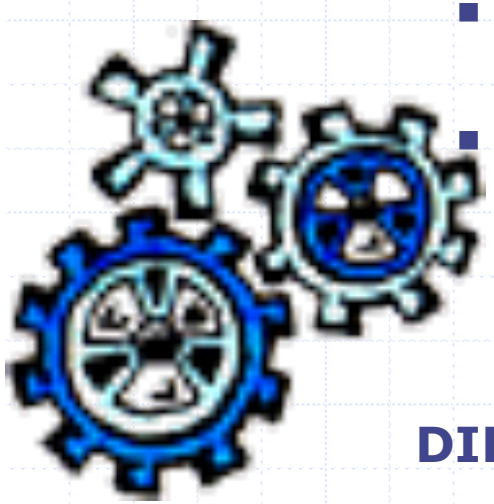
Data Mining & Machine Learning

- Fosca Giannotti, ISTI-CNR, fosca.giannotti@isti.cnr.it
- Dino Pedreschi, Dipartimento di Infomatica , dino.pedreschi@di.unipi.it
- Tutor: Letizia Milli, Dipartimento di Informatica



DIPARTIMENTO DI INFORMATICA - Università di Pisa

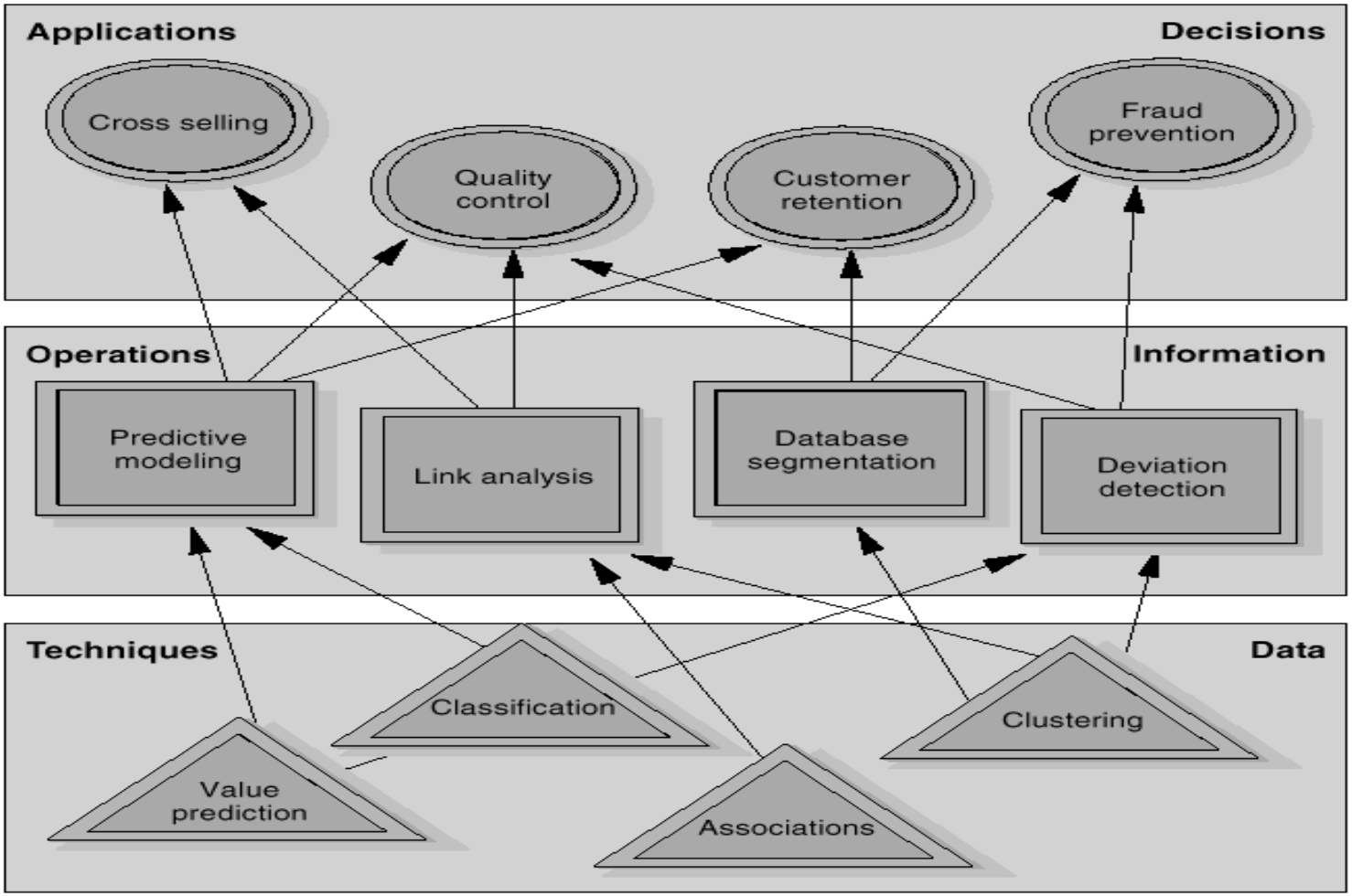
Master Big Data 2015

Master MAINS 2015

Lezione 2 Case Studies outline

- Competitive Intelligence
- Fraud Detection,
- Health care – Athereosclerosy prevention,
- Customer Segmentation
- Target Marketing

Applications, operations, techniques



L'Oreal, a case-study on competitive intelligence:

Source: DM@CINECA

http://open.cineca.it/datamining/dmCineca/

A small example

- Domain: technology watch - a.k.a. competitive intelligence
  - Which are the emergent technologies?
  - Which competitors are investing on them?
  - In which area are my competitors active?
  - Which area will my competitor drop in the near future?
- Source of data:
  - public (on-line) databases

The Derwent database

- Contains all patents filed worldwide in last 10 years
- Searching this database by keywords may yield thousands of documents
- Derwent document are semi-structured: many long text fields

- Goal: analyze Derwent document to build a model of competitors' strategy

Structure of Derwent documents

Raccolta dei Documenti

esempio di documento brevettuale

1/3881 - (C) Derwent Info 1994

AN: 94-364398 [45]

TI: Television with function for enlarging picture by variation of deflection frequency - has microprocessor for controlling system synchronous signal output, horizontal and vertical frequency drive circuit, sync. signal counter, signal detector.

DC: W03

PA: (GLDS) GOLDSTAR CO LTD

IN: O.KEITH

NP: 1

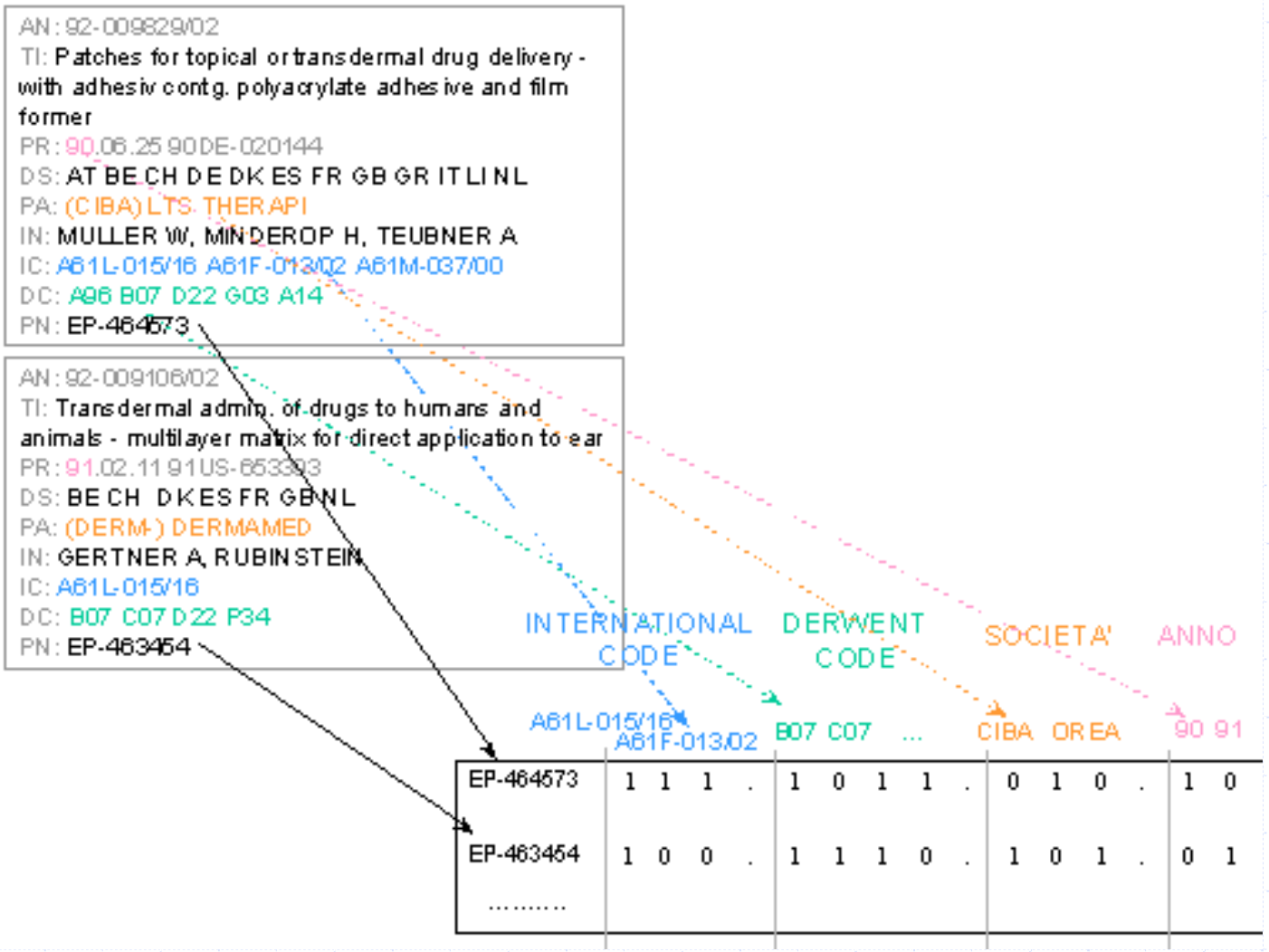
PR: 88KR-011143 880831

IC: H04N-005/262; C08J-005/18; G11B-005/704

PN: KR940043 B1 940120 DW9445

AB: ..... abstract .....

Demographic clustering: data structure



IBM-IM demographic clustering

- Designed for categorical variables
- Similarity index:
  - increases with number of common values on same attribute
  - decreases with number of different values on same attribute
- # of clusters is not fixed a priori
  - only upper bound set

Demographic clustering: parameters

	W_1	W_2	...								W_n
Doc i	1	1	1	1	0	1	1	0	1	0	1	0
Doc j	1	0	0	1	1	1	0	1	0	0	0	1

$$N_{11} = \sum_{k=1}^{m} x_{ik}\, x_{jk}$$

$$N_{10} = \sum_{k=1}^{m} x_{ik}\,(1 - x_{jk})$$

$$N_{01} = \sum_{k=1}^{m} (1 - x_{ik})\, x_{jk}$$

$$N_{00} = \sum_{k=1}^{m} (1 - x_{ik})\,(1 - x_{jk})$$

- Indice di Somiglianza

$$s(i,j) = \frac{a\, N_{11}}{b\, N_{11} + c\,(N_{10} + N_{01})}$$

- **Condorcet** $a=b=1$ $c=1/2$
- **Dice** $a=b=1$ $c=1/4$

- Soglia di Somiglianza

se $s(i,j) > \alpha$ → Doc_i e Doc_j sono simili α in $[0,1]$

- default: $\alpha = 0.5$

- Sistema di ponderazione

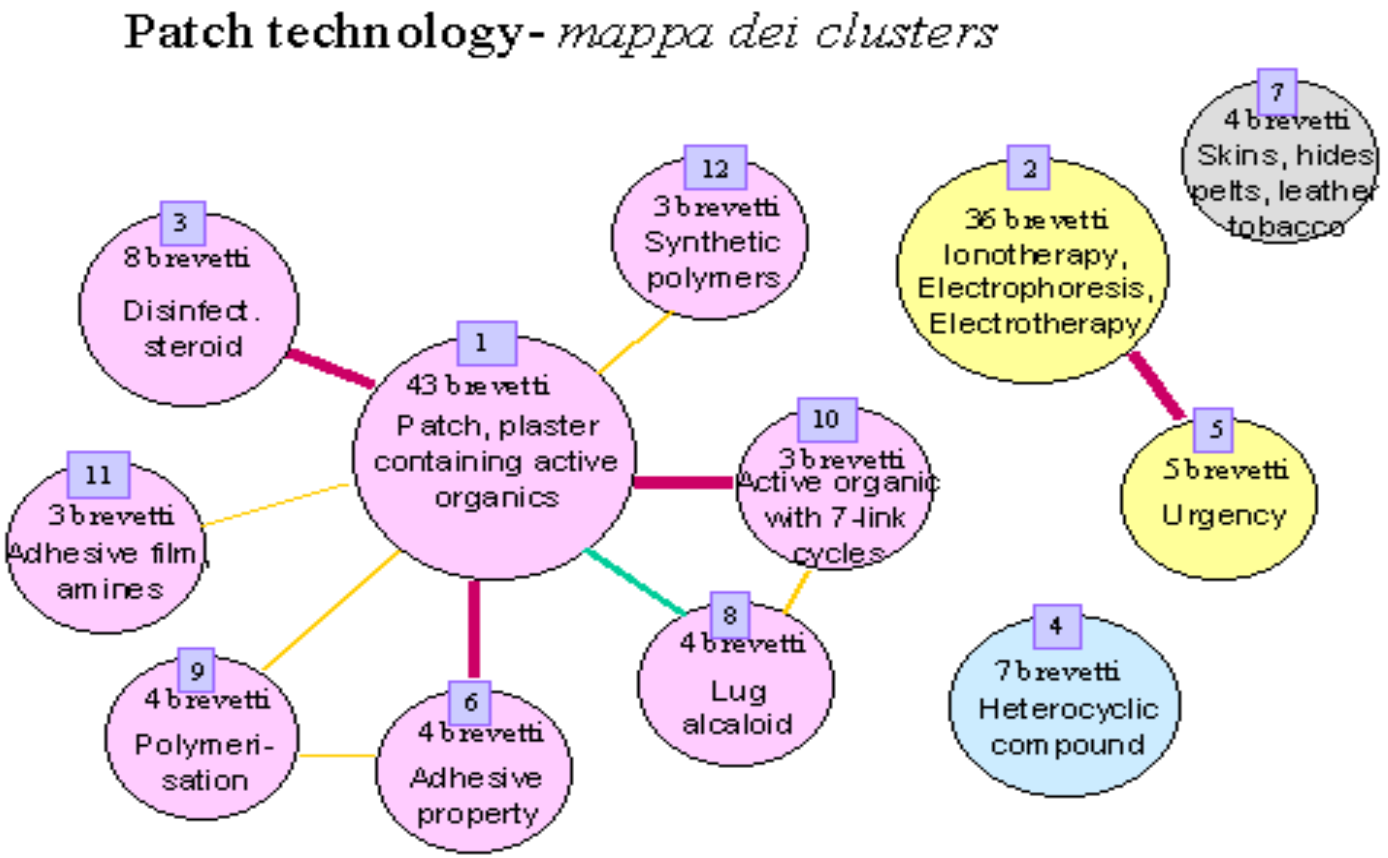
$$N_{11} = \sum_{k=1}^{m} x_{ik}\, x_{jk}\, w_k \quad (N_{10} = .. N_{01} = ...)$$

- $w_k = 1 / x_{.k}$
- $w_k = \log(N / x_{.k})$

Example dataset

- Patents in the area: patch technology (cerotto medicale)
  - 105 companies from 12 countries
  - 94 classification codes
  - 52 Derwent codes

Clustering output



Zoom on cluster 2

Patch technology- *descrizione del cluster n.2*

Classificazione Internazionale:

A61N-001/30 Electrotherapy; Appliances of electrical power by contact electrodes; Ionotherapy or electrophorese devices
A61M-037/00 Therapeutic patch

Classificazione Derwent:

S05 Electromedical
P34 Health, Electrotherapy

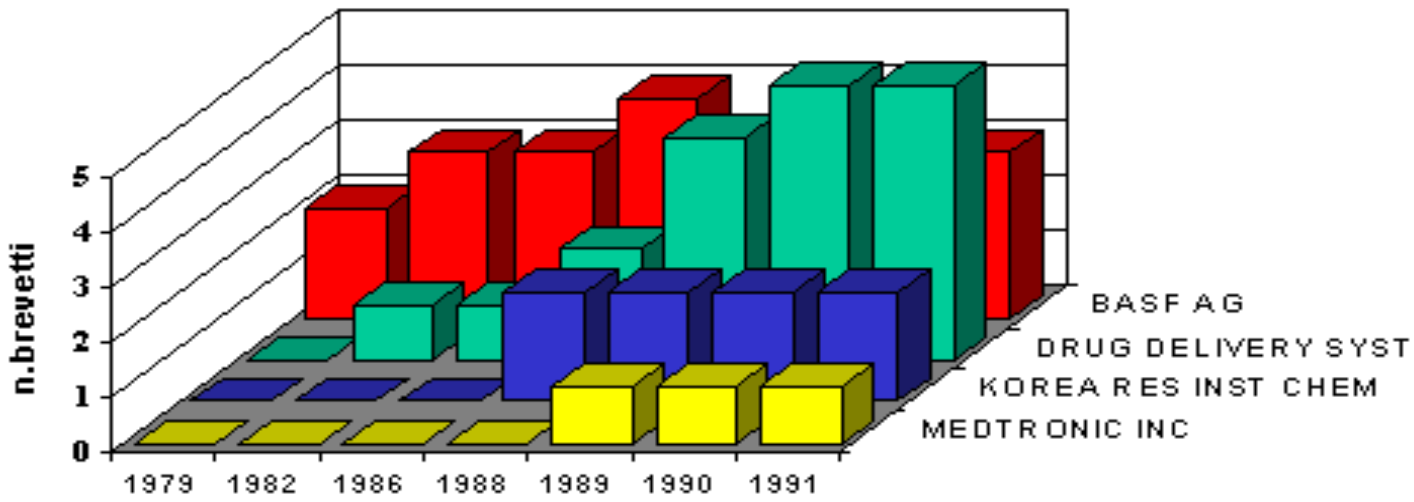
Società proprietarie:

DRUG DELIVERY SYST	42%
BASF AG	36%
KOREA RES INST CHEM	16%
MEDTRONIC INC	6%

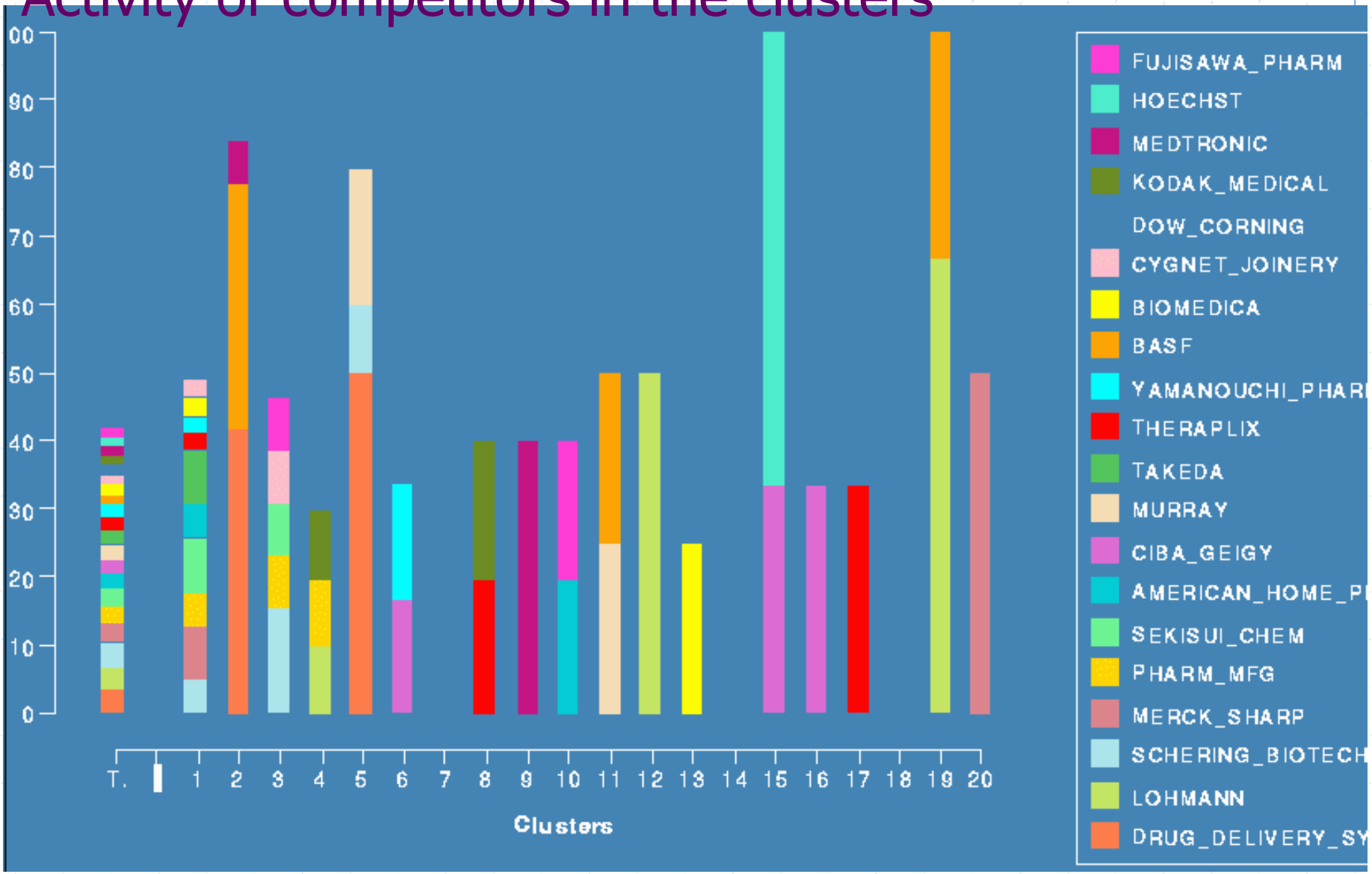
Anno	n. brevetti
1979	2
1982	4
1986	4
1988	8
1989	10
1990	11
1991	11

Zoom on cluster 2 - profiling competitors

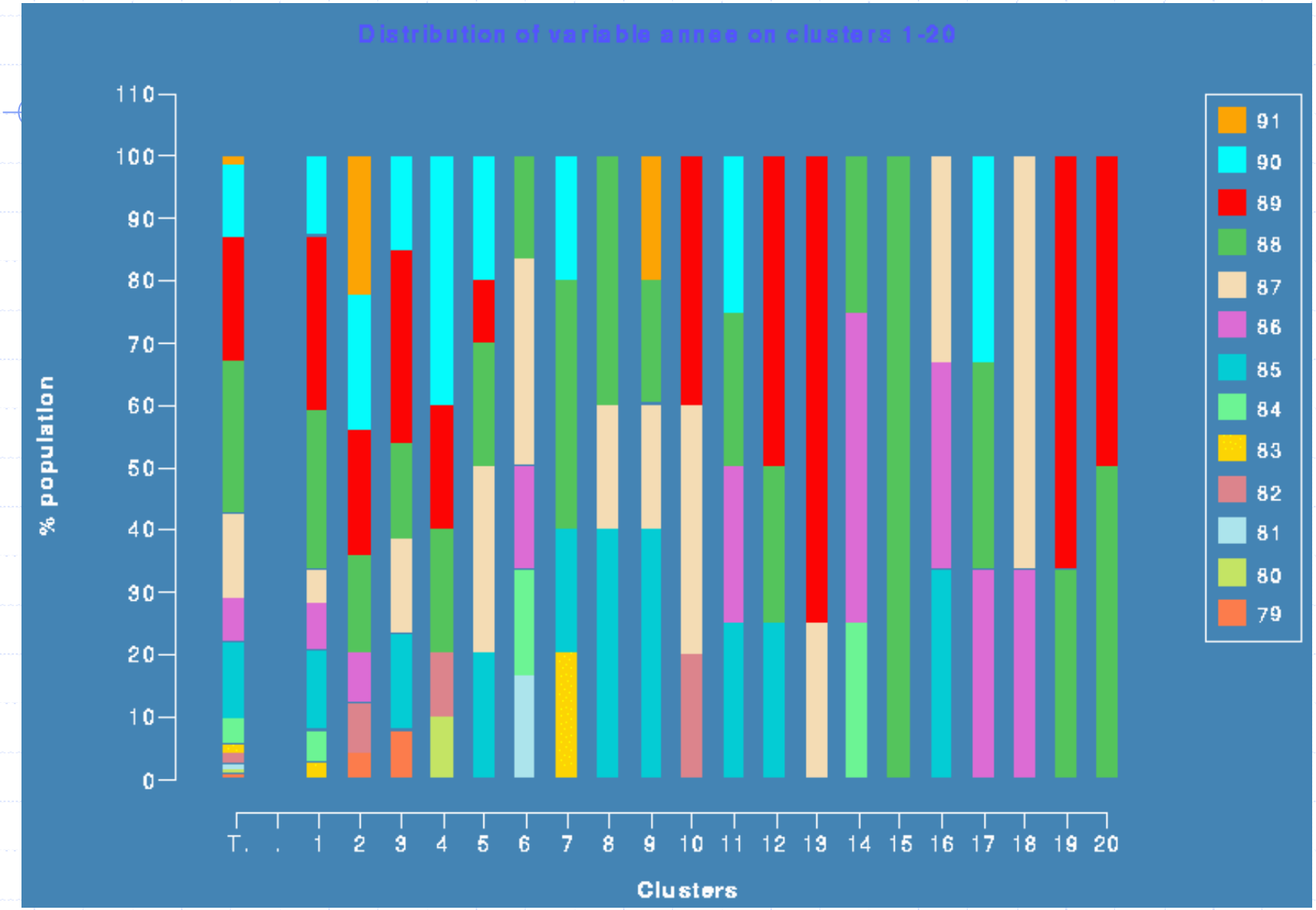
Patch technology- *cluster n.2 –*
attività della concorrenza nel tempo



Activity of competitors in the clusters



Temporal analysis of clusters



Fraud detection and audit planning

Source: Ministero delle Finanze

Progetto Sogei, KDD Lab. Pisa

Fraud detection

- A major task in fraud detection is constructing *models* of fraudulent behavior, for:
  - preventing future frauds (*on-line* fraud detection)
  - discovering past frauds (*a posteriori* fraud detection)
- analyze historical audit data to plan effective future audits

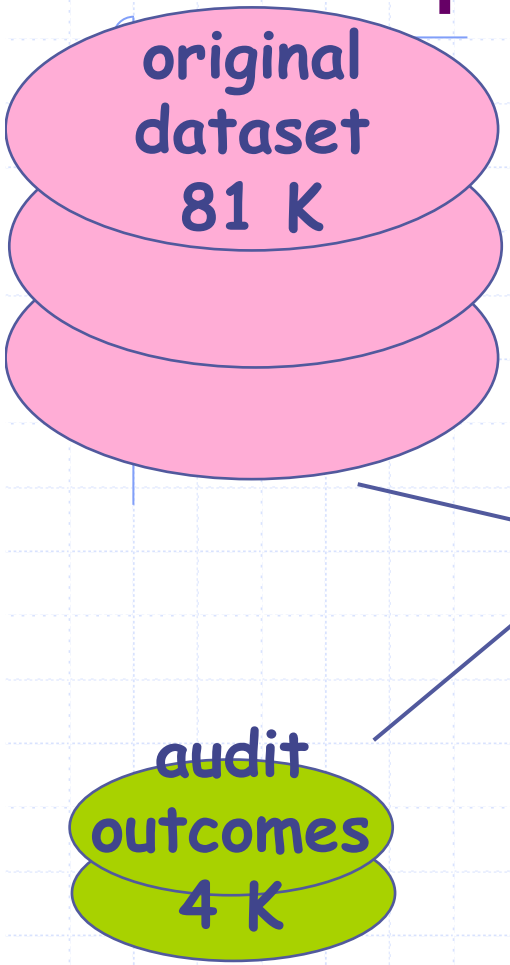
Audit planning

- Need to face a trade-off between conflicting issues:
  - *maximize audit benefits*: select subjects to be audited to maximize the recovery of evaded tax
  - *minimize audit costs*: select subjects to be audited to minimize the resources needed to carry out the audits.

Available data sources

- Dataset: tax declarations, concerning a targeted class of Italian companies, integrated with other sources:
  - social benefits to employees, official budget documents, electricity and telephone bills.
- Size: 80 K tuples, 175 numeric attributes.
- A subset of 4 K tuples corresponds to the *audited* companies:
  - outcome of audits recorded as the *recovery* attribute (= *amount of evaded tax ascertained*)

Data preparation

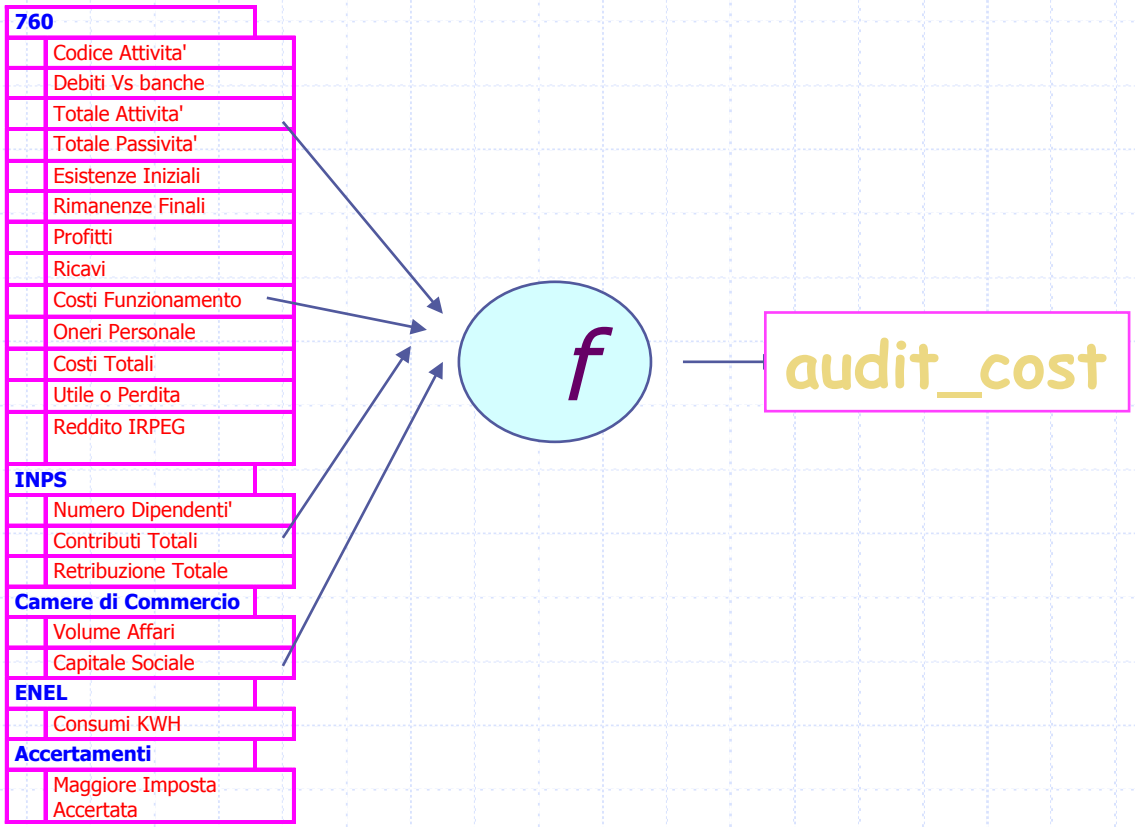


data consolidation
data cleaning
attribute selection

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

Cost model

- A derived attribute audit_cost is defined as a function of other attributes



Cost model and the target variable

- recovery of an audit after the audit cost
  actual_recovery = recovery - audit_cost

- target variable (class label) of our analysis is set as the Class of Actual Recovery (c.a.r.):

- $c.a.r. = \begin{cases} negative & \textbf{if} \quad actual\_recovery \leq 0 \\ positive & \textbf{if} \quad actual\_recovery > 0. \end{cases}$

Quality assessment indicators

- The obtained classifiers are evaluated according to several indicators, or metrics
- Domain-independent indicators
  - confusion matrix
  - misclassification rate
- Domain-dependent indicators
  - audit #
  - actual recovery
  - profitability

Domain-dependent quality indicators

- audit # (of a given classifier): number of tuples classified as positive =
  # (FP ∪ TP)
- actual recovery: total amount of actual recovery for all tuples classified as positive
- profitability: average actual recovery per audit
- relevance: ratio between profitability and misclassification rate

The REAL case

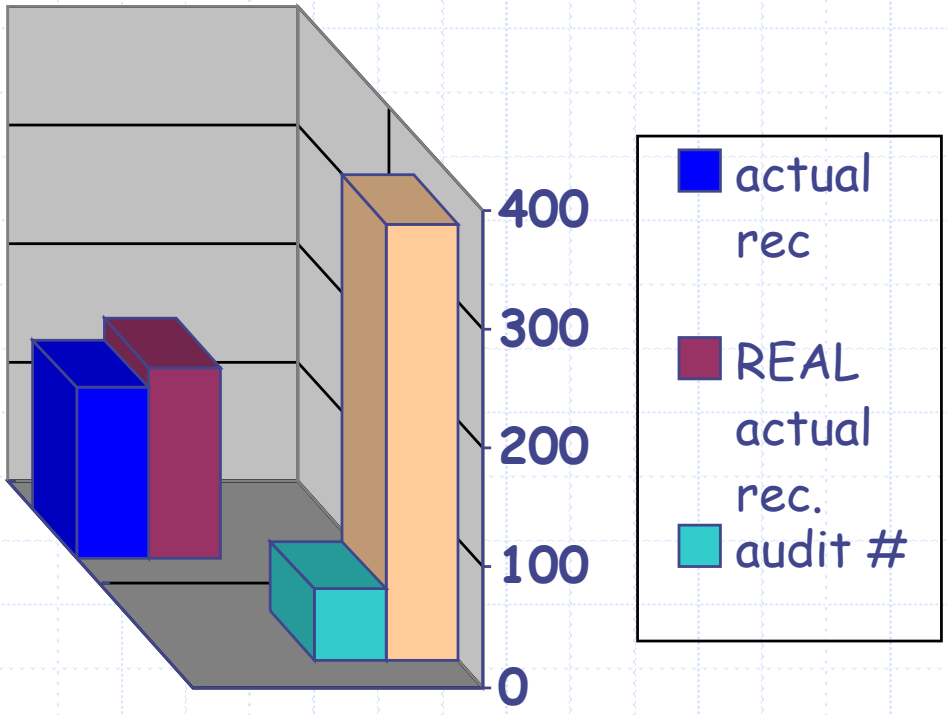
- Classifiers can be compared with the REAL case, consisting of the whole test-set:

- audit # (REAL) = 366
- actual recovery(REAL) = 159.6 M euro

Model evaluation: classifier 1 (min FP)

- no replication in training-set (unbalance towards negative)
- 10-trees adaptive boosting

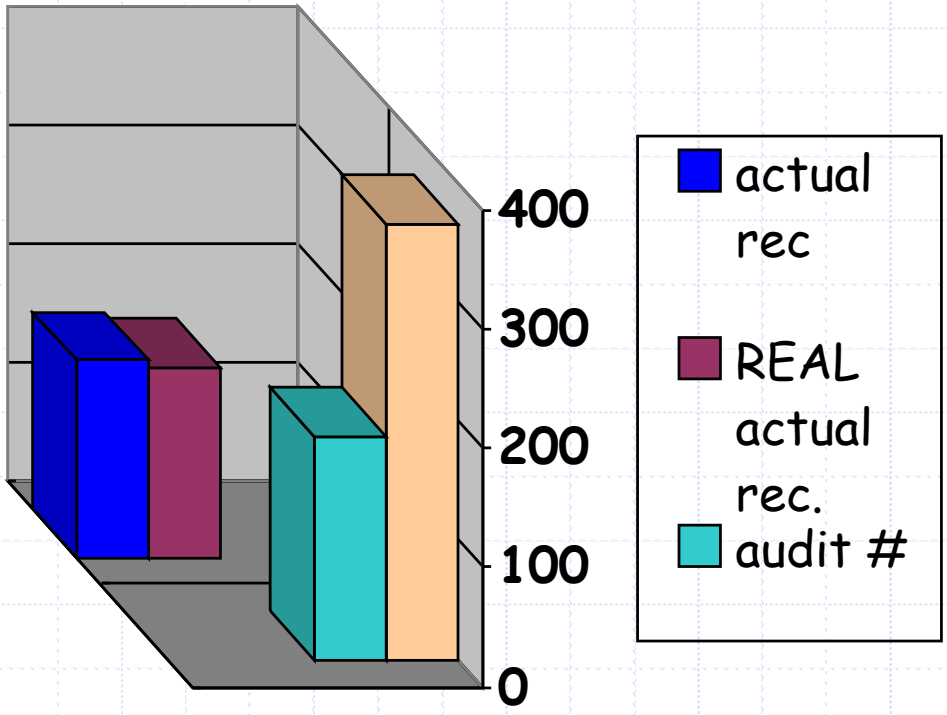
- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.*= 141.7 Meuro
- *profitability* = 2.401



Model evaluation: classifier 2 (min FN)

- replication in training-set (balanced neg/pos)
- misc. weights (trade 3 FP for 1 FN)
- 3-trees adaptive boosting

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



Atherosclerosis prevention study

2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC)

Atherosclerosis prevention study:

- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.

- Used for Discovery Challenge at PKDD 00-02-03-04

Atherosclerosis prevention study:

- Study on 1400 middle-aged men at Czech hospitals
  - Measurements concern development of cardiovascular disease and other health data in a series of exams
- The aim of this analysis is to look for associations between medical characteristics of patients and death causes.
- Four tables
  - Entry and subsequent exams, questionnaire responses, deaths

The input data

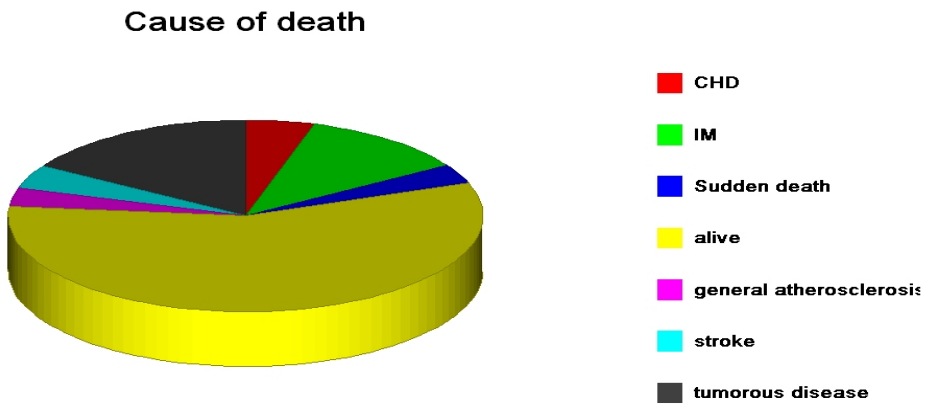
Data from Entry and Exams		
General characteristics	**Examinations**	**habits**
Marital status	Chest pain	Alcohol
Transport to a job	Breathlessness	Liquors
Physical activity in a job	Cholesterol	Beer 10
Activity after a job	Urine	Beer 12
Education	Subscapular	Wine
Responsibility	Triceps	Smoking
Age		Former smoker
Weight		Duration of smoking
Height		Tea
		Sugar
		Coffee

The input data

DEATH CAUSE	PATIENTS	%
myocardial infarction	80	20.6
coronary heart disease	33	8.5
stroke	30	7.7
other causes	79	20.3
sudden death	23	5.9
unknown	8	2.0
tumorous disease	114	29.3
general atherosclerosis	22	5.7
TOTAL	389	100.0

Data selection

- When joining “Entry” and “Death” tables we implicitely create a new attribute “Cause of death”, which is set to “alive” for subjects present in the “Entry” table but not in the “Death” table.
- We have only 389 subjects in death table.



The prepared data

Patient	General characteristics		Examinations		Habits		Cause of death
	Activity after work	Education	Chest pain	…	Alcohol	…..	
1	moderate activity	university	not present		no		Stroke
2	great activity		not ischaemic		occasionally		myocardial infarction
3	he mainly sits		other pains		regularly		tumorous disease
……	……..	……..	………..	..	…	……	alive
389	he mainly sits		other pains		regularly		tumorous disease

Descriptive Analysis/ Subgroup Discovery / Association Rules

Are there strong relations concerning death cause?

General characteristics (?) ⇒ Death cause (?)

Examinations (?) ⇒ Death cause (?)

Habits (?) ⇒ Death cause (?)

Combinations (?) ⇒ Death cause (?)

Example of extracted rules

- Education(university) & Height<176-180> ⇒Death cause (tumouros disease), *16 ; 0.62*
- It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.

Example of extracted rules

- Physical activity in work(he mainly sits) & Height<176-180> ⇒ Death cause (tumouros disease), *24; 0.52*
- It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.

Example of extracted rules

- Education(university) & Height<176-180>
  ⇒Death cause (tumouros disease),
  16; 0.62; +1.1;
- the relative frequency of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients

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

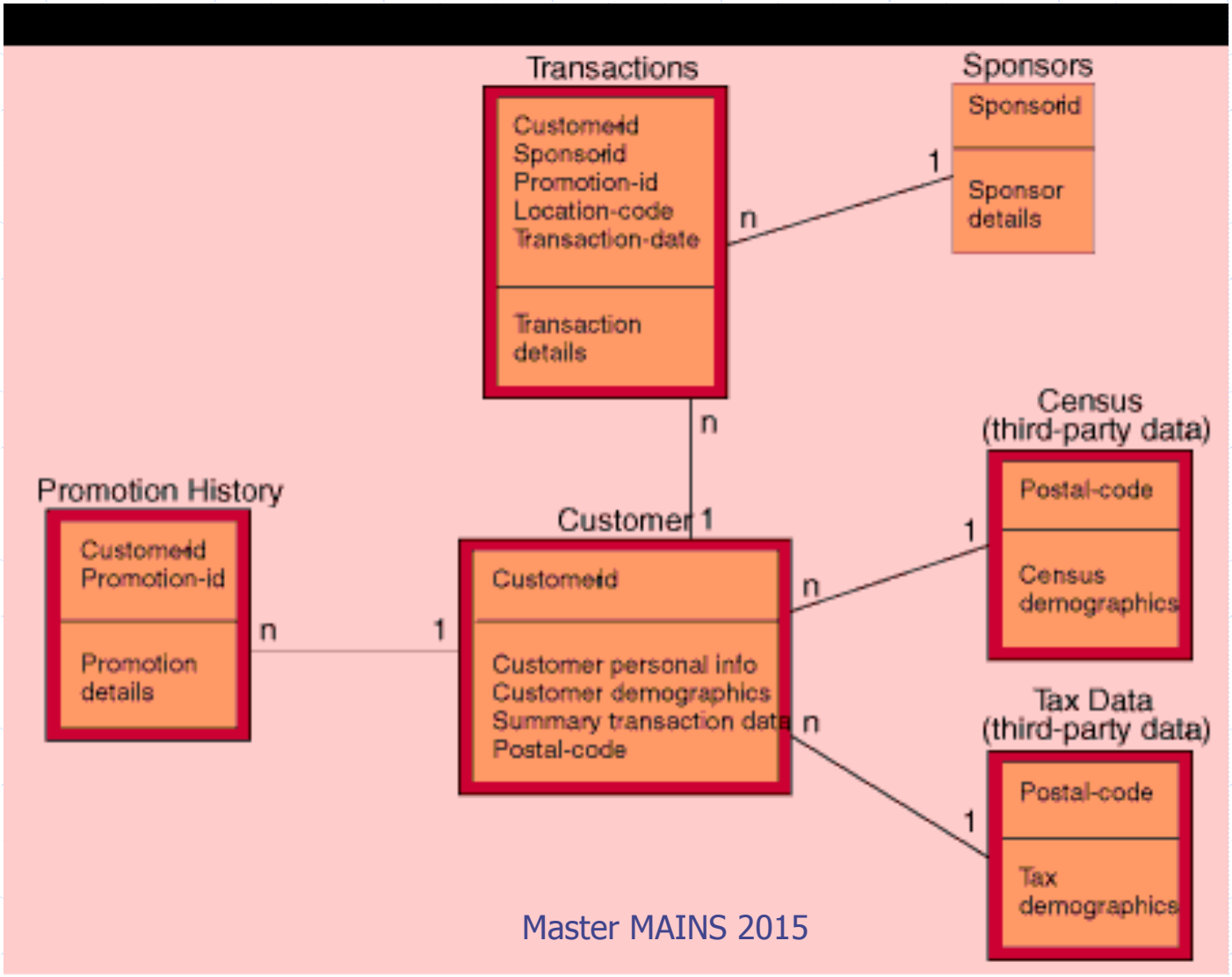
Stato dell'arte prima del data mining

- 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!

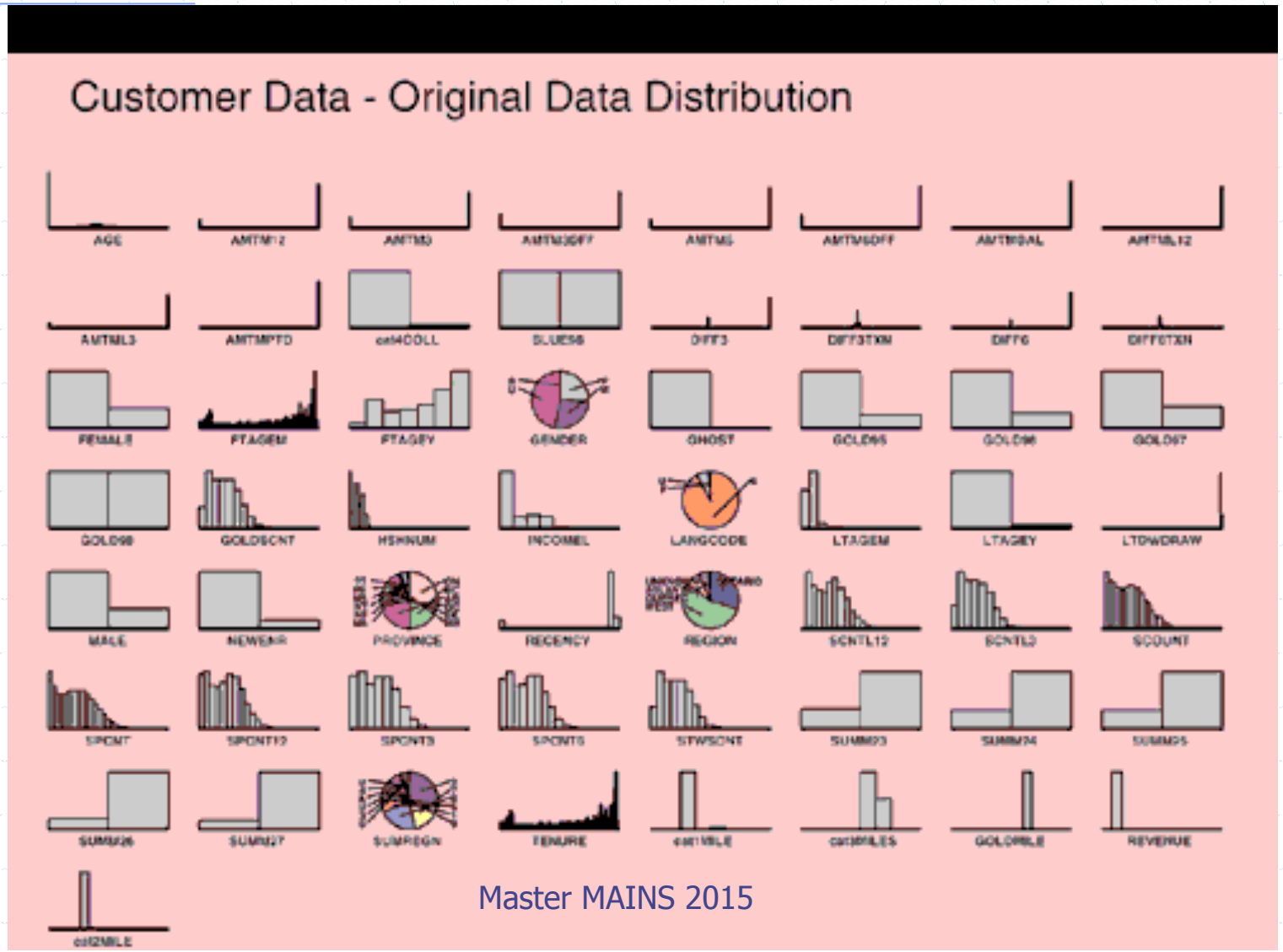
Sorgente dei dati nel DW



Preparazione dei dati

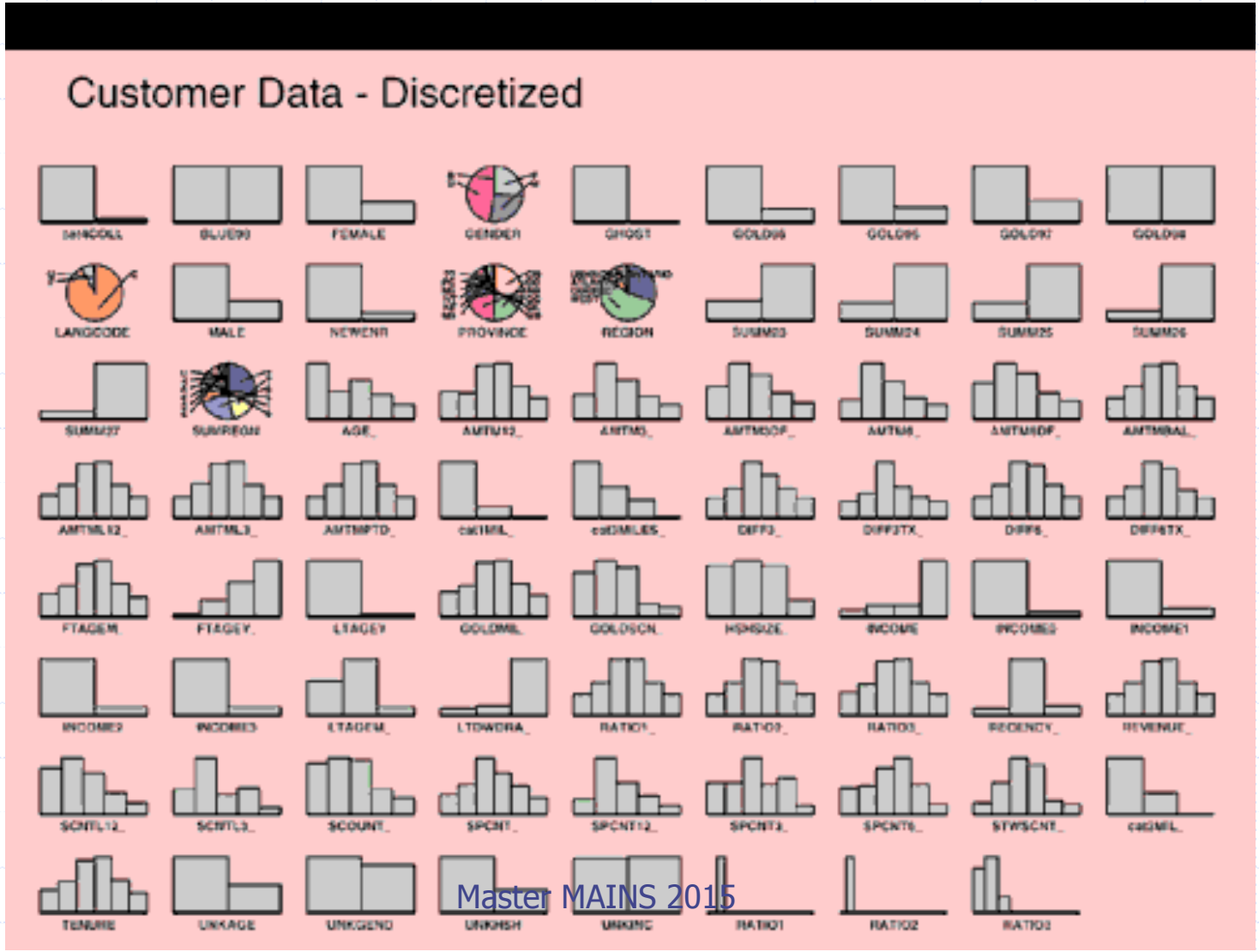
- Creazione delle variabili economiche di ciascun **cliente**, mediante aggregazione dei propri acquisti
  - Volume di spesa
  - Durata del suo ciclo di vita
  - Numero di compagnie sponsor in cui ha acquistato
  - Numero di compagnie sponsor in cui ha acquistato negli ultimi 12 mesi
  - Distanza (in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

I dolori della pulizia dei dati: prima …



Master MAINS 2015

… e dopo la cura



Prima e dopo la cura

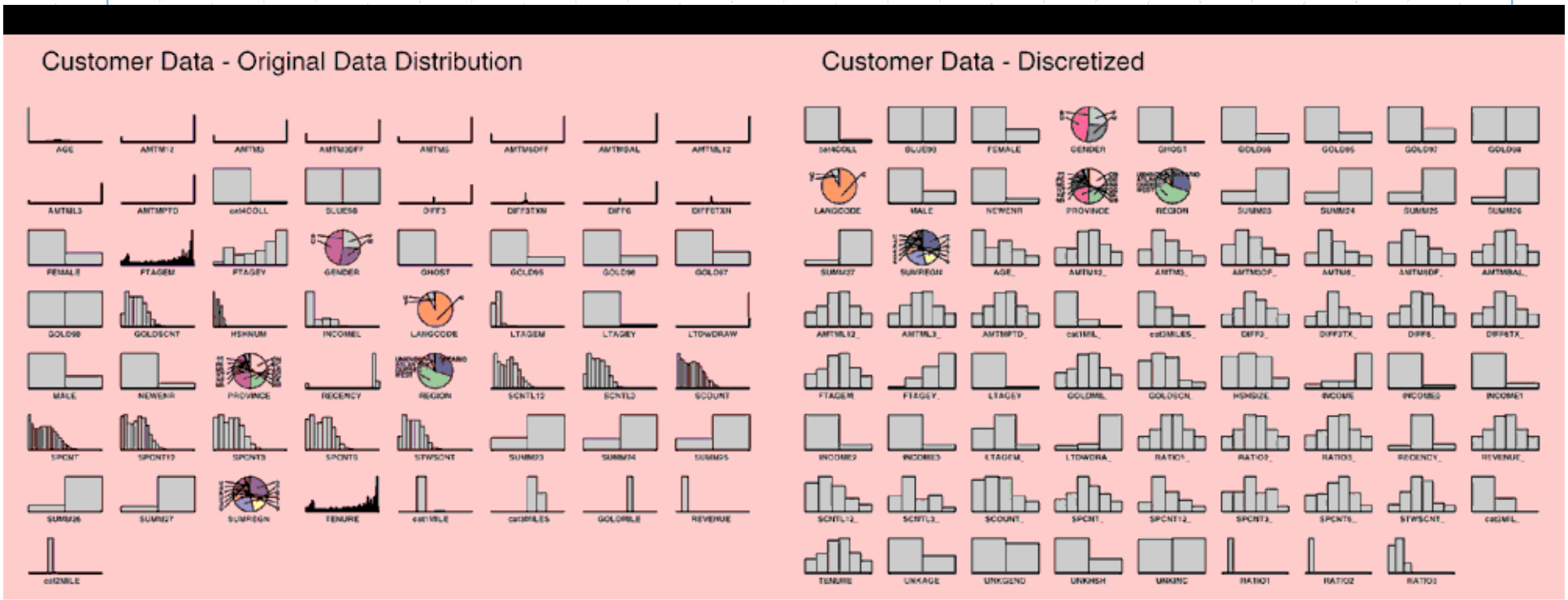
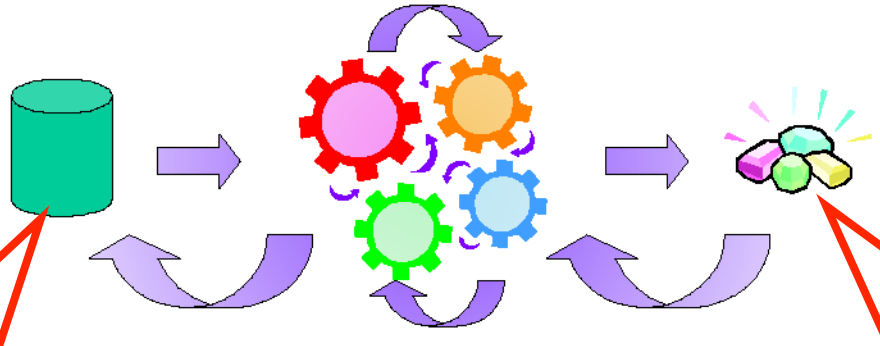


Figure3. *Original data.*

Figure 4. *Discretized data.*

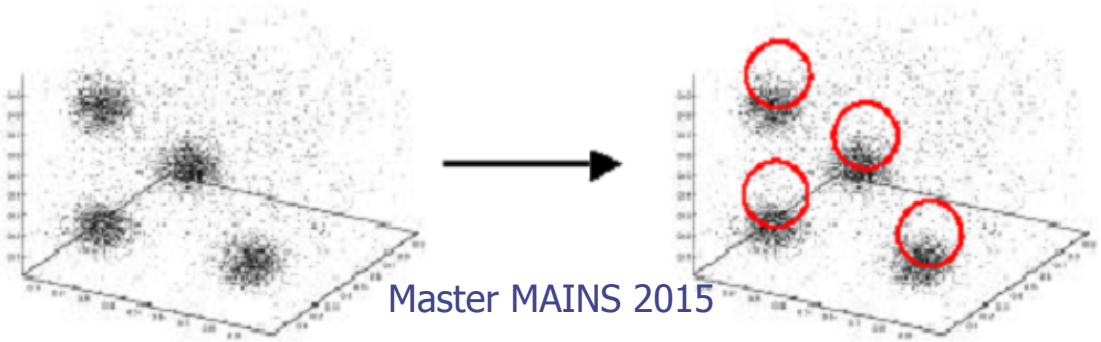
Estrazione del modello di clustering

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



Clustering/segmentation methodology

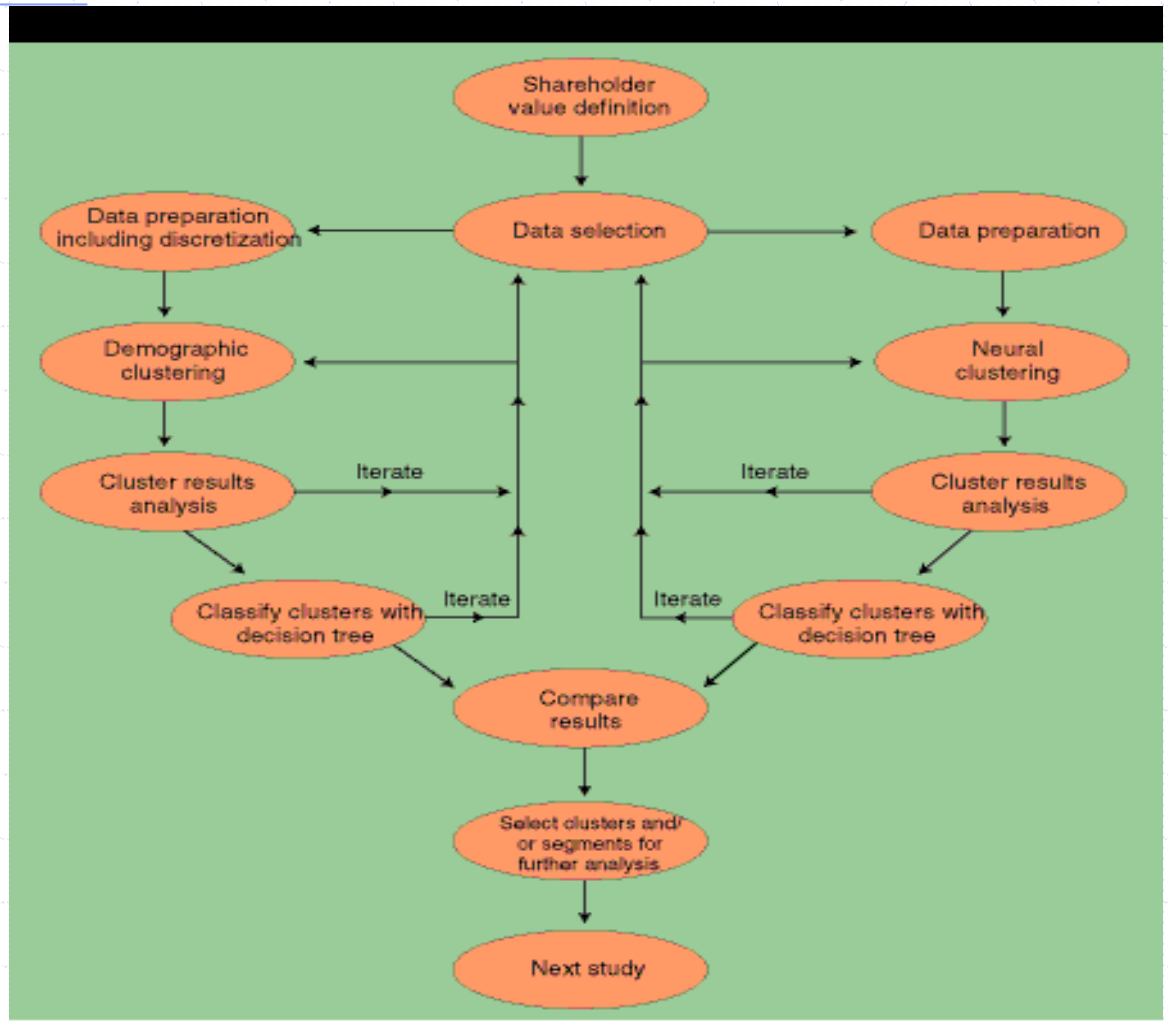
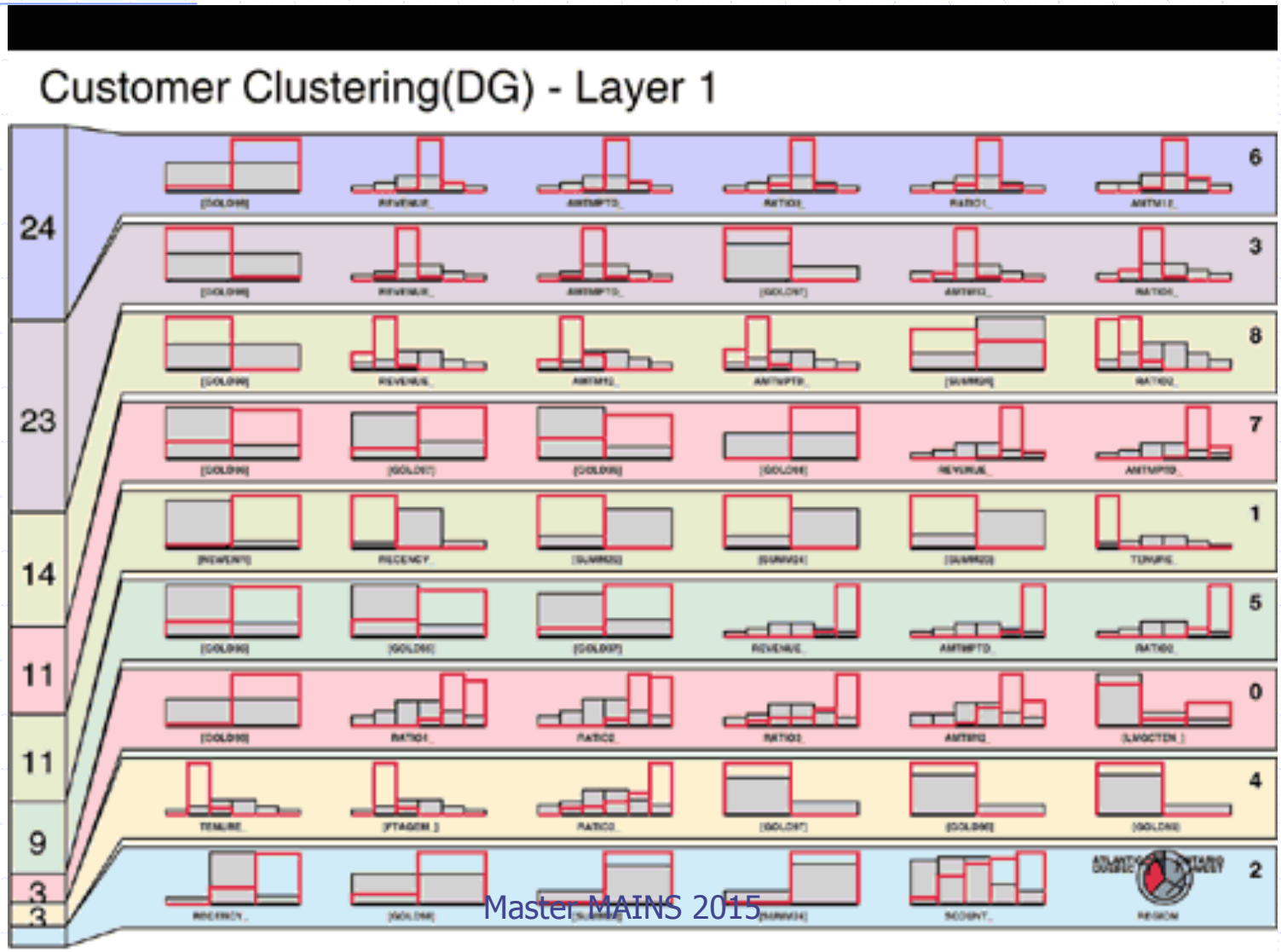


Figure 6. *Clustering workflow.*

Output del clustering



Visualization of clusters

- horizontal strip = a cluster
- clusters are ordered from top to bottom in order of size
- variables are ordered from left to right in order of importance to the cluster, based on a chi-square test between variable and cluster ID.
- other metrics include entropy, Condorcet criterion, and database order.

Visualization of clusters

- variables used to define clusters are without brackets, while the supplementary variables appear within brackets.
- numeric (integer), discrete numeric (small integer), binary, and continuous variables have their frequency distribution shown as a bar graph.

- red bars = distribution of the variable within the current cluster.
- gray solid bars = distribution of the variable in the whole universe.

Visualization of clusters

- Categorical variables are shown as pie charts.
- inner pie = distribution of the categories for the current cluster
- outer ring = distribution of the variable for the entire universe.

- The more different the cluster distribution is from the average, the more interesting or distinct the cluster.

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

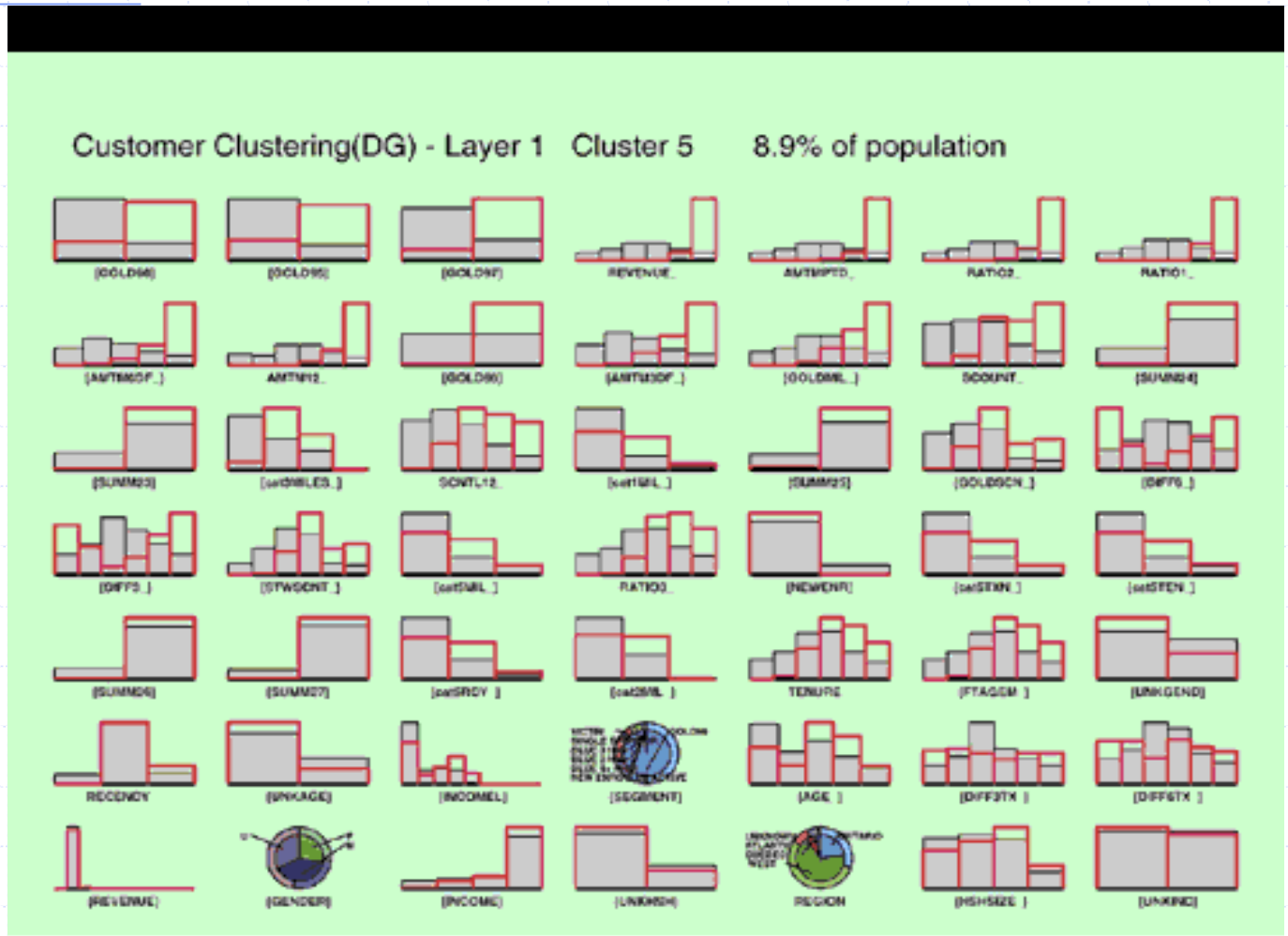


Figure 8. *Cluster 5 output.*

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 cluste*

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.

Follow-up

- Reactions from The Loyalty Group
  - visualization of results allowed for meaningful and actionable analysis.
  - original segmentation methodology validated, but that refinements to the original segmentation could prove valuable.
  - decision to undertake further data mining projects, including
    - predictive models for direct mail targeting,
    - further work on segmentation using more detailed behavioral data,
    - opportunity identification using association algorithms within the segments discovered.

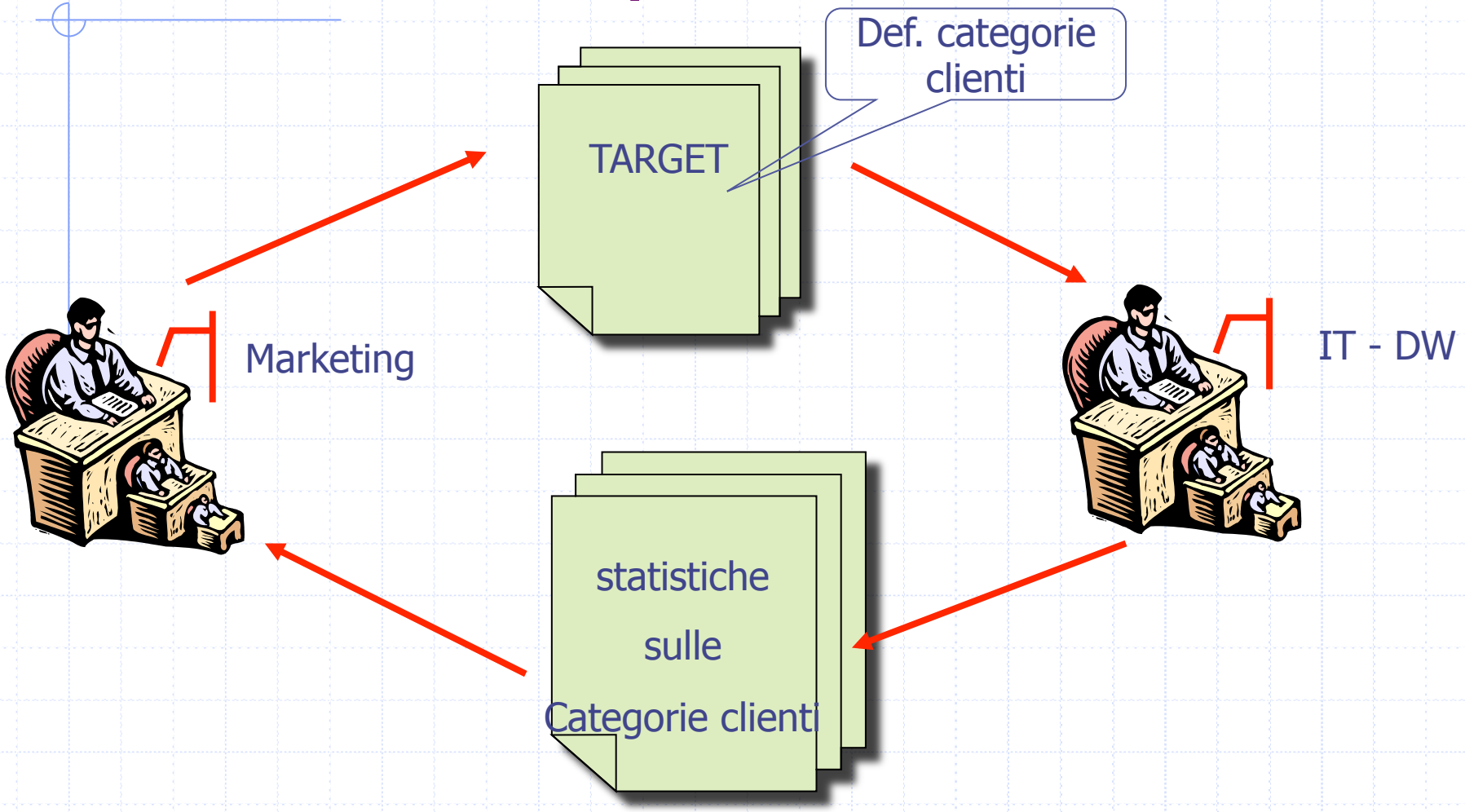
Analisi previsionale per l'ottimizzazione della postalizzazione delle promo

KDD Lab. Pisa

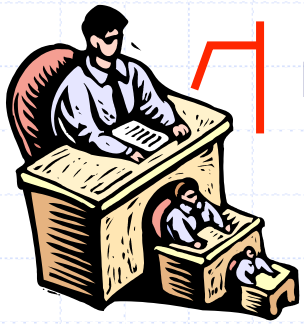
Postalizzazione di promozioni

- 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

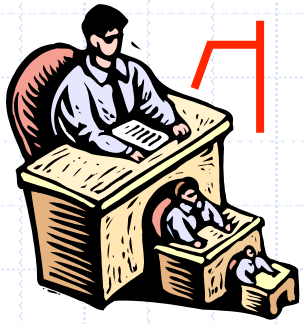


selezionare i clienti e postalizzare



Marketing

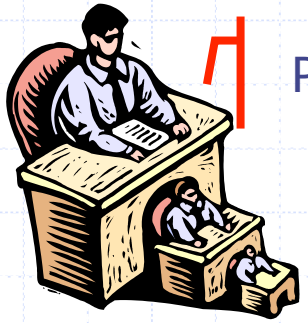
TARGET



IT - DW

Def. dettagli promo

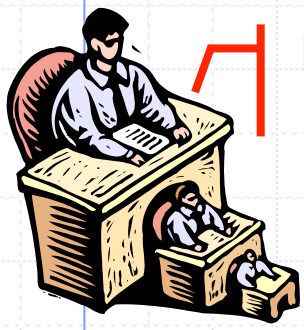
promozione



Postalizzazione

Estrazione Categorie Clienti

Erogare premi e raccogliere dati



Marketing sviluppo

promozione

Def. dettagli promo

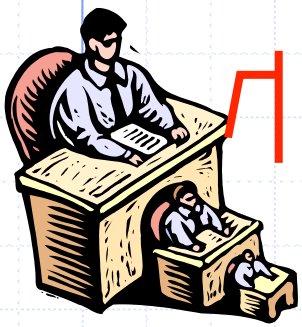
Sistema automatico aggiornamento DATA Mart

Mercati

Sistema automatico attivazione mercati

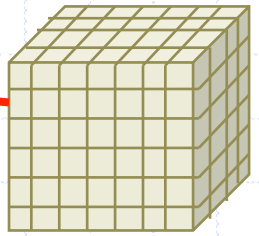
DATA MART PROMOZIONI

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

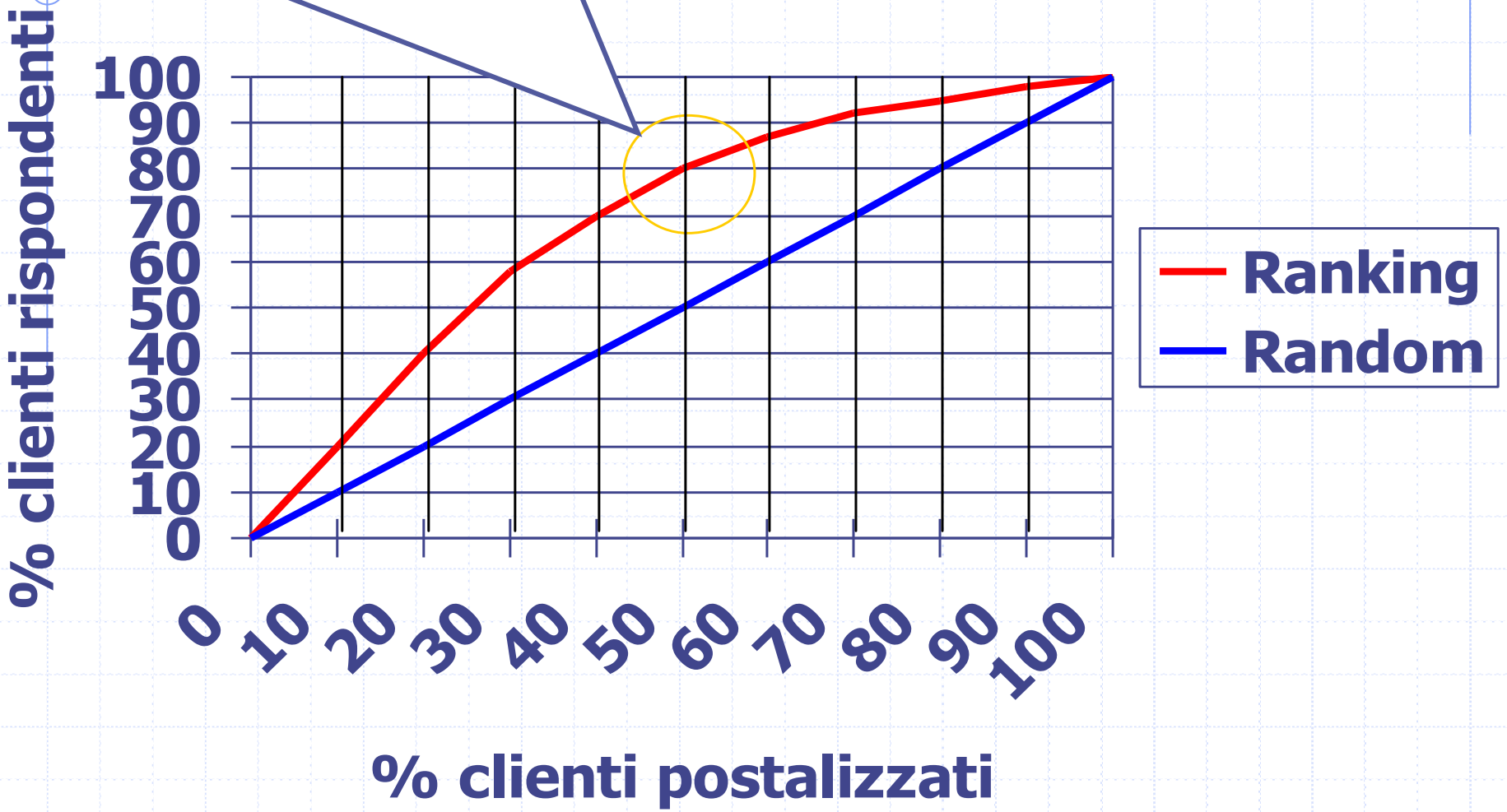
Lift Chart


% 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

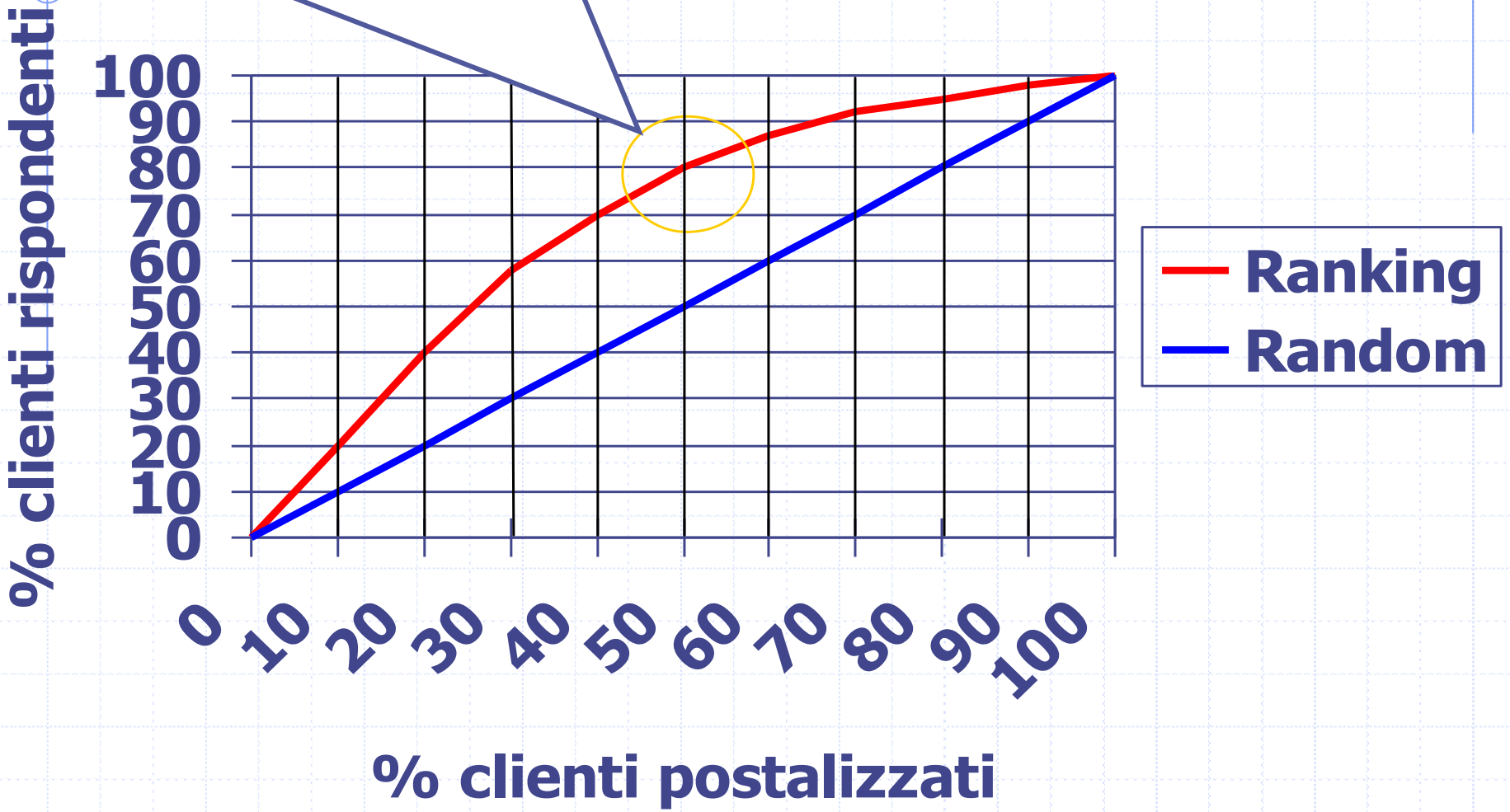
LIFT CHART

- 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)

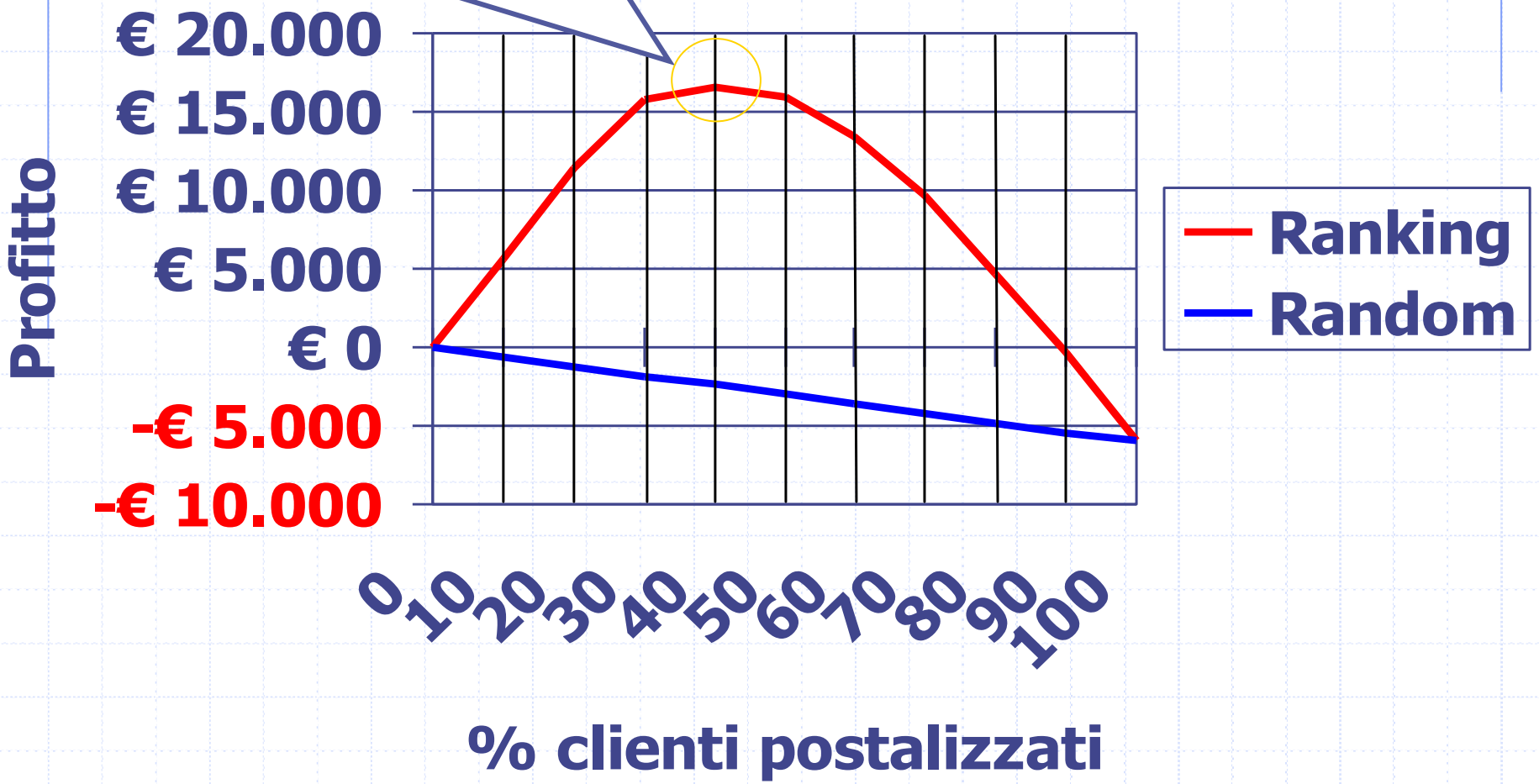
- Il Lift Chart rappresenta un aiuto grafico per ragionare sul rapporto ottimale fra costi di postalizzazione e percentuale di redemption
  - a fronte di sostanziali riduzioni di postalizzati (=budget) permette di ridurre di poco il numero di redenti
  - a parità di budget, permette di incrementare il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

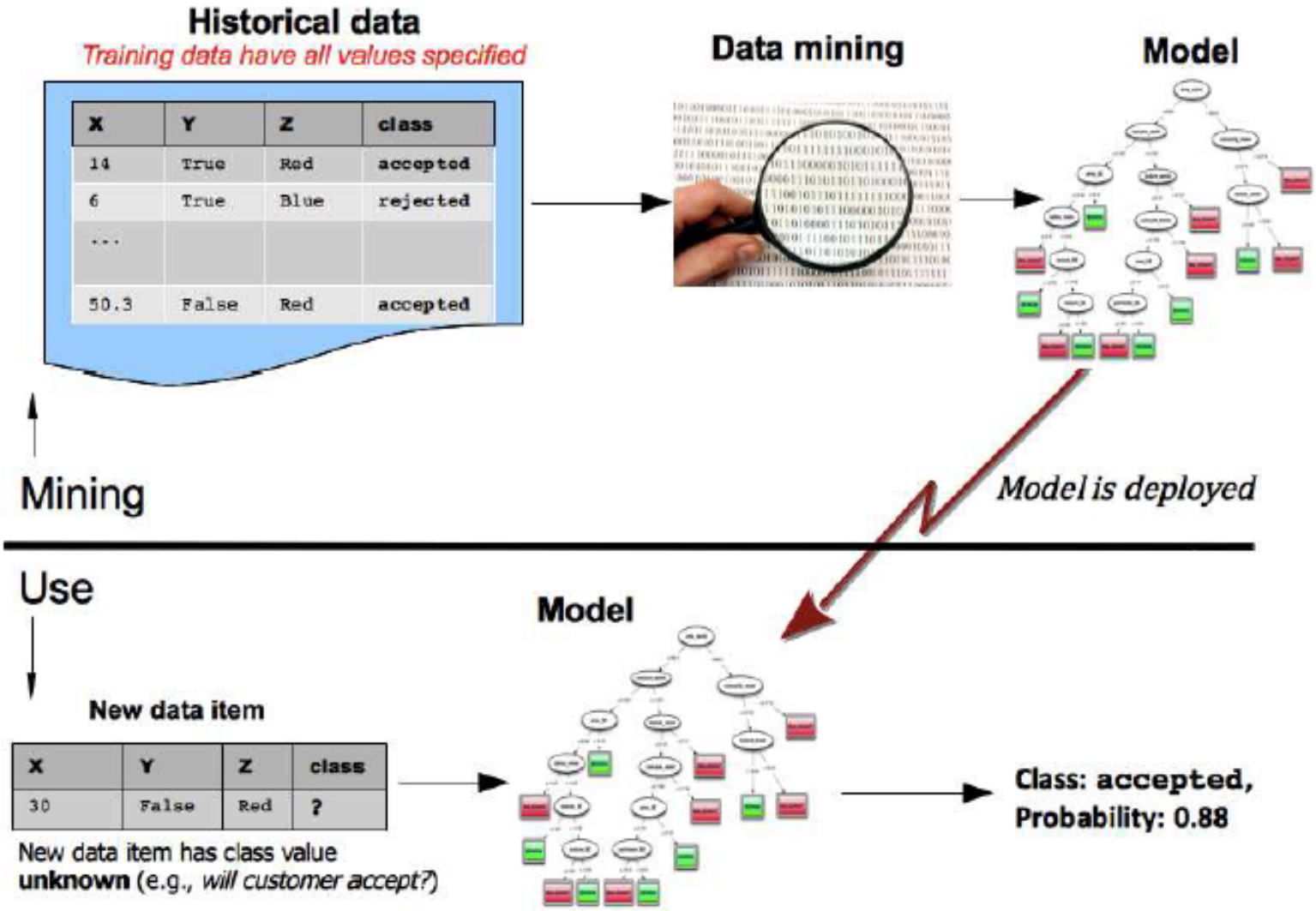
Dietro le quinte

- Il lift-chart della scheda promo e gli elenchi dei clienti da postalizzare sono calcolati ad a cura dell' ufficio marketing
richiesta dell' utente marketing/sviluppo, a partire dai modelli predittivi che risiedono sul server (di progetto o di DW)
- I modelli predittivi sono riaggiornati periodicamente, ad ogni richiesta dell' utente sulla base a cura dell' ufficio IT/DW contenuto attuale del DW, mediante tecniche di data mining

On-line

Off-line

Data mining and its use



Seminar 1 Case studies - Bibliography

- Artif. Intell. Med. 2001 Jun;22(3), Special Issue on Data Mining in Medicine
- Klosgen, Zytkow, Handbook of Data Mining, Oxford, 2001
- Micheael, J. A. Berry, Gordon S. Linoff, Mastering Data Mining, Wiley, 2000
- ECML/PKDD2004 Discovery Challenge homepage[http://lisp.vse.cz/challenge/ecmlpkdd2004/]
- On the road to knowledge: mining 21 years of UK traffic accident reports, in *Data Mining and Decision Support: Aspects of Integration and Collaboration*, pages 143--155. Kluwer Academic Publishers, January 2003