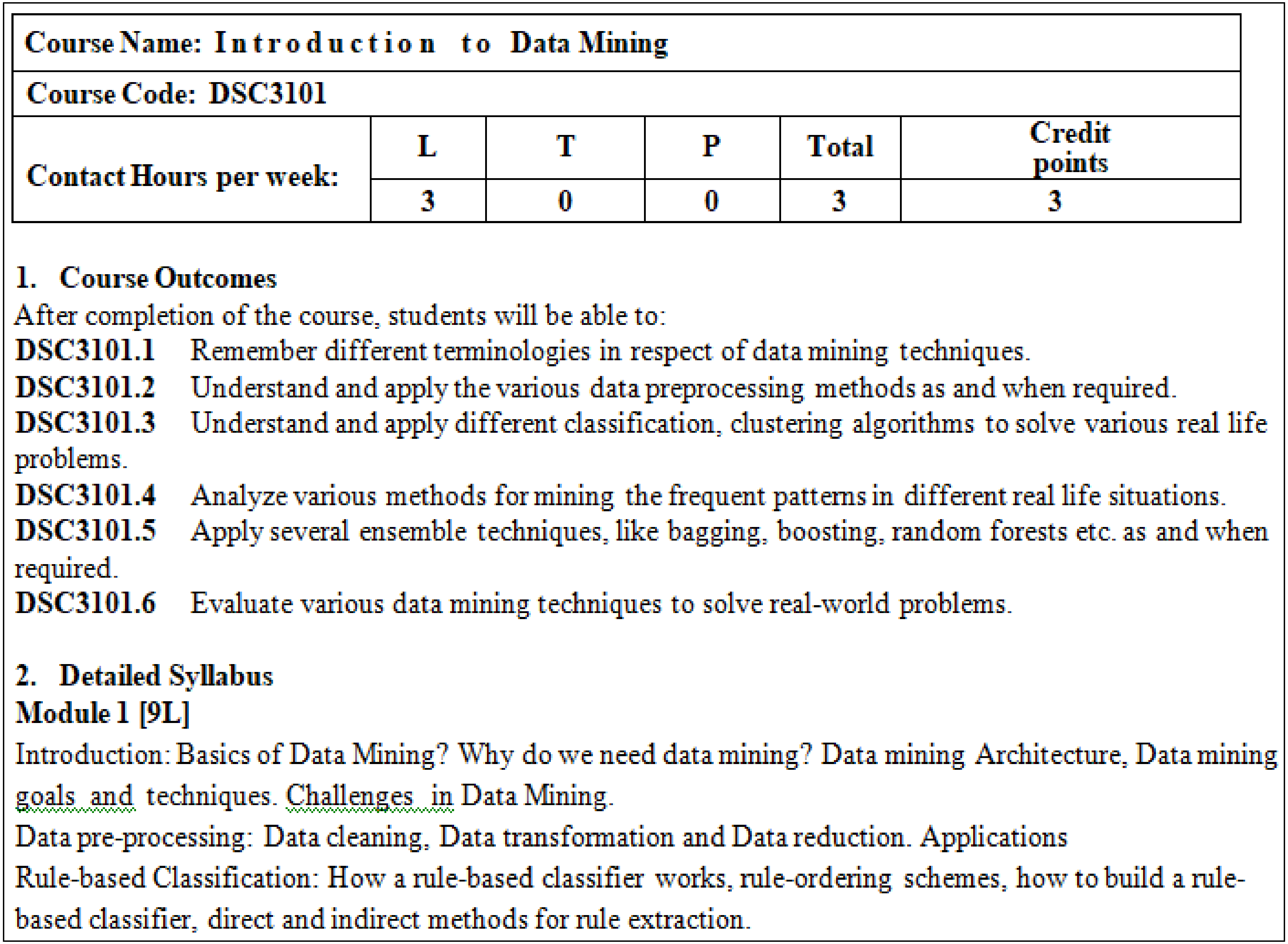
Module I - Introduction

Introduction to Data Mining (DSC3101) – CSE(DS) – 3rd yr (5th Sem) Prepared by Dr. Nilina Bera, CSE(DS)



Course Name: Introduction to Data Mining					
Course Code: DSC3101					
Contact Hours per week:	L	T	P	Total	Credit points
	3	0	0	3	3

1. Course Outcomes

After completion of the course, students will be able to:

DSC3101.1 Remember different terminologies in respect of data mining techniques.

DSC3101.2 Understand and apply the various data preprocessing methods as and when required.

DSC3101.3 Understand and apply different classification, clustering algorithms to solve various real life problems.

DSC3101.4 Analyze various methods for mining the frequent patterns in different real life situations.

DSC3101.5 Apply several ensemble techniques, like bagging, boosting, random forests etc. as and when required.

DSC3101.6 Evaluate various data mining techniques to solve real-world problems.

2. Detailed Syllabus

Module 1 [9L]

Introduction: Basics of Data Mining? Why do we need data mining? Data mining Architecture, Data mining goals and techniques. Challenges in Data Mining.

Data pre-processing: Data cleaning, Data transformation and Data reduction. Applications

Rule-based Classification: How a rule-based classifier works, rule-ordering schemes, how to build a rule-based classifier, direct and indirect methods for rule extraction.

Module 1: Introduction to Data Mining & Rule-Based Classification

1. Basics of Data Mining

❑Data mining is the process of discovering meaningful patterns from large datasets.

❑Part of the larger Knowledge Discovery in Databases (KDD) process.

❑Involves data cleaning, integration, selection, transformation, mining, pattern evaluation, and knowledge presentation.

2. Why Do We Need Data Mining?

❑To extract valuable insights from massive, unstructured data.

❑Enhances business decision-making, scientific research, and predictive analytics.

3. Data Mining Architecture

Data Sources: Databases, data warehouses, flat files, web data.

Data Warehouse Server: Performs data preprocessing.

Data Mining Engine: Core mining component (classification, clustering, etc.)

Pattern Evaluation Module: Evaluates interestingness of patterns.

User Interface: Allows user interaction and visualization.

4. Data Mining Goals and Techniques

Goals: Prediction, description, classification, clustering, pattern discovery.

Techniques: Classification, regression, clustering, association rule mining, anomaly detection.

5. Challenges in Data Mining

Handling noise and incomplete data.

Scalability to huge datasets.

Data heterogeneity.

Model interpretability and actionability.

Privacy and ethical issues.

6. Data Preprocessing

Cleaning: Handling missing values, noise removal.

Transformation: Normalization, encoding, discretization.

Reduction: Dimensionality reduction (PCA), data cube aggregation, sampling.

7. Applications

Healthcare diagnostics, fraud detection, customer segmentation, recommender systems, bioinformatics.

8. Rule-Based Classification

Uses IF-THEN rules for classification.

Rule ordering is critical: priority rules, conflict resolution.

Direct Methods: Generate rules directly from data (e.g., RIPPER).

Indirect Methods: Extract rules from other models (e.g., decision trees).

[it is a procedure]

What is Data Mining ?

Example :

❑ A bank wants to search new ways to increase revenues from its credit card operations.

❑ They want to check whether usage would double if fees were halved.

❑ Bank has multiple years of record on average credit card balances, payment amounts, credit limit usage, and other key parameters.

❑ They create a model to check the impact of the proposed new business policy.

❑ The data results show that cutting fees in half for a targeted customer base could increase revenues by Rs 10 crores.

Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

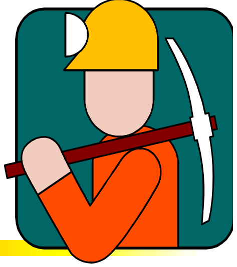
- Why data mining? Why we are motivated to explore data mining & KDD?
- There have been an explicit growth in the volume of data → Why? It's because of ease of storage, ease of data collection, more computerization of various companies and daily life.
- Thus, business like e-commerce, Amazon, Flipkart, Myntra whatever you use on the web, Google or any search engine, bank transactions, stock markets; We have various scientific areas like remote sensing images, biological data, scientific simulation various social media sites like news, digital photography, YouTube video etc.
- Thus, there is a huge volume of data, but it is the data is *not* giving us enough information that can be further processed to derive knowledge.
- *So, this gave rise to the method of automated analysis of such massive data to gather meaningful knowledge, and this gave birth to the subject of data mining. There is an alternate name to data mining which is knowledge discovery in databases*.

The world is data rich but information poor.

Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

What Is Data Mining?



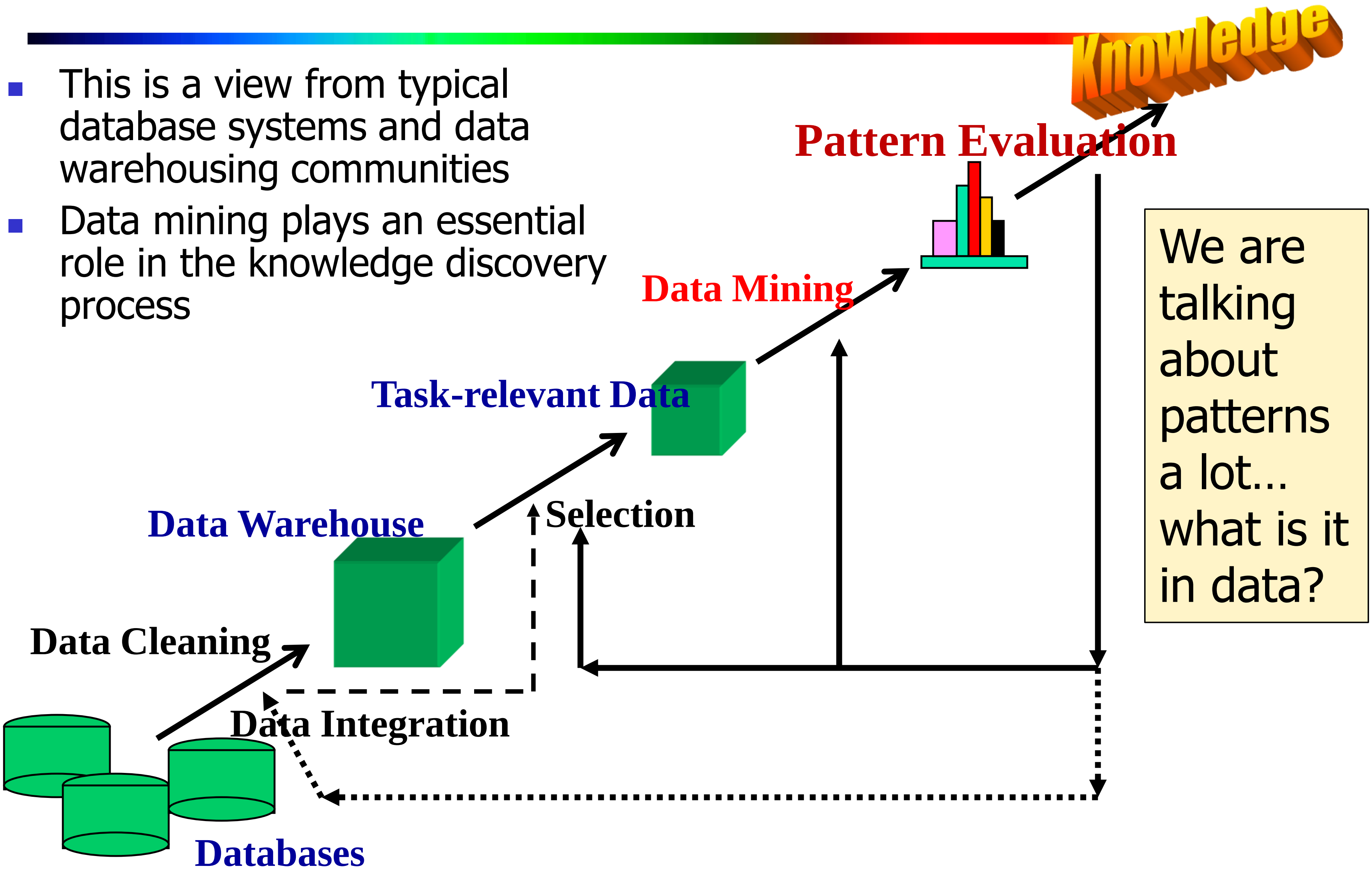
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

What Is Data Mining? Some interesting thoughts

- Definition : First it is a non-trivial process; which means that it is not obvious, i.e., the knowledge is not obvious it has to be extracted. It is implicit in the sense that the knowledge is inbuilt in the data you extract it only from the data. There is a novelty part which means that the knowledge has to be a new knowledge and unknown knowledge previously. And finally, it has to be potentially useful, i.e., this has to be useful knowledge depending on the application. So, the knowledge often takes the form of *patterns* in data, some regularity or some kind of structure in the data and from huge amount of data that is also an important aspect.
- Let's talk about what is *NOT* data mining → plain search like we do in Google or any search engine; similarly query processing in a RDBMS system, say transaction processing. So, for example, query how much balance you have in your account → these are not data mining. So, the plain querying in Google as well as say you are booking a ticket in Indian railway irctc site and you want to find out if reservations are available on a particular day → is obviously a query to the web-server and is not data mining.
- *Data mining would be from historical data sets.*

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



We are talking about patterns a lot… what is it in data?