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.