SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

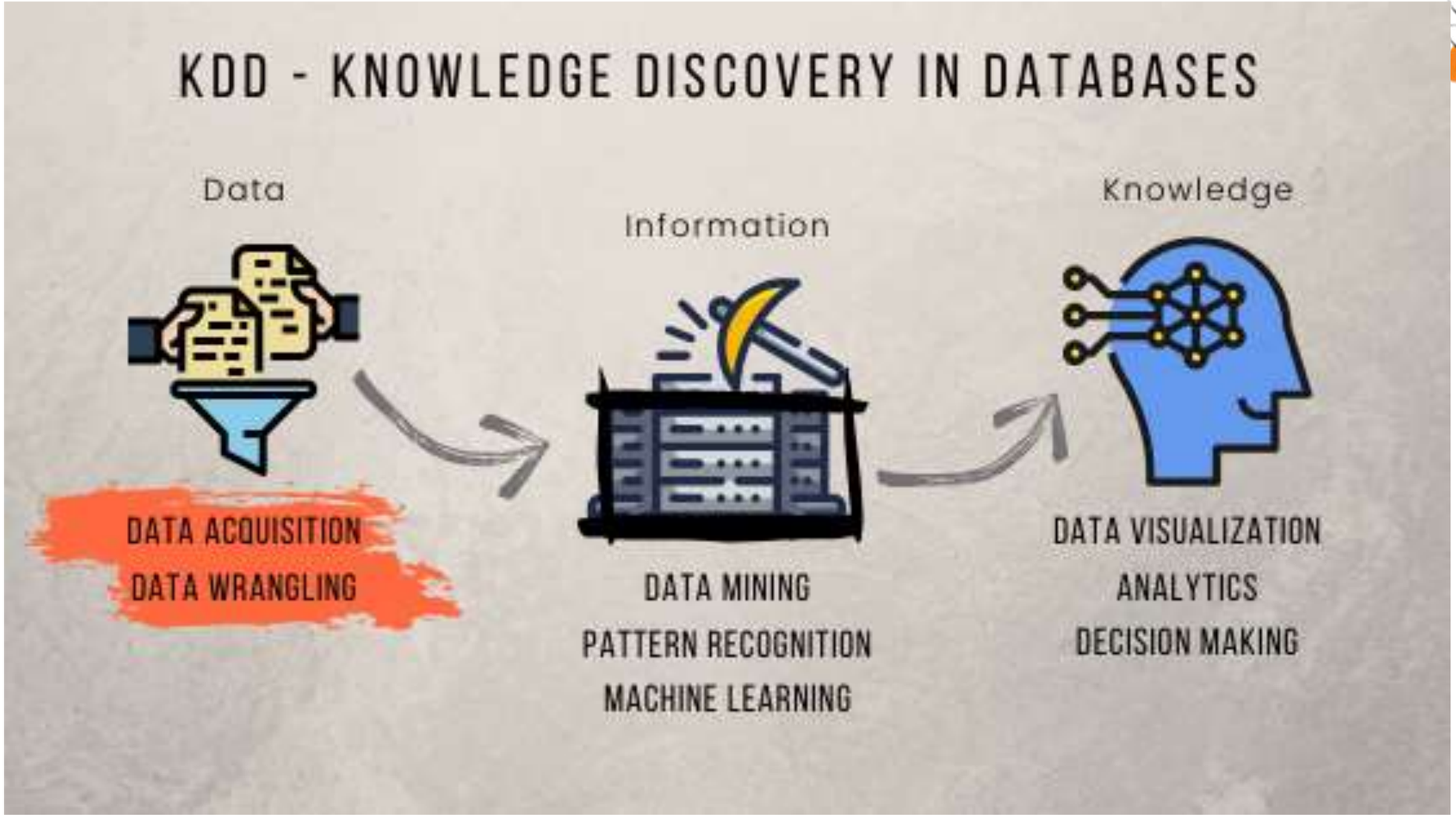
DEPARTMENT OF COMPUTER SCIENCE AND TECHNOLOGY

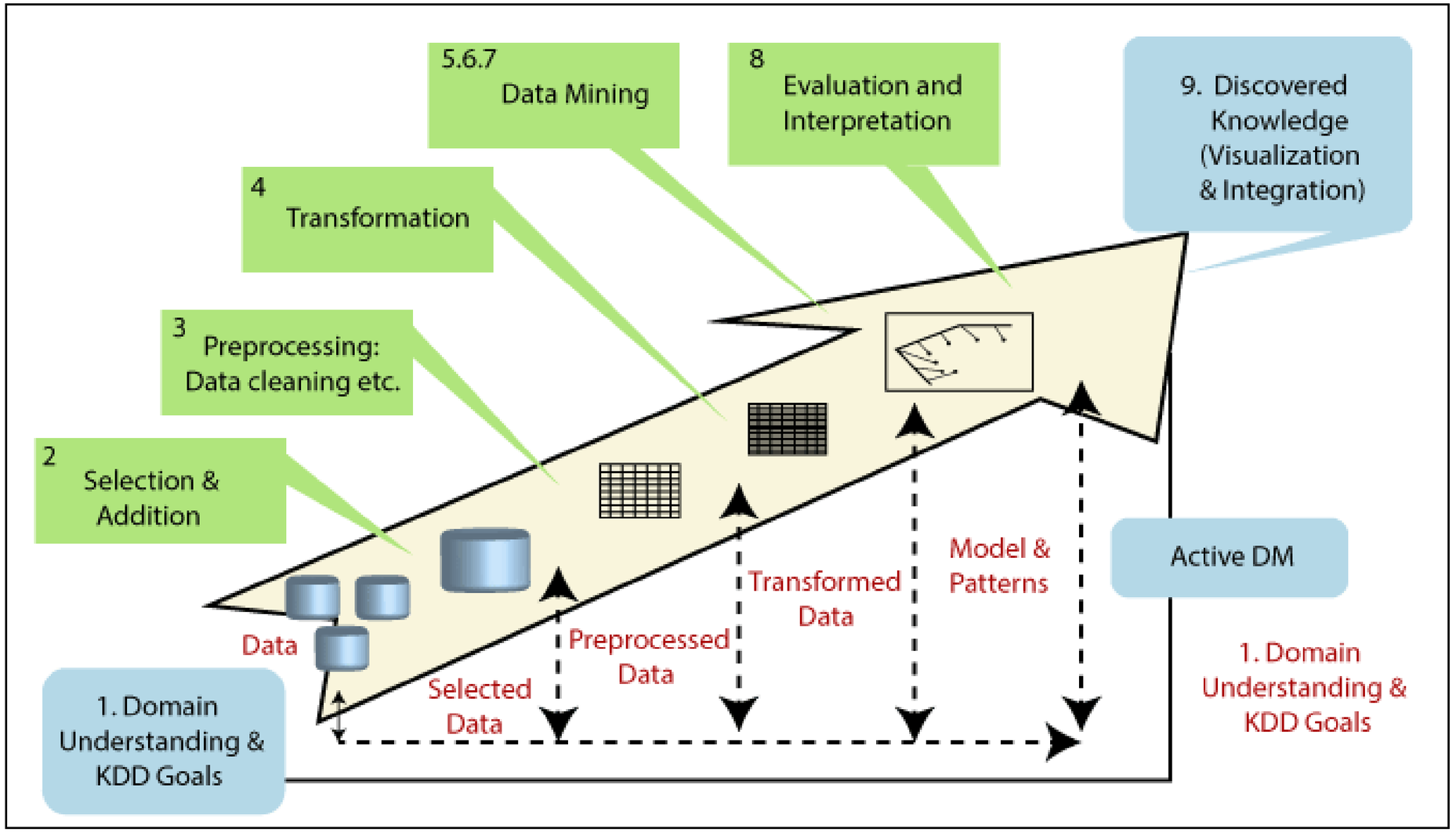
COURSE NAME :19CS407 DATA ANALYTICS WITH R

II YEAR /IV SEMESTER

Unit 1- Introduction

Topic : KDD Process


KDD - KNOWLEDGE DISCOVERY IN DATABASES
Data
Information
Knowledge
DATA ACQUISITION
DATA WRANGLING
DATA MINING
PATTERN RECOGNITION
MACHINE LEARNING
DATA VISUALIZATION
ANALYTICS
DECISION MAKING


5.6.7
Data Mining
8
Evaluation and
Interpretation
9. Discovered
Knowledge
(Visualization
& Integration)
4
Transformation
3
Preprocessing:
Data cleaning etc.
2
Selection &
Addition
Active DM
Model &
Patterns
Transformed
Data
Preprocessed
Data
Selected
Data
Data
1. Domain
Understanding &
KDD Goals
1. Domain
Understanding &
KDD Goals

The KDD Process

- Intended to be a methodology that could cope with all the processes necessary to extract knowledge from data, the KDD process proposes a sequence of nine steps.

- In spite of the sequence, the KDD process considers the possibility of going back to any previous step in order to redo some part of the process.

Learning the application domain

- ✓ What is expected in terms of the application domain?
- ✓ What are the characteristics of the problem; its specificities?
- ✓ A good understanding of the application domain is required

Creating a target dataset

- ✓ What data are needed for the problem?
- ✓ Which attributes?
- ✓ How will they be collected and put in the desired format (say,a tabular data set)?
- ✓ Once the application domain is known, the data analyst team should be able to identify the data necessary to accomplish the project

Data cleaning and pre-processing:

- ✓ How should missing values and/or outliers such as extreme values be handled?
- ✓ What data type should we choose for each attribute?
- ✓ It is necessary to put the data in a specific format, such as a tabular format.

  - ✓ Cleaning in case of Missing values.
  - ✓ Cleaning noisy data, where noise is a random or variance error.
  - ✓ Cleaning with Data discrepancy detection and Data transformation tools

Data reduction and projection

- Which features should we include to represent the data? From the available features, which ones should be discarded?
- Should further information be added, such as adding the day of the week to a timestamp?
- This can be useful in some tasks. Irrelevant attributes should be removed.

Choosing the data mining function

- Which type of methods should be used?
- Four types of method are: summarization, clustering, classification and regression.
- The first two are from the branch of descriptive analytics while the latter two are from predictive analytics.

Choosing the data mining algorithm(s)

- ✓ Given the characteristics of the problem and the characteristics of the data, which methods should be used?
- ✓ It is expected that specific algorithms will be selected

Data mining

- ✓ Given the characteristics of the problem, the characteristics of the data, and the applicable method type, which specific methods should be used?
- ✓ Which values should be assigned to the hyper-parameters?
- ✓ The choice of method depends on many different factors: interpretability, ability to handle missing values, capacity to deal with outliers, computational efficiency, among others.

Interpretation

- ✓ What is the meaning of the results?
- ✓ What is the utility for the final user?
- ✓ To select the useful results and to evaluate them in terms of the application domain is the goal of this step.
- ✓ It is common to go back to a previous step when the results are not as good as expected

Using discovered knowledge

- ✓ How can we apply the new knowledge in practice?
- ✓ How is it integrated in everyday life?
- ✓ This implies the integration of the new knowledge into the operational system or in the reporting system

Assessment 1

To create your own KDD Process

References

1. João Moreira, Andre Carvalho, Tomás Horvath – “A General Introduction to Data Analytics” – Wiley -2018

Thank You