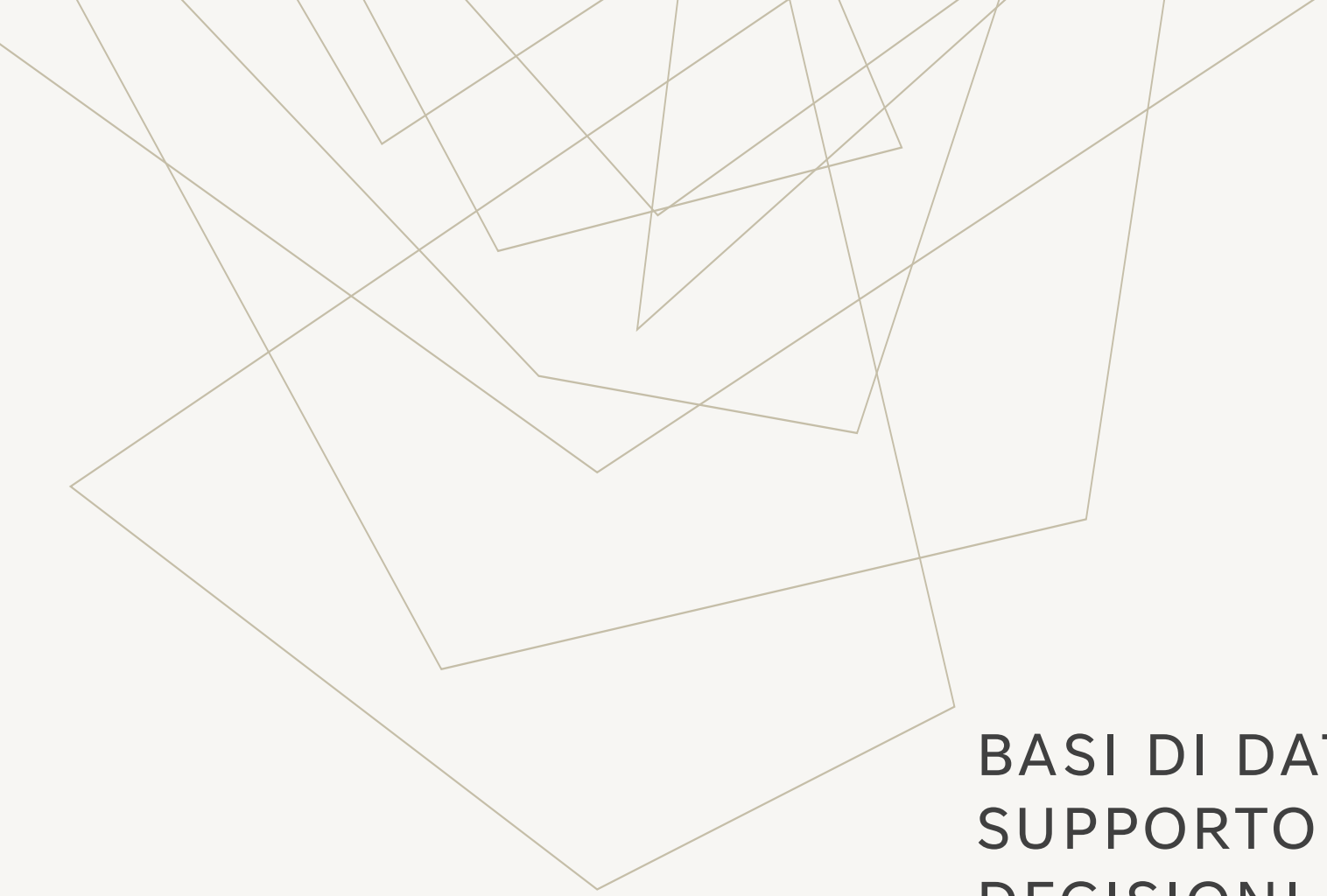


BASI DI DATI PER IL SUPPORTO ALLE DECISIONI

Vincenzo Calabrò

Limiti delle basi di dati relazionali

Le basi di dati sono adatte per la gestione efficiente di dati in linea (OLTP), ma offrono supporto limitato all'analisi dei dati (OLAP):

- SQL non adatto agli analisti di alto livello
- complessità e rigidità delle applicazioni
  - difficile ottimizzare applicazioni in modo che soddisfino sia le esigenze della gestione in linea sia quelle di analisi

Necessità di soluzioni che rendano i dati prodotti per la gestione operativa utilizzabili anche per la gestione strategica

OLTP: On Line Transaction Processing

Tradizionale elaborazione di transazioni, che realizzano i processi operativi

- operazioni predefinite e relativamente semplici
- ogni operazione coinvolge “pochi” dati
- dati di dettaglio, aggiornati
- proprietà ACIDe delle transazioni essenziali

OLAP: On Line Analytical Processing

Elaborazione di operazioni per il supporto alle decisioni

- operazioni complesse e casuali
- ogni operazione può coinvolgere molti dati
- dati aggregati, storici, anche non aggiornati
- proprietà ACIDe non rilevanti
  - tipicamente operazioni di lettura

OLTP vs OLAP

- la configurazione di sistemi dedicati a uno solo dei due compiti è un problema gestibile

- è estremamente difficile far convivere i due carichi di lavoro
  - disomogeneità di utenti e requisiti
  - differenze tecniche

OLTP vs OLAP: differenze (1)

	OLTP	OLAP
Utente	impiegato	dirigente
Funzione	operazioni giornaliere	supporto alle decisioni
Progettazione	orientata all'applicazione	orientata ai dati
Dati	correnti, aggiornati, dettagliati, relazionali, omogenei	storici, aggregati, multidimensionali, eterogenei
Uso	ripetitivo	casuale
Accesso	read-write, indicizzato	read, sequenziale
Unità di lavoro	transazione breve	interrogazione complessa
Accesso record	decine	milioni
N. utenti	migliaia	centinaia
Dimensione	100MB - 1GB	100GB - 1TB
Metrica	throughput	tempo di risposta

OLTP vs OLAP: differenze (2)

- conflitto di lock
  - OLTP: tante transazioni rapide con lock esclusivi
  - OLAP: poche transazioni lunghe con lock condivisi
  - OLTP+OLAP:
    - le transazioni OLTP sono molto rallentate
    - o le query OLAP non riescono ad essere eseguite
- uso degli indici
  - OLTP: pochi e solo se servono
  - OLAP: tanti per coprire ogni esigenza
  - OLTP+OLAP
    - o le transazioni OLTP rallentano per l'aggiornamento di molti indici
    - o le query OLAP non hanno a disposizione gli indici necessari

OLTP vs OLAP: differenze (3)

- precomputazione di query
  - OLTP: molto raramente, per problemi di consistenza e di carico
  - OLAP: aspetto chiave per abbassare i tempi di risposta

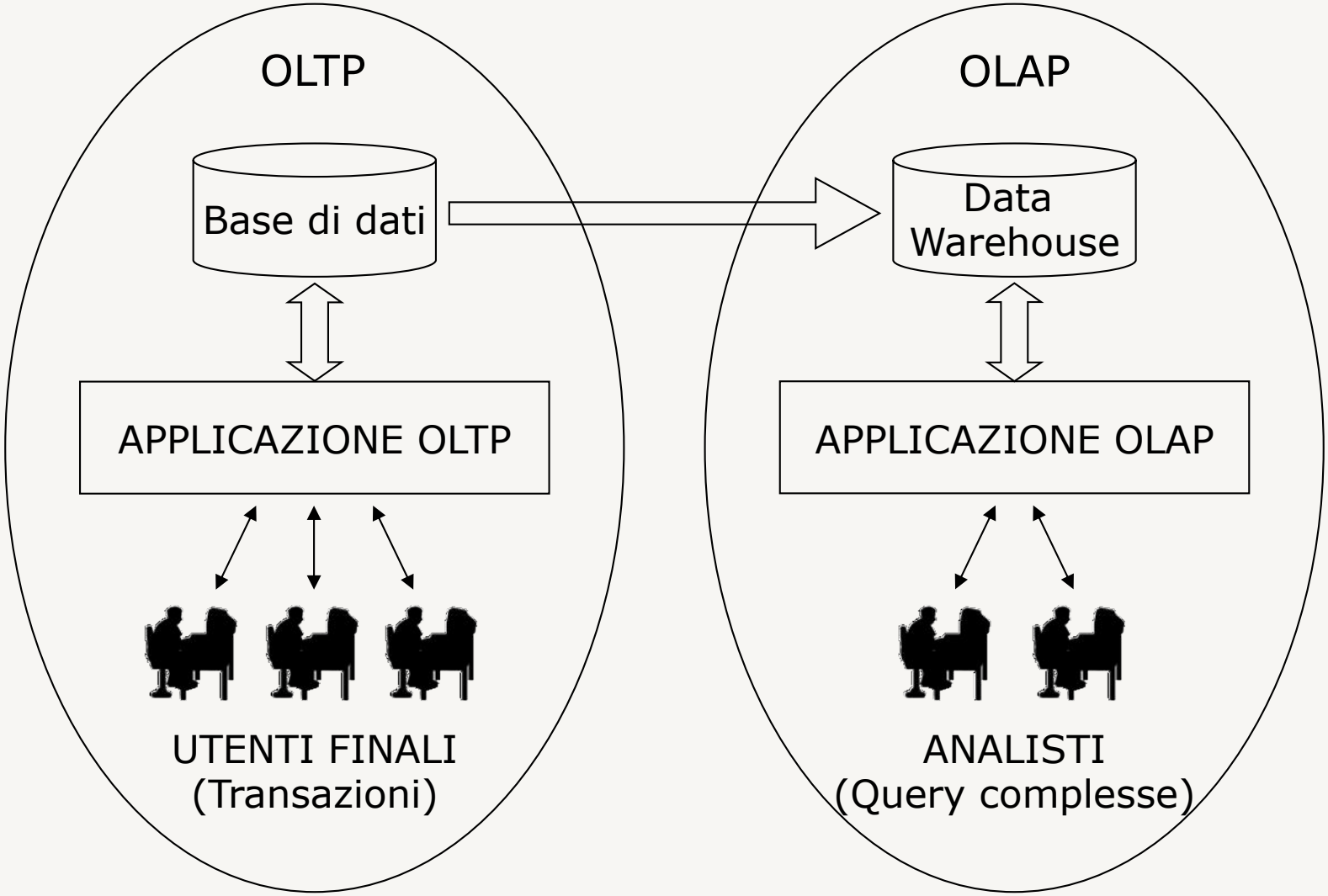
- modello logico
  - OLTP: elevata frammentazione e tante tabelle, generalmente normalizzate
  - OLAP: poche tabelle denormalizzate

OLTP vs OLAP: soluzione

- hanno un conflitto intrinseco
  - non sparisce con l'aumentare della potenza di calcolo

- soluzione
  - separare i due ambienti
    - basi di dati (OLTP)
    - data warehouse (OLAP)

OLTP e OLAP: separazione degli ambienti



Data Warehouse: caratteristiche (1)

Base di dati per il supporto alle decisioni (OLAP)

- integrata
- dati in forma aggregata
- dati storici/temporali
- fuori linea e non volatile
- autonoma

Data Warehouse: caratteristiche (2)

Integrata

- riunisce i dati di diverse sorgenti informative preesistenti
- rappresenta i dati in modo univoco, riconciliando le eterogeneità delle diverse rappresentazioni
  - nomi
  - struttura
  - codifica

Data Warehouse: caratteristiche (3)

Forma aggregata

- i dati delle sorgenti sono aggregati sulla base di opportune coordinate
  - es., tempo, collocazione geografica, tipologia di prodotto
- è orientata ai dati (non alle applicazioni)
  - le basi di dati operazionali sono costruite a supporto dei singoli processi operativi o applicazioni
  - la data warehouse è costruita attorno alle principali entità del patrimonio informativo aziendale

Data Warehouse: caratteristiche (4)

Dati storici/temporali

- tiene l'evoluzione storica delle informazioni con un ampio orizzonte temporale e indicazione di elementi di tempo
  - è di interesse l'evoluzione storica delle informazioni con un orizzonte temporale dell'ordine degli anni
  - le basi di dati operazionali mantengono il valore corrente delle informazioni (orizzonte temporale di pochi mesi)

Data Warehouse: caratteristiche (5)

Fuori linea e non volatile

- tipicamente formata in modo asincrono e periodico rispetto alle sorgenti
  - operazioni di accesso e interrogazione “diurne”
  - operazioni di caricamento e aggiornamento dei dati “notturne”
  - le operazioni possono riguardare milioni di record (in una base di dati operazionale, vengono acceduti, inseriti, modificati, cancellati pochi record alla volta)

Data Warehouse: caratteristiche (6)

Autonoma

- fisicamente separata dalle sorgenti informative
  - ragioni tecniche
  - non esiste un'unica base di dati operazionale che contiene tutti i dati di interesse
  - la data warehouse deve essere integrata
  - i dati di interesse sarebbero comunque diversi
    - devono essere mantenuti dati storici
    - devono essere mantenuti dati aggregati
  - l'analisi dei dati richiede organizzazioni speciali dei dati e metodi di accesso specifici
  - degrado generale delle prestazioni senza la separazione

OLAP e controllo di gestione

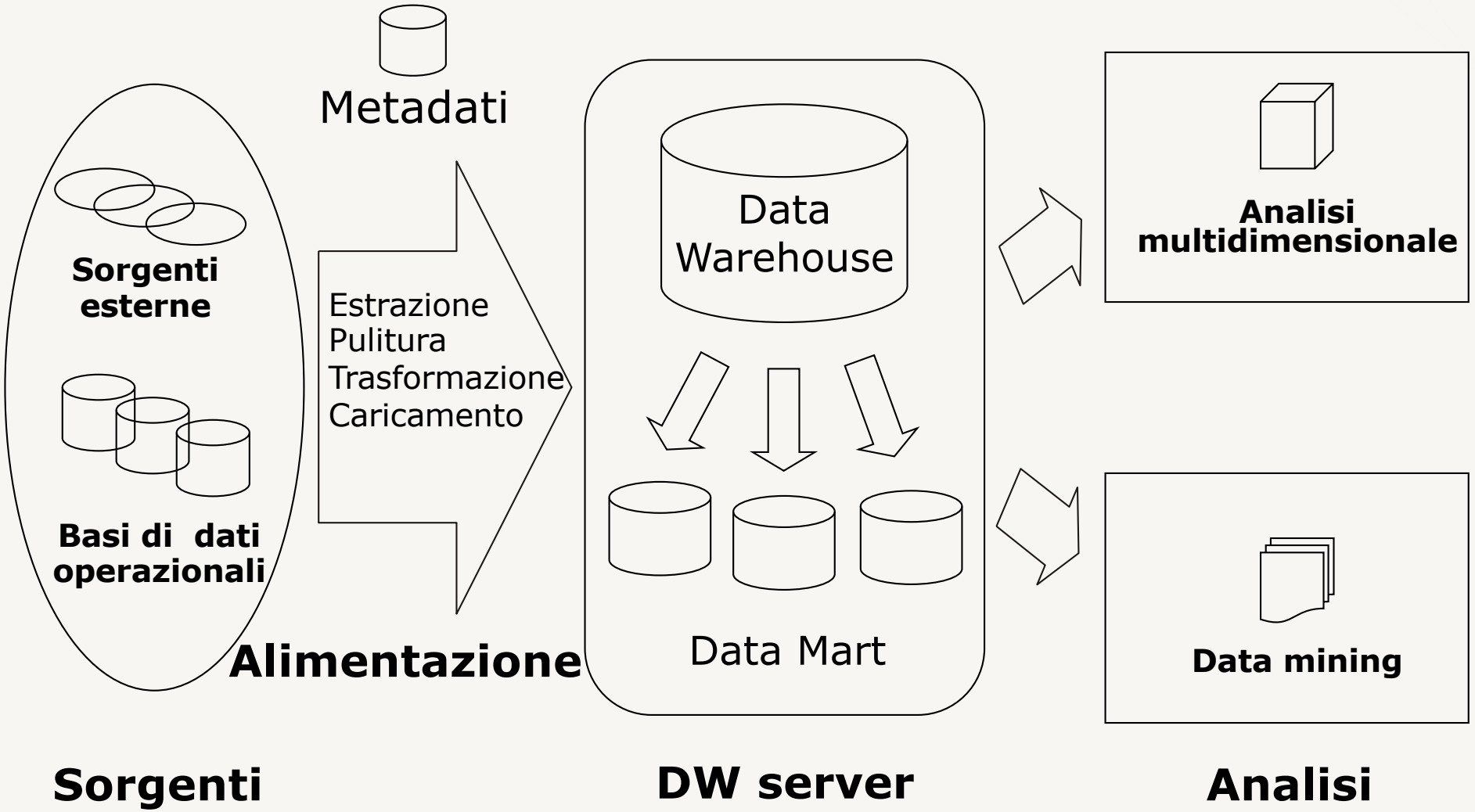
Le applicazioni per il controllo di gestione

- hanno alcune caratteristiche in comune con le applicazioni OLAP
  - principalmente lettura di dati
  - necessità di svolgere aggregazioni
  - confronti con serie storiche
- sono una cosa diversa!
  - staticità delle query (non query ad hoc, ma report fissi)
  - necessità di mantenere il dettaglio dello schema
- possono condividere informazioni con l'OLAP, ma devono essere progetti diversi

Data warehouse: architettura (1)

- sorgenti dei dati
- data warehouse server (dedicato all'OLAP)
  - spesso gestisce anche viste materializzate (data mart)
  - può avere associati metadati e strumenti per l'assistenza allo sviluppo
- sistema di alimentazione
  - estrazione dei dati dalle sorgenti
  - pulizia (elimina errori e inconsistenze) e trasformazione
  - caricamento nella data warehouse
- strumenti di analisi
  - analisi multidimensionale (operazioni interattive di aggregazione/disaggregazione dei dati)
  - data mining (ricerche sofisticate per inferire nuovi dati e correlazioni fra i dati)

Data warehouse: architettura (2)



Modello multidimensionale

Rappresentazione dei dati ad alto livello

- prescinde dai criteri di memorizzazione
- favorisce l'analisi

Concetti rilevanti

- fatto: concetto sul quale centrare l'analisi
- misura: proprietà atomica di un fatto da analizzare
- dimensione: descrive una prospettiva lungo la quale effettuare l'analisi

Fatti, misure, dimensioni: esempi

Catena di negozi

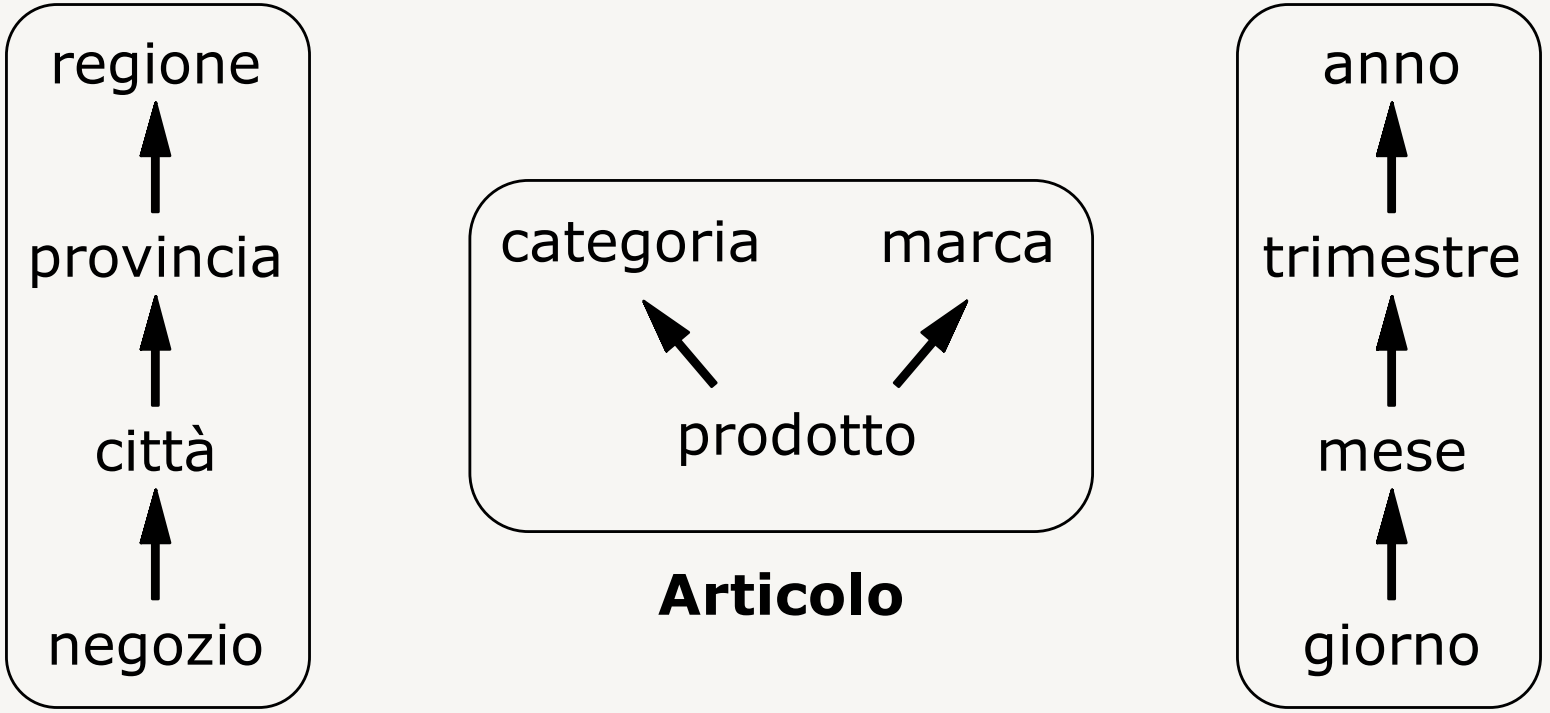
- fatto: vendita
- misure: quantità venduta, incasso
- dimensioni: articolo, tempo, luogo

Compagnia telefonica

- fatto: telefonata
- misure: costo, durata
- dimensioni: chiamante, chiamato, tempo

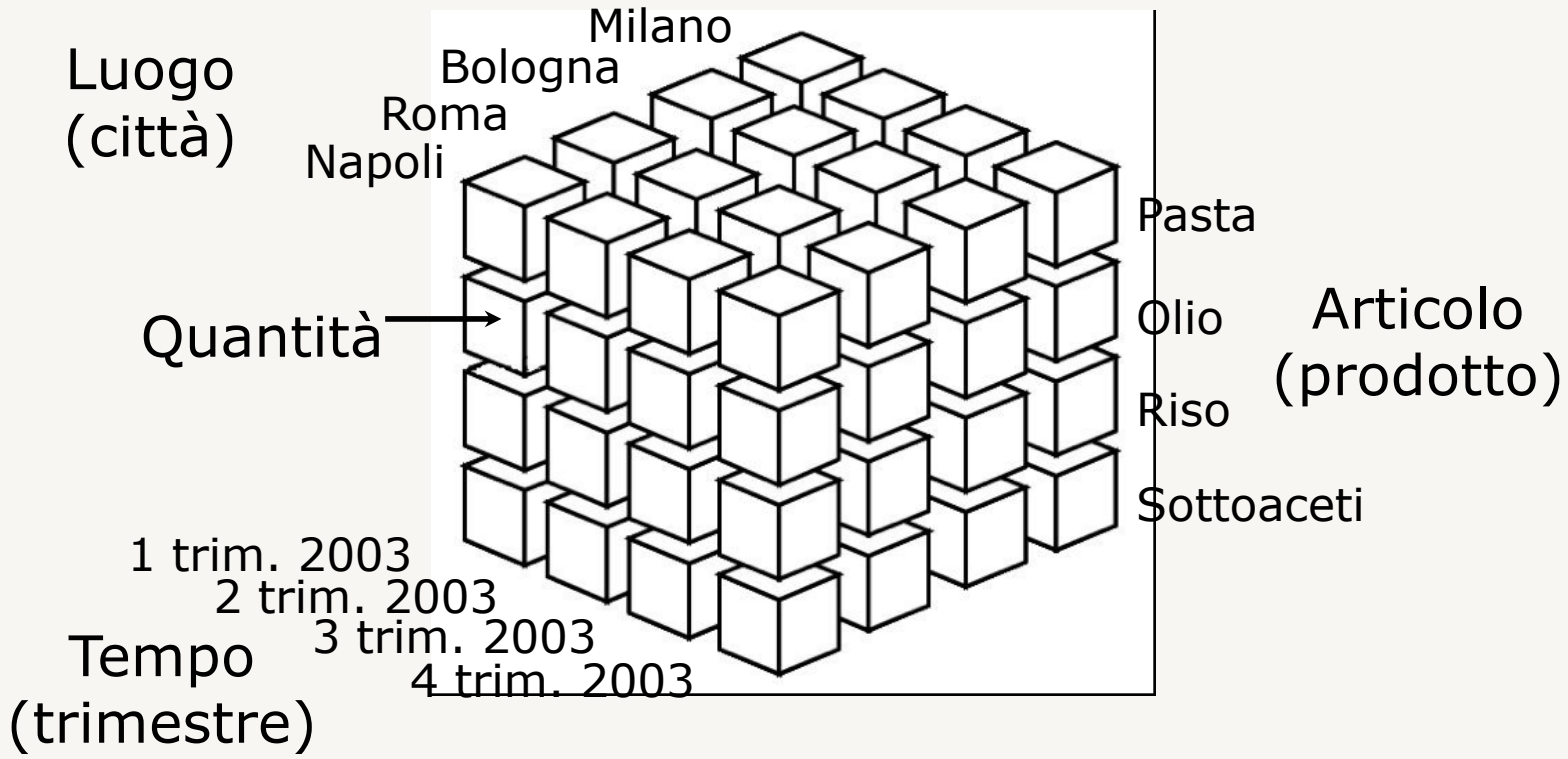
Dimensioni e gerarchie di livelli

Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione per i dati della dimensione stessa



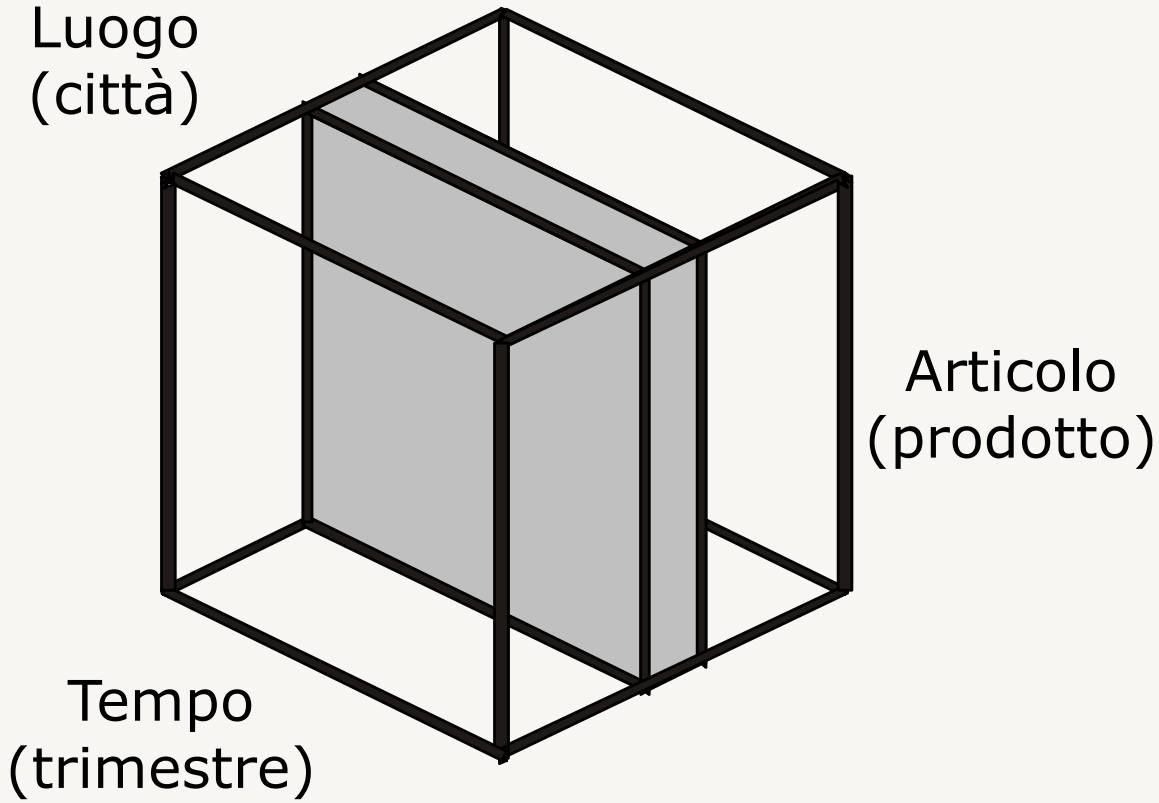
Modello multidimensionale: esempio (1)

- fatto: vendita
- misura: quantità
- dimensioni: articolo, tempo, luogo



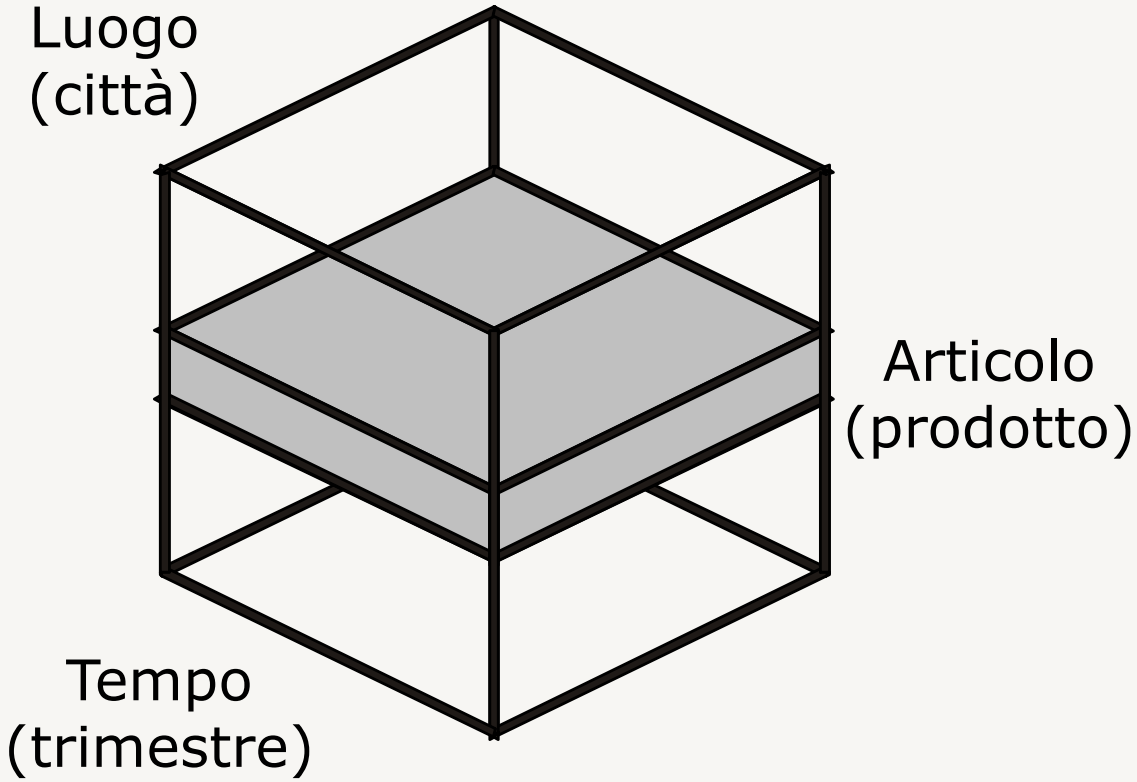
Modello multidimensionale: esempio (2)

Il manager di una città esamina la vendita dei prodotti in tutti i periodi relativamente alla città di sua competenza



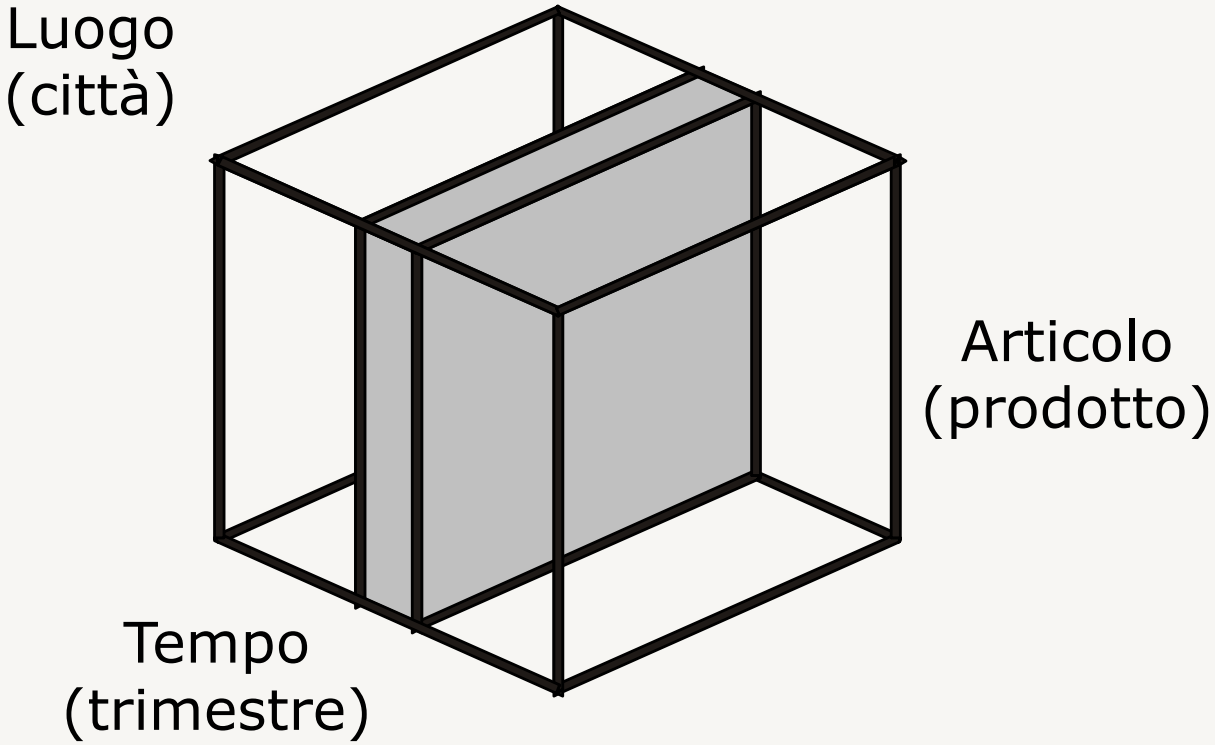
Modello multidimensionale: esempio (3)

Il manager di prodotto esamina la vendita di un prodotto in tutti i periodi e in tutte le città



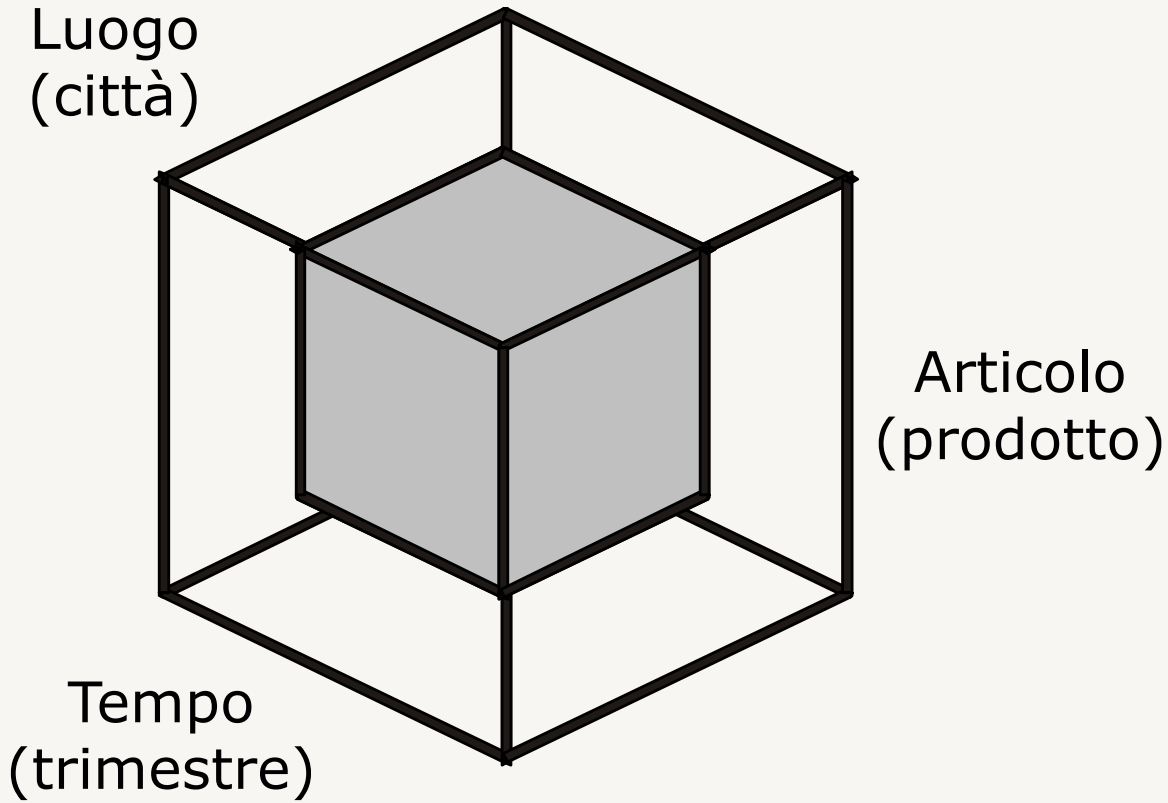
Modello multidimensionale: esempio (4)

Il manager finanziario esamina la vendita dei prodotti in tutte le città relativamente al periodo corrente e quello precedente



Modello multidimensionale: esempio (5)

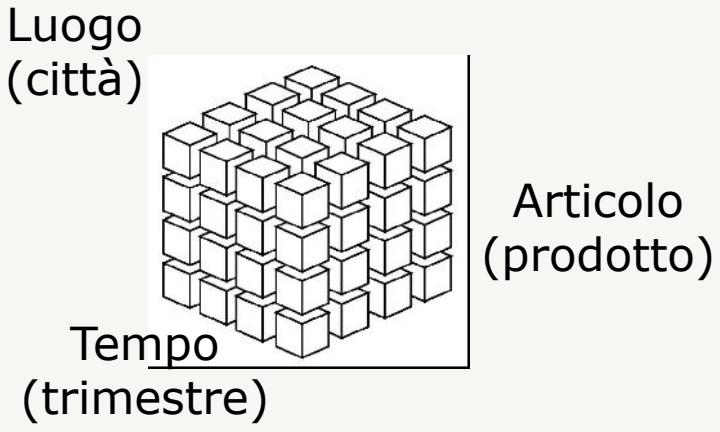
Il manager strategico si concentra su una categoria di prodotti, una area geografica e un orizzonte temporale medio



Operazioni su dati multidimensionali

- Slice-and-dice: seleziona e proietta
  - seleziona un sottoinsieme delle celle di un cubo
- Roll-up: aggrega i dati
  - applica una funzione aggregativa (tipicamente somma) sui dati aggregati di un cubo
  - è possibile effettuarla su misure additive (su cui ha senso fare la somma lungo una dimensione)
- Drill-down: disaggrega i dati
  - è l'operazione inversa del roll-up
  - aggiunge dettaglio ad un cubo disaggregandolo lungo una o più dimensioni

Slice-and-dice: esempio



seleziona le vendite di pasta per trimestre e città

Pasta	1 trim. 2003	2 trim. 2003	3 trim. 2003	4 trim. 2003
Milano	130000	125000	110000	145000
Bologna	125000	125000	135000	110000
Roma	80000	85000	90000	85000
Napoli	70000	75000	70000	90000

Drill-down: esempio

Pasta	1 trim. 2003	2 trim. 2003	3 trim. 2003	4 trim. 2003
Milano	130000	125000	110000	145000
Bologna	125000	125000	135000	110000
Roma	80000	85000	90000	85000
Napoli	70000	75000	70000	90000

Drill-down sul Luogo (da città a negozio)

Pasta	1 trim. 2003	2 trim. 2003	3 trim. 2003	4 trim. 2003
Milano-1	70000	65000	40000	75000
Milano-2	60000	60000	70000	70000
Bologna-1	60000	60000	55000	30000
Bologna-2	30000	35000	35000	40000
Bologna-3	35000	30000	45000	40000
Roma-1	40000	40000	45000	35000
…	…	…	…	…

Roll-up: esempio

Pasta	1 trim. 2003	2 trim. 2003	3 trim. 2003	4 trim. 2003
Milano	130000	125000	110000	145000
Bologna	125000	125000	135000	110000
Roma	80000	85000	90000	85000
Napoli	70000	75000	70000	90000

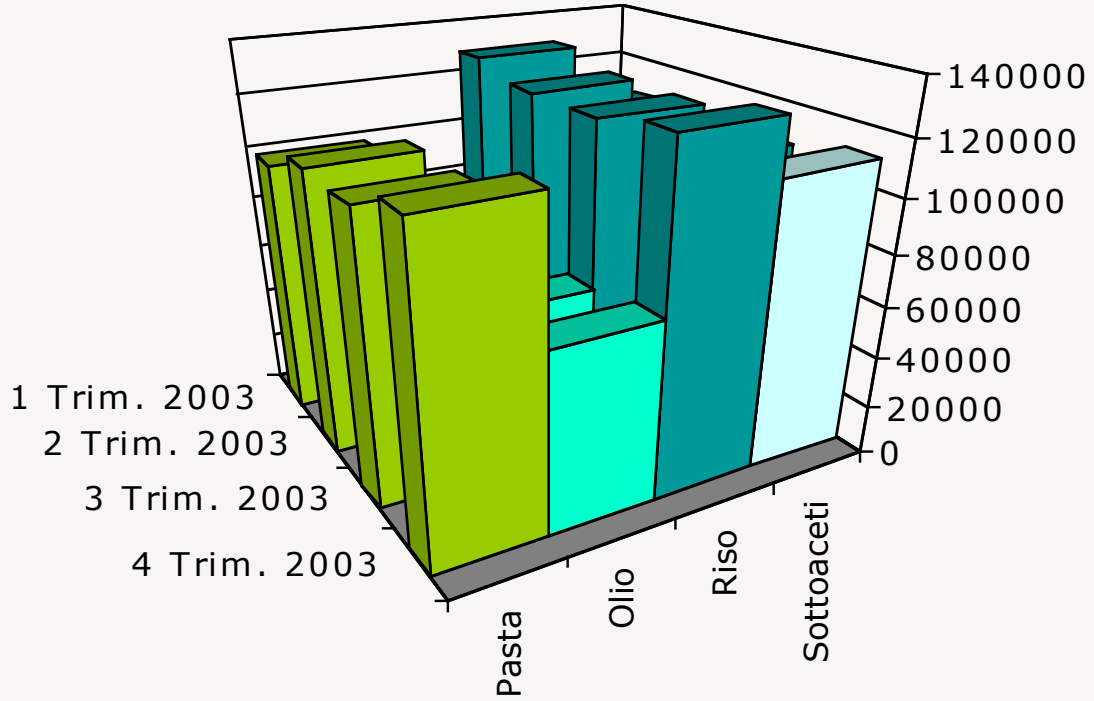
Roll-up sul Tempo (da trimestre a anno)

Pasta	2003
Milano	510000
Bologna	495000
Roma	340000
Napoli	305000

Visualizzazione dei dati

I dati vengono visualizzati in veste grafica per essere facilmente comprensibili

- tabelle, istogrammi, grafici, torte, superfici 3D, bolle, area in pila, ecc.



Realizzazione di DW: approcci

- ROLAP: Relational OLAP
  - utilizza tecnologia relazionale, opportunamente adattata e estesa
  - dati memorizzati in tabelle e operazioni di analisi espresse come istruzioni SQL
  - strutture di accesso ai dati speciali per ottimizzare le operazioni

- MOLAP: Multidimensional OLAP
  - memorizza i dati direttamente in forma multidimensionale
  - strutture dati tipicamente proprietarie

Relational OLAP

- generalmente costruito in modo incrementale, per collezione di data mart settoriali

- utilizza uno schema dimensionale o a stella
  - basato su modello concettuale a stella
    - restrizioni sulla struttura dello schema
  - dimensionale: contiene dimensioni di analisi
  - stella: struttura “stellare” dello schema

Schema a stella: componenti

Tabella dei fatti

- una tabella principale che memorizza i fatti del data mart

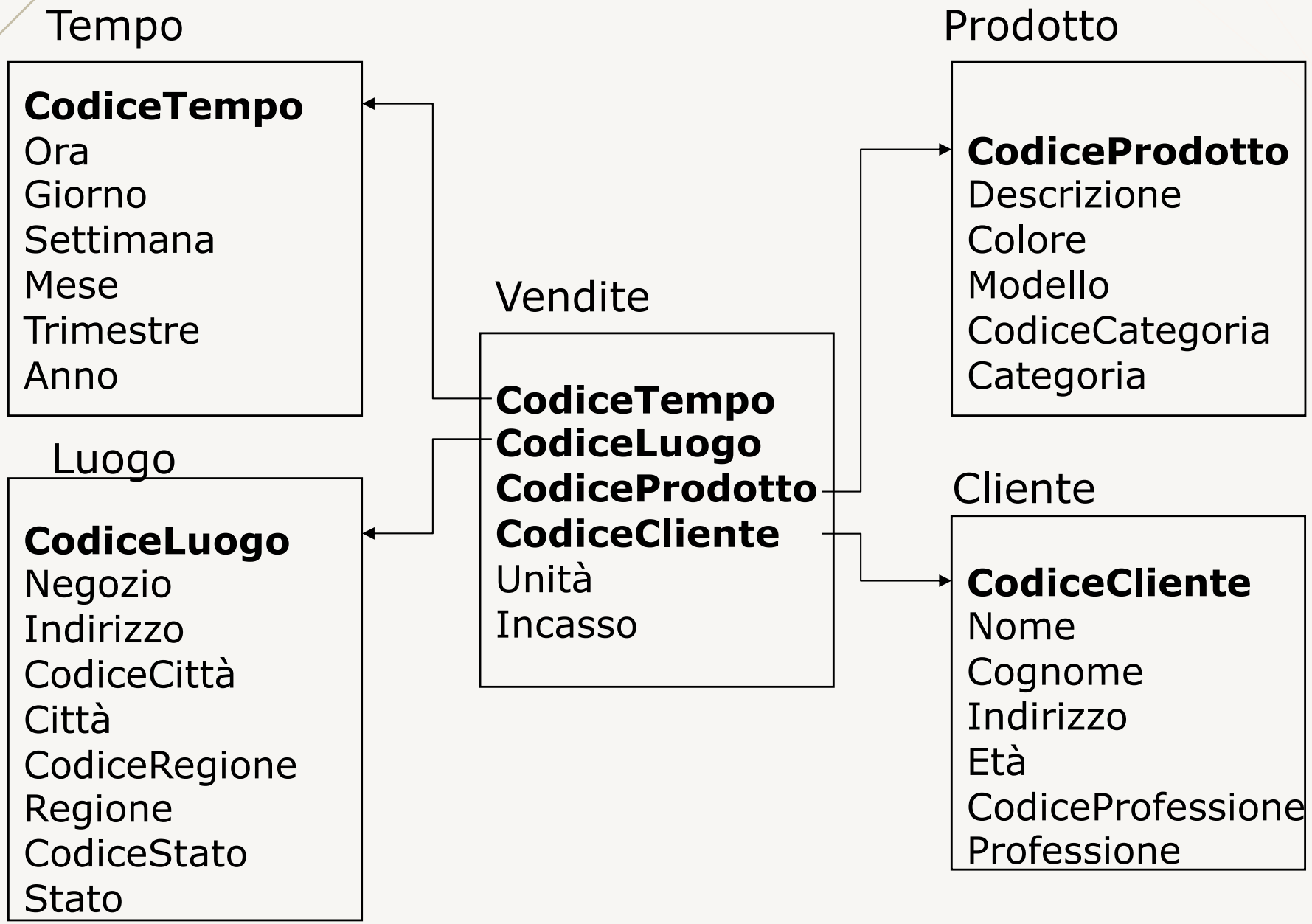
Tabelle dimensione

- relazioni ausiliarie che memorizzano i dati relativi alle dimensioni dell'analisi

Vincoli di integrità referenziale

- ognuno collega un attributo della tabella dei fatti a una tabella dimensione

Schema a stella: esempio



Schema a stella: caratteristiche

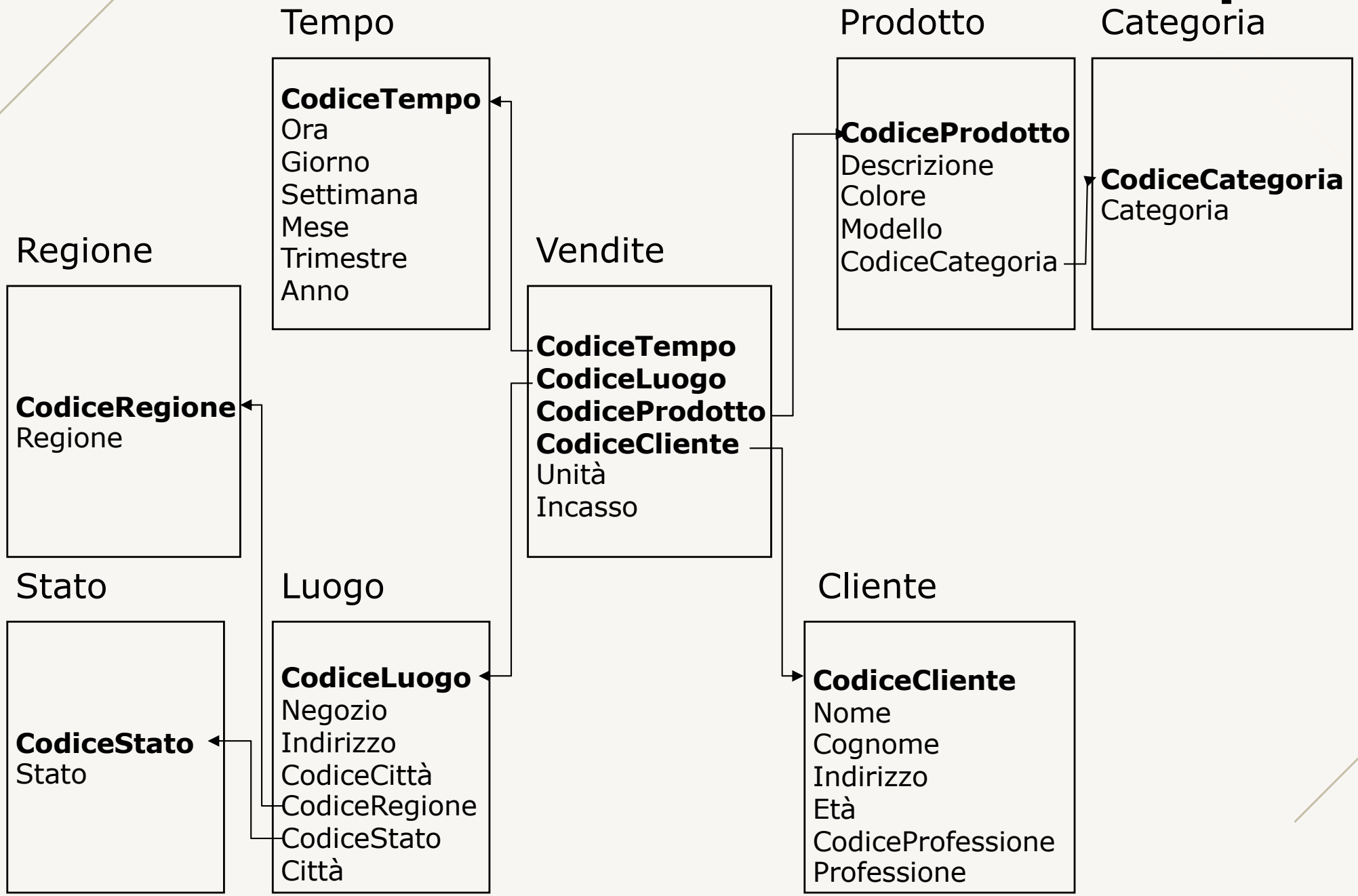
- tabella dei fatti
  - la chiave è composta da attributi che sono riferimenti alle chiavi delle tabelle dimensione
  - gli attributi non chiave rappresentano le misure del fatto e sono solitamente numerici
  - è in forma normale (Boyce-Codd)
- tabelle dimensione
  - chiave semplice (un solo attributo)
  - gli attributi non chiave rappresentano i livelli della dimensione o qualche loro proprietà e sono tipicamente testuali/descrittivi
  - generalmente non in forma normale
    - ridondanza ma maggiore efficienza

Schema a fiocco di neve (snow flake)

Risulta da una normalizzazione (anche parziale) di uno schema a stella

- permette di evitare ridondanze eccessive nelle dimensioni
- normalizzazione da fare con attenzione
  - porta degrado nelle prestazioni
- dalla tabella dei fatti si raggiungono tutte le tabelle delle dimensioni
  - muovendosi lungo associazioni n:1

Schema a fiocco di neve: esempio



Operazioni su ROLAP

- dati presentati all'utente con un modello di alto livello (multidimensionale)

- le interrogazioni sullo schema multidimensionale sono trasformate dal sistema in istruzioni SQL

Operazioni su ROLAP: esempio (1)

Roll-up

```
select D1.L1,…,Dn.Ln,Aggr1(F.M1),…,Aggrk(F.M1)
from Fatti as F,Dimens1 as D1,…,Dimensn as Dn
where Join-predicate(F,D1) …
      and Join-predicate(F,Dn)
      and selection-predicate
group by D1.L1,…, Dn.Ln
order by D1.L1,…, Dn.Ln
```

- `F.Mi`: misura *i*-esima della tabella dei fatti `F`
- `Di.Li`: livello della *i*-esima tabella dimensione
- `Aggri`: funzione aggregativa
- `Join-predicate(F,Di)` predicato di join fra `Fatti` e `Di`
- `selection-predicate`: condizione di selezione sulle tabelle dimensione

Operazioni su ROLAP: esempio (2)

Seleziona le vendite complessive del 2003 per categoria di articolo e trimestre

```
select P.Categoria, T.Trimestre, sum(V.Unità)
from Vendite as V, Prodotto as P, Tempo as T
where V.CodiceProdotto = P.CodiceProdotto
      and V.CodiceTempo = T.CodiceTempo
      and T.Anno = 2003
group by P.Categoria, T.Trimestre
order by P.Categoria, T.Trimestre
```

Aggregazione in SQL

Lo standard SQL offre operatori per aggregazioni

- esprime tutte le possibili aggregazioni delle tuple di una tabella
- utilizza il nuovo valore polimorfo ALL
  - presente in tutti i domini
  - corrisponde all'insieme di tutti i possibili valori del dominio
- opzioni:
  - with data cube: aggrega su tutte le possibili dimensioni
  - with roll up: aggrega in modo progressivo su una dimensione per volta
    - secondo ordine specificato

With cube: esempio (1)

Modello	Anno	Colore	Unità
fiat	1994	rosso	50
fiat	1995	rosso	85
ford	1994	rosso	80

```
select Modello, Anno, Colore, sum(Unità)
from Vendite as V, Prodotto as P, Tempo as T
where V.CodiceProdotto = P.CodiceProdotto
      and V.CodiceTempo = T.CodiceTempo
      and P.Modello in {'fiat','ford'}
      and P.Colore = 'rosso'
      and T.Anno between 1994 and 1995
group by Modello, Anno, Colore
with cube
```

With cube: esempio (2)

Modello	Anno	Colore	sum(Unità)
fiat	1994	rosso	50
fiat	1995	rosso	85
fiat	1994	ALL	50
fiat	1995	ALL	85
fiat	ALL	rosso	135
fiat	ALL	ALL	135
ford	1994	rosso	80
ford	1994	ALL	80
ford	ALL	rosso	80
ford	ALL	ALL	80
ALL	1994	rosso	130
ALL	1995	rosso	85
ALL	ALL	rosso	215
ALL	1994	ALL	130
ALL	1995	ALL	85
ALL	ALL	ALL	215

With roll up: esempio (1)

Modello	Anno	Colore	sum(Unità)
fiat	1994	rosso	50
fiat	1995	rosso	85
ford	1994	rosso	80

```
select Modello, Anno, Colore, sum(Unità) from
Vendite as V, Prodotto as P, Tempo as T
where V.CodiceProdotto = P.CodiceProdotto  and
V.CodiceTempo = T.CodiceTempo
      and P.Modello in {'fiat','ford'}
      and P.Colore = 'rosso'
      and T.Anno between 1994 and 1995
group by Modello, Anno, Colore
with roll up
```

With roll up: esempio (2)

Modello	Anno	Colore	sum(Unità)
fiat	1994	rosso	50
fiat	1995	rosso	85
ford	1994	rosso	80
fiat	1994	ALL	50
fiat	1995	ALL	85
ford	1994	ALL	80
fiat	ALL	ALL	135
ford	ALL	ALL	80
ALL	ALL	ALL	215

Ottimizzazioni

- indici bitmap
  - permettono di valutare in modo efficiente disgiunzioni o congiunzioni di predicati di selezione, oppure operazioni insiemistiche di unione e intersezione
- indici join
  - precomputano il join fra le tabelle dimensione e la tabella dei fatti
- materializzazione delle viste
  - vengono precalcolate le viste che servono per rispondere alle query più frequenti

Data mining

Tecniche di analisi per estrarre informazione non esplicita nei dati

- è una materia interdisciplinare:
  - statistica, algoritmi, intelligenza artificiale, reti neurali, …

Data mining: processo

Svolto in modo iterativo e adattativo con costruzione progressiva della conoscenza

- comprensione del dominio applicativo
- preparazione dell'insieme di dati
  - selezione del sottoinsieme dei dati della DW
  - possibile discretizzazione
- scoperta dei pattern
- valutazione dei pattern
- utilizzo dei risultati

Data mining: applicazioni

Regole di associazione

- ricercano regolarità nei dati

Regole di classificazione

- classificano nuovi fenomeni in classi predefinite

Regole di associazione

- ricercano regolarità nei dati
- struttura
  - premessa della regola
  - conseguenza della regola

Esempio

- pannolini → birra
  - il 30% delle transazioni che contiene pannolini contiene anche birra
  - il 2% tra tutte le transazioni contiene sia pannolini sia birra

Regole di associazione: proprietà

Supporto

- probabilità che siano presenti in una transazione entrambi gli elementi di una regola
  - il 2% tra tutte le transazioni contiene sia pannolini sia birra (supporto 0.02)

Confidenza

- probabilità che sia presente in una transazione la conseguenza di una regola, essendo presente la premessa
  - il 30% delle transazioni che contiene pannolini contiene anche birra (confidenza 0.30)

Regole di associazione: esempio

No.	Prodotto
1	Pasta
1	Ragù
2	Pasta
2	Acqua
3	Pasta
3	Acqua
3	Passata

Premessa	Conseguenza	Sup.	Conf.
Pasta	Ragù	0.33	0.33
Pasta	Acqua	0.66	0.66
Pasta	Passata	0.33	0.33
Ragù	Pasta	0.33	1
Acqua	Pasta	0.66	1
Acqua	Passata	0.33	0.5
Passata	Pasta	0.33	1
Passata	Acqua	0.33	1
{Pasta, Acqua}	Passata	0.33	0.5
{Pasta, Passata}	Acqua	0.33	1
{Acqua, Passata}	Pasta	0.33	1

Data mining per associazione

Estrarre tutte le regole con supporto e confidenza superiori a valori prefissati

- esempi di applicazione
  - analisi di mercato
    - prodotti acquisiti insieme o in sequenza
  - analisi di comportamento
    - individuare usi illeciti di credit card
  - previsione
    - prevedere il costo delle cure mediche
  - controllo
    - errori di produzione

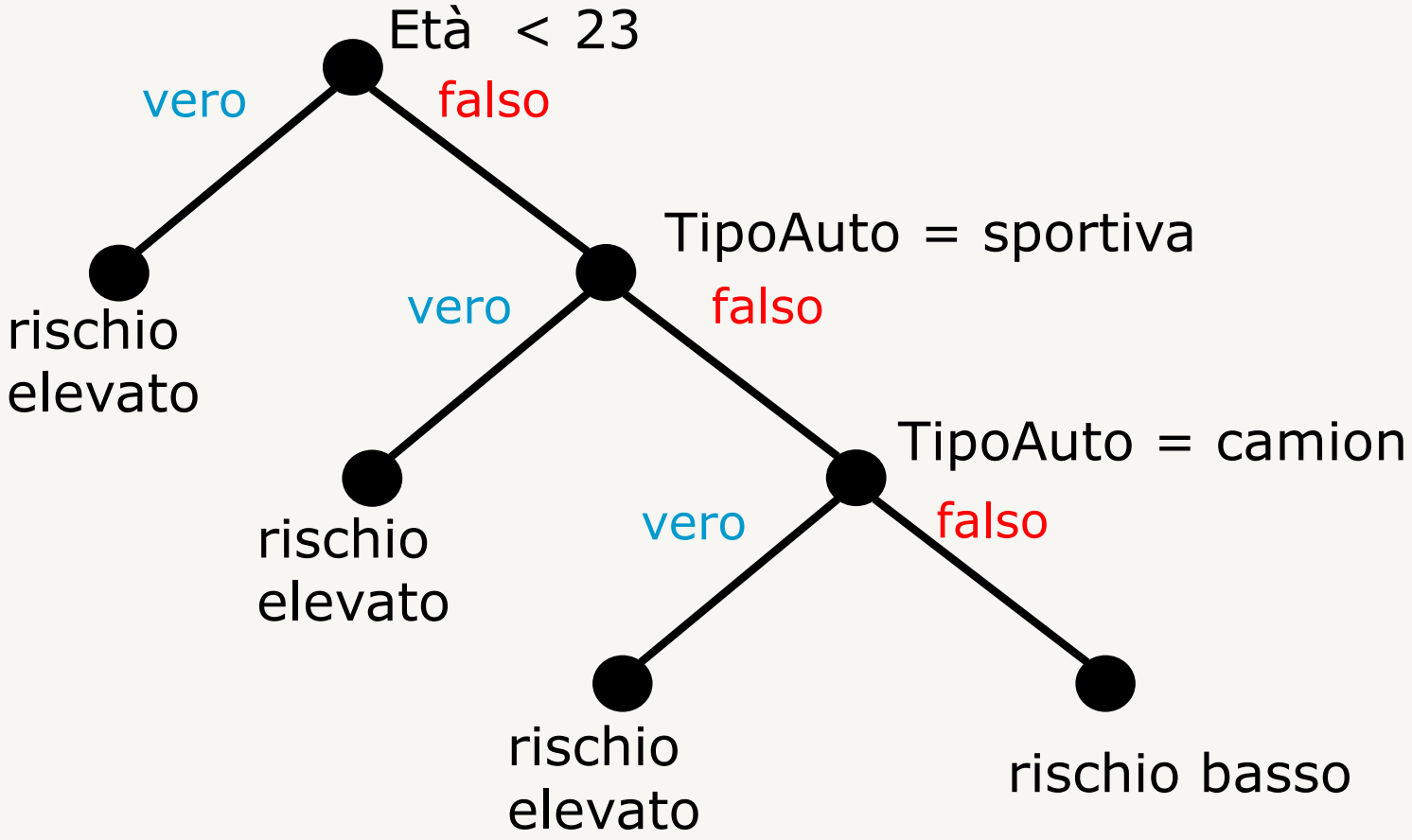
Classificazione

Catalogazione di un fenomeno particolare in una classe predefinita

- fenomeno da classificare, presentato sotto forma di fatto elementare (tupla)
- classificatore, algoritmo che svolgono la classificazione
  - costruito automaticamente tramite un insieme di dati di prova (training set)
  - applicato per la classificazione di fenomeni generici
  - rappresentato come albero di decisione

Classificazione: esempio

fenomeno: tupla di POLIZZA(NumPatente,Età,TipoAuto)



VINCENZO CALABRÒ

LinkedIn vincenzocalabro

www.vincenzocalabro.it