Knowledge Discovery in Databases

Knowledge Discovery in Databases

- “E’ il processo non banale di identificazione di pattern nei dati che siano validi, nuovi, potenzialmente utili e comprensibili” [Fay96]
- Dati: un insieme di record F in un database
- Pattern: una espressione E in un linguaggio L che descrive i fatti in un sottoinsieme F_E di F. E e’ chiamato un pattern se è piu’ semplice della enumerazione dei fatti di F_E
- Processo: e’ composto da più passi. Deve essere non banale, ovvero deve richiedere una ricerca o un’inferenza, non può essere il calcolo di una quantità predefinita
- Validi: i pattern scoperti devono essere validi su nuovi dati con un certo grado di certezza

Knowledge Discovery in Databases

- Nuovi: i pattern scoperti non devono essere precedentemente noti
- Potenzialmente utili: devono potenzialmente condurre a azioni utili
- Comprensibili: i pattern devono essere comprensibili a un essere umano in modo da facilitare la comprensione dei dati sottostanti.
- Validità, novità, utilità e comprensibilità devono poter essere misurate mediante funzioni di E ed F
- Per la comprensibilità, si possono adottare misure di semplicità che possono essere sintattiche (ad esempio il numero di bit di un pattern) o semantiche

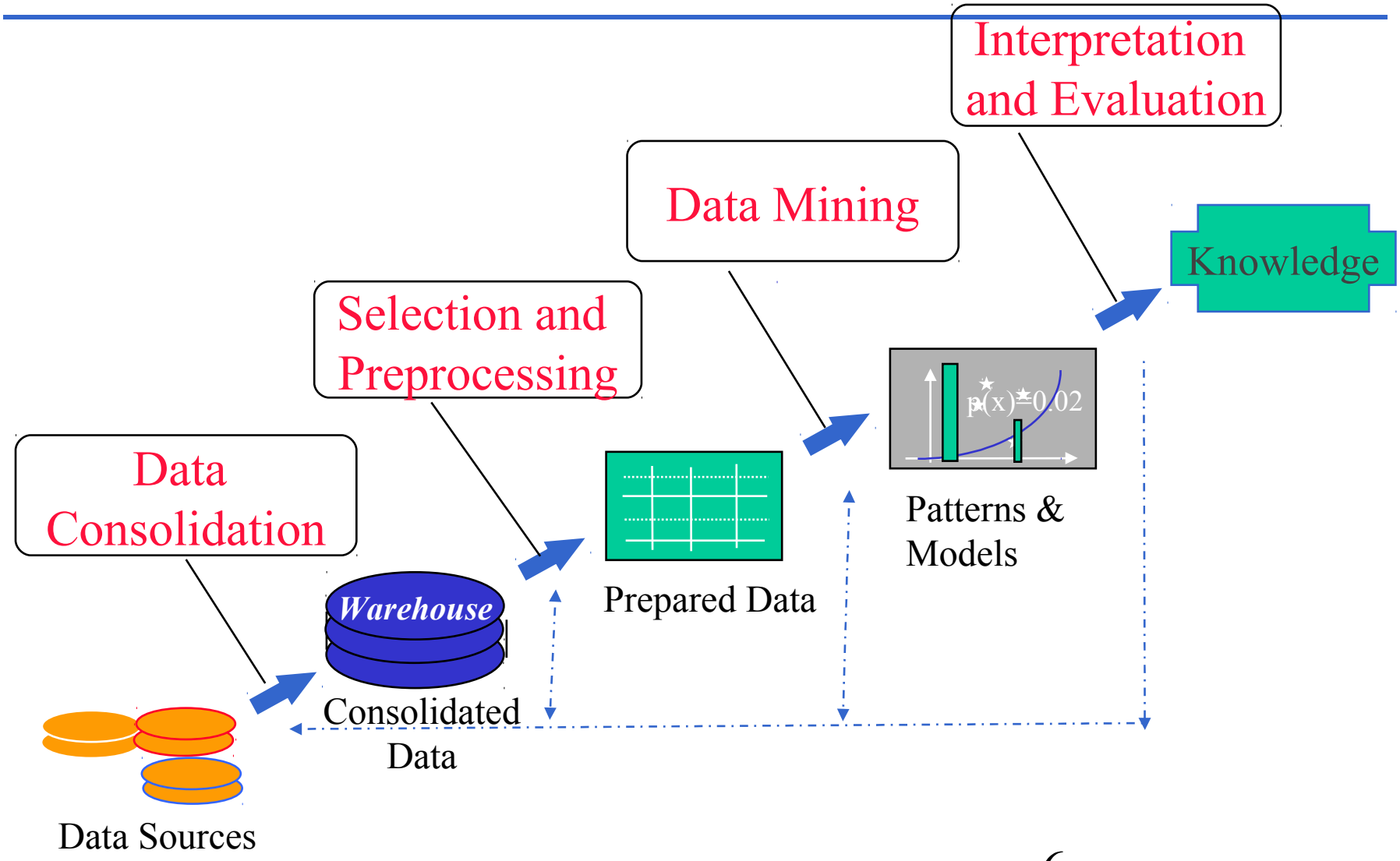
Knowledge Discovery in Databases

- Interesse: misura complessiva del valore di un pattern. E' una funzione della validità, novità, utilità e semplicità (oltre che dei dati e del pattern) i=I(E,F,V,N,U,S).
- Definizione di conoscenza: un pattern $E \in L$ è chiamato conoscenza se, per qualche soglia k specificata dall'utente I(E,F,V,N,U,S)>k

Data mining

- E' un passo del processo di KDD che consiste nell'applicare algoritmi che estraggono pattern E_j da F sotto limitazioni accettabili del tempo di calcolo.
- Lo spazio dei possibili pattern è spesso infinito e l'estrazione dei pattern richiede una qualche forma di ricerca
- Il data mining non è che un passo del KDD
- Possiamo quindi anche definire il process di KDD come:
  - "Il processo dell'impiego di metodi di data mining per estrarre ciò che viene definito conoscenza secondo la definizione delle misure e delle soglie, usando il database F insieme ad ogni necessario preprocessing, campionamento e trasformazione di F." [Fay96]

Knowledge Discovery in Databases



Processo di KDD

- Contiene diversi passi:
  1. Sviluppo di una comprensione del dominio applicativo, della conoscenza a priori rilevante e degli obiettivi dell'utente finale
  2. Consolidamento dei dati
  3. Selezione e preprocessing
  4. Scelta del compito di data mining: scelta tra classificazione, regressione, clustering, ecc.
  5. Scelta dell'algoritmo di data mining
  6. Data mining: applicazione dell'algoritmo
  7. Interpretazione e valutazione dei pattern e possibile ritorno ai passi da 1 a 6 per ulteriori iterazioni

Processo di KDD

8. Consolidamento della conoscenza scoperta: incorporamento della conoscenza all'interno di un sistema oppure all'interno di un documento da mostrare all'utente. Questo include anche la ricerca e la risoluzione di potenziali conflitti con la conoscenza precedentemente creduta.

- Il processo di KDD
  - è iterativo in quanto i passi possono essere ripetuti
  - è interattivo perché richiede l'intervento dell'analista per molte decisioni e per le scelta di vincoli da imporre agli algoritmi.

Consolidamento

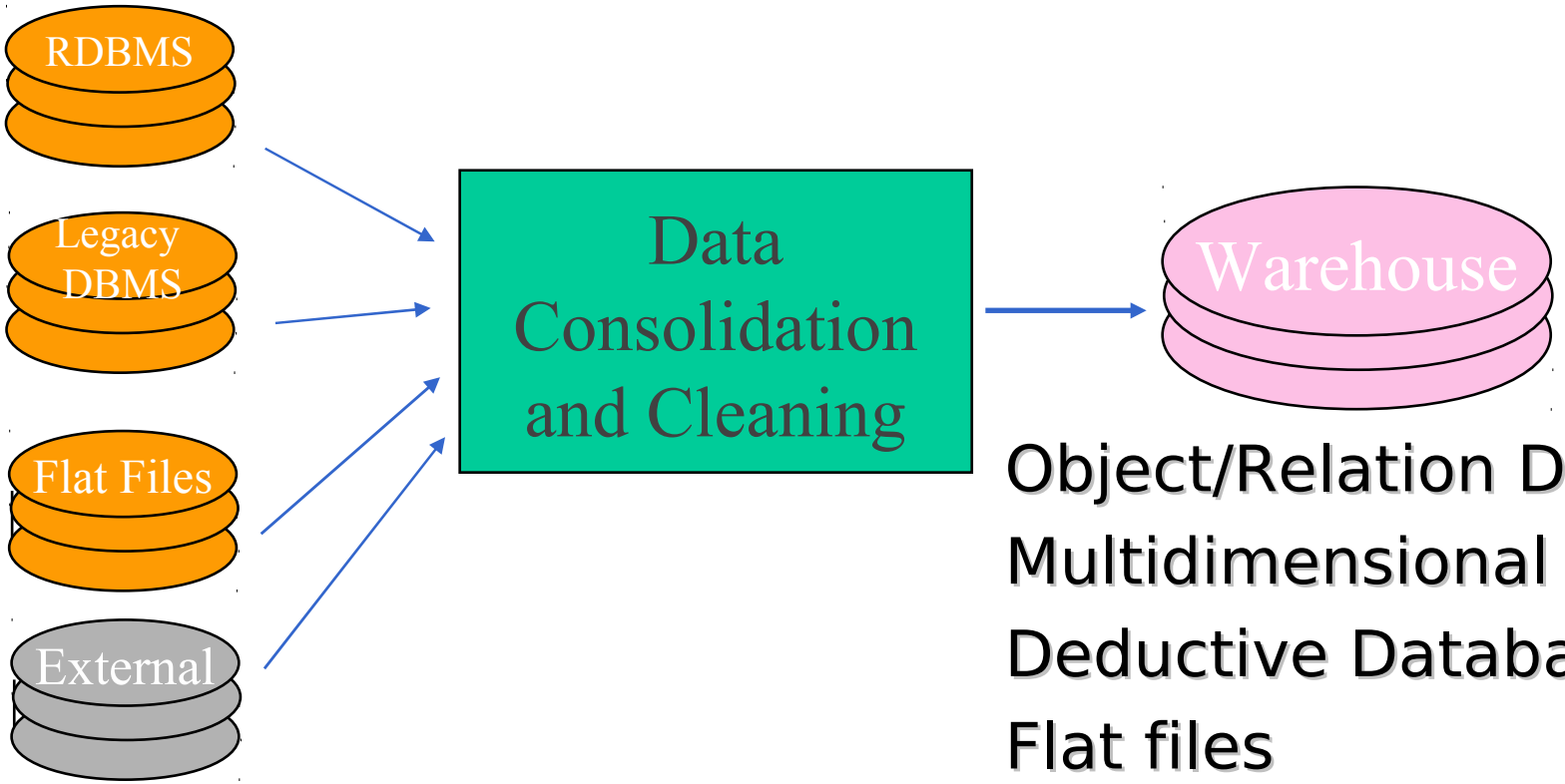
- Garbage in => garbage out
- La qualità dei risultati è in relazione diretta con la qualità dei dati
- Il 50%-70% degli sforzi sul processo di KDD viene speso nella preparazione e nel consolidamento dei dati
- Importante giustificazione per Data Warehouses Aziendali

Consolidamento

- Determinazione di un insieme preliminare di attributi
- Consolidamento dei dati in un database
  - sorgenti interne ed esterne
- Scelta della strategia da adottare per i valori nulli (eliminazione, stima o nessun trattamento)
- Rimozione di outliers e rumore (eccezioni ovvie, picchi)

Consolidamento

Da sorgenti eterogenee a data repositories consolidati



Object/Relation DBMS
Multidimensional DBMS
Deductive Database
Flat files

Selezione e preprocessing

- Selezione di un campione di record nel caso in cui sia impossibile utilizzare l'intero database
- Riduzione della dimensionalità degli attributi
  - Rimozione di attributi ridondanti e/o correlati
  - Combinazione di attributi (somma, moltiplicazione, differenza)
- Riduzione dei domini degli attributi
  - Raggruppamento di valori per gli attributi discreti
  - Quantizzazione degli attributi continui

Selezione e preprocessing

- Trasformazione dei dati: normalizzazione dei valori, ad es. nelle reti neurali l'ingresso deve essere compreso all'interno di un dominio di valori tra 0 e 1 oppure tra –1 e 1

- Codifica dei dati
  - La rappresentazione deve essere adeguata al tool di data mining che verrà usato

Compiti del Data Mining

- Predizione:
  - Classificazione: apprendimento di una funzione che assegni ad ogni oggetto una classe da un insieme predefinito di classi (ad esempio, apprendimento di alberi di decisione, regole di produzione)
  - Regressione: apprendimento di una funzione che assegni a ogni oggetto un valore reale (ad esempio, regressione lineare)

Compiti del Data Mining

- Descrizione:
  - Clustering: scoperta di sottogruppi di dati tali che i dati all'interno di uno stesso sottogruppo siano simili tra loro e siano dissimili da quelli negli altri gruppi
  - Scoperta di regole associative: scoperta di regole che descrivono regolarità nei dati

Bibliografia

- [Fay96] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *From Data Mining to Knowledge Discovery: An Overview*, in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (editori) Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, 1996.