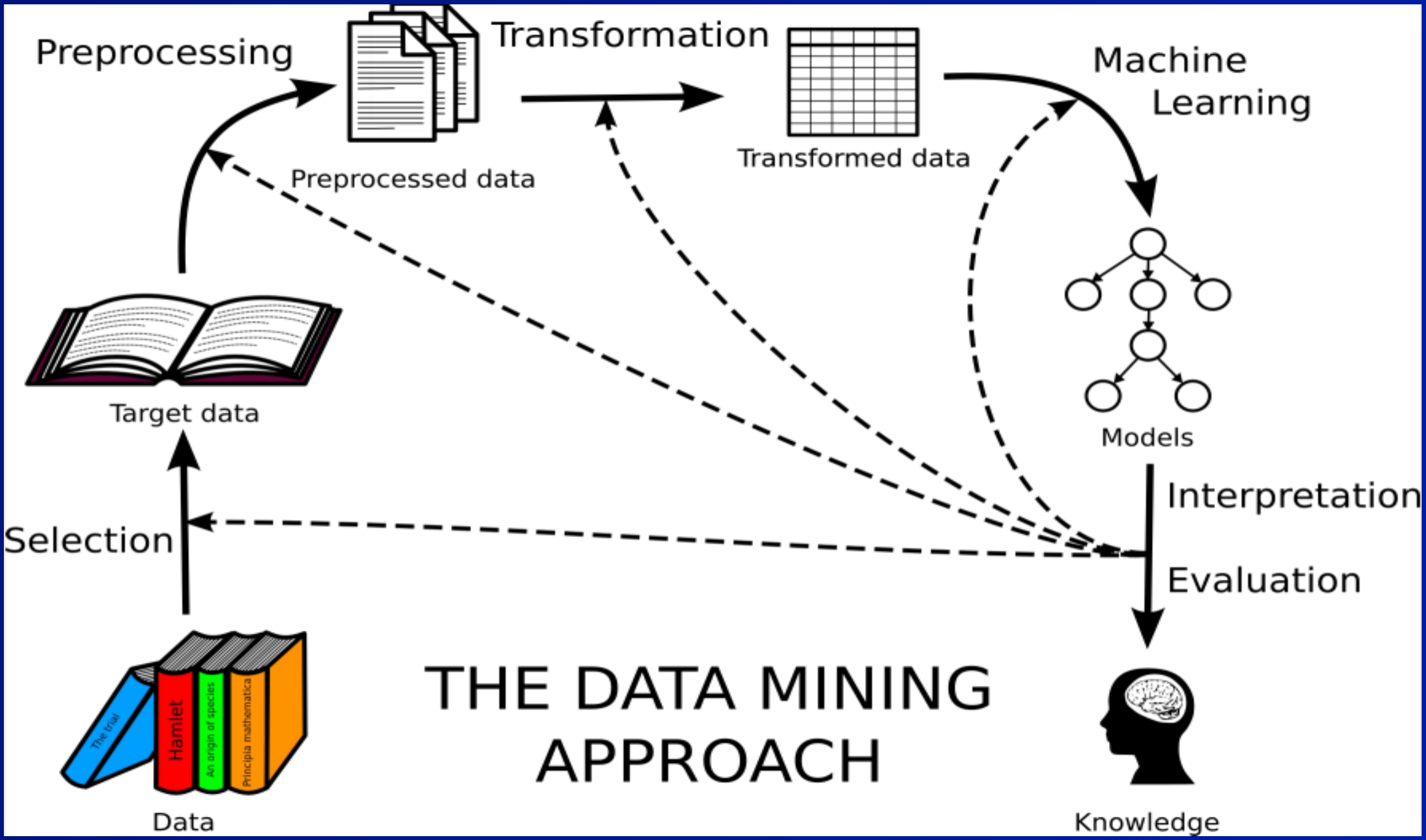
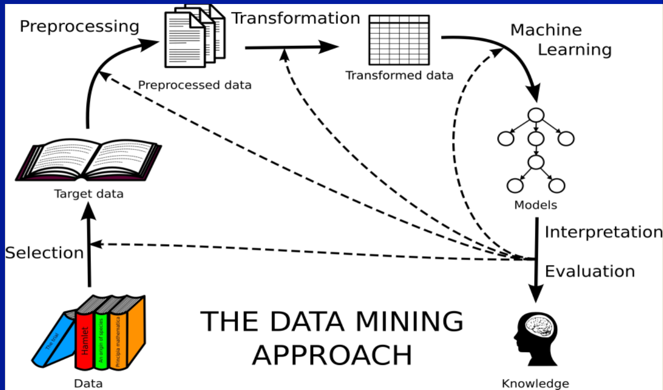
KDD PROCESS

KNOWLEDGE DISCOVERY IN DATABASES



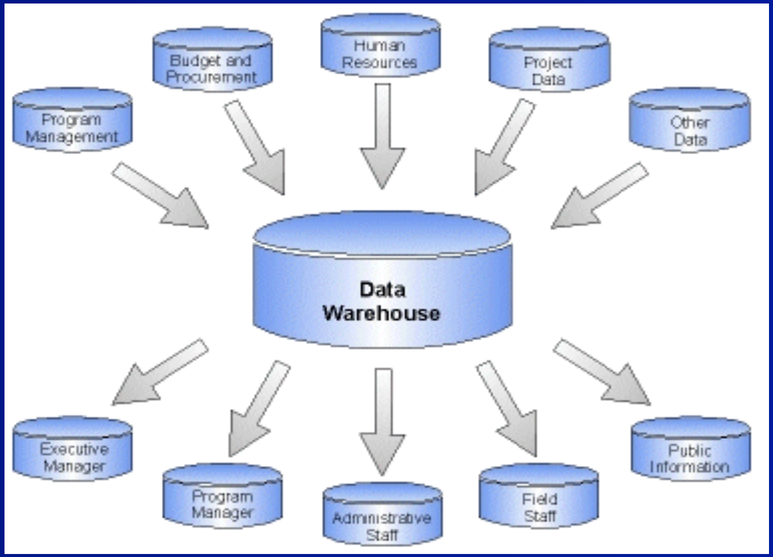
KDD PROCESS
KNOWLEDGE DISCOVERY IN DATABASES

1. Data collection and integration
2. Data cleaning, selection and transformation – Data preprocess
3. Model construction – Machine Learning
4. Model evaluation and interpretation
5. Model use and difusion



DATA COLLECTION AND INTEGRATION

- The culture of data-saving is exponentially increasing in current world
- Data Warehouse: a well-known concept in IT and database systems
- Instead of data warehouse system, many file-flat systems: text files, excel files, etc.
- Many “data streaming” scenarios: telecommunication, robotics, energy consumption, on-line sale systems…



DATA CLEANING, SELECTION AND TRANSFORMATION

- More than the algorithm itself → the quality of the final model depends on the quality of the data
  - Detection of errors in data [e.g. negative age]
  - Removal of non-sense features [e.g. date, identifier]
  - Outlier detection [e.g. fraud detection]
  - How to deal with missing values [e.g. missing at random?]
  - Random selection of samples in huge databases
  - Construction of new features which could help the model construction phase [e.g. Cartesian product]
  - Discretization of continuous variables
  - Feature selection: removal of irrelevant and redundant features
  - …

MODEL INDUCTION – MACHINE LEARNING

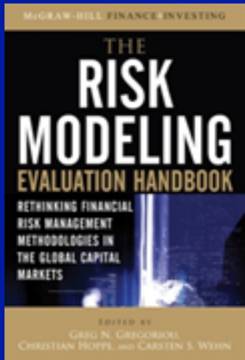
- A huge number of data analysis algorithms have been proposed in the last decade
- A large batery of user-friendly softwares: MLC++, Mineset, R-project, RapidMiner, WEKA…
- “No free lunch theorem”: WOLPERT, David H., 1996. The lack of *a priori* distinctions between learning algorithms. *Neural Computation*, **8**(7), 1341–1390

- Descriptive models:
  - Association rules: A-priori algorithm, Eclat…
  - Clustering: hierarchical, partitional, EM, SOM…
- Predictive models:
  - Regression: lineal regression, additive regression…
  - Supervised classification: Bayesian network classifiers, K-NN, neural networks, combination of classifiers, decision trees…

MODEL EVALUATION AND INTERPRETATION

- Universal evaluation (performance estimation) techniques for data models
  - K-fold cross-validation, bootstrapping, hold-out…

- Performance scores-metrics to asses the goodness of each type of model:
  - Supervised classification: correctly classification percentage, confusion matrix, ROC curves…
  - Regression: MSE…
  - Clustering: intra-class homogeneity, inter-class heterogeneity
  - Association rules: coverage, confidence…



MODEL USE AND DIFFUSION

- Integration of the model in the company *know-how* system?
  - Difficult task
  - Personal attitude, a generational issue?

- Updating the model:
  - More data [e.g. data streaming scenarios]
  - Current classification techniques