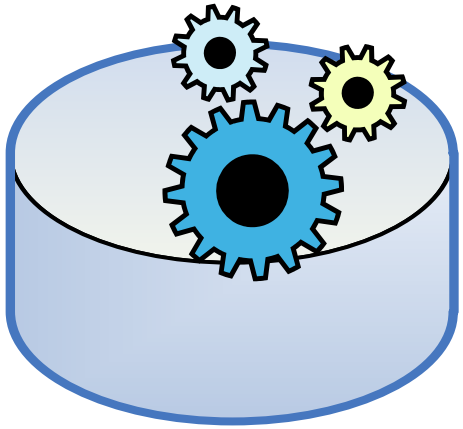
Recap – 1st Lecture

Data Base and Data Mining Group of Politecnico di Torino

Danilo Giordano

Politecnico di Torino

Data mining

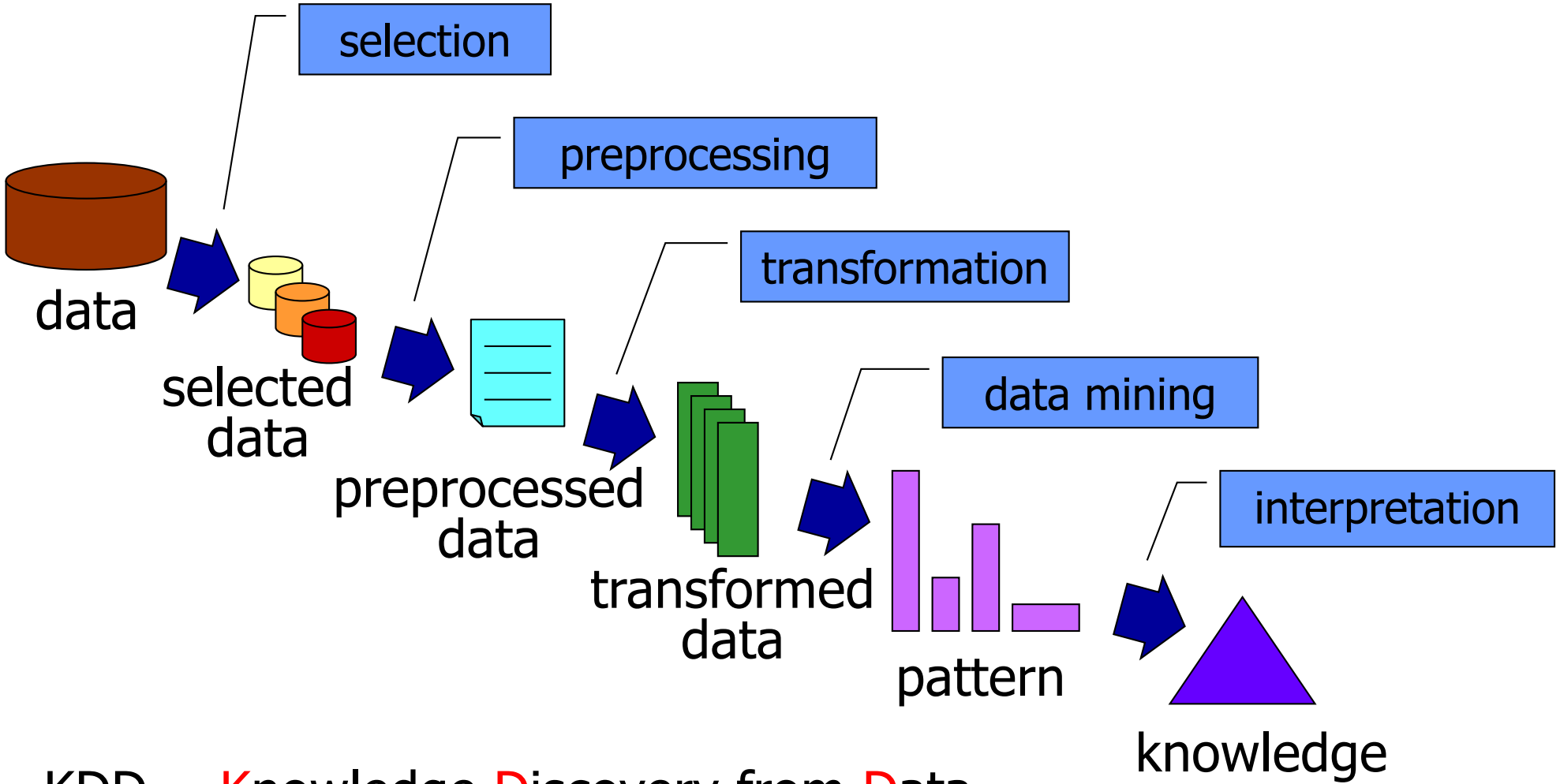


- Non trivial extraction of
  - implicit
  - previously unknown
  - potentially useful

  information from available data
- Extraction is automatic

  performed by appropriate algorithms
- Same pipeline is used to address for very different tasks
  - Users profiling vs identification of key genes of a pathology
  - Brand reputation vs gene ontology
- Limitation of data mining solutions in the big data era

Recap - Knowledge Discovery from Data



KDD = Knowledge Discovery from Data

Preprocessing

- Clean the data
  - Remove noise or outlier
    - Transform the data to detect noise e.g., Fourier transformation
    - Aggregate data samples to get a more "stable" data
- Reduce the dimensionality
  - Data sampling: reduce the number of data instances
    - Random sampling/Stratified Random Sampling, …
  - Feature Selection/Creation: remove the number of attributes per instance
    - Filtering, Embedded, Wrapper, Correlation, PCA, Domain Knowledge combinations, …
  - Discretization: reduce the possible values of an attribute
    - Equally spaced interval, same frequency, clustering, …

Preprocessing

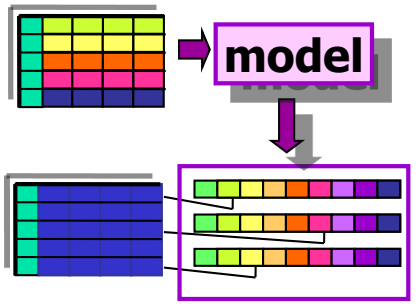
- Transform the data
  - Custom functions: transform data into an other domain
    - Fourier Function, exponential, logarithm, …
  - Normalization: have all the data in an equal dimensional space
    - min-Max, z-score, decimal scaling
- Document data: to represent a document with features
  - Translate each word in a ' term'
    - Terms are defined to avoid singular/plural, verb tenses
  - Remove stop words
    - Articles, too common words
  - Each term is a component (attribute) of the vector
    - the value of each component is the number of times the corresponding term occurs in the document

Distance/Similarity

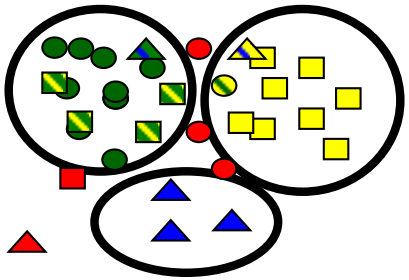
- All features should be in the same dimensional space
  - Otherwise the comparison is biased
- Numeric Distance
  - Euclidean Distance, Minkowski Distance
- Vector Distance: Distance between documents
  - **Simple Matching**: evaluating what is present or not present in both vectors
  - **Jaccard Index**: counting evaluating **only** what is present is both vectors
  - **Cosine similarity**: evaluating the distance between the value of the component of the vectors
- Combine distances based on weights may bias your analysis

Data mining methodologies

- **Classification** - Predictive methods
  - Study previous historical data to create a model
  - Use this model to predict discrete class labels



- **Clustering** - Descriptive methods
  - detecting groups of similar data objects
  - identifying exceptions and outliers



- **Association rules** - Descriptive methods
  - extraction of frequent correlations or pattern from a transactional database

Association rule

diapers $\Rightarrow$ beer