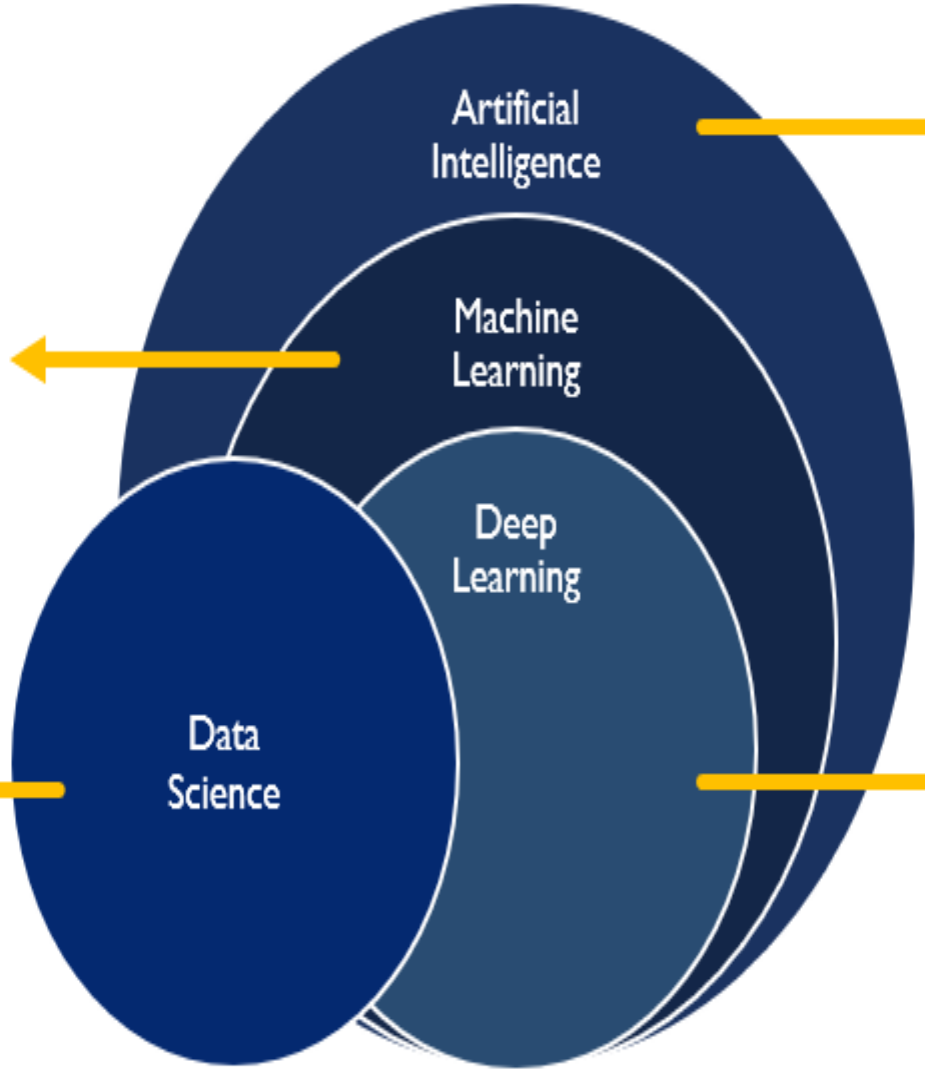
Data Science Workshop

What is Data Science?

- **Definition**: Data Science is the field that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.
- AI vs ML vs DS
- **Importance**:
  - Transforms raw data into actionable insights.
  - Helps businesses and industries make informed decisions.
  - Drives innovation in fields like healthcare, finance, and marketing.

Artificial Intelligence means creating smart machines to mimic human behavior

Machine Learning is a subset of AI, and builds a model based on training data to make predictions

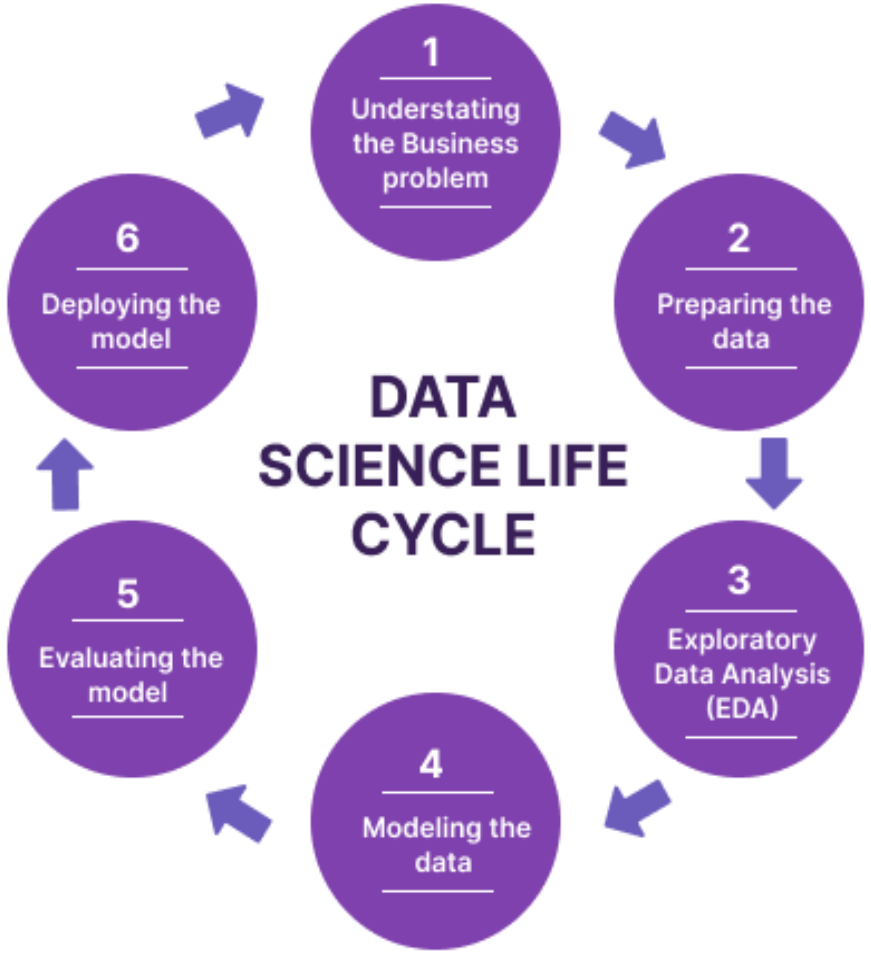


Deep learning is a subset of ML, a class of ML algorithms to solve complex problems.

Data Science is a subset of AI. It is an area of statistics, scientific methods, etc. to extract meaning and insights from data..

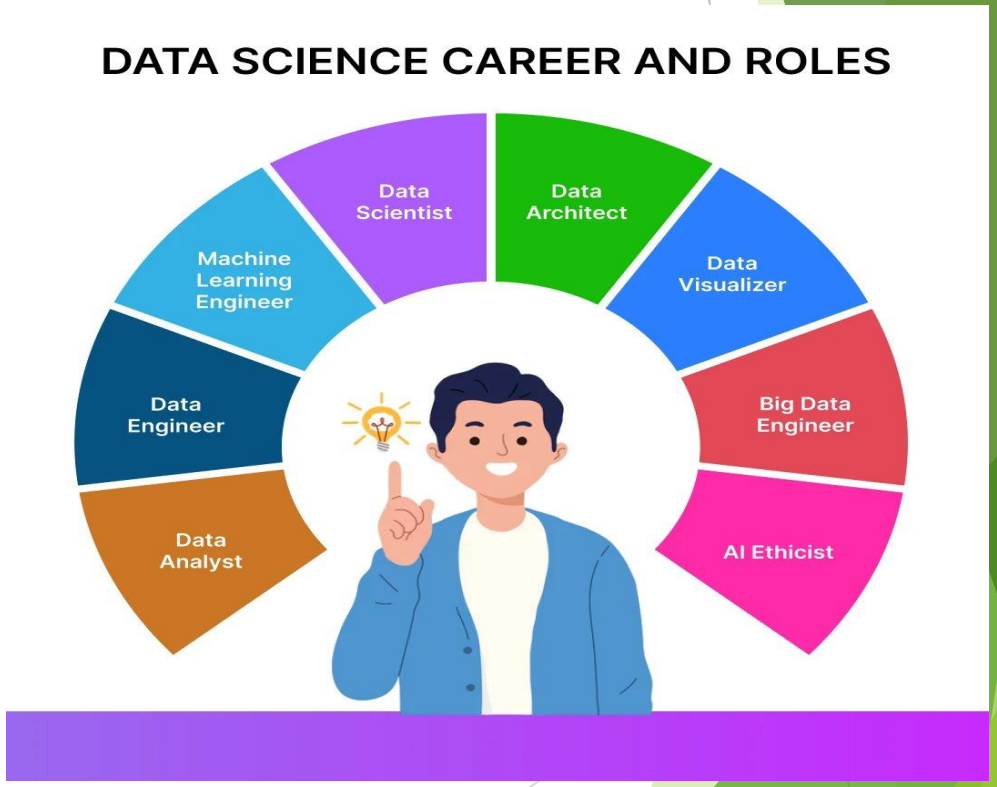
Data Science Lifecycle?

- **Problem Definition**: Understanding the business problem.
- **Data Collection**: Gathering relevant data from various sources.
- **Data Cleaning**: Removing inconsistencies and handling missing data.
- **Exploratory Data Analysis (EDA)**: Visualizing data and uncovering patterns.
- **Model Building**: Applying machine learning or statistical models.
- **Model Evaluation**: Assessing model performance.
- **Deployment**: Implementing the model in a production environment.



Key Roles in Data Science

- **Data Scientist**: Focuses on creating models, interpreting data, and generating actionable insights.
- **Data Analyst**: Focuses on analyzing data and providing visual reports.
- **Data Engineer**: Works on designing and managing data pipelines and infrastructures.
- **Machine Learning Engineer**: Specializes in deploying machine learning models at scale.
- **Business Analyst**: Works closely with stakeholders to understand business requirements and translate them into data problems.



Essential Tools in Data Science

- **Python**:
  - Libraries like Pandas, NumPy, Matplotlib, and Scikit-Learn for data manipulation and analysis.
  - Python is widely used due to its simplicity and rich ecosystem for data science.

- **R**:
  - A programming language used primarily for statistical analysis and visualization.
  - Libraries like ggplot2, dplyr, and caret are popular in R.

- **SQL**:
  - Structured Query Language is essential for querying relational databases.
  - Helps in data extraction, manipulation, and aggregation.

Outcome of Workshop

- **Skills Acquired**:
  - Understand the Data Science lifecycle from problem definition to model deployment.
  - Clean and preprocess data to handle inconsistencies, missing values, and outliers.
  - Perform Exploratory Data Analysis (EDA) to uncover insights and patterns.
  - Apply machine learning algorithms to solve real-world problems.
  - Effectively communicate data-driven insights to diverse audiences.

- **Tools and Libraries**:
  - Proficiency in **Python** programming.
  - Expertise in libraries like **NumPy**, **Pandas**, **Matplotlib**, and **Seaborn** for data analysis and visualization.

- **Capstone Project**:
  - Students will complete a capstone project on Titanic survival prediction
  - Github repo: GitHub Repository.

Setup

- **Colab / Jupyter notebook**: Best for writing and debugging python code.
- **Vscode / pycharm**: Code editor for writing production level code.
- **Github Account**: Use for maintaining the code.