# The Dawning of the Age of Stochasticity

David Mumford

ABSTRACT. For over two millennia, Aristotle's logic has ruled over the thinking of western intellectuals. All precise theories, all scientific models, even models of the process of thinking itself, have in principle conformed to the straight- jacket of logic. But from its shady beginnings devising gambling strategies and counting corpses in medieval London, probability theory and statistical inference now emerge as better foundations for scientific models, especially those of the process of thinking and as essential ingredients of theoretical mathematics, even the foundations of mathematics itself. We propose that this sea change in our perspective will affect virtually all of mathematics in the next century.

## 1. Introduction

This paper is based on a lecture delivered at the conference "Mathematics towards the Third Millennium", held at the Accademia Nazionale dei Lincei, May 27-29, 1999[1]. I would like to congratulate the seven very enterprising and very energetic Professors from the University of Rome, Tor Vergata, all women, who conceived and orchestrated that meeting. I am especially impressed by their achievement in getting a dozen mathematicians to speak, not about the latest advances in their field but to address larger issues and talk about ideas as well as theorems. Their invitation tempted me to try to formulate more clearly some ideas that I've been trying to put together for the last ten years. I could not resist the great fun of formulating a long term view out of them which is, no doubt, simplistic and which certainly stretches beyond my area of expertise. To quantify the hubris of this talk, let me borrow Karen Uhlenbeck's statistic defined in her talk at this conference: I wish to make assertions which cover some 2400 years; take as a yardstick the length of my own research experience – about 40 years; thus the hubris quotient of this talk is 60!

[1] This paper is reproduced here with the permission of the Accademia. References below to "other talks" all refer to this conference.

This paper is a meant to be a polemic which argues for a very fundamental point: that stochastic models and statistical reasoning are more relevant i) to the world, ii) to science and many parts of mathematics and iii) particularly to understanding the computations in our own minds, than exact models and logical reasoning. My points will be laid out as follows: in §2, I will argue that all mathematics arises by abstracting some aspect of our experience and that, alongside the mathematics which arises from objects and their motions in the material world, formal logic arose, in the work of Aristotle, from observing thought itself. However, there can be other ways of abstracting the nature of our thinking process and one of these leads to probability and statistics. In §3, I will give a quick look at the 2400 years since Aristotle, noting some high points in the development of these two strands. Precise logic-based models and precise logic-based mathematics have held the high ground and deeply influenced our thinking. Stochastic theories emerged much more slowly and only in the last century have begun to show their real depth. In §4, I want to look at the standard reductionist approach to probability. The basic object of study in probability is the *random variable* and I will argue that it should be treated as a basic construct, like spaces, groups and functions, and it is artificial and unnatural to define it in terms of measure theory. In §5, we pursue this point further and, building on inspiring work of Jaynes and Freiling, propose that probabilities and random variables can be built into the foundations of mathematics, resulting in a more intuitive and powerful formalism. In §6, we look at the impact of stochastic models on mainstream mathematics, especially on the theory of ordinary and partial differential equations. We argue that stochastic differential equations are more fundamental and relevant to modeling the world than deterministic equations. Finally, in §7, we return to modeling thought and examine recent stochastic approaches to artificial intelligence, vision and speech. We ask: do these offer a better chance of success, e.g. at duplicating human abilities with a computer, than logic based approaches. I believe so, although this is not yet clear.

I also have to confess at the outset to the zeal of a convert, a born-again believer in stochastic methods. Last week, Dave Wright reminded me of the advice I had given a graduate student during my algebraic geometry days in the 70's: 'Good grief, don't waste your time studying statistics – it's all cookbook nonsense'. I take it back! I would like to warmly thank some of the many people who have helped me either through discussions of these ideas or with the details of this article, especially Shlomo Sternberg, Rohit Parikh, Persi Diaconis, Ulf Grenander, Stuart Geman, David Fowler, and Stephen Stigler.

## 2. The taxonomy of mathematics

I want to begin by setting probability and statistics in their places as a part of mathematics. First, I want to quote a definition of what is mathematics due to Davis and Hersh in their very penetrating book "The Experience of Mathematics" (Davis-Hersh, 1980, p.399): *'The study of mental objects with reproducible properties is called mathematics.'* I love this definition because it doesn't try to limit mathematics to what has been called mathematics in the past but really attempts to say why certain communications are classified as math, others as science, others as art, others as gossip. Thus reproducible properties of the physical world are science whereas reproducible mental objects are math. Art lives on the mental plane (the real painting is not the set of dry pigments on the canvas nor is a symphony the sequence of sound waves that convey it to our ear) but, as the post-modernists insist, is reinterpreted in new contexts by each appreciator. As for gossip, which includes the vast majority of our thoughts, its essence is its relation to a unique local part of time and space.

Expanding on the Davis and Hersh definition, one can ask what are the various primitive elements of human experience which lead to the diverse types of reproducible mental objects, which in turn embody the great divisions of mathematics? The classical subdivisions of mathematics are geometry, algebra, and analysis. Let's look at each of them and try to name the corresponding experiences and the resulting mental objects.

Geometry is the most obvious: an infant at the age of 3-6 months is working intensely at integrating the two senses of vision and touch with its own simple muscular movements, learning that moving its hand and arm appropriately leads to the sensation of gripping the rattle and the sight of its displacement. Put succinctly, let me say that the perception of space (through senses and muscular interaction) is the primitive element of our experience on which geometry is based. One of the simplest mental objects this leads to is 'the stretched string' as Davis and Hersh call it, the origin of ruler and compass constructions. The paradigmatic object of its formal study is a space $M$ made up of points with various sorts of structure.

Analysis, I would argue, is the outgrowth of the human experience of force and its children, acceleration and oscillation. An example is the falling of the apple onto Newton's head. This primitive experience gives rise to the paradigmatic mental object consisting of a function and its derivatives, originally functions describing some physical quantity evolving in time.

Algebra seems to stem from the grammar of actions, i.e., the fact that we carry out actions in specific orders, concatenating one after the other, and making various 'higher order' actions out of simpler more basic ones. The simplest example, the one first acquired by children, is counting itself, which may be part of the grammar of dexterous manipulations if piling pebbles in heaps is used or part of the grammar of language when words are used. The paradigmatic mental object here is a set of things with a law of composition.

Enough for the 'classical' divisions of mathematics. I believe there is a fourth branch of human experience which creates reproducible mental objects, hence creates math: our experience of thought itself through our conscious observation of our mind at work. Instead of observing the world and finding there the germs of geometry and analysis, or observing our actions and finding algebra, we observe our mind at work. In the hands of Aristotle, this lead to the creation of formal logic in which propositions are the basic mental objects. Logic was the reproducible formalization constructed to model the raw stream of thoughts passing through our consciousness.

But is this right? The alternate view for which I will argue is that thought is the weighing of relative likelihoods of possible events and the act of sampling from the 'posterior', the probability distribution on unknown events, given the sum total of our knowledge of past events and the present context. If this is so, then the paradigmatic mental object is not a proposition, standing in all its eternal glory with its truth value emblazoned on its chest, but the random variable $x$, its value subject to probabilities but still not fixed. We will focus on random variables in §4. The simplest example where human thinking is clearly of this kind may well be the case where the probabilities can be made explicit: gambling. Here we are quite conscious that we are weighing likelihoods (and even calculating them if we are mathematically inclined). If we accept this, the division of mathematics corresponding to this realm of experience is not logic but probability and statistics.

## 3. A brief history of logic vs. statistics

It is entertaining to make a timeline and trace some of the high points in the evolution of these two conflicting views of the nature of thought. Starting in the high period of ancient Athens, here are some quotes from Plato, put into the mouth of Socrates:

> *If Theodorus, or any other geometer, were prepared to rely on plausibility when he was doing geometry, he'd be worth absolutely nothing.* (The dialog with Theaetetus, 162e, c. 360 B.C.)

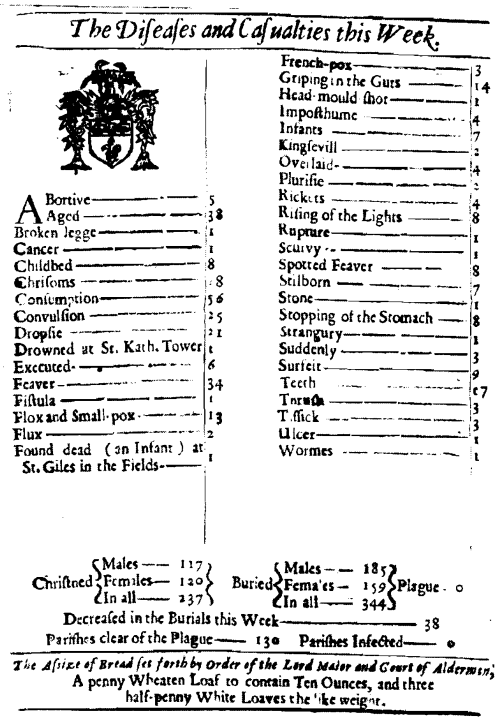
In the Republic VII, 529c, Plato goes a bit far even for the tastes of the purest contemporary mathematicians by arguing that astronomers are better off *not looking at the stars* (?!):

> *The sparks that paint the sky, since they are decorations on a visible surface, we must regard, to be sure, as the fairest and most exact of material things; but we must recognize that they fall far short of the truth .... both in relation to one another and as vehicles of the things they carry and contain. These can be apprehended only by reason and thought, but not by sight. .... It is by means of problems, then, as in the study of geometry, that we will pursue astronomy too, and we will let be the things in the heavens, if we are to have a part in the true science of astronomy.*

In the same vein, it is interesting that some of the worst mistakes made by Aristotle arose because, although he wrote extensively about biology, he never consulted practising physicians such as Hippocrates and his school for real data about the human body. Thus he believed that the heart, not the brain, was the seat of thought, something readily disproven by observing the effects of trauma to the brain (see the excellent article by Charles Gross (1995)).

Skipping ahead to the Renaissance, Cardano (1500-1571) is a remarkable figure. On the one hand, because of his book *Ars Magna*, 1545, he is often called the inventor of $i$. He appears to be a superb practitioner of the formalism of algebra, following the consequences of its logical rules a bit further than those before him. But he was also an addicted gambler and wrote the first analysis of the laws of chance in *Liber de Ludo Aleae*, which, however, he was ashamed to publish! It did not appear until 1663, about the time Jacob Bernoulli began to work. In the 17th century, we find Newton and Leibniz squarely in the logic camp, Newton believing that Euclidean geometry was the only reliable language for trustworthy proofs and Leibniz foreshadowing modern AI in his PhD thesis *De Arte Combinatoria*. In the stat camp, we have true empiricists beginning to gather and analyze statistics. Graunt assembled his mortality tables in London (see figure 1 from the year 1665) and Jacob Bernoulli proved the law of large numbers, justifying the use of empirical estimates.

The Reverend Thomas Bayes lived in the 18th century (1701 or 1702-1761). He argued for the introduction of *a priori* (or 'prior') probabilities, probabilities that one assigns to unknown events based on experience of related but not identical events or just expressing a neutral agnostic view. These probabilities should then be modified by new observations, leading to better



The Diſeaſes and Caſualties this Week.

| | | | |
|---|---|---|---|
| ABortive | 5 | French-pox | 3 |
| Aged | 38 | Griping in the Guts | 14 |
| Broken legge | 1 | Head-mould ſhot | 1 |
| Cancer | 1 | Impoſthume | 4 |
| Childbed | 8 | Infants | 7 |
| Chriſoms | 8 | Kingſevill | 2 |
| Conſumption | 56 | Overlaid | 4 |
| Convulſion | 25 | Pluriſie | 2 |
| Dropſie | 21 | Rickets | 4 |
| Drowned at St. Kath. Tower | 1 | Riſing of the Lights | 8 |
| Executed | 6 | Rupture | 1 |
| Feaver | 34 | Scurvy | 1 |
| Fiſtula | 1 | Spotted Feaver | 8 |
| Flox and Small-pox | 13 | Stilborn | 7 |
| Flux | 2 | Stone | 1 |
| Found dead (an Infant) at St. Giles in the Fields | 1 | Stopping of the Stomach | 8 |
| | | Strangury | 1 |
| | | Suddenly | 3 |
| | | Surfeit | 9 |
| | | Teeth | 67 |
| | | Thruſh | 3 |
| | | Tiſſick | 3 |
| | | Ulcer | 1 |
| | | Wormes | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Chriſtned | Males | 117 | Buried | Males | 185 | Plague 0 |
| | Females | 120 | | Females | 159 | |
| | In all | 237 | | In all | 344 | |

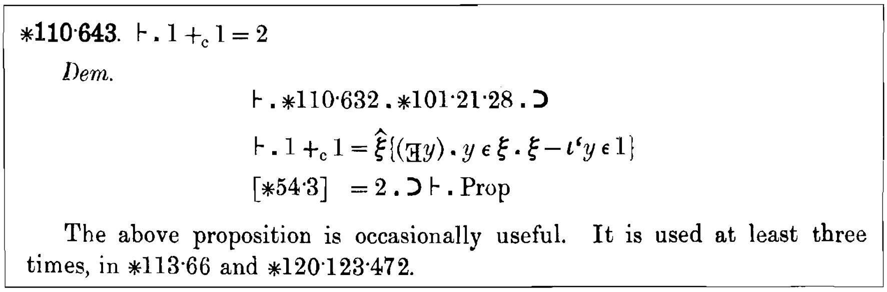
Decreaſed in the Burials this Week — 38

Pariſhes clear of the Plague — 130 Pariſhes Infected — 0

*The Aſsize of Bread ſet forth by Order of the Lord Maior and Court of Aldermen;*
A penny Wheaten Loaf to contain Ten Ounces, and three
half-penny White Loaves the like weight.

FIGURE 1. Graunt was one of the first people to realize the usefulness of empirical data: here is a week in the life and death of medieval London (photograph courtesy of Stephen Stigler).

and better *a posteriori* probabilities as data is accumulated. To demonstrate the central importance of Bayes's work, let me describe the lead article in the Business Section of the L.A. Times of 10/28/96. It featured a picture of Bayes with the headline "The future of software may lie in the obscure theories of an 18th century cleric named Thomas Bayes". The article went on to say, "Asked recently when computers would finally begin to understand human speech, Gates began discussing the critical role of 'Bayesian systems'. ... Is Gates onto something? is this alien-sounding technology Microsoft's new secret weapon?" In speech recognition, the prior probabilities may be generic models of human speech and the posterior probabilities the much more accurate model of one person's speech after training. Although the Times labelled them 'obscure theories', a growing school of researchers today (myself among them) believes Bayesian statistics is the key to the effective use of statistical inference in complex situations.



**$*110{\cdot}643$.** $\vdash . 1 +_{c} 1 = 2$

*Dem.*

$$\vdash . *110{\cdot}632 . *101{\cdot}21{\cdot}28 . \supset$$
$$\vdash . 1 +_{c} 1 = \hat{\xi}\{(\exists y) . y \in \xi . \xi - \iota\text{‘}y \in 1\}$$
$$[*54{\cdot}3] \quad = 2 . \supset \vdash . \text{Prop}$$

The above proposition is occasionally useful. It is used at least three times, in $*113{\cdot}66$ and $*120{\cdot}123{\cdot}472$.

FIGURE 2. A crowning achievement in the reductionist approach to the foundations of mathematics. The above theorem occurs some thousand odd pages into the monumental work *Principia Mathematica* of Russell and Whitehead, building purely on logic and set theory. Reproduced with permission of Cambridge University Press.

Gauss is interesting because of his immense abilities both in pure logical deduction and in applied statistics. Indeed, he invented the method of least squares to deal with redundant but inaccurate data, leading to the rediscovery of Ceres, and proved the central limit theorem which justified the method. Perhaps his most famous hypothesis testing experiment was to test the euclidean nature of our 3-dimensional world. He did this by measuring the three angles in the triangle formed by the 3 peaks of Brocken, Hohehagen and Inselsberg: it came out 14.85 arc-seconds higher than $\pi$, but within experimental error of $\pi$. The logic camp flourished in the rest of the 19th century, with Dedekind's cuts to arithmetize the real numbers, Boole's logic, Frege's formalization of predicate calculus and Cantor's formalization of set theory. It is not uninformative to reproduce here a high point of this school: Russell and Whitehead's demonstration that 1+1=2 (this is Theorem 110.643 of Principia Mathematica). See figure 2 and note their comment on the result in the next paragraph! But the gathering of empirical statistics also flourished in the 19th century, notably in the hands of Francis Galton, who liked to measure so much about people that he is not now considered very 'politically correct'[2].

Moving to our century, I think the most significant trend has been the development of more complex and truly interesting probability models with much deeper applications to the sciences. Thus Galton was pretty much limited to fitting Gaussian distributions to scalar or low-dimensional data sets. A huge leap was made when Gibbs introduced very high-dimensional

[2] A personal note: my grandfather, Alfred A. Mumford, was a physician at Manchester Grammar School for many years and fascinated by the correlations he observed in the meticulous measurements and health records he made of the boys. (Mumford-Young 1923) is cited in the classical statistics textbook of Snedecor and Cochran.

probability models in physics, e.g. for gases, starting statistical mechanics. Keynes wrote both on the foundations of probability and of economics and sought to clarify what was the correct use of probabilistic reasoning in the real world. Wiener applied stochastic methods to signal prediction and control theory. Shannon applied stochastic methods to data compression and identified the key role played by the entropy of a probability distribution. Grenander applied stochastic methods first to algebraic structures and later to the patterns they create in the world, especially in vision. All these together have given us powerful tools and inspiring examples of applied stochastic methods.

While all these really exciting uses were being made of statistics, the majority of statisticians themselves, led by Sir R.A. Fisher, were tying their hands behind their backs, insisting that statistics couldn't be used in any but totally reproducible situations and then only using the empirical data. This is the so-called 'frequentist' school which fought with the Bayesian school which believed that priors could be used and the use of statistical inference greatly extended. This approach denies that statistical inference can have anything to do with real thought because real-life situations are always buried in contextual variables and cannot be repeated. Fortunately, the Bayesian school did not totally die, being continued by DeFinetti, E.T. Jaynes, and others. I will describe some of Jaynes's ideas below. The new applications of Bayesian statistics to vision, speech, expert systems and neural nets have now started an explosive growth in these ideas.

## 4. What is a 'random variable'?

This is actually a quote from David Kazhdan: when he transplanted Gel'fand's seminar to Harvard, he called it the 'Basic Notions Seminar' and asked everyone to describe a notion they knew best which everyone should learn. He gave Persi Diaconis the topic which is the title of this section. I like his idea: a random variable is not such an easy thing to describe. It is the core concept in probability and statistics and, as such, appears in many guises. Let's make a list:

- There are empirical random variables. These arise, for example, by taking a sample of people and tabulating their heights and weights; taking a random image and measuring the intensity of its pixels; taking a sample of stocks and tabulating their prices; throwing a dart at a dart board and measuring where it lands.
- There are elementary random variables. For example, a random sample from a finite set with the uniform distribution; a random normally distributed real number; a random sample from Brownian motion.

- There are truly complex random variables. One example would be the solution of a stochastic PDE with a white noise driving term. Another would be a random manifold created by some construction using elementary random elements of some kind. Gromov described some of these in his lecture.
- A doctor's diagnosis can be viewed as a random sample from his posterior probability distribution on the state of your body, given the combination of a) his personal experience, b) his knowledge from books, papers and other doctors, c) your case history and d) your test results. See the very influential article (Lauritzen-Spiegelhalter 1988).
- A novel can be viewed as a random sample from the author's posterior probability distribution on stories, conditioned on all the things the author has observed or learned about the nature of the real world. This will be developed in the last section.
- It can be viewed as an undefined operation in the axiomatization of mathematics: see the next section.
- Perhaps an observation in quantum mechanics is a 'non-commutative random variable', if we use the perspective A. Connes discussed in his talk?

When probability is built on top of measure theory, the usual formal definition of a random variable with values in a set $X$ is that it is a measurable function $x : \Omega \to X$ from a probability space $\Omega$ to $X$. The probability space itself, however, usually plays almost no role and $x$ acts as though it is a floating member of the set $X$ (like a generic point in algebraic geometry). Thus, i) for empirical random variables, $\Omega$ is essentially unknowable; ii) for the elementary random variables, $\Omega = X$; iii) for the complex random variables, $\Omega$ is some big product of the probability spaces from which all the random elements in the construction have been drawn; iv) for the novelist or doctor, $\Omega$ is the full probability model that he/she has constructed of how the world works.

There are two approaches to developing the basic theory of probability. One is to use wherever possible the reduction to measure theory, eliminating the probabilistic language. Then $\Omega$ is dropped and $X$ is endowed with the measure $p(x)$ or $p(x)dx$ given by the direct image under the map $x$ of the probability measure on $\Omega$. The other is to put the concept of 'random variable' on center stage and work with manipulations of random variables wherever possible. Here is one example contrasting these two styles.

Consider the concept of 'infinite divisibility' (ID) of a real-valued random variable $x$. One can be classical and denote the probability density of $x$ by $p(x)$. Then $x$ is ID if, for every $n$, there is a probability density $q_n(x)$ such that $p = q_n * \cdots * q_n$ ($n$ factors $q_n$). Alternately, one can say that, for every $n$, $x \sim y_1 + \cdots + y_n$ where $y_i$ are independent identically distributed random variables (and $\sim$ means having the same law).

This is little more than a simple change of notation but consider what happens when you state the Levy-Khintchine theorem in the two corresponding ways. The first way of stating this theorem says that $x$ is ID if and only if the Fourier transform $\hat{p}(\xi)$ of $p(x)$ can be written:

$$\hat{p}(\xi) = e^{ia\xi - b\xi^2 - c\int\left(e^{i\xi y} - 1 - \frac{i\xi y}{1+y^2}\right)d\mu(y)}.$$

The second way writes the same condition directly in terms of the random variable $x$ as follows:

$$x \sim a + bx_{\text{normal}} + c\sum(x_i - \text{convergence factor } c_i).$$

where $x_{\text{normal}}$ is a standard normal variable and $\{x_i\}$ are a Poisson process from a density $\nu$. Now these look quite different! For my part, I find the second way of stating the Levy-Khintchine theorem infinitely clearer: making the random variables explicit tells you the real stochastic meaning of the result.

## 5. Putting random variables into the foundations

The reductionist approach defines random variables in terms of measures, which are defined in terms of the theory of the reals, which are defined in terms of set theory, which is defined on top of predicate calculus. I'd like to propose instead that it should be possible to put random variables into the very foundations of both logic and mathematics and arrive at a more complete and more transparent formulation of the stochastic point of view. I do not have a complete formulation of this, but a sketch which draws on two sources I find very provocative. The first is the development by E.T. Jaynes of the foundations of Bayesian probability and statistics (Jaynes 1996-2000); the second is a beautiful stochastic argument due to Christopher Freiling to *disprove the continuum hypothesis* (Freiling 1986).

First Jaynes: as we have seen, the probability space $\Omega$ needed for the random variables in applications like medical diagnosis is impossible to pin down precisely. Too many fragments of experience may guide the physician and we can never make his/her probability table explicit. This problem was at the root of the frequentist's complaint about Bayesian methods. Jaynes has, I believe, the most convincing answer. His theory starts with the assumption that agents like us assign to various events $A$ plausibilities which lie in some unknown linearly ordered set, call it $\mathcal{P}\ell$. In fact, we assign plausibilities not only to events by themselves, but also to conditional events – if $B$ is known to happen, then what is the plausibility of $A$ as well being true? Denote this plausibility by

$$p(A|B) \in \mathcal{P}\ell.$$

Jaynes's result is that with a few reasonable axioms, one can deduce that there is an order isomorphism $\mathcal{P}\ell \cong [0,1]$ under which $p$ becomes a probability distribution on the algebra of $A$'s (in particular, $p(A|B) = p(A \wedge B)/p(B)$). We may summarize this result as saying that probabilities are the *normative* theory of plausibility, i.e., if we enforce natural rules of internal consistency on any home-spun idea of plausibility, we end up with a true probability model. For details, see his fascinating book (Jaynes 1996-2000, chapters 1,2) which apparently is going to finally appear posthumously.

This leads to the following proposal for a stochastic predicate calculus. It should have the syntax of standard predicate calculus except that we have two kinds of variables in it: the ordinary predicates and constants and quantifiable free variables $x$ but also a set of random constants $\underline{x}$. In addition, it comes with a truth value function $p$ mapping all formulas $F$ without free variables to real numbers between 0 and 1. If the formula $F$ has only ordinary variables in it, then $p(F) \in \{0,1\}$. Formal semantics for this theory would make the random constants functions on probability spaces so that a formula would define a subset of the product of these spaces, hence have a probability.

Stochastic formal number theory would be expressive enough to add an axiom of continuity for $p$:

$$p\left(\exists n \underline{F}(n)\right) = \text{l.u.b.}_m\, p\left((\exists n \leq m)\underline{F}(n)\right).$$

We also want axioms giving us the basic elementary random variables. Thus if $\mathcal{N}$ is the predicate defining natural numbers, Bernoulli random variables are given by the meta-axioms:

$$(\forall a \in \mathcal{P}\ell)(\exists \underline{x}_a) \ni \mathcal{N}(\underline{x}_a) \wedge [p(\underline{x}_a = 0) = 1 - a] \wedge [p(\underline{x}_a = 1) = a].$$

In fact, one wants countable families of independent Bernoulli variables. In the same vein, the basic axiom of stochastic analysis should be the existence of the continuum defined by i) a predicate $\mathcal{C}$, ii) a linear ordering $<(c_1, c_2)$ of numbers $c_1, c_2$ for which $\mathcal{C}(c)$ is true, iii) a random $\underline{x}_0$ satisfying $p(\mathcal{C}(\underline{x}_0)) = 1$, and iv) finally an axiom:

$$(\forall a \in \mathcal{P}\ell, a \neq 0,1)(\exists ! c) \ni \mathcal{C}(c) \wedge [p(<(\underline{x}_0, c)) = a]$$

In english, what we have in mind is that we can order the continuum in such a way that its one-sided intervals give all possible probabilities between 0 to 1 exactly once, i.e., loosely speaking, a continuum is exactly a thing you can throw darts at. The dart game (formally, Lebesgue measure on $(0,1)$) is given by the basic random variable $\underline{x}_0$, which connects syntactic real number variables to semantic plausibility value variables through the above axiom. This embeds measure theory into the very foundations of the theory.

This leads us to the stunning result of Christopher Freiling (1986): using the idea of throwing darts, we can disprove the continuum hypothesis. Why his theorem is not universally known and considered on a par with the results of Gödel and Cohen, I do not know. Here is the argument in classical language (see figure 3). Two dart players independently throw darts at a dartboard. If the continuum hypothesis is true, the points $P$ on the surface of a dartboard can be well-ordered so that for every $P$, the set of $Q$ such that $Q < P$, call it $S_Q$, is countable. Let players 1 and 2 hit the dart board at points $P_1$ and $P_2$. Either $P_1 < P_2$ or $P_2 < P_1$. Assume the first holds. Then $P_1$ belongs to a countable subset $S_{P_2}$ of the points on the dartboard. As the two throws were independent, we may treat throw 2 as taking place first, then throw 1. After throw 2, this countable set $S_{P_2}$ has been fixed. But every countable set is measurable and has measure 0. Thus the probability of throw 1 landing on $S_{P_2}$ is 0. The same argument shows that the probability of $P_2$ landing on $S_{P_1}$ is 0. Thus almost surely neither happened and this contradicts the assumption that the dartboard is the first uncountable cardinal!

So what is 'wrong' with this? We have treated random variables, throws of the dart, as real things! If we try to rewrite this argument in classical measure theory, we will need to assume that the graph of the well-ordering is measurable, which, of course, should not be done. So do we throw out the proof? Freiling used the argument to motivate a new axiom of set theory which disproves the continuum hypothesis. I believe we should go much further: his 'proof' shows that if we make random variables one of the basic elements of mathematics, it follows that the C.H. is false and we will get rid of one of the meaningless conundrums of set theory. The continuum hypothesis is surely similar to the scholastic issue of how many angels can stand on the head of a pin: an issue which disappears if you change your point of view.

This calls for the most difficult part of this proposed reformulation of the foundations: we need to decide how to define stochastic set theory. Clearly we must drop either the axiom of choice or the power set axiom. But the existence of random objects is a sort of axiom of random choice and my belief at this point is that it is better to drop the power set axiom. What mathematics really needs, for each set $X$, is not the huge set $2^X$ but the set of sequences $X^{\mathbb{N}}$ in $X$. Moreover, since $p(\underline{x}_0 \in A) \in \mathcal{P}\ell$ must be defined for every subset $A$ of $\mathcal{C}$, it is necessary that every definable subset of $\mathcal{C}$ is measurable. This is not my area but it seems to me that the results of (Shelah and Woodin, 1990) make this not obviously inconsistent or unworkable! It would be exciting to pursue this approach.

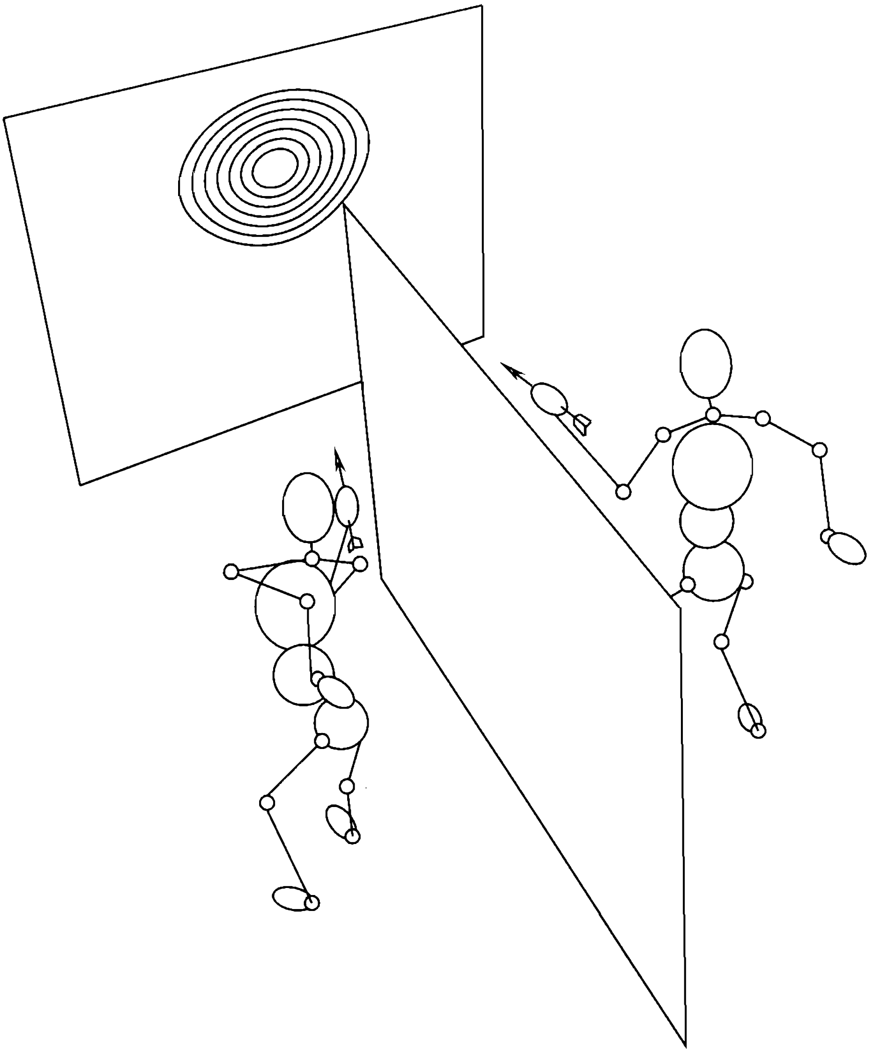


FIGURE 3. Two dart players face off. This led C. Freiling to his argument against the continuum hypothesis (drawing by I. Trotts).

## 6. Stochastic methods have invaded classical mathematics

It may be useful to look at the degree to which many areas of classical mathematics have been transformed and deepened by the use of stochastic methods.

Graph theory is a clear example from the area of combinatorics. The invasion started with Erdös's introduction of random graphs with a fixed number of vertices and edges, which led to the beautiful discovery of the phase transition phenomenon: that the random graph becomes connected almost surely within a very narrow band of edge numbers. An interesting example is use of stochastic methods to construct graphs with given degree and *girth* (the minimum size of a cycle) which are roughly of the minimum possible order. This approach is now called the 'probabilistic method', described, e.g., in (Spencer, 1994). In another direction, there is the elegant theory of random branching processes, which developed from the question of how likely it is that a given line of descent of the nobility would become extinct.

Perhaps the most convincing case for the importance of stochastic methods is in the theory of ODE's and PDE's. Differential equations were developed to model nature with the full understanding that every specific equation was a partial representation of reality that modeled some effects but not others. The original case was, of course, the 2-body problem and Newton's laws of motion. This predicted wonderfully planetary motion and, with perturbations, models the full set of planets for moderate periods of time (e.g., maybe $10^8$ years). But going out further (maybe to $10^9$ years), the unmodeled effects begin to add up and the approximation is not useful. So where does this leave the mathematical study of the 3-body problem? It makes the classical deterministic analysis of the 3-body gravitational equations about as relevant to the world as the continuum hypothesis! A major step in making the equation more relevant is to add a small stochastic term. Even if the size of the stochastic term goes to 0, its asymptotic effects need not. It seems fair to say that *all* differential equations are better models of the world when a stochastic term is added and that their classical analysis is useful only if it is stable in an appropriate sense to such perturbations.

What is more important to the mathematician is that the nature of the analysis of a differential equation shifts when they are considered stochastically. For classical differential equations with well-behaved solutions, it generally makes little difference whether we add a stochastic term or not: an attractive fixed point remains an attractive fixed point (though it gets a bit 'blurred' – the solution will jiggle around the fixed point a bit). But when the equation leads to some sort of 'chaotic' or turbulent behavior, we get a very different and hopefully much more satisfactory picture of the equation through its stochastic analysis. Instead of focussing on describing the pathologies of the strange attractors to which the classical solution tends asymptotically, the center of attention is now the existence of an invariant probability measure in which almost all solutions spend their whole lives. This idea originated in statistical mechanics, in the study of Brownian motion and the Ising model. Unfortunately, many of the 'results' in these theories are either heuristically justified by physicist-style reasoning or are still in the stage of dreams (as discussed in the talk of Talagrand). What we hope will happen, and has been proven at least in some cases, is that almost all random orbits have similar structure which can be described in great detail and which give real insight into the differential equation.

A startlingly beautiful successful example is the analysis of the stochastic Burger's equation by Weinan E, Y. Sinai and others (E et al, 1997). Whereas the usual Burger's equation can develop a huge mess of shocks, in the periodic stochastic case, there turns out to be almost surely one and only one shock wave which persists for all time (past and future) and which absorbs all other shocks[3]. The grand challenge (as our funding agencies like to say) is to analyze the stochastic Navier-Stokes equation, possibly leading to an understanding of turbulence, as discussed in Fefferman's talk.

Mathematical physics has lept ahead of pure mathematics in the use of stochastic methods: a central element in string theory is the introduction of random Riemann surfaces via a probability measure on the moduli space and Hawking has considered random topologies on space-time.

## 7. Thinking as Bayesian inference

I want to conclude with some description of the area that I know best: the modeling of thought as a computational process. I want to begin by contrasting the idea of reasoning with logic and reasoning with likelihoods with two examples. The example of the use of logic is an amusing syllogism taken from the Boston Driver's Handbook. The reader may entertain himself/herself by checking the logic!

**Premises:**
a) Tolstoi was a genius,
b) Tolstoi can only be truly appreciated by geniuses,
c) No genius is without some eccentricity,
d) Tolstoi sang the blues,
e) Every eccentric blues singer is appreciated by some half-wit,
f) Eccentrics think they own the road.

$\Longrightarrow$

**Consequence:** There is always some half-wit who thinks he owns the road.

[3] I am sorry that this group is not into computer simulations, so I cannot show you here an impressive simulation.

Although absurd, I think this syllogism points out well the fact that precise reasoning is seldom appropriate in real life – generalizations usually apply only in various contexts with various degrees of plausibility and stringing many of them together is bound to create nonsense. Thus (d) for instance might be considered an acceptable metaphor and (f) is acceptable common usage meant to apply only to a certain class of eccentrics (possibly disjoint from the eccentrics in (c)). But it makes no sense to reason with them logically. Contrast the above with Judea Pearl's example of common-sense reasoning from his book (Pearl 1988):

a) Watson phones Holmes in his office and states the burglar alarm in Holmes's house is going off. Holmes prepares to rush home.
b) Holmes recalls Watson is known to be a practical joker hence doubts his statement.
⇒
c) Holmes phones Mrs. Gibbon, another neighbor. She is tipsy and rants about crime, making Holmes think she has heard the alarm.
d) Holmes remembers the alarm manual said it might have been triggered by an earthquake.
e) Holmes realizes that if there had been an earthquake, it ought to be mentioned on the radio.
⇒
f) Holmes turns on his radio to check.

In Pearl's analysis, Holmes's mental processes are modeled by a 'Bayesian net', shown in figure 4. Such a net is a directed graph whose nodes represent events which may or may not be true. The edges represent causation and have conditional probabilities attached to them. This set of conditional probabilities is called the 'prior', the probabilistic information that Holmes brings to the table before his phone rings. In the figure there are 6 vertices representing the 2 known events – the testimony of Watson and Gibbons – and 4 events whose occurence Holmes is weighing. At each stage of his thinking, Holmes has some evidence – vertices whose truth value is known – and has his priors and needs to compute the 'posterior', the updated probabilities of all the events, given the evidence. Note how his reasoning goes up and down in this graph, seeking to fix better probabilities to the unknown events by, e.g., phoning Mrs. Gibbons and by turning on his radio. See Pearl's book, p.42-52, for details on this example and the algorithm for 'belief updating'.

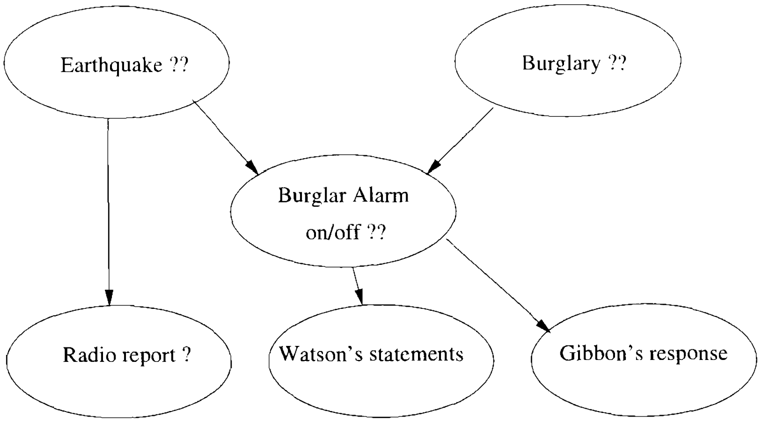


FIGURE 4. The Bayesian belief net for Pearl's anecdote about Holmes's burglar alarm.

One of the central problems in understanding thinking has been to formalize inductive learning. Although logic offers a system for formalizing deduction, induction has been much harder to understand from a logical perspective. I believe by far the most convincing formal definition of induction is the one discovered by Leslie Valiant and now known as the 'PAC' (or 'probably approximately correct') model. I want to give this definition in full because it also illustrates how probabilistic methods extend naturally to learning as well as inference.

Here is the setup: $\Omega$ is a set of possibilities and $\mathcal{C}$ is a class of predicates $P : \Omega \to \{0,1\}$. One of the $P \in \mathcal{C}$ is true and the problem is to estimate the true $P$ on the basis of examples $(y, P(y))$. The class $\mathcal{C}$ is said to be PAC-learnable if there is an algorithm which computes a guess $\hat{P}_\mathcal{D} : \Omega \to \{0,1\}$ from $n$ examples $\mathcal{D} = (y_1, \cdots, y_n; P(y_1), \cdots, P(y_n))$ and which satisfies the following:

$$\forall \epsilon_1, \epsilon_2 \, \exists n \, \forall \text{ prob. distr. } \pi \text{ on } \Omega$$

$$y_1, \cdots, y_n \in \Omega, \text{ iid wrt } \pi \Rightarrow \text{Prob}_\mathcal{D}\Big(\text{Prob}_y(\hat{P}_\mathcal{D}(y) \neq P(y)) < \epsilon_2\Big) > 1-\epsilon_1.$$

Note what this means: you have no idea which examples are common and which are rare in real life, but what you must rely on is that your learning examples are drawn from the same distribution as your test examples. Then there is small probability $\epsilon_1$ of being given really misleading examples; but, if you are given typical examples, then you only make $\epsilon_2$ errors after seeing enough examples. I find this very convincing as the 'right' way to formalize inductive learning.

Returning to thinking as a whole, which includes learning models, storing models, and reasoning from models, let's consider the hypothesis that thinking is accomplished by constructing probability models and using Bayesian inference. I believe there are three major obstacles that have to be overcome to make it plausible that this can work in real situations and not just in toy examples like that of Pearl. The first is that in the real world, there are millions of random variables to be considered and full probability tables for all possible values of these variables are much much too huge to be stored. We need some restricted class of probability models which seem expressive enough to model reality but which are succinct enough to be storeable. Secondly, we have to show inference at reasonable speed is feasible with such models. Thirdly, we have to show that the parameters in these classes of probability distributions are PAC-learnable. This is a tall order but major work has been done and some very interesting progress seems to be being made.

Gibbs made the first major step to creating huge but workable probability models. His idea is to consider models such that the logarithm of the probability is the sum of terms each involving only a small number of random variables at a time:

$$Prob(\{x_k\}) = \frac{1}{Z} e^{-\sum_C E_C(x_C)}$$

where $Z$ is a constant, $x_C = \{x_k \mid k \in C\}$ and the sets $C$ are supposed to be 'small' sets of the variables. Such 'Gibbsian' models have been extremely widely used in AI, vision, speech, and neural networks. In the continuous domain, such models may be viewed as natural generalizations of Gaussian models: Gaussian models are precisely those such that log-probability is the sum of terms involving only two variables at a time and of the form $ax_i, bx_i^2$ or $cx_ix_j$. But general Gibbsian models may be highly non-Gaussian, nonparametric, and with mixed continuous and discrete variables.

Wavelet expansions of images of the real world are examples which lead directly to non-Gaussian Gibbs probability distributions. The key reason wavelet expansions are preferred to, e.g., Fourier expansions, for images is that the wavelet coefficients of natural images are sparse. This means that typically a relatively small number of the wavelet coefficients are large and most are near zero. More explicitly, they behave as though sampled from a non-Gaussian distribution like $p(I) = \frac{1}{Z} e^{-a \sum_\alpha \sqrt{|c_\alpha|}}$ where $c_\alpha$ are the wavelet coefficients of the image $I$.

Gibbsian models alone do not seem to be expressive enough for the full real world: it seems that the needed probability models must also incorporate 'dynamic links', further variables which bind or compose parts into wholes in a grammatical fashion. Some of these variables identify 'slot fillers', e.g.,

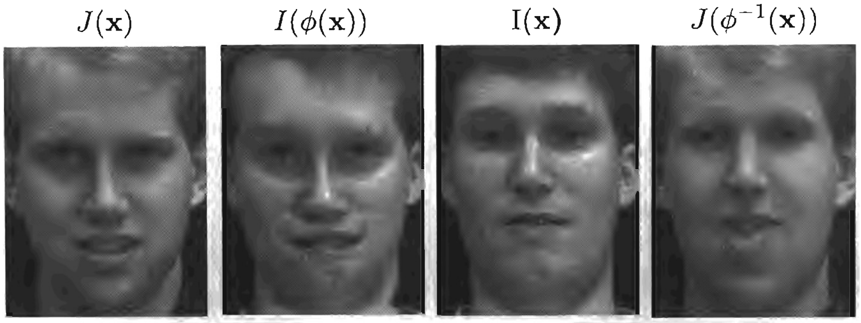


FIGURE 5. Example from the work P. Hallinan on aligning faces by diffeomorphisms. The two faces are given by images $I$ and $J$ and the warping is given by the map $\phi$. Reproduced with permission by AKPeters.

pointers to the word which is the subject of a sentence or the point on the retina is the nose of a face being perceived. Other links are needed to group related objects like things with common motion or the pixels imaging the same object in the left and right eyes. Developing probability models with such dynamic links is a major area of research today.

Face recognition is a simple example where dynamic link variables may be used. One can seek to identify faces by forming a universal 'template' face and warping this template onto all perceived faces by a suitable diffeomorphism, called the 'rubber mask technique' by (Widrow 1973). Differing illumination also causes large changes in the image of a face, so the random variables in this model are both the coordinates of the warping applied to reference points in the template and shading coefficients expressing how the face is illuminated. The log-probability is then a sum of terms expressing the goodness of fit of the warping of the observed image with a sum of templates representing faces under different lighting conditions (Hallinan et al. 1999). Some examples are shown in figure 5.

Is it practical to make inferences on the basis of these complex models? Very often, the inference one wants to make is to find the MAP estimate for the relevant unobserved random variables $x_S$, with the probability distribution conditioned on all observations $\widehat{x_T}$. Here MAP stands for 'Maximum A Posteriori' probability, the most probable set of values of these variables and we are seeking:

$$\text{argmax}_{x_S} p(x_S \mid \widehat{x_T}).$$

This is an optimization problem and there are three basic techniques for solving or approximately solving such problems: gradient descent, dynamic programming, and Monte Carlo Markov chains. Unfortunately, they all run into problems when the model gets complex: gradient descent gets lost in

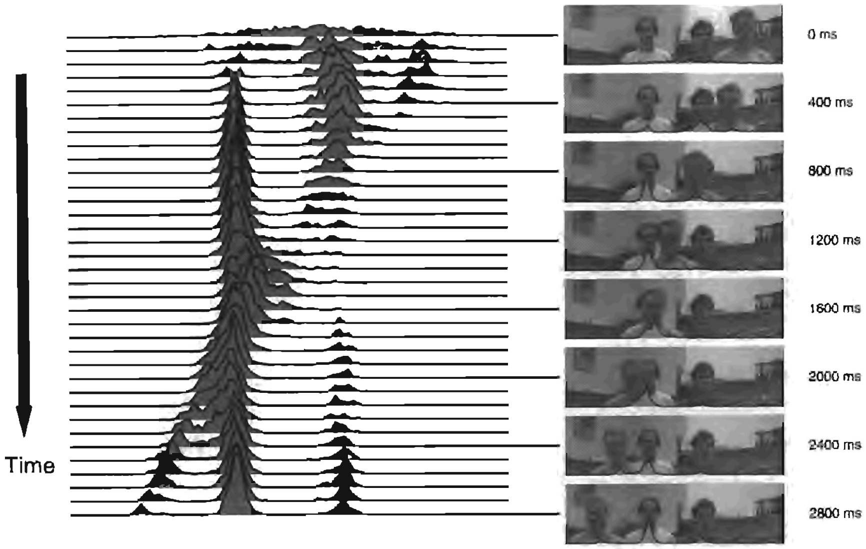


FIGURE 6. Example from the work of M. Isard and A. Blake on tracking moving faces in a cluttered environment using particle filtering. On the right are the images; on the left are smoothed multi-modal probability distributions estimating the conditional probability of a face at each location, given the present and past image sequence. Reproduced with permission of Kluwer Academic Publishers.

local optima; dynamic programming only works when there is a natural linear ordering of the variables, decoupling non-adjacent variables; and Monte Carlo Markov chains tend to be very slow. Nonetheless, these have been the workhorses in the field until recently. Speech recognition, for example, got where it is by total reliance on dynamic programming techniques and is weak where these methods fail.

A new idea to tame stochastic methods has recently been explored by several groups. This has been called 'particle filtering' and 'factored sampling' (Grenander et al., 1991), (Gordon et al., 1993), (Kanizawa et al., 1995) and (Blake-Isard, 1996), and is a Monte Carlo method which works by computing with a moderate sized sample $\{x_S^{(\alpha)}\}$ (perhaps 100 or 1000 $\alpha$'s) from the distribution, not just with one sample at a time as in Monte Carlo Markov chains. The point is to make a weak approximation:

$$p(\cdot \mid \widehat{x_T}) \sim \sum_{\alpha} w_\alpha \delta_{x_S^{(\alpha)}}$$

which is to say, for some class of nice random variables $f$ on our probability space:

$$\mathrm{Exp}(f \mid \widehat{x_T}) \approx \sum_{\alpha} w_\alpha f(x_S^{(\alpha)}).$$

The hope is that many multi-modal probability distributions can be approximated by weighted samples in this way, at least for the random variables of interest. More than that, one hopes that maintaining this sample will allow the robust merging of new data into a situation where a previously less likely option is changed into the most likely option. An example showing the successful tracking of multiple moving people, from the work of Blake and Isard, is shown in figure 6. Standard classical techniques, like the Kalman filter, based on Gaussian models, typically fail in cases like this.

This discussion has been aimed at giving a flavor of research in the application of stochastic methods to modeling intelligent behaviour. This is very much an on-going enterprise. All too often, various schools studying the problem of modeling thought have announced that they had the key and that the full solution of reproducing intelligent behaviour was just a matter of a few more years of research! As all these pronouncements in the past have flopped, I refrain from making any claims now except to say that the ideas just sketched seem to me on the right track.

My overall conclusion is that I believe stochastic methods will transform pure and applied mathematics in the beginning of the third millenium. Probability and statistics will come to be viewed as the natural tools to use in mathematical as well as scientific modeling. The intellectual world as a whole will come to view logic as a beautiful elegant idealization but to view statistics as the standard way in which we reason and think.

## References

Davis, P. and Hersh, R., 1980, *The Mathematical Experience*, Birkhauser-Boston.

E, Weinan, Khanin, K., Mazel, A., and Sinai, Ya, 1997, Probability distribution functions for the random forced Burger's equation, Phys. Rev. Letters, **78**, 1904-1907.

Freiling, C., 1986, Axioms of symmetry: throwing darts at the real line, *J. Symb. Logic*, **51**, 190-200.

Gordon, N., Salmond, D., and Smith, A., 1993, Novel Approach to nonlinear/non-Gaussian Bayesian State Estimation, *IEEE Proc. F*, **140**, pp. 107-113.

Grenander, U., Chow, Y. and Keenan, D., 1991, *HANDS, A Pattern Theoretic Study of Biological Shapes*, Springer.

Gross, C., 1995, Aristotle on the Brain, *The Neuroscientist*, **1**, 245-250.

Hallinan, P., Gordon, G., Yuille, A., Giblin, P., and Mumford, D., 1999, *Two and Three-dimensional Patterns of the Face*, AKPeters.

Isard, M. and Blake, A., 1996, Contour tracking by stochastic propagation of conditional density, *Proc. Eur. Conf. Comp. Vision*, pp.343-356.

Jaynes, E. T., 1996-2000, *Probability Theory: The Logic of Science*, available at `http://bayes.wustl.edu/etj/prob.html`. To be published by Camb. Univ. Press.

Kanizawa, K., Koller, D., and Russell, S., 1995, Stochastic simulation algorithms for dynamic probabilistic networks, *Proc. Conf. Uncertainty in A.I.*, pp.346- 351.

Lauritzen, S. and Spiegelhalter, D., 1988, Local computations with probabilities on graphical structures, *J.Royal Stat. Soc.*, **B50**, 157-224.

Mumford, A. A. and Young, M., 1923, *Biometrika*, **15**, p.109-115.

Pearl, J., 1988, *Probabilistic Reasoning in Intelligent Systems*, Morgan-Kaufman.

Russell, B. and Whitehead, A. N., 1912, *Principia Mathematica, vol.2*, Cambridge Univ. Press.

Shelah, S. and Woodin, W. H., 1990, Large cardinals imply that every reasonable definable set is Lebesgue measurable, *Israel J. Math.*, **70**, 381-394.

Spencer, J., 1994 (2nd edition), *Ten Lectures on the Probabilistic Method*, SIAM.

Widrow, B., 1973, The rubber mask technique, *Pattern Recognition*, **5**, 175-211.