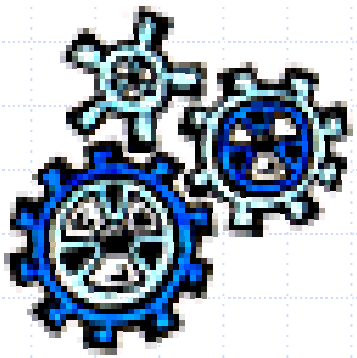
Preparazione e caratteristiche dei Dati per Data Mining

Fosca Giannotti
f.giannotti@isti.cnr.it

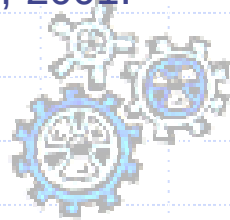
Mirco Nanni
nanni@isti.cnr.it



ISTI-CNR Pisa

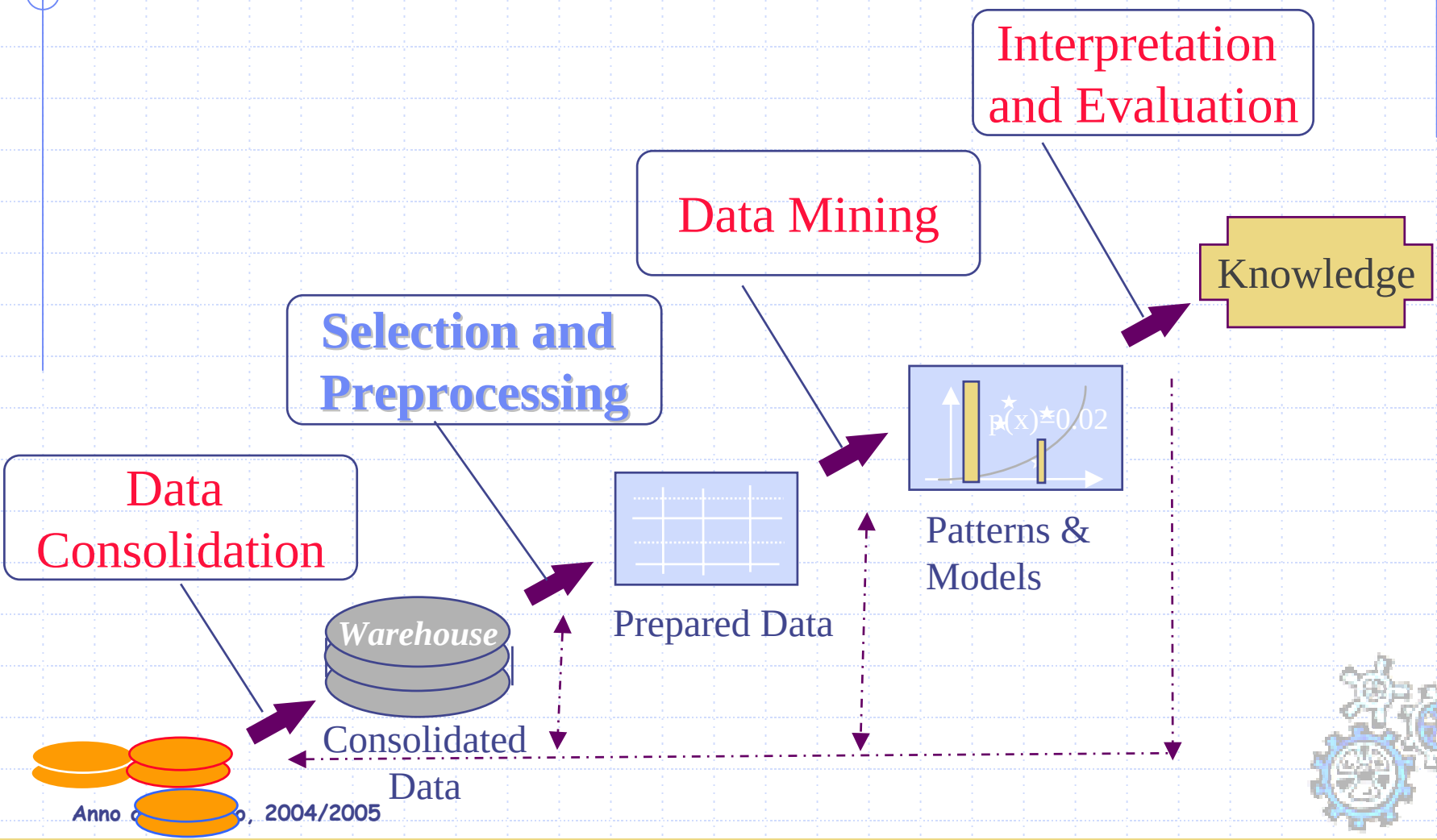
Materiale

- **Lucidi delle lezioni** (Slides PowerPoint):
  - Primo autore: G. Manco Revisione: M. Nanni
  - Versione attuale: In distribuzione
- **Testi di Riferimento**
  - J. Han, M. Kamber. ***Data Mining: Concepts and Techniques***. Morgan Kaufmann, 2000.
  - Dorian Pyle. ***Data Preparation for Data Mining.*** Morgan Kaufmann, 1999.
  - D. Hand, H. Mannila, P. Smyth. ***Principles of Data Mining***. MIT Press, 2001.



Anno accademico, 2004/2005

Il Processo di KDD



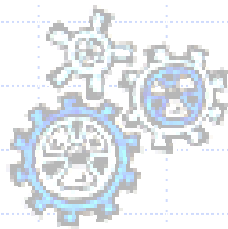
I Contenuti

- Introduzione e Concetti di Base
  - Motivazioni
  - Il punto di partenza: dati consolidati, Data Marts
- Data Selection
  - Manipolazione di Tabelle
- Information Gathering
  - Misurazioni
  - Visualizzazioni
  - Statistiche
- Data cleaning
  - Trattamento di valori anomali
  - Identificazione di Outliers
  - Risoluzione di inconsistenze
- Data reduction
  - Campionamento
  - Riduzione di Dimensionalità
- Data transformation
  - Normalizzazioni
  - aggregazione
  - Discretizzazione
- Data Similarity
  - Similarity and Dissimilarity (on Single attribute)
  - Distance (Many attributes)
  - Distance on Binary data (Simple matching; Jaccard)
  - Distance on Document Data
- Data Esploration (lumtidimensional array)

Anno accademico, 2004/2005

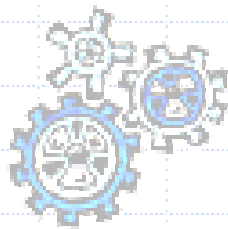
Problemi tipici

- Troppi dati
  - dati sbagliati, rumorosi
  - dati non rilevanti
  - dimensione intrattabile
  - mix di dati numerici/simbolici
- Pochi dati
  - attributi mancanti
  - valori mancanti
  - dimensione insufficiente



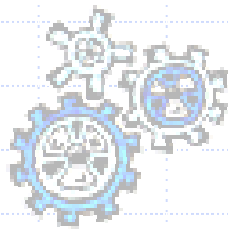
Il Data Preprocessing è un Processo

- Accesso ai Dati
- Esplorazione dei Dati
  - Sorgenti
  - Quantità
  - Qualità
- Ampliamento e arricchimento dei dati
- Applicazione di tecniche specifiche



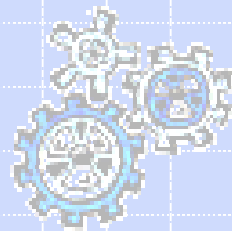
Il Data Preprocessing dipende (ma non sempre) dall'Obiettivo

- Alcune operazioni sono necessarie
  - Studio dei dati
  - Pulizia dei dati
  - Campionamento
- Altre possono essere guidate dagli obiettivi
  - Trasformazioni
  - Selezioni



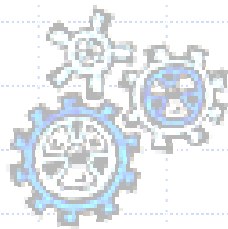
Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection ⇦



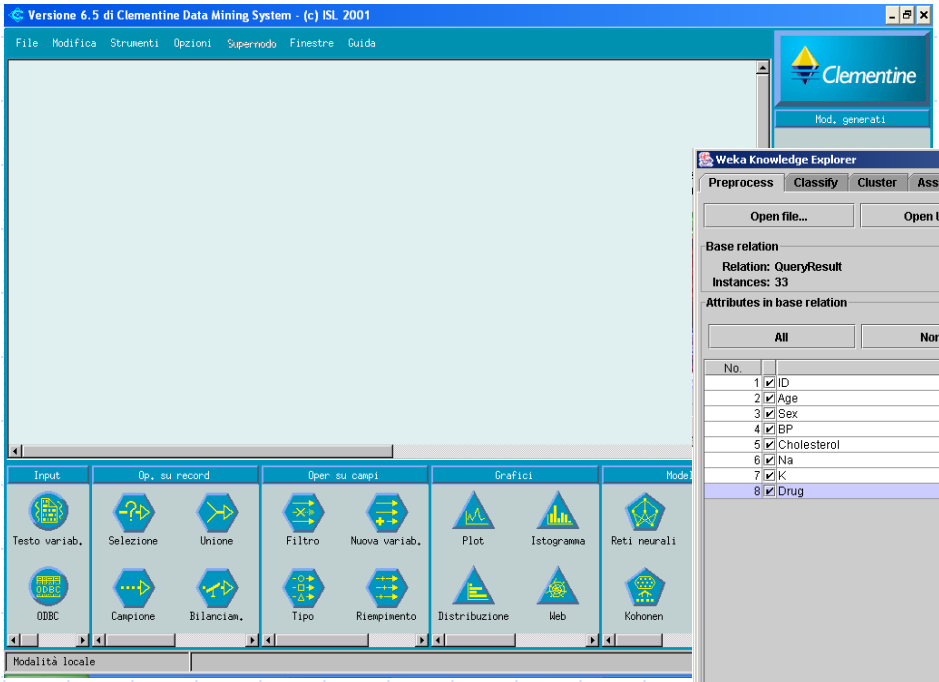
E’ sempre necessario SQL?

- I moderni tools raggruppano una serie di operazioni in maniera uniforme
- La metafora di interazione è visuale
  - Esempi che vedremo:
    - Clementine
    - Weka
- SQL è più generico
  - Ma anche più difficile da usare

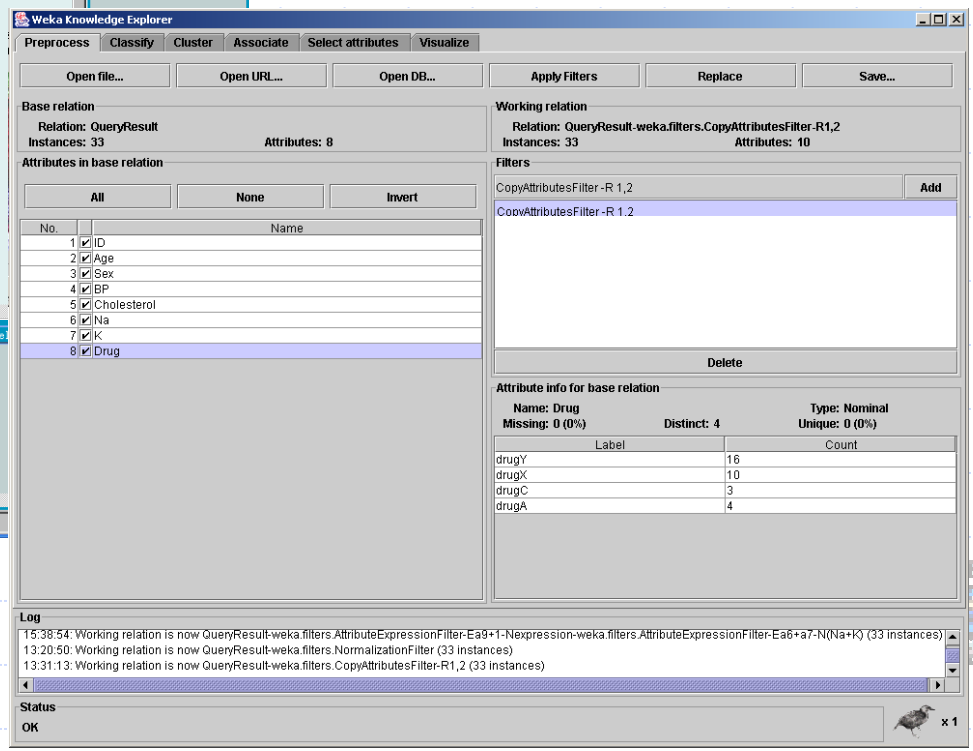


Es. due piattaforme per DM

Clementine



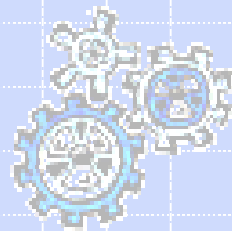
Weka



Anno accademico, 2004/2005

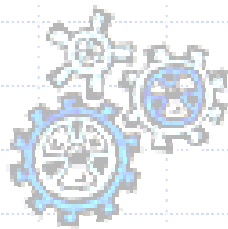
Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering ⇦



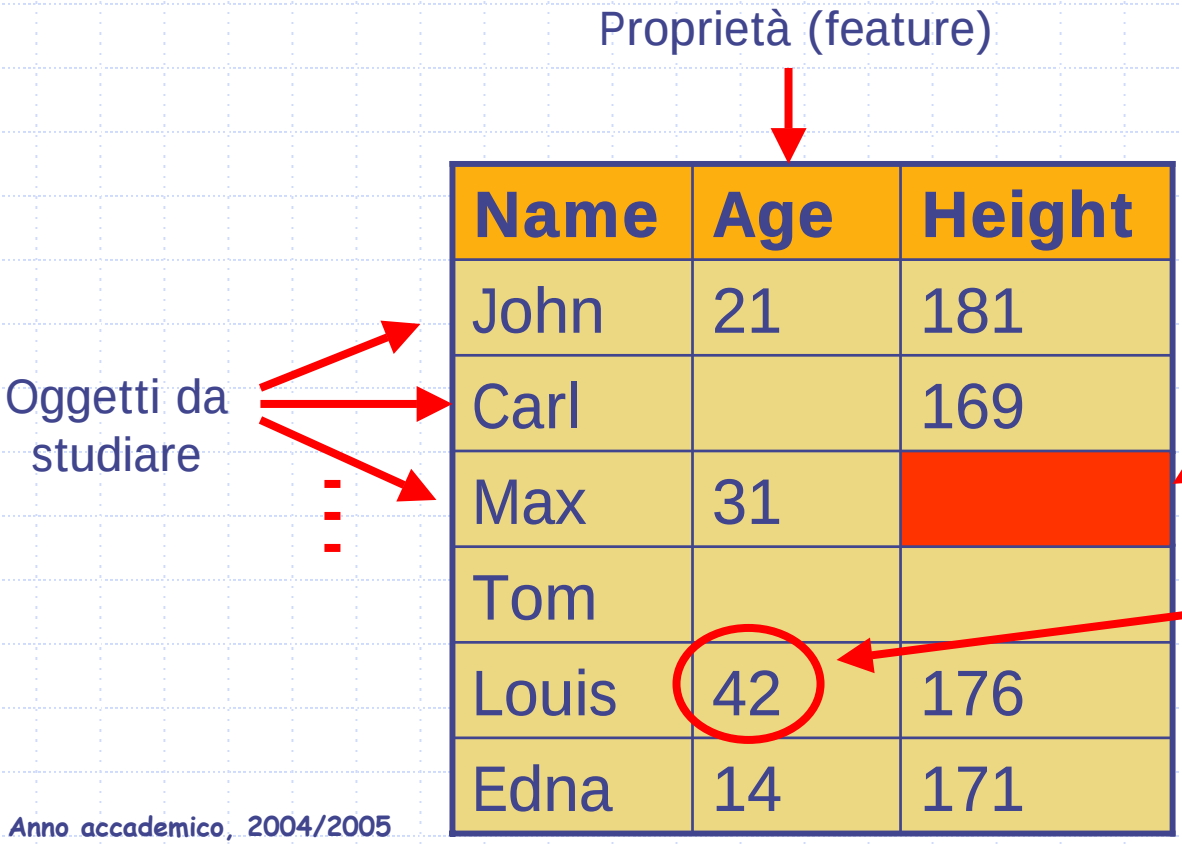
Oggetti, Proprietà,Misurazioni

- Il mondo reale consiste di oggetti
  - Automobili, Vigili, Norme, ...
- Ad ogni oggetto è associabile un insieme di proprietà (features)
  - Colore, Cilindrata, Proprietario, ...
- Su ogni proprietà è possibile stabilire delle misurazioni
  - Colore = rosso, Cilindrata = 50cc, Proprietario = luigi, ...



La Nostra Modellazione

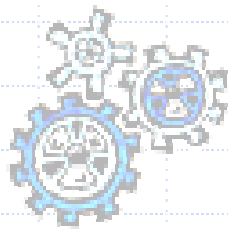
- La realtà è descritta da una tabella



Anno accademico, 2004/2005

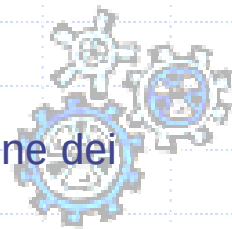
Tipi di misure

- Misure Discrete (simboliche, categoriche,qualitative)
  - Nominali → identificatori univoci (Cod. Fiscale)
  - Ordinali → è definito un ordine (low < high)
  - Binarie → due soli valori (T/F, 1/0,...)
- Misure Continue
  - Interval-Based → Scalabili di fattore costante (es.: misure in MKS e CGS)
  - Ratio-Scaled → Scalabili linearmente (ax+b) (es.: temperature °C e °F)



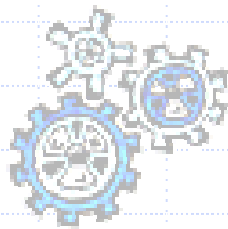
Caratteristiche delle Variabili (dei data sets)

- **Sparsità**
  - Mancanza di valore associato ad una variabile
    - Un attributo è sparso se contiene molti valori nulli
- **Monotonicità**
  - Crescita continua dei valori di una variabile
    - Intervallo [-∞, ∞] (o simili)
  - Non ha senso considerare l'intero intervallo
- **Outliers**
  - Valori singoli o con frequenza estremamente bassa
  - Possono distorcere le informazioni sui dati
- **Dimensionalità delle variabili**
  - Il numero di valori che una variabile può assumere può essere estremamente alto
    - Tipicamente riguarda valori categorici
- **Dimensionalità degli oggetti**
  - Il numero di attributi che un oggetto ha può essere estremamente alto
    - Es. prodotti di un market basket
- **Anacronismo**
  - Una variabile può essere contingente: abbiamo i valori in una sola porzione dei dati



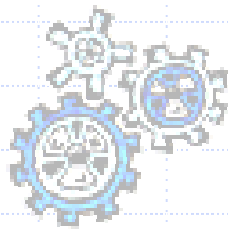
Bias

- Un fattore esterno significativo e rilevante nei dati
  - Comporta problemi (espliciti o impliciti) nei dati
  - Molti valori della variabile **`Velocità`** in una tabella **`Infrazioni`** è alto
- Il problema è sistematico
  - Appare con una certa persistenza
    - Il misuratore della velocità è tarato male
- Il problema può essere trattato
  - Il valore è suscettibile di una distorsione, che deve essere considerata
    - Considera solo i valori che vanno oltre una certa tolleranza



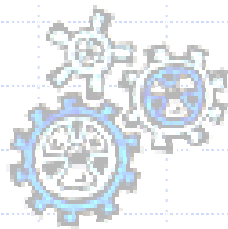
Descrizione dei dati

- Grafici
  - Distribuzione frequenze
  - Correlazione
  - Dispersione
- Misure
  - Media, mediana, quartili
  - Varianza, deviazione standard
  - Forma, simmetria, curtosi



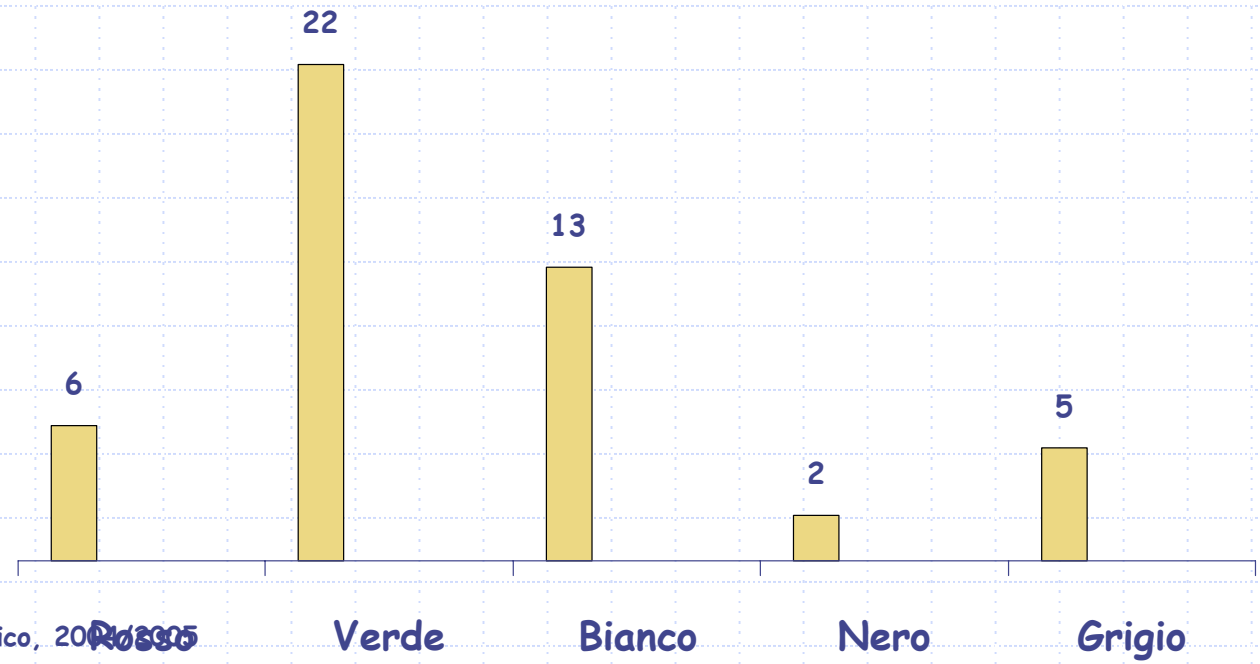
Visualizzazione dati qualitativi

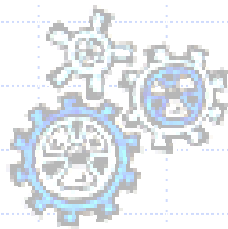
- Rappresentazione delle frequenze
  - Diagrammi a barre
  - Ortogrammi
  - Aerogrammi
- Correlazione
  - Web diagrams
- Ciclicità
  - Diagrammi polari



Diagrammi di Pareto

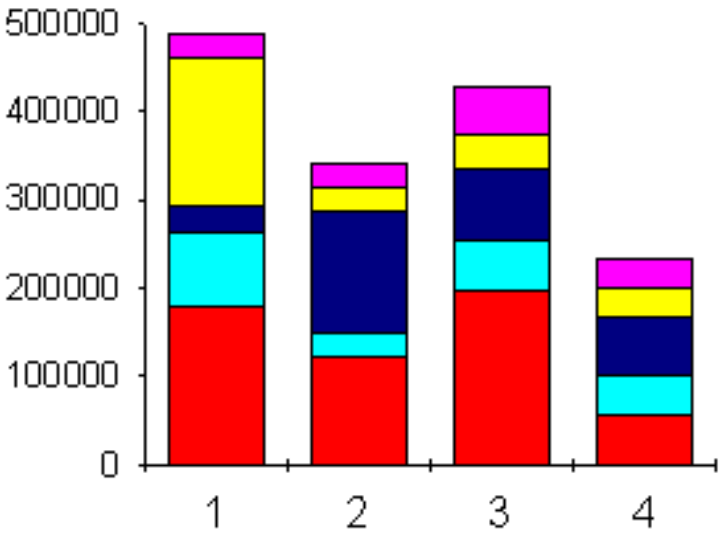
- Diagammi a barre distanziate
- Un assortimento di eventi presenta pochi picchi e molti elementi comuni

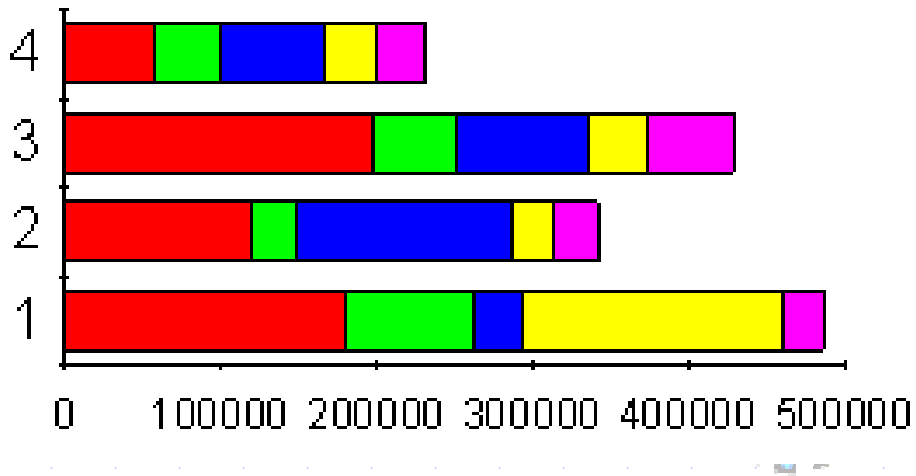




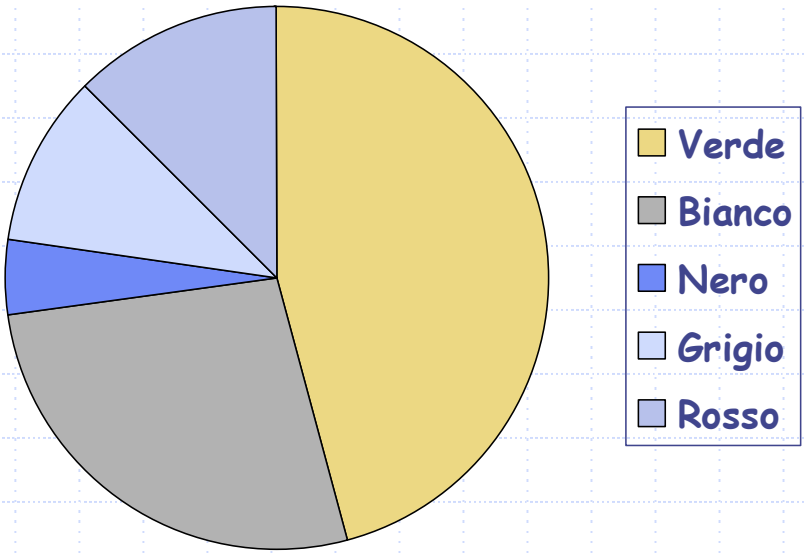
Ortogrammi

- Ogni colonna indica la la distribuzione interna per un dato valore e la frequenza





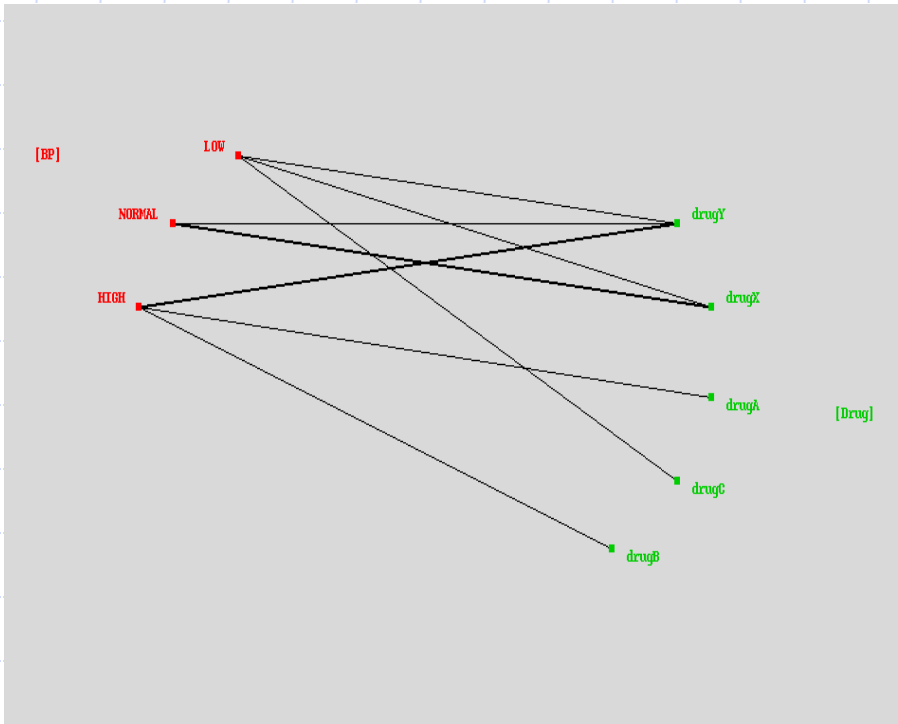
Aerogrammi

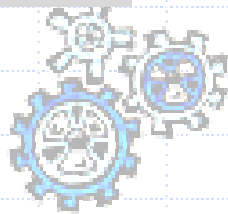


- Rappresentazioni a torta
- frequenza della distribuzioni

Web

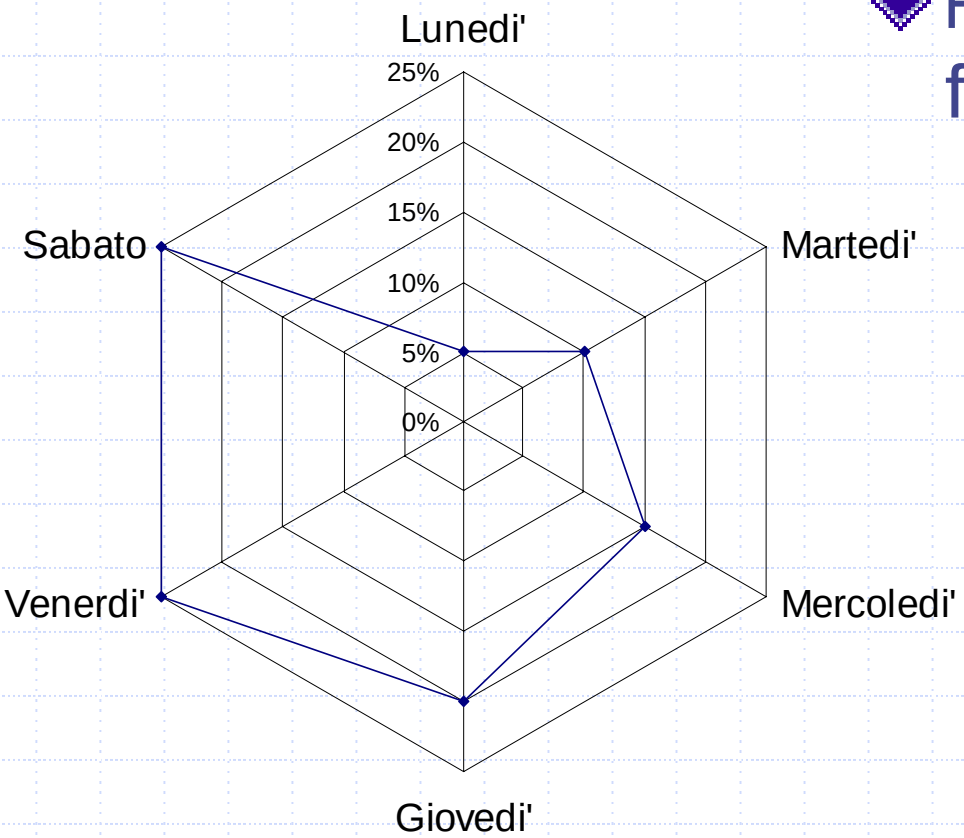
- Visualizzano correlazioni tra valori simbolici





Anno accademico, 2004/2005

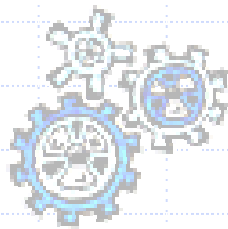
Diagrammi polari



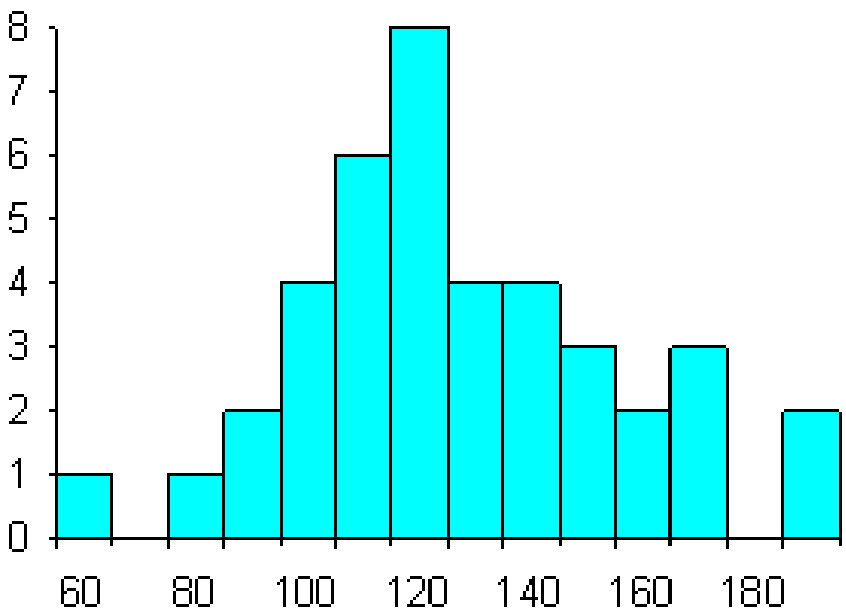
- Rappresentano fenomeni ciclici
  - E.g., concentrazione delle vendite nell'arco settimanale

Dati Quantitativi

- Istogrammi
- Poligoni
- Stem and leaf
- Dot Diagrams
- Diagrammi quantili

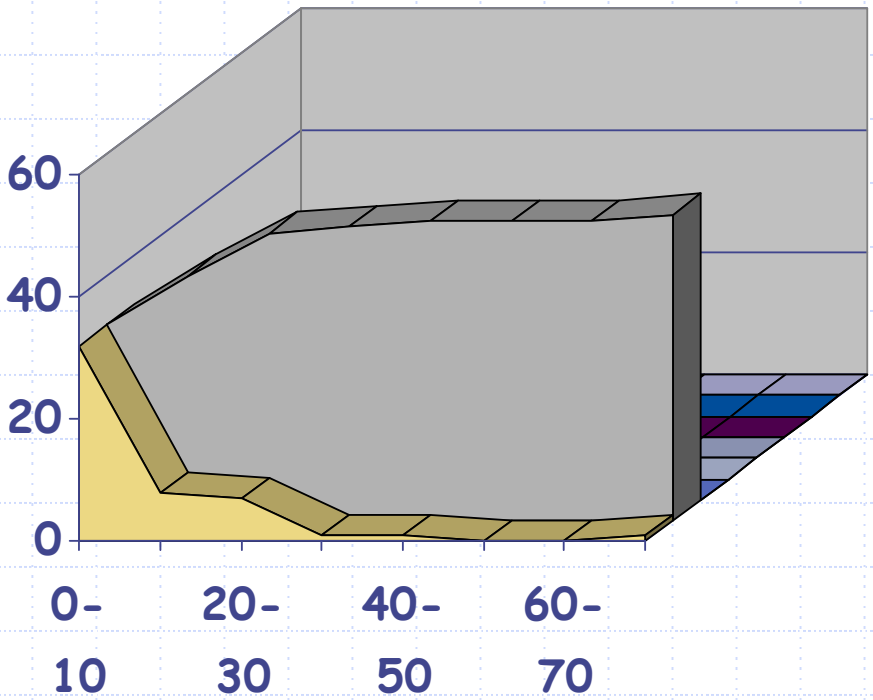


Istogrammi



- Rappresentazioni a barre
- Evidenziano la frequenza su intervalli adiacenti
  - La larghezza di ogni rettangolo misura l'ampiezza degli intervalli
  - Quale larghezza?

Poligoni

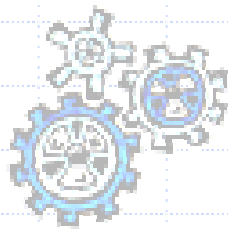


- Per la descrizione di frequenze cumulative
- I punti sono uniti tramite linee

Rappresentazione “Stem & Leaf”

10-19	2 7 5
20-29	9 19 5 3 4 7 1 8
30-39	4 9 2 4 7
40-49	4 8 2
50-59	3

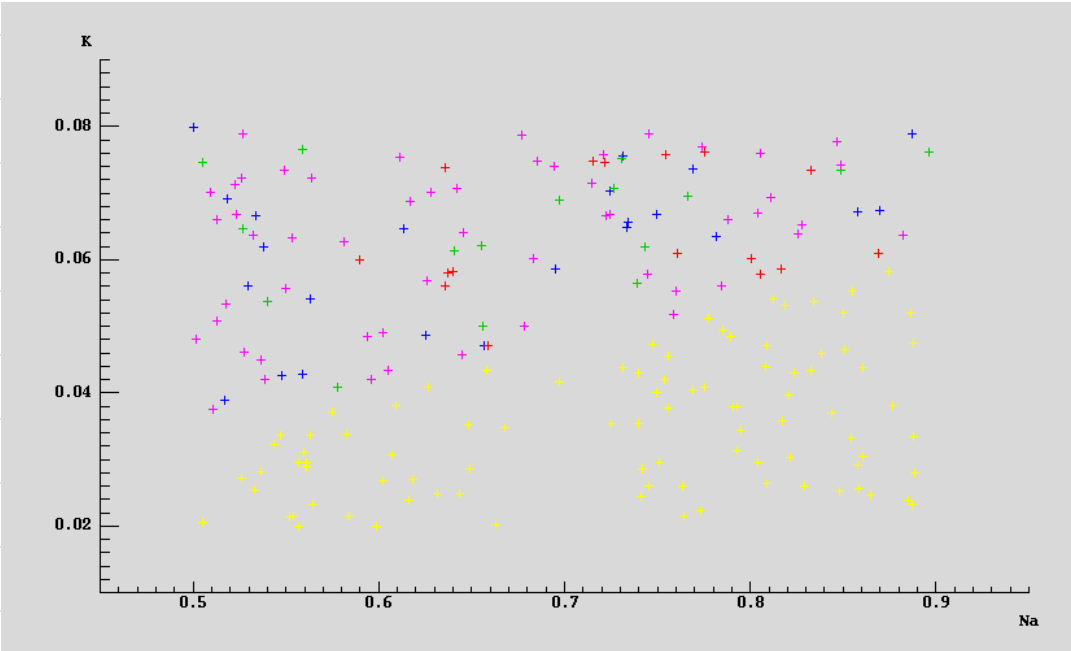
- Simile a istogrammi
- Evita la perdita di informazione
- Utile per pochi dati

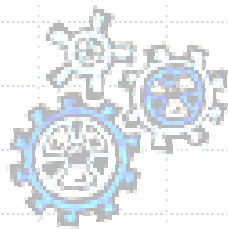


Dot Diagrams, Scatters

Weka

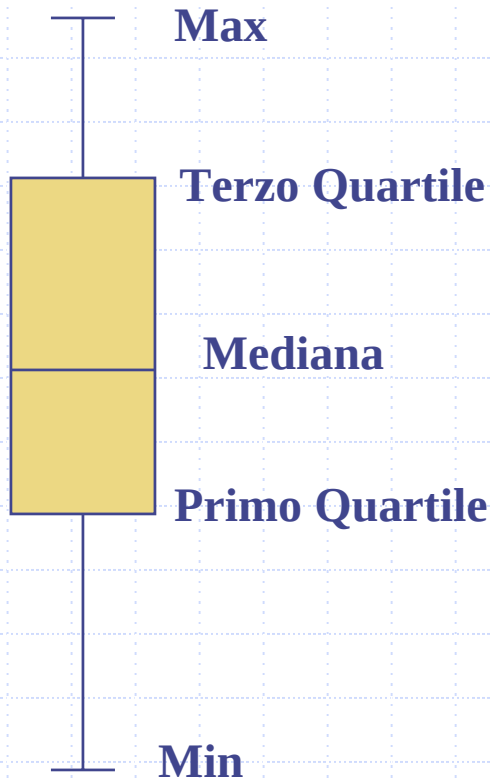
- Visualizza la Dispersione plot dei dat





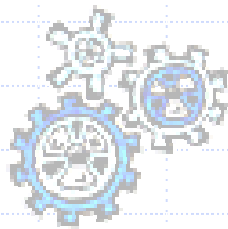
Rappresentazioni Boxplot

- Rappresentano
  - il grado di dispersione o variabilità dei dati (w.r.t. mediana e/o media)
  - la simmetria
  - la presenza di valori anomali
- Le distanze tra i quartili definiscono la dispersione dei dati



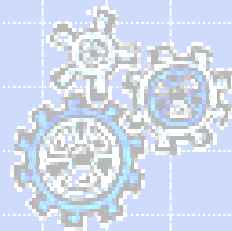
Misure descrittive dei dati

- **Tendenza centrale o posizione**
  - Media aritmetica, geometrica e armonica, mediana, quartili, percentili, moda
- **Dispersione o variabilità**
  - Range, scarto medio, varianza, deviazione standard
- **Forma della distribuzione**
  - Simmetria (medie interquartili, momenti centrali, indice di Fisher) e curtosi (indice di Pearson, coefficiente di curtosi)



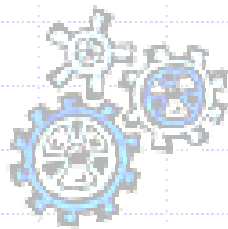
Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning ⇦



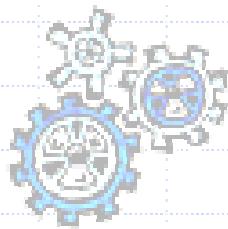
Data Cleaning

- Trattamento di valori anomali
- Trattamento di outliers
- Trattamento di tipi impropri



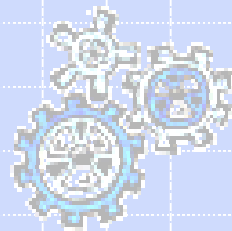
Valori Anomali

- Valori mancanti
  - NULL
- Valori sconosciuti
  - Privi di significato
- Valori non validi
  - Con valore noto ma non significativo



Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction ⇦



Anno accademico, 2004/2005

Trattamento di valori nulli

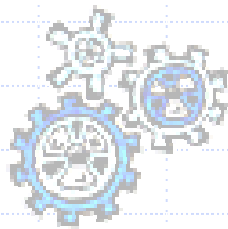
1. Eliminazione delle tuple
2. Sostituzione dei valori nulli

N.B.: può influenzare la distribuzione dei dati numerici

- Utilizzare media/mediana/moda
- Predirre i valori mancanti utilizzando la distribuzione dei valori non nulli
- Segmentare i dati e utilizzare misure statistiche (media/moda/mediana) di ogni segmento
- Segmentare i dati e utilizzare le distribuzioni di probabilità all'interno dei segmenti
- Costruire un modello di classificazione/regressione e utilizzare il modello per calcolare i valori nulli

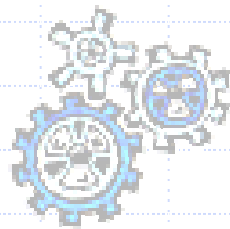
Data Reduction

- Riduzione del volume dei dati
  - Verticale: riduzione numero di tuple
    - Data Sampling
    - Clustering
  - Orizzontale: riduzione numero di colonne
    - Seleziona un sottinsieme di attributi
    - Crea un nuovo (e piccolo) insieme di attributi



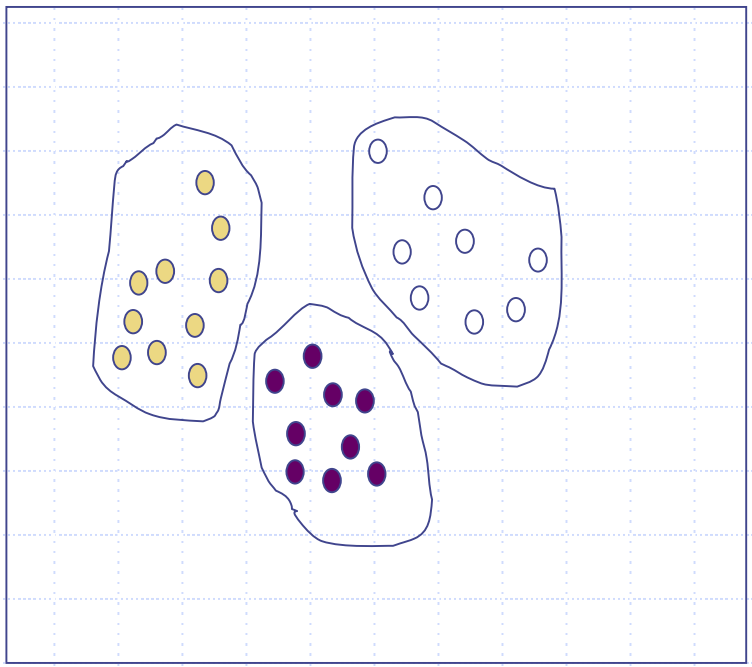
Sampling
(Riduzione verticale)

- Riduce la complessità di esecuzione degli algoritmi di Mining
- Problema: scegliere un sottoinsieme rappresentativo dei dati
  - La scelta di un campionamento casuale può essere problematica per la presenza di picchi
- Alternative: Schemi adattativi
  - **Stratified sampling**:
    - Approssimiamo la percentuale di ogni classe (o sottopopolazione di interesse rispetto all'intero database)
    - Adatto a distribuzioni con picchi: ogni picco è in uno strato
  - Possiamo combinare le tecniche random con la stratificazione
- N.B.: Il Sampling potrebbe non ridurre I tempi di risposta se i dati risiedono su disco (page at a time).

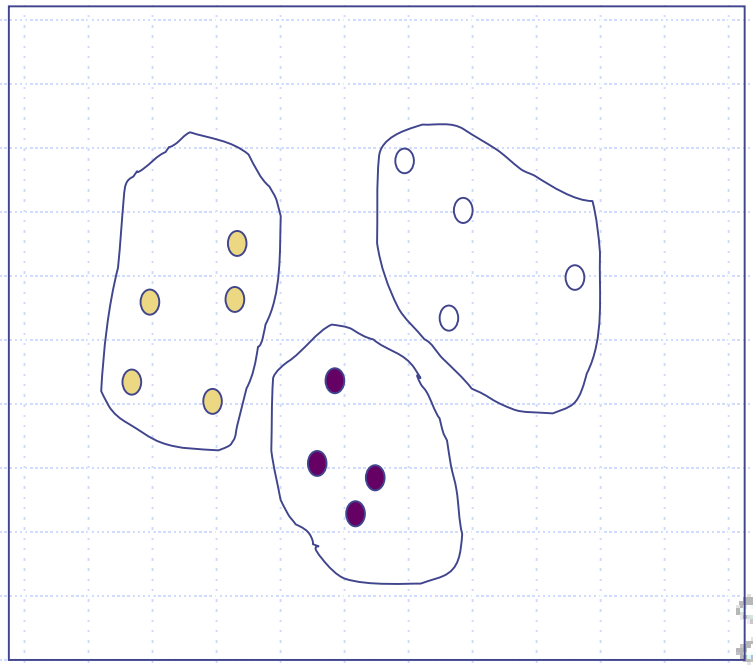


Sampling

Raw Data



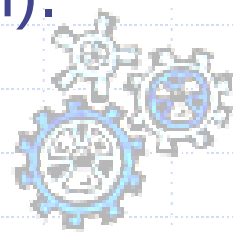
Cluster/Stratified Sample



Riduzione Dimensionalità

(Riduzione orizzontale)

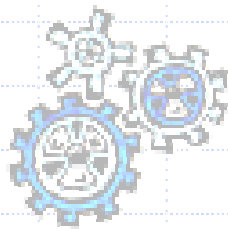
- Selezione di un sotto-insieme di attributi
  - Manuale
    - In seguito a analisi di significatività e/o correlazione con altri attributi
  - Automatico
    - Selezione incrementale degli attributi “migliori”
    - “Migliore” = rispetto a qualche misura di significatività statistica (es.: information gain).



Riduzione Dimensionalità

(Riduzione orizzontale)

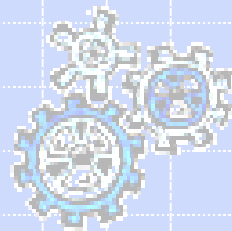
- Creazione di nuovi attributi con i quali rappresentare le tuple
  - Principal components analysis (PCA)
    - Trova le combinazioni lineari degli attributi nei k vettori ortonormali più significativi
    - Proietta le vecchie tuple sui nuovi attributi
  - Altri metodi
    - Factor Analysis
    - Decomposizione SVD



Anno accademico, 2004/2005

Outline del Modulo

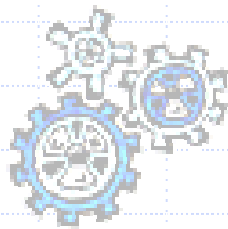
- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation ⇦



Anno accademico, 2004/2005

Data Transformation: Motivazioni

- Dati con errori o incompleti
- Dati mal distribuiti
  - Forte asimmetria nei dati
  - Molti picchi
- La trasformazione dei dati può alleviare questi problemi



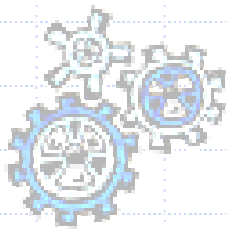
Obiettivi

- Vogliamo definire una trasformazione T sull'attributo X:

$$Y = \mathrm{T}(X)$$

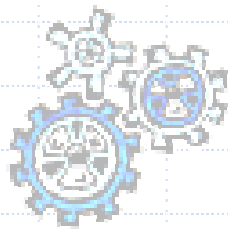
tale che:

- Y preserva l'informazione "rilevante" di X
- Y elimina almeno uno dei problemi di X
- Y è più "utile" di X



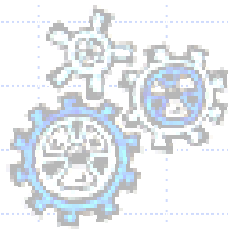
Obiettivi

- Scopi principali:
  - stabilizzare le varianze
  - normalizzare le distribuzioni
  - linearizzare le relazioni tra variabili
- Scopi secondari:
  - semplificare l'elaborazione di dati che presentano caratteristiche non gradite
  - rappresentare i dati in una scala ritenuta più adatta.



Perché normalità, linearità, ecc.?

- Molte metodologie statistiche richiedono correlazioni lineari, distribuzioni normali, assenza di outliers
- Molti algoritmi di Data Mining hanno la capacita` di trattare automaticamente non-linearita' e non-normalita'
  - Gli algoritmi lavorano comunque meglio se tali problemi sono trattati

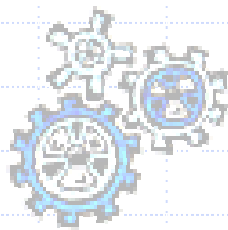


Metodi

- Trasformazioni esponenziali

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

- con a,b,c,d e p valori reali
  - Preservano l'ordine
  - Preservano alcune statistiche di base
  - sono funzioni continue
  - ammettono derivate
  - sono specificate tramite funzioni semplici



Migliorare l'interpretabilita`

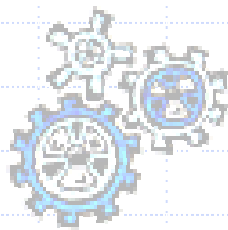
- Trasformazioni lineari

1€ = 1936.27 Lit.

- *p=1, a= 1936.27 ,b =0*

ºC= 5/9(ºF -32)

- *p = 1, a = 5/9, b = -160/9*



Normalizzazioni

- min-max normalization

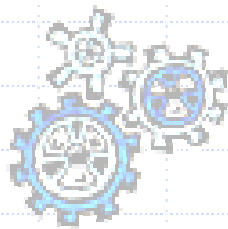
$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization tramite decimal scaling

$$v' = \frac{v}{10^j}$$ dove j è il più piccolo intero tale che Max($|v'|$)<1

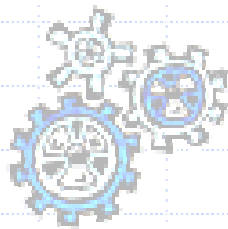


Stabilizzare varianze

- Trasformazione logaritmica

$$T(x) = c \log x + d$$

  - Si applica a valori positivi
  - omogeneizza varianze di distribuzioni lognormali
  - E.g.: normalizza picchi stagionali

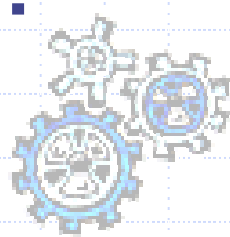


Trasformazione logaritmica: esempio

Bar	*Birra*	*Ricavo*
A	Bud	20
A	Becks	10000
C	Bud	300
D	Bud	400
D	Becks	5
E	Becks	120
E	Bud	120
F	Bud	11000
G	Bud	1300
H	Bud	3200
H	Becks	1000
I	Bud	135

2300	Media
2883,3333	Scarto medio assoluto
3939,8598	Deviazione standard
5	Min
120	Primo Quartile
350	Mediana
1775	Secondo Quartile
11000	Max

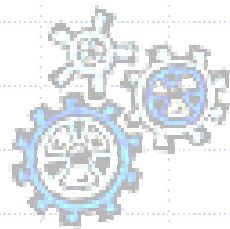
Dati troppo dispersi!!!



Trasformazione Logaritmica: esempio

Bar	*Birra*	*Ricavo (log)*
A	Bud	1,301029996
A	Becks	4
C	Bud	2,477121255
D	Bud	2,602059991
D	Becks	0,698970004
E	Becks	2,079181246
E	Bud	2,079181246
F	Bud	4,041392685
G	Bud	3,113943352
H	Bud	3,505149978
H	Becks	3
I	Bud	2,130333768

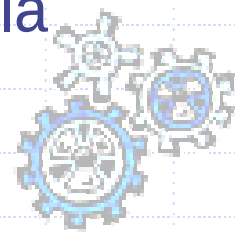
Media	2,585697
Scarto medio assoluto	0,791394
Deviazione standard	1,016144
Min	0,69897
Primo Quartile	2,079181
Mediana	2,539591
Secondo Quartile	3,211745
Max	4,041393



Stabilizzare varianze

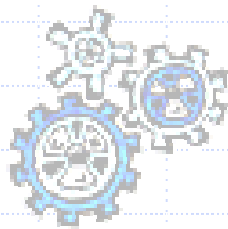
$$T(x) = ax^p + b$$

- Trasformazione in radice
  - $p = 1/c$, c numero intero
  - per omogeneizzare varianze di distribuzioni particolari, e.g., di Poisson
- Trasformazione reciproca
  - $p < 0$
  - per l'analisi di serie temporali, quando la varianza aumenta in modo molto pronunciato rispetto alla media



Simmetria

- Si ha simmetria quando media, moda e mediana coincidono
  - condizione necessaria, non sufficiente
  - Asimmetria sinistra: moda, mediana, media
  - Asimmetria destra: media, mediana, moda



Asimmetria dei dati

- Simmetria e Media interpercentile

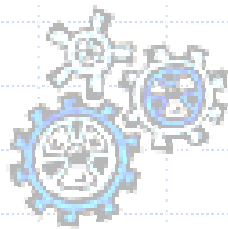
$$M - x_p = x_{1-p} - M \Leftrightarrow \frac{x_{1-p} + x_p}{2} = M$$

- Se la media interpercentile è sbilanciata, allora la distribuzione dei dati è asimmetrica
  - sbilanciata a destra

$$\bar{x}_p > M$$

  - sbilanciata a sinistra

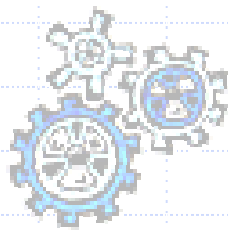
$$\bar{x}_p < M$$



Asimmetria nei dati: esempio

- Verifichiamo la simmetria (valori di un unico attributo)

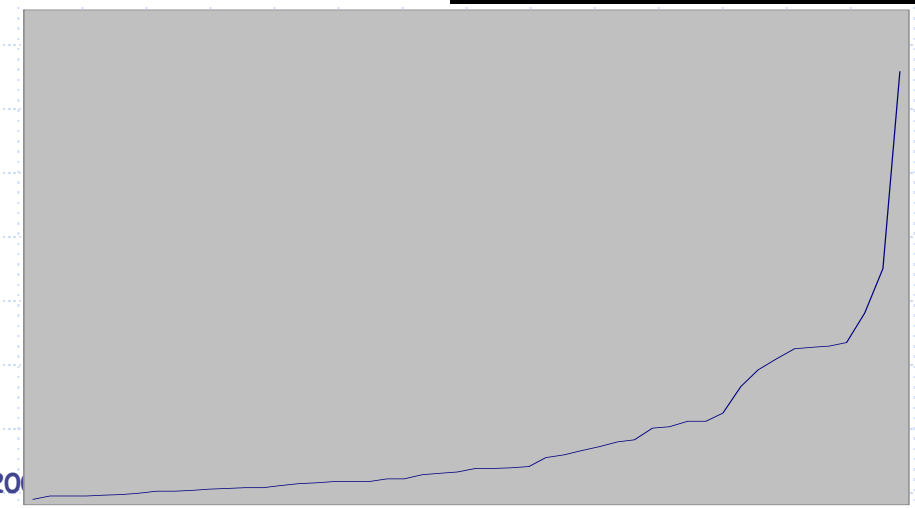
2.808	14.001	4.227	5.913	6.719
3.072	29.508	26.463	1.583	78.811
1.803	3.848	1.643	15.147	8.528
43.003	11.768	28.336	4.191	2.472
24.487	1.892	2.082	5.419	2.487
3.116	2.613	14.211	1.620	21.567
4.201	15.241	6.583	9.853	6.655
2.949	11.440	34.867	4.740	10.563
7.012	9.112	5.732	4.030	28.840
16.723	4.731	3.440	28.608	995

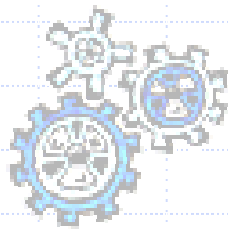


Asimmetria : esempio

- I valori della media interpercentile crescono col percentile considerato
- Distribuzione sbilanciata a destra

Percentile	Media	Low	High
M	6158	6158	6158
F	9002	3278	14726
E	12499	2335	22662
D	15420	2117	28724
C	16722	2155	31288
1	39903	995	78811





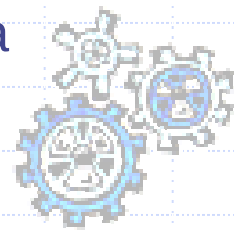
Creare simmetria nei dati: Trasformation plot

- Trovare una trasformazione T_p che crei simmetria
  - Consideriamo i percentili x_U e x_L
  - I valori c ottenuti tramite la formula

$$\frac{x_U + x_L}{2} - M = (1-c)\frac{(x_U - M)^2 + (M - x_L)^2}{4M}$$

suggeriscono dei valori adeguati per p

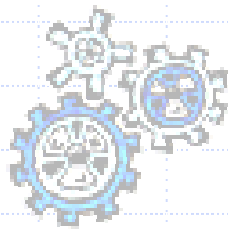
- Intuitivamente, compariamo la differenza assoluta e relativa tra mediana e medie interpercentili
- il valore medio (mediano) dei valori di c è il valore della trasformazione



Trasformation plot: esempio

$(x_L-x_U)/2-M$	$((M- x_L)^2+(x_U -M)^2)/4M$	c
2844.5	3317.5	0.14258
6341	11652.8	0.45583
9262.7	21338.8	0.56592
10564.3	26292.5	0.59820

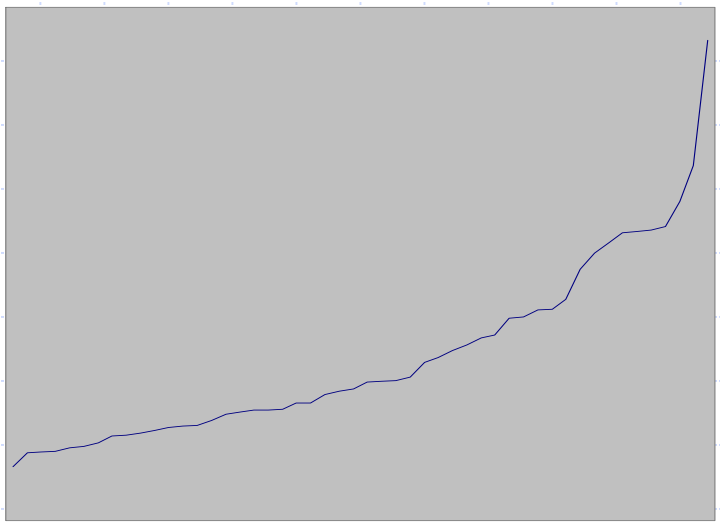
- Calcolando la mediana dei valori *c* otteniamo $p=0.5188$
- Proviamo con $p=1/2$...



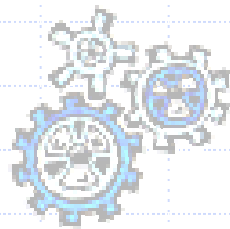
Trasformazione 1: radice quadrata

$$T(x) = \sqrt{x}$$

Percentile	Media	Low	High	
M	78,42283	78,42283	78,42283	0,50000
F	89,28425	57,23633	121,33217	0,25000
E	99,37319	48,27950	150,46688	0,12500
D	107,58229	45,68337	169,48122	0,06250
C	110,87427	45,05801	176,69054	0,03125
1	156,13829	31,54362	280,73297	



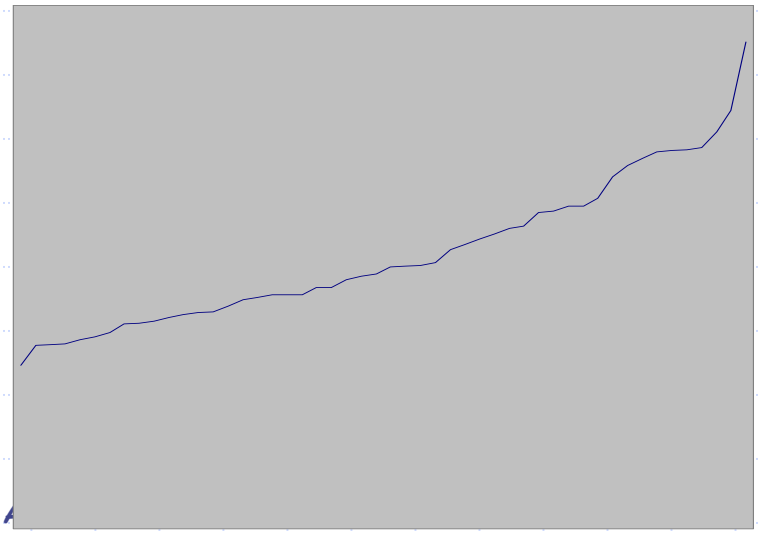
- La curva si tempera,ma i valori alti continuano a produrre differenze notevoli
- Proviamo a diminuire p...



Trasformazione 2: radice quarta

$$T(x) = \sqrt[4]{x}$$

Percentile	Media	Low	High	
M	8,85434	8,85434	8,85434	0,50000
F	9,28978	7,56489	11,01467	0,25000
E	9,60590	6,94676	12,26503	0,12500
D	9,88271	6,74694	13,01849	0,06250
C	9,97298	6,65710	13,28886	0,03125
1	11,18573	5,61637	16,75509	

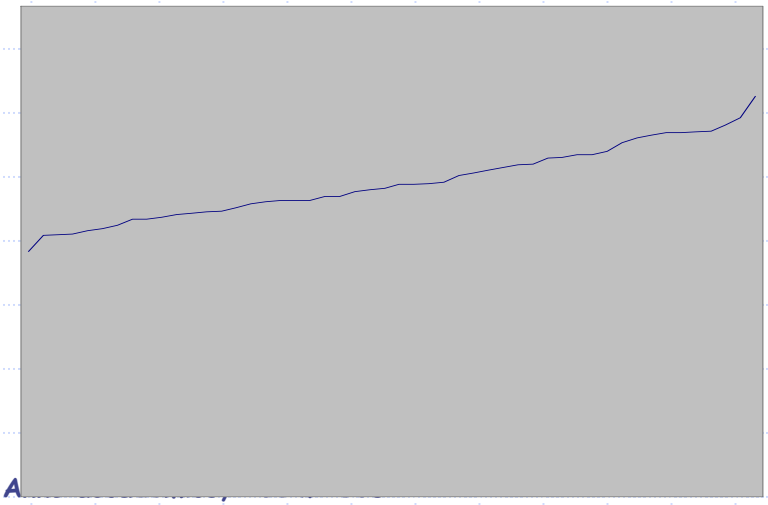


- I valori alti continuano ad influenzare
- Proviamo con il logaritmo…

Trasformazione 3: logaritmo

$$T(x) = \log x$$

Percentile	Media	Low	High	
M	3,78836502	3,78836502	3,78836502	0,50000
F	3,84144850	3,51507795	4,16781905	0,25000
E	3,86059853	3,36672764	4,35446943	0,12500
D	3,88578429	3,31332721	4,45824138	0,06250
C	3,88573156	3,27798502	4,49347811	0,03125
1	3,94720496	2,99782308	4,89658684	



- Abbiamo ottenuto simmetria!

Semplificare le relazioni tra attributi

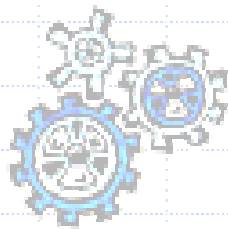
- Esempio: caso della regressione
  - La formula

$$y = \alpha x^{p}$$

puo' essere individuata studiando la relazione

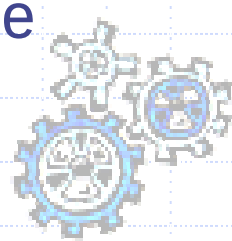
$$z = \log \alpha + pw$$

dove $z = \log y$ e $w = \log x$



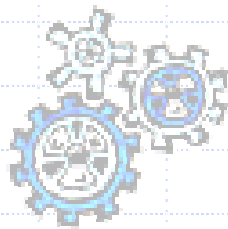
Discretizzazione

- Unsupervised vs. Supervised
- Globale vs. Locale
- Statica vs. Dinamica
- Task difficile
  - Difficile capire a priori qual'è la discretizzazione ottimale
    - bisognerebbe conoscere la distribuzione reale dei dati



Discretizzazione: Vantaggi

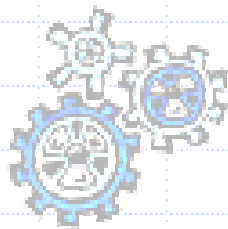
- I dati originali possono avere valori continui estremamente sparsi
- I dati discretizzati possono essere più semplici da interpretare
- Le distribuzioni dei dati discretizzate possono avere una forma “Normale”
- I dati discretizzati possono essere ancora estremamente sparsi
  - Eliminazione della variabile in oggetto



Unsupervised Discretization

- Caratteristiche:
  - Non etichetta le istanze
  - Il numero di classi è noto a priori

- Tecniche di *binning*:
  - **Natural binning** → Intervalli di identica ampiezza
  - **Equal Frequency binning** → Intervalli di identica frequenza
  - **Statistical binning** → Uso di informazioni statistiche (Media, varianza, Quartili)



Natural Binning

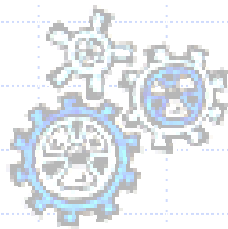
- Semplice
- Ordino i valori, quindi divido il range di valori in *k* parti della stessa dimensione

$$\delta = \frac{x_{\max} - x_{\min}}{k}$$

- l'elemento x_j appartiene alla classe i se

$$x_j \in [x_{min} + i\delta,\ \ x_{min} + (i+1)\delta)$$

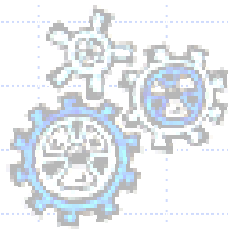
- Puo` produrre distribuzioni molto sbilanciate



Esempio

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- δ =(160-100)/4 = 15
- classe 1: [100,115)
- classe 2: [115,130)
- classe 3: [130,145)
- classe 4: [145, 160]



Equal Frequency Binning

- Ordino e Conto gli elementi, quindi definisco k intervalli di f elementi, dove:

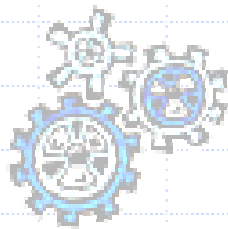
$$f = \frac{N}{k}$$

(N è il numero di elementi del campione)

- l'elemento x_i appartiene alla classe j se

$$j \times f \leq i < (j+1) \times f$$

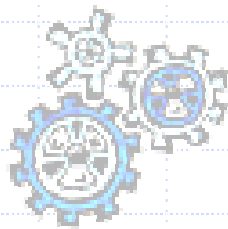
- Non sempre adatta ad evidenziare correlazioni interessanti



Esempio

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- $f = 12/4 = 3$
- classe 1: {100,110,110}
- classe 2: {120,120,125}
- classe 3: {130,130,135}
- classe 4: {140,150,160}



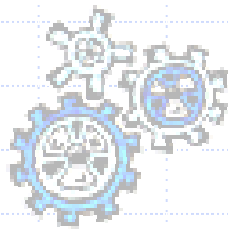
Quante classi?

- Se troppo poche
  => perdita di informazione sulla distribuzione
- Se troppe
  => disperde i valori e non manifesta la foma della distribuzione
- Il numero ottimale C di classi è funzione del numero N di elementi (Sturges, 1929)

$$C = 1 + \frac{10}{3}\log_{10}(N)$$

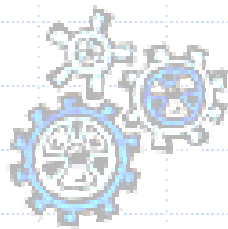
- L'ampiezza ottimale delle classi dipende dalla varianza e dal numero dei dati (Scott, 1979)

$$h = \frac{3{,}5 \cdot s}{\sqrt{N}}$$



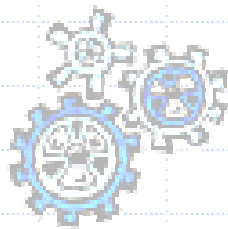
Supervised Discretization

- Caratteristiche:
  - La discretizzazione ha un obiettivo quantificabile
  - Il numero di classi non è noto a priori
- Tecniche:
  - ChiMerge
  - Discretizzazione basata sull'Entropia
  - Discretizzazione basata sui percentili

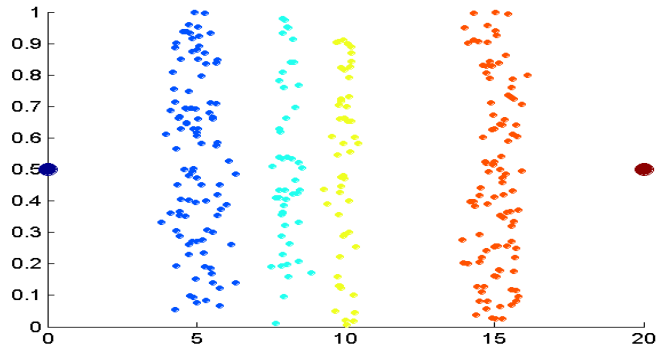


Supervised Discretization: ChiMerge

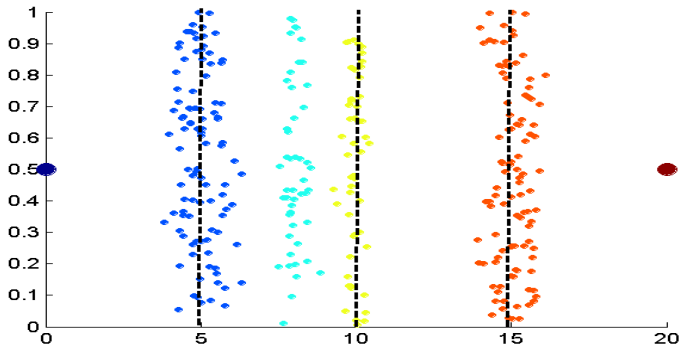
- Procedimento Bottom-up:
  - Inizialmente, ogni valore è un intervallo a se'
  - Intervalli adiacenti sono iterativamente uniti se sono simili
  - La similitudine è misurata sulla base dell'attributo target, contando quanto i due intervalli sono "diversi"



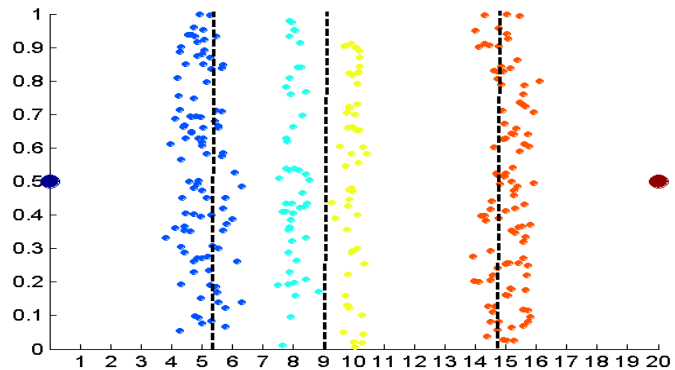
Labels



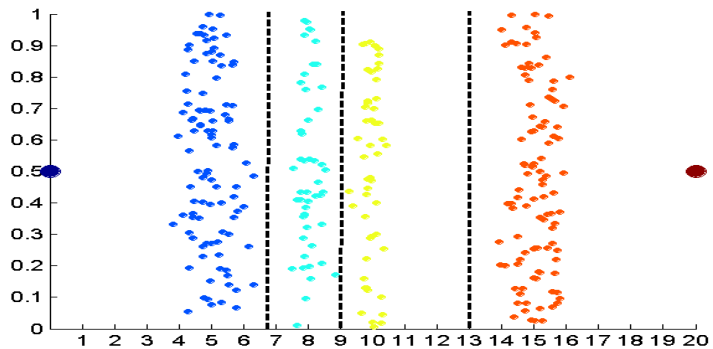
Data



Equal interval width



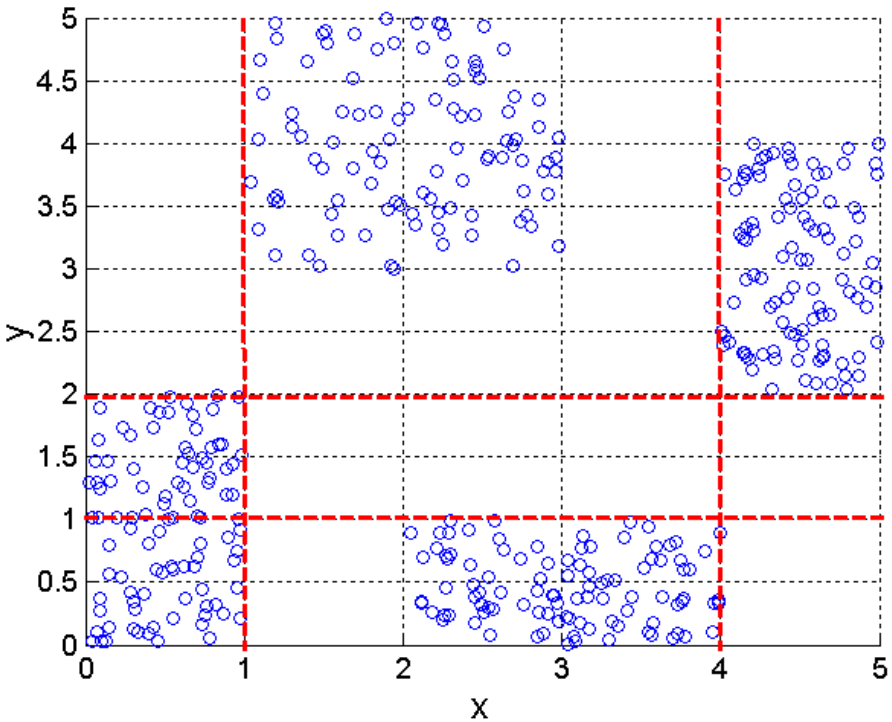
Equal frequency



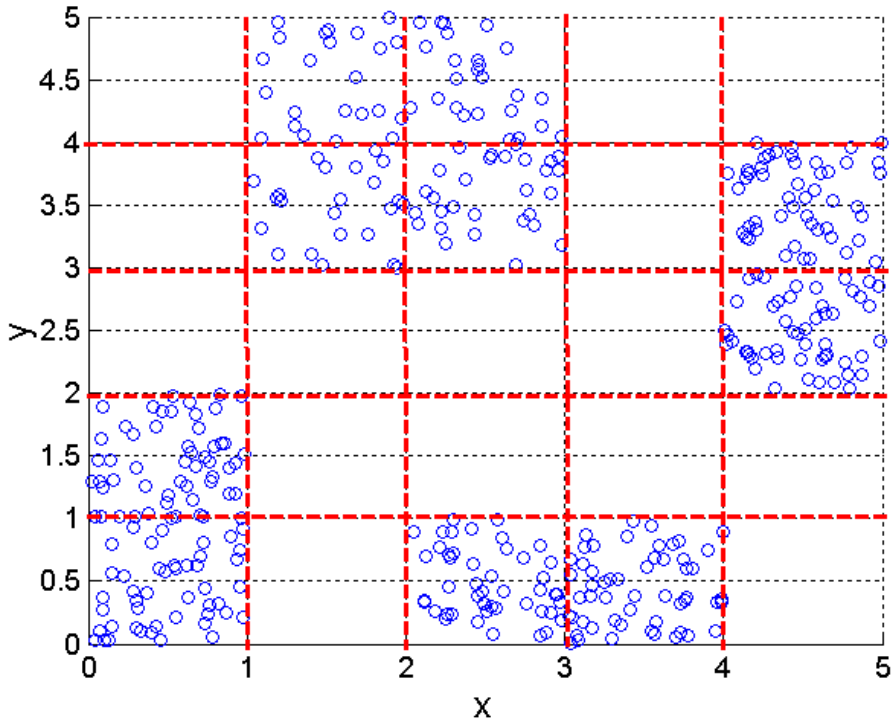
K-means

Discretization Using Class Labels

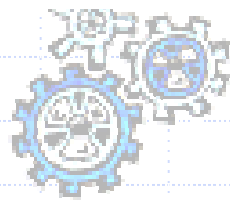
Entropy based approach



3 categories for both x and y

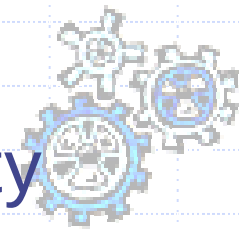


5 categories for both x and y



Similarity and Dissimilarity

- Similarity
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity



Similarity/Dissimilarity for ONE Attribute

- *p* and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

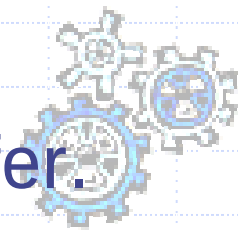
Many attributes: Euclidean Distance

- Euclidean Distance

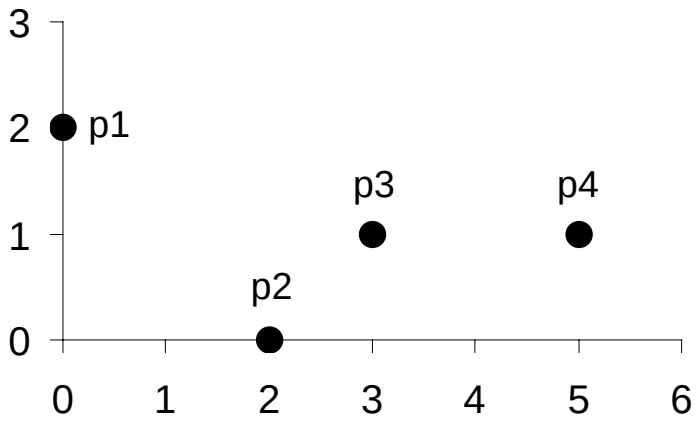
$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the value of kth attributes (components) or data objects p and q.

- Standardization is necessary, if scales differ.



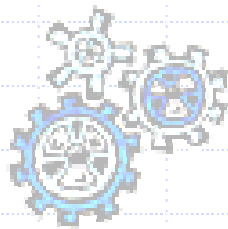
Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

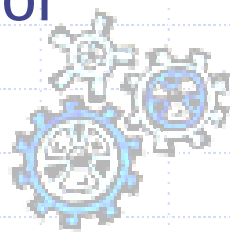


Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

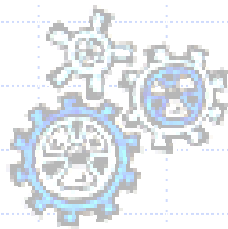
$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) or data objects p and q.



Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.



Minkowski Distance

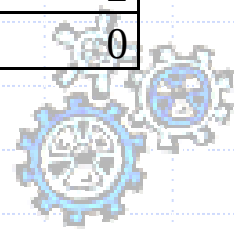
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

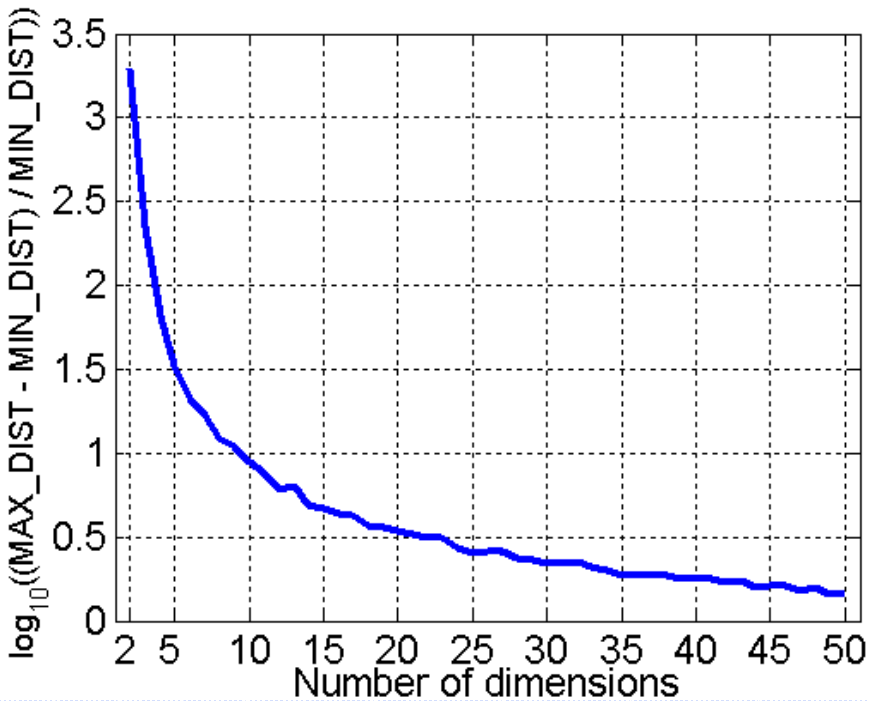
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

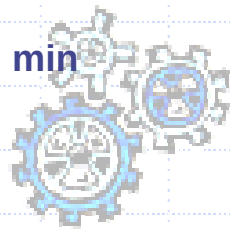


Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful



- **Randomly generate 500 points**
- **Compute difference between max and min distance between any pair of points**



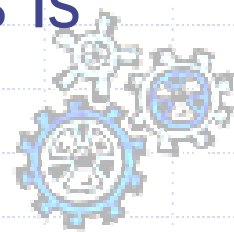
Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q. (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. (Triangle Inequality)

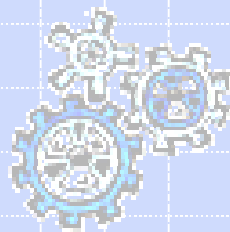
where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is a metric



Outline del Modulo

- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation
- Data similarity



Anno accademico, 2004/2005

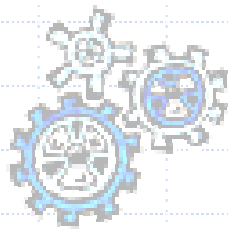
Common Properties of a Similarity

- Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.

2. $s(p, q) = s(q, p)$ for all p and q. (Symmetry)

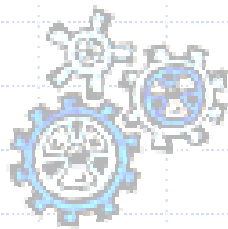
where $s(p, q)$ is the similarity between points (data objects), p and q.



Binary Data

Categorical	insufficient	sufficient	good	very good	excellent
p1	**0**	0	1	0	0
p2	**0**	0	1	0	0
p3	**1**	0	0	0	0
p4	**0**	1	0	0	0

item	bread	butter	milk	apple	tooth-past
p1	**1**	1	0	1	0
p2	**0**	0	1	1	1
p3	**1**	1	1	0	0
p4	**1**	0	1	1	0



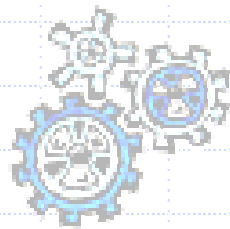
Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1
  M_{10} = the number of attributes where p was 1 and q was 0
  M_{00} = the number of attributes where p was 0 and q was 0
  M_{11} = the number of attributes where p was 1 and q was 1
- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes
  = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values
  = $(M_{11}) / (M_{01} + M_{10} + M_{11})$



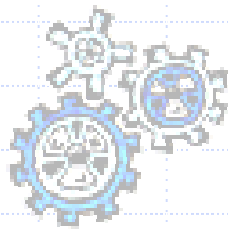
SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0
q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)
M_{10} = 1 (the number of attributes where p was 1 and q was 0)
M_{00} = 7 (the number of attributes where p was 0 and q was 0)
M_{11} = 0 (the number of attributes where p was 1 and q was 1)

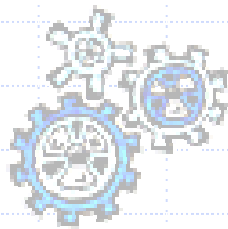
SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0



Document Data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

where $\bullet$ indicates vector dot product and $||d||$ is the length of vector d.

- Example:

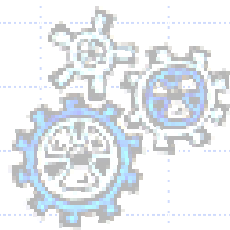
d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{0.5}$ = (42) $^{0.5}$ = 6.481

$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{0.5}$ = (6) $^{0.5}$ = 2.245

$\cos(d_1, d_2) = .3150$



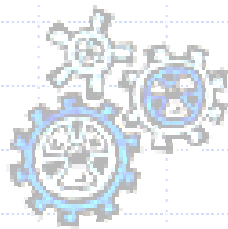
Correlation

- Correlation measures the linear relationship between objects (binary or continuos)
- To compute correlation, we standardize data objects, p and q, and then take their dot product (covariance/standard deviation)

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$



General Approach for Combining Similarities

- Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k, in the range $[0, 1]$.
2. Define an indicator variable, δ_k, for the k_{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

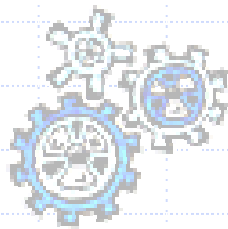
$$similarity(p, q) = \frac{\sum_{k=1}^{n} \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use weights w_k which are between 0 and 1 and sum to 1.

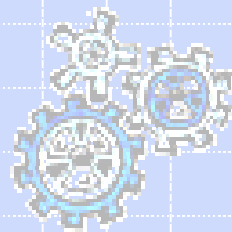
$$similarity(p,q) = \frac{\sum_{k=1}^{n} w_k \delta_k s_k}{\sum_{k=1}^{n} \delta_k}$$

$$distance(p,q) = \left(\sum_{k=1}^{n} w_k |p_k - q_k|^r \right)^{1/r}$$



Outline del Modulo

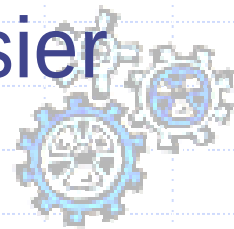
- Introduzione e Concetti di Base
- Data Selection
- Information Gathering
- Data cleaning
- Data reduction
- Data transformation
- Data similarity
- Data Exploration (Multidimensional array)



Anno accademico, 2004/2005

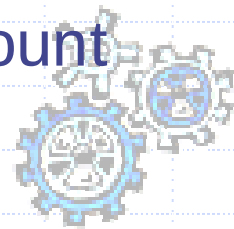
OLAP

- On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database.
- Relational databases put data into tables, while OLAP uses a multidimensional array representation.
  - Such representations of data previously existed in statistics and other fields
- There are a number of data analysis and data exploration operations that are easier with such a data representation.



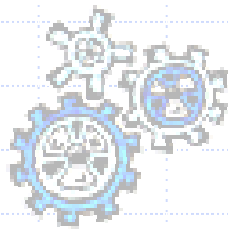
Creating a Multidimensional Array

- Two key steps in converting tabular data into a multidimensional array.
  - First, identify which attributes are to be the **dimensions** and which attribute is to be the target attribute (**Measure**) whose values appear as entries in the multidimensional array.
    - The attributes used as dimensions must have discrete values
    - The target value is typically a count or continuous value, e.g., the cost of an item
    - Can have no target variable at all except the count of objects that have the same set of attribute values



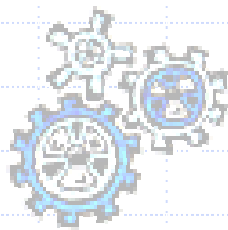
Creating a Multidimensional Array

- Second, find the value of each entry in the multidimensional array by summing the values (of the target attribute) or count of all objects that have the attribute values corresponding to that entry.



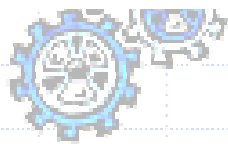
Example: Iris data

- The attributes, petal length, petal width, and species type can be converted to a multidimensional array
  - First, we discretized the petal width and length to have categorical values: *low*, *medium*, and *high*
  - petal length, petal width, and species type are the dimensions
  - Count is the measure



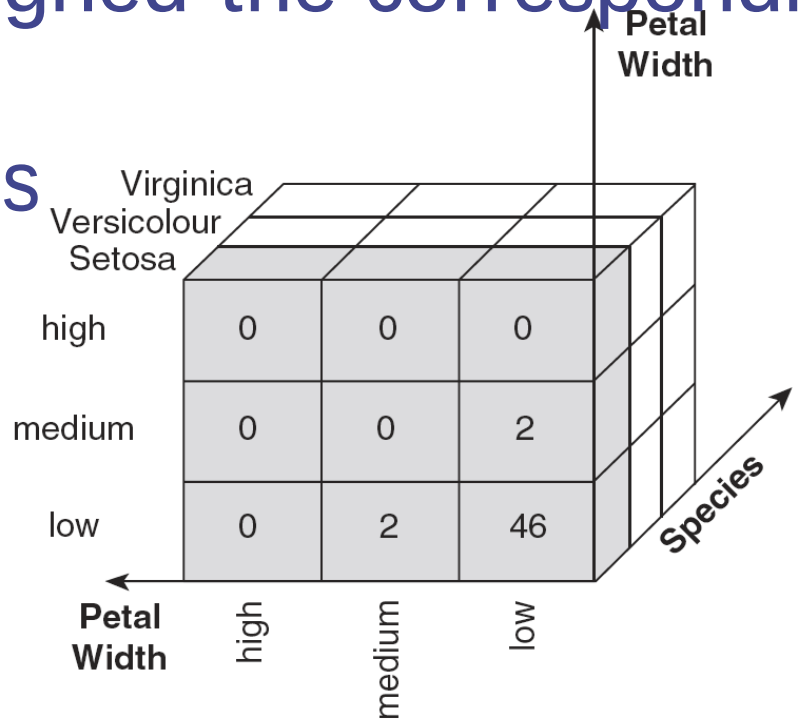
Example: Iris data

Petal Length	Petal Width	Species Type	Count
low	low	Setosa	46
low	medium	Setosa	2
medium	low	Setosa	2
medium	medium	Versicolour	43
medium	high	Versicolour	3
medium	high	Virginica	3
high	medium	Versicolour	2
high	medium	Virginica	3
high	high	Versicolour	2
high	high	Virginica	44



Example: Iris data (continued)

- Each unique tuple of petal width, petal length, and species type identifies one element of the array.
- This element is assigned the corresponding count value.
- The figure illustrates the result.
- All non-specified tuples are 0.

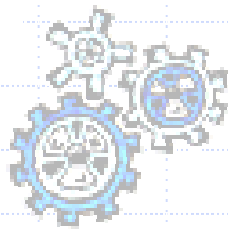


- Slices of the multidimensional array are shown by the following cross-tabulations
- What do these tables tell us?

		Width		
		low	medium	high
Length	low	46	2	0
	medium	2	0	0
	high	0	0	0

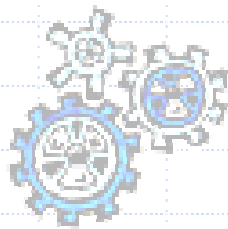
		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	43	3
	high	0	2	2

		Width		
		low	medium	high
Length	low	0	0	0
	medium	0	0	3
	high	0	3	44



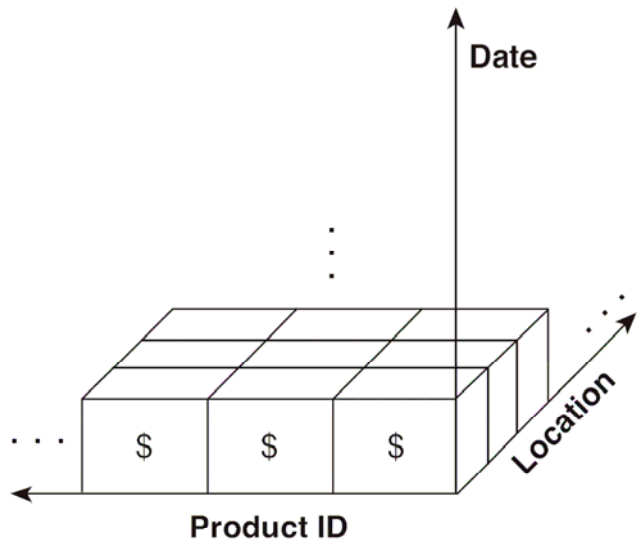
OLAP Operations: Data Cube

- The key operation of a OLAP is the formation of a data cube
- A data cube is a multidimensional representation of data, together with all possible aggregates.
- By all possible aggregates, we mean the aggregates that result by selecting a proper subset of the dimensions and summing over all remaining dimensions.
- For example, if we choose the species type dimension of the Iris data and sum over all other dimensions, the result will be a one-dimensional entry with three entries, each of which gives the number of flowers of each type.



Data Cube Example

- Consider a data set that records the sales of products at a number of company stores at various dates.
- This data can be represent as a 3 dimensional array
- There are 3 two-dimension aggregates (3 choose 2), 3 one-dimensional aggrega and 1 zero-dimensional aggregate (the overall total)

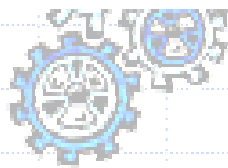


Anno accademico, 2004/2005

Data Cube Example

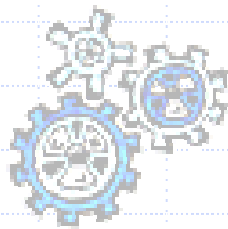
- The following figure table shows one of the two dimensional aggregates, along with two of the one-dimensional aggregates, and the

product ID \ date	Jan 1, 2004	Jan 2, 2004	...	Dec 31, 2004	total
1	$1,001	$987	...	$891	$370,000
⋮	⋮			⋮	⋮
27	$10,265	$10,225	...	$9,325	$3,800,020
⋮	⋮			⋮	⋮
total	$527,362	$532,953	...	$631,221	$227,352,127



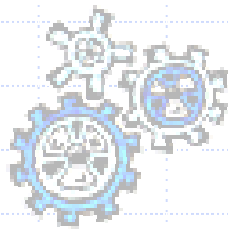
OLAP Operations: Slicing and Dicing

- Slicing is selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions.
- Dicing involves selecting a subset of cells by specifying a range of attribute values.
  - This is equivalent to defining a subarray from the complete array.
- In practice, both operations can also be accompanied by aggregation over some dimensions.



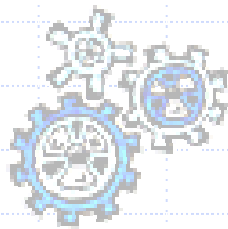
OLAP Operations: Roll-up and Drill-down

- Attribute values often have a hierarchical structure.
  - Each date is associated with a year, month, and week.
  - A location is associated with a continent, country, state (province, etc.), and city.
  - Products can be divided into various categories, such as clothing, electronics, and furniture.
- Note that these categories often nest and form a tree or lattice
  - A year contains months which contains day
  - A country contains a state which contains a city



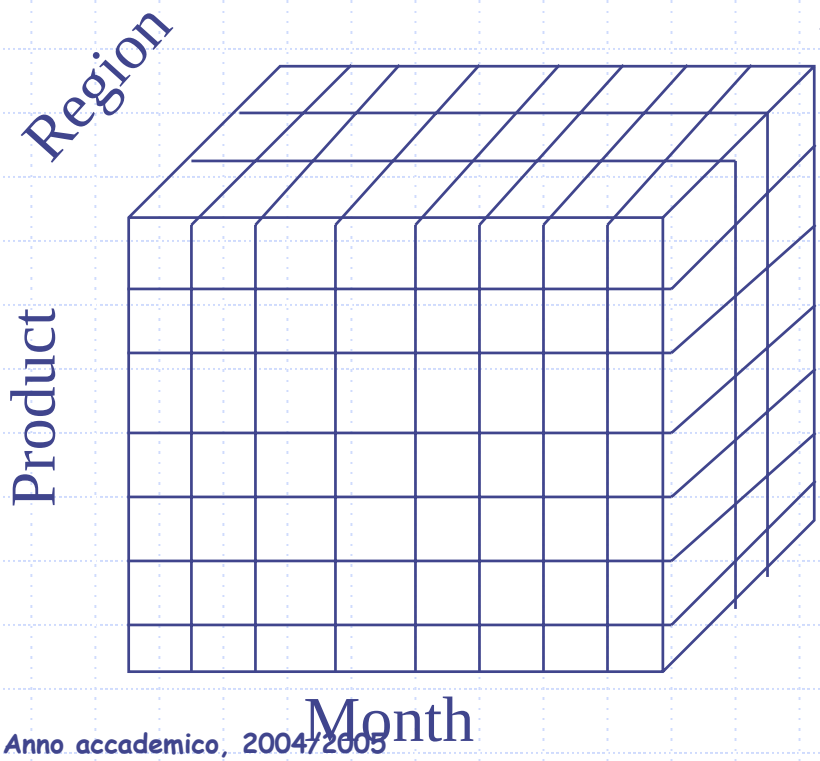
OLAP Operations: Roll-up and Drill-down

- This hierarchical structure gives rise to the roll-up and drill-down operations.
  - For sales data, we can aggregate (roll up) the sales across all the dates in a month.
  - Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down) into daily sales totals.
  - Likewise, we can drill down or roll up on the location or product ID attributes.

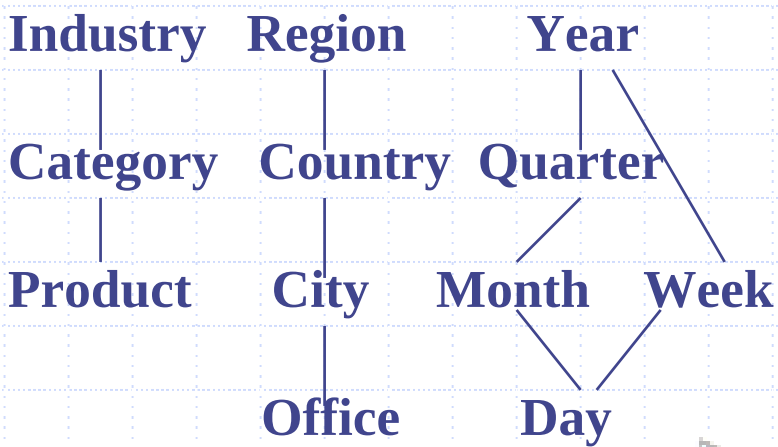


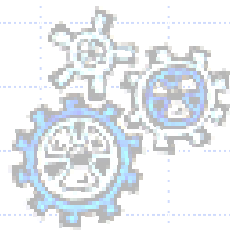
Multidimensional Data

- Sales volume as a function of product, month, and region

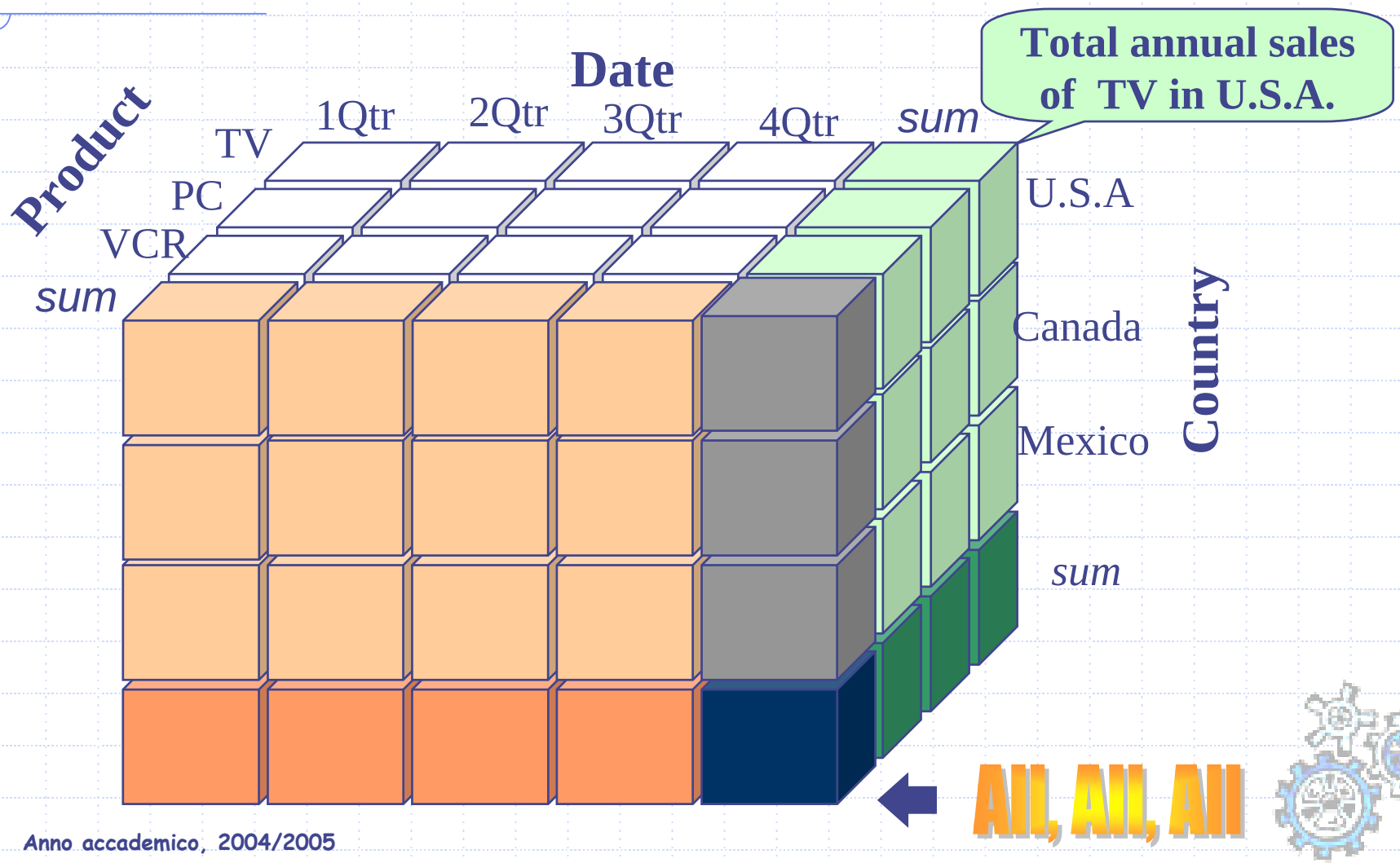


Dimensions: Product, Location, Time
Hierarchical summarization paths



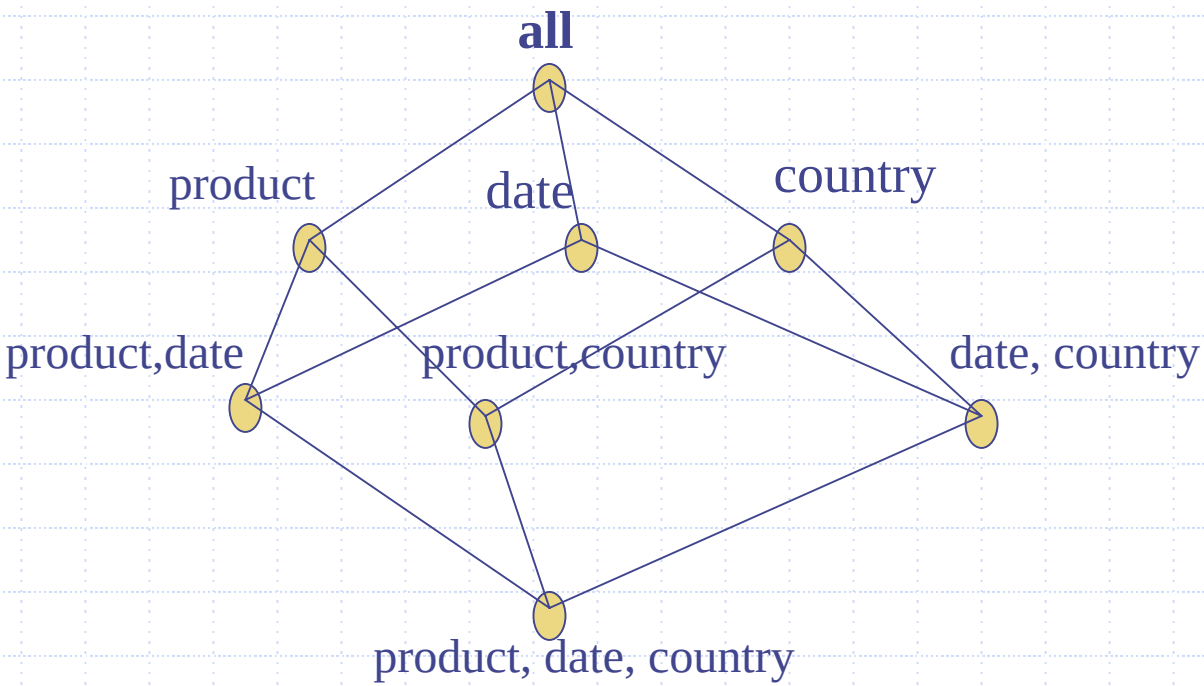


A Sample Data Cube



Anno accademico, 2004/2005

Cuboids Corresponding to the Cube

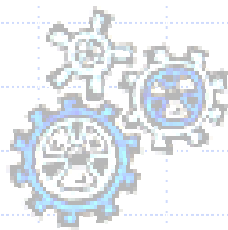


0-D(apex) cuboid

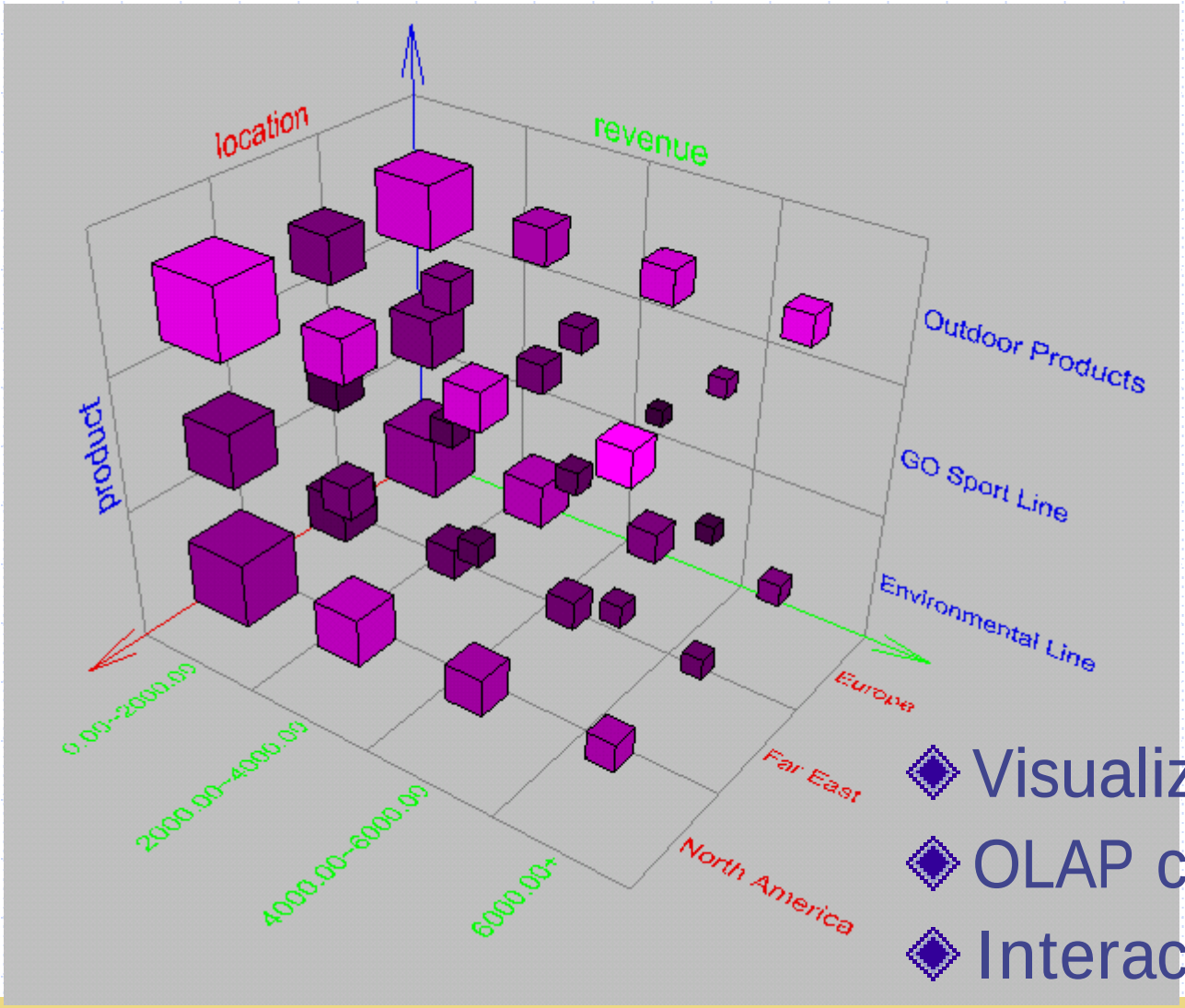
1-D cuboids

2-D cuboids

3-D(base) cuboid



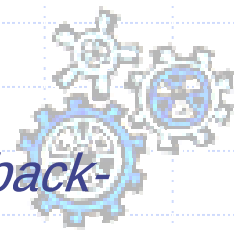
Browsing a Data Cube



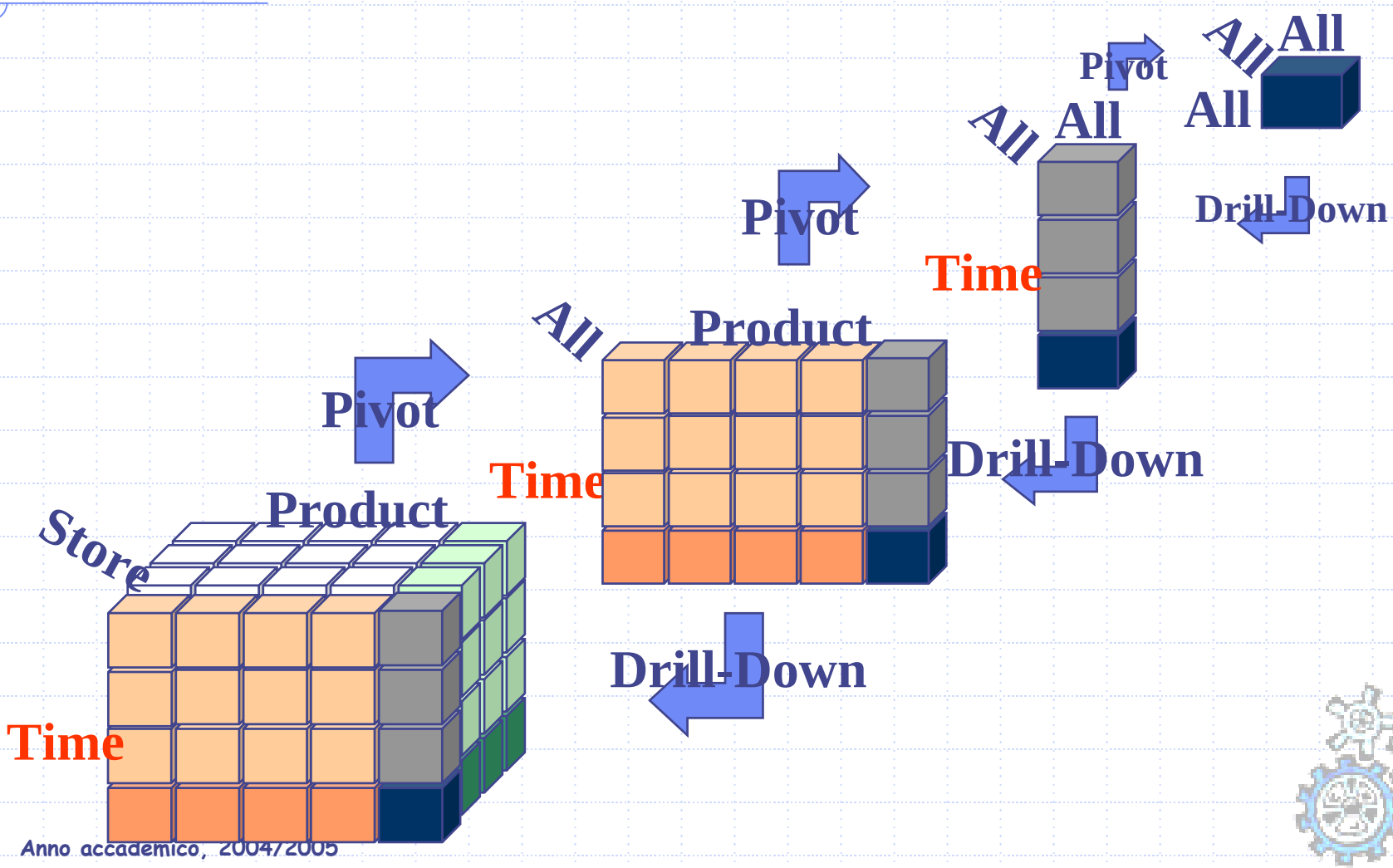
- Visualization
- OLAP capabilities
- Interactive manipulation

Typical OLAP Operations

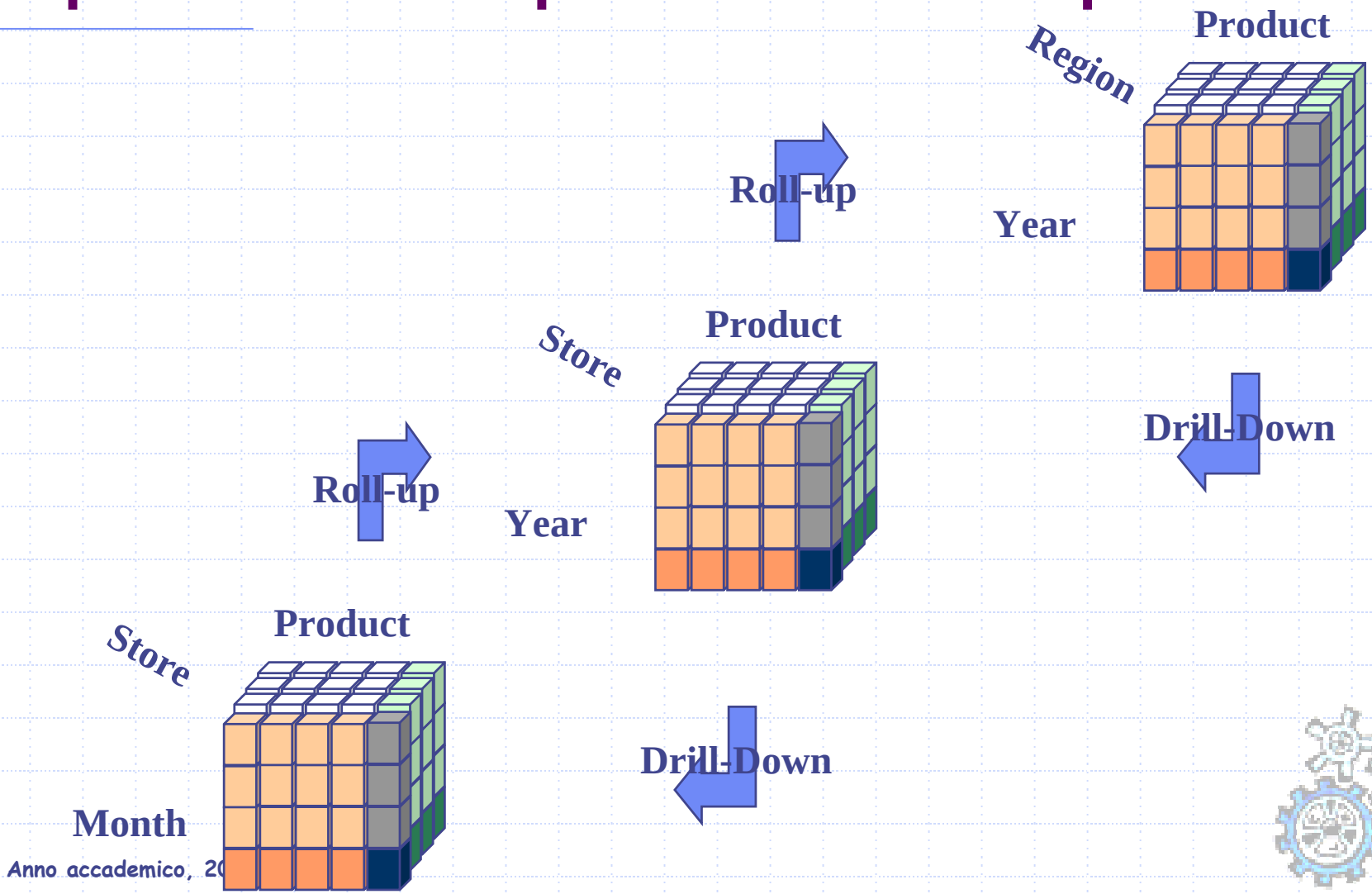
- Roll up (drill-up): summarize data
  - *by climbing up hierarchy or by dimension reduction*
- Drill down (roll down): reverse of roll-up
  - *from higher level summary to lower level summary or detailed data, or introducing new dimensions*
- Slice and dice:
  - *project and select*
- Pivot (rotate):
  - *reorient the cube, visualization, 3D to series of 2D planes.*
- Other operations
  - *drill across: involving (across) more than one fact table*
  - *drill through: through the bottom level of the cube to its back-end relational tables (using SQL)*



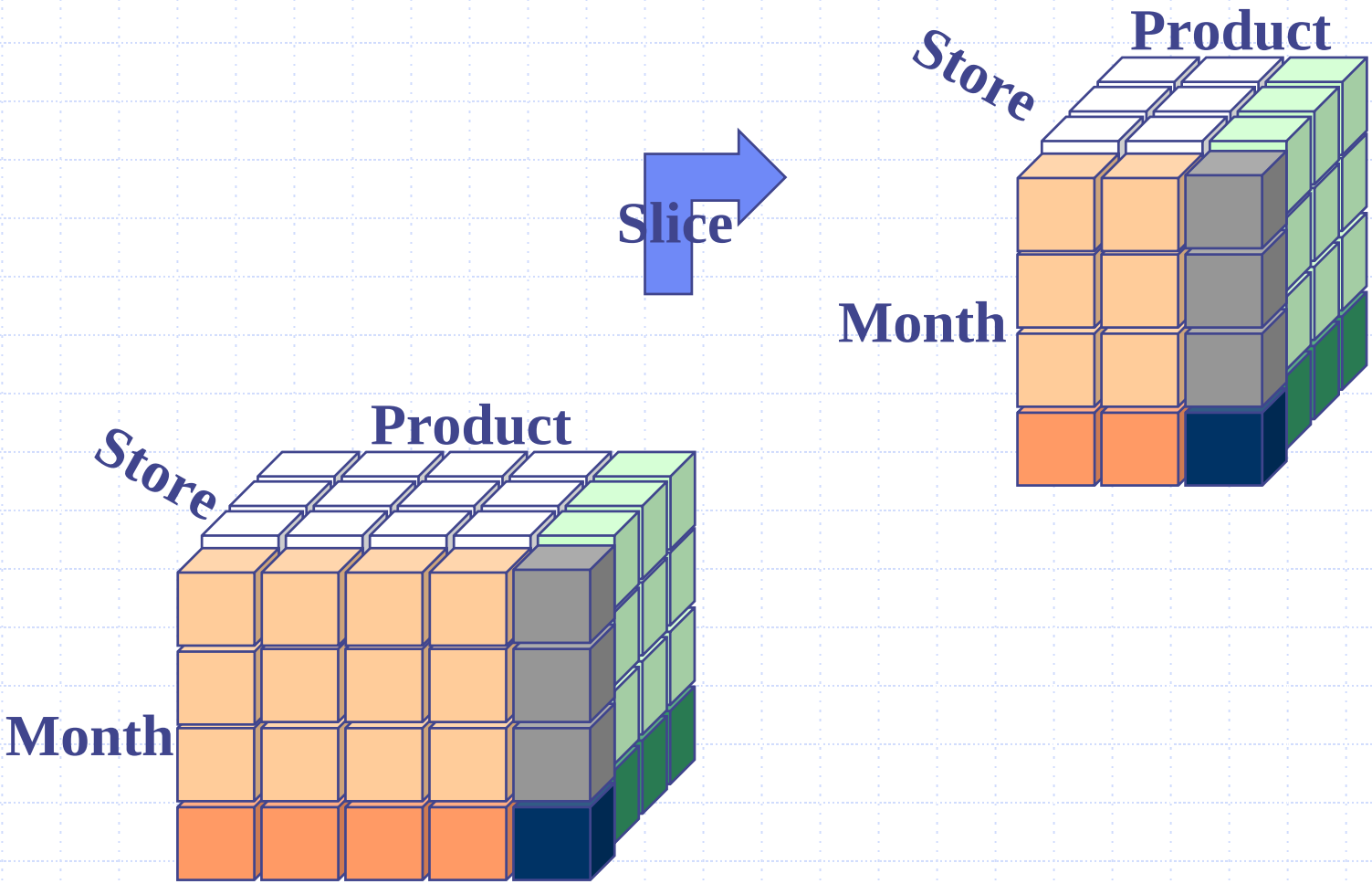
Operazioni tipiche: Pivot



Operazioni tipiche: Roll-Up

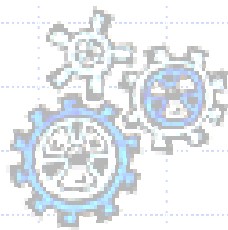


Operazioni tipiche: Slice and Dice



ChiMerge: criterio di similitudine

- Basato sul test del Chi quadro
- k = numero di valori differenti dell'attributo target
- A_{ij} = numero di casi della j-esima classe nell'i-esimo intervallo
- R_i = numero di casi nell'i-esimo intervallo ($\sum_{j=1}^{k} A_{ij}$)
- C_j = numero di casi nella j-esima classe ($\sum_{i=1}^{2} A_{ij}$)
- E_{ij} = frequenza attesa di A_{ij} $(R_i * C_j / N)$

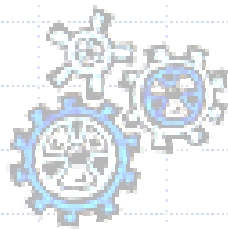


Test del Chi Quadro per la discretizzazione

	1	2	…	K	Total
1	$\mathbf{A_{11}}$	$\mathbf{A_{12}}$	…	$\mathbf{A_{1k}}$	$\mathbf{R_1}$
2	$\mathbf{A_{21}}$	$\mathbf{A_{22}}$	…	$\mathbf{A_{2k}}$	$\mathbf{R_2}$
Total	$\mathbf{C_1}$	$\mathbf{C_2}$	…	$\mathbf{C_k}$	**N**

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

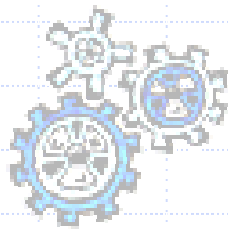
- Si individua quanto “distinti” sono due intervalli
- k-1 gradi di liberta`
- La significativita` del test è data da un threshold δ
  - Probabilita` che l’intervallo in questione e la classe siano indipendenti



Esempio

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

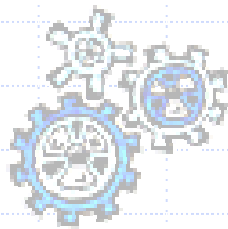
- Discretizzazione w.r.t. Beer
- threshold 50% confidenza
- Vogliamo ottenere una discretizzazione del prezzo che permetta di mantenere omogeneita` w.r.t. Beer



Esempio: Chi Values

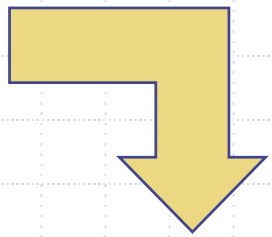
	Bud	Becks
100	1	0
110	2	0
120	1	1
125	1	0
130	2	0
135	1	0
140	0	1
150	0	1
160	0	1

Scegliamo gli elementi adiacenti con Chi-Value minimo



Esempio: passo 1

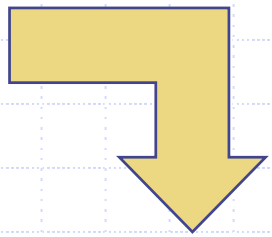
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150	0	1	0
160	0	1	1.38629



	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150-160	0	2	1.38629

Esempio: passo 2

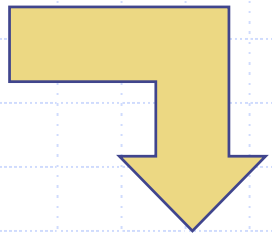
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	2
140	0	1	0
150-160	0	2	1.38629



	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	4
140-150-160	0	3	1.38629

Esempio: passo 3

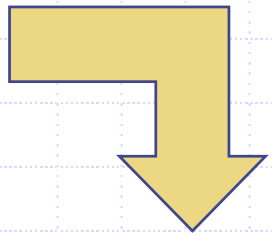
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130	2	0	0
135	1	0	4
140-150-160	0	3	1.38629



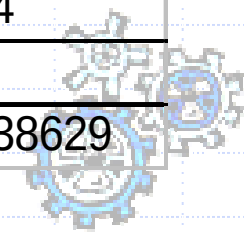
	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130-135	3	0	6
140-150-160	0	3	1.38629

Esempio: passo 4

	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	0.75
125	1	0	0
130-135	3	0	6
140-150-160	0	3	1.38629

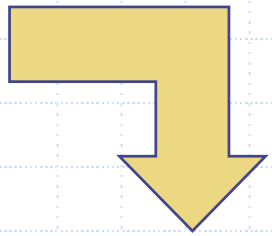


	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	2.4
125-130-135	4	0	7
140-150-160	0	3	1.38629



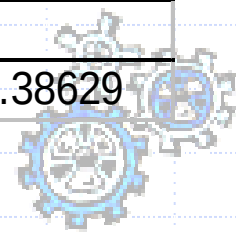
Esempio: passo 5

	Bud	Becks	Chi Value
100	1	0	0
110	2	0	1.33333
120	1	1	2.4
125-130-135	4	0	7
140-150-160	0	3	1.38629



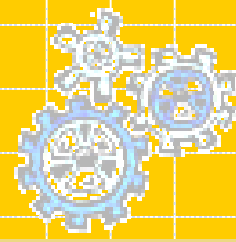
Tutti i valori sono oltre il 50% di confidenza (1.38)

	Bud	Becks	Chi Value
100-110	3	0	1.875
120	1	1	2.4
125-130-135	4	0	7
140-150-160	0	3	1.38629



Esercitazione Clementine

- Allegato Esercitazioni
  - Esercitazione 2



Anno accademico, 2004/2005

Appendice

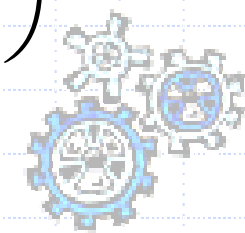
Misure descrittive dei dati

Media Aritmetica

- Per effettuare la correzione di errori accidentali
  - permette di sostituire i valori di ogni elemento senza cambiare il totale
    - Sostituzione di valori NULL
- Monotona crescente

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{1}{n+k}\left(\sum_{i=1}^{n} x_i + k\bar{x}\right) = \bar{x}$$



Media Geometrica

$$x_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

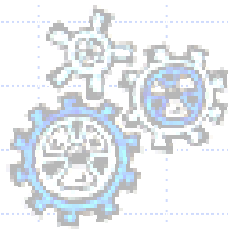
- Per bilanciare proporzioni
- dati moltiplicativi

Prodotto	***Variazioni Prezzi***	
	1996	*1997*
A	100	200
B	100	50
Media	100	125

$$x_g = 100$$

- La media aritmetica dei logaritmi è il logaritmo della media geometrica
- Monotona crescente

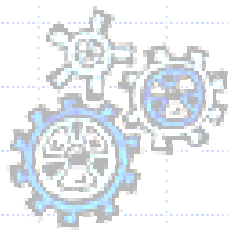
$$\log x_g = \frac{1}{n} \sum_{i=1}^{n} \log x_i$$



Media Armonica

- Monotona decrescente
- Per misure su dimensioni fisiche
- E.g., serie temporali

$$x_a = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$



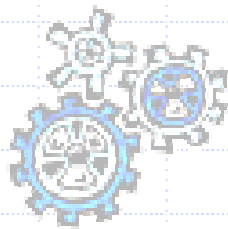
Mediana

- Il valore centrale in un insieme ordinato di dati
- Robusta
  - poco influenzata dalla presenza di dati anomali

1 7 12 18 23 34 54

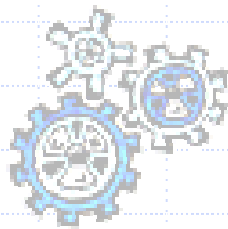
$$\bar{x} = 21.3$$

$$M = 23$$



Mediana e Quartili

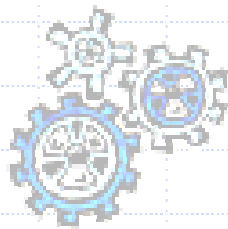
- Divide un insieme di dati a meta`
  - statistica robusta (non influenzata da valori con rilevanti differenze)
  - ulteriori punti di divisione
- interquartili
  - mediane degli intervalli dei dati superiore e inferiore
  - Un quarto dei dati osservati è sopra/sotto il quartile
- percentili
  - di grado p: il p% dei dati osservati è sopra/sotto il percentile
  - mediana: 50-esimo percentile
  - primo quartile: 25-esimo percentile
  - secondo quartile: 75-esimo percentile
- max, min
  - range = max-min



Percentili

- Rappresentati con x_p
- Utilizziamo le lettere per esprimerli

Etichetta	***P***
M	½= 0.5
F	¼=0.25
E	1/8=.125
D	1/16=0.625
C	1/32=0.3125
B	1/64
A	1/128
Z	1/256
Y	1/512
X	1/1024



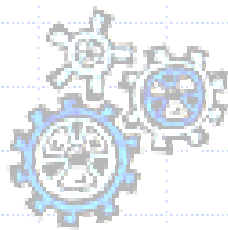
Moda

- Misura della frequenza dei dati

$$a\ a\ b\ b\ c\ c\ a\ d\ b\ c\ a\ e\ c\ b\ a\ a$$

moda = a ($f = 6$)

- Significativo per dati categorici
- Non risente di picchi
- Molto instabile



Range, Deviazione media

- Intervallo di variazione

$$r = max\text{-}min$$

- Scarti interquantili

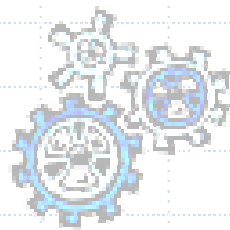
$$r_p = x_{100-p} - x_p$$

- Scarto medio assoluto

$$S_n = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}\right|$$

- Scarto medio assoluto dalla mediana
  - *In generale,* $S_{.5} \leq S_n$

$$S_M = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - M\right|$$



Varianza, deviazione standard

- misure di mutua variabilità tra i dati di una serie
- Devianza empirica

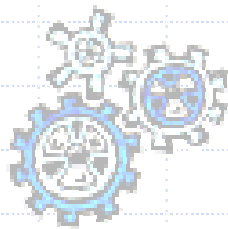
$$dev = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Varianza

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

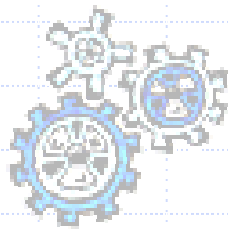
- Coefficiente di variazione
  - misura relativa

$$V = \frac{s}{\bar{x}}$$



Simmetria

- Si ha simmetria quando media, moda e mediana coincidono
  - condizione necessaria, non sufficiente
  - Asimmetria sinistra: moda, mediana, media
  - Asimmetria destra: media, mediana, moda



Simmetria (Cont.)

- Indici di asimmetria
  - medie interquartili

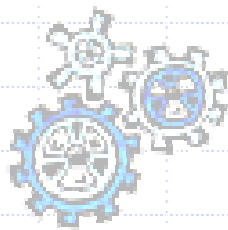
$$\bar{x}_p = (x_{1-p} + x_p)/2$$

  - Momenti centrali

$$m_k = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^k$$

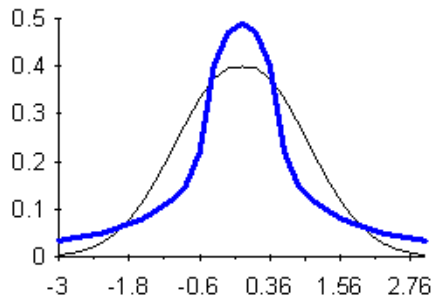
- indice di Fisher
  - γ nullo per distribuzioni simmetriche
  - $\gamma > 0$: sbilanciamenti a destra
  - $\gamma < 0$: sbilanciamento a sinistra

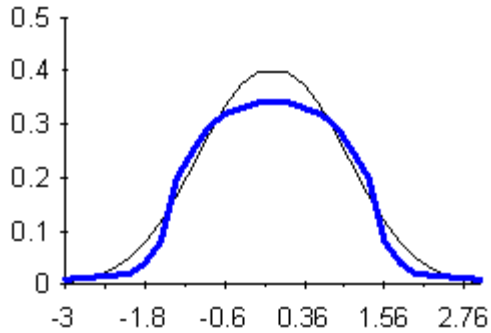
$$\gamma = \frac{m_3}{\hat{s}^3}$$



Curtosi

- Grado di appiattimento della curva di distribuzione rispetto alla curva normale
  - mesocurtica: forma uguale alla distribuzione normale;
  - leptocurtica:una frequenza minore delle classi intermedie, frequenza maggiore delle classi estreme e dei valori centrali;
  - platicurtica:una frequenza minore delle classi centrali e di quelle estreme, con una frequenza maggiore di quelle intermedie
    - numero più ridotto di valori centrali.





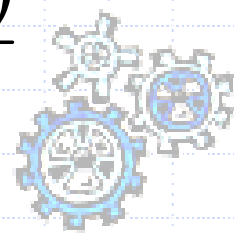
Curtosi (cont.)

- Indice di Pearson
  - $\beta=3$: distribuzione mesocurtica
  - $\beta>3$: distribuzione leptocurtica
  - $\beta<3$: distribuzione platicurtica

$$\beta = \frac{m_4}{\hat{s}^4}$$

- Coefficiente di curtosi
  - Una distribuzione leptocurtica ha K ~ 1/2
  - platicurtosi: k~0

$$K = \frac{\frac{1}{2}\left(x_{.75} - x_{.25}\right)}{\left(x_{.90} - x_{.10}\right)}$$



Coefficienti di Correlazione

- Covarianza

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Coefficiente di Pearson

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y}$$