**Characteristics of Data**

*What is****Dimensionality?***

→ The dimensionality of a data set is the number of attributes that the objects in the data set have.

In a particular data set if there are high number of attributes (also called high dimensionality), then it can become difficult to analyse such a data set. When this problem is faced, it is referred to as **Curse of Dimensionality.**

In order to understand what the hell is this **Curse of Dimensionality,**we first need to understand the other two characteristics of Data.

*What is****Sparsity?***

→ For some data sets, such as those with asymmetric features, most attributes of an object have values of 0; in many cases fewer than 1% of the entries are non-zero. Such a data is called **sparse data**or it can be said that the data set has **Sparsity.**

*What is****Resolution****?*

→ The patterns in the data 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 may 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.

Now, coming back to the **Curse of Dimensionality,**it means many types of Data Analysis becomes difficult as the dimensionality (number of attributes in the data set) of the data set increases. Specifically, as dimensionality increases, the data becomes increasingly sparse in the space that it occupies. For classification, this can mean that there are not enough data objects to allow the creation of a model that reliably assigns a class to all possible objects. For clustering, the definitions of density and the distance between points, which are critical for clustering, become less meaningful.

Central Tendency and Dispersion

**Converting Data to Information:**  The goal is to find out as much data as possible and convert it into meaningful information that can be used by the concerned personnel to make meaningful decisions about the process. However for that one needs to learn how to statistically deal with huge amounts of data.

Data primarily needs to be understood for its two characteristics viz central tendency and dispersion. Data tends to be centred around a point known as average. The degree to which it is spread out from that point is also important because it has an important bearing on the probability. It is for this reason that we use the following characteristics to make sense of the data involved:

**Measures of Central Tendency:** Different types of data need different measures of central tendency. Some of the important measures, commonly used are as follows:

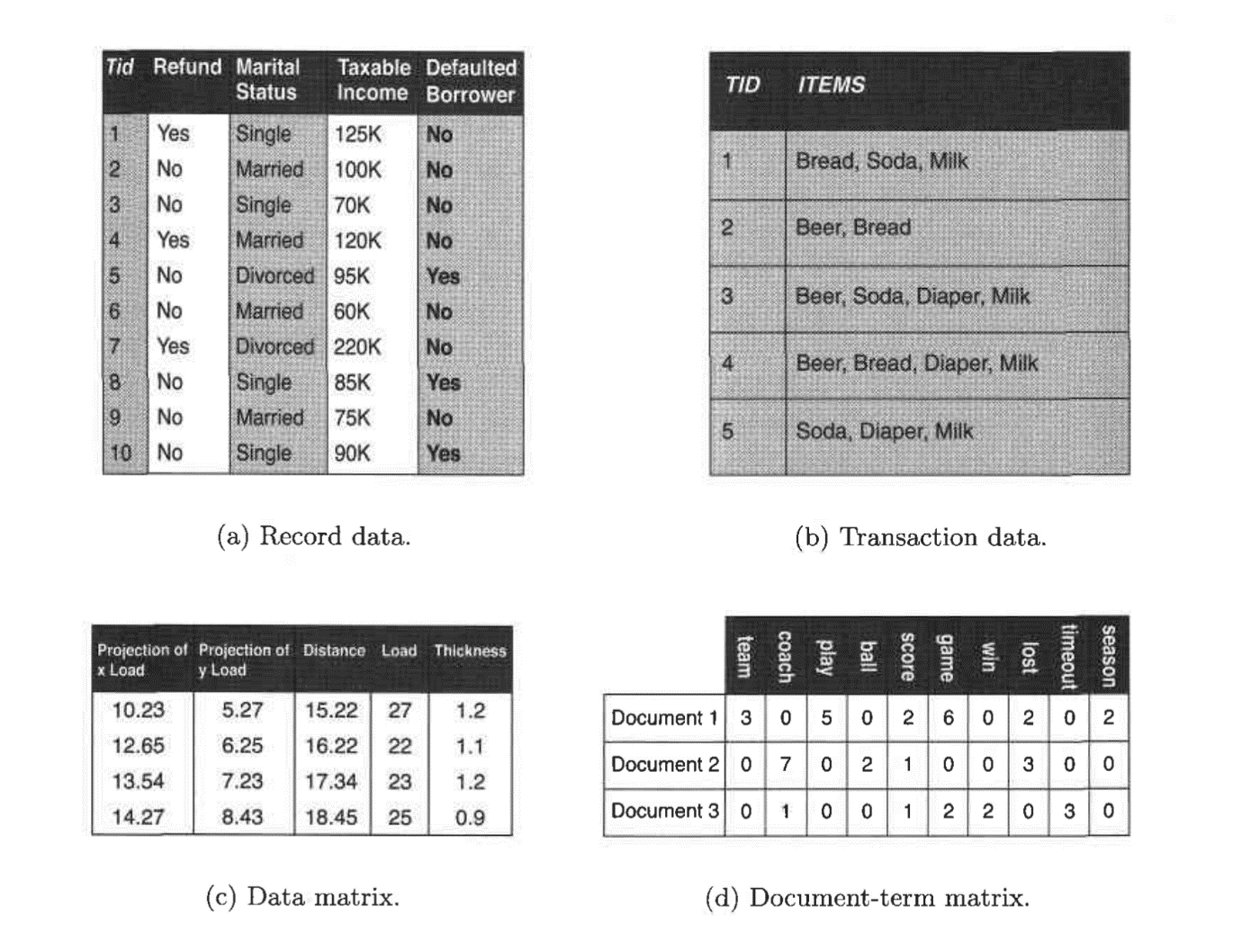
* **Mean:** This is most probably the arithmetic mean or simply the average of the data points involved. It could also be the geometric or harmonic mean however that is unusual. This is the most popular measure of central tendency. Many statistical techniques have evolved that use the mean as the primary measure to understand the centrality of a given set of data points.
* **Median:** If all the data points given in a particular data set were arranged in ascending or descending order, the value in the centre is called the median. In case where data sets have an odd number of elements like 7, the median is the 4th item because it has 3 data points on each side. In case the number is even like 8, then the median is the average of 4th and 5th data point. Median is used where there are outliers i.e. big numbers that impact the mean giving a false picture of the data involved.
* **Mode:** This is the value of the most frequently occurring item in the data set. This is the value of the most expected number to occur.

**Measures of Dispersion:** The degree of spread determines the probability and the level of confidence that one can have on the results obtained from the measures of central tendency. Common measures of dispersion are as follows:

* **Range:** The two endpoints between which all the values of a data set fall is called a range. It is important because it exhaustively includes all the possibilities.
* **Quartiles:** The data set is divided into 4 sets and the number of elements is each set is studied to give us data about quartiles. Similar measures include the deciles and the percentiles. However quartiles remain most widely used.
* **Standard Deviation:** A complex formula is used to work out standard deviation of a given set of data. However standard deviation is like the mean, it is the most important measure of dispersion and is used exhaustively in almost every statistical technique.

Finally, coming on the types of **Data Sets,**we define them into *three*categories namely, **Record Data, Graph-based Data,**and **Ordered Data.**Let’s have a look at them one at a time.

***Record Data***



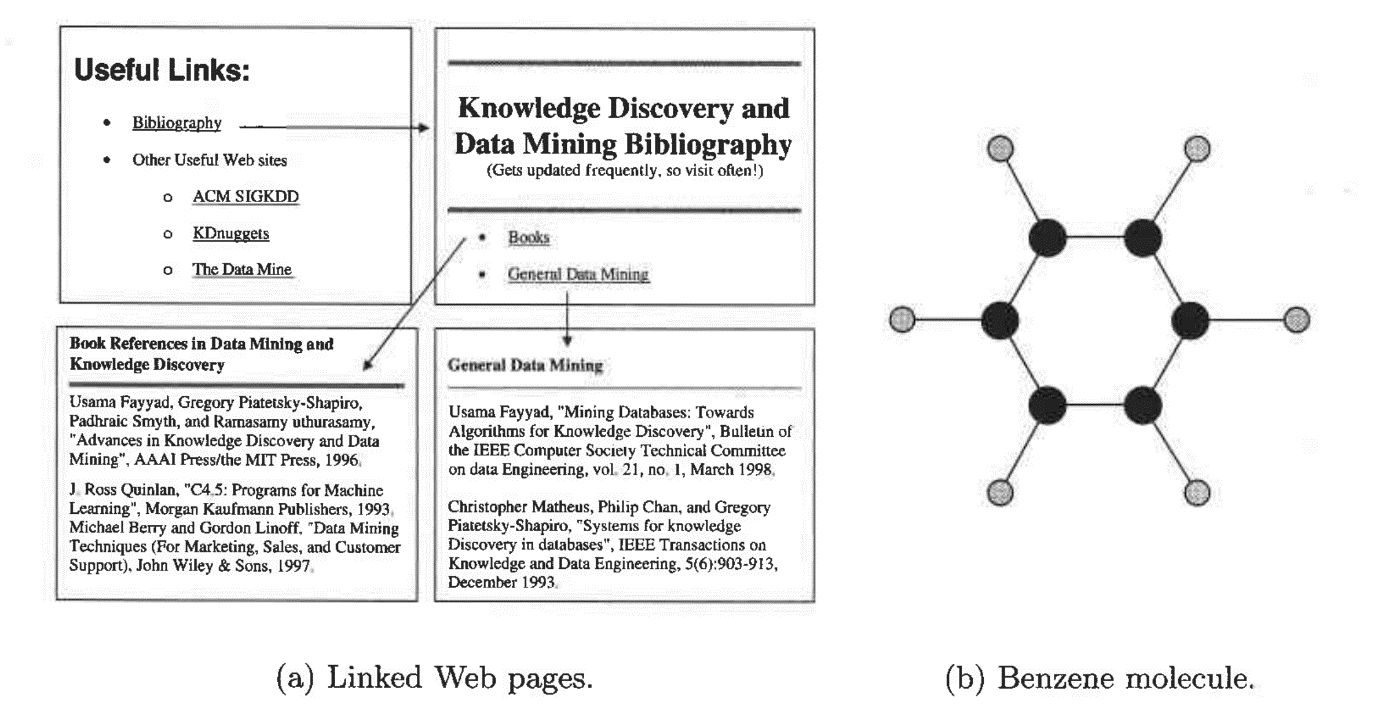
→ Majority of *Data Mining*work assumes that data is a collection of records (data objects).

→ The most basic form of record data has no explicit relationship among records or data fields, and every record (object) has the same set of attributes. Record data is usually stored either in flat files or in relational databases.

There are a few variations of **Record Data,**which have some characteristic properties.

1. **Transaction or Market Basket Data:**It is a special type of record data, in which each record contains a set of items. For example, shopping in a supermarket or a grocery store. For any particular customer, a record will contain a set of items purchased by the customer in that respective visit to the supermarket or the grocery store. This type of data is called **Market Basket Data.**Transaction data is a collection of sets of items, but it can be viewed as a set of records whose fields are asymmetric attributes. Most often, the attributes are binary, indicating whether or not an item was purchased or not.
2. **The Data Matrix:** If the data objects in a collection of data all have the same fixed set of numeric attributes, then the data objects can be thought of as points (vectors)in a multidimensional space, where each dimension represents a distinct attribute describing the object. A set of such data objects can be interpreted as an m X n matrix, where there are n rows, one for each object, and n columns, one for each attribute. Standard matrix operation can be applied to transform and manipulate the data. Therefore, the data matrix is the standard data format for most statistical data.
3. **The Sparse Data Matrix:** A sparse data matrix (sometimes also called **document-data matrix**)is a special case of a data matrix in which the attributes are of the same type and are asymmetric; i.e., only non-zero values are important.

***Graph-based Data***

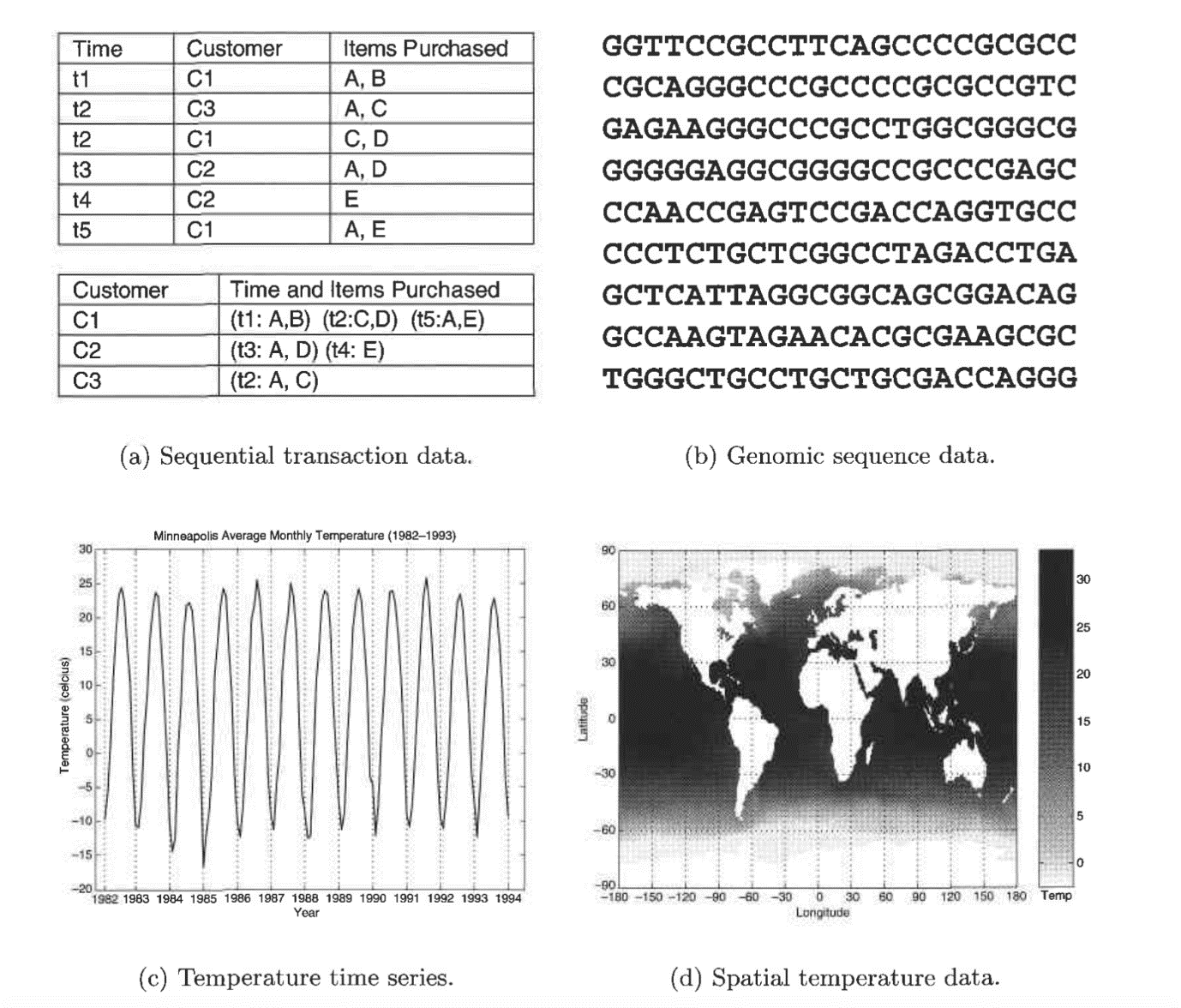


Introduction to Data Mining — Pang-Ning Tan, Michael Steinbach, Vipin Kumar

This can be further divided into types:

1. **Data with Relationships among Objects:** The data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links between objects and link properties, such as direction and weight. Consider Web pages on the World Wide Web, which contain both text and links to other pages. In order to process search queries, Web search engines collect and process Web pages to extract their contents.
2. **Data with Objects That Are Graphs:** If objects have structure, that is, the objects contain sub objects that have relationships, then such objects are frequently represented as graphs. For example, the structure of chemical compounds can be represented by a graph, where the nodes are atoms and the links between nodes are chemical bonds.

***Ordered Data***



Introduction to Data Mining — Pang-Ning Tan, Michael Steinbach, Vipin Kumar

For some types of data, the attributes have relationships that involve order in time or space. As you can see in the picture above, it can be segregated into *four*types:

1. **Sequential Data:**Also referred to as **temporal** data, can be thought of as an extension of record data, where each record has a time associated with it. Consider a retail transaction data set that also stores the time at which the transaction took place
2. **Sequence Data:** Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters. It is quite similar to sequential data, except that there are no time stamps; instead, there are positions in an ordered sequence. For example, the genetic information of plants and animals can be represented in the form of sequences of nucleotides that are known as genes.
3. **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 over time. For example, a financial data set might contain objects that are time series of the daily prices of various stocks.
4. **Spatial Data:** Some objects have spatial attributes, such as positions or areas, as well as other types of attributes. An example of spatial data is weather data (precipitation, temperature, pressure) that is collected for a variety of geographical locations.

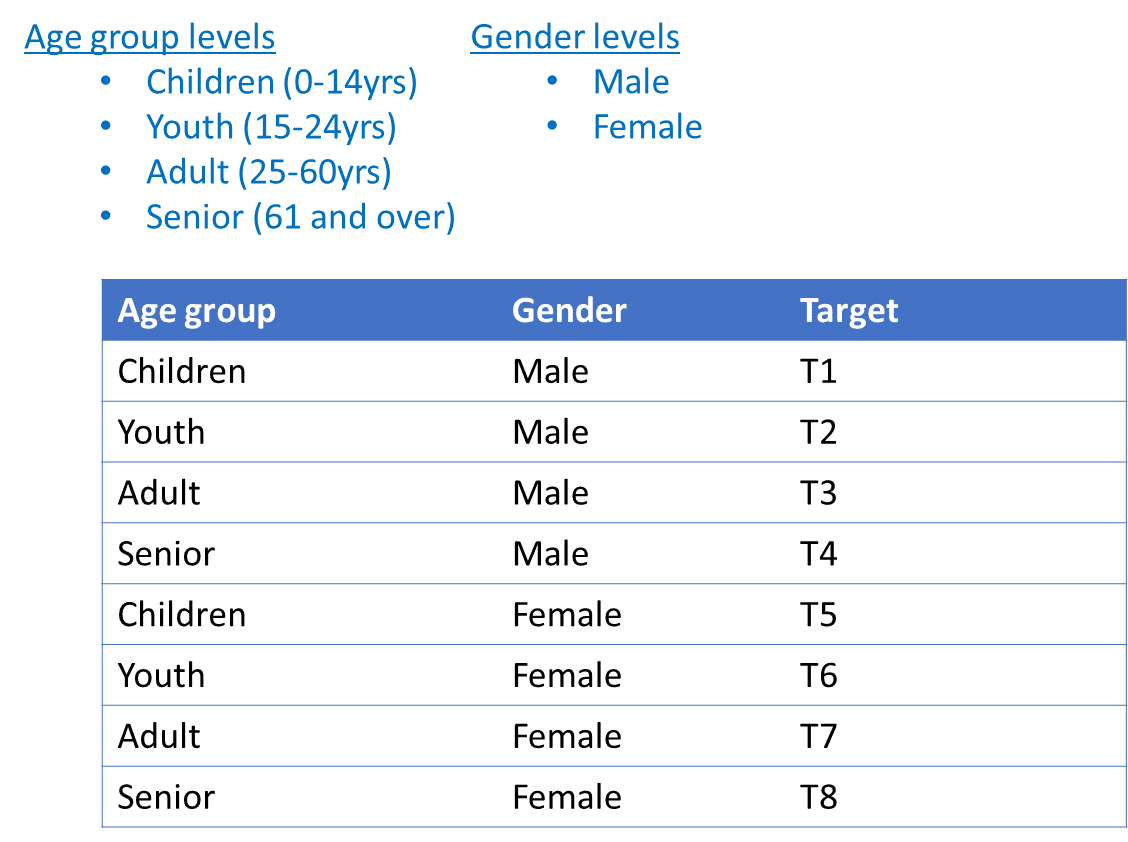
Additional notes: **Characteristics of Data**

Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of attributes/features that exist in a dataset. A dataset with a large number of attributes, generally of the order of hundred or more, is referred to as high dimensional data. Some of the difficulties that come with high dimensional data manifest during analyzing or visualizing the data to identify patterns, and some manifest while training machine learning models. The difficulties related to training machine learning models due to high dimensional data is referred to as ‘**Curse of Dimensionality**’. The popular aspects of curse of dimensionality; ‘data sparsity’ and ‘distance concentration’ are discussed in the following sections.

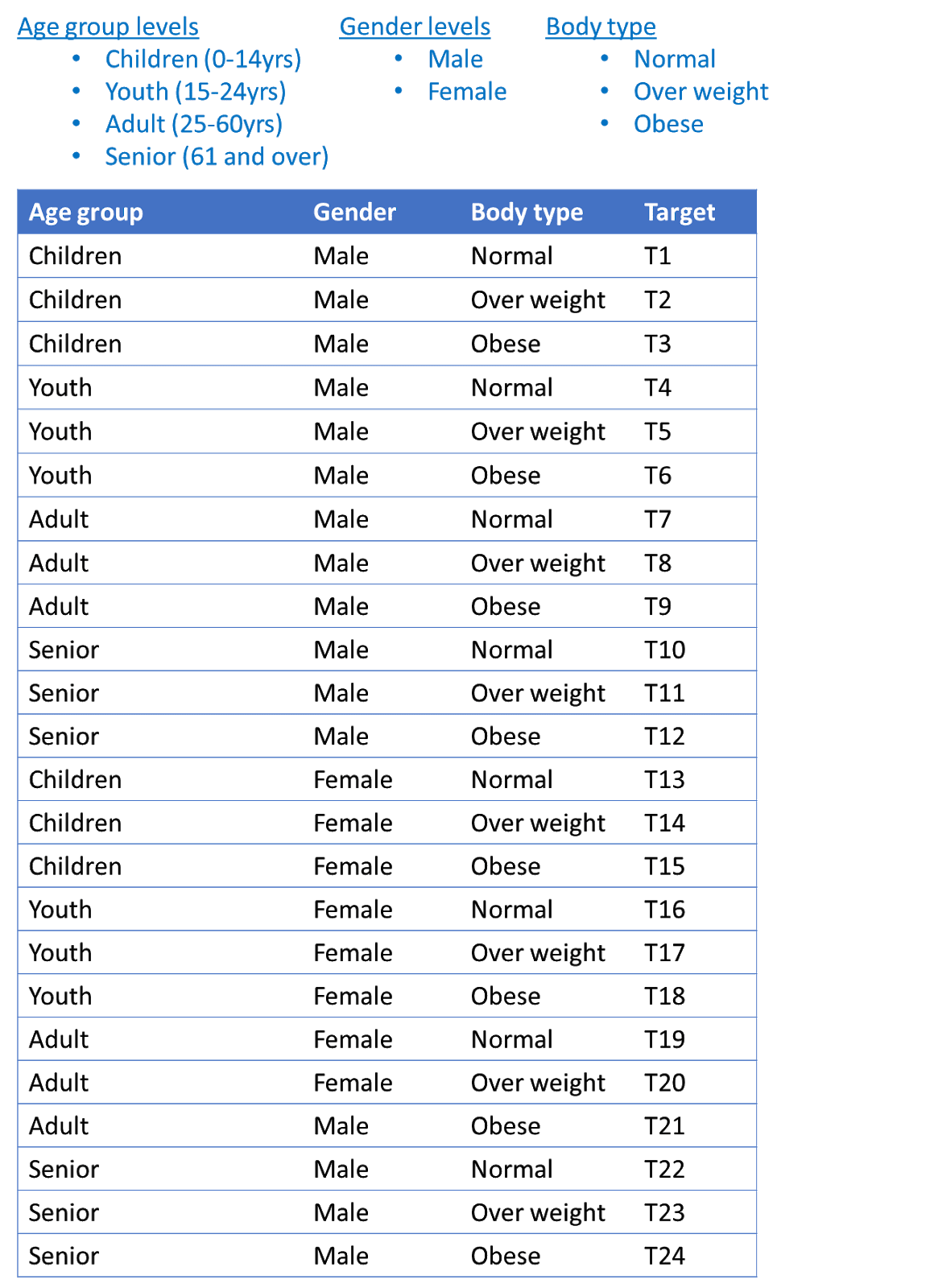
**Data Sparsity**

The supervised [machine learning](https://www.mygreatlearning.com/blog/machine-learning-tutorial-for-complete-beginners/) models are trained to predict the outcome for a given input data sample accurately. While training a model, the available data is used such that part of the data is used for training the model, and a part of the data is used to evaluate how the model performs on unseen data. This evaluation step helps us establish whether the model is generalized or not. Model generalization refers to the models’ ability to predict the outcome for an unseen input data accurately. It is important to note that the unseen input data has to come from the same distribution as the one used to train the model. A generalized model’s prediction accuracy on the unseen data should be very close to its accuracy on the training data. An effective way to build a generalized model is to capture different possible combinations of the values of predictor variables and the corresponding targets.

For instance, if we are trying to predict a target, that is dependent on two attributes: gender and age group, we should ideally capture the targets for all possible combinations of values for the two attributes as shown in figure 1. If this data is used to train a model that is capable of learning the mapping between the attribute values and the target, its performance could be generalized. As long as the future unseen data comes from this distribution (a combination of values), the model would predict the target accurately.

**Figure 1. Combination of values of 2 attributes for generalizing a model**

In the above example, we assume that the target value depends on gender and age group only. If the target depends on a third attribute, let’s say body type, the number of training samples required to cover all the combinations increases phenomenally. The combinations are shown in figure 2. For two variables, we needed eight training samples. For three variables, we need 24 samples.

**Figure 2. Combination of values of 3 attributes for generalizing a model**

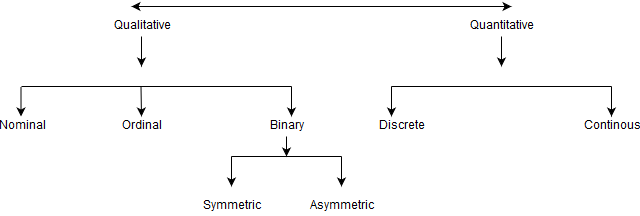
The above examples show that, as the number of attributes or the dimensions increases, the number of training samples required to generalize a model also increase phenomenally.

In reality, the available training samples may not have observed targets for all combinations of the attributes. This is because some combination occurs more often than others. Due to this, the training samples available for building the model may not capture all possible combinations. This aspect, where the training samples do not capture all combinations, is referred to as ‘**Data** **sparsity**’ or simply ‘**sparsity’** in high dimensional data. Data sparsity is one of the facets of the curse of dimensionality. Training a model with sparse data could lead to high-variance or overfitting condition. This is because while training the model, the model has learnt from the frequently occurring combinations of the attributes and can predict the outcome accurately. In real-time when less frequently occurring combinations are fed to the model, it may not predict the outcome accurately.

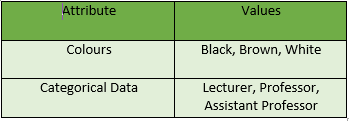
**Attribute**

It can be seen as a data field that represents characteristics or features of a data object. For a customer object attributes can be customer Id, address etc. We can say that a **set of attributes used to describe a given object are known as attribute vector or feature vector.**

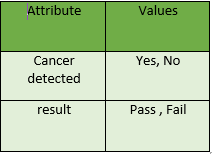
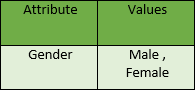
**Type of attributes :**  
1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).  
2. Quantitative (Discrete, Continuous)

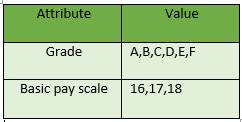


1. **Nominal Attributes – related to names :**The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represents some category or state and that’s why nominal attribute also referred as **categorical attributes** and there is no order (rank, position) among values of nominal attribute.

Example :  


1. **Binary Attributes :** Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false.  
   i)**Symmetric :** Both values are equally important (Gender).  
   ii)**Asymmetric :** Both values are not equally important (Result).

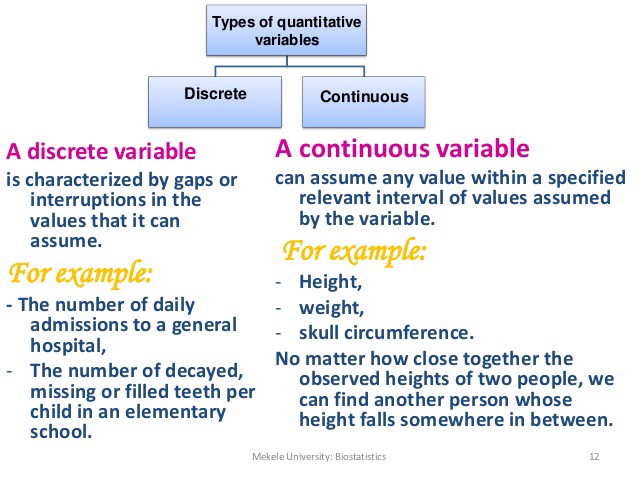


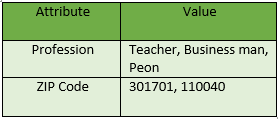
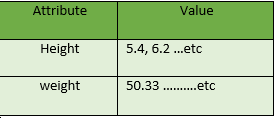
1. **Ordinal Attributes :**The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don’t indicate how important it is.  
   

**Quantitative Attributes**

1. **Numeric :** A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval** and **ratio**.  
   i) An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point or we can call zero point. Data can be added and subtracted at interval scale but can not be multiplied or divided. Consider a example of temperature in degrees Centigrade. If a days temperature of one day is twice than the other day we cannot say that one day is twice as hot as another day.

ii) A**ratio-scaled** attribute is a numeric attribute with an fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range and Five number summary can be given.



1. **Discrete/Categorized :**Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countably infinite set of values.  
   Example  
   
2. **Continuous** : Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.  
   Example :  
   

**Different types of attributes in a data mining data set are:**

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(=,≠)
   * Examples: zip codes, employees ID numbers.
2. Ordinal:
   * The values of an ordinal attribute provide enough information to order objects(<, >)
   * Examples: Hardness of minerals, street numbers
3. Interval:
   * For interval attributes, the differences between values are meaningful,i.e. a unit of measurement exists(+,-)
   * Examples: Calendar dates, Temperature in Celsius or Fahrenheit.
4. Ratio:
   * For ration variables, both differences and ratios are meaningful(\*,/)
   * Examples: Temperature in Kelvin, counts, age.