Introduction of Data Science

BY:

DR. VIPIN KUMAR

DEPARTMENT OF COMPUTER SCIENCE &IT

MAHATMA GANDHI CENTRAL UNIVERSITY

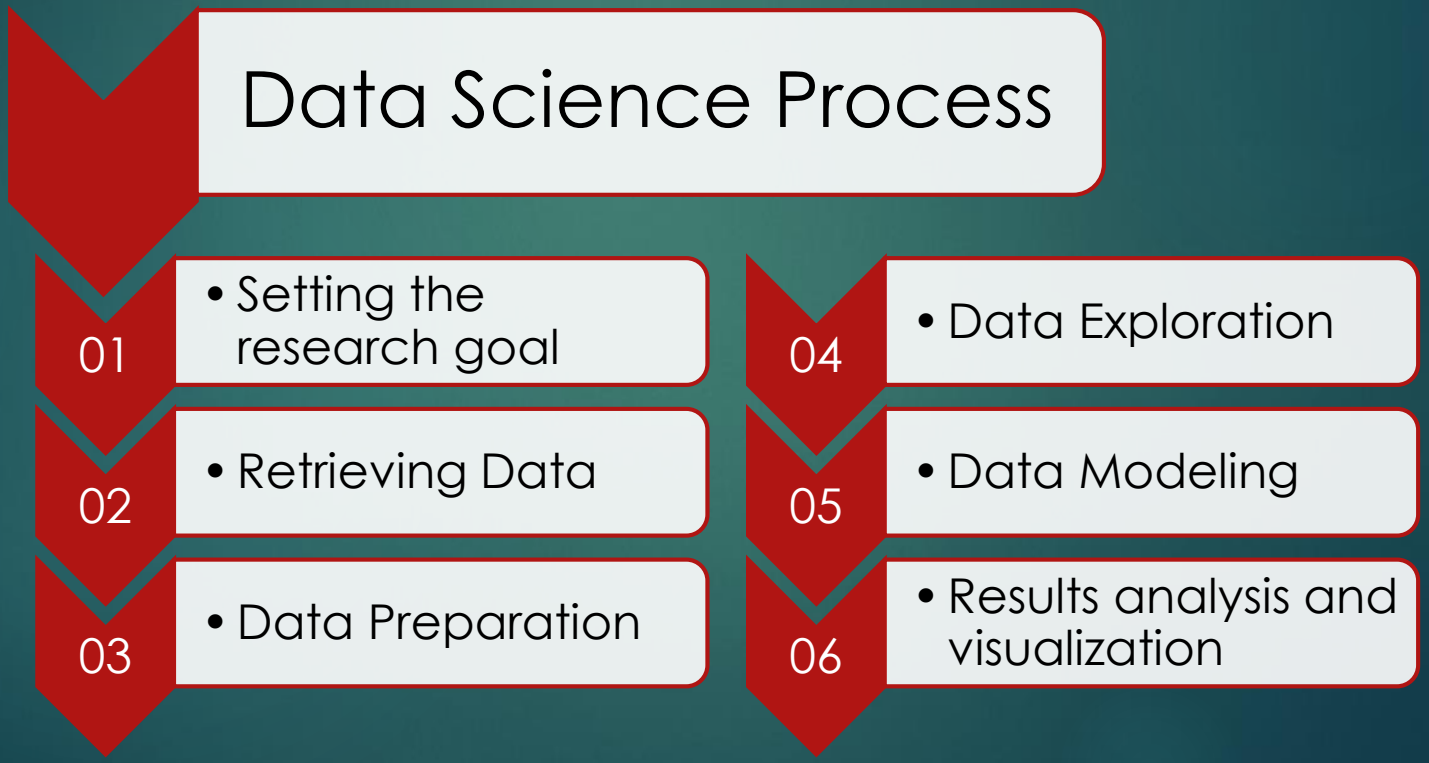
MOTIHARI, BIHAR

Outline…

- DEFINING THE DATA SCIENCE
- UNDERSTANDING OF TYPES OF DATA
- DATA SCIENCE PROCESS
- DATA SCIENCE IN PERSPECTIVE OF BIG DATA
- APPLICATION OF DATA SCIENCE

Defining Data Science

- **A PROCESS OF FINDING THE KNOWLEDGE (HIDDEN PATTERN) FROM THE RAW DATA USING PRINCIPLE OF MACHINE LEARNING, ALGORITHMS AND VARIOUS TOOLS.**



1. Setting the Research Goal

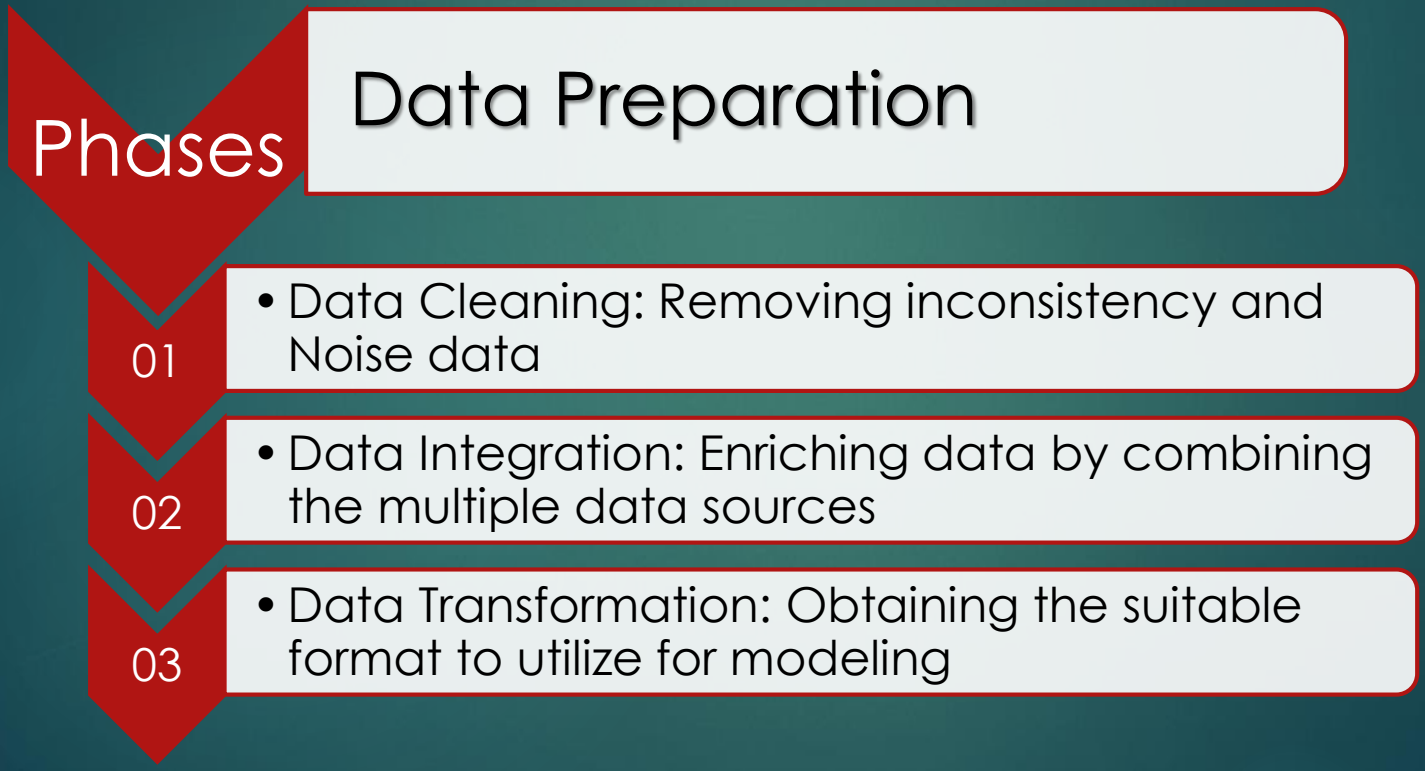
- DATA SCIENCE RESEARCH GOAL IS MOSTLY OBTAIN AS PER ORGANIZATION REQUIREMENT.
- PREPARING THE CHARTER WITH SOME MAJOR QUESTIONS AND THEIR ANSWERS AS:
  - What is going to research?
  - How the organization will get benefit from it?
  - What are the resources and data required?
  - What are the time table and deliverable?

2. Retrieving the Data

- DATA COLLECTION IS THE SECOND STEPS OF DATA SCIENCE PROCESS.
- COLLECTING THE REQUIRED DATA AS PER PROJECT CHARTER BY CHECKING THE DATA EXISTENCE, ACCESS, AND QUALITY WITHIN AND OUTSIDE OF THE ORGANIZATION.
- DEALING WITH DIFFERENT TYPES OF DATA FORMAT AND DATABASE.
- ACCESSING THE THIRD PARTY RESOURCE TO ENRICH THE QUALITY OF INFORMATION.

3. Data Preparation

- **PREPARING A GOOD QUALITY OF DATA IN REQUIRED FORMAT USING COMMON AND DOMAIN SPECIFIC PREPROCESSING STEPS.**



4. Data Exploration

- UNDERSTANDING THE DATA USING STATISTICAL ANALYSIS AND VISUALIZATION.
- DETECTING THE NOISE AND OUTLIERS.
- UNDERSTANDING THE VARIABLE INTERACTIONS.
- TRYING TO SENSE THE DISTRIBUTION OF THE DATA
- THIS STEP SPECIALLY KNOWN AS EXPLORATORY DATA ANALYSIS (EDA).

5. Building the Model

- THIS STEP USE THE PREVIOUS EXPERIENCES OF THE DOMAIN TO BUILD THE MODELS.
- WHILE BUILDING THE MODEL, IT UTILIZES THE STATISTICS, OPERATION RESEARCH METHODS, OPTIMIZATION AND MACHINE LEARNING ALGORITHMS.
- IN ITERATIVE PROCESS, HYPERPARAMETER TUNING IS DONE FOR SELECTING THE FINAL MODEL.
- FINAL MODEL GOT SELECTED BASED ON PERFORMANCE OF MODEL ON VALIDATION SET OF THE DATA.

5. Result Analysis and Visualization

- THIS STEPS INVOLVES THE RESULTS ANALYSIS AND VISUALIZATION
- THERE ARE TWO WAY TO ANALYZE THE RESULTS
  - Quantitative measures
  - Graphical measures
  - Statistical measures
- SOME TIME, IT IS IMPORTANT TO VISUALIZE THE RESULTS DYNAMICALLY THAT SHOWS THE REAL TIME BEHAVIOR OF RESULTS.
- BUSINESS INTELLIGENCE TOOLS ARE UTILIZED FOR VISUALIZATION OF RESULTS LIKE: MICROSOFT POWER BI, TABLEAU DESKTOP, GOOGLE CHART, MICROSOFT BI ETC.

Data Science vs Business Intelligence

- **DATA SCIENCE:**
  1. It analyze the previous and current experiences for predicting the future outcome.
  2. It has informed decision.
  3. "What" and "How" types question can be answered through data science.
- **BUSINESS INTELLIGENCE:**
  1. It analyze the historical data to find the trends.
  2. It provide the dashboard to answer the queries.
  3. It integrates the multiple sources to run the queries based on business question.

Data Science vs Data Analytics

- DATA SCIENCE HAVE THREE PRIMARY OBJECTIVES:
  1. Exploratory data analysis
  2. Machine Learning
  3. Data Product Engineering
- DATA ANALYTICS HAS TWO OBJECTIVES:
  1. Business administration
  2. Exploratory data analysis

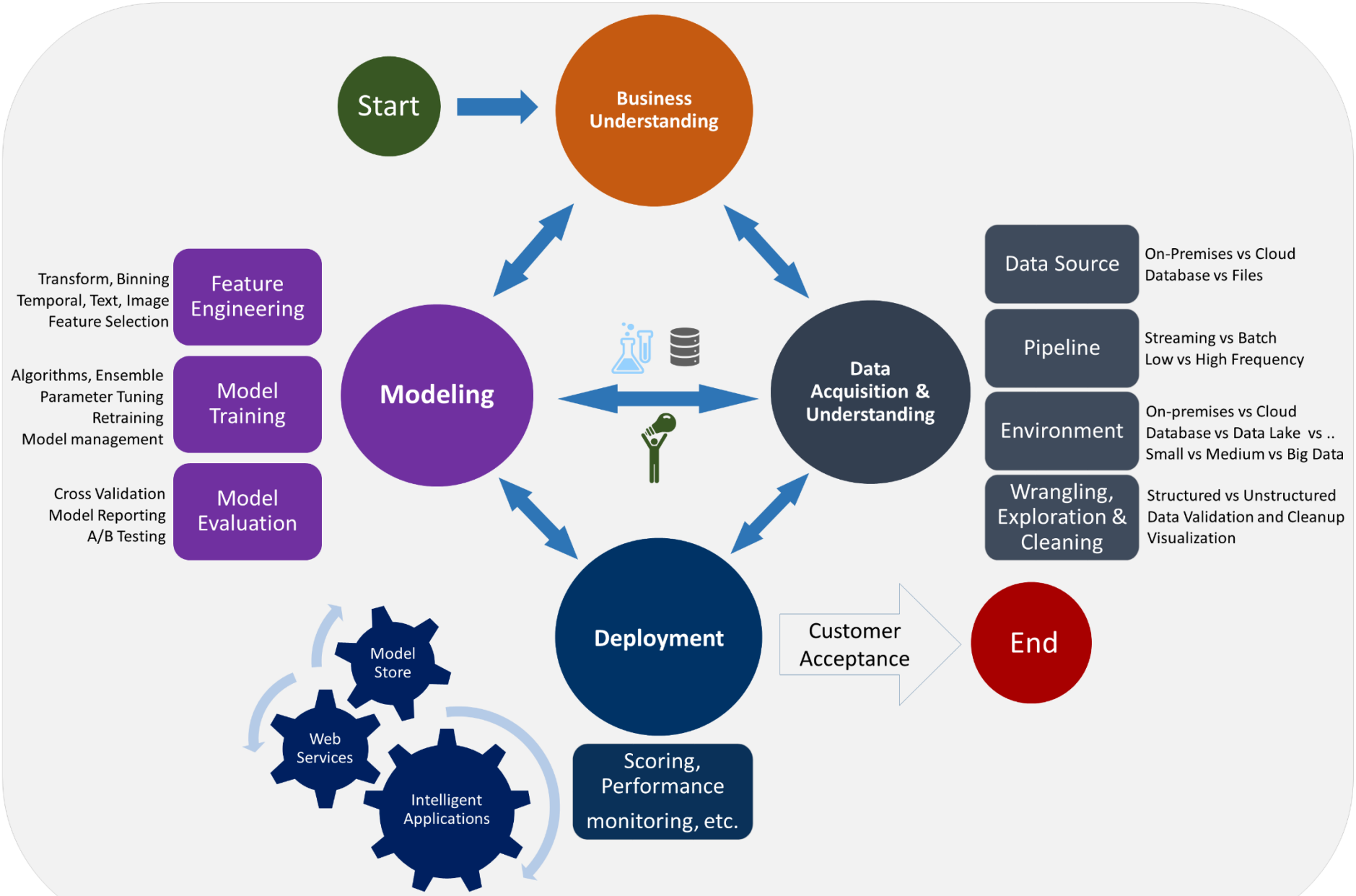
Data Science vs Business Intelligence

- **DATA SCIENCE:**
  1. It analyze the previous and current experiences for predicting the future outcome.
  2. It has informed decision.
  3. "What" and "How" types question can be answered through data science.
- **BUSINESS INTELLIGENCE:**
  1. It analyze the historical data to find the trends.
  2. It provide the dashboard to answer the queries.
  3. It integrates the multiple sources to run the queries based on business question.

Life Cycle of Data Science

- **THERE ARE FIVE LIFE CYCLE STAGES:**
  1. Business understanding
  2. Data acquisition
  3. Modeling of machine learning algorithm
  4. Deployment in the environment
  5. Customer acceptance

Data Science Lifecycle



Reference: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

Application of Data Science

- **FINANCE:** IT AUTOMATED FINANCE INDUSTRIES AND RISK ANALYTICS WITH KEY ROLL OF ALGORITHMIC TRADING. DATA SCIENCE ANALYSES THE COSTUMER BEHAVIOR BY BOOSTING SOCIAL MEDIA INTERACTION USING SENTIMENT ANALYSIS.
- **TRANSPORT:** IT HELPS TO PROVIDE BETTER EXPERIENCE WITH LOW CAST TO THE CUSTOMER FOR TRANSPORTATION COMPANIES. IT ALSO HELPS TO PROVIDE THE SELF-DRIVING CAR WITH SAFETY.
- **HEALTHCARE:** .DATA SCIENCE HAS BEEN UTILIZED IN MANY AREA OF HEALTH CARE LIKE DRUGS DISCOVERY, MEDICAL IMAGE ANALYSIS, DIAGNOSIS OF DISEASES WITH PREDICTIVE MODELING, ETC.
- **BANKING:** PROVIDE THE BETTER DECISION CAPABILITY FOR FRAUD DETECTION, RISK MODELING, MANAGEMENT OF CUSTOMER DATA, CUSTOMER SEGMENTATION, REAL-TIME PREDICTIVE ANALYTICS, ETC.
- **MANUFACTURING:** IT PROVIDE THE CONTINUOUS MONITORING SYSTEM USING INTERNET OF THINGS DEVICE WHICH HELPS TO REDUCE THE COST, PRODUCT OPTIMIZATION AND ENHANCING THE PROFITS.

Bibliography

- O'NEIL, CATHY, AND RACHEL SCHUTT. *DOING DATA SCIENCE: STRAIGHT TALK FROM THE FRONTLINE*. " O'REILLY MEDIA, INC.", 2013.
- CIELEN, DAVY, ARNO MEYSMAN, AND MOHAMED ALI. *INTRODUCING DATA SCIENCE: BIG DATA, MACHINE LEARNING, AND MORE, USING PYTHON TOOLS*. MANNING PUBLICATIONS CO., 2016.
- MITCHELL, TOM M. "MACHINE LEARNING." (1997).
- PROVOST, FOSTER, AND TOM FAWCETT. *DATA SCIENCE FOR BUSINESS: WHAT YOU NEED TO KNOW ABOUT DATA MINING AND DATA-ANALYTIC THINKING*. " O'REILLY MEDIA, INC.", 2013.
- AYODELE, TAIWO OLADIPUPO. "INTRODUCTION TO MACHINE LEARNING." *NEW ADVANCES IN MACHINE LEARNING* (2010): 1-9.
- HTTPS://WWW.EDUREKA.CO/BLOG/WHAT-IS-DATA-SCIENCE/
- HTTPS://DOCS.MICROSOFT.COM/EN-US/AZURE/MACHINE-LEARNING/TEAM-DATA-SCIENCE-PROCESS/LIFECYCLE
- HTTPS://WWW.DEZYRE.COM/ARTICLE/LIFE-CYCLE-OF-A-DATA-SCIENCE-PROJECT/270
- HTTPS://DATA-FLAIR.TRAINING/BLOGS/DATA-SCIENCE-APPLICATIONS/

Thank you