Intro to Data Science

Lecture 02

Data pre-processing

- Data in the real world is dirty:
  - Incomplete:

    lacking attribute values or lacking certain attributes of interest

    e.g. Occupation=“ ”, year_salary = “13.000”, ...
  - Noisy:

    containing errors or outliers

    e.g., Salary=“-10”, Family=“Unknown”, ...
  - Inconsistent:

    containing discrepancies in codes or names

    e.g. Age=“42” Birthday=“03/07/1997”

    e.g. Previous rating “1,2,3”, Present rating “A, B, C”

Data issues

- Incomplete data may come from “Not applicable” data value when collected:
  - Different considerations between the time when the data was collected and when it is analyzed.
    - Modern life insurance questionnaires would now be: Do you smoke?, Weight?, Do you drink?, …
  - Human/hardware/software problems: forgotten fields…/limited space…/year 2000 problem … etc.
- Noisy data (Incorrect values) may come from-
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission etc.

Data issues

Inconsistent data may come from

- Integration of different data sources
  - e.g. Different customer data, like addresses, telephone numbers; spelling conventions
- Functional dependency violation
  - e.g. Salary changed, while derived values like tax or tax deductions, were not updated

Duplicate records also need data cleaning-

- Which one is correct?
- Is it really a duplicate record?
- Which data to maintain?
  - Jan Jansen, Utrecht, 1-1 2008, 10.000, 1, 2, ...
  - Jan Jansen, Utrecht, 1-1 2008, 11.000, 1, 2, ...

Data pre-processing

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
- A multi-dimensional measure of data quality
  - accuracy, completeness, consistency, timeliness, believability, value added, interpretability, accessibility

Major Tasks of Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, files, or notes
- Data transformation
  - Normalization (scaling to a specific range)
  - Aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: with particular importance, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression

Data Cleaning

- Tasks of Data Cleaning:
  - Fill in missing values
  - Identify outliers and smooth noisy data
  - Correct inconsistent data

Data cleaning:

Fill-in missing values

- Ignore the row:
  - usually done when class label is missing (assuming the task is classification not effective in certain cases)
- Fill in the missing value manually:
  - tedious + infeasible?
- Use a global constant to fill in the missing value:
  - e.g., “unknown”, a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Data cleaning:

Manage Noisy Data

- Binning Method:
  - First sort data and partition into bins then one can smooth by bin means, smooth by bin median, etc
- Clustering:
  - Detect and remove outliers
- Semi Automated:
  - Computer and Manual Intervention
- Regression
  - Use regression functions

Data cleaning:

Manage Noisy Data

Binning:

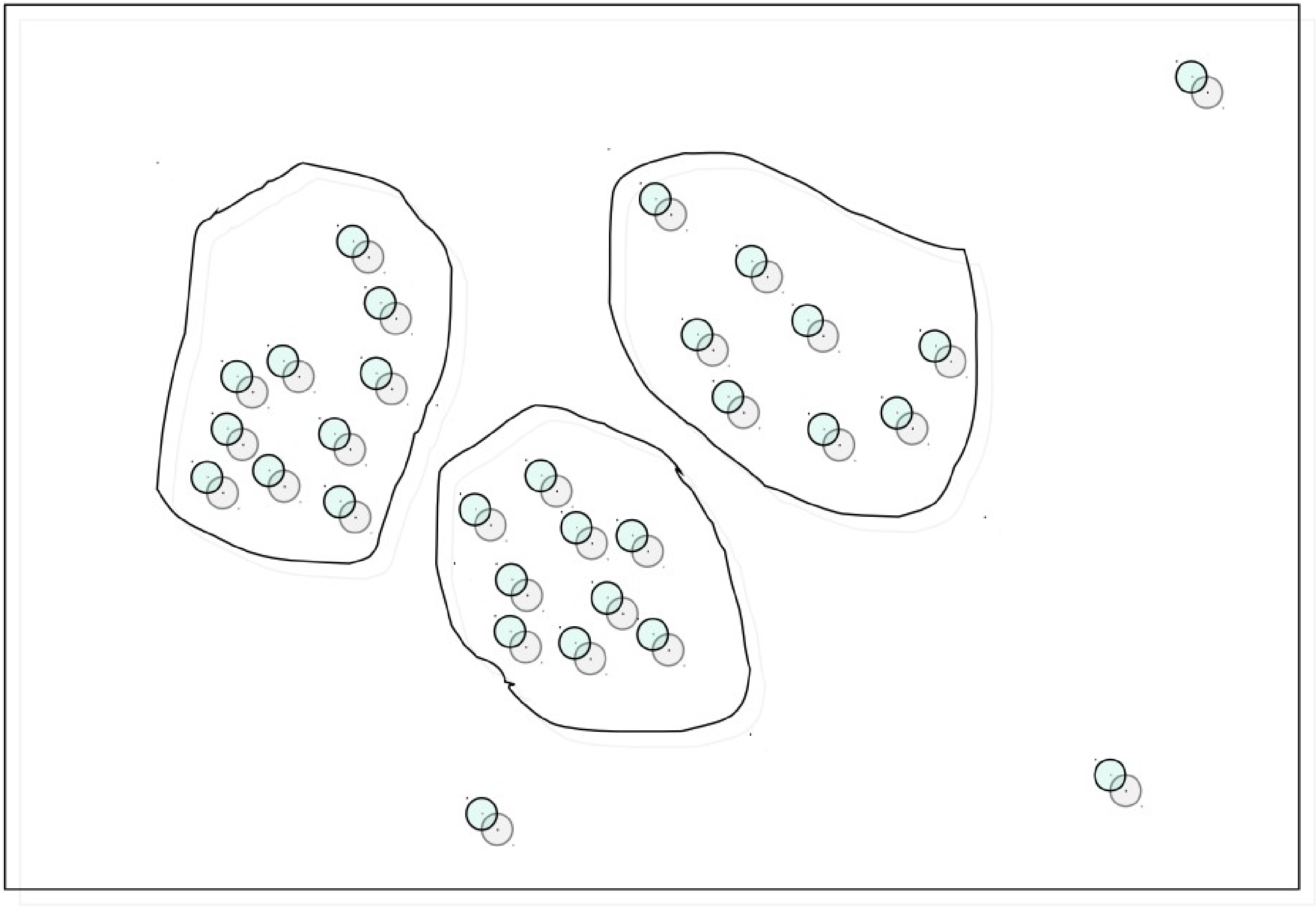
Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into equal-frequency (equi-depth) bins:
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29

Data cleaning:

Manage Noisy Data

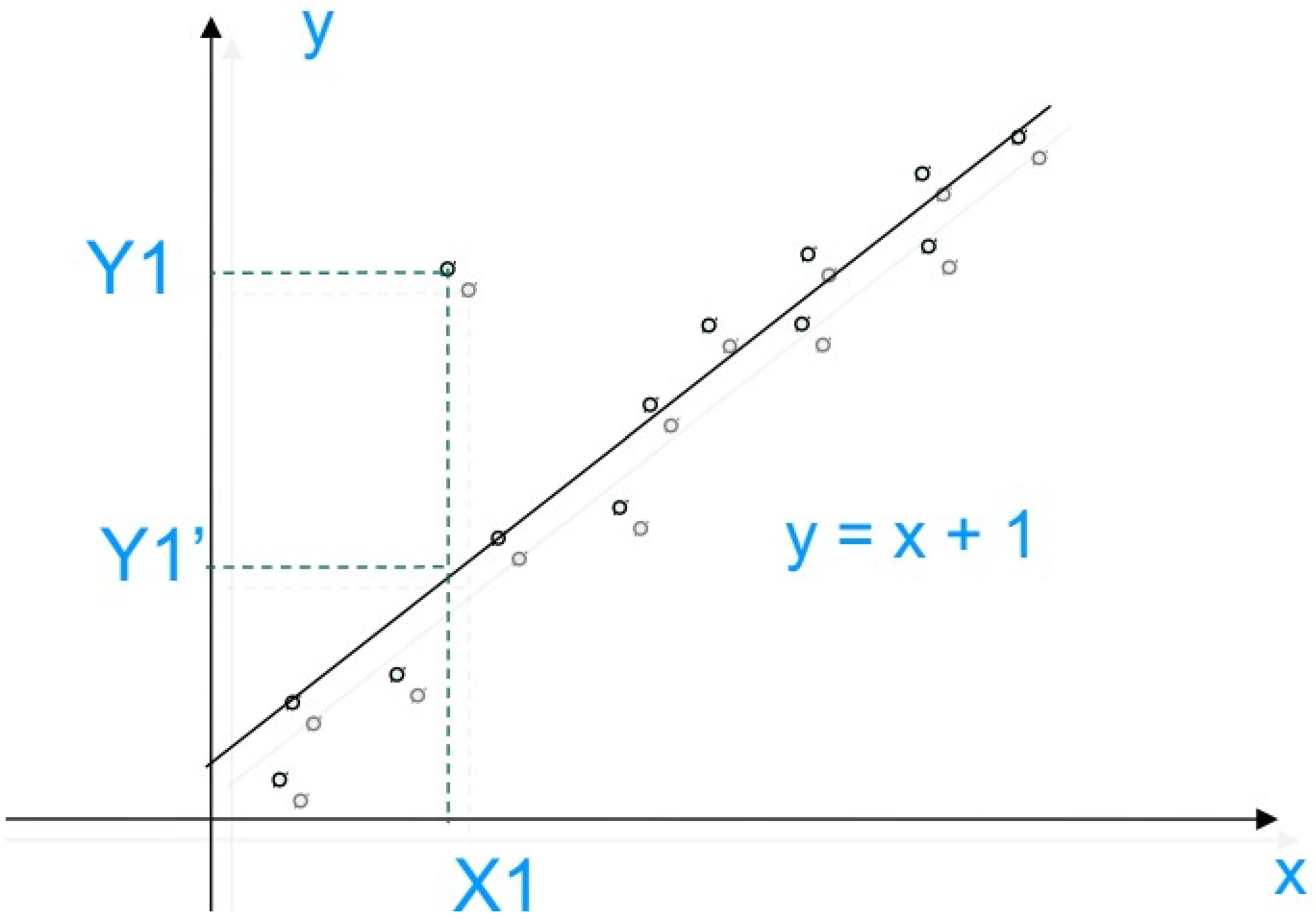
Clustering:



Data cleaning:

Manage Noisy Data

Regression Analysis



Data cleaning:

Inconsistant Data

- Manual correction using external references
- Semi-automatic using Knowledge engineering tools:
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

Data integration and transformation

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration
  - Integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units, different currency

Manage Data Integration

- Redundant data occur when integrating multiple DBs
  - The same attribute may have different names in different databases
  - One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis

$$r_{A,B} = \frac{\Sigma(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A \sigma_B}$$

Use chi 2 test to measure redundancy of categorical data.

- Careful integration helps reduce redundancies and inconsistencies and improve mining speed and quality

Manage Data Transformation

- Smoothing: remove noise from data (binning, clustering, regression)
- Aggregation: summarization, data cube construction
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Manage Data Reduction

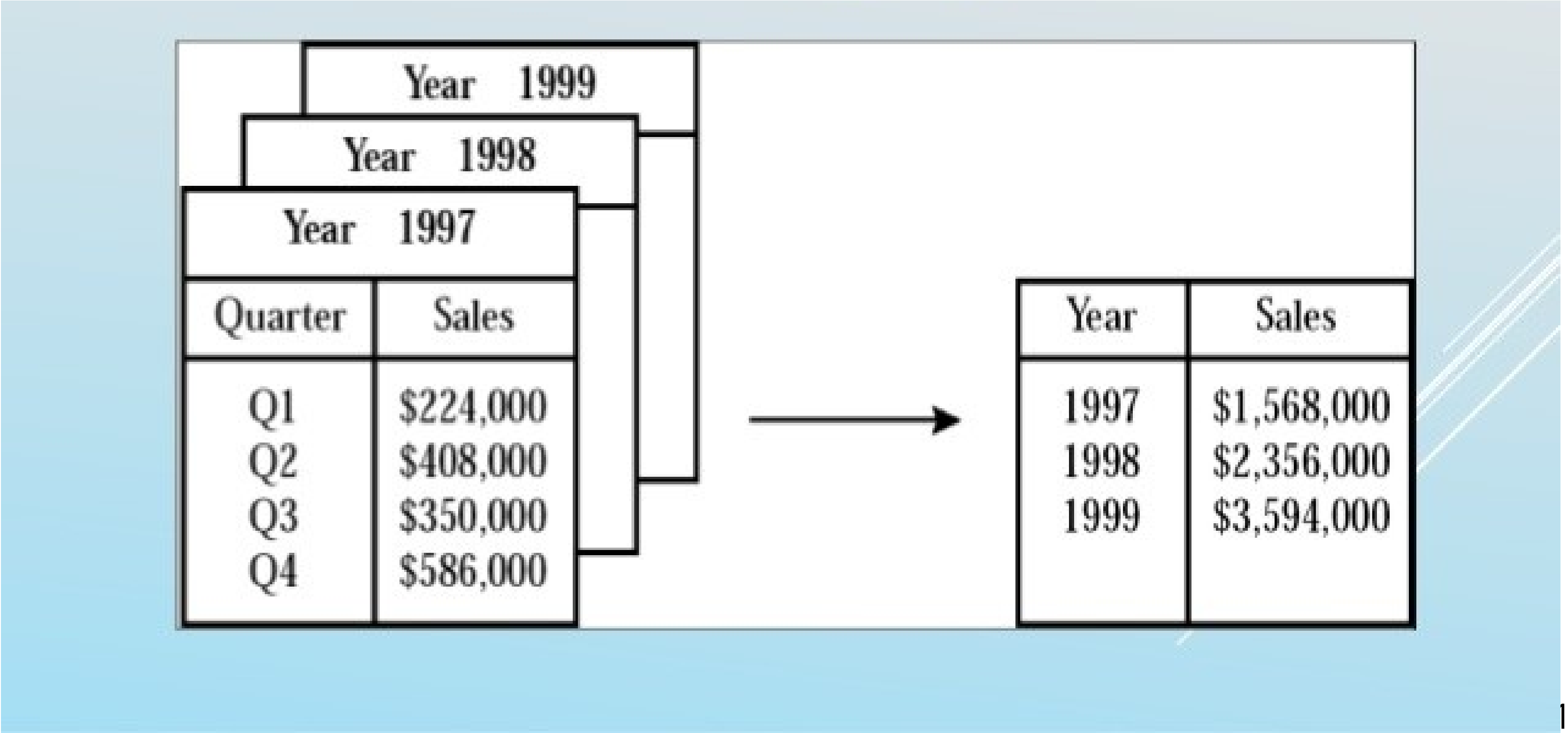
- Data reduction: reduced representation, while still retaining critical information

  - Data cube aggregation
  - Dimensionality reduction
  - Data compression
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Data cube aggeregation

Reduce data to concept level.

- Multiple levels of aggregation in data cubes:
  - Further reduce the size of data to deal with



Year 1999

Year 1998

Year 1997

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
1997	$1,568,000
1998	$2,356,000
1999	$3,594,000

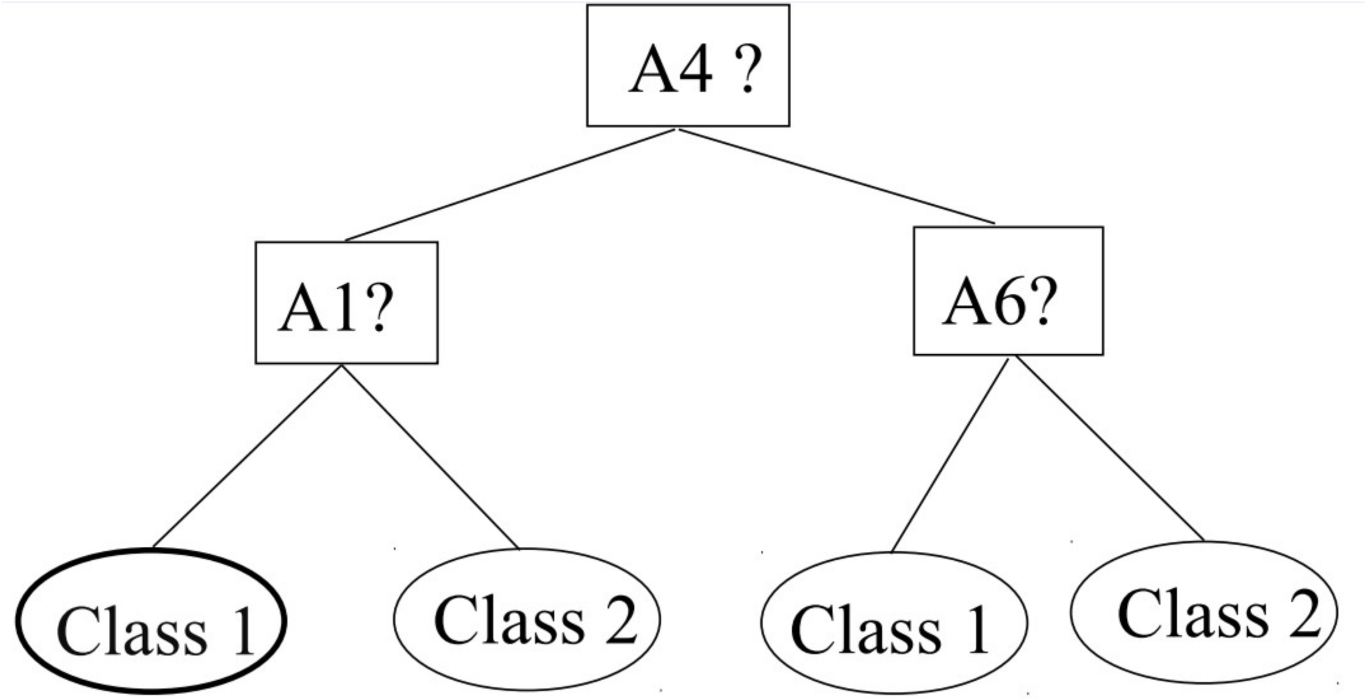
Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video, image compression
- Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time

Dimension reduction

For instance use Decision trees

- Initial attribute set: {A1, A2, A3, A4, A5, A6}



Reduced attribute set: {A1, A4, A6}

Numerosity Reduction

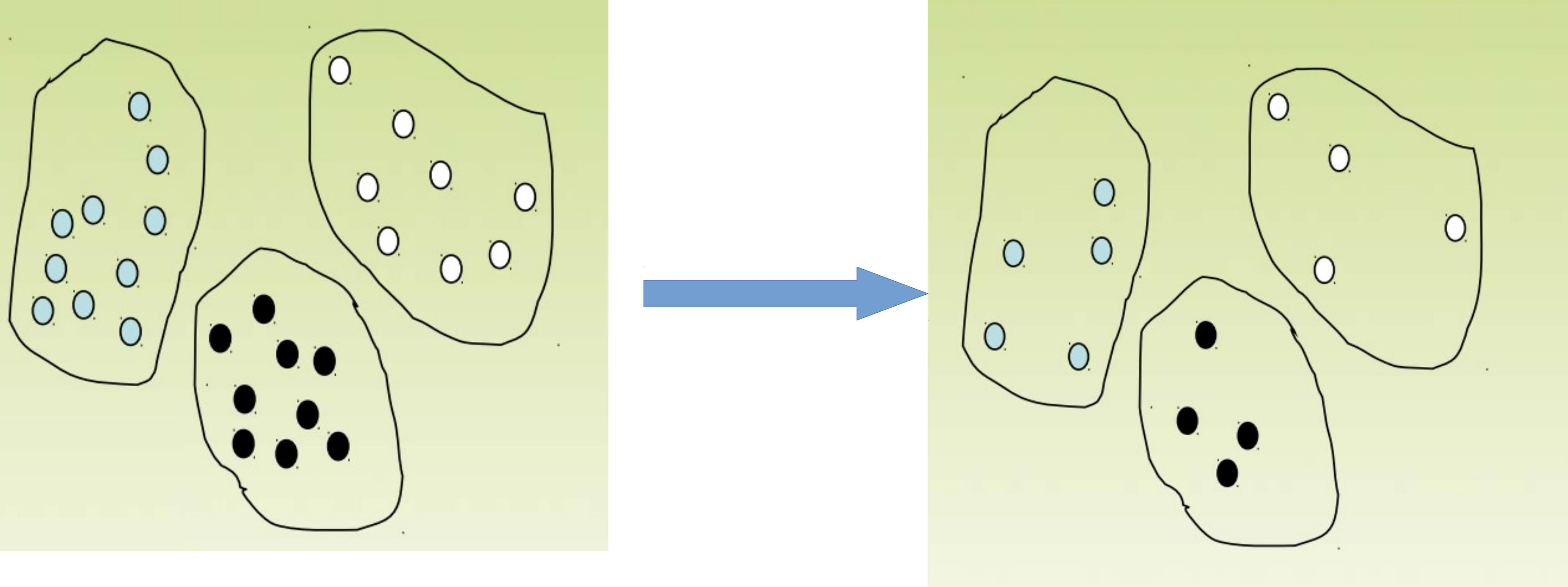
For instance use clustering:

  - Partition data set into clusters, and one can store cluster representation only

- Can be very effective if data is clustered.
- Can have hierarchical clustering and be stored in multidimensional index tree structures.

Numerosity Reduction

- Using clustering



Numerosity Reduction

Use proximity measure to select samples:

- Proximity is used to refer to Similarity or Dissimilarity.
    - Similarity: Numeric measure of the degree to which the two objects are alike.
    - Dissimilarity: Numeric measure of the degree to which the two objects are different.

Euclidean Distance to measure proximity

Euclidean Distance:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

- Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q.
- Standardization is necessary, if scales differ.