

Data Mining & Knowledge Discovery

Introduction to Data & Attributes
Data Quality

Md. Biplob Hosen

Lecturer, IIT-JU

Email: biplob.hosen@juniv.edu

Data-related Issues Important for Data Mining

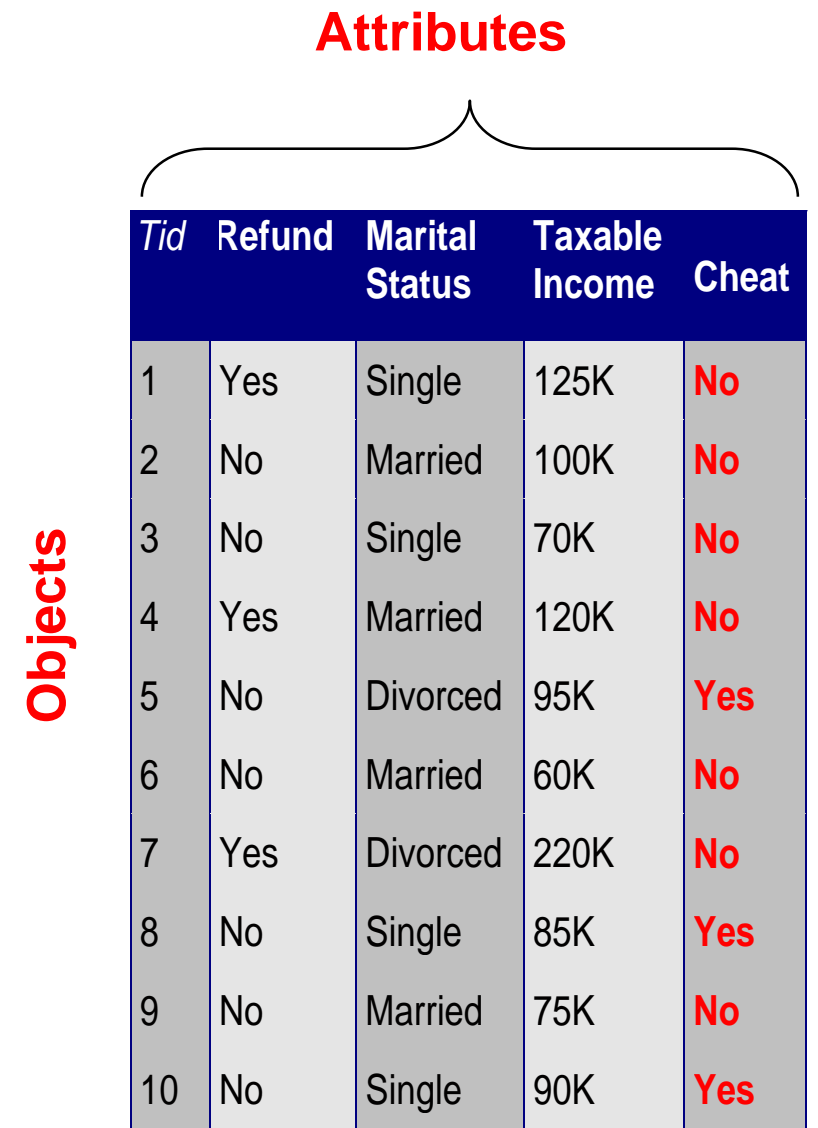
- **The Type of Data:** Datasets differ in a number of ways. For example, the attributes used to describe data objects can be of different types—quantitative or qualitative—and data sets often have special characteristics; e.g., some datasets contain time series or objects with explicit relationships to one another. Not surprisingly, the type of data determines which tools and techniques can be used to analyze the data. Indeed, new research in data mining is often driven by the need to accommodate new application areas and their new types of data.
- **The Quality of the Data:** Data is often far from perfect. While most data mining techniques can tolerate some level of imperfection in the data, a focus on understanding and improving data quality typically improves the quality of the resulting analysis. Data quality issues that often need to be addressed include the presence of noise and outliers; missing, inconsistent, or duplicate data; and data that is biased or, in some other way, unrepresentative of the phenomenon or population that the data is supposed to describe.

Data-related Issues Important for Data Mining

- **Preprocessing Steps:** To make the data more suitable for data mining. Often, the raw data must be processed in order to make it suitable for analysis. While one objective may be to improve data quality, other goals focus on modifying the data so that it better fits a specified data mining technique or tool. For example, a continuous attribute, e.g., length, sometimes needs to be transformed into an attribute with discrete categories, e.g., short, medium, or long, in order to apply a particular technique. As another example, the number of attributes in a data set is often reduced because many techniques are more effective when the data has a relatively small number of attributes.
- **Analyzing Data in Terms of Its Relationships:** One approach to data analysis is to find relationships among the data objects and then perform the remaining analysis using these relationships rather than the data objects themselves. For instance, we can compute the similarity or distance between pairs of objects and then perform the analysis—clustering, classification, or anomaly detection—based on these similarities or distances. There are many such similarity or distance measures, and the proper choice depends on the type of data and the particular application.

Data Mining: Data

- Collection of **data objects** and their **attributes**.
- An attribute is a property or characteristic of an object.
- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, dimension, or feature.
- A collection of attributes describe an object.
- Object is also known as record, point, case, sample, entity, or instance.



The diagram illustrates the relationship between data objects and their attributes. A bracket labeled "Attributes" in red spans the header row of the table. The word "Objects" is written vertically in red to the left of the table rows. The table itself has a dark blue header and light gray data rows.

Attributes				
<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Mining: Attributes

- An attribute is a property or characteristic of an object that can vary, either from one object to another or from one time to another.
- For example, eye color varies from person to person, while the temperature of an object varies over time.
- Note that eye color is a symbolic attribute with a small number of possible values {brown, black, blue, green, hazel, etc.} , while temperature is a numerical attribute with a potentially unlimited number of values.
- A **measurement scale** is a rule (function) that associates a numerical or symbolic value with an attribute of an object.
- For instance, we step on a bathroom scale to determine our weight, we classify someone as male or female, or we count the number of chairs in a room to see if there will be enough to seat all the people coming to a meeting.
- In all these cases, the “physical value” of an attribute of an object is mapped to a numerical or symbolic value.

Type of an Attribute

- There are different types of attributes
 - **Nominal:** Examples - ID numbers, eye color, zip codes.
 - **Ordinal:** Examples - rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}.
 - **Interval:** Examples - calendar dates, temperatures in Celsius or Fahrenheit.
 - **Ratio:** Examples - length, counts, elapsed time (e.g., time to run a race).

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
 - Distinctness: $= \neq$
 - Order: $< >$
 - Differences are meaningful: $+ -$
 - Ratios are meaningful: $* /$
- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & meaningful differences
- Ratio attribute: all 4 properties/operations

Categorization of Attributes

		Attribute Type	Description	Examples	Operations
Categorical	Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: { <i>male</i> , <i>female</i> }	mode, entropy, contingency correlation, χ^2 test
		Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, { <i>good</i> , <i>better</i> , <i>best</i> }, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric	Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
		Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Temperature Scales

- Temperature provides a good illustration of some of the concepts that have been described.
- First, temperature can be either an interval or a ratio attribute, depending on its measurement scale.
- When measured on the Kelvin scale, a temperature of 2 is, in a physically meaningful way, twice that of a temperature of 1 .
- This is not true when temperature is measured on either the Celsius or Fahrenheit scales, because, physically, a temperature of 1 Fahrenheit (Celsius) is not much different than a temperature of 2 Fahrenheit (Celsius).
- The problem is that the zero points of the Fahrenheit and Celsius scales are, in a physical sense, arbitrary, and therefore, the ratio of two Celsius or Fahrenheit temperatures is not physically meaningful.

Discrete and Continuous Attributes

Discrete Attribute:

- Has only a finite or countably infinite set of values.
- Examples: zip codes, counts, or the set of words in a collection of documents.
- Often represented as integer variables.
- Note: binary attributes are a special case of discrete attributes and assume only two values, e.g., true/false, yes/no, male/female, or 0/1.

Continuous Attribute:

- Has real numbers as attribute values.
- Examples: temperature, height, or weight.
- Practically, real values can only be measured and represented using a finite number of digits.
- Continuous attributes are typically represented as floating-point variables.

Asymmetric Attributes

- For asymmetric attributes, only presence—a non-zero attribute value—is regarded as important.
- Consider a data set in which each object is a student and each attribute records whether a student took a particular course at a university.
- For a specific student, an attribute has a value of 1 if the student took the course associated with that attribute and a value of 0 otherwise.
- Because students take only a small fraction of all available courses, most of the values in such a data set would be 0.
- Therefore, it is more meaningful and more efficient to focus on the non-zero values.
- To illustrate, if students are compared on the basis of the courses they don't take, then most students would seem very similar, at least if the number of courses is large.
- Binary attributes where only non-zero values are important are called asymmetric binary attributes.
- This type of attribute is particularly important for association analysis.

General Characteristics of Datasets

- **Dimensionality:** The dimensionality of a dataset is the number of attributes that the objects in the dataset possess. Analyzing data with a small number of dimensions tends to be qualitatively different from analyzing moderate or high-dimensional data. Indeed, the difficulties associated with the analysis of high-dimensional data are sometimes referred to as the curse of dimensionality. Because of this, an important motivation in preprocessing the data is dimensionality reduction.
- **Distribution:** The distribution of a data set is the frequency of occurrence of various values or sets of values for the attributes comprising data objects. Equivalently, the distribution of a data set can be considered as a description of the concentration of objects in various regions of the data space.

General Characteristics of Datasets

- **Resolution:** It is frequently possible to obtain data at different levels of resolution, and often the properties of the data are different at different resolutions.
- For instance, the surface of the Earth seems very uneven at a resolution of a few meters, but is relatively smooth at a resolution of tens of kilometers.
- The patterns in the data also depend on the level of resolution. If the resolution is too fine, a pattern may not be visible or may be buried in noise; if the resolution is too coarse, the pattern can disappear.
- For example, variations in atmospheric pressure on a scale of hours reflect the movement of storms and other weather systems. On a scale of months, such phenomena are not detectable.

Types of datasets

- Record
 - Data Matrix
 - Document Data
 - Transaction Data
- Graph
 - World Wide Web
 - Molecular Structures
- Ordered
 - Spatial Data
 - Temporal Data
 - Sequential Data
 - Genetic Sequence Data

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes.

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.
- Such a data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a 'term' vector.
- Each term is a component (attribute) of the vector.
- The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	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

Transaction Data

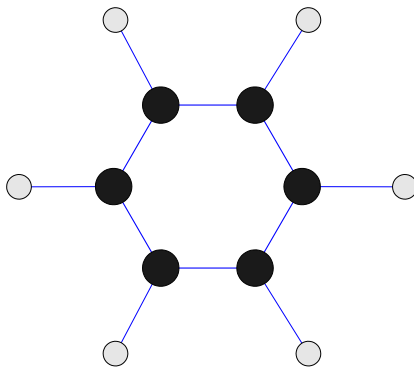
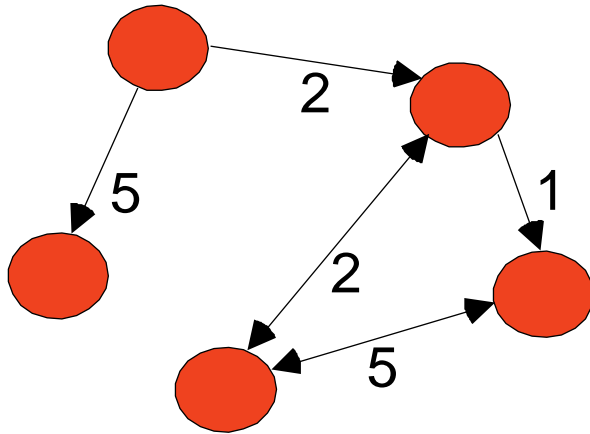
A special type of data, where -

- Each transaction involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.
- Can represent transaction data as record data.

<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

Graph Data

- Examples: Generic graph, a molecule, and webpages.



Useful Links:

- [Bibliography](#)
- Other Useful Web sites
 - [ACM SIGKDD](#)
 - [KDnuggets](#)
 - [The Data Mine](#)

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- [Books](#)
- [General Data Mining](#)

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support)", John Wiley & Sons, 1997.

General Data Mining

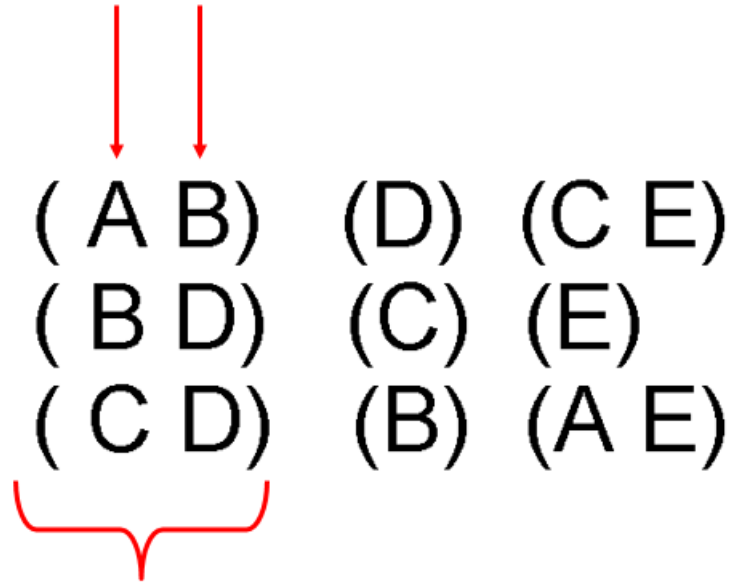
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequences of transactions.

Items/Events



An element of
the sequence

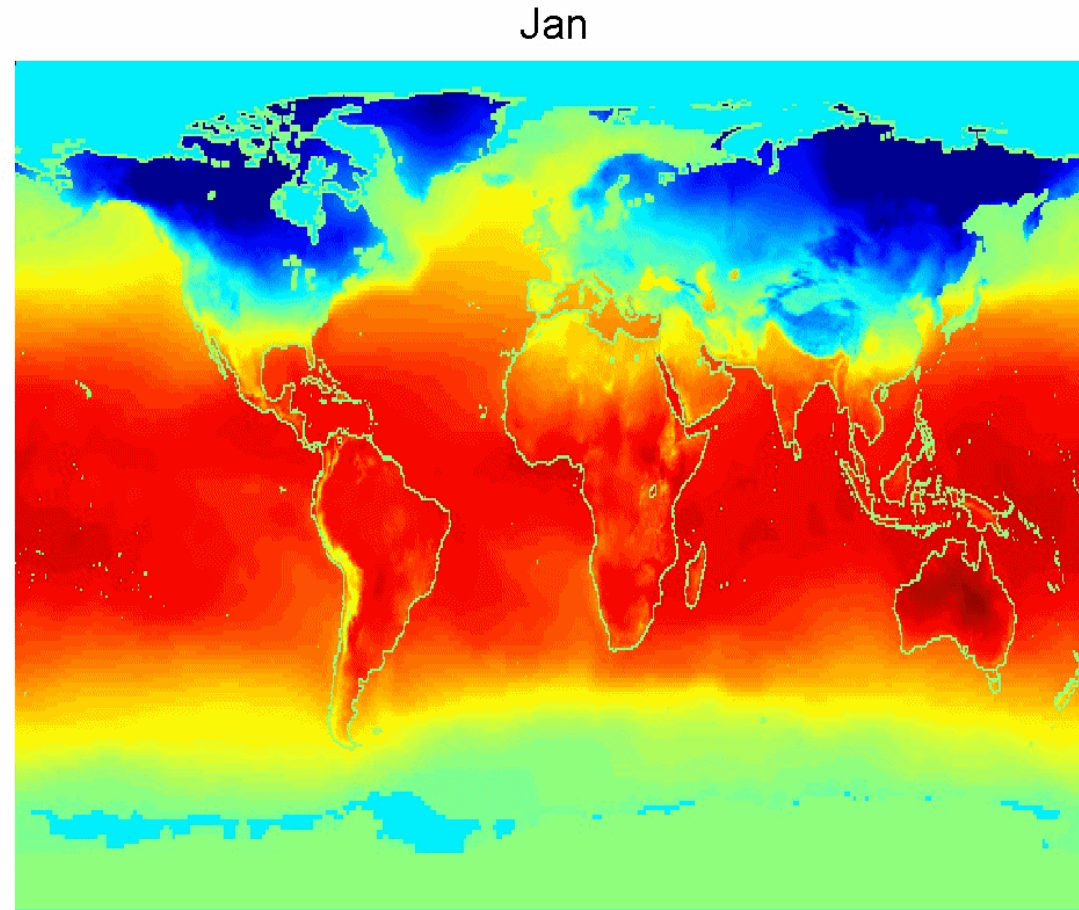
- Genomic sequence data

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Ordered Data

- Spatio-Temporal Data

**Average Monthly
Temperature of
land and ocean**

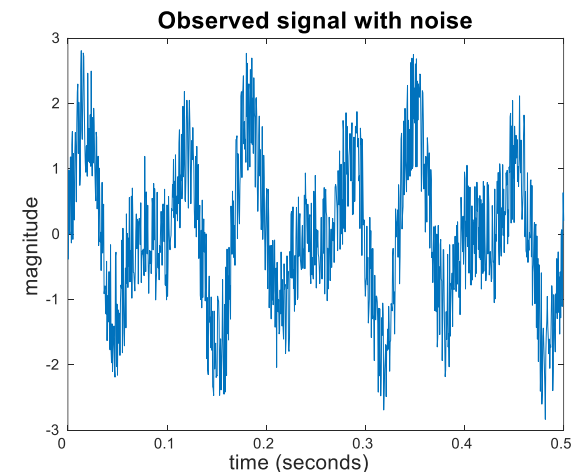
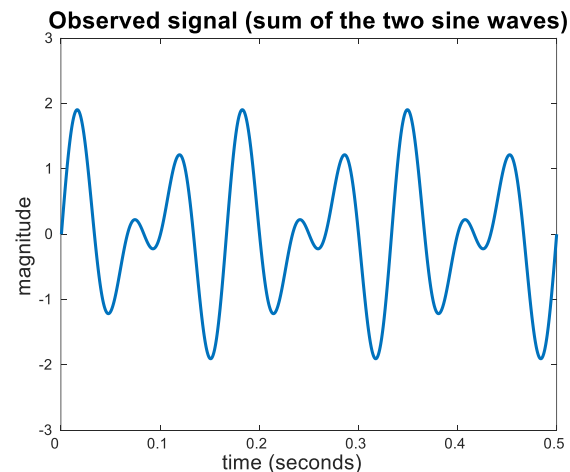
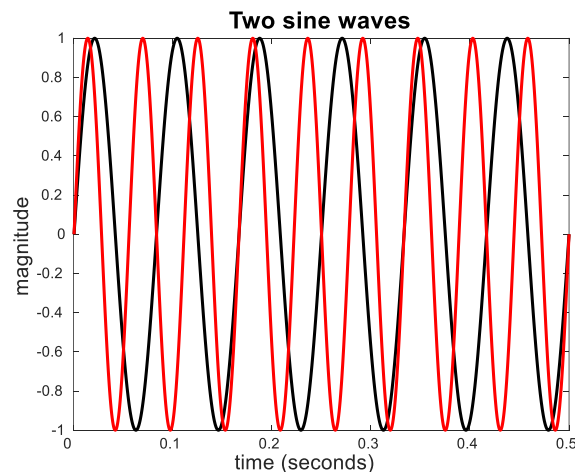


Data Quality

- Poor data quality negatively affects many data processing efforts.
- Data mining example: a classification model for detecting people who are loan risks is built using poor data:
- Some credit-worthy candidates are denied loans.
- More loans are given to individuals that default.
- Main Issues -
 - What kinds of data quality problems?
 - How can we detect problems with the data?
 - What can we do about these problems?
- Examples of data quality problems:
 - Noise and outliers
 - Wrong data
 - Missing values
 - Duplicate data

Noise

- For objects, noise is an extraneous object.
- For attributes, noise refers to modification of original values.
- Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen.
- The figures below show two sine waves of the same magnitude and different frequencies, the waves combined, and the two sine waves with random noise.
- The magnitude and shape of the original signal is distorted.



Outliers

- Outliers are either (1) data objects that, in some sense, have characteristics that are different from most of the other data objects in the data set, or (2) values of an attribute that are unusual with respect to the typical values for that attribute.
- Alternatively, they can be referred to as anomalous objects or values.
- It is important to distinguish between the notions of noise and outliers.
- Unlike noise, outliers can be legitimate data objects or values that we are interested in detecting.
- For instance, in fraud and network intrusion detection, the goal is to find unusual objects or events from among a large number of normal ones.

Missing Values

- Reasons for missing values:
 - Information is not collected
 - (e.g., people decline to give their age and weight)
 - Attributes may not be applicable to all cases.
 - (e.g., annual income is not applicable to children)
- Handling missing values:
 - Eliminate data objects or variables.
 - Estimate missing values.
 - Example: time series of temperature
 - Ignore the missing value during analysis.

Duplicate Data

- A data set can include data objects that are duplicates, or almost duplicates, of one another. Many people receive duplicate mailings because they appear in a database multiple times under slightly different names.
- To detect and eliminate such duplicates, two main issues must be addressed.
- First, if there are two objects that actually represent a single object, then one or more values of corresponding attributes are usually different, and these inconsistent values must be resolved.
- Second, care needs to be taken to avoid accidentally combining data objects that are similar, but not duplicates, such as two distinct people with identical names.
- The term **deduplication** is often used to refer to the process of dealing with these issues.
- In some cases, two or more objects are identical with respect to the attributes measured by the database, but they still represent different objects.
- Here, the duplicates are legitimate, but can still cause problems for some algorithms if the possibility of identical objects is not specifically accounted for in their design.

Issues Related to Applications

- **Timeliness:** Some data starts to age as soon as it has been collected. In particular, if the data provides a snapshot of some ongoing phenomenon or process, such as the purchasing behavior of customers or web browsing patterns, then this snapshot represents reality for only a limited time. If the data is out of date, then so are the models and patterns that are based on it.
- **Relevance:** The available data must contain the information necessary for the application. Consider the task of building a model that predicts the accident rate for drivers. If information about the age and gender of the driver is omitted, then it is likely that the model will have limited accuracy unless this information is indirectly available through other attributes.
- Making sure that the objects in a data set are relevant is also challenging. A common problem is sampling bias, which occurs when a sample does not contain different types of objects in proportion to their actual occurrence in the population. For example, survey data describes only those who respond to the survey.

Issues Related to Applications

- **Knowledge about the Data:** Ideally, data sets are accompanied by documentation that describes different aspects of the data; the quality of this documentation can either aid or hinder the subsequent analysis.
- For example, if the documentation identifies several attributes as being strongly related, these attributes are likely to provide highly redundant information, and we usually decide to keep just one. (Consider sales tax and purchase price.)
- If the documentation is poor, however, and fails to tell us, for example, that the missing values for a particular field are indicated with a -9999, then our analysis of the data may be faulty.

Thank You!