Part 2
Data

NGURAH AGUS SANJAYA ER, S.KOM, M.KOM
E-mail: agus.sanjaya@cs.unud.ac.id

Outline

- Attributes & Objects
- Types of Data
- Data Quality
- Data Preprocessing

What is Data?

- Data set → a 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, or feature**
- A collection of attributes describe an object
  - **Object is also known as record, point, case, sample, entity, or instance**

Attributes

Objects

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

Attribute

- A property or characteristic of an object that may vary, either from one object to another or from one time to another
- Example: eye color varies from person to person, while the temperature of an object varies over time
- Eye color use symbolic attribute, temperature is a numerical attribute (potentially unlimited number of values)
- To discuss and more precisely analyze the characteristics of objects → assign numbers or symbols to them
- To do this in a well-defined way → measurement scale

Measurement Scale

- A rule (function) that associates a numerical or symbolic value with an attribute of an object

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - **Same attribute can be mapped to different attribute values**
    - Example: height can be measured in feet or meters

  - **Different attributes can be mapped to the same set of values**
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

- How to identify the type of an attribute?
  - **Identify the properties of numbers that correspond to underlying properties of the attribute**
    - Example: attribute such as length has many properties of numbers. It makes sense to compare and order objects by length, as well as to talk about the differences and ratios of length
- The following properties (operations) of numbers are typically used to describe attributes:
  - **Distinctness (= and ≠)**
  - **Order (<, ≤, >, and)**
  - **Addition (+ and -)**
  - **Multiplication (* and /)**

Types of Attributes (2)

- 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 in {tall, medium, short}
  - **Interval**
    - Examples: calendar dates
  - **Ratio**
    - Examples: length, time, counts

Properties of Attribute Values / 1

- The type of an attribute depends on which of the following properties it possesses:
  - **Nominal attribute: distinctness**
  - **Ordinal attribute: distinctness & order**
  - **Interval attribute: distinctness, order & addition**
  - **Ratio attribute: all 4 properties**

Properties of Attribute Values / 2

Attribute Type	Description	Examples	Operations
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. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Properties of Attribute Values / 3

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

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**

- 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

- Only presence (a non-zero attribute value) is regarded
- Examples:
  - **Words present in documents**
  - **Items present in customer transactions**

	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

Types of data sets

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

Important Characteristics of Structured Data

- Dimensionality
  - **Curse of Dimensionality**
- Sparsity
  - **Only presence counts**
- Resolution
  - **Patterns depend on the scale**

Record Data

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

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 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

Sparse Data Matrix

- 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.**

Document Data	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 or Market Basket Data

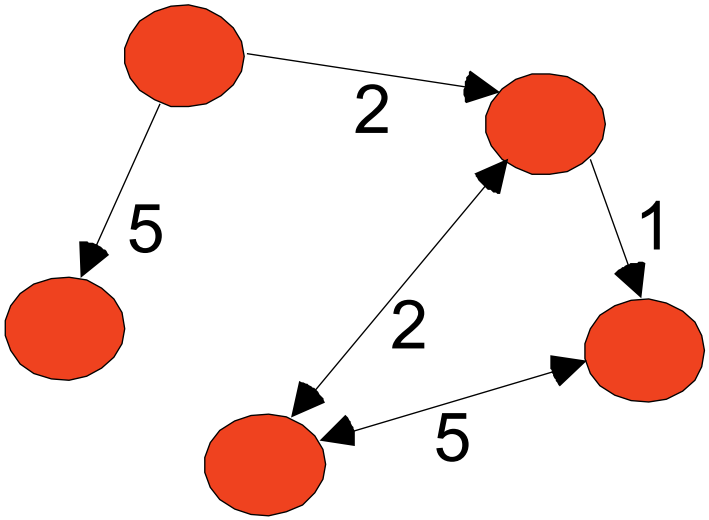
- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

Transaction Data

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

Graph Data

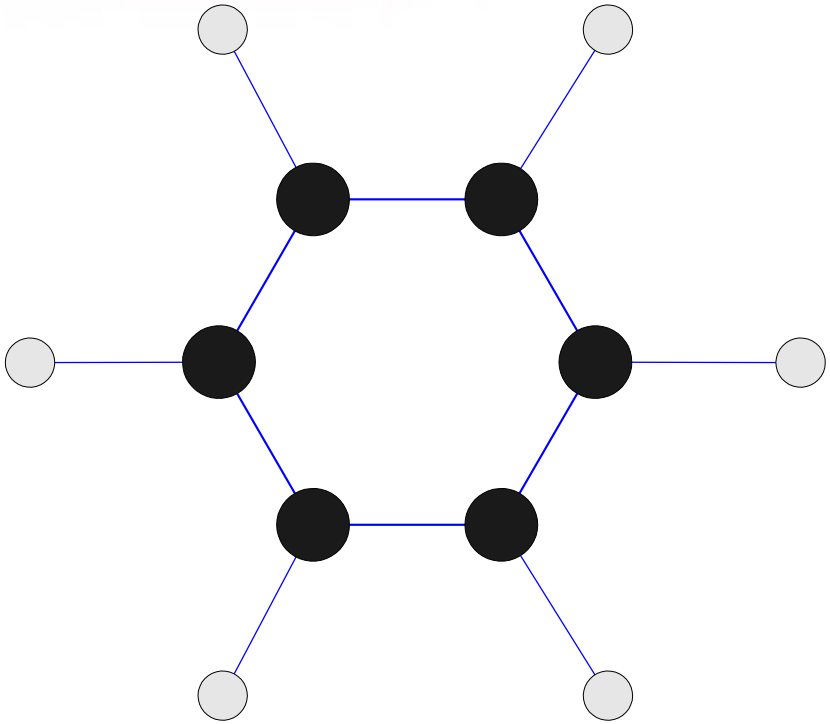
- Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

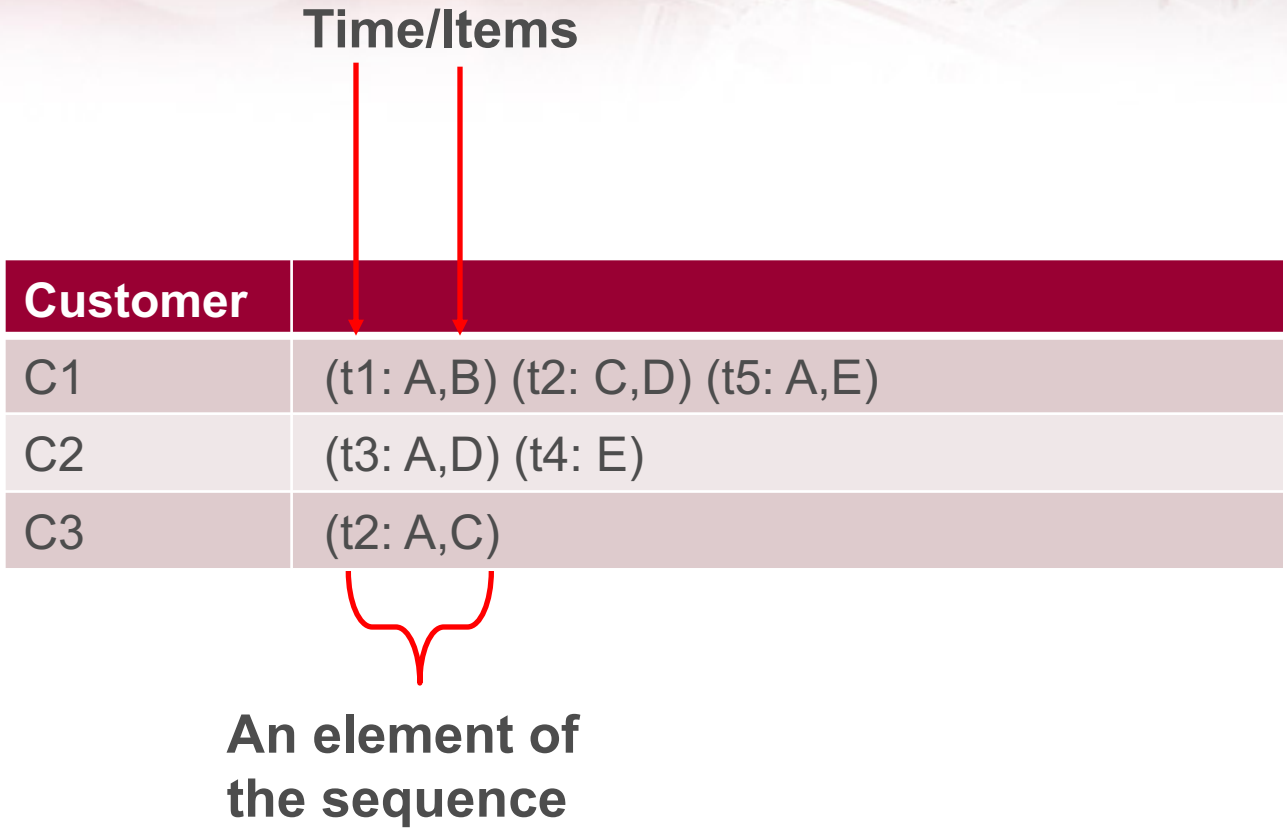
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data / 1

- Sequences of transactions



Ordered Data / 2

- Genomic sequence data

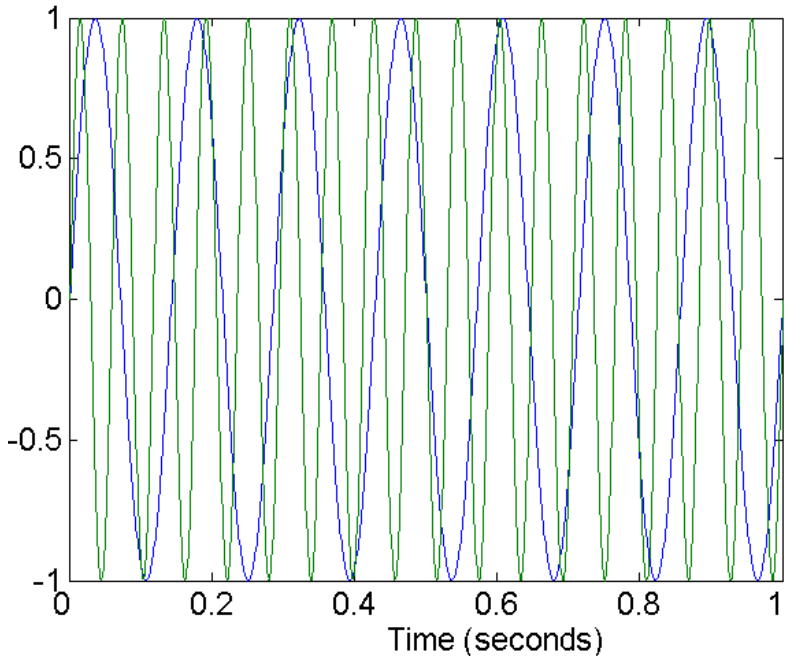
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Data Quality

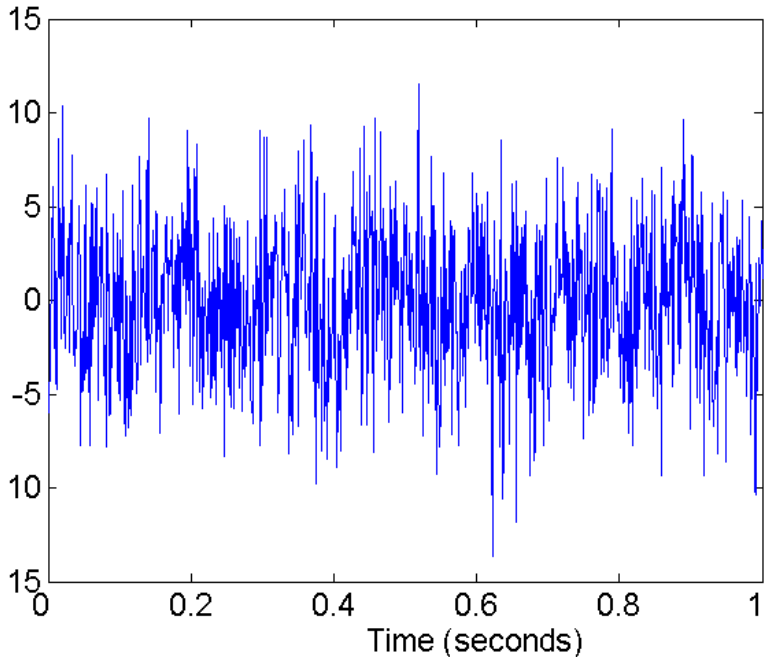
- Data collection issues
  - **Measurement error → any problem that arise due to the measurement process**
  - **Data collection error → error because of omitting data objects or attribute values as well as inappropriately including a data object**
- Examples
  - **Noise**
  - **Outlier**
  - **Missing values**

Noise

- Random component of a measurement error.
  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**



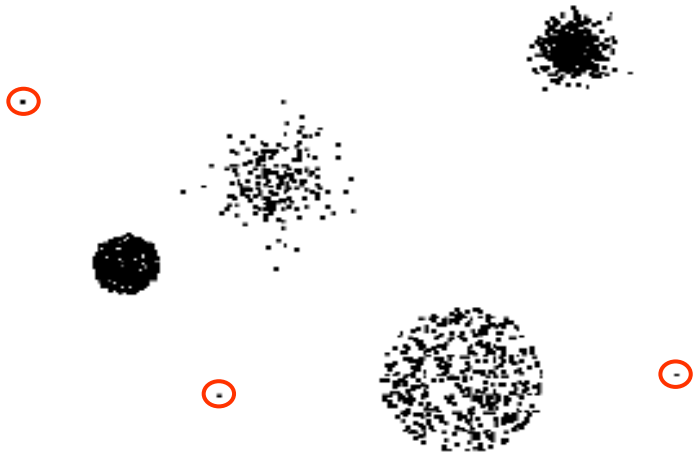
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Data objects with characteristics that are considerably different than most of the other data objects in the data set
- Values of an attribute that are unusual with respect to the typical values for that attribute



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 attributes**
  - **Estimate missing values**
  - **Ignore the missing value during analysis**
  - **Replace with all possible values (weighted by their probabilities)**

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - **Major issue when merging data from heterogeneous sources**
- Examples:
  - **Same person with multiple email addresses**
- Data cleaning
  - **Process of dealing with duplicate data issues**

Data Quality (2)

- Issues related to applications: specific to particular applications and fields
  - **Timeliness → data starts to age as soon as it has been collected**
  - **Relevance → available data must contain the information necessary for the application**
  - **Knowledge about the data → data must be accompanied with documentation that describes different aspects of the data**