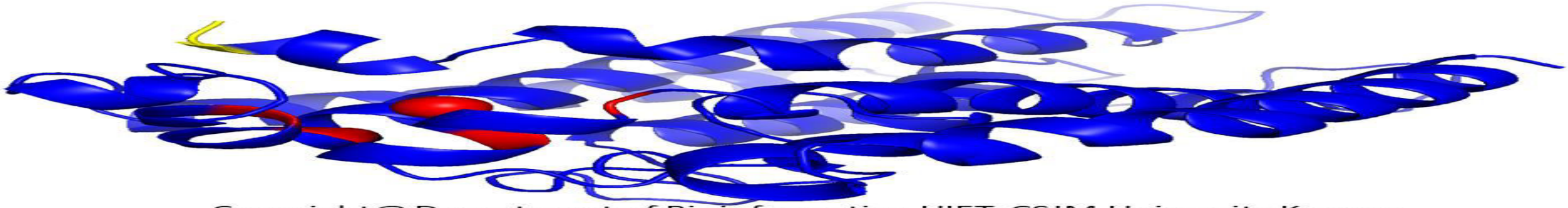
Department of Bioinformatics, University Institute of Engineering & Technology, Chhatrapati Shahuji Maharaj University, Kanpur



Copyright@ Department of Bioinformatics, UIET, CSJM University Kanpur

A Value Added Course on "Data Mining & Data Analytics June 10-June 22 2022

Patron

Honorable Vice Chancellor Dr Vinay Kumar Pathak

CSJMU

&

Dr Brishti Mitra

Director, University Institute of Engineering & Technology

Coordinator

Mamta Sagar

Head, Department of bioinformatics, UIET

Lecture 1 on Data Mining Introduction & What is Data?

Mamta Sagar

Department of Bioinformatics, UIET, CSJMU

Data Mining : An Introduction

- Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data.
- Within these masses of data lies hidden information of strategic importance.
- But when there are so many trees, how do you draw meaningful conclusions about the forest?

- The newest answer is data mining, which is being used both to increase revenues and to reduce costs.
- The potential returns are enormous. Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud.

- Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.

- Data mining applications Data mining is increasingly popular because of the substantial contribution it can make. It can be used to control costs as well as contribute to revenue increases.
- Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers.
- By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics.

- By profiling customers who have bought a particular product it can focus attention on similar customers who have not bought that product (cross-selling).
- By profiling customers who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one.
- Data mining offers value across a broad spectrum of industries. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services.

- Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications.
- Companies active in the financial markets use data mining to determine market and industry characteristics as well as to predict individual company and stock performance.
-

- Retailers are making more use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons.
- Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

What is Data ?

- Data is a collection of facts, such as numbers, words, measurements, observations or just descriptions of things.
- Data or Datum?
- The singular form is "datum", so we say "that datum **is** very high".
- "Data" is the plural so we say "the data **are** available", but data is also a **collection** of facts, so "the data **is** available" is fine too

- **Data** can be defined as a representation of facts, concepts, or instructions in a formalized manner, which should be suitable for communication, interpretation, or processing by human or electronic machine.
- Data is represented with the help of characters such as alphabets (A-Z, a-z), digits (0-9) or special characters (+,-,/,*,<,>,= etc.)

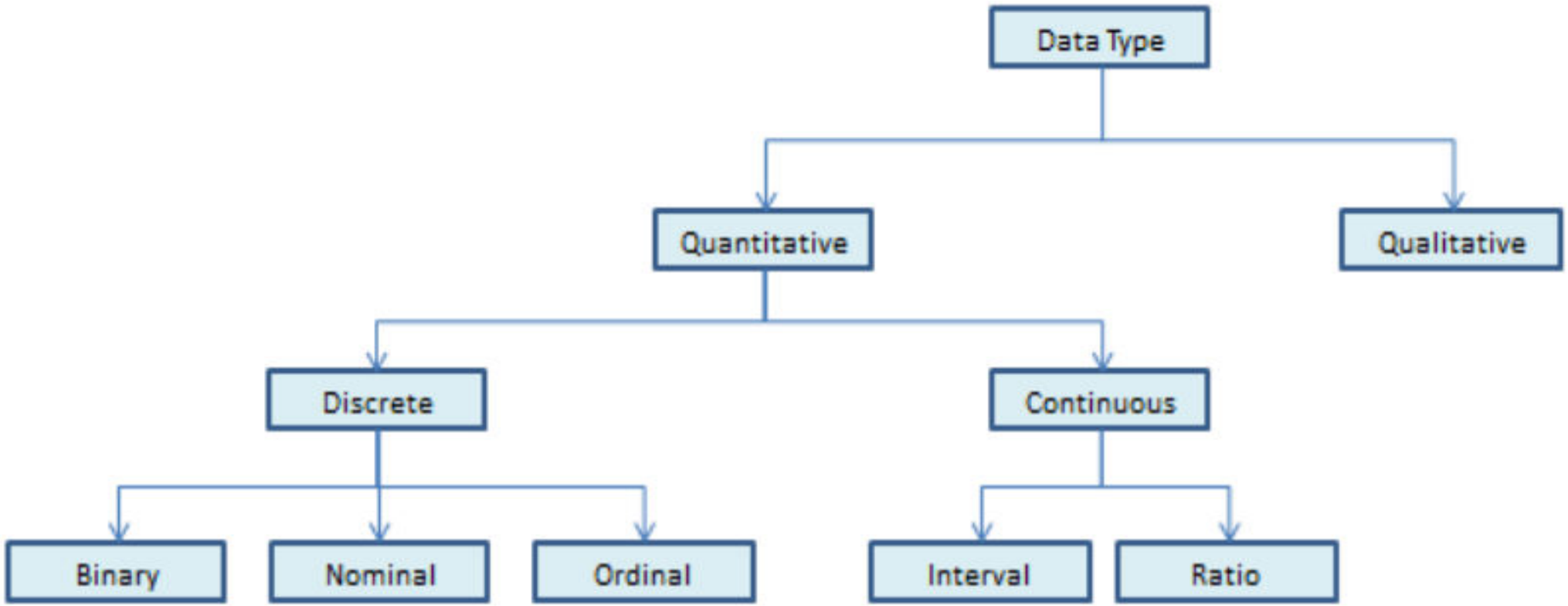
- **Qualitative vs Quantitative**
- Data can be qualitative or quantitative.
- **Qualitative data** is descriptive information (it *describes* something)
- **Quantitative data** is numerical information (numbers)

- **Quantitative data** can be Discrete or Continuous:
- **Discrete data** can only take certain values (like whole numbers)
- **Continuous data** can take any value (within a range)
- **Discrete data** is counted, **Continuous data** is measured

- **Qualitative**:
- He is brown and black
- He has long hair
- He has lots of energy
- **Quantitative**:
- Discrete:
  - He has 4 legs
  - He has 2 brothers
- Continuous:
  - He weighs 25.5 kg
  - He is 565 mm tall

- Data can be continuous, having any numerical value (e.g., quantity sold) or categorical, fitting into discrete classes (e.g., red, blue, green). Categorical data can be further defined as either ordinal, having a meaningful order (e.g., high/medium/low), or nominal, that is unordered (e.g., postal codes).

1) *Different types of data (source http://www.sigmamagic.com/forum/archives/176)*



Data Collection

- You want to find how many cars pass by a certain point on a road in a 10-minute interval.
- So: stand near that road, and count the cars that pass by in 10 minutes.
- You might want to count many 10-minute intervals at different times during the day, and on different days too!

The nature of data

- Diversity: ranging from gene & protein- pathways-networks-cell tissues-organism population, different trees in a forest
- Variability: Different individuals & species vary tremendously. For ex. Structure & function of organs vary across age & gender

Data sources in life science

- Nucleic acid, protein
- Biological data base are autonomous
- Heterogeneous
- Dynamic
- Computational tools are required specific I/O formats and broad domain knowledge

Modeling & simulation of protein- protein interaction, cellular process need more experimental observation to fill in missing quantitative details for mature efforts

Challenges in information integration

- Heterogeneity
- Data Complexity

References

- Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation.
- https://www.mathsisfun.com/data/data.html
- https://www.tutorialspoint.com/computer_fundamentals/computer_data.htm