Business Intelligence: Data Mining and Optimization for Decision Making

Carlo Vercellis

Politecnico di Milano, Italy.

WILEY

A John Wiley and Sons, Ltd., Publication

Business Intelligence

Business Intelligence: Data Mining and Optimization for Decision Making

Carlo Vercellis

Politecnico di Milano, Italy.

WILEY

A John Wiley and Sons, Ltd., Publication

This edition first published 2009

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Vercellis, Carlo.
Business intelligence : data mining and optimization for decision making / Carlo Vercellis.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-51138-1 (cloth) – ISBN 978-0-470-51139-8 (pbk. : alk. paper)
1. Decision making–Mathematical models. 2. Business intelligence. 3. Data mining. I. Title.
HD30.23.V476 2009
658.4′038–dc22

2008043814

A catalogue record for this book is available from the British Library.

ISBN: 978-0-470-51138-1 (Hbk)
ISBN: 978-0-470-51139-8 (Pbk)

Typeset in 10.5/13pt Times by Laserwords Private Limited, Chennai, India
Printed in the United Kingdom by TJ International, Padstow, Cornwall

Contents

Preface

Since the 1990s, the socio-economic context within which economic activities are carried out has generally been referred to as the *information and knowledge society*. The profound changes that have occurred in methods of production and in economic relations have led to a growth in the importance of the exchange of intangible goods, consisting for the most part of transfers of information. The acceleration in the pace of current transformation processes is due to two factors. The first is *globalization*, understood as the ever-increasing interdependence between the economies of the various countries, which has led to the growth of a single *global economy* characterized by a high level of integration. The second is the new *information technologies*, marked by the massive spread of the Internet and of wireless devices, which have enabled high-speed transfers of large amounts of data and the widespread use of sophisticated means of communication.

In this rapidly evolving scenario, the wealth of development opportunities is unprecedented. The easy access to information and knowledge offers several advantages to various actors in the socio-economic environment: *individuals*, who can obtain news more rapidly, access services more easily and carry out on-line commercial and banking transactions; *enterprises*, which can develop innovative products and services that can better meet the needs of the users, achieving competitive advantages from a more effective use of the knowledge gained; and, finally, the *public administration*, which can improve the services provided to citizens through the use of e-government applications, such as on-line payments of tax contributions, and e-health tools, by taking into account each patient's medical history, thus improving the quality of healthcare services.

In this framework of radical transformation, methods of governance within complex organizations also reflect the changes occurring in the socio-economic environment, and appear increasingly more influenced by the immediate access to information for the development of effective action plans. The term *complex organizations* will be used throughout the book to collectively refer to a diversified set of entities operating in the socio-economic context, including enterprises, government agencies, banking and financial institutions, and non-profit organizations.

The adoption of low-cost massive data storage technologies and the wide availability of Internet connections have made available large amounts of data that have been collected and accumulated by the various organizations over the years. The enterprises that are capable of transforming data into *information* and *knowledge* can use them to make quicker and more effective decisions and thus to achieve a competitive advantage. By the same token, on the public administration side, the analysis of the available information enables the development of better and innovative services for citizens. These are ambitious objectives that technology, however sophisticated, cannot perform on its own, without the support of competent minds and advanced analysis methodologies.

Is it possible to extract, from the huge amounts of data available, knowledge which can then be used by *decision makers* to aid and improve the governance of the enterprises and the public administration?

Business intelligence may be defined as a set of mathematical models and analysis methodologies that systematically exploit the available data to retrieve information and knowledge useful in supporting complex decision-making processes.

Despite the somewhat restrictive meaning of the term *business*, which seems to confine the subject within the boundaries of enterprises, business intelligence systems are aimed at companies as well as other types of complex organizations, as mentioned above.

Business intelligence methodologies are interdisciplinary and broad, spanning several domains of application. Indeed, they are concerned with the representation and organization of the decision-making process, and thus with the field of decision theory; with collecting and storing the data intended to facilitate the decision-making process, and thus with data warehousing technologies; with mathematical models for optimization and data mining, and thus with operations research and statistics; finally, with several application domains, such as marketing, logistics, accounting and control, finance, services and the public administration.

We can say that business intelligence systems tend to promote a scientific and rational approach to managing enterprises and complex organizations. Even the use of an electronic spreadsheet for assessing the effects induced on the budget by fluctuations in the discount rate, despite its simplicity, requires on the part of decision makers a mental representation of the financial flows.

A business intelligence environment offers decision makers information and knowledge derived from data processing, through the application of mathematical models and algorithms. In some instances, these may merely consist of the calculation of totals and percentages, while more fully developed analyses make use of advanced models for optimization, inductive learning and prediction.

In general, a model represents a selective abstraction of a real system, designed to analyze and understand from an abstract point of view the operating behavior of the real system. The model includes only the elements of the system deemed relevant for the purpose of the investigation carried out. It is worth quoting the words of Einstein on the subject of model development: 'Everything should be made as simple as possible, but not simpler.'

Classical scientific disciplines, such as physics, have always made use of mathematical models for the abstract representation of real systems, while other disciplines, such as operations research, have dealt with the application of scientific methods and mathematical models to the study of artificial systems, such as enterprises and complex organizations.

'The great book of nature', as Galileo wrote, 'may only be read by those who know the language in which it was written. And this language is mathematics.' Can we apply also to the analysis of artificial systems this profound insight from one of the men who opened up the way to modern science?

We believe so. Nowadays, the mere intuitive abilities of decision makers managing enterprises or the public administration are outdone by the complexity of governance of current organizations. As an example, consider the design of a marketing campaign in dynamic and unpredictable markets, where however a wealth of information is available on the buying behavior of the consumers. Today, it is inconceivable to leave aside the application of advanced inferential learning models for selecting the recipients of the campaign, in order to optimize the allocation of resources and the redemption of the marketing action.

The interpretation of the term *business intelligence* that we have illustrated and that we intend to develop in this book is much broader and deeper compared to the narrow meaning publicized over the last few years by many software vendors and information technology magazines. According to this latter vision, business intelligence methodologies are reduced to electronic tools for querying, visualization and reporting, mainly for accounting and control purposes. Of course, no one can deny that rapid access to information is an invaluable tool for decision makers. However, these tools are oriented toward business intelligence analyses of a *passive* nature, where the decision maker has already formulated in her mind some criteria for data extraction. If we wish business intelligence methodologies to be able to express their huge strategic potential, we should turn to *active* forms of support for decision making, based on the systematic adoption of mathematical models able to transform data not only into *information* but also into *knowledge*, and then knowledge into actual competitive advantage. The distinction between passive and active forms of analysis will be further investigated in Chapter 1.

One might object that only simple tools based on immediate and intuitive concepts have the ability to prove useful in practice. In reply to this objection, we cannot do better than quote Vladimir Vapnik, who more than anyone has contributed to the development of inductive learning models: 'Nothing is more practical than a good theory.'

Throughout this book we have tried to make frequent reference to problems and examples drawn from real applications in order to help readers understand the topics discussed, while ensuring an adequate level of methodological rigor in the description of mathematical models.

Part I describes the basic components that make up a business intelligence environment, discussing the structure of the decision-making process and reviewing the underlying information infrastructures. In particular, Chapter 1 outlines a general framework for business intelligence, highlighting the connections with other disciplines. Chapter 2 describes the structure of the decision-making process and introduces the concept of a decision support system, illustrating the main advantages it involves, the critical success factors and some implementation issues. Chapter 3 presents data warehouses and data marts, first analyzing the reasons that led to their introduction, and then describing on-line analytical processing analyses based on multidimensional cubes.

Part II is more methodological in character, and offers a comprehensive overview of mathematical models for pattern recognition and data mining. Chapter 4 describes the main characteristics of mathematical models used for business intelligence analyses, offering a brief taxonomy of the major classes of models. Chapter 5 introduces data mining, discussing the phases of a data mining process and their objectives. Chapter 6 describes the activities of data preparation for business intelligence and data mining; these include data validation, anomaly detection, data transformation and reduction. Chapter 7 provides a detailed discussion of exploratory data analysis, performed by graphical methods and summary statistics, in order to understand the characteristics of the attributes in a dataset and to determine the intensity of the relationships among them. Chapter 8 describes simple and multiple regression models, discussing the main diagnostics for assessing their significance and accuracy. Chapter 9 illustrates the models for time series analysis, examining decomposition methods, exponential smoothing and autoregressive models. Chapter 10 is entirely devoted to classification models, which play a prominent role in pattern recognition and learning theory. After a description of the evaluation criteria, the main classification methods are illustrated; these include classification trees, Bayesian methods, neural networks, logistic regression and support vector machines. Chapter 11 describes association rules and the Apriori algorithm. Chapter 12 presents the best-known clustering models: partition methods, such

as K-means and K-medoids, and hierarchical methods, both agglomerative and divisive.

Part III illustrates the applications of data mining to relational marketing (Chapter 13), models for salesforce planning (Chapter 13), models for supply chain optimization (Chapter 14) and analytical methods for performance assessment (Chapter 15).

Appendix A provides information and links to software tools used to carry out the data mining and business intelligence analyses described in the book. Preference has been given to *open source* software, since in this way readers can freely download it from the Internet to practice on the examples given. By the same token, the datasets used to exemplify the different topics are also mostly taken from repositories in the public domain. Appendix B includes a short description of the datasets used in the various chapters and the links to sites that contain these as well as other datasets useful for experimenting with and comparing the analysis methodologies.

Bibliographical notes at the end of each chapter, highly selective as they are, highlight other texts that we found useful and relevant, as well as research contributions of acknowledged historical value.

This book is aimed at three main groups of readers. The first are students studying toward a master's degree in economics, business management or other scientific disciplines, and attending a university course on business intelligence methodologies, decision support systems and mathematical models for decision making. The second are students on doctoral programs in disciplines of an economic and management nature. Finally, the book may also prove useful to professionals wishing to update their knowledge and make use of a methodological and practical reference textbook. Readers belonging to this last group may be interested in an overview of the opportunities offered by business intelligence systems, or in specific methodological and applied subjects dealt with in the book, such as data mining techniques applied to relational marketing, salesforce planning models, supply chain optimization models and analytical methods for performance evaluation.

At Politecnico di Milano, the author leads the research group *MOLD – Mathematical modeling, optimization, learning from data*, which conducts methodological research activities on models for inductive learning, prediction, classification, optimization, systems biology and social network analysis, as well as applied projects on business intelligence, relational marketing and logistics. The research group's website, *www.mold.polimi.it*, includes information, news, in-depth studies, useful links and updates.

A book free of misprints is a rare occurrence, especially in the first edition, despite the efforts made to avoid them. Therefore, a dedicated area for errata and corrigenda has been created at *www.mold.polimi.it*, and readers are welcome

to contribute to it by sending a note on any typos that they might find in the text to the author at *carlo.vercellis@polimi.it*.

I wish to express special thanks to Carlotta Orsenigo, who helped write Chapter 10 on classification models and discussed with me the content and the organization of the remaining chapters in the book. Her help in filling gaps, clarifying concepts, and making suggestions for improvement to the text and figures was invaluable.

To write this book, I have drawn on my experience as a teacher of graduate and postgraduate courses. I would therefore like to thank here all the many students who through their questions and curiosity have urged me to seek more convincing and incisive arguments.

Many examples and references to real problems originate from applied projects that I have carried out with enterprises and agencies of the public administration. I am indebted to many professionals for some of the concepts that I have included in the book: they are too numerous to name but will certainly recognize themselves in some statements, and to all of them I extend a heartfelt thank-you.

All typos and inaccuracies in this book are entirely my own responsibility.

Part I

Components of the decision-making process

1

Business intelligence

The advent of low-cost data storage technologies and the wide availability of Internet connections have made it easier for individuals and organizations to access large amounts of data. Such data are often heterogeneous in origin, content and representation, as they include commercial, financial and administrative transactions, web navigation paths, emails, texts and hypertexts, and the results of clinical tests, to name just a few examples. Their accessibility opens up promising scenarios and opportunities, and raises an enticing question: is it possible to convert such data into information and knowledge that can then be used by decision makers to aid and improve the governance of enterprises and of public administration?

Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. This opening chapter will describe in general terms the problems entailed in business intelligence, highlighting the interconnections with other disciplines and identifying the primary components typical of a business intelligence environment.

1.1 Effective and timely decisions

In complex organizations, public or private, decisions are made on a continual basis. Such decisions may be more or less critical, have long- or short-term effects and involve people and roles at various hierarchical levels. The ability of these *knowledge workers* to make decisions, both as individuals and as a community, is one of the primary factors that influence the performance and competitive strength of a given organization.

Business Intelligence: *Data Mining and Optimization for Decision Making* C. Vercellis

Most knowledge workers reach their decisions primarily using easy and intuitive methodologies, which take into account specific elements such as experience, knowledge of the application domain and the available information. This approach leads to a stagnant decision-making style which is inappropriate for the unstable conditions determined by frequent and rapid changes in the economic environment. Indeed, decision-making processes within today's organizations are often too complex and dynamic to be effectively dealt with through an intuitive approach, and require instead a more rigorous attitude based on analytical methodologies and mathematical models. The importance and strategic value of analytics in determining competitive advantage for enterprises has been recently pointed out by several authors, as described in the references at the end of this chapter. Examples 1.1 and 1.2 illustrate two highly complex decision-making processes in rapidly changing conditions.

Example 1.1 – Retention in the mobile phone industry. The marketing manager of a mobile phone company realizes that a large number of customers are discontinuing their service, leaving her company in favor of some competing provider. As can be imagined, low customer loyalty, also known as customer *attrition* or *churn*, is a critical factor for many companies operating in service industries. Suppose that the marketing manager can rely on a budget adequate to pursue a customer retention campaign aimed at 2000 individuals out of a total customer base of 2 million people. Hence, the question naturally arises of how she should go about choosing those customers to be contacted so as to optimize the effectiveness of the campaign. In other words, how can the probability that each single customer will discontinue the service be estimated so as to target the best group of customers and thus reduce churning and maximize customer retention? By knowing these probabilities, the target group can be chosen as the 2000 people having the highest churn likelihood among the customers of high business value. Without the support of advanced mathematical models and data mining techniques, described in Chapter 5, it would be arduous to derive a reliable estimate of the churn probability and to determine the best recipients of a specific marketing campaign.

Example 1.2 – Logistics planning. The logistics manager of a manufacturing company wishes to develop a medium-term logistic-production plan. This is a decision-making process of high complexity which includes,

among other choices, the allocation of the demand originating from different market areas to the production sites, the procurement of raw materials and purchased parts from suppliers, the production planning of the plants and the distribution of end products to market areas. In a typical manufacturing company this could well entail tens of facilities, hundreds of suppliers, and thousands of finished goods and components, over a time span of one year divided into weeks. The magnitude and complexity of the problem suggest that advanced optimization models are required to devise the best logistic plan. As we will see in Chapter 14, optimization models allow highly complex and large-scale problems to be tackled successfully within a business intelligence framework.

The main purpose of business intelligence systems is to provide knowledge workers with tools and methodologies that allow them to make *effective* and *timely* decisions.

Effective decisions. The application of rigorous analytical methods allows decision makers to rely on information and knowledge which are more dependable. As a result, they are able to make better decisions and devise action plans that allow their objectives to be reached in a more effective way. Indeed, turning to formal analytical methods forces decision makers to explicitly describe both the criteria for evaluating alternative choices and the mechanisms regulating the problem under investigation. Furthermore, the ensuing in-depth examination and thought lead to a deeper awareness and comprehension of the underlying logic of the decision-making process.

Timely decisions. Enterprises operate in economic environments characterized by growing levels of competition and high dynamism. As a consequence, the ability to rapidly react to the actions of competitors and to new market conditions is a critical factor in the success or even the survival of a company.

Figure 1.1 illustrates the major benefits that a given organization may draw from the adoption of a business intelligence system. When facing problems such as those described in Examples 1.1 and 1.2 above, decision makers ask themselves a series of questions and develop the corresponding analysis. Hence, they examine and compare several options, selecting among them the best decision, given the conditions at hand.

If decision makers can rely on a business intelligence system facilitating their activity, we can expect that the overall quality of the decision-making process will be greatly improved. With the help of mathematical models and algorithms, it is actually possible to analyze a larger number of alternative

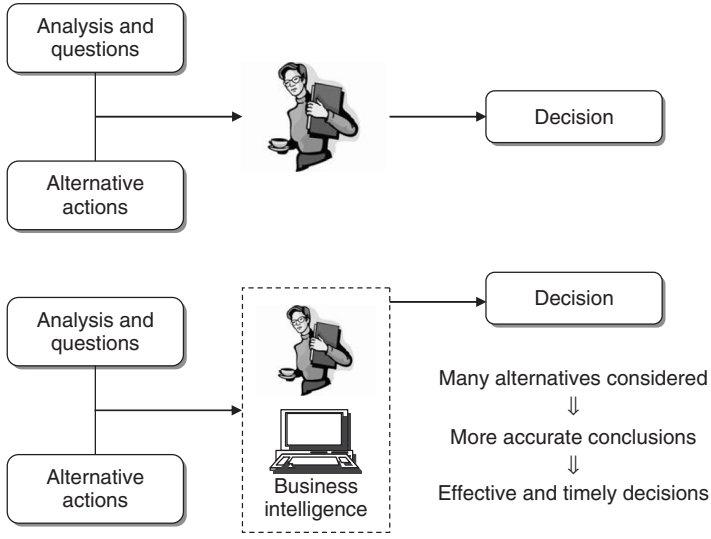

Figure 1.1 Benefits of a business intelligence system

actions, achieve more accurate conclusions and reach effective and timely decisions. We may therefore conclude that the major advantage deriving from the adoption of a business intelligence system is found in the increased *effectiveness* of the decision-making process.

1.2 Data, information and knowledge

As observed above, a vast amount of data has been accumulated within the information systems of public and private organizations. These data originate partly from internal transactions of an administrative, logistical and commercial nature and partly from external sources. However, even if they have been gathered and stored in a systematic and structured way, these data cannot be used directly for decision-making purposes. They need to be processed by means of appropriate extraction tools and analytical methods capable of transforming them into information and knowledge that can be subsequently used by decision makers.

The difference between *data*, *information* and *knowledge* can be better understood through the following remarks.

Data. Generally, data represent a structured codification of single primary entities, as well as of transactions involving two or more primary entities. For example, for a retailer data refer to primary entities such as customers, points of sale and items, while sales receipts represent the commercial transactions.

Information. Information is the outcome of extraction and processing activities carried out on data, and it appears meaningful for those who receive it in a specific domain. For example, to the sales manager of a retail company, the proportion of sales receipts in the amount of over €100 per week, or the number of customers holding a loyalty card who have reduced by more than 50% the monthly amount spent in the last three months, represent meaningful pieces of information that can be extracted from raw stored data.

Knowledge. Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions. Therefore, we can think of knowledge as consisting of information put to work into a specific domain, enhanced by the experience and competence of decision makers in tackling and solving complex problems. For a retail company, a sales analysis may detect that a group of customers, living in an area where a competitor has recently opened a new point of sale, have reduced their usual amount of business. The knowledge extracted in this way will eventually lead to actions aimed at solving the problem detected, for example by introducing a new free home delivery service for the customers residing in that specific area. We wish to point out that knowledge can be extracted from data both in a *passive* way, through the analysis criteria suggested by the decision makers, or through the *active* application of mathematical models, in the form of inductive learning or optimization, as described in the following chapters.

Several public and private enterprises and organizations have developed in recent years formal and systematic mechanisms to gather, store and share their wealth of knowledge, which is now perceived as an invaluable intangible asset. The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as *knowledge management*.

It is apparent that business intelligence and knowledge management share some degree of similarity in their objectives. The main purpose of both disciplines is to develop environments that can support knowledge workers in decision-making processes and complex problem-solving activities. To draw a boundary between the two approaches, we may observe that knowledge management methodologies primarily focus on the treatment of information that is usually unstructured, at times implicit, contained mostly in documents, conversations and past experience. Conversely, business intelligence systems are based on structured information, most often of a quantitative nature and usually organized in a database. However, this distinction is a somewhat fuzzy one: for example, the ability to analyze emails and web pages through text mining methods progressively induces business intelligence systems to deal with unstructured information.

1.3 The role of mathematical models

A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms. In some instances, this activity may reduce to calculations of totals and percentages, graphically represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models.

In general terms, the adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations. Even the use of a spreadsheet to estimate the effects on the budget of fluctuations in interest rates, despite its simplicity, forces decision makers to generate a mental representation of the financial flows process.

Classical scientific disciplines, such as physics, have always resorted to mathematical models for the abstract representation of real systems. Other disciplines, such as operations research, have instead exploited the application of scientific methods and mathematical models to the study of artificial systems, for example public and private organizations. Part II of this book will describe the main mathematical models used in business intelligence architectures and decision support systems, as well as the corresponding solution methods, while Part III will illustrate several related applications.

The rational approach typical of a business intelligence analysis can be summarized schematically in the following main characteristics.

- First, the objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
- Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
- Finally, *what-if* analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

Although their primary objective is to enhance the effectiveness of the decision-making process, the adoption of mathematical models also affords other advantages, which can be appreciated particularly in the long term. First, the development of an abstract model forces decision makers to focus on the main features of the analyzed domain, thus inducing a deeper understanding of the phenomenon under investigation. Furthermore, the knowledge about the domain acquired when building a mathematical model can be more easily transferred in the long run to other individuals within the same organization, thus allowing a sharper preservation of knowledge in comparison to empirical decision-making processes. Finally, a mathematical model developed for a

specific decision-making task is so general and flexible that in most cases it can be applied to other ensuing situations to solve problems of similar type.

1.4 Business intelligence architectures

The architecture of a business intelligence system, depicted in Figure 1.2, includes three major components.

Data sources. In a first stage, it is necessary to gather and integrate the data stored in the various primary and secondary sources, which are heterogeneous in origin and type. The sources consist for the most part of data belonging to operational systems, but may also include unstructured documents, such as emails and data received from external providers. Generally speaking, a major effort is required to unify and integrate the different data sources, as shown in Chapter 3.

Data warehouses and data marts. Using extraction and transformation tools known as *extract, transform, load* (ETL), the data originating from the different sources are stored in databases intended to support business intelligence analyses. These databases are usually referred to as *data warehouses* and *data marts*, and they will be the subject of Chapter 3.

Business intelligence methodologies. Data are finally extracted and used to feed mathematical models and analysis methodologies intended to support decision makers. In a business intelligence system, several decision support applications may be implemented, most of which will be described in the following chapters:

- multidimensional cube analysis;
- exploratory data analysis;

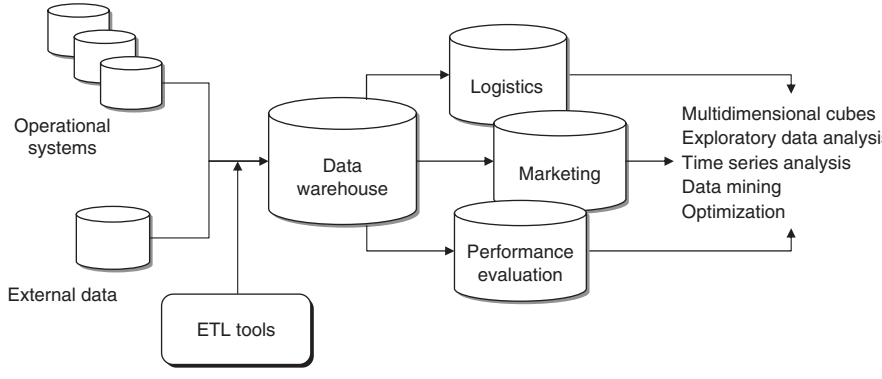

Figure 1.2 A typical business intelligence architecture

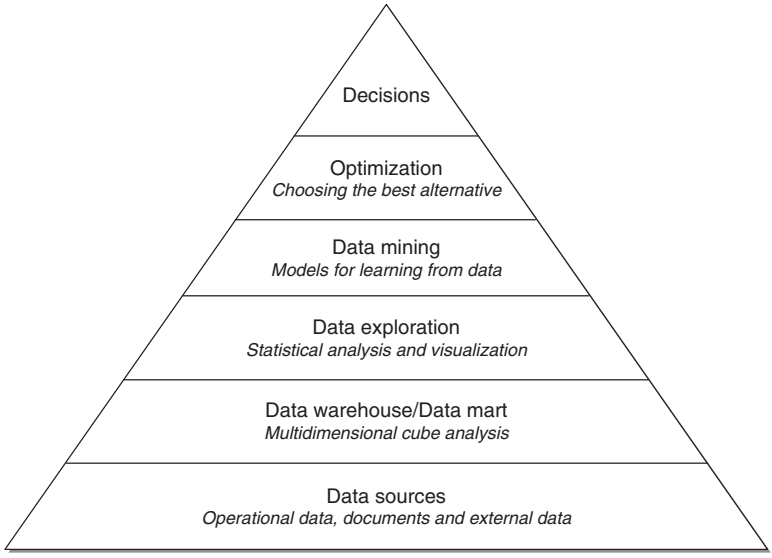

Figure 1.3 The main components of a business intelligence system

- time series analysis;
- inductive learning models for data mining;
- optimization models.

The pyramid in Figure 1.3 shows the building blocks of a business intelligence system. So far, we have seen the components of the first two levels when discussing Figure 1.2. We now turn to the description of the upper tiers.

Data exploration. At the third level of the pyramid we find the tools for performing a *passive* business intelligence analysis, which consist of query and reporting systems, as well as statistical methods. These are referred to as passive methodologies because decision makers are requested to generate prior hypotheses or define data extraction criteria, and then use the analysis tools to find answers and confirm their original insight. For instance, consider the sales manager of a company who notices that revenues in a given geographic area have dropped for a specific group of customers. Hence, she might want to bear out her hypothesis by using extraction and visualization tools, and then apply a statistical test to verify that her conclusions are adequately supported by data. Statistical techniques for exploratory data analysis will be described in Chapters 6 and 7.

Data mining. The fourth level includes *active* business intelligence methodologies, whose purpose is the extraction of information and knowledge from data.