

# Data Mining

## Types of Data and Data Quality

Slides by Tan, Steinbach, Kumar adapted by Pimprapai Thainiam

# Data

There are four main data-related issues that are important for successful data mining:

## 1. The types of data:

The attributes used to describe data objects can be of different types:

- Quantitative: continuous value, discrete value, interval data, ratio data
- Qualitative: nominal data, ordinal data
- Special characteristics: time series data, objects with explicit relationship to one another

The type of data determines which tools and techniques can be used to analyze the data.

<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

## 2. The quality of data:

The real-world datasets are 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 are:

- Noise and Outliers
- Missing value
- Inconsistent value
- Duplicate data

# Data

## 3. Data preprocessing:

Often, we need to preprocess the data to make it more suitable for data analysis.  
The two goals of data preprocessing are

- To improve data quality
- To modify the data so that it better fits a specified data mining technique

We can either

- change the type of a particular attribute
  - reduce number of attributes
  - reduce number of data objects
  - create new attributes
- etc.

# Data

## 4. Analyzing data in term of its relationship between data objects

Some data mining techniques can only perform analysis using relationships between the data objects rather than the data objects themselves. Thus, we have to compute relationships of the data objects first and then perform the remaining analysis.

There are two types of relationships:

- 1) Similarity
- 2) Dissimilarity (Distance)

	1	2	3	4
1	1	0.5	1	1
2	0.5	1	1	1
3	1	1	1	1
4	1	1	1	1

Tid	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

# Topics

- **Types of Data**
  - **Attributes and Measurement**
  - Types of Data Sets
- Data Quality

# What is Data?

- A data set can be viewed as a collection of **data objects** and **their attributes**
- **An object** can be described by a collection of attributes
  - Object is also known as record, **point**, **case**, **sample**, **entity**, or **instance**
- **An attribute** is **a property or characteristic of an object** that may vary, either from one object to another or one time to another
  - Example1: age, height, weight, gender etc.
  - Example2: temperature, wind speed etc.
  - Attribute is also known as **variable**, **field**, **characteristic**, or **feature**

Attributes				
Tid	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

# What are Attribute and Attribute Values?

- **Attribute** is a **property or characteristic of an object** that may vary, either from one object to another or one time to another; for example,
  - Eye color varies from person to person; it is a **symbolic attribute**.
  - Temperature of an object varies over time; it is a **numerical attribute**.
- **Attribute values** are **symbols or numbers** assigned to an attribute; for example, brown eyes, body temperature of 38 degree Celsius.

<i>Tid</i> Refund Marital Status Taxable Income Cheat					Attributes
1	Yes ✓	Single ✓	125K ✓	No ✓	Objects
2	No	Married	100K	No	
3	No	Single	70K	No	

Attribute Values

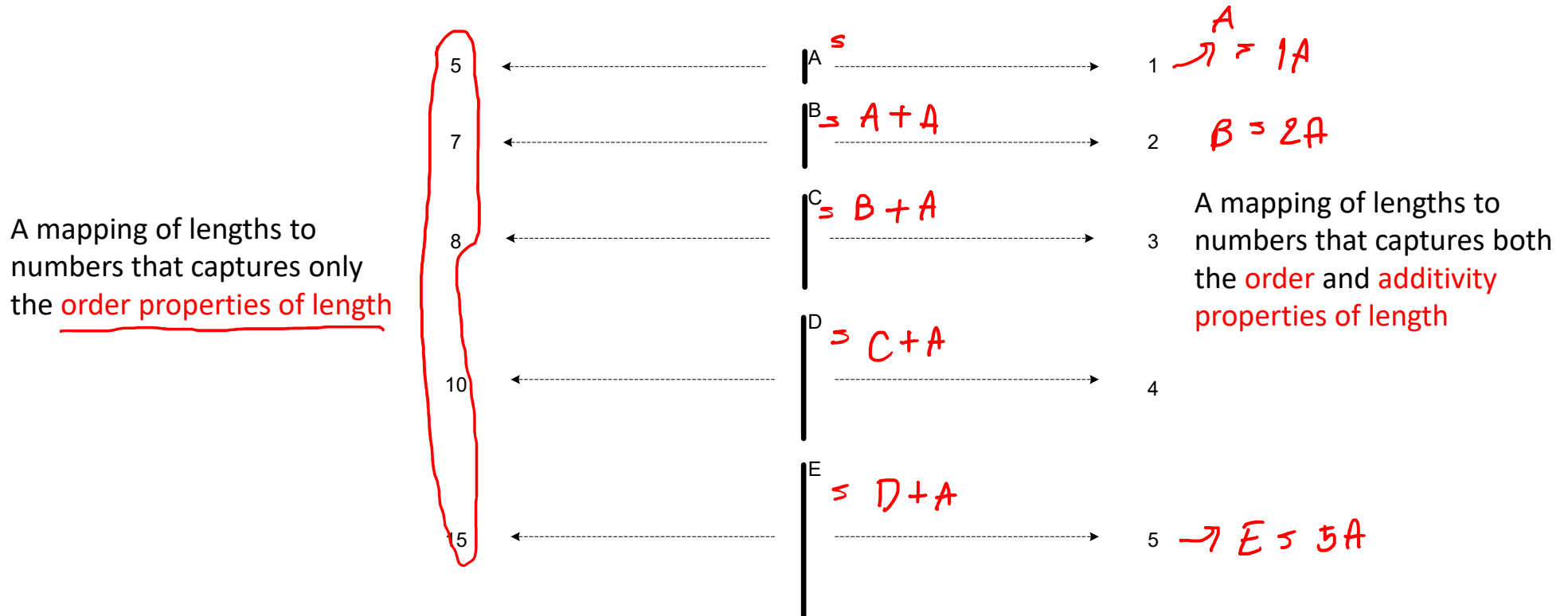


# What are Attribute and Attribute Values?

- A **measurement scale** is a rule (function) that associates a numerical or symbolic value with an attribute of an object.
- The **process of measurement** is the application of a measurement scale to associate a value with a particular attribute of a specific object; for example, we step on a bathroom scale to determine our weight, we classify some one as male or female.
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values by using different measurement scale; for example, height can be measured in feet or meters  
*Handwritten: 5 - 7 feet 150 - 200*
  - Different attributes can be mapped to the same set of values but properties of attribute values can be different; for example, attribute values for ID and age (in years) are integers  
*Handwritten: No. Child, No. pets = {0, 1, 2, 3, 4, ..., 10}*

# What are Attribute and Attribute Values?

The measurement of the length of line segments on two different scales of measurement.



# Properties of Numbers

- The following properties (operations) of numbers are typically used to describe attributes:

1. Distinctness : = and  $\neq$
2. Order :  $\leq$ ,  $\geq$ ,  $<$ , and  $>$
3. Addition : + and -
4. Multiplication :  $\times$  and  $\div$

obj.1 obj.2  
ex.  $A \neq A \leftarrow$  attribute value  
 $small < medium, small \neq medium.$   
birthdate  $\rightarrow 6 \neq 10, 6 < 10, 6 \neq 10$   
age  $\rightarrow 30 \neq 15, 30 \neq 15, 30 > 15,$   
 $30 \neq 15$

# Types of Attributes

- Given these four properties of numbers, we can define four types of attributes:
  1. **Nominal:** The values of a nominal attribute are just different names, i.e., nominal attributes **provide only enough information to distinguish one object from another**, for examples, ID numbers, eye color, zip codes, sex
  2. **Ordinal:** The values of an ordinal attribute **provide enough information to order objects**, for example, satisfactory in 7 points rating scale, grades, height in {tall, medium, short}, hardness of minerals in {good, better, best}, grades, street numbers
  3. **Interval:** For interval attributes, **the differences between values are meaningful**, i.e., a unit of measurement exists, for example, calendar dates, year, temperature.  
 $\hookrightarrow \frac{10}{5} = 2 \times$  no meaning in this value
  4. **Ratio:** For ratio variables, **both differences and ratios are meaningful**, for example, length, distance, time, counts, age, area, mass.  
 $\hookrightarrow 5 \in \text{obj 1} \quad \frac{10}{5} = 2 \rightarrow \text{obj 2 is 2 times longer than obj 1,}$   
 $10 \in \text{obj 2}$

# Types of Attributes

Attribute Type		Possessing Property			
		Distinctiveness (= and $\neq$ )	Order ( $\leq$ , $\geq$ , $<$ , and $>$ )	Addition (+ and $-$ )	Multiplication ( $\times$ and $\div$ )
Categorical (Qualitative)	Nominal	✓			
	Ordinal	✓	✓		
Numerical (Quantitative)	Interval	✓	✓	✓	
	Ratio	✓	✓	✓	✓

↪ integer, continuous

Note that quantitative attributes can be integer or continuous

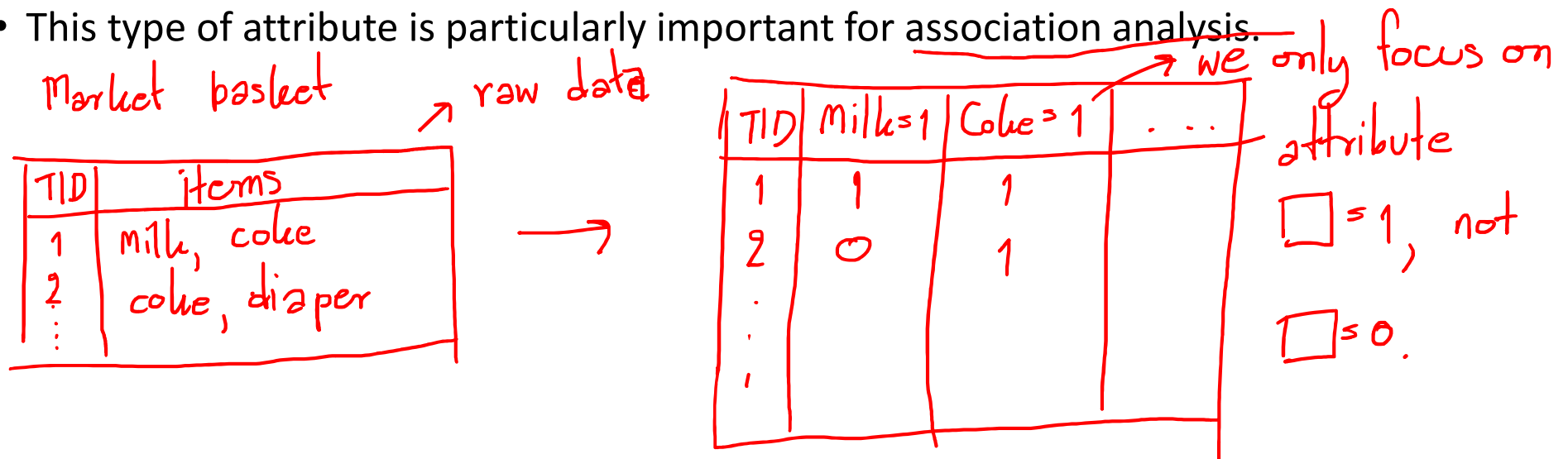
# ~~Discrete~~ Describing and Continuous Attributes

An independent way of distinguishing between attributes is by the number of values they can take

- 1. Discrete Attribute:** A discrete attribute has a finite or countably finite set of value which is often represented as integer variables. Such attributes **can be categorical**, such as zip codes or ID number, **or numeric**, such as counts. Note that binary attributes are a special case of discrete attributes and has only two values, e.g., true/false, yes/no, 0/1.
- 2. Continuous Attribute:** A continuous attribute is one whose value are real numbers which are typically represented as floating-point variables. Practically, real values can only be measured and represented using a finite number of digits. For example, temperature, height, or weight.

# Asymmetric Attributes

- For asymmetric attributes, only presence (a non-zero attribute value) is regarded as important because it is more meaningful and more efficient to focus on the non-zero values.
- Binary attributes where only non-zero values are important are called **asymmetric binary attributes**.
- This type of attribute is particularly important for association analysis.



# Topics

- **Types of Data**
  - Attributes and Measurement
  - **Types of Data Sets**
- Data Quality



# General Characteristics of Data Sets

Three characteristics that apply to many data sets are

1. **Dimensionality:** The dimensionality of a data set is the number of attributes that the objects in the data set possess. The difficulties associated with analyzing high-dimensional data are referred to as **the curse of dimensionality**. Thus, it is important to apply **dimensionality reduction technique** in order to reduce dimensions of the data set.

0	0	1	0	0
0	0	0	0	0
1	0	0	0	0

2. **Sparsity:** Most attributes of an object have value of zero. (fewer than 1% of the entries are non zero)

3. **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.

# Types of Data Sets

**Three main groups of data are**

1. Record data
2. Graph based data
3. Ordered data

# Record Data

**1. Record data:** Data that consists of a collection of records (data objects), each of which consists of a fixed set of data fields (attributes), e.g., transaction or market basket data, data matrix, sparse data matrix.

<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

# Record Data

## Transaction Data

- **Transaction data** is considered as a special type of record data, where each record (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.


5  
transactions  
, baskets

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

# Record Data

## Transaction Data

- Transaction data can be converted to a normal record data whose attributes are asymmetric attributes. The attributes are binary, indicating whether or not an item was purchased.



TID	Bread	Coke	Milk	Beer	...	Diaper
1	1	1	1	0		0
2	1	0	0	1		0
3	0	1	1	1		1
4	1	0	1	1		1
5	0	1	1	0		1

# Record Data

## Data Matrix

- **Data matrix** refers to data set where data objects have **the same fixed set of numeric attributes**. The data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.
- Such 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.

4 attributes

5 objects	Sepal Length	Sepal Width	Petal Length	Petal Width
	5.6	2.7	4.2	1.3
	4.5	3.0	5.8	2.2
	5.2	3.1	4.8	1.4
	6.1	4.1	1.7	0.8
	4.1	2.8	4.0	1.2

# Record Data

## Sparse Data Matrix

- A **sparse data matrix** is a special case of a data matrix in which the **attributes are of the same type and are asymmetric**.
- A document can be represented as a term vector, where each term is an attribute of the vector and the value of each attribute is the number of times the corresponding term occur in the document.
- This representation of a collection of documents is called a **document term matrix**.

	team	coach	play	ball	score	game	win	lost	timeout
Document 1	3	0	5	0	2	6	0	2	0
Document 2	0	7	0	2	1	0	0	3	0
Document 3	0	1	0	0	1	2	2	0	3

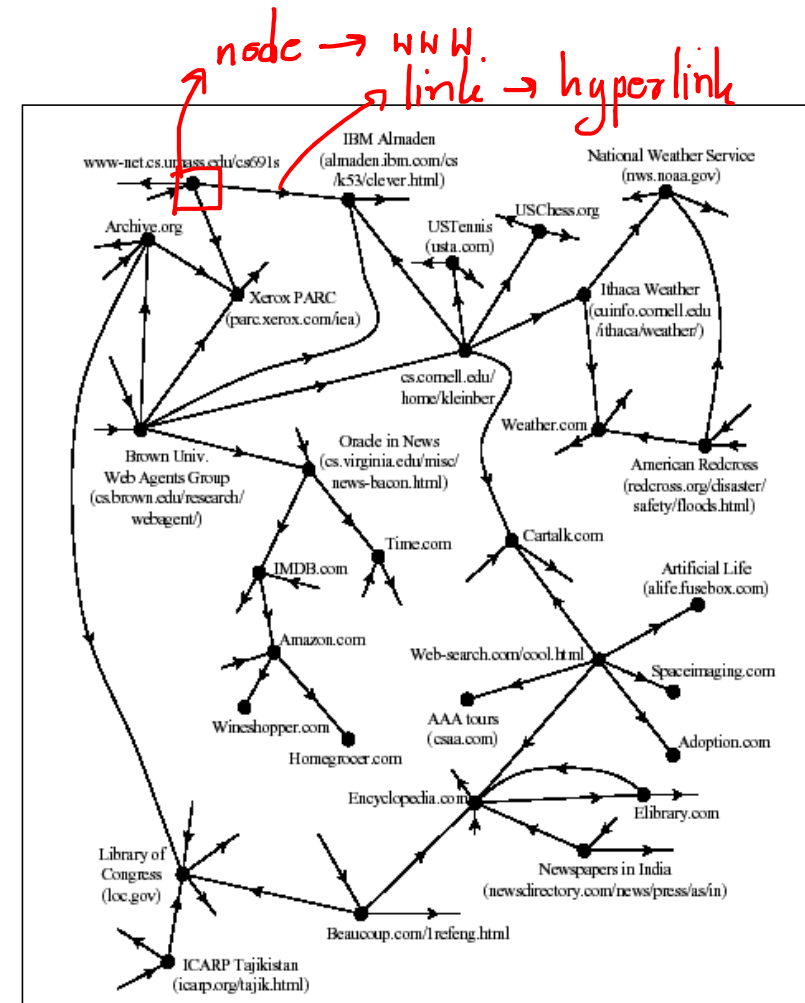
- Converted transaction data and document-term data matrix are examples of sparse data matrix.

# Graph Based Data

## 2. Graph based data: We consider two specific cases:

**(1) Data with relationships among data objects:** The relationships among objects frequently convey important information. The data is often represented as a graph where data objects are mapped to nodes of the graph while the relationship among objects are captured by the links between objects and link properties, such as direction and weight, e.g., World Wide Web Graph.

The world-wide web can be viewed as a directed graph in which the vertices are static pages, and two vertices are connected by an edge if one page contains a hyperlink to the other.

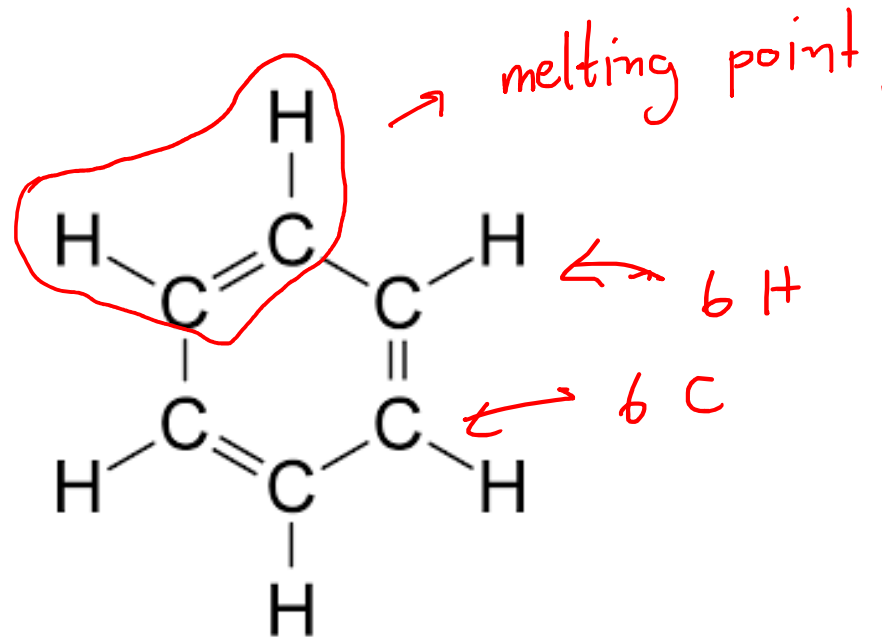




# Graph Based Data

**(2) Data with objects that are graphs:** If objects have structure, that is, the objects contain subobjects that have relationships, then such objects are frequently represented as graphs, e.g., the structure of chemical compound.

Benzene Molecule:  $C_6H_6$



# Ordered Data

3. **Ordered data:** Data that the attributes have relationships that involve order in time or space, e.g.,

- Sequential data
- Sequence data
- ~~T~~ime series data
- Spatial data

space related  $\nwarrow$   
space  $\leftarrow$  spatio,  
time related  $\leftarrow$  temporal

# Ordered Data

## Sequential Data

- **Sequential data**, also referred to as temporal data, can be thought of as **an extension of record data, where each record has time associate with it.**

*time related.*

Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

# Ordered Data

## Sequence Data

- **Sequence data** consists of a data set that is a **sequence of individual entities**, such as a sequence of words or letters, genetic sequence.
- It is quite similar to sequential data, except that there are no time stamps; instead **there are positions in an ordered sequence**.

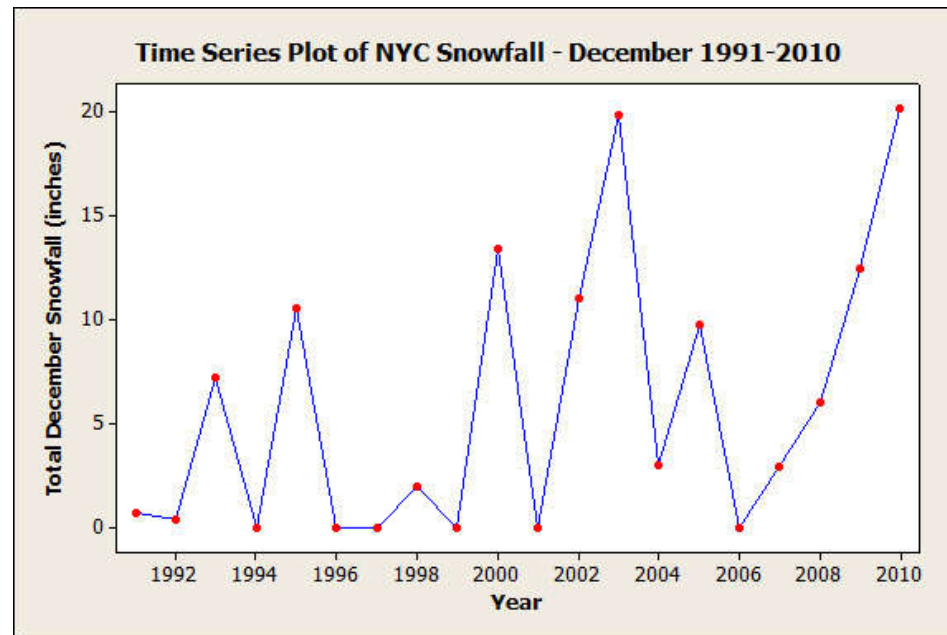
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

↪ genomic data

# Ordered Data

## Time Series Data

- **Time series data** is a special type of sequential data in which **each record is a time series**, i.e., a series of measurements taken overtime. For example, daily prices of various stocks, average monthly temperature of Bangkok.

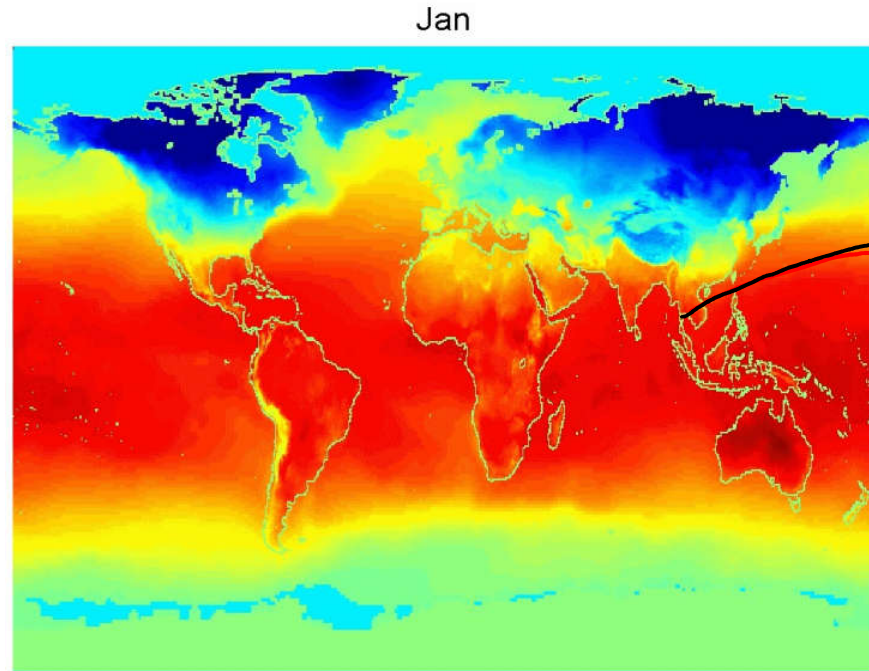


# Ordered Data

## Spatial Data

- Some objects **have spatial attributes**, such as positions or areas, as well as other types of attributes, e.g., weather data that is collected for a variety of geographical locations.

Average Monthly Temperature  
of land and ocean



temperature,  
m  
Thailand

# Topics

- Types of Data
  - Attributes and Measurement
  - Types of Data Sets
- **Data Quality**

# Data Quality

- Data mining focuses on
  - (1) The detection and correction of data quality problems, is often called **data cleaning**.
  - (2) The use of algorithms that can tolerate poor data quality
- **Examples of data quality problems:**
  - Noise
  - Outliers
  - Missing values
  - Inconsistent Values
  - Duplicate data

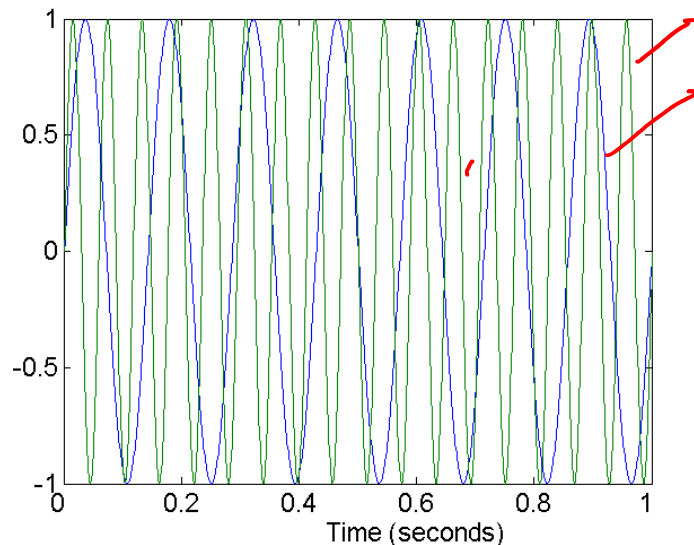


# Noise

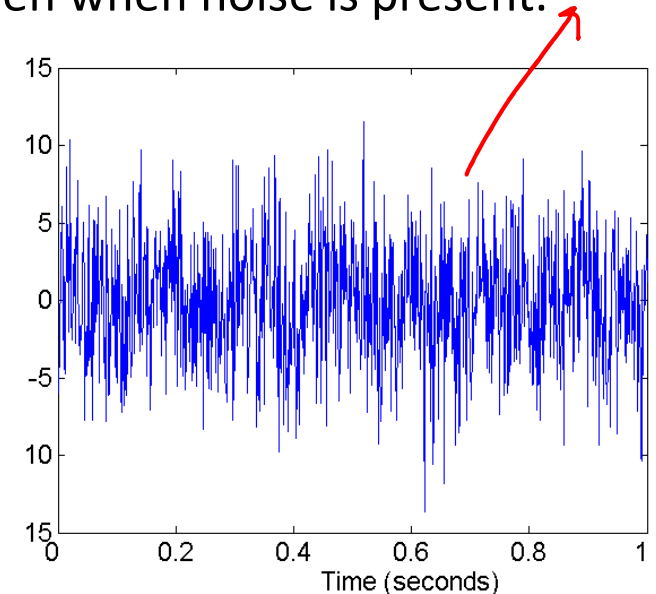
- **Noise** refers to distortion of a value or the addition of spurious objects, for example, distortion of a person's voice when talking on a poor phone, snow on television screen.
- Noise can be reduced by implementing techniques from signal or image processing.
- The elimination of noise is difficult; much work in data mining focuses on devising **robust algorithms** that produce acceptable results even when noise is present.



Noise in a  
time series  
context



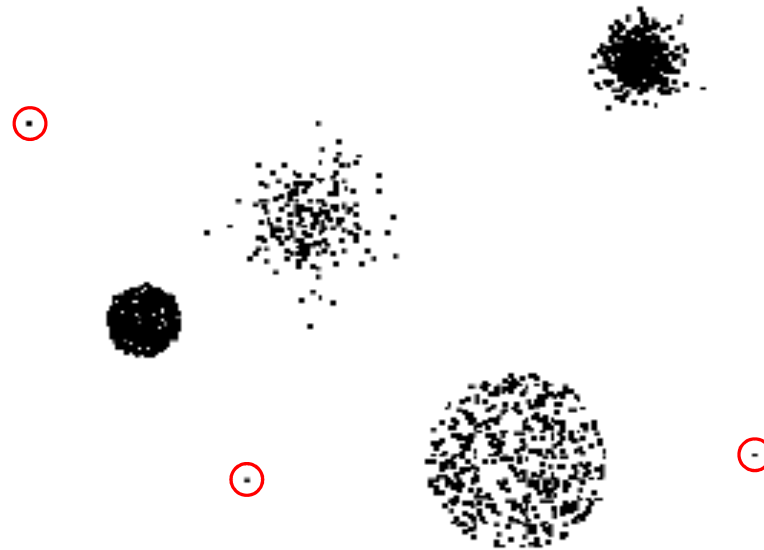
Two Sine Waves



Two Sine Waves + Noise

# Outliers

- Outliers are either
  - (1) Data objects that have characteristics that are different from most of the other data objects in the data set → anomalous objects
  - (2) Values of an attribute that are unusual with respect to the typical values of that attribute → anomalous values



# 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)

- **Correcting missing values:**

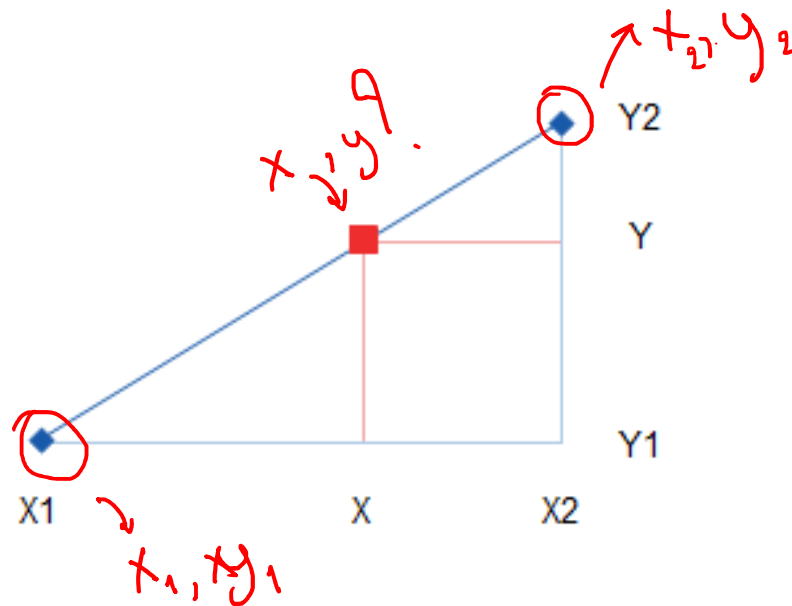
**1) Eliminate Data Objects or Attributes:** This method is simple and **effective when the object (attribute) contains several attributes (objects) with missing values, and not many objects have missing values.** It is not very effective when there are many objects have missing values. Eliminating attributes should be done with caution because the eliminated attributes may be the ones that are critical to the analysis.

# Missing Values

2) **Estimate Missing Values:** The missing value of an object can be reliably estimated. Techniques used to estimate missing value are

**For time series data;**

**(1) Interpolation** uses 2 nearest neighbors



$$\frac{(X - X1)}{(X2 - X1)} = \frac{(Y - Y1)}{(Y2 - Y1)}$$

$$Y = Y1 + (X - X1) \frac{(Y2 - Y1)}{(X2 - X1)}$$

# Missing Values

For record data, we can estimate a missing data value by using similar data points

*all data objects have the same values except the missing one.*

(2) Average value of nearest neighbor data objects is used to estimate a missing value for continuous attributes

(3) Most commonly occurring attribute value of nearest neighbor data objects is used to estimate a missing value for categorical attributes

TID	Sex	Age	Race	Marital Status
1	Male	30	Thai	Single
2	Female	35	X	Single
3	Female	35	USA	Single
4	Male	Y	Thai	Single
5	Female	35	<del>USA</del> Thai	Single
6	Male	32	Thai	Single
7	Female	35	<del>USA</del> Thai	Single

X = USA

$$Y = (30 + 32) / 2 = 31$$

# Missing Values

↗ Used when there is no nearest neighbor data object

**3) Fill in with Mean or Mode :** Similar to 2) but use all data objects to estimate the missing values when there is no similar data points to that missing data point. This method will fill in all missing values in that particular attribute with one value.

**(1) Average (mean)** used for numerical attributes

**(2) Most commonly occurring attribute value (mode)** for categorical attributes

**4) Ignore the Missing Value During Analysis:** Many data mining approaches can be modified to ignore missing value. Ignoring missing values does cause some deviation in the results.

# Inconsistent Values

- **Reasons for inconsistent values:**

- Entering the data manually (e.g. transpose two digits in a zip code, 10502 instead of 10520)
- Misspelling when the information was entered manually (e.g. Bamgkok instead of Bangkok, or bangkok instead of Bangkok)
- Misreading when the information was scanned from a hand written form (e.g. Bamgkok instead of Bangkok, or bangkok instead of Bangkok)

- **Two inconsistent types**

1. **Inconsistent values within an attribute**

- for example, two digits in a zip code, 10502 instead of 10520.

*no, ~~10502~~ 10502 in Ladkrabang district*

2. **Inconsistent values between attributes**

- for example, a person height is 160 cm. but his weight is 20kg.

# Inconsistent Values

- **Detecting inconsistent values:** Depending on attributes' characteristic, for example,
  - A person's height should not be negative
  - A person's age should not be more than 120
  - zip codes should have 5 digits
- **Correcting inconsistent values:** The correction of an inconsistency requires additional or redundant information.
  - 1) **Double-check with the original data source**, e.g., a hand written form
  - 2) **Check validity of a value of an attribute with another attribute**, e.g. no 10502 (zip code attribute) in Ladkrabang City (city attribute)
  - 3) **Check validity of a value by considering its attribute's characteristic**, e.g. a person height



# Duplicate Data

- A data set may include data objects that are duplicates, or almost duplicates of one another.
- **Reasons for duplicate data:**
  - Human error of having customers providing slightly different information at different points in time, e.g., a consumer lists his name as Jonathan Smith on one form and Jon Smith on another
  - Changing demographic data, e.g., name changes, address changes
  - Merging data from heterogenous sources

# Duplicate Data

- **Issues of duplicate data:**

- (1) Two objects that are identical representing a single object (duplicates) → easy to be detected
- (2) Two objects that are similar representing a single object (almost duplicates) → inconsistent problem
- (3) Two objects that are identical with respect to the attributes but they still represent different objects (legitimate duplicate) → hard to prove, usually be detected as duplicates

# Duplicate Data

- **Detecting duplicate data:**
  - Set a matching level, e.g., 90% or 100% of two objects will be considered as duplicates
- **Correcting duplicate data:**
  - 1) Remove a duplicated object
  - 2) Solve inconsistent problem and merge to one object