Analisi dei Dati ed Estrazione della Conoscenza

Gianluca Amato

Corso di Laurea Specialistica in Economia Informatica
Università “G. D'Annunzio” di Chieti-Pescara
ultimo aggiornamto: 16/03/09

Di cosa ci occupiamo (1)

- Esplosione dei dati
  - La società produce una grande quantità di dati, grazie anche allo sviluppo si sistemi automatici di raccolta e immagazzinamento
    - **Azienda**: transazioni (supermercato), e-commerce
    - **Scienza**: simulazioni scientifiche, osservazioni astronomiche
    - **Società**: YouTube, reti sociali (FaceBook)
  - I dati grezzi sono inutili
    - Cosa ne facciamo di tutti questi dati a un così alto livello di dettaglio?
  - 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
    - Oppure individuare i "leader" in una rete sociale, le persone più influenti che possono condizionare le scelte degli altri (si pensi a campagne di marketing)

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, ovvero un sistema di gestione per basi di dati ottimizzato per query complesse
    - specializzato per funzionalità di aggregazione (somme, conteggi) analoghe a quelle di SQL
    - dotato di 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
    - sarà questo l'argomento principale del nostro corso!

Programma del Corso

- Introduzione
- Cenni di statistica descrittiva
- Preparazione dei dati
- Data Warehouse e OLAP
- Estrazione della conoscenza
  - Regressione, Classificazione, Associazione, Raggruppamento
- Valutazione della conoscenza
  - abbiamo estratto informazioni interessanti?
  - possiamo utilizzarla per fare previsioni?

Impostazione del Corso

- Parte teorica
  - lezioni
  - esercitazioni
- Parte pratica
  - il sistema Weka
    (Waikato Environment for Knowledge Analysis)
    http://www.cs.waikato.ac.nz/ml/weka/
  - il linguaggio R
    http://www.r-project.org/

Dove studiare (1)

- Testi adottati
  - J. Han, M. Kamber
    Data Mining: Concepts and Techniques (2nd edition)
    Morgan Kaufmann
  - Barbara Pacini, Meri Raggi
    Statistica per l'analisi operativa dei dati
    Carocci editore
- Ma come? Un libro è 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 (2nd edition)
    Morgan Kaufmann

Dove trovarmi?

- Al solito posto per chi lo sa...
  - Dipartimento di Scienze
    sezione staccata di viale Pindaro 87
    telefono: 085-453-7686
    email: amato@sci.unich.it
    home page: http://sci138.sci.unich.it

- Orario di ricevimento:
  - Martedì 10:00-12:30
  - potrebbe cambiare in futuro.. consultare la mia home page per eventuali aggiornamenti

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

Cosa si intende per Data Mining?

Cosa si intender per 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 presenti implicitamente nei dati.
  - vedremo in seguito vari modi di rappresentare questa informazione.
  - i risultati del processo di estrazione di informazione prendono il nome di pattern o di modelli.

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

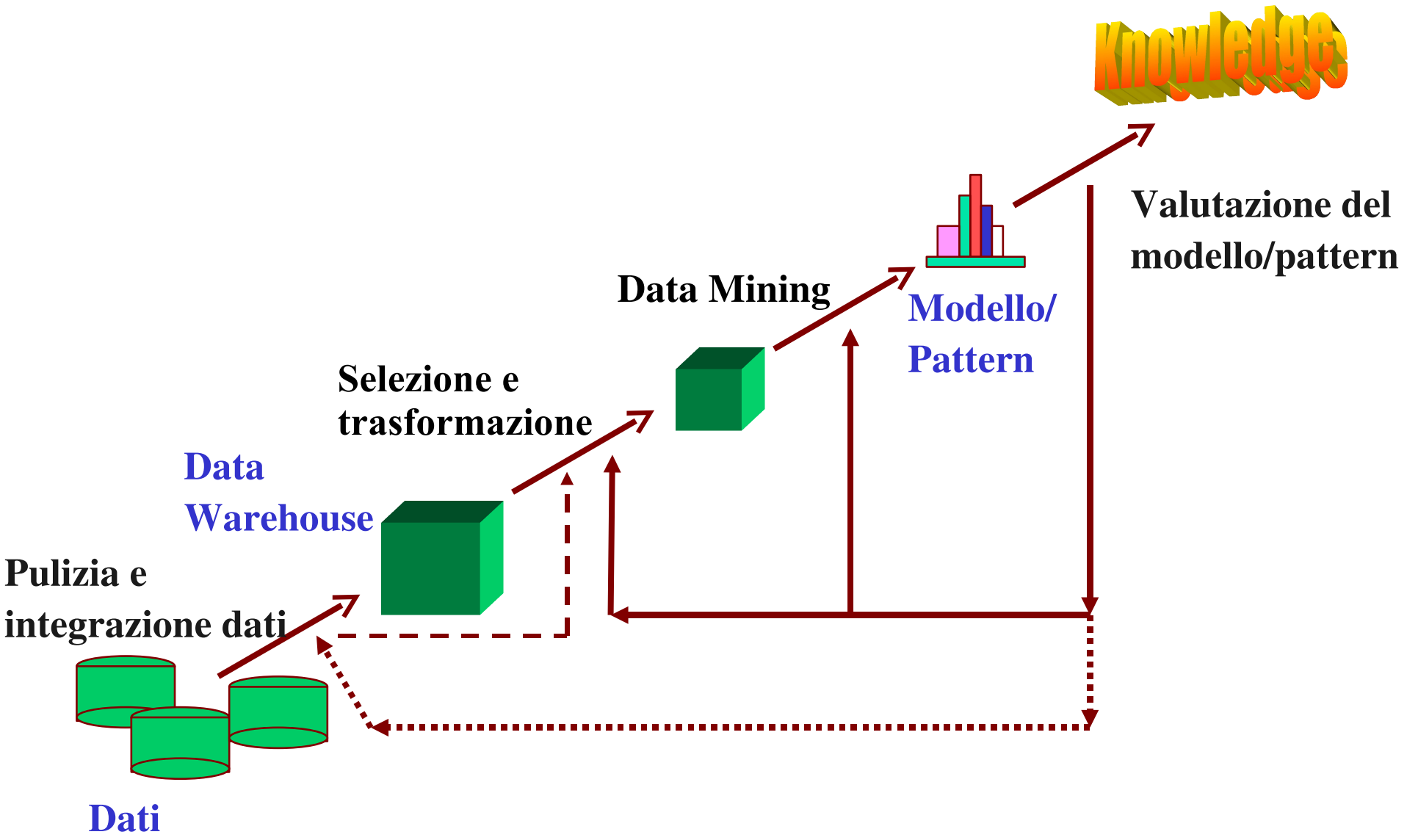
Cosa c'è di nuovo? (2)

- Nel data mining i dati sono memorizzati in forma elettronica e la ricerca è automatica o semi-automatica
  - neanche questo è particolarmente nuovo
  - gli 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, da molti esperti di data mining, alla ricerca di informazione all'interno di database di vaste dimensioni

Sinonimi 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 del processo di "Knowledge Discovery in Database"

Knowledge Discovery in Database



Passi per il 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 e reduction)
  - eliminare attributi ridondanti
  - discretizzare i dati numerici
  - ...ed altro

Passi per il 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/model evaluation)
- Visualizzazione dei risultati, eliminazione pattern ridondanti
- Uso della conoscenza acquisita

Tipi di analisi e di modelli

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

Descrizione di concetti (1)

- si distingue in
  - caratterizzazione di concetti: 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.
  - confronto di concetti: 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.

Descrizione di concetti (2)

- Supponiamo di avere i seguenti dati, che rappresentano i laurati 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
    - contiene(A,"computer") => contiene(A,"software")
      [supporto = 1%, confidenza = 50%]

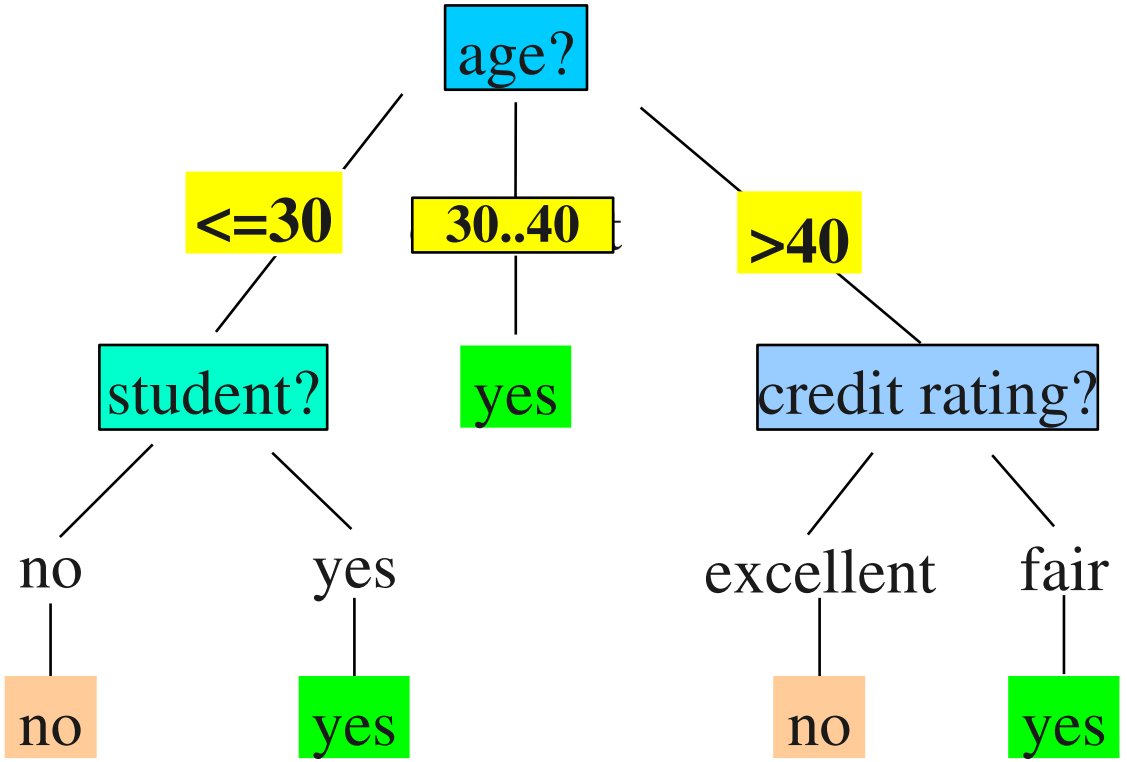
      dove A è una variabile che rappresenta un acquisto.
- Il supporto è la percentuale di transazioni che hanno sia un computer che uno scontrino
- La confidenza è la percentuale di transazioni che hanno computer e software rispetto a tutte le transazioni che hanno un computer.

Classificazione (1)

- La classificazione è il processo che consiste nel trovare un modello che descrive delle classi di dati, allo scopo di predirre 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



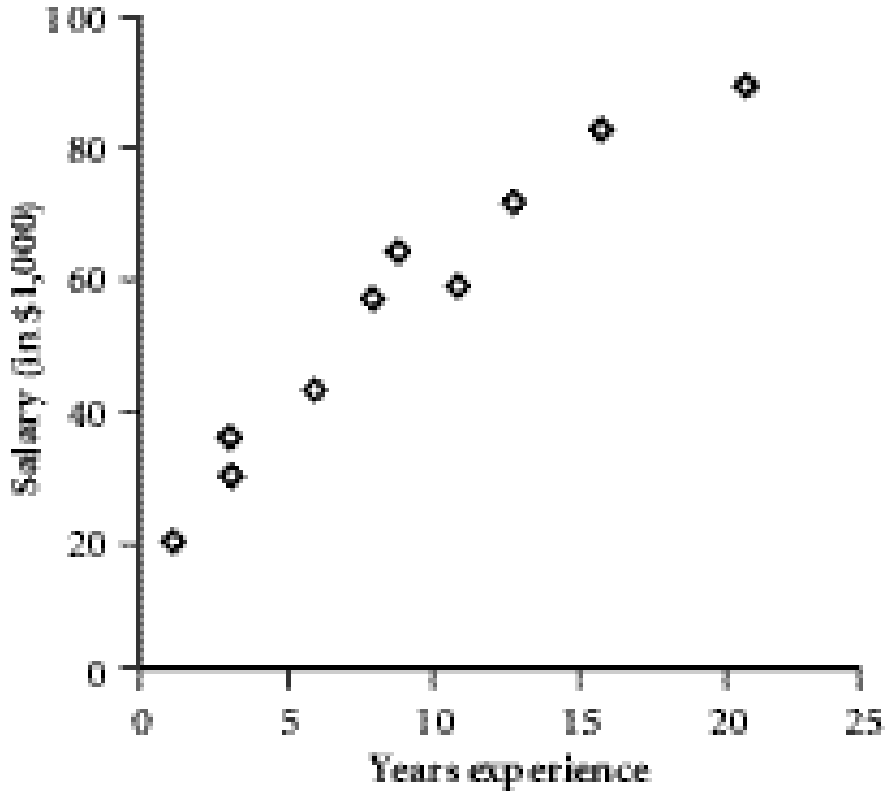
Classificazione (3)

- 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*"

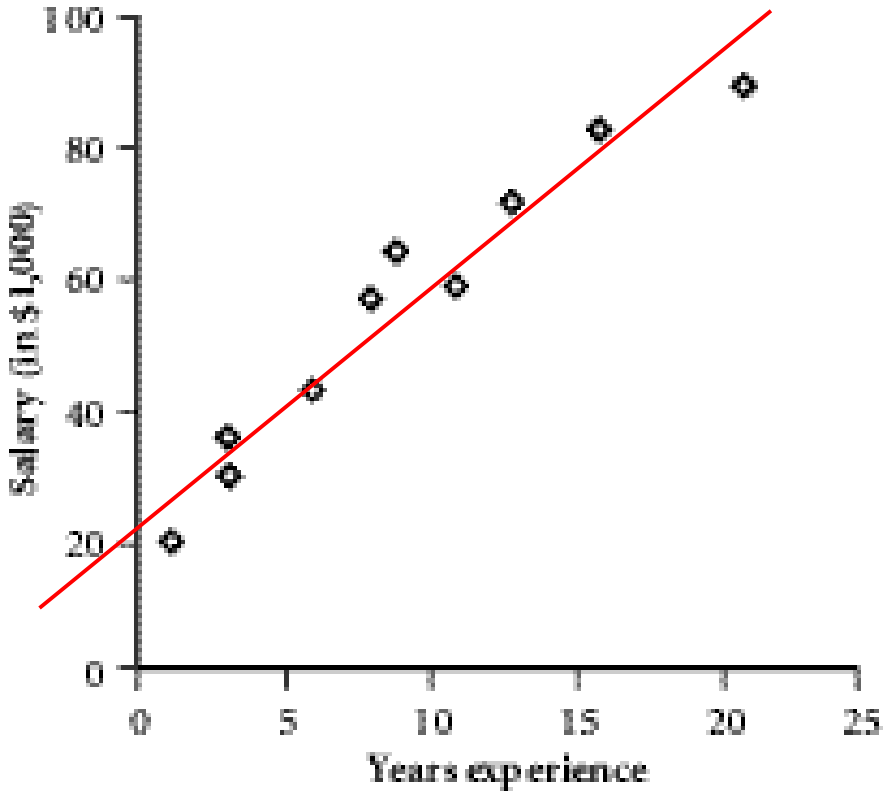
Predizione (1)

- 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..



Predizione (2)

- Si vuole trovare un modello per determinare il valore del salario per gli altri casi
- Un metodo possibile: la regressione lineare
  - Salary = 4.6 * years + 22.5

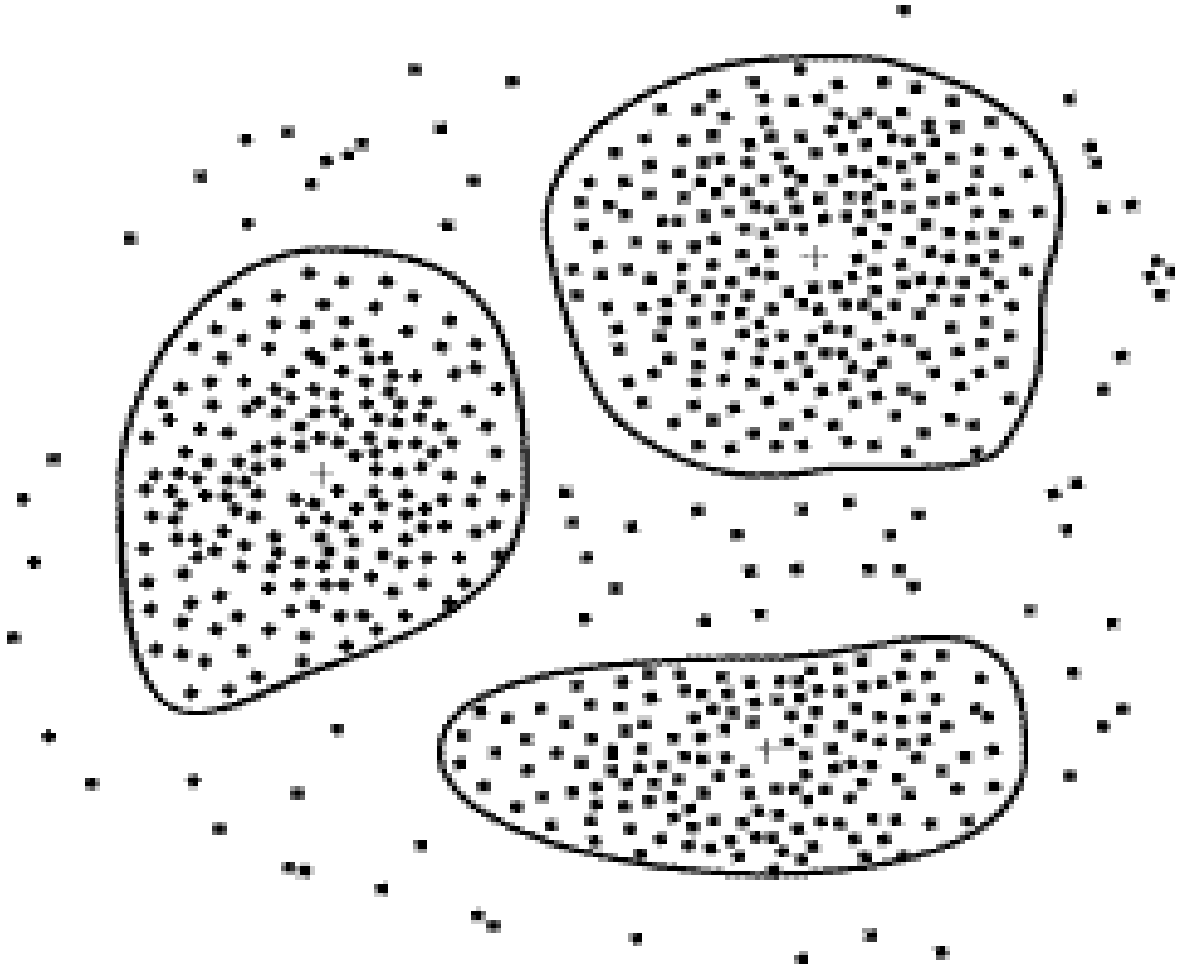


Analisi di Raggruppamento

- L'analisi di raggruppamento (clustering) analizza i dati senza consultare nessuna informazione nota sulla classe.
  - le analisi di raggruppamento generano automaticamente classi interessanti
  - gli oggetti vengono messi nella classe giusta in base alla similarità con altri oggetti
- Esempio: segmentazione dei clienti della AllElectronics per adottare politiche di marketing mirate.

Analisi di Raggruppamento (2)

- Clienti della AllElectronics divisi per posizione geografica



Riconoscimento di Outlier

- Consiste nella ricerca di di oggetti che si discostano dal comportamento o modello dei dati
  - tipicamente gli algoritmi di data-mining sono progettati per ignorare gli outlier, considerati a stregua di rumore
  - talvolta, tuttavia, la ricerca degli outlier è lo scopo principale dell'algoritmo.
- Esempio: frodi con carta di credito
  - acquisti estremamente costosi rispetto a quelli abituali
  - variazione del luogo degli acquisti, della loro tipologia o frequenza

Pattern interessanti?

- Una analisi di dati può produrre migliaia di pattern.. occorre scegliere quelli interessanti
  - ci sono alcuni parametri oggettivi, basati su statistiche o sulla struttura del pattern
    - ad esempio i valori di supporto e confidenza per l'analisi di associazione
  - ma in ultima analisi l'utilità di un pattern è qualcosa di puramente soggettivo
- Come indirizzare l'analisi su quelli interessanti?
  - non c'è un metodo generale
  - è possibile focalizzare la ricerca di risultati interessanti fornendo dei parametri oggettivi da rispettare
    - ad esempio fornendo una confidenza e supporto minimo
  - occorrono comunque una certa quantità di tentativi per ottenere dei pattern veramente utili

Sorgenti dei dati

Database relazionali

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

items_sold

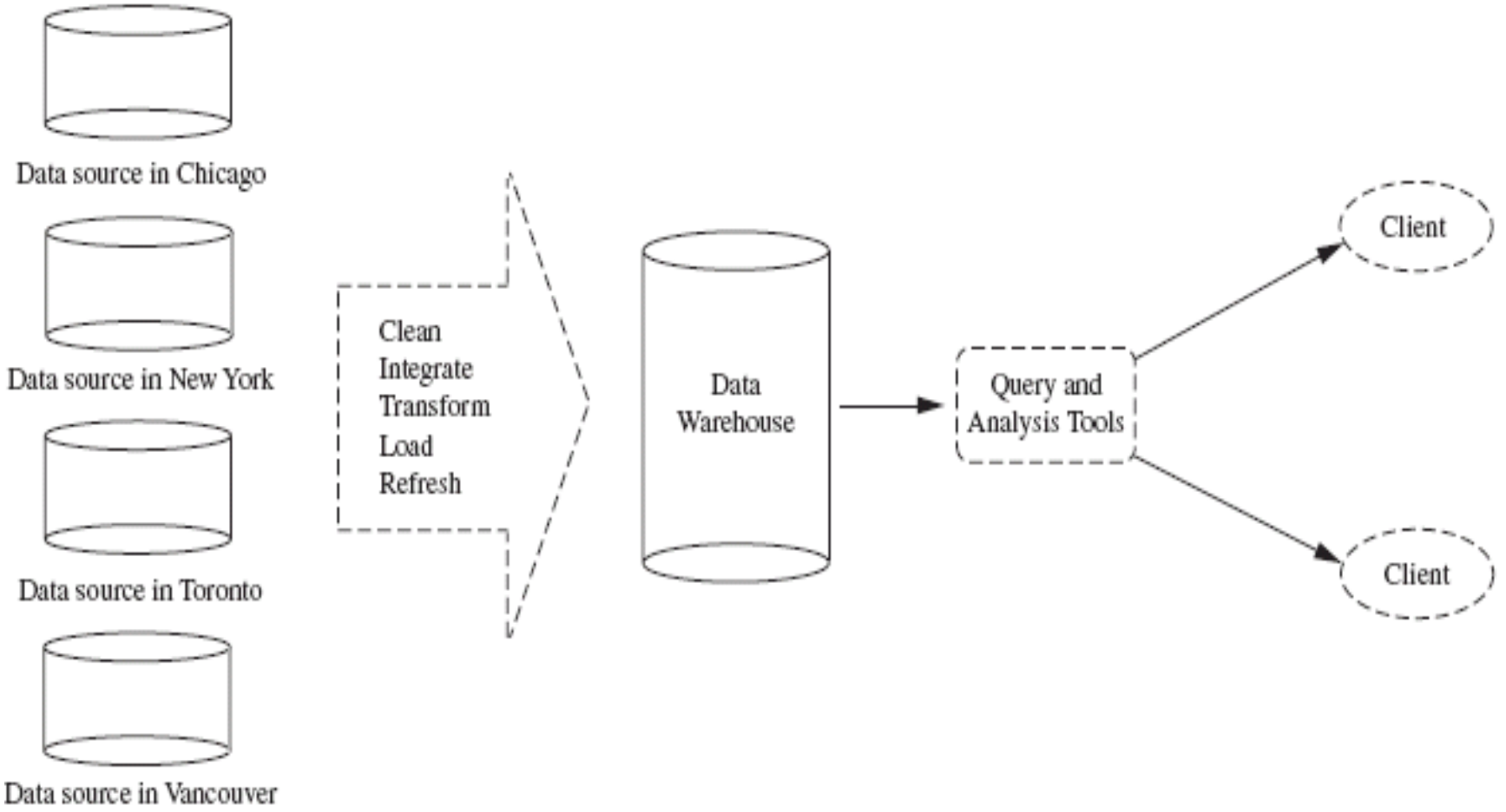
trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

works_at

empl_ID	*branch_ID*
E55	B1
. . .	. . .

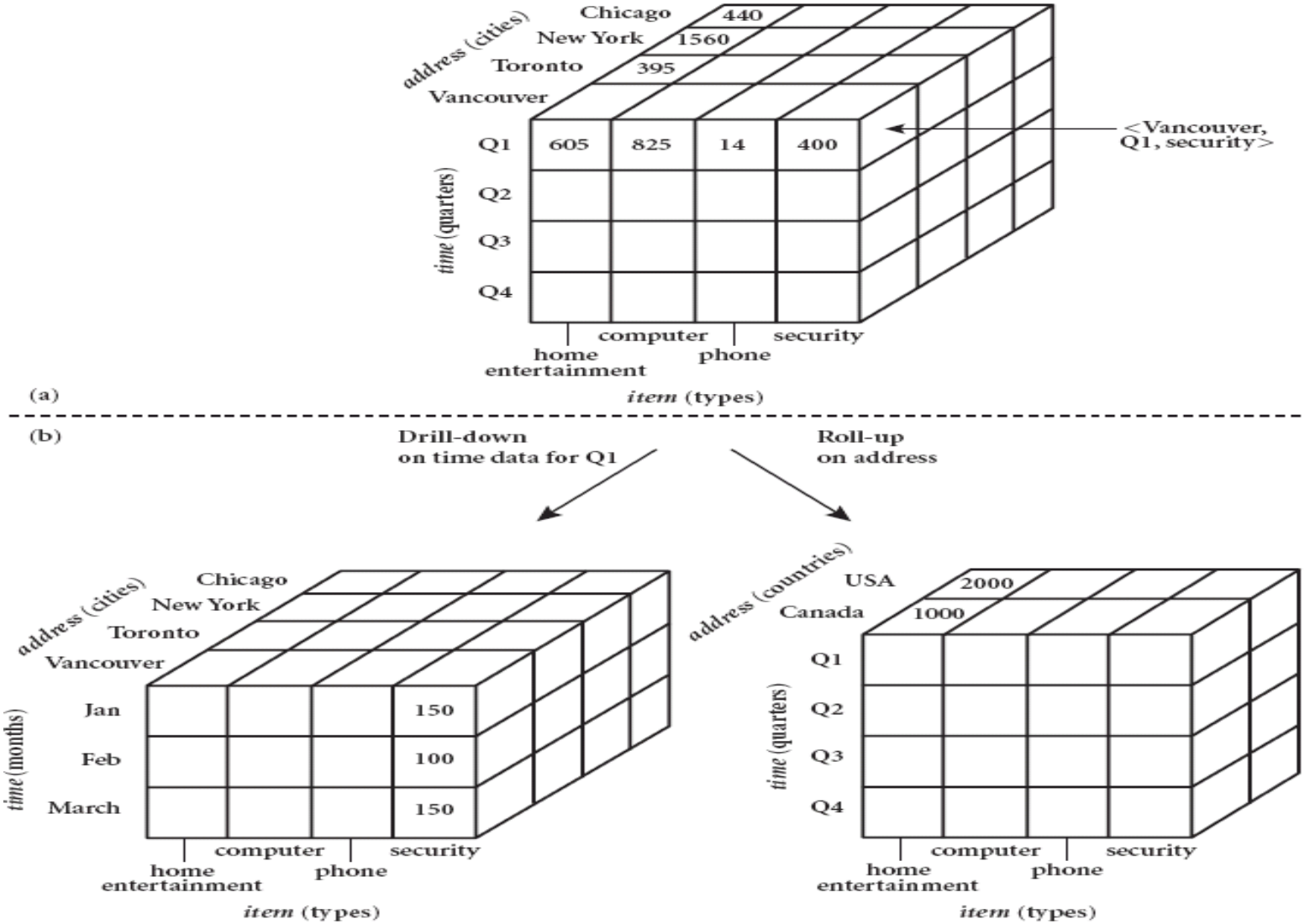
Data Warehouse (1)

- Dati integrati da sorgenti diverse



Data Warehouse (2)

- Dati aggregati



Database transazionali

- Ogni riga è una transazione che contiene uno o più oggetti

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
. . .	. . .

Altre sorgenti (1)

- Database a oggetti-relazionali
  - Caratteristiche
    - Esistono relazioni di ereditarietà tra le tabelle
    - Le righe delle tabelle sono oggetti che rispondono a metodi
    - Possono rappresentare più facilmente oggetti complessi
  - Servono tecniche speciali per trattare oggetti e tipi di dati complessi
- Database temporali
  - Esempi
    - Dati dell'andamento della borsa, affluenza di clienti in una banca
  - Esempi di analisi:
    - Predizione dell'andamento della borsa
    - Predizione dell'andamento dell'afflusso di clienti in una banca, in modo da razionalizzare l'orario lavorativo dei dipendenti

Altre sorgenti (2)

- Database spaziali
  - Esempio principale: GIS (Geographic Information System)
    - Mappe del territorio con informazioni su vari fenomeni sociali, urbanistici, ambientali
  - Esempi di analisi
    - descrivere le caratteristiche delle abitazioni che si trovano vicino a un parco
    - determinare come varia il tasso di povertà nelle città, sulla base della loro vicinanza alle autostrade.

Altre sorgenti (3)

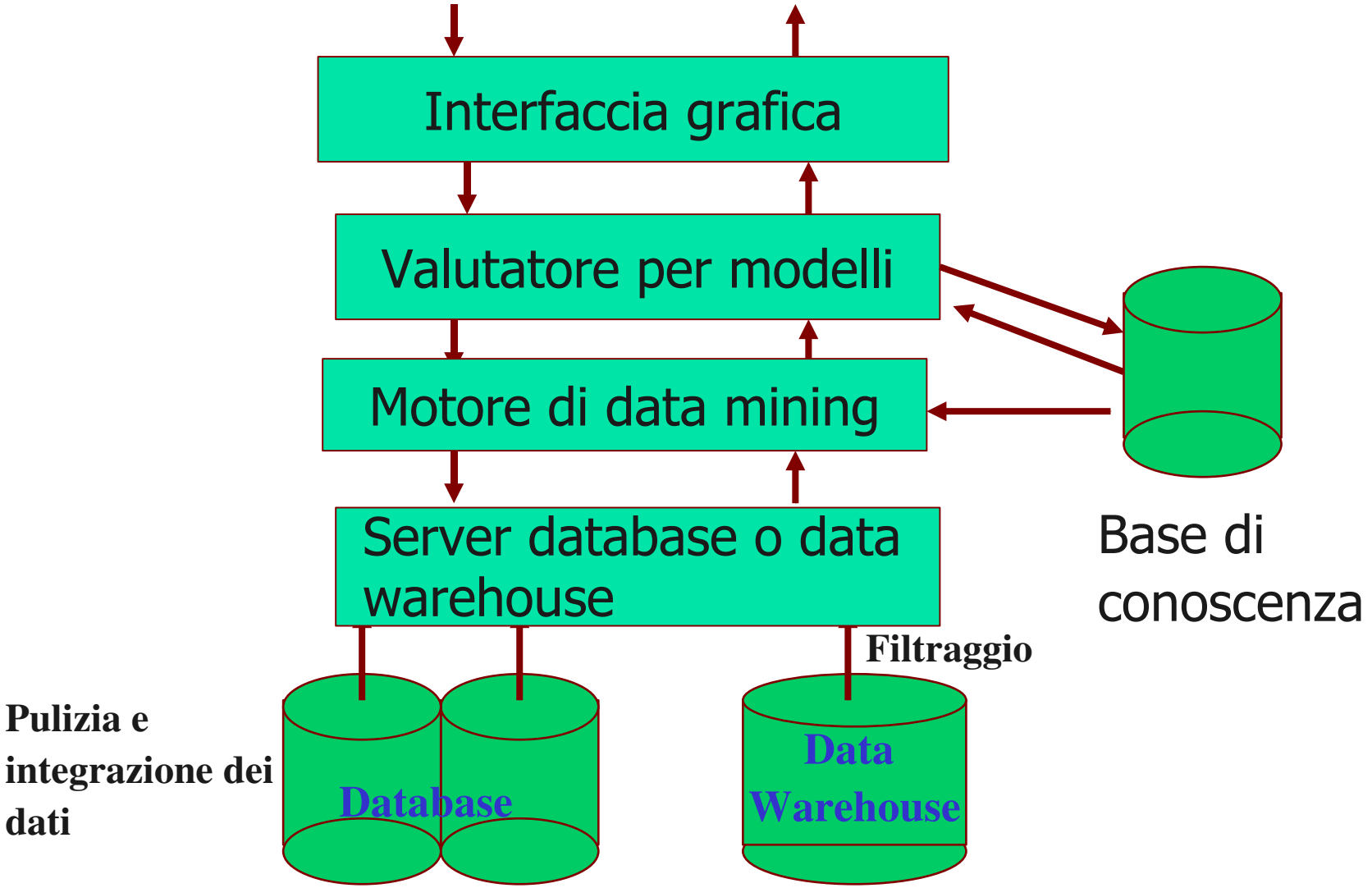
- Database testuali
  - Contengono testi più o meno lunghi e più o meno strutturati (in capitoli, paragrafi, sezioni, etc...)
  - Esempi di analisi:
    - Determinare una descrizione concisa di insiemi di documenti tramite parole chiavi
    - Raggruppare documenti tra di loro simili
- Database multimediali
  - Contengono immagini, audio o video
  - Necessitano di tecniche di immagazzinamento e ricerca specifici a causa della enorme mole di dati

Altre sorgenti (4)

- Flussi continui di dati
  - Caratteristiche
    - I dati sono spesso troppo voluminosi per poter essere memorizzati
    - ... oppure si possono memorizzare, ma solo per poco tempo
    - Le analisi devono essere veloci, spesso in tempo reale
  - Esempi di analisi:
    - Determinazione di intrusioni nei sistemi informatici sulla base delle anomali nei flussi di messaggi prodotti dallo stesso

Architettura

Architettura di un sistema di data mining



Base di conoscenza

- I sistemi di data mining possono usare altre informazioni oltre al tipo degli attributi:
  - informazioni dimensionali, in modo da non confrontare dati espressi con unità di misura diverse (cosa vuol dire che 3 Km è minore di 5 Litri?)
  - ordinamenti circolari: indicare se un attributo è soggetto a particolare circolarità dei dati
    - gli angoli vanno da 0 a 360° (o da 0 a 2π) e poi ricomnciano da 0.
    - ci si può riferire allo "stesso giorno nella prossima settimana" o alla "prossima domenica"
  - gerarchie di concetti: alcuni attributi possono essere trattati a vari livelli di dettaglio
- Tutte queste informazioni prendono il nome di metadati e consentono di aumentare l'efficienza del sistema di data mining.

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 in 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

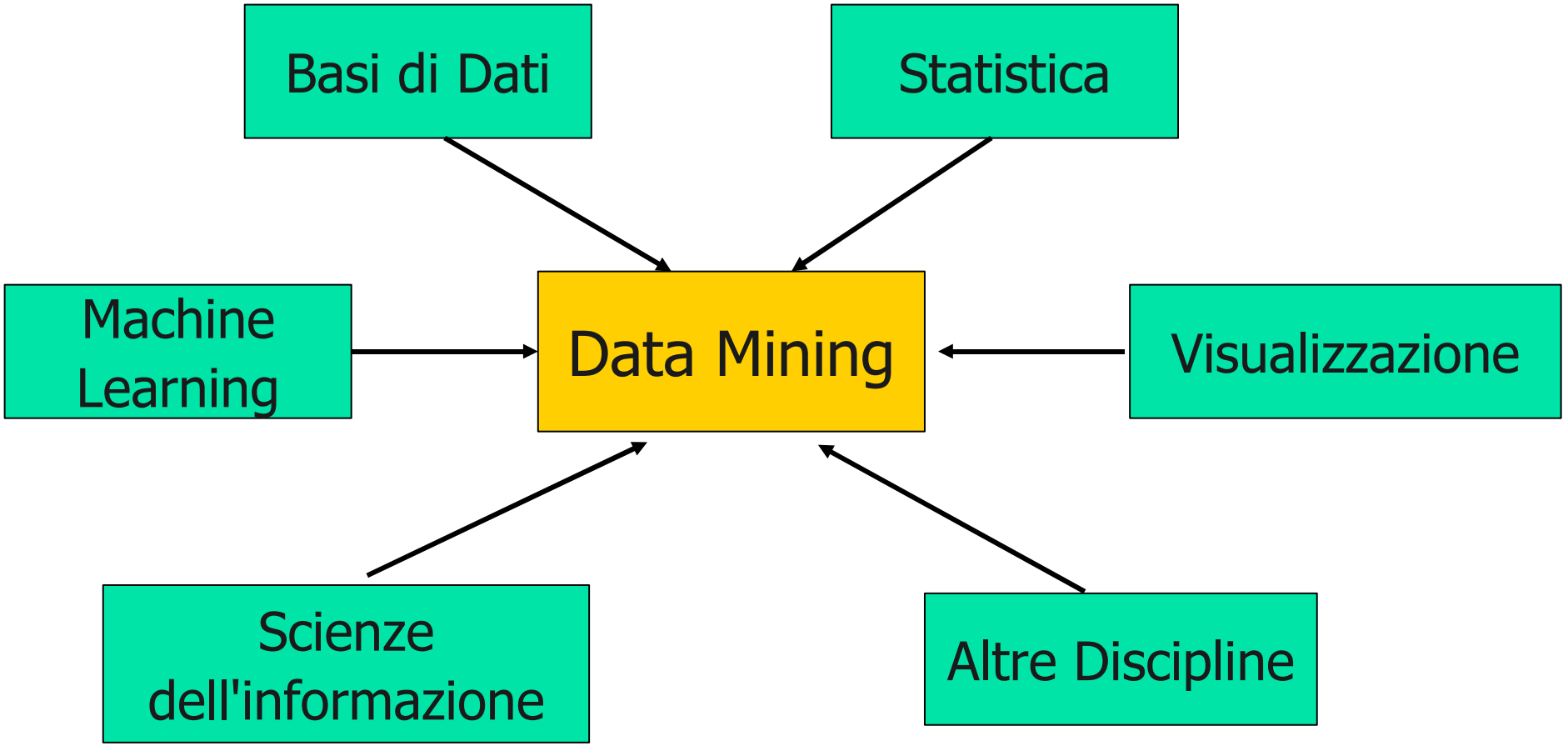
- Una industria ha migliaia di apparati elettromeccanici. Quando un guasto si verifica, il tipo di guasto viene identificato da una serie di sensori che misurano le vibrazioni in punti diversi dell'impianto
- La identificazione avveniva con un esperto umano.. come automatizzarla?
  - un sistema esperto 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

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 tecnice 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 perché:
  - 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 contengono spesso dei dati che non sono stati raccolti appositamente a scopo di analisi

Data Mining e Statistica (2)

- La ricerca nell'ambito del data mining è stata vista per lungo con sospetto dagli statistici
  - da cui i termini dispregiativi di "data fishing" e "data dredging".
- Queste le critiche
  - nel data mining non vi è un unico modello di riferimento, ma numerosi modelli in competizione. È sempre possibile trovare un modello complesso che si adatti bene ai dati
  - l'abbondanza di dati può portare a pattern in realtà inesistenti
- Tuttavia
  - le tecniche moderne pongono molta attenzione alla generalizzabilità dei pattern, preferendo modelli più semplici
  - 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

Data Mining ed Etica (1)

- L'uso del data mining ha una 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
  - Tuttavia, molto 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
  - pulizia, integrazione, selezione e trasformazione
  - data mining
  - valutazione e visualizzazione dei pattern
- funzionalità del data mining: classificazione, descrizione di concetti, associazioni, raggruppamenti, previsioni..
- applicazioni del data mining
- relazione del datamining con altre discipline e problemi di etica