L 1-Sources and nature of data sets Attributes types

What is Data?

- 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

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, shirt color etc.
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short},shirt size in {small, medium, large}
  - Interval(equal distance between the values)

    It is measured along a scale, in which each point is placed at an equal distance from one another
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit etc.
  - Ratio(any fraction)
    - Examples: shirt price, length, weight etc.

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

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*}	
Ordinal	The values of an ordinal attribute provide enough information to order objects (< >).	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	Monetary quantities, mass, length, electrical current	

Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another. Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal.

Interval data cannot be multiplied or divided, however, it can be added or subtracted. Interval data is measured on an interval scale. A simple example of interval data: The difference between 100 degrees Fahrenheit and 90 degrees Fahrenheit is the same as 60 degrees Fahrenheit and 70 degrees Fahrenheit.

The difference between interval and ratio data is simple. Ratio data has a defined zero point. Income, height, weight, annual sales, market share, product defect rates, time to repurchase, unemployment rate, and crime rate are examples of ratio data.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably 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.

Types of data sets

- **Record**
  - **Data Matrix**
  - **Document Data**
  - **Transaction Data**
- **Graph**
  - **World Wide Web**
- **Ordered**
  - **Genetic Sequence 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

Patient	Pain	Fever	Running nose	Fatigueness
XYZ	Y	Y	Y	Y
MNO	N	Y	N	Y

Text 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										
Document 2										

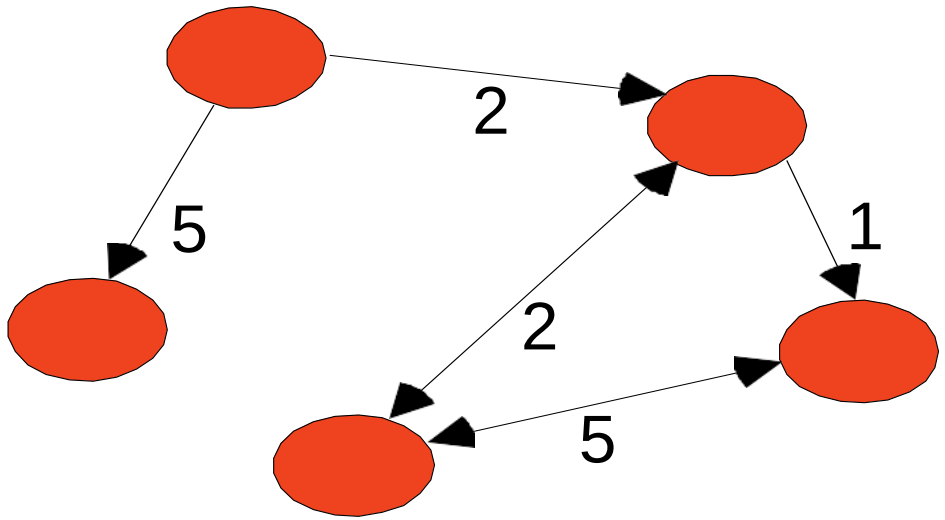
Transaction Data

- 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 Data

- Examples: Facebook graph and HTML Links



Ordered Data

- Genomic sequence data

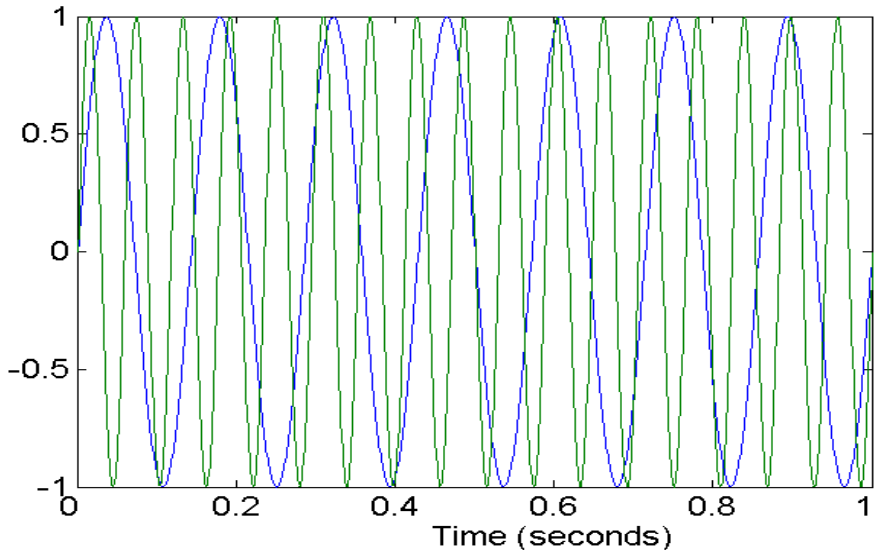
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Data Quality

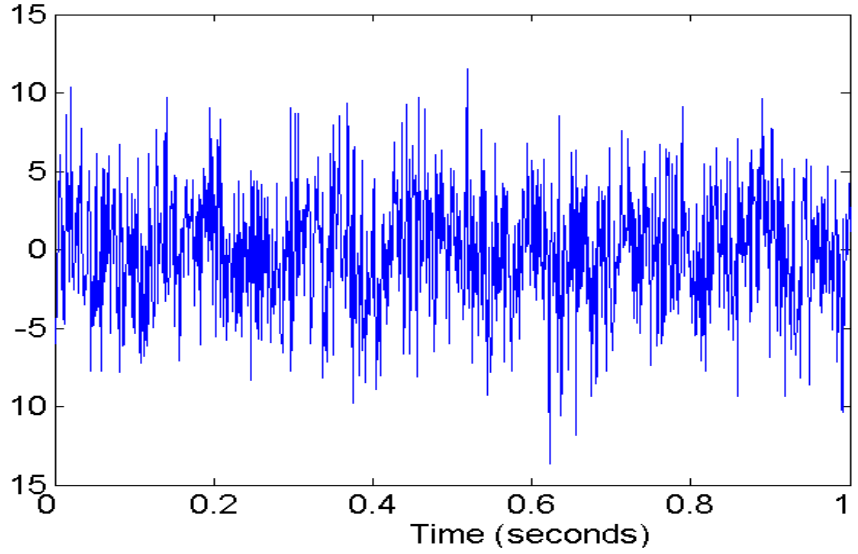
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- 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



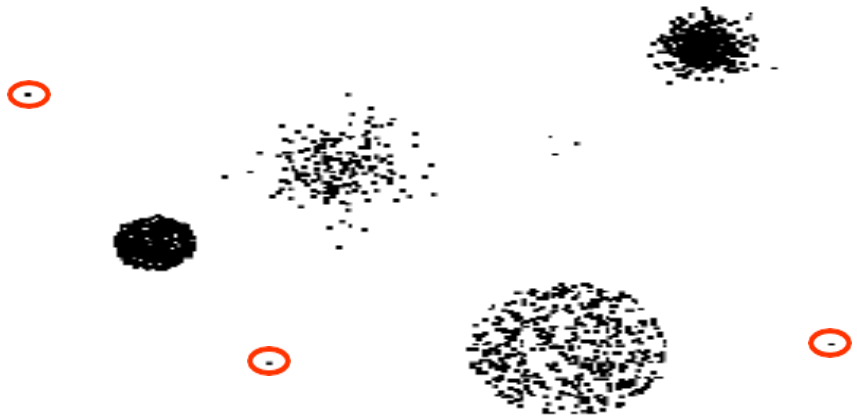
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



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)

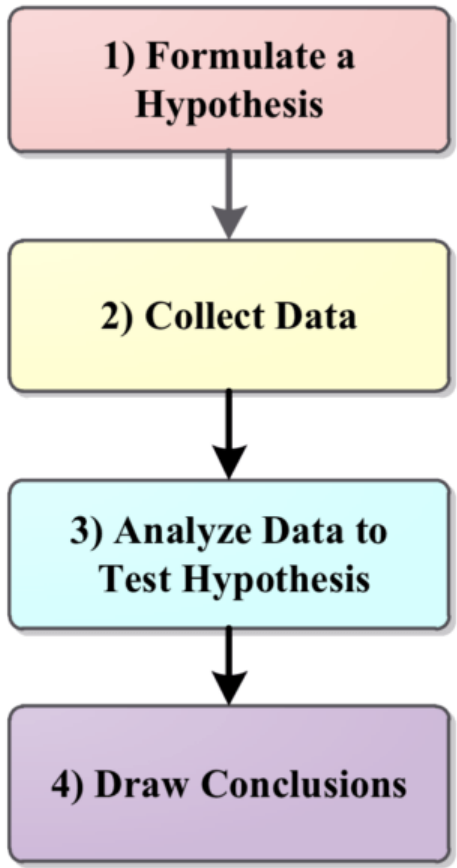
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data or Inconsistency

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

Difference between Data Analytics and Data Mining:

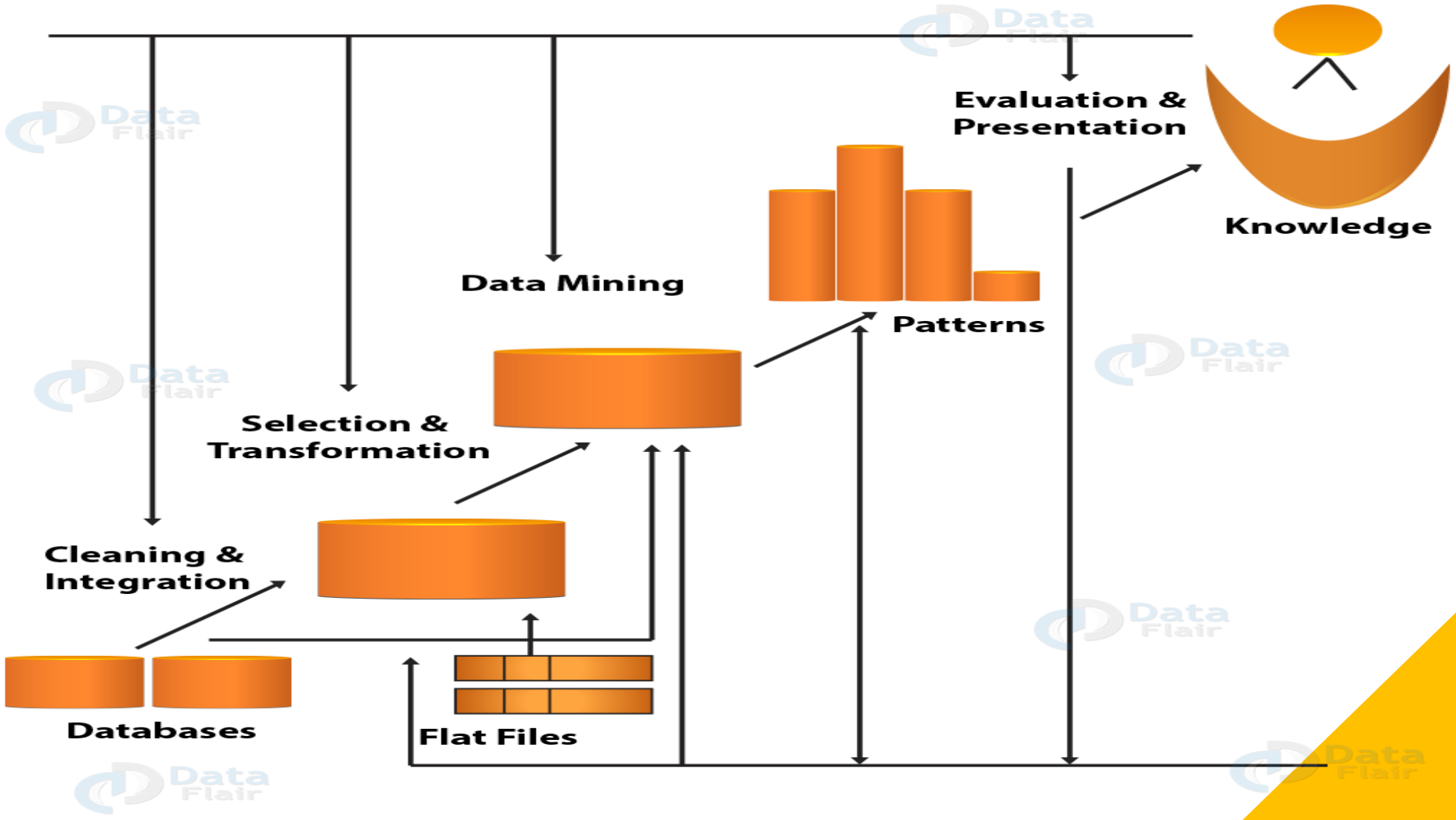
A key difference between data analytics and data mining is that data mining does not require any preconceived hypothesis or notions before tackling the data. It simply compiles it into useful formats. However, data analysis does need a hypothesis to test, as it is looking for answers to particular questions.


1) Formulate a Hypothesis
2) Collect Data
3) Analyze Data to Test Hypothesis
4) Draw Conclusions

Data mining can be undertaken by a single specialist with excellent technological skills. With the right software, they are able to collect the data ready for further analysis. At this stage, a larger team simply isn't required. From here, a data mining specialist will usually report their findings to the client, leaving the next steps in someone else's hands.

However, when it comes to data analytics, a team of specialists may be needed. They need to assess the data, figure out

Knowledge Discovery Database(KDD)



- **Data Integration**

First of all the data is collected and integrated from all the different sources.

- **Data Selection**

Generally, we may not all the data we have collected in the first step. Also, in this step, we select only those data which we think useful for data mining.

- **Data Cleaning**

Generally, the data we have collected is not clean and may contain errors, missing values, noisy or inconsistent data. Therefore we need to apply different techniques to get rid of such anomalies.

- **Data Transformation**

Basically, the data even after cleaning is not ready for mining. Also, we need to transform them into forms appropriate for mining. Thus, the techniques used to do this are smoothing, aggregation, normalization etc.

- **Data Mining**

As now in this step, we are ready to apply **data mining techniques** on the data. Basically, it is to discover the interesting patterns. Hence, clustering and association analysis are among the many different techniques present. Also, as we used for data mining.

- **Pattern Evaluation and Knowledge Presentation**

Generally, this step includes visualization, transformation, removing redundant patterns from the patterns we generated.

- **Decisions / Use of Discovered Knowledge**

As this step is beneficial to us. Also, it helps to use the knowledge acquired to take better decisions.

patterns, and draw conclusions. They may use machine learning to help with the processing, but this still has a human element involved. Data analytics teams need to know the right questions to ask –

e.g. relation between reader gender and English paper

e.g. relation between trained and trained students versus committing errors

e.g. inoculated with vaccine, not inoculated vaccine versus died of disease , survived

Data mining usually does not need any visualizations, bar charts, graphs etc., whereas these visualizations /are the bread and butter of data analysis. Without a good representation of the data in question, all the efforts which are put into the analysis of the data would not come to fruition.

Quiz

1. Which of the following is usually the last step in the data mining process?

A. Visualization

B. Preprocessing

C. Modeling

D. Deployment

2. Name of a movie, can be considered as an attribute of type?

A. Nominal

B. Ordinal

C. Interval

D. Ratio

3. User rating given to a movie in a scale 1-10, can be considered as an attribute of type?

A. Nominal

B. Ordinal

C. Interval

D. Ratio

4. Which of the following operations cannot be performed on interval attributes?

A. Distinctness

B. Order

C. Addition

D. Multiplication

5. Which of the following operations can be performed on ratio attributes?

A. Addition

B. Multiplication

C. Both of the above

D. None of the above

6. Sales database of items in a supermarket can be considered as an example of:

A. Record data

B. Tree data

C. Graph data

D. None of the above

7. Rows of a data matrix storing record data usually represents?

A. Metadata

B. Objects

C. Attributes

D. Aggregates

8. Which of the following is an example of continuous attribute?

A. Weight of a person

B. Shoe size of a person

C. Gender of a person

D. None of the above