MUST

Wisdom & Virtue

MIRPUR UNIVERSITY OF SCIENCE AND TECHNOLOGY (MUST), MIRPUR

DEPARTMENT OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY

DATA MINING

BCS-3605

Lecture 01

Dr Yasir Mehmood

(Assistant Professor)

COURSE DESCRIPTION

A Purpose of this course is to…

provide a comprehensive introduction to the data mining process; build theoretical and conceptual foundations of key data mining tasks such as data preprocessing, discovering the interested patterns from huge data set, classification and clustering techniques, other algorithms related to data mining (fuzzy logic, genetic algorithm , decision tree and neural network)

COURSE OBJECTIVES

***Objective** of this course is to…*

- understand what data mining is and what is not data mining with practical examples from real life
- to understand the knowledge discovery process to perform data mining
- to know about the types and characteristics of attributes and basic statistical description of data
- to know how to prepare data before mining
- what the frequent patterns are and how it can be mined from huge data.
- can differentiate classification, regression, clustering and outliers.
- data Mining Trends and Research Frontiers

COURSE SCHEDULE

Till Mid Term

- Concepts and basics of Data mining
- Getting to Know Your Data
- Data pre-processing and pre-mining,(noisy and missing data, data normalization and discretization)
- Outlier detection

After Mid Term

- Data mining learning methods
- Data mining classes (association rule mining, clustering, classification),
- Fundamental of other algorithms related to data mining (fuzzy logic, genetic algorithm and neural network),
- Decision trees, rules, patterns and trends

COURSE TEXTBOOKS

- *Data Mining: Concepts and Techniques, 3rd Edition Jiawei Han, Micheline Kamber, Jian Pei; , 2011*
- *Data Mning: Concepts, Models, Methods, and Algorithms, 2nd Edition, Mehmed Kantatardzic, 2011. (Reference Book)*
- *Data Mining, Introductory and Advanced Topics, 2006, Margaret H. Dunham and S. Sridhar, Pearson Education. (Reference Book)*
- *Principles of Data Mining, 2007, Max Bramer, Springer-Verlag. (Reference Book)*

GRADING POLICY

Mid Term: *[30%]*

Quizzes: *[10%]*

Assignments: *[10%]*

Final Exam: *[50%]*

HOMEWORK & ASSIGNMENT POLICY

- *Assignments will be on group bases.*
- *Each group have to prepare and submit assignment report in proper format.*
- *Do not copy past data in report, plagiarism policy will be strictly implemented.*
- *Late submission without justification and prior approval will not be entertained.*

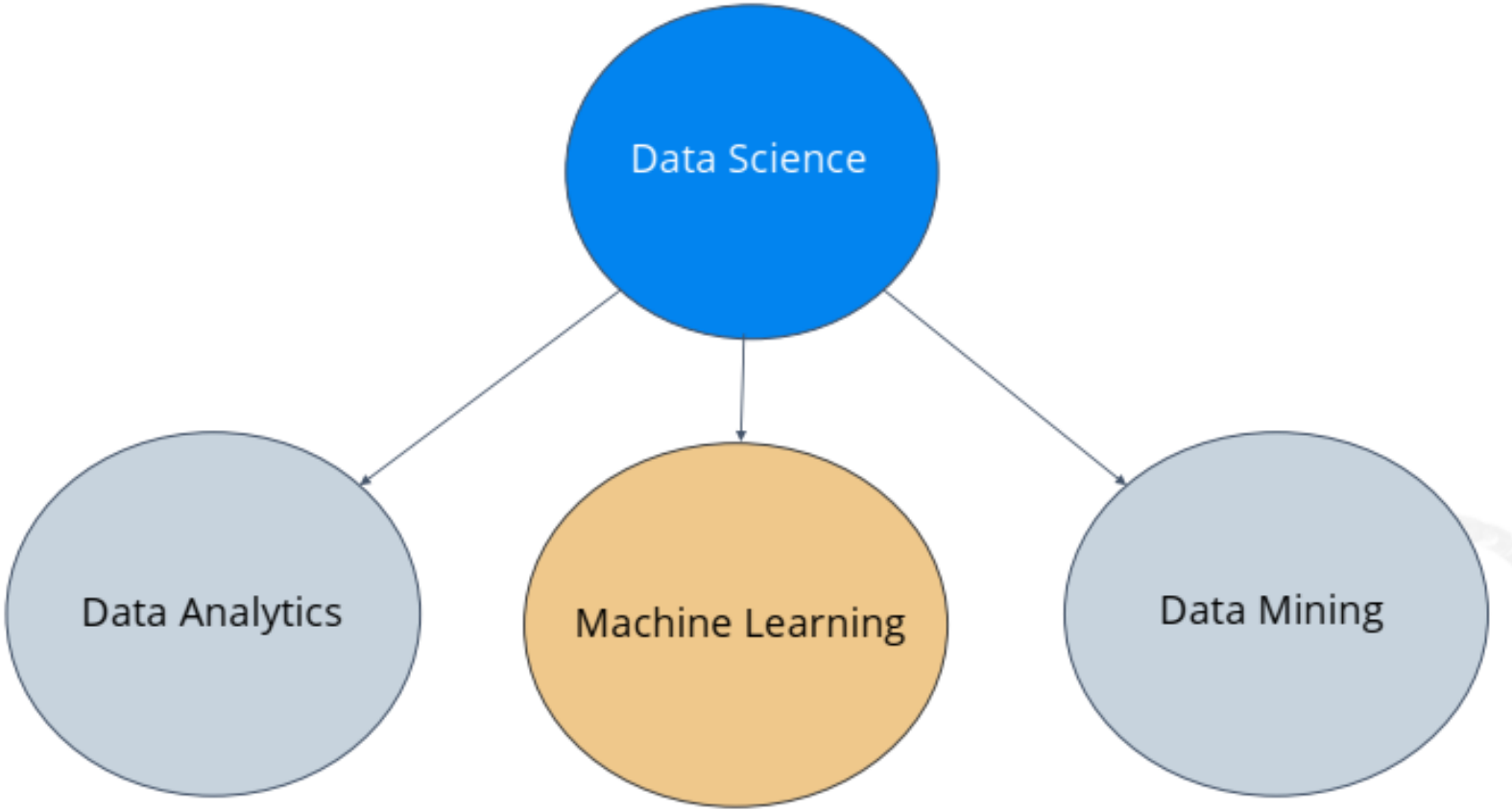
Agenda of Today's Lecture

- ***Data Science***
- ***Why Data Mining***
- ***Moving toward the Information Age***

Data Science

Data science is the study of data, which involves gathering, storing, analyzing, and plotting data, to effectively extract useful information.

Types of Data Science



Types of Data Science

Data Analytic

Data analytics is the process of examining and analyzing raw data sets to:

Draw conclusion

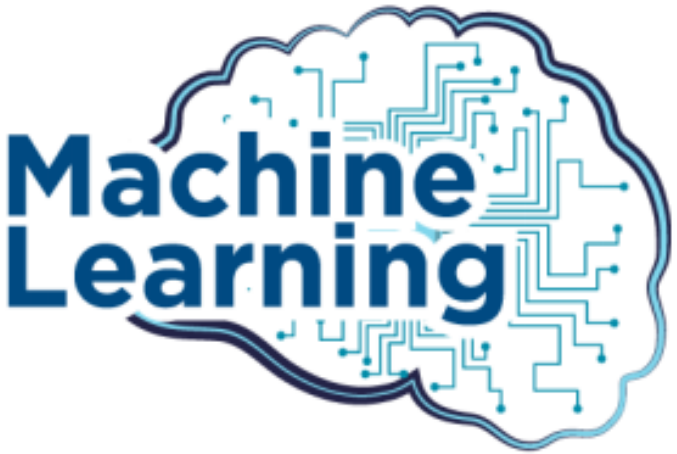
Derive information

Derive insights from raw data sources

Types of Data Science

Machine Learning

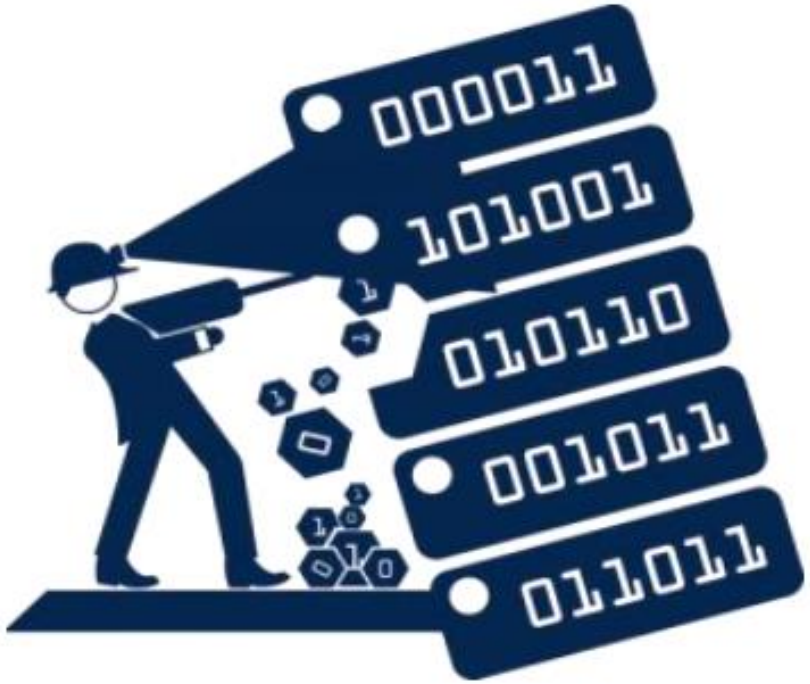
Learns from patterns in the past using a set of algorithms



Predicts outcomes accurately

Types of Data Science

Data Mining

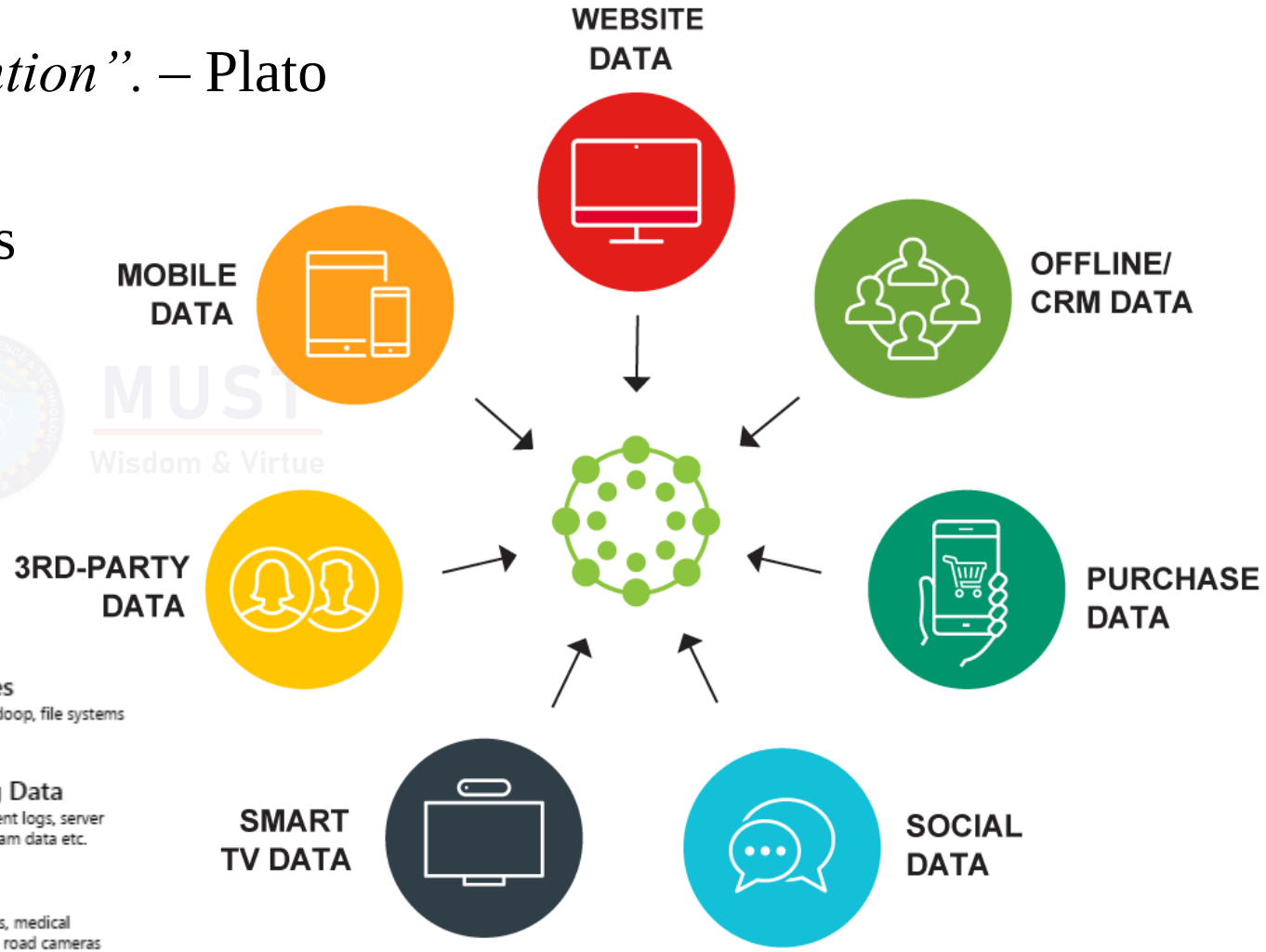


- Data mining is the process of analyzing data from different perspectives.
- It summarizes data into useful information.
- It helps increase revenue and cut costs.

Why Data Mining

"Necessity, who is the mother of invention". – Plato

We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need.



Archives
Scanned documents, statements, medical records, e-mails etc..

Docs
XLS, PDF, CSV, HTML, JSON etc.

Business Apps
CRM, ERP systems, HR, project management etc.

Media
Images, video, audio etc.

Social Networks
Twitter, Facebook, Google+, LinkedIn etc.

Public Web
Wikipedia, news, weather, public finance etc

Data Storages
RDBMS, NoSQL, Hadoop, file systems etc.

Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

Sensor Data
Smart electric meters, medical devices, car sensors, road cameras etc.

Why Data Mining Cont'd

- The New York Stock Exchange generates about one terabyte of new trade data per day.
- The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day.
- A single Jet engine can generate 10+terabytes of data in 30 minutes of flight time.

Why Data Mining Cont'd

Why Data Mining Cont'd

We are actually living in the data age.

The explosive growth of data: from terabytes to petabytes into CN, WWW, storage device, satellite data, business transactions, health care, science and engineering, medicine, and almost every other aspect of daily life.

- ***Businesses:*** worldwide generate gigantic data sets, including
  - sales transactions,
  - stock trading records,
  - product descriptions,
  - sales promotions,
  - company profiles and performance, and customer feedback

Why Data Mining Cont'd

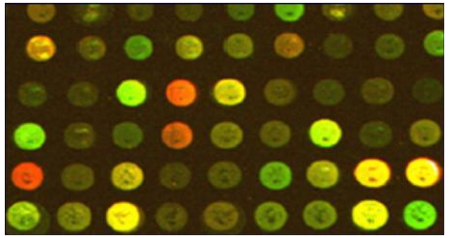
- ***Scientific and engineering practices:*** generate high orders of petabytes of data in a continuous manner, from
  - remote sensing,
  - process measuring,
  - scientific experiments,
  - system performance,
  - engineering observations,
  - simulation and
  - surveillance.

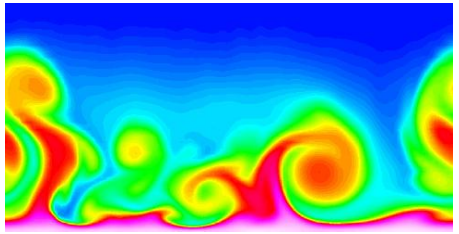
Why Data Mining Cont'd

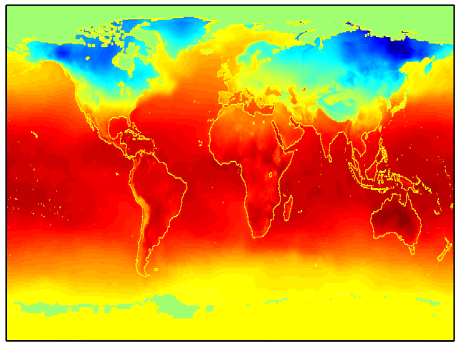
- ***Telecommunication networks:*** carry tens of petabytes of data traffic every day
- ***Medical and health industry:*** generates tremendous amounts of data from
  - medical records,
  - patient monitoring, and
  - medical imaging.
- ***Web searches engines:*** process tens of petabytes of data daily.

  Social media: become increasingly important data sources
  - Pics, videos,
  - blogs, social media
- The list of sources that generate huge amounts of data is endless.

Why Data Mining Cont'd







Moving toward the Information Age

We are drowning in data but starving for knowledge!

- Traditional techniques are infeasible for raw data
- Now computers have become more powerful and cheaper

Welding gap between data in information- Data Mining

- Automated analysis of massive data sets
- Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data mining may help scientists

- In classifying and segmenting data
- In Hypothesis Formation

Moving toward the Information Age Cnot'd

Data mining turns a large collection of data into knowledge. A search engine (e.g., Google) receives hundreds of millions of queries every day. Each query can be viewed as a transaction where the user describes her or his information need. What novel and useful knowledge can a search engine learn from such a huge collection of queries collected from users over time? Interestingly, some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone. For example, Google's *Flu Trends* uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms. A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, *Flu Trends* can estimate flu activity up to two weeks faster than traditional systems can.[2] This example shows how data mining can turn a large collection of data into knowledge that can help meet a current global challenge. ■

Reference

Data Mining Concepts and Techniques Third Edition

1.1 Why data mining

ASSIGNMENT #01

How a search engine (e.g.,Google) responds to our query in nano second with millions of related web pages?

Due date: before the start of next class.

THANKS