Data Mining2

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 2011/2012

Data Mining

- Acronimo: DM
- Orario: Mrcoledi14-16 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
- Blog per la discussione su privacy & DM
  - Vari articoli e libri messi a disposizione sul wiki per la discussione anna.monreale@isti.cnr.it
  - http://hd.media.mit.edu/wef_globalit.pdf

Data Mining- teoria

- Mining di pattern frequenti e regole associative
- Mining di dati sequenziali,
- Mining di serie temporali ed motifs
- Mining di grandi grafi e reti
- Rilevazione di Anomalie e Outliers.
- Mining di dati spazio temporali (Mobility DM)
- Impatto sociale del data mining - Data mining e protezione della privacy

Data Mining – Casi di studio

- Data Mining e Rilevamento di frodi:
  - Sogei1, DIVA (progetto 1)
- Data Mining per il CRM
  - 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: i flussi turistici.
- E-commerce
  - analisi da dati da siti E-comemrce: e-marketing
- Mobilità e trasporti:
  - esplorazione, e postprocessing per la validazione dei comportamenti di mobilità: progetto3

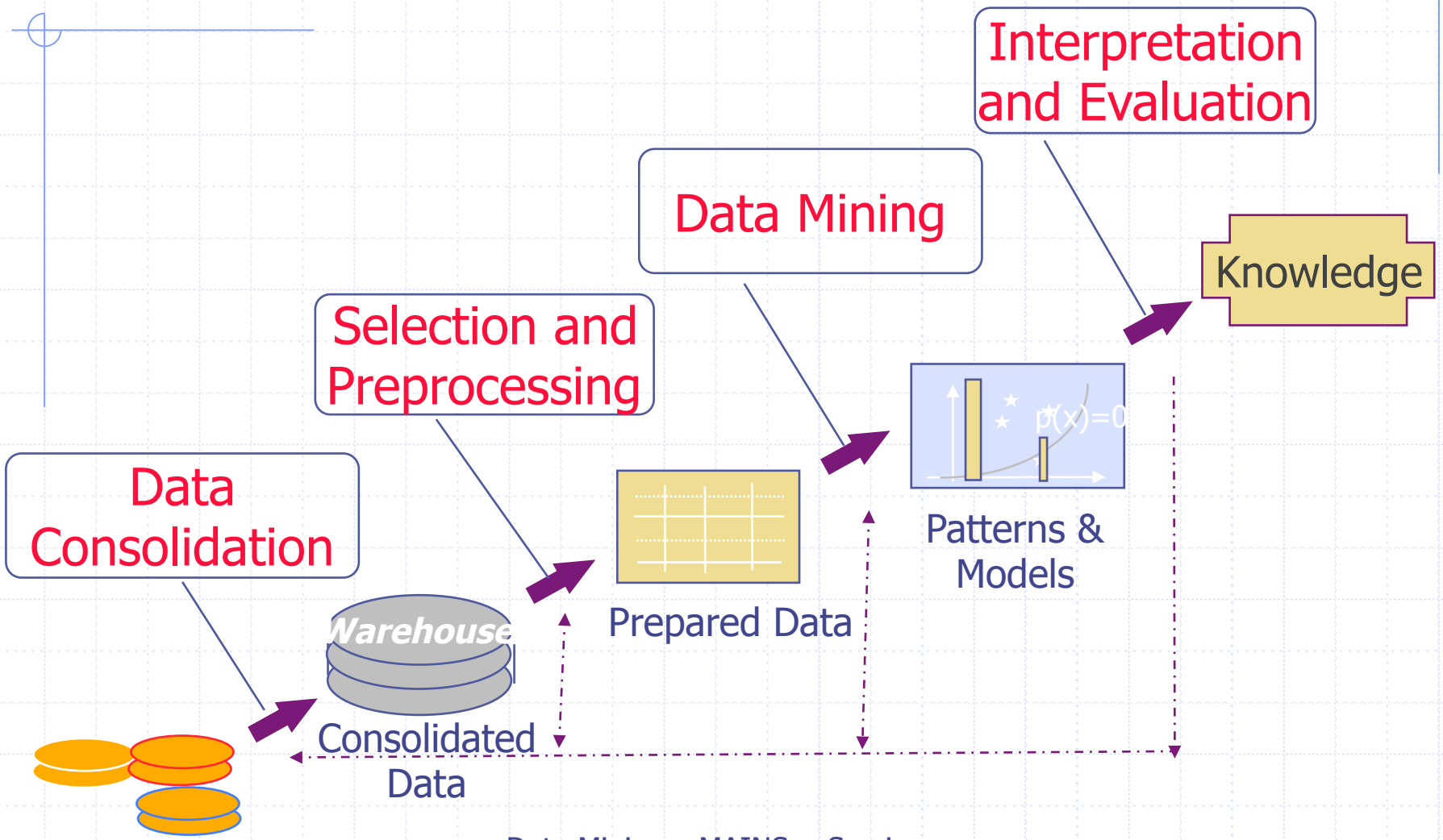
Modalità di valutazione

- Progetti in itinere (Analisi di piccoli datasets)
- Progetto finale
  - 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

Sommario lezione 1

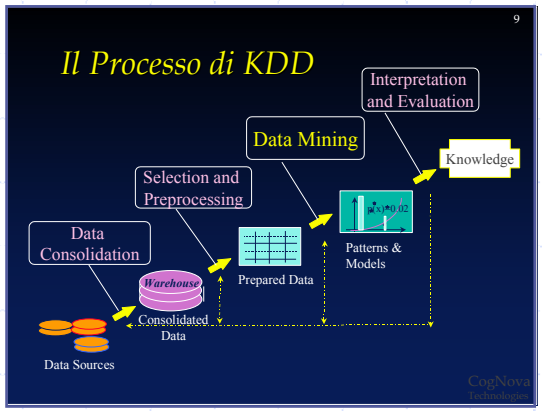
- Il processo KDD
- Es.1. Individuazione ed uso di segmentazione di clienti
- Es. 2. Ottimizzazione di servizio di marketing
- Il CRISP model

The KDD process



Il ciclo *virtuoso della filiera BI*

Problema



Conoscenza

Identificare il problema e le opportunità

Utilizzare la conoscenza

Strategia

Misurare gli effetti dell' azione

Risultati

AIR MILES

un caso di studio di customer segmentation

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

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

Clustering & segmentazione dei clienti

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

Segmenti

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

Segmenti di comportamento di acquisto

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

The Loyalty Group in Canada

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

Acquisizione dei dati

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

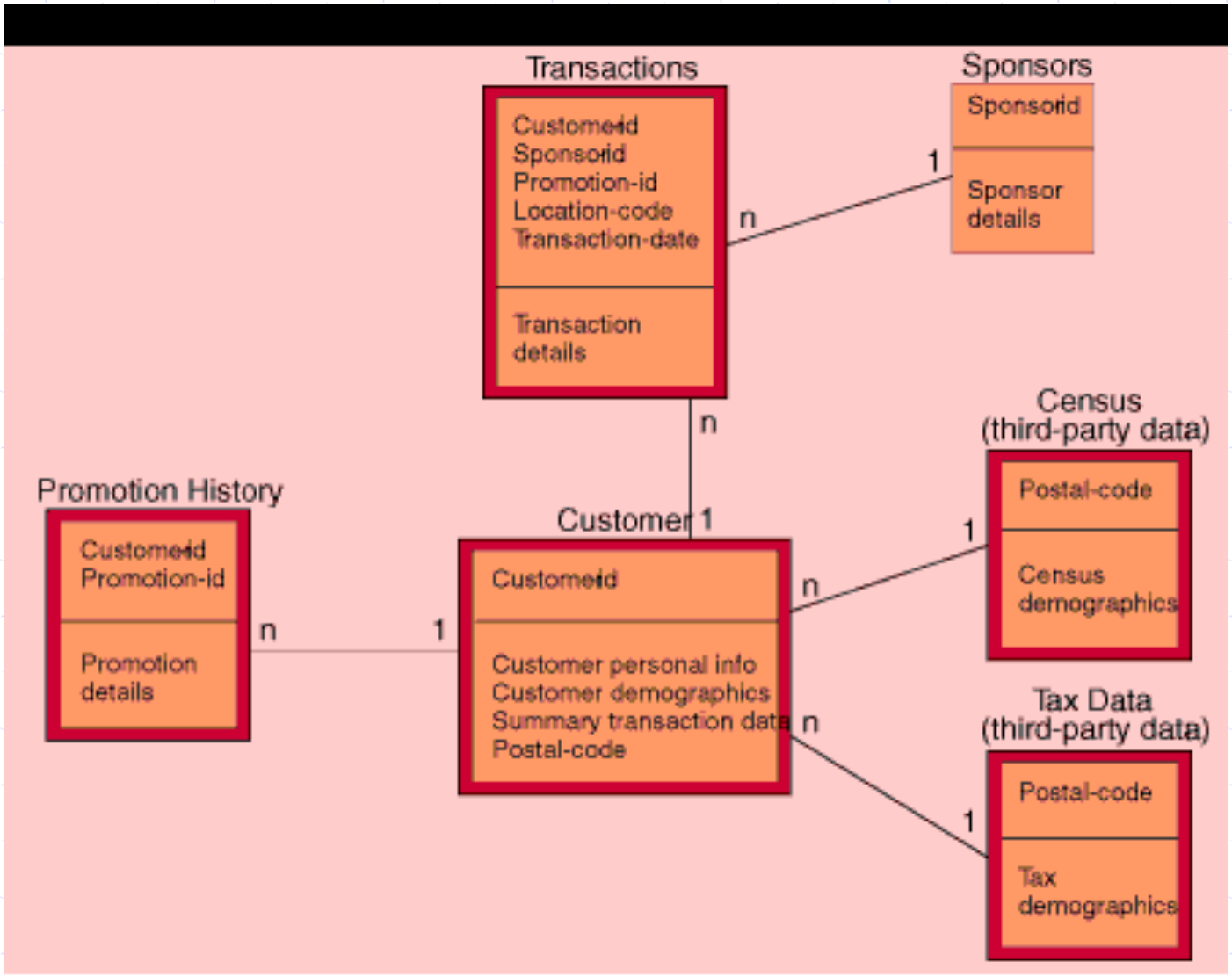
Stato dell'arte prima del data mining

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

Una esperienza di Data mining

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

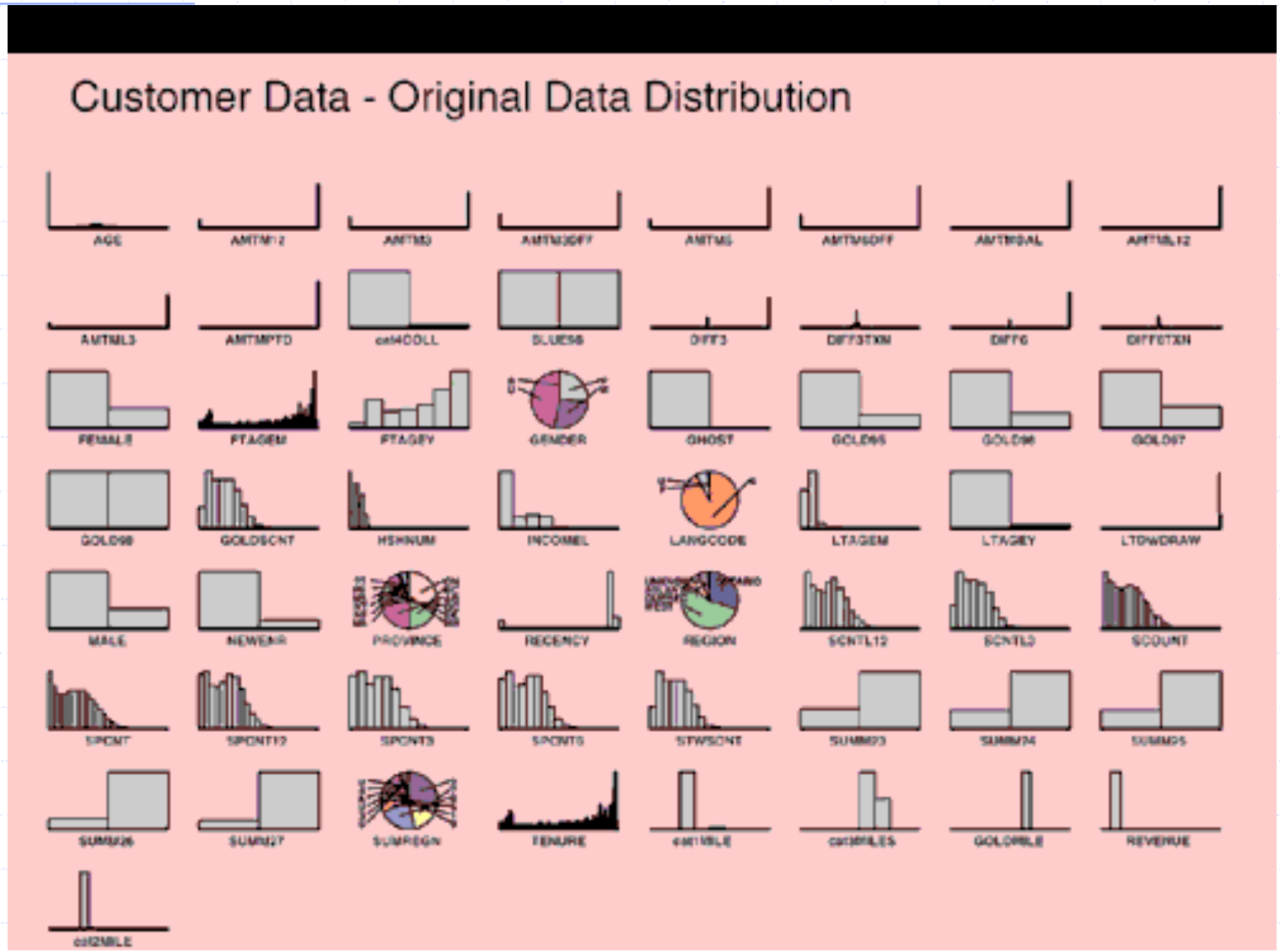
Sorgente dei dati nel DW



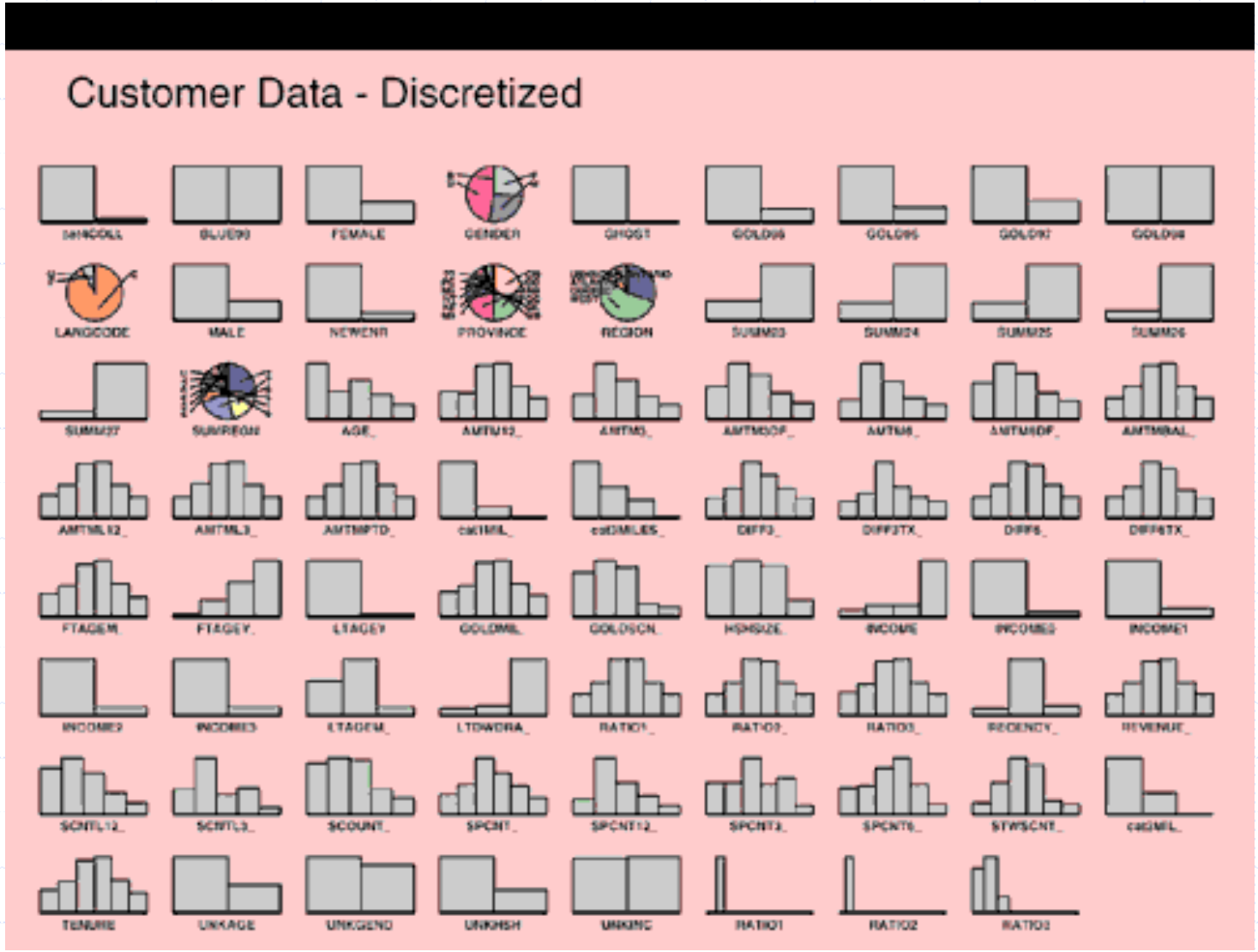
Preparazione dei dati

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

I dolori della pulizia dei dati: prima …



… e dopo la cura



Prima e dopo la cura

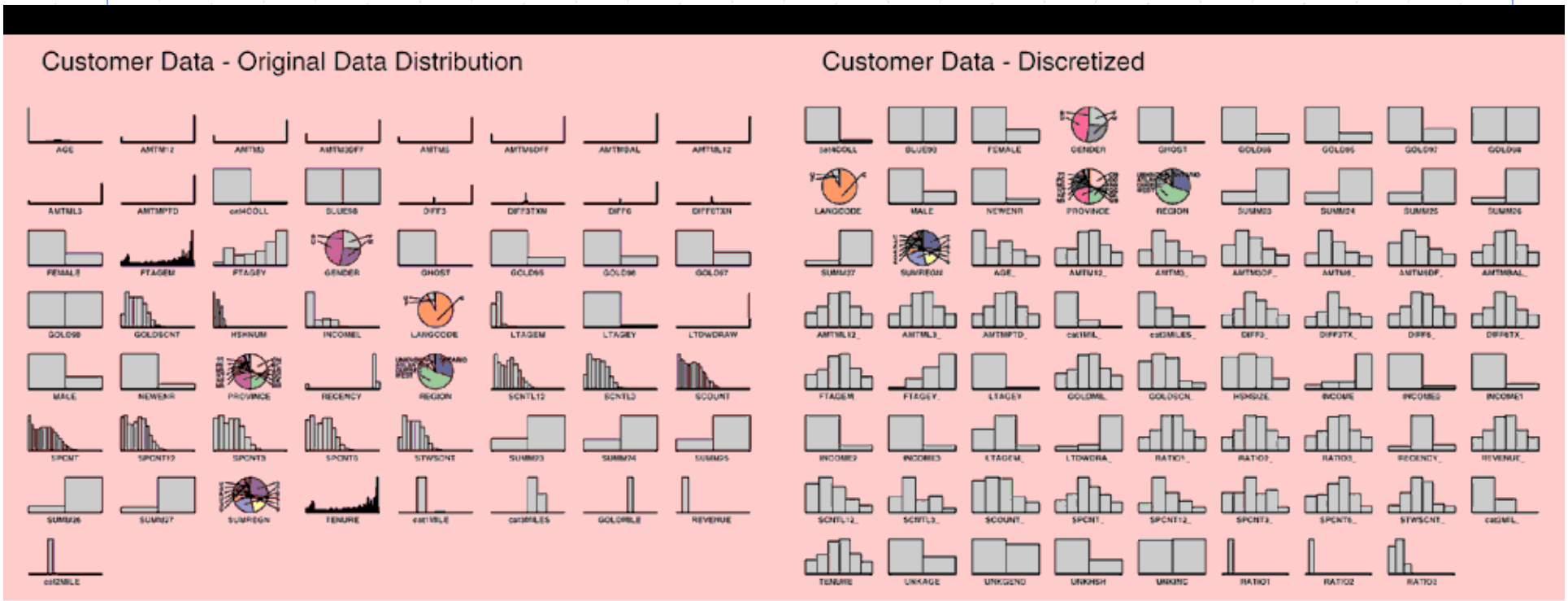
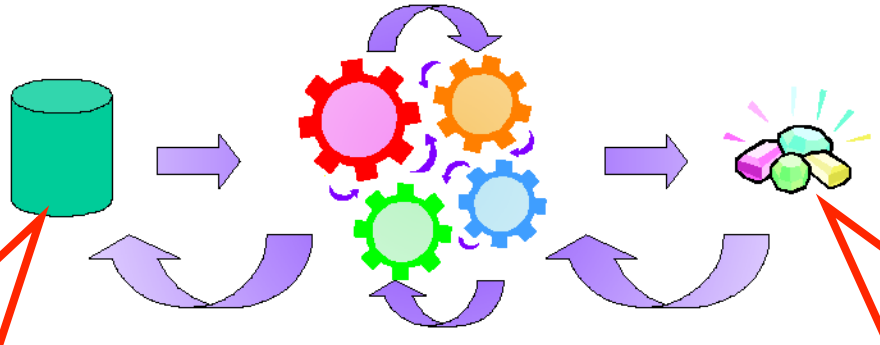


Figure3. *Original data.*

Figure 4. *Discretized data.*

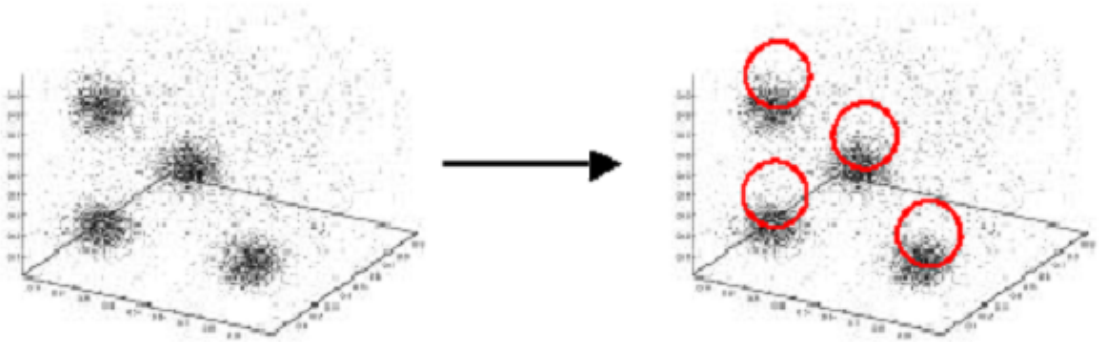
Estrazione del modello di clustering

Clustering = raggruppamento di oggetti simili in gruppi omogenei

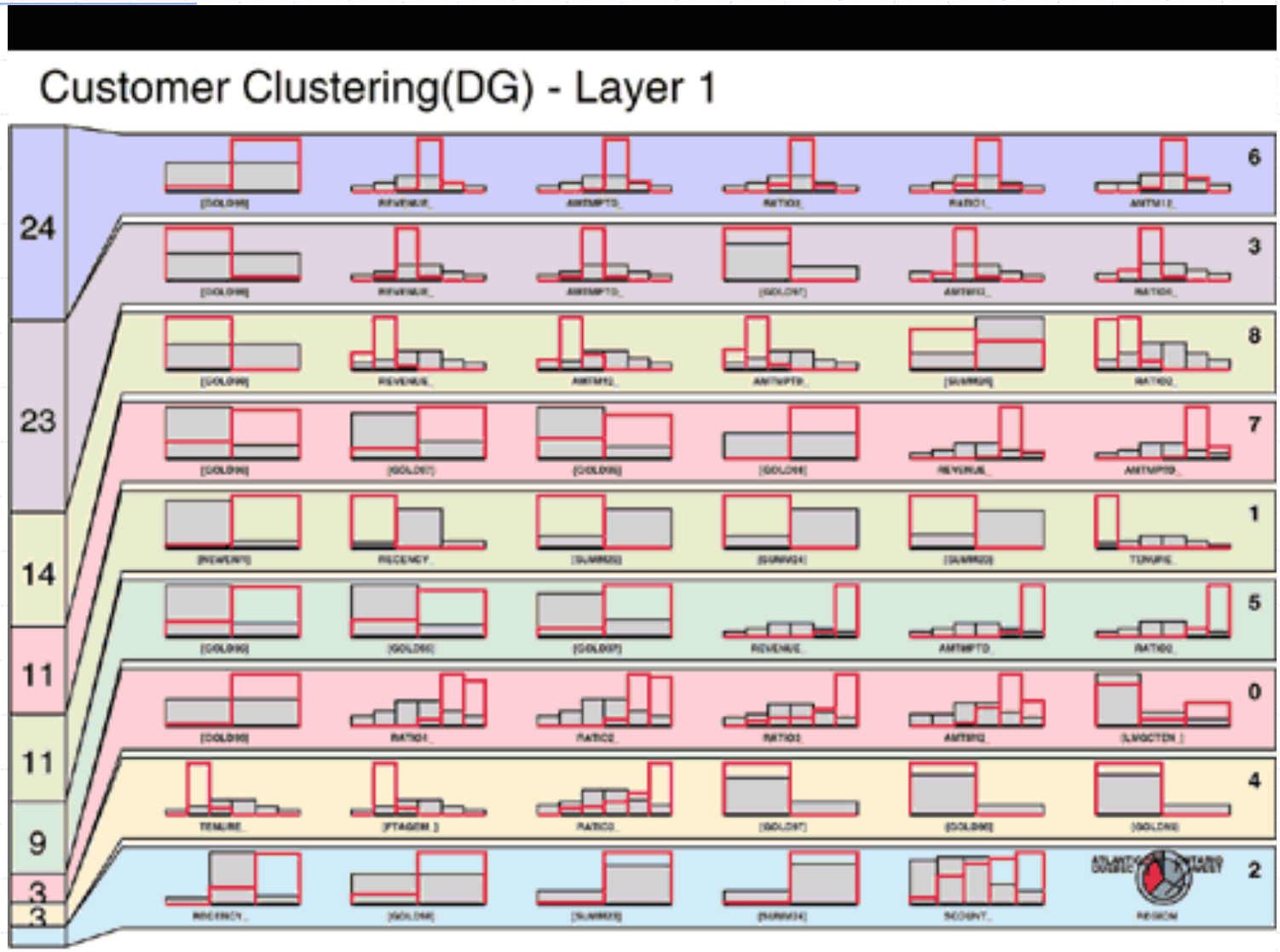


Dati in input: variabili economiche di ciascun cliente

Modello in output: segmentazione dei clienti simili in clusters



Output del clustering



Analisi qualitativa dei cluster

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

Analisi qualitativa dei cluster

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

Analisi del cluster 5 – clienti Platinum

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

Vista dettagliata del cluster 5

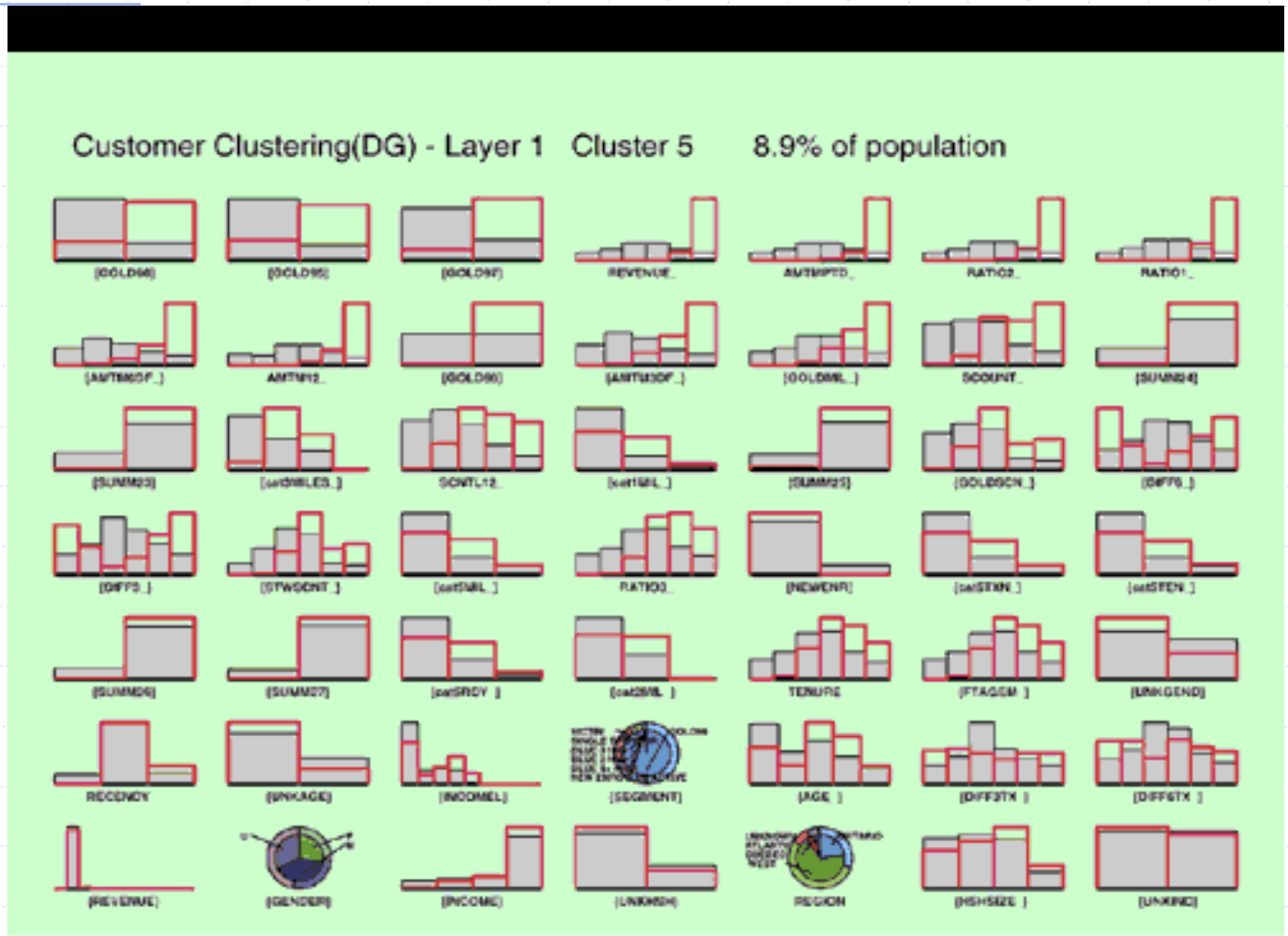


Figure 8. *Cluster 5 output.*

Analisi dei cluster

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

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

Table 1. *Profiling a cluster.*

Analisi dei cluster

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

Opportunità di business

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

Opportunità di business (2)

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

Una buona pratica di mining

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

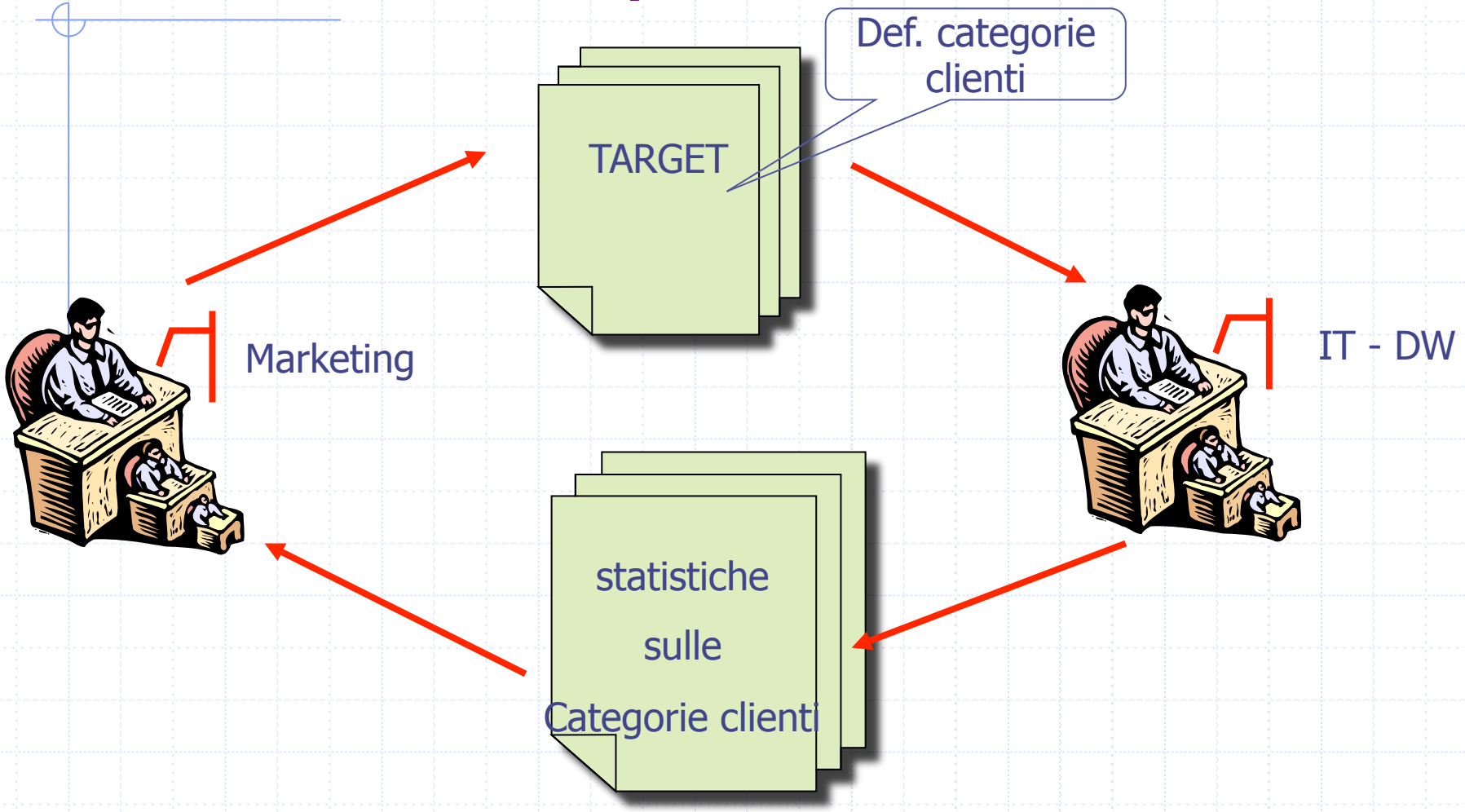
Analisi previsionale per l'ottimizzazione della postalizzazione delle promo

KDD Lab. Pisa

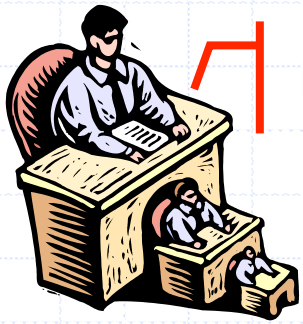
Postalizzazione di promozioni

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

Inventare la promozione

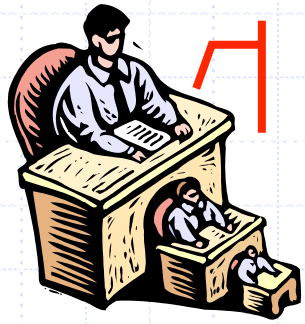


selezionare i clienti e postalizzare



Marketing

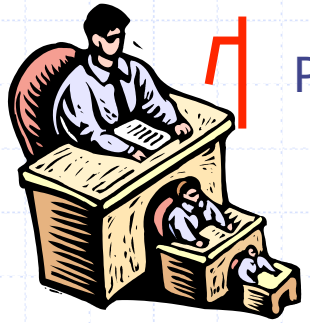
TARGET



IT - DW

Def. dettagli promo

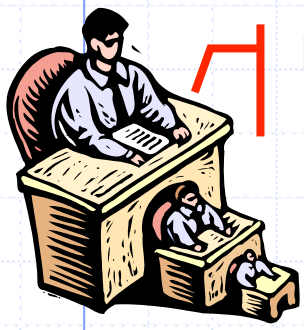
promozione



Postalizzazione

Estrazione Categorie Clienti

Erogare premi e raccogliere dati

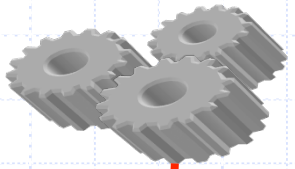


Marketing sviluppo

promozione

Def. dettagli promo

Sistema automatico aggiornamento DATA Mart

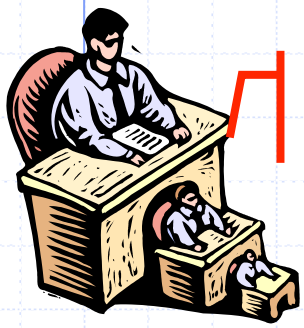


Mercati

Sistema automatico attivazione mercati

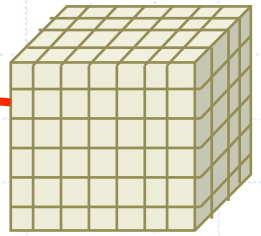
DATA MART PROMOZIONI

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

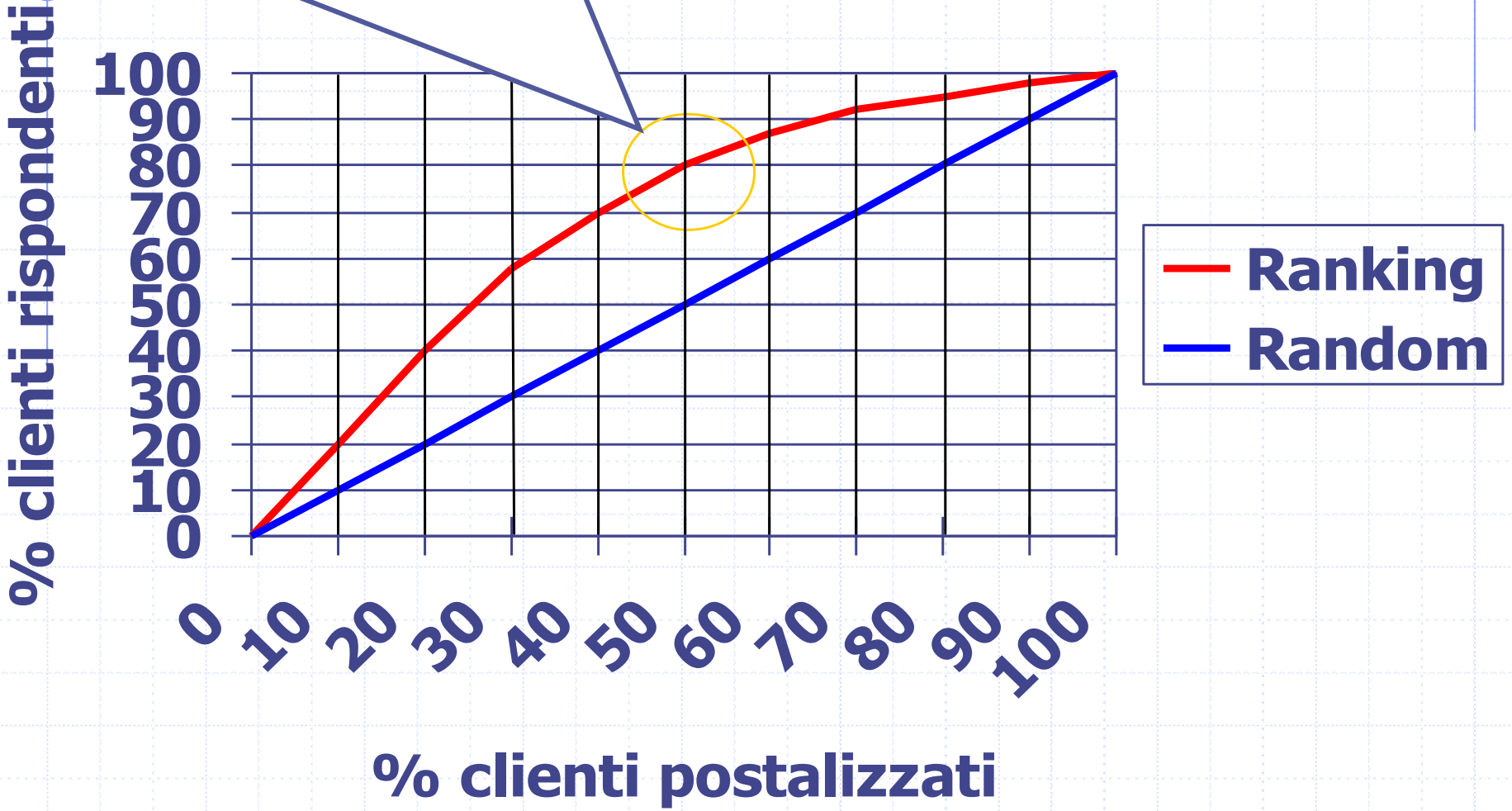

% clienti rispondenti
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
% clienti postalizzati
Ranking
Random

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)

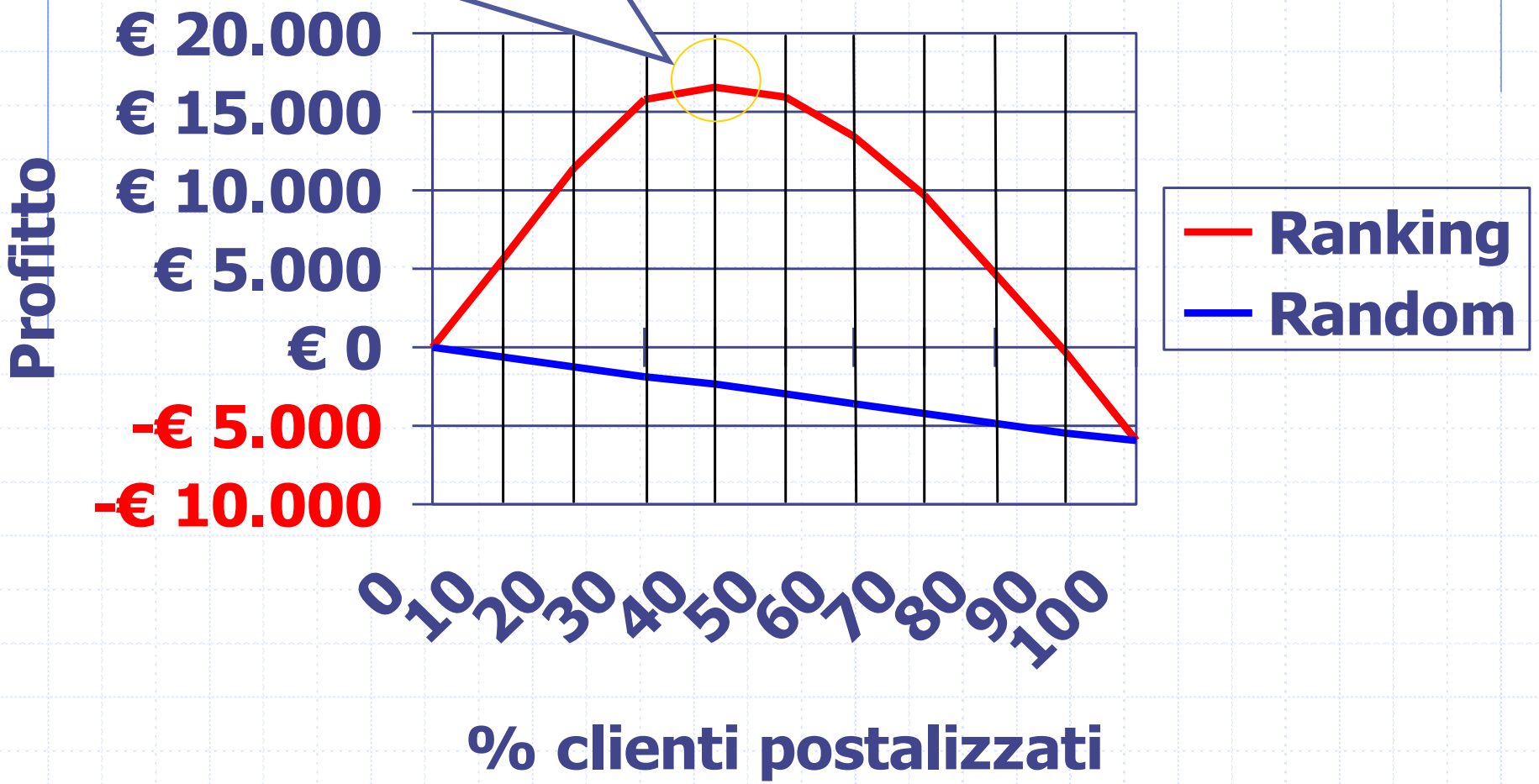
- Il Lift Chart rappresenta un aiuto grafico per ragionare sul rapporto ottimale fra costi di postalizzazione e percentuale di redemption
  - a fronte di sostanziali riduzioni di postalizzati (=budget) permette di ridurre di poco il numero di redenti
  - a parità di budget, permette di incrementare il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

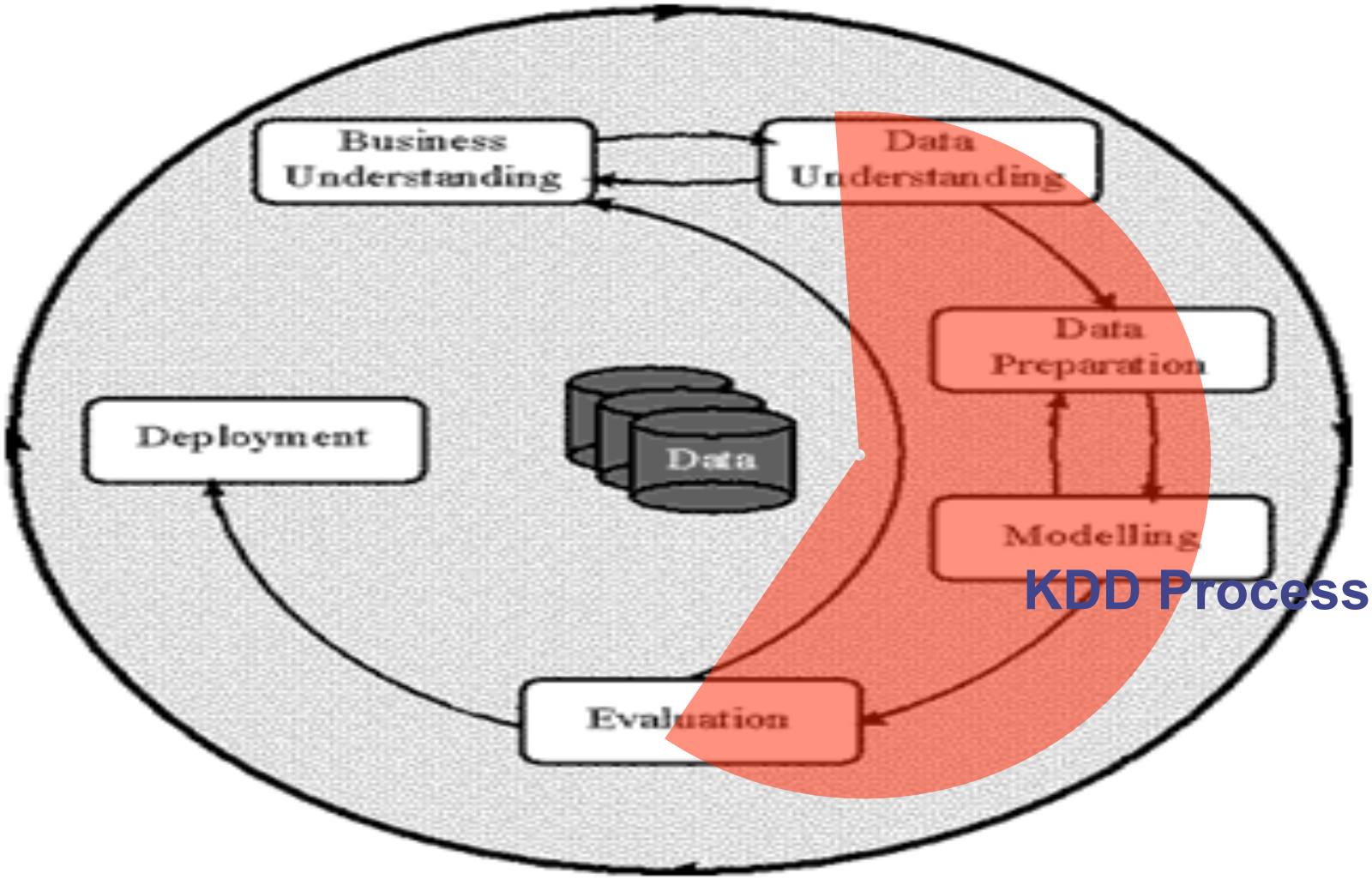
Dietro le quinte

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

On-line

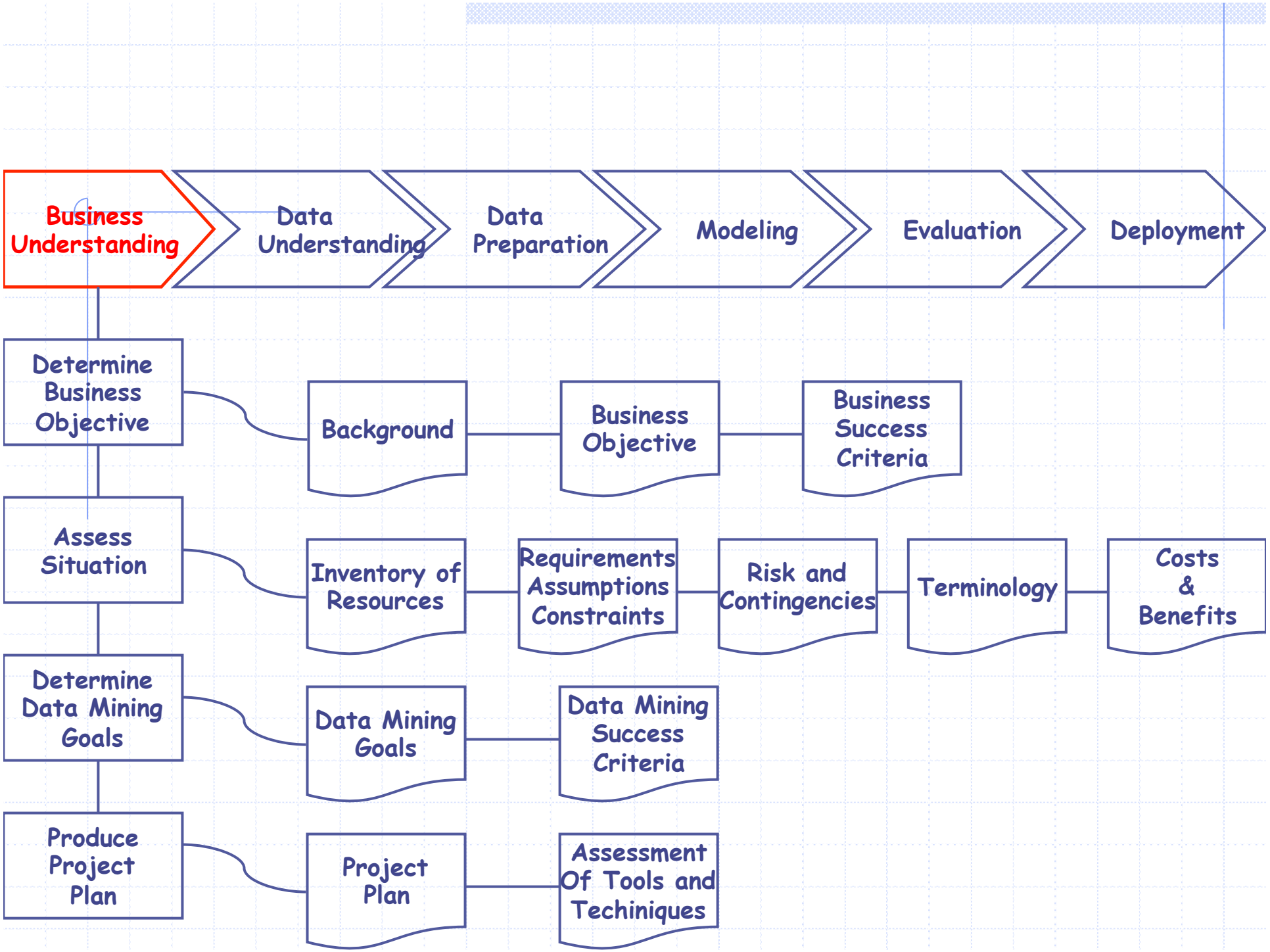
Off-line

CRISP-DM: The life cicle of a data mining project



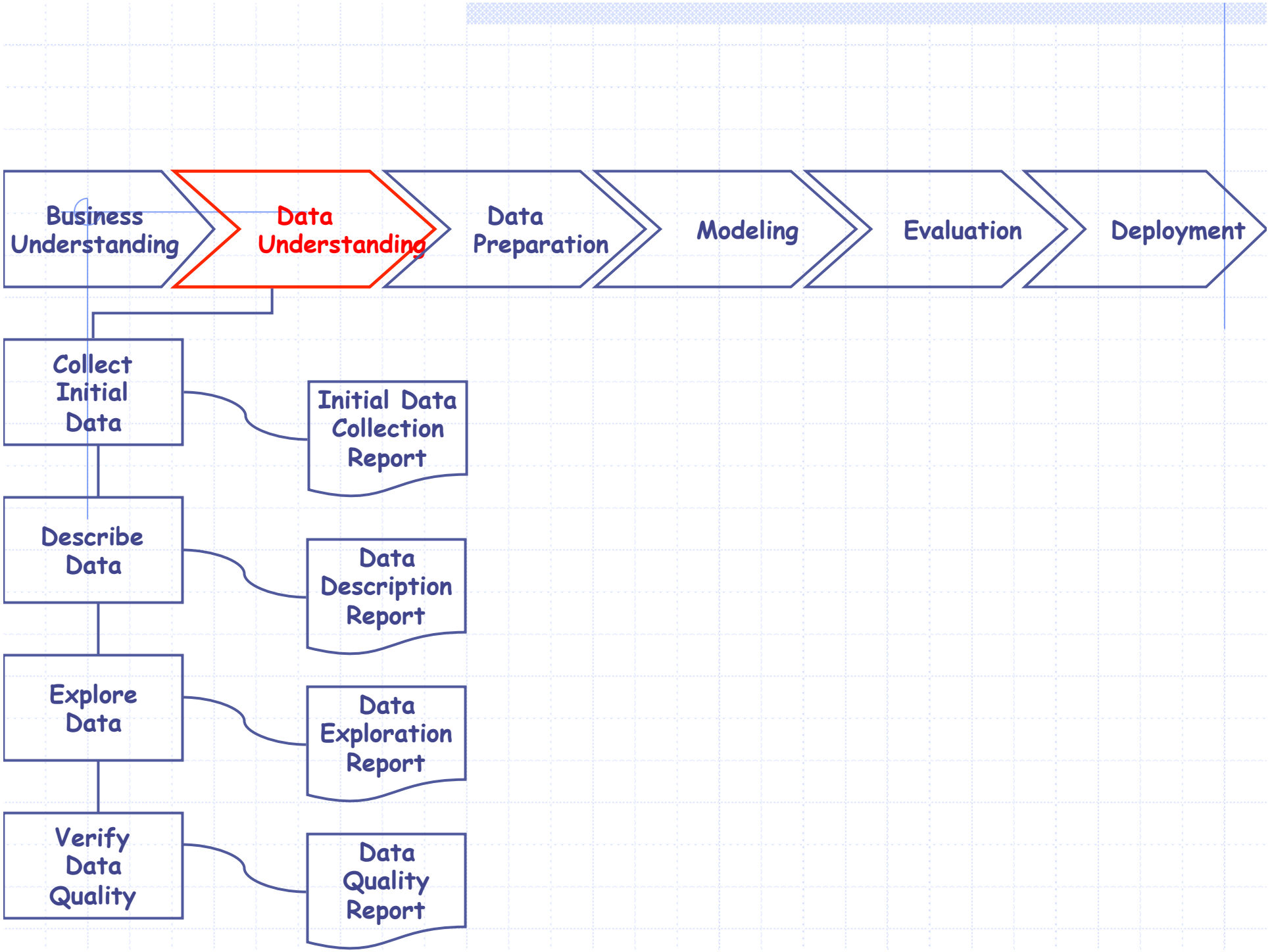
Business understanding

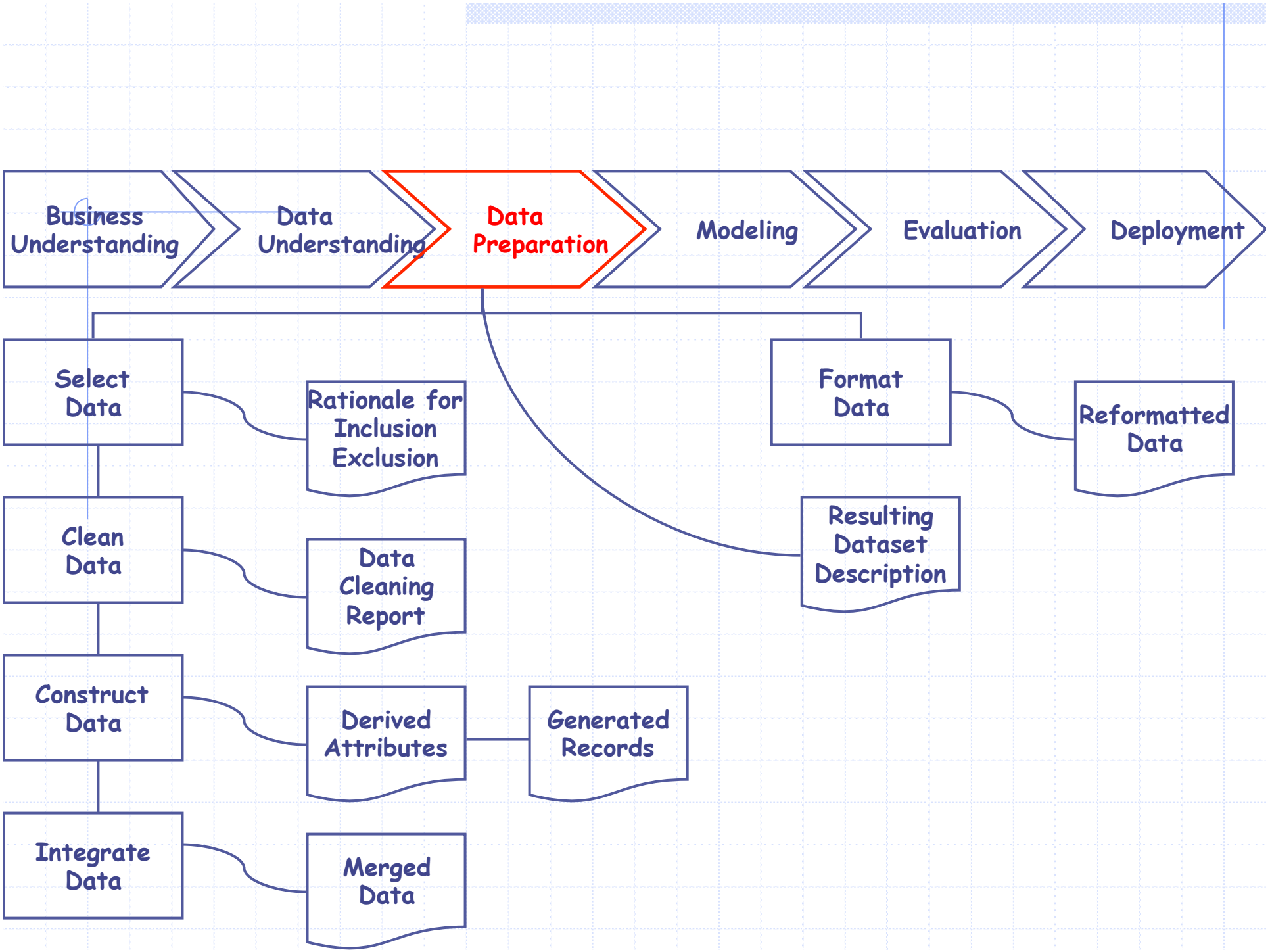
- Understanding the project objectives and requirements from a business perspective.
- then converting this knowledge into a data mining problem definition and a preliminary plan.
  - **Determine the Business Objectives**
  - **Determine Data requirements for Business Objectives**
  - **Translate Business questions into Data Mining Objective**


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Determine Business Objective
Background
Business Objective
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements Assumptions Constraints
Risk and Contingencies
Terminology
Costs & Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Assessment Of Tools and Techiniques

Data understanding

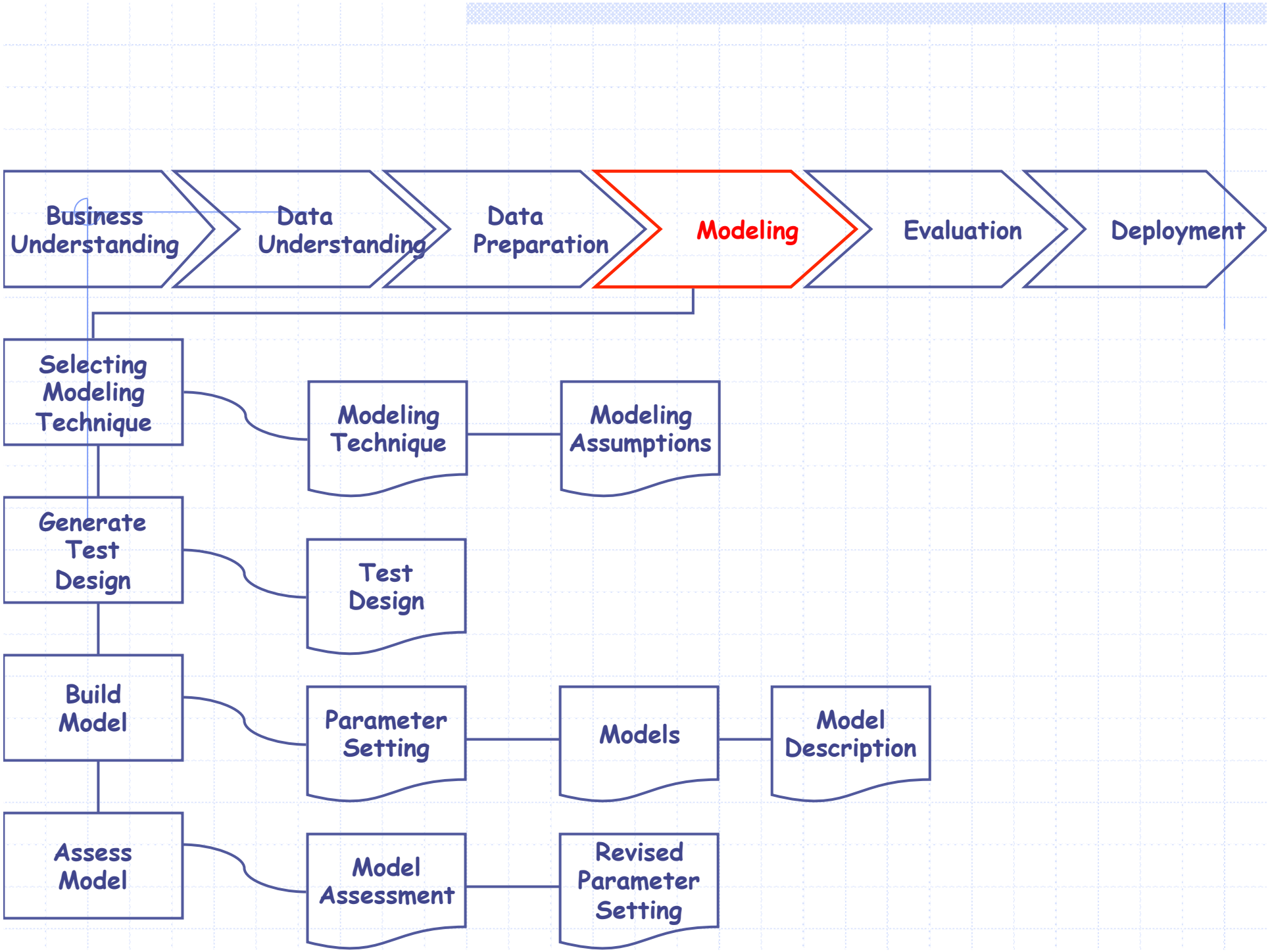
- **Data understanding:** characterize data available for modelling. Provide assessment and verification for data.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Select Data
Rationale for Inclusion Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data
Resulting Dataset Description

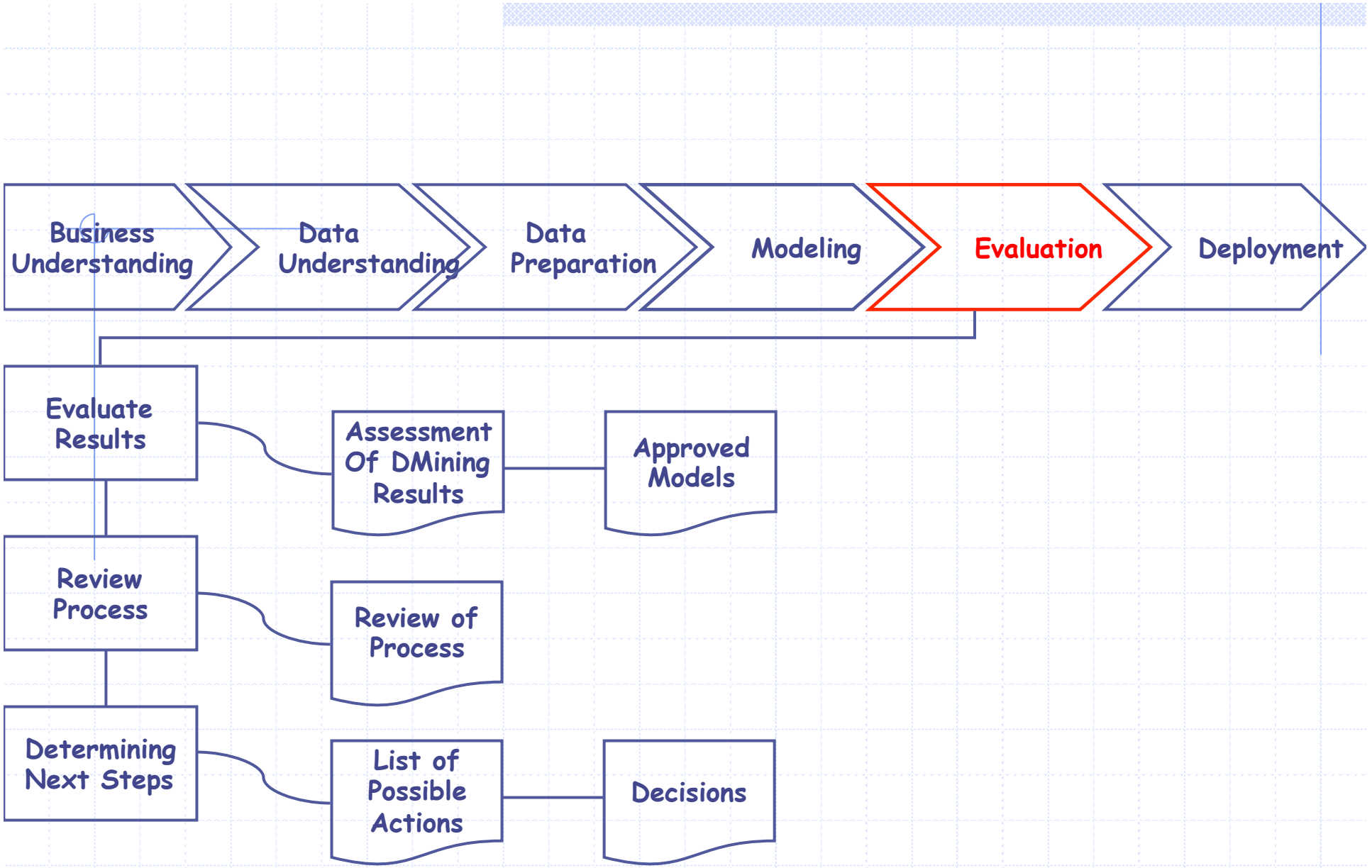
Modeling:

- In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
- Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data.
- Therefore, stepping back to the data preparation phase is often necessary.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Selecting Modeling Technique
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Setting
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Setting

Evaluation

- At this stage in the project you have built a model (or models) that appears to have high quality from a data analysis perspective.
- Evaluate the model and review the steps executed to construct the model to be certain it properly achieves the business objectives.
- A key objective is to determine if there is some important business issue that has not been sufficiently considered.

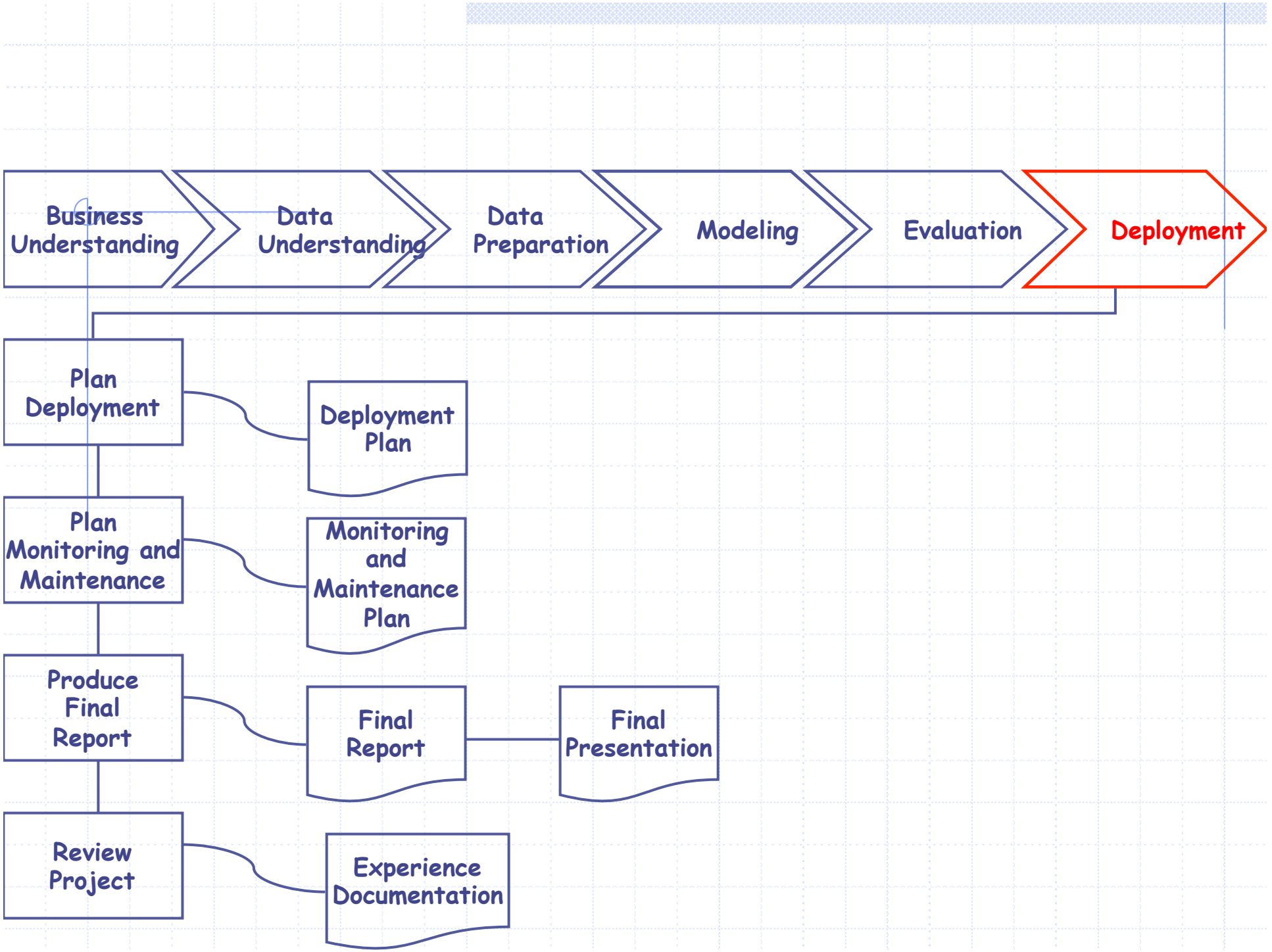

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Evaluate Results
Assessment Of DMining Results
Approved Models
Review Process
Review of Process
Determining Next Steps
List of Possible Actions
Decisions

Deployment:

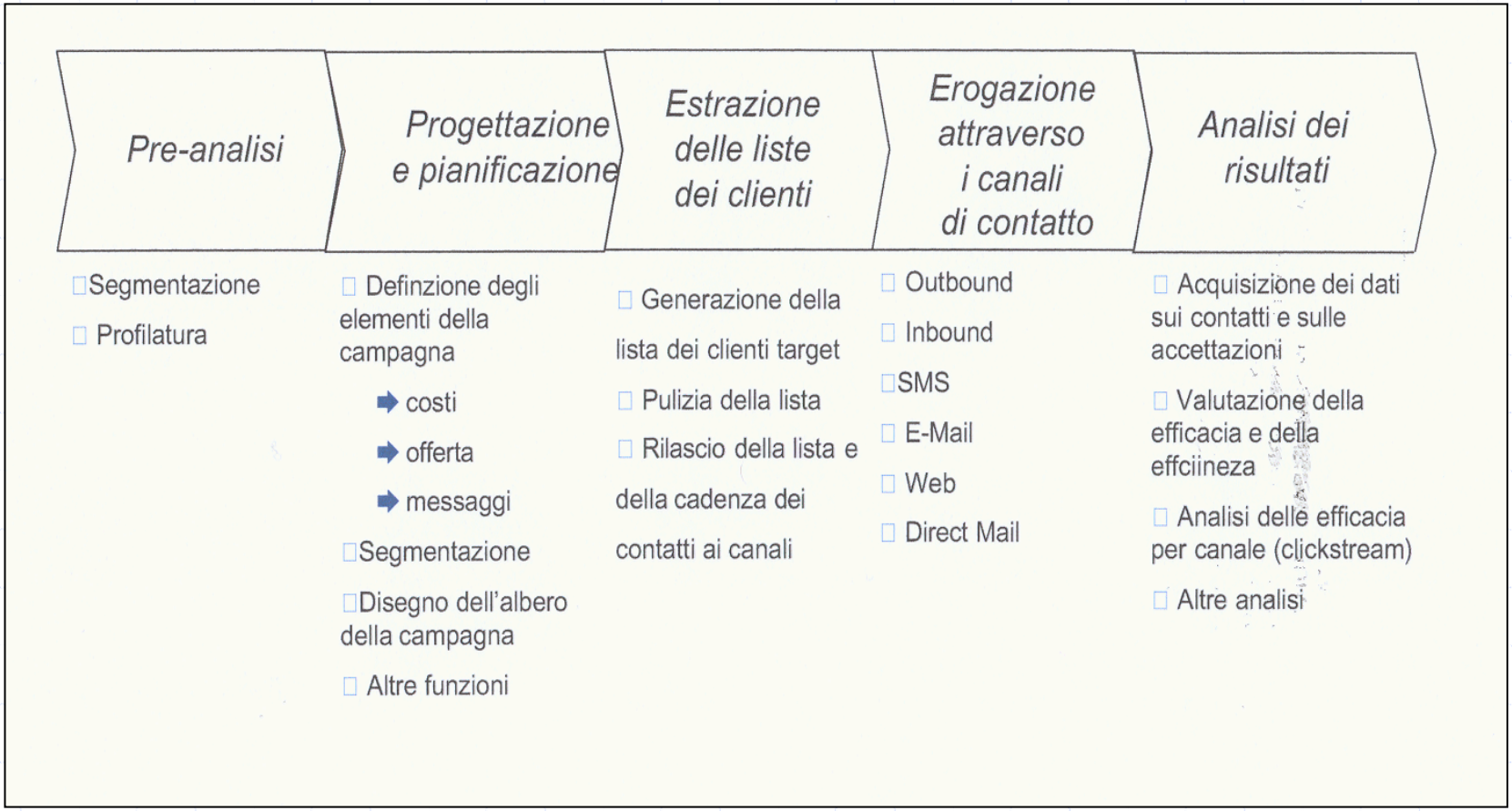
- The knowledge gained will need to be organized and presented in a way that the customer can use it.
- It often involves applying "live" models within an organization's decision making processes, for example in real-time personalization of Web pages or repeated scoring of marketing databases.

Deployment:

- It can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.
- In many cases it is the customer, not the data analyst, who carries out the deployment steps.


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

Es: Automatic Target Marketing



Mining Based Decision Support System: Adaptive Architecture