Data Mining

Sergio Greco
DIMES, Università della Calabria

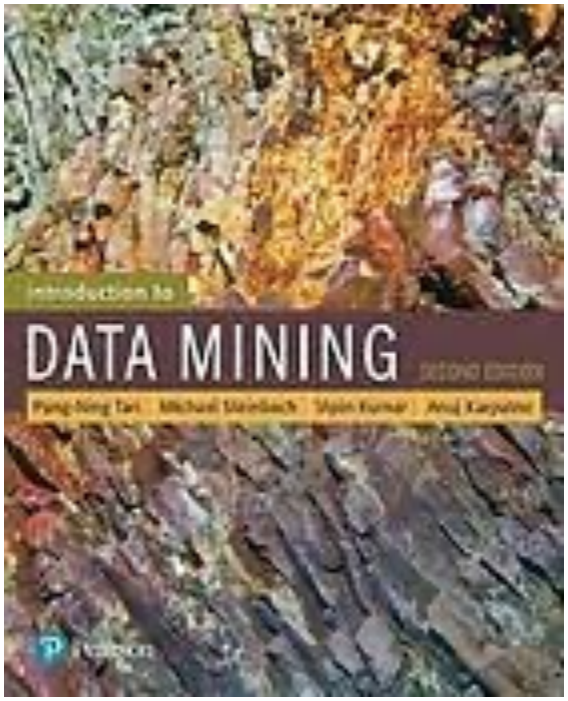
Alcune informazioni

Modalità didattiche e materiale

- Docenti:
  - Prof. Sergio Greco,
  - Dr. Domenico Mandaglio
  - Dr. Reza Shahbazian
- Numero di CFU: 6
- Orario: 4/6 ore per settimana
- Sito del corso: Teams (codice a03kab9)

Modalità didattiche e materiale

- Lezioni in aula ed esercitazioni utilizzando alcune librerie tramite programmi Python.
- Il corso introduce i concetti di base, descrive le tecniche di mining da applicare a dati strutturati
- **Libro di testo**:
  - ✓ Pang-Ning Tan, Michael Steinbach, Anui Karpatne, Vipin Kumar Introduction *to Data Mining*. Pearson International, second edition.



Modalità di esame

- L'esame consta di una prova in laboratorio (2 ore circa) e di una prova orale.

- La prova in laboratorio consiste nel completamento di uno jupyter notebook (maggiori dettagli in seguito…)

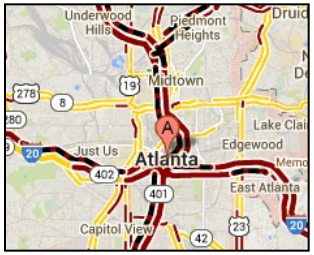
Motivazione

Ovunque abbiamo grandi quantità di dati

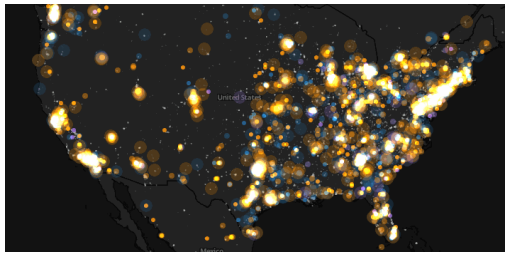
- Enorme crescita dei dati in database commerciali e scientifici a causa dei progressi nella generazione di dati e tecnologie di raccolta.
- Nuovo mantra:
  Raccogli tutti i dati che puoi quando e dove possibile.
- Aspettative:
  I dati raccolti avranno valore sia per lo scopo per cui sono stati collezionati o per uno scopo non previsto.

Cyber Security

E-Commerce

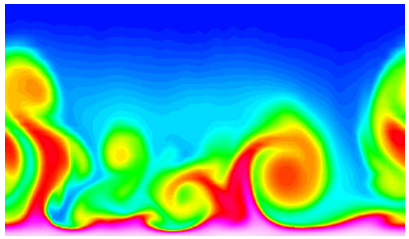


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

Data Mining – Punto di vista commerciale

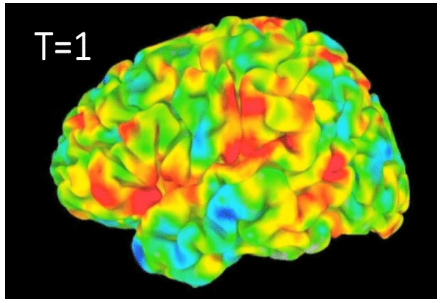
- Molti dati vengono raccolti e immagazzinati
  - Dati Web
    - Yahoo ha peta byte di dati web
    - Facebook ha miliardi di utenti attivi
  - Acquisti e e-commerce
    - Amazon gestisce milioni di visite al giorno
  - Transazioni bancarie / con carta di credito

- I computer sono diventati più economici e più potenti

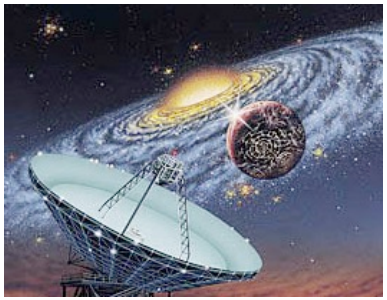
- La pressione competitiva è forte
  - Bisogna fornire servizi personalizzati migliori per un vantaggio competitivo (e.g. Customer Relationship Management)

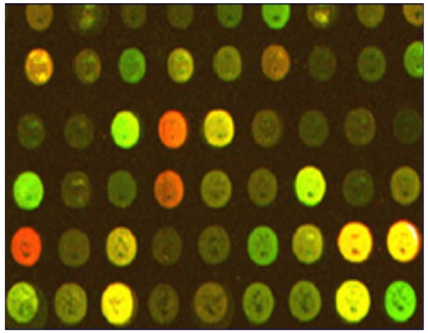
Data Mining – Punto di vista scientifico

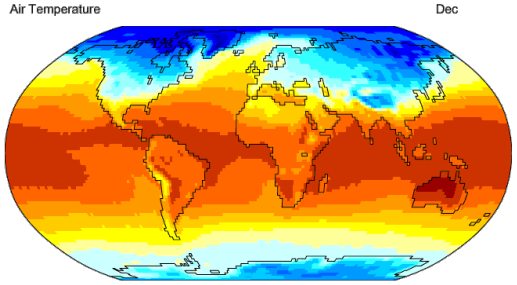
- Dati raccolti e archiviati sempre più velocemente
  - sensori remoti su un satellite
    - Archivi della NASA EOSDIS hanno petabyte di dati di osservazioni della terra per anno
  - telescopi che scansionano i cieli
    - Dati del sondaggio Sky
  - Dati biologici
  - Simulazioni scientifiche
    - terabyte di dati generati in poche ore
- Il data mining aiuta gli scienziati
  - nell'analisi automatizzata di enormi set di dati
  - Nella formazione di ipotesi.



fMRI Data from Brain


Sky Survey Data

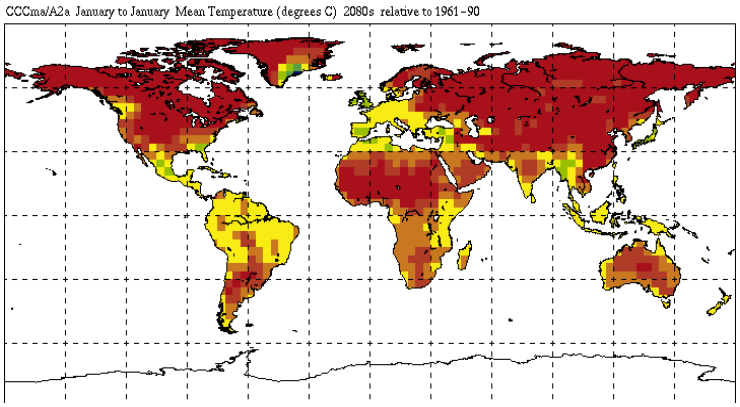

Gene Expression Data



Surface Temperature of Earth

Grandi opportunità per problemi rilevanti

Migliorare le cure sanitarie e ridurre I costi



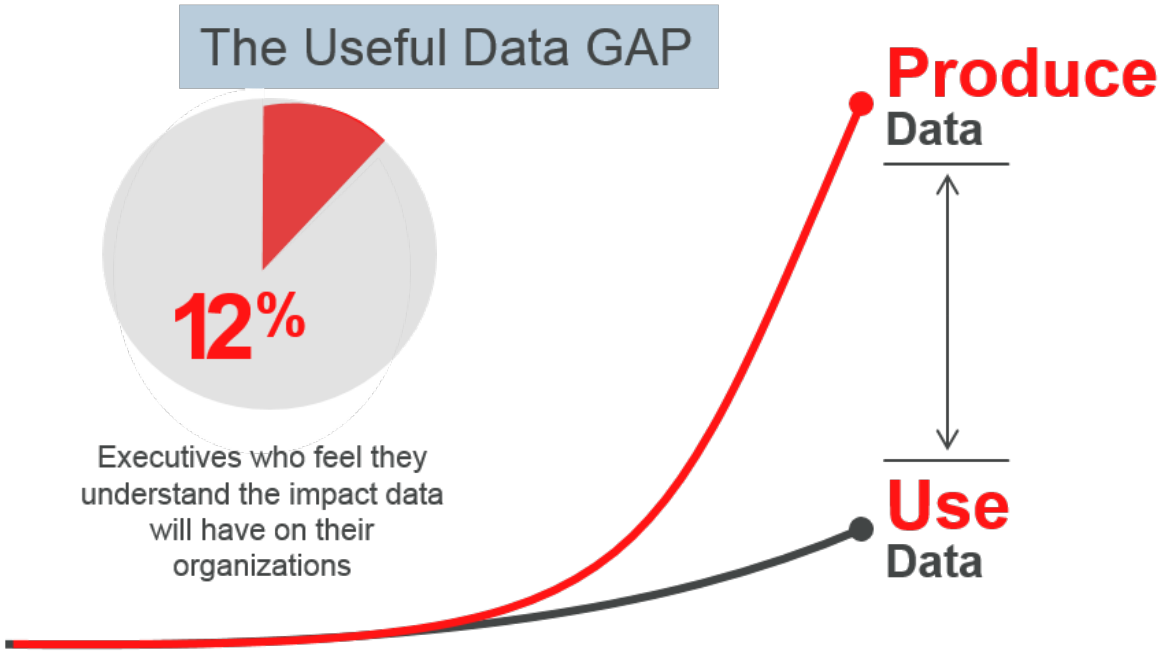
Prevedere l'impatto del cambio del clima

Trovare fonti alternative di energia verde

Migliorare la produzione Agricola

Data mining su grandi data set

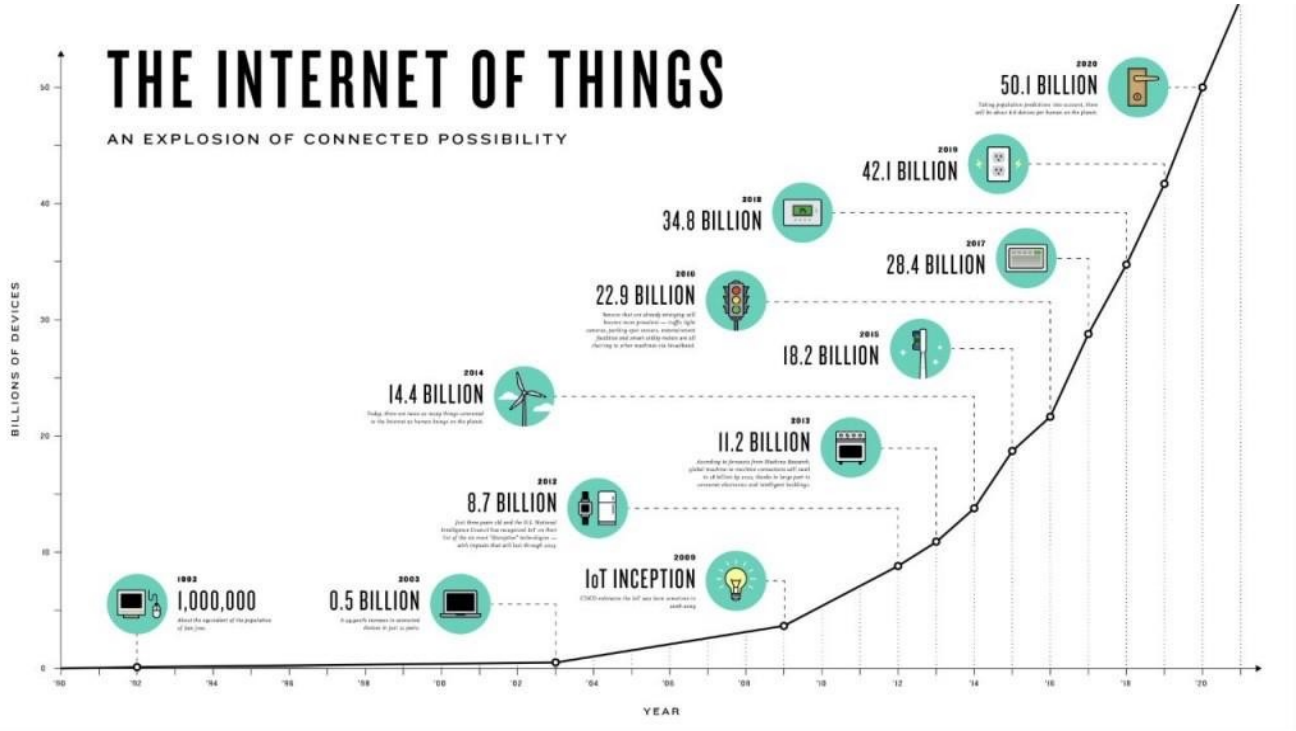
- Molte delle informazioni presenti sui dati non sono direttamente evidenti
- Le analisi guidate dagli uomini possono richiedere settimane per scoprire informazioni utili
- Larga parte dei dati non sono di fatto mai analizzate



Da: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

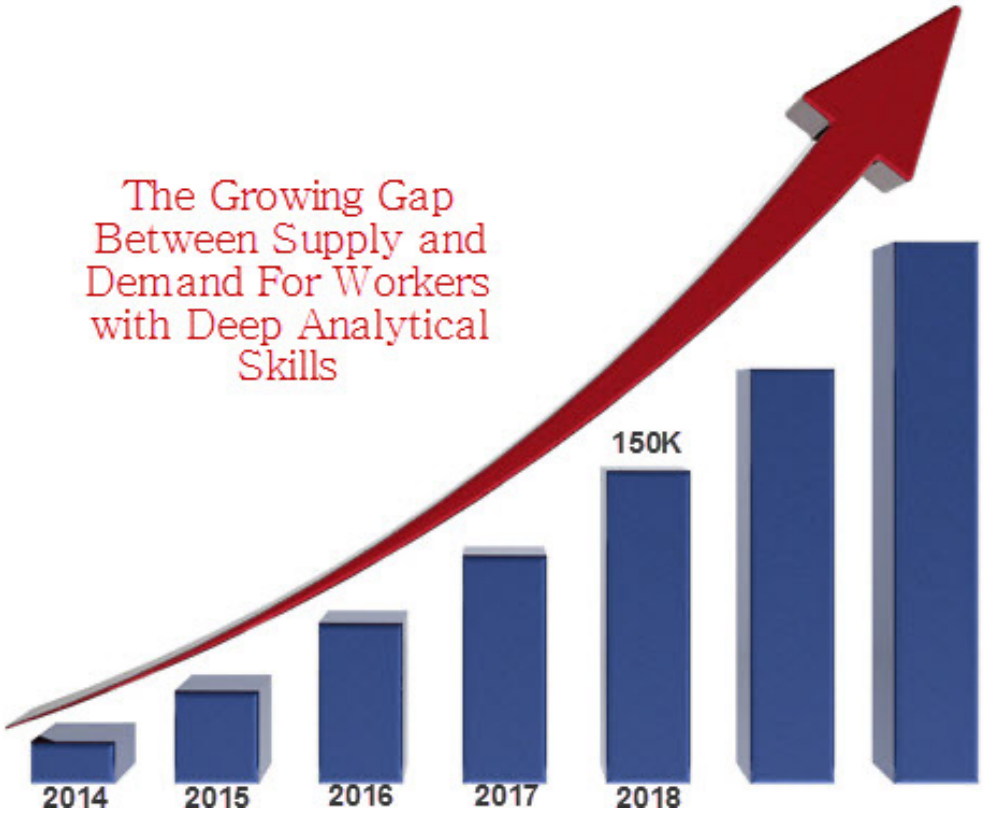
Data mining e i Big Data

- La nuova frontiera è rappresentata dall'analisi dei Big Data. Dati generati da:
  - ✓ Sensori e applicazioni
  - ✓ Comunicazioni tra applicazioni e utenti
  - ✓ Digital Footprint



Crescente ricerca di esperti

- La necessità di esperti nell'analisi dei dati aumenta in modo esponenziale

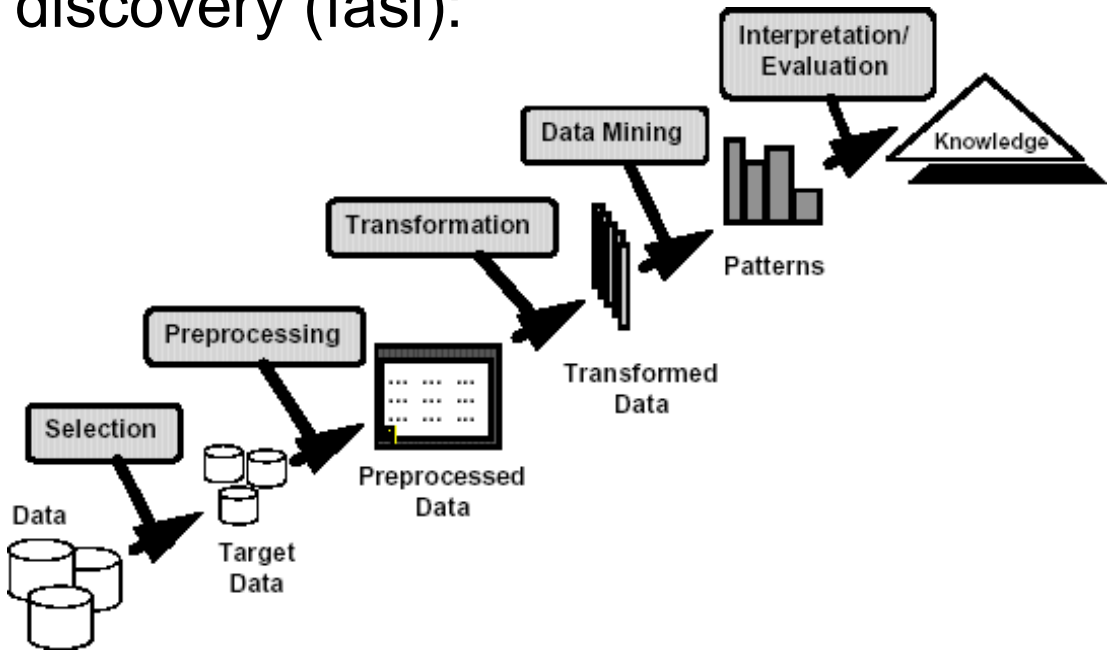


Cosa è e di cosa si occupa il Data Mining

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

✓ Knowledge discovery (fasi):



Pattern

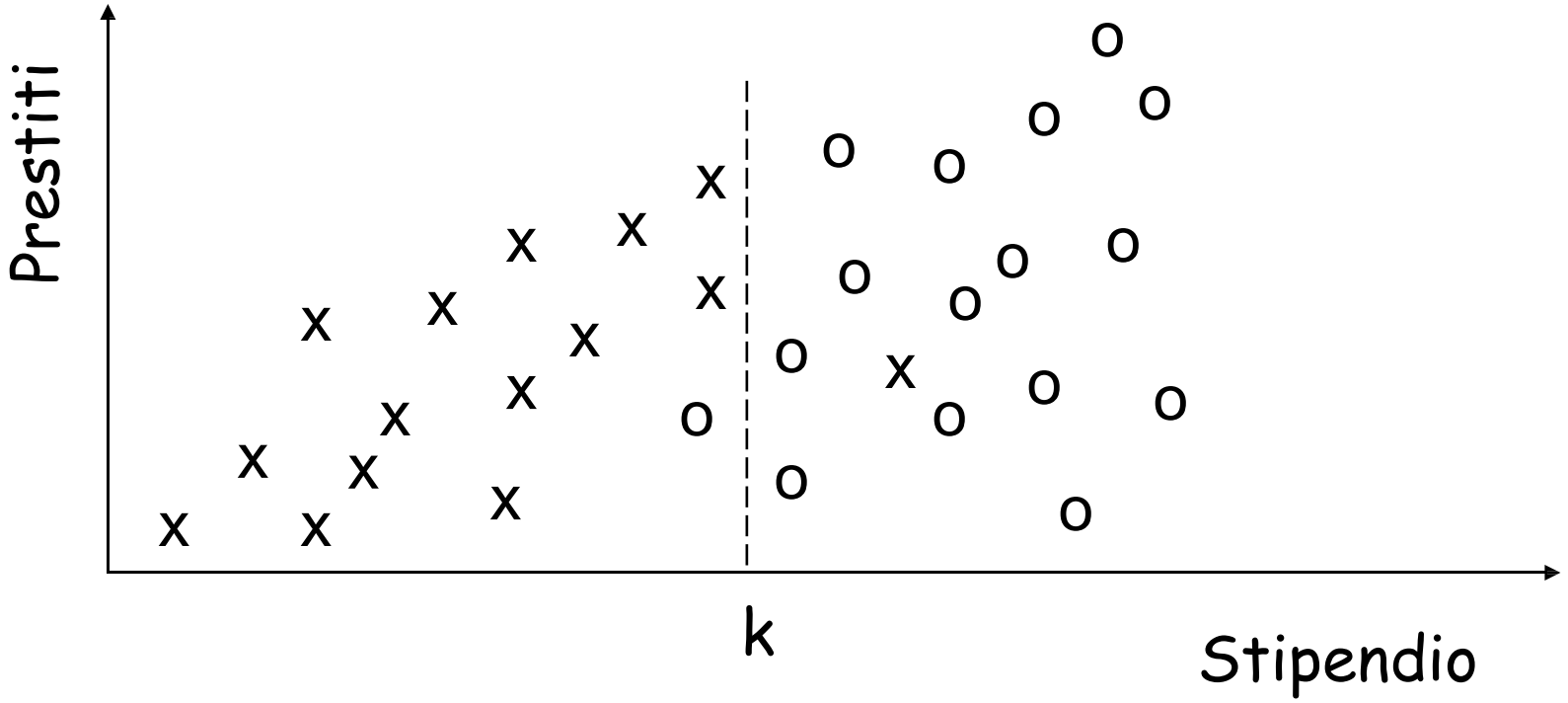
- Un **pattern** è una rappresentazione sintetica e ricca di semantica di un insieme di dati;
- Esprime in genere un modello ricorrente nei dati, ma può anche esprimere un modello eccezionale
- Un pattern deve essere:
  - ✓ **Valido** sui dati con un certo grado di confidenza
  - ✓ **Comprensibile** 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?

- Eseguire una query su una base di dati (e.g. cercare un numero nell'elenco telefonico)
- Interrogare un motore di ricerca per cercare informazioni su "Amazon"

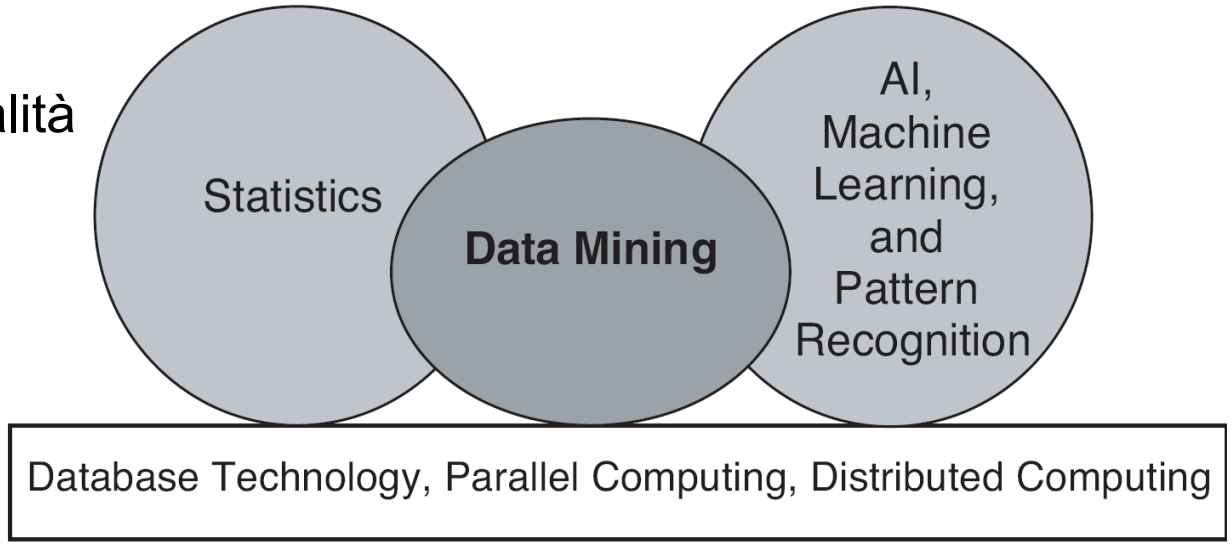
Cosa è Data Mining?

- Certi cognomi sono più comuni in alcune aree geografiche (es. Filice, Spadafora e Greco nella prov. di CS)
- Raggruppare i documenti restituiti da un motore di ricerca in base a informazioni di contesto

Origini del Data Mining

- Trae ispirazione da diverse aree quali machine learning/AI, pattern recognition, statistica, e database systems

- Le tecniche tradizionali sono inadeguate a causa di
  - Grandi dimensioni
  - Elevata dimensionalità
  - Eterogeneità
  - Complessità
  - Distribuzione



- Una componente chiave di settori emergent quali *data science* e *data-driven discovery*

Attività tipiche del Data Mining

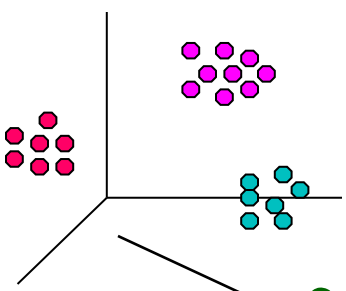
- Sistemi di predizione
  - ✓ Utilizzare alcune variabili per predire il valore incognito o futuro di altre variabili.

- 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]

Data Mining Tasks ...

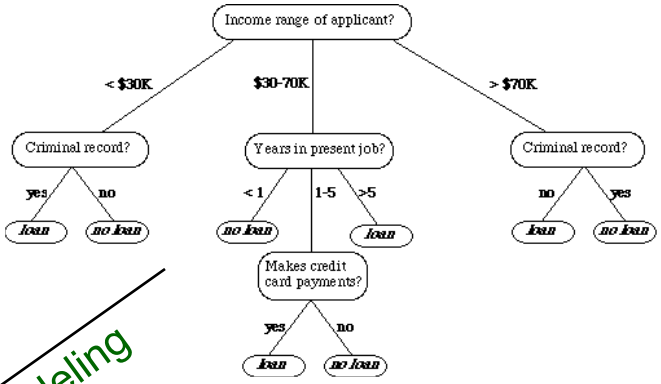


Clustering

Data

Tid	Refund	Marital Status	Taxable Income	Cheat
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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Predictive Modeling



Association Rules

Anomaly Detection

Tecniche e applicazioni del Data Mining

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.

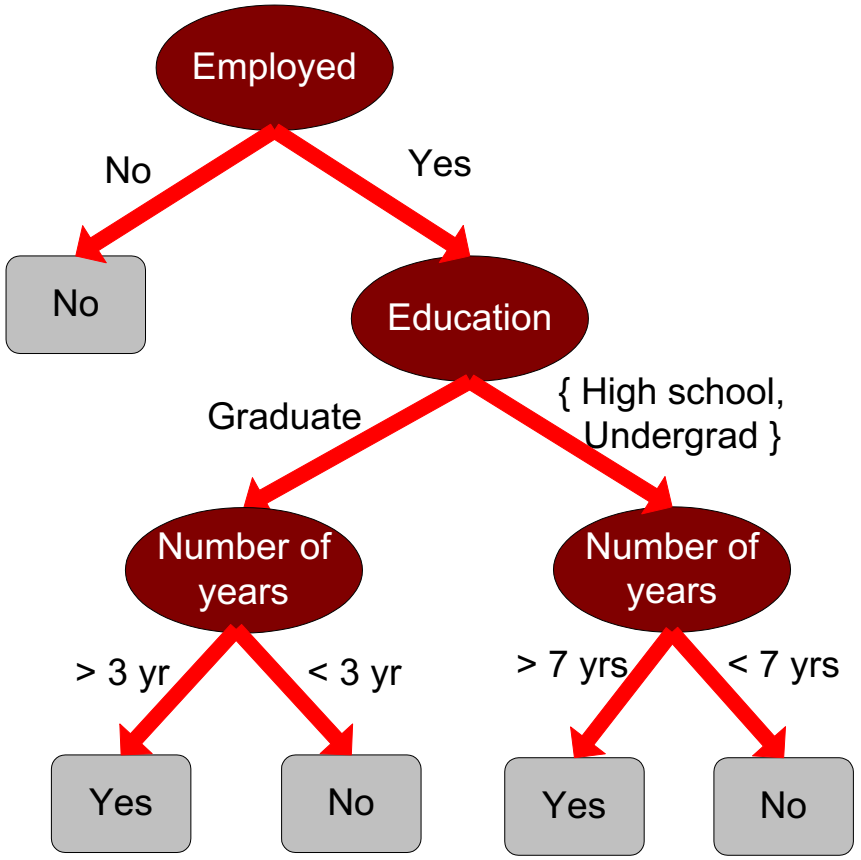
Modello Predittivo: Classificazione

- Trova un modello per l'attributo di classe in funzione dei valori di altri attributi

Modello per la predizione della garanzia del credito

Class

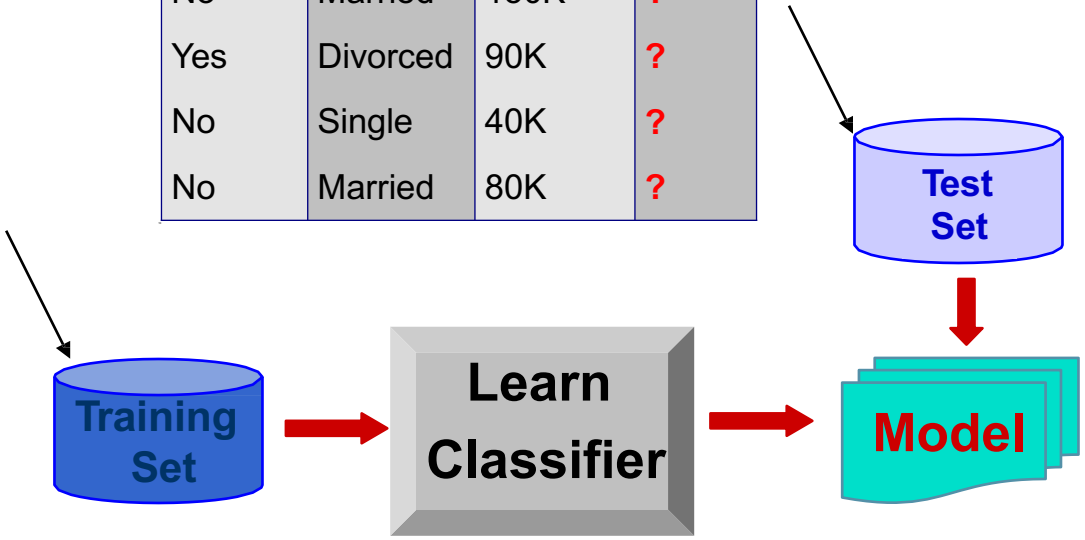
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…



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 (Drop-out Risk)
  - ✓ 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

Classificazione: Applicazione 4

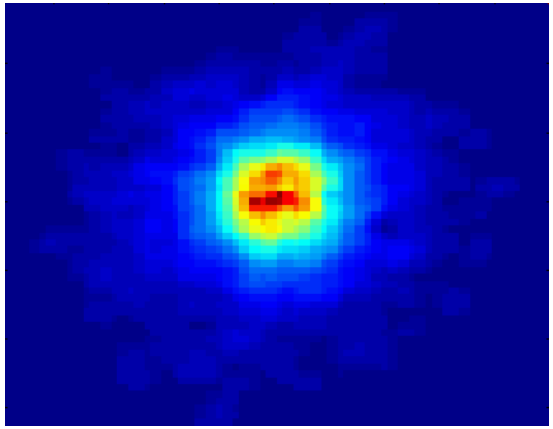
- **Classificare le "osservazioni del cielo"**
  - **Obiettivo**: Predire la classe (***stella*** o ***galassia***) di un oggetto celeste, in particolare quelli visivamente deboli, basandosi sulle immagini delle osservazioni telescopiche (e.g. osservatorio di Palomar).
    - 3000 immagini con 23,040 x 23,040 pixels per immagine.
  - **Approccio**:
    - Partiziona le immagini.
    - Misura gli attributi (features) - 40 per ciascun oggetto.
    - Genera un modello per la classe basato su tali attributi.
    - Success Story: Trovate 16 nuove quasar ad elevato spostamento verso il rosso (oggetti lontani difficili da trovare).

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classificazione delle galassie

Courtesy: http://aps.umn.edu

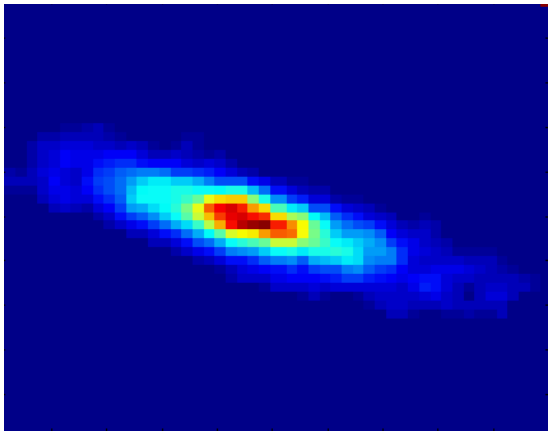
Early



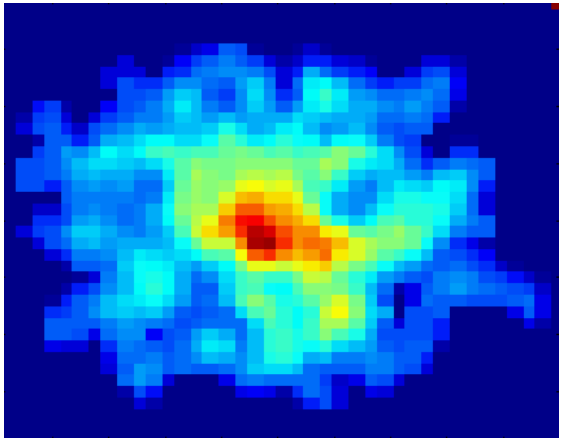
Class:
- Stadio di formazione

Attributi:
- Imaggini features,
- Caratteristiche della luce, onde ricevute, ecc.

Intermediate



Late



Dimensione dei dati:
- 72 milioni di stelle, 20 milioni di galassie
- Catalogo degli oggetti : 9 GB
- Database di immagini: 150 GB

Clustering: Definizione

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

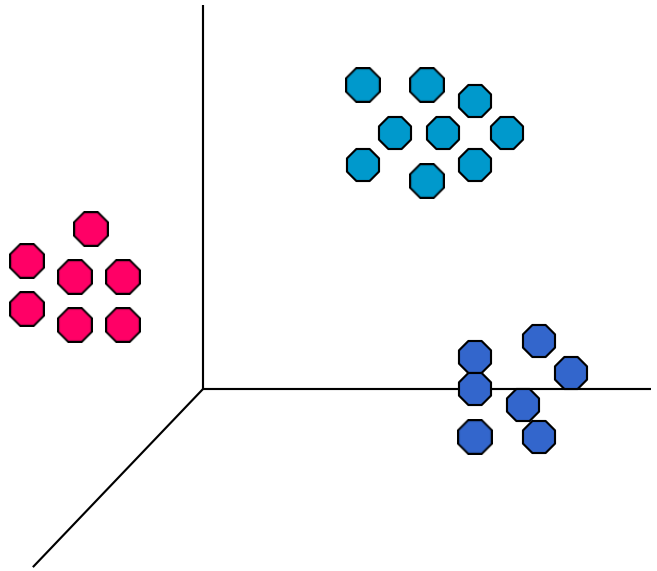
- Misure di similarità:
  - ✓ La distanza euclidea è applicabile se gli attributi dei punti assumono valori continui
  - ✓ Sono possibili molte altre misure che dipendono dal problema in esame

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 (**cluster**) 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 (significativi) 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 i**ncrementare 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.
  - ✓ **Esempio:**

    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.

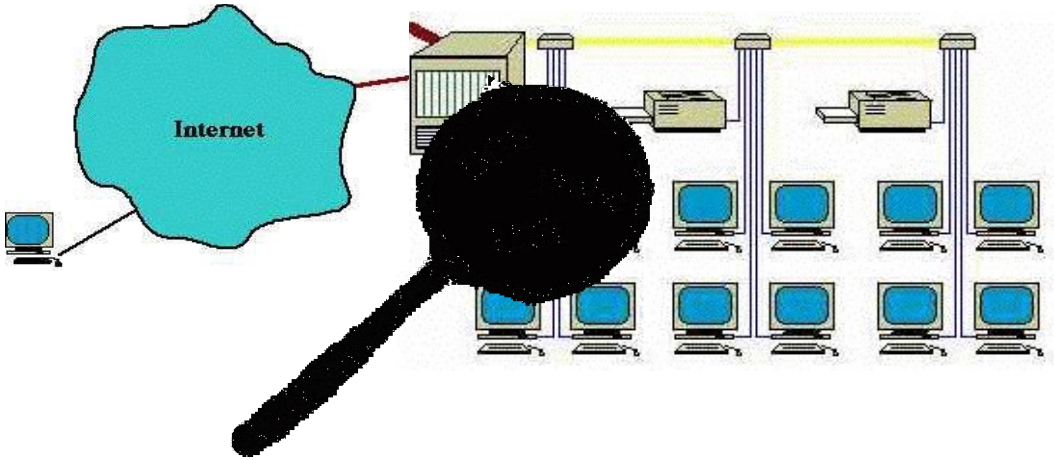
Regressione

- 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

anomalie e scostamenti

- Identificazione di scostamenti dal normale comportamento
- Applicazioni:
  - ✓ Identificazioni di frodi nell'uso delle carte di credito
  - ✓ Identificazioni di intrusioni in rete

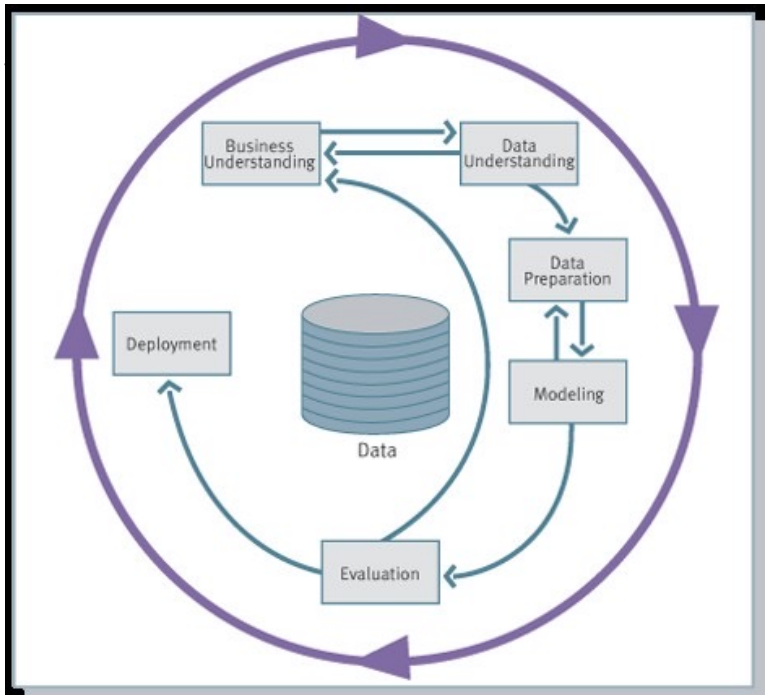


Scommesse del Data Mining

- Scalabilità
- Multidimensionalità del data set
- Complessità ed eterogeneità dei dati
- Qualità dei dati
- Proprietà dei dati
- Mantenimento della privacy
- Processamento in tempo reale

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 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