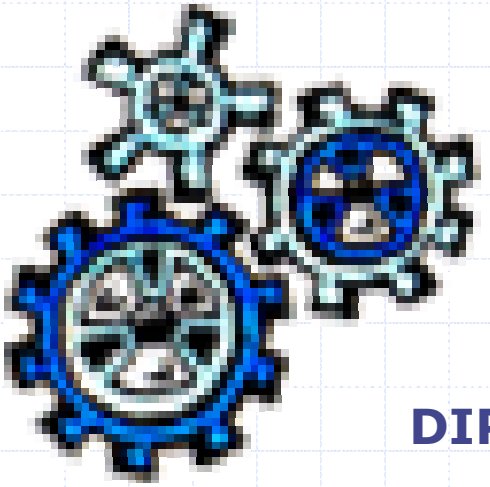
Data Mining

Fosca Giannotti and Mirco Nanni
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/



DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2007/2008

Data Mining

- Acronimo: DM
- Orario: Mercoledi14-16 aula C1, Venerdi 9-11aula B1
- Docenti:
  - Fosca Giannotti, ISTI-CNR, fosca.giannotti@isti.cnr.it
  - Mirco Nanni, ISTI-CNR, mirco.nanni@isti.cnr.it
- Ricevimento:
  - Giannotti: mercoledì 15-17, ISTI, Area Ricerca CNR, località San Cataldo, Pisa (prenotazione per e-mail)

Data Mining

- Riferimenti bibliografici
  - **Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to DATA MINING, Addison Wesley, ISBN 0-321-32136-7, 2006**
  - Barry Linoff Data Mining Techniques for Marketing Sales and Customer Support, John Wiles & Sons, 2002
- I lucidi utilizzati nelle lezioni saranno resi disponibili attraverso il sito web del corso: http://didawiki.cli.di.unipi.it

Data Mining- teoria

- Mining di pattern frequenti
- Mining di dati sequenziali,
- Mining di serie temporali e dati spazio temporali
- Mining di grandi grafi e reti
- il processo KDD: lo standard CRISP.
- Impatto sociale del data mining - Data mining e protezione della privacy
- Opinion Mining

Data Mining – Case di studio

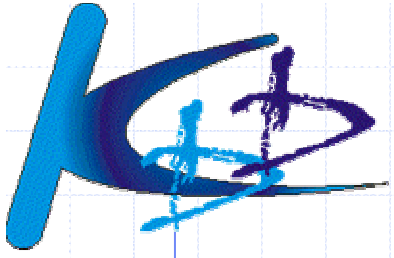
- Rilevamento di frodi: Sogei1, DIVA (progetto 1)
- Grande distribuzione: data set COOP, TargetMarketing: PromoRank, ChurnAnalysis: coop (progetto 2)
- Sanità, case study su fascicolo sanitario elettronico
- Industria delle telecomunicazioni: analisi da dati GSM: ORANGE, i flussi turistici.
- Mobilità e trasporti: esplorazione, e postprocessing per la validazione dei comportamenti di mobilità. progetto3

Analisi dei Dati ed Estrazione di conoscenza

- Una parte centrale dove si continua l'ntroduzione delle principali tecniche di datamining (regole associative, sequential pattern mining, serie temporali, graph mining, spazio.temporal data mining). Di queste tecniche si studieranno gli aspetti formali e implementativi;
- Una parte più metodologica dove: si visiteranno alcune casi di studio nell'ambito del marketing, del supporto alla gestione clienti e dell'evasione fiscale. Questi saranno la base per la realizzazione di tre progetti che costituiscono la valutzione del corso.
- L'ultima parte del corso ha l'obiettivo di introdurre gli aspetti di privacy ed etici inerenti all'utilizzo di tecniche inferenza sui dati e dei quali l'analista deve essere a conoscenza.
- Una lezione (se c'è tempo) su un tema avanzato: Opinion Mining

Modalità di valutazione

- Progetti (Analisi dei dati): 60%
  - 3 Progetti: Si dovranno fare gruppi da due-tre. Gli studenti di un gruppo riceveranno lo stesso voto. La divisione del lavoro è loro responsabilità. I progetti, corredati di relazione, debbono essere presentati con relazioni scritte. Per ogni progetto sono previste sempre due fasi: esplorazione e data preparation ed analisi
  - Discussione orale sui progetti

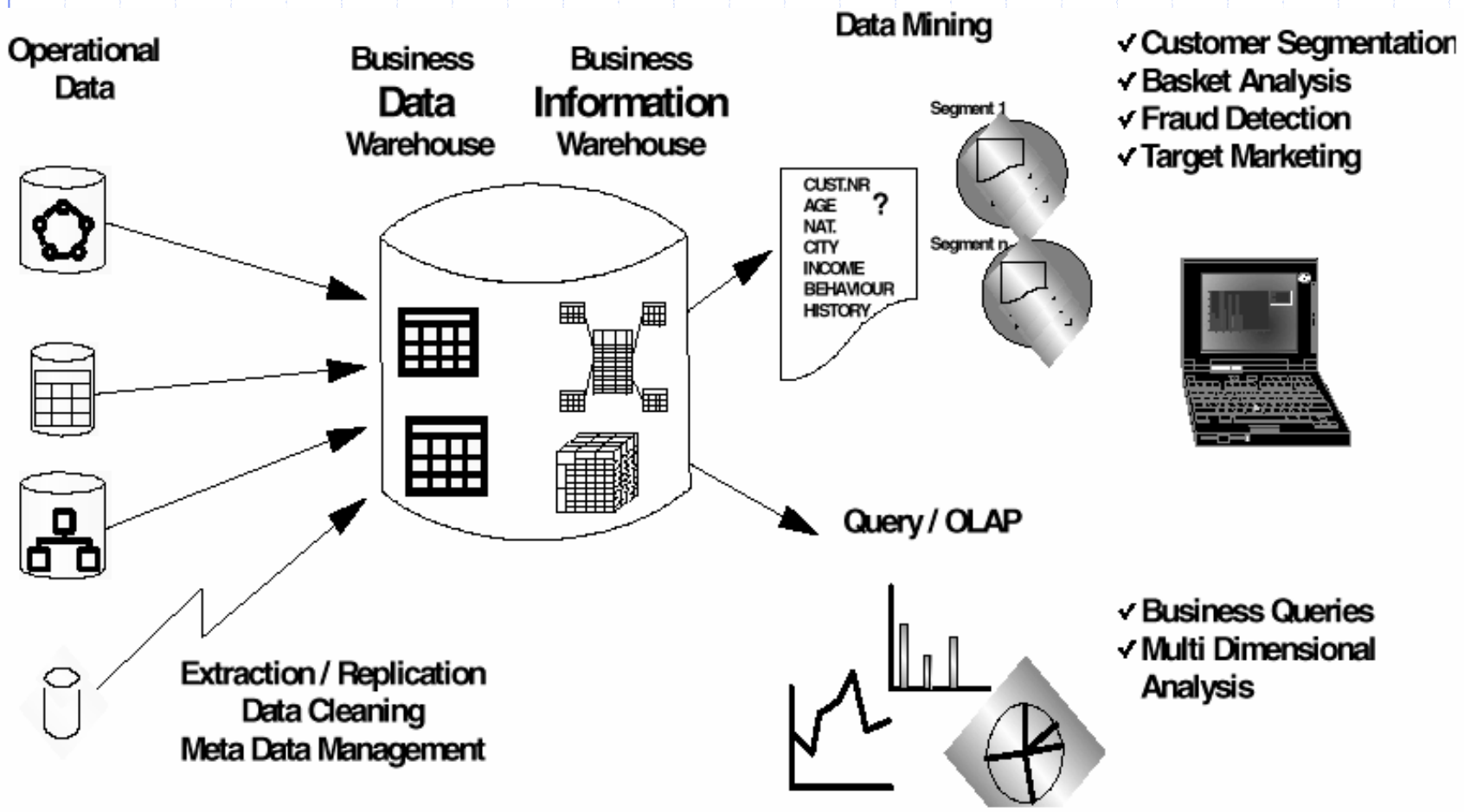


Lezione 1- Data Mining 2010-2011

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

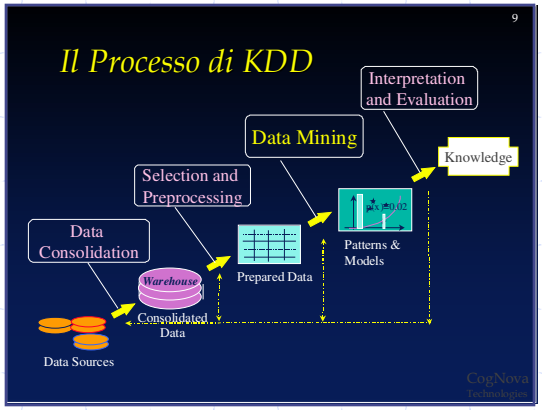
http://www.db2mag.com/98fsaar.html

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.

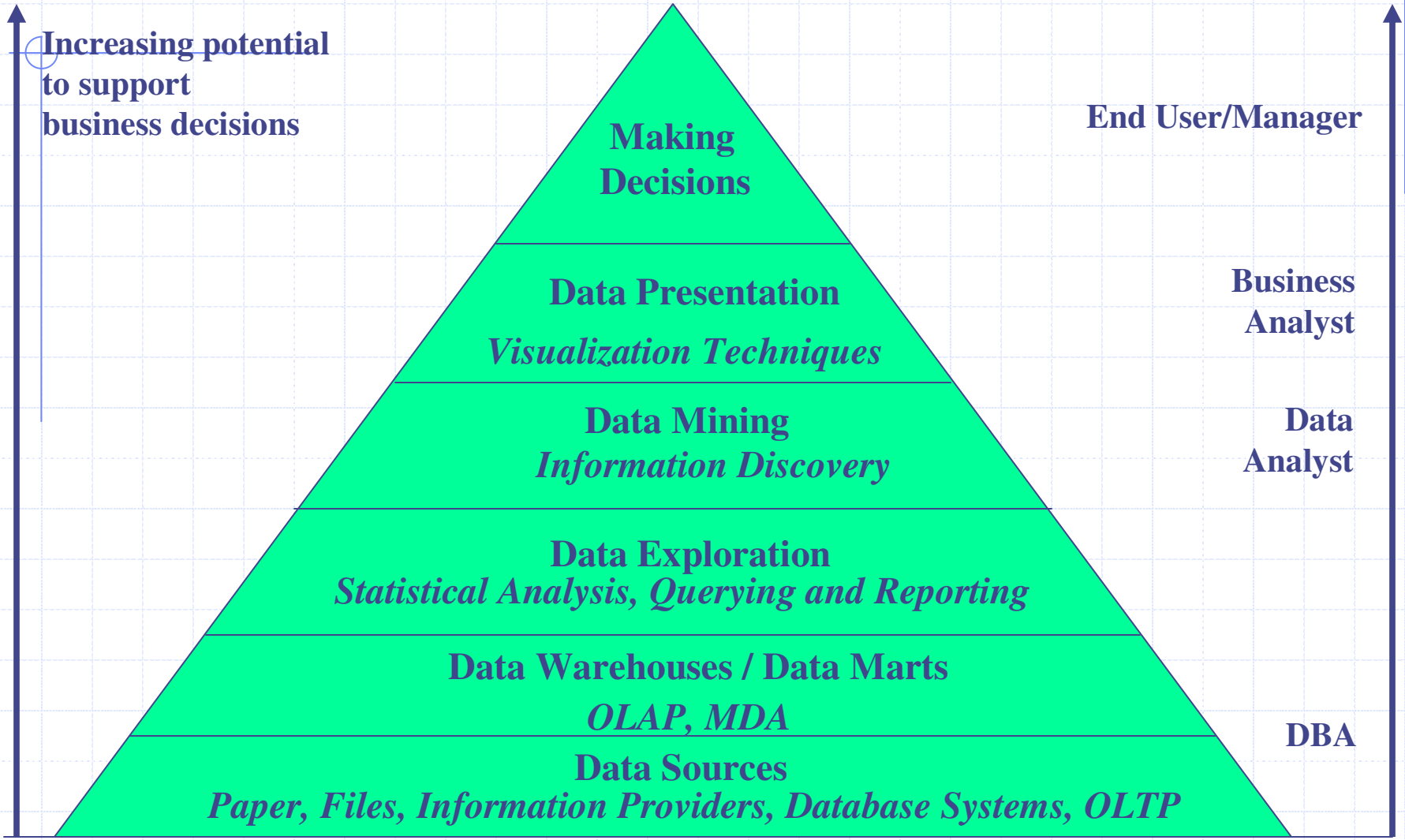
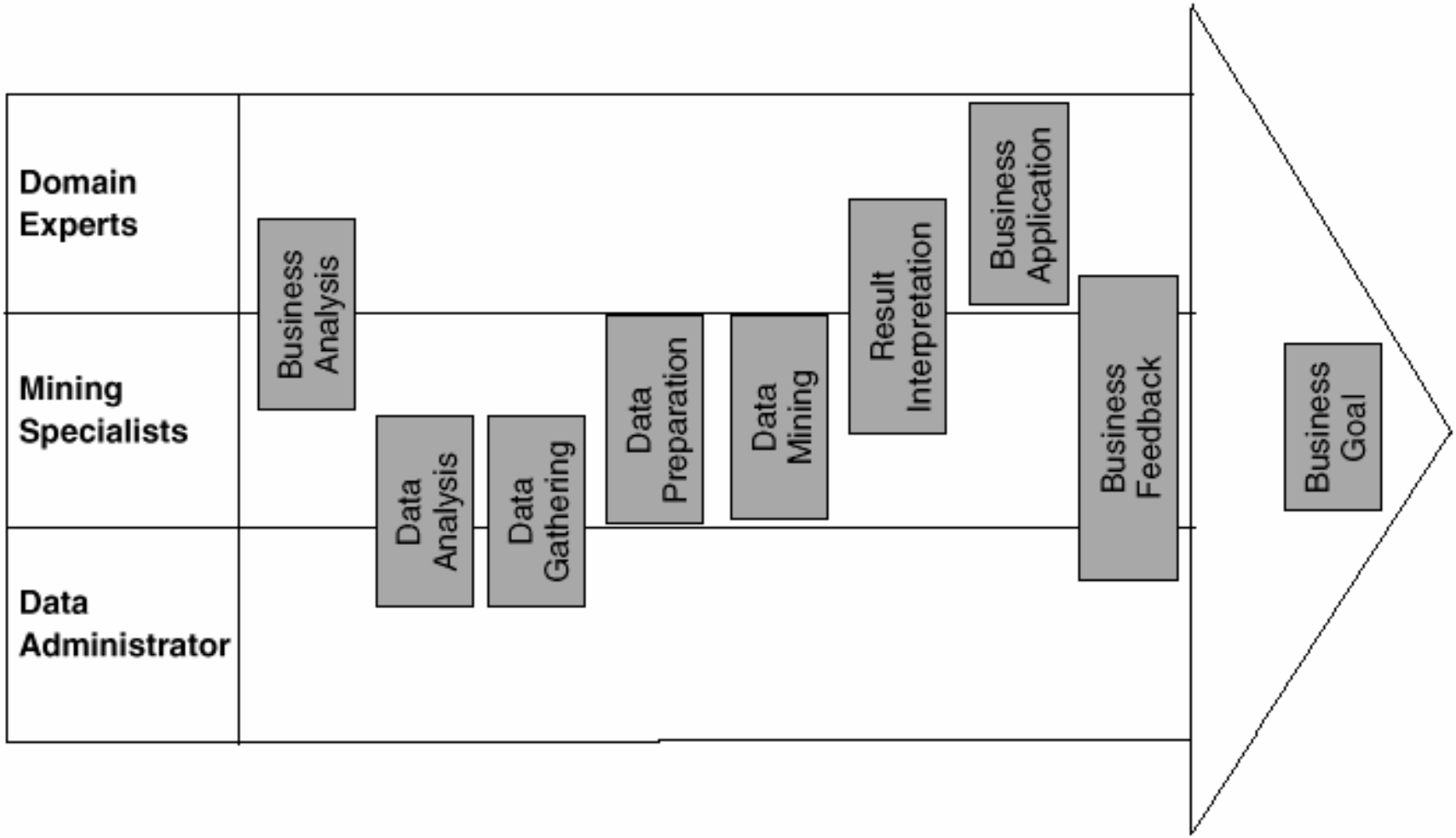


Figure nel processo di KDD



Intelligence/Value

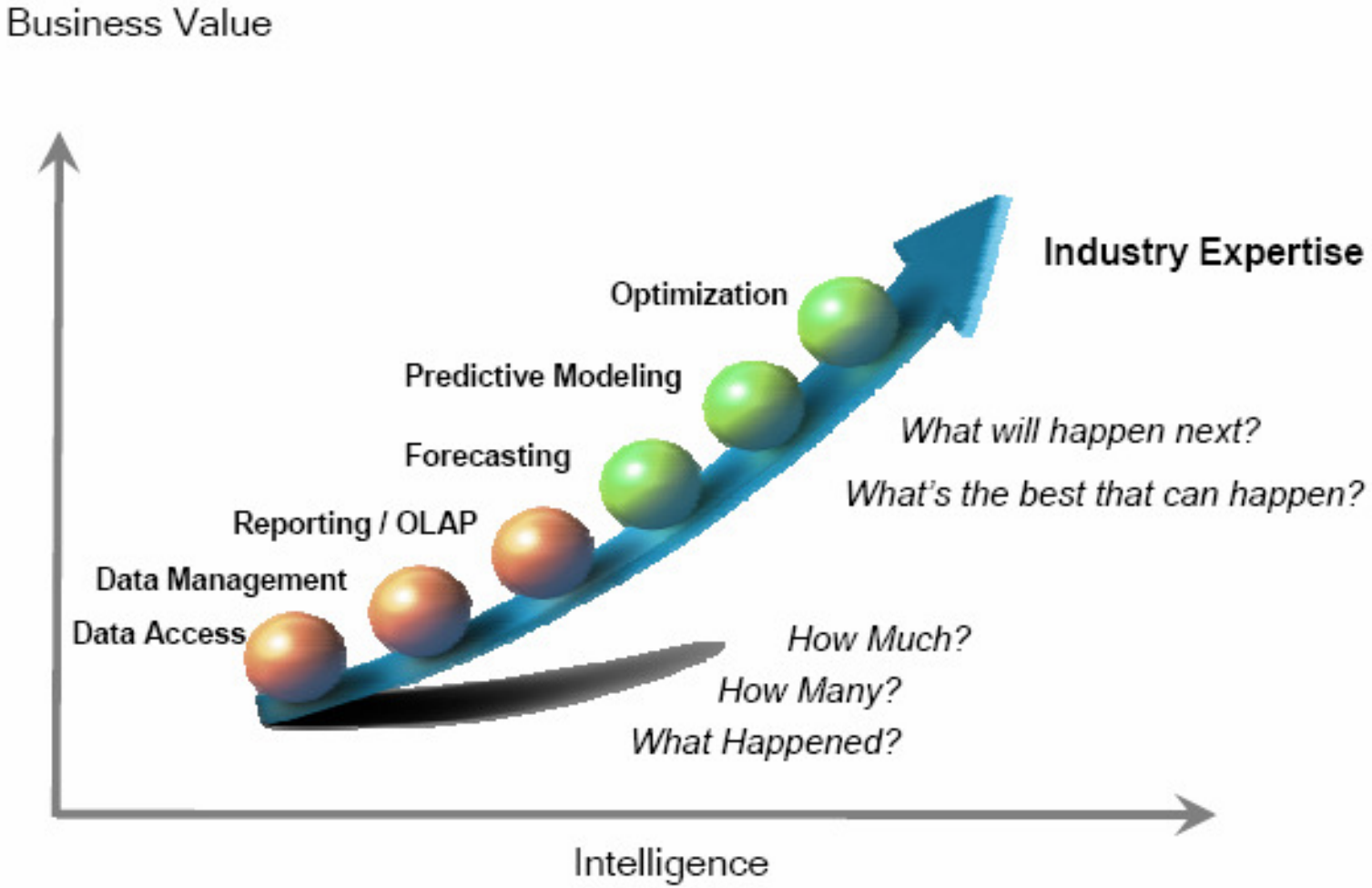
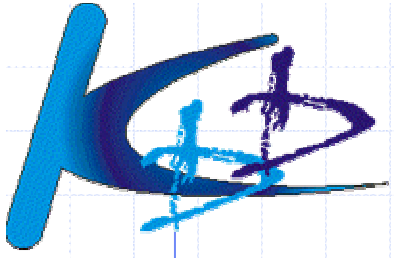


Figure 1: Business value increases exponentially with intelligence.



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

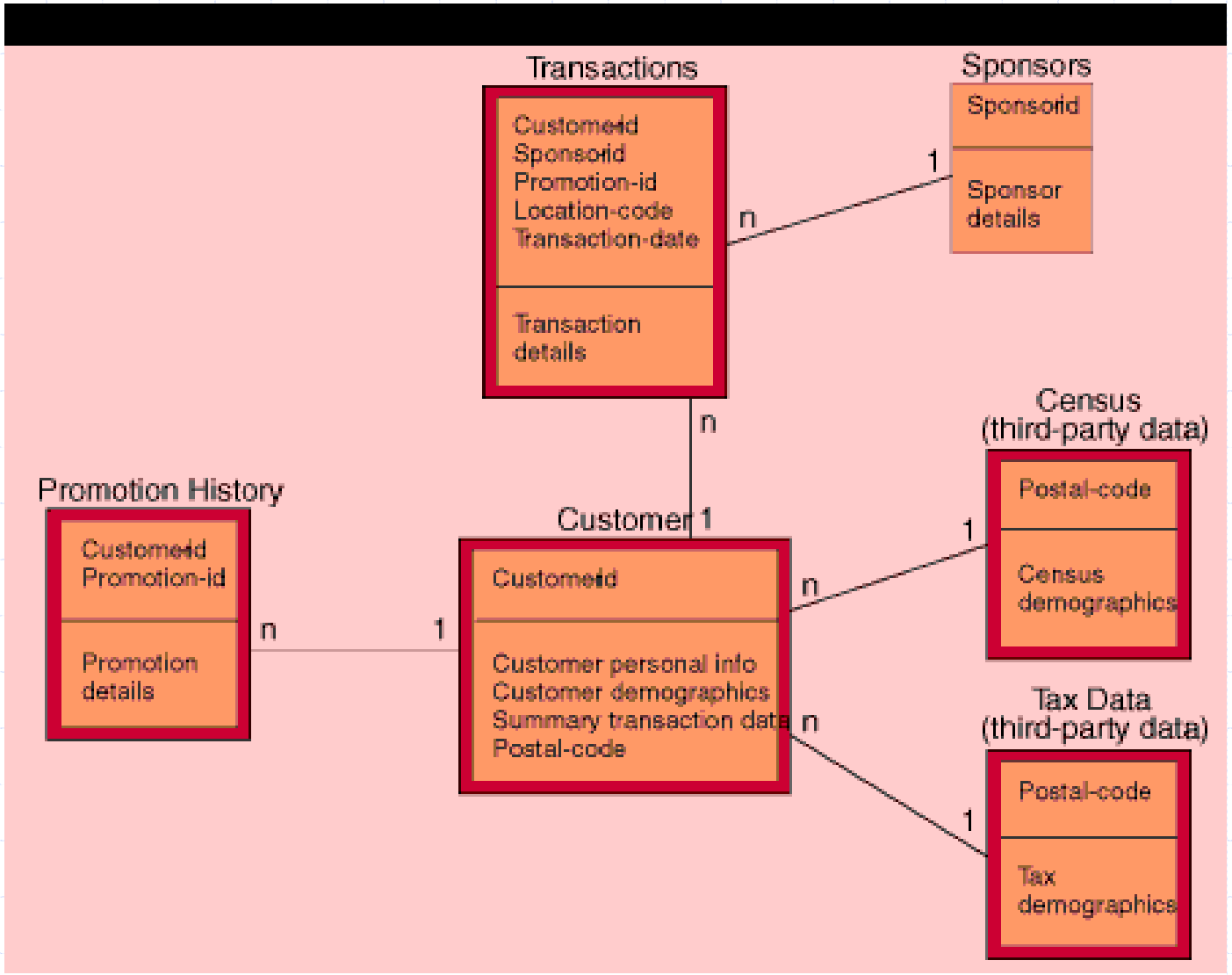
Stato dell'arte prima del data mining

- 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!

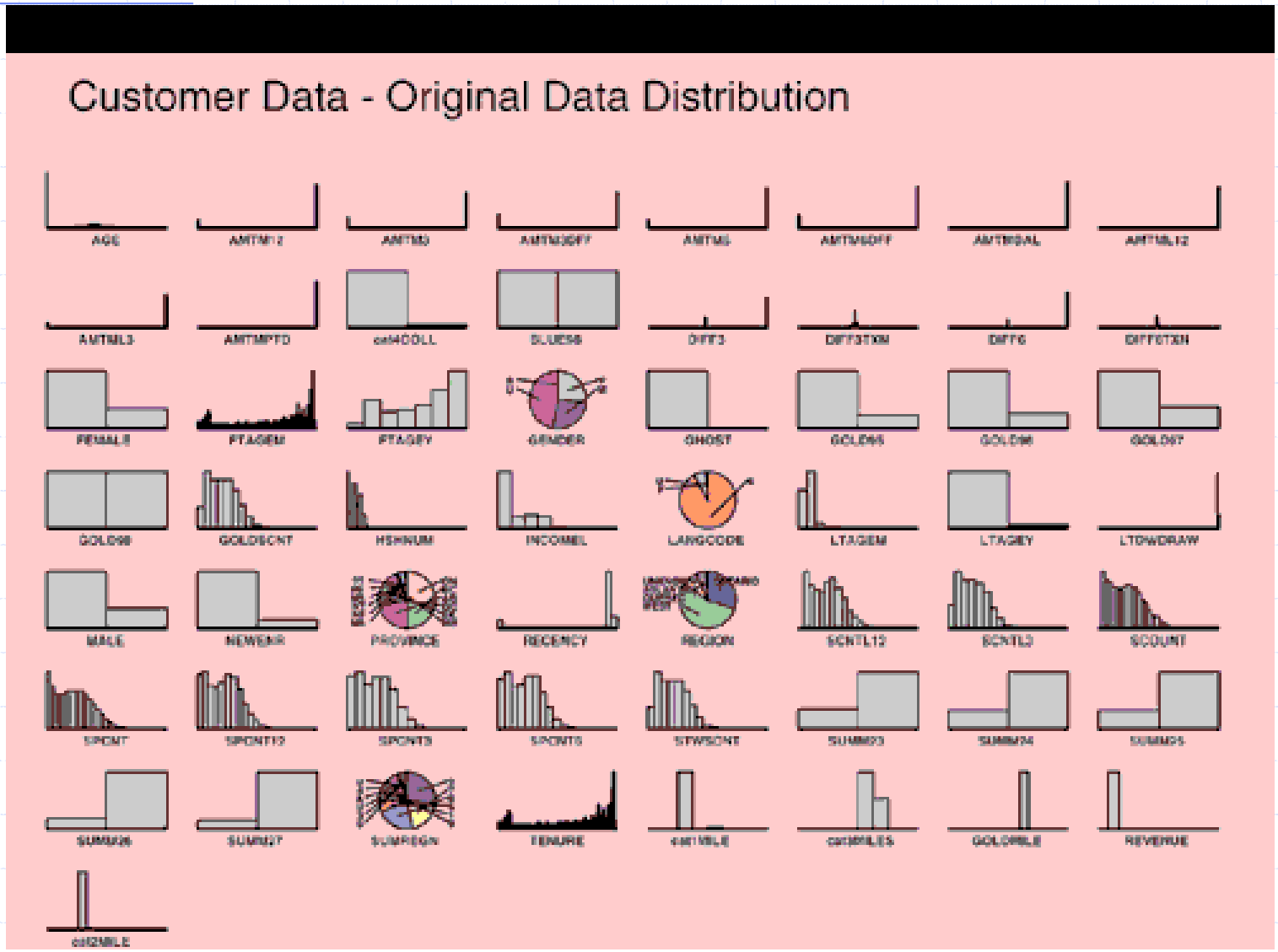
Sorgente dei dati nel DW



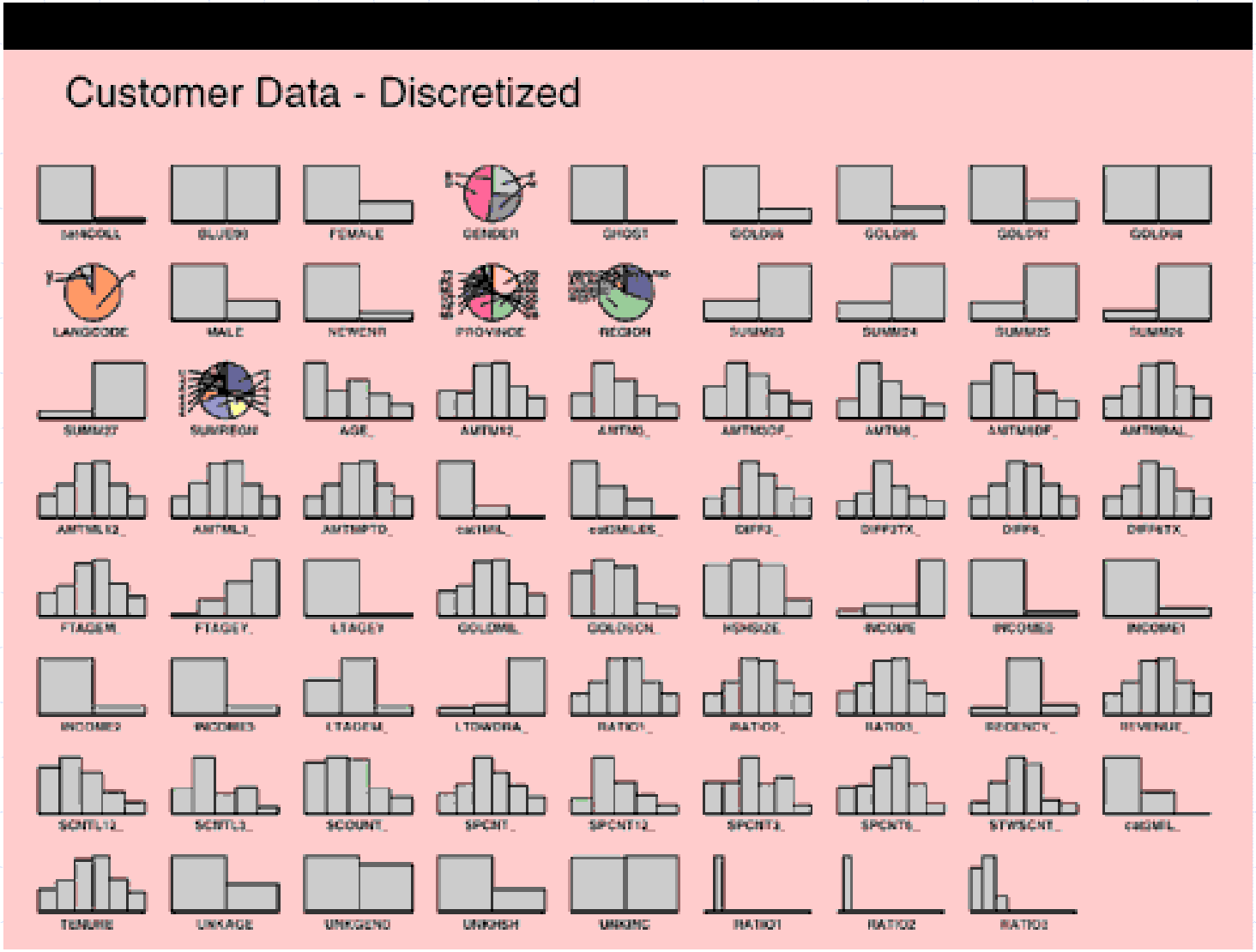
Preparazione dei dati

- 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

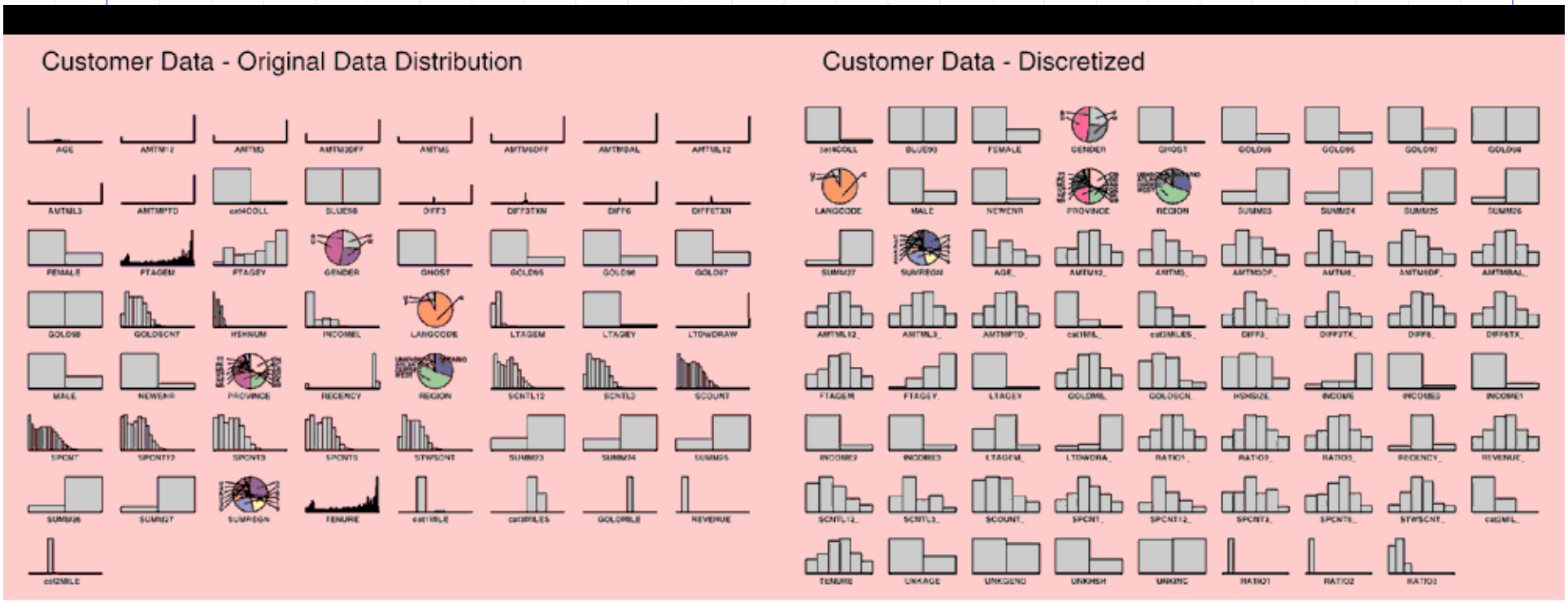
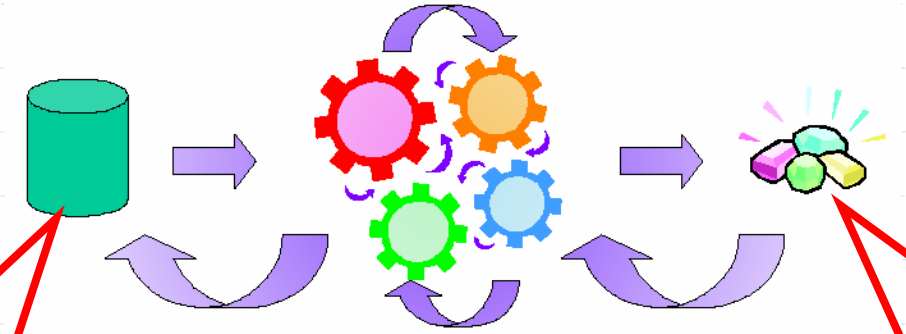


Figure3. *Original data.*

Figure 4. *Discretized data.*

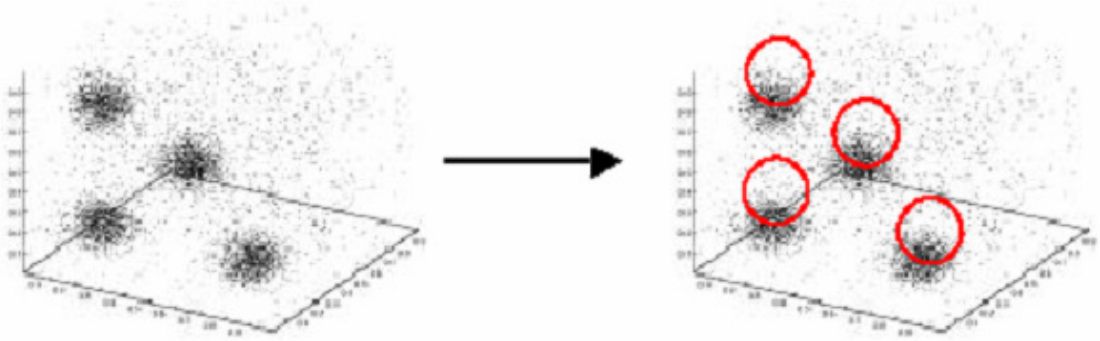
Estrazione del modello di clustering

Clustering = raggruppamento di oggetti simili in gruppi omogenei

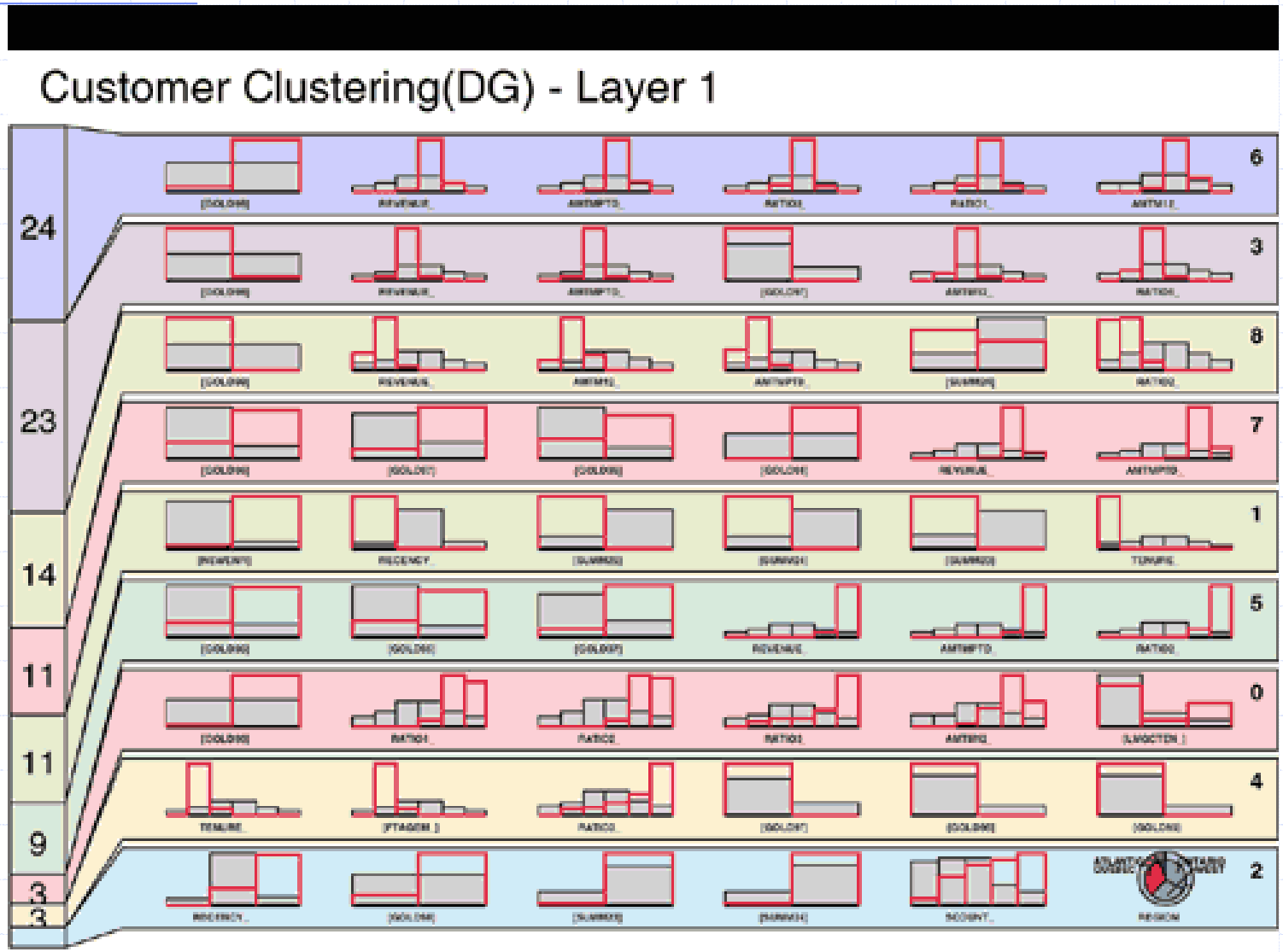


Dati in input: variabili economiche di ciascun cliente

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

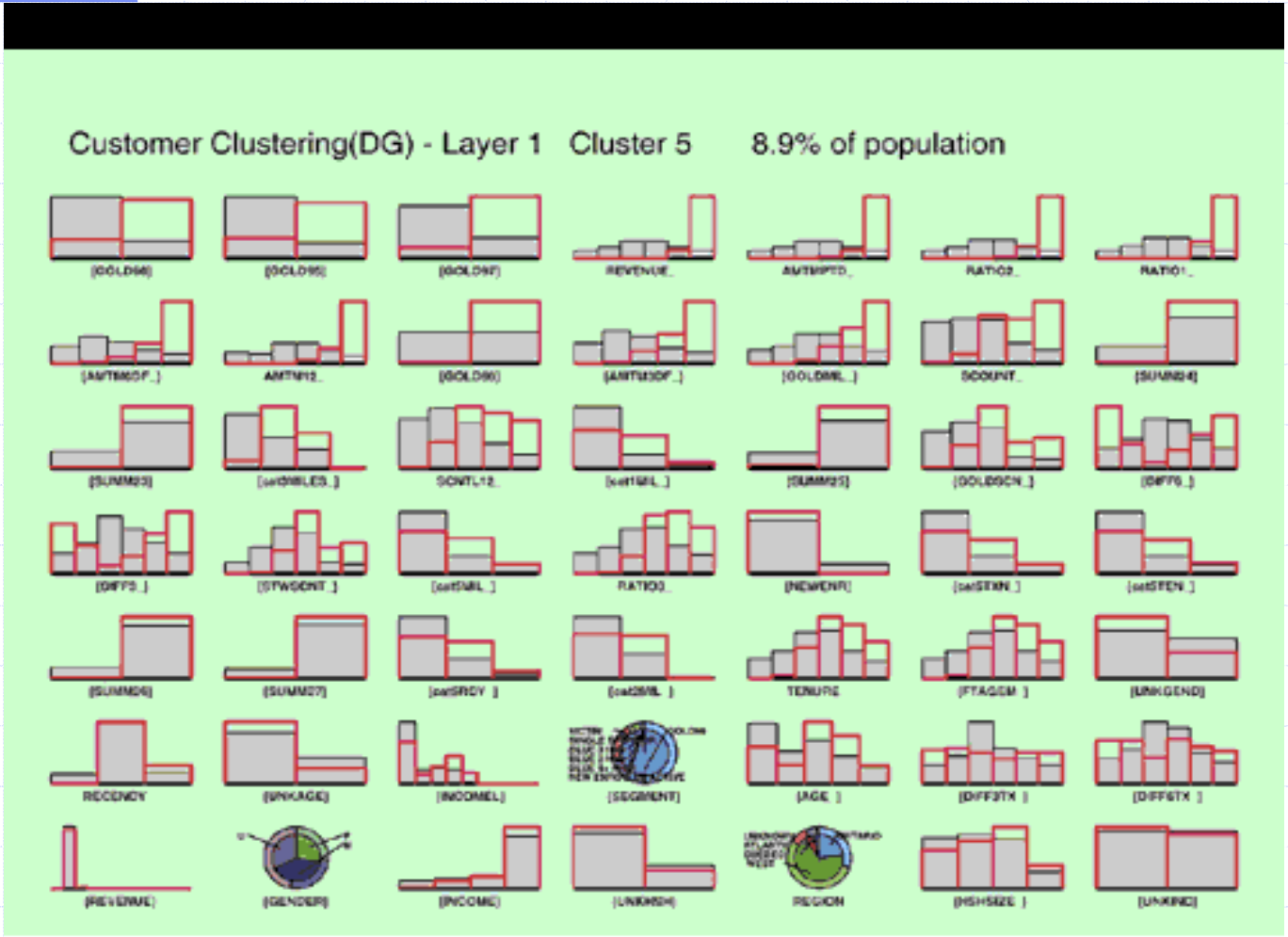


Figure 8. *Cluster 5 output.*

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.

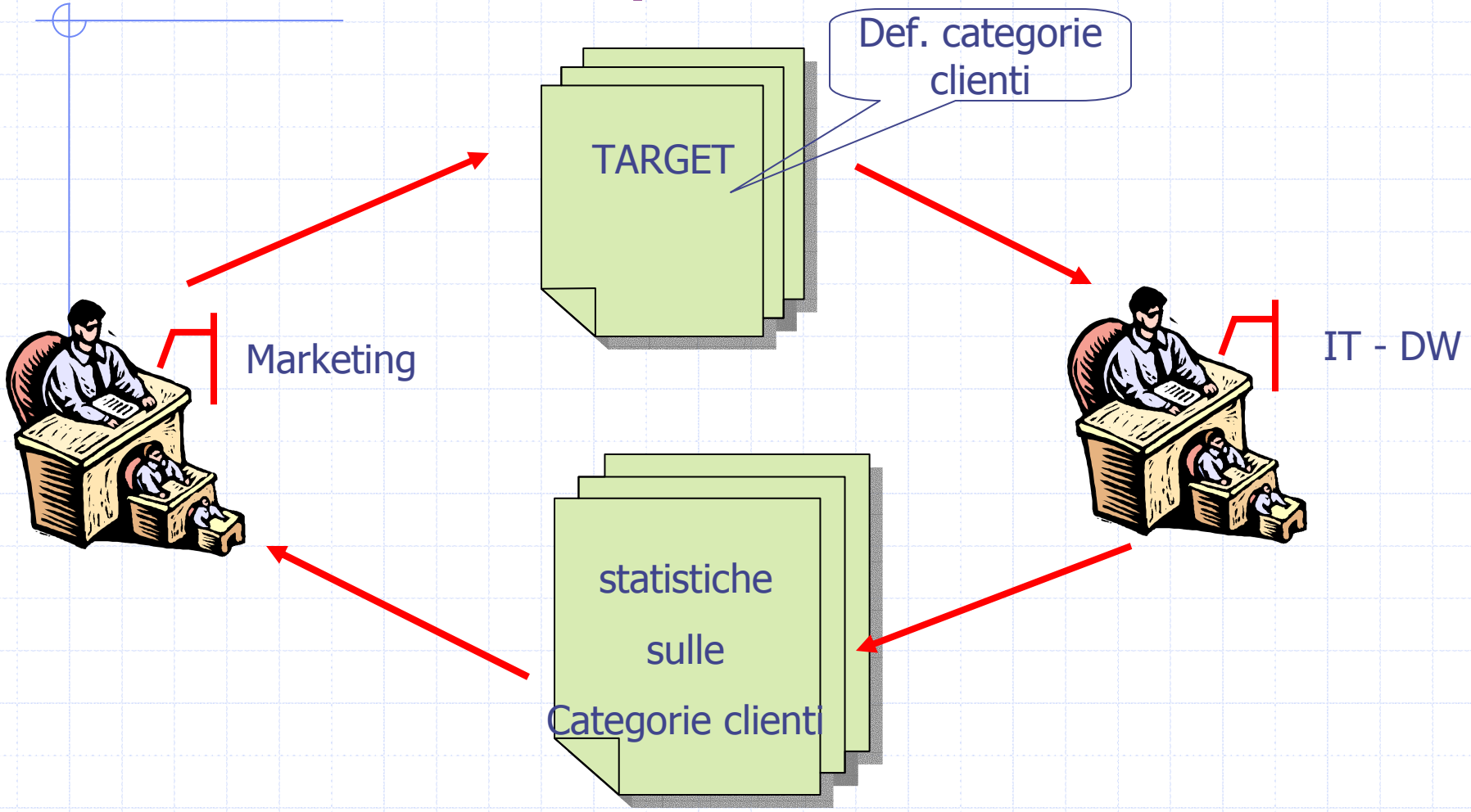
Analisi previsionale per l'ottimizzazione della postalizzazione delle promo

KDD Lab. Pisa

Postalizzazione di promozioni

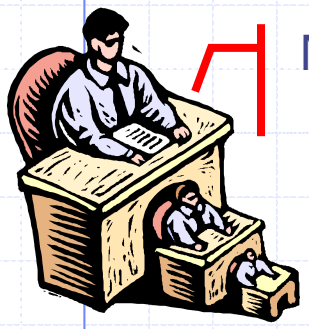
- Il processo decisionale:
  - Inventare la promozione
  - Selezionare il target
  - Contattare il target
  - Consegnare i premi
  - Tenere traccia dei redenti
  - Valutare a posteriori l'efficacia intervento
- Gli attori
  - Ufficio Marketing, Ufficio IT/DW, Postalizzatore, Ufficio IT/DW , Ufficio Marketing

Inventare la promozione



selezionare i clienti e postalizzare

Erogare premi e raccogliere dati



Marketing sviluppo

promozione

Def. dettagli promo

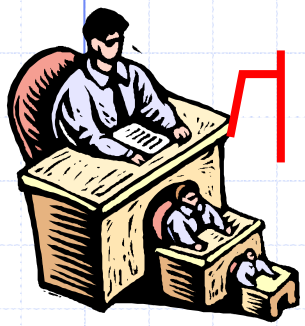
Sistema automatico aggiornamento DATA Mart

Sistema automatico attivazione mercati

DATA MART PROMOZIONI

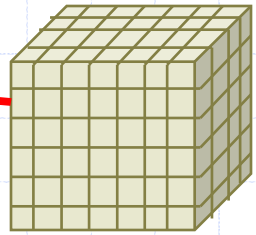
Mercati

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

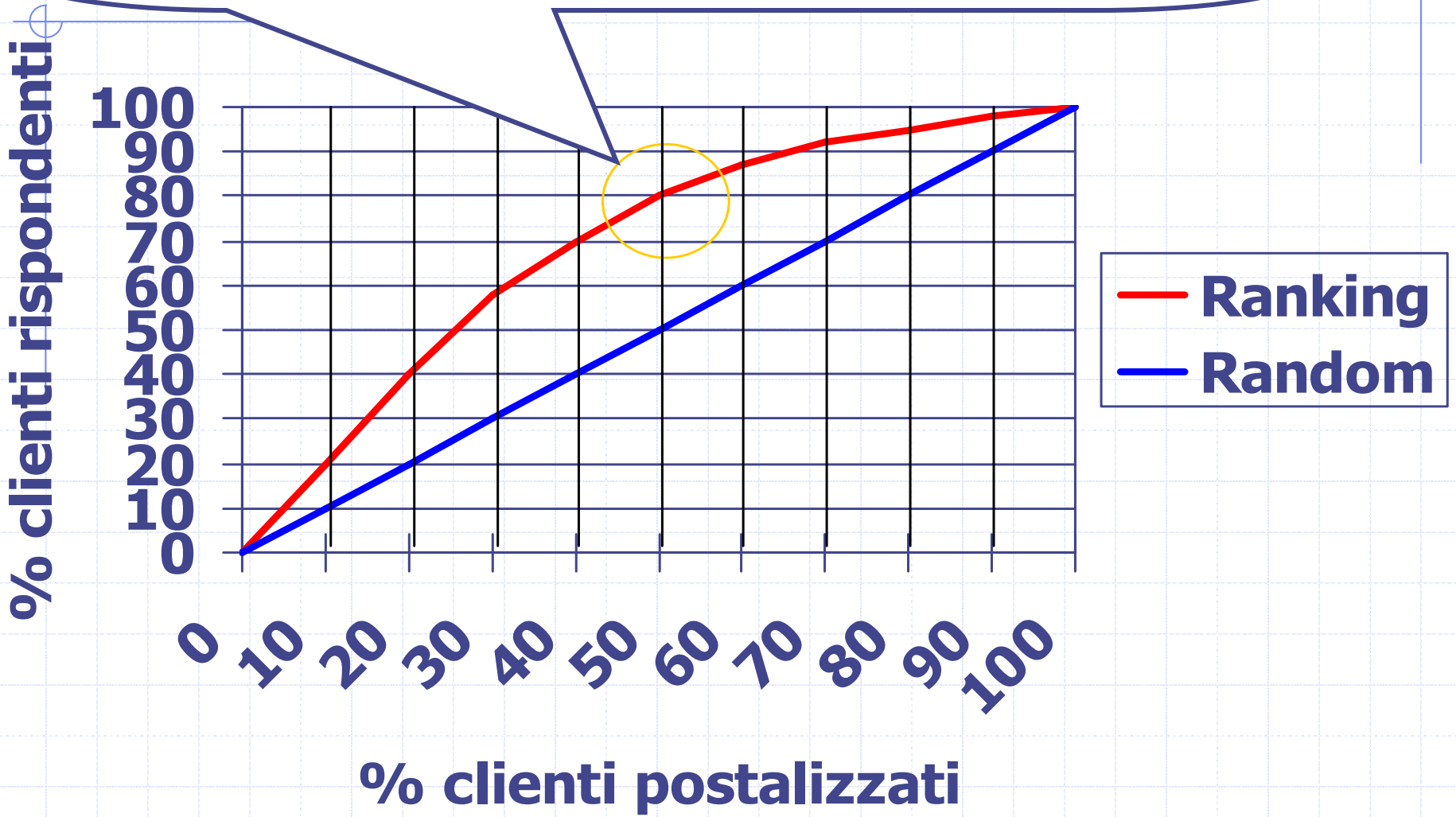
Lift Chart


% 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

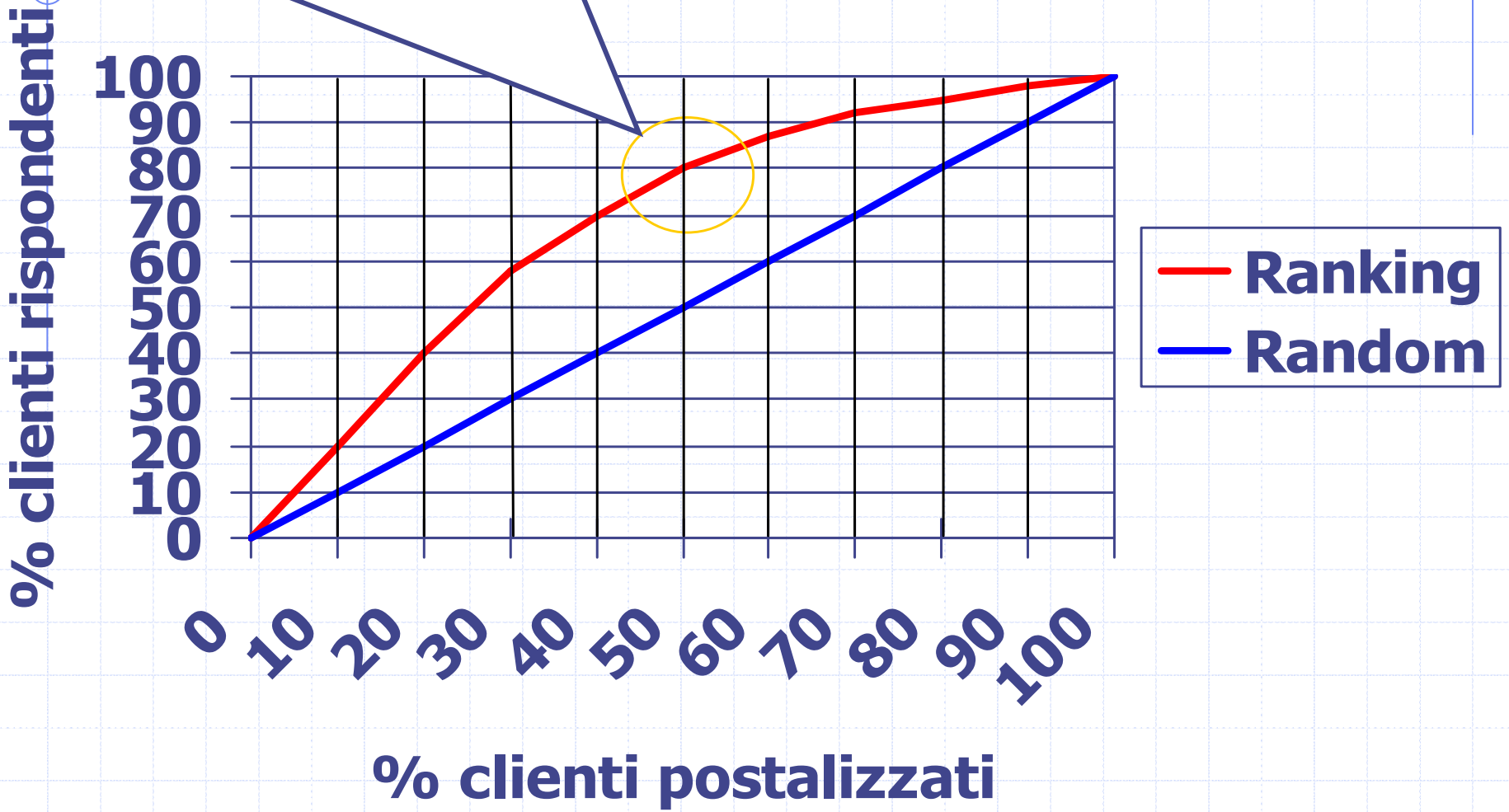
LIFT CHART

- 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)

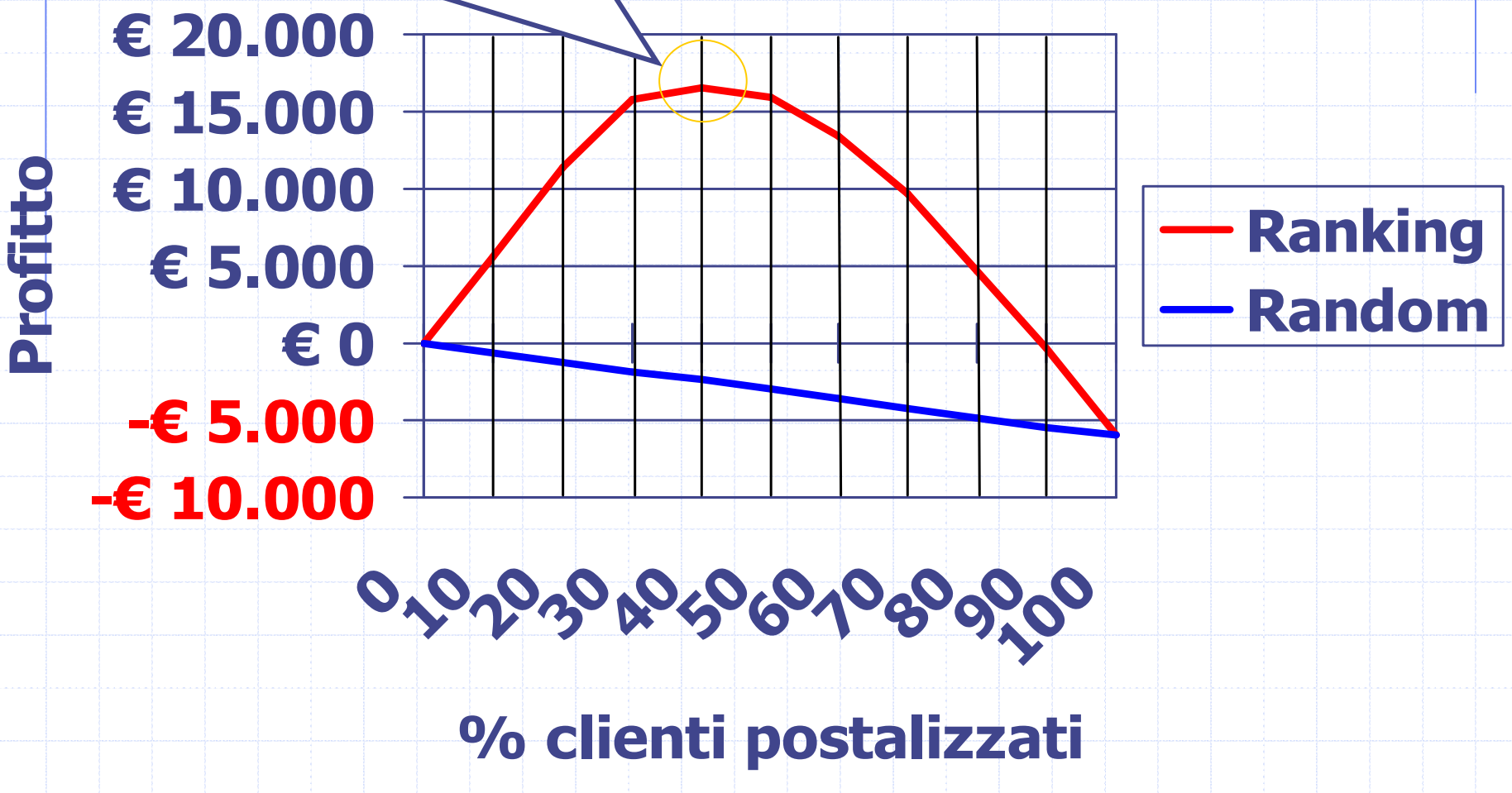
- Il Lift Chart rappresenta un aiuto grafico per ragionare sul rapporto ottimale fra costi di postalizzazione e percentuale di redemption
  - a fronte di sostanziali riduzioni di postalizzati (=budget) permette di ridurre di poco il numero di redenti
  - a parità di budget, permette di incrementare il 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]iesta dell[illegible]
  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

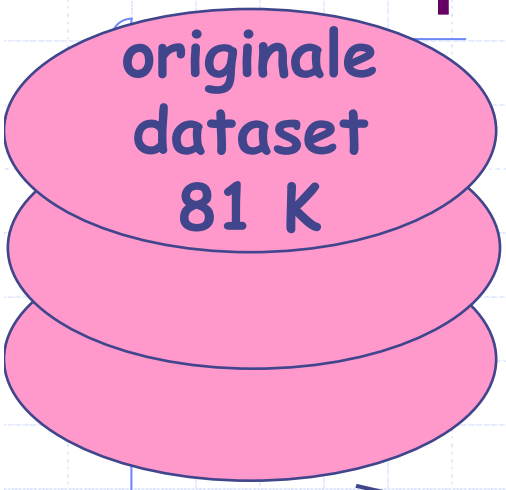
Pianificazione di verifiche

- 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

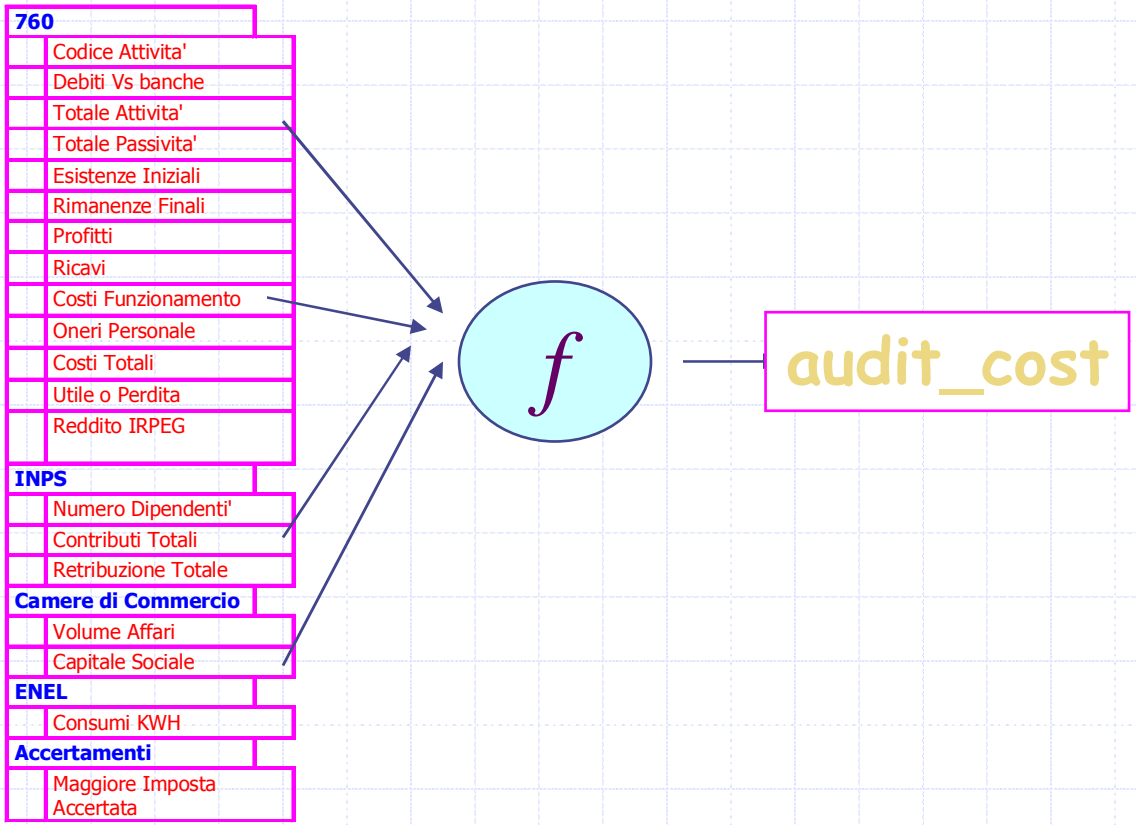


data consolidation
data cleaning
attribute selection

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

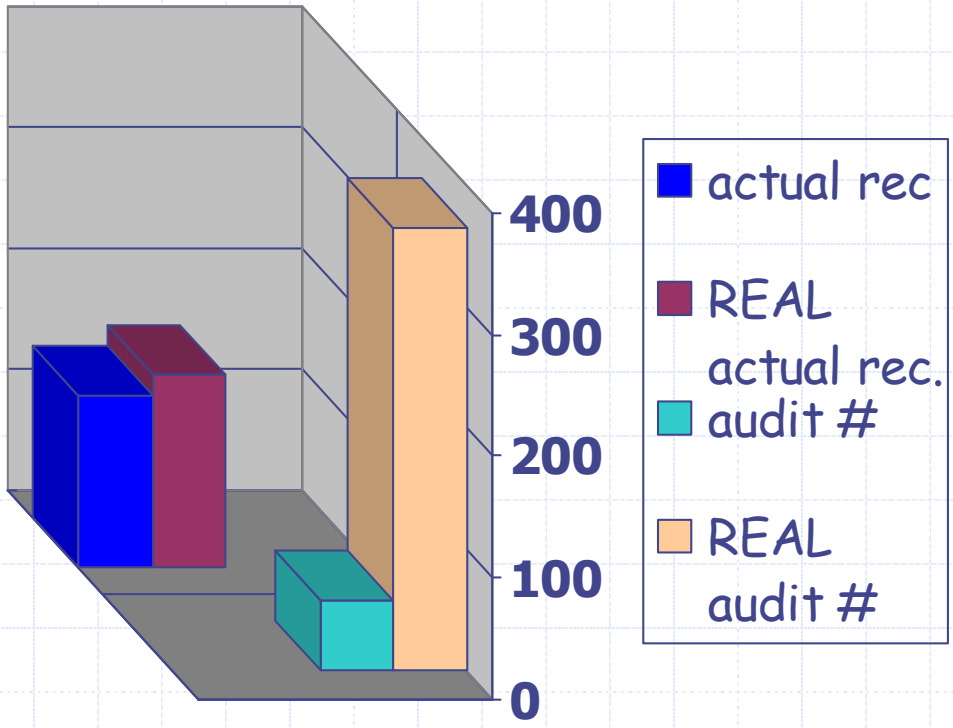
- audit # (di un dato classificatore): numero di tuple classificate come positive = # (FP ∪ TP)
- actual recovery: ammontare totale del recupero effettivo per tutte le tuple classificate come positive
- profitability: recupero effettivo medio per verifica
- relevance: rapporto tra profitability e l'errore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

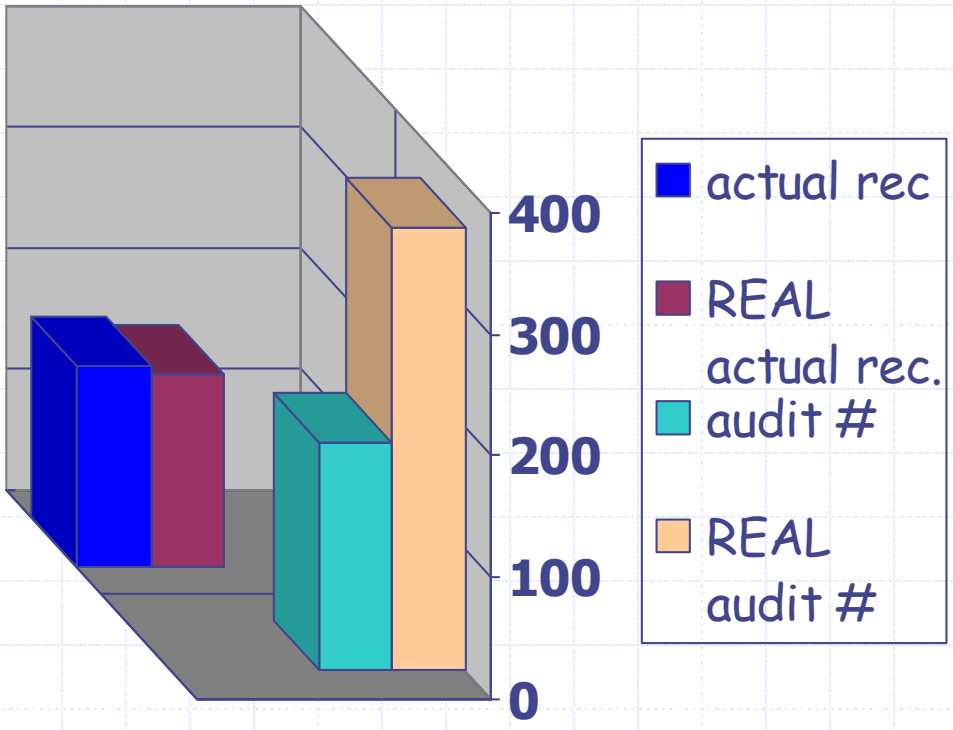
Classificatore 1 (min FP)

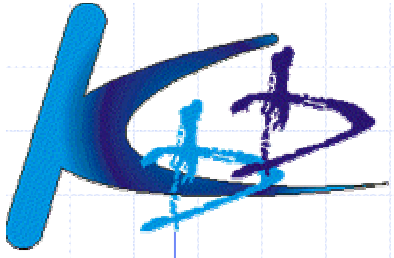
- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.*= 141.7 Meuro
- *profitability* = 2.401



Classificatore 2 (min FN)

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



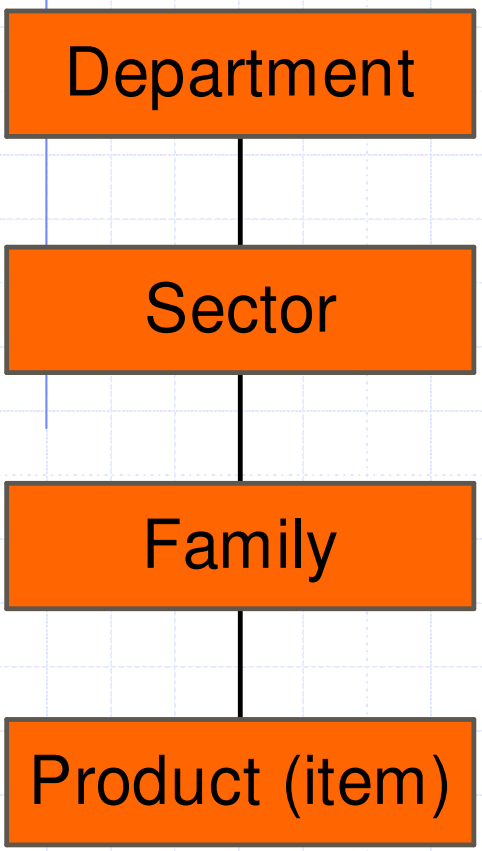


Market Basket Analysis presso la COOP

DataSift e COOl Patterns

KDD Lab. Pisa

Datasift – COOP (’96-’99)



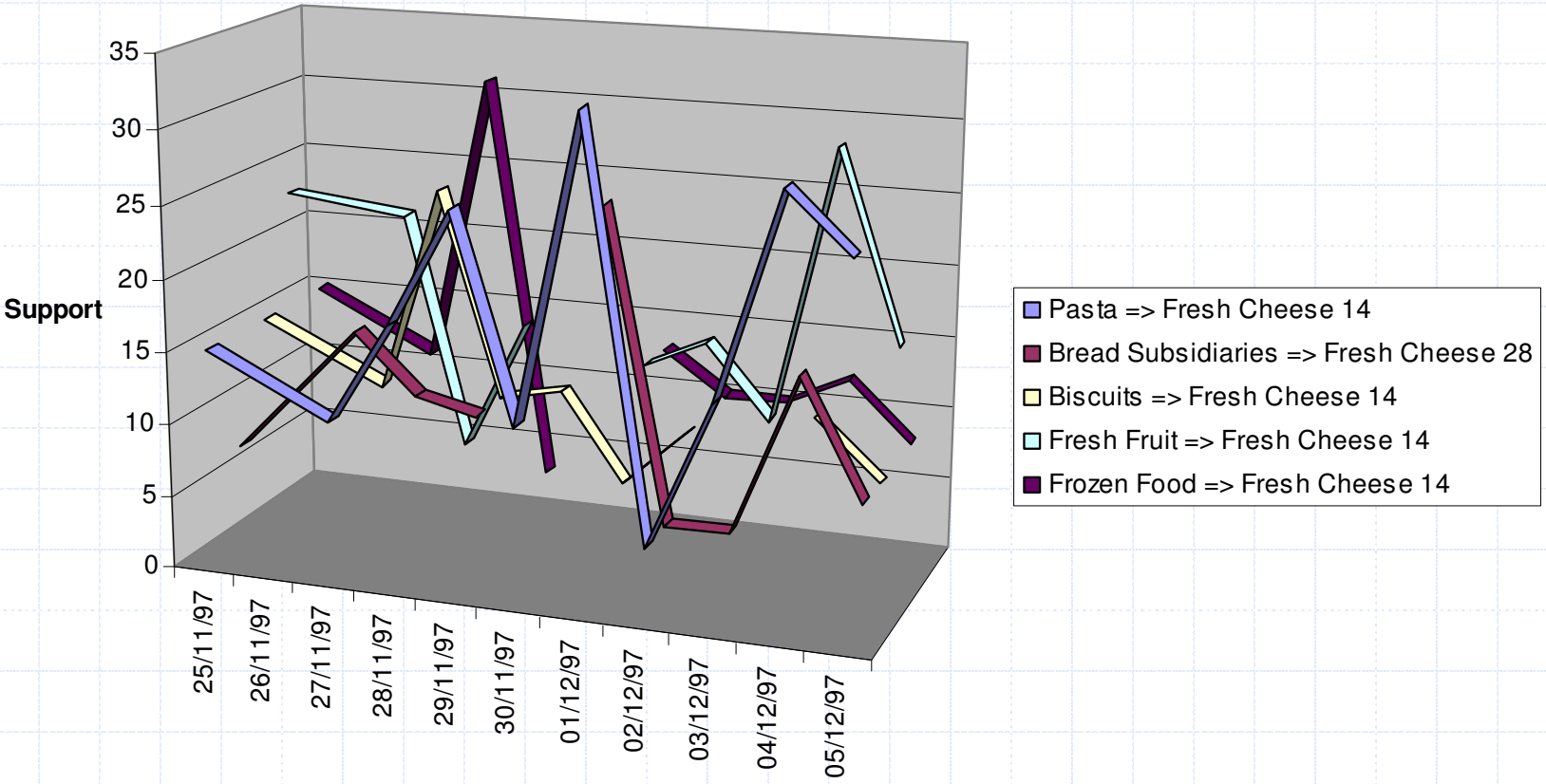
- Progetto pionieristico di Market Basket Analisys a partire da dati di vendita (**scontrini**)
- Estrazione di **regole associative**
- Ragionamento sulle regole estratte ai diversi livelli della **gerarchia dei prodotti**
- Studio dell’effetto delle **promozioni** sulla dinamica temporale delle regole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



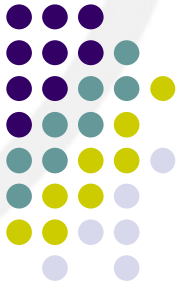
COOL PATTERNS

Progetto "COOL PATTERNS"

Analisi delle vendite nella grande distribuzione

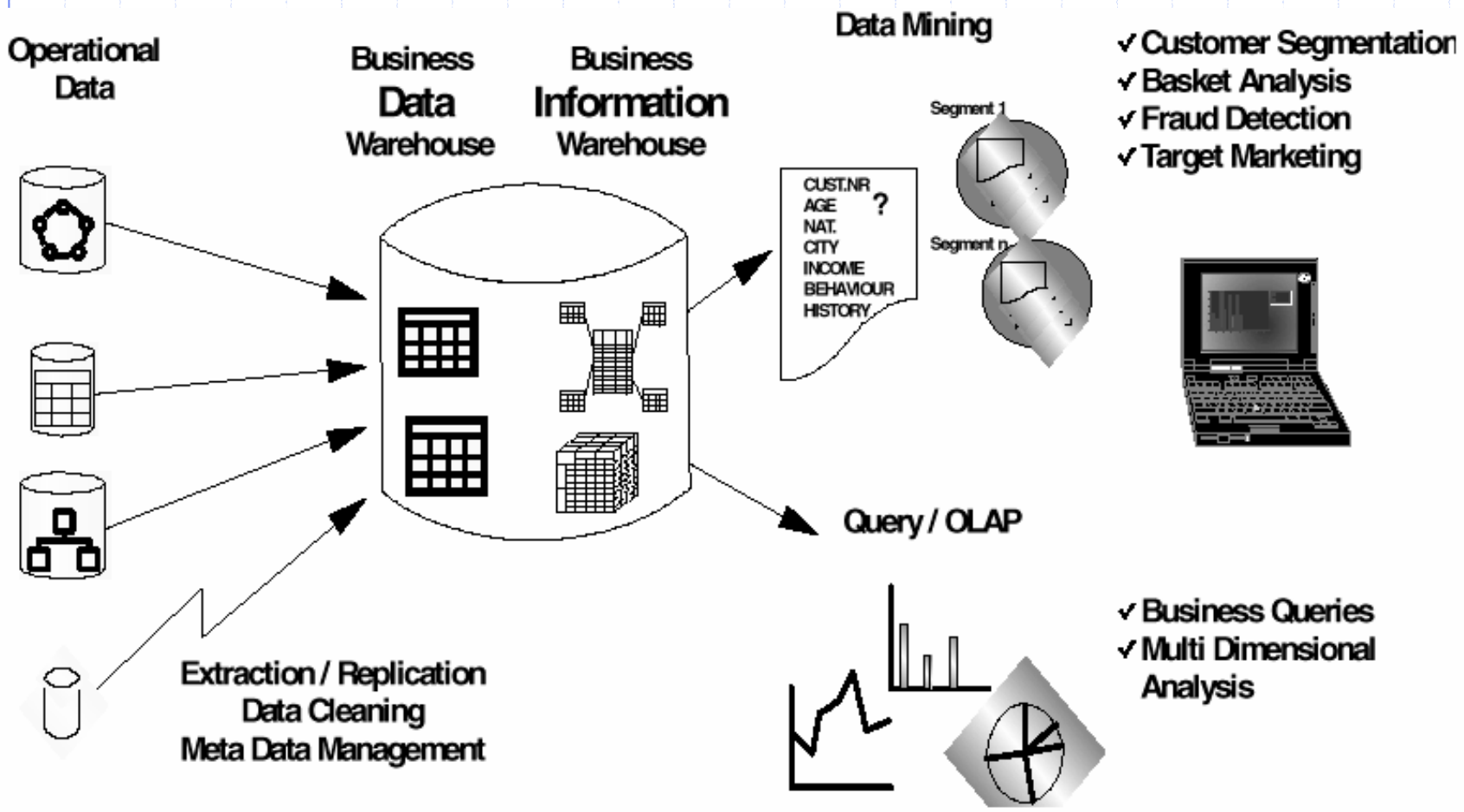
Analisi dei Dati ed
Estrazione di Conoscenza
2004/2005

Federico Colla



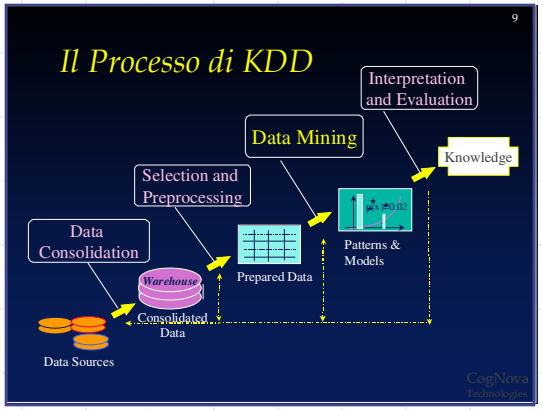
… per concludere, debrief!

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.

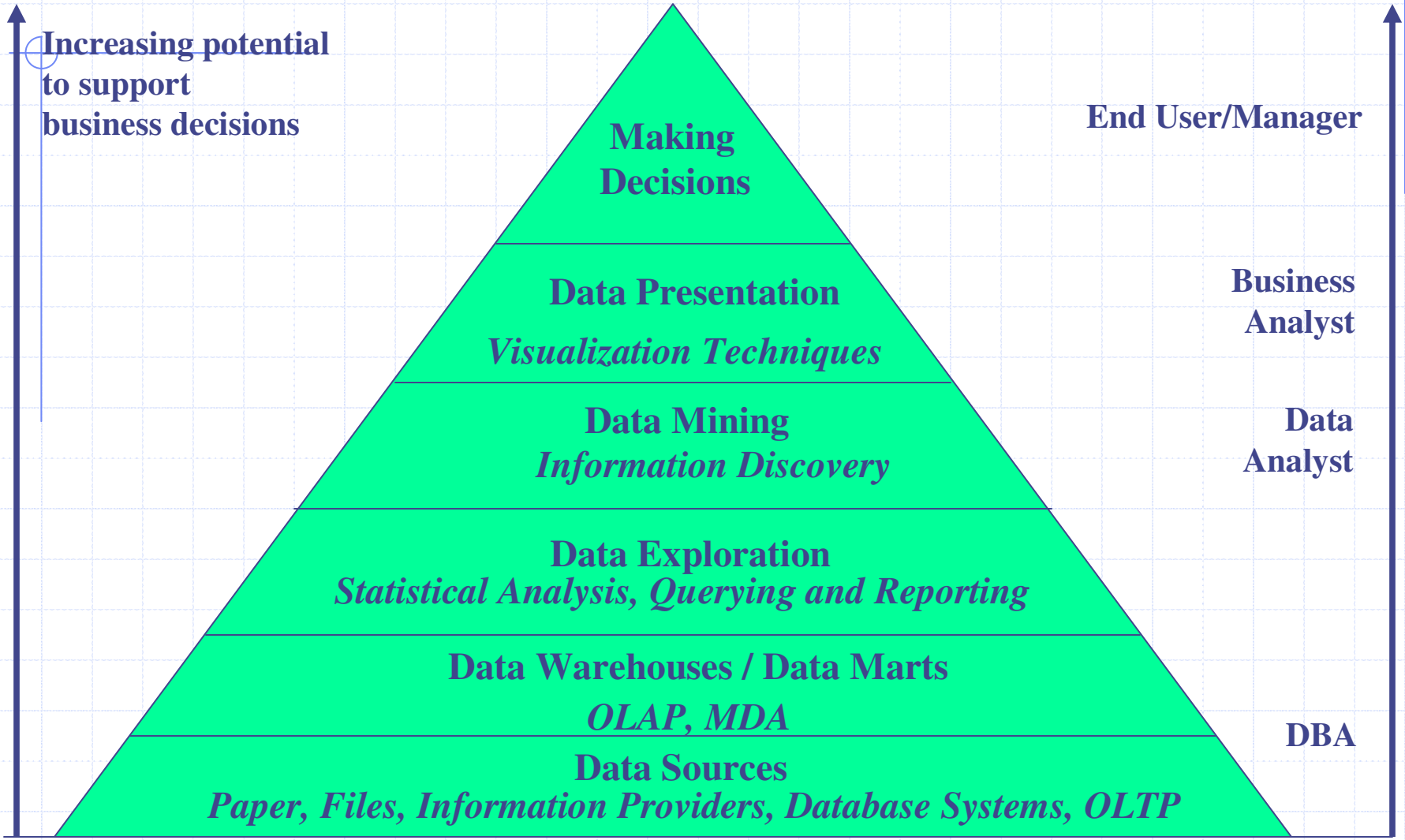
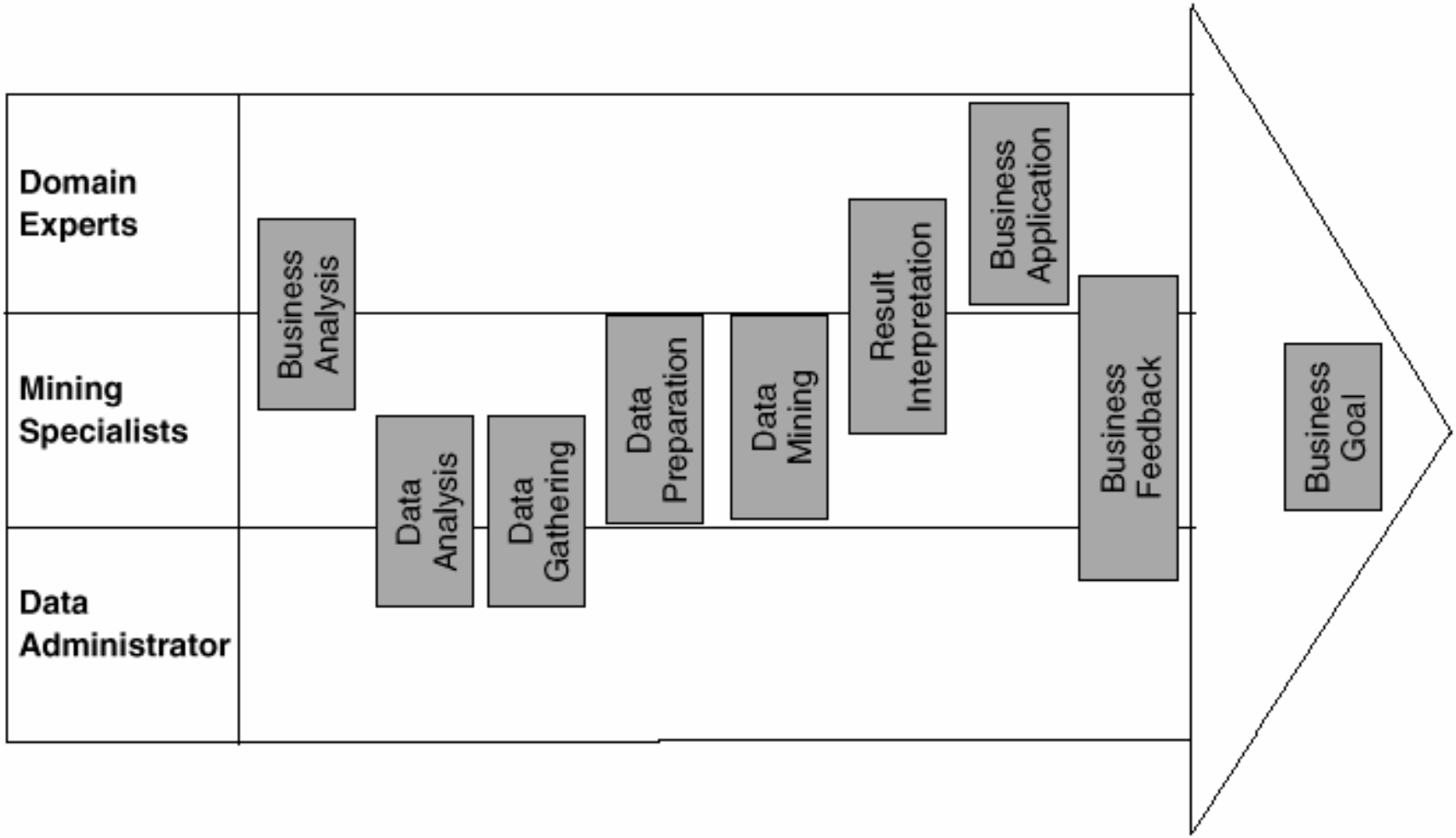


Figure nel processo di KDD



Intelligence/Value

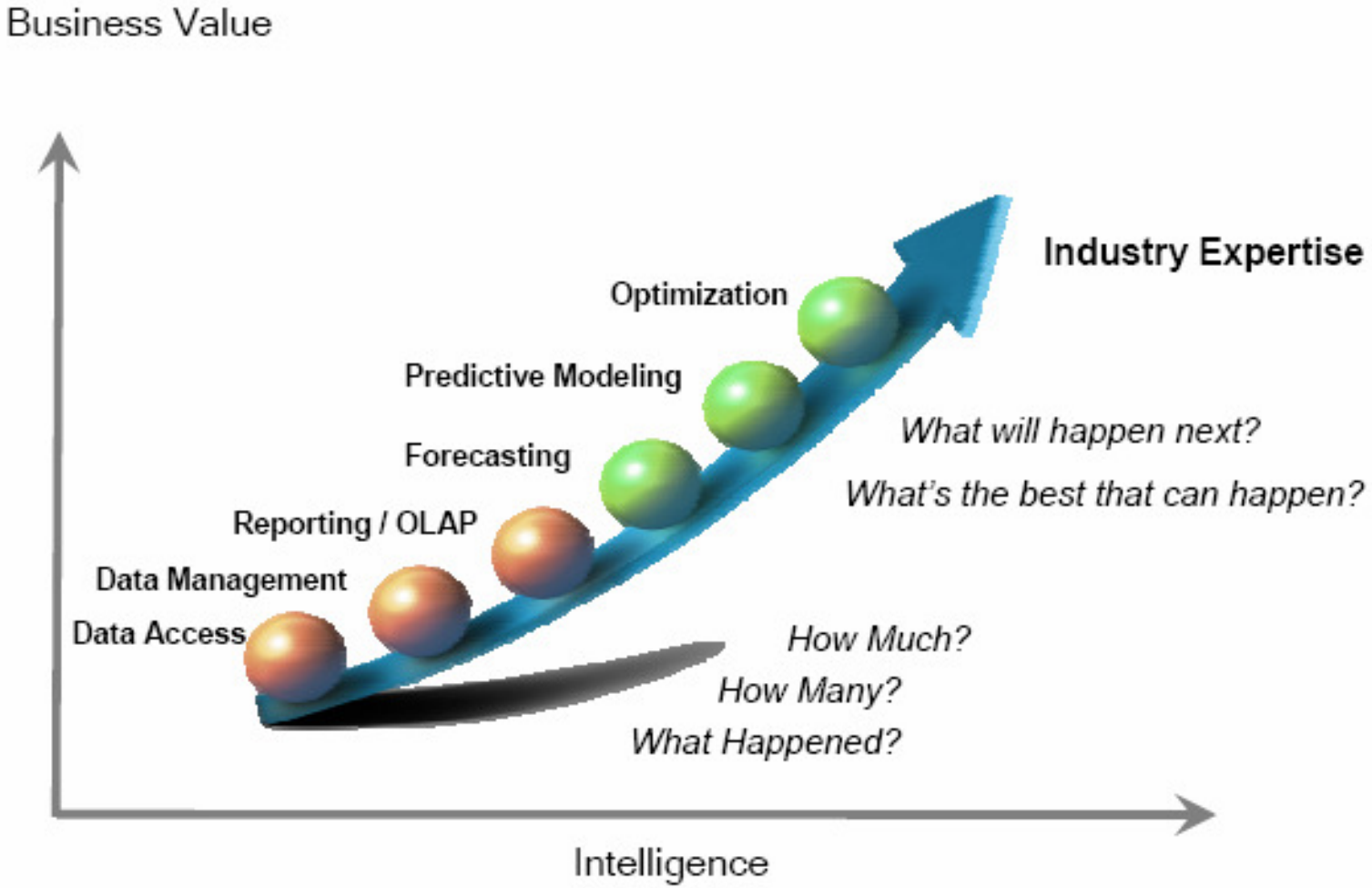


Figure 1: Business value increases exponentially with intelligence.

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.

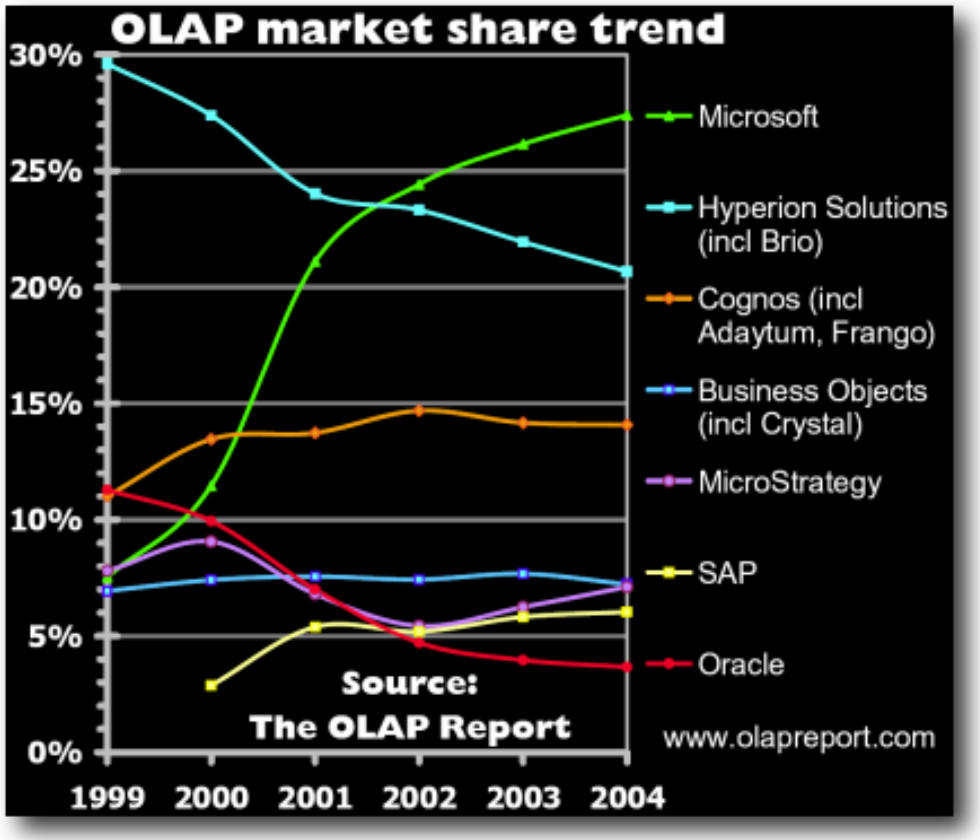
- Tecnici IT:
  - Da progettisti e amministratori DB
  - A progettisti e amministratori DW e creatori di cubi tematici
- Analisti (dei dati e del business)
  - Da estensori manuali di rapporti
  - A creatori di rapporti e cruscotti interattivi
- Utenti finali (manager in senso lato, ad ogni livello)
  - 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

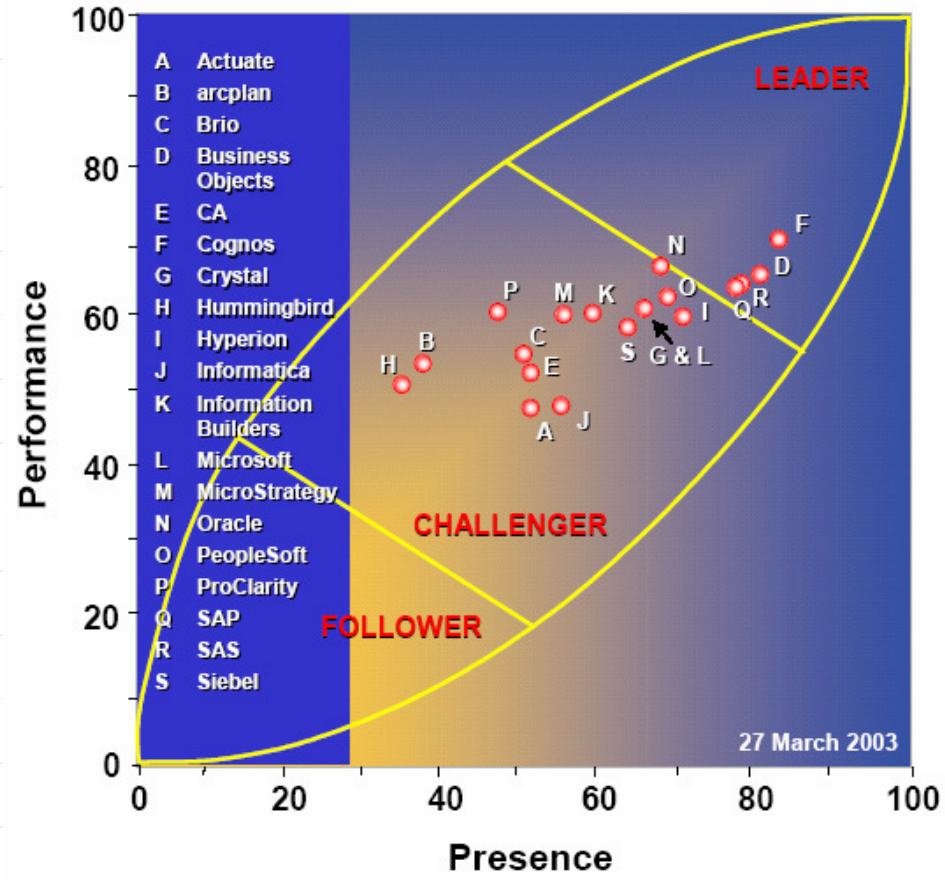
I principali vendor di B.I.

OLAP Market Share



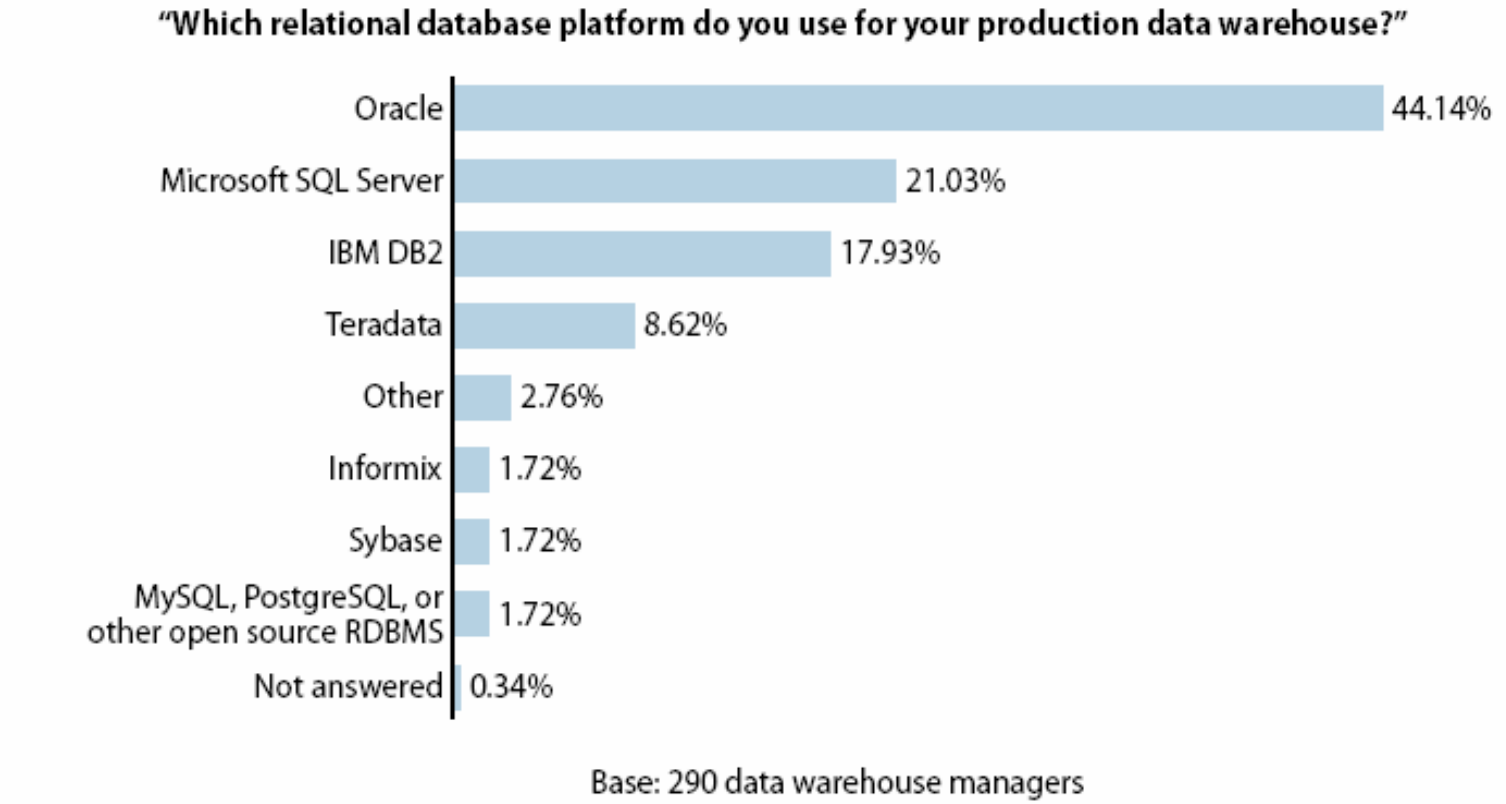
- Olap report: http://www.olapreport.com

Prodotti OLAP



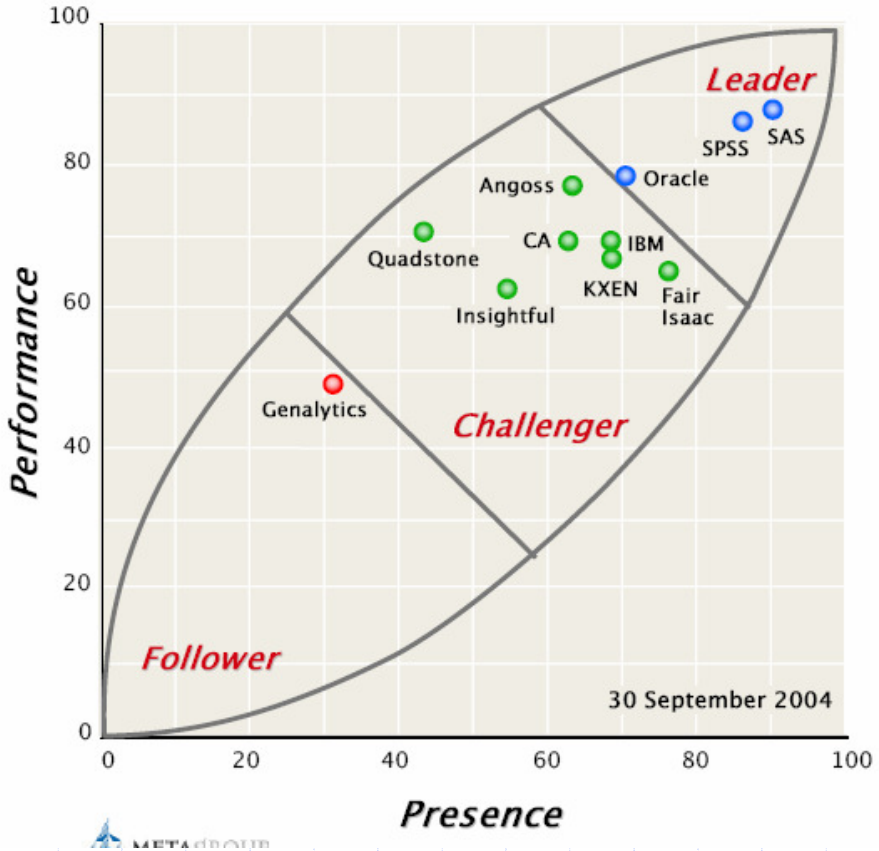
- METAspectrum evaluation 2003

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