The Data Mining Workflow

ABSTRACT: In a lot of the lectures the speakers described a similar workflow and software stack for data mining processes. Lennart Heuckendorf gave an overview over the software stack typically used. This poster generalizes the typical data mining workflow seen in the lectures and shows typical software used in each step.

Raw Data

Event Data

Excel Sheets

Databases

1. Data Integration & Data Cleaning

Remove inconsistencies and bring data from heterogeneous sources into an uniform format.
Therefore one needs to deal e. g. with missing values, mapping attribute names and duplicates.

Target Data

Data Warehouse

Selections &

Projections

Relevant Data

2. Data Mining

Depending on the task, techniques are chosen and implemented to find new information in the data. The techniques reach from simple statistics, clustering, outlier mining and classification techniques to using neural networks.

Patterns

R (Language)

Patterns

Python

3. Data Visualization

Visualize the patterns one found, or try to find new information through visual data mining.

Visualization

Tableau

Clusters

Graphs

Bar Plots

Gain & **Apply** Knowledge

Lasse Steffen, Bachelor

Hasso Plattner Institute, Potsdam, Germany

E-Mail: lasse.steffen@student.hpi.de

Credit for the Icon goes to Smashicon , Freepik, Vectors market, Wissawa Khamsriwath and Lyolya