Most companies own huge databases containing:
- operational data
- textual documents
- experiment results

These databases are a potential source of useful information.

Data analysis
- Information is “hidden” in huge datasets
  - not immediately evident
  - human analysts need a large amount of time for the analysis
  - most data is never analyzed at all

The Data Gap

Data mining
- Non trivial extraction of
  - implicit
  - previously unknown
  - potentially useful information from available data
- Extraction is automatic
  - performed by appropriate algorithms
- Extracted information is represented by means of abstract models
  - denoted as pattern

Example: profiling
- Consumer behavior in e-commerce sites
  - Selected products, requested information, ...
- Search engines and portals
  - Query keywords, searched topics and objects
- Social network data
  - Facebook, google+ profiles
  - Dynamic data: posts on blogs, FB, tweets
- Maps and georeferenced data
  - Localization, interesting locations for users

Example: profiling
- User/service profiling
  - Recommendation systems
  - Advertisements
- Market basket analysis
  - Correlated objects for cross selling
    - User registration, fidelity cards
- Context-aware data analysis
  - Integration of different dimensions
    - E.g., location, time of the day, user interest
- Text mining
  - Brand reputation, sentiment analysis, topic trends
Data mining fundamentals

Example: biological data
- Microarray
  - expression level of genes in a cellular tissue
  - various types (mRNA, DNA)
- Patient clinical records
  - personal and demographic data
  - exam results
- Textual data in public collections
  - heterogeneous formats, different objectives
  - scientific literature (PUBMed)
  - ontologies (Gene Ontology)

Biological analysis objectives
- Clinical analysis
  - detecting the causes of a pathology
  - monitoring the effect of a therapy
  ⇒ diagnosis improvement and definition of new specific therapies
- Bio-discovery
  - gene network discovery
  - analysis of multifactorial genetic pathologies
- Pharmacogenesis
  - lab design of new drugs for genic therapies

Knowledge Discovery Process
KDD = Knowledge Discovery from Data

Preprocessing
- data cleaning
  - reduces the effect of noise
  - identifies or removes outliers
  - solves inconsistencies
- data integration
  - reconciles data extracted from different sources
  - integrates metadata
  - identifies and solves data value conflicts
  - manages redundancy

Real world data is "dirty"
Without good quality data, no good quality pattern

Data mining origins
- Draws from
  - statistics, artificial intelligence (AI)
  - pattern recognition, machine learning
  - database systems
- Traditional techniques are not appropriate because of
  - significant data volume
  - large data dimensionality
  - heterogeneous and distributed nature of data

Analysis techniques
- Descriptive methods
  - Extract interpretable models describing data
  - Example: client segmentation
- Predictive methods
  - Exploit some known variables to predict unknown or future values of (other) variables
  - Example: "spam" email detection
Data mining fundamentals

Classification

- Objectives
  - prediction of a class label
  - definition of an interpretable model of a given phenomenon

- Approaches
  - decision trees
  - bayesian classification
  - classification rules
  - neural networks
  - k-nearest neighbours
  - SVM

Classification

- Requirements
  - accuracy
  - interpretability
  - scalability
  - noise and outlier management

- Applications
  - detection of customer propension to leave a company (churn or attrition)
  - fraud detection
  - classification of different pathology types
  - ...

Clustering

- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers

- Approaches
  - partitional (K-means)
  - hierarchical
  - density-based (DBSCAN)
  - SOM

- Requirements
  - scalability
  - management of
    - noise and outliers
    - large dimensionality
    - interpretability
Data mining fundamentals

### Clustering
- Applications
  - customer segmentation
  - clustering of documents containing similar information
  - grouping genes with similar expression pattern
  - ...

### Association rules
- Objective
  - extraction of frequent correlations or pattern from a transactional database

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diapers, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diapers, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diapers, Milk</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Association rule
  - diapers ⇒ beer
  - 2% of transactions contain both items
  - 30% of transactions containing diapers also contain beer

Tickets at a supermarket counter

### Association rules
- Applications
  - market basket analysis
  - cross-selling
  - shop layout or catalogue design

### Other data mining techniques
- Sequence mining
  - ordering criteria on analyzed data are taken into account
  - example: motif detection in proteins
- Time series and geospatial data
  - temporal and spatial information are considered
  - example: sensor network data
- Regression
  - prediction of a continuous value
  - example: prediction of stock quotes
- Outlier detection
  - example: intrusion detection in network traffic analysis

### Open issues
- Scalability to huge data volumes
- Data dimensionality
- Complex data structures, heterogeneous data formats
- Data quality
- Privacy preservation
- Streaming data

Elena Baralis
Politecnico di Torino