University of Science, VNU-HCM
Faculty of Information Technology

Introduction to Data Science Course

Data Preprocessing

Le Ngoc Thanh
lnthanh@fit.hcmus.edu.vn
Department of Computer Science

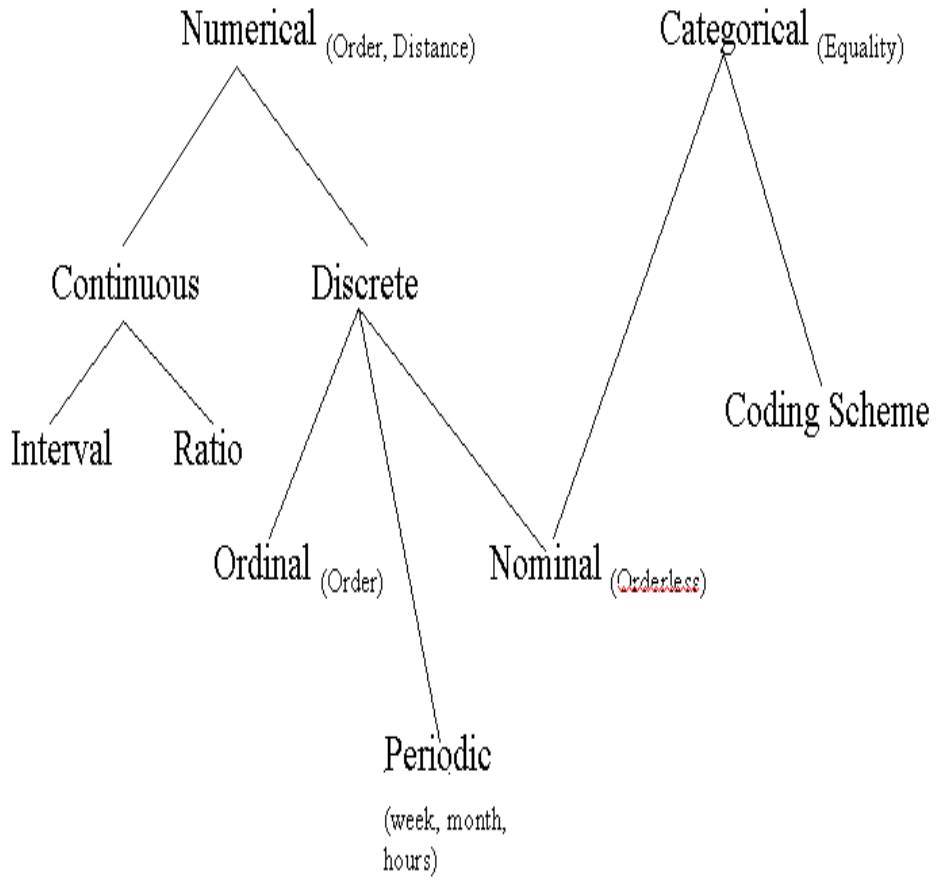
Ho Chi Minh City

Contents

- **Why need to preprocess data?**
- Data cleaning
- Data integration
- Data reduction
- Data transformation

Data

- Attribute (Key) - Value
- Data types
  - numeric, categorical
  - static, dynamic (time)
- Other data types
  - Distributed data
  - Text data
  - Web data, metadata
  - Pictures, audio / video
  - ....



Data quality

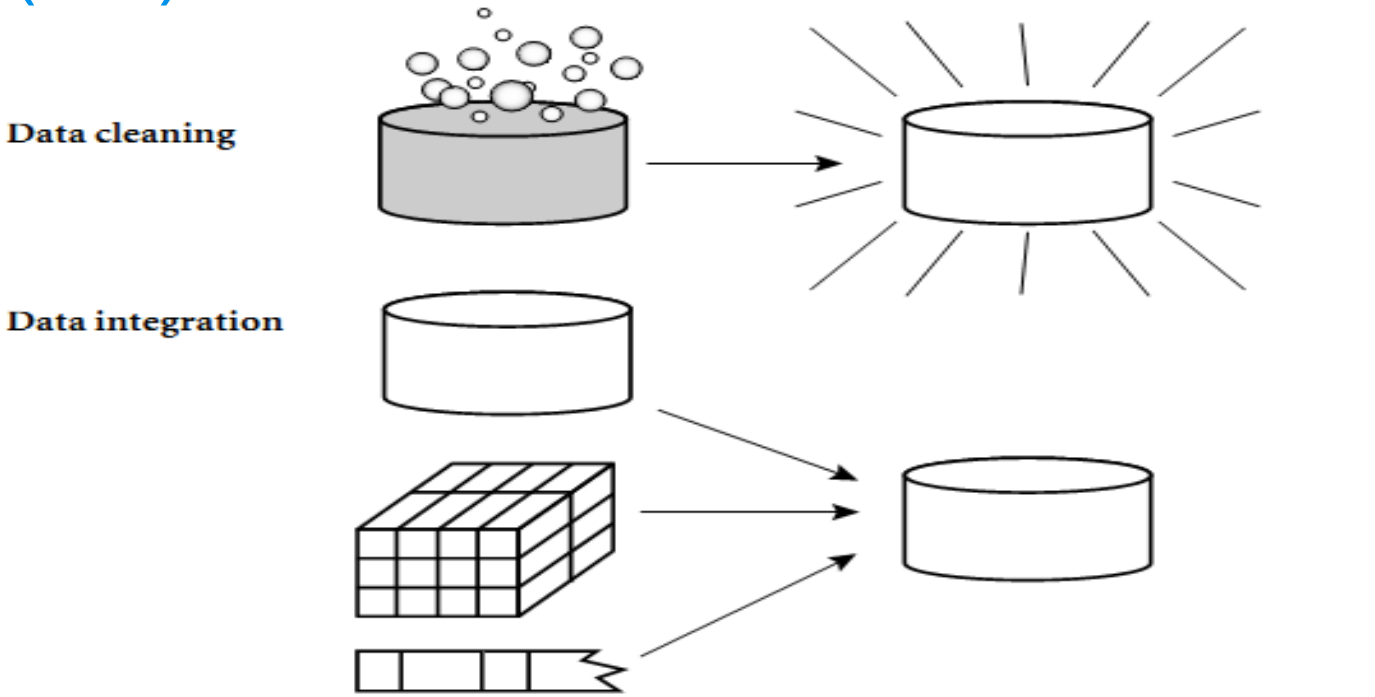
- Missing, incomplete: missing attribute value, missing attributes of interest, or only contains integrated data
  - Example : age, weight = “ ”
- Noise: contain errors or outliers
  - Example: salary =“-100 000”
- Conflict: there is inconsistency in the code or in the name
  - Example: age =42 , birth = 03/07/1997; US=USA?

Consequences of data quality

- The right decision must be based on accurate data
  - For example, duplication or lack of data can lead to inaccurate statistics, or even misleading.
- Data warehouse needs consistent integration of quality data

"Poor quality data -> not good exploitation"

Solutions? (1/2)



Solutions? (2/2)

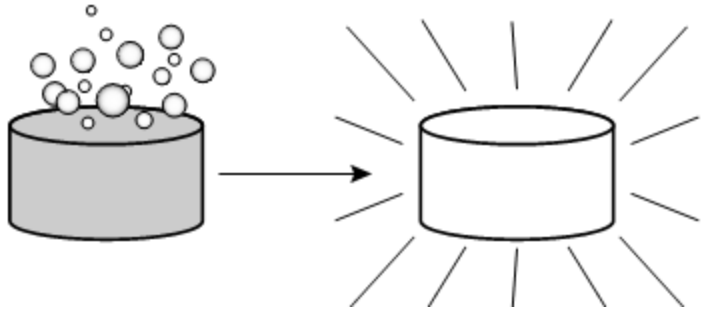
- Data Cleaning
  - Fill in missing values, eliminate noise data, identify and eliminate discrepancies, noise data, and resolve conflicting data
- Data Intergration
  - Synthesize, integrate DL from many databases, different files.
- Data Transformation
  - Aggregation.
- Data Reduction
  - Reduce the data size but ensure analytical results.

Contents

- Why need to prepare data?
- **Data cleaning**
- Data integration
- Data reduction
- Data transformation

Data cleaning

- Data cleaning is the most important task
- Data cleaning is the process:
  - Fill in the missing values
  - Identify and eliminate noise data
  - Resolve conflicting data



Fill the missing value(1/2)

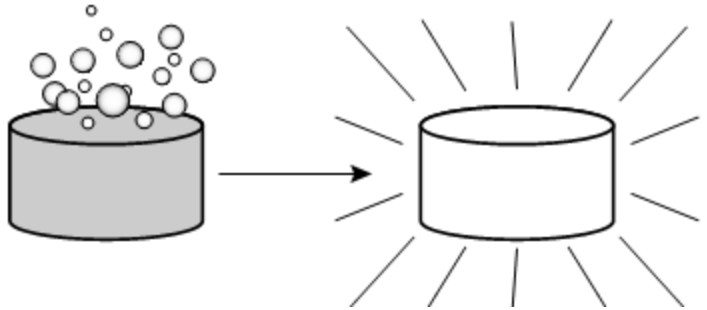
- Delete missing items:
  - Commonly used when class labels are missing (in classification)
  - Ease, but not efficiency, especially when the ratio of missing values is high.
- Fill in missing values manually: tasteless and not feasible
- Fill missing values automatically:
  - Replaced by a common constant. For example, "don't know". Can become new class in data

Fill the missing value (2/2)

- Fill missing values automatically:
  - Replaced with the property's mean
  - Replaced with the property's mean in a class
  - Replace with the most likely value: infer from a Bayesian formula, decision tree or EM algorithm (Expectation Maximization)

Data cleaning

- Data cleaning is the most important task
- Data cleaning is the process:
  - Fill in the missing values
  - Identify and eliminate noise data
  - Resolve conflicting data



Noise reduction

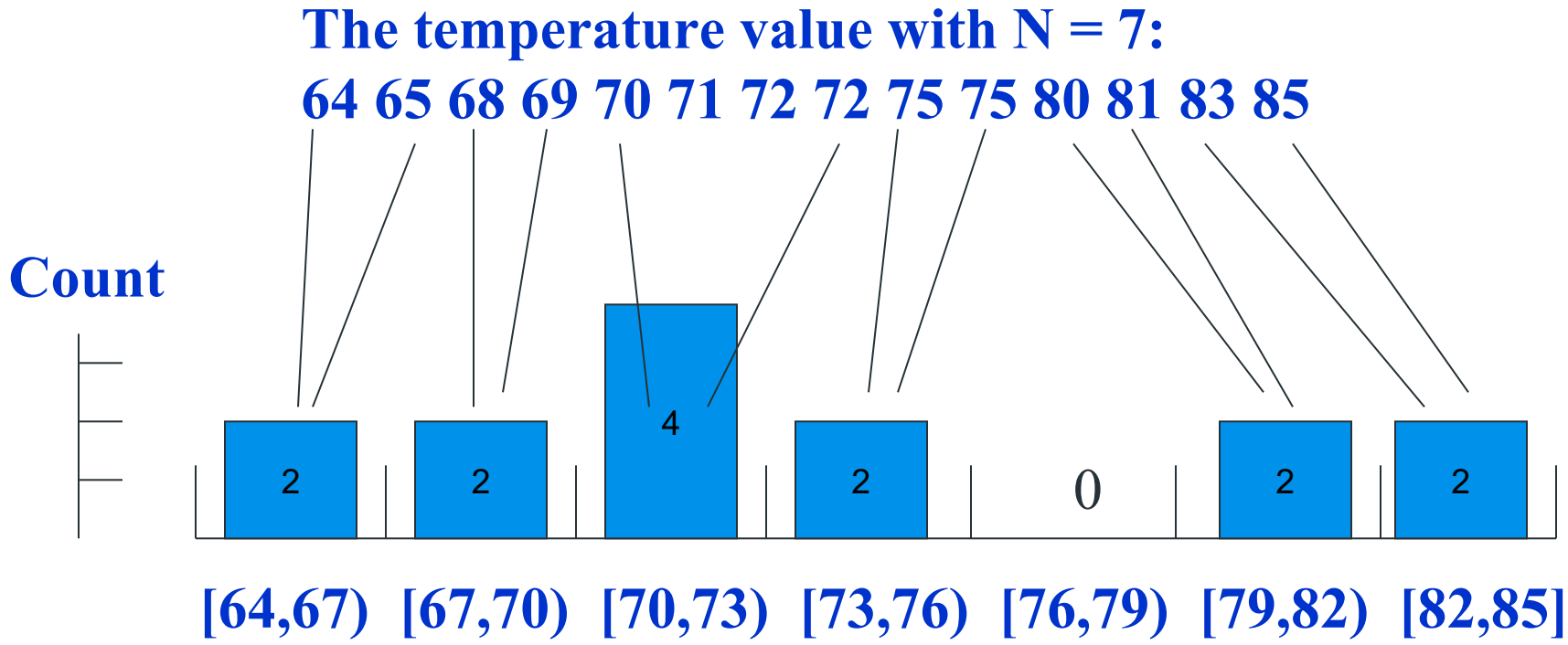
- The basic methods of noise reduction:
  - Binning method:
    - Sort and divide data into equal-width or equal-depth bins
    - Noise reduction by mean, median, margin, ...
  - Clustering method:
    - Detect and remove outliers
  - Regression method:
    - Fit data into the regression function

Noise reduction– Binning (1/4)

- Binning method
  - Divide data into equal-width bins:
    - Divide the range of values into N about the same size
    - The width of each interval = (maximum value - minimum value) / N
  - Divide data into equal-depth bins:
    - Divide the range of values into N ranges that each contain approximately the same number of samples

Noise reduction– Binning (2/4)

- Example about equal-width:

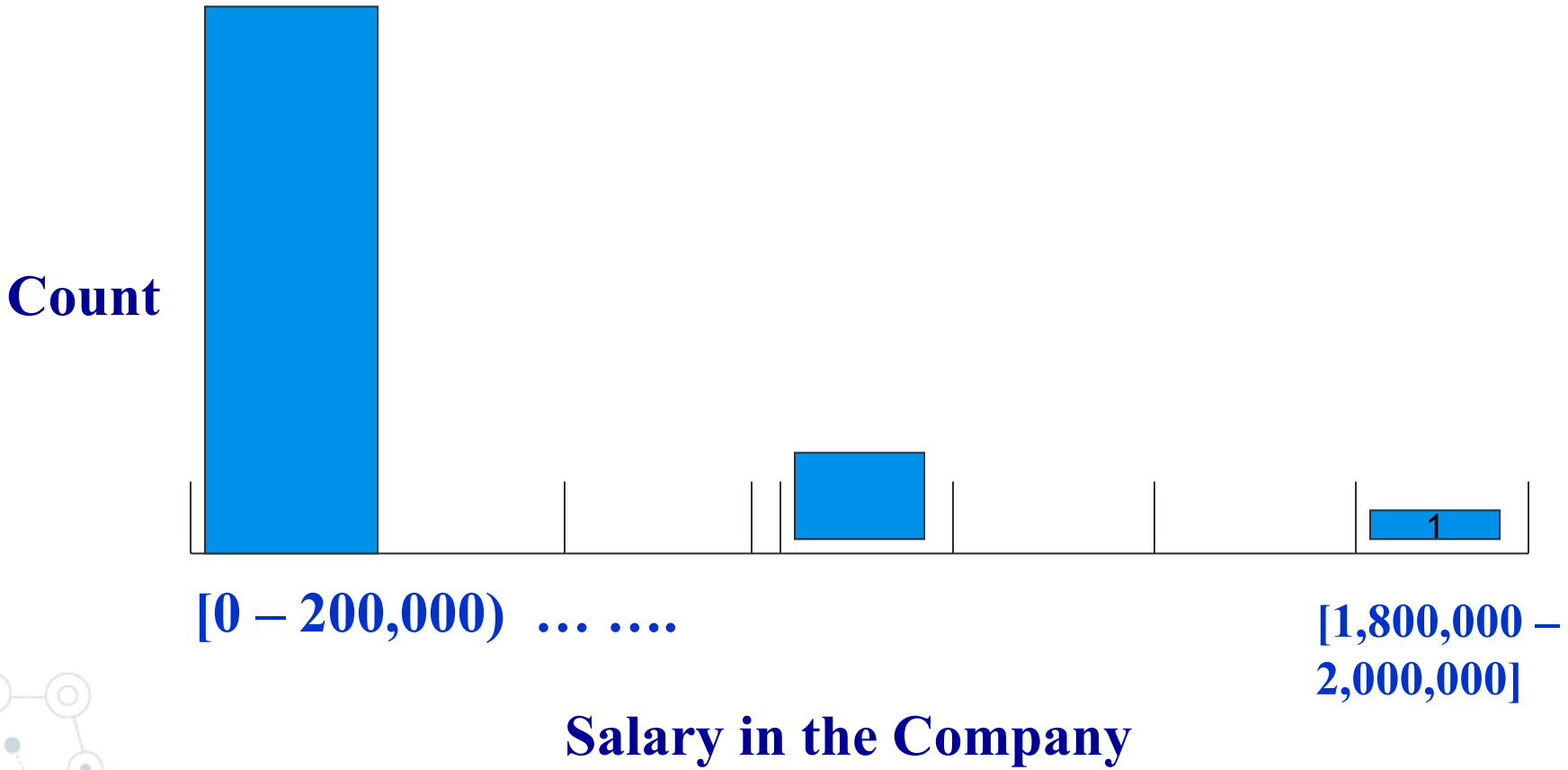


Left Bound <= value < Right Bound

Divide the range of values into N intervals.
The width of each interval = (maximum value - minimum value) / N.

Noise reduction– Binning (3/4)

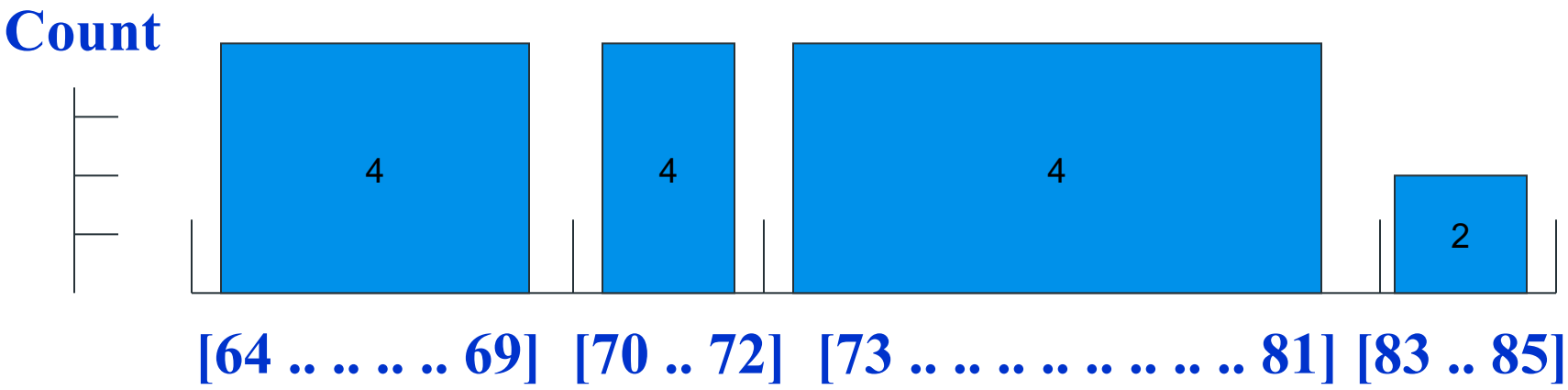
- But not good for skewed data



Noise reduction– Binning(4/4)

- Example about equal-depth:

The temperature value with N = 4:
64 65 68 69 70 71 72 72 75 75 80 81 83 85



Depth = 4, except for the last bin

Divide the range of values into N ranges that each contain approximately the same number of samples

Noise reduction with split bins

- Sorted prices:
  - 4, 8, 15, 21, 21, 24, 25, 28, 34
- Divide data into an equal-depth bins with N = 3
  - Bin 1: 4, 8, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 28, 34

→ What to do with the split bins?

Noise reduction with split bins

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

Smoothing by median:
- Bin 1: 8, 8, 8
- Bin 2: 21, 21, 21
- Bin 3: 28, 28, 28

Smoothing by mean:
- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

Smoothing by margin:
- Bin 1: 4, 4, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 25, 34

Exercises

- Prices :

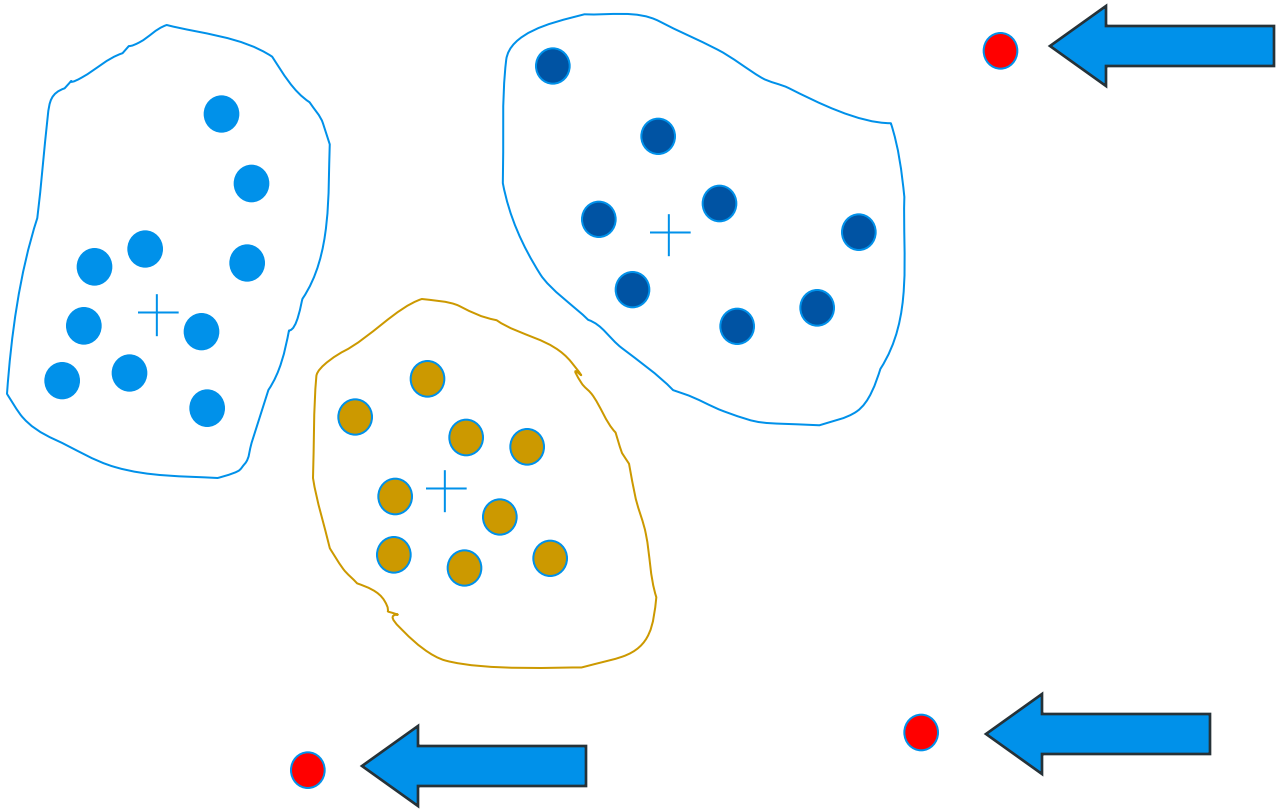
  15, 17, 19, 25, 29, 31, 33, 41, 42, 45, 45, 47, 52, 52, 64

- Use the binning method with equal-width and equal-depth with four bins:
  - Calculate the value of the bin according to the median smoothing.
  - Calculate the value of the bin according to the margin smoothing.
  - Calculate the value of the bin according to the mean smoothing.
  - Give some comments on results.

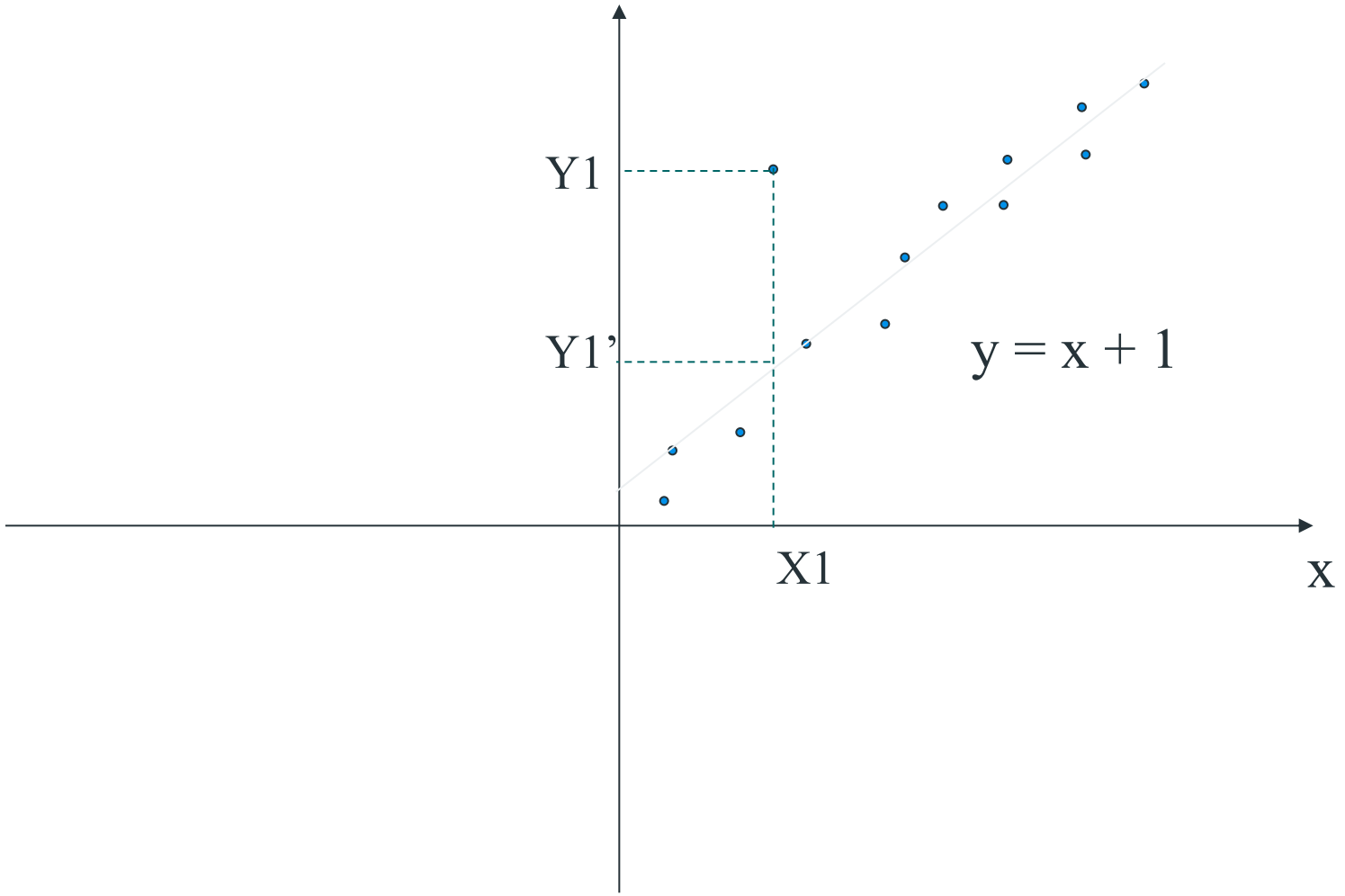
Noise reduction?

- The basic methods of noise reduction:
  - Binning method:
    - Sort and divide data into equal-width or equal-depth bins
    - Noise reduction by mean, median, margin, ...
  - Clustering method:
    - Detect and remove outliers
  - Regression method:
    - Fit data into the regression function

Noise reduction – clustering

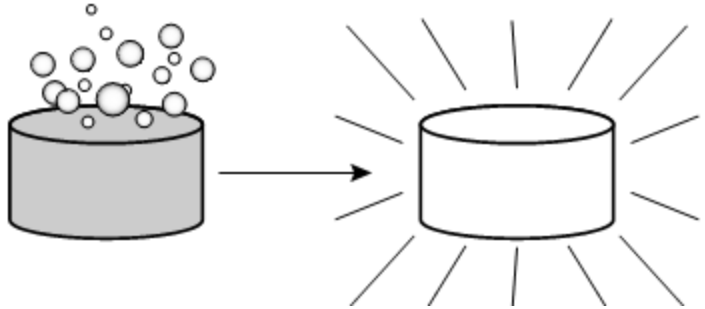


Noise reduction – regression



Data cleaning

- Data cleaning is the most important task
- Data cleaning is the process:
  - Fill in the missing values
  - Identify and eliminate noise data
  - Resolve conflicting data



Resolve conflicts

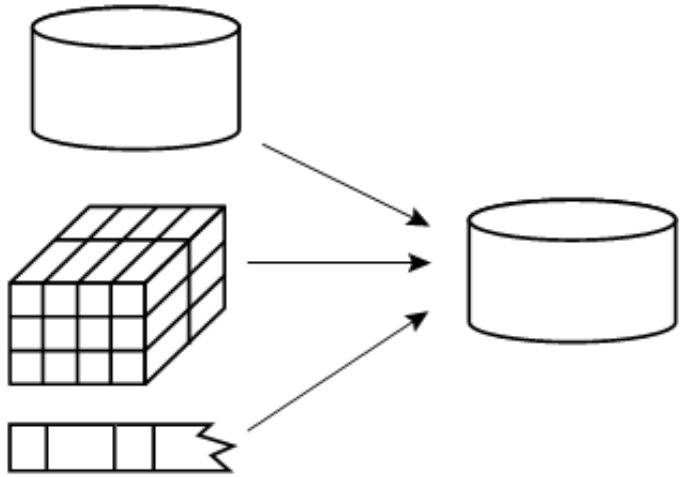
- How to handle conflicting data?
- Give examples of each conflict resolution method.

Contents

- Why need to prepare data?
- Data cleaning
- **Data integration**
- Data reduction
- Data transformation

Data Integration

- Select and aggregate data from many different sources into one database
- What problems occur when selecting and aggregating data?



Data integration process (1/4)

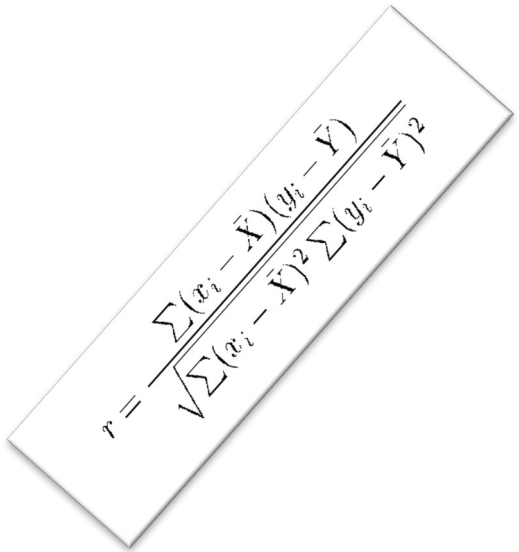
- Process:
  - Select only required data for the data mining process.
  - Matches the data schema
  - Eliminate redundant and duplicate data
  - Detect and resolve data inconsistencies

Data integration process (2/4)

- Schema Matching
  - Entity recognition problem
    - How do entities from multiple data sources become relevant
    - US=USA; customer_id = cust_number
  - Metadata

Data integration process (3/4)

- Eliminate redundant and duplicated data
  - An attribute is redundant if it can be inferred from other properties
  - The same property can have multiple names in different databases
  - Some records in the data are repeated
  - Use correlation analysis
    - r=0: X and Y are not correlated
    - r>0: positive correlation. X↑↔Y↑
    - r<0: negative correlation . X↓↔ Y ↑



Data integration process (4/4)

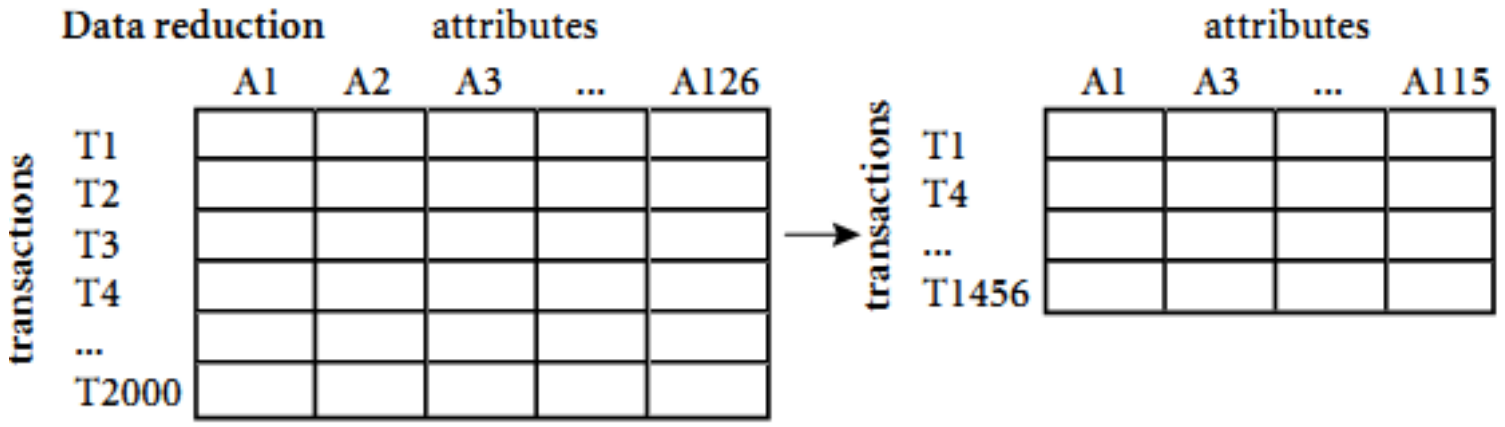
- Resolve inconsistencies in data
  - For example, weight is measured in kilograms or pounds
  - Define standards and mapping based on metadata

Contents

- Why need to prepare data?
- Data cleaning
- Data integration
- **Data reduction**
- Data transformation

Data reduction

- The data may be too large for some data mining applications: time consuming.
- Data reduction is the process of reducing data (size) so that the same (or almost the same) analysis result is obtained.



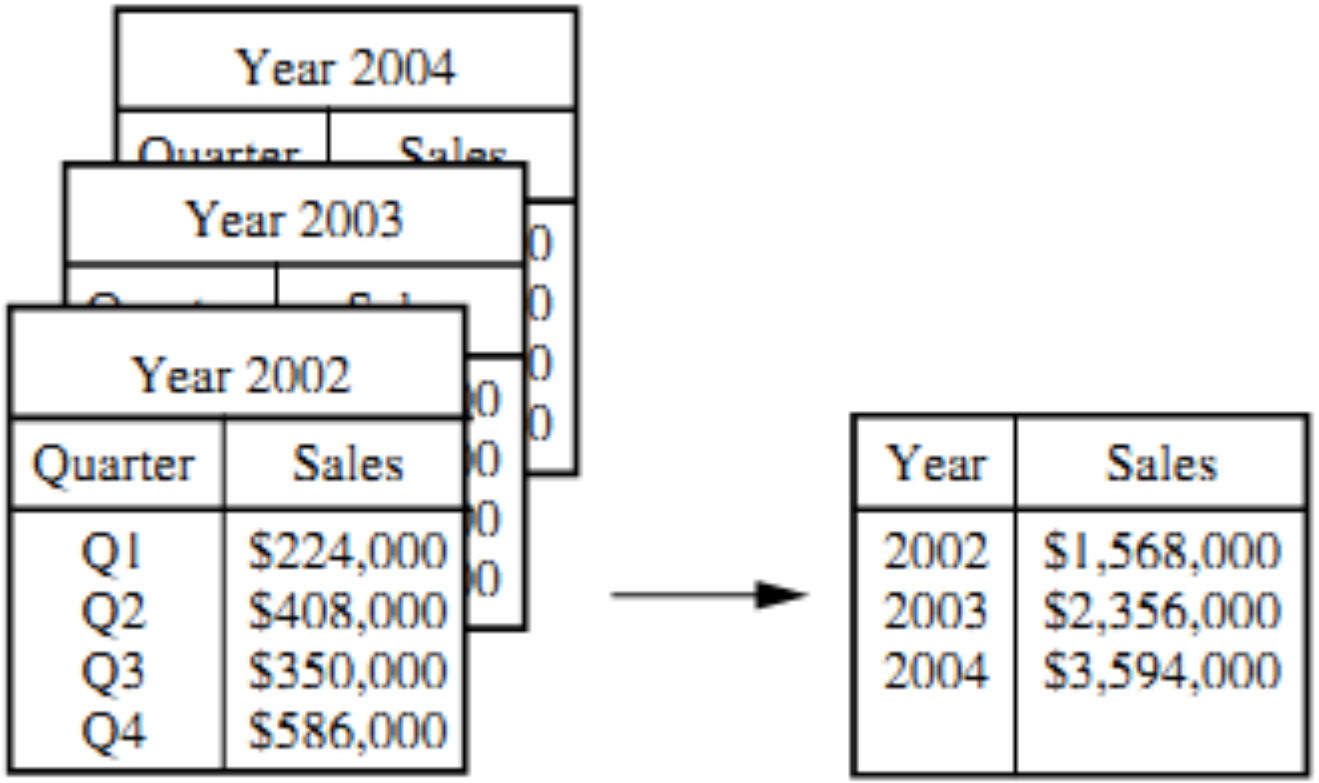
Methods of data reduction

- Methods:
  - Aggregation
  - Dimensionality reduction
  - Data compression
  - Numerosity reduction
  - Discretization and Concept hierarchies

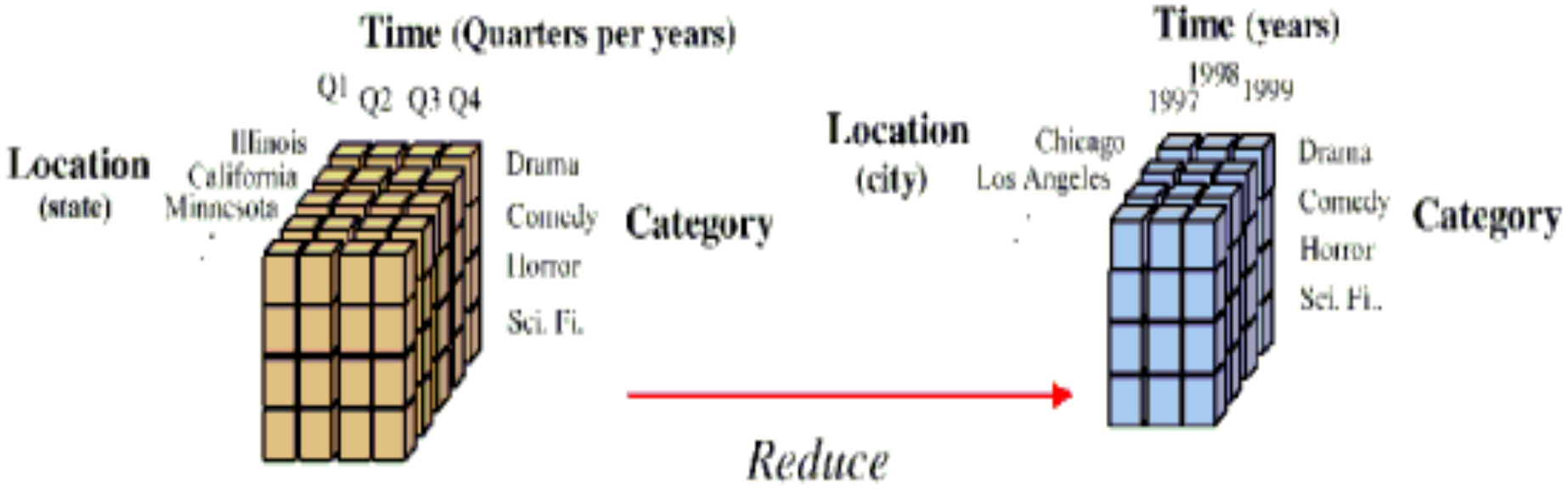
Data reduction – Aggregation (1/3)

- Aggregation
  - Combination of 2 or more attributes (object) into 1 attribute (object)
    - Example: cities integrated into regions, regions and water, …
  - Aggregate low-level data into high-level data:
    - Decrease data set size: reduce the number of attributes
    - Increase the interestingness of the sample

Data reduction – Aggregation (2/3)



Data reduction – Aggregation (3/3)



Data reduction – Dimensionality reduction (1/6)

- Dimensionality reduction
  - Feature selection (subset of attributes)
    - Choose m from n attributes
    - Remove irrelevant, redundant attributes
  - How to define irrelevant attributes?
    - Statistics
    - Information gain

Data reduction – Dimensionality reduction (2/6)

- How to reduce the data dimension?
  - Brute Force
    - There are 2d attribute subsets of d attributes
    - Computational complexity is too high
  - Heuristic method
    - Stepwise forward selection
    - Stepwise backward elimitation
    - Combine two methods
    - Inductive decision tree

Data reduction – Dimensionality reduction (3/6)

- Heuristic - Stepwise forward
  - Step 1: choose the best single attribute
  - Step 2: Choose the best attribute from the rest,...
- Example with initial attribute set:

  {A1,A2,A3,A4,A5,A6}
  - Result ={}
    - S1: Result = {A1}
    - S2: Result = {A1,A4}
    - S3: Result = {A1,A4,A6}

Data reduction – Dimensionality reduction (4/6)

- Heuristic - Stepwise backward
  - Step 1: removes the worst single attribute
  - Step 2: continues to remove the worst of the remaining attributes, …
- Example with initial attribute set:

  {A1,A2,A3,A4,A5,A6}
  - Result ={A1,A2,A3,A4,A5,A6}
    - S1: Result = {A1,A3,A4,A5,A6}
    - S2: Result = {A1,A4,A5,A6}
    - S3: Result = {A1,A4, A6}

Data reduction – Dimensionality reduction (5/6)

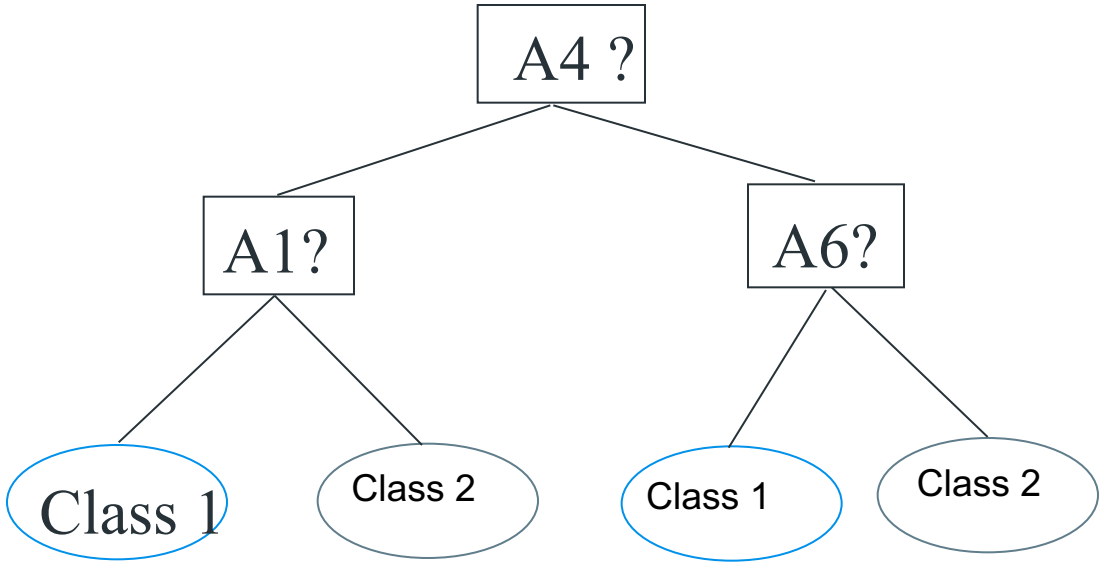
- Heuristic – Combine Forward and Backward
  - Step 1: select the best single attribute and the worst single attribute type
  - Continue to choose the best attribute and the worst attribute type among the rest, …
- Example with initial attribute set: {A1,A2,A3,A4,A5,A6}
  - Result = {A1,A2,A3,A4,A5,A6}
    - S1: Result = {A1,A3,A4,A5,A6}
    - S2: Result = {A1,A4,A5,A6}
    - S3: Result = {A1,A4, A6}

Data reduction – Dimensionality reduction (6/6)

- Heuristic – Inductive decision tree
  - Step 1: build decision tree
  - Step 2: removes any properties that are not present on the tree
- Example with initial attribute set:

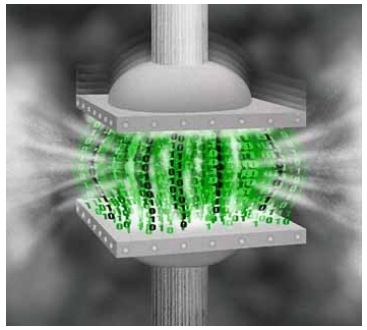
  {A1,A2,A3,A4,A5,A6}

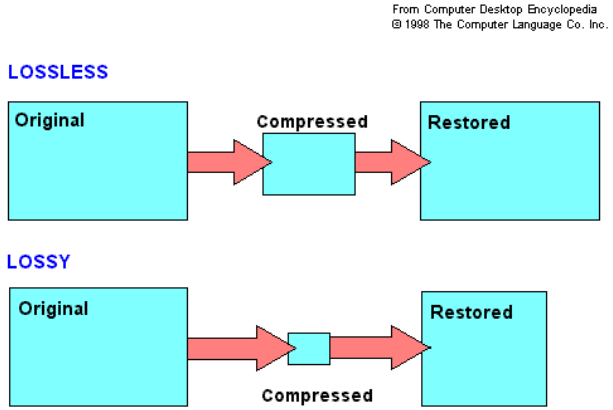
  ⇒ Result = {A1, A4, A6}



Data reduction – Compression

- Data Compression:
  - Encrypt or transform data
  - Lossless compression
    - Data can be recovered
  - Lossy compression
    - Data cannot be fully recovered
  - Using wavelet transforms, principal component analysis (PCA), ...





Data reduction – Numerosity reduction

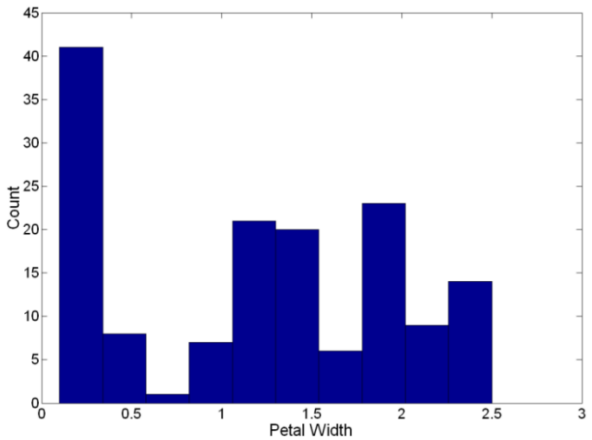
- Numerosity reduction: selects a different representation of the data ("less than")
- Some methods:
  - Parameter method:
    - Use a mathematical model to store parameters
    - Regression model and log-linear
  - Non-parametric method:
    - Do not use a mathematical model but save the reduced representation
    - Graphs, grouping, sampling

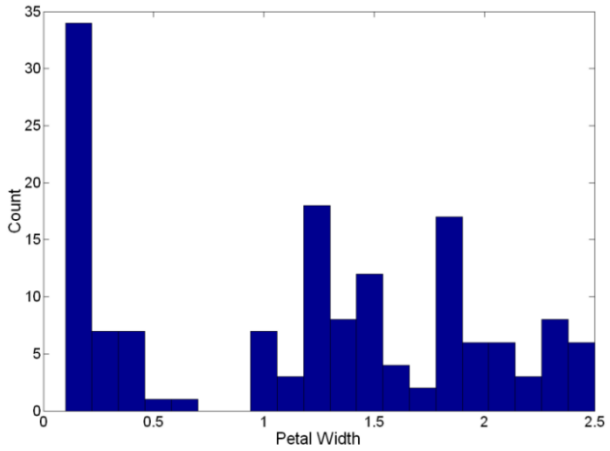
Data reduction – Numerosity reduction

- Linear regression: $Y = \alpha + \beta X$
- Multi linear regression: $Y = b0 + b1 X1 + b2 X2$
- Log-linear model:
  - Probaility: $p(a, b, c, d) = \alpha ab\ \beta ac\ \chi ad\ \delta bcd$

Data reduction – Numerosity reduction

- Histogram
  - Common methods for data reduction
  - Divide the data into bins and the height of the column is the number of objects in each bin. Store only the average of each bin.
  - The shape of the chart depends on the number of bins

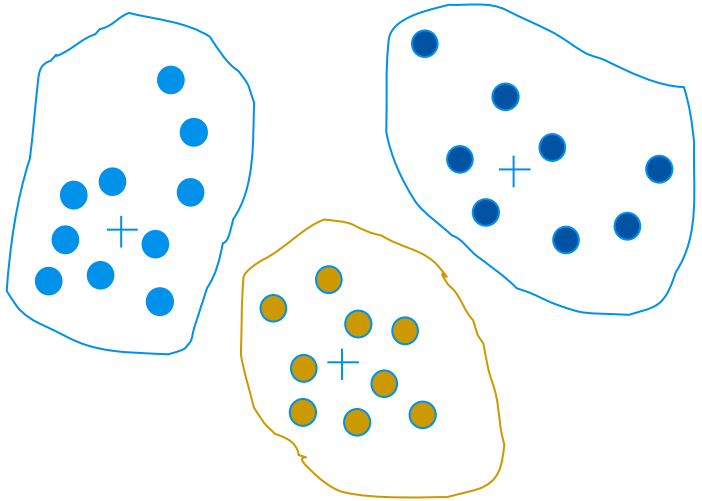




Data reduction – Numerosity reduction

Clustering

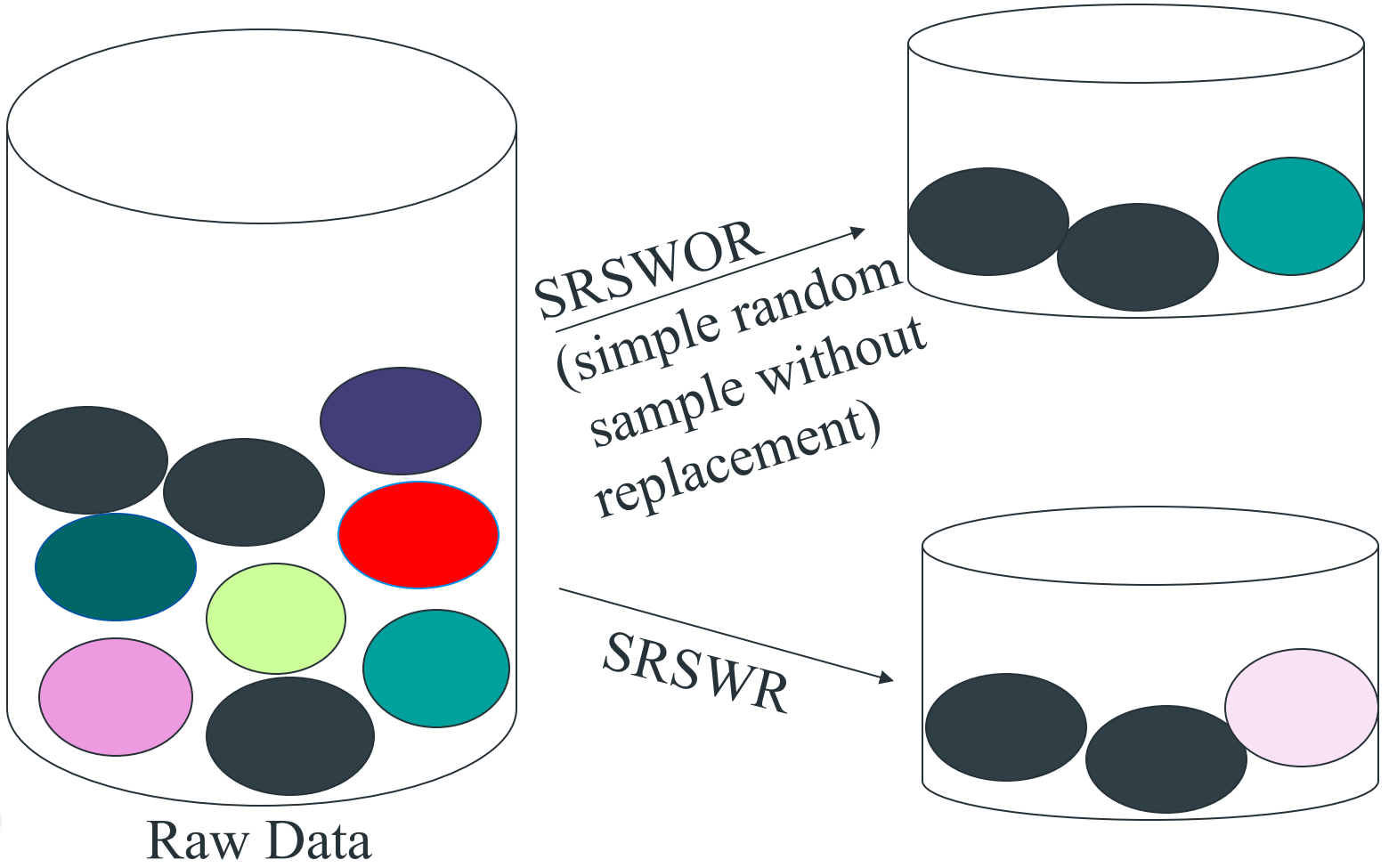
- Divide data into groups and save group representations.
- Very effective if the data is grouped but vice versa when the data is scattered
- Lots of clustering algorithms.



Data reduction – Numerosity reduction

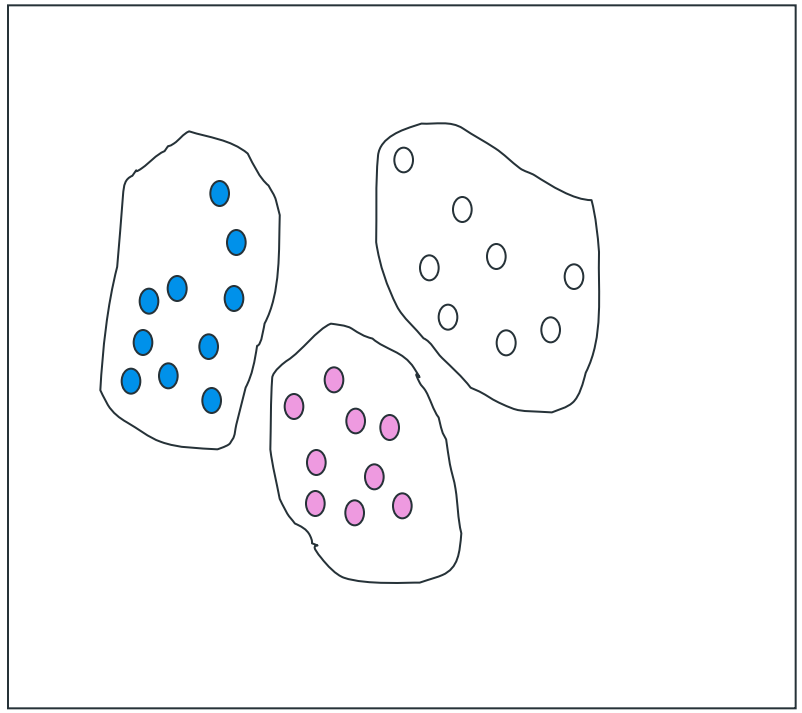
- Sampling
  - Use a much smaller random sample set instead of large data set.
  - Simple random sample without replacement (SRSWOR)
  - Simple random sample with replacement (SRSWR)
  - Group / hierarchical sampling method

Data reduction – Numerosity reduction

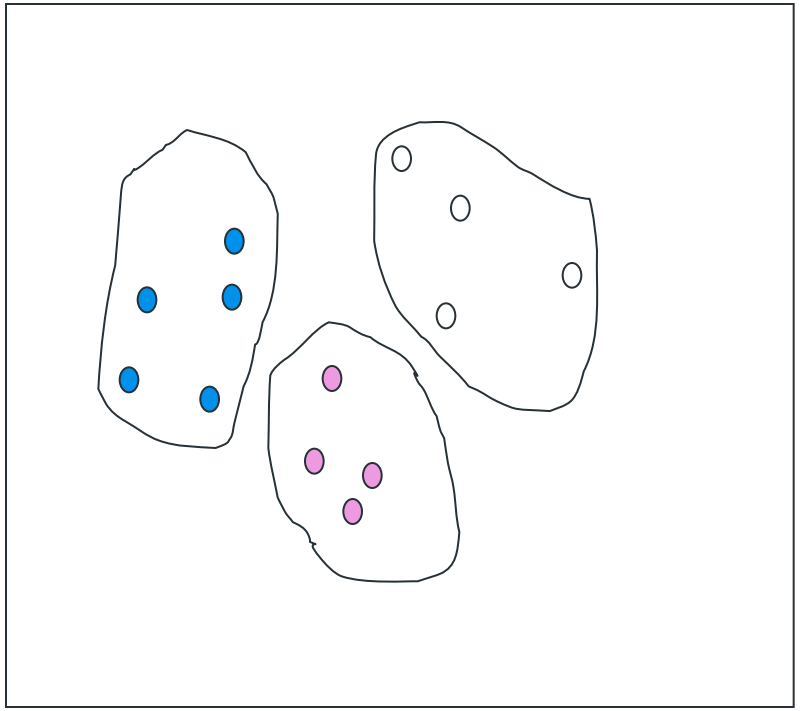


Data reduction – Numerosity reduction

Raw Data



Cluster/Stratified Sample



Data reduction – Discretization and Concept hierarchies

- **Discretization**:
  - Converts the property value domain (contiguous) by dividing the value domain into intervals.
  - Store labels of ranges instead of actual values
  - Suitable for continuous numeric data.
  - Methods: binning, chart analysis, grouping, discrete by entropy, natural segmentation.

Data reduction – Discretization and Concept hierarchies

- **Concept hierarchies**:
  - Gather and replace a low-level concept with a higher-level concept.
  - Suitable for non-numeric data: create a hierarchy.

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Attributes:
- Outlook (overcast, rain, sunny)
- Temperature real
- Humidity real
- Windy (true, false)
- Play (yes, no)

Standard Spreadsheet Format

OutLook	OutLook	OutLook	Temp	Humidity	Windy	Windy	Play	Play
overcast	rain	sunny			TRUE	FALSE	yes	no
0	0	1	85	85	0	1	1	0
0	0	1	80	90	1	0	0	1
1	0	0	83	78	0	1	1	0
0	1	0	70	96	0	1	1	0
0	1	0	68	80	0	1	1	0
0	1	0	65	70	1	0	0	1
1	0	0	64	65	1	0	1	0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	no
sunny	80	90	true	no
overcast	83	78	false	yes
rain	70	96	false	yes
rain	68	80	false	yes
rain	65	70	true	no
overcast	64	65	true	yes
sunny	72	95	false	no
sunny	69	70	false	yes
rain	75	80	false	yes
sunny	75	70	true	yes
overcast	72	90	true	yes
overcast	81	75	false	yes
rain	71	80	true	no

Attributes:

- Outlook (overcast, rain, sunny)
- Temperature real
- Humidity real
- Windy (true, false)
- Play (yes, no)

Standard Spreadsheet Format

OutLook	OutLook	OutLook	Temp	Humidity	Windy	Windy	Play	Play
overcast	rain	sunny			TRUE	FALSE	yes	no
0	0	1	85	85	0	1	1	0
0	0	1	80	90	1	0	0	1
1	0	0	83	78	0	1	1	0
0	1	0	70	96	0	1	1	0
0	1	0	68	80	0	1	1	0
0	1	0	65	70	1	0	0	1
1	0	0	64	65	1	0	1	0
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

Data reduction – Discretization and Concept hierarchies

- Example:
  - Converts the logical value to 1.0
  - Converts a date value to a number
  - Converts columns with large numeric values into a set of values in a smaller range, for example dividing them by a certain factor.
  - Group of values has the same semantics as: Activity before August Revolution is group 1; from 01/08/45 - 31/06/54; group 2; from 01/07/54 - 30/4/75 is group 3, ...
  - Substitute the value of age into young, middle-aged, old

Contents

- Why need to prepare data?
- Data cleaning
- Data integration
- Data reduction
- **Data transformation**

Data transformation

- Data transformation: convert data into a form that is suitable and convenient for algorithms
- Data transformation process :
  - Smoothing
  - Aggregation
  - Generalization
  - Normalization
  - Attribute construction

Data transformation process

- Smoothing: the process of removing noise from the data.
- Integration: summarizing or integrating data.
- Generalization: replacing low-level concepts with high-level concepts.
- Normalization: attribute data should be returned to a small range of values like 0 to 1.
- Attribute construction: new properties are created and added to a given set of properties

Conclusion

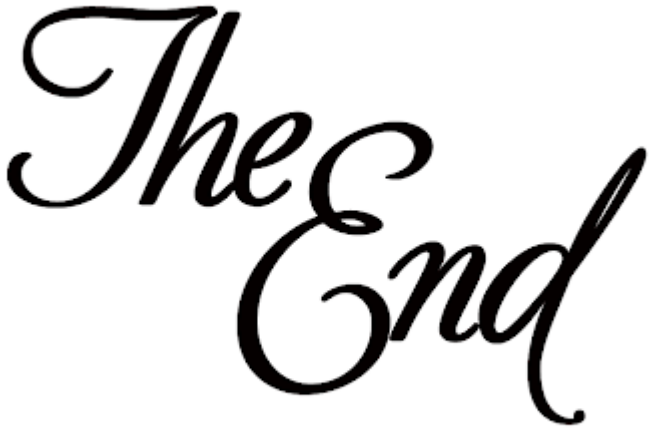
- Data is often missing, noisy, inconsistent, and multidimensional.
- Good data is the key to creating reliable and valid models.
- Data preparation includes the following processes:
  - Cleaning
  - Selection
  - Reduction
  - Transformation

Exercises

- Why is preparing data so urgent and time-consuming?
- How to solve the problem of missing values in database records?
- Assuming the database has Age attribute with the values in the records (ascending): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70
  - Denoise data by mean of bins with the number of bins n = 4. Explain the effectiveness of this technique with the above data.
  - Plot the equal-width histogram with the width = 10

Exercises

- Why do we need to select / integrate data? Please describe the data selection process.
- Why need to data reduction? Can data reduction process lose information? If yes, please state how to fix it.
- Learn about the data transformation processes. Give examples for each direction.


The End