Technology Overview: A Report on Data Mining

CSCS TR-95-02

May 29, 1995

Abstract:

In this report we present a technology assessment of a spectrum of mathematical methods and software techniques for data mining, i.e., for finding patterns and regularities in sets of data. After an introduction to the subject, the second part of the paper describes the essentials of a variety of methods for data mining through supervised and un-supervised learning. In the third part of the report we present results from the application of data mining techniques to different application areas such as chemical and pharmaceutical chemistry, retail and consumer banking, remote sensing, and the handling of engineering and maintenance data.

Karsten M. Decker
CSCS–ETH, Swiss Scientific Computing Center, CH-6928 Manno

Sergio Focardi
The Intertek Group, F-75015 Paris

Technical Report

Table of Contents

List of Figures

1 Introduction

This report is intended as a technology assessment of a spectrum of mathematical methods and software techniques used for finding patterns and regularities in sets of data. This process is commonly referred to as data mining. Historically, the development of the science of statistics has resulted in a number of methods for the analysis of data. Statistical methods have been widely applied to analyzing data with the objective of finding correlations and dependencies among sets of data.

So what's new ? Three factors have recently changed the panorama of data analysis. First, the availability of large amounts of computing power at a low cost. When most statistical methods now in use were developed, in the 1960s and 1970s, the computing power of mainframes was equivalent to that of today's PCs; today's more powerful machines have computing power several orders of magnitude greater than that of the then-available mainframes. Hand in hand with the falling cost of processing power is the falling cost of storage. As Peter Lazard, head of technology strategy and research at Abbey National, remarked, until recently the UK bank was very careful about accumulating data due to the cost of storage; now data is cheap and the information that can be mined from it very important. A second factor is the accumulation of data at an explosive rate. This accumulation of data is a consequence of the low cost of storage combined with the present high level of automation of experiments and of the introduction of electronic data-gathering devices, be they electronic point-of-sale or remote-sensing devices. The third factor is the introduction of a new set of logical methods developed by the artificial-intelligence community, statisticians, and by physicists working in the domain of non-linear dynamics.

In the past decades scientists working in a number of related disciplines, from psychology to artificial intelligence, have tried to develop general models of cognition, and in particular of learning. For instance, SOAR, developed by Allen Newell and John Laird is one such model. See [1] for a comprehensive discussion. This was a vast and difficult endeavor that produced a number of by-products of technological interest. In particular, the techniques of representation of knowledge and the notions and techniques of learning point to the possibility of important applications for the representation and analysis of data. Rumelhart and McLelland [2] discuss learning paradigms of parallel distributed processing as implemented, for instance, in neural networks. Patrick Winston's Artificial Intelligence [3] is a standard textbook for paradigms of knowledge representation. In addition to the traditional statistical analysis of data, we are now seeing the increasing use of new methods, essentially methods of learning and knowledge representation. There is hardly an industrial sector where technology innovators are not experimenting with or already using regularly one or more of these methodologies.

These new logic methods are computationally intensive. This helps explain that only now is there an explosion of interest in them. In the 1960s and 1970s, most AI techniques showed poor results basically because of lack of computing power that made it impossible to go beyond the stage of toy demonstrations. This clearly limited the industrial and even scientific interest in these technologies. Now, however, impressive demonstrations are making powerful proofs of concept. This in turn is attracting research funds and the best intellects, producing results that fuel additional interest.

This report puts the problem of data mining in the general framework of machine learning. The global perspective of this paper is the following: starting from a set of data, automatic methodologies allow to develop an optimal representation of the structure of the given data.

The knowledge thus acquired can be extended to larger sets of data under the assumption that these larger sets have a structure similar to the sample data. Alternatively, the knowledge can be linked with other knowledge related to the same set (Fig. 1).

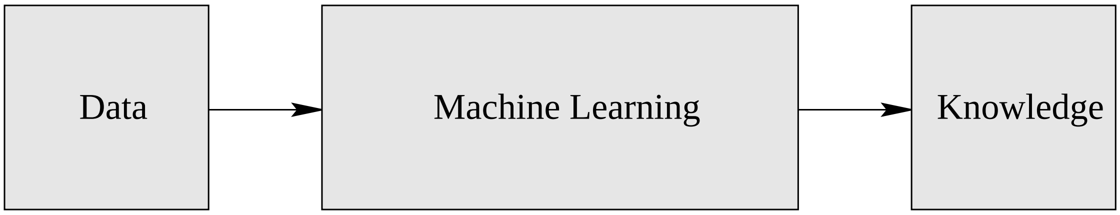

Figure 1: Data machine learning knowledge.

We would like to stress the notion of knowledge and knowledge processing as data mining works at creating the best representation of data in a process similar to the discovery process of science. The idea of automatic discovery of scientific knowledge is largely due to Herbert Simon. See [4]. A number of researchers have tried to apply the notions of automatic discovery to practical problems. Tetrad, [5], for example is a computer program for the discovery of causal relationships. Subsequent developments in machine learning, and thus data mining, are steps in the development of automatic science and automatic discovery. Although many applications of data mining are mundane endeavors (e.g., targeting a direct mail campaign), the enabling technique is the development of organized scientific representation and knowledge of the underlying phenomena.

While we put data mining in the framework of machine learning and the automatic production of scientific knowledge and discovery, machine learning has, however, a wider scope as part of the attempt to understand cognition, science being only a fraction of cognition. The language developed by the machine-learning community has a rich variety of meanings that might make the industrial user of these technologies sceptical or, occasionally, over-optimistic. This report will try to indicate the meanings and the techniques that are important to the representation and discovery of structures of data, separating them from other points of view that, although conceptually very important, are not significant for this endeavor.

On the other hand, the notion of automatic discovery and automatic science should not frighten anyone. We should understand that most of the high-tech products we currently use were made possible by consistent efforts at automating the process of discovery. Virtually any electronic device in use, for instance, has been designed with automatic procedures that have produced an optimal layout of components that would be impossible for a human designer to find. Methodologies of this type are now being introduced in areas that were traditionally the domain of hunch and intuition. Their use will produce significant changes in the type, quality, and structure of services provided, similar to the changes computer-aided design and engineering have already produced in areas dominated by the physical sciences. In particular, patterns of employment in the service area will be significantly affected by new methodologies.

There is an urgent need to understand the change and to make the new technology work for the increase of the wealth and the welfare of developed societies. In modern, post-industrial societies, services are responsible for the largest fraction of employment, with upwards 60% of the active population working in the service area. The consequences of the adoption of new scientific methodologies in areas traditionally far from them should be clearly understood from a global social point of view.

2 The Technology of Data Mining

In this and the following paragraphs, we will explore the problems that arise in data mining and how they are addressed, showing how the same type of problem presents itself in sectors as diverse as chemistry, retail, banking, and science.

Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data. There are two basic ways of performing data mining and data analysis. The first makes use of supervised learning, the other is based on un-supervised learning. In supervised learning, patterns are found exploring a number of known cases that show or imply well defined patterns: based on those cases, generalizations are formed. With un-supervised learning, data patterns are found starting from some logical characterization of the regularities. We can represent data mining as a dual process that can be either synthetic, generalizing from known cases, or analytic, expanding some high-level implicit description, as depicted Fig. 2.

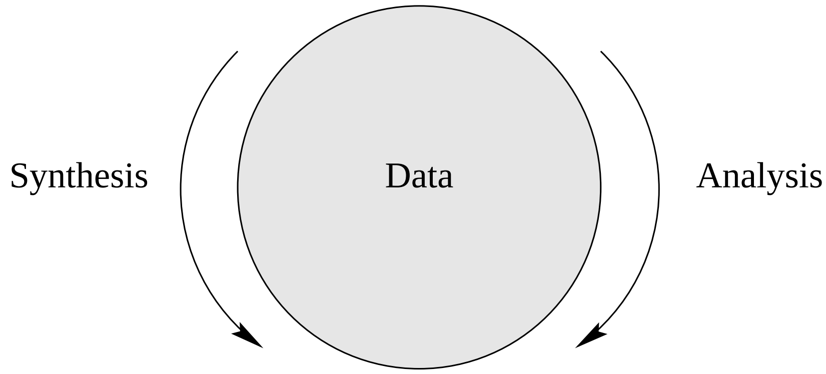

Figure 2: Representation of data mining as a dual process.

Supervised and un-supervised learning have other meanings in other contexts. In particular, as cognition paradigms and as paradigms of self-organization, they can be understood differently. This might create confusion and scepticism when machine learning is applied to rather mundane endeavors as retail or maintenance. It is therefore appropriate to remark that, in the context of data mining, machine learning is a set of techniques that perform two fundamental tasks:

1. they generalize from a set of known examples;
2. they detail structure from its descriptions.

Data mining allows to produce scientific knowledge in areas that are not yet suitable for the traditional approach of mathematical physics. Norman Packard has shown in [6] and [7] how learning and optimization algorithms can be used to produce optimal modeling of experimental data in the absence of previous theoretical explanations. Knowledge thus acquired has been applied to a number of different domains and used for engineering purposes. We currently use expressions such as rational drug design and financial engineering; we could easily talk of retail engineering. Data mining is a building block of these engineering methods as it provides basic knowledge on which exploratory search and design optimization can be based (Fig 3).

The objective of this report is to offer a unified view of data-mining methods. To this end, we will explore the methods of data mining and how they are applied to different sectors, showing

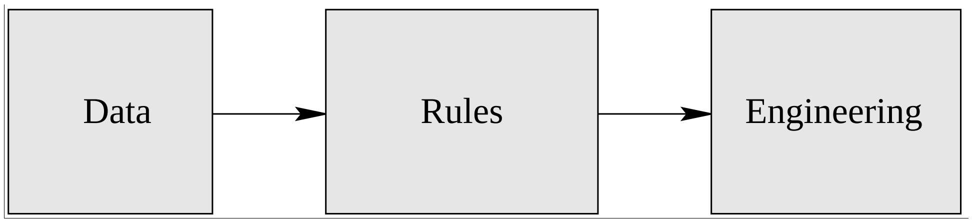

Figure 3: Data rules engineering.

how the same problem structure and solution is found in applications as diverse as chemistry, finance, engineering, and retail. Then we will analyze each sector to see how the techniques are being applied. This will inevitably involve some overlapping and repetition, but it allows to show how sectors apparently unrelated solve similar types of problems.

2.1 Data Mining Through Supervised Learning

The basic notion of supervised learning in data mining is that of the *classifier*. Classifiers are systems that classify input data, or sets of input data, in a number of possible categories. See [8] for a review of machine learning concepts and techniques. There is possible confusion around the notion of the classifier as the term *classifier system* is used in a broader context. See [9] for a discussion of different concepts of classifier systems. As depicted Fig. 4, a classifier is a system that receives some input and produces, as output, values that encode a classification.

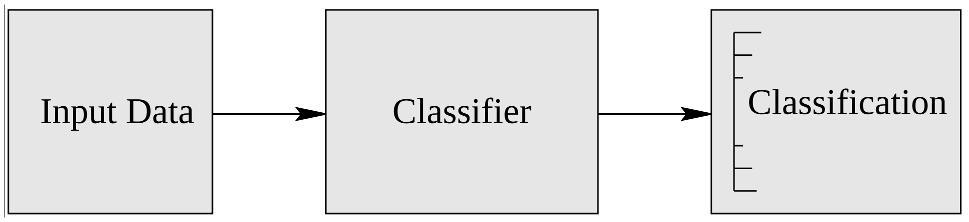

Figure 4: Classification of input data by a classifier.

In supervised learning, classifiers are trained and tested on a set of sample and test data. The classifier creates its own internal structure or rules based on the cases that are presented. The classifier is then used to classify inputs that it has most probably not seen before. The classifier then becomes a *generalizer*, as it generalizes from a set of samples to a set of external data.

There is immediately one distinction that has to be made. A generalizer is trained to produce a set of classification values on a fixed number of samples. When it generalizes, it might be restrained to use only the same classification values or it might produce additional values. A rule-based system such as Ross Quinlan's *ID3*, described in [10], uses a fixed number of generalization values while a neural network can in principle produce many more interpolation values not seen in the training cases.

We can thus define a generalizer as a system that implements a function of N variables $y = f(\hat{x})$ from a finite set of m couples $(y_j, \hat{x}_j), j = 1, \ldots, m$, where $\hat{x}_j$ is a vector of N values. The case where y is itself a vector can be easily derived.

The $\hat{x}_j$ can be a variety of parameters that identify individuals and situations. They can be a set of parameters to characterize a bank's client, such as sex, age, professional code, income level, but they can also be a history of past data, such as the purchase records of a client of a specific retail store. In other applications, the $\hat{x}_j$ can be preprocessed data as moving averages or computed derivatives on historical data. It is not necessary that variables be numerical ordered variables: they can be categorical variables without any notion of ordering. For instance, an input variable could be a classification of customers according to sex, which clearly implies no order.

The problem of generalization is made difficult by the fact that the number of possible different shapes of a function is enormous, even if we make strict assumptions about the shapes. If the set of samples were exhaustive of all possible cases, then the problem would be reduced to the construction of a look-up table. In this case, every possible pattern that could be presented to the classifier is known and classification is achieved simply finding the appropriate entry in the look-up table.

At the other extreme, if there is no constraint on the shape of possible functions, the problem of generalization is not solvable, not even in principle. Even if we had a set of infinite sample data, it would not be possible to reconstruct an arbitrary function. This can be seen in a variety of ways: the simplest is that there is an infinity of functions that are simply not computable. This is one of many results related to the un-solvability of the halting problem, and to the theorem of Goedel.

Every practical generalization problem lies somewhere in between completely constrained classification and completely free generalization. Data mining is based on the assumption that some level of generalization is possible. Stating this problem in full generality is not easy. One possible approach is to cast this problem in the framework of the theory of probability by evaluating the probability of correct generalization of approximation schemes. The paper by Abu-Mustafa, [11] provides a short but comprehensive introduction to this approach. Another framework for it is the notion of algorithmic compression introduced by Chaitin [12] and the Minimum Description Length principle. See Rissanen [13]. Wolpert [14] provides a number of recent advances on the mathematical aspects of the problem of generalization.

In practical applications, data mining is based on two assumptions:

1. the functions that one wants to generalize can be approximated through some relatively simple computational model with a certain level of precision;

2. the sample data set contains enough information to perform the generalization.

In other words, data mining assumes that regularities exist and that they are sufficiently simple to be approximated with manageable computational procedures.

In practical applications, there is the need to be precise about the complexity of the model that must be constructed. A naive view is that a complex model has better generalization abilities than a simple model. This is not generally true as there is a relationship between the true dimensionality of a problem, the complexity of a model, and the number of samples that are necessary for appropriate learning. A measure of the complexity of a model is given by the number of free parameters that need to be specified to construct the model. The dimensionality of a problem is the number of free, independent parameters that are necessary to describe the dynamics of phenomena. In the most general case, the number of parameters can be infinite, even a continuous of parameters, while only problems of finite dimensionality, i.e., problems identified by a finite number of parameters, can be represented by a model of finite length. The possibility of infinite dimensionality is another way of saying that the problem of generalization cannot be solved in general with algorithms of finite length, as any computer program is.

There are two major strategies for understanding the complexity of a problem, direct and indirect characterization. Direct characterization attempts to find a priori evaluations of the complexity of a data set. An important conceptual tool for understanding the dimensionality of a data set, such as a time series, is the notion of state-space embedding. State-space embedding consists in mapping a time series into an N-dimensional space, taking N consecutive past values of the series. Geometrical properties of the trajectories in the embedding space reveal important properties of the time series. Sauer, Yorke, and Casdagli provide in [15] a complete account of the problem of state-space embedding. [16], edited by Weigend and Gershenfeld, provides a good introduction to the problem of forecasting by analyzing the results and the methods used in a competition on forecasting organized by The Santa Fe Institute. A different approach to the problem of indirect characterization makes use of concepts of information and entropy. See Fraser [17].

Indirect characterization works essentially by optimizing models and inferring the dimensionality of the problem by the dimensionality of the optimized algorithm. Optimization algorithms work by either progressively reducing the complexity of oversized models or increasing the complexity of undersized models. At each step results are tested through out of sample testing. These ideas have been extensively used for neural networks. There is a vast literature on different methods for optimizing neural networks and other approximation schemes through out of sample performance measurements. See, for instance, Weigend, Huberman, and Rumelhart, [18], for a description of these algorithms. There are also commercial products that implement these methodologies, as the Siemens package for "pruning" their neural networks.

Data-mining applications are based on the assumption that the dimensionality of the problem and the complexity of the model match. This implies one of two possible situations:

- the model and problem have the same dimensionality and the model faithfully represents the problem;
- the problem has a higher dimensionality than the model, but the model is still able to generalize well as it captures the essential features of the problem.

Most practical problems fall into the second class: phenomena have a higher dimensionality than the model; this appears as random fluctuations in the space of unresolved dimensions. A good fit occurs when unresolved dimensions produce small fluctuations.

A satisfactory model must then have a complexity sufficiently high to capture the most important features of the phenomena. To build such a model, a sufficient number of examples is necessary. Exactly how many depends on the structure of the model and the training procedure, but there is a strict relationship between the complexity of a model and the number of samples on which it must be trained. If a complex model is trained on a number of samples that is too small, typically the model will fit the samples perfectly but will generalize poorly. This fact is known as over-fitting. Complex problems of data mining are feasible only when a large number of sample data has been accumulated.

Creating a good model is thus an interplay between

1. the dimensionality of the problem,
2. the complexity of the model,
3. the choice of features,
4. the number of available samples, and
5. the generalization ability of the model.

Before exploring the major types of problems in data mining, we have to make a few preliminary remarks on two technologies ubiquitous in data mining: *neural networks* and *induction trees*.

As far as data mining is concerned, neural networks are trainable function approximators. They are devices that can compute functions that approximate the values of the function provided by the samples. The remarkable point is that approximation is achieved through a procedure that adaptively changes parameters in the network in function of the samples. The book by Hertz, Krogh, and Palmer, [19], provides a good introduction to neural networks, showing how they can be analyzed in the framework of statistical mechanics.

Induction trees solve the same problem from a different stand-point: they find the regions where a sample function assumes values that lay within a certain range. Induction trees perform this task implementing a cascade of tests on the distribution of data among various partitions. As this region can be described by partition rules, they are said to induce rules. Ross Quinlan, [10] has developed the notion of induction trees and applied them to many problems. Induction trees can also be seen from a statistical perspective. See [20].

Neural networks and induction trees serve basically the same purpose: the former finds the value of a function for an arbitrary input, the latter finds the regions where a function will display an arbitrary output. They show, however, different abilities to handle specific data sets.

In many applications different specific algorithms cooperate to solve problems, as each algorithm may be best suited to certain regions of data or to perform specific functions. In consequence, an important aspect of research is related to technologies to hybridize different algorithms. David Wolpert, [21], has described a technique, called “stacking” to superimpose different layers of generalizers to improve global performance. See Goonatilake, [22] for a review of concepts of hybridization and recent advances in this area.

Let’s now explore the major categories of problems in data mining.

2.1.1 Selection of One Binary Feature

The selection of a binary feature is the problem of partitioning a set into two subsets that reflect the presence or absence of a particular quality. The training set is formed by a sample of individuals for which the classification is known. The problem is to extend this classification to the entire set. Individuals are characterized by a set of features that might be actual data or historical patterns. The correct classification is known from samples with a certain number of possible combinations of features and needs to be extended to combinations of features that do not appear in the sample list. This problem appears in a vast class of cases including targeting direct mail campaigns, fraud detection, bankruptcy prediction, detecting “interesting” sonar patterns, substructure molecular search, survival prediction.

The abstract problem is to generalize a step function which has values 0 or 1 on two sets, or collections of sets, of the sample space, to a step function that assumes values 0 or 1 on a partition of the entire space. Although the abstract problem is the same in all cases, its characteristics are different and thus require different techniques in function of the available data.

The problem of *targeting direct mail* can be very different in function of the data that the mailer possesses. In the simplest case, data are descriptive features such as address, sex, family status, income and job classification, that might have been gathered by any number of means, plus a history of the results of previous mailings.

However, the use of credit cards and electronic point-of-sale (POS) devices has allowed to gather historical data on the spending of individuals. A large retail store might have a complete history of the purchases of each of its credit-card clients, in the range of millions of individuals. And historical data on the purchases of a significant fraction of the entire population are available to the generic credit-card companies.

With this large amount of data, it is possible to create a much larger space of features. As remarked above, it is important to note that expanding the space of features requires the availability of more samples for training as there is a combinatorial explosion of the model space. So to make profitable use of historical data requires gathering data on a very large amount of individuals.

In all cases, results will be cast in the language of probability as it is not possible to predict individual responses to a mailing but only the probability of response for each individual or for each subset of selected individuals. From this probability distribution it is possible to engineer the mailing campaign in function of commercial goals. It is, for instance, possible to make very selected mailings to target groups with a high response probability relative to a specific offer or to optimize the mailing in other ways.

How are machine learning techniques used ? There are issues of features selection, pre and post-processing of data and of basic methods.

The most widely used techniques are neural networks and identification trees. Neural networks might create a function equivalent to a probability distribution over the entire feature space, starting with a sharp distribution of respondents/non respondents in the training cases. In this mode, they receive as inputs the recorded features of each client, including his/her history of past responses, and produce as output values that encode numbers proportional to the probability of answer to the next mailing. The final classification into two sets, the ones who will receive the mailing and the ones that will not, depends on the post-processing of probability distribution, for instance choosing a threshold of probability to classify an individual as a potential respondent. Ultimately, this is an optimization job.

Identification trees might also create a distribution of probability over the partitions of the features space that they create. As with neural networks, they can alternatively be trained to produce a binary classification of respondents/non-respondents in which case the probability thresholds are implicitly defined in the training procedure.

If a probability distribution is defined over the set of partitions, it is possible to optimize the mailing in function of many criteria. If not, i.e., if training produces a yes/no answer, it is necessary to specify from the start the objective of the mailing.

The selection of features is an important process for two reasons. First, it is necessary to constrain the feature space to really important features to avoid unnecessary computations.

It is also important not to miss important features. It might also be important to construct appropriate features from historical data. Historical data on each individual need to be in a uniform format, and eventually some preprocessing, such as computing slopes of curves, must be done. If data are available, feature selection can be done adaptively from the performance of the classifier.

The problems of *fraud detection* and *bankruptcy prediction* have different characteristics. Here the goal is to make sharp predictions of relatively rare events. Clearly one does not want to stop service to or to irritate a good client, while at the same time wants to avoid the cost and problems of a defector. As the cost of a fraud or bankruptcy is high relative to the gain on the same account, there is no or little room for optimization. Thus the objective is not to optimize on a probability distribution but to sharpen as much as possible the evaluation of potential problems.

The comparison of the problems of fraud detection and of target mailing elucidates one fundamental issue of data mining: better generalizations imply better knowledge. Increments of knowledge can come from three sources: gathering more data, so that more complex models can be built; improving the model, so that more information is gained from samples; or gaining more theoretical insight, so that the problem can be better constrained.

In direct mail, it is essentially the amount of data that matters, and so the possibility of handling it flexibly. Practical applications, such as the one implemented by Epsilon Data Management of Cambridge, MA, currently uses up to 500 features with plans to go as high as 2500 features. Fraud detection is a more difficult problem and requires some basic research on what data are really important and how to represent them. Commercial implementations such as HNC or Nestor use from thirty to seventy features, but there is research work going on at Los Alamos for a consortium of banks to significantly reduce dimensionality.

2.1.2 Multiple Classifications

Multiple classification is a natural extension of the previous set of problems as the classification is not only in terms of a yes/no answer but involves a rating or a multiple classification in a number of predetermined classes. Typical examples include medical diagnosis, equipment testing, credit rating, bond rating, stock selection, classification of risks.

The abstract problem is to generalize a step function that can assume one of N different values on a set of training points to a step function that can assume one of N different values on a partition of the entire space.

Medical tests are among the first sets of data that were mined. The problem is to infer one of a possible set of diagnoses from a set of symptoms and medical tests. Each variable can be a continuous variable or a categorical variable such as the classification of a certain symptom. The problem is to formulate a diagnosis from a set of known cases.

The problem of equipment testing is similar. A number of instrument readings and condition checks are used to diagnose an actual fault or to predict a potential fault. Instrumental readings can themselves be the product of some machine-learning technique. As in the case of medical diagnosis, a set of past successful equipment diagnoses is used to train the system.

The problem of diagnosis has been handled with expert systems. However, as experts are difficult to find, techniques of machine learning were applied successfully. The major challenge in diagnosis is clearly the accuracy of classification. Only a few systems, and only in the area of equipment maintenance, are used with full confidence, acting blindly on the diagnosis. In most cases, these systems are used as expert assistants to professionals. For this reason, diagnostic predictions must be easy to understand, so in general rule-induction systems are used.

Another version of the multiple classification problem is the problem of rating which is very common in business applications. Banks need to rate the credit worthiness of clients, investors need to rate securities to manage their portfolios, and so on. Baestaens [23] has applied neural networks to the problem of credit rating. One potential difficulty of rating systems is that they are trained on data that are themselves subjective. For instance, a credit rating system might be trained on a past history of rating given by professionals. Rating of this type might reflect individual biases of the professionals or be faulty in other ways. As a consequence, a machine-learning system might not really be given the chance to improve working on factual data. Another significant problem is the lack of sufficient data. A stock rating application might be trained on several dozen samples, a small training set that does not allow for a rich feature space.

Rating systems are a relatively common application typically implemented using neural nets or induction trees. Neural networks can be trained to produce sharp classifications or a multivariate distribution of probability. In the latter case, the rating must then be extracted fixing thresholds.

2.1.3 Function Reconstruction

Function reconstruction is a further evolution of previous generalizers: one has a number of values of a function and wants to reconstruct the entire function, at least a certain region of it. The need to solve this problem in its full generality arises in a number of instances. The most important is *time-series forecasting*, treated separately in the next paragraph. We have seen previously how a neural network can reconstruct the entire probability distribution of the response to a mailing. Problems of this sort arise in the context of, for instance, computing demand surfaces and price elasticities, in machine tuning, and in interpreting experiments.

The problem of computing demand surfaces is very common in business. The demand surface is a function that relates sales to a number of parameters such as the price of goods, advertising spent on the product, and the income of buyers. Once a demand surface is known, one can use standard optimization algorithms for business engineering. The same problems arise in many different contexts in business, when a certain business quantity clearly depends on other quantities but it is not known exactly how. For instance, the amount and location of shelf space in a retail store affects the sales of products exposed: one would like to know exactly in what measure. The same applies to changes in advertising budgets or price reductions during a promotion.

There is a vast literature on how to use these functions when available, but there is very little knowledge on how to construct these functions. These phenomena clearly escape any reliable standard mathematical formulation for instance in terms of differential equations. In addition, there is no way to perform adequate experiments: if one would start experimenting with business parameters, one risks bankruptcy well before finding any solution.

However, sets of past data under different conditions are available: past response to promotions, changes in sales in consequence to changes in pricing and/or advertising budgets, and so on. The problem with these data is that typically more than one parameter changes at the same time. The problem then becomes one of function estimation based on a selection of values.

One criticism that can be immediately raised against this approach is that past business data cannot be used reliably as the business models themselves change over time. In the end, this is not true over relatively short periods of time where all relevant changes can be parameterized. It is interesting to compare this statement with an equivalent statement regarding a complex physical system. In this case, one assumes no fixed functional relationship but a higher level system of laws that are formulated as differential equations. In physical systems, one assumes that fundamental differential equations are stable, a statement that has not been proved false.

For economic systems there is no such general model. However, one might try to write down a model as a family of functions with a number of degrees of freedom and adjust the free parameters in function of measured data. A business model is one such model where all the structural changes are reflected in a number of basic parameters.

There are many implementations of this type of business modeling. Some use neural networks, other use induction trees. More recently, there are business models that make use of hybrid techniques. These packages have reached a significant level of user friendliness through human interfaces in the form of spreadsheets and through the use of problem description languages.

Machine tuning is the prototype of a number of data-mining applications where a set of data relative to a physical process is available and one wants to reconstruct the entire function. In particular, for equipment tuning a set of past values of parameter adjustment is known, together with the relative performance measure. The problem is to reconstruct the tuning surface as performance in function of adjustments. Once the tuning surface is known, it is possible to find the optimal tuning setting through standard optimization algorithms. This problem is perfectly similar to the previous problem. Neural networks and identification trees have been used successfully to solve a variety of these problems.

2.1.4 Forecasting

Forecasting is the application of data mining that has attracted the highest level of attention given the potentially enormous financial implications of successful forecasting. The basic idea of forecasting is to use a large number of past values of one or more time series to discover the future values of one series.

In principle, the problem of forecasting is a problem of function reconstruction. However, most series of economic interest, in particular time series of financial data, have a low level of forecastability. For this reason, forecasting financial time series is a discipline in itself, as much theory and empirical testing is being used to extract a modest amount of information from what was believed, until recently, to be pure noise. In fact, standard economic theory predicts that an efficient market immediately and efficiently reacts to external inputs so that there is no residual forecastability. Recent analysis has shown that this theory is not correct and that financial time series can be forecastable to some extent. A significant amount of research in time-series forecasting is now being devoted to finding the regions that offer the best forecastability. Exploring these techniques is beyond the scope of this report. See [16] for a survey of forecasting techniques through the results of a competition held at The Santa Fe Institute.

However, not all business series are so difficult to forecast. As already seen, we can predict the return on a mailing campaign with good accuracy. Another important area of forecasting is *sales forecasting*. With the large amount of data generated by credit cards and electronic POS devices, sales forecasting can be done at the level of individuals. The ability to forecast at the level of individuals is an advantage. First, it offers better forecasting abilities as the average of individual forecasting gives a better prediction than forecasting of the average, i.e., forecasting data taken in aggregate. A second important advantage is the ability to forecast the spending of consumer groups that might be chosen. This allows to profile spending patterns according to different groupings.

Neural networks are becoming a very popular technique for forecasting. Historical data on purchases either of individuals or of preselected groups of individuals are fed into neural networks to teach patterns of spending. Once the network is trained, the most recent historical data are used to predict future purchases. In this way, a complete profile of individual spending can be constructed. From these data various aggregates can be built. Aggregates can be selected not only through a simple mechanism of thresholds, but also recognizing the shape of spending. For instance, one might want to categorize clients on the basis of the seasonality of their purchases.

Induction trees can also be used to forecast sales. Induction trees do not compute individual spending, but group clients according to a forecasted level of spending. In other words, they explore past records to determine how to segment clients that show a specific level of purchases. Aggregated data can then be computed easily. The grouping of clients can be more complex, for instance classifying not one single purchase datum but a sale pattern.

In addition to business data, physical phenomena can be forecasted as well. Many physical phenomena are known only in their global behavior and there is no hint of how to compute their future values from first principles. This is often the case in environmental sciences where a detailed description of the natural phenomena is often beyond reach. In other instances, although precise calculations or experiments can be performed, it is desirable to gain a quick insight into the possible outcome of experiments in order to limit the time and cost of performing them. An example is the prediction of chemical activity of compounds from a number of computed or measured quantities. In this case, it would be possible to make either a detailed calculation or to perform an exhaustive set of experiments, but it is highly desirable to limit these expensive tasks to only promising compounds. As a rich history of past cases is often available, it pays to try a machine-learning approach.

2.2 Un-Supervised Learning

In the above sections we have described a number of problems of supervised learning. The following paragraphs will describe how data are mined with un-supervised learning. While in supervised learning the patterns are known on the samples and need to be generalized, in un-supervised learning the patterns are not known in their structure. The starting point of un-supervised learning is some description of the type of patterns required.

At the one extreme, there is the simple data inquiry where one single datum is located. One level above, there are more complex inquiries where combinations of conditions are imposed and all the data that satisfy those combinations are retrieved. One might, for instance, require all clients whose exposure is above a certain level in a certain job code. Inquiries like this may be very complex, but they are characterized by a combination of conditions that select single individuals.

There is another level where no individual information is specified but only descriptions of relationships between individuals or descriptions of sets of individuals are given. In this case, the learning algorithm searches the data base trying to create the partitions that match the described relationships. We will now explore a number of major problems in this area. Kohonen [24] is a tutorial on self-organization and associative memory. Kohonen himself has done pioneering work on self organizing maps. It has to be remarked that the field of self organizing systems has conceptual implications that go well beyond the technology of data mining.

2.2.1 Clustering and Segmentation

Clustering and segmentation is the problem of creating a partition of the data base so that all members of each set of the partition are similar according to some metric. If we have a set of balls, some of them white and some red, a natural partition is to group together all the white balls and all the red balls. If in addition the balls come in two sizes, big and small, it is natural to create a partition of four sets, big-red, big-white, small-red, small-white.

Suppose now that one wants to partition the children of a school. There is no clear criteria. One could start by defining intervals of continuous variables, such as age, and create all the possible partitions. Even for a small number of features, the global number of partitions is very high, and not very useful. So one has to ask what is the objective of the partitioning. For example, one such objective could be to match students to create the best possible classes. One criteria could be the homogeneity of classes. First of all a measure of homogeneity is needed. A class is homogeneous if all of its members are similar according to some similarity metric. One has therefore to define what makes two students similar. Of course one might have an a priori notion of classification of students; in this case the grouping is trivial. However, one might have only an idea of what makes two students similar in various degrees.

Clustering according to *similarity* is a simple but powerful concept that appears not only in marketing and social sciences but also in physical sciences. Chemists, for instance, find it useful to rate molecules according to structural similarity criteria. The key point is to translate an intuitive measure of similarity into a quantitative measure.

If a measure of similarity is available, there are a number of techniques for forming clusters. One class of techniques consists in forming rules that dictate membership in the same group based on the level of similarity between members. These rules are not trivial, as similarity is not an equivalence relation: if A is similar to B and B is similar to C, it does not follow that A is similar to C. In order to form clusters, similarity constraints must be combined so that they create an equivalence relation. After an equivalence relation is formed, a set can be clustered in equivalence classes. There are many ways a similarity rating can be organized into an equivalence relation. The choice of the methods depends on the type of clustering that is sought and of computational considerations.

Similarity is one possible consideration in forming clusters. A different important criteria is to build *set functions* that measure some property of partitions as functions of some parameter of the partition. Suppose a set A is given, and suppose that for each partition of A into disjoint sets A_j one can define a function $F(A_j)$ that measure some property. One can then build a global function $F(A)$ as the sum of $F(A_j)$:

$$F(A) = \sum_j F(A_j). \tag{1}$$

Now, if the partitioning in sets A_j can be made to correspond to a vector of parameters $\hat{x}$, then a function $G(\hat{x})$ exists such that

$$F(A) = G(\hat{x}). \tag{2}$$

It is then possible to define a clustering rule as optimal partitioning, i.e., as the search for the $\hat{x}$ that maximizes $G(\hat{x})$ and, consequently, $F(A)$. This notion of clustering finds important applications in data mining as it translates naturally important problems. For instance, Norman Packard, co-founder of The Prediction Company [25], applied this idea to select partitions that carry the most information content, an idea that plays a central role in analyzing chaotic systems. Packard later applied this technique to the problem of forecasting currency exchange rates.

The two above techniques are significantly different in the results they provide. Clustering according to similarity rules produces clusters of arbitrary shapes, while the maximization of set functions is constrained by the need to build an effective correspondence between partitions and a set of indices so that a numerical function can effectively be maximized. They might also be very different in terms of the computational needs of the algorithms. Clustering according to similarity is a technique used in a large number of data mining applications in different sectors: segmenting the client base and clustering molecules are two applications similar in their abstract conception although very different in their objectives and areas of interest.

Clustering according to the optimization of set functions is used in data analysis. There are important business applications. For instance, setting insurance tariffs is a problem of cluster optimization. In this application, customers are segmented according to a number of parameters and the optimal tariff segmentation is determined. Another application is the selection of the best areas for business development. In this case, the optimal clustering of factors in function of their location has to be found.

Clustering can also be implemented in supervised mode starting from a set of given clusters. New members are then added to each class in function of their similarity with the members of the classes. The n-nearest neighbors method puts each new member in that class that appears most frequently among its n nearest neighbors.

3 Applications Areas for Data Mining

3.1 Introduction

The previous section explored the basic technologies used in, and offered a framework for, data mining. This section explores a number of sectors where data mining is being applied, and illustrates a number of major application areas. The perspective of this section is practical: it shows how data-mining applications are being effectively implemented.

The following sectors will be covered: chemical/pharmaceutical, retail and consumer banking, financial applications, remotely sensed data, engineering and maintenance data. Although this is not intended to be an exhaustive report, it includes a number of applications that are significant for the field. The reader might find of interest the perspectives of the European Community [26] and of the Japanese Real World Computing Program [10] and [27] on research and applications of adaptive methods.

As we shall see, different sectors offer a different picture in terms of the state of the art of the applications. This situation is due to various factors, some of them of a technical nature, some of a management nature. The previous section showed that there is a natural framework for grouping applications that are structurally similar though at first sight they might appear distant. This suggests that it should be possible to benefit from a cross fertilization of ideas. The research community, industrial users, and suppliers were interviewed to this end.

3.2 Chemical/Pharmaceutical Sector

In the chemical and pharmaceutical sectors, data mining techniques are used as part of the solution of the global problem of chemical information management. The chemical and pharmaceutical industry produces a staggering amount of information, either in the form of company-proprietary information or as published results. Companies such as Sandoz and Pfizer cited the need to build ever bigger databases, entering and integrating additional data as biological information. The question of how to mine all this data to help researchers check their hypotheses is identified as a key problem. In more general terms, the ability to access and exploit tremendous amounts of data has become a critical factor of success and players as Akzo Pharma are putting this high on their priority list.

Let's go back and state the problem of information management in chemistry. There are several sources of information for the chemical and pharmaceutical industry. First there are commercial databases of chemical products and related information. These databases contain information about chemical compounds whose structure and properties are in the public domain. Second there are results produced by universities and research laboratories that are published world-wide in scientific and technical magazines. Another source of information is company proprietary data generated by performing experiments on compounds, and field data on the behavior of commercial products.

Data fall into four major categories, each subject to different treatment: chemico-physical data on the structure and properties of molecules, data on the chemico-physical behavior of compounds, data from field tests of compounds, and finally additional information on patents and other facts of economic and legal relevance.

Chemical information management addresses basically two types of needs: first, to help the discovery of new useful compounds, second, to analyze and interpret field data to evaluate and improve the usefulness of products and their contribution to the company's profit. Let's start by exploring the use of information technology, and in particular data mining, in the discovery process.

The chemical discovery process is a design process that starts with functional specifications of a product and produces a structural description of a molecule and a manufacturing process. In this respect, it is not basically different from any other design process that starts with some functional description of the desired product and produces a structural description of it that might lead to the manufacturing process.

As in any design process there is no straight path from functions to structure; it has to be constructed through creative discovery. The process of creative design is generally implemented through an adaptive search: a large number of potential solutions are created and then adaptively modified to fit the design goals and design constraints.

A peculiarity of chemical design or discovery is the huge amount of potential solutions and the relatively little space left for optimization. A molecule might be a very useful compound while similar molecules might be totally useless. Finding the molecules that perform a specified task is a difficult endeavor that implies exploring thousands of potentially useful compounds. This situation is unmatched by other design processes where the basic features of the design are fixed in the early design phase and the rest of the time and resources are spent optimizing.

As searching is expensive and the time factor is fundamental in producing profitable compounds that must be patented before competition, it is clear that chemical and pharmaceutical companies need to reduce the space of search and to increase the speed of search for new compounds. Chemical and pharmaceutical companies are very eager to exploit the techniques of automatic discovery in the race to find new chemicals and new drugs. In the last few decades, chemical discovery has evolved towards the notion of rational material and drug design. The pillars of rational design are four: a better understanding of physical and biological processes, the accuracy and speed of experiments, simulation, and the use of information methodologies.

The understanding of physical and biological processes brings functional specifications of chemicals much closer to their structural descriptions. For instance, in drug design functional specifications of new drugs are cast in the language of molecular activity as knowledge of biological processes is now at the level of molecular activities. This greatly reduces the space of search for new compounds. This is true also for materials and industrial chemical design, although this is a much less constrained process.

Experiments can now be performed at a much higher speed than it was possible two or three decades ago. In a number of cases, the gain in speed is many orders of magnitude. In some areas of biotechnology, single tests can be performed at speeds in the range of thousands of compounds tested per day. The amount of information that it is possible to extract from single experiments has also grown at a fast speed. Physical experiments can give quite an accurate description of the structure of complex molecules in a short time.

Computer simulation is a technique that allows to gain deep insight into the molecular structure of compounds from purely theoretical considerations. It allows for rapid evaluation of physical properties of new molecules without the need to conduct experiments.

All these advances have produced an explosion of the amount of data that any researcher working on new molecules needs to manage. If properly used, these data can be very helpful in the rational design of drugs, materials, and industrial compounds.

There are many challenges in handling chemical information. First there is the problem of the language used to describe molecules and the amount of information that descriptions carry. In chemistry there is no uniform notation. There is a basic *2-D* notation that describes the fundamental structure of a molecule, but there are many different notations and chemical languages to describe this two-dimensional structure. This variety of languages and notations is a challenge in itself as chemists need to translate notations and descriptions.

2-D descriptions are not very close to physical reality. To gain additional insight in the structure of a molecule, a *3-D* representation is needed. In this representation, a molecule is represented as a spatial three-dimensional structure of atoms and bonds. Software programs able to translate a *2-D* structure into a *3-D* structure are now available. This translation is not one-to-one as a *2-D* representation can be expanded in a large variety of possible *3-D* structures under different physical assumptions. Thus *3-D* databases can be very large, of the order of tens of millions of molecular shapes.

In addition to structural information, data bases contain data on physical properties of compounds, commercial names, patents, and so on.

How are chemists using all this information ? How are they mining these databases searching for information that can help in the discovery process ? Operations on data bases fall into five major categories:

1. clustering molecules for similarity,
2. searching all molecules that contain a specified pattern,
3. searching all molecules that bind in a specified way with a target molecule,
4. predicting the results of experiments on new molecules from stored data,
5. checking patents and other data.

Similarity evaluation is an important step in the process of discovery as it allows to explore different compounds that might share important beneficial features. It is also important to discovering chemicals that are unique, as they don't cluster. Clustering is basically the process of grouping together all the molecules that present a certain level of similarity among themselves. The definition of similarity is somehow arbitrary; consequently there are different measures of similarity that are used in chemistry, each leading to a different clustering of molecules.

Commercially available products implement a measure of similarity starting from some type of idealized representation of molecules. Similarity is then measured by concepts such as the Euclidean metric or Tanimoto coefficients, that give a measure of the number of bits that two descriptions have in common. Clustering is intended to help the imagination of chemists, so it should be very intuitive to be effective. See [28] and [29] for the description of two commercial products used for *3-D* and similarity searching.

Clustering is a data-mining technique. Given a data base of molecules and a similarity metric, molecules are grouped together in sets that maximize the level of similarity. It is a purely data-processing method that allows to discover important chemical properties. Jarvis [30] describes the techniques for clustering of similar molecules.

Searching molecules that contain a specified pattern as a substructure is another important conceptual tool in the chemical discovery process. It is basically a pattern recognition problem. In this case, a query to a chemical data base is generated specifying a substructure; the system must retrieve all the molecules that are similar to the that structure according to the chosen metric. See [31].

A different type of search is performed when searching for molecules that bind with a target molecule. In principle, this problem could be solved exhaustively, but the computational cost would be prohibitive. Different types of searches must therefore be implemented to limit the search space. There are a variety of techniques implemented in commercially available products, mainly to search conformational space.

Lastly, data bases can be searched to obtain predictions of the chemical activity of new compounds based on the activity of known compounds. In the past, expert systems have been used to predict chemical activities. Research efforts are under way in a number of companies to use neural networks in this application. At the Los Alamos National Laboratory, a neural network was developed to predict the level of toxicity from an Ion Mobility Spectrometer IMS measurements, as described in [32]. Glaxo is experimenting with neural nets in spectral determination. Akzo Pharma is using them together with genetic algorithms in QSAR (Quantitative Structure Activity Relationships) to reduce the data needed to find significant relationships. However, most experiments with neural nets and GAs are in the early stages; a final evaluation of the technologies is still to be made. It is nevertheless fair to say that while industry finds that these techniques tend to work, the lack of explanation as to why is a point of concern here as in other industrial sectors where experimentation is at an early stage.

The general findings of the field research can be summarized as follows. First, nearly all respondents pointed out that building a coherent data base (data warehouse) is still a major problem in view of the staggering amount of data and differences in the formats. Decisions about products and techniques have been made over a number of years and the integration of new software products into a coherent data management system is a major preoccupation.

The general tendency is towards an ever greater use of commercial software products, limiting internal development to areas that are of crucial importance for the company and/or where no commercial product is available. Amongst the products most widely used are DayLight, MDL, and Tripos suites of products. Areas where in-house development will continue due to the competitive edge to be gained include similarity/dissimilarity searches.

The data-mining applications most widely used are similarity searching, *3-D* searching and the prediction of chemical activity. There are some in-house research and development efforts on neural networks and genetic algorithms as mentioned above, particularly in the area of forecasting chemical activity.

Chemists are often reluctant in experimenting with computational techniques. They need to understand according to their mental models, and adaptive methods represent a significant departure from standard chemical thinking. It was pointed out that chemical information management systems must help chemists in their work, offering information that helps the creative process: only after acquiring familiarity with simple tools, as those based on spreadsheets, will researchers turn to even more sophisticated computational methods.

The second important area for data mining is the analysis of field data, in particular clinical tests and data coming from the use of industrial chemicals. Potentially this is an area of large use of sophisticated data-mining techniques, given the size of data generated and the importance of finding correlations. However, the findings of the field research show that most companies use standard statistical methods available in commercial statistical packages. This is understandable, as the primary need of pharmaceutical company is to get drugs approved; data analysis must be presented in standard forms. However, it seems that data-mining techniques could bring significant benefits if properly integrated into the corporate discovery process.

3.3 Retail and Consumer Banking Sectors

The retail sector is becoming an important area for the use of sophisticated data-mining techniques. This statement needs a word of qualification. Only few retail chains have the size and the resources to undertake advanced data-mining projects. However, where these techniques are used, they are implemented using some of the most sophisticated methodologies available. Implementations are often the result of in-house development, eventually with the cooperation of universities and research centers.

To understand the use of data mining in retail, it is important to appreciate the size of the operations carried out by a large retailer. Companies as Wal-Mart or K-Mart in the USA or Marks & Spencer in the UK typically carry up to two hundred thousand different products, in hundreds of different outlets located in areas characterized by different consumer behavior, income and so on. Marks & Spencer, for instance, has 600 outlets that are visited weekly by 14 millions potential consumers. Wal-Mart, the second largest buyer in the USA after the federal government, has 2000 retails shops plus nearly 400 business retail shops.

In retail, the sale of each item is sensitive to a number of different factors: marketing factors, such as advertising campaigns and shelf space, exogenous factors such as fads and disposable household income, competition, and many others. There are internal correlations between the sale of different products, for instance between dresses and accessories. Relationships among sales parameters can be highly non linear: a small change in apparel sizes might produce substantial changes in sales. In addition, pricing is not a constant, but must follow a precise dynamic to exploit the waxing and waning of fads, seasonal factors, etc. Prices might start low to jump start sales, rise to exploit that fraction of early buyers that are willing to pay more, drop as sales lag. This is the well-known bell-shaped curve of retail pricing, aimed at appealing to different segments of the consumer population. The curve parameters must be optimized to maximize profit.

In brief, organizing sales in a large retail chain is a complex engineering task, characterized by difficult dynamic optimization problems. According to A. C. Nielsen's European Vice President for Statistical Operations Oreste Assereto, in the past retailers and marketeers were interested basically in tracking market share evolution; today they want to be able to evaluate the impact on sales of any number of variables, with the objective of forecasting.

This explosion in the number of different parameters and the potentially complex relationships among them has rendered increasingly difficult the problem of understanding the dynamics of marketing. It is, however, possible to assume that purchasing patterns are a rather stable phenomenon over relatively long periods of time. To the casual observer, the madness of fads might seem nearly random while it follows rather precise, though complex, rules. It is then appropriate to assume that consumer behavior can be studied with scientific methodologies, replacing the hunches of shop managers with precise quantitative evaluations and optimization. This is precisely what leading retailers as Wal-Mart are doing.

In the past, given the cost of computing power and the paucity of data, only a rough analysis of statistically aggregate data was possible. However, the advent of electronic point-of-sale (POS) devices and the widespread use of credit cards has put an amazing quantity of data at the disposal of the analyst. It is possible to follow purchasing patterns at the individual level as purchases are typically tracked by credit card numbers and individually registered at the POS. In addition, sales data are now commercially available, as data from POSs are purchased by marketing organizations and then made commercially available. A large retail chain now has precise individual data on a large fraction of its customers, i.e., the vast majority that pays with credit cards, and has access to consolidated data that can be purchased through marketing organizations.

Wal-Mart uses the massively parallel Teradata machine on customer files containing over 2 Terabytes of data on individual purchases, and data are growing at the rate of 20 million transactions per day. Telephone companies, a kind of retailer albeit of a special nature, can generate Gigabytes of data per day related to calls placed. To handle these data, AT & T has more than 50 Teradata machines with over 3000 disk drives; the telecommunications giant represents the single largest user of the powerful Teradata parallel processors.

The size of these data sets has suggested the possibility of tackling the problem of retail from a scientific point of view. The fundamental assumption is that consumer behavior can be categorized with a high level of precision, provided that a sufficient number of parameters can be taken into consideration. It is important to understand exactly what is meant by this assumption and why so many data are needed.

The scientific approach to retail starts from assumptions that are similar to statistical mechanics, with one important difference: the dynamic of each individual customer is not known exactly, not even in principle. In statistical mechanics, the trajectory of each particle follows well known rules. Rules might be cast in the framework of quantum-mechanical probabilities, but they are known a priori.

The data mining approach to retail assumes that consumers follow statistically stable but unknown rules. Not only rules are not known but there is no attempt to discover them through individual experiments: the objective is not to find psychological characterization of customers but to understand with great precision how they behave in aggregate.

A retailer, for instance, is interested in understanding what proportion of early buyers of a certain product line that fall within a certain parameters range will respond to a campaign to renew purchases at later stages in the sales cycle. It is clear that there are precise assumptions about the behavior of consumers in aggregate behind these questions, but there are no further assumptions about the individual psychological profile.

The reason why so many data are really necessary is that a rich set of rules generates an enormous space of possible scenarios. This point can be easily misunderstood. If one wants to evaluate a single statistical parameter, a relatively small sample can give a very accurate answer. At elections, exit polls give very accurate answers based on small samples of the population because aggregation regards a choice among a small set of possible candidates. In retail there are hundreds of different possible parameters to generate rules with mutual influence: the space of possible scenarios is thus enormous. One needs a very large amount of data to create a set of rules with forecasting.

For the same reason the problem of retail cannot be tackled with the classical methods of mathematical physics. For instance, it would be fruitless to regard clients as a continuum in a state space governed by differential equations. There is no way to simplify the problem at individual level as the "atoms", i.e., individual buyers, are complex structures themselves. On the other hand, if aggregates are considered, the problem is still unmanageably complex by human standards.

The solution to the problem is the use of machine-learning techniques that can effectively find rules exploring billions of individual relationships. However, it might happen that out of this effort, with time, a smaller set of rules can be distilled and that, finally, a concise description of the problem in mathematical physical terms can be found. This, however, is by no means a task of today's retailers, whose objective is to maximize sales and profits through an optimization of marketing actions. To this end, we can say that the applications of data mining in retail fall into two major categories:

- sales forecasting;
- optimization of marketing strategies.

3.3.1 Forecasting Sales

Forecasting sales is a problem of forecasting a time series from past data. This problem has a long history in statistics. The first attempts to solve the problem of forecasting time series made use of linear models, i.e., functions that are a linear combination of available data containing parameters to be determined. Linear methods of statistical forecasting were cast in a sound theoretical and practical framework by Box and Jenkins in the early 1970s. They devised a methodology, named *Box-Jenkins methodology*, for forecasting time series based on linear methods. The methodology includes model selection, determination of free parameters and diagnostic tests. The most widely used method of forecasting, the Box-Jenkins approach is included in every statistical package, such as *SAS* or *SPSS*.

Linear methods have limits when the phenomena to be forecasted are complex and non linear, as is the case of the sales data of large retailers. The availability of large computing power has allowed the use of adaptive methods. To solve the problem of sales forecasting, neural networks and induction trees have been used with success.

Induction trees explore the past sales data to extract prediction rules for the future. The principle of induction is rather simple but its application to sales forecasting can be complex. Data are not homogeneous and need to be preprocessed. This is a difficult task in itself. The major problem neural networks meet in retail is the fact that their predictions are not easily understandable. A neural network is a function approximator and it produces as output a function that minimizes a certain functional: this is not an intuitive process. Genetic algorithms have also been used to generate induction trees and to optimize prediction networks.

The most important use of sales forecasting is for the optimization of purchases and stocks. If a retail chain can predict with accuracy sales per item and per location it can optimize purchases of individual items, taking into consideration parameters such color and size. In addition, it can optimize the level of stocks to be kept in each location in order to minimize the cost of stocking while, at the same time, ensuring that customers find the products they want. The latter is an important factor not only in maximizing sales but also in attracting and keeping clients.

In an industry characterized by very low margins, the ability to accurately forecast sales and in consequence to place orders to suppliers is critical. The objective of Wal-Mart is to completely automate the ordering process, basing it on automatic sales forecasting.

A number of applications have been implemented in Europe. SearchSpace, a London-based company specialized in machine-learning methods, has developed an application using neural networks to forecast sales for a high-street retail organization. In Germany, the Karstadt retail chain uses a neural network system developed by Neurotec to predict the sales of the 200 000 items carried in their stores in an effort to optimize orders. Neurotec has also developed an application based on neural networks to predict the daily sales of newspapers for the publishing company Axel Springer Verlag ; the objective is to forecast the number of copies that will be sold daily in an effort to reduce production costs without losing sales.

3.3.2 Retail Engineering and Strategies Optimization

The engineering and optimization of marketing strategies is a conceptually more complex problem, based on the ability to make forecasts under different sets of conditions and to generate business models. We will explore the approach of targeting marketing actions such as direct mail and to generate global business models.

Targeting direct mail is a case in point. Direct mail campaigns are expensive to produce, so it is important to limit the actual mailing to those individuals most likely to buy. This involves establishing linkages to find subsets of potential clients most likely to respond. The same concept can be applied to any marketing action that is being performed on large sets of potential clients where it is clearly important to reduce the target to the individuals that are most likely to respond favorably.

Targeting direct mail has been achieved through the use of classification trees and of neural networks. As remarked in the previous section, targeting direct mail is the problem of optimizing from a distribution of the probability of answers. A pioneer of these techniques is the Cambridge, MA based Epsilon Data Management. Epsilon handles direct mail campaigns for some of America's biggest direct mailers, including American Express. Some years ago, Epsilon started using a Thinking Machines parallel computer to analyze the vast amounts of data from American Express in targeting direct mail campaigns. Later bought by American Express itself, Epsilon is now developing a suite of applications based on adaptive methodologies for the retail market.

In Europe, SearchSpace has developed an application based on neural networks to target direct mail campaigns for the same high-street retailer mentioned above. Retailers as Marks & Spencer, which has to date relied on traditional statistical methods as SAS, are now evaluating other techniques as neural networks to establish linkages for marketing purposes. Thanks to the company's 3.7m proprietary card holders, the retailer has data on purchases and can, for example, mine its database for those who, say, purchased a suit two years ago as likely candidates for a direct mail campaign aimed at attracting customers on a suit promotion.

3.3.3 Marketing Focalization

Marketing focalization is the broad problem of recognizing specific market segments that respond to particular characteristics. This problem arises when one wants to implement specific marketing strategies for narrow market segments, or when it is desirable to avoid negative events, such as credit-card frauds or non-renewal of subscriptions.

Conceptually the problem is one of supervised learning: generally a set of past examples is available and the problem is to generalize to new events. The technical problem is to make the generalization sharp and effective. Take for instance the problem of credit card frauds. Clearly is not desirable to stop service to clients that are not likely to default, but at the same time it is vital to avoid expensive frauds.

So although conceptually similar to the problem of targeting direct mail, the technological challenge is different, as it is the sharpness of generalization that matters. A number of techniques are being used, among which neural networks and identification trees.

An important research effort is under way in this sector at the Los Alamos National Laboratory as described in [32]. A consortium of banks including Citicorp, Chase Manhattan Bank and others, is financing a research effort at a number of national laboratories (including Los Alamos, Oak Ridge, Sandia, Lawrence Livermore) to develop techniques of data mining for the banking sector. At Los Alamos, the problem of fraud detection related to credit cards is being tackled with an internally developed neural network using the techniques of radial basis functions.

3.3.4 Customer Profiling

Customer base segmentation is an important tool to engineer marketing strategies or product offerings according to definite types of customer behavior. The objective of segmenting the market is to exploit differences in the demand curve of different market segments.

Conceptually the problem of market segmentation can be cast in the different frameworks of supervised or un-supervised learning. If a similarity measure and segmentation examples are known, clustering can be achieved with k-nearest neighbors type of techniques. In this case, there are sample clusters of customers and all the other clients are included in the cluster that includes the largest number of their k-nearest neighbors. This is a supervised approach to clustering useful when market classification is known on some examples.

If no classification is known but a good measure of similarity between customers exists, clusters are formed maximizing the degree of similarity between customers in a same partition. One technique consists in measuring the global similarity of a partition and using optimization strategies such as GAs to maximize the global similarity of a partition. A different approach consists in imposing additional constraints on similarity so that relations become transitive and equivalence classes can be built. Standard statistical techniques, such as principal components analysis, can also be used.

The IBM research center in Paris has developed a proprietary technology for customer segmentation, as described in [33]. Neurotec has developed an application for a large German insurance company using GAs to optimize the clustering of clients into different tariff classes. In the UK, the consumer bank Abbey National has purchased a 15 million $ Teradata system for customer profiling and direct mail campaigns. The bank's objective is to bring all customer data into a single relational database in an effort to better understand their customers and how their products perform. Ultimately, the bank's objective is to improve their product mix and their bottom line through offers targeted at the most profitable businesses. While the bank is now using traditional techniques for customer profiling, they are looking at other techniques as neural networks. An often cited obstacle to the introduction of these new methodologies remains the difficulty of building a good business case for them.

3.3.5 Sensitivity and Elasticity Analysis

Sensitivity and elasticity analysis is a very important tool for optimizing marketing strategies. Sensitivity analysis is the evaluation of the size of the change of a variable in function of another one. An example is the price elasticity of demand, i.e., the curve that says how much the demand changes for a change in the price of some item. The same concept can be applied to any other business variable, depending on one or more variable. More generally, it is very important to understand quantitatively the response of clients to specific marketing actions, such as advertising, promotions, and so on.

If a lot of historical data are available, computing sensitivities is a problem of generalization of a function from a number of sample values. Conceptually, it can be seen rather naturally as a problem of function approximation and generalization; it can also be regarded as a problem of finding a set of prediction rules. The latter formulation is much less natural from a mathematical point of view but it has the advantage of being more understandable by humans.

There are a number of business modeling products that offer the ability to compute sensitivities making use of neural-network techniques, among them *4-Thought* of Right Information Systems in London and *Quantum Leap* from Quantum Development Corporation in Wilmington, DE.

3.4 Financial Applications

The use of data-mining techniques in the financial sector falls into two basic sectors: consumer banking, with applications similar to those of retailing as we have seen in the previous chapter, and financial forecasting. Financial forecasting is a sort of data mining, as forecasts are based on exploring large sets of past data. However, the technologies used are quite specific and the underlying economic theory quite important: we will touch on them only briefly.

The reason for the peculiarities of financial forecasting are the limited level of forecastability of financial time series. In fact, there is still a lively debate if financial time series are forecastable at all and to what extent new sophisticated methods can really be beneficial.

Most applications of data mining in this report are relative to sets of data where patterns are clear although complex and difficult to find. We have already seen two significant areas, chemico/pharmaceutical where data to be mined can, in principle, be quite accurately described and the process of mining is a computational tool, and retail where regularities are believed to be quite strong and stable though complex.

Financial applications of data mining, on the other hand, deal with financial series that are highly volatile. The application of data mining depends on a good deal of sophisticated theory on how it is possible to extract a limited amount of forecastability from otherwise random data. As it is beyond the scope of this report to discuss this theory, we touch only marginally the problem of financial time series. Most applications in financial forecasting use one or more of the following three basic methods: chaos theory, neural networks, statistical methods. The literature on financial applications of data mining is very large. Dr. Paul Refenes, one of the founders of the NeuroForecasting Club in London provides in [34] a brief introduction to some of the major problems of data mining in financial applications.

3.5 Remotely Sensed Data

Remotely sensed data are probably the single largest source of data that is accumulated daily. A large fraction of data come from satellite images; other data are provided by earth and sea sensors, and from radar. In addition, data on the atmosphere are accumulated from a variety of sources, from balloons to airplanes.

Historically, the primary motivation around this vast effort of collecting data was military surveillance. Military surveillance includes not only the inspection of satellite images that provide a rather detailed picture of what is going on in a potentially hostile country from a military point of view, but also a number of data on other parameters that might be of military interest.

More recently, remotely sensed data have been used to understand and eventually control environmental situations. This might include monitoring phenomena such as ozone layer depletion, ice cap contraction, desertification of fertile areas, but also local phenomena such as floods, changes in vegetation and similar. Presently, remotely sensed data are also considered a commercial product to be sold to entities potentially interested in their exploitation.

In principle, exploring remotely sensed data is an ideal area for the use of data-mining techniques as data are abundant and most phenomena stable and identifiable, although potentially complex. In fact, it is likely that many military applications make use of these technologies.

Most existing non classified applications of automatic analysis of remotely sensed data are in the area of pattern recognition, using fairly standard statistical recognition methods. This is understandable as the primary task is surveillance. However there are a number of applications that make use of machine-learning techniques, in particular in the area of pattern recognition and forecasting.

Neural networks are currently used as recognition tools. The Los Alamos National Laboratory has developed an application for automatic recognition of moving objects from satellite images [32]. This application makes use of an internally developed CLSN-type network that is trained to recognize special shapes in a very noisy environment. TICA Technologies, a Cambridge, MA company, has developed an application to recognize sonar patterns, using a neural network optimized by genetic algorithms. Neurotec has developed an application for the European Joint Research Center at Ispra, Italy, for the prediction of up-welling. Basically this application uses a neural network to predict the temperature of the water under conditions that are particularly suitable for fishing. EOS, Earth Observation Science, a UK-based company, has developed a number of applications for analyzing and recognizing a broad variety of natural phenomena.

These applications are representative of machine learning in remotely sensed data. It is possible to conceive a much broader spectrum of usage of data-mining techniques. Clearly forecasting environmental phenomena could witness the wide use of neural networks.

3.6 Engineering and Maintenance Data

The use of data-mining techniques is very important not in the engineering process itself but more in functions such as machine tuning, analysis of maintenance data, quality control.

The problem of machine tuning is very delicate for certain types of equipment that it is not possible or practical to test exhaustively. Often, in such cases, experienced technicians are able to fine tune equipment but the tuning is a long and expensive process that does not seem to follow any easily recognizable rule. In these cases, a large set of data on past successful tuning is accumulated and the problem is to extract meaningful rules from this set of data.

One solution is the use of neural networks to reconstruct the tuning surface. This was done at Los Alamos for the fine tuning of a negative ion source. Tuning the equipment was a time consuming activity that required a team of experienced technicians. Based on past data a CNSN neural network was able to reconstruct the surface that gives performance in function of parameter settings. After reconstructing the tuning surface an optimization procedure was able to fine tune the ion source. See [32].

Another important application was done at Westinghouse for the fine tuning of a nuclear plant. In that case, there was a history of past tuning cases but it was clearly not practical and even dangerous to run experiments. A rule-induction system was able to mine the past data to find rules for the optimal tuning of the plant. This application is described in [35].

4 Acknowledgement

Sect. 3 of this report is based on interviews with persons from the chemical and pharmaceutical sectors, retail, banking, and finance, the research community, government, and hardware, software, and service suppliers. Our objective was to give a broad overview of how data mining is used to make sense of large data sets in areas as divers as drug design and financial forecasting, retail and natural phenomena. The organizations interviewed reflect the breadth of areas scanned - in the Unites States and in Europe - and the variety of experience and evaluation.

We wish to thank persons from the organizations listed below who contributed to this report by sharing with us their experience and their views:

Academia: Carnegie Mellon University, Erasmus University Rotterdam, London Business School, Los Alamos National Laboratory, The Sante Fe Institute, Swiss Federal Institute of Technology Zurich, University College London, Wharton School of Economics.

Hardware, software, and service suppliers: AT & T Global Information Systems, DayLight Chemical Information Systems, Epsilon Data Management, IBM Power Parallel Systems, IBM Scientific Center Paris, InferOne-NeuroAgent, A.C. Nielsen, NeuralWare, Neurotec, The Prediction Company, Quantum Development, SearchSpace, Siemens-Nixdorf, Thinking Machines, Tica Technologies.

Chemical and pharmaceutical sector: Akzo Pharma, Glaxo, Pfizer, Sandoz.

Retail: Epsilon/American Express, Marks & Spencer.

Banking: Abbey National, Credit Lyonnais, ING Bank.

Remotely-sensed data: Earth Observation Science, Joint Research Center Ispra.

References

[1] J. Laird, A. Newell, and P. Rosenbloom. Soar: An Architecture for General Intelligence. *Artificial Intelligence*, 33(1):1–64, 1987.

[2] D. Rumelhart and J. McClelland. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, MA, 1986.

[3] Patrick Henry Winston. *Artificial Intelligence.* Addison Wesley, Reading, MA, 1992.

[4] Herbert Simon. *Models of Discovery.* D. Reidel, Dordrecht, Holland, 1977.

[5] C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structures.* Academic Press, Orlando, FL, 1987.

[6] J. L. Breeden and N. H. Packard. A Learning Algorithm for Optimal Representations of Experimental Data. Technical Report CCSR-92-11, University of Illinois Urbana-Champaign, 1992.

[7] T. P. Meyer and N. H. Packard. Local Forecasting of High Dimensional Chaotic Dynamics. Technical Report CCSR-91-1, University of Illinois, Urbana-Champaign, 1991.

[8] S. M. Weiss and C. A. Kulikowski. *Computer Systems That Learn.* Morgan Kaufmann, San Mateo, CA, 1991.

[9] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, MA, 1992.

[10] J. Quinlan. Induction of Decision Trees. *Machine Learning*, 1:81–106, 1986.

[11] Y. S. Abu-Mostafa. The Vapnik-Chervonenkis Dimension: Information versus Complexity in Learning. *Neural Computation*, 1:312–317, 1989.

[12] G. J. Chaitin. *Algorithmic Information Theory.* Cambridge University Press, Cambridge, 1990.

[13] J. Rissanen. *Stochastic Complexity in Statistical Enquiry.* World Scientific Publication Corp., New Jersey, 1989.

[14] *The Mathematics of Generalization*, 1994.

[15] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal Statistical Physics*, 65(3/4):579–616, 1991.

[16] Andreas S. Weigend and Neil A. Gershenfeld, editors. *Time Series Prediction: Forecasting the Future and Understanding the Past.* Addison Welsey, Reading, MA, 1994.

[17] A. M. Fraser. Reconstructing attractors from scalar time series. *Physics D*, 34:391–404, 1989.

[18] A. S. Weigend, B. A. Huberman, and D. E. Rumelhart. Predicting the Future: a Connectionist Approach. *Intl. J. Neur. Sys.*, 1:193–209, 1990.

[19] John Hertz, Anders Krogh, and Richard Palmer. *Introduction to the Theory of Neural Computation.* Addison Wesley, Redwood, CA, 1991.

[20] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Monterey, CA, Wadsworth, 1984.

[21] D. H. Wolpert. Stacked Generalization. *Neural Networks*, 5:241–259, 1992.

[22] Suran Goonatilake and Sukhdev Khebbal. Intelligent Hybrid Systems. In *Proceedings of the First Singapore International Conference on Intelligent Systems, Singapore*, September 1992.

[23] Dirk Emma Baestaens and Willem van den Bergh. Qualitative Credit Assessment Using a Neural Classifier. In A. P. Refenes, editor, *Proceedings of the First International Workshop on Neural Networks in the Capital Markets*, London, November 1993.

[24] T. Kohonen. *Self-Organization and Associative Memory.* Springer, Berlin, 1989.

[25] N. H. Packard. A Genetic Learning Algorithm for the Analysis of Complex Data. *Complex Systems*, 4:543–572, 1990.

[26] European Commission. Exprit IT R&D Programme: High-Performance Computing and Networking, Summaries of the Esprit Projects. Technical Report EUR 15375 en/4, European Commission, Luxembourg, 1994.

[27] Real-World Computing Partnership. Introduction to Real-World Computing Program. Technical Report, Machinery and Information Industries Bureau and MITI, Tokyo, June 1994.

[28] Y. C. Martin, E. B. Danaher, C. S. May, and D. Weininger. MENTHOR, A Data Base System for the Storage and Retrieval of 3-Dimensional Molecular Structures and Associated Data Searchable by Sub-Structural, Biological, Physical, or Geometric Properties. *Journal Computer-Aided Molecular Design*, 2:15–29, 1988.

[29] J. H. Van Drie, D. Weininger, and Y. C. Martin. ALADDIN: An Integrated Tool for Computer-assisted Molecular Design and Pharmacophore Recognition from Geometric, Steric, and Substructural Searching of 3-Dimensional Molecular Structures. *Journal Computer-Aided Molecular Design*, 3:225–251, 1989.

[30] R. A. Jarvis and E. A. Patrick. Clustering Using a Similarity Measure Based on Shared Near Neighbors. *IEEE Transactions on Computers*, C22:1025–1034, 1973.

[31] Tad Hurst. Flexible 3-D Searching: The Directed Tweak Technique. Technical Paper, Tripos Associates, St. Louis, MO, 1994.

[32] Roger D. Jones. Nonlinear Adaptive Computation. Technical Report, Los Alamos, 1993.

[33] Pierre Michaud. Aggrégation à la Majorité II: Analyse du Résultat d'une Vote. Technical Report Etude F.052, IBM Centre Scientifique de Paris, May 1985.

[34] A. N. Refenes and M. Azema-Barac. Neural Network Applications in Financial Asset Management. *Neural Computing and Applications*, 2:13–39, 1994.

[35] W. J. Leech. A Rule Based Process Control Method With Feedback. In *Proceedings of the International Conference and Exhibit, Instrument Society of America*, 1986.

[36] Donald Berry and Bernard Lindgren. *Statistics: Theory and Methods*. Brooks/Cole, Pacific Grove, CA, 1990.

[37] Patrick Billingsley. *Probability and Measure*. Wiley, New York, NY, 1986.

[38] Lawerence Davis. *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, NY, 1991.

[39] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, NY, 1973.

[40] B. Everitt. *Cluster Analysis*. Heinemann, London, 1980.

[41] D. Farmer and J. Sidorowich. Predicting Chaotic Time Series. *Physical Review Letters*, 59:845–848, 1987.

[42] V. W. Fisanick, K. Cross, J. Forman, and A. Rusinko III. Experimental System for Similarity and 3-D Searching of CAS Registry Substances. *Journal Chem. Inf. Computer Science*, 33:548–559, 1993.

[43] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York, NY, 1979.

[44] C. W. J. Granger and P. Newbold. *Forecasting Economic Time Series*. Academic Press, New York, NY, 1986.

[45] M. James. *Classification Algorithms*. Wiley, New York, NY, 1985.

[46] Steven O. Kimbrough and J. R. Oliver. On Automating Candle Lighting Analysis: Insight From Search with Genetic Algorithms and Approximate Models. In *Proceedings of the 27th Hawaii International Conference on Systems Sciences*, January 1994.

[47] Kenneth E. Kinnear, editor. *Advances in Genetic Programming.* MIT Press, Cambridge, MA, 1994.

[48] J. R. Koza. *Genetic Programming.* MIT Press, Cambridge, MA, 1993.

[49] Zbigniew Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer, Berlin, 1992.

[50] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[51] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, CA, 1988.

[52] Heinz Mühlenbein and Dirk Schlierkamp-Voosen. The Science of Breeding and its Application to the Breeder Genetic Algorithm. Technical Paper, Project Sifoga, Real-World Computing Program, GMD, 1994.

[53] Lynn Nadel and Daniel L. Stein, editors. *1992 Lectures in Complex Systems.* Addison Wesley, Reading, MA, 1993.

[54] Real-World Computing Program. The Master Plan for the Real-World Computing Program. Technical Report, Machinery and Information Industries Bureau and MITI, Tokyo, May 1992.

[55] Ian Stewart. *Does God Play Dice ?, The Mathematics of Chaos.* Blackwell, Cambridge, MA, 1989.

[56] J. M. T. Thompson and H. B. Stewart. *Nonlinear Dynamics and Chaos.* Wiley, New York, NY, 1986.

[57] W. H. Zurek, editor. *Complexity, Entropy and the Physics of Information.* Addison Wesley, Reading, MA, 1990.