Selected Topics-3

Data Categorization

Dr Amira Abdelatey
amira.mohamed@ci.menofia.edu.eg

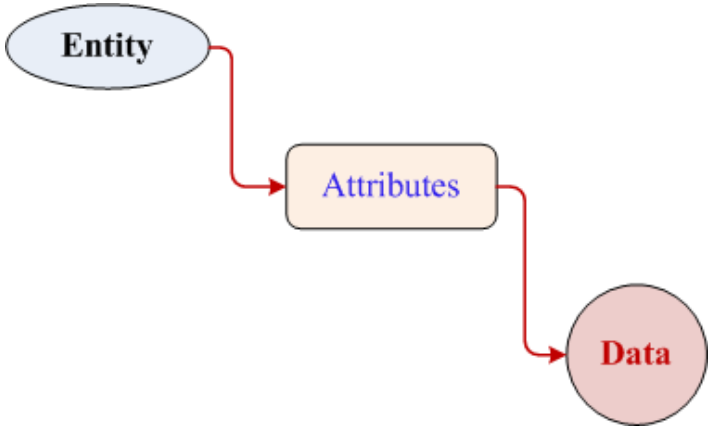
- ✓ **Face Mask**
- ✓ **Respect Social Distancing**

Just a minute to mark your attendance

Today’s discussion…

- Data in data analytics
- NOIR topology
- Nominal scale
  - Binary
    - Symmetric
    - Asymmetric
- Ordinal scale
- Interval and ration scale
- Types of Datasets
- 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.

NOIR

Classification of scales of Measurement

NOIR classification

- The mostly recommended scales of measurement are

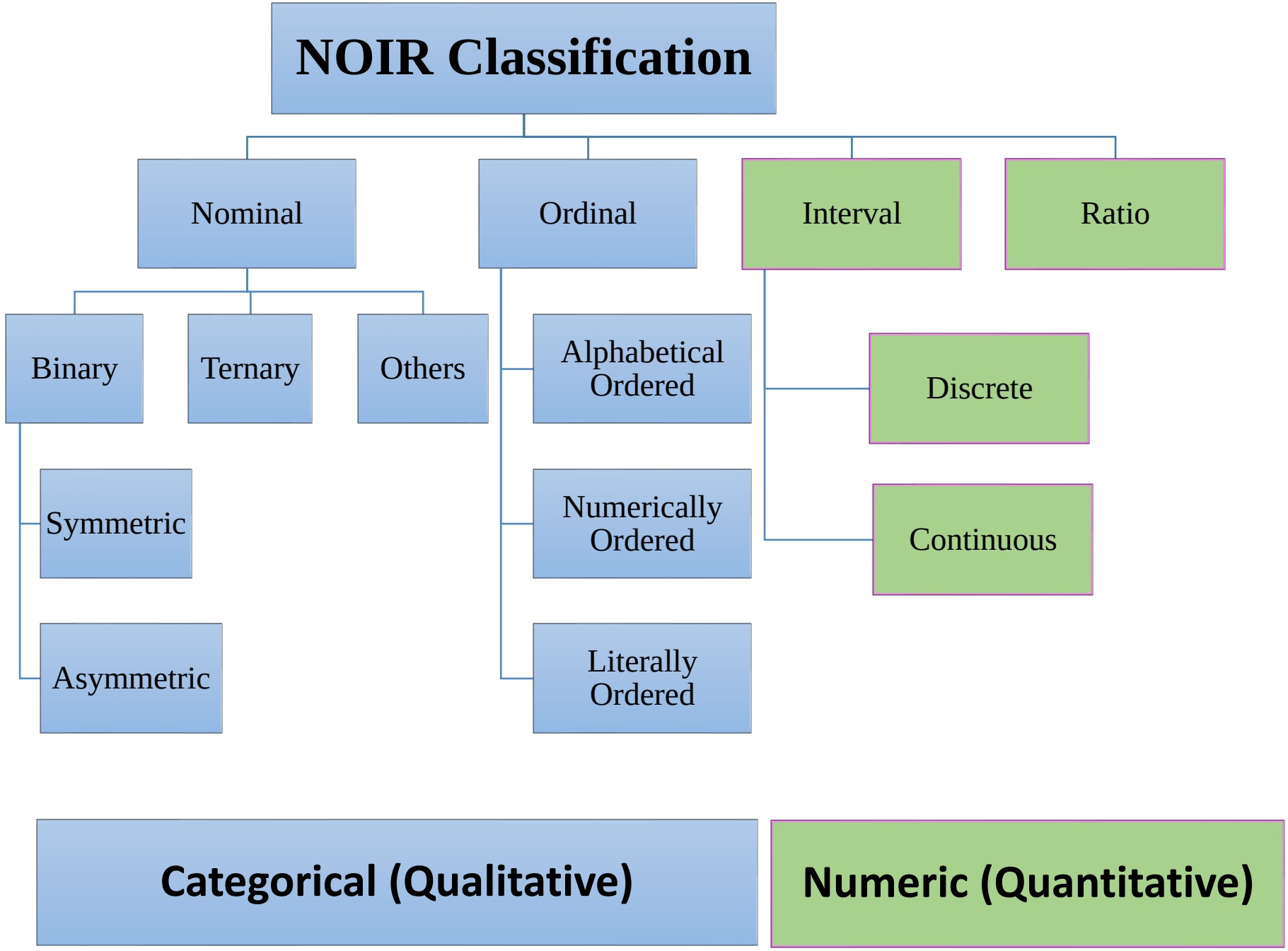
N: **N**ominal

O: **O**rdinal

I: **I**nterval

R: **R**atio

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


NOIR Classification
Nominal
Ordinal
Interval
Ratio
Binary
Ternary
Others
Symmetric
Asymmetric
Alphabetical Ordered
Numerically Ordered
Literally Ordered
Discrete
Continuous
Categorical (Qualitative)
Numeric (Quantitative)

Data Categorization

There are two general types of data – **quantitative and qualitative** and both are equally important. You use both types to demonstrate effectiveness, importance or value.

	Quantitative Variables	Qualitative Variables
Definition	*Take on numeric values*	*Take on names or labels*
Examples	Number of students in a class	Eye color
	Number of square feet in a house	Gender
	Population size of a city	Breed of dog
	Age of an individual	Level of Education
	Height of an individual	Marital status

Properties of data

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

#	Property	Operation	Type
1.	Distinctiveness	= and ≠	Categorical (Qualitative)
2.	Order	< , ≤ , > , ≥	
3.	Addition	+ and -	Numerical (Quantitative)
4.	Multiplication	* and /	

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 (≠).
- 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: medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)

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)

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 (< , ≤ , > , ≥).

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.)
- Calculations based on order are permitted (such as count, min, max, etc.).
- 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^{o} does not mean absence of temperature, that is, no heat!

Operation on Interval data

- We can add to or from interval data.
  - For example: date1 + x-days = date2
- Subtraction can also be performed.
  - For example: current date – date of birth = age
- Negation (changing the sign) and multiplication by a constant are permitted.

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.

Discrete and Continuous Data

Discrete data can only take on certain individual values.

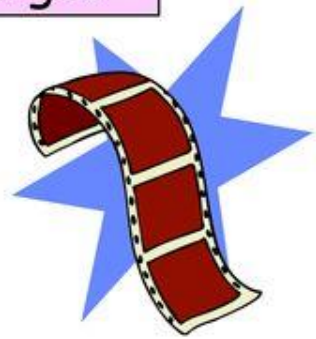
Continuous data can take on any value in a certain range.

Example 1

Number of pages in a book is a **discrete variable**.

Example 2

Length of a film is a **continuous variable**.



Example 3

Shoe size is a **Discrete variable**. E.g. $5, 5\frac{1}{2}, 6, 6\frac{1}{2}$ etc. Not in between.

Example 4

Temperature is a **continuous variable**.

Example 5

Number of people in a race is a **discrete variable**.

Example 6

Time taken to run a race is a **continuous variable**

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.
- 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.

Types of dataset: (1)Record data

- Relational records:
  - Relational tables, highly structured

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

Types of dataset: (1)Record data

- Data matrix, e.g., numerical matrix, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Types of dataset: (1)Record data

- Transaction data

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

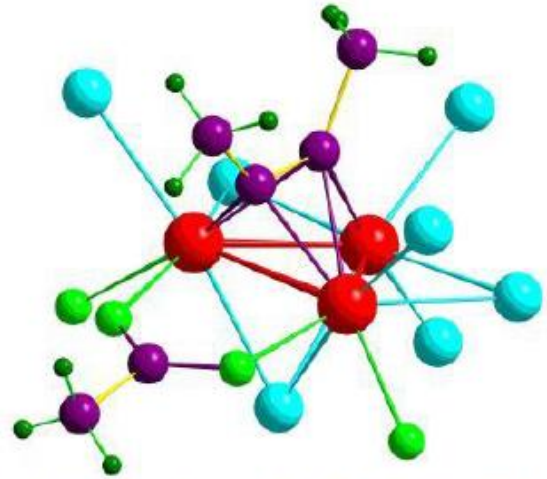
Types of dataset: (1)Record data

- Document data: Term-frequency vector (matrix) of text documents

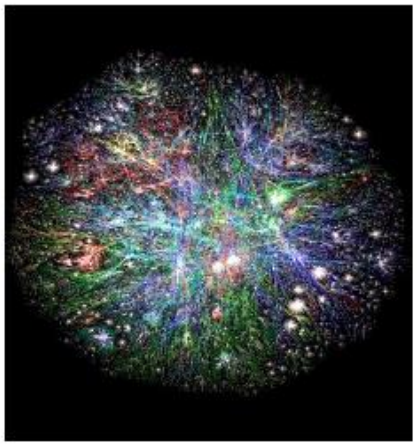
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

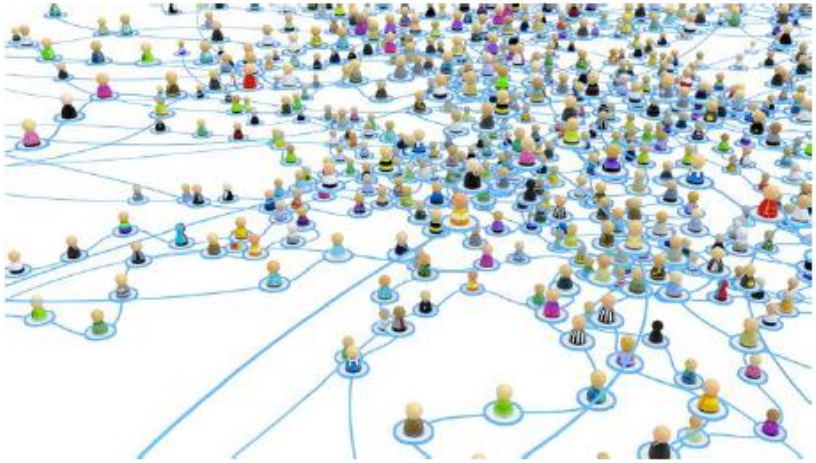
Types of dataset: (2) graphs and networks

- Transportation network



- World Wide Web



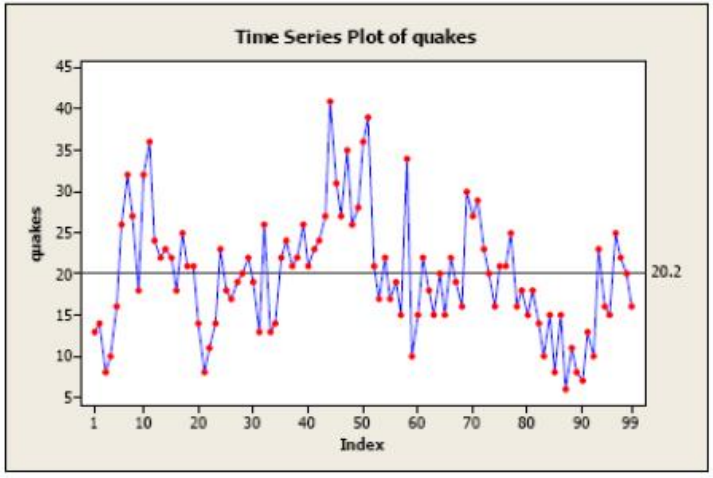


- Molecular Structures
- Social or information networks

Types of dataset: (3) ordered data

- Video data: sequence of images

- Temporal data: time-series



- Sequential Data: transaction sequences

- Genetic sequence data

Data Cube

Multidimensional Data Modeling

Concept of data cube

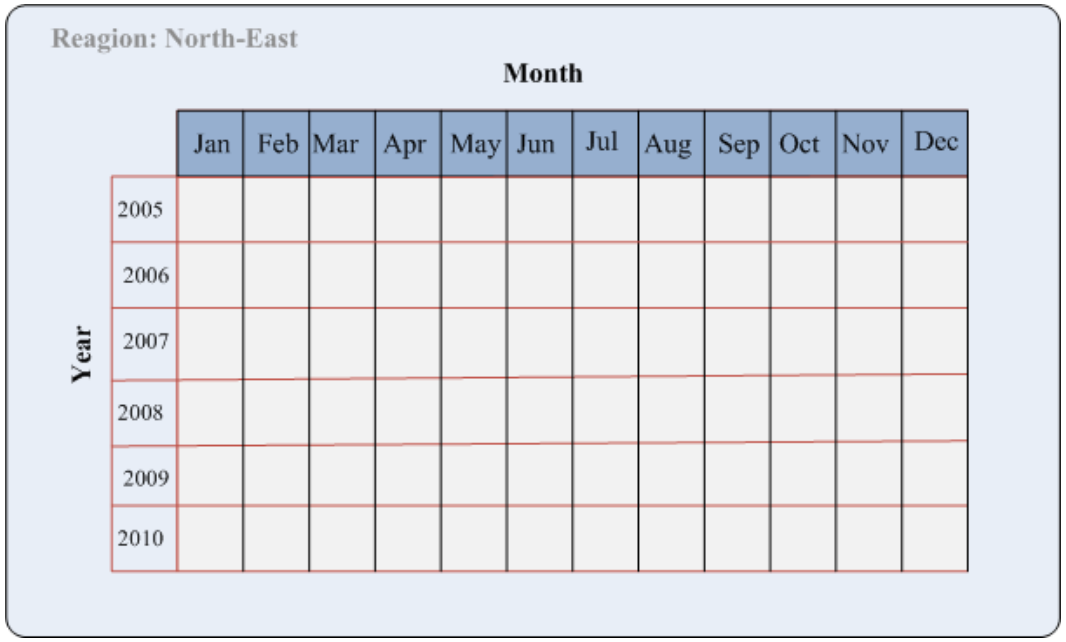
- A multidimensional data model views data in the form of a cube.

- A data cube is characterized with two things
  - **Dimension:** the perspective or entities with respect to which an organization wants to keep record.
  - **Fact:** The actual values in the record

Example.

- Rainfall data of Metrological Department
  - Time (Year, Season, Month, Week, Day, etc.)
  - Location (Country, Region, State, etc.)

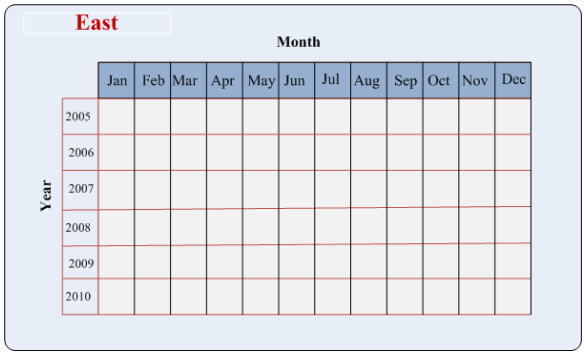
2-D view of rainfall data

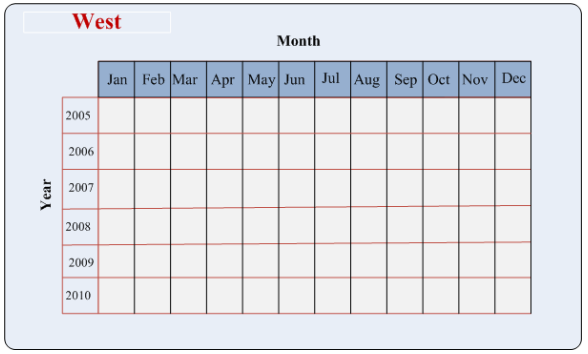


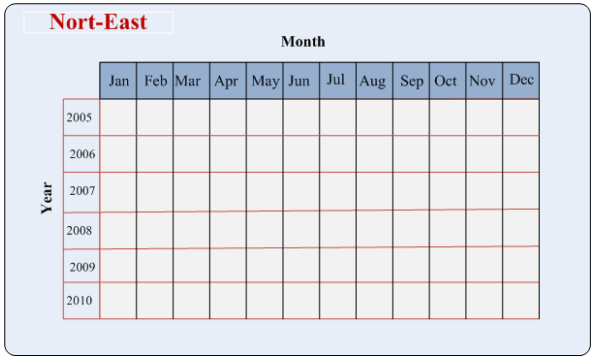
- In this 2-D representation, the rainfall for “North-East” region are shown with respect to different months for a period of years

3-D view of rainfall data

- Suppose, we want to represent data according to times (Year, Month) as well as regions of a country say East, West, North, North-East, etc.

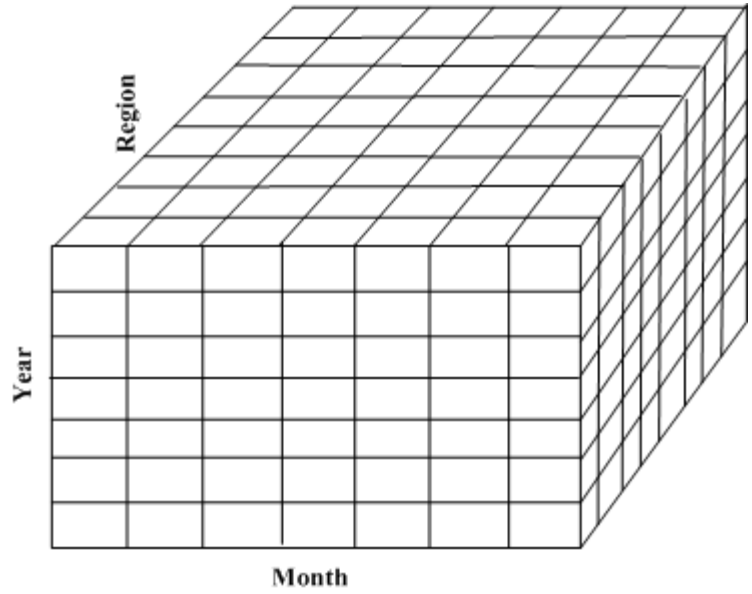






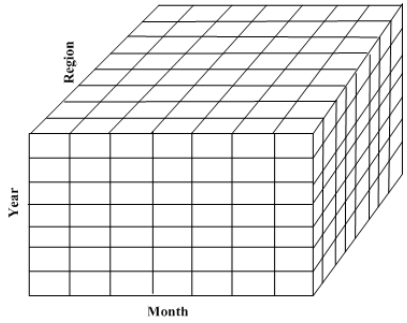
- A 2-D view of 3-D rainfall data

3-D view of rainfall data

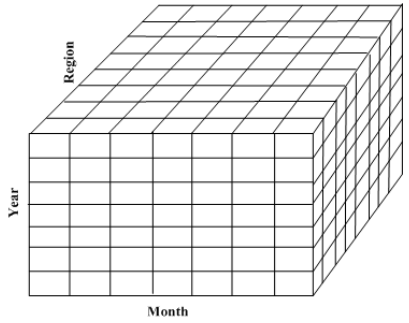


- Data cube: This enables us a 3-D view of the rainfall data

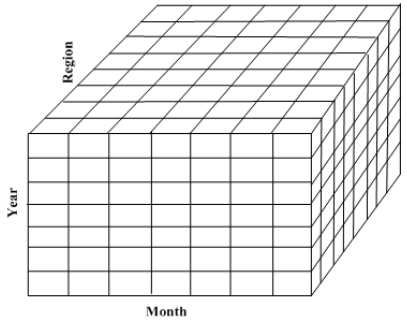
3-D view of rainfall data



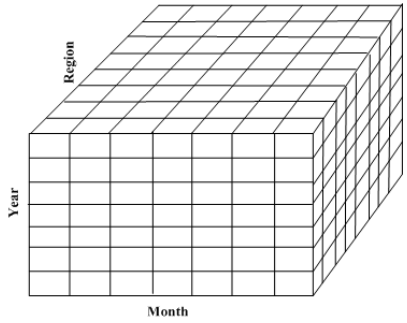
India



China



Russia

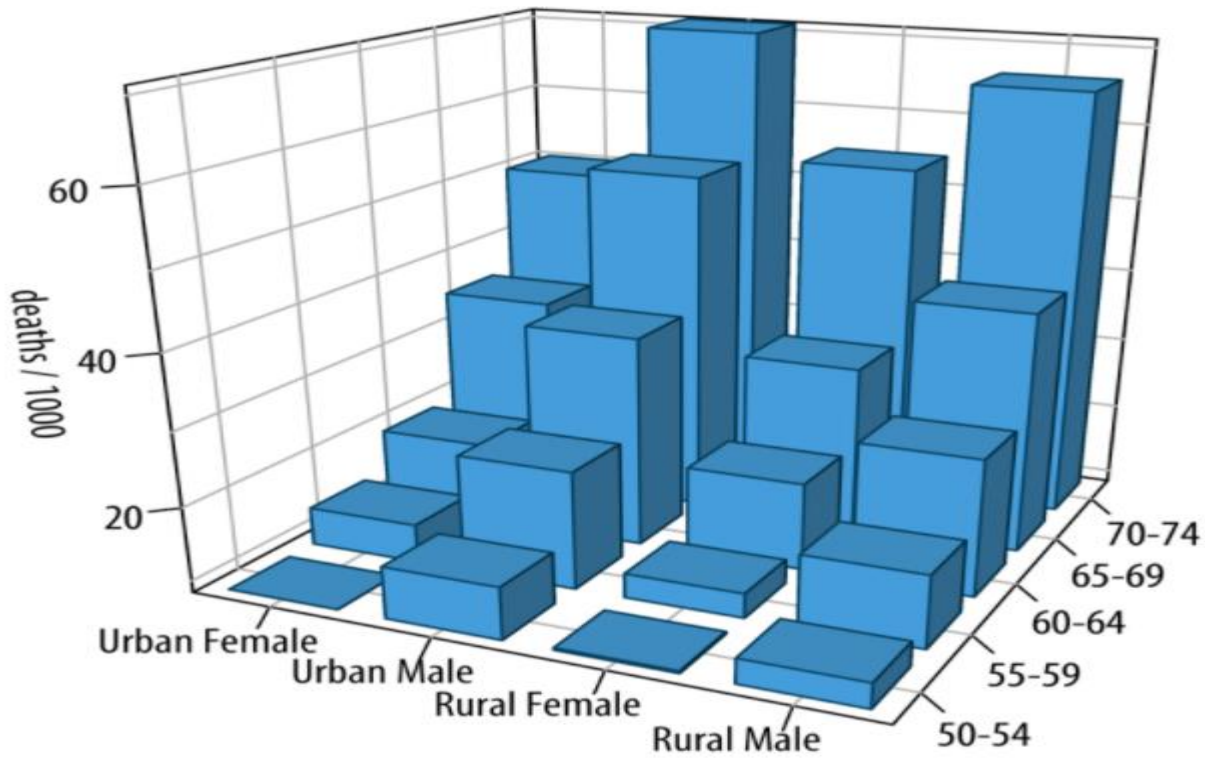


Pakistan

- Data cube: This enables us a 3-D view of the rainfall data for a continent say?

3-D view of rainfall data

- What is the data cube representation of rainfall data of the entire world?



Reference

- The detail material related to this lecture can be found in

  Data Mining: Concepts and Techniques (3rd Edn.) by Jiawei Han, Michelline Kamber and Jian Pei, Morgan Kaufmann (2014).

Any question?

Questions of the day…

1. Consider an image as an entity.
   - What are the attributes you should think to represent an image?
   - Categorize each attribute according to the NOIR data classification.
   - Suppose, two images are given. Give an idea to check if two images are identical or not.

2. How you can convert a data of interval type to ordinal type? Give an example. What are the issues of such transformation? Whether the reverse is possible or not? Justify you answer.

Questions of the day…

3. What are the different properties used to categorize the data according to NOIR data categorization?

4. Given an entity say “STUDENT” with the following attributes. Identify the NOIR category to which each of them belongs.

Scholarship amount	Name	RollNo	DoB	Aaadhar No.	Gender	Mobiloe No.	Email Id

Questions of the day…

5. Give the concept of data cube to represent hyper-dimensional data? Also, explain with suitable diagrams the following.
   - Roll up
   - Drill down
   - Slice
6. Using the concept of data cube, how YouTube can archive videos of all type?
7. Give FOUR differences between data of types “interval” and “ratio-scale”