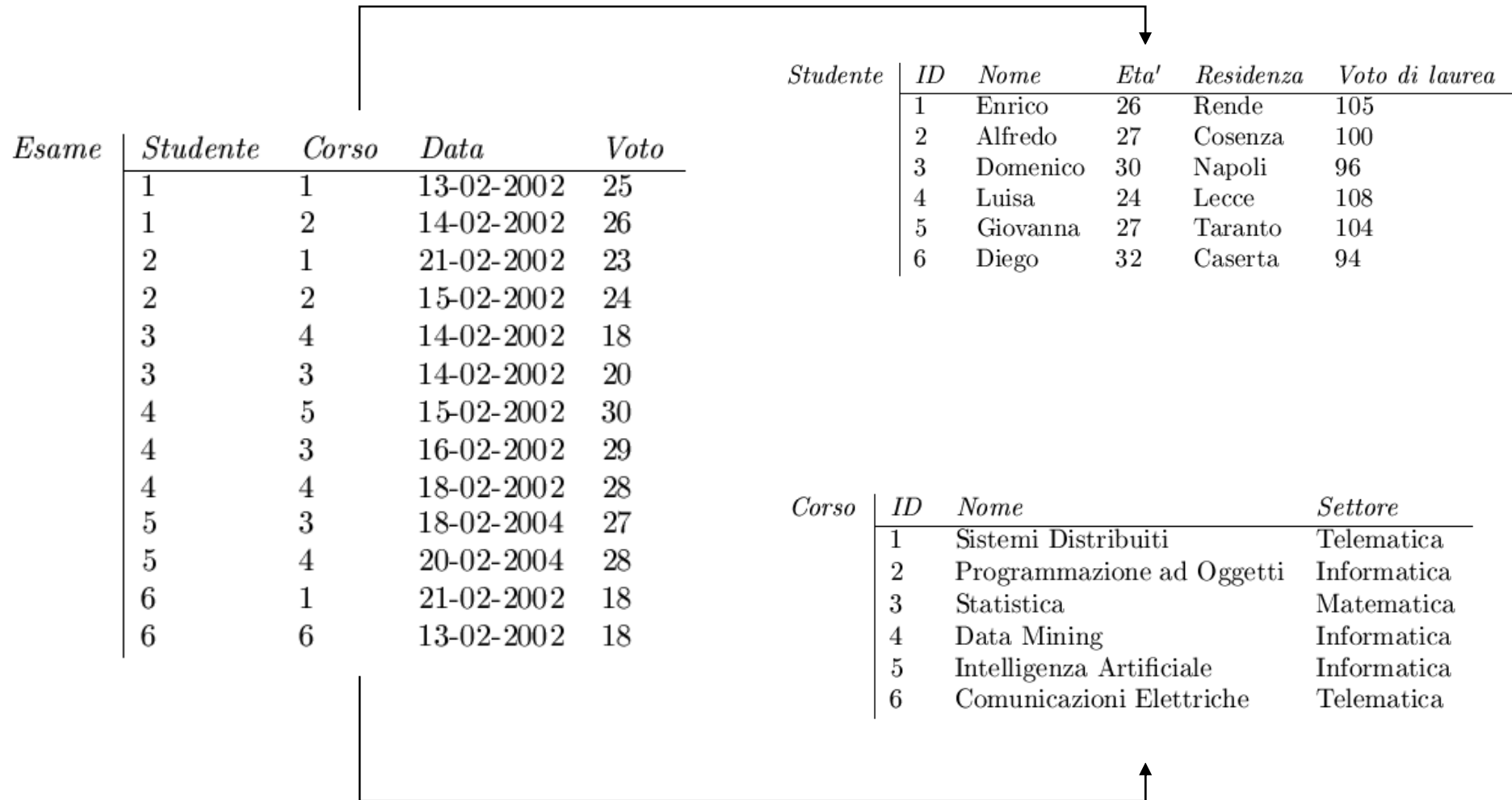


Tecniche Di Data Mining per l'Analisi dei Dati

Giuseppe Manco



Dall'inferenza all'induzione...



• Normalizzazione

$$\forall s, e, d, v \quad (\text{Esame}(s, e, d, v) \rightarrow \exists x, y, z, c (\text{Studente}(s, x, y, z) \wedge \text{Corso}(e, c))))$$

– Essenzialmente, per il Transaction Processing

• Queries (semplici)

- Che voti hanno avuto negli esami di informatica gli studenti di Cosenza?
- Che media hanno gli studenti del corso di Data Mining?
- Qual è lo studente che ha avuto il voto più alto?

Dalle queries al supporto alle decisioni

- **Quanti sono gli studenti che hanno ottenuto un voto di laurea superiore a 100 e che hanno fatto esami sia in informatica che in statistica?**
- **Qual è l'andamento temporale della media dei voti in Informatica e Matematica, rispetto alla media in telematica?**

L'analisi dei dati

- Come sono fatti gli studenti che hanno i voti alti?

					Studente				
					ID	Nome	Eta'	Residenza	Voto di laurea
					1	Enrico	26	Rende	105
					2	Alfredo	27	Cosenza	100
					3	Domenico	30	Napoli	96
					4	Luisa	24	Lecce	108
					5	Giovanna	27	Taranto	104
					6	Diego	32	Caserta	94

Esame	Studente	Corso	Data	Voto
5	3	18-02-2004	27	
5	4	20-02-2004	28	
6	1	21-02-2002	18	
6	6	13-02-2002	18	

			Corso		
ID	Nome	Settore			
1	Sistemi Distribuiti	Telematica			
2	Programmazione ad Oggetti	Informatica			
3	Statistica	Matematica			
4	Data Mining	Informatica			
5	Intelligenza Artificiale	Informatica			
6	Comunicazioni Elettriche	Telematica			

L'analisi dei dati

- **Come sono fatti gli studenti che hanno i voti alti?**

<i>ID</i>	<i>Nome</i>	<i>Eta'</i>	<i>Residenza</i>	<i>Telematica</i>	<i>Informatica</i>	<i>Matematica</i>	<i>Voto di laurea</i>
1	Enrico	26	Rende	Si	Si	No	Alto
2	Alfredo	27	Cosenza	Si	Si	No	Medio
3	Domenico	30	Napoli	No	Si	Si	Basso
4	Luisa	24	Lecce	No	Si	Si	Alto
5	Giovanna	27	Taranto	No	Si	No	Alto
6	Diego	32	Caserta	Si	No	No	Basso

L'analisi dei dati

- **Come sono fatti gli studenti che hanno i voti alti?**

<i>ID</i>	<i>Nome</i>	<i>Eta'</i>	<i>Residenza</i>	<i>Telematica</i>	<i>Informatica</i>	<i>Matematica</i>	<i>Voto di laurea</i>
1	Enrico	26	Calabria	Si	Si	No	Alto
2	Alfredo	27	Calabria	Si	Si	No	Medio
3	Domenico	30	Campania	No	Si	Si	Basso
4	Luisa	24	Puglia	No	Si	Si	Alto
5	Giovanna	27	Puglia	No	Si	No	Alto
6	Diego	32	Campania	Si	No	No	Basso

L'analisi dei dati

- **Come sono fatti gli studenti che hanno i voti alti?**

<i>ID</i>	<i>Nome</i>	<i>Eta'</i>	<i>Residenza</i>	<i>Telematica</i>	<i>Informatica</i>	<i>Matematica</i>	<i>Voto di laurea</i>
1	Enrico	24-26	Calabria	Si	Si	No	Alto
2	Alfredo	27-29	Calabria	Si	Si	No	Medio
3	Domenico	30-32	Campania	No	Si	Si	Basso
4	Luisa	24-26	Puglia	No	Si	Si	Alto
5	Giovanna	27-29	Puglia	No	Si	No	Alto
6	Diego	30-32	Campania	Si	No	No	Basso

L'analisi dei dati

- Quali esami vengono sostenuti insieme di solito?

Esame					Studente						
ID	Sist. Distr.	Progr. a Oggetti	Statistica	Data Mining	Int. Artificiale	Com. El.	ID	Nome	Eta'	Residenza	Voto di laurea
1	Si	Si	No	No	No	No	1	Enrico	26	Rende	105
2	Si	Si	No	No	No	No	2	Alfredo	27	Cosenza	100
3	No	No	Si	Si	No	No	3	Domènec	28	Montebelluna	98
4	No	No	Si	Si	Si	No	4	Enrico	27	Montebelluna	100
5	No	No	Si	Si	No	No	5	Enrico	27	Montebelluna	100
6	Si	No	No	No	No	Si	6	Enrico	27	Montebelluna	100

5	3	18-02-2004	27	1	Sistemi Distribuiti	Telematica
5	4	20-02-2004	28	2	Programmazione ad Oggetti	Informatica
6	1	21-02-2002	18	3	Statistica	Matematica
6	6	13-02-2002	18	4	Data Mining	Informatica
				5	Intelligenza Artificiale	Informatica
				6	Comunicazioni Elettriche	Telematica

L'analisi dei dati

- Ci sono tendenze nel comportamento degli studenti?

Esame		Studente	Corso	Data	Voto	
1	1					
1	1	ID	Eta'	Telematica	Informatica	Matematica
2	2	1	26	25	26	0
2	2	2	27	23	24	0
3	3	3	30	0	18	20
3	3	4	24	0	28.5	30
4	4	5	27	0	27	28
4	4	6	32	18	0	18
5	5					
5	5	4	20-02-2004	28		
6	6	1	21-02-2002	18		
6	6	6	13-02-2002	18		

Studente	ID	Nome	Eta'	Residenza	Voto di laurea
1	1	Enrico	26	Rende	105
2	2	Alfredo	27	Cosenza	100
3	3	Domenico	30	Napoli	96
				ce	108
				anto	104
				serta	94

	Settore	
1	Sistemi Distribuiti	Telematica
2	Programmazione ad Oggetti	Informatica
3	Statistica	Matematica
4	Data Mining	Informatica
5	Intelligenza Artificiale	Informatica
6	Comunicazioni Elettriche	Telematica

L'analisi dei dati

- **Ci sono tendenze nel comportamento degli studenti?**

<i>ID</i>	<i>Eta'*</i>	<i>Telematica</i>	<i>Informatica</i>	<i>Matematica</i>
1	0.5	25	26	0
2	0.625	23	24	0
3	0.75	0	18	20
4	0	0	28.5	30
5	0.625	0	27	28
6	1	18	0	18

L'analisi dei dati

- **Ci sono tendenze nel comportamento degli studenti?**

<i>ID</i>	<i>Eta'</i> *	<i>Telematica</i>	<i>Informatica</i>	<i>Matematica</i>
4	0	0	28.5	30
1	0.5	25	26	0
2	0.625	23	24	0
5	0.625	0	27	28
3	0.75	0	18	20
6	1	18	0	18

L'analisi dei dati

- **Ci sono tendenze nel comportamento degli studenti?**

<i>ID</i>	<i>Eta'</i> *	<i>Telematica</i> *	<i>Informatica</i>	<i>Matematica</i>
4	0	-	1	1
1	0.5	1	0.76	-
2	0.625	0.7	0.57	-
5	0.625	-	0.85	0.83
3	0.75	-	0	0.16
6	1	0	-	1

L'analisi dei dati

- **Ci sono tendenze nel comportamento degli studenti?**

<i>ID</i>	<i>Eta'*</i>	<i>Telematica*</i>	<i>Informatica</i>	<i>Matematica</i>
1	0.5	1	0.76	-
2	0.625	0.7	0.57	-
4	0	-	1	1
5	0.625	-	0.85	0.83
3	0.75	-	0	0.16
6	1	0	-	1

Obiettivi del corso

- **Introdurvi agli aspetti principali del processo di Knowledge Discovery**
 - **Teoria e applicazioni del Data Mining**
- **Fornire una sistematizzazione della miriade di concetti che sono presenti in quest'area, secondo le seguenti linee**
 - **Il processo di Knowledge Discovery**
 - **I metodi, applicati a casi paradigmatici**

Organizzazione del corso

- **Teoria di base del Knowledge Discovery**
 - **Modellazione predittiva**
 - **Modellazione descrittiva**
- **Una (non così) profonda introduzione ai tools di data mining**
 - **Weka (Machine Learning in Java)**
- **Casi di studio**
 - **In laboratorio**
- **Valutazione**
 - **Mid-term**
 - **Due compiti (metà ottobre, fine novembre)**
 - **Progetto: esperienza su un caso reale**
 - **Implementare un algoritmo e utilizzarlo per analizzare un insieme di dati**
 - **In gruppi di 2/3 persone**
 - **Attività seminariale**
 - **Presentazione (powerpoint) di una tecnica studiata in letteratura**
 - **Attività singola**

Materiale didattico

- **Riferimento principale:**
 - <http://www.icar.cnr.it/manco/>
 - Questi lucidi e altro materiale di approfondimento
 - Dispense (in corso di preparazione)
- **Libri di riferimento**
 - Tan, Steinbach, Kumar, *Introduction to Data Mining*, Addison-Wesley, 2005
 - J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufman, 2000
 - I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools with Java Implementation*. Morgan Kaufman, 1999
 - T. Mitchell, *Machine Learning*, McGraw-Hill, 1997
 - D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*, MIT Press, 2001
 - R.J. Roiger, M. W. Geatz, *Introduzione al Data Mining*, McGraw-Hill, 2004
- **Riferimenti bibliografici (articoli di survey e/o su argomenti specializzati):**
 - Distribuiti a lezione

Corpo docente

- **Giuseppe Manco**
 - manco@icar.cnr.it
 - **0984/831728**
 - **Ricevimento: Martedì, 12:30-13:30**
- **Antonio Locane**
 - locane@exeura.it
 - **0984/493026**

Outline

- **Motivazioni**
- **Aree applicative**
- **Il Processo di Knowledge Discovery**
- **Una (breve) rassegna dei passi di knowledge discovery**

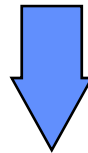
L'evoluzione della tecnologia dei databases: dalla gestione all'analisi dei dati

- **1960s:**
 - Collezioni di dati, creazione dei databases, IMS and network DBMS.
- **1970s:**
 - Modello relazionale,, implementazione dei DBMS relazionali.
- **1980s:**
 - RDBMS, modelli dei dati avanzati (relazionali estesi, OO, deduttivi, etc.) e orientati alle applicazioni (spaziali, scientifici,, etc.).
- **1990s:**
 - Data mining e data warehousing, multimedia databases, tecnologia Web.

Motivazioni

“Necessity is the Mother of Invention”

- *Il problema dell'esplosione dei dati:*
 - I meccanismi di collezione automatica dei dati insieme alla maturità della tecnologia database e ad internet, portano alla memorizzazione di una grossa quantità di dati.
- *We are drowning in information, but starving for knowledge!* (John Naisbett)



- *Data warehousing e data mining :*
 - On-line analytical processing
 - Estrazione di conoscenza interessante da grandi collezioni di dati.

Un po' di numeri...

- **1 Bit = Binary Digit**
- **8 Bits = 1 Byte**
- **1000 Bytes = 1 Kilobyte**
- **1000 Kilobytes = 1 Megabyte**
- **1000 Megabytes = 1 Gigabyte**
- **1000 Gigabytes = 1 Terabyte**
- **1000 Terabytes = 1 Petabyte**
- **1000 Petabytes = 1 Exabyte**
- **1000 Exabytes = 1 Zettabyte**
- **1000 Zettabyte = 1 Yottabyte**
- **1000 Yottabyte = 1 Brontobyte**

Esempi di grandi collezioni

- **Il Very Long Baseline Interferometry (VLBI) europeo ha 16 telescopi, ognuno dei quali produce 1 Gigabit/second di dati astronomici su una finestra di osservazione di 25 giorni**
 - **Come gestire la memorizzazione e l'analisi?**
- **AT&T gestisce miliardi di chiamate al giorno**
 - **Una tale quantità di dati non può essere memorizzata – l'analisi deve essere effettuata “on the fly”, sui flussi di dati che si producono**

I più grandi databases del 2003

- **Databases commerciali:**
 - Winter Corp. 2003 Survey: France Telecom ha il più grande DB per il supporto alle decisioni, ~30TB;
AT&T ~ 26 TB
- **Web**
 - Alexa internet archive: 7 anni di dati, 500 TB
 - Google searches 4+ miliardi di pagine, centinaia di TB
 - IBM WebFountain, 160 TB (2003)
 - Internet Archive (www.archive.org), ~ 300 TB

5 milioni di terabytes creati nel 2002

- **Una stima di UC Berkeley del 2003 : 5 exabytes (5 million terabytes) di nuovi dati creati nel 2002.**

www.sims.berkeley.edu/research/projects/how-much-info-2003/

- **Gli USA producono ~40% dei nuovi dati in tutto il mondo**

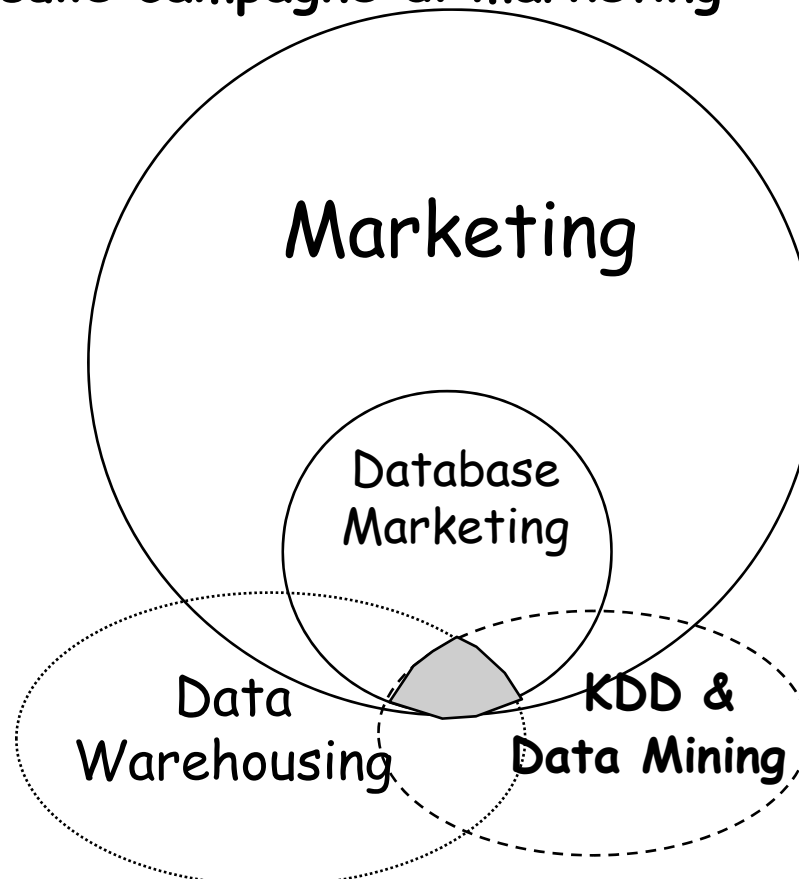
Il tasso di crescita dei dati

- **Nel 2002 è stato generato il doppio dei dati del 1999 (tasso di crescita ~30%)**
- **Quanti di questi dati potranno essere guardati da un umano?**
- **Analisi dei dati automatizzata è un requisito ESSENZIALE per capire il senso e l'utilità dei dati.**

A cosa serve il Data Mining?

Ampliare le conoscenze su cui basare le proprie decisioni.

Esempio: impatto sulle campagne di marketing



Aree di applicazioni

- **Scienza**
 - astronomia, bioinformatica, medicina, ...
- **Commercio**
 - Campagne pubblicitarie, CRM (Customer Relationship management), investimenti, manufacturing, sports/intrattenimento, telecomunicazioni, e-Commerce, marketing mirato, tutela della salute, ...
- **Web:**
 - search engines, bots, ...
- **Governo**
 - Applicazione della legge, profilazione di evasori, anti-terrorismo

Data Mining per la modellazione dei clienti

- **obiettivi:**
 - **Predizione della perdita del cliente (attrition)**
 - **Marketing mirato:**
 - **Vendite incrociate (cross-sell), acquisizione dei clienti**
 - **Rischio di credito**
 - **Rilevazione delle frodi**
- **Industrie interessate**
 - **Banche, telecomunicazione, grande distribuzione,**
...

Un caso di studio su Customer Attrition

- **Situazione: il tasso di attrition per gli acquirenti di cellulari è all'incirca del 25-30% per anno!**

Obiettivo:

- **Sulla base dell'informazione sui clienti collezionata negli N mesi precedenti, predire chi probabilmente verrà perduto il prossimo mese.**
- **Dare anche una stima del valore del cliente, e quale può essere un'offerta di ritenzione che sia vantaggiosa**

Risultati

- **Verizon Wireless ha costruito un data warehouse di clienti**
- **Ha identificato i potenziali “abbandonatori”**
- **Ha sviluppato modelli regionali**
- **Ha individuato i clienti con alta propensione ad accettare un’offerta**
- **Ha ridotto il tasso di attrition da oltre il 2%/mese a to meno del 1.5%/mese**
 - **(impatto significativo, su >30 M clienti)**

Caso di studio: Stimare il rischio di credito

- **Situazione:** un individuo richiede un prestito
- **Task:** Cosa deve fare la banca?
- **Nota:** Le persone che godono di una buona situazione personale non necessitano il prestito, e le persone che hanno una pessima situazione verosimilmente non pagheranno. I migliori clienti della banca sono nel mezzo

Caso di studio - e-commerce di successo

- **Una persona acquista un libro su Amazon.com.**
- **Task: Raccomanda altri libri che questa persona verosimilmente acquisterà**
- **Amazon effettua il raggruppamento sulla base degli acquisti:**
 - **Chi ha acquistato “Advances in Knowledge Discovery and Data Mining”, ha anche acquistato “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”**
- **Il meccanismo di raccomandazione è particolarmente efficace**

Caso di studio: e-commerce fallimentare (KDD-Cup 2000)

- **Data:** dati di clickstream e acquisti da Gazelle.com, rivenditore on-line di attrezzature per jogging
- **Q:** caratterizzare i visitatori che spendono più di \$12 in media
- **Dataset** di 3,465 acquisti, 1,831 clienti
- **Vendite totali-- \$Y,000**
- **Obitorio:** Gazelle.com fuori dal mercato, agosto 2000

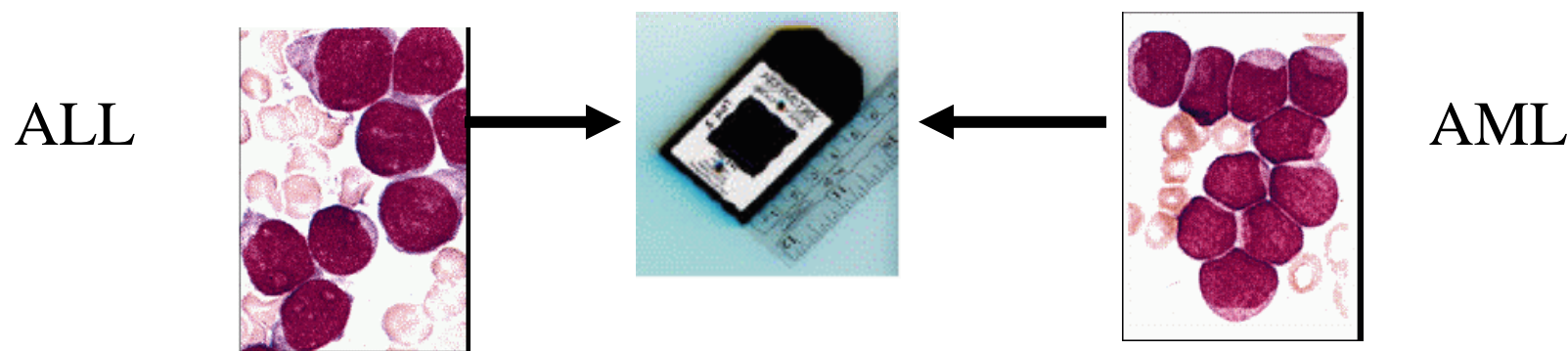
Caso di studio: Genomic Microarrays

Avendo a disposizione un microarray di dati per un certo numero di pazienti, possiamo

- **Diagnosticare accuratamente la malattia?**
- **Predirre il risultato di un trattamento?**
- **Raccomandare il miglior trattamento?**

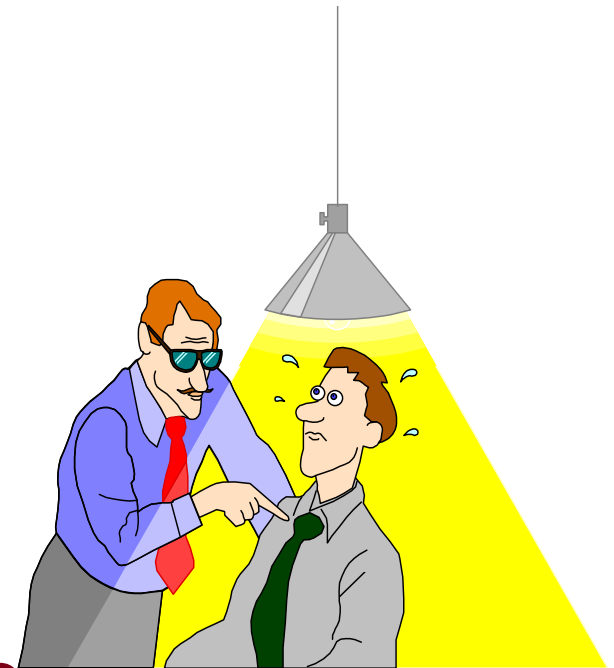
Example: ALL/AML data

- 38 casi (+ 34 per testare), ~ 7,000 geni
- 2 Classi: Leucemia acuta linfoplastica (ALL) ,
Leucemia acuta mieloide (AML)
- Costruzione di un modello diagnostico



Caso di studio: Sicurezza e rilevazione di frodi

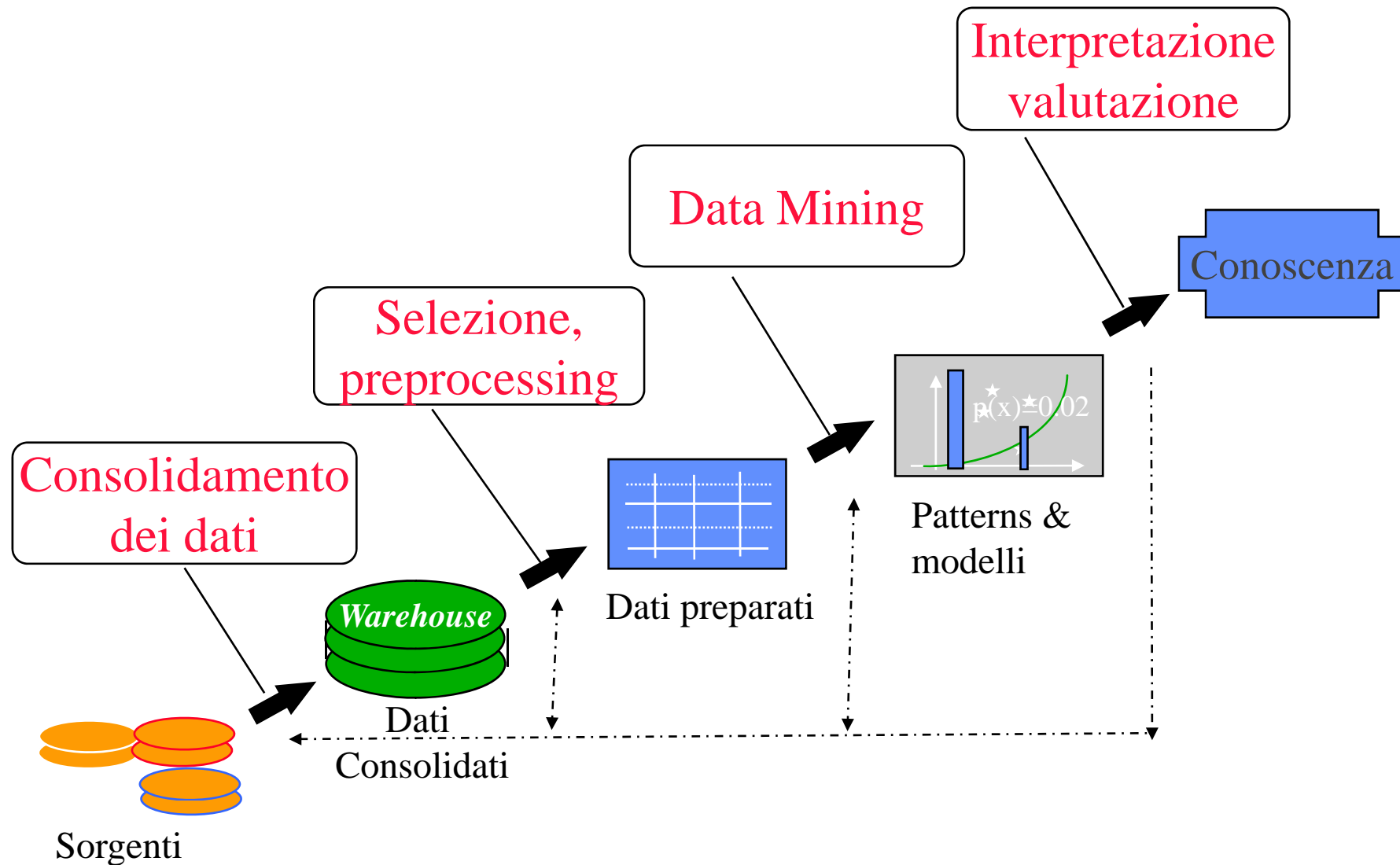
- **Clonazione di carte di credito**
- **Identificazione di operazioni di lavaggio di denaro sporco**
- **Frode al sistema di sicurezza**
 - **NASDAQ KDD system**
- **Frodi telefoniche**
 - **AT&T, Bell Atlantic, British Telecom/MCI**
- **Identificazione del Bio-terrorismo alle Olimpiadi di Salt Lake City, 2002**



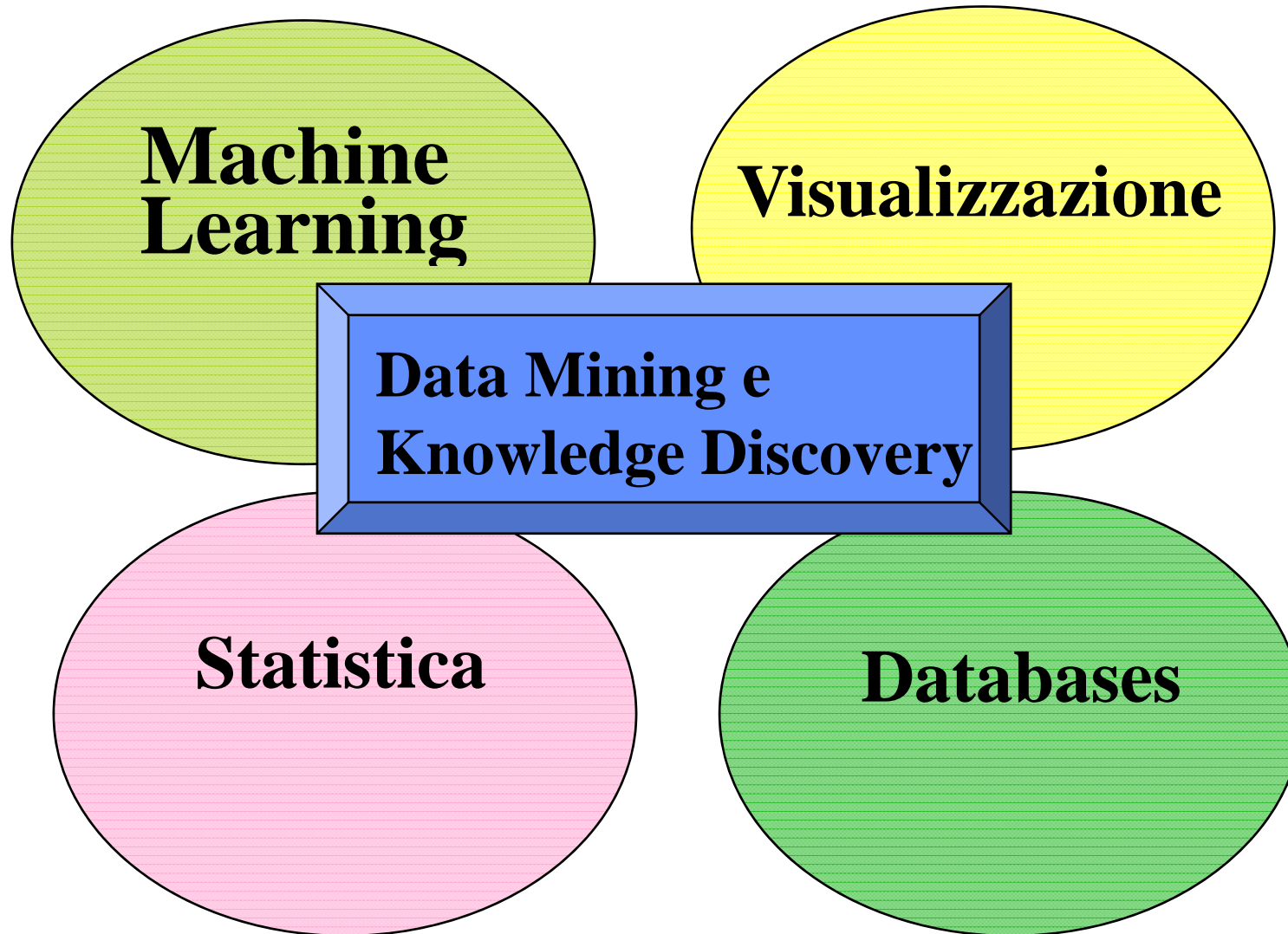
Cos'è il Knowledge Discovery? Un processo

- **La selezione e il processamento dei dati per:**
 - **L'identificazione di pattern nuovi, accurati e utili**
 - **La modellazione di fenomeni reali.**
- **Data mining** è una componente significativa del processo di KDD - la scoperta automatica di patterns è lo sviluppo di modelli predittivi e descrittivi.

Il processo di KDD



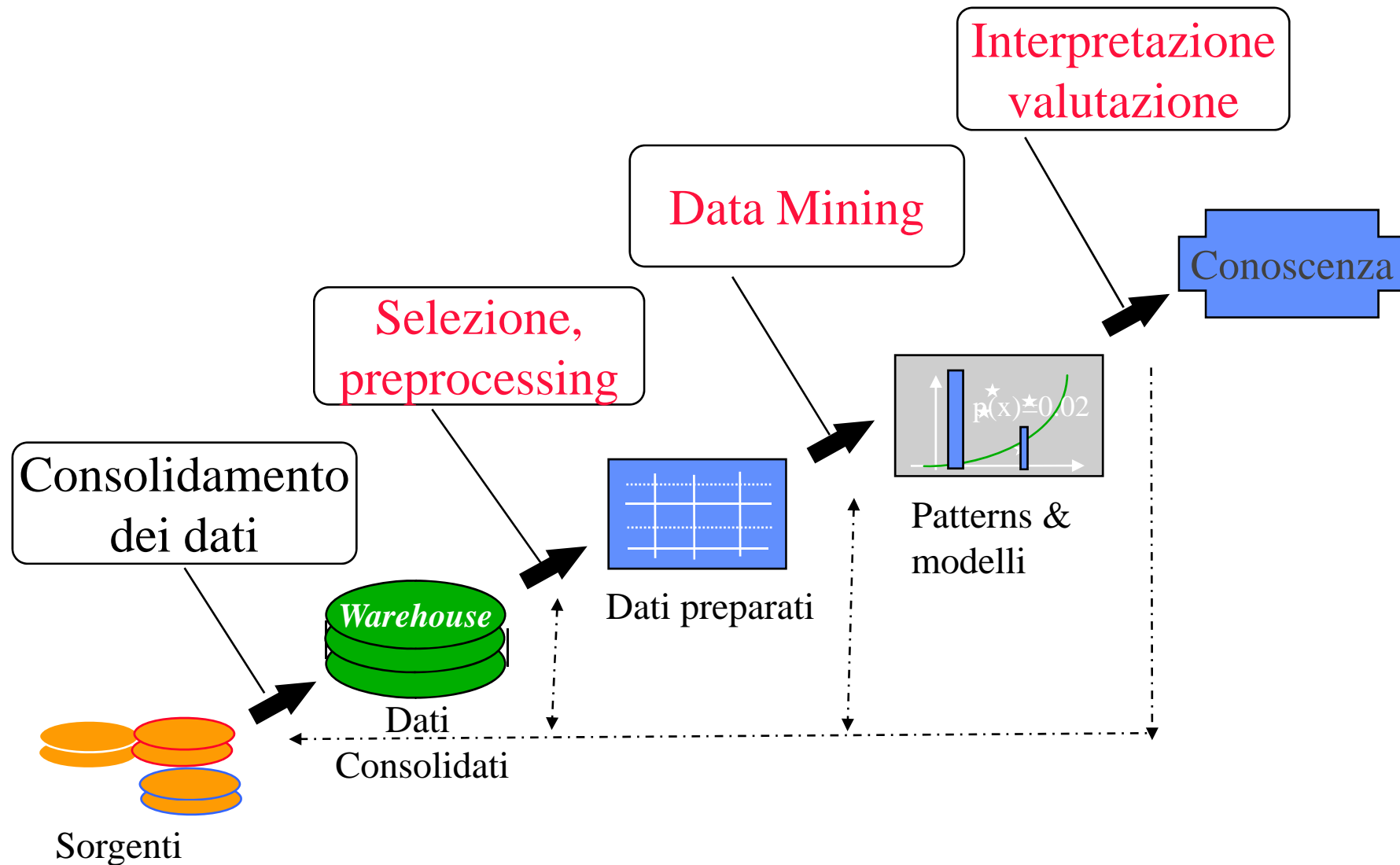
Campi correlati



Statistica, Machine Learning e Data Mining

- **Statistica:**
 - Basata sulla teoria
 - Focalizzata al test di ipotesi
- **Machine learning/apprendimento automatico**
 - Euristico
 - Mirato al miglioramento delle performance di apprendimento
 - Spazia anche nella robotica– non rilevante per il data mining
- **Data Mining e scoperta di conoscenza**
 - Integra teoria ed euristiche
 - Si concentra sull'intero processo: pulizia, apprendimento, integrazione e visualizzazione dei risultati
- **Le distinzioni non sono nette**
- **Le tecniche tradizionali non sono applicabili direttamente**
 - Dimensione, dimensionalità
 - eterogeneità

Il processo di KDD



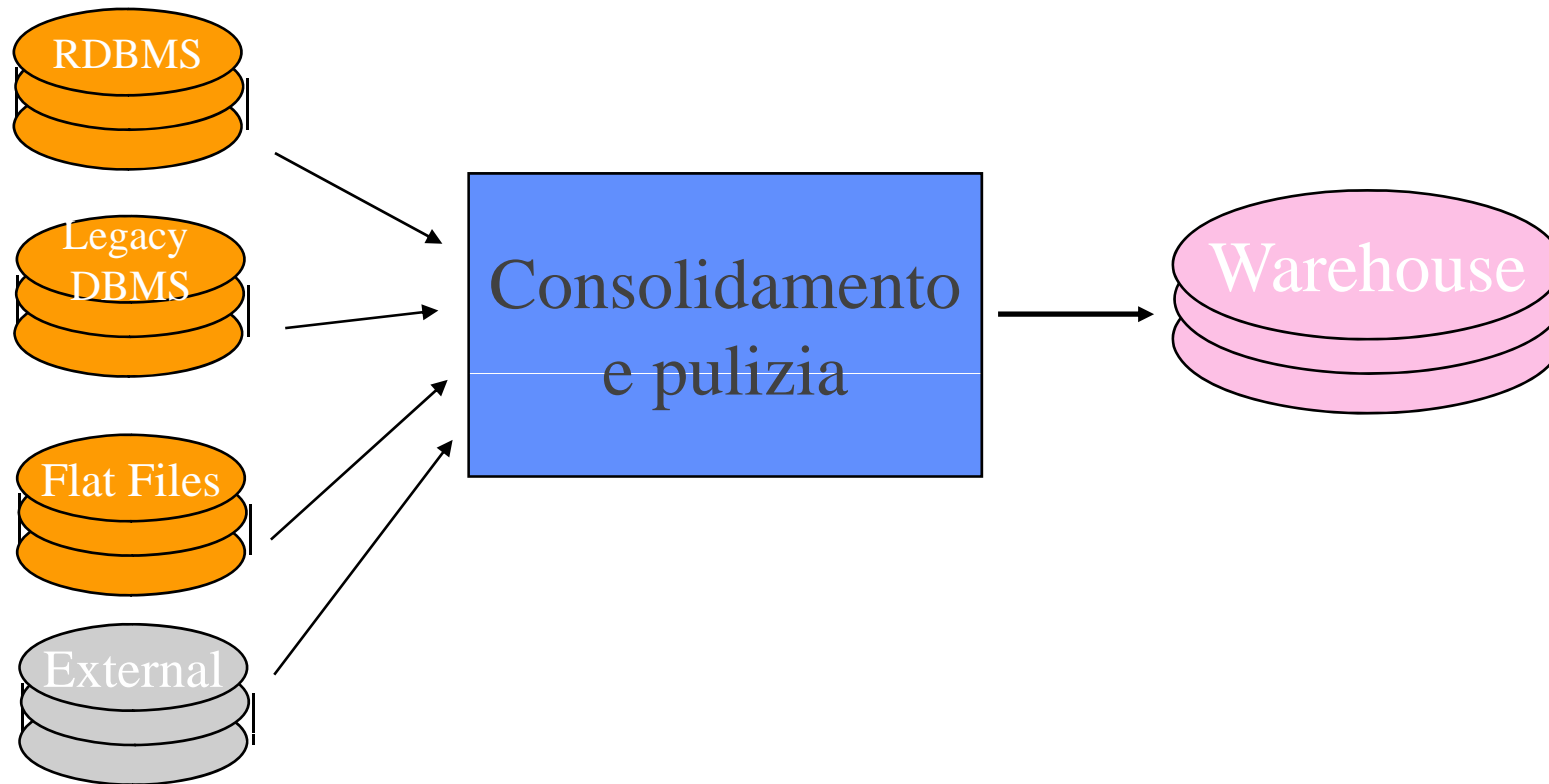
Consolidamento e preparazione

Garbage in → **Garbage out**

- **La qualità dei risultati è correlata alla qualità dei dati**
- **Il 50%-70% dello sforzo riguarda il consolidamento e la preparazione**

Consolidamento

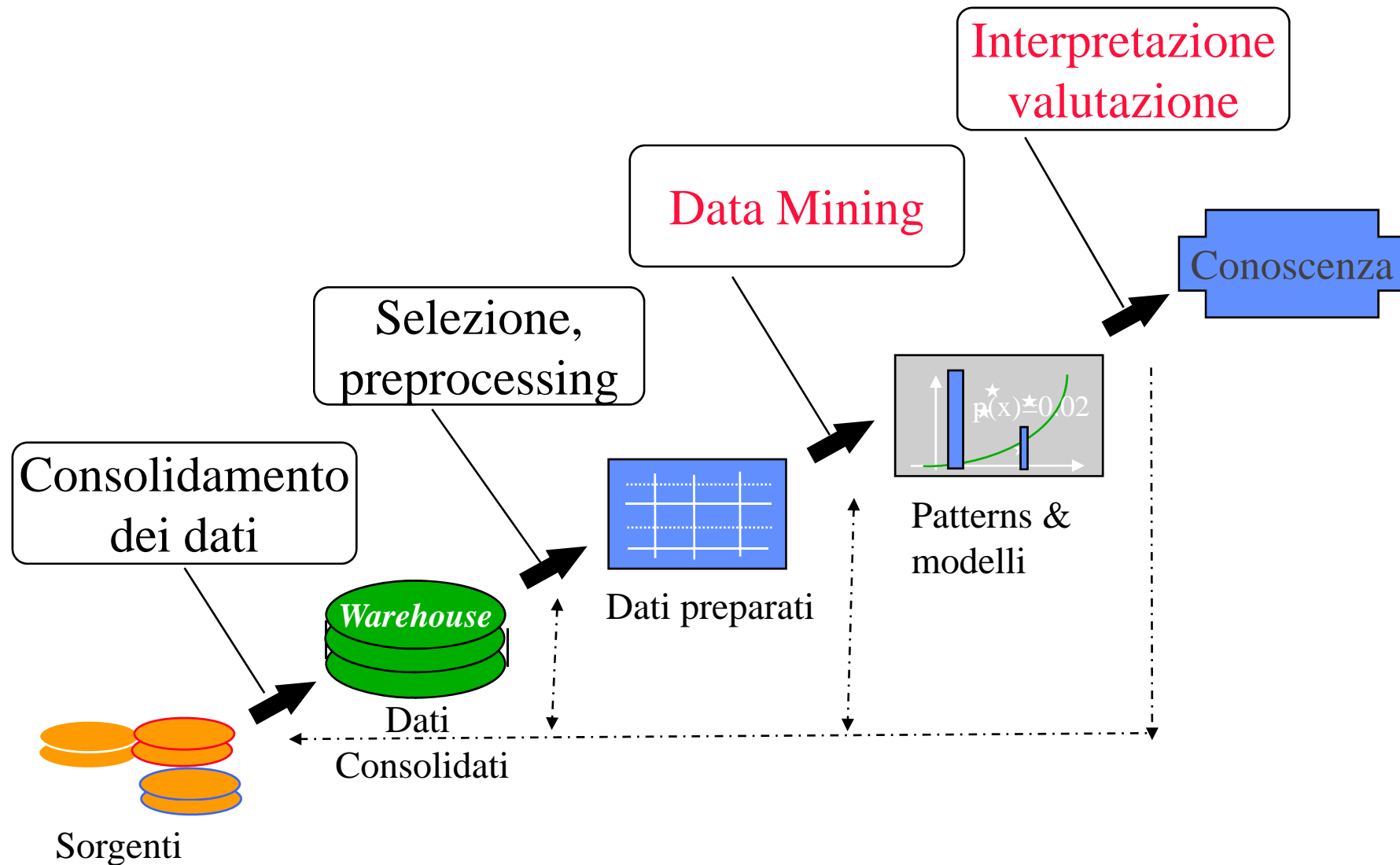
Da sorgenti eterogenee a una repository consolidata



Consolidamento

- **Determinare una lista preliminare di attributi**
- **Consolidare i dati in una tabella**
- **Eliminare o stimare i valori mancanti**
- **Rimozione di *outliers***

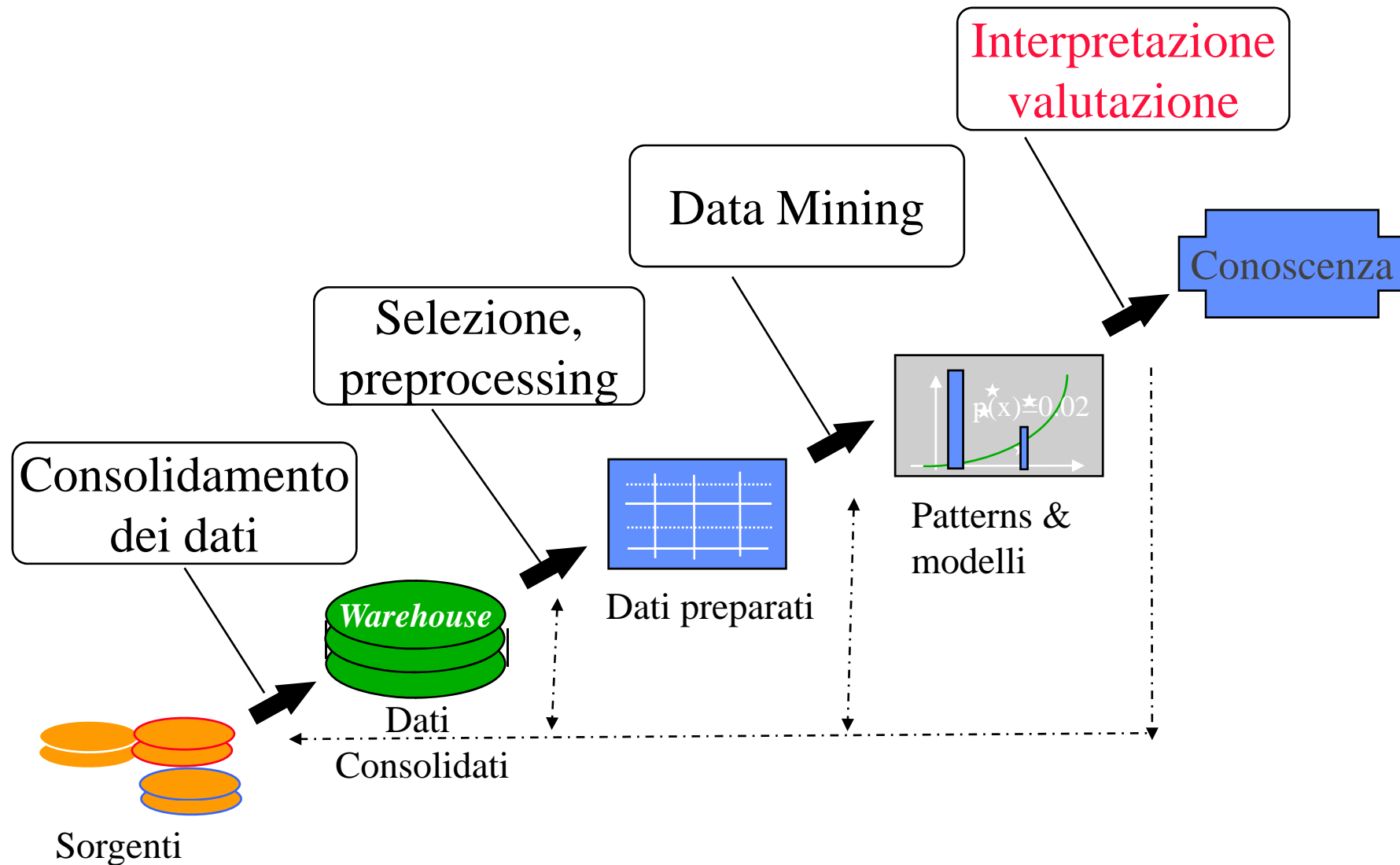
Il processo di KDD



Selection, preprocessing

- **Generazione di un campione**
 - Scelta del metodo di campionamento
 - Analisi della complessità del campione
 - Trattamento dell'influenza del campionamento
- **Riduzione della dimensionalità degli attributi**
 - Rimozione di attributi ridondanti e/o correlati
 - Combinazione di attributi
- **Riduzione dei range**
 - Raggruppamento di valori discreti
 - Discretizzazione di valori numerici
- **Transformazione dei dati**
 - de-correlare e normalizzare i valori

Il processo di KDD

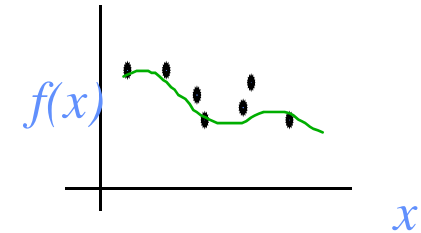


Task e metodi di Data mining

- **Predizione(classificazione)**

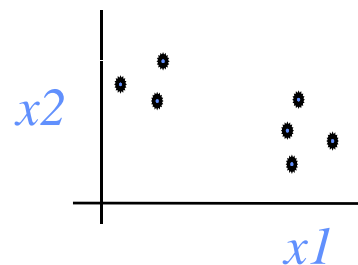
- regressione, reti neurali, algoritmi genetici, alberi di decisione

if age > 35
and income < \$35k
then ...

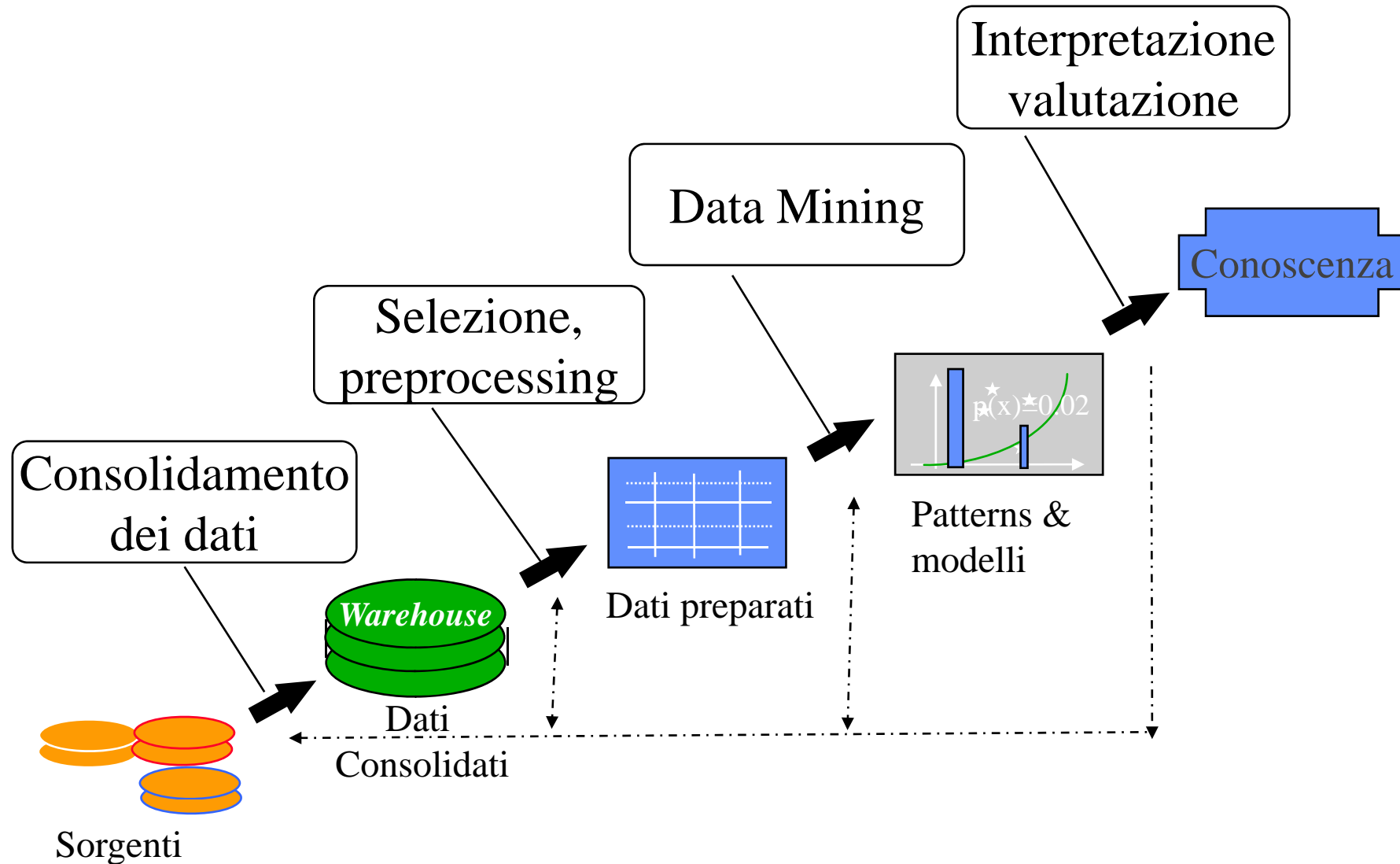


- **Descrizione**

- decision trees, regole associative
- clustering analysis



Il processo di KDD

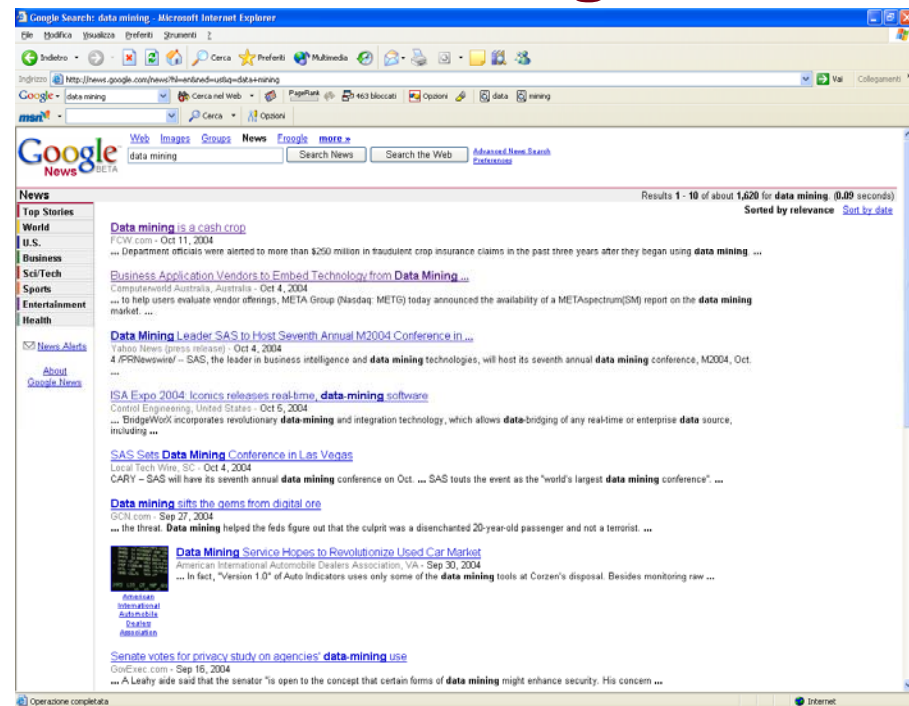


Quanto sono interessanti I patterns?

- **Misure di interesse:**
 - **Comprensibilità**
 - **Validità su dati nuovi (in accordo ad un certo grado di certezza).**
 - **Utilità**
 - **Novità, validazione di ipotesi**
- **Misure oggettive/soggettive**
 - **Oggettive: basate sulla statistica e sulla struttura dei patterns**
 - **Soggettive: basate sulla conoscenza dei dati: nuove, inaspettate, ...**

Esercizio: Data Mining nelle News

- Usa la search engine di Google (news.google.com) per identificare storie recenti che riguardano l'applicazione di tecniche di data mining
- Esempio:



- Riportare una breve descrizione delle storie