CHANDIGARH UNIVERSITY

Discover. Learn. Empower.

Apex Institute of Technology

Department of Computer Science & Engineering

Introduction To Data Science (CST-292)

Dr. Jitender Kaushal
Associate Professor
(E14621)
AIT-CSE, CU

DISCOVER . LEARN . EMPOWER

Syllabus

Unit-1		Contact Hours:15 hours
	Big data why and where, Types of Big data , problem with traditional databases , Characteristics of Big data ,Basic Architecture of Big Data, Application of Big Data ,Business profiles in Big data	
Unit-2		Contact Hours: 15 hours
	Advantages of Big Data processing , Data life cycle (Business Understanding, Data Understanding, Data Preparation , Modeling , Evaluation , Deployment) ,Big data analytics - methodology ,introduction to data warehousing and data mart , Difference between ETL and ELT , OLAP and OLTP, Types of Analytics(prescriptive, predictive, descriptive)with example,	
Unit-3		Contact Hours: 15 hours
	Technologies for Handling Big Data, Understanding Hadoop Ecosystem, Architecture of HDFS	

Introduction to Data Science: Course Objectives

COURSE OBJECTIVES

The Course aims to:

- This course brings together several key big data problems and solutions.
- To recognize the key concepts of Extraction, Transformation, and Loading
- To prepare a sample project in the Hadoop Environment

("An open-source software framework that supports the processing and storage of extremely large datasets in a distributed computing environment.")

COURSE OUTCOMES

On completion of this course, the students shall be able to:-

CO1	Identify and describe the importance of Big data analysis over Conventional Database management System.

Contents to be Covered

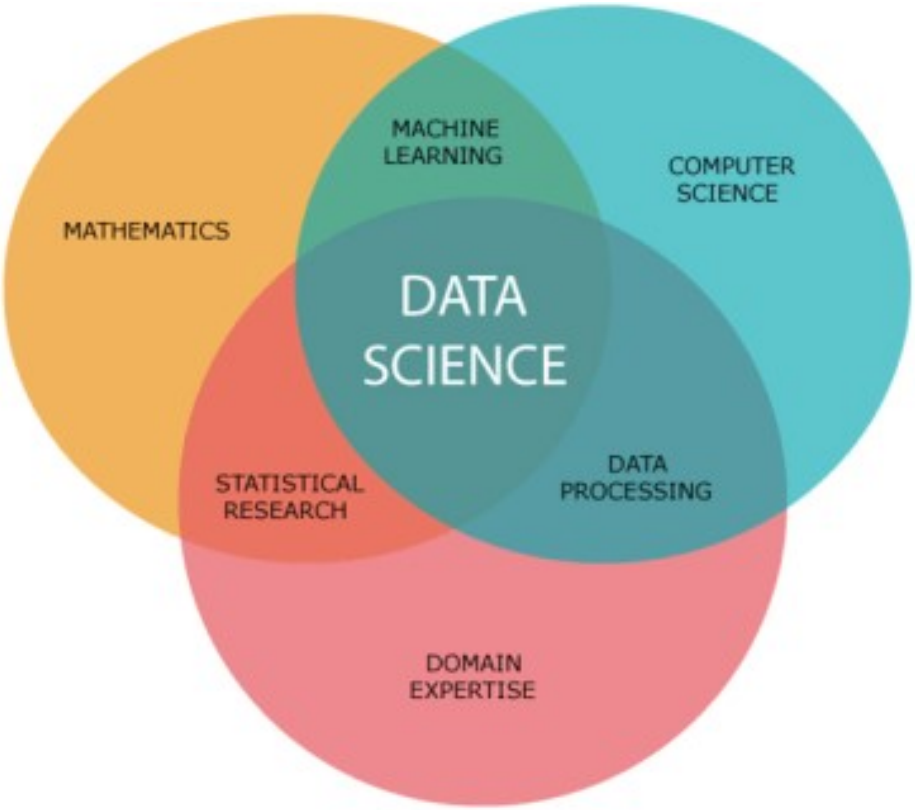
- What are Data and Information?
- Difference between Data and Information.
- Define Data Science
- Define Traditional Data and Big Data
- Why is big data analytics important?

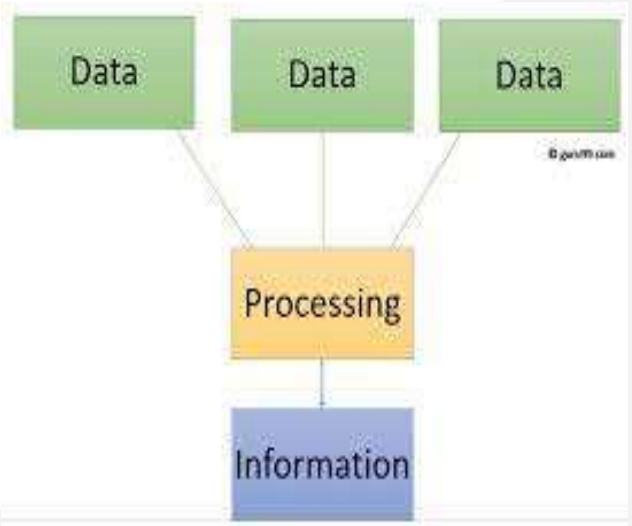
Available on: https://data.gov.in/

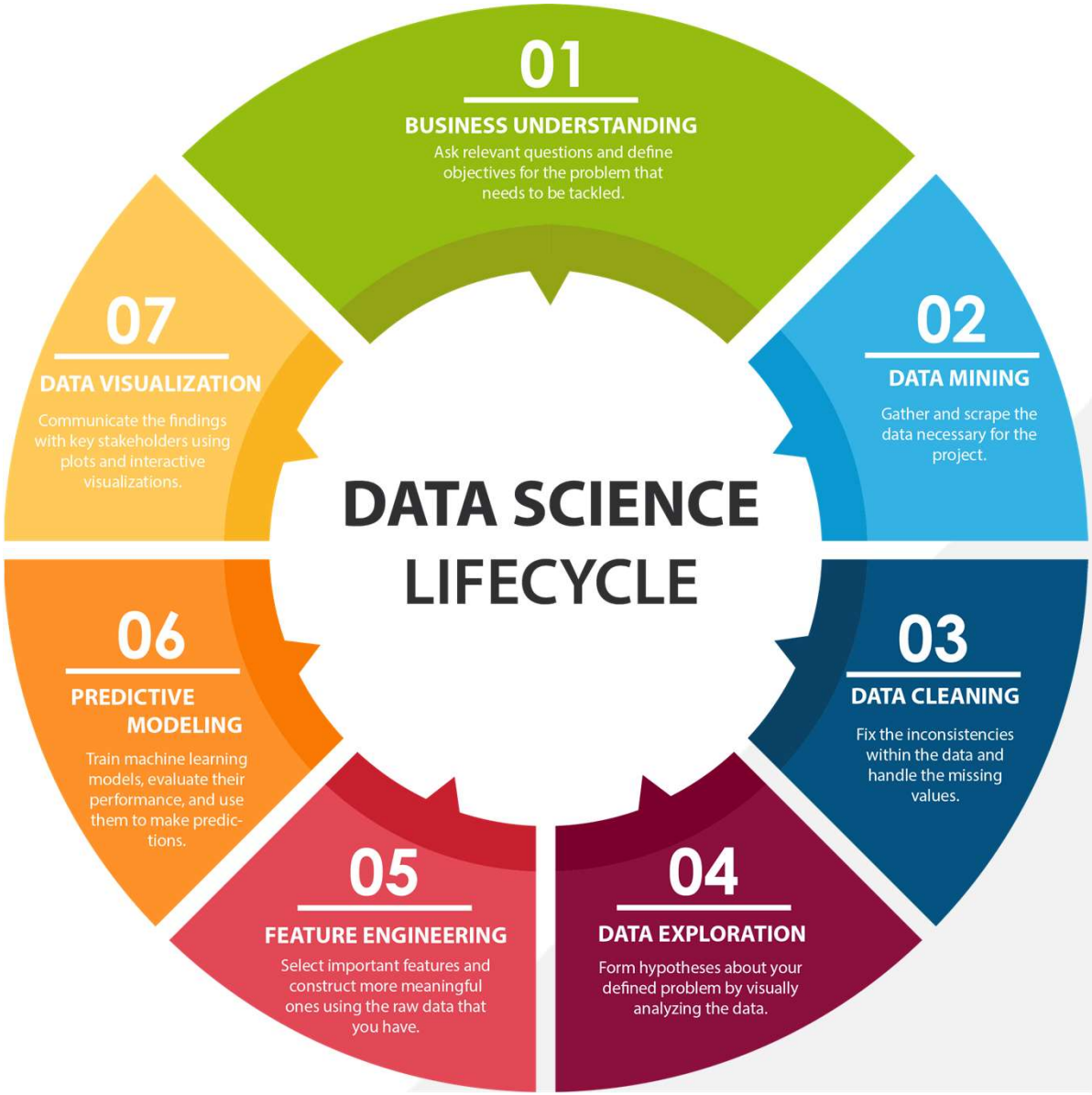
What Is Data and Information?

- **Data:** Data can come in the form of **text, observations, figures, images, numbers, graphs, or symbols**. For example, data might include individual prices, weights, addresses, ages, names, temperatures, dates, or distances. Data is a raw form of knowledge and, on its own, doesn't carry any significance or purpose.
- **Information:** Information is defined as classified or organized data that has some meaningful value for the user. Information is also the processed data used to make decisions and take action. Processed data must meet the following criteria for it to be of any significant use in decision-making:
- Accuracy: The information must be accurate.
- Completeness: The information must be complete.
- Timeliness: The information must be available when it's needed.

Diagram:


MACHINE LEARNING
COMPUTER SCIENCE
MATHEMATICS
DATA SCIENCE
STATISTICAL RESEARCH
DATA PROCESSING
DOMAIN EXPERTISE


Data
Data
Data
Processing
Information


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE

The Key Differences Between Data vs Information

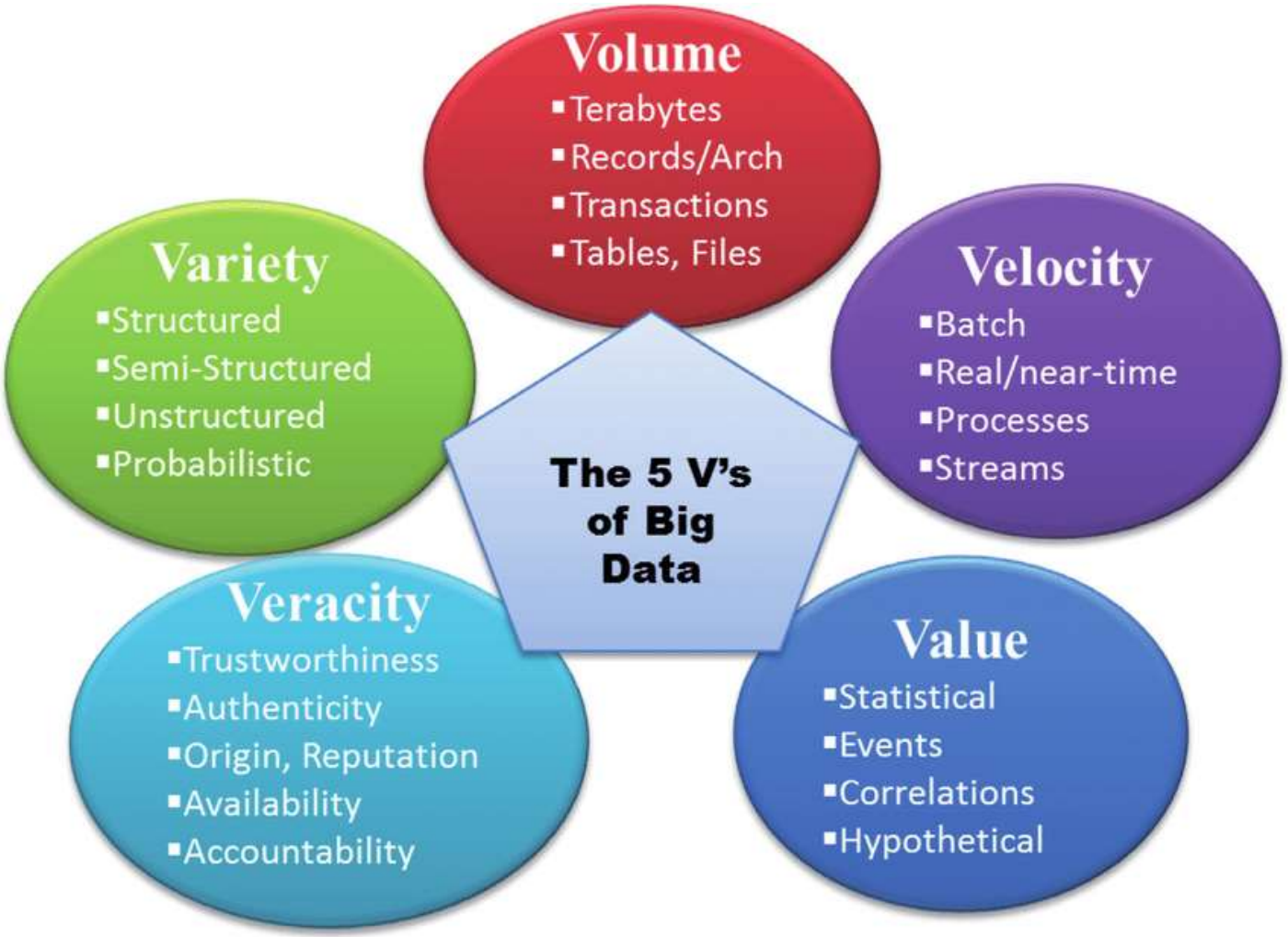
- Data is a collection of facts, while information puts those facts into context.
- While data is raw and unorganized, information is organized.
- Data points are individual and sometimes unrelated. Information maps out that data to provide a big-picture view of how it all fits together.
- Data, on its own, is meaningless. When it's analyzed and interpreted, it becomes meaningful information.
- Data does not depend on information; however, information depends on data.
- Data typically comes in the form of graphs, numbers, figures, or statistics. Information is typically presented through words, language, thoughts, and ideas.
- Data isn't sufficient for decision-making, but you can make decisions based on information.

What Is Data Science?

- Data science is a field that deals with unstructured, structured data, and semi-structured data. It involves practices like data cleansing, data preparation, data analysis, and much more.
- Data science is the combination of: statistics, mathematics, programming, and problem-solving, the ability to look at things differently; and the activity of cleansing, preparing, and aligning data. This umbrella term includes various techniques that are used when extracting insights and information from data.

What is Traditional Data & Big Data?

- **Traditional data:** Traditional data is the structured data that is being majorly maintained by all types of businesses starting from very small to big organizations. In a traditional database system, a centralized database architecture is used to store and maintain the data in a fixed format or fields in a file. For managing and accessing the data Structured Query Language (SQL) is used. [Structured Query Language is a standard Database language which is used to create, maintain and retrieve the relational database.]
- **Big data:** We can consider big data as the upper version of traditional data. Big data deals with too large or complex data sets which is difficult to manage in traditional data-processing application software. It deals with large volumes of both structured, semi-structured and unstructured data. Volume, Velocity and Variety, Veracity and Value refer to the 5'V characteristics of big data. Big data not only refers to large amount of data it refers to extracting meaningful data by analyzing the huge amount of complex data sets. semi-structured



Real-time is instant, whereas near-time is delayed (whether that's by a few milliseconds or a few hours).

The difference between Traditional data and Big data are as follows:

Traditional Data	Big Data
• Traditional data is generated in enterprise level.	• Big data is generated outside the enterprise level.
• Its volume ranges from Gigabytes to Terabytes.	• Its volume ranges from Petabytes to Zettabytes or Exabytes.
• Traditional database system deals with structured data.	• Big data system deals with structured, semi-structured, database, and unstructured data.
• Traditional data is generated per hour or per day or more	• But big data is generated more frequently mainly per seconds.
• Traditional data source is centralized and it is managed in centralized form.	• Big data source is distributed and it is managed in distributed form.

Structured data is usually easier to search and use, while unstructured data involves more complex search and analysis.

Multiples of Bits		
Unit (Symbol)	**Value (SI)**	**Value (Binary)**
Kilobit (Kb) (Kbit)	10^{3}	2^{10}
Megabit (Mb) (Mbit)	10^{6}	2^{20}
Gigabit (Gb) (Gbit)	10^{9}	2^{30}
Terabit (Tb) (Tbit)	10^{12}	2^{40}
Petabit (Pb) (Pbit)	10^{15}	2^{50}
Exabit (Eb) (Ebit)	10^{18}	2^{60}
Zettabit (Zb) (Zbit)	10^{21}	2^{70}
Yottabit (Yb) (Ybit)	10^{24}	2^{80}

The difference between Traditional data and Big data are as follows:

Traditional Data (Small Data)

- Data integration is very easy.
- Normal system configuration is capable to process traditional data.
- The size of the data is very small.
- Traditional data base tools are required to perform any data base operation.
- Normal functions can manipulate data.

Big Data

- Data integration is very difficult.
- High system configuration is required to process big data.
- The size is more than the traditional data size.
- Special kinds of database tools are required to perform any database schema-based operation.
- Special kinds of functions can manipulate data.

The difference between Traditional data and Big data are as follows:

Traditional Data

- Traditional data is stable and inter relationship.
- Traditional data is in manageable volume.
- It is easy to manage and manipulate the data.
- Its data sources includes ERP transaction data (Enterprise Resource Planning), CRM transaction data (Customer relationship management), financial data, organizational data, web transaction data etc.

Big data

- Big data is not a stable and unknown relationship.
- Big data is in huge volume which becomes unmanageable.
- It is difficult to manage and manipulate the data.
- Its data sources include social media, device data, sensor data, video, images, audio, etc.

Why is big data analytics important?

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, **leads to smarter business moves**, **more efficient operations**, **higher profits** and **happier customers**. Businesses that use big data with advanced analytics gain value in many ways, such as:

- **Reducing cost.** Big data technologies like cloud-based analytics can significantly reduce costs when it comes to storing large amounts of data (for example, a data lake). Plus, big data analytics helps organizations find more efficient ways of doing business.
- **Making faster, better decisions.** The speed of in-memory analytics – combined with the ability to analyze new sources of data, such as streaming data from IoT – helps businesses analyze information immediately and make fast, informed decisions.
- **Developing and marketing new products and services.** Being able to gauge customer needs and customer satisfaction through analytics empowers businesses to give customers what they want, when they want it. With big data analytics, more companies have an opportunity to develop innovative new products to meet customers' changing needs.

Key points of the lecture

- Data and Information?
- Difference between Data and Information.
- Define Data Science.
- Define Traditional Data and Big Data.
- Why is big data analytics important?

Homework

- Differentiate between Traditional Data and Big Data.
- Explain Overfitting in Big Data.
- What are the different big data processing techniques?

Thank you

E-Mail: jitender.e14621@cumail.in