Informazioni sul corso

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

Modalità didattiche e materiale

- Lezioni in aula e in laboratorio utilizzando il software open source Weka
- Il corso è composto da due moduli
  - ✓ Data mining (36 ore): introduce i concetti di base, descrive le tecniche di mining da applicare a dati strutturati
  - ✓ Text mining (18 ore): descrive come le tecniche di mining devono essere specializzate per operare efficacemente su dati testuali

Modalità didattiche e materiale

- Lezioni in aula e in laboratorio utilizzando il software open source Weka
- Tutti gli argomenti del corso coperti dalle slide scaricabili dal sito del docente
- Il **libro di testo** per il modulo di **data mining** è:
  - ✓ Pang-Ning Tan, Michael Steinbach, Vipin Kumar *Introduction to Data Mining*. Pearson International, 2006.
- Il **libro di testo** per il modulo di **text mining** è:
  - ✓ Christopher Manning, Hinrich Schutze, Prabhakar Raghavan. Introduction to Information Retrieval. Cambridge University Press, 2008. (disponibile on line)
- Ulteriori dettagli sul software Weka
  - ✓ Ian H. Witten and Eibe Frank*Data Mining: Practical Machine Learning Tools and Techniques, IInd Ed.*. Morgan Kaufmann, 2005.

Modalità di esame

- L'esame consta di un elaborato e da una prova orale su tutti gli argomenti del corso.
  - ✓ Durante la prova orale potrà essere richiesto l'utilizzo del software Weka
- La scelta dell'elaborato deve essere concordata con il docente
  - ✓ Determina un punteggio aggiuntivo ai fini dell'esame [0..4] punti
  - ✓ Implementazione di un algoritmo tra quelli presenti in letteratura
  - ✓ Analisi di un data set con tecniche di mining

Il profilo Data & Knowledge Engineering

- Il profilo studia la modellazione e gli algoritmi necessari alla costruzione e allo sfruttamento della conoscenza al servizio di applicazioni aziendali e scientifiche avanzate. Gli ambiti applicativi di riferimento sono:
  - ✓ Business Intelligence
  - ✓ Semantic Web
  - ✓ Internet-of-Things

Figure professionali DKE

- Data Scientist
- Progettista e consulente nel settore della Business Intelligence e degli Analytics
- Esperto di semantic web e di sistemi IoT
- Esperto delle tecnologie in ambito Big Data
- Project manager di progetti ad elevato contenuto tecnologi

Gli altri corsi del profilo DKE

- Big Data (Prof. Enrico Gallinucci)
- Business Intelligence (Prof. Rizzi)
- Project Management (Prof. Boschetti)
- Sistemi di Supporto alle Decisioni (Prof. Maniezzo)
- Semantic Web (Prof. Carbonaro)

- Esiste un accordo Erasmus specifico con l'Universidad Politecnicade Catalunya (Barcellona) in cui ha sede un master specializzato sulle tematiche proprie del prolo.

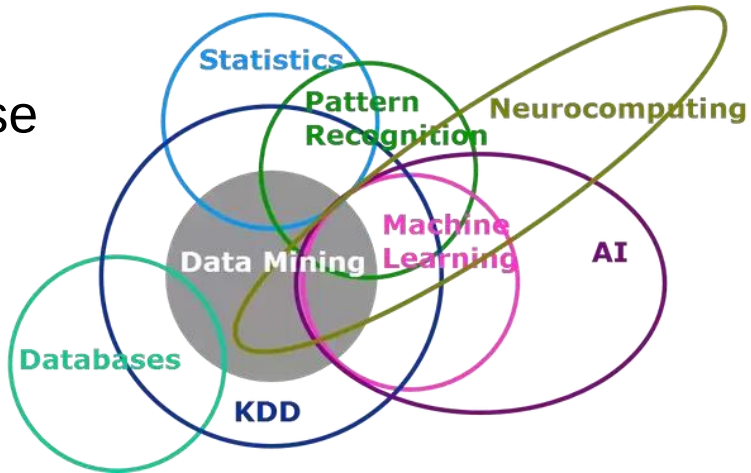
Introduzione al Data Mining

Prof. Matteo Golfarelli
Alma Mater Studiorum - Università di Bologna

AI, Machine Learning & Data Mining

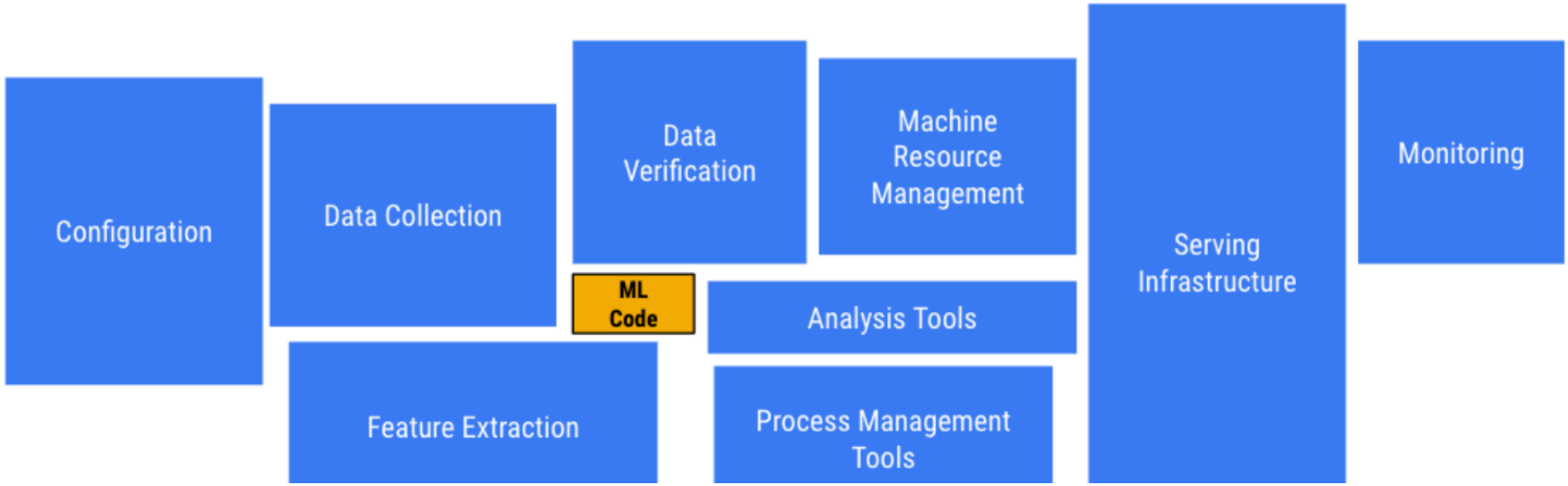
- Sebbene fortemente interrelati tra loro, il termine machine learning è formalmente distinto dal termine Data Mining con il quale si indica il processo computazionale di scoperta di pattern in grandi dataset utilizzando metodi di machine learning, intelligenza artificiale, statistica e basi di dati.

- A parte la fase di analisi vera e propria, il data mining copre aspetti di:
- Gestione del dato e pre-processing
- Modellazione
- Identificazione di metriche di interesse
- Visualizzazione



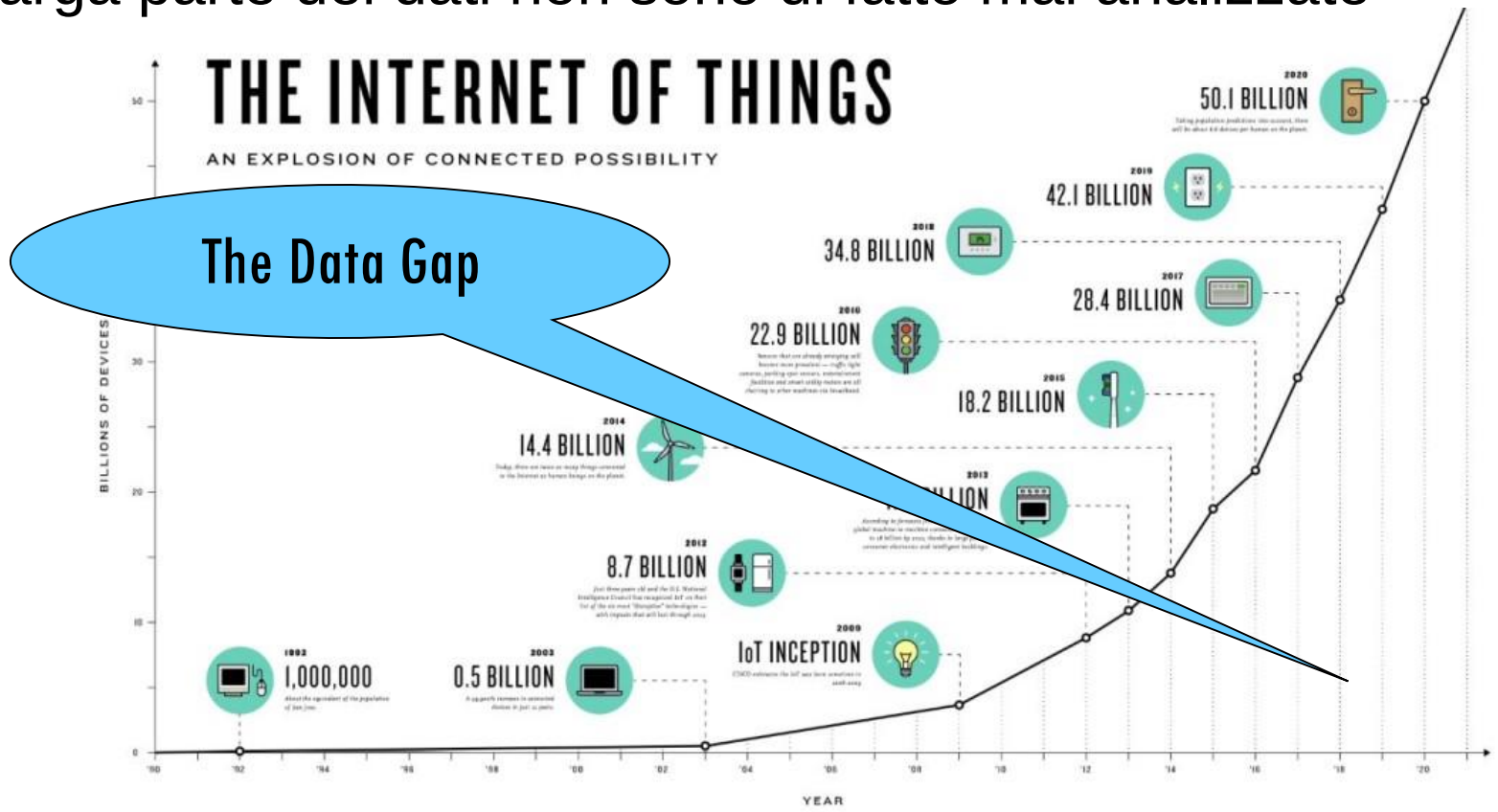
AI, Machine Learning & Data Mining

Il ruolo del Machine Learning in un progetto reale di data mining è reso bene dalla seguente immagine che elenca le attività necessarie. A rettangoli più grandi corrispondono alle attività a cui è dedicato più tempo.



Data mining su grandi data set

- Molte delle informazioni presenti sui dati non sono direttamente evidenti
- Le analisi guidate dagli uomini possono richiedere settimane per scoprire informazioni utili
- Larga parte dei dati non sono di fatto mai analizzate

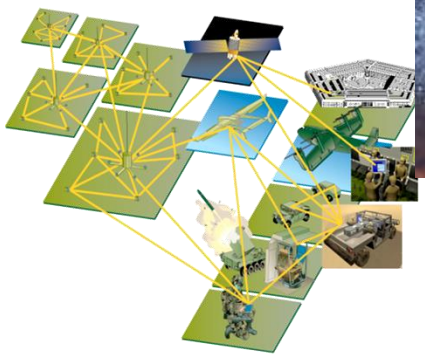


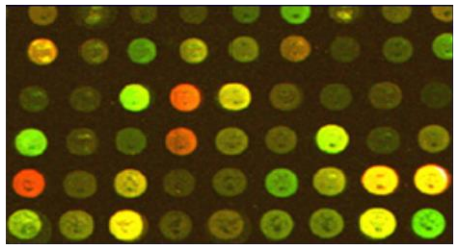
Perché fare data mining?

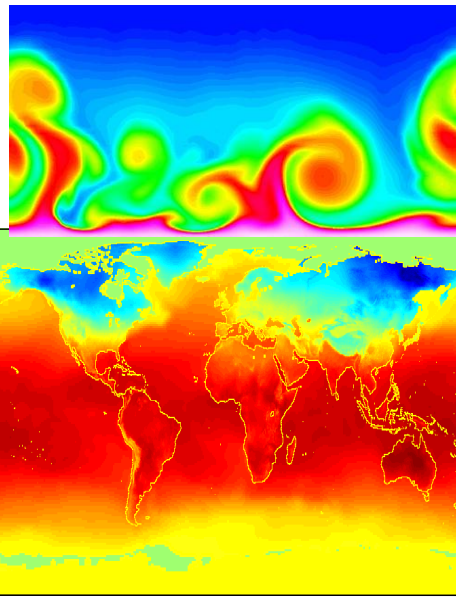
- La quantità dei dati memorizzata su supporti informatici è in continuo aumento
  - ✓ Pagine Web, sistemi di e-commerce
  - ✓ Dati relativi ad acquisti/scontrini fiscali
  - ✓ Transazioni bancarie e relative a carte di credito
- L'hardware diventa ogni giorno più potente e meno costoso
- La pressione competitiva è in continua crescita
  - ✓ La risorsa informazione è un bene prezioso per superare la concorrenza

Perché fare data mining?

- I dati prodotti e memorizzati crescono a grande velocità (GB/ora)
  - ✓ Sensori posti sui satelliti
  - ✓ Telescopi
  - ✓ Microarray che generano espressioni genetiche
  - ✓ Simulazioni scientifiche che producono terabyte di dati
- Le tecniche tradizionali sono inapplicabili alle masse di dati grezzi
- Il Data mining può aiutare gli scienziati a:
  - ✓ Classificare e segmentare i dati
  - ✓ Formulare ipotesi

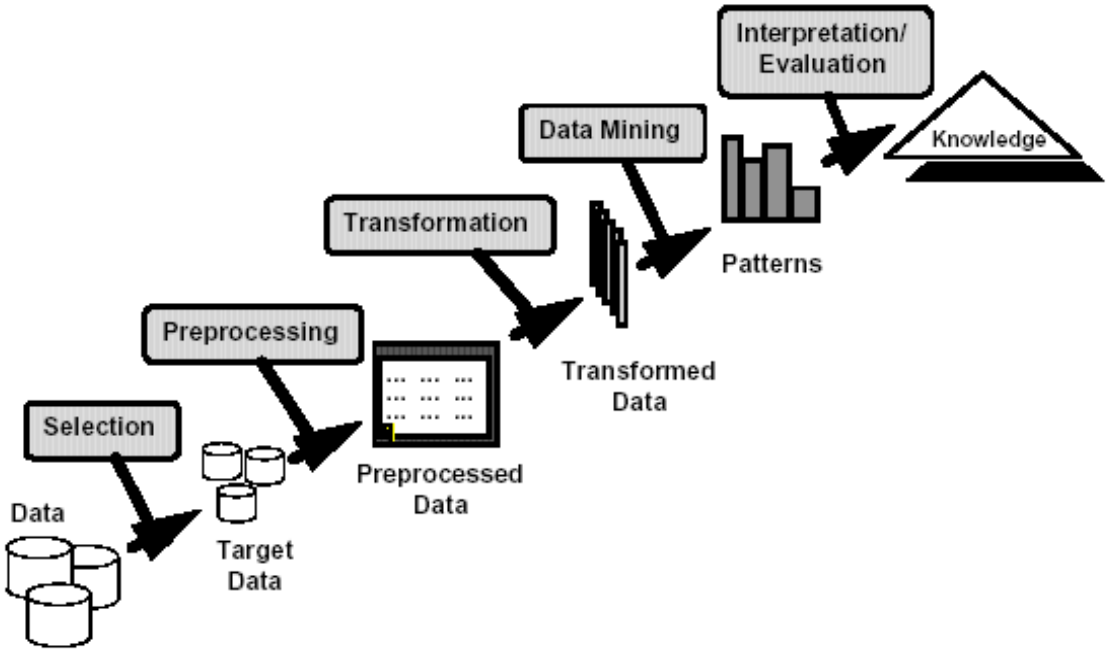






Cosa è il Data Mining?

- Alcune definizioni
  - ✓ Estrazione complessa di informazioni implicite, precedentemente sconosciute e potenzialmente utili dai dati.
  - ✓ Esplorazione e analisi, per mezzo di sistemi automatici e semi-automatici, di grandi quantità di dati al fine di scoprire pattern significativi

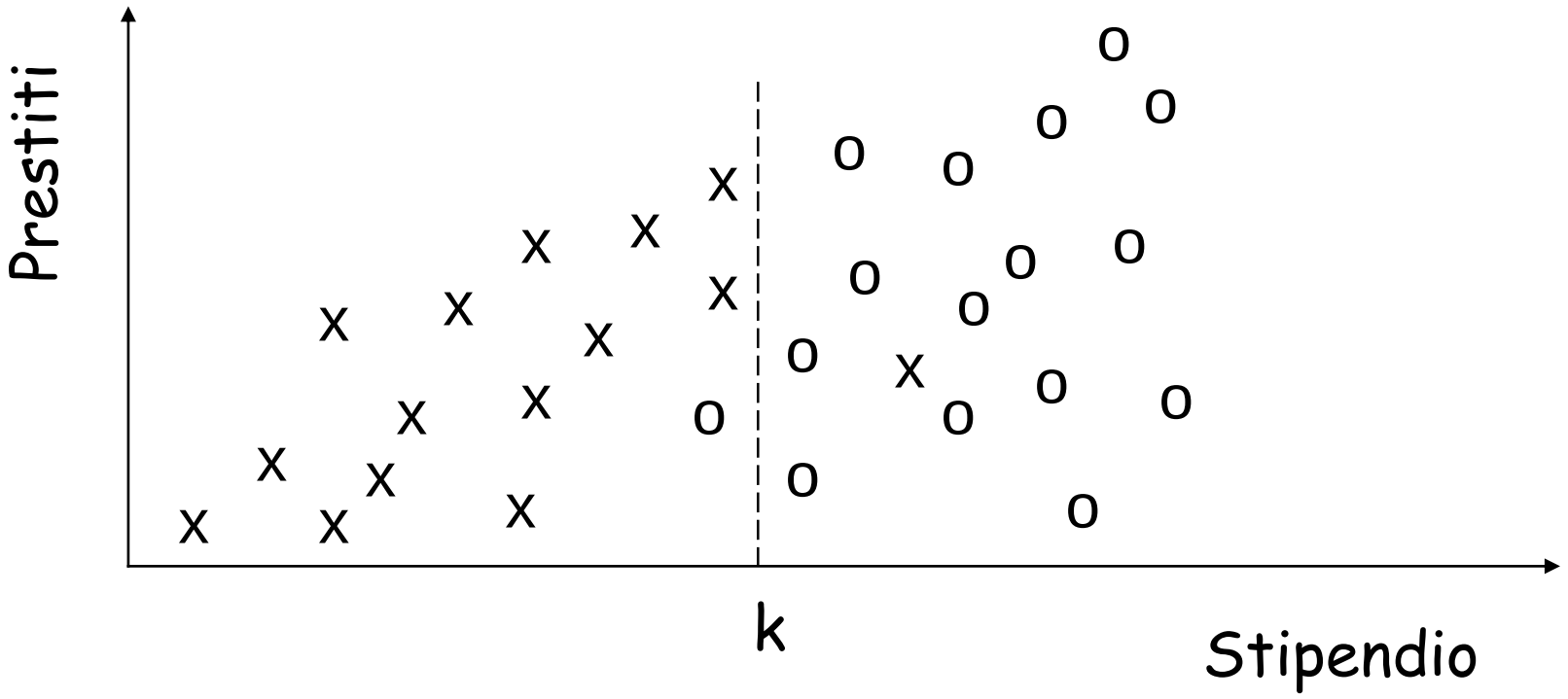


Pattern

- Un pattern è una rappresentazione sintetica e ricca di semantica di un insieme di dati; esprime in genere un modello ricorrente nei dati, ma può anche esprimere un modello eccezionale
- Un pattern deve essere:
  - ✓ Valido sui dati con un certo grado di confidenza
  - ✓ Comprensibile dal punto di vista sintattico e semantico, affinché l'utente lo possa interpretare
  - ✓ Precedentemente sconosciuto e potenzialmente utile, affinché l'utente possa intraprendere azioni di conseguenza

Esempio

Persone che hanno ricevuto un prestito
- x: hanno mancato la restituzione di rate
- o: hanno rispettato le scadenze



- Pattern:
  - ✓ **IF** stipendio < k **THEN** pagamenti mancati

Tipi di pattern

- Regole associative
  - ✓ consentono di determinare le regole di implicazione logica presenti nella base di dati, quindi di individuare i gruppi di affinità tra oggetti
- Classificatori
  - ✓ consentono di derivare un modello per la classificazione di dati secondo un insieme di classi assegnate a priori
- Alberi decisionali
  - ✓ sono particolari classificatori che permettono di identificare, in ordine di importanza, le cause che portano al verificarsi di un evento
- Clustering
  - ✓ raggruppa gli elementi di un insieme, a seconda delle loro caratteristiche, in classi non assegnate a priori
- Serie temporali
  - ✓ Permettono l'individuazione di pattern ricorrenti o atipici in sequenze di dati complesse

Cosa NON è Data Mining?

Cosa NON è Data Mining?

- Cercare un numero nell'elenco telefonico
- Interrogare un motore di ricerca per cercare informazioni su "Amazon"

Cosa è Data Mining?

- Certi cognomi sono più comuni in certe regioni (es. Casadei, Casadio, … in Romagna)
- Raggruppare i documenti restituiti da un motore di ricerca in base a informazioni di contesto (es. "Amazon rainforest", "Amazon.com")

Le origini del Data Mining

- Questa disciplina trae ispirazioni dalle aree del machine learning/intelligenza artificiale, pattern recognition, statistica e basi di dati
- Le tradizionali tecniche di analisi risultano inidonee per molteplici motivi
  - ✓ Quantità dei dati
  - ✓ Elevata dimensionalità dei dati
  - ✓ Eterogeneità dei dati

Attività tipiche del Data Mining

- Sistemi di predizione
  - ✓ Utilizzare alcune variabili per predire il valore incognito o futuro di altre variabili.
- Sistemi di descrizione
  - ✓ Trovare pattern interpretabili dall'uomo che descrivano I dati

Attività tipiche del Data Mining

- Classificazione [Predittiva]
- Clustering [Descrittiva]
- Ricerca di regole associative [Descrittiva]
- Ricerca di pattern sequenziali [Descrittiva]
- Regressione [Predittiva]
- Individuazione di deviazioni [Predittiva]

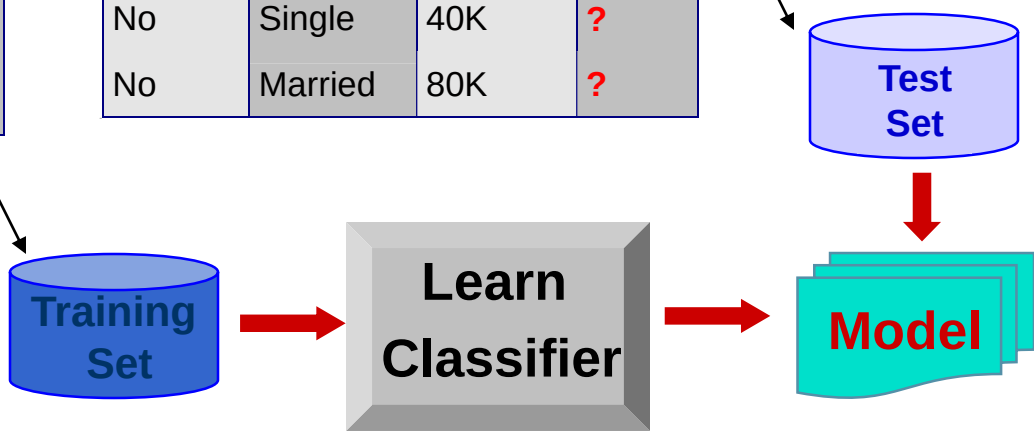
Classificazione: Definizione

- Data una collezione di record (*training set*)
  - ✓ Ogni record è composto da un insieme di *attributi*, di cui uno esprime la *classe* di appartenenza del record.
- Trova un *modello* per l'attributo di classe che esprima il valore dell'attributo in funzione dei valori degli altri attributi.
- Obiettivo: record non noti devono essere assegnati a una classe nel modo più accurato possibile
  - ✓ Viene utilizzato un *test set* per determinare l'accuratezza del modello. Normalmente, il data set fornito è suddiviso in training set e test set. Il primo è utilizzato per costruire il modello, il secondo per validarlo.

Classificazione: Esempio

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classificazione: Applicazione 1

- Direct Marketing
  - ✓ Obiettivo: Ridurre il costo della pubblicità via posta *definendo* l'insieme dei clienti che, con maggiore probabilità, compreranno un nuovo prodotto di telefonia
  - ✓ Approccio:
    - Utilizza i dati raccolti per il lancio di prodotti similari
    - Conosciamo quali clienti hanno deciso di comprare e quali no Questa informazione *{compra, non compra}* rappresenta *l'attributo di classificazione*
    - Raccogli tutte le informazioni possibili legate ai singoli compratori: demografiche, stile di vita, precedenti rapporti con l'azienda
      - Attività lavorativa svolta, reddito, età, sesso, ecc.
    - Utilizza queste informazioni come attributi di input per addestrare un modello di classificazione

Classificazione: Applicazione 2

- Individuazione di frodi
  - ✓ Obiettivo: predire l'utilizzo fraudolento delle carte di credito
  - ✓ Approccio:
    - Utilizza le precedenti transazioni e le informazioni sui loro possessori come attributi
      - Quando compra l'utente, cosa compra, paga con ritardo, ecc.
    - Etichetta le precedenti transazioni come fraudolenti o lecite
    - Questa informazione rappresenta l'attributo di classificazione
    - Costruisci un modello per le due classi di transazioni
    - Utilizza il modello per individuare comportamenti fraudolenti delle prossime transazioni relative a una specifica carta di credito

Classificazione: Applicazione 3

- Individuazione dell'insoddisfazione del cliente:
  - ✓ Obiettivo: Predire clienti propensi a passare a un concorrente.
  - ✓ Approccio:
    - Utilizza i dati relativi agli acquisti dei singoli utenti (presenti e passati) per trovare gli attributi rilevanti
      - Quanto spesso l'utente contatta l'azienda, dove chiama, in quali ore del giorno chiama più di frequente, quale è la sua situazione finanziaria, è sposato, ecc.
    - Etichetta gli utenti come fedeli o non fedeli
    - Trova un modello che definisca la fedeltà

Clustering: Definizione

- Dato un insieme di punti, ognuno caratterizzato da un insieme di attributi, e avendo a disposizione una *misura di similarità* tra i punti, trovare i sottoinsiemi di punti tali che:
  - ✓ I punti appartenenti a un sottoinsieme sono più simili tra loro rispetto a quelli appartenenti ad altri cluster

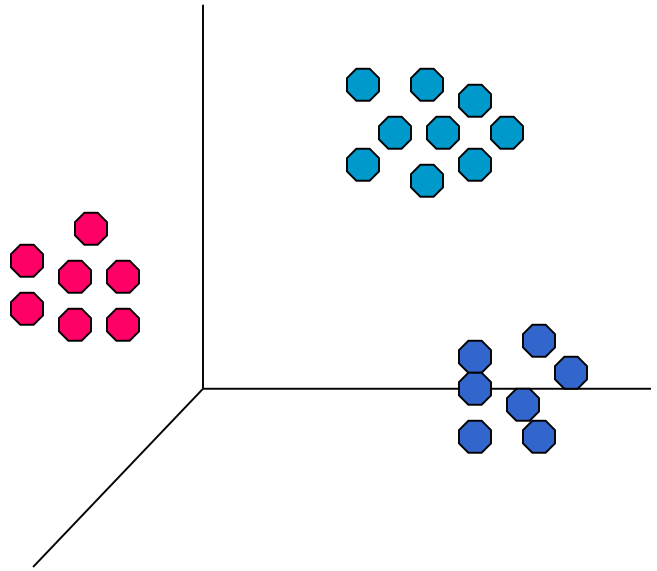
  - Misure di similarità:
  - ✓ La distanza euclidea è applicabile se gli attributi dei punti assumono valori continui
  - ✓ Sono possibili molte altre misure che dipendono dal problema in esame

Rappresentazione del clustering

- Rappresentazione di un clustering nello spazio 3d costruito utilizzando la distanza euclidea come misura di similarità

Le distanze intra-cluster sono minimizzate

Le distanze inter-cluster sono massimizzate



Clustering: Applicazione 1

- Segmentazione del mercato:
  - ✓ Obiettivo: suddividere i clienti in sottoinsiemi distinti da utilizzare come target di specifiche attività di marketing
  - ✓ Approccio:
    - Raccogliere informazioni sui clienti legati allo stile di vita e alla collocazione geografica
    - Trovare cluster di clienti simili
    - Misurare la qualità dei cluster verificando se il pattern di acquisto dei clienti appartenenti allo stesso cluster è più simile di quello di clienti appartenenti a cluster distinti

Clustering: Applicazione 2

- Clustering di documenti:
  - ✓ Obiettivo: trovare sottogruppi di documenti che sono simili sulla base dei termini più rilevanti che in essi compaiono
  - ✓ Approccio: Identificare i termini che si presentano con maggiore frequenza nei diversi documenti. Definire una misura di similarità basata sulla frequenza dei termini e usarla per creare i cluster.

Clustering di documenti

- Punti da clusterizzare: 3204 articoli del Los Angeles Times.
- Misura di similarità: numero di parole comuni tra due documenti (escluse alcune parole comuni).

Categoria	*# articoli*	*#correttamente classsificati*	*%correttamente classsificati*
Finanza	555	364	66%
Esteri	341	260	76%
Cronaca nazionale	273	36	13%
Cronaca locale	943	746	79%
Sport	738	573	78%
Intrattenimento	354	278	79%

Regole associative: Definizione

- Dato un insieme di record ognuno composto da più elementi appartenenti a una collezione data
  - ✓ Produce delle regole di dipendenza che predicono l'occorrenza di uno degli elementi in presenza di occorrenze degli altri.

TID	*Record*
1	Pane, Coca Cola, Latte
2	Birra, Pane
3	Birra, Coca Cola, Pannolini, Latte
4	Birra, Pane, Pannolini, Latte
5	Birra, Pannolini, Latte

Regola:

{Latte} --> {Coca Cola}

{Pannolini, Latte} --> {Birra}

Regole associative: applicazione 1

- Marketing e promozione delle vendite:
  - ✓ Si supponga di avere scoperto la regola associativa

    {Bagels, … } --> {Potato Chips}
  - ✓ Potato Chips come conseguente: l'informazione può essere utilizzata per capire quali azioni intraprendere per incrementare le sue vendite
  - ✓ Bagels come antecedente: l'informazione può essere utilizzata per capire quali prodotti potrebbero essere condizionati nel caso in cui il negozio interrompesse la vendita dei Bagel

Regole associative: Applicazione 2

- Disposizione della merce.
  - ✓ Obiettivo: identificare i prodotti comprati assieme da un numero sufficientemente elevato di clienti.
  - ✓ Approccio: utilizza i dati provenienti dagli scontrini fiscali per individuare le dipendenze tra i prodotti.
  - ✓ Una classica regola associativa
    - Se un cliente compra pannolini e latte, allora molto probabilmente comprerà birra.
    - Quindi non vi stupite se trovate le casse di birra accanto ai pannolini!

Regole associative: Applicazione 3

- Gestione dell'inventario:
  - ✓ Obiettivo: un'azienda che effettua riparazione di elettrodomestici vuole studiare le relazioni tra i malfunzionamenti denunciati e i ricambi richiesti al fine di equipaggiare correttamente i propri veicoli e ridurre le visite alle abitazioni dei clienti.
  - ✓ Approccio: elabora i dati relativi ai ricambi utilizzati nei precedenti interventi alla ricerca di pattern di co-occorrenza.

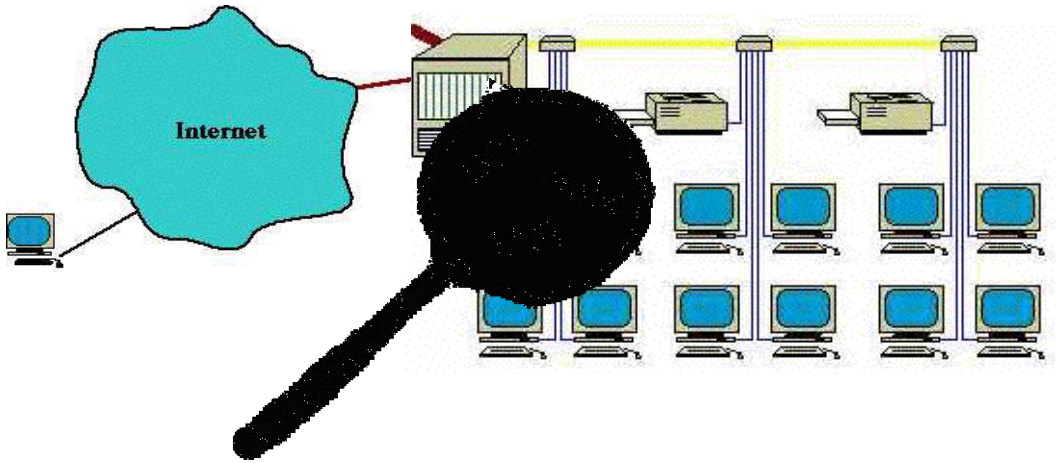
Regressione

- Predire il valore di una variabile a valori continui sulla base di valori di altre variabili assumendo un modello di dipendenza lineare/non lineare.
- Problema ampiamente studiato in statistica e nell'ambito delle reti neurali.
- Esempi:
  - ✓ Predire il fatturato di vendita di un nuovo prodotto sulla base degli investimenti in pubblicità.
  - ✓ Predire la velocità del vento in funzione della temperatura, umidità, pressione atmosferica
  - ✓ Predizione dell'andamento del mercato azionario.

Identificazione di comportamenti anomali e scostamenti

- Identificazione di scostamenti dal normale comportamento
- Applicazioni:
  - ✓ Identificazioni di frodi nell'uso

  - ✓ Identificazioni di intrusioni in rete

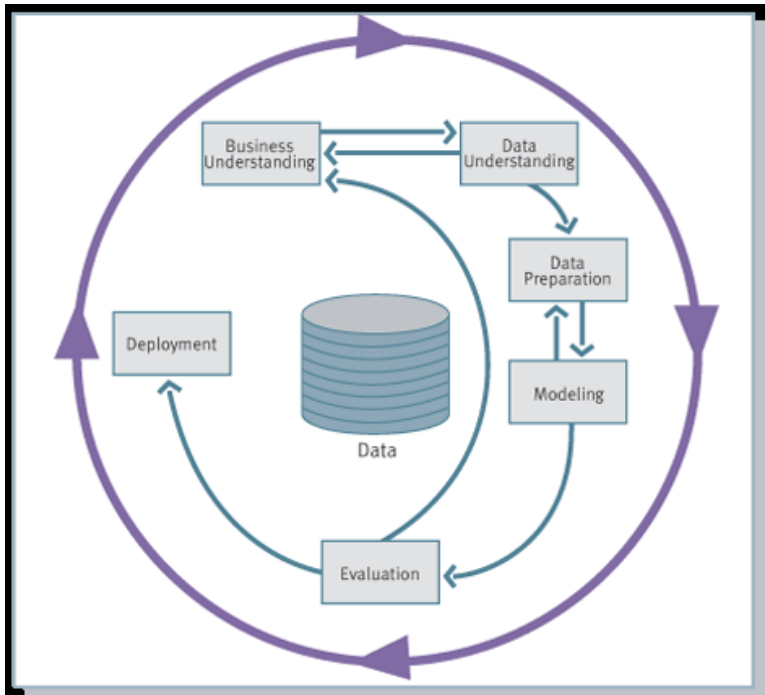


Scommesse del Data Mining

- Scalabilità
- Multidimensionalità del data set
- Complessità ed eterogeneità dei dati
- Qualità dei dati
- Proprietà dei dati
- Mantenimento della privacy
- Processing in real time

CRISP-DM: un approccio metodologico

- Un progetto di Data mining richiede un approccio strutturato in cui la scelta del miglior algoritmo è solo uno dei fattori di successo
- La metodologia **CRISP-DM** è una delle proposte maggiormente strutturate per definire i passi fondamentali di un progetto di Data Mining
- Le sei fasi del ciclo di vita non sono strettamente sequenziali. Tornare su attività già svolte è spesso necessario
- http://www.crisp-dm.org/



CRISP-DM: le fasi

1) Comprensione del dominio applicativo: capire gli obiettivi del progetto dal punto di vista dell'utente, tradurre il problema dell'utente in un problema di data mining e definire un primo piano di progetto

2) Comprensione dei dati: raccolta preliminare dei dati finalizzata a identificare problemi di qualità e a svolgere analisi preliminari che permettano di identificarne le caratteristiche salienti

3) Preparazione dei dati: comprende tutte le attività necessarie a creare il dataset finale: selezione di attributi e record, trasformazione e pulizia dei dati

CRISP-DM: le fasi

4) Creazione del modello: diverse tecniche di data mining sono applicate al dataset anche con parametri diversi al fine di individuare quella che permette di costruire il modello più accurato

4) Valutazione del modello e dei risultati: il modello/i ottenuti dalla fase precedente sono analizzati al fine di verificare che siano sufficientemente precisi e robusti da rispondere adeguatamente agli obiettivi dell'utente

5) Deployment: il modello costruito e la conoscenza acquisita devono essere messi a disposizione degli utenti. Questa fase può quindi semplicemente comportare la creazione di un report oppure può richiedere di implementare un sistema di data mining controllabile direttamente dall'utente