Analisi dei dati ed estrazione di conoscenza

Mastering Data Mining

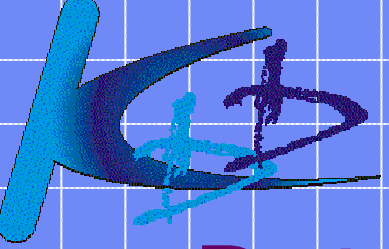
Fosca Giannotti

Pisa KDD Lab, ISTI-CNR & Univ. Pisa

http://www-kdd.isti.cnr.it/

DIPARTIMENTO DI INFORMATICA - Università di Pisa

anno accademico 2005/2006



Data Warehousing e Data Mining per la Business Intelligence:
una panoramica sulle idee e le applicazioni in ambito *retail*

Obiettivi del seminario

- Introdurre i concetti di base della business intelligence e del processo di estrazione di conoscenza.
- Fornire gli strumenti necessari per orientarsi tra le molteplici tecnologie coinvolte dall'analisi esplorativa dei dati (Data Warehousing ed OLAP) all'analisi previsionale (Data Mining).
- Comprendere le funzionalità e le soluzioni che è possibile aspettarsi in risposta ad esigenze nei diversi settori del *retail*, in particolare nel CRM, attraverso la discussione di alcuni casi di studio concreti.
- Capire quali sono le figure professionali coinvolte in un ambiente di business intelligence e quale è l'impatto organizzativo

Agenda del Seminario

- **Business Intelligence: cos'è, a quali esigenze risponde, come si colloca nell'organizzazione aziendale** ⇦
- B.I. ed estrazione di conoscenza dalle basi di dati – glossarietto minimo
- Esempi, casi di studio, buone pratiche di B.I. con strumenti di data warehouse
- Esempi, casi di studio, buone pratiche di B.I. con strumenti di data mining

"We are drowning in information, but starving for knowledge"

- Ogni organizzazione, pubblica o privata, raccoglie ogni giorno grandi quantità di dati
  - le tecnologie delle basi di dati e delle reti,
  - l'avvento del web,
  - la crescente capacità di memorizzazione
- Nei diversi contesti – la grande distribuzione, la medicina, la scienza o l'amministrazione pubblica – la mole di dati immagazzinati è **utile alla gestione, ma spesso non alle attività decisionali e strategiche**.

Sistemi informativi: tombe di dati o miniere di conoscenza?

- In queste **miniere di dati** giace, spesso nascosta, una ricchezza potenzialmente inestimabile in termini di **conoscenza strategica**.
- Queste **pepite** di informazione, se estratte, possono essere utilizzate
  - per aumentare efficacia ed efficienza dei processi,
  - per migliorare la qualità dei servizi,
  - per raggiungere un vantaggio competitivo.

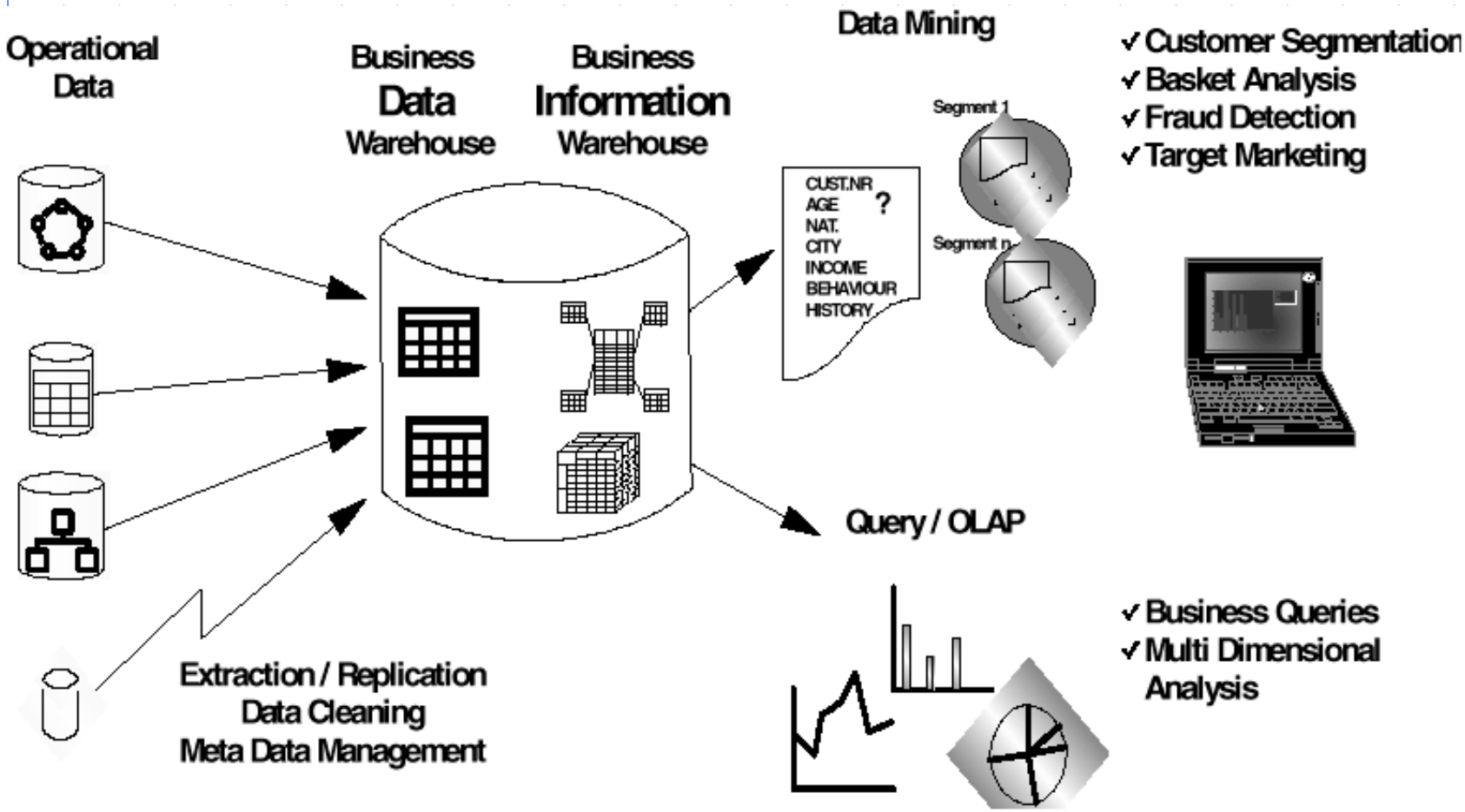
Business Intelligence: cos'è

- L'insieme delle tecnologie e dei processi che aiutano l'azienda a trasformare il proprio patrimonio informativo in conoscenza utile ai processi decisionali *(L'intelligence diventa business grazie all'informatica ANSA.it 14/6/05)*
- Un insieme di concetti, strumenti e metodologie volti a favorire i processi decisionali all'interno delle aziende (Gartner group 1989)

Business Intelligence: dove si colloca

- Tutte le aree funzionali
  - Vendite e Marketing, Amministrazione, Risorse umane, Servizi ai clienti, Relazioni con i fornitori
- Non solo per i livelli superiori della gerarchia organizzativa per la definizione di strategie aziendali, ma ..
- .. ricerca intelligente di dati, produzione e analisi dell'informazione appaiono ora fondamentali
  - per la produttività e l'efficienza di tutti i livelli di organizzazione aziendale.

La piattaforma BI



L’ambiente di BI: aspetti cruciali

- *Ampiezza*: integra funzioni e tecnologie da diversi comparti dell’azienda.
  - Mette insieme dati da ogni angolo dell’azienda.
- *Profondità*: raggiunge tutti quelli che ne hanno bisogno.
  - Servono interfacce appropriate e strumenti per utenti con necessità completamente diverse a tutti i livelli dell’organizzazione.

L’ambiente di business intelligence: aspetti cruciali

- *Completezza*: è una piattaforma integrata dall’inizio alla fine.
  - È una catena di applicazioni e tecnologie che lavorano su un insieme di dati comuni per creare una unica verità
- *Previsionale*: non solo osservazione dell’attuale, ma anche previsione basandosi su tecniche di analisi innovative.

Agenda del Seminario

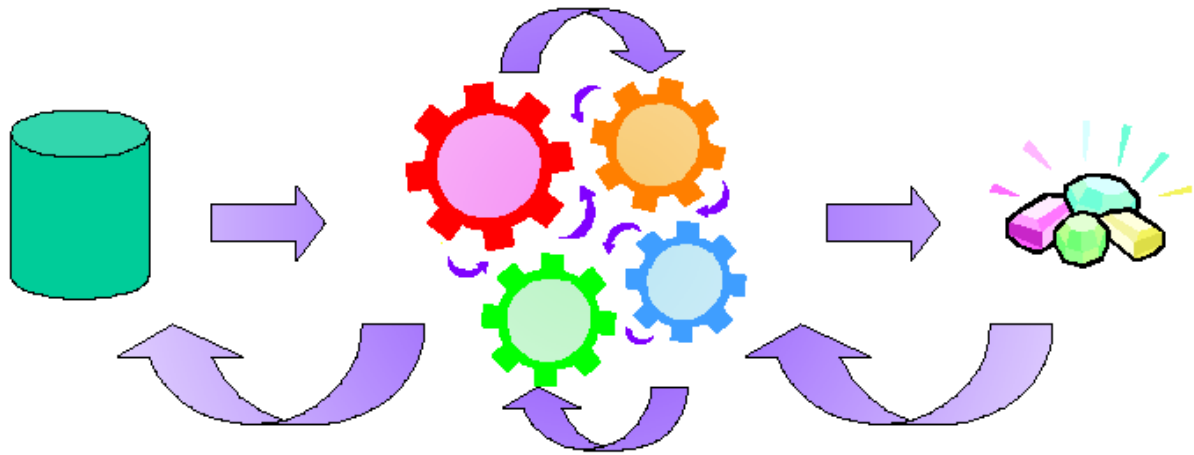
- Business Intelligence: cos'è, a quali esigenze risponde, come si colloca nell'organizzazione aziendale
- **B.I. ed estrazione di conoscenza dalle basi di dati – glossarietto minimo**
- Esempi, casi di studio, buone pratiche di B.I. con strumenti di data warehouse
- Esempi, casi di studio, buone pratiche di B.I. con strumenti di data mining

B.I. – glossarietto minimo

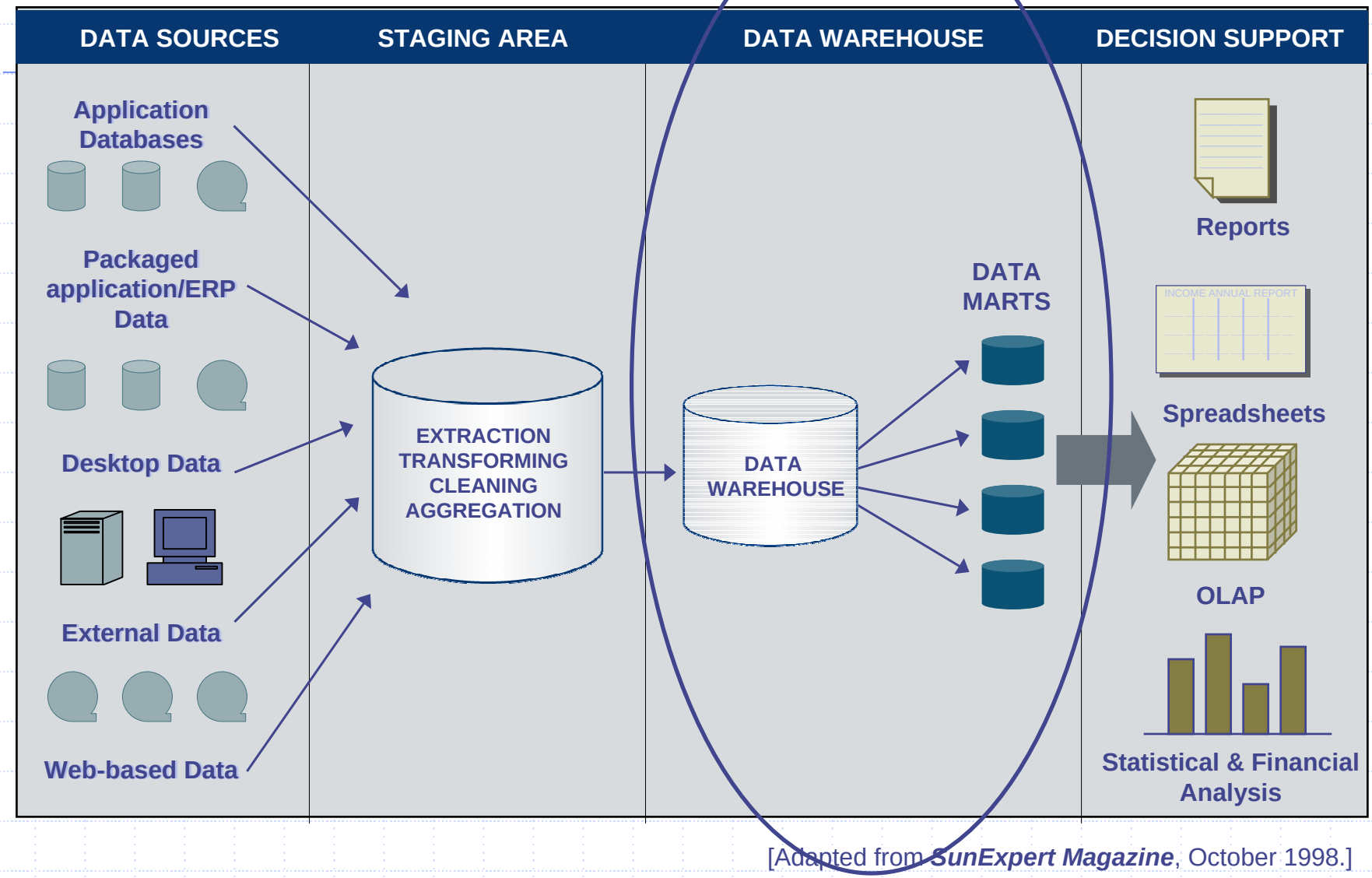
- Data Warehouse, Data Mart
- OLAP ed analisi multidimensionale
- Reportistica Avanzata.
- Dashboards – Cruscotti Aziendali
- Data mining – Strumenti Previsionali
- Applicazioni verticali

Il processo di BI in pratica

- E’ un processo di estrazione di conoscenza (KDD: Knowledge Discovery in Databases)
- KDD è un processo ITERATIVO
  - arte + ingegneria piuttosto che scienza



La base della BI



[Adapted from ***SunExpert Magazine***, October 1998.]

Cosa è il data warehouse

- Definito in molti modi
  - Un DB di supporto alle decisioni mantenuto separatamente dai DB operazionali dell'azienda.
  - Un processo di elaborazione dell'informazione che fornisce il basamento per dati consolidati e storici per l'analisi.

- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

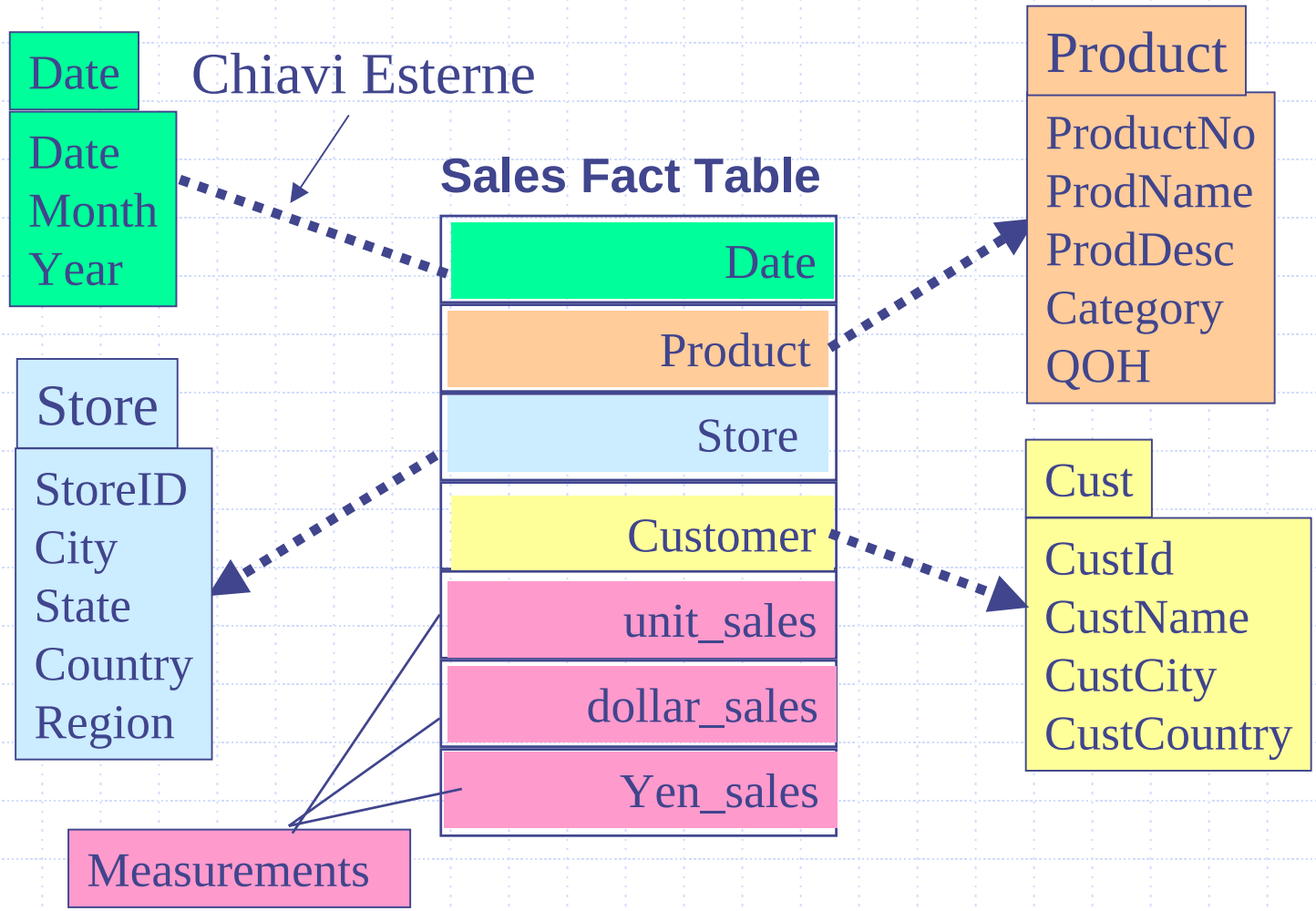
Data Mart

- Data warehouse che mette insieme i dati necessari ad una area funzionale
- Implementato creando *viste* specifiche alle applicazioni
- Viste materializzate dipartimentali che focalizzano su soggetti determinati:
  - Vendite e Marketing, Amministrazione, Risorse umane, Servizi ai clienti, Relazioni con i fornitori

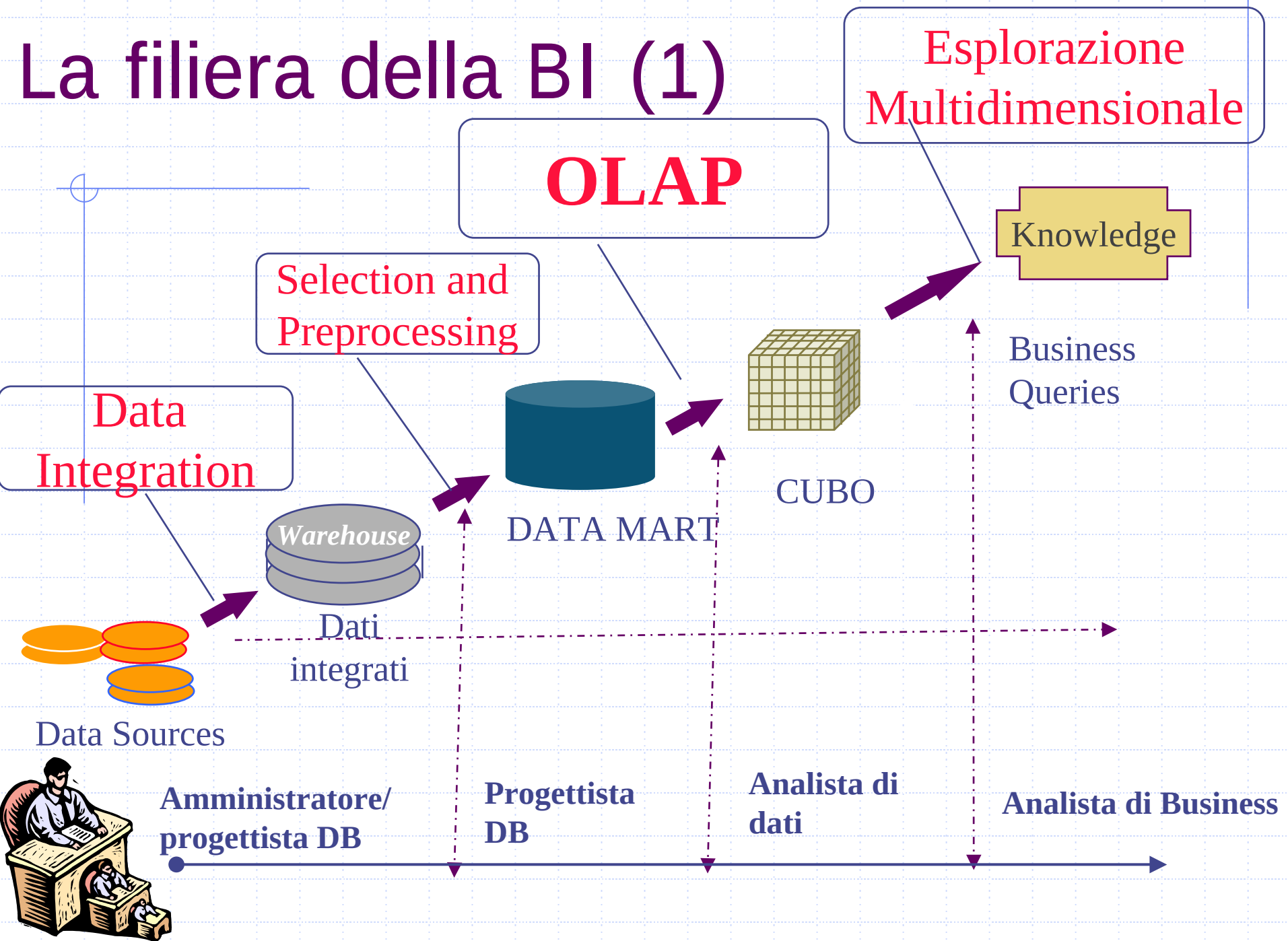
Il modello multi-dimensionale

- Un fatto è un evento di interesse per l'impresa (*vendite, spedizioni, acquisti*)
- Le misure sono attributi che descrivono quantitativamente il fatto (*unità vendute, prezzo unitario*)
- Una dimensione determina la granularità minima di rappresentazione dei fatti (*il prodotto, il negozio, la data*)
- Una gerarchia determina come le istanze di un fatto possono essere aggregate e selezionate - descrive una dimensione

Esempio di Star Schema



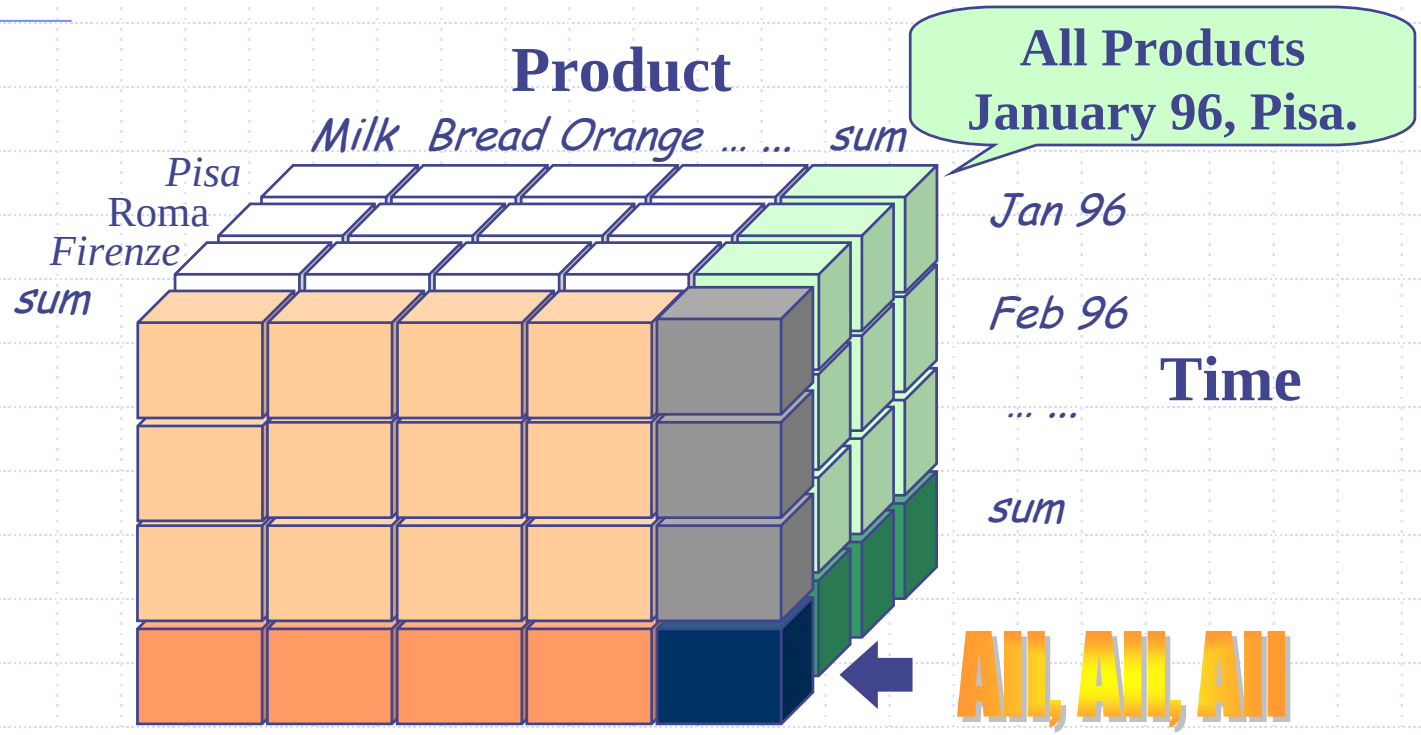
La filiera della BI (1)



OLAP e analisi multidensionale

- OLAP: On-Line Analytical Processing
- Analisi interattiva dei dati multi-dimensionali
- Le dimensioni definiscono la struttura della navigazione, ovvero i diversi punti di osservazione dei dati
- Le misure definiscono l'aspetto quantitativo dei dati osservati
- Le gerarchie sulle varie dimensioni definiscono il livello di granularità da cui si osservano le misure, quindi ricalcolandole secondo una adeguata aggregazione.

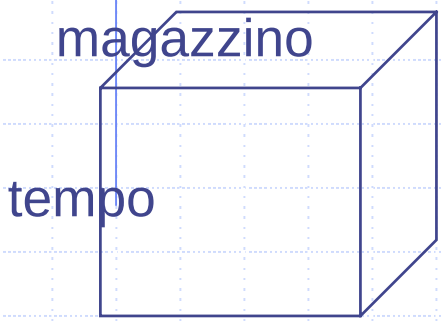
OLAP: Data Cubes

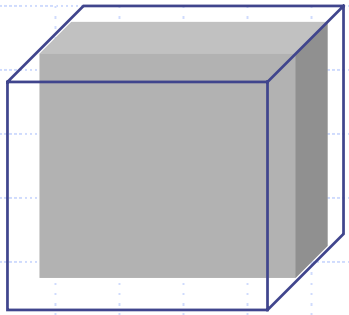


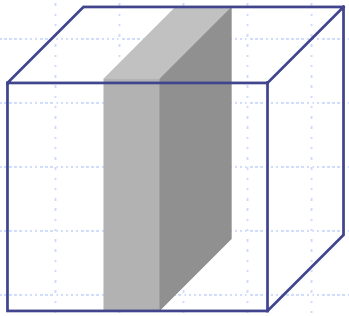
Ogni dimensione contiene una gerarchia di valori
una cella del cubo contiene valori aggregati
(count, sum, max, etc.)

OLAP

Il manager regionale esamina
la vendita dei prodotti
in tutti i periodi relativamente
ai propri mercati

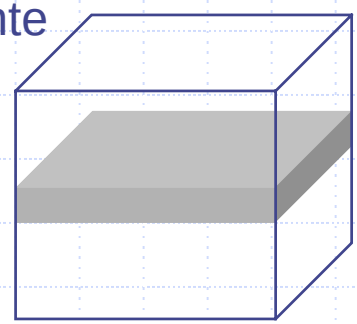


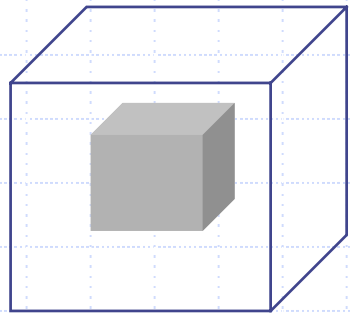




Il manager di prodotto esamina
la vendita di un prodotto
in tutti i periodo e in tutti i mercati

Il manager finanziario esamina la
vendita dei prodotti in tutti i mercati
relativamente al periodo corrente e
quello precedente



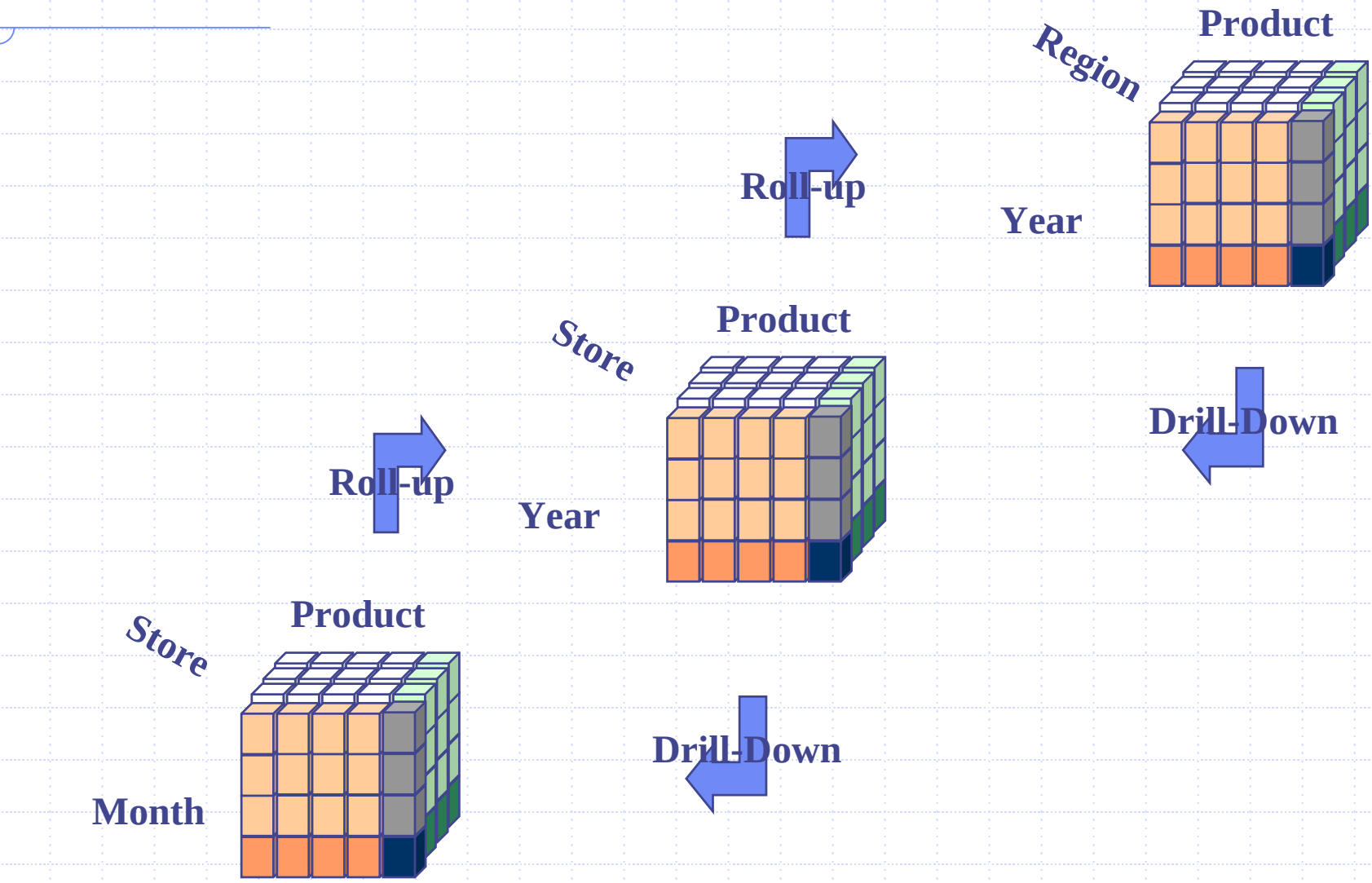


Il manager strategico si concentra su
una categoria di prodotti,
un'area regionale e un orizzonte
temporale medio

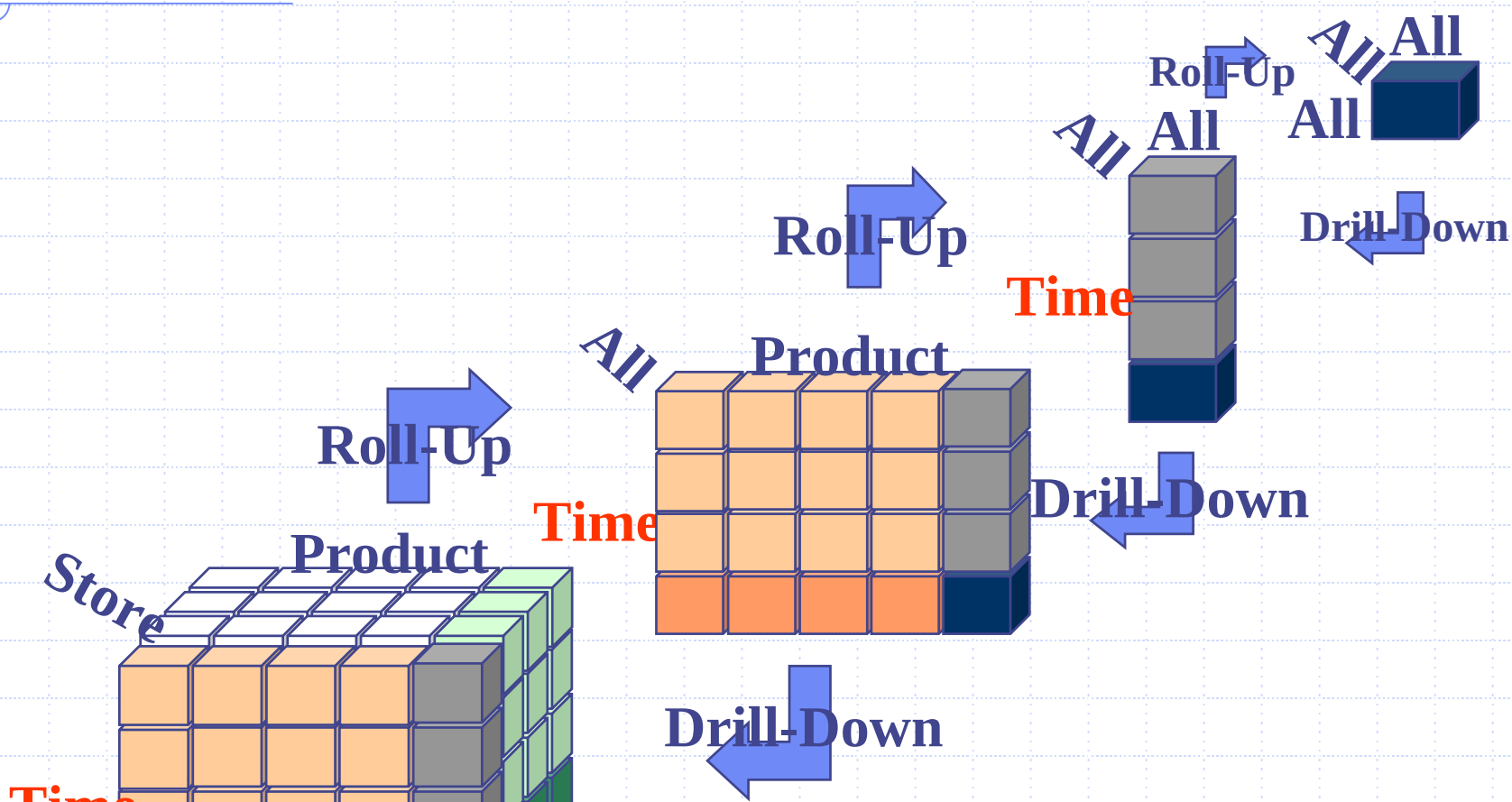
Operazioni tipiche

- Roll up: riassumi i dati
  - *il volume totale di vendite per categoria di prodotto e per regione*
- Roll down, drill down, drill through: passa da un livello di dettaglio basso ad un livello di dettaglio alto
  - *per un particolare prodotto, trova le vendite dettagliate per ogni venditore e per ogni data*
- Slice and dice: select & project
  - *Vendite delle bevande nel West negli ultimi 6 mesi*
- Pivot: riorganizza il cubo

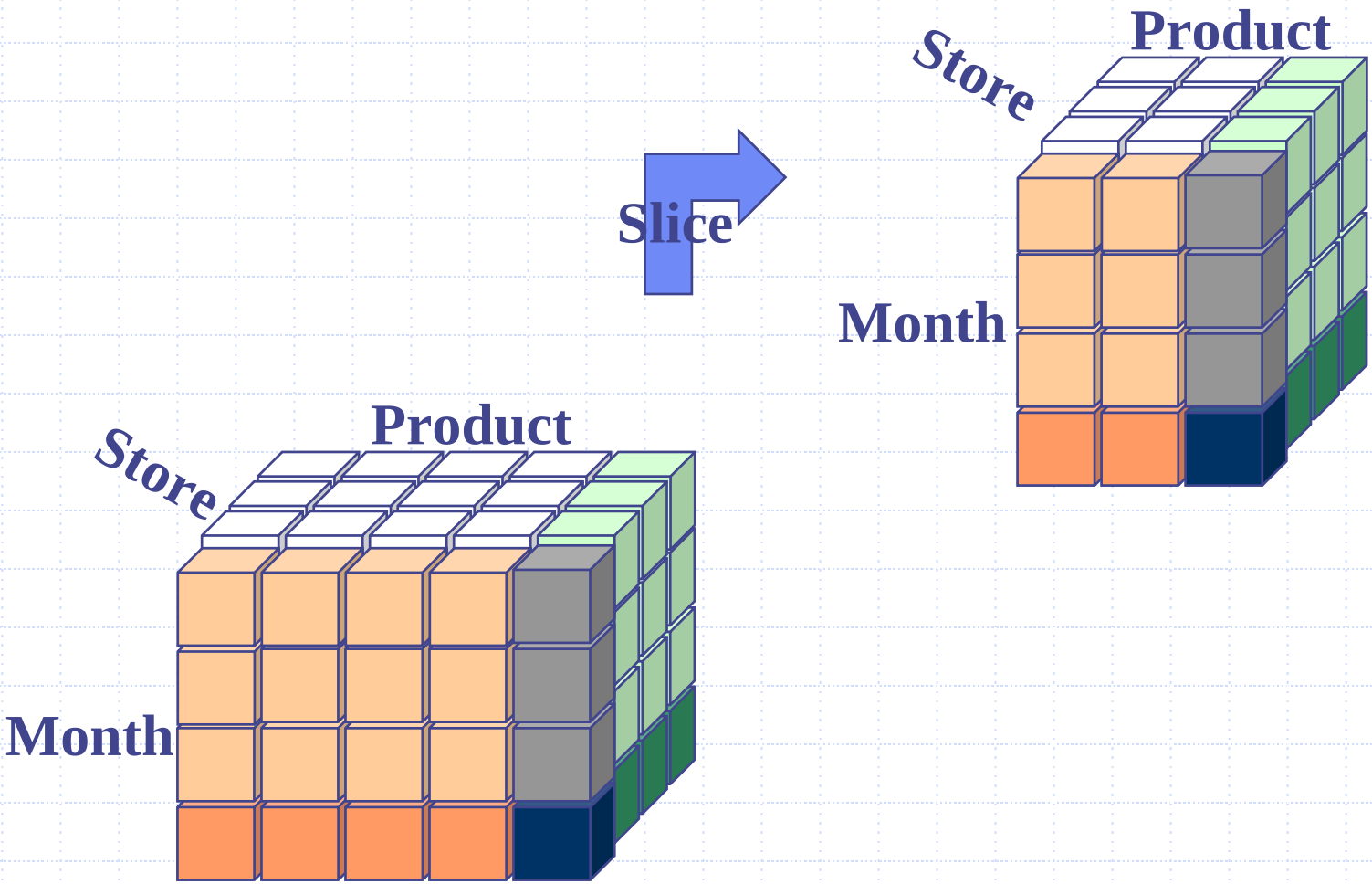
Operazioni tipiche: Roll-up



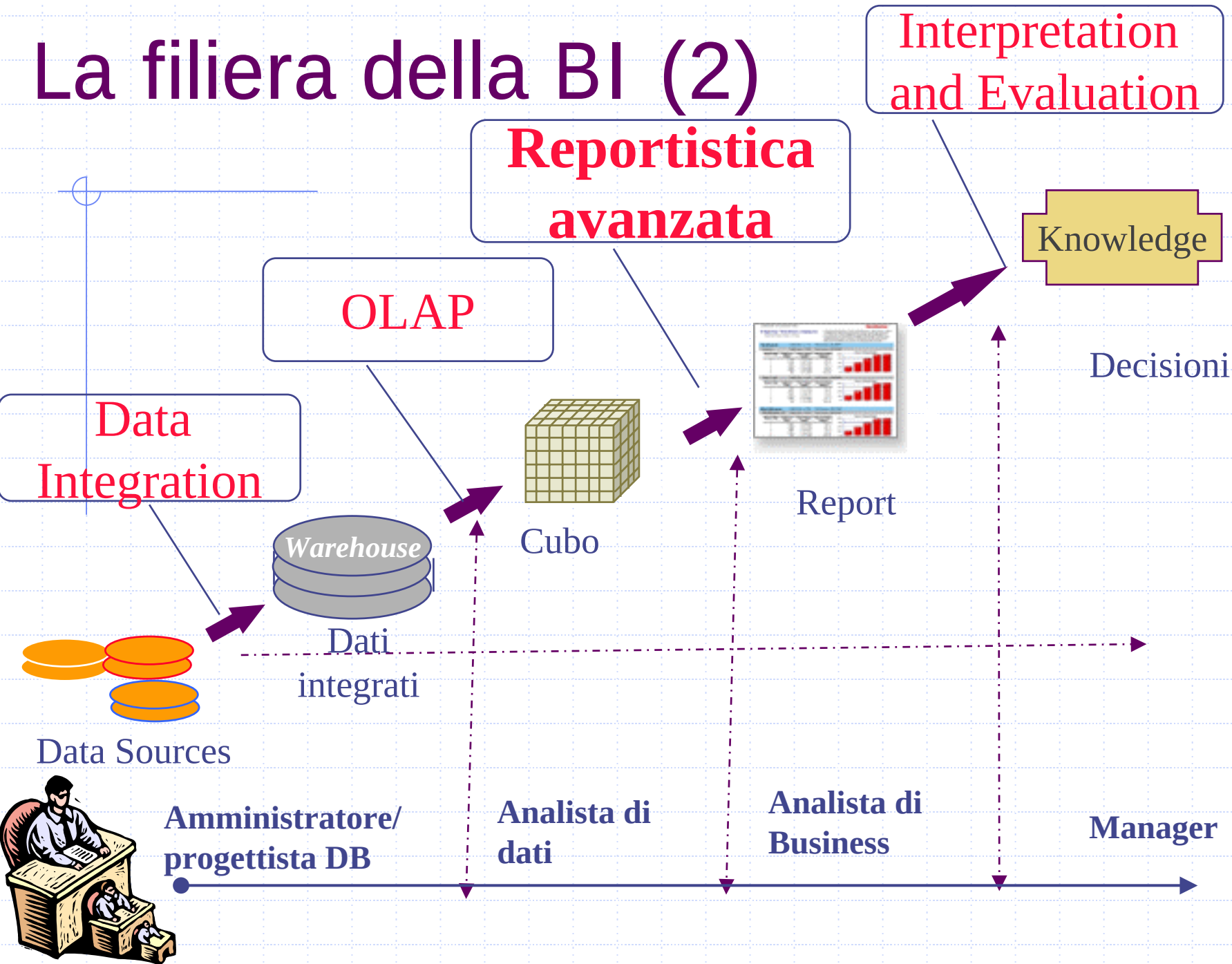
Operazioni tipiche: Roll-Up e Drill-Down



Operazioni tipiche: Slice and Dice

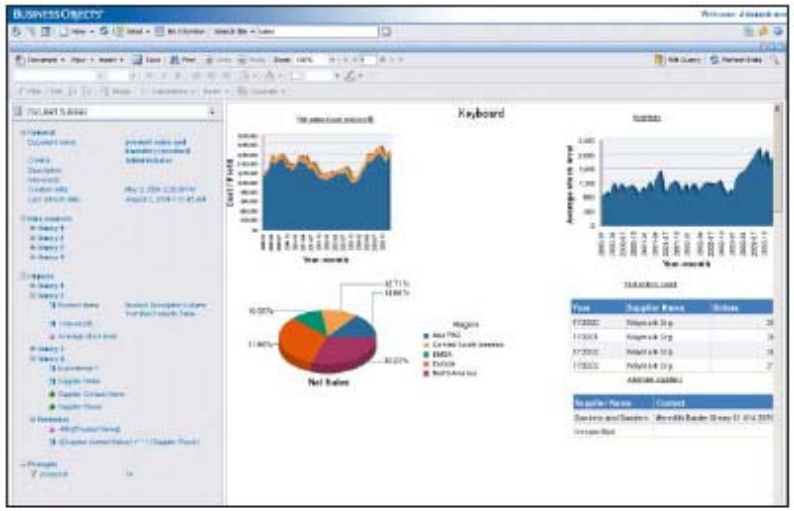


La filiera della BI (2)



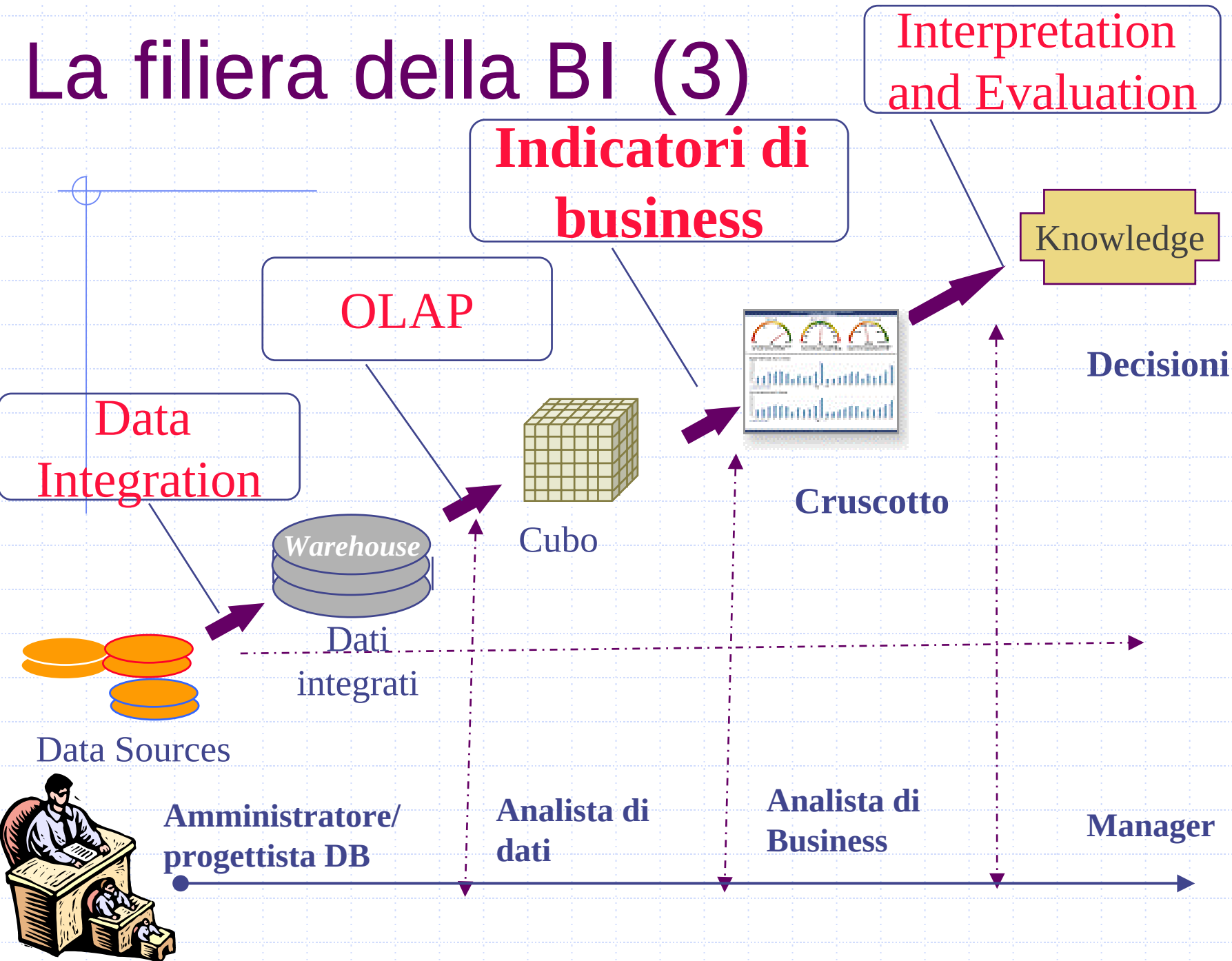
Reportistica avanzata

- Evoluzione del rapporto cartaceo. Es. Report di fidelizzazione COOP
- Sintesi in un rapporto di alcune navigazioni di un cubo. Es., per il manager di prodotto
- Il report è interattivo, secondo modalità preconfezionate di navigazione



Shield users from data complexity so they can focus on turning information into extreme insight.

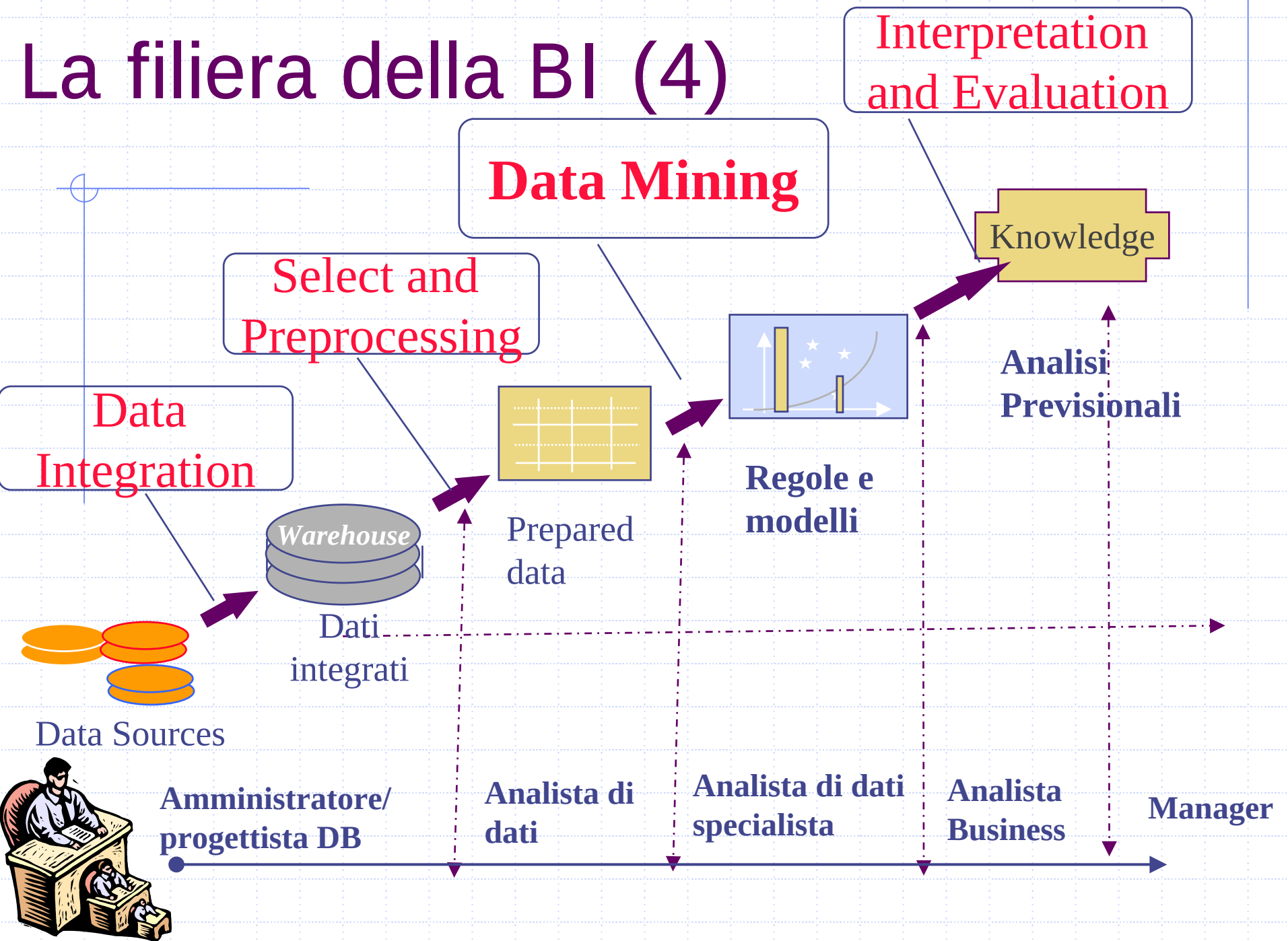
La filiera della BI (3)



Cruscotto/ Dashboard/ Scorecard

- Si utilizzano modelli specifici per il *retail* per definire indicatori interessanti
- Si confezionano con interfacce di immediato impatto visuale ed un insieme limitato di manopole di navigazione

La filiera della BI (4)

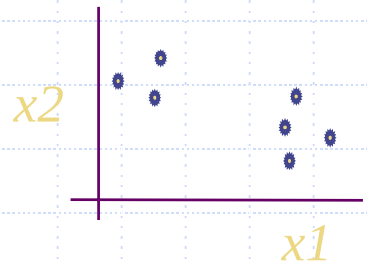


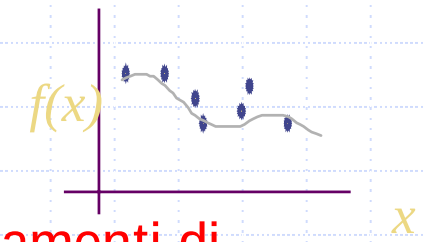
Dal data warehouse al data mining

- La complessità dei dati rende spesso difficile l'analisi dei dati coi metodi tradizionali
  - statistici
  - database, data warehouse
- È spesso impossibile prefigurare ipotesi da validare
- È spesso necessario lasciare che la conoscenza **emerga** dall'informazione grezza
  - Analisi previsionale ed esplorativa, analisi di trend

Modelli di Data Mining

- Esplorazione automatica/Discovery
  - *e.g.,* Scoperta di nuovi segmenti di mercato
  - clustering
- Predizione/Classificazione
  - *e.g..* Previsione delle vendite o della redemption
  - regressione, reti neurali, algoritmi genetici, alberi di decisione
- Spiegazione/Descrizione
  - *e.g..* Caratterizzazione di gruppi di clienti e comportamenti di acquisto
  - alberi di decisione, regole di associazione





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

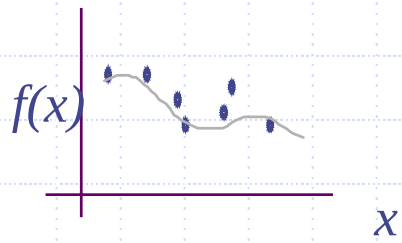
Previsione e classificazione

- Apprendimento di un modello predittivo a partire dai dati storici
- Uso del modello per la Classificazione di nuovi dati in cui la classe non è nota
- Molti metodi:
  - Reti neurali
  - Alberi di decisione
  - Algoritmi genetici
  - ...

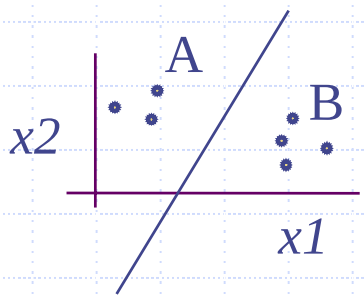
Induzione di modelli = apprendimento

Obbiettivo: *Costruire un modello generale o un'ipotesi a partire da esempi specifici*

- Regressione, stima del valore di una variabile numerica (es., il margine) sulla base dello storico



- Classificazione dei dati sulla base dei valori di una variabile categorica *target* presente nei dati storici (es., redento o no)

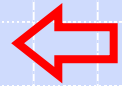


Applicazioni verticali che **possono** contenere mining

- Customer Retention
  - Identificare pattern che portano il cliente alla “defezione” (churn)
- Customer Service
  - Servizi di recommendation del prodotto
- Marketing
  - Targeting delle promozioni
  - Analisi di redemption per le promozioni
- Risk Assessment, Fraud Detection
  - Trovare pattern sospetti

Agenda del Seminario

- Business Intelligence: cos'è, a quali esigenze risponde, come si colloca nell'organizzazione aziendale
- B.I. ed estrazione di conoscenza dalle basi di dati – glossarietto minimo
- **Esempi, casi di studio, buone pratiche di B.I. con strumenti di data warehouse**
- **Esempi, casi di studio, buone pratiche di B.I. con strumenti di data mining**



Esempi di BI esplorativa

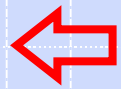
- navigazione OLAP
  - un cubo Sales (Vendite) attraverso interfacce usuali
- reportistica avanzata
  - Il report di fidelizzazione Unicoop
  - Analisi di redemption
- cruscotti
  - Performance aziendale

Esempi di BI previsionale

- Segmentazione clienti, una compagnia aerea
- Modelli predittivi di redemption nel retail
  - per l'ottimizzazione postalizzazione promozioni
- Rilevamento frodi fiscali
  - Ottimizzazione degli accertamenti
- Market Basket Analysis con dati di scontrino UniCOOP .

Agenda del Seminario

- Business Intelligence: cos'è, a quali esigenze risponde, come si colloca nell'organizzazione aziendale
- B.I. ed estrazione di conoscenza dalle basi di dati – glossarietto minimo
- **Esempi, casi di studio, buone pratiche di B.I. con strumenti di data warehouse**
- Esempi, casi di studio, buone pratiche di B.I. con strumenti di data mining



Navigazione OLAP sul cubo delle Vendite

- Demo di strumenti di navigazione
  - Basati su Excel con tabelle pivot o simili
  - Basati su grafici con pulsanti di navigazione
- Dietro le quinte: collegamento con il server OLAP (cubo Sales)

Reportistica avanzata: un esempio fatto in casa

- Per capirci, consideriamo il **report di fidelizzazione** prodotto dal settore marketing di **Unicoop Tirreno** ogni quadrimestre.
- Analizza la classificazione dei soci nei diversi negozi della rete.

Classificazione dei soci

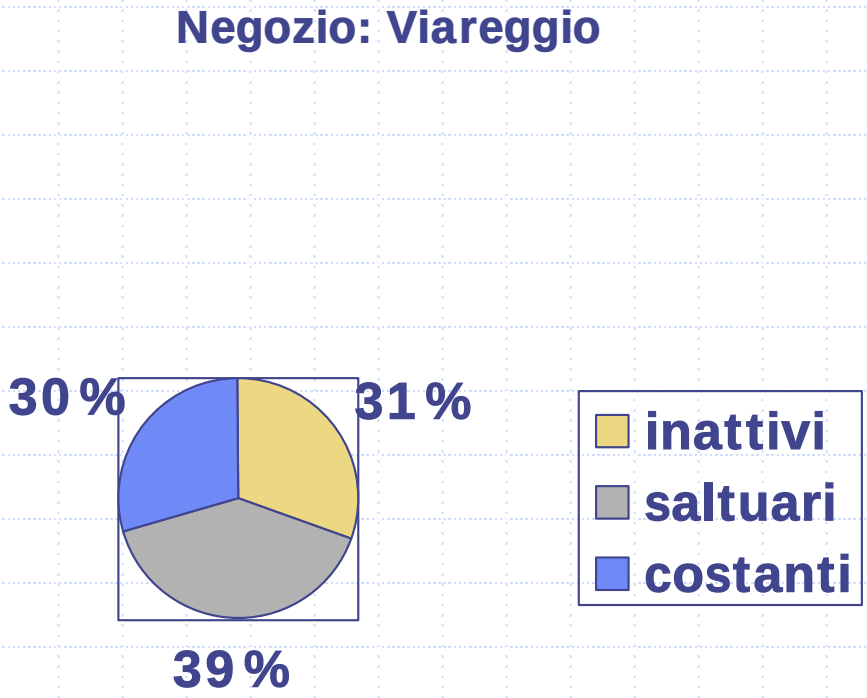
- **Costanti**: negli ultimi 4 mesi hanno fatto almeno 2 spese al mese per almeno 3 mesi su 4
- **Saltuari**: negli ultimi 4 mesi hanno fatto la spesa, ma non raggiungono la soglia dei costanti
- **Inattivi**: negli ultimi 4 mesi non hanno mai fatto la spesa

Soci costanti

- Un socio costante è classificato come **completo** per un certo reparto (es. ortofrutta) se acquista in tale reparto con una frequenza superiore ad una soglia stabilita (specifica del reparto)
- I soci costanti sono classificati in 5 **classi di spesa**
- Si tiene traccia dell'incidenza dei reparti freschi sul totale della spesa alimentare

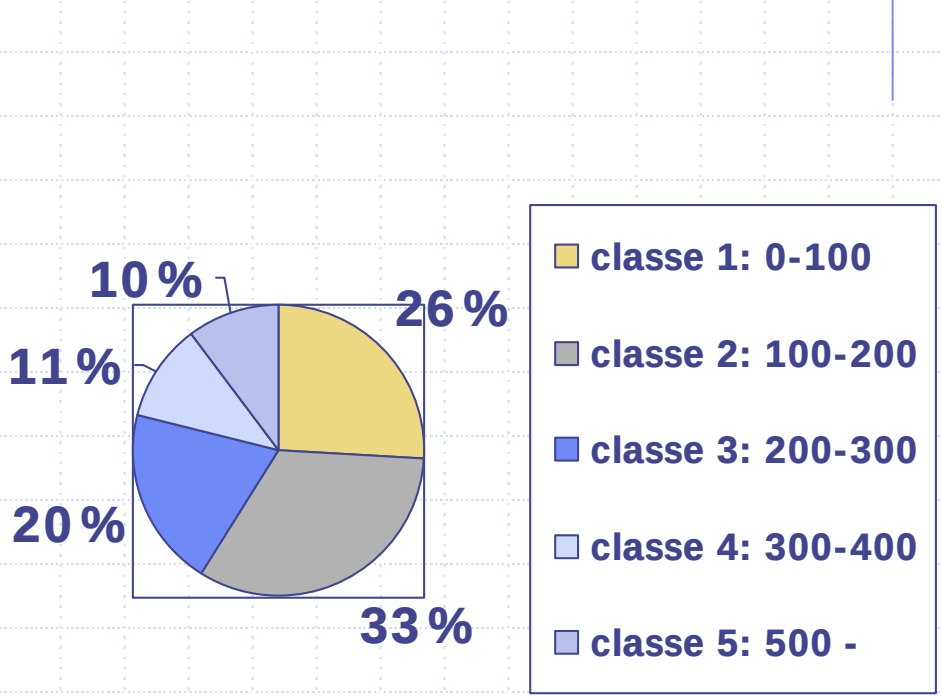
Report di fidelizzazione

- Riporta la classificazione dei soci
  - per ogni area geografica
  - per ogni negozio della rete



Report di fidelizzazione

- Riporta la suddivisione dei soci costanti in classi di spesa
  - per ogni area geografica
  - per ogni negozio della rete



Report di fidelizzazione

- Riporta la percentuale dei soci costanti con spesa completa
  - per ogni reparto
  - per ogni area geografica
  - per ogni negozio della rete
- Riporta l'incidenza dei freschi sulla spesa alimentare dei soci costanti
  - per ogni reparto
  - per ogni area geografica
  - per ogni negozio della rete

... complessivamente

- Circa 30 pagine
- Anche se esauriente, è certo possibile esplorare gli stessi dati da altre dimensioni:
  - Diversi intervalli temporali, tendenze, raffronti non solo con la volta precedente, diverse aggregazioni geografiche, diverse aggregazioni sui soci
- Richiede un lavoro non trascurabile, integrando a mano molte interrogazioni sul database ed analizzandone i risultati
- Sarebbe forse desiderabile poterlo ripetere con maggiore frequenza

Dalla carta al report avanzato

- Gli strumenti di BI consentono di creare un **report di fidelizzazione interattivo**, con una struttura analoga a quello di carta, ma navigabile sulle dimensioni:
  - Classificazione dei soci
    - navigabile sulla dimensione geografica e temporale
  - Classe di spesa dei soci costanti
    - navigabile sulla dimensione geografica e temporale
  - Percentuale dei soci costanti con spesa completa
    - navigabile sulla dimensione geografica, temporale e dei reparti, ma anche delle classi di spesa dei soci (analogamente per l'incidenza dei freschi)

Report interattivo

- Il **report di fidelizzazione interattivo** può essere prodotto a partire da un cubo delle vendite disponibile nel data warehouse
- è una interfaccia intelligente verso quei dati, aggregati al fine di avere un quadro dell'andamento della fidelizzazione dei soci
- Una volta disegnato, può quindi essere **ricalcolato**, quando desiderato, in funzione dei nuovi dati via via disponibili nel DW

Report interattivo

- Può essere **distribuito** alle diverse figure interessate
  - per via elettronica (web, Excel, o anche cartacea)
  - anche in forme differenziate: il manager di negozio vede il rapporto solo a livello di negozio, il manager regionale anche a livello regionale, ...
- Il personale del servizio marketing può essere scaricato di una parte routinaria del proprio lavoro ed assolvere alla funzione di disegnare nuovi report sempre più raffinati e rispondenti alle esigenze degli utenti.

... dice il saggio:

- Le organizzazioni complesse hanno una naturale propensione a creare conoscenza e a diffonderla al proprio interno per assolvere meglio alle proprie funzioni
  - ... altrimenti Unicoop non sentirebbe il bisogno di produrre un report di fidelizzazione
- Spesso però questo è un lavoro faticoso, episodico, non valorizzato come strategico
- Il messaggio autentico della BI è: **creare le condizioni perché il management della conoscenza faccia sistema**

Esempio: analisi di redemption

- **Dati sorgente**: scontrini di vendita con registrazione delle promozioni “redente”
- **Data mart**: acquisti dei clienti con indicazione di promo, clienti postalizzati, clienti redenti
- **Obiettivo di analisi**: valutare l’efficacia delle campagne promozionali
- **Esempio di report**: confronto fra le diverse promozioni/campagne rispetto al rapporto fra clienti (soci) postalizzati e clienti che rispondono alla promo (redenti)

Elementi di un report per l'analisi di redemption

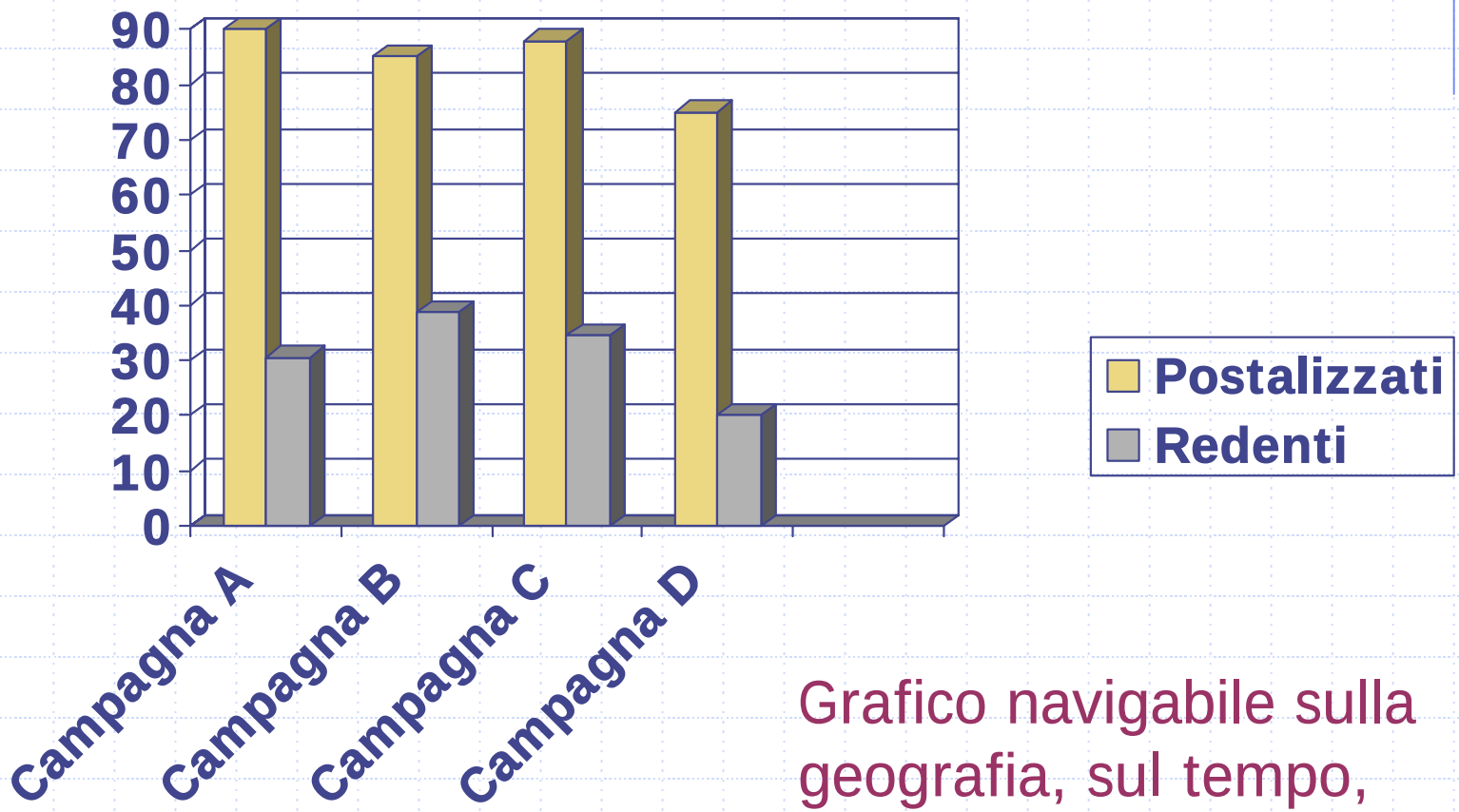
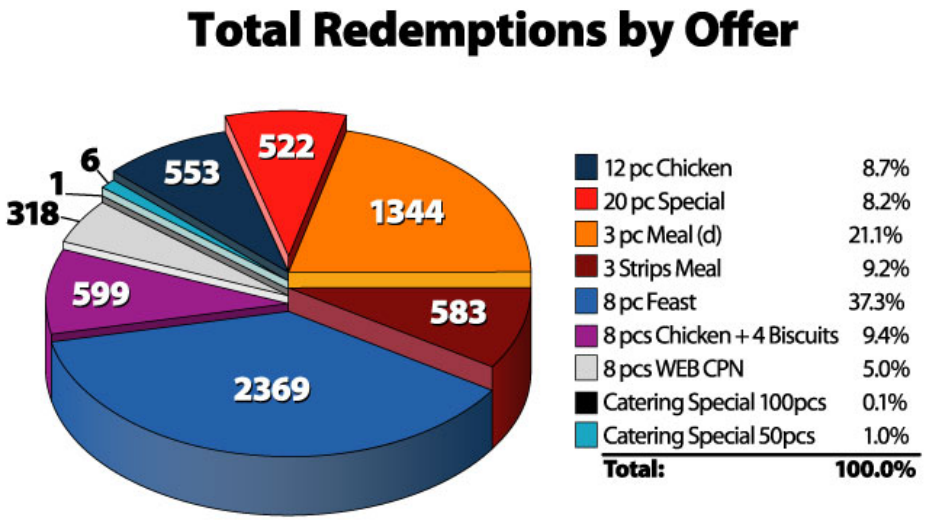


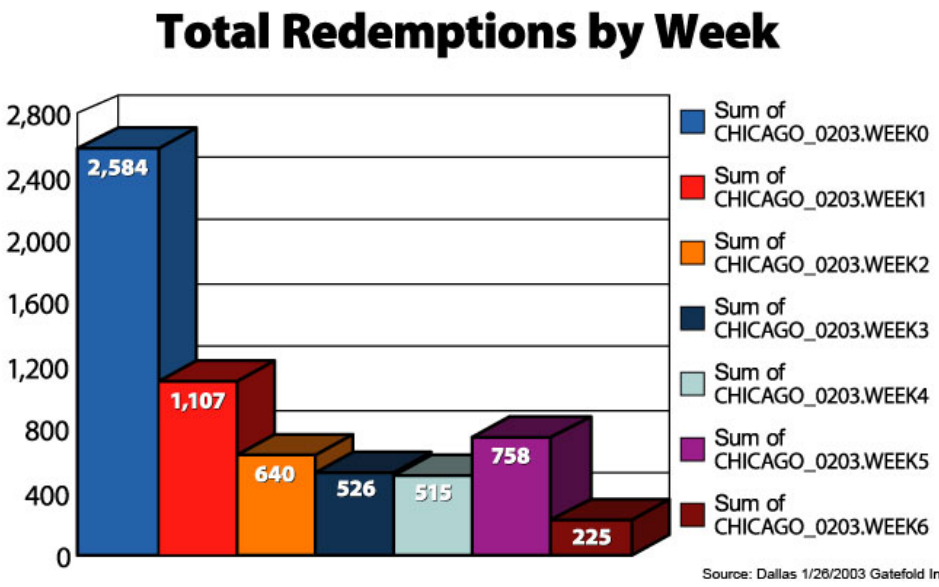
Grafico navigabile sulla geografia, sul tempo, sul tipo di promo, ...

Elementi di un report per l'analisi di redemption

Dimensione: tempo



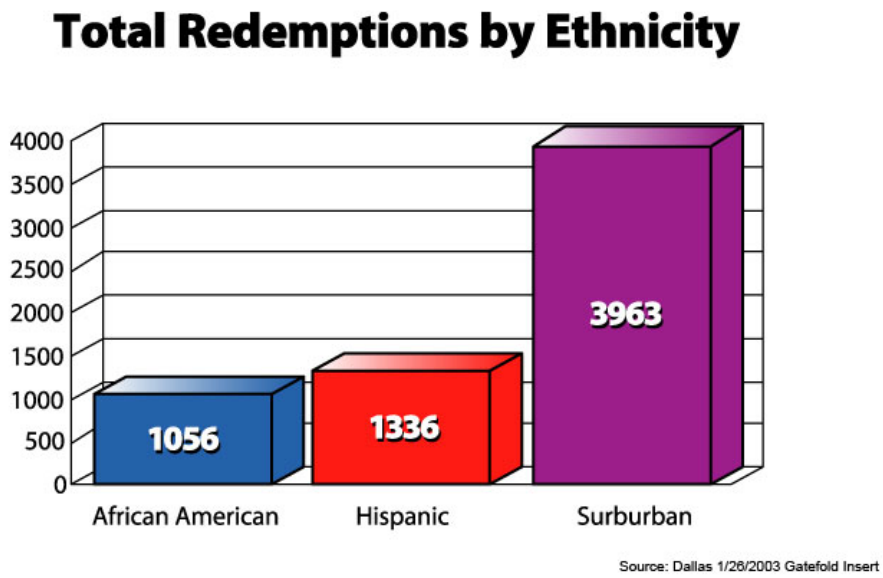
Source: Dallas 1/26/2003 Gatefold

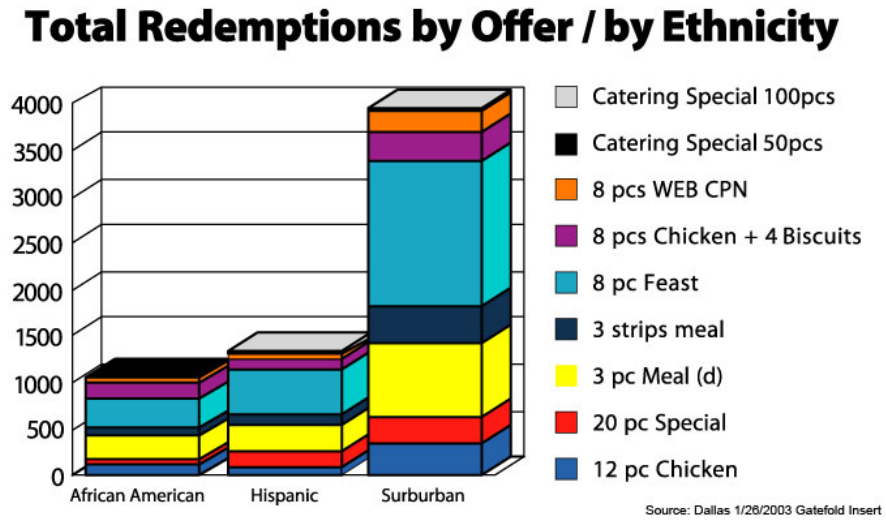


Source: Dallas 1/26/2003 Gatefold Insert

Dimensione tipo promo

Elementi di un report per l'analisi di redemption





Dimensione: demografica

Dimensione: incrocio promo/demo

Cruscotti aziendali

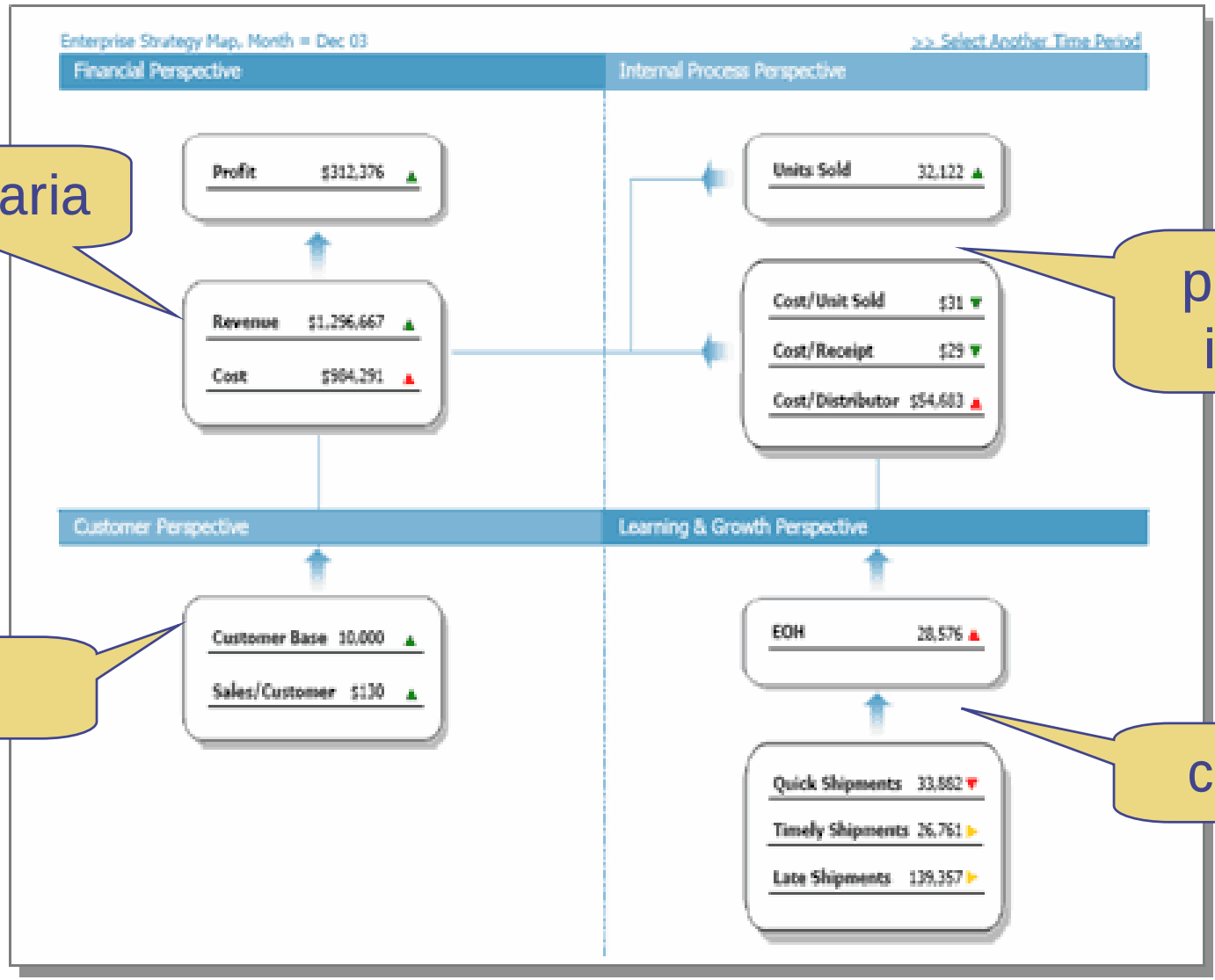
- Le alte figure direzionali (commerciale, marketing, risorse umane, finanziario, ...) e i decision-makers hanno bisogno di rapporti
  - molto sintetici e di rapido impatto
  - aggiornati alla situazione corrente
  - flessibili
- in grado di mettere in luce in estrema sintesi gli **indicatori chiave** della **performance aziendale**
  - Key Performance Indicators

Cruscotti aziendali

- I cruscotti (**dashboard**, **scorecards**) sono finalizzati a comunicare lo stato del business e monitorare l'andamento progressivo
- Forte impatto visuale
  - Una figura vale mille parole ...
  - Layout grafico a zone
- Confronto su diverse (poche) dimensioni rilevanti
  - Territorio, tempo, divisioni dell'impresa

Dashboard per il management della performance aziendale

Scorecard: indicatori standard visti da 4 prospettive



finanziaria

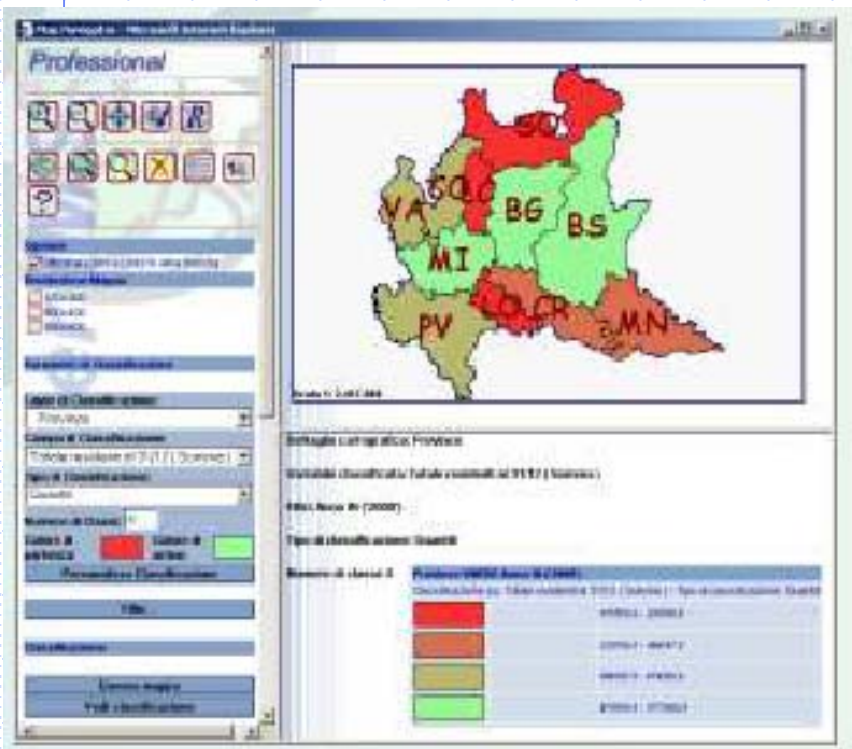
processi interni

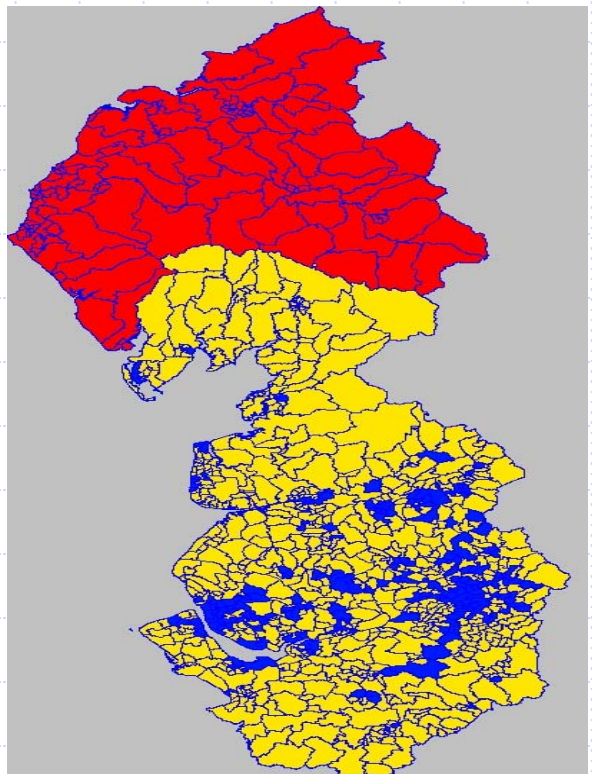
clienti

crescita

Reportistica cartografica

- Navigare la dimensione geografica mediante zoom-in (drill-down) e zoom-out (roll-up) su mappe
- Colori delle zone = visualizzazione del range di valori una misura (classe)



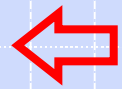


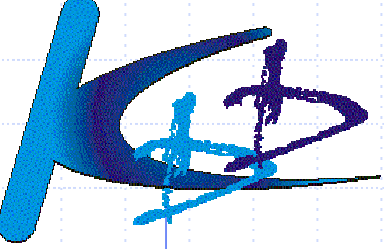
Dashboard & scorecards ...

- ... sono la **punta di un iceberg**
- Rappresentano uno dei prodotti finali della filiera della BI, dai dati grezzi alla conoscenza
- La loro semplicità e immediatezza si basa su un grande lavoro dietro le quinte:
  - integrazione, aggregazione, analisi e sintesi dei dati sorgente

Agenda del Seminario

- Business Intelligence: cos'è, a quali esigenze risponde, come si colloca nell'organizzazione aziendale
- B.I. ed estrazione di conoscenza dalle basi di dati – glossarietto minimo
- Esempi, casi di studio, buone pratiche di B.I. con strumenti di data warehouse
- **Esempi, casi di studio, buone pratiche di B.I. con strumenti di data mining**





AIR MILES

un caso di studio di customer segmentation

G. Saarenvirta, “Mining customer data”, DB2 magazine on line, 1998

http://www.db2mag.com/98fsaar.html

Clustering & segmentazione dei clienti

- Obiettivo: analizzare i dati di acquisto dei clienti per
  - Comprendere i comportamenti di acquisto
  - Creare strategie di business
  - Mediante la suddivisione dei clienti in **segmenti** sulla base di variabili di valore economico:
    - volume di spesa
    - margine
    - frequenza di spesa
    - “recency” di spesa (distanza delle spese più recenti)
    - misure di rischio di defezione (perdita del cliente, churn)

Segmenti

- Clienti high-profit, high-value, e low-risk
  - In genere costituiscono dal 10% al 20% dei clienti e creano dal 50% all'80% del margine
  - Strategia per il segmento: **ritenzione**!
- Clienti low-profit, high-value, e low-risk
  - Strategia per il segmento: **cross-selling** (portare questi clienti ad acquistare altri prodotti a maggior margine)

Segmenti di comportamento di acquisto

- All’interno dei segmenti di comportamento di acquisto, si possono creare sottosegmenti demografici.
- I dati demografici non sono usati, di solito, insieme a quelli economici per creare i segmenti
- I sottosegmenti demografici invece usati per scegliere appropriate **tattiche** (pubblicità, canali di marketing, campagne) per implementare le **strategie** identificate a livello di segmenti.

The Loyalty Group in Canada

- Gestisce lo AIR MILES Reward Program (AMRP) per conto di più 150 compagnie in tutti i settori - finanza, credit card, retail, gas, telecom, ...
- coinvolge il 60% delle famiglie canadesi
- è un programma **frequent-shopper**:
  - Il consumatore accumula punti che può redimere con premi (biglietti aerei, hotel, autonoleggio, biglietti per spettacoli o eventi sportivi, ...)

Acquisizione dei dati

- Le compagnie partner catturano i dati di acquisto e li trasmettono a The Loyalty Group, che
- immagazzina le transazioni in un DW e usa i dati per iniziative di marketing, oltre che per la gestione dei premi.
- Il DW di The Loyalty Group conteneva (al 2000)
  - circa 6.3 milioni di clienti
  - circa un 1 miliardo di transazioni

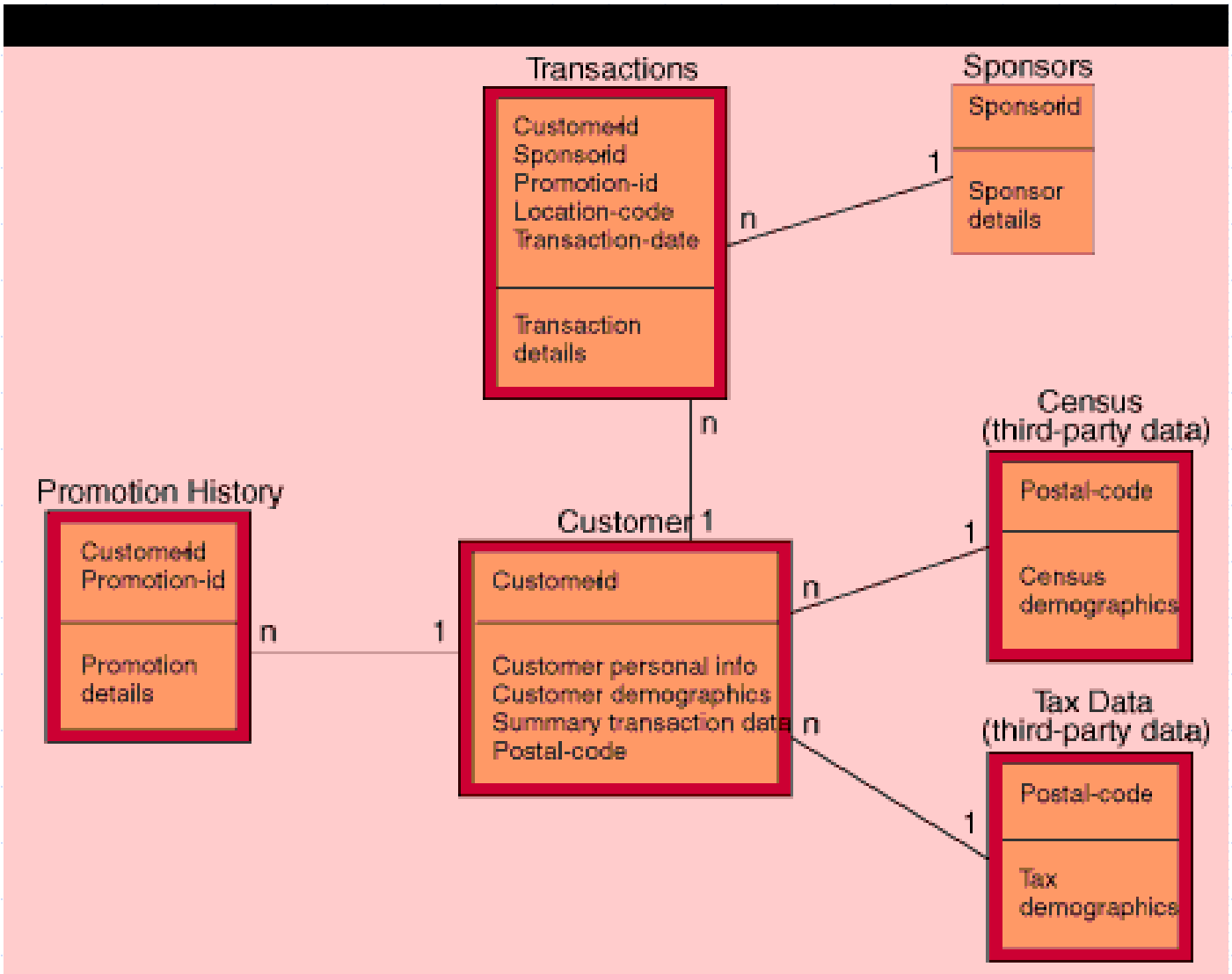
Stato dell'arte prima del data mining

- The Loyalty Group impiega tecniche analitiche standard per la segmentazione dei clienti
  - Recency, Frequency, Monetary value (RFM) analysis
- In sostanza, un modello fatto di regole generali che vengono imposte ai dati per creare i segmenti
- Analogo delle regole di classificazione dei soci Unicoop:
  - Socio costante: ha fatto almeno 2 spese al mese per almeno 3 degli ultimi 4 mesi

Una esperienza di Data mining

- Obiettivo:
  - creare una segmentazione dei clienti
  - a partire dai dati su clienti e loro acquisti nel DW
  - usando il **clustering**, una tecnica di data mining
  - e confrontare i risultati con la segmentazione esistente sviluppata con l'analisi RFM.
- ... lasciare che **i segmenti emergano direttamente dai comportamenti di acquisto simili effettivamente riscontrati nella realtà**, senza imporre un modello preconfezionato ...
- ... e vedere che succede!

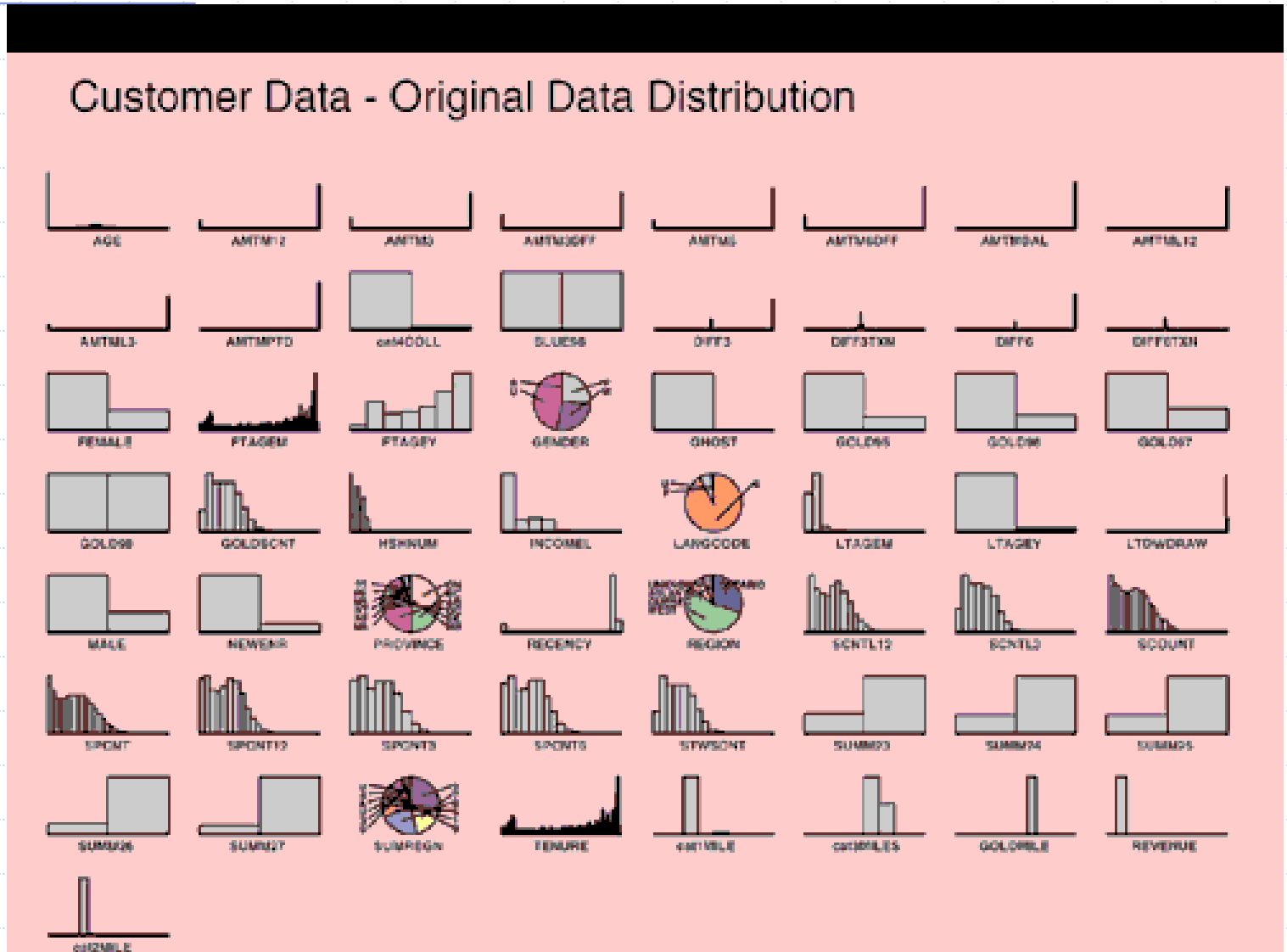
Sorgente dei dati nel DW



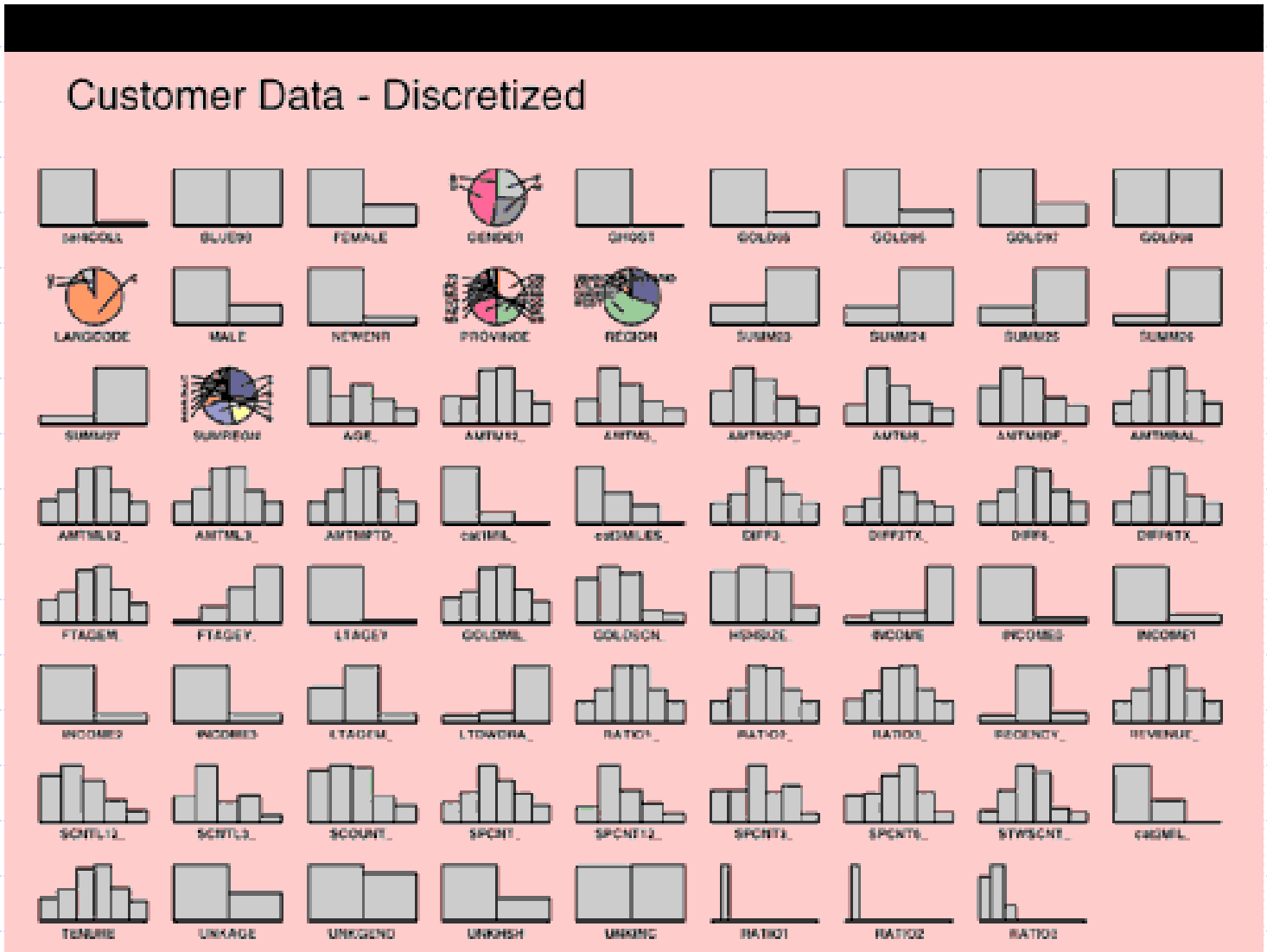
Preparazione dei dati

- Creazione delle variabili economiche di ciascun **cliente**, mediante aggregazione dei propri acquisti
  - Volume di spesa
  - Durata del suo ciclo di vita
  - Numero di compagnie sponsor in cui ha acquistato
  - Numero di compagnie sponsor in cui ha acquistato negli ultimi 12 mesi
  - Distanza (in mesi) dall'ultimo acquisto
  - ...
- Circa 100 variabili economiche derivate dai dati di acquisto nel DW!

I dolori della pulizia dei dati: prima ...



... e dopo la cura



Prima e dopo la cura

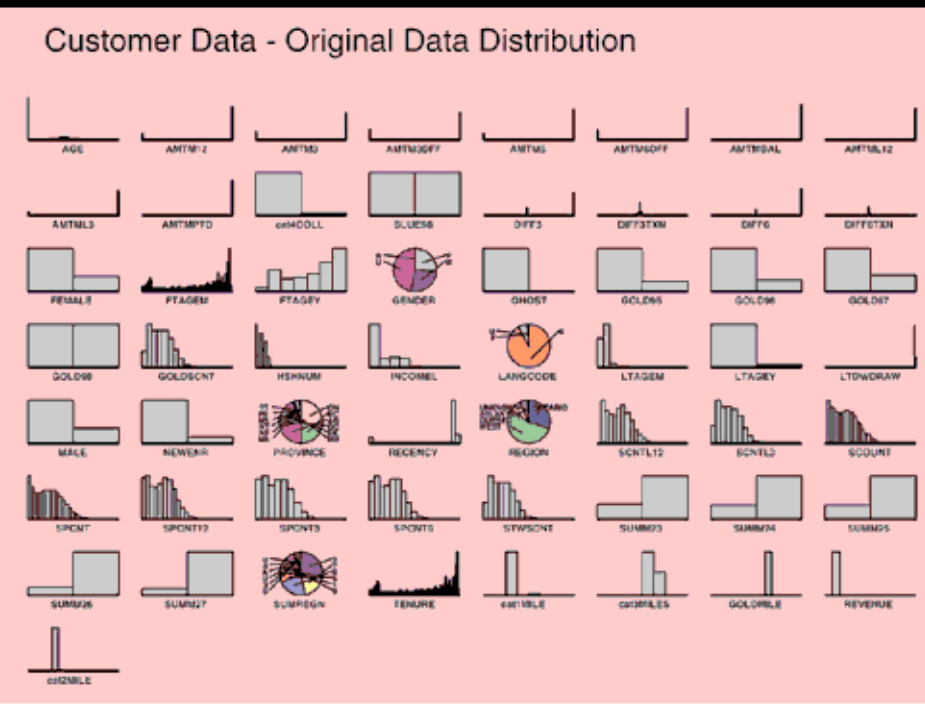


Figure3. *Original data.*

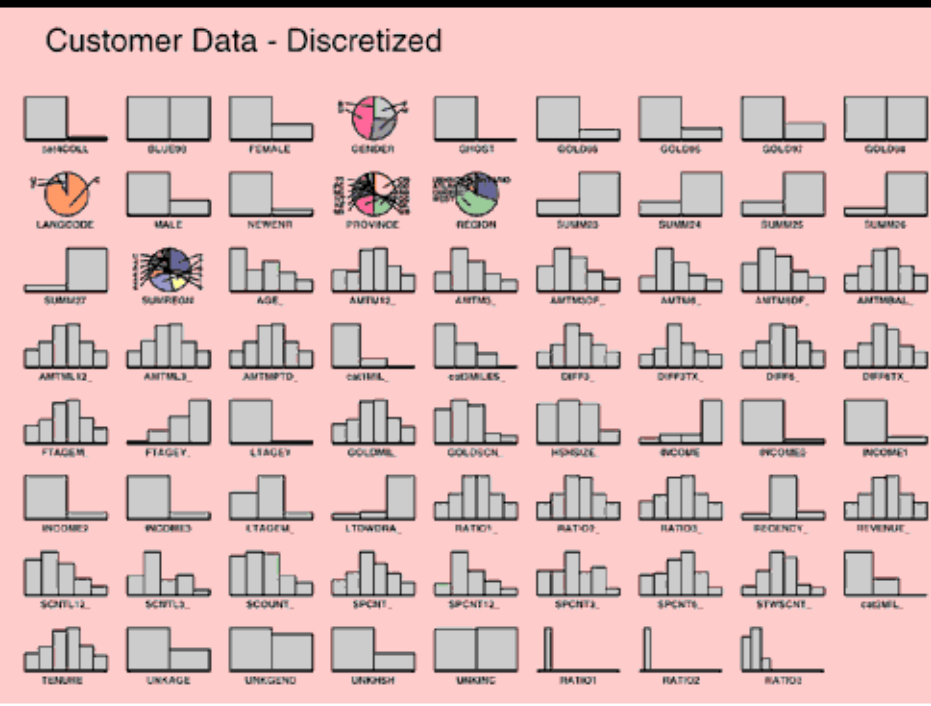
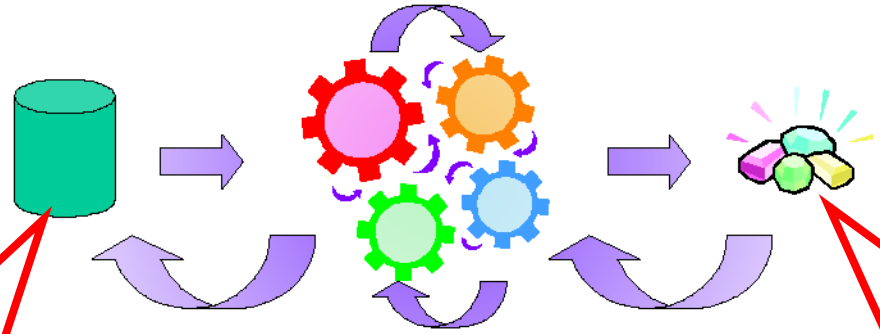


Figure 4. *Discretized data.*

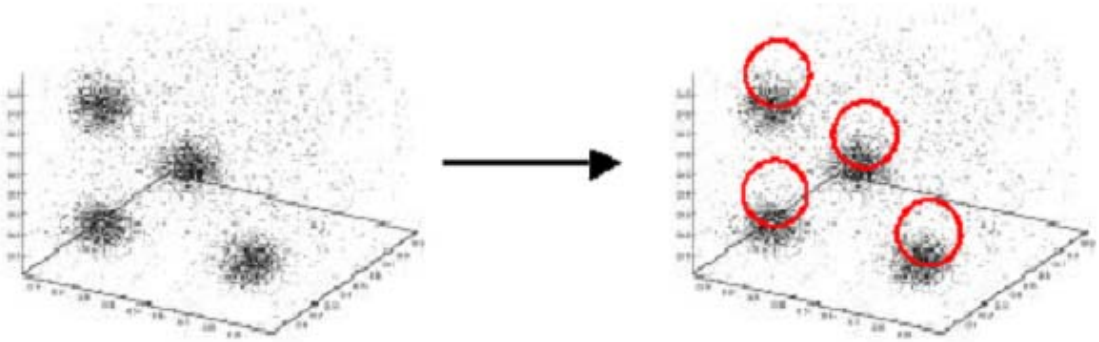
Estrazione del modello di clustering

Clustering = raggruppamento di oggetti simili in gruppi omogenei

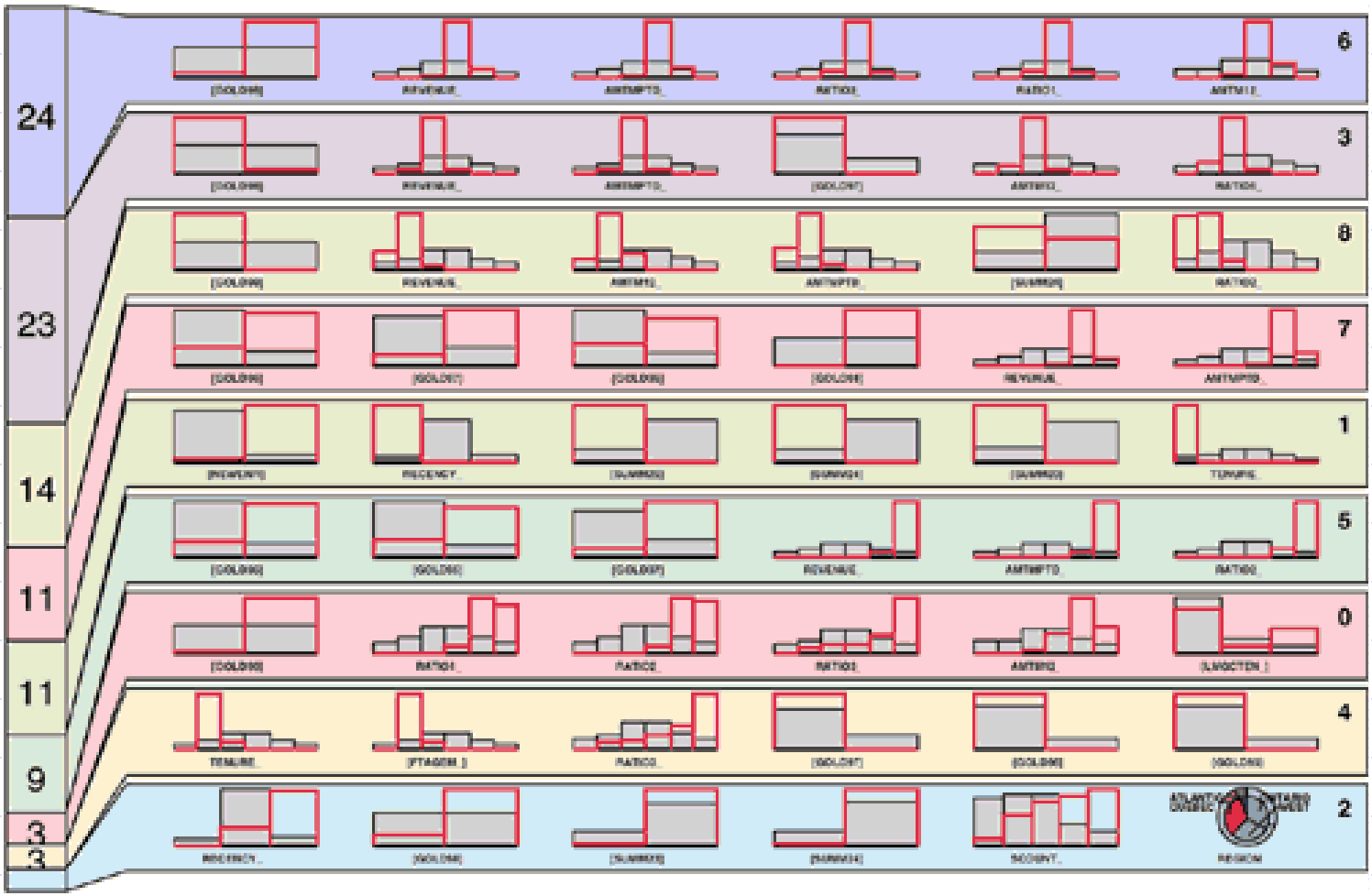


Dati in input: variabili economiche di ciascun cliente

Modello in output: segmentazione dei clienti simili in clusters



Output del clustering



Analisi qualitativa dei cluster

- La variabile Gold98 indica se il cliente è o meno uno migliori clienti, secondo la segmentazione preesistente creata con le tecniche RFM.
- Nel clustering non viene usata: serve solo a “spiegare” i clienti del cluster.
- Il modello di clustering conferma la definizione esistente: tutti i cluster hanno quasi tutti clienti Gold oppure non Gold.

Analisi qualitativa dei cluster

- Ma il risultato non si limita a validare il concetto esistente di cliente Gold:
  - Crea un sottosegmento dei clienti Gold, raffinando la conoscenza preesistente
  - In pratica, è stato scoperto un sottosegmento di clienti Platinum
- **Cluster 5**
  - Quasi tutti clienti Gold98, con molte variabili economiche nei percentili alti

Analisi del cluster 5 – clienti Platinum

- 9 % della popolazione
- volume di spesa totale e mensile, durata, punti redenti, ... sono tutti al di sopra del 75esimo percentile, alcuni addirittura sopra il 90esimo
- Mette in luce un segmento di clienti molto redditizio

Vista dettagliata del cluster 5

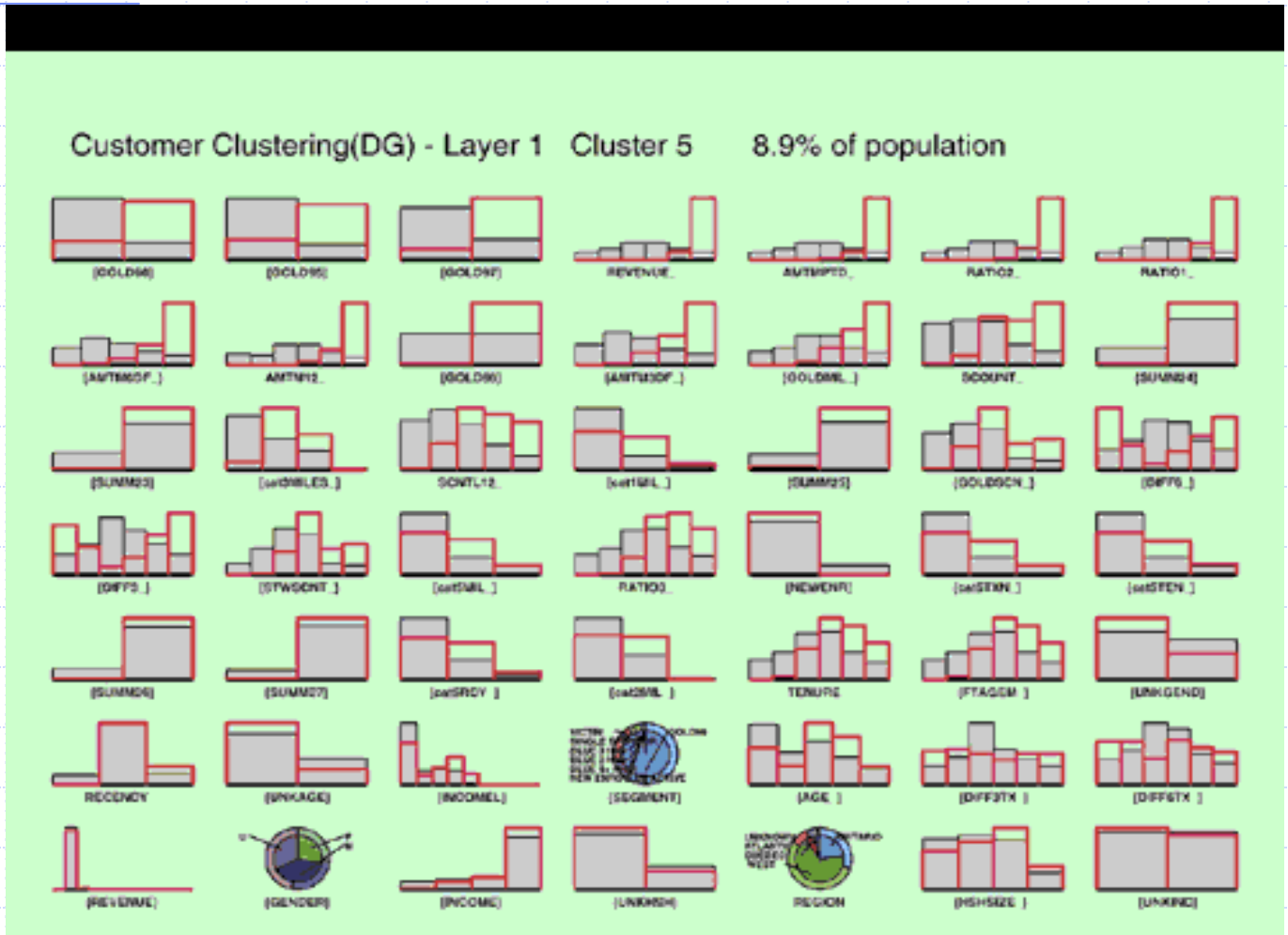


Figure 8. *Cluster 5 output.*

Analisi dei cluster

- Obiettivo: un rapporto che valuti quantitativamente il valore potenziale dei cluster trovati mediante indicatori calcolati per aggregazione sui clienti di ciascun cluster.

CLUSTERID	REVENUE	CUSTOMERS	PRODUCT INDEX	LEVERAGE	TENURE
5	34.74%	8.82%	1.77	3.94	60.92
6	26.13%	23.47%	1.41	1.11	57.87
7	21.25%	10.71%	1.64	1.98	63.52
3	6.62%	23.32%	.73	.28	47.23
0	4.78%	3.43%	1.45	1.40	31.34
2	4.40%	2.51%	1.46	1.75	61.38
4	1.41%	2.96%	.99	.48	20.10
8	.45%	14.14%	.36	.03	30.01
1	.22%	10.64%	.00	.02	4.66

Table 1. *Profiling a cluster.*

Analisi dei cluster

- ***leverage*** = rapporto fra
  - *revenue* (ricavo) e
  - popolazione del cluster.
- Il cluster 5 il più redditizio.
- ***product index*** = rapporto fra
  - numero medio di prodotti acquistati dai clienti del cluster e
  - numero medio di prodotti acquistati dai clienti in generale
- La redditività del cliente aumenta con la *tenure* (durata)
- NOTA: questa non è altro che analisi OLAP con la nuova dimensione della segmentazione appena scoperta!!

Opportunità di business

- Migliori clienti (clusters 2, 5 e 7):
  - indicazione: ritenzione!!
- Clusters 6 e 0
  - indicazione: cross-selling
  - Goal: cercare di convertire i clienti dei clusters 6 e 0 ai clusters 2, 5 o 7.
  - Si può procedere a studiare quali siano i prodotti maggiormente acquistati nei vari clusters per trovare prodotti candidati al cross-selling ...

Opportunità di business (2)

- Clusters 3 e 4
  - indicazione: cross-selling verso i clusters 2, 6 e 0
- Cluster 1
  - indicazione: attendere, potrebbe essere un nuovo segmento di clienti
- Cluster 8
  - indicazione: nessun investimento di marketing (maledetti cherry-peakers!)

Una buona pratica di mining

- Reazioni di The Loyalty Group ai risultati del progetto
  - La visualizzazione dei risultati supporta un livello di analisi significativa e utile alle decisioni.
  - La segmentazione preesistente viene confermata, ma anche raffinata attraverso sottosegmenti sconosciuti a priori, e potenzialmente utili e proficui.
  - Decisione di intraprendere nuovi progetti di mining:
    - Messa a regime della segmentazione usando clustering su dati più completi sui comportamenti di acquisto,
    - Modelli predittivi per direct mail targeting,
    - Identificazione di opportunità di cross selling usando regole di associazione frequenti nei segmenti scoperti.

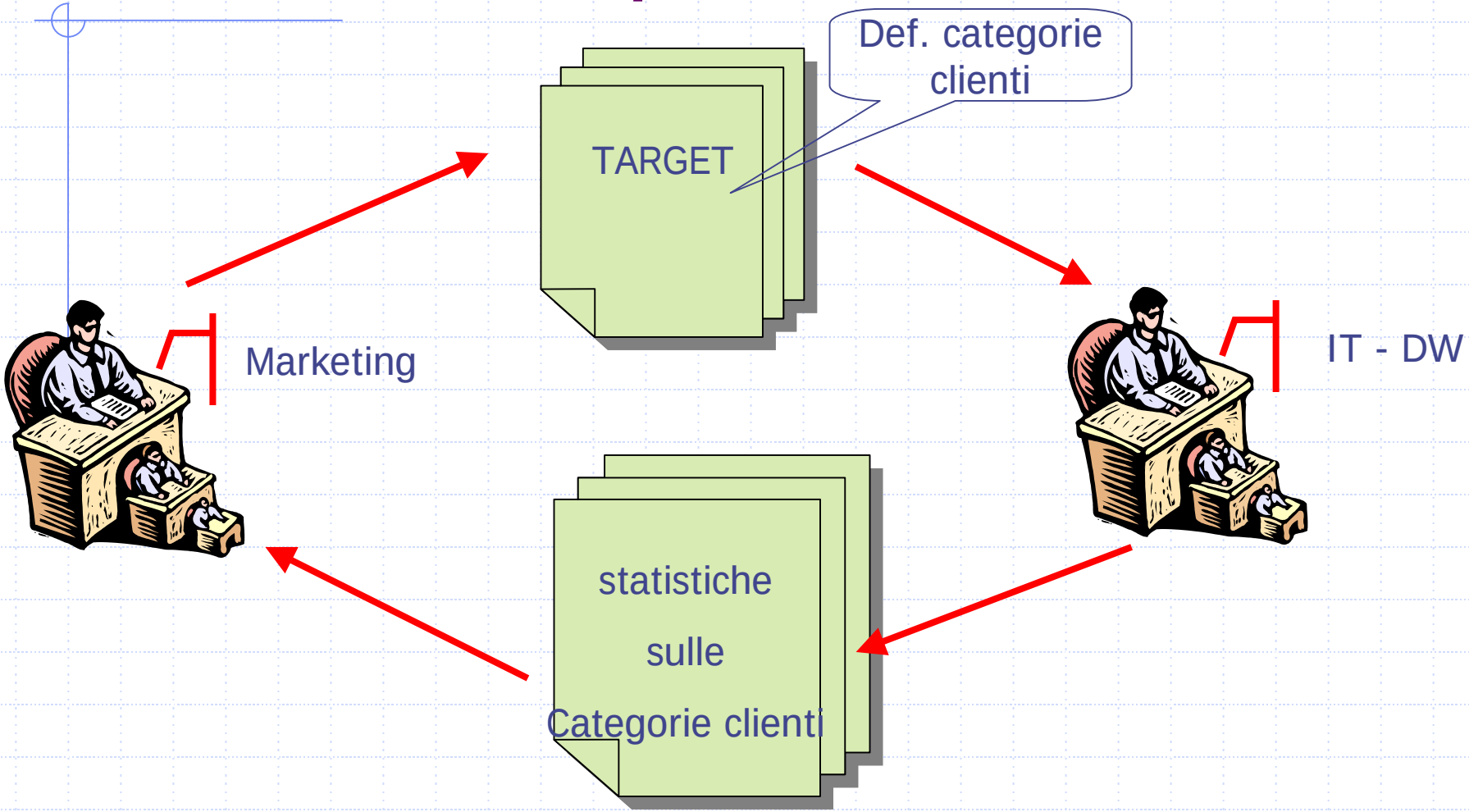
Analisi previsionale per l'ottimizzazione della postalizzazione delle promo

KDD Lab. Pisa

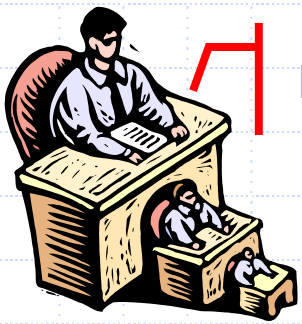
Postalizzazione di promozioni

- Il processo decisionale:
  - Inventare la promozione
  - Selezionare il target
  - Contattare il target
  - Consegnare i premi
  - Tenere traccia dei redenti
  - Valutare a posteriori l'efficacia intervento
- Gli attori
  - Ufficio Marketing, Ufficio IT/DW, Postalizzatore, Ufficio IT/DW , Ufficio Marketing

Inventare la promozione

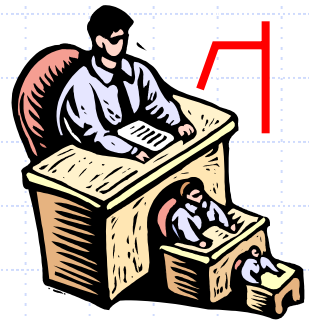


selezionare i clienti e postalizzare



Marketing

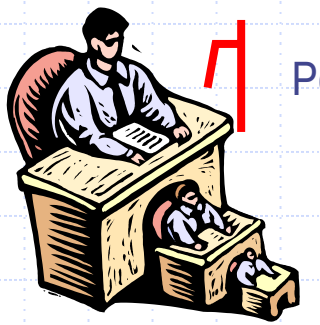
TARGET



IT - DW

Def. dettagli promo

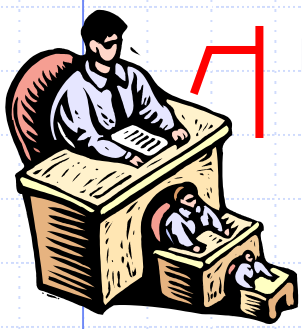
promozione



Postalizzazione

Estrazione Categorie Clienti

Erogare premi e raccogliere dati

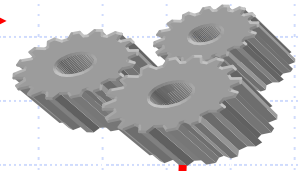


Marketing sviluppo

promozione

Def. dettagli promo

Sistema automatico aggiornamento DATA Mart

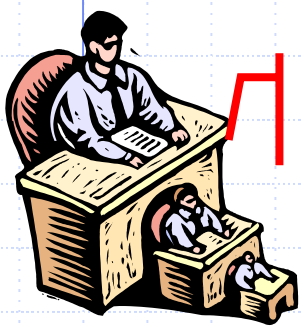


Sistema automatico attivazione mercati

DATA MART PROMOZIONI

Mercati

Analizzare i risultati della promozione



Marketing
analisi

analisi redemption

DATA MART
PROMOZIONI

Gli attori

- Ufficio Marketing inventa la promozione e produce
  - Regole di estrazione delle categorie dei clienti destinatari (**Definizione Target**)
  - Dettagli promozione, tipi di premi per categoria di clienti (**Definizione Promozione**)
  - Diffusione delle informazioni sulla promozione verso i mercati ed il DW
- Ufficio IT/DW produce
  - Statistiche relative alle regole di estrazione
  - Crea le associazione nel DW per la raccolta dati
  - Attiva le procedure di premio nei mercati

Gli attori

- Ufficio Postalizzazione riceve/accede
  - la descrizione promozione e produce, a partire dalle tabella categorie-clienti del DW, il materiale da postalizzare
- Ufficio Marketing/Analisi produce
  - analisi di redemption sulla base di una vista multidimensionale creato dal DW a partire dai dati di vendita per le promozioni di interesse

Promozione

- Definisce per ogni promozione:
  - regole discriminanti per le categorie (costanti, saltuari, inattivi) (da clusterizzazione RFM periodica)
  - Regole discriminanti per sottogruppi di ogni cluster (ulteriori aspetti del comportamento di acquisto)
  - Regole di promozione per ogni categoria (premi, buoni sconto, etc.)

La postalizzazione: è possibile migliorare?

- Nella situazione attuale vengono postalizzati tutti i clienti individuati nelle varie categorie della promozione.
- Se fosse possibile stimare la **probabilità di risposta** (redemption) dei clienti alla promozione, potremmo decidere di postalizzare un sottoinsieme dei clienti, quelli a maggiore probabilità
- Problemi da risolvere:
  - Come stimare la probabilità di redemption?
  - Quale sottoinsieme scegliere?

Ranking dei clienti

- Stima della probabilità di redemption di ciascun cliente sulla base di un **modello previsionale** sviluppato con tecniche di data mining a partire dai dati storici disponibili nel DW
- Ordinamento (ranking) dei clienti in base a questa probabilità

Selezione dei clienti da postalizzare

- Una volta ottenuto il ranking, occorre un criterio per scegliere:
  - La porzione di clienti da postalizzare per raggiungere un rapporto ottimale fra
    - costo di postalizzazione e
    - raggiungimento di clienti ad alta probabilità di redemption
  - La modulazione di postalizzazione fra le varie categorie di clienti definite per la promo
    - costanti, saltuari, inattivi, ...

Come ci si inserisce nel processo decisionale delle promozioni

- Nella preparazione della definizione della Promozione
- Per ogni **gruppo** di clienti della promozione è disponibile un meccanismo per l'analisi di previsione della redemption e di ottimizzazione della postalizzazione
- Meccanismo di base:
  - LIFT CHART

Lift Chart


% clienti rispondenti
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
% clienti postalizzati
Ranking
Random

LIFT CHART

- Asse **X**: percentuali di clienti postalizzati (rispetto al totale del gruppo)
- Asse **Y**: percentuale dei clienti rispondenti che sono raggiunti dalla postalizzazione
- Linea **BLU**: andamento di Y in funzione di X, rispetto ad una scelta **casuale** dei clienti
- Linea **ROSSA**: andamento di Y in funzione di X, rispetto al ranking dei clienti col modello di data mining

Postalizzando il primo 50% dei clienti secondo il ranking si **stima** di raggiungere l'80% dei clienti che redimeranno.

Con la metà dei costi di postalizzazione si **stima** di raggiungere l'80% dei clienti che redimeranno.

Leggere il Lift Chart (1)

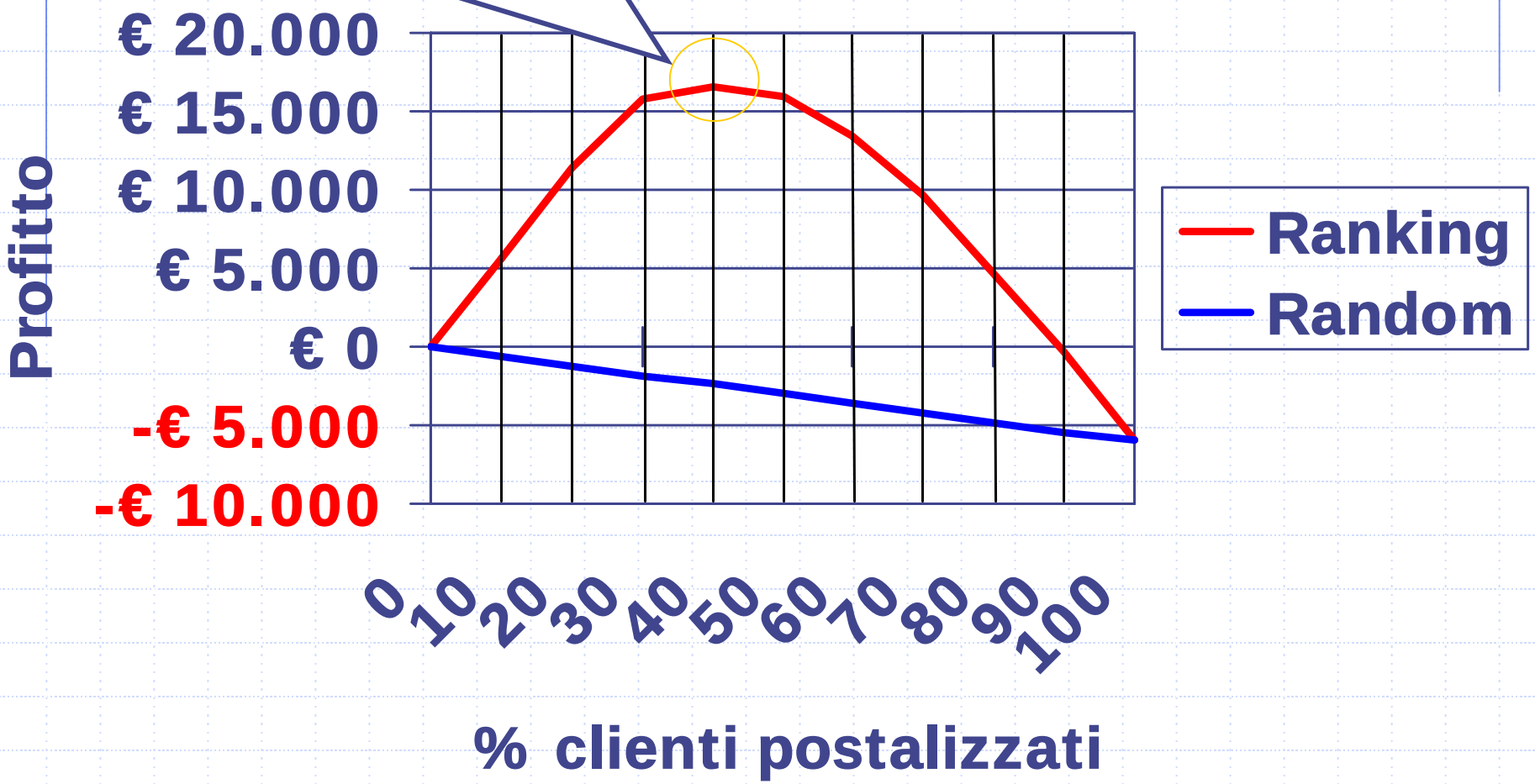
- Il Lift Chart rappresenta un aiuto grafico per ragionare sul rapporto ottimale fra costi di postalizzazione e percentuale di redemption
  - a fronte di sostanziali riduzioni di postalizzati (=budget) permette di ridurre di poco il numero di redenti
  - a parità di budget, permette di incrementare il numero di promozioni oppure di allargare la numerosità delle classi di clienti.

Leggere il Lift Chart (2)

- A partire dal Lift Chart è possibile costruire modelli economici della postalizzazione. **A titolo di esempio:**
  - C = costo unitario di postalizzazione, es. 2,30€
  - B = beneficio unitario di redenzione, es. 6,00€
  - N = numero postalizzabili, es. 30.000
  - T = numero rispondenti postalizzando tutti (stima sulla base dello storico di promozioni simili), es. 10.500 (pari al 35% di 30.000)
  - Profitto = Beneficio – Costo
    - Postalizzando una percentuale P
    - Beneficio = B x T x Lift(P) / 100
    - Costo = C x N x P / 100

Postalizzando il primo 40% dei clienti secondo il ranking si **stima** di massimizzare il beneficio

C=2,30€ B=6,00€ N=30.000 T=10.500.



Le nuove funzionalità per l'ufficio marketing

- Nuova funzionalità per il decisore:
  - accedere al meccanismo di analisi previsionale mediante lift-chart separato per ogni gruppo di clienti
  - modulare la scelta del sottoinsieme di clienti da postalizzare in base:
    - Al ragionamento sul lift-chart, combinato con
    - L'obiettivo di dirigere la promozione in modo preferenziale verso determinati gruppi di clienti (fedeli vs. occasionali, etc.)
  - verificare le conseguenze delle scelte di postalizzazione operate in termini complessivi (copertura, risparmio, etc.), ed eventualmente modificarle

Ma dov'è il **data mining**?!?

- Risposta: **dietro le quinte**!
- Il ranking dei clienti rispetto alla probabilità di redemption è il risultato dello sviluppo di una serie di modelli predittivi che classificano i clienti come rispondenti o meno in base allo storico delle promozioni desumibile dal venduto nel datamart dei Fidelizzati

Dietro le quinte

- Il lift-chart della scheda promo e gli elenchi [illegible]ti da post[illegible] a cura dell'ufficio marketing [illegible]d
  [illegible]hiesta dell'utente marketing/sviluppo, a partire dai modelli predittivi che risiedono sul server (di progetto o di DW)
- I modelli predittivi sono riaggiornati [illegible]mente, ad ogni richiesta dell'utente [illegible]ulla base a cura dell'ufficio IT/DW contenuto attuale del DW, mediante tecniche di data mining

On-line

Off-line

Rilevamento di frodi fiscali e pianificazione degli accertamenti

Sorgente: Ministero delle Finanze

Progetto Sogei, KDD Lab. Pisa

Lotta all'evasione – Min. Finanze/SOGEI ('98-'99)

- **Pianificazione di accertamenti fiscali**
- **Obiettivo**: costruire un modello predittivo che individui una porzione di contribuenti su cui risulti vantaggioso effettuare un controllo fiscale.
  - Estrazione di **alberi di decisione**
- **Dataset**:
  - dati storici provenienti da fonti diverse (mod. 760, mod. 770, INPS, ENEL, SIP, Camere del Commercio)
  - dati storici sui risultati degli accertamenti pregressi.
- Variabile da predire: imposta recuperata al netto delle spese di accertamento.
- Valutazione dei modelli estratti rispetto ad **indici** generali (accuratezza) e specifici di dominio (redditività)

Rilevamento di frodi

- Obiettivo generale:
  - Determinare *modelli* per la previsione del comportamento fraudolento per:
  - Prevenire frodi future (rilevamento di frodi *on-line*)
  - Scoprire frodi passate (rilevamento frodi *a posteriori*)
- Obiettivo specifico:
  - Analizzare i dati storici sulle verifiche per pianificare verifiche future più EFFICACI

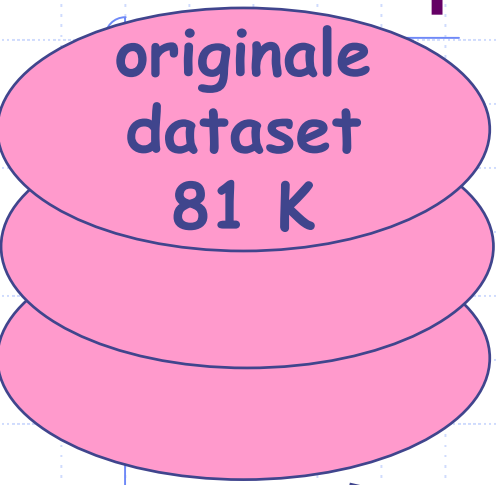
Pianificazione di verifiche

- C’è un trade-off tra:
  - *Massimizzare i benefici della verifica*: selezionare quei contribuenti che massimizzano il recupero di tasse evase.
  - *Minimizzare il costo della verifica* : selezionare quei contribuenti che minimizzano le risorse necessarie alla verifica.

Available data sources

- Dataset: Dichiarazioni dei redditi, su una classe selezionata di aziende italiane integrate con altre sorgenti:
- Contributi INPS per dipendenti, consumi ENEL e telefonici..
- Dimensione: 80 K tuple, 175 numerici attribute.
- Un sottoinsieme di 4 K tuples corrisponde ad aziende verificate:
  - I risultati delle verifiche sono memorizzati nell'attributo: *recovery* (= *amount of evaded tax ascertained*)

Data preparation

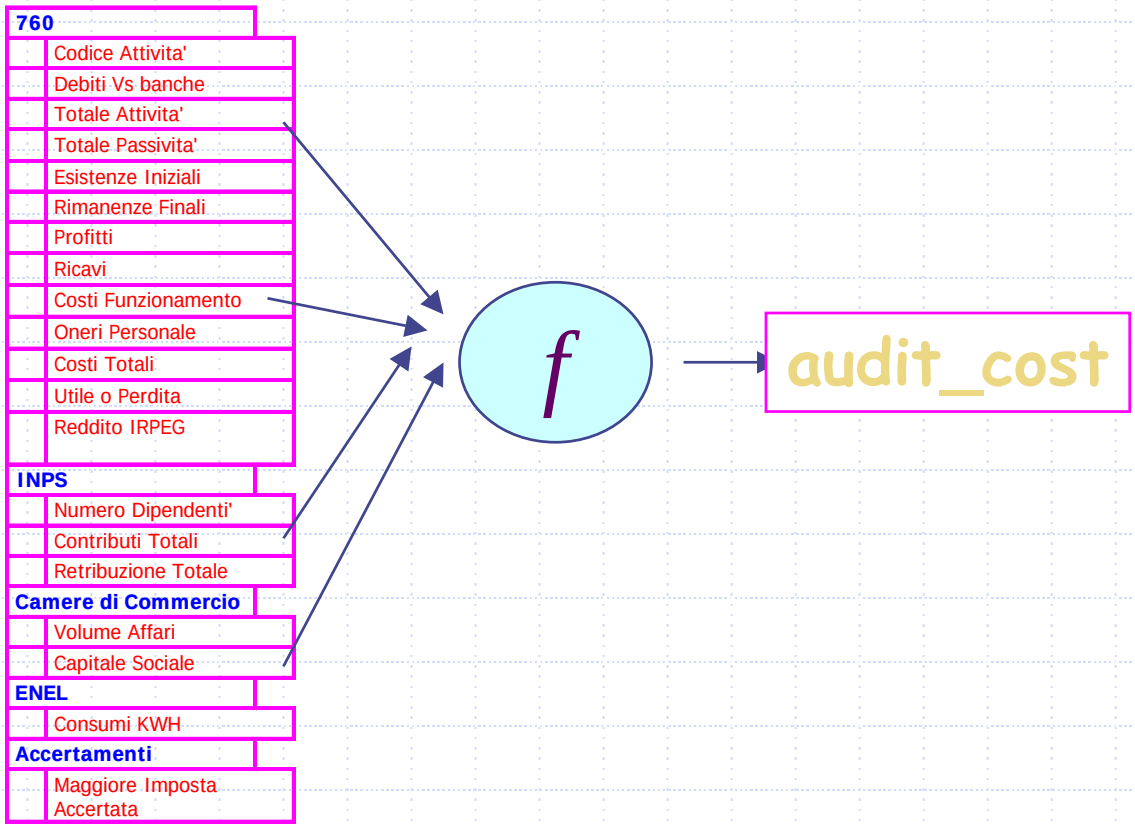


data consolidation
data cleaning
attribute selection

TAX DECLARATION	
	Codice Attivita'
	Debiti Vs banche
	Totale Attivita'
	Totale Passivita'
	Esistenze Iniziali
	Rimanenze Finali
	Profitti
	Ricavi
	Costi Funzionamento
	Oneri Personale
	Costi Totali
	Utile o Perdita
	Reddito IRPEG
SOCIAL BENEFITS	
	Numero Dipendenti'
	Contributi Totali
	Retribuzione Totale
OFFICIAL BUDGET	
	Volume Affari
	Capitale Sociale
ELECTRICITY BILLS	
	Consumi KWH
AUDIT	
	Recovery

Modello di costo

- si definisce l'indicatore audit_cost come funzione di altri attributi



Modello dei costi e variabile target

- Recupero di una verifica
  - actual_recovery = recovery - audit_cost
- La variabile target (class label) della nostra analisi: Class of Actual Recovery (c.a.r.):
- *c.a.r. =*

$$\begin{cases} \textit{negative} & \textbf{\textit{if}} \quad \textit{actual\_recovery} \leq 0 \\ \textit{positive} & \textbf{\textit{if}} \quad \textit{actual\_recovery} > 0. \end{cases}$$

Indicatori di qualità

- Si costruiscono vari classificatori che sono valutati secondo diverse metriche:
- Domain-independent indicators
  - confusion matrix
  - misclassification rate
- Domain-dependent indicators
  - audit #
  - actual recovery
  - profitability
  - relevance

Indicatori Domain-dependent

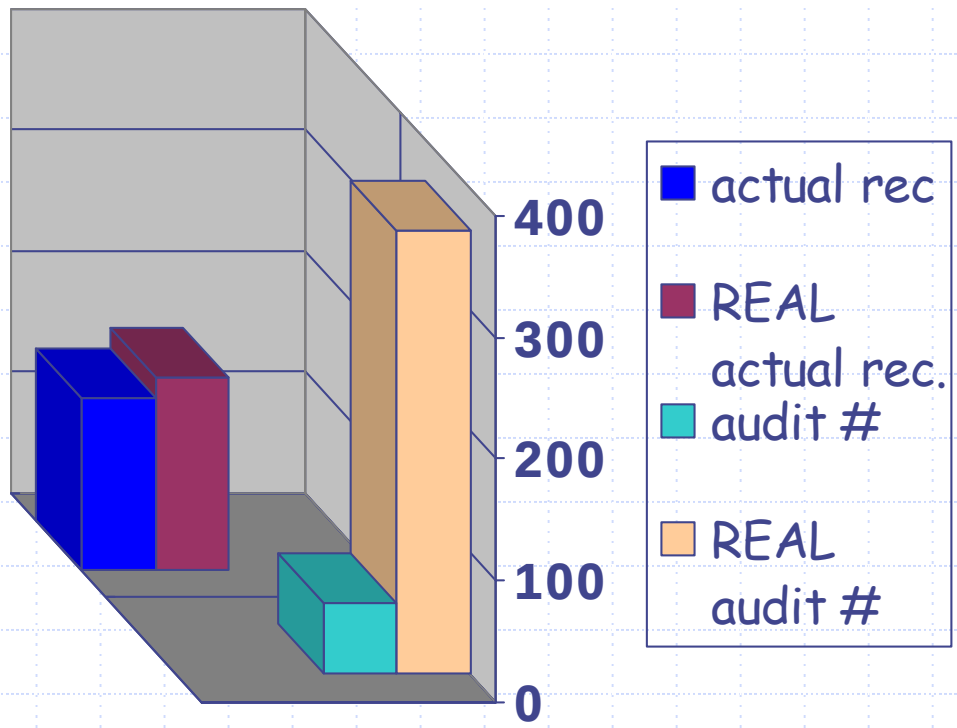
- audit # (di un dato classificatore): numero di tuple classificate come positive = # ($FP \cup TP$)
- actual recovery: ammontare totale del recupero effettivo per tutte le tuple classificate come positive
- profitability: recupero effettivo medio per verifica
- relevance: rapporto tra profitability e l'errore di classificazione

Il caso REAL

- I Classificatori sono confrontati con l'intero test-set, cioè gli accertamenti veramente condotti.
- audit # (REAL) = 366
- actual recovery(REAL) = 159.6 M euro

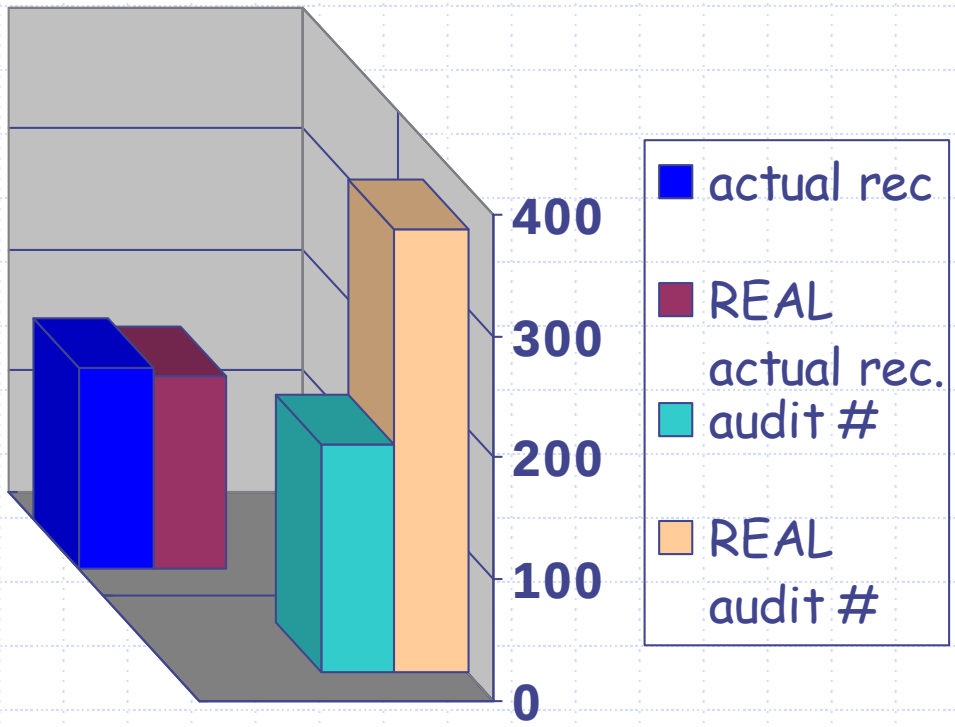
Classificatore 1 (min FP)

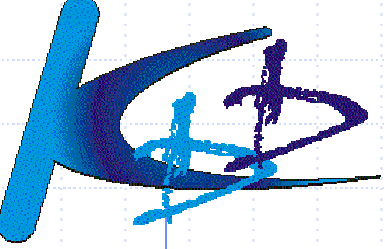
- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.*= 141.7 Meuro
- *profitability* = 2.401



Classificatore 2 (min FN)

- *misc. rate* = 34%
- *audit #* = 188 (98 FP)
- *actual rec.*= 165.2 Meuro
- *profitability* = 0.878



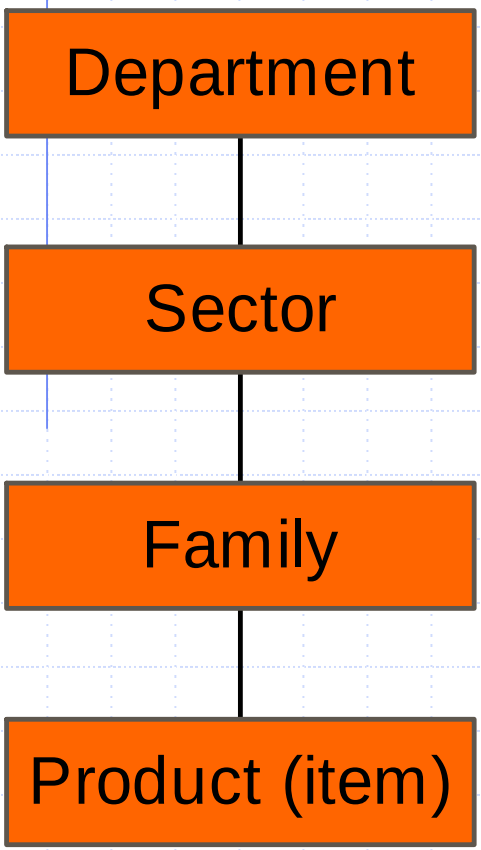


Market Basket Analysis presso la COOP

DataSift e COOl Patterns

KDD Lab. Pisa

Datasift – COOP (’96-’99)



- Progetto pionieristico di Market Basket Analisys a partire da dati di vendita (**scontrini**)
- Estrazione di **regole associative**
- Ragionamento sulle regole estratte ai diversi livelli della **gerarchia dei prodotti**
- Studio dell’effetto delle **promozioni** sulla dinamica temporale delle regole estratte.
- Data Mining **Query Language**

Quali strumenti per MBA?

- Regole associative
  - A->B (chi compra A frequentemente compra anche B)
- Gli analisti di marketing sono interessati a regole business del tipo:
  - L'assortimento è adeguato per un certo target di clienti del supermercato?
  - La campagna promozionale è stata efficace nello stabilire un certo comportamento (desiderato) d'acquisto?

REGOLE DI BUSINESS: ragionamento temporale sulle RA

- Quali regole sono generate/confermate dalla promozione?
- Come cambiano le regole nel tempo?

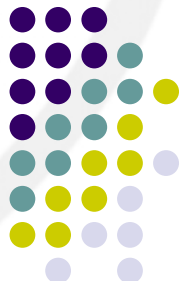
COOL PATTERNS

Progetto "COOL PATTERNS"

Analisi delle vendite nella grande distribuzione

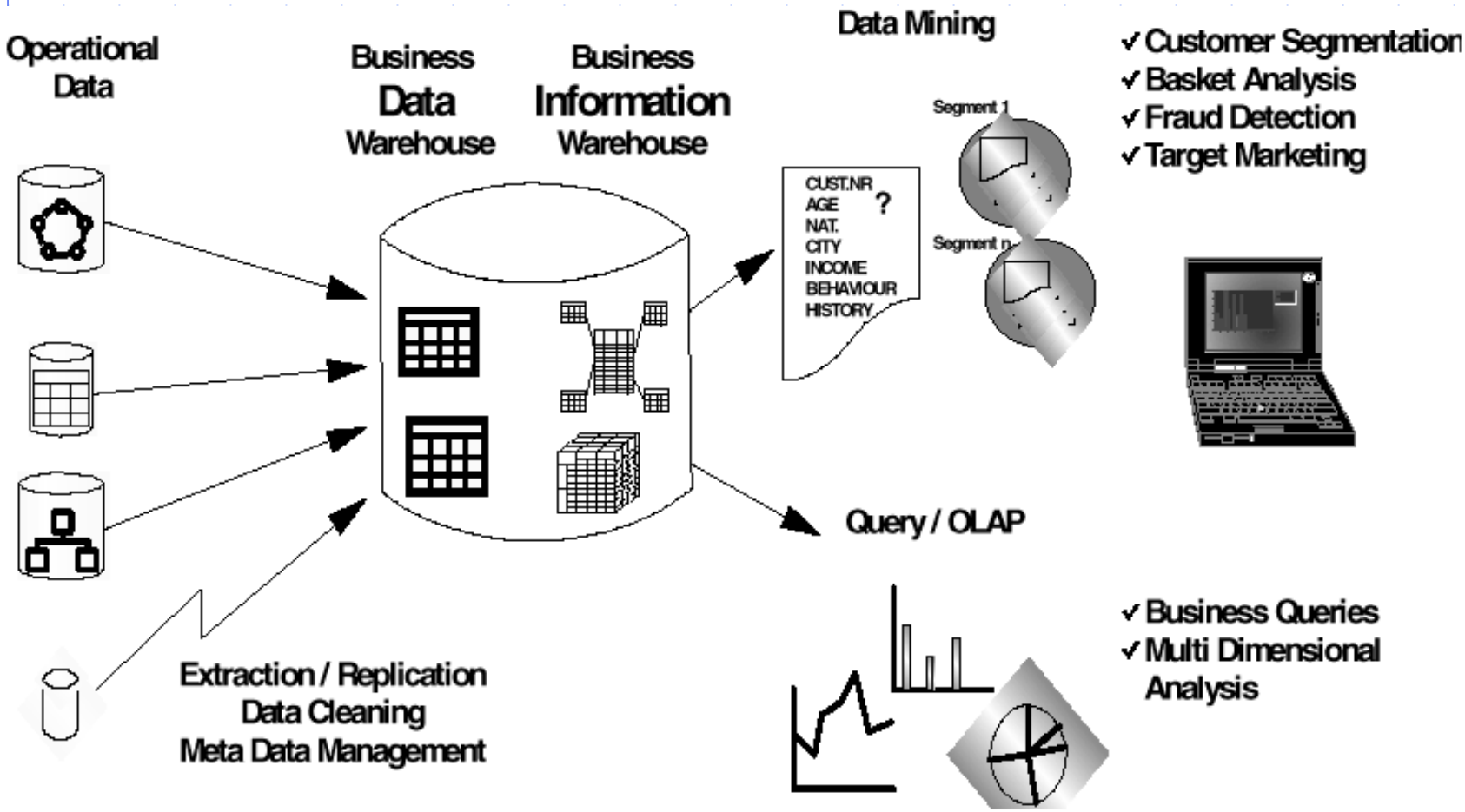
Analisi dei Dati ed
Estrazione di Conoscenza
2004/2005

Federico Colla



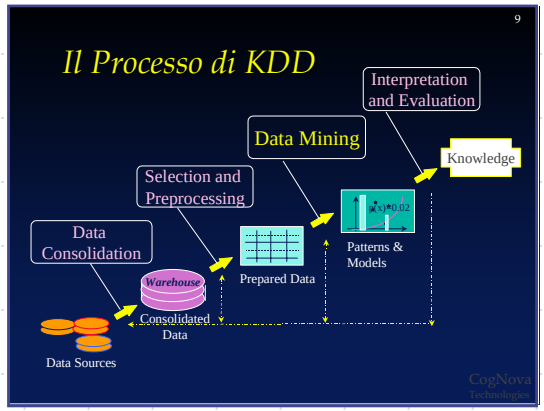
… per concludere, debrief!

La piattaforma abilitante per la B.I.



Il ciclo *virtuoso della filiera BI*

Problema



Conoscenza

Identificare il problema e le opportunità

Utilizzare la conoscenza

Strategia

Misurare gli effetti dell'azione

Risultati

Figure per la B.I.

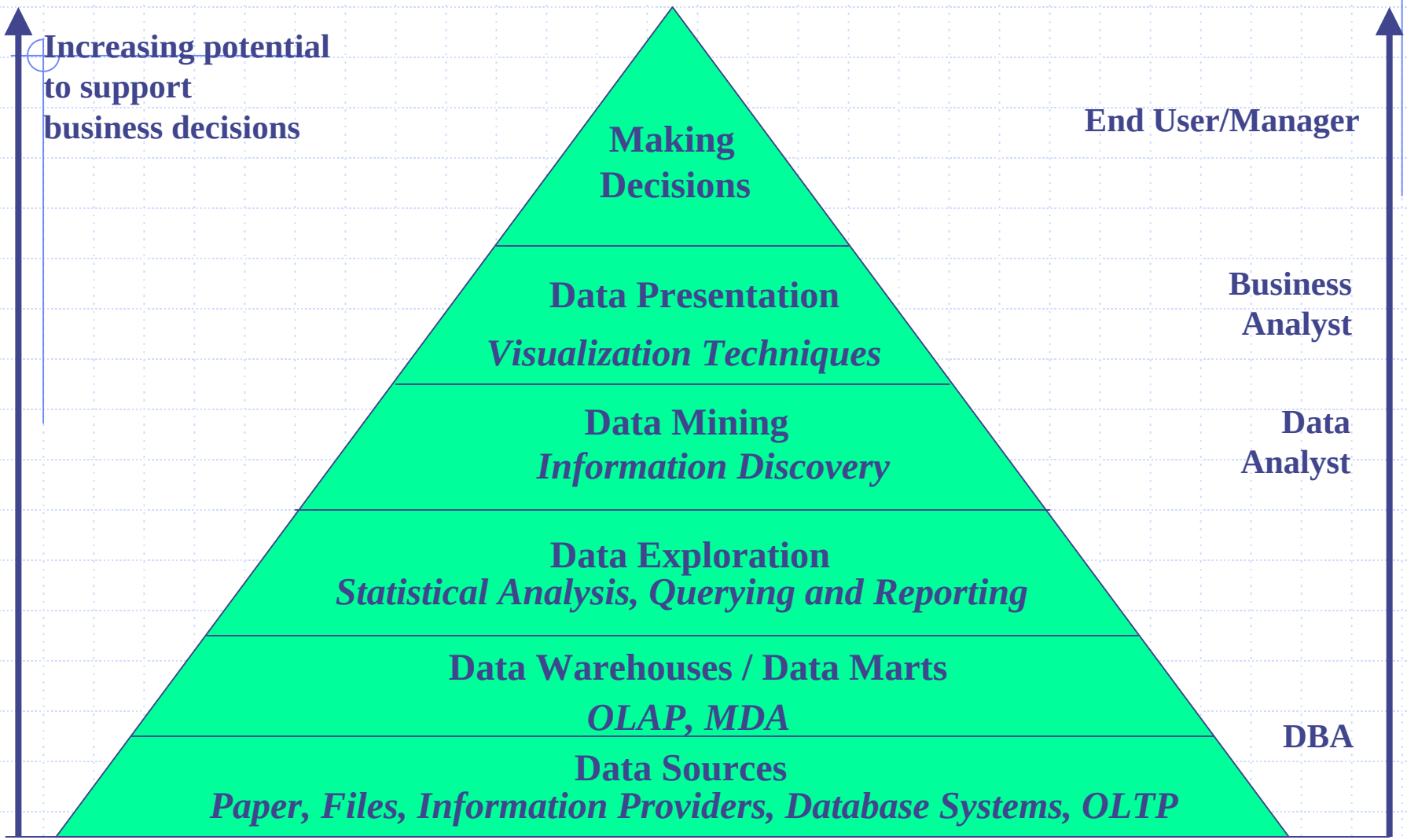
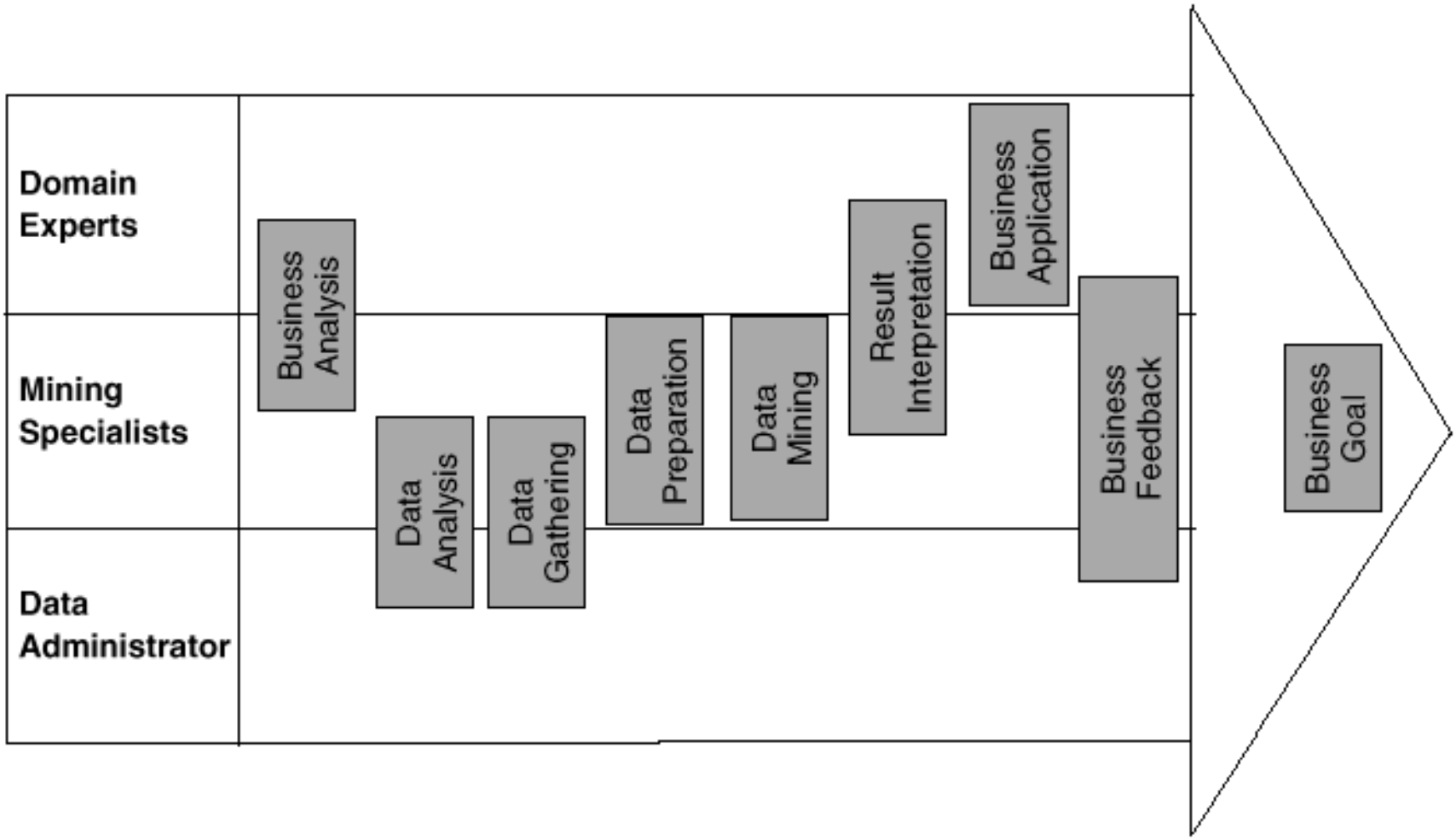


Figure nel processo di KDD



Intelligence/Value

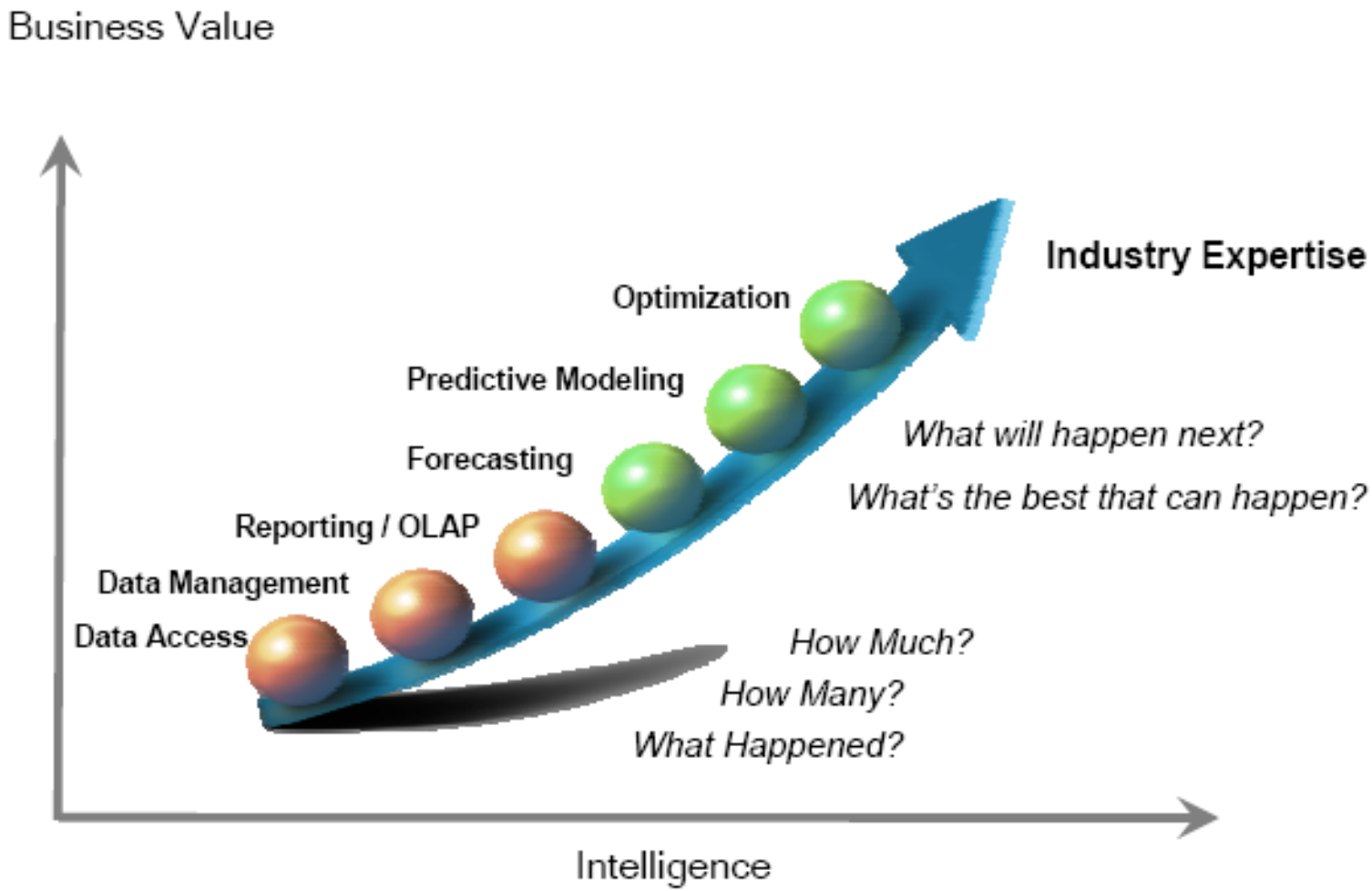


Figure 1: Business value increases exponentially with intelligence.

Business Intelligence come cultura aziendale

- Dice il saggio: Se una soluzione di B.I. non ti aiuta a prendere buone decisioni, *velocemente, facilmente e con fiducia*, non è né buona né intelligente
- B.I. come strategia aziendale piuttosto che come tattica per un singolo problemino
- Non paga come soluzione spot

Investire nella B.I.

- La B.I. non è un investimento puramente tecnologico, ma sui tre piani
  - **Competenze, Organizzazione, Tecnologie**
- Il segreto del successo è usarla come leva dell'evoluzione professionale delle diverse figure coinvolte
  - Tecnici IT (amministratori e progettisti database)
  - Analisti (dei dati e del business)
  - Utenti finali (manager in senso lato, ad ogni livello)

Le capacità professionali di questi tre gruppi di figure devono crescere insieme per la (e grazie a) la diffusione della B.I. in azienda

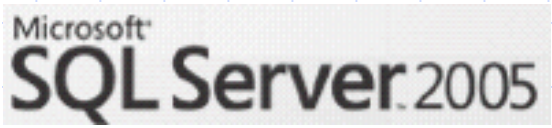
Nuove competenze per la B.I.

- Tecnici IT:
  - Da progettisti e amministratori DB
  - A progettisti e amministratori DW e creatori di cubi tematici
- Analisti (dei dati e del business)
  - Da estensori manuali di rapporti
  - A creatori di rapporti e cruscotti interattivi
- Utenti finali (manager in senso lato, ad ogni livello)
  - Da consumatori di rapporti cartacei o, al massimo, di fogli Excel
  - A navigatori di rapporti multi-dimensionali e di tabelle pivot di Excel

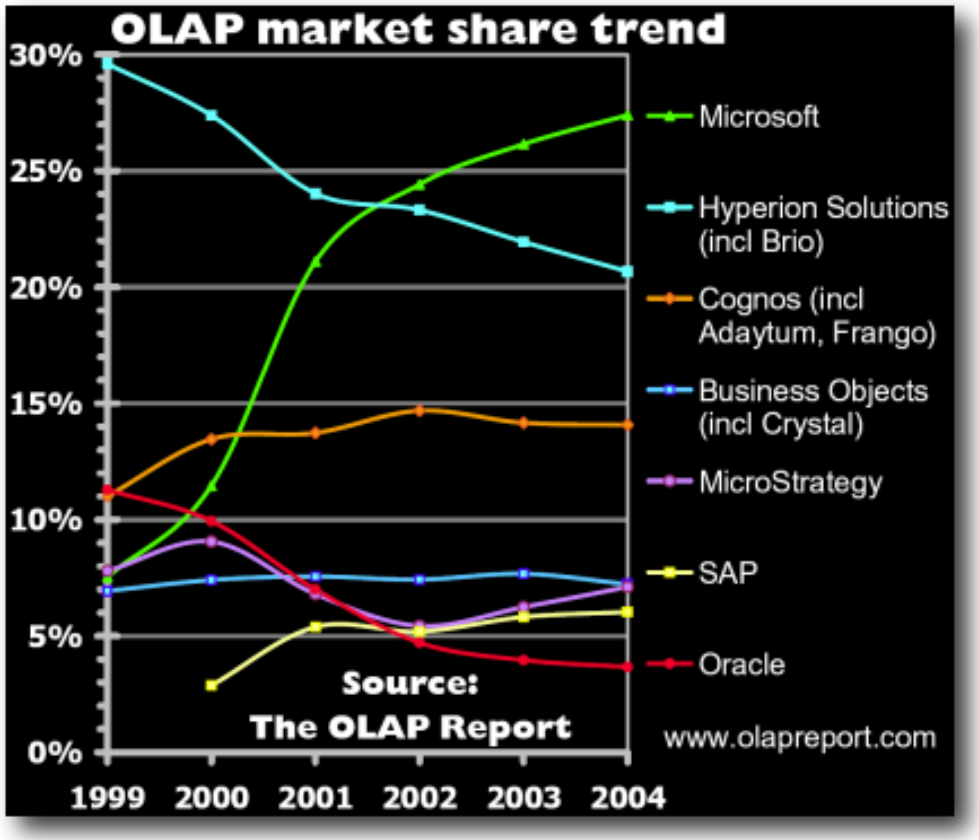
Business Intelligence: è un business essa stessa

- Previsione: il mercato della B.I. nel 2009
- a livello mondiale: 2.3 miliardi di dollari con una crescita annua del 6%
- in Europa: 852,5 milioni di dollari, 5.6% di crescita annua (1/3 del mercato mondiale)
- Stima Gartner group

I principali vendor di B.I.

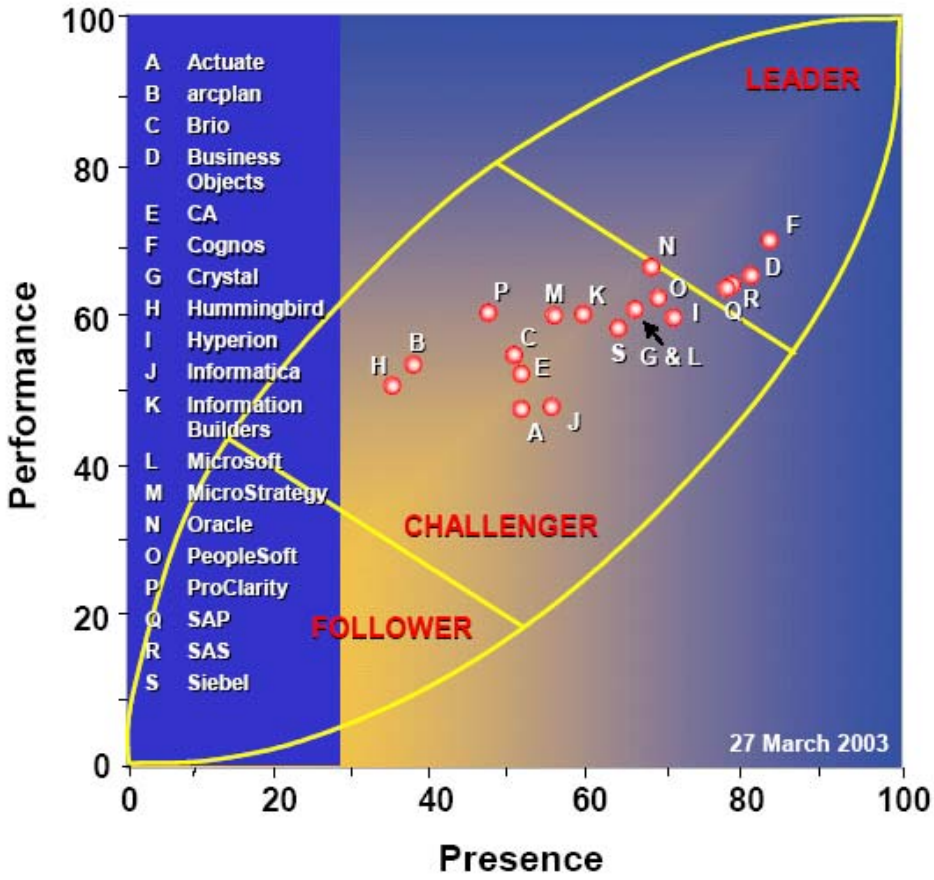


OLAP Market Share



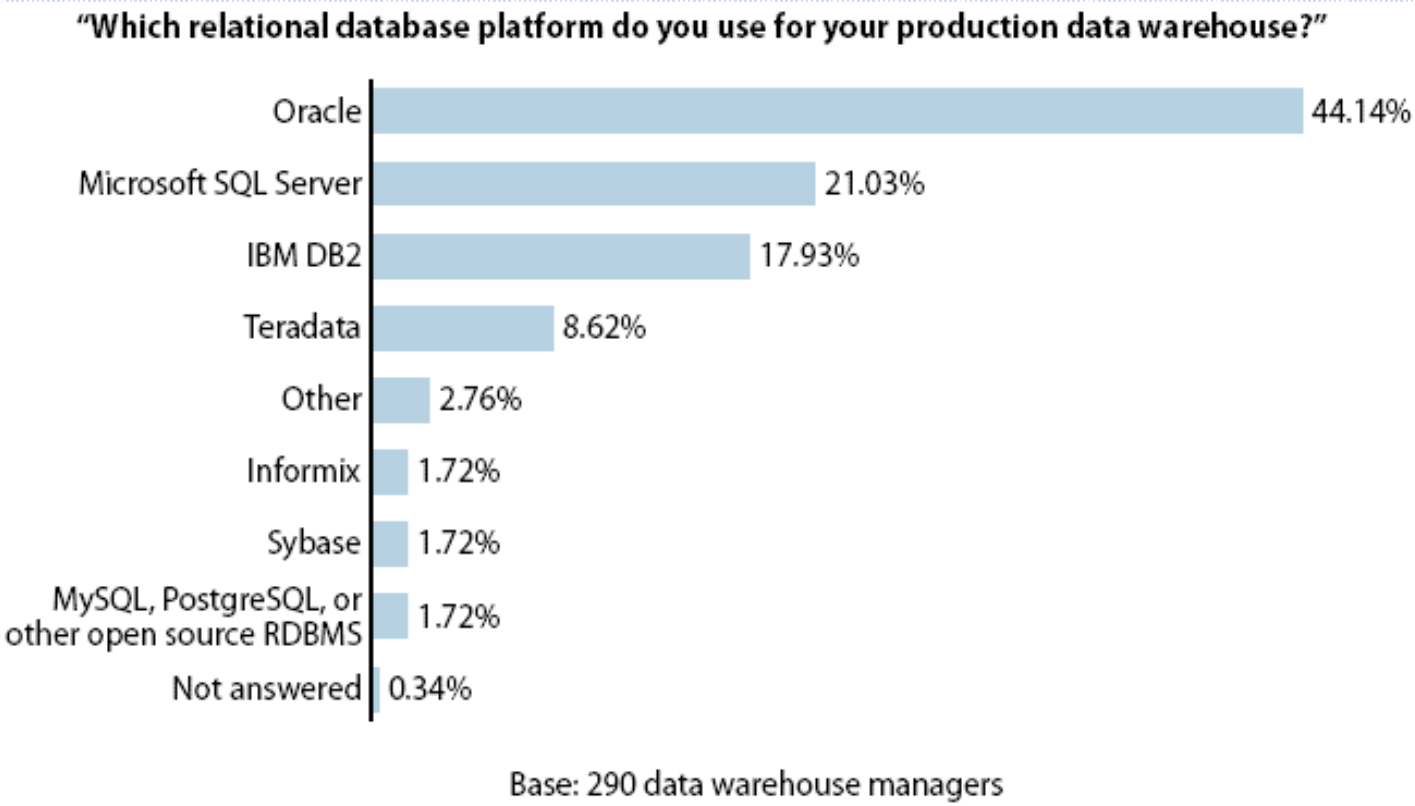
- Olap report: http://www.olapreport.com

Prodotti OLAP



- METAspectrum evaluation 2003

Integrazione RDBMS-OLAP



- TDWI-Forrester Survey 2004

Prodotti Data Mining

- METAspectrum evaluation 2004

Una metafora fotografica

- Mastering data mining (and BI)
- Padroneggiare la BI = padroneggiare l'arte della fotografia
- Dal libro *Mastering Data Mining*
  - Barry & Linoff, 2002

Usare una Polaroid

- Acquisire analisi preconfezionate da aziende esterne del settore, ad esempio Nielsen
- Acquisire informazione statistica aggregata, ad esempio dall'ISTAT
- Acquisire i risultati di ricerche (survey) demografiche, di mercato, studi di settore, ...

Usare una “automatica”

- Acquisire soluzioni software che inglobano, dietro le quinte, meccanismi e tecnologie di B.I., mirati a specifiche applicazioni
- Prodotti verticali “preconfezionati”
  - Sistema di alert per Credit Card Fraud detection
  - Sistema previsionale per Churn Management (gestione delle defezioni dei clienti)
- Sistemi di Customer Relationship Management (ad esempio, Decisionhouse)

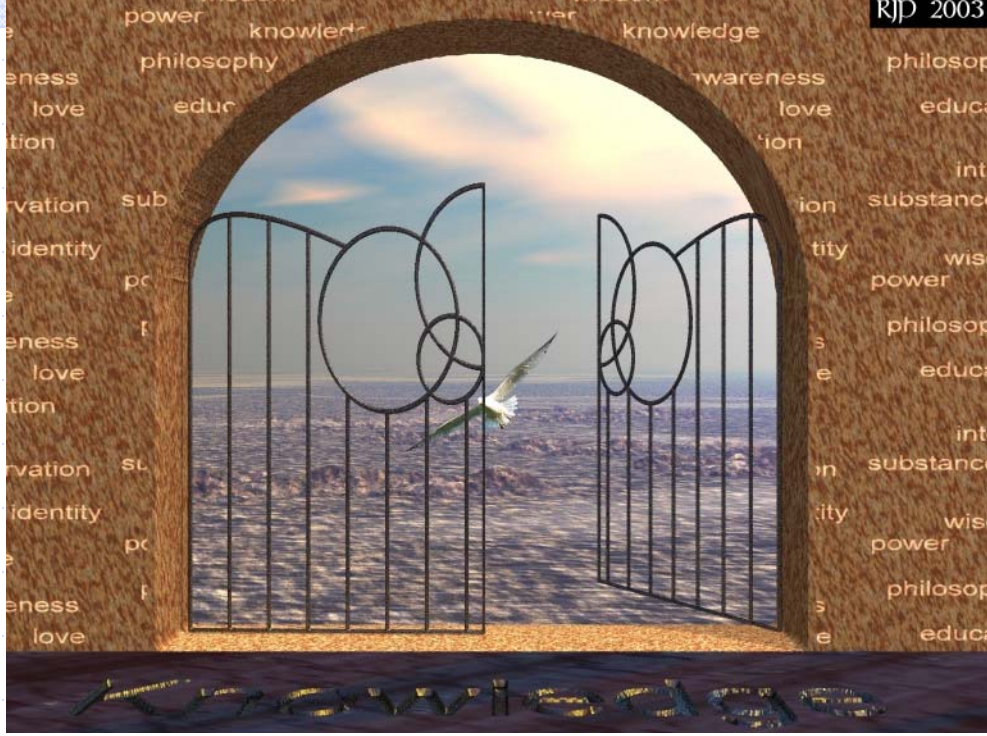
Assumere un fotografo professionista

- Dotarsi di consulenti esterni per compiti di analisi avanzata, ad esempio analisi previsionale.
- Valevole nella fase iniziale
- Fallisce quando tutti i modelli, i dati e la conoscenza generata rimane nelle mani degli esterni
- Il punto è **come** usare l'esperienza esterna
- "Un profeta di un'altra terra può avere più successo nel persuadere il management a seguire una nuova strada".
- Progetti pilota con laboratori di ricerca orientati al trasferimento tecnologico

Costruire la propria camera scura e diventare un fotografo esperto

- **Sviluppare in casa** le competenze.
- Un obiettivo di medio periodo, da raggiungere gradualmente.
- Chi conosce sia i dati che il business produce modelli migliori. E **conoscenza** più utile.

Conoscenza



Science is built up with facts,
as a house is with stones.
But a collection of facts
is no more a science
than a heap of stones is a house.

Henri Poincaré,
La Science et l'hypothèses, 1901

Stile toscano

Considerate la vostra semenza:
fatti non foste a viver come dati
ma per seguir virtute e canoscenza

Dante, Inferno, canto XXVI