b \wedge c)$. Then we can use 4.1 above to rewrite the degree $p(a \wedge b \wedge c|i)$ two different ways in terms of F. Consistency requires that they are equal

$$F[p(a \wedge b|i), p(c|(a \wedge b) \wedge i)] = F[p(a|i), p((b \wedge c)|a \wedge i)]. \tag{4.2}$$

Writing $p(a \wedge b|i)$ and $p((b \wedge c)|a \wedge i)$ above in terms of F, and substituting $x = p(a|i)$, $y = p(b|a \wedge i)$, and $z = p(c|a \wedge b \wedge i)$, we get a simple *functional equation*

$$F[F[x, y], z] = F[x, F[y, z]], \tag{4.3}$$

which has as its solution † $F[x, y] = xy$. So that from (4.1) we have the familiar *product rule* of probability theory

$$p(a \wedge b|i) = p(a|i)p(b|i \wedge a), \tag{4.4}$$

which we see is required by consistency with associativity (Smith & Erickson 1990).

The *sum rule* of probability is derived similarly by noting that the degree to which the premise i implies the complement of a statement a' depends on the degree to which i implies the original statement a: $p(a'|i) = G[p(a|i)]$. However, the complement of the complement of a statement is the original statement so we have the functional relation $p(a|i) = G[G[p(a|i)]]$, which leads to

$$p(a|i) + p(a'|i) = 1. \tag{4.5}$$

From this point on we will recognize that probability is a real number describing the degree of implication on a lattice of logical statements. Thus we have developed probability theory as a description of one's state of knowledge regarding logical propositions. As described in the introduction, this approach is much more general than frequentist statistics, which regards probability as representing the frequency of event occurrences.

Last, we consider commutativity of the conjunction, where $p(a \wedge b|i) = p(b \wedge a|i)$. Applying the product rule to both sides we get $p(a|i)p(b|a \wedge i) = p(b|i)p(a|b \wedge i)$. Solving for $p(b|a \wedge i)$ gives *Bayes' Theorem*

$$p(b|a \wedge i) = p(b|i)\frac{p(a|b \wedge i)}{p(a|i)}, \tag{4.6}$$

† There is a more general solution, but it only serves to set the scale and offset of the logarithm of the probability.

which allows one to write $p(b|a \wedge i)$ in terms of $p(a|b \wedge i)$ thus turning around the inference. It is important to realize that these rules are the only logically consistent way to manipulate probabilities. Any other rules will eventually lead to a contradiction violating logical consistency.

5. Automating Inference

The humble origin of Bayes' Theorem belies the power that this relation wields. First we consider a hypothesis about a situation we wish to understand. This hypothesis could be a simple statement as in the Stolen Tarts example, or it could be a compound hypothesis formed by taking the logical conjunction of several hypotheses. This is useful in science when one has a parameterized *model* of a situation. We can conjoin hypotheses describing the model like $h1 =$ '*Parameter $p = 2.9!$*', and $h2 =$ '*Parameter $q = 6.4!$*' to form $model = h1 \wedge h2$. In such a situation, we have a hypothesis space defined by all possible hypotheses we could consider. In addition to our hypothesis, we may have some acquired data $d =$ '*I measured r to have a value of 2.3!*'. The premise i represents our knowledge about the problem prior to obtaining new data. Rewriting Bayes' Theorem (4.6) by replacing a with *data* and b with *model* we get

$$p(model|data \wedge i) = p(model|i)\frac{p(data|model \wedge i)}{p(data|i)}. \tag{5.1}$$

The first term on the right $p(model|i)$, called the *prior probability* or *prior*, represents the degree to which we believe the model is correct given only our prior information i. The term in the numerator $p(data|model \wedge i)$ is called the *likelihood*, which represents the degree to which we believe that the situation described by the model could have resulted in the observed data. The term in the denominator $p(data|i)$ is called the *evidence* and it represents the degree to which we believe the data could have been observed based only on our prior information. Finally, the result on the left $p(model|data \wedge i)$ is called the *posterior probability*, which describes how our initial state of knowledge $p(model|i)$ is updated with the acquisition of new information. Bayes' Theorem is thus a learning rule that allows us to improve our state of knowledge as we gain new information!

Keep in mind that these probabilities are not to be thought of as frequencies of event occurrences, but rather they represent the degree to which one believes a logical statement is true. This results in a much broader range of application. To utilize this methodology we must assign values to the necessary priors and likelihoods that appear on the right-hand side of the equation. The assignment of priors often causes concern among those who practice frequentist statistics. However, they assign likelihoods, which they call sampling distributions, and neglect priors, which assumes that they are the same for all hypotheses. How to assign prior probabilities in a logically consistent manner is an area of ongoing study. Jaynes (1967) generalized Bernoulli's probability assignments by accounting for symmetries in one's state of knowledge, which can be used in conjunction with Maximum Entropy to incorporate known constraints (Jaynes 1957, 1979).

We will now demonstrate how this is used to develop a machine learning system that can use data to understand a physical system. We recommend that the inter-
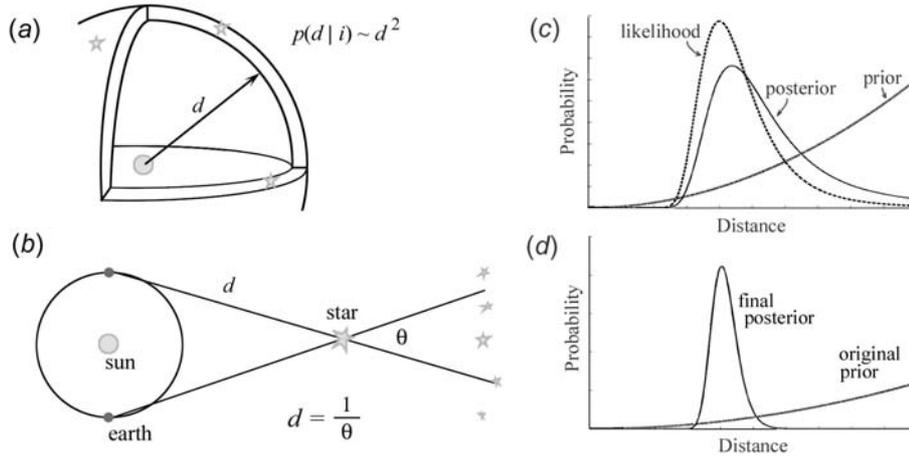
Figure 2. a. Stars uniformly distributed in space are more probable to be distant from earth than near. b. The origin of stellar parallax. c. The resulting posterior is narrower than the prior—we have learned something. d. After incorporating 5 new measurements, the posterior has grown in amplitude (compare to prior), reflecting the high probability of the solutions, and has narrowed, reflecting a more refined solution.

ested reader consult these practical works and tutorials (Bretthorst 1988; Loredo 1990; Hanson 1993; Skilling 1998; Dose 2002), and especially Sivia (1996).

(*a*) *Stellar Distance from Parallax*

Consider a nearby star to which we would like to determine the distance d. Hypothesizing a value for this distance will constitute our model of this situation. Before we begin thinking about measurements, we know that nearby stars are approximately uniformly distributed in space. Thus the star has an equal chance to be in any given volume element of space up to a maximum resolvable distance. These little volume elements of space are the equal probability cases of Bernoulli. However, we want a probability for the distance to the star. At a given d, the star must be in a thin shell with radius d. The volume of these shells gets bigger with d^2 (figure 2a). So a prior probability $p(d|i) \propto d^2$ reflects our expectation that stars are uniformly distributed in space.

As the Earth orbits the Sun, the star's apparent position in the sky changes with respect to the more distant stars (figure 2b). This measured angular position change θ is called the parallax and will constitute our data. Parallax is inversely related to distance $\theta = \frac{1}{d}$, with 1 milli-arcsecond of angle corresponding to 1 parsec (3.26 lightyears). We can use this relation to predict a value for the parallax given a hypothesized value for d. With some knowledge of the errors of our measurements we can write the likelihood $p(\theta|d \wedge i)$ as a Gaussian distribution centered about the parallax predicted by d. One can think of the likelihood in terms of the *forward problem* where we start with our model, and compute what it predicts we should observe. The difference between the prediction and the observed data is represented by the likelihood. In cases where no analytic equation exists (like $\theta = \frac{1}{d}$), complex simulations must be employed to make predictions from the model.

Using Bayes' Theorem, we plot the posterior probability (figure 2c), which is proportional to the product of the prior and likelihood, as a function of all the
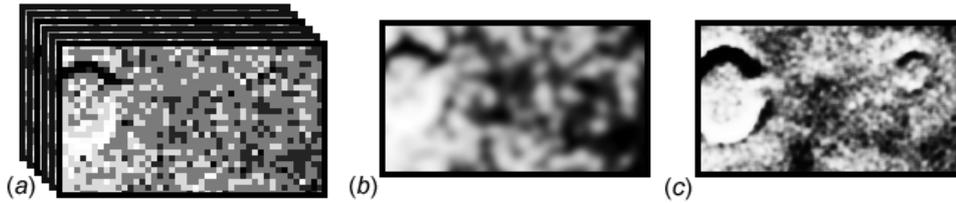
Figure 3. a. From a set of Viking Orbiter images of the Martian surface, b. the average reveals no new information, c. while the inferred high-resolution model reveals a crater not resolvable in an individual image.

possible values of distance. Plotted as a function of d the likelihood no longer looks symmetric, due to the inverse relation between d and θ. The important point is that the posterior is narrower than the prior, which means that we have ruled out possibilities and have learned something from the data. The posterior for the first datum point can now be used as a prior for a newly acquired second piece of data (figure 2c). Repeating this process several times results in a more certain value for the distance to the star (figure 2d). Bayes' Theorem thus allows us to automate an inference procedure, taking into account prior knowledge as well as new data.

We are currently working on a more complex version of this problem where we are estimating the distances to planetary nebulae, which are the outer atmospheres of sun-like stars that have been thrown off during their collapse. These clouds of gas expand in time and we can use multiple images taken over time along with the Doppler shifts in their spectral lines due to their expansion velocity to simultaneously model the three-dimensional structure of these objects while estimating their distances from Earth (Knuth & Hajian 2002).

(b) Super-Resolution Imaging of Martian Surface

In this example we discuss a project from Cheeseman *et al.* (1994) where multiple images of a planetary surface are used together to infer a super-resolved image. Figure 3a shows one image from a set of images taken by the Viking Orbiter of a particular area on the surface of Mars. These images were taken under similar lighting conditions, however the camera position and orientations were slightly different each time. To obtain a better picture of the surface, one is tempted to average together the images (figure 3b). However, one does much better by considering a *model image* with a much higher resolution described by a set of model pixels, or *mixels*. The idea is to use the Viking images as data to infer the most probable mixel intensities in our model while simultaneously inferring the camera properties for each Viking image. Such super-resolution imaging works because each pixel in each image is an independent datum point describing the intensity emitted by a small patch of the surface. Given hypothesized mixel intensities and camera properties, one can make predictions about the pixel intensities in the Viking images, which are represented by the likelihood function. The prior probability used reflects the expectation that neighboring mixel intensities are correlated. Using Bayes' Theorem the posterior probability for any super-resolved model can be computed, and using efficient search algorithms, the most probable model can be found (figure 3c). This technique is remarkable as it reveals a newly discovered crater!

6. Questions and the Logic of Inquiry

While the mathematics of inference has become well-understood in the twentieth century, we are only beginning to understand the mathematics of inquiry. Cox (1979), in his last work, defined a question as the set of all possible statements that answer it. To be assured that this set contains *every* statement that answers the question, the set must also include all statements that imply any statement already in the set, as these statements also answer the question. In lattice theory, such a set is called a *down-set* as it is generated by a given set of elements in the lattice and everything below them. Several key down-sets in the statement lattices are shown in figure 4.

Given this definition, two questions are equivalent if they are answered by the same down-set of statements. Two such questions are '*Is it raining?*' and '*Is it not raining?*'. They are both answered by the same down-set generated by the statements '*It is raining!*' and '*It is not raining!*', and thus are equivalent as they ask the same thing. Furthermore, we can impose an ordering relation on questions, as the set of answers to one question may be a subset of the set of answers to another. Consider the question $T =$'*Who stole the tarts made by the Queen of Hearts all on a summer day?*', which I will write concisely using the set (see the down-set in lower right corner of figure 4)

$$T = \{a = \text{`Alice stole the tarts!'} \qquad\qquad (6.1)$$
$$k = \text{`The Knave of Hearts stoke the tarts!'}$$
$$n = \text{`No one stole the tarts!'}\}$$

We can also consider the *binary question* $B =$'*Did or did not Alice steal the tarts?*', which can be written

$$B = \{a = \text{`Alice stole the tarts!'} \qquad\qquad (6.2)$$
$$a' = \text{`Alice did not steal the tarts!'}\}.$$

As the defining set of T is exhaustive, the statement '*Alice did not steal the tarts!*' is equivalent to the statement '*Either the Knave of Hearts or no one stole the tarts!*', written $a' = k \vee n$. As B is a down-set (see figure 4), it must contain all the statements that imply a', which are k and n. Thus the set T is a subset of B, and by answering the question T, we will have also answered the question B. The converse is not true as if we obtain as an answer to B that '*Alice did not steal the tarts!*', then we still will have not answered T.

At this stage, we are well on our way to constructing the lattice of all questions that can be asked relative to the issue '*Who stole the tarts?*'. With the ordering relation '*is a subset of*', or equivalently '*answers*', we can show that the conjunction or meet of two questions is the intersection of the down-sets of statements answering each question, so that $X \wedge Y \equiv X \cap Y$. This results in a question which asks what the two ask jointly, thus earning it the name of the *joint question*. Similarly the disjunction or join of two questions, called the *common question*, is formed from the union of the two down-sets of statements answering each question, $X \vee Y \equiv X \cup Y$, and as such asks what the two questions ask in common.

The lattice of questions can be formed by considering all the possible down-sets of the assertion lattice and ordering them appropriately. As writing out a question
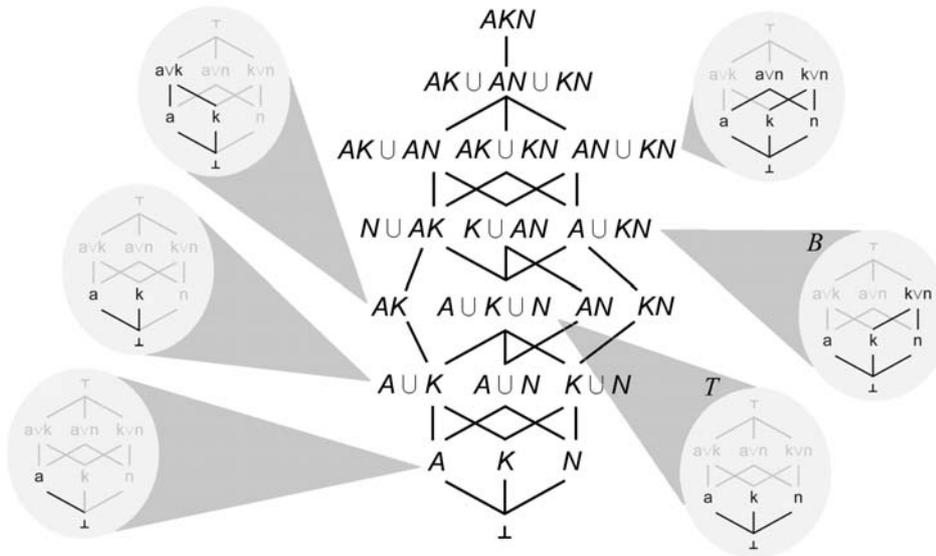
Figure 4. The lattice of all questions (center) generated by three mutually exclusive statements. Examples of the down-sets defining several questions are shown, including T and B discussed in the text (right). The ordering relation 'is a subset of' is applied to the down-sets of statements, so that lower questions answer higher questions.

in terms of its assertions can be quite lengthy, we will use the following notation where A represents the down-set formed by descending the assertion lattice from the element a (figure 4), AK represents the down-set formed from the element $a \vee k$ (figure 4), and AKN represents the down-set formed from the element $a \vee k \vee n$. The elements of the question lattice are then formed from all possible disjunctions of the questions A, K, N, AK, AN, KN, AKN. As an example, the binary question B in the example above can be written as $B \equiv A \cup KN$ denoting that its possible answers derive from a and $k \vee n$.

Figure 4 (center) shows the question lattice for the three mutually exclusive statements in our stolen tarts example (Knuth 2002). This lattice is not Boolean, which indicates that questions do not possess complements. This particular lattice structure is known as the *free distributive lattice*, and as it is associative and distributive, it possesses a measure *analogous to* probability, which following its own sum and product rules describes the degree to which one question answers another. For this reason it is called *bearing* or *relevance*. With this measure, one can compute the relevance that a question Q has on an outstanding issue I, denoted $b(Q|I)$. The notation, introduced by Robert Fry, represents an upside-down p reflecting the inherent relationship between relevance on the question lattice and probability on the statement lattice. Thus we are introduced to the calculus of inquiry.

While the exact relationship between probabilities on the statement lattice and the relevances on the question lattice are still being explored, the results we have obtained to date (Cox 1979; Fry 1999; Knuth 2002) suggest that relevance can be represented in terms of the entropy of the probabilities, and that the calculus of inquiry is a generalization of information theory. This has intuitive appeal as the probabilities then represent what is known, while entropies or relevances represent

what is not known. It also make sense from an information-theoretic standpoint where the design of a communication channel can be interpreted as the design of a question to be asked of the transmitter. This is also suggested by the relationships between the lattices (Knuth 2002) as the map from the statements to the questions acts like an exponential function, whereas the inverse map acts like the logarithm (Davey & Priestley 2002).

When probabilities of what is known can be used to compute relevances among that which we desire to know, we will be able to design machines that can identify missing information and request it. To do this, one will not need to draw all these elaborate lattice diagrams. All that will be necessary is the calculus of probability and relevance. This will be especially important in scientific problems where one is unable to ask a single question that will resolve an issue directly, and is instead limited to a given set of experimental questions.

7. Generalization of the Methodology

There are two very important realizations to be made. First, the concept of generalizing inclusion on a lattice to a degree of inclusion can be made on *any* kind of lattice. Thus we expect that there are other rules analogous to the sum and product rules we introduced here that exist in other disciplines. Ariel Caticha (1998) has shown how the sum and product rules can be derived from associativity and distributivity, respectively, thus indicating that any lattice that has the distributive property has associated with it a degree of inclusion that follows a sum and product rule (Knuth 2003, unpublished work). In addition, the cross-ratio in projective geometry has been shown to have the same form as the odds-ratio in Bayesian inference (Rodrìguez 1991), which is now believed to derive from the fact that the projective lattice also exhibits associativity (Knuth 2002). Fry has also demonstrated that this methodology is applicable to the area of control in cybernetic systems (Fry 2002). Second, degrees of inclusion do not need to be represented by real numbers. Complex numbers and quaternions also conform to the consistency requirements (Youssef 1994; Youssef 2001, unpublished work), as do the more general Clifford algebras (Rodrìguez 1991), which are multivectors in the geometric algebra (Hestenes & Sobczyk 1984) described in an earlier Millennium Issue (Lasenby *et al.* 2000). Furthermore, Caticha (1998) has derived the calculus of wavefunction amplitudes and the Schrodinger equation entirely by constructing a poset of experimental *setups* and using Cox's arguments of consistency with degrees of inclusion represented with complex numbers. This work leads to a very satisfying description of quantum mechanics in terms of measurements while explaining how it looks like probability theory—yet is not. We expect that the generalizations of lattice theory described here will not only identify unrecognized relationships among disparate fields, but also allow new measures to be developed and understood at a very fundamental level.

8. Automating Both Inference and Inquiry

Automation of both inference and inquiry will allow machines to learn from data, as well as ask relevant questions to obtain new data that could aid their inferences. This promises to automate the scientific method within a framework defined by a

set of possible experiments (questions that can be asked of a system) and a set of hypothesized theoretical models. Imagine a robot that has drilled through several kilometers of Europa's icy crust to emerge into an immense ocean far from any possible intervention by its human creators on Earth. Having been programmed to resolve the issue 'Is there life in Europa's ocean?', the machine, taking into account all of the relevant data on Europa that its creators have gathered, calculates the most relevant experimental question to ask given what it knows. Each experimental apparatus it carries also comes with its own energy cost, which the machine may also take into account. What is learned in its earlier experiments will help it decide which successive experiments to perform to resolve the scientific issue.

While the creation of independently behaving, learning machines will undoubtedly find great use in areas of science where humans are unable to perform experiments, they will most likely pervade our lives in ways we have not yet imagined—most probably with the equal potential of being annoying as being helpful. While the methodology necessary to construct such thinking machines is becoming clear, we have yet to find a way to automatically generate new hypotheses for the machines to entertain. Such flashes of inspiration serving to change the way we perceive the world are not obviously related to any logical procedure, and often occur through generalizations and analogies, which at this point seem to fall outside the domain of the methodology discussed in this paper.

References

Bayes, T. 1763 An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, **53**, 370–418.

Bernoulli, J. 1713 *Ars conjectandi*. Basel:Thurnisiorum.

Boole, G. 1854 *An investigation of the laws of thought*. London:Macmillan.

Bretthorst, G. L. 1988 *Bayesian spectrum analysis and parameter estimation*. Springer Lecture Notes in Statistics, no. 48.

Burg, J. P. 1967 Maximum entropy spectral analysis. Paper presented at 37th meeting of the Society of Exploration Geophysicists, Oklahoma City. Reprinted 1978 in *Modern Spectrum Analysis* (ed. D. Childers). New York:IEEE Press.

Caticha, A. 1998 Consistency, amplitudes and probabilities in quantum theory. *Phys. Rev.* **A57**, 1572.

Cheeseman, P., Kanefsky B., Kraft R., Stutz J., Hanson R., 1996 Super-resolved surface reconstruction from multiple images. In *Maximum Entropy and Bayesian Methods*, (ed. G. R. Heidbreder). pp. 293–308, Dordrecht:Kluwer.

Cox, R. T. 1946 Probability, frequency, and reasonable expectation. *Am. J. Physics* **14**, 1–13.

Cox, R. T. 1961 *The algebra of probable inference*. Baltimore:Johns Hopkins Press.

Cox, R. T. 1979 Of inference and inquiry. In *The Maximum Entropy Formalism* (eds. R. D. Levine & M. Tribus). pp. 119–167, Cambridge:MIT Press.

Davey B. A. & Priestley H. A. 2002 *Introduction to lattices and order*. Cambridge:Cambridge Univ. Press.

Dose, V. 2002 Bayes in five days. Reprint no. 83, Centre for Interdisciplinary Plasma Science.

Fry R. L. 1999 Maximum entropy and Bayesian methods. Electronic course notes (525.475), Johns Hopkins University.

Fry R. L. 2002 The engineering of cybernetic systems. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Baltimore MD, USA, August 2001* (ed. R. L. Fry). New York:AIP, 497–528.

Hanson, K. M. 1993 Introduction to Bayesian image analysis. In *Medical Imaging: Image Processing*, (ed. M.H. Loew). Proc. SPIE 1898, 716–731.

Hardy M., 2002 Scaled Boolean algebras. *Adv. in Appl. Math.* **29**(2), 243–292.

Hestenes D. & Sobczyk G. 1984 *Clifford algebra to geometric calculus: a unified language for mathematics and physics*. Dordrecht:Reidel.

Jaynes, E. T. 1957 Information theory and statistical mechanics. *Phys. Rev.* **106**, 620.

Jaynes, E. T. 1968 Prior Probabilities. *IEEE Trans. on Systems Science and Cybernetics* **SSC-4**, 227.

Jaynes, E. T. 1979 Where do we stand on maximum entropy? In *The Maximum Entropy Formalism* (eds. R. D. Levine & M. Tribus), pp. 15–118, Cambridge:MIT Press.

Jaynes, E. T. 2003 *Probability theory: the logic of science*. Cambridge:Cambridge Univ. Press.

Jeffreys, H. 1939 *The theory of probability*, Oxford:Clarendon Press.

Knuth K. H. & Hajian A. R. 2002 Hierarchies of models: toward understanding planetary nebulae. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Moscow ID, USA, August 2002* (ed. C. Williams). New York:AIP, in press.

Knuth K. H. 2002 What is a question? In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Moscow ID, USA, August 2002* (ed. C. Williams). New York:AIP, in press.

Laplace P. S. 1812 Théorie analytique des probabilitiés. Paris:Courcier Imprimeur.

Lasenby J., Lasenby A. N., Doran C. J. L. 2000 A unified mathematical language for physics and engineering in the 21st century. *Phil. Trans. R. Soc. Lond. A* **358**, 21–39.

Loredo T. 1990 From Laplace to supernova SN1987A: Bayesian inference in astrophysics. In *Maximum Entropy and Bayesian Methods*, (ed. P. Fougere), Dordrecht:Kluwer, 81–142.

Rodríguez C. C. 1991 From Euclid to entropy. In *Maximum Entropy and Bayesian Methods, Laramie Wyoming 1990*, (eds. W. T. Grandy & L.H. Schick). Dordrecht:Kluwer, 343–348.

Rodríguez C. C. 1998 Unreal probabilities: partial truth with Clifford numbers. In *Maximum Entropy and Bayesian Methods*, (eds. W. von der Linden, V. Dose, R. Fischer, and R. Preuss). Dordrecht:Kluwer, 247–270.

Shannon, C. E. 1948 A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 & 623–656.

Sivia D. S. 1996 *Data analysis: a Bayesian tutorial*, Oxford:Oxford Science Pub.

Skilling, J. 1998 Probabilistic data analysis: an introductory guide. *J. Microscopy* **190**(1), 28–36.

Smith, C. R., and Erickson G. J. 1990 Probability theory and the associativity equation. In *Maximum Entropy and Bayesian Methods*, (ed. P. Fougere). Dordrecht:Kluwer, 17–30.

Tribus, M 1961 Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications, Princeton:Van Nostrand.

Tribus, M. & McIrvine E. C. 1971 Energy and information. *Scientific American*, September **225**(3), 179–184.

Youssef, S. 1994 Quantum mechanics as complex probability theory. *Mod. Phys. Lett.* **A9**, 2571–2586.