# Machine Learning

## Lecture 4: Data

COURSE CODE: CSE451

2023

# Course Teacher

**Dr. Mrinal Kanti Baowaly**

Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: mkbaowaly@gmail.com

# DATA

- Data can be any unprocessed fact, value, text, sound, picture or video that is not being interpreted and analyzed
- Data is the most important part of all Data Mining, Machine Learning, Artificial Intelligence
- Without data, we can’t train any model and all modern research and automation will go vain
- Big Enterprises are spending loads of money just to gather as much certain data as possible
- Example: Facebook acquires WhatsApp by paying a huge price of $19 billion

# Information and Knowledge

- Information: Processed, organized, or structured data to provide context and meaning.
- Knowledge: Combination of inferred information, experiences, learning and insights. Knowledge is **useful** and **actionable** information that can lead to impact.
- Machine Learning is a tool for turning information into knowledge

# Types of Data (Variable) in Statistics

<table>
<tr><td colspan="2">Numeric / Quantitative<br>Take numerical values only and the values reflect the actual measurement (with units) of the subjects or object</td><td colspan="2">Categorical / Qualitative<br>Contains categories only. Each category represents a particular characteristic of interest within a group of subjects or objects.<br>Typically text data, but numerically coded by statistical packages</td></tr>
<tr><td>Continuous<br>Takes any value in a range of values and additional values can be taken that fit between each consecutive value<br><br>Height (metres)<br>Age (days, months or years<br>Temperature (C)</td><td>Discrete<br>Normally takes integer values typically counts<br><br>Number of Children in family<br>Number of Asthma attacks</td><td>Ordinal<br>Categories are mutually exclusive & have a ranked order, however, each interval may not be equally spread<br><br>Cancer Staging: 0 1 2 3 4<br><br>Likert scale: strongly agree, agree, neutral, disagree, strongly disagree</td><td>Nominal<br>Categories are mutually exclusive but have no implicit order<br><br>Blood Group: A, B, AB, O<br><br>Eye Colour: Blue, Brown, Green</td></tr>
<tr><td colspan="3"></td><td>Binary<br>Same as Nominal but only 2 possibilities<br>Gender: Male, Female<br>Status: Dead, Alive</td></tr>
</table>

# Quantitative data vs Qualitative data

Quantitative data

- Number-based, countable, or measurable, also known as numerical data
- Tell us how many, how much, or how often in calculations
- Analyzed using statistical analysis
- Examples: measurable such as distance, area, time, speed, height, length, weight, cost; counts such as the number of website visitors, sales, or email sign-ups etc.

Qualitative data

- Interpretation-based, descriptive, and relating to language but not measured or counted, also known as categorical data
- Analyzed by grouping it in terms of meaningful categories
- Can help us to understand why, how, or what happened behind certain behaviors
- Examples: Employee ID, text, documents, color, marital status, nationality, gender, grades, education level, etc.

# Discrete data vs Continuous data

## Discrete Data

- Can be counted
- Has only a finite or countably infinite set of values
- Examples: the number of students in a class, the number of words in a document, the number of heads in 100 coin flips
- Often represented as integer variables.

## Continuous Data

- Can only be measured
- Has any value (real number) within a range
- Examples: temperature, height, or weight.
- represented as real or floating-point variables.

# Nominal data vs Ordinal data

Nominal Data

- Qualitative or categorical data
- Can’t be quantified, neither have any implicit ordering
- No numeric operations can be performed
- Examples: Colour of hair (White, Red, Brown, Black, etc.), Marital status (Single, Widowed, Married), Nationality (Indian, German, American), Gender (Male, Female, Others), Eye Color (Black, Brown, etc.)

Ordinal Data

- Qualitative or categorical data
- Have some kind of ranked order, and it is possible to assign numbers to the data
- It is possible to compare one item with another in terms of ranking.
- Examples: Grades in the exam (A, B, C, D, etc.), Ranking in a competition (First, Second, Third, etc.), Economic Status (High, Medium, and Low), Education Level (Higher, Secondary, Primary)

# What is Data set?

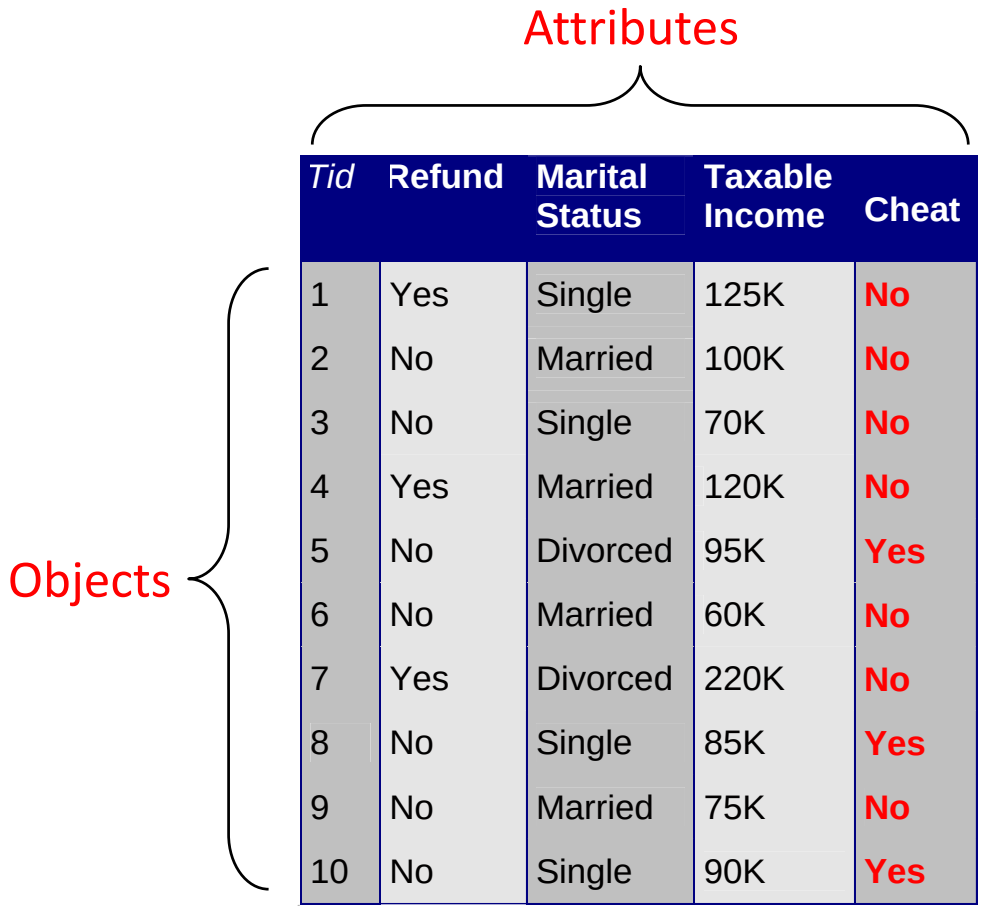
Collection of data objects and their attributes

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




| *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 |

# Types of Data sets

**1. Record**

- Data Matrix
- Document Data
- Transaction Data

**2. Graph**

- Generic
- World Wide Web
- Molecular Structures

**3. Ordered**

- Sequential Transaction Data
- Time Series Data
- Sequence Data
- Spatial and Spatio-Temporal Data

# 1. 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 $m \times n$ 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

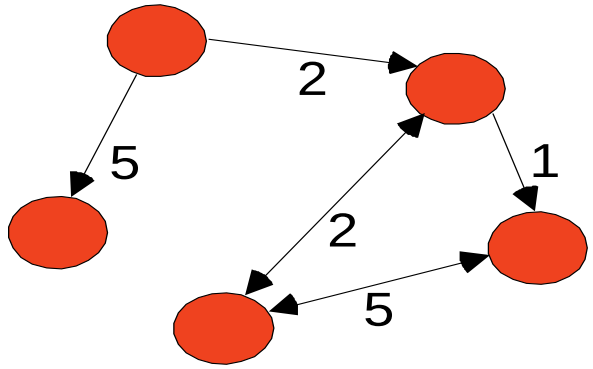
A special type of record data, where

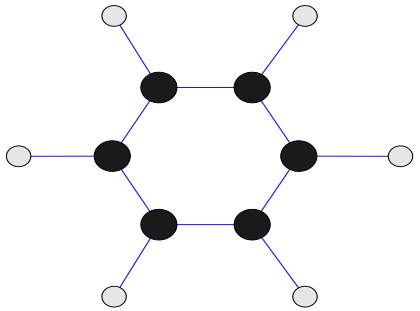
- each record (transaction) involves a set of items.
- For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

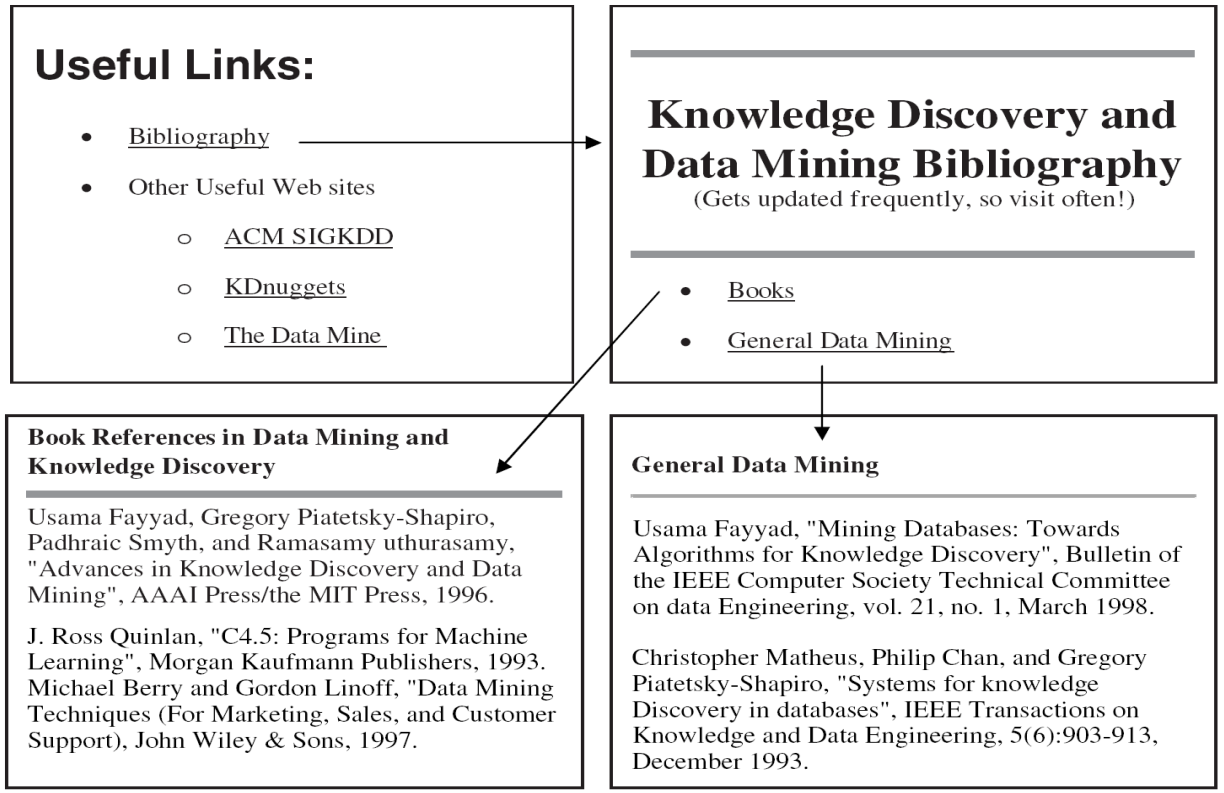
# 2. Graph Data

Examples: Generic graph, linked webpages/social networks, and a molecule



Benzene Molecule: C6H6



# 3. Ordered Data

**Sequential Transaction Data:**

- An extension of transaction data, where each transaction has a time associated with it.
- It is possible to find patterns such as "people who by DVD players, tend to buy DVDs immediately following the purchase."

| 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) |

(a) Sequential transaction data.

# 3. Ordered Data (Cont.)

**Sequence Data:**

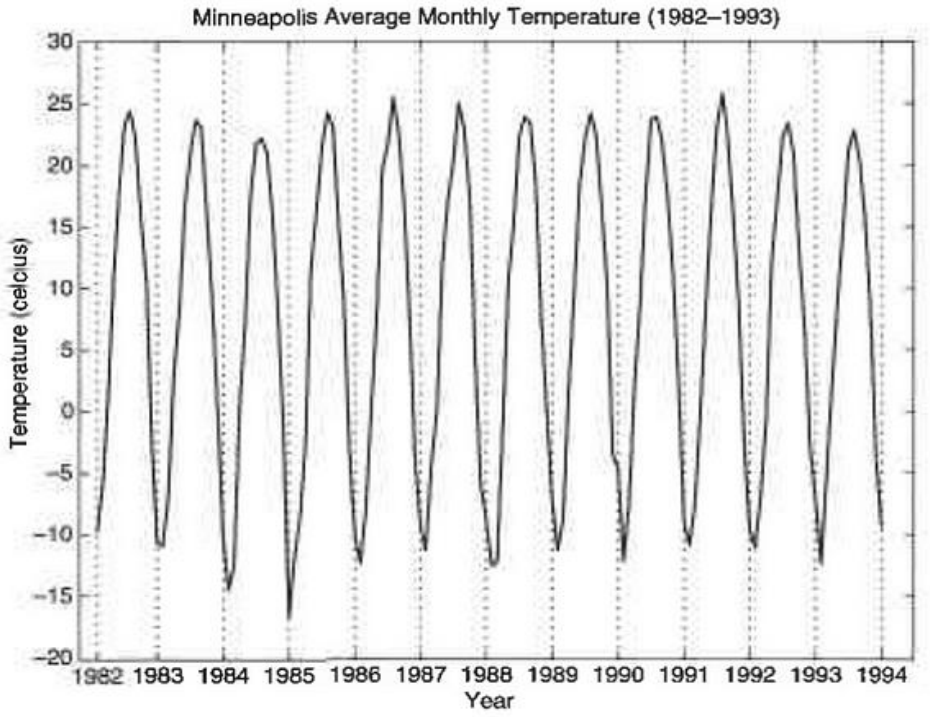
- Sequence of individual entities, such as sequence of words or letters.
- Have no time stamps; instead, there are positions in the ordered sequence. For example, the genomic sequence data have sequence of nucleotides (A, T, C, and G) that make up an organism's DNA.
- Enable advancements in biology, medicine, agriculture, and various other fields.


(b) Genomic sequence data.

# 3. Ordered Data (Cont.)

**Time Series Data:**

- Each record is a time series, i.e. a series of data collected at consistent intervals over a set period rather than just collecting the data intermittently or randomly
- One of the study's main goal is to predict future value

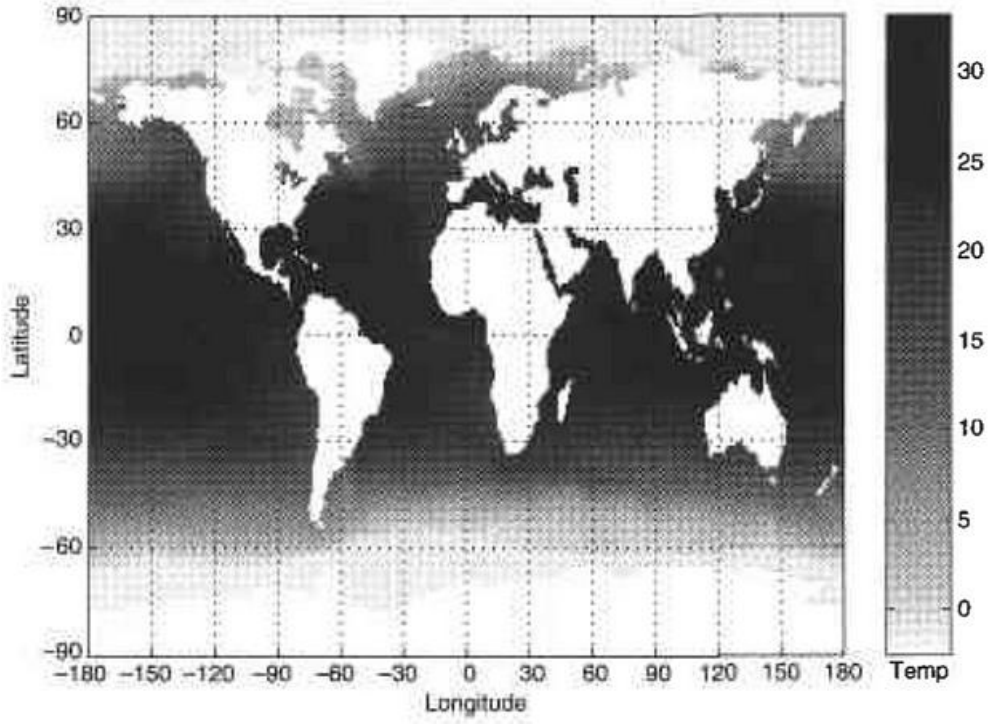



(c) Temperature time series.

# 3. Ordered Data (Cont.)

**Spatial and Spatio-Temporal Data:**

- Spatial data: have spatial attributes, such as locations or areas, for example, weather data.
- Spatio-temporal data: when spatial data are collected over time, for example, tracking the trajectories of objects such as vehicles, in time and space.




(d) Spatial temperature data.

# Test Your Understanding

- Take part in the following Quiz Test on Types of Data
- Click here

# How to get datasets for Machine Learning

- Popular sources for Machine Learning datasets
  - Kaggle Datasets
  - UCI Machine Learning Repository
  - Datasets via AWS
  - Google's Dataset Search Engine
  - Microsoft Datasets
  - Government Datasets
  - Computer Vision Datasets
  - Scikit-learn dataset

Source: JavaTPoint

# End of Lecture-4