Exemplar Case Studies

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

Master BigData 2018

Outline

Case studies from real data:

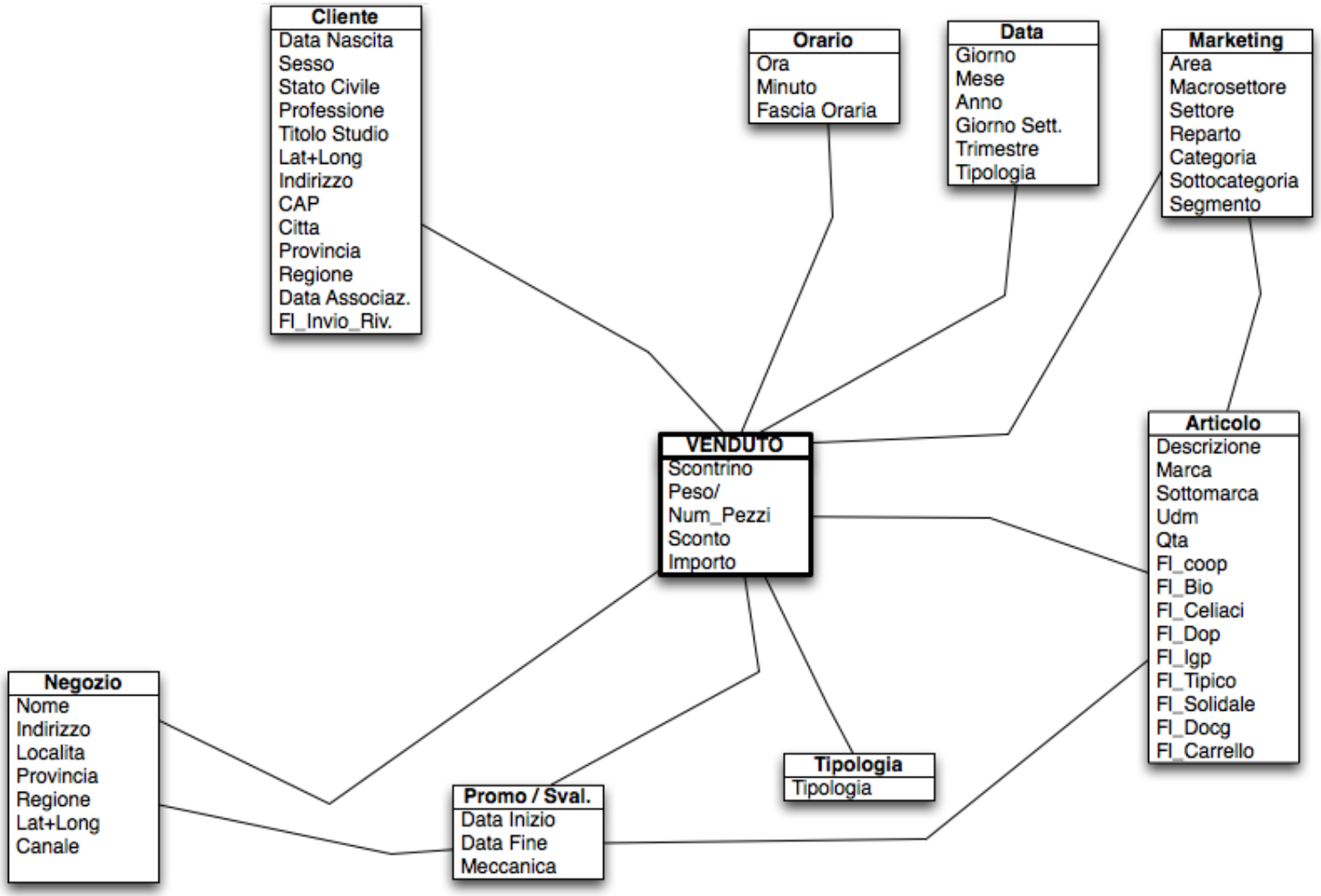
- CaseStudy1 - Churn Analysis (30)

Case studies from real data:

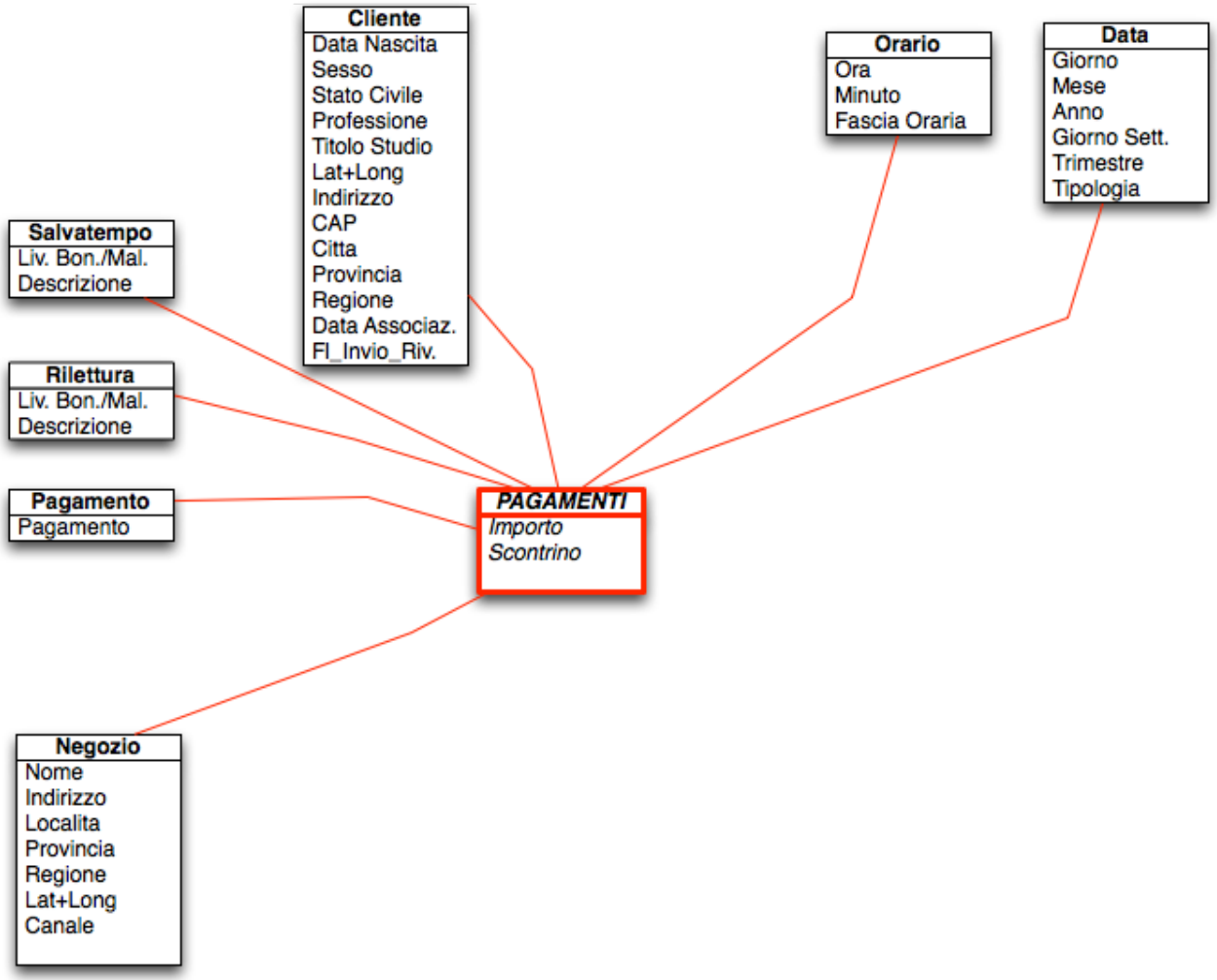
CaseStudy - Fraude detection

CaseStudy1 - back to Audit Planning(20)

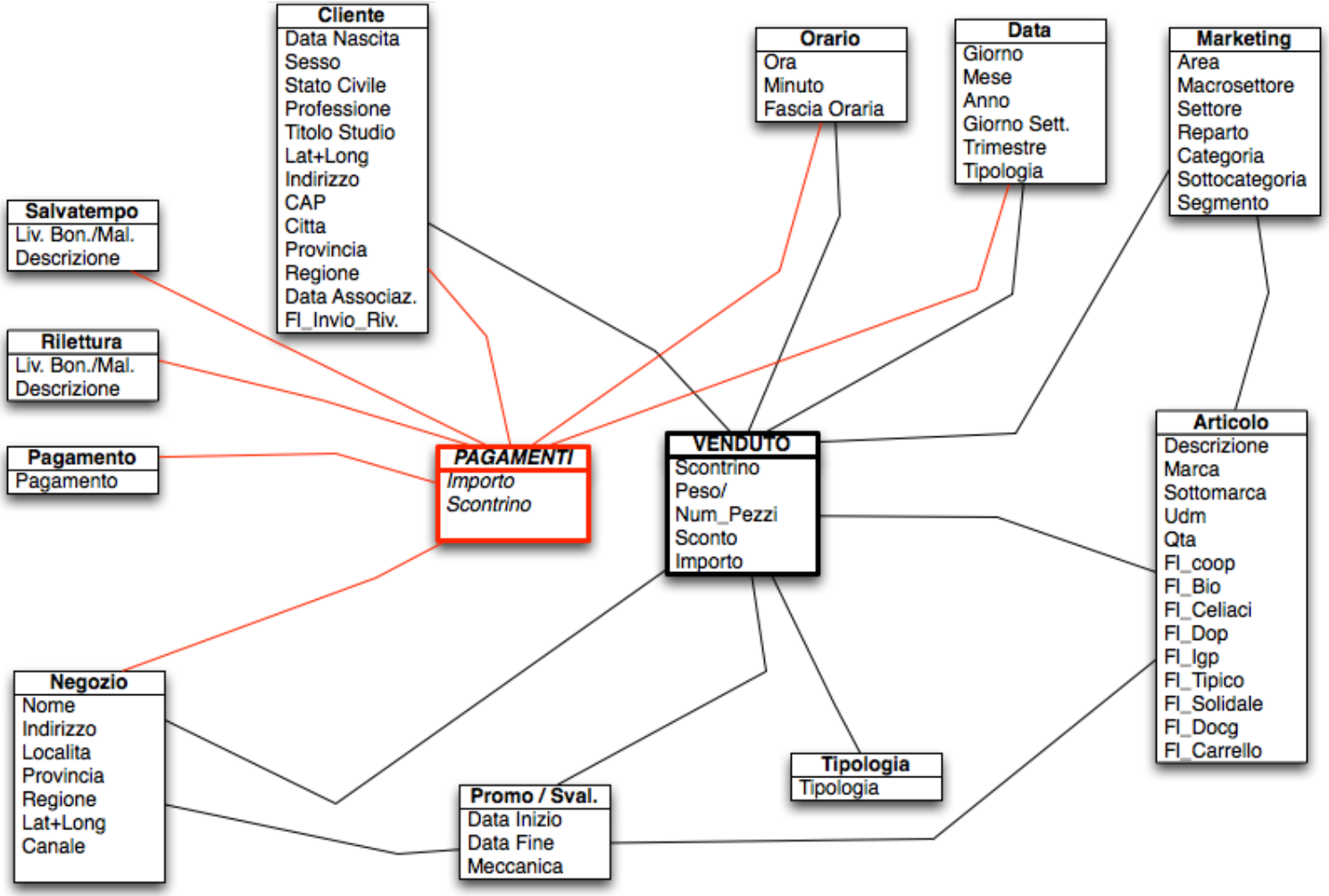
Il Data Warehouse (1)



Il Data Warehouse (2)

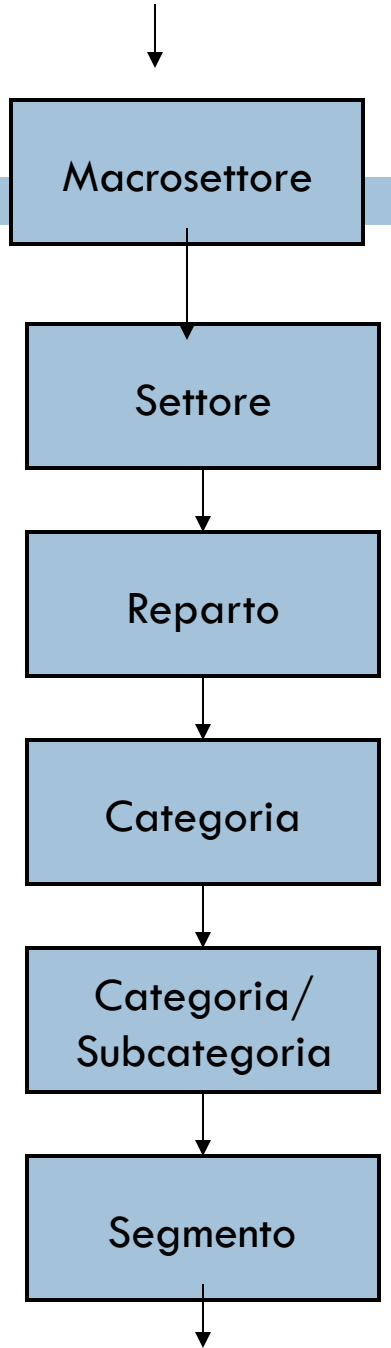


Il Data Warehouse (3)



La gerarchia Marketing

- **Area** – 3 valori
- **Macrosettore** – 4 valori
- **Settore** – 13 valori
- **Reparto** – 76 valori
- **Categoria** – 529 valori
- **Sottocategoria** – 2665 valori
- **Segmento** – 7656 valori
- **Articolo** – 571092 valori



La gerarchia Marketing - Esempi

I primi 3 livelli

Area	Macrosettore	Settore
Alimentari	Freschi	Freschi Freschissimi
	Generi vari	Grocery Alimentari
Altro	Altro	Confezionato per vendita Erogazione carburanti Iniziative speciali Non definito
Non alimentari	No Food	Casa Chimica Multimedia Persona Salute Stagionali e Brico

Esempio: coche decaffeinate

Area	Alimentari
Macrosettore	Generi vari
Settore	Grocery alimentari
Reparto	Liquidi
Categoria	Bibite
Sottocategoria	Cole
Segmento	Decaffeinata

COCA-COLA SENZA CAFFEINA LATTINA ML.330X6
COCA COLA SENZA CAFFEINA BOTTIGLIA PET ML.500
COCA COLA SENZA CAFFEINA LATTINA ML.330
COCA COLA SENZA CAFFEINA PET LT.1,5
COCA-COLA SENZA CAFFEINA LATTINA ML.330
COCA COLA SENZA CAFFEINA PET ML.500X4
COCA COLA SENZA CAFFEINA LATTINA ML.330 X6
COCA COLA SENZA CAFFEINA PET LT. 1,5 + ML. 250 OMAGGIO
BIBITA COCA COLA SENZA CAFFEINA CLUSTER LT.1,5 X 6
COLA SENZA CAFFEINA COOP NO COLOR. NUOVA RICETTA BOTTIGLIA PET LT
COLA SENZA CAFFEINA COOP PET. LT. 1,5
COLA SENZA CAFFEINA HAPPYHAND PET LT 1,5
COLA SENZA CAFFEINA PERLA PET 1,5 L
BIBITA PEPSI BOOM JUNIOR PET ML.330X4
BIBITA PEPSI BOOM PET LT.1,5
PEPSI BOOM PET LT.2
BIBITA PEPSI BOOM LATTINA ML. 330

Churn analysis

Introduction

Challenge: FAST identification

- Churn – defezione - abbandono
- In alcuni casi, nel momento in cui la defezione si manifesta, è troppo tardi per intervenire
  - non è più possibile recuperare il defezionante, o
  - non è più conveniente recuperarlo
- Fondamentale identificare il defezionante *immediatamente*, o addirittura *in anticipo*
- Nuova formulazione del problema:
  - Churn Analysis = *Previsione dell'abbandono*

Interruzione del rapporto: modalità

- Interruzione esplicita
  - Tipica dei rapporti che richiedono un contratto o impegni da parte del fruitore
    - Es.: Telefonia, nei casi in cui è previsto un canone
    - Es.: Tesseramenti rinnovabili non gratuiti
- Interruzione implicita
  - Tipica dei rapporti non formalizzati o privi di costi per il fruitore
    - Es.: Tessere sconto e carte fedeltà

Interruzione implicita

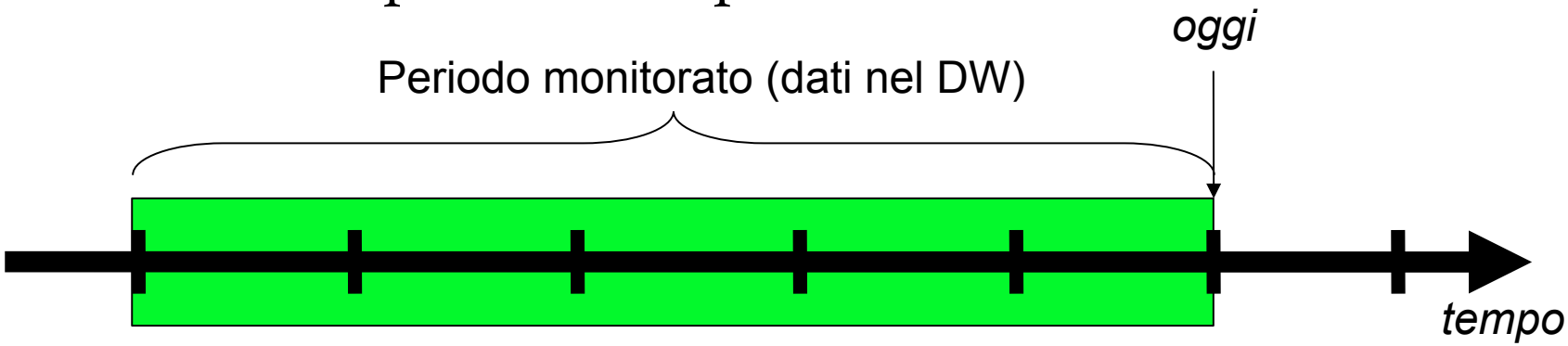
- E' la situazione più comune nel settore delle vendite al dettaglio
  - Carte fedeltà che non richiedono rinnovi né costi
  - Il defezionante semplicemente non la usa più
- Domanda: è sempre facile capire quando il cliente/fruitore ha abbandonato?
  - Non fa acquisti per un mese?
  - Non fa acquisti per un anno?
  - Visita il punto vendita meno di 2 volte al mese?
  - Spende meno del 50% di quanto faceva 3 mesi fa?

Abbandono “soft”

- Nozione alternativa di abbandono:
  - Passaggio da un tipo di rapporto ad uno diverso
  - Caso estremo: da “fedele” a “abbandono totale”
- Situazione naturale nella vendita al dettaglio
  - Il segmento “fedele” fornisce (parziali) garanzie su un indotto minimo dell'attività
  - Il degrado del cliente “fedele” a cliente “saltuario” ha effettivi negativi sulla gestione aziendale
    - Valgono le stesse motivazioni dell'abbandono “hard”

Previsione dell'abbandono

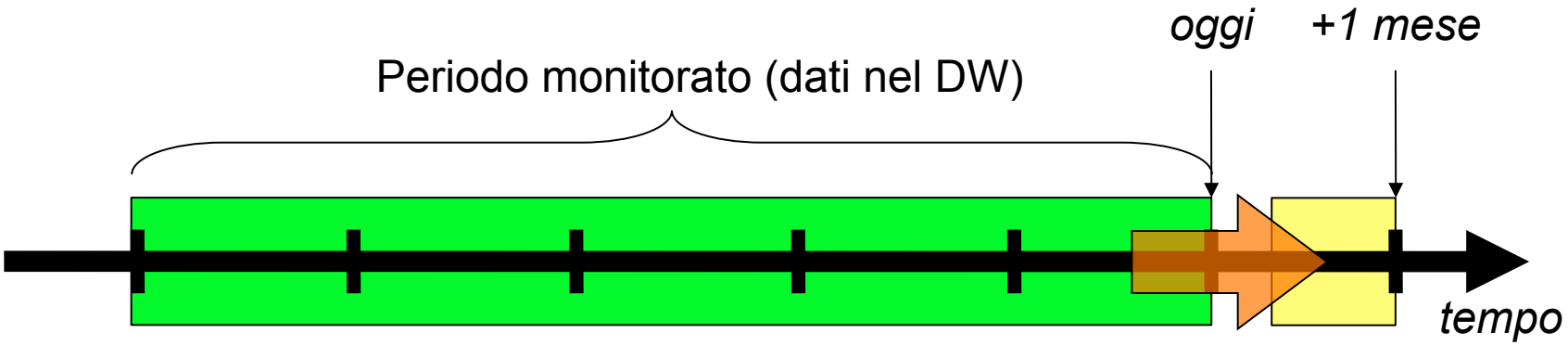
- Il tracciamento del cliente ci consente di ricostruire la sua "storia" per un certo periodo



- La nozione di abbandono adottata sarà valutata su un segmento recente di tale storia

Previsione dell'abbandono

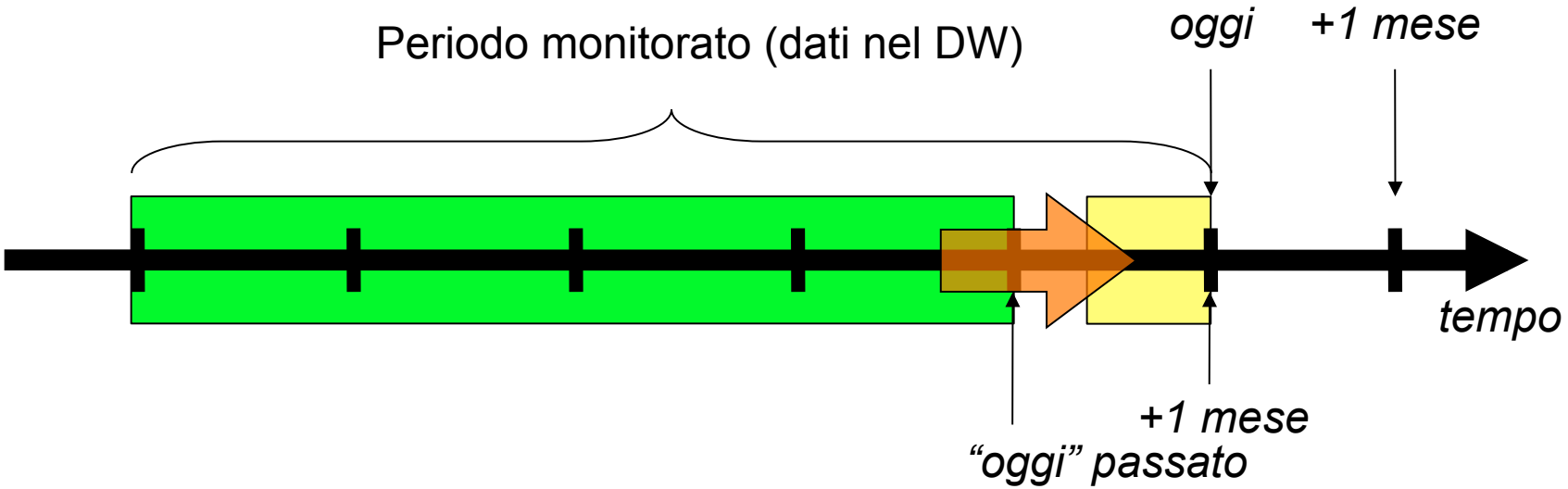
- Obiettivo: previsione dello stato di abbandono *futuro,* conoscendo la storia recente del cliente



  - La storia recente fornisce indizi sul comportamento che il cliente si presta a tenere
    - alcuni indizi permettono di discriminare i futuri defezionanti, altri no
    - alcuni indizi sono espliciti nei dati a disposizione, altri vanno derivati da essi

Previsione dell'abbandono

- Come determinare *oggi* le correlazioni tra situazione attuale e stato futuro?
  - Cerchiamo queste correlazioni nel *passato*
  - Le relazioni "passato → oggi" verranno sfruttate per predire il futuro dall'oggi



Schema dell'applicazione

- Definizione/estrazione delle variabili di lavoro
  - Variabili predittive: gli *indizi* disponibili *oggi/passato*
  - Variabile target: lo stato di abbandono *futuro/oggi*
- Estrazione del modello predittivo
  - Ricerca di correlazioni tra variabili predittive e variabile target, da sfruttare in fase di predizione
- Applicazione del modello predittivo
  - Le relazioni variabile predittiva → target vengono applicate alla situazione odierna (in termini di variabili predittive) per stimare la var. target

BICOOP – Churn Analisys

Definizione del concetto di abbandono e creazione di modelli previsionali

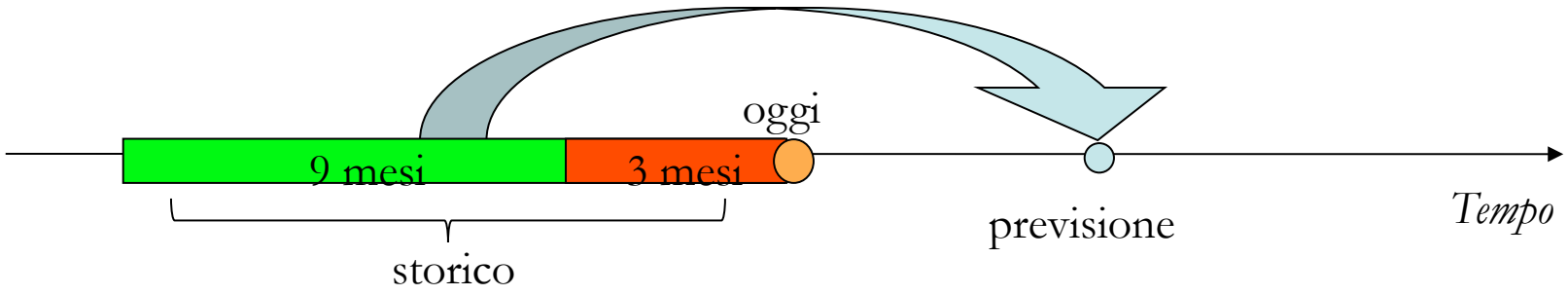
Problem Definition:

- Estimate the probability of churn on the base of DW evidences:
  - Detailed buying records
  - Demographic data
- Churn risk definition
  - For a client the churn risk appears when a dramatic decrease of her/his expenditure measures:
    - Number of visits
    - Total amount of expenditure value
    - Number of items bought

Analisi previsionale

- Raccolta dei dati storici per l'estrazione di:
  - Variabili di vendita e anagrafiche, i predittori (periodo verde)
  - Variabili obiettivo (periodo rosso)

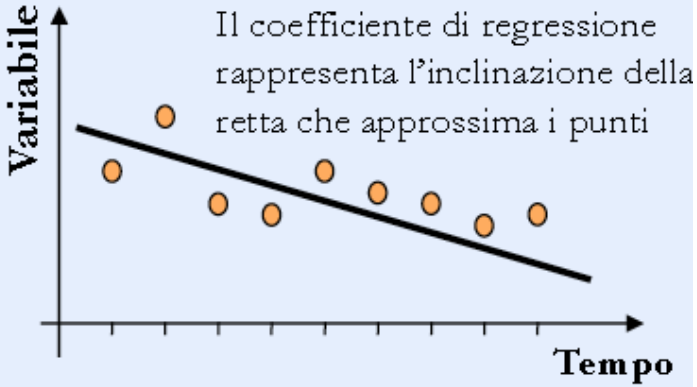
- Costruzione di un modello predittivo
  - Addestrato in modo opportuno su dati storici
  - Utilizzabile per ottenere informazioni previsionali



Data preparation

Si sono estratte dal data warehouse, per il periodo di 9 mesi (Dicembre 2006 – Agosto 2007) le seguenti informazioni:

- Dati anagrafici (sesso, età, professione etc.)
- Dati di spesa
  - Globale
  - Settori specifici: fresco, carni, pesce, ortofrutta
  - Pesata (abbattimento no-food)
- Trend di spesa:
  - Tipologia cliente (per ogni mese)
  - Regressione spese
  - Regressione spesa
  - Regressione battute



Preparazione dati – target (periodo rosso)

- Si sono estratte dal data warehouse, per il periodo di 3 mesi (Settembre 2007 – Novembre 2007) le seguenti informazioni:
  - Numero di spese
  - Variazione di spesa rispetto al periodo verde
    - Volume di spesa
    - Battute di cassa
    - Numero di visite

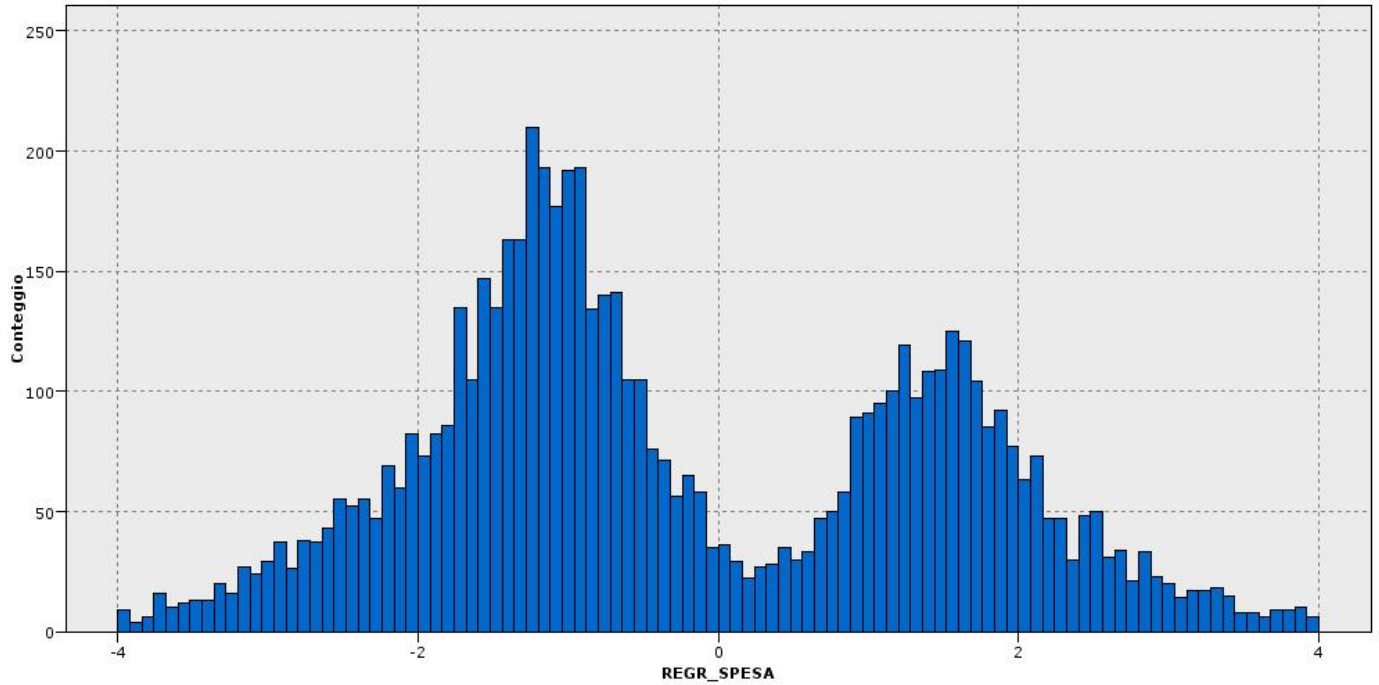
Dataset

- Il dataset così ottenuto presenta una riga per ogni cliente che ha effettuato almeno una spesa nei nove mesi di osservazione. In tutto abbiamo ottenuto:
  - 517.000 righe
  - 47 attibuti

Predittori Anagrafici	Predittori di spesa	Predittori di trend	Variabili target
CLIENTE_ID	DATA_ULTIMA_SPESA	TIPOLOGIA_01	T_NUM_SPESE
SESSO	NUM_SPESE	TIPOLOGIA_02	T_RAPP_SPESE
STATO_CIVILE	SPESA_TOT	TIPOLOGIA_03	T_RAPP_SPESA
PROFESSIONE	SPESA_TOT_PESATA	TIPOLOGIA_04	T_RAPP_BATTUTE
TITOLO_STUDIO	SPESA_MEDIA	TIPOLOGIA_05	
PROVINCIA	SPESA_MEDIA_PESATA	TIPOLOGIA_06	
REGIONE	BATTUTE	TIPOLOGIA_07	
ANNO_SOCIO	FRESCHI_TOT	TIPOLOGIA_08	
FASCIA_ANNO_SOCIO	FRESCHI_SPESE	TIPOLOGIA_09	
FL_INVIO_RIVISTA	CARNI_TOT	TIPOLOGIA_MEDIA	
COD_NEGOZIO	CARNI_SPESE	TIPOLOGIA_ZERI	
ETA	PESCE_TOT	REGR_NUM_SPESE	
ETA_FASCIA	PESCE_SPESE	REGR_SPESA	
	ORTOFRUTTA_TOT	REGR_SPESA_PESATA	
	ORTOFRUTTA_SPESE	REGR_BATTUTE	

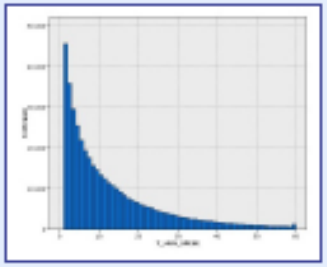
Data Exploration

- Distribuzione trend di spesa

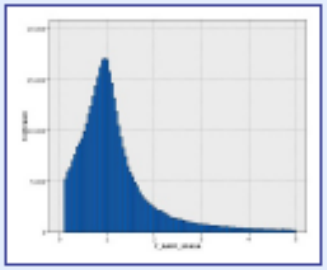


Trend dei clienti con spesa totale > 400€

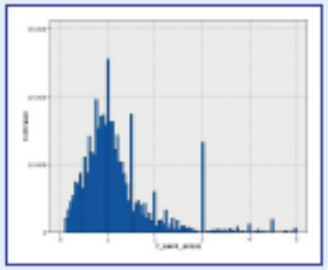
Funzioni Obiettivo



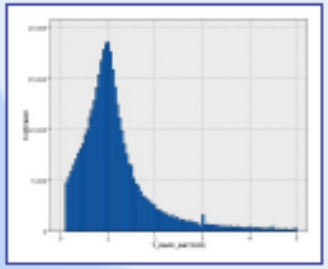
NUM_SPESE: spese del cliente nel periodo target



RAPP_SPESE: rapporto tra il numero delle spese del periodo target e quello del periodo d'osservazione



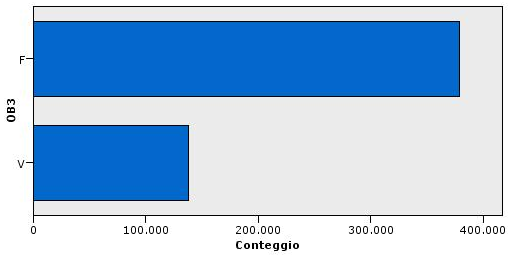
RAPP_SPESA: rapporto tra la spesa del periodo target e quella del periodo d'osservazione



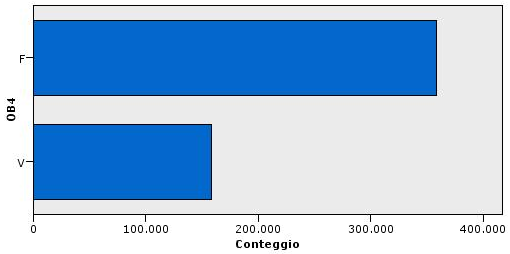
RAPP_BATTUTE: rapporto fra le battute di cassa del periodo target e quelle del periodo d'osservazione

F. Obiettivo – Thresholds

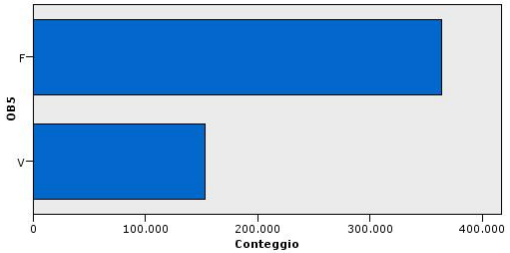
- Scelta una soglia di allarme per indicare un possibile cliente defezionario i rapporti si trasformano in tre indicatori di abbandono
- Abbiamo scelto come soglia una diminuzione sulle 3 misure del 50%
- Otteniamo le seguenti distribuzioni: F (Basso rischio , V (Alto rischio Abbandoni)



OB1: RAPP_SPESE



OB2: RAPP_SPESA

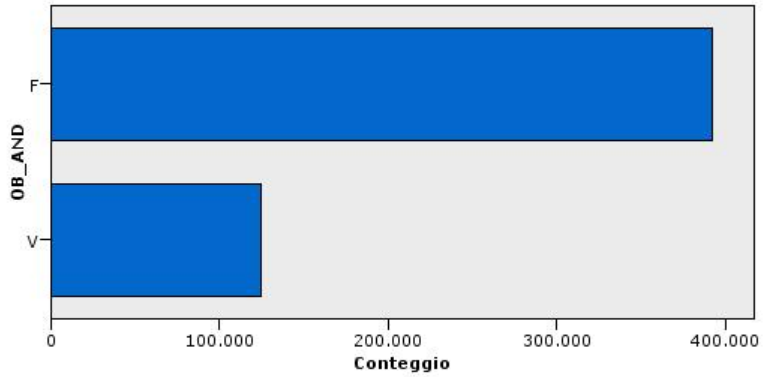


OB3: RAPP_BATTUTE

F. Obiettivo – Sintesi

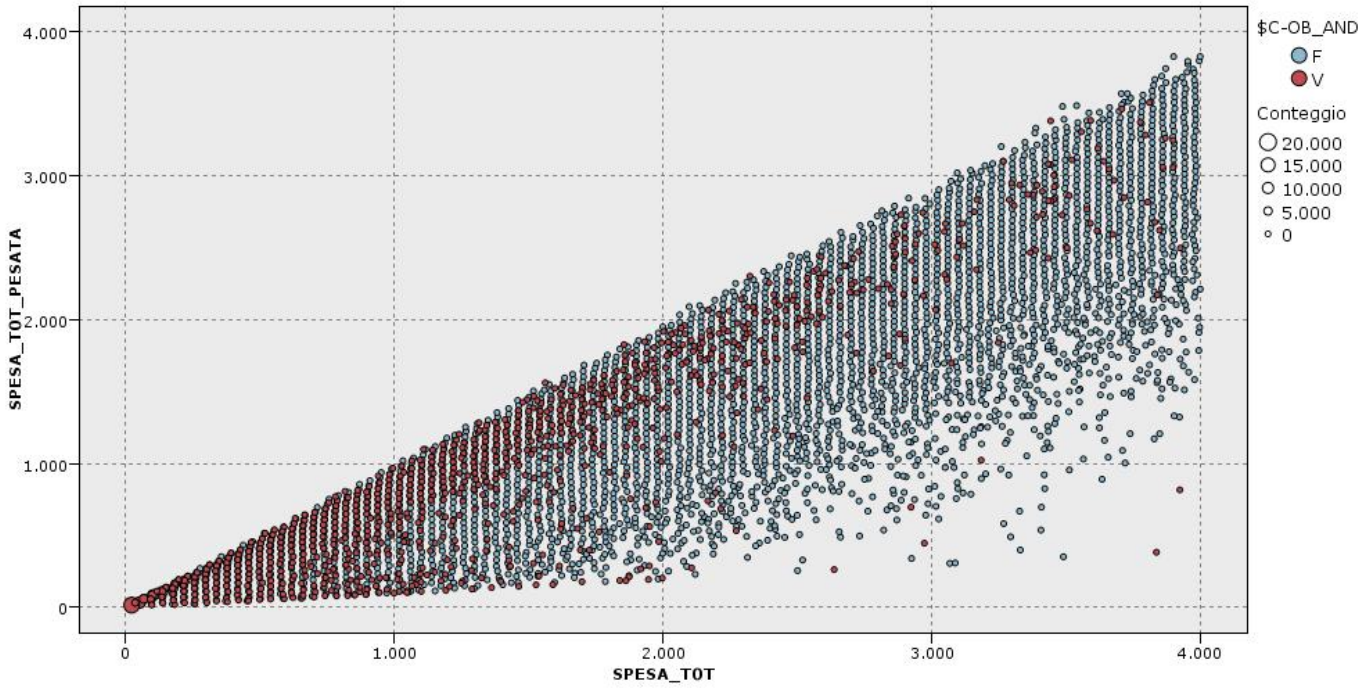
Per la funzione obiettivo finale si è deciso di considerare come potenziali defezionari tutti i clienti che superato la soglia di allarme, in ognuno dei tre indicatori OB1, OB2, OB3:

OB_AND: OB1 and OB2 and OB3



Modello previsionale e Risultati

- Distribuzione spesa totale vs. spesa pesata



Modello previsionale e Risultati

- Esempio di regole associative:

se REGIONE = TOSCANA
e NUM_SPESE <= 128
e TIPOLOGIA_01 = 7
e TIPOLOGIA_09 = 0
e TIPOLOGIA_ZERI > 2
e REGR_BATTUTE <= -0,98
allora V (confidenza 82,8%)

se DATA_ULTIMA_SPESA > 183
e NUM_SPESE <= 21
e TIPOLOGIA_ZERI > 1
e REGR_NUM_SPESE <= -0,02
e REGR_BATTUTE <= -0,98
allora V (confidenza 92%)

Modello previsionale
Risultati Globali

- Correttezza generale del modello:
  - 81.06% sul training set (70% del dataset, 360.000 righe)
  - 80.94% sul test set (30% del dataset, 155.000 righe)
- Matrici di confusione:

Valori Predetti

Valori Reali

Training Set	F	V
F	256.608	17.920
V	50.540	36.466

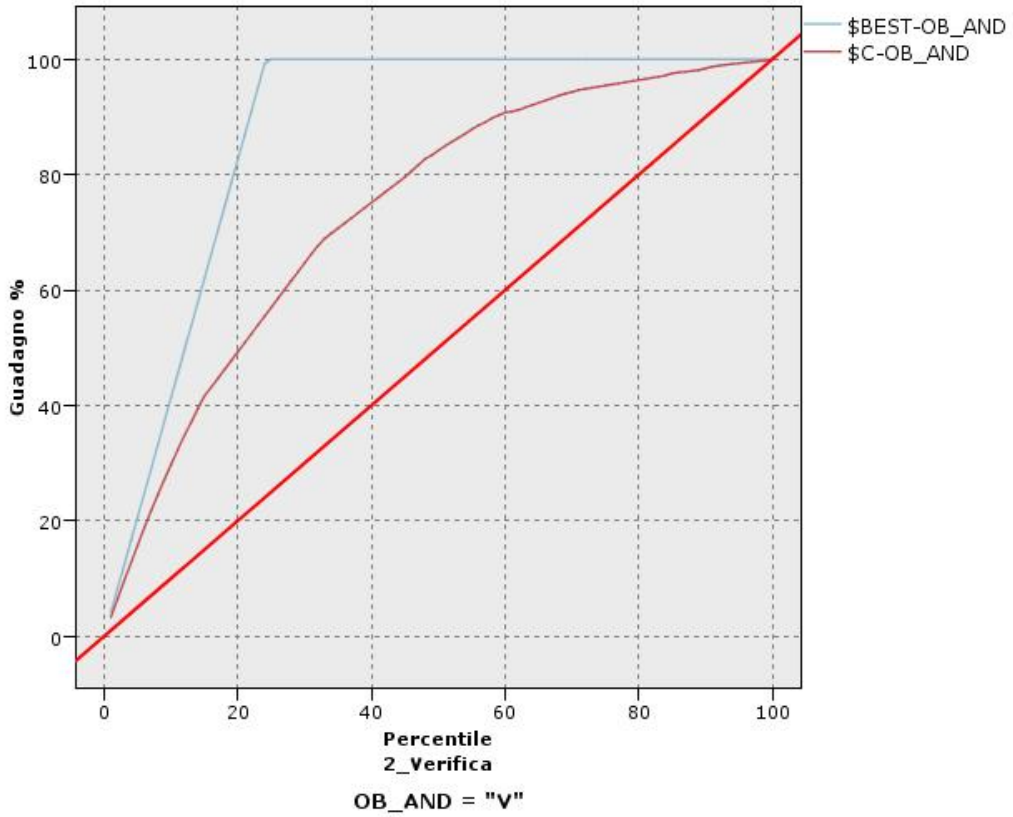
Test Set	F	V
F	110.029	7.767
V	21.855	15.734

66.9%

Con un guadagno netto del **42.8%**

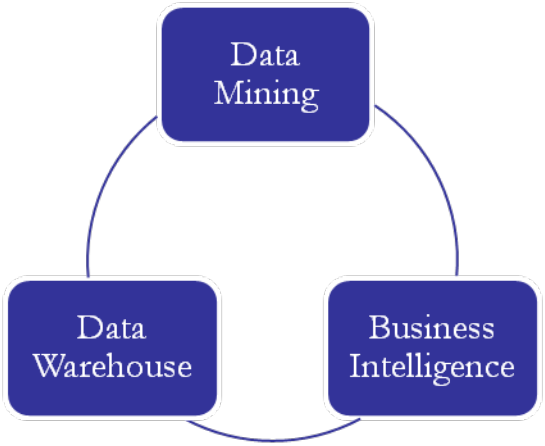
Modello previsionale
Risultati Globali

- Lift chart



Scenario d'uso – Esempio

- Creare un ambiente aperto e dinamico nel quale i dati forniti dal data warehouse vengono elaborati e trasformati in modelli di tipo previsionale.



- I modelli previsionali possono essere usati per arricchire il data warehouse, innestando un circolo virtuoso di informazioni utilizzabili anche direttamente in ambienti di Business Intelligence.

Conclusioni

- Per concludere:
  - Sono stati utilizzati dati provenienti dal data warehouse, risparmiando tempo e ottenendo dati di buona qualità
  - Abbiamo usato tecniche di mining avanzate per generare modelli predittivi, principalmente regole associative e alberi di decisione.
  - I risultati ottenuti sono soddisfacenti e si intravedono buone prospettive di miglioramento
- Possibili sviluppi futuri
  - Sperimentazione di altri tipi di analisi: sub group analysis, market segmentation, clustering ect.
  - Consolidamento e validazione dei risultati ottenuti
  - Incrementare la collaborazione con gli esperti del dominio per una migliore taratura del problema, delle definizioni usate e delle funzioni obiettivo
  - Integrazione dei dati previsionali forniti dai modelli predittivi all'interno della struttura di business intelligence aziendale

Analisi della defezione nella grande distribuzione tramite indicatori geografici e socio demografici/1

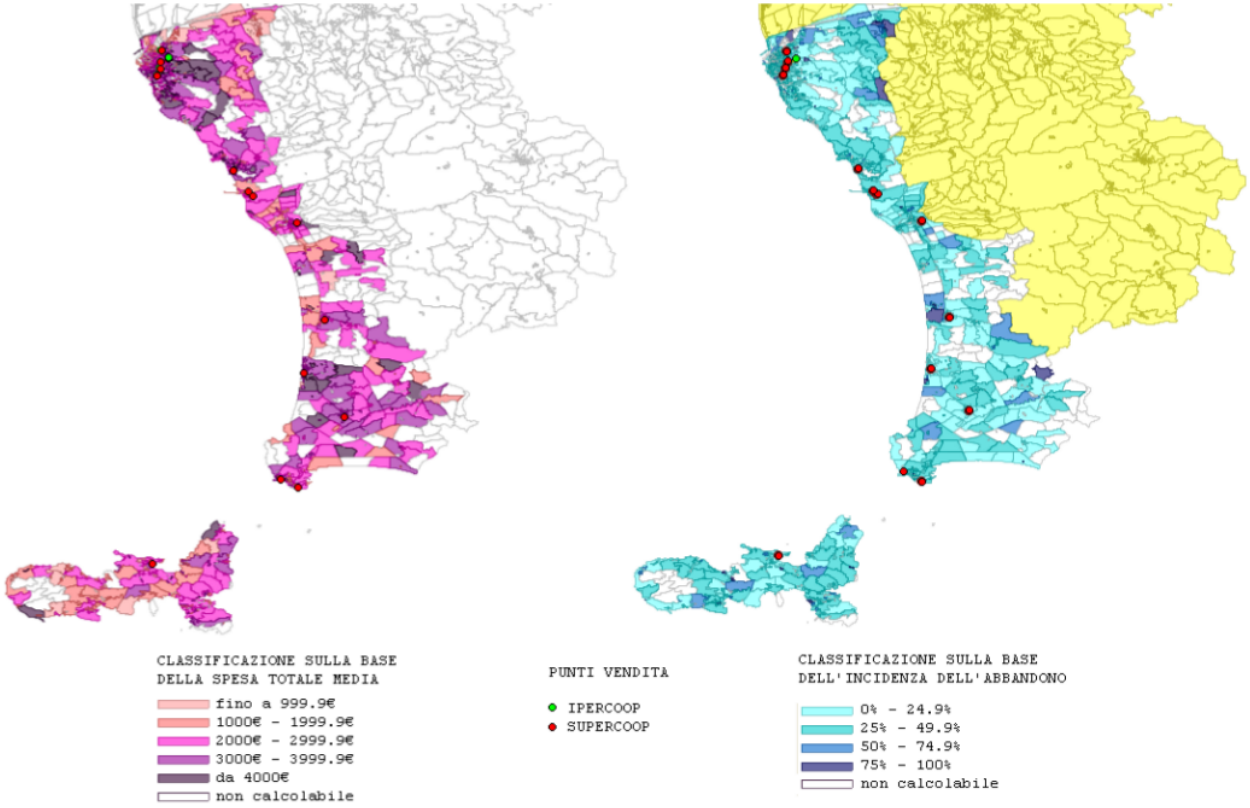


Figura 5.3: Provincia di Livorno con colorazione tematica in base all'incidenza del fenomeno dell'abbandono e alla media degli acquisti totali.

Analisi della defezione nella grande distribuzione tramite indicatori geografici e socio demografici/2

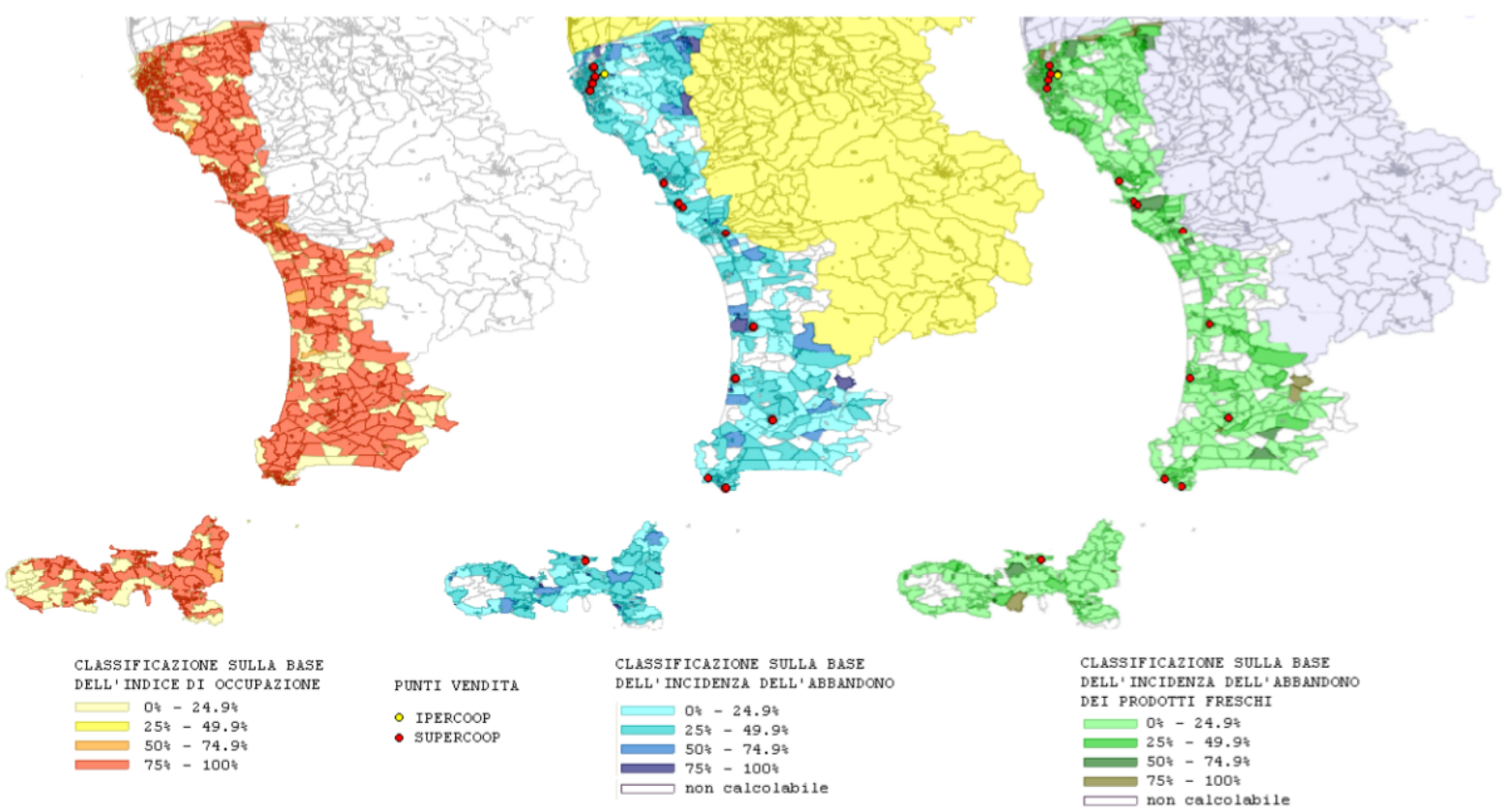


Figura 5.14: Provincia di Livorno con colorazione tematica in base al livello di occupazione e all'incidenza dell'abbandono.

Analisi Spazio-Temporale della defezione nella grande distribuzione / Eventi individuali e fenomeni collettivi nella grande distribuzione. Analisi spazio-temporale dei dati di vendita

- Varie tipologie di eventi individuali
  - Abbandono
  - Fidelizzazione ad un prodotto
  - Ecc.
- Analisi spaziale e temporale alla ricerca di grandi gruppi di eventi co-localizzati e contemporanei
- Metodi:
  - Clustering density-based
  - SatScan (metodo statistico usato in epidemiologia)

Analisi Spazio-Temporale della defezione nella grande distribuzione / Eventi individuali e fenomeni collettivi nella grande distribuzione. Analisi spazio-temporale dei dati di vendita

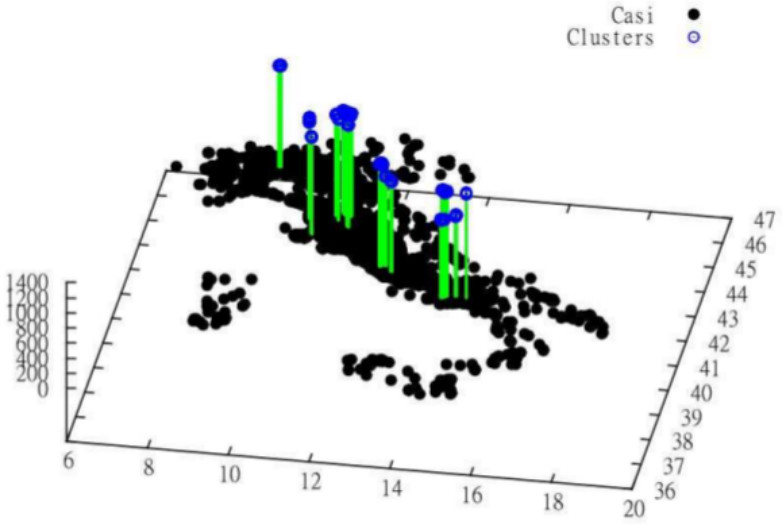


Fig.42 – Apertura di un nuovo centro commerciale a Monteforte Irpino.

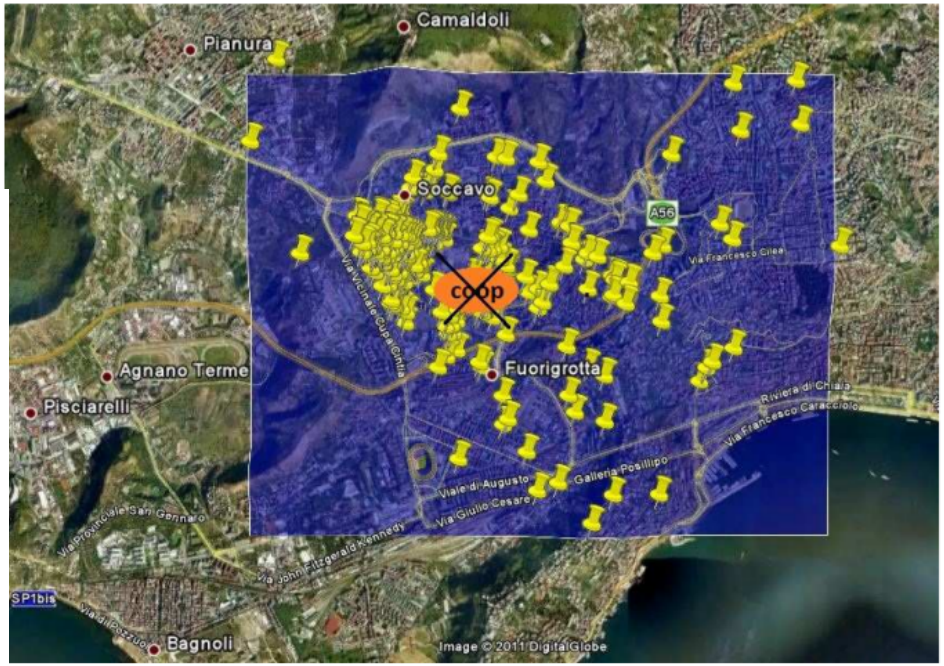


Fig.38 – Chiusura del supermercato Unicoop di Soccavo.

INTELLIGENT AUDIT PLANNING

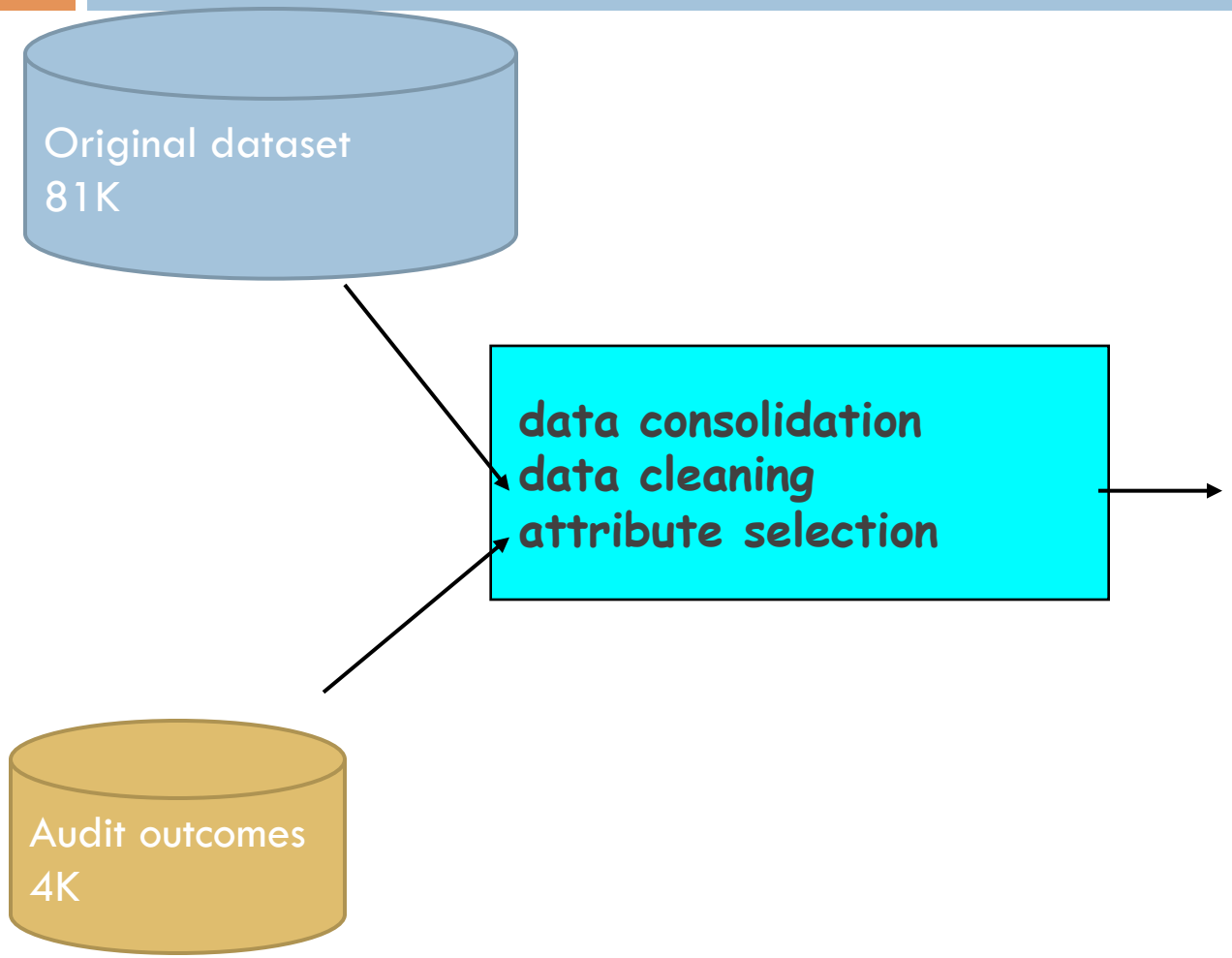
Sorgente: Ministero delle Finanze

Progetto Sogei, KDD Lab. Pisa

Audit planning

- Need to face a trade-off between several conflicting optimization issues:
  - Maximize audit benefits
    - True positive rate
    - Profitability
      - Amount recovered
      - Assessed/Declared
      - …
  - Minimizing audit costs
    - False positive rate
    - No. of filed cases
  - Continuous monitoring, gaining control of socio-economic realities

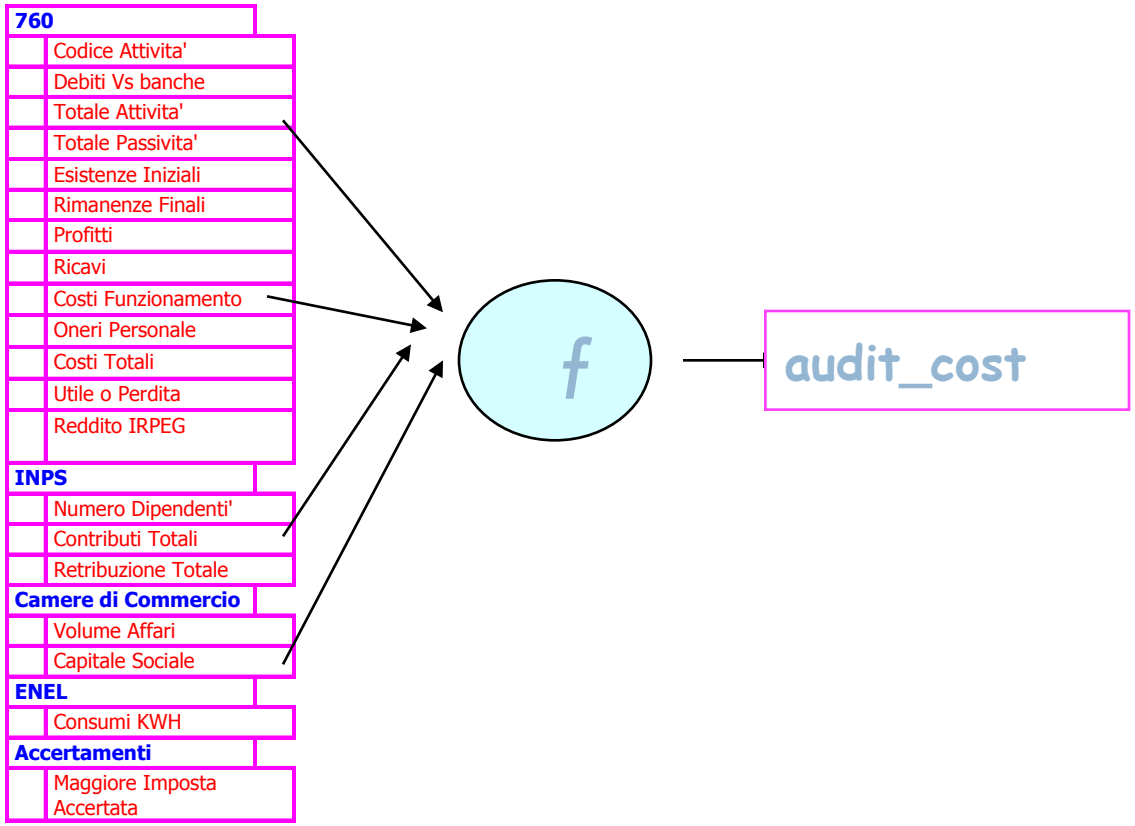
Data preparation



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 of our analysis is set as the Class of Actual Recovery (c.a.r.):

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

Classification Methodologies

- Decision trees
  - Simple, widely accepted models, with high explanatory capability
- Advanced mechanisms used:
  - pruning factor
  - misclassification weights
  - boosting factor

Quality assessment indicators

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

Evaluation measures (1): Domain – independent measures

		PREDICTED VALUE	
		Positive	Negative
ACTUAL VALUE	Positive	TP	FN
	Negative	FP	TN

- TP (TN) True Positive (Negative) instances
- FP (FN) False Positive (Negative) instances

Domain-independent measures

- Accuracy = (TP+TN)/(TP+TN+FP+FN)
  - Misclassification Rate:(FP+FN)/(TP+TN+FP+FN)
- audit # (of a given classifier)
  - number of tuples classified as positive = FP + TP

		PREDICTED VALUE	
		Positive	Negative
ACTUAL VALUE	Positive	TP	FN
	Negative	FP	TN

Evaluation measures (2): Domain – Dependent measures

- 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

- Predictive c.a.r
- Classifiers compared with the REAL case, consisting of the whole test-set:

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

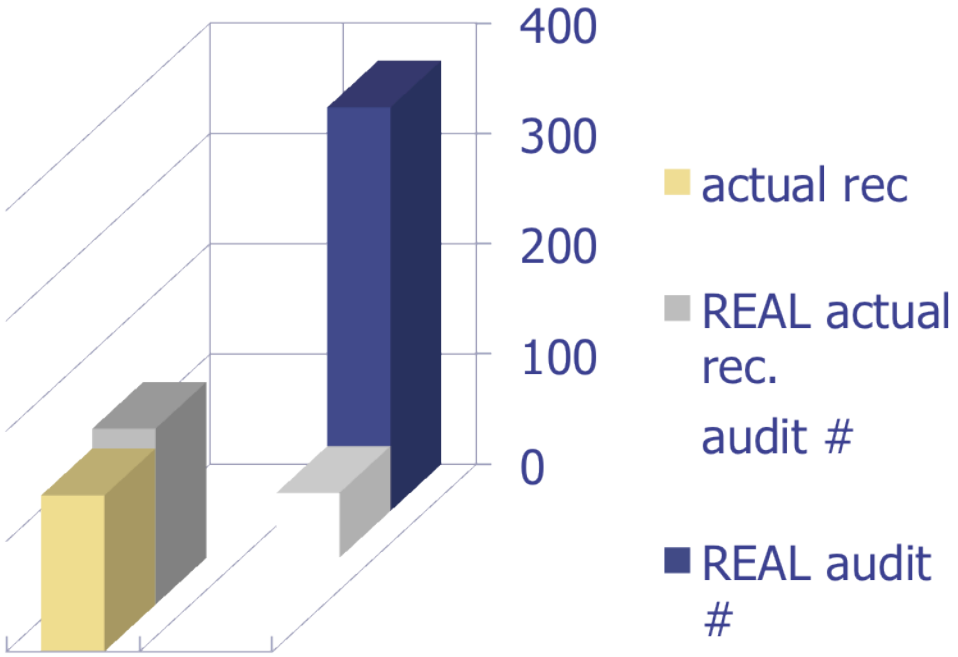
Controlling classifier construction

- *maximize audit benefits*: minimize FN
- *minimize audit costs*: minimize FP
- hard to get both!
  - unbalance tree construction towards either negatives or positives
- which parameters may be tuned?
  - misclassification weights, e.g., trade 1 FN for 10 FP
  - replication of minority class
  - boosting and pruning level

Model evaluation: classifier 1 (min FP)

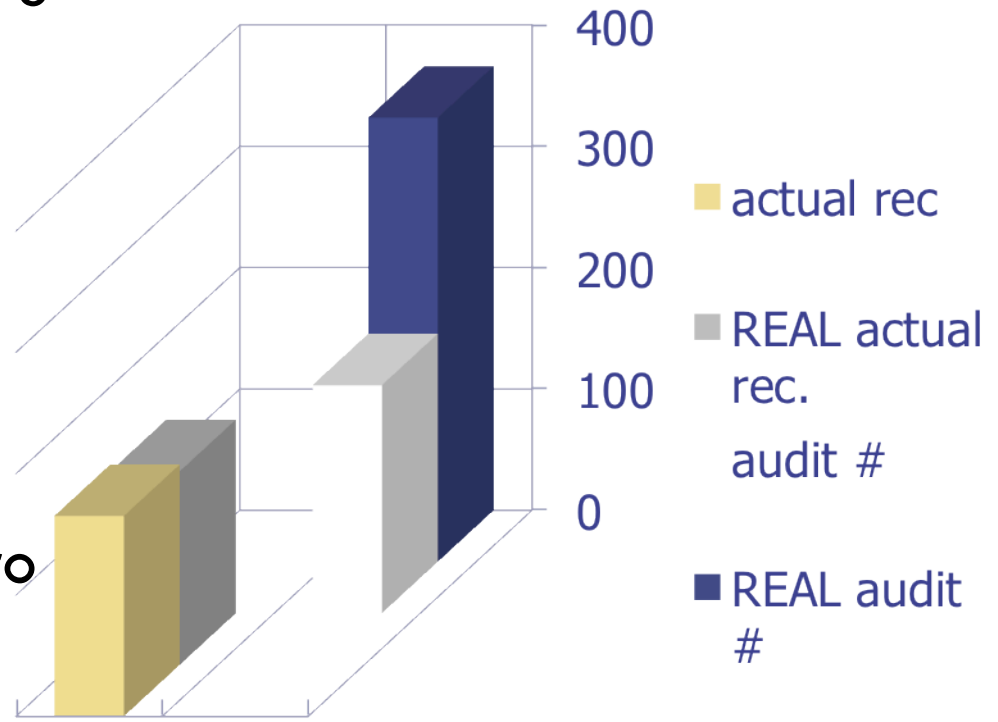
- no replication in training-set (unbalance towards negative)
- 10-trees adaptive boosting

- *misc. rate* = 22%
- *audit #* = 59 (11 FP)
- *actual rec.*= 141,7 Meuro
- *profitability* = 2,401



Model evaluation: classifier 2 (min FN)

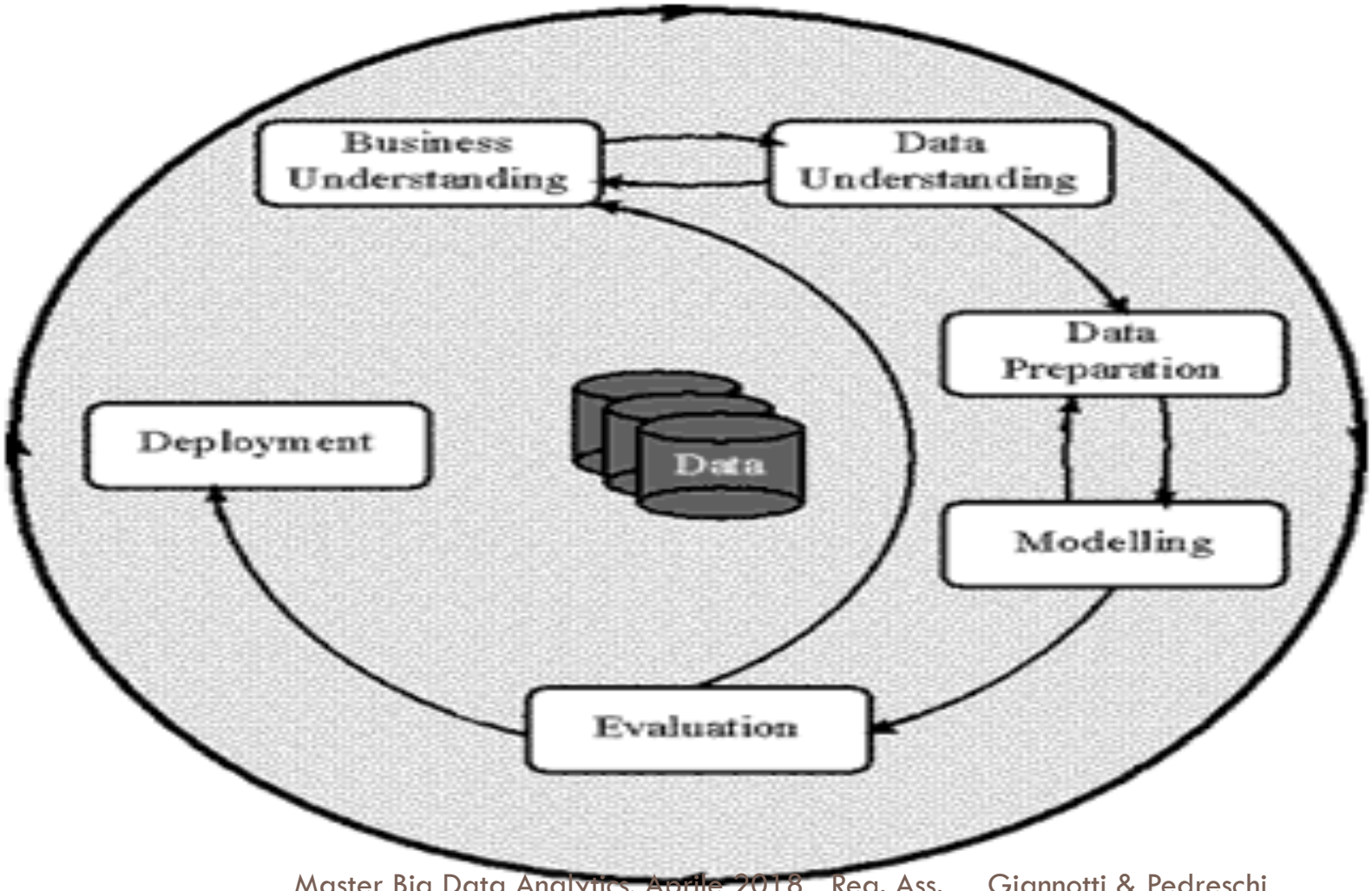
- replication in training-set (balanced neg/pos)
- misc. weights (trade 3 FP for 1 FN)
- 3-trees adaptive boosting



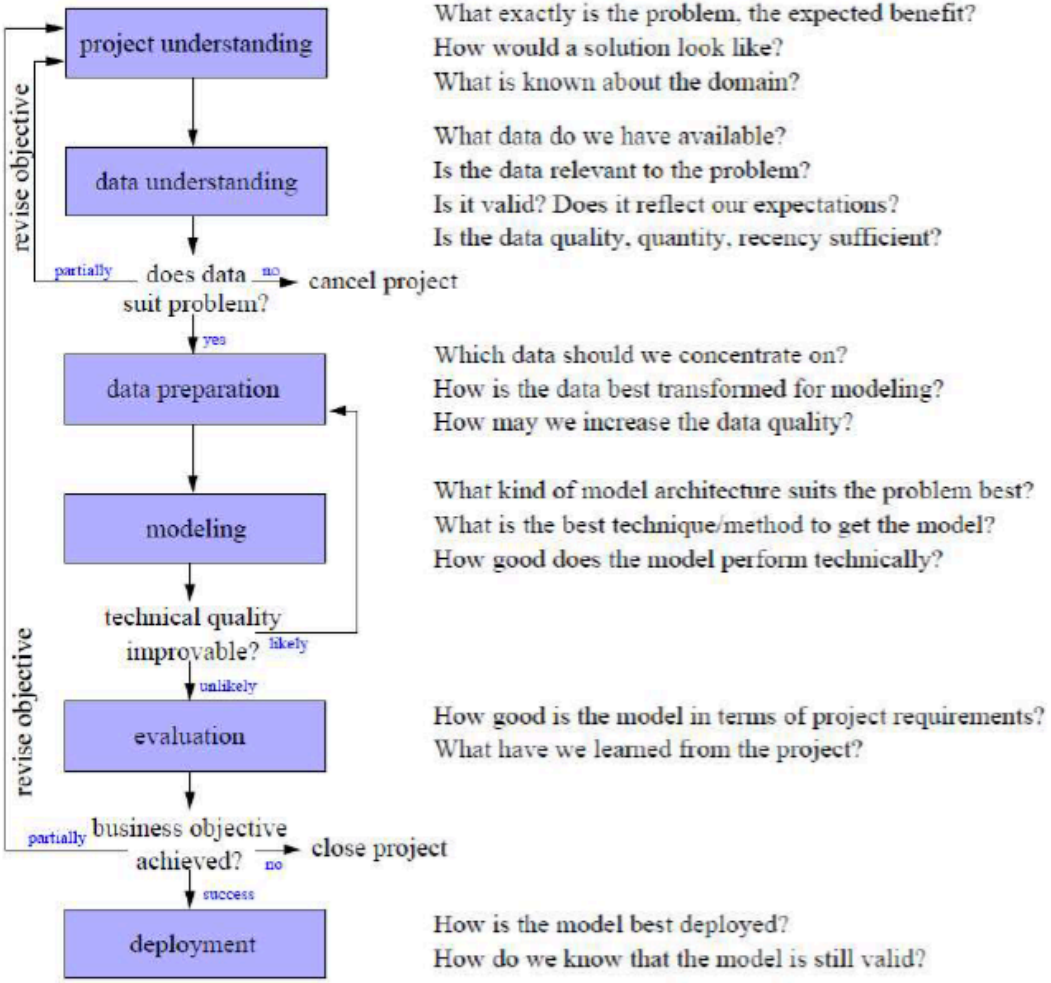
- *misc. rate* = 34%
- *audit #* = 188 (98 FP)
- *actual rec.*= 165.2 Meuro
- *profitability* = 0.878

CRISP-DM: THE LIFE CICLE OF A DATA MINING PROJECT

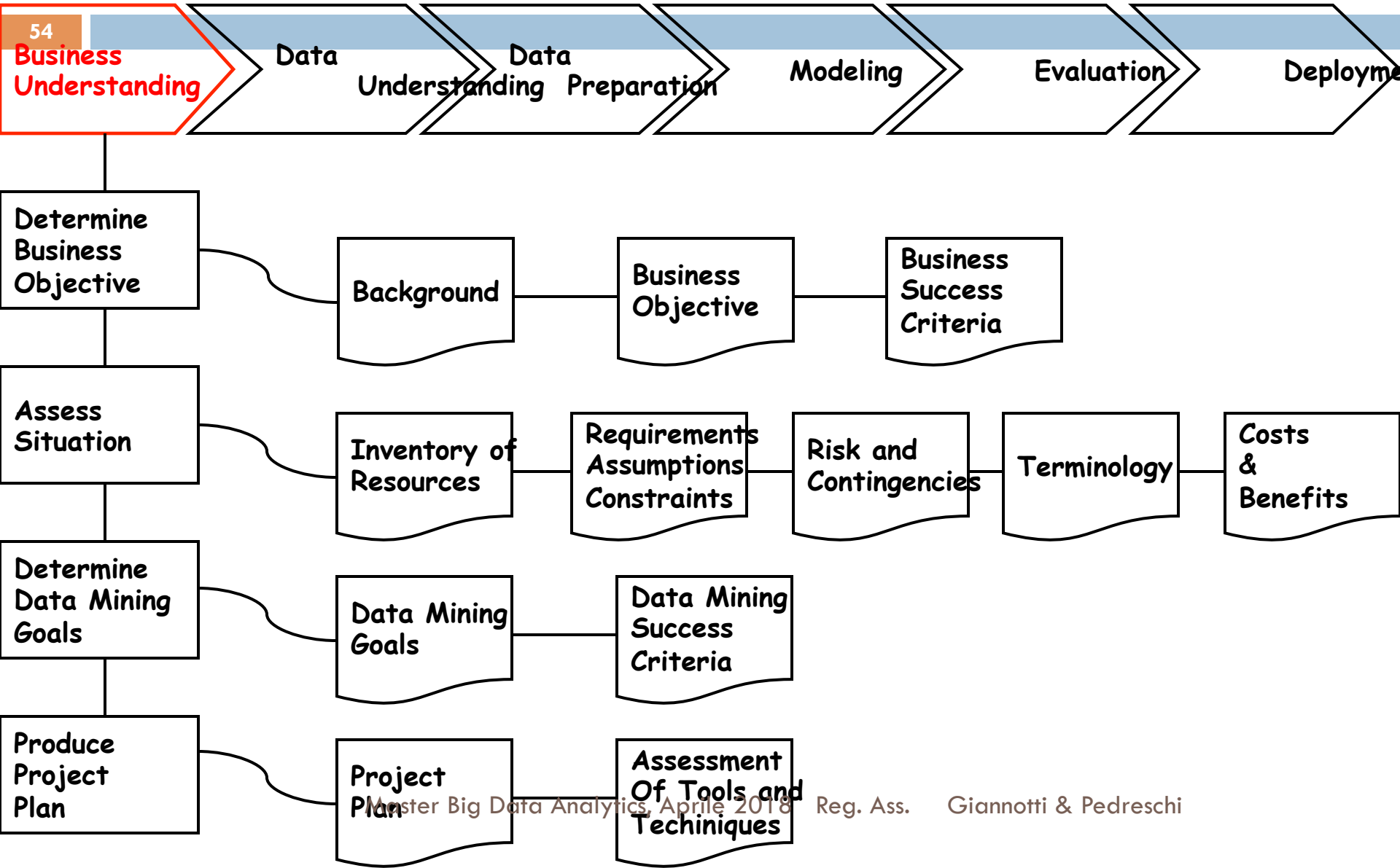
CRISP Methodology

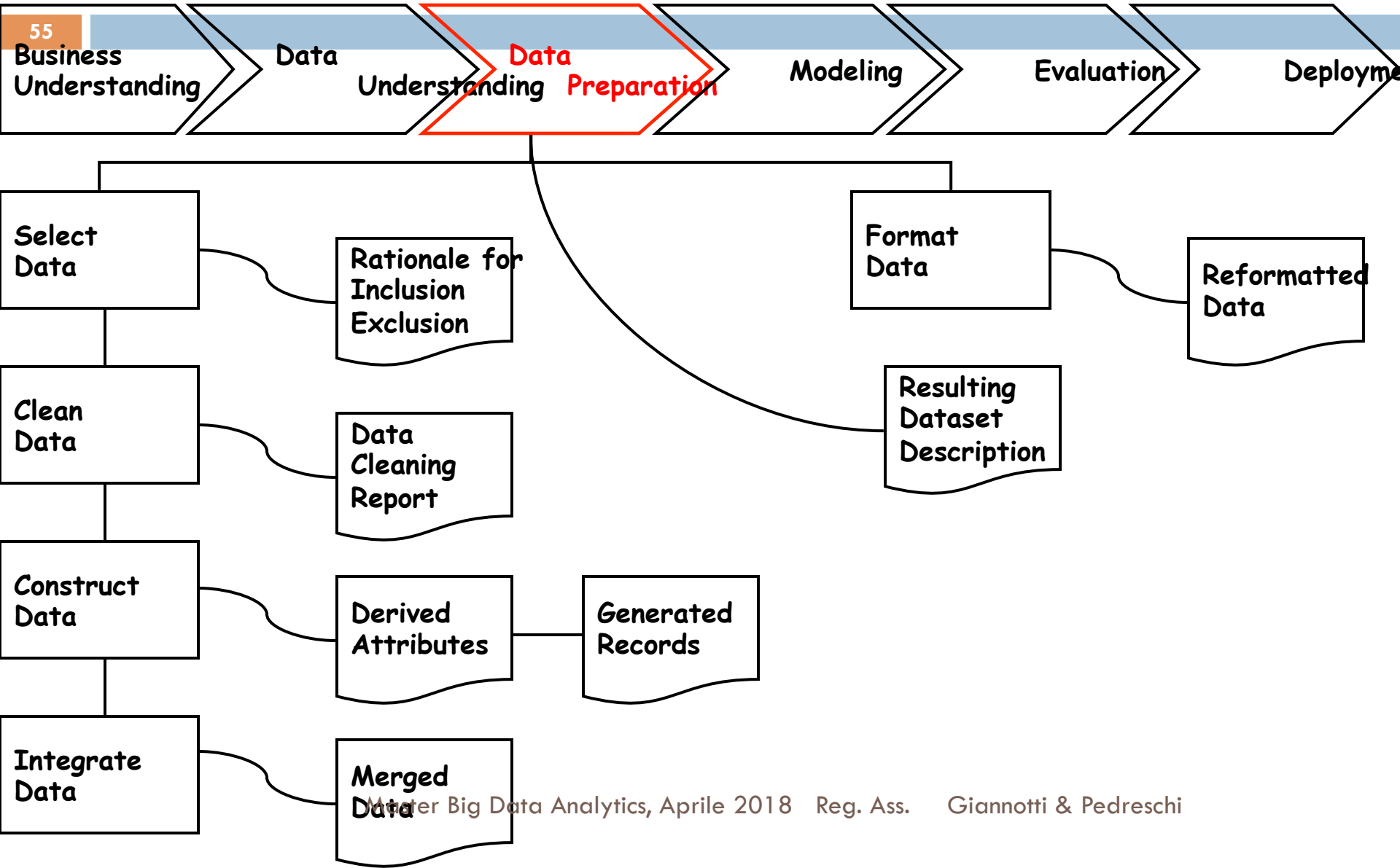


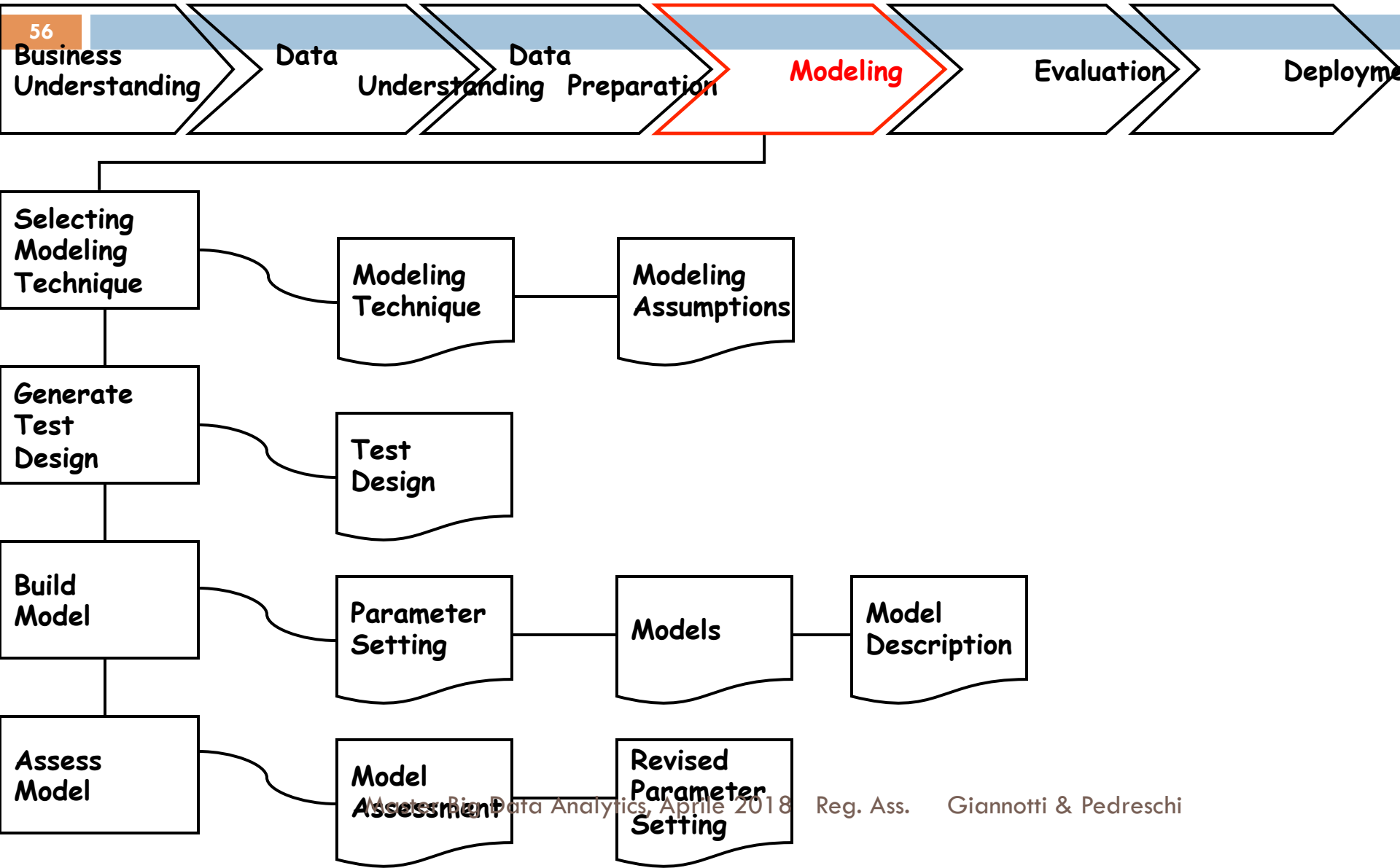
CRISP

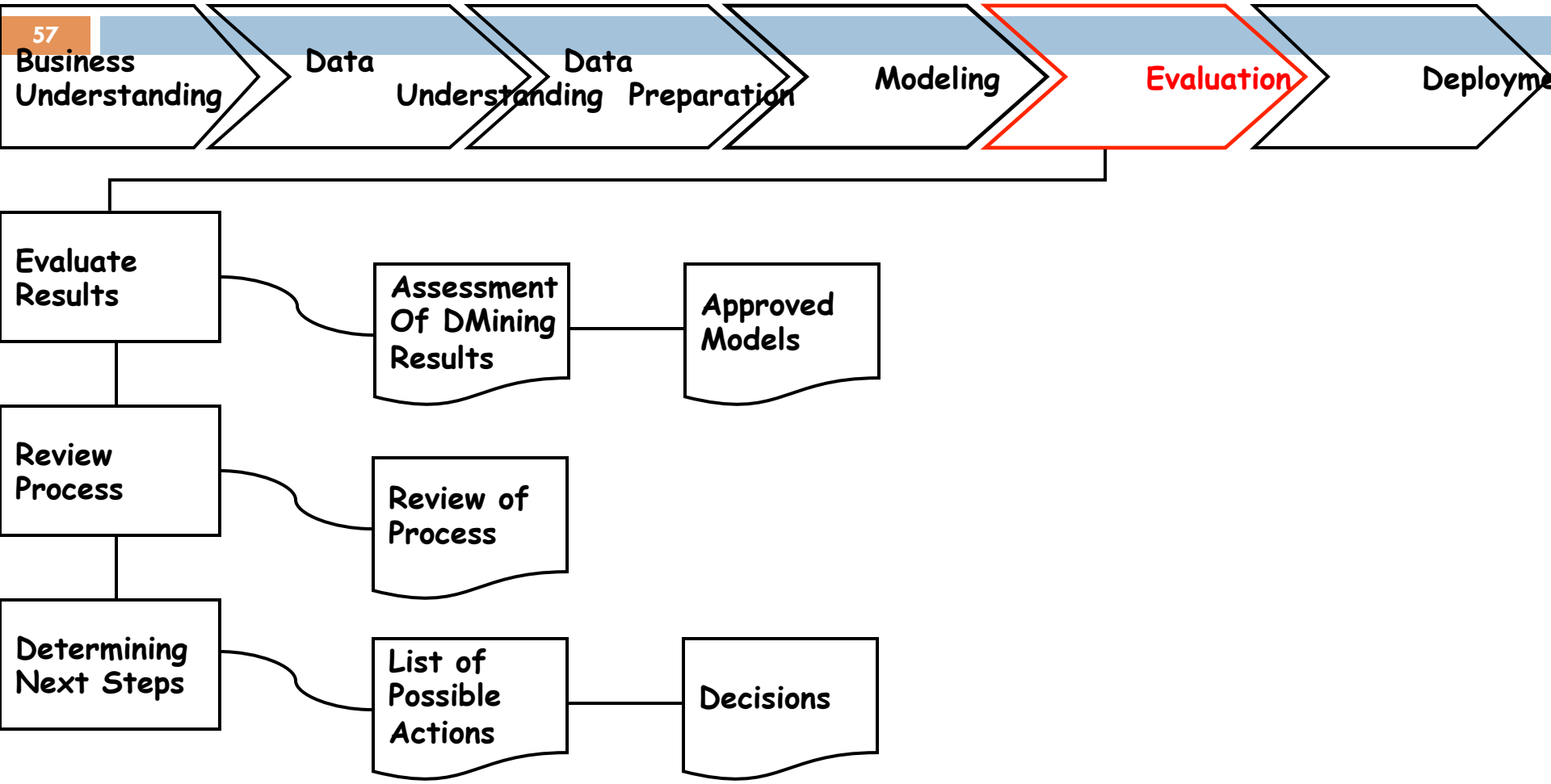


- **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- Iteration as a rule
- Process of data exploration

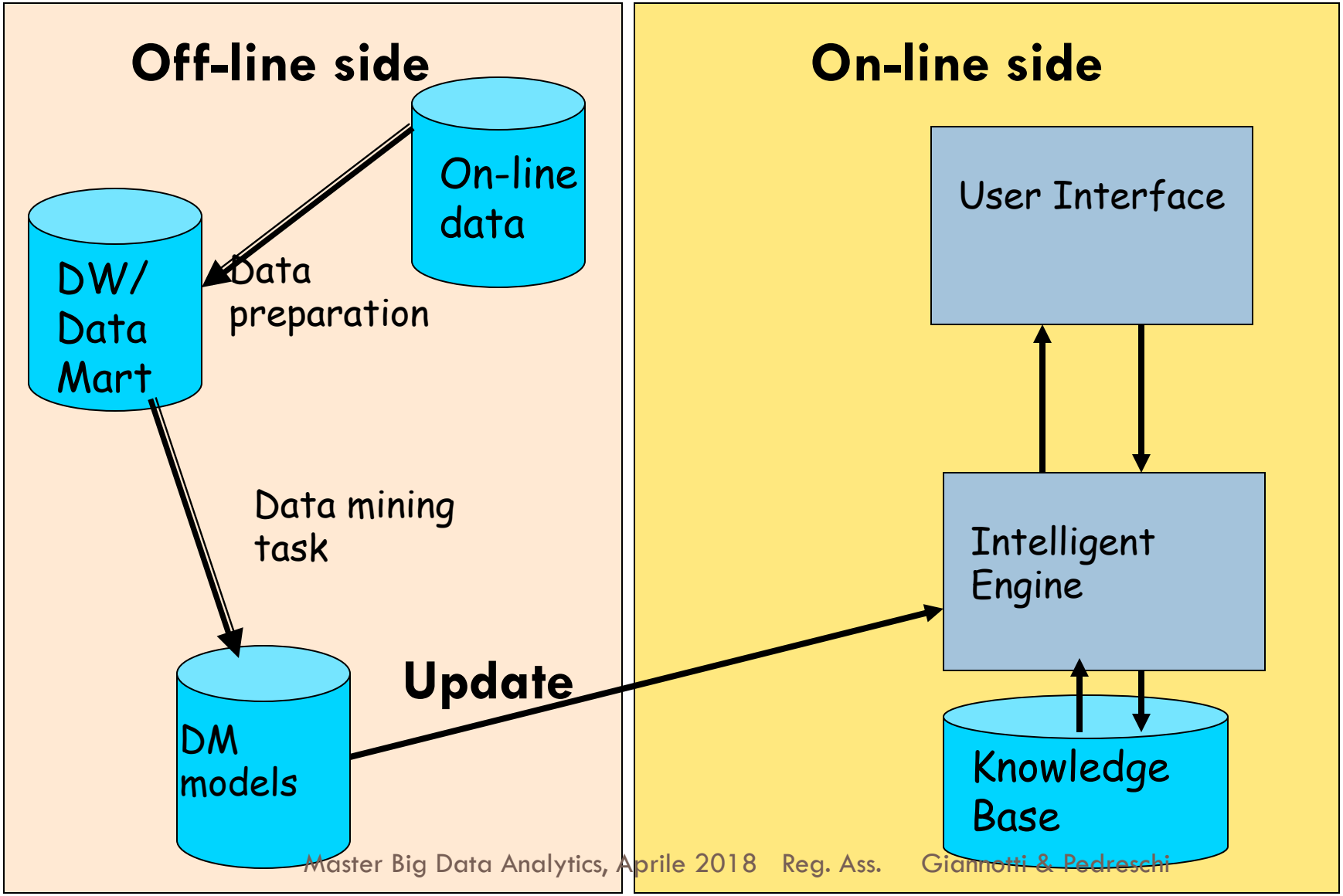

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
Determine Business Objective
Background
Business Objective
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements Assumptions Constraints
Risk and Contingencies
Terminology
Costs & Benefits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


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
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


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
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


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

Mining Based Decision Support System: Adaptive Architecture



The Big Data Analytics technologies

Visual workflow Tools

Specialized Libraries

Programming Languages