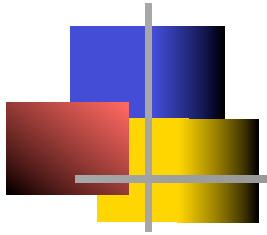


di.unito.it

DIPARTIMENTO DI INFORMATICA

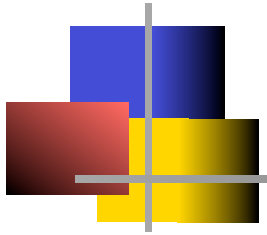
Introduzione alla *Business Intelligence*

ovvero ai concetti di base dell'apprendimento automatico a supporto della competizione aziendale

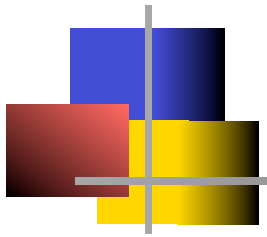
Rosa Meo

Università di Torino

Introduzione



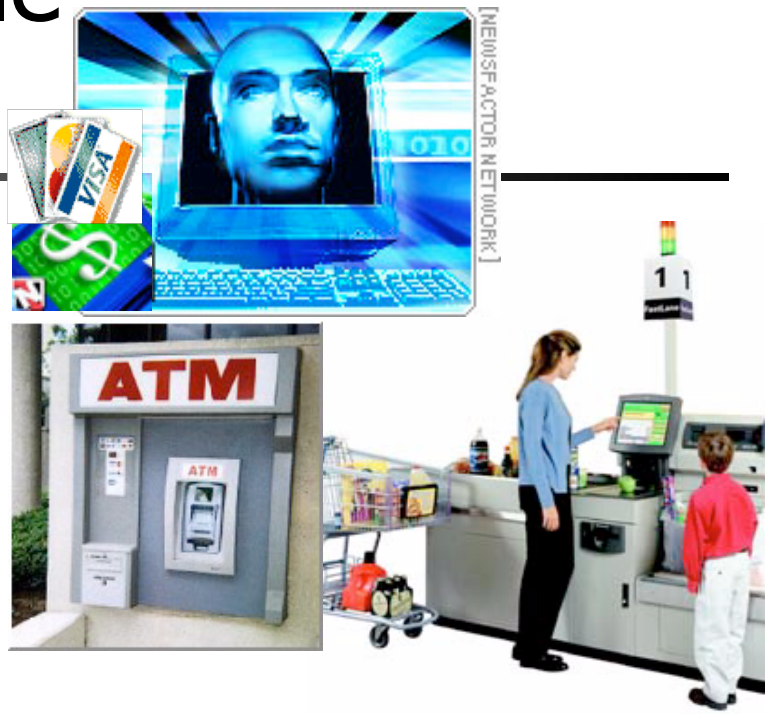
- Motivazioni: Perchè fare business intelligence?
- Cos'è?
- Su quale tipo di dati?
- Funzionalità
- Esempi
- Esperimenti con *Weka*



Business Intelligence

- Con *Business Intelligence* si intende la disciplina che studia e applica tecniche informatiche avanzate per analizzare i dati a disposizione delle aziende ai fini di aumentarne la competitività e migliorare la comprensione di fenomeni di interesse
- Si parla anche dei sistemi a supporto delle decisioni aziendali
- In generale, l'analisi di dati (specie se disponibili in grossi volumi) si denota con il termine *Data Mining* (ovvero *scavare nei dati* alla ricerca di preziose informazioni).

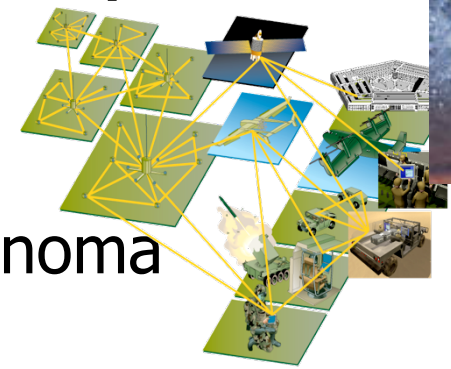
Perchè scavare nei dati? punto di vista commerciale

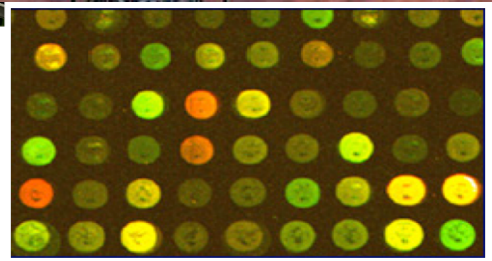


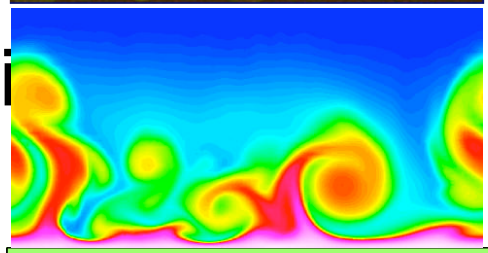
- Molti dati vengono collezionati e memorizzati
  - Pagine Web, e-commerce
  - Acquisti/supermercati
  - Transazioni bancomat/carte di credito
- I computer sono divenuti meno costosi e più potenti
- La pressione competitiva è più forte
  - Occorre fornire un servizio migliore al cliente e più specifico sulle sue esigenze (Customer Relationship Management)

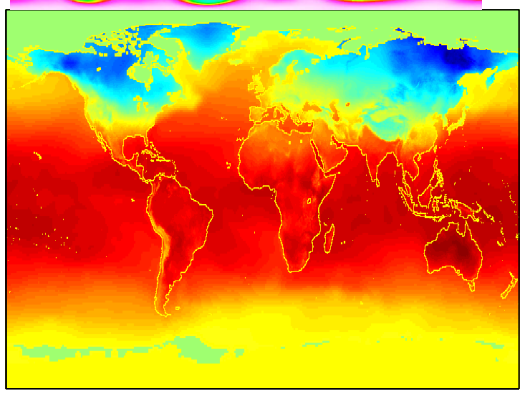
Scavare nei dati? Punto di vista scientifico

- I dati vengono collezionati e memorizzati a enormi velocità (GB/ora)
  - Sensori remoti sui satelliti
  - telescopi di monitoraggio
  - Microarray per analisi del genoma
  - Simulazioni scientifiche che generano terabyte di dati







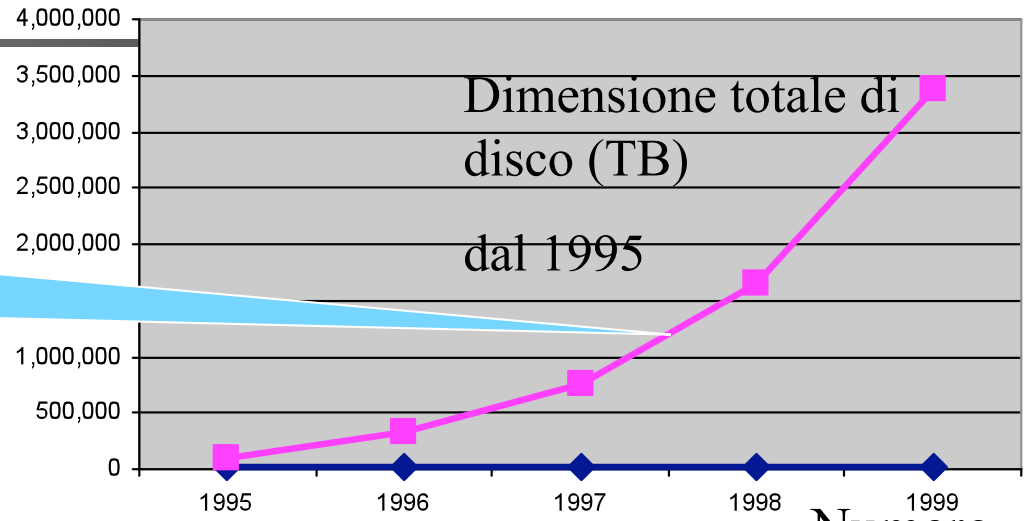


- Le tecniche tradizionali sono insufficienti sui dati grezzi
- Il Data mining aiuta gli scienziati:
  - A classificare e separare in categorie i dati
  - Nella formulazione delle ipotesi

Motivazioni: "La necessità è la madre delle invenzioni"

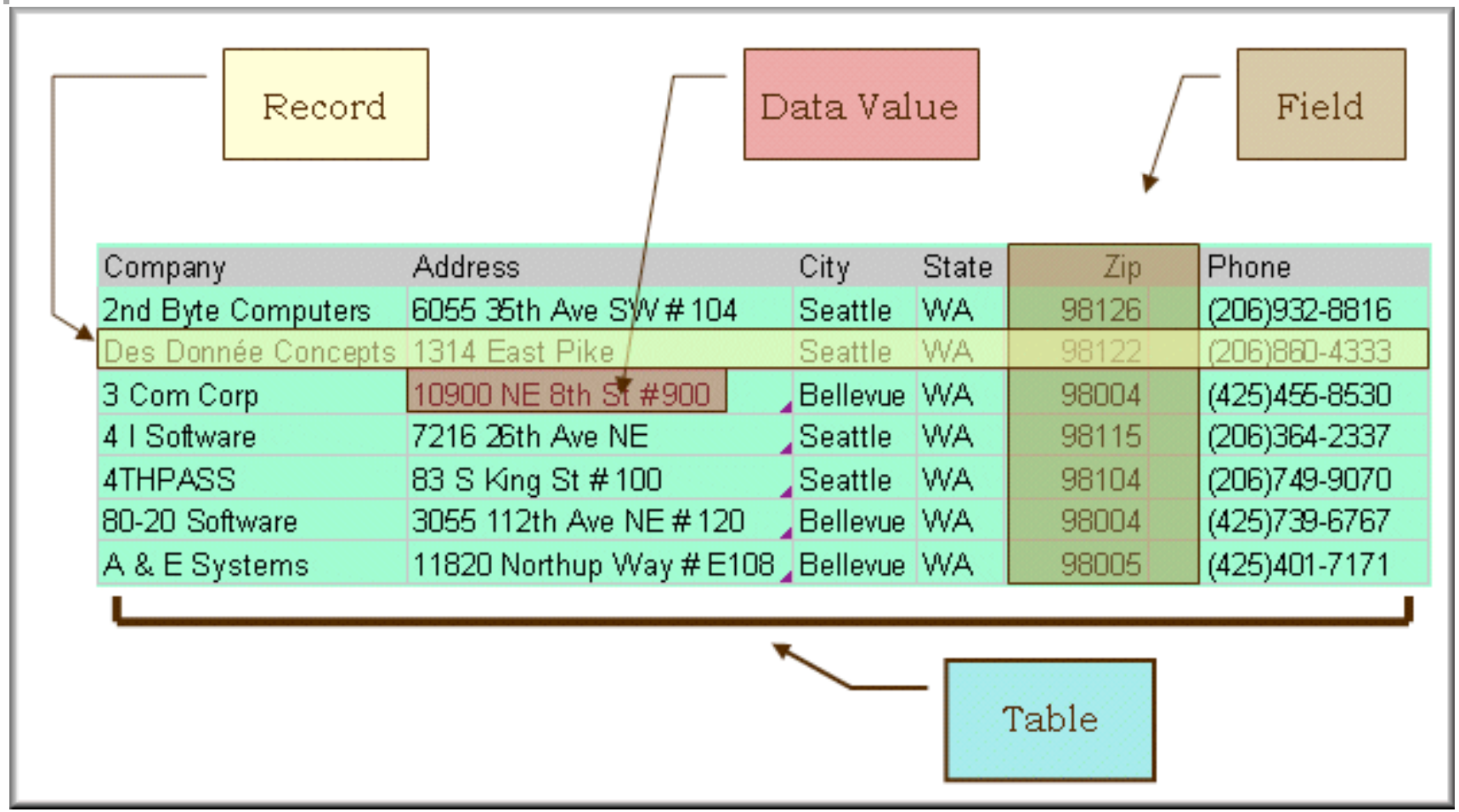
- Esplosione dei dati

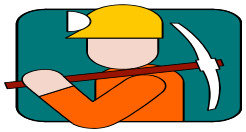
Data Gap



- Stiamo annegando nei dati, ma bramiamo per acquisire conoscenza da essi!
- Soluzione: Data warehousing e data mining
  - Analisi statistica su dati a più dimensioni da grossi volumi di dati
  - Estrazione di elementi di conoscenza (regole, regolarità, pattern, vincoli)

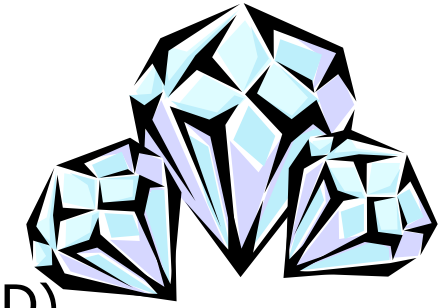
Basi di dati relazionali





Cos'è il Data Mining?

- Data mining (Knowledge Discovery in Databases):
  - Estrazione di informazione di interesse, di elevato valore (non nota e non ovvia, implicita, e potenzialmente utile) da grandi volumi di dati



- Nomi alternativi e la loro storia:
  - Data mining: un equivoco?
  - La scoperta di conoscenza (mining) nei dati (KDD), estrazione di conoscenza, analisi dei dati o riconoscimento di pattern, archeologia dei dati, business intelligence, ecc.

Perchè Data Mining? — Potenziali Applicazioni

- Analisi delle basi di dati e supporto alla decisione
  - Analisi e gestione del mercato aziendale
    - Publicità mirata, gestione della relazione col cliente, analisi degli acquisti e delle preferenze del cliente, cross selling, segmentazione del mercato

  Analisi e gestione del rischio
    - Previsione, mantenimento della clientela, miglioramento della stipula di contratti, controllo della qualità, analisi della concorrenza aziendale

  - Riconoscimento e gestione delle frodi (anche informatici, virus)
- Altre applicazioni: analisi dei testi (newsgroup, email, documenti) e delle pagine web, dei motori di ricerca su internet

Funzionalità del Data Mining

- Metodi predittivi
  - Usano il valore osservato di alcune variabili in esempi a disposizione per predire il valore sconosciuto di una variabile di interesse in esempi futuri

- Metodi Descrittivi
  - Trovano *pattern* o schemi interpretabili dall'utente per la descrizione dei dati

Da [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Obiettivi del Data Mining

- Classificazione [Metodo predittivo]
- Clustering [Metodo descrittivo]
- Estrazione di regole associative [Metodo descrittivo]
- Estrazione di pattern sequenziali [Metodo descrittivo]
- Regressione [Metodo predittivo]
- Scoperta di anomalie [Metodo predittivo]

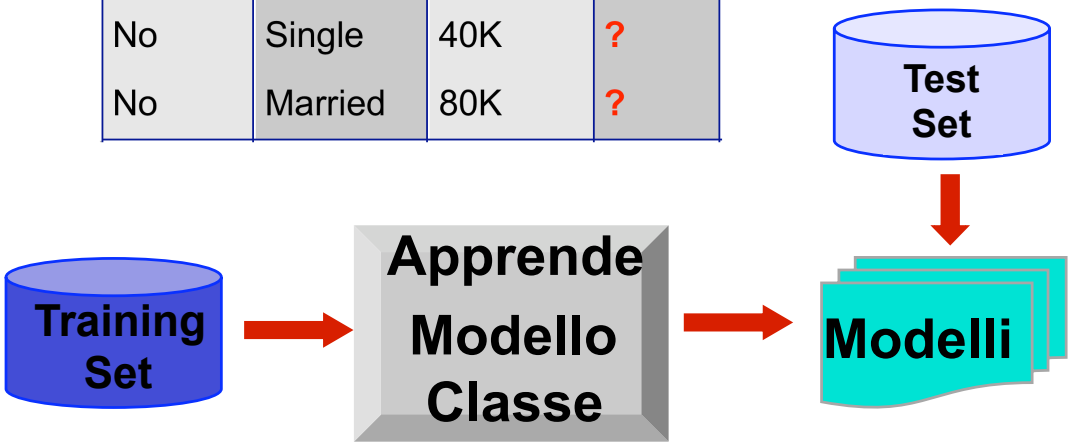
Classificazione: Definizione

- Data una collezione di esempi (*training set*)
  - Ciascun esempio contiene un set di *attributi*; uno è la classe.
- L'obiettivo è trovare un *modello* per la classe degli esempi come una funzione dei valori degli altri attributi.
- Scopo del modello: predire la classe degli esempi futuri il più accuratamente possibile.
  - Si usa un *test set* per determinare l'accuratezza del modello. Di solito, il data set dato è diviso in training e test set, dove il training set viene usato per costruire il modello mentre il test set viene usato per validare il modello stesso.
  - Presentazione: albero di decisione, regole di classificazione, reti neurali
  - Potrebbe essere usato anche per determinare il valore mancante di alcuni attributi

Classificazione: Esempio

Oid	Rim-borso (categorico)	Stato matrim. (categorico)	Imponibi-le (tasse) (continuo)	Frode (classe)
O1	Yes	Single	125K	No
O2	No	Sposato	100K	No
O3	No	Single	70K	No
O4	Yes	Sposato	120K	No
O5	No	Divorziato	95K	Yes
O6	No	Sposato	60K	No
O7	Yes	Divorziato	220K	No
O8	No	Single	85K	Yes
O9	No	Sposato	75K	No
O10	No	Single	90K	Yes

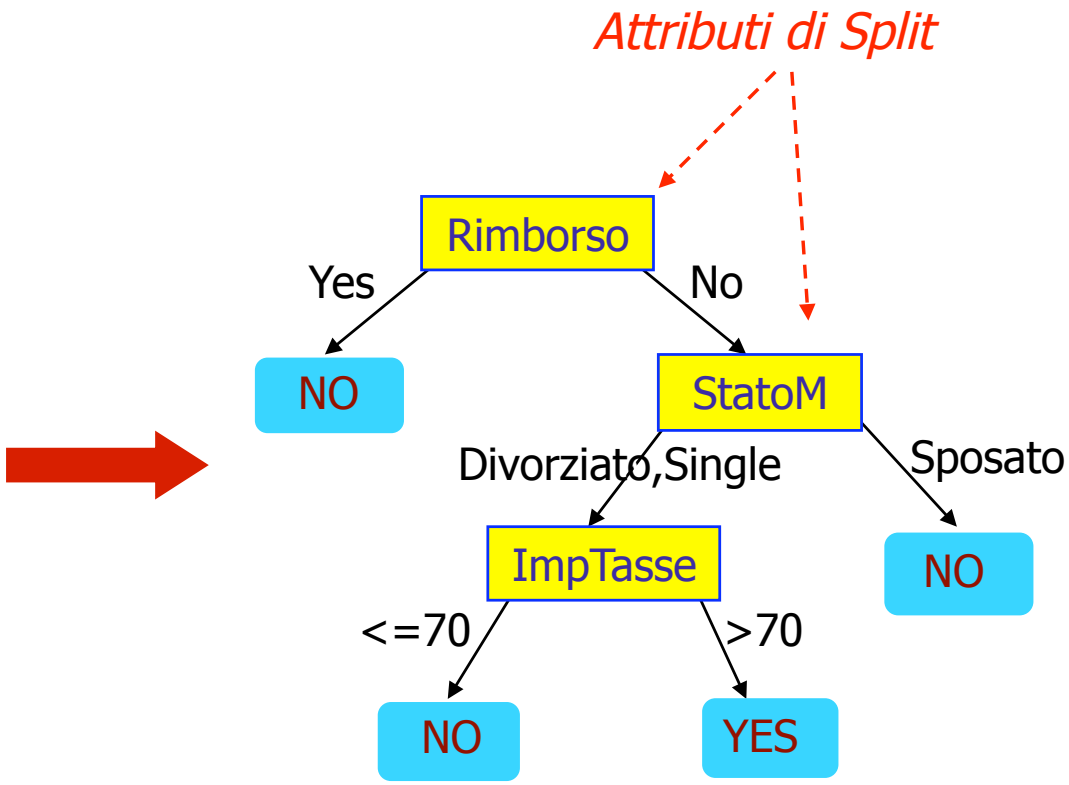
Rim-borso	Stato Matrim.	Imponib (tasse)	Frode
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Esempio di Modello: Albero di Decisione

Oid	Rim-borso	Stato matrim.	Imponibi-le (tasse)	Frode
O1	Yes	Single	125K	No
O2	No	Sposato	100K	No
O3	No	Single	70K	No
O4	Yes	Sposato	120K	No
O5	No	Divorziato	95K	Yes
O6	No	Sposato	60K	No
O7	Yes	Divorziato	220K	No
O8	No	Single	85K	Yes
O9	No	Sposato	75K	No
O10	No	Single	90K	Yes

Training set



Modello: Albero di Decisione

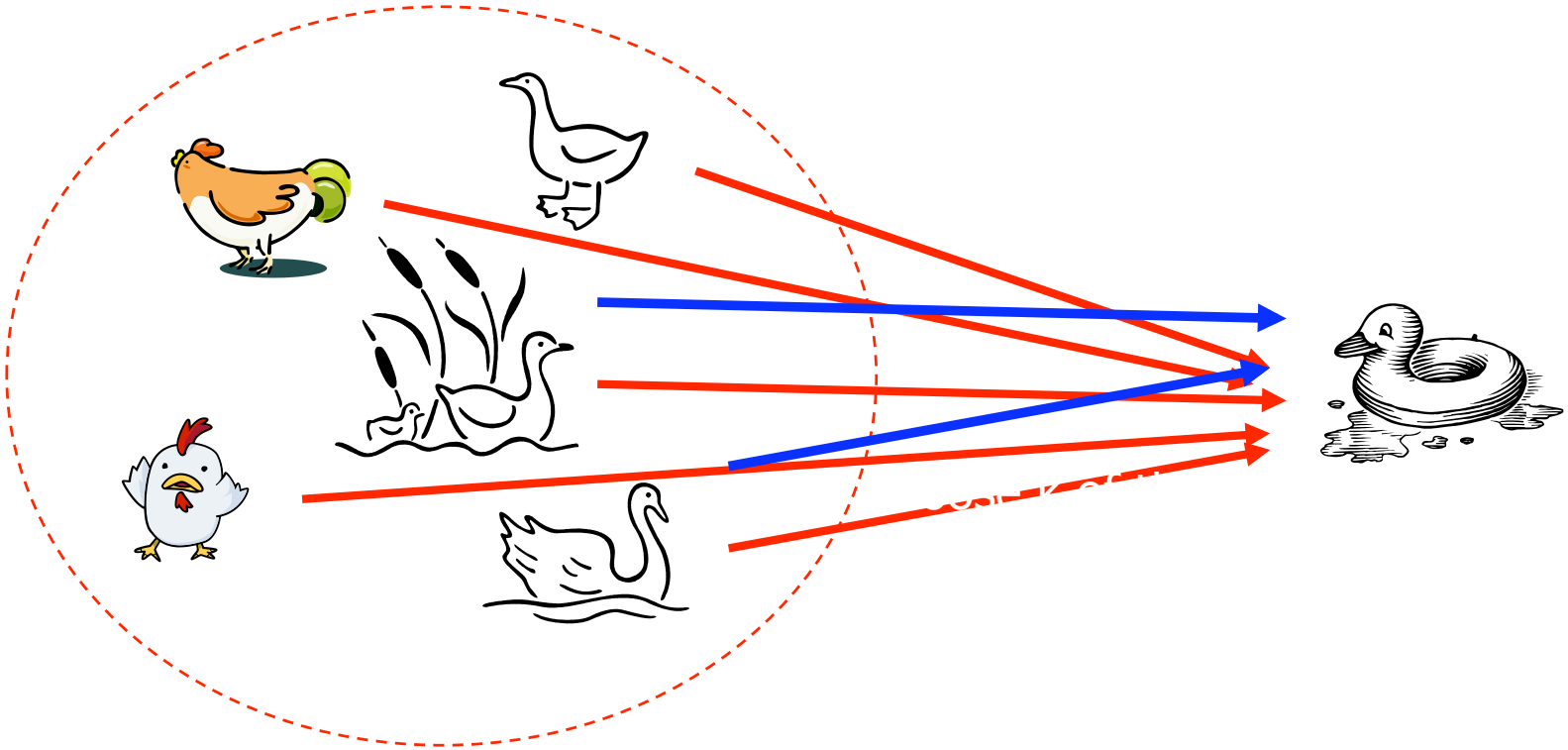
Esempio 2: Classificatore a Regole

Nome	Sangue	Partorisce	Vola	Vive in acqua	Classe
uomo	caldo	yes	no	no	mammifero
pitone	freddo	no	no	no	rettile
salmone	freddo	no	no	yes	pesce
delfino	caldo	yes	no	yes	mammifero
pipistrello	caldo	yes	yes	no	mammifero
aquila	caldo	no	yes	no	uccello
...	...	...	...	...	...

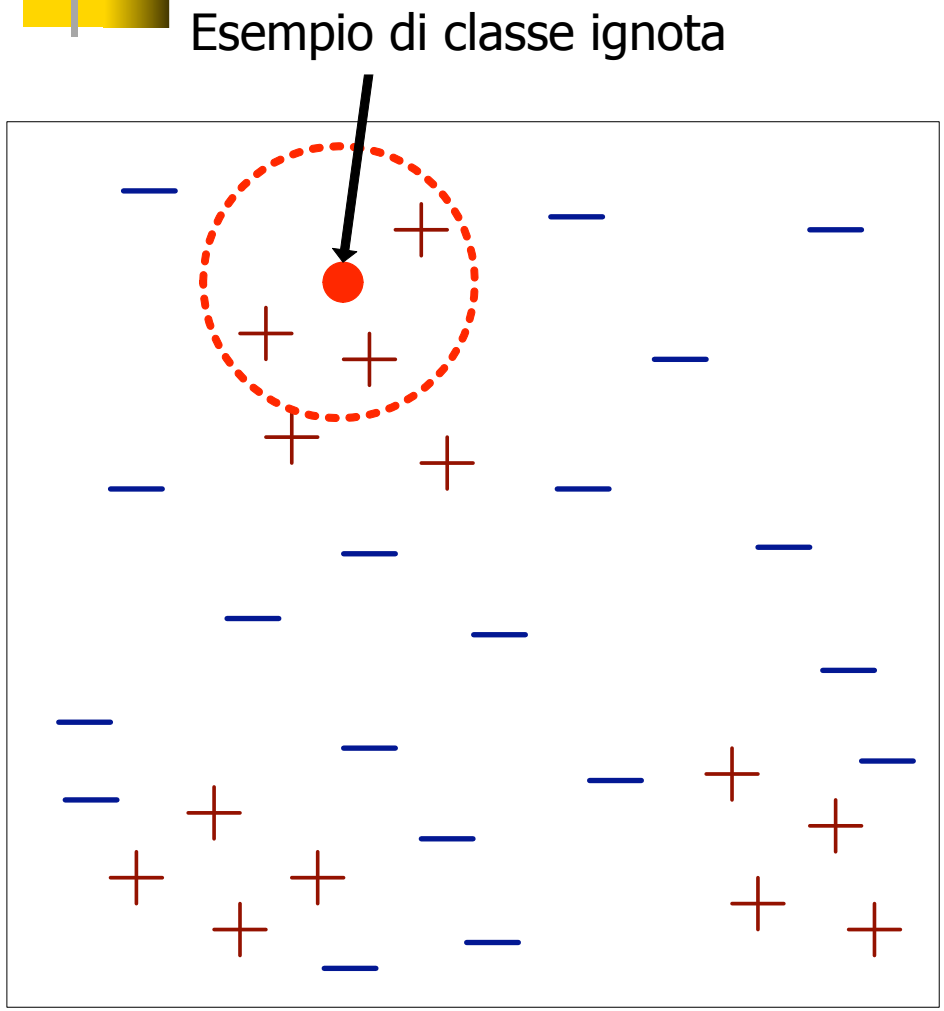
R1: (Partorisce = no) E (Vola = yes) → (Classe = uccello)
R2: (Partorisce = no) E (Vive in acqua = yes) → (Classe = pesce)
R3: (Partorisce = yes) E (Sangue = caldo) → (Classe = mammifero)
R4: (Partorisce = no) E (Vola = no) → (Classe = rettile)

Esempio 3: Classificatore sugli esempi simili

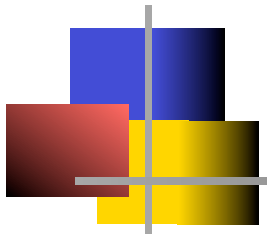
- Idea base:
  - Se cammina come un'oca, fa "quack"come un'oca, allora probabilmente è un'oca



Classificatore sugli esempi simili



- Richiede tre cose
  - L'insieme degli esempi
  - Una funzione di distanza tra gli esempi
  - Il valore di *k*, il numero dei vicini da considerare
- Per classificare un esempio:
  - Calcolare la distanza con altri esempi nel training set
  - Identificare i *k* esempi più vicini
  - Usare la classe dei vicini per determinare la classe dell' esempio ignoto (dalla maggioranza)



Classificazione: Applicazione 1

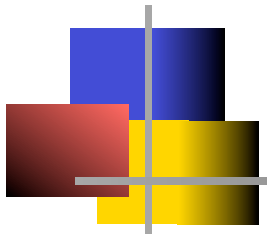
- Pubblicità mirata
  - Obiettivo: Ridurre il costo della pubblicità cercando di capire meglio a quali clienti fare pubblicità (quelli che verosimilmente potrebbero comprare il nuovo prodotto).
  - Approccio:
    - Usiamo le informazioni di un prodotto simile già venduto
    - Sappiamo quali clienti lo hanno già comprato. Questa informazione ci dà l'attributo di classe.
    - Collezionare informazioni sullo stile di vita, sulle caratteristiche demografiche della popolazione e su come i nostri clienti hanno interagito con l'azienda.
      - Tipo di lavoro, dove vivono, quanto guadagnano, ecc.
    - Usiamo questa informazione come attributi di input per costruire il modello della classe.

Da [Berry & Linoff] Data Mining Techniques, 1997

Classificazione: Applicazione 2

- Riconoscimento di frodi
  - Obiettivo: Predire i casi fraudulenti nelle transazioni con carta di credito.
  - Approccio:

    - Usiamo le informazioni sulle carte di credito e sui clienti dei conti in banca.
      - quando il cliente compra, che cosa compra, quanto spesso paga in ritardo, ecc.
    - Etichettiamo le transazioni passate come frode o no. Questa etichetta forma l'attributo di classe.
    - Apprendiamo un modello per l'attributo di classe delle transazioni.
    - Usiamo questo modello per riconoscere le frodi osservando le nuove transazioni con carta di credito su un conto bancario.



Classificazione: Applicazione 3

- Abbandono dei clienti:
  - Obiettivo: predire se un cliente verosimilmente lascerà l'azienda per la concorrenza.
  - Approccio:
    - Usiamo le informazioni dettagliate sull'interazione dei clienti presenti e passati con l'azienda.
      - Quanto spesso il cliente effettua chiamate, dove chiama, quando chiama, informazioni sul suo stato finanziario, situazione famigliare, ecc.
    - Etichettare i clienti come fedeli o no.
    - Trovare un modello per la fedeltà dei nostri clienti.

From [Berry & Linoff] Data Mining Techniques, 1997

Classificazione: Applicazione 4

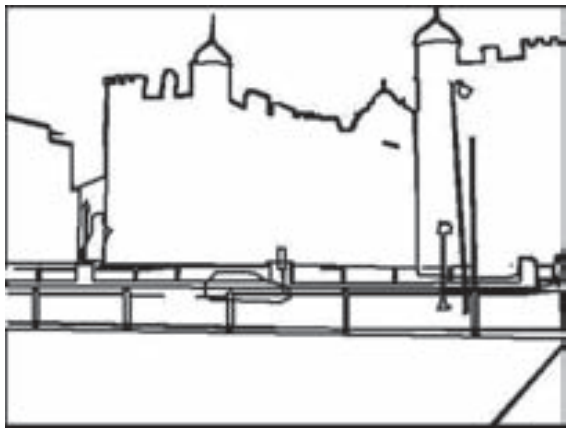
- Catalogo delle stelle nella volta celeste
  - Obiettivo: predire la classe (stella o galassia) dell'oggetto celeste, sulla base delle immagini catturate dai telescopi (Osservatorio Palomar).
    - 3000 immagini con 23,040 x 23,040 pixel ciascuna.

  - Approccio:
    - Segmentare l'immagine.
    - Misurare gli attributi dell'immagine (features) - 40 attributi sono stati selezionati.
    - Costruire un modello della classe sulla base di questi attributi.
    - Storia di successo: trovati 16 nuovi quasar, alcuni dei quali sono i più distanti e difficili da scoprire!

Da [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Segmentazione delle immagini

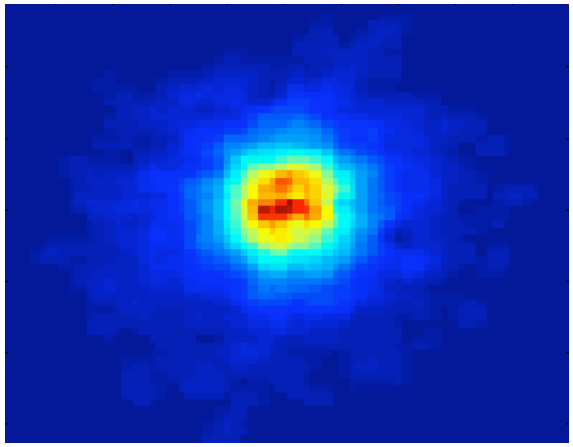
- Nella visione tramite computer la segmentazione delle immagini si riferisce al processo di partizionamento di una immagine digitale in porzioni distinte che condividono le stesse caratteristiche (come colore, intensità, ..) con lo scopo di identificare i contorni degli oggetti.



Classificazione delle galassie

Da: http://aps.umn.edu

Origine



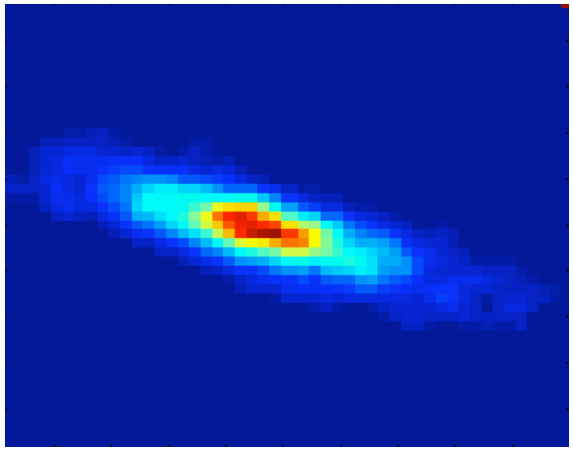
Attributo di classe:

- **fase della stella**

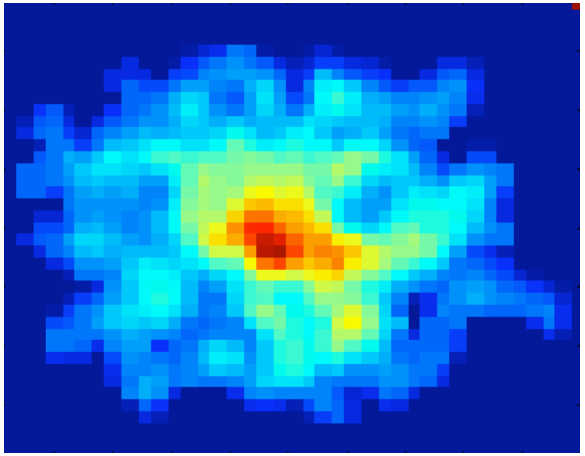
Attributi:

- **Caratteristiche delle immagini,**
- **caratteristiche della luce e delle forme d'onda, ecc.**

Intermedio

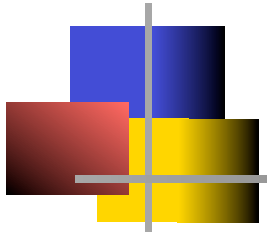


Finale



Dimensione dei dati:

- **72 milioni di stelle, 20 milioni di galassie**
- **Catalogo: 9 GB**
- **Immagini: 150 GB**



Clustering: Definizione

- Dato un insieme di punti, ciascuno descritto da un insieme di attributi, e una misura di similarità tra i punti, trovare cluster di punti tali che:
  - I punti nei cluster sono simili tra di loro.
  - I punti in cluster diversi sono meno simili.

- Misure di similarità:
  - Distanza euclidea (se gli attributi descrittivi dei punti sono a valori continui).
  - Altre misure che dipendono dallo specifico problema.

Distanza euclidea

$$\overline{C}^2=\overline{A}^2+\overline{B}^2=(x_0-x_1)^2+(y_0-y_1)^2$$

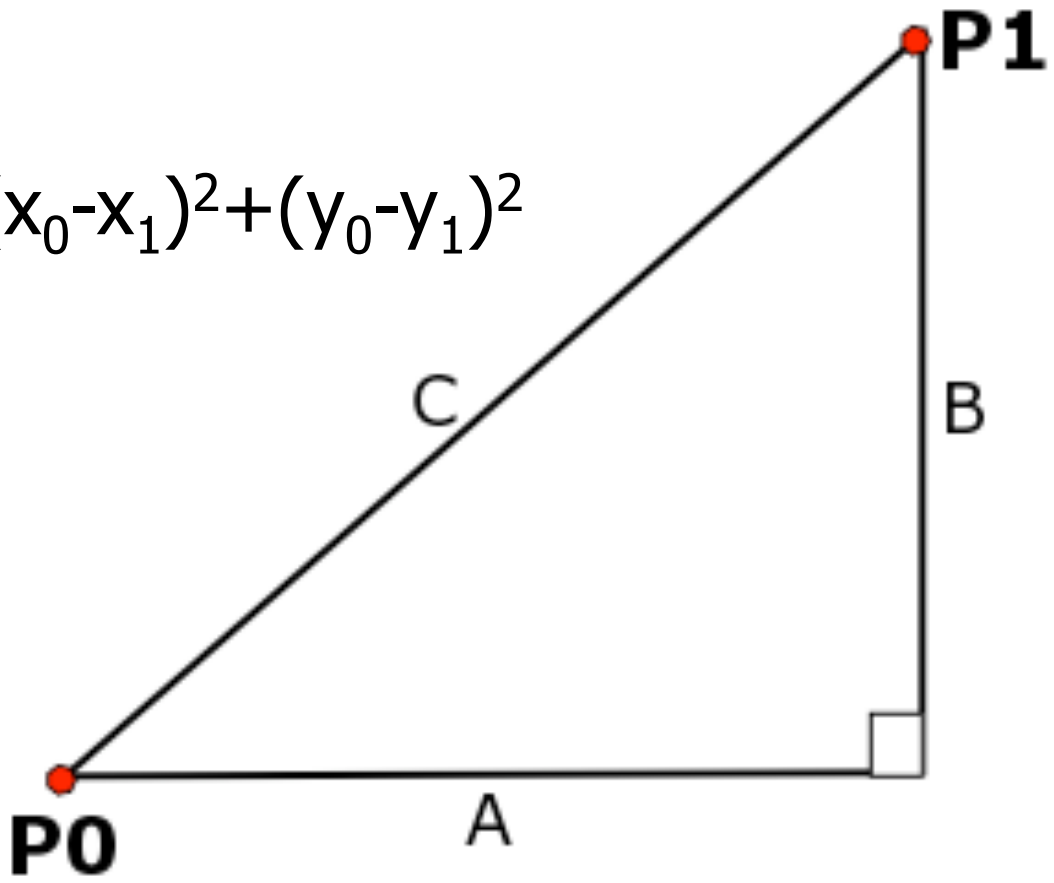
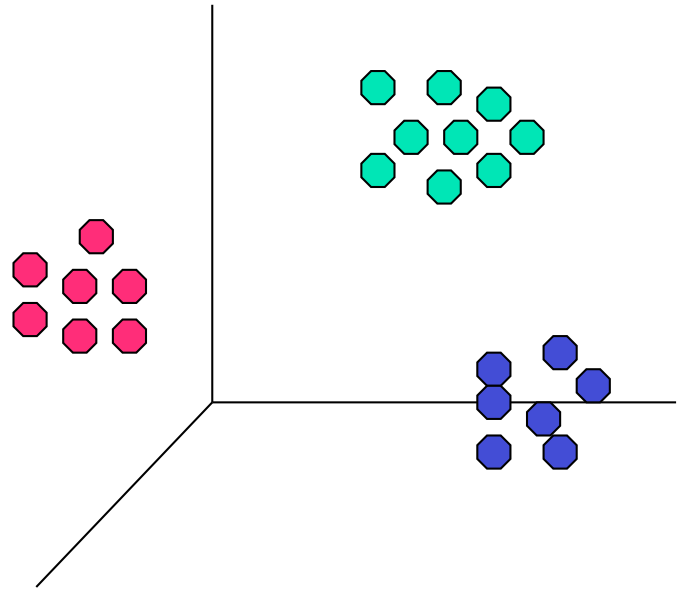


Illustrazione del Clustering

- Cluster basati sulla distanza euclidea in uno spazio a 3-D.

Le distanze **Intra-cluster** vengono *minimizzate*

Le distanze **Inter-cluster** vengono *massimizzate*



La distanza tra oggetti su attributi categorici

Oid	Rim-borso	Stato matrim.	Imponibile (tasse)	Frode
O1	Yes	Single	125K	No
O2	No	Sposato	100K	No
O3	No	Single	70K	No
O4	Yes	Sposato	120K	No
O5	No	Divorziato	95K	Yes
O6	No	Sposato	60K	No
O7	Yes	Divorziato	220K	No
O8	No	Single	85K	Yes
O9	No	Sposato	75K	No
O10	No	Single	90K	Yes

$$d(Oi.cat, Oj.cat)^2 = \begin{cases} 0 \text{ se } O1.cat = O2.cat \\ 1 \text{ se } O1.cat \neq O2.cat \end{cases}$$

$$dist(O1, O2)^2 = d(O1.Rimbor - O2.Rimbor)^{2+} d(O1.StatoM - O2.StatoM)^{2+} d(O1.ImTasse - O2.ImTasse)^{2+} d(O1.Frode - O2.Frode)^{2} =$$

$$1+1+25^2+0 = 627$$

Clustering: Applicazione 1

- Segmentazione del mercato:
  - Scopo: suddividere il mercato in porzioni separate tali che i clienti di ciascuna porzione potrebbero essere selezionati come destinatari a cui proporre un ben distinto insieme di offerte.
  - Metodo:
    - Collezionare diversi attributi descrittivi dei clienti basati su informazioni geografiche e sul loro stile di vita.
    - Trovare in base a tali attributi i cluster dei clienti simili.
    - Misurare la qualità dei cluster osservando gli *schemi* (pattern) di acquisto di prodotti dei clienti nello stesso cluster e confrontarli con gli schemi dei clienti degli altri cluster.

Clustering: Applicazione 2

- Clustering dei documenti:
  - Scopo: Trovare gruppi di documenti che sono simili basati sulle parole importanti che contengono.
  - Metodo: Identificare le parole che occorrono frequentemente in ciascun documento.
    Formare una misura di similarità tra documenti basata sulle frequenze dei termini.
    Usarla per formare i cluster.
  - Vantaggio: l'*Information Retrieval* può utilizzare i cluster così formati per mettere in relazione un nuovo documento ai cluster di documenti o un termine ai documenti clusterizzati.

Illustrazione dei cluster di documenti

- Punti da organizzare in cluster: 3204 articoli del *Los Angeles Times*.
- Misura di similarità: quante parole in comune ci sono tra due documenti (dopo aver applicato un filtro sulle *stop words*).

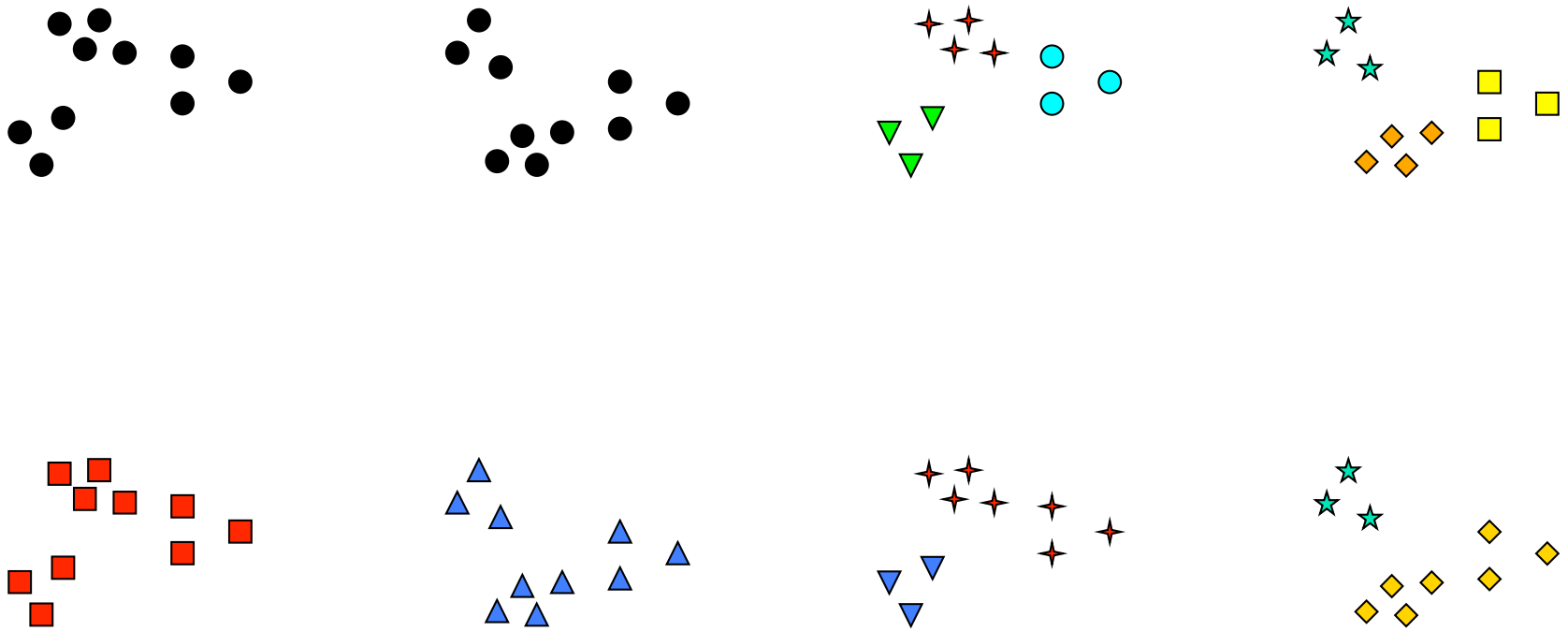
Categoria	Articoli totali	Articoli nel cluster
Finanza	555	364
Estero	341	260
Nazionale	273	36
Locale	943	746
Sport	738	573
Cultura e spettacolo	354	278

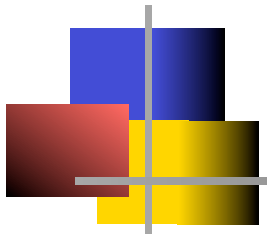
Cluster dei titoli di S&P 500

- Osserviamo l'andamento del prezzo dei titoli giorno per giorno: evento UP se il prezzo è salito rispetto al giorno precedente oppure DOWN.
- Misura di similarità: due titoli sono simili se gli eventi UP/DOWN appaiono frequentemente in entrambi nello stesso giorno.

	Cluster	*Industry Group*
1	Applied-Matl, Bay-Network, 3-COM, Cabletron-Sys, CISCO,HP, DSC-Comm, INTEL, LSI-Logic, Micron-Tech, Texas-Inst, Tellabs-Inc, Natl-Semiconduct, Oracle, SGI, Sun	Technology1-DOWN
2	Apple-Comp, Autodesk, DEC, ADV-Micro-Device, Andrew-Corp, Computer-Associate, Circuit-City, Compaq, EMC-Corp, Gen-Inst, Motorola, Microsoft, Scientific-Atl	Technology2-DOWN
3	Fannie-Mae, Fed-Home-Loan, MBNA-Corp, Morgan-Stanley	Financial-DOWN
4	Baker-Hughes, Dresser-Inds, Halliburton-HLD, Louisiana-Land, Phillips-Petro, Unocal, Schlumberger	Oil-UP

La nozione di cluster può essere ambigua

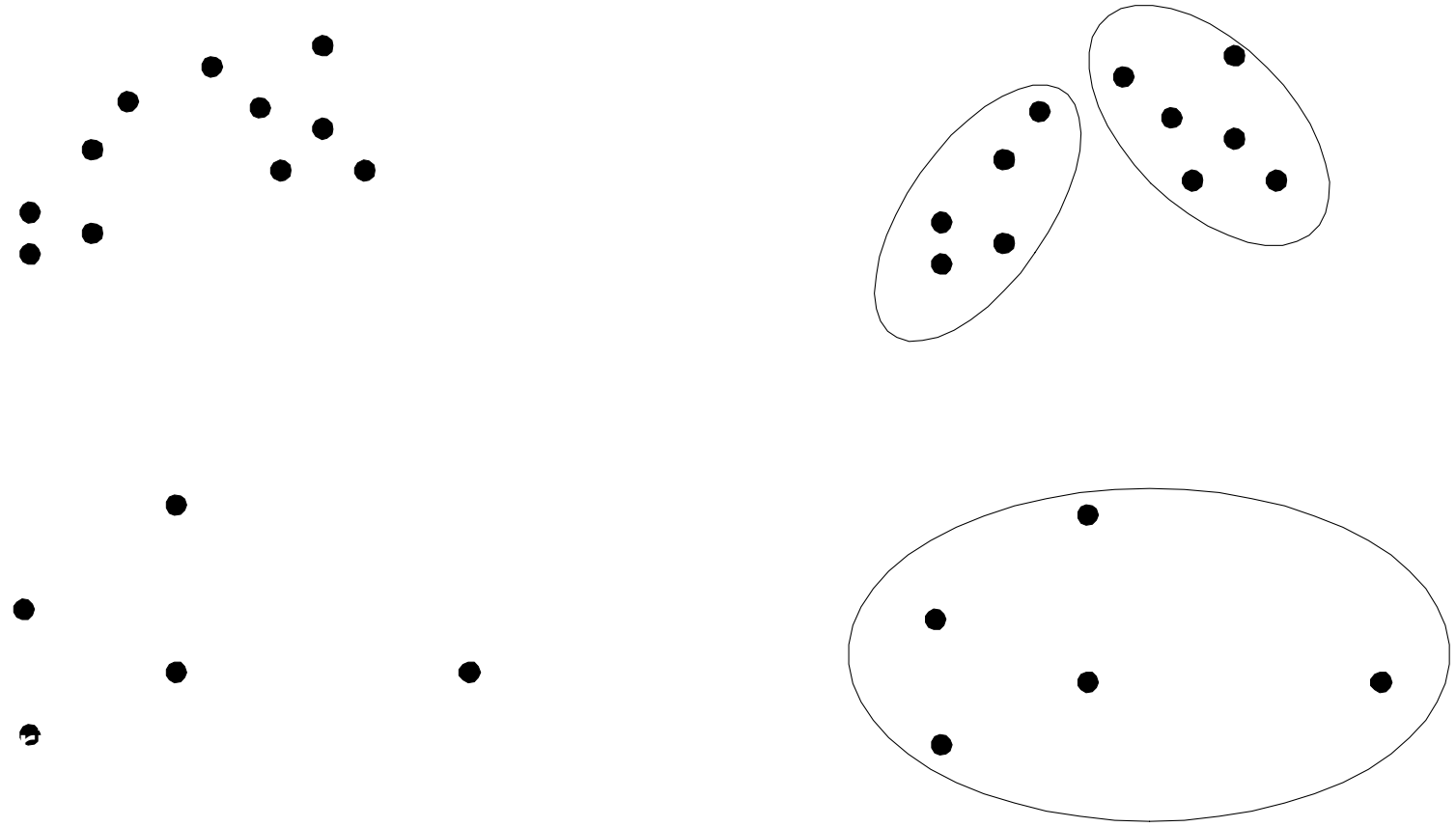




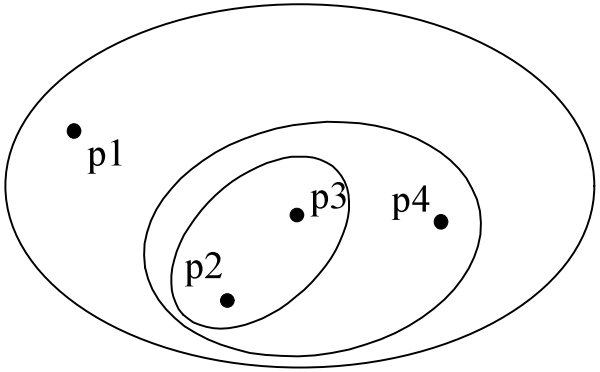
Tipi di clustering

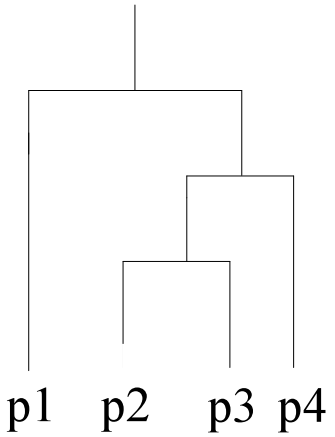
- Un clustering è un insieme di cluster
- Distinzione importante tra insiemi di cluster gerarchici e partizionali
- Clustering partizionale
  - Una suddivisione degli oggetti in sottoinsiemi non sovrapposti (clusters) tali che ciascun oggetto sta in un insieme soltanto
- Clustering gerarchico
  - Un insieme di cluster *annidati* organizzati in un albero (gerarchia)

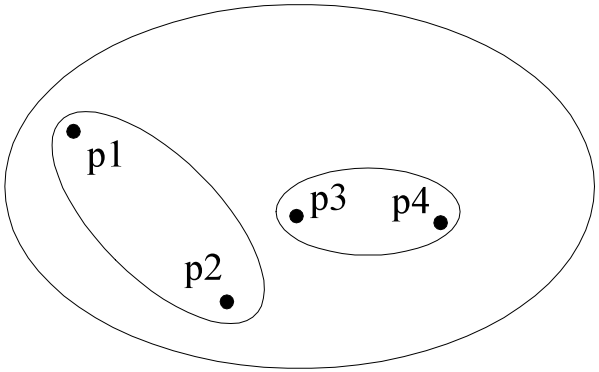
Clustering partizionale

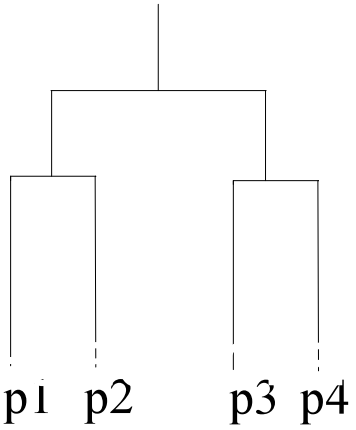


Clustering gerarchico

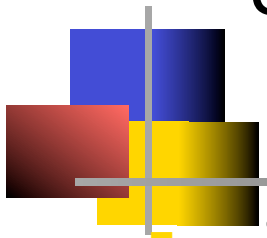








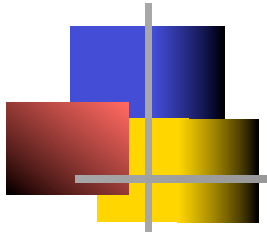
Un algoritmo di clustering: *K-medie* (K-means)



- Segue il metodo di clustering partizionale
- Ciascun cluster è associato ad un centroide (baricentro)
- Ciascun punto è assegnato al cluster con il baricentro più vicino
- Il numero dei cluster deve essere specificato: **K**
- L'algoritmo di base è molto semplice:

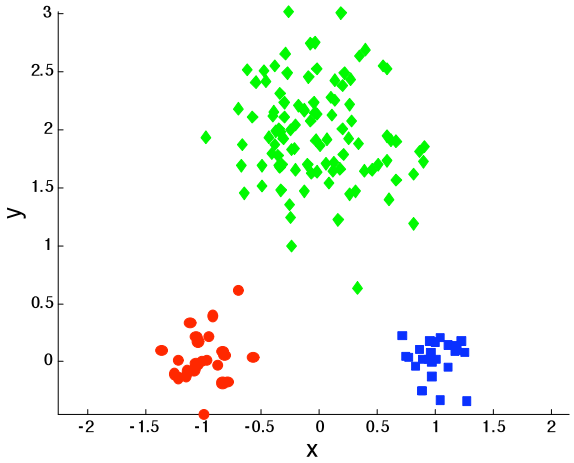
1. Selezionare K punti come centri dei K cluster
2. Ripetere:
    1. Formare K cluster assegnando ciascun punto al cluster il cui centro è più vicino
    2. Ricalcolare il centro di ciascun cluster
3. Finchè la posizione dei centri rimane stabile

K-medie – Ulteriori dettagli

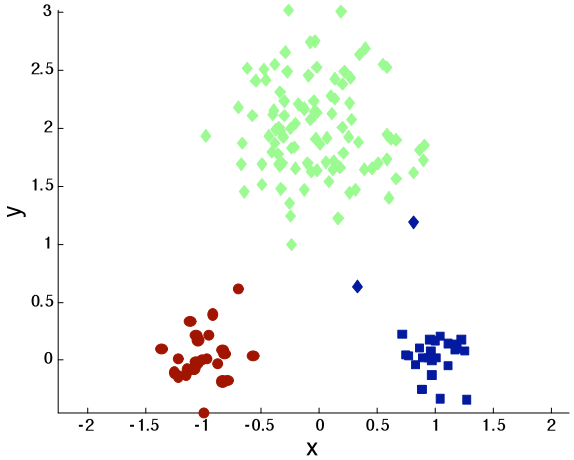


- I centri iniziali sono spesso scelti a caso.
  - I cluster prodotti possono variare da una esecuzione all'altra.
- Il centro è (tipicamente) la media dei punti nel cluster.
- La 'Distanza' tra punti può essere misurata dalla distanza euclidea (o dalla similarità del coseno, correlazione, ecc).
- K-medie convergerà per le più comuni misure di distanza.
- La maggior parte della convergenza avviene nelle prime iterazioni.
- La complessità è lineare in: **n**, **K**, **I**, **d**
  - **n** = numero dei punti, **K** = numero dei cluster,
    I = numero delle iterazioni, **d** = numero degli attributi

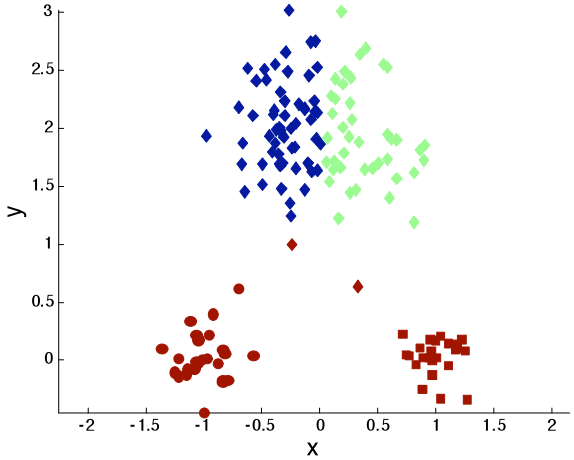
Due diversi clustering generati da K-medie



Punti originari

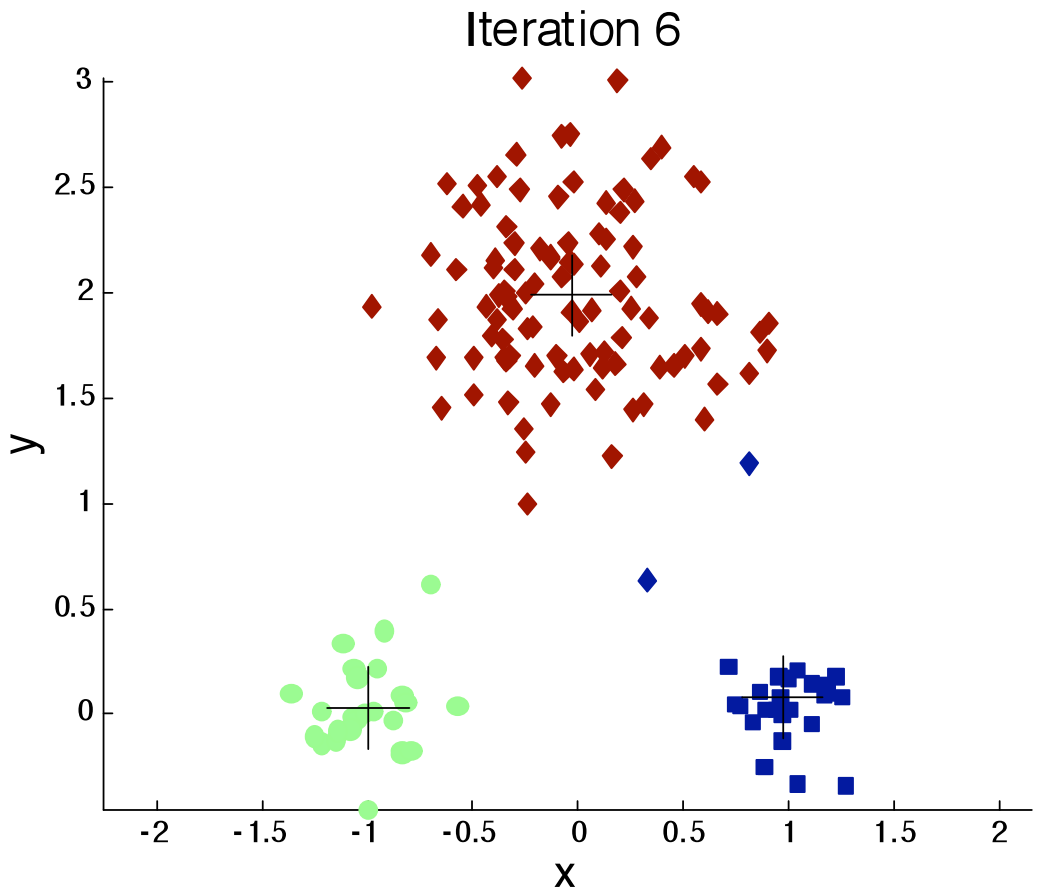


Clustering ottimale

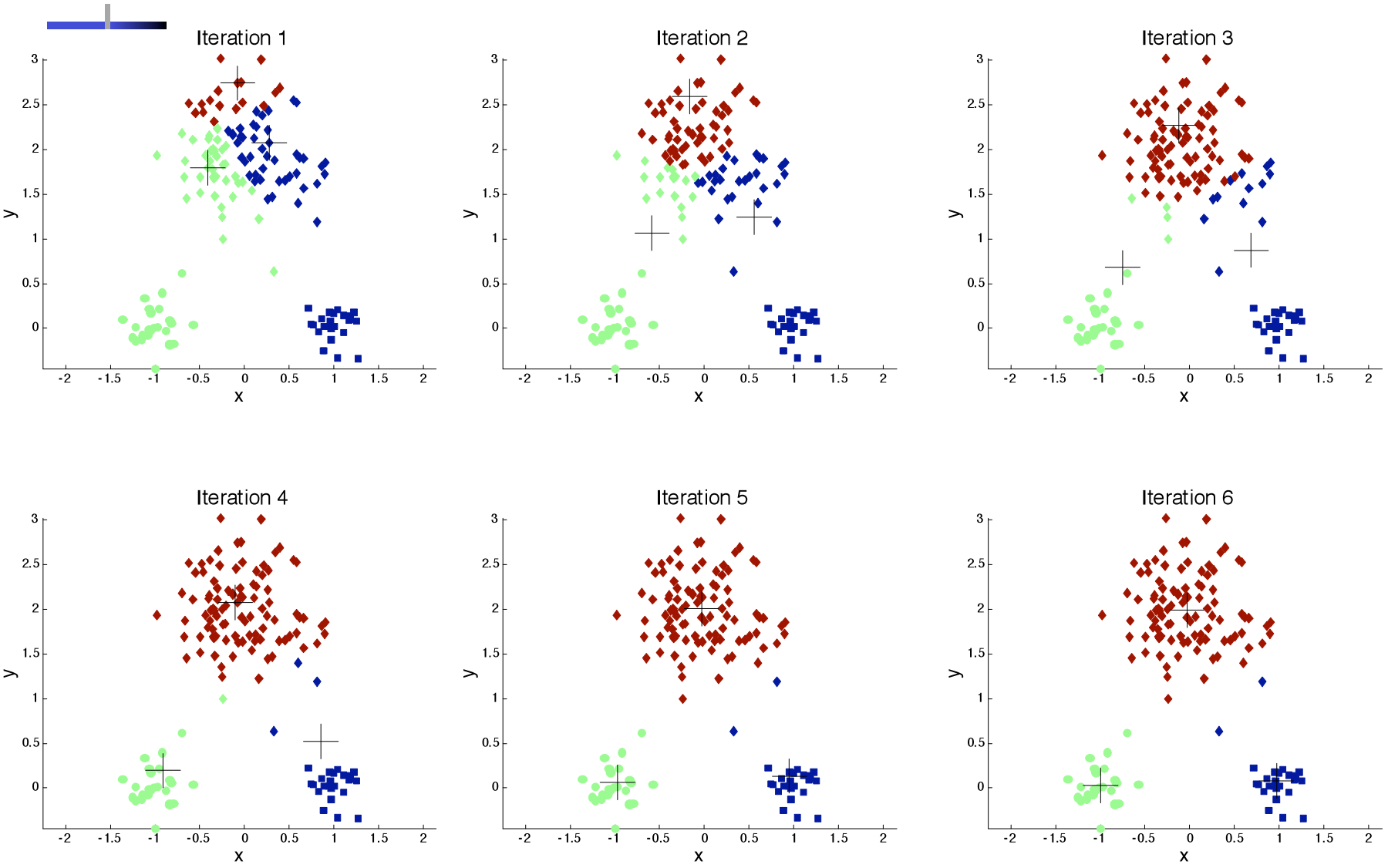


Clustering sotto-ottimale

Esempio di esecuzione

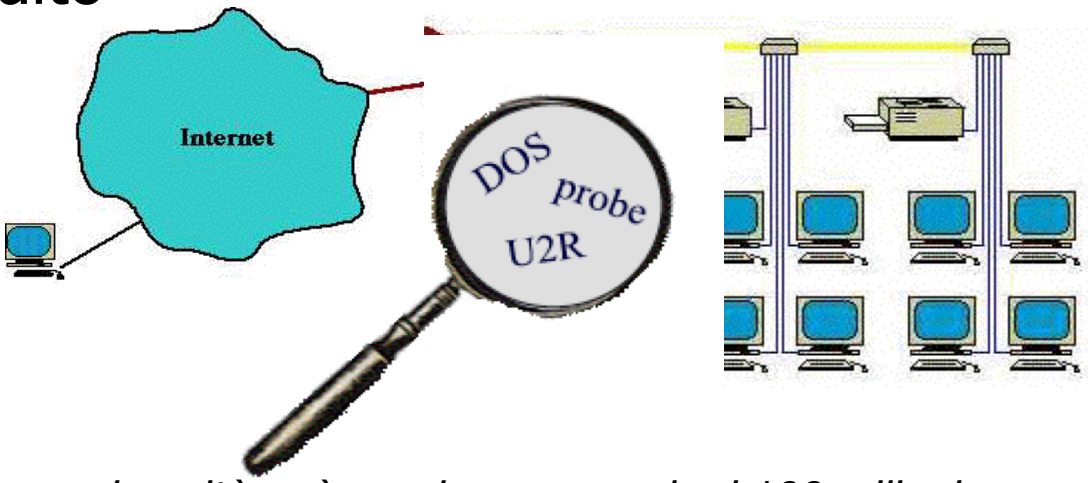


Importanza della scelta iniziale dei centri



Riconoscimento delle deviazioni/anomalie

- Riconoscere le deviazioni significative dal comportamento normale
- Applicazioni:
  - Frodi nelle carte di credito
  - Intrusione nelle reti di computer



Il volume di traffico tipico nella rete di una università può raggiungere anche i 100 milioni di connessioni al giorno

Riconoscimento e gestione delle frodi (1)

- Applicazioni
  - Ampiamente usata in sanità, vendita al dettaglio, servizi bancari con carta di credito, reti di telecomunicazione (frodi nelle carte telefoniche), ecc.
- Metodo
  - Usare i dati dello storico per costruire modelli di comportamento fraudolento e usare il clustering per identificare casi simili
- Esempi
  - Assicurazioni auto: scoperto un gruppo di persone che inscenano incidenti per ottenere il rimborso
  - Riciclaggio di denaro: scoperte le transazioni monetarie sospette (rete USA del crimine contro il tesoro e finanza)
  - Assicurazioni mediche: scoperta una rete di pazienti finti e dottori compiacenti

Riconoscimento e gestione delle frodi (2)

- Riconoscimento di trattamento medico inappropriato
  - La commissione per l'assicurazione medica australiana ha identificato molti casi di richieste di test medici fasulle (risparmio di 1milione di dollari all'anno).
- Riconoscimento di frodi telefoniche
  - Costruire il modello di chiamata: destinazione della chiamata, durata, ora del giorno e giorno della settimana. Analizzare le chiamate che deviano dalla norma.
  - British Telecom ha identificato un gruppo di clienti con chiamate frequenti intra-gruppo (specialmente cellulari), e ha fermato una frode multi milionaria.
- Vendite al dettaglio
  - Gli analisti stimano che il 38% del calo nelle vendite sia dovuto a dipendenti disonesti.

Altre applicazioni

- Sport

IBM Advanced Scout ha analizzato le statistiche di NBA (tiri bloccati, assist, e falli) per riuscire a ottenere un vantaggio competitivo su New York Knicks e Miami Heat

- Astronomia
  - JPL e l'Osservatorio di Palomar hanno scoperto ben 22 quasar con un classificatore

- Aiuto a chi naviga su Internet
  - IBM Surf-Aid applica algoritmi di data mining ai log dei server Web per scoprire le preferenze e il comportamento degli utenti che navigano sui siti Web dell'azienda, analizzando l'efficacia della pubblicità, migliorando l'organizzazione del sito, ecc.

Scoperta delle regole di association: definizione

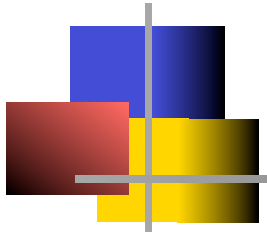
- Dato un insieme di record, ciascuno dei quali contiene un elenco dei prodotti acquistati da una certa offerta;
  - Produrre le regole di dipendenza che predirranno la presenza di un prodotto in una transazione di acquisto sulla base della presenza di altri.

TID	*Elenco Prodotti*
1	Pane, Coca-Cola, Latte
2	Birra, Pane
3	Birra, Coca-Cola, Pannoloni, Latte
4	Birra, Pane, Pannoloni, Latte
5	Coca-Cola, Pannoloni, Latte

Regole:

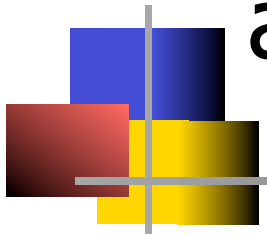
{Latte} --> {Coca-Cola}
{Pannoloni, Latte} --> {Birra}

Associazioni per la descrizione di concetti



- Descrizione di concetti: caratterizzazione e discriminazione
  - Generalizzare, riassumere, estrarre caratteristiche di contrasto tra concetti alternativi, ad es., regioni aride/umide
  - Metodo: organizzare in gruppi gli oggetti aventi le stesse caratteristiche di interesse;
  - Descrivere i gruppi tramite associazioni tra caratteristiche descrittive e il concetto di interesse
- Associazione (correlazione tra valori di attributi)
  - Associazioni multi-dimensionali / a singola dimensione
  - età="20..29" ^ stipendio= "20..29K" → acquisto="PC" [frequenza = 2%, confidenza = 60%]
  - acquisto="computer" → acquisto="software" [1%, 75%]

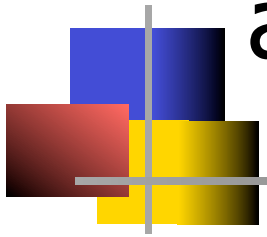
Scoperta di regole di associazione: applicazione 1



- Marketing e promozione delle vendite:
  - la regola scoperta sia

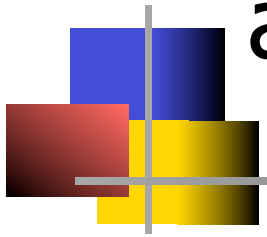
    {Pizzette, … } --> {Patatine}
  - Patatine come conseguente => può essere usata per determinare cosa dovrebbe essere fatto per aumentare le vendite.
  - Pizzette nell'antecedente => può essere usata per determinare di quali prodotti la vendita verrebbe influenzata se il negozio smettesse di vendere pizzette.
  - Pizzette nell'antecedente *e* Patatine nel conseguente => può essere usata per determinare quali prodotti potrebbero essere venduti con le Pizzette per promuovere le vendite di Patatine!

Scoperta di regole di associazione: applicazione 2



- Gestione degli scaffali del supermercato.
  - Scopo: identificare prodotti che vengono comprati insieme da un numero sufficiente di clienti.
  - Metodo: analizzare i dati delle vendite accumulati nei punti vendita in modo automatico tramite la lettura dei codici a barre dei prodotti e trovare le dipendenze tra i prodotti.
  - Una regola classica --
    - Se un cliente compra pannoloni e latte, allora verosimilmente comprerà anche birra.
    - Non siate sorpresi di trovare birre in pacchi da 6 accanto ai pannoloni per bambini!

Scoperta di regole di associazione: applicazione 3



- Gestione dell'inventario:
  - Scopo: una compagnia di riparazione di pezzi di elettrodomestici desidera anticipare la previsione delle necessità di riparazione e tenere pronti i veicoli con i pezzi di ricambio per ridurre il costo delle visite ai clienti.
  - Metodo: analizzare i dati sugli elettrodomestici e sui pezzi richiesti nelle precedenti visite presso i clienti e scoprire gli schemi di co-occorrenza tra elettrodomestici e pezzi.

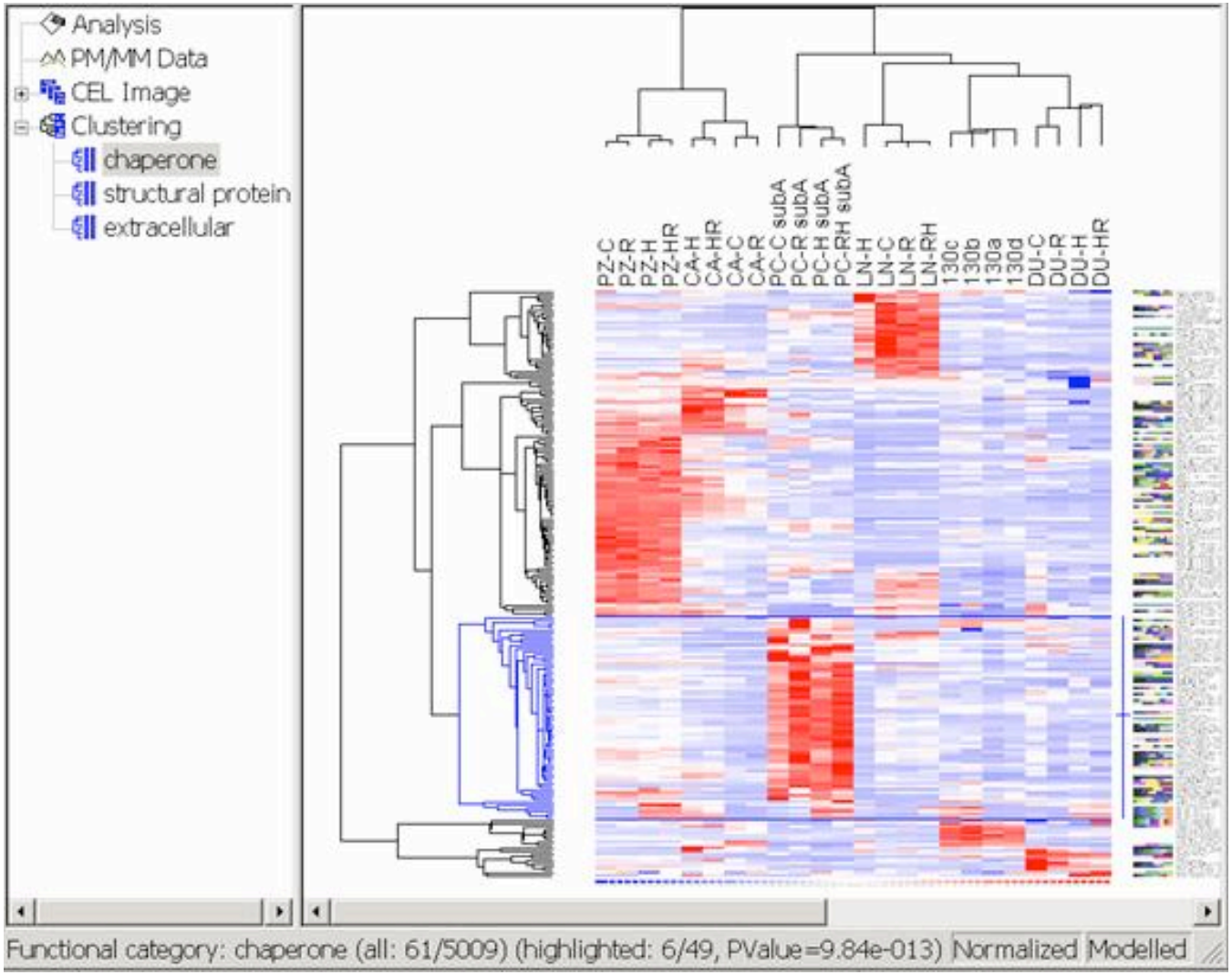
Scoperta di regole di associazione: applicazione 4

- Micro-array di espressione di DNA
  - Un Micro-array di DNA misura l' espressione di un gene in una cellula tramite la misurazione della quantità di RNA-messaggero presente per quel gene.
    - Le sequenze di nucleotidi per alcune migliaia di geni contenute in campioni di interesse (e in campioni di controllo sperimentale) vengono depositate su un vetrino
    - Poi, le sequenze sono attivate con il DNA (ibridizzate).
    - Infine attraverso una tecnica a fluorescenza si misurano le intensità di rosso/verde presenti nelle posizioni dove RNA si è manifestato.
    - Il risultato è una matrice di numeri reali (tipicamente variabile da –6 a 6) che misura il livello di espressione di ciascun gene nel campione di interesse relativamente al campione di controllo.

Scoperta di correlazioni tra geni

- Scopo:
  - Quali campioni sono più simili tra di loro in termini dei loro profili di espressione sui geni?
- Metodo: analizzare i dati e scoprire le correlazioni tra geni espressi nei diversi campioni
  - Quali geni sono più simili tra di loro, in termini dei propri profili di espressione attraverso i campioni?
- Metodo: analizzare i dati e scoprire i campioni cellulari che presentano gli stessi geni espressi
  - Certi geni presentano un livello di espressione molto alto (basso) per una certa tipologia di campioni (es., tumorali)?
- Metodo: Partizionare i campioni per tipologia (es., tumorale, sano).
  Trovare i geni che hanno gli stessi livelli di espressione nelle partizioni. Confrontare i livelli di espressione dei geni nella stessa partizione dei campioni cellulari.

Clustering gerarchico di geni

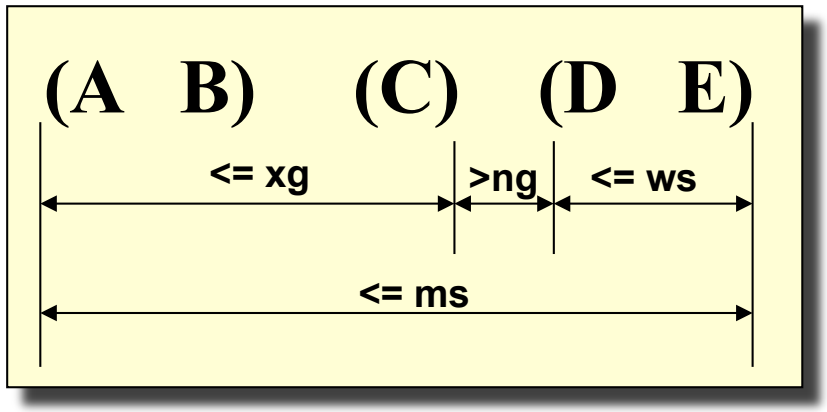


Scoperta di pattern sequenziali: definizione

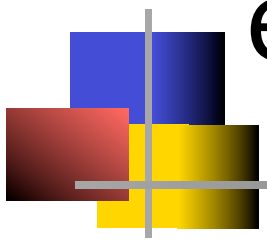
- Sia dato un insieme di oggetti rappresentativi di eventi, dove ciascun oggetto è associato ad un istante temporale in una dimensione temporale; trovare le regole che predicono una forte dipendenza sequenziale tra i diversi eventi.

$$(A\ \ B)\ \ \ (C) \longrightarrow (D\ \ E)$$

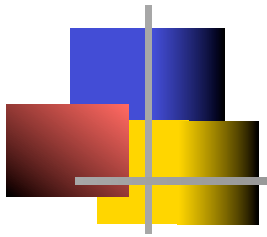
- Le regole vengono formate tramite l'estrazione di patterns. Le occorrenze di eventi nei patterns sono governate da vincoli temporali.



Scoperta di pattern sequenziali: esempi



- Nelle reti di telecomunicazione, si analizzano i log degli allarmi,
  - (problema all'Inverter Eccessiva_Corrente)
    (Allarme_Raddrizzatore) --> (Allarme_Incendio)
- Nelle sequenze degli acquisti nelle transazioni di cassa,
  - Nel negozio di libri:
    (Introduzione_alla_programmazione) (Programmare in C++) -->
    (Perl_per_principianti)
  - Nel negiozio di articoli sportivi:
    (scarpe) (racchetta) --> (completo sportivo)

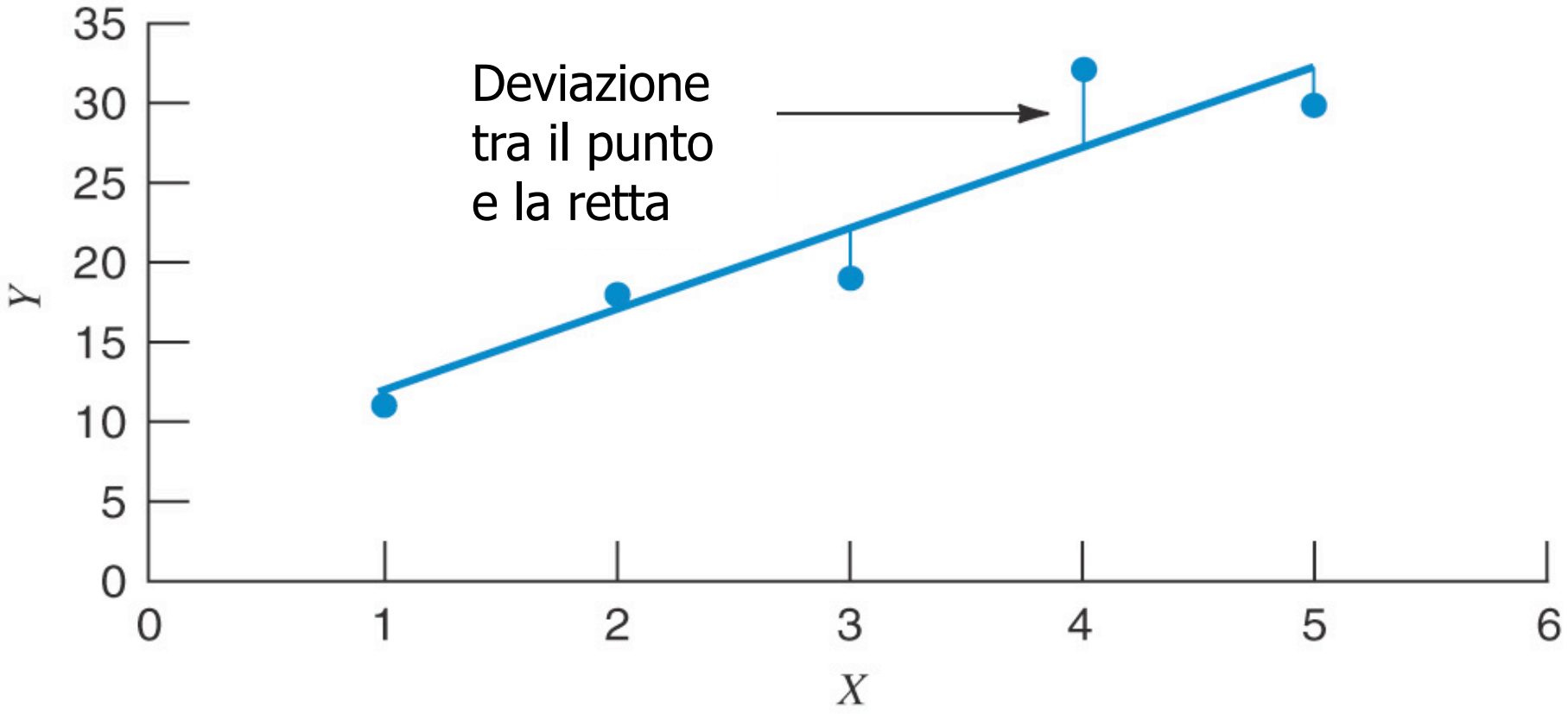


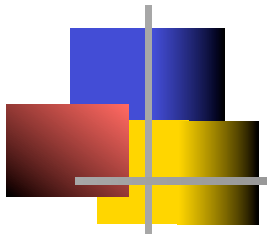
Regressione

- Predire il valore di una variabile data a valori continuisulla base dei valoridi altre variabili, assumendo un modello di dipendenza lineare o non-lineare.
- Molto studiata in statistica, nelle reti neurali.
- Esempi:
  - Predire l'incasso conseguente alle vendite di un nuovo prodottosulla base della spesa in pubblicità.
  - Predire la velocità del vento, in funzione della temperatura, umidità, pressione, ecc.
  - Predizione di una serie temporale (es., negli indici di borsa).

Regressione

Linea dei minimi scostamenti quadratici



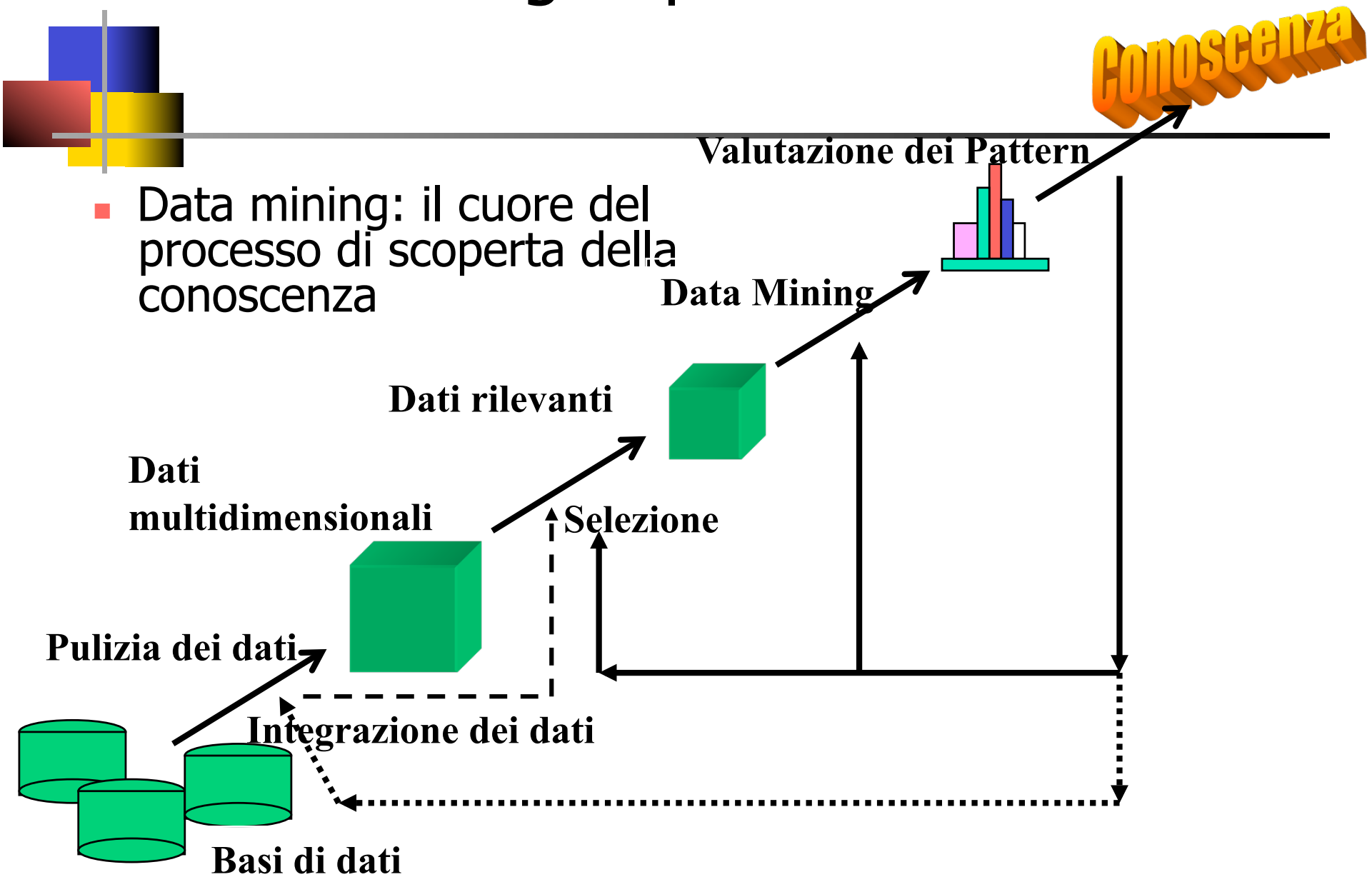


Sfide del Data Mining

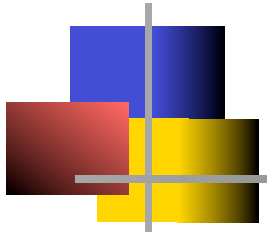
- Scalabilità
- Dimensionalità
- Dati complessi ed eterogenei
- Qualità dei dati
- Diritti sui dati e distribuzione
- Attenzione alla privacy
- Dati in Streaming

Data Mining: il processo di KDD

- Data mining: il cuore del processo di scoperta della conoscenza



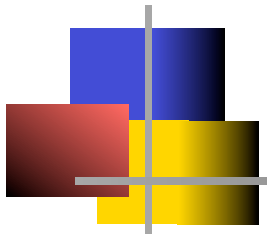
Passi del processo di KDD



- Comprendere il dominio applicativo:
  - La conoscenza a priori e gli obiettivi dell'applicazione
- Creare un set di dati rilevanti per l'obiettivo d'interesse: selezione dei dati
- Pulizia dei dati: (potrebbe riguardare anche 60% dello sforzo!)
- Riduzione dei dati e trasformazione:
  - trovare le caratteristiche utili, riduzione del numero delle caratteristiche, reppresentazione alternativa dei dati.
- Scelta delle funzioni di data mining
  - riassunto, classificazione, regressione, associazione, clustering.
- Scelta degli algoritmi di data mining
- Data mining: ricerca dei pattern di interesse
- Valutazione dei pattern e presentazione della conoscenza
  - visualizzazione, trasformazione, rimozione dei pattern ridondanti
- Uso della conoscenza estratta

Tutti i pattern sono interessanti?

- Un sistema di data mining potrebbe generare migliaia di patterns; not tutti sono interessanti.
  - Metodo suggerito: centrato sull'interazione uomo-macchina, basato sull'interrogazione, focalizzato all'obiettivo
- **Misure di interesse**: un pattern è interessante se è facilmente compreso dall'utente, valido su nuovi dati (di test) con un certo livello di confidenza, potenzialmente utile, nuovo, o serve per validare qualche ipotesi che l'utente desidera confermare
- **Misure di interesse obiettive o soggettive:**
  - Obiettive: basate su statistiche e strutture dei pattern, es., frequenza, errore, ecc.
  - Soggettive: utili in riferimento alle conoscenze dell'utente, sorpresa, ecc.

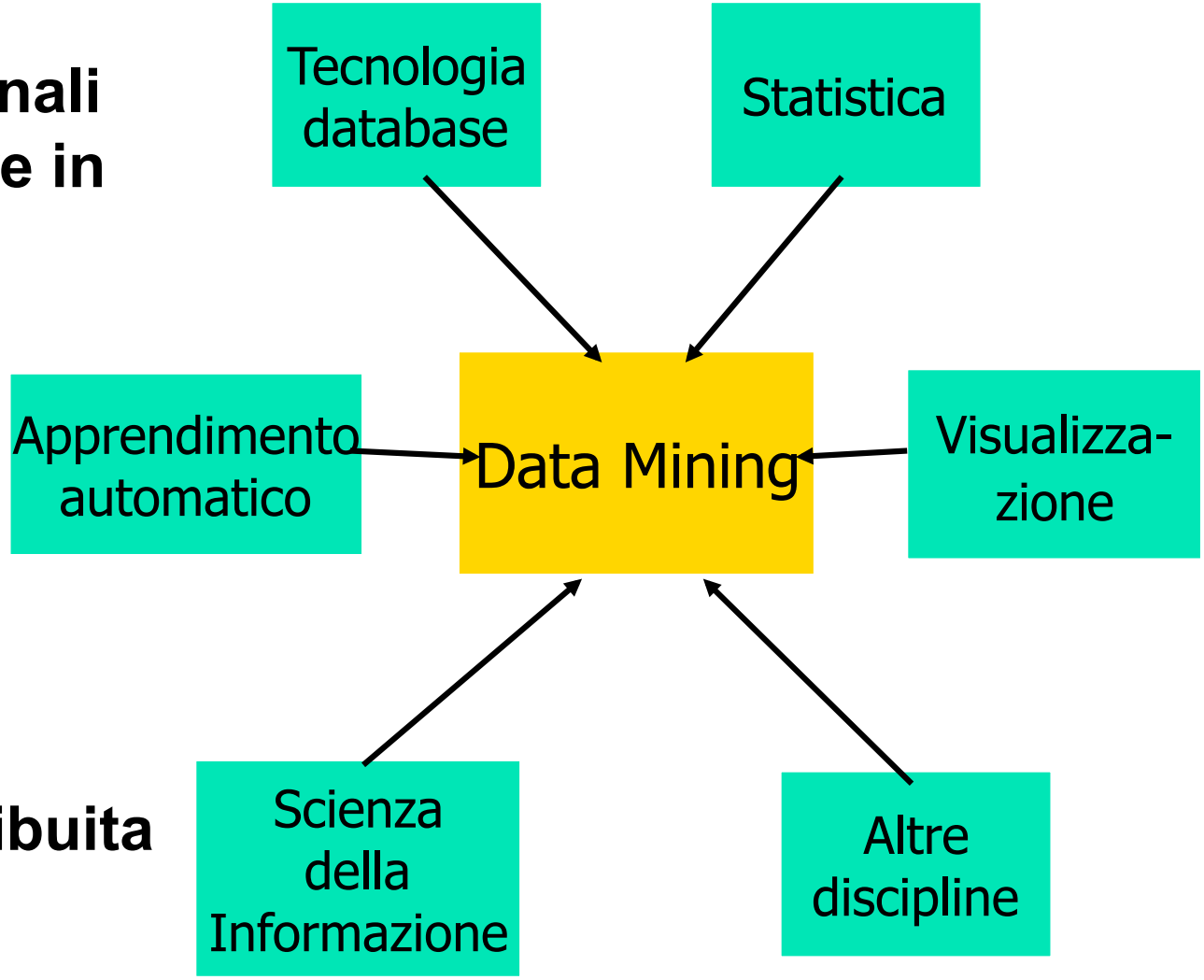


Possiamo trovare *tutti* i pattern interessanti e *solo* quelli?

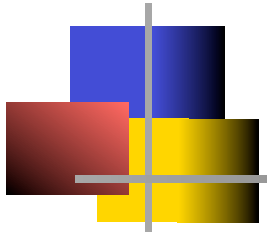
- Trovare tutti i pattern interessanti: completezza
  - Può un sistema di data mining trovare tutti i pattern interessanti? (Richiamo)
  - Associazione / classificazione / clustering
- Ricerca dei soli pattern interessanti: ottimizzazione
  - Può un sistema di data mining trovare solo i pattern interessanti? (Precisione)
  - Metodo
    - 1. Genera tutti i pattern e poi elimina quelli inutili.
    - 2. Genera solo i pattern interessanti — ottimizzazione.

Data Mining: alla confluenza di molte discipline

- **Le tecniche tradizionali potrebbero non essere in grado di trattare:**
  - **Enormi volumi di dati**
  - **Elevata dimensionalità**
  - **Dati di natura eterogenea e distribuita**



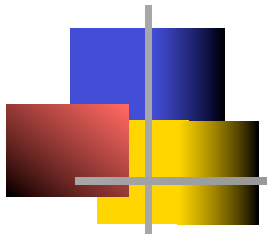
Principali tematiche del Data Mining (1)



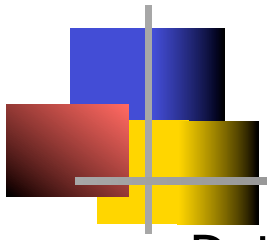
- Metodologia di indagine e interazione con l'utente
  - Estrazione di tipi diversi di conoscenza dai dati
  - Estrazione interattiva di conoscenza a molteplici livelli di astrazione
  - Incorporare la conoscenza del dominio
  - Linguaggi di interrogazione e analisi ad-hoc (specifiche per dominio)
  - Visualizzazione e presentazione dei risultati dell'analisi
  - Gestione di dati rumorosi (sporchi) e incompleti
  - Valutazione dei pattern di conoscenza: il problema della specifica dei pattern interessanti

  Efficienza e scalabilità
  - Efficienza e scalabilità degli algoritmi
  - Metodi paralleli, distribuiti e incrementali

Principali tematiche del Data Mining (2)



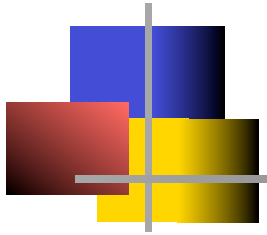
- Tematiche relative alla diversità dei tipi di dati
  - Gestione di dati relazionali (a tabella) e più complessi (documenti)
  - Estrarre informazione da dati eterogenei e da Internet
- Tematiche relative ad applicazioni con impatto sociale
  - Applicazione della conoscenza estratta
    - Metodi specifici a un certo dominio di applicazione
    - Metodi di interrogazione intelligente (ottimizzato)
    - Controllo del processo e supporto alle decisioni (cruscotti decisionali)
  - Integrazione della conoscenza estratta con quella preesistente: un problema di fusione di conoscenza
  - Protezione della sicurezza, integrità e privacy dei dati



Riassunto

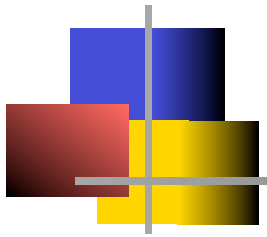
- Data mining: scoperta di pattern interessanti da grossi volumi di dati
- Una naturale evoluzione della tecnologia delle basi di dati, molto diffuse e con ampie applicazioni
- Il processo di KDD include la pulizia dei dati, integrazione, selezione, transformazione, il passo di data mining vero e proprio, la valutazione dei pattern estratti e la presentazione dei risultati
- L'analisi può essere rivolta a grandi varietà di dati
- Funzionalità di Data mining: caratterizzazione, discriminazione, associazione, classificazione, clustering, analisi delle anomalie e dei trend, ecc.
- Classificazione dei sistemi di data mining
- Principali tematiche del data mining

Una breve storia del Gruppo di interesse in Data Mining (SIG)

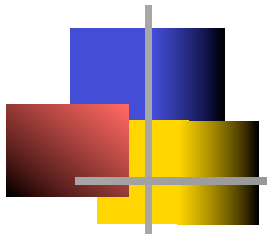


- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, e SIGKDD Explorations
- Altre conferenze in data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, ecc.

Dove trovare materiale e riferimenti bibliografici?



- Data mining and KDD (SIGKDD member CDROM):
  - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery
- Database field (SIGMOD member CD ROM):
  - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
- AI and Machine Learning:
  - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics:
  - Conference proceedings: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization:
  - Conference proceedings: CHI, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.



Riferimenti bibliografici

- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.
- T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.
- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.