Filippo Geraci

DATA MINING ED INFORMATION RETRIEVAL

Definizioni

- **Data mining**: attività di scoperta di informazione latente all'interno di un certo insieme di dati (tipicamente molto grande)
- **Information retrieval (IR)**: insieme delle tecnologie utilizzate per reperire una specifica informazione all'interno di un certo insieme di dati (tipicamente molto grande)

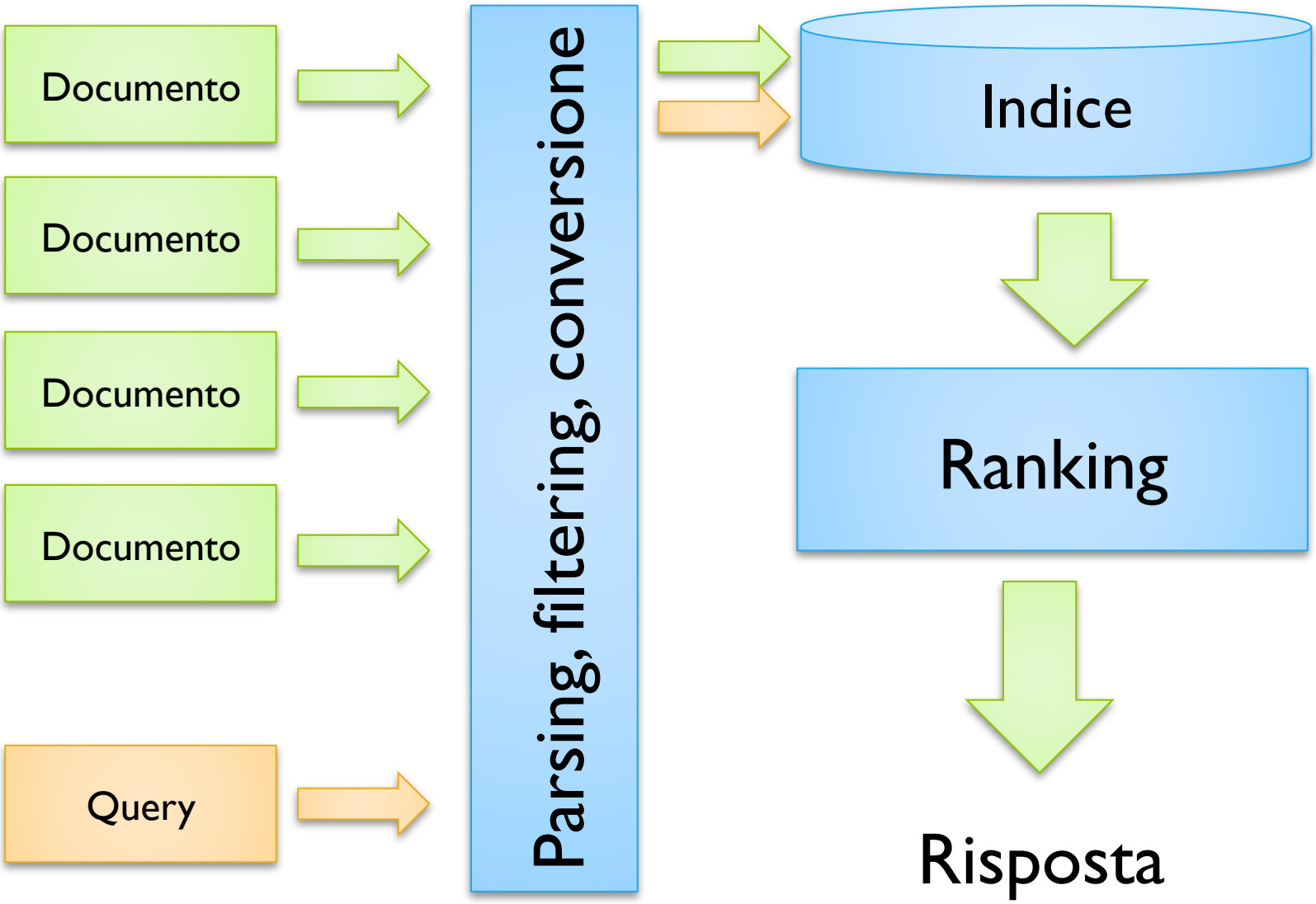
Definizioni

- **Data mining**: attività di scoperta di informazione latente all'interno di un certo insieme di dati (tipicamente molto grande)
- **Information retrieval (IR)**: insieme delle tecnologie utilizzate per reperire una specifica informazione all'interno di un certo insieme di dati (tipicamente molto grande)

Information retrieval (IR)

Parte I

Piattaforma di IR

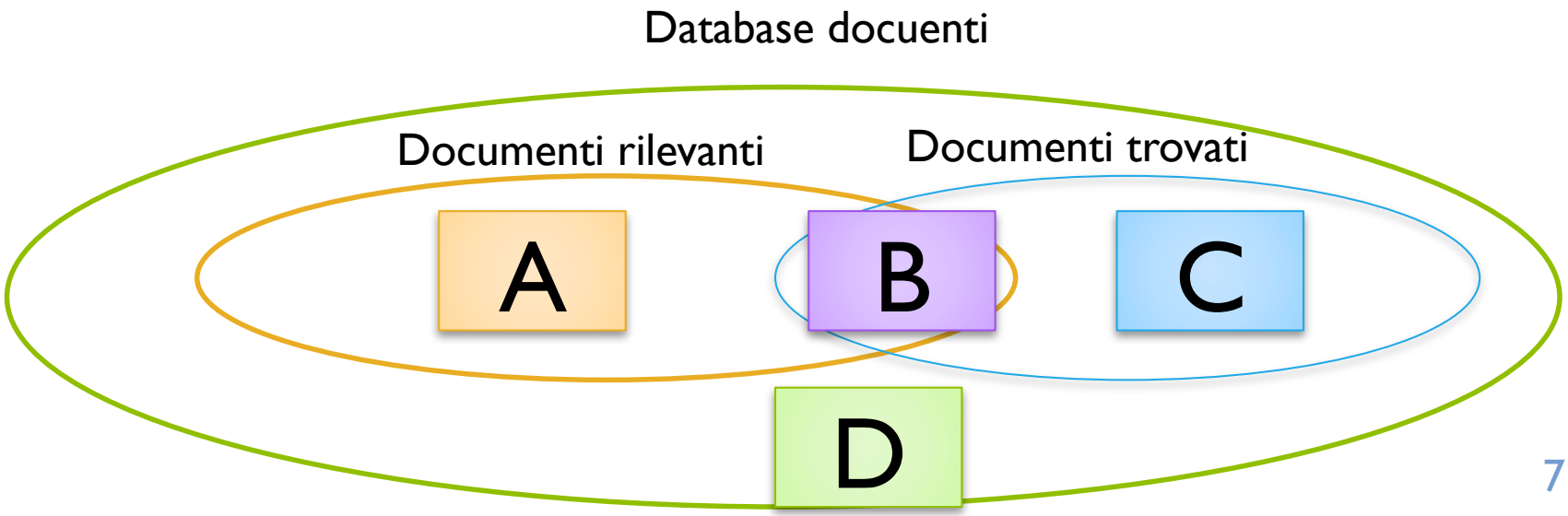


Problemi di progettazione delle piattaforme di IR

- Indicizzazione dei documenti e delle query
  - Qual è il miglior modo di rappresentarle?
- Valutazione della query (processo di retrieval)
  - Qual è il livello di affinità tra query e dicumento?
- Valutazione del sistema
  - Quant'è efficiente?
  - Quanti dei documenti trovati sono rilevanti? (Precision)
  - Riesce a trovare tutti i documenti rilevanti? (recall)

Precision e recall

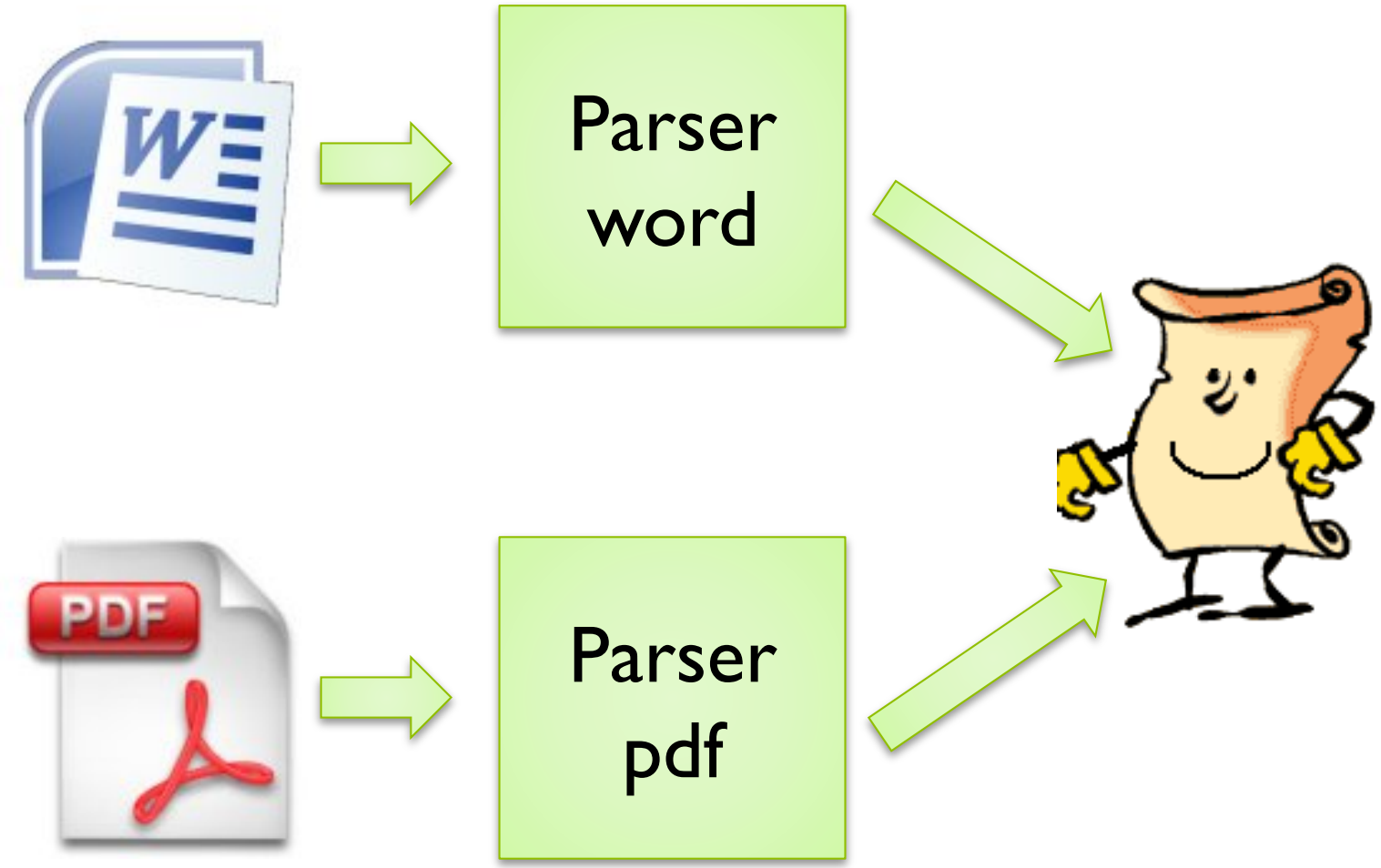
- **Precision**: percentuale di documenti rilevanti tra quelli restituiti (B/C)
- **Recall**: percentuale rilevanti resituiti rispetto al totale contenuto nella raccolta (B/A)
- Precision e Recall sono tra loro collegate, per esempio aumentando il numero di documenti ritrovati aumenta la recall e diminuisce la precision.



Tipi di documenti

- I documenti tipicamente trattati sono di tre tipi:
  - Testi, video/immagini, suoni
- Ogni tipo di documento utilizza tecniche diverse
  - In comune rimane lo schema della piattaforma di IR
- Spesso servono molti parser per supportare formati diversi dello stesso tipo di documento

Tipi di documenti



Misure di distanza

- In information retrieval si calcola spesso la distanza fra due elementi
  - Maggiore similarità vuol dire minore distanza e viceversa
- Tipicamente gli elementi sono rappresentati da vettori n-dimensionali
- Il passaggio da un documento ad un vettore dipende dal tipo di documento

Correlazione del coseno

- Cosine similarity

$$s(o_a, o_b) = \frac{o_a \cdot o_b}{\|o_a\| \cdot \|o_b\|}$$

- Si trasforma in una distanza

$$d(o_a, o_b) = \sqrt{1 - s^2(o_a, o_b)}$$

Dove o_a, o_b sono due vettori

Coefficiente di Jaccard

- Siano o_a, o_b sono due insiemi di features, il coefficiente di jaccard si definisce come:

$$J(o_a, o_b) = \frac{\#(o_a \cap o_b)}{\#(o_a \cup o_b)}$$

- Per due vettori il coefficiente può essere generalizzato come:

$$GJC(o_a, o_b) = \frac{\min_{i=1}^{m}(o_{a,i}, o_{b,i})}{\max_{i=1}^{m}(o_{a,i}, o_{b,i})}$$

Distanza di Minkowski

- Siano o_a, o_b due vettori, la distanza di minkowski è definita come:

$$L_p(o_a, o_b) = (\sum_{i=1}^{m} |o_{a,i} - o_{b,i}|^p)^{1/p}$$

- p=1 si riduce alla manatthan distance
- p=2 si riduce alla norma euclidea
- Se p è infinito si ha la norma infinito

$$L_\infty(o_a, o_b) = \max_{i=1}^{m}(o_{a,i}, o_{b,i})$$

Text information retireval

Esempio

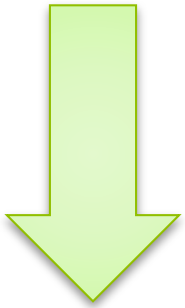
- L'esame di S.I.A è costituito da una prova scritta ed una prova orale. Il voto può variare da 18 a 30.

Trattamento del testo

- Normalizzazione del testo
  - Rimozione della punteggiatura
  - Rimozione numeri e sigle
  - Conversione dei caratteri in minuscolo
  - Ordinamento in ordine lessicografico
    - Rimuovo duplicati, ma ne ricordo la cardinalità
    - Devo rricordare le posizioni nella fase di indicizzazione
- Rimozione delle stop words
  - Vengono rimosse parole poco significative
    - Esempio: connettivi, articoli, preposizioni
  - Liste solitamente note a priori, oppure basate su analisi delle frequenze delle parole nel testo

Esempio

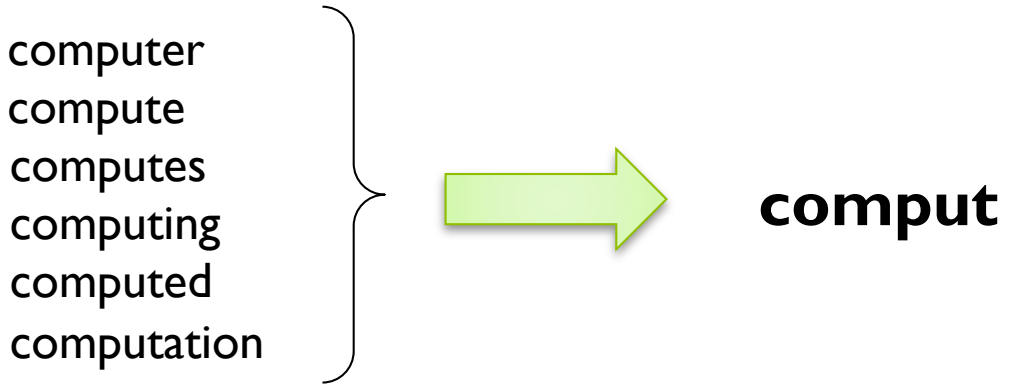
- ~~L'~~esame ~~di S.I.A è~~ costituito ~~da una~~ prova scritta ~~ed una~~ prova orale~~. Il~~ voto ~~può~~ variare ~~da 18 a 30.~~



- costituito esame orale prova (2) scritta variare voto

Stemminng

- La stessa parola può apparire in diverse forme pur mantenendo lo stesso significato
- Usato per incrementare l'efficacia in una ricerca e ridurre la dimensione dell'indice
- Stem: porzione della parola rimasta dopo aver rimosso i suoi suffissi o prefissi
- Nasce il problema della correttezza
- Serve un algoritmo per ogni lingua



Stemming - controindicazioni

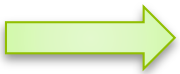
- Stemming aggressivo
  - organization → organ
  - army → arm
- Stemming timido
  - european → europe
  - creation → create

Pesatura dei termini:il modello TF-IDF

- Term frequency: numero di volte in cui un termine compare nel documento

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

- Document frequency: numero di documenti in cui compare Il termine

INVERSO  $\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$

- Il TF-IDF è un peso associato ad ogni termine in un documento

$$(\text{tf-idf})_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$$

Pesatura dei termini:il modello TF-IDF

- Uno score alto nel modello TF-IDF si ottiene da un valore alto di term frequency in un dato documento ed un valore basso del document frequency nell'intero dataset.
- Documenti memorizzati come matrice
- Ogni documento è un vettore in uno spazio n-dimensionale

Due aspetti della rappresentazione dei documenti

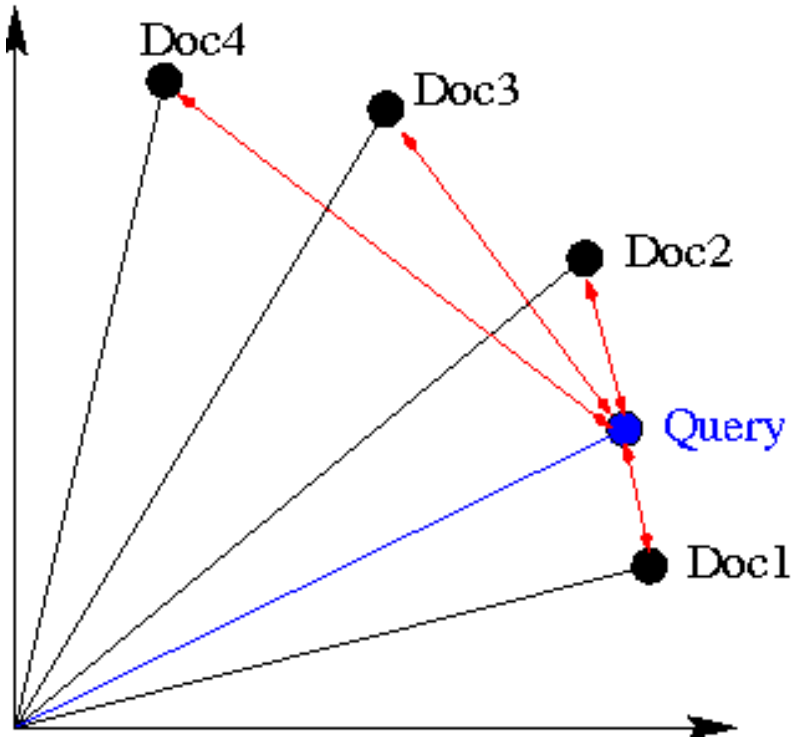
- Descrittività del modello
  - Qual'è il contesto del documento?
  - Importante per recall
- Discriminazione
  - Come distinguo questo documento dagli altri?
  - Importante per precision
- La rimozione di stopwords migliora la discriminazione riducendo il numero di parole in comune tra documenti dissimili
- Lo stemming migliora la decrittività rendendo uguali parole simili
- La migliore rappresentazione dei documenti deve bilanciare questi aspetti

Modelli di retrieval

- Data una query q ed un documento d viene assegnato un punteggio alla rilevanza della coppia (q, d)
- Questo valore è una stima della rilevanza di d per q
- Modelli di retrieval:
  - Booleano (basato su logica booleana query simili ad una formula)
  - Spazio vettoriale (dove documenti e query sono vettori e la similarità è il coseno dell'angolo)
  - Probabilistico (query in linguaggio naturale, cerca di valutare la probabilità che q ed d siano correlate)

Spazio vettoriale

- Proiettare i documenti e la query in punti di uno spazio euclideo ad alta dimensione e valutare la distanza



Problema

Vota Antonio, vota Antonio, vota Antonio, vota Antonio, vota Antonio, vota Antonio

- La correlazione tra query e documenti dipende solo dal testo
- Non si considerano i sinonimi

Indici inversi

credevo che l'azzurro dei tuoi occhi per me fosse sempre cielo

il pomeriggio è troppo azzurro e lungo per me

per sognare un cielo azzurro all'orizzonte senza nuvole

I	Azzurro	3	5	5
2	Cielo	11		4
3	Credevo	1		
4	Fosse	9		
5	Lungo		7	
6	Me	8	9	
7	Nuvole			9
8	Occhi	6		
9	Orizzonte			7
10	Per	7	8	1
11	Pomeriggio		2	
12	Sempre	10		
13	Senza			8
14	Sognare			2
15	Troppo		4	
16	Tuoi	5		

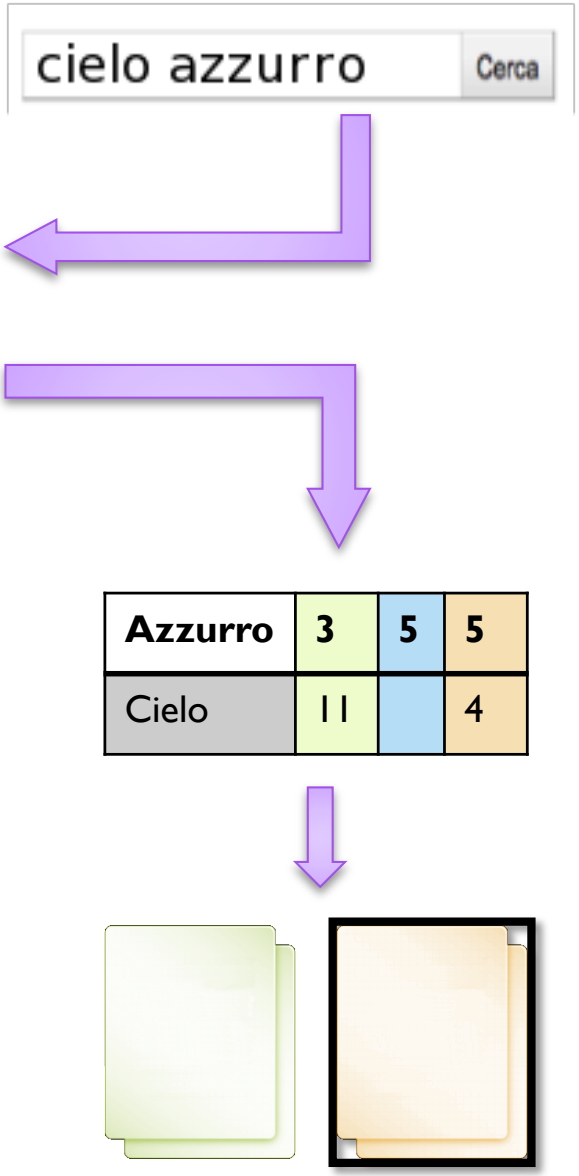
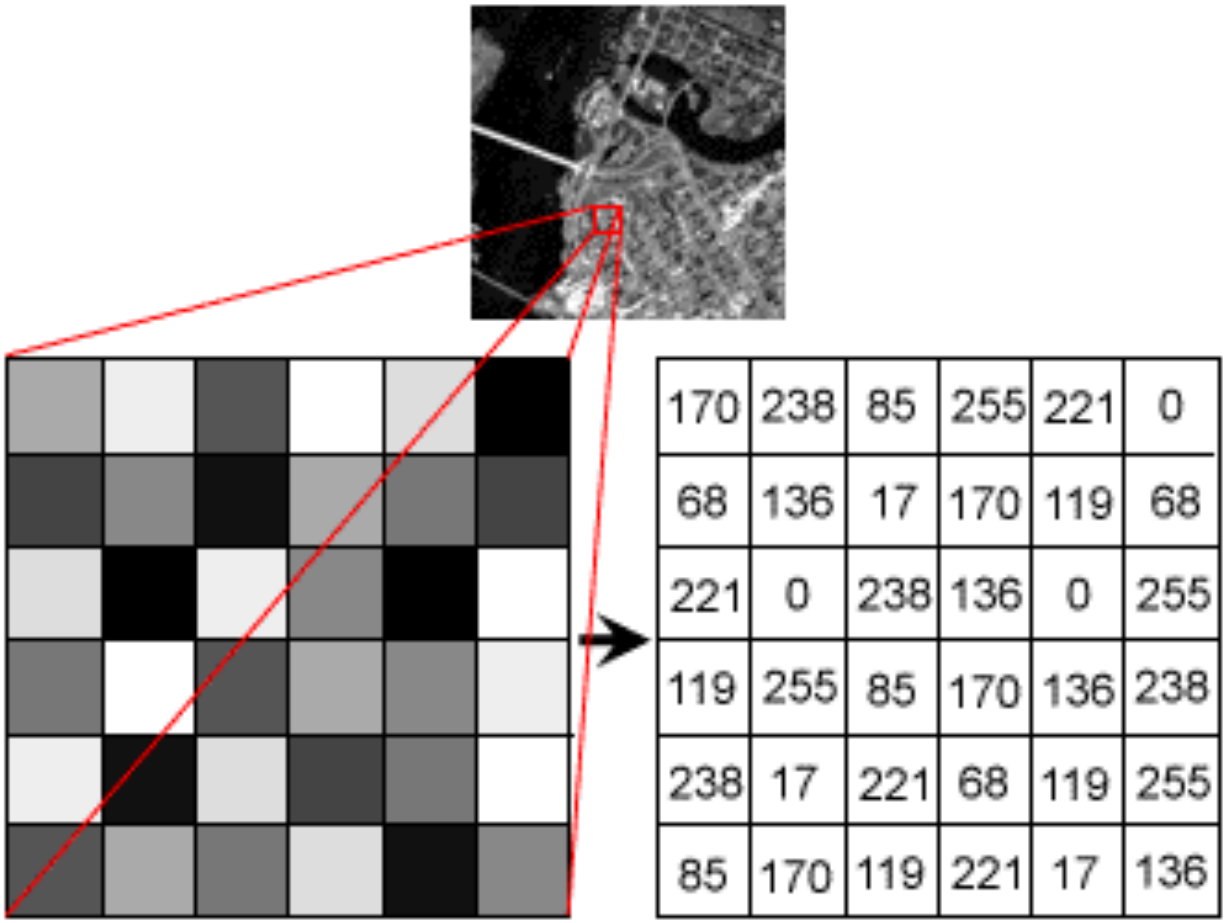


Image and video information retireval

Video ed immagini

- Un video è costituito da una sequenza di immagini (frames)
- Per ogni frame posso usare le tecniche di rappresentazione delle immagini
- Quali frame rappresentare?
  - Tutti: costosissimo (un video ha almeno 25 frames per secondo)
  - Uno ogni k (scelgo k in base alle mie esigenze)
  - I più rappresentativi (Bisogna individuarli)

Rappresentazione immagini



170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

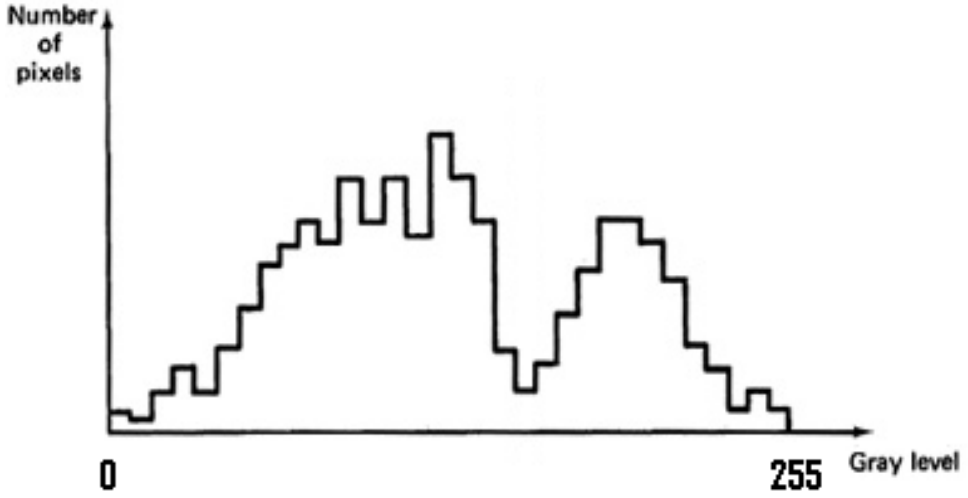
Fonte: http://hosting.soonet.ca/eliris/remotesensing/LectureImages/pixel.gif

Rappresentazione tramite Feature di un'immagine

- Una feature è una rappresentazione, tramite un vettore di valori numerici, di una immagine
- In genere una feature è una caratteristica facilmente misurabile dell'immagine
- L'immagine in esame viene quindi descritta usando i valori di un insieme di feature pre-scelte f1, …, fn

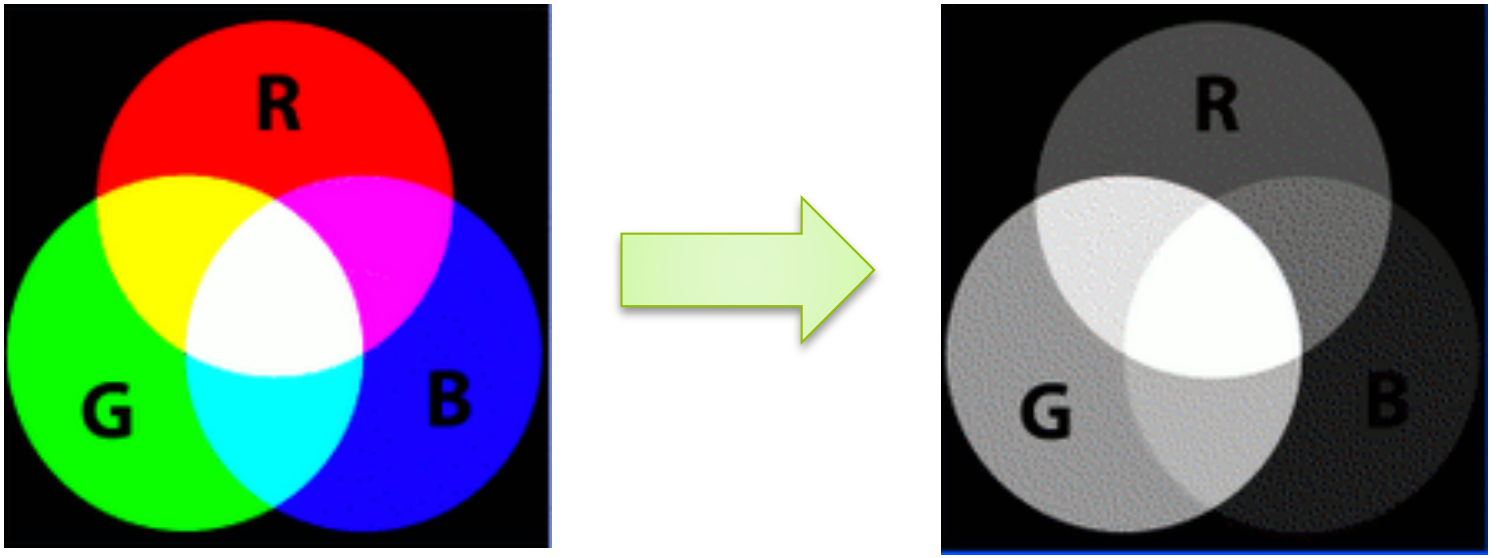
Istogrammi dell'intensità di grigio

- Istogramma dell'intensità dei pixel
  - Divido il range [0, 255] in *k* bin
  - Assegno ogni pixel al bin corrispondente
  - $f = (v_0, \ldots, v_{k-1})^T$, dove:
  - v_i = *# pixel nell'i-esimo bin*

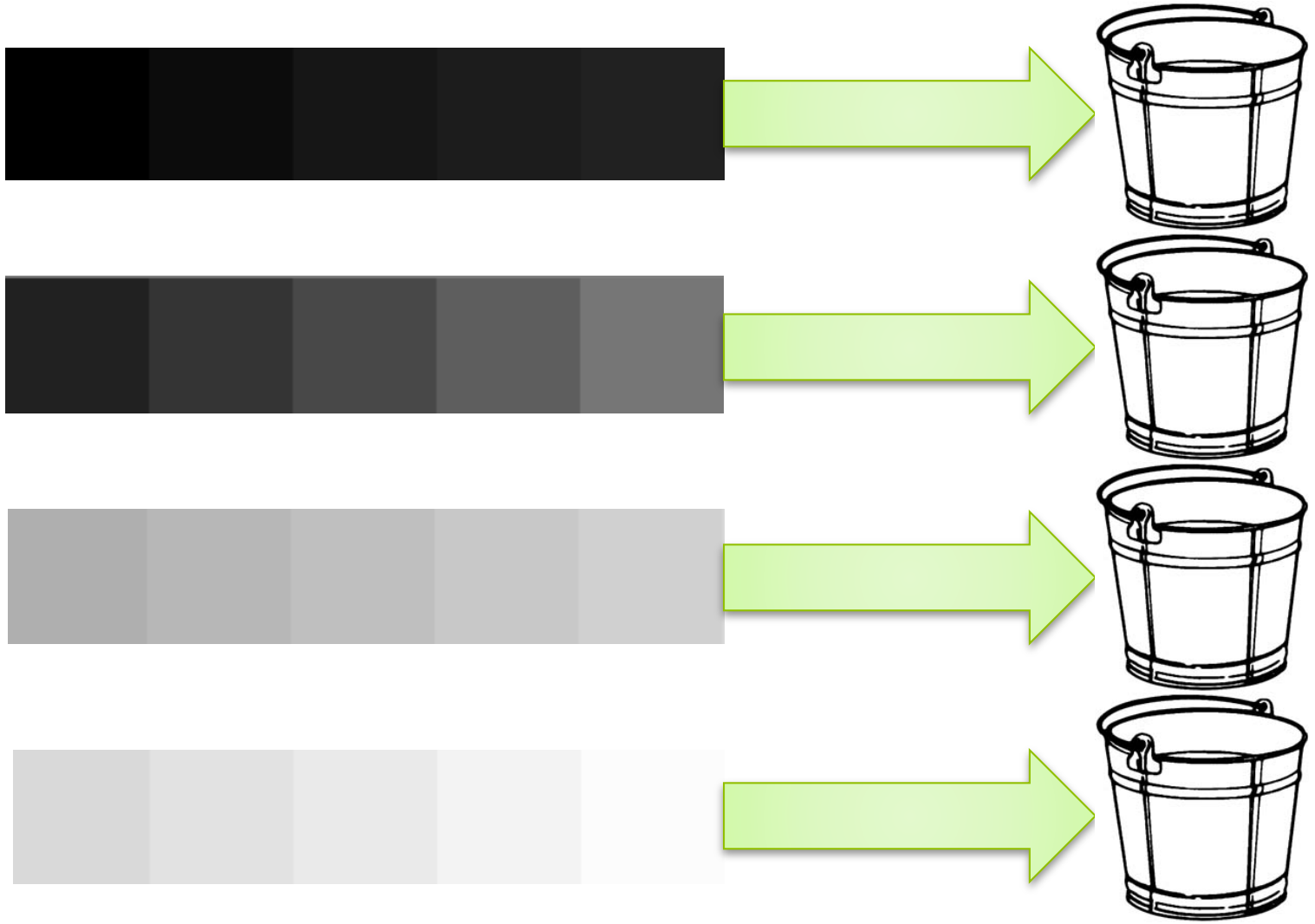


Da RGB a scala di grigi

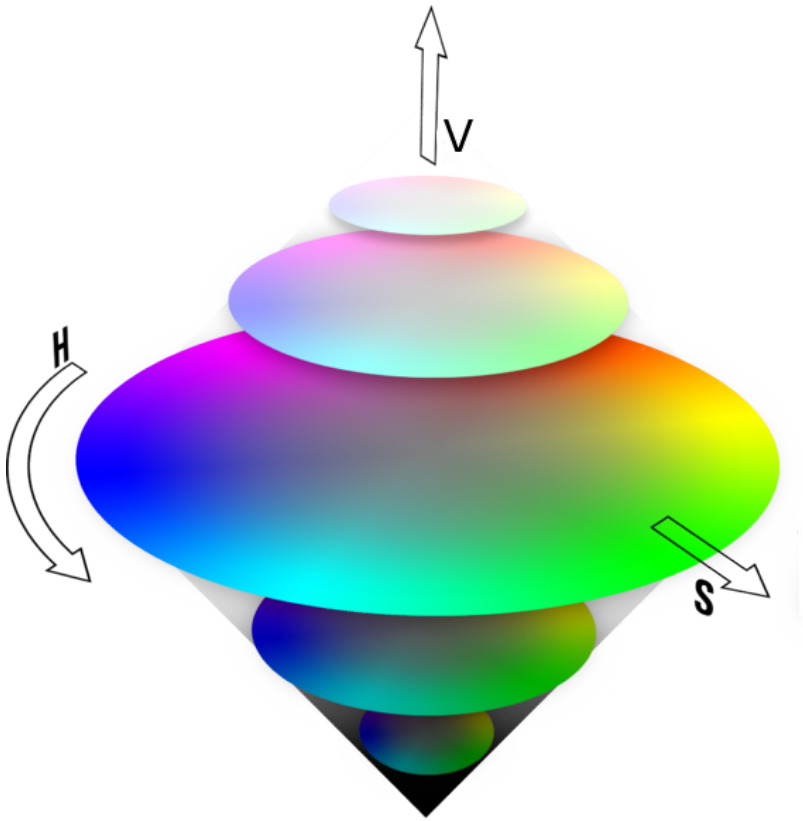
- Gray = R *0.33 + G *0.33 + B *0.34
- Gray, R,G,B compresi tra 0 e 255



Esempio

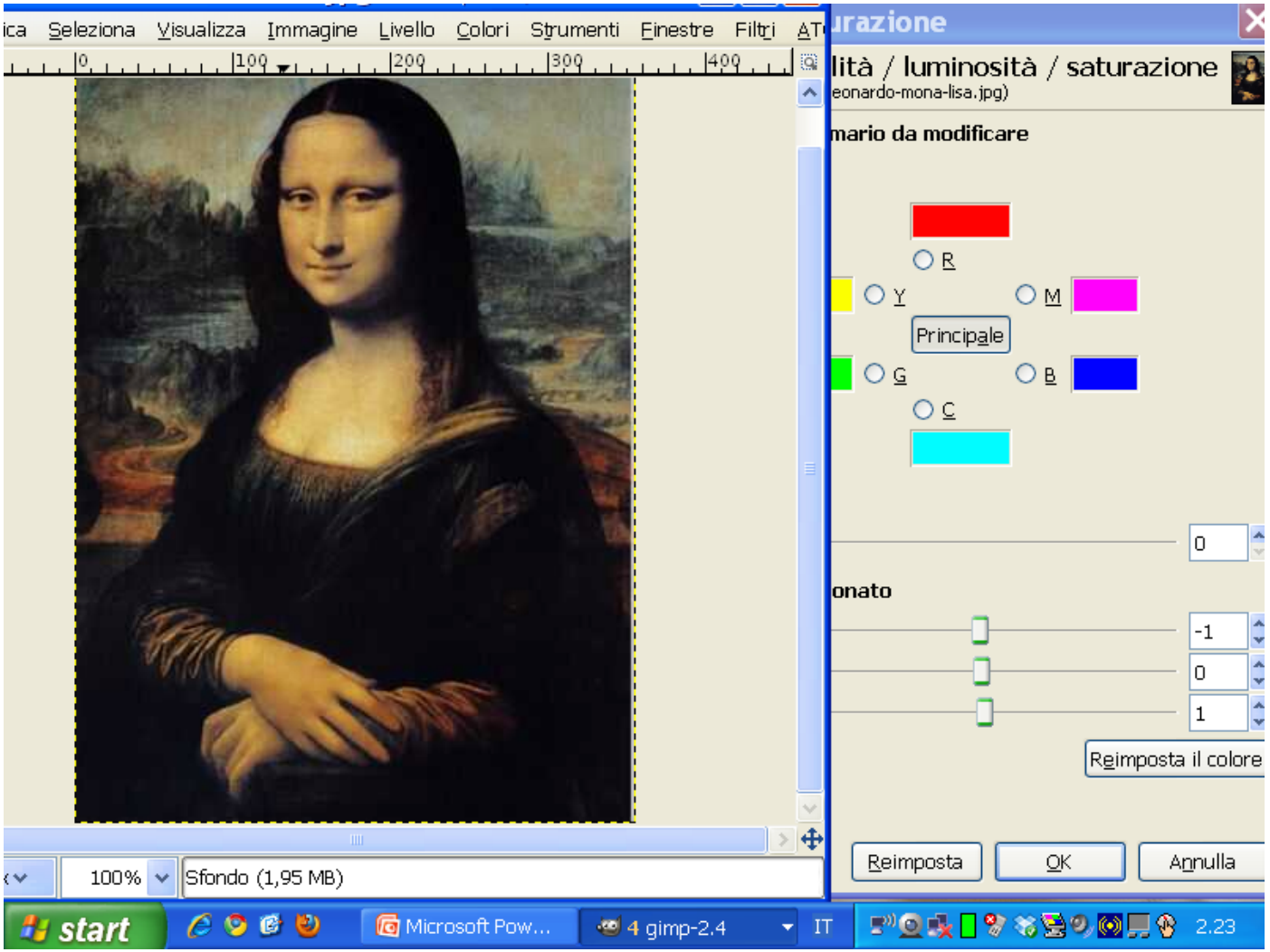


Lo spazio di colore HSV

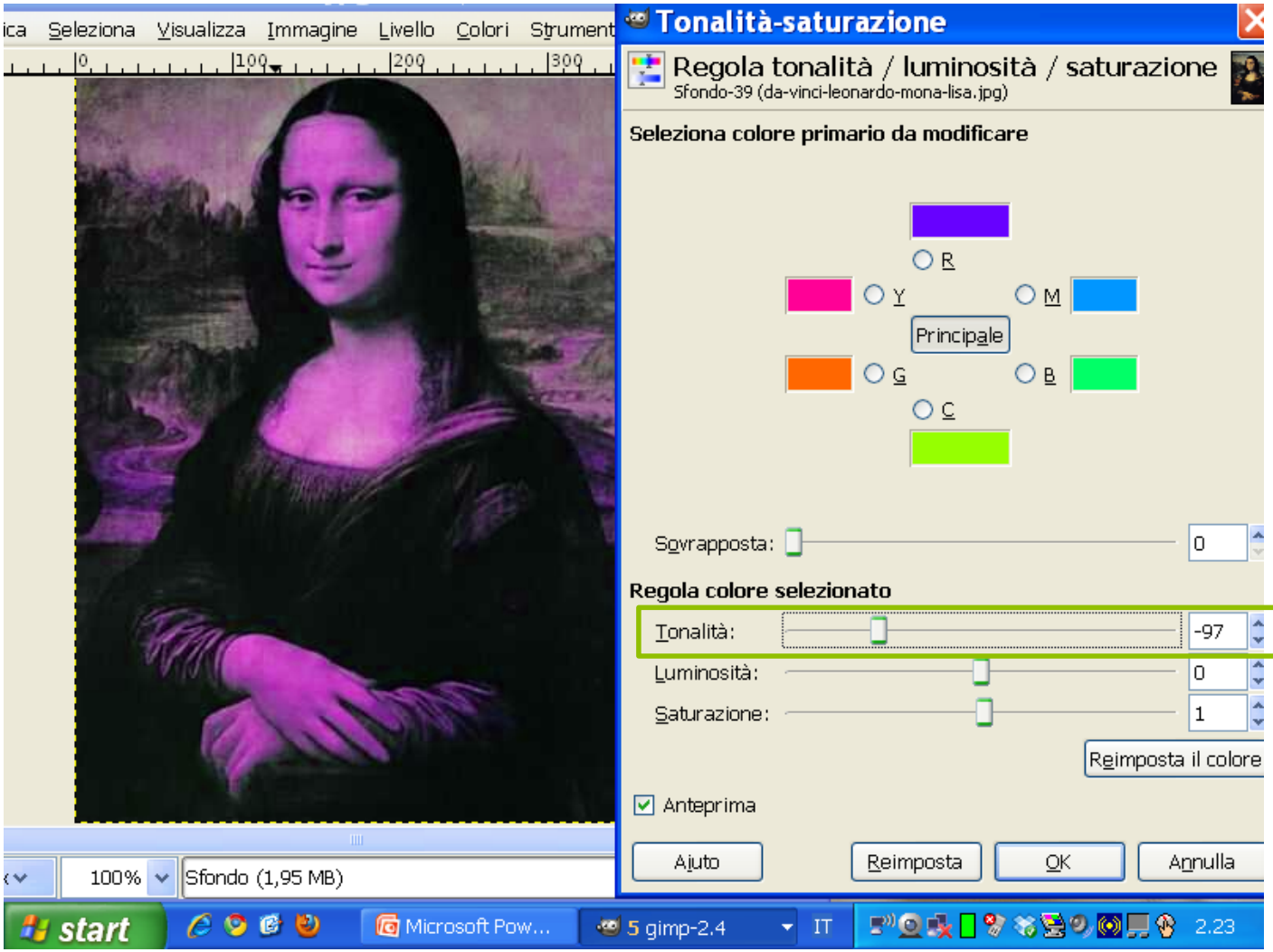


- Posso usare lo stesso metodo dei vettori di bin per ognuna delle caratteristiche di
- Hue, saturation e value

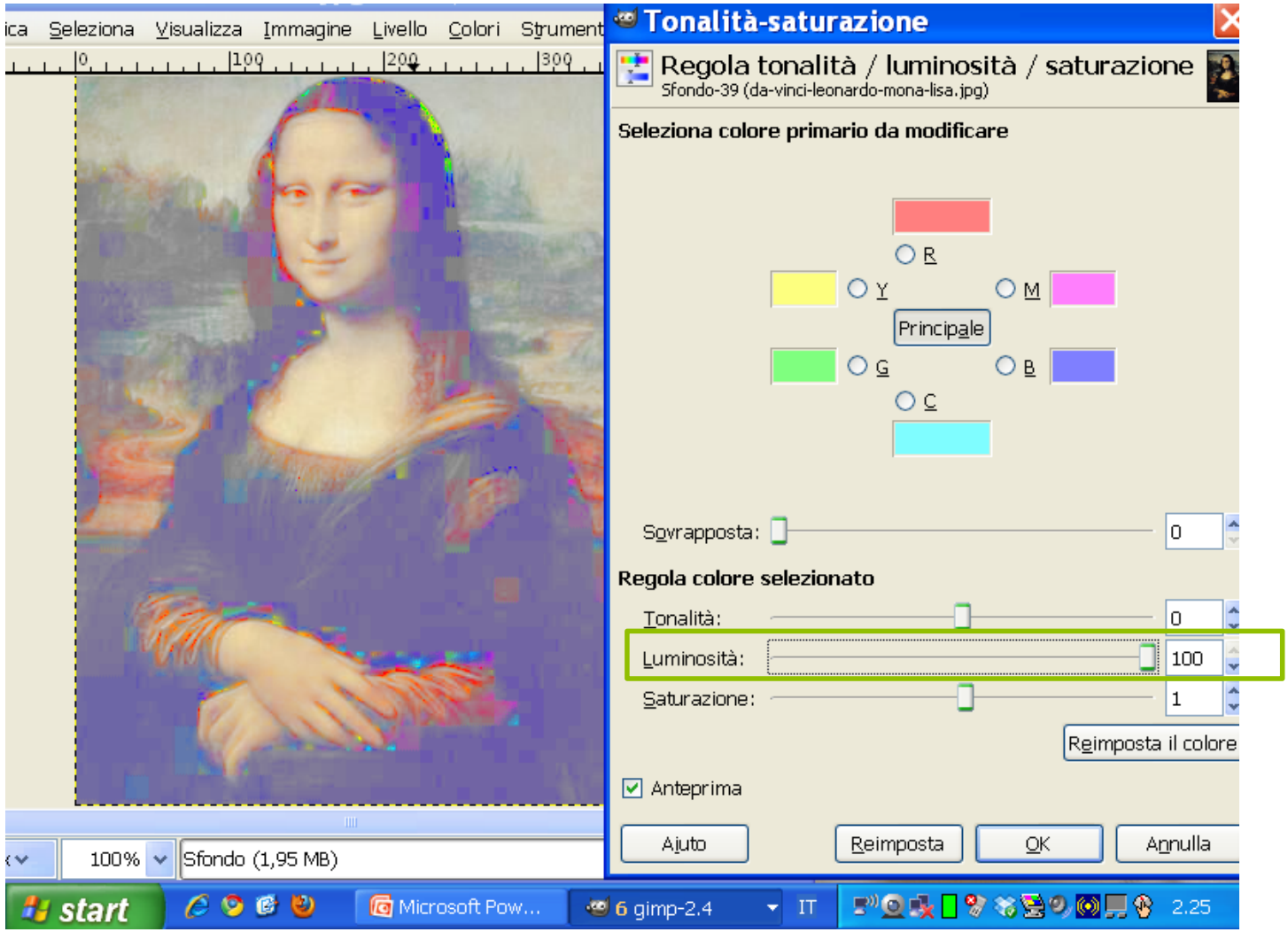
Esempio HSV


Seleziona
Visualizza
Immagine
Livello
Colori
Strumenti
Finestre
Filtri
100%
Sfondo (1,95 MB)
lità / luminosità / saturazione
mario da modificare
R
Y
M
Principale
G
B
C
onato
Reimposta il colore
Reimposta
OK
Annulla
start
Microsoft Pow...
4 gimp-2.4
IT
2.23

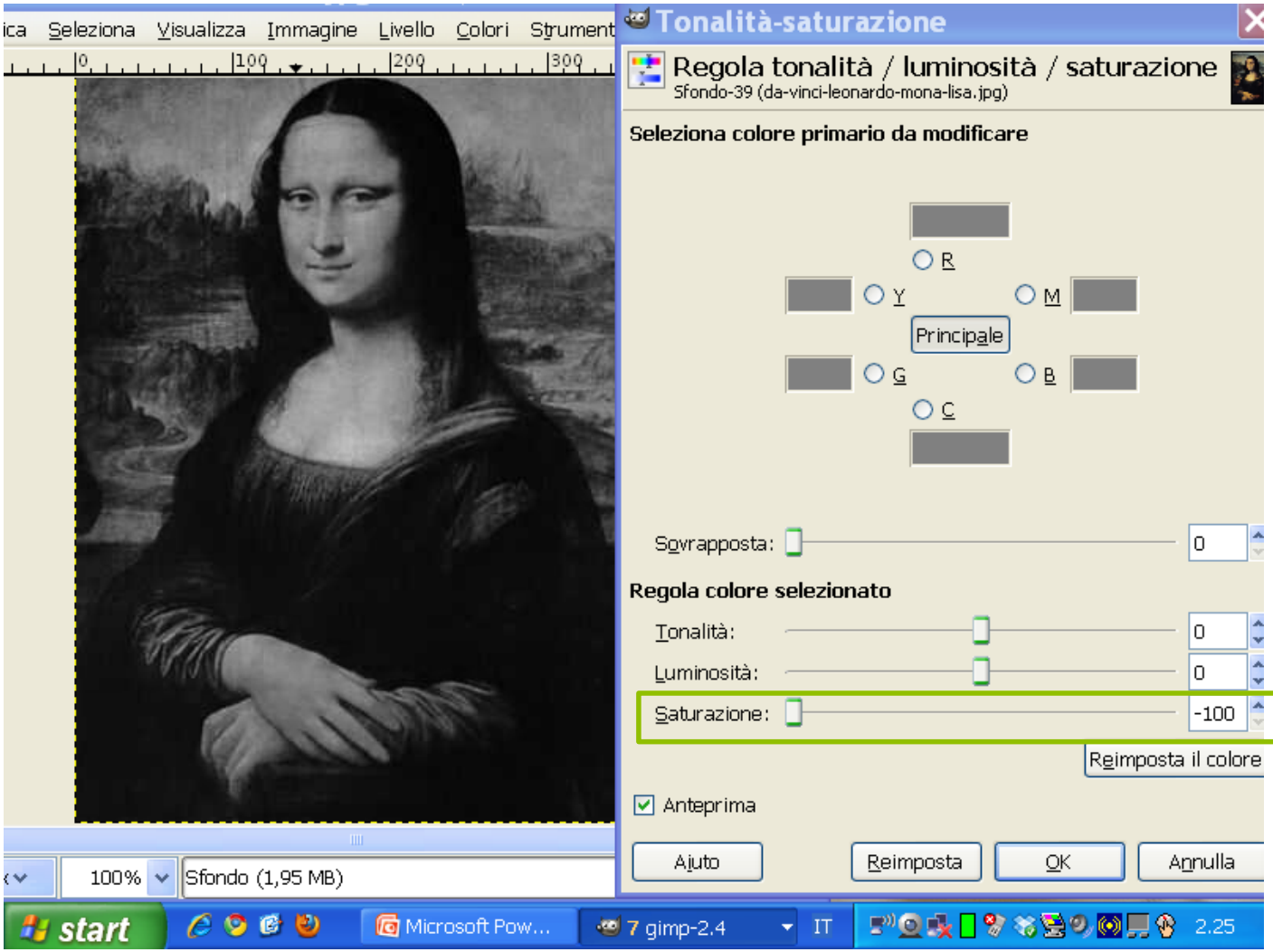
Esempio HSV


Seleziona
Visualizza
Immagine
Livello
Colori
Tonalità-saturazione
Regola tonalità / luminosità / saturazione
Sfondo-39 (da-vinci-leonardo-mona-lisa.jpg)
Seleziona colore primario da modificare
R
Y
M
Principale
G
B
C
Sovrapposta:
0
Regola colore selezionato
Tonalità:
-97
Luminosità:
0
Saturazione:
1
Reimposta il colore
Anteprima
Aiuto
Reimposta
OK
Annulla
100%
Sfondo (1,95 MB)
start
Microsoft Pow...
5 gimp-2.4
IT
2.23

Esempio HSV


Seleziona
Visualizza
Immagine
Livello
Colori
Tonalità-saturazione
Regola tonalità / luminosità / saturazione
Sfondo-39 (da-vinci-leonardo-mona-lisa.jpg)
Seleziona colore primario da modificare
R
Y
M
Principale
G
B
C
Sovrapposta:
0
Regola colore selezionato
Tonalità:
0
Luminosità:
100
Saturazione:
1
Reimposta il colore
Anteprima
Aiuto
Reimposta
OK
Annulla
100%
Sfondo (1,95 MB)
start
Microsoft Pow...
6 gimp-2.4
IT
2.25

Esempio HSV


Seleziona
Visualizza
Immagine
Livello
Colori
Tonalità-saturazione
Regola tonalità / luminosità / saturazione
Sfondo-39 (da-vinci-leonardo-mona-lisa.jpg)
Seleziona colore primario da modificare
R
Y
M
Principale
G
B
C
Sovrapposta:
0
Regola colore selezionato
Tonalità:
0
Luminosità:
0
Saturazione:
-100
Reimposta il colore
Anteprima
Aiuto
Reimposta
OK
Annulla
100%
Sfondo (1,95 MB)
start
Microsoft Pow...
7 gimp-2.4
IT
2.25

Spazio delle features

- Supponendo di utilizzare *n* features che sono vettori di k elementi a valori in ***R***, una immagine *I* può essere rappresentata tramite il vettore:

$$x(I) = (f_1(I)^T \quad \dots \quad f_n(I)^T)^T$$

- *che* è un punto in $\boldsymbol{R}^{n*k}$

Clustering e classificazione

Parte 2

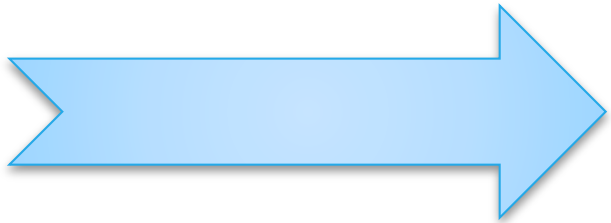
Clustering

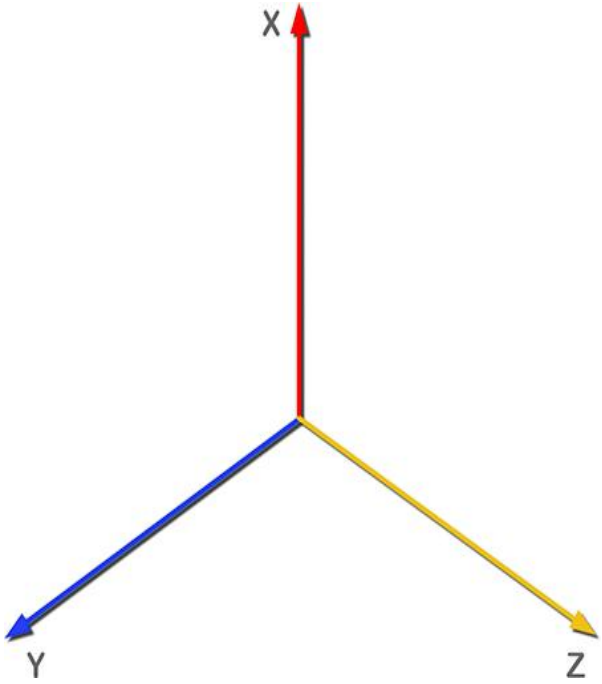
- Suddivide un insieme di oggetti in gruppi omogenei
- Procedure indipendenti dal tipo di dati
  - Utilizzabile anche quando non sono disponibili informaioni a-priori sul dominio dei dati;
  - Non possono essere introdotte ottimizzazioni dipendenti dal particolare problema

Cluster hypothesis

- Se due oggetti sono molto simili ed il primo e' anche simile ad un terzo oggetto, molto probabilmente anche tra il secondo ed il terzo oggetto esiste una similarità

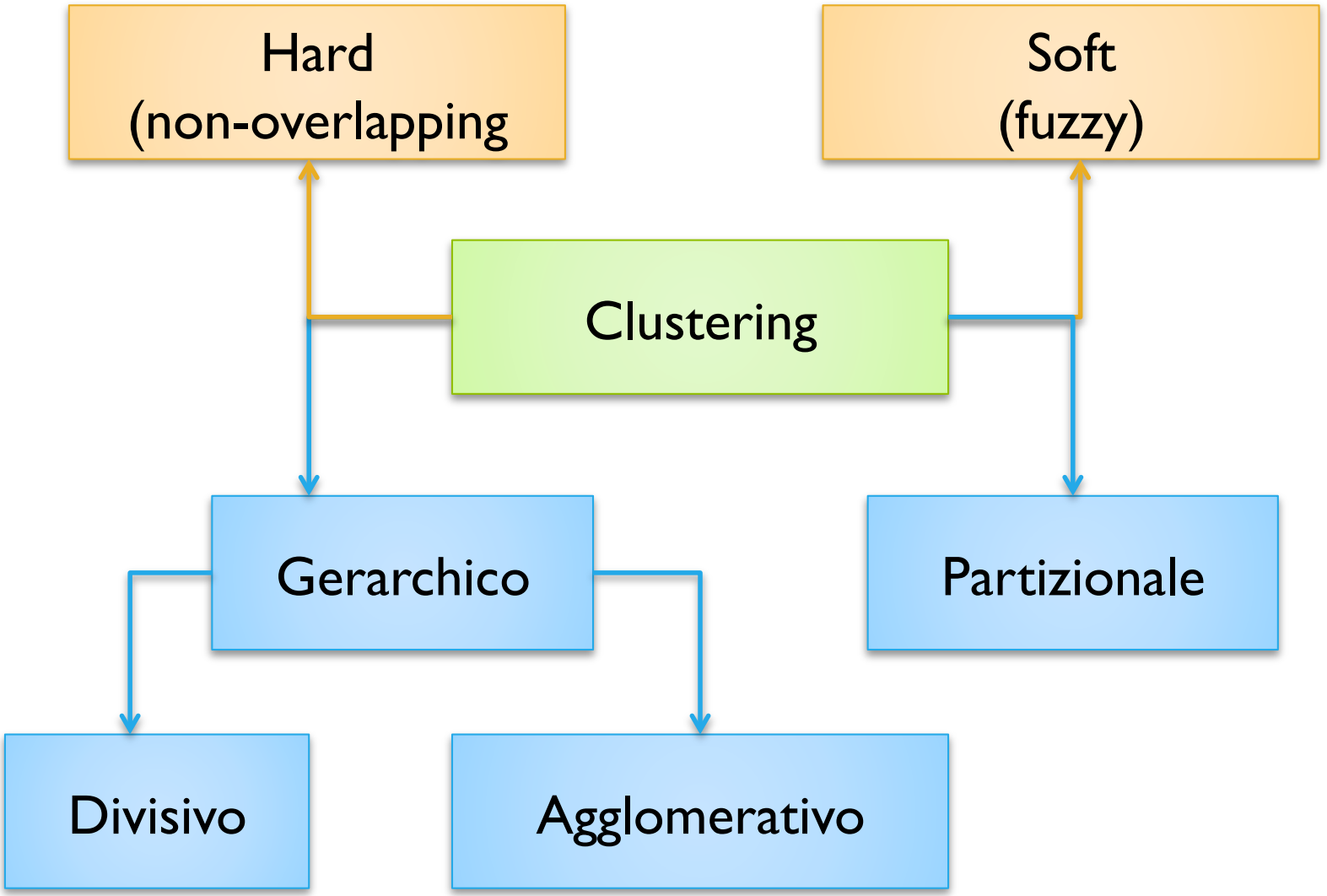
Come funziona





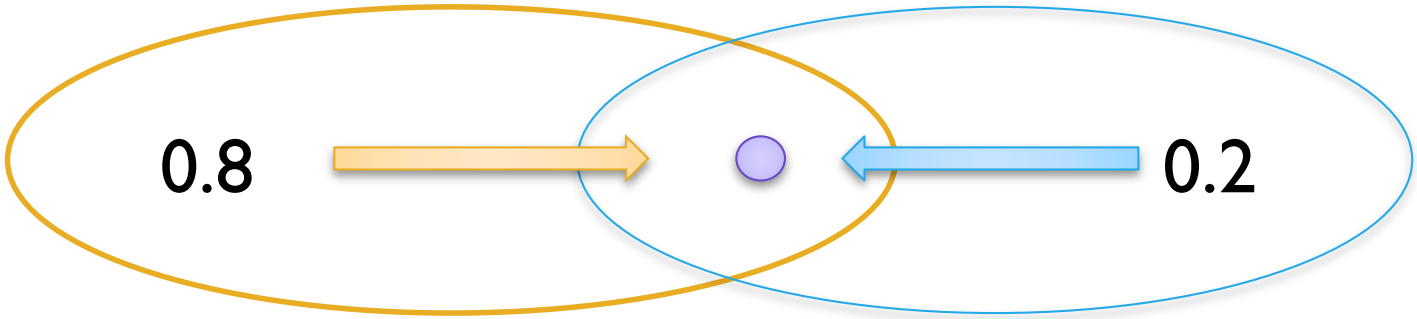
- Tramite tecnologie di information retrieval si trasformano i dati in punti appartenenti ad uno spazio vettoriale (in alcuni casi euclideo) e si lavora su quello
- Indipendente dal tipo di dato

Tassonomia degli algoritmi di clustering



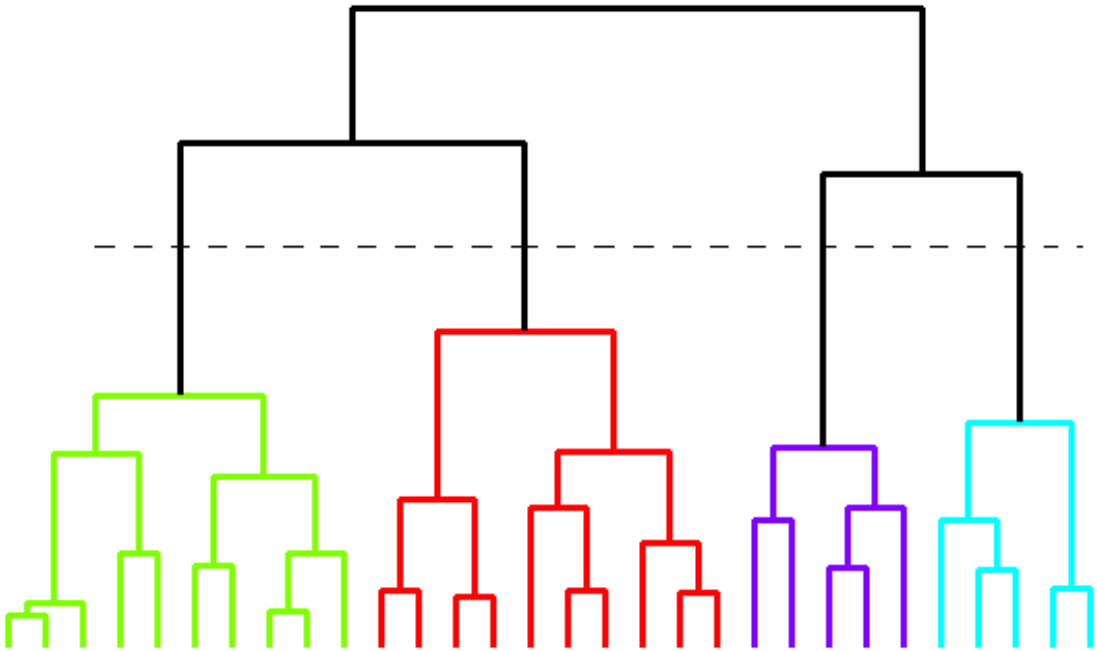
Hard e soft clustering

- Hard clustering: un elemento appartiene solo e soltanto ad un cluster
- Soft clustering: un elemento può appartenere a più cluster
  - Solitamente viene associato un valore di peso (probabilità) di appartenenza dell'elemento ad ogni cluster



Clustering gerarchico

- Richiede l'intera matrice delle distanze
- Produce dendrogramm
- Tagliando l'albero a vari livelli si ottiene un numero diverso di cluster



Clustering gerarchico divisivo

1. Per ogni cluster seleziona una coppia di punti diametrali
2. Se un cluster soddisfa il criterio di partizionamento allora:
   a) Dividi il cluster usando come centri la coppia selezionata nel punto 1
   b) Torna al punto 1
3. Se non ci sono più cluster da dividere fermati

Criteri di partizionamento

- Ogni cluster fino a che tutti i cluster non hanno un solo elemento (albero completo)
- Il più grande (albero bilanciato)

Clustering gerarchico agglomerativo

1. All'inizio ad ogni punto si associa un cluster
2. Ad ogni iterazione
   a) Costruisci la matrice delle distanze tra cluster
   b) Unisci i cluster in accordo al criterio di **linkage**
3. Ricomincia dal punto 2 finché tutti i punti sono in un unico cluster

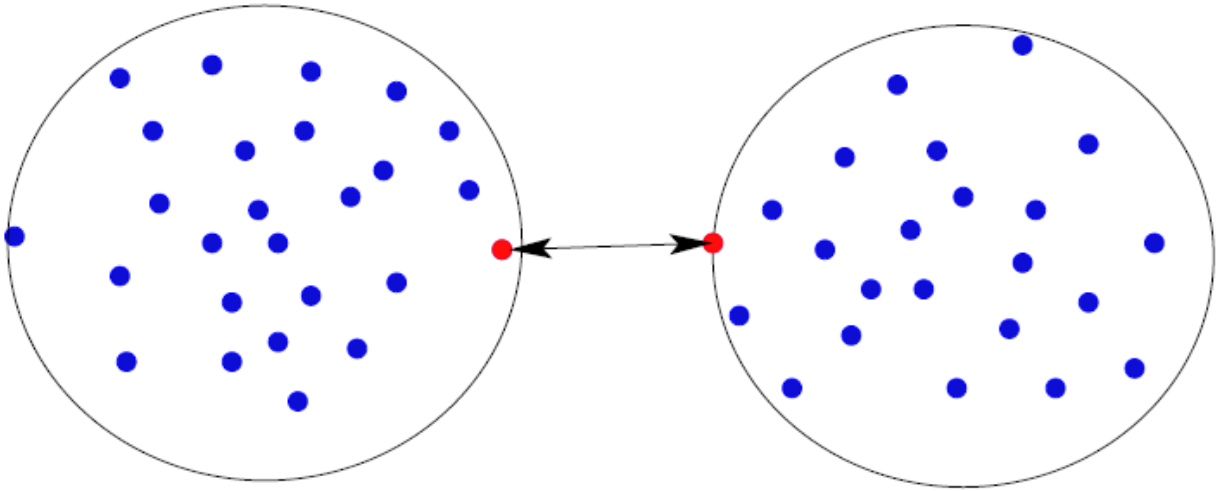
---

- HAC è l'algoritmo gerarchico più usato
- Computazionalmente costoso
- Produce intero albero

Criteri di linkage: single linkage

$$d(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p, q)$$

- Distanza tra i due elementi più vicini appartenenti a cluster diversi

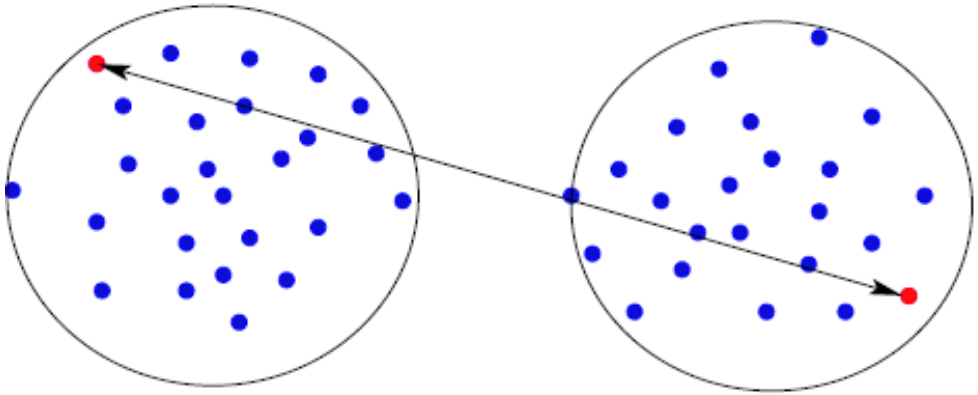


- Forza due cluster ad unirsi solo in base a due elementi non considerando gli altri
  - Crea effetto concatenazione

Criteri di linkage: complete linkage

$$d(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

- Distanza tra i due elementi più lontani appartenenti a cluster diversi

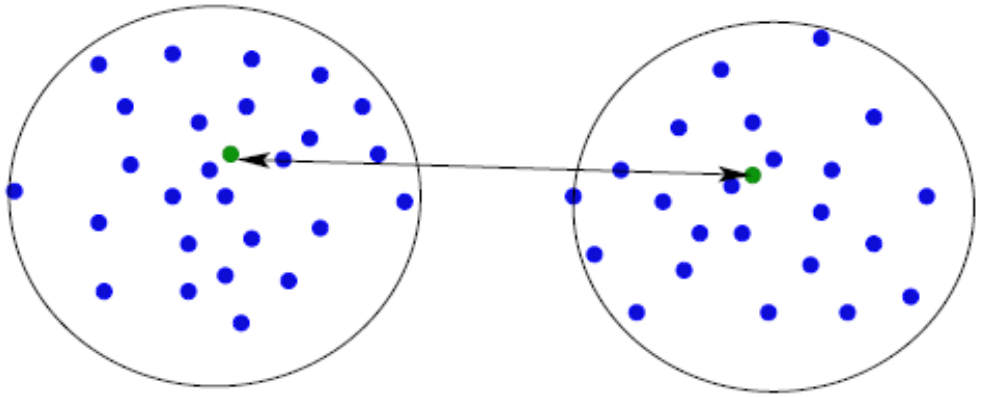


- Crea cluster molto compatti
- Non tollerante a dati contententi "rumore"

Criteri di linkage: complete linkage

$$d(C_i, C_j) = \frac{1}{|\ C_i\ ||\ C_j\ |} \sum_{p \in C_i} \sum_{q \in C_j} d(p, q)$$

- Media tra le distanze di tutti gli elementi appartenenti a cluster diversi



- Robusto rispetto a concatenazione e "rumore"
- Computazionalemte costoso

Clustering partitivo

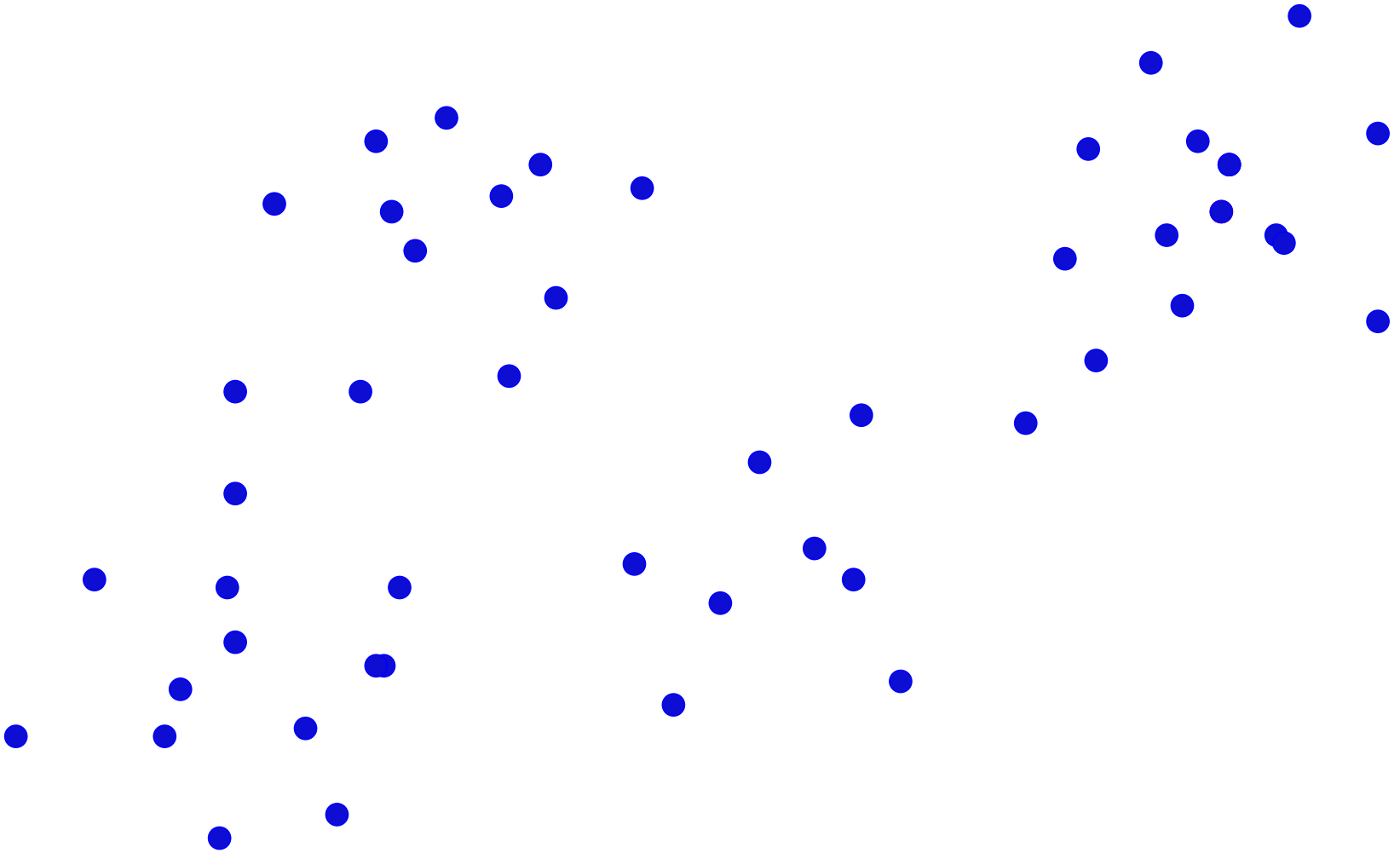
- Algoritmo FPF (furthest point first)
  - Il modo più intuitivo e veloce
  - Minimizza il massimo raggio dei cluster

$$\min_{j} \max_{x \in C_j} M_v(x, \mu_j)$$

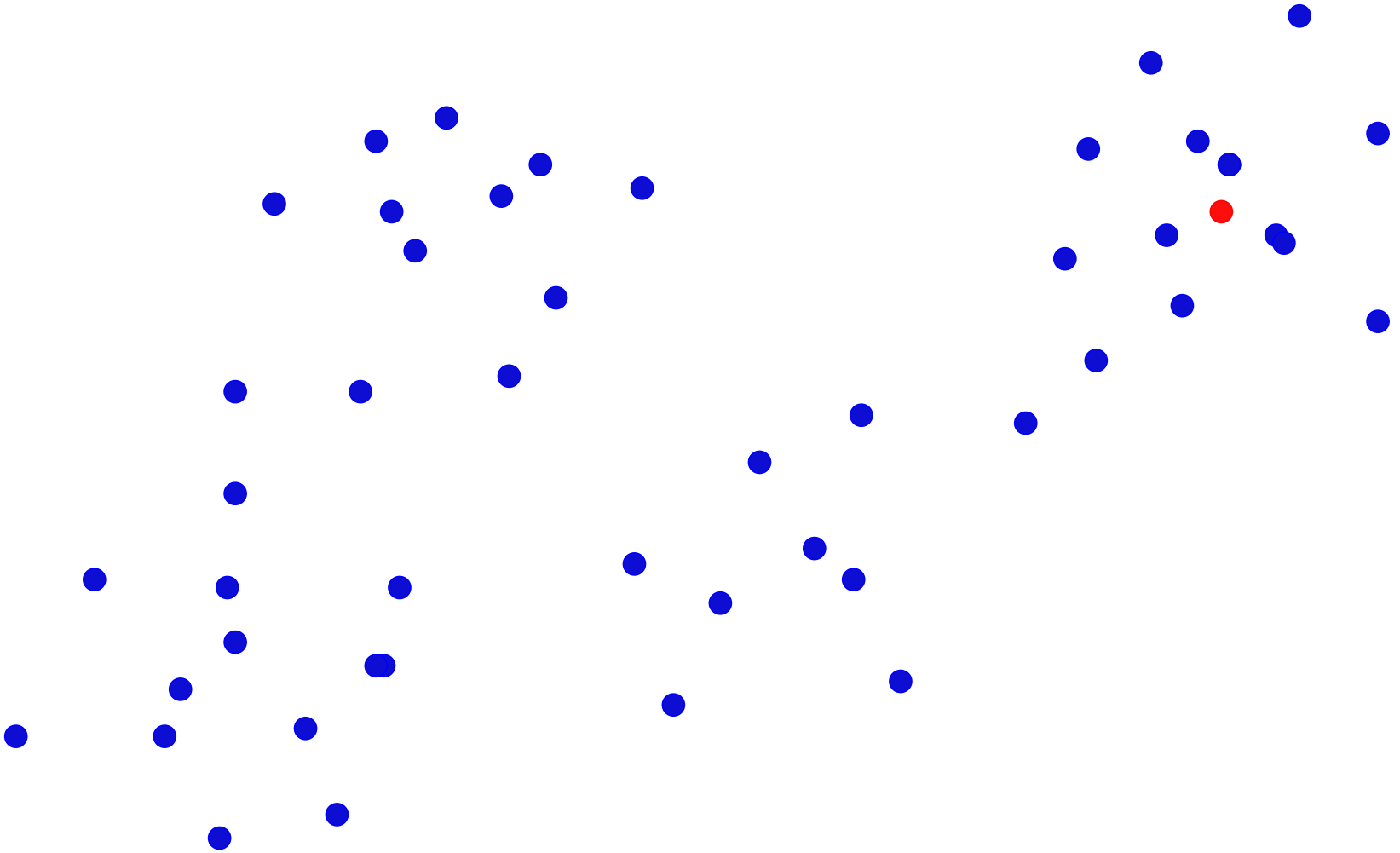
- Algoritmo k-means
  - Il più usato
  - Minimizza la somma dei quadrati delle distanze dei punti dal centroide di riferimento

$$\min \sum_{j} \sum_{x \in C_j} (M_v(x, \mu_j))^2$$

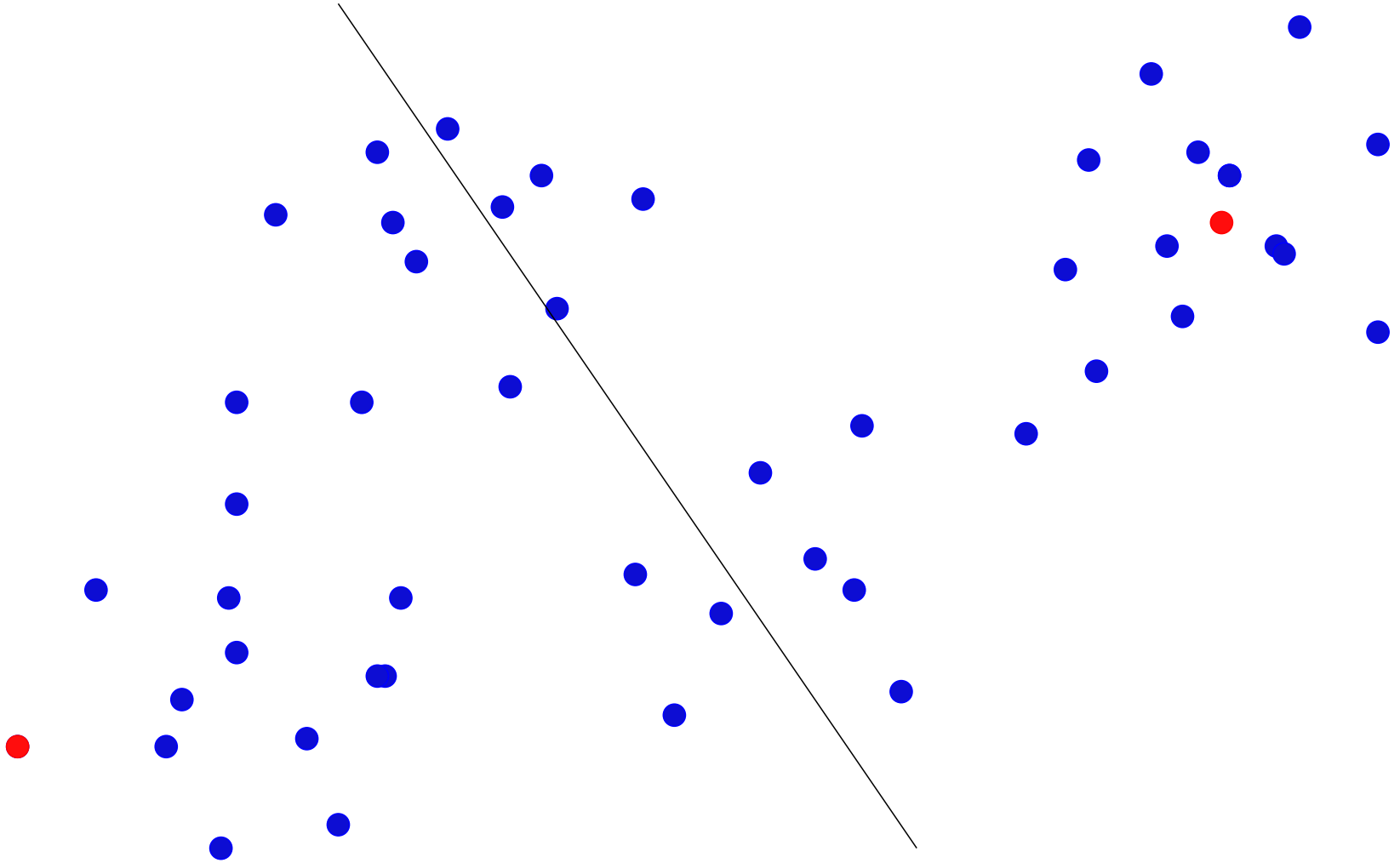
FPF con un esempio



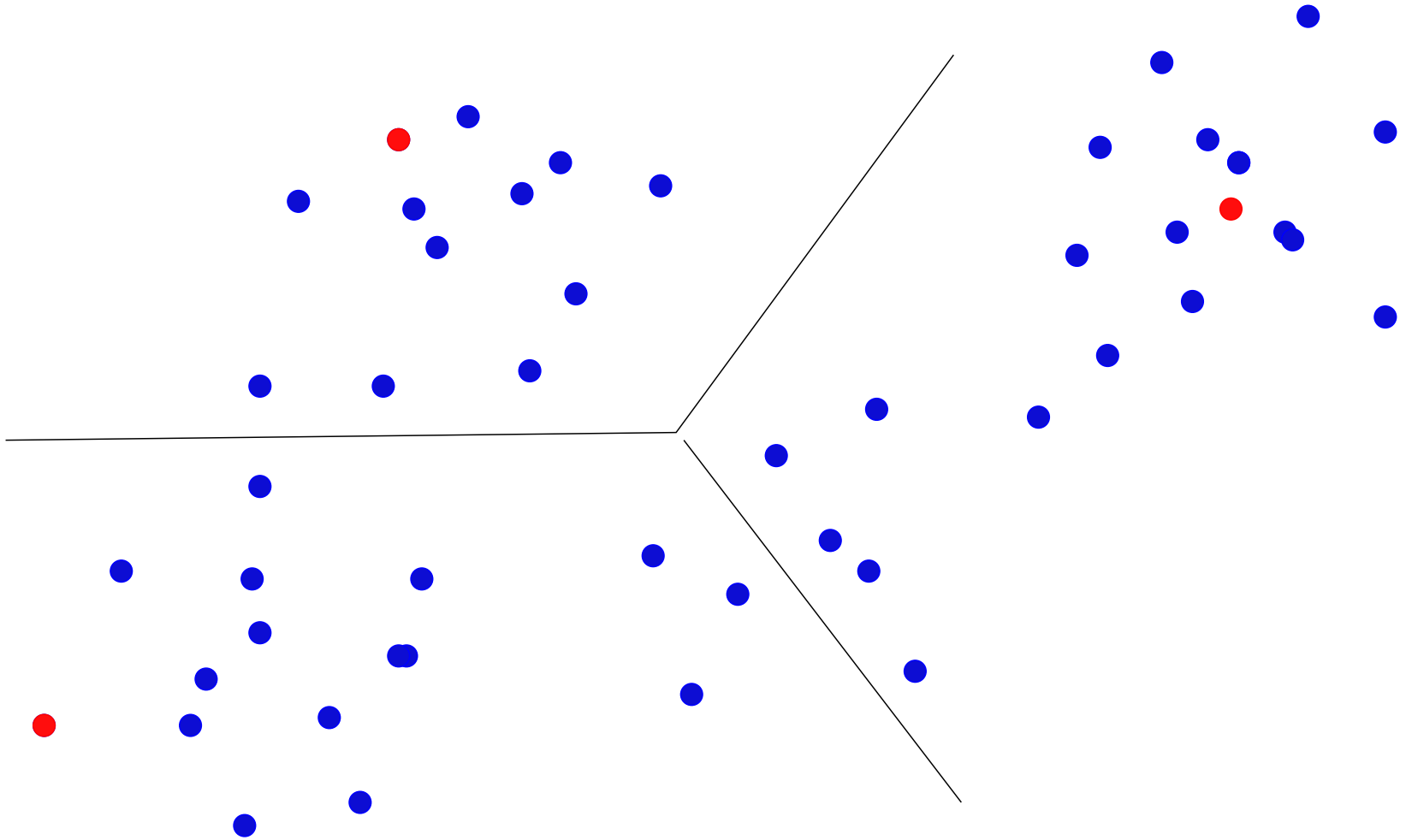
FPF con un esempio



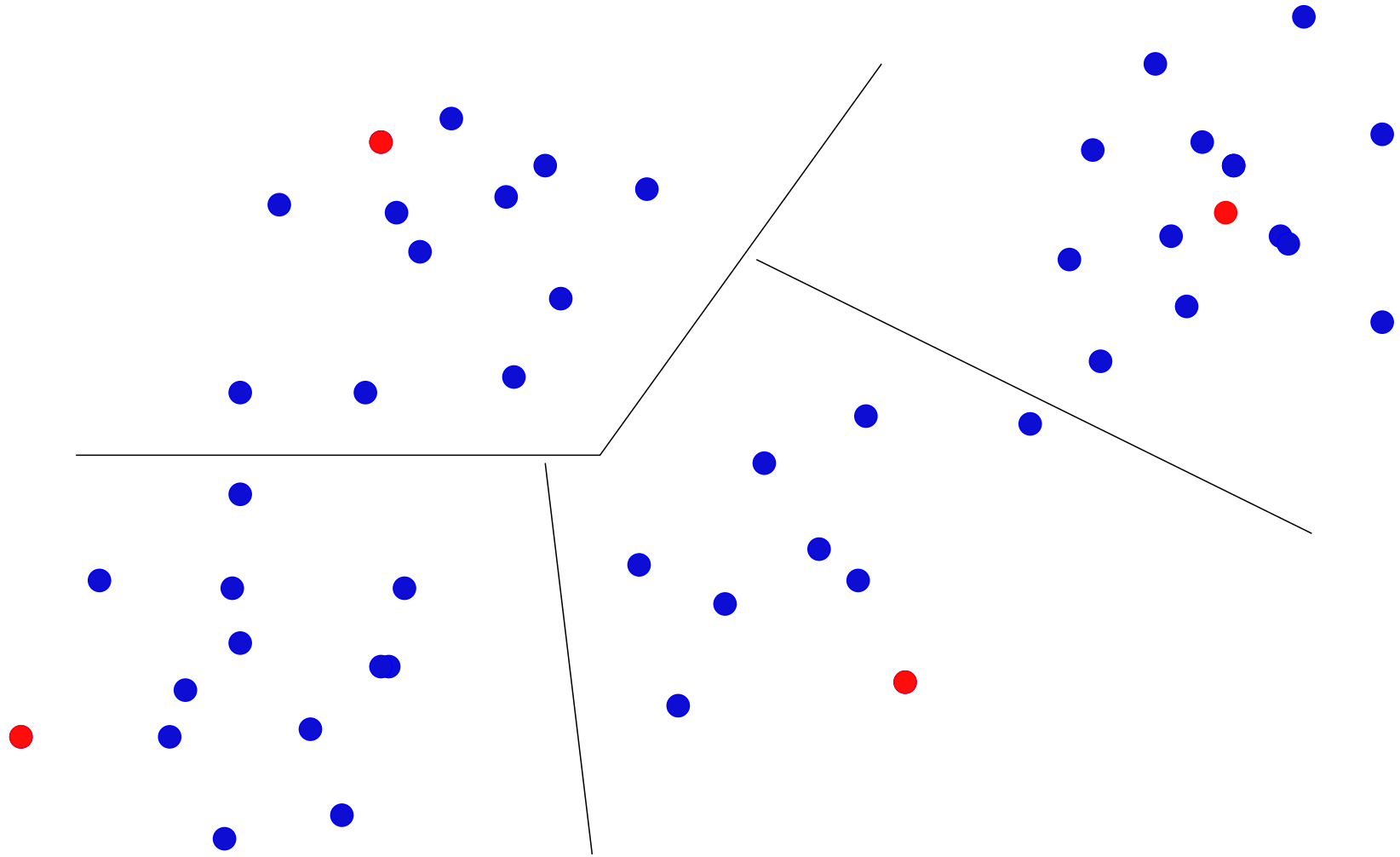
FPF con un esempio



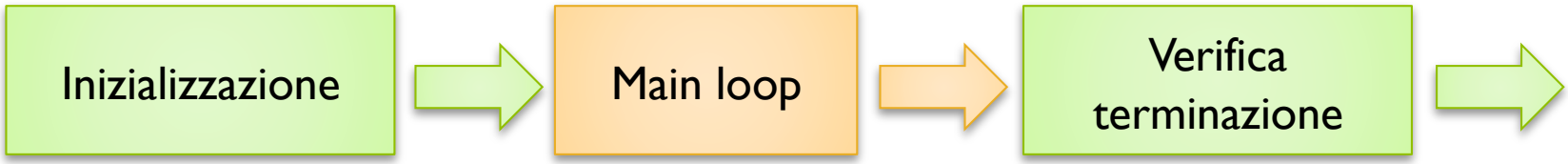
FPF con un esempio



FPF con un esempio



K-means

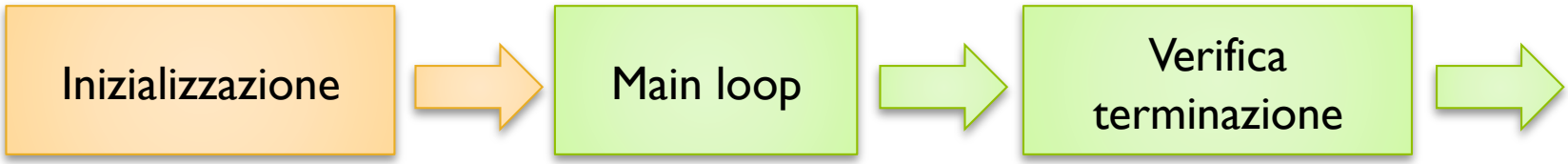


Algoritmo k-means

- Assegna ogni punto al centroide più vicino
- Aggiorna il centroide
  1. Per ogni dimensione del punto calcola il valore intermedio

- In uno spazio euclideo equivale al baricentro

K-means

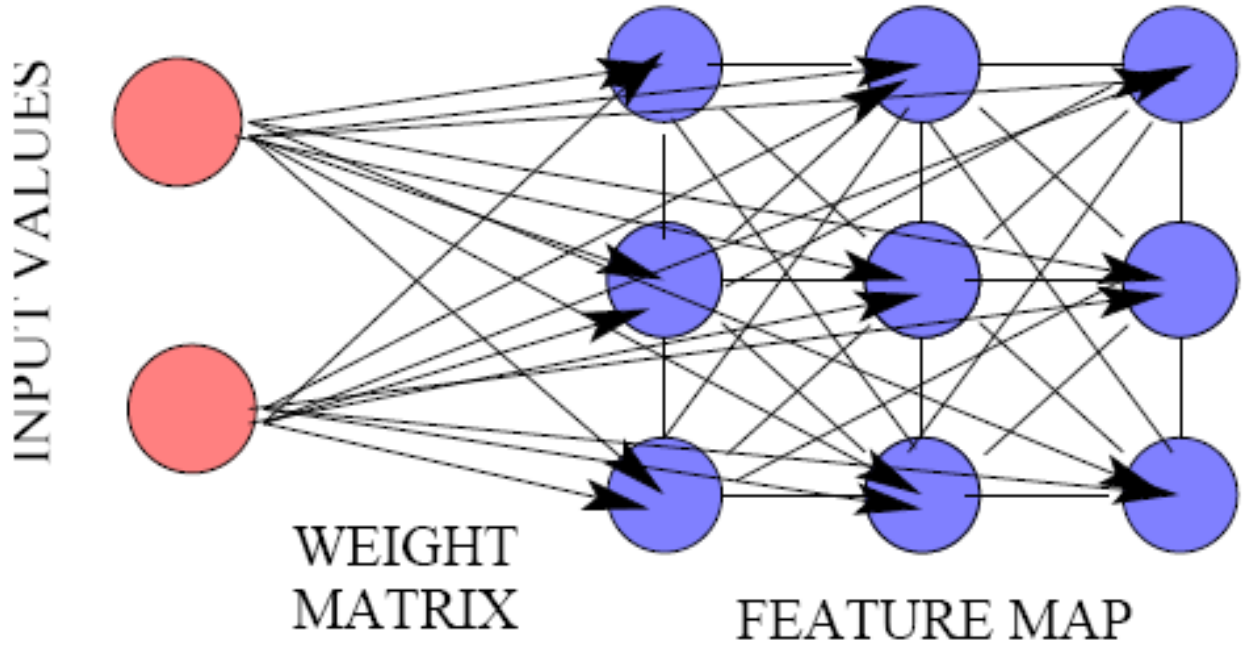


Inizializzazione

- Punti casuali
- Centri calcolati con FPF
- One pass k-means (McQueen)
  1. Centroidi inizialmente scelti casualmente
  2. Aggiungo un punto al cluster con il centroide più vicino
  3. Aggiorno il centroide
  4. Ricomincio dal punt 2 finche non finisco i punti da assegnare

Self organizing maps

- Rete neurale Single layer feed-forward
- Un vettore dei pesi per ogni neurone



Self organizing maps

- Inizializzazione:
  - Valori casuali piccoli
- Lavora in due fasi
  - **Training**: modifica i pesi della rete utilizzando degli esempi
    - Se non sono disponibili si usano come esempi i dati di input
  - **Mapping**: assegna i dati di input ai nodi in base ai pesi imparati
    - Confronta vettore dei pesi con vettore del dato in input.

Quanti cluster fare?

- Dipende dal problema in esame
  - Se sono pochi il contenuto non è omogeneo
  - Se sono troppi si rischia di "spezzare" cluster omogenei
- Dunque?
  - Soluzioni dipendenti dal problema
    - Possono non essere possibili
  - Approcci teorici
    - Sempre disponibili
    - Indipendenti dal problema

Esempio di uso del clustering: analisi associativa

- Permette di identificare condizioni che si verificano contemporaneamente con elevata frequenza
- Rileva pattern che si ripetono su determinati attributi e ne deriva regole di implicazione del tipo A ⇒ B
- Esempi
  - compra(farina, biscotti) ⇒ compra(latte)
  - compra(X, "divano 2 posti") ⇒ compra (X, "poltrona")
  - fatturato (X, "> 100M") ∧ struttura(X, "Spa") ⇒ compra(X, "Jaguar")
- Applicazioni
  - market basket analysis
  - profili clienti (abitudini di acquisto)
  - ottimizzazione delle manutenzioni

Significatività delle associazioni

- Viene valutata in base a:
  - **Confidenza**: misura la certezza del pattern
  - **Supporto**: misura la frequenza con cui il pattern è presente sulla base di dati
- Esempio:
  - Compra(X, "divano 2 posti") ⇒ Compra (X, "poltrona") [c.85%;s.30%]
  - L'85% di tutti coloro che comprano un divano 2 posti compra anche una poltrona
  - Nel 30% delle vendite il cliente ha comprato sia un divano a due posti che una poltrona

Analisi delle associazioni: il problema del carrello

- Data la registrazione delle "transazioni" di un supermercato:
  - una transazione è un insieme di oggetti acquistati contemporaneamente da un utente
- trovare gli oggetti che più di frequente sono stati acquistati insieme

TID	CID	Data	Prod.	Q.tà
111	201	5/1/05	farina	2
111	201	5/1/05	lievito	1
111	201	5/1/05	latte	3
111	201	5/1/05	carne	6
112	105	7/1/05	farina	1
112	105	7/1/05	lievito	1
112	105	7/1/05	latte	2
113	106	7/1/05	farina	2
113	106	7/1/05	latte	1
114	201	8/1/05	farina	3
114	201	8/1/05	lievito	2
114	201	8/1/05	carne	6
114	201	8/1/05	vino	6

Classificazione

Classificazione e predizione

- Costruzione di modelli per
  - Predire gli eventi futuri
  - Stimare il valore di elementi non noti
- Classificazione
  - Definizione di criteri che permettono di assegnare un soggetto ad una classe
- Predizione
  - Calcolo di funzioni di tendenza continue tramite l'interpolazione dei dati noti

Classificazione e predizione

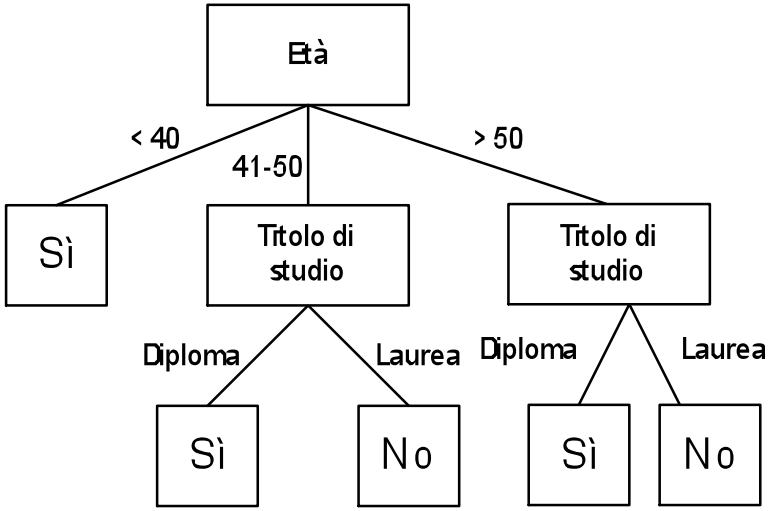
- Costruzione "basata su esempi"
  - Il modello deriva da un sottoinsieme significativo dei dati esistenti
  - L'efficacia viene testata su un sottoinsieme diverso (disgiunto) dei dati
  - Se il modello si rivela efficace può essere usato come 'predittore'
- Applicazioni
  - Propensione all'acquisto dei clienti
  - Qualità dei fornitori
  - Affidabilità dei prodotti

Classificazione

- Permette di indicare l'appartenenza di un elemento ad una certa classe
- Diversi tipi di modelli
  - Analisi statistiche
  - Regole associative
  - **Alberi di decisione**
  - Reti bayesiane (Naïve Bayes ne è un caso)
  - Reti neurali

Alberi di decisione

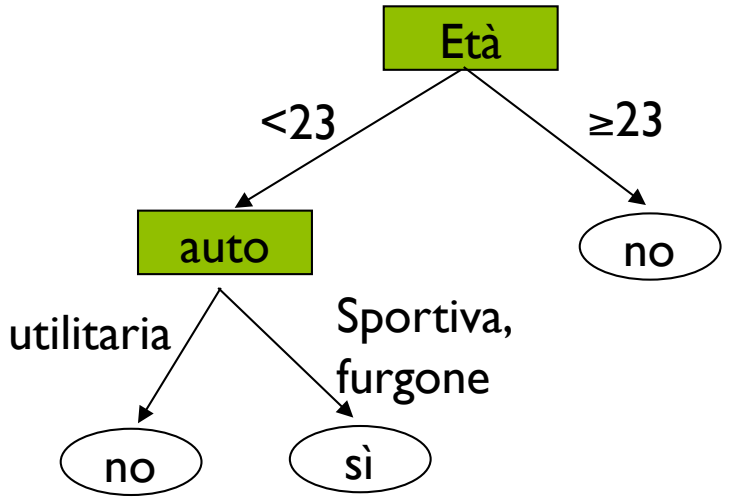
- Struttura di classificazione basata sulla valutazione di condizioni del tipo if-then-else
- Nodi interni
  - Attributi del soggetto da classificare
- Archi in uscita
  - Etichettati con i valori che l'attributo può assumere
- Nodi foglia
  - Classi
- La classificazione avviene seguendo un percorso guidato dai valori assunti dagli attributi dell'elemento da classificare



Esempio di albero di decisione che suddivide in due classi

Alberi di decisione

- Rappresentano un insieme di regole che permettono di fare la predizione automaticamente
- Sono costruiti automaticamente usando i dati disponibili
- Altro esempio: le caratteristiche di rischio dei clienti dell'assicurazione



Alberi di decisione: creazione

- La realizzazione dell'albero si basa su due fasi
  - Costruzione
  - Raffinamento
- Costruzione
  - Si cerca un buon criterio C per dividere dataset nei sottoinsiemi D1, D2
    - un buon criterio dovrebbe minimizzare la profondità dell'albero
  - Si costruisce un nodo che usa il criterio C e si applica ricorsivamente l'algoritmo a D1 e D2
- Raffinamento
  - L'albero costruito viene semplificato eliminando i rami meno importanti

Valutazione piattaforme di data mining

- Sia i sistemi di clustering che quelli di classificazione devono essere valutati prima di affidare le decisioni strategiche aziendali al sistema
- La bontà del sistema dipende:
  - Dal modello
  - Dagli algoritmi usati
  - Dal numero di cluster (solo clustering)
  - Dal training (solo classificazione)

Strategie di valutazione

Misure interne

- Modelli matematici
- Sempre disponibili
- Non correlati al problema

Misure esterne

- Confronto con il ground truth
- Spesso non disponibili
- Correlate al problema

Misure interne: omogeneità e separazione

- Gruppi molto omogenei indicano buon clustering/ classificazione
- Gruppi molto separati indicano buon clustering/ classificazione

$$Homogeneity = \frac{1}{\#mates} \sum_{o_i, o_j \ mates, i<j} Sim(o_i, o_j)$$

$$Separation = \frac{2}{n(n-1) - 2\#mates} \sum_{o_i, o_j \ non-mates, i<j} Sim(o_i, o_j)$$

- Due oggetti sono mates se appartengono allo stesso gruppo
- Alta omogeneità, bassa separazione indicano migliori risultati

Misure esterne: precision e recall per raggruppamenti

- Dato un raggruppamento c_j ed una classe GT_i si definiscono

$$precision(GT_i, c_j) = \frac{|GT_i \cap c_j|}{|c_j|}$$

$$recall(GT_i, c_j) = \frac{|GT_i \cap c_j|}{|GT_i|}$$

- La precision misura la probabilità che un elemento della classe GT_i sia anche nel gruppo c_j
- La recall misura la probabilità che un elemento nel gruppo c_j sia anche della classe GT_i

F-measure

- Dato un clustering/classificazione ed il suo ground truth,precision e recall possono essere combinate insieme:
- F-measure per una classe

$$F(GT_i, c_j) = 2\frac{precision(iGT, c_j)recall(GT_i, c_j)}{precision(GT_i, c_j) + recall(GT_i, c_j)}$$

- Estensione all'intero raggruppamento

$$F = \sum_i \frac{|GT_i|}{n} \max_j (F(GT_i, c_j)),$$