# 1

# Overview of data mining



This first chapter defines data mining and sets out its main applications and contributions to database marketing, customer relationship management and other financial, industrial, medical and scientific fields. It also considers the position of data mining in relation to statistics, which provides it with many of its methods and theoretical concepts, and in relation to information technology, which provides the raw material (data), the computing resources and the communication channels (the output of the results) to other computer applications and to the users. We will also look at the legal constraints on personal data processing; these constraints have been established to protect the individual liberties of people whose data are being processed. The chapter concludes with an outline of the main factors in the success of a project.

## 1.1 What is data mining?

Data mining and statistics, formerly confined to the fields of laboratory research, clinical trials, actuarial studies and risk analysis, are now spreading to numerous areas of investigation, ranging from the infinitely small (genomics) to the infinitely large (astrophysics), from the most general (customer relationship management) to the most specialized (assistance to pilots in aviation), from the most open (e-commerce) to the most secret (prevention of terrorism, fraud detection in mobile telephony and bank card applications), from the most practical (quality control, production management) to the most theoretical (human sciences, biology, medicine and pharmacology), and from the most basic (agricultural and food science) to the most entertaining (audience prediction for television). From this list alone, it is clear that the applications of data mining and statistics cover a very wide spectrum. The most relevant fields are those where large volumes of data have to be analysed, sometimes with the aim of rapid decision making, as in the case of some of the examples given above. Decision assistance is becoming an objective of data mining and statistics; we now expect these techniques to do more than simply provide a model of reality to help us to understand it. This approach is not completely new, and is already established in medicine, where some treatments have been developed on the basis of statistical analysis, even though the biological mechanism of the disease is little understood because of its

*Data Mining and Statistics for Decision Making*, First Edition. Stéphane Tufféry.

complexity, as in the case of some cancers. Data mining enables us to limit human subjectivity in decision-making processes, and to handle large numbers of files with increasing speed, thanks to the growing power of computers.

A survey on the www.kdnuggets.com portal in July 2005 revealed the main fields where data mining is used: banking (12%), customer relationship management (12%), direct marketing (8%), fraud detection (7%), insurance (6%), retail (6%), telecommunications (5%), scientific research (4%), and health (4%).

In view of the number of economic and commercial applications of data mining, let us look more closely at its contribution to 'customer relationship management'.

In today's world, the wealth of a business is to be found in its customers (and its employees, of course). Customer share has replaced market share. Leading businesses have been valued in terms of their customer file, on the basis that each customer is worth a certain (large) amount of euros or dollars. In this context, understanding the expectations of customers and anticipating their needs becomes a major objective of many businesses that wish to increase profitability and customer loyalty while controlling risk and using the right channels to sell the right product at the right time. To achieve this, control of the information provided by customers, or information about them held by the company, is fundamental. This is the aim of what is known as customer relationship management (CRM). CRM is composed of two main elements: operational CRM and analytical CRM.

The aim of analytical CRM is to extract, store, analyse and output the relevant information to provide a comprehensive, integrated view of the customer in the business, in order to understand his profile and needs more fully. The raw material of analytical CRM is the data, and its components are the data warehouse, the data mart, multidimensional analysis (online analytical processing[1]), data mining and reporting tools.

For its part, operational CRM is concerned with managing the various channels (sales force, call centres, voice servers, interactive terminals, mobile telephones, Internet, etc.) and marketing campaigns for the best implementation of the strategies identified by the analytical CRM. Operational CRM tools are increasingly being interfaced with back office applications, integrated management software, and tools for managing workflow, agendas and business alerts. Operational CRM is based on the results of analytical CRM, but it also supplies analytical CRM with data for analysis. Thus there is a data 'loop' between operational and analytical CRM (see Figure 1.1), reinforced by the fact that the multiplication of communication channels means that customer information of increasing richness and complexity has to be captured and analysed.

The increase in surveys and technical advances make it necessary to store ever-greater amounts of data to meet the operational requirements of everyday management, and the global view of the customer can be lost as a result. There is an explosive growth of reports and charts, but 'too much information means no information', and we find that we have less and less knowledge of our customers. The aim of data mining is to help us to make the most of this complexity.

It makes use of databases, or, increasingly, data warehouses,[2] which store the profile of each customer, in other words the totality of his characteristics, and the totality of his past and

[1] Data storage in a cube with $n$ dimensions (a 'hypercube') in which all the intersections are calculated in advance, so as to provide a very rapid response to questions relating to several axes, such as the turnover by type of customer and by product line.

[2] A *data warehouse* is a set of databases with suitable properties for decision making: the data are thematic, consolidated from different production information systems, user-oriented, non-volatile, documented and possibly aggregated.

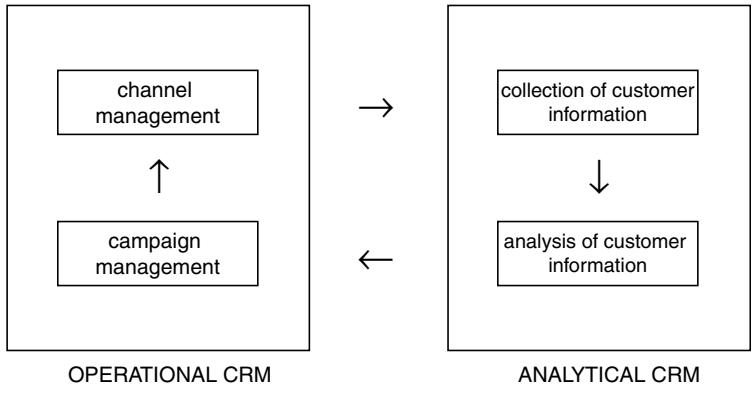


**Figure 1.1** The customer relationship circuit.

present agreements and exchanges with the business. This global and historical knowledge of each customer enables the business to consider an individual approach, or 'one-to-one marketing',[3] as in the case of a corner shop owner 'who knows his customers and always offers them what suits them best'. The aim of this approach is to improve the customer's satisfaction, and consequently his loyalty, which is important because it is more expensive (by a factor of 3–10) to acquire a new customer than to retain an old one, and the development of consumer comparison skills has led to a faster customer turnover. The importance of customer loyalty can be appreciated if we consider that an average supermarket customer spends about €200 000 in his lifetime, and is therefore 'potentially' worth €200 000 to a major retailer.

Knowledge of the customer is even more useful in the service industries, where products are similar from one establishment to the next (banking and insurance products cannot be patented), where the price is not always the decisive factor for a customer, and customer relations and service make all the difference.

However, if each customer were considered to be a unique case whose behaviour was irreducible to any model, he would be entirely unpredictable, and it would be impossible to establish any proactive relationship with him, in other words to offer him whatever may interest him at the time when he is likely to be interested, rather than anything else. We may therefore legitimately wish to compare the behaviour of a customer whom we know less well (for a first credit application, for example) with the behaviour of customers whom we know better (those who have already repaid a loan). To do this, we need two types of data. First of all, we need 'customer' data which tell us whether or not two customers resemble each other. Secondly, we need data relating to the phenomenon to be predicted, which may be, for example, the results of early commercial activities (for what are known as propensity scores) or records of incidents of payment and other events (for risk scores). A major part of data mining is concerned with modelling the past in order to predict the future: we wish to find rules concealed in the vast body of data held on former customers, in order to apply them to new customers and take the best possible decisions. Clearly, everything I have said about the customers of a business is equally applicable to bacterial strains in a laboratory, types of

[3] Or, more modestly and realistically, 'one-to-few'.

fertilizer in a plantation, chemical molecules in a test tube, patients in a hospital, bolts on an assembly line, etc. So the essence of data mining is as follows:

> Data mining is the set of methods and techniques for exploring and analysing data sets (which are often large), in an automatic or semi-automatic way, in order to find among these data certain unknown or hidden rules, associations or tendencies; special systems output the essentials of the useful information while reducing the quantity of data.
>
> Briefly, data mining is the art of extracting information – that is, knowledge – from data.
>
> Data mining is therefore both descriptive and predictive: the descriptive (or exploratory) techniques are designed to *bring out information that is present* but buried in a mass of data (as in the case of automatic clustering of individuals and searches for associations between products or medicines), while the predictive (or explanatory) techniques are designed to *extrapolate new information based on the present information*, this new information being qualitative (in the form of classification or scoring[4]) or quantitative (regression).

The rules to be found are of the following kind:

- Customers with a given profile are most likely to buy a given product type.
- Customers with a given profile are more likely to be involved in legal disputes.
- People buying disposable nappies in a supermarket after 6 p.m. also tend to buy beer (a example which is mythical as well as apocryphal).
- Customers who have bought product A and product B are most likely to buy product C at the same time or *n* months later.
- Customers who have behaved in a given way and bought given products in a given time interval may leave us for the competition.

This can be seen in the last two examples: we need a history of the data, a kind of moving picture, rather than a still photograph, of each customer. All these examples also show that data mining is a key element in CRM and one-to-one marketing (see Table 1.1).

## 1.2 What is data mining used for?

Many benefits are gained by using rules and models discovered with the aid of data mining, in numerous fields.

### 1.2.1 Data mining in different sectors

It was in the *banking sector* that risk scoring was first developed in the mid-twentieth century, at a time when computing resources were still in their infancy. Since then, many data mining techniques (scoring, clustering, association rules, etc.) have become established in both retail and commercial banking, but data mining is especially suitable for retail banking because of

[4] The statistical technique is called 'classification' or 'discrimination'; the application of this technique to certain business problems such as the selection of customers according to certain criteria is called 'scoring'.

**Table 1.1** Comparison between traditional and one-to-one marketing.

| Traditional marketing | One-to-one marketing |
|---|---|
| Anonymous customer | Individually identified customer |
| Standard product | Personalized product and service |
| Serial production | Bespoke production |
| Mass advertising | Individual message |
| Unilateral communication | Interactive communication |
| Achievement of a sale, high take-up | Development of customer loyalty, low attrition rate |
| Market share | Customer share |
| Broad targets | Profitable niches |
| Segmentation by job and RFM | Statistical, behavioural segmentation |
| Traditional distribution channels, disconnected from each other | New, interconnected channels (telephone platforms, Internet, mobile telephones) |
| Product-oriented marketing | Customer-oriented marketing |

the moderate unitary amounts, the large number of files and their relatively standard form. The problems of scoring are generally not very complicated in theoretical terms, and the conventional techniques of discriminant analysis and logistic regression have been extremely successful here. This expansion of data mining in banking can be explained by the simultaneous operation of several factors, namely the development of new communication technology (Internet, mobile telephones, etc.) and data processing systems (data warehouses); customers' increased expectations of service quality; the competitive challenge faced by retail banks from credit companies and 'newcomers' such as foreign banks, major retailers and insurance companies, which may develop banking activities in partnership with traditional banks; the international economic pressure for higher profitability and productivity; and of course the legal framework, including the current major banking legislation to reform the solvency ratio (see Section 12.2), which has been a strong impetus to the development of risk models. In banks, loyalty development and attrition scoring have not been developed to the same extent as in mobile telephones, for instance, but they are beginning to be important as awareness grows of the potential profits to be gained. For a time, they were also stimulated by the competition of on-line banks, but these businesses, which had lower structural costs but higher acquisition costs than branch-based banks, did not achieve the results expected, and have been bought up by insurance companies wishing to gain a foothold in banking, by foreign banks, or by branch-based banks aiming to supplement their multiple-channel banking system, with Internet facilities coexisting with, but not replacing, the traditional channels.

The *retail industry* is developing its own credit cards, enabling it to establish very large databases (of several million cardholders in some cases), enriched by behavioural information obtained from till receipts, and enabling it to compete with the banks in terms of customer knowledge. The services associated with these cards (dedicated check-outs, exclusive promotions, etc.) are also factors in developing loyalty. By detecting product associations on till receipts it is possible to identify customer profiles, make a better choice of products and arrange them more appropriately on the shelves, taking the 'regional' factor into account in

the analyses. The most interesting results are obtained when payments are made with a loyalty card, not only because this makes it possible to cross-check the associations detected on the till receipts with sociodemographic information (age, family circumstances, socio-occupational category) provided by the customer when he joins the card scheme, but also because the use of the card makes it possible to monitor a customer's payments over time and to implement customer-targeted promotions, approaching the customer according to the time intervals and themes suggested by the model. Market baskets can also be segmented into groups such as 'clothing receipt', 'large trolley receipt', and the like.

In *property and personal insurance*, studies of 'cross-selling', 'up-selling' and attrition, with the adaptation of pricing to the risks incurred, are the main themes in a sector where propensity is not stated in the same terms as elsewhere, since certain products (motor insurance) are compulsory, and, except in the case of young people, the aim is either to attract customers from competitors, or to persuade existing customers to upgrade, by selling them additional optional cover, for example. The need for data mining in this sector has increased with the development of competition from new entrants in the form of banks offering what is known as 'bancassurance' (bank insurance), with the advantage of extended networks, frequent customer contact and rich databases. The advantages of this offer are especially great in comparison with 'traditional' non-mutual insurance companies which may encounter difficulties in developing marketing databases from information which is widely diffused and jealously guarded by their agents. Furthermore, the customer bases of these insurers, even if not divided by agent, are often structured according to contracts rather than customers. And yet these networks, with their lower loyalty rates than mutual organizations, have a real need to improve their CRM, and consequently their global knowledge of their customers. Although the propensity studies for insurance are similar to those for banking, the loss studies show some distinctive features, with the appearance of the Poisson distribution in the generalized linear model for modelling the number of claims (loss events). The insurers have one major asset in their holdings of fairly comprehensive data about their customers, especially in the form of home and civil liability insurance contracts which provide fairly accurate information on the family and its lifestyle.

The opening of the *landline telephone* market to European competition, and the development of the *mobile telephone* market through maturity to saturation, have revived the problems of 'churning' (switching to competing services) among private, professional and business customers. The importance of loyalty in this sector becomes evident when we consider that the average customer acquisition cost in the mobile telephone market is more than €200, and that more than a million users change their operator every year in some countries. Naturally, therefore, it is churn scoring that is the main application of data mining in the telephone business. For the same reasons, operators use *text mining* tools (see Chapter 14) for automatic analysis of the content of customers' letters of complaint. Other areas of investigation in the telephone industry are non-payment scoring, direct marketing optimization, behavioural analysis of Internet users and the design of call centres. The probability of a customer changing his mobile telephone is also under investigation.

Data mining is also quite widespread in the *motor industry.* A standard theme is scoring for repeat purchases of a manufacturer's vehicles. Thus, Renault has constructed a model which predicts customers who are likely to buy a new Renault car in the next six months. These customers are identified on the basis of data from concessionaires, who receive in return a list of high-scoring customers whom they can then contact. In the production area, data mining is used to trace the origin of faults in construction, so that these can be minimized. Satisfaction

studies are also carried out, based on surveys of customers, with the aim of improving the design of vehicles (in terms of quality, comfort, etc.). Accidents are investigated in the laboratories of motor manufacturers, so that they can be classified in standard profiles and their causes can be identified. A large quantity of data is analysed, relating to the vehicle, the driver and the external circumstances (road condition, traffic, time, weather, etc.).

The *mail-order* sector has been conducting analyses of data on its customers for many years, with the aim of optimizing targeting and reducing costs, which may be very considerable when a thousand-page colour catalogue is sent to several tens of millions of customers. Whereas banking was responsible for developing risk scoring, the mail-order industry was one of the first sectors to use propensity scoring.

The *medical sector* has traditionally been a heavy user of statistics. Quite naturally, data mining has blossomed in this field, in both diagnostic and predictive applications. The first category includes the identification of patient groups suitable for specific treatment protocols, where each group includes all the patients who react in the same way. There are also studies of the associations between medicines, with the aim of detecting prescription anomalies, for example. Predictive applications include tracing the factors responsible for death or survival in certain diseases (heart attacks, cancer, etc.) on the basis of data collected in clinical trials, with the aim of finding the most appropriate treatment to match the pathology and the individual. Of course, use is made of the predictive method known as *survival analysis*, where the variable to be predicted is a period of time. Survival data are said to be 'censored', since the period is precisely known for individuals who have died, while it is only the minimum survival time that is known for those who remain. We can, for example, try to predict the recovery time after an operation, according to data on the patient (age, weight, height, smoker or non-smoker, occupation, medical history, etc.) and the practitioner (number of operations carried out, years of experience, etc.). *Image mining* is used in medical imaging for the automatic detection of abnormal scans or tumour recognition. Finally, the deciphering of the genome is based on major statistical research for detecting, for example, the effect of certain genes on the appearance of certain pathologies. These statistical analyses are difficult, as the number of explanatory variables is very high with respect to the number of observations: there may be several tens of millions of genes (genome) or pixels (image mining) relating to only a few hundred individuals. Methods such as partial least squares (PLS) regression or regularized regression (ridge, lasso) are highly valued in this field. The tracing of similar sequences ('sequence analysis') is widely used in genomics, where the DNA sequence of a gene is investigated with the aim of finding similarities between the sequences of a single ancestor which have undergone mutations and natural selection. The similarity of biological functions is deduced from the similarity of the sequences.

In cosmetics, Unilever has used data mining to predict the effect of new products on human skin, thus limiting the number of tests on animals, and L'Oréal, for example, has used it to predict the effects of a lotion on the scalp.

The *food industry* is also a major user of statistics. Applications include 'sensory analysis' in which sensory data (taste, flavour, consistency, etc.) perceived by consumers are correlated with physical and chemical instrumental measurements and with preferences for various products. Discriminant analysis and logistic regression predictive models are also used in the drinks industry to distinguish spirits from counterfeit products, based on the analysis of about ten molecules present in the beverage. Chemometrics is the extraction of information from physical measurements and from data collected in analytical chemistry. As in genomics, the number of explanatory variables soon becomes very great and may justify the use of PLS

regression. Health risk analysis is specific to the food industry: it is concerned with understanding and controlling the development of microorganisms, preventing hazards associated with their development in the food industry, and managing use-by dates. Finally, as in all industries, it is essential to manage processes as well as possible in order to improve the quality of products.

Statistics are widely used in *biology*. They have been applied for many years for the classification of living species; we may, for example, quote the standard example of Fisher's use of his linear discriminant analysis to classify three species of iris. Agronomy requires statistics for an accurate evaluation of the effects of fertilizers or pesticides. Another currently fashionable use of data mining is for the detection of factors responsible for air pollution.

### 1.2.2 Data mining in different applications

In the field of customer relationship management, we can expect to gain the following benefits from statistics and data mining:

- identification of prospects most likely to become customers, or former customers most likely to return ('winback');
- calculation of profitability and *lifetime value* (see Section 4.2.2) of customers;
- identification of the most profitable customers, and concentration of marketing activities on them;
- identification of customers likely to leave for the competition, and marketing operations if these customers are profitable;
- better rate of response in marketing campaigns, leading to lower costs and less customer fatigue in respect of mailings;
- better cross-selling;
- personalization of the pages of the company website according to the profile of each user;
- commercial optimization of the company website, based on detection of the impact of each page;
- management of calls to the company's switchboard and direction to the correct support staff, according to the profile of the calling customer;
- choice of the best distribution channel;
- determination of the best locations for bank or major store branches, based on the determination of store profiles as a function of their location and the turnover generated by the different departments;
- in the retail industry, determination of consumer profiles, the 'market basket', the effect of sales or advertising; planning of more effective promotions, better prediction of demand to avoid stock shortages or unsold stock;

- telephone traffic forecasting;
- design of call centres;
- stimulating the reuse of a telephone card in a closely identified group of customers, by offering a reduction on three numbers of their choice;
- winning on-line customers for a telephone operator;
- analysis of customers' letters of complaint (using text data obtained by text mining – see Chapter 14);
- technology watching (use of text mining to analyse studies, specialist papers, patent filings, etc.);
- competitor monitoring.

In operational terms, the discovery of these rules enables the user to answer the questions 'who', 'what', 'when' and 'how' – who to sell to, what product to sell, when to sell it, how to reach the customer.

Perhaps the most typical application of data mining in CRM is propensity scoring, which measures the probability that a customer will be interested in a product or service, and which enables targeting to be refined in marketing campaigns. Why is propensity scoring so successful? While poorly targeted mailshots are relatively costly for a business, with the cost depending on the print quality and volume of mail, unproductive telephone calls are even more expensive (at least €5 per call). Moreover, when a customer has received several mailings that are irrelevant to him, he will not bother to open the next one, and may even have a poor image of the business, thinking that it pays no attention to its customers.

In *strategic marketing*, data mining can offer:

- help with the creation of packages and promotions;
- help with the design of new products;
- optimal pricing;
- a customer loyalty development policy;
- matching of marketing communications to each segment of the customer base;
- discovery of segments of the customer base;
- discovery of unexpected product associations;
- establishment of representative panels.

As a general rule, data mining is used to gain a better understanding of the customers, with a view to adapting the communications and sales strategy of the business.

In *risk management*, data mining is useful when dealing with the following matters:

- identifying the risk factors for claims in personal and property insurance, mainly motor and home insurance, in order to adapt the price structure;

- preventing non-payment of bills in the mobile telephone industry;
- assisting payment decisions in banks, for current accounts where overdrafts exceed the authorized limits;
- using the risk score to offer the most suitable credit limit for each customer in banks and specialist credit companies, or to refuse credit, depending on the probability of repayment according to the due dates and conditions specified in the contract;
- predicting customer behaviour when interest rates change (early credit repayment requests, for example);
- optimizing recovery and dispute procedures;
- automatic real-time fraud detection (for bank cards or telephone systems);
- detection of terrorist profiles at airports.

Automatic fraud detection can be used with a mobile phone which makes an unusually long call from or to a location outside the usual area. Real-time detection of doubtful bank transactions has enabled the Amazon on-line bookstore to reduce its fraud rate by 50% in 6 months. Chapter 12 will deal more fully with the use of risk scoring in banking.

A recent and unusual application of data mining is concerned with *judicial* risk. In the United Kingdom, the OASys (Offenders Assessment System) project aims to estimate the risk of repeat offending in cases of early release, using information on the family background, place of residence, educational level, associates, criminal record, social workers' reports and behaviour of the person concerned in custody and in prison. The British Home Secretary and social workers hope that OASys will standardize decisions on early release, which currently vary widely from one region to another, especially under the pressure of public opinion.

The *miscellaneous applications* of data mining and statistics include the following:

- road traffic forecasting, day by day or by hourly time slots;
- forecasting water or electricity consumption;
- determining whether a person owns or rents his home, when planning to offer insulation or installation of a heating system (Électricité de France);
- improving the quality of a telephone network (discovering why some calls are unsuccessful);
- quality control and tracing the causes of manufacturing defects, for example in the motor industry, or in companies such as the one which succeeded in explaining the sporadic appearance of defects in coils of steel, by analysing 12 parameters in 8000 coils during 30 days of production;
- use of survival analysis in industry, with the aim of predicting the life of a manufactured component;
- profiling of job seekers, in order to detect unemployed persons most at risk of long-term unemployment and provide prompt assistance tailored to their personal circumstances;

- pattern recognition in large volumes of data, for example in astrophysics, in order to classify a celestial object which has been newly discovered by telescope (the SKICAT system, applied to 40 measured characteristics);
- signal recognition in the military field, to distinguish real targets from false ones.

A rather more entertaining application of data mining relates to the prediction of the audience share of a television channel (BBC) for a new programme, according to the characteristics of the programme (genre, transmission time, duration, presenter, etc.), the programmes preceding and following it on the same channel, the programmes broadcast simultaneously on competing channels, the weather conditions, the time of year (season, holidays, etc.) and any major events or shows taking place at the same time. Based on a data log covering one year, a model was constructed with the aid of a neural network. It is able to predict audience share with an accuracy of $\pm 4\%$, making it as accurate as the best experts, but much faster.

Data mining can also be used for its own internal purposes, by helping to determine the reliability of the databases that it uses. If an anomaly is detected in a data element X, a variable 'abnormal data element X (yes/no)' is created, and the explanation for this new variable is then found by using a decision tree to test all the data except X.

## 1.3 Data mining and statistics

In the commercial field, the questions to be asked are not only 'how many customers have bought this product in this period?' but also 'what is their profile?', 'what other products are they interested in?' and 'when will they be interested?'. The profiles to be discovered are generally complex: we are not dealing with just the 'older/younger', 'men/women', 'urban/rural' categories, which we could guess at by glancing through descriptive statistics, but with more complicated combinations, in which the discriminant variables are not necessarily what we might have imagined at first, and could not be found by chance, especially in the case of rare behaviours or phenomena. This is true in all fields, not only the commercial sector. With data mining, we move on from 'confirmatory' to 'exploratory' analysis.[5]

Data mining methods are certainly more complex than those of elementary descriptive statistics. They are based on artificial intelligence tools (neural networks), information theory (decision trees), machine learning theory (see Section 11.3.3), and, above all, inferential statistics and 'conventional' data analysis including factor analysis, clustering and discriminant analysis, etc.

There is nothing particularly new about exploratory data analysis, even in its advanced forms such as multiple correspondence analysis, which originated in the work of theoreticians such as Jean-Paul Benzécri in the 1960s and 1970s and Harold Hotelling in the 1930s and 1940s (see Section A.1 in Appendix A). Linear discriminant analysis, still used as a scoring method, first emerged in 1936 in the work of Fisher. As for the evergreen logistic regression,

[5] In an article of 1962 and a book published in 1977, J.W. Tukey, the leading American statistician, contrasts *exploratory* data analysis, in which the data take priority, with *confirmatory* data analysis, in which the model takes priority. See Tukey, J.W. (1977) *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Pierre-François Verhulst anticipated this in 1838 and Joseph Berkson developed it from 1944 for biological applications.

The reasons why data mining has moved out of universities and research laboratories and into the world of business include, as we have seen, the pressures of competition and the new expectations of consumers, as well as regulatory requirements in some cases, such as pharmaceuticals (where medicines must be trialled before they are marketed), or banking (where the equity must be adjusted according to the amount of exposure and the level of risk incurred). This development has been made possible by three major technical advances.

The first of these concerns the storage and calculation capacity offered by modern computing equipment and methods: data warehouses with capacities of several tens of terabytes, massively parallel architectures, increasingly powerful computers.

The second advance is the increasing availability of 'packages' of different kinds of statistical and data mining algorithms in integrated software. These algorithms can be automatically linked to each other, with a user-friendliness, a quality of output and options for interactivity which were previously unimaginable.

The third advance is a step change in the field of decision making: this includes the use of data mining methods in production processes (where data analysis was traditionally used only for single-point studies), which may extend to the periodic output of information to end users (marketing staff, for example) and automatic event triggering.

These three advances have been joined by a fourth. This is the possibility of processing data of all kinds, including incomplete data (by using imputation methods), some aberrant data (by using 'robust' methods), and even text data (by using 'text mining'). Incomplete data – in other words, those with missing values – are found less commonly in science, where all the necessary data are usually measured, than in business, where not all the information about a customer is always known, either because the customer has not provided it, or because the salesman has not recorded it.

A fifth element has played a part in the development of data mining: this is the establishment of vast databases to meet the management requirements of businesses, followed by an awareness of the unexploited riches that these contain.

## 1.4 Data mining and information technology

An IT specialist will see a data mining model as an *IT application*, in other words a set of instructions written in a programming language to carry out certain processes, as follows:

- providing an output data element which summarizes the input data (e.g. a segment number);
- or providing an output data element of a new type, deduced from the input data and used for decision making (e.g. a score value).

As we have seen, the first of these processes corresponds to descriptive data mining, where the archetype is *clustering*: an individual's membership of a cluster is a summary of all of its *present* characteristics. The second example corresponds to predictive *data mining*, where the archetype is *scoring*: the new variable is a probability that the individual will behave in a certain way *in the future* (in respect of risk, consumption, loyalty, etc.).

Like all IT applications, a data mining application goes through a number of phases:

- development (construction of the model) in the decision-making environment;
- testing (verifying the performance of the model) in the decision-making environment;
- use in the production environment (application of the model to the production data to obtain the specified output data).

However, data mining has some distinctive features, as follows:

- The development phase cannot be completed in the absence of data, in contrast to an IT development which takes place according to a specification; the development of a model is primarily dependent on data (even if there is a specification as well).
- Development and testing are carried out in the same environment, with only the data sets differing from each other (as they must do!).
- To obtain an optimal model, it is both normal and necessary to move frequently between testing and development; some programs control these movements in a largely automatic way to avoid any loss of time.
- The data analysis for development and testing is carried out using a special-purpose program, usually designed by SAS, SPSS (IBM group), KXEN, Statistica or SPAD, or open source software (see Chapter 5).
- All these programs benefit from graphic interfaces for displaying results which justify the relevance of the developments and make them evident to users who are neither statisticians nor IT specialists.
- Some programs also offer the use of the model, which can be a realistic option if the program is implemented on a server (which can be done with the programs mentioned above).
- The conciseness of the data mining models: unlike the instructions of a computer program, which are often relatively numerous, the number of instructions in a data mining model is nearly always small (if we disregard the instructions for collecting the data to which the model is applied, since these are related to conventional data processing, even though there are special purpose tools), and indeed conciseness (or 'parsimony') is one of the sought-after qualities of a model (since it is considered to imply readability and robustness).

To some extent, the last two points are the inverse of each other. On the one hand, data mining models can be used in the same decision-making environment and with the same software as in the development phase, provided that the production data are transferred into this environment. On the other hand, the conciseness of the models means that they can be exported to a production environment that is different from the development environment, for example an IBM and DB2 mainframe environment, or Unix and Oracle. This solution may provide better performance than the first for the periodic processing of large bodies of data without the need for bulky transfers, or for calculating scores in real time (with inputting face to face with the customer), but it requires an export facility. The obvious advantage of the first

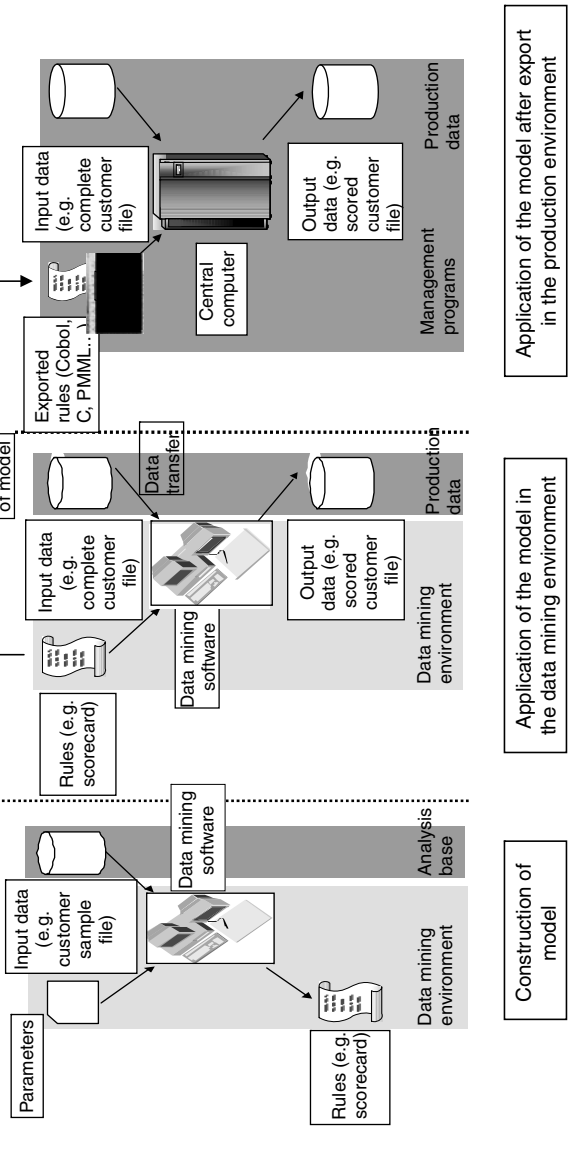


**Figure 1.2** IT architecture for data mining.

solution is a gain in time in the implementation of the data mining processes. In the first solution, the data lead to the model; in the second, the model leads to the data (see Figure 1.2).

Some models are easily exported and reprogrammed in any environment. These are purely statistical models, such as *discriminant analysis* and *logistic regression*, although the latter requires the presence of an exponential function or the power function at least (which, it should be noted, is provided even in Cobol). These standard models are concise and high-performing, provided that they are used with care. In particular, it is advisable to work with a few carefully chosen variables, and to apply these models to relatively homogeneous populations, provided that a preliminary segmentation is carried out.

Here is an example of a logistic regression model, which supplies the 'score' probability of being interested in purchasing a certain product. The ease of export of this type of model will be obvious.

```
logit = 0.985 - (0.005*variable_W) + (0.019* variable_X) +
(0.122* variable_Y) - (0.002* variable_Z);
score = exponential(logit) / [1 + exponential(logit)];
```

Such a model can also be converted to a scoring grid, as shown in Section 12.8.

Another very widespread type of model is the *decision tree*. These models are very popular because of their readability, although they are not the most robust, as we shall see.

A very simple example (Figure 1.3) again illustrates the propensity to buy a product. The aim is to extend the branches of the tree until we obtain terminal nodes or *leaves* (at the end of the branches, although the leaves are at the bottom here and the root, i.e. the total sample, is at the top) which contain the highest possible percentage of 'yes' (propensity to buy) or 'no' (no propensity to buy).

The algorithmic representation of the tree is a set of rules (Figure 1.4), where each rule corresponds to the path from the root to one of the leaves. As we can see in this very simple example, the model soon becomes less concise than a statistical model, especially as real trees often have at least four or five depth levels. Exporting would therefore be rather more difficult if it were a matter of copying the rules 'manually', but most programs offer options for automatic translation of the rules into C, Java, SQL, PMML, etc.

Some clustering models, such as those obtained by the *moving centres* method or variants of it, are also relatively easy to reprogram in different IT environments. Figure 1.5 shows an example of this, produced by SAS, for clustering a population described by six variables into three clusters. Clearly, this is a matter of calculating the Euclidean distance separating each individual from each of the three clusters, and assigning the individual to the cluster to which he is closest (where CLScads[_clus] reaches a minimum).

However, not all clustering models can be exported so easily. Similarly, models produced by neural networks do not have a simple synthetic expression. To enable any type of model to be exported to any type of hardware platform, a universal language based on XML was created in 1998 by the Data Mining Group (www.dmg.org): it goes by the name of Predictive Model Markup Language (PMML). This language can describe the data dictionary used (variables, with their types and values) and the data transformations carried out (recoding, normalization, discretization, aggregation), and can use tags to specify the parameters of various types of model (regressions, trees, clustering, neural networks, etc.). By installing a PMML interpreter or relational databases, it is possible to deploy data mining models in an operating environment which may be different from the development environment. Moreover, these models can

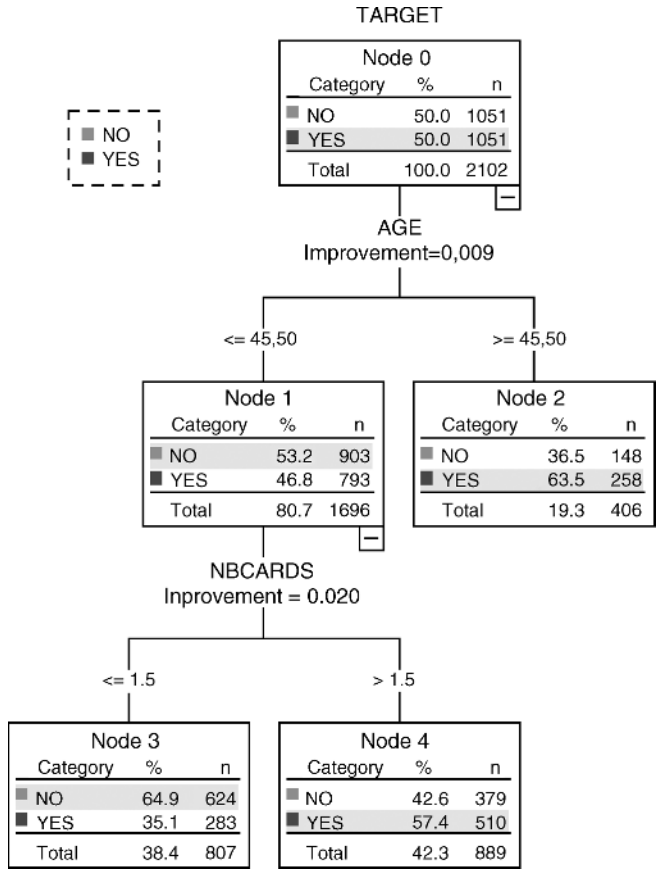


**Figure 1.3** Example of a decision tree generated by Answer Tree.

be generated by different data mining programs (SAS, IBM SPSS, R, for example), since the PMML language tends to spread slowly, even though it remains less widespread and possibly less efficient than C, Java and SQL.

In R, for example, a decision tree is exported by using the *pmml* package (which also requires the *XML* package). The first step is to create the model, using the *rpart* tree function (Figure 1.6). The *pmml* package currently allows the export of models produced by linear regression, logistic regression, support vector machines, neural networks (the *nnet* package), decision tree (the *rpart* package), random forests and *k*-means.

I have mentioned three software packages in this section: SAS, IBM SPSS and R. These will be described in detail, with other data mining software, in Chapter 5.

## 1.5 Data mining and protection of personal data

Statisticians and data miners must comply with their national law with regard to the processing of personal data. Such data are defined as those which can be directly or indirectly related to an individual physical person by using his civil status, another identifier such as the

```
/* Node 3 */
DO IF (SYSMIS(age) OR (VALUE(age) LE 45.5)) AND
((VALUE(nbcards) LE 1.5) OR SYSMIS(nbcards)).
COMPUTE nod_001 = 3.
COMPUTE pre_001 = 'N'.
COMPUTE prb_001 = 0.649318.
END IF.
EXECUTE.

/* Node 4 */
DO IF (SYSMIS(age) OR (VALUE(age) LE 45.5)) AND
((VALUE(nbcards) GT 1.5) OR SYSMIS(nbcards)).
COMPUTE nod_001 = 4.
COMPUTE pre_001 = 'O'.
COMPUTE prb_001 = 0.573678.
END IF.
EXECUTE.

/* Node 2 */
DO IF (VALUE(age) GT 45.5).
COMPUTE nod_001 = 2.
COMPUTE pre_001 = 'O'.
COMPUTE prb_001 = 0.635468.
END IF.
EXECUTE.
```

**Figure 1.4** Example of SPSS code for a decision tree.

telephone number of a customer or an assured party, or any other element belonging to him (voice, image, genetic or biometric fingerprints, address, etc.).[6]

Most countries have passed laws to restrict the collection, storage, processing and use of personal data, especially sensitive data, relating to health, sexual orientation, criminal convictions, racial origin, political opinions and religious faith. This is also the case in the European Union member states that have adopted European Directive 95/46/EC of 24 October 1995 into their national law. According to Article 6 of this directive, personal data must be:

(a) processed fairly and lawfully;

(b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards;

[6] This does not apply to files on physical persons that are anonymized by the removal of all identifiers that could be used to trace them. Such files may be useful for statistical research.

```
************************************;
*** begin scoring code for clustering;
************************************;
label _SEGMNT_ = 'Cluster ID'
    Distance = 'Distance to Cluster Seed';
drop _nonmiss; _nonmiss = n(
  AGE , SAVINGS, NBPROD , EXPENDIT , INCOME , NBCARDS);
if _nonmiss = 0 then do;
  _SEGMNT_ _ = .; distance = .;
end;
else do;
  array _CLScads[3] _temporary_;
  drop _clus;
  do _clus = 1 to 3; _CLScads[_clus] = 0; end;
  if n(AGE) then do;
    _CLScads[1] + (AGE - 69.111111111 )**2;
    _CLScads[2] + (AGE - 70.095238095 )**2;
    _CLScads[3] + (AGE - 43.473900586 )**2;
  end;
  if n(SAVINGS) then do;
    _CLScads[1] + (SAVINGS - 383125.82523 )**2;
    _CLScads[2] + (SAVINGS - 256109.6931 )**2;
    _CLScads[3] + (SAVINGS - 14778.055064 )**2;
  end;
  if n(NBPROD) then do;
    _CLScads[1] + (NBPROD - 14.055555556 )**2;
    _CLScads[2] + (NBPROD - 15.476190476 )**2;
    _CLScads[3] + (NBPROD - 8.8776628878 )**2;
  end;
  if n(EXPENDIT) then do;
    _CLScads[1] + (DEPENSES - 5091.3631019 )**2;
    _CLScads[2] + (EXPENDIT- 3699.0411688 )**2;
    _CLScads[3] + (EXPENDIT - 2296.6205468 )**2;
  end;
  if n(INCOME) then do;
    _CLScads[1] + (INCOME - 3393.2086589 )**2;
    _CLScads[2] + (INCOME - 3247.8545619 )**2;
    _CLScads[3] + (INCOME - 1863.3223265 )**2;
  end;
  if n(NBCARDS) then do;
    _CLScads[1] + (NBCARDS - 1 )**2;
    _CLScads[2] + (NBCARDS - 1.2380952381 )**2;
    CLScads[3] + (NBCARDS - 1.4502304722 )**2;
  end;
_SEGMNT_ _ = 1; distance = _CLScads[1];
  do _clus = 2 to 3;
    if _CLScads[_clus] < distance then do;
      _SEGMNT_ = _clus; distance = _CLScads[_clus];
    end;
  end;
  distance = sqrt(distance*6/_nonmiss);
end;
**********************************;
*** end scoring code for clustering;
 **********************************;
```

**Figure 1.5** Example of SAS code generated by SAS Enterprise Miner.

```
  > library(pmml)
  Loading required package: XML
  > pmml(titanic.rpart)
  <PMML version=''3.2'' xmlns=''http://www.dmg.org/PMML-3_2''
xmlns:xsi=''http://www.w3.org/2001/XMLSchema-instance''
xsi:schemaLocation=''http://www.dmg.org/PMML-3_2
http://www.dmg.org/v3-2/pmml-3-
2.xsd''>
  <Header copyright=''Copyright (c) 2010 Stéphane''
description=''RPart Decision Tree Model''>
  <Extension name=''timestamp'' value=''2010-06-27 16:30:10''
extender=''Rattle''/>
   <Extension name=''description'' value=''Stéphane''
extender=''Rattle''/>
   <Application name=''Rattle/PMML'' version=''1.2.15''/>
   </Header>
  <DataDictionary numberOfFields=''4''>
  <DataField name=''survived'' optype=''continuous''
dataType=''double''/>
   <DataField name=''class'' optype=''continuous''
dataType=''double''/>
 <DataField name=''age'' optype=''continuous'' dataType=''double''/>
 <DataField name=''sex'' optype=''continuous'' dataType=''double''/>
  </DataDictionary>
  <TreeModel modelName=''RPart_Model'' functionName=''regression''
algorithmName=''rpart'' splitCharacteristic=''binarySplit''
missingValueStrategy=''defaultChild''>
   <MiningSchema>
    <MiningField name=''survived'' usageType=''predicted''/>
    <MiningField name=''class'' usageType=''active''/>
    <MiningField name=''age'' usageType=''active''/>
    <MiningField name=''sex'' usageType=''active''/>
   </MiningSchema>
   <Node id=''1'' score=''0.323034984098137'' recordCount=''2201''
defaultChild=''2''>
    <True/>
    <Node id=''2'' score=''0.212016175621028'' recordCount=''1731''
defaultChild=''4''>
     <SimplePredicate field=''sex'' operator=''greaterOrEqual''
value=''0.5''/>
    <Node id=''4'' score=''0.202759448110378'' recordCount=''1667''>
     <SimplePredicate field=''age'' operator=''greaterOrEqual''
value=''0.5''/>
    </Node>
    <Node id=''5'' score=''0.453125'' recordCount=''64''
defaultChild=''10''>
     <SimplePredicate field=''age'' operator=''lessThan''
value=''0.5''/>
     <Node id=''10'' score=''0.270833333333333'' recordCount=''48''>
      <SimplePredicate field=''class'' operator=''greaterOrEqual''
value=''2.5''/>
```

**Figure 1.6** Exporting a model into PMML in R software.

```
     </Node>
     <Node id=''11'' score=''1'' recordCount=''16''>
      <SimplePredicate field=''class'' operator=''lessThan''
value=''2.5''/>
     </Node>
    </Node>
   </Node>
   <Node id=''3'' score=''0.731914893617021'' recordCount=''470''
defaultChild=''6''>
    <SimplePredicate field=''sex'' operator=''lessThan''
value=''0.5''/>
    <Node id=''6'' score=''0.459183673469388'' recordCount=''196''>
     <CompoundPredicate booleanOperator=''surrogate''>
      <SimplePredicate field=''class'' operator=''greaterOrEqual''
value=''2.5''/>
      <SimplePredicate field=''age'' operator=''lessThan''
value=''0.5''/>
     </CompoundPredicate>
    </Node>
    <Node id=''7'' score=''0.927007299270073'' recordCount=''274''>
     <CompoundPredicate booleanOperator=''surrogate''>
      <SimplePredicate field=''class'' operator=''lessThan''
value=''2.5''/>
      <SimplePredicate field=''age'' operator=''greaterOrEqual''
 value=''0.5''/>
    </CompoundPredicate>
   </Node>
  </Node>
 </Node>
</TreeModel>
</PMML>
>
```

**Figure 1.6** (*Continued*).

(c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed;

(d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified;

(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed.

Article 25 adds that personal data may be transferred to a third country only if the country in question ensures an adequate level of the protection of the data.

Unless an exception is made, for example in the case of a person working in the medical field, no statistician or other person is allowed to work on the aforementioned sensitive data unless they have been anonymized[7]. In some cases, the secrecy surrounding these data is reinforced by specific regulations, such as those relating to medical or banking confidentiality. Clearly, the disclosure of medical information about a patient may be harmful to him in terms of his image, but also, and more seriously, if it creates difficulties when he is seeking work or insurance cover. In another field, the abuse of Internet surfing data may be an intrusion into private life, because such data may reveal preferences and habits, and may lead to unwelcome selling operations. In banking, the disclosure of confidential data may expose their owner to obvious risks of fraud. Banking data, especially those relating to the use of bank cards, could also lead to drift, if they are used to analyse the lifestyle, movements and consumption habits of cardholders. This could be used to create a customer profile, which is exploited for gain, but which bears no relationship to the purpose for which these data were collected. As a general rule, therefore, all such statistical and computerized processing is highly restricted, in order to avoid cases in which customers innocently providing information simply for the purposes of managing their contracts find that, unbeknown to them, the information has been used in other ways for commercial purposes. Close attention is paid to the interconnection of computer files, which would make it possible to link different kinds of information, collected for different purposes, in such a way that disclosure would lead to the abusive or dangerous retention of files on individuals. Sadly, history has shown that the fear of 'Big Brother' is not unjustified. In this context, the sensitive nature of data on origin or political or religious opinions is evident, but legislators have generally considered that the storage of personal data on an individual could affect the freedom of the individual, even if none of the data are sensitive when taken in isolation.

This is particularly true of data relating to different aspects of an individual which can be used to create a 'profile' of the individual, using automatic processing, which may cause him to lose a right, a service or a commercial offer. In this area, risk scoring and behavioural segmentation are processes that are monitored and regulated by the authorities. The data miner must be aware of this, and must be careful to use only the data and processes that are legally available.

This fear of the power of files has even led some countries, including France, to restrict the dissemination of geomarketing data, such as those obtained from censuses, to a sufficiently coarse level of resolution to prevent these data from being applied too accurately to specific individuals (see Section 4.2.1). Geodemographic databases also contain hundreds of pieces of data on the income, occupational and family circumstances, consumption habits and lifestyles of the inhabitants of each geographical area. If each geographical area contains only a few tens of families, the data on a given family would clearly be very similar to the mean data in the geographical area, and the profile of each family could therefore be fairly accurately estimated.

The principles stated above are those that apply in the European Union, in the European Economic Area (EU, Iceland, Liechtenstein, Norway), in Switzerland, in Canada and in

[7] Current cryptographic methods, such as *hashing*, can be used for *secure matching* of files, and for reconciling anonymity with the need to cross-check files containing data on a single person or to update them subsequently. Each identifier is associated with a personal code, such that the identifier can be used to retrieve the personal code, but not vice versa. This personal code is then used to match the files. These mechanisms have been used in the medical field for several years.

Argentina. In the United States, personal data protection is much less closely regulated, and there is no equivalent to Directive 95/46/EC. Very few states have followed California in passing appropriate legislation, and data protection is based on the reactions of citizens rather than the law. However, some processes are subject to more restrictions than elsewhere in some areas; for example, the granting of credit has been covered by the Equal Credit Opportunity Act since 1974. This law prohibits the use of certain variables in the lending criteria, and therefore in scoring systems as well. Ethnic origin, nationality and religion must not be used. Neither must the sex of the applicant. Age can be used only if it does not penalize older persons. Income can be taken into account, but its origin (wages, pensions, social security, etc.) must not. No distinction must be made between married and unmarried applicants. However, occupation, seniority in employment, and owner-occupier status may be taken into account. Unless credit is refused, there is no obligation to tell an applicant what his score is (it is available only once per year on the Annualcreditreport.com site), but some organizations sell this information to applicants and also offer advice on improving their scores.

In another area, the *Safe Harbor Privacy Principles*, which US organizations are asked to comply with, were developed by the US Department of Commerce in response to Article 25 of Directive 95/46/EC, to protect the security of exchanges with those organizations that agree to observe these principles. However, it appears that these principles are not always followed in practice, and the exchange of data between Europe and the USA is a regular source of conflict. The history of Passenger Name Record (PNR) data is one example of this.

Following the attacks on the USA on 11 September 2001, the Information Awareness Office was established in January 2002, under the aegis of the US Department of Defense, to provide permanent automatic surveillance of all possible kinds of information which might be evidence of preparation for terrorist activities. This project, called *Total Information Awareness* (TIA), and renamed *Terrorism Information Awareness* in 2003, was intended to enable significant links to be created between police and judicial data and behaviour such as applications for passports, visas, work permits and driving licences, the use of credit cards, the purchase of airline tickets, the hiring of cars, and the purchase of chemical products, weapons, etc. TIA included other aspects such as automatic recognition of a human face in a crowd, automatic transcription of verbal and written communications in foreign languages, surveillance of medical databanks to detect a possible biological attack, and the surveillance of certain movements on the stock exchanges. This project led to protests by defenders of individual liberties, and its funding was stopped in September 2003. It reappeared shortly afterwards in other forms such as the CAPPS (Computer Assisted Passenger Prescreening System) and CAPPS2 of the Transportation Security Administration. This programme was more closely targeted than the preceding ones, since it was intended to be applied to users of air transport only: on embarking, each passenger supplied his name, address and telephone number, permitting automatic consultation of several hundred databases of various kinds (government and private). The result of this interrogation could trigger an alert and result in a passenger being searched. The initial aim was to extend this from the USA to Europe, but the US companies that had agreed to participate in the tests were boycotted by consumers, and the programme was finally wound up by Congress because of the dangers it posed to private life. Furthermore, opposition developed in Europe against the transmission of PNR data to the US customs and security services. These were data exchanged in standardized form between the stakeholders in air transport. They could be used by CAPPS2, and their extent and period of retention were critical. In 2004, CAPPS2 was replaced by *Secure Flight* and *VRPS* (Voluntary Registered Passenger System), and agreement was reached, with some difficulty, in August

2007 between the European Union and the United States on the transmission of PNR data to the authorities. The number and nature of these data (bank card number, telephone number, address, etc.) were always considered excessive by the Europeans, especially since their period of retention was 15 years, the right of access and correction was at the discretion of the Americans, access to sensitive data (ethnic origin, political opinions, state of health, etc.) was possible, the purposes of use could be extended, and the data could be transferred to third countries by the Americans: these were all infringements of Directive 95/46/EC. Further disputes between the two continents arose from the 'no-fly lists', or lists of passengers named as undesirable by the USA, supplied by the US authorities to the airlines who were asked to ensure that any passengers flying to the USA did not appear on these lists. The legal basis and quality of these lists were questioned by the Europeans, who pointed out, among other things, the risks of confusion of similar names in these lists, which are updated daily but still contain 65 000 names.

Any statistician or data miner dealing with personal data, even if these are rarely as sensitive as PNR data, must be careful to respect the private life of the persons concerned and must avoid using the data in a way which might cause undue offence. This is particularly true in a world in which the resources of information technology are so vast that enormous quantities of information can be collected and stored on each of us, almost without any technical limitation. This is even more important in the field of data mining, a method which is often used to help with decision making, and which could easily be transformed, if care is not taken, into a tool which takes decisions automatically on the basis of collected information. New problems have arisen concerning our individual data, which no longer simply form the basis for global, anonymous analysis, but can also be used to make decisions which can change the lives of individuals.

For further details on the protection of personal data and the problems it entails in modern society, see Chapter 4 of David Hand's book, *Information Generation: How Data Rule Our World*,[8] as well as the websites of the national data protection authorities.

## 1.6 Implementation of data mining

The main factors in the success of a project are:

(i) precise, substantial and realistic targets;

(ii) the richness, and above all the quality, of the information collected;

(iii) the cooperation of the departmental and statistical specialists in the organization;

(iv) the relevance of the data mining techniques used;

(v) satisfactory output of the information generated and correct integration into the information system where appropriate;

(vi) analysis of the results and feedback of experience from each application of data mining, to be used for the next application.

I will examine these matters in detail later on, especially in the next chapter and Chapter 13.

[8] Hand, D.J. (2007) *Information Generation: How Data Rule Our World.* Oxford: Oneworld Publications.

Data mining can be applied in different ways in a business. The business may entirely outsource the data mining operation, as well as its computer facilities management, supplying raw commercial files as required to specialist service providers. The service providers then return the commercial files supplemented with information such as customers' scores, their behavioural segments, etc. Alternatively, the business may subcontract the essentials of the data mining operation, with its service providers developing the data mining models it requires, but then take over these models and apply them to its own files, possibly modifying them slightly. Finally, the business may develop its own data mining models, using commercial software, possibly with the assistance of specialist consultants – and this book, of course! These different approaches are described more fully in Section 12.6.