Unit 1
Introduction to Data Science

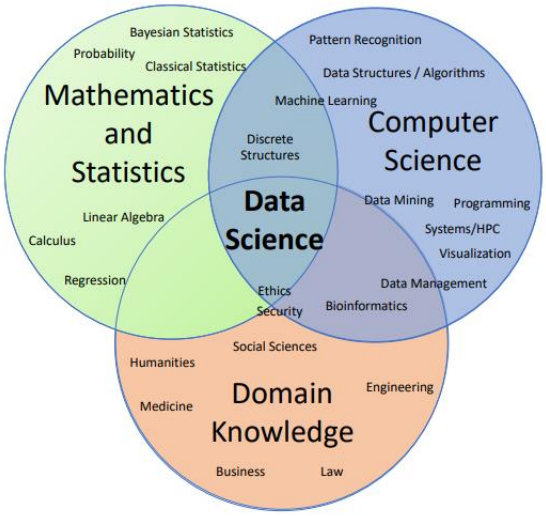
1.1 Overview of data science

The term "data science" was coined in 2001, attempting to describe a new field. Some argue that it's nothing more than the natural evolution of statistics, and shouldn't be called a new field at all. But others argue that it's more interdisciplinary. For example, in The Data Science Design Manual (2017), Steven Skiena says the following.

"I think of data science as lying at the intersection of computer science, statistics, and substantive application domains."

Regardless of whether data science is just a part of statistics, and regardless of the domain to which we're applying data science, the goal is the same: to turn data into actionable value.

Data Science is the science which uses computer science, statistics and machine learning, visualization and human- computer interactions to collect, clean, integrate, analyze, visualize, and interact with data to create data products.



Data Science is an interdisciplinary field that focuses on extracting meaningful insights from data. It combines elements of statistics, computer science, and domain expertise to analyze structured and unstructured data using various techniques such as machine learning, data mining, and predictive modeling. The ultimate goal of data science is to use data to inform decisions, solve problems, and generate actionable insights.

Importance of Data Science:

In today's era, data is growing day by day, As the world generates an unprecedented amount of data every day. Everyone depends on it to make decisions, improve how they work, and understand what's going on. and the key to making sense of all this data is something called Data Science. It helps us to find important patterns and information hidden in big and complicated data sets.

1. Enhancing Decision-Making

Data science plays a pivotal role in enhancing decision-making processes within organizations. Data scientists can provide actionable insights that help businesses make informed decisions by analyzing historical data and identifying patterns. This can range from understanding consumer behavior to predicting market trends, thus enabling companies to strategize effectively.

2. Improving Operational Efficiency

Operational efficiency is critical for any organization aiming to maximize productivity and reduce costs. Data science contributes significantly by identifying bottlenecks and inefficiencies in processes. For instance, in manufacturing, data analysis can reveal production line inefficiencies, while in logistics, it can optimize routes and reduce delivery times.

3. Driving Innovation

Innovation is essential for staying competitive in today's market. Data science fosters innovation by uncovering new opportunities and enabling the development of novel products and services. Through techniques such as machine learning and predictive analytics, organizations can experiment with new ideas and approaches, ultimately leading to breakthroughs and advancements.

Examples of Real-World Use Cases

- ✓ Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to

- ✓ Supply Chain Optimization: Retailers like Amazon use data science to optimize inventory management, demand forecasting, and logistics.
- ✓ Risk Assessment: Credit scoring agencies and banks evaluate creditworthiness by analyzing customer credit history, income, and spending patterns.
- ✓ Dynamic Pricing: E-commerce platforms use data science to adjust prices dynamically based on demand, competition, and customer behavior.
- ✓ Targeted Advertising: Google Ads and Facebook Ads analyze user activity to deliver personalized ads based on preferences and search history.
- ✓ Game Strategy: Teams like those in the NBA and MLB use data analytics to optimize strategies, player positions, and training regimens.

1.2 Jargons of data science (terminology)

Jargons in Data Science refer to the specialized terms and phrases that are commonly used by professionals within the field. These terms may be unfamiliar to those outside the domain and can sometimes make discussions or learning more challenging for beginners. However, understanding these jargons is crucial for anyone looking to get involved in data science, as they represent key concepts and techniques used in analyzing data, building models, and drawing insights.

✓ Data

The raw facts or figures that can be processed to extract meaningful insights. Data can be structured (e.g., tables) or unstructured (e.g., text, images).

✓ Metadata

Metadata is simply defined as data about data.

✓ Algorithm

Algorithms are repeatable sets, usually expressed mathematically, of instructions that humans or machines can use to process given data. Typically, algorithms are constructed by feeding them data and adjusting variables until the desired result is achieved.

✓ Exploratory data analysis (EDA)

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

EDA helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

✓ Big Data:

Big data can be defined as very large volumes of data available at various sources in varying degree of complexity, genertaed at different speed i.e velocities and varying degree of ambiguity, which cannot be processed using traditional technologies, processing methods or soultions.

3V's are the characteristics of Big Data. i.e Volume, Velocity and Variety.

Volume refers to the quantities of data larger than conventional relational database infrastructure. The velocoty refer to the speed of generation of data. And Variety refers to heterogeneous sources and the nature of data both structure and unstructred. Tools used for Bigdata are Hadoop, Spark, Flink etc.

✓ Data Warehouse

It can be referred to as electronic storage, where businesses store a large amount of data and information. e.g Amazon Redshift, Google BigQuery, Microsoft Azure Synapse Analytics, Teradata.

Data Warehousing is the process of collecting, transforming, storing, and managing data within the data warehouse for business intelligence purposes.

✓ Data Analytics

Data Analytics focuses on analyzing data to gain insights, identify patterns, and support decision-making. It typically involves using statistical techniques and software tools to examine datasets and generate reports or visualizations.

While Data Science is broader and includes data preparation, algorithm building, and programming, Data Analytics is often more focused on analyzing data and producing actionable insights.

✓ Artificial Intelligence (AI)

AI refers to the creation of machines or systems that can perform tasks that would typically require human intelligence. These tasks include problem-solving, decision-making, speech recognition, image recognition, and language understanding. AI systems use algorithms and models that allow them to simulate human cognition. AI is often applied in complex scenarios like autonomous vehicles, recommendation systems, and virtual assistants.

✓ Machine Learning (ML)

Machine Learning is a subset of AI that involves creating algorithms that learn from data and make predictions or decisions without being explicitly programmed. ML models are designed to identify patterns in large

datasets and improve their accuracy over time as they are exposed to more data. ML includes several approaches, such as supervised learning, where the model is trained on labeled data, and unsupervised learning, where the model finds patterns in unlabeled data. ML is widely used in applications like fraud detection, image recognition, and natural language processing.

1.3 Modern Data Ecosystem

A modern data ecosystem, sometimes referred to as a "technology stack," contains these fundamental elements:

- ✓ Data types and Sources.
- ✓ The people who use it.
- ✓ The technology and infrastructure that supports it.
- ✓ The processes that facilitate it.

In simple terms, a modern data ecosystem is a complex, interconnected network of people, processes, and technologies that work together to manage, analyze, and derive insights from data.

It involves several key components:

1. Datatypes and Sources:

✓ Types of Data:

- Structured data: Organized data typically found in relational databases (e.g., tables, spreadsheets).
- Unstructured data: Data without a predefined format, such as text, images, videos, and social media posts.
- Semi-structured data: Data that doesn't fit into a traditional database but still contains tags or markers, such as JSON, XML, or CSV.

✓ Data Sources:

Internet of Things (IoT): Sensors, smart devices, and connected equipment generating real-time data.

Social Media: Platforms like Twitter, Facebook, or LinkedIn that provide valuable unstructured data.

Sensors: Devices that collect environmental, industrial, or personal data for analysis

2. Stakeholders/ People using it:

- ✓ Data Engineers: Design, build, and maintain the infrastructure that allows for data collection, storage, and processing.
- ✓ Data Analysts: Analyze and interpret data to provide actionable insights to stakeholders.

- ✓ Data Scientists: Use advanced analytics, machine learning, and statistical techniques to predict trends, build models, and create algorithms for business solutions.

3. Technologies and Infrastructure:

- ✓ Cloud Computing: Platforms like AWS, Microsoft Azure, and Google Cloud Platform (GCP) provide scalable storage, computing, and analytics capabilities.
- ✓ Data Storage:
  - Databases: Traditional relational databases like MySQL, PostgreSQL, or Oracle, which store structured data. Non-relational databases like MongoDB, Cassandra, or Couchbase that support semi-structured and unstructured data.
  - Data Lakes: Large, centralized repositories where both structured and unstructured data can be stored in its raw form (e.g., AWS S3, Azure Data Lake).
  - Data warehouses: Data warehouses are centralized repositories designed for structured, historical data analysis. They offer fast query performance and support complex data models. e.g. Snowflake, Google Bigquery, Amazon Redshift etc.
- ✓ Business intelligence (BI): Business intelligence tools are software applications that enable users to explore, analyze, and visualize data to gain insights. They provide interactive dashboards, reports, and data mining capabilities. e.g Power BI, Tebleau, Qlik, Looker etc.

4. The Processes That Facilitate It

Processes are the policies, standards, and workflows that ensure the data ecosystem operates efficiently and securely. These are the workflows and practices that guide how data is handled within the ecosystem.

- ✓ Data Transformation: The process of cleaning, structuring, and enriching raw data so that it is usable and meaningful for analysis.
- ✓ Data Visualization: The process of presenting data in a graphical format to make it easier to understand, interpret, and communicate insights.
- ✓ Data Governance: Establishes policies and practices for how data is collected, stored, processed, and accessed, ensuring compliance with regulations like GDPR and CCPA, and managing data quality and lineage.
- ✓ Data Integration: Combines data from various sources into a unified view, typically by extracting, transforming, and loading (ETL) data into centralized repositories like data lakes or warehouses for analysis.

- ✓ Data Quality Management: Ensures the accuracy, completeness, reliability, and timeliness of data to support correct decision-making.
- ✓ Data Security and Privacy: Protects sensitive data from unauthorized access and ensures compliance with privacy regulations, using methods like encryption, access controls, and anonymization.
- ✓ Collaboration and Workflow Management: Encourages cross-functional teamwork and manages data workflows using tools like Jupyter Notebooks and platforms like Airflow or Kubernetes.

1.4 Data science lifecycle

The life cycle of data science outlines the steps/phases, from start to finish, that projects usually follow when they are executed.

It provides a framework for the each phase from the creation of the project until its completion. Figure below shows the steps of data science life cycle.

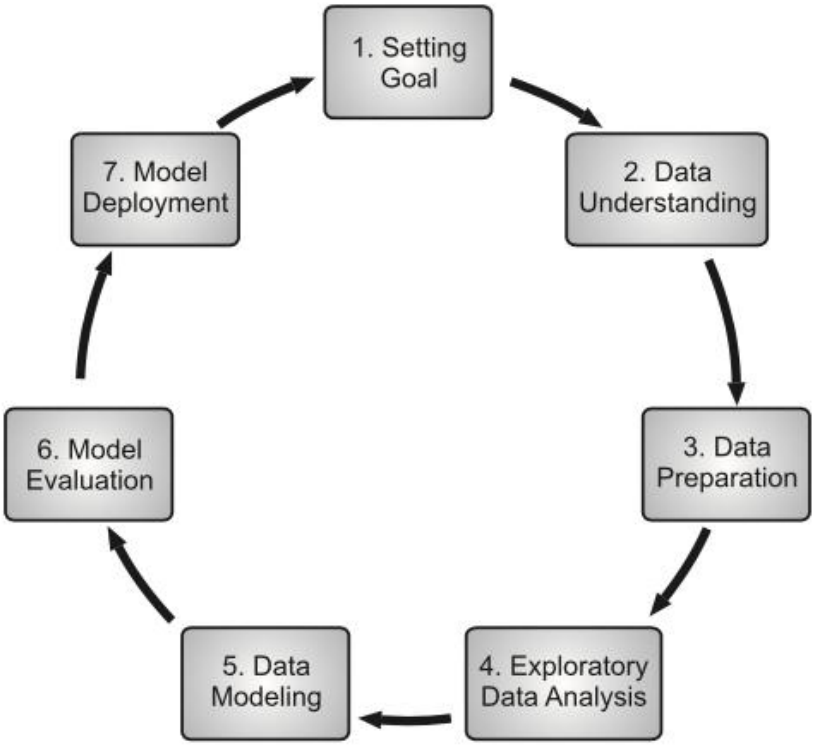


Fig. 1.4: Life Cycle of Data Science

Reference: Foundation of datascience, Nirali Publication 2nd Edition

Step 1: Setting Goal

- ✓ The entire cycle revolves around the business or research goal. What will we solve if we do not have a precise problem? It is essential to understand the business objective clearly because that will be the final goal of the analysis.

- ✓ Only we can set the specific goal of analysis in synchronised with the business objective after proper understanding.

Step 2: Data Understanding

- ✓ Data understanding involves the collection of all the available data.
- ✓ We need to understand what data is present and what data could be used for given problem.
- ✓ The data understanding step also involves describing the data, their structure, their relevance, their data type.

Step 3: Data Preparation

- ✓ The data preparation step includes selecting the relevant data, integrating the data by merging the data sets, cleaning them, treating the missing values by either removing them or imputing them, treating erroneous data by removing them and checking outliers using box plots and handle them.
- ✓ This step is used for constructing new data derive new features from existing ones.
- ✓ Data preparation is the most time consuming yet arguably the most important step in the entire life cycle. The model will be as good as the data.

Step 4: Exploratory Data Analysis

- ✓ This step involves getting some idea about the solution and factors affecting it before building the actual model.
- ✓ The distribution of data within different feature variables is explored graphically using bar-graphs; relations between different features are captured through graphical representations like scatter plots and heat maps.
- ✓ Many other data visualization techniques are extensively used to explore every feature individually and combine them with other features.

Step 5: Data Modeling

- ✓ Data modeling is the heart of data analysis. A model takes the prepared data as input and provides the desired output.
- ✓ Data modeling step includes choosing the appropriate type of model, whether the problem is a classification problem, or a regression problem or a clustering problem.
- ✓ After choosing the model family, amongst the various s amongst that family, we need to choose the algorithms to implement and implement them carefully.

- ✓ We need to tune the hyper parameters of each model to achieve the desired performance.
- ✓ We also need to make sure there is a correct balance between performance and generalizability. We do not want the model to learn the data and perform poorly on new data.

Step 6: Model Evaluation and Tuning

- ✓ In this step, the model is evaluated for checking if it is ready to be deployed. The model is tested on unseen data, evaluated on a carefully thought out set of evaluation metrics.
- ✓ We also need to make sure that the model conforms to reality. If we do not obtain a satisfactory result in the evaluation, we must re-iterate the entire modeling process until the desired level of metrics is achieved.
- ✓ Any data science solution, a machine learning model, just like a human, should evolve, should be able to improve itself with new data, adapt to a new evaluation metric.
- ✓ We can build multiple models for a certain phenomenon, but a lot of them may be imperfect. Model evaluation helps us choose and build a perfect model.

Step 7: Model Deployment and Monitoring

- ✓ The model, after a rigorous evaluation, is finally deployed in the desired format and channel. This is the final step in the data science life cycle.
- ✓ One goal of a project is to change a process and/or make better decisions. We may still need to convince the business that our findings will indeed change the business process as expected.
- ✓ The importance of this step is more apparent in projects on a strategic and tactical level. Certain projects require us to perform the business process over and over again, so automating the project will save time.
- ✓ Continuously monitor the model's performance and make improvements as needed.

1.5 Trends, markets and applications of data science

Data science is a rapidly evolving field, and several key trends are shaping its future. These trends reflect advancements in technology, the increasing availability of data, and the growing need for data-driven insights across industries.

1. AI Integration: AI is becoming an essential component in data science workflows. From automating data processing to enhancing decision-making

with machine learning algorithms, AI integration allows companies to unlock deeper insights, automate processes, and improve operational efficiency. AI tools such as GPT (like ChatGPT) and other machine learning models are now being widely used for data analysis, content generation, and even customer interaction.

2. Edge Computing: As data generation moves closer to the source (e.g., IoT devices), edge computing is becoming crucial. By processing data on devices rather than in centralized cloud servers, companies reduce latency, improve real-time decision-making, and enhance privacy by keeping sensitive data local. This trend is especially significant in industries like manufacturing, healthcare, and automotive.

3. Natural Language Processing (NLP): NLP is transforming how businesses extract value from unstructured data, such as text and voice. NLP enables systems to understand, interpret, and respond to human language. It's being applied in chatbots, customer service automation, sentiment analysis, and content generation. With models like Llama, GPT, NLP's capabilities are continuously improving.

4. Explainable AI (XAI): As AI models become more complex, the need for transparency and interpretability increases. Explainable AI seeks to make machine learning models more understandable for humans, allowing stakeholders to trust the model's decisions. This is critical in sectors like healthcare, finance, and legal systems, where the consequences of AI decisions can be significant.

5. Automation and AutoML:
Tools that automate parts of the data science process, such as feature selection, model building, and hyperparameter tuning (AutoML), are making data science more accessible to non-experts. This is reducing the time it takes to deploy AI solutions.

Market and applications of Data science includes Healthcare(Predictive Analytics, Medical Imaging, Drug Discovery), Finance(fraud detection, Credit scoring), Retail and E-commerce(Recommnedations, Supply Chain Managament, Customer sentiment analysis), Manufacturing and Industries(aquality control),, Telecommunications(Network Optimization) etc.

✓ Predictive Analytics:

In various industries (e.g., healthcare, finance, marketing), data science helps predict future outcomes by analyzing historical data, improving decision-making processes.

✓ Natural Language Processing (NLP):

NLP techniques are used for sentiment analysis, chatbots, and automatic translation. This has applications in customer support, marketing, and healthcare.

✓ Customer Segmentation:

Using clustering techniques, businesses can segment their customer base, helping tailor marketing efforts, improve customer experiences, and optimize products.

✓ Fraud Detection:

Data science is used to detect fraud patterns in sectors like banking, insurance, and e-commerce. Machine learning algorithms analyze transaction data to identify anomalies.

✓ Image Recognition and Speech Recognition:

Image and speech recognition is a prominent application area for Data Science. When we upload an image on any social media, it automatic tagging suggestion uses image recognition algorithm, which is part of data science.

1.6 Tools and technologies in data science

Following are the overview of the tools and technologies in Data Science:

1. Programming Languages:

- ✓ Python: Widely used for its simplicity and extensive libraries for data analysis, machine learning, and visualization.
- ✓ R: A programming language focused on statistical computing and graphics, often used in academic and research settings.
- ✓ SQL: Essential for querying relational databases and handling large datasets.

2. Libraries and Frameworks:

- ✓ Pandas: Provides data structures and data analysis tools for handling structured data.
- ✓ NumPy: A library for numerical computing, including support for arrays and matrices.

- ✓ Scikit-learn: A comprehensive machine learning library that supports various algorithms for data mining and data analysis.
- ✓ TensorFlow: An open-source framework for machine learning and deep learning, particularly for neural networks.
- ✓ PyTorch: A deep learning framework similar to TensorFlow, known for its dynamic computation graph and strong community.

3. Visualization Tools:

- ✓ Matplotlib: A popular Python library for creating static, animated, and interactive visualizations.
- ✓ Seaborn: Built on top of Matplotlib, Seaborn provides a high-level interface for drawing attractive and informative statistical graphics.
- ✓ Tableau: A leading data visualization tool that allows users to create interactive, shareable dashboards.
- ✓ Power BI: A business analytics tool from Microsoft for creating interactive reports and visualizations.

4. Big Data Tools:

- ✓ Hadoop: An open-source framework for distributed storage and processing of large datasets using a network of computers.
- ✓ Spark: A fast, in-memory data processing engine with support for a wide range of analytics tasks, including machine learning and graph processing.

5. Cloud Platforms:

- ✓ AWS (Amazon Web Services): Offers scalable cloud storage, computing power, and machine learning services.
- ✓ Azure: Microsoft's cloud platform, offering a suite of tools for building, deploying, and managing applications and services.
- ✓ Google Cloud: Provides tools for computing, data storage, and machine learning, including services like BigQuery.

6. Collaboration Tools:

- ✓ Jupyter Notebook: An open-source web application for creating and sharing documents that contain live code, equations, visualizations, and narrative text.
- ✓ Google Colab: A cloud-based notebook environment that allows you to write and execute Python code in a collaborative manner.

1.7 Data scientist and their roles

A Data Scientist is a professional who uses advanced analytical techniques, algorithms, machine learning, and statistical tools to extract insights and knowledge from structured and unstructured data. They are tasked with solving complex problems and making data-driven decisions in various industries, from healthcare and finance to marketing and technology. They are responsible for developing Operational Models.

Data scientist roles and responsibilities include

1. Data Collection and Cleaning:

- ✓ Data Acquisition: Gathering data from various internal and external sources (databases, APIs, etc.).
- ✓ Data Preprocessing: Handling missing values, outliers, and duplicates.
- ✓ Data Transformation: Converting data into suitable formats and types for analysis (e.g., normalization, encoding categorical variables).
- ✓ Data Quality Assurance: Ensuring that the data is clean, accurate, and consistent for reliable analysis.

2. Model Building and Evaluation:

- ✓ Model Selection: Choosing the right machine learning algorithms (e.g., regression, classification, clustering).
- ✓ Training Models: Applying the chosen algorithms to the data and training models to identify patterns and make predictions.
- ✓ Hyperparameter Tuning: Adjusting the parameters of the models to optimize their performance.
- ✓ Model Evaluation: Assessing model performance using metrics such as accuracy, precision, recall, F1 score, and cross-validation to ensure the model's reliability and generalizability.

3. Collaborate with Business and IT teams

ASSIGNMENT:

1. What is data science? Why the learning of data science is important?? Explain in detail.

2. What are the key components of a modern data ecosystem, and how do they work together?

3. Differentiate between Data Scientist ,Data Analyst and Data Engineer.

4. With the help of diagram describe lifecycle of data science.

5. What are applications of data science? Explain.

6. Explain the tools and technologies widely used in data science.

7. Explain Data Scientist and their roles and responsibilities.

(Submit within 1 week after the assigned date)