National Science Foundation

Division of Mathematical Sciences

Mathematics and Science

Dr. Margaret Wright
Prof. Alexandre Chorin

April 5, 1999

PREFACE

Today's challenges faced by science and engineering are so complex that they can only be solved through the help and participation of mathematical scientists. All three approaches to science, observation and experiment, theory, and modeling are needed to understand the complex phenomena investigated today by scientists and engineers, and each approach requires the mathematical sciences. Currently observationalists are producing enormous data sets that can only be mined and patterns discerned by the use of deep statistical and visualization tools. Indeed, there is a need to fashion new tools and, at least initially, they will need to be fashioned specifically for the data involved. Such will require the scientists, engineers, and mathematical scientists to work closely together.

Scientific theory is always expressed in mathematical language. Modeling is done via the mathematical formulation using computational algorithms with the observations providing initial data for the model and serving as a check on the accuracy of the model. Modeling is used to predict behavior and in doing so validate the theory or raise new questions as to the reasonableness of the theory and often suggests the need of sharper experiments and more focused observations. Thus, observation and experiment, theory, and modeling reinforce each other and together lead to our understanding of scientific phenomena. As with data mining, the other approaches are only successful if there is close collaboration between mathematical scientists and the other disciplinarians.

Dr. Margaret Wright of Bell Labs and Professor Alexandre Chorin of the University of California-Berkeley (both past and present members of the Advisory Committee for the Directorate for Mathematical and Physical Sciences) volunteered to address the need for this interplay between the mathematical sciences and other sciences and engineering in a report to the Division of Mathematical Sciences. Their report identifies six themes where there is opportunity for interaction between the mathematical sciences and other sciences and engineering, and goes one to give examples where these themes are essential for the research. These examples represent only a few of the many possibilities. Further, the report addresses the need to rethink how we train future scientists, engineers, and mathematical scientists.

The report illustrates that some mathematical scientists, through collaborative efforts in research, will discover new and challenging problems. In turn, these problems will open whole new areas of research of interest and challenge to all mathematical scientists. The fundamental mathematical and statistical development of these new areas will naturally cycle back and provide new and substantial tools for attacking scientific and engineering problems.

The report is exciting reading. The Division of Mathematical Sciences is greatly indebted to Dr. Wright and Professor Chorin for their effort.

Donald J. Lewis
Director (1995-1999)
Division of Mathematical Science
National Science Foundation

1 Overview

Mathematics and science[1] have a long and close relationship that is of crucial and growing importance for both. Mathematics is an intrinsic component of science, part of its fabric, its universal language and indispensable source of intellectual tools. Reciprocally, science inspires and stimulates mathematics, posing new questions, engendering new ways of thinking, and ultimately conditioning the value system of mathematics.

Fields such as physics and electrical engineering that have always been mathematical are becoming even more so. Sciences that have not been heavily mathematical in the past---for example, biology, physiology, and medicine---are moving from description and taxonomy to analysis and explanation; many of their problems involve systems that are only partially understood and are therefore inherently uncertain, demanding exploration with new mathematical tools. Outside the traditional spheres of science and engineering, mathematics is being called upon to analyze and solve a widening array of problems in communication, finance, manufacturing, and business. Progress in science, in all its branches, requires close involvement and strengthening of the mathematical enterprise; new science and new mathematics go hand in hand.

The present document cannot be an exhaustive survey of the interactions between mathematics and science. Its purpose is to present examples of scientific advances made possible by a close interaction between science and mathematics, and draw conclusions whose validity should transcend the examples. We have labeled the examples by words that describe their scientific content; we could have chosen to use mathematical categories and reached the very same conclusions. A section labeled “partial differential equations” would have described their roles in combustion, cosmology, finance, hybrid system theory, Internet analysis, materials science, mixing, physiology, iterative control, and moving boundaries; a section on statistics would have described its contributions to the analysis of the massive data sets associated with cosmology, finance, functional MRI, and the Internet; and a section on computation would have conveyed its key role in all areas of science. This alternative would have highlighted the mathematical virtues of generality and abstraction; the approach we have taken emphasizes the ubiquity and centrality of mathematics from the point of view of science.

2 Themes

As Section 3 illustrates, certain themes consistently emerge in the closest relationships between mathematics and science:

- modeling
- complexity and size
- uncertainty
- multiple scales
- computation
- large data sets.

[1] For compactness, throughout this document “mathematics” should be interpreted as “the mathematical sciences”, and “science” as “science, engineering, technology, medicine, business, and other applications”.

2.1 Modeling

Mathematical modeling, the process of describing scientific phenomena in a mathematical framework, brings the powerful machinery of mathematics---its ability to generalize, to extract what is common in diverse problems, and to build effective algorithms---to bear on characterization, analysis, and prediction in scientific problems. Mathematical models lead to “virtual experiments” whose real-world analogues would be expensive, dangerous, or even impossible; they obviate the need to actually crash an airplane, spread a deadly virus, or witness the origin of the universe. Mathematical models help to clarify relationships among a system's components as well as their relative significance. Through modeling, speculations about a system are given a form that allows them to be examined qualitatively and quantitatively from many angles; in particular, modeling allows the detection of discrepancies between theory and reality.

2.2 Complexity and Size

Because reality is almost never simple, there is constant demand for more complex models. However, ever more complex models lead eventually---sometimes immediately---to problems that are *fundamentally different*, not just larger and more complicated. It is impossible to characterize disordered systems with the very same tools that are perfectly adequate for well-behaved systems. Size can be regarded as a manifestation of complexity because substantially larger models seldom behave like expanded versions of smaller models; large chaotic systems cannot be described in the same terms as small-dimensional chaotic systems.

2.3 Uncertainty

Although uncertainty is unavoidable, ignoring it can be justified when one is studying isolated, small-scale, well-understood physical processes. This is not so for large-scale systems with many components, such as the atmosphere and the oceans, chemical processes where there is no good way to determine reaction paths exactly, and of course in biological and medical applications, or in systems that rely on human participation. Uncertainty cannot be treated properly using *ad hoc* rules of thumb, but requires serious mathematical study. Issues that require further analysis include: the correct classification of the various ways in which uncertainty affects mathematical models; the sensitivities to uncertainty of both the models and the methods of analysis; the influence of uncertainty on computing methods; and the interactions between uncertainty in the models themselves and the added uncertainty arising from the limitations of computers.

Uncertainty of outcome is not necessarily directly related to uncertainty in the system or in the model. Very noisy systems can give rise to reliable outcomes, and in such cases it is desirable to know how these outcomes arise and how to predict them. Another extreme can occur with strongly chaotic systems: even if a specific solution of a model can be found, the probability that it will actually be observed may be nil; thus it may be necessary to predict the average outcome of computations or experiments, or the most likely outcome, drawing on as yet untapped resources of statistics.

2.4 Multiple Scales

The need to model or compute on multiple scales arises when occurrences on vastly disparate scales (in space, time, or both) contribute simultaneously to an observable outcome. In turbulent combustion, for example, the shape of the vessel is important and so are the very small fluctuations in temperature that control the chemical reactions. Multiple scales are inherent in complex systems, a topic of great importance across science, whenever entities at microscales and macrolevels must be considered together.

When it is known in advance that phenomena on different scales are independent, one may rely on a separate model on each scale; but when different scales interact, or when the boundaries between scales become blurred, models are needed that allow interactions between scales without an undue sacrifice of structure or loss of information at any scale. A related complication is that the finiteness of computers limits the range of scales that can be represented in a given calculation; only mathematical analysis can overcome this built-in restriction.

2.5 Computation

Experiment and theory, the two classical elements of the scientific method, have been joined by computation as a third crucial component. Computations that were intractable even a few years ago are performed routinely today, and many people pin their hopes for mastering problem size and complexity on the continuing advent of faster, larger computers. This is a vain hope if the appropriate mathematics is lacking. For more than 40 years, gains in problem-solving power from better mathematical algorithms have been comparable to the growth of raw computing speed, and this pattern is likely to continue. In many situations, especially for multiscale and chaotic problems, fast hardware alone will never be sufficient; methods and theories must be developed that can extract the best possible numerical solutions from whatever computers are available.

It is important to remember that no amount of computing power or storage can overcome uncertainties in equations and data; computed solutions cannot be understood properly unless the right mathematical tools are used. A striking visualization produced over many days of computation is just a pretty picture if there are flaws in the underlying mathematical model or numerical methods, or if there are no good ways to represent, manipulate, and analyze the associated data.

It is also worthy of note that computation has come to permeate even the traditional core mathematical areas, which allot expanding roles for computation, both numerical and symbolic.

2.6 Large Data Sets

The enormous sets of data that are now being generated in many scientific areas must be displayed, analyzed, and otherwise “mined” to exhibit hidden order and patterns. However, large data sets do not all have similar characteristics, nor are they used in the same way. Their quality ranges from highly accurate to consistently noisy, sometimes with wide variations within the same data set. The definition of an “interesting” pattern is not the same nor even similar in different scientific fields, and may vary within a given field. Structure emerges in the small as well as in the large, often with differing mathematical implications. Large data sets that need to be analyzed in real time---for instance, in guiding surgery or controlling aircraft---pose further challenges.

3 Examples

The examples in this section, described for a general scientific audience, illustrate the scientific and technological progress that can result from genuine, continuing, working relationships between mathematicians and scientists. Certain well publicized pairings, such as those between modern geometry and gauge field theory, cryptography and number theory, wavelets and fingerprint analysis, have been intentionally omitted---not to slight their remarkable accomplishments, but rather to demonstrate the breadth and power of connections between mathematics and science over a wide range of disparate, often unexpected, scientific applications.

3.1 Combustion

Combustion, a critical and ubiquitous technology, is the principal source of energy for transportation, for electric power production, and in a variety of industrial processes. Before actually building combustion systems, it is highly desirable to predict operating characteristics such as their safety, efficiency, and emissions. Mathematicians, in collaboration with scientists and engineers, have played and continue to play a central role in creating the analytical and computational tools used to model combustion systems. Two examples---modeling the chemistry of combustion and engineering-scale simulation---illustrate the ties between mathematics and practical combustion problems.

Modeling the chemistry of combustion. To model combustion it is necessary to understand the detailed chemical mechanisms by which fuel and air react to form combustion products. For a complex hydrocarbon fuel such as gasoline, whose burning involves thousands of distinct chemical species, one must identify the reactions that are most important for the combustion process. The rates of reaction, which are sensitive functions of temperature and pressure, must also be estimated, along with their energetics, e.g. the heats of formation of the various species.

For more than twenty years, mathematicians and chemists have worked together on computational tools that have become critical to the development of reaction mechanisms. The need for robust and accurate numerical solvers in combustion modeling was clearly understood as early as the 1970s. In response to this need, algorithms and software for solving stiff systems of ordinary differential equations were developed and combined into integrated packages for chemically reacting systems, such as the Chemkin package developed at the Sandia National Laboratory. Given arbitrarily complex chemical reaction mechanisms specified in a standard format, Chemkin automatically generates an interface to numerical methods that compute various chemically reacting systems. These include spatially homogeneous systems as well as a variety of one-dimensional systems, such as premixed flames, opposed-flow diffusion flames, and detonation waves.

The mathematical and numerical analysis embodied in Chemkin has been a key ingredient in designing and evaluating mechanisms, including those in wide laboratory use. The existence of a reliable and generalizable mathematical model facilitates the testing of new ideas in mechanism design, since the effects of modifying a chemical mechanism can be assessed directly. Finally, the mathematical software is not only sufficiently robust to model arbitrarily complex chemical reaction mechanisms, but also accurate enough so that the numerical error is negligible relative to laboratory measurements.

Chemkin represents an amalgam of mathematical analysis, numerical methods, and software development. The history of Chemkin illustrates the fact that in many application areas advanced mathematical ideas are more likely to be used by scientists and engineers if they are embodied in software.

Engineering-scale simulation. The goal in this area is to represent the three-dimensional fluid dynamics and other physical processes as they occur in combustion devices such as internal combustion engines, industrial and utility burners, and gas turbines. Two issues make these simulations particularly challenging. The first is the number and complexity of the physical processes that must be represented, which include fluid dynamics, heat and mass transport, radiative heat transfer, chemical kinetics, turbulence and turbulent combustion, and a variety of multiphase fluid flow phenomena. The second is the enormous range of length and time scales in such systems. The relevant physical processes must operate simultaneously on scales ranging from the smallest turbulent fluctuations (10^{-6} meters) up to a utility boiler (100 meters).

Mathematicians have consistently been at the forefront in developing innovative methods for modeling engineering combustion problems. Within computational fluid dynamics, a huge field that encompasses numerous applications, many of the mathematical methods have arisen as a direct response to specific difficulties presented by combustion problems. Examples include novel discretization techniques, such as high-order accurate finite-difference methods and vortex methods; adaptive gridding techniques, which estimate the error as a calculation is running and locally increase or decrease the grid density to maintain a uniform level of accuracy; and new methods for problems in complex geometries, such as the overset grid and embedded boundary methods.

A major mathematical contribution has been asymptotic analysis that makes possible an understanding of the coupling between different physical processes in these complex systems; insights from asymptotic analysis are used to find stable and accurate representations of these processes in terms of simpler subprocesses. Examples include the use of low Mach-number asymptotics to eliminate zero-energy acoustic waves while retaining the bulk effects of compression and expansion due to heat release, and front-tracking methods based on a separation-of-scales analysis for thin premixed flames.

Today, packages such as Chemkin are part of the standard toolkit for combustion researchers and engineers. New numerical methods for engineering-scale simulations of combustion systems have been extensively implemented as research codes, and are slowly making their way into production engineering software.

Looking ahead, the requirements of combustion simulation suggest promising directions for mathematics research that will make new science possible. Even with the most powerful computers, it is impossible to represent directly all of the processes involved at all of the relevant length scales. Instead, one needs to introduce sub-grid models that capture the effect on the large scales of all the scales below the resolution limit of the calculation. In the area of chemical reaction mechanisms, this corresponds to the development of *reduced* mechanisms, i.e., reaction mechanisms with a few tens of species that accurately represent energy release and emissions. The systematic development of reduced mechanisms will involve a variety of mathematical tools, from statistical analysis and optimization to dynamical systems.

For engineering-scale simulations, modeling at the sub-grid scale is a central requirement for future progress. The development of sub-grid models for turbulent combustion is particularly difficult, since chemical reactions are sensitive to small-scale fluctuations in temperature and composition. The effect of these fluctuations must be separated from the larger-scale dynamics representable on the grid. There has been renewed progress in turbulence modeling in recent years, based on ideas from mathematical statistical mechanics, and extension of these ideas to turbulent combustion represents a substantial mathematical challenge; any successes will have enormous practical consequences.

3.2 Cosmology

Cosmology, which once consisted of speculations based on extremely scarce observations, has become a science rich in both data and theory. The relativistic “hot big bang” model for the expanding universe is widely accepted today and supported by a substantial body of evidence; just as significantly, no data are inconsistent with this model. But the standard cosmology leaves unanswered certain key questions about the nature and evolution of the universe, including the quantity and composition of energy and matter, and the origin and nature of the density perturbations that seeded all the structure in the universe. While a promising paradigm for extending the standard cosmology---inflation plus cold dark matter---is being developed and tested, many fundamental cosmological issues remain to be resolved or clarified. (“Inflation” refers to the quantum-mechanical fluctuations occurring during a very early burst of expansion driven by vacuum energy; cold dark matter consists of slowly moving elementary particles left over from the earliest fiery moments of the universe.) Mathematical progress in two broad areas will be essential for cosmology: techniques for dealing with massive data sets and large-scale, nonlinear, multiscale modeling and numerical simulation.

Massive data sets. As cosmology moves toward becoming an exact science, major mathematical challenges arise in coping with, displaying, understanding, and explaining the unprecedented avalanche of high-quality data expected during the next few years. To mention only a few sources, NASA's MAP and the European Space Agency's Planck Surveyor will map the full sky to an angular resolution of 0.1°, allowing determination of the mass distribution in the universe before nonlinear structures formed. The Sloan Digital Sky Survey will obtain the redshifts of a million galaxies over 25% of the northern sky, and the Two-Degree Field Survey will collect 250,000 redshifts in many 2° patches of the southern sky, together covering around 0.1% of the observable universe and mapping structures well beyond the largest presently known size. In addition, experiments at accelerators, nuclear reactors, and large-underground detectors are planned or in place to search for neutralinos, explore the entire theoretically favored mass range, and pursue neutrino mass. The quantity, quality, and nature of the data require connections between mathematics and cosmology. Although some generic principles of data analysis have emerged, the various features to be "mined" in cosmological data differ from one another in ways whose definition remains far from precise. The patterns of interest change from application to application, and may even vary when several uses are made of the same data set. In contrast to data from other scientific areas, the cosmological data are likely to be of very high quality; thus it will be important to squeeze every possible insight from each data set.

A further striking feature of cosmological data is the vastness of the scale ranges in almost every dimension. Data will be gathered not only on the scale of galaxies, but also from particle physics; the "hot" part of big bang cosmology implies the need for physics of ever-higher energies and ever-shorter times.

Finally, astronomical data not only arrive at very high speed, but patterns detected in real time may be used to control subsequent data collection adaptively---for example, to concentrate on regions where something interesting is being observed. Careful mathematical analysis will be needed because techniques appropriate for "on the fly" data mining are quite different from those used to examine data at leisure.

Modeling and simulation. The mathematical models in cosmology typically involve highly nonlinear coupled partial differential equations that cannot conceivably be solved analytically---for instance, the equations may model turbulence in nuclear explosions that occur when stars blow themselves apart. Small differences in the mathematical form of these equations can lead to big variations in the predicted phenomena. Cosmological models need to be complex enough to capture all the phenomena reflected in the data, yet amenable to analysis. Important modeling questions arise in the inverse problem, reasoning backwards from observations and images to find the laws that created them. The hope is that, by varying the initial conditions and the parameters embedded in mathematical models, simulations can reveal the fundamental parameters that define the universe, such as the mean density and Einstein's cosmological constant Λ.

Like the associated data, cosmological models contain enormous ranges of scales that pose difficulties for both mathematical analysis and numerical solution. Creating *a priori* cutoffs that define different scale regimes is a common tactic, but it breaks down as the ends of the scales approach each other---when the noise for a large scale becomes comparable to the signal for the next-smaller scale. Subtle mathematical modeling is essential to separate the phenomena that can be ignored from those that count.

Carefully executed large-scale simulations match observations well, and have become a standard tool in modern astrophysics. Cosmological calculations consume a large portion of the available supercomputer cycles in the United States, and worldwide as well. This is because solving the complex partial differential equations of cosmology over the wide multidimensional range of scales for problems of realistic size is a massive undertaking at the edge of current mathematical and computational capabilities.

To illustrate these points, consider the formation and evolution of galaxy clusters, the largest objects in the universe. For a simulation to be credible, enormous dynamic ranges in size and density are required to resolve individual galaxies within a cluster; the range of mass is perhaps 10^9, over a time period of 10 billion years. One approach is to begin with a “box” (part of the universe) that is initialized with a large number (say, 10 million) of uniformly distributed particles, and then to follow the motion of each particle as its position and velocity are perturbed following theoretical predictions.

This approach poses formidable difficulties for numerical methods in addition to those arising from the already-mentioned nonlinearities and ranges of scale: the particles move non-uniformly, model geometries are highly complex, and there is a demand for ever-finer resolution. A fruitful arena for mathematical analysis is the effect of decisions about partition into scales on numerical accuracy; here the recent mathematical work on particle methods and on fast summation and multipoles may be of key importance.

Since cosmological calculations will continue to tax the capabilities of the highest-performance available hardware, further mathematical and algorithmic ingenuity is needed to make the implementations of these simulations run efficiently on parallel machines without inordinate specialization for a particular hardware configuration. Taking advantage of new computer architectures without unduly compromising generality is a problem for all applications that strain today's high-performance computers.

3.3 Finance

Modern finance, although not a science in the traditional sense, is intertwined with mathematics, and the connection is not limited to theory---mathematics is a central feature in the day-to-day functioning of the world's financial markets. Mathematics and finance are tightly connected in the two areas of derivative securities and risk management.

Derivative securities. In recent years, headlines about business have repeatedly mentioned “derivatives”. A financial derivative is an instrument that derives its value from other, more fundamental instruments, such as stocks, bonds, currencies, and commodities (any one of which is called an *underlying*). Typical derivatives include options, futures, interest rate swaps, and mortgage-backed securities. The Nobel-prize-winning papers on option pricing containing the famous Black-Scholes partial differential equation were published in 1973 as the Chicago Board of Options Exchange was being established, and within months the Black-Scholes model became a standard tool on the trading floor. Worldwide, the volume of trade in derivatives has since grown to rival the volume of trade in equities. One of the reasons for this phenomenal growth is the existence of reasonably reliable mathematical models to guide their pricing and trading.

In theory, derivatives are redundant because they can be synthesized by dynamic trading in the underlying instruments. Trading in derivatives thus rests on the possibility of finding the fair price of a derivative. Under standard assumptions, the unique fair price of an option can be found from the Black-Scholes equation. However, certain key parameters need to be determined before this equation can be used in practical settings.

One of these parameters, the *volatility*, has been the subject of intense mathematical and algorithmic attention for almost twenty years. The original Black-Scholes model requires the estimation of a constant volatility derived from a diffusion model of the underlying's price process. Multiple approaches have been devised to calculate this form of volatility---for example, using weighted past data, or selecting the *implied volatility* corresponding to a specific similar traded option with the same underlying. (The implied volatility of a traded option is the value that, substituted into the Black-Scholes equation, produces the known price of the option; implied volatility is calculated by solving a one-dimensional inverse problem.)

The classical Black-Scholes model has known limitations that are often displayed through the “smile effect”---the characteristic U-shaped curve that relates implied volatility for comparable options to the price associated with buying or selling the underlying. Existing models of volatility are not completely satisfactory, with *hedging* playing a major role in the difficulties. Hedging is related to the sensitivity of the option's value to different parameters; the choice of volatility may have a large effect on the hedging strategy. There is extensive mathematical research today on formulating stochastic models of volatility (with, in some cases, widely differing time scales as a key feature), and on modeling volatility as a two-dimensional surface that depends on properties of the underlying. In addition, new approaches involving approximation and optimization are being developed for calculating volatility.

Today's derivative models include heavy doses of continuous-time martingale theory, changes of measure for stochastic integration, the fine structure of stochastic processes, supermartingales and potential theory, stochastic calculus, and partial differential equations. The continuing creation of more complex derivatives calls for new mathematics in these areas as well as in simulation, large-scale optimization, and real-time analysis of large data sets.

In the financial world, a few minutes or even seconds may make a major difference in profit, so that an ideal financial model should be able to make accurate predictions in the very short term. However, the relationship between models and data is very different in finance than in experimental sciences: the world's markets do not lend themselves to meaningful large-scale experiments designed to test or validate models. Thus models are of necessity evaluated by their ability to track the huge quantities of financial data generated throughout the day around the world. A further contrast to other scientific areas is that neither the quantity nor quality of financial data is likely to improve, so that new models must make do with the same forms of data that are available today.

Risk management. The creation of an international financial system in which large volumes of capital move quickly has led to the new challenge of *risk management*. Despite some spectacular failures to manage risk in derivatives markets, such as the 1998 debacle of Long Term Capital Management, derivative securities are too useful to disappear. Hence strategies are needed for managing the risks associated with derivatives and other financial instruments.

The long-standing assumption, originating with Markowitz in the 1950s, that stock returns are normally distributed is known to be an inadequate approximation to reality in times of crisis. Indeed, repeated studies have found stock returns to have “fatter tails” than the normal distribution and models based on the normal assumption can err by ten to twenty standard deviations.

In addition, the implicit idea in the Black-Scholes model and its successors is to synthesize derivative securities or a portfolio of derivative securities, thereby allowing institutions to hedge the risk associated with their business by owning financial instruments which offset this risk. But liquidity can disappear in times of crisis, so that hedging may become impossible. Even in normal times, some derivative securities cannot be hedged, for example a security that offsets the default risk of a corporate bond.

A yet-to-be developed mathematical theory would show how to decouple a portfolio into two parts, one part whose risk can be hedged and another part that is "purely unhedgeable". One possible strategy is to project the space of all portfolios onto the subspace of hedgeable portfolios, but the complexity and difficulties of such an approach are daunting from both theoretical and computational perspectives.

The standard portfolio risk measure is *value at risk*, "VaR", which is the probability that the portfolio will lose money exceeding a specified threshold within a specified time period. There are several objections, conceptual as well as practical, to VaR, primarily that it assigns the same risk measure to all portfolios with the same probability of loss exceeding the threshold, regardless of the distribution of loss above the threshold. It also fails to satisfy the desirable mathematical property of subadditivity, since the sum of the VaRs of two portfolios can be less than the VaR of their sum; this encourages institutions to play accounting games, subdividing dangerous positions into smaller ones entirely for risk-accounting purposes. A further disadvantage is that VaR assumes normally distributed returns, and hence tends to give optimistic values in the tail of the loss distribution, which is where risk matters most.

The mathematics of risk management is in its infancy, building on ideas such as extreme value theory from actuarial science. The distributions describing extremal events are well understood, but it is not yet known how to build extreme-value models based on large numbers of jointly distributed random variables. A fundamental open problem in this area is defining how to measure risk in any particular model. An appealing approach, currently under active exploration, is to devise mathematical properties that are desirable in a risk measure and then define a set of risk measures that possess these properties.

The problems of quantifying, computing, and managing risk are likely to pose substantial mathematical challenges into the foreseeable future.

3.4 Functional Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is a well-established tool for analyzing the structure of the brain. Starting in the early 1990s, functional MRI (fMRI; the "f" is by convention lower-case) began to be used to study brain *dynamics*. The underlying principle of fMRI is related to the properties of blood within active areas of the brain. "Blue" blood (deoxygenated hemoglobin) is more paramagnetic than "red" blood (oxygenated hemoglobin), so that the MR signal from blue blood is stronger. In the late 1980s, positron emission tomography research showed that, although active areas of the brain require a larger supply of blood, the corresponding increase in available oxygen is not used. Consequently, the blood leaving active regions of the brain contains relatively more oxygen and interferes less with the local magnetic field, which means that the MR signal in the vicinity of active regions shows an apparent gain. By comparing the averages of MR images taken at close time intervals while a brain function is activated and deactivated, the active areas can be identified.

To test the brain function of working memory, for example, a subject is presented with a sequence of letters arriving at one per second, and is asked to press a button when a letter is repeated. Next, the subject is asked to press the button when a letter is repeated with a different letter in between ("one back"); and so on. Most people cannot sustain "four back" for more than a few minutes, and the work in the brain as *n* in "*n* back" increases can be measured and correlated with *n*.

In studying brain function through fMRI, the key is to *compare* images rather than to study an individual image in detail. Images are compared by determining the voxels that are "significant", i.e., those that have changed by more than a given tolerance between images. The MRI used to observe brain structure requires approximately 20 minutes to produce a single fine, detailed image. By contrast, in fMRI a sequence of images is collected rapidly, to observe the brain dynamics, and there is an obvious tradeoff between the time per image and image quality. The role of mathematics (in the form of statistics) in fMRI is to analyze the image data. The data sets collected are extremely large (a typical experiment produces between 0.5 and 500 gigabytes of data), and also extremely noisy, thus presenting multiple statistical difficulties.

Many sources of noise are present in fMRI data; some are understood, while others remain mysterious. The signal induced by neural activation has approximately the same magnitude as noise in the experiments, which means that many images need to be acquired in each experiment to obtain a meaningful result. The noise process in fMRI has a complex distributional structure that is not yet fully understood---for example, signal variance depends on the mean in an unknown, nonlinear way, and significant spatial correlations exist that depend on how the data are collected. Outliers of various forms are frequent, and the variations between individual brains are enormous. Most importantly, statistical analysis of fMRI data needs to be built on a detailed understanding of the structure of the noise, which means understanding the scientific elements of fMRI: physics, MRI technology, and theories of brain functioning. Consequently, statisticians in this area necessarily work with teams of physicists, electrical engineers, psychologists, neurologists, technologists, and computer scientists.

To obtain the best possible images, the data need to be corrected to reduce the effects of noise, which arises from at least two sources: the hardware and the subject. In the hardware, there may be a lack of uniformity in the main magnetic field, or a lack of linearity in the gradient field. In addition, the analog-to-digital converter may be miscalibrated, or mistimings of resonant gradients may cause "ghosts" in the images. The main source of variation originating in the subject is movement of the brain, which can result from, for example, rigid motion of the head, the almost-periodic compression and vertical movement caused by the cardiac cycle, and distortions caused by respiration.

To deal with the noise, two approaches are being taken simultaneously: removing or reducing the noise at its source through engineering; and, through mathematics, modeling the data and noise, then adjusting the predicted variation. With the latter approach, the goal is to develop a mathematical model that accurately relates the data to parameters of interest, but this remains a daunting task. Substantial progress has been made by successively estimating and correcting for each effect known to cause noise. To date, these effects include analog-to-digital miscalibration, gradient mistimings, receiver drift, subject head motion, and shot noise. After these corrections, the images are reconstructed by a fast Fourier transform and then the (still unexplained) voxel-wise trend over time is removed. Finally, statistical methods such as *t*-tests are used to assess the effect of the experimental paradigm.

New statistical and computational techniques have already contributed substantially to the quality of fMRI data. It is now possible, for instance, to estimate and correct for rigid motions of the brain as small as 50 microns. Statistical models can also account for differential brain response, and have extended motion correction between images using a fully three-dimensional method. Incremental task effects from a variety of administered cognitive tests have been quantified by novel statistical methods, and statistical methods of spatial growth curves have been extended to quantify changes in the pattern of activation over time. More powerful statistical tests are still needed; *t*-tests are often sufficient, but subtler methods will be called for as MR techniques and the cognitive questions become more complex.

Contributions from statistics have answered several important questions about fMRI data---for example, how to make multiple comparisons while retaining the power of statistical tests, and what happens if the same experiment is repeated. However, statisticians working on fMRI have found that every answer leads to a new question, and that substantial mathematical challenges arise from every new question, with no end in sight.

There has been tremendous progress not only in conceptual techniques for modeling and resolving the noisy data, but also in numerical and computational algorithms. Several years ago, processing the data from a 15-minute experiment required 12 hours of computation; now it takes three seconds. Concurrently, there have been continuing, rapid gains in the achievable spatial resolution---for example, an eight-fold improvement between 1996 and 1997. Most of the gains in speed and accuracy are attributable to better mathematical algorithms, not to increased computational power.

The cognitive science driving fMRI has also advanced; one interesting discovery was that reading more complex sentences causes greater brain activity in precisely the ways predicted by theory and earlier, more primitive external measurements of eye movements. In ongoing projects, fMRI is being used to study the cognitive and brain activity characteristics of high-functioning autistic subjects, and to examine brain plasticity and rehabilitation in aphasia therapy.

One final point of interest is that certain aspects of the statistical techniques developed in the context of fMRI generalize to analysis of seismic data collected by geophysicists in oil exploration.

3.5 Hybrid System Theory and Air Traffic Management

Hybrid system theory, a field of applied mathematics abutting control theory and computer science, has an enormous potential for impact on practical problems. Hybrid systems can be loosely defined as systems that allow the interaction of discrete events and continuous dynamics; hybrid system theory attempts to prove properties such as reachability and stability. Discrete event models naturally accommodate linguistic and qualitative information, and are used to model modes of operation of a single system, for example an aircraft or the interaction of several aircraft. The continuous dynamics in a hybrid system model physical processes, such as the continuous response of an aircraft to changes in the positions of aileron and throttle.

Hybrid systems are good models of complex reactive systems, in which physical processes interact with man-made automated environments; algorithms developed to analyze and control the behavior of hybrid systems may therefore be used in the design of automatic controls for these systems. A common real-world example of a hybrid system arises when advanced automation is introduced into manually operated systems in order to enhance performance and flexibility while significantly reducing the workload of human operators. Accompanying this increase in automation, however, is the necessity of ensuring that the automated system always performs as expected. This is especially crucial for *safety-critical* systems: if a telephone switch crashes or a power grid node goes down, lives are usually not lost; if an error occurs in the automated avionics in a commercial jet, the results could be disastrous.

Many of today's safety-critical systems are growing at a rate that will make their manual operation extremely difficult if not impossible in the near future. The air traffic control system is an example of such a system. Air traffic in the United States is expected to grow by 5% annually for the next 15 years, and rates of growth across the Pacific Rim are expected to be more than 15% a year. Even with today's traffic, ground holds and airborne delays in flights due to congestion have become so common that airlines pad their flight times with built-in allowances. Aging air traffic control equipment certainly contributes to these delays: the plan view displays used by controllers to look at radar tracks and flight information are the very same ones that were installed in the early 1970s, and they fail regularly. The computer systems that calculate radar tracks and store flight plans were designed in the 1980s, using software written in 1972.

The introduction of new computers, display units, and communication technologies for air traffic controllers will help alleviate the problems caused by failing equipment, yet the Federal Aviation Administration admits that any significant improvement will require that many of the basic practices of air traffic control be automated. For example, today's airspace has a rigid route structure based on altitude and on ground-based navigational “fixes”. The current practice of air traffic controllers is to route aircraft along predefined paths connecting fixes, to manage the complexity of route planning for several aircraft at once. The rigid structure puts strict constraints on aircraft trajectories, which could otherwise follow wind-optimal or “user-preferred” routes (routes that are shorter or involve lower fuel consumption because of tailwinds). Also, while a data link between aircraft and ground is being considered as a replacement for the current voice communication over radio channels between pilot and controller, there is a limit to the amount of information processing that a controller can perform with these data. Recent studies indicate that, if there is no change to the structure of air traffic control, then by the year 2015 there could be a major accident every 7 to 10 days; obviously this cannot be permitted to happen.

The main goal of air traffic control is to maintain safe separation between aircraft while guiding them to their destinations. However, its tight control over the motion of every aircraft in the system frequently causes bottlenecks to develop. Uncertainties in positions, velocities, and wind speeds, along with the inability of a single controller to handle large numbers of aircraft at once, lead to overly conservative controller actions and procedures. An example is the set of methods used by air traffic controllers to predict and avoid conflicts between aircraft. If a controller predicts that the separation between two aircraft will become less than the regulatory separation, the controller will issue a directive to one or both of the pilots to alter their paths, speed, or both. Often the resolution is not needed, and usually it is too drastic. Also, user-preferred routes are disallowed because of the requirement that prescribed jetways be used.

As a result of all these difficulties, there is a widely perceived need in the air traffic, airline, and avionics communities for an architecture that integrates data storage, processing, communications, and display into a safe and efficient air traffic management system; a new air traffic system has been proposed that involves the Global Positioning System and a datalink communication protocol called Automatic Dependent Surveillance for aircraft-aircraft and aircraft-ground communication. While the degree of decentralization and level of automation in such a system are still under debate, the integrity of any automated functionality in a new air traffic management system depends on a *provably safe* design as well as high confidence that the control actions will not fail.

This level of reliability requires accurate models, techniques for verifying that the design is safe to within the accuracy of these models, and procedures for synthesizing the system's control actions. Hybrid system researchers have designed models and control laws for two systems: a provably safe algorithm for resolving trajectory conflicts between aircraft, and a provably safe algorithm for a single aircraft to switch between different flight modes. A rigorous notion of “safety” in each case is crucial. In the conflict resolution problem, the system is safe if the aircraft always maintain minimum separation from each other. In the mode-switching problem, system safety means that the state of the aircraft remains within minimum and maximum bounds imposed on its velocities, angles, etc., so that the aircraft does not stall and plunge out of the sky.

The hybrid system associated with air traffic control models the discrete dynamics with finite-state automata whose transition functions describe the mode-switching logic, and uses nonlinear ordinary differential equations to model the continuous dynamics. The system includes continuous as well as discrete variables to model parameters that the designer may manipulate (such as a flight mode switch in an on-board flight management system) and disturbance parameters that the designer must control against (such as an aircraft entering the five-mile-radius protected zone around another aircraft). Using analysis based on traditional discrete and continuous optimal control techniques, and on two-person zero-sum game theory for automata and continuous dynamical systems, partial differential equations can be derived whose solution describes exactly those states (aircraft positions, velocities, accelerations, and modes of operation) that the system may reach from a given initial state. By analyzing these reachable states, it is possible to determine automatically those configurations that the system must be prevented from entering if safety is to maintained.

Ten years ago such a method would have been prohibitively computationally expensive, but advances in computational power and new fast methods for integrating partial differential equations have made such solutions feasible even for real-time applications such as on-board autopilots and computer-aided tools for air traffic controllers. The same approach has been applied to design conflict resolution maneuvers for multiple aircraft and to verify the mode-switching logic for vertical flight modes in an aircraft's flight management system.

3.6 Internet Analysis, Reliability, and Security

The Internet is one of the most talked-about and written-about phenomena of the late twentieth century. Data traffic on the Internet has grown exponentially since the early 1980s---there were 235 IP hosts on the Internet in 1982, 100,000 in 1989, and more than 30 million in 1998. The most optimistic extrapolations
have consistently underpredicted the continuing expansion of the Web, which is known within the Internet research community as a “success disaster”; because the Internet has succeeded beyond anyone's expectations, it is not prepared or able to cope with the consequences. Problems with the Internet are likely to escalate as popularity of the Web spreads; the efficiency, reliability, and security of the Internet are becoming important to an increasing fraction of the population. All of these areas are obvious candidates for new connections between mathematics and communications technology.

There is a long history of involvement by mathematics in the development of existing voice communication networks---in fact, traditional teletraffic theory is widely regarded as one of the most successful applications of mathematical techniques in industry. Mathematical models of voice traffic and call arrivals at network links have been available for at least 50 years. These models typically involve only a few parameters, they are associated with intuitively satisfying physical interpretations, and their predictions have consistently matched measured data. Their well understood mathematical structure has led to further applications of mathematics in telecommunications---for example, in designing highly optimized network management and control systems.

But any expectation that the known mathematics of teletraffic theory can be generalized to Internet traffic is doomed to disappointment. In almost every dimension, Internet traffic is *completely different* from voice traffic. Because computers do not communicate with other computers in the same way as humans speaking on the telephone, the old mathematical properties no longer apply. Most strikingly, both length and transmission rates for data traffic range across scales that are unimaginable for voice connections: data connections may last for days, and high-end users are already transmitting data at hundreds of megabits per second, with higher rates regularly becoming available. Furthermore, data network traffic displays *multiscale burstiness*---it arrives in fits and starts, interspersed with gaps, and this burstiness persists over three orders of magnitude in time scales. Standard voice traffic, by contrast, is bursty when observed over short time intervals such as 100 milliseconds, but is essentially smoothed out over longer periods of (say) one hour.

Existing networks, designed for voice traffic, are under stress. Information on the Internet is sent using the Internet Protocol (IP); when too many data packets arrive, routers keep them in buffers until traffic is reduced. If traffic is heavy for a sustained period, buffers fill up and packets are “dropped”. From an engineering perspective, Internet traffic plays havoc with standard voice network design: there is a need for big buffers in routers and switches to avoid loss of data packets when buffers overflow; links may be saturated without warning at any time, so that safe operating points must be chosen conservatively; and individual users may experience poor response even though overall network performance is satisfactory. Internet users today routinely encounter delays in access and sluggish performance that are essentially unknown in voice communication, and these problems are likely to become more severe as Internet traffic grows.

Many networking experts argue that the mathematics needed to model the Internet will be radically different from traditional teletraffic theory, and the topic of Internet-based mathematics is in a state of lively ferment. For example, the multiple time and rate scales observed in Internet traffic have led to work on scaling phenomena, a multidisciplinary field that includes mathematicians, network engineers, physicists, and control theorists. Much press has been devoted to the idea that Internet traffic processes can be modeled effectively in terms of fractal and multifractal scaling behavior---ideas that have been embraced by some but rejected by others. Approaching the problem from another angle, work on Internet analysis has been done using renormalization group techniques and mean-field theory. For the problem of controlling data networks, mathematicians have begun looking at paradigms of pattern formation, self-organization, and adaptation.

Irrespective of the mathematical constructs used, a universal theme in modeling and analysis of the Internet is the importance of *data.* Because the Internet's behavior is emphatically not captured by standard teletraffic models, researchers in this area rely on large quantities of multidimensional data gathered over wide-ranging time scales. The size and complexity of these data sets create a further mathematical challenge of devising methods that can meaningfully manage, manipulate, represent, and visualize the data. An important issue in handling these particular large data sets is their inherent extreme variability in scale.

Additional mathematical questions related to the Internet arise from concerns about reliability and security. As Internet connectivity expands, there are more and more opportunities for damage by malicious users---for example, targeted sites can be and have been crippled by deliberate saturation. The Internet's history of functioning without regulation means that systems are needed to *detect* attacks across the network in real time. Network intrusion detection is being approached by designing monitors that can be added to a network without modifying the hosts; such a property is essential when dealing with several thousand heterogeneous, individually administered hosts.

Of course, any network monitor can itself be subject to attacks intended to subvert the monitoring; hackers attempting to break in might well attack the monitor also. Such attacks may take several forms, each progressively more subtle and difficult to detect: overload attacks, where the strategy is to overburden the monitor so that it cannot keep up with the data stream; crash attacks, in which the goal is to knock the monitor out of service; and subterfuge attacks, in which the attacker tries to mislead the monitor about the meaning of the traffic that the monitor is analyzing. Each of these forms of attack calls for a different mathematical model that allows the attack to be detected in real time and then protects against it.

Mathematics is also needed to define and verify protective techniques such as congestion control. The end-to-end congestion control techniques of the Transmission Control Protocol (TCP) have been critical in the robustness of the Internet, but the Internet has ceased to be a small, close user community. Hence it is no longer possible to rely on end-nodes to cooperate in achieving end-to-end congestion control, nor on developers to include congestion control in their Internet applications.

Several distinct varieties of congestion arise from *unresponsive flows* that do not use end-to-end congestion control, implying that they do not reduce their load on the network when subjected to packet drops. Without congestion control, well-behaved traffic will reduce its sending rates in response to congestion, leading to a situation in which the uncooperative flows shut out the responsive traffic. In addition to this kind of unfairness, congestion collapse---a decrease in useful work by the network because of an increase in load---may occur in various forms. For example, “classical” congestion collapse occurs when there is unnecessary retransmission of packets. Undelivered packets can cause congestion collapse when bandwidth is wasted by transmitting packets that are dropped before they reach their destination; the latter situation is exacerbated by applications that willfully raise their sending rate as more packets are dropped. Research on congestion control involves queueing theory, scheduling algorithms, and fairness metrics. Inevitably, further mathematical complexity will be needed to blend modeling and network measurements as well as (eventually) policy issues.

3.7 Materials Science

Mathematical and computational techniques are assuming an increasing role in materials science, as illustrated by two areas---new materials and multiscale phenomena.

The search for new materials. Since prehistoric times, the *systematic method*---changing synthesis or processing variables over a limited range and then measuring properties of the resulting samples---has been the main tool in materials science. The classical variables are composition, heat treatment time and temperature, and quantities that influence formation of a specimen into a certain shape. With modern methods of synthesis, this process encompasses a wide range of controllable variables associated with thin films, composites, microscale and nanoscale devices, and electronic, magnetic, and dielectric materials.

Despite its successes, the systematic method can be inefficient or inappropriate in some situations. A common example occurs when the tolerances that define a new material are tight relative to the possible range of variable values. Consider shape memory materials, a class of materials that can undergo large plastic deformations, but recover their original shape upon heating. These materials are part of a revolution in biomedical technology, the $500 million, two-year-old technology of *stents.* Stents, placed in the coronary artery using a guidewire (often made out of shape memory material as well), in many cases allow an outpatient procedure rather than difficult bypass operations. The most important shape memory material in stents, an alloy of nickel and titanium, shows crucial differences in behavior as its composition varies from 50.2% to 50.6% nickel; furthermore, there are interesting alloys of nickel, titanium, and copper, and even quaternary alloys. If 0.1% is conservatively regarded as an acceptable tolerance, then it becomes extremely difficult to make many samples of slightly different composition and test their properties. If the parameters involved in heat treatment are varied as well, the systematic method is simply not practical.

The systematic method is also unlikely to discover entirely unexpected behavior---for example, a previously unknown microelectronic property that occurs in a film having a certain precise thickness, configuration, orientation or defect structure. The special behavior could not be inferred by looking at a trend based on small changes from a known sample; in such circumstances, the only path to new materials is through mathematics.

In cases where the systematic method cannot be used to find new materials, mathematical theory is playing an ever-growing role on two fronts separated by huge length and time scales. The first stems from improvements in continuum theories of materials. There is an emerging understanding of how to model and simulate accurately the growth of a new phase, including its complex geometrical shape and topology. An instance of this work is the development of materials with large magnetostriction. (Magnetostrictive materials convert magnetic energy to mechanical energy, and vice versa.) In the 1960s, the class of "giant" magnetostrictive materials was discovered using an ingenious strategy that relied on the inherently large magnetostriction of some rare earth metals. Recently, guided by gains in understanding of the theory of micromagnetics, predictions were made of a new class of materials with even larger magnetostriction. The mathematical theory not only directly predicted the mechanism of magnetostriction, but also guided the alloy development and subsequent experiments that revealed the effect. The resulting class of materials shows a magnetostrictive effect 50 times that of giant magnetostrictive materials.

The other development, perhaps more spectacular in the long run, is the use of density functional theory. Density functional theory, based on the observation of W. Kohn (the 1998 co-winner of the Nobel Prize in chemistry) and his colleagues that the density of electrons in quantum mechanics is subject to a variational principle, has as input only fundamental atomic information, in particular the locations of nuclei and their charges. Thus density functional theory can directly predict material properties from knowledge of the composition. Recently, unexpected new phases of even common materials have been predicted theoretically using density functional theory, and have subsequently been observed at the predicted temperature, pressure and composition. The key mathematical problems for density functional theory are understanding the quantum mechanical foundations, passing to simpler models of atomic forces, improving methods for including statistical mechanics to produce predictions at finite temperature. The future integration of density functional theory and materials science is likely to lead to major advances in "new materials from theory", and may one day surpass even the time-honored systematic method.

Multiscale phenomena. Quantum mechanics cannot deal effectively today with some of the most interesting and useful characteristics of materials---properties that are *structure-sensitive*, meaning that they are affected, often profoundly, by the microstructure of the material. Examples of structure-sensitive properties are strength, plasticity, magnetic hysteresis, dielectric constant, optical properties of liquid crystals, superconductivity, and almost any property associated with a phase transformation. The relevant microstructural features are, for example, a precipitate produced by a diffusional phase transition, a magnetic domain, a vortex, a point or line defect, or a dislocation tangle.

Unfortunately, the smallest microstructural features of general interest in materials science are much too small for the application of density functional theory. Furthermore, these microstructural features frequently act collectively in a way that cannot be discovered by analyzing only one of them. The gap in the time scales is even wider:

kinetic processes have time scales that range between milliseconds, seconds, and days, yet the analysis of even a microsecond event is far beyond the capability of first-principles computations with only a few atoms.

Despite these difficulties, there is hope because of the recent appearance of mathematical methods suited to the passage from one scale to another. When properties exhibit large spatial or temporal fluctuations on one scale governed by a certain set of partial differential equations, it is now becoming understood how to derive equations appropriate to a larger scale, using weak convergence methods, homogenization, Young measures, and various notions of convergence of operators. While these methods have mainly been applied to derive one continuum theory from another, they could well serve more generally for problems of change of scale, such as the direct passage from density functional theory to continuum theory. The dream of researchers in this area is to have the coefficients of macroscale differential equations evaluated directly by atomic-scale computations with an input of only fundamental constants.

The other opportunity for multiscale methods comes because it is becoming possible to reproducibly synthesize structures with an atomic-scale dimension. The subsequent investigation of the unexpected properties and possible applications of these nanoscale structures has given rise to the huge, dynamic field of nanotechnology. Properties that are known to be structure-sensitive on the macroscale are susceptible to unusual behavior at the microscale or nanoscale. Qualitatively, something strange is expected when the size of the structure is decreased below the typical size of the feature that gives rise to the structural sensitivity. But, quantitatively, there is a conspicuous absence of mathematical theory that can be used to predict the behavior of such structures; when this theory becomes available, important breakthroughs are likely.

3.8 Mixing in the Oceans and Atmospheres

At first blush it would appear that mixing in the atmosphere or ocean is straightforward and of little mathematical interest. After all, children who make chocolate milk from a powder quickly learn that the longer and more energetically they stir, the more evenly the chocolate powder is spread and dissolved in the milk. While that common-sense lesson is valid, the oceans and atmosphere are, in some sense, less vigorously stirred, so that the mixing is incomplete.

A careful look at mixing in oceans, atmospheres, and laboratory experiments reveals “islands” of unmixed fluid that nothing from the outside seems capable of penetrating; thus there are clearly demarked contours that act as barriers to mixing. While this phenomenon results in pretty pictures of laboratory experiments, the consequences can be a matter of life or death for fish whose survival depends upon the correct mixing of nutrients, chemicals, plankton, other fish, and even their own larvae or juveniles. Similarly, the spread of pollution and greenhouse gases in the atmosphere depends on the vagaries of natural mixing. When mixing changes in oceans or atmospheres, there is an immediate and large impact. For example, the changed mixing of nutrients for anchovies in Monterey Bay led to the disappearance not only of the anchovies, but also of the active warehouses and factories of Cannery Row. Our ability to predict the effects of pollution, global and long-term changes in climate, and the health of our oceans depends on our ability to understand and model the vagaries of mixing.

Using sophisticated ideas with origins in classical mechanics, nonlinear dynamics, and chaos theory, mathematicians have been able to show that mixing is far more complex than a fast diffusion process (i.e., how ink spreads in non-moving water). Mixing occurs at unequal rates depending upon direction and locations. It is profoundly affected by the state of the fluid and by the locations of eddies and currents. The mathematics of mixing shows that, while large eddies or vortices make the mixing and transport of chemicals and pollutants very efficient around and along their outside edges, the edges themselves act as barriers to the mixing of chemicals into or out of the vortices.

An interesting example is the "ozone hole" over Antarctica, a region of the atmosphere where ozone is nearly completely destroyed due to a chemical reaction in the upper atmosphere's clouds. Since the hole is surrounded by ozone and the atmosphere is highly stirred from atmospheric turbulence, it is natural to ask why the surrounding ozone does not mix into the hole. The answer is that the hole is at the center of a large vortex (the Antarctic stratospheric polar vortex), and mathematical models correctly predict that its outer edge acts as a strong barrier to mixing even though the atmosphere is turbulent and the edge of the vortex is constantly changing position in response to the turbulence. The vortex is crucial to maintenance of the hole. Each spring the stratospheric vortex breaks up due to warming of the ground below; this destroys not only the vortex, but also its edge---the barrier to mixing. Thus the ozone is replenished in the hole and the hole goes away.

The effects of barriers to mixing can be appreciated on even a larger scale. It has long been recognized that the equator hampers mixing in the atmosphere between the northern and southern hemispheres. Mathematical analysis is beginning to explain the selective permeability of this barrier, which is a complex and interesting function of location (with respect to continents and other topographic features), time of year, and time with respect to longer-term cycles (such as that associated with *El Niño)* that occur in the ocean-atmosphere system. An important question being addressed is how man-made greenhouse gases such as carbon dioxide, which are predominantly created in the north, spread south of the equator.

The incomplete mixing caused by stirring has consequences beyond the distribution of pollutants. Not only are tangible things such as chemicals mixed by stirring, but so is "vorticity" (the amount of spin or rotation of a small parcel of fluid). If the fluid is not quiet or rotating as a solid body, the vorticity changes as a function of position in the fluid, and is likely to change with time as well. What makes the mixing of vorticity so fascinating is that it is a highly nonlinear process in the sense of having substantial feedback onto itself, since the locations of vortices, especially of their edges, determine where mixing of vorticity occurs. In particular, if a double row of oppositely-signed vorticity accumulates in sheets, it produces an ocean current or jet stream.

Observations show that the complex mixing due to stirring often divides oceans and atmospheres into separate regions ("patches") such that the fluid within each region is well mixed, but there is very little mixing among the regions. Mathematical theories based on statistical mechanics have recently taken away the mystery from the most visually striking examples of these patches: Jupiter's Great Red Spot and the horizontal, multi-colored stripes of Saturn and Jupiter.

The Red Spot, seen through some of the first telescopes 360 years ago, is a very robust vortex that owes its existence to the continuous accumulation and mixing together of small patches of vorticity. The stripes of Jupiter and Saturn are alternating jet streams. The east-west bands of Saturn appear multi-colored because the chemicals within each band are well mixed but there is little mixing of chemicals from band to band.

Numerical simulations of mixing on supercomputers have contributed several insights. If the Red Spot were artificially broken apart, the pieces would simply mix back together. If Jupiter or Saturn initially had no motion in its atmosphere with respect to an observer on the planet, then mixing of vorticity would start spontaneously and create, after several decades, the jet streams we see today. An atmosphere initially at rest with respect to an observer on the planet contains a continuous distribution of vorticity, maximized at the north pole, decreasing to zero at the equator, and reaching its minimum at the South pole. Since different fluid elements thus have different values of vorticity, the distribution can be mixed so that there are alternating sheets of vorticity with opposite sign (and consequently jet streams).

Saturn and Jupiter both display rings; Saturn's are easily seen with a pair of binoculars, but Jupiter's ring is so faint that it was not detected until the 1970s. These rings are made of small particles, are extraordinarily thin, and have very sharp, well-defined edges. The rings of Saturn consist of a nest of several rings with well-defined gaps in between. Several properties of the rings are explained by the mathematics of mixing and of nonlinear dynamical systems. As with the ozone hole over Antarctica, it might appear that the gaps would quickly fill with particles that are continuously bumped into them via interactions with other particles. However, gravitation from the moons controls the mixing in such a way that these narrow gaps are kept free of particles and the edges of the rings are well-defined.

The mathematical theory of mixing in nonlinear systems allows us to understand and predict much of what we see in nature. Although *controlling* nature with these theories is well beyond present capabilities, the same mathematics is being used in practical engineering problems. In micro-electrical and mechanical systems, there is an ever-increasing desire to miniaturize both electronic and mechanical components. Millions of motors can now be created on the head of a pin and used as pumps to deliver medicines, carry out chemical reactions, act as precise carburetors for engines, and so on. For these applications it is necessary to mix one or more chemicals together. Early on, researchers believed that everything in these devices would mix together easily because of the very small distances that the constituents had to travel, but this belief failed to take into account the many barriers to mixing. Fortunately, the mathematics reveals how to destroy these barriers. One such method, known as chaotic mixing, requires stirring the fluid at two incommensurate frequencies. This can be done in one of these devices by creating two small heating elements via the same type of lithographic techniques used to build electronic chips. When the elements are supplied with voltage, they boil tiny amounts of fluid, producing two bubbles. Oscillating the voltage makes the two bubbles oscillate, which in turn provides stirring at any desired frequency. Thus an application of mathematical mixing that was originally inspired by the study of natural phenomena solves a critical problem in state-of-the-art engineering.

3.9 Physiology

With a few notable exceptions such as Helmholtz, Frank, Hodgkin, and Huxley, physiology and the mathematical sciences have not been closely linked until recently. Many, perhaps most, physiologists have regarded their science as primarily descriptive, with little scope or need for mathematics; mathematicians trained in the traditional way almost invariably “speak physics”, and may be reluctant to enter a field in which competence demands a significant degree of specialized scientific study of an unfamiliar kind.

But this situation is changing from both directions. Mathematical models and computational simulation offer means for characterizing and analyzing processes in physiology that are individually complex to begin with and whose interactions add further complexity; in return, physiology provides a rich, fascinating field of science with opportunities for new applications and new mathematics. A prize-winning book on mathematical physiology[2] stresses the importance of increased connections between mathematics and physiology: “... teaching physiology without a mathematical description of the underlying dynamical processes is like teaching planetary motion to physicists without mentioning or using Kepler's laws; you can observe that there is a full moon every 28 days, but without mathematics you cannot determine when the next total lunar or solar eclipse will be nor when Halley's comet will return".

One area among many in physiology where the mathematical sciences are beginning to make major contributions is *integrative biology*, the system-level study of how complex, spatially distributed biological systems manage to perform their functions. Mathematical models are being developed that analyze the following aspects of complex physiological systems, to mention just a few:

- the macroscopic behavior of lung tissue based on the microstructure of respiratory regions;
- the self-organization of cells and molecules in the immune system that underlies responses to attacking pathogens; and
- the control of cells in a developing system so that they “know” where they should go and what to do at their destination.

A pervasive example of integrative behavior is *movement*: living organisms internally move things such as nutrients, blood, oxygen, and pigment. Somehow, based on principles that remain unknown, living creatures self-organize a movement that achieves a prescribed result. Two instances in cell biology of self-organizing behavior related to movement involve *centering*: a nucleus spends most of its time at the center of its cell, yet it cannot directly sense the cell membrane nor evaluate its distances from points on the membrane; and cell division requires chromosomes to be aligned along a central plane during cell division. Current biological models rely on the unrealistic assumption of non-local dynamics, so an obvious question is whether global behavior such as centering can be achieved entirely through local interactions.

[2] J. Keener and J. Sneyd, *Mathematical Physiology*, Springer-Verlag, Berlin, 1998

To answer this question, mathematical models constructed from local processes can be studied to see whether (and, if so, why) they result in centering behavior. An illustration of the role of mathematics is provided by recent work on modeling properties of cells in the black tetra, a small colorful fish popular in home aquariums. Melanophore cells create the tetra's colors and patterns through self-organizing behavior that depends on the interactions of microtubules, dynein, and pigment. *Microtubules* are long tubelike protein polymers with a polar structure (a difference between the two ends). *Dynein* is a molecular motor that transforms stored energy into mechanical work. When activated by adrenaline, dynein moves along microtubules, always in the same direction--- toward the “minus end”, away from the “plus end”. Dynein has an affinity for pigment molecules, and will drag them along as it moves.

In melanophore cells, microtubules are normally arranged in a radial pattern with minus ends near the nucleus and plus ends near the membrane. If the dynein motors are activated, pigment tends to aggregate in a small region around the cell nucleus. The macroscopic effect of pigment aggregation is to change the intensity of the cell's (and, in the large, the tetra's) color.

Recent experiments were designed to understand the dynamics of how pigment centers around the nucleus. A fragment of a melanophore cell was sliced off, separating it from the nucleus. Following the cut, the dynein was activated in the fragment. Very soon, a pigment aggregate formed near the “minus” edge of the fragment. A slower process then occurred in which the pigment aggregate drifted toward and eventually stopped at the “center” of the fragment. In this final state, the microtubules had actually rearranged themselves in a radial pattern within the fragment. Numerous other experiments demonstrated that the radial array of microtubules did not form unless the dynein was activated and pigment was present.

A mathematical model of this process begins with several facts and assumptions. In the absence of pigment, microtubules grow (at the plus end) and shrink (at the minus end) at the same rate. Dynein, even when carrying pigment, moves much faster than the growth/shrinkage rate of microtubules without pigment. Plus ends of microtubules are stabilized when they reach the cell boundary; minus ends tend to be “caught” in regions of high pigment concentration. Nucleation (the appearance of new microtubules) occurs on pigment particles. Together, these assumptions descriptively explain the fast initial movement of pigment to the cell boundary followed by the slow centering of the pigment aggregate.

The challenge for mathematical modeling is to translate these assumptions into a form that captures, both qualitatively and quantitatively, the observed relationships among microtubules, dynein, and pigment. Work in this direction has begun with a one-dimensional version of the problem. Although overly simplistic, it is nonetheless appropriate for a long thin fragment in which almost all microtubules run down the long axis and there is little variation in pigment along the thin axis. The main parameters are the fragment length, the speed of dynein, the plus end growth rate, and the diffusion coefficient for pigment. The pigment concentration is treated as a function of position and time; microtubules are described by their plus and minus ends and orientation (left- or right-moving).

To define the cell dynamics, the shrinkage and nucleation rates are described as functions of pigment concentration. Growth at the plus end of a microtubule is interpreted as moving that end at a particular velocity, and nucleation is treated as a reaction term. Pigment flux is determined by diffusion and motion along microtubules. Using conservation principles, analysis of the flux due to microtubules, and pigment dynamics, a system of partial differential equations and boundary conditions has been defined. A crucial feature of the model is that all relationships are local, as they are in the theory being represented.

Even with this relatively simple formulation, centering of the pigment aggregate within the fragment occurs consistently, and simulations have satisfactorily matched appropriate experimental observations. The ability to vary the mathematical parameters and initial conditions in the model allows "virtual experiments" in which essentially every conceivable combination of pigment distribution and microtubule orientation can be tried. The next step is, of course, to refine and extend the model to convey the full set of known properties of the melanophore cells, with the ultimate goal of understanding centering.

An implicit but crucial general point is that mathematics and physiology must be intimately connected to succeed in this kind of endeavor. The example of melanophore in the black tetra clearly illustrates that serious knowledge of physiology is required to create even an elementary mathematical model of pigment centering.

3.10 Diagnosis Using Variational Probabilistic Inference

The rapid growth of the information sciences is leading to new challenges for mathematics. Although in many cases entirely new mathematical theories must be formulated, the reservoir of mathematical knowledge is vast and what is called for is sometimes the discovery of appropriate analogies so that old ideas can be applied in new ways. In a recent success story, a difficult problem in probabilistic diagnosis has been solved via the use of techniques originally developed for statistical physics, quantum mechanics, and mechanical engineering.

The problem of diagnosis is an instance of the general problem of "inductive inference" or, more informally, "reasoning backward". Consider, for example, the problem of diagnostic reasoning in medicine. A doctor observes a set of symptoms in a patient and wishes to infer the disease (or diseases) that could be responsible. In general the doctor must utilize basic medical knowledge inductively to uncover an explanation of a pattern of symptoms. Basic medical knowledge consists of biologically-based, causal theories specifying the way in which various diseases affect the organism and lead to various symptoms. From this knowledge, in the form of disease-to-symptom relationships, the doctor must reason backwards to make predictions about symptom-to-disease relationships.

Backward reasoning can be complex computationally. A major source of complexity is that causally unrelated diseases (i.e., with unrelated biological origins) can become strongly dependent diagnostically. Suppose that two unrelated diseases have a predicted symptom in common and that the symptom is in fact observed; then the two diseases compete to explain it, i.e., additional evidence that one of the diseases is present tends to reduce our belief in the presence of the other disease. In general, a disease can "explain away" a symptom, decreasing the need to posit some other disease as the explanation of the symptom. This changed belief in a disease can then "flow forward", lowering or raising the support for other symptoms, which---by the same explaining-away mechanism---can affect the belief in yet other diseases. The fact that different diseases have common sets of symptoms can lead to a tangled web of interdependencies.

Scientists have started to build probabilistic tools for diagnosis not only in medicine but in many other domains, including manufacturing, transportation, and communications. Building these tools has improved understanding of the mathematical issues underlying backward reasoning. Significant progress has been made in an area known as *graphical modeling*, where, in the past ten years, a general mathematical theory has emerged that yields a clear specification of the complexity of diagnosis in probabilistic systems and allows optimal algorithms to be defined. Many classical probabilistic tools, including the Kalman filter (used in control and estimation theory) and the hidden Markov model (used in speech recognition and molecular biology), are special cases of this general methodology. But the theory applies much more widely, providing a general understanding of probabilistic inference in arbitrary probabilistic networks.

A particularly challenging instance of a complex probabilistic knowledge base is the "Quick Medical Reference" (QMR) database for diagnosis in internal medicine. This database, developed at the University of Pittsburgh with 25 person-years of effort, is one of the largest probabilistic databases in existence, and contains a significant fraction of the diseases in internal medicine. The QMR database is organized as a probabilistic network in which approximately 600 binary-valued nodes representing diseases are linked to approximately 4000 binary-valued nodes representing symptoms.

Unfortunately, when one analyzes the QMR network from the viewpoint of the recently developed theory of inference, one finds that *exact* diagnostic reasoning is infeasible computationally. For a set of typical symptoms, it has been estimated that calculation of the exact probabilities of diseases would require approximately 50 years on current computers. Research on QMR and related large-scale diagnostic systems has consequently lain fallow for want of efficient algorithms.

The general mathematical problem underlying probabilistic inference, hinted at in the earlier discussion of explaining away, takes the form of a set of nonlinear equations in which each equation can have an exponentially large number of terms. Roughly speaking, to determine the probability of a disease in the QMR network, given a set of symptoms, one must take the product over the probabilities of the observed symptoms (a nonlinear operation) and then take the sum over all configurations of other diseases (a sum involving 2^{599} terms). The actual computation is not as bad as this, given that the network is not fully connected (e.g., some diseases have zero probability of producing certain symptoms), but it is still intractable.

There is a way out of this computational dilemma: viewing the problem as numerical, with a need to find accurate statistical estimates, rather than as symbolic, with a need to compute a large number of terms. The fact that there are so many terms in the sums to be calculated offers hope that laws of large numbers will come into play, rendering the system probabilistically simple despite its apparent symbolic complexity.

This point of view is of course natural in the context of the highly interacting systems in statistical physics, and one might hope that the tools developed in physics could be employed in the service of large-scale diagnostic inference problems. In fact, a number of useful analogies can be drawn between graphical models and statistical physics models. The major technical difficulty arises because the graphical models studied in diagnostic reasoning are generally based on directed graphs (graphs in which the nodes are linked by arrows), whereas in statistical physics the graphs tend to be undirected (a consequence of Newton's third law). Once this technical hurdle is overcome, many ideas from the physics context can be exploited in diagnosis. In particular, the mean field approach in statistical physics has a natural analogue for graphical models.

More broadly, mean field theory can be viewed as a variational method in which a nonlinear system with strong couplings is approximated by a variational principle. Variational principles are highly successful in mechanics, where variational finite element methods characterize the global state of stress or strain of a piece of material. These methods can also provide useful insight into approximation methods for diagnostic reasoning.

Researchers have recently developed an approximate approach to probabilistic inference known as *variational inference*, which is very similar to mean field theory and finite element analysis. Rather than performing inference directly on a dense probabilistic network, the variational approach considers a simplified network in which some of the links are missing. Roughly speaking, a variational parameter is introduced for each missing link; this parameter captures in an approximate way the high-order probabilistic dependencies induced when that link is present in the network. The simplified network is chosen so as to obtain bounds on the probabilities of interest rather than exact values.

The advent of variational methods in probabilistic inference has created new mathematical problems. Some of these are analogous to problems in statistical physics and finite element analysis, and solutions in these domains may prove useful in variational inference. For example, variational methods can fail when there are deterministic relationships between nodes in a network. This is conceptually similar to the difficulty posed by incompressible media in finite element analysis, where solution methods are available and may be broadly useful.

The variational approach has been highly effective for the QMR database, where it can yield accurate estimates of disease probabilities within less than a second of computer time. It has also been applied successfully to a number of other graphical models in which exact inference is intractable. Applications to diagnosis, pattern recognition, statistical genetics, and error-correcting codes are currently being explored. A particularly interesting application is to learning theory, where one would like to find out the parameters of a graph based on data; there are many interesting relationships between inference and learning that variational methods may help us to understand.

3.11 Iterative Control of Nuclear Spin

Nuclear spins play a central role in nuclear magnetic resonance (NMR), spectroscopy, and magnetic resonance imaging (MRI). Control of nuclear spins is tantamount to control of the parameters that determine the features and information content of NMR spectra and MRI images. Recent research has led to development and implementation of an approach for *iterative control* of nuclear spins.

In general terms, control takes a system from a given initial state to a desired final state under the action of a control propagator. The system may be a robot, a vehicle or spacecraft, a molecule, or a system of nuclear spins. Traditional differential control involves the feedback adjustment of the parameters of an evolving system in order to prevent deviations from a prescribed trajectory. Such control necessitates comparison of the actual evolving trajectory with the prescribed trajectory, i.e. it is necessary to “see where you're going”.

In the novel iterative schemes, by contrast, the propagator that induces the desired trajectory is chosen as the stable fixed point in “propagator space” of the iterative map that is applied between stages of the system. This choice ensures that any initial propagator, regardless of errors or perturbations, will always converge to the desired final state.

With this approach, it is not necessary to “see where you're going”. Thus, instead of tailored differential control for each member of an ensemble that may experience different errors, the same control sequence can be applied “blindly” to the whole ensemble. There is, of course, a price to pay for this broadband privilege---the trajectory from initial to final state may be considerably longer and more complex. However, convergence to the desired final state with predetermined precision is assured. Clearly there are circumstances in which differential control is more appropriate, and there are others where iterative control is superior.

Systems containing nuclear spins are often well suited to iterative control because they involve large ensembles with broad ranges of control parameters and errors. The new stable, indeed “super stable”, fixed points for such systems have been obtained through dynamical systems theory.

Iterative sequences derived from these mathematical models have been implemented in NMR and MRI through collaborations between mathematicians and scientists. With the resulting enhanced instruments, precise and selective control of the states of nuclear spins can be achieved. On the microscopic scale, for example, iterative decoupling sequences permit elimination of the effects of spin-spin interactions. As a result, the NMR spectra are enormously simplified, allowing the structures of molecules in solution and in materials to be determined. On the macroscopic scale, iterative excitation in MRI makes it possible to elicit and to selectively enhance or suppress signals from particular regions of the images of organisms, consequently providing spatially selective biomedical information.

In recent years, NMR has emerged---beyond its role as a diagnostic analytical tool for molecules, materials, and organisms---as a potentially powerful environment for implementation of quantum computing. The nuclear spins are, after all, quantum systems with a natural binary basis, namely the two quantum states “up” and “down” in a magnetic field. The spins can therefore function as “qubits” whose entangled quantum states are manipulated in quantum logic gates by means of delicately controlled radiofrequency pulses, as in NMR spectroscopy.

Enormous potential advantage of quantum computing over classical computing is foreseen because quantum algorithms involve participation of all qubits at the same time. This is a uniquely quantum phenomenon akin to capitalizing on the simultaneous existence of the alive and dead quantum “Schrödinger cat”. Iterative control schemes currently under development should make it possible to overcome the effects of decoherence, thus allowing the implementation of extended quantum computation algorithms even in the presence of imperfect quantum logic gates and interactions with the environment.

3.12 Moving Boundaries and Interfaces

Many physical problems involve moving boundaries. *Dynamic* boundaries change position and shape in response to the particular physics at work: examples are breaking waves in the ocean, dancing flames in the fireplace, and milk swirling in a cup of tea. *Static* boundaries, such as tumors in medical scans and cartoon characters against a background animation, can be just as perplexing: try finding edges in a picture of a dalmatian lying on a rug with spots! Surprisingly, many other interesting problems, such as negotiating a robot around obstacles and finding the shortest path over a mountain range, can also be cast as evolving boundary problems.

The physics and chemistry that drive a boundary or interface may be difficult to describe, but even when the speed and direction of a moving interface are well understood, following its shape can be difficult. The first concern is what to do when sharp corners appear, as they do in, for example, the intricate patterns of a snowflake. Second, distant edges can blend together: the “edge” of a forest fire changes as separate fires burn together and sparks carried by the wind ignite distant regions. Finally, in three dimensions (and higher), even finding a nice way to represent---let alone move---an undulating boundary is a challenge.

One technologically important example of interface motion involves the manufacture of computer chips. In the etching and deposition process, a layer of metal is deposited on a silicon wafer, etched away, and then the process is repeated numerous times until a final profile is obtained. As device sizes get smaller and smaller, using trial and error to obtain the correct design becomes impractical. Instead, one would like to simulate these processes as accurately as possible in order to test various layering strategies and resulting device characteristics. In recent years, the application of new mathematical and numerical algorithms for interface motion has afforded real breakthroughs in this area. Before these techniques, complex problems involving the evolution of profiles in two dimensions were difficult; now, fully three-dimensional simulations involving a wide range of physical effects are easily within grasp. The new algorithms have been incorporated into the simulation packages at many major semiconductor manufacturers in the United States, and are part of the production environment in various chip lines today.

These computational techniques, known as level set methods and fast marching methods, rest on a fundamental shift in how evolving fronts are viewed. Rather than focus on the evolving front itself, these techniques discretize the region in which the front moves. Each point in that space keeps track of either its distance to the front or of the time when the front passes over it; the accumulation of all this information gives an accurate portrait of the moving interface. The key is to define equations for the time at which the front passes over each point and then to solve these equations.

The equations which keep track of the front at each grid point in the domain are variants of the Hamilton-Jacobi equations; these equations have a long history in such areas as optics, wave propagation, and control theory. While they can be very complex, their derivatives bear a resemblance to hyperbolic conservation laws and to the equations of fluid mechanics, allowing use of the knowledge acquired in those well-developed fields. The main breakthrough in modeling interface motion was the realization that schemes from fluid mechanics could be unleashed onto the equations of moving fronts. The result is a wide range of computational tools for tracking evolving interfaces with sharp corners and cusps, with topological changes, and in the presence of three-dimensional complications. These schemes have found their way into a vast number of applications, including fluid mechanics, dendrite solidification and the freezing of materials, image processing, medical imaging, combustion, and robotic navigation.

Some of the most complex interface applications appear in simulating the manufacture of computer chips. To begin, a single crystal ingot of silicon is extracted from molten pure silicon. This silicon ingot is then sliced into several hundred thin wafers, each of which is polished to a smooth finish. A thin layer of crystalline silicon is oxidized, a light-sensitive “photoresist” is applied, and the wafer is covered with a pattern mask that shields part of the photoresist. This pattern mask contains the layout of the circuit itself. Under exposure to a light or an electron beam, the unshielded photoresist polymerizes and hardens, leaving an unexposed material that is etched away in a dry etch process, revealing a bare silicon dioxide layer. Ionized impurity atoms such as boron, phosphorus, and argon are implanted into the pattern of the exposed silicon wafer, and silicon dioxide is deposited at reduced pressure in a plasma discharge from gas mixtures at a low temperature. Finally, thin films like aluminum are deposited by processes such as plasma sputtering, and contacts to the electrical components and component interconnections are established. The result is a device that carries the desired electrical properties.

This sequence of events produces considerable changes in the surface profile as it undergoes various processes of etching and deposition. Describing these changes is known as the “surface topography problem” in microfabrication and requires an analysis of the effects of many factors, such as the visibility of the etching/deposition source from each point of the evolving profile, surface diffusion along the front, complex flux laws that produce faceting, shocks and rarefactions, material-dependent discontinuous etch rates, and masking profiles. The physics and chemistry that contribute to the motion of the interface are areas of active research. Once empirical models are formulated, one is left with the problem of tracking the evolving front.

Here is where level set methods and fast marching methods come into play: they provide the means to follow the evolving profile as it is shaped by the etching and deposition process, and they capture some of the most subtle effects. For example, visibility has a key role; if part of the evolving surface causes a shadow zone that blocks the effects of the etching or deposition beam, the motion is reduced. Computing this shadow zone was formerly a very expensive proposition; however, the fast marching method yields an elegant and fast way to do it.

Another example is the complex manufacturing process called ion-milling, in which a beam of reactive ions acts like a sandblaster and etches away at a surface. The etching rate depends on, among other things, the angle at which the beam hits the surface. The most effective etching angle is not always directly straight down; the “yield function” relates how much material is removed to the incoming angle. Interestingly enough, this process produces beveled, rounded edges in some areas and sharp cusps in others. While these are difficult problems to model, they are easily handled by level set and fast marching methods.

4 Education

The importance of strong ties between mathematics and science is self-evident from the examples presented---which, we stress again, are only a tiny sample from a very large pool. Unfortunately, there is a clear shortage of people able to bridge the gap between mathematics and the sciences, and one of the challenges that must be faced is how to educate more.

It is obvious to us that students of mathematics should be able to understand problems in science, and that students of science should understand the power and roles of mathematics. Each area of science has its own unique features, but the different areas share common features that are often of a mathematical nature.

The themes of *modeling*, *computation*, and *problem solving* are especially relevant to education.

- **Modeling.** Students in science and mathematics need to be educated in modeling far beyond the simple paradigm exemplified by ``do this experiment, plot the data, and observe that they lie almost on a straight line''. Given a physical problem and/or data, students should learn to construct a mathematical model, explain why the model is appropriate, perform mathematical analysis or a computational simulation, devise experiments to check the accuracy of their model, and then improve the model and repeat the process.
- **Computation.** The view that ``anyone can compute'' is just as wrong as the statement that ``anyone can build a telescope''. One has to learn how. Much of the current teaching of computation is flawed; a ``cookbook'' strategy of using canned programs without attention to fundamentals is completely inadequate. At the other extreme, scientists should not waste their time implementing outmoded methods or reinventing known algorithms and data structures. Students in science and mathematics need to be aware of the intellectual content and principles of modern computer science.
- **Problem-solving.** In traditional academic presentations of scientific and mathematical problems, the context is stripped away and simplified so that students can focus on the essentials. But, especially when developing mathematical insights, students must learn how to approach ill-defined, poorly formulated problems---an area in which education is lacking. There are no shortcuts; the only way to learn is by direct experience.

We offer a number of recommendations for education in mathematics and science. Our primary focus is education for students who specialize in mathematics or science; we cannot begin to address the national problem of mathematics and science education for all.

1. Support curriculum development in areas that are essential for connections between mathematics and science. Every curriculum-related activity should include production of Web-based materials.

(a) Create modeling courses for high school, undergraduate, and graduate students. Unlike many other skills, modeling can be taught (at an elementary level) to students in high school. At the undergraduate level, there would be enormous benefits if a one-year modeling course were part of the core curriculum in science, engineering, mathematics, and computer science. Graduate modeling courses would deepen the scientific knowledge of mathematics students while enriching the mathematical skills of science students.

(b) Support development of courses that tie core computer science to science, engineering, and mathematics. Programming, numerical analysis, data structures, and algorithms---each of which is a topic with serious mathematical content---should be part of the education of every scientist and mathematician.

(c) Encourage experiments in activities (courses, summer or short-term workshops) that teach scientific and mathematical problem solving. Such programs could involve not only techniques and direct experience of problem solving, but also group projects that teach students how to work collaboratively with others and how to present their work.

2. Encourage students to undertake programs of study, at both undergraduate and graduate levels, which combine mathematics and science. That this can be done at the graduate level has been shown by the successful Computational Science Graduate Fellowship program of the Department of Energy, which requires students to undertake a demanding interdisciplinary program in exchange for a generous fellowship.

3. Support summer institutes in (i) mathematical topics that address scientific applications and (ii) scientific topics with mathematical content.

The NSF Research Experiences for Undergraduates (REU) program has been extremely successful in exposing students to research at an early stage. REU and other institutes have become important for top undergraduates interested in science and mathematics, and it is now common to prepare for graduate school by attending a summer school or institute. However, these programs are overwhelmingly devoted to highly specialized subjects. In part this is understandable; the organizers want to give the students a taste of research, which is more easily done in a narrow area. But because those summer institutes often determine the direction students will take, NSF should ensure that there are high-quality institute programs with a multidisciplinary emphasis centered on connections between mathematics and science.

Certain emerging areas (such as mathematical biology) are not yet widely covered in graduate programs. Carefully designed summer institutes would help to broaden the education of graduate students whose home institutions lack offerings in such fields.

4. Fund research groups that include both (i) a genuine collaboration between scientists and mathematicians, and (ii) a strong educational program for graduate students, postdoctoral fellows, and possibly undergraduates. To be effective, such funding should be as long-term as possible; if funding is only short-term, researchers are unlikely to

make the huge investment of time needed to develop group structures that will sustain multidisciplinary collaborations.

5. Fund postdoctoral fellowships in environments that combine excellence in science with excellence in mathematics. Efforts to create industrial postdoc programs could be expanded to create joint university/national lab postdoctoral fellowships, as well as short-term fellowships for scientists in mathematics programs with a strong applied component.

Beyond the postdoctoral level, there should be programs to encourage and support faculty who would like to become active in collaborations outside their own discipline. The existing NSF program in this vein, Interdisciplinary Grants in the Mathematical Sciences (IGMS), is small and imposes relatively strict requirements on qualification and support by the home department.

6. Develop a program of group grants for mathematics and science departments that encourage the creation of new courses, experimentation with instructional formats, and coordinated programs of hands-on experiments, modeling, and computation. Departments that receive such grants should have substantial science requirements for undergraduate degrees in mathematics, and substantial mathematics requirements for undergraduate degrees in science. Many, if not most, U.S. undergraduates in mathematics take no, or almost no, science courses. In certain areas of science and engineering, undergraduates take only minimal, and sometimes outdated, mathematics courses; even worse, those courses may give students no understanding of the ties between their fields and mathematics. These unfortunate situations are likely to be corrected only if there is an incentive for departments to change their basic programs.

5 Conclusions

Strong ties between mathematics and the sciences exist and are thriving, but there need to be many more. To enhance scientific progress, such connections should become pervasive, and it is sound scientific policy to foster them actively.

It is especially important to make connections between mathematics and the sciences more timely. Scientists and engineers should have access to the most recent mathematical tools, while mathematicians should be privy to the latest thinking in the sciences. In an earlier era of small science, Einstein could use the geometry of Levi-Civita within a few years of its invention. With today's vastly expanded scientific enterprise and increased specialization, new discoveries in mathematics may remain unknown to scientists and engineers for extended periods of time; already the analytical and numerical methods used in several scientific fields lag well behind current knowledge. Similarly, collaborations with scientists are essential to make mathematicians aware of important problems and opportunities.

6 References and URLs

Combustion

[1] Information about Chemkin™, a registered trademark of Sandia National Laboratories:

http://stokes.lance.colostate.edu/CHEMKIN_Collection.html
http://www.sandia.gov/1100/CVDwww/chemkin.htm
http://www.sandia.gov/1100/CVDwww/theory.htm

Cosmology

[2] M. S. Turner and J. A. Tyson (1999), *Cosmology at the Millennium*, working paper.

[3] Web sites about mathematical models and numerical simulation:

http://star-www.dur.ac.uk/~frazerp/virgo/aims.html
http://phobos.astro.uwo.ca/~thacker/cosmology/

Finance

[4] I. Karatzas and S. E. Shreve (1998), *Methods of Mathematical Finance*, Springer-Verlag, New York.

[5] T. F. Coleman (1999), An inverse problem in finance, *Newsletter of the SIAM Activity Group on Optimization*.

Functional Magnetic Resonance Imaging

[6] W. F. Eddy (1997), Functional magnetic resonance imaging is a team sport, *Statistical Computing and Statistical Graphics Newsletter*, Volume 8, American Statistical Association.

[7] Information about functional image analysis software:

http://www.stat.cmu.edu/~fiasco

Hybrid System Theory and Air Traffic Management

[8] C. Tomlin, G. J. Pappas, and S. Sastry (1998), Conflict resolution for air traffic management: a case study in multi-agent hybrid systems, *IEEE Transactions on Automatic Control*, 43, 509---521.

Internet Analysis, Reliability, and Security

[9] Willinger and V.\ Paxson (1998), Where mathematics meets the Internet, *Notices of the American Mathematical Society* 45, 961---970.

[10] The Web site of the Network Research Group, Lawrence Berkeley Laboratory:

http://www-nrg.ee.lbl.gov

Materials Science

[11] Research trends in solid mechanics (G. J. Dvorak, ed), United States National Committee on Theoretical and Applied Mechanics, to appear in *International Journal of Solids and Structures*, 1999.

[12] G. Friesecke and R. D. James (1999), A scheme for the passage from atomic to continuum theory for thin films, nanotubes and nanorods, preprint.

Mixing in the Oceans and Atmospheres

[13] P. S. Marcus (1993), Jupiter's great red spot and other vortices, *The Annual Review of Astronomy and Astrophysics* 31, 523---573.

Physiology

[14] J. Keener and J. Sneyd (1998), *Mathematical Physiology*, Springer-Verlag , Berlin.

[15] Details about modeling melanophore in the black tetra (the home page of Eric Cyntrynbaum, the University of Utah):

http://www.math.utah.edu/~eric/research

Diagnosis Using Variational Probabilistic Inference

[16] T. S. Jaakkola, T. S. and M. I. Jordan (1999). Variational methods and the QMR-DT database, submitted to *Journal of Artificial Intelligence Research*.

[17] M. I. Jordan (1998), *Learning in Graphical Models*, MIT Press, Cambridge, Massachusetts.

Iterative Control of Nuclear Spins

[18] R. Tycko, J. Guckenheimer, and A. Pines (1985), Fixed point theory of iterative excitation schemes in NMR, *J. Chem. Phys.* 83, 2775---2802.

[19] A. Lior, Z. Olejniczak, and A. Pines (1995), Coherent isotropic averaging in zero-field NMR, *J. Chem. Phys.* 103, 3966---3997.

Moving Boundaries and Interfaces

[20] J. A. Sethian (1996), *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Sciences*, Cambridge University Press.

Acknowledgements

This document was assembled and written by Alexandre J. Chorin and Margaret H. Wright. They gratefully acknowledge help from:

Dr. Phillip Colella,
Professor Thomas F. Coleman,
Professor William F. Eddy,
Professor John Guckenheimer,
Professor Richard D. James,
Professor Michael Jordan,
Professor James Keener,
Professor Philip Marcus,
Dr. Andrew M. Odlyzko,
Professor Alexander Pines,
Professor Shankar Sastry,
Professor James Sethian,
Professor Steven E. Shreve,
Professor Claire Tomlin, and
Dr. J. A. Tyson.