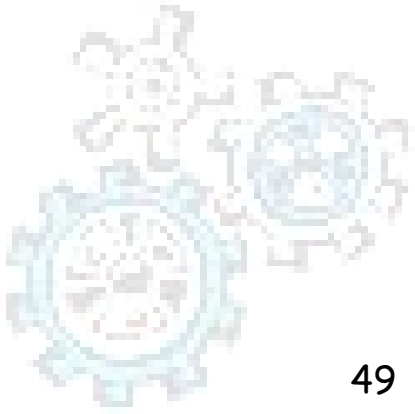
L'Oreal, a case-study on competitive intelligence:

Source: DM@CINECA

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

A small example

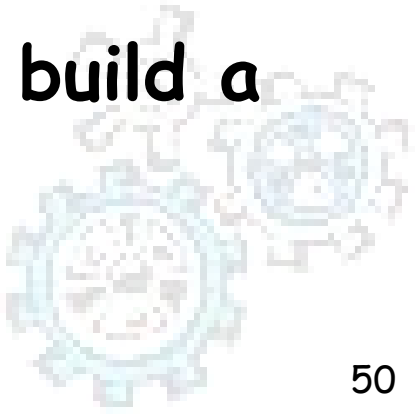
- Domain: technology watch - a.k.a. competitive intelligence
  - Which are the emergent technologies?
  - Which competitors are investing on them?
  - In which area are my competitors active?
  - Which area will my competitor drop in the near future?
- Source of data:
  - public (on-line) databases



The Derwent database

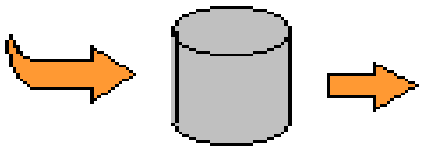
- Contains all patents filed worldwide in last 10 years
- Searching this database by keywords may yield thousands of documents
- Derwent documents are semi-structured: many long text fields

- Goal: analyze Derwent documents 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

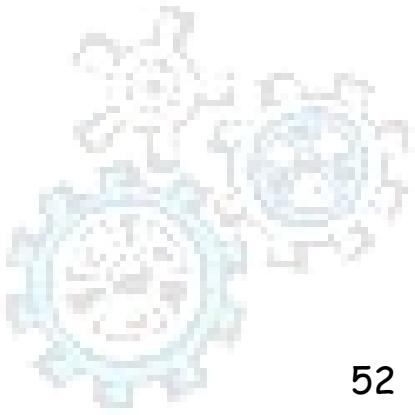
IC: H04N-005/262; C08J-005/18; G11B-005/704

PN: KR940043 B1 940120 DW9445

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

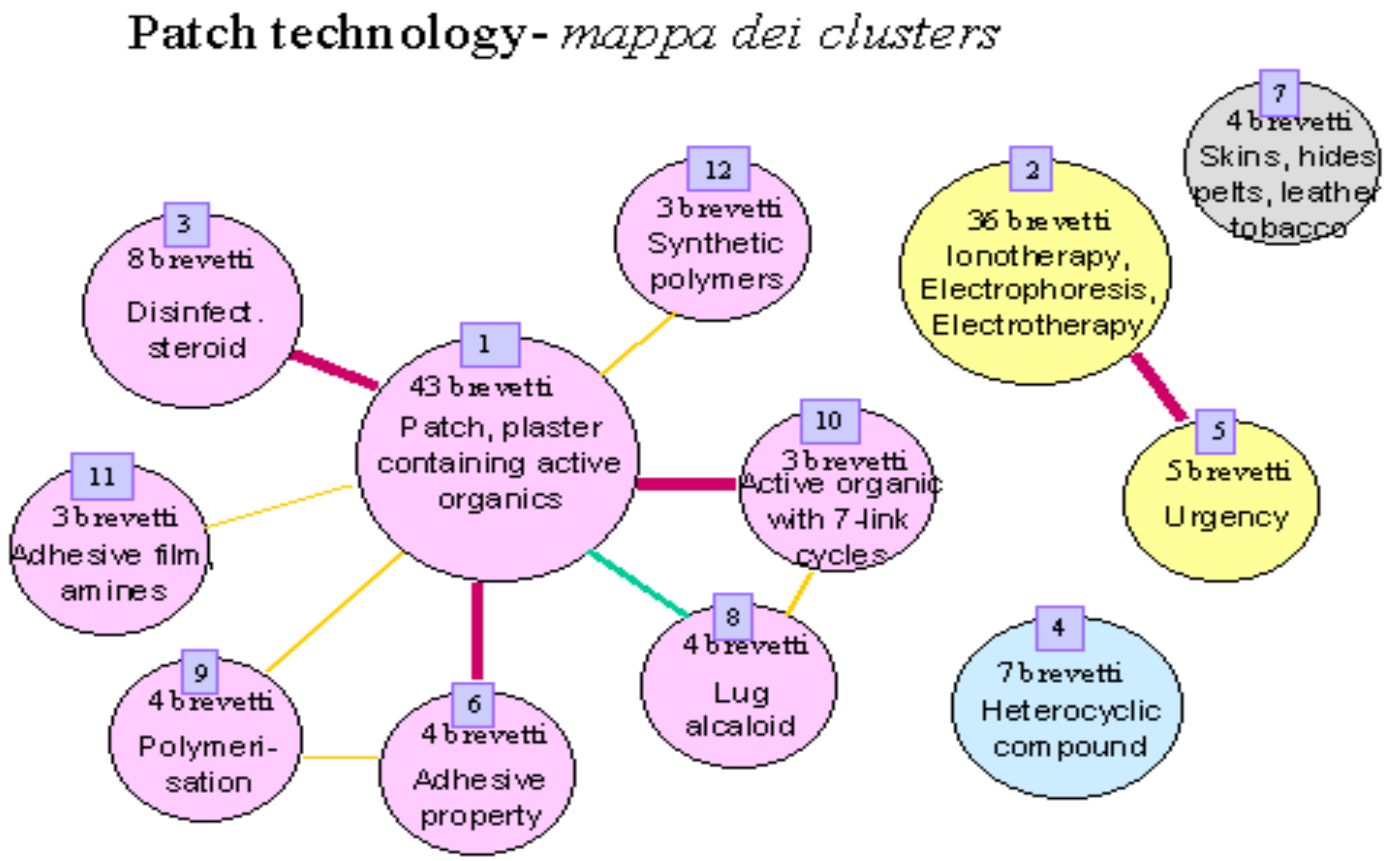
Example dataset

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



Clustering output

- Clusters patents with similar sets of keywords in the same group
- Groups are linked if they share some keywords



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

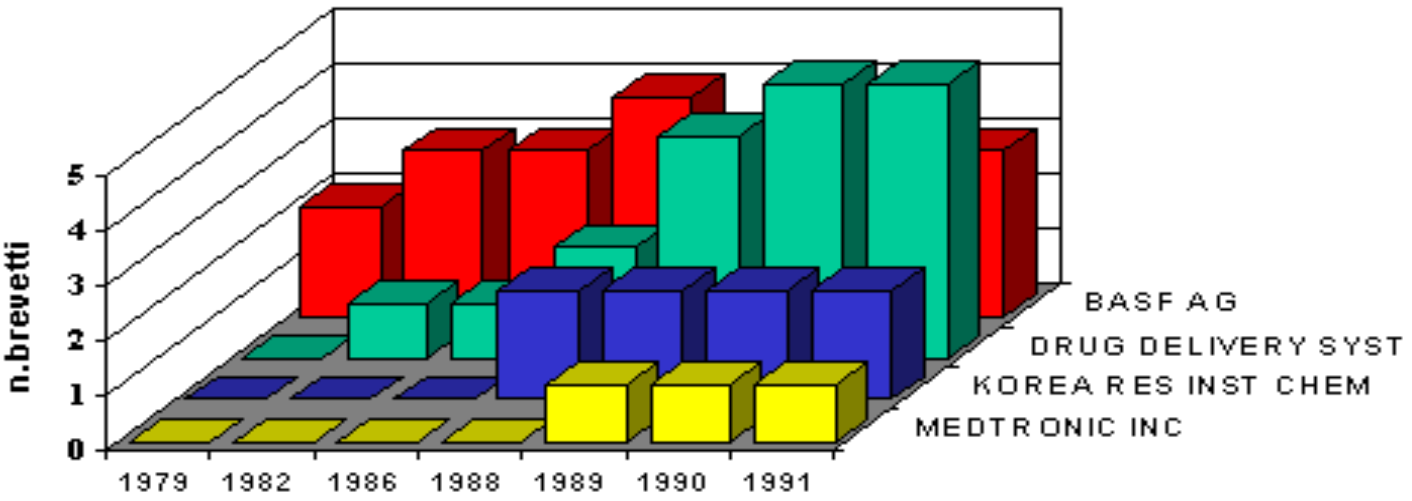
Società proprietarie:

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

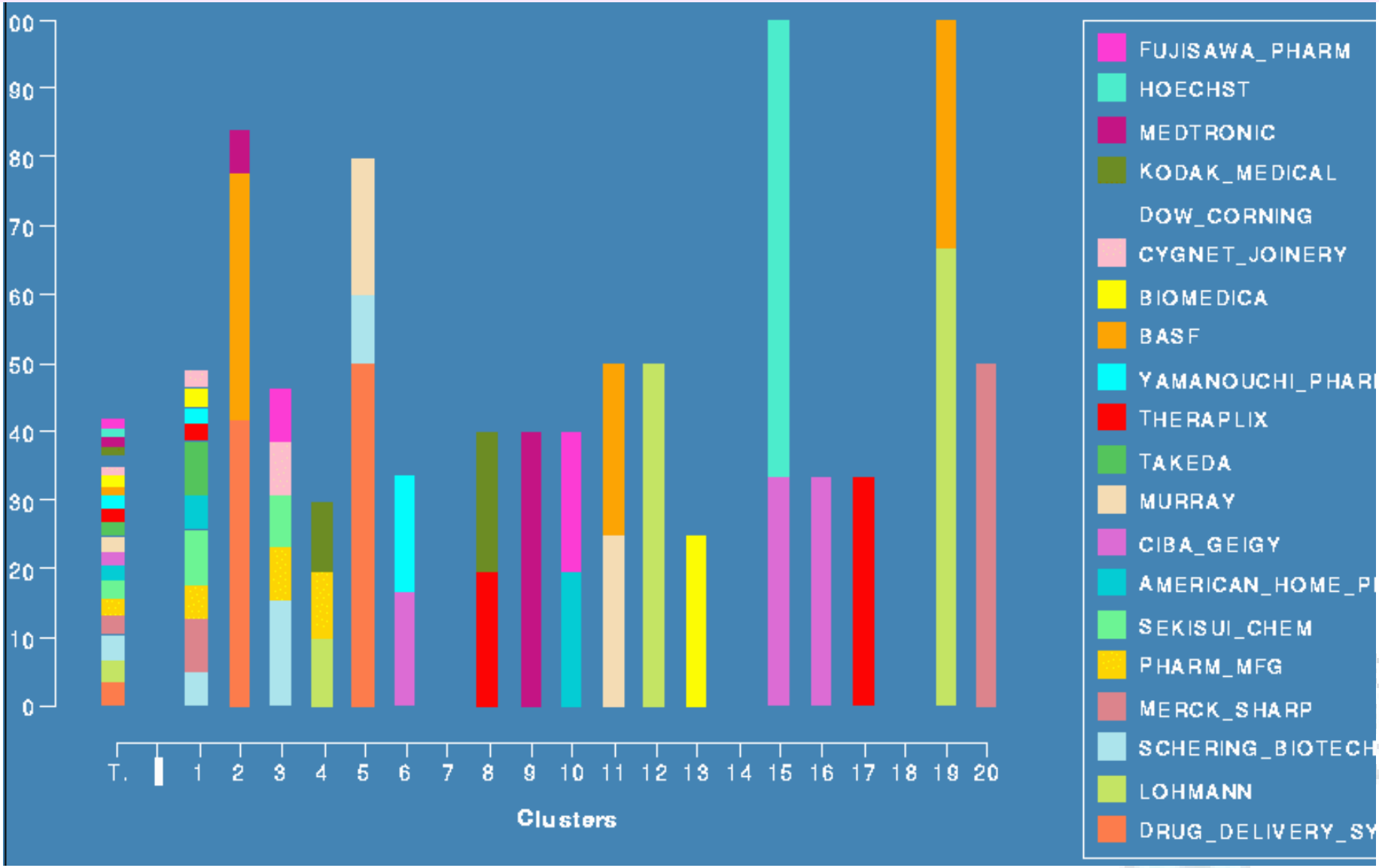
Anno	n. brevetti
1979	2
1982	4
1986	4
1988	8
1989	10
1990	11
1991	11

Zoom on cluster 2 - profiling competitors

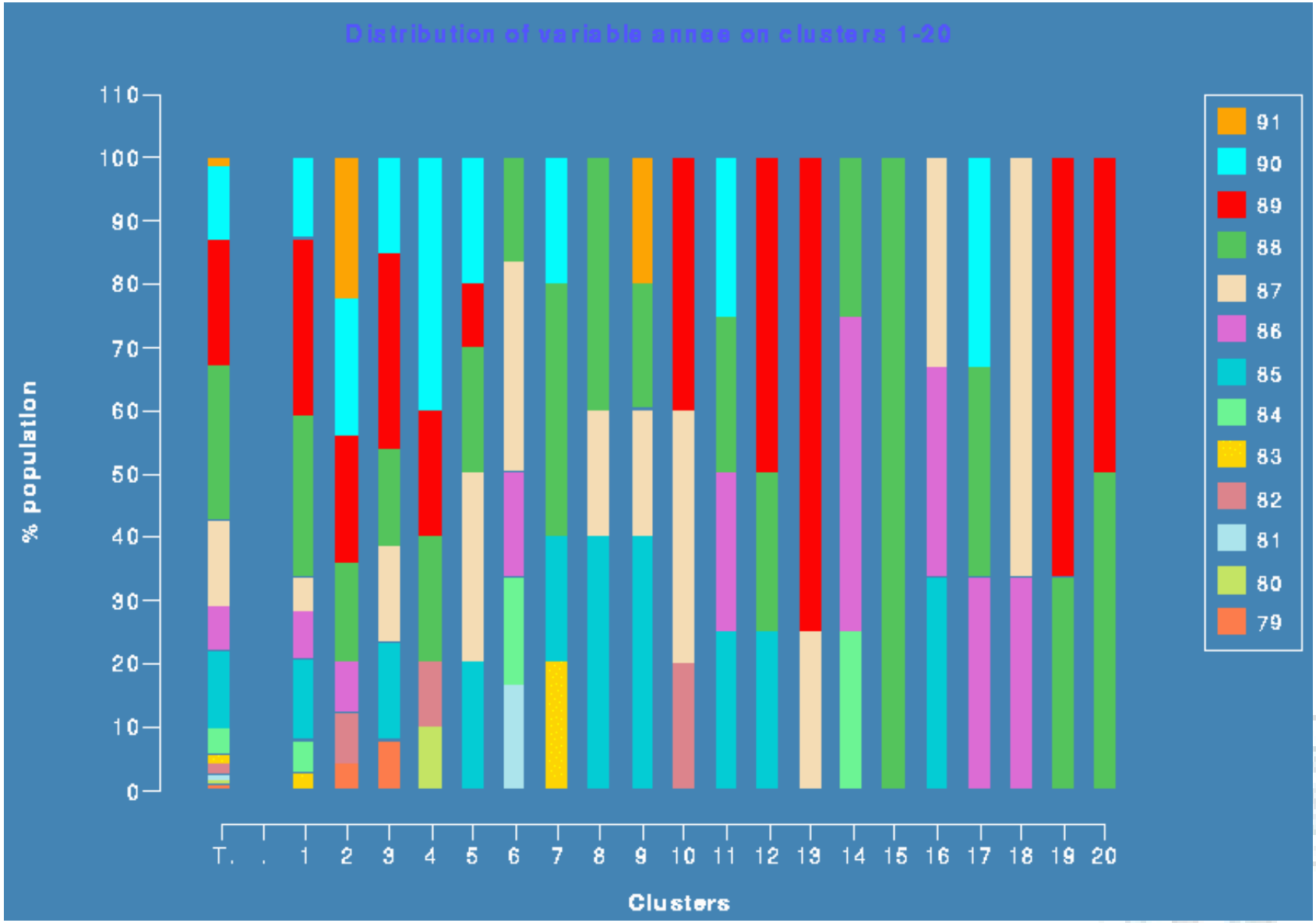
Patch technology- *cluster n.2 -*
attività della concorrenza nel tempo



Activity of competitors in the clusters



Temporal analysis of clusters



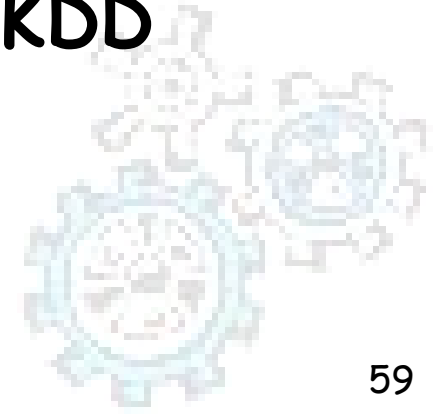
Atherosclerosis prevention study

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:

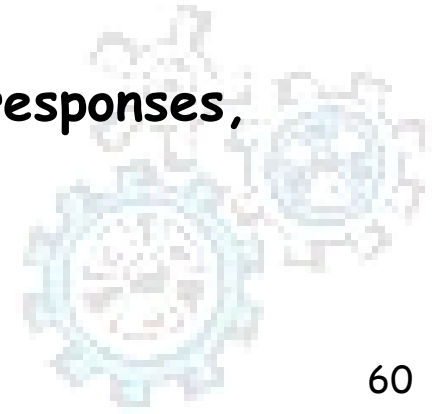
- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.

- 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

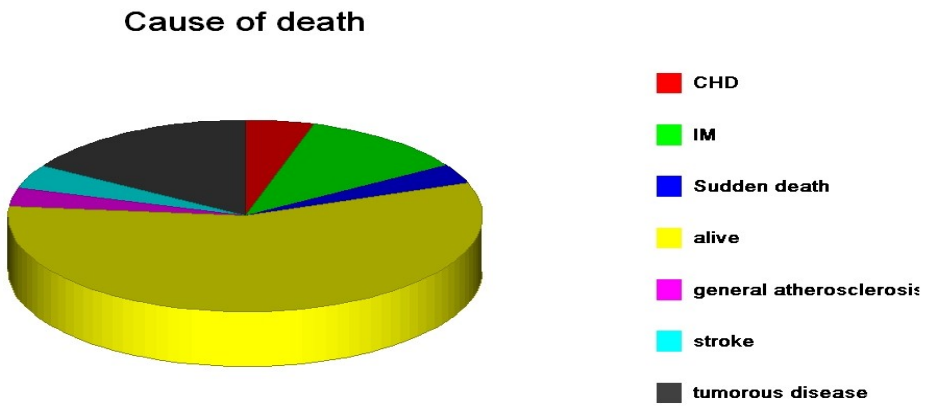
Data from Entry and Exams		
General characteristics	**Examinations**	**habits**
Marital status	Chest pain	Alcohol
Transport to a job	Breathlesness	Liquors
Physical activity in a job	Cholesterol	Beer 10
Activity after a job	Urine	Beer 12
Education	Subscapular	Wine
Responsibility	Triceps	Smoking
Age		Former 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

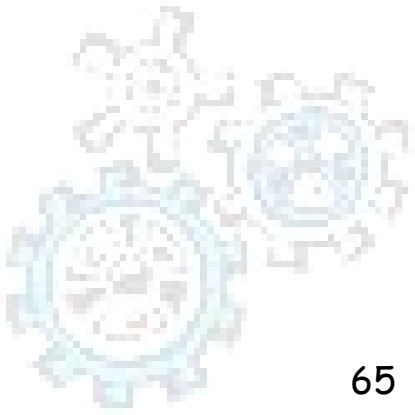
Are there strong relations concerning death cause?

General characteristics (?) ⇒ Death cause (?)

Examinations (?) ⇒ Death cause (?)

Habits (?) ⇒ Death cause (?)

Combinations (?) ⇒ Death cause (?)



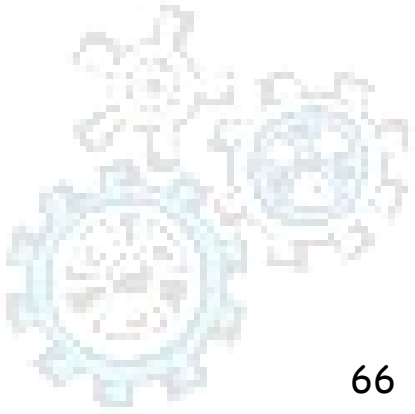
Example of extracted rules

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.**



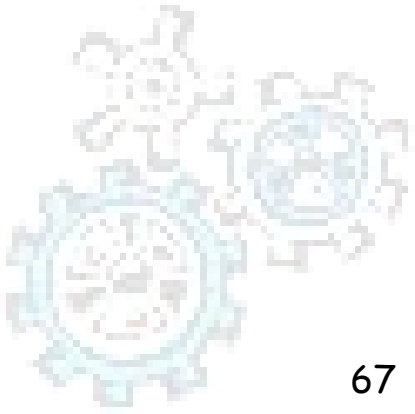
Example of extracted rules

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

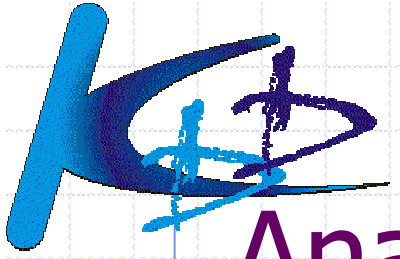
Education(university) & Height<176-180>

⇒

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**



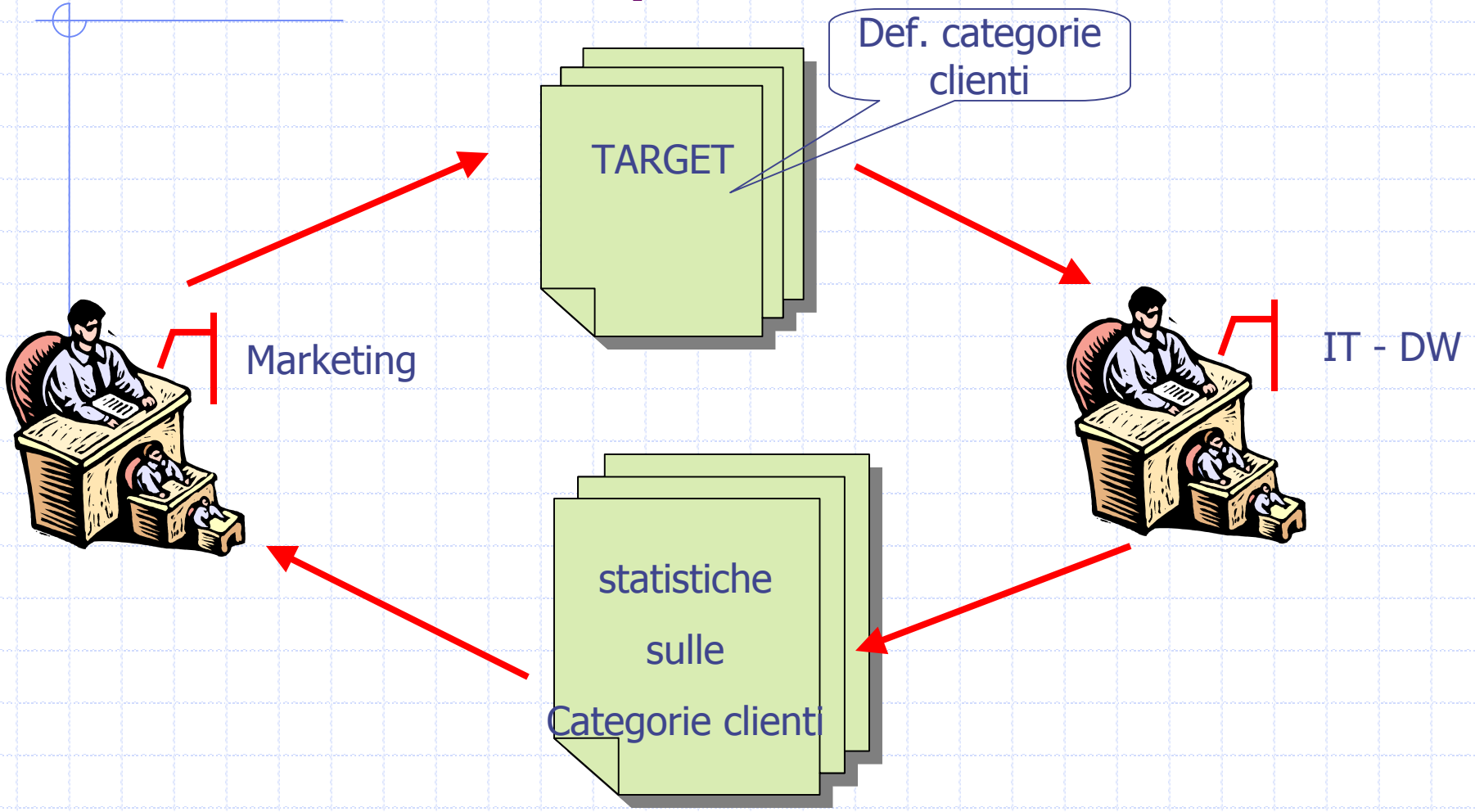
Analisi previsionale per l'ottimizzazione della postalizzazione delle promo

KDD Lab. Pisa

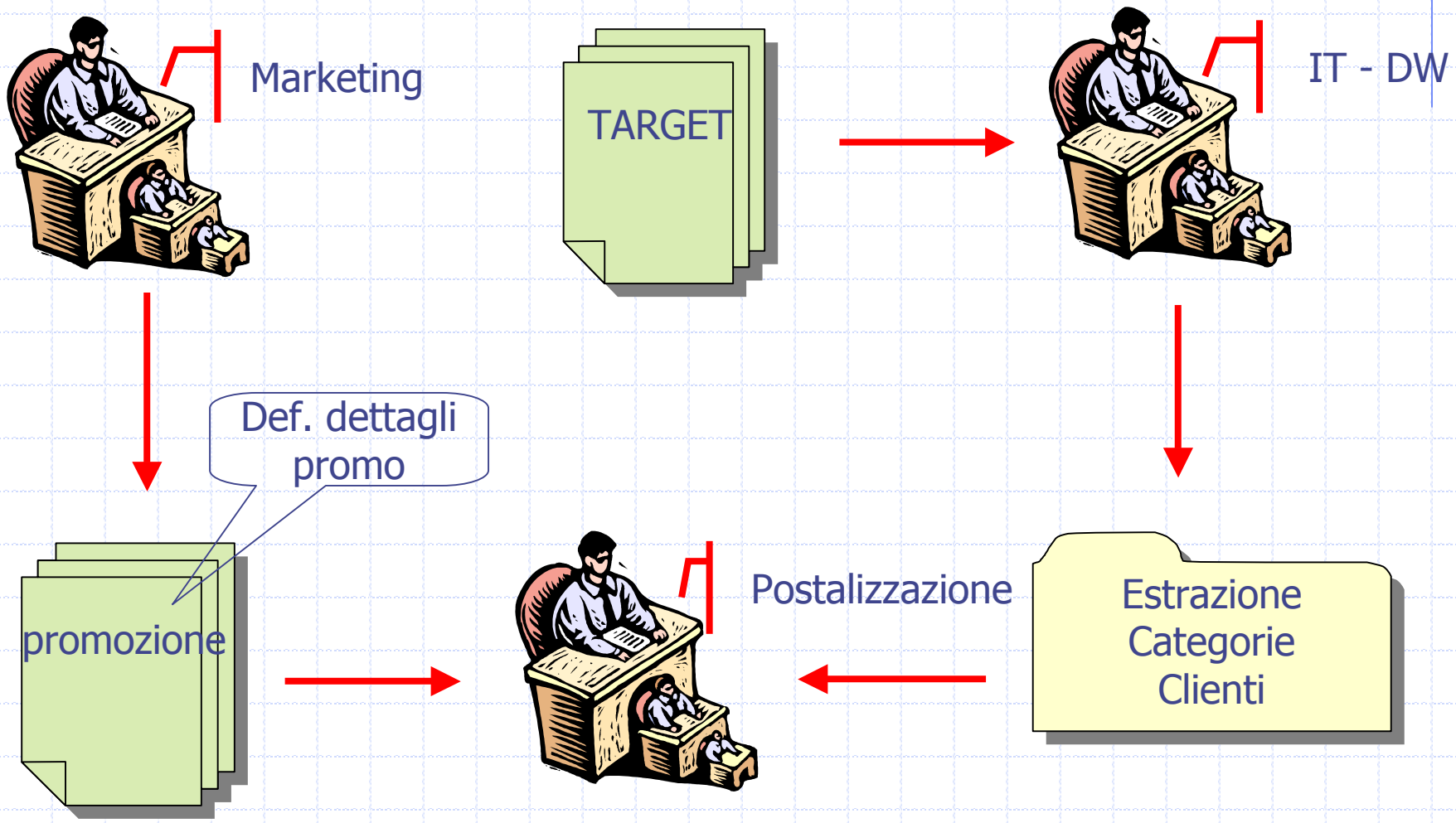
Postalizzazione di promozioni

- 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

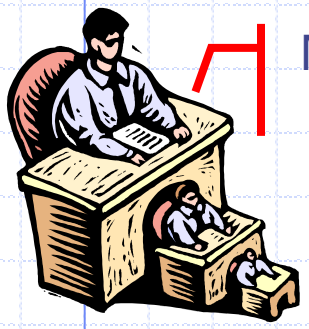
Inventare la promozione



selezionare i clienti e postalizzare



Erogare premi e raccogliere dati



Marketing sviluppo

promozione

Def. dettagli promo

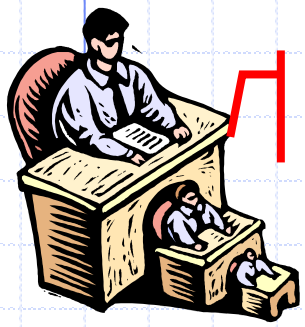
Sistema automatico aggiornamento DATA Mart

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

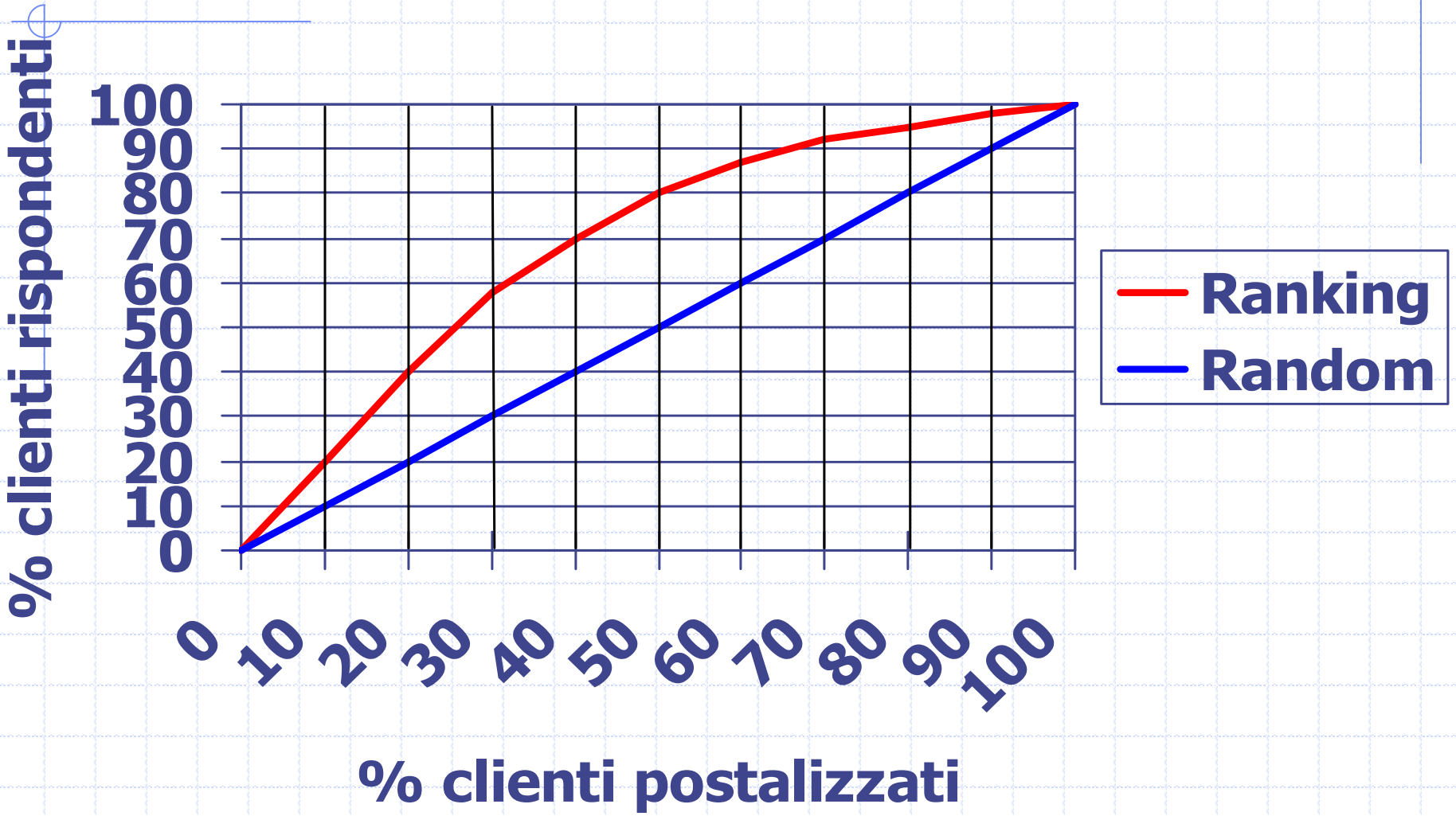
Selezione dei clienti da postalizzare

- 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

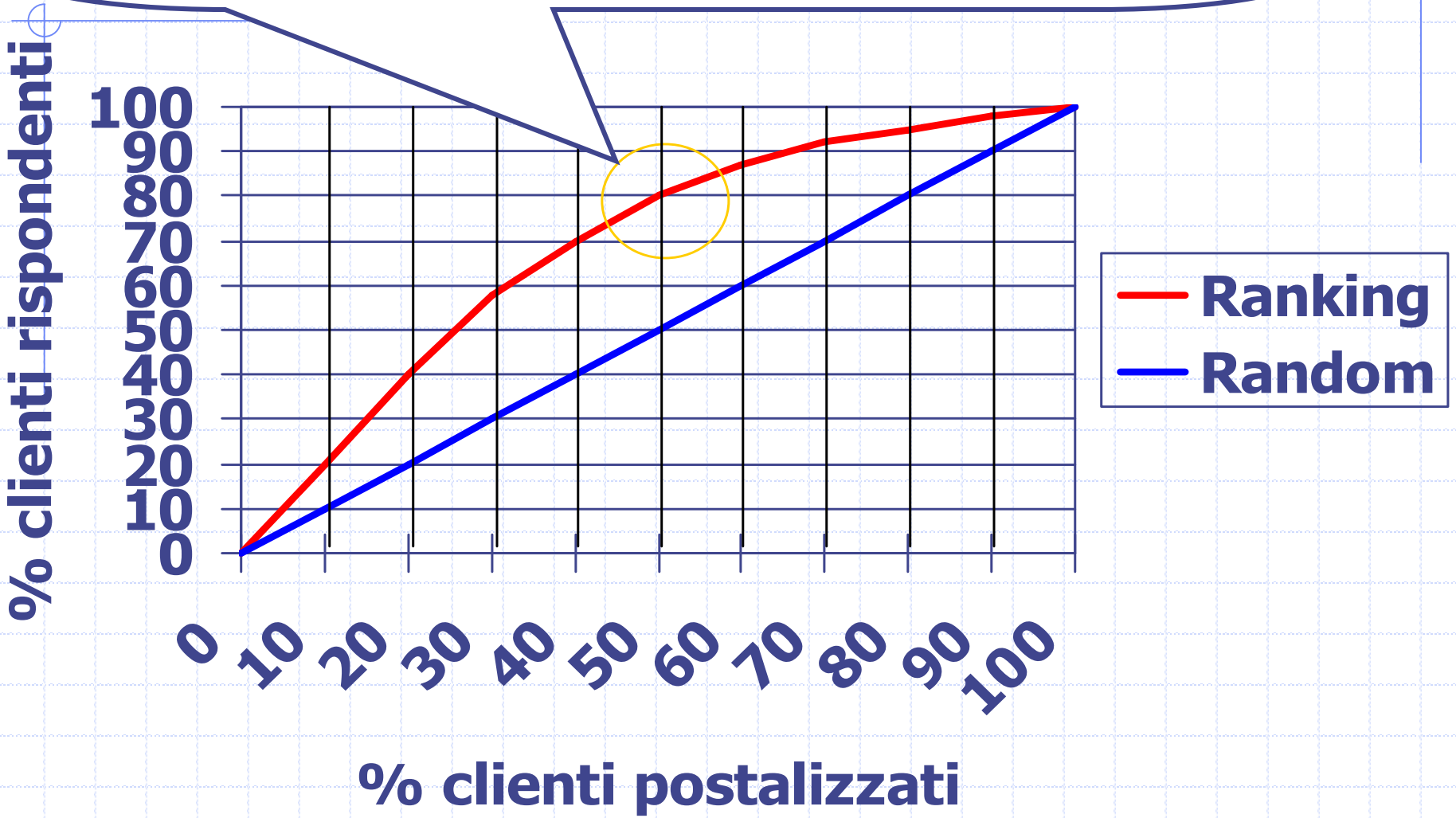
Lift Chart


% 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

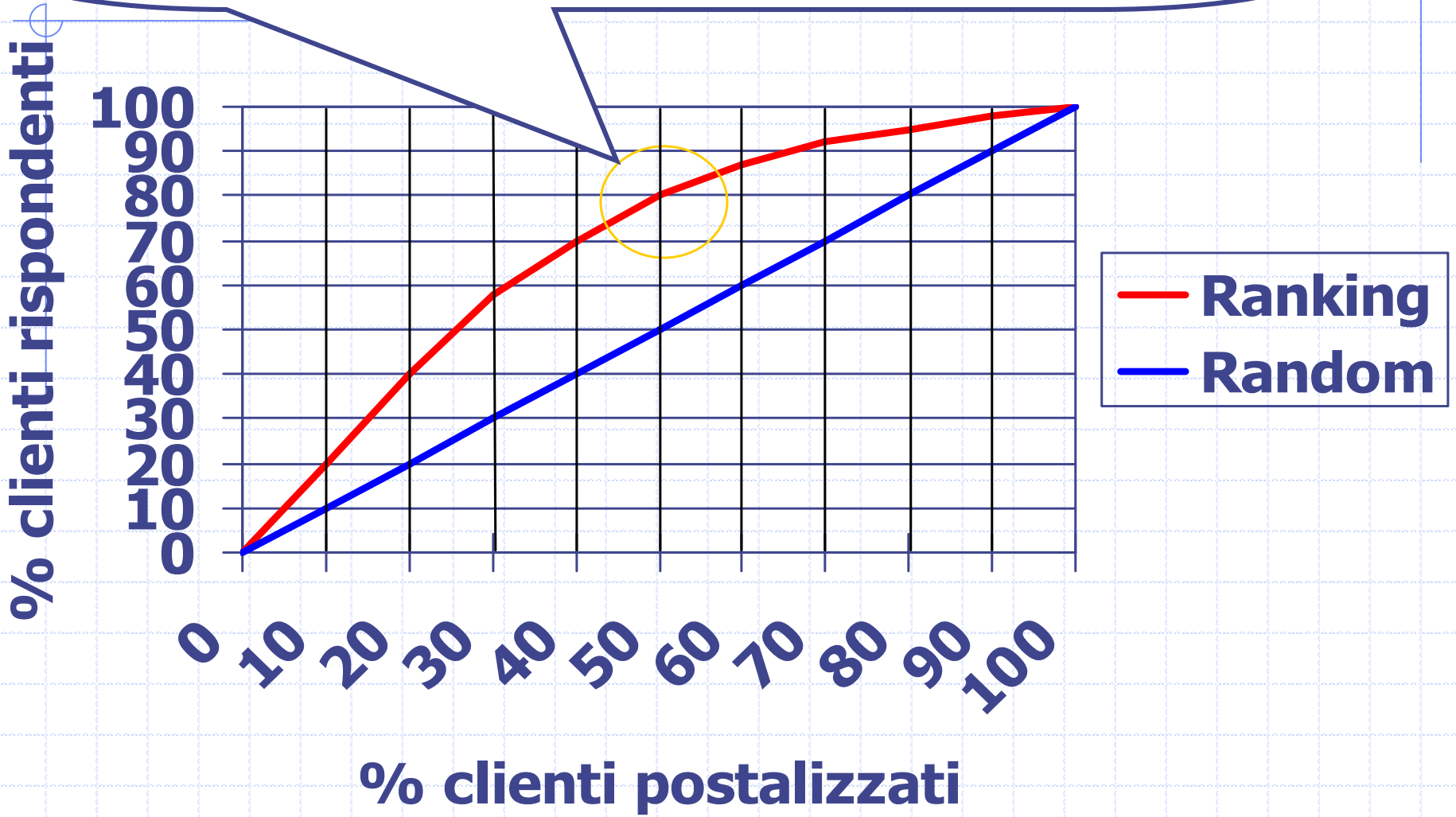
LIFT CHART

- 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)

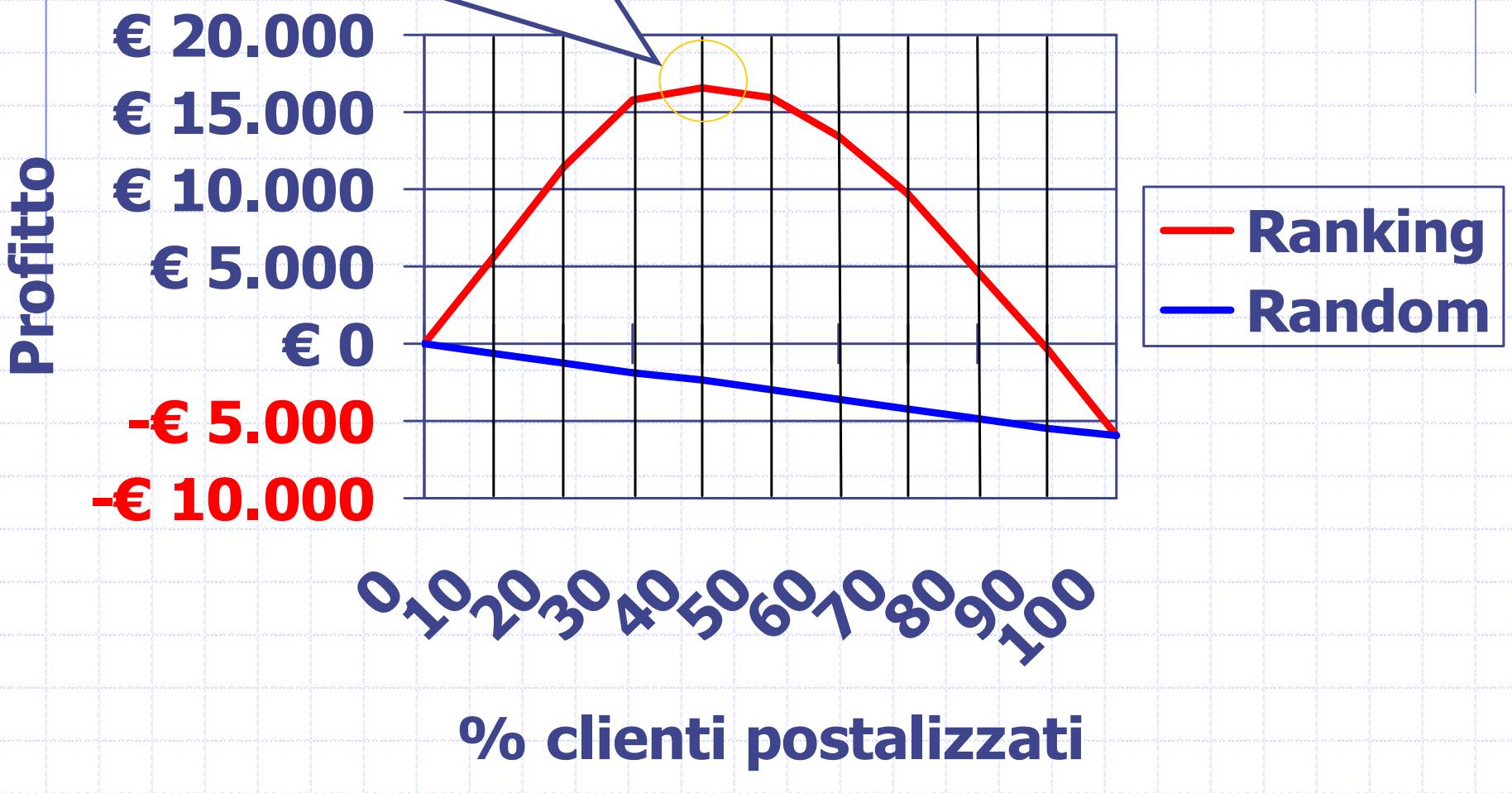
- Il Lift Chart rappresenta un aiuto grafico per ragionare sul rapporto ottimale fra costi di postalizzazione e percentuale di redemption
  - a fronte di sostanziali riduzioni di postalizzati (=budget) permette di ridurre di poco il numero di redenti
  - a parità di budget, permette di incrementare il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