Master of Technology in Knowledge/Software Engineering

KE5107: Data Mining Methodology and Methods

Data Preparation and Transformation

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
E-mail: zhenzhen@nus.edu.sg

Module Objectives

- To use different methods to prepare and transform the data for data mining

Agenda

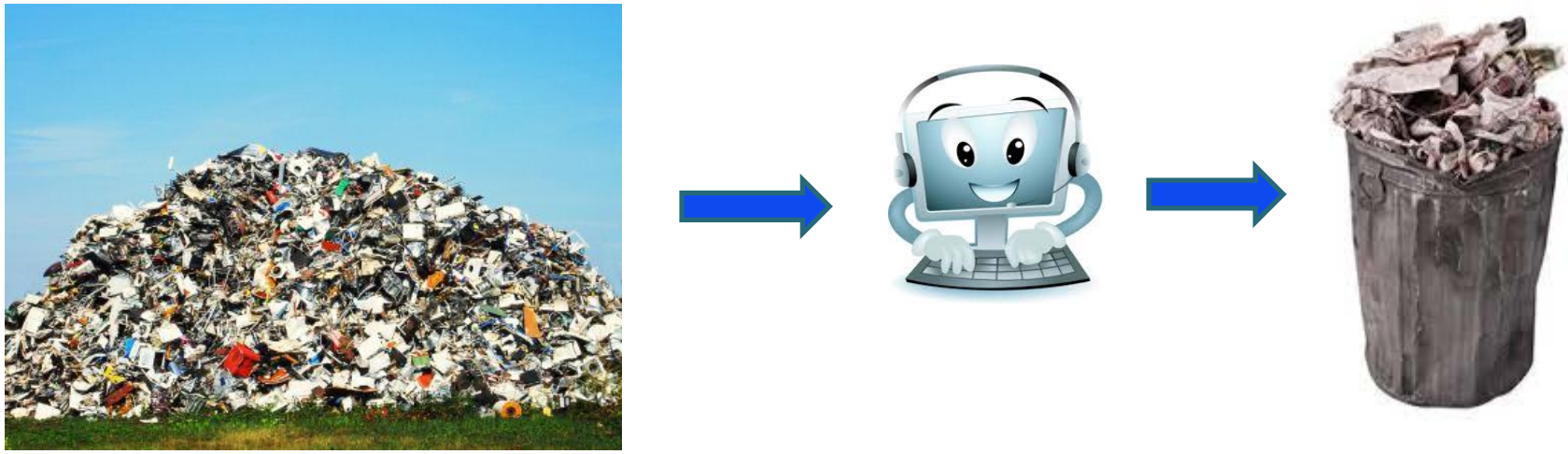
- The need to prepare data
- Major tasks in data preprocessing
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction

Data Preparation and Transformation

- Data Cleaning and Preparation = Data Preprocessing, a critical phase in data mining
- 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
    - e.g., Salary="-10"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1988"

Quality of Data Matters

- No quality data, no quality results
  - e.g., duplicate or missing data may cause incorrect or even misleading statistics



Major Tasks in Data Preprocessing

1. Data cleaning
   - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
2. Data integration
   - Integration of multiple databases
3. Data transformation
   - Normalization and aggregation
4. Data reduction
   - Obtains reduced representation in volume but produces the same or similar analytical results

1. Data Cleaning

- Data may not be perfectly collected, or collected with the right purpose.
- Many reasons exist for data to be dirty:
  - Data entry errors
  - Misplaced decimal points
  - Inherent error in counting or measuring devices
  - External factors, etc.
- Data exploration can discover anomalous patterns, leading to the questioning of data quality
  - E.g. categories with very low frequency counts ➔ mistyping?
  - Name and addresses recorded in multiple ways in data integrated from multiple sources (can be up to 20~30 variations)
  - Missing data

Data Cleaning Tasks

- Data cleaning tasks
  - Handle missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Values

- Common feature of any dataset
- Various reasons:
  - Information not available
  - Lost data
  - Purposefully left out with a reason
- Might be marked with a special value
  - E.g. *"9999", "1 Jan 1900", "*", "?", "#", "$"*, etc
- The presence of missing values in data can make problems for the modeling tools.

Handling Missing Values

- Simply ignore the data sample with missing values
  - Throwing away data, especially poor when the percentage of missing values per attribute varies considerably

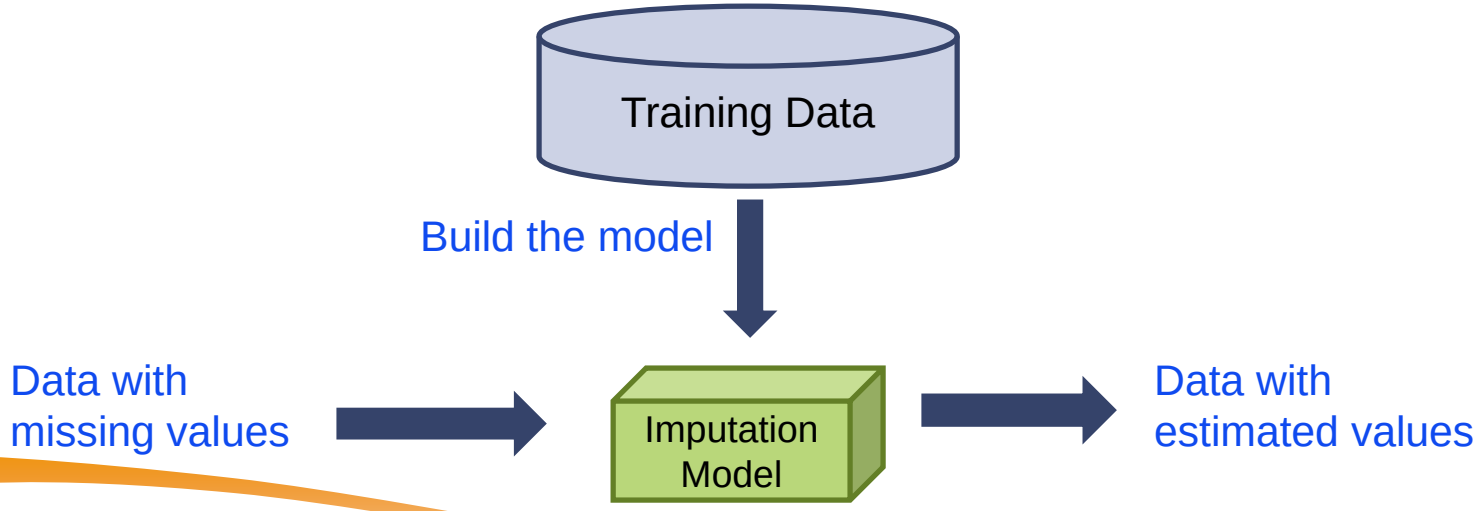
- Fill in the missing values – **data imputation**
  - Principle: Avoid adding bias and distortion to the data
  - Understand why the data is missing can help guide the imputation
    - E.g. for '*rainfall*' variable, a missing value may mean no rain recorded on that day ➔ 0

Data Imputation

- Fill in the missing values automatically
  - a global constant : e.g., “unknown”
    - Straight forward approach
    - Modeling algorithms may mistakingly treat “unknown” as a concept
  - the attribute mean (or median, mode)
    - Simple and quick, been found empirically useful, though not always satisfactory
  - the attribute mean for all samples belonging to the same class
    - Better estimate than attribute mean

Data Imputation

- Train a prediction model (E.g. regression model, decision tree, k-Nearest Neighbor) to predict the **most probable** value
  - Use variables containing values to estimate the variable with missing values
  - Can produce good estimates
  - Need training data and additional modeling



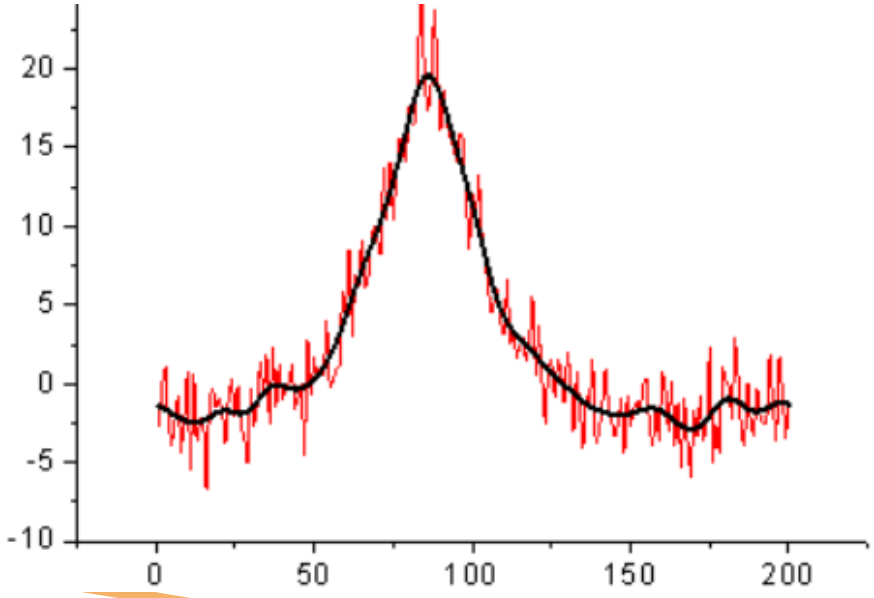
Noisy Data

- Noise: random error or variance in a measured variable

- Incorrect attribute values may have been entered due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

Handling Noisy Data

- Binning method:
  - first sort data and partition into bins
  - then one can smooth the data by bin means, by bin median, by bin boundaries, etc.
  - Equal width binning, equal frequency binning , quantile binning, etc.



Handling Noisy Data

- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Outliers

- Observations that "deviate so much from other observations as to arouse suspicion that it was generated by a different mechanism". (Hawkins, 1980)
- Appearing at the maximum or minimum end of a variable, skewing or distorting the distribution
  - E.g. extreme weather conditions on a particular day, a very wealthy person financially very different from the rest of the population, etc.

Dealing with Outliers

- Can be rare, unusual, infrequent events we are interested in. They should be identified for further investigation.
  - E.g. frauds in income tax, insurance, banking, etc.

- Otherwise, outliers usually should be removed to avoid adversely affecting the modeling result (though some algorithms, like random forests and support vector machines can be robust to outliers)

2. Data Integration

- Data integration
  - combines data from multiple sources into a coherent store

- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources may be different
  - possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integrating multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another database, e.g., monthly revenue and annual revenue

- Redundant data may be detected by correlation analysis

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

3. Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization
- Generalization: concept hierarchy climbing
- Log Transformation
- Square Root Transformation
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation

- Log Transformation
  - Makes a skewed attribute more symmetric
  - Reduces the magnitudes
  - Common bases 10, 2, *e*
- Square Root Transformation
  - More discrete values to normal form
- Categorical variable to indicator or dummy variables with value of 0 and 1 (some modelling algorithms include this step)
  - E.g.

Obs.	Colour	Colour_Red	Colour_Green	Colour_Blue
1	Green	0	1	0
2	Blue	0	0	1
3	Blue	0	0	1
4	Red	1	0	0
5	Green	0	1	0
6	Red	1	0	0

Data Transformation: Normalization

- Reduces outlier distortion and enhances linear predictability

- Required before distance-measure-based clustering , to ensure all variables have approximately the same scale
  - E.g. variable *Age* vs *Income*: a distance of 10 "years" may be more significant than a distance of $1000, yet $1000 swamps 10 when they are added in calculating distance

- Normally re-center and rescale the data to be around zero, in the range from 0 to 1, etc.

Common Normalization Methods

- ***min-max*** normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- ***z-score*** normalization: centered around zero

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by **decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max($|v'|$)<1

Attribute/Feature Construction

Product Holdings	Purchased Service
1. ProdA + ProdC	Y
2. ProdB + ProdC	N
3. ProdA + ProdD	N
4. ProdB + ProdD	Y
…	

ProdA	ProdB	ProdC	ProdD	Svc
1. 1	0	1	0	Y
2. 0	1	1	0	N
3. 1	0	0	1	N
4. 0	1	0	1	Y
…				

4. Data Reduction

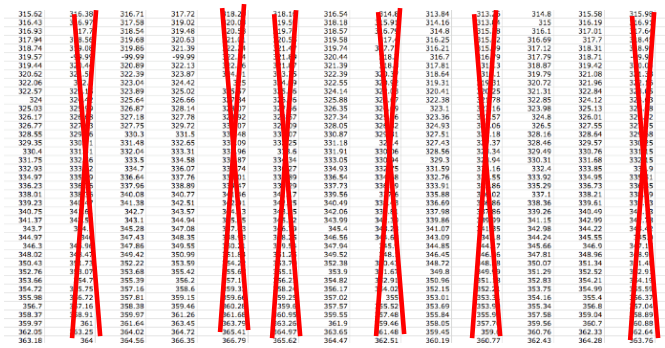
- Complex data analytics may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Dimensionality reduction—reduce number of attributes
  - Numerosity reduction—fit data into models

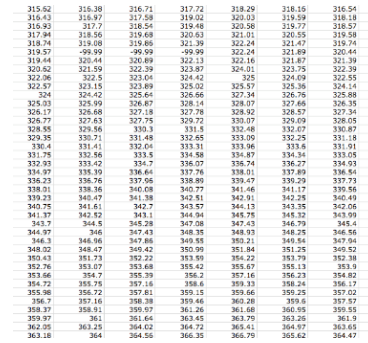
Dimensionality Reduction

- Rational: not all variables contribute useful information
  - E.g. CRM data supplied by French telecom company Orange for 2009 KDD Cup has 15,000 variables. Are they all useful?
  - Irrelevant variables add noise to the data and increases the time and resources required for model building and scoring.
  - Noise increases the size of model and affect model accuracy
  - Groups of correlated variables (possibly measuring the same underlying feature)
  - Some algorithms, like nearest-neighbor learning, are susceptible to irrelevant features

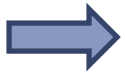
Dimensionality Reduction

- **Feature selection** (i.e., attribute subset selection)
  - selecting the most relevant attributes





- **Feature Extraction**
  - combining attributes into a new reduced set of features



Feature Selection

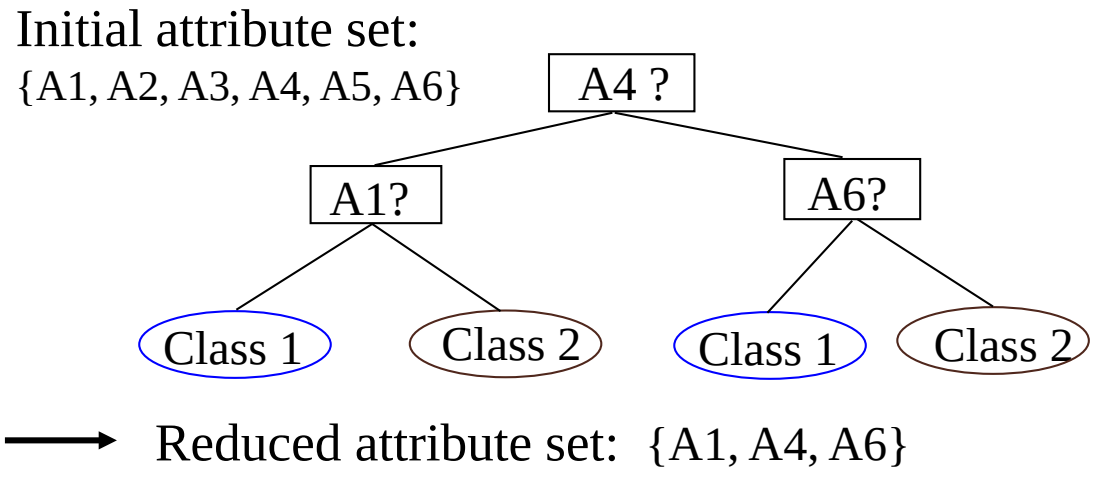
- Select a minimum set of features
  - such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
- Reduce number of features, easier to understand
- Need a good knowledge of the data set and relevance of variables to the problem

Feature Selection

- Heuristic methods searching the attribute space (due to exponential # of choices):
  - Greedy search
  - Need evaluation measure (model accuracy)
  - Step-wise forward selection: start with a best feature and add to it the feature that together with the first provides the best result
  - Step-wise backward elimination: start with all features and remove the features one at a time for improved performance
  - Combining forward selection and backward elimination

Feature Selection

- Wrapper methods - a learning algorithm is wrapped into the selection procedure
  - decision-tree induction
    - E.g. use decision-tree on the full dataset to filter the variables to be used for nearest-neighbor algorithm
  - Algorithms that output coefficients of attributes (e.g. linear SVM)



Feature Selection with Correlation Analysis

- Use correlation analysis to select a subset of attributes that individually correlate well with the target but have little inter-correlation
- Correlation analysis – given two attributes, such analysis measures how strongly one attribute implies the other, based on the available data.
  - Numeric data: correlation coefficient (such as *Pearson's coefficient*)
  - Categorical data: the χ^2 (chi-square) test
- Attributes found correlated with the target attribute are important.
- Attributes strongly correlated with each other are redundant.

Feature Extraction

- Also attribute reduction process by combining the original attributes
- Leading to a much smaller and richer set of attributes
- Methods exist which work well for linear between-variable relationships
  - Principle component analysis
  - Factor analysis

Principal Component Analysis

- Given *N* data vectors from *k*-dimensions, find *c* <= *k* orthogonal vectors that can be best used to represent data
  - The original data set is reduced to one consisting of *N* data vectors on *c* principal components (reduced dimensions)
  - A *component* is an artificially constructed variable that is fitted to all of the original variables in a dataset in such a way that it extracts the highest possible amount of variability
- Each data vector is a linear combination of the *c* principal component vectors
- Works for numeric data only
- Used when the number of dimensions is large
- Doesn't work well with nonlinear relationships

Factor Analysis

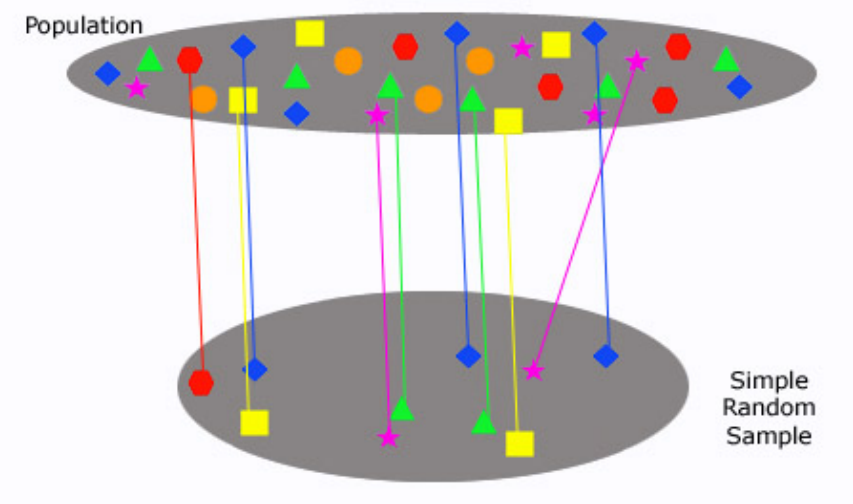
- Also a technique for reducing the complexity of high-dimensional data
- Statistical method to describe variability among observed, *correlated* variables in terms of a potentially lower number of unobserved variables called factors.
  - E.g. for psychiatric data, not possible to measure a factor of interest directly (such as intelligence), but possible to measure other quantities (like student exam scores, IQ test scores) that reflect the factor of interest.
  - Assumes that an unobserved variable is linearly related to two correlated variables
- Related to PCA, but differ in how the reduced dimensionality is constructed
  - Factor analysis uses regression modeling techniques where as PCA is a descriptive technique.
  - PCA explains variability; factor analysis explains correlation

Numerosity Reduction

- Data replaced or estimated by alternative, smaller data representations.
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

Sampling

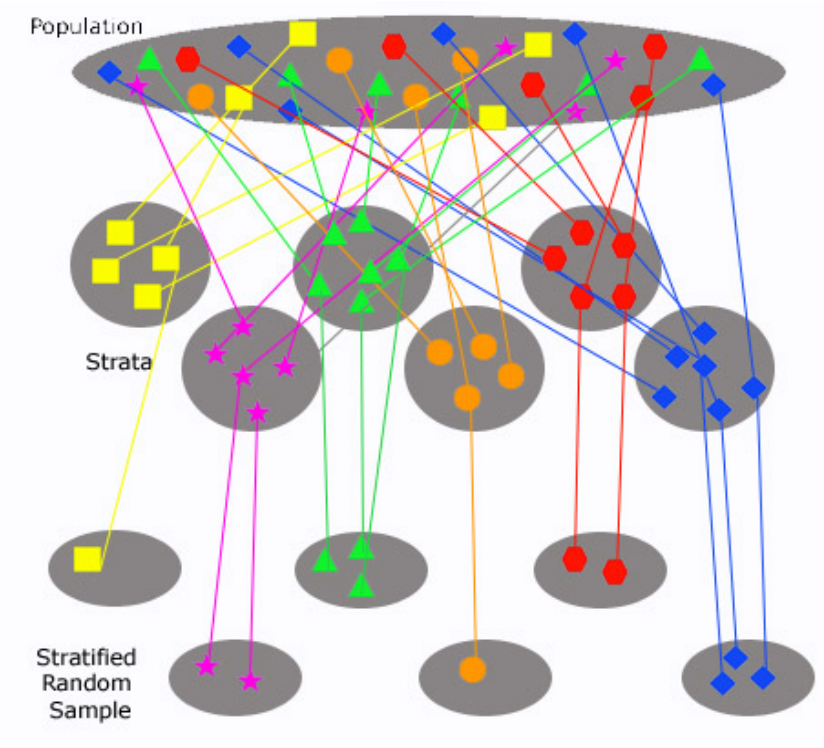
- Simple random sampling
  - without replacement
    - A sampled tuple doesn't participate in the next selection.
    - All tuples are equally likely to be sampled.
    - It avoids choosing any member of the population more than once.
  - with replacement
    - A sampled tuple is put back to the population for next selection



http://facultyweb.berry.edu/nmiller/classinfo/211/sampling_and_error_printout.htm

Sampling

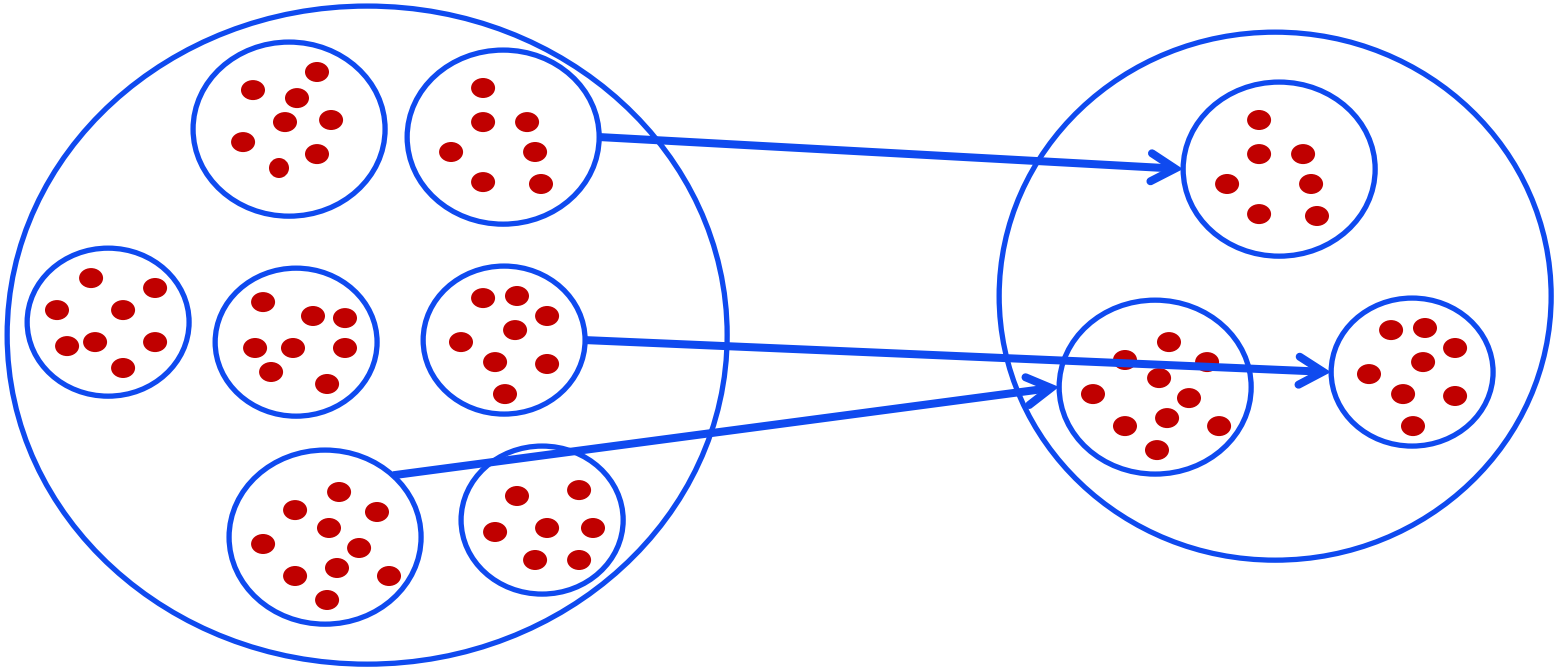
- Stratified sample
  - Divide data into mutually disjoint strata, and take simple random sample from each stratum
  - It helps ensure a representative sample, especially when the data are skewed



http://facultyweb.berry.edu/nmiller/classinfo/211/sampling_and_error_printout.htm

Sampling

- Cluster sampling
  - Group data into clusters, and take simple random sample of clusters instead of individuals



Summary

- Data needs to be prepared into a form suitable for further analysis and modeling.
- It's a tedious and time consuming process but essential for successful data mining.
- Data preparation improves the quality of data and consequently helps improve the quality of data mining results.

References

1. Jiawei Han and Micheline Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann Publishers, August 2000.

2. Tom Mitchell, *Machine learning*, McGraw Hill, 1997.

3. David Hand, Heikki Mannila, Padhraic Smyth, *Principles of data mining* , Cambridge, Mass. : MIT Press, c2001.

4. Graham Williams, *Data mining with Rattle and R*, Springer, 2011.