JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

Programme: BCA

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

MODULE 1: Introductory Concepts

Dr. Ananta Charan Ojha, Professor

Session -0

- Course Information
- Trends in Computer Applications

Introduction to Data Analytics (22BCAD4C01)

Course Objectives

COB1:	Provide knowledge on basics of data science, data analytics and its process.
COB2:	Make students understand various techniques for data pre-processing, and exploratory data analysis.
COB3:	Familiarize machine learning techniques for model building and solving real-world problems.

Introduction to Data Analytics (22BCAD4C01)

Course Outcomes: After completion of this course student will be able to:

Sl. No.	Course Outcome	Description	Bloom's Taxonomy Level
1.	CO1	Describe data science and analytics, data science process and its applications.	L2
2.	CO2	Choose various techniques to prepare data for analysis.	L3
3.	CO3	Examine data using EDA	L4
4.	CO4	Develop models using machine learning techniques and modelling process.	L6
5.	CO5	Assess models using techniques of model evaluation and selection.	L5

Course Contents

Module	Details	Contact Hours
I	**Introductory Concepts** Overview of Data Science and Data Analytics, Types of Analytics: Descriptive, Diagnostics, Predictive and Prescriptive; Data Ubiquity, Nature of Data: Structured, Unstructured, Big Data; Advantages of Data-Driven Decisions, Data Science Process, Applications of Data Science in various fields, Data Science Roles, Data Security, Privacy, and Ethical Issues.	12
II	**Data Preparation** Data Collection Methods: Primary, Secondary data; Pre-Processing: Data Cleaning, Data Integration and Transformation, Data Discretization; Dimensionality Reduction, PCA, Feature Engineering and Selection.	12

Course Contents

III	**Exploratory Data Analysis** Exploratory Data Analysis (EDA) Descriptive Statistics: Mean, Standard Deviation, Skewness and Kurtosis; Types of EDA: Univariate, Bi-Variate, Multi-Variate; Visualizing EDA: Histogram, Scattered Plot, Box Plots, Pivot Table, Heat Map; Correlation Statistics, Statistical Significance and Hypothesis Testing, ANOVA.	**12**
IV	**Machine Learning and Model Building** Machine Learning in Data Science, Types of Machine Learning, Supervised, Unsupervised, Modelling Process, Simple and Multiple Regression, Time Series Analysis, Classification, Prediction, Clustering, Decision Tree, k-Nearest Neighbour, Association Rules Mining.	12
V	**Model Evaluation** Model Accuracy, Overfitting, Under Fitting, Bias-Variance Trade-off, Evaluation Metrics, Confusion Metrix, Methods for evaluating Accuracy, Cross Validation, Model Evaluation using Visualization: Residual Plot, Distribution Plot, ROC curve, Model Selection.	12

Text Books, References

Text Book (TB):

1. Vijay Kotu and Bala Deshpande, (2019). Data Science: Concepts and Practice, Morgan Kaufmann Publications.
2. João Moreira, Andre Carvalho, Tomás Horvath, (2019). A General Introduction to Data Analytics, John Wiley & Sons.

Reference Books (RB):

1. Davy Cielen, Arno D. B. Meysman, Mohamed Ali, (2016). Introducing Data Science, by Manning Publications.
2. Foster Provost and Tom Fawcett, (2013). Data Science for Business, by O'Reilly.
3. EMC Education Services (Editor), (2015). Data Science and Big Data Analytics, Wiley.
4. Rachel Schutt and Cathy O'Neil, (2014). Doing Data Science, O'Reilly.
5. Peter Bruce and Andrew Bruce, (2017). Practical Statistics for Data Scientists, O'Reilly.

Web Video Links:

1. https://www.linkedin.com/learning/data-science-foundations-fundamentals-5/the-fundamentals-of-data-science?autoAdvance=true&autoSkip=false&autoplay=true&resume=false&u=92695330
2. https://www.linkedin.com/learning/introduction-to-data-science-2/beginning-your-data-science-exploration?autoAdvance=true&autoSkip=false&autoplay=true&resume=true&u=92695330
3. https://www.linkedin.com/learning/paths/master-excel-for-data-science

Activities / Tools

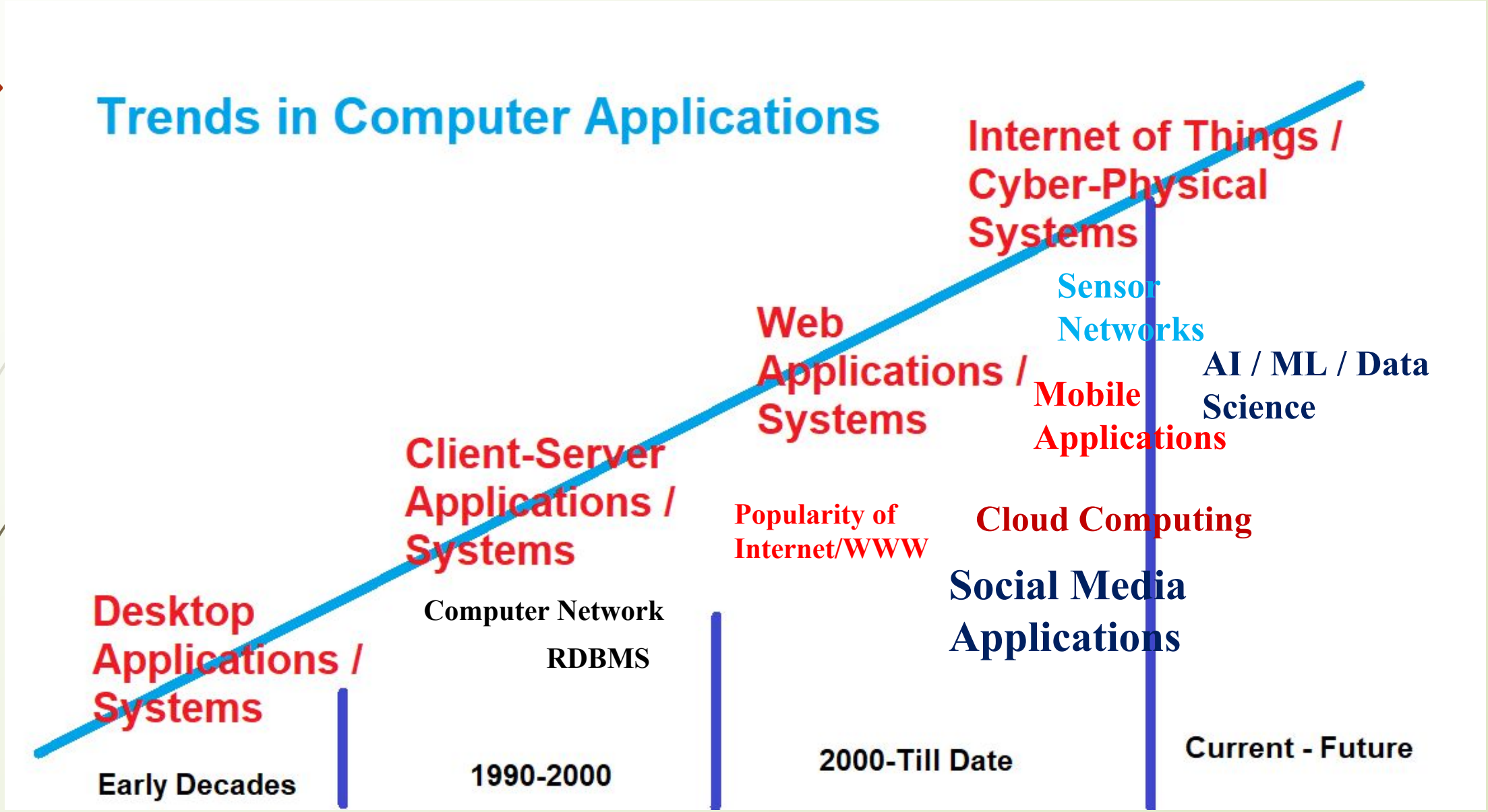
- **Activity-1**: Online Certification (minimum 15 hrs of study)
- **Activity-2**: Problems-Solving in Data Analytics (EDA , Model Building)

Tools to be Used for Hand-On:

- **Weka, Microsoft Excel**

Assessment Scheme: IA: ESE - 50:50

Sl. No.	Assessment Instrument	Formative/ Summative	Frequency	Weightage (%)	CO
1	Class Participation	Formative	Continuous	5	CO1, CO2, CO3, CO4, CO5
2	Activity-1	Formative	1	15	CO1, CO2, CO3, CO4, CO5
3	Activity-2		1	15	CO2, CO3, CO4
4	Test-1		1	5	CO1, CO2
5	Test-2	Formative	1	10	CO3, CO4, CO5
6	End Semester Exam	Summative	1	50	CO1, CO2, CO3, CO4, CO5
	Total			100	


Trends in Computer Applications
Internet of Things / Cyber-Physical Systems
Sensor Networks
AI / ML / Data Science
Web Applications / Systems
Mobile Applications
Client-Server Applications / Systems
Popularity of Internet/WWW
Cloud Computing
Social Media Applications
Desktop Applications / Systems
Computer Network
RDBMS
Early Decades
1990-2000
2000-Till Date
Current - Future

THANK YOU

Any questions…?

JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

Programme: BCA

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

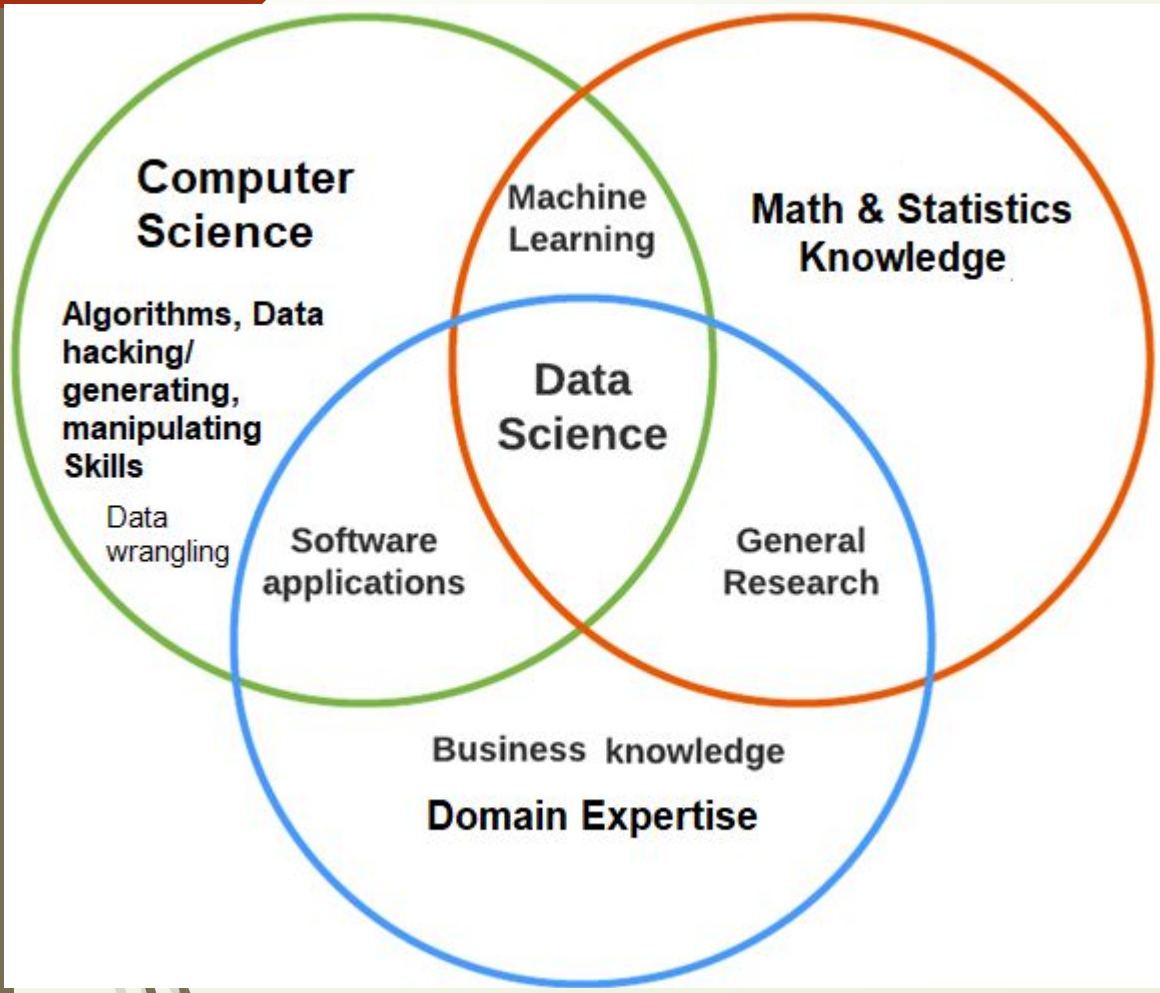
MODULE 1: Introductory Concepts

Dr. Ananta Charan Ojha, Professor

Session -1

- Overview of Data Science and Data Analytics
- Data Science vs Data Analytics
- Type of Data Analytics

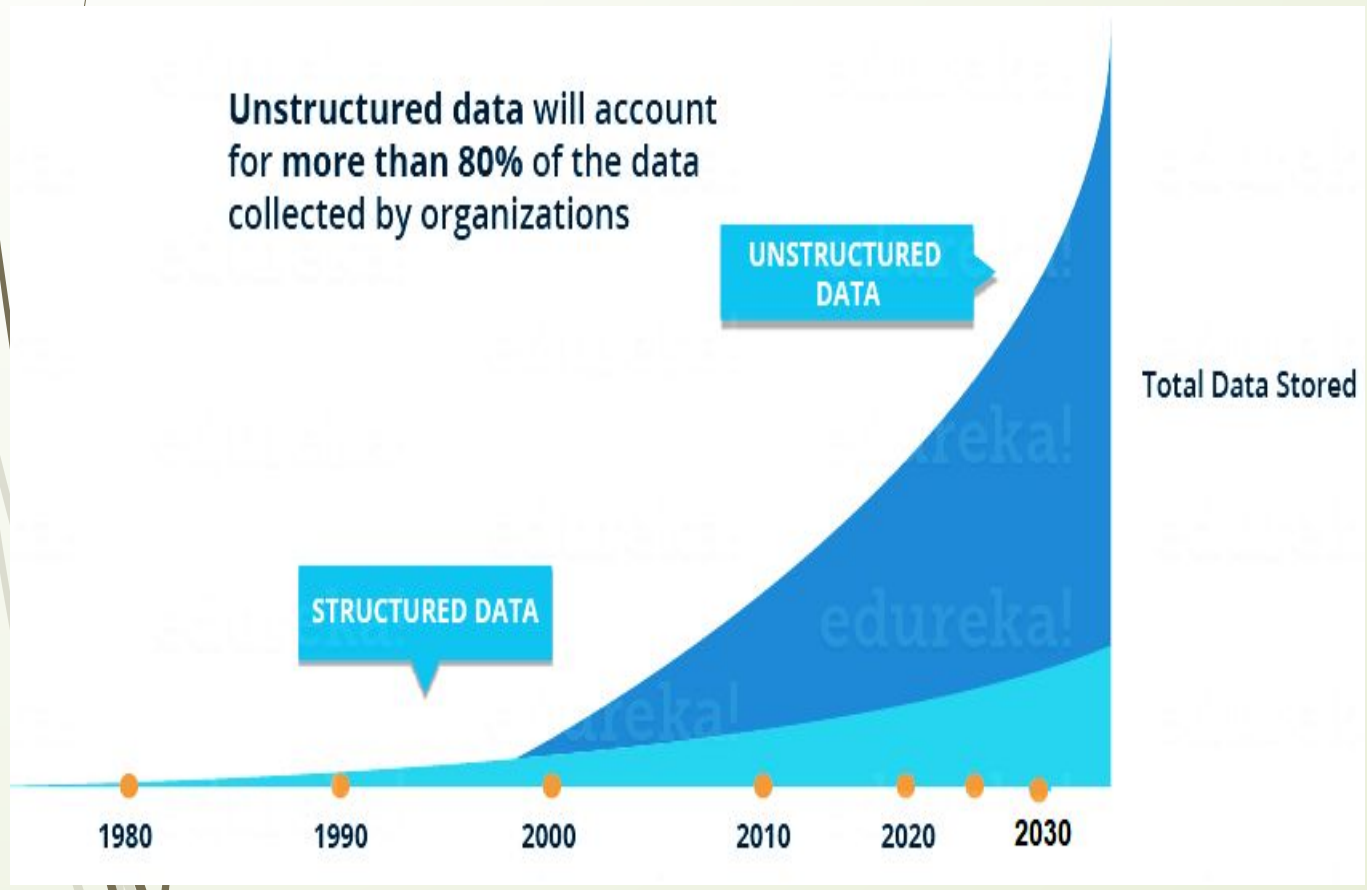
What is Data Science?



Modified Data Science Venn Diagram originally suggested by Drew Conway (2013)

- Data Science is multidisciplinary field that uses methods to analyze data and extract useful knowledge from it.
- It is a combination of multiple disciplines – Mathematics, Statistics, Optimization, Computer Science, Information Science, Data Mining, Data Visualization, Machine Learning, and Artificial Intelligence as well as domain expertise.
- It operates on massive amount of structured, and unstructured data.
- It uses a fusion of analytical methods, domain expertise, and a variety of tools and technologies.
- Data science includes descriptive, diagnostic, predictive, and prescriptive capabilities. This means that with data science, organizations can use data to figure out what happened, why it happened, what will happen, and what they should do about the anticipated result.

Why Data Science?



- Traditionally, the data that we had was mostly structured and small in size, which could be analyzed by using simple BI tools.
- Unlike data in the traditional systems which was mostly structured, today most of the data is unstructured or semi-structured.
- Let's have a look at the data trends in the image given below which shows that by 2025, more than 80 % of the data will be unstructured.
- This data is generated from different sources like transactional logs, text files, multimedia forms, sensors, machines and instruments.
- Simple BI tools are not capable of processing this huge volume and variety of data.
- This is why we need more complex and advanced analytical tools and algorithms for processing, analyzing and extracting meaningful insights out of it.
- Today, the whole world contributes to massive data growth in colossal volumes. The World Economic Forum states that by the end of 2025, the daily global data generation will reach 463 exabytes (2^{60} bytes) of data!

Evolution of Data Science

- The term 'Data Science' has been evolved over several decades. Although Computer Science and Information Technology has been there for storing, and manipulating data, reporting and disseminating information, the term 'Data Science' has never been in the limelight.
- However, the term became buzzword only after Harvard Business Review published an article "**Data Scientist: The Sexiest Job of the 21st Century" in October 2012.** The article by Thomas H. Davenport and D.J. Patil described a data scientist as a high-ranking professional with the training and curiosity who can make discoveries and find treasure out of messy, unstructured data.

Age of Number-Crunching: 1960s-1970s

- At that time, it was called **data processing**. It involved getting data into a database and reporting them, but not *analyzing* them further. Although computer programming was used in managing data but to statisticians it was not a data science because it didn't involve much analysis of data.
- Statistics had been used mostly in military applications and more mundane logistics and demographic reporting. Then the dominance of deterministic engineering applications, industrial optimization grew and drew most of the public's attention.
- In the late 1960s statistical-software-packages, most notably BMDP (Bio-Medical Data Package) and later SPSS (**Statistical Package for the Social Sciences**) and SAS (**Statistical Analysis System**), were developed and applied statisticians became very important people in 1970s.

Age of Data-Wrangling: 1980s-2000s

- Statistical analysis changed a lot after the 1970s. PC sales had reached almost a million per year by 1980. Now companies had IT Departments. From the early 1990s, sales of PCs had been fueled by Pentium-speed, GUIs, the Internet, and affordable, user-friendly software, including spreadsheets with statistical functions.
- Statistics was going through a phase of explosive evolution. By the mid-1980s, statistical analysis was no longer considered the exclusive domain of professionals.
- Another major event in the 1983 was the introduction of Lotus 1–2–3. In 1985, Microsoft Excel was introduced and became the prominent spreadsheet software within a decade surpassing Lotus 1-2-3. The popularity of Microsoft Excel fueled data analysis tasks in management information reporting systems. **BI** (Business Intelligence) emerged in 1989, mainly in major corporations.
- Against that backdrop of applied statistics came the explosion of data wrangling capabilities **(i.e., transforming or mapping data from one form to another with an intent of making it suitable for analysis)**.
- Relational databases and Sequel (SQL) data retrieval became the trend. This led to data warehousing, and **Data Mining**. It brought black-box modeling / machine learning models. R programming impacted profoundly the statistical analysis and data mining task.
- In the late 90s, instant messaging, blogging, and social media evolved and became very popular in a short span of time. The amount of data generated by and available from the Internet skyrocketed. **Big Data** became big and inflicted challenges for organizations. Technologies such as Hadoop, Spark, NoSQL databases like MongoDB, Cassandra evolved to manage it.
- In 2001, William S. Cleveland coined the term “Data Science”. Shortly thereafter, in April 2002, the publications of the “CODATA Data Science Journal” by the International Council for Science: Committee on Data for Science and Technology and in January 2003, the “Journal of Data Science” by Columbia University, respectively kickstarted the journey of Data Science.
- Government organizations and corporates started funding for activities related to data science and big data.

Age of Data Science: 2010s- Present

- The major technological advances in NoSQL databases, artificial intelligence and machine learning, the surge in social media computing, cloud computing, Internet of things, led to a revolution.
- In 2012, Harvard Business Review (HBR) published an article that declared *data scientist* to be the sexiest job of the 21st century.
- Organizations started pouring money into data science programs in anticipation of the money that would be generated from it.
- Six years later in 2018, KDnuggets (a leading site on AI, Analytics, Big Data, Data Mining, Data Science, and Machine Learning) described Data Science as an interdisciplinary field at the intersection of Statistics, Computer Science, Machine Learning, and Business Expertise, quite a bit more specific than the HBR article.
- A few others also joined to describe what data science actually was. Everybody wanted to be on the bandwagon / craze that was trendy, prestigious, and lucrative.

Data Science vs Data Analytics

- Data Science is an umbrella field that encompasses Data Analytics. Data Analytics is a branch of Data Science that focuses on more specific answers to the questions that Data Science brings forth.
- While Data Science deals with extracting hidden patterns and insights from large complex datasets, Data Analytics is designed to uncover the specifics of extracted insights.
- For example, if data science creates customer segmentations from the past transactional data based on the hidden patterns by applying clustering algorithms, then data analytics interpret the clusters in the business context to answer specific questions, and enable business decisions.
- Data science is centred around building, cleaning, and organizing datasets from raw data by using algorithms, building statistical & machine learning models to lay the groundwork for all of the analysis that need to be carried out for organizational decision-making.
- Applying data analytics tools and methodologies in a business setting is typically referred to as **business analytics**. The main goal of business analytics is to extract meaningful insights from data that an organization can use to inform its strategy and, ultimately, reach its objectives.
- Data Analytics in business organizations let to
  - Improved efficiency and productivity
  - Better financial performance
  - Identification and creation of new product and service revenue
  - Improved customer acquisition and retention
  - Improved customer experiences

Types of Analytics

- **Data analytics** refers to the process and practice of analyzing data to answer questions, extract insights, and identify trends. This is done using an array of tools, techniques, and frameworks that vary depending on the type of analysis being conducted.

- The four major types of analytics include:

1. **Descriptive analytics**, which looks at data to examine, understand, and *describe* something that's already happened.
2. **Diagnostic analytics**, which goes deeper than descriptive analytics by seeking to understand the *why* behind what happened.
3. **Predictive analytics**, which relies on historical data, past trends, and assumptions to answer questions about what will happen in the future.
4. **Prescriptive analytics**, which aims to identify specific actions that an individual or organization should take to reach future targets or goals.

Descriptive Data Analytics

- Descriptive analytics is the process of using current and historical data to identify trends and relationships.
- Descriptive data analytics is the simplest type of analytics and the foundation (or backbone) the other types are built on. It allows you to pull trends and relationships from raw data and concisely describe what happened or is currently happening but doesn't dig deeper.
- Descriptive analytics answers the question, *"**What happened?**". It also answers question of when happened, where happened, how many happened…*
- For example, imagine you're analyzing your company's data and find there's a seasonal surge in sales for one of your products: a video game console. Here, descriptive analytics can tell you, "This video game console experiences an increase in sales in October, November, and December each year."

Diagnostic Data Analytics

- Diagnostic data analytics is the process of using data to determine the causes of trends and correlations between variables. It can be viewed as a logical next step after using descriptive analytics to identify trends.
- Diagnostic analytics addresses the next logical question, ***"Why did this happen?"***
- Taking the analysis a step further, this type includes comparing coexisting trends or movement, uncovering correlations between variables, and determining causal relationships where possible.
- Continuing the aforementioned example, you may dig into video game console users' demographic data and find that they're between the ages of 8 and 18. The customers, however, tend to be between the ages of 35 and 55. Analysis of customer survey data reveals that one primary motivator for customers to purchase the video game console is to gift it to their children. The spike in sales in the October, November and December months may be due to the festivals and holidays that include gift-giving.

Predictive Data Analytics

- Predictive analytics is the use of data to predict future trends and events. It uses historical data to forecast potential scenarios that can help drive strategic decisions.
- It answers the question, ***"What might happen in the future?"***
- The predictions could be for the near future—for next day, next month or the more distant future - for the upcoming year.
- By analyzing historical data in tandem with industry trends, you can make informed predictions about what the future could hold for your company.

- For instance, knowing that video game console sales have spiked in October, November, and December every year for the past decade provides you with ample data to predict that the same trend will occur next year.
- Predictive analysis can be conducted manually or using machine-learning algorithms. Either way, historical data is used to make assumptions about the future.

Prescriptive Data Analytics

- Prescriptive analytics is the use of data that provides information on not just what will happen in your company, but *how it could happen better* if you did x, y, or z.
- Beyond providing information, prescriptive analytics goes even one step further to recommend actions you should take to optimize a process, campaign, or service to the highest degree.
- Prescriptive analytics answers the question, ***"What should we do next?"***
- Prescriptive analytics is "the future of data analytics". This type of analysis goes beyond explanations and predictions to recommend the best course of action moving forward.
- In the video game example: What should your team decide to do given the predicted trend in seasonality due to winter gift-giving?
  - Perhaps you decide to run an A/B test (a method of two-sample hypothesis testing for comparing the outcomes of two different choices, A and B) with two ads: one that caters to product end-users (children) and one targeted to customers (their parents).
  - The data from that test can inform how to capitalize on the seasonal spike and its supposed cause even further. Or, may be you decide to increase marketing efforts in September with festival / holiday-themed messaging to try to extend the spike into next months.

THANK YOU

Any questions…?

JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

Department of BCA

INTRODUCTION TO DATA ANALYTICS (22BCAD4C01)

MODULE 1: Introductory Concepts

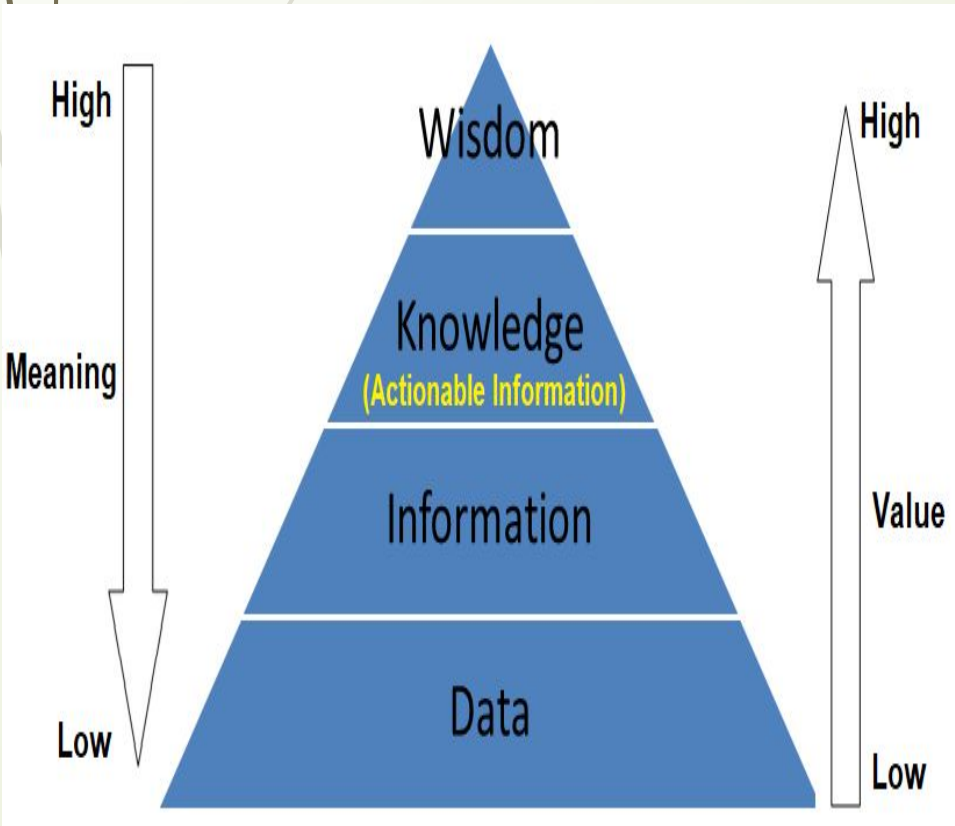
Dr. Ananta Charan Ojha, Professor

Session -2

- DIKW Pyramid
- Data Ubiquity
- Nature of Data
- Big Data
- Data-Driven Decision

DIKW Pyramid

- The DIKW (Data, Information, Knowledge, Wisdom) Pyramid represents the hierarchical relationships among data, information, knowledge and wisdom.
- The more we enrich our data with meaning and context, the more knowledge and insights we get out of it so we can take better, informed and data-driven decisions.



- **Data**: Data is a collection of facts, figures in a raw or unorganized form doesn't provide any meaning; doesn't answer any questions nor draw any conclusion. Raw data is not useful nor valuable without a context. **Example**: Rs. 5.32 crores.
- **Information**: provides meaning to data in a context, describes what is the data in a context. Information is a contextual data that has a meaning. **Example**: Rs. 5.32 crores is the total sales of last month of ABC Enterprises. Sales figures of different months.

This is data that has been "cleaned" of errors and further processed in a way that makes it easier to measure, visualize and analyze for a specific purpose.

- **Knowledge**: "How" is the information relevant to our goals? "How" are the pieces of this information connected to other pieces to add more meaning and value? And, maybe most importantly, "how" can we apply the information to achieve our goal?

We turn it into knowledge. We discovery of hidden patterns in data, relationships among variables in data. **Example**: When you compare the monthly sales, there is a decreasing trend of sales over last couple of months.

- **Wisdom**: Wisdom is the top of the DIKW hierarchy and to get there, we must answer questions such as 'why do something' and 'what is best'. In other words, wisdom is knowledge applied in action.

It provides understanding, explanation, why it is happening. **Example**: why is the decrease in sales?

It helps in doing the right thing, tells what is the best thing to do, wisdom to enable right decisions. **Example**: what is the right action to nullify the factors that cause decrease in sales and increase the sales.

The four types of data analytics add values to the raw data and contribute to the DIKW pyramid. You process the data and get information; when you derive hidden patterns from the data, you get knowledge; you understand it and apply the knowledge to take right decision- you become intelligent and wise.

Data Ubiquity

- According to experts "What we're seeing right now is the end of the era of software. Hardware had its 20- to 30-year run. Software had its 20- to 30-year run".
- So, what's next? An era that experts call the "age of data ubiquity," one in which a new generation of data-centric apps exploit massive data sets generated by both individuals and enterprises.
- The IDC estimates that by 2025 there will be over 40 billion IoT devices on earth, generating almost half the world's total digital data.
- The bulk of data generated comes from three primary sources: social data, machine data and transactional data.
- **Transactional data** is generated from all the daily transactions that take place both online and offline. Invoices, payment orders, storage records, delivery receipts – all are characterized as transactional data.
- **Machine data** is generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behavior. This type of data is expected to grow exponentially as the internet of things grows, become more pervasive and expands around the world. Sensors such as medical devices, smart meters, road CCTV cameras, satellites, and drones will deliver high velocity, value, volume and variety of data in the very near future.
- **Social data** comes from the Posts, Likes, Tweets & Retweets, Comments, Video Uploads, and general media (e.g., image, text docs) that are uploaded and shared via the world's favorite social media platforms. This kind of data provides invaluable insights into consumer behavior and sentiment and can be enormously influential in marketing analytics. The public web is another good source of social data.

Nature of Data

Data can be distinguished along many dimensions. The most important one is the degree of organization.

- **Structured Data**
  - Structured data has a high degree of organization. The data that has been formatted and transformed into a well-defined data model / schema, consisting of rows and columns so that it can be machine readable and extracted through algorithms easily. Examples: Relational Data, ExcelSheet, CSV etc. Structured data is generated by both humans and machines.
- **Unstructured Data**
  - Unstructured data has an internal structure but is not structured via predefined data models or schema. It may be textual or non-textual, and human- or machine-generated. This data is difficult to process due to its complex arrangement and no specific formatting. Examples: **Audio, Video, Image, Text files, PDF files etc.**
- **Semi-Structured Data**
  - Semi-structured data has some degree of organization in it. It is not as rigorously formatted as structured data, but also not as messy as unstructured data. This degree of organization is typically achieved with some sort of tags and markings that identify separate data elements, which enables data analysts to determine information grouping and hierarchies. These can be comma or colons or anything else for that matter. Examples: **Markup languages, HTML, XML; Open standard JSON (JavaScript Object Notation); Data stored in** NoSQL databases.

Big Data

Big data refers to data that is so large, fast or complex that it's very difficult or impossible to process using traditional methods.

- The concept of big data gained momentum in the early 2000s when industry analyst Doug Laney articulated the definition of big data as the three V's: volume, velocity and variety.

Big Data Characteristics

- **Volume:** The quantity of generated and stored data. The size of the data determines the value and potential insight, and whether it can be considered big data or not. The size of big data is usually larger than terabytes and petabytes.
  - Examples: Date such as Twitter data feeds, clickstreams on a web page or a mobile app, or sensor-enabled equipment.
- **Velocity:** The speed at which the data is generated and (perhaps) processed. Big data is often available in real-time or near real-time and requires real-time evaluation and action. Two kinds of velocity related to big data are the frequency of generation and the frequency of handling, recording, and publishing.
  - Examples: RFID tags, sensors and smart meters are driving the need to deal with these torrents of data in near-real time.
- **Variety:** The type and nature of the data. Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured and semi-structured data types, such as text documents, emails, videos, audios, stock ticker data and financial transactions that require additional pre-processing to make it suitable for storage and analysis.
- **Veracity:** The truthfulness or reliability of the data, which refers to the quality of data. Because data comes from so many different sources, the data quality of captured data can vary greatly; it's difficult to link, match, cleanse and transform data across systems; it can be erroneous affecting an accurate analysis.
- **Value:** The worth in information that can be achieved by the processing and analysis of large datasets. Value also can be measured by an assessment of the usability of information that is retrieved from the analysis of big data.
- **Variability:** In addition to the increasing velocities and varieties of data, data flows are unpredictable – changing often and varying greatly. It's challenging, but businesses need to know when something is trending in social media, and how to manage daily, seasonal and event-triggered peak data loads.

Data-Driven Decisions

- Data-driven decision-making is the process of using data to inform your decision-making process and validate a course of action before committing to it.

- **Benefits**

  1. You'll Make More Confident Decisions

  2. You'll Become More Proactive

  - For example, by identifying business opportunities before your competitor does, or by detecting threats before they grow too serious.

  3. You Can Realize Cost Savings and profitability

  - ✔ Improved efficiency and productivity in organizational processes
  - ✔ Better financial performance
  - ✔ Identification and creation of new product and service revenue
  - ✔ Improved customer acquisition and retention
  - ✔ Improved customer experiences
  - ✔ Competitive advantage

- **How to become more Data-Driven**

  1. Look for Patterns Everywhere

  2. Tie Every Decision Back to the Data

  3. Visualize the Meaning Behind the Data

THANK YOU

Any questions…?

JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

Programme: BCA

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

MODULE 1: Introductory Concepts

Dr. Ananta Charan Ojha, Professor

Session -3

- Data Science Process

Data Science Process

- A data science process consists of several activities in multiple stages that are used to find a solution for a business problem at hand.

1. **Frame the Business Problem**
2. **Data Acquisition**
3. **Data Preparation**
4. **Data Exploration**
5. **Feature Engineering**
6. **Model Building**
7. **Presentation and Model deployment**

1. Frame the Business Problem

- Before solving a problem, the pragmatic thing to do is to know what exactly the problem is. You begin the process of data science by asking the right questions to find what the problem is.
- Let's take a very common problem – The sales problem of a company that deals with a product.
- For analysis of the problem, you need to start by asking a lot of questions:

  - Who are the customers? How to identify them?
  - What products they are interested in?
  - Why are they interested in your products?
  - What are the markets? How do you approach the target market?
  - What information do you have about the target market?
  - What is the sale process right now?

- After a discussion with the marketing team, you decide to focus on the problem: "How can we identify potential customers who are more likely to buy our product?"
- The next step for you is to figure out what all data you have available with you to answer the above questions.

2. Data Acquisition

- After defining the problem, you will need to collect the required data to derive insights and turn the business problem into a probable solution.
- The required data may be available in the organization, if not it can be outsourced / purchased from external sources.
- If you think the data available is not sufficient, then you must make arrangements to collect new data.
- Many companies store the sales data they have in customer relationship management (CRM) systems.

3. Data preparation

- The data you have collected may contain errors like invalid entries, inconsistence values, missing values / null values, duplicate values, and many more.
- First, you need to make sure the data is clean and free from all possible errors.
- Since data collection is an error-prone process, in this phase you enhance the quality of the data and prepare it for use in subsequent steps.
- This phase consists of three subphases:
  - *data cleansing* removes false values from a data source and inconsistencies across data sources,
  - *data integration* enriches data sources by combining information from multiple data sources, and
  - *data transformation* ensures that the data is in a suitable format for use in your models.

4. Data exploration

- This is one of the most crucial steps in a data science process. This step often popularly known as EDA (Exploratory Data Analysis).
- Data exploration is concerned with building a deeper understanding of your data.
- You try to understand how variables interact with each other, the distribution of the data, and whether there are outliers. Identify correlation and trends between the dependent and independent variable.
- To achieve this, you mainly use descriptive statistics, and data visualization techniques.
- Descriptive statistics describe the data based on its properties such as
  - Measures of Frequency
  - Measures of Central Tendency
  - Measures of Dispersion or Variation
  - Measures of Position
- Data visualisation makes the patterns and trends identification much easier rather than just looking at thousands of rows on a dataset and just using statistics.

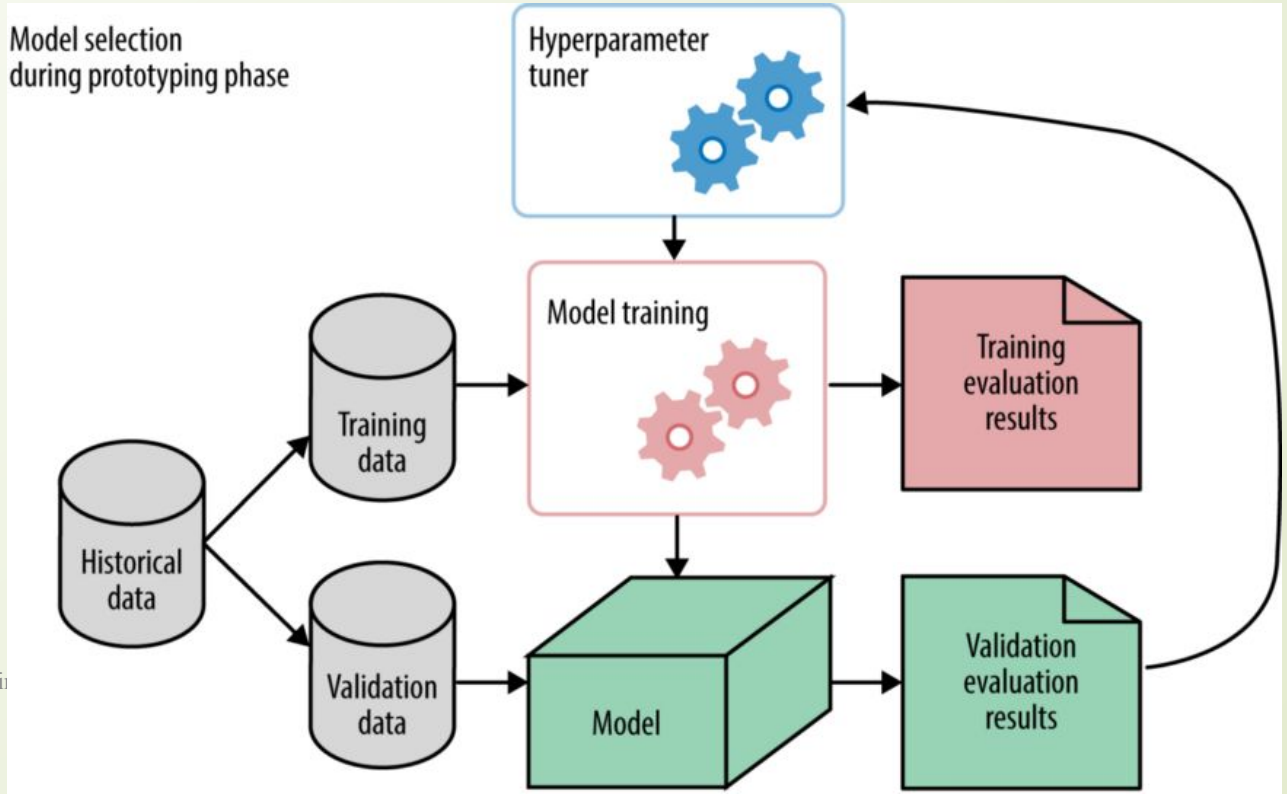
5. Feature Engineering

- **Feature engineering** is the process of using domain knowledge to extract **features** from raw data and to transform data into a form where a model can understand better.
- After transforming the data to the right format and dealing with potential data hazard, most of the time especially with a high dimensional dataset, we end up with many features. We cannot feed all the features to the machine learning model, that is not how it works, that would overfit the model hugely. Instead, we have to choose the right number of features. This is called **Feature Selection**.

6. Model Building

- Once the data with reduced features are ready to be modelled, the data is divided into training and test sets. Model building (see Figure) involves model training and evaluation in iteration. We create a **Baseline model and** keep increasing or decreasing the complexity of the model by **Hyperparameter Tuning to get desired accuracy.**

- Model is trained using training set. We need to evaluate the performance of the Trained model, on test / validation dataset.



7. Presentation and Model deployment

- Finally, you present the results to your organization. These results can take many forms, ranging from presentations to research reports. Sometimes you'll need to automate the execution of the process and use the outcome from your model. You deploy the model into production that may be on real-time mode or batch mode.

- Once the model is deployed, it is monitored for its performance. If required you may go back to one or more previous stages to recalibrate the model to improve its performance.

- Due to the iterative nature of the process, it is sometimes called Data Science process life cycle.

THANK YOU

Any questions…?

JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

Programme: BCA

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

MODULE 1: Introductory Concepts

Dr. Ananta Charan Ojha, Professor

Session -4

- Data Science Applications

Applications of Data Science

- Data science has been the most popular field with applications in a wide range of domains. It has been revolutionizing the way we perceive data and data-driven decisions.
- Some Popular Sectors:
  - Banking, Finance and Insurance
  - Sales and Marketing
  - Healthcare and Pharma
  - Internet and Social media
  - Travel and Transport

Banking, Finance & Insurance

- Data science has enabled financial institutions to be more secure and manage their resources efficiently. It also enables them to make smarter and more strategic decisions and be saved from fraud. It also helps manage customer data, risk analysis and modeling, predictive analysis, and much more.

- **Personalized Services** (personalized services offerings and interactions based on customer profiling)
- **Risk Analysis and Management** (Loan approvals, Credit card approval, Customer credit scoring)
- **Fraud Detection** (Credit Card Fraud Detection, Money laundering, Enhancing auditing by finding irregularities, Suspicious financial transactions)
- **Predicting stock performance and pricing**
- **Economic and financial forecasting of organizations**
- **Bankruptcy prediction**
- **Mortgage underwriting**
- **Foreign exchange rate forecasting**
- **Insurance Claim Prediction**

Marketing and Sales

- **Sales forecast :** predicting future sales
- **Customer Sentiment Analysis:** analysing **t**he feedbacks and reviews to understand what customers desire and why
- **Customer Churn prevention:** identify trends and features in the behavior, communication, and ordering of customers who have ceased shopping through customer relationship management information.
- **Inventory Management:** identify buying patterns, optimize inventory management and timely delivery
- **Cross-sell recommendations:** cross-selling invites customers to buy related or complementary items. Cross-selling requires that a consumer who has previously purchased or intends to buy the extra product being offered.
- **Predicting customer lifetime value (CLV):** to know how valuable a customer is during his association; Several metrics, such as the buying pattern, gross value, frequency of purchase, mean order value, etc. are used to measure customer lifetime value to the company and loyalty programs are recommended.
- **Customer Segmentation and Target Marketing**: personalised product and service offerings

Healthcare and Pharma

- **Disease Diagnosis:** identifying diseases accurately
- **Medical Prognosis:** treatment sensitivity, life expectancy, survivability, and disease progression
- **Medical Image Analysis**
  - With the help of medical image analysis, a machine predicts diseases such as cancer, tumor, organ delineation, and many others.
- **Genomics and Proteomics:** identification, quantification, profiling new genes, proteins
- **Drug Development:** target molecule identification
- **Clinical Trails**: conducted to identify the dose-toxicity, side effects, and drug effectiveness
- **Epidemic Outbreak and Control:** cluster identification, contact tracing, peak prediction

Internet and Social Media

- **Internet Search**
  - Search engines make use of data science algorithms to deliver the best result for our searched query in a fraction of seconds.
  - Google gives accurate autocomplete prompts in search box as if it is reading our mind about search context
- **Digital Advertisements / Targeted Advertising**
  - Data science algorithms are used to determine the banners or advertisements to be displayed on various websites you visit. They can be tailored to a user's previous actions or browsing patterns.
- **Price Comparison Websites**
  - These websites provide the convenience of comparing the price of a product from multiple vendors at one place. These websites are being driven by lots and lots of data which is fetched and processed in seconds. PriceGrabber, PriceRunner, Junglee, Shopzilla, DealTime are some examples of price comparison websites. Now a days, price comparison website can be found in almost every domain such as technology, hospitality, automobiles, durables, apparels etc.
- **Sentiment Analysis:** predicting positive and negative sentiment analysis
- **Influencer Marketing**: identifying influencers and promotion effectiveness
- **Others**

- Google maps predicts which route will be faster and shows such an accurate estimated time of reaching destination.
- Gmail filters spam mails to the spam folder in your gmail account.
- Amazon or Netflix are so good in recommending you what you may like to buy or watch movie.
- “Friends you may know” feature by Facebook.

Travel and Transport

- **Airline Route Planning**
  - Predict flight delay
  - Decide which class of airplanes to buy
  - Whether to directly land at the destination, or take a halt in between (For example: A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
  - Effectively drive customer loyalty programs
- **Popular cab services**
  - Ola, Uber employ data science to improve price and delivery routes, as well as optimal resource allocation, by combining numerous factors such as consumer profiles, geography, economic indicators, and logistical providers.
- **Dynamic Pricing**
- **Faster Route and ETA** (Estimated Time of arrival) prediction

THANK YOU

Any questions…?

JAIN DEEMED-TO-BE UNIVERSITY | SCHOOL OF COMPUTER SCIENCE AND IT

School of Computer Science & IT

BCA Programme

INTRODUCTION TO DATA ANALYTICS (23BCAD4C01)

MODULE 1: Introductory Concepts

Dr. Ananta Charan Ojha, Professor

Session -5

- Data Science Job Roles
- Data Ethics

Data Science Job Roles

- Data science is a rapidly growing field. The hot new field promises to revolutionize various sectors from business to government, health care to academia.
- There is a rising demand for data science jobs around the world. These job opportunities would continue to surge in coming years.
- Data science roles and responsibilities are diverse and skills required for them vary considerably.
  - **Data Scientist, Data Analyst, Data Engineer, Data Architect, Machine Learning Scientist, Machine Learning Engineer, Business Intelligence Developer, Database Administrator etc.**
- The two most common and high paid job titles are
  - Data Scientists and
  - Data Analysts.

Major Roles and Responsibilities

- Extracting / Gathering data from various sources using automated tools
- Developing and maintaining databases
- Processing, cleansing and verifying the integrity of data.
- Performing Exploratory Data Analysis
- Discover business insights using machine learning tools and techniques.
- Identifying new trends and patterns in data to make predictions for the future.
- Develop visualization and Presentation KPIs
  - (**Example**: **Key performance indicators** are leading indicators/ metrics by which the sales team and their leaders evaluate the performance and how effective their efforts the sales and organizational goals.)

Major Skill Sets

- Programming skills in Python, R, Perl, SQL, MS Excel,
- Knowledge of Statistics, Machine Learning, Model building
- Big data tools like Hadoop, Pig, Hive and Spark
- Knowledge of SQL and NoSQL databases, Data Modelling and Design
- Data Visualization, and tools like Tableau, Microsoft Power BI
- Analytical and Creative Thinking / Structured Thinking

What is Data Ethics?

- Data ethics encompasses the moral obligations of gathering, protecting, and using **personally identifiable information (PII)** and how it affects individuals.
  - For instance, your company may collect and store data about customers' mobile number, email address etc. on your e-commerce website when they purchase your product. If you're a digital marketer, you likely interact with this data daily.
- Data ethics asks, 'Is this the right thing to do?'
- Any instances of unethical data collection, storage, or use can lead to legal issues.
- By following ethical practices, you can protect your customers' safety and save your organization from legal issues.

Principles of Data Ethics

- Here are five principles of data ethics to apply at your organization.

1. Ownership

- The first principle of data ethics is that an individual has ownership over his personal information. Just as it's considered stealing to take an item that doesn't belong to you, it's unlawful and unethical to collect someone's personal data without his consent.
- Never assume a customer is OK with you collecting his data; always ask for permission to avoid ethical and legal dilemmas.

2. Transparency

- In addition to owning their personal information, data subjects have a right to know how you plan to collect, store, and use it. When gathering data, exercise transparency.
- Withholding or lying about your company's methods or intentions is deception and both unlawful and unfair to your data subjects.

3. Privacy

- Another ethical responsibility that comes with handling data is ensuring data subjects' privacy. Even if a customer gives your company consent to collect, store, and analyze their personally identifiable information (PII), that doesn't mean they want it publicly available.
- PII is any information linked to an individual's identity. Some examples of PII include:
  - Full name, Birthdate, Address, Phone number, Aadhar Number,
  - PAN Card Number, Credit card information, Bank account number, Passport number

- To protect individuals' privacy, ensure you're storing data in a secure database so it doesn't end up in the wrong hands.
- Data security methods that help protect privacy include dual-authentication password protection and file encryption.

- Data security focuses on systems in place that prevent malicious external attempts to access, steal, or destroy data, whereas data privacy focuses on the ethical and legal use and access to sensitive data and PII. Data security and data privacy work together to ensure your customers' safety and anonymity.

- One way to protect privacy is by de-identifying a dataset.
- A dataset is de-identified when all pieces of PII are removed, leaving only anonymous data. This enables analysts to find relationships between variables of interest without attaching specific data points to individual identities.

4. Intention

- In the context of ethics, intentions matter.
- Before collecting data, ask yourself why you need it, what you'll gain from it, and what changes you'll be able to make after analysis.
- If your intention is to hurt others, profit from your subjects' weaknesses, or any other malicious goal, it's not ethical to collect their data.
- When data could be sensitive, so collecting it when it's unnecessary isn't ethical.
- Strive to collect the minimum viable amount of data, so you're taking as little as possible from your subjects while making a difference.

5. Outcomes

- Even when intentions are good, the outcome of data analysis may cause inadvertent / unintentional harm to individuals or groups of people.
- For example, Voters' mood / opinion survey may negatively impact a party during election season.

THANK YOU

Any questions…?