Unit 1

Data Types & Collection

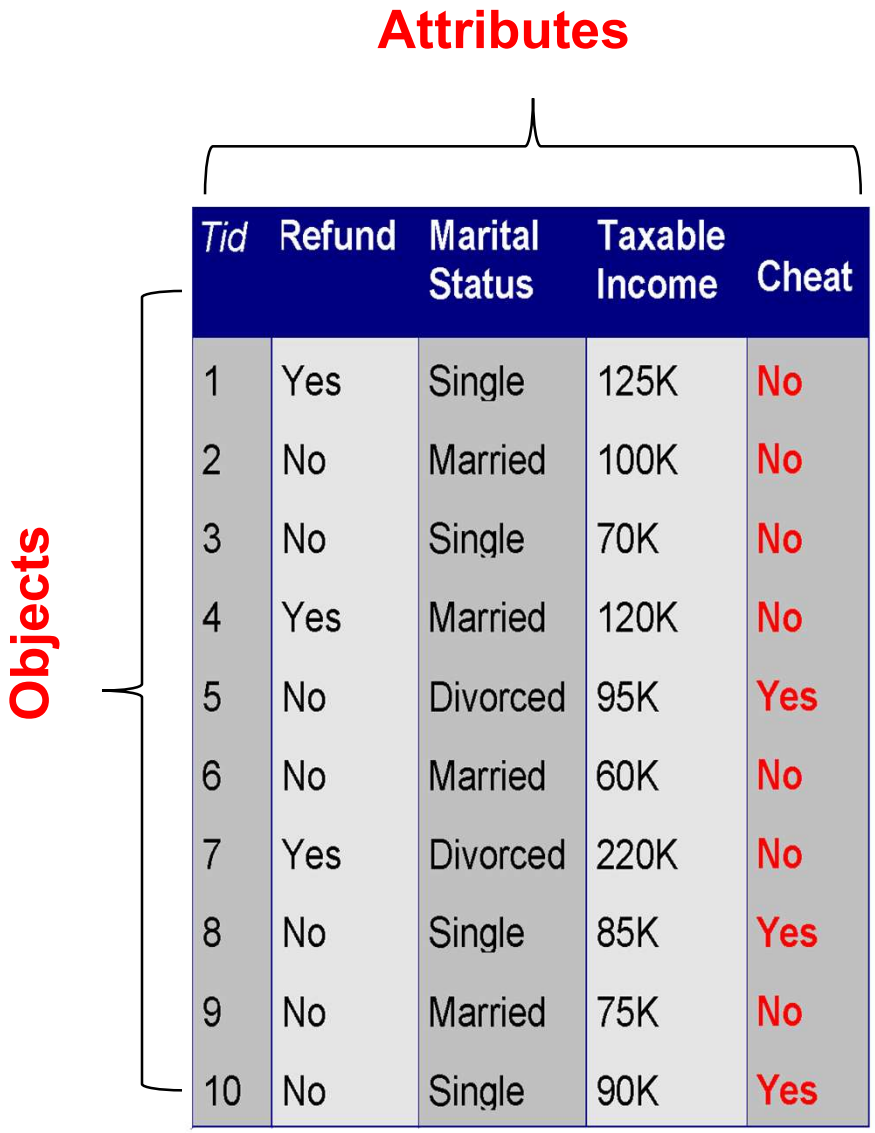
UNIT-I DATA TYPES & COLLECTION: Types of Data: Attributes and Measurement, What is an Attribute?, The Type of an Attribute, The Different Types of Attributes, Describing Attributes by the Number of Values, Asymmetric Attributes, Binary Attribute (Pg.No:22-29, Text Book-1), Nominal Attributes, Binary Attributes, Ordinal Attributes, Numeric Attributes, Discrete versus Continuous Attributes (Pg. No. 39-44, Text-2), Types of Data Sets, General Characteristics of Data Sets, Record Data, Transaction or Market Basket Data, The Data Matrix, The Sparse Data Matrix, Graph Based Data, Graph-Based Data, Ordered Data. Handling Non-Record Data, Data Quality, Measurement and Data Collection Issues, Precision, Bias and Accuracy. (Pg. No. 29-39, Text-1)

Out line

- Types of Data
  - Attributes & Measurements
  - Types of Data Sets
- Data Quality
  - Data Measurement and Data Collection Issues

What is Data?

- Collection of ***Data Objects*** and their ***Attributes***
- An ***attribute*** is a property or characteristic of an object that may vary, either from one objeect to another or from one time to another.
  - 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*



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 Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- 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

Measurement

- *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.

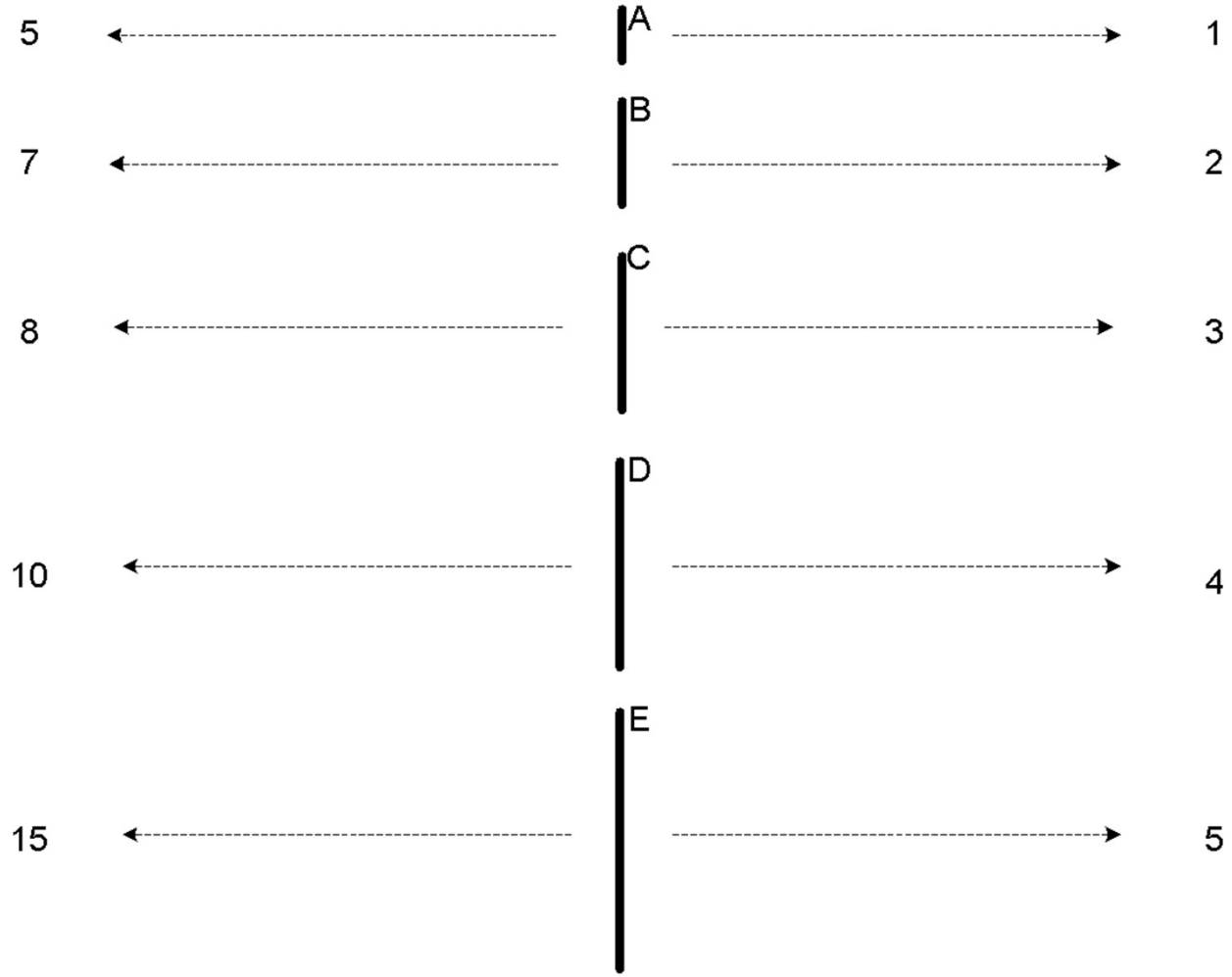
- In all these cases the "physical value" of an attribute of an object is mapped to a numerical or symbolic value.
- With this background, we can now discuss the type of an attribute, a concept that is important in determining if a particular data analysis technique is consistent with a specific type of attribute.

Type of an attribute

- The values used to represent an attribute may have properties that are not properties of the attribute itself, and vice versa
- Example 1: Employee Age and ID Number
- Example 2: Length of Line Segments

Example 2: Length of Line Segments

- The way you measure an attribute may not match the attributes properties.



This scale preserves only the ordering property of length.

This scale preserves the ordering and additvity properties of length.

Properties of Attribute Values

- The type of an attribute depends on which of the following properties/operations it possesses:
  - Distinctness : = and ≠
  - Order : < , <=, > and >=
  - Addition : + and -
    - (Meaningful Differences)
  - Multiplication : * and /
    - (Meaningful Differences)

Types of Attributes

- There are 4 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: temperature in Kelvin, length, time, counts

Nominal Attribute:
Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Gender of the Person.

Ordinal Attribute:
The ordinal attribute value provides sufficient information to order the objects. Such as Rankings, Grades, Height

Interval Scaled attribute:
It is measured on a scale of equal size units, these attributes allows us to compare such as temperature in C or F and thus values of attributes have order.

Ratio Scaled attribute:
Both differences and ratios are significant for Ratio. For eg. age, length, Weight.

- Nominal attribute : distinctness
- Ordinal attribute : distinctness & order
- Interval attribute : distinctness, order & meaningful differences
- Ratio attribute : all 4 properties/operations

Difference Between Ratio and Interval

- Is it physically meaningful to say that a temperature of 10 ° is twice that of 5° on
  - the Celsius scale?
  - the Fahrenheit scale?
  - the Kelvin scale?

- Consider measuring the height above average
  - If Bill's height is three inches above average and Bob's height is six inches above average, then would we say that Bob is twice as tall as Bill?
  - Is this situation analogous to that of temperature?

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, 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, *t* and *F* 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

	Attribute Type	Transformation	Comments
Categorical Qualitative	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}.
Numeric Quantitative	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 as important
  - Words present in documents
  - Items present in customer transactions

- If we met a friend in the grocery store would we ever say the following?

  "I see our purchases are very similar since we didn't buy most of the same things."

- We need two asymmetric binary attributes to represent one ordinary binary attribute
  - Association analysis uses asymmetric attributes

- Asymmetric attributes typically arise from objects that are sets

Critiques …

- Not a good guide for statistical analysis
  - May unnecessarily restrict operations and results
    - Statistical analysis is often approximate
    - Thus, for example, using interval analysis for ordinal values may be justified
  - Transformations are common but don't preserve scales
    - Can transform data to a new scale with better statistical properties
    - Many statistical analyses depend only on the distribution

More Complicated Examples

- ID numbers
  - Nominal, ordinal, or interval?
- Number of cylinders in an automobile engine
  - Nominal, ordinal, or ratio?
- Biased Scale
  - Interval or Ratio

Key Messages for Attribute Types

- The types of operations you choose should be "meaningful" for the type of data you have
  - Distinctness, order, meaningful intervals, and meaningful ratios are only four properties of data
  - The data type you see – often numbers or strings – may not capture all the properties or may suggest properties that are not there
  - Analysis may depend on these other properties of the data
    - Many statistical analyses depend only on the distribution
  - Many times what is meaningful is measured by statistical significance
  - But in the end, what is meaningful is measured by the domain

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 Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

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

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

Graph Based Data

1, Data with Relationships among objects

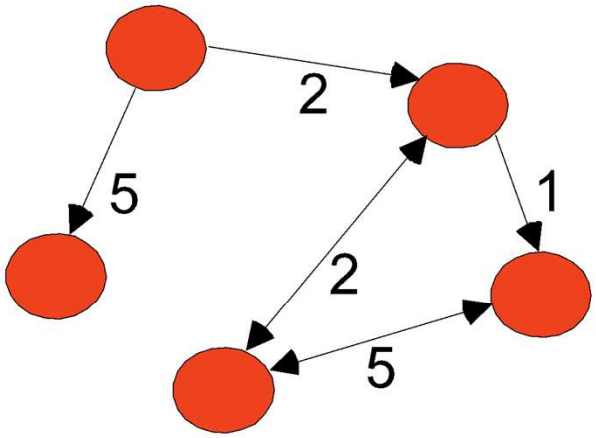
ex: Web pages

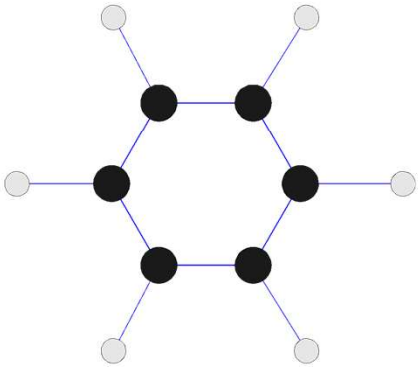
2, Data with Objects that are Graphs

ex: Molecule

Graph Data

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6

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.

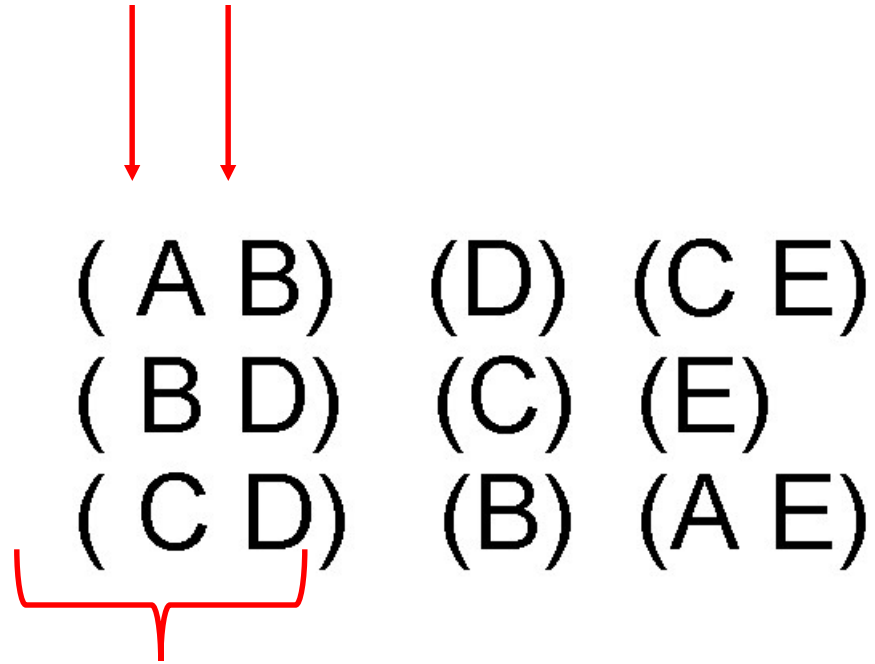
Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

1. Sequential Data
2. Sequence Data
3. Time Series Data
4. Spatial Data

Ordered Data

- Sequential Data : Also referred as *temporal data*
- Sequences of transactions



Ordered Data

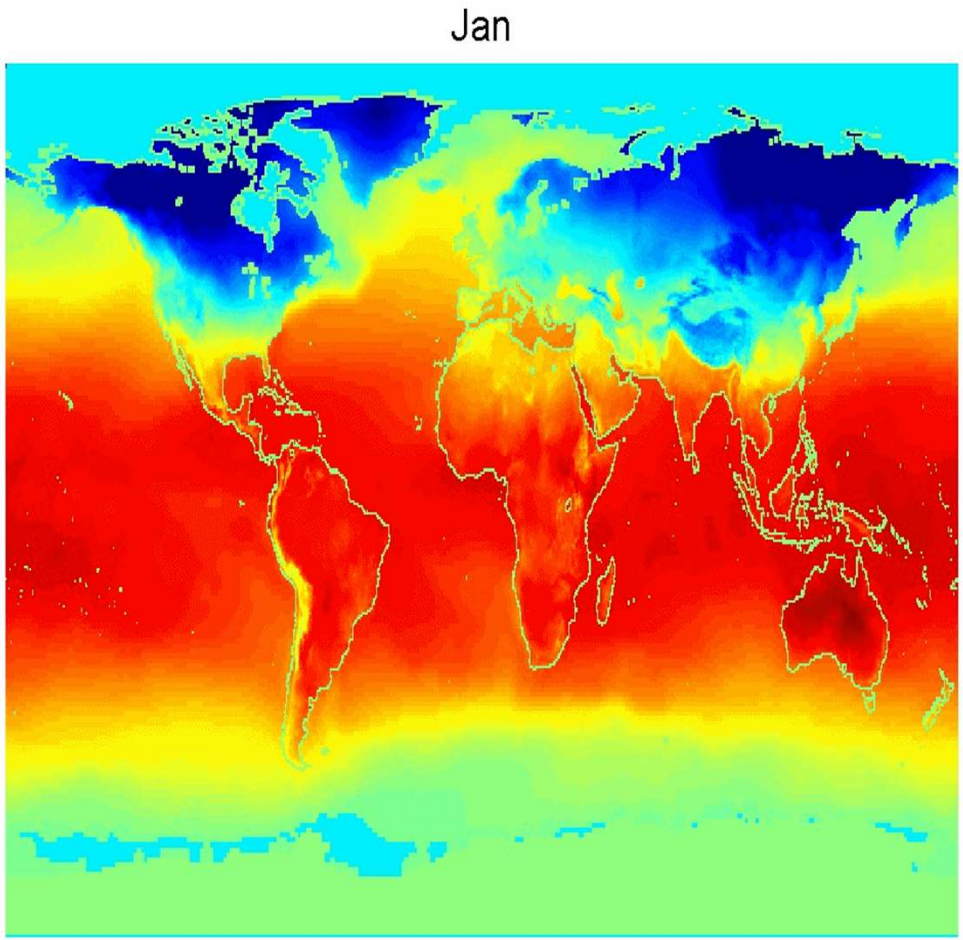
- Sequence Data: Ex: Genome sequencing

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
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

"The most important point is that poor data quality is an unfolding disaster.

  - Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

Thomas C. Redman, DM Review, August 2004

- 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

Data Quality …

- 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
  - Missing values
  - Duplicate data
  - Wrong data

Measurement and Data Collection Errors

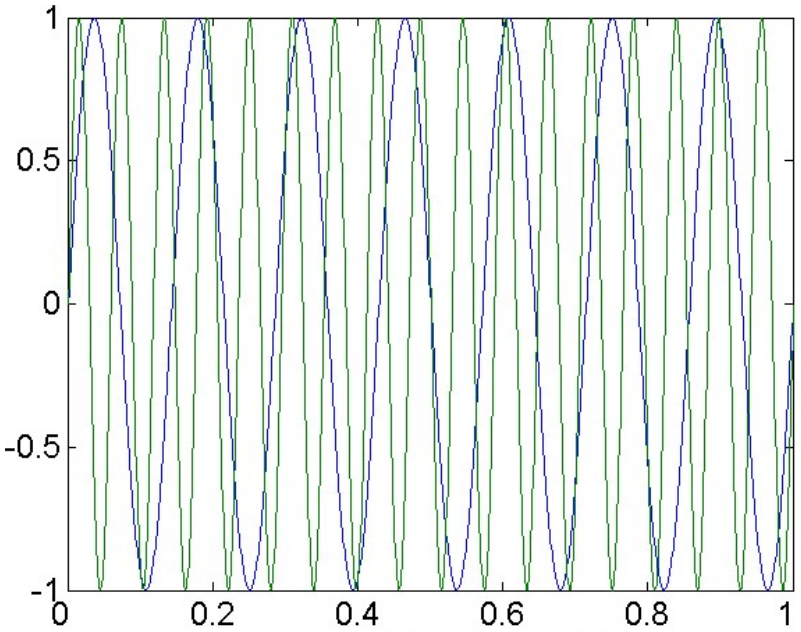
- The term Measurement Error refers to any problem resulting from the measurement process.
- A common problem is that the value recorded differs from the true value to some extent.
- For continuous attributes, the numerical difference of the measured and true value is called the *error*.
- The term Data Collection Error refers to errors such as omitting data objects or attribute values, or inappropriately including a data object.
- For example, a study of animals of a certain species might include animals of a related species that are similar in appearance to the species of interest.

Noise and Artifacts

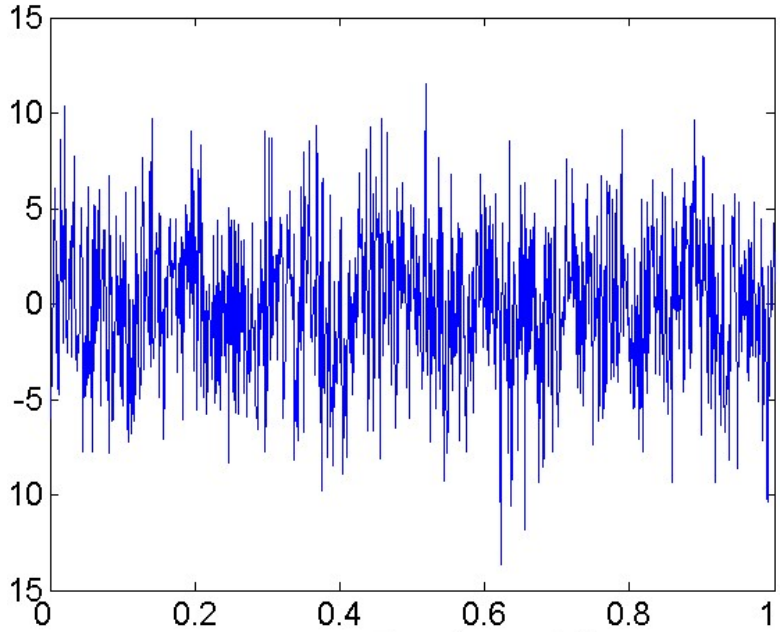
- Noise is the random component of a measurement error. It may involve the distortion of a value or the addition of spurious objects.
- The term noise is often used in connection with data that has a spatial or temporal component.
- In such cases, techniques from signal or image processing can frequently be used to reduce noise and thus, help to discover patterns (signals) that might be "lost in the noise."
- Nonetheless, the elimination of noise is frequently difficult, and much work in data mining focuses on devising robust algorithms that produce acceptable results even when noise is present.

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



Two Sine Waves



Two Sine Waves + Noise

- Data errors may be the result of a more deterministic phenomenon, such as a streak in the same place on a set of photographs. Such deterministic distortions of the data are often referred to as *artifacts*.
- A **data artifact** is a **data** flaw caused by equipment, techniques or conditions.
- Common sources of **data** flaws include hardware or software errors, conditions such as electromagnetic interference and flawed designs such as an algorithm prone to miscalculations

Precision, Bias, and Accuracy

Precision: The closeness of repeated measurements (of the same quantity) to one another.

Precision is often measured by the standard deviation of a set of values

Bias: A systematic quantity being measured.

Bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured.

Bias can only be determined for objects whose measured quantity is known by means external to the current situation

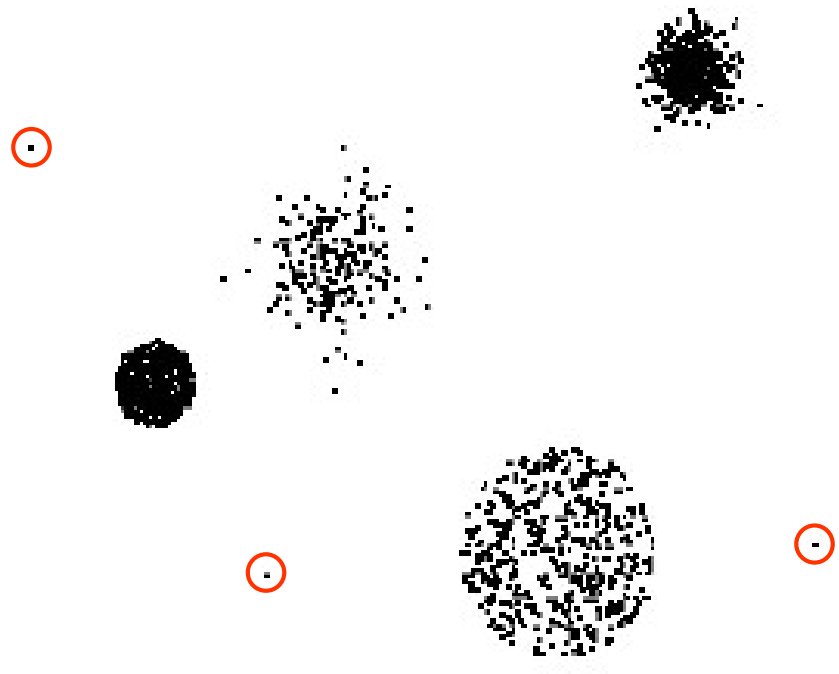
- For example: Suppose that we have a standard laboratory weight with a mass of 1g and want to assess the precision and bias of our new laboratory scale.
- We weigh the mass five times, and obtain the following five values: {1.015,0.990, 1.013, 1.001,0.986}.
- The mean of these values is 1.001, and hence, the bias is 0.001. The precision, as measured by the standard deviation, is 0.013.
- Accuracy: The closeness of measurements to the true value of the quantity being measured.
- Accuracy depends on precision and bias, but since it is a general concept, there is no specific formula for accuracy in terms of these two quantities

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.
- It is important to distinguish between the notions of noise and outliers.
- Outliers can be legitimate data objects or values. Thus, unlike noise, outliers may sometimes be of interest.
- In fraud and network intrusion detection, for example, the goal is to find unusual objects or events from among a large number of normal ones.

Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection
- Causes?



Missing Values

- It is not unusual for an object to be missing one or more attribute values.
- In some cases, the information was not collected; e.g., some people decline to give their age or weight.
- In other cases, some attributes are not applicable to all objects; e.g., often, forms have conditional parts that are filled out only when a person answers a previous question in a certain way, but for simplicity, all fields are stored.
- Regardless, missing values should be taken into account during the data analysis.

Eliminate Data Objects or Attributes

- A simple and effective strategy is to eliminate objects with missing values.
- However, even a partially specified data object contains some information, and if many objects have missing values, then a reliable analysis can be difficult or impossible.
- Nonetheless, if a data set has only a few objects that have missing values, then it may be expedient to omit them.
- A related strategy is to eliminate attributes that have missing values. This should be done with caution, however, since the eliminated attributes may be the ones that are critical to the analysis.

Estimate Missing Values

- Sometimes missing data can be reliably estimated.
- For example, consider a time series that changes in a reasonably smooth fashion, but has a few, widely scattered missing values.
- In such cases, the missing values can be estimated (interpolated) by using the remaining values.
- As another example, consider a data set that has many similar data points. In this situation, the attribute values of the points closest to the point with the missing value are often used to estimate the missing value.

- If the attribute is continuous, then the average attribute value of the nearest neighbors is used.
- If the attribute is categorical, then the most commonly occurring attribute value can be taken.
- For a concrete illustration, consider precipitation measurements that are recorded by ground stations. For areas not containing a ground station, the precipitation can be estimated using values observed at nearby ground stations..

Inconsistent Values

- Data can contain inconsistent values. Consider an address field, where both a zip code and city are listed, but the specified zip code area is not contained in that city.
- It may be that the individual entering this information transposed two digits, or perhaps a digit was misread when the information was scanned from a handwritten form.
- Some types of inconsistencies are easy to detect. For instance, a person's height should not be negative.
- In other cases, it can be necessary to consult an external source of information.

- For example, when an insurance company processes claims for reimbursement, it checks the names and addresses on the reimbursement forms against a database of its customers.
- A product code may have "check" digits, or it may be possible to double-check a product code against a list of known product codes, and then correct the code if it is incorrect, but close to a known code. The correction of an inconsistency requires additional or redundant information.

Duplicate Data

- A data set may 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 the values of corresponding attributes may differ, 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.
- 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 may still cause problems for some algorithms if the possibility of identical objects is not specifically accounted for in their design