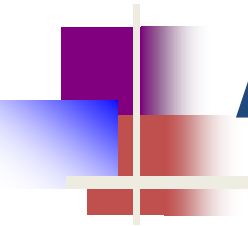


INTRODUCTION TO DATA MINING

Pronalaženje skrivenog znanja

Bojan Furlan

INTRODUCTION

- For what data mining can be used for?
  - Business
    - **Customer Relationship Management (CRM)**: make customer's individual profile,
    - **Churn Analysis**: instead of sending an offer to all people, send offers to *loyal* customers.
    - **Market basket analysis**: identify the *purchase patterns*.
    - **Fraud detection**: insurance,
    - **Risk management**: should the loan be approved for *this* customer?

INTRODUCTION

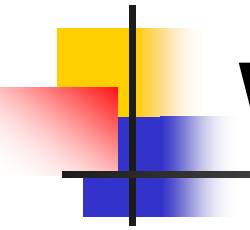
- For what data mining can be used for?
  - Science and engineering
    - **Bioinformatics, genetics**, **medicine**:
      e.g. cancer risks,
    - **Education**:
      e.g. expertise finding systems,
    - **Electrical power engineering**:
      e.g. smart grid, predictive maintenance
  - Surveillance, etc.

WHAT IS DATA MINING?

- Definitions of data mining:
  - Data mining is the process of discovering meaningful new correlations, patterns and trends by shifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.
  - Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.
  - Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases.

IR vs. ML vs. DM

- **Information retrieval** is about finding something that *already* is part of your data, as fast as possible.
  - e.g. Search engines – find a document within many documents
- **Machine learning** are techniques to generalize existing knowledge to *new* data, as accurate as possible.
  - e.g. computer vision – object recognition
- **Data mining (or Knowledge Discovery in Data)** is primarily about discovering something *hidden* in data, that you did *not* know before, as "new" as possible.
  - CRM – customer analysis
- They intersect and often use techniques of one another:
  - DM and IR both use index structures to accelerate processes.
  - DM uses a lot of ML techniques
    - e.g. a pattern in the data set, might be a new knowledge.

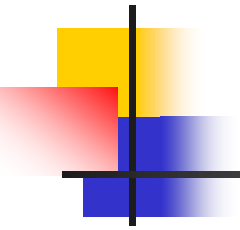


WHY DATA MINING?

- A lot of information, but no knowledge
- Big money spent to collect data,
  but not taking advantage of the information hidden deep inside.
- Data mining uncovers patterns and trends from existing databases
  - Companies are more competitive: better insight to make better decisions.

NEED FOR HUMAN DIRECTION OF DATA MINING

- DM algorithms are automated, but…?
- Imagine a black box capable of answering any question it is asked. Any question.
- Will this eliminate need for human participation?
- Opposite, a human interface problem:
  - How do I *phrase* the question correctly?
  - How do I *set up the parameters* for the particular case I am interested in?
  - How do I get the results in *reasonable time* and in a *form* that I can *understand*?



DANGER! DATA MINING IS EASY TO DO BADLY

- A little knowledge is dangerous:
  - analyses carried out on unpreprocessed data can lead to erroneous conclusions
  - inappropriate approach to data set
  - these errors can lead to very expensive failures.
- The best way: instead of black-box, apply a "white-box" methodology.

CROSS-INDUSTRY STANDARD PROCESS (CRISP-DM)

- A standard that is industry-neutral, tool-neutral, and application-neutral.
- A data mining project has a life cycle consisting of six phases.
- The phase sequence is adaptive:
  - the next phase depends on the outcomes of the preceding phase.

CROSS-INDUSTRY STANDARD PROCESS (CRISP-DM)

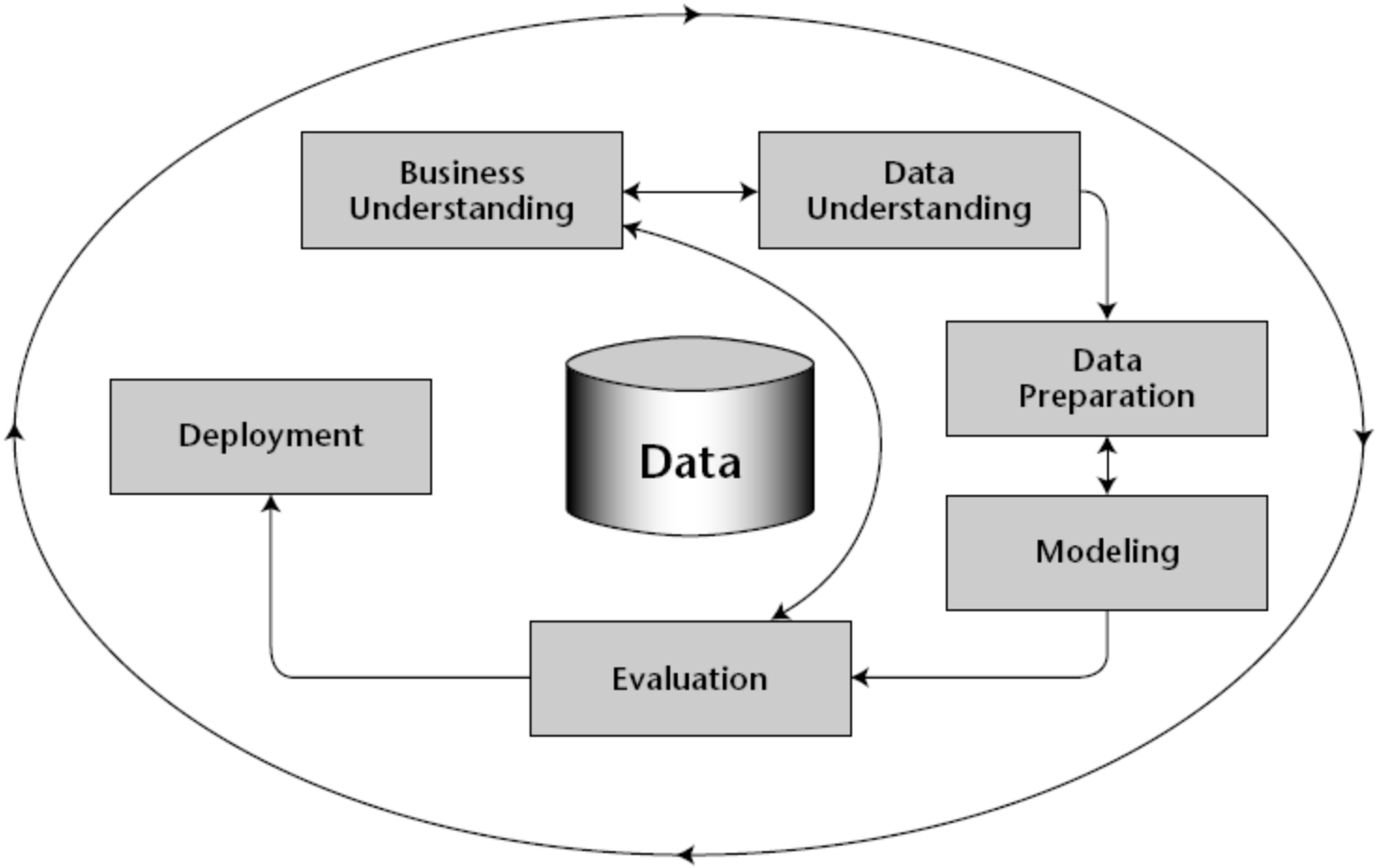


Figure 1.9 Phases of CRISP-DM reference model

CROSS-INDUSTRY STANDARD PROCESS: CRISP-DM

- CRISP-DM: The Six Phases

1. Business understanding phase.
   a. Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole.
   b. Translate these goals and restrictions into the formulation of a data mining problem definition.
   c. Prepare a preliminary strategy for achieving these objectives.
2. Data understanding phase
   a. Collect the data.
   b. Use exploratory data analysis to familiarize yourself with the data and discover initial insights.
   c. Evaluate the quality of the data.
   d. If desired, select interesting subsets that may contain actionable patterns.
3. Data preparation phase
   a. Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive.
   b. Select the cases and variables you want to analyze and that are appropriate for your analysis.
   c. Perform transformations on certain variables, if needed.
   d. Clean the raw data so that it is ready for the modeling tools.

CROSS-INDUSTRY STANDARD PROCESS: CRISP-DM

4. Modeling phase
   a. Select and apply appropriate modeling techniques.
   b. Calibrate model settings to optimize results.
   c. Remember that often, several different techniques may be used for the same data mining problem.
   d. If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique.
5. Evaluation phase
   a. Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field.
   b. Determine whether the model in fact achieves the objectives set for it in the first phase.
   c. Establish whether some important aspect of the business or research problem has not been accounted for sufficiently.
   d. Come to a decision regarding use of the data mining results.
6. Deployment phase
   a. Make use of the models created: Model creation does not signify the completion of a project.
   b. Example of a simple deployment: Generate a report.
   c. Example of complex deployment: Implement a parallel data mining process in another department.
   d. For businesses, the customer often carries out the deployment based on your model.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

WHAT TASKS CAN DATA MINING ACCOMPLISH?

Description

- Example: Uncover evidence why those who have been laid off are less likely to support the present incumbent in the presidential election?
- Descriptions often suggest possible *explanations* for patterns and trends:
  - For example, those who are laid off are now less well off financially than before the incumbent was elected, and so would tend to prefer an alternative.
- Data mining models should be as transparent as possible:
  - Some data mining methods are more transparent than others and suited to *interpretation*.
  - Decision trees provide an intuitive and human friendly explanation of their results.
  - Neural networks are opaque to nonspecialists, due to the nonlinearity and complexity of the model.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

Classification

- In classification, there is a target categorical variable, such as:
  - income, which could be partitioned into three classes or categories: high, middle, and low.
- Classification refers to assigning cases into categories based on available attributes:
  - new case contains a set of attributes, and the class attribute (target) is unknown.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

TABLE 1.1 Excerpt from Data Set for Classifying Income

Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software engineer	High
002	28	M	Marketing consultant	Middle
003	35	M	Unemployed	Low
⋮				

- For example, in the starting data set older females may be associated with the high-income bracket. This data set is called the *training set*.
- Then the algorithm would look at new records, which has no information about income bracket.
- Based on the classifications in the training set, the algorithm would assign classifications to the new records.
- For example, a 63-year-old female professor might be classified in the high-income bracket.

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- Other examples of classification tasks:
  - Determining whether a particular credit card transaction is fraudulent
  - Assessing whether a mortgage application is a good or bad credit risk
  - Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat

WHAT TASKS CAN DATA MINING ACCOMPLISH?

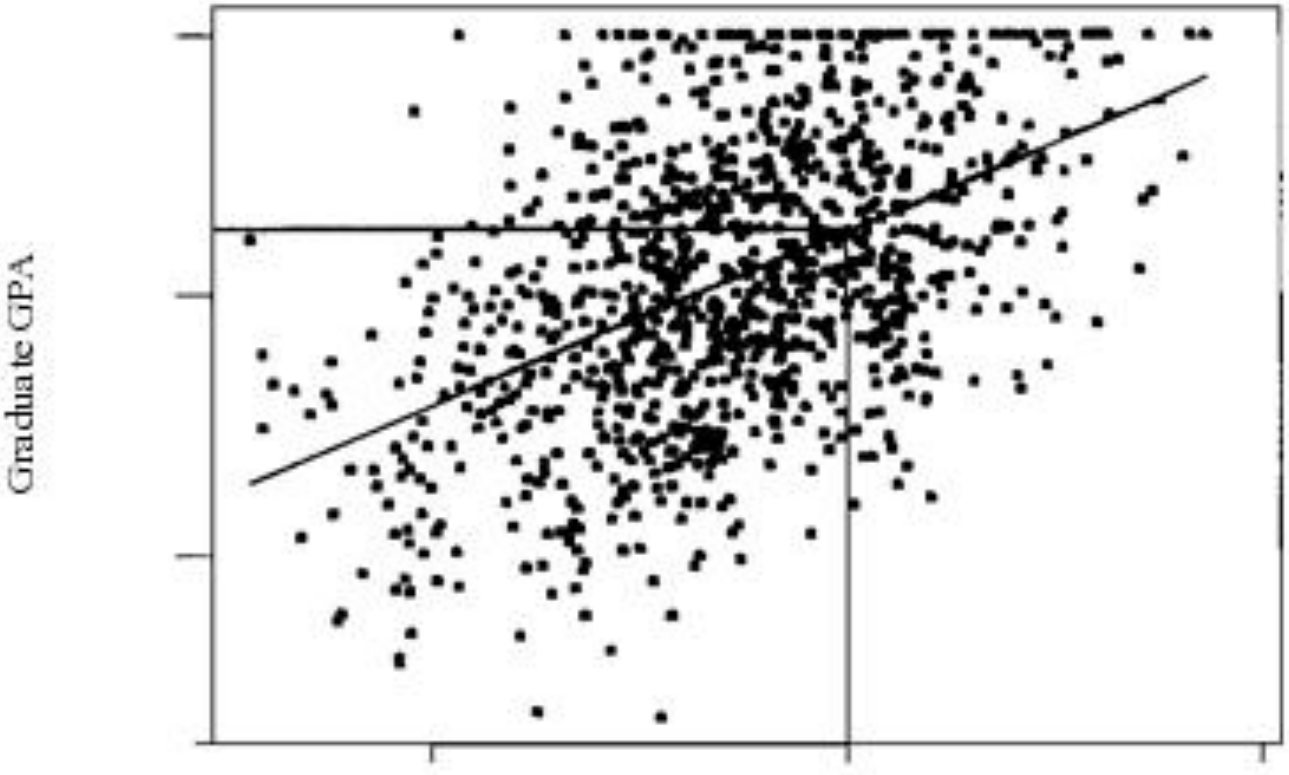
- **Estimation**

  Estimation is similar to classification except that the target variable is numerical.

  For example:

  - Estimating the systolic blood pressure based on the patient's age, gender, body-mass index, and blood Na levels.
  - The relationship between systolic blood pressure and the predictor variables in the training set would provide us with an estimation model.
  - We can then apply that model to new cases.
- Other examples of estimation tasks:
  - Estimating the grade-point average (GPA) of a graduate student, based on that student's undergraduate GPA.

WHAT TASKS CAN DATA MINING ACCOMPLISH?



WHAT TASKS CAN DATA MINING ACCOMPLISH?

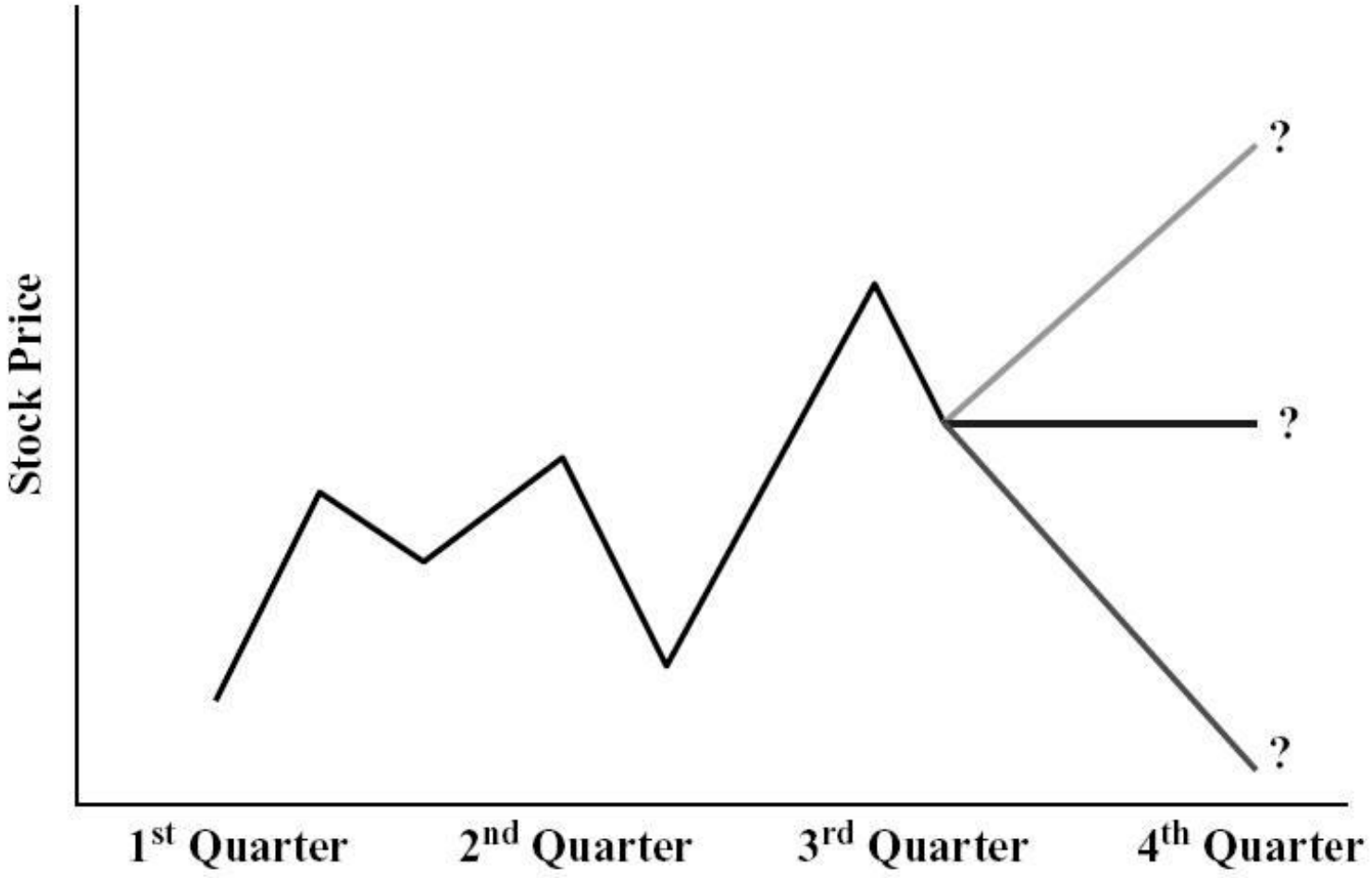
Prediction

Prediction is similar to classification and estimation, except that for prediction, the results lie in the future.

- Any of the methods and techniques used for classification and estimation may also be used for prediction.

- Examples of prediction:
  - Predicting the percentage increase in traffic deaths next year if the speed limit is increased
  - Predicting the price of a stock three months into the future

WHAT TASKS CAN DATA MINING ACCOMPLISH?



Predicting the price of a stock three months in the future

WHAT TASKS CAN DATA MINING ACCOMPLISH?

- **Clustering** refers to the grouping of records, observations, or cases into classes of similar objects.
- A cluster is a collection of records that are similar to one another, and dissimilar to records in other clusters.
- Clustering differs from classification: there is no target variable for clustering.
- Clustering algorithms seek to segment the entire data set into homogeneous subgroups or clusters:
  - where the similarity of the records within the cluster is maximized and the similarity to records outside the cluster is minimized.
- Examples of clustering include:
  - As a dimension-reduction tool when the data set has hundreds of attributes
  - For gene expression clustering, where very large quantities of genes may exhibit similar behavior

WHAT TASKS CAN DATA MINING ACCOMPLISH?

Association

- Data mining association task is to find which attributes "go together."
- Goal is to find rules that are quantifying the relationship between two or more attributes.
- Association rules represent "If antecedent, then consequent," with a measure of the *support* (zastupljenosti) and *confidence* (pouzdanosti).
- Examples of association tasks:
  - Finding out which items in a supermarket are purchased together and which items are never purchased together
  - Example: a supermarket may find that of the 1000 customers, 200 bought diapers, and of those 200 who bought diapers, 50 bought beer.
  - Thus, the association rule would be "If buy diapers, then buy beer" with a support of 200/1000 = 20% and a confidence of 50/200 = 25%.