

Analisi dei Dati ed Estrazione della Conoscenza

Gianluca Amato

Corso di Laurea in Economia Informatica
Università "G. D'Annunzio" di Chieti-Pescara

Anno Accademico 2004/05

Di cosa ci occupiamo (1)

- Esplosione dei dati
 - La società produce una grande quantità di dati, grazie anche allo sviluppo di sistemi automatici di raccolta e immagazzinamento degli stessi
 - In un supermercato, è possibile memorizzare con un database tutti gli scontrini prodotti e le informazioni sulle offerte utilizzate dai clienti.
 - I dati grezzi sono inutili
 - Cosa ne facciamo di tutti questi scontrini memorizzati?
 - Stiamo affogando nei **dati**, ma c'è carenza di **informazioni**
 - Una informazione utile sarebbe sapere quali prodotti vengono acquistati in coppia più spesso, in modo da spostare la loro collocazione in maniera opportuna.

Di cosa ci occupiamo (2)

- Soluzioni
 - **Data Warehouse**: raccolta organica di informazioni da più sorgenti di dati anche eterogenee (database aziendali, database di altre aziende, internet)
 - **OLAP**: on-line analytical processing.. sistema di gestione per basi di dati ottimizzato per query complesse
 - specializzato per funzionalità di aggregazione (somme, conteggi) analoghe a quelle di SQL
 - interfaccia che consente velocemente e in maniera intuitiva di vedere i dati sotto diverse angolazioni (attributi su cui raggruppare, granularità degli attributi)
 - **Data Mining**: estrazione di informazioni interessanti (regole, associazioni, vincoli) dai dati

Programma del Corso (molto di massima)

- Introduzione
- Data Warehouse e OLAP
- Preparazione dei dati
 - i dati provenienti dai database sono spesso inconsistenti
 - eliminazione delle inconsistenze, discretizzazione, normalizzazione
- Estrazione della conoscenza
 - Regressione, Classificazione, Associazione, Raggruppamento
- Valutazione della conoscenza
 - abbiamo estratto informazioni interessanti?
 - possiamo utilizzarla per fare previsioni?

Impostazione del Corso

- Lezioni teoriche
- Esercitazioni in laboratorio
 - utilizziamo il sistema Weka (Waikato Environment for Knowledge Analysis)
<http://www.cs.waikato.ac.nz/ml/weka/>
 - da un certo punto in poi del corso (prima bisogna introdurre le conoscenze fondamentali)
- Esame
 - non ho ancora deciso...
 - ...ma probabilmente ci sarà una prova scritta.

Dove studiare (1)

- Lucidi del corso
- Testo di riferimento (che seguirò molto da vicino)
 - J. Han, M. Kamber
Data Mining: Concepts and Techniques
Morgan Kaufmann
- Ma come? In inglese?
 - I testi in italiano che ho trovato hanno un taglio marcatamente statistico e poco informatico.
 - Avere dimestichezza con l'inglese tecnico è fondamentale per chi vuole lavorare in un settore tecnologico.

Dove studiare (2)

- Altri testi di consultazione
 - Ian Witten, Eibe Frank
Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
Morgan Kaufmann
 - Paolo Giudici
Data Mining: metodi statistici per le applicazioni aziendali
McGraw Hill

Dove trovarmi?

- Al solito posto per chi lo sa...
 - per chi non lo sa:

Dipartimento di Scienze
sezione staccata di viale Pindaro 87
telefono: 085-4546425
email: amato@sci.unich.it
- Orario di ricevimento:
 - giovedì ore 16:00 - 18:00

Ringraziamenti

- Questi lucidi sono stati prodotti saccheggiando le presentazioni associate ai libri di testo di cui si è parlato prima.
- Le presentazioni originali si possono trovare nei seguenti siti web:
 - [il sito web del libro di Han e Kamber](#)
 - [il sito web del libro di Frank e Witten](#)

Introduzione

Data Mining e Knowledge Discovery in Databases

Cos'è il Data Mining?

- Col termine data mining si intende:
 - **estrazione di informazione interessante dai dati contenuti in una (potenzialmente ampia) base di dati.**
- Cosa vuol dire **informazione**?
 - con informazione intendiamo l'insieme delle regolarità e dei modelli (**pattern**) presenti implicitamente nei dati.
 - vedremo in seguito vari modi di esprimere questa informazione.

Informazione interessante

- Cosa vuol dire **interessante**? In prima analisi
 - **nuova**: non è qualcosa di già noto o conoscenza comune
 - oppure anche **attesa**, se si tratta di convalidare una ipotesi fatta a priori.
 - **implicita**: presente nei dati analizzati, ma non immediatamente accessibile;
 - **potenzialmente utile**: può essere utilizzata per prendere delle decisioni;
 - **comprensibile agli uomini**: la forma in cui la conoscenza è estratta deve essere interpretabile facilmente dagli esseri umani;

Cosa c'è di nuovo? (1)

- Gli uomini sono andati alla scoperta di regolarità da quando la vita umana ha avuto inizio
 - i cacciatori cercano regolarità nelle migrazioni degli animali
 - i politici cercano regolarità nell'opinione degli elettori
 - quale azione posso intraprendere per guadagnare il 5% dei voti?
 - un fisico cerca regolarità nei fenomeni naturali
 - scopre così che una mela che si stacca dall'albero viene attratta sulla terra

Cosa c'è di nuovo? (2)

- Nel data mining i dati sono memorizzati in forma elettronica e la ricerca è automatica o semi-automatica
 - neanche questo è particolarmente nuovo
 - economisti, statistici hanno sempre lavorato all'idea che regolarità potessero essere trovate con mezzi automatici
- Quello che è nuovo è l'enorme aumento delle opportunità per applicare questa ricerca di informazioni
 - causata, come abbiamo detto, dalla crescita delle basi di dati negli anni recenti
 - da cui l'accento posto, nel data mining, alla ricerca di informazione all'interno di un database

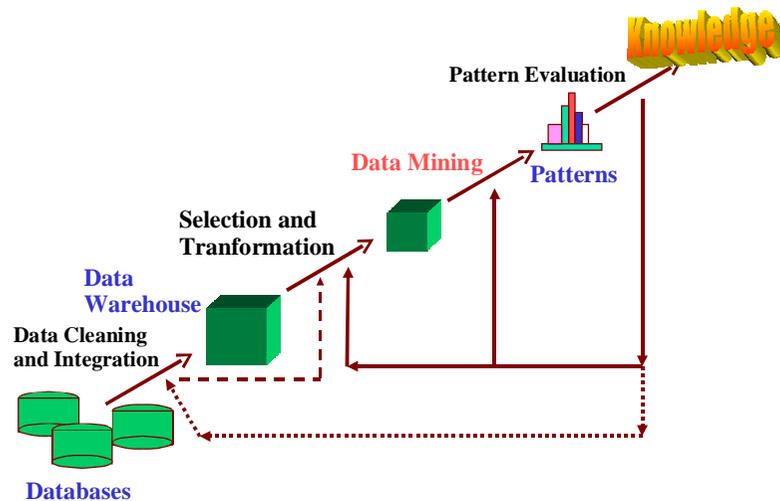
Cosa non è il Data Mining

- Sistemi esperti
 - i sistemi esperti sono sistemi che utilizzano la conoscenza (tipicamente fornita da un esperto umano) per analizzare dei dati e trarre nuove conclusioni (ad esempio per diagnosi mediche)
 - nei sistemi data mining la conoscenza non è data a priori ma inferita analizzando i dati (ad esempio analizzando le cartelle cliniche dei pazienti precedenti)
 - i sistemi esperti ve li fa Andrea Roli

Altri termini per Data Mining

- Il termine è etimologicamente errato.. bisognerebbe parlare di “**knowledge mining**”
- Molti termini hanno dei significati uguali o simili a “data mining”
 - knowledge mining from database
 - knowledge extraction (estrazione della conoscenza)
 - data/pattern analysis
 - data archeology
 - **knowledge discovery in database** (KDD)
- Per alcuni, però, il data mining è solo uno dei passi nel processo di “knowledge discovery in database”

Knowledge Discovery in Database



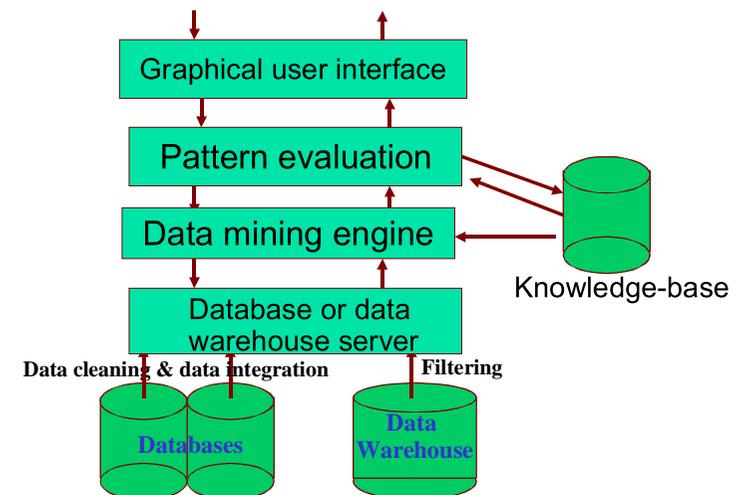
Passi per un processo di KDD /1

- Acquisire informazioni sul dominio applicativo
- Pulire i dati a disposizione (**data cleaning**)
 - può anche rivelarsi l'operazione più faticosa!!!
- Integrare i dati provenienti da sorgenti diverse (**data integration**)
- Selezionare i dati di interesse (**data selection**)
- Trasformare i dati (**data transformation and reduction**)
 - eliminare attributi ridondanti
 - discretizzare i dati numerici
 - ...ed altro

Passi per un processo di KDD /2

- Scegliere il tipo di analisi da effettuare
 - classificazione, associazione, clustering, regressione lineare, etc...
- Scegliere l'algoritmo da utilizzare
 - lo stesso tipo di analisi può essere svolta da algoritmi diversi con risultati diversi
- **Data Mining!**
- Valutazione dei risultati ottenuti (**pattern evaluation**)
- Visualizzazione dei risultati, eliminazione pattern ridondanti
- Uso della conoscenza acquisita

Architettura di un sistema di data mining



Introduzione

Funzionalità di un software per Data Mining

Funzionalità del Data Mining

- I pattern ottenuti da un processo di data mining possono essere di due tipi: **descrittivi** e **predittivi**
 - descrittivi: si ottiene una caratterizzazione delle proprietà dei dati selezionati per l'analisi
 - predittivi: si ottiene un sistema che è in grado, sulla base dei dati attuali, di fare previsioni sui dati futuri.
- I pattern inoltre possono essere degli oggetti più o meno comprensibili.
 - le reti neurali possono essere addestrate per effettuare delle previsioni, ma il loro funzionamento è oscuro
 - le regole associative, che vedremo tra breve, hanno invece un significato intuitivo chiarissimo
 - ci interesseranno di più pattern facilmente comprensibili, detti **pattern strutturali**

Concept Description (1)

- si distingue in
 - **concept characterization**: consiste nel riassumere le caratteristiche generali di un insieme di dati
 - ad esempio, caratterizzare i clienti che hanno speso più di 1000€ presso la AllElectronics nell'ultimo anno
 - il risultato può essere un profilo di utente dai 40 ai 50 anni, occupato, non sposato.
 - **concept comparison** (o **discrimination**): fornisce una descrizione che confronta due o più insiemi di dati.
 - ad esempio, caratterizzare i clienti che comprano regolarmente alla AllElectronics contrapposti a quelli che comprano di rado.

Concept Description (2)

- Supponiamo di avere i seguenti dati, che rappresentano i laureati in una università americana:

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...
Removed	Retained	Sci.Eng, Bus	Country	Age range	City	Removed	Excl, VG,...

- Otteniamo

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...

- Alcuni attributi sono stati rimossi, altri generalizzati (non si conta il corso di laurea ma solo la facoltà)

Analisi delle Associazioni

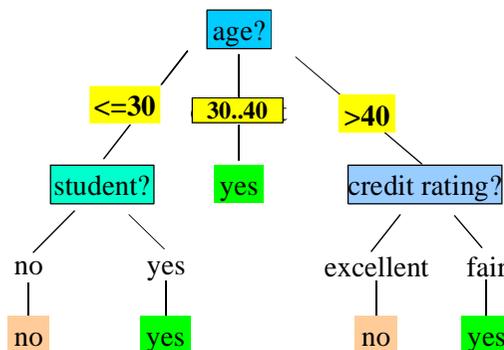
- Consiste nella scoperta di **regole associative**
 - ad esempio, se stiamo analizzando quali oggetti vengono comprati assieme più frequentemente alla AllElectronics, possiamo ottenere
 - `contains(T,"computer") => contains(T,"software")`
[support = 1%, confidence = 50%]
- Il **supporto** è la percentuale di transazioni che hanno sia un computer che uno scontrino
- La **confidenza** è la percentuale di transazioni che hanno computer e software rispetto a tutte le transazioni che hanno un computer.

Classificazione (1)

- La classificazione è il processo che consiste nel trovare un modello che descrive delle classi di dati, allo scopo di predire il valore della classe su dati sconosciuti.
 - alla AllElectronics vogliono classificare i clienti in coloro che con alta probabilità acquistano computer e coloro che non lo fanno.
- Si distingue una fase di addestramento (**training**) sui dati che si conoscono e una fase in cui, quando nuovi dati arrivano, si utilizza il modello prodotto per capire la classe del nuovo dato.

Classificazione (2)

- Risultato espresso come **Albero di Decisione**



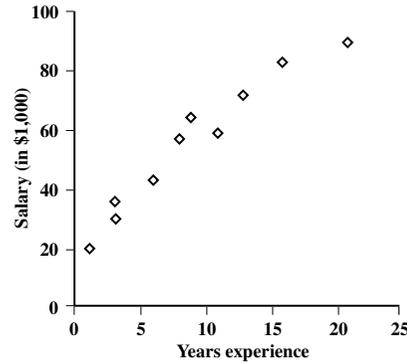
Classificazione (3)

- Risultato espresso come **Regole di decisione**

```
IF age = "<=30" AND student = "no" THEN buys_computer = "no"
IF age = "<=30" AND student = "yes" THEN buys_computer = "yes"
IF age = "31..40" THEN buys_computer = "yes"
IF age = ">40" AND credit_rating = "excellent" THEN buys_computer = "yes"
IF age = ">40" AND credit_rating = "fair" THEN buys_computer = "no"
```

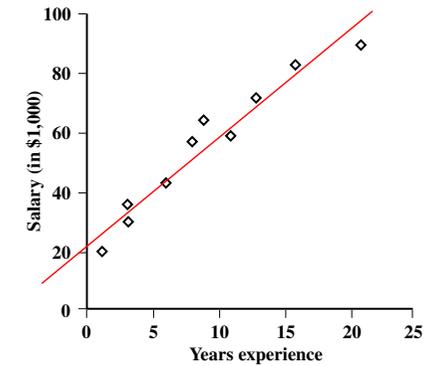
Predizione (1)

- Come la classificazione ma quello che si vuole ottenere è un dato continuo (un numero reale)
 - si hanno a disposizione dei dati relativi al salario di un laureato in base al numero di anni di esperienza..
 - si vuole trovare un modello per determinare il valore del salario per gli altri casi



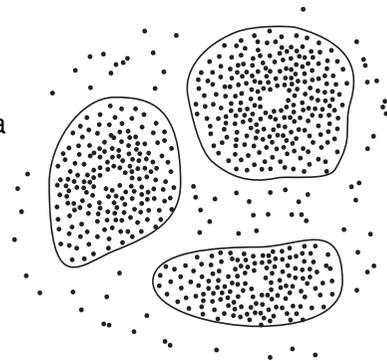
Predizione (2)

- Un metodo possibile: la **regressione lineare**
 - che dovrete conoscere dal corso di statistica
 - $\text{salary} = 23.6 * \text{years} + 3.5$



Analisi di Raggruppamento (clustering)

- A differenza di classificazione e predizione il clustering analizza i dati senza consultare nessuna informazione nota sulla classe.
 - il clustering genera automaticamente classi interessanti
 - gli oggetti vengono messi nella classe giusta in base alla similarità con altri oggetti



Otteniamo pattern interessanti?

- Raffiniamo quanto detto nella [definizione di data mining](#).
- Una analisi di dati può produrre migliaia di pattern.. occorre scegliere quelli interessanti
 - ci sono alcuni parametri **oggettivi**, basati su statistiche e struttura del pattern come **supporto** e **confidenza**,
 - ma in ultima analisi l'utilità di un pattern è qualcosa di puramente **soggettivo**

Possiamo ottenere tutti e soli i risultati interessanti?

- Non esistono metodi generali
- E' possibile focalizzare la ricerca di risultati interessanti fornendo dei parametri oggettivi da rispettare (ad esempio confidenza e supporto minimo)
- Occorre comunque una certa quantità di tentativi per ottenere dei pattern veramente utili.

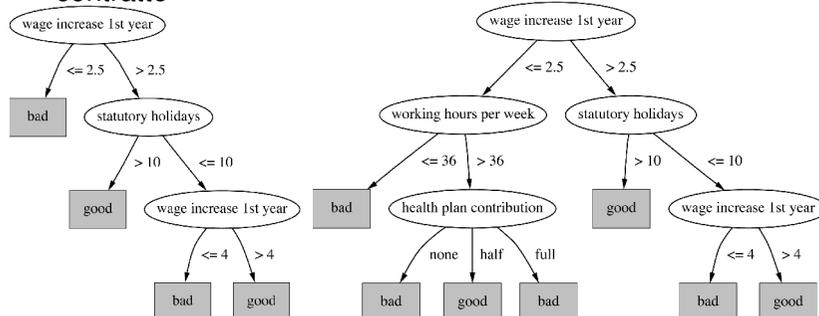
Validità pattern e overfitting (1)

- Il seguente insieme di dati riassume il risultato delle contrattazioni per i contratti collettivi di lavoro in Canada nell'anno 1987-88.

attributo	1	2	3	...
durata	1	2	3	
incremento retrib.(I anno)	2.00%	4.00%	4.30%	
incremento retrib. (II anno)	?	5.00%	4.40%	
incremento retrib. (III anno)	?	?	?	
ore di lavoro per settimana	28	35	38	
giorni di vacanza	11	15	12	
contributi per assicuraz. medica	nessuno	?	completi	
.....				
accettabilità del contratto	no	sì	sì	

Validità pattern e overfitting (2)

- Ecco due possibili alberi di classificazione per la bontà del contratto



- Il secondo albero è più preciso nei dati usati per l'addestramento, ma si rivela meno accurato quando viene applicato ad altri dati (**overfitting**)

Introduzione

Esempi Applicativi

Concessione di prestiti

- Una banca che concede un prestito vorrebbe essere sicura che questo verrà ripagato
- Di solito si procede a far riempire un questionario e a calcolare un parametro numerico di affidabilità. Nel 90% dei casi questo porta ad una decisione.
- E il restante 10%?
 - non concedere il prestito?
 - intervista con un membro della banca?
 - il 50% degli intervistati a cui viene concesso il prestito non paga.
 - un sistema di classificazione che produce un insieme di regole decisionali!
 - la percentuale di non-paganti scende al 25%

Marketing

- Market Basket Analysis
 - uso delle tecniche di associazione per trovare gruppi di prodotti che vengono acquistati insieme
 - un ulteriore valore aggiunto se riusciamo a distinguere i vari clienti
 - da qui la diffusione di carte fedeltà
- Fedeltà del cliente
 - una banca può individuare i clienti che hanno più probabilità di cambiare banca
 - si analizzano i comportamenti del cliente alla ricerca di cambiamenti nel suo modo di agire
 - tecniche simili consentono di riconoscere le frodi effettuate con le carte di credito

Diagnosi

- Una industria ha migliaia di apparati elettromeccanici. Quando un guasto si verifica, il tipo di guasto viene identificato da una serie di sensori che misurano le vibrazioni in punti diversi dell'impianto
- La identificazione avveniva con un esperto umano.. come automatizzarla?
 - un sistema esperto può andare bene, ma ogni apparecchiatura ha bisogno di regole diverse..
 - a un sistema di data mining sono stati forniti dati su circa 300 malfunzionamenti per l'addestramento. Il risultato, **dopo una serie di tentativi**, è stato un insieme di regole:
 - comprensibili dall'esperto
 - che mettevano in luce nuove relazioni prima nascoste

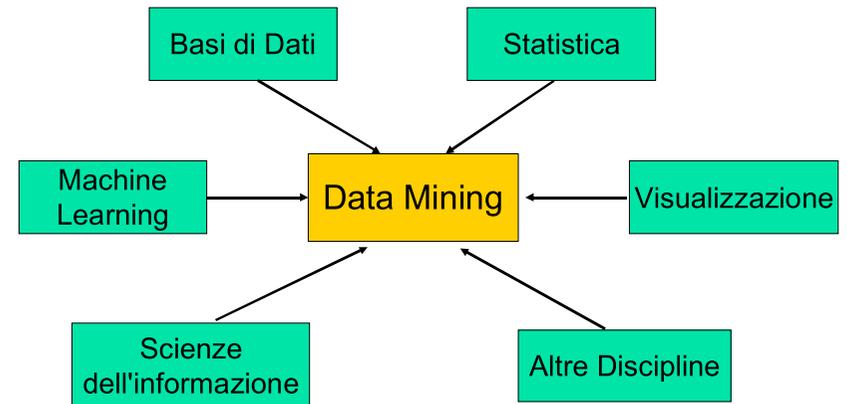
È veramente la manna dal cielo?

- nell'esempio precedente abbiamo evidenziato **“dopo una serie di tentativi”**
- come tutte le tecnologie emergenti, il data mining è circondato da molta enfasi...
- ...ma una analisi alla cieca difficilmente produce un risultato utile
- **La bontà delle analisi prodotte da un processo di data mining dipende dalle capacità degli esseri umani che guidano questo processo!**
 - per questo serve una conoscenza dei metodi di data mining, degli algoritmi utilizzati e dei risultati che è possibile ottenere

Introduzione

Data Mining e altre discipline

Data Mining: confluenza di varie discipline



Data Mining e Machine Learning

- Il machine learning (apprendimento automatico) è un filone di studi, collegato all'informatica e all'intelligenza artificiale, che si occupa di ricavare delle regolarità dai dati.
- L'apprendimento automatico è dunque una delle basi tecniche del data mining
- I metodi di data mining differiscono da quelli di apprendimento automatico puro in quanto:
 - focalizzati ad estrarre informazione dai database
 - si occupano tipicamente dell'analisi di grandi moli di dati, quindi sono di interesse per il data mining soltanto gli algoritmi di machine learning scalabili

Data Mining e Statistica (1)

- La statistica si è sempre occupata di metodologie per l'analisi dei dati: recentemente molti statistici si stanno interessando al data mining.
- Dunque anche la statistica fornisce basi tecniche al data mining, sia per il processo di costruzione di pattern che per il processo di verifica della validità di quest'ultimi.
- I metodi di data mining differiscono da quelli puramente statistici perchè:
 - focalizzati ad estrarre informazione dai database
 - si occupano tipicamente dell'analisi di grandi moli di dati
 - sì, sono gli stessi punti del lucido precedente...
 - i database su cui operano non contengono dati raccolti apposta: spesso contengono dati raccolti per altri scopi.

Data Mining e Statistica (2)

- Il termine data mining è stato per lungo tempo inteso dagli statistici con una accezione negativa, come sinonimo di “data fishing” o “data dredging”.
- Queste le critiche
 - nel data mining non vi è un unico modello di riferimento, ma numerosi modelli in competizione. E` sempre possibile trovare un modello complesso che si adatti bene ai dati
 - l'abbondanza di dati può portare a trovare dei pattern in realtà inesistenti

Data Mining e Statistica (3)

- Tuttavia
 - le moderne tecnologie di data mining pongono molta attenzione alla generalizzabilità dei risultati, penalizzando anche i modelli più complessi..
 - molti risultati di interesse per un'applicazione non sono noti a priori, mentre i metodi statistici hanno di solito bisogno di una ipotesi di ricerca data a priori.
- Alcuni autori distinguono due approcci al data mining:
 - analisi bottom-up (esplorative): cercano di ampliare la conoscenza su nuovi aspetti di un fenomeno
 - analisi top-down (confermative): mirano a confermare o smentire fatti ipotizzati
 - qui si utilizzano tipicamente metodi statistici

Data Mining ed Etica (1)

- L'uso del data mining ha serie implicazioni etiche.
- Quando applicato a persone, il data mining è usato per **discriminare**:
 - chi ottiene il prestito? certe discriminazioni (ad esempio in base a sesso e razza) sono eticamente poco corrette o anche illegali
- Dipende dall'applicazione
 - le stesse informazioni per scopi medici sono ok
- Alcuni attributi possono essere correlati ad informazioni problematiche
 - ad esempio, il luogo di residenza può essere correlato al gruppo etnico

Data Mining ed Etica (2)

- Domande importanti che nascono nelle applicazioni
 - chi ha il diritto di accedere ai dati?
 - per che scopo erano stati raccolti i dati?
- Tutte le analisi dovrebbero avvenire con il consenso esplicito delle persone coinvolte

Riassumendo

- data mining: scoperta di informazione interessante da basi di dati
- KDD: processo di analisi dei dati che include
 - data cleaning, integration, selection and transformation,
 - data mining
 - pattern evaluation and visualization
- funzionalità del data mining: classificazione, descrizione di concetti, associazioni, clustering, previsioni..
- applicazioni del data mining
- relazione del datamining con altre discipline e problemi di etica