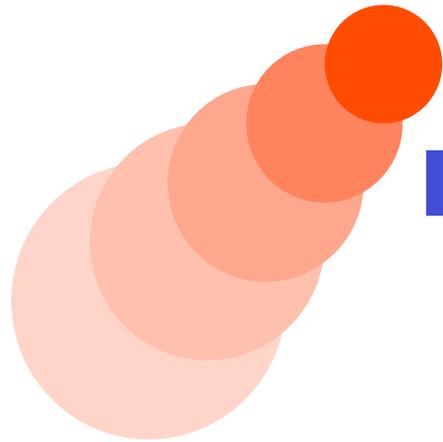


Data Warehousing, Data Mining & Business Intelligence



Introduzione

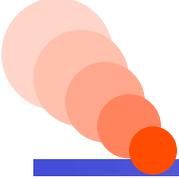
Paolo G. Franciosa

Dipartimento di Statistica, Probabilità e Statistiche Applicate

Università "La Sapienza"

paolo.franciosa@uniroma1.it

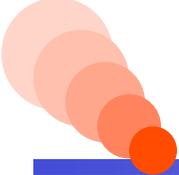
Questo materiale deriva dalla traduzione e adattamento delle presentazioni
pubblicate dal prof. Jiawei Han



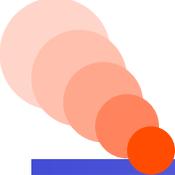
Motivazioni

- Esplosione della quantità di dati disponibili
 - Evoluzione della tecnologia dei DBMS
 - Trasmissione di dati, protocolli di comunicazione e accesso a DB
 - Diffusione di dispositivi di data collection (badge, bar code, cards, telecomunicazioni, reti, ...)
- Necessità di archiviazione per scopi di analisi (Data Warehousing)
- Disponibilità di sistemi di potenza adeguata al trattamento di grandi quantità di dati

Motivazioni (cnt.)



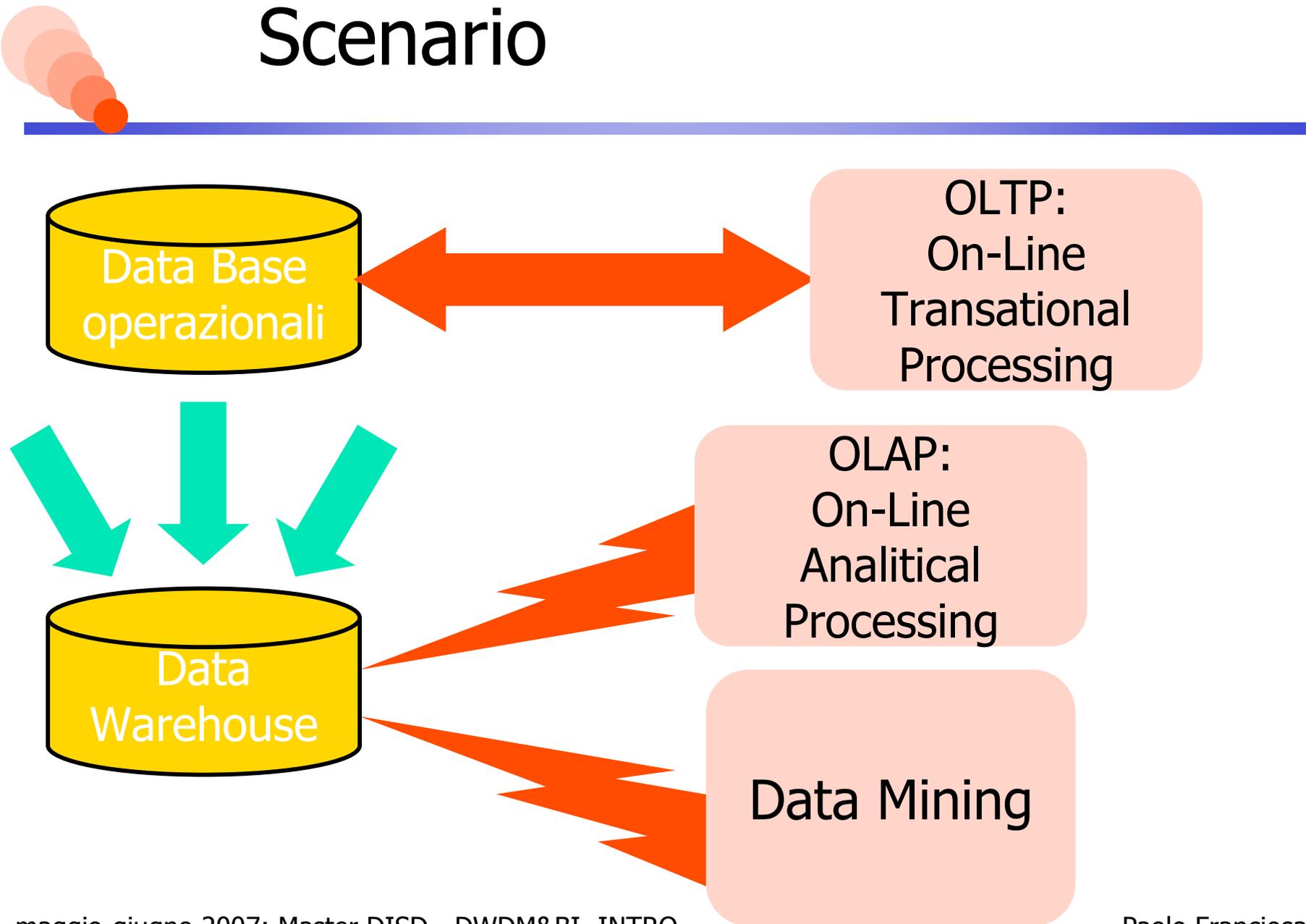


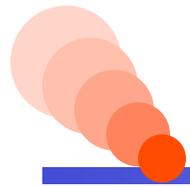


Una prospettiva storica

- 1960:
 - Raccolta di dati, data bases, modelli gerarchici e reticolari
- 1970:
 - Modello relazionale dei dati [Codd], primi DBMS relazionali
- 1980:
 - Diffusione di DBMS relazionali commerciali, estensioni del modello relazionale (Object Oriented DB, Geographical Information Systems, DB deduttivi, etc.)
- 1990—2000:
 - Data mining e data warehousing, database multimediali, interesse verso informazioni semistrutturate (World Wide Web)

Scenario





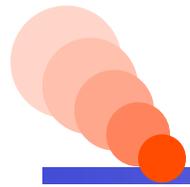
OLTP vs. OLAP e DM

➤ Caratteristiche dei dati:

- schema articolato
- situazione corrente
- schema semplice (star ...)
- dati storici

➤ Dati coinvolti nelle operazioni:

- volume ridotto
- aggiornati
- volume elevato
- non necessariamente aggiornati



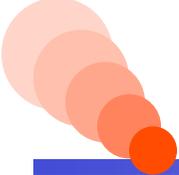
OLTP vs. OLAP e DM

➤ Caratteristiche delle operazioni:

- snelle
- predeterminate
- ripetitive, quotidiane
- aggiornamento
- onerose
- indagini ad-hoc
- occasionali
- lettura

➤ Requisiti delle operazioni:

- tempo reale
- robustezza
- concorrenza
- iterative ma non in real time
- robustezza non critica
- utenti isolati, read only



OLAP vs. Data Mining

OLAP

Analisi del contenuto
del DB/DW
attraverso lo studio
di aggregazioni
guidate dall'utente
Reportistica

DM

“Estrazione di
conoscenza
(non banale, implicita,
nuova,
potenzialmente utile)
da grandi quantità di
informazioni” [Han]

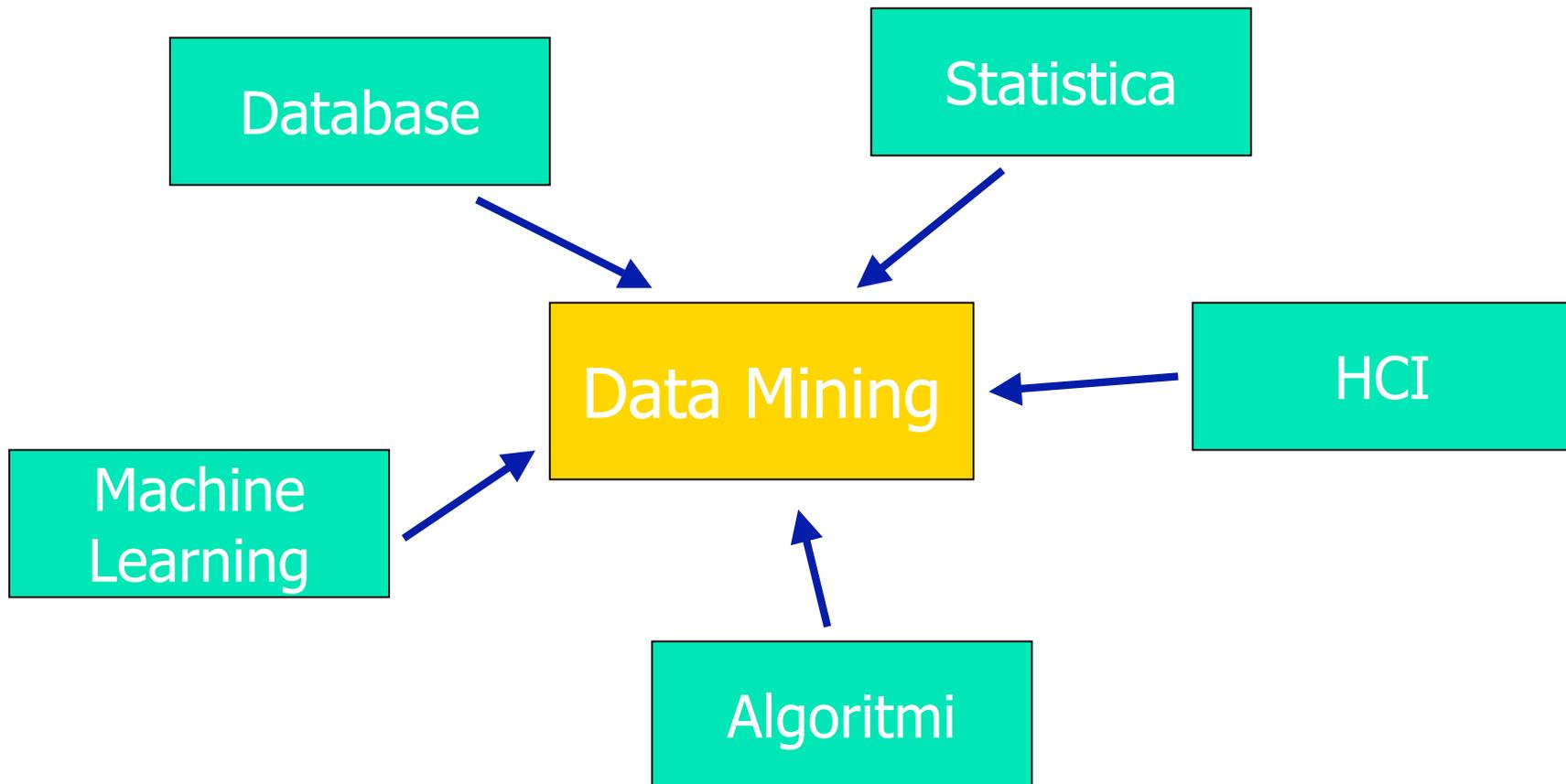
Data Mining: origini

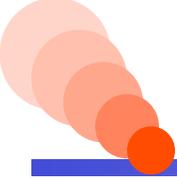


- Data mining (parte del processo di KDD-Knowledge Discovery in Databases):
 - Estrazione di informazioni o esempi utili (non banali, implicite, sconosciute e potenzialmente utili) da DB di grandi dimensioni
- Nomi alternativi
 - knowledge discovery (mining) in databases (KDD)
 - knowledge extraction
 - data/pattern analysis
 - data archeology
 - data dredging
 - information harvesting
 - business intelligence
 - ... machine learning?



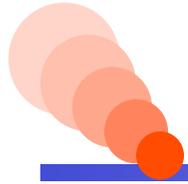
Data Mining: confluenza di molte discipline





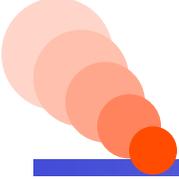
Applicazioni potenziali

- Analisi di DB e Sistemi di Supporto alle Decisioni
 - Market analysis and management
 - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - forecasting, customer retention, quality control, competitive analysis
 - **Fraud detection**
- Altre applicazioni
 - text mining (news groups, email, documents) and Web analysis.
 - intelligent query answering
 - stream analysis, log data analysis



Market Analysis and Management (1)

- Sorgenti informative
 - registratori di cassa, transazioni con carte di pagamento, buoni sconto, adesione a campagne promozionali, ogni operazione che permetta di “tracciare” il cliente
- Target marketing
 - individuare classi di clienti con caratteristiche simili: interessi, potenziale di spesa, abitudini di acquisto, in particolare per **intraprendere iniziative mirate**
- Scoprire modifiche nel comportamento del cliente
 - prevedere/spiegare abbandoni, adeguamento dell’offerta
- Cross-market analysis
 - scoprire associazioni tra interessi, sfruttare sinergie



Fraud Detection (1)

➤ Applicazioni

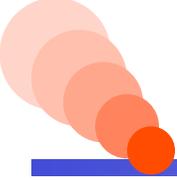
- sanità, vendite, uso carte di credito, telefonia, etc.

➤ Approccio

- costruire un modello dei comportamenti fraudolenti usando dati storici
- identificare similarità con comportamenti fraudolenti usando DM

➤ Esempi

- assicurazioni auto: individuare gruppi di truffatori
- riciclaggio di denaro: individuare movimenti di denaro sospetti (US Treasury's Financial Crimes Enforcement Network)
- assicurazioni sanitarie: scoprire pazienti "di professione" , medici conniventi

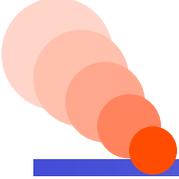


Fraud Detection (2)

- Frodi verso compagnie telefoniche e clienti
 - modello dell'utente: destinazione delle chiamate, durata, ora e giorno della settimana
 - vengono evidenziati comportamenti anomali dell'utente

- Retail
 - si stima che il 38% delle flessioni delle vendite sia dovuto a impiegati disonesti

Altre applicazioni



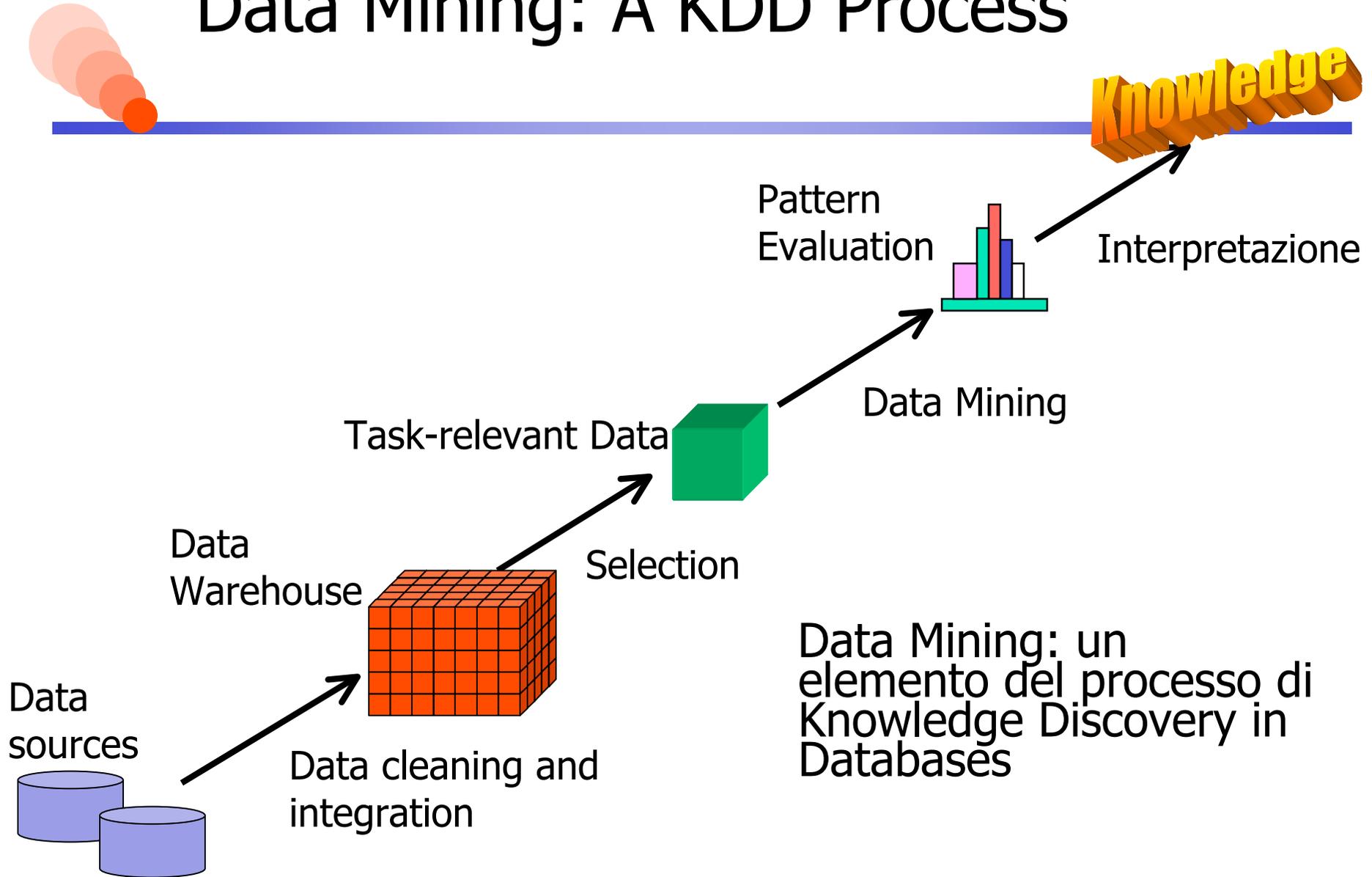
➤ **Astronomia**

- JPL (Jet Propulsion Laboratory) e l'osservatorio di monte Palomar hanno scoperto 22 quasar usando tecniche di data mining

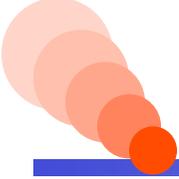
➤ **Profilo dei visitatori di siti web**

- dallo studio dei log degli accessi, in particolare verso siti commerciali, è possibile dedurre preferenze, gradimento di pagine, e trarre informazioni utili sull'efficacia della organizzazione del sito web

Data Mining: A KDD Process



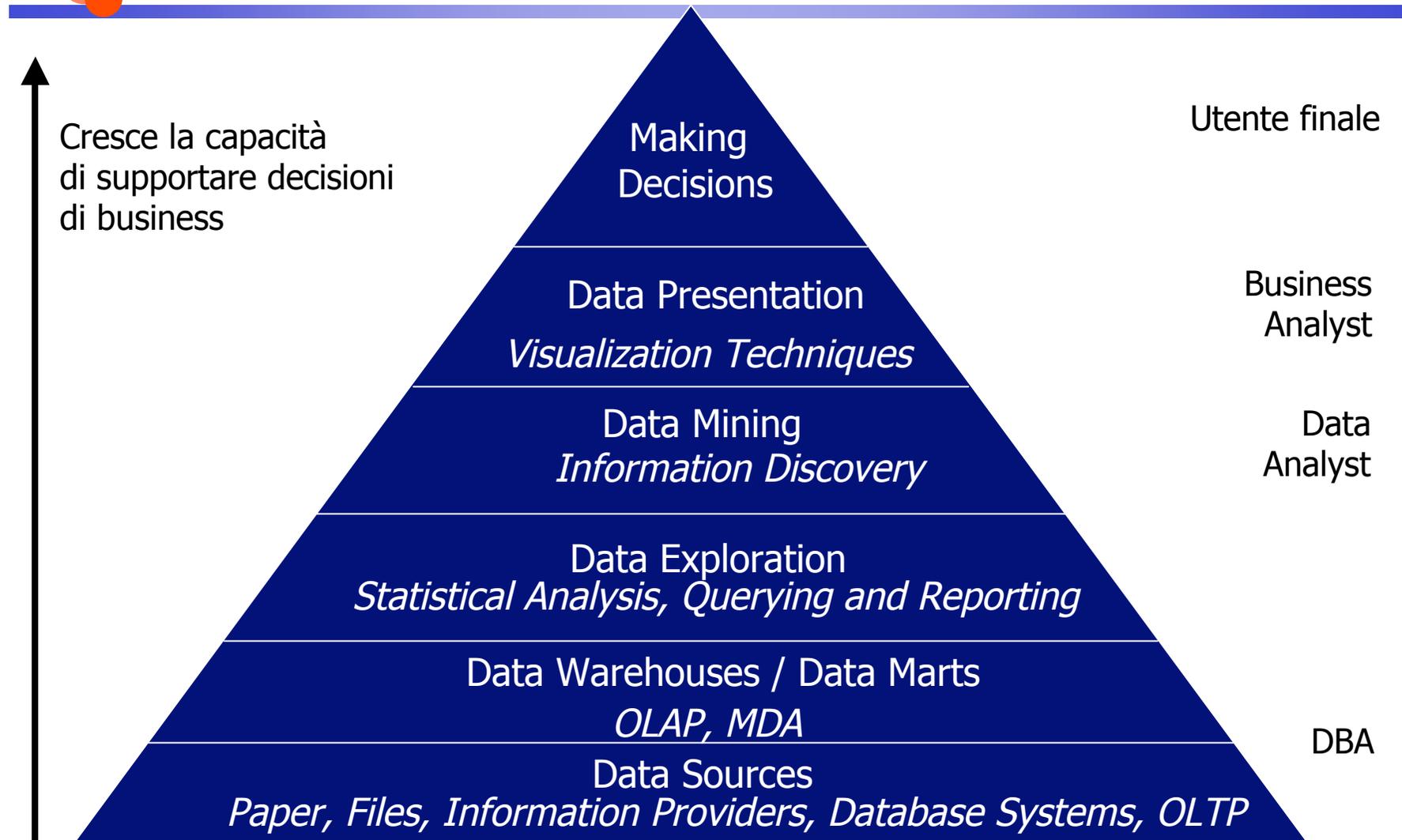
Data Mining: un elemento del processo di Knowledge Discovery in Databases

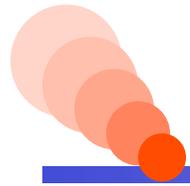


Fasi del processo KDD

- Conoscenza del contesto applicativo e dello scopo dell'analisi
- Scelta dei **dati**
- **Data cleaning** and **preprocessing**: (non sottovalutare)
- **Trasformazione dei dati**:
 - scala, unità di misura, codifiche, eliminazione di conoscenze implicite
- Scelta delle **funzioni di Data Mining**
 - aggregazioni (OLAP), classificazione, regressione, regole associative, clustering.
- Scelta degli **algoritmi/tools** di data mining
- **Estrazione dei pattern** di interesse
- **Interpretazione** e **Valutazione dell'utilità dei pattern**
 - visualizzazione, trasformazione, eliminazione dei pattern banali, già noti o inutili
- **Ricaduta** della conoscenza acquisita sul contesto applicativo

Data Mining e Business Intelligence

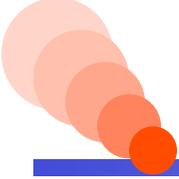




Data Mining: su quale tipo di dati?

- DB relazionali
- Data Warehouses
- DB Transazionali
- DB non convenzionali
 - Object-oriented e object-relational
 - DB spaziali (GIS)
 - Serie temporali
 - Testo e DB multimediali
 - DB specialistici
 - WWW

Funzionalità di Data Mining (1)

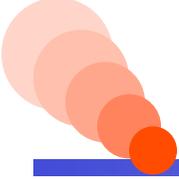


➤ Descrizione di concetti: caratterizzare e discriminare

- Generalizzare, aggregare e distinguere caratteristiche dei dati:
p.es. zone aride da zone umide

➤ Associazione (correlazione e causalità)

- associazioni multi-dimensionali vs. mono-dimensionali
- $\text{età}(X, "20..29") \wedge \text{reddito}(X, "20..29K") \rightarrow \text{acquista}(X, "PC")$
[supporto = 2%, confidenza = 60%]
- $\text{contiene}(x, "computer") \rightarrow \text{contiene}(x, "software")$ [1%, 75%]



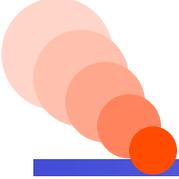
Funzionalità di Data Mining (2)

➤ Classificazione e Predizione

- Trovare modelli (funzioni) che descrivono e distinguono classi o concetti a fini predittivi
- p.es. classificare regioni in base al clima, o classificare autoveicoli in base alla frequenza dei rifornimenti
- Presentazione: alberi di decisione, regole di classificazione, reti neurali
- Predizione: prevedere valori mancanti o sconosciuti a partire da attributi noti

➤ Cluster analysis

- le classi sono sconosciute: vogliamo raggruppare i dati in modo da formare nuove classi, p. es. raggruppare edifici per trovare regole di distribuzione
- il principio di base: massimizzare la similarità intra-cluster e minimizzare la similarità inter-cluster



Funzionalità di Data Mining (3)

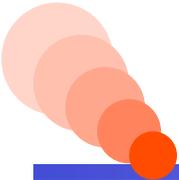
➤ Analisi degli outlier

- Outlier: un dato che non rispetta il comportamento generale
- Può essere considerato rumore, eccezione, ma è invece rilevante nella analisi di eventi rari (p. es. frodi)

➤ Analisi delle tendenze

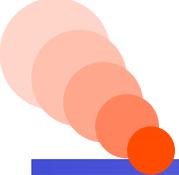
- Trend e deviation: analisi regressiva
- Mining di pattern sequenziali, periodicità

➤ Altre analisi statistiche



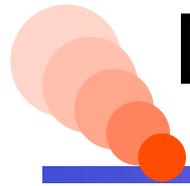
Rilevanza dei pattern

- Un algoritmo di data mining può generare migliaia di pattern, non tutti interessanti
 - necessario intervento umano
- Misure di interesse: un pattern è **interessante** se è:
 - facilmente comprensibile
 - **valido su dati nuovi o di test con un certo grado di certezza**
 - potenzialmente utile
 - nuovo, o che conferma ipotesi non certe



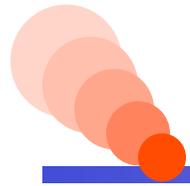
Data Mining: schemi di classificazione

- Funzionalità generali
 - Data mining descrittivo
 - Data mining predittivo
- Altre classificazioni in base a:
 - tipo di database da utilizzare
 - tipo di conoscenza da scoprire
 - tipo di tecnica da applicare
 - dominio applicativo



Data Mining: schemi di classificazione

- tipo di database da utilizzare
 - relazionale, transazionale, o-o, spaziale, serie temporali, testo, multimediale, www, ...
- tipo di conoscenza da scoprire
 - caratterizzazione, discriminazione, associazione, classificazione, clustering, tendenza, studio di outlier, ...
- tipo di tecnica da applicare
 - DB query oriented, OLAP, apprendimento automatico, statistica, visualizzazione, reti neurali, ...
- dominio applicativo
 - vendite, telecomunicazioni, banche, DNA mining, borsa, Web mining, Weblog analysis



OLAP Mining: integrazione di DM e DW

- Sistemi DM, DW e DBMS possono essere più o meno strettamente accoppiati
- OLAM (On-Line Analytical Mining)
 - integrazione di tecnologie OLAP e DM
- Uso di gerarchie nell'attività di mining
 - applicare tecniche di DM a diversi livelli di astrazione, sfruttando primitive di drilling/rolling, pivoting, slicing/dicing, etc.
- Integrazione di più funzioni di mining
 - p.es. cercare regole associative separatamente all'interno di ciascun cluster