Data Science (CS3206)

Lecture #3

Data Categorization

Quote of the day..

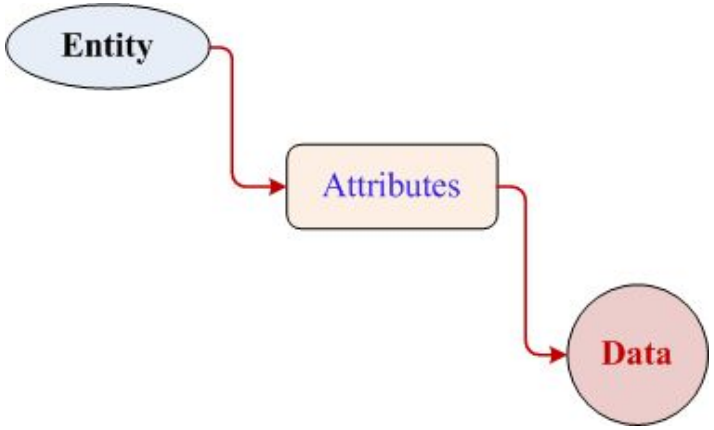
- ‘The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn.’

–Alvin Toffler

Today's discussion…

- NOIR topology
- Nominal scale
  - Binary
    - Symmetric
    - Asymmetric
- Ordinal scale
- Interval and Ratio 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.).

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	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

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

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 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

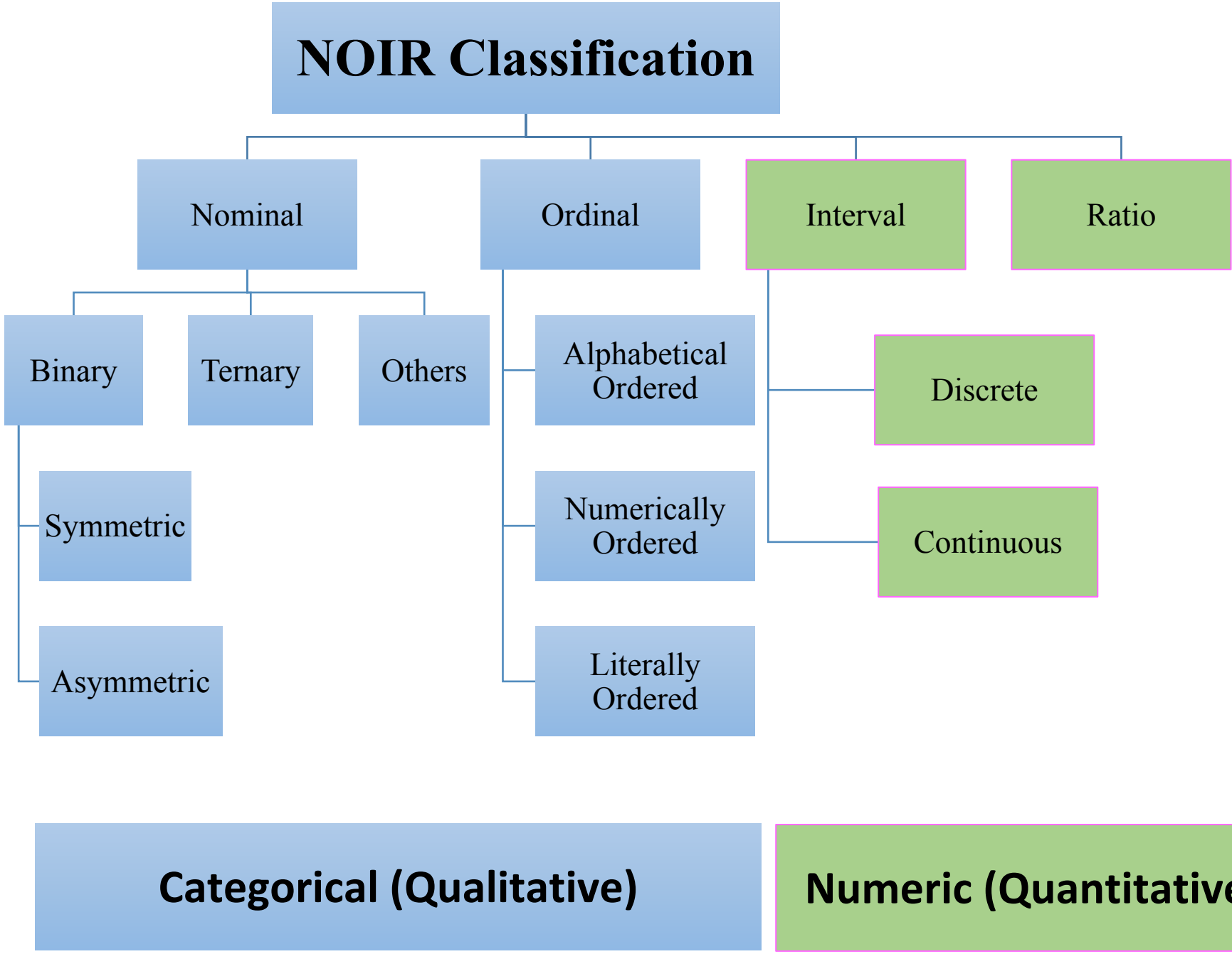
N: **N**ominal

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
Alphabetical Ordered
Discrete
Symmetric
Numerically Ordered
Continuous
Asymmetric
Literally Ordered
Categorical (Qualitative)
Numeric (Quantitative)

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 /	

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.

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: 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)

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.) 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:
  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; 0^0 does not mean absence of temperature, that is, no heat!

Operation on Interval data

- We can add to or from interval data.
  - For example: 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.
- 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,weight, height, latitude, longitude, weather, 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.

Which of the following classifications of variable types is **false**?

A. Whether a student has previously taken a statistics course → categorical
B. Customer satisfaction: very unsatisfied, unsatisfied, satisfied, very satisfied → ordinal categorical
C. Population of each state in the US → continuous numerical
D. Student height → continuous numerical

Data Cube

Multidimensional Data Modeling

Concept of data cube

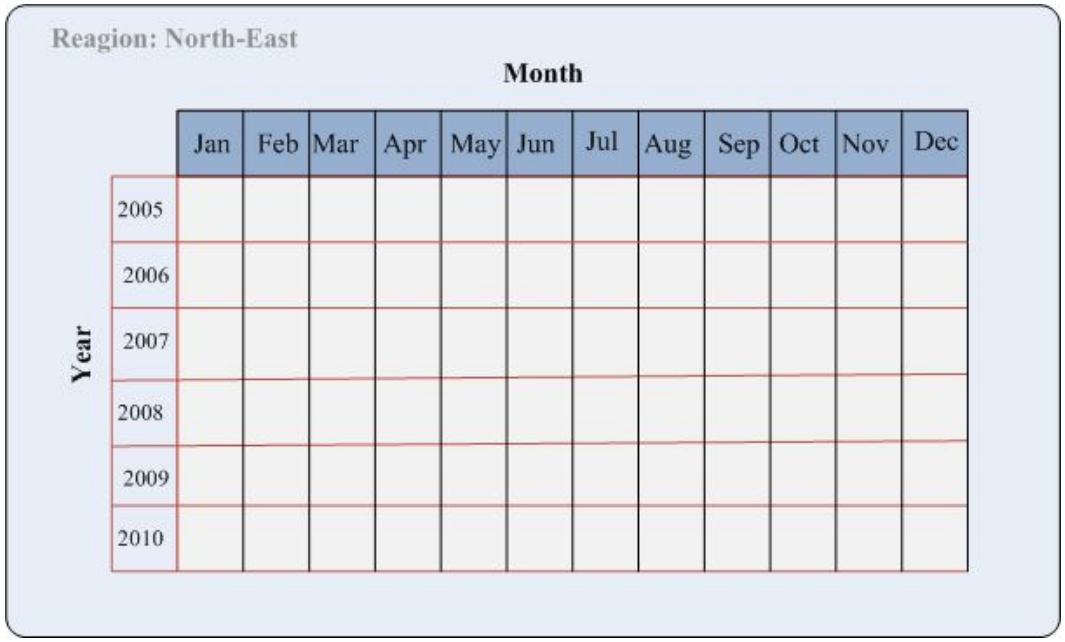
- A multidimensional data model views data in the form of a cube.

- 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.)

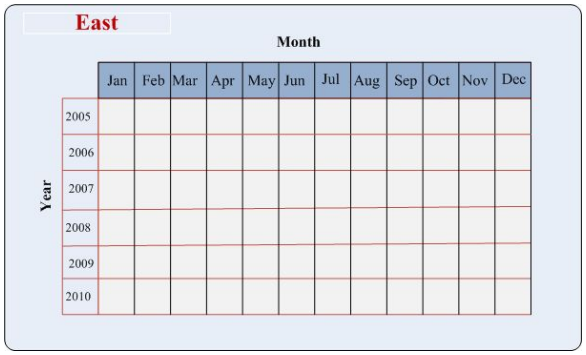
2-D view of rainfall data

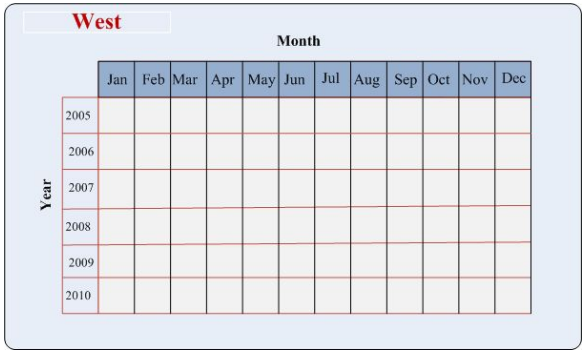


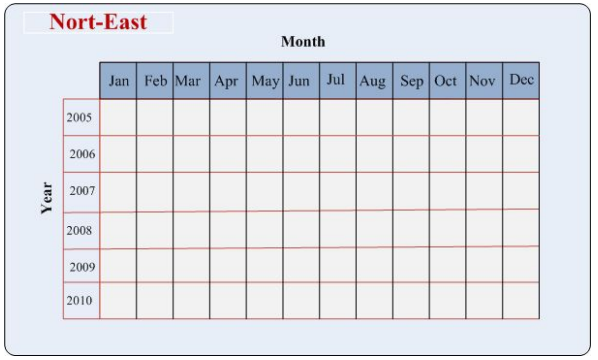
- In this 2-D representation, the rainfall for “North-East” region are shown with respect to different months for a period of years

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.

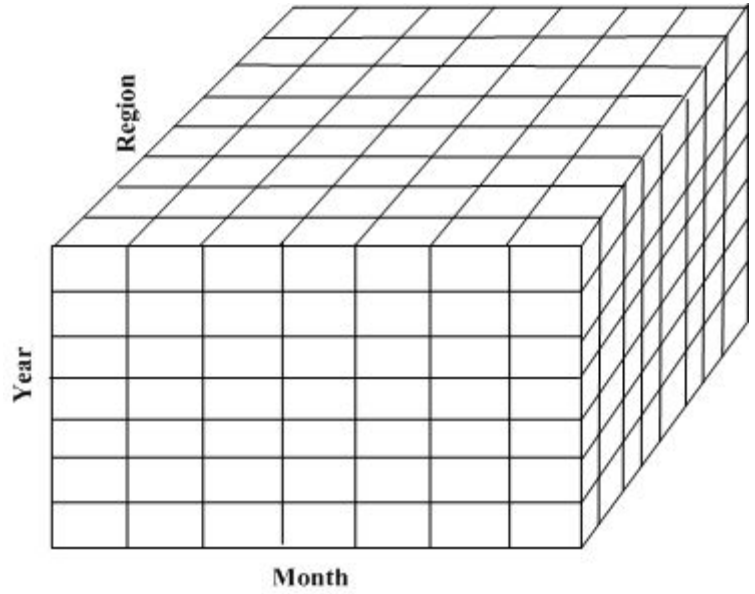






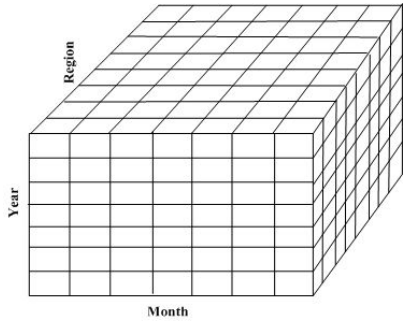
- A 2-D view of 3-D rainfall data

3-D view of rainfall data

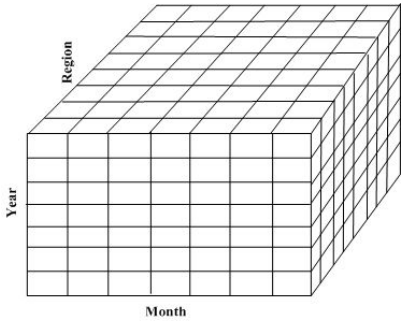


- Data cube: This enables us a 3-D view of the rainfall data

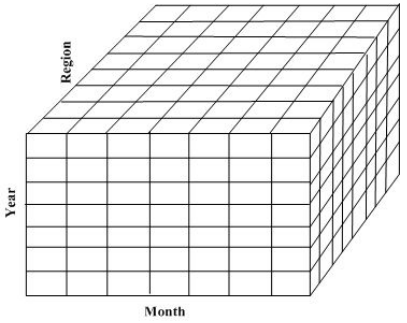
3-D view of rainfall data



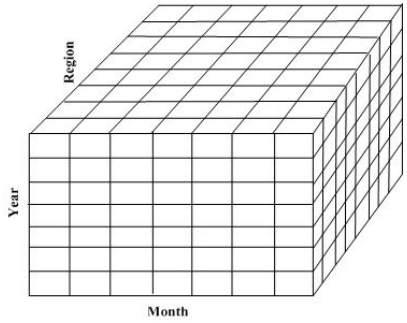
India



China



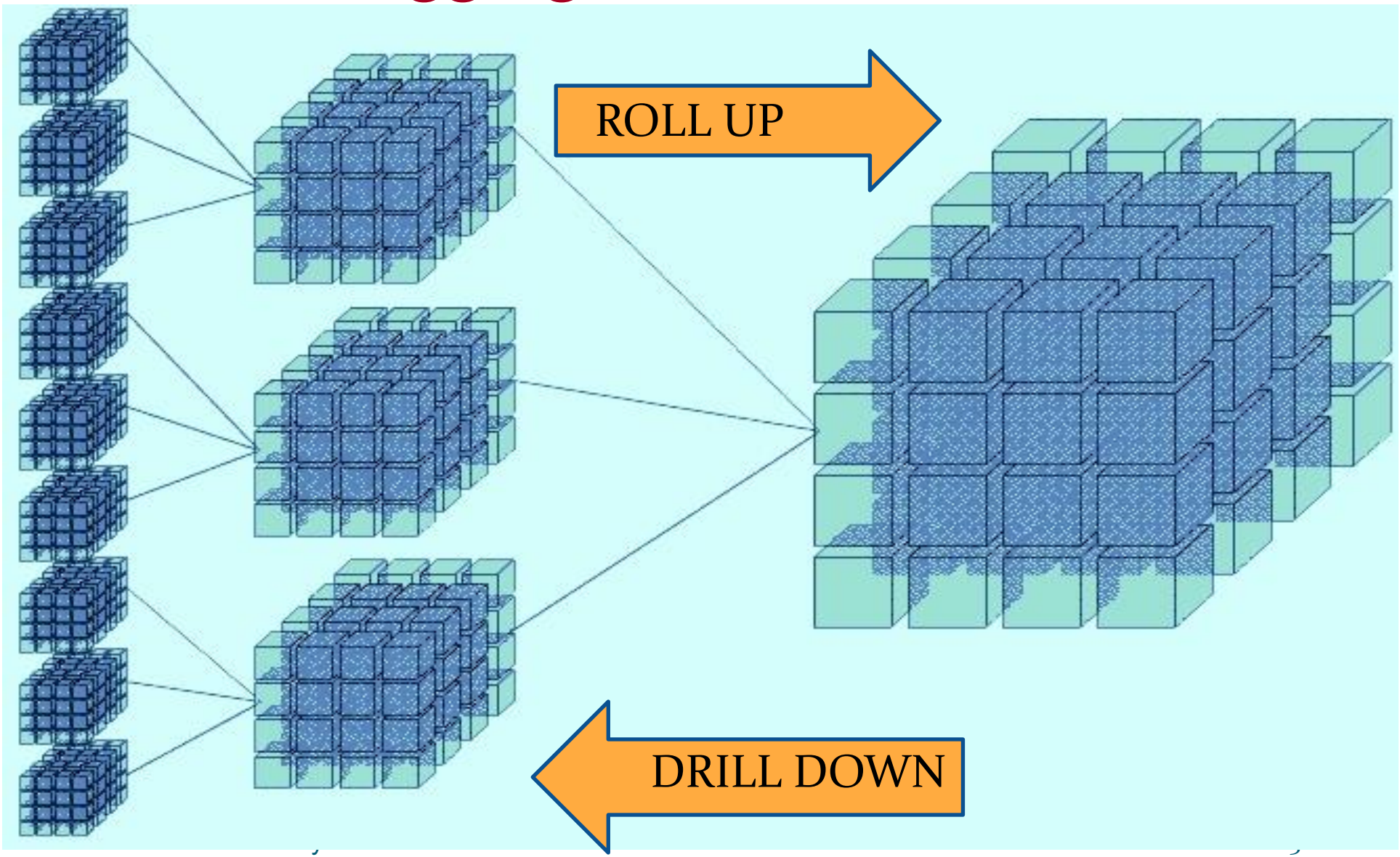
Russia



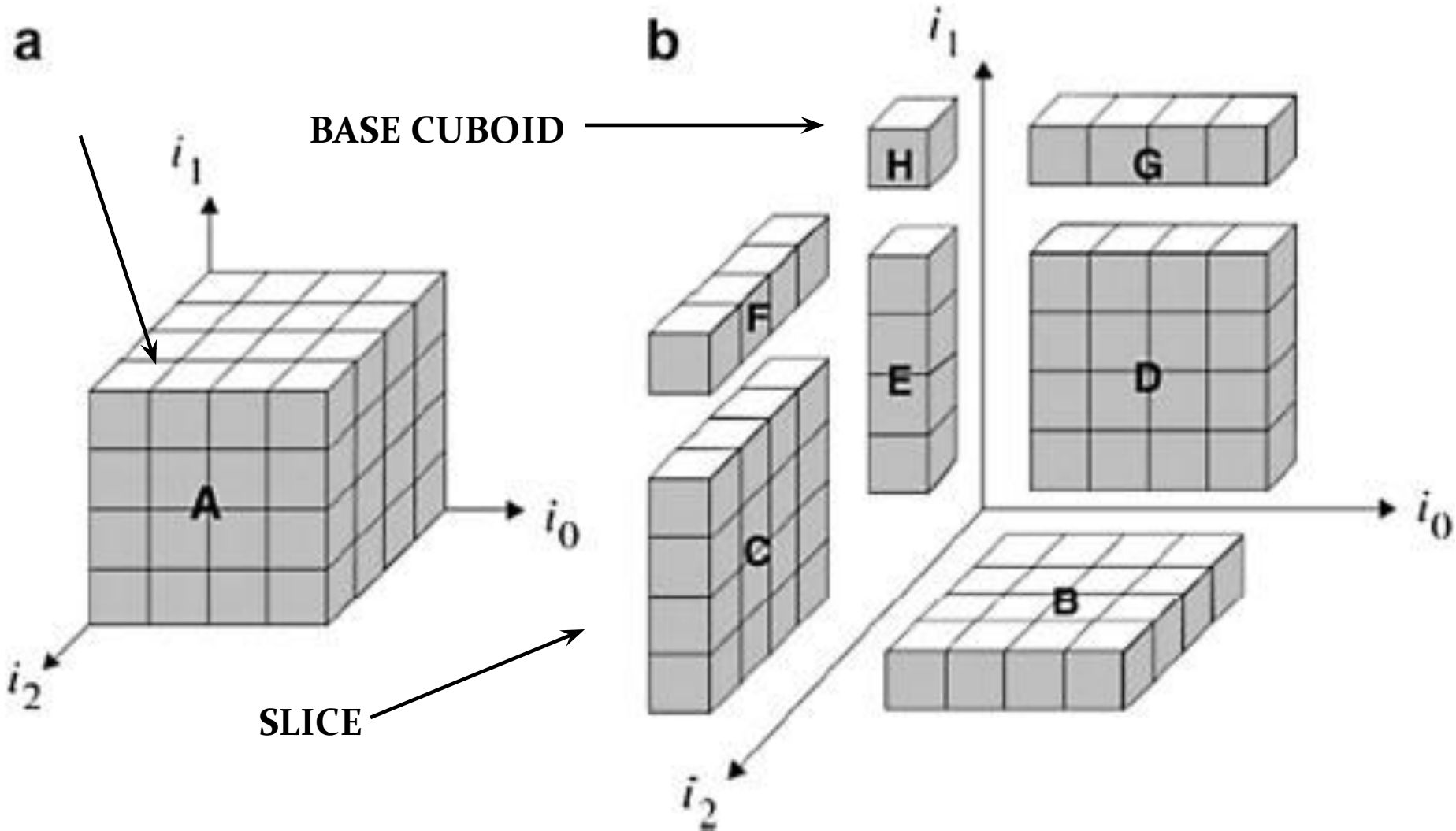
Pakistan

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

Data cube aggregation



Data cube segregation



Data representation

- How a document (e.g., text) can be represented?

THE SILVER CHAIR

by C.S. Lewis

CHAPTER ONE
BEHIND THE GYM

IT was a dull autumn day and Jill Pole was crying behind the gym.

She was crying because they had been bullying her. This is not going to be a school story, so I shall say as little as possible about Jill's school, which is not a pleasant subject. It was "Co-educational," a school for both boys and girls, what used to be called a "mixed" school; some said it was not nearly so mixed as the minds of the people who ran it. Those people had the idea that boys and girls should be allowed to do what they liked. And unfortunately what ten or fifteen of the biggest boys and girls liked best was bullying the others. All sorts of things, horrid things, went on which at an ordinary school would have been found out and stopped in half a term, but at this school they weren't. Or even if they were, the people who did them were not expelled or punished. The Head said they were interesting psychological cases and sent for them and talked to them for hours. And if you know the right sort of things to say to the Head the main result was that you became rather a favourite than otherwise.

That was why Jill Pole was crying on that dull autumn day on the damp little path which runs between the back of the gym and the shrubbery. And she hadn't nearly finished her cry when a boy came round the corner of the gym whistling, with his hands in his pockets. He nearly ran into her.

"Can't you look where you're going?" said Jill Pole.

"All right," said the boy, "you needn't start -" and then he noticed her face. "I say, Pole," he said, "what's up?"

Jill only made faces; she isn't one to make when you're trying to say something but find that if you speak you'll start crying again.

"It's Thats, I suppose – as usual," said the boy grimly, digging his hands farther into his pockets.

Jill nodded. There was no need for her to say anything, even if she could have said it. They both knew.

"Now, look here," said the boy, "there's no good in all -"

He meant well, but he did talk rather like someone beginning a lecture. Jill suddenly flew into a temper (which is quite a likely thing to happen if you have been interrupted in a cry).

Data representation

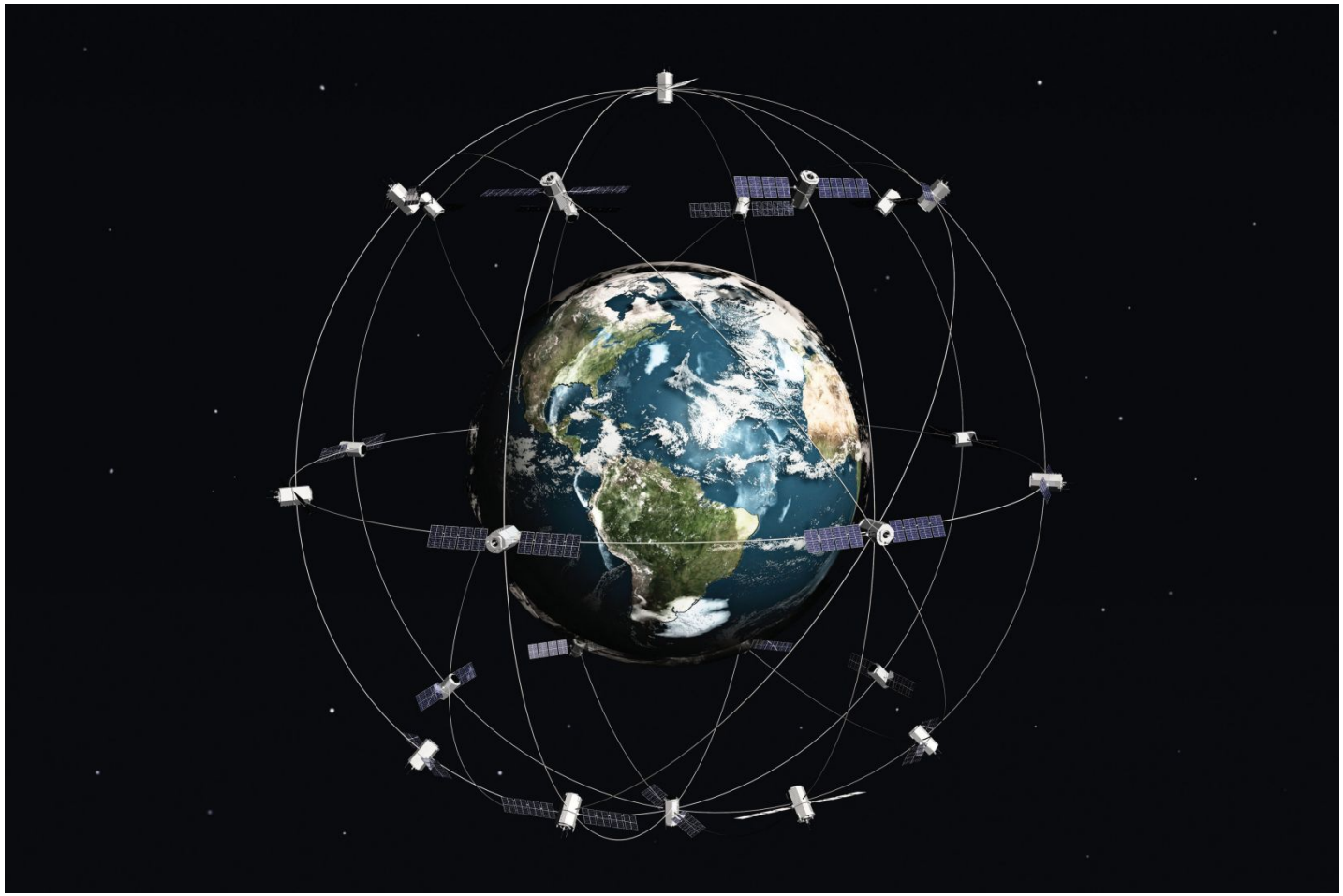
- How an image can be represented?

Data representation

- How a video can be represented?

Data representation

- How the streaming data from an artificial earth satellite can be represented?



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).

Present size of the digital universe is in the order of

(a) Terabyte (TB)

(b) Petabyte (PB)

(c) Exabyte (EB)

(d) Zetabyte (ZB)

Which is/are the source of data in data analytics?

(a) Scientific instruments

(b) Social media

(c) Mobile devices

(d) Sensor networks

Elastic is a tool for

(a) Strong big data

(b) Processing data with scalable architecture

(c) A distributed file system

(d) Cloud security

MapReduce is meant for

(a) Data visualizations

(b) Massive parallel programming

(c) Query reporting

(d) Data storage in Cloud

Which data scale uses “zero point as origin”?

A. Nominal
B. Ordinal
C. Interval
D. Ratio

Which operation cannot be carried out on “Ordinal” data?

(a) To find the minimum

(b) To find the mean

(c) To find the mode

(d) To find the median

Map from entries in Column A to appropriate entries in Column B in the following table.

	Column A		Column B
(p)	Pig	(w)	Data storage
(q)	HDFS	(x)	Data process server
(r)	EC_2	(y)	Tool from parallel programming
(s)	ZooKeeper	(z)	Data analysis technique

(a) (p)-(y), (q)-(w), (r)-(x), (s)-(z)

(b) (p)-(x), (q)-(w), (r)-(y), (s)-(z)

(c) (p)-(w), (q)-(y), (r)-(z), (s)-(x)

(d) (p)-(y), (q)-(x), (r)-(z), (s)-(w)

Consider the data about all students in a course stored with the following structure.

Table Q.22

Name	Roll No	Category*	Mark1	Mark2	Total	Grade
…	…	…	…	…	…	…
…	…	…	…	…	…	…
…	…	…	…	…	…	…

**Category denotes whether a student belongs to UG or PG*

If the structure is used to store the data of 100 students, then the dimension of the data is

(a) 2 (b) 7 (c) 100 (d) 200 (e) 700

According to NOIR classification, the attribute "Category" in Table. Q22 can be categorized as

(a) Categorical

(b) Symmetric binary

(c) Asymmetric binary

(d) Ordinal

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”