Corso di Laurea Magistrale in Economia

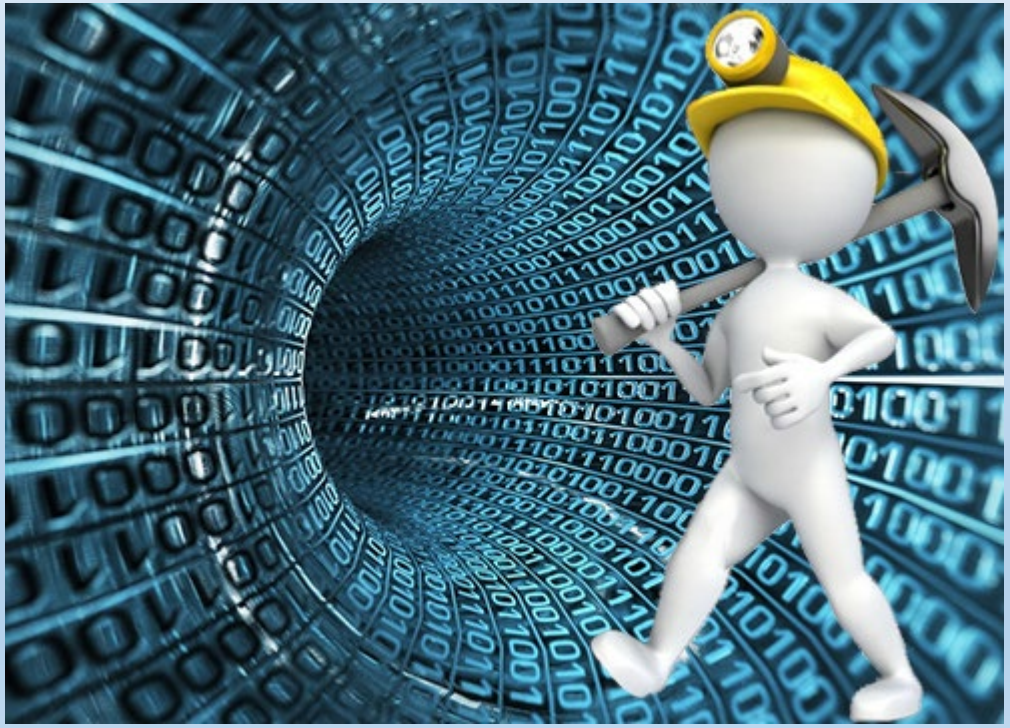
Data Science

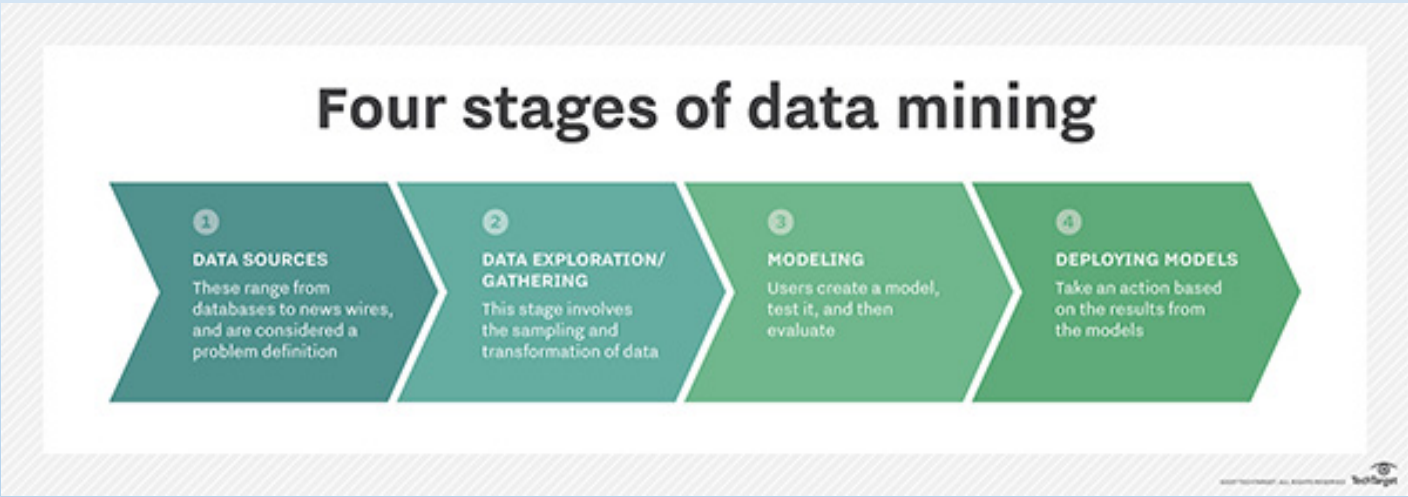
A.A. 2018/2019

Lez. 5 –Data Mining

Data Mining

- Processo di esplorazione e analisi di un insieme di dati, generalmente di grandi dimensioni, per individuare eventuali regolarità, estrarre conoscenza e ricavare regole ricorrenti significative





Data Mining

- Il processo di *Data Mining* si basa su metodi di «apprendimento induttivo», che si propongono di ricavare regole generali a partire da una sequenza di esempi (osservazioni passate).
- Le analisi di *Data Mining* si propongono di trarre conclusioni a partire da un campione di osservazioni e di generalizzare le conclusioni all'intera popolazione.
- Spesso la raccolta dei dati avviene indipendentemente dalle analisi di *Data Mining,* pertanto non segue schemi di campionamento predeterminati.

Orientamenti di indagine

- **Interpretazione**
  - Individuare schemi di regolarità e criteri comprensibili
  - Generazione di regole originali per incrementare il livello di conoscenza del sistema analizzato
  - Esempio: azienda GDO che raggruppa i clienti in base al profilo d'acquisto per evidenziare nuove nicchie di mercato e indirizzare le promozioni
- **Predizione**
  - Prevedere il valore che una variabile casuale assumerà in futuro oppure stimare la probabilità di accadimento di eventi futuri
  - Predizioni sulla base del valore di alcune variabili associate a entità presenti in un database
  - Esempio: valore delle vendite di un prodotto nelle settimane successive
- In alcuni casi, un modello sviluppato per la predizione può essere efficace per l'interpretazione

Modelli e metodi di *data mining*

- Esistono numerose metodologie che possono essere impiegate per obiettivi di *data mining*:
  - *Machine learning*
  - *Knowledge discovery in database*
  - *Statistical learning theory* (calcolo delle prob.+ottimizzazione+statistica)

- Passi per lo sviluppo di un modello di *data mining*
  - Scelta classe di modelli
  - Definizione metrica di valutazione di efficacia e accuratezza
  - Progettazione algoritmo di calcolo

Data Mining, statistica classica e OLAP

- Ruolo attivo dei modelli di *data mining*, rispetto alle tecniche statistiche e OLAP.
- In Statistica, il decisore ha formulato un'ipotesi che cerca di confermare attraverso l'analisi campionaria
- Nelle analisi OLAP, il decisore esprime ipotesi sulle quali basare i criteri di estrazione e visualizzazione.
  - Flusso *top-down*
- I modelli di *data mining* generano predizioni e interpretazioni che costituiscono nuova conoscenza
  - Flusso *bottom-up*

Data Mining, statistica classica e OLAP

OLAP	statistics	data mining
extraction of details and aggregate totals from data	verification of hypotheses formulated by analysts	identification of patterns and recurrences in data
information	validation	knowledge
distribution of incomes of home loan applicants	analysis of variance of incomes of home loan applicants	characterization of home loan applicants and prediction of future applicants

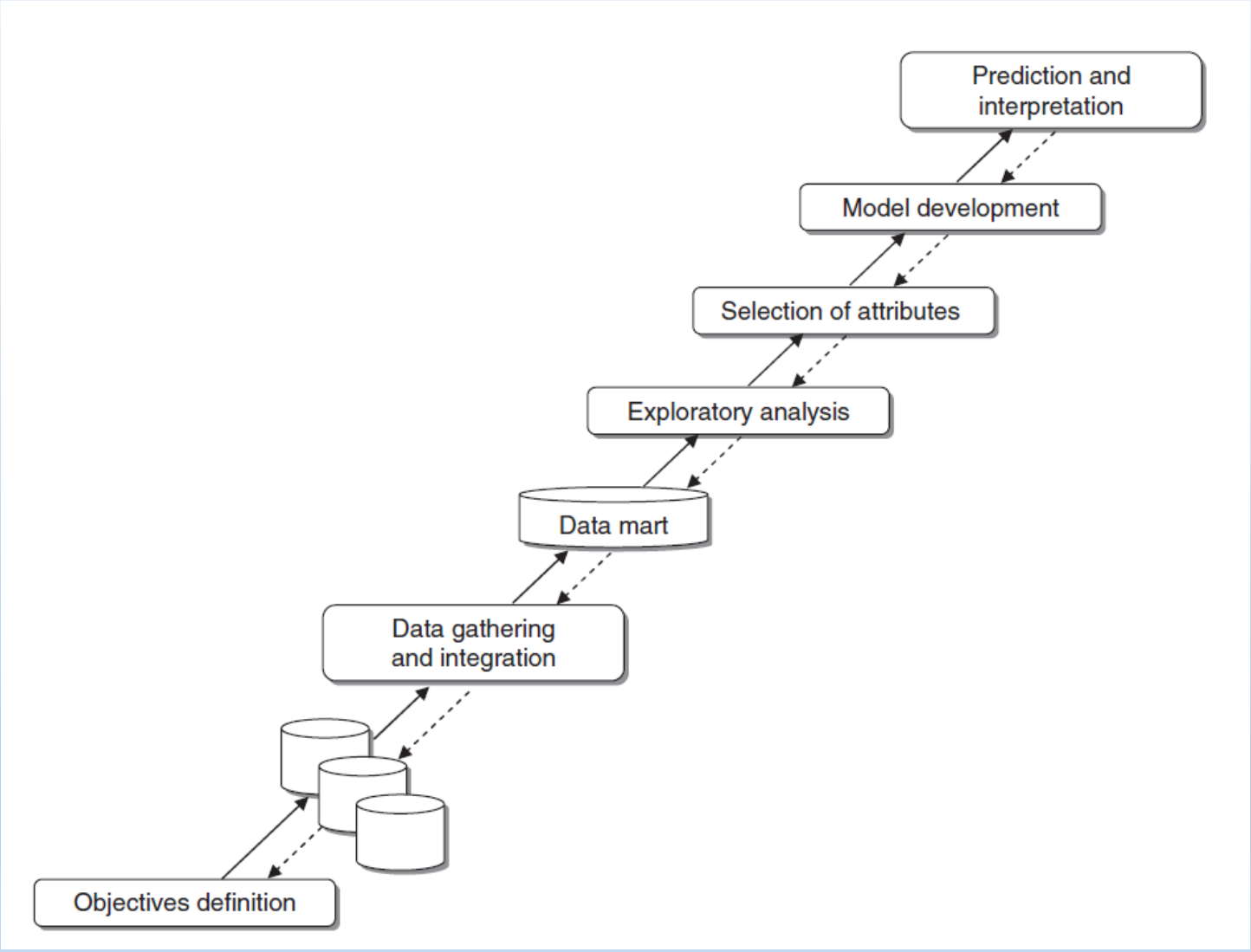
Applicazioni di *Data Mining*

- Marketing relazionale
  - Identificazione segmenti di clienti che risponderanno ad azioni di vendita mirate (*cross-selling*)
  - Identificazione segmenti di clienti cui rivolgere azioni di *retention*
  - Stima tasso di risposte positive a campagne di marketing
  - Interpretazione comportamenti d'acquisto
  - Analisi delle combinazioni di acquisto dei clienti *(market based analysis)*
- Identificazione di frodi
  - Assicurative (ad es., denuncia sinistri fasulli)
  - Bancarie (ad es., utilizzo illegale carte di credito)
- Valutazione del rischio
  - Ad esempio, modello di apprendimento per stabilire sulla base delle caratteristiche del richiedente il rischio legato a un mutuo

Applicazioni di *Data Mining*

- Text mining
  - Analisi applicate a testi di varia natura (dati non strutturati) per classificare articoli, libri, documenti
  - Filtri per messaggi email
- Riconoscimento immagini
  - Riconoscimento caratteri di scrittura
  - Identificare volti
  - Segnalare comportamenti anomali mediante videocamere
- Web mining
  - Analisi *clickstream* (sequenza pagine visitate) (utile ad es. per e-commerce)
- Diagnostica medica
  - Analisi risultati test clinici

Processo di *Data Mining*



Definizione degli obiettivi

- Le analisi si svolgono in specifici contesti applicativi
- Necessità di formulare obiettivi di indagine plausibili e ben definiti
- Se il problema non è ben delineato si rischia di vanificare l'attività
- Collaborazione tra esperti del contesto e analisti di *data mining*

Raccolta dati ed integrazione

- Diverse fonti
- Necessità di integrazione
- In alcuni casi, dati già disponibili in *data warehouse*
- Pericolo di eccessiva aggregazione
- Dataset estratti in accordo a distribuzioni ignote

Analisi esplorativa

- Analisi preliminare dei dati, per conoscenza delle informazioni disponibili e per validazione
- Verifica della distribuzione dei valori per ciascun attributo, per evidenziare anomalie e valori mancanti
- Ad esempio, date di nascita fuori dal range di valori ammissibili, valori di spesa negativi, ecc.

Selezione degli attributi

- Valutazione della rilevanza degli attributi rispetto agli obiettivi.
- Rimozione attributi non rilevanti
- Passaggio cruciale per l'efficacia delle analisi

Sviluppo e validazione dei modelli

- *Training* dei modelli utilizzando un campione di record estratti dal dataset
- Valutazione dell'accuratezza predittiva sul resto dei dati
- Articolazione del dataset:
  - *Training set*, di dimensioni contenute, per identificare uno specifico modello all'interno di una classe
  - *Test set*, per valutare l'accuratezza del modello identificato

Predizione e interpretazione

- Modello prescelto incorporato nelle procedure di supporto alle decisioni
- Cicli di retroazione per modificare scelte precedenti
- Coinvolgimento diverse figure:
  - Esperto dell'ambito di applicazione
  - Esperto di sistemi informativi aziendali
  - Esperto di teoria dell'apprendimento e statistica

Metodologie di analisi

Distinzione in base alla presenza di una variabile target

- **Apprendimento supervisionato** (analisi dirette)
  - Presenza di un attributo target che per ciascun record rappresenta la classe di appartenenza oppure una grandezza misurabile
  - Processi orientati alla predizioni e all'interpretazione in riferimento ad un target

- **Apprendimento non supervisionato** (analisi indirette)
  - Le analisi mirano a individuare ricorrenze, affinità e difformità presenti nel *dataset*
  - Possibilità di individuare raggruppamenti di record (*cluster*) che risultano omogenei

Metodologie di analisi

Funzionalità principali:

1) **Caratterizzazione e discriminazione**
2) **Classificazione**
3) **Modelli di stima**
4) **Modelli di serie storiche**
5) **Regole associative**
6) **Clustering**
7) **Descrizione e visualizzazione**

Caratterizzazione e discriminazione

- In presenza di un attributo target, prima di procedere allo sviluppo di modelli di classificazione risulta utile un'analisi di natura esplorativa, avente un duplice scopo:
  - «Caratterizzazione» confrontando la distribuzione dei valori degli attributi per i record appartenenti ad una medesima classe
  - «Discriminazione» mediante il confronto tra la distribuzione dei valori degli attributi per i record di una classe e i record di un'altra
- Le informazioni così acquisite vengono di solito presentate agli utilizzatori mediante semplici grafici (ad es. istogrammi)
- Il valore delle informazioni generate può guidare la fase di selezione degli attributi

Classificazione

- Disponibilità di un insieme di osservazioni (record) di cui è nota la classe di appartenenza
- Ogni osservazione è descritta mediante un numero di attributi il cui valore è noto
- Un algoritmo di classificazione utilizza le osservazioni disponibili (passate) per identificare una funzione che consenta di assegnare la classe di appartenenza alle osservazioni future, di cui siano noti i valori degli attributi
- L'attributo target, il cui valore deve essere predetto, assume un numero finito e piuttosto limitato di valori (anche binario)

Modelli di stima

- Vengono utilizzati quando la variabile target assume valori continui
- Sulla base degli attributi esplicativi disponibili si cerca di predire il valore della variabile target per ciascuna osservazione
- Un problema di classificazione può essere ricondotto ad un problema di stima e viceversa
- Ad esempio, una società di telefonia mobile, interessata alla classificazione dei clienti sulla base della fedeltà, può ricavare un problema di stima valutando la probabilità che ciascun cliente rimanga fedele

Modelli di serie storiche

- Serie storica: sequenza di valori della variabile target nel tempo
- I modelli di serie storiche studiano fenomeni caratterizzati da una dinamica temporale e si propongono di predire il valore della variabile target per uno o più periodi futuri

Regole associative

- Mediante le regole associative si mira a identificare associazioni ricorrenti tra gruppi di record in un dataset
- Ad esempio, le aziende della GDO le utilizzano per pianificare la disposizione dei prodotti sugli scaffali per promuovere le vendite «incrociate» (*cross-selling*)

Clustering

- Segmentazione di una popolazione eterogenea in un certo numero di sottogruppi, contenenti osservazioni aventi caratteristiche affini
- Le osservazioni di *cluster* differenti presentano elementi caratteristici *distintivi*
- Nel clustering non eistono classi predefinite o esempi di riferimento che indicano l'appartenenza a una certa classe
- Gli oggetti vengono raggruppati in base alla loro reciproca omogeneità

Descrizione e visualizzazione

- La descrizione efficace e sintetica delle informazioni è molto utile, in quanto può offrire suggerimenti per una spiegazione delle relazioni tra i dati e costruire un punto di partenza dei fenomeni
- Non sempre è facile costruire una visualizzazione significativa dei dati
- Lo sforzo di rappresentazione è giustificato dalla notevole capacità di sintesi delle informazioni ottenuta mediante un grafico ben progettato.