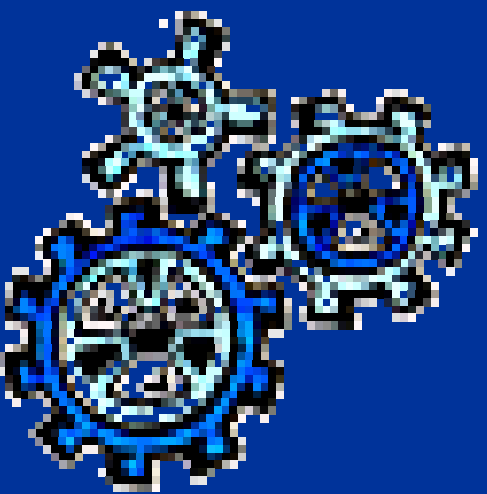
Data Mining per l'analisi dei dati



Pisa KDD Lab, ISTI-CNR & Univ. Pisa

http://www-kdd.isti.cnr.it/

Firenze, 11-12 Maggio 2006

Case Study

Data Warehousing e Data Mining per lo studio dei fattori di rischio relativo allo stato di salute dei residenti del comune di Pisa

Alessio Uva

Università degli Studi di Pisa
Laurea Specialistica in Informatica per l'Economia e per l'Azienda
Anno accademico 2004/05

Data Warehousing e Data Mining per lo studio dei fattori di rischio relativo allo stato di salute dei residenti del comune di Pisa

Candidato

Alessio Uva

Relatori

Prof.ssa Fosca Giannotti

Dott.ssa Maria Angela Vigotti

Controrelatore

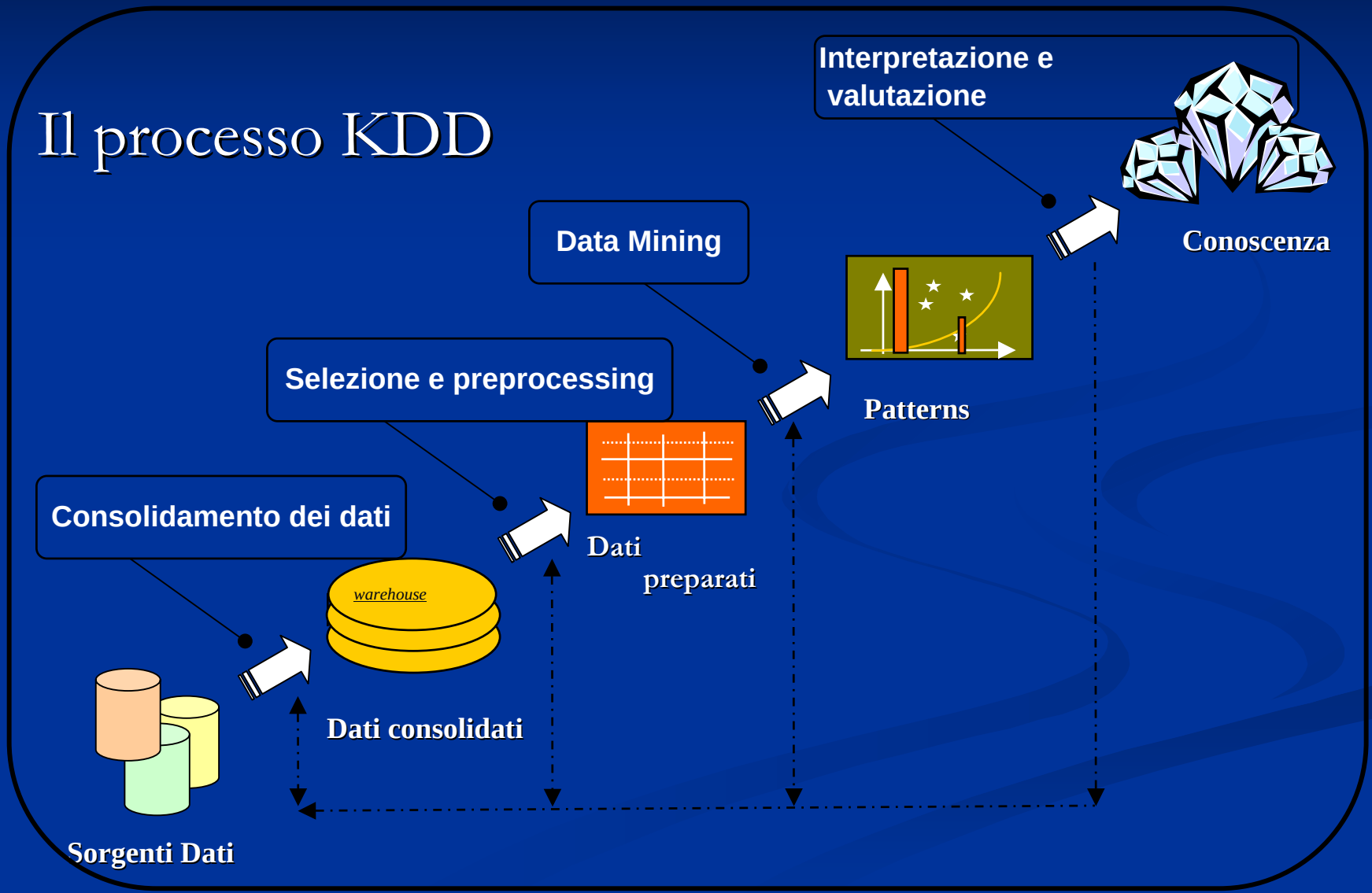
Prof.re Carlo Bianchi

Introduzione al KDD

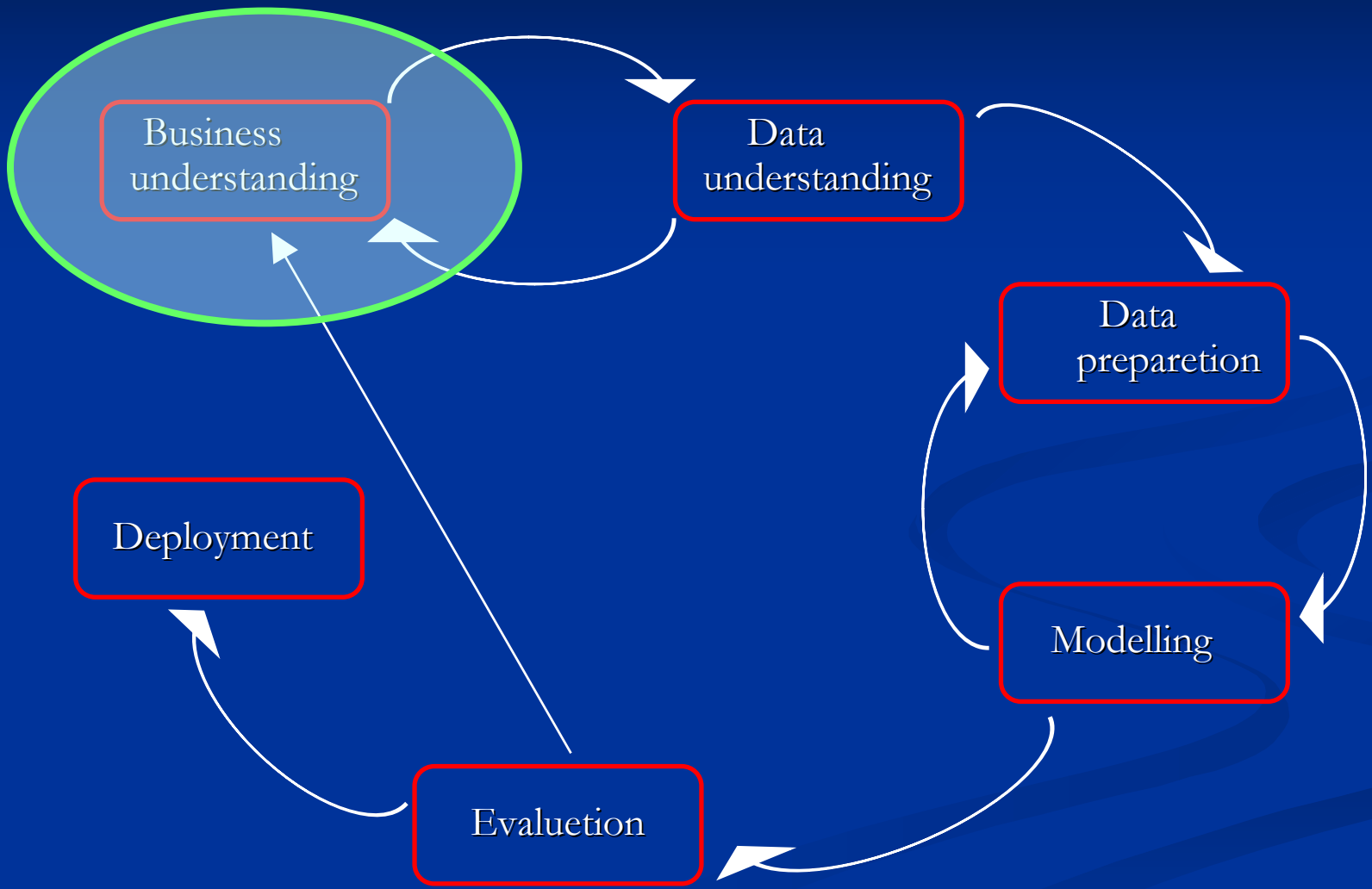
- Knowledge Discovery in Database come tecnologia fondamentale per ottenere conoscenza utile dai dati.

"KDD si pone come processo di selezione, esplorazione e modellazione di grandi masse di dati, al fine di scoprire regolarità o relazioni non note a priori, ed allo scopo di ottenere informazione utile da trasformare in conoscenza per supportare le decisioni"

Introduzione (2)



Crisp-DM model

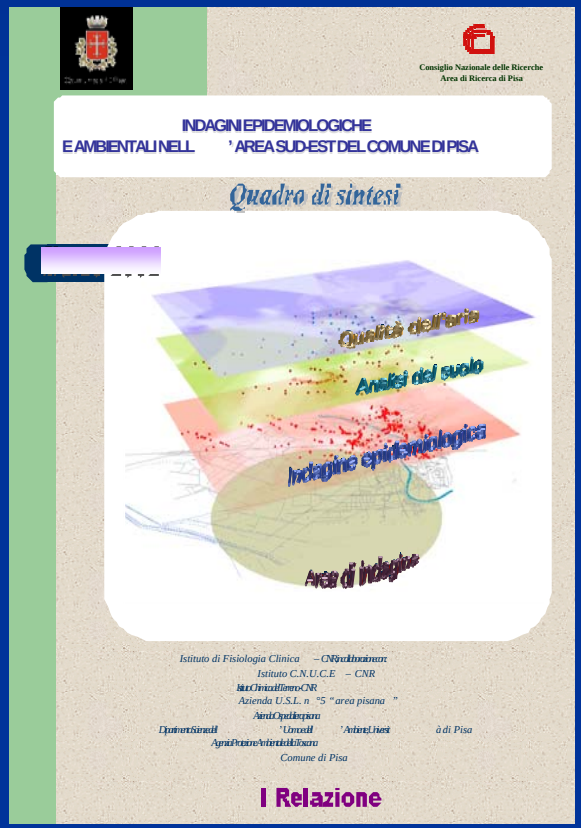


Il contesto

- Questo lavoro si inserisce in un più ampio contesto di ricerca, regionale e nazionale, in campo epidemiologico.
- Presso l'Istituto di Fisiologia Clinica sono stati effettuati studi statistici con l'obiettivo di osservare molti fenomeni di un determinato contesto ambientale per cercare di capire le correlazioni di alcune patologie di una determinata popolazione
- Gli strumenti sono essenzialmente statistici e l'osservazione dei dati mira a verificare ipotesi che lo studioso si è posto in precedenza

Il contesto

Indagine sull'Area di Ospedaletto, Pisa 1990-2000



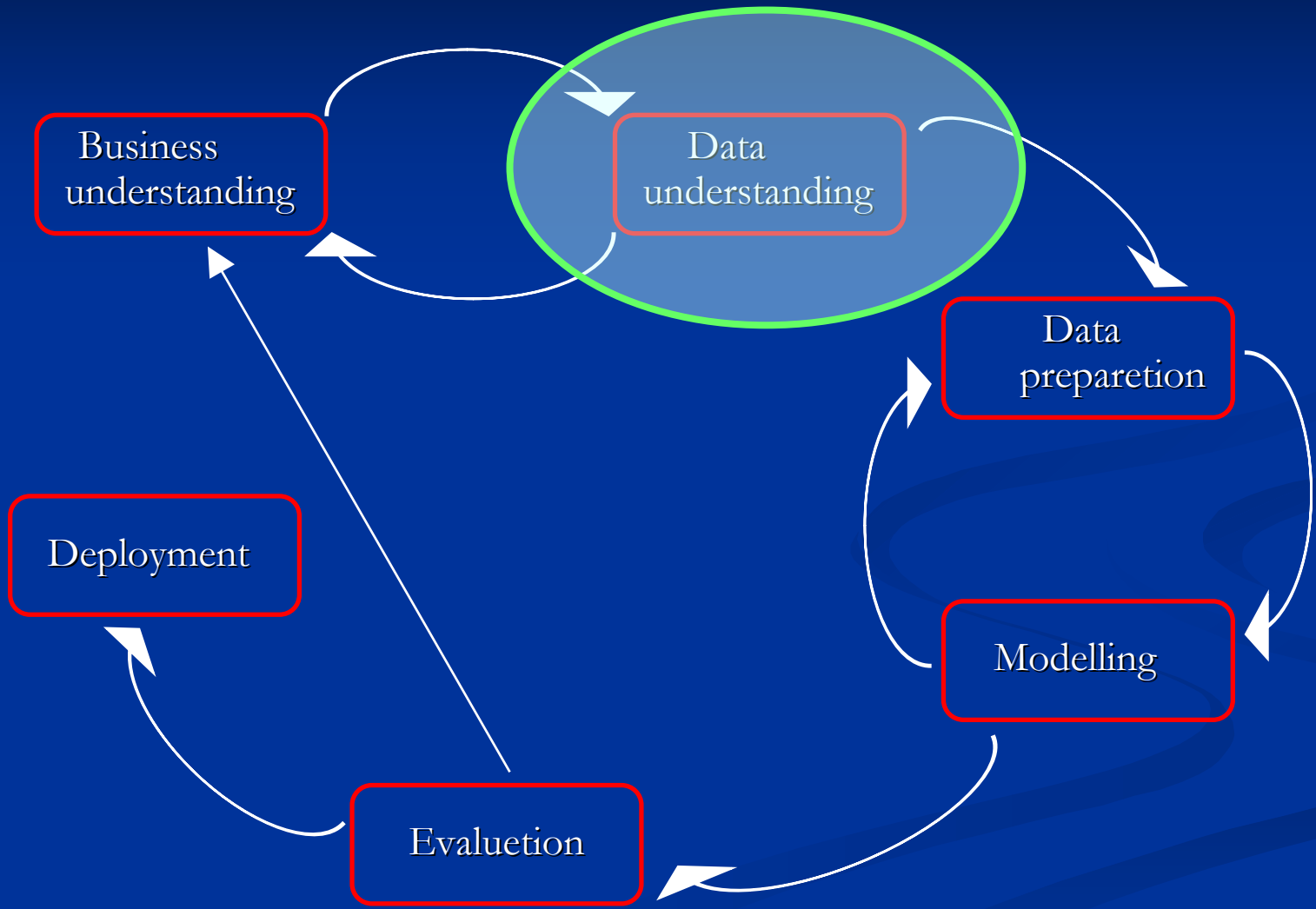
Misa-2
Effetti acuti dell'inquinamento Atmosferico, 1998-2002

Comune di Pisa	
Ricostruzione Popolazione Residente 1990 -1999	G. Bertelloni
Suddivisione territoriale, Confini sub-aree	M. Bonfanti
ASL-5 di Pisa	
Ricoveri Ospedalieri (Sistema Regionale SDO)	M.Perco, A.Calgaro
Dati di Mortalita' (Registro Mortalita' Regionale)	E. Virgone
ARPA-T - Pisa	
Dati giornalieri di inquinamento	L.Senatori
IFC-CNR di Pisa	
Accoppiamento dati anagrafici, decessi e ricoveri	A.M. Romanelli , M. Raciti, M. A. Protti , R. Cosio, M. Franchini
Elaborazioni, mappe per subaree	L.Marchini, F. Minichilli
Coordinamento generale	F.Bianchi
ISTI-CNR di Pisa	
Controllo di qualita' stradario di Pisa ed Elaborazioni GIS	R. Fresco, R. della Maggiore
Universita' di Pisa	
Coordinamento ed elaborazioni analisi	M. A. Vigotti
Elaborazioni DWH e data mining sui ricoveri	A. Uva
Aereonautica Militare -Dati meteorologici giornalieri	
Regione Toscana - ISTAT	
Variabili censuarie 1991 per sezione di Censimento	P. Baldi, M. Franci

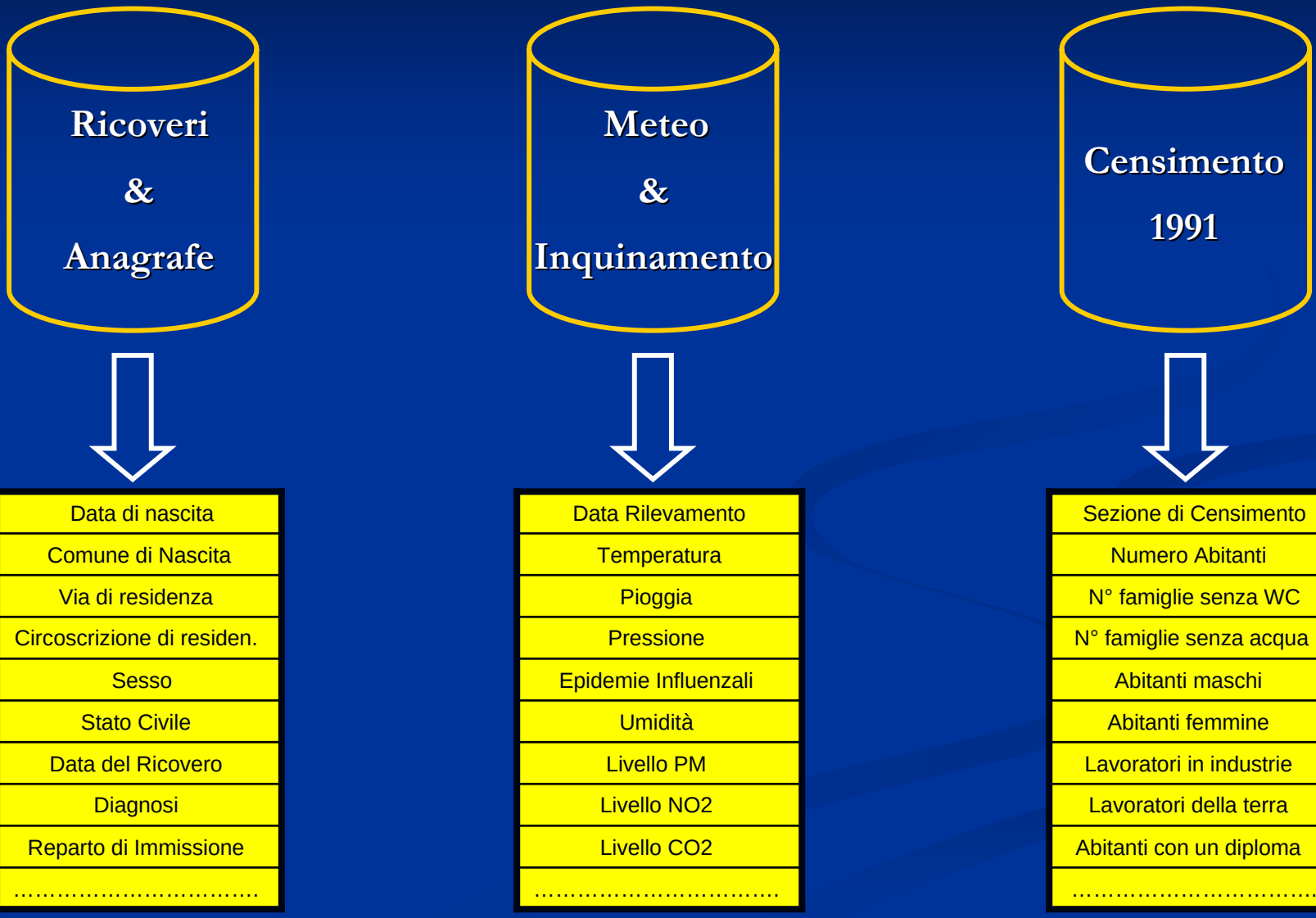
Obiettivi

- Fornire una base di dati omogenea: creazione di un Data Warehouse da utilizzare come sorgente unica di dati e come ambiente di Analisi OLAP
- Applicare tecniche di Data Mining su un date set completo, mostrando come il data mining sia un metodo valido per ottenere risultati rilevanti da confrontare con quelli ottenuti con metodologie statistiche.

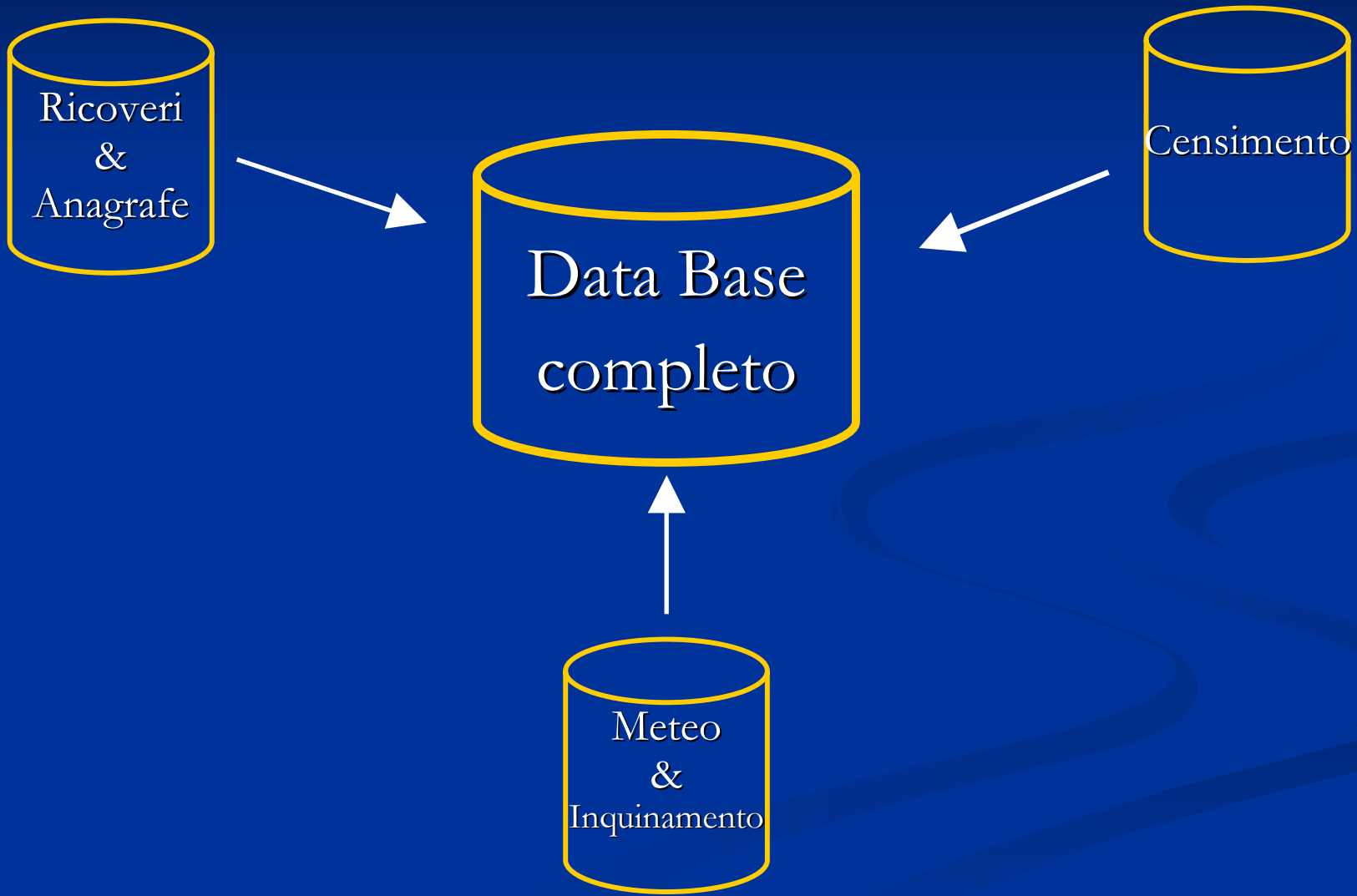
Crisp-DM model



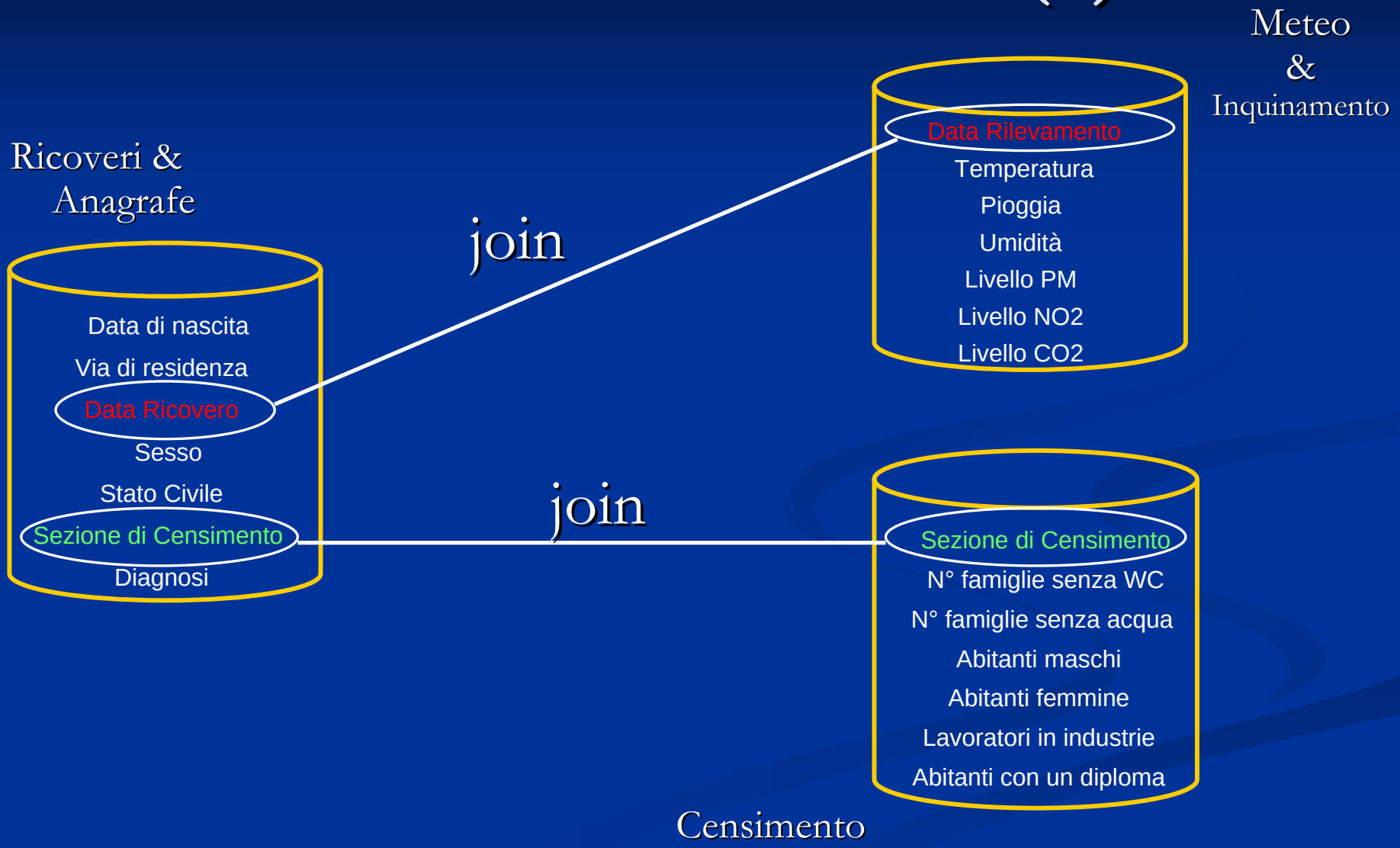
Comprensione dei dati



Un unico Data Base (1)



Un unico Data Base (2)



Un unico Data Base (3)

- Abbiamo ottenuto un data base Access, organizzato per "ricovero": ogni record rappresenta un ricovero di un singolo soggetto avvenuto ad una determinata data

- Oltre alle informazioni anagrafiche del soggetto ricoverato ora possiamo usufruire delle informazioni :
  - sullo stato di inquinamento dell'area nei tre giorni precedenti il ricovero
  - sulle caratteristiche della zona di residenza dell'individuo ricoverato

Osservazioni sui dati

- Per studiare in maniera più precisa gli effetti delle sostanze inquinanti sulla popolazione, in collaborazione con l'esperta del dominio si è deciso di creare dei campi che riportassero le informazioni dei rilevamenti nei tre giorni antecedenti il ricovero.

Data	Livello SO_2	Media SO_2	Livello NO_2	Media NO_2
1/1/1998	2.0	-	0.9	-
2/1/1998	1.9	-	4.1	-
3/1/1998	5.8	-	3.0	-
4/1/1998	3.3	**3.23**	5.8	3
5/1/1998	4.2	3,66	2.1	4.3
6/1/1998	5.0	4.43	1.1	3.6

(2.0 + 1.9 + 5.8) / 3

Data Warehousing

- Costruzione e popolamento del Data Warehouse
- Sono state scelte le misure e le dimensioni.
- Misura: “count”
- Dimensioni: censimento, diagnosi, luogo di nascita, luogo di residenza, meteo e inquinamento, tempo, persona, ospedale
- SQL Server 2000

Ricoveri, 1998-2002: valori% per diagnosi e livelli di PM10

Pisa

Livello PM-10	Tutti i ricoveri	Malattie Respir.	Malattie Endocr.	Malattie del Sangue
Altissimo	7,8	9,8	10,1	9,3
Alto	24,1	27,8	25,3	26,2
Normale	39,9	39,8	39,4	35,5
Basso	28,2	25,4	25,2	28,9
	100,0	100,0	100,0	100,0

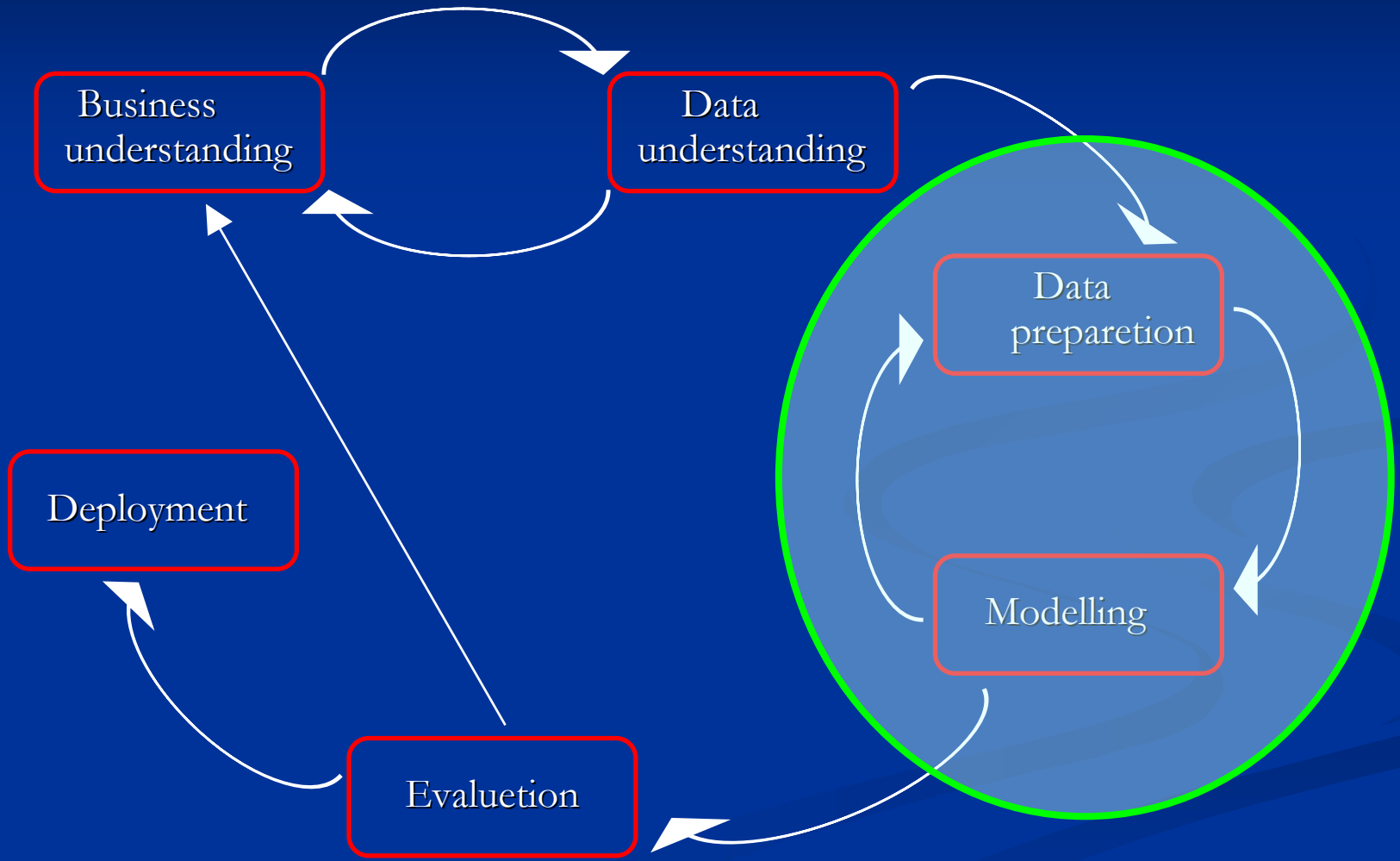
Circoscrizione n.5 Don Bosco, Cisanello

Altissimo	7,7	11,5	9,5	9,0
Alto	23,8	28,0	24,5	22,4
Normale	40,7	34,2	39,0	33,7
Basso	27,8	26,2	27,0	34,8

Circoscrizione n.5 Bambini 2-10 anni

Altissimo	9,3	16,5
Alto	25,2	29,1
Normale	37,8	27,8
Basso	27,6	26,9

Crisp-DM model



Preprocessing (1)

- Discretizzazione:
  - Età: i valori dell'attributo sono stati divisi in 7 fasce
  - Livello sostanze inquinanti: i valori degli attributi contenenti le informazioni sulle rilevazioni delle sostanze inquinanti e meteo sono stati divisi in 3 o 4 fasce, a seconda dell'attributo (Altissimo, Alto, Medio, Basso)
  - Alcuni attributi contenenti le informazione censuarie sono stati ricalcolati in base alla percentuale sulla popolazione e successivamente discretizzati.

Proproccesing (2)

- Altre operazioni di preprocessing sono state realizzate tenendo presente le caratteristiche dell'algoritmo di mining utilizzato
  - C5.0 (alberi di decisione) predilige attributi continui, mentre l'algoritmo Apriori (regole associative) utilizza solo attributi categorici.

Data Mining

Software utilizzati:

- Clementine 6.5
- KDDML
- Knolwledge Studio

Algoritmi di mining utilizzati:

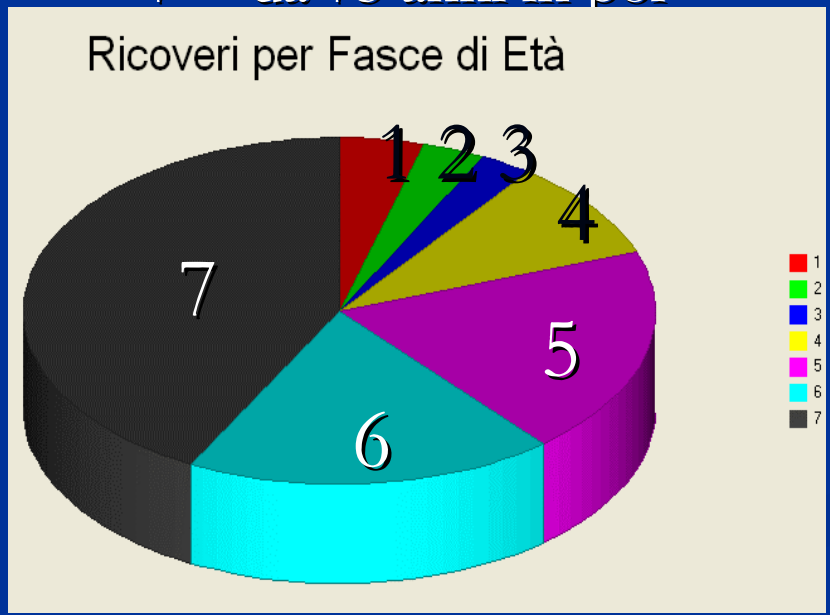
- C 5.0 per gli alberi di decisione
- Apriori per le regole associative
- K-means per il clustering

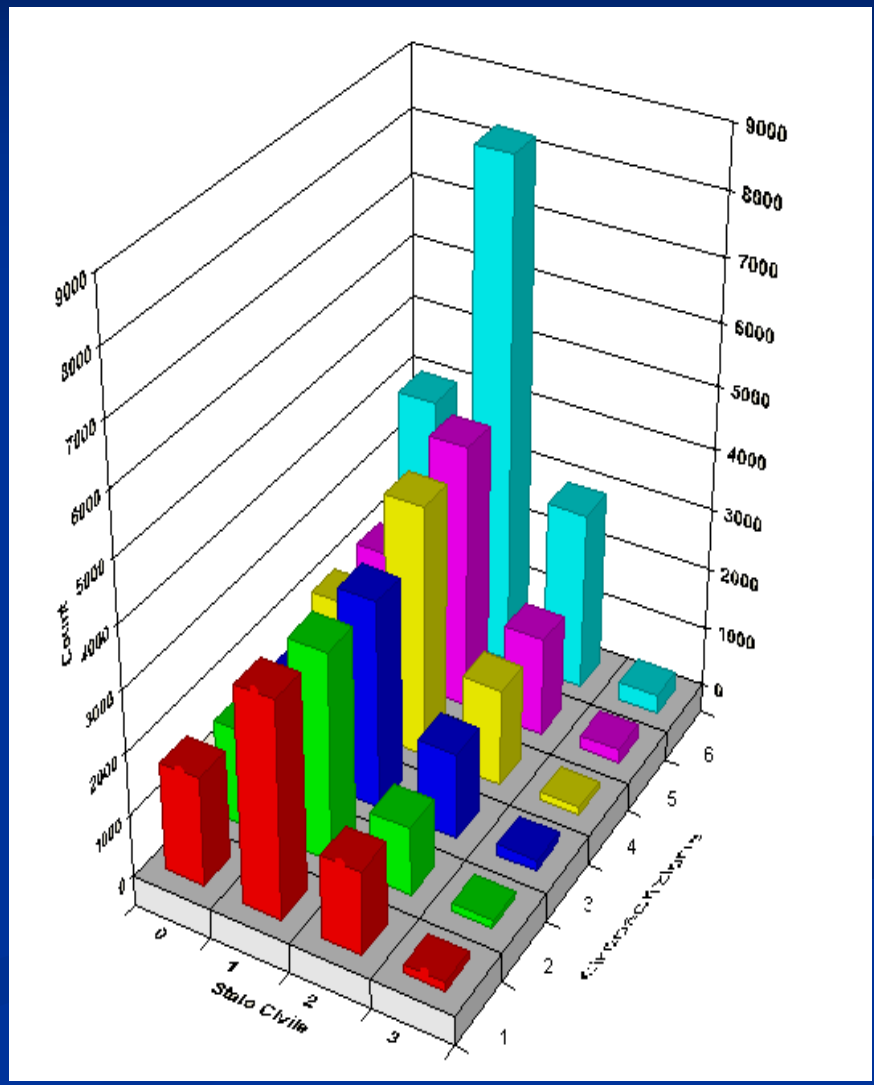
Studio degli attributi

Attribute Info							
Field Name	**Type**	**Missed Values**	**Cardinality**	**Mean**	**Varian**	**Min**	**Max**
Circoscrizione	nominal	0 (0%)	6: 15437 (30%) 5: 8292 (16%) 4: 7889 (16%) 3: 6410 (13%) 2: 6152 (12%) 1: 6697 (13%)	n/a	n/a	n/a	n/a
Sesso	nominal	0 (0%)	F: 27734 (55%) M: 23143 (45%)	n/a	n/a	n/a	n/a
Eta	numeric	0 (0%)	n/a	54.2	614	0.0	106
TempDisc	nominal	111 (0%)	Freddissimo: 2117 (4%) Fresco: 14029 (28%) Caldissimo: 4026 (8%) Mite: 14608 (29%) Caldo: 10207 (20%) Freddo: 5779 (11%)	n/a	n/a	n/a	n/a
LivelloPm	nominal	0 (0%)	Basso: 14371 (28%) Altissimo: 3946 (8%) Normale: 20308 (40%) Alto: 12252 (24%)	n/a	n/a	n/a	n/a
LivelloPress	nominal	3670 (7%)	Alta: 12933 (27%) Normale: 16391 (35%) Bassa: 17883 (38%)	n/a	n/a	n/a	n/a

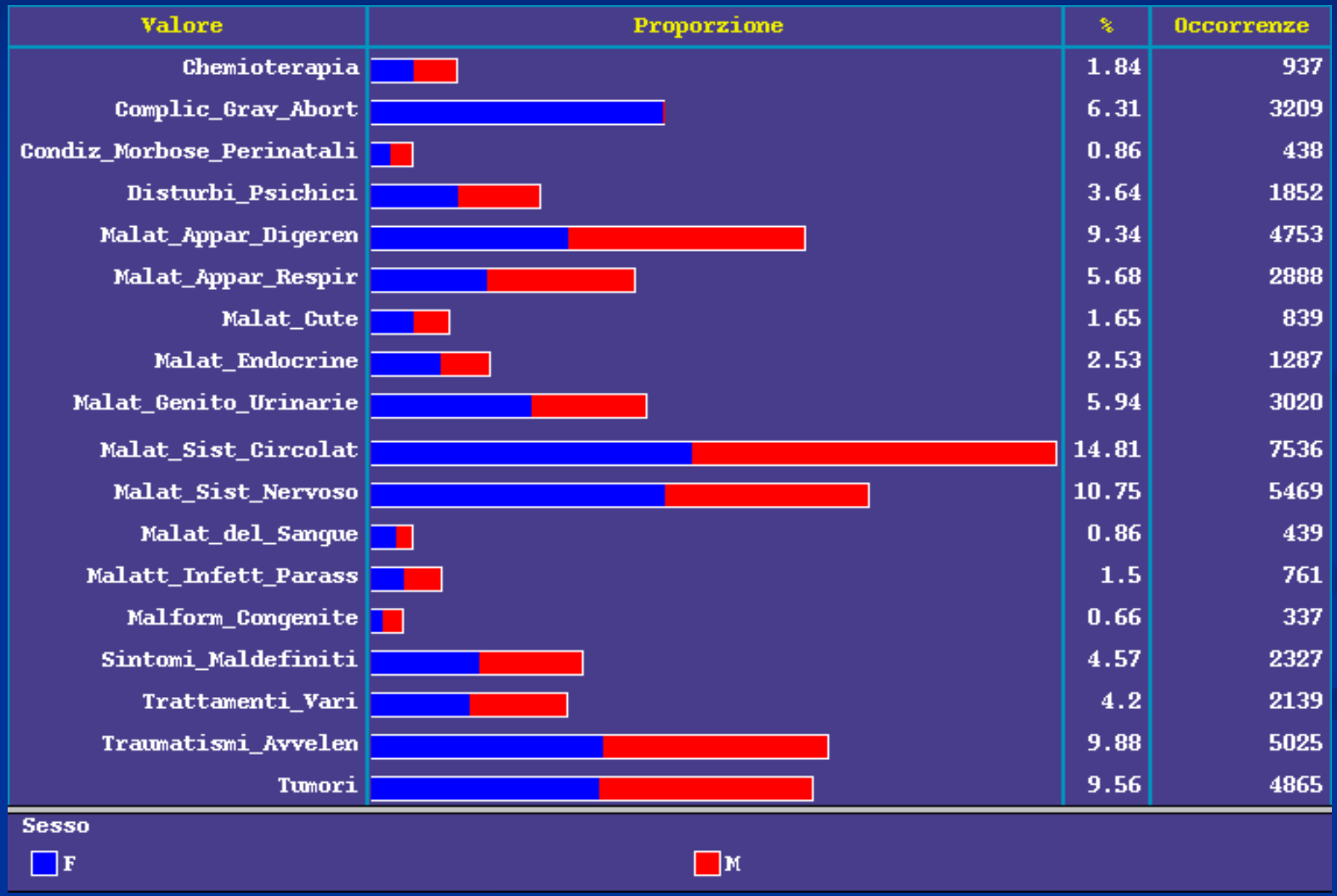
Analisi delle distribuzioni (1)

- 1 = da 0 a 1 anni
- 2 = da 2 a 10 anni
- 3 = da 11 a 18 anni
- 4 = da 18 a 35 anni
- 5 = da 35 a 55 anni
- 6 = da 55 a 75 anni
- 7 = da 75 anni in poi





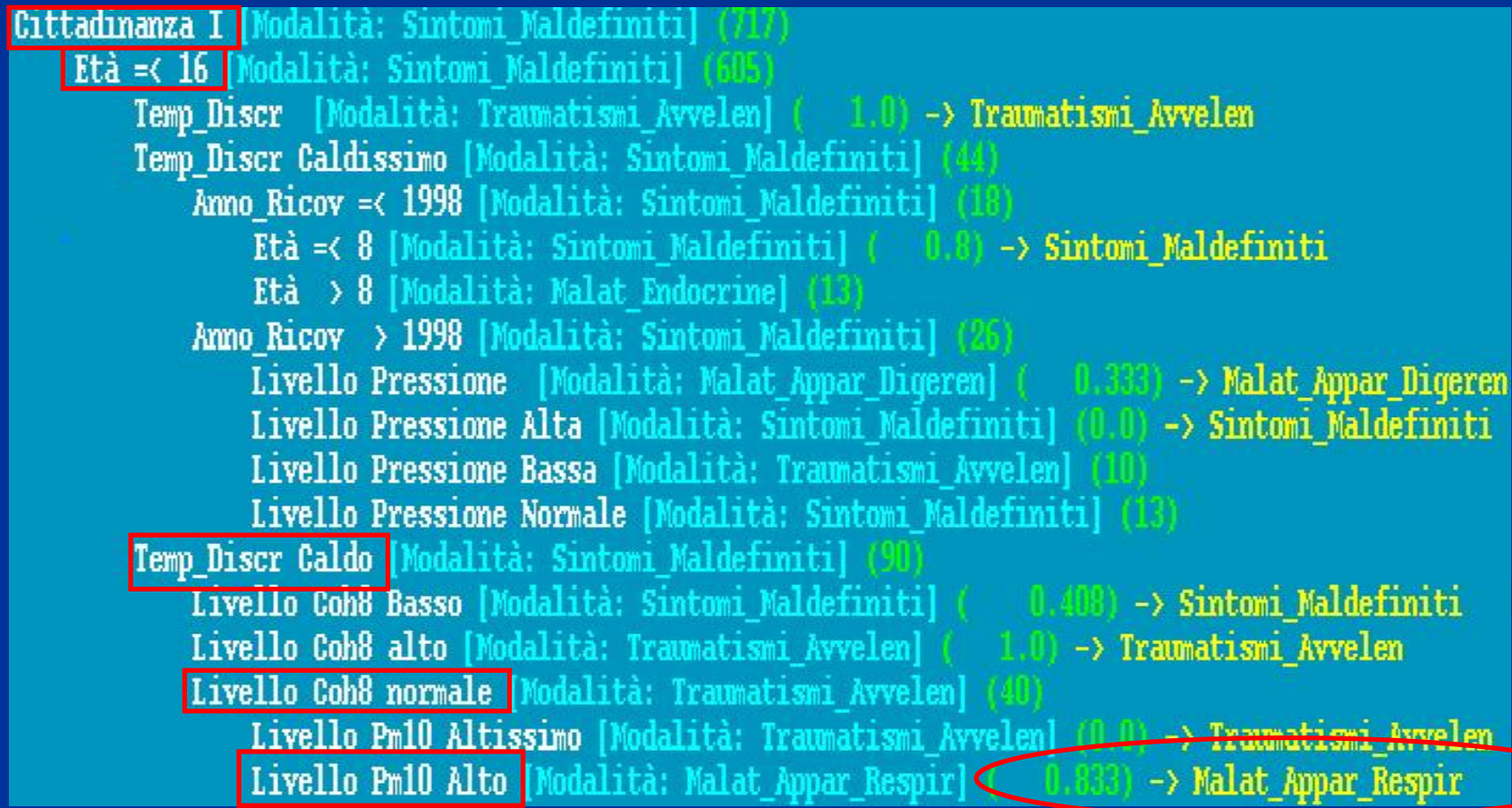
Analisi delle distribuzioni (2)



Valore	Proporzione	%	Occorrenze
Chemioterapia		1.84	937
Complic_Grav_Abort		6.31	3209
Condiz_Morbose_Perinatali		0.86	438
Disturbi_Psichici		3.64	1852
Malat_Appar_Digeren		9.34	4753
Malat_Appar_Respir		5.68	2888
Malat_Cute		1.65	839
Malat_Endocrine		2.53	1287
Malat_Genito_Urinarie		5.94	3020
Malat_Sist_Circolat		14.81	7536
Malat_Sist_Nervoso		10.75	5469
Malat_del_Sangue		0.86	439
Malatt_Infett_Parass		1.5	761
Malform_Congenite		0.66	337
Sintomi_Maldefiniti		4.57	2327
Trattamenti_Vari		4.2	2139
Traumatismi_Avvelen		9.88	5025
Tumori		9.56	4865

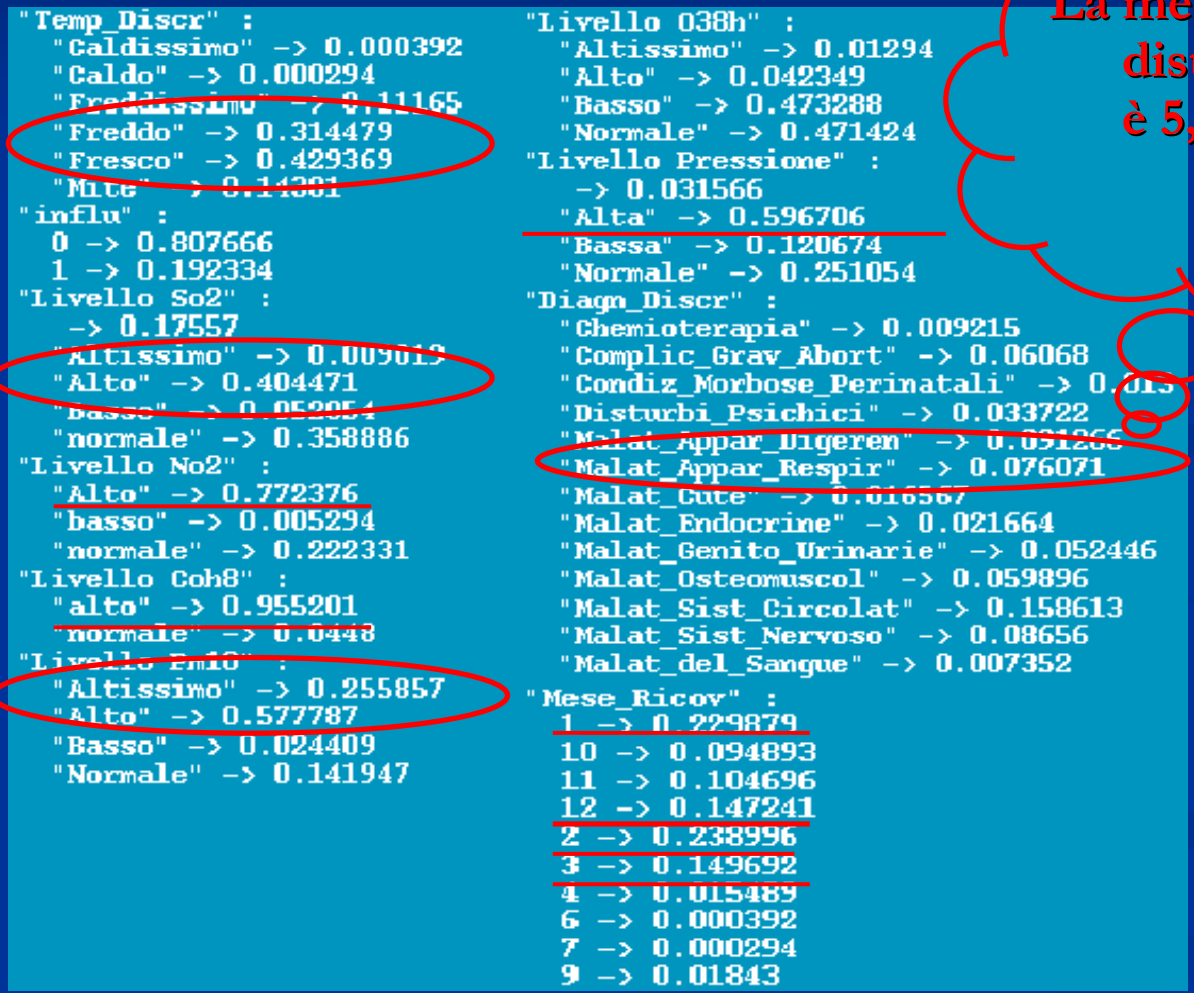
Albero di decisione

- Algoritmo utilizzato : C5.0
- Attributo target : diagnosi



Clustering

- Algoritmo utilizzato : K-means



La media della distribuzione è 5,7%

Questo cluster descrive una situazione tipicamente invernale, in cui il freddo e la pressione alta contribuiscono al ristagnare delle sostanze inquinanti nelle città.

In tali condizioni la % di ricoveri per malattie dell'apparato respiratorio risulta più alta (7,6%) rispetto alla media della distribuzione (5.7%)

Regole Associative

- Per la scoperta di interessanti regole associative siamo passati ad analisi più specifiche.
- Sono stati creati dei nuovi campi:

Attributi	Valori
Diagnosi	Malattie sistema circolatorio, disturbi psichici, tumori, malattie apparato digerente.
Malat_app_resp	SI / NO
Malat_sist_circol	SI / NO
Malat_app_digeren	SI / NO

Regole Associative (2)

- Algoritmo utilizzato : Apriori
- Attributo target : Malattie apparato circolatorio

Istan	Confid	Cons.	Ant. 1	Ant. 2	Ant. 3	Ant. 4	Ant. 5
6	83,3%	Circolatorio	Sesso = M	Stato civ = 2	Cicoscr = 4	Mese = 2	
6	83,3%	Circolatorio	Sesso = M	Stato civ = 2	Mese = 1	Influ = 1	Giorno_sett=4
17	82,4%	Circolatorio	PM = Alto	Oh38 = Alto	Temp =Calda	Stato civ=2	Circoscr = 1
21	71,4%	Circolatorio	PM = Alto	Età = 7	Temp =Calda	Stato civ=2	Circoscr = 1

Conclusioni

Cosa è stato fatto:

- E' stata creata una sorgente unica di dati, facilmente aggiornabile e adatta alle analisi OLAP e di Data Mining
- Sono state effettuate analisi che hanno dimostrato l'efficacia delle tecnologie utilizzate
- Parte dei risultati prodotti sono già stati visionati con interesse dagli esperti del dominio e da rappresentanti del comune e della provincia.

Cosa c'è da fare:

- Aggiornare il Data Warehouse con i dati del censimento del 2001
- Data l'ampia dimensione del Data Set si possono fare ulteriori analisi e creare ulteriori indici per le analisi OLAP
- Introdurre dati spaziali GIS per georeferenziare la popolazione e creare mappe dei risultati