# Lab 2: Data preprocessing

19.03.2015

1. Data preparation for knowledge discovery is a crucial issue. A lot of time and effort is put there.
2. Data in the real world is dirty:
   - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
   - noisy: containing errors or outliers
   - inconsistent: containing discrepancies (disagreements) in codes or names
3. Why is data preprocessing important? Low quality data implies low quality mining results.
4. Data extraction, cleaning and transformation comprises the majority of the work of building a data warehouse.
5. Basic problems in Data Cleaning:
   - Data acquisition / integration and metadata
   - Unified formats and other transformations
     - Normalization
     - Discretization – transformation of numerical values into codes / values of ordered subintervals defined over the domain of an attribute

$$Ent(S,T) = \frac{\|S_1\|}{\|S\|} Ent(S_1) + \frac{\|S_2\|}{\|S\|} Ent(S_2)$$

$$Ent(S) = \sum_k (-p_k \log_2(p_k))$$

Rysunek 1: Entropy-Based Discretization

   - Redundant data – Large number of redundant data may slow down or confuse data mining process. Redundant data may be detected thanks to correlation analysis.
   - Erroneous values
     - Inaccurate values
     - Noisy data and incorrect attribute values
     - Duplicate records
     - Outliers – graphical identification
   - Missing values – Missing data may need to be inferred.
     - Ignore / Delete the instance (not effective).
     - Fill in the missing values manually.
     - Fill in a more advanced way: the attribute mean or the most common value.
   - Data validation and statistics
   - Attribute selection – Filter vs. wrapper approach
6. Good data preparation is key to producing valid and reliable models!