Introduzione

Le profonde trasformazioni intervenute nei modi di produzione e nelle relazioni economiche hanno attribuito una crescente rilevanza agli scambi di beni immateriali, riconducibili in larga misura a trasferimenti di informazioni. A partire dall'ultimo decennio del secolo scorso il contesto socio-economico entro il quale si svolgono le nostre attività viene infatti indicato come *società dell'informazione e della conoscenza*.

Due fattori hanno contribuito più di altri ad accrescere la rapidità dei processi di transizione in atto, se confrontati a fasi di evoluzione economica del passato. Da un lato la globalizzazione, intesa come crescente interdipendenza fra le economie dei diversi paesi, che ha condotto allo sviluppo di un'unica *economia globale* segnata da un alto livello di integrazione. Dall'altro le nuove tecnologie dell'informazione, caratterizzate dalla massiccia diffusione di *internet* e di dispositivi *wireless*, che hanno consentito di trasferire grandi quantità di dati ad elevata velocità e di sviluppare forme evolute di comunicazione.

Si tratta di uno scenario in rapida evoluzione che presenta opportunità di sviluppo senza precedenti. L'accesso alle informazioni e alla conoscenza presenta vantaggi per i diversi attori dell'ambiente socio-economico: da un lato gli individui, che possono acquisire più liberamente notizie, accedere a servizi con maggiore facilità, effettuare operazioni commerciali e bancarie on-line; dall'altro le imprese, che possono sviluppare prodotti innovativi e servizi più vicini alle esigenze degli utenti, e trarre vantaggi competitivi da un impiego efficace delle conoscenze maturate; infine la pubblica amministrazione, che può migliorare i servizi ai cittadini mediante attività di *e-government*, come pagamenti di tributi fiscali on-line, e di *e-health*, tenendo conto della storia terapeutica di ciascun paziente per migliorare l'assistenza sanitaria.

In questo quadro di radicale trasformazione anche i processi di governo delle strutture a organizzazione complessa riflettono i mutamenti del contesto circostante e appaiono sempre più condizionati dall'accesso tempestivo alle informazioni per lo sviluppo di piani d'azione efficaci. Con il termine *strutture a organizzazione complessa* vogliamo riferirci a un insieme articolato di attori operanti nell'ambiente socio-economico, comprendente le imprese, gli enti della pubblica amministrazione, gli istituti del sistema bancario e finanziario, le associazioni non lucrative.

L'adozione di tecnologie di memorizzazione di massa a basso costo e l'ampia diffusione della connettività hanno reso disponibili grandi moli di dati che si sono accumulate presso le diverse organizzazioni. Le imprese capaci di trasformare i dati in informazioni e in conoscenza sono in grado di elaborare decisioni più tem-

pestive ed efficaci, e di conseguire un differenziale competitivo. Analogamente, sul versante della pubblica amministrazione, l'analisi delle informazioni facilita lo sviluppo di servizi migliori e innovativi per i cittadini. Si tratta di compiti ambiziosi che la tecnologia, per quanto evoluta, non può svolgere da sola, senza il contributo di menti preparate e di adeguate metodologie di analisi.

È possibile estrarre conoscenze utili per il *decision making* dalle ingenti moli di dati disponibili presso le imprese e la pubblica amministrazione?

Con il termine *business intelligence* intendiamo riferirci a un insieme di modelli matematici e metodologie di analisi che esplorano i dati per ricavare informazioni e conoscenze utilizzabili nel corso dei processi decisionali.

A dispetto del carattere in certo modo restrittivo del termine *business*, che sembra confinare la tematica al solo ambito aziendale, le analisi di business intelligence si applicano sia alle imprese sia agli altri tipi di organizzazioni complesse cui abbiamo fatto riferimento in precedenza.

Le metodologie di business intelligence hanno un'ampia portata e una natura interdisciplinare. Esse riguardano infatti la rappresentazione e l'articolazione dei processi decisionali nelle organizzazioni, e quindi la teoria delle decisioni; la raccolta e la conservazione dei dati destinati a facilitare i processi decisionali, e quindi le tecniche di data warehousing; i modelli matematici per l'analisi dei dati, e quindi le metodologie della ricerca operativa e della statistica; e infine gli ambiti prevalenti di applicazione, quali il marketing, la logistica, il controllo di gestione, i sistemi finanziari, i servizi, la pubblica amministrazione.

Possiamo affermare in termini generali che le analisi di business intelligence tendono a promuovere un orientamento scientifico e razionale nella gestione delle imprese e delle organizzazioni a struttura complessa. Persino l'utilizzo di un foglio elettronico per valutare gli effetti provocati sul budget dalle variazioni nei tassi di sconto, a dispetto della sua semplicità, richiede da parte dei *decision maker* una rappresentazione mentale del processo dei flussi finanziari.

Un ambiente di business intelligence offre ai decision maker le informazioni e le conoscenze ricavate a partire dai dati, mediante l'applicazione di modelli matematici e di algoritmi. In alcuni casi questi ultimi possono ridursi al calcolo di totali e percentuali, visualizzati mediante semplici istogrammi, mentre le analisi più evolute utilizzano sofisticati modelli di ottimizzazione, di apprendimento induttivo e di previsione.

In generale, un modello rappresenta un'astrazione selettiva di un sistema reale, progettato per analizzare e comprendere da un punto di vista astratto il funzionamento di un sistema concreto, del quale contiene solo gli elementi ritenuti rilevanti ai fini dell'indagine svolta. Possiamo ricordare ciò che Einstein osservava a proposito dell'elaborazione di modelli: "bisognerebbe rendere tutto il più semplice possibile, ma non troppo semplice".

Le discipline scientifiche tradizionali, come la fisica, hanno sempre fatto ricorso a modelli matematici per la rappresentazione astratta di sistemi reali, mentre altre discipline, come la ricerca operativa, si sono occupate dell'applicazione di metodi scientifici e modelli matematici allo studio di sistemi artificiali, quali le imprese e le organizzazioni a struttura complessa.

"Il grande libro della natura - scriveva Galileo - può essere letto soltanto da coloro che conoscono il linguaggio in cui fu scritto. E questo linguaggio è la matematica". Possiamo adattare anche all'analisi dei sistemi artificiali questa profonda intuizione di uno degli uomini che aprirono la strada alla scienza moderna?

Noi crediamo di sì. La complessità di governo delle attuali organizzazioni sovrasta ormai le doti di sola intuizione dei decision maker impegnati nelle imprese e nella pubblica amministrazione. Ricorrendo a un esempio, la progettazione di una campagna di marketing in mercati complessi e imprevedibili, dove sono tuttavia disponibili molte informazioni sui comportamenti d'acquisto dei consumatori, non può prescindere dall'utilizzo di adeguati modelli di apprendimento inferenziale per la selezione dei destinatari, in modo da ottimizzare le risorse impiegate.

L'interpretazione del concetto di business intelligence che abbiamo illustrato e che intendiamo sviluppare nel corso del testo appare molto più estesa e approfondita rispetto a un'accezione riduttiva diffusa in questi anni da parte di produttori di software commerciale e periodici di area informatica. Questa visione tende a ridurre le metodologie di business intelligence a strumenti informatici di interrogazione, visualizzazione e reporting, orientati in prevalenza al controllo di gestione. Non si può negare che l'accesso tempestivo e flessibile alle informazioni offra ai decision maker un prezioso ausilio. Si tratta tuttavia di analisi di business intelligence di tipo *passivo*, nel corso delle quali chi interroga i dati ha già formulato nella sua mente un criterio di estrazione. Se si vuole che le metodologie di business intelligence siano in grado di esprimere le loro enormi potenzialità è necessario volgere lo sguardo a forme *attive* di supporto alle decisioni, basate sull'impiego di modelli matematici capaci di trasformare i dati non soltanto in *informazione* ma anche in *conoscenza*, e la conoscenza in concreto vantaggio competitivo. Riprenderemo con maggiore ampiezza la distinzione tra analisi passive e attive nel corso del capitolo 1.

Qualcuno potrebbe obiettare che soltanto strumenti semplici basati su concetti immediati e intuitivi sono in grado di rivelarsi utili in pratica. Rispondiamo citando una frase di Vladimir Vapnik, colui che più di altri ha contribuito allo sviluppo dei modelli di apprendimento induttivo: "nothing is more practical than a good theory".

Il testo si articola in tre parti, corredate da un'appendice. Abbiamo cercato di utilizzare con frequenza riferimenti a problemi reali ed esemplificazioni che rendessero più agevole la comprensione delle tematiche affrontate, mantenendo tuttavia il necessario rigore metodologico nella descrizione dei modelli matematici.

La prima parte è dedicata allo studio dei componenti di base che costituiscono un ambiente di business intelligence, all'articolazione dei processi decisionali e alle infrastrutture informative. Il capitolo 1 traccia un quadro generale delle problematiche di business intelligence, ponendo in luce i nessi con altri ambiti disciplinari. Il capitolo 2 descrive la struttura dei processi decisionali e introduce i sistemi di supporto alle decisioni, illustrandone i principali vantaggi, i fattori critici e le problematiche realizzative. Il capitolo 3 introduce i data warehouse e

i data mart, analizzando le motivazioni che hanno condotto alla loro definizione, per descrivere in seguito le analisi OLAP svolte mediante cubi multidimensionali.

La seconda parte ha carattere metodologico e presenta un'ampia rassegna dei modelli matematici di apprendimento inferenziale e dei metodi di data mining. Il capitolo 4 descrive le caratteristiche principali dei modelli matematici per le analisi di business intelligence, offrendo una breve tassonomia dei modelli più comunemente utilizzati. Il capitolo 5 introduce le tematiche di data mining, mostrandone gli obiettivi e l'articolazione in fasi. Il capitolo 6 descrive le attività che permettono di predisporre i dati per le analisi di business intelligence e data mining: la validazione, l'identificazione di anomalie, la trasformazione e la riduzione. Il capitolo 7 illustra in dettaglio l'analisi esplorativa dei dati, svolta mediante metodi grafici e indicatori statistici, per comprendere le caratteristiche degli attributi presenti in un dataset e per individuare l'intensità delle relazioni che li legano. Il capitolo 8 descrive i modelli di regressione semplice e multipla, discutendo la valutazione dei modelli di regressione, e presentando i principali criteri per la verifica di significatività e di accuratezza. Il capitolo 9 illustra i modelli per l'analisi di serie storiche, analizzando i metodi di scomposizione, i modelli di smoothing esponenziale e i modelli autoregressivi. Il capitolo 10 è dedicato ai modelli di classificazione che occupano una posizione preminente nella teoria dell'apprendimento. Dopo avere descritto i criteri di valutazione, vengono illustrati i principali metodi di classificazione: gli alberi di classificazione, i metodi bayesiani, le reti neurali, la regressione logistica e le support vector machines. Il capitolo 11 descrive le regole associative e l'algoritmo Apriori. Nel capitolo 12 vengono presentati i più noti modelli di clustering: i metodi di partizione delle K-medie e dei K-medoidi e i metodi gerarchici di agglomerazione e di divisione.

La terza parte illustra le applicazioni di data mining al marketing relazionale (capitolo 13), i modelli per la pianificazione delle reti di vendita (capitolo 13), i modelli di ottimizzazione della supply chain (capitolo 14) e i metodi analitici per la valutazione delle prestazioni (capitolo 15).

L'appendice fornisce informazioni e links relativi agli strumenti software utilizzati per svolgere le analisi di business intelligence descritte nel testo. Si è preferito utilizzare esclusivamente software *open source*, che i lettori possono liberamente scaricare dal web per svolgere gli esempi proposti. In linea con questa scelta, anche i dataset utilizzati per esemplificare gli argomenti trattati provengono in prevalenza da siti di pubblico dominio. L'appendice descrive brevemente i dataset impiegati e fornisce links a siti che contengono questi e altri dataset utili per sperimentare e confrontare i metodi di analisi.

Il volume si rivolge a tre tipologie principali di lettori. Da un lato gli studenti dei corsi di laurea magistrale a indirizzo economico-gestionale o scientifico che seguono insegnamenti relativi alle metodologie di business intelligence, ai sistemi di supporto alle decisioni, ai modelli matematici per le decisioni. Dall'altro, gli allievi dei corsi master impegnati in orientamenti di studio di natura economico-gestionale.

Infine, il testo può risultare utile a professionisti che desiderano aggiornare il proprio bagaglio di conoscenze e disporre di un riferimento sistematico e con-

creto. I lettori appartenenti a questo terzo gruppo possono essere interessati a una panoramica delle opportunità offerte dalle analisi di business intelligence, oppure a specifici aspetti metodologici e applicativi trattati nel testo, come le tecniche di data mining applicate al marketing relazionale, i modelli per la pianificazione delle reti di vendita, i modelli di ottimizzazione della supply chain e i metodi analitici per la valutazione delle prestazioni.

Presso il Politecnico di Milano, l'autore è responsabile del gruppo di ricerca *MOLD - Mathematical modeling, optimization, learning from data*, che svolge attività di ricerca metodologica sui modelli di apprendimento induttivo, di previsione, di ottimizzazione, e progetti applicativi su temi di business intelligence, marketing relazionale, logistica. Il sito del gruppo di ricerca, all'indirizzo *www.mold.polimi.it*, contiene informazioni, segnalazioni, approfondimenti, links utili e aggiornamenti.

È raro che un libro sia privo di refusi, soprattutto nella sua prima edizione, nonostante l'impegno per evitarli. Sul medesimo sito *www.mold.polimi.it* sarà quindi disponibile un'errata corrige, cui i lettori potranno cortesemente contribuire segnalando eventuali imprecisioni e refusi all'indirizzo email dell'autore: *carlo.vercellis@polimi.it*.

Voglio esprimere un ringraziamento particolare a Carlotta Orsenigo. Ha collaborato alla stesura del capitolo 10, relativo ai modelli di classificazione, ha discusso con me l'impostazione degli altri capitoli del libro, ha svolto un lavoro intelligente e assiduo nel colmare le lacune, chiarire l'esposizione, suggerire miglioramenti al testo e alle figure.

Nella stesura del libro ho beneficiato dell'esperienza didattica universitaria e post-universitaria. Un ringraziamento collettivo va dunque ai tanti studenti che attraverso le loro domande e la loro curiosità mi hanno stimolato a cercare argomentazioni più convincenti e incisive.

Molti degli esempi e dei riferimenti a problemi reali traggono spunto da progetti applicativi che ho potuto svolgere presso imprese ed enti pubblici. Sono debitore nei confronti di numerosi professionisti di alcune delle idee che ho trasferito nel libro: a tutti loro, che non posso nominare personalmente ma che certamente si riconosceranno in alcune affermazioni, un ringraziamento sentito.

Ogni refuso o imprecisione presente nel testo è mia esclusiva responsabilità.