

Data Analytics

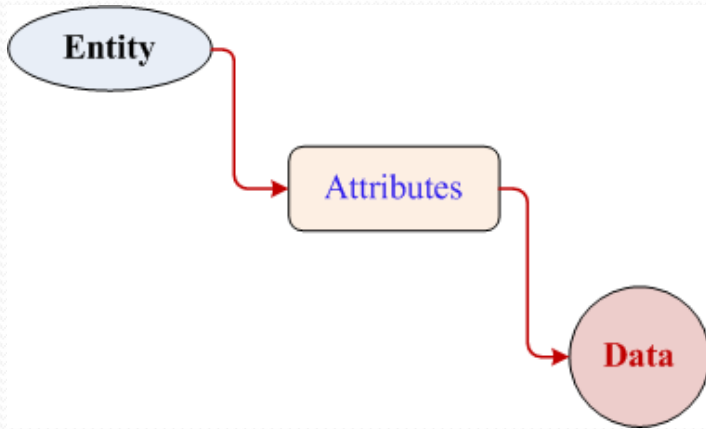
Data Categorization

MSAJCE

Today's discussion...

- Data in data analytics
- NOIR topology
- Nominal scale
 - Binary
 - Symmetric
 - Asymmetric
- Ordinal scale
- Interval and ration scale
- Multidimensional Data Model

Data in Data Analytics



NAME	AGE	GENDER	SALARY	EMPLOYER
:				
:				
ABCD	34	F	40000	XYZ
:				
:				

- **Entity:** A particular thing is called entity or object.
- **Attribute.** An attribute is a measurable or observable property of an entity.
- **Data.** A measurement of an attribute is called data.
- Note
 - Data defines an **entity**.
 - Computer can manage all type of data (e.g., audio, video, text, etc.).

Data in Data Analytics

- In general, there are many types of data that can be used to measure the properties of an entity.
- A good understanding of data **scales** (also called scales of measurement) is important.
- Depending the scales of measurement, different technique are followed to derive hitherto unknown knowledge in the form of
 - patterns, associations, anomalies or similarities from a volume of data.

NOIR

Classification of scales of Measurement

NOIR classification

- The mostly recommended scales of measurement are

N: Nominal

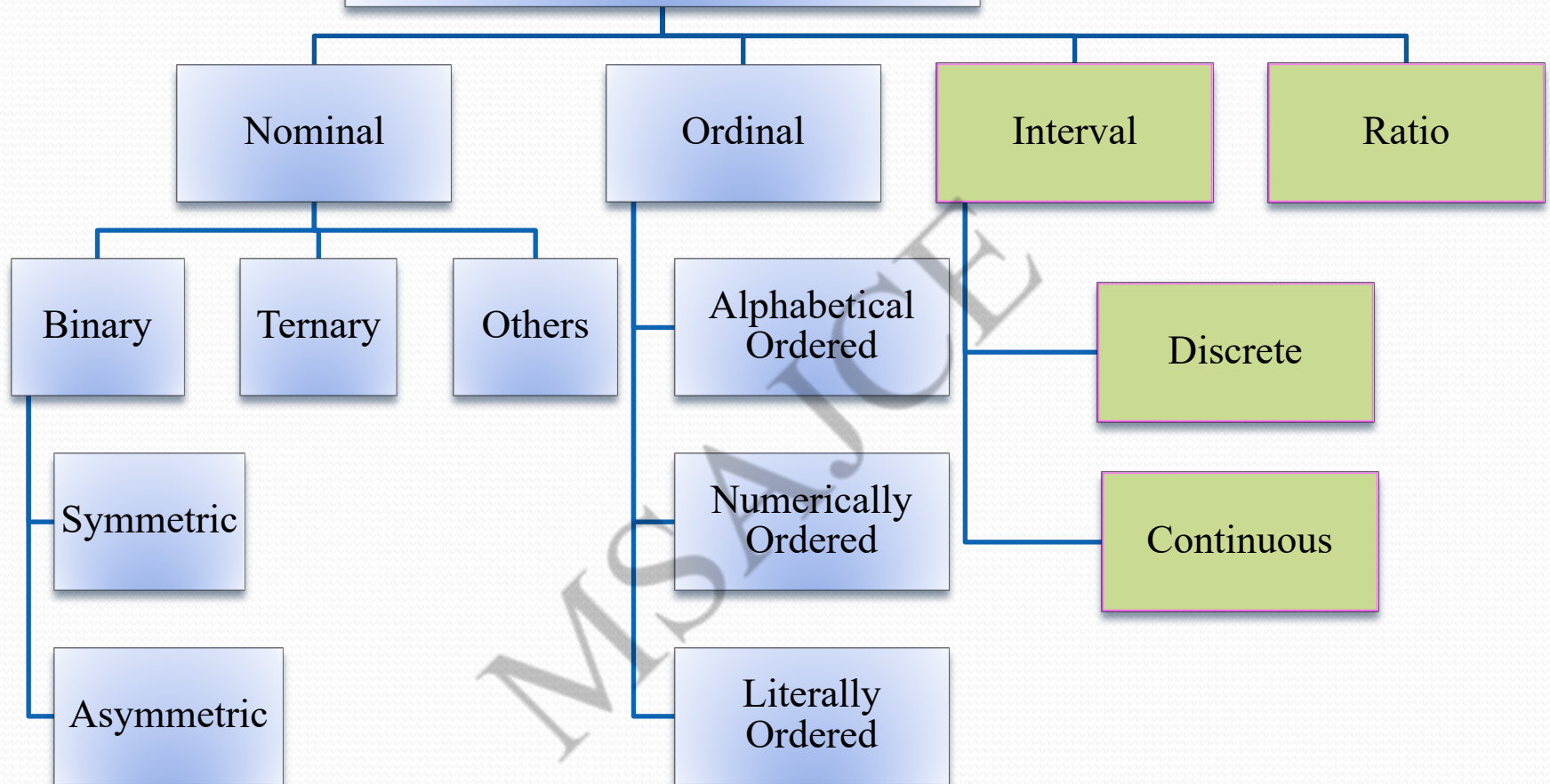
O: Ordinal

I: Interval

R: Ratio

The NOIR scale is the **fundamental building block** on which the **extended data types** are built.

NOIR Classification



Categorical (Qualitative)

**Numeric
(Quantitative)**

Properties of data

- Following FOUR properties (operations) of data are pertinent.

#	Property	Operation	Type
1.	Distinctiveness	= and $\neq$	Categorical (Qualitative)
2.	Order	$<$, $\leq$, $>$, $\geq$	
3.	Addition	+ and -	Numerical (Quantitative)
4.	Multiplication	* and /	

NOIR summary

- ✓ Nominal (with distinctiveness property only)
- ✓ Ordinal (with distinctive and order property only)
- ✓ Interval (with additive property + property of Ordinal data)
- ✓ Ratio (with multiplicative property + property of Interval data)
- Further, nominal and ordinal are collectively referred to as **categorical or qualitative data**. Whereas, interval and ratio data are collectively referred to as **quantitative or numeric data**.

Which of the following employment classifications best describes your area of work?

- 1. Educator
- 2. Construction worker
- 3. Manufacturing worker
- 4. Lawyer
- 5. Doctor
- 6. Other

Nominal scale

- **Definition**

A variable that takes a value **among a set of mutually exclusive codes** that have no logical order is known as a nominal variable.

- **Examples**

Gender	Used letters or numbers { M, F } or { 1, 0 }
--------	--

Blood groups	Used string { A , B , AB , O }
--------------	-----------------------------------

Rhesus (Rh) factors	Used symbols { + , - }
---------------------	---------------------------

Country code	?? ????
--------------	------------

Nominal scale

Note

- The nominal scale is used to label data categorization using **a consistent naming convention**.
- The labels can be numbers, letters, strings, enumerated constants or other keyboard symbols.
- Nominal data thus makes “**category**” of a set of data.
- The number of categories should be two (binary) or more (ternary, etc.), but **countably finite**.

Nominal scale

Note

- A nominal data **may be numerical in form**, but the numerical values have no mathematical interpretation.
 - For example, 10 prisoners are 100, 101, ... 110, but; $100 + 110 = 210$ is meaningless. They are simply labels.
- Two labels **may be identical** (=) or dissimilar ($\neq$).
- These labels **do not have any ordering** among themselves.
 - For example, we cannot say blood group B is better or worse than group A.
- Labels (from two different attributes) **can be combined to** give another nominal variable.
 - For example, blood group with Rh factor (A+ , A- , AB+, etc.)

Binary scale

- **Definition**

A nominal variable with **exactly two mutually exclusive categories** that have **no logical order** is known as binary variable

- **Examples**

Switch: {ON, OFF}

Attendance: {True, False}

Entry: {Yes, No}

etc.

Note

- A Binary variable is a special case of a nominal variable that takes **only two possible** values.

Symmetric and Asymmetric Binary Scale

- Different binary variables may have unequal importance.
- If two choices of a binary variable have **equal importance**, then it is called symmetric binary variable.
 - Example: Gender = {male , female}
// usually of equal probability.
- If the two choices of a binary variable have **unequal importance**, it is called asymmetric binary variable.
 - Example: Food preference = {V , NV}

Operations on Nominal variables

- Summary statistics applicable to nominal data are **mode**, contingency **correlation**, etc.
- Arithmetic (+, -, * and /) and logical operations (<, >, ≠ etc.) **are not permitted**.
- The allowed operations are : accessing (read, check, etc.) and re-coding (into another non-overlapping symbol set, that is, one-to-one mapping) etc.
- Nominal data can be visualized using line charts, bar charts or pie charts etc.
- Two or more nominal variables can be combined to generate other nominal variable.
 - Example: Gender (M,F) × Marital status (S, M, D, W)

Survey

This computer tutorial is

—	—	—	—	—
not	somewhat	moderately	very	extremely
helpful	helpful	helpful	helpful	helpful
1	2	3	4	5

Ordinal scale

- **Definition**

Ordered nominal data are known as ordinal data and the variable that generates it is called ordinal variable.

- Example:

Shirt size = { S, M, L, XL, XXL }

Note

The values assumed by an ordinal variable can be ordered among themselves as each pair of values can be compared literally or using relational operators ($<$, $\leq$, $>$, $\geq$).

Operation on Ordinal data

- Usually relational operators can be used on ordinal data.
- Summary measures **mode** and **median** can be used on ordinal data.
- Ordinal data can be ranked (numerically, alphabetically, etc.) Hence, we can find any of the **percentiles measures** of ordinal data.
- Calculations based on order are permitted (such as count, min, max, etc.).
- Spearman's R can be used as a measure of the strength of association between two sets of ordinal data.
- Numerical variable can be transformed into ordinal variable and vice-versa, but with a loss of information.
 - For example, Age [1, ... 100] = [young, middle-aged, old]

Interval scale

- **Definition**

Interval-scale variables are **continuous measurements** of a **roughly linear scale**.

- Example:
weight, height, latitude, longitude, weather, temperature, calendar dates, etc.

Note

- Interval data are with well-defined interval.
- Interval data are measured on a numeric scale (with +ve, 0 (zero), and -ve values).
- Interval data **has a zero point on origin**. However, the origin does not imply a true absence of the measured characteristics.
 - For example, temperature in Celsius and Fahrenheit; 0° does not mean absence of temperature, that is, no heat!

Operation on Interval data

- We can add to or from interval data.
 - For example: $\text{date}_1 + x\text{-days} = \text{date}_2$
- Subtraction can also be performed.
 - For example: $\text{current date} - \text{date of birth} = \text{age}$
- Negation (changing the sign) and multiplication by a constant are permitted.
- All operations on ordinal data defined are also valid here.
- Linear (e.g. $cx + d$) or Affine transformations are permissible.
- Other one-to-one non-linear transformation (e.g., log, exp, sin, etc.) can also be applied.

Operation on Interval data

Note

- Interval data can be transformed to nominal or ordinal scale, but with loss of information.
- Interval data can be graphed using histogram, frequency polygon, etc.

Ratio scale

- **Definition**

Interval data with a clear definition of “zero” are called ratio data.

- Example:

Temperature in Kelvin scale, Intensity of earth-quake on Richter scale, Sound intensity in Decibel, cost of an article, population of a country, etc.

Note

- All ratio data are interval data but the reverse is not true.
- In ratio scale, both differences between data values and ratios (of non-zero) data pairs are meaningful.
- Ratio data may be in linear or non-linear scale.
- Both interval and ratio data can be stored in same data type (i.e., integer, float, double, etc.)

Operation on Ratio data

- All arithmetic operations on interval data are applicable to ratio data.
- In addition, multiplication, division, etc. are allowed.
- Any linear transformation of the form $(ax + b)/c$ are known.

Ratio
Interval
Ordinal
Nominal