La Sapienza
Università degli Studi di Roma

Dipartimento
di Informatica e Sistemistica

I vantaggi ottenibili nei campi applicativi attraverso l'uso di tecniche di data mining

Renato Bruni
bruni@dis.uniroma1.it

Antonio Sassano
sassano@dis.uniroma1.it

Seminario Metodi e Algoritmi per il Controllo e la Correzione dei Dati delle Famiglie, Roma, 29 Marzo 2004

Sommario

- Cosa si intende con Data Mining?
- Il processo di Knowledge Discovey
- Diversi aspetti del Data Mining
- Vari algoritmi di Data Mining

Cos'è il Data Mining?

- Problema dell'**esplosione dei dati**:

  L'uso dell'archiviazione informatica unita alla tecnologia dei database porta ad avere collezioni di informazione di dimensioni impressionanti.

- We are drowning in data, but starving for **knowledge**!

  Avere molti dati è un vantaggio, ma complica la loro gestione.
  Per utilizzarli occorrono strumenti sempre più sofisticati.

- **Data mining** (knowledge discovery in databases):

  Estrazione di informazioni non ovvie, precedentemente non note e potenzialmente utili (regole, regolarità, pattern, etc. = conoscenza) contenute in grandi quantità di dati.

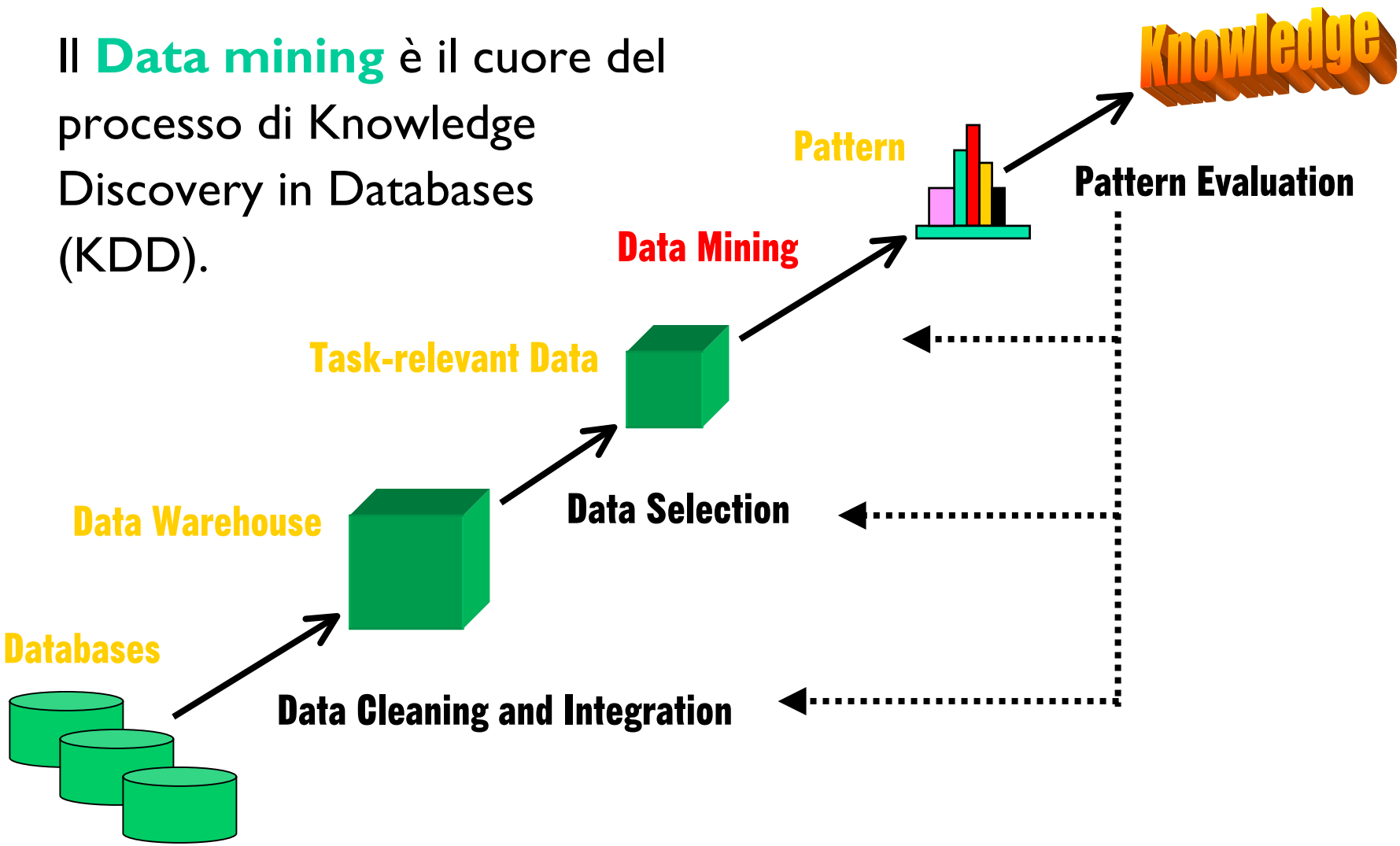
Perché Data Mining? - Applicazioni

Presentare **conoscenza** anziché soltanto **dati** è essenziale in una serie di situazioni:

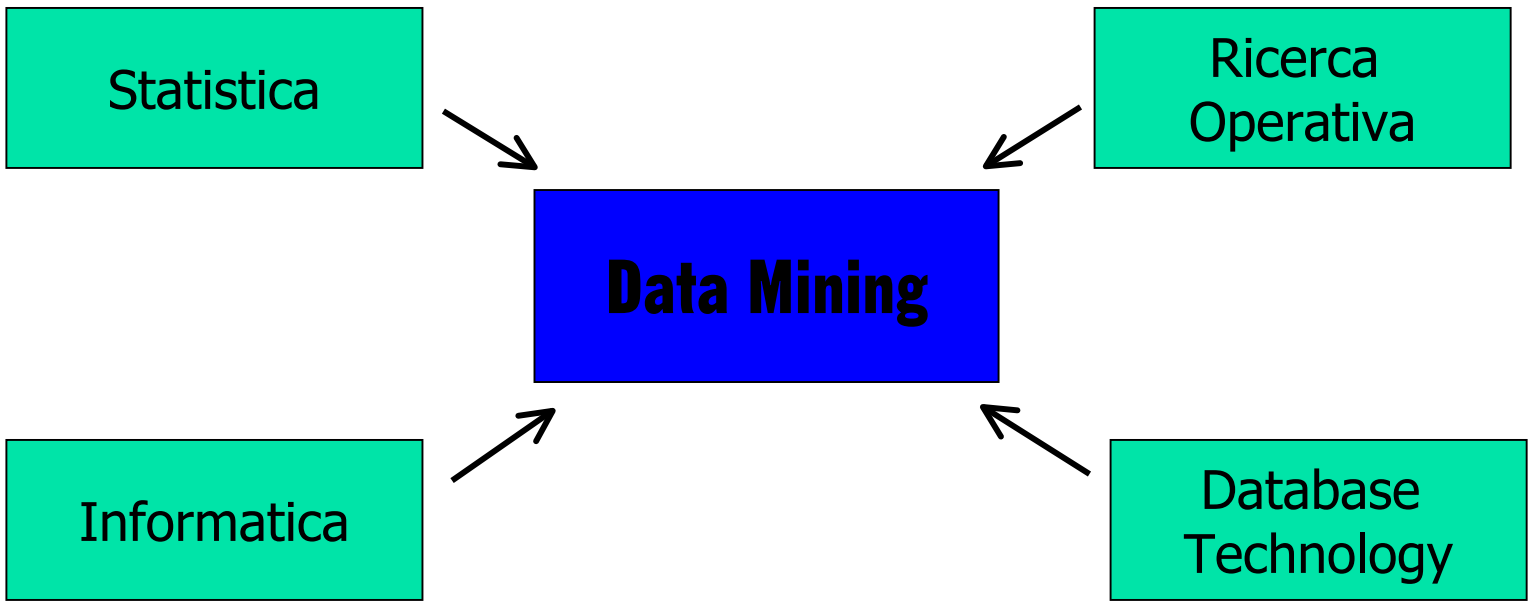
- Analisi di Database (Estrazione di regole, Associazioni)
- Analisi di Mercato (Customer profiling, Marketing)
- Analisi di Rischio (Finance planning, Investimenti)
- Individuazione di Frodi (Carte di credito, Sofisticazioni alimentari)
- Supporto alle Decisioni (Resource management, Allocazione)
- Analisi Mediche (Diagnosi, Gestione donatori)
- Text mining (news-group, email, documents) nel Web
- Analisi di Politiche Economiche o Sociali (Rule learning)
- Analisi di Eventi Rari …

Il Data Mining nel processo di KDD

Il **Data mining** è il cuore del processo di Knowledge Discovery in Databases (KDD).



Data Mining: Confluenza di Discipline



Servono modelli adeguati, algoritmi efficienti, gestione delle informazioni evoluta, presentazione dei risultati fruibile.
I quattro aspetti sono sinergici e **complementari**.

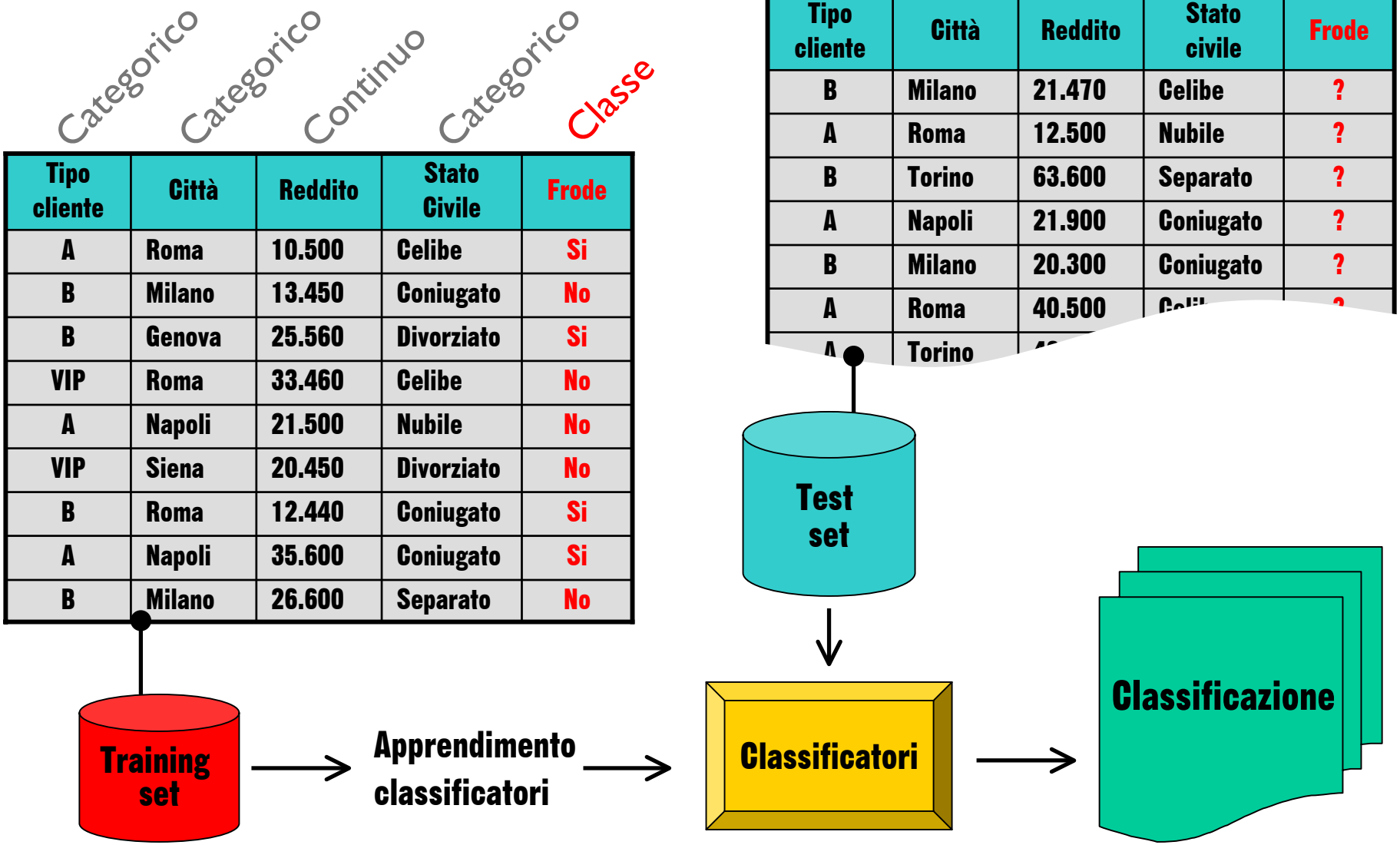
Diversi Aspetti del Data Mining

- **Classificazione**: apprendimento di una funzione per mappare oggetti in un insieme predefinito di classi.
- **Regressione**: apprendimento di una funzione per mappare un oggetto in un valore reale.
- **Clustering**: identificazione di una collezione di gruppi di oggetti simili.
- **Apprendimento di Dipendenze e Associazioni**: identificazione di dipendenze significative tra gli attributi dei dati.
- **Apprendimento di Regole e Sommarizzazione**: individuazione di una descrizione compatta di un insieme o sottoinsieme di dati.

Es. Classificazione: Fraud Detection

Tipo cliente (Categorico)	Città (Categorico)	Reddito (Continuo)	Stato Civile (Categorico)	Frode (Classe)
A	Roma	10.500	Celibe	Si
B	Milano	13.450	Coniugato	No
B	Genova	25.560	Divorziato	Si
VIP	Roma	33.460	Celibe	No
A	Napoli	21.500	Nubile	No
VIP	Siena	20.450	Divorziato	No
B	Roma	12.440	Coniugato	Si
A	Napoli	35.600	Coniugato	Si
B	Milano	26.600	Separato	No

Tipo cliente	Città	Reddito	Stato civile	Frode
B	Milano	21.470	Celibe	?
A	Roma	12.500	Nubile	?
B	Torino	63.600	Separato	?
A	Napoli	21.900	Coniugato	?
B	Milano	20.300	Coniugato	?
A	Roma	40.500		
A	Torino			



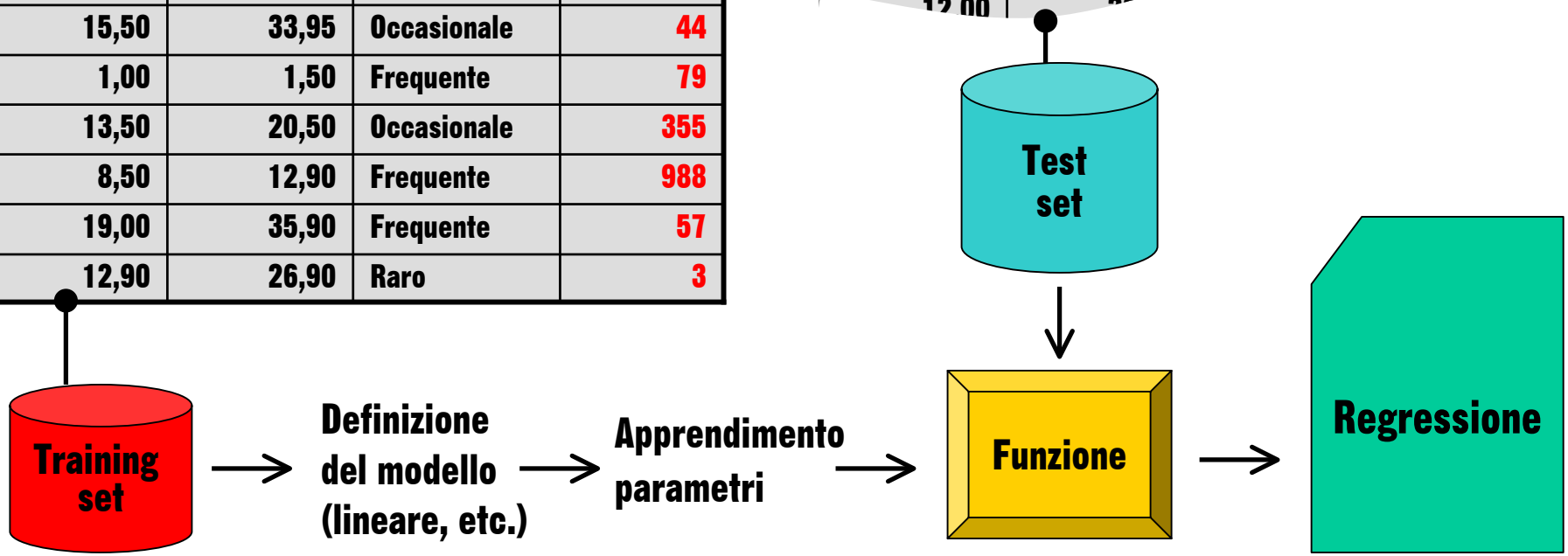
Es. Regressione: Predizione Vendite

Variabili predittrici

Variabile dipendente (numerica)

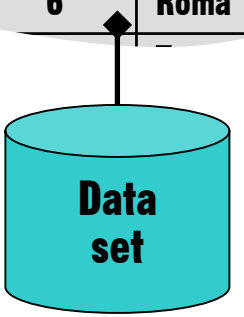
Costo Materiale	Prezzo di Vendita	Uso	Vendite
5,00	11,50	Frequente	154
6,00	12,80	Raro	21
15,50	25,50	Frequente	234
15,50	33,95	Occasionale	44
1,00	1,50	Frequente	79
13,50	20,50	Occasionale	355
8,50	12,90	Frequente	988
19,00	35,90	Frequente	57
12,90	26,90	Raro	3

Costo Materiale	Prezzo di Vendita	Uso	Vendite
10,00	19,90	Frequente	?
5,50	11,00	Occasionale	?
14,50	25,90	Occasionale	?
63,00	128.00	Raro	?
2,50	4,90	Frequente	?
24,00	49,90	Occasi...	
12.00			



Es. Clustering: Market Segmentation

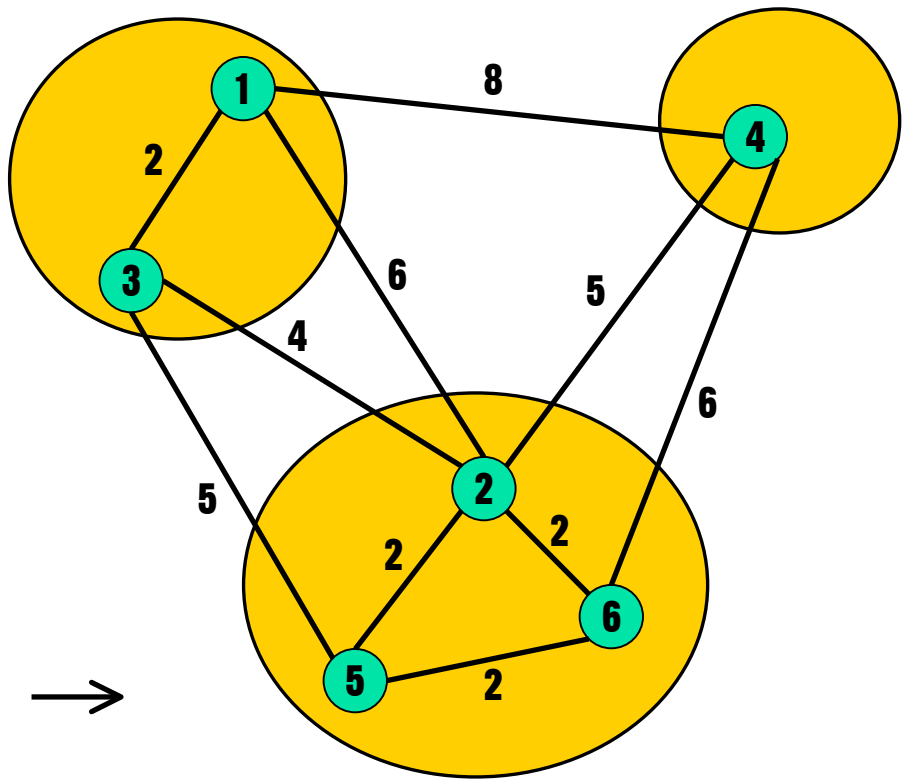
ID cliente	Città	Reddito	Stato civile	Pagamento
1	Milano	21.470	Celibe	2.500
2	Roma	12.500	Nubile	400
3	Torino	63.600	Separato	250
4	Napoli	21.900	Coniugata	12.000
5	Milano	20.300	Coniugato	645
6	Roma	40.500		



Calcolo della similitudine tra clienti ottenendo delle distanze →

Partizionamento in k gruppi minimizzando le distanze tra record dello stesso gruppo →

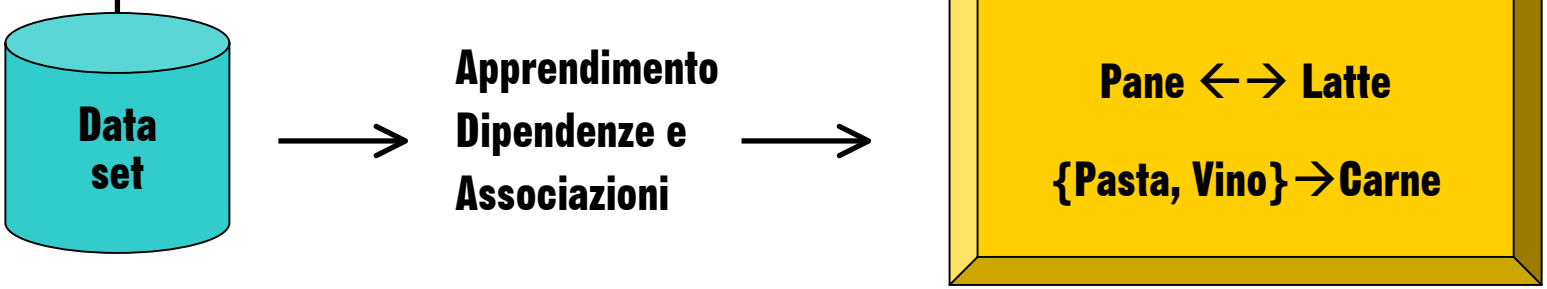
Voglio dividere i clienti in k=3 gruppi che verranno trattati diversamente.



Es. Associazione: Acquisti di Beni

ID	Oggetti Acquistati
1	Pane, Latte, Uova
2	Pane, Pasta, Biscotti, Latte
3	Carne, Formaggio
4	Pane, Formaggio, Latte
5	Pasta, Pane, Vino, Latte, Carne
6	Pasta, Vino, Carne

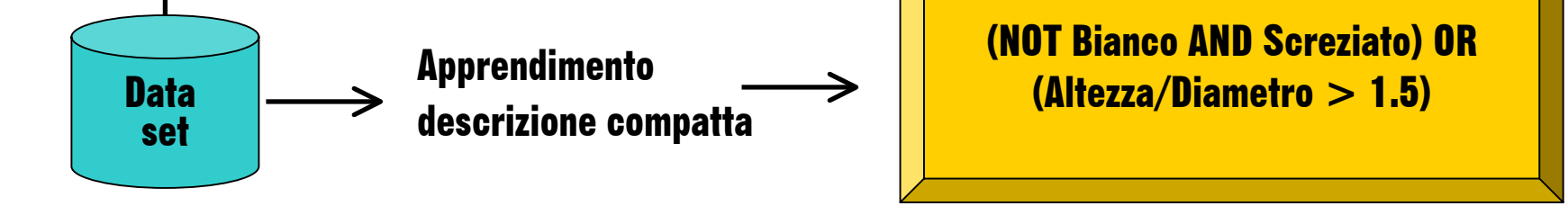
Dato un insieme di record, ognuno contenente oggetti appartenenti a un elenco, individuare regole di dipendenza per predire l'occorrenza di in oggetto in base all'occorrenza di altri oggetti.



Es. Sommarizzazione: Funghi Velenosi

Categorico	Categorico	Continuo	Continuo
Colore	**Superficie**	**Diametro**	**Altezza Gambo**
Rosso	Screziata	13	5
Bianco	Liscia	4	7
Grigio	Screziata	10	8
Grigio	Liscia	6	12
Rosso	Screziata	10	10
Bianco	Screziata	5	9
Grigio	Liscia	6	10
Bianco	Screziata	3	6
Rosso	Liscia	10	16

Dati record rappresentanti funghi velenosi, trovarne una descrizione compatta.
La conoscenza così trovata può poi essere usata in molti modi.



"Tortura i dati finché non confessano"

Tecniche di Data Mining (Cenni)

Sono state proposte molte e diverse tecniche, aventi ognuna specifiche **caratteristiche** e **vantaggi**.

Alberi di Decisione: Classificazione, Sommarizzazione, es: C4.5, CART, IC3, Entropia, CHAID.

Analisi logica e Programmazione Intera: Classificazione, Apprendimento di Regole, es: LAD.

Teoria dei Grafi: Clustering, Classificazione, es: B&C.

Rappresentazioni analitiche dei dati (OLAP): Sommarizzazione, Database streaming.

Reti neurali (ANN): Classificazione, es: Perceptron, a singolo strato, multi-strato, Backpropagation, Radial-Basis Function (RBF) networks, es: SNNS, Nevprop.

Metodi Bayesiani: Regressione, Classificazione, Bayesian Learning, Bayesian belief network, Bayesian Classifiers, Maximum Likelihood.

Support Vector Machines (SVM): Classificazione, Pattern recognition, es: RSVM.

Association/Pattern Discovery: Regole di associazione e dipendenze, pattern sequenziali, es: CN2.

Conclusioni

- In molti casi è **essenziale** essere in grado di estrarre della conoscenza dai dati, cioè fare Data Mining. Chi possiede i dati **potrebbe identificarsi** con chi ne estrae la conoscenza.
- All'aumentare delle **dimensioni** degli insiemi di dati queste operazioni richiedono strumenti automatici sempre più **sofisticati**.
- Per fare Data Mining sono necessarie competenze di **varie discipline**: in particolare, sono necessari modelli adeguati, algoritmi efficienti, implementazione evoluta, tecnologia di gestione dei dati sofisticata.
- Siamo di nuovo a un **inizio**?

References

- 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.
- D. Hand, H. Mannila and P. Smyth. Principles of Data Mining. The MIT Press, 2001.
- V.N. Vapnik. Statistical Learning Theory. Wiley & Sons, 1998.
- T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning. Springer 2001.
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.