Introduction to Data Mining

Methodology CRISP-DM

PL02

DATA MINING

APPLICATIONS GENERATE HUGE AMOUNTS OF DATA

WWW, computer systems/programs, biology experiments, Business transactions, Scientific computation and simulation, Medical and person data, Surveillance video and pictures, Satellite sensing, Digital media

TECHNOLOGIES ARE AVAILABLE TO COLLECT AND STORE DATA

Bar codes, scanners, satellites, cameras etc.

Databases, data warehouses, variety of repositories …

We are drowning in data, but starving for knowledge!

DATA MINING

DATA MINING (KNOWLEDGE DISCOVERY FROM DATA):

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

KEY CHARACTERISTICS:

Combination of Theory and Application

Engineering Process

- CRISP-DM

Collection of Functionalities

- Different Tasks and Algorithms

Interdisciplinary Field

DATA MINING

APLICATIONS OF DATA MINING:

Customer relationship management: develop loyalty, implemente customer focused strategies;

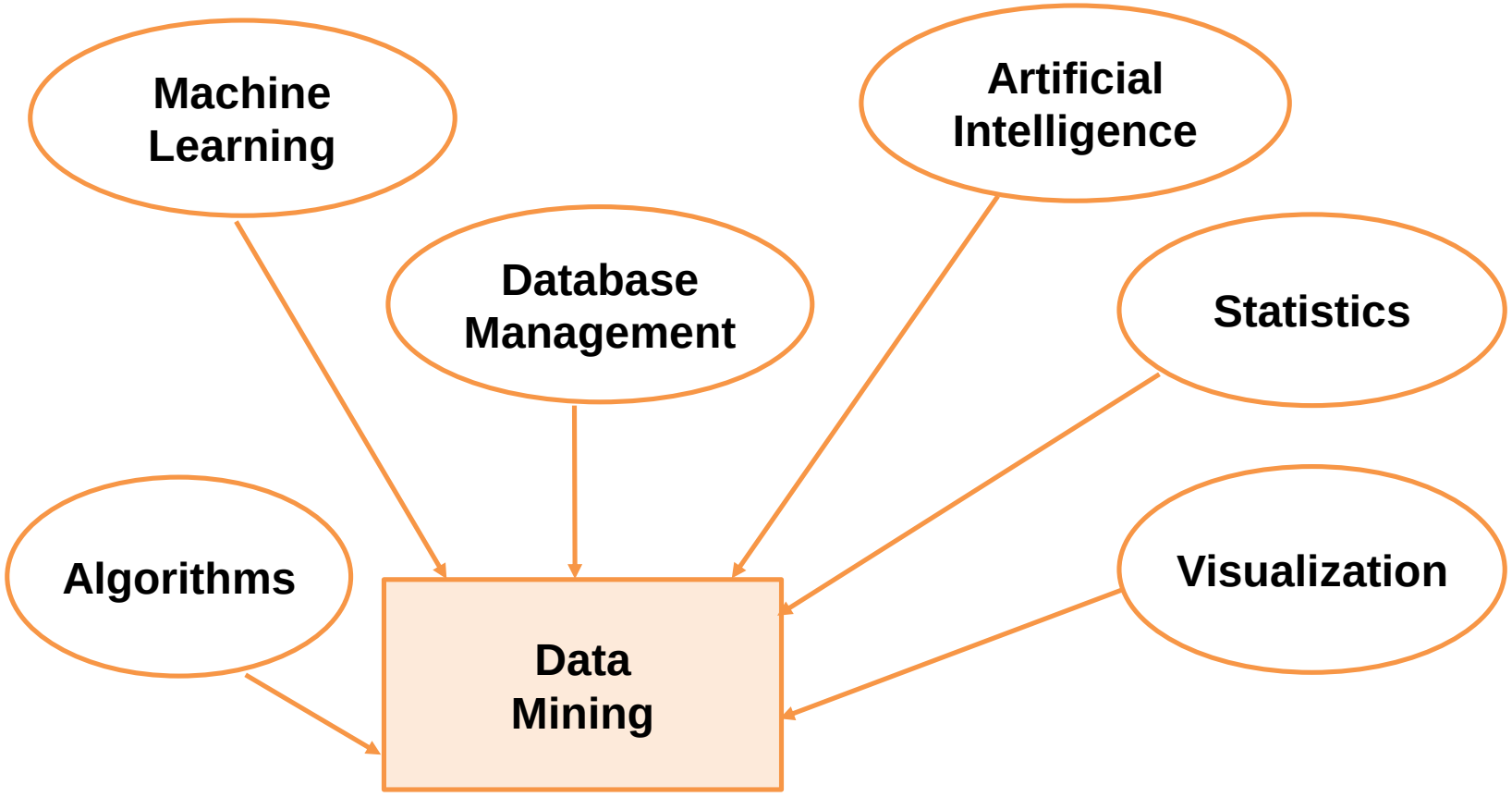
Financial data Analysis: finding patters, causalities and correlations in business information and marker prices;

Supermarket basket analysis: understand buyer's needs and change the store layout accordingly;

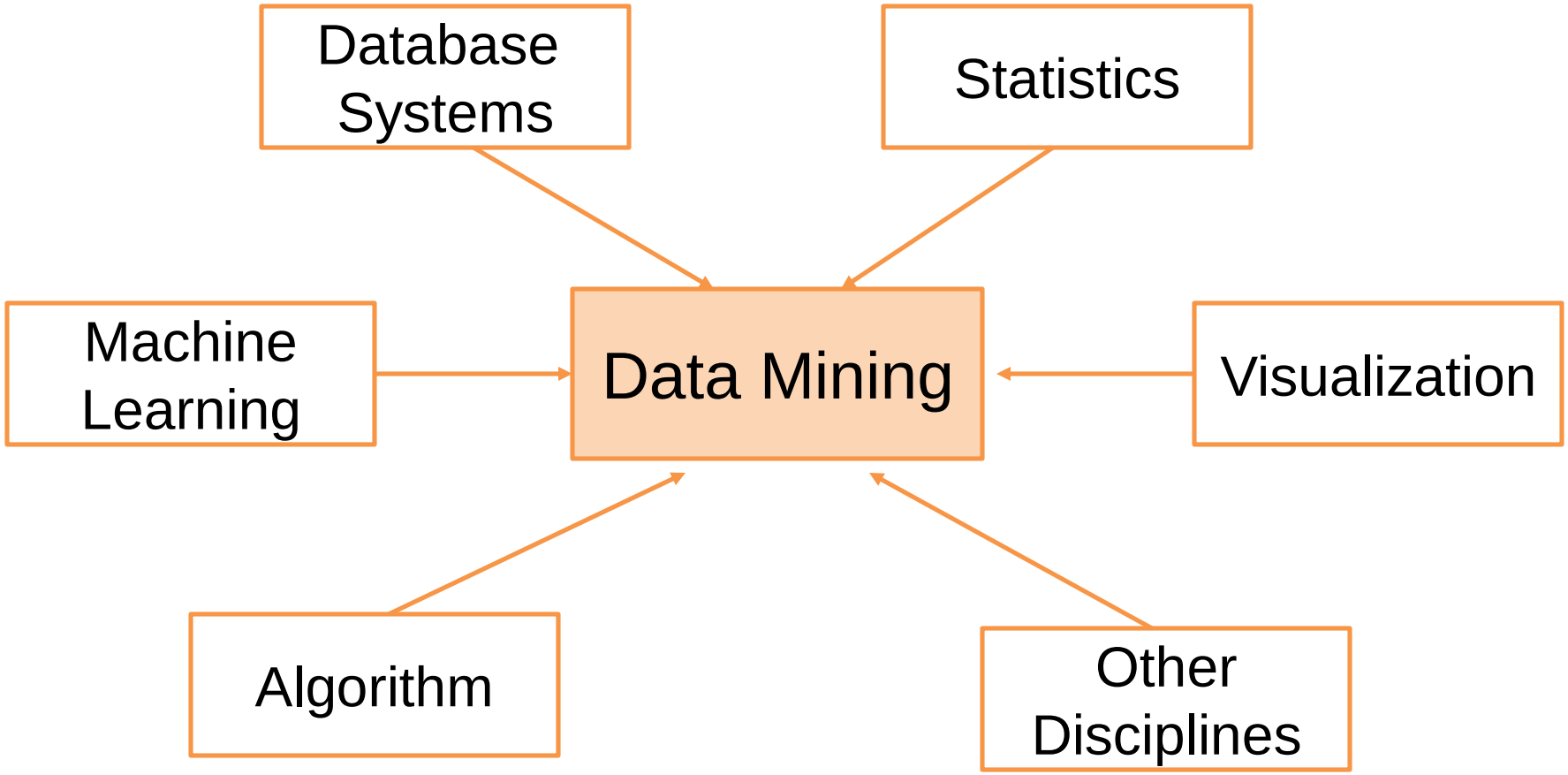
Healthcare

improve care and reduce costs. Predict the number of patients in the ER, and the length of stay of those patients.

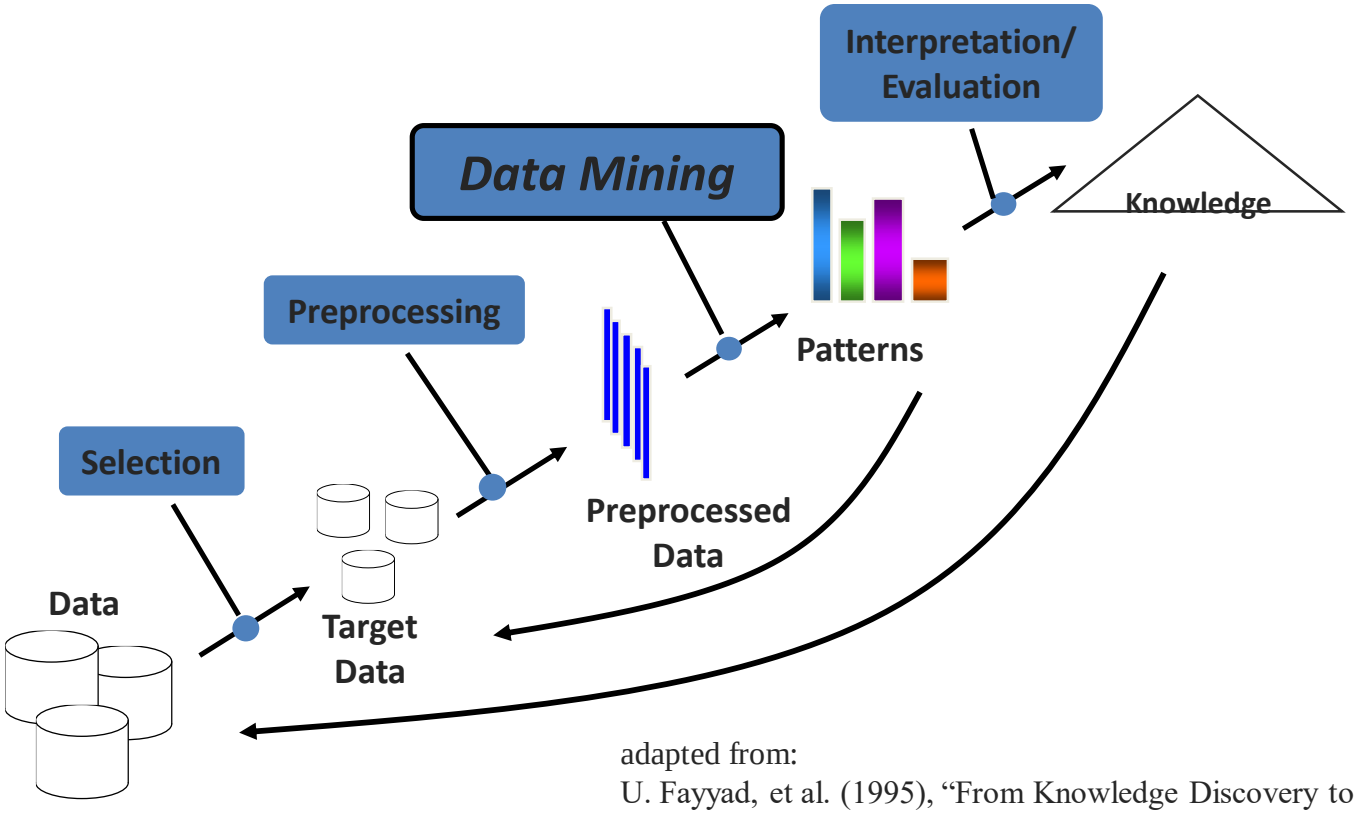
DATA MINING


Machine Learning
Artificial Intelligence
Database Management
Statistics
Algorithms
Visualization
Data Mining

DATA MINING


Database Systems
Statistics
Machine Learning
Data Mining
Visualization
Algorithm
Other Disciplines

DATA MINING



adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

DATA MINING

- Cluster
- Classify
  - Categorical, Regression
- Summarize
  - Summary statistics, Summary rules
- Link Analysis / Model Dependencies
  - Association rules
- Sequence analysis
  - Time-series analysis, Sequential associations
- Detect Deviations

DATA MINING

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Diaper → Beer [0.5%, 75%]
- Classification and Prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Predict some unknown or missing numerical values

DATA MINING

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

CRISP-DM

CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

CRISP-DM

European Community funded effort to develop framework for data mining tasks

Goals:

- Encourage interoperable tools across entire data mining process
- Take the mystery/high-priced expertise out of simple data mining tasks

CRISP-DM

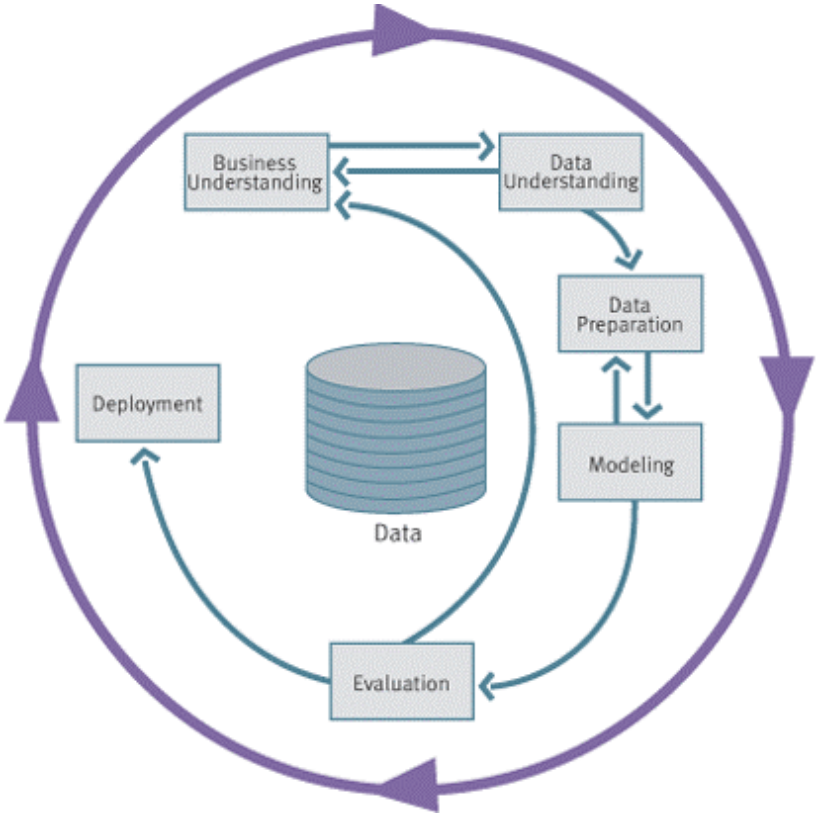
The data mining process must be reliable and repeatable by people with little data mining background !!

CRISP-DM

FEATURES:

- Framework for recording experience
- Allows projects to be replicated
- Aid to project planning and management
- “Comfort factor” for new adopters
- Demonstrates maturity of Data Mining
- Reduces dependency on “stars”

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Situation Assessment** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goal** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Asessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

CRISP-DM

- **Business Understanding**
  - Understanding project objectives and requirements
  - Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization
  - Identify data quality issues
  - Initial, obvious results
- **Data Preparation**
  - Record and attribute selection
  - Data cleansing
- **Modeling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives
  - Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice
  - Set up for repeated/continuous mining of the data

CRISP-DM PHASE 1 - BUSINESS UNDERSTANDING

- Statement of Business Objective
  - States goal in business terminology
- Statement of Data Mining objective
  - States objectives in technical terms
- Statement of Success Criteria

Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives

What the client really wants to accomplish?

Uncover important factors (constraints, competing objectives)

CRISP-DM PHASE 1 - BUSINESS UNDERSTANDING

Determine business objectives

- Key persons and their roles? Is there a steering committee. Internal sponsor (financial, domain expert).
- Business units impacted by the project (sales, finance,...)? Business success criteria and who assesses it?
- Users' needs and expectations.
- Describe problem in general terms. Business questions, Expected benefits.

Assess situation

- Are they already using data mining.
- Identify hardware and software available. Identify data sources and their types (online, experts, written documentation).
- Identify knowledge sources and types (online, experts, written documentation)
- Describe the relevant background.

CRISP-DM PHASE 1 - BUSINESS UNDERSTANDING

Determine data mining goals

- Translate the business questions to data mining goals

  (e.g., a marketing campaign requires segmentation of customers in order to decide whom to approach in this campaign; the level/size of the segments should be specified).

- Specify data mining problem type

  (e.g., classification, description, prediction and clustering).

- Specify criteria for model assessment.

Produce project plan

- Define initial process plan; discuss its feasibility with involved personnel.
- Put identified goals and selected techniques into a coherent procedure.
- Estimate effort and resources needed; Identify critical steps.

CRISP-DM PHASE 2 – DATA UNDERSTANDING

- Acquire the data
- Explore the data (query & visualization)
- Verify the quality

Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

CRISP-DM PHASE 2 – DATA UNDERSTANDING

Collect data

- List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).

Describe data

- Check data volume and examine its gross properties.•
- Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
- Understand the meaning of each attribute and attribute value in business terms.
- For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

CRISP-DM PHASE 2 – DATA UNDERSTANDING

Explore data

Analyze properties of interesting attributes in detail

Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses

Verify data quality

Identify special values and catalogue their meaning.

Does it cover all the cases required? Does it contain errors and how common are they?

Identify missing attributes and blank fields. Meaning of missing data.

Do the meanings of attributes and contained values fit together?

Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).

Check for plausibility of values, e.g. all fields have the same or nearly the

same values.

CRISP-DM PHASE 3 – DATA PREPARATION

Construct data

Derived attributes.

Background knowledge .

How can missing attributes be constructed or imputed?

Integrate data

Integrate sources and store result (new tables and records).

Format Data

Rearranging attributes (Some tools have requirements on the order of the attributes, e.g. first field being a unique identifier for each record or last field being the outcome field the model is to predict).

Reordering records (Perhaps the modeling tool requires that the records be sorted according to the value of the outcome attribute).

Reformatted within-value (These are purely syntactic changes made to satisfy the requirements of the specific modeling tool, remove illegal characters, uppercase lowercase).

CRISP-DM PHASE 4 – MODELING

- Select the modeling technique
  - Based upon the data mining objective
- Generate test design
  - Procedure to test model quality and validity
- Build model
  - Parameter settings
- Assess model (rank the models)

Various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often necessary

CRISP-DM PHASE 4 – MODELING

Select modeling technique

- Select technique
- Identify any built-in assumptions made by the technique about the data (e.g. quality, format, distribution).
- Compare these assumptions with those in the Data Description Report and make sure that these assumptions hold.
- Preparation Phase if necessary.

Generate test design

- Describe the intended plan for train, test and evaluate the models.
- How to divide the dataset into training, test and validation sets.
- Decide on necessary steps (number of iterations, number of folds etc.).
- Prepare data required for test

CRISP-DM PHASE 4 – MODELING

Build model

- Set initial parameters and document reasons for choosing those values.
- Run the selected technique on the input dataset. Post-process data mining results (eg. editing rules, display trees).
- Record parameter settings used to produce the model.
- Describe the model, its special features, behavior and interpretation.

Assess model

- Evaluate result with respect to evaluation criteria. Rank results with respect to success and evaluation criteria and select best models.
- Interpret results in business terms. Get comments by domain experts.
- Check plausibility of model.
- Check model against given knowledge base (discovered info. novel and useful?)
- Check result reliability. Analyze potentials for deployment of each result.

CRISP-DM PHASE 5 – EVALUATION

- More thoroughly evaluate model
- Decide how to use results

- Methods and criteria depend on model type:

e.g., coincidence matrix with classification models, mean error rate with regression models

Interpretation of model: important or not, easy or hard depends on algorithm

Determine if there is some important business issue that has not been sufficiently considered.

A decision on the use of the data mining results should be reached

CRISP-DM PHASE 5 – EVALUATION

Evaluate results

- Understand data mining result. Check impact for data mining goal.
- Check result against knowledge base to see if it is novel and useful.
- Evaluate and assess result with respect to business success criteria
- Rank results according to business success criteria. Check result impact on initial application goal.
- Are there new business objectives? (address later in project or new project?)
- State conclusions for future data mining projects.

Review of process

- Summarize the process review (activities that missed or should be repeated).
- Overview data mining process. Is there any overlooked factor or task?
- (did we correctly build the model? Did we only use attributes that we are allowed to use and that are available for future analyses?)
- Identify failures, misleading steps, possible alternative actions, unexpected paths
- Review data mining results with respect to business success

CRISP-DM PHASE 5 – EVALUATION

Determine next steps

- Analyze potential for deployment of each result. Estimate potential for improvement of current process.
- Check remaining resources to determine if they allow additional process iterations (or whether additional resources can be made available).
- Recommend alternative continuations. Refine process plan.

Decision

- According to the results and process review, it is decided how to proceed to the next stage (remaining resources and budget)
- Rank the possible actions. Select one of the possible actions.
- Document reasons for the choice.

CRISP-DM

WHY?

The data mining process must be reliable and repeatable by people with little data mining skills

CRISP-DM provides a uniform framework for

- guidelines
- experience documentation

CRISP-DM is flexible to account for differences

- Different business/agency problems
- Different data

EXAMPLES

Step Towards Prediction of Perineal Tear

Francisca Fonseca (2017)

Predicting the need of Neonatal Resuscitation using Data Mining

Ana Morais (2017)

Understanding Stroke in Dialysis and Chronic Kidney Disease

Mariana Rodrigues (2017)

(...)

EXERCISES

FE01

Introduction to Data Mining

Methodology CRISP-DM

PL02