

DATA WAREHOUSE E KNOWLEDGE DISCOVERY

Prof. Fabio A. Schreiber

**Dipartimento di Elettronica e Informazione
Politecnico di Milano**

DATA WAREHOUSE (DW)

- **TECNICA PER ASSEMBLARE E GESTIRE CORRETTAMENTE DATI PROVENIENTI DA SORGENTI DIVERSE AL FINE DI OTTENERE UNA VISIONE DETTAGLIATA DI UN SISTEMA ECONOMICO**
- **E' UNA RACCOLTA DI DATI**
 - INTEGRATA
 - PERMANENTE
 - VARIABILE NEL TEMPO
 - ORIENTATA AD UN PRECISO ARGOMENTO**A SUPPORTO DI DECISIONI MANAGERIALI**
- **E' L'ELEMENTO DI SEPARAZIONE TRA I CARICHI DI LAVORO OLTP E QUELLI DSS (OLAP)**

ON LINE TRANSACTION PROCESSING (OLTP)

- SONO APPLICAZIONI TIPICHE **DELL'ELABORAZIONE DI DATI GESTIONALI (EDP)**
- LE **TRANSAZIONI** DEVONO AVERE **PROPRIETA' ACID**
 - STRUTTURATE E RIPETITIVE
 - BREVI E ISOLATE
- I **DATI** DEVONO ESSERE **DETTAGLIATI E AGGIORNATI** E L'ACCESSO AVVIENE PER LO PIU' MEDIANTE LA CHIAVE PRIMARIA
- LE DIMENSIONI DELLE BASI DI DATI VARIANO TRA **10² MBYTE E 10 GBYTE**
- LA PRINCIPALE METRICA DI PRESTAZIONE E' IL **THROUGHPUT DELLE TRANSAZIONI**

ON LINE ANALYTICAL PROCESSING (OLAP)

- SONO APPLICAZIONI TIPICHE DEI **SISTEMI DI SUPPORTO ALLE DECISIONI**
- IL CARICO E' FORMATO DA **INTERROGAZIONI MOLTO COMPLESSE** CHE ACCEDONO A MILIONI DI RECORD
- I **DATI** SONO DI TIPO **STORICO, AGGREGATI** A PARTIRE DA VARIE FONTI
- LE **DIMENSIONI** DEL WAREHOUSE RAGGIUNGONO FACILMENTE IL **TBYTE**
- LE PRESTAZIONI CONSIDERATE SONO IL **THROUGHPUT DELLE INTERROGAZIONI E IL LORO TEMPO DI RISPOSTA**

MODELLI DEI DATI PER OLAP

- DEVONO SUPPORTARE **ANALISI E CALCOLI SOFISTICATI** SU DIVERSE DIMENSIONI E GERARCHIE
- IL MODELLO LOGICO DEI DATI PIU' ADATTO E' UNA STRUTTURA MULTIDIMENSIONALE - IL **DATA CUBE**
- LE **DIMENSIONI** DEL CUBO SONO COSTITUITE DAGLI ATTRIBUTI SECONDO I QUALI SI VOGLIONO FARE LE RICERCHE (**CHIAVI**)
- OGNI DIMENSIONE PUO' RAPPRESENTARE A SUA VOLTA UNA **GERARCHIA**
 - DATA {GIORNO - MESE - TRIMESTRE- ANNO}
 - PRODOTTO {NOME - TIPO - CATEGORIA}
(LAND ROVER - FUORISTRADA - AUTOVEICOLI)
- LE **CELLE** DEL CUBO CONTENGONO I VALORI **METRICI** RELATIVI AI VALORI DIMENSIONALI

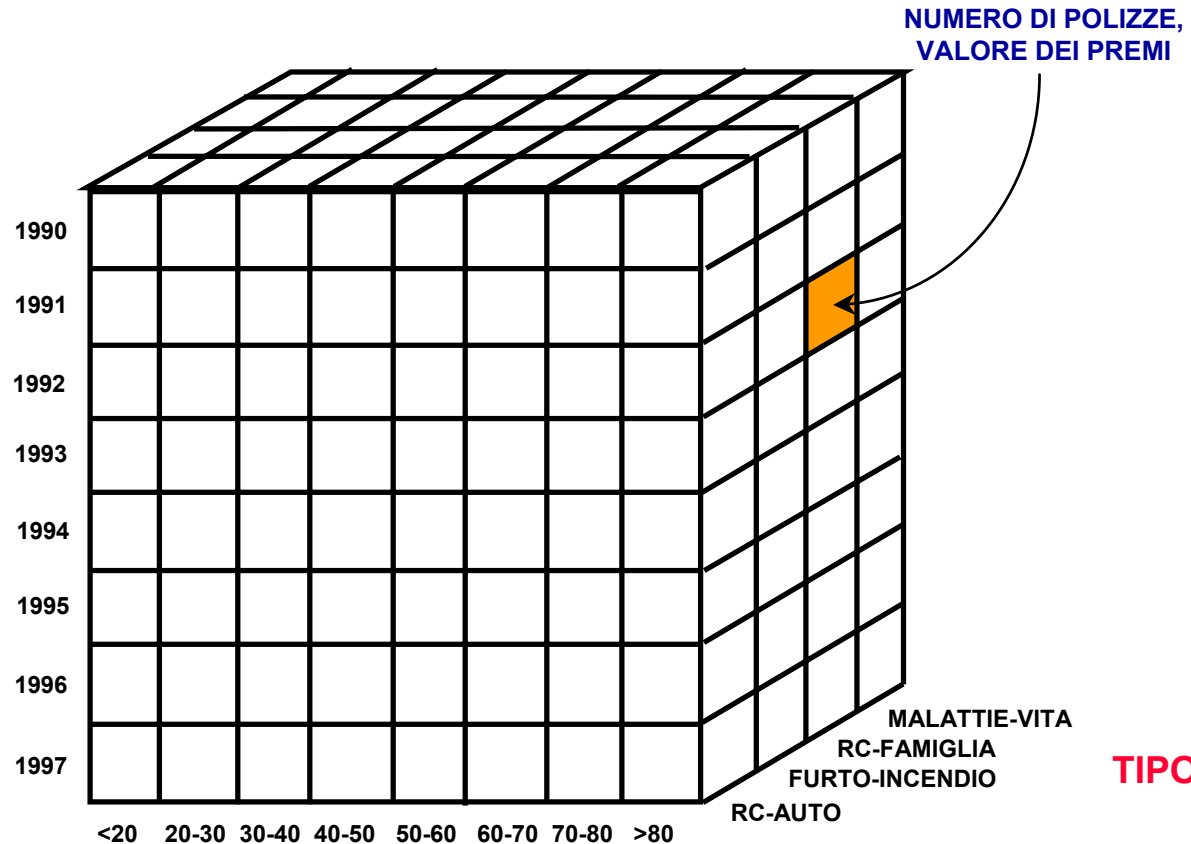
MODELLI LOGICI DEI DATI PER OLAP

ESEMPIO PER UNA COMPAGNIA DI ASSICURAZIONI

DIMENSIONI

VALORI METRICI

ANNO

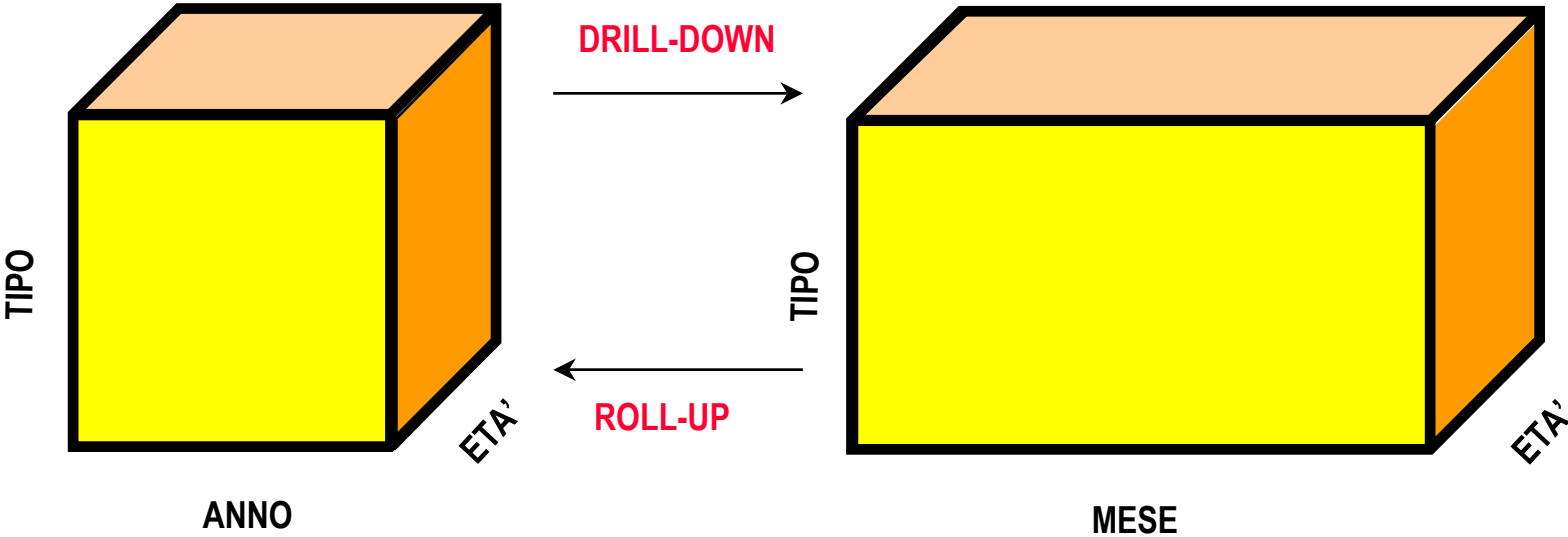


ETA'

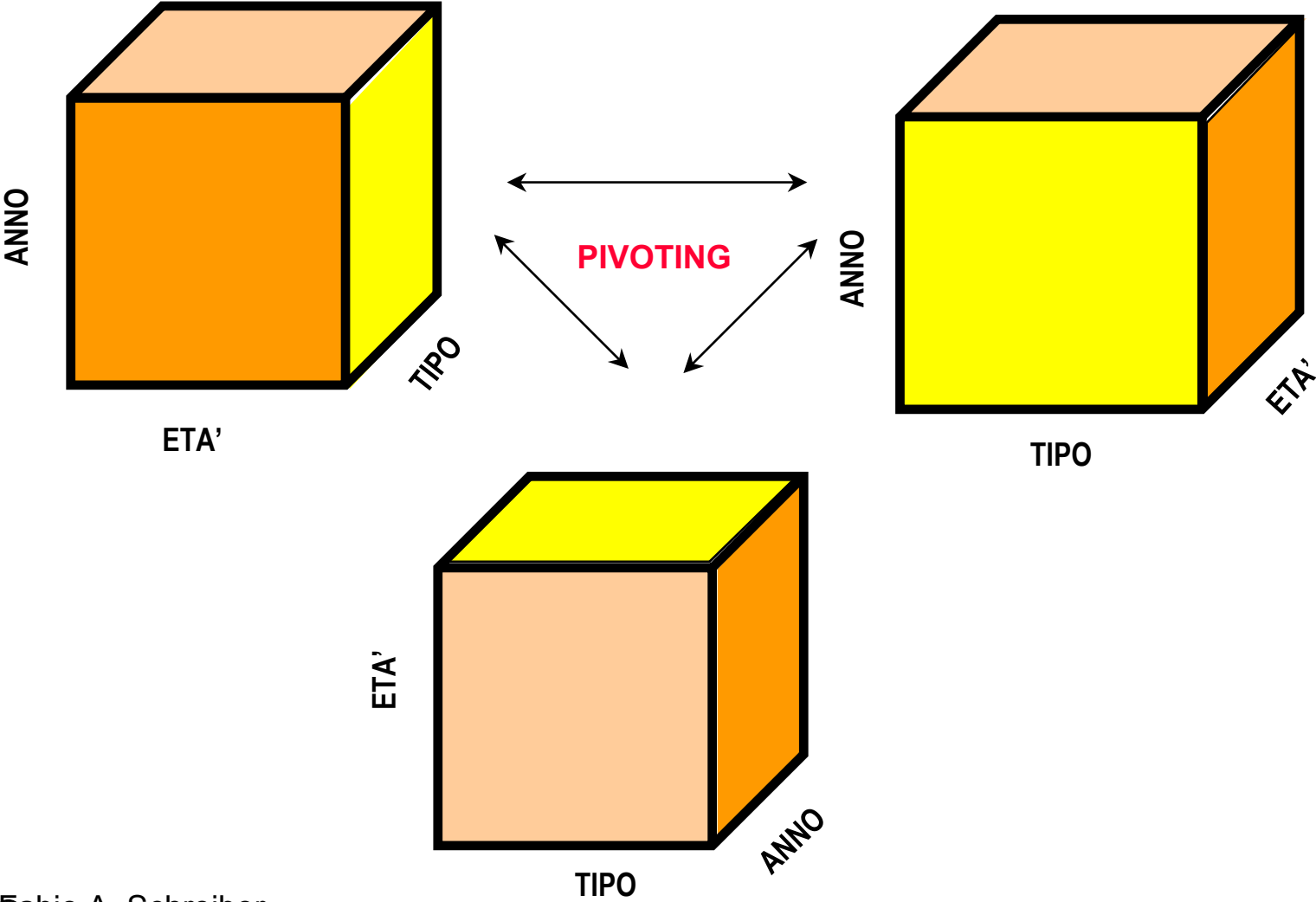
OPERAZIONI OLAP SUL WAREHOUSE

- **ROLL-UP** (ACCUMULARE)
 - AUMENTA IL LIVELLO DI AGGREGAZIONE DEI DATI
- **DRILL-DOWN** (PERFORARE)
 - AUMENTA IL LIVELLO DI DETTAGLIO DEI DATI
- **SLICE-AND-DICE** (AFFETTARE E TAGLIARE A CUBETTI)
 - SELEZIONA E PROIETTA RIDUCENDO LA DIMENSIONALITA' DEI DATI
- **PIVOTING** (FAR PERNO)
 - SELEZIONA DUE DIMENSIONI ATTORNO ALLE QUALI AGGREGARE I DATI METRICI
- **RANKING** (ATTRIBUIRE UNA CLASSE DI MERITO)
 - ORDINA I DATI SECONDO CRITERI PREDEFINITI
- **OPERAZIONI TRADIZIONALI (SELEZIONE, ATTRIBUTI CALCOLATI, ECC.)**

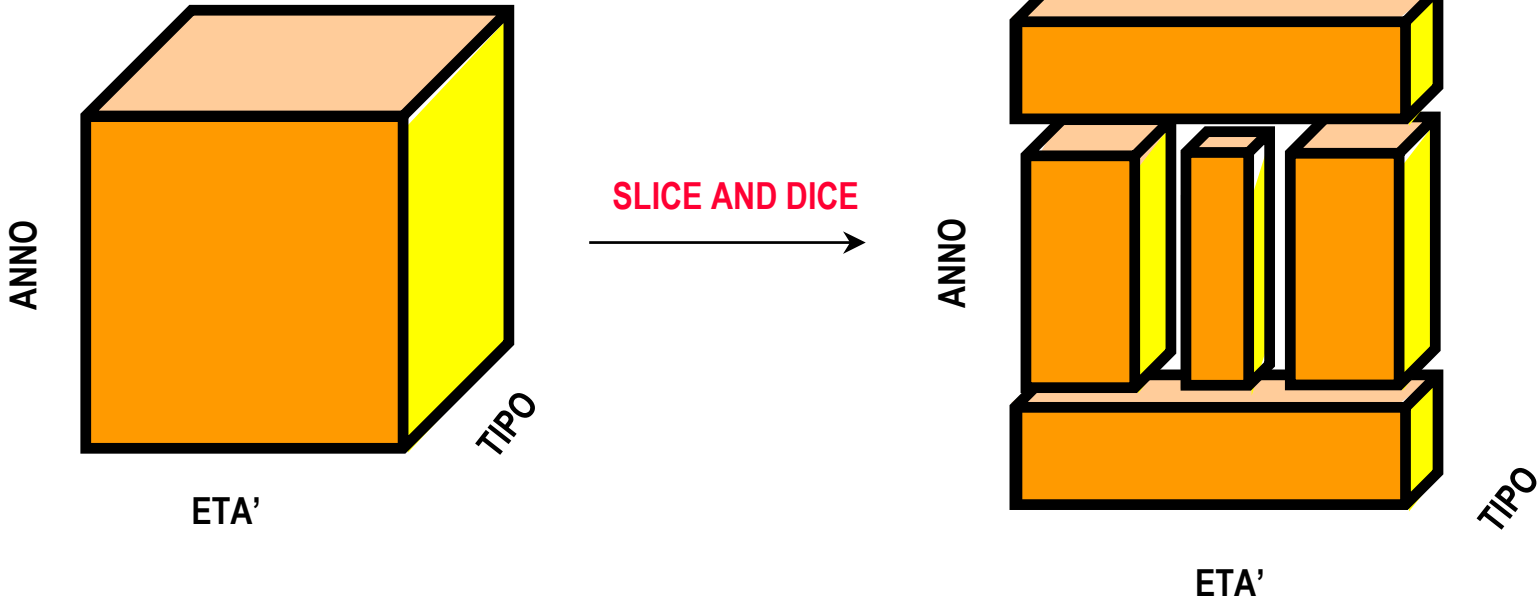
OPERAZIONI OLAP



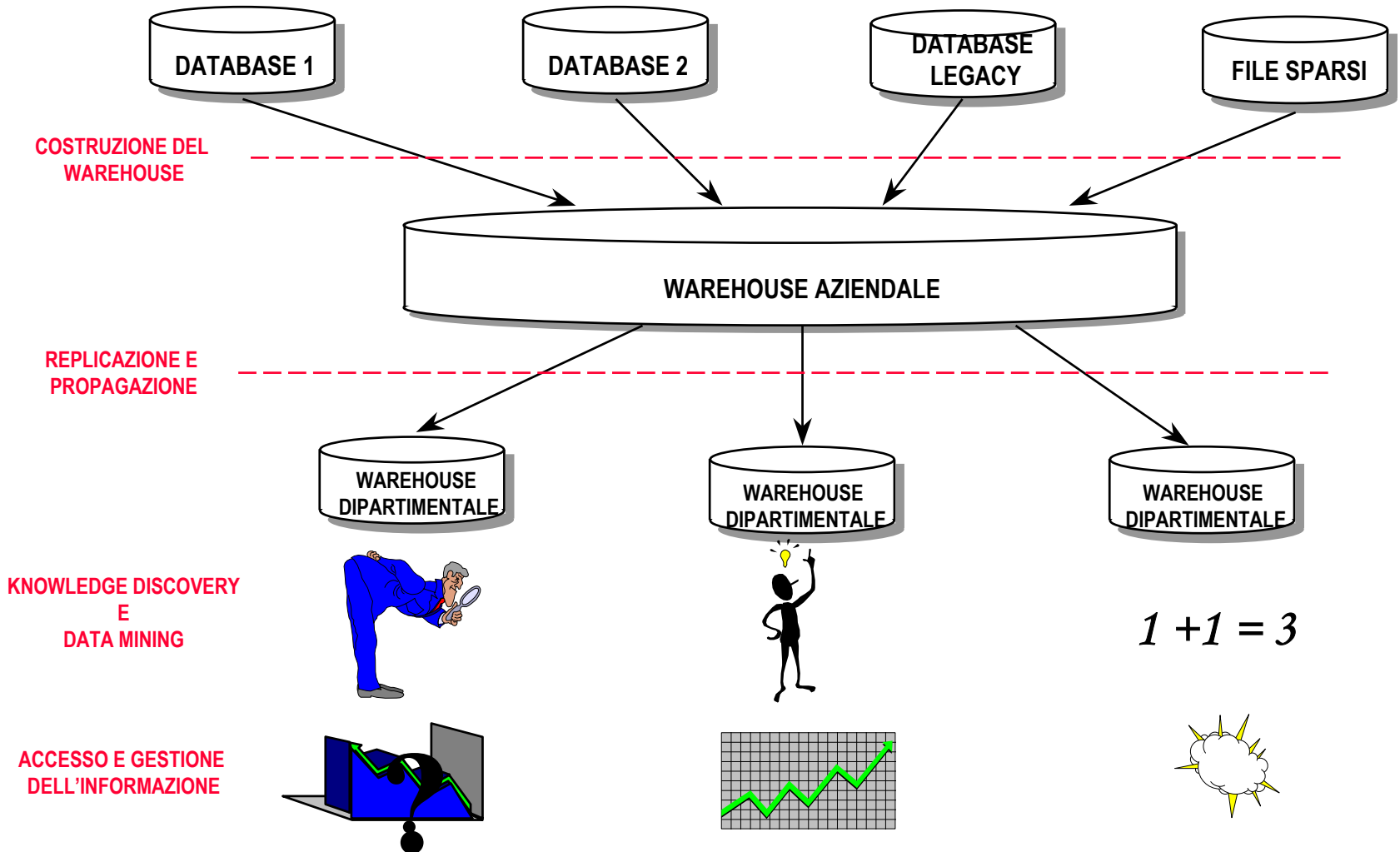
OPERAZIONI OLAP



OPERAZIONI OLAP



STRUTTURA DI UN DW



COSTRUZIONE DEL WAREHOUSE

- I DATI PROVENGONO DA **SORGENTI DIVERSE E “SPORCHE”**
 - SISTEMI LEGACY NON DOCUMENTATI
 - SISTEMI DI PRODUZIONE SENZA CHECK DI INTEGRITA' INTERNI
 - SORGENTI ESTERNE CON DUBBIE CARATTERISTICHE DI QUALITA'
 - SCARSA QUALITÀ DEL DATA ENTRY
 - ASSENZA DI DATI IN ALCUNI CAMPI
- E' INDISPENSABILE **RESTITUIRE LA QUALITA'** AI DATI PER POTERVI BASARE **DECISIONI AFFIDABILI**

COSTRUZIONE DEL WAREHOUSE

- **STRUMENTI PER LA QUALITA' DEI DATI**
 - **PER LA MIGRAZIONE**
 - TRASFORMANO E RIFORMATTANO I DATI DALLE DIVERSE FONTI
 - **PER LA PULIZIA (SCRUBBING)**
 - USANO LA CONOSCENZA DEL DOMINIO PER PULIRE E OMOGENEIZZARE
Jerry L. Jonson, 16 Clarke St., Altuna, PA = Gerry L. Johnson, 16 Clark Street, Altoona, Penn ???
 - **PER IL CONTROLLO (AUDITING)**
 - SCOPRONO REGOLE E RELAZIONI TRA I DATI E NE VERIFICANO IL RISPETTO
- **STRUMENTI PER IL CARICAMENTO DEI DATI**
 - VERIFICANO VIOLAZIONI DI INTEGRITA' REFERENZIALE
 - ORDINANO, AGGREGANO, COSTRUISCONO DATI DERIVATI
 - COSTRUISCONO INDICI E ALTRI PERCORSI DI ACCESSO
 - COSTITUISCONO UN CARICO BATCH MOLTO PESANTE
 - NECESSITA' DI PARALLELIZZARE O RENDERE INCREMENTALE L'OPERAZIONE DI CARICAMENTO

COSTRUZIONE DEL WAREHOUSE

- **CARICAMENTO INCREMENTALE**
 - ATTUATO DURANTE L'AGGIORNAMENTO CARICANDO SOLO LE **TUPLE NUOVE O MODIFICATE**
 - ENTRA IN **CONFLITTO** CON IL FUNZIONAMENTO ORDINARIO
 - RICHIEDE **TRANSAZIONI CORTE** (<1000 RECORD)
 - NECESSITA DI **COORDINAMENTO** PER GARANTIRE LA **CONSISTENZA** DEGLI INDICI E DEI DATI DERIVATI
- **AGGIORNAMENTO**
 - VIENE FATTO **PERIODICAMENTE** IN BASE ALLE ESIGENZE APPLICATIVE
 - USO DI **SERVER DI DUPLICAZIONE**
 - PER **SPEDIZIONE DI DATI**: USANO TRIGGER PER AGGIORNARE, AD OGNI VARIAZIONE DELLA TABELLA SORGENTE, UNA TABELLA LOG DI SNAPSHOT, CHE VIENE QUINDI PROPAGATA
 - PER **SPEDIZIONE DI TRANSAZIONI**: VIENE MONITORATO IL LOG STANDARD E LE VARIAZIONI SULE TABELLE REPLICATE VENGONO TRASFERITE AL SERVER DI DUPLICAZIONE.

METODOLOGIA DI PROGETTAZIONE DI UN DW

- **ANALISI DEI DATI IN INGRESSO**
 - SELEZIONE DELLE SORGENTI INFORMATIVE RILEVANTI
 - TRADUZ. IN MODELLO CONCETTUALE DI RIFERIMENTO (E-R)
 - ANALISI DELLE SORGENTI INFORMATIVE
 - IDENTIFICAZIONE DI **FATTI, MISURE, DIMENSIONI**
- **INTEGRAZIONE IN SCHEMA CONCETTUALE GLOBALE**
- **PROGETTAZIONE DEL DATA WAREHOUSE**
 - **CONCETTUALE**
 - INTRODUZIONE DI DATI **AGGREGATI**, DATI **STORICI**, ECC.
 - **LOGICA**
- **PROGETTAZIONE DEI DATA MART (BD MULTIDIMENS.)**
 - IDENTIFICAZ DI FATTI E DIMENSIONI
 - RISTRUTTURAZIONE DELLO SCHEMA E-R
 - DERIVAZIONE DI UN GRAFO DIMENSIONALE
 - TRADUZIONE NEL MODELLO LOGICO

FREQUENTI MOTIVI DI FALLIMENTO

- **NON CONSIDERARE LA QUALITA' DEI DATI**
 - ACCURATEZZA
 - COMPLETEZZA
 - CONSISTENZA
 - TEMPESTIVITA'
 - DISPONIBILITA'
- **NON MEMORIZZARE I DATI NECESSARI**
 - IGNORARE I DATI CONTENUTI IN FONTI ESTERNE
 - IGNORARE I DATI “SOFT” (p.e. giudizi soggettivi)

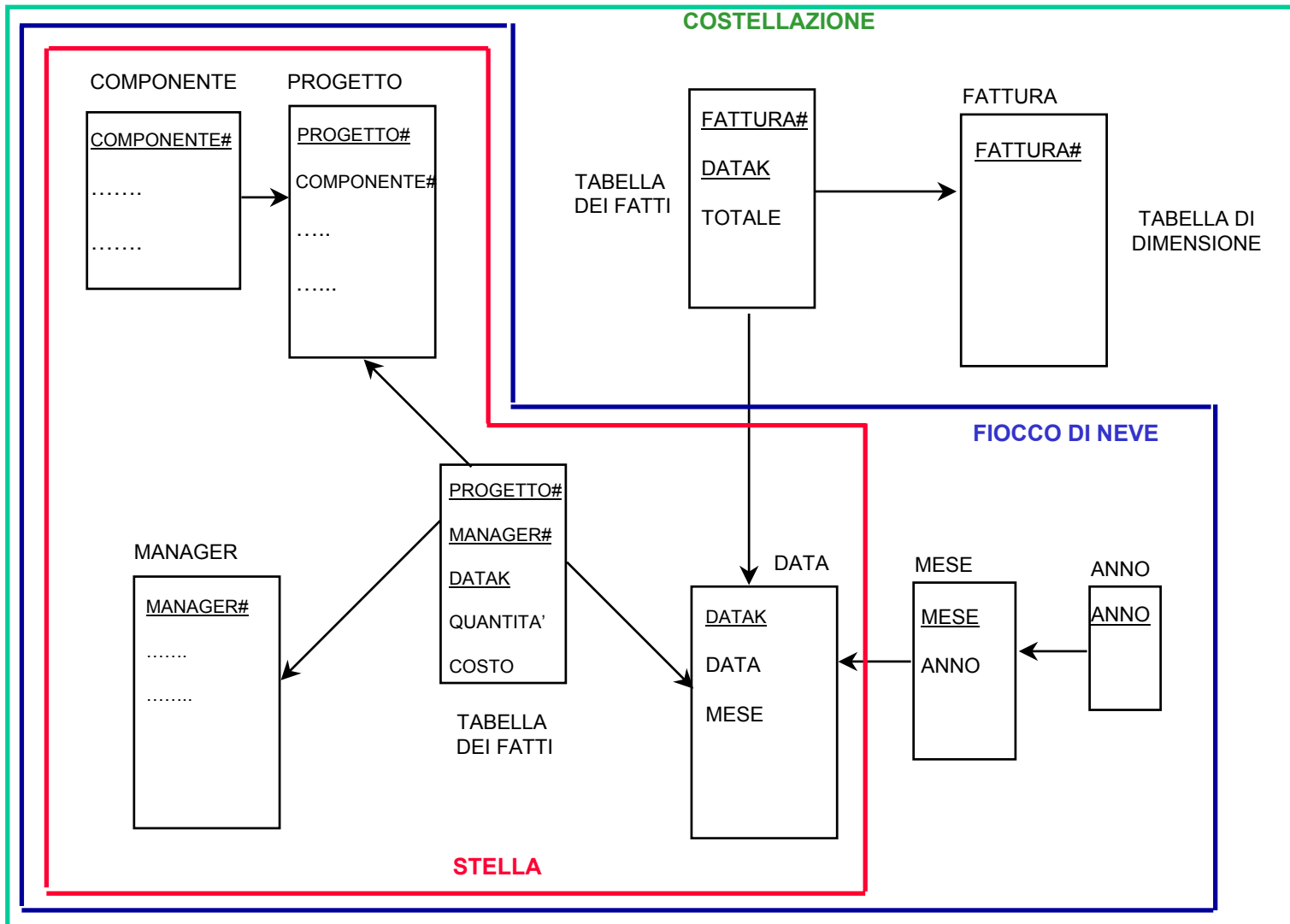
SERVER OLAP MULTIDIMENSIONALE (MOLAP)

- IMPLEMENTA **DIRETTAMENTE** IL MODELLO A CUBO
 - STRUTTURE A **MATRICE MULTIDIMENSIONALE**
 - OTTIME PER STRUTTURE **DENSE**
 - LA RICERCA PER INDIRIZZO SI RIDUCE AD UN CALCOLO ALGEBRICO
- PRESTAZIONI **ELEVATE E COSTANTI** PER L'ELABORAZIONE DELLE INTERROGAZIONI
 - METODI DI ACCESSO **SPECIALIZZATI**
 - AGGREGAZIONE E COMPILAZIONE ESEGUITE IN PRECEDENZA
- **LIMITATA SCALABILITA'** A CAUSA DELLE PREELABORAZIONI
- RICHIEDE MAGGIORI CAPACITA' DA PARTE DELL'AMMINISTRATORE DEI DATI

SERVER OLAP RELAZIONALE (ROLAP)

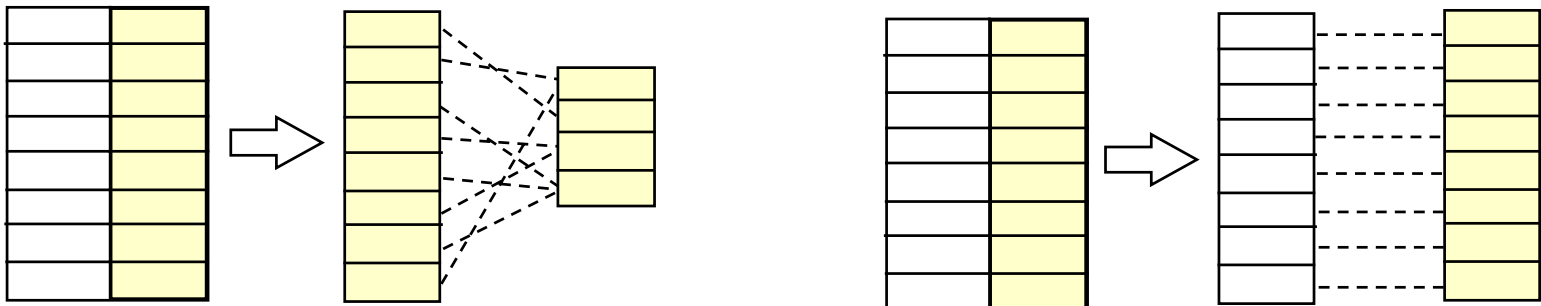
- UTILIZZA UN **RDBMS STANDARD** PER REALIZZARE LA STRUTTURA MULTIDIMENSIONALE, APPLICANDO L'OPERAZIONE DI **GROUP_BY**
- LO **SCHEMA** ASSUME UNA CONFIGURAZIONE **A STELLA** O A **FIOCCO DI NEVE** (NORMALIZZA LE GERARCHIE)
 - **TABELLA CENTRALE DEI FATTI**
 - LE TUPLE SONO COSTITUITE DAI **PUNTATORI** (CHIAVI ESTERNE) ALLE TABELLE DI DIMENSIONE E DAI **VALORI PER LE COORDINATE DESCRITTE** $f=(k_1, \dots, k_n, v_1, \dots v_m)$
 - **TABELLE DI DIMENSIONE**
 - CONTENGONO LE TUPLE CON GLI **ATTRIBUTI RELATIVI A QUELLA DIMENSIONE** $d_1=(k_1, a_1, \dots , a_n)$
 - **COSTELLAZIONE DI FATTI**
 - **PIU' TABELLE DEI FATTI CONDIVIDONO TABELLE DI DIMENSIONE DI UGUALE STRUTTURA**

SCHEMI A STELLA



PRINCIPALI PROBLEMATICHE DI UN DW

- PROGETTO DELLE **STRUTTURE LOGICHE** PER OTTIMIZZARE LE INTERROGAZIONI
 - NECESSITA' DI **MINIMIZZARE I JOIN**
 - DENORMALIZZAZIONE CON RIPETIZIONE DI DATI
 - **RIDUZIONE DELLE DIMENSIONI** DELLE TABELLE
 - PARTIZIONAMENTO **ORIZZONTALE**
 - PARTIZIONAMENTO VERTICALE PER **SPEZZETTAMENTO DELLE RIGHE** (UTILE PER DRILL-DOWN)



PRINCIPALI PROBLEMATICHE DI UN DW

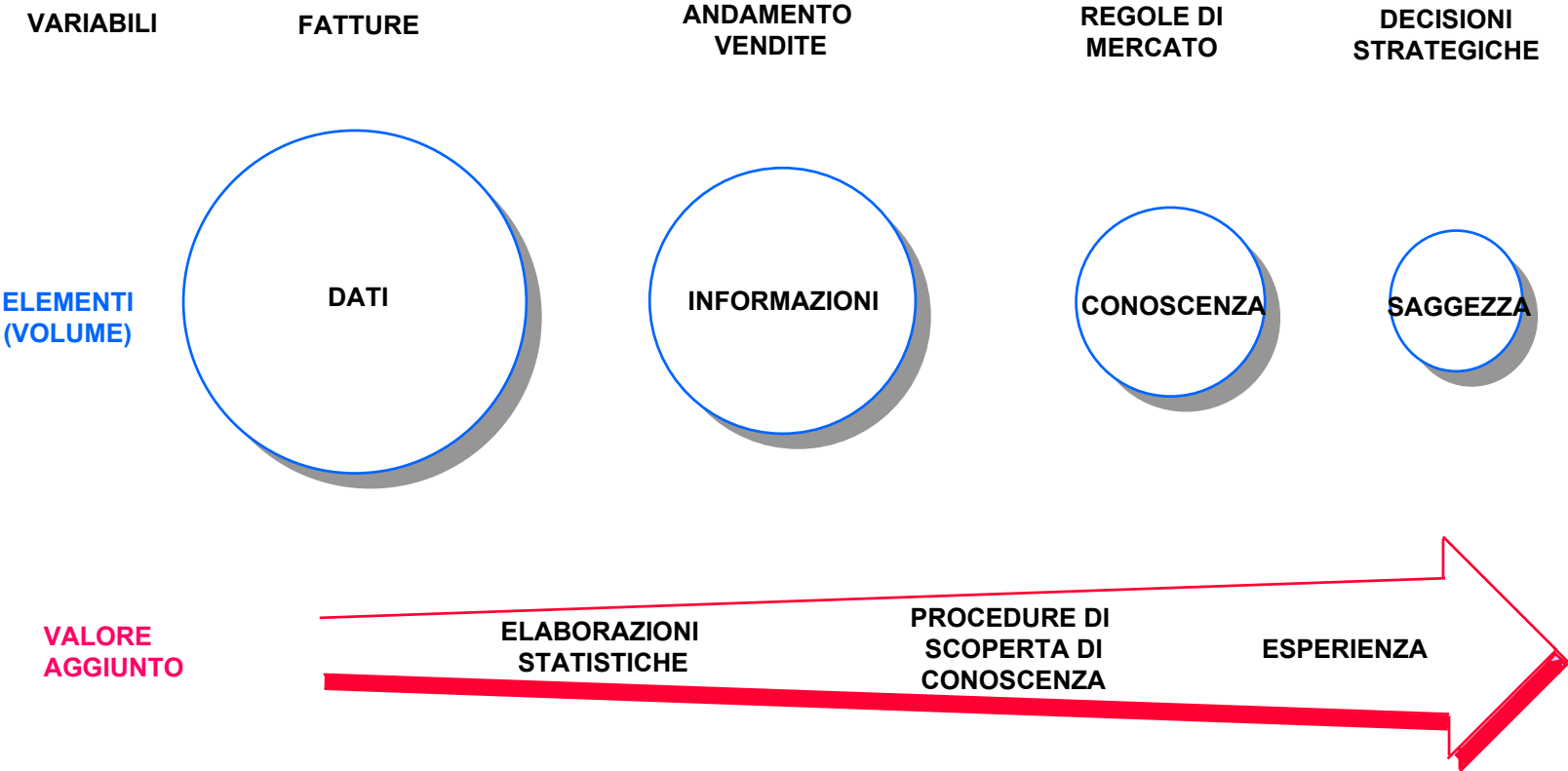
- **PROGETTO DELLE STRUTTURE FISICHE**
 - SCELTA DEGLI **INDICI**
 - SCELTA DELLE **VIEW DA MATERIALIZZARE**
- **MANUTENZIONE DELLE VIEW E DEI METADATI**
- **GESTIONE DELLA REPLICAZIONE**
 - COME E QUANDO FARE GLI AGGIORNAMENTI
- **GESTIONE DELLA CONSISTENZA**
- **REALIZZAZIONE DELLE APPLICAZIONI**

LA SCOPERTA DI CONOSCENZA (KNOWLEDGE DISCOVERY)

***OGNI CONOSCENZA UMANA
PARTE DA **INTUIZIONI**, PROCEDE
ATTRAVERSO **CONCETTI** E
CULMINA IN **IDEE*****

E. Kant

LA GERARCHIA DELLA CONOSCENZA



KNOWLEDGE DISCOVERY E DATA MINING

- **SCOPERTA DI CONOSCENZA NELLE BASI DI DATI (KDD)**
 - IDENTIFICARE LE INFORMAZIONI PIU' SIGNIFICATIVE
 - PRESENTARLE IN MODO OPPORTUNO ALL'UTENTE
- **DATA MINING**
 - APPLICAZIONE DI ALGORITMI AI DATI GREZZI AL FINE DI ESTRARNE CONOSCENZA (RELAZIONI, PERCORSI, ...)
 - OBIETTIVO **PREDITTIVO** (ANALISI DEI SEGNALI, RICONOSCIMENTO DEL PARLATO, ECC.)
 - OBIETTIVO **DESCRITTIVO** (SISTEMI DI SUPPORTO ALLE DECISIONI)

IL PROCESSO DI SCOPERTA DI CONOSCENZA (1)

ANCHE IN PRESENZA DI EFFICACI STRUMENTI RICHIEDE

- **COMPETENZA DELLE TECNICHE** UTILIZZATE
- **OTTIMA CONOSCENZA DEL DOMINIO** DI APPLICAZIONE

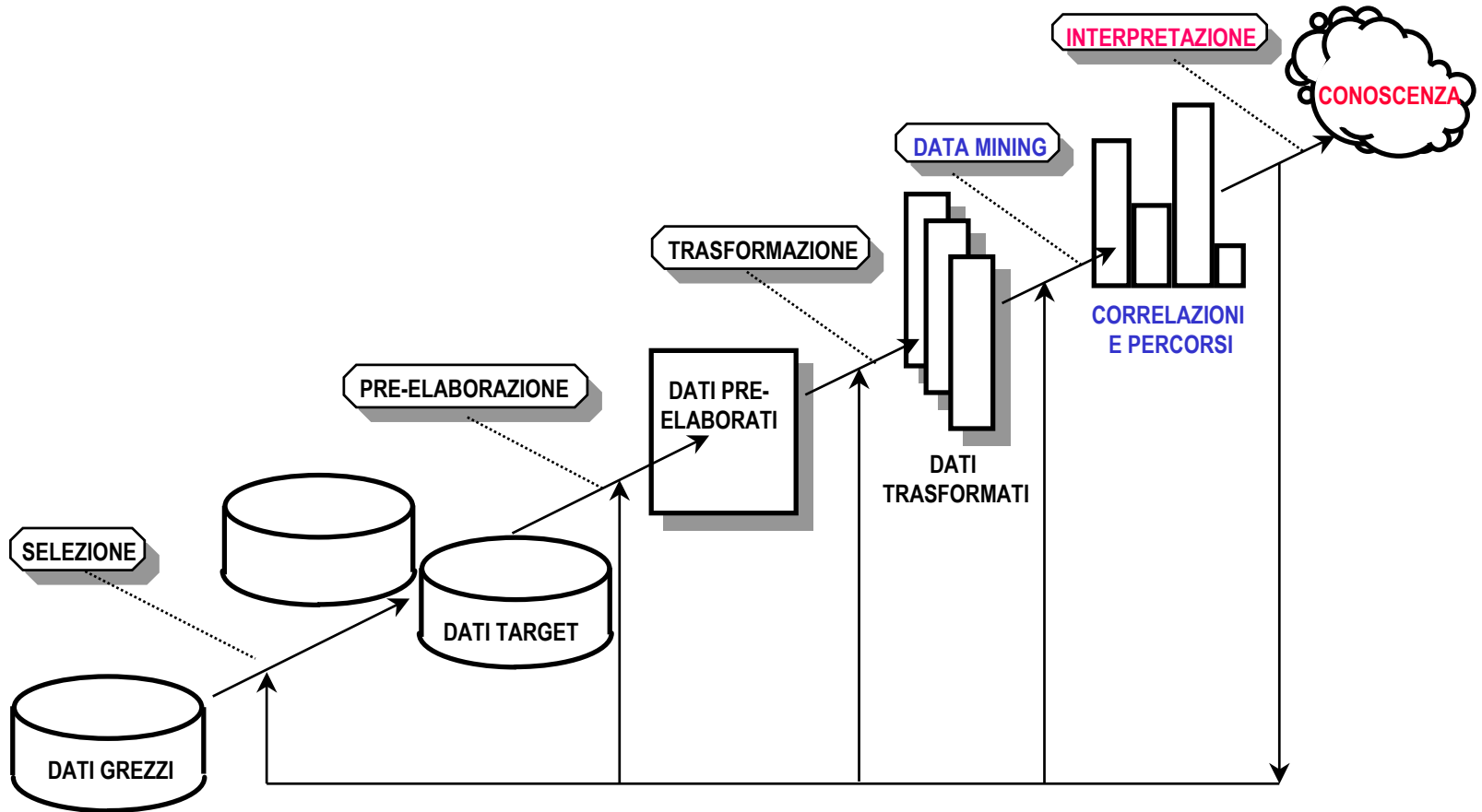
PASSI SUCCESSIVI

- **SELEZIONE**
 - **SCELTA DEI DATI CAMPIONE SUI QUALI FOCALIZZARE L'ANALISI**
- **PRE-ELABORAZIONE**
 - **CAMPIONAMENTO DEI DATI PER RIDURNE IL VOLUME**
 - **PULIZIA DI DATI ERRATI E/O MANCANTI**

IL PROCESSO DI SCOPERTA DI CONOSCENZA (2)

- **TRASFORMAZIONE**
 - OMOGENEIZZAZIONE E/O CONVERSIONE DEI TIPI DI DATI
- **DATA MINING**
 - SCELTA DEL TIPO DI METODO/ALGORITMO
- **INTERPRETAZIONE E VALUTAZIONE**
 - FILTRAGGIO DELL'INFORMAZIONE OTTENUTA
 - EVENTUALE RAFFINAMENTO CON RIPETIZIONE DI PASSI PRECEDENTI
 - PRESENTAZIONE VISUALE (GRAFICA O LOGICA) DEL RISULTATO DELLA RICERCA

IL PROCESSO DI SCOPERTA DI CONOSCENZA (3)



da: G. Piatetsky-Shapiro 1996

APPLICAZIONI DEL DATA MINING

- **VENDITA AL DETTAGLIO E PER CORRISPONDENZA**
 - QUALI “OFFERTE SPECIALI” FARE
 - COME DISPORRE LE MERCI SUGLI SCAFFALI
- **MARKETING**
 - PREVISIONI DI VENDITA
 - PERCORSI DI ACQUISTO DEI PRODOTTI
- **BANCHE**
 - CONTROLLO DEI PRESTITI
 - USO (ED ABUSO) DELLE CARTE DI CREDITO
- **TELECOMUNICAZIONI**
 - AGEVOLAZIONI TARIFFARIE

APPLICAZIONI DEL DATA MINING

- **ASTRONOMIA E ASTROFISICA**
 - CLASSIFICAZIONE DI STELLE E GALASSIE
- **RICERCA CHIMICO FARMACEUTICA**
 - SCOPERTA DI NUOVI COMPOSTI
 - RELAZIONI TRA COMPOSTI
- **BIOLOGIA MOLECOLARE**
 - PATTERN NEI DATI GENETICI E NELLE STRUTTURE MOLECOLARI
- **TELERILEVAMENTO E METEOROLOGIA**
 - ANALISI DEI DATI SATELLITARI
- **STATISTICA ECONOMICA E DEMOGRAFICA**
 - ANALISI DEI CENSIMENTI

STRUTTURA DEGLI ALGORITMI DI DATA MINING

- **RAPPRESENTAZIONE DEL MODELLO**

FORMALISMI PER LA RAPPRESENTAZIONE E LA DESCRIZIONE DEI PERCORSI INDIVIDUABILI

- **VALUTAZIONE DEL MODELLO**

STIMA STATISTICA O LOGICA DELLA RISPONDENZA DI UN PERCORSO AI CRITERI DI RICERCA

- **METODO DI RICERCA**

–DEI PARAMETRI

RICERCA DEI PARAMETRI CHE OTTIMIZZANO I CRITERI DI VALUTAZIONE, FISSATI L'INSIEME DELLE OSSERVAZIONI E LA RAPPRESENTAZIONE DEL MODELLO

–DEL MODELLO

I PARAMETRI VENGONO APPLICATI A MODELLI DI UNA STESSA FAMIGLIA, CHE SI DIVERSIFICANO PER IL TIPO DI RAPPRESENTAZIONE, PER VALUTARNE LA QUALITA'

TIPI DI INFORMAZIONI OTTENUTE

- **ASSOCIAZIONI**
 - INSIEME DI REGOLE CHE SPECIFICA L'OCCORRENZA CONGIUNTA DI DUE (O PIU') ELEMENTI
- **SEQUENZE**
 - POSSIBILITA' DI STABILIRE CONCATENAZIONI TEMPORALI DI EVENTI
- **CLASSIFICAZIONI**
 - RAGGRUPPAMENTI DI ELEMENTI IN CLASSI SECONDO UN MODELLO PREDEFINITO
- **RAGGRUPPAMENTI (CLUSTER)**
 - RAGGRUPPAMENTI DI ELEMENTI IN CLASSI NON DEFINITE A PRIORI
- **TENDENZE (TREND)**
 - SCOPERTA DI ANDAMENTI TEMPORALI CARATTERISTICI CON VALENZA PREVISIONALE

TECNICHE (MODELLI) PER DATA MINING

- **RETI NEURALI**

- STRUMENTO MOLTO POTENTE
- CAPACITA' DI APPRENDIMENTO
- LAVORANO IN MODO OPACO

- **INDUZIONE**

- ALBERI DI DECISIONE (GERARCHIA DI DECISIONI *if - then*)
- INDUZIONE DI REGOLE (INSIEMI NON GERARCHICI DERIVATI DAGLI ALBERI DI DECISIONE)
- SONO AUTOESPLICATIVE

TECNICHE (MODELLI) PER DATA MINING

- **SCOPERTA DI REGOLE**

- REGOLE DI ASSOCIAZIONE (EVENTI SIMULTANEI $X \Rightarrow Y$)
- ASSOCIAZIONI SEQUENZIALI

- **VISUALIZZAZIONE DEI DATI**

- I DATI VENGONO PREPARATI E PRESENTATI IN VESTE GRAFICA PER EVIDENZIARE EVENTUALI IRREGOLARITA' O PATTERN STRANI

TECNICHE PER DATA MINING

INFORMAZIONI TECNICHE	ASSOCIAZIONI	SEQUENZE	CLASSIFICAZIONI	CLUSTER	TENDENZE
RETI NEURALI	—	—	×	×	×
INDUZIONE	—	—	×	×	×
SCOPERTA DI REGOLE	×	×	—	—	—
VISUALIZZAZIONE DEI DATI	×	×	—	×	—

IL “CESTINO DELLA SPESA”

- E' IL **MODELLO PIU' NOTO** SUL QUALE SI APPLICANO LE TECNICHE DI DATA MINING
- E' USATO PRINCIPALMENTE, MA NON SOLO, PER I PROBLEMI DI **VENDITA AL DETTAGLIO**
- LO SCOPO E' QUELLO DI SCOPRIRE **PATTERN RICORRENTI** NEI DATI (REGOLE DI ASSOCIAZIONE)

IL “CESTINO DELLA SPESA”

$$I = \{i_1, \dots, i_k\}$$

INSIEME DI k ELEMENTI (**ITEM**)

$$B = \{b_1, \dots, b_n\}$$

INSIEME DI n SOTTOINSIEMI (**BASKET**) DI I

$$b_i \subseteq I$$

- I
 - PRODOTTI IN UN SUPERMERCATO
 - PAROLE IN UN DIZIONARIO
- B
 - ACQUISTI DI UN SINGOLO CLIENTE
 - SINGOLO DOCUMENTO IN UN CORPUS
- REGOLA DI ASSOCIAZIONE $i_1 \Rightarrow i_2$
 - i_1 E i_2 COMPAIONO ASSIEME IN **ALMENO** $s\%$ DEGLI n BASKET (**SUPPORTO**)
 - DI TUTTI I BASKET CHE CONTENGONO i_1 **ALMENO** $c\%$ CONTENGONO ANCHE i_2 (**CONFIDENZA**)

ESEMPI DI REGOLE

- **REGOLE CHE HANNO “DIET COKE” COME CONSEGUENTE**
 - AIUTANO A DEFINIRE LE STRATEGIE PER **AUMENTARE LE VENDITE** DI UN PRODOTTO
- **REGOLE CHE HANNO “GRISSINI” TRA ANTECEDENTI**
 - AIUTANO A CAPIRE L’IMPATTO DELLA **CESSAZIONE DELLA VENDITA** DI UN PRODOTTO
- **REGOLE CHE HANNO “WURSTEL” TRA GLI ANTECEDENTI E “SENAPE” COME CONSEGUENTE**
 - EVIDENZIANO GLI **ABBINAMENTI** DI PRODOTTI PRESENTI NELL’ANTECEDENTE CHE **INDUCONO LA VENDITA DEL CONSEGUENTE**

ESEMPI DI REGOLE

- **REGOLE CHE COLLEGANO ELEMENTI NELLO SCAFFALE A A ELEMENTI NELLO SCAFFALE B**
 - AIUTANO UNA PIU' EFFICACE ALLOCAZIONE DEI PRODOTTI NEGLI SCAFFALI
- **LE MIGLIORI k REGOLE CHE HANNO “GRISSINI” NEL CONSEGUENTE**
 - IN TERMINI DEL FATTORE DI **CONFIDENZA**
 - INTERMINI DEL FATTORE DI **SUPPORTO**

QUALCHE PROBLEMA

$c \rightarrow$ NEL CESTINO C'E' CAFFE'

$t \rightarrow$ NEL CESTINO C'E' THE

$\bar{c} \rightarrow$ NEL CESTINO NON C'E' CAFFE'

$\bar{t} \rightarrow$ NEL CESTINO NON C'E' THE

	c	\bar{c}	Σ righe
t	20	5	25
\bar{t}	70	5	75
Σ colonne	90	10	100

$t \Rightarrow c$ E' VERA???

$$s = 20\%$$

$$c = P[t \wedge c] / P[t] = 20/25 = 80\%$$

FORSE SI, MA ... COLORO CHE
COMPRANO COMUNQUE IL CAFFE'
 SONO PARI AL **90% !!!**

ATTENZIONE! ESISTE UNA **CORRELAZIONE** TRA THE E CAFFE'

$$r = P[t \wedge c] / (P[t] \times P[c]) = 0.89$$

DATA MINING IN AMBIENTE RELAZIONALE

- ESTENSIONE DI SQL CON OPERATORI DI DATA MINING INTEGRATI NEL LINGUAGGIO
- INTEGRARE UN SERVER SQL CON UN MOTORE DI DATA MINING
 - **MINE RULE** PER LE REGOLE DI ASSOCIAZIONE
 - **MINE CLASSIFICATION** PER I PROBLEMI DI CLASSIFICAZIONE
 - **MINE INTERVAL** PER LA DISCRETIZZAZIONE DI ATTRIBUTI CONTINUI

MINE RULE: ESEMPIO

MINE RULE AssociazioneSemplice AS

**SELECT DISTINCT 1..n item AS BODY, 1..1 item AS HEAD, SUPPORT, CONFIDENCE
FROM Purchase**

GROUP BY transaction

EXTRACTING RULES WITH SUPPORT: 0.1, CONFIDENCE: 0.2

TR	CLIENTE	OGGETTO	DATA	PREZZO	QUANT.
1	Rossi	pantal. da ski	17/12	90.000	1
	Rossi	scarponi mont.	17/12	180.000	1
2	Bianchi	camicia sport	18/12	70.000	2
	Bianchi	scarpe marron	18/12	250.000	1
	Bianchi	giacca	18/12	400.000	1
3	Rossi	giacca	18/12	400.000	1
4	Bianchi	camicia sport	19/12	70.000	3
	Bianchi	giacca	19/12	400.000	2

MINE RULE: ESEMPIO

CORPO	TESTA	SUPPORTO	CONFIDENZA
pantaloni da ski	scarponi montagna	0.25	1
scarponi montagna	pantaloni da ski	0.25	1
camicia sport	scarpe marron	0.25	0.5
camicia sport	giacca	0.5	1
scarpe marron	camicia sport	0.25	1
scarpe marron	giacca	0.25	1
giacca	camicia sport	0.5	0.66
giacca	scarpe marron	0.25	0.33
camicia sport, scarpe marron	giacca	0.25	1
camicia sport, giacca	scarpe marron	0.25	0.5
scarpe marron, giacca	camicia sport	0.25	1

PROBLEMA DI CLASSIFICAZIONE

CIASCUN ELEMENTO (RECORD) DI UN INSIEME DI DATI E' ASSOCIATO, IN BASE A ESPERIENZE O OSSERVAZIONI PREGRESSE, AD UNA **CARATTERISTICA DISTINTIVA** CHIAMATA **CLASSE**

- **FASE 1: APPRENDIMENTO (TRAINING)**
 - COSTRUZIONE DI UN MODELLO SULL'INSIEME NOTO (TRAINING SET) IN MODO CHE **OGNI CLASSE SIA UNA PARTIZIONE** DELL'INSIEME
- **FASE 2: APPLICAZIONE**
 - IL MODELLO INDIVIDUATO VIENE UTILIZZATO PER CLASSIFICARE **NUOVI DATI**

PROBLEMA DI CLASSIFICAZIONE: ESEMPIO

- **FASE 1: APPRENDIMENTO**

```
MINE CLASSIFICATION CarInsuranceRules AS  
SELECT DISTINCT RULES ID, *, CLASS  
FROM CarInsurance  
CLASSIFY BY Risk
```

- **FASE 2: APPLICAZIONE**

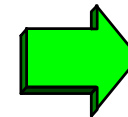
```
MINE CLASSIFICATION TEST ClassifiedApplicants AS  
SELECT DISTINCT *, CLASS  
FROM Applicants  
    USING CLASSIFICATION FROM CarInsuranceRules AS RULES
```

PROBLEMA DI CLASSIFICAZIONE: ESEMPIO

AGE	CAR TYPE	RISK
17	sports	high
43	family	low
68	family	low
32	truck	low
23	family	high
18	family	high
20	family	high
45	sports	high
50	truck	low
64	truck	high
46	family	low
40	family	low

1. IF Age \leq 23 THEN Risk IS High;
2. IF CarType = sports THEN Risk IS High;
3. IF CarType IN {family, truck} AND Age > 23 THEN Risk IS Low;
4. DEFAULT Risk IS Low

AGE	CAR TYPE
22	family
60	family
35	sports



MINE
CLASSIFICATION
TEST

AGE	CAR TYPE	CLASS
22	family	high
60	family	low
35	sports	high

PROBLEMA DI DISCRETIZZAZIONE

SI TRATTA DI **TRASFORMARE UN ATTRIBUTO NUMERICO IN UNO CATEGORIALE** FRAZIONANDO IL DOMINIO NUMERICO IN INTERVALLI CHE, ASSOCIATI CIASCUNO AD UNA ETICHETTA DI CLASSE, COSTITUISCONO IL DOMINIO CATEGORIALE

- **METODI DI DISCRETIZZAZIONE**

- **NON SUPERVISIONATI**

- VIENE CONSIDERATA SOLAMENTE LA DISTRIBUZIONE DEI VALORI DELL'ATTRIBUTO

- **SUPERVISIONATI**

- SI CERCA DI CONSERVARE QUANTA PIU' INFORMAZIONE POSSIBILE RELATIVA ALLE CLASSI DI UN ATTRIBUTO ASSOCIATO ALLE TUPLE

PROBLEMA DI DISCRETIZZAZIONE

DISCRETIZZAZIONE SEMPLICE

- INTERVALLI DI **UGUALE AMPIEZZA**

IL DOMINIO NUMERICO VIENE SUDDIVISO IN UN ASSEGNATO NUMERO DI INTERVALLI DI UGUALE AMPIEZZA

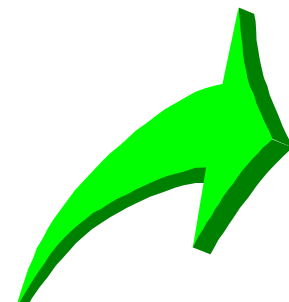
- INTERVALLI DI **UGUALE FREQUENZA**

GLI INTERVALLI SONO PIU' STRETTI DOVE I VALORI SONO DENSII E PIU' AMPI DOVE I VALORI SONO SPARSI

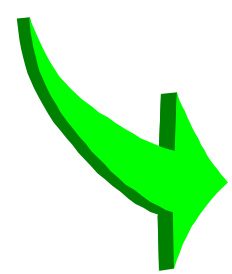
```
MINE INTERVAL IntervalliDiUgualeAmpiezza AS  
SELECT DISTINCT ID, LOWER, UPPER  
GENERATING AssicurazioneAutoDiscreta  
FROM AssicurazioneAuto  
DISCRETIZE Età BY WIDTH USING 3 INTERVALS
```

PROBLEMA DI DISCRETIZZAZIONE: ESEMPIO

AGE	CAR TYPE	RISK
17	sports	high
43	family	low
68	family	low
32	truck	low
23	family	high
18	family	high
20	family	high
45	sports	high
50	truck	low
64	truck	high
46	family	low
40	family	low



ID	LOWER	UPPER
1	17	34
2	34	52
3	52	68



AGE	CAR TYPE	RISK
1	sports	high
---	-----	----
1	truck	low
2	family	low
---	-----	----
2	truck	low
3	-----	----
3	family	low