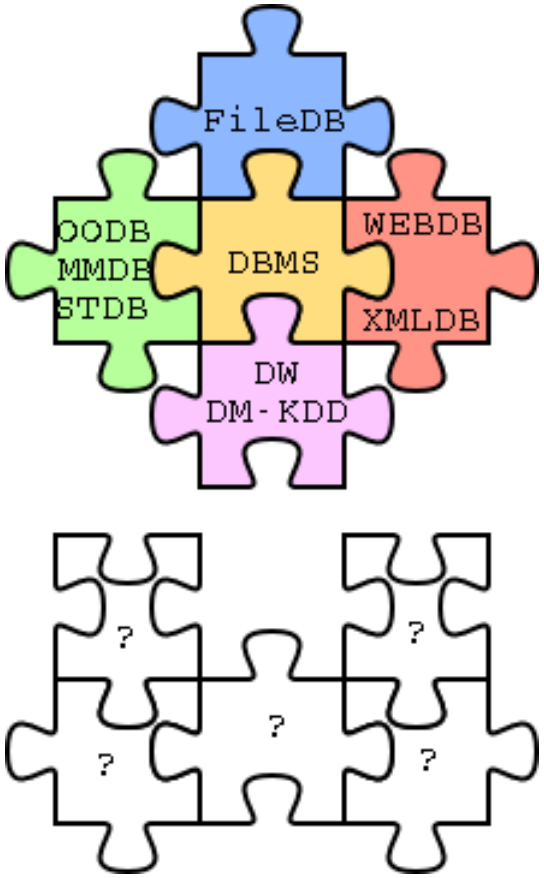
Introduzione al Data Mining

Sistemi informativi per le Decisioni

Slide a cura di Prof. Claudio Sartori

Evoluzione della tecnologia dell'informazione (IT) *(Han & Kamber, 2001)*

- Percorso evolutivo iniziato negli anni '60
- Ogni stadio è caratterizzato da un nuovo insieme di funzionalità



IT: stadi evolutivi *(Han & Kamber, 2001)*

- Anni ’60: Raccolta dati e creazione dei database
  - Elaborazione elementare di file
- ’70 - primi ’80: DBMS
  - Modelli dei dati e sistemi gerarchici, reticolari, relazionali
  - Modelli concettuali (ER)
  - Indici
  - Linguaggi di interrogazione non procedurali (SQL)
  - Ottimizzazione delle interrogazioni
  - Gestione delle transazioni e delle autorizzazioni

IT: stadi evolutivi *(Han & Kamber, 2001)*

- Metà ’80 - presente: Sistemi database avanzati
  - Modelli e sistemi Object-Oriented, Object-Relational, deduttivi
  - Sistemi orientati all’applicazione: Spazio-temporali, statistico-scientifici, multimediali, basati sulla conoscenza (KBMS)
- ’90 - presente: Sistemi database basati sul Web
  - Sistemi di database XML
  - Web mining
- Tardi ’80 - presente: Data Warehousing e Data Mining
  - Data Warehouse, On-Line Analytical Processing
  - Data Mining, Knowledge Discovery in Databases

“Tombe” dei dati?
La necessità è la madre delle invenzioni

- Esplosione dei dati
  - Strumenti di raccolta automatica dei dati, maturità della tecnologia database
  - Enormi quantità di dati memorizzati e disponibili
- La capacità di raccogliere e memorizzare dati ha largamente superato la capacità umana di analizzarli
- Archivi di dati ⇔ cimiteri di dati
- Anneghiamo nei dati ma siamo affamati di conoscenza
- Tuttavia, i dati contengono informazioni di grande interesse economico e scientifico: la ricerca in DM e KDD ha come scopo la progettazione di strumenti per trasformare i dati in informazione
- Data Warehousing e Data Mining
  - Integrazione e analisi/sintesi
  - Estrazione di conoscenza interessante e non nota a priori da grandi basi di dati

Una storia...

- Il dipartimento dell'agricoltura degli Stati Uniti ogni anno eroga indennizzi per danni da maltempo a centinaia di migliaia di agricoltori
- Una frazione delle richieste di indennizzo è fraudolenta
- Un'analisi a campione delle richieste per verificarne l'autenticità ha un costo molto elevato rispetto alla resa
- Un progetto di Data Mining volto a individuare le frodi ha reso oltre venti volte il suo costo

Definizione di *Knowledge Discovery*

"The nontrivial extraction of
implicit, previously unknown and
potentially useful
information from data"

W. Frawley, G. Piatetsky-Shapiro, and C. Matheus:
"Knowledge Discovery in Databases: An Overview".
AI Magazine, Fall 1992, pgs 213-228

Knowledge Discovery

- scoprire (e presentare) “conoscenza” in una forma facilmente comprensibile, e utilizzabile a scopi gestionali/decisionali
  - tecniche statistiche
  - di visualizzazione
  - di machine learning
- scalabilità
  - efficienza computazionale su DB di notevoli dimensioni (Giga/Tera-bytes)

Knowledge Discovery (ii)

- non solo algoritmi
  - processo complesso di manipolazione dei dati
  - data integration
    - formato eterogeneo (es.: rappresentano gli stessi dati con schemi differenti)
    - riconciliazione delle varie fonti
  - data cleaning
    - dati affetti da “rumore” (errori, dati non interessanti, ecc.)
    - pre-processing per “pulire” i dati

Interdisciplinarietà

- Database e Data Warehousing
- Statistica
- Apprendimento automatico
- Acquisizione della conoscenza
- Metodi di visualizzazione

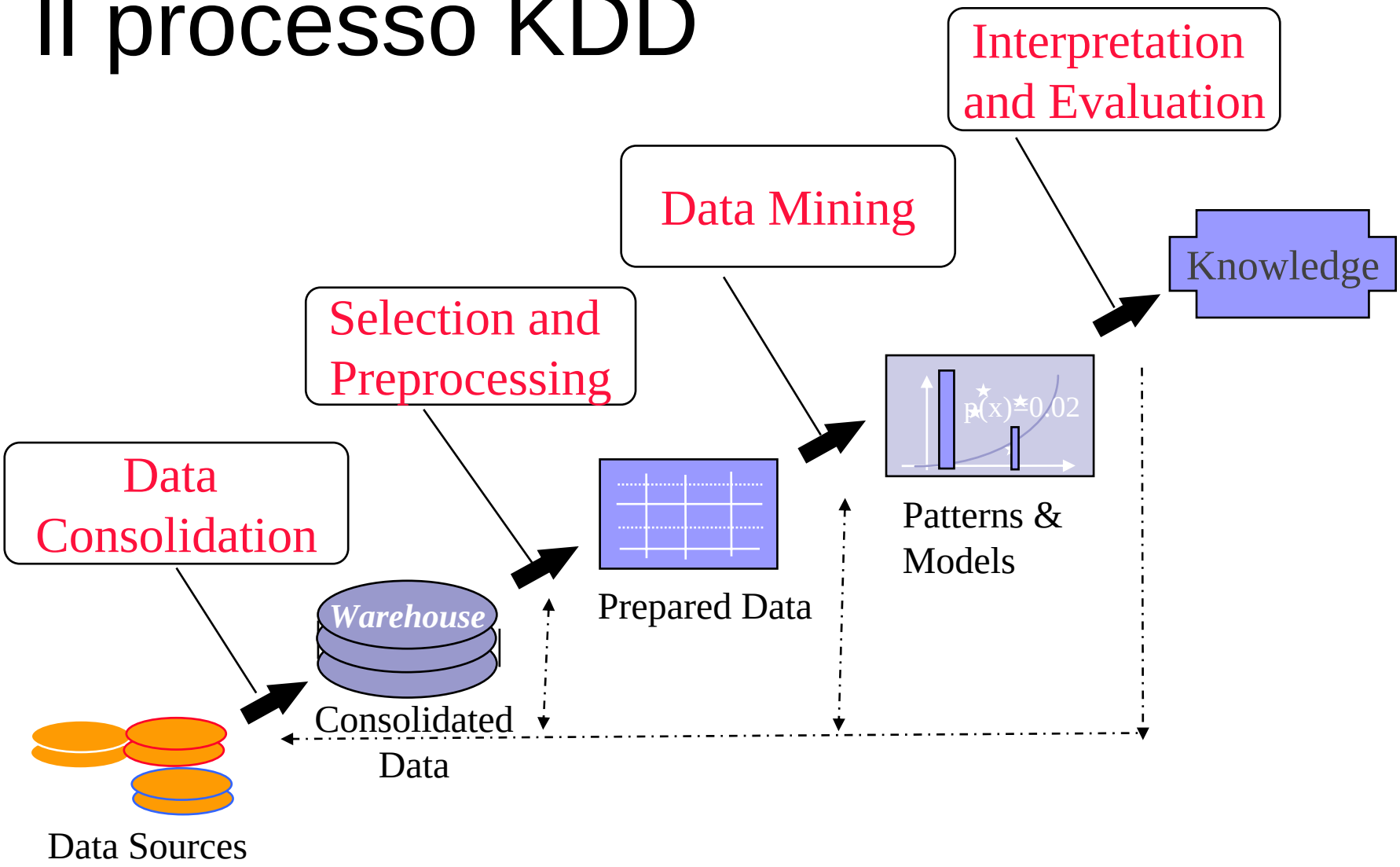
Fattibilità

- Abbondanza di dati
- Disponibilità di potenza di calcolo
- Forte fondamento matematico
  - Apprendimento automatico e logica inferenziale
  - Statistica e sistemi dinamici
  - DBMS

Knowledge Discovery in Databases

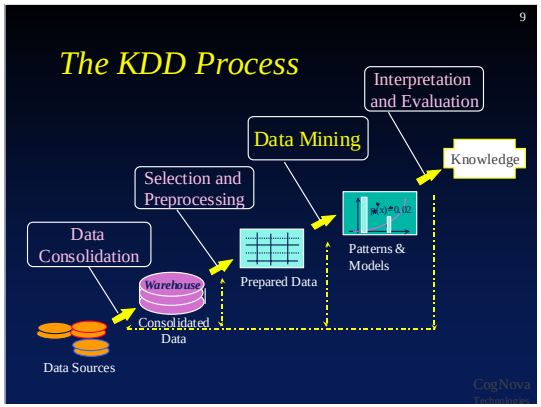
- Un processo
  - Selezione ed elaborazione di dati per
    - Identificare schemi (pattern) nuovi, accurati e utili
    - Modellare fenomeni del mondo reale
  - Data Mining è il componente principale del processo
  - Sviluppo di modelli predittivi ed esplorativi

Il processo KDD



Il ciclo virtuoso

Problema



Conoscenza

Identificare problema o opportunità

Agire in base alla conoscenza

Strategia

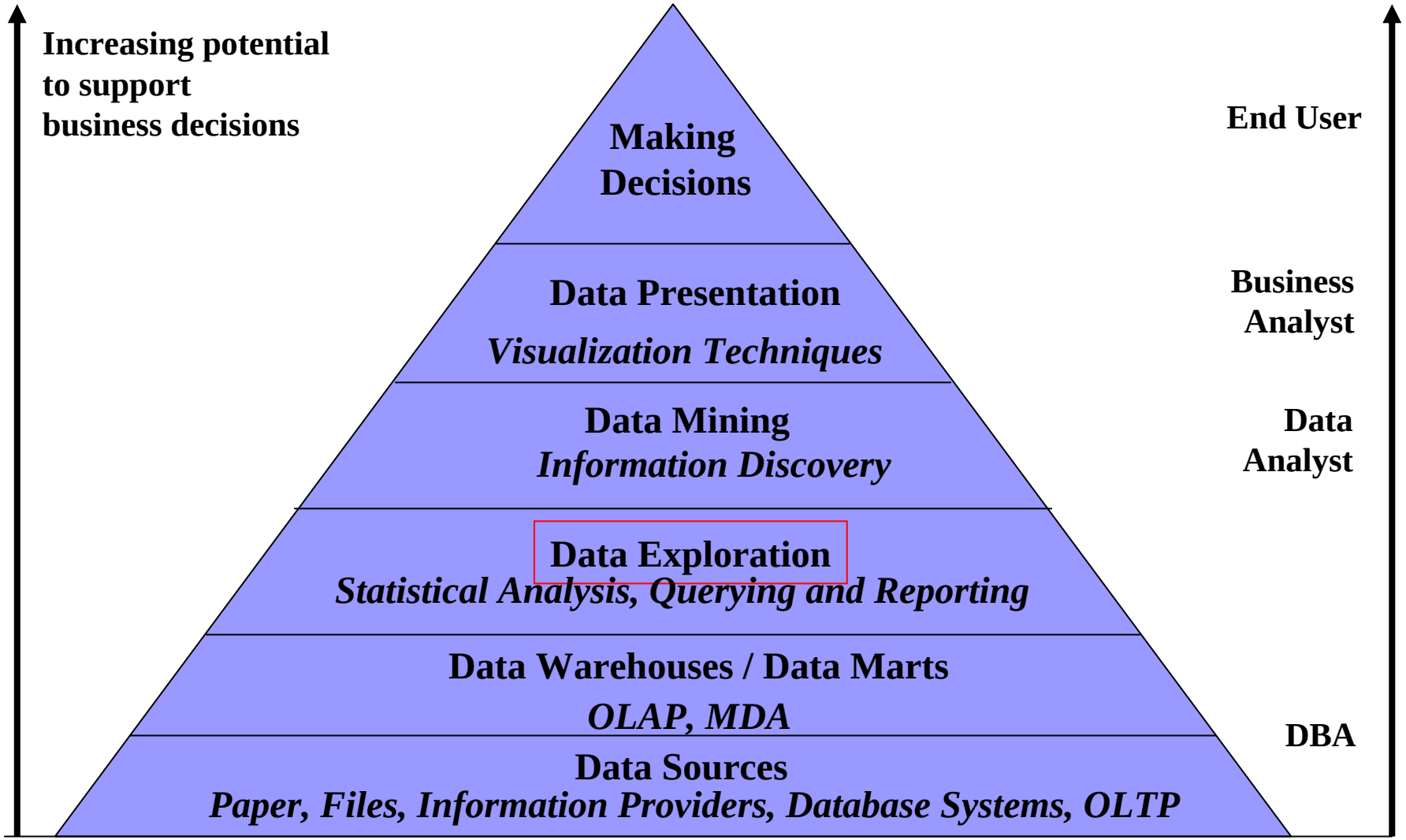
Misurare l'effetto dell'azione

Risultati

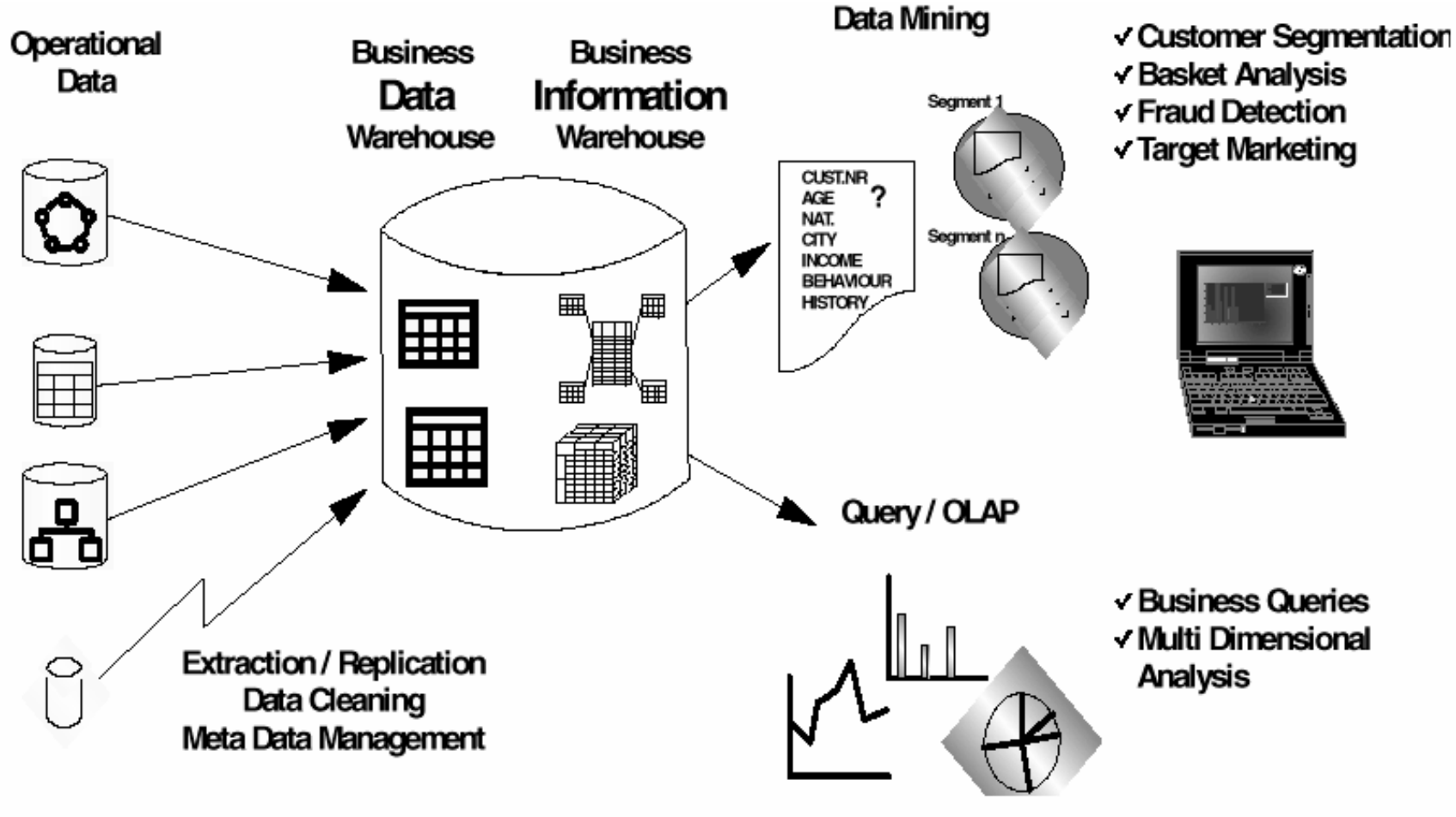
Passi del processo KDD

- Conoscere il dominio applicativo
  - conoscenza pregressa rilevante, obiettivi
- Consolidamento dei dati
  - creare un insieme di dati target
- Selezione e pre-processing
  - data cleaning (fino al 60% dello sforzo)
  - riduzione e proiezione dei dati
    - trovare caratteristiche utili, riduzione di dimensionalità, rappresentazione di invarianti
- Scegliere le funzioni di data mining
  - sommari, classificazione, regressione, associazione, clustering
- Scegliere gli algoritmi di mining
- Individuare i pattern interessanti
- Interpretazione e valutazione
  - visualizzazione, rimozione di pattern ridondanti

Data Mining e Business Intelligence



Ambiente di *Business Intelligence*



La generazione dei dati

- La tecnologia dei dispositivi automatici di raccolta e memorizzazione dei dati ha compiuto sostanziali progressi, sia incrementando prestazioni e capacità che abbattendo i costi
  - Lettori di codici a barre
  - Stazioni di misura atmosferiche
  - Satelliti geografici
- Tali progressi permettono la memorizzazione di enormi moli di dati negli archivi di un numero sempre crescente di organizzazioni pubbliche e private

Esempio: dati gestionali

- Una catena di ipermercati ha un flusso di informazioni completamente automatizzato
  - lettura di codici a barre alle casse
  - lettura di codici a barre nei magazzini
  - ...
- I dati relativi alle vendite possono essere interpretati per migliorare l'efficienza/efficacia dell'azione commerciale
  - offerte speciali
  - andamenti stagionali
  - fidelizzazione clienti

Esempio: dati geografici

- Earth Observing System (NASA)
  è un sistema di raccolta di dati satellitari che memorizza 1450 dataset di 350 Gbyte ciascuno al giorno.
- La quantità di dati generati in 2 settimane equivale a quella generata da LandSAT in 17 anni.

Esempio: dati astronomici

- Nel 1988 in un articolo del Wall Street J. alcuni scienziati impegnati in ricerche spaziali manifestavano sfiducia circa la possibilità di analizzare i dati provenienti dallo spazio
- Nello stesso articolo, la quantità di dati analizzabili realisticamente (al tasso di generazione del 1988) era stimata essere intorno al 10%
- La sonda planetaria Magellano ha inviato 1000 miliardi di byte in 5 anni

Esempio: dati atmosferici

- Tutti i database al National Center for Atmospheric Research
  NCAR,
  http://www.ucar.edu/ucar/news.html
  - 1986: 1 terabyte = 1024 gigabyte = 2^{40} byte = circa 10^{12} byte
  - 2003: 1 petabyte = 1024 terabyte

I pattern “scoperti” sono tutti interessanti?

- Un sistema DM può generare migliaia di pattern: alcuni possono non essere interessanti.
  - Approccio tipico: Human-centered, query-based, focused mining
- Misure di “interestingness”: un pattern è interessante se è:
  - facilmente comprensibile per un essere umano
  - valido su nuovi dati con un certo grado di sicurezza
  - potenzialmente utile o nuovo
  - valida qualche ipotesi che l’utente ha formulato
- Misure obiettive/soggettive:
  - Oggettive: basate su statistiche e/o strutture dei pattern
    - supporto, confidenza...
  - Soggettive: basate sulla conoscenza dell’utente sui dati
    - novità, utilizzabilità...

Possiamo trovare tutti e solo i pattern interessanti?

- Completezza (tutti):
  - Associazioni vs. classificazione vs. clustering
- Ottimizzazione (solo):
  - Approcci
    - genera tutti i pattern e filtra quelli non interessanti
    - genera solo i pattern interessanti – mining query optimization