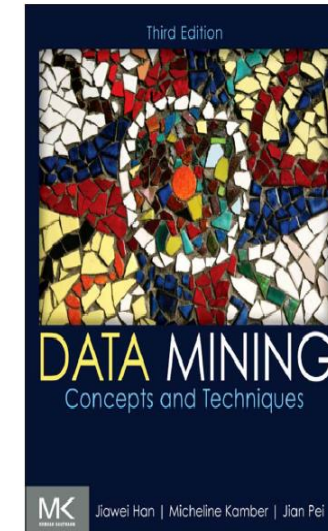


Data Mining & Business Intelligence (IT647)
Selected Topics In Data Science (IT688)
Information Technology Department
College Of Computer
Qassim University
DR.MONA ALSALEEM

Data Mining

— Chapter 2 —
Data Preprocessing

Text Book: “Data Mining: Concepts and Techniques”, Third Edition



Outlines

Getting to know your Data

- Data Objects
- Attribute Types

Data Preprocessing: An Overview

- Data Quality
- Major Tasks in Data Preprocessing
 - Data Cleaning
 - Data Integration
 - Data Reduction
 - Data Transformation

Summary

Getting to now your Data

Types of Data Sets

Relational Datasets

Person:

| Pers_ID | Surname | First_Name | City |
|---------|-----------|------------|----------|
| 0 | Miller | Paul | London |
| 1 | Ortega | Alvaro | Valencia |
| 2 | Huber | Urs | Zurich |
| 3 | Blanc | Gaston | Paris |
| 4 | Bertolini | Fabrizio | Rom |

no relation

Car:

| Car_ID | Model | Year | Value | Pers_ID |
|--------|-------------|------|--------|---------|
| 101 | Bentley | 1973 | 100000 | 0 |
| 102 | Rolls Royce | 1965 | 330000 | 0 |
| 103 | Peugeot | 1993 | 500 | 3 |
| 104 | Ferrari | 2005 | 150000 | 4 |
| 105 | Renault | 1998 | 2000 | 3 |
| 106 | Renault | 2001 | 7000 | 3 |
| 107 | Smart | 1999 | 2000 | 2 |

- Relational tables, highly structured
- Data stored in tables (rows and columns), like databases.
- Each row = one record (e.g., a person or a car).
- Each column = one attribute (e.g., name, city, year).

Used in:

- SQL databases
- Business systems
- Student records, hospital systems
- Easy to query, filter, and analyze.

Types of Data Sets

Transactional Datasets

| <i>TID</i> | <i>Items</i> |
|------------|---------------------------|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

- Each record shows a set of items bought or occurring together.
- Common in shopping or event logs.

Used in:

- Market basket analysis
- Recommendation systems
- Finding patterns like:
 - “Customers who buy diapers often buy milk.”
- Focus is on item combinations, not numerical values.

Types of Data Sets

Data matrix, e.g., numerical matrix

| Product | China | England | France | Japan | USA | Total |
|-------------------------------|-------|---------|--------|-------|----------|----------|
| Active Outdoors Crochet Glove | 12.00 | 4.00 | 6.00 | 1.00 | 240.00 | 257.00 |
| Active Outdoors Lycra Glove | 10.00 | 6.00 | 8.00 | 3.00 | 323.00 | 339.00 |
| Influx Crochet Glove | 3.00 | 6.00 | 8.00 | 2.00 | 132.00 | 149.00 |
| Influx Lycra Glove | 5.00 | 4.00 | 6.00 | 2.00 | 143.00 | 145.00 |
| Yamagh Pro Helmet | 5.00 | 5.00 | 7.00 | 3.00 | 335.00 | 352.00 |
| Yamagh Vertigo Helmet | 4.00 | 1.00 | 22.00 | 0.00 | 474.00 | 499.00 |
| Xtrem Adult Helmet | 8.00 | 6.00 | 7.00 | 2.00 | 251.00 | 274.00 |
| Xtrem Youth Helmet | 1.00 | 1.00 | 8.00 | 2.00 | 251.00 | 274.00 |
| Total | 24.00 | 43.00 | 54.00 | 14.00 | 1,972.00 | 2,085.00 |

- Purely numerical table.
- Rows = objects
- Columns = features/variables
- Cells = numbers

Used in:

- Machine learning
- Statistics
- Scientific data analysis
- Ideal for algorithms like regression, clustering, classification.

Types of Data Sets

Document data:

Term-frequency vector (matrix) of text documents

| | team | coach | pts | ball | score | game | wt | lost | timeout | season |
|------------|------|-------|-----|------|-------|------|----|------|---------|--------|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 0 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

- Text documents converted into numbers.
- Shows how often words appear in each document.

Used in:

- Text mining
- NLP (Natural Language Processing)
- Search engines and chatbots
- Turns text into a format computers can analyze.

Types of Data Sets

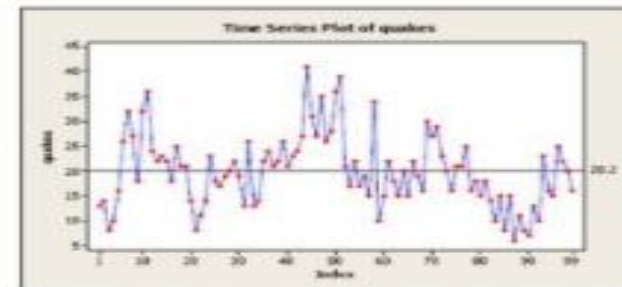
Social or information networks



Video data: sequence of images



Temporal data: time-series



Data Objects and Attribute Types

A *data object* represents an entity

Examples:

- sales database: customers, sales
- medical database: patients, treatments
- university database: students, professors, courses

Also called samples, instances, data points, objects, tuples

Data objects are described by *attributes*

| | | |
|---------------|---|--------------|
| database rows | → | data objects |
| columns | → | attributes |

Data Objects and Attribute Types

What Is an Attribute?

- Attribute (or dimensions, features, variables);
 - A data field, representing a characteristic or feature of a data object;
 - Attributes describing a customer object can include, for example, customer ID, name, and address.
- Observed values for a given attribute are known as observations;
- A set of attributes used to describe a given object is called an attribute vector.

Data Objects and Attribute Types

Attribute Types

- The type of an attribute is determined by the set of possible values the attribute can have.
- Types:
 - Nominal (e.g., red, blue)
 - Binary (e.g., {true, false})
 - Ordinal (e.g., {junior, senior})
 - Numeric: quantitative
 - Discrete and Continuous

Data Objects and Attribute Types

Nominal Attributes

- Nominal means “relating to names.”
- The values of a nominal attribute are symbols or names of things.
- Each value represents some kind of category, code, or state
- Nominal attributes are also referred to as categorical

Examples

- `hair_color = {black, blond, brown, red, ...}`
- The attribute marital status can take on the values single, married, divorced, and widowed.

Data Objects and Attribute Types

Binary Attributes

- A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present.
- Binary attributes are referred to as Boolean if the two states correspond to true and false.
- A binary attribute is *symmetric* if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1;
 - e.g., gender
- A binary attribute is *asymmetric* if the outcomes of the states are not equally important, such as the positive and negative outcomes of a medical test for HIV.
 - e.g., medical test (positive vs. negative)
 - Fraud ,spam detection

Data Objects and Attribute Types

Ordinal Attributes

- An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them,

Examples

- *drink size* corresponds to the size of drinks available at a fast-food restaurant. It has three possible values: small, medium, and large,
- *grade (e.g., A+, A, B+, B, C+, C and so on),*
- *Professional ranks* can be enumerated in a sequential order: for example, assistant, associate, and full for professors,

Data Objects and Attribute Types

Numeric Attributes

- A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.
 - Interval-Scaled Attributes
 - Measured on a scale of equal-sized units,
 - Values have order and can be positive, 0, or negative,
 - E.g., Calendar dates , temperature
 - No true zero-point , in temperature example neither 0 ° C nor 0 ° F indicates “no temperature.”
 - We can compare between values the years 2002 and 2010 are eight years apart.
 - we can compute their mean value, in addition to the median and mode measures of central tendency.

Data Objects and Attribute Types

Numeric Attributes

- Ratio-Scaled Attributes

- is a numeric attribute with an inherent **zero-point**
- the values are ordered, and we can also compute the difference between values, as well as the mean, median,
- E.g., length (length can be measured in meters or in cm)
- examples include attributes to measure weight, height, latitude and longitude
monetary quantities (e.g., you are 100 times richer with \$100 than with \$1).

Data Objects and Attribute Types

Discrete versus Continuous Attributes

■ Discrete Attribute

- Has only a finite set of values
 - E.g., profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

■ Continuous Attribute

- Has real numbers as attributes values
 - E.g., height, weight, temperature
- Real values can be measured and represented using a finite number of digits
- Continuous values are represented as floating-point variables

Data Preprocessing: An Overview

Data Preprocessing: An overview

Data Quality: Why processing the Data?

- Data have quality if they satisfy the requirements of the intended use.
- Six data quality factors:
 - Accuracy,
 - Completeness,
 - Consistency,
 - Timeliness,
 - Validity,
 - Uniqueness.

Data Preprocessing: An overview

Data Quality: Why processing the Data?

| Factor | How it's measured |
|--------------|---|
| Accuracy | How well does a piece of information reflect reality? |
| Completeness | Does it fulfill your expectations of what's comprehensive? |
| Consistency | Does information stored in one place match relevant data stored elsewhere? |
| Timeliness | Is your information available when you need it? |
| Validity | Is information in a specific format, does it follow business rules, or is it in an unusable format? |
| Uniqueness | Is this the only instance in which this information appears in the database? |

Data Preprocessing: An overview

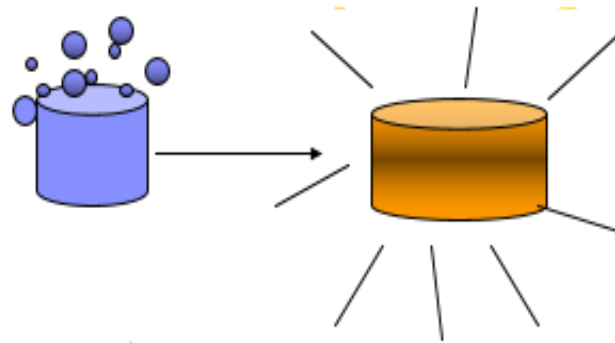
Major Tasks in Data Preprocessing

- **Data cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
 - Integration of multiple databases, data cubes, or files
- **Data reduction**
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- **Data transformation**
 - Normalization
 - Concept hierarchy generation

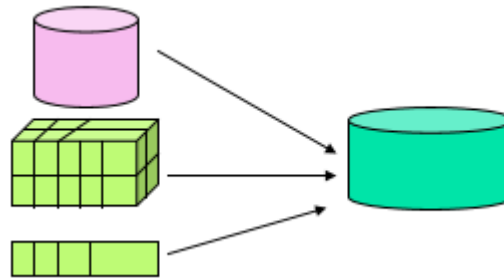
Data Preprocessing: An overview

Forms of Data Preprocessing

Data Cleaning



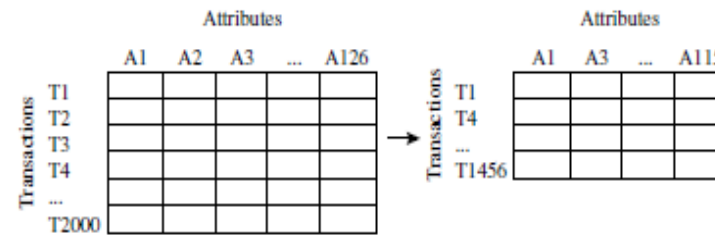
Data integration




Data Preprocessing: An overview

Forms of Data Preprocessing

Data Reduction

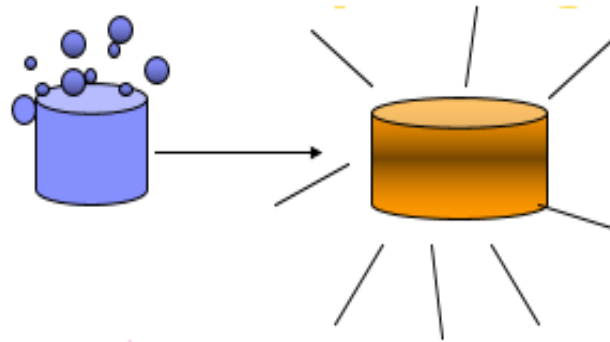


Data transformation

-2, 32, 100, 59, 48  -0.02, 0.32, 1.00, 0.59, 0.48

Task 1: Data Cleaning

Data Cleaning



“If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output.”

Task 1: Data Cleaning

- Data in the Real World is dirty because many things:
 - Incomplete (missing data): lacking attribute values, lacking certain attributes of interest
 - e.g., *Occupation*=“ ” (missing data)
 - Noisy: containing noise, errors, or outliers
 - e.g., *Salary*=“-10” (an error)
 - Inconsistent: containing discrepancies in codes or names, e.g.,
 - Age=“42”, Birthday=“03/07/2010”
 - Was rating “1, 2, 3”, now rating “A, B, C”
 - Conflict between duplicate records

Task 1: Data Cleaning

- **How to clean data?**
- By applying Data cleaning tasks, which are:
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data
 - Resolve redundancy caused by data integration

Task 1: Data Cleaning

Missing Values

Missing data may be due to:

- Information is not collected (e.g., people decline to give their age and weight and salary)
 - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
 - equipment malfunction
 - inconsistent/ **contradiction** with other recorded data and thus it will be deleted
 - data not entered due to misunderstanding
 - certain data may not be considered important at the time of entry
-
- **How to deal with Missing data?** Missing data may need to be inferred or concluded.

Task 1: Data Cleaning

- Method of Handling and filling the **missing values**
 - Ignore the tuple
 - Fill in the missing value manually
 - Use a global constant
 - Use the attribute mean for all samples belonging to the same class
 - Eliminate Data Objects.
 - Estimate Missing Values
 - Ignore the Missing Value During Analysis
 - Replace with all possible values (weighted by their probabilities)

Task 1: Data Cleaning

Noisy Data

- Noise is a random error or variance in a measured variable
- **Reasons for noisy data:**
 - The data collection instruments used may be faulty
 - Human or computer errors occurring at data entry
 - Errors in data transmission can also occur
 - There may be technology limitations
 - Duplicate records also need data cleaning to prepare it for a data warehouse

Task 1: Data Cleaning

Noisy Data

- Data smoothing techniques:
 - Binning (smoothing by bin means, smoothing by bin medians, smoothing by bin boundaries)
 - Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34
 - Regression :
 - a technique that conforms data values to a function. Linear regression, Multiple linear regression.

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

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

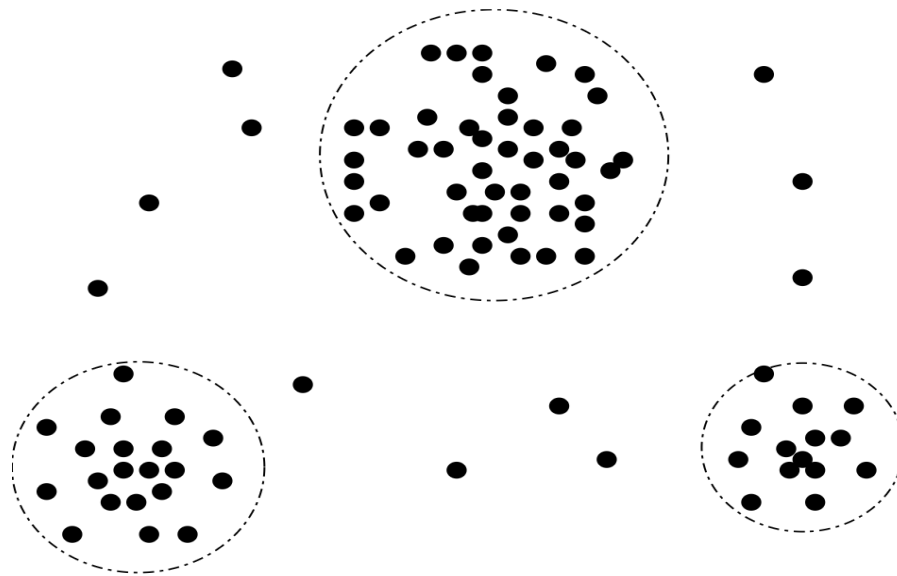
Task 1: Data Cleaning

Noisy Data

- Data smoothing techniques:

- Outlier Analysis:

- example clustering



A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

Task 2: Data Integration

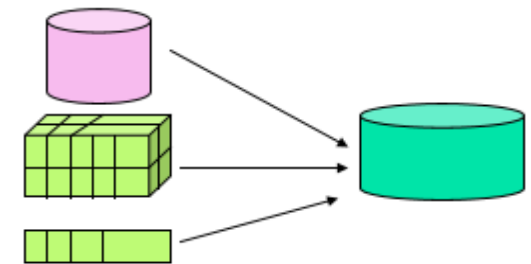
- Data mining often requires data integration—the merging of data from multiple data stores.
- Combines data from multiple sources into a coherent store DW

- **Entity identification problem**

- Identify real world entities from multiple data sources, e.g.,
how can the data analyst or the computer be sure that *customer_id* in one database and *cust_number* in another refer to the same attribute?

- Detecting and resolving data value conflicts.
 - Possible reasons: different representations, different scales, e.g., metric vs. British units. kilogram & ton

Data integration



Task 2: Data Integration

- **Redundancy and Correlation Analysis**
 - Redundant data occur often when integration of multiple databases, Object identification: The same attribute or object may have different names in different databases.
 - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue and student average...
 - Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*.
 - Careful integration of the data from multiple sources may help reduce/avoid redundancies, inconsistencies and improve mining speed and quality.

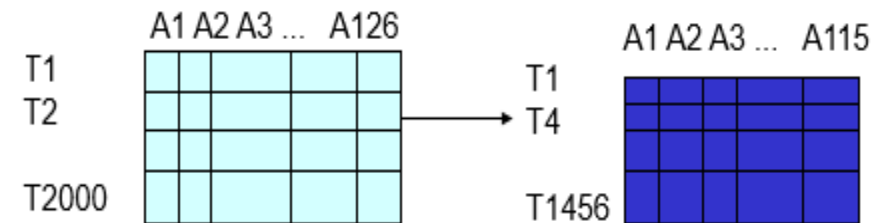
Task 3: Data Reduction

- Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.
- Why data reduction? — A database/data warehouse may store terabytes of data.
- Complex data analysis may take a very long time to run on the complete data set.
- **Data reduction strategies**

(1) **Dimensionality reduction:** Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration. ,

- e.g., remove unimportant attributes, dimensionality reduction.

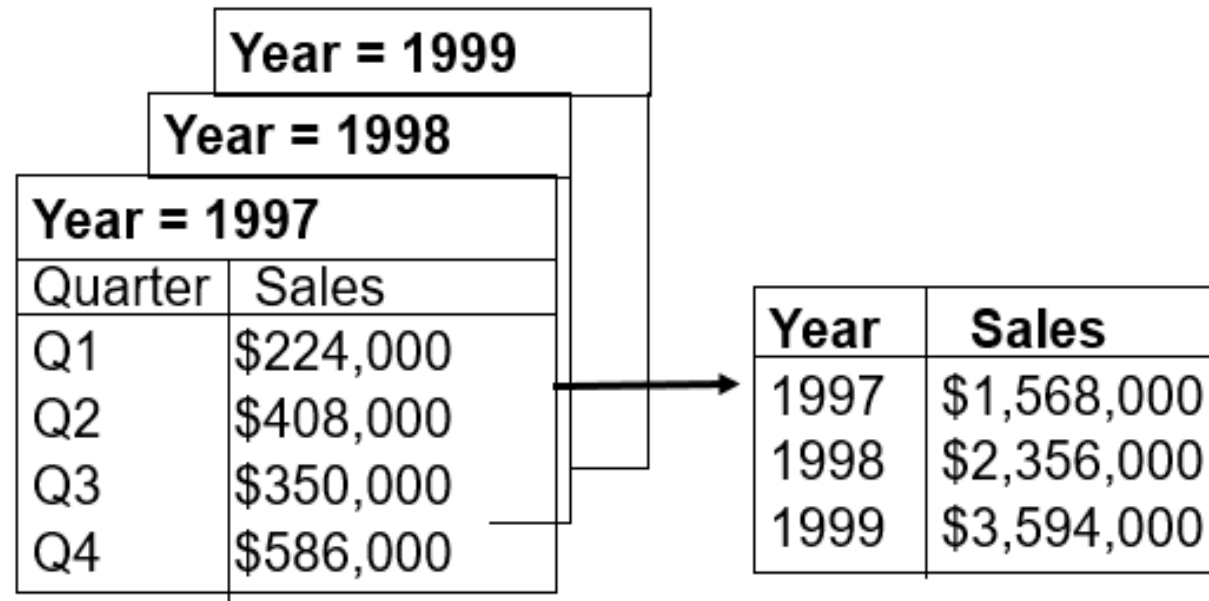
(pca , wavelet transform)



Data reduction

Task 3: Data Reduction

Data Cube Aggregation



Sales data for company *All Electronics* for 1997 - 1999

Task 3: Data Reduction

(2) **Numerosity reduction** : replace the original data volume by alternative, smaller forms of data representation. These techniques may be parametric or nonparametric.

- For **parametric methods**, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. Regression and log-linear models are examples.
- **Nonparametric methods** for storing reduced representations of the data include histograms, clustering, sampling and data cube aggregation.

Task 3: Data Reduction

(3) Data compression:

In data compression, transformations are applied so as to obtain a reduced or “compressed” representation of the original data.

- If the original data can be *reconstructed* from the compressed data without any information loss, the data reduction is called lossless.
- If, instead, we can *reconstruct only an approximation* of the original data, then the data reduction is called lossy.

Task 4: Data Transformation

- In data transformation, the data are transformed or consolidated into appropriate forms for mining.
- Strategies for data transformation:
 - **Smoothing**, which works to remove noise from the data. Techniques include binning, regression, and clustering.
 - **Attribute construction** (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.
 - **Aggregation** where summary or aggregation operations are applied to the data. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.
 - **Normalization** where the attribute data are scaled so as to fall within a smaller range, such as 0.0 to 1.0.

Summary

- Data Objects and Attribute Types are defined
- An overview of Data Preprocessing is presented
 - Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.
- Data integration combines data from multiple sources to form a coherent data store.
- Data reduction techniques obtain a reduced representation of the data while minimizing the loss of information content. These include methods of dimensionality reduction, numerosity reduction, and data compression.
- Data transformation routines convert the data into appropriate forms for mining.

Thank you

